Introduction to Data Science CapStone (Jonathan Dinh Jhd9252)

Pre-Processing (Dimension Reduction, Data Cleaning, Data Transformation, and Missing Data): Out of 523269 possible entries in the dataset, around 62.6% of it is missing (or 327526 entries). I am choosing to remove missign entries row-wise according to each question. Inputation of missing data by mean, median, mode or regression would reduce the variance too much, and cause high correlation with dependent variables by selection. To begin, I separated out the 3 categories of sensation seeking, movie experience and personality. After scaling the data to standard with mean = 0 and std = 1, I conducted PCA to extract the synthetic vartiables of lower-dimensional space by the Kaiser criterion. As result, sensation seeking had 6 PCs, personality had 7 PCs, and movie experience had 2PCs.

Question 1: What is the relationship between sensation seeking and movie experience?

From pre-processing, the data of sensation seeking and movie experience is already scaled to standard, and we have the principal components. To find an insight into the relationship of sensation seeking and movie experience, we first find the average per category, per row. The correlation then, between sensation seeking and movie experience is 0.00742265. This fails to provide any evidence that there is a linear relationship between sensation seeking and movie experience.

Question 2: Is there evidence of personality types based on data of participants? If so characterize these types both quantitaively and narratively.

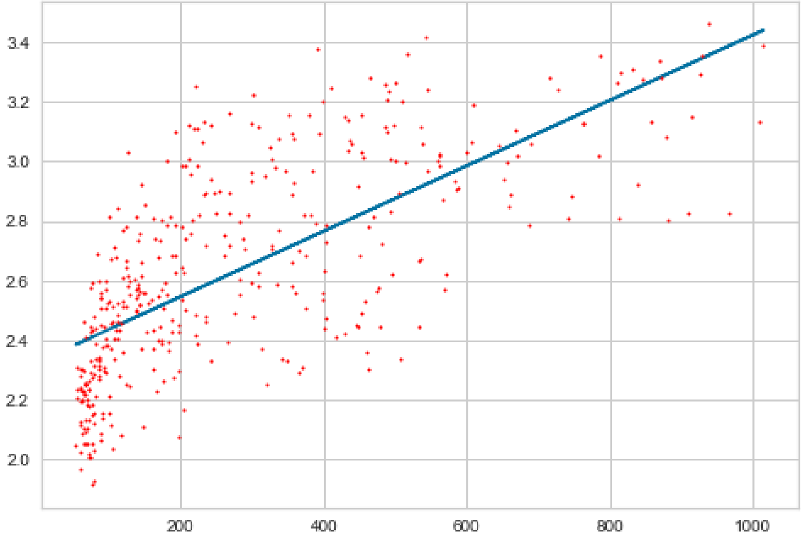
The first step is to ensure the data is scaled to standard, and PCA conducted. This was done in pre-processing. The find evidence of personality types, I used kmeans unsupervised clustering, as the input PCA had no labeling. To obtain the optimal amount of clusters, I used the elbow method ranging from k=(1,10) to obtain k=3 optimal clusters. From there, I fitted and predicted using the principal components as inputs and labeled each partipant a cluster. From interpreting the the scree plot of personality principal components;

PCA 1	Does a thorough Job
PCA 2	Careless
PCA 3	Shy/Inhibited, Worries a lot, Efficient
PCA 4	Is talkative
PCA 5	Original, comes up with new ideas
PCA 6	Talkative
PCA 7	Depressed, Generates a lot of enthusiasm

Characterizing by highest and lowest values, participants in cluster 1 low level carelessness (-1.39667) and high level of shyness, worry and efficiency (1.30212). Cluster 2 is characterized by extremely high level of thoroughness at job (2.26012) and relatively low level of shyness, worry and efficiency (-0.662832). Cluster 3 is characterized by extremely poor thoroughness at job (-2.66913) and slight carelessness (0.688983). Therefore this is some evidence to support that there are personality types.

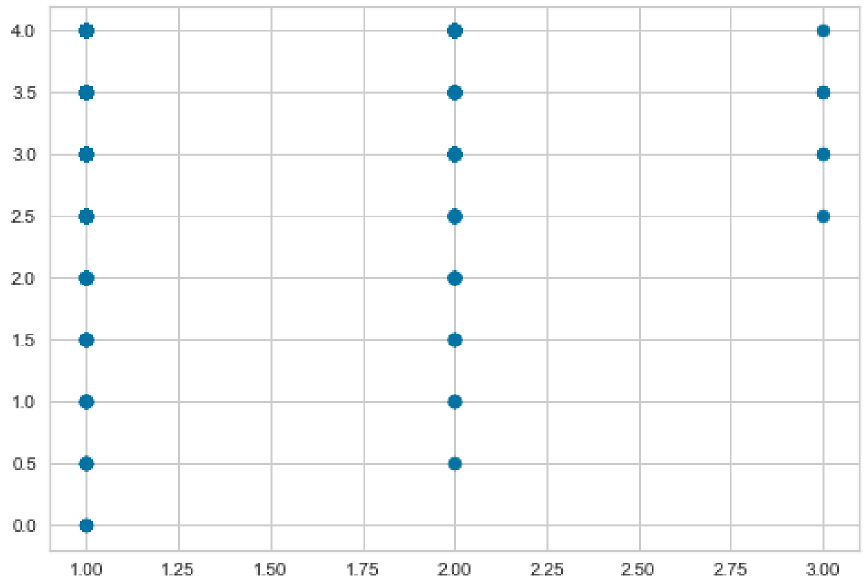
Question 3: Are movies that are more popular rated higher than movies that are less popular?

Operationalized popularity of each movie by the number of ratings recorded. This goes off the assumption that more popular movies will have more people selecting to watch. To find the relationship between popularity and movie ratings, I sought the correlation between the two. Using the mean rating of each movie and its number of ratings, there is a high positive correlation of around 0.7. This indicates a strong positive linear relationship.



Q4: Is enjoyment of 'Shrek (2001)' gendered, i.e. do male and female viewers rate it differently?

First I separated out the self-identified genders, removing and missing data. From the descriptive statistics, the average rating for self-identified females is 3.15545 and the standard deviation is 0.906547. For self-identified males, the average rating is 3.08299 while the standard deviation is 0.824975. The sample size of self-identified gender of 'Other' is low at n = 6, with average rating of 3.25 and standard deviation of 0.524404. The below plot shows the x-axis as gender, and the y-axis as rating. We see that generally, the participants who identified other had the smallest range of ratings, followed by men, then women. In addition, doing an independent parametric t-test in pairs (female-male, female-other, male-other) resulting in p-values of 0.24834907946281018, 0.679928726393761, 0.47941799800855334 respectively. These are all greater than designated alpha level 0.05, and therefore this is no significant indication that there is any difference in enjoyment of Shrek due to gender. Conducted using Welch's t-test for unequal variances.



Q5. Do people who are only children enjoy ‘The Lion King (1994)’ more than people with siblings?

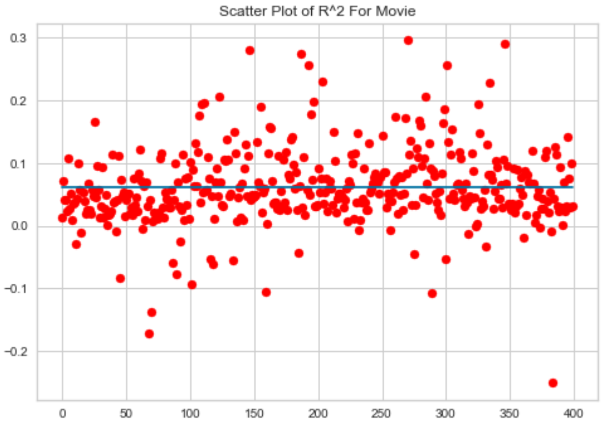
Excluded missing entries from pertaining data, recorded entries of ‘-1’. I separated out the participants who recorded as an only child, and those that were not. From basic descriptive statistics, the average rating given by those who are an only child for The Lion King is 3.34768, with a standard deviation of 0.816483. For those that are not an only child, the average rating is 3.48196, with a standard deviation of 0.718194. Conducting an independent parametric t-test, resulting in a p-value of 0.06102886373552747, indicating no significant evidence of difference in rating due to siblingship. This was done with the Welch’s t-test with the assumption of unequal variance. This is due to extremely small sample sizes, unequal sample sizes. Alternatively however, if there was the assumption of equal variance among those who were an only child and those who were not, we obtain the p-value of 0.0402670552626826 which would support rejecting the null hypothesis that there is no difference in enjoyment due to siblingship.

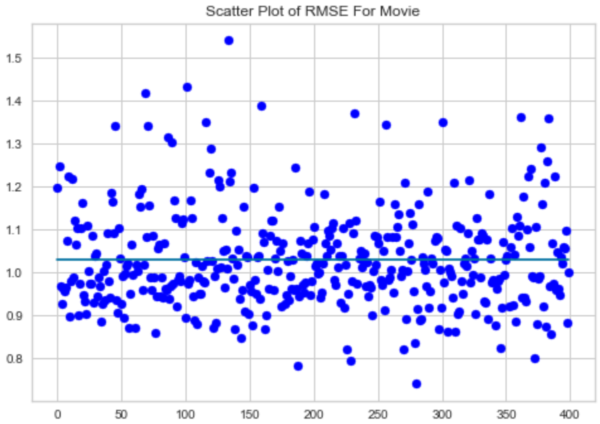
Q6. Do people who like to watch movies socially enjoy ‘The Wolf of Wall Street (2013)’ more than those who prefer to watch them alone?

After separating out those with different preferences in viewer company, those that enjoy movies alone on average rate The Wolf of Wall Street as 3.14377 stars with a standard deviation of 0.869886. Those who enjoy watching movies with others on average rated the movie as 3.03333 with standard deviation of 0.921047. Conducting an independent parametric t-test assuming unequal variance, we have a p-value of 0.12139103950020742 which does not support the rejection of the null hypothesis. Therefore we fail to reject the null as there is no substantial evidence that supports a difference in populatino rating due to viewing preference.

Q8. Build a prediction model of your choice (regression or supervised learning) to predict movie ratings (for all 400 movies) from personality factors only. Make sure to use cross-validation methods to avoid overfitting and characterize the accuracy of your model.

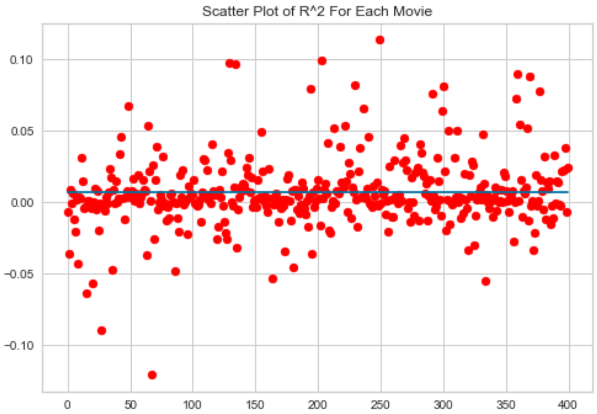
A multiple linear regression model is based on $y = a + B_1X_1 + B_2X_2 + \cdots + B_nX_n$. Since the variables attributed to personality has been scaled and undergone PCA, the same must be done to all 400 movies. I then drop the rows that have missing personality entries, as those are our independent variables. In order to predict 400 movie ratings from personality, I utilize a loop. For each loop, I drop entries row-wise that are missing ratings for that movie, and split the data into x_train, x_test, y_train, y_test with a 0.7:0.3 training to test ratio in order to cross-validate. From there I fit the model, predict and obtain the $R^2 and\ RMSE$ for each regression. The average $R^2 and\ RMSE$ for these regressions is 0.064 and 1.03 respectively. Generally, this means that the model on average can only explain for around 6.4% of the variance. However this does not mean the regression model is not an adequate fit for the data. Human preference or behavior will often have an a lower R^2. But with and average RMSE of around 1, taking into account the scale of the dependent variable (rating), as well as the R^2, we can characterize the model as not accurate.

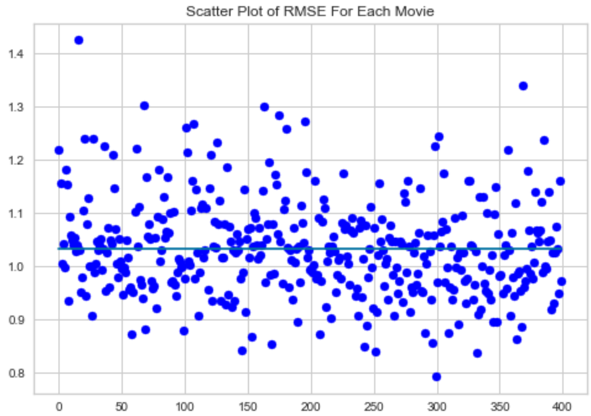




Q9. Build a prediction model of your choice (regression or supervised learning) to predict movie ratings (for all 400 movies) from gender identity, sibship status and social viewing preferences (columns 475-477) only. Make sure to use cross-validation methods to avoid overfitting and characterize the accuracy of your model.

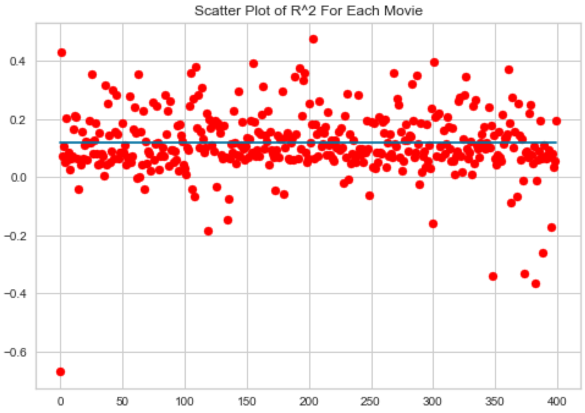
Following the same vein as Question 8, I created a new dataframe with the required ratings, gender, siblingship and social viewing preferences data. The independent variables are categorical and not on a numerical scale, and therefore do not require PCA. The movie ratings are already scaled. I conducted a multiple linear regression per movie column. The average $R^2 and\ RMSE$ respectively is 0.0073, 1.027763 respectively. This indicates that predicting movie ratings from gender identity, siblingship and social viewing preference is even less accurate than using personality.

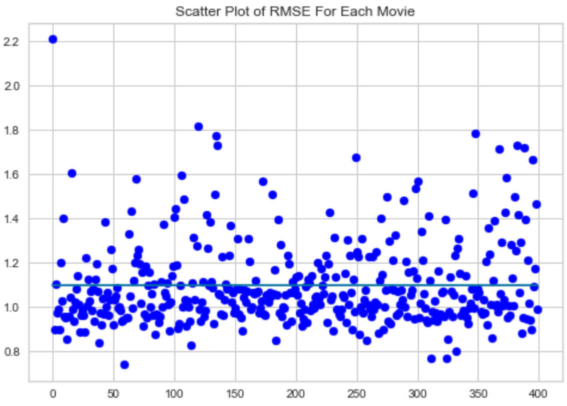




Q10. Build a prediction model of your choice (regression or supervised learning) to predict movie ratings (for all 400 movies) from all available factors that are not movie ratings (columns 401-477). Make sure to use cross-validation methods to avoid overfitting and characterize the accuracy of your model.

From Question 8 and Question 9, I assume that taking in account more independent and important variables into account will provide a higher average R^2 and lower $RMSE$. The movie ratings are scaled to standard. Sensation seeking, personality and movie experience are scaled and PCA'd. Gender identity, siblingship and social viewing preference do not need to be cleaned or transformed. I conducted a multiple linear regression with 18 independent variables and 1 dependent variable (rating) per participant, per movie. The average R^2 of the model is 0.12346, and the $RMSE$ is 1.0905035. Compared to the previous prediction models, this model is able to explain around twice as much variance as Q8, and 24 times as much variance as Q9. However the $RMSE$ is the same, however, suggesting that the goodness of fit of this model is around the same as the others. From the scatter plot, both $R^2and\ RMSE$ are tigher around the mean than the other models.


Scatter Plot of R^2 For Each Movie


Scatter Plot of RMSE For Each Movie