ECE 20875 Final Project
Name: Nur Nadhira Aqilah Binti Mohd Shah
Purdue Username: mohdshah
Github Username: nadflop
Project Path 2
Date: 6/30/2020

The dataset I'm working on, which is contained in the 'behavior-performace.txt' is for an online course on how students watched videos and how they performed on in-video quizzes. In the dataset, the data is sorted by the studentID and videoID. The features that are considered for my analysis are:

fracSpent	Fraction of time student spent watching the video
fracComp	Fraction of the video the student watched
fracPaused	Fraction of time student spent paused on the video
numPauses	Number of times student paused the video
avgPBR	Average playback rate that student used while watching the video
numRWs	Number of times student rewind the video
numFFs	Number of times student fast forward the video

For each analysis, here's the number of students considered during the analysis:

Analysis	Total Number of Students	Criteria/Condition
Clustering	1535	Watched >= 5 videos
Predict Average Performance	96	Watched at least half of the videos
Predict Student's Performance based on video	3977	All

ANALYSIS CHOSEN FOR EACH QUESTION

Problem 1: K-Means Algorithm

I am using a multivariate k-means algorithm to analyze and cluster the student based on their video-watching behavior. I am using this method because it gives reliable results and the algorithm is a fast, robust and simple to implement. Since we have a lot of features to consider when clustering this dataset, using this algorithm allows us to group the data points into distinct non-overlapping subgroups. Besides, we can tell if the data is well clustered in two ways: 1) by looking at the distance of each point to its cluster 2) using the Elbow test where we are plotting the number of clusters to its squared sum error

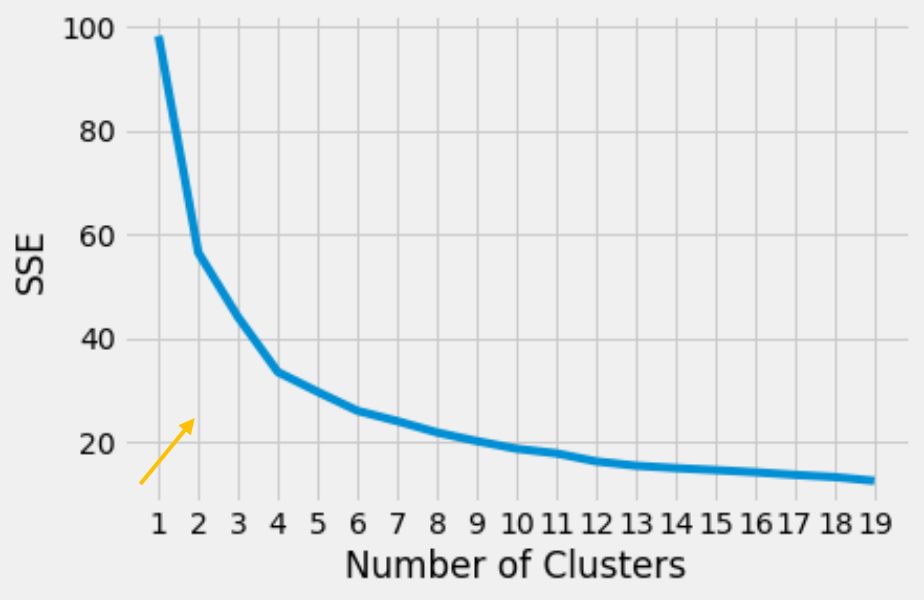
Problem 2: Ridge Regression

In this problem since we are required to predict the average performance of a student based on the student's behavior, I choose the ridge regression. Ridge regression guarantees the best model for predictions on unseen data. In the dataset given, we have a larger number of predictors (p) than number of observations (n) so using ridge regression is better since it avoids overfitting by adding penalty to models that have too large coefficients. Besides, this method also accounts for multiple predictors when predicting the outcome, which is returned in a continuous value. From here, we can compute the mean squared error and accuracy of the model more accurately and we can use these values to see how good the model is at predicting a student's performance.

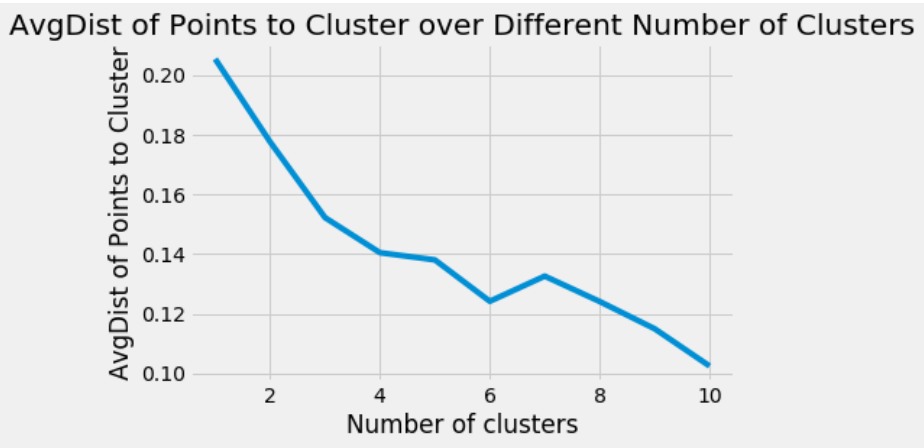
Problem 3: Ridge Regression

I am also using ridge regression to make predictions and the reasons are similar to the second problem. Yet, the difference between these two problems is that the data will be grouped based on the videoID instead of studentID. By doing this, we can analyze the individual training model per videoID which can tell us how well the student's behavior can be used to predict the performance for different videos.

RESULTS OF EACH ANALYSIS
Problem 1
Question: How well can the students be clustered by their behavior?



The graph above shows the number of clusters vs the sum of squared error/sum of squared distances. From this graph, we can conduct the Elbow Test to determine which is used to determine the *k* for our k-means test. From the graph, we can see that the 'elbow' is 6. To verify if the value of *k* is indeed true, we can run another test where we compute the average distance of points of each cluster over different number of clusters.



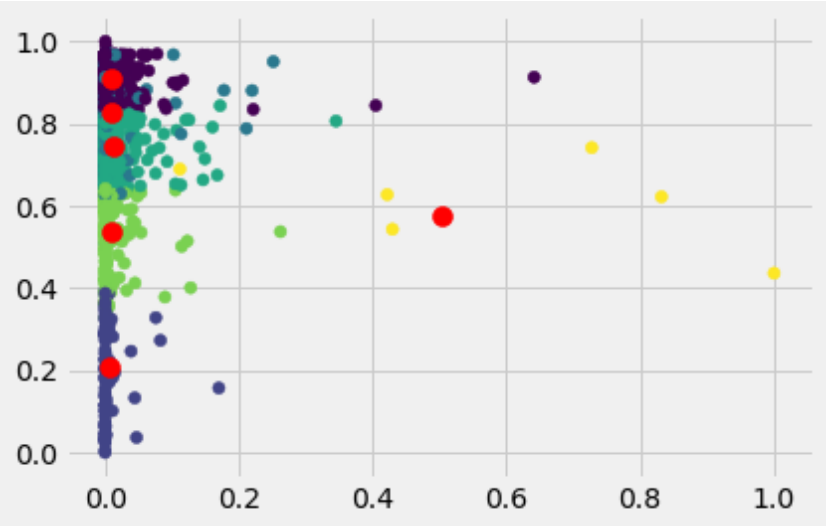
This graph shows the average distance of all points to its cluster. At *k*=6, we can see how the average distance is low and we can also see how there's a change in gradient of average distance after *k*=6. This might be because there are less centroids chosen in the dataset and all of the points are being over-clustered.

I also printed the clusters and its points to verify if that's really the case:

```
[Cluster: 202 points and center = [0.0103454637804909, 0.8317628075392793, 0.007633759216855106,
0.006082402031496612, 0.7881969330893543, 0.005317167132771142, 0.021744548201995638]
 Cluster:  523  points  and  center  =  [0.006980433581101852,  0.918847143520079,
0.008262593395833503,  0.0015385518406064694,  0.5252085780399622,  0.0030879499672826533,
0.009291681844796743]
 Cluster:  39  points  and  center  =  [0.008722112012289722,  0.11702122291714151,
0.014300912187981272,  0.0009556262483283687,  0.4966719264105385,  0.005129336095055687,
0.0391146189251841]
 Cluster:  392  points  and  center  =  [0.011676587797159157,  0.7754736557208437,
0.011445509943106452,  0.0020056656550147206,  0.5132731962642735,  0.008313355684237606,
0.026625511460403203]
 Cluster: 31 points and center = [0.00717430134106923, 0.5769447261429842, 0.010310109673514778,
0.0018594881853444712, 0.5884369295339135, 0.022795785867705934, 0.3301095110159342]
 Cluster:  235  points  and  center  =  [0.025426826105011836,  0.6130236484219402,
0.03254229059000822,  0.0018208113514683313,  0.5046435116464587,  0.017488910126997964,
0.037654896980504594]
 Cluster: 86 points and center = [0.00760675059721079, 0.40380599638755615, 0.01760481637375734,
0.0015796130075975657, 0.49343785910513965, 0.015531229553661265, 0.057743607742102174]
 Cluster:  23  points  and  center  =  [0.004694639290504925,  0.27905731414910323,
0.00649497061517761,  0.0009690407742771781,  0.3066009211687411,  0.0030970393646166552,
0.02271164440660757]
 Cluster:  3  points  and  center  =  [0.0005697944314155415,  0.05049570058348304,
0.00081523293010176,  0.00019746843809901517,  0.12183435228633333,  0.0003010752688172043,
0.002336448598130841]
 Cluster: 1 points and center = [0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0]]
```

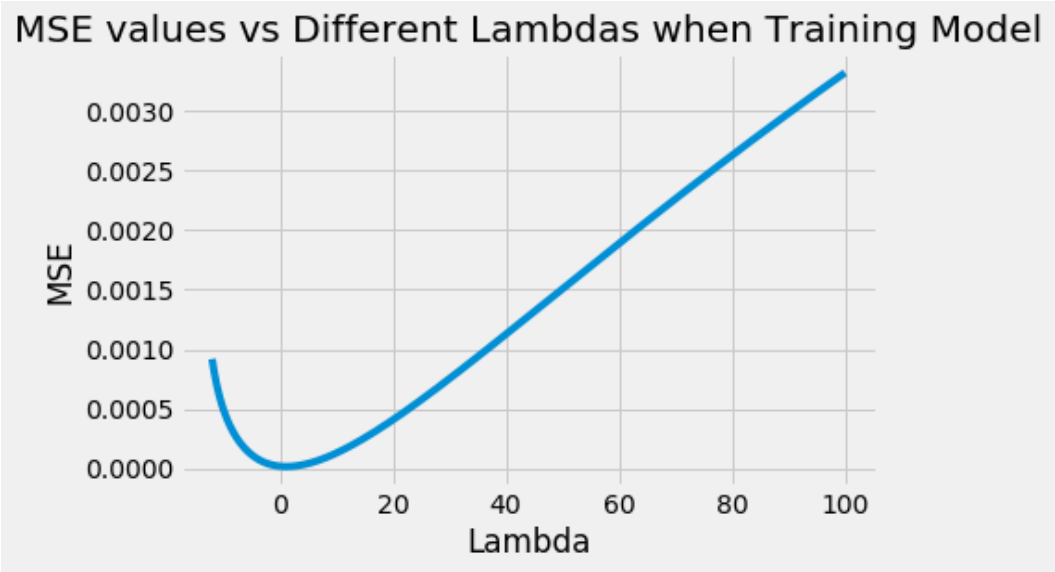
We can conclude that the number of clusters that best represent the data is really 6, with the average distance of points to clusters being around 0.12414156582672652.

I also tried plotting the dataset on a scatterplot using the library in sklearn and we can see how the centers chosen are quite near to each other, but nevertheless the algorithm still managed to cluster the dataset.



Problem 2

Question: Can student's video-watching behavior be used to predict a student's performance (i.e., average score s across all quizzes)?



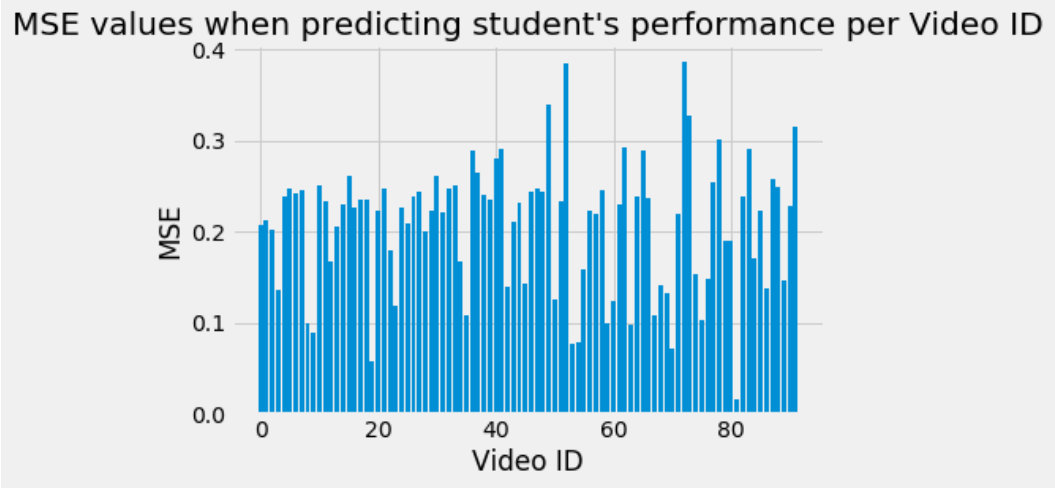

```
Number of Training Data: 96
Number of Testing Data:9
Threshold Value: 0.2
Model Accuracy: 88.89 %
Model   Coefficients:   [[-2.97423909e-04   -1.51506971e-04   2.31026746e-04
7.21512232e-04 4.47641269e-04 1.11853422e-03 -1.42557738e-03 1.59736245e-01]]
MSE: 0.00
R^2: 0.05
Best Lambda Value: 1.4791083881682072
```

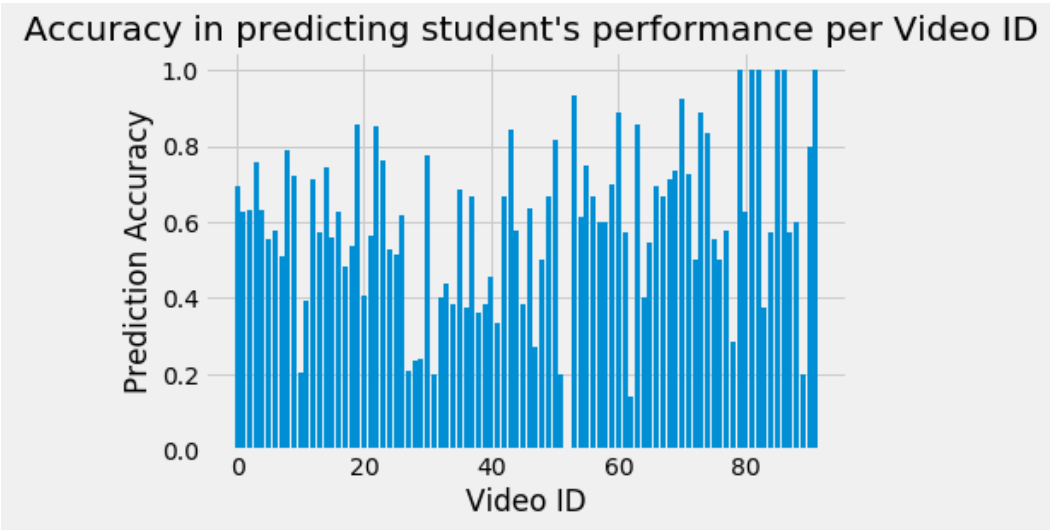
In this problem, the data was split into the training and testing data with a ratio of 9:1. As we can see from the output generated by the program, the accuracy is around 88.89% which is quite high and also indicated that the model is accurate in predicting the performance of student based on the behavior. Moreover, we can also see how the best lambda value for the model is almost near to 1, which is our goal in any ridge regression analysis.

The R-squared value of 5% means that the model doesn’t really explain the variations in the response variable around its mean, which says that the regression model doesn’t really fit the observations, which if we think about it makes sense since the data being analyzed is related to human behavior which is less predictable and there’s always other factors that contributes to the observation seen.

Problem 3

Question: How well can you predict a student's performance on a particular in-video quiz question (i.e., whether they will be correct or incorrect) based on their video-watching behaviors while watching the corresponding video?





The graphs above shows how accurate student's watching behavior can be used to predict their performance for each video.

From the graph, we can conclude that videos [19, 22, 43, 50, 53, 60, 63, 70, 73, 74, 79, 81, 82, 85, 86, 90, 91]have an accuracy rate over 80% which means that these videos can be used to predict student's performance.

For videos [0, 1, 2, 3, 4, 8, 9, 12, 14, 16, 23, 26, 30, 35, 37, 42, 46, 49, 54, 55, 56, 59, 66, 67, 68, 69, 71, 80], their accuracy rate is

between 60-80% which means that the student's performance can be fairly predicted.

Other videos might have a smaller prediction accuracy since it is actually difficult to predict a student's performance through their video watching behavior alone. Besides, we should also consider other factors that can also contribute to a student's performance in a quiz, such as a student might have a better background knowledge prior to watching the video or the video itself could also not cover all the necessary information needed to answer the quiz.

PROJECT CONCLUSION

Based on the analysis done, we can say that the students can be clustered well based on the featured behavior and the best result is achieved when $k = 6$. This is verified by the elbow test and the average distance of the points with their respective clusters, which is 0.12414156582672652.

Furthermore, we can also see how student's video watching behavior can be used to predict their performance in a quiz, where the accuracy is 88.89% using the given data and trained model.

Yet, we can't really do the same when predicting a student's performance per video, since the accuracy percentage isn't that consistent. On top of that, there are only a few videos that have an accuracy more than 80%, so if someone needs to do a prediction based on a video, they should only consider videos that have an accuracy rate more than 80%.