**Allan Flores**

**Syracuse iSchool**

**MSc. In Applied Data Science**

**SUID: 833497871**

Github Link

**PORTFOLIO MILESTONE**

**PROJECTS**

1. **DATA VISUALIZATION**

Project Title:

INBOUND CROSSING AT THE US-MEXICO AND US-CANADA BORDER

Background:

The main goal of this project is to provide insight on some questions in relation to people crossing the US-Canada and US-Mexico border from years 1996 to 2018. The idea is to present a poster that can walk through readers some questions showing through visual communication. Moreover, some readers are drawn to paying more close attention to colors, maps, graphs, and some facts which the poster which reflect on. The output can also be educational to students who are interested at a quick glance some facts on immigration which is one of the political highlights of the country.

Data Questions:

How many have entered the United States from the Canadian and Mexican ports from Y1996 to Y2018?

What is the split of border crossing coming from Mexico and Canada?

Which states do people cross from?

What are the top 10 port of entries?

What are the mode of entries?

How many pedestrians have crossed the border? What is the trend?

Which states are pedestrians crossing from?

What are number of pedestrians crossing by port?

Data Source:

* Primary source of dataset is from Kaggle competition titled “Border Crossing Entry Data.”
  + https://www.kaggle.com/akhilv11/border-crossing-entry-data

Pre-Processing, Tools and Techniques:

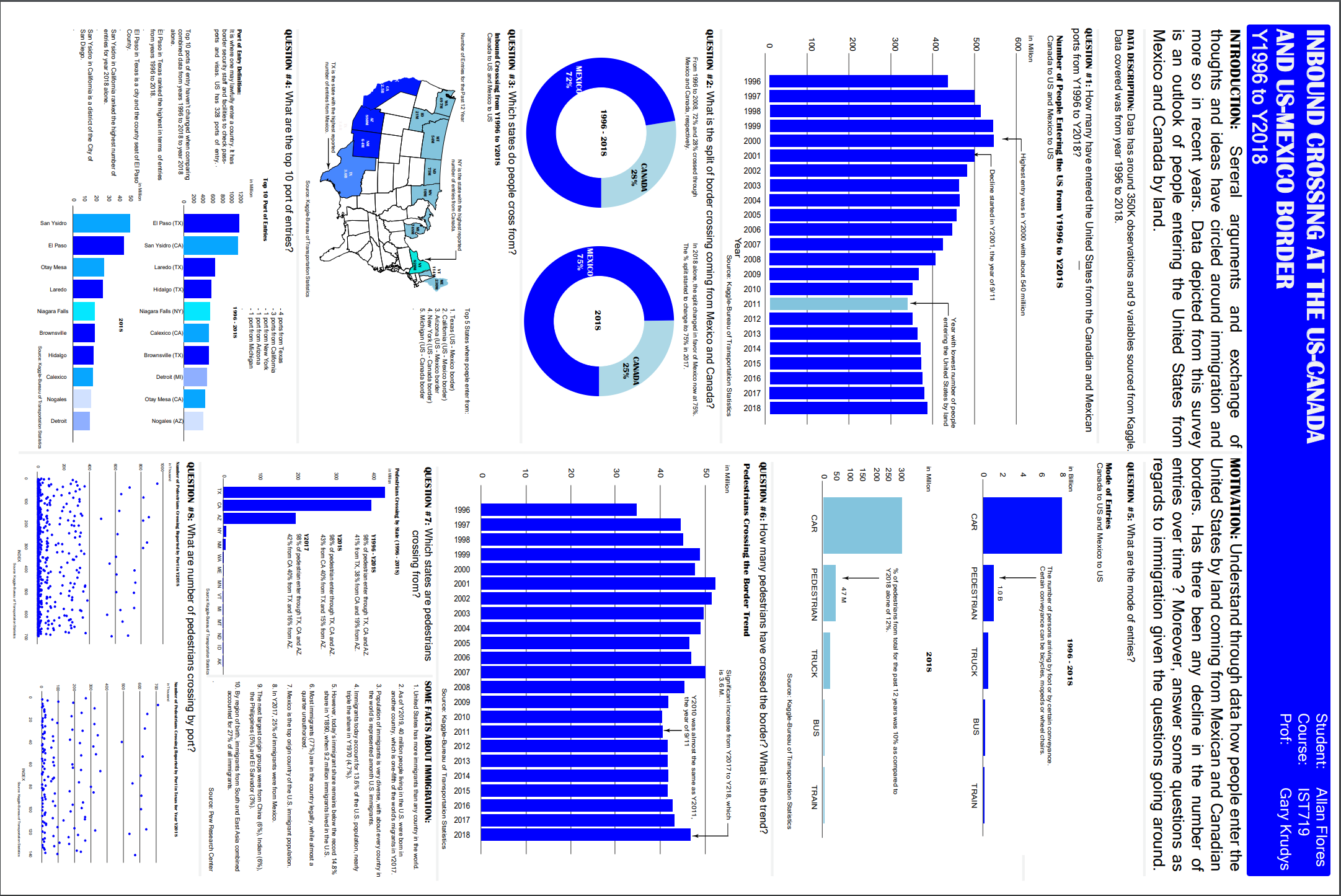
No observations were taken out from the original dataset. Used packages/libraries like gplots, dplyr, tidyverse, ggplot2, ggmap, maps, RColorBrewer, mapdata, and viridis. In using the “ggmap,” it was required to use the Google API. Adobe Illustrator software was used in designing and producing a posted. In the pre-processing, some descriptions were either shortened or abbreviated.

Insights Gained:

Conveying messages with data science can also be enhanced through other ways like using the Adobe Illustrator to produce a poster that can be readable to readers especially if the objective is to catch more attention. Some readers interpret or derive information through data but some are through visuals. The presentation of a poster can address both with the flexibility of changing some of the features and colors. It widens the range of potential readers of a particular report.

Excerpts from the Output:

Below was the final poster produced as requirement of the project with the aid of Adobe Illustrator.



1. **BIG DATA ANALYTICS**

Project Title:

KOBE BRYANT SHOT PREDICTION AND FINANCIAL IMPACT TO LAKERS

Background:

Sports analytics has become a popular topic in the data science field where the term has been popularized with the release of the 2011 film, Moneyball. The game of sports maybe different among popular sports of baseball, hockey, football, and basketball, but the underlying principle governing the idea of analyzing and predicting outcome in sports performance is identical.

Kobe Bryant was an icon in the history of basketball and we want to leverage on data collected from his entire career, together with his team, the Lakers. Moreover, look into the financial impact of Kobe to the Lakers.

Data Questions:

Using Kobe Bryant’s stats throughout 20 years of his NBA career, we want to build several models to predict whether his shot will make or miss it.

The other part of the project is to determine if there is any correlation between Kobe Bryant’s performance and the revenue and profitability of the NBA team LA Lakers using exploratory and/or Machine Learning Techniques.

Data Sources:

Part 1: Analyses of shots made by Kobe Bryant (sourced from Kaggle dataset)

* Primary source of dataset is from Kaggle competition titled “Kobe Bryant Shot Selection, for the player’s statistics.
  + <https://www.kaggle.com/c/kobe-bryant-shot-selection/overview/description>.

Part 2: Analyses of Kobe’s performance in relation to team’s performance, revenue and profitability

* Annual Income and Expense data
  + From Rodney Fort’s sports financial data.
  + <https://sites.google.com/site/rodswebpages/codes>
  + <https://drive.google.com/drive/folders/1pr_yPm9oPLcfCtWOtnrLoJdvzFDbYAAk>
* Annual Ticket Price and Fan Cost Data
  + From Rodney Fort’s sports financial data.
* Lakers Attendance
  + Data provided by the Association of Professional Basketball Research
  + <https://www.apbr.org/attendance.html>
* Lakers Annual Record
  + Sourced from Sports Reference API, Robert Clark
  + https://sportsreference.readthedocs.io/en/stable/nba.html
* Lakers Team Data
  + Sourced from Sports Reference API, Robert Clark
  + <https://sportsreference.readthedocs.io/en/stable/nba.html>
* Player’s Career Data – Kobe Bryant Data
  + From sportsreference.nba.roster import Player
  + Selected player, “bryanko01” to extract Kobe’s statistics

Pre-Processing,Tools and Techniques:

Part 1: Analyses of shots made by Kobe Bryant (sourced from Kaggle dataset)

* Removed from the dataframe for column, “shot\_made\_flag” with null values leaving us 25,697 shots from 30,697.
* This is 16% reduction from the original number of observations.
* Converted game date to date time variable for future analyses.
* Grouped game date at yearly level to plot the number of games per year.
* Used Linear Discriminant Analysis, Ada Boost Classifier, Logistics Regression, Radom Forest Classifier, Support Vector Machine and Kneighbors Classifier

Part 2: Analyses of Kobe’s performance in relation to team’s performance, revenue and profitability

* Combined all relevant team, and Kobe data through merge function
  + Operating income and franchise value
  + Fan cost and ticket price index
  + Attendance
  + Record (wins, losses, wins ratio)
  + Team data (team statistics)
  + Kobe data (Kobe’s statistics including his salary)
* Generated a normalized data frame to be used for some analyses
* OLS Regression

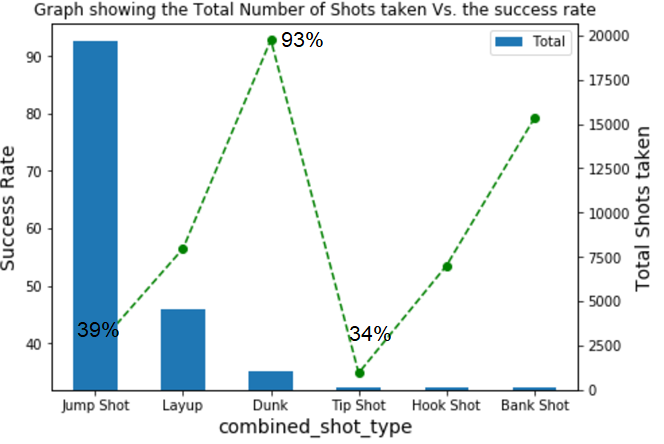
Insights Gained:

Kobe’s jump shots with majority of the shots only had 39% success rate, while dunk had the highest success rate at 93%. His shots were more successful when executed from the center of the court.

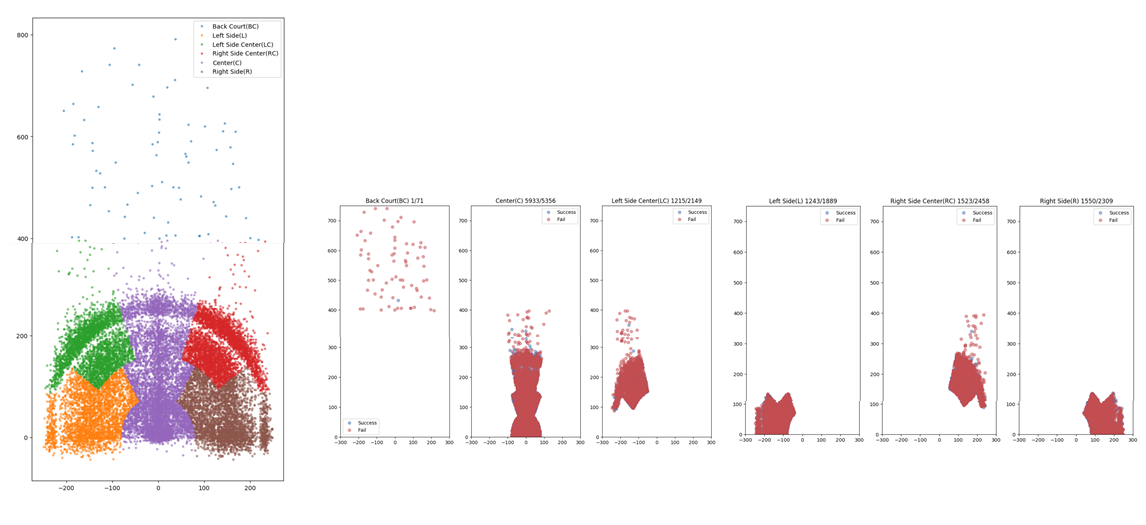
Ticket prices and general fans costs were positively correlated with Kobe’s salary. Attendance grew sharply with the opening of Staples Center, and held flat throughout Kobe’s career. Kobe created a “surplus” of $3M during 2000 – 2001 season, winning the NBA championship. The surplus Kobe created started to decline in Y2002 until his retirement, but Lakers franchise value started to increase in Y2010.

Excerpts from the Output:

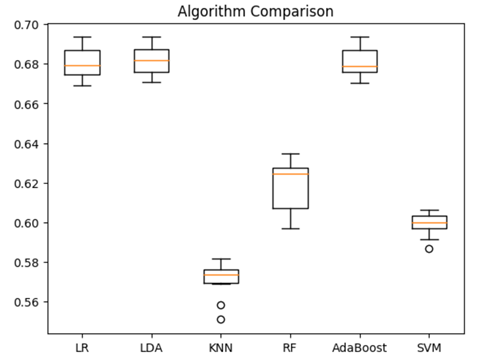
Majority of the shot type taken by Kobe was a jumpshot with 39% accuracy :

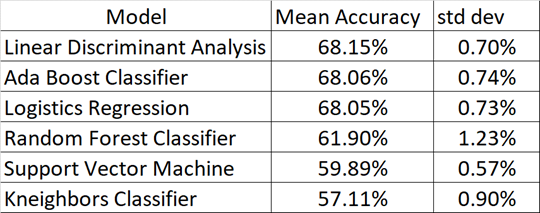


More successful shots were taken from the center of the court:

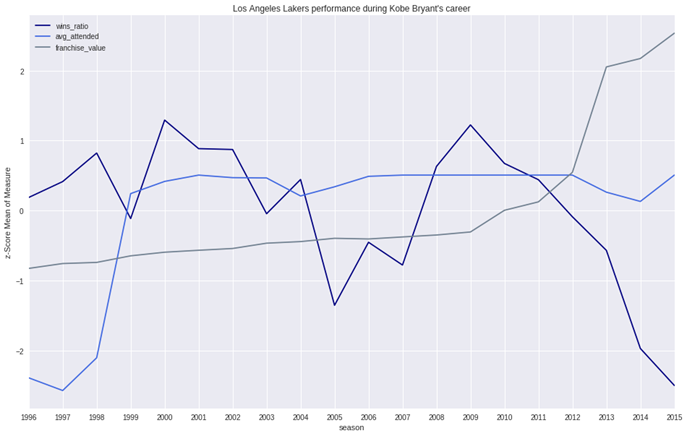


Shot prediction models:





Lakers performance during Kobe’s career:

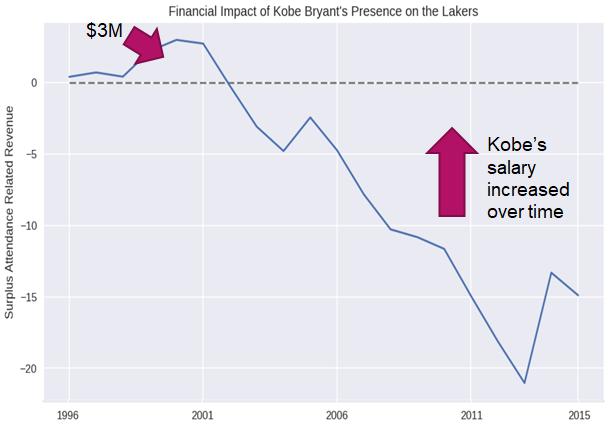


Financial impact of Kobe Bryant’s presence on the Lakers:

Given that Kobe’s play, especially defensively, made an impact on the overall team performance **and**  overall team performance draws attention to the franchise

(ex. attendance, merchandise sales, etc.):

* The financial impact of Kobe Bryant is being assessed as the difference between expected and actual attendance, weighted by Kobe’s use throughout the season (minutes played, injuries, etc.), and reduced by Kobe’s salary.
* This measure shows a net gain for Lakers from 1996 through 2002, then declining through the end of Kobe’s career as his salary increases.
* Even though the Lakers won five championships throughout Kobe’s career (2000 - 2002, 2009, 2010), an attendance capacity diminishes the financial impact of one player.
* Albeit, this is a conservative view of a player’s financial impact.



1. **TIME SERIES FORECASTING**

Project Title:

Recommend Three Zip Codes to Invest

Background:

Given the available resources to invest in real estate requires scrutiny of information and data. Some of the factors may include goals with the property, the real estate market, income stability, and life situation among other things. There are several uncertainties in the market but one can use his or her judgment by looking at historical data as one of the bases in making a decision.

The scope covers recommendation of the three zip codes for the Syracuse Real Estate Investment Trust (SREIT) based on the median home values from 1996 to 2019. Crime rates and median household income were taken into consideration in filtering the desirable zip codes for potential investments.

Years of historical data until 2019 will help up assess the trend over a period of time and generate a forecast for Y2020. Using the Prophet (fbprophet) model as forecasting technique will help us assess the best zip codes based on several assessment parameters.

Data Questions:

1. What 3 zip codes provide the best investment opportunity for the Syracuse Real Estate Investment Trust (SREIT)?

Data Sources:

Three sets of data were used in delivering the requirements for this exercise.

* Zillow Median Home Values:
  + Data from files.zillowstatic.com/research/public/Zip/Zip\_Zhvi\_SingleFamilyResidence.csv”
* Median Household Income for Y2018:
  + Extracted manually from https://www.ers.usda.gov/data-products/county-level-data-sets/download-data/
* Crime Rates and Population:
  + Extracted from Kaggle with Y2016 crime rates by county in the United States. Title from Kaggle: “United States Crime Rate by County”

Pre-Processing,Tools and Techniques:

* Dropped Y1996 data from the Zillow Median Home Values data.
* Merged data frames of Zillow, Crime Rate and Median Household Income.
* Keep only the rows with at least 60% non-NA Values.
* Removed rows with crime rate of at least 1.27%.
* Removed rows with median household income of at most $59,000.
* Generated the Top 10 zip codes based on highest growth rates from Jan 1997 to Dec 2019.
* Use fbprophet forecasting procedure for forecasting time series data implemented in Python.

Insights Gained:

Out of the Top 20 zip codes generated from the original Zillow data prior to down sampling, most of the zip codes or cities were coming from California. None of these cities came out from the final Top 10 zip codes as result of the data downsizing process. This was mainly due to the effects of crime rates and median household income criteria. Also worth mentioning that none of the selected Arkansas metro areas made it to the final Top 10 zip codes.

Using fbprophet as forecasting algorithm helped us use the outputs in determining the forecast in the next 12 months of Y2020. Main considerations were growth rates from Dec 2019 to Dec 2020, absolute change or increase in value and MAPE. The forecast error from a 365 days horizon were from 1.5% to 6% which was relatively an acceptable forecast error range. Although RMSE and MSE graphs were not shown from this report, their ranges were also relatively acceptable.

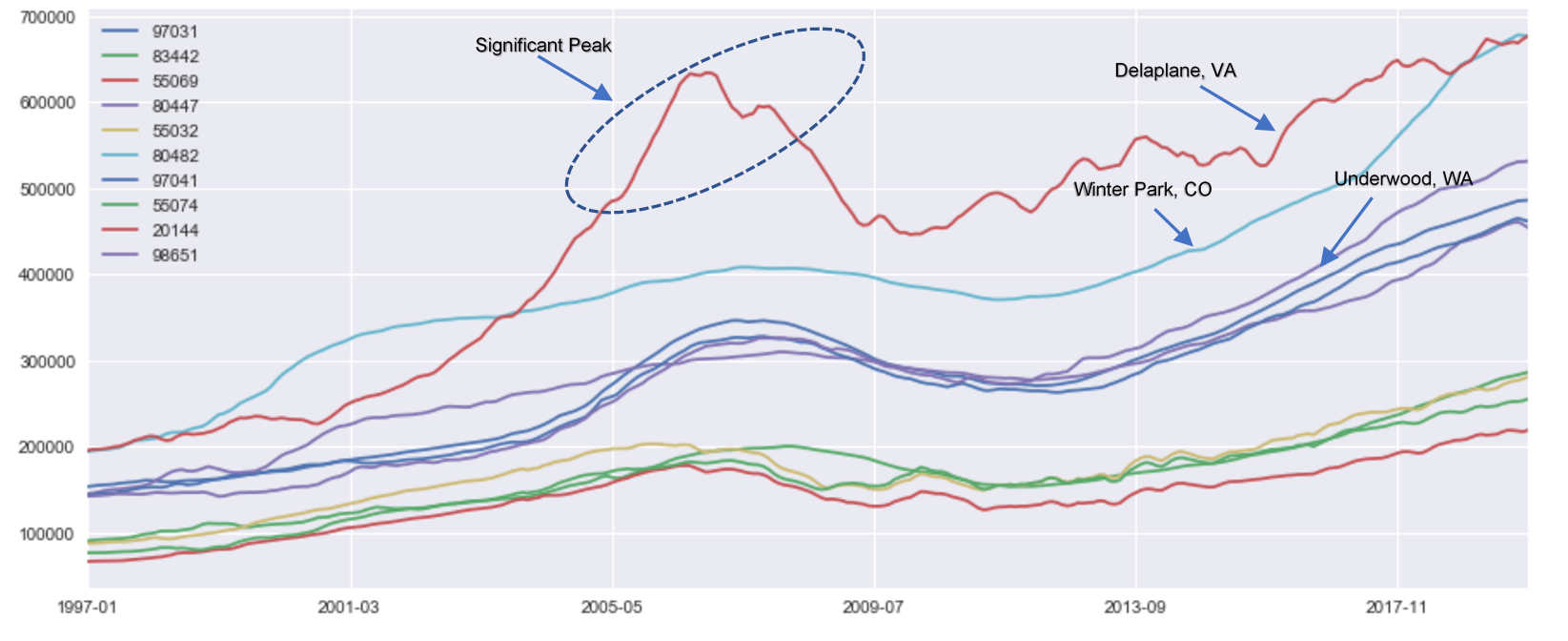
The final Top 3 zip codes recommended were 98651-Underwood, WA, 80482-Winter Park, CO and 97031-Hood River, OR. The combined growth rate from Dec 2019 to Dec 2020 was estimated to be around 9.75% and absolute change or increase of $165K in a 12-month period. In short, investing a total of $1,693K as of Dec 2019 median home value for the 3 zip codes was forecasted to be $1,858K by Dec 2020.

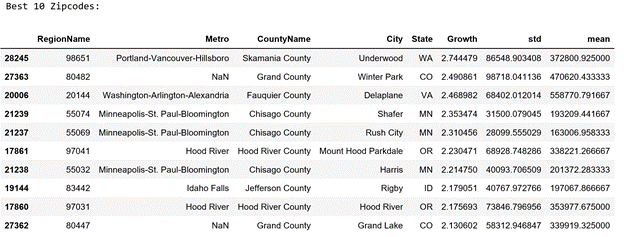
Moreover, as another option, we can replace 97031-Hood River, OR with 97041-Mount Hood Parkdale, OR, since their values were not far from each other. This may require further research or data should we entertain this route.

In the interest of time in getting more data or information, these analyses were scoped down to only using crime rates and median household income in filtering the data. Information like economic growth rates, unemployment rate, infrastructure development scoring, and investment scoring can be used. Moreover, recommending the Top 3 zip codes also depends on the available resources as a consideration that wasn’t provided for this case study.

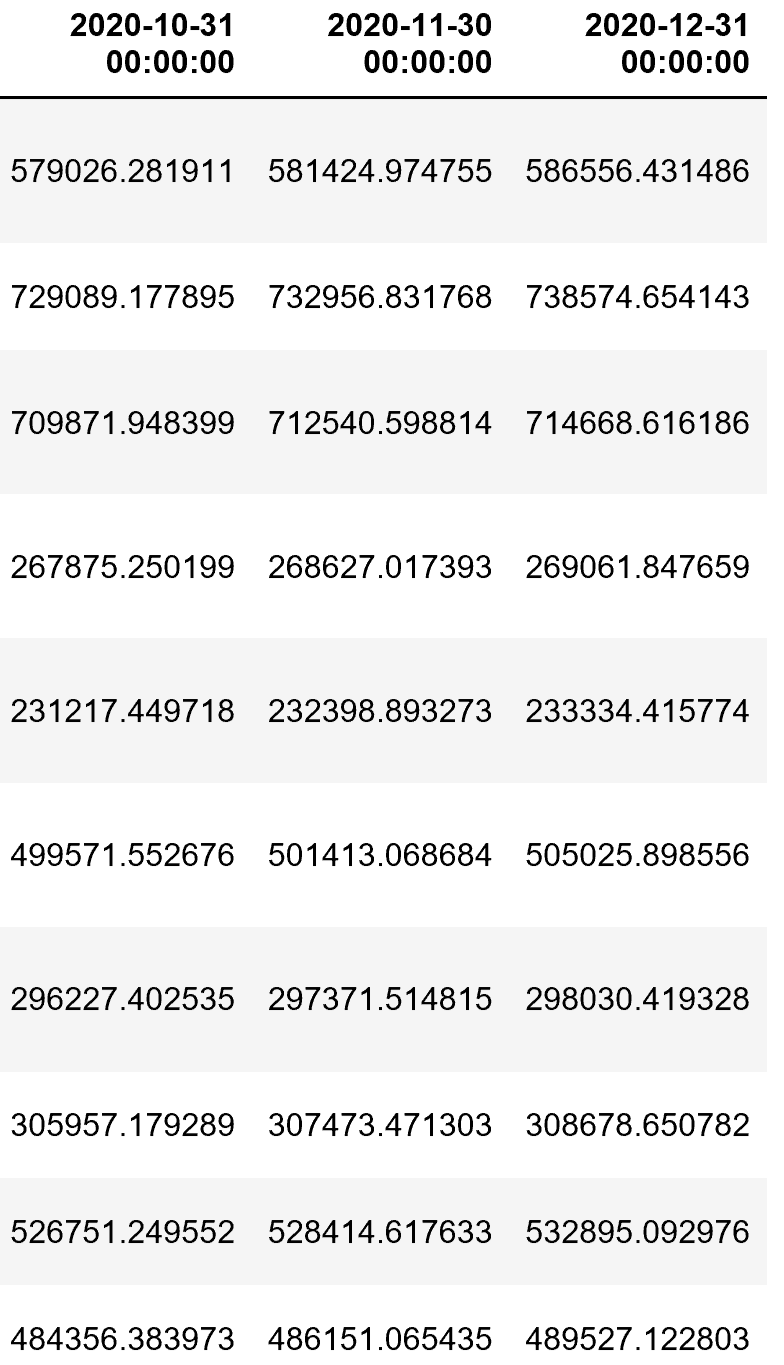
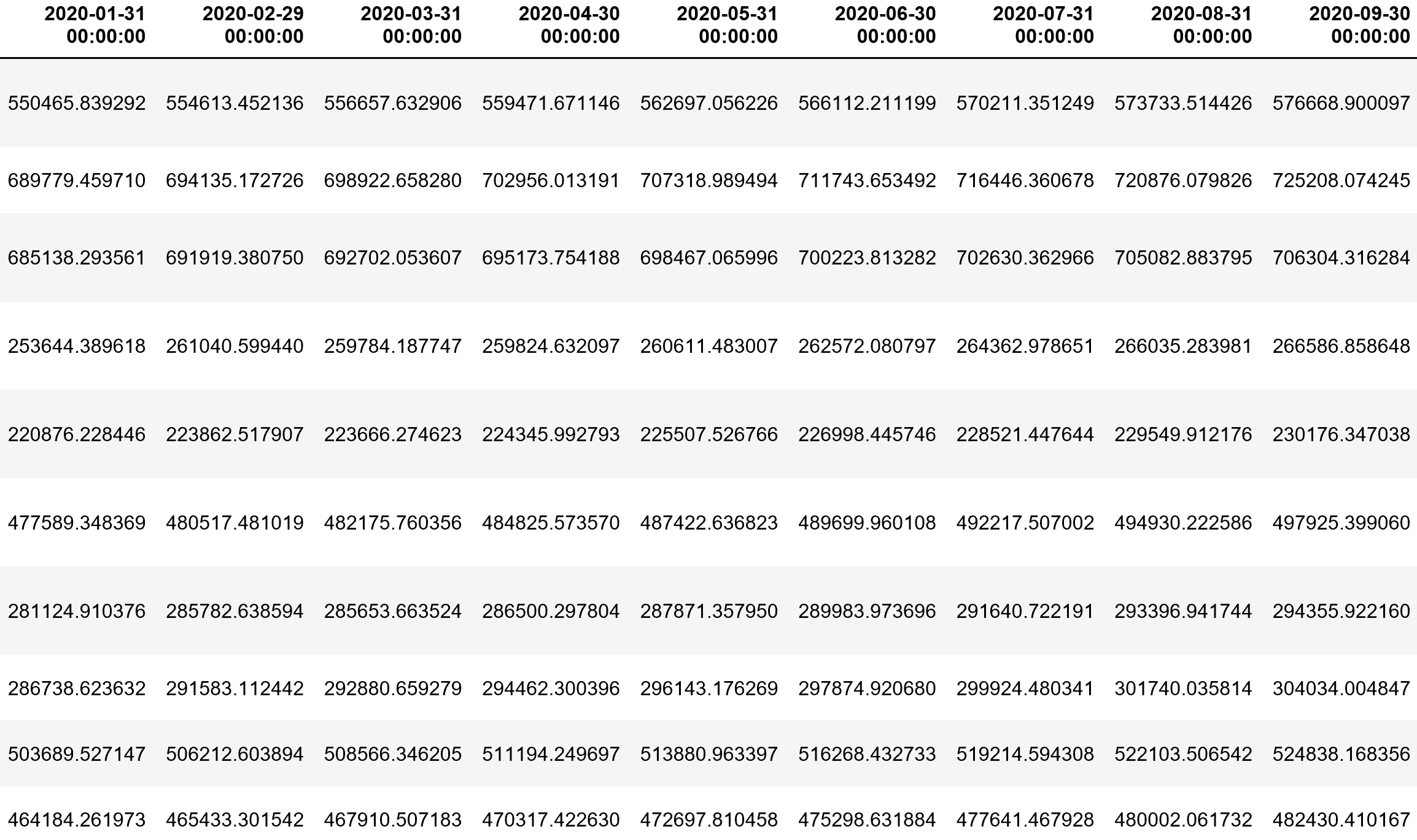
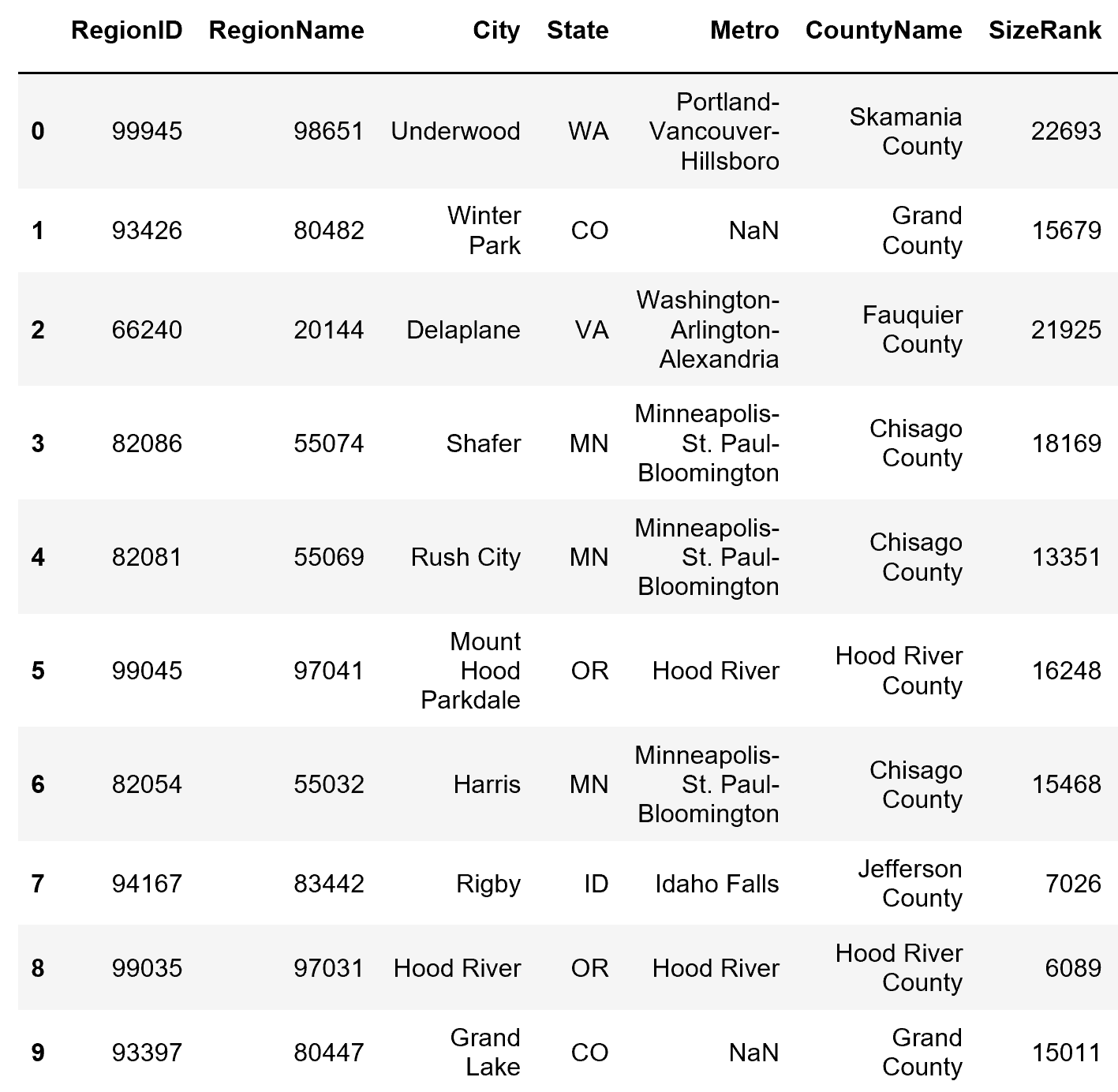
Excerpts from the Output:

Historical trend of Top 10 Median Home Values:





Year 2020 Forecast of the Top 10 Zip codes:



Summary of Results of the Top 10 Zip codes:

