Sunny Son

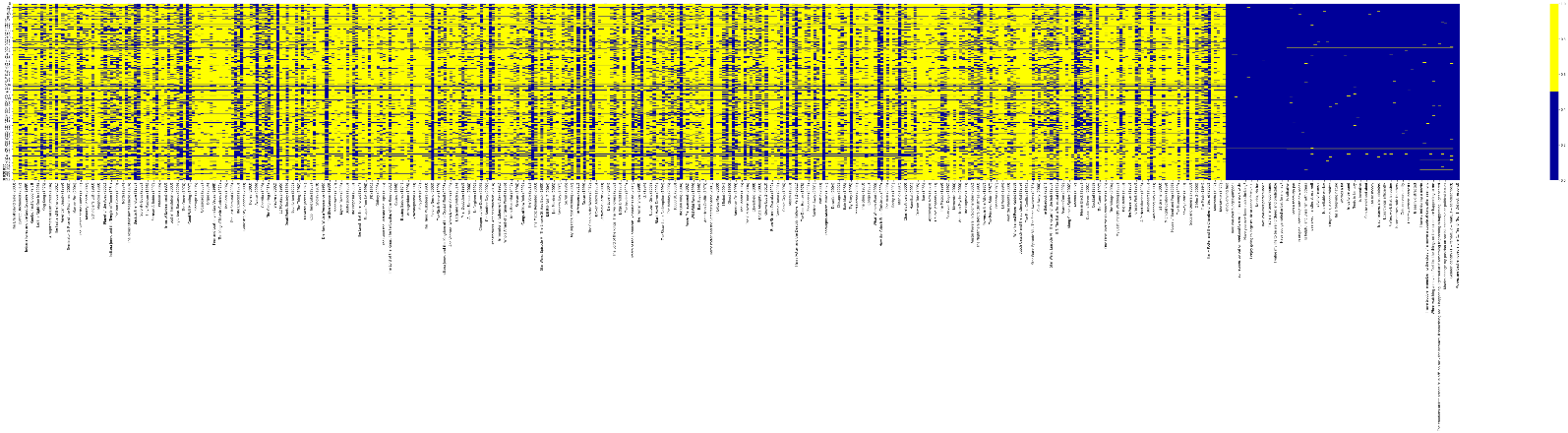
Introduction to Data Science, Section 001

Professor Wallisch Pascal

24 December, 2021

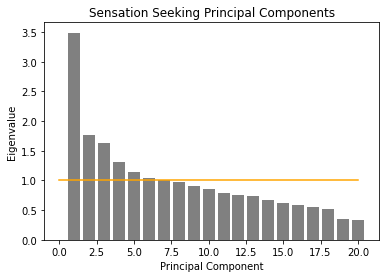
Introduction to Data Science Capstone Project

**Data Cleaning & Handling**

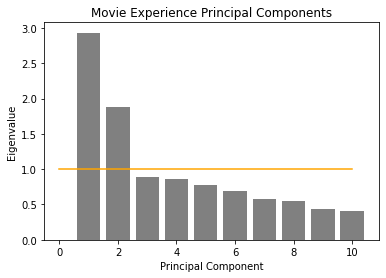
Before manipulating/analyzing/interpreting the data, the data frame was first interpreted as a heatmap of missing vs. presented data (shown below).

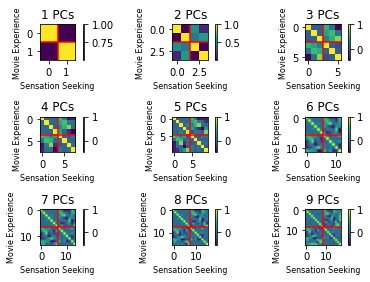
For data cleaning and transformation, a mix between imputation and row-wise removal was implemented. The reasoning behind this was motivated by the cause of the missing values as "item non-response," with movie ratings themselves being sparse across all 400 columns, but latter survey questions having few missing responses (removed in place, across all features). With the fact that these characteristics are designed to be unique, a k-NN imputation method was implemented on them, by creating a basic mean impute then using the resulting complete list to construct a K-d Tree. It then uses the resulting K-d Tree to compute nearest neighbors (NN). After it finds the k-NNs, the weighted average of values is taken. (Badr, 2019)

For dimensional reduction, the process was only carried out on non-movie ratings data, specifically the characteristic data columns, the gender identity column, the only-child column, and the social viewing preference column. The method implemented was the PCA, with further uses being made on movies data as required.

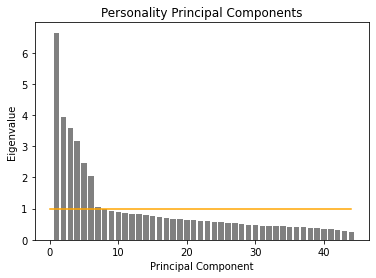


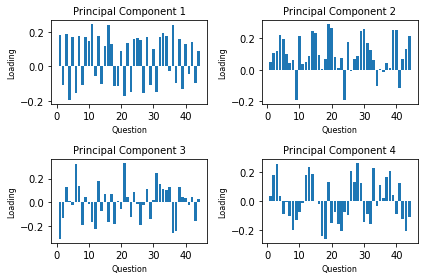
**Question 1: What is the relationship between sensation seeking and movie experience?**

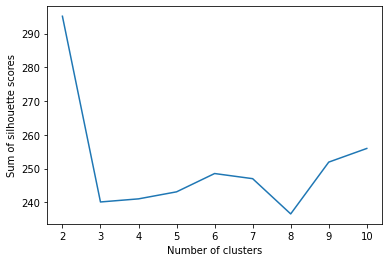
Initially, we took each characteristic and imputed values (as stated above in the Data Cleaning & Handling section), before carrying out a PCA (Principal Component Analysis) on either predefined characteristic. By the Kaiser Criterion, Sensation Seeking is supposed to have 6 factors, and for Movie Experience, the Critical Number is 2. However, I wanted to compare the data accounting for as much of the variance as possible, to compare the correlation matrices between (up to 10) of the principal components of each characteristic. The resulting plot is shown below:

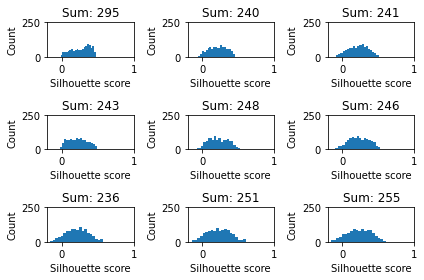
The interpretation for this graph can be given by the thought that, as more granularity is given to the amount of variance we want to account for (through the addition of more principal components for each characteristic). Looking at the bottom left (and top right) quadrants of each graph (bounded by the red lines), we see the correlation between the principal components of Sensation Seeking and that of Move Experience.

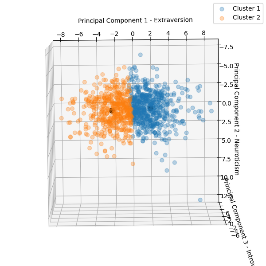
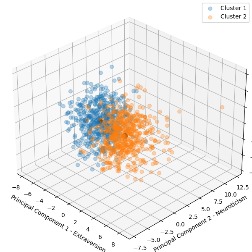
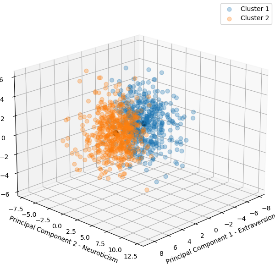
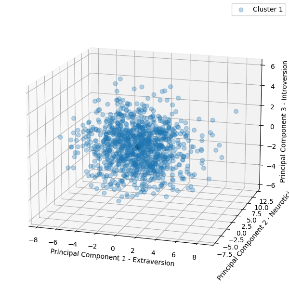
Through perceiving the actual correlations, the quadrants stabilize around 0.4, with sporadic correlations which are negative. In general, there is not a *strong positive* relationship between Sensation Seeking and Movie Experience, but in general a *positive* relationship *does* exist.

**Question 2: Is there evidence of personality types based on the data of these research participants? If so, characterize these types both quantitatively and narratively.**

Interpreting “personality types” as ‘clusters,’ we seek to find specific clusters within the Personality Characteristic questions, indicating predetermined propensities for answers based on each other. We first run a PCA on all 44 of the of the imputed (as described in the data cleaning section) columns of data. Next, I decided to utilize the first 3 PCs (instead of the 7 shown right) as to be able to accurately visualize the output. These PCs are (in order) then interpreted as extraversion, neuroticism, and introversion. Finally, a k-Means clustering algorithm was applied on the data to find the optimal number of centroids.

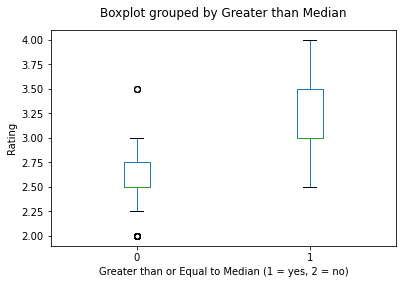
Looking at the silhouette plot of the k-Means (shown below) as well as the sum of silhouette scores (below), we can see that the optimal number of “personality types,” as evidenced by silhouette scores, is 2.



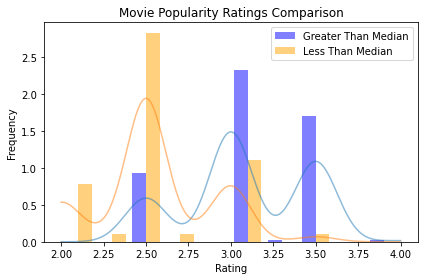
****

However, by looking at the plotted 3d graph of the first 3 PCs for personality, we don’t see any *distinctively* separable or *clustered* regions of data. Therefore, while silhouette suggests the optimal number of centroids is 2, by visualization the number is 1.

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

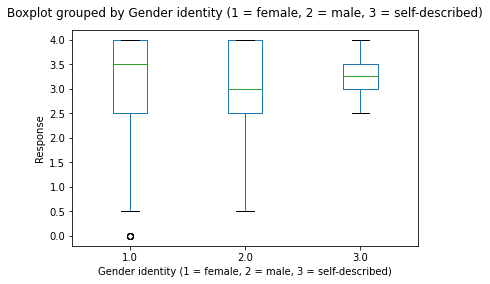
Firstly, the number of non-NaN entries is calculated for each movie (across axis=0), and the median of that series is then set as the deciding factor for whether a movie is classified as “greater than median” or “less than median.” This is the method of operationalizing for “movie popularity,” with the definition being that the top half of movies, ordered by viewership, are popular. Furthermore, for each “greater than median” and “less than median,” set of data, the median of each movie was then taken to create two arrays of data.

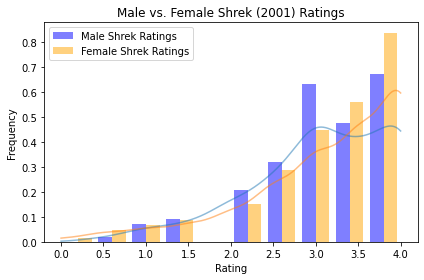
First, a Kolmogorov-Smirnoff test was carried out on the two sets of “greater than median” and “less than median” data. The *U*-statistic and *p*-value for the test resulted in 0.575, 6.7854964852395645e-31. We can interpret this as saying that, with a significance level of 0.05, we can reject the null hypothesis (that “movies greater than median” (movies that are more popular) and “movies less than median” (movies that are less popular) come from the same underlying distribution.

Then, a Mann-Whitney U test was carried out on both sets of data, with the alternative hypothesis being that the movies defined as “more popular,” i.e., had more than the median number of ratings, returning a *U*-statistic and *p*-value of 33427.5, 9.929258851707232e-35. We can interpret the results of this test as saying that with a significance level of 0.05, we can reject the null hypothesis that movies that are more popular is not rated higher than movies that are less popular.

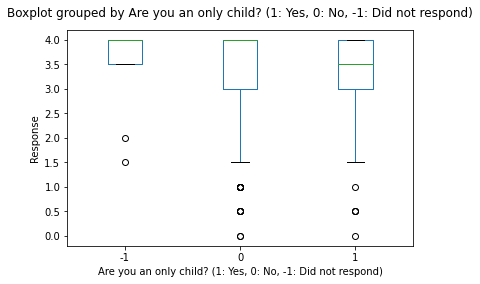
**Question 4: Is enjoyment of ‘Shrek (2001)’ gendered, i.e. do male and female viewers rate it differently?**

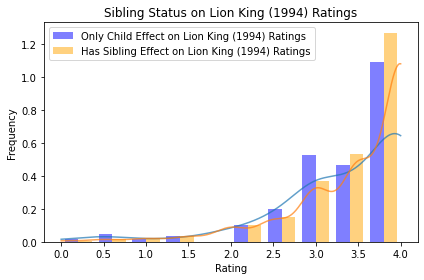
To carry out this comparison, the column of all movie ratings for “Shrek (2001)” was concatenated with the column of data for Gender Identity, and any/all NaN values from either column were dropped (using row-wise removal of missing data). Next, all remaining “Shrek (2001)” ratings associated with a female rating were grouped, and vice versa (male “Shrek (2001)” ratings).

 First, a Kolmogorov-Smirnoff test was carried out on the two sets of “male Shrek (2001) ratings” and “female Shrek (2001) ratings” data. The *D*-statistic and *p*-value for the test resulted in 0.09796552051512596, 0.056082040722863824. We can interpret this as saying that, with a significance level of 0.05, we **cannot** reject the null hypothesis that “female viewers” and “male viewers” come from the same underlying distribution.

 Then, a Mann-Whitney U test was carried out on both sets of data, with the alternative hypothesis being that the median ratings of “Shrek (2001)” by female viewers being different than the median ratings of “Shrek (2001)” by male viewers, returning a *U*-statistic and *p*-value of 82232.5, 0.050536625925559006. We can interpret the results of this test as saying that with a significance level of 0.05, we **cannot** reject the null hypothesis that there is a difference between median ratings of “Shrek (2001)” for female and male viewers.

**Question 5: Do people who are only children enjoy ‘The Lion King (1994)’ more than people with siblings?**

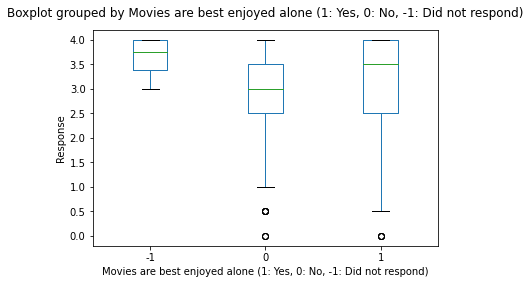
 To carry out this comparison, the column of all movie ratings for “The Lion King (1994)” was concatenated with the column of data for “Are you an only child?”, and any/all NaN values from either column were dropped (using row-wise removal of missing data). Next, all remaining “The Lion King (1994)” ratings associated with an “only child rating” were grouped, and vice versa (“has siblings”, “The Lion King (1994)” ratings).

 First, a Kolmogorov-Smirnoff test was carried out on the two sets of “only child The Lion King (1994) ratings” and “has siblings The Lion King (1994) ratings” data. The *D*-statistic and *p*-value for the test resulted in 0.09913292824469175, 0.15449987478607996. We can interpret this as saying that, with a significance level of 0.05, we **cannot** reject the null hypothesis that “only child ratings for Lion King (1994)” and “has sibling ratings for Lion King (1994)” come from the same underlying distribution.

Then, a Mann-Whitney U test was carried out on both sets of data, with the alternative hypothesis being that the median ratings of “The Lion King (1994)” by only child viewers being different than the median ratings of “The Lion King (1994)” by has-sibling viewers, returning a *U*-statistic and *p*-value of 52929.0, 0.978419092554931. We can interpret the results of this test as saying that with a significance level of 0.05, we **cannot** reject the null hypothesis that there is a difference between median ratings of The Lion King (1994) between viewers that are an only child and viewers that have siblings.

**Question 6: Do people who like to watch movies socially enjoy ‘The Wolf of Wall Street (2013)’ more than those who prefer to watch them alone?**

To carry out this comparison, the column of all movie ratings for “The Wolf of Wall Street (2013)” was concatenated with the column of data for “Movies are best enjoyed alone?”, and any/all NaN values from either column were dropped (using row-wise removal of missing data). Next, all remaining “The Wolf of Wall Street (2013)” ratings associated with an “prefers watching movies alone” were grouped, and vice versa (“prefers watching movies socially”, “The Wolf of Wall Street (2013)” ratings).

 First, a Kolmogorov-Smirnoff test was carried out on the two sets of “prefers watching movies alone, The Wolf of Wall Street (2013) ratings” and “prefers watching movies socially, The Wolf of Wall Street (2013) ratings” data. The *D*-statistic and *p*-value for the test resulted in 0.06420695504664971, 0.49837656470415104. We can interpret this as saying that, with a significance level of 0.05, we **cannot** reject the null hypothesis that “prefers watching movies alone, The Wolf of Wall Street (2013)” and “prefers watching movies socially, The Wolf of Wall Street (2013)” come from the same underlying distribution.

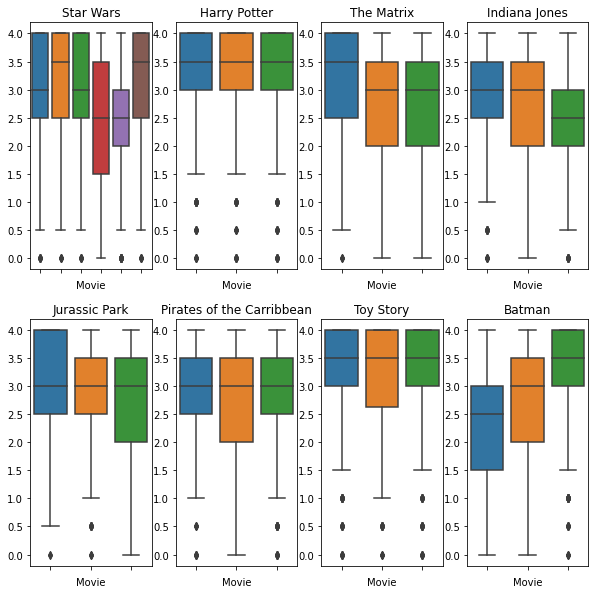
 Then, a Mann-Whitney U test was carried out on both sets of data, with the alternative hypothesis being that the median ratings of “The Wolf of Wall Street (2013)” by alone preference viewers being different than the median ratings of “The Wolf of Wall Street (2013)” by social preference viewers, returning a *U*-statistic and *p*-value of 49303.5, 0.9436657996253056. We can interpret the results of this test as saying that with a significance level of 0.05, we **cannot** reject the null hypothesis that there is a difference between median ratings of The Wolf of Wall Street (2013) between viewers that prefer to watch movies alone and viewers that prefer to watch movies socially.

**Question 7: There are ratings on movies from several franchises ([‘Star Wars’, ‘Harry Potter’, ‘The Matrix’, ‘Indiana Jones’, ‘Jurassic Park’, ‘Pirates of the Caribbean’, ‘Toy Story’, ‘Batman’]) in this dataset. How many of these are of inconsistent quality, as experienced by viewers?**

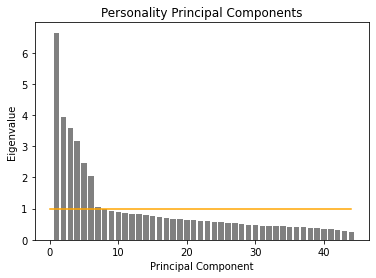
In this case, I interpreted “inconsistent quality” as having one or movie in any given franchise, be shown significant, under a Kruskal-Wallis ANOVA test on ranks, to not have been derived from the same underlying distribution.

* Star Wars Franchise: *H-*statistic, *p*-value = 230.5841753686405, 8.01647736660335e-48
* Harry Potter Franchise: *H*-statistic, *p*-value = 1.9994627621643, 0.3679782738223946
* The Matrix Franchise: *H*-statistic, *p*-value = 48.37886652130624, 3.123651788077413e-11
* Indiana Jones Franchise: *H*-statistic, *p*-value = 15.180674501620055, 0.0005053106088122954
* Jurassic Park Franchise: *H*-statistic, *p*-value = 46.59088064385204, 7.636930084365822e-11
* Pirates of the Caribbean Franchise: *H*-statistic, *p*-value = 20.64399756002606, 3.2901287079094474e-05
* Toy Story Franchise: *H*-statistic, *p*-value = 24.385994936261316, 5.065805156542464e-06
* Batman Franchise: *H*-statistic, *p*-value = 190.53496872634642, 4.2252969509030006e-42

Interpreting the results, we can see that for all the franchises, only the Harry Potter franchise is such that the p-value result of the Kruskal-Wallis test is not significant (p > 0.05), meaning that we cannot reject the null hypothesis (the movies are of similar quality as operationalized by rating). For all other franchises, we can see that the p-values are significant (< 0.05), and thus we have evidence to reject the null hypothesis (that the movies in each franchise are of similar quality) in favor of the alternative (not every movie in each franchise are of similar quality) for **every franchise except the Harry Potter franchise**.

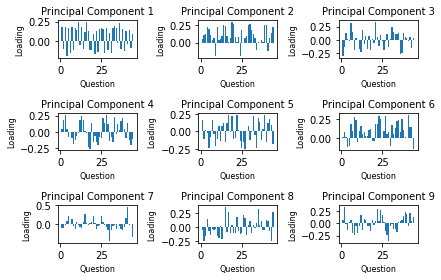


Looking at the boxplot above, it is easy to see that only the Harry Potter franchise is of similar (consistent) quality, as opposed to every other franchise compared where there exist large deviations.

**Question 8: 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.**

Looking at the graphs to the right, it is easy to see that the number of principal components we need to fulfill the Kaiser Criterion is 7 (This is again, based off the answer in Question 2, and utilizes the k-d Tree imputed data for all questions in the characterizes). The first 7 principal components will then serve as the predictors to a multivariate regression.

In terms of model, I decided on the Extreme Gradient Boosting algorithm, mostly due to its portability and robust features in the face of sparse data. The regressor format of the module was called and was then iterated through each of the 400 movies, appending to a dictionary of movie names (as keys) and model itself, the r^2 score, and k-fold cross-validated RMSE (both as a mean of the 5 splits provided) as values.

* For the r^2 score, cross\_val\_score from the sklearn.model\_selection was utilized on all the training data, with the parameter scoring='r2'
* For the RMSE value, cross\_val\_score from the sklearn.model\_selection was utilized on all the training data, with the parameter scoring=' neg\_root\_mean\_squared\_error'

Lastly, a final function was called to conglomerate and print the arithmetic average of all 400 models (one on each movie), and obtained a grand r^2 score and cross-validated RMSE of:

Score: 0.25849655194674614, Cross Validated RMSE: 1.2417111645824612

**Question 9: 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.**

For this question, I did not find it unreasonable to predict the rating for each movie based off the 3 qualities of gender identity, sibship status, and social viewing preferences. The methodologies were again very similar, starting with k-d Tree imputation for row-wise variables.

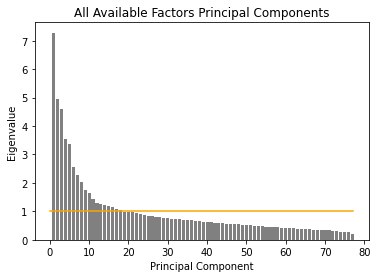
Similar to the previous question, the XGBoost algorithm was implemented, due to the robustness of the model to sparse data. The regressor format of the module was called and was then iterated through each of the 400 movies, appending to a dictionary of movie names (as keys) and model itself, the r^2 score, and k-fold cross-validated RMSE (both as a mean of the 5 splits provided) as values.

* For the r^2 score, cross\_val\_score from the sklearn.model\_selection was utilized on all the training data, with the parameter scoring='r2'
* For the RMSE value, cross\_val\_score from the sklearn.model\_selection was utilized on all the training data, with the parameter scoring=' neg\_root\_mean\_squared\_error'

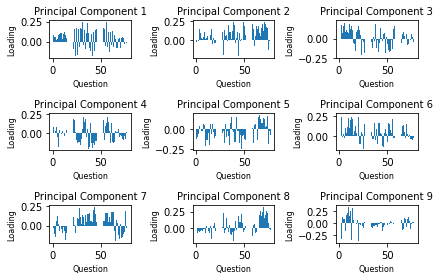
Lastly, a final function was called to conglomerate and print the arithmetic average of all 400 models (one on each movie), and obtained a grand r^2 score and cross-validated RMSE of:

Score: 0.010021745028667588, Cross Validated RMSE: 1.1190860185949998

**Question 10: 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.**

****

For this question, I at first attempted to complete a PCA on the given data. However, I computed the PCA and noticed the amount of variation explained by each, having none being greater than 10%. Alternatively, I decided to run a regression on each of the movies (as I did previously in question 8 and 9) against all 77 predictors of “all available factors.”

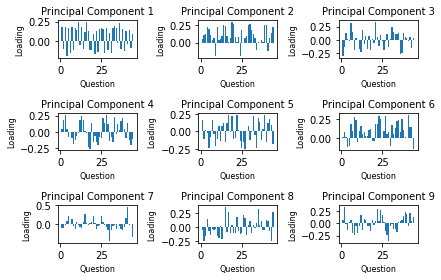
****Similar to the previous two questions, the XGBoost algorithm was implemented, due to the robustness of the model to sparse data. The regressor format of the module was called and was then iterated through each of the 400 movies, appending to a dictionary of movie names (as keys) and model itself, the r^2 score, and k-fold cross-validated RMSE (both as a mean of the 5 splits provided) as values.

* For the r^2 score, cross\_val\_score from the sklearn.model\_selection was utilized on all the training data, with the parameter scoring='r2'
* For the RMSE value, cross\_val\_score from the sklearn.model\_selection was utilized on all the training data, with the parameter scoring=' neg\_root\_mean\_squared\_error'

Lastly, a final function was called to conglomerate and print the arithmetic average of all 400 models (one on each movie), and obtained a grand r^2 score and cross-validated RMSE of:

Score: 0.17024514525726686, Cross Validated RMSE: 1.1913448102730877

**Question 11 (Extra credit): Tell us something interesting about this dataset that is not trivial and not already part of an answer (implied or explicitly) to these enumerated questions.**

I was really interested why there did not seem to be ***significant*** signs of“personality types” as suggested in Question 2. The only plausible answer was that the principal components (derived from the underlying questions), which are more so independent of one another, represent *independent* components of how a *personality* is characterized.

Inspecting all the loadings matrices for each individual principal component, we can see that:

* Chart, scatter chart

  Description automatically generatedPrincipal Component 1 relates *highly* to Extraversion
* Principal Component 2 relates *highly* to Neuroticism
* Principal Component 3 relates *highly* to Introversion

Then, we can arguably reason that, based on the results of the personality test, *extraversion,* and *introversion* **do not exist on the same spectrum**, and are rather independent factors. This hypothesis is further supported by the existence of a lack of dependance (correlation) between introversion and extraversion (shown right).

**Works Cited**

Badr, Will. “6 Different Ways to Compensate for Missing Data (Data Imputation with Examples).” *Medium*, Towards Data Science, 12 Jan. 2019, <https://towardsdatascience.com/6-different-ways-to-compensate-for-missing-values-data-imputation-with-examples-6022d9ca0779>.

**Appendix**

\*\*Please see the code attached to this submission to understand any of the outputs/interpretations\*\*