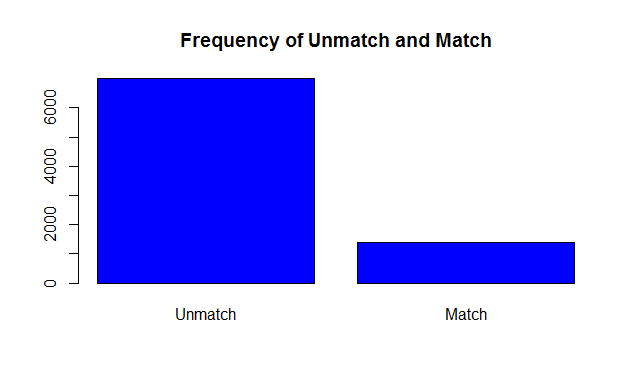
**Part A: Describe the dataset in your own words.**

This dataset is call speed dating data which include all the data that are related to speed dating experiment. Speed dating is a method of meeting potential partners in which each participant has only a few minutes to talk to each other before being moved on to the next one. At the end of the event, participants will decide whether they would dates again. This dataset was collected and done by Columbia Business School professors Ray Fisman and Sheena Iyengar in years 2002-2004. All the data was gathered from participants that involved in this experiment.

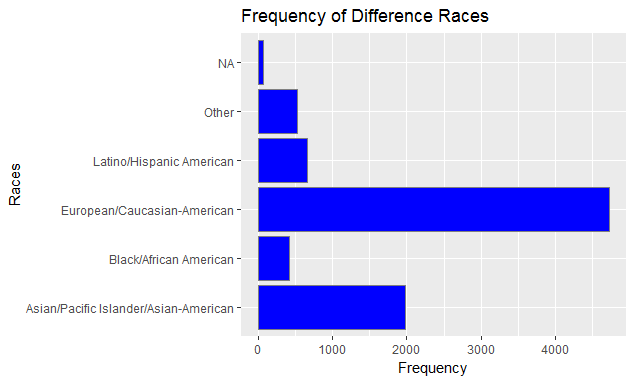
In this dataset, there are total number of 195 variables or columns are created and it had total number of 8378 observations or rows are recorded based on the experiment. Types of attributes in this dataset which are category attribute and continuous attribute. For the continuous attribute, it have binary type data example like 0/1, discrete type data and continuous type data. Because this dataset had involved binary type data, for this reason it can’t figure out whether it is normal distributed.

Some important variables in this dataset which are likes gender, age, race, goal, match and so on. The key variable or class attribute in this dataset which is Match. Table below shows the descriptions for each important variables in this dataset.

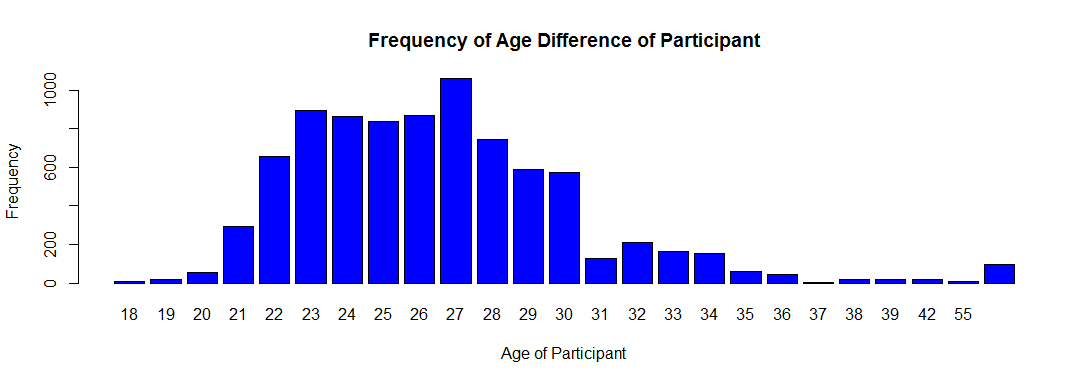
|  |  |  |
| --- | --- | --- |
| Variable | Description | Note |
| Gender | Gender of the participant | Female= 0, Male= 1 |
| Age  Age\_o | Age of the participant  Age of partner |  |
| Race  Race\_o | Race of the participant  Race of the partner | Black/African American= 1  European/Caucasian-American= 2  Latino/Hispanic American= 3  Asian/Pacific Islander/Asian-American= 4  Native American= 5  Other= 6 |
| Goal | Primary goal for participating in this event | Seemed like a fun night out= 1  To meet new people= 2  To get a date= 3  Looking for a serious relationship= 4  To say I did it= 5  Other= 6 |
| Match | Determine whether the participant is successfully match with the partner | Yes= 1, No= 0 |



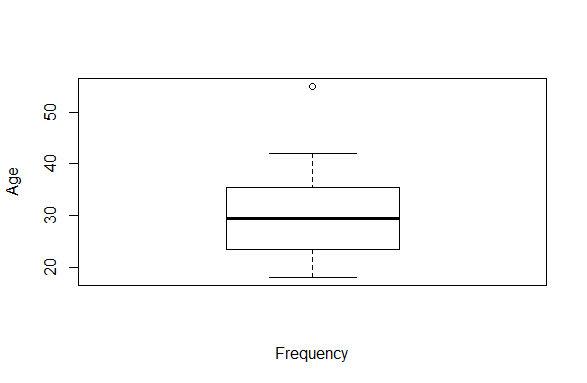
Bar chart above shows that the frequency of un-matched or matched result in this speed dating experiment. Based on the result, there are total number of 6998 participants which are un-matched with theirs partner throughout this experiment while only total number of 1380 participants are matched with theirs partner. From the result, we can say that majority of participants are unsuccessfully matched with others, because the frequency of un-matched result is more than five times the frequency of matched result.



Bar chart above shows that the frequency of difference races of the participants that participate in this experiment. Based on the bar chart, we can see that the most high frequency of races is European/Caucasian-American which have 4727 of participants than follow by Asian/Pacific Islander/Asian-American, Latino/Hispanic American, Other and last is Black/African American which only have 420 participants. Unfortunately, they are total of 63 participants are not showing theirs races in this dataset. Based on the result, we can say that the race of European/Caucasian-American have more courage to try this experiment compare to others race.



Bar chart above shows that the frequency of age difference of participant that participate in this experiment. The age group that participate in this experiment are from the smallest age which is 18 to the largest age which is 55. Based on the dataset, the most high frequency of age that participate in this experiment which is falls on age 27 while the most low frequency which is falls on age 37. After calculation, the mean of age that participate in this experiment which is 26.35893. Based on the result, we can say that most of the participant that participate in this experiment which are adult and it is falls on age group of 21 to 30.



Above is the box-and-whisker plots which shows that the outliers for the age differences of the speed dating participant. In the plots, it shows the age 55 is the outlier of the age for the participant, this is because in this data set the average age for all the participants are 26.35893, therefore this makes that age 55 have some serious age gap compare to the average age.

**Section B: Possible Insight on the Speed Dating Dataset**

There are many different exploratory ideas one can get within this speed dating dataset. The idea here is to focus more on the chances of getting a match with a speed date partner and go on a second date.

**Insight:**

Do people who go on dates frequently have higher chance to get second date than those who does not?

For the attribute go on dates frequently in the dataset have a fairly vague meaning, it can mean how frequent that a person went on dates with different person, or it can also mean that how frequent a person went on dates with the same person.

People who goes on date frequently would have slightly more advantage and experience in terms of what to expect in dating, they may even come up with interesting topic or pickup lines that may interest their speed dating partners to leave a good impression that is nice enough to land them a chance for second date.

For this insight, we only consider the case where by the people went on dates with the same person each time which is an advantage for this speed dating case. Besides that, in order to further describe the participant that are successful in matching we also take in consideration of the participant’s activity of interest.

**Section C: Type of data mining technique**

For the insight that has been discussed earlier, we think that classification would be relevant as the technique for extracting the rules from the training data that will predict the categorical labels of the “match” column in the data set. This process is also known as supervised learning because the class label is known for each of the training samples.

At the initial step, the original data set will be treated with pre-processing tasks to deal with possible data quality issues like duplication of records and missing values. After getting the clean and tidy data, the next step is to split the data into 2/3 where it will be used as the training samples and the remaining 1/3 as the testing samples. Random sampling without replacement is applied in this case as the technique for choosing both the training and testing data sets. In addition to that, the original data set comes with 195 columns of attributes, however the main focus for this discussion is to build a model that will use the dater’s frequency of going out on a date, several basic information like gender and age, and field of interest in different activities as the predictor columns to predict the class label for the “match” column (with 1=yes, and 0=no).

These predictor columns are named as iid, gender, age, date, sports, dining, art, gaming, clubbing, reading, movies, concerts and match in the data set.

Table below shows the example of the training data where the records have been labelled with 0 or 1 in the “match” column (the meaning of each columns is described at below).

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| idd | gender | age | date | sports | dining | art | gaming |
| 1 | 0 | 21 | 3 | 6 | 8 | 4 | 5 |
| 2 | 1 | 21 | 1 | 8 | 6 | 6 | 7 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| clubbing | reading | movies | concerts | shopping | match |
| 8 | 3 | 9 | 7 | 10 | 0 |
| 4 | 1 | 5 | 3 | 6 | 1 |

iid Unique subject number

gender Gender of the subject

* 0= Female
* 1= Male

age Age of the subject

date Frequency of going dates

* 1= Several times a week
* 2= Twice a week
* 3= Once a week
* 4= Twice a month
* 5= Once a month
* 6= Several times a year
* 7= Almost never

How interested are you in the following activities on a scale of 1-10

sports Playing sports/ athletics

dining Dining out

art Art

gaming Gaming

clubbing Dancing/clubbing

reading Reading

movies Movies

converts Going to concerts

shopping Shopping

match Whether the dating match or not

* 0= Yes
* 1= No

The goal of classification is to create a set of classification rules called a classifier or model that will do the work of classifying or categorizing a given data record using the classification algorithm. Therefore, the choice of algorithm is very important depending on the goal and the nature structure of the data set. In this task, we are going to learn a classifier model using the Naive Bayes algorithm. The reason of choosing this algorithm is because it is one of the most efficient and effective algorithm for supervised learning. It assumes that the features are independent. Each feature contributes to the probability to identify the classification independently. For instance, a fruit can be classified as an apple based on the color, circle shape and the diameter. Although the features exist relationship among themselves, the features contribute independently in the probability to identify the fruit is apple (Ray, 2015).

Just to recall, in the process of building a classifier, the attribute to be predicted are known in the training set, but not known in the testing set. The predictor columns are used to determine the value of the column “match” in this case. After the classifier model has been generated, the testing set will be used to test the classifier for accuracy, robustness, and speed. To visualize the performance of this classifier, we can adopt receiver operating characteristic (ROC) to plot a two-dimensional graph in which false positive rate is plotted on the x-axis and true positive rate is plotted on the y-axis. ROC is suitable in this case because our classifier is a binary system, which mean that the result of the output is either 0 or 1.

**Reference:**

Ray, S. (2015). *6 easy steps to learn naive bayes algorithm (with code in python)*. Retrieved from <https://analyticsvidhya.com/blog/2015/09/naive-bayes-explained/>

Fawcett, T. (2003). *ROC Graphs: Notes and Practical Considerations for Data Mining Researchers.* Retrieved from <http://www.hpl.hp.com/techreports/2003/HPL-2003-4.pdf/>

**Section D: Describe data quality issues, and be specific.**

Attribute Career is having misspelling issue. For example, one of the column in the dataset is career. There are several choices for this column but it is inconsistent because there are different terms that are referring to the same thing. For instances, law, Law, lawyer and Lawyer. It is sometimes caused by the computer program. Another common reason for misspelling is where words are spelt differently from their pronunciation. Sometimes, data entry clerk may be tired or lose some focus, so they are unaware of their spelling mistake. In order to solve this problem, we can use “adist” function in R to look for the minimum character substitution required to convert these terms into a consistent and correct spelling.

Attribute career\_c is having missing data issue. Attribute career and career\_c is correlated, there are many career options in the list and each of them is labelled. In addition, when the user fill in their career, the attribute career\_c will fill in with the number list of the options. For example, lawyer in the career option is “1”, so in the attribute career\_c, it should fill in with “1”. In the data set, there are some value missing when the user did fill in career but in the attribute career\_c didn’t show up anything. Missing data occurs for many reason in survey. Sometimes, daters may refuse to answer a question due to privacy issues. Perhaps, the respondent just lost of interest or do not have enough time to complete the questionnaire. Every survey question without an answer is a missing data point. Beside that, missed observations may occur when data entry. There is a way to solve the problem which is using imputation.

Attribute data\_3 is having outlier issue. In this attribute, there is two option for the user to select which 1=yes and 2=no. In the data set, there is invalid number show on the column which is ‘0’, the range of the options is from 1 to 2, so 0 is invalid in this case. This will occur due to the reason of entering the wrong value at the beginning. In order to identify these outliers, we could use whisker box plot.