Pt

SUID # **833497871**

Portfolio milestone  
MSc in Applied Data Science

Allan Flores

**TABLE OF CONTENTS**

GitHub Link…………………………………………………………………………………….3

Reflection………………………………………………………………………………………3

Projects

Data Visualization………………………………………………………………………..5

Big Data Analytics………………………………………………………………………..7

Time Series Forecasting……………………………………………………………….11

Conclusion……………………………………………………………………………………15

Video Link…………………………………………………………………………………….15

GitHub Link

**REFLECTION**

I have been a supply chain management practitioner covering demand and supply planning, purchasing, distribution and warehousing, transportation, and customer service for over 20 years having worked with multinational companies like Nestle, Novartis Pharmaceuticals, Sanofi Pharmaceuticals, and Mead Johnson Nutrition in both local and global settings. I have managed several projects with the end goal of always contributing to the growth and profitability with identified key performance indicators and process performance indicators. I have led implementations of integrated computerized planning systems like SAP, BPCS among other things and we have always looked into process improvements and optimization. It’s been a never ending routine, but we continue to explore other opportunities to make it better. This was when the idea of learning data science started.

The CEO of the company I work for encouraged me to take any course in data science so I can eventually start and organize the company’s Data Science department. I have searched for several training programs until I came across the MSc in Applied Data Science offered by Syracuse University. It convinced me after so much research about this field that it will be complimentary to my supply chain management experience. We always dealt with enormous amount of tables, data and variables but were not always confident of the right approach. We have so much data in SAP for instance, but most of the time clueless on how to use some of them. As an example, we always use MS Excel in crunching the downloaded data to derive some information from it, but we knew that efficiency and its value for decision-making were always compromised. I work for a billion dollar company which has been successful in doing business the traditional way with its leading brands, but leadership team is now cognizant that to sustain its leadership and continuous growth, one of the key strategies must be focused on data science.

The courses I have taken to date from both Whitman and Data Science (iSchool) gave me the opportunity to learn from both my professors and classmates. From the Whitman school in particular, we have learned the value of optimization using different tools, dashboards that can be used in conveying the relevant information and statistics that help in the analyses of data (especially in Exploratory Data Analysis). And from iSchool, we learned from using both Python and R languages, help us make predictions using the most appropriate model/s, recommend clusters, deal with big sets data and use vector graphics editor like the Adobe Illustrator to effectively convey the information. These are only some of the many tools and applications of data science.

Though there have been several projects and key lab works delivered for the whole program, I have selected three that can put together what I have learned so far as I complete this program by June 2020. Let me start with Data Visualization from the IST719 using R language. This is where I learned skills and techniques in the visualization of large data sets using Adobe Illustrator. This helped me explore further other ways of using the right visual in presenting data and information that were complete and easy to comprehend or interpret using the right graphical display, color and font among other things. I see the application of this in presenting our semi-annual or annul dashboard in varying poster formats that will convey the story to all target readers. I can also see this applied by our Sales and Marketing team showcasing marketing strategy by channel for instance. Employees in our organization have been inundated by the details of electronic dashboard that they see and at the end of the day don’t capture the message. By the process of designing a poster, conveying that right information can be more relevant and effective.

Next is Big Data Analytics from IST718. This is where we encountered more exercises in dealing with several huge data sets, cleaning data with different kinds or types of issues, merging data from different sources, choosing the right graph without inundating readers with irrelevant visuals, and more importantly is in addressing the business questions that can be used for decision making, In a nutshell, we learned here using Python language, the different models that can be used in predicting from logistics regression to neural network. From here, we used the outcome in making recommendations of which model to use with the highest accuracy and performance and integrating them with some financial data to derive other relevant information. I can see the application of this in my career (eventually as Head of Data Science) in predicting the success or failure of a marketing campaign or product launches for instance, helping HR identify high potential employees, assisting Finance in predicting a successful investment, and aiding supply chain in identifying suppliers that have the potential to be business partners. These are only some of the many applications that I can start initiating.

The last from my list is also coming from Big Data Analytics (IST719). It isn’t a project but was one of our major lab works. This involved time series forecasting using a forecasting procedure, Prophet in Python. Just like any other project that we dealt with, this required several data cleansing and preparation merging with other data that can help us make sound financial recommendations based on trend, potential and growth. This is relevant in my field after implementing several forecasting tools globally. The use of this tool can help us in the organization validate some trends with our huge data sets. Apparently, we can use this in strategically locating our distribution centers incorporating other data to assess the financial impact say of demographics and infrastructure development score.

The first challenge that I see is always in the cleansing of data since this is where we encounter a lot of challenges before we can proceed with the next steps. Moreover is in the integration or merging of other data to come up with the right findings and recommendations. It is imperative that we should have a strong understanding of the business process and challenges to be able to synthesize the presentation. This is where I see the added value of a data scientist comes in. It is where you continue to wear two hats, the business side and the technical side. The whole program did not only exposed me to Python, R, SQL, optimization tools, statistics, dashboards, but enhanced my ability to put all things together in making the outcome more relevant and valuable. This way that the company will not see data science as a fad but rather a game changer in action!

**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:

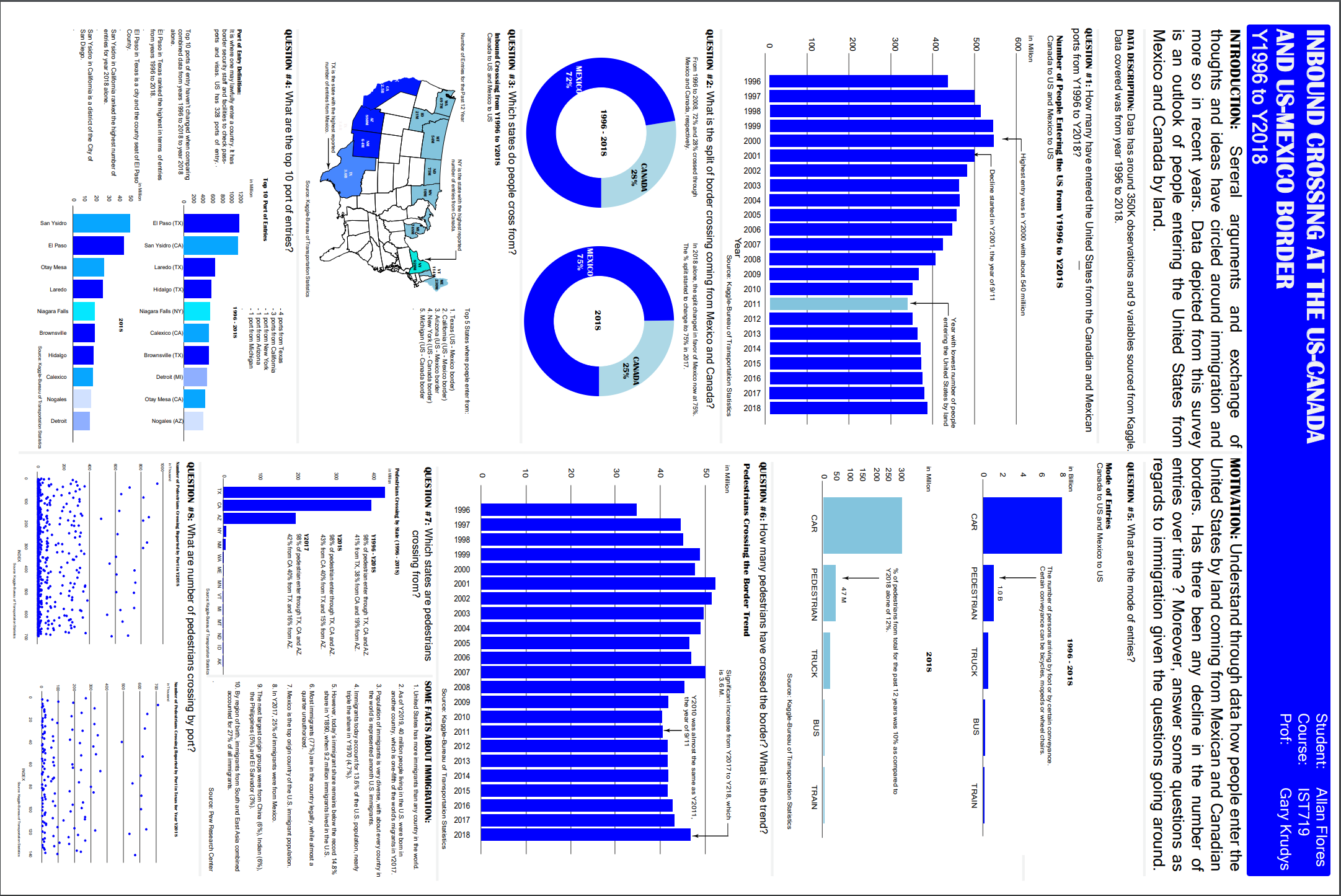
No observations were taken out from the original dataset. Used packages/libraries like gplots, dplyr, tidyverse, ggplot2, ggmap, maps, RColorBrewer, mapdata, and viridis. In using the “ggmap,” it was required to use the Google API. Adobe Illustrator software was used in designing and producing a posted.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

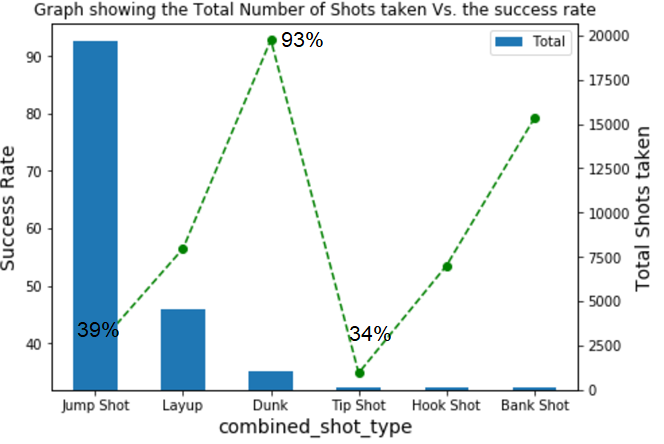
Insights Gained:

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.

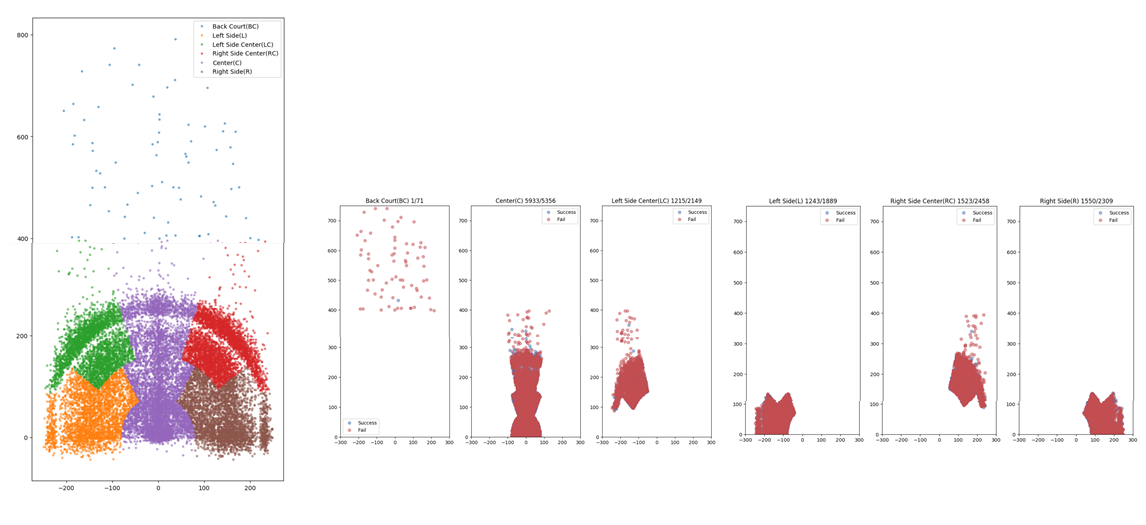
Ticket prices and general fans costs were positively correlated with Kobe’s salary. Attendance grew sharply with the opening of Staples Center, and held flat throughout Kobe’s career. Kobe created a “surplus” of $3M during 2000 – 2001 season, winning the NBA championship. The surplus Kobe created started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

Excerpts from the Output:

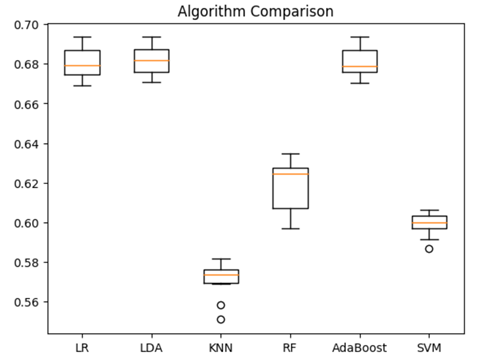
Majority of the shot type taken by Kobe was a jumpshot with 39% accuracy :

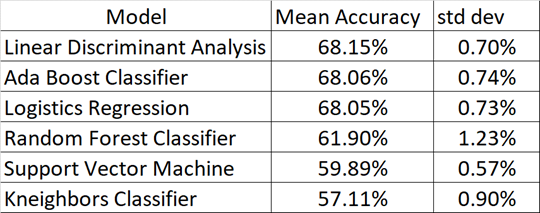


More successful shots were taken from the center of the court:

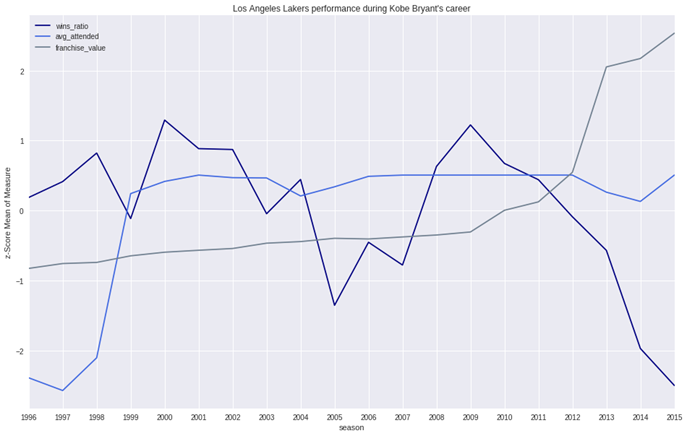


Shot prediction models:





Lakers performance during Kobe’s career:

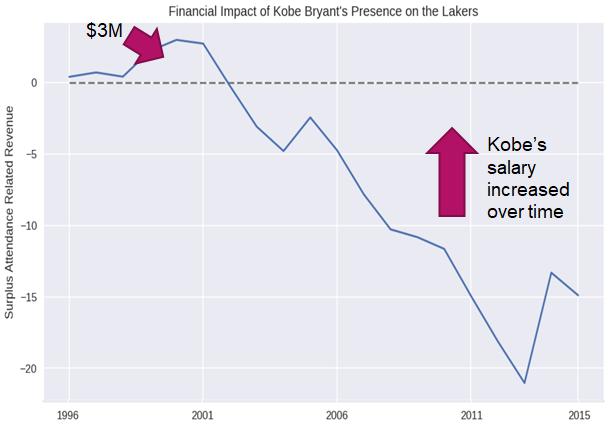


Financial impact of Kobe Bryant’s presence on the Lakers:

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.

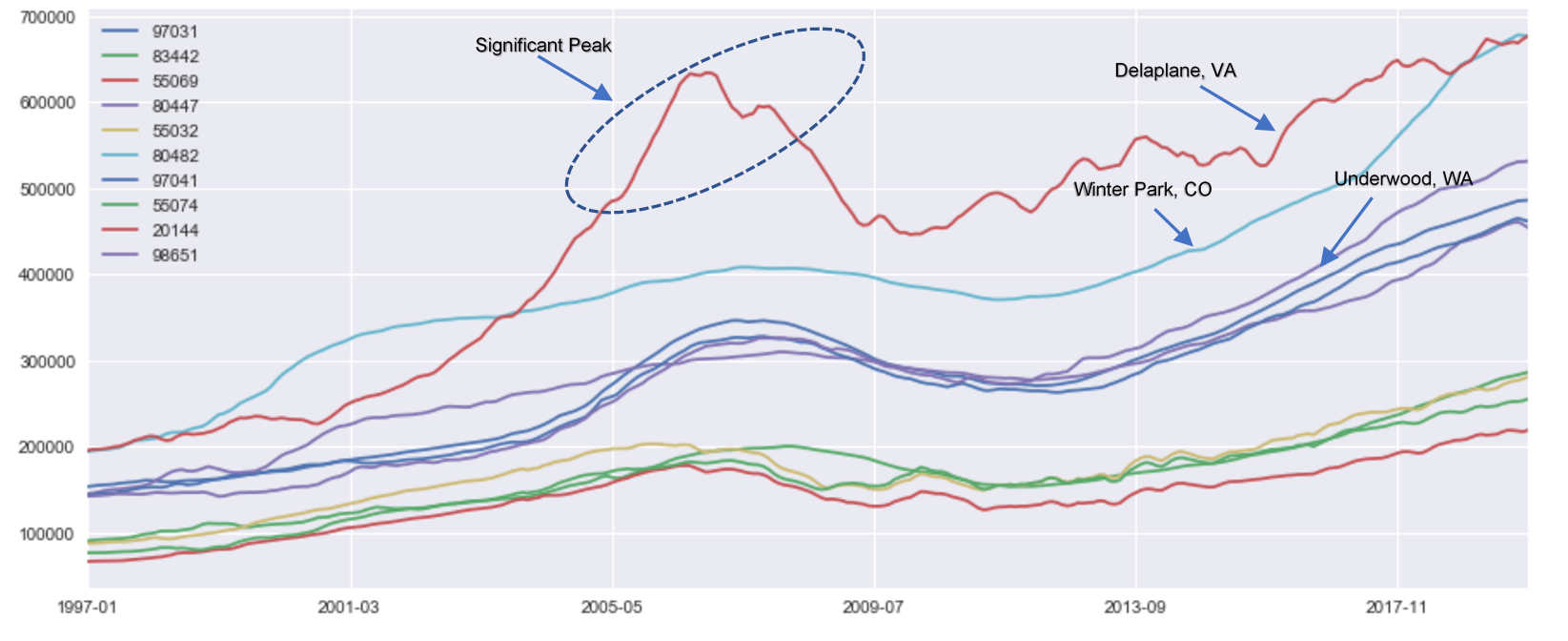
The final Top 3 zip codes recommended were 98651-Underwood, WA, 80482-Winter Park, CO and 97031-Hood River, OR. The combined growth rate from Dec 2019 to Dec 2020 was estimated to be around 9.75% and absolute change or increase of $165K in a 12-month period. In short, investing a total of $1,693K as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

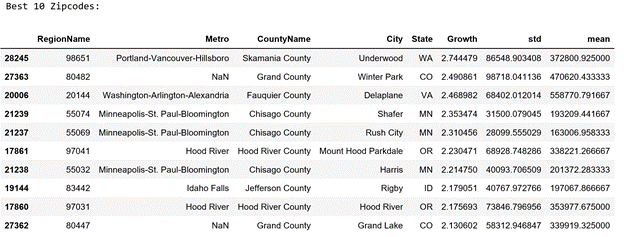
Moreover, as another option, we can replace 97031-Hood River, OR with 97041-Mount Hood Parkdale, OR, since their values were not far from each other. This may require further research or data should we entertain this route.

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.

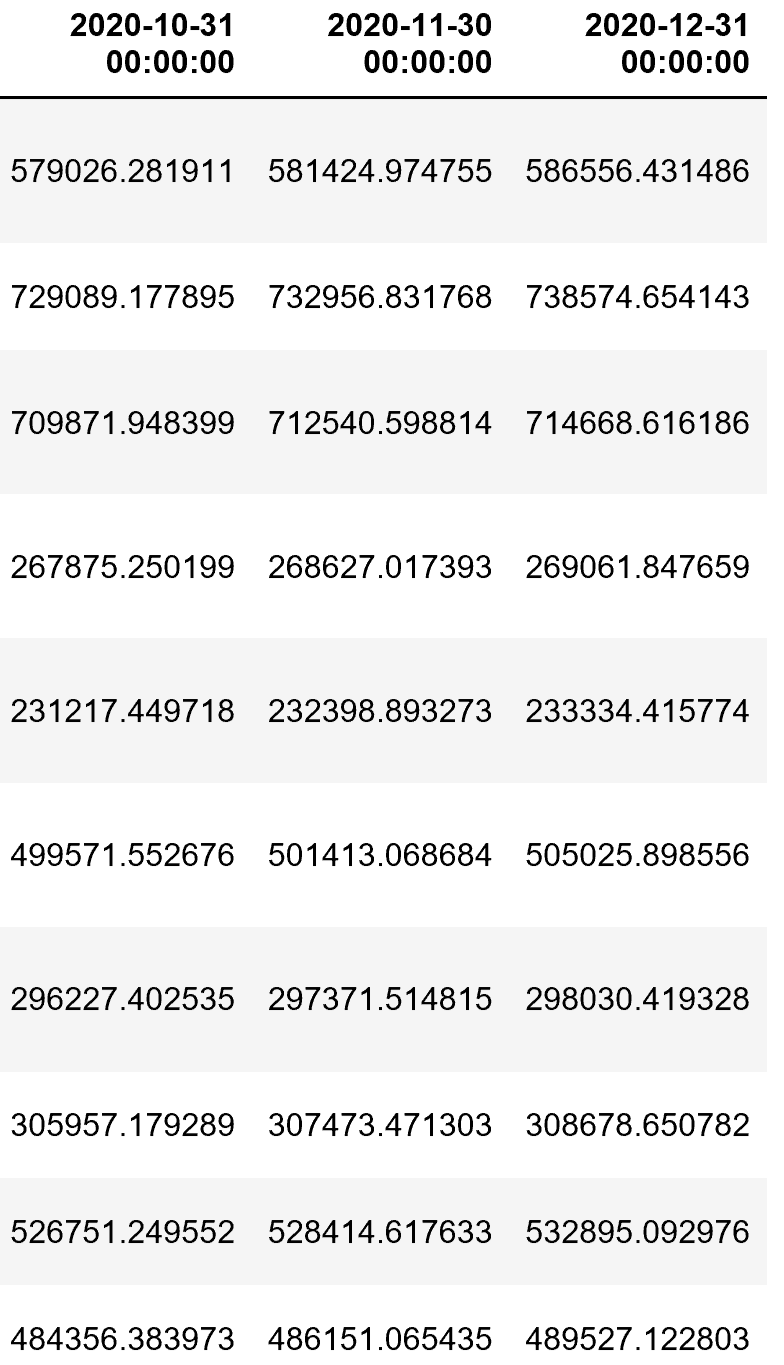
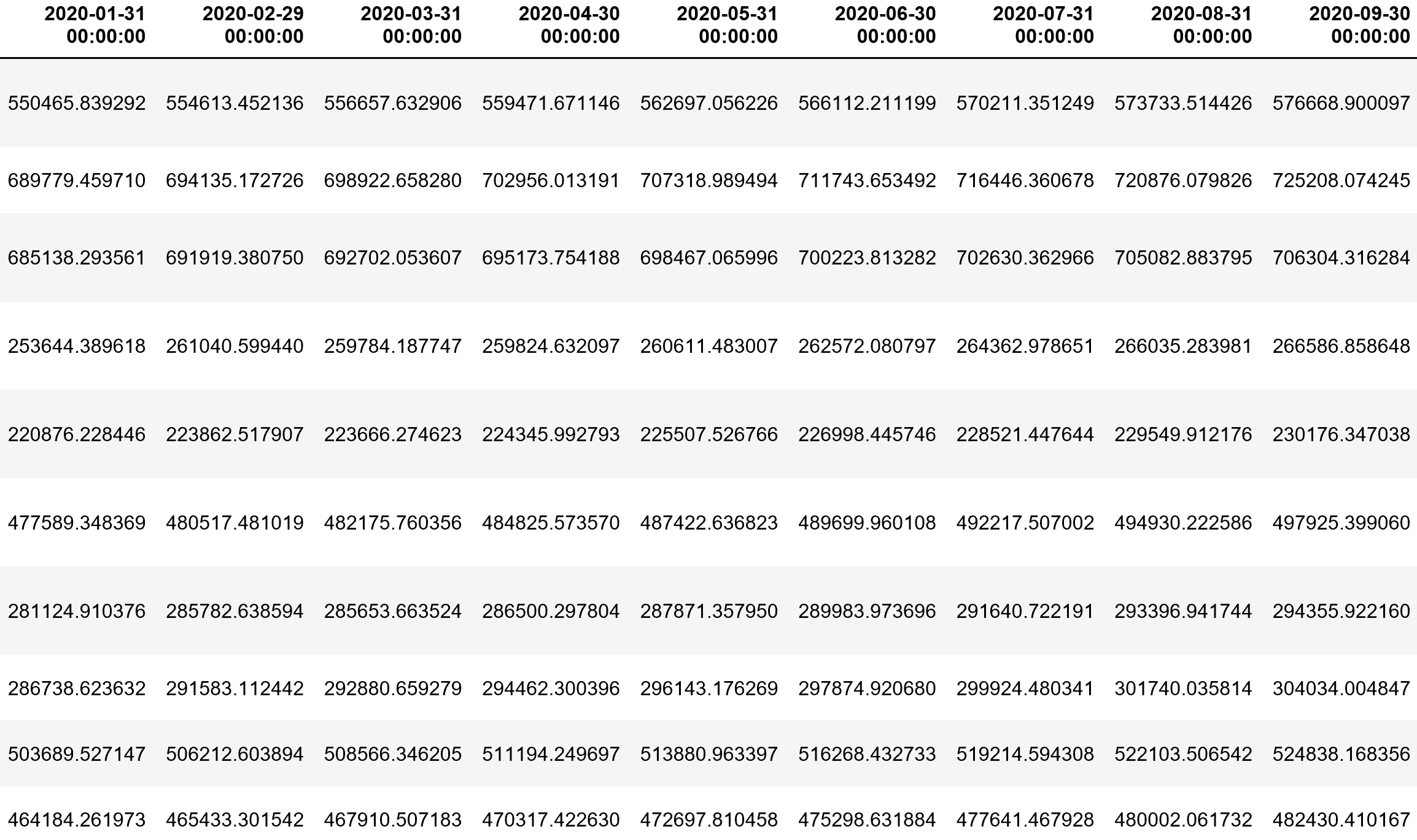
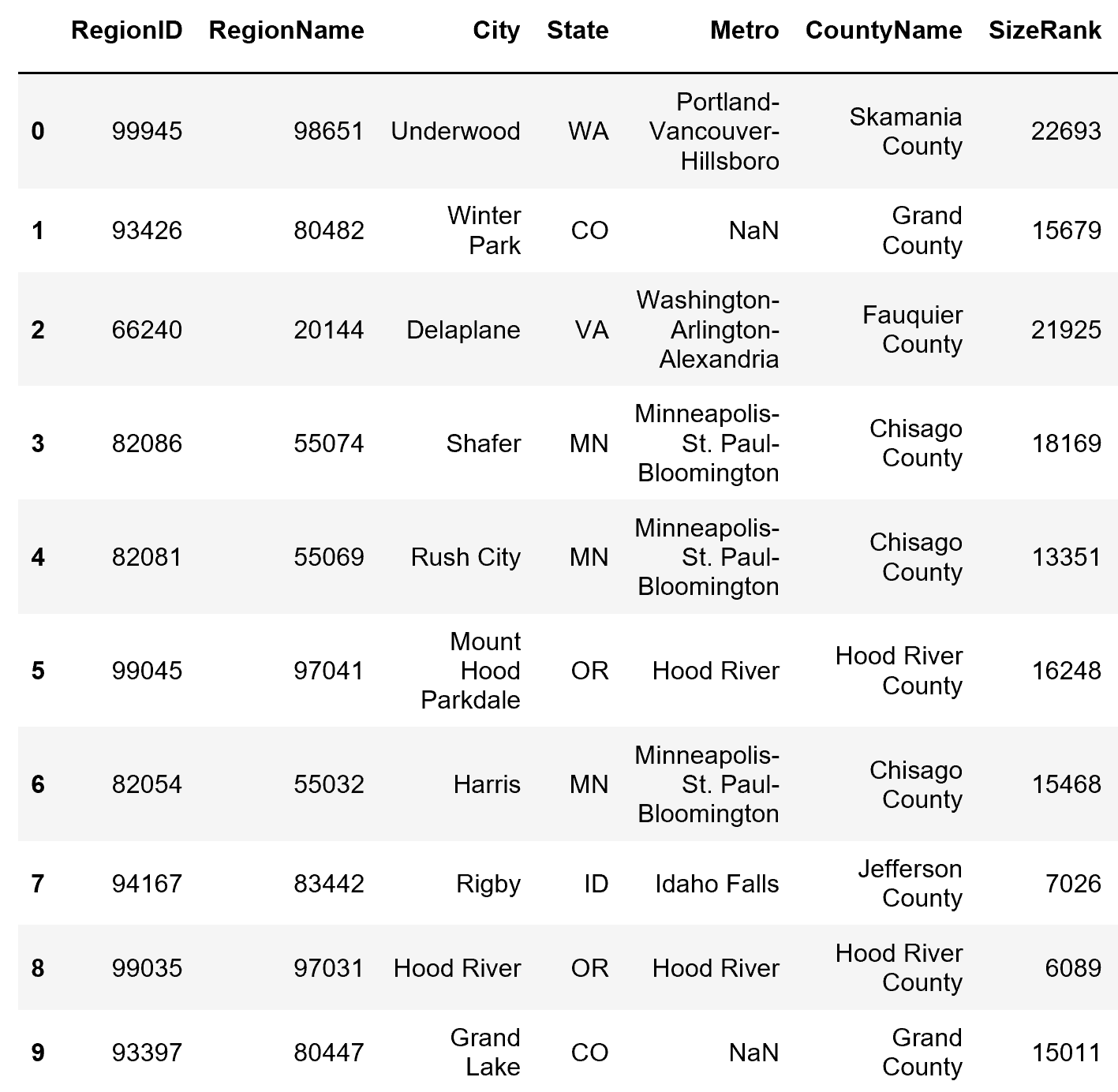
Excerpts from the Output:

Historical trend of Top 10 Median Home Values:

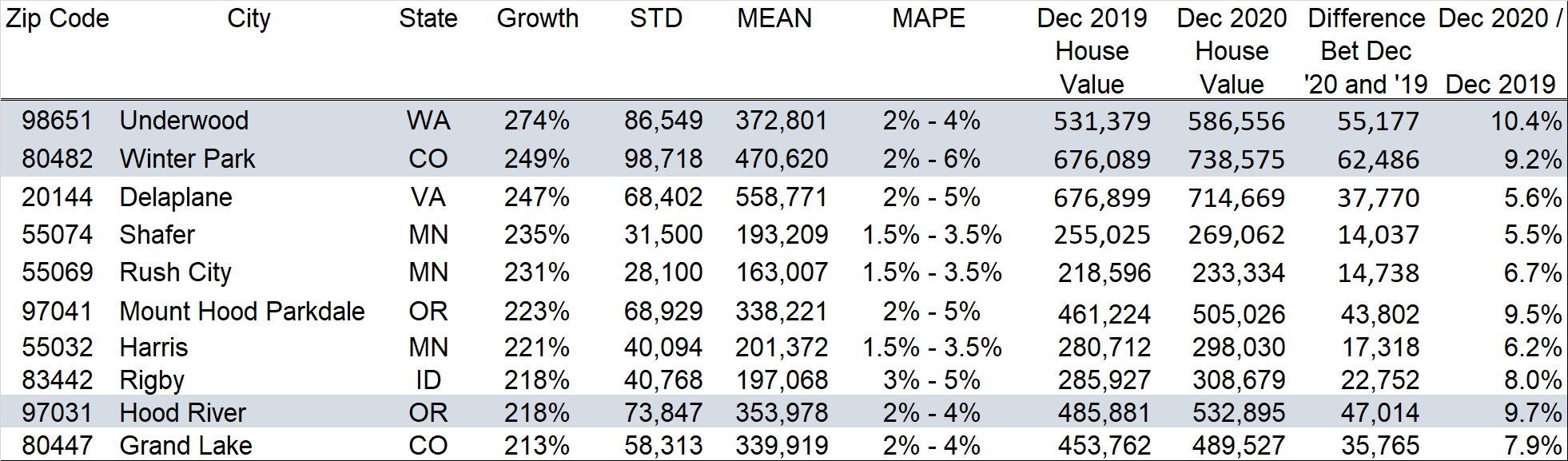




Year 2020 Forecast of the Top 10 Zip codes:



Summary of Results of the Top 10 Zip codes:



**CONCLUSION**

The program in general expanded immensely my approach now towards the use and interpretation of data in the organization and how they are applied by other institutions or entities that we deal with. We are mostly confined with using data internally without optimizing the use of external data by integrating them all together to unravel important findings that can make a difference in running an organization with the end goal of contributing to its growth and profitability. It doesn’t only make me focus on supply chain management as my field of expertise, but rather be more cognizant of other functions like sales, marketing, finance and human resources where data science can be applied. The approach is now all encompassing than thinking in silos. Moreover, this will help me educate and share my experiences with my team and colleagues in the organization to enforce common understanding and application of data science.

In talking about the content of the program or courses offered, I should say that we went through a “break-in” through the first two courses like the Data Admin Concept and Database Management and Data Analysis (IST 659) and Decision Making (MBC 638). It helped me in my preparation for the rest of the courses having taken basic statistics and basic programming 30 years ago. The challenges started after taking courses which required working on projects and cases studies using R and Python languages. It required me to do a lot of research to facilitate in surviving the rigors of these courses. It could have been easier by not spending excessive number of hours by gaining some knowledge of the programming languages prior to enrolling with the program. I also find it a smart idea to have taken the elective course, Statistical Methods in Information Science and Technology (IST 772) as it enhanced my interpretation of data with the application of both the Frequentist and Bayesian statistics. Lastly in my opinion, there should a review of the syllabus of some courses in the Whitman school like the Principles of Management Science (SCM 702) where we were thought of several optimization tools using MS Excel where some cases may be more complex in the real world inundated with more variables or attributes.

In summing up, I not only gained the knowledge and experience from fellow classmates and distinguished professors from different backgrounds, but I have expanded my network in the data science field. Our established network of data scientists is one of the key results of the program as we expand our resources and getting continuous updates in the different applications and challenges of data science. I am grateful to Syracuse University in accepting my application to be a part of a great institution!

**PORTFOLIO MILESTONE VIDEO LINK**

Due to the size of the file, please copy the link below for access.

https://drive.google.com/drive/folders/1UGk8KsntAbc\_\_kvbq20T3ugUyT8pyZB-?usp=sharing