SMPE Project 2016-2017

Students: Katherine Chao, Elio Keddisseh, Andres Pulido

Dataset

In this report we are exploring the Speed Dating Experiment dataset from Columbia Business School. The dataset is built from data gathered from 2002-2004 speed dating events.
In addition to the dataset csv file, we have a doc file that explains each one of the 195 variables. The data is pretty large (8378 records), and because it was formed by the users who filled the information on papers by hand, it is worth mentioning that we have mistakes and missing values. Moreover, all participants were students in graduate and professional schools at Columbia University, which may lead to a biased conclusion.

Dataset and Dataset Key Legend can be found in our repository or downloaded from:
http://www.stat.columbia.edu/~gelman/arm/examples/speed.dating/

Data Description

The participants had a four minute conversation with their pairs, for each conversation for one individual we have one record. Within each record we can see if there is a match (1) or not (0) from the variable “match“. In addition there was objective (such as Race, Career, Education field, SAT Score and more) and subjective (such as personality information and favorite activities) measures.
Subjective data were on a scale from 1 to 10.
For our analysis we used the following variables in order to answer our hypothesis questions:

1. age (age of the participant)
2. race (race of the participant)
3. gender (Male 1 or female 0)
4. field (Field of study)
5. goal (goal for participating to the speed dating experiment)
6. samerace (Pairs were from the same race. Yes = 1 / No = 0)
7. dec_o (Decision of the partner at the end of the night)
8. Interest of participants in different activities (Gaming, clubbing and many more)
9. go_out (number of times a participant usually go out in a week)

We also created a table in order to sum up and aggregate some of the data. Table desirability contains the desirability ratio which is based on the total number of rounds for each person, the number of matches they got, which then we combined with different variables to compute different correlations.

Replication

In order to replicate our work, the following material is needed:

1. Latest version of R (Can be downloaded from here)
2. The following libraries:
    a. library(dplyr)
    b. library(ggplot2)
    c. library(plotly)
3. Dataset (Can be downloaded from here)

Hypothesis

This large amount of data can answer a lot of different questions indeed, but for our project we decided to focus on the following:

1. Does race have any effect on the choice of the partner ?
2. How does occupation affect the decision of the partner?
3. What are the activities that make a person more or less desirable?
4. Are there certain character traits that make a person more or less desirable?

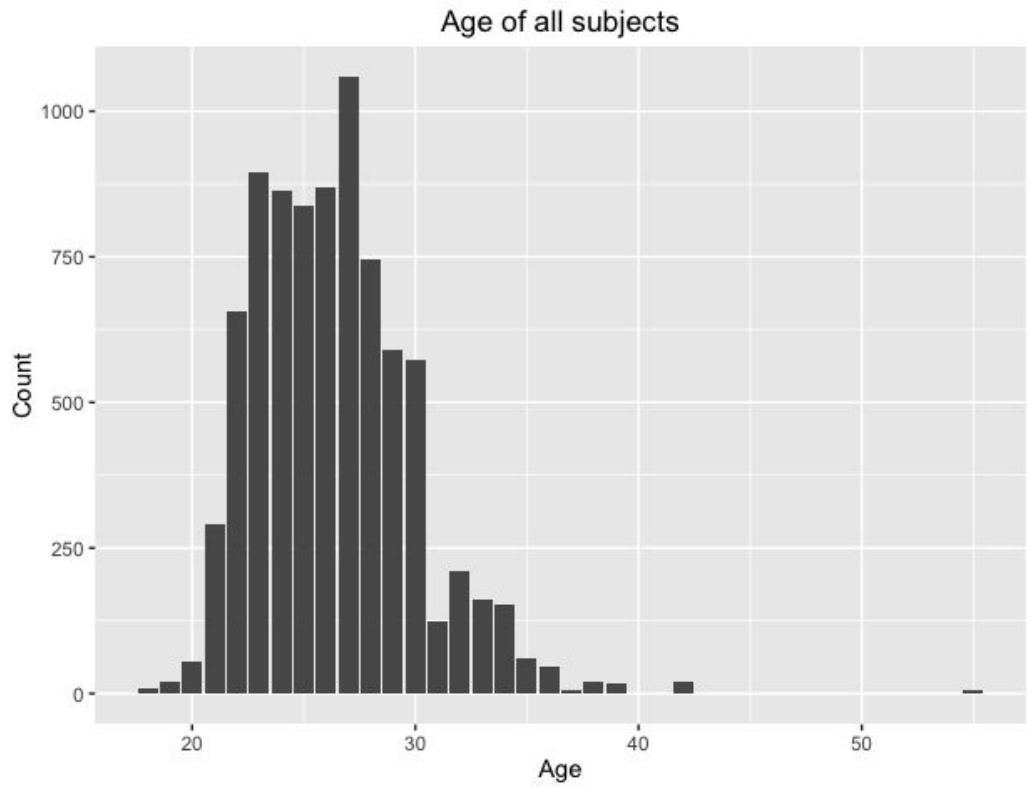
Statistics of the entire data set

First we wanted to examine the general composition of the subjects in the experiment as a whole to get an idea of their identities and motivations. We did not filter the dataset at all for the following analysis.

Around how old are all the subjects?

We can see that most subjects are in their 20's, which makes sense given the context (students). The mode is 27 years of age.

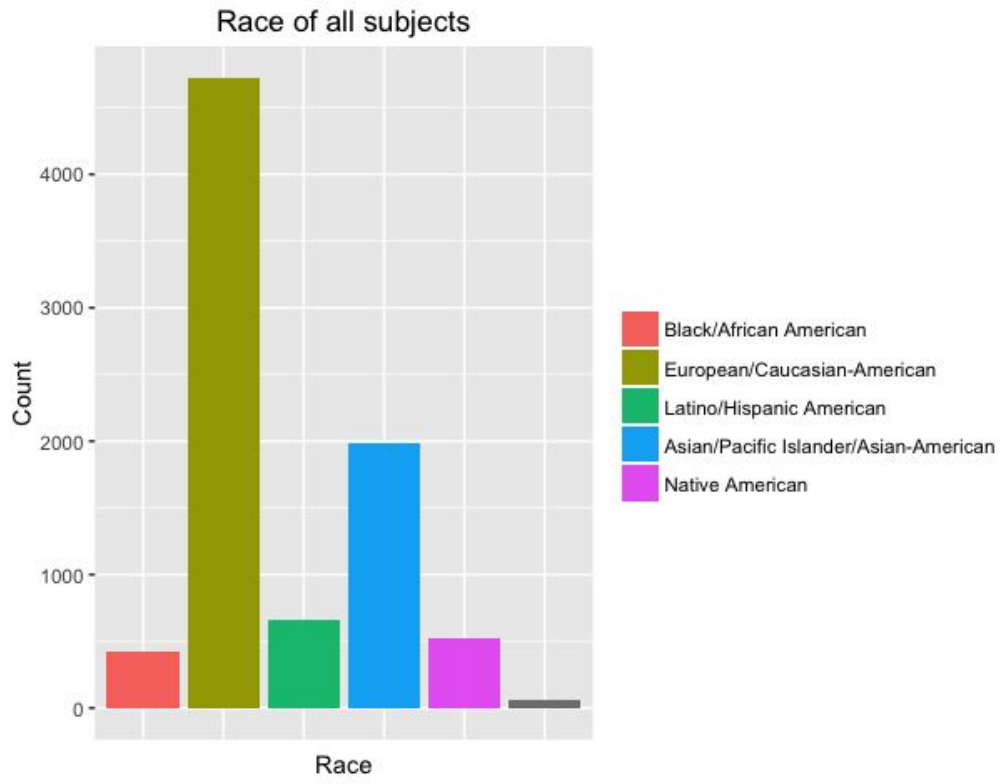
```
ggplot(data=mydata, aes(x=mydata$age, fill = mydata$age)) +
geom_bar(stat="count", position = "stack") +
  guides(color = "legend") +
  theme(axis.text.x=element_blank(), axis.ticks.x=element_blank(),
legend.position = "right") +
  labs(fill = "") +
  xlab("Age") + ylab("Count") +
  ggtitle("Age of all subjects")
```




```
ggplot(data=mydata, aes(x=mydata$race, fill = mydata$race)) +
geom_bar(stat="count", position = "stack") +
  guides(color = "legend") +
  theme(axis.text.x=element_blank(), axis.ticks.x=element_blank(),
legend.position = "right") +
  labs(fill = "") +
  xlab("Race") + ylab("Count") +
  ggtitle("Race of all subjects")
```

What race are all the subjects?

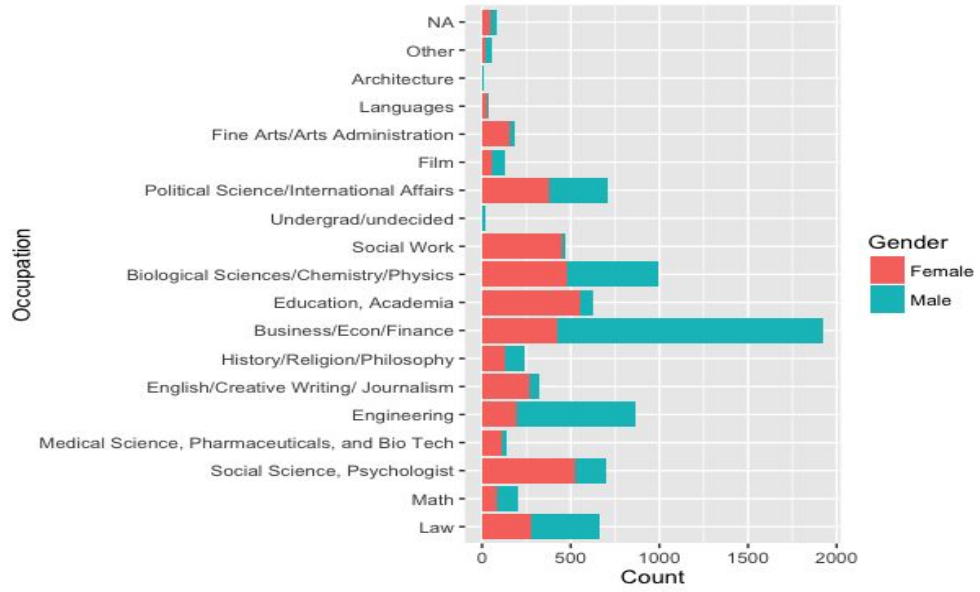
The subjects were all fairly diverse, with Caucasian being the largest demographic and with quite a lot of Asians as well. The least represented demographic is Black.



What field are people in (with a breakdown by gender)?

Here we can see that Business/Econ/Finance is by far the most represented field, and with a large proportion being male. Engineering is also another prominently male-majority field. Female-dominated fields include Social Work and English.

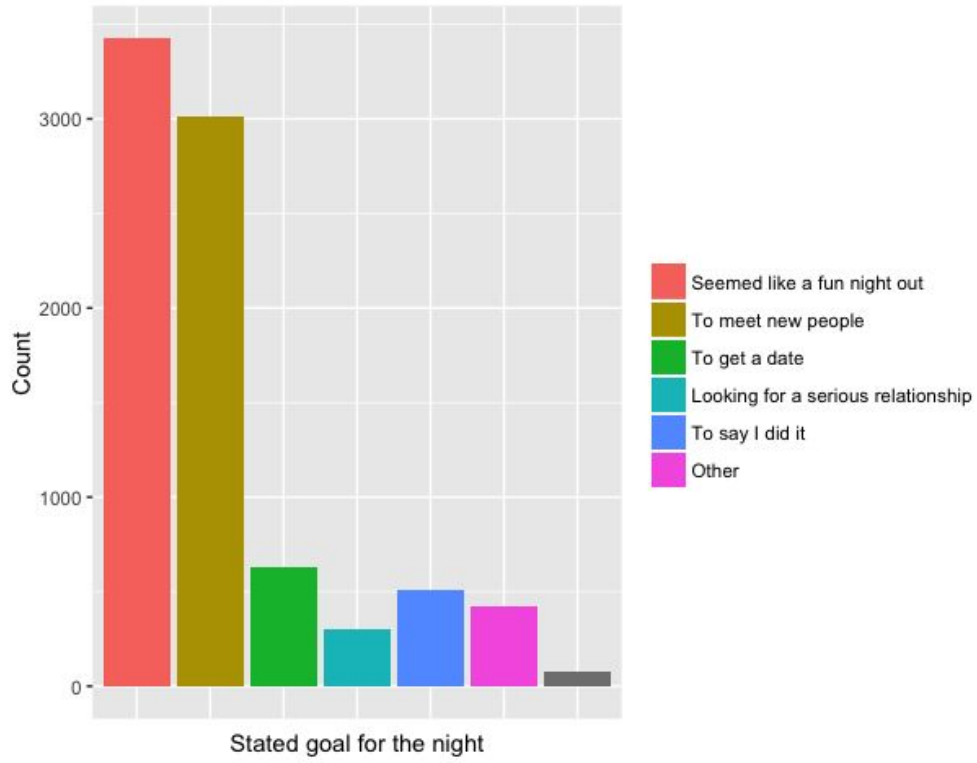
```
ggplot(data=mydata, aes(x=field_cd, fill = factor(gender))) +
geom_bar(stat="count") +
  coord_flip() +
  xlab("Occupation") + ylab("Count") + labs(fill = "Gender")
```



What are the intended goals of people attending this event?

People overwhelming chose to attend this event either because it seemed like a fun night out or to meet new people.

```
ggplot(data=mydata, aes(x=mydata$goal, fill = mydata$goal)) +
geom_bar(stat="count", position = "stack") +
  guides(color = "legend") +
  theme(axis.text.x=element_blank(), axis.ticks.x=element_blank(),
legend.position = "right") +
  labs(fill = "") +
  xlab("Stated goal for the night") + ylab("Count")
```



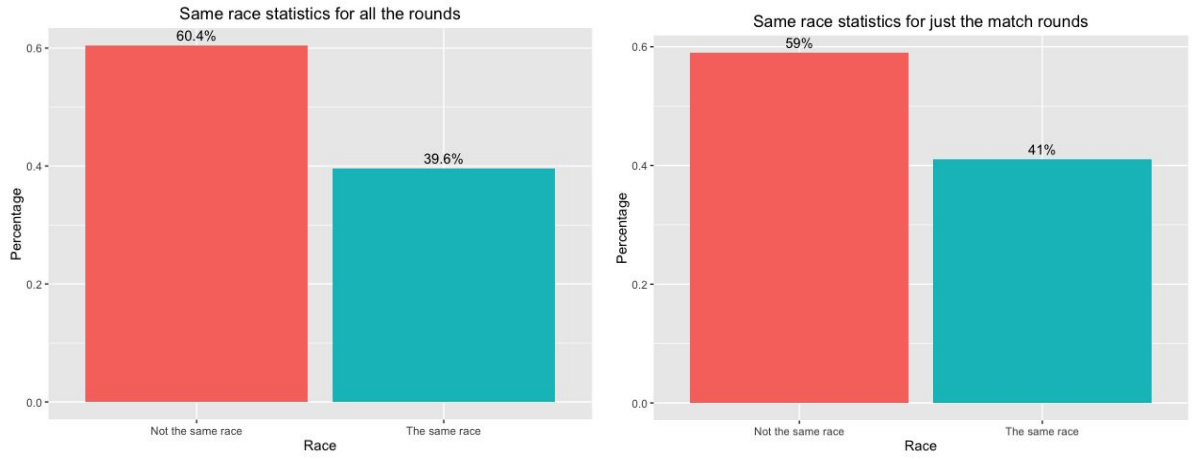
Statistics of the matches

Next we wanted to study those that got a match (a match being defined as a round in which both the person and their partner having mutual attraction and agreeing to see each other for a next date).

Is being the same race a good predictor of a mutual match?
Comparing the rounds of all the people vs. the rounds of only the people who were matches, we see that the breakdown in terms of whether the people were the same race or not is more or less the same. Being of the same or different race does not seem to affect whether the two were a match or not.

```
ggplot(data=mydata, aes(x=samerace, fill = mydata$samerace)) + geom_bar(aes(y =
(..count..)/sum(..count..)), stat="count") +
  geom_text(aes(y = ((..count..)/sum(..count..)), label =
scales::percent((..count..)/sum(..count..))), stat = "count", vjust = -.5) +
  theme(legend.position="none") +
  xlab("Race") + ylab("Percentage") +
  ggtitle("Same race statistics for all the rounds")
```

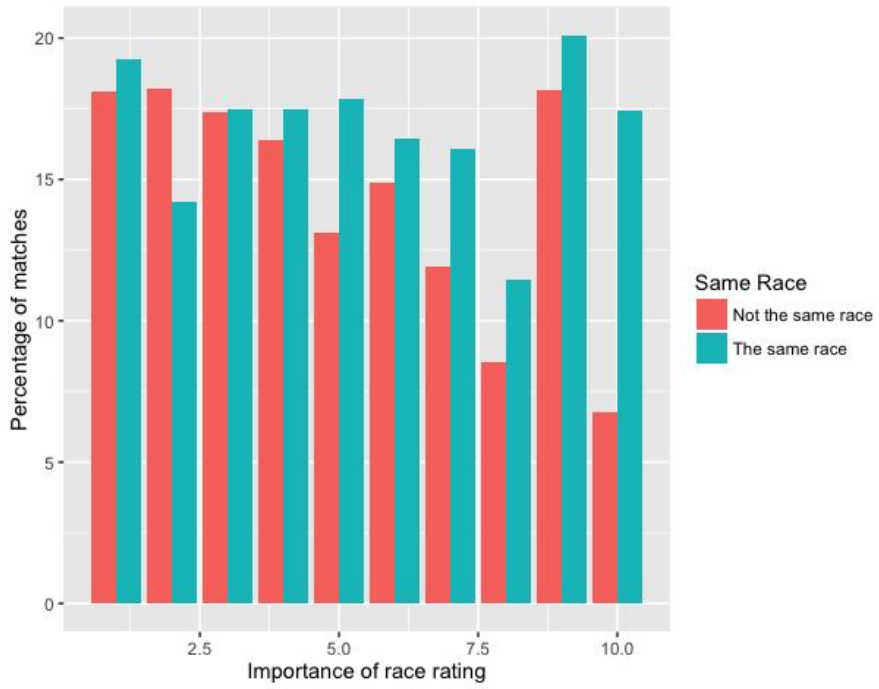
```
ggplot(data=areMatches, aes(x=samerace, fill = areMatches$samerace)) +
geom_bar(aes(y = (..count..)/sum(..count..)), stat="count") +
  geom_text(aes(y = ((..count..)/sum(..count..)), label =
scales::percent((..count..)/sum(..count..))), stat = "count", vjust = -.5) +
  theme(legend.position="none") +
  xlab("Race") + ylab("Percentage") +
  ggtitle("Same race statistics for just the match rounds")
```



Do people who say they prefer the same race really choose people of the same race?
In the study, people were asked to rate on a scale from 1 (low importance) to 10 (high importance) how important it was to them that the other person was the same race. We thought it would be interesting to see what people expected of themselves vs. their actual actions. It turns out that people more or less stuck to what they expected of themselves, with people who rated this very highly having a large gap between matching with people of the same race as opposed to not.

```
imprace.importance <- mydata %>%
  group_by(samerace, imprace) %>%
  summarise(sum.match = sum(match),
            total = n())
imprace.importance <- imprace.importance[imprace.importance$imprace > 0 &
!is.na(imprace.importance$imprace),]
ggplot(imprace.importance, aes(x = imprace, y = (sum.match / total) * 100,
fill = factor(samerace))) +
  geom_bar(stat = "identity", position = "dodge") +
  xlab("Importance of race rating") + ylab("Percentage of matches") +
```

```
labs(fill = "Same Race")
```



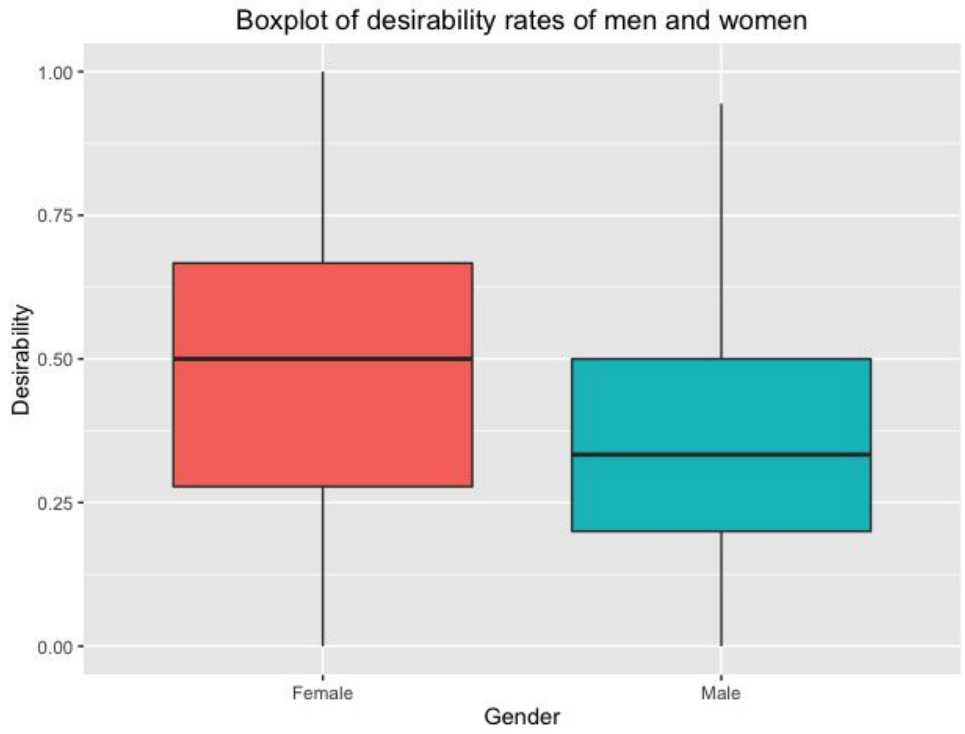
Statistics of desirability

Finally, we wanted to be able to do analysis on the characteristics of those that had the highest or lowest desirability. To that end, we computed a “desirability ratio” which is based on the total number of rounds a person had and the number, out of those rounds, that liked them. For example, if someone saw 10 people in total and 5 people liked them, then their desirability would be 0.5.

```
desirability <- aggregate(mydata$dec_o ~ mydata$iid, mydata, sum)
a = c(desirability$`mydata$iid`)
b = c(desirability$`mydata$dec_o`/table(mydata$iid))
desirability = data.frame(a, b)

names(desirability)[names(desirability) == 'a'] <- 'iid'
names(desirability)[names(desirability) == 'b'] <- 'des'
desirability_merge <- merge(mydata, desirability, by="iid", all=FALSE)
```

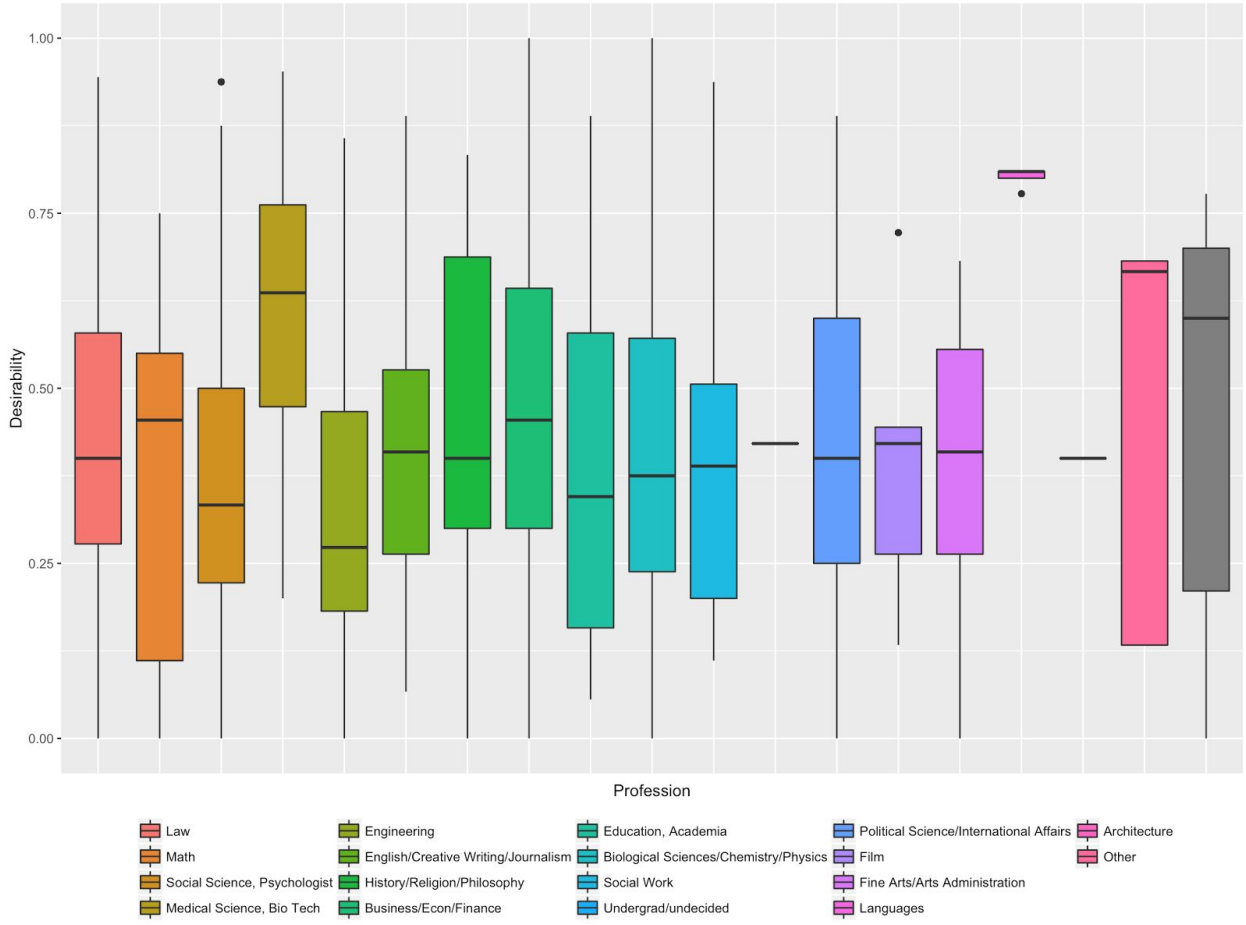
How desirable are the females compared to the desirability of the males?
We were first interested in whether it was easier for men or women to get dates. It turns out that gender does indeed matter and we can see that women are desired at higher rates than men. The top 3rd quartile man has just about the same desirability rate (~50%) as the average woman.

```
ggplot(desirability_merge, aes(x = desirability_merge$gender, y =
desirability_merge$des, fill = desirability_merge$gender)) +
  xlab("Gender") + ylab("Desirability") +
  theme(legend.position="none") +
  geom_boxplot() +
  ggtitle("Boxplot of desirability rates of men and women")
```



What professions are the most desirable?
Our next analysis concerned the profession of the person. We can see from the graph that those in Medicine had the highest desirability rate while Engineering had the lowest median desirability of all the professions.

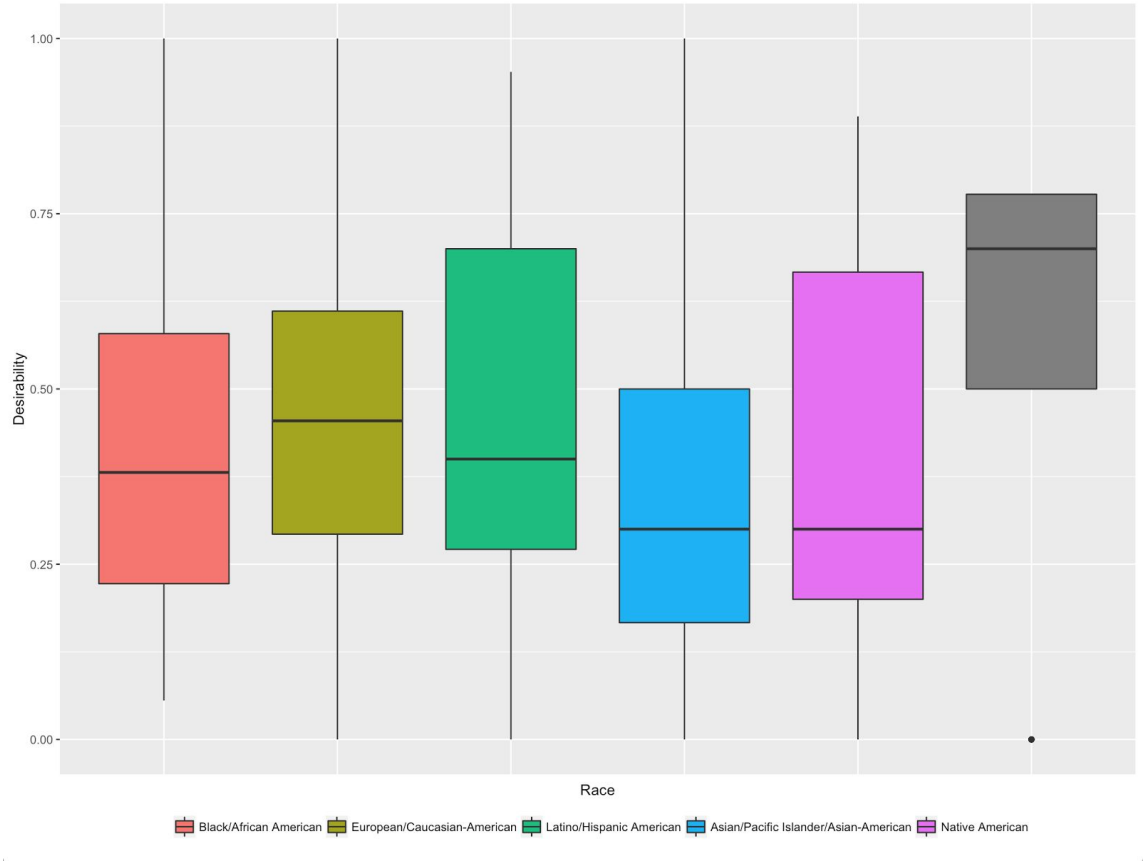
```
ggplot(desirability_merge, aes(x = desirability_merge$field_cd, y =
desirability_merge$des, fill = desirability_merge$field_cd)) +
  xlab("Profession") + ylab("Desirability") +
  theme(axis.text.x=element_blank(), axis.ticks.x=element_blank(),
legend.position = "bottom") +
  geom_boxplot() +
  labs(fill = "")
  ggtitle("Boxplot of desirability rates by profession")
```



What about race?

Are there big variations in desirability among each of the races? It turns out that there are some slight differences, with Latino and Native American having the highest variability and Caucasians with the lowest variability. The median for Caucasians was the highest.

```
ggplot(desirability_merge, aes(x = desirability_merge$race, y =
desirability_merge$des, fill = desirability_merge$race)) +
  xlab("Race") + ylab("Desirability") +
  theme(axis.text.x=element_blank(), axis.ticks.x=element_blank(),
legend.position = "bottom") +
  geom_boxplot() +
  labs(fill = "")
  ggtitle("Boxplot of desirability rates by race")
```

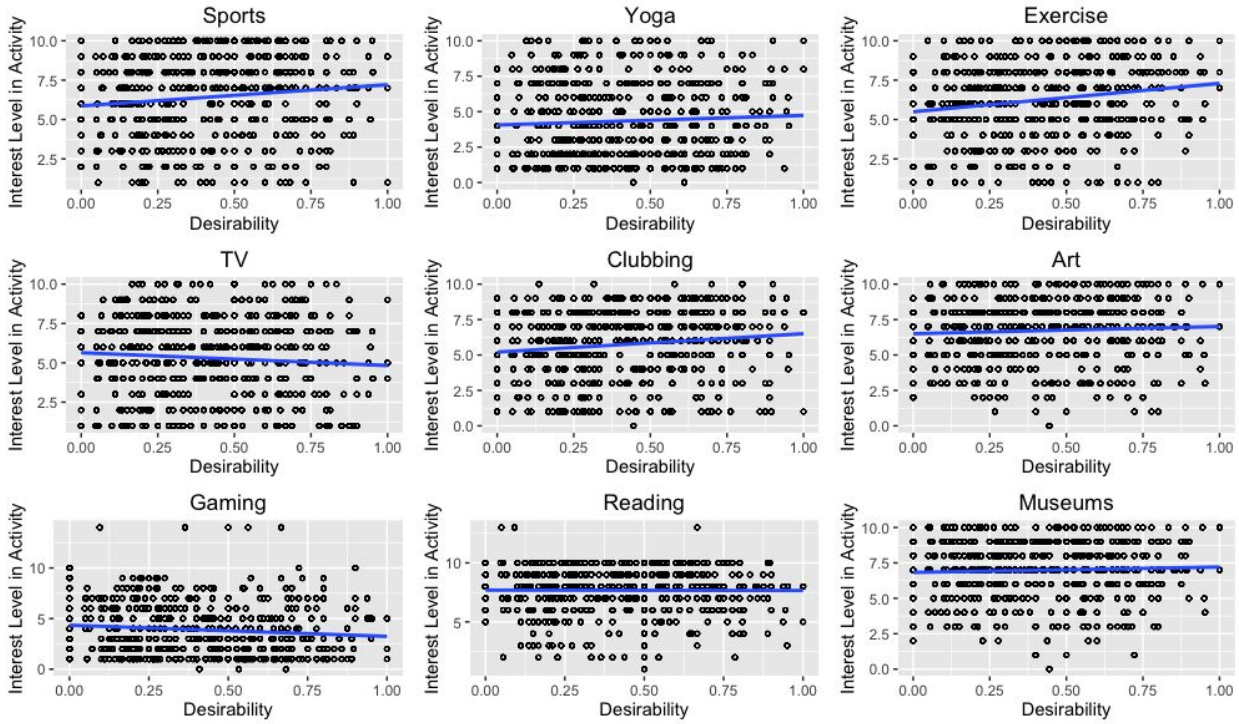


Are hobbies a good predictor of desirability?

People in the study were asked to rank on a scale of 1-10 their interest in certain hobbies. Maybe an interest in some activities made some people more attractive (or less attractive). Plotted below is a scatterplot between the desirability rate and their interest in a certain hobby. For the most part, the regression lines are pretty flat, with a slightly high positive correlation in the hobbies Clubbing, Sports, and Exercise and a slightly negative correlation in the hobbiest TV and Gaming.

```
sportsp <- ggplot(desirability_merge, aes(x=desirability_merge$des,
y=desirability_merge$sports)) +
  geom_point(shape = 1, size = 1) +
  xlab("Desirability") + ylab("Interest Level in Activity") +
  ggtitle("Sports") +
  geom_smooth(method=lm)

yogap <- ggplot(desirability_merge, aes(x=desirability_merge$des,
y=desirability_merge$yoga)) +
  geom_point(shape = 1, size = 1) +
  xlab("Desirability") + ylab("Interest Level in Activity") +
  ggtitle("Yoga") +
  geom_smooth(method=lm)

exercisep <- ggplot(desirability_merge, aes(x=desirability_merge$des,
y=desirability_merge$exercise)) +
  geom_point(shape = 1, size = 1) +
  xlab("Desirability") + ylab("Interest Level in Activity") +
  ggtitle("Exercise") +
  geom_smooth(method=lm)

tvp <- ggplot(desirability_merge, aes(x=desirability_merge$des,
y=desirability_merge$tv)) +
  geom_point(shape = 1, size = 1) +
  xlab("Desirability") + ylab("Interest Level in Activity") +
  ggtitle("TV") +
  geom_smooth(method=lm)

clubbingp <- ggplot(desirability_merge, aes(x=desirability_merge$des,
y=desirability_merge$clubbing)) +
  geom_point(shape = 1, size = 1) +
  xlab("Desirability") + ylab("Interest Level in Activity") +
  ggtitle("Clubbing") +
  geom_smooth(method=lm)
```

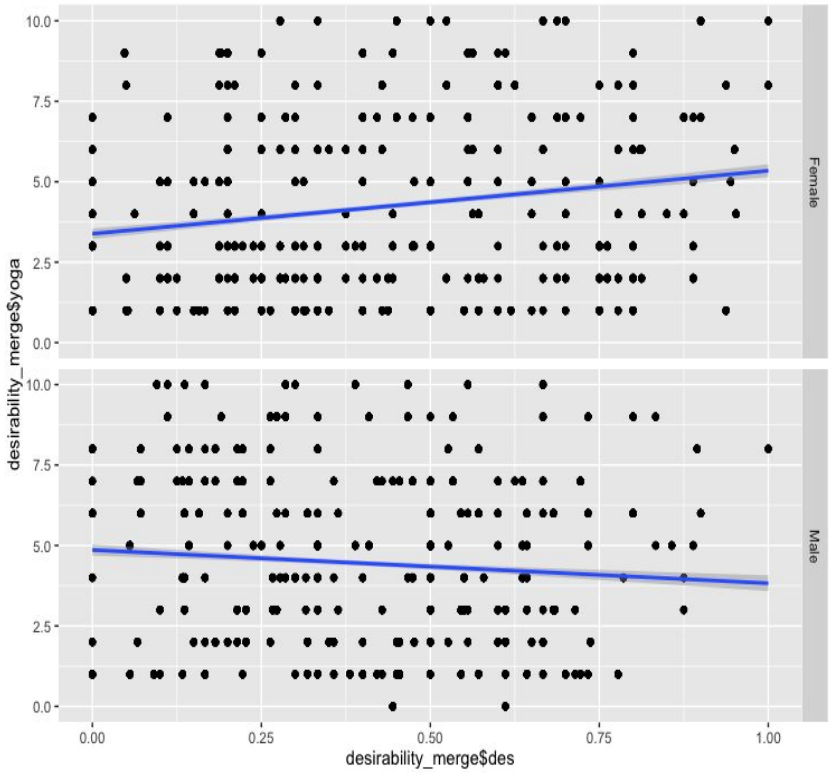
```
artp <- ggplot(desirability_merge, aes(x=desirability_merge$des,
y=desirability_merge$art)) +
  geom_point(shape = 1, size = 1) +
  xlab("Desirability") + ylab("Interest Level in Activity") +
  ggtitle("Art") +
  geom_smooth(method=lm)

gamingp <- ggplot(desirability_merge, aes(x=desirability_merge$des,
y=desirability_merge$gaming)) +
  geom_point(shape = 1, size = 1) +
  xlab("Desirability") + ylab("Interest Level in Activity") +
  ggtitle("Gaming") +
  geom_smooth(method=lm)

readingp <- ggplot(desirability_merge, aes(x=desirability_merge$des,
y=desirability_merge$reading)) +
  geom_point(shape = 1, size = 1) +
  xlab("Desirability") + ylab("Interest Level in Activity") +
  ggtitle("Reading") +
  geom_smooth(method=lm)

museumsp <- ggplot(desirability_merge, aes(x=desirability_merge$des,
y=desirability_merge$museums)) +
  geom_point(shape = 1, size = 1) +
  xlab("Desirability") + ylab("Interest Level in Activity") +
  ggtitle("Museums") +
  geom_smooth(method=lm)

grid.arrange(sportsp, yogap, exercisep, tvp, clubbingp, artp, gamingp,
readingp, museumsp, ncol=3)
```



And does this attractiveness by hobby vary by gender?

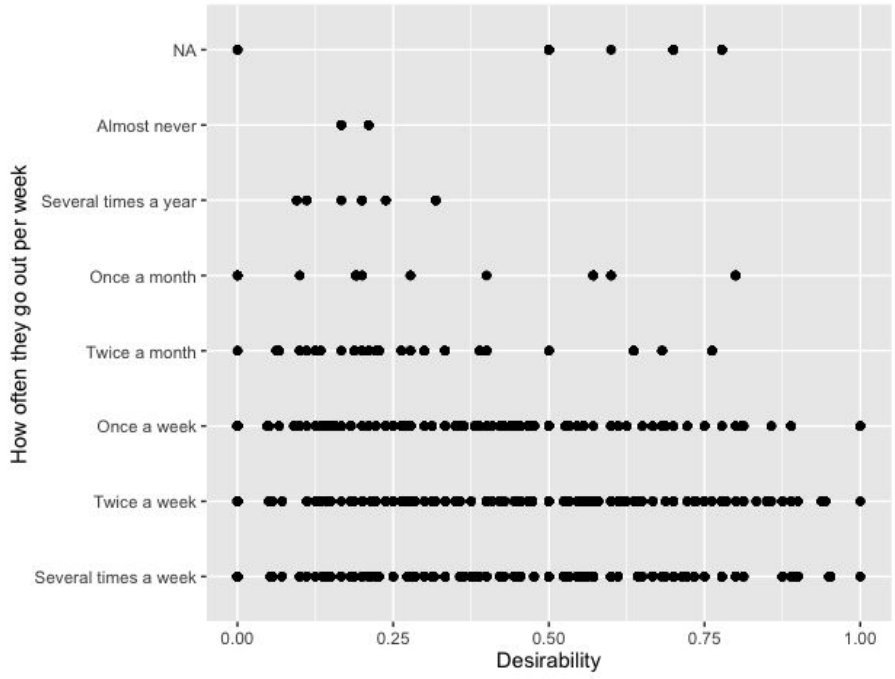
It turns out that it does not very much, except in the case of yoga, in which women had a slightly positive correlation between a high interest in yoga and higher desirability, while men had a slightly negative correlation.

```
ggplot(desirability_merge, aes(x=desirability_merge$des,
y=desirability_merge$yoga)) +
  geom_point() +
  facet_grid(gender ~ .) +
  geom_smooth(method=lm)
```



Is going out a lot (ie. being more social) a good predictor of desirability?
Next, we wanted to see the relationship between going out and desirability. The participants were asked to rank the frequency that they went out in the study. Because it's categorical data, can't draw regression curve but we can see a slight correlation between the amount one goes out and their desirability rate.

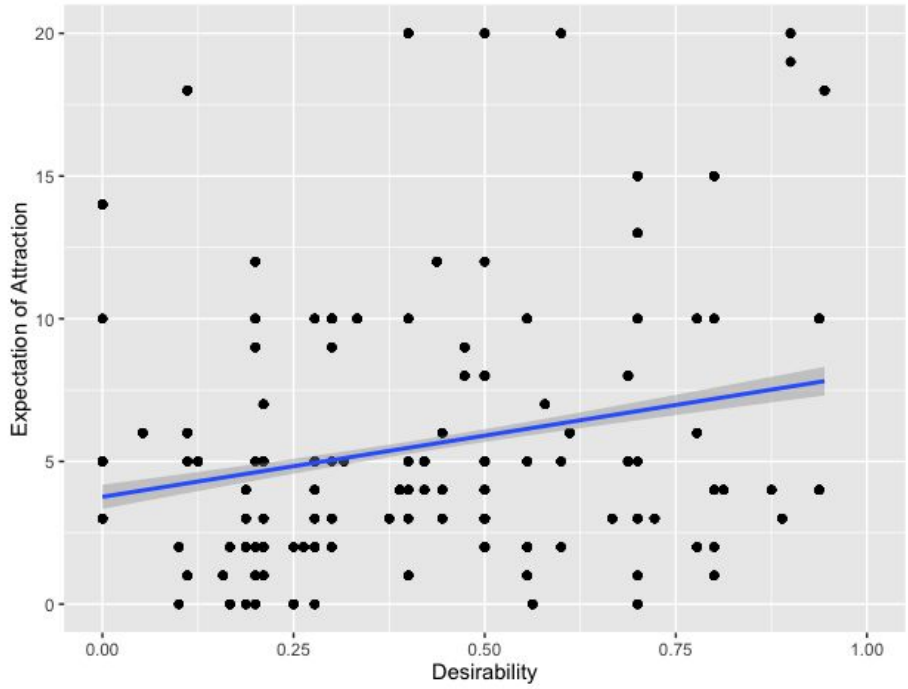
```
ggplot(desirability_merge, aes(x=desirability_merge$des,
y=desirability_merge$go_out)) +
  geom_point() +
  xlab("Desirability") + ylab("How often they go out")
```



Are people delusional or realistic about themselves? (Is your expectation of your own attraction a good indicator of your actual desirability?)

Finally, we wanted to see whether people's perception of their own desirability was accurate or not. People in the study were asked to predict how many people, out of 20, would want to see them for another date. We can see that there is indeed a slight positive correlation between what people expected and their actual desirability, although it seemed overall people underestimated their own attractiveness when compared to the rate that they actually attracted people.

```
ggplot(desirability_merge, aes(x=desirability_merge$des,
y=desirability_merge$expnum)) +
  geom_point() +
  xlab("Desirability") + ylab("Expectation of Attraction")
```

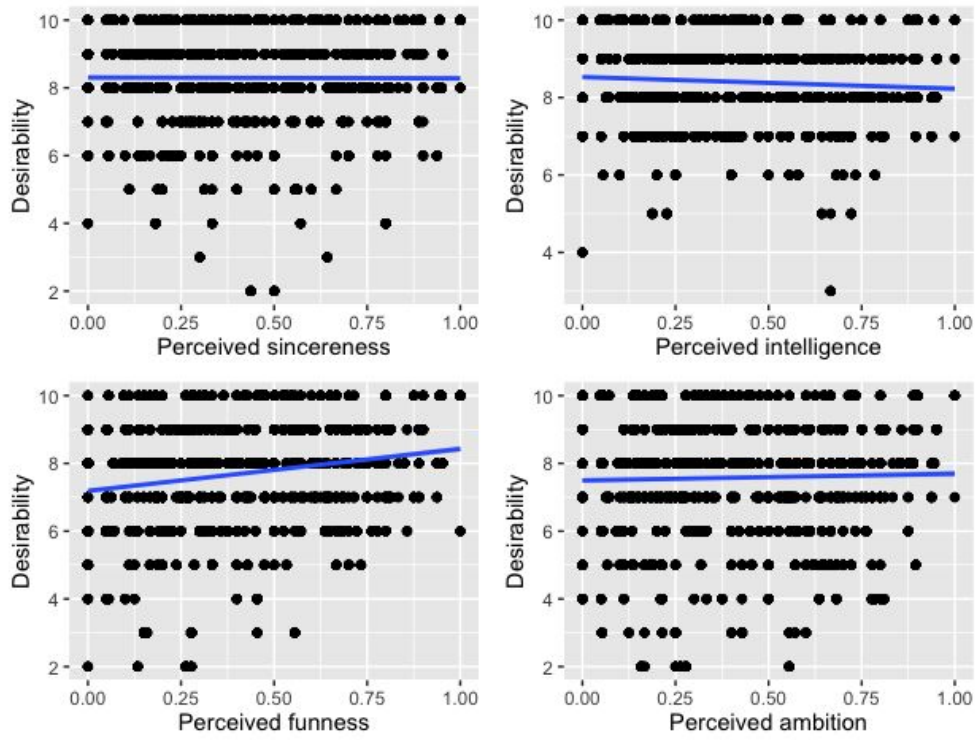


Furthering this question, what qualities that people believe they are the strongest in reflect how desirable other people find them?

The participants were asked to rate themselves, on a scale of 1-10, based on what they thought of in terms of the strength of the following qualities: sincerity, intelligence, funness, and ambition. We were curious as to whether a certain character that people thought they possessed would influence their desirability. It turns out that there doesn't seem to be much of a correlation, although intelligence showed a slight negative correlation (those that thought they were more intelligent were less desirable) and funness showed a positive correlation (those that thought of themselves as more fun were more desirable).

```
sincp <- ggplot(desirability_merge, aes(x=desirability_merge$des,
y=desirability_merge$sinc3_1)) +
  geom_point() +
  xlab("Perceived sincereness") + ylab("Desirability") +
  geom_smooth(method=lm)

intp <- ggplot(desirability_merge, aes(x=desirability_merge$des,
y=desirability_merge$intel3_1)) +
  geom_point() +
  xlab("Perceived intelligence") + ylab("Desirability") +
  geom_smooth(method=lm)
```

```
funp <- ggplot(desirability_merge, aes(x=desirability_merge$des,
y=desirability_merge$fun3_1)) +
  geom_point() +
  xlab("Perceived funness") + ylab("Desirability") +
  geom_smooth(method=lm)

ambp <- ggplot(desirability_merge, aes(x=desirability_merge$des,
y=desirability_merge$amb3_1)) +
  geom_point() +
  xlab("Perceived ambition") + ylab("Desirability") +
  geom_smooth(method=lm)

grid.arrange(sincp, intp, funp, ambp, ncol=2)
```



Conclusion

Despite all the regression lines that we plotted, it turns out that most of them were pretty flat and therefore there doesn't seem to be huge factors studied by the experiment that determine a person's desirability. Things that were clear was that gender and profession had large differences in attractiveness, but everything else (race, hobbies, your own perception of yourself and what you value) had weak effects on attractiveness.

Further analysis can be done on the differences in preferences by gender and race as opposed to the entire group of participants as a whole.