Process book Data Processing final project

November 4
Before today I had another project idea which was related to the game of life, however it was rejected because it did not include a data set. I searched for data sets until I found one I was interested in. I proposed another project idea where I would do linear regressions to predict a show's popularity. This idea was approved

November 13
I loaded libraries that I may need for the project, such as pandas, matplotlib, numpy and sklearn. I loaded the data files with pandas and wrote the code for the first plotting

November 26
Today I created the linear regression model for score and popularity, I fit the model and created predictions. Then I plotted it and computed the mse and r2 as statistics to evaluate the model. I already did a linear regression as part of the transformations module, so to learn something new I decided to split the data for train and test. 20% of the data will be using for testing the linear regression model, while the rest will be used for training the model.

Some resources that I used
- https://www.geeksforgeeks.org/machine-learning/how-to-split-a-dataset-into-train-and-test-sets-using-python/
- https://www.geeksforgeeks.org/python/what-is-the-difference-between-transform-and-fit_transform-in-sklearn-python/

December 5
I had the deadlines mixed up and started the project earlier than I was supposed to. Today in class after finishing all other assignments I discussed with the professor my idea further. The professor suggested me to add some other elements to my project if I had the opportunity, such as a decision tree, and doing a linear regression with two variables (scores and episodes in one linear regression).

I also calculated the Standard Error of the slope. To remember how to calculate it, I referred to materials that I had in another course and wrote the formula manually in python.

December 6
Today I started to work on the linear regression for genres and popularity. TF-IDF is something I've never done before, and it was suggested to me by the professor when I proposed my project. Before starting I read the documentation for the function TfidVectorizer and I watched some YouTube video tutorials to understand TF-IDF better. After watching the video tutorials, I did the linear regression as I did for popularity and score, however I could not plot the linear regression for genres and popularity. This was because linear regression only works with 2 features, while the TF-IDF matrix produced

95 features. Because of this, I decided to discard plotting the linear regression for genres and popularity and do clustering and principal component analysis to visualize the data.

I also revised and looked at clustering and PCA tutorials, to expand on the brief introduction I had in another course and to learn how to code it in python.

These are some of the resources that helped:

- PCA explanation: https://www.youtube.com/watch?v=g-Hb26agBFg
- How to do PCA in python: https://www.youtube.com/watch?v=8klqIM9UvAc
- Clustering: https://www.youtube.com/watch?v=4b5d3muPQmA
- Plus more resources from my past course
- https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

December 7

Today I worked more on the linear regression for genres and popularity and run a clustering analysis. I created a clustering Kmeans model and did an elbow plot. Based on the elbow plot I picked 3 clusters. I fit the data to the Kmeans model and got the average popularity per cluster. Doing a clustering analysis to create a plot still was not enough, as the data was 95 dimensions and it needs to be 2 to be able to plot it. So I used PCA to make the clusters plottable. I learned how to do a PCA and clustering analysis in python.

Some resources that helped me:

- https://www.geeksforgeeks.org/machine-learning/k-means-clustering-introduction/
- https://www.w3schools.com/python/python_ml_k-means.asp
- https://numpy.org/doc/stable/user/basics.indexing.html

December 8

Today I started writing on a new document the interpretation for the score and popularity model but I struggled to interpret the linear regression and graph. The graph had a huge variance, but the standard error was small. The R^2 indicated a very poor performance of the model (the model cannot be used to predict popularity) but the p-value was significant, meaning that there was a relationship between score and popularity. I showed the graph and statistical results to a friend (who has experience in statistics) and he told me that while there was a relationship between the variables, the model could still fail. His advice allowed me a better understanding of the data and to interpret the linear regression plot.

I also had the proof of concept meeting with the professor in class. I showed the professor my code so far for the linear regressions and I explained that my model failed. She told me that I was on good track and to do another linear regression with 2 variables and to maybe add a decision tree. She also told me to write a report where I explained my analysis in detail, and she advised me to focus more in depth on the concepts that I'm

working with currently instead of adding more features and analysis. I decided to also place significant focus on explaining my code and theory in the report.

I transformed the popularity variable so that it would have a smaller range as advised by the professor.

December 9
Today I started writing the report for the project, I explained the theory and my results for the linear regression for score and popularity.

I also wanted to measure to run a t-test for the linear regression for popularity and genres, however since the linear regression has 95 coefficients I could not. I tried to research other ways to do a linear regression but the only other solution would have been to do it for every coefficient. I decided to exclude the t-test from this analysis and focus on more interpretable methods for this multi-dimensional data.

As an alternative test, did an anova test to determine whether the means of the clusters were significantly different.

I also wanted to know which cluster represented which genre (previously the clusters were labeled 0, 1 and 2). I also made it so the clusters would be visible in the diagram so that it would be possible to know which cluster is which only by looking at the diagram.
I used the coordinates of the cluster centers to assign the label. I looped through each cluster, got the cluster centers and added them to the diagram

Helpful resources:
- https://www.geeksforgeeks.org/python/how-to-perform-a-one-way-anova-in-python/
- https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

December 10
I wrote more for the project report. I wrote the part for the linear regression for score and genres.

I did the regression for score, number of episodes and popularity. Since there are 2 predictors I cannot create a plot for the linear regression. I initially planned to try to create a 3d plot for the linear regression, but because of the lack of time I decided to only do a decision tree to visualize the data.

December 11
My original plan was to calculate the standard error and do a t-test for all the models and use those statistics to compare the models to each other. However, since I excluded the t test and standard error for the linear regressions with more than 1 predictor, I decided to use the adjusted R^2 to compare the different models.

I also created a decision tree for the linear regression of popularity and score. I researched how decision trees worked and their documentation to learn how to interpret them and how to do them in python

Some resources that helped:

- https://www.youtube.com/watch?v=LDRbO9a6XPU
- https://scikit-learn.org/stable/modules/tree.html

December 12

Today I created a decision tree for also the last linear regression. This regression tree was different as I had to learn to interpret the decision tree even if there were different variables in the nodes. I added some last comments to the code and finished writing my report. I decided to not add a decision tree for the linear regression of genres and popularity since the many variables in this linear regression would be very hard to interpret.

Personal notes:

I initially planned for this project to be simpler than it was and to simply learn how to run statistical analysis I was already a bit familiar with, but in python. However, I encountered challenges from the structure of the TF-IDF matrix, which did not allow me to do the analysis as I originally intended. This unpredicted structural issue forced me to run additional analysis to visualize and interpret the data. The linear regression I spent the most amount of time on was the one for genres and popularity. This unpredicted challenge however was positive as not only I learned how to do t-tests and TF-IDF, but also to review old statistical methods I studied in the past and do them in python. Decision trees were also something new I learned. While decision trees are not a complex method, I was surprised by how much interpretation they allow for the data.