* C5
S4
O5

Nice report.
You've obviously put a lot of work in to it.

```
# Waiter Tip Survey
#
# Statictics 480
# Homework 8
#
#
================================================================================
#
# Load data
completetipdata <- read.csv(file.choose())

# ------------------------------------------------------------------------------
# PRESENTATION OF THE DATA
# ------------------------------------------------------------------------------
#
# This dataset present 2557 observations and 77 variables. There are several types of variables:
# 54 discrete or integer variables, 12 factors, 9 continuous or numerical variables, and 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
# waiter (experience, habits ...). The last category regroups tipping information (amount, form
# ...).
#
# These information can be verified using the following code:
#
# Dimension of the dataset

dim(completetipdata)

# ---------
#
# Display the structure of the dataset

str(completetipdata)

# ---------
#
# Display the first 6 rows

head(completetipdata)

# ---------
#
# Summarize the different variables

summary(completetipdata)
```

Numerous path can be followed to analyze this dataset because of the amount of information
it contains. However, in this assignment, we will focus our analysis on the relation between
the waiter and the tip amount. A more complete analysis of the dataset will be performed
later (project #2).
#

QUESTIONS ABOUT THE DATA

#
When looking at the different variables available in categories 2 and 3 (respectively waiter
and tip), two groups of questions come to mind.
#
The first is about the relation between the gender, the ethnic origin, the experience and the
age of the waiter on the amount of tip. Those first questions would identify if disparities exist
among waiters.
#
The second interesting element to investigate is the relation between the behavior of the
waiter and the tip he/she receives. The dataset contains a lot of information about the waiter
behavior. In this study, we will focus on the gesture, the facial expression (smile), and the
writing on the bill. This second set of questions should allow us to determine what are the
waiter practices that impact the tip amount.
#
The answer of those questions would give a good first overview of the findings that will
be made by the completion of a deeper analysis of the dataset.
#

PREPARATION OF THE DATA

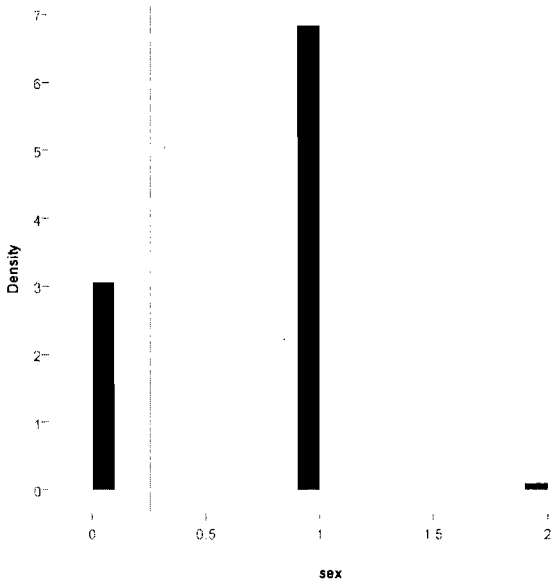
#
As mentioned earlier, we will not use all the 77 variables. Thus, it is necessary to create a new
dataset that will contain only the variables of interest for the analysis. This will reduce the
probability of error and make the calculations more readable.
#
Thus, we need now to select the variables needed to perform the analysis we planned. The
following code performs the process.

```
data <- completetipdata[,c("sex","race","birth_yr","yrs_experience","pcttip","squatt",
"touch","smile","draw","thanks")]
```

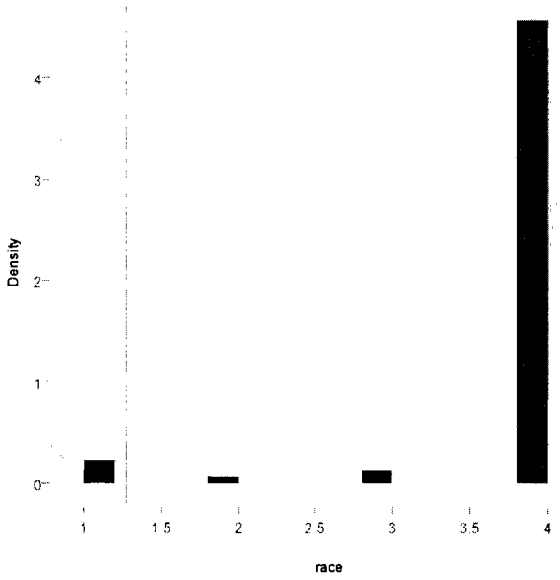
Now, it is necessary to investigate the presence of outliers in the data. A good way is to plot
the data.

Load the graphic package
```
library(ggplot)
```

```
qplot(sex,data=data,type="histogram")
```

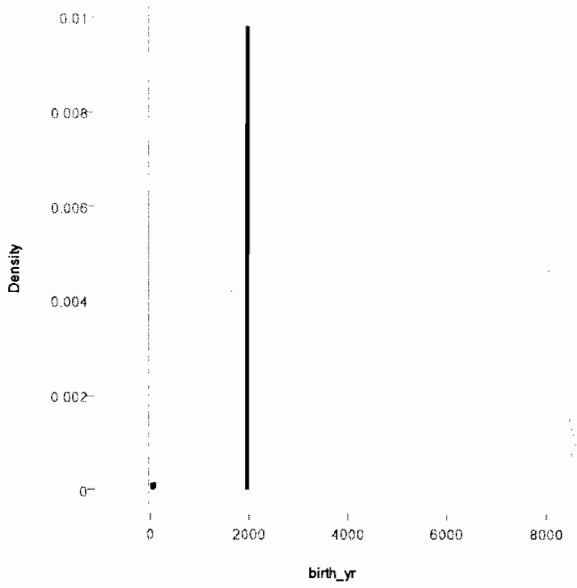



```
# We can see that there is three sex categories: 0, 1, and 2. By exploring the data, we can see that
# 0 is for male, 1 for female, and that 2 does not have any meaning and need to be removed.
# This is done by the following code.

data <- data[data$sex %in% 0:1, ]

# Other variables need to be checked for errors and outliers. First, the race.

qplot(race,data=data,type="histogram")
```



The variable race does not appear containing any outlier. From the previous survey data, we
can deduce that 1 corresponds to Asian, 2 to Black, 3 to Hispanic, and 4 to White.
#
Now, let's investigate the birth years.

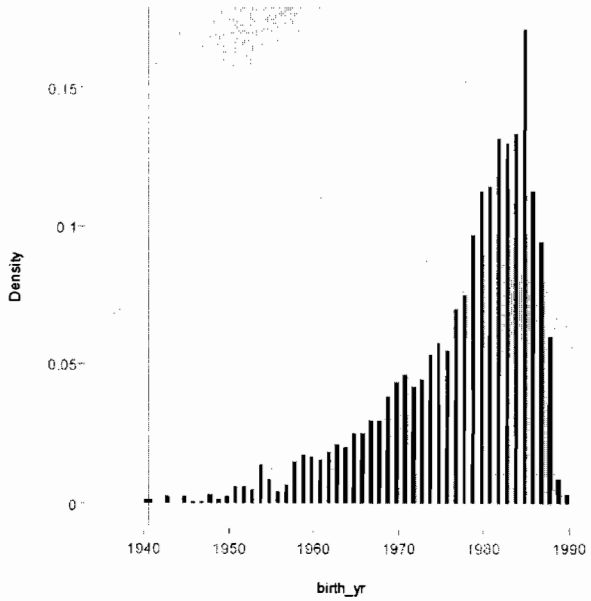
```
qplot(birth_yr,data=data,type="histogram", breaks = 5000)
```



We can see that most of the values are between 1900 and 2000 but some are very close to 0.
This is because people used for instance 81 instead of 1981. Those elements need to be
removed.

```
data <- data[data$birth_yr %in% 1900:1995, ]
qplot(birth_yr,data=data,type="histogram",breaks=100)
```

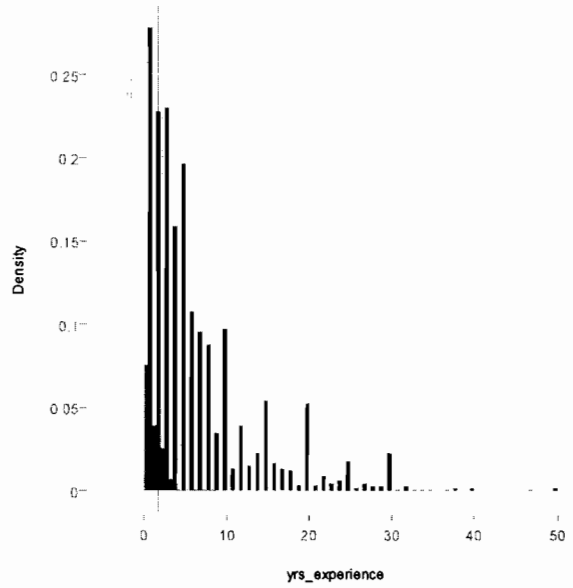
? better to specify seq(1940, 1999, by=1) so you know exactly where the bins are.



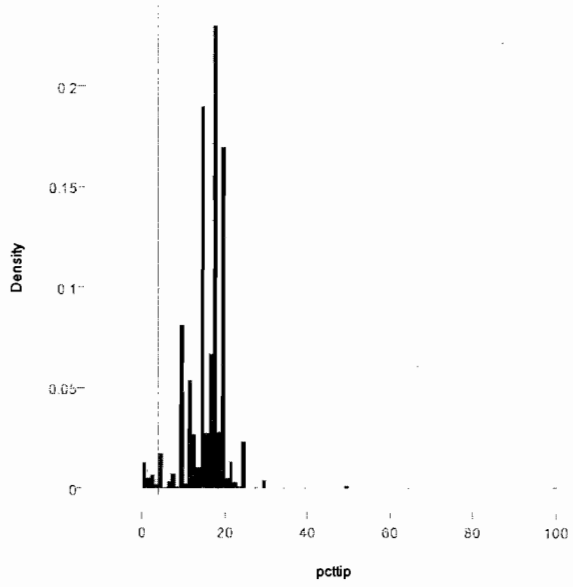
We can see that now, erroneous data were removed. To continue the check of the data, we
will plot the waiter's experience and the percentage of tip.

```
qplot(yrs_experience,data=data,type="histogram",breaks=100)
```

↑ too many! default 30 is sufficient.



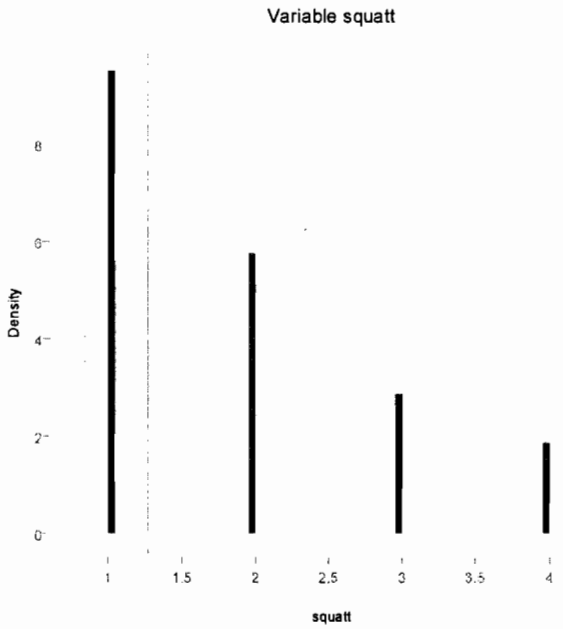

```
qplot(pcttip,data=data,type="histogram",breaks=100)
```



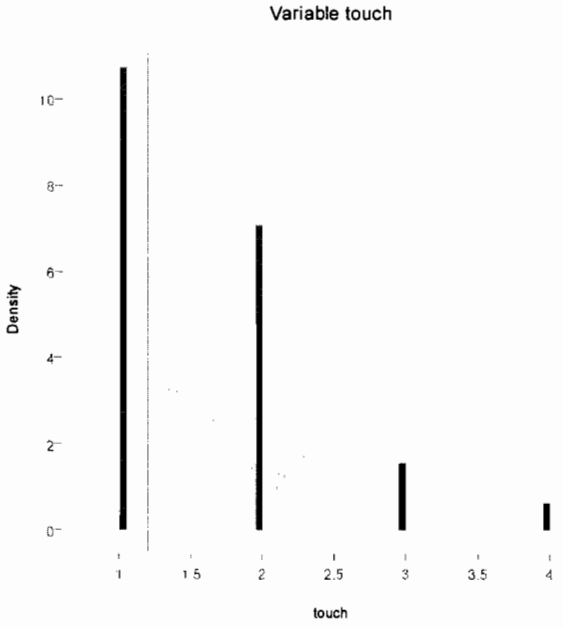
These two variables do not present any obvious outlier.

We now need to check discrete variables that present values of 1 for "never", 2 for
"Sometimes", 3 for "Often", and 4 for "Always".

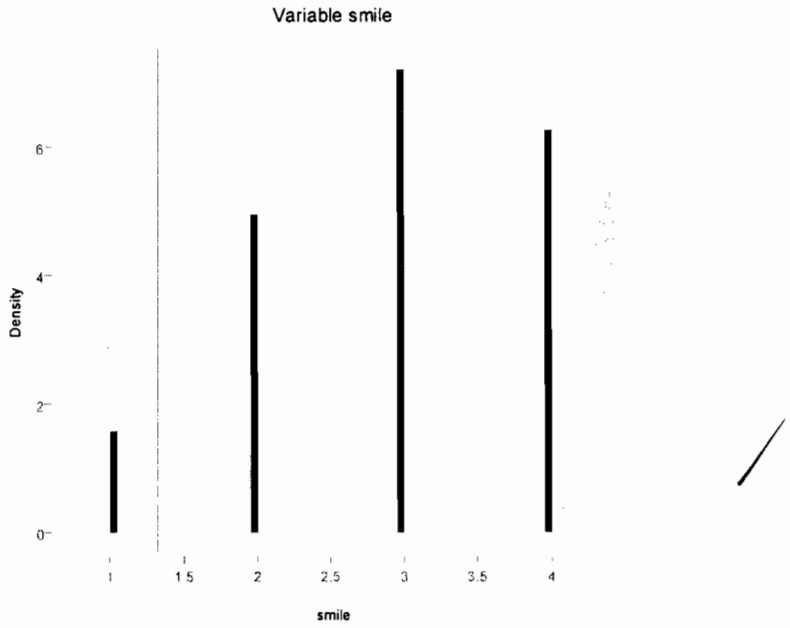
```
qplot(squatt,data=data,type="histogram",breaks=100, main="variable squatt")
```



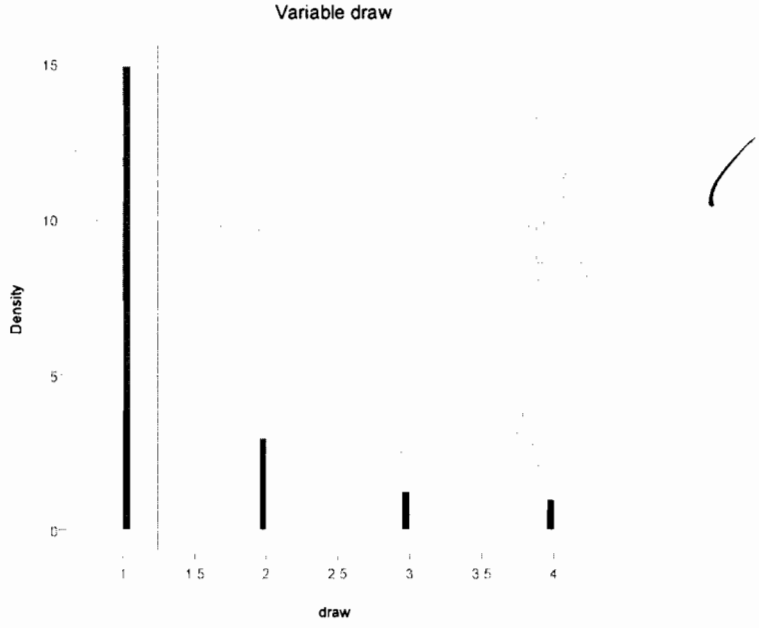

```
qplot(touch,data=data,type="histogram",breaks=100,main="Variable touch")
```



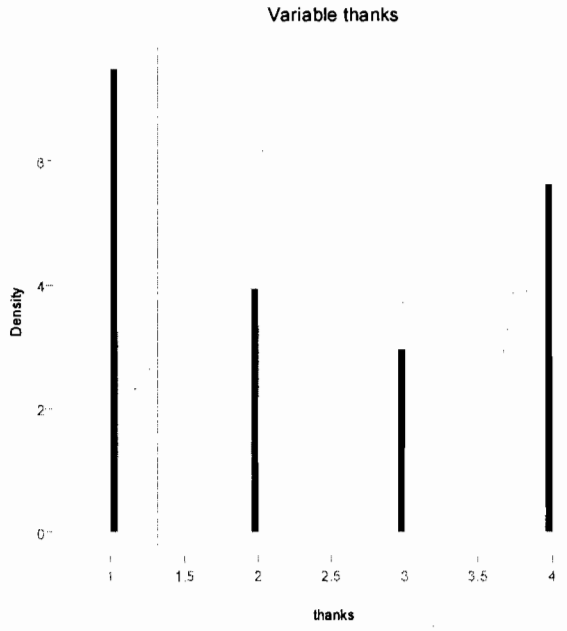
```
qplot(smile,data=data,type="histogram",breaks=100,main="Variable smile")
```




```
qplot(draw,data=data,type="histogram",breaks=100,main="Variable draw")
```



```
qplot(thanks,data=data,type="histogram",breaks=100,main="Variable thanks")
```



Those variables do not present erroneous entries. We can now consider that the dataset
is clean. It does not mean that there is no outlier but at least, there is no false entry.

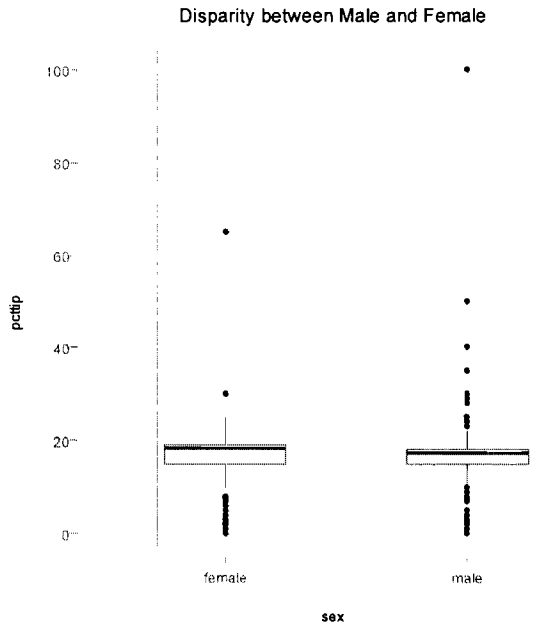
That we can detect by looking at variables individually.

In this assignment, we will use the equivalent of pivot tables in Excel. This is done in R by
using the package "reshape". Thus, it is necessary to load the package and also to prepare the
data by melting them (order the observation). This is done by the following lines of code.

Load reshape. This package is also loaded with ggplot.

```
library(reshape)

datamelt<-melt(data,measure.var=c(5),id.var=c(1,2,3,4,6,7,8,9,10),preserve.na=F)
head(datamelt)

table<-cast(datamelt,sex+race+birth_yr+yrs_experience+squatt+touch+smile+draw+
thanks~variable,mean,na.rm=T)
```

```
# ------------------------------------------------------------------------------------------------------------
# DOES IT EXIST DISPARITY BETWEEN WAITERS REGARDING THE TIPS?
# ------------------------------------------------------------------------------------------------------------
#
# In this part, we are interested in finding if the gender, the origin, the experience and the age of
# the waiter have an impact on the amount of tip.
#
# Let's first investigate the relation between sex and tip.

table$sex=factor(table$sex,levels=c(0,1),labels=c("female","male"))
qplot(sex,pcttip,data=table,type="boxplot",rm.na=T,main="Disparity between Male and
Female")
```




```
# We can see that the median for male is larger than the one for female. It means that males are
# more likely to receive bigger tips than females. We can also observe that one of the value for
# the female is abnormally high. This could be considered as an outlier. The next table shows
# the mean and standard deviation of the both populations.

table1<-cast(datamelt,sex~variable,c(mean,sd))
table1$sex<-factor(table1$sex,levels=c(0,1),labels=c("male","female"))
table1

      sex          pcttip_mean  pcttip_sd
1     male         16.52207     4.653207
2     female       16.03950     5.347817


# We can see that male present a higher average. It would be interesting to determine if this
# difference is significant. This is done by using a t.test.
```

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

Given the outliers is this test valid?

```
        Welch Two Sample t-test

data:  female$value and male$value
t = -2.1031, df = 1441.63, p-value = 0.03563
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.93268768 -0.03245786
sample estimates:
mean of x mean of y
 16.03950  16.52207
```

With a p-value of 0.036, we can reject the non-hypothesis. A significant difference exists
between males and females regarding the amount of tip at 95% confidence interval.

Lets now investigate the impact of the race on the amount of tip. It is first necessary to
determine the proportion of each race on the dataset to ensure the validity of the conclusions.
As shown in page 3, we know that there is much more white waiter that participated to this
survey than other races.

```
ratio <- function(a){
      sum1<-data[data$race %in% a,]
      sum2<-data[data$race %in% 1:4,]
      (sum(sum1$race)/sum(sum2$race))*100
}
```

or table(data$race) / nrow(race).

```
ratio(1)
[1] 1.236643
ratio(2)
[1] 0.6963621
ratio(3)
[1] 2.017049
ratio(4)
[1] 96.04995
```

The proportion of white is much higher and as a consequence, results for this race will be
more trustable than of other races. The following table and graph present the tip mean of the
different races:

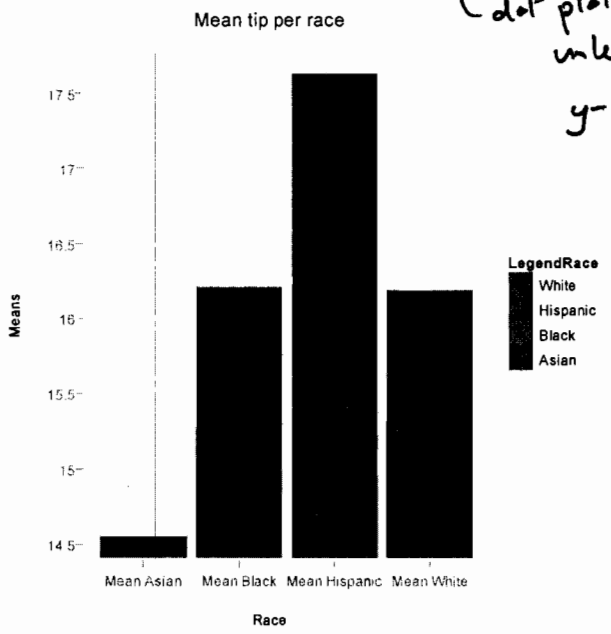
```
table2<-cast(datamelt,race~variable,mean)
table2$race<-c("Asian","Black","Hispanic","White")
table2
```

	race	pcttip
1	Asian	14.55208
2	Black	16.20690
3	Hispanic	17.62963

```
4       White          16.18111
```

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

qplot(Race, Means, data=df, colour=LegendRace, type="bar", main = "Mean tip per race")
```

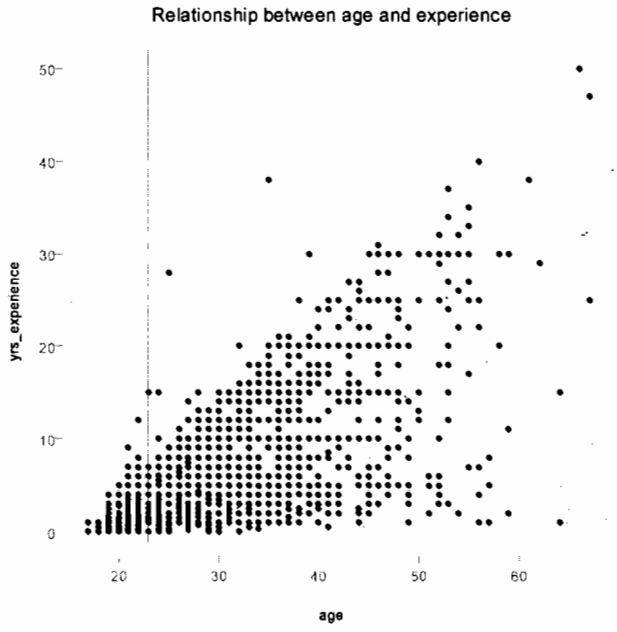
(dot plot is appropriate unless you ensure the y-range is from 0.



It is surprising to see that the tip average for Black and Hispanic waiters is higher than for the
White. It could be explained by the fact that those two races (Black and Hispanic) present a
reduced number of observations and could be as a consequence biased. However, it is
interesting to see that Asian earn the less even if the number of Asian waiter is not
representative of the entire population because of their reduced number in the dataset.

It is now time to analyze the relation between the age and the year of experience of the waiter
on the tip amount. It is first necessary to investigate the relation between age and experience.

```
table$birth_yr<-(2007-table$birth_yr)
age<-table$birth_yr
qplot(age,yrs_experience,data=table,type="point",main="Relationship between age and
experience")
```



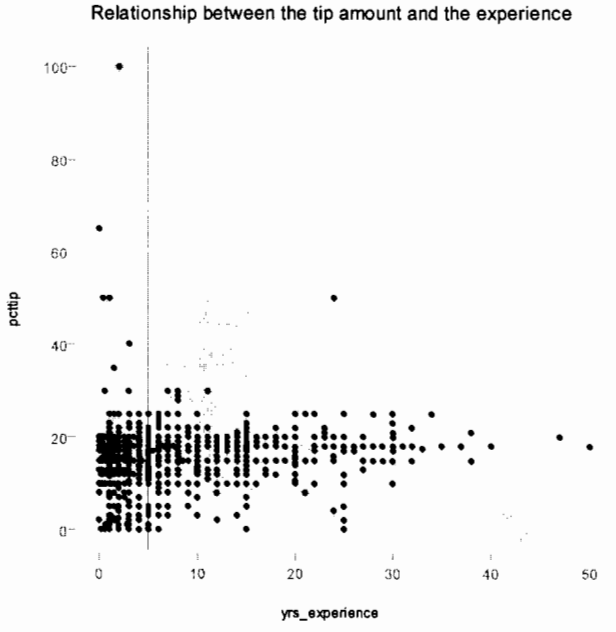
A smooth would be useful as there is a lot of overplotting.

How would you deal with that?

We first can see that some outlier are present in the data. For example, someone says that
he/she is 35 years old and has 38 years of experience. After, the data appears consistent.
There is a very large variability of years of experience for each age but most of the observation
reveal that waiter are mostly young with less than 5 years of experience.

The following plot shows the relation between the years of experience and the tip amount.

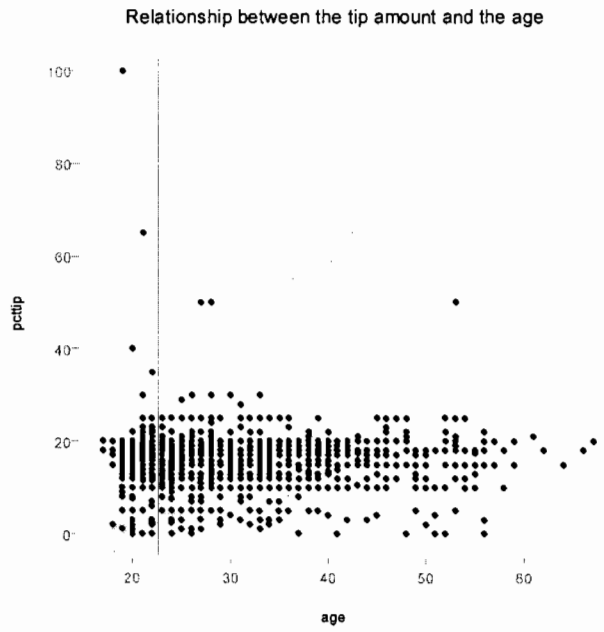
```
qplot(yrs_experience,pcttip,data=table,type="point",main="Relationship between the tip
amount and the experience")
```



We can see that there is no clear pattern. We may be able to say that the more experience the

closer to 20% is the average tip. We can observe that some people report very low tip
amounts. It would be necessary to investigate if these low values are typing errors or true
values because we can clearly see that two waiter with 15 and 25 years of experience present a
tip amount of $0. It is hard to believe. But it could be wrong to remove these values, they
maybe correspond to some restaurants in countries where tipping is not common. It appears
that a upper tip limit exist. Very rarely are the tips that exceed 30%. Higher values could be
considered as outlier or not following the regular US tipping rules. The same observations can
be done for the age of the waiters.

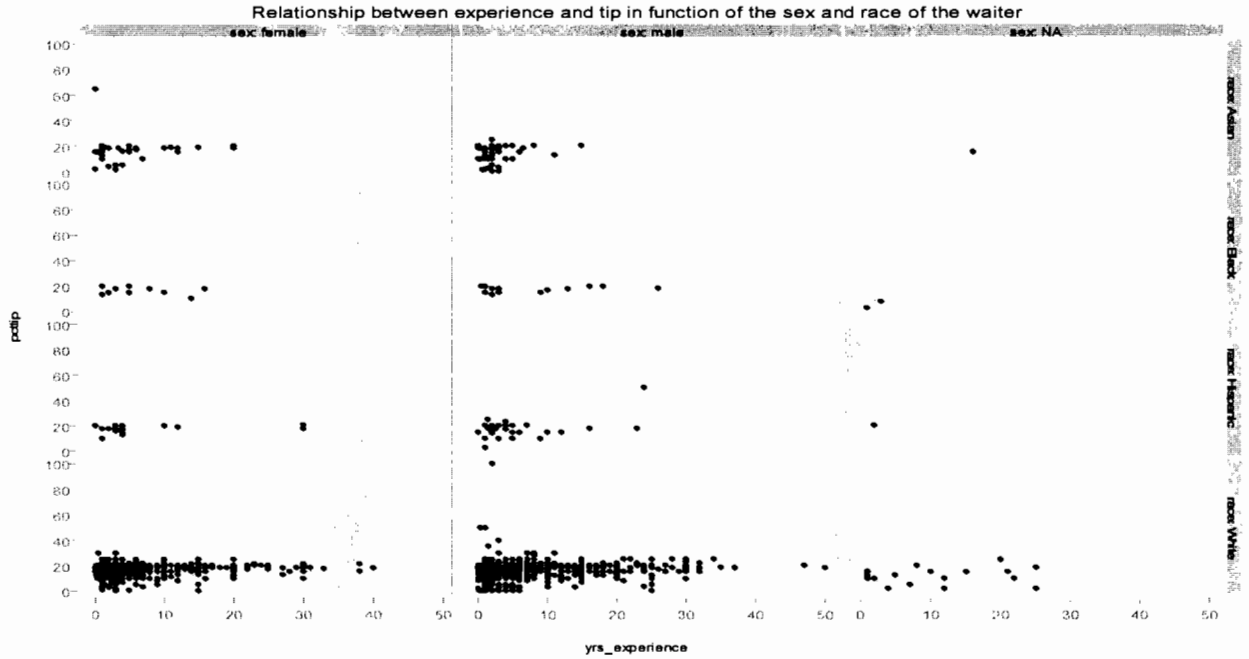
```
qplot(age,pcttip,data=table,type="point", main="Relationship between the tip amount
and the age")
```



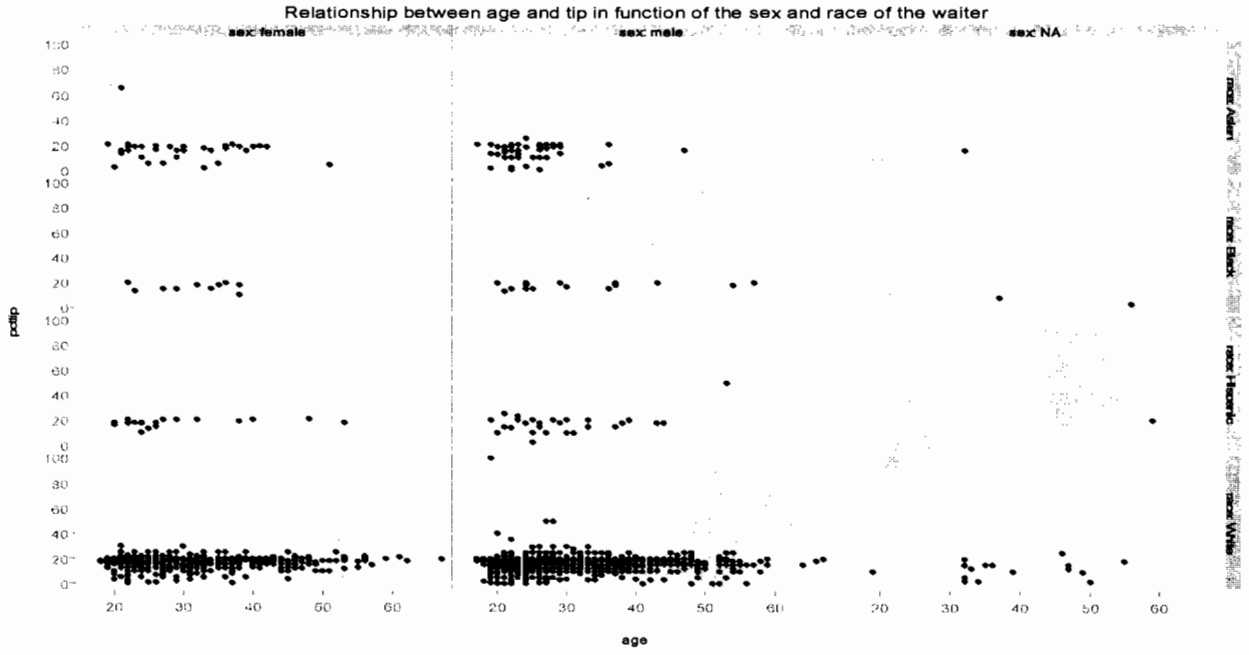
Thus, it appears that the age of the waiter or its experience do not impact the tipping habits of
costumers. It could be interesting to subset those results by race and sex.

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

qplot(yrs_experience,pcttip,data=table,type="point",facet = race~sex,main=
"Relationship between experience and tip in function of the sez and race of the
waiter")
```

Interesting.




```
qplot(age,pcttip,data=table,type="point",facet = race~sex, main= "Relationship between
age and tip in function of the sex and race of the waiter")
```



Again, we do not observe any major trend except the fact that the older or the more
experience a waiter has, the higher the average tip amount is.

As a conclusion of this first part, we can say that the amount of tip is significantly higher for
males than for females. When analyzing per race, we observed that Hispanic and black people

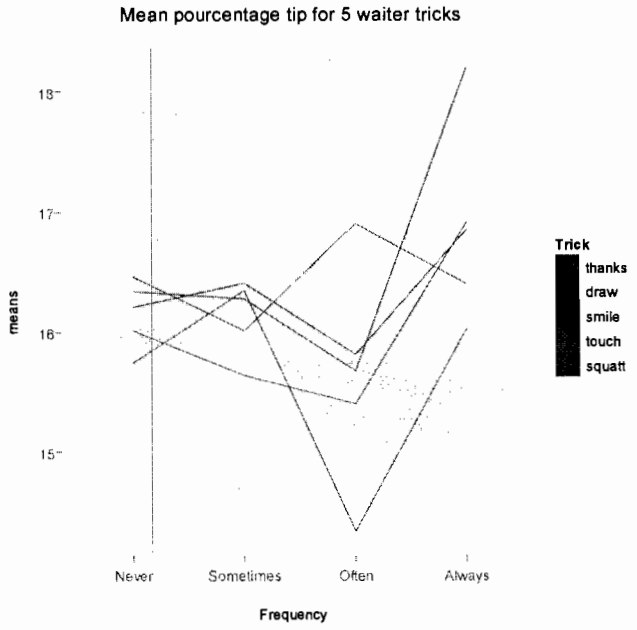
present a higher tip average than white but because the number of white people to participate
in this survey is much larger than for other categories, we cannot really compare them. Also,
we found that the age and the experience of the waiter had an impact on the tip amount that
tended to be higher with a high experience or age. However, several analysis showed that the
preparation of the dataset was not complete and that it would be necessary to clean it more to
avoid any misinterpretation due to outliers.
#
#

IS THERE A RELATION BETWEEN WAITER'S TRICKS AND THE TIPS?

#
To analyze this element, we will take into account five waiter tricks. Those are squat, touch,
smile, draw, and thanks. They respectively correspond to if the waiter sits on the table or
interact with his body, if the waiter touches its costumer, if the waiter smiles, if the waiter
draws on the bill, and if he writes thanks on the bill.
#
The next graph presents the average percentage tip for each "waiter trick" regarding the
frequency at which they do it (never to always).

```
table3<-cast(datamelt,squatt~variable,mean)
table4<-cast(datamelt,touch~variable,mean)
table5<-cast(datamelt,smile~variable,mean)
table6<-cast(datamelt,draw~variable,mean)
table7<-cast(datamelt,thanks~variable,mean)

table3$squatt <- factor(table3$squatt, levels=c(1,2,3,4), labels=c("Never","Sometimes"
,"Often","Always"))
table4$touch <- factor(table4$touch, levels=c(1,2,3,4), labels=c("Never","Sometimes"
,"Often","Always"))
table5$smile <- factor(table5$smile, levels=c(1,2,3,4), labels=c("Never","Sometimes"
,"Often","Always"))
table6$draw <- factor(table6$draw, levels=c(1,2,3,4), labels=c("Never","Sometimes"
,"Often","Always"))
table7$thanks <- factor(table7$thanks, levels=c(1,2,3,4), labels=c("Never","Sometimes"
,"Often","Always"))

squatt<-table3$pcttip
touch<-table4$pcttip
smile<-table5$pcttip
draw<-table6$pcttip
thanks<-table7$pcttip

Frequency <- rep(c("Never","Sometimes","Often","Always"),5)
Frequency <- factor(Frequency,level=c("Never","Sometimes","Often","Always"))
Trick <- rep(c("squatt","touch","smile","draw","thanks"),4)
Trick <- factor(Trick,level=c("squatt","touch","smile","draw","thanks"))

df<-data.frame(
       Frequency,
       Trick,
       means = c(squatt,touch,smile,draw,thanks)
)
```

```
qplot(Frequency,means,data=df,colour=Trick, id=Trick,type = "line",
main = "Mean pourcentage tip for 5 waiter tricks")
```



Nice graph!

We can see that waiters who are always smiling, squatting, and drawing receive more money
that waiter that do it Never, Sometimes, and Often. It is very surprising to see that for these
same three categories, waiter that reported "Often", receive less than waiter doing it Never
and Sometime. It would be interesting to investigate if some outliers are influencing this
category. Concerning waiter touching costumers and writing thanks on the bill, we do not
see a large increase of the tip with an increase of the frequency. Here again, we can observe
that the variable "Often" presents some strange patterns. ✓

CONCLUSIONS

#
This first investigation of the waiter dataset allowed us to determine some important elements
for the future analysis. It is necessary to clean the dataset before drawing calculations because,
as we saw, some elements are very surprising. Thus, a first element would be to separate US's
data with other countries' data because other countries' data are a source of noise and
outliers.

Good point.

Those data were collected through a website. Thus, people could have entered erroneous data
and statisticians have no way to verify them. Thus, all conclusions must mention this element. ✓

Thus, we were able to find that the sex and the race of the waiter appear having an impact on
the tip amount. Also, we saw that waiters always squatting, smiling, and drawing earn more
money. We also found that the tip amount has an upper limit of 20-22%. But those
conclusions must be taken with care because it included US + other countries data and
we found some points that could be considered as outliers.