C 5
5 5
0 5

Great report.

Server Survey Data Report
Stat 480

Mar 29, 2007

Introduction

The dataset is a survey of tipping habits of people and how they could be related to the behavior of servers. It contains an extensive range of data for both the server and the consumer, including attributes such as their race and age group. Tipping is a very interesting phenomenon as in some countries like US and Canada it is socially mandatory to tip while in most European and Asian countries, tipping is more a choice.

The mean tips received by male and female servers were also calculated. I also explored the affect that years of experience of a server might have on the tip and the relation between total bill and tip. For the exploration of the server-survey data, I decided to concentrate on various server attributes and race of the customers, as one race might have a habit of tipping more compared to the others. After an initial exploration of server gesture attributes, I decided to report 5 different gestures that might have impact on the tip. These attributes were joking, introducing themselves, selling (suggesting menu items), thanking and talking about weather with the customers. All these attributes can be considered as engaging in a conversation with the customers that would give a personal touch. These customers might feel more comfortable and have a better overall experience of the restaurant than other customers, and more likely tip higher.

For the customer race, I explored the tipping habits of people from different races. I wanted to determine whether servers at a restaurant with higher proportion of customers from a certain race were tipped more. The different customer races given in the dataset were Asian, Black, Hispanic and White.

Data Exploration

Load data in R, named object ss

> ss <- read.csv(file.choose())

Use str to display the structure of dataset ss

> str(ss)

```
'data.frame':   2618 obs. of  77 variables:
 $ source                : int  2 2 2 2 2 2 2 2 2 2 ...
 $ remoteip              : Factor w/ 2479 levels "105901586","105904156",..: 1919 ...
 $ datercvd              : num  2.01e+13 2.01e+13 2.01e+13 2.01e+13 2.01e+13 ...
 $ submit_time           : Factor w/ 10 levels " ","8/10/2006",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ when_employed         : int  1 1 1 1 1 3 3 NA 1 1 ...
```

```
$ rest                          : Factor w/ 1904 levels "'T Voorhuys",..: 1554 1552 1630 5...
$ city                          : Factor w/ 1321 levels "----"," ","2020 K St. Washington",: ...
$ state                         : Factor w/ 346 levels " ","(Not US)",..: 245 244 193 298 239...
$ mos_current                   : num  187 156 240 180 1 42 84 96 60 10 ...
$ xtra_months                   : num  180 156 24 12 12 6 0 240 12 12 ...
$ more_mos                      : Factor w/ 47 levels " ","????","1 year",..: 1 1 1 1 1 1 1 1 1 1
$ asian_prop                    : num  1 1 3 1 NA 15 4 5 5 2 ...
$ black_prop                    : num  1 5 6 20 NA 10 5 5 10 10 ...
$ hispanic_prop                 : num  1 45 6 30 NA 5 1 5 5 1 ...
$ white_prop                    : num  97 65 85 59 100 70 90 85 80 87 ...
$ breakfast                     : int  0 0 0 0 0 0 0 0 0 0 ...
$ lunch                         : int  1 1 1 1 1 1 0 1 1 1 ...
$ dinner                        : int  1 1 0 1 1 1 1 1 1 1 ...
$ late_night                    : int  0 0 0 0 0 0 0 0 0 0 ...
$ busy                          : int  2 2 2 3 2 2 2 2 2 2 ...
$ ppbill                        : num  16.0 20.0 14.4 100.0 45.0 ...
$ pcttip                        : num  18 12 18 20 15 22 20 21 12 20 ...
$ big_tips                      : num  50 5 20 50 10 70 95 80 8 80 ...
$ comparative_tips              : int  2 3 5 3 3 3 4 3 3 3 ...
$ flair                         : int  4 1 4 2 2 1 1 2 1 1 ...
$ intro                         : int  4 1 4 4 4 1 4 2 4 2 ...
$ selling                       : int  4 4 3 4 4 4 4 3 4 3 4 ...
$ squatt                        : int  1 1 4 2 1 1 2 1 2 1 ...
$ touch                         : int  4 1 2 2 2 1 4 1 2 1 ...
$ jokes                         : int  4 1 2 3 3 3 3 3 2 1 ...
$ repeat.                       : int  4 1 4 3 3 2 2 2 3 2 ...
$ customer_name                 : int  4 3 2 3 2 2 3 2 3 1 ...
$ draw                          : int  2 1 2 3 1 1 2 1 1 1 ...
$ smile                         : int  4 3 3 3 3 3 3 3 3 2 ...
$ thanks                        : int  2 1 4 4 2 1 2 1 1 1 ...
$ weather                       : int  3 1 1 1 3 1 1 1 1 2 ...
$ complement                    : int  4 3 3 3 3 2 1 4 2 3 ...
$ happy                         : int  7 7 6 7 5 2 5 7 6 3 ...
$ yrs_experience                : num  25 17 26 10 10 15 12 30 20 19 ...
$ effect_sz                     : int  5 5 3 3 3 5 3 5 3 4 ...
$ men                           : int  2 3 3 3 2 3 3 3 2 3 ...
$ women                         : int  2 1 2 1 1 2 1 2 2 2 ...
$ teenagers                     : int  2 1 1 1 1 2 1 1 1 1 ...
$ young_adults                  : int  2 1 1 2 2 2 2 1 1 2 ...
$ middle_aged_adults            : int  2 2 2 2 3 2 3 3 2 3 ...
$ elderly_adults                : int  2 2 1 2 1 2 1 1 1 2 ...
$ cash_customers                : int  2 2 3 3 2 3 2 2 2 2 ...
$ charge_customers              : int  2 3 2 3 2 NA 3 2 3 3 ...
$ smokers                       : int  2 3 3 3 2 1 3 3 2 2 ...
$ regulars                      : int  3 3 3 3 2 3 3 3 2 3 ...
$ first_timers                  : int  2 2 3 2 2 2 2 2 3 2 ...
$ asians                        : int  2 2 2 2 1 2 2 2 2 1 ...
$ blacks                        : int  2 1 1 1 2 2 2 1 1 1 ...
$ hispanics                     : int  2 1 3 1 2 2 2 1 1 2 ...
$ whites                        : int  2 2 2 2 2 3 3 3 2 2 ...
$ foreigners                    : int  2 1 1 2 1 1 1 1 1 1 ...
$ couples                       : int  2 3 2 3 2 3 2 2 3 2 ...
$ onetops                       : int  2 3 1 2 3 3 2 3 3 3 ...
$ kids                          : int  2 2 1 2 1 2 2 1 2 1 ...
$ business_people               : int  3 99 3 3 3 3 3 3 2 3 ...
$ extraverted_enthusiastic      : int  7 7 6 7 6 6 7 6 4 5 ...
$ critical_quarrelsome          : int  1 6 1 1 4 2 3 1 5 2 ...
$ dependable_selfdisciplined    : int  7 7 7 7 7 5 6 6 6 7 ...
$ anxious_easily_upset          : int  7 2 5 2 1 2 5 1 2 1 ...
$ open_to_new_experiences_complex : int  7 6 NA 4 6 6 7 6 3 6 ...
$ reserved_quiet                : int  2 1 NA 2 2 5 1 5 4 3 ...
$ sympathetic_warm              : int  7 2 6 6 6 3 7 5 5 5 ...
$ disorganized_careless         : int  6 1 2 2 2 5 1 1 1 1 ...
$ calm_emotionally_stable       : int  7 7 7 4 7 7 6 7 5 6 ...
$ conventional_uncreative       : int  2 1 1 1 3 1 2 6 3 3 ...
$ birth_yr                      : int  1947 1969 1953 1973 1970 1961 1975 1954 1960 ...
$ sex                           : int  0 1 1 1 0 0 1 0 0 1 ...
$ hair                          : int  NA 2 NA 1 NA 2 2 2 2 1 ...
$ hair_other                    : Factor w/ 117 levels " ","\\Black\\\" or REALLY DARK
$ married                       : int  0 0 0 1 0 0 0 0 0 0 ...
```

```
$ race                  : int  NA 4 2 4 4 4 4 4 4 ...
$ race_other            : Factor w/ 93 levels " ","1/2 Asian 1/2 Caucasian",..: 76...
```

The dataset contains 77 variables.

Load the Reshape package

```
> library(reshape)
Attaching package: 'reshape'
    The following object(s) are masked _by_ .GlobalEnv :
     tips
```

The percent tip (pcttip) variable was specified as measured variable using melt function. All other variables are then considered as id variables.

```
> ssm<-melt(ss, measure.var=22, preserve.na=FALSE)
```

Examine the counts for factor levels of sex.

```
> table(ssm$sex, exclude=NULL)

   0    1    2 <NA>
 714 1606  142  156
```

0 corresponds to male and 1 corresponds to female servers. It seems that value 2 was entered in error as there can only be 2 genders, either male or female.

Thus, level 2 and NA can be removed to make the analysis easy.

```
> ssm <- ssm[ssm$sex %in% 0:1,]
```

Check the sex levels again.

```
> table(ssm$sex, exclude=NULL)

   0    1
 714 1606
```

Levels given for sex of servers can be encoded as factors.

```
> ssm$sex <- factor(ssm$sex, levels=c(0,1), labels=c("male", "female"))
> table(ssm$sex, exclude=NULL)

  male female
   714   1606
```

Calculate mean, minimum and maximum percent tip for males and female servers

```
> cast(ssm, sex~variable, min, na.rm=TRUE)
    sex pcttip
1   male      1
2 female      0

> cast(ssm, sex~variable, max, na.rm=TRUE)
    sex pcttip
1   male     30
2 female    100

> cast(ssm, sex~variable, mean, na.rm=TRUE)
    sex   pcttip
1   male 16.50
2 female 16.09
```

It shows that on average male servers get higher tips than females but it is not a considerable difference. The maximum tip earned is higher for female waiter.

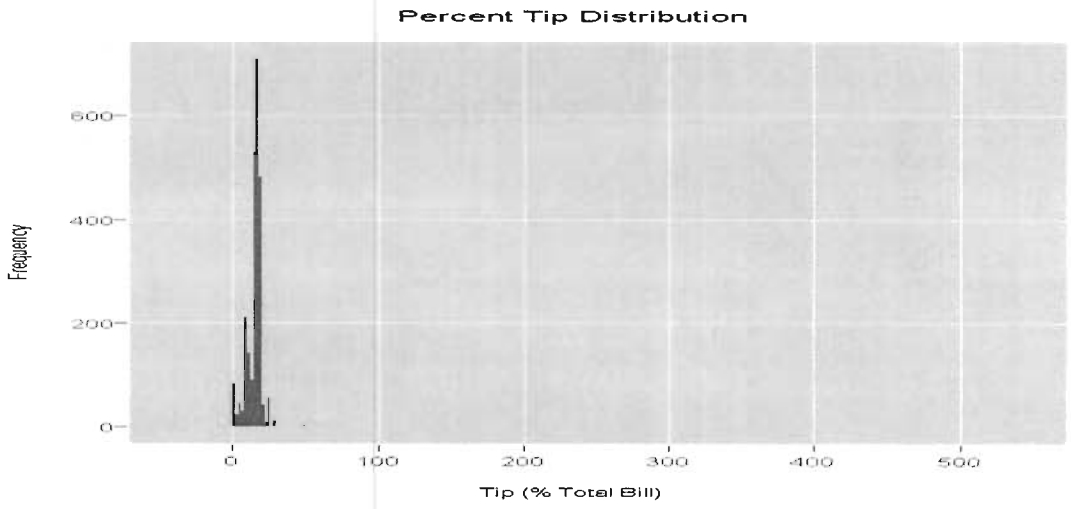
There are some records where people tipped a lot. To investigate tips over 50% of total bill, use the following syntax:

```
ss <- ss[!is.na(ss$pcttip) &  ss$pcttip > 50, ]

90 100 500
 1   2   1
```

This shows that there are 4 records where the tip was greater than 50 % of total bill

Plot a histogram for distribution of tips

```
> qplot(pcttip, data=ss, type= "histogram", scale = "count", breaks= seq(0,500, by = 2),
xlab = "Tip (% Total Bill)", main = "Percent Tip Distribution")
```



Clearly, there are outliers present, as we see a tip of 500% in the histogram. This is a self- reported dataset and therefore, many errors can occur. Some servers reported very high tips of more than 100%. This is very unlikely. In order for accurate data analysis, all tips greater than 50 % were removed.

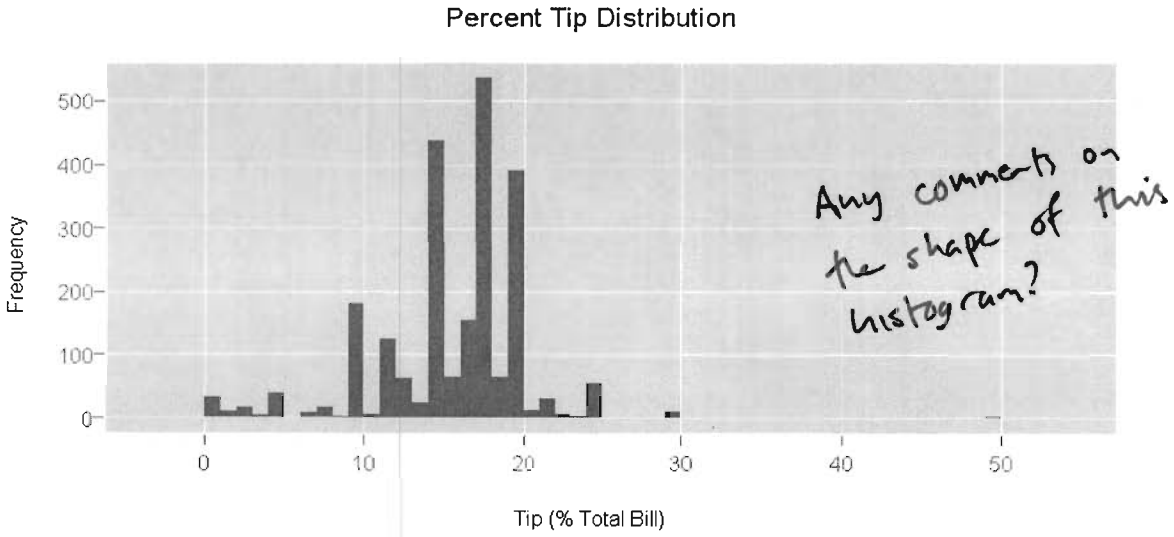
Good.

> table(ss$pcttip, exclude=NULL)

```
   0 1e-05 0.005  0.08   0.1  0.12  0.13  0.15  0.18   0.2  0.21  0.25   0.5     1
  25     1     1     1     3     1     1     5     3     4     1     1     1    16
 1.5     2   2.5     3     4     5     6     7   7.5     8     9    10    11    12
   2    17     2    18     5    45     1    10     1    20     2   209     7   136
12.5    13  13.5    14  14.5    15 15.59    16  16.5 16.78    17  17.5    18  18.5
   2    63     1    25     1   464     1    65     3     1   159     5   541     1
18.9    19  19.5    20    21    22    23    24    25    28    29    30    35    40
   1    65     1   414    12    31     7     2    57     1     1    10     1     1
  50    65    70    90   100   500  <NA>
   3     1     1     1     2     1   134
```

> tip <- !is.na(ss$pcttip) & ss$pcttip > 50
*> ss$pcttip[tip] <- ss$pcttip[tip] * 1*
> ss[!is.na(ss$pcttip) & ss$pcttip > 50,]
> ss[!is.na(ss$pcttip) & ss$pcttip > 50, "pcttip"] <- NA

> qplot(pcttip, data=ss, type= "histogram", scale = "count", breaks= seq(0,50, by = 1), xlab = "Tip (% Total Bill)", main = "Percent Tip Distribution")



Gesture of the servers

I decided to study the influence of gesture of the servers on the tip. I considered the gestures that imply that the servers engaged in a conversation with the customers:

1. Servers sharing jokes (jokes)
2. Servers introducing themselves (intro)
3. Servers trying to sell food items (selling)

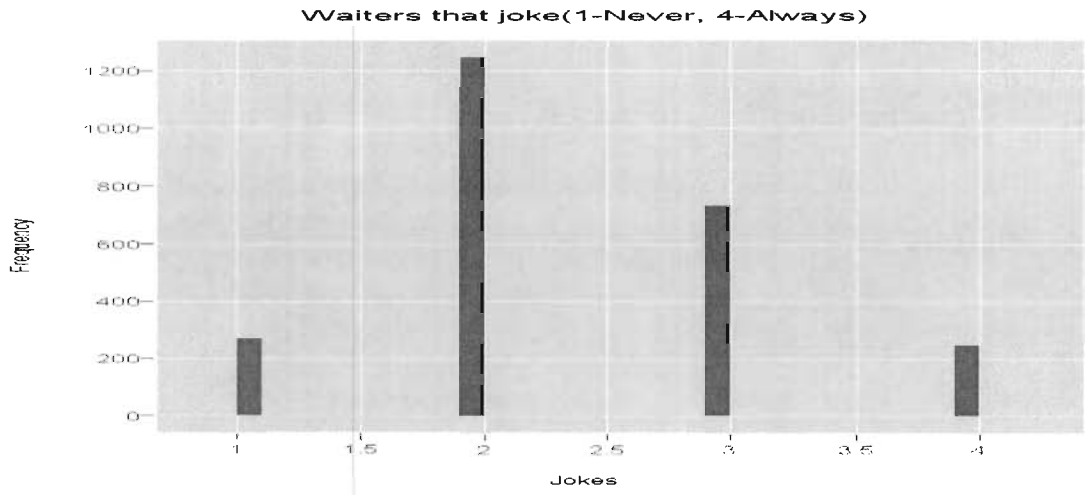
4. Servers thanking the customers (thanks)
5. Servers talking about weather (weather)

All these attributes are ranked from 1 to 4. 1 represents that the server never engages in the above actions and 4 represents that the server always engages in these actions with the customers.

1. Servers who share jokes with the customers

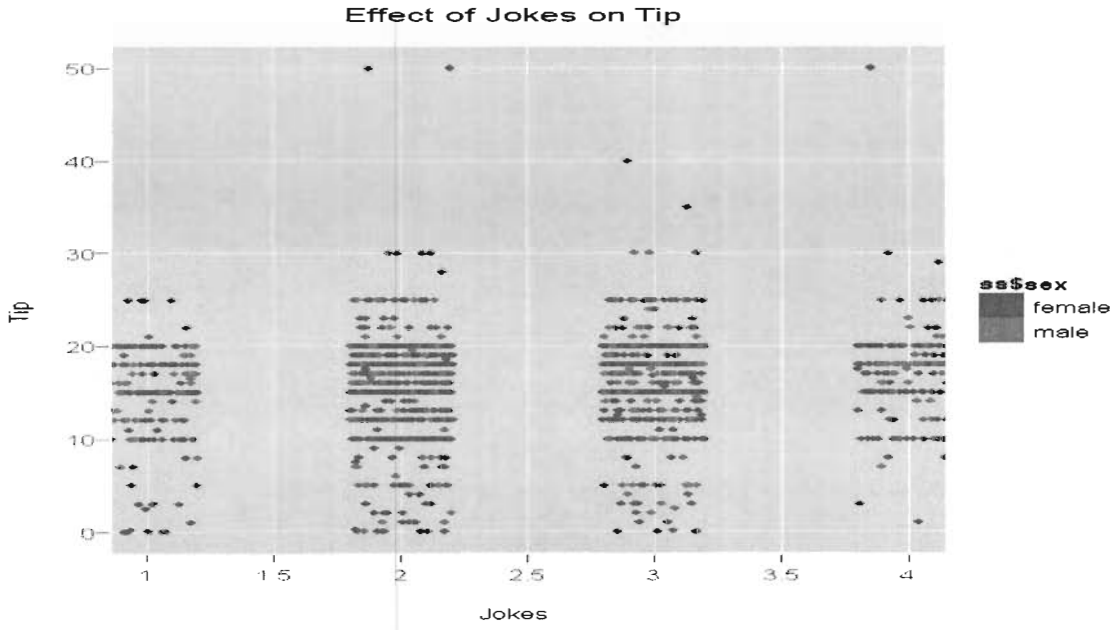
I removed the all the data other than the numbers from 1 to 4 as follows. The chart shows the histogram for the number of servers who like to share jokes with the customers. The frequency of number 2 is the highest which implies that only sometimes, most of the servers like to share jokes with their customers.

```
> gjoke <- !is.na(ss$jokes) & ss$jokes > 4
> ss$jokes[gjoke] <- ss$jokes[gjoke] * 1
> ss[!is.na(ss$jokes) & ss$jokes > 4, ]
> ss[!is.na(ss$jokes) & ss$jokes > 4, "jokes"] <- NA

> qplot(ss$jokes, type="histogram", breaks=seq(1, 4, by=0.1), scale="count", xlab
="Jokes", main="Waiters that joke(1-Never, 4-Always)")
```



Affect of sharing jokes by male and female servers on the tip

```
> qplot(ss$jokes, ss$pcttip, type="jitter", colour = ss$sex, xlab="Jokes", ylab="Tip",
main = "Influence of Jokes on Tip")
```



The above chart shows that there is no affect of sharing jokes by the servers on the tip given by the customers. Add a smooth to double check.

2. Servers who introduce themselves and try to sell to the customer

> ssm<-melt(ss, measure.var=22, preserve.na=FALSE)
> cast(ssm, intro~selling | variable, mean)

$pcttip

	intro	X1	X2	X3	X4
1	1	16.4	15.8	16.4	17.6
2	2	13.4	15.8	15.7	16.5
3	3	18.1	14.9	15.3	17.2
4	4	15.7	16.7	16.0	16.4

Good experimentation with cast.

> cast(ssm, .~variable | selling, mean)

$`1`

	value	pcttip
1	value	15.8

$`2`

	value	pcttip
1	value	15.9

$`3`

	value	pcttip
1	value	15.9

$`4`
  value pcttip
1 value   16.6

> *cast(ssm, .~variable | intro, mean)*
$`1`
  value pcttip
1 value   16.5

$`2`
  value pcttip
1 value   15.7

$`3`
  value pcttip
1 value   15.9

$`4`
  value pcttip
1 value   16.3

Servers that introduced themselves to the customer did not get considerably more tip than the ones who did not introduce themselves nor did that occasionally, while the servers who always tried to sell to the customers or suggested items from the menu received more tip than others. The combination of these two attributes is interesting. I found that the servers who often introduced themselves and never suggested menu items received the highest tip. The ones who received lowest tip sometimes introduced themselves but never suggested menu items.

Looking at the numbers would be useful here too.

3. Servers who thanked the customers and talked about the weather

> *cast(ssm, thanks+weather~variable , mean, margins = "weather", subset = thanks %in% c(1,2,3,4))*

	thanks	weather	pcttip
1	1	1	16.4
2	1	2	16.3
3	1	3	16.8
4	1	4	18.8
5	2	1	16.3
6	2	2	16.3
7	2	3	16.1
8	2	4	14.0
9	3	1	15.5
10	3	2	15.8
11	3	3	14.9

I really like how you have investigated multiple variables at a time.

12	4	1	15.8
13	4	2	16.0
14	4	3	16.4
15	4	4	17.7

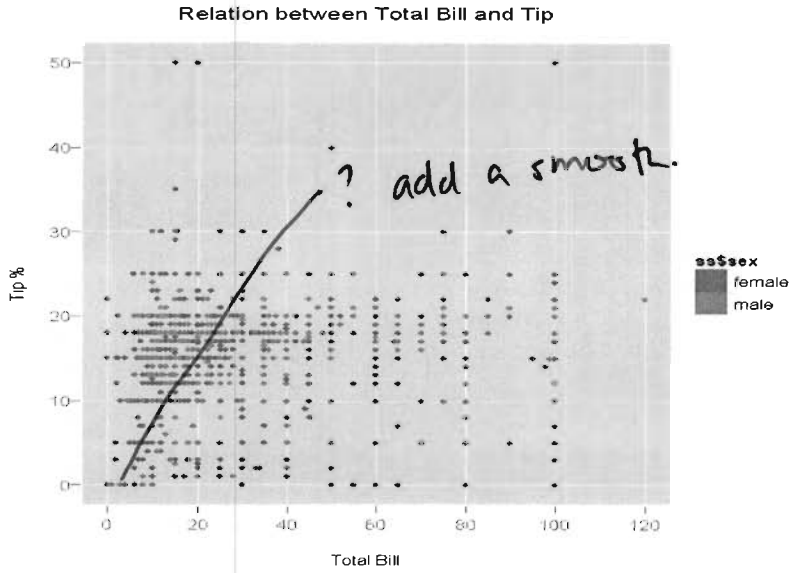
The servers who always talked about weather and never thanked the customers received more tips while the servers that sometimes thanked the customer and never talked about the weather received the less tip than others. This finding is very unusual. I would expect that a server who thanks the customer and engages in a conversation is likely to get more tips. But, there are many factors that affect the tipping criteria of the customer, so we cannot draw any conclusion based on a few actions of the server.

Relationship between total bill and tip

One might expect that the people who spend more in a restaurant are likely to tip more than the ones who spend less.

In order to examine the trend between total bill per person and the percent tip, I neglected the total bill values above $150.

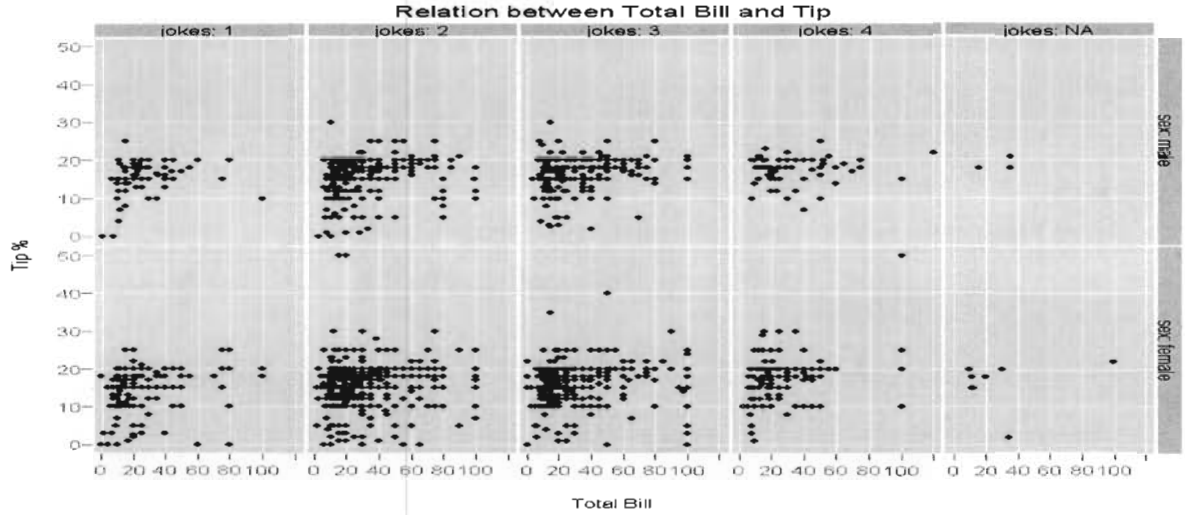
```
> bill <- !is.na(ss$ppbill) & ss$ppbill > 150
> ss$ppbill[bill] <- ss$ppbill[bill] * 1
> ss[!is.na(ss$ppbill) &  ss$ppbill > 150, ]
> ss[!is.na(ss$ppbill) &  ss$ppbill > 150, "ppbill"] <- NA

> qplot(ss$ppbill, ss$pcttip, type="jitter", colour=ss$sex, xlab="Total Bill", ylab="Tip %", main= "Relation between Total Bill and Tip")
```



The plot shows no pattern for the relation between total bill and the tip. In some cases, female servers might receive more tip than the male servers for the same amount of bill but the number of these records is very less.

Relation between total bill, tip categorized by the sex of the server and joking

> qplot(ppbill, pcttip, data = ss, type="jitter", facet = sex~jokes, xlab="Total Bill", ylab="Tip %", main= "Relation between Total Bill and Tip")



There are a few records present that indicate that the female servers who always joke with customers receive more tip than the male servers for the total bill amount from $ 10 to $ 40. For a total bill of $ 100, the female servers who often joke receive fewer tips than the male servers. Again, there is no solid pattern for this relationship.

Affect of years of experience of server

> cast(ssm, yrs_experience~sex | variable, mean)

$pcttip

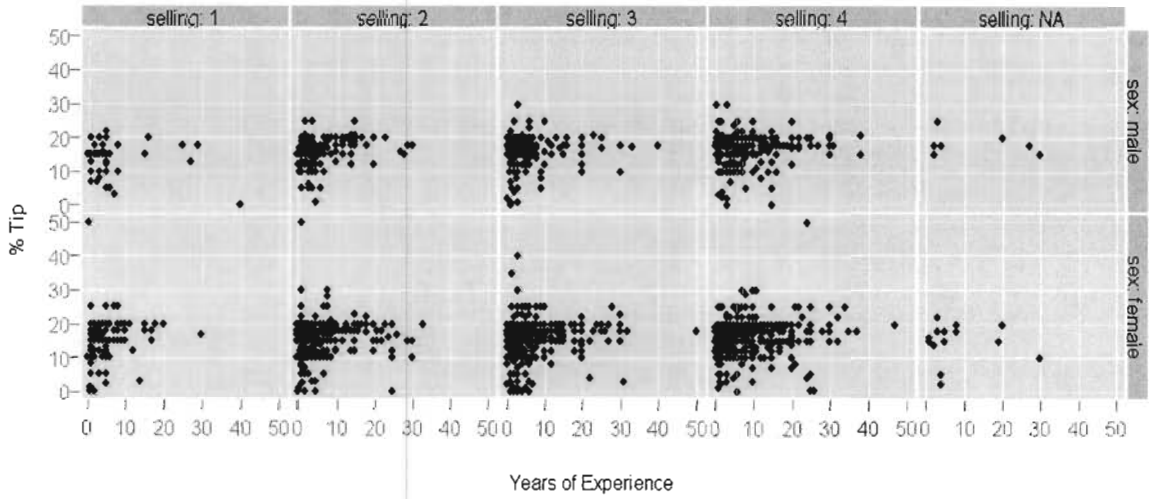
	yrs_experience	male	female
14	1.00	16.19	14.6
15	1.20	NA	6.5
16	1.25	19.00	NA
17	1.30	17.00	NA
18	1.50	9.33	15.0
19	1.60	NA	15.0
20	2.00	16.21	16.0
21	2.50	14.00	15.9
22	2.80	NA	1.0
23	3.00	16.21	15.9
24	3.50	16.00	18.0
25	4.00	16.33	15.7
26	4.33	NA	15.0
27	5.00	16.05	15.9
28	5.50	NA	17.0
29	6.00	16.55	16.2
30	7.00	15.70	16.2
31	8.00	16.58	17.8
32	8.50	NA	15.0
33	9.00	17.33	18.2
34	10.00	17.78	16.4

35	11.00	18.50	18.5
36	12.00	17.50	16.9
37	13.00	18.67	17.7
38	14.00	17.00	17.3
39	15.00	16.25	17.0
40	16.00	18.00	18.0
41	17.00	18.67	17.1
42	18.00	18.00	18.3
43	19.00	18.00	20.0
44	20.00	17.50	16.6
45	21.00	NA	11.5
46	22.00	18.50	9.5
47	23.00	21.00	20.0
48	24.00	NA	25.8
49	25.00	NA	14.0
50	26.00	NA	17.0
51	27.00	13.00	NA
52	28.00	15.00	25.0
53	29.00	18.00	15.0
54	30.00	16.67	18.8
55	31.00	18.00	NA
56	34.00	NA	25.0
57	35.00	NA	18.0
58	38.00	18.00	NA

The records with more than 1 year of experience are shown above. There is no obvious pattern for the relation between tip received and the years of experience of male and female servers.

Affect of years of experience of server combined with selling and sex

> qplot(yrs_experience, pcttip, data=ss, facet=sex~selling, xlab="Years of Experience", ylab="% Tip")



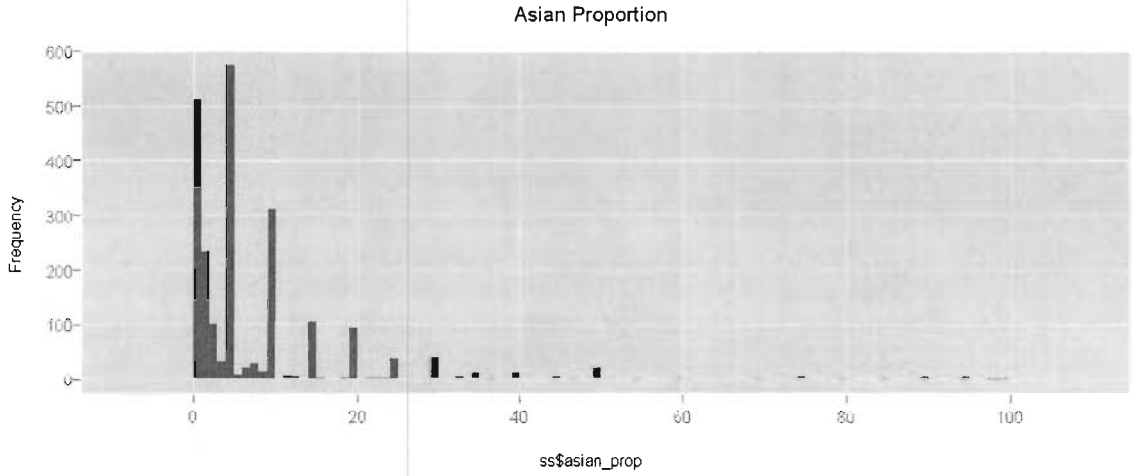
This chart shows that female servers who always suggested menu items (selling) received more tips than male servers regardless of number of years of experience.

Affect of proportion of different races in a restaurant on tips

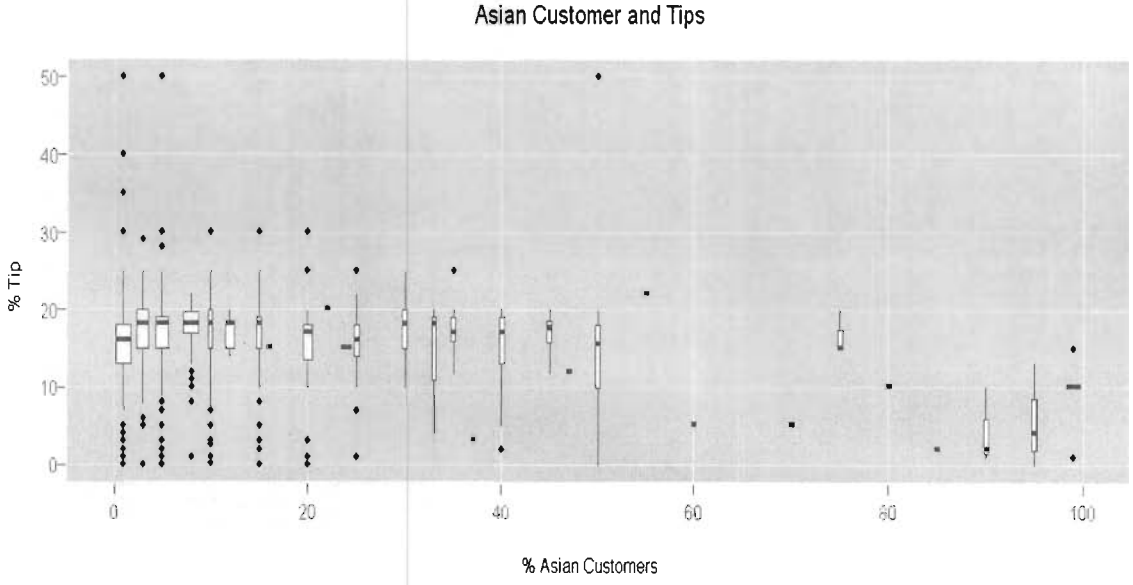
The proportions given for different races were converted into percentage values.

Asian

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

> qplot(ss$asian_prop, type="histogram", scale="count",breaks=seq(0, 100, by=1))
```

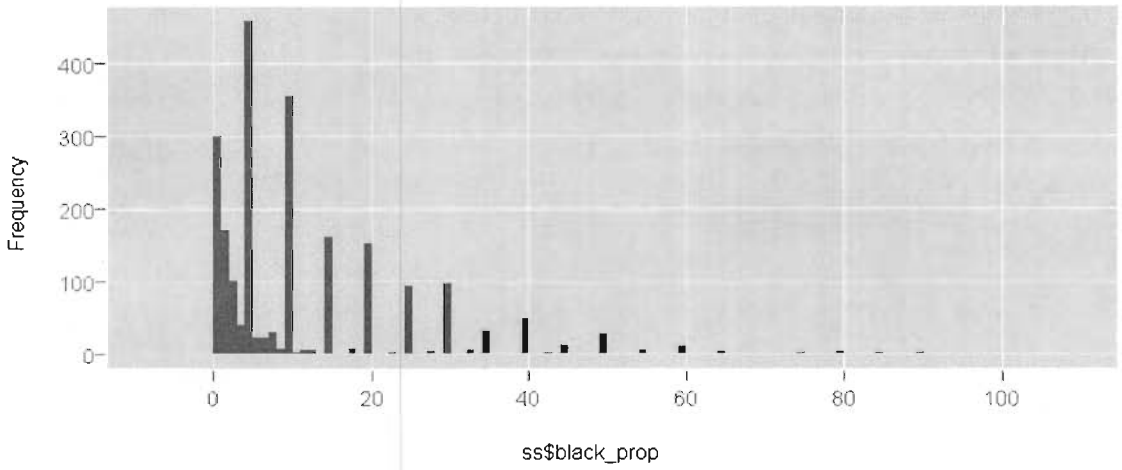



```
> qplot(ss$asian_prop, ss$pcttip, type="boxplot", main = "Asian Customer and Tips",
xlab="% Asian Customers", ylab="% Tip")
```

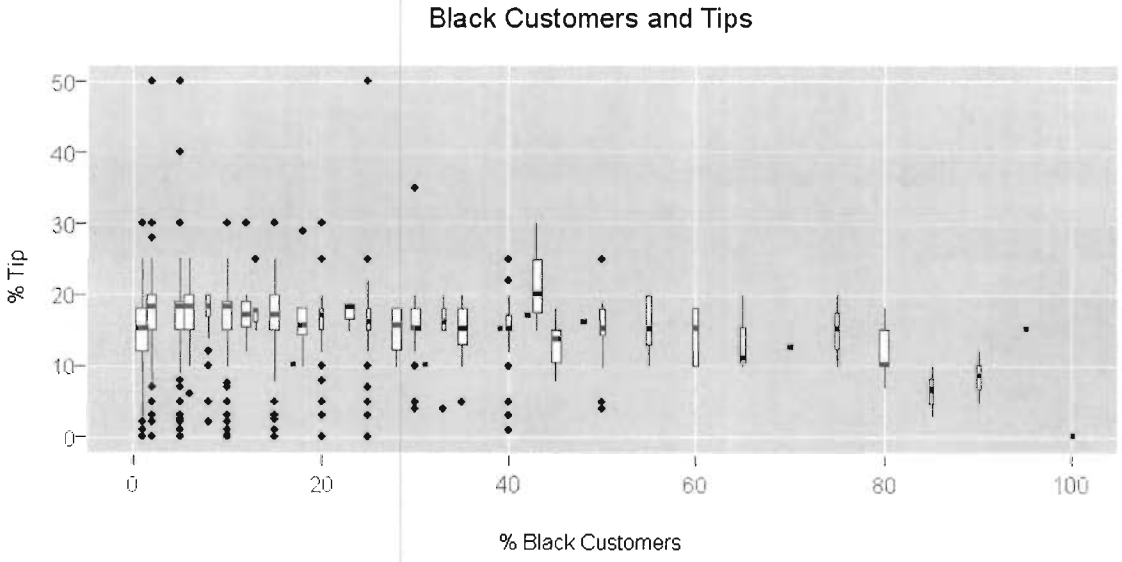


Black

```
> propsb <- !is.na(ss$black_prop) & ss$black_prop < 1
> ss$black_prop[propsb] <- ss$black_prop[propsb] * 100
> ss[!is.na(ss$black_prop) &  ss$black_prop > 100, ]
> ss[!is.na(ss$black_prop) &  ss$black_prop > 100, "black_prop"] <- NA
> qplot(ss$black_prop, type="histogram", scale="count",breaks=seq(0, 100, by=1))
```



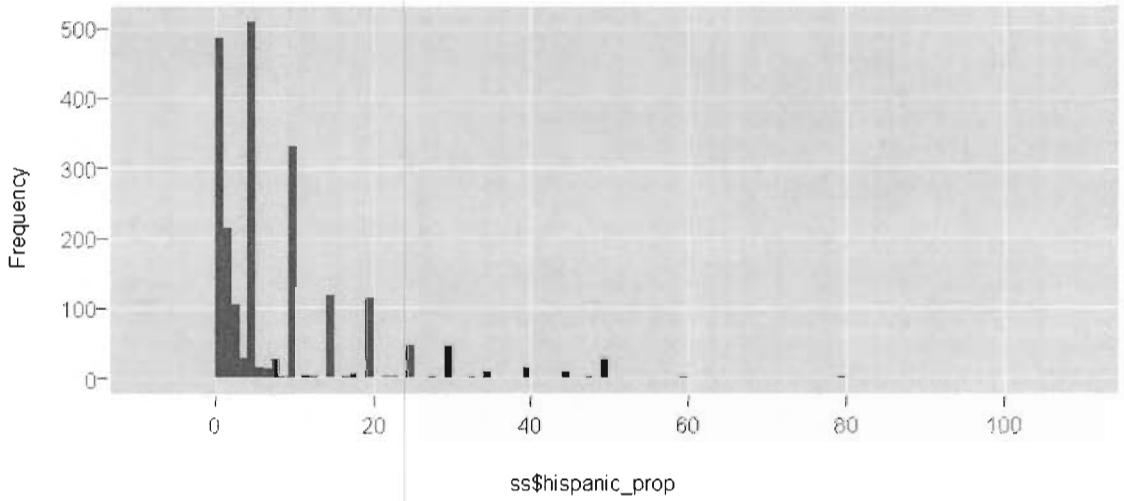

```
> qplot(ss$black_prop, ss$pcttip, type="boxplot", main = "Black Customers and Tips",
xlab="% Black Customers", ylab="% Tip")
```



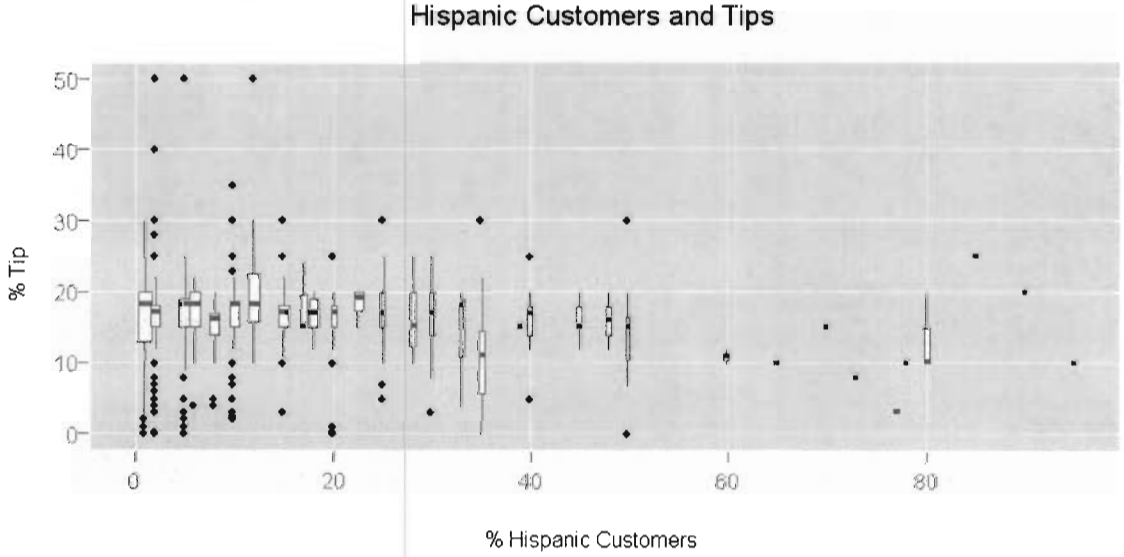
Hispanic

```
> propsh <- !is.na(ss$hispanic_prop) & ss$hispanic_prop < 1
> ss$hispanic_prop[propsh] <- ss$hispanic_prop[propsh] * 100
> ss[!is.na(ss$hispanic_prop) &  ss$hispanic_prop > 100, ]
```

```
> ss[!is.na(ss$hispanic_prop) &  ss$hispanic_prop > 100, "hispanic_prop"] <- NA
> qplot(ss$hispanic_prop, type="histogram", scale="count",breaks=seq(0, 100, by=1))
```

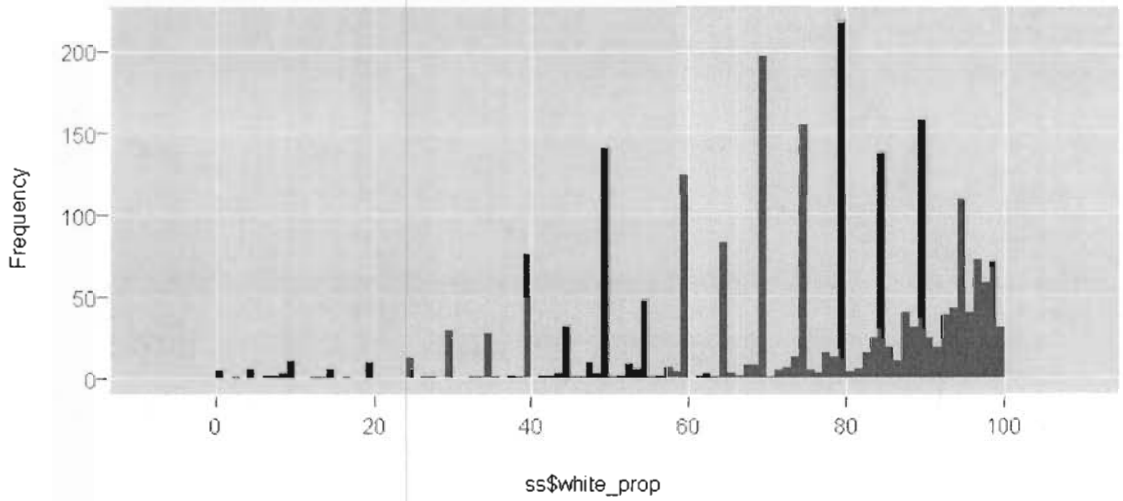



```
> qplot(ss$hispanic_prop, ss$pcttip, type="boxplot", main = "Hispanic Customers and Tips", xlab="% Hispanic Customers", ylab="% Tip")
```

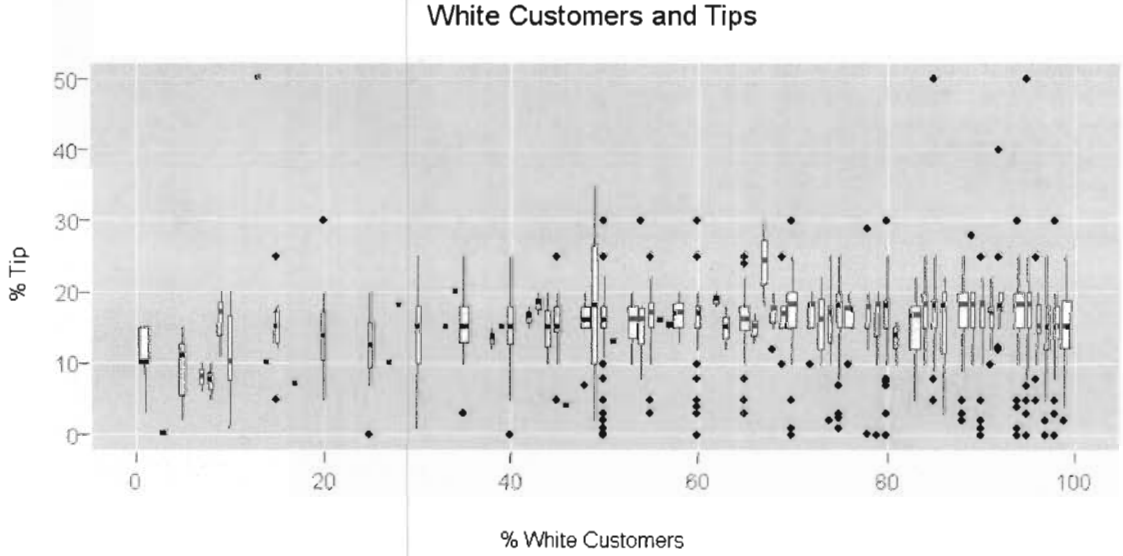


White

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

> qplot(ss$white_prop, type="histogram", scale="count",breaks=seq(0, 100, by=1))
```



> qplot(ss$white_prop, ss$pcttip, type="boxplot", main = "White Customers and Tips", xlab="% White Customers", ylab="% Tip")



These plots show that the percentage of white customers is highest in all restaurants. The average tips were between 15 and 20 % for all restaurants regardless of the race of customers. Because the percentage of white customers was highest in the restaurants, a few records were present for other races. Thus, this data cannot be used to make a compelling case of evidence for tipping habits of people from different races.

Conclusions

Most of the data available is reported by US servers with high proportions of white customers. In US, it is socially customary to leave a tip. Most people would tip anywhere between 10 to 20 %. The amount people leave for tip depends on a large variety of variables, some of which could be the type of restaurant (fast-food, casual dining, etc), quality of food, behavior of the server (a combination of a wide variety of attributes), age of the customer (younger customers- students, etc often tip less) and the overall dining experience of the customer. Even if the overall experience of the customer is not acceptable, customers in US would still leave some minimal tip. Customers from other races like Asians are not historically known to leave a tip, but if present in US would do so.

An attempt was made to carry out an initial exploration of the data to identify any trends. The dataset contains records from several different countries. For a more in-depth analysis, the data should be divided by different variables, such as, country, type of restaurant, age of the customer and also time of the day (customers might tip more for dinner than they do for breakfast).