DSC 465 Final Report
Group: Dating Data Wizards
Members: Jessica Izquierdo, Areli Rios-Hanhan, Zehua Li, Yan Yu, Tianyi Tan

I. Introduction

Speed dating can be an exciting adventure. People who participant such an event often wonder: will the person sitting across from me be a good fit? How similar are my date and I? Can she or he be the one? This project attempts to understand human behavior and reveal speed dating insights.

The dataset used for this project was compiled by Columbia Business School professors through a series of speed dating sessions[1]. The dataset contains 8,387 observations of 195 variables. In the event, participants were asked to have a speed date (four minutes per session) with multiple participants and use one minute to answer questions about their experience with each date. The questionnaire includes demographics, self-rating and peer review of six attributes (i.e. attractiveness, sincerity, intelligence, fun, ambition, and shared interests), field of study and intended career, hometown, interests and hobbies, and other miscellaneous questions.

There are three types of variables in the dataset: categorical variables (e.g. race, gender, and career), numerical variables (e.g. rating and income), and ordinal variables (e.g. position: station number of the partner). With data visualizations, we attempt to answer the following questions: whether racial preference exists in a match? Which cities have the highest success rate of the matches? How did participants rate themselves across different attributes, and how did their dates rated participants? Are there any discrepancies? What attributes contribute to the success of dating?

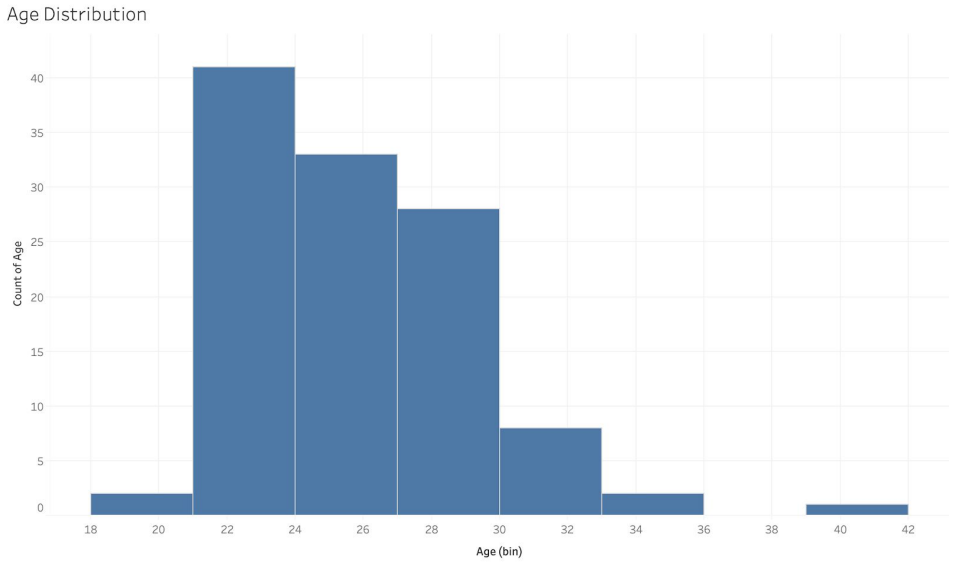
II. Exploratory Analysis

Our initial exploratory analysis focused on two avenues: (1) visualization of individuals involved in the speed dating (e.g. distribution of demographic information such as age, field/career, race, income, goal of participating in event, interests; differences between the individual perceives himself/herself vs. perceived by others); and (2) visualizations of two gender groups. Initial exploration involved cleaning data to remove duplicates and edit field names when needed. For example, variable "gender" was shown as a 0 or 1; Tableau interpreted this variable as a measure. We converted it to be a discrete dimension and change the alias names to corresponding category ("Female" or "Male").

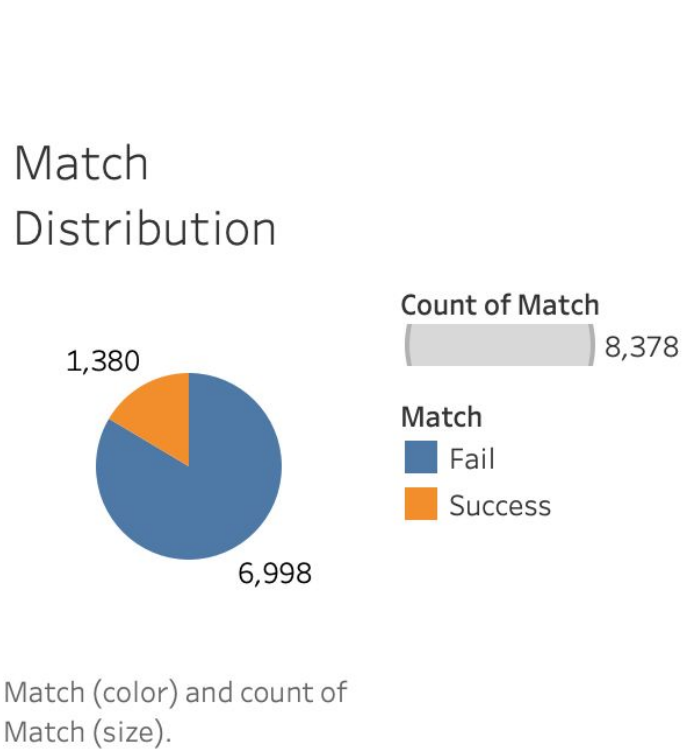
For the first avenue, Tableau was frequently used to generate graphs. For example, we used histogram to plot distribution of ages for all participants. We learned that the majority were between 21 to 30 years-old, and that participation skewed toward a younger crowd. Another graph was created to visualize match distribution, which illustrates that the number of failed dates is slightly more than five times the number of successful dates.

---

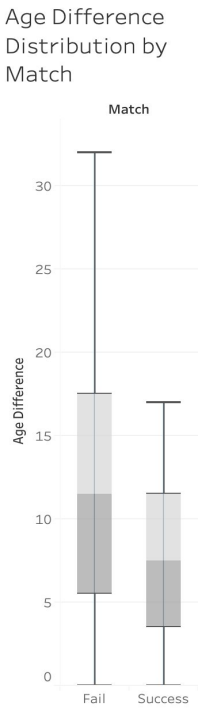
[1] Kaggle Data set: https://www.kaggle.com/annavictoria/speed-dating-experiment



The trend of count of Age for Age (bin).



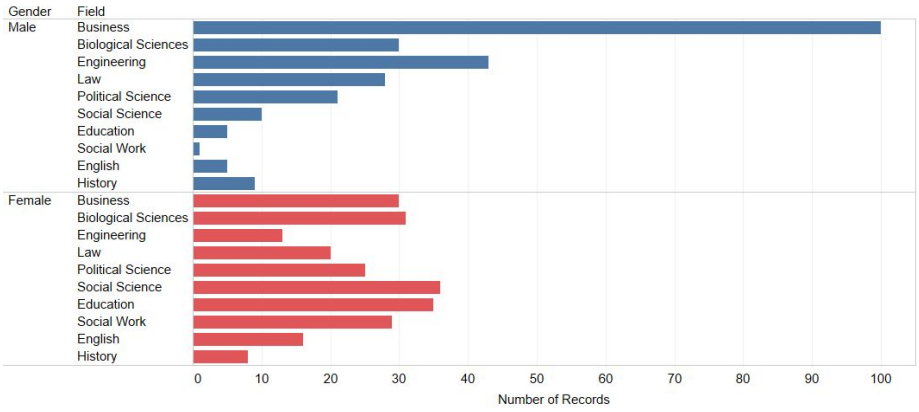
Match (color) and count of Match (size).



Age Difference for each Match.

For the second avenue, we attempted to visualize all the attributes of the dataset to understand the characteristics of both gender groups. For example, the first bar graph illustrates the top 10 fields of work for both genders. Fields such as business, engineering, and biological sciences are popular among male group, while fields such as social science, education, and biological sciences are popular among female group. The latter two bar graphs set up a good foundation for explanatory analysis.

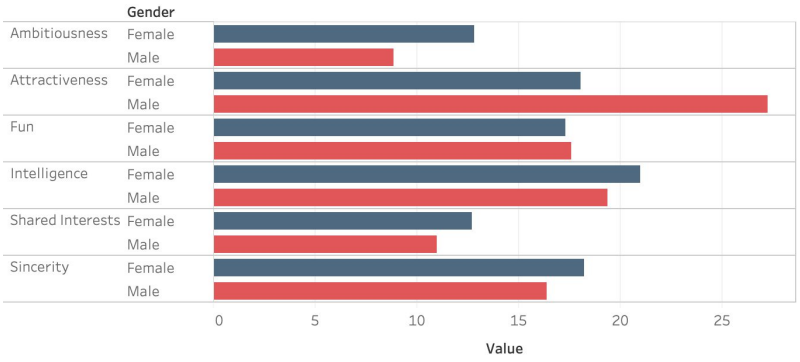
Top 10 Fields of Work - Male and Female



Count of Field for each Field broken down by Gender. Color shows details about Gender. The view is filtered on Field, which keeps 10 of 19 members.

Gender
Male
Female

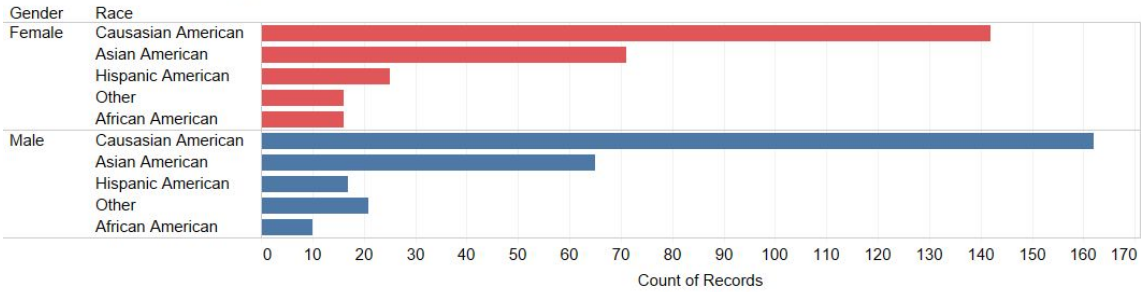
What Qualities Are Most Important to Each Gender? (Avg score)



Ambitiousness, Attractiveness, Fun, Intelligence, Shared Interests and Sincerity for each Gender. Color shows details about Gender. The data is filtered on Gender, which ranges from Female to Male.

Gender
Female
Male

Participants Race by Gender



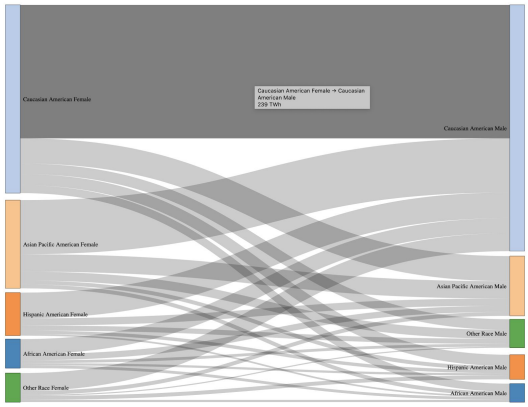
Count of Race for each Race broken down by Gender. Color shows details about Gender.

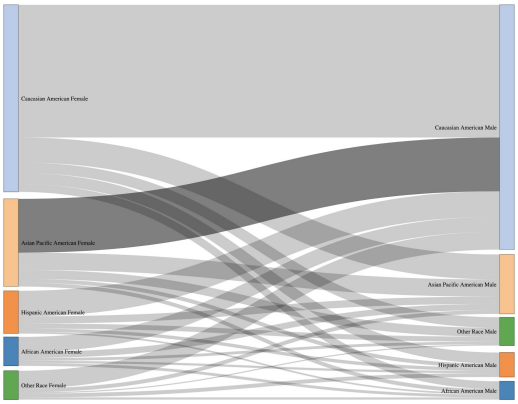
Gender
Female
Male

III. Explanatory Analysis

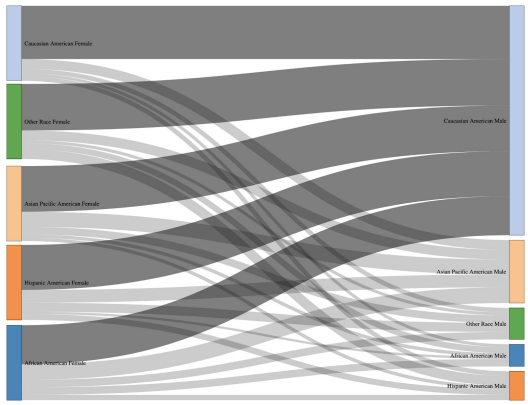
Visualization 1 - Sankey Diagram

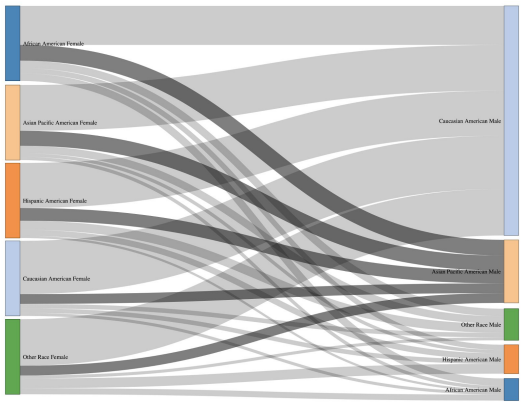
Sankey diagram 1 (which uses absolute numbers of matches)



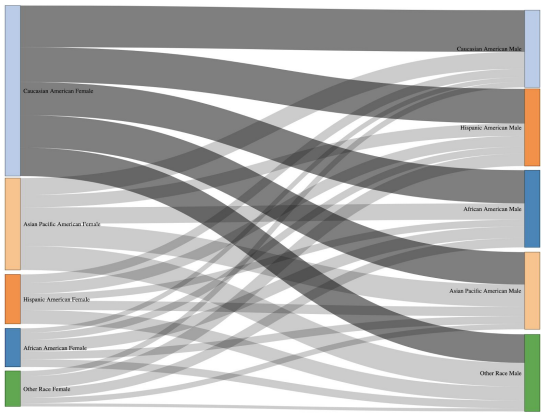


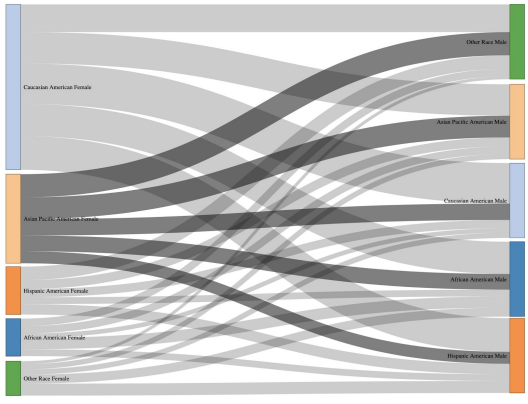
Sankey diagram 2 (which normalizes each racial group on the female side to be 100 people)





Sankey diagram 3 (which normalizes each racial group on the male side to be 100 people)



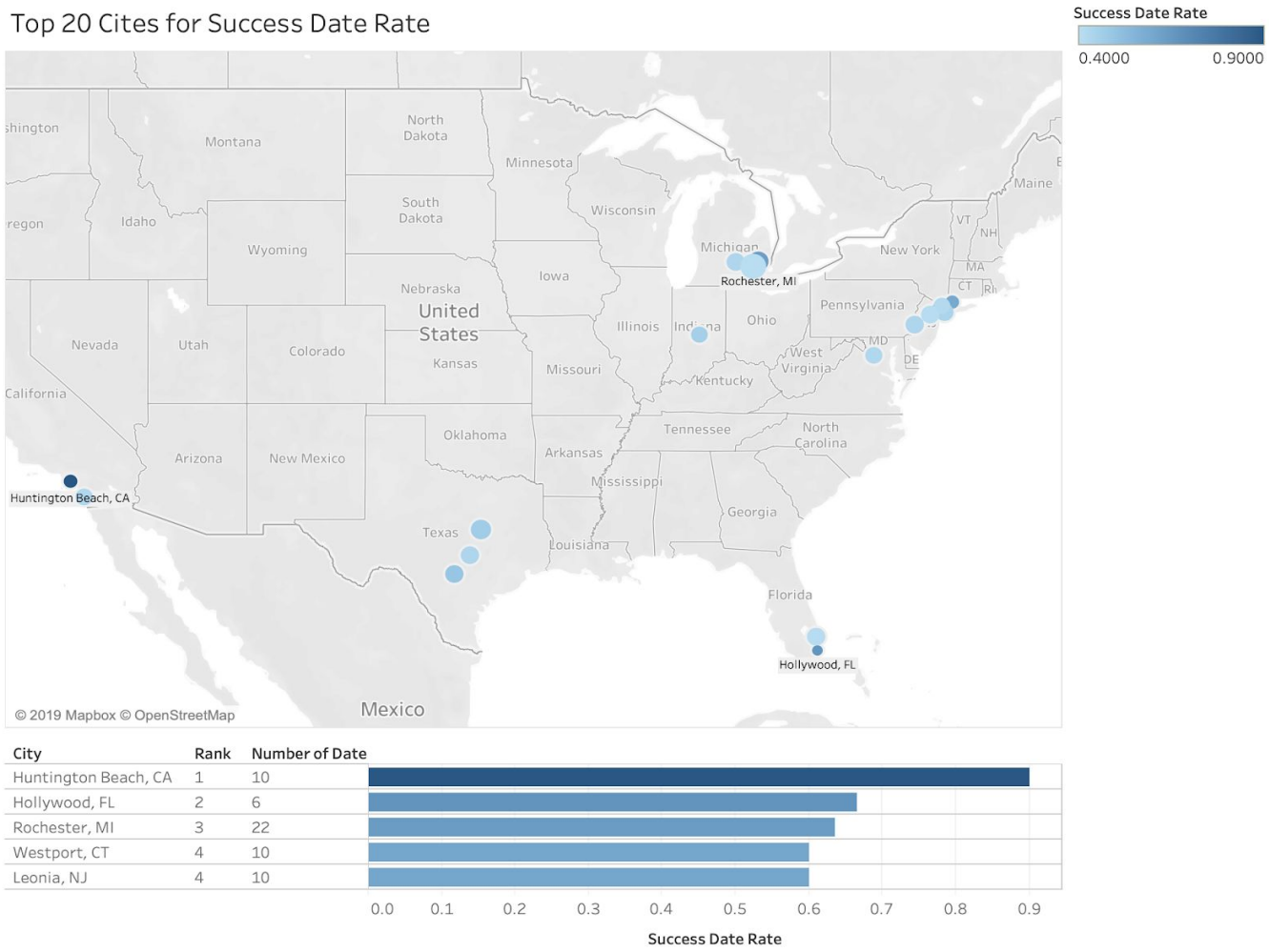


Among 8,378 observations of the dataset regarding speed dating conducted at Columbia University, 8,242 observations have complete racial information about participants and their counterparts. The dataset duplicates observations. For example, it would record an observation of Participant A dated Partner B and another observation of B (as a participant) dated A (as B's counterpart). Thus, 4,121 speed dates (i.e. 4,121 females with 4,121 males) were conducted, with a total of 676 turned out to be successful matches. Created with networkD3 package, interactive Sankey diagram attempts to visualize racial pairing of 676 successful matches.

The first interactive Sankey diagram uses absolute number of 676 successful matches. It places female counterparts of successful matches on the left-hand side and their male counterparts on the right-hand side. On each side, there are five tabs representing five different racial groups. A pathway shown in grey connects a tab on the left-hand side to a tab on the right-hand side. After laying the cursor on a path, a message box will pop up and show information about the pathway (e.g. Caucasian American Female —> Caucasian American Male 239, meaning 239 successful matches were formed between 239 pairs of Caucasian American females and Caucasian American males). The tabs can be moved upward or downward to rearrange the order. After arranging the order in terms of total number of people in each racial group for each gender, the diagram shows that the greatest number of matches is between Causation American males and Caucasian American females (239), and the second greatest number of matches is between Asian Pacific American females and Caucasian American males (97). This is typical in that both races are the largest populations of Ivy League schools such as Columbia University.

The second interactive Sankey diagram normalizes each racial group on the female to be 100 people. After arranging the order of the tabs, the diagram shows that over 50% of female participants, whichever race they are, formed successful matches with Caucasian American males: Caucasian American females (71%), Other Race female (62%), Asian Pacific American female (61%), Hispanic American female (60%), and African American female (52%). Successful matches with Asian Pacific males ranks the second: African American female (21%), Asian Pacific American female (20%), Hispanic American female (17%), Caucasian American females (13%), and Other Race female (13%). The third interactive Sankey diagram normalizes each racial group on the male to be 100 people. After arranging the order of the tabs, we can see that over one third of male participants, whichever race they are, formed successful matches with Caucasian American females: Caucasian American male (54%), Hispanic American male (45%), African American male (43%), Asian Pacific American male (42%), and Other Race male (37%). Successful matches with Asian Pacific females ranks the second: Other Race male (31%), Asian Pacific American male (29%), Caucasian American male (22%), African American male (21%), and Hispanic American male (16%). Based on Sankey Diagrams 2 and 3, it is obvious that both genders of Caucasian participants are popular racial groups in the speed dating event. Female participants from each racial group show more concentrated preferences than male participants from corresponding racial group.

Visualization 2 - Geographical Visualization

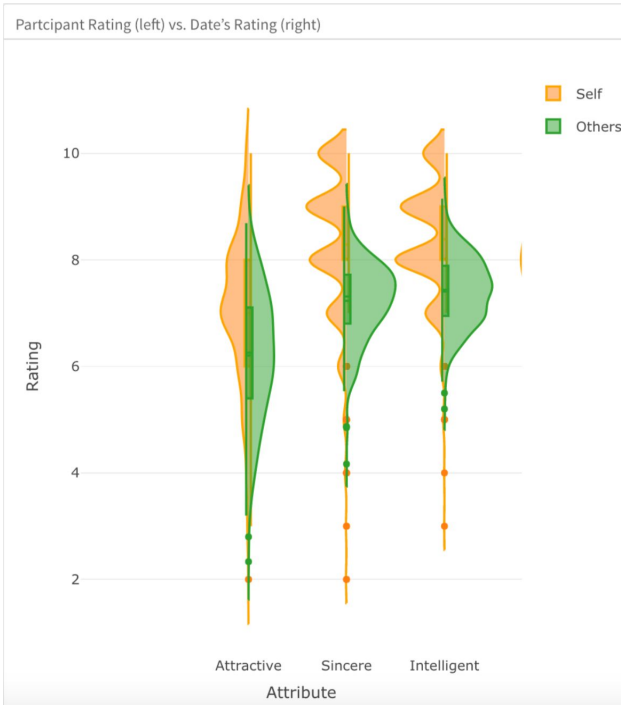


The dataset includes hometown for each participant. It would be interesting to unfold what cities have the highest success data rate. In this graph, we combined two visualization techniques: geographic plot and bar chart. Success date rate for each city was calculated by grouping data by cities and dividing the number of successful matches by the total round of matches. Success data rate for each city was calculated and ranked from the highest rate to the lowest. In particular, participants from Huntington, CA have the highest success date rate (90%). In other words, they typically scored 9 out of 10 dates. Top ranking ones are plotted as glyphs in the U.S. geographic map. The size represents the number of dates; the greater the number of dates, the bigger the size. The color shown in saturation format represents the success date rate. The higher the rate, the darker the color. The bar chart shows the top five findings with brief description of the city and its exact ranking, number of dates, and success rate. In addition to Huntington, CA, the other four cities with top success data rate are: Hollywood, FL; Rochester, MI; Westport, CT; and Leonia, NJ. After examining participants by city, we found some participants had high ratings in certain attributes such as attractiveness, while other participants have average ratings. It would be worthwhile for future research to explore factors contributing to high success rate.

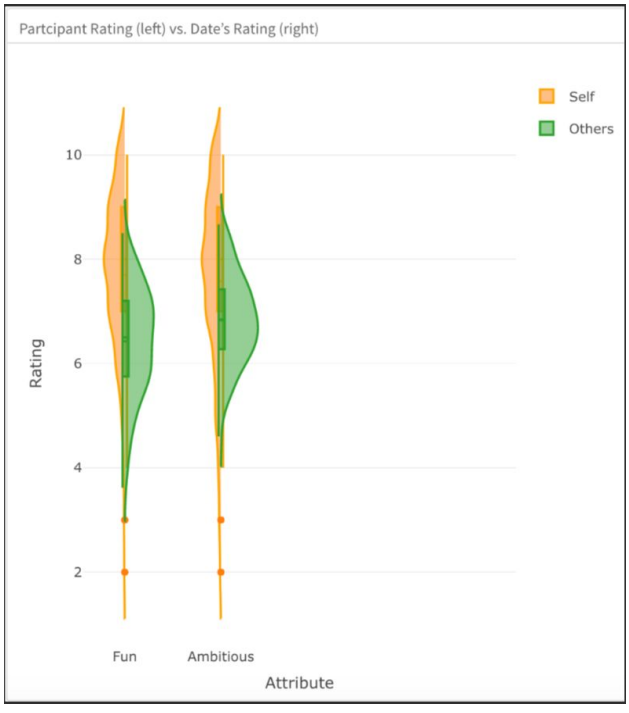
Visualization 3 - FlexDashboard

Link to dashboard: http://rpubs.com/izquierdo_jess/553064

First three attributes of violin plot:



Last two attributes of violin plot:



As part of the speed dating experiment, participants were asked to rate themselves on five different attributes on a scale of 1-10 : Attraction, Ambition, Intelligence, Fun, and Sincerity. Their dates were then asked to rate them on those same attributes. The visualization above attempts to capture some of the differences in the way participants rated themselves versus how their date rated them. It helps highlight the idea that in general people tend to overestimate their attributes i.e. the participants gave themselves higher scores for each of the attributes compared to how their dates rated them on those same attributes.

Because we had two visualizations that explored similar aspects of the data, we decided to merge them into a FlexDashboard. The first chart in the dashboard is a violin plot which helps analyze the distribution differences of the participant (orange / left side of the plot) versus the date's score of the participant (green / right side of the plot). The side-by-side comparison let's us see the differences in the shape of the distributions for both. The plot is produced by using R to pre-process the data and use the R package Plotly to plot the with different hues to help distinguish between the two groups. We understand that people with color blindness might have a hard time distinguishing the two colors, but were limited to color choices when using Plotly.

The second chart in the dashboard is a radar chart that compares the average rating across each of the attributes for both parties. The orange shading (the shading that extends out the most) represents the participant's rating of themselves while the green shading (the smaller one) represents the date's rating of the participant. This chart allows us to see that for all attributes, the participant rated themselves higher (the orange shading extends out more than the green). It also helps illustrate that the largest gap in rating exists in the Fun attribute. The hexagon, created in R, includes an axis going through the middle of it that represents the average score. Additionally, we opted to make the color transparent to allow the axis to still be seen through the shading.

Last, we decided to include the average gap between the participant's rating and the date's rating. For these we opted to use Value Boxes on the side with each attribute having its own box. We used white text on a dark gray background to make the text stand out.

We decided to keep the color of the dashboard simple and used gray so that only the colors used in the two charts stand out. Additionally, we made sure that each of the charts uses the same color for each category and that axis were labeled whenever possible. We increased the font on the charts so that labels were easy to see and made the font in the value boxes white for this same reason.
Both charts combined help the audience conclude that people do in fact overestimate their own attributes. While thinking highly of ourselves is not necessarily a flaw, this visualization helps show that often other people's perception of us isn't always the same.

Visualization 4 - Parallel Coordinate Plot

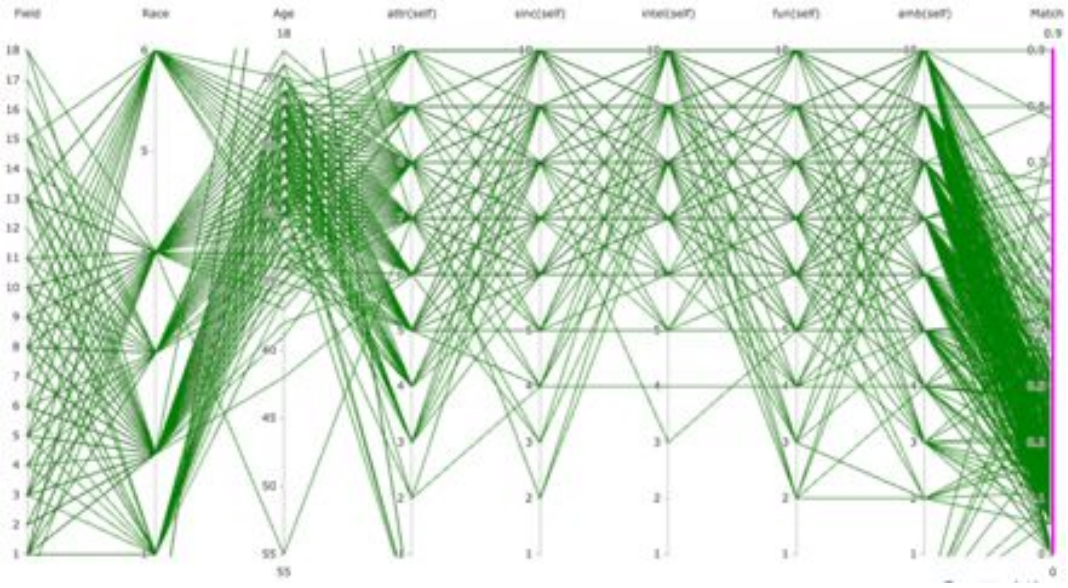
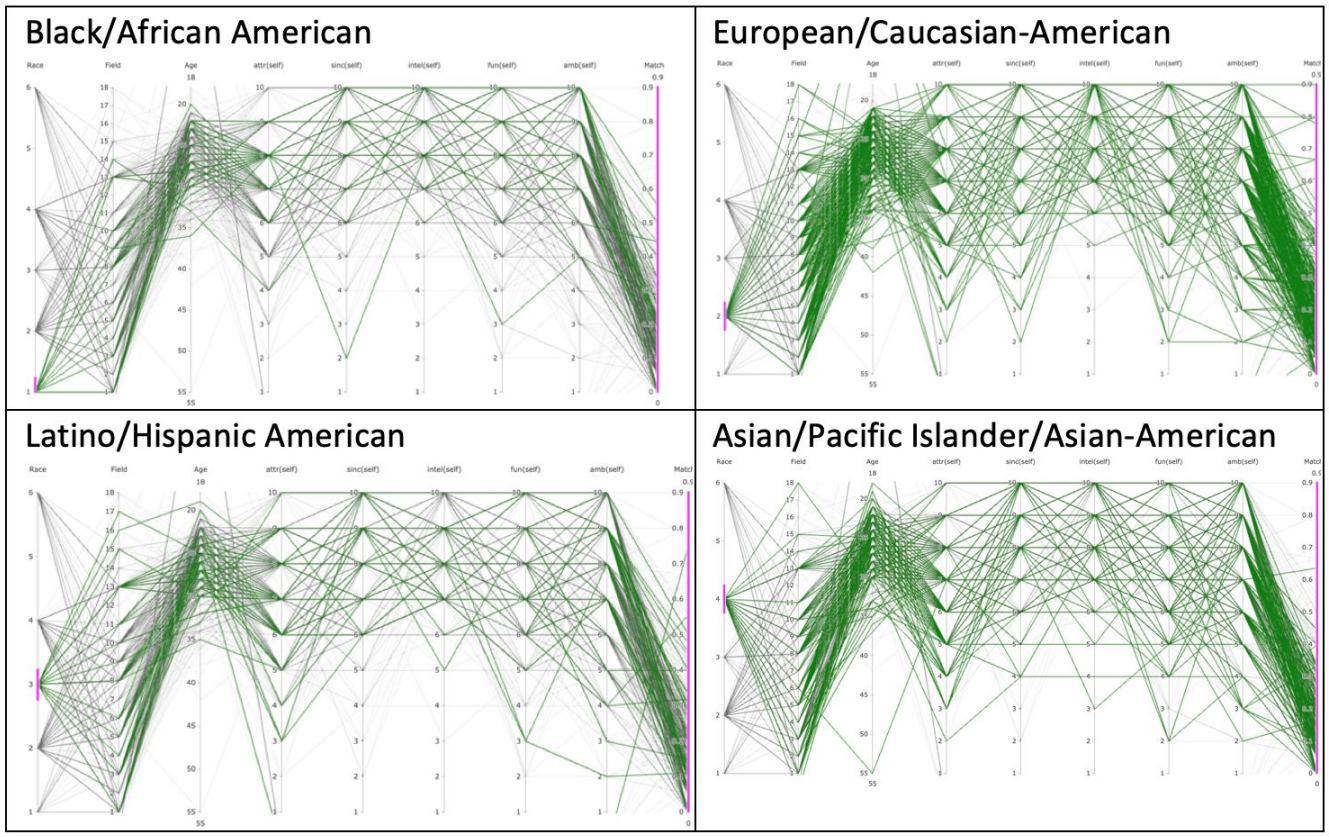


Figure 1:



The plot visualizes the pattern of high-dimensional data. Nine features are included: demographics of participants (categorical variables): field of study, race, age, self-perception (ordinal variables) of their attractiveness, sincere, intelligence, fun, ambition and calculated success rate (continuous variables) (matching percentage = match count/total rounds). The matching percentage (success rate) is an important variable to draw a conclusion about whether each feature contributes to high or low success rate.

We have also applied parallel coordinate plot for nine features. It is usually used for high-dimensional data to show the patterns of complex relationships. The parallel coordinate enables more features to be included in a clearer way and also plot is interactive produced by R code. By pulling the feature around vertically and selecting different values for each different feature, it is a lot of fun and also help identify patterns even with many instances and features involved.

The color palette chosen for the plot were originally sequential colors with different hues and saturation to indicate the percentage of matching. However, the best sequential color in Plotly is still making the graph more complicated and the messages a bit unclear for audience. Following the advice from Professor, we decide to change the color to pure green which is most sensitive to human cones and serves the purpose of highlighting the lines selected for comparison.

The plot is produced by using R to pre-process the data and use R package Plotly to plot which is interactive. Codes are attached in the appendix. It is important to use the interactive plot to show the pattern by pulling different columns and selecting different options for each column. By interactively playing with the plot, it can help us answer some interesting questions. It can also be related to the split violin plot that was created and complement each other. This plot takes a deep step to analyze at the individual level and take a broader view with more features and observe how they interact with each other.

This plot shows interesting patterns of demographics, self-perceptions and dating success by answering the following three questions:

1. Are people of different races perceiving themselves differently in general? (Figure 1)

Four of the six options of races are analyzed which have enough instances to show patterns and specified ethnicity. By pulling ‘race’ to the first vertical line and select different values of this feature, the four graphs show the patterns of 4 different races. In terms of the general trend of different attributes, there are similarities among different races. For example, people of different races have the most diversity in self-perceived ambitions while most of the people of different races rate themselves above 5 for intelligence. We might conclude that people of different races might have similar perception of themselves.

2. Are people from certain field of study more successful in dating?

By pulling field to the first vertical line and selecting matching percent over 50%, the graph shows that participants of field of study from 1 to 13 all have high matching percent. The participant with 90% matching rate studies (field 8) business/econ/finance. Participants with over 80% matching rates also

study (field 1,3,4,8,13) Law, Social Science, Medical Science, Business and Political Science/International Affairs. People who are good at communication or studying science might win themselves a date. By selecting the matching rate ranging from 0 to 10%, we can see that no particular field of study might result in low matching rate. Thus, never blame your major!

3. Are people who have a high self-perception for different attributes more successful in dating?

Using the same graph with high matching rate selected (50% to 90%), we can observe that participants who have matching rate over 50% tends to rate themselves high in all 5 attributes. Thus, confident people or people with high value in the five attributes tend to have successful matching in speed dating. Self-perceived intelligence ratings are at least 7 shown in the graph which means people think themselves or are really smart tend to secure a date while pattern is not as clear for ambition.

IV. Conclusion

Different visualizations help us understand various aspects of the speed dating dataset. Aside from general information about the participants, it enabled us to dive deeper into racial pairing of successful matches, examine success rate by geographical locations, and compare self-rating with peer view. One interesting conclusion is that Caucasian participants are popular racial groups in the speed dating event; racial preferences are more obvious in female group. Another interesting conclusion is that in general, participants rated themselves higher than what their dates rated them. In addition, we found that people of different races tend to have similar perception of themselves. While we have not found a magic formula to help future participants succeed in the dating world or find soul mates, we have learned more about human behavior and gained speed dating insights through data visualizations.

Appendix A - Individual Reports

Jessica Izquierdo

If there is one thing I learned about data visualization from both the project and from class, it is that Audience is key and a major driver around the decisions that surround data visualization. Working with my team helped me to think about aspects of data that I might have not considered on my own. Everyone brought a unique perspective and skill, and we learned about new types of visualizations from one another that I can see myself utilizing in the future.

Working on the Speed Dating dataset also helped sharpen skills having to do with tuning out noise in data. For example, our dataset consisted of almost 200 columns. This gave us a lot to work with, but it also meant that there would be a lot that would not be relevant towards the message we were trying to convey through our visualizations. It forced us to try to answer a question with the data and focus on the fields that would help us answer those questions.

One other thing the course and the project taught me is how easy it is to (purposely or accidentally) mislead the audience with the visualizations used. It helped to highlight the importance of the little stuff such as legends, axis titles, and even colors. It helped me with understanding why Chart Junk isn't always a good option when I think it might "enhance" the visualization, as well as helped me be mindful about all the decisions that went into why my visualizations looked the way they did.

My contribution to the data visualizations used centered around the FlexDashboard. The violin chart was created by a fellow team member, and the remainder of the visuals and creation of the dashboard was my own. Both Tianti and my visualizations focused around looking at how the subjects rated their own attributes vs how their dates rated them, so it made sense to try to integrate the two in a dashboard format (and would allow us to incorporate other things such as descriptions), especially when we were trying to narrow down the top 4 visualizations we would use in the presentation and the final paper.

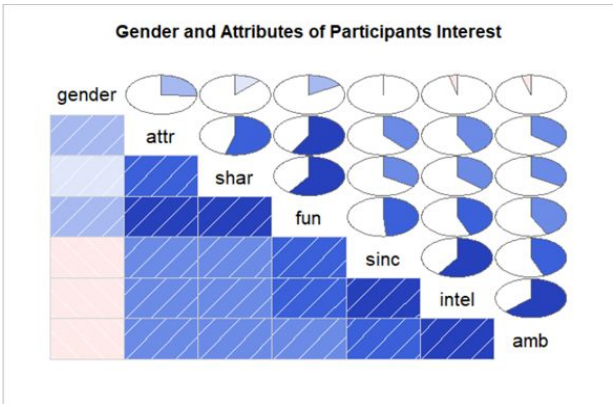
I also contributed to each of the project deliverables, specifically in PD2, I contributed to five of the initial visualizations: Top 10 Fields of Work for Women, Top 10 Fields of Work for Men, What Qualities are Most Important to Each Gender, Age Distribution by Gender, and Income by Gender. I was also responsible for creating the Slack page for our project so all of our work was in one place and to help us maintain organized as much as possible. By having several channels for each of the phases of the project (dataset selection, project deliverables, presentation, final paper, etc), it made it easy to go back and look for past work that we did. In the presentation, I helped with the slide regarding the FlexDashboard as well as the one that gave a general overview of the data we were using.

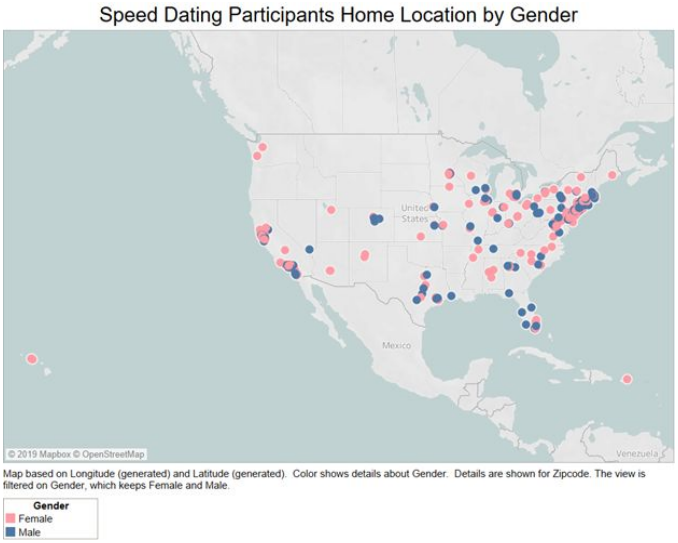
Last, I contributed to the final paper by providing an explanation for the FlexDashboard visualization piece, as well as helping put incorporating one of team member's work and doing general review of the paper.

Areli Rios-Hanhan

In this project, our team the dating data wizard used data about a speed dating experiment. My focus in the project was to view the attributes around gender, how do men and women experience the speed dating event. We did a lot of data cleaning; removing attributes that didn't make sense in the data, as well as cleaning up duplicate data that would be unrelated for some of the graphs. The data for this project turned out to be very interesting and challenging to understand at times.

My contribution to the group project was a graph in RStudio called "Correlogram", this type was used to see a visual representation of the correlation between gender and the different type of attributes the speed dating participants listed as "interest" for a match. The attributes are (Ambition, Intelligence, Sincerity, Fun, Shared Interest, and Attraction). The listed attributes help is showing based on gender there is a correlation between each attribute, however, some have a higher correlation than other. Attraction is an attribute with gender that shows a higher correlation. This leads one to believe, both genders are interested in having an attraction to the participant in the speed dating event. This graph is also displaying the correlation between each attribute which in a way can be helpful to see the differences between each one. However, this graph was not used for the presentation or final project. I also created a graph of a map in Tableau. The visualization was put together by using the longitude and latitude measured from the zip code the participants provided. The zip code measure was used to provide a detailed map in the visualization. The gender dimension was used to provide a color display of where Male and Female participants were originally living. The reason for choosing this type of visualization was because I wanted to see the common locations where the participants originated from and if those participants were male or female. This graph was not used for the presentation or final report.





In the group project, I also helped contribute to the P2 assignment. I helped by creating a visualization on matching by rounds filtered by gender. Since I was part of the I avenue focused on gender, I create various graphs in tableau to begin to get an idea of how the event was viewed by the different genders. Many of the graphs were simple bar charts with averages on a lot of the surveys taken by the participants. I also contributed on P3 assignment, I put together the format of the presentation that would be used for this assignment. Although, my visualization was not used for the presentation, I helped write out one of the slides with a visualization. I stayed engaged with the group to try to help

with anything that was needed to complete the assignments. Lastly, I helped with the final project report to write the initial parts of telling a story and help the team write the analysis.

In conclusion, this project and class was a learning curve for me. I had very little experience with tableau and no experience with RStudio. I found it a bit challenging to try to come up with some of the visualization but did my best to research how to use both tools. I can honestly say the class pushed me to think out of the box to create some of visualizations and the many techniques that can be used. It also gave me ideas of how to leverage other tools to display the data. The data we picked for this project helped give an interesting perspective to generate visualizations. It was fun to use data about speed dating, it helped write a story about it and make it interesting for everyone to view.

Zehua (KC) Li

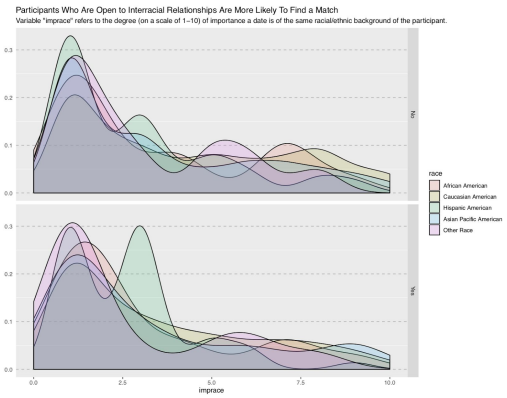
During the group project, I tried my best to be a supportive teammate. Knowing that our group was interested in topics related to social media, hotel rating, and dating matters, I researched multiple datasets for the group's selection and examined variables to see possible visualizations that could be created from each dataset. After the group decided which dataset to use, I conducted a comprehensive research on published articles and blogs regarding the selected dataset and composed a list of summarized ideas and links to visualizations. My proposed two avenues of data investigation (i.e. visualization of individual participants and visualization of two gender groups) was accepted by group members. I volunteered to be the group liaison and checked in with the team at least a couple of times a week. If there is any question that could not be solved internally, I would email Professor Brown for advice. In addition to being an active member in group discussion and project deliverables, I attempted to explore data visualization options, such as suggesting combination charts and researching for appropriate visualization techniques not specifically covered in class.

Visualization is an effective tool to understand a dataset. Besides looking at popular variables such as how speed dating participants rated themselves and how their partners perceived the participants, I also paid attention to other demographic information (e.g. participants' primary goal in participating the speed dating event and feedback on four minutes per date illustrated in PD2). Finally, I found racial aspects of speed dating to be worth exploring further. Four variables were associated with racial aspects: "race" (i.e. race of participant), "race_o" (race of partner), "imprace" (levels of importance (on a scale of 1-10) to a participant that a date should be of the same racial background as the participant), and "samerace" (i.e. whether participant and the partner were the same race). Several techniques were applied during the exploration stage. For example, the density plot reveals that participants who found a match generally were more open-minded and did not consider dating partners of the same racial background to be very important.

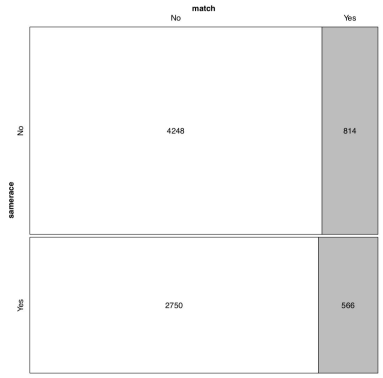
Mosaic plot was also used since it integrally exhibits relationship between multiple categorical variables. Mosaic Plot 1 illustrates that the proportion of grey area in same-race-dating category is slightly larger than that in not-the-same-race-dating category, indicating all other things equal, the factor of participant being the same race as the partner contributes to the match rate ("Yes" to variable "match" divided by the total rounds of speed dating). Mosaic Plot 2 shows that for both genders, Caucasian Americans are the largest participants, followed by Asian Pacific Americans. The racial pairing of 676 successful matches interested me the most. Initially, I chose interactive chord diagram, which could be confusing since it arranged different racial groups for each gender radically around the circle and had no clear division to separate two genders. Through research and contemplation, I found interactive Sankey diagram appropriate for visualization. Originally used to visualize energy flows, Sankey diagram with two gender groups residing on each side of the diagram is a cleaner and more innovative way to display racial pairing. After taking Professor Brown's suggestion regarding normalization, racial preference in each gender group becomes more obvious.

Data visualization is a powerful and valuable tool. It takes an active thinking process to explore and understand insights, progress from simple graphs to more complex ones, and find appropriate visualization techniques to deliver an intriguing story to targeted audience. I am glad that I chose to have this course during my second quarter of the program, learn the wisdom of data visualization, and have a pleasant team working experience.

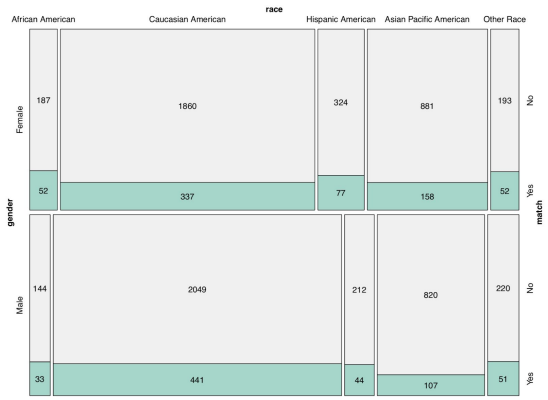
Density Plot



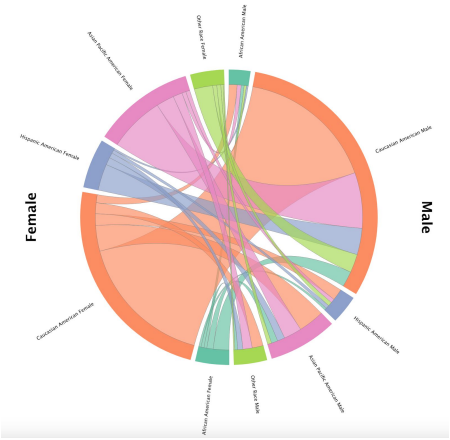
Mosaic Plot 1



Mosaic Plot 2



Chord Diagram



References:

Berhane, Fisseha. *Interactive Chord Diagrams in R/Shiny*. Retrieved from: https://datascience-enthusiast.com/R/Interactive_chord_diagrams_R.html.

Kuzmin, Alex, et al. *Alternative platforms - D3.js in R*. Retrieved from: https://bbolker.github.io/stat744/platform/part1.html.

Fisman, Raymond J., et al. *Racial Preferences in Dating: Evidence from a Speed Dating Experiment*. Retrieved from: https://www0.gsb.columbia.edu/mygsb/faculty/research/pubfiles/867/datingFULL-EK1.pdf.

Yan Yu

For this final project, I participated in the discussion of data selection, exploratory data analysis, data visualization generation and final presentation. My contribution is finding the error when we exploring the data. In our speed dating experiment dataset, each observation is a date not an individual. An individual could have multiple date. This is very important during the exploratory data analysis stage. For example, when we analyze the fields for individuals, we need to remove the replicates so we can have the actual number, otherwise the number could be hugh. I made age distribution by histogram math distribution by pie chart and age different by boxplot in the exploratory analysis by Also I made the visualization 2, the geographic plot. There are several things I did for making the geographic plot. The dataset only has the zip code, so I transfer the zip code to the actual city first and then generate the longitude and latitude for each city. At the end, I input the data into Tableau.

It is very helpful for me to cooperate with my teammates in this project. Because we would be able to discuss topics together in different aspects. Also we can be each other's audience to inspect the works. Presentation is also a very important part. Even though a picture is more than a thousand of words, we still need to explain the purpose of the plot and the variables. My teammate, Jocelyn did a fantastic job on this. She not only explained the plot in detail but also make it understandable and attractive. I need to learn from her skill.

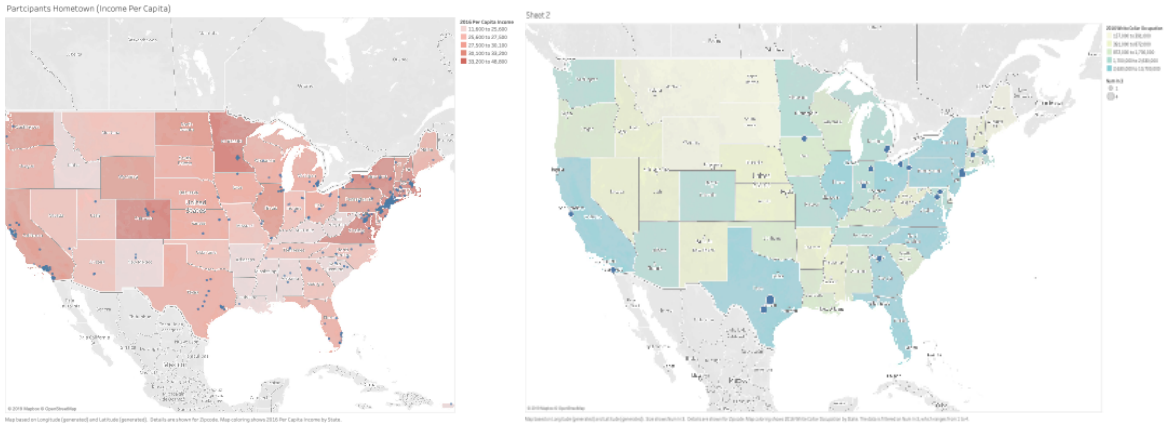
In this course, I have learned different data visualization techniques such as violin plot and network plot. Also when to use them and their functions. Besides the techniques, I know the common mistakes that I used to make. It is very important to know your audience and deliver your information effectively. One of the mistakes I always have is I am likely to make the plot complex but hard to process for the audience. In this project, we try to tell a story based on those data visualization methods. We adjusted a lot since the first draft, because we need to make the plots more accurate and concise. Knowing how to use data visualization methods properly is very important in both the data exploration and data exploration. Visualization is one of the most efficient ways to deliver information and make conclusions. I am glad that I took this class and implemented those techniques in my project. I believe it will be beneficial in my further data science study and work.

Tianyi Tan

In our Dating Data Wizards team, we worked as a team to explore different aspect of the speed-dating dataset. Specifically, I worked with Yan Yu to analyze the data of individual participants in the dataset. I also took the responsibility of in-class presentation with all the visualizations created by the team. For the final report, I am responsible for detailed description of visualization I have created. In our team, we help each other with the improvement of each visualization.

I have created two visualizations that are included in the final report: split violin plot and the interactive parallel coordinate plot. All the features used are pre-processed using R. Both charts are produced using R codes and Plotly package. I have also provided some exploratory analysis on geographic data in the dataset. Choropleth maps with glyphs on the map identifying the hometown of the participants and sequential colors used for income per capita for different states. We can see from the map that participants of the speed dating tend to grow up in the states that have high income per capita. Note: it only includes the participants in the United States.

I have also used choropleth graph to see the number of matches the participant have been on a date. The size of the points shows the number of matches the participant have been on a date. The color shows the white-collar occupation by states. We can observe that there are not so many of 'success' in terms of converting matches to dates in the speed dating in general. The successful participants tend to in the states that has more white-collar occupations.



In this project, I have learned different techniques to send clear messages to the audience. During the process of communicating the visualizations to the group and the whole class, I have learned how to convey the encoded information from the graph verbally and the importance of producing clear and uncluttered visualization. Through the feedback of the professor, I have learned how to improve the visualization from different perspectives such as changing color palette for the parallel coordinate plot. Together with this team of wonderful teammates, I have learned so much when we discuss different visualizations. I also have a better understanding of how great teamwork will turn the project into a fun and rewarding journey to learn!

Appendix B - Code of Exploratory and Explanatory Data Analysis

Visualization 1: Sankey Diagram

```
Speed.Dating.Data <- read.csv("~/Downloads/speed-dating-experiment/Speed Dating Data.csv")
df<-Speed.Dating.Data

# Mosaic Plot 1: match vs. samerace
df$match<-as.factor(df$match)
summary(df$match)
#0    1
#6998 1380
levels(df$match)<-c("No","Yes")
summary(df$match)
#No  Yes
#6998 1380

df$samerace<-as.factor(df$samerace)
summary(df$samerace)
#0    1
#5062 3316
levels(df$samerace)<-c("No","Yes")
summary(df$samerace)
#No  Yes
#5062 3316

library(vcd)
mosaic(match~samerace,data=df, labeling = labeling_values,highlighting_fill=c("white","#bdbdbd"))

# Density plot
library(ggridges)
library(ggplot2)
library(ggpubr)
df$race<-as.factor(df$race)
summary(df$race)
levels(df$race)<-c("African American","Caucasian American","Hispanic American","Asian Pacific American","Other Race")
summary(df$race)
df1<-subset(df,df$samerace=="No")
ggplot(df1,aes(x=imprace,fill=race))+geom_density(stat="density", alpha=I(0.2))+facet_grid(df1$match)+theme_cleveland()+
  labs(title="Participants Who Are Open to Interracial Relationships Are More Likely To Find a Match",
       subtitle="Variable \"imprace\" refers to the degree (on a scale of 1-10) of importance a date is of the same racial/ethnic background
of the participant.")

# Mosaic Plot 2: match vs. race + gender
df$race<-as.factor(df$race)
summary(df$race)
#1    2    3    4    6    NA's
#420 4727  664 1982  522  63
df$race_o<-as.factor(df$race_o)
summary(df$race_o)
#1    2    3    4    6    NA's
#420 4722  664 1978  521   73

df<-df[complete.cases(df$race,df$race_o),]
summary(df$match)
#No  Yes
#6890 1352
summary(df$samerace)
#No  Yes
#4926 3316

df$gender<-as.factor(df$gender)
summary(df$gender)
```

```
#0    1
#4121 4121
levels(df$gender)<-c("Female","Male")

summary(df$race)
#1    2    3    4    6
#416 4687  657 1966  516
levels(df$race)<-c("African American","Caucasian American","Hispanic American","Asian Pacific American","Other Race")
summary(df$race)
#African American     Caucasian American     Hispanic American Asian Pacific American     Other Race
#416                  4687                   657                1966                    516

Female<-subset(df,df$gender=="Female")
Male<-subset(df,df$gender=="Male")

summary(Female$race)
#African American     Caucasian American     Hispanic American Asian Pacific American     Other Race
#239                  2197                   401                1039                    245

summary(Female$race_o)
#1    2    3    4    6
#177 2490  256  927  271

summary(Male$race)
#African American     Caucasian American     Hispanic American Asian Pacific American     Other Race
#177                  2490                   256                927                     271

summary(Male$race_o)
#1    2    3    4    6
#239 2197  401 1039  245

library(vcd)
mosaic(match~gender+race,data=df, labeling = labeling_values,highlighting_fill=c("#f0f0f0","#99d8c9"))

# chord diagram
df<-Speed.Dating.Data
df$race<-as.factor(df$race)
summary(df$race)
#1    2    3    4    6    NA's
#420 4727  664 1982  522  63
df$race_o<-as.factor(df$race_o)
summary(df$race_o)
#1    2    3    4    6    NA's
#420 4722  664 1978  521   73
df<-df[complete.cases(df$race,df$race_o),]
summary(df$race)
#1    2    3    4    6
#416 4687  657 1966  516
summary(df$race_o)
#1    2    3    4    6
#416 4687  657 1966  516

df<-subset(df,df$match==1)
df$gender<-as.factor(df$gender)
summary(df$gender)
#0   1
#676 676

Female<-subset(df,df$gender==0)
Male<-subset(df,df$gender==1)
```

```
summary(Female$race)
#1   2   3   4   6
#52 337  77 158  52

summary(Female$race_o)
#1   2   3   4   6
#33 441  44 107  51

summary(Male$race)
#1   2   3   4   6
#33 441  44 107  51

summary(Male$race_o)
#1   2   3   4   6
#52 337  77 158  52

F1<-subset(Female,Female$race==1)
summary(F1$race_o)
#1  2  3  4  6
#5 27  4 11  5

F2<-subset(Female,Female$race==2)
summary(F2$race_o)
#1   2   3   4   6
#14 239  20  45  19

F3<-subset(Female,Female$race==3)
summary(F3$race_o)
#1  2  3  4  6
#3 46  6 13  9

F4<-subset(Female,Female$race==4)
summary(F4$race_o)
#1  2  3  4  6
#7 97  7 31 16

F6<-subset(Female,Female$race==6)
summary(F6$race_o)
#1  2  3  4  6
#4 32  7  7  2

library(chorddiag)
Match<-matrix(c(5,14,3,7,4,
             27,239,46,97,32,
             4,20,6,7,7,
             11,45,13,31,7,
             5,19,9,16,2),
           byrow=TRUE,
           nrow=5,ncol=5)

colnames(Match)<-c("African American Female","Caucasian American Female","Hispanic American Female",
                "Asian Pacific American Female","Other Race Female")
row.names(Match)<-c("African American Male","Caucasian American Male","Hispanic American Male",
                 "Asian Pacific American Male","Other Race Male")
library(chorddiag)
groupColors <- c("#66c2a5", "#fc8d62", "#8da0cb", "#e78ac3","#a6d854")
chorddiag(Match,type = "bipartite",margin=150,groupnameFontsize = 10, groupColors = groupColors,
       categoryNames = c("Male","Female"),groupnamePadding = 25,showTicks = FALSE)

#sankey diagram
Speed.Dating.Data <- read.csv("~/Downloads/speed-dating-experiment/Speed Dating Data.csv")
```

```
df<-Speed.Dating.Data

df$race<-as.factor(df$race)
summary(df$race)
#1    2    3    4    6    NA's
#420 4727  664 1982  522  63
df$race_o<-as.factor(df$race_o)
summary(df$race_o)
#1    2    3    4    6    NA's
#420 4722  664 1978  521   73

df<-df[complete.cases(df$race,df$race_o),]
summary(df$race)
#1    2    3    4    6
#416 4687  657 1966  516
summary(df$race_o)
#1    2    3    4    6
#416 4687  657 1966  516

df<-subset(df,df$match==1)
df$gender<-as.factor(df$gender)
summary(df$gender)
#0   1
#676 676

Female<-subset(df,df$gender==0)
Male<-subset(df,df$gender==1)

summary(Female$race)
#1   2   3   4   6
#52 337  77 158  52

summary(Female$race_o)
#1   2   3   4   6
#33 441  44 107  51

summary(Male$race)
#1   2   3   4   6
#33 441  44 107  51

summary(Male$race_o)
#1   2   3   4   6
#52 337  77 158  52

F1<-subset(Female,Female$race==1)
summary(F1$race_o)
#1  2  3  4  6
#5 27  4 11  5

F2<-subset(Female,Female$race==2)
summary(F2$race_o)
#1   2   3   4   6
#14 239  20  45  19

F3<-subset(Female,Female$race==3)
summary(F3$race_o)
#1  2  3  4  6
#3 46  6 13  9

F4<-subset(Female,Female$race==4)
summary(F4$race_o)
```

```
#1  2  3  4  6
#7 97  7 31 16

F6<-subset(Female,Female$race==6)
summary(F6$race_o)
#1  2  3  4  6
#4 32  7  7  2

library(networkD3)
df1<-as.data.frame(c("African American Female","Caucasian American Female","Hispanic American Female",
                     "Asian Pacific American Female","Other Race Female",
                     "African American Male","Caucasian American Male","Hispanic American Male",
                     "Asian Pacific American Male","Other Race Male"))
names(df1)<-"names"
source<-c(1,2,3,4,5,1,2,3,4,5,1,2,3,4,5,1,2,3,4,5)
target<-c(6,6,6,6,6,7,7,7,7,7,8,8,8,8,8,9,9,9,9,9,10,10,10,10,10)

#Use absolute numbers of people in pairings
count<-c(5,14,3,7,4,27,239,46,97,32,4,20,6,7,7,11,45,13,31,7,5,19,9,16,2)
df2<-data.frame(source,target,count)
df2$ncount<-df2$count+1
sankeyNetwork(Links = df2-1, Nodes = df1, Source = 'source',
              Target = 'target', Value = 'ncount', NodeID = 'names',
              units = 'TWh', fontSize = 12, nodeWidth = 30)

# Normalize each racial group on the female side to be 100 people
pct1<-c(10,4,3,4.5,8,52,71,60,61,62,7,6,8,4.5,13,21,13,17,20,13,10,6,12,10,4)
df2<-data.frame(source,target,pct1)
df2$ncount<-df2$pct1+1
sankeyNetwork(Links = df2-1, Nodes = df1, Source = 'source',
              Target = 'target', Value = 'ncount', NodeID = 'names',
              units = 'TWh', fontSize = 12, nodeWidth = 30)

# Normalize each racial group on the male side to be 100 people
pct2<-c(15,43,9,21,12,6,54,11,22,7,9,45,14,16,16,10,42,12,29,7,10,37,18,31,4)
df2<-data.frame(source,target,pct2)
df2$ncount<-df2$pct2+1
sankeyNetwork(Links = df2-1, Nodes = df1, Source = 'source',
              Target = 'target', Value = 'ncount', NodeID = 'names',
              units = 'TWh', fontSize = 12, nodeWidth = 30)
```

Visualization 2 - Geographical Visualization

This plot is made in Tableau. Geographic information, success date rate, and number of total date were imported into Tableau. We converted the zip code to actual cities and generated the longitude and latitude for each city. To calculate the success date rate for each city, we grouped the data by cities and divided the number of successful date by the number of total dates. Next, we filtered the top 20 cities to display on the map. For bar chart, the top 5 were displayed. And we indicate the rate by the color, from light to dark, and the number of date by the size from small to big. The color is also reflected in the bar chart too. Lastly, we used the dashboard to combine the two plots and use the bar chart for reference purpose.

Visualization 3: FlexDashboard

```
---
title: 'People Tend to Overestimate Their Attributes'
output:
  flexdashboard::flex_dashboard: null
  theme: journal
  pdf_document: default
  html_document:
  df_print: paged
  orientation: rows
```

```
 vertical_layout: fill
---
 Each individual in the speed dating experiment was asked to rate themselves on 5 different attributes. The radar graph shows the average particpant's perception of themselves vs. what their date actually rated them as. The split violin graph also helps capture the participant's rating of themselves vs. their date's rating of them.

<style>
.navbar {
  background-color:gray;
  border-color:white;
}
.navbar-brand {
color:white!important;
}
</style>

```{r setup, include=FALSE}
library(flexdashboard)
library(ggplot2)
library("dplyr")
library("ggplot2")
library("fmsb")
library("knitr")
library("rmarkdown")
library("latexpdf")
library("fmsb")
library("tibble")
library("stringr")
library("grid")
library("gridBase")
library("scales")
setwd("/Users/jessicaizquierdo/Documents")
dating = read.csv("dating.csv")
rawdat = read.csv("dating.csv")
```

Row {data-height=650}

Partcipant Rating (left) vs. Date's Rating (right)

```{r, fig.height=15, fig.width=20}


#r remove variables that will not be used
dat <-
  rawdat %>%
  select(-id, -idg, -condtn, -round, -position, -positin1, -order, -partner, -tuition, -undergra, -mn_sat)


#```{r Clean Data}
#Ratings by other ppl and self-rating

at00 <-
  dat %>%
  select(iid, pid, dec, gender, attr_o, sinc_o, intel_o, fun_o, amb_o, attr3_1, sinc3_1, intel3_1, fun3_1, amb3_1) %>%
  filter(!pid == "NA")

#drop rows where all attributes were rated NA (col 4-9)
#Since in the instructions it clearly outlined that not attributes will be discussed during
```
```

```
#the couple's meetings, we cannot do a full NA drop in the data. The workaround here is we will assign all NA values to 1000, and drop the rows if all the attributes add up to 5000. Rows with 1 or 2 NAs will add up to be less than 6000 and will not be dropped. Finally the rows with 1000 will be converted back to NA.

at00[is.na(at00)] <- 1000

at00$total <- rowSums(at00[,c("attr_o", "sinc_o", "intel_o", "fun_o", "amb_o")])
at00$total1 <-rowSums(at00[,c("attr3_1","sinc3_1","intel3_1","fun3_1","amb3_1")])

at00 <-
  at00 %>%
  filter(!total == "5000")

at00 <-
  at00 %>%
  filter(!total1 == "5000")

at00[at00 == "1000"] <- NA

at00$total <- rowSums(at00[,c("attr_o", "sinc_o", "intel_o", "fun_o", "amb_o")], na.rm=TRUE)

at00$total1 <- rowSums(at00[,c("attr3_1","sinc3_1","intel3_1","fun3_1","amb3_1")], na.rm=TRUE)

at00 <-
  at00 %>%
  filter(!total == "0")

#Finally, it is important to realize that the attributes are evaluated for the opposite gender.
#Another column for the partner is generated

at00 <-
  at00 %>%
  mutate(pgender = ifelse(gender == 0, 1, 0))

Others <-
  at00 %>%
  group_by(iid) %>%
  summarise(Attractive = mean(attr_o), Sincere = mean(sinc_o), Intelligent = mean(intel_o), Fun = mean(fun_o), Ambitious =
mean(amb_o))

Self <-
  at00 %>%
  select(iid,attr3_1, sinc3_1, intel3_1, fun3_1, amb3_1) %>%
  unique()

# merge the two table (left outer join)

Others_Self <- merge(x=Others, y=Self, by = "iid", all.x=TRUE)

library(data.table)

setnames(Others_Self, old=c("attr3_1", "sinc3_1", "intel3_1", "fun3_1", "amb3_1"), new=c("Attr(Self)", "Sinc(Self)", "Intel(Self)",
"Fun(Self)", "Amb(Self)"))

library(dplyr)
library(tidyr)
library(plotly)

plot_data <- Others_Self %>%
  gather(variable, value, -iid)
```

```
plot_data <-
  plot_data %>%
  mutate(self = ifelse(variable == 'Attr(Self)' |variable == 'Sinc(Self)'|variable == 'Fun(Self)'|variable == 'Amb(Self)'|variable == 'Intel(Self)',
"Self", "Others"))

splitplot <- plot_data

splitplot$variable[splitplot$variable == 'Attr(Self)'] <- "Attractive"
splitplot$variable[splitplot$variable == 'Sinc(Self)'] <- "Sincere"
splitplot$variable[splitplot$variable == 'Intel(Self)'] <- "Intelligent"
splitplot$variable[splitplot$variable == 'Fun(Self)'] <- "Fun"
splitplot$variable[splitplot$variable == 'Amb(Self)'] <- "Ambitious"


df <- splitplot

p <- df %>%
  plot_ly(type = 'violin') %>%
  add_trace(
    x = ~variable[df$self == "Self"],
    y = ~value[df$self == "Self"],
    legendgroup = 'Yes',
    scalegroup = 'Yes',
    name = 'Self',
    side = 'negative',
    box = list(
      visible = T
    ),
    meanline = list(
      visible = T
    ),

    line = list(
      color = 'orange'
    )
  ) %>%
  add_trace(
    x = ~variable[df$self == "Others"],
    y = ~value[df$self == "Others"],
    legendgroup = 'No',
    scalegroup = 'No',
    name = 'Others',
    side = 'positive',
    box = list(
      visible = T
    ),
    meanline = list(
      visible = T
    ),
    line = list(
      color = 'seagreen1'
    ),
    fillcolor="seagreen1"

  ) %>%
  layout(
    xaxis = list(
      title = "Attribute"
    ),
    yaxis = list(
      title = "Rating",
```

```
    zeroline = F
  ),
  violingap = 0,
  violingroupgap = 0,
  violinmode = 'overlay'
 )

p

```
Row {data-height=50}

Average Perception on Five Attributes

``` {r, fig.height=15, fig.width=17}


#Assign Gender and Age Group | calculate differences
#3_ is the score subjects gave themselves and _o is the score subject was given by partner
dating$Gender2 = ifelse(dating$gender ==1,"Male","Female")
dating$AgeGroup[dating$age <27 & dating$age>17 ] = "18 - 26"
dating$AgeGroup[dating$age <56 & dating$age>26 ] = "27 - 55"
dating$attrDiff = abs(dating$attr3_1 -dating$attr_o)
dating$sincDiff = abs(dating$sinc3_1 -dating$sinc_o)
dating$intelDiff = abs(dating$intel3_1 -dating$intel_o)
dating$funDiff = abs(dating$fun3_1 -dating$fun_o)
dating$ambDiff = abs(dating$amb3_1 -dating$amb_o)

#head(dating)

#Create subsets
datingAll = subset(dating,,select = c("Gender2","age","AgeGroup","attr_o","sinc_o","intel_o","fun_o",
                                      "amb_o","attr3_1","sinc3_1","fun3_1","intel3_1","amb3_1","attrDiff","sincDiff","intelDiff",
                                      "funDiff","ambDiff"))

#Find means for attraction, combine into table
Attraction= c((mean(datingAll$attr3_1 ,na.rm = T)),(mean(datingAll$attr_o,na.rm = T)))
Sincerity= c((mean(datingAll$sinc3_1 ,na.rm = T)),(mean(datingAll$sinc_o,na.rm = T)))
Fun= c((mean(datingAll$fun3_1 ,na.rm = T)),(mean(datingAll$fun_o,na.rm = T)))
Intelligence= c((mean(datingAll$intel3_1 ,na.rm = T)),(mean(datingAll$intel_o,na.rm = T)))
Ambition= c((mean(datingAll$amb3_1 ,na.rm = T)),(mean(datingAll$amb_o,na.rm = T)))
Combined= t(data.frame(Attraction,Sincerity,Fun,Intelligence,Ambition))
colnames(Combined) = c("Own Perception","Date Perception")
#Combined

range = data.frame(Attraction = c(10, 0),Sincerity = c(10, 0),Fun = c(10, 0),Intelligence = c(10, 0),Ambition = c(10, 0))
wC2=t(Combined)
#wC2
data1 = rbind(range,wC2)
#data1
radarchart(data1,
        axistype = 1,
        pcol = c(adjustcolor("#FF9900", 0.7), adjustcolor("#7fbf7b", 0.7)),
        pfcol = c(adjustcolor("#FF9900", 0.7), adjustcolor("#7fbf7b", 0.7)),
        cglcol = "gray50",
        centerzero = TRUE,
        seg = 5,
        vlcex = 4,
        palcex = 4,
        plty = 1,
        plwd = 3,
```

```
        cglty = 1,
        pty=32,
        axislabcol="black",
        caxislabels=seq(0,10,2)
        )

legend("topright",
     c("Own Perception", "Date's Perception"), box.lty=1, box.lwd=1,cex = 3,box.col = "gray50",
     fill = c(adjustcolor("#FF9900", 0.7), adjustcolor("#7fbf7b", 0.8)))
```

```
Row
-----------------------------------------------------------------------
### Avg Attr. Gap

```{r}
dating$attrDiff = abs(dating$attr3_1 -dating$attr_o)
comments = mean(dating$attrDiff,na.rm = T)
valueBox(round(comments,2), color = "gray")

```

Avg Sinc. Gap

```{r}
dating$sincDiff = abs(dating$sinc3_1 -dating$sinc_o)
comments1 = mean(dating$sincDiff,na.rm = T)
valueBox(round(comments1,2), color = " gray")

```

Avg Fun Gap

```{r}
dating$funDiff = abs(dating$fun3_1 -dating$fun_o)
comments2 = mean(dating$funDiff,na.rm = T)
valueBox(round(comments2,2), color = "gray")

```

Avg Intel. Gap

```{r}
dating$intelDiff = abs(dating$intel3_1 -dating$intel_o)
comments3 = mean(dating$intelDiff,na.rm = T)
valueBox(round(comments3,2), color = "gray")

```

Avg Amb. Gap

```{r}
dating$ambDiff = abs(dating$amb3_1 -dating$amb_o)
comments4 = mean(dating$ambDiff,na.rm = T)
valueBox(round(comments4,2), color = "gray")

```

Visualization 4: Parallel Coordinate Plot
```

```
---
    1. title: " Interactive Parallel Coordinate Plot”
output: word_document
---
```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
library(fmsb)
library(dplyr)
library(tibble)
library(stringr)
library(ggplot2)
library(grid)
library(gridBase)
library(scales)
library(psych)
library(plotly)
```
R Markdown
```{r Load Data and Cleaning }
setwd("~/Dropbox/Homework/DSC465/HW3")
rawdat <-
read.csv('Speed Dating Data.csv', header = T, stringsAsFactors = F)
```
```{r remove variables that will not be used}
dat <-
rawdat %>%
  select(-id, -idg, -condtn, -round, -position, -positin1, -order, -partner, -tuition, -undergra, -mn_sat)
```
```{r Prepare the Features}
at00 <-
dat %>%
  select(iid, pid, age, field_cd,race,attr3_1,sinc3_1,intel3_1,fun3_1,amb3_1,attr1_1,sinc1_1,intel1_1,fun1_1,amb1_1,shar1_1,match)
%>%
  filter(!pid == "NA") %>%
```
```{r Calculate Match Percentage}
match_df <-
at00%>%
  group_by(iid) %>%
  summarise(matchcount = sum(match)/n())
```
```{r Remove Duplicates}
features_df <-
  at00 %>%
select(iid,age,field_cd,race,attr3_1,sinc3_1,intel3_1,fun3_1,amb3_1,attr1_1,sinc1_1,intel1_1,fun1_1,amb1_1,shar1_1) %>%
    unique()
```
```{r Merge Dataset}
df_all <-merge(x = features_df, y = match_df, by = "iid", all = TRUE)
```
```{r}
df_all$highmatch <- ifelse(as.numeric(df_all$matchcount) > 0.14, 1,0)
```
```{r}
describe(df_all)
```
```{r Plot}
p <- df_all %>%
  plot_ly(width = 1000, height = 600) %>%
  add_trace(type = 'parcoords',
        #line = list(color = ~matchcount,
```
```

```
            #colorscale = "Viridis",
            #showscale = TRUE
            #),
        line = list(color = 'green',showscale = TRUE),


        dimensions = list(
          list(range = c(1,18),
             tickvals=c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18),
             label = 'Field', values = ~field_cd)
          list(range = c(1,6),
             tickvals=c(1,2,3,4,5,6),
             label = 'Race', values = ~race),
          list(range = c(1,10),
             tickvals=c(1,2,3,4,5,6,7,8,9,10),
             label = 'attr(self)', values = ~attr3_1),
          list(range = c(55,18),
             label = 'Age', values = ~age)
          list(range = c(1,10),
             tickvals=c(1,2,3,4,5,6,7,8,9,10),
             label = 'sinc(self)', values = ~sinc3_1),
          list(range = c(1,10),
             tickvals=c(1,2,3,4,5,6,7,8,9,10),
             label = 'intel(self)', values = ~intel3_1       ),
          list(range = c(1,10),
             tickvals=c(1,2,3,4,5,6,7,8,9,10),
             label = 'fun(self)', values = ~fun3_1),
          list(range = c(1,10),
             tickvals=c(1,2,3,4,5,6,7,8,9,10),
             label = 'amb(self)', values = ~amb3_1       ),
          list(range = c(~min(matchcount),~max(matchcount)),
             constraintrange = c(0,1),
             label = 'Match%', values = ~matchcount)
          )
        )
p

```
```{r}
# Create a shareable link to your chart
# Set up API credentials: https://plot.ly/r/getting-started
Sys.setenv("plotly_username"="ttanjocelyn")
Sys.setenv("plotly_api_key"="OXWKncxtmPfJWNwhCmcu")
chart_link = api_create(p, filename="parcoords-advanced")
chart_link
```
```
```