**Angel Robinson Capstone Project Deliverable**

In this project, I was asked to help two Data Analyst answer questions about business trends and derive recommendations for better performance. The Movies Data Set is originally comprised of 4803 rows and 20 columns. After cleaning, 3229 rows and 16 columns remained. The data contained genre names, popular production companies, movie titles along with their budget and revenue. I was asked to discover what makes a movie successful at the box office, which genres are most favored, as well as how budget affects a movie’s popularity. The Bank Churn Data Set contained 10127 rows and 20 columns of data. This data set took a dive into customer age, marital status, utilization ratio as well as their attrition level. I was required to analyze the data to better understand what kinds of customers attrite their membership and why. I utilized Python to remove outliers and further rid the data of null values. Following data cleaning, 8875 rows × 20 columns remained, and visualizations were produced to further examine ways to reduce customer loss.

To begin my descriptive analysis of the Bank Churn Data Set, I imported the Pandas Library and used multiple built-in functions such as the describe method to provide the central tendencies for qualifying nominal columns. A column with numerical data and at least a maximum and minimum value are considered qualifying columns for data analysis. The drop method was used to prevent certain columns from displaying in this data frame. I then identified outliers using a boxplot and nested for loop to extract the following measures: Maximum, Minimum, IQR, Q1, Q3, and Outliers. Next, I extracted the outliers by index and proceeded to use the drop method to remove them from the dataset. Following the removal of outliers, I used the dropna() method to drop the remaining values that were null in the dataset. These values were not detrimental to the data, so they were safely excluded.

In the data cleaning and preparation phase, I used multiple cross tabs as well as scatterplots to demonstrate the relationship between across many variables. For instance, in the Bank Churn Set, I implemented a crosstab of the Attrition Level and Education Level. During my analysis of the Movies Data Set, I realized that further investigation was required to make accurate recommendations. Showcasing the relationship between Popularity and Budget was not enough. I explored the relationship between Popularity and Runtime, as well as Revenue and Budget. These extra steps helped me better understand the trends and point the companies in a clearer direction.

Upon my analysis of the Movies Data Set, I can infer movies with a lower budget, generate less revenue. The correlation between popular movies and budget demonstrates how much a low budget can affect a movie’s success. Furthermore, certain genres have been proven to be more successful in the box office. This is evident in the analysis of the Top 10 movies that have produced the highest revenue. 80% of these movies were apart of the Action genres. Of the movies with the lowest budgets, Comedy and Drama were the top two most occurring genres. Production Companies should widen their budget for said genres and recalibrate their marketing tactics surrounding these genres. I would also advise companies to create movies with a Runtime of around 150 minutes. The correlation between Popularity and Runtime revealed that movies approximate to this Runtime are most popular.

After assessing the correlation between variables in the Bank Churn Data, the focus of this company should shift towards women, married couples, and low-income families across all regions. When analyzing the correlation between gender and attrition, Women account for 62% of attrition which should be addressed. Over 43% of attrite customers earn less than 40k a year. Married couples account for 24% of past customers. To prevent further loss, the attention must be on these groups. The company should also implement better marketing techniques in the Wales region. This area accounts for the highest number of attrite customers, however it only accounts for 10% of the customers across all 4 regions. There are many opportunities to reverse customer loss by understanding the history and dynamics of the regions and customers they serve.