**Data Analysis Project 2**

**Data Cleaning**

I replaced the NaN value in dataframe with a blend(50/50) of the mean of each columns and each row. Then I export the imputed data to a new csv for further use; it saves my time when importing the data that has already been handled.

**Question 1:**

In a loop of 400\*399 times, I choose one movie and use the other 399 movies to create 399 linear regression model and then find the max cod in 399 movies in each movies of the 400 movies. The largest cod corresponds to the best predictor

The average COD of those 400 simple linear regression models is **0.42378171067196035**. The histogram of these 400 COD is below:

Chart, histogram

Description automatically generated

Tables of 10 movies that are most easily predicted (top 10 rows) and 10 movies(row10-19) that are hardest to predict with their cod and best predictors:

Table

Description automatically generated

**Question 2:**

In a loop of 20 times, I add features such as gender, sib, social and rate of best predictors to each movies (10 easiest predicted and 10 hardest predicted). Then I performed multiple regression models for each movie and get 20 new R2 score.

R2 doesn’t changed too much. Here is a scatter plot of the old R2(x axis) and new R2(y axis)

Chart, scatter chart

Description automatically generated

Text

Description automatically generatedText

Description automatically generated

**Question 3:**

After picking 30 movies with cod in the middle range(Y) and 10 movies randomly(X). I split training and testing data in 80/20. I performed 30 Ridge regression with Grid Search to find the best alpha. I set the possible parameter of alpha from 1 to 200 in integers and use “neg\_mean\_sqaured\_error” as scoring method in Grid Search.

Finally, I get the betas in each of these 30 models

In the table below, I have 30 movies used as Y and each model’s best alpha, RSME, and the coefficient beta of 10 movies(input)

Table

Description automatically generated

Table

Description automatically generated

**Question 4:**

Same as question 3 but this time I used lasso regression and set the alpha grid from 0.0001 to 0.1 with 1000 values in between. I got the table below:

Table

Description automatically generated

Table

Description automatically generated

**Question 5:**

I calculate each user's average movie enjoyment (using only real, non-imputed data). Put these averages in a logistic regression model as the predictor variable X. Then I pick 4 movies in the middle of the score range. They are **Fahrenheit 9/11 (2004), Happy Gilmore (1996), Diamonds are Forever (1971) and Scream (1996)** and label the rating data as enjoyed or not enjoyed according to the median. Then, for each of these four movies, I perform logistic regression model with grid search to find the penalty hyperparameter. The AUC score for these two movies are **[0.9545454545454546, 0.8584078119827872, 0.9540153833429824, 0.8590909090909089]** and betas are [**10.28532939, 5.96005022, 10.33151309, 4.97839277].** Logistic graph is the following:

Graphical user interface

Description automatically generated

Graphical user interface

Description automatically generated

**Extra Credit:**

Find if there is a linear correlation between if the viewer cry and the rating of the movies:

I choose if the viewer cry(rated from 1 to 5) as X and the average rating of the movie as Y. I split the test and train data in portion of 20% and 80%. Then I fit a linear regression model with cry and ratings and use test data to get R2 score. The r2\_score is -0.0227 which means the relationship with cry and movie ratings cannot modeled as a linear model. There are other model that works better.