Onboarding Journal - Kevin Zhou, Greg Bodik, Marvin Yang

Week of March 1 (Meeting Feb 29, Workshop Webscraping)
Objectives: Come up with three questions of interest and rank them. Gather a dataset for at least the question you are most interested in.
Progress:

- Kevin: Looked at possible project ideas such as potential prediction projects based on sports and outcomes of games, yelp reviews, etc
- Marvin: Explored possible data sets for the semester project, and focused on something like Yelp reviews. Considered focusing project on subset of Yelp reviews such as restaurants only.
- Greg: Started brainstorming possible project ideas and coming up with questions of interest.
    - Possible ideas: online shopping patters, ranking systems (ex. yelp), spotify and duolingo datasets
    - Possible Questions:
        - What patterns exist in online shopping patterns? [Online shopping data set]
        - How do websites rank their products? [Reddit data set]
        - Something to do with dialogue between opponents in sports? [Tennis dataset]
    - Consolidated team ideas into powerpoint for meeting.

Week of March 8 (Meeting March 7, Workshop EDA and feature selection)
Objectives:

- Kevin:
    - Try to see if the restaurant data is viable
    - Come up with central questions about the dataset such as "what characteristics can be used to predict good Yelp reviews"
- Marvin:
    - Look at the Yelp data set and see if it is viable
- Greg:
    - Drop all non-restaurant features from yelp data, then inner join all the datasets.

    - Start EDA to find some areas of interest in the data
    - Re-evaluate our question, potentially "What gives restaurants positive reviews..."

Progress:

- Kevin:
    - Tried to create some visualizations on the Yelp data (bar graphs) to pique interest but they weren't very effective and we decided to scrap the Yelp dataset idea
- Marvin:
    - Other team members found working with the data set to be too slow, so went to look for another data set on kaggle. Looked at some popular ones but the team eventually decided on a movie metadata set.
- Greg:
    - Tried working with Yelp data, too slow 😔 started looking for new data sets
    - Found some more data on store sales (may to compare to online ones, college acceptances, others.
    - Started working on the movie data set.
        - Looked at distributions of countries represented and removed non-US films
        - Looked at distributions of movie earnings, years of movie release, duration
        - Plotted basic correlation matrix to start identifying features.

Week of March 15 (Meeting on March 14)

Objectives: Create normalized features for the data

- Kevin: Make visualizations to look for which features could be most related
- Marvin: Explore the data set for relevant features
- Greg: Find features that account for most variance, explore data more

Progress:

- Kevin:
    - Decided that gross revenue as our dependent variable could make sense
    - Plotted features against each other, mainly with scatter plots
    - Started to look at linear regression models with these different features using code from the INFO 1998 class that I was in the previous semester, but most of these ended up not being too useful because they were done without the focused goal of looking at how characteristics impacted gross revenue
- Marvin:
    - Created visualizations using pyplot to see if any features are obviously related

    - Used scatterplots and investigated duration, and various facebook likes vs gross
    - Confirmed findings by using sklearn.selectkbest to select the 10 best features based on their chi^2 test score the following were the most relevant:
        - Budget
        - Number of Reviews
        - Number of Likes
- Greg:
    - We determined that we wanted to investigate gross revenue of films further.
    - Found content_rating and imdb score to reliably relate to revenue, made associated boxplots and scatter plots.
    - Also found relationship between budget, num voted users.
    - Used chi^2 to confirm features (Marvin also did this independently)
    - Isolated all the features we identified into a separate data set, performed z normalization on all of them.
    - Features identified: We have for sure:
        1. content_rating
        2. imdb_score
        3. budget
        4. num_voted_users

        from the chi-squared test:

        5. movie_facebook_likes
        6. cast_total_facebook_likes

        super maybe from random graphs:

        7. num_critic_reviews_for
    - Created slides to display results.

Week of March 29 (Meeting March 28, Workshop: D3)
Objectives: Create an interactive D3 vis for the data
Progress: We all discussed possible visualizations and settled on a boxplot of the distribution of movie revenues of a certain content rating. When that rating is changed, the boxplot will expand or contract.

- Kevin:
    - Fixed the math in the code that was used to calculate the quartiles for the boxplots
    - Decided to make the boxplots' axes consistent across the different content ratings so that we could better visually compare the boxplots against each other from G to PG to PG-13 to R without having to reorient ourselves to a new axis

    - Changed the values for the axis so that it fit better with our data
    - Fixed formatting of code to make it more readable and streamline
- Marvin:
    - Read up online and looked at a lot of examples to try and figure out how to make the d3 visualization the team wanted
    - Found out about drop down menus and proceeded to make a visualization that would change the box plot displayed based on a drop down menu that selected the content rating (the boxplots were against gross)
    - Lots of referencing others code and changing to fit my needs
- Greg:
    - Isolated the revenue values of each movie for each content rating for use in d3.
    - Added jitter of points to Marvin's boxplot
    - Added slider to control jitter values.
        - Moving the slider adjusts the bound that a random number [0,1] is multiplied by to get the displacement from the central x-axis for each point.
        - Since we took the approach of having four separate boxplots and a jitter slider to spread out the points, Eric recommends adding hover over features so that you can get a better idea of what the actual revenue number for a point is.
        - Will implement this in the future.

Week of April 5 (Meeting on April 4, Workshop: Supervised ML)
Objectives: Define a clear prediction problem from your dataset and investigate using one of the ML models discussed.
Progress:

- Kevin:
    - Created several regression models this time focused on comparing characteristics to gross revenue
        - Made models for IMDB rating, total Facebook likes, cast's total Facebook likes, budget, number of users who voted, and number of critic reviews
    - Tried to see if polynomial regression models fit the data with different characteristics, but they weren't much better than the linear models if at all better
        - (tried degrees ranging from 2 to 5)
- Marvin:
    - Wanted to use regression to find relationship between features and gross

    - Looked at and created some linear regression using what we learned in the workshop
- Greg:
    - Focused on how/if a movie's content rating could be accurately predicted by the other features identified.
    - Trained KNN to predict content rating using all data and just the features identified, max accuracy was only 0.51.
    - Made multiple linear regression to predict gross revenue from the other features (r^2 = 0.516)
    - Used decisionTreeRegressor and RandomForestRegressor to try to predict revenue. Default decision tree was overfitting a lot. Decision tree had a slightly better test accuracy of 0.6 with MSE 2.01
    - (Need to go back and optimize these models later)
    - Corresponding slides

Week of April 12 (Meeting on April 11, Workshop: Unsupervised ML)
Objectives: Apply an unsupervised learning algorithm to your data
Progress:

- Kevin:
    - Tried to make an elbow plot for kmeans clustering but there were some errors in my code and we ended up going with Greg's
- Marvin:
    - Took what we learned in the workshop and explored the data a little on my own (not too sure what I was doing honestly). I'm not sure how significant my results were but used some of the functions from the workshop
- Greg:
    - Used KElbowVisualizer from yellowbrick.cluster to generate elbow plot for kmeans clustering
    - Used kmeans with 3 clusters
    - Used cluster_centers to inspect centroid locations
    - clusters do not seem disparate
    - UPDATE: see week of 5/3 for more detailed analysis/explanation of the clustering I did at a later date.

Week of April 19 (Meeting on April 18, Workshop: NLP)
Objectives: Add to the project somehow (NLP isn't applicable to everyone)
Progress:

- Kevin:
    - Fixed some visualization errors with the regression models
        - This was mostly fixing axis titles and colors

        - I also had the code print out the test and train scores as well as the slope and intercepts of the lines
- Greg:
    - Made a d3 histogram to visualize the distribution of movie revenues
        - Number of bins can be adjusted to variably display the data
        - Matched styled to previous d3
    - Added hover over to display to revenue values from previous boxplot
        - (See Week of March 29 above)
    - Made style consistent across both plots.

Week of April 26 (Meeting on April 25)
Objectives:
Progress:

- Kevin:
    - Looked at previous visualizations to see what could be used and prepare for the presentation
    - Started making heat maps with different features such as imdb score and gross revenue of curiosity as to what the more common scores for certain revenues was and also a more clear visual way of seeing this
        - Made them with the scaled values though, so it didn't have as much of a real world message (more on this in the next week entry)
- Marvin:
    - Started reviewing old code considering what could be used for the presentation
- Greg:
    - Started trying to optimize depth of decision tree classifier to predict content rating
    - Plotted testings scores, training scores, and training - testing to find optimal depth
    - (Should probably also do k-fold cv to investigate this further)

Week of May 3 (Meeting on May 2)
Objectives: Amalgamate work thus far into presentation for Wed 5/6
Progress:

- Kevin:
    - Worked on the presentation by adding visualizations, models, and write ups
    - Made a correlation matrix with all of the selected features to use in the feature selection portion of the presentation

    - Made individual linear regression models for movies after 2010 so we could compare them to the linear regression models for those features with the overall data (the features used here were imdb rating, budget, the number of users who voted, and the cast's total Facebook likes
    - Made the heat maps using the data before it was scaled, giving a more direct translation to real world information. For example, you could see through the heap map for the imdb ratings that the most common data was between an imdb rating of 6 and 7 with a gross revenue under 100,000,000. The issue with the heat maps is that there's too many movies that gross under 100,000,000, which is pretty low on our y-axis but is where much of the movies lie (as can be seen in our histograms), making it hard for the blocks above that to be of relevance in the heat map. However, this fact helps to bolster what was shown in the histogram and gives some more real world contextualization to the project
- Marvin:
    - Reviewed old code and visualizations and chose what would be good for the presentation
    - Edited code for visualizations of what we did for feature selection, mostly clarifying scatter plots and changing the alpha values.
    - Created slides for data source and features selection
- Greg:
    - Made detailing the data preprocessing we performed, our motivations in doing the project, the central questions we derived from this process.
    - Made slides for all of the multiparameter ML models and provided interpretation for each.
    - Re-segmented the data in all-film and only new-film categories, and re-performed all existing analyses on these new data.
    - Regular train-test-split showed strangely higher number of optimal neighbors for knn.
        - Added 10-fold cv to make sure this is accurate (it was). Optimal k is ~35-40 for both partitions of the data
    - Previous kmeans clustering was hard to interpret because we based our judgment of success based on the geometry of cluster centers.
        - kmeans .score() function gave a value large in magnitude (relative to normalized data input) so we interpreted this to mean that on average, points were very far from their cluster centroids, which would mean the data isn't cluster able.

        - This isn't necessarily the case. Although our clustering wasn't really "high dimensional," involving only 8 dimensions instead of a potential 100, *A Few Useful Things to Know about Machine Learning* (2012) doi:10.1145/2347736.2347755, states "In high dimensions, most of the mass of a multivariate Gaussian distribution is not near the mean, but in an increasingly distant “shell” around it; and most of the volume of a highdimensional orange is in the skin, not the pulp."
        - Thus our high score for the kmeans objective makes relative sense and doesn't mean that clustering won't work.
        - On the same k-means clustering with 3 general clusters (as per the elbow plot) I did another analysis where I just looked at the distribution of movie gross revenue between the 3 clusters. When looking at all the films and just the new films subset, there emerged a cluster with a median gross rev 1.5 sds above the mean, and 2 clusters with median gross rev ~1 sd below the mean of the data.
    - Added interpretation of each step described above to the presentation and provided conclusions at the end.
    - See slides here

Week of May 10 (Meeting on May 9, Last meeting of the semester)
Objectives: Prepare for presentation on 5/13, keeping in mind suggestions made by Prof. Rzeszotarski
Progress:

- Kevin:
    - Started thinking about what potential next steps could be, such as potentially finding a more effective cutoff year than 2010 for our models. The impetus for this is mainly just because 2010 was an arbitrary selection but done with the thought of accounting for the fact that things like Facebook are more prevalent in recent times than they would've been before
    - Looked into potentially putting the information in our presentation onto a website, but we decided that that wasn't necessary and that our project works well in presentation form
- Marvin:
- Greg:
    - We were cautioned about the risk of not accounting for inflation when analyzing gross revenue values for movies made across a span of many years.
        - To investigate the potential effects of this I plotted the distribution of film revenues for each year individually.
        - As shown below they are relatively consistent.

        - Since the data comes from IMDb I looked up how they obtain their revenue data.
            - According to IMDb: "...gross figures are cumulative (i.e. they reflect the total amount of money grossed by a movie up to a certain date in the specified territory)"
            - That ^ site links to Box Office Mojo, a revenue reporting service made by the professional version of IMDb — IMDb pro. According to the FAQ page of Box Office Mojo, when calculating gross revenue, inflation in movie ticket prices over time is a factor. More specifically, the "'*Top Lifetime Adjusted Grosses' chart adjusts for ticket price inflation using estimated number of tickets sold*" although Box Office Mojo does stipulate that "*Adjusting for ticket price inflation is not an exact science and should be used for a general idea of what a movie might have made if released in a different year, assuming it sold the same number of tickets.*"
            - This means we can be relatively confident in the validity of comparing gross between movies made in different years because they were all cumulatively amassed and their revenues adjusted to the currency value of the day on which the data were aggregated. Of course this rests upon the original data set being amassed at approximately one time point (which is reasonable to assume).
            - Despite this, if I were to do the same analysis again, instead of doing a standard Z-normalization on the column of all gross revenues, I might individually normalize each subset of revenues from a particular year. That way each normalized value would reflect how well a movie grossed only in relation to those released in the same year.
        - Looking at the IMDb site I also saw some additional metrics pro users (industry professionals) can access including more granular information about revenue such as opening weekend rev, first month etc. It would be interesting to see how well initial reactions predict long term film success.
    - We were also warned about the difficulty of drawing solid conclusions when using social media metrics for platforms that have not come into widespread popularity until recently. This is indeed probably the biggest limitation in the data we have and the approach we took to analyze it. It's not fair use Facebook likes (for example) to predict revenue of films released in 1995 because those after-the-fact likes are probably coming from a small subset of die-hard fans that are still watching and engaging with the movie. Another way to form a model would be to have the coefficient given to Facebook likes (and other

such metrics) be a function of when that movie was released, to try to account for this difference.

- Some other limitations/inconsistencies I thought of:
  - People don't go to movie theaters as much anymore
  - Data amassed over time is not a good indicator of initial public response
  - many more
- Wrote final write up and made/inserted associated figures.