CS210 Final Report Project
Group Members: Jonas Azhak, Paolo Gervasoni, Armaan Saleem

Predicting Legal Case Outcomes Based on Attorney Factors

Project Definition:

This project aims to solve the problem of getting a more data-based approach on predicting the outcome of a legal case based on observable factors from an attorney, like the amount of experience they have (years), their win rate prior to the case, and their education level (JSD, LLM, or JD). However, because of the potential legal concerns of actually acquiring this data, the approach for getting data was to simulate a dataset, which amounted to 200 attorneys which all had random years of experience, win rate, and education level.

The approach for visualizing the data and actually using it was 2 main things, which are an exploratory data analysis and a logistic regression model. The dataset simulation was also made into a CSV file, however the CSV file acted as a data medium which turned into the EDA and logistic model, both of which provide better visualizations and opportunities to see graphs and more accurately predicted outcomes respectively.

Some previously scrapped approaches, which involved using more advanced techniques and models, were scrapped and instead we opted to strictly use the EDA and logistic regression model to more closely align with what this course taught us and be more detailed with the fewer models we did use rather than spreading broadly across multiple. These models are sufficient in showing how data can be made, used, processed, and also interpreted for a final review.(Changes made from the project proposal).

Novelty & Importance:

This project matters because it provides a quantifiable way of determining whether or not an attorney would win a case (as well as the probability that they would via logistic regression) and provides a valuable opportunity for clients to get the attorney they think is right for them while also allowing attorneys to market themself according to their own value.

The project was also naturally exciting, because ironically the lack of accessibility to real life data instead allowed us to see the project in its very infancy from its initial simulated data into graphs and models which could more accurately predict the outcome of a case based on the factors we chose. It instilled a sense of responsibility and excitement which allowed us to add and remove features if we pleased to see how an outcome would be affected.

Naturally as mentioned before, the factors we chose are limited in the interest of time and lowering unnecessary complications. In general however, these datasets in question are hardly accessible unless directly involved in the case, which limits the data you can realistically use and shows a flaw in approaching this with real data as there are immense confidentiality issues which need to be rectified first in order to actually initialize even an infantile version of a data-approached setup to predicting case outcomes.

Previous attempts and works exist which dictate similar things to our project, however because of the aforementioned legal issues that come with trying to access data like this, it is difficult to see a large-scale example of our project which doesn't cross any legal boundaries. In light of this, we hope our simplified and less overreaching approach of organizing and presenting data can provide a beacon of light or act as a stepping stone for future projects or data-based and more quantifiable approaches in the future of the legal world.

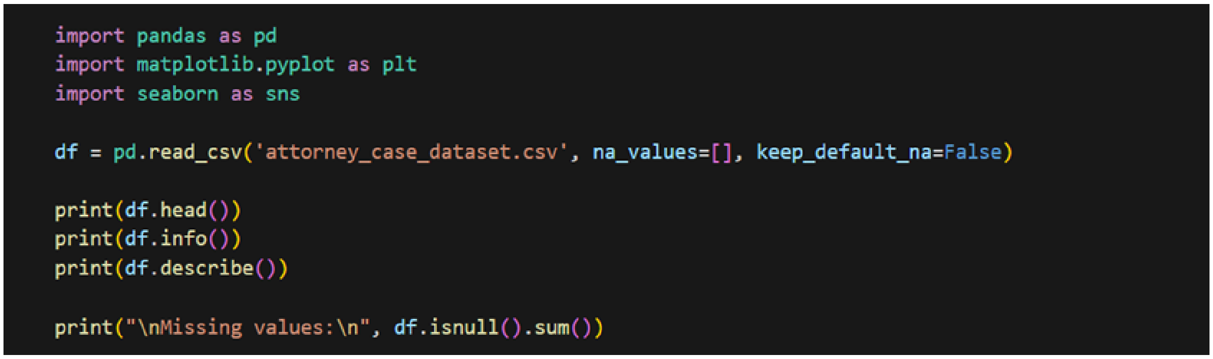
Progress and Contribution:

The data we used was simulated, because of the legal issues that would come with trying to accrue actual data, although it may have been feasible at some point, it was simulated regardless in case getting actual data would take too much time before we could work on our project. The factors within this simulated data are Experience_Years represented from 0-30 years, Win_Loss_Ratio from .2-.9, and Education_Level from JSD (Doctor of Juridical Science), LLM (Master of Laws), JD (Juris Doctor).

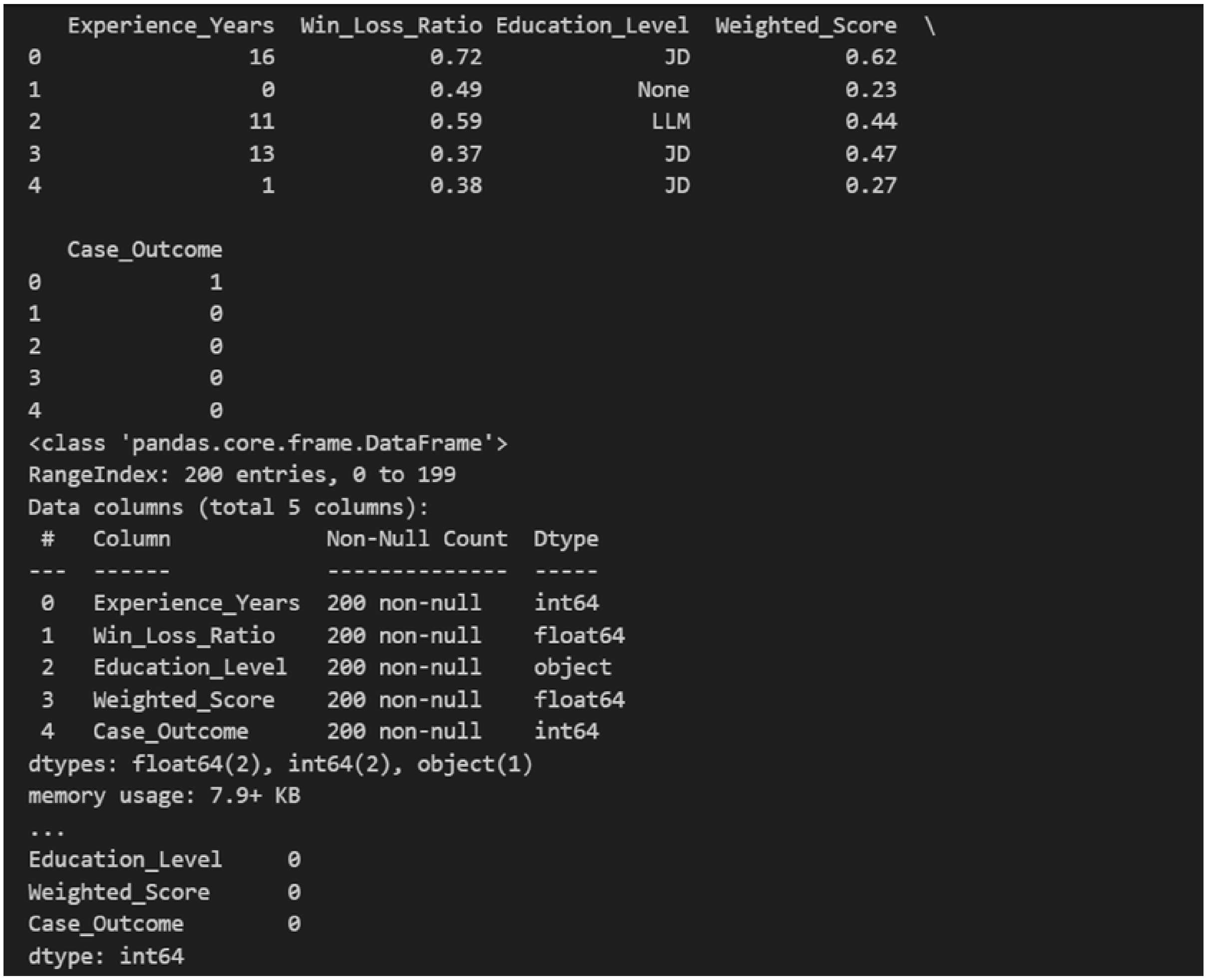
The data was created/simulated via NumPy and Pandas using Jupyter Notebook as the medium. From the factors we used, we made a weighting system which essentially weighed the factors on importance (In order: experience, win rate, education level) and would have the more important factors have more 'pull' on the final score, which would be checked against a value (0.6). If the overall score (designated by variable 'score') was larger than .6, then the last column in our CSV (Case_Results) would show a 1 representing that the attorney would win the case. This was our most binary and basic approach to representing this data in use before working with logistic regression and does not represent our final result, as the logistic model would instead work with probability and would use our simulated data to find the most optimal set of values from the factors to win while also being realistically feasible (not having the highest value in each category).

The techniques, models, and files used were as follows: NumPy and Pandas vs Jupyter Notebook to simulate the data, a CSV file to represent our data before it was used, an exploratory data analysis which helped visualized our data, and a logistic regression model which provided a more accurate prediction of outcomes based on probability rather than a binary system which strictly stated a win or loss.

As part of our analytical process, Exploratory Data Analysis (EDA) played a critical role in validating the structure and predictive potential of our simulated dataset. Using Python libraries such as Pandas, Seaborn, and Matplotlib, we created a series of visualizations to better understand the relationships between attorney attributes and legal case outcomes.

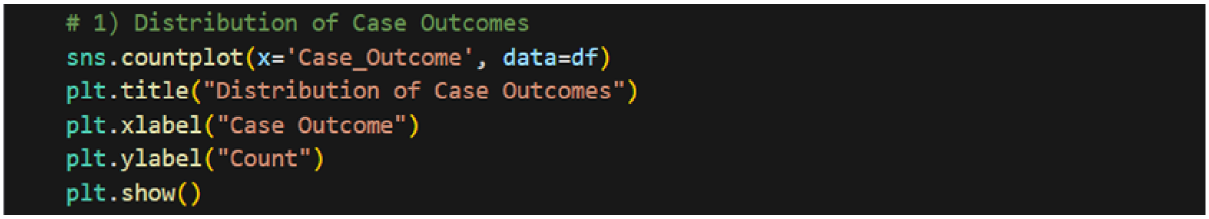


```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

df = pd.read_csv('attorney_case_dataset.csv', na_values=[], keep_default_na=False)

print(df.head())
print(df.info())
print(df.describe())

print("\nMissing values:\n", df.isnull().sum())
```

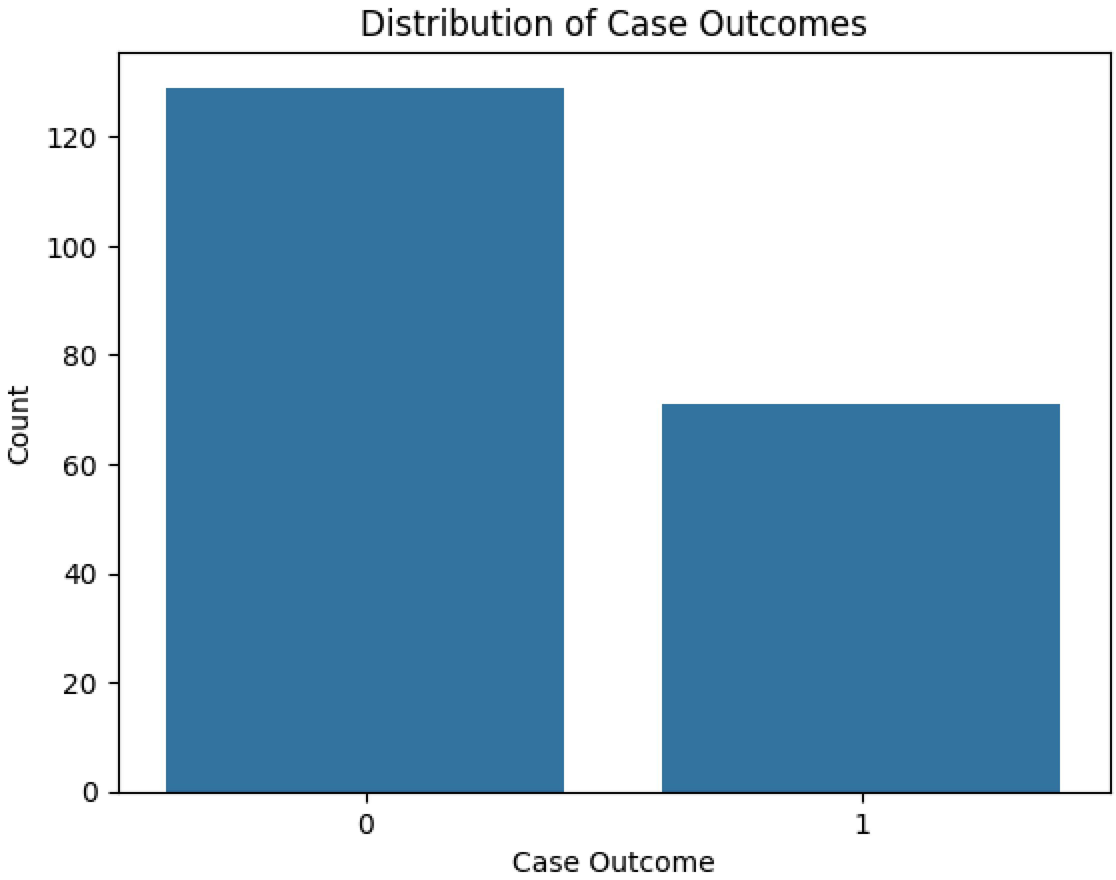


```
   Experience_Years  Win_Loss_Ratio Education_Level  Weighted_Score  \
0                16            0.72              JD            0.62
1                 0            0.49            None            0.23
2                11            0.59             LLM            0.44
3                13            0.37              JD            0.47
4                 1            0.38              JD            0.27

   Case_Outcome
0             1
1             0
2             0
3             0
4             0
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 5 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Experience_Years  200 non-null    int64
 1   Win_Loss_Ratio    200 non-null    float64
 2   Education_Level   200 non-null    object
 3   Weighted_Score    200 non-null    float64
 4   Case_Outcome      200 non-null    int64
dtypes: float64(2), int64(2), object(1)
memory usage: 7.9+ KB
...
Education_Level     0
Weighted_Score      0
Case_Outcome        0
dtype: int64
```

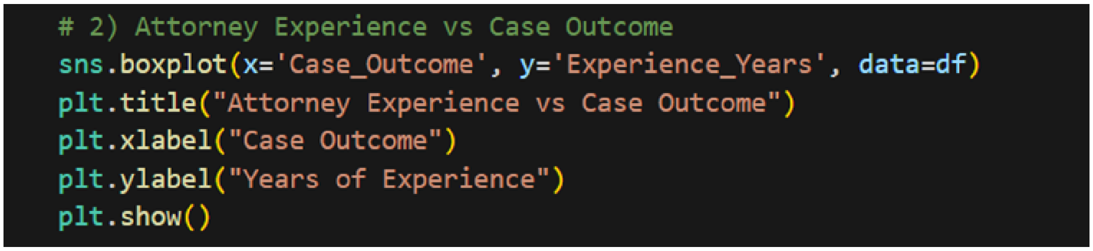
We began with a count plot of the case outcomes, which illustrated an imbalance in the dataset: a much larger number of cases were lost than won. This trend highlighted the importance of considering class imbalance during model evaluation and tuning.



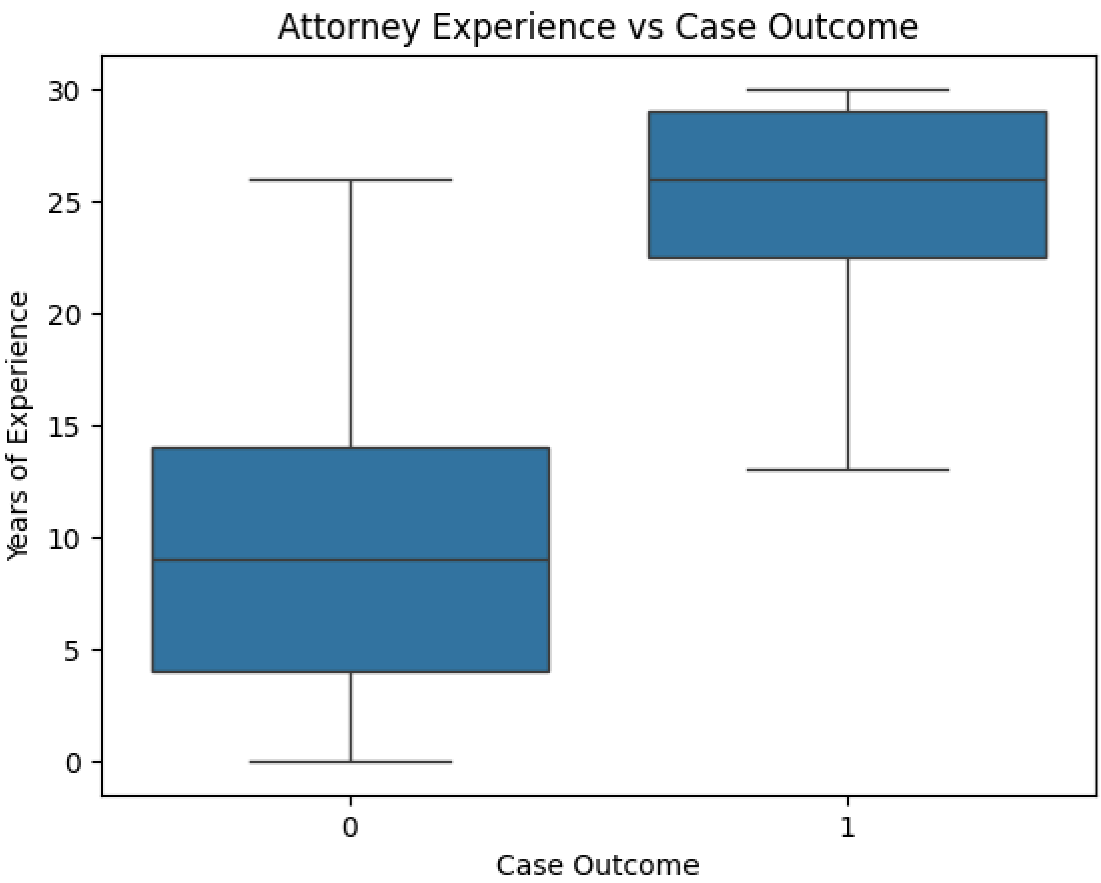
```
    # 1) Distribution of Case Outcomes
    sns.countplot(x='Case_Outcome', data=df)
    plt.title("Distribution of Case Outcomes")
    plt.xlabel("Case Outcome")
    plt.ylabel("Count")
    plt.show()
```



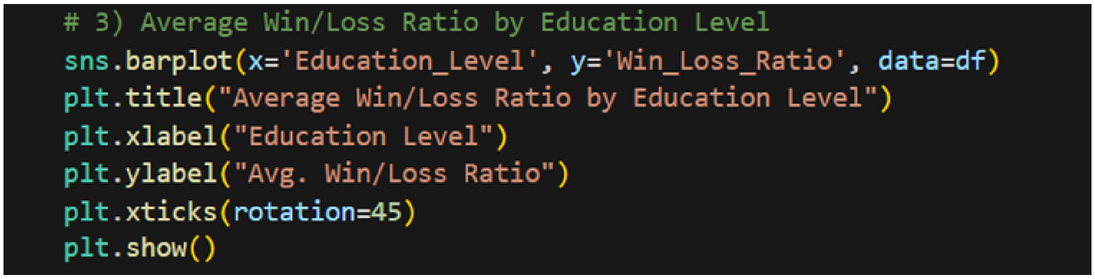
Next, we created a box plot comparing years of experience to case outcome. This visualization showed that attorneys who won cases generally had more years of experience, often well above 20 years, while those who lost tended to have fewer than 10 years. This supports the idea that experience is positively correlated with success.



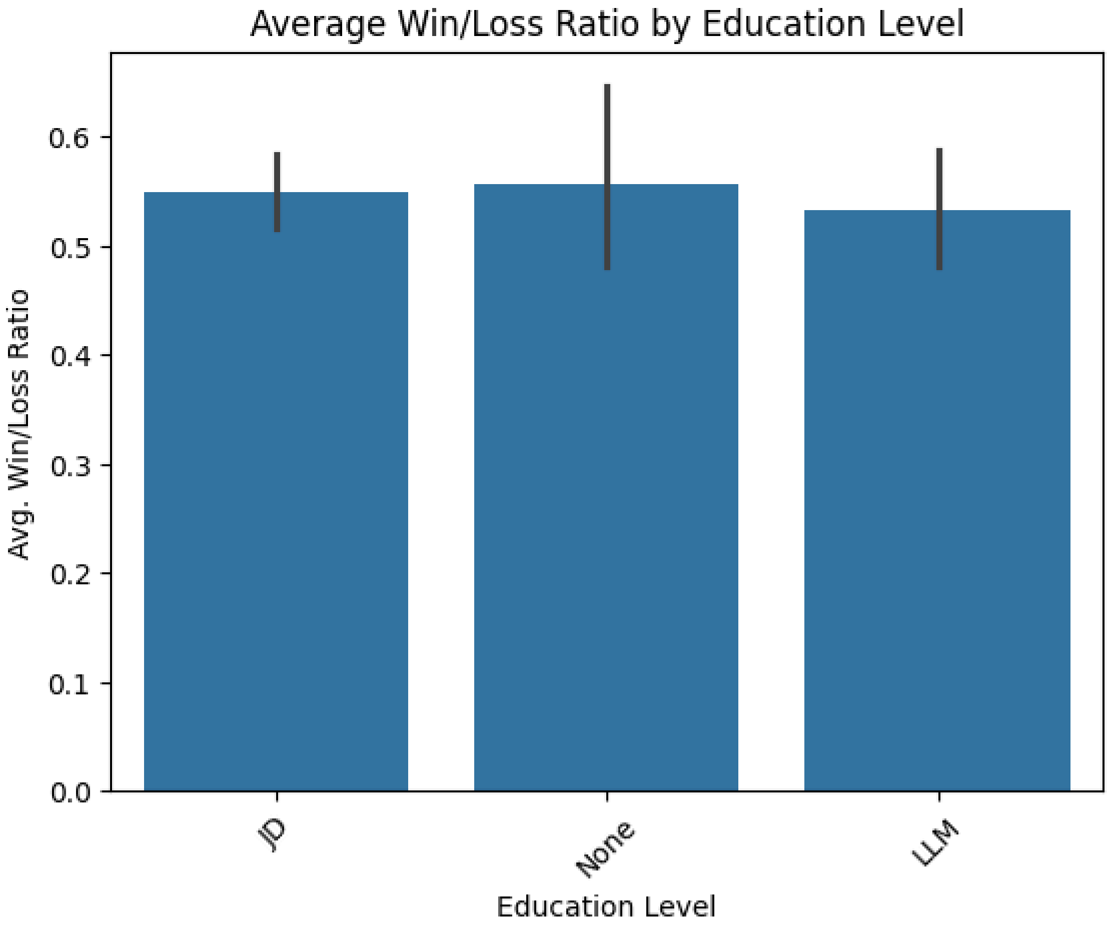
```
# 2) Attorney Experience vs Case Outcome
sns.boxplot(x='Case_Outcome', y='Experience_Years', data=df)
plt.title("Attorney Experience vs Case Outcome")
plt.xlabel("Case Outcome")
plt.ylabel("Years of Experience")
plt.show()
```



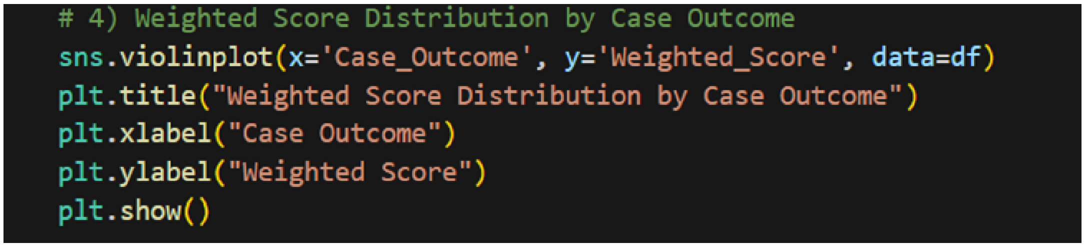
We then analyzed education by plotting the average win/loss ratio by education level. Interestingly, attorneys with JD, LLM, or even no listed degree ("None") performed at very similar levels, with only minor differences between the groups. This revealed that, in this dataset, education level alone did not show a strong effect on winning rates, suggesting that experience and prior performance metrics are more predictive.



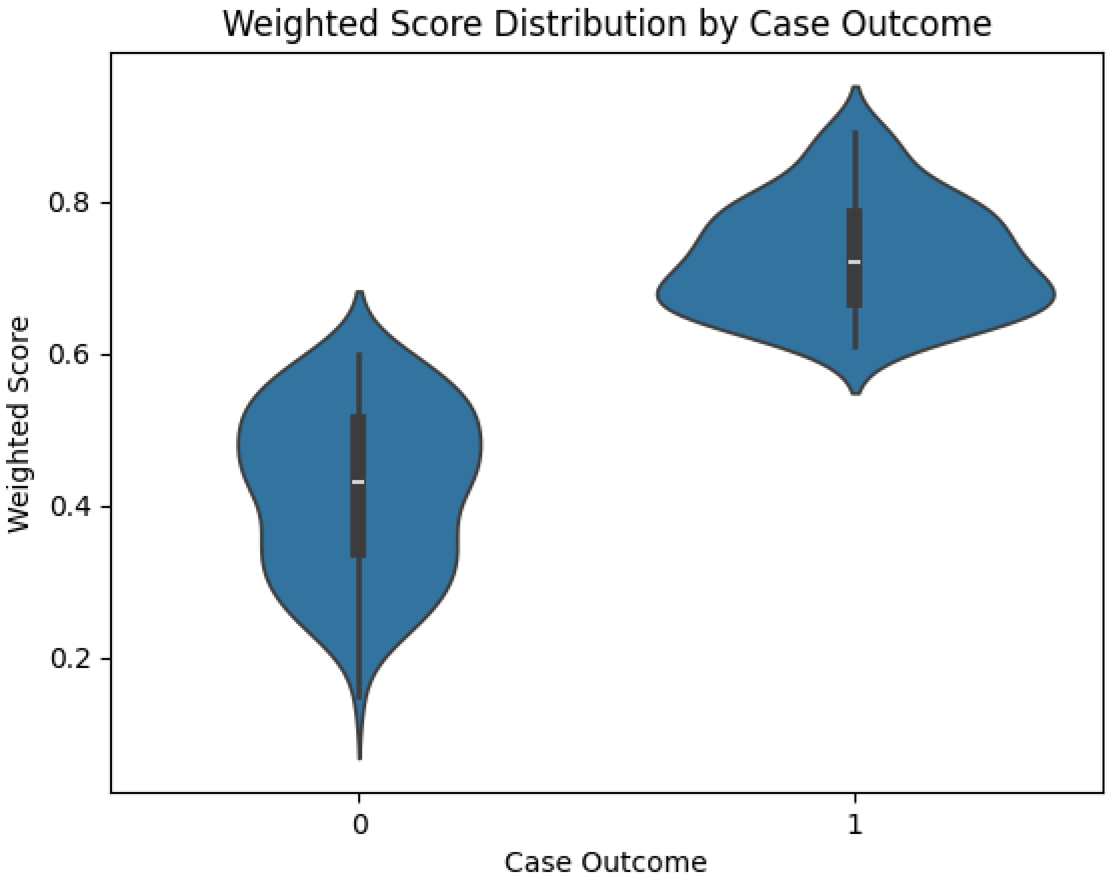
```
# 3) Average Win/Loss Ratio by Education Level
sns.barplot(x='Education_Level', y='Win_Loss_Ratio', data=df)
plt.title("Average Win/Loss Ratio by Education Level")
plt.xlabel("Education Level")
plt.ylabel("Avg. Win/Loss Ratio")
plt.xticks(rotation=45)
plt.show()
```



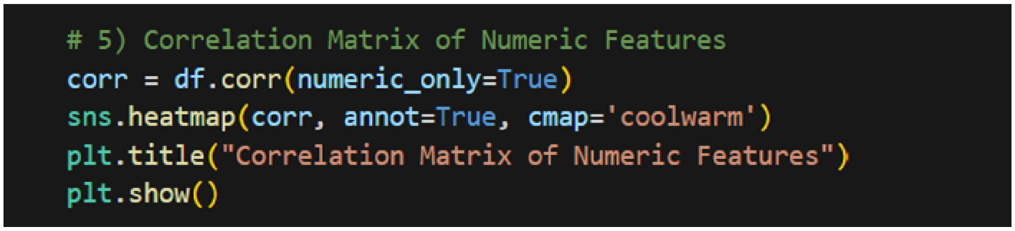
A violin plot of weighted score versus case outcome further confirmed our findings: attorneys who won consistently had higher weighted scores, with scores typically centered around 0.7 – 0.8, compared to lower scores for losing attorneys. This validated our approach of combining different features into a single weighted score.



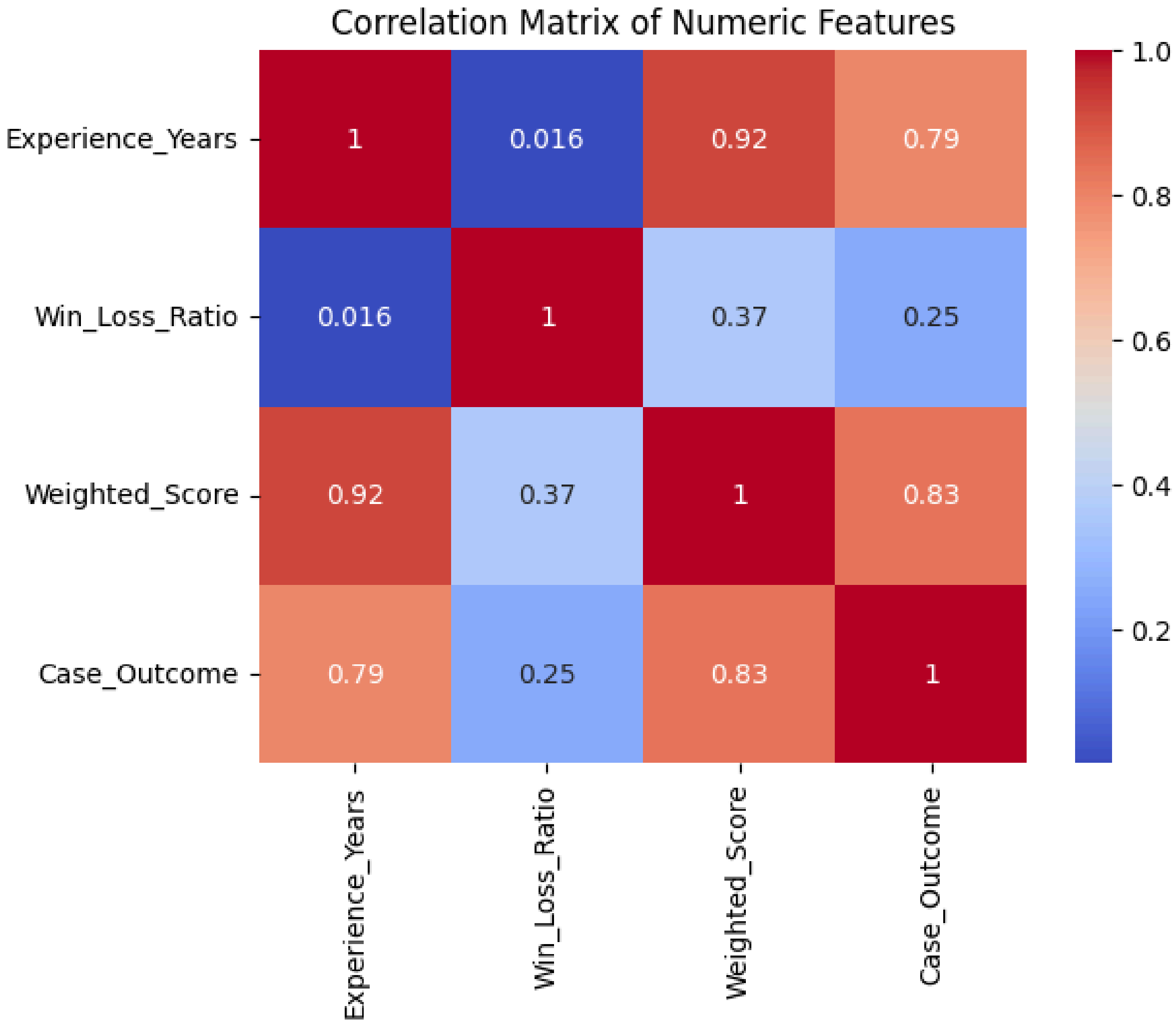
```
# 4) Weighted Score Distribution by Case Outcome
sns.violinplot(x='Case_Outcome', y='Weighted_Score', data=df)
plt.title("Weighted Score Distribution by Case Outcome")
plt.xlabel("Case Outcome")
plt.ylabel("Weighted Score")
plt.show()
```



Finally, a correlation matrix highlighted strong positive correlations among key features. Years of experience and weighted score showed a very high correlation (0.92), and weighted score also correlated strongly with case outcome (0.83). These insights confirmed that the weighted score is a meaningful and useful composite predictor.



```
# 5) Correlation Matrix of Numeric Features
corr = df.corr(numeric_only=True)
sns.heatmap(corr, annot=True, cmap='coolwarm')
plt.title("Correlation Matrix of Numeric Features")
plt.show()
```



Together, these analyses confirmed that our simulated data reflected logical patterns, allowing us to confidently proceed to logistic regression modeling.

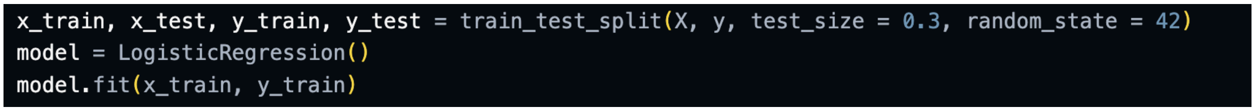
We hypothesized that lawyers with more pull and emphasis on their experience could from a binary perspective (either 1 or 0, win or loss) would win more often even if their other 2 factors, win rate and education were lacking in comparison. Of course, they have their own pull and their own importance within deciding the outcome of the case, however within the CSV file especially, it was apparent that most if not all attorneys who were victorious were not lacking in the years of experience column.

Some advantages and disadvantages of this project include that our weighting system and factors/data were controlled. While the data was randomized it wasn't real life data, so some factors in comparison to real life may have been more randomized and thus may not have been a very accurate representation of reality, however the controlled setting and manner of our project made it easier to visualize everything, hypothesize the result, and also work with the logistic regression model for probability.

Some changes in our project compared to our proposal mainly include not using advanced techniques like SVMs among other models and techniques which didn't seem to provide much more value for what we were going for as the EDA and logistic regression model were perfectly capable of what this project aimed to uncover. In the interest of time and not overcomplicating it the best approach for this project seemed to be choosing fewer models but delving deeper into them instead of spreading out amongst several models/techniques with a shallow understanding or use.

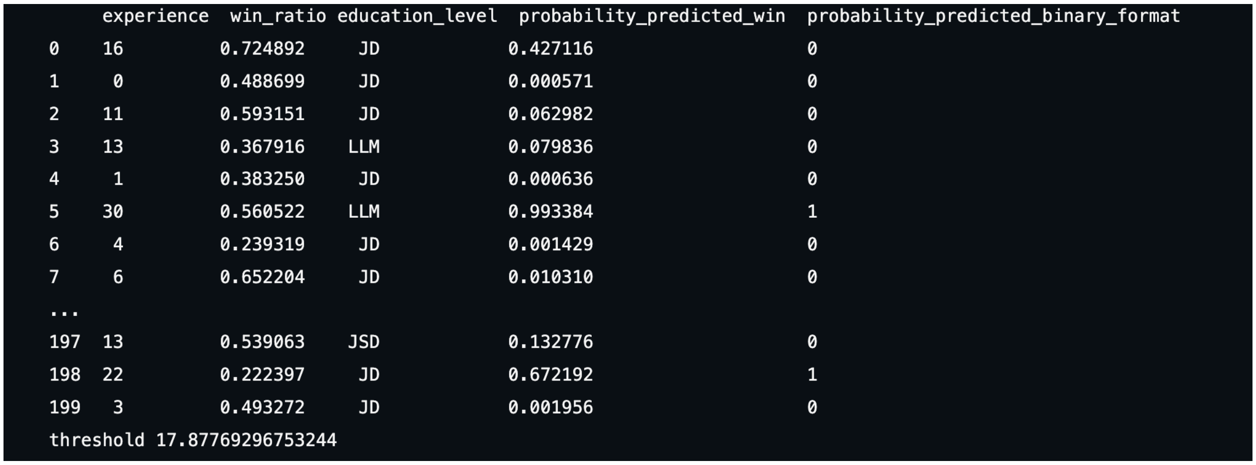
Apart from using an Explanatory Data Analysis to help us understand and organize our information, we also created a Logistic Regression Model which visualized our data, and returned us a probability of each of the 200 lawyers' likelihood to win a case. We decided to use a logistic regression model over a linear regression model, because a logistic regression model would show us the probability of the outcomes. A linear regression model shows a continuous number, which was not fully supportive of our data, and would not result in the precise conclusion that we came to.

In order for us to program the model, we had to import numpy, pandas, matplot, seaborn, and sklearn. These all came in handy because their libraries made it possible for us to display graphs, write readable code, and ensure conciseness and precision. We started by using the coefficients: experience, education level, and win ratio to format our model. We then trained this model with a 80% training, 20% testing split. The training portion would create the model and work through the data, and the testing portion was responsible for ensuring it was working properly and compared correctly to the data we gave it. This can be seen below:



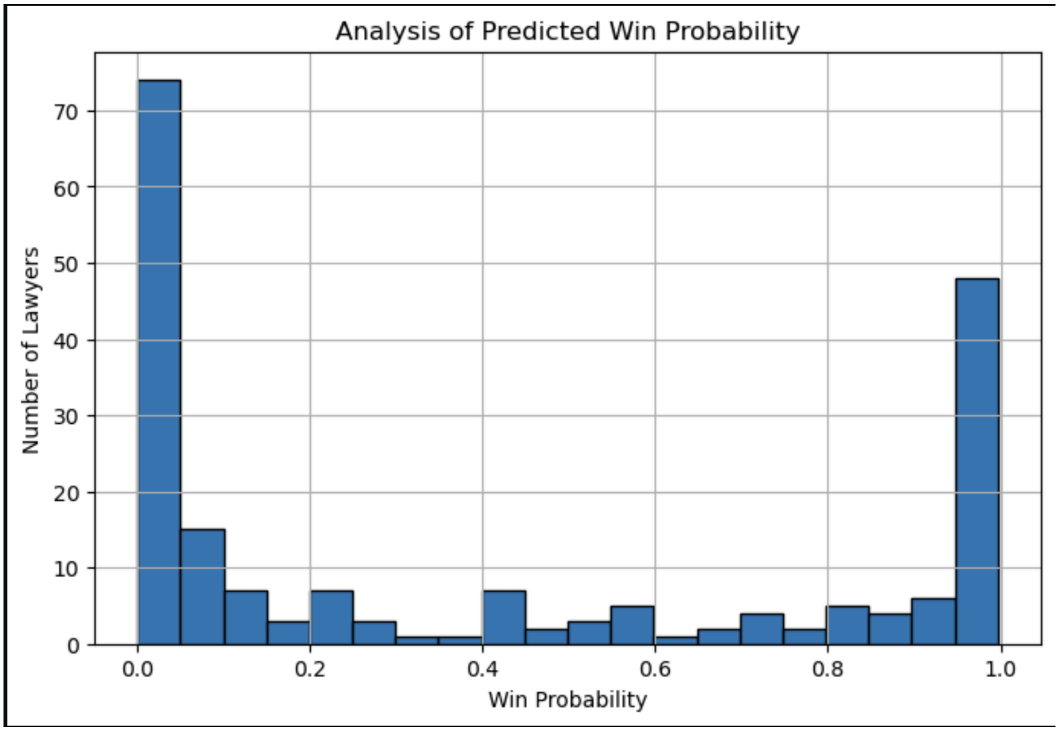
```
x_train, x_test, y_train, y_test = train_test_split(X, y, test_size = 0.3, random_state = 42)
model = LogisticRegression()
model.fit(x_train, y_train)
```

After doing this, we printed out the results in both decimal form and binary form. We decided to do this to return a surplus of information in different ways so visualization of our results would be easier. Printing them out in binary gave us the idea of which coefficient combinations would result in a win or a loss, but doing so in precise decimals provided us the exact number resulting from the three coefficients.

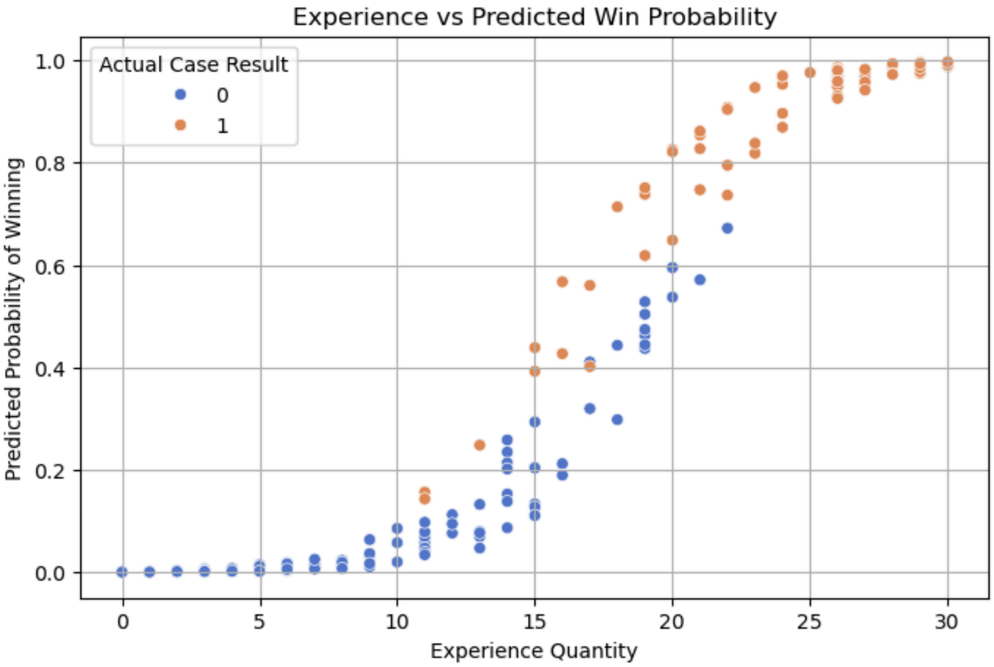


```
       experience  win_ratio education_level  probability_predicted_win  probability_predicted_binary_format
   0    16         0.724892     JD           0.427116                   0
   1     0         0.488699     JD           0.000571                   0
   2    11         0.593151     JD           0.062982                   0
   3    13         0.367916    LLM           0.079836                   0
   4     1         0.383250     JD           0.000636                   0
   5    30         0.560522    LLM           0.993384                   1
   6     4         0.239319     JD           0.001429                   0
   7     6         0.652204     JD           0.010310                   0
   ...
   197  13         0.539063    JSD           0.132776                   0
   198  22         0.222397     JD           0.672192                   1
   199   3         0.493272     JD           0.001956                   0
   threshold 17.87769296753244
```

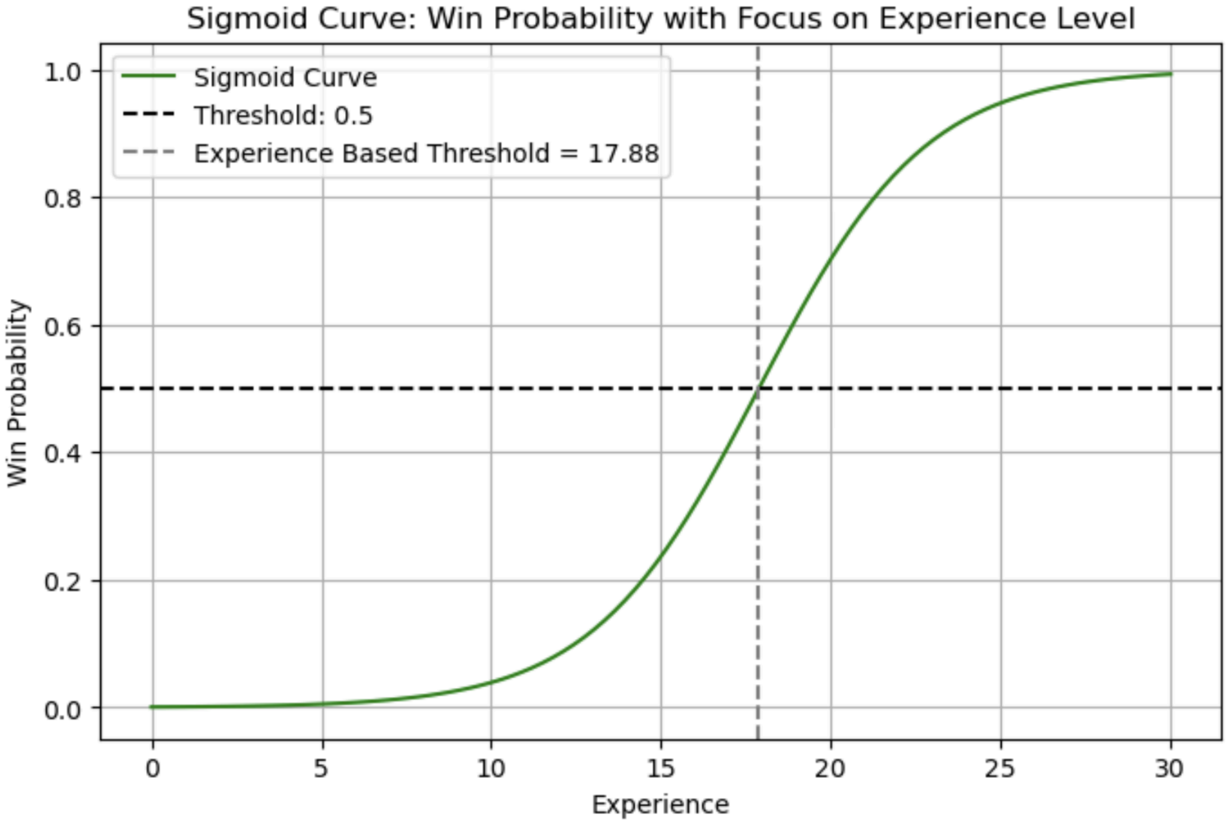
We then coded our graphs. We chose to code and display three different graphs to provide an array of information and visualization. Our first graph is a histogram, directly imaging the 200 lawyers, and where they fall under the realm of win probability. We did this through the libraries provided through matyplot.pyplot and sklearn. After displaying this graph, we were shown that a large surplus of the lawyers were less probable to win a case, versus those who were more probable. This histogram gave us a general understanding of how the coefficients influence a case.



Next, we programmed and displayed a scatter plot. This scatter plot specifically targeted the experience age of each lawyer and how it correlated with their probability of winning a case. We found that the information returned back to us from the scatter plot resembled an S curve, similar to the findings we would have with the sigmoid curve. Denoting the 0(likely to lose a case) and the 1(likely to win a case) with different colors, we discovered that the more experience one had in law, the more likely they are to win. This brought us closer to a correct hypothesis.



Lastly, we programmed and displayed our sigmoid curve, which provided us the most information out of the three graphs. This sigmoid curve was heavily influenced by the experience age of the lawyer, and that is because the experience age was the most versatile, and was a continuous number. We attempted to create the sigmoid curve with the other two coefficients being the lead, but found it less informative. This was due to the fact that there are only three education levels a lawyer could be, making education level more finite than any of the other coefficients, therefore unhelpful as the lead. The win ratio we found to be high, which made it very difficult for the graph to include and withold all ends of the lawyers' other two demographics. This resulted in a less productive graph, so we isolated that idea as well. Finding that experience age was the best option, we followed the route and used a finite mean number from the other two coefficients to return our graph.



The graph contains three lines. The green S curved line, represents the sigmoid curve, and displays the experience and other coefficients increasing the win probability. The black dotted line shows the basis threshold, that we marked as 0.5. This threshold was integral in showing us that any data under that line would be less probable to win a case. Turning to a more specific line, the grey dotted line represents a threshold ONLY for the experience age. We found that threshold to be approximately 17.88 years of experience. This showed us that any experience age lower than that would have a small chance of winning a case, and vice versa.

Creating the logistic regression model, we found three different ways to display the information, and it gave us more clarity on how the three coefficients individually affected a lawyer's turnout of a case. This validated our understanding of why a logistic regression model would be more useful than a linear regression model, and conclusively useful in supporting our hypothesis.

Conclusion

All in all, our project gives a data based approach on predicting outcomes of a legal case based on some of a lawyers demographic. We hypothesised that a lawyer with more experience, a higher education, and a larger number of past cases won would result in them having a likelier chance of winning a future case. We used an explanatory data analysis and logistic regression model to validate our hypothesis. We utilized information we learned from lecture slides and videos, and we found our hypothesis to be true as can be seen by the graphs provided above. The difficulties we found, included deciding on valuable coefficients to use. This stems from law being a confidential and legal based occupation, and many demographics being confidential. We decided on the three listed above, and then had to find correct ranges for all three as ranges are capable of changing data very easily. Our project would be useful in the real world since it would help a client choose the right attorney for them. It would also give students in law school or even lawyers a better array of expectations for their future endeavors. With more demographics, including the location of the case, the client, the jury, and the judge, the project could be used to influence the entire court room.