Statistics 480

Statistical Computing Applications

Project 2 – Server survey

April 2007

Introduction

During summer 2006, Professor William Michael Lynn of the School of Hotel Administration at Cornell University, Ithaca, NY, collected tipping information from more than 2,400 waiters and waitresses from all over the world. Through an online survey, he collected data not only on waiters/waitresses habits and characteristics, but also on restaurants where they work and their clientele. With seventy-seven variables and more than 2,400 observations, this dataset is a marvelous source of information to study the complex relationships between the waiters'/waitresses' behaviors and the tip they receive, to observe the tipping trends from a geographical perspective, to evaluate the influence of ethnical parameters on tipping habits …

In this study, we will focus our analysis on the servers. We will develop a guide that will provide servers information on how they should behave to increase their tip. We are conscious that an extensive study of these elements would require the use of all available variables (especially those in relation with customers) but this would make the report more complex whose primary goal is to provide useful and accessible information.

Thus, we will first provide a description of the global dataset. This could be useful to future investigators that may choose to focus their analysis on a different topic. We will then conduct the analysis of three main elements. Primarily, we will investigate relationships between the percentage of tip from a world perspective by comparing tips received in the USA and Canada where tipping is a cultural element with the tips received in other countries. We will complete this geo-tipping analysis by focusing on the tipping habits in the USA by mapping tips and correlating them to the gross state product. The second part of this study will focus on the servers' behaviors. We will investigate the impact of the gesture, the attitude, and writing on the bill on the percentage tip and the relationship between customers' ethnicity and the tip. Finally, we will try to evaluate the impact of the classiness of the restaurant on the tip and study the case of one restaurant in more depth. The conclusion will be designed as a server guide that summarizes elements that impact the more the tip, the dos and don'ts.

Material and Methods

As mentioned in the introduction, the dataset contains 77 variables and 2483 observations. There are four types of variables: (1) 54 discrete or integer variables, (2) 12 factors, (3) 9 continuous or numerical variables, and (4) 1 logical variable. These variables can be regrouped into three main categories. One that describes the restaurant (type of restaurant, localization, clientele ...). The second category describes the servers (experience, habits ...). The last category regroups tipping information (amount, form …). A list of the variables and their description is provided in appendix 1. It would take too long to present all of them here. This is why we will just describe four important one that will give to the reader an idea of the type of information available in this dataset. These variables are the gender, the ethnical group, the age, and the experience of the servers.

This report was developed using R 2.4.1 (The R Development Core Team). The R code used to develop calculations is provided in appendix 2. In the report, we will only provide graphical outputs and their interpretations.

Results

General description of the dataset

Distribution of the servers' genders

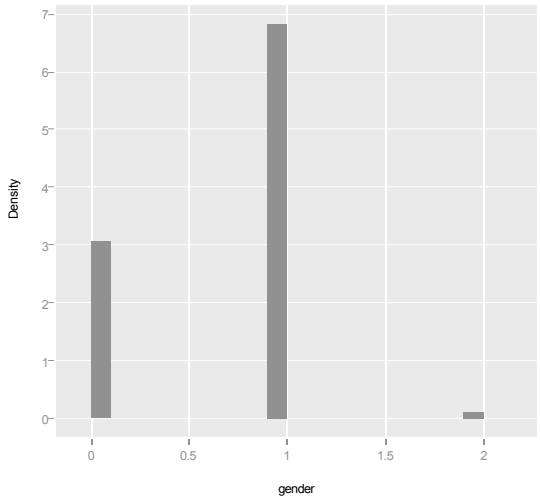


Figure 1: Server's gender distribution

This graph displays the gender of the servers. From information in the dataset, we were able to determine that 0 corresponds to males and 1 to females. However, one can see that several observations present a value of 2. This value has no meaning and needs to be removed. When investigating the relation between gender and the percentage of tip received, we get the following representation:

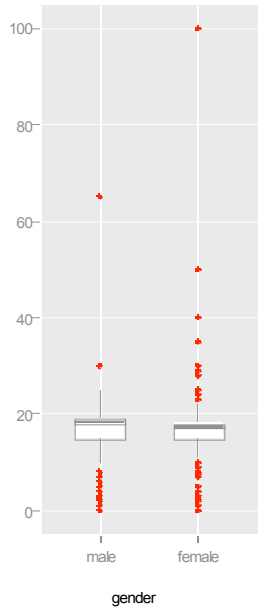


Figure 2: Relation between tip and gender

We can see that the range of tip for males is larger than the one for females. We can also observe that the median value is higher for males than for females. An interesting element would be to investigate if the difference between both genders is significant. Table 1 presents the average tip and standard deviation for each gender.

Table 1: Mean and standard deviation of the percentage tip by gender

Gender	Mean	Standard deviation
males	16.51	4.65
females	16.03	5.35

We can see that there is a difference between means of both genders and that the standard deviation for females is larger than for males. This last element is certainly due to the value of 100% of tip reported by one waitress. To determine if the difference between percentage tip is significant, we used a t-test at 95% confidence interval. With a p-value of 0.03, we can reject the null hypothesis and conclude that on the entire dataset, there is a significant difference between both genders and that the tip average for males is higher than for females.

Distribution of the servers' ethnical groups

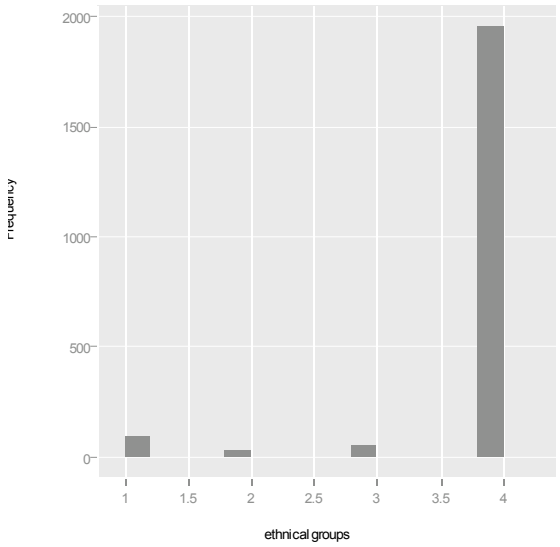


Figure 3: Server's ethnical group

The dataset presents four ethnical groups: Asian (1), Black (2), Hispanic (3), and White (4). The graph shows clearly that the proportion of White servers is the largest. Table 2 presents the proportion of each ethnical group.

Table 2: Proportion of each ethnical group in the dataset

	Asian	Black	Hispanic	White
Proportions	1.24%	0.70%	2.02%	96.05%

Thus, with more than 96% of the observations from the White population, conclusions that we will draw will more likely be applicable to White waiter/waitress but may not be completely accurate for other ethnical groups.

When looking at the relation between the percentage tip and the ethnical groups, we can obtain the following graph.

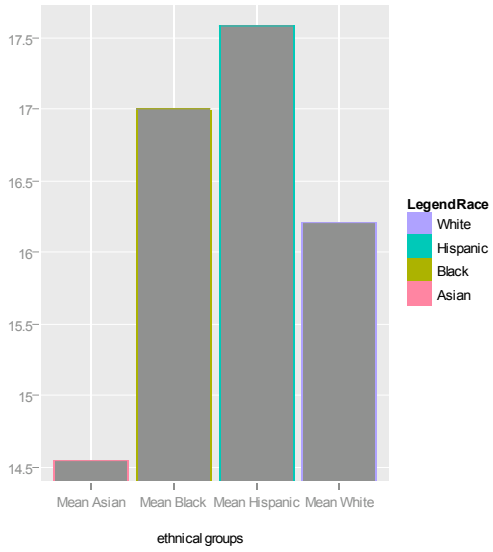


Figure 4: Mean tip per ethnical group

Here, we can see that the average tip for Black and Hispanic servers is much higher than for White. This result is certainly related to the proportion of servers in each category. With 96% of the data, we will be more likely to believe that the mean tip for a White server is around 16.2% while with only 0.70% and 2.02% of the observation for respectively Black and Hispanic, results cannot be generalized.

Distribution of the servers' age and experience

The age and the experience of a server is an important parameter that is most likely to influence the tip amount. The next plot shows the relation between the age and the experience as reported by servers through the online survey.

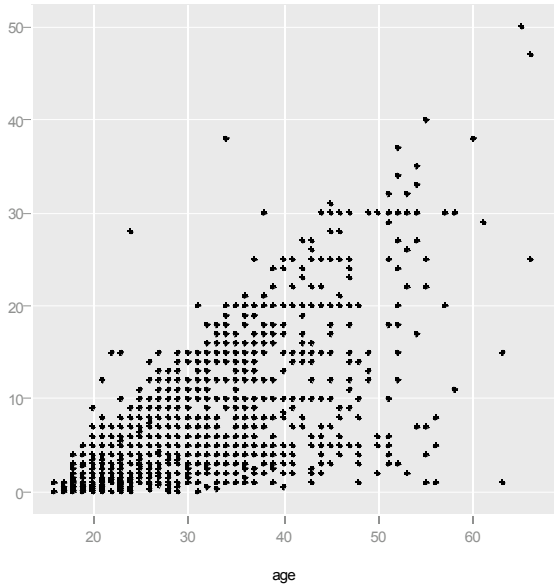


Figure 5: Relation between experience and age

We can see that some servers reported erroneous values for either age or experience because this graph shows for instance someone 34 years old and having been server for 38 years. This graph also shows that most of the servers are between 13 and 30 years of age. To correct the dataset, we removed people that presented a difference between the age and the experience lower than 13.

The next two graphs show the relation between the percentage of tip and the age and the experience respectively.

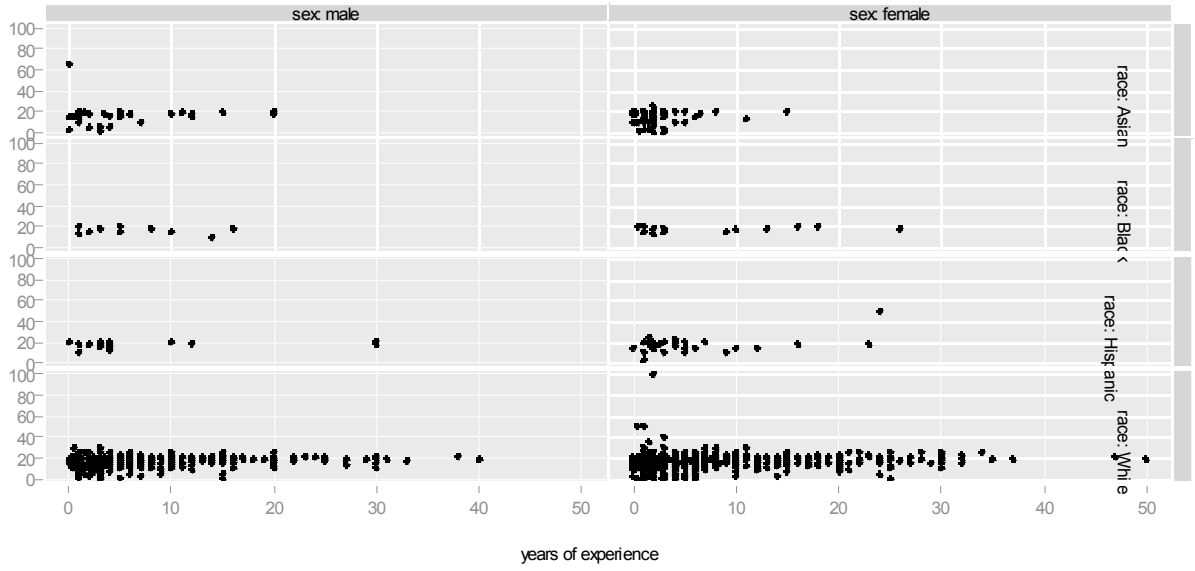


Figure 6: Relationship between experience and tip as a function of gender and ethnical group of the waiter

We can observe that an upper limit for tips exists. This limit of 24—26% applies to all servers, with or without a lot of experience. We can see in some cases that people with limited experience (lower than 5 years) and mostly women reported values between 40 and 55 %. These

extra-tips exist only for White females. For males, an Asian reported a tip of 64%, but tips higher than 26% are rather rare and only for White males.

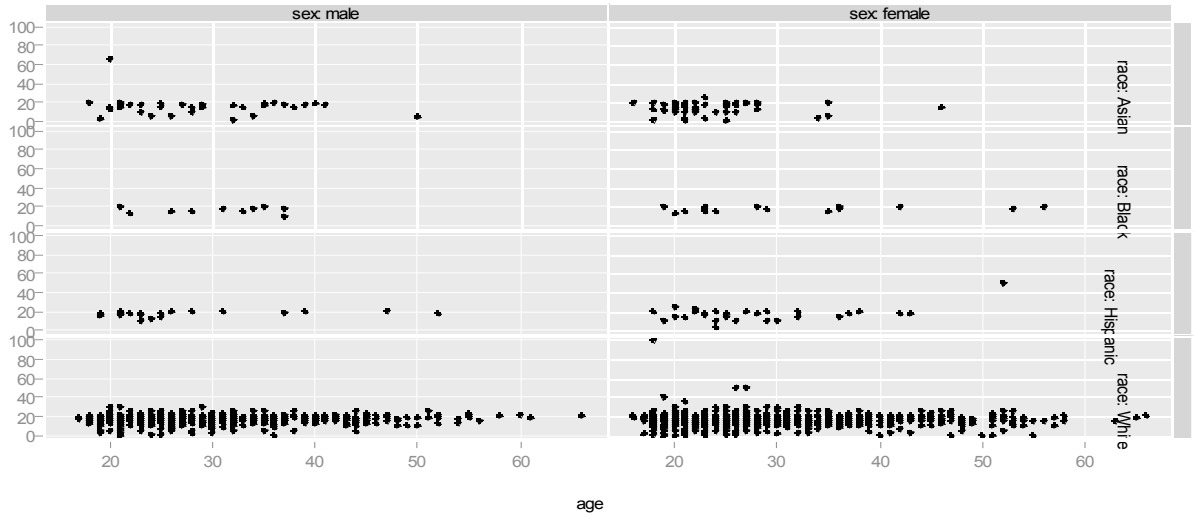


Figure 7: Relationship between age and tip in function of the gender and ethnical group of the waiter

For this second graph reporting tips as a function of the age, similar observations can be made. This plot brings the fact that only servers younger than 30 years old have extra-tips (except one 53 years old Hispanic female). We can also see that no matter the age, the sex or the experience, some servers reported a tip amount of 0%. This is certainly due to the fact that the dataset contains information from people from countries where the tip is not a common practice because the salary of the servers is included in the price of the meal.

The following plot shows the percentage tip as a function of age, sex, and ethnical group for the US servers only.

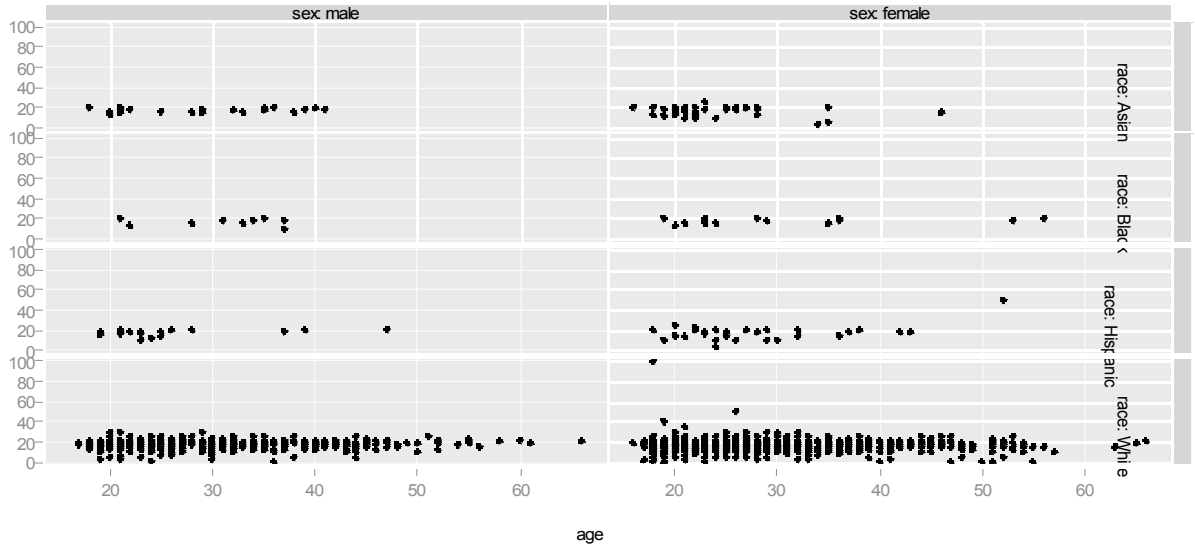


Figure 8: Relationship between age and tip in function of the gender and ethnical group of the waiter

Surprisingly, values with a 0% tip still exist. We can see two possible reasons. The first would be that servers did not enter any value when completing the survey or that they work in a restaurant where there is not tip.

When analyzing observations that have a 0% tip, we found eight observations among which two are fast food (Subway, Mc Donald's). Others appear as independent restaurants that do not belong to a chain. Their values were considered as outliers and were removed from the dataset.

As a conclusion of this presentation of the dataset, we can say that the original dataset presents a lot of outliers or erroneous data entries. By the presentation of those four variables (gender, ethnical group, experience and age), we were able to remove numerous outliers. It does not mean that the dataset is not clear from other erroneous values and attention will be paid to this element when investigating other variables. Thus, we were able to show that males appear to earn significantly more than females. The dataset containing most records from White people, our conclusions will be more likely trustable for a generalization to White servers but not for Asian, Black, and Hispanic servers. In terms of age and experience, we found that an upper tip limit exists at about 24-26%, but that young White females present some higher observations.

We will now investigate more in depth features of the dataset by starting with describing the tip habits of tipping and non-tipping countries.

Global tipping habits

Tipping is not a common practice across countries. North America tips while Europe, Asia and the South Pacific countries do not. Tipping in North America is mandatory while is it considered as a sign of customer satisfaction in other countries. Thus, we compared the average tip in the US and Canada with the average tips in the rest of the world.

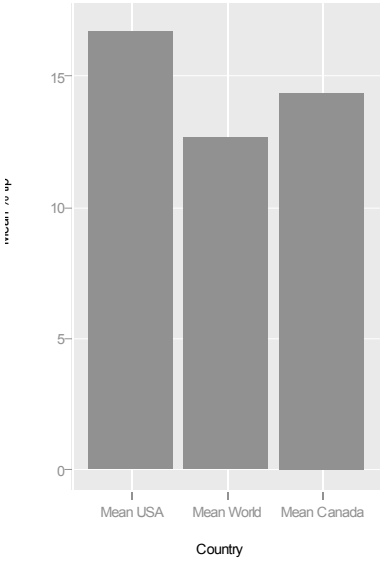


Figure 9: Average tip between tipping and non-tipping countries

Surprisingly, the average tip for non-tipping countries is rather high. It can be explained by the fact that the dataset contains few records from these countries (6% of the observations) and that servers that usually receive tip are the only one that posted the data. The difference between Canada and the US can be explained by several facts. The first is that only 4.5% of records are from Canadian servers and that it is not representative of the Canadian reality. A second reason could be that Canada is more protective towards servers than the US government and that the base server's salary in Canada is higher.

Because 89.5% of the records are from US servers, all conclusions about tipping habits will only be applicable to the US. Thus, the following investigations will be restricted to US customers and servers. We will start our analysis by comparing tipping habits among US states.

Relation between localization, richness of a state and the tip amount

All guides recommend their readers the best location to go for practicing an activity. In this section, we will study the average tip per state and try to draw conclusions by relating the tip amount with the gross state product.

The map below presents the tip average per US state.

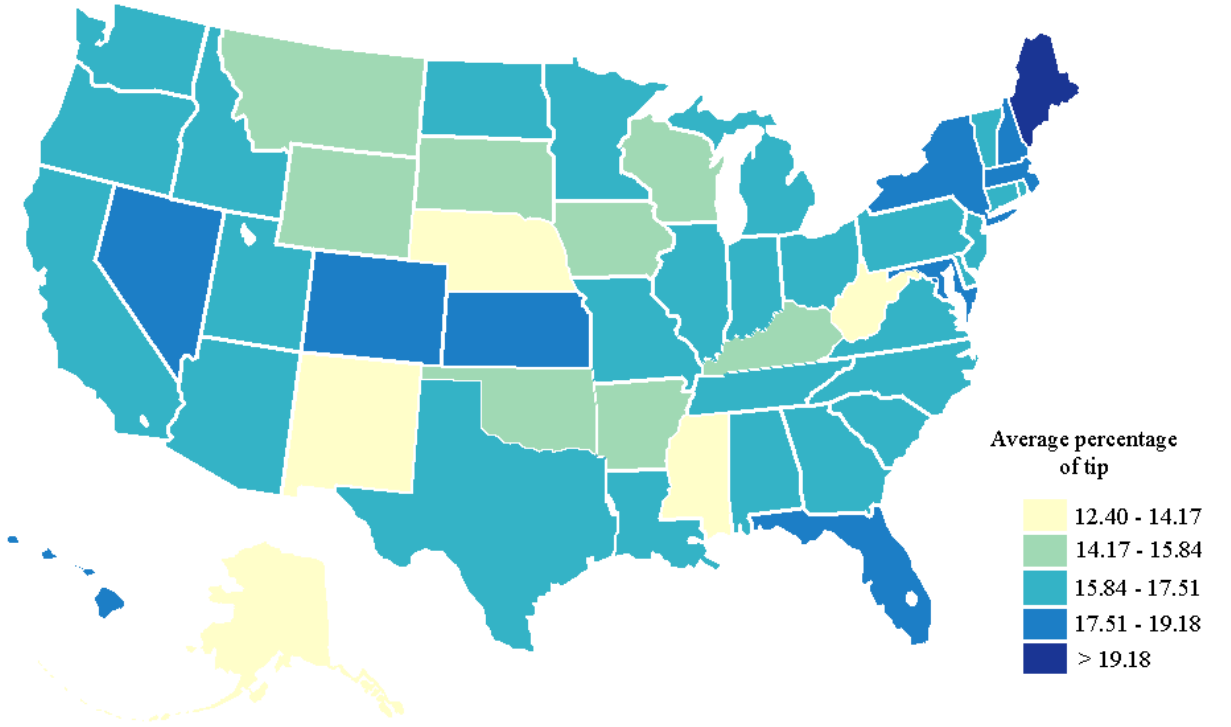


Figure 10: Tip amount per state

We can see that the states that tip the most are those on the east coast (north), Florida, Nevada, Hawaii, Colorado, and Kansas. This fact is understandable for the north east states because they group the biggest cities of the USA; they are economic centers with a high population. High tips in Florida, Colorado and Hawaii are understandable because they are large tourist places for US and non-US citizens. The state of Nevada with Las Vegas has a high tip average certainly because of the money generated by gambling and games. However, there is no obvious reason that can explain why the state of Kansas is among the highest tipping states.

On the contrary, when looking at the states that tip less, we see that they are mostly in the center of the country. They are all important agricultural states with no major cities except Chicago or Minneapolis. Some states appear behaving differently such as North Dakota that tips higher than most of the states in the Midwest.

To investigate this element more in depth, we will try to correlate the gross state product (GSP) of each state with the tip amount. The GSP is a measurement of the economic output of a state. It is the sum of all value added by industries within the state (wikipedia.com).

The next table presents the GSP for each state, sorted in decreasing order.

Table 3: States ranked by GSP

Rank	State	Percentage tip (%)	GSP($)	Rank	State	Percentage tip (%)	GSP($)
1	dc	18.50	139,849	27	pa	17.13	37,416
2	de	17.33	63,004	28	ia	15.23	37,323
3	ak	13.70	54,713	29	oh	16.84	37,133
4	ct	16.62	52,149	30	in	16.22	36,850
5	ma	18.43	48,803	31	tn	16.48	36,782
6	wy	15.00	47,623	32	mi	16.02	36,282
7	nj	17.32	47,242	33	ks	18.26	36,188
8	ny	17.89	47,031	34	mo	16.00	35,740
9	mn	16.71	44,073	35	nd	17.50	35,662
10	co	17.58	43,764	36	la	17.42	35,544
11	va	16.68	43,713	37	vt	17.00	35,401
12	nv	17.97	42,498	38	fl	17.65	35,051
13	ca	17.13	42,386	39	ut	17.25	34,100
14	il	16.08	41,987	40	az	16.56	33,841
15	md	17.87	41,483	41	nm	13.30	33,444
16	wa	17.40	40,774	42	me	20.83	32,896
17	tx	16.83	40,193	43	ky	14.43	32,112
18	nh	17.83	40,090	44	ok	14.29	31,740
19	hi	18.00	39,804	45	sc	16.71	31,323
20	ne	13.48	38,902	46	al	15.93	31,239
21	ri	16.83	38,747	47	id	17.42	31,186
22	sd	14.83	38,539	48	ar	14.69	30,077
23	ga	16.16	38,094	49	mt	14.35	29,758
24	nc	16.78	37,933	50	wv	14.00	27,532
25	wi	15.48	37,746	51	ms	12.40	26,582
26	or	17.02	37,483				

We can see that except the District of Columbia that has a GSP higher than $100,000, all other states have GSP between $25,000 and $65,000 per capita. A quick look at the table shows that there is no apparent relation between the tip amount and the richness of a state. The following graph details the trend.

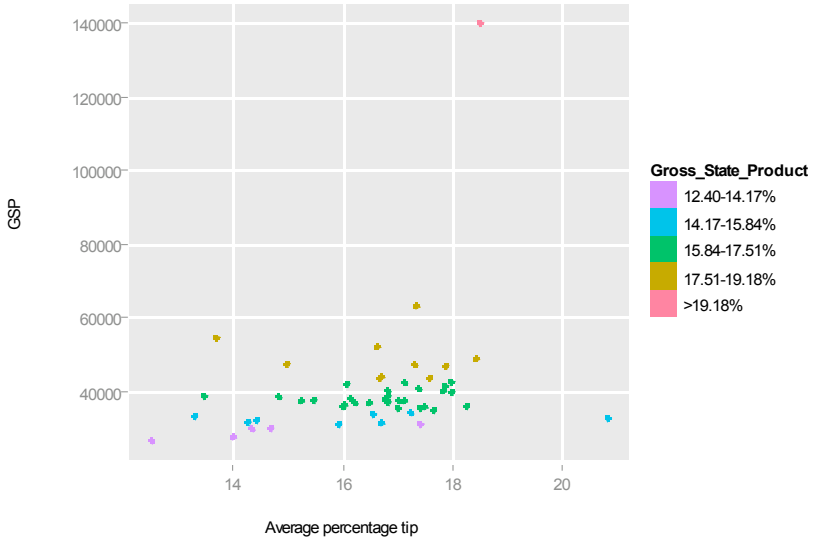


Figure 11: Relation between GSP and the average tip per state

This plot does not show any major trend except a correlation between the GSP and the average percentage tip for states belonging in the lowest tip category (r = 0.88). The fact that we do not see any major trend can be related to the upper tip limit that was determined earlier. In the previous section, we determined a tip limit of 26-28%. When working with tip means in the USA, we found an upper limit of 18.5%. Thus, no matter the richness of the people in a state, the tip amount will most likely have an average of 18.5% or lower. One exception exists; Maine belongs in the state category with the second lowest GSP but tips higher than any other state with an average percentage tip of 20.8. There does not appear to exist any reason for this fact except maybe a cultural element in this state or a more restrictive law for restaurant regarding the server's income that require bigger tips from customers.

As a conclusion, a server wanting to travel and work should target states of the north east or the west coast, Nevada, Colorado, and Florida. He/she must not base his/her choice on the prosperity of a state but more on the habits of its inhabitants.

To carry on the analysis of the dataset, we will now focus on the servers' traits and their influence on the percentage tip.

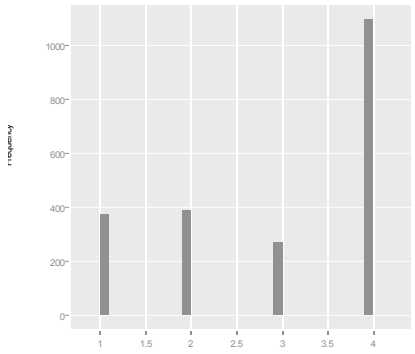
Influence of server traits on the amount of tip

We decided to study the influence of different traits of servers on the amount of tip. One would expect that server traits such as engaging in a conversation with the customers, thanking or complementing the customers, sharing jokes, etc. would get them higher tips. The dataset contains 14 variables describing the server traits. These traits were assigned values from 1 to 4 describing how often the server presented each trait (1- "Never", 2- "Sometimes", 3- "Often" and 4- "Always").

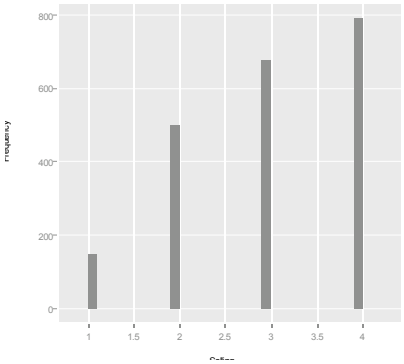
We decided to consider six of these traits that implied that the server engaged in a conversation with the customer:

1. Servers that introduced themselves (intro)
2. Servers that tried suggestive selling (selling)
3. Servers that shared jokes (jokes)
4. Servers that addressed customers by their name (customer_name)
5. Servers that complemented the customers (weather)
6. Servers that thanked the customers (thanks)

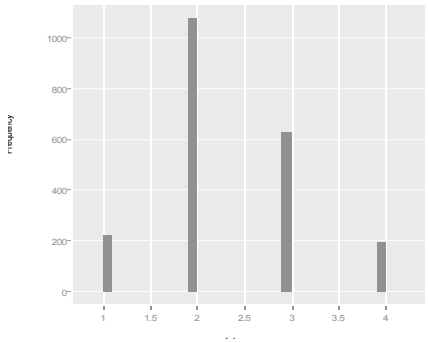
First of all, we needed to make sure that the dataset is clean and there are no missing values for these variables. The following histograms investigate the completeness of the data.



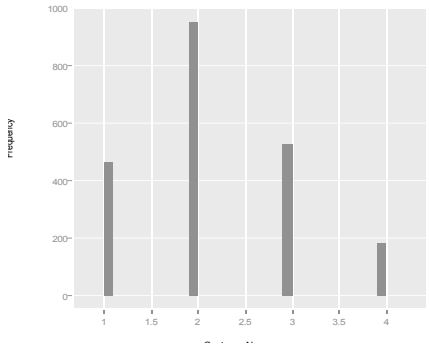
Introduce themselves



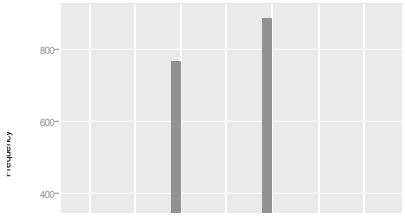
Suggestive Selling



Jokes



Customer Name



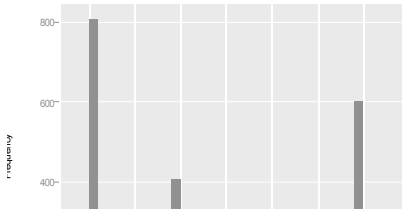


Figure 12: Frequency distribution of six server traits

Histograms show that the data is clean and there is no outlier. Next, we investigated the relationship between average tips received by the servers based on the six selected traits. Figure 12 shows the relationship between average tip and frequency of six server traits.

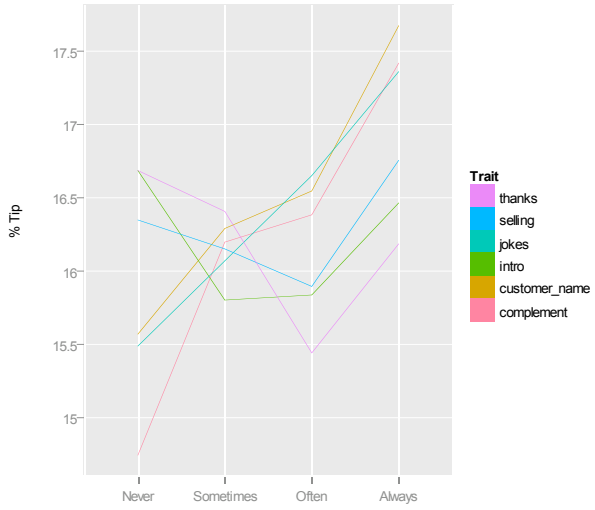


Figure 13: Relationship between tip and frequency of server traits

The above figure shows that the servers who always share jokes, complement their customers and address them by their name get higher tip than the ones who show these traits less often. Servers who always try suggestive selling also get considerably higher tip than others. Those who never introduce themselves and never thank the customers get higher tip than servers who frequently display these traits. These findings are very interesting but quite surprising. It would be a common sense to believe that a server introducing himself and thanking clients is more likely to get higher tip than those who appear less friendly. We can also observe that the variable "Often" presents some strange patterns compared to the general trend. These facts are hardly understandable except if when filling the survey, some servers understood 1 as "Always" and 4 as "Never" and not 1 as "Never" and 4 as "Always". However,

this element can neither be proved nor corrected because we do not have any way to confirm our hypothesis but this could be the reason for the unexpected results that will follow.

Thus, we decided to investigate a combination of traits categorized by gender and ethnical group for both servers and customers. Believing that the introduction of the server to the customer is one of the most important factors that impact the tip, we grouped the variable "introduction" with each of the other variables. The following graphs show the influence of combination of two server traits on tip for both male and female servers.

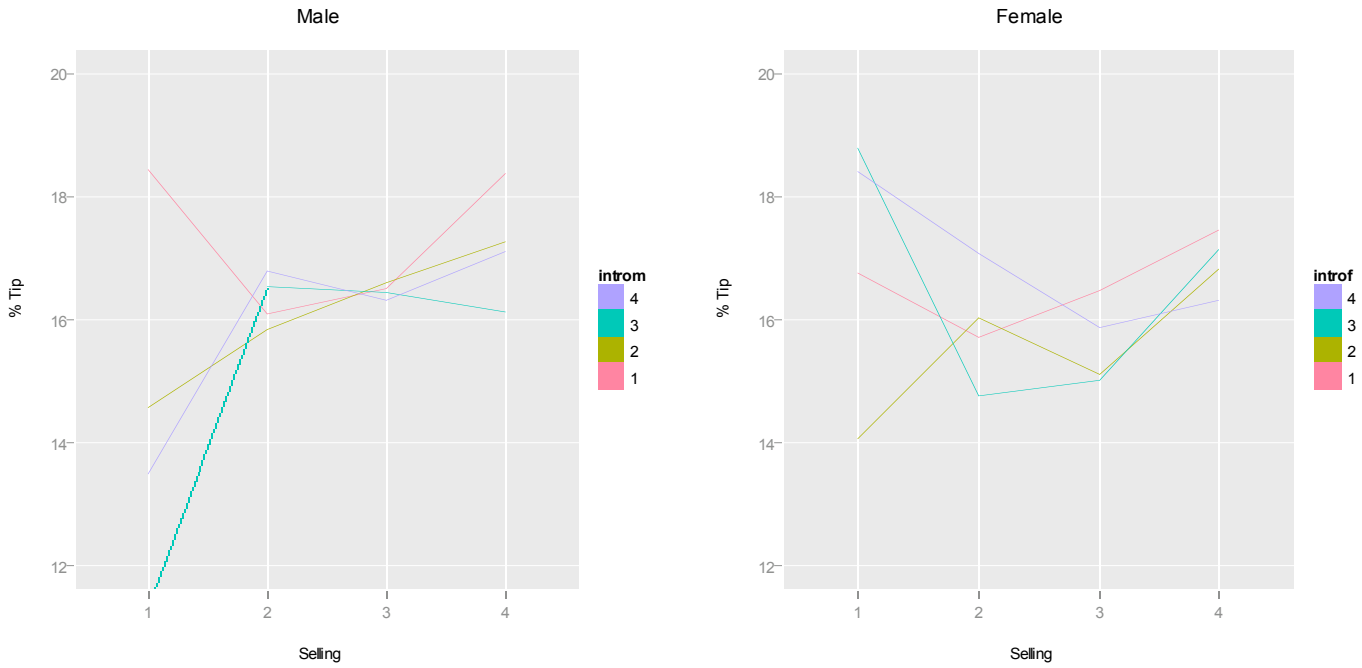


Figure 14: Relationship between tip and frequency of Introduction and Suggestive Selling for male and female servers

Figure 14 shows the influence of introduction combined with suggestive selling on the percentage of tip for male and female servers. The average tip for male servers was highest when they always tried suggestive selling and never introduced themselves. The contrary was observed for females. When often introducing themselves and never trying suggestive selling both females and males had lowest tip.

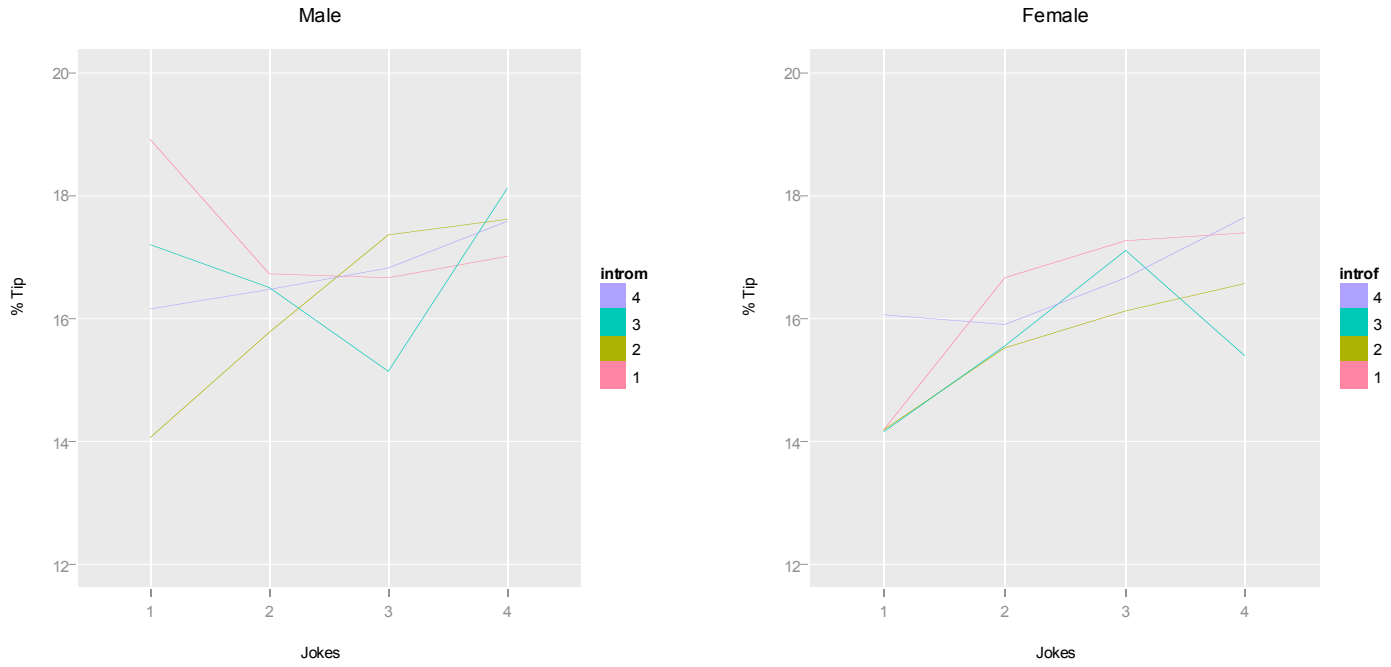


Figure 15: Relationship between tip and frequency of Introduction and Sharing Jokes for male and female servers

Figure 15 shows the influence of introduction combined with sharing jokes on the percentage of tip for male and female servers. The average tip for male servers was highest when they always shared jokes and never introduced themselves. Although, for female servers, no major trend could be observed except that when they often introduced themselves and never shared jokes, the average tip was lowest.

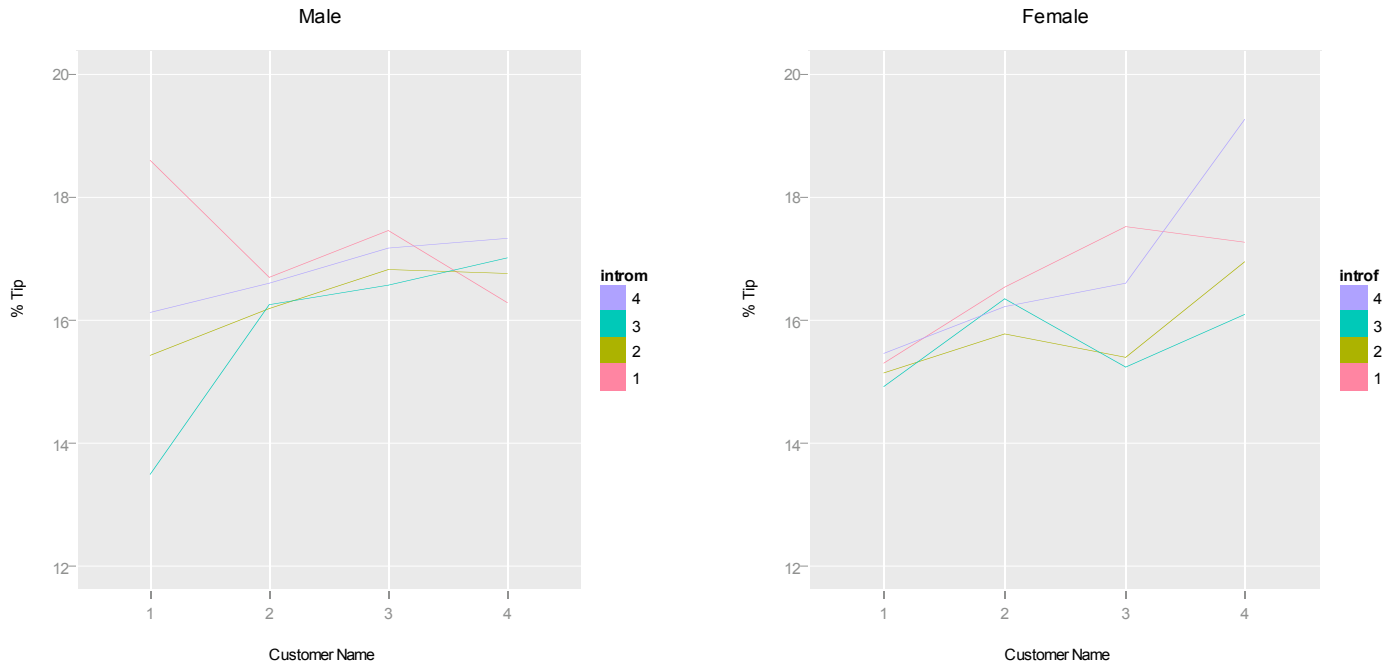


Figure 16: Relationship between tip and frequency of Introduction and Addressing Customers by their name for male and female servers

Figure 16 shows the influence of introduction combined with addressing customers by their name on the percentage of tip for male and female servers. The average tip for male servers was highest when they often used customer names and sometimes introduced themselves. While, for female servers, the average tip was highest when they both always used customer names and introduced themselves. It also appears for males that when they never addressed customers by their name, the tip is lowest but for females, the tip was lowest when they often do it.

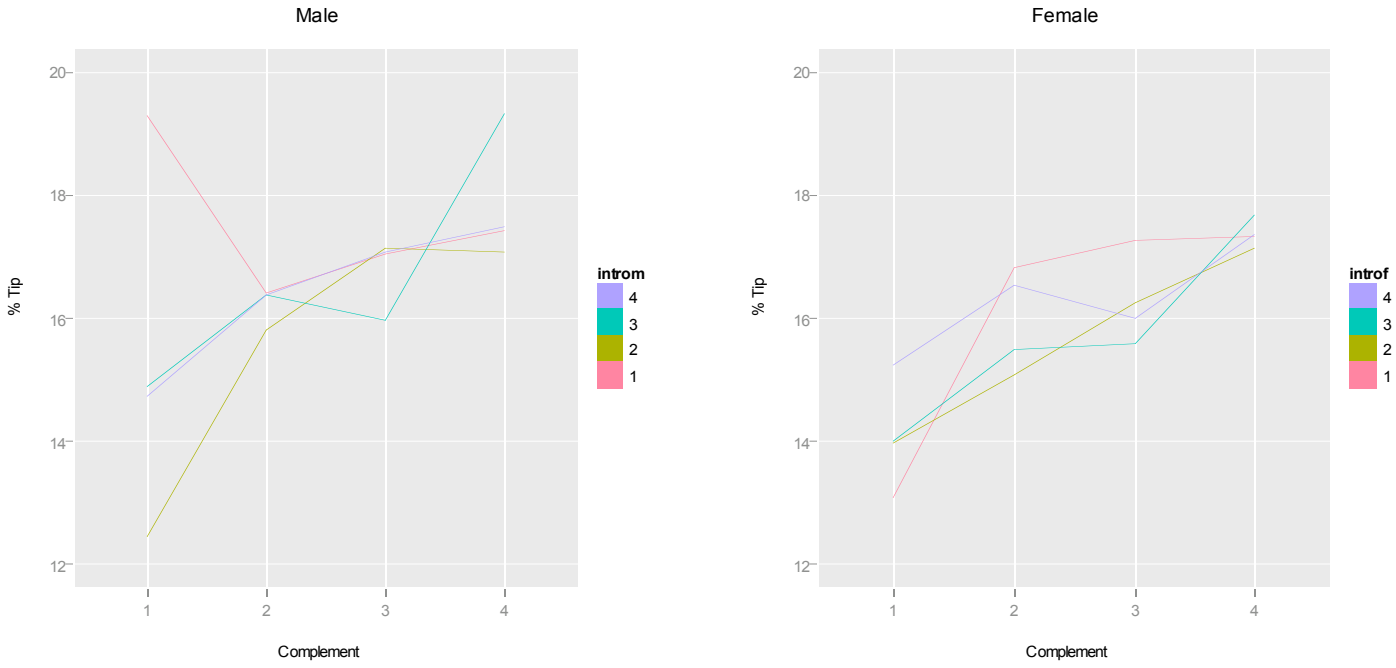


Figure 17: Relationship between tip and frequency of Introduction and Complementing the Customers for male and female servers

Figure 16 shows the influence of introduction combined with complementing the customers on the percentage of tip for male and female servers. The average tip for male servers was highest when they always introduced themselves and always complemented the customer. While, for female servers, the average tip was highest when they often introduced themselves and always complemented the customer. Both males and females had lowest tip when they never complemented the customer.

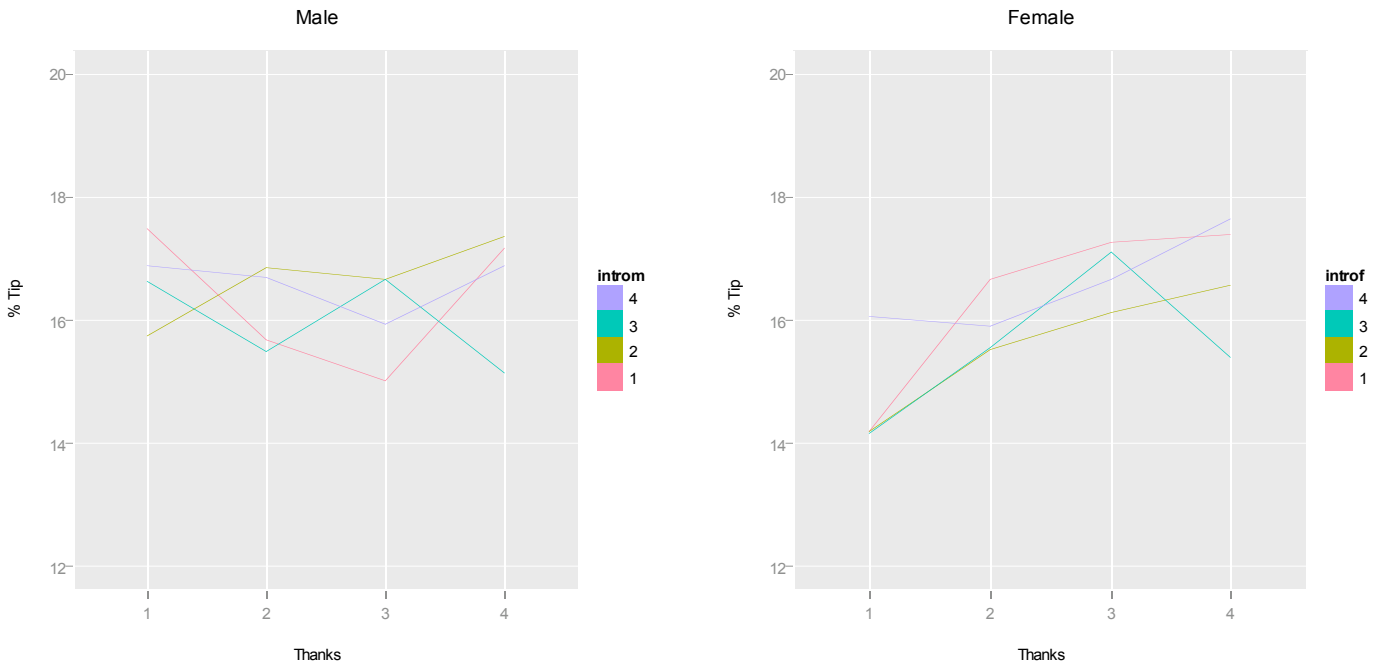


Figure 18: Relationship between tip and frequency of Introduction and Thanking the Customers for male and female servers

Figure 18 shows the influence of introduction combined with thanking the customers on the percentage of tip for male and female servers. The average tip for male and female servers was highest when they never introduced themselves and always thanked the customer. However, it shows that when they both often thanked the customer, they get a lower tip.

From these last results, we can see more clearly that our assumption stating that while filling the survey, they confused 1 and 4 might be correct. However, in the case that the dataset is correct, we were still able to show that the servers that always tried suggestive selling, shared jokes, used customer names, complemented and thanked the customers had higher tips. Depending on the sex of the server, the impact of these variables may vary.

We decided to test the significance of our results. t-tests were conducted to determine if the difference in average tip of servers who always practice each of the six traits is significant at 95% confidence interval. Table 4 shows the p-values for each set of traits.

Table 4. P-Values for tips of servers with different traits (Frequency = Always)

Server Trait 1	Server Trait 2	p-value
Introduction	Selling	0.320
Introduction	Jokes	0.003
Introduction	Customer Name	0.223
Introduction	Complement	0.010
Introduction	Thanks	0.329
Selling	Jokes	0.019
Selling	Customer Name	0.379
Selling	Complement	0.060
Selling	Thanks	0.063
Jokes	Customer Name	0.647
Jokes	Complement	0.498
Jokes	Thanks	0.001
Customer Name	Complement	0.967
Customer Name	Thanks	0.119
Complement	Thanks	0.002

The average tips of servers who always practiced the following traits were significantly different from each other:

1. Introduced themselves and shared jokes.
2. Introduced themselves and complemented the customers.
3. Tried suggestive selling and shared jokes.
4. Shared jokes and thanked the customers.
5. Complemented and thanked the customers.

Here again, the results must be considered with care because of the possible mistake in reporting the data by the servers.

The investigation was carried on by analyzing the impact of relation between the ethnical group of the customer and servers on the tip.

Influence of the ethnicity of customers on Server's Tip

We decided to investigate the influence of ethnicity of customers on the tip of servers from different ethnical groups. However, because 96.05% of the servers are White, we do not have enough data to provide helpful information for servers from other ethnical groups (a total of 125 observations for ethnical groups other than Whites). Thus, the next graph shows the average tip of White servers depending on the proportion of customers from various ethnicities.

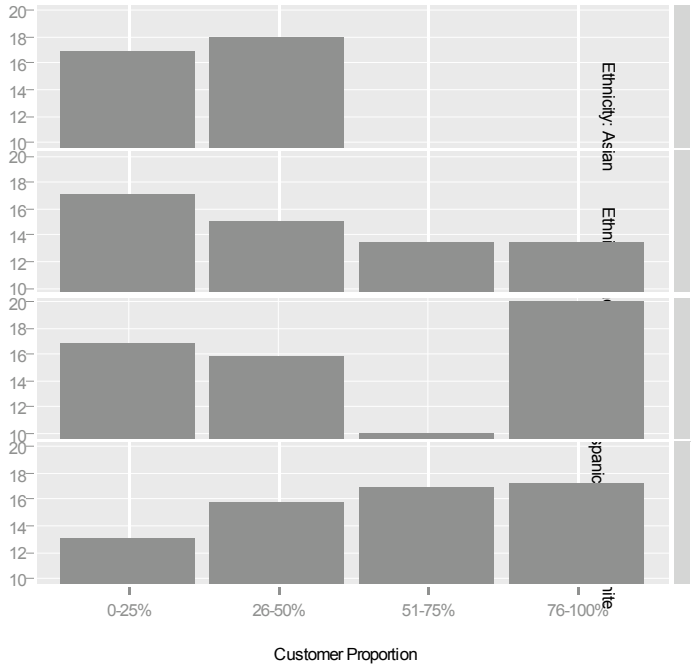


Figure 19: Relation between Customer Proportion and Average Tip grouped by Ethnical Origin of Customers for White Servers

Figure 19 shows that there is no data available for restaurants where proportion of Asian customers is more than 51%. We do not observe any specific trends for Asian and Hispanic customers regarding the tipping of White servers. But, there is an interesting pattern for Black and White customers. We can see that for higher proportion of White customers, White servers receive higher tips while the contrary is true for restaurants where the Black customers are in majority. This finding could be related to the fact that this dataset presents a majority of restaurants with a larger proportion of White customers as shown by figure 20 below.

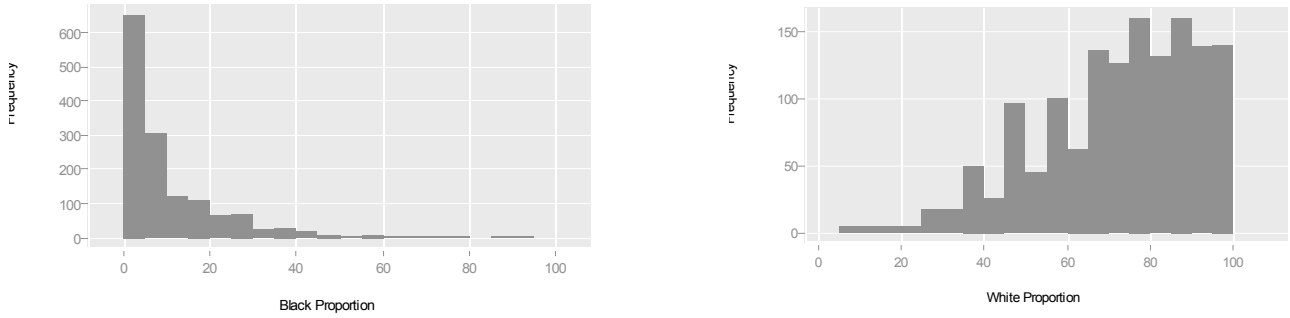


Figure 20: Distribution of proportion of Black and White customers

Relationship between the classiness of a restaurant and tip

Until now, we investigated the world and US tipping habits, the server's traits and their impact on the tip, and the effect of the customer's ethnicity on tips of White servers. In this last section, we will approach the restaurant variable group through the study of the classiness of the restaurant and its effect on server tips.

Among all available records, we selected restaurants that have 10 and more observations. Table 5 presents the average tip for the 11 selected restaurants.

Table 5: average tip per restaurant with more than 10 records

Restaurants	Mean % Tip
Applebee's	15.27
Bob evans	15.70
Carrabbas	14.82
Cheesecake factory	17.46
Chili's	15.97
Denny's	16.20
Olive garden	15.52
Outback steakhouse	15.34
Red lobster	15.32
Ruby Tuesday	14.31
Tgi Fridays	15.75

From our customer point of view, we created an order of classiness for these restaurants: (1) Cheesecake Factory, (2) Carrabba's, (3) Olive garden, (4) Red Lobster, (5) Chili's, (6) Applebee's, (7) Ruby Tuesday, (8) Outback steakhouse, (9) Tgi Fridays, (10) Denny's, and (11) Bob evans. Figure 21 shows the relation between our ranking of restaurants' classiness and their mean tip.

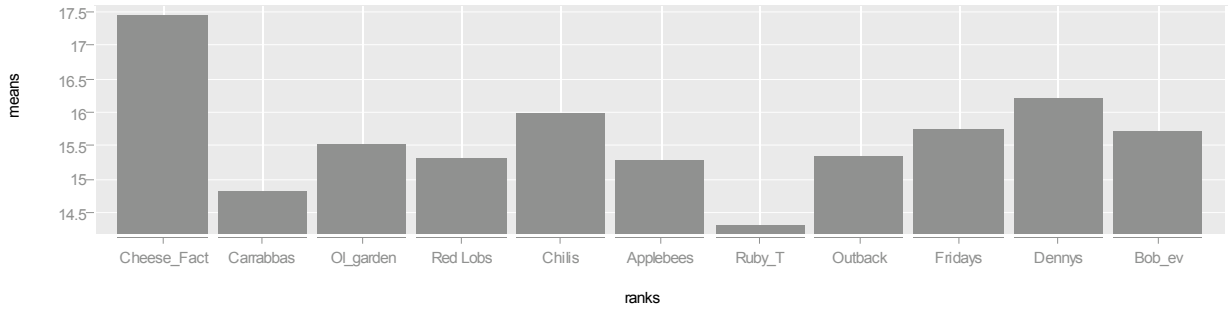


Figure 21: Average tip per restaurant

From the graph, we do not see any trend between our expectations of the effect of classiness on the actual average tip given by customers except for the Cheesecake Factory. The ten other places can be considered as family restaurants while Cheesecake Factory is more likely

to be classified as a casual dining place. People frequenting this place go for business meetings, dates and anniversaries while family restaurants are visited by regulars and their families and friends. Thus, the high tip average of the Cheesecake Factory could be explained by the fact that people will tip more on special occasions than on a regular basis.

To conclude on the relationship between the restaurant and tip amount, we decided to investigate the average tip at Applebee's in different US states.

Applebee's tip habits

We chose Applebee's because it has the most records (n=54). The map below shows the average tips at Applebee's locations in different states.

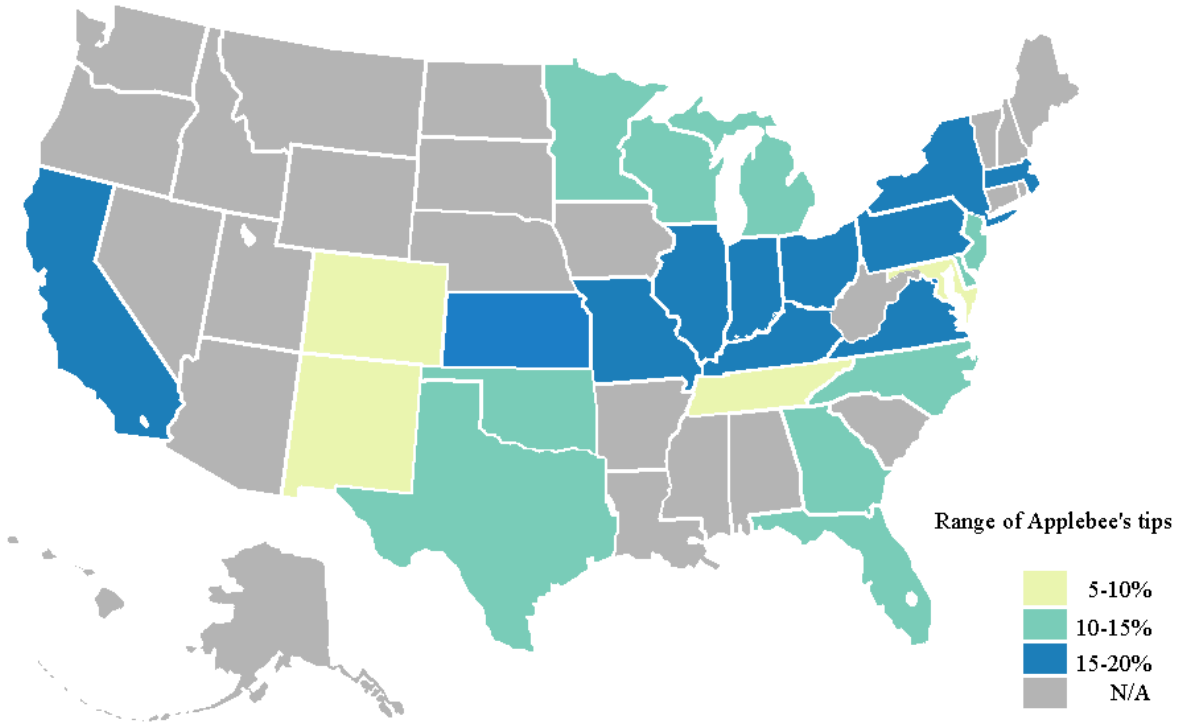


Figure 22: Average tips at Applebee's locations in different states

There are many states with no records for Applebee's. Therefore, we would not be able to generalize our conclusions for entire US. We can see that four states present a very low average tip (5-10%), that is unusually low compared to the average tipping rate for all restaurants (Colorado = 17.58%, Maryland = 17.87%, New Mexico = 13.30%, and Tennessee = 16.48%). We could say that the way Applebee's is appreciated by customers can vary from state to state. So, it would be interesting to do the similar investigation using another restaurant. However, this conclusion should be considered with care because for each state, there are between 1 and 6 records and no general trend can be confirmed from this sparse data.

Conclusions

The analysis of the server survey data allowed us to come up with very interesting conclusions regarding server behavior, customer behavior for US and the rest of the world. However, the dataset contains many erroneous and missing values. Also, most of the data were posted by White US servers and therefore, conclusions will mostly be applicable for White US servers.

To earn higher tip, a server should:

- Work in north-east states, Florida, Nevada, Hawaii, Colorado, and Kansas.
- Address customers by their name, share jokes and complement their choice of food.
- Work in high-class restaurants.

Furthermore, we were able to observe a significant difference between the tip received by male and female servers. Males receive significantly higher tips than females. But young White female are likely to get unusually high tips than others. However, the study of Applebee's tips showed that the customer behavior can vary from state to state for the same restaurant.

Finally, readers should pay attention to the fact that these conclusions are derived from an online survey where there was no way to verify the accuracy of the information. Further research on this topic should be conducted in a controlled environment to ensure the accuracy of the data. It would be interesting to design the survey to include the restaurants from a specific category (fast-food, casual-dining, etc) to customize the advice to servers depending on their working environment. Attention should be paid to collect data equally from all ethnical groups and genders to ensure the accuracy of final conclusions.

Appendixes

Appendix 1: Description of variables

source:	N/A
remoteip:	I.P. address of location from where data is submitted
datercvd:	Date of data collection
submit_time:	Time of data collection
when_employed:	N/A
rest:	Restaurant name
city:	City
state:	State
mos_current:	Months server has been working at specified restaurant
extra_months:	Extra months server would be working at same restaurant
more_mos:	N/A
asian_prop:	Proportion of Asian customers
black_prop:	Proportion of Black customers
hispanic_prop:	Proportion of Hispanic customers
white_prop:	Proportion of White customers
breakfast:	Servers who work at restaurant during breakfast shift
lunch:	Servers who work at restaurant during lunch shift
dinner:	Servers who work at restaurant during dinner shift
late_night:	Servers who work at restaurant at late night shift
busy:	How busy is the shift that the server works
ppbill:	Total bill per person
pcttip:	Tip as percentage of total bill
big_tips:	Big tips as percentage of total bill for server
comparative:	How often the server get tips compared to other servers at same restaurant
flair:	How often the server engages in flair with customers
intro:	How often the server introduces himself/herself to the customers
selling:	How often the server tries suggestive selling
squatt:	How often the server squats at customer table
touch:	How often the server touches the customers
jokes:	How often the server shares jokes with customers
repeat.:	How often the server repeats the order
customer_name:	How often the server addresses customer by their names
draw:	How often the server draws on the check
smile:	How often the server smiles
thanks:	How often the server writes thanks on the check
weather:	How often the server talks about weather with the customers
complement:	How often the server complements the customers
happy:	How happy the server is while waiting tables
yrs_experience:	Years of experience of the server
effect_sz:	Effect of server's quality of service on the size of tip
men:	What kind of tippers men are?
women:	What kind of tippers women are?
teenagers:	What kind of tippers teenagers are?
young_adults:	What kind of tippers young adults are?

middle_aged_customers:	What kind of tippers middle aged customers are?
elderly_customers:	What kind of tippers elderly customers are?
cash_customers:	What kind of tippers cash customers are?
charge_customers:	What kind of tippers charge customers are?
smokers:	What kind of tippers smokers are?
regulars:	What kind of tippers regular customers are?
first_timers:	What kind of tippers first time customers are?
asians:	What kind of tippers Asian customers are?
blacks:	What kind of tippers Black customers are?
hispanics:	What kind of tippers Hispanic customers are?
whites:	What kind of tippers White customers are?
foreigners:	What kind of tippers foreigners are?
couples:	What kind of tippers couples are?
onetops:	N/A
kids:	What kind of tippers kids are?
business_people:	What kind of tippers business customers are?
extraverted_enthusiastic:	To what extent this trait describes the server
critical_quarrelsome:	To what extent this trait describes the server
dependable_selfdisciplined:	To what extent this trait describes the server
anxious_easily_upset:	To what extent this trait describes the server
open_to_new_experiences_complex:	To what extent this trait describes the server
reserved_quiet:	To what extent this trait describes the server
sympathetic_warm:	To what extent this trait describes the server
disorganized_careless:	To what extent this trait describes the server
calm_emotionally_stable:	To what extent this trait describes the server
conventional_uncreative:	To what extent this trait describes the server
birth_yr:	Birth year of the server
sex:	Sex of the server
hair:	N/A
hair_other:	Hair color of the server
married:	Is the server currently married
race:	Ethnical group of the server
race_other:	N/A

Appendix 2: R Code

General description of the data

Presentation of the servers' gender

```
# Load data
completetipdata <- read.csv(file.choose())

# Select variables of interest in the dataset
data <- completetipdata[,c("sex","race","birth_yr","yrs_experience",
"pcttip","squatt","touch","smile","draw","thanks","rest","state")]

# Load the plot tools
library(ggplot)

# Figure 1: Server's gender distribution
qplot(sex,data=data,type="histogram", xlab="gender")

# Delete observations with sex = 2
data <- data[data$sex %in% 0:1, ]

# Create a new table with variables necessary to present gender data
datamelt<-melt(data,measure.var=c(5),id.var=c(1,2,3,4,6,7,8,9,10,11,12),
preserve.na=F)
table<-cast(datamelt,sex+race+birth_yr+yrs_experience+squatt+touch+smile+draw+
thanks+state+rest~variable, mean,na.rm=T)

# Associate 0 to male and 1 to female
table$sex=factor(table$sex,levels=c(0,1),labels=c("male","female"))

# Figure 2: Relation between tip and gender
qplot(sex,pcttip,data=table,type="boxplot",rm.na=T, xlab="gender", ylab="percentage
tip")

# t.test between males and females tips
table1<-cast(datamelt,sex~variable,c(mean,sd))
table1$sex<-factor(table1$sex,levels=c(0,1),labels=c("male","female"))

female <- datamelt[datamelt$sex %in% 1, ]
male <- datamelt[datamelt$sex %in% 0, ]
t.test(female$value,male$value,conf.level = 0.95)
```

Presentation of the servers' ethnical groups

```
# Figure 3: Server's ethnical group
qplot(race,data=table,type="histogram",scale="count",xlab="ethnical groups")

# Table 2: Proportion of each ethnical group in the dataset
ratio <- function(a){
       sum1<-table[table$race %in% a,]
       sum2<-table[table$race %in% 1:4,]
       (sum(sum1$race)/sum(sum2$race))*100
}

ratio(1)
ratio(2)
ratio(3)
ratio(4)

# Preparation of the data
```

```
table2<-cast(datamelt,race~variable,mean)
table2$race<-c("Asian","Black","Hispanic","White")

df<-data.frame(
       Race = c("Mean Asian","Mean Black","Mean Hispanic","Mean White"),
       Means = table2$pcttip,
       LegendRace = c("Asian","Black","Hispanic","White")
       )
```

Figure 4: Mean tip per ethnical group

```
qplot(Race, Means, data=df, colour=LegendRace, type="bar",xlab="ethnical
groups",ylab="mean percentage tip")
```

Presentation of the servers' age and experience

Delete erroneous dates

```
table <- table[table$birth_yr %in% 1900:1995, ]
```

Create the variable age

```
table$age<-(2006-table$birth_yr)
table$age_exp_relation<-table$age-table$yrs_experience
```

Figure 5: Relation between experience and age

```
qplot(age,yrs_experience,data=table,type="point")
```

Delete people who present less than 13 years between the age and experience

```
table<-table[table$age_exp_relation > 13 & !is.na(table$age_exp_relation), ]
```

Set ethnical group names

```
table$race <- factor(table$race, levels=c(1,2,3,4),
labels=c("Asian","Black","Hispanic","White"))
```

#Figure 6: Relationship between experience and tip in function of the gender and
ethnical group of the waiter

```
qplot(yrs_experience,pcttip,data=table,type="point",xlab="years of
experience",ylab="percentage of tip",facet = race~sex)
```

Figure 7: Relationship between age and tip in function of the gender and ethnical
group of the waiter

```
qplot(age,pcttip,data=table,type="point",xlab="age",ylab="percentage of tip",facet =
race~sex)
```

Set state names to lower case

```
table$state <- tolower(table$state)
```

Create a vector with the state names

```
states <- c("alabama"="al", "alaska"="ak", "arizona"="az", "arkansas"="ar",
"california"="ca", "colorado"="co", "connecticut"="ct", "delaware"="de", "district of
columbia"="dc", "florida"="fl", "georgia"="ga", "hawaii"="hi", "idaho"="id",
"illinois"="il", "indiana"="in", "iowa"="ia", "kansas"="ks", "kentucky"="ky",
"louisiana"="la", "maine"="me","maryland"="md", "massachusetts"="ma", "michigan"="mi",
"minnesota"="mn", "mississippi"="ms", "missouri"="mo", "montana"="mt",
"nebraska"="ne", "nevada"="nv", "new hampshire"="nh", "new jersey"="nj", "new
mexico"="nm", "new york"="ny", "north carolina"="nc", "north dakota"="nd",
"ohio"="oh", "oklahoma"="ok", "oregon"="or", "pennsylvania"="pa", "rhode island"="ri",
"south carolina"="sc", "south dakota"="sd", "tennessee"="tn", "texas"="tx",
```

```
"utah"="ut", "vermont"="vt", "virginia"="va", "washington"="wa", "west virginia"="wv",
"wisconsin"="wi", "wyoming"="wy")
```

Create a binary vector with 1 for observations corresponding to a US waiter
```
longstates <- table$state %in% names(states)
table$longstates<-longstates
table$longstates<- table$longstates*1
```

Create a vector with the state names abreviation
```
states2 <- c("al"="al","ak"="ak","az"="az","ar"="ar","ca"="ca","co"="co",
"ct"="ct","de"="de","dc"="dc","fl"="fl","ga"="ga","hi"="hi","id"="id","il"="il","in"="
in","ia"="ia","ks"="ks","ky"="ky","la"="la","me"="me","md"="md","ma"="ma","mi"="mi","m
n"="mn","ms"="ms","mo"="mo","mt"="mt","ne"="ne","nv"="nv","nh"="nh","nj"="nj","nm"="nm
","ny"="ny","nc"="nc","nd"="nd","oh"="oh","ok"="ok","or"="or","pw"="pw","pa"="pa","ri"
="ri","sc"="sc","sd"="sd","tn"="tn","tx"="tx","ut"="ut","vt"="vt","va"="va","wa"="wa",
"wv"="wv","wi"="wi","wy"="wy")
```

Create a binary vector similar to the previous but for observation presenting an abbreviated
US state name
```
shortstates <- table$state %in% names(states2)
table$shortstates<-shortstates
table$shortstates <- table$shortstates*1
table$USstates <- table$longstates-table$shortstates
table$USstates <- (table$USstates)^2
```

Sort observation between US and non-US data
```
tableUS<-table
tableUS<-tableUS[tableUS$USstates %in% 1,]
tableNONUS<-table
tableNONUS<-tableNONUS[tableNONUS$USstates %in% 0,]
```

Figure 8: Relationship between age and tip in function of the gender and ethnical
group of the waiter
```
qplot(age,pcttip,data=tableUS,type="point",xlab="age",ylab="percentage of tip",facet =
race~sex)
```

Global tipping habits

Calculate mean tip for US and Non US countries
```
meanUS<-mean(tableUS$pcttip)
meanNONUS<-mean(tableNONUS$pcttip)
```

Create a vector of Canadian states
```
canada<-c("ontario"="canada","ontario"="canada","new
brunswick"="canada","canada"="canada","ontario"="canada","ontario
(canada)"="canada","ontario"="canada","bc"="canada","nova
scotia"="canada","manitoba"="canada","bc (canada)"="canada","nova scotia
(canada)"="canada","british columbia"="canada","bc"="canada","ontario
(canada)"="canada","bc"="canada","on"="canada","manitoba canada"="canada","british
columbia (canada)"="canada","ontario canada"="canada","bc canada"="canada"
,"manitoba"="canada","alberta (canada)"="canada", "saskatchewan"="canada","manitoba"
="canada","b.c. canada"="canada","saskatchewan"="canada","bc canada"="canada",
"alberta"="canada","bc"="canada","manitoba"="canada","b.c. canada"="canada","manitoba"
="canada","british columbia"="canada","bc"="canada","ontario canada"="canada",
"bc"="canada","ontario"="canada","alberta"="canada","bc"="canada","british
columbai"="canada","manitoba"="canada","on"="canada","alberta"="canada","canada"="cana
da","alberta"="canada","alberta"="canada","ontario"="canada","ontario"="canada","ontar
```

```
io"="canada","ontario"="canada","canada"="canada","ontario"="canada","bc"="canada","on
tario"="canada","ontario"="canada","canada"="canada","ontario"="canada","bc
(canada)"="canada","alberta"="canada","british columbia"="canada","alberta"="canada"
,"ontario"="canada","british columbia (canada)"="canada","new brunswick"=
"canada","ontario"="canada","quebec (canada)"="canada","canada"="canada","quebec
(canada)"="canada","ontario"="canada","ontario"="canada","ontario"="canada","nova
scotia"="canada","british columbia"="canada","bc"="canada","quebec"=
"canada","quebec"="canada","ontario"="canada","quebec"="canada","ontario"="canada"
,"alberta"="canada","nova scotia"="canada","ontario"="canada","british
columbia"="canada","bc canada"="canada","quebec"="canada","ontario"="canada"
,"ontario"="canada","bc (canada)"="canada","on canada"="canada","alberta"
="canada","canada"="canada","alberta canada"="canada","alberta"="canada"
,"ontario"="canada")
```

Select Canadian states from the dataset

```
canadastate <- tableNONUS$state %in% names(canada)
canadastate<-canadastate*1
tableNONUS$canadastate<-canadastate

tableNONUS<-tableNONUS[tableNONUS$canadastate %in% 1,]
```

Calcuate mean tip for Canada and plot data

```
meancanada<-mean(tableNONUS$pcttip)
means<- c(meanUS, meanNONUS, meancanada)
Country = c("Mean USA","Mean World","Mean Canada")
Country<-factor(Country, level=c("Mean USA","Mean World","Mean Canada"))
df<-data.frame{
       Country,
       means
}
```

Figure 9: Average tip between tipping and non-tipping countries

```
qplot(Country, means, type="bar", data=df, ylim=c(0,17), ylab="Mean % tip")
```

Relation between localization, richness of a state and the tip # amount

Delete erroneous tip amounts

```
tableUS <-tableUS[tableUS$pcttip %in% 1:100, ]
```

Create a GSP vector

```
GSP<-c(31239,54713,33841,30077,42386,43764,52149,63004,139849,35051,38094,39804,
31186,41987,36850,37323,36188,32112,35544,32896,41483,48803,36282,44073,26582,35740,29
758,38902,42498,40090,47242,33444,47031,37933,35662,37133,31740,37483,37416,38747,3132
3,38539,36782,40193,34100,35401,43713,40774,27532,37746,47623)
```

Transform the dataset in a pivot table

```
dataUS1<-
melt(tableUS,measure.var=c(3),id.var=c(1,2,4,5,6,7,8,9,10,11,12,13,14,15,16,17,
18,19,20,21,22,23,24,25,26,27,28,29,30,31),preserve.na=F)
dataUS1answer<-cast(dataUS1,state~variable,mean)
```

Correct the states name

```
dataUS1answer$state<-
factor(dataUS1answer$state,levels=c("alabama","alaska","arizona","arkansas","californi
a","colorado","connecticut","delaware","district of columbia","florida",
"georgia","hawaii","idaho","illinois","indiana","iowa","kansas","kentucky","louisiana"
,"maine","maryland","massachusetts","michigan","minnesota","mississippi","missouri","m
```

```
ontana","nebraska","nevada","new hampshire","new jersey","new mexico","new
york","north carolina","north dakota","ohio","oklahoma"
,"oregon","pennsylvania","rhode island","south carolina","south dakota"
,"tennessee","texas","utah","vermont","virginia","washington","west
virginia","wisconsin","wyoming","al","ak","az","ar","ca","co","ct","de",
"dc","fl","ga","hi","id","il","in","ia","ks","ky","la","me","md","ma","mi","mn","ms","
mo","mt","ne","nv","nh","nj","nm","ny","nc","nd","oh","ok",
"or","pa","ri","sc","sd","tn","tx","ut","vt","va","wa","wv","wi","wy"),labels=c("al","
ak","az","ar","ca","co","ct","de","dc","fl","ga","hi","id","il","in","ia","ks","ky","l
a","me","md","ma","mi","mn","ms","mo","mt","ne","nv","nh","nj","nm","ny","nc","nd","oh
","ok", "or","pa","ri","sc","sd","tn","tx","ut","vt",
"va","wa","wv","wi","wy","al","ak","az","ar","ca","co","ct","de","dc","fl","ga","hi","
id","il","in","ia","ks","ky","la","me","md","ma","mi","mn","ms","mo","mt","ne","nv","n
h","nj","nm","ny","nc","nd","oh","ok", or","pa","ri","sc","sd"
,"tn","tx","ut","vt","va","wa","wv","wi","wy"))
```

Calculate the mean tip per state, and provide the ranking for the map

```
dataUS2<-melt(dataUS1answer,measured.var=c(2),id.var=c(1))
dataUS2answer<-cast(dataUS2,state~variable,mean)
dataUS2answer$GSP<-GSP
dataUS2answer<-dataUS2answer[order(dataUS2answer$GSP, decreasing=T),]
```

Table 3: States ranked by GSP

```
dataUS2answer
```

Figure 11: Relation between GSP and the average tip per state

```
a<-rep(c(1),1)
b<-rep(c(2),10)
c<-rep(c(3),27)
d<-rep(c(4),8)
e<-rep(c(5),5)

Gross_State_Product<-c(a,b,c,d,e)

dataUS2answer$Gross_State_Product<-Gross_State_Product
dataUS2answer$Gross_State_Product<-factor(dataUS2answer$Gross_State_Product,
levels=c(1,2,3,4,5),labels=c(">19.18%","17.51-19.18%","15.84-17.51%","14.17-
15.84%","12.40-14.17%"))

qplot(pcttip,GSP,data=dataUS2answer,type="point",xlab="Average percentage
tip",colour=Gross_State_Product)
```

Calculate correlation between GSP and Tip

```
correl<-function(x){
sub<-dataUS2answer[dataUS2answer$Gross_State_Product %in% x,]
cor(sub$GSP,sub$pcttip)
}

correl(">19.18%")
correl("17.51-19.18%")
correl("15.84-17.51%")
correl("14.17-15.84%")
correl("12.40-14.17%")
```

Influence of server traits on the amount of tip

Label gender of servers as male and female

```
data$sex<-factor(data$sex, levels=c(0,1), labels=c("male","female"))
dataUS$sex<-factor(dataUS$sex, levels=c(0,1), labels=c("male","female"))
```

Plot the server traits to make sure that the dataset is clean and there are no missing
values

Figure 12: Frequency distribution of six server traits
Frequency of introduction
```
qplot(data$intro, type="histogram", breaks=seq(1, 4, by=0.1), scale="count", xlab
="Intro", main="Introduce themselves(1-Never, 4-Always)")
```

Frequency of suggestive selling
```
qplot(data$selling, type="histogram", breaks=seq(1, 4, by=0.1), scale="count", xlab
="Selling", main="Suggestive Selling(1-Never, 4-Always)")
```

Frequency of sharing jokes with the customers
```
qplot(data$jokes, type="histogram", breaks=seq(1, 4, by=0.1), scale="count", xlab
="Jokes", main="Jokes(1-Never, 4-Always)")
```

Frequency of addressing customers by their names
```
qplot(data$customer_name, type="histogram", breaks=seq(1, 4, by=0.1), scale="count",
xlab ="Customer Name", main="Customer Name(1-Never, 4-Always)")
```

Frequency of complementing the customers
```
 qplot(data$complement, type="histogram", breaks=seq(1, 4, by=0.1), scale="count",
xlab ="Complement", main="Complement(1-Never, 4-Always)")
```

Frequency of thanking the customers
```
qplot(data$thanks, type="histogram", breaks=seq(1, 4, by=0.1), scale="count", xlab
="Thanks", main="Thanks(1-Never, 4-Always)")
```

Create a dataset for comparing the six different traits displayed by the servers and their affect
on tip
```
traitUS<-data[,c("pcttip","intro","selling","jokes","customer_name","thanks",
"complement","sex", "race","asian_prop","black_prop","hispanic_prop","white_prop")]
```

Melt this dataset to include tip as measured variable and all others as id variables
```
traitUSm<-melt(traitUS,measure.var=1,preserve.na=FALSE)
```

Figure 13: Relationship between tip and frequency of server traits

#Creating a line plot for comparing the mean tip for servers displaying the six traits. Each trait
is cast with variable to calculate the mean tip for each server trait.

```
options(digits=3)

t1<-cast(traitUSm,intro~variable,mean)

t2<-cast(traitUSm,selling~variable,mean)

t3<-cast(traitUSm,jokes~variable,mean)

t4<-cast(traitUSm,customer_name~variable,mean)

t5<-cast(traitUSm,complement~variable,mean)

t6<-cast(traitUSm,thanks~variable,mean)
```

Label the trait levels 1,2,3,4 as never, sometimes, often and always for each trait

```
t1$intro <- factor(t1$intro, levels=c(1,2,3,4), labels=c("Never","Sometimes","Often",
"Always"))

t2$selling <- factor(t2$selling, levels=c(1,2,3,4), labels=c("Never","Sometimes"
,"Often","Always"))

t3$jokes<- factor(t3$jokes, levels=c(1,2,3,4), labels=c("Never","Sometimes"
,"Often","Always"))

t4$customer_name <- factor(t4$customer_name , levels=c(1,2,3,4),
labels=c("Never","Sometimes"
,"Often","Always"))

t5$complement <- factor(t5$complement, levels=c(1,2,3,4), labels=c("Never","Sometimes"
,"Often","Always"))

t6$thanks <- factor(t6$thanks, levels=c(1,2,3,4), labels=c("Never","Sometimes"
,"Often","Always"))
```

Assign variable names for percentage tip from each category of server trait

```
intro<-t1$pcttip
selling<-t2$pcttip
jokes<-t3$pcttip
customer_name<-t4$pcttip
complement<-t5$pcttip
thanks<-t6$pcttip
```

Create a vector to assign the labels (selected for trait frequencies above) to each of the six traits
and factor them

```
Frequency <- rep(c("Never","Sometimes","Often","Always"),6)
Frequency <- factor(Frequency,level=c("Never","Sometimes","Often","Always"))
```

Create a vector to assign the trait names to each of the four frequencies and factor them

```
Trait <- rep(c("intro","selling","jokes","customer_name","complement","thanks"),4)
Trait <-
factor(Trait,level=c("intro","selling","jokes","customer_name","complement","thanks"))
```

Create vectors to repeat each trait four times, each for four different frequencies and combine
these vectors into a matrix to include all traits

```
a<-rep("intro",4)
b<-rep("selling",4)
c<-rep("jokes",4)
d<-rep("customer_name",4)
e<-rep("complement",4)
f<-rep("thanks",4)

Trait<-c(a,b,c,d,e,f)
```

Create a data frame from the frequency, trait matrices and percentage tip per trait variables
created above

```
 df<-data.frame(
      Frequency,
      Trait,
      means = c(intro,selling,jokes,customer_name,complement,thanks)
)
```

Plot a line chart to investigate the relation between server traits and percentage tips

```
qplot(Frequency,means,data=df,colour=Trait, id=Trait,type ="line",
```

```
main = "Relationship between Tip and Server Traits", ylab="% Tip", xlab="")
```

Combine each trait with "introduction" and calculate the mean categorized by gender of the
server. Plot line charts for each combination of traits and both genders.

Figure 14: Relationship between tip and frequency of Introduction and Suggestive Selling
for male and female servers

```
intro_sellingUS<-cast(traitUSm,intro+selling~variable|sex,mean)

introm<-intro_sellingUS$male$intro
sellingm<-intro_sellingUS$male$selling
is_tipm<-intro_sellingUS$male$pcttip
df <- data.frame(introm=factor(introm),sellingm=factor(sellingm),is_tipm)
qplot(sellingm, is_tipm, data=df, type="line", colour=introm, id=introm, main="Male",
ylim=c(12,20),xlab="Selling", ylab="% Tip")


introf<-intro_sellingUS$female$intro
sellingf<-intro_sellingUS$female$selling
is_tipf<-intro_sellingUS$female$pcttip
df <- data.frame(introf=factor(introf),sellingf=factor(sellingf),is_tipf)
qplot(sellingf,    is_tipf,    data=df,    type="line",    colour=introf,    id=introf,
main="Female",ylim=c(12,20),xlab="Selling", ylab="% Tip")
```

Figure 15: Relationship between tip and frequency of Introduction and Sharing Jokes for
male and female servers

```
intro_jokesUS<-cast(traitUSm,intro+jokes~variable|sex,mean)
intro_jokesUS

introm<-intro_jokesUS$male$intro
jokesm<-intro_jokesUS$male$jokes
ij_tipm<-intro_jokesUS$male$pcttip
df <- data.frame(introm=factor(introm),jokesm=factor(jokesm),ij_tipm)
qplot(jokesm,    ij_tipm,    data=df,    type="line",    colour=introm,    id=introm,
main="Male",ylim=c(12,20),xlab="Jokes", ylab="% Tip")
introf<-intro_jokesUS$female$intro
jokesf<-intro_jokesUS$female$jokes
ij_tipf<-intro_jokesUS$female$pcttip
df <- data.frame(introf=factor(introf),jokesf=factor(jokesf),ij_tipf)
qplot(jokesf,    ij_tipf,    data=df,    type="line",    colour=introf,    id=introf,
main="Female",ylim=c(12,20),xlab="Jokes", ylab="% Tip")
```

Figure 16: Relationship between tip and frequency of Introduction and Addressing
#Customers by their name for male and female servers

```
intro_custnameUS<-cast(traitUSm,intro+customer_name~variable|sex,mean)
intro_custnameUS

introm<-intro_custnameUS$male$intro
namem<-intro_custnameUS$male$customer_name
in_tipm<-intro_custnameUS$male$pcttip
df <- data.frame(introm=factor(introm),namem=factor(namem),in_tipm)
qplot(namem,    in_tipm,    data=df,    type="line",    colour=introm,    id=introm,
main="Male",ylim=c(12,20),xlab="Customer Name", ylab="% Tip")


introf<-intro_custnameUS$female$intro
namef<-intro_custnameUS$female$customer_name
```

```
in_tipf<-intro_custnameUS$female$pcttip
df <- data.frame(introf=factor(introf),namef=factor(jokesf),in_tipf)
qplot(namef,    in_tipf,    data=df,    type="line",    colour=introf,    id=introf,
main="Female",ylim=c(12,20),xlab="Customer Name", ylab="% Tip")
```

Figure 17: Relationship between tip and frequency of Introduction and Complementing the
Customers for male and female servers

```
intro_complementUS<-cast(traitUSm,intro+complement~variable|sex,mean)
intro_complementUS

introm<-intro_complementUS$male$intro
compm<-intro_complementUS$male$complement
ic_tipm<-intro_complementUS$male$pcttip
df <- data.frame(introm=factor(introm),compm=factor(compm),ic_tipm)
qplot(compm,    ic_tipm,    data=df,    type="line",    colour=introm,    id=introm,
main="Male",ylim=c(12,20),xlab="Complement", ylab="% Tip")


introf<-intro_complementUS$female$intro
compf<-intro_complementUS$female$complement
ic_tipf<-intro_complementUS$female$pcttip
df <- data.frame(introf=factor(introf),compf=factor(compf),ic_tipf)
qplot(compf,    ic_tipf,    data=df,    type="line",    colour=introf,    id=introf,
main="Female",ylim=c(12,20),xlab="Complement", ylab="% Tip")
```

Figure 18: Relationship between tip and frequency of Introduction and Thanking the
Customers for male and female servers

```
intro_thanksUS<-cast(traitUSm,intro+thanks~variable|sex,mean)
intro_thanksUS

introm<-intro_thanksUS$male$intro
thanksm<-intro_thanksUS$male$thanks
it_tipm<-intro_thanksUS$male$pcttip
df <- data.frame(introm=factor(introm),thanksm=factor(thanksm),it_tipm)
qplot(thanksm,    it_tipm,    data=df,    type="line",    colour=introm,    id=introm,
main="Male",ylim=c(12,20),xlab="Thanks", ylab="% Tip")


introf<-intro_thanksUS$female$intro
thanksf<-intro_thanksUS$female$thanks
it_tipf<-intro_thanksUS$female$pcttip
df <- data.frame(introf=factor(introf),thanksf=factor(thanksf),it_tipf)
qplot(thanksf,    ij_tipf,    data=df,    type="line",    colour=introf,    id=introf,
main="Female",ylim=c(12,20),xlab="Thanks", ylab="% Tip")
```

Investigate the difference in the mean tip for each trait with frequency = “Always (4)”. First,
select the tips when the frequency of each trait is equal to 4.

```
Intro_always<-traitUS[traitUS$intro %in% 4, ]
Selling_always<-traitUS[traitUS$selling %in% 4, ]
Jokes_always<-traitUS[traitUS$jokes %in% 4, ]
Name_always<-traitUS[traitUS$customer_name %in% 4, ]
Complement_always<-traitUS[traitUS$complement %in% 4, ]
Thanks_always<-traitUS[traitUS$thanks %in% 4, ]
```

Table 4: P-Values for tips of servers with different traits (Frequency = Always)

Conduct t-test for all combinations of traits and extract the p-value to determine if the mean
tip of the servers who always display each trait is significantly different or not

```
t.test(Intro_always$pcttip,Selling_always$pcttip)$p.value
t.test(Intro_always$pcttip,Jokes_always$pcttip)$p.value
t.test(Intro_always$pcttip,Name_always$pcttip)$p.value
t.test(Intro_always$pcttip,Complement_always$pcttip)$p.value
t.test(Intro_always$pcttip,Thanks_always$pcttip)$p.value
t.test(Selling_always$pcttip,Jokes_always$pcttip)$p.value
t.test(Selling_always$pcttip,Name_always$pcttip)$p.value
t.test(Selling_always$pcttip,Complement_always$pcttip)$p.value
t.test(Selling_always$pcttip,Thanks_always$pcttip)$p.value
t.test(Jokes_always$pcttip,Name_always$pcttip)$p.value
t.test(Jokes_always$pcttip,Complement_always$pcttip)$p.value
t.test(Jokes_always$pcttip,Thanks_always$pcttip)$p.value
t.test(Name_always$pcttip,Complement_always$pcttip)$p.value
t.test(Name_always$pcttip,Thanks_always$pcttip)$p.value
t.test(Complement_always$pcttip,Thanks_always$pcttip)$p.value
```

Influence of the ethnicity of customers on Server's Tip

Load data
```
completetipdata <- read.csv(file.choose())
```

Select variables of interest
```
data <- completetipdata[,c("city","state","pcttip","rest","flair","intro"
,"selling","squatt","touch","jokes","repeat.","customer_name","draw","smile","thanks",
"weather","complement","happy","yrs_experience","birth_yr","sex","race","asian_prop","
black_prop","hispanic_prop","white_prop")]
```

Delete erroneous data
Sex
```
data <- data[data$sex %in% 0:1, ]
```
Age
```
data <- data[data$birth_yr %in% 1900:1995, ]
```

Proportion of customers
```
props <- !is.na(data$asian_prop) & data$asian_prop< 1
data$asian_prop[props] <- data$asian_prop[props] * 100
data[!is.na(data$asian_prop) &  data$asian_prop > 100, ]
data[!is.na(data$asian_prop) &  data$asian_prop > 100, "asian_prop"] <- NA


props <- !is.na(data$black_prop) & data$black_prop< 1
data$black_prop[props] <- data$black_prop[props] * 100
data[!is.na(data$black_prop) &  data$black_prop > 100, ]
data[!is.na(data$black_prop) &  data$black_prop > 100, "black_prop"] <- NA

props <- !is.na(data$hispanic_prop) & data$hispanic_prop< 1
data$hispanic_prop[props] <- data$hispanic_prop[props] * 100
data[!is.na(data$hispanic_prop) &  data$hispanic_prop > 100, ]
data[!is.na(data$hispanic_prop) &  data$hispanic_prop > 100, "hispanic_prop"] <- NA

props <- !is.na(data$white_prop) & data$white_prop< 1
data$white_prop[props] <- data$white_prop[props] * 100
data[!is.na(data$white_prop) &  data$white_prop > 100, ]
data[!is.na(data$white_prop) &  data$white_prop > 100, "white_prop"] <- NA
```

Label the ethnical groups of servers as Asian, Black , Hispanic and White for the US data

```
dataUS$race <- factor(dataUS$race, levels = c(1, 2, 3, 4), labels = c("Asian", "Black", "Hispanic", "White"))
```

Create a dataset for investigating the affect of the relation between ethnical groups of
customers and the ethnical groups and gender of the servers

```
cust_server <- dataUS[, c("asian_prop", "black_prop", "hispanic_prop", "white_prop", "pcttip", "race", "sex")]
```

Create a variable to add all the customer proportions

```
cust_server$customer_prop <- cust_server$asian_prop + cust_server$black_prop + cust_server$hispanic_prop + cust_server$white_prop
```

Omit the values where the total proportion exceeds 100%

```
cust_server<-cust_server[cust_server$customer_prop %in% 100,]
```

Choose data only for white servers

```
cust_server<-cust_server[cust_server$race %in% "White",]
```

Figure 19: Relation between Customer Proportion and Average Tip grouped by Ethnical
Origin of Customers for White Servers

Divide the customer proportions into groups of 4, 0-25%, 26-50%, 51-75% and 76-100%

```
asian<-cust_server[cust_server$asian_prop %in% 0:25,]
asian2<-cust_server[cust_server$asian_prop %in% 26:50,]
asian3<-cust_server[cust_server$asian_prop %in% 51:75,]
asian4<-cust_server[cust_server$asian_prop %in% 76:100,]

black<-cust_server[cust_server$black_prop %in% 0:25,]
black2<-cust_server[cust_server$black_prop %in% 26:50,]
black3<-cust_server[cust_server$black_prop %in% 51:75,]
black4<-cust_server[cust_server$black_prop %in% 76:100,]

hispanic<-cust_server[cust_server$hispanic_prop %in% 0:25,]
hispanic2<-cust_server[cust_server$hispanic_prop %in% 26:50,]
hispanic3<-cust_server[cust_server$hispanic_prop %in% 51:75,]
hispanic4<-cust_server[cust_server$hispanic_prop %in% 76:100,]

white<-cust_server[cust_server$white_prop %in% 0:25,]
white2<-cust_server[cust_server$white_prop %in% 26:50,]
white3<-cust_server[cust_server$white_prop %in% 51:75,]
white4<-cust_server[cust_server$white_prop %in% 76:100,]
```

Create function to calculate the mean tip for each group of customer proportion

```
means<-function(x){
       a<-summary(x$pcttip)
       a[4]
}

a<-means(asian)
b<-means(asian2)
c<-means(asian3)
d<-means(asian4)
```

```
e<-means(black)
f<-means(black2)
g<-means(black3)
h<-means(black4)

i<-means(hispanic)
j<-means(hispanic2)
k<-means(hispanic3)
l<-means(hispanic4)

m<-means(white)
n<-means(white2)
o<-means(white3)
p<-means(white4)
```

Create plots for relation between percentage tip of White servers for each ethnical group of # customer proportions. Also, create histograms presenting the distribution of Black and White # customers in the dataset

```
means_prop=c(a,b,c,d,e,f,g,h,i,j,k,l,m,n,o,p)
Ethnicity<-c(1,1,1,1,2,2,2,2,3,3,3,3,4,4,4,4)
Ethnicity<-
factor(Ethnicity,levels=c(1,2,3,4),labels=c("Asian","Black","Hispanic","White"))
prop<-c(1,2,3,4,1,2,3,4,1,2,3,4,1,2,3,4)
prop<-factor(prop,levels=c(1,2,3,4),labels=c("0-25%","26-50%","51-75%","76-100%"))

qplot(prop,means_prop,type="bar",facet=Ethnicity~., xlab="Customer Proportion", ylab=
"Average Tip %")

qplot(black_prop, type="histogram", data=cust_server, breaks=30, scale="count", xlab=
"Black Proportion")

qplot(white_prop, type="histogram", data=cust_server, breaks=30, scale="count", xlab=
"White Proportion")
```

Age

```
data$age<-(2006-data$birth_yr)
data$age_exp_relation<-data$age-data$yrs_experience
data<-data[data$age_exp_relation > 13 & !is.na(data$age_exp_relation), ]
```

Sort between US and non US observations

```
data$state <- tolower(data$state)
data$rest<-tolower(data$rest)

states <- c("alabama"="al", "alaska"="ak", "arizona"="az", "arkansas"="ar",
"california"="ca", "colorado"="co", "connecticut"="ct", "delaware"="de", "district of
columbia"="dc", "florida"="fl", "georgia"="ga", "hawaii"="hi", "idaho"="id",
"illinois"="il", "indiana"="in", "iowa"="ia", "kansas"="ks", "kentucky"="ky",
"louisiana"="la", "maine"="me","maryland"="md", "massachusetts"="ma", "michigan"="mi",
"minnesota"="mn", "mississippi"="ms", "missouri"="mo", "montana"="mt", "nebraska"
="ne", "nevada"="nv", "new hampshire"="nh", "new jersey"="nj", "new mexico"="nm", "new
york"="ny", "north carolina"="nc", "north dakota"="nd", "ohio"="oh", "oklahoma"="ok",
"oregon"="or", "pennsylvania"="pa", "rhode island"="ri", "south carolina"="sc", "south
dakota"="sd", "tennessee"="tn", "texas"="tx", "utah"="ut", "vermont"="vt", "virginia"
="va", "washington"="wa", "west virginia"="wv", "wisconsin"="wi", "wyoming"="wy")

longstates <- data$state %in% names(states)
data$longstates<-longstates
data$longstates<- data$longstates*1
```

```
states2 <- c("al"="al","ak"="ak","az"="az","ar"="ar","ca"="ca","co"="co","ct"="ct",
"de"="de","dc"="dc","fl"="fl","ga"="ga","hi"="hi","id"="id","il"="il","in"="in","ia"="
ia","ks"="ks","ky"="ky","la"="la","me"="me","md"="md","ma"="ma","mi"="mi","mn"="mn","m
s"="ms","mo"="mo","mt"="mt","ne"="ne","nv"="nv","nh"="nh","nj"="nj","nm"="nm","ny"="ny
","nc"="nc","nd"="nd","oh"="oh","ok"="ok","or"="or","pw"="pw","pa"="pa","ri"="ri","sc"
="sc","sd"="sd","tn"="tn","tx"="tx","ut"="ut","vt"="vt","va"="va","wa"="wa","wv"="wv",
"wi"="wi","wy"="wy")

shortstates <- data$state %in% names(states2)
data$shortstates<-shortstates
data$shortstates <- data$shortstates*1
data$USstates <- data$longstates-data$shortstates
data$USstates <- (data$USstates)^2


dataUS<-data
dataUS<-dataUS[dataUS$USstates %in% 1,]
dataNONUS<-data
dataNONUS<-dataNONUS[dataNONUS$USstates %in% 0,]

library(ggplot)
dataUS1<-
melt(dataUS,measure.var=c(3),id.var=c(1,2,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20
,21,22,23,24,25,26,27,28,29,30,31),preserve.na=F)
```

Add GSP variable

```
GSP<-
c(31239,54713,33841,30077,42386,43764,52149,63004,139849,35051,38094,39804,31186,41987
,36850,37323,36188,32112,35544,32896,41483,48803,36282,44073,26582,35740,29758,38902,4
2498,40090,47242,33444,47031,37933,35662,37133,31740,37483,37416,38747,31323,38539,367
82,40193,34100,35401,43713,40774,27532,37746,47623)

dataUS1answer<-cast(dataUS1,state~variable,mean)
```

Correct state names

```
dataUS1answer$state<-factor(dataUS1answer$state,levels
=c("alabama","alaska","arizona","arkansas","california","colorado",
"connecticut","delaware","district of columbia","florida","georgia","hawaii",
"idaho","illinois","indiana","iowa","kansas","kentucky","louisiana","maine","maryland"
,"massachusetts","michigan","minnesota","mississippi","missouri","montana","nebraska",
"nevada","new hampshire","new jersey","new mexico","new york" ,"north carolina","north
dakota","ohio","oklahoma","oregon", "pennsylvania","rhode island","south
carolina","south dakota","tennessee"
,"texas","utah","vermont","virginia","washington","west virginia",
"wisconsin","wyoming","al","ak","az","ar","ca","co","ct","de","dc","fl","ga","hi","id"
,"il","in","ia","ks","ky","la","me","md","ma","mi","mn","ms","mo","mt","ne","nv","nh",
"nj","nm","ny","nc","nd","oh","ok", "or","pa","ri","sc","sd","tn", "tx","ut",
"vt","va","wa","wv","wi","wy"),labels=c("al","ak","az","ar","ca","co","ct","de","dc","
fl","ga","hi","id","il","in","ia","ks","ky","la","me","md","ma","mi","mn","ms","mo","m
t","ne","nv","nh","nj","nm","ny","nc","nd","oh","ok","or","pa","ri","sc","sd","tn","tx
","ut","vt","va","wa","wv","wi", "wy","al","ak","az","ar", "ca","co","ct",
"de","dc","fl","ga","hi","id","il","in","ia","ks","ky","la","me","md","ma","mi","mn","
ms","mo","mt","ne","nv","nh","nj","nm","ny","nc","nd","oh","ok", "or","pa","ri","sc",
"sd","tn","tx","ut","vt","va","wa","wv","wi","wy"))

dataUS2<-melt(dataUS1answer,measured.var=c(2),id.var=c(1))
dataUS2answer<-cast(dataUS2,state~variable,mean)

dataUS2answer$GSP<-GSP
```

```
dataUS2answer[order(dataUS2answer$GSP, decreasing=T),]
dataUS2answer$state <- factor(dataUS2answer$state, levels=dataUS2answer$state
[order(dataUS2answer$GSP, decreasing=T)])
```

Create a table with restaurants

```
dataUS3<-cast(dataUS1,rest~variable,c(sum,length))

dataUS3<-dataUS3[dataUS3$pcttip_length > 2 & !is.na(dataUS3$pcttip_length), ]
```

Correct restaurant names

```
dataUS3$rest<-factor(dataUS3$rest,levels=c("applebee's","applebees","banfi's","banfi's
restaurant","bennigans","bertuccis","bob evans","bonefish grill", "buca di beppo",
"buffalo wild wings","california pizza kitchen","carrabba's",
"carrabba's italian grill","champps americana","cheesecake factory",
"chili's","chilis","cracker barrel","denny's","dennys","friendly's","ground
round","hooters","houlihans","ihop","joe's crab shack","joes crab shack","johnny
carino's","logan's roadhouse","lone star","lonestar steakhouse","macaroni
grill","mimi's cafe","o'charley's","olive garden",
"on the border","outback","outback steakhouse","pappasito's cantina","perkins","pizza
hut","rainforest cafe","red lobster","red robin","romano's macaroni grill","ruby
tuesday",
"ruth's chris steak house","smokey bones","tgi fridays","the cheesecake
factory","waffle house"), labels=c("applebees","applebees","banfis",
"banfis","bennigans","bertuccis","bob evans","bonefish grill","buca di beppo","buffalo
wild wings","california pizza kitchen","carrabbas",
"carrabbas","champps americana","cheesecake factory","chilis","chilis",
"cracker barrel","dennys","dennys","friendlys","ground round",
"hooters","houlihans","ihop","joes crab shack","joes crab shack","johnny
carinos","logans roadhouse","lone star","lone star","macaroni grill","mimis cafe","o
charleys","olive garden","on the border","outback", "outback","pappasitos
cantina","perkins","pizza hut","rainforest cafe","red lobster","red robin","romanos
macaroni grill","ruby tuesday",
"ruths chris steak house","smokey bones","tgi fridays","cheesecake factory","waffle
house"))
```

Sort data by tip amount

```
dataUS3[order(dataUS3$pcttip_length,decreasing=T),]
```

Calculate a mean tip per restaurant

```
dataUS3answer<-melt(as.data.frame(dataUS3),id.var=c(1),preserve.na=F)
dataUS4<-cast(dataUS3answer, fun=sum)
dataUS4$mean<-(dataUS4$pcttip_sum/dataUS4$pcttip_length)
```

Sort data by tip amount

```
dataUS4[order(dataUS4$pcttip_length,decreasing=T),]
```

Create a variable with restaurant presenting more than 10 records

```
restaurants<-dataUS4[dataUS4$pcttip_length %in% 10:41,]
restaurants<-restaurants[,c("rest","mean")]
```

Create the classiness ranking

```
ranks<-c(1,2,3,4,5,6,7,8,9,10,11)
ranks<-factor(ranks,levels=c(1,2,3,4,5,6,7,8,9,10,11),
labels=c("Cheese_Fact","Carrabbas",
"Ol_garden","Red Lobs","Chilis","Applebees","Ruby_T","Outback",
"Fridays","Dennys","Bob_ev"))
means<-c(17.46,14.82,15.52,15.32,15.97,15.27,14.31,15.34,15.75,16.20,15.70)
```

Figure 20: Relation between Customer Proportion and Average Tip grouped by

Ethnical Origin of Customers for White Servers

```
qplot(ranks,means,type="bar")
```

Correct applebee's spelling

```
Apple<-c("applebee"="applebee","applebees"="applebees","applebee's"="applebee's",
"appleby's grill and bar"="appleby's grill and bar","applebee's bar and
grill"="applebee's bar and grill","applebee's neighborhood grill and bar"="applebee's
neighborhood grill and bar","applebees nieghborhood bar and grill"="applebees
nieghborhood bar and grill")
```

Select Applebee's data

```
rest1<-data[data$rest %in% names(Apple),]

rest1answer<-rest1[,c("state","pcttip")]
rest1answer<-rest1answer[order(rest1answer$state,decreasing=F),]
```

Set the right state names

```
rest1answer$state<-factor(rest1answer$state,levels=c("california","colorado",
"delaware","delaware","florida","florida","florida","ga","georgia","georgia","il","in"
,"indiana","kansas","kentucky","ma","ma","marlyand","massachusetts","mi","minnesota","
minnesota","minnesota","missouri","missouri","missouri","missouri","mo","new
mexico","new york","nj","north carolina","ny","ny","oh","ohio", "ohio",
"ohio","ohio","ohio","ohio","oklahoma","oklahoma","pa","pa",
"texas","tn","tn","tx","va","virginia","wi","wi","wi"),labels=c("ca","co","de","de","f
l","fl","fl","ga","ga","ga","il","in","in","ks","ky","ma","ma","md","ma","mi","mn","mn
","mn","mo","mo","mo","mo","mo","nm","ny","nj","nc","ny","ny","oh","oh","oh","oh","oh"
,"oh","oh","ok","ok","pa","pa","tx","tn","tn","tx","va","va","wi","wi","wi"))
```

Sort data to create the map

```
rest1answer<-melt(rest1answer, measured.var=c(2), id.var=c(1))
rest_tip<-cast(rest1answer, state~variable, mean)
```