**Analysis of Sports Gambling in the U.S.**

Project Deliverable 2

May 1, 2021

Michael Alberts (11758253)

Jeffrey Bridges (10622476)

Ryan Hartquist (20183791)

Andrew Kuchel (11147971)

Table of Contents

[I. Executive Summary 1](#_Toc70794824)

[II. Statement of Scope 2](#_Toc70794825)

[a. Project Objectives 2](#_Toc70794826)

[b. Unit of analysis 2](#_Toc70794827)

[c. Variables 3](#_Toc70794828)

[III. Project Schedule 3](#_Toc70794829)

[IV. Data Preparation 5](#_Toc70794830)

[a. Data Access 5](#_Toc70794831)

[b. Data Consolidation 6](#_Toc70794832)

[c. Data Cleaning 7](#_Toc70794833)

[d. Data Transformation 8](#_Toc70794834)

[e. Data Reduction 9](#_Toc70794835)

[f. Data Dictionary 11](#_Toc70794836)

[V. Descriptive Statistics and Analysis 15](#_Toc70794837)

[VI. Visualizations 16](#_Toc70794838)

[VII. Modeling Techniques 22](#_Toc70794839)

[a. Multiple Regression 22](#_Toc70794840)

[b. Neural Network 24](#_Toc70794841)

[c. Extreme Gradient Boosting regressor (xgboost) 25](#_Toc70794842)

[VIII. Data Splitting and Sub-Sampling 26](#_Toc70794843)

[a. Justification and Explanation 26](#_Toc70794844)

[b. Split Comparison 27](#_Toc70794845)

[IX. build models/Model Results 28](#_Toc70794846)

[a. Regression 28](#_Toc70794847)

[b. Neural Network 30](#_Toc70794848)

[C. XGBOOST 31](#_Toc70794849)

[X. Model Assessment 33](#_Toc70794850)

[a. Assessment Choice 33](#_Toc70794851)

[b. Model Comparison 34](#_Toc70794852)

[c. Final Choice 35](#_Toc70794853)

[XI. Predicting Bet Revenue for states without legalized gambling 36](#_Toc70794854)

# Executive Summary

In the United States, only 18 states along with the District of Columbia and Puerto Rico allow legal bets on sporting events. Six more states have pending litigation to approve sports betting in 2021. Before a supreme court ruling in 2018, the Professional and Amateur Sports Protection Act (PASPA) made most sports betting illegal. Nevada was grandfathered an exempt status and was the only state with significant legalized sports gambling until the recent ruling. Since that ruling to strike down PASPA, those states which have passed legislation to legalize sports betting have realized an additional revenue stream for their governments. The main goals of this project are: to generate a model for revenue from states that have legalized sports betting that will evaluate the most significant variables and estimate potential revenue in states that have yet to legalize sports betting. The results of the analysis will identify trends and growth in revenue from legalized sports betting states and show a pathway to maximize a substantial source of income for state governments.

# Statement of Scope

We will implement a model to estimate revenue in any states that have not yet adopted legislation to legalize sports betting. Oklahoma is uniquely positioned to adopt legalized sports gambling due to its already established framework for Native American gaming institutions. This will involve scraping public revenue data from state governments to gather crucial variables for analysis. Additionally, we will scrape demographic data to profile users, understand state population nuances that may impact revenue numbers and behavioral activity in each state.

## Project Objectives

* Create a model of additional state revenue generated by the legalization of sports betting.
  + Compare and analyze the additional revenue generated in states with legalized sports betting
  + Compare and analyze the demographics between states and correlate that to states where gambling has already been legalized to overall revenue and sports betting revenue
  + Identify the most important variables in the model to help states maximize opportunity
  + Utilize the model to predict sports betting revenue for states that have not legalized sports betting

## Unit of analysis

The unit of analysis for our project is a state. We are generating a model capable of assessing the possible revenue a state could generate by the legalization of sports betting by analyzing the performance of states that have already legalized and realized tax revenue from sports betting.

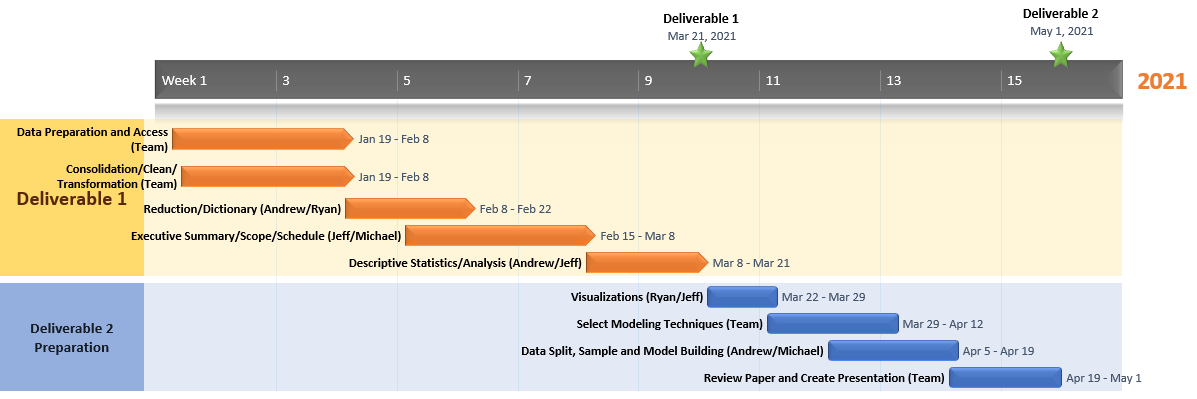
## Variables

The main target variable is the revenue a state would generate on either a yearly or monthly basis based on the adoption of legislature that would allow for legal sports gambling in a state. Other potential variables of interest include calculating the number of potential sports bettors in a state to give governments an idea of the scope of potential opportunities for other revenue streams related to them (Advertising opportunities, opportunities for additional casinos to place bets, and tourism to said casinos). Predictor variables will relate to state population demographics and the proportion of revenues currently generated by general gambling.

# Project Schedule

The project is scheduled for three parts. Deliverable 1, deliverable 2, and the final presentation. Since the beginning of the semester, the team has been preparing the data by using data consolidation, cleaning, and transformation. After “data wrangling” was complete, we turned our attention to see if any data reduction was warranted. Several techniques were utilized including PCA, Factor Analysis, and Clustering. A data dictionary was also created. Once the data was prepared, descriptive statistics and analysis of the variables were conducted to conclude the objectives for deliverable 1.

After Deliverable 1 is complete, Jeff and Ryan will build visualization to help explore the data. The team will discuss which models should be built and the team will select the best model and any flaws associated with other models. Andrew/Michael will come up with the appropriate data splitting and sub-sampling numbers. Once the data splitting is done, it will be time to build the models. After the models are built, we will have a team meeting to assess the models and move into the final analysis. As our schedule shows below, this should be concluding no later than April 19th, 2021 to have plenty of time to review the report and begin creating a presentation. Our goal as a team is to have the presentation recorded by May 3rd to allow plenty of time for corrections.



# Data Preparation

## Data Access

A major source of compiled data to utilize is LegalSportsReport.com (<https://www.legalsportsreport.com/>). LegalSportsReport.com covers the legal online sports wagering industry, including sports betting sites and daily fantasy sports in the US. This information was collected through a web crawler to be further parsed and analyzed through individual efforts with analytics software. For each state that has legalized sports betting, the data includes: Handle, Revenue, Hold Percentage, and State Tax Generated by month and year. The handle is the total amount of money bet through the sportsbook. The revenue is the amount in dollars that the sportsbook collects after paying out winnings (this is the number that taxes are based on). Hold percentage is the percentage of the betting revenue divided by the handle. The state tax generated is the taxes the state receives from the sportsbook. This number can vary by each state and its tax rules. A data frame with overall totals for all variables above was also created. This frame sums up the totals for Handle, Revenue, Hold Percentage, and State Tax Generated for a state since the inception of legalized betting in that state.

The other portion of data concerns gathering data from states to equate potential revenue numbers for states where they to legalize wagering, the number of potential sports gamblers, and potential increases in state budgets. The data gathered for each state includes a percentage age breakdown of age distributions, state population, median household income, per capita income, total tax dollars collected, and per capita taxes collected for 2017, 2018, and 2019. State names and age breakdowns were gathered with a web scraper from the Kaiser Family Foundation (<https://www.kff.org/>). KFF provides estimates on demographic breakdowns based on available census and survey data that they collect. Due to the nature of the KFF website, with ever-updating information and graphics, Selenium had to be utilized to successfully gather the appropriate data. Data was individually pulled for each age bracket. State population data, median income, and per capita income data were pulled from Wikipedia utilizing the html\_table node (<https://en.wikipedia.org/wiki/List_of_states_and_territories_of_the_United_States_by_population>, <https://en.wikipedia.org/wiki/List_of_U.S._states_and_territories_by_income>). Tax revenue was gathered from The Federation of Tax Administrators (https://www.taxadmin.org). The data on this website is divided into separate web pages by year, so a different URL per year of data was necessary. Data points extracted from this source were total tax revenue by year/state and per capita taxes by year/state. The data was organized cleanly on the website, and the only cleansing required was numeric transformation.

## Data Consolidation

Most of the data consolidation was performed by the web scraping scripts as written. For the gambling revenue data, based on the 18 states that currently allow it, the data was assembled into a data frame for each data set, they could be used in total or separately. Different data sets could be consolidated using the State or Jurisdiction name as an identifier. For the state population data, each column of the data was scraped using a separate query. Due to pulling data from multiple sources, a column with the state name was used with each query to act as a primary key for that column to make sure the order of the data stayed consistent throughout the merging process. Initially, the population and income columns of this data frame included data for U.S. Territories, Washington D.C., as well as overall data. Because we are focused on states, these columns were removed in the data consolidation process. Once each column was reduced to 50 rows (one for each state), the data was sorted by state alphabetically to ensure all states lined up correctly and then merged into one data frame.

## Data Cleaning

After the data extraction, the data had to be cleaned to perform an analysis. Some of the key aspects performed to clean the data include:

* Addressing missing data by relabeling to a usable format
* Removing extra spaces and characters to maintain variable integrity
* Convert scraped data from “Character” class-type to numeric where appropriate
* Creating complete data frames that could be manipulated to perform analysis

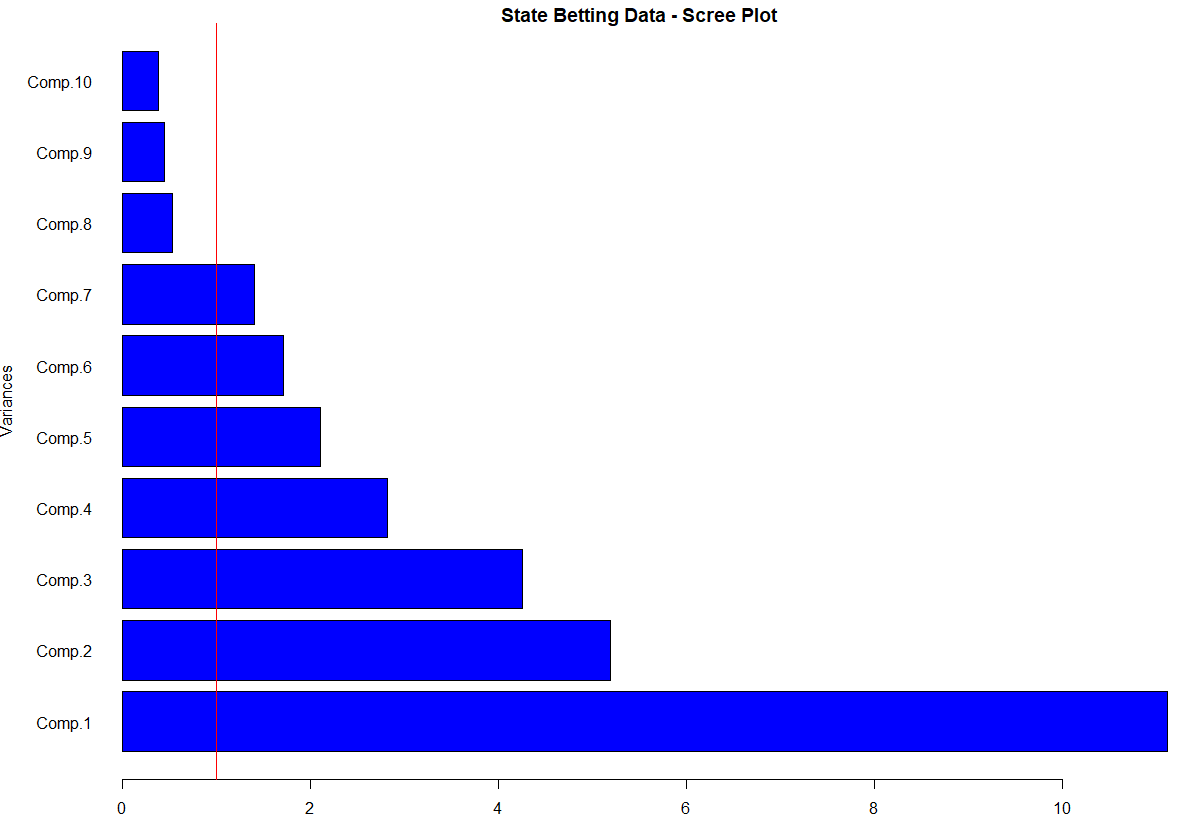
Due to the nature of our project, there were very few missing values in our data sets. The state demographic data frame pulls data from widely available sources where missing values would not be accepted. The sports betting and revenue data frames are set up in a way where the data is only pulled once a state legalizes sports betting and starts to release revenue numbers. As an example, the data frame containing revenue numbers for Indiana starts in September 2019 to the present, while the data frame for Mississippi begins in August 2018 to present. Some states reported no revenues for March 2020-June 2020 due to the coronavirus pandemic and casinos being closed. All zero values were relegated to missing data as it was considered atypical and an irregular variation in the data.

## Data Transformation

One variable we calculated was the percent of taxes realized due to sports betting compared to overall state tax income. This would help a state realize how much impact legalized gambling could have on its budget. While this number appears small compared to overall revenue for states, the important thing that it is showing is the growth of that percentage year over year. Even when taking into account the uniqueness of the year 2020 from a sports gambling perspective (no sports to gamble on for multiple months due to the pandemic) for most states, and the fact that most sports books took in very little money for a few months earlier this year, we are still able to see projected growth that should only increase. To be able to perform this analysis, we first took our merged data frame with a row for each month of a year that a state had shown money gathered from sports betting and summed the money data by state and year to get an amount that each state had generated over 2018, 2019, and 2020, and for the first few months of 2021, if the data was available. After that, the demographic data and sports betting data frames were merged using the state name as the primary key. Each of the ‘Total Taxes Collected’ columns, which represent the total amount from gambling, were divided by the ‘Total Taxes’ columns to generate the 4 ‘Percent of taxes from betting’ columns. As the group gets farther into the analysis portion of the report, we will determine the need for normalization or transformation of the data. One of the variables we plan to calculate is the potential number of gamblers per state. Once more information is gathered, we will convert the age demographics based on research to the potential number of gamblers per state.

## Data Reduction

The team performed a PCA (Principal Components Analysis) along with a Factor Analysis to determine if any variable reduction was appropriate. The PCA creates linear combinations of variables, determining which variables are similar to one another and creating new columns of data. Each component is assessed via eigenvalues (the square of the standard deviation). Although we do not use the output of the columns generated by the PCA, we use the interpretation of the analysis to inform which variables may be removed. Below is a scree plot of the components ordered by the eigenvalues, where a significant eigenvalue is greater than 1. We can see that the PCA suggests that the 6 columns of age-based demographic data along with population are the most influential.



Following the PCA, the team performed a Factor Analysis, which determines how similarly the variables behave. The results of the factor analysis produce scores for each variable and factor, where the absolute value scores between .7 and 1 are more influential. These results more or less agree with the PCA in that the highest representative scores from each factor suggest the same variables (except the demographic age group 19-25 and Population).



The team concluded that, although the PCA and Factor Analysis suggest reducing the dataset to ~7 variables, the intention of our model is to predict potential revenue. To provide the regression model with the necessary evaluation parameters, we will include all the variables and evaluate the statistical significance and variable dominance.

Data gathered for the state of Nevada will likely have to be removed from any analysis. Before 2018, Nevada was the only state with legalized sports betting. Sports betting was legalized in Nevada in 1949 and has always been a source of revenue for the state as a tourist destination, unlike other states which could only legalize sports betting starting in 2018. Some other states or jurisdictions may also behave so differently than what appears typical that including in analysis distorts the model in an unbeneficial way.

## Data Dictionary

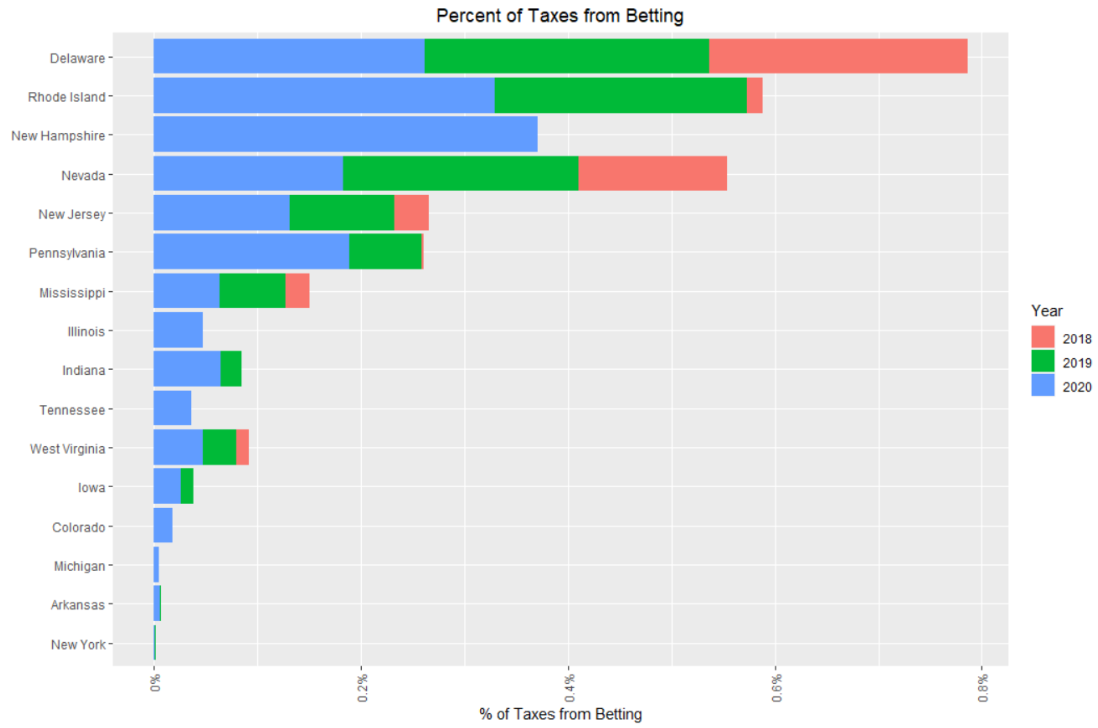
| **Attribute Name** | **Description** | **Data Type** | **Source** |
| --- | --- | --- | --- |
| **Handle** | Amount wagered over the time period. | Numeric | <https://www.legalsportsreport.com/sports-betting/revenue/> |
| **Bet Revenue** | Amount of money kept by sportsbooks out of the amount wagered. | Numeric | <https://www.legalsportsreport.com/sports-betting/revenue/> |
| **Hold %** | How much revenue sportsbooks keep as a function of handle. | Numeric | <https://www.legalsportsreport.com/sports-betting/revenue/> |
| **Taxes** | Taxes collected by state and local jurisdictions; or state share of proceeds in revenue-sharing markets. | Numeric | <https://www.legalsportsreport.com/sports-betting/revenue/> |
| **State Name** | State name. Used as primary key to link different data frames together. | Character | https://www.kff.org/other/state-indicator/distribution-by-age/ |
| **0-18** | Percentage of state population that falls between ages 0 and 18. | Numeric | https://www.kff.org/other/state-indicator/distribution-by-age/ |
| **19-25** | Percentage of state population that falls between ages 19 and 25. | Numeric | https://www.kff.org/other/state-indicator/distribution-by-age/ |
| **26-34** | Percentage of state population that falls between ages 26 and 34. | Numeric | https://www.kff.org/other/state-indicator/distribution-by-age/ |
| **35-54** | Percentage of state population that falls between ages 35 and 54. | Numeric | https://www.kff.org/other/state-indicator/distribution-by-age/ |
| **55-64** | Percentage of state population that falls between ages 55 and 64. | Numeric | https://www.kff.org/other/state-indicator/distribution-by-age/ |
| **65+** | Percentage of state population that falls over age 65 | Numeric | https://www.kff.org/other/state-indicator/distribution-by-age/ |
| **Population** | State populations based on 2019 data | Numeric | <https://en.wikipedia.org/wiki/List_of_states_and_>  territories\_of\_the\_United\_States\_by\_population |
| **Median Household Income** | Average Median Household Income per household per state | Numeric | <https://en.wikipedia.org/wiki/List_of_U.S._>  states\_and\_territories\_by\_income |
| **Per Capita Income** | Average per person income per state | Numeric | <https://en.wikipedia.org/wiki/List_of_U.S._>  states\_and\_territories\_by\_income |
| **Total Taxes 2017** | Total taxes in millions each state collected- 2017 | Numeric | https://www.taxadmin.org/2017-state-tax-revenue |
| **Per Capita Taxes 2017** | Average per person taxes per state | Numeric | https://www.taxadmin.org/2017-state-tax-revenue |
| **Total Taxes 2018** | Total taxes in millions each state collected- 2018 | Numeric | https://www.taxadmin.org/2018-state-tax-revenue |
| **Per Capita Taxes 2018** | Average per person taxes per state | Numeric | https://www.taxadmin.org/2018-state-tax-revenue |
| **Total Taxes 2019** | Total taxes in millions each state collected- 2019 | Numeric | https://www.taxadmin.org/2019-state-tax-revenue |
| **Per Capita Taxes 2019** | Average per person taxes per state | Numeric | https://www.taxadmin.org/2019-state-tax-revenue |
| **Per Capita Taxes 2019** | Average per person taxes per state | Numeric | https://www.taxadmin.org/2019-state-tax-revenue |
| **Percent of Taxes from Betting 2018** | Percentage of state income gathered through sports wagers | Numeric | Calculated from Total Taxes per year and taxes collected through sports gambling |
| **Percent of Taxes from Betting 2019** | Percentage of state income gathered through sports wagers | Numeric | Calculated from Total Taxes per year and taxes collected through sports gambling |
| **Percent of Taxes from Betting 2020** | Percentage of state income gathered through sports wagers | Numeric | Calculated from Total Taxes per year and taxes collected through sports gambling |
| **Projected Percent of Taxes from Betting 2021** | Percentage of state income gathered through sports wagers | Numeric | Calculated from Total Taxes per year and taxes collected through sports gambling |

# Descriptive Statistics and Analysis

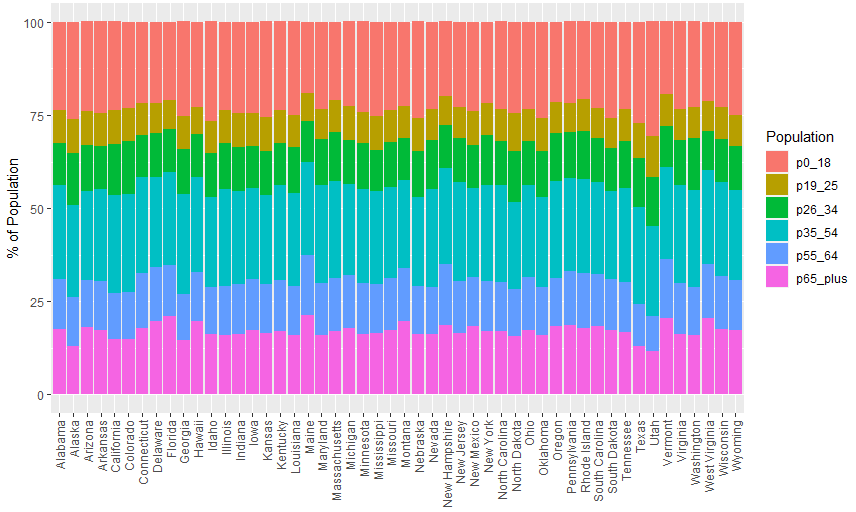
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **Mean** | **Median** | **Minimum** | **Maximum** | **Standard Deviation** |
| Handle | $ 132,039,160 | $ 34,462,821 | $ 0 | $ 996,300,794 | $ 198,208,988 |
| Bet Revenue | $ 9,441,834 | $ 3,073,950 | $ (890,623) | $ 82,640,979 | $ 14,112,448 |
| Hold (%) | 8.367 | 7.805 | -22.60 | 37.1 | 5.949 |
| Taxes | $ 1,367,016 | $ 459,478 | $ (454,218) | $ 17,000,911 | $ 2,224,245 |
| 0-18 | 23.56 | 23.5 | 19 | 30.8 | 2.095 |
| 19-25 | 8.646 | 8.6 | 7.2 | 10.9 | 0.606 |
| 26-34 | 12.14 | 11.95 | 10.5 | 14.1 | 0.8395 |
| 35-54 | 25.11 | 25.1 | 23.2 | 26.7 | 0.9016 |
| 55-64 | 13.51 | 13.5 | 9.5 | 16.3 | 1.0944 |
| 65+ | 17.07 | 17.05 | 11.5 | 21.3 | 1.9453 |
| Population | 6,550,675 | 4,558,234 | 578,759 | 39,512,223 | 7,389,282 |
| Median Household Income | $ 61,549 | $ 59,761 | $ 44,097 | $ 83,242 | $ 10,184 |
| Per Capita Income | $ 28,445 | $ 27,564 | $ 21,036 | $ 39,373 | $ 4,205 |
| Total Taxes 2017 | $ 18,858 | $ 11,506 | $ 1,190 | $ 155,932 | $ 24,794 |
| Per Capita Taxes 2017 | $ 2,919 | $ 2,785 | $ 1,608 | $ 5,015 | $ 835 |
| Total Taxes 2018 | $ 20,450 | $ 11,709 | $ 1,642 | $ 175,017 | $ 27,747 |
| Per Capita Taxes 2018 | $ 3,146 | $ 2,950 | $ 2,075 | $ 5,533 | $ 951 |
| Total Taxes 2019 | $ 21,631 | $ 12,322 | $ 1,781 | $ 188,235 | $ 29,505 |
| Per Capita Taxes 2019 | $ 3,353 | $ 3,136 | $ 2,086 | $ 6,521 | $ 1,036 |
| Percent of Taxes from Betting 2018 | 0.06833 | 0.02308 | 0.00223 | 0.24959 | 0.09315 |
| Percent of Taxes from Betting 2019 | 0.09536 | 0.06435 | 0.00065 | 0.27447 | 0.10381 |
| Percent of Taxes from Betting 2020 | 0.11138 | 0.05572 | 0.00120 | 0.37052 | 0.12041 |
| Projected Percent of Taxes from Betting 2021 | 0.02287 | 0.01986 | 0.00015 | 0.06286 | 0.01925 |

The only categorical variable in our project is State Name. It has 50 unique values. Most numeric values have very large standard deviations, which is to be expected when dealing with a data set revolving around states, due to vast population differences between states that would lead to vastly different scales of income for them. The demographic data on state populations have relatively small standard deviations, though. The descriptive statistics for Percent of Taxes from betting is showing growth year over year, which is the main point we wanted to illustrate. The values for 2019 and 2020 are very close, even with the issues surrounding 2020 sports gambling that was discussed earlier.

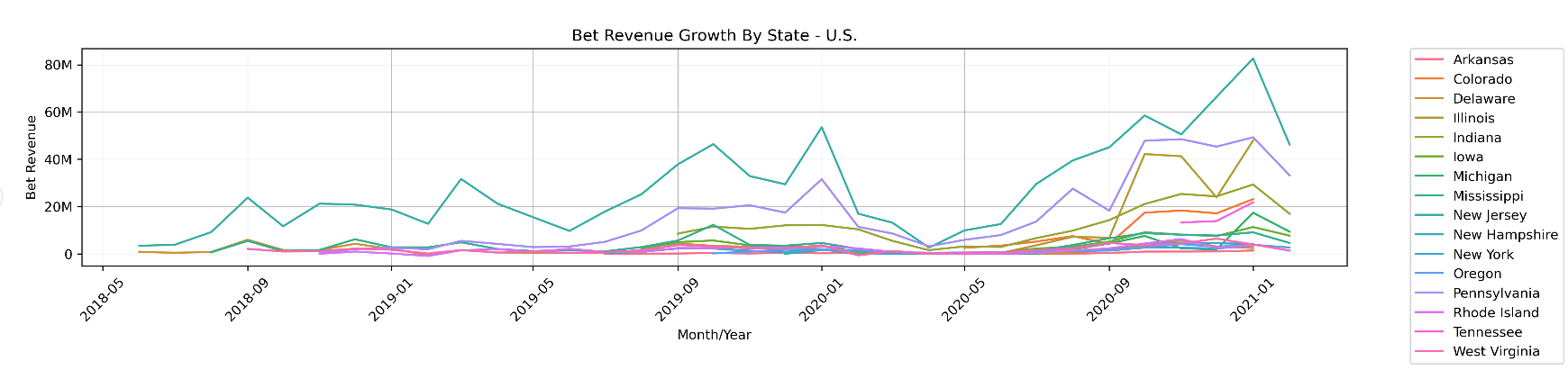
# Visualizations



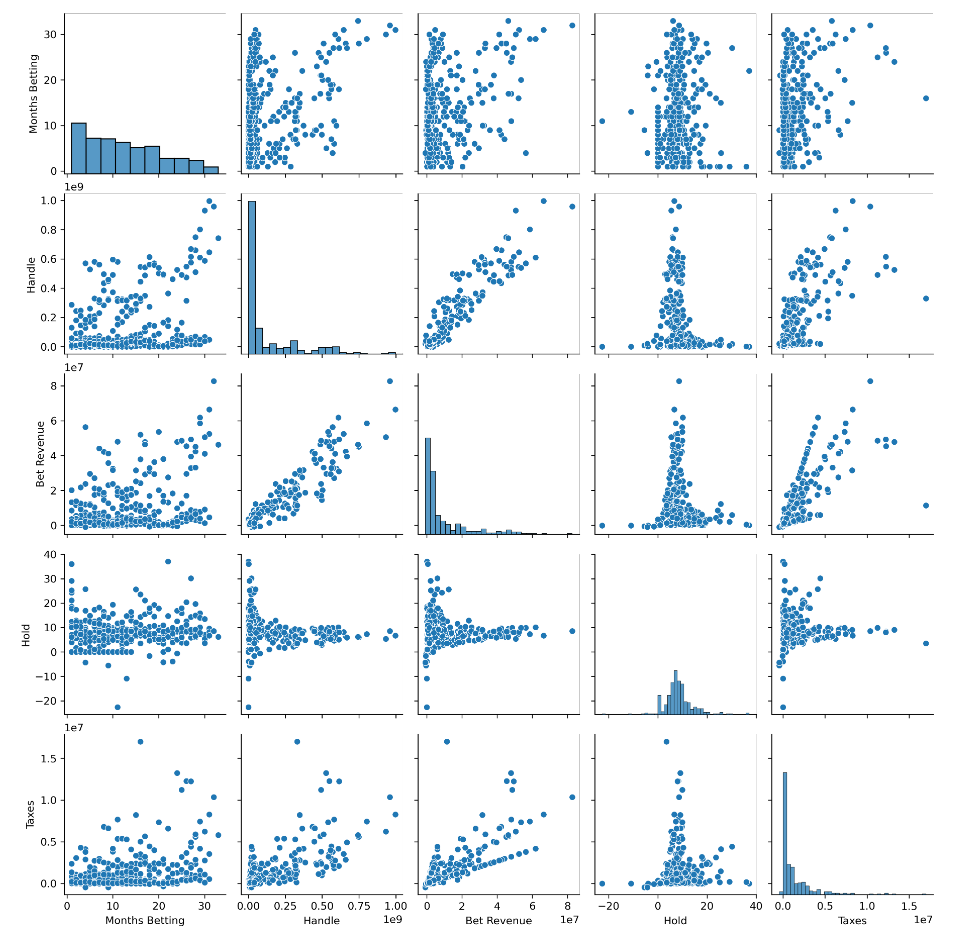
The first visualization story shows the states that have legalized sports betting and their percent of taxes from betting for 2018, 2019, and 2020. This is a very important visualization because it shows the enormity of the impact that Sports Betting can have on a state’s budget. Currently, Delaware has the highest percentage of taxes from betting. Delaware taxes the winnings on sports betting at a rate of 50%. This generates a significant amount of revenue for the state and its budget. As new states come online there will be an opportunity to generate significant tax percentages based on each states’ policies. While these numbers appear small at the moment, we are still in the very early stages of states taking full advantage of these new income sources and would expect these numbers to keep growing at a rapid pace for all states.

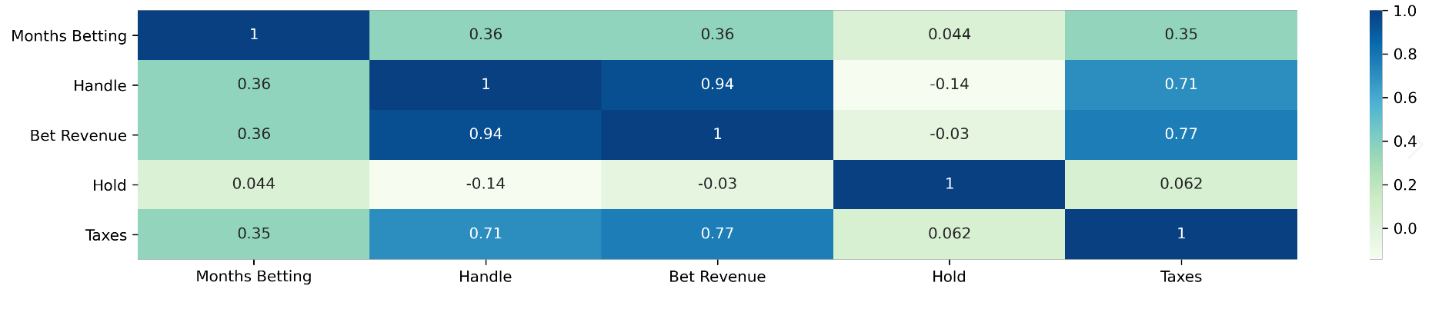


A state population is a significant variable in the regression model. Age demographics are key in predicting the overall amount that will ultimately lead to tax revenue for these states. The demographics for age are split into the following categories for ages 0-18, 19-25, 26-34, 35-54, 55-64, and 65 plus. Maine is the state with the highest percent of people in the 65+ category. Most age categories are statistically significant in our regression model when predicting bet revenue, but some of the ranges have a negative impact on the State’s bet revenue. States that are still deciding on whether to legalize sports betting will have plenty of data to determine which demographics contribute. Gambling hotlines will be implemented in most states as sports betting is introduced, and the age demographics could help to target the ranges that might benefit from that information.

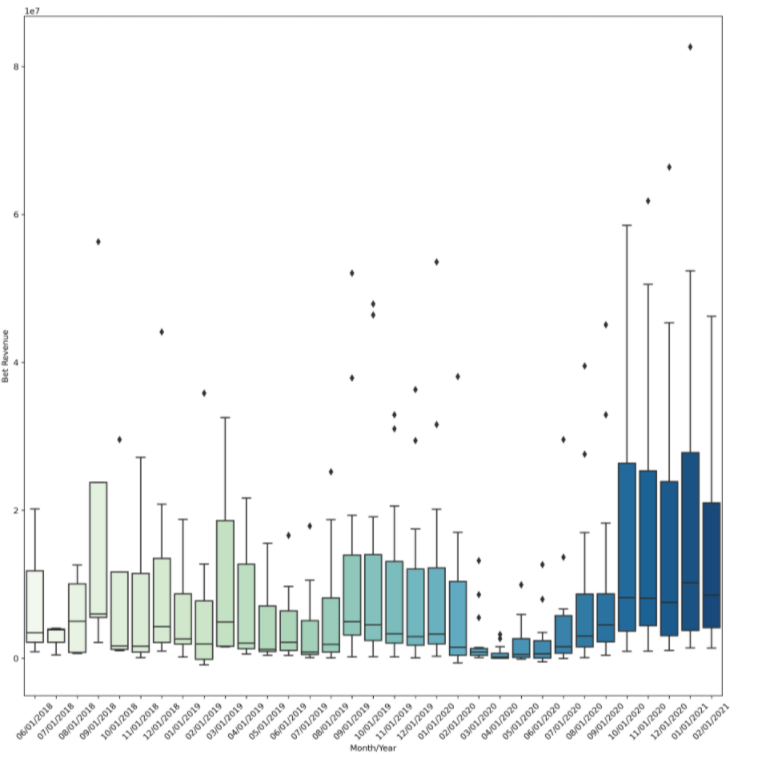


The growth of Bet Revenue for the states that participate in sports betting can be seen in the visual above. New Jersey shows that since 2018 there has been continuous growth month over month in Bet Revenue. The Covid-19 pandemic is seen from around March 2020 to May 2020 which affected the sports betting industry, but growth has increased even more from that point onwards. This visual portrays the story of the recovery that occurs post Covid-19 and the emergence of new states post-Covid-19. When examining the data from a visual perspective, this visual helps to lead to a realization that predicting Bet Revenue can be challenging due to the number of months of data for the new states, the Covid-19 pandemic occurring right in the middle, and the massive growth is seen month over month in Bet Revenue as it gains traction in the respective states.

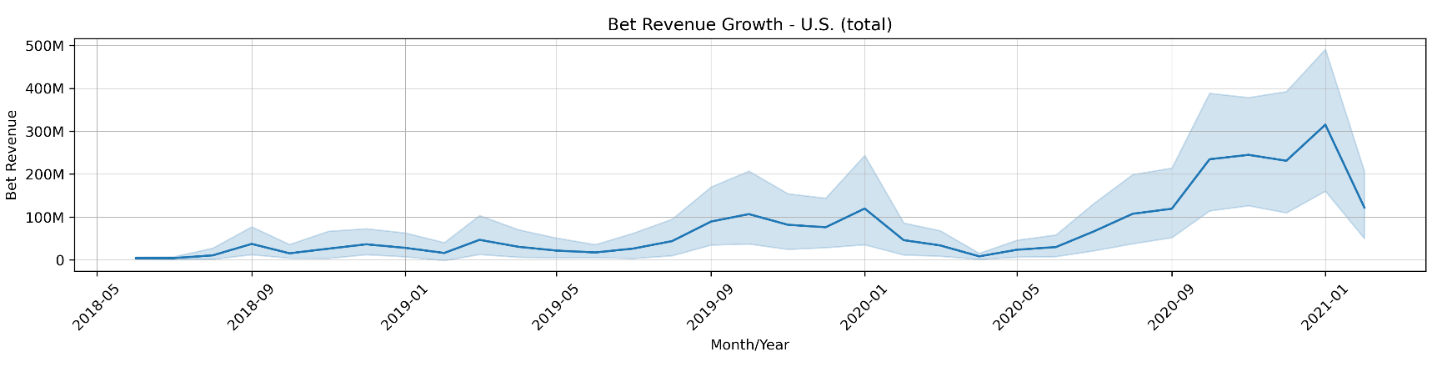




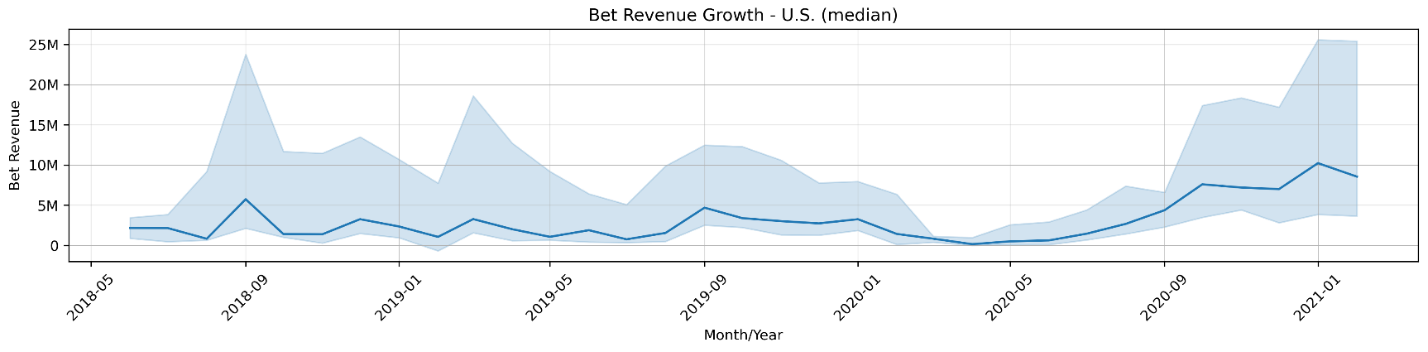
The pairwise scatter plots and correlation matrix can be used to illustrate the relationships between target and predictor variables. The relationship between Bet Revenue and Taxes is positive as Bet Revenue goes up the collected Taxes goes up. This will be the most important takeaway from the relationships as states examine the implications of legalizing sports betting.



This plot shows how total Bet Revenue varies over time from legalization through February 2021. This helps to illustrate and again, reinforce, our point of how volatile the industry is in terms of its growth, and unexpected decline due to outside factors. Since the end of the stoppage of sports due to the pandemic, growth has been exponential. Month over month the average sports betting revenues continue to climb higher.



Growing Bet Revenue can be seen over the time period of May 2018 to February 2021. During this time the total bet revenue growth gained in value as new states came online. The total growth of the betting revenue started in the tens of millions and is now approaching half a billion dollars. As more states join in, this can become a massive tax revenue generator across the US.



Median Bet Revenue is shown above and can be seen trending upwards over time. Once again the implication from Covid-19 are shown, but as things re-opened the growth takes off and is higher than the pre-pandemic era.

# Modeling Techniques

To create multiple predictive models for additional revenue generated by the legalization of sports betting, several techniques can be utilized. Multiple Regressions, as well as basic Neural Networks and decision trees, are utilized with techniques capable of handling the data to create a model with the training data available and allow for predictions to be made for states without current sports betting legalization.

## Multiple Regression

There are two main advantages to analyzing data using a multiple regression model. The first is the ability to determine the relative influence of one or more predictor variables on the target value. The second advantage is the ability to identify outliers or anomalies.

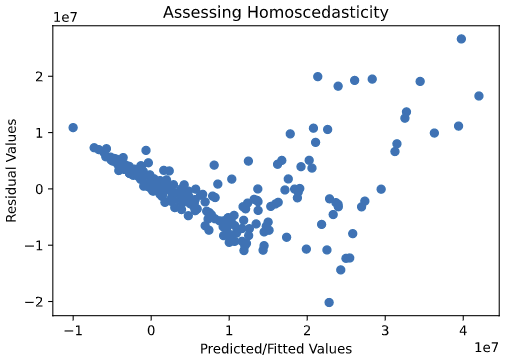
The disadvantages of using a multiple regression model, which is typical for most models, depending on the data being used. Two examples of this are using incomplete data and falsely concluding that a correlation is causation.

The ability to observe the relative weight for each variable makes regression invaluable. This will help states evaluate the importance of some predictor variables that involve decision making such as whether to permit online betting as an option.

There are four assumptions associated with a linear regression model:

**Linearity**: The relationship between X and the mean of Y is linear. This is evaluated by plotting a predictor variable against the target variable. This test is provided in the paired plot display in the “Visualizations” section.

**Homoscedasticity**: The variance of residual is the same for any value of X. This is evaluated by plotting residuals and/or standardized residuals vs fitted values from the model. This plot also assists in checking for linearity. A non-constant variance test can also be performed.

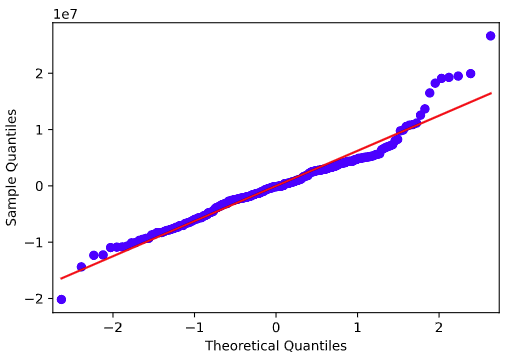


The variance of residuals for the model has a ‘bowed’ shape that suggests a non-linear relationship between predictor and target values.

**Independence**: Observations are independent of each other. This is evaluated with a Durbin Watson Test to look for autocorrelation.

This value for the Durbin-Watson test suggests there is no autocorrelation in the sample.

**Normality**: For any fixed value of X, Y is normally distributed. This is evaluated using a Q-Q plot by comparing the residuals to ‘ideal’ normal observations. We can also look at the distribution of studentized residuals.



The tail ends of the Q-Q plot deviate from normal.

All of these assumptions were assessed in python using graphs and the test statistics printed out in the regression summary. Choosing a regression model allows us to have a good “baseline” for performance and the additional benefit of being able to see what variables are importing influencers in predicting our target variable. It is a robust model and allows us to get an insight into the data.

## Neural Network

The main advantage to analyzing data with a neural network model is that neural networks can handle nonlinear data with a large number of inputs without making assumptions about the underlying pattern for the data. Also, the learning methods are robust to noise in the training data, making the final output more reliable even with errors in the input data.

The biggest disadvantage to a neural network is the unexplained functioning of the network. The model functions as a ‘black-box’ with only input and output values, with no discernible reasoning given without also creating a surrogate model to help decipher with a best guess.

An assumption we have to make for the model is that the activation function is accurate for the actual formation of the output values. This is evaluated by creating multiple models and testing with validation and test data sets to compare accuracy.

Due to the exponential growth sports betting revenues have recently shown, along with the unprecedented decline in revenues for a short period in 2020 due to the pandemic, we strongly believe that our data is currently exhibiting a non-linear trend, which makes a neural network a great choice to try and model our data.

## c. Extreme Gradient Boosting regressor (xgboost)

The XGBoost model is an implementation of gradient boosted decision trees designed for speed and performance. The “gradient” in the model refers to a gradient descent algorithm to minimize the loss when adding new models. Boosting is an ensemble model technique where each time a model is added, it corrects errors made by existing models, ultimately added together to make a final prediction.

The advantage of an XGBoost model is both speed and performance. Although our dataset is not particularly large, we were curious to analyze how a model such as this would perform. This method handles missing data and regularization to avoid overfitting/bias (something not provided by previous gradient methods).

A disadvantage of XGBoost is that one has to manually create dummy variables and/or label encoding for categorical variables before feeding them into the model. Another disadvantage is that it requires more involved hyperparameter tuning to tune the model properly.

# Data Splitting and Sub-Sampling

## Justification and Explanation

The total data set was randomly split in a 75-15-10 percent fashion between training, validation, and testing data sets respectively. The total data set is only 314 rows of data meaning the training set needed to be relatively large to maintain adequate data for model creation. This does leave the possibility that the validation and testing sets could be dominated by specific states if the split is not evenly stratified. A split comparison was created below to help indicate any potential significant differences in the data sets. Using the random state during the split allows for repeatable results. The training data set is used to train all potential models while the test and validation data sets are used to assess and compare the model performance.

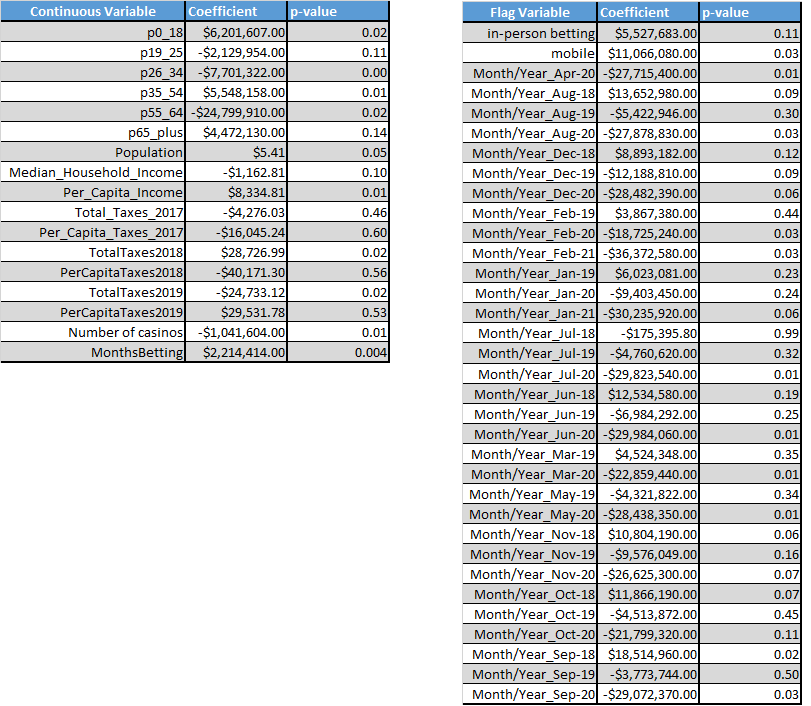
## Split Comparison



Originally, with all data included, potential significant differences between the training, test, and validation data sets were observed. Nevada does not follow the typical progression that recent history has displayed with the legalization of sports betting for all states. With such a small data set and short history overall of legalized sports betting, Nevada’s long-term experience disrupts the models and creates significant differences between data sets if it is not uniformly included or not included across all data sets. With the removal of Nevada, the three data sets are similar. Since the validation and test data appear to have a potentially significant difference in mean for ‘Bet Revenue’, a t-test was performed. The resulting p-value of 0.19 suggests that we cannot reject the null hypothesis that the averages are the same.

# build models/Model Results

## Regression

Ordinary least squares is the statistical method used to estimate the parameters of the linear regression model. Below are the coefficients and p-values calculated for the used independent variables. Overall, the model was statistically significant with a p-value of 4.58E-33 with an F-statistic of 10.64.

The continuous variable coefficients show how Bet Revenue would change, in a positive or negative adjustment, based on the value for each of the variables. For example, the model suggests that Bet Revenue can be increased by $5.41 for every person in the population. The flag variables are simple adjustments based on status and do not scale. This model shows a benefit to having both in-person and mobile betting. Many of the p-values for our continuous variables are significant with an alpha of 0.05.

Since the values for each variable will have different scales, the most impactful variables are not evident. We can standardize the variable ranges then prepare the model again to view the standardized coefficients.



By utilizing the standardized coefficients, we can compare across the continuous variables to see the most impactful in total, or a subset i.e. population age groups. It appears that per capita income and taxes are the most significant variables when it comes to generating tax revenue from sports betting. More importantly, it shows that the 65 plus age group has the largest positive impact among age groups when standardized. The MonthsBetting variable being 1.50 indicates that the longer a state has gambling, the more money they will make every month. This relatively high value of this variable leads us to believe that this is still an extremely fast-growing industry and also indicating, that we have not yet reached the “peak” of the amount of bet revenue a state will get in a month. This leads to modeling challenges, but with more data as time goes on, this number will probably decrease once the market becomes saturated. In-person and mobile betting also display large positive numbers. This indicates that any new states wanting to implement sports betting should strongly look at adding a mobile betting option. The ease of access to mobile betting is one way to quickly increase bet revenue numbers for a state.

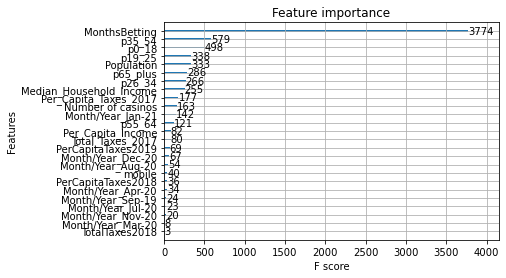
## Neural Network

As a whole, the neural network model is considered a black box for diagnostic and interpretable outputs. We can determine the parameters and bias values for the different layers, but those values do not have a managerial significance for insight. The usefulness of the model must be determined in the validation and testing assessment. To build our neural network model, the first step was to scale all of the predictor variables for use in the model. Many different iterations of models were tested with different solvers, activation functions, and the number and size of hidden layers. The model that we landed on had 3 hidden layers of 100 neurons each, adam solver, and relu activation function.

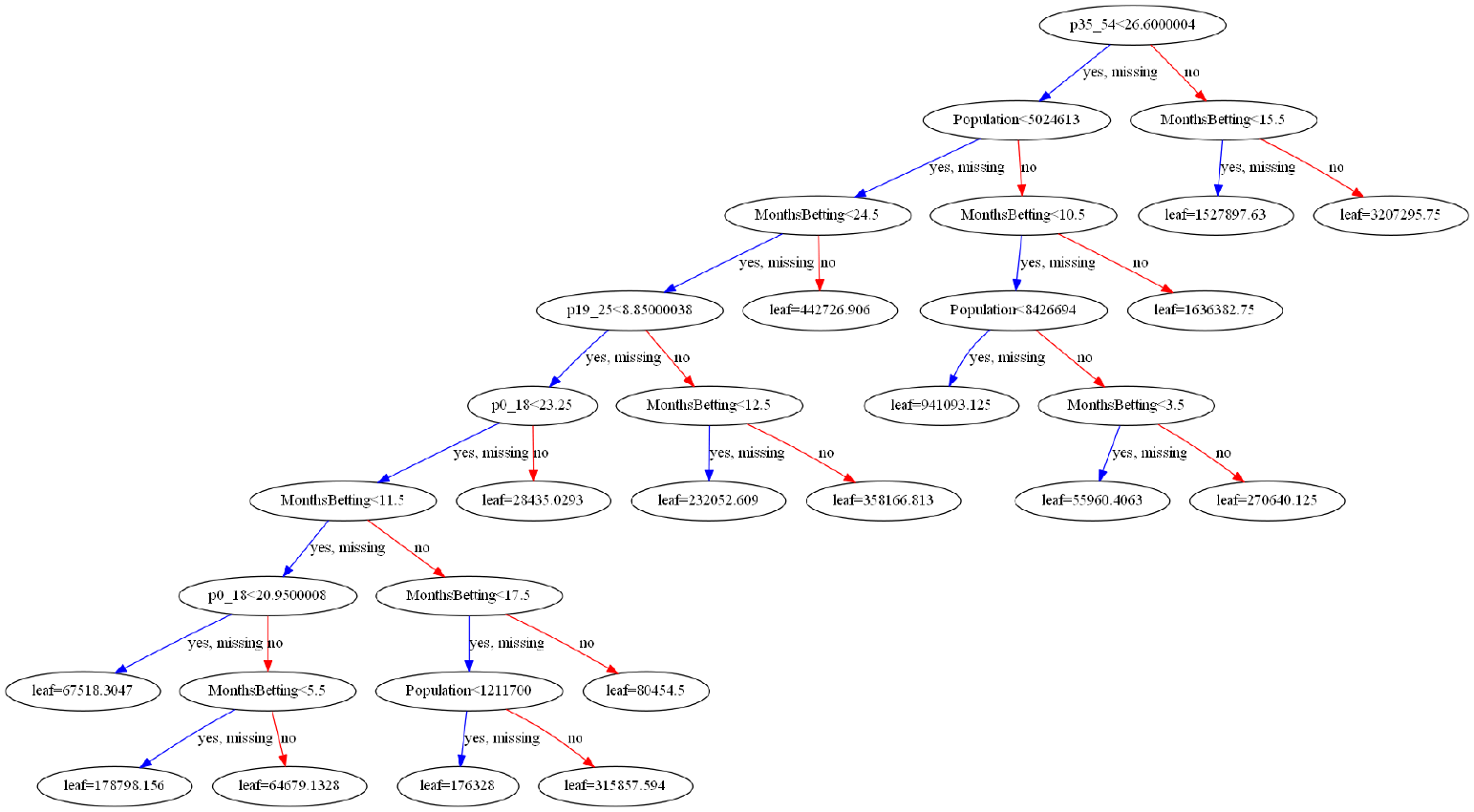


## C. XGBOOST

By analyzing the XGBoost model feature F-Scores, we see that *Months Betting* is the most significant feature, weighting a much higher score than any other variable. Other notably significant variables are the age-based variables as well as *Population*.



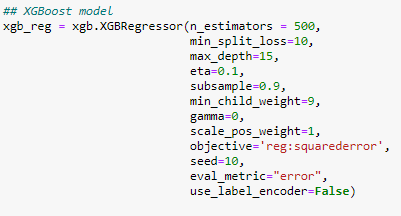
Below is the tree generated by the model. We can see that based on the feature importance plot above, this is also how the tree split. *P35­\_54* (age bracket) is the first split, with the subsequent splits consisting of either *Months Betting* or the other age-based variables.



Part of some of the elusiveness of a gradient boosted model is the hyperparameter tuning taking place. Below, you can see the parameters defining the model. Notable parameters include the tree-based parameters, *n\_estimators, min\_split\_loss,* and *max\_depth.* The most sensitive parameter in terms of weight on the model scoring is *min\_child\_weight*. According to the XGBoost documentation:

“*minimum sum of instance weight (hessian) needed in a child. If the tree partition step results in a leaf node with the sum of instance weight less than min\_child\_weight, then the building process will give up further partitioning. In linear regression mode, this simply corresponds to minimum number of instances needed to be in each node. The larger, the more conservative the algorithm will be.”*

We discovered that minor adjustments to this parameter caused major over/underfitting in the model. As stated in the documentation, the larger the value, the more conservative the algorithm will be – a larger value indicating a higher weight on children and less sensitivity in the model. We chose these parameters as a trial-and-error exercise to create the best R-squared value with the least amount of overfitting.



This model produced R-squared values for training, validation, and testing of .993, .793, and .745 respectively. This indicates that the model is performing at an adequate accuracy level. At first glance, there appears to be some slight overfitting; however, the .793 R-squared for validation indicates that we are still achieving acceptable goodness of fit.

# Model Assessment

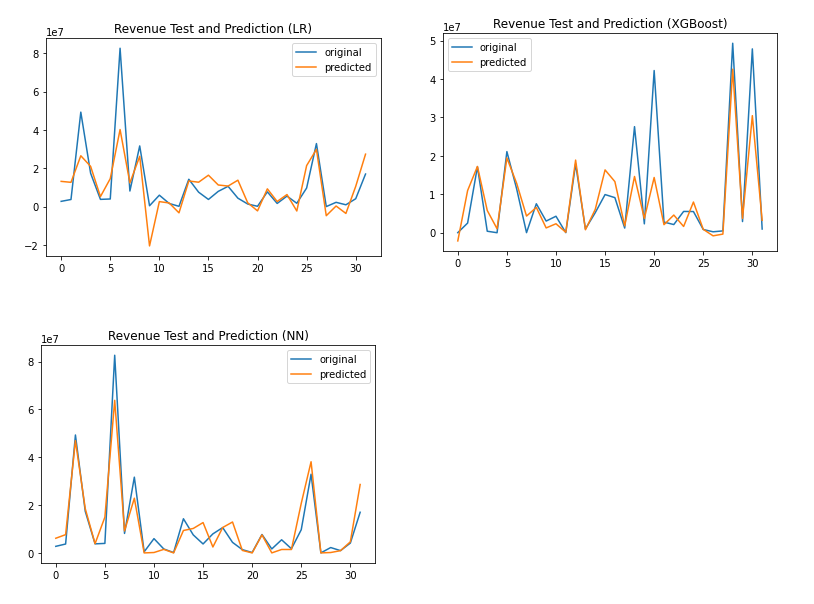
## Assessment Choice

Since the objective is to create a model to most accurately predict sports betting revenue, the r-squared value allows for comparison between models’ performance. The r-squared value is the percentage of the variance explained by the collection of predictor variables. We can perform predictions with the validation and test data sets and compare them to the true values to determine if the model generated continues to explain variance on unseen data.

## Model Comparison



The models were compared on r-squared values for all 3 data sets. The regression model is the lowest overall model on all metrics. The XGBoost model performs the best on the training data, but significantly worse on the validation and test data. This is probably an indication of this model overfitting to the training data. The Neural Network model performs slightly worse than XGBoost on training data, but much closer to the training data on its validation and test data sets. This indicates that this model is not overfitting as much to the training data as the XGBoost model. MAE and MSE follow similar patterns to r-squared and were included as an additional metric to assess models on.



The prediction plots for each of our models help reinforce the points made above for each model. The test data was plotted along with the predicted values for each point in the test data sets. The deviation between the two lines at a given point shows the residuals (actual-predicted) and the closer the lines are to each other indicate a better prediction. The Neural Network picks up many more of the nuances than the other two models. The regression model tends to underfit on most points and the XGBoost model does well on predicting the lower value bet revenues but struggles where the actual bet revenue for a month was very high.

## Final Choice

All three of our predictive models exhibit results that are useful in their own ways. The regression model allows us to see the important variables in predicting Bet Revenue, even if it does not perform the best. The Neural Network model allows us to get better results based on the non-linear trend of our data. The XGBoost model allows us to develop “rules” to quantify how much Bet Revenue will be generated. Based on these factors, and the ones listed in the model comparison section, we have selected the neural network model as our best model. It is hard to ignore the much better r-squared performance on our validation and test data when compared to the other models. With our project goal of showing states how much-increased tax revenue they can get from betting revenue if legalized, we feel it is more important to show the model that is the most accurate, as compared to the two models that give insight into what drives the higher Bet Revenue numbers, at this point in our analysis.

# Predicting Bet Revenue for states without legalized gambling

We used our neural network model and predictor variables from states with no sports betting to see what kind of monthly revenue these states could get for legalizing sports gambling. A few assumptions had to be made to achieve this. We assumed that all states had legalized sports betting 3 months prior, allowed both in-person and mobile betting, and that we were predicting what revenue would have been for January 2021. The results are in the graph below. California and Texas would generate the most revenue for their states. Oklahoma would generate around $18 million a month in bet revenue based on our model which would lead to a significant increase in tax revenue for the state over time.

