



MS in Applied Data Science  
Syracuse University  
Portfolio Milestone

**ALLAN FLORES**

**SUID: 833497871**

# Background

- B.S. in Industrial Engineering
- Supply Chain Management Director
- Monde Nissin Corporation, Mead Johnson Nutrition, Nestle, Sanofi, Novartis, Saputo Foods Dairy

# Project Covered

- Kobe Bryant Shot Selection (Big Data - Python)
- Inbound Crossing at the US-MEXICO AND US-CANADA Border (Visualization - R)
- Investing in Top 3 Zip Codes (Time Series Analysis - Python)

# Kobe Bryant Shot Prediction and Financial Impact to Lakers



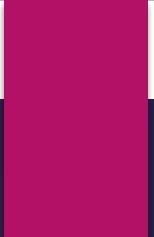
# Motivation

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.

Therefore, we would like to get an experience in sports analytics.



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 our 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.

## Business Questions

# **Collection of Data – 2 Parts**

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

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

# The Collection of Data – Part 1

## Kobe Bryant Shot Selection Data

- Primary source of dataset is from Kaggle competition titled. “Kobe Bryant Shot Selection” (<https://www.kaggle.com/c/kobe-bryant-shot-selection/overview/description>) for the player’s statistics.
  - Each of the observation represents a shot taken by Kobe Bryant over the data collection period.
  - Period covered from years 1996 to 2005

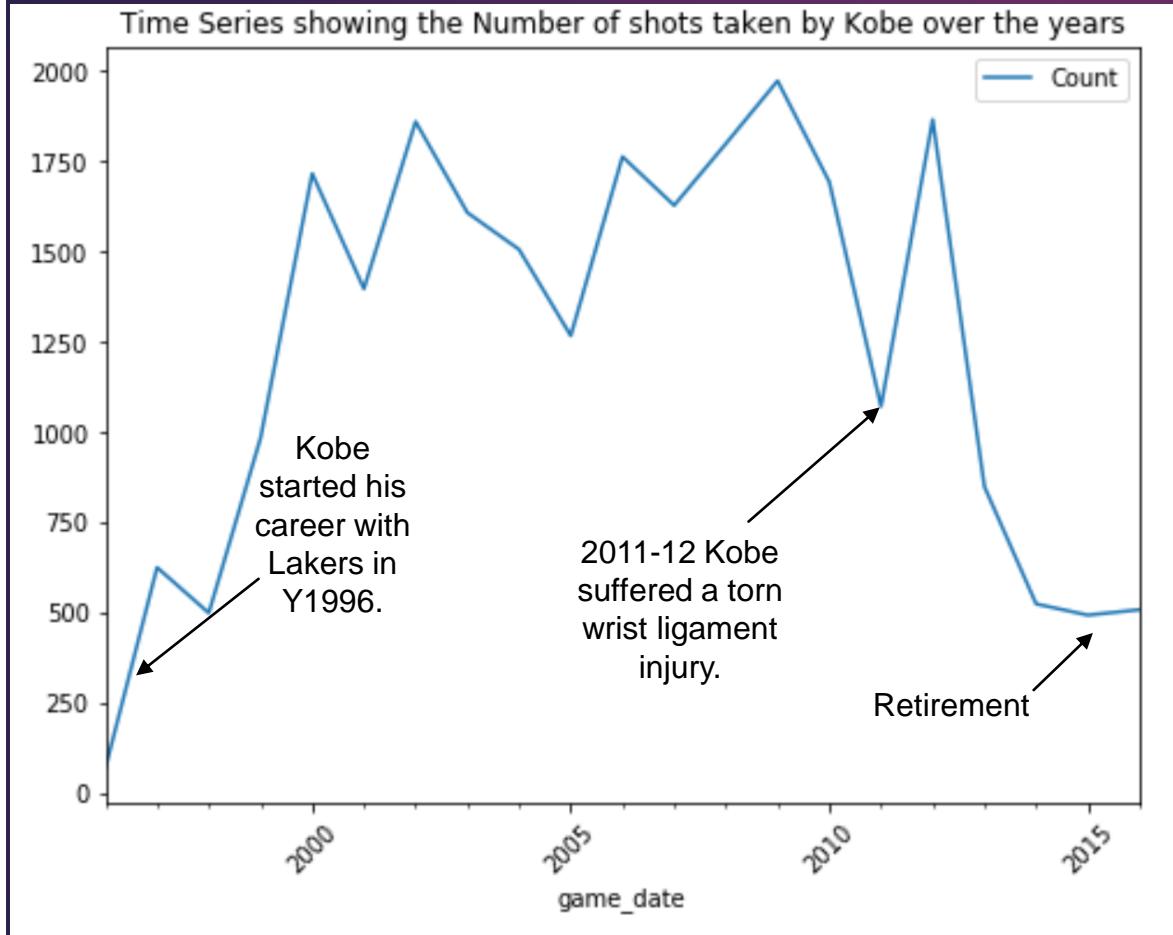
| Column             | Sample Data        |
|--------------------|--------------------|
| action_type        | Slam Dunk Shot     |
| combined_shot_type | Dunk               |
| game_event_id      | 253                |
| game_id            | 20000068           |
| lat                | 34.0443            |
| loc_x              | -1                 |
| loc_y              | 0                  |
| lon                | -118.2708          |
| minutes_remaining  | 7                  |
| period             | 3                  |
| playoffs           | 0                  |
| season             | 2000-01            |
| seconds_remaining  | 53                 |
| shot_distance      | 0                  |
| shot_made_flag     | 1                  |
| shot_type          | 2PT Field Goal     |
| shot_zone_area     | Center(C)          |
| shot_zone_basic    | Restricted Area    |
| shot_zone_range    | Less Than 8 ft.    |
| team_id            | 1610612747         |
| team_name          | Los Angeles Lakers |
| game_date          | 36838              |
| matchup            | LAL @ SAS          |
| opponent           | SAS                |
| shot_id            | 102                |

# Pre-Processing and Data Cleaning – Part 1

| Column             | Sample Data        |
|--------------------|--------------------|
| action_type        | Slam Dunk Shot     |
| combined_shot_type | Dunk               |
| game_event_id      | 253                |
| game_id            | 20000068           |
| lat                | 34.0443            |
| loc_x              | -1                 |
| loc_y              | 0                  |
| lon                | -118.2708          |
| minutes_remaining  | 7                  |
| period             | 3                  |
| playoffs           | 0                  |
| season             | 2000-01            |
| seconds_remaining  | 53                 |
| shot_distance      | 0                  |
| shot_made_flag     | 1                  |
| shot_type          | 2PT Field Goal     |
| shot_zone_area     | Center(C)          |
| shot_zone_basic    | Restricted Area    |
| shot_zone_range    | Less Than 8 ft.    |
| team_id            | 1610612747         |
| team_name          | Los Angeles Lakers |
| game_date          | 36838              |
| matchup            | LAL @ SAS          |
| opponent           | SAS                |
| shot_id            | 102                |

- 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.

# Exploring the Dataset – Part 1



There was a steady increase in the number of shots until year 2000.

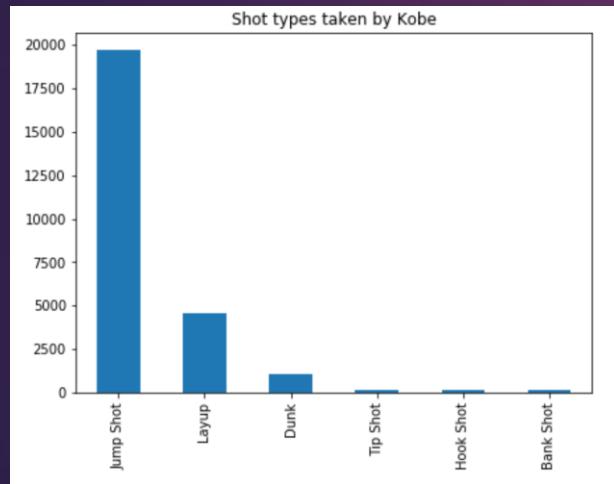
His shots stabilized with some variations from years 2000 to 2011.

His shots started to decline in Y2011 until his retirement in Y2016.

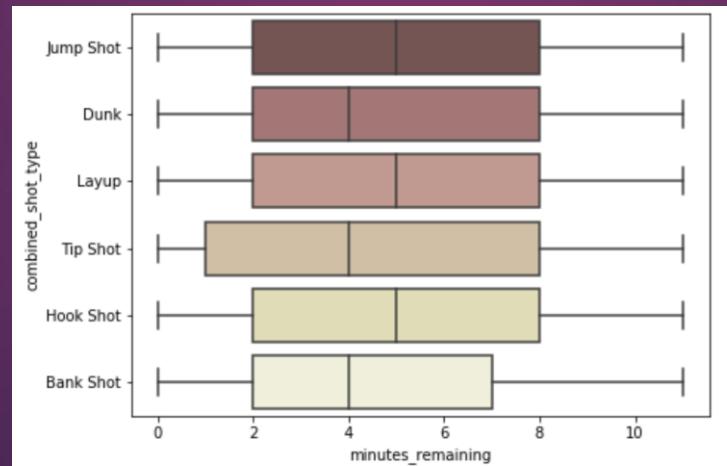
It was around Y2011-12 when Kobe suffered from a torn wrist ligament.

# Exploring the Dataset – Part 1

Majority of the shot type taken by Kobe was a jump shot.

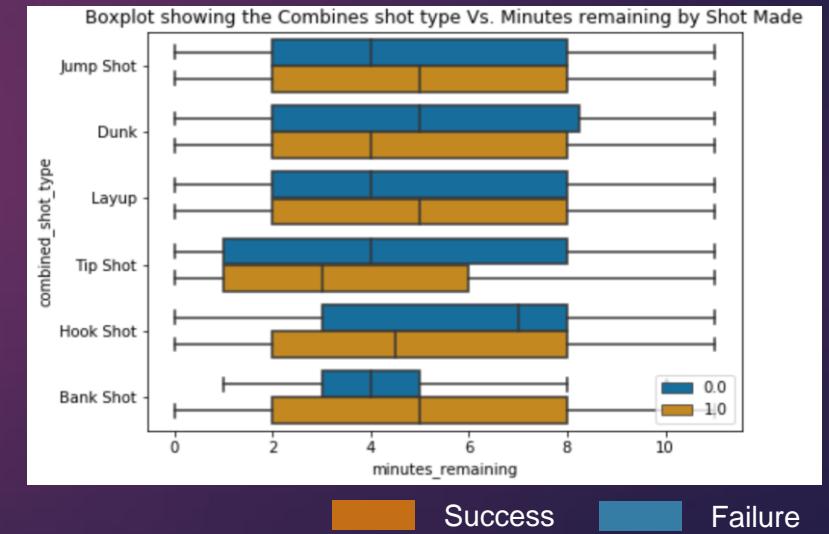


Kobe took different kinds of shots at an average of 4 – 5 minutes remaining. The tip shop had the highest distribution.



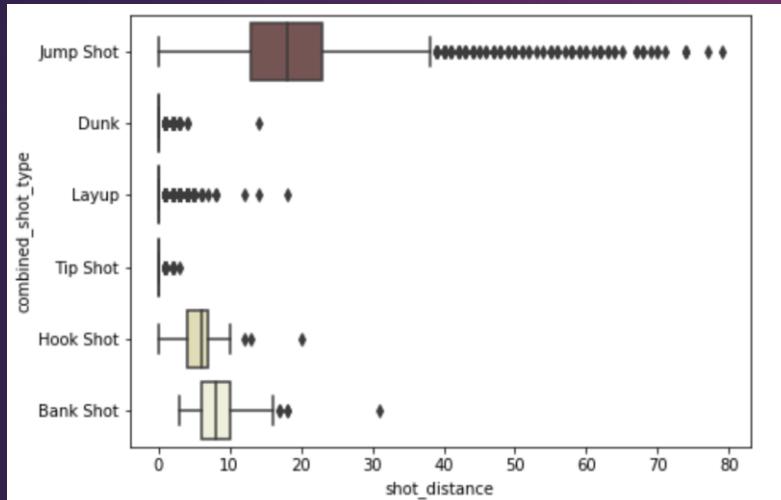
With 4 - 5 minutes remaining, the tip shop had the highest distribution, while bank shots had the lowest distribution.

Tip shots had the lowest median in terms of successful shots.

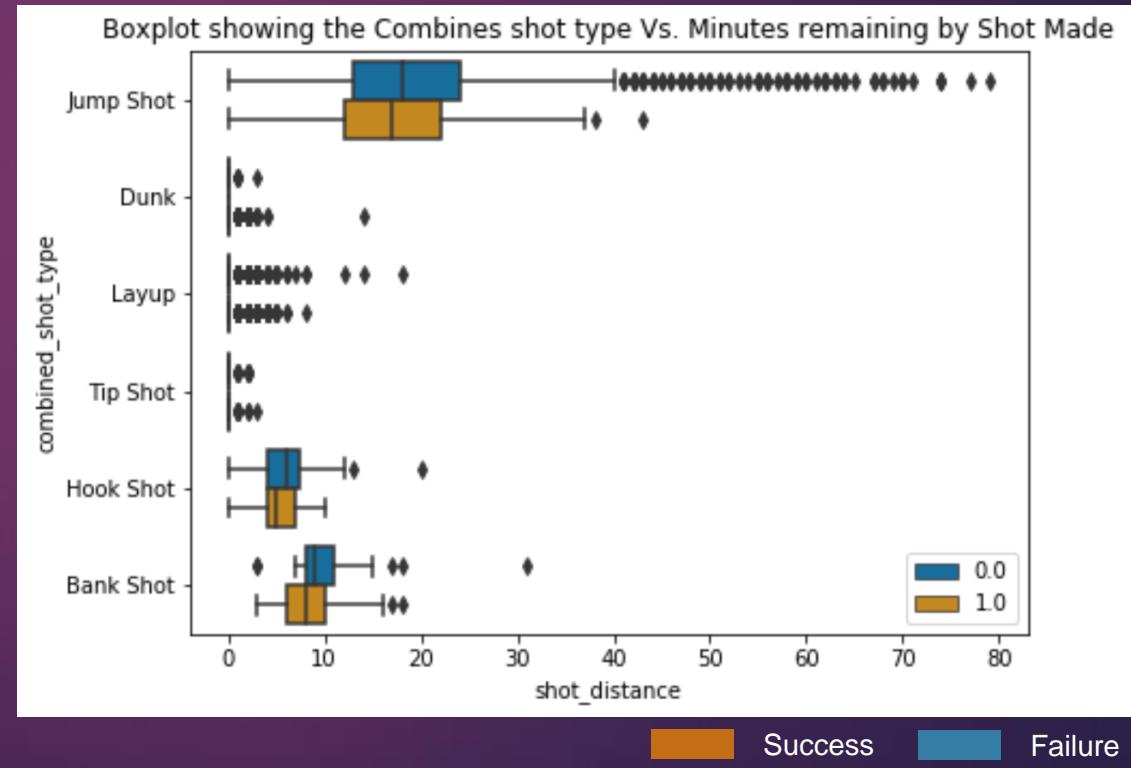


# Exploring the Dataset – Part 1

Jump shot had the widest distribution in terms of distance at a median of ~20 meters away from his court.



In general, the distribution of failed shots as compared to successful shots were wider except for dunk and tip shot.

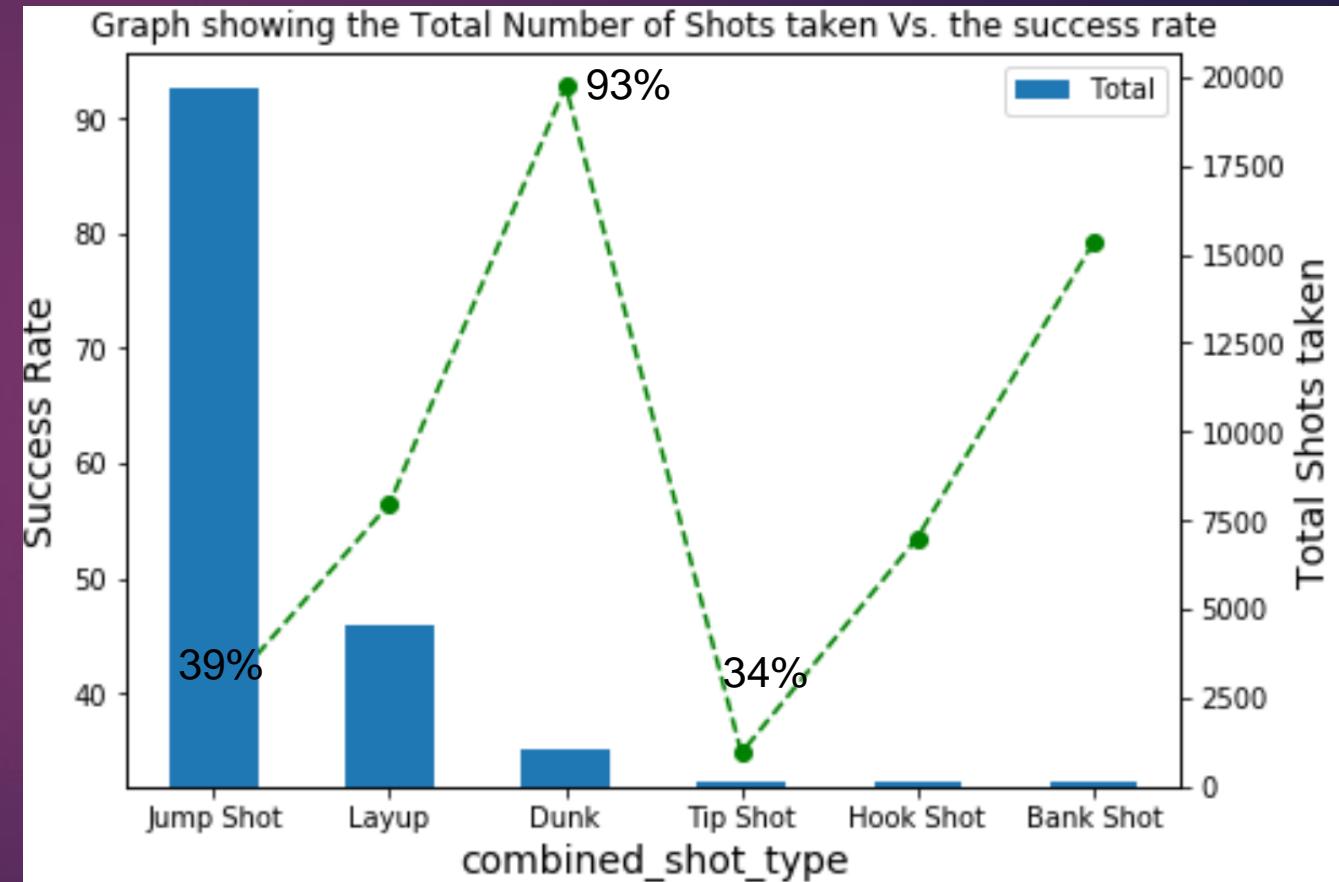


# Exploring the Dataset – Part 1

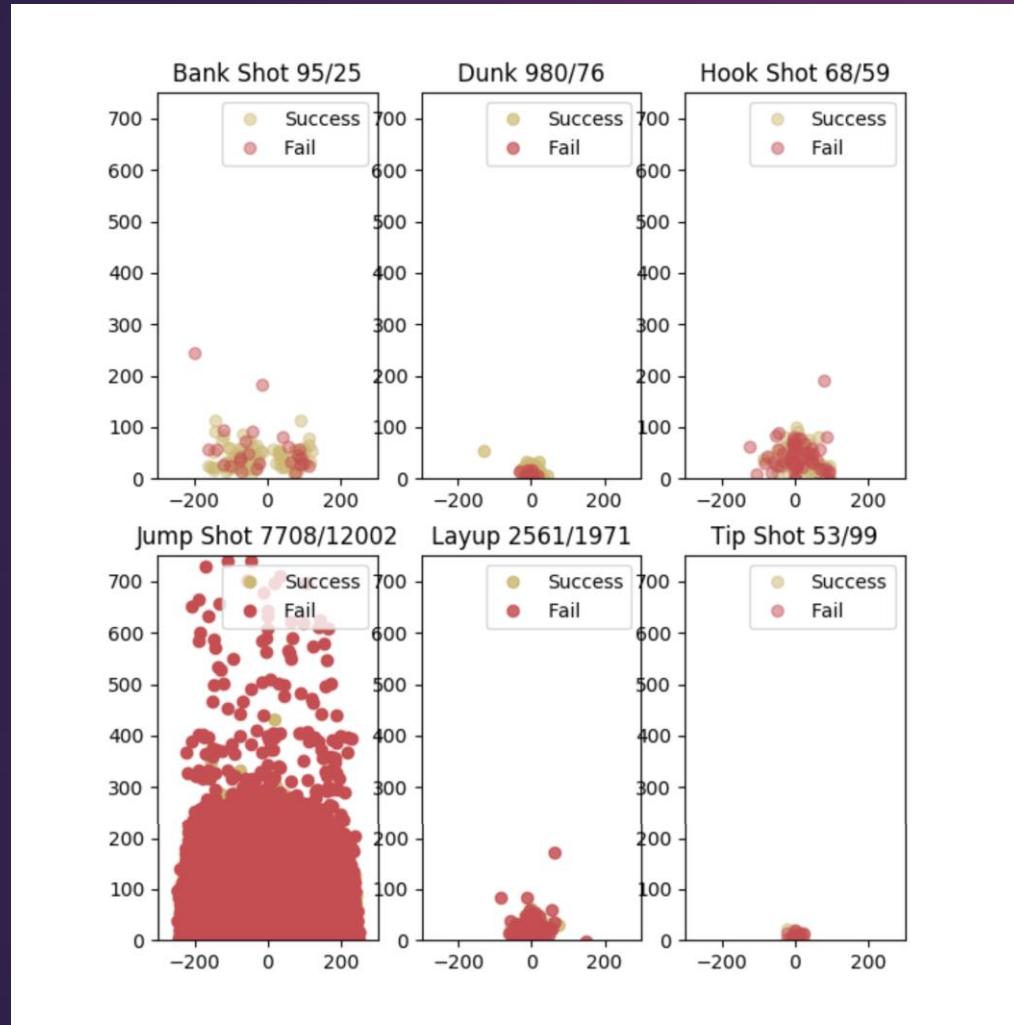
Jump shots with majority of the shots only had 39% success rate.

Tip shots had 34% success rate.

Dunk had the highest success rate of 93%.



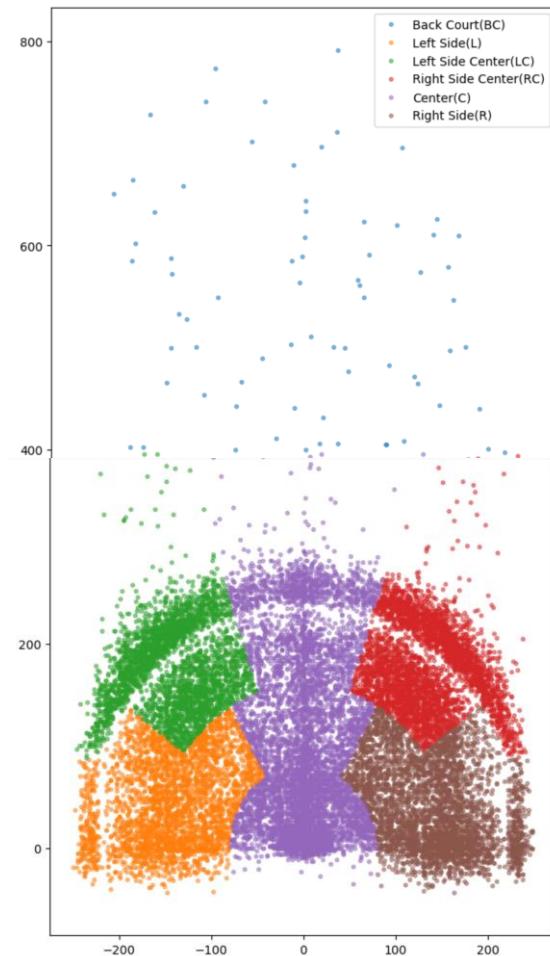
# Exploring the Dataset – Part 1



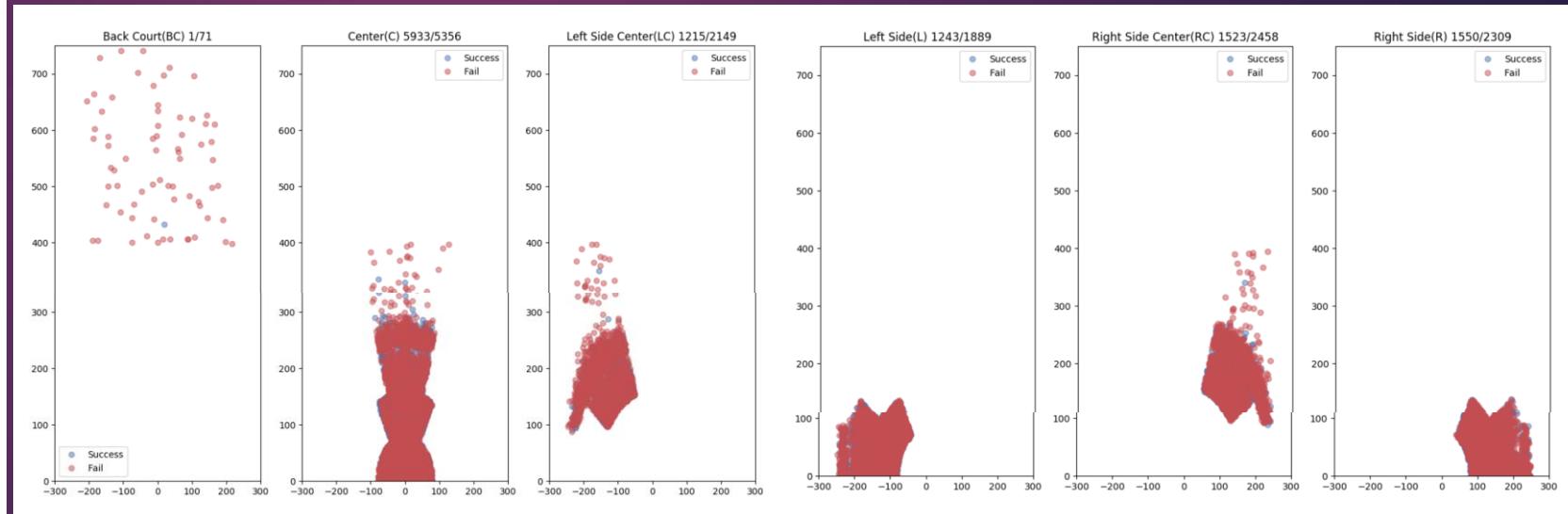
More successful shots for bank shot and dunk.

Less successful for jump shot, layup, tip shot, and hook shot.

# Exploring the Dataset – Part 1

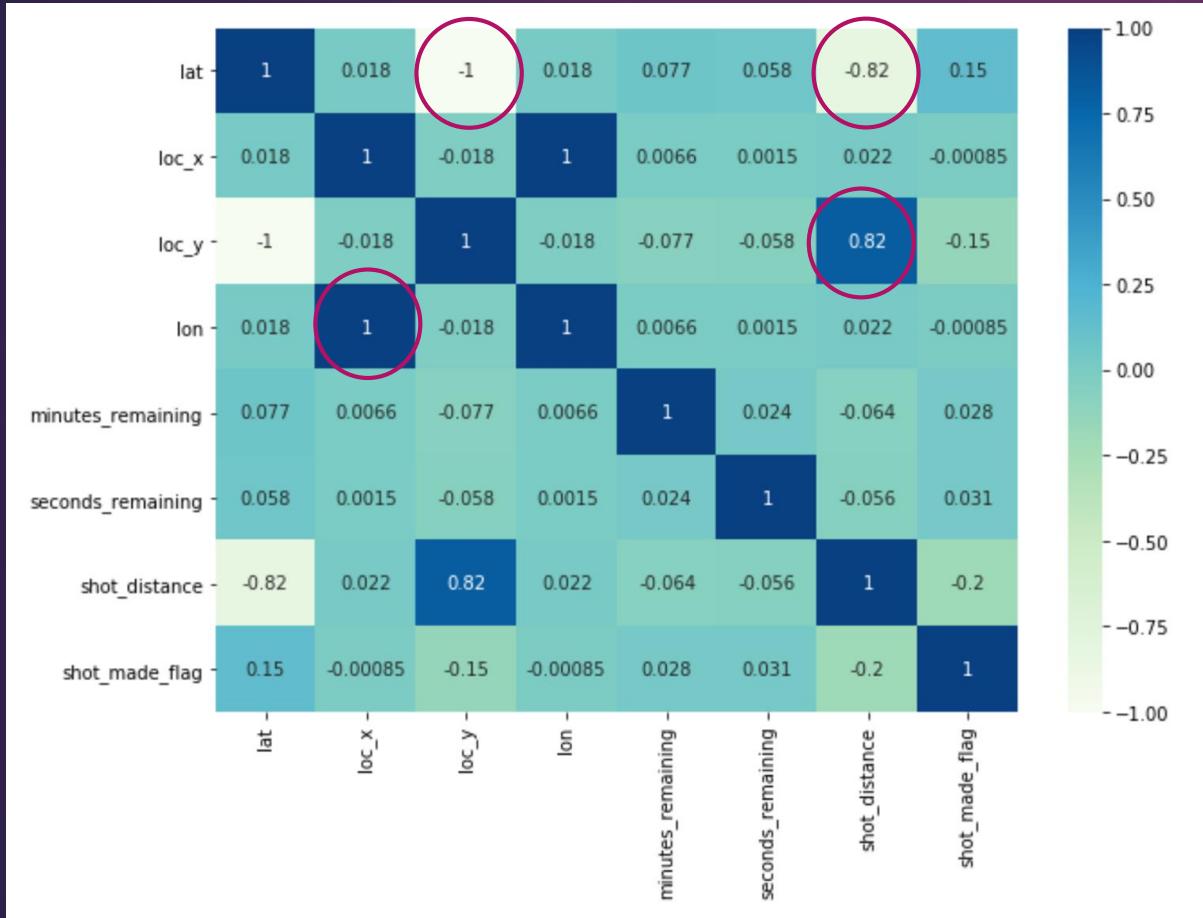


More successful shots can be seen when taken at the center of the court.



# Model – Part 1

## Feature Selection:



Removed highly correlated values.

lat and loc\_y have a perfect negative correlation.

lon and loc\_x have a perfect +ve correlation.

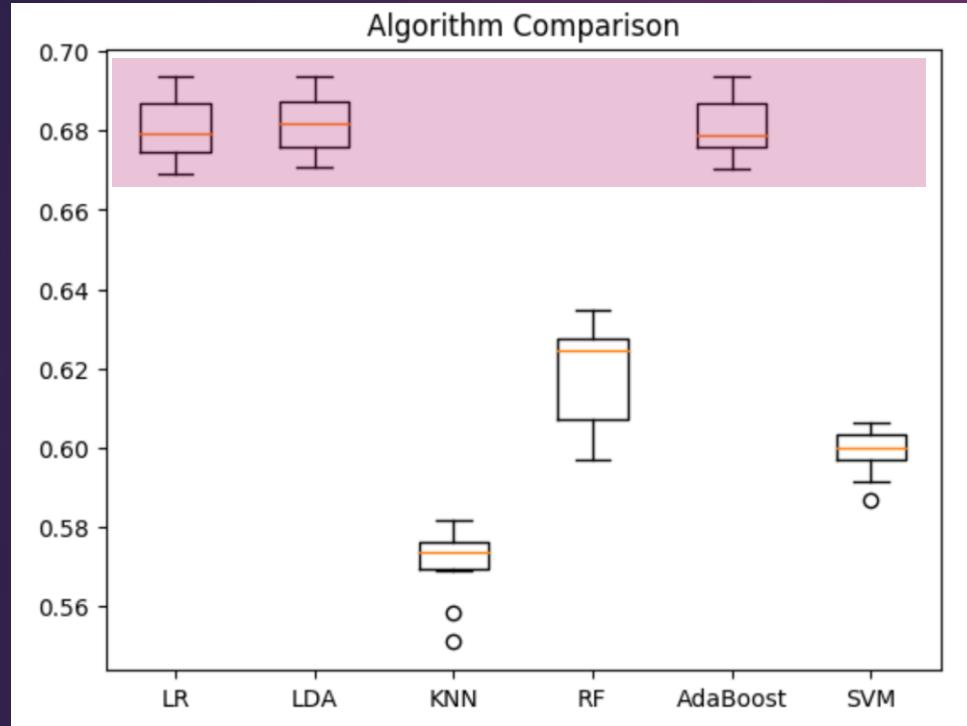
Shot distance is also correlated to lat and loc\_y as well with a strength of -0.82 and +0.82, respectively.

# Model Part 1 – OLS Regression

Loc\_y and lat explain the 66.9% of the variance in shot distance.

| OLS Regression Results |                  |                     |           |       |        |        |
|------------------------|------------------|---------------------|-----------|-------|--------|--------|
| Dep. Variable:         | shot_distance    | R-squared:          | 0.669     |       |        |        |
| Model:                 | OLS              | Adj. R-squared:     | 0.669     |       |        |        |
| Method:                | Least Squares    | F-statistic:        | 5.201e+04 |       |        |        |
| Date:                  | Sat, 29 Feb 2020 | Prob (F-statistic): | 0.00      |       |        |        |
| Time:                  | 10:03:59         | Log-Likelihood:     | -79792.   |       |        |        |
| No. Observations:      | 25697            | AIC:                | 1.596e+05 |       |        |        |
| Df Residuals:          | 25695            | BIC:                | 1.596e+05 |       |        |        |
| Df Model:              | 1                |                     |           |       |        |        |
| Covariance Type:       | nonrobust        |                     |           |       |        |        |
|                        |                  |                     |           |       |        |        |
|                        | coef             | std err             | t         | P> t  | [0.025 | 0.975] |
| Intercept              | 0.0047           | 4.18e-05            | 113.649   | 0.000 | 0.005  | 0.005  |
| lat                    | 0.1616           | 0.001               | 113.566   | 0.000 | 0.159  | 0.164  |
| loc_y                  | 0.0873           | 0.000               | 229.093   | 0.000 | 0.087  | 0.088  |
| Omnibus:               | 3839.416         | Durbin-Watson:      | 1.917     |       |        |        |
| Prob(Omnibus):         | 0.000            | Jarque-Bera (JB):   | 5807.890  |       |        |        |
| Skew:                  | 1.100            | Prob(JB):           | 0.00      |       |        |        |
| Kurtosis:              | 3.764            | Cond. No.           | 3.26e+16  |       |        |        |

# Model Part 1 – LR, LDA, KNN, RF, AdaBoost, SVM



LDA, AdaBoost and LR are the most accurate models in predicting shots.

| Model                        | Mean Accuracy | std dev |
|------------------------------|---------------|---------|
| Linear Discriminant Analysis | 68.15%        | 0.70%   |
| Ada Boost Classifier         | 68.06%        | 0.74%   |
| Logistics Regression         | 68.05%        | 0.73%   |
| Random Forest Classifier     | 61.90%        | 1.23%   |
| Support Vector Machine       | 59.89%        | 0.57%   |
| Kneighbors Classifier        | 57.11%        | 0.90%   |

# The Collection of Data – Part 2

## Annual Income and Expense Data

- From Rodney Fort's sports financial data

- <https://sites.google.com/site/rodswebpages/codes>

redirect to : [https://drive.google.com/drive/folders/1pr\\_yPm9oPLcfCtWOtnrLoJdvzFDbYBAk](https://drive.google.com/drive/folders/1pr_yPm9oPLcfCtWOtnrLoJdvzFDbYBAk)

- Lakers financial performance by year from 1996 to 2015

| season | gate_revenue | other_revenue | total_revenue | player_expense | other_expense | total_expense | operating_income | franchise_value | financial_rank |
|--------|--------------|---------------|---------------|----------------|---------------|---------------|------------------|-----------------|----------------|
| 2011   | 74.0         | 123.0         | 197.0         | 76.0           | 73.2          | 149.2         | 47.8             | 1000.0          | 2.0            |
| 2012   | 94.0         | 201.0         | 295.0         | 110.0          | 118.6         | 228.6         | 66.4             | 1350.0          | 2.0            |
| 2013   | 96.0         | 197.0         | 293.0         | 89.0           | 99.9          | 188.9         | 104.1            | 2600.0          | 1.0            |
| 2014   | 98.0         | 206.0         | 304.0         | 76.0           | 94.5          | 170.5         | 133.4            | 2700.0          | 2.0            |
| 2015   | 99.0         | 234.0         | 333.0         | 84.0           | 129.8         | 213.8         | 119.2            | 3000.0          | 1.0            |

## Annual Ticket Price and Fan Cost Data

- From Rodney Fort's sports financial data
- Lakers game performance from 1996 to 2015

| season | attendance | fan_cost_index | ticket_price_index | win_ratio |
|--------|------------|----------------|--------------------|-----------|
| 1996   | 697132     | 232.56         | 38.39              | 0.683     |
| 1997   | 691994     | 250.60         | 41.65              | 0.744     |
| 1998   | 430007     | 295.44         | 51.11              | 0.620     |
| 1999   | 771420     | 427.56         | 81.89              | 0.817     |
| 2000   | 776336     | 446.76         | 87.69              | 0.683     |

# The Collection of Data – Part 2

## Lakers Attendance

- Data provided by the Association of Professional Basketball Research
- <https://www.apbr.org/attendance.html>
- Years covered from 1960 to 2015

 Columns selected in the final data frame

| season | home_games | total_attended | avg_attended |
|--------|------------|----------------|--------------|
| 2011   | 33         | 626901         | 18997        |
| 2012   | 41         | 778877         | 18997        |
| 2013   | 41         | 771974         | 18829        |
| 2014   | 41         | 768244         | 18738        |
| 2015   | 41         | 778877         | 18997        |

## Lakers Annual Record

- Sourced from Sports Reference API, Robert Clark
- <https://sportsreference.readthedocs.io/en/stable/nba.html>
- Installed 'sportsreference' and chose Lakers annual records from 1996 to 2016
- Data selected:
  - Wins
  - Losses
  - Wins Ratio = Wins / (Wins + Losses) – *Derived column*

# The Collection of Data – Part 2

## Lakers Team Data

- Sourced from Sports Reference API, Robert Clark
- <https://sportsreference.readthedocs.io/en/stable/nba.html>
- Installed ‘sportsreference’ and chose Lakers annual records from 1996 to 2016

Data selected (excerpt):

|      |  | abbreviation | name               | assists | blocks | defensive_rebounds | field_goal_attempts | field_goal_percentage | field_goals | free_throw_attempts | free_throw_percentage |
|------|--|--------------|--------------------|---------|--------|--------------------|---------------------|-----------------------|-------------|---------------------|-----------------------|
| 1996 |  | LAL          | Los Angeles Lakers | 2080    | 516    |                    | 2303                | 6706                  | 0.480       | 3216                | 2049                  |
| 1997 |  | LAL          | Los Angeles Lakers | 1845    | 575    |                    | 2414                | 6642                  | 0.454       | 3018                | 2330                  |
| 1998 |  | LAL          | Los Angeles Lakers | 2009    | 556    |                    | 2471                | 6536                  | 0.481       | 3146                | 2743                  |
| 1999 |  | LAL          | Los Angeles Lakers | 1095    | 287    |                    | 1482                | 3935                  | 0.468       | 1841                | 1503                  |
| 2000 |  | LAL          | Los Angeles Lakers | 1921    | 534    |                    | 2738                | 6836                  | 0.459       | 3137                | 2368                  |

Final data selected:

| TEAM                                   |
|--|
| Team Rank                              |
| Team Games Played                      |
| Team Minutes Played                    |
| Team Points                            |
| Team Total Rebounds                    |
| Team Blocks                            |
| Team Steals                            |
| Team Turnovers                         |
| Team Two Point Field Goal Attempts     |
| Team Two Point Field Goal Percentage   |
| Team Three Point Field Goal Attempts   |
| Team Three Point Field Goal Percentage |
| Team Free Throw Attempts               |
| Team Personal Fouls                    |
| Team Opponent Points                   |
| Team Opponent Field Goal Attempts      |
| Team Opponent Field Goal Percentage    |
| Team Opponent Turnovers                |
| Team Opponent Personal Fouls           |

# The Collection of Data – Part 2

Data selected (excerpt):

| team_abbreviation | player_id | nationality | position                 | height | weight | assist_percentage | assists | block_percentage | blocks | box_plus_minus |      |
|-------------------|-----------|-------------|--------------------------|--------|--------|-------------------|---------|------------------|--------|----------------|------|
| season            |           |             |                          |        |        |                   |         |                  |        |                |      |
| 1996              | LAL       | bryanko01   | United States of America | SF     | 6-6    | 212               | 13.8    | 91               | 1.6    | 23             | -0.1 |
| 1997              | LAL       | bryanko01   | United States of America | SF     | 6-6    | 212               | 16.0    | 199              | 1.3    | 40             | 1.4  |
| 1998              | LAL       | bryanko01   | United States of America | SF     | 6-6    | 212               | 17.5    | 190              | 1.9    | 50             | 2.1  |
| 1999              | LAL       | bryanko01   | United States of America | SF     | 6-6    | 212               | 22.4    | 323              | 1.7    | 62             | 5.1  |
| 2000              | LAL       | bryanko01   | United States of America | SF     | 6-6    | 212               | 23.0    | 338              | 1.1    | 43             | 4.8  |

Final data selected:

|                                    |
|------------------------------------|
| Kobe Salary                        |
| Kobe Games Started                 |
| Kobe Games Played                  |
| Kobe Minutes Played                |
| Kobe Win Share Per 48 Minutes      |
| Kobe Win Shares                    |
| Kobe Value Over Replacement Player |
| Kobe Usage Percentage              |
| Kobe Box Plus Minus                |
| Kobe Defensive Box Plus Minus      |
| Kobe Offensive Box Plus Minus      |
| Kobe Defensive Win Shares          |
| Kobe Offensive Win Shares          |
| Kobe True Shooting Percentage      |
| Kobe Player Efficiency Rating      |
| Kobe Points                        |
| Kobe Personal Fouls                |

## Player's Career Data – Kobe Bryant Data

- From  
'sportsreference.nba.roster'  
import Player
- Selected player, "bryanko01"  
to extract Kobe's statistics

# Pre-Processing and Data Cleaning – Part 2

---

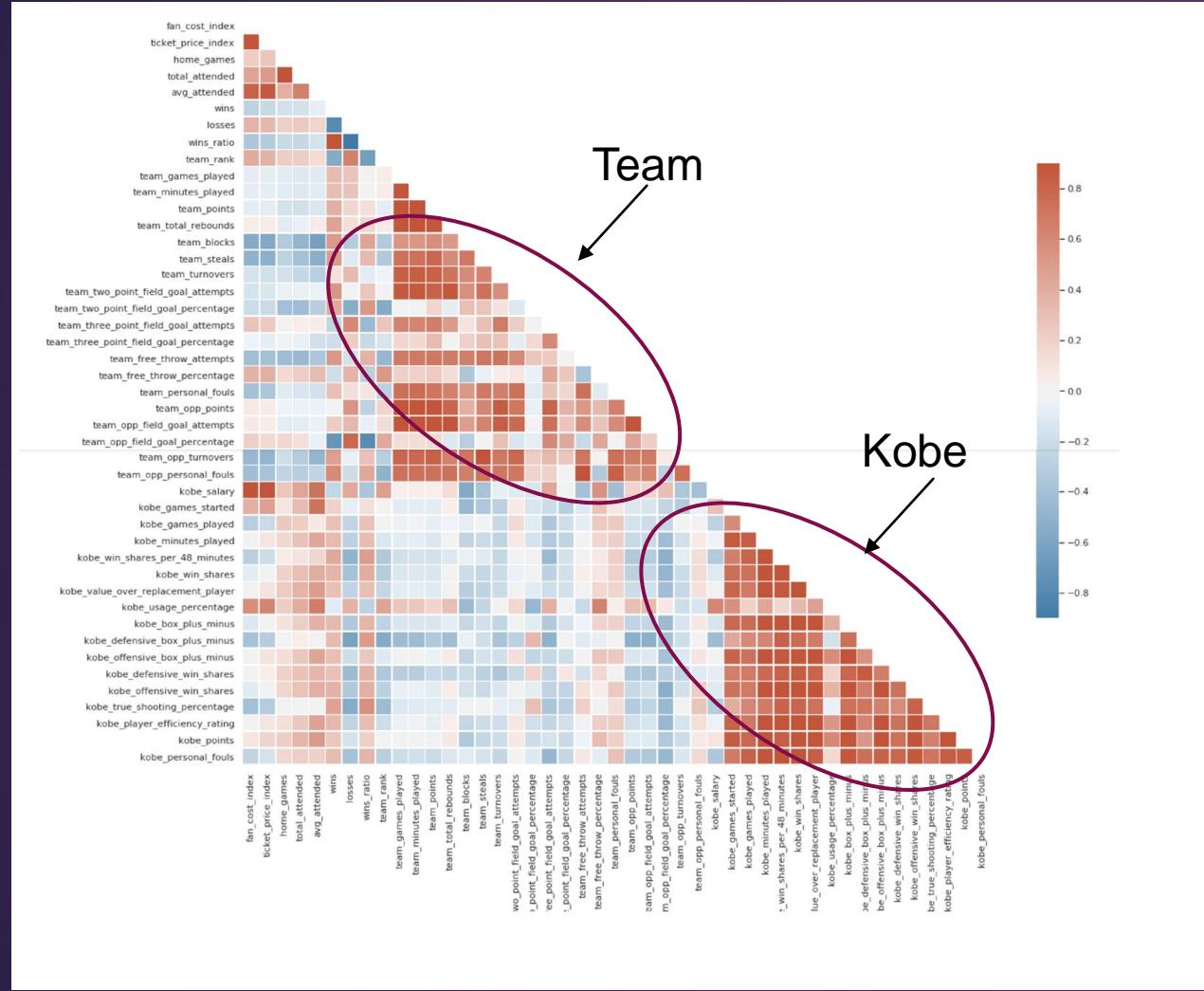
- Combined all relevant team, and Kobe data through merge function
  - Operating income and franchise value
  - Fan cost and ticket price index
  - Attendance
  - Record (wins, losses, wins ratio)
  - Team data (team statistics)
  - Kobe data (Kobe's statistics including his salary)
- Generated a normalized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

# Pre-Processing and Data Cleaning – Part 2

## Final Data Frame

| GENERAL            | TEAM                                   | KOBE                               |
|--------------------|--|------------------------------------|
| Operating Income   | Team Rank                              | Kobe Salary                        |
| Franchise Value    | Team Games Played                      | Kobe Games Started                 |
| Fan Cost Index     | Team Minutes Played                    | Kobe Games Played                  |
| Ticket Price Index | Team Points                            | Kobe Minutes Played                |
| Home Games         | Teal Total Rebounds                    | Kobe Win Share Per 48 Minutes      |
| Total Attended     | Team Blocks                            | Kobe Win Shares                    |
| Average Attended   | Team Steals                            | Kobe Value Over Replacement Player |
| Wins               | Team Turnovers                         | Kobe Usage Percentage              |
| Losses             | Team Two Point Field Goal Attempts     | Kobe Box Plus Minus                |
| Wins Ratio         | Team Two Point Field Goal Percentage   | Kobe Defensive Box Plus Minus      |
|                    | Team Three Point Field Goal Attempts   | Kobe Offensive Box Plus Minus      |
|                    | Team Three Point Field Goal Percentage | Kobe Defensive Win Shares          |
|                    | Team Free Throw Attempts               | Kobe Offensive Win Shares          |
|                    | Team Personal Fouls                    | Kobe True Shooting Percentage      |
|                    | Team Opponent Points                   | Kobe Player Efficiency Rating      |
|                    | Team Opponent Field Goal Attempts      | Kobe Points                        |
|                    | Team Opponent Field Goal Percentage    | Kobe Personal Fouls                |
|                    | Team Opponent Turnovers                |                                    |
|                    | Team Opponent Personal Fouls           |                                    |

# Exploring the Dataset – Part 2



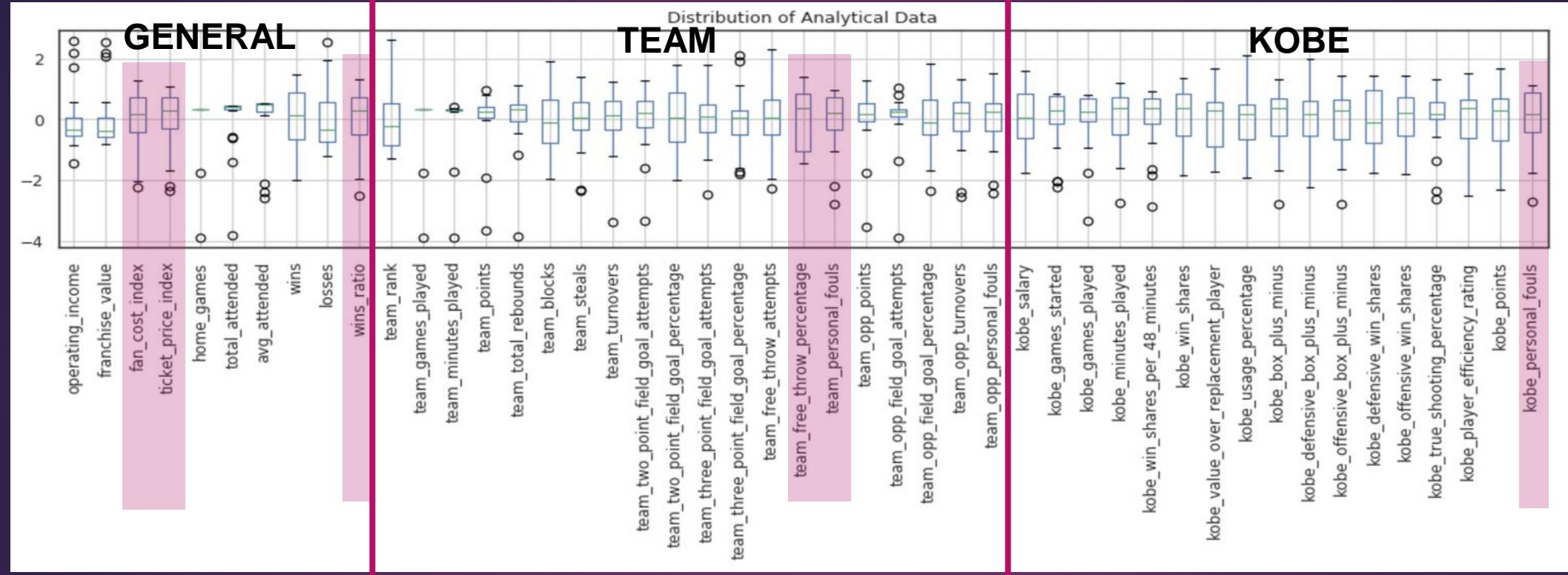
The team performance are correlated with itself and not with Kobe's.

Ticket prices and general fan costs are positively correlated with Kobe's salary.

Opponent's field goal % is closely correlated with Lakers' wins and losses.

Avg. attendance is slightly negatively correlated with team blocks and slightly positively correlated with Kobe's games played.

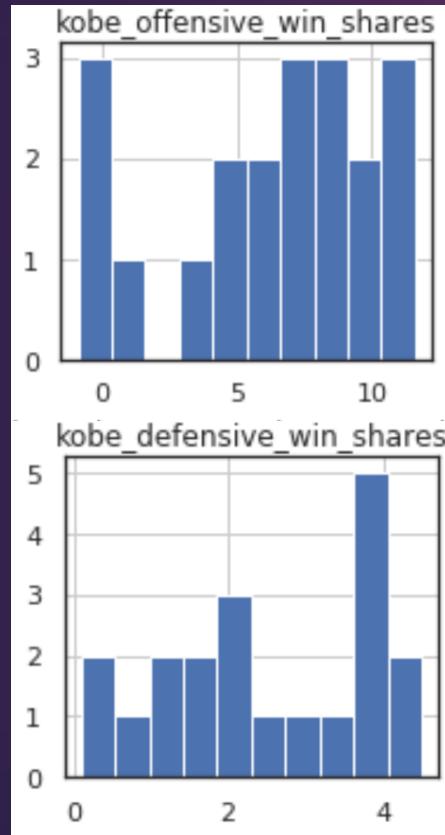
# Exploring the Dataset – Part 2



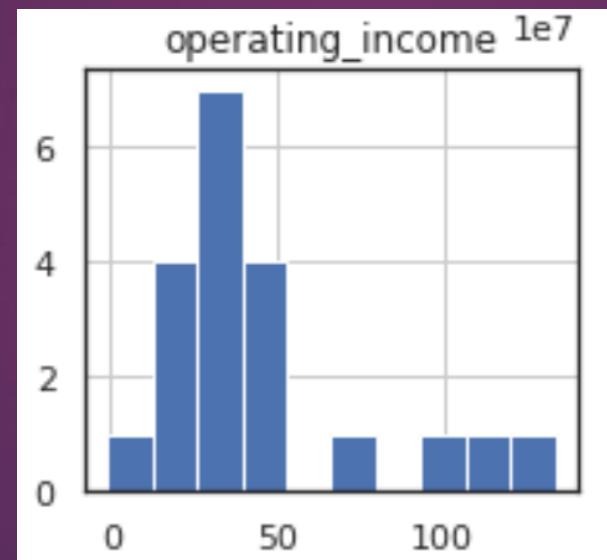
- Most team performance are normally distributed except for personal fouls and free throw percentage.
- Ticket price and fan cost indices are negatively skewed, but attendance varies only slightly throughout time.
- Most of the Lakers final ranking are on the lower side outside of some good years.
- Kobe's statistics are all relatively normally distributed expect for personal fouls, that are negatively skewed.
- All aggregate measures of Kobe's performance show positive results for most of his career.

# Exploring the Dataset – Part 2

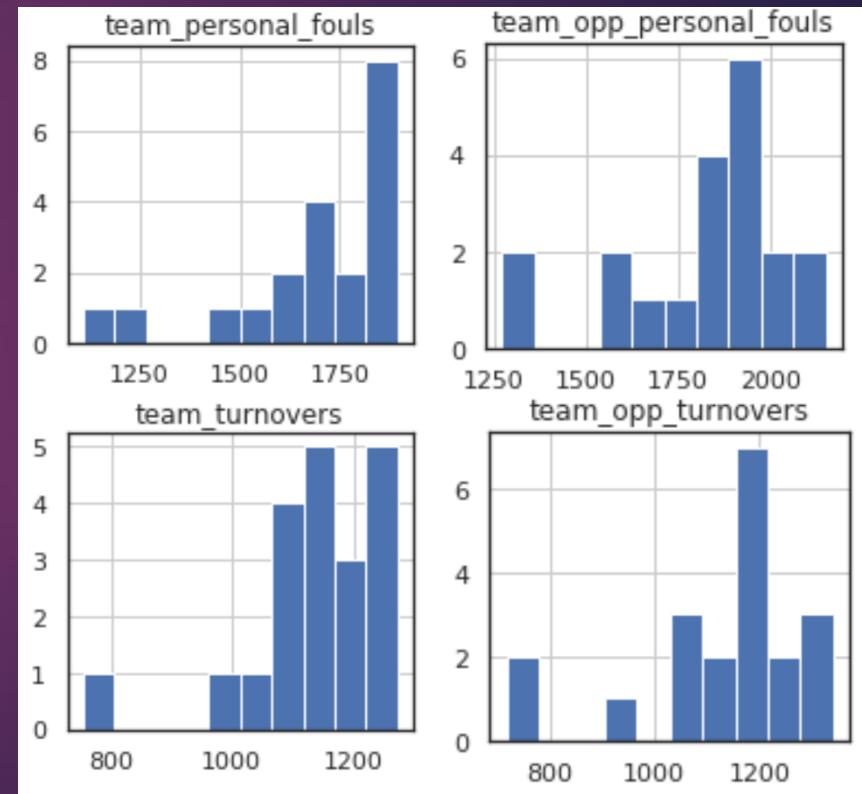
Kobe appears to have a stronger offensive win shares rate than defensive win shares rate.



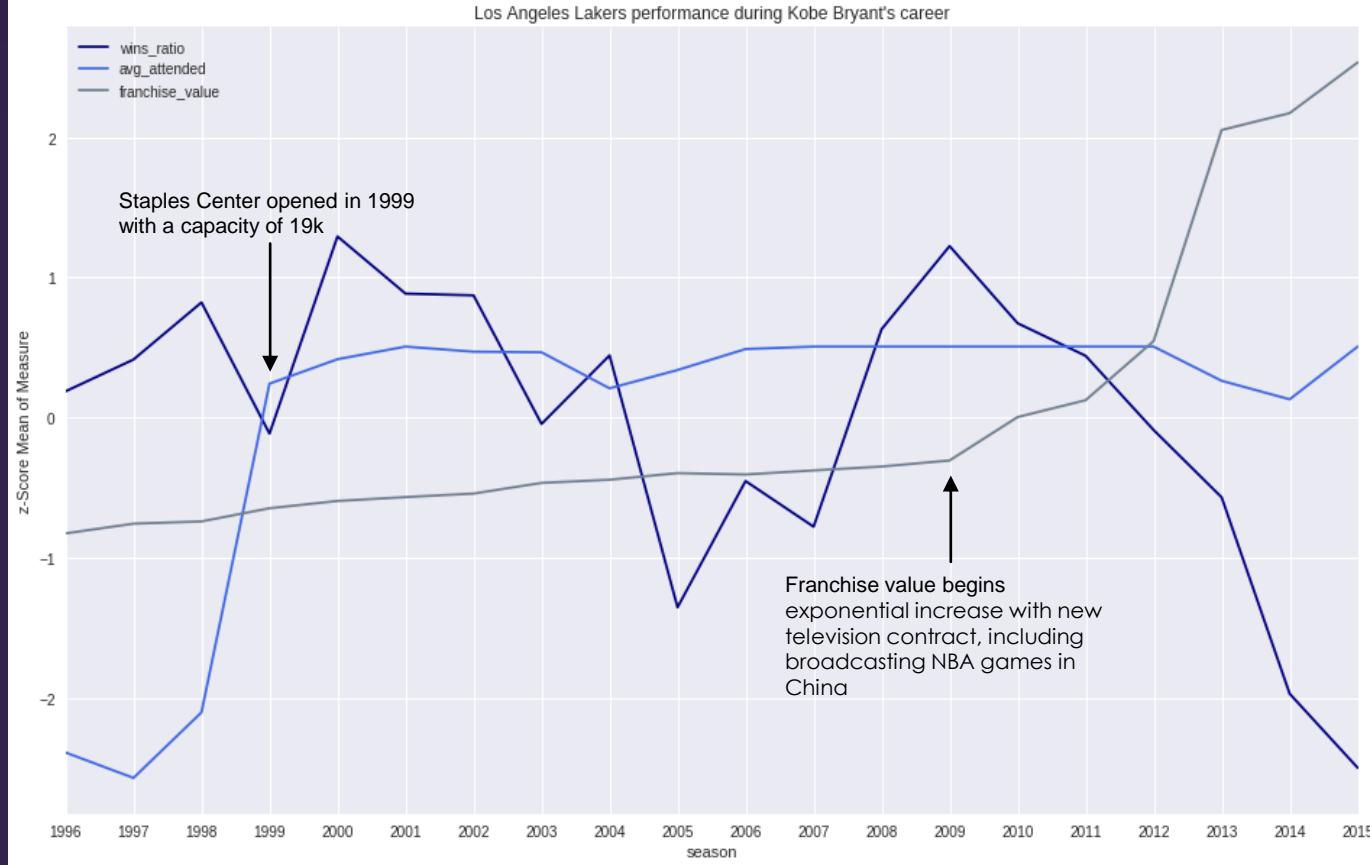
Operating income centralizes between \$0 to \$50M during most years except for four high years.



Team personal fouls and turnovers are negatively skewed, more so than their opponents.



# Exploring the Dataset – Part 2



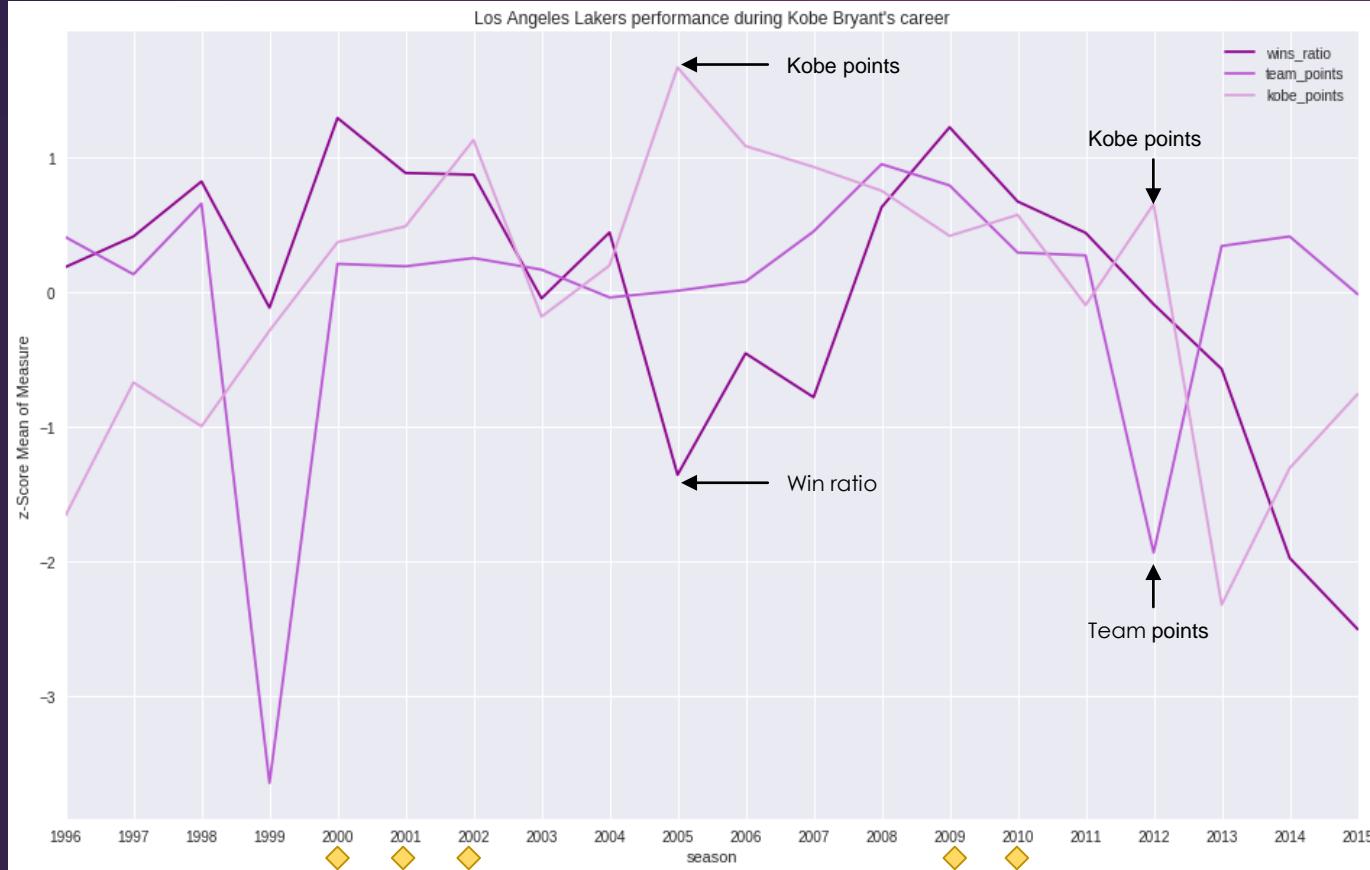
Average attendance totals appear to be relatively flat throughout much of Kobe's career.

Franchise value increases steadily from 1996, with an exponential increase starting in 2009.

Franchise value appears unaffected by changes in attendance, wins, etc.

Attendance appears to fluctuate most at the end of Kobe's career.

# Exploring the Dataset – Part 2



Team wins fell off sharply toward the end of Kobe's career, more often from his injuries and inversion of points.

Kobe's points were inverted from team points in most of the final years of his career.

The first year at the Staples Center showcased a mediocre lack of team wins and the lowest amount of team points throughout all of Kobe's career.

# Exploring the Dataset – Part 2

|                                      | finals_ind | 0         | 1 |
|--------------------------------------|------------|-----------|---|
| team_rank                            | 0.594464   | -0.891697 |   |
| home_games                           | -0.060324  | 0.090485  |   |
| team_points                          | 0.055756   | -0.083633 |   |
| team_total_rebounds                  | 0.090354   | -0.135531 |   |
| team_blocks                          | -0.055033  | 0.082550  |   |
| team_steals                          | -0.116010  | 0.174014  |   |
| team_turnovers                       | 0.183330   | -0.274995 |   |
| team_two_point_field_goal_attempts   | -0.015260  | 0.022890  |   |
| team_three_point_field_goal_attempts | 0.308013   | -0.462019 |   |
| team_free_throw_attempts             | -0.176891  | 0.265336  |   |
| team_personal_fouls                  | -0.101932  | 0.152898  |   |
| kobe_games_started                   | -0.253937  | 0.380906  |   |
| kobe_minutes_played                  | -0.240543  | 0.360815  |   |
| kobe_value_over_replacement_player   | -0.243825  | 0.365738  |   |
| kobe_defensive_box_plus_minus        | -0.426647  | 0.639970  |   |
| kobe_offensive_box_plus_minus        | -0.195789  | 0.293684  |   |

When the Lakers finish as one of the top 4 teams in the NBA (`finals_ind = 1`), some notable differences emerge in team performance.

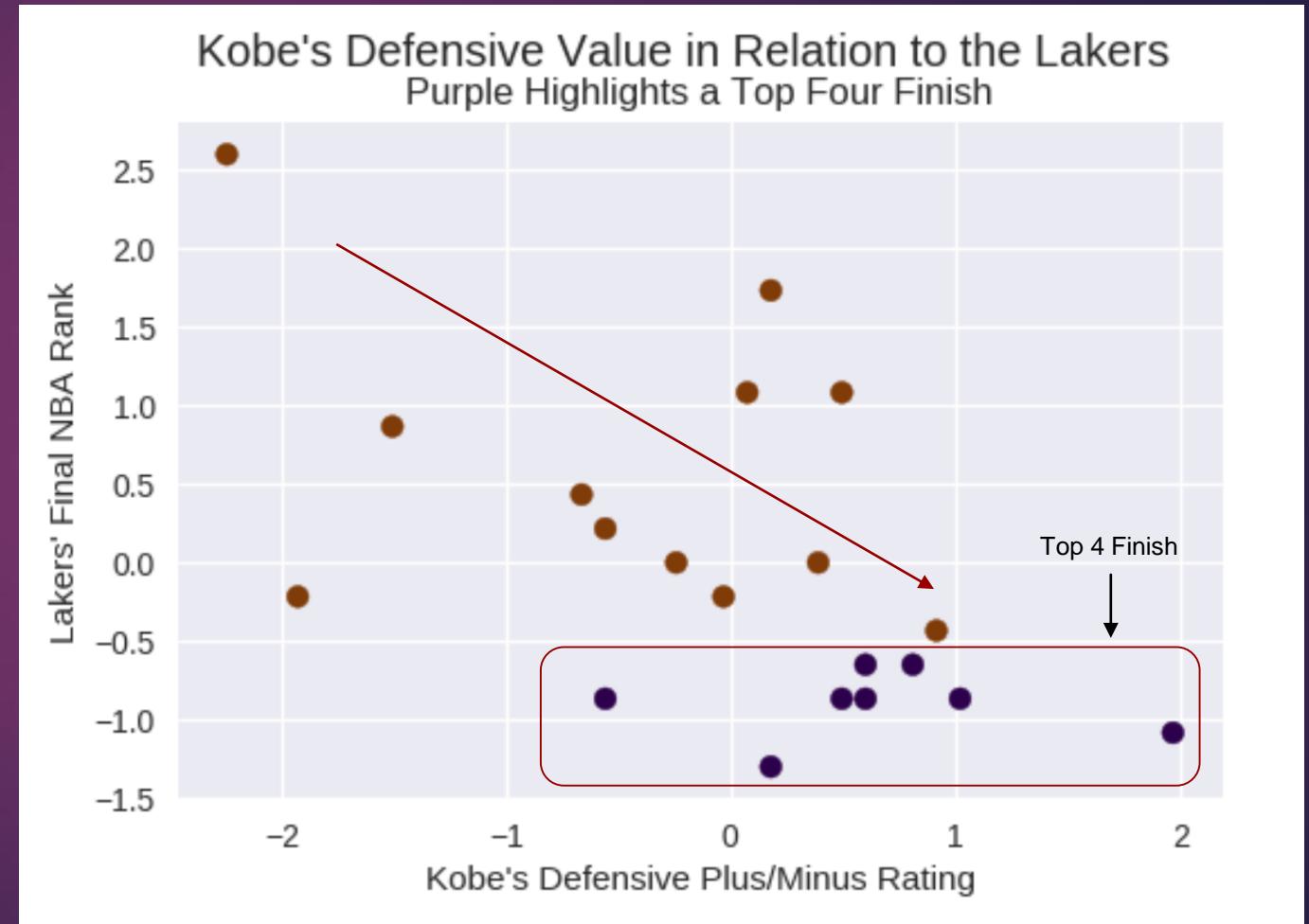
In years the Lakers finished in the top 4, team points were slightly lower, but steals and rebounds were much improved from other years.

Kobe's measures are generally more susceptible to change. Years in which his 'Defensive Plus/Minus' was significantly better signaled a top four team finish.

# Exploring the Dataset – Part 2

Expanding upon the prior slides, Kobe's defensive performance appears linearly correlated with better overall team performance (as measured by final NBA rank).

The linear relationship is not perfect, but does highlight an overall trend in most years. This measure can be affected by additional support from other star players on the Lakers (ex. Shaquille O'Neal)

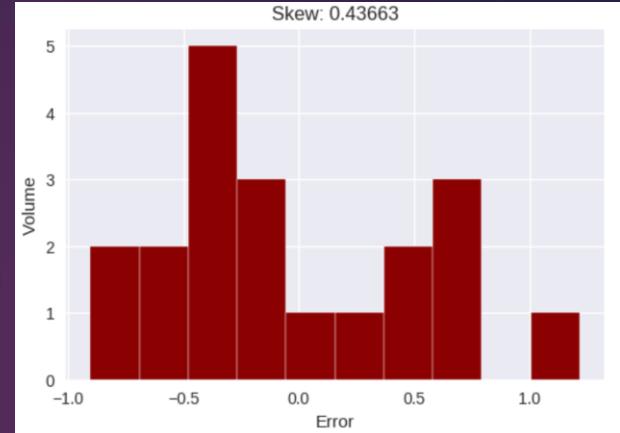
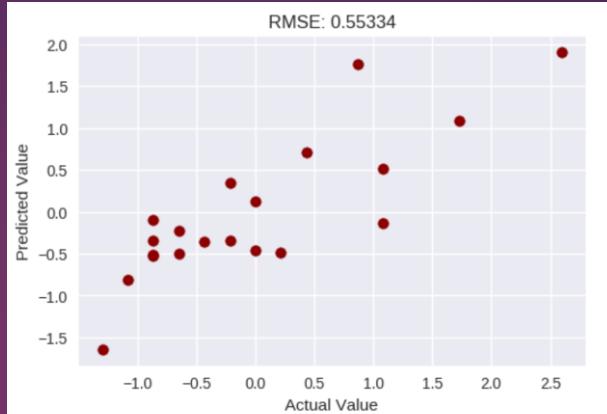


# Model – OLS Regression – Part 2

| Version     | Description                    | R-squared |
|-------------|--------------------------------|-----------|
| Version 1.0 | First run                      | 0.846     |
| Version 1.1 | Top 3 highest p-values removed | 0.754     |
| Version 1.2 | Next top 2 highest p-values    | 0.712     |
| Version 1.3 | P-value above 0.1 removed      | 0.694     |
| Version 1.4 | P-value above 0.05 removed     | 0.692     |

Version 1.4:

| OLS Regression Results        |                  |                     |           |       |        |        |
|-------------------------------|------------------|---------------------|-----------|-------|--------|--------|
| Dep. Variable:                | team_rank        | R-squared:          | 0.692     |       |        |        |
| Model:                        | OLS              | Adj. R-squared:     | 0.549     |       |        |        |
| Method:                       | Least Squares    | F-statistic:        | 4.860     |       |        |        |
| Date:                         | Thu, 05 Mar 2020 | Prob (F-statistic): | 0.00819   |       |        |        |
| Time:                         | 02:12:35         | Log-Likelihood:     | -16.613   |       |        |        |
| No. Observations:             | 20               | AIC:                | 47.23     |       |        |        |
| Df Residuals:                 | 13               | BIC:                | 54.20     |       |        |        |
| Df Model:                     | 6                |                     |           |       |        |        |
| Covariance Type:              | nonrobust        |                     |           |       |        |        |
|                               |                  |                     |           |       |        |        |
|                               | coef             | std err             | t         | P> t  | [0.025 | 0.975] |
| Intercept                     | -3.469e-18       | 0.154               | -2.25e-17 | 1.000 | -0.333 | 0.333  |
| team_blocks                   | 0.1527           | 0.261               | 0.586     | 0.568 | -0.410 | 0.716  |
| team_turnovers                | 0.2646           | 0.276               | 0.959     | 0.355 | -0.331 | 0.861  |
| team_free_throw_attempts      | -0.8256          | 0.281               | -2.937    | 0.012 | -1.433 | -0.218 |
| team_personal_fouls           | 0.0500           | 0.265               | 0.189     | 0.853 | -0.522 | 0.622  |
| kobe_games_started            | 0.3366           | 0.223               | 1.512     | 0.154 | -0.144 | 0.818  |
| kobe_defensive_box_plus_minus | -0.6817          | 0.199               | -3.420    | 0.005 | -1.112 | -0.251 |
|                               |                  |                     |           |       |        |        |
| Omnibus:                      | 0.975            | Durbin-Watson:      | 2.252     |       |        |        |
| Prob(Omnibus):                | 0.614            | Jarque-Bera (JB):   | 0.896     |       |        |        |
| Skew:                         | 0.440            | Prob(JB):           | 0.639     |       |        |        |
| Kurtosis:                     | 2.453            | Cond. No.           | 3.94      |       |        |        |



Linear relationship exists between Kobe's and the Lakers performance and the final NBA ranking.

Team Free Throw Attempts and Kobe's Defensive Plus or Minus Ratio are the strongest predictors.

Model underpredicts by approximately half of a std dev.

RMSE is bimodal with a higher propensity to predict a slightly better than actual rank.

# Time Series – Part 2

Prior to the opening of the Staples Center, the Lakers could have experienced a variety of attendance results. The time series was not stationary, even after log transformation.

As shown below, four periods of attendance change occurred from 1970 through 1993. If included the opening of the Staples Center, in 1999, would have been noted as a significant change.

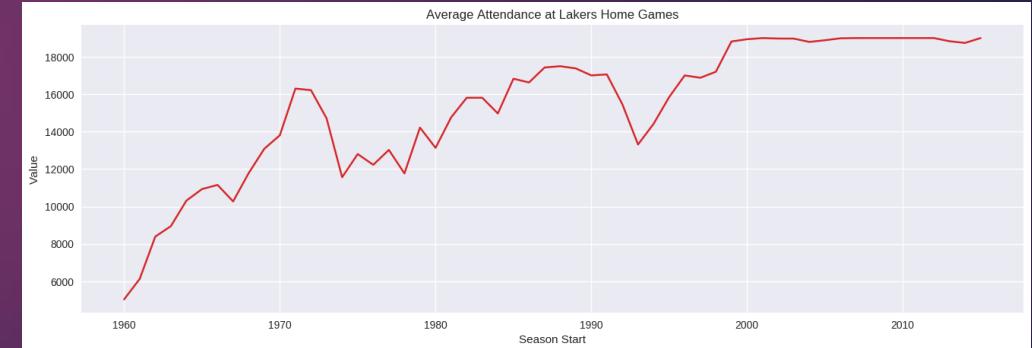
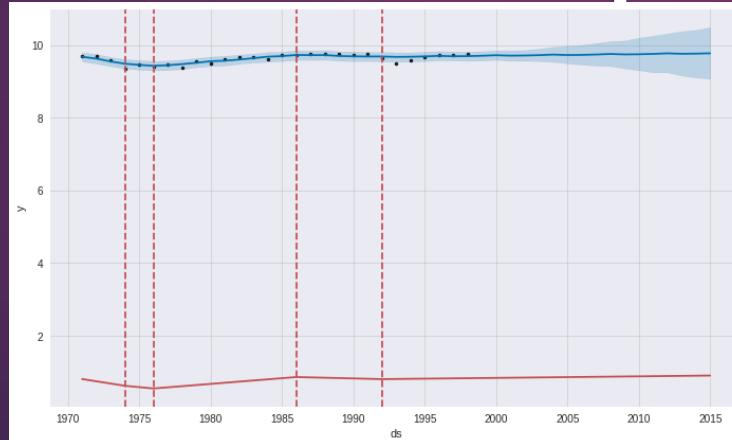
The Lakers ultimately averaged a near-sellout during the peak of Kobe's career, only slightly decreasing at the tail end. The FB Prophet forecast produced a RMSE of +1,424 people from '99

```
-- ACTUAL DATA VALUES --
ADF Statistic: -1.884731
p-value: 0.339278

-- Critical Values --
1%: -3.578
5%: -2.925
10%: -2.601

-- LOG TRANSFORMED (e) DATA VALUES --
ADF Statistic: -2.091609
p-value: 0.247902

-- Critical Values --
1%: -3.578
5%: -2.925
10%: -2.601
```

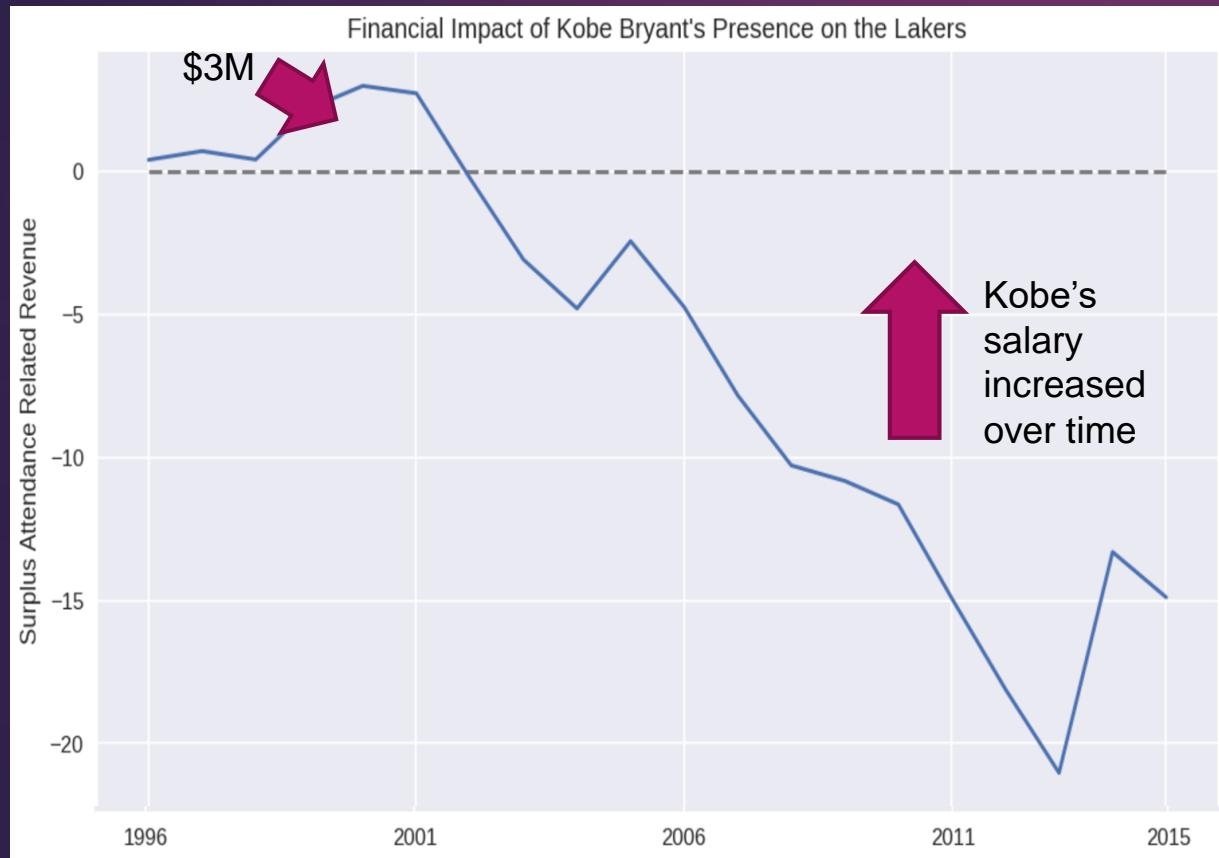


Data Source: Attendance – Part  
2

# Exploring the Dataset – Part 2

```
# Create a calculated ticket revenue surplus, sans Kobe's Salary  
finance['kobe_surplus'] = ((finance.home_games * finance.error * finance.fan_cost_index * (finance.kobe_usage_percentage / 100)) - finance.kobe_salary) / 1000000
```

Million USD



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.

# Conclusion

Kobe's jump shots with majority of the shots only had 39% success rate, while dunk had the highest success rate at 93%.

Shot 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.

## Recommendations

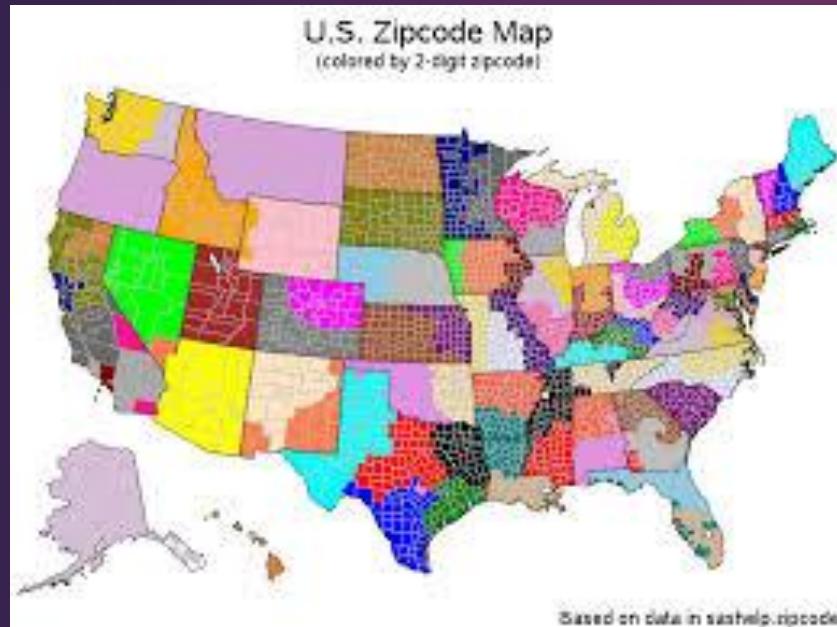
Explore the upward trend of the franchise value after Y2010 .

Comparison of his statistics and impact to team's profitability as compared to other basketball legends and teams (e.g. Michael Jordan, LeBron James, 1988 to 1990 Detroit Pistons, etc.).

Kobe's defensive winning style as compared to other player's style in winning a game.

Extrapolate on how a weighted shot has an impact to overall team win as compared to other measurements.

Extrapolate how intermingling younger players, with lower salaries, can augment star players to create a cost effective but high performing (championship) team.



# Three Zip Codes in the US to Invest

# Motivation

Volatility of the housing market in the US.

Application of latest technique in forecasting with data science.

Which three zip codes to invest given historical data and other factors to consider.

Factors to consider in recommending a decision.

## **Business Questions**

# The Collection of Data

---

## Zillow Median Home Values

- [files.zillowstatic.com/research/public/Zip/Zip\\_Zhvi\\_SingleFamilyResidence.csv](https://files.zillowstatic.com/research/public/Zip/Zip_Zhvi_SingleFamilyResidence.csv)

## Median Household Income for Y2018

- <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 and Data Cleaning

---

## Dealing with Missing Values, Cleaning and Transforming Data

### Zillow Data:

- Dropped data from Apr 1996 to Dec 1996.
- Added “Growth” rate column by taking the Dec 2019 data and dividing it by the Jan 1997 data.
- Added “std” or standard deviation column by taking data for the past 10 years from Jan 2010 to Dec 2019.
- Added “mean” column by taking data for the past 10 years from Jan 2010 to Dec 2019.

### Median Household Income for Y2018:

- Retained only columns County Name, State and Median Income.

### Crime Rates:

- Retained only columns County Name, State and Crime Rate Income.

# Pre-Processing and Data Cleaning

---

## Merging of Data

First merge:

- Merged Zillow and Crime Rate data frames joining through “County Name” and “State” columns.
- 746 observations were dropped from 30,434 to 29,688.

Second merge:

- Merged “First merge” with the Median Household Income joining through “County Name” and “State” columns.
- This time, 29,519 observations were left or only 169 observations were dropped from the first merge.

# Pre-Processing and Data Cleaning

---

## Further Cleaning Transformation After Merging Data

- Kept rows with at least 60% non-NA values, which resulted to 23,044 observations left.
- Removed rows with crime rates of more than 1.27% (taken from 25% quantile of Crime Rate data), which resulted to 3,954 observations.
- Removed rows with median income of at most \$59,000 (taken from 75% quantile of Median Household Income data), which resulted to 1,151 observations.

# Pre-Processing and Data Cleaning

---

## Transformation After Choosing Top 10 Zip Codes

- Dates as columns had been reset as index.
- Re-assigned index as “ds” which is a requirement in using fbprophet.
- Converted new index to datetime.
- Re-assigned zip codes as “y.”
- Generated fbprophet output **functions** (can be seen from Jupyter Notebook file) for the Top 10 Zip Codes

# Exploratory Data Analysis

Best 20 Zipcodes:

| RegionName | Metro | CountyName                     | City               | State           | Growth | mean      | std          |               |
|------------|-------|--------------------------------|--------------------|-----------------|--------|-----------|--------------|---------------|
| 2331       | 90403 | Los Angeles-Long Beach-Anaheim | Los Angeles County | Santa Monica    | CA     | 13.914982 | 1.092774e+06 | 518560.866240 |
| 9350       | 91108 | Los Angeles-Long Beach-Anaheim | Los Angeles County | San Marino      | CA     | 13.776618 | 9.497870e+05 | 470036.233530 |
| 9981       | 90211 | Los Angeles-Long Beach-Anaheim | Los Angeles County | Beverly Hills   | CA     | 12.554974 | 9.871943e+05 | 494881.970778 |
| 1088       | 90020 | Los Angeles-Long Beach-Anaheim | Los Angeles County | Los Angeles     | CA     | 12.272559 | 1.488784e+06 | 635028.946798 |
| 331        | 90027 | Los Angeles-Long Beach-Anaheim | Los Angeles County | Los Angeles     | CA     | 11.977903 | 7.680990e+05 | 335516.443857 |
| 253        | 90004 | Los Angeles-Long Beach-Anaheim | Los Angeles County | Los Angeles     | CA     | 11.318958 | 7.775213e+05 | 338011.493077 |
| 2251       | 90266 | Los Angeles-Long Beach-Anaheim | Los Angeles County | Manhattan Beach | CA     | 10.948580 | 1.071792e+06 | 507941.211850 |
| 2733       | 90048 | Los Angeles-Long Beach-Anaheim | Los Angeles County | Los Angeles     | CA     | 10.497167 | 9.961686e+05 | 406054.897367 |
| 1626       | 94610 | San Francisco-Oakland-Hayward  | Alameda County     | Oakland         | CA     | 10.286459 | 8.363412e+05 | 401629.738911 |
| 4468       | 91030 | Los Angeles-Long Beach-Anaheim | Los Angeles County | South Pasadena  | CA     | 10.195877 | 6.273837e+05 | 268614.706156 |
| 2106       | 90069 | Los Angeles-Long Beach-Anaheim | Los Angeles County | West Hollywood  | CA     | 9.604911  | 1.297958e+06 | 533324.387651 |
| 7396       | 90212 | Los Angeles-Long Beach-Anaheim | Los Angeles County | Beverly Hills   | CA     | 9.271800  | 1.453807e+06 | 614666.562734 |
| 43         | 90046 | Los Angeles-Long Beach-Anaheim | Los Angeles County | Los Angeles     | CA     | 9.170598  | 9.438518e+05 | 372731.539802 |
| 1307       | 90005 | Los Angeles-Long Beach-Anaheim | Los Angeles County | Los Angeles     | CA     | 9.067142  | 8.978202e+05 | 344777.433684 |
| 589        | 90036 | Los Angeles-Long Beach-Anaheim | Los Angeles County | Los Angeles     | CA     | 8.657940  | 9.829735e+05 | 378354.297132 |
| 2032       | 90291 | Los Angeles-Long Beach-Anaheim | Los Angeles County | Los Angeles     | CA     | 8.409081  | 1.038788e+06 | 412799.270720 |
| 514        | 11216 | New York-Newark-Jersey City    | Kings County       | New York        | NY     | 8.332922  | 1.047424e+06 | 381081.572991 |
| 2814       | 90068 | Los Angeles-Long Beach-Anaheim | Los Angeles County | Los Angeles     | CA     | 8.272667  | 8.431533e+05 | 319368.055118 |
| 5310       | 90210 | Los Angeles-Long Beach-Anaheim | Los Angeles County | Beverly Hills   | CA     | 8.216722  | 2.298412e+06 | 868016.499644 |
| 8143       | 90401 | Los Angeles-Long Beach-Anaheim | Los Angeles County | Santa Monica    | CA     | 8.152619  | 1.228189e+06 | 414800.255759 |

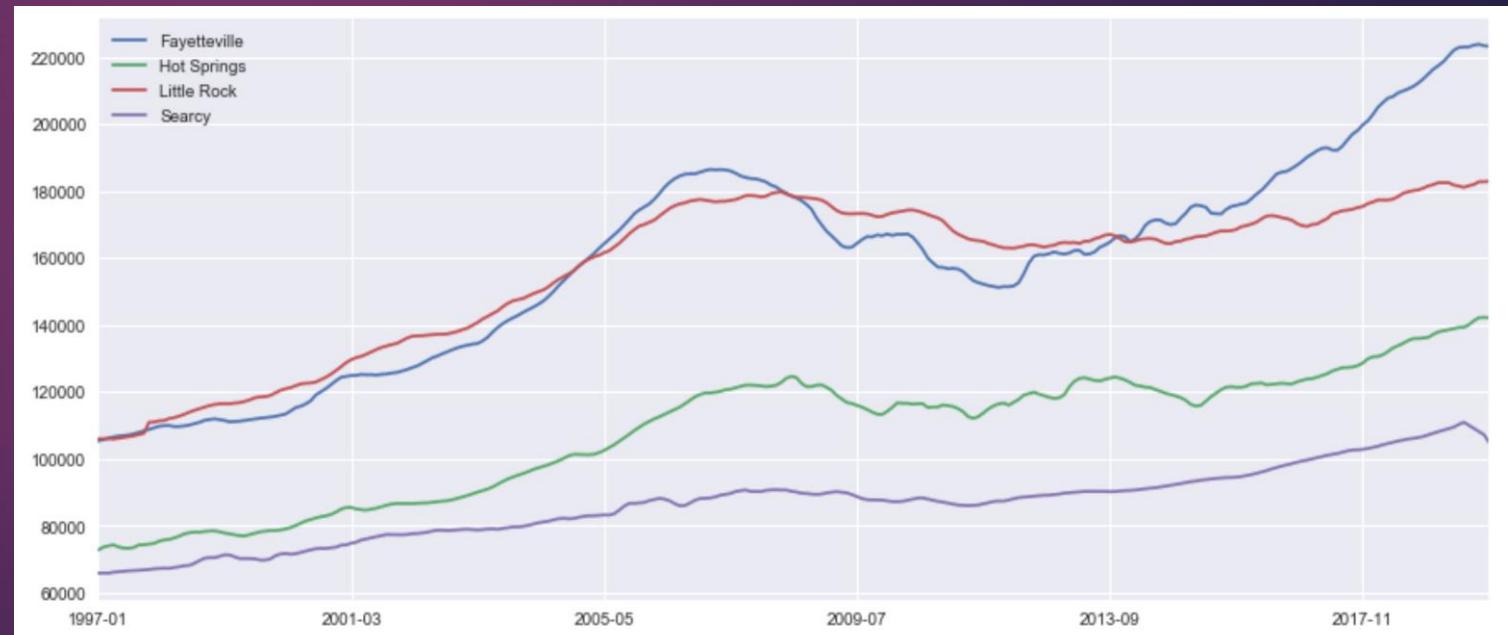
From the top 20 zip codes nationwide based on growth rates, majority of the zip code were all out of California and only 1 few from New York.

# Exploratory Data Analysis

Time Series plot for Hot Springs, Little Rock, Fayetteville, and Searcy in Arkansas.

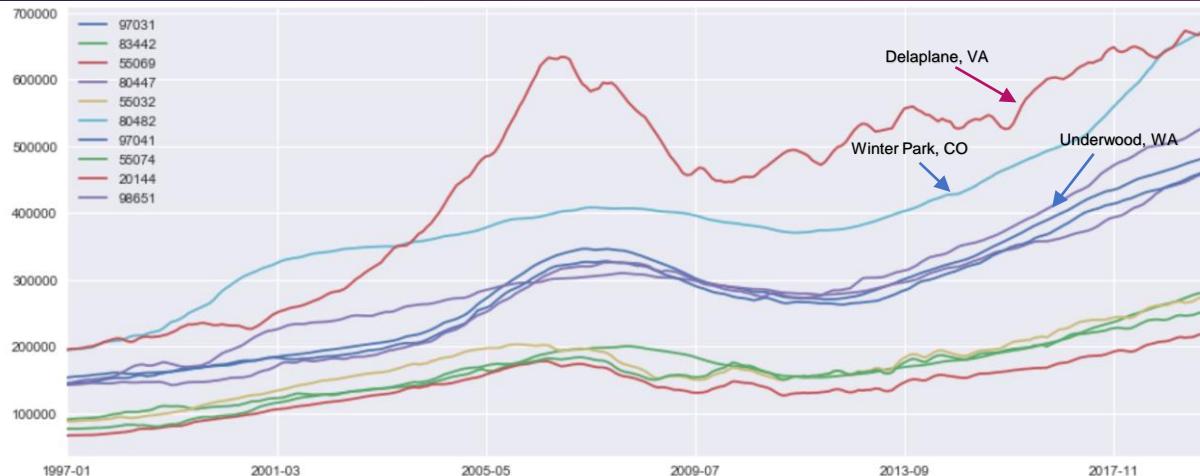
Best Zipcodes:

| City         | Growth   | std          | mean          |
|--------------|----------|--------------|---------------|
| Fayetteville | 1.130409 | 21893.753876 | 179309.294444 |
| Hot Springs  | 0.996877 | 7905.089083  | 123511.337500 |
| Little Rock  | 0.638650 | 7509.868394  | 170605.392308 |
| Searcy       | 0.557487 | 7454.500756  | 95253.145833  |



# Exploratory Data Analysis

Top 10 Zip Codes from the final data frame after merging all the data sorting from the highest growth rate.



| Best 10 Zipcodes: |            |                                  |                   |                     |       |          |              |               |  |
|-------------------|------------|----------------------------------|-------------------|---------------------|-------|----------|--------------|---------------|--|
|                   | RegionName | Metro                            | CountyName        | City                | State | Growth   | std          | mean          |  |
| 28245             | 98651      | Portland-Vancouver-Hillsboro     | Skamania County   | Underwood           | WA    | 2.744479 | 86548.903408 | 372800.925000 |  |
| 27363             | 80482      |                                  | Grand County      | Winter Park         | CO    | 2.490861 | 98718.041136 | 470620.433333 |  |
| 20006             | 20144      | Washington-Arlington-Alexandria  | Fauquier County   | Delaplane           | VA    | 2.468982 | 68402.012014 | 558770.791667 |  |
| 21239             | 55074      | Minneapolis-St. Paul-Bloomington | Chisago County    | Shafer              | MN    | 2.353474 | 31500.079045 | 193209.441667 |  |
| 21237             | 55069      | Minneapolis-St. Paul-Bloomington | Chisago County    | Rush City           | MN    | 2.310456 | 28099.555029 | 163006.958333 |  |
| 17861             | 97041      |                                  | Hood River County | Mount Hood Parkdale | OR    | 2.230471 | 68928.748286 | 338221.266667 |  |
| 21238             | 55032      | Minneapolis-St. Paul-Bloomington | Chisago County    | Harris              | MN    | 2.214750 | 40093.706509 | 201372.283333 |  |
| 19144             | 83442      |                                  | Jefferson County  | Rigby               | ID    | 2.179051 | 40767.972766 | 197067.866667 |  |
| 17860             | 97031      |                                  | Hood River County | Hood River          | OR    | 2.175693 | 73846.796956 | 353977.675000 |  |
| 27362             | 80447      |                                  | Grand County      | Grand Lake          | CO    | 2.130602 | 58312.946847 | 339919.325000 |  |

From the top 10 zip codes nationwide based on growth rates, 3 came out of Minnesota, 2 from Oregon, 2 from Colorado, 1 from Washington, 1 from Virginia, and 1 from Idaho .

Top 3 in terms of growth rates were Underwood, WA, Winter Park, CO and Delaplane, VA, which grew by around 250% from Jan 2017 to Dec 2019.

# Models

Showing results from running the fbprophet of top 3 zip codes based on highest growth rates from Jan 1997 to Dec 2019. (**Note: Results from the remaining zip codes can be found from the Jupyter Notebook file.**)

## Zip Code 98651 – Underwood, Washington



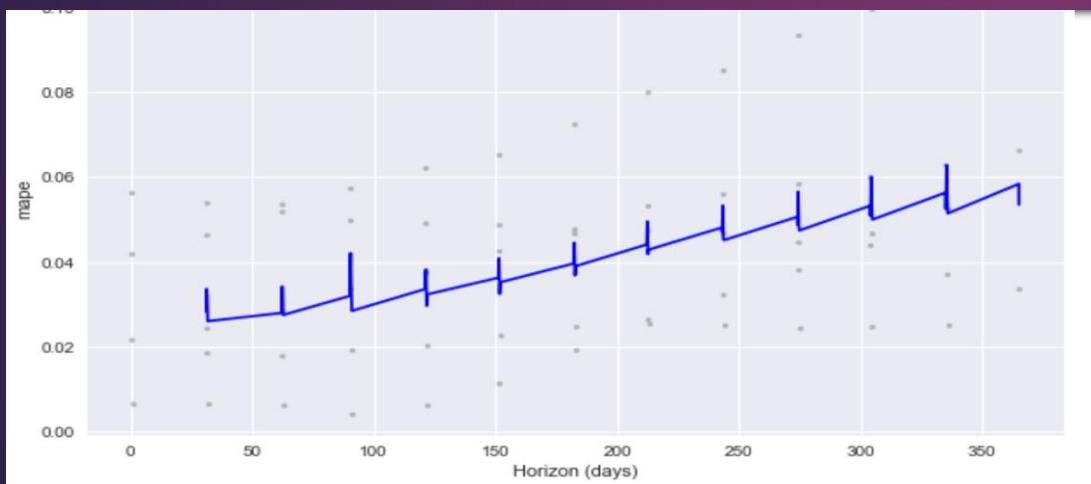
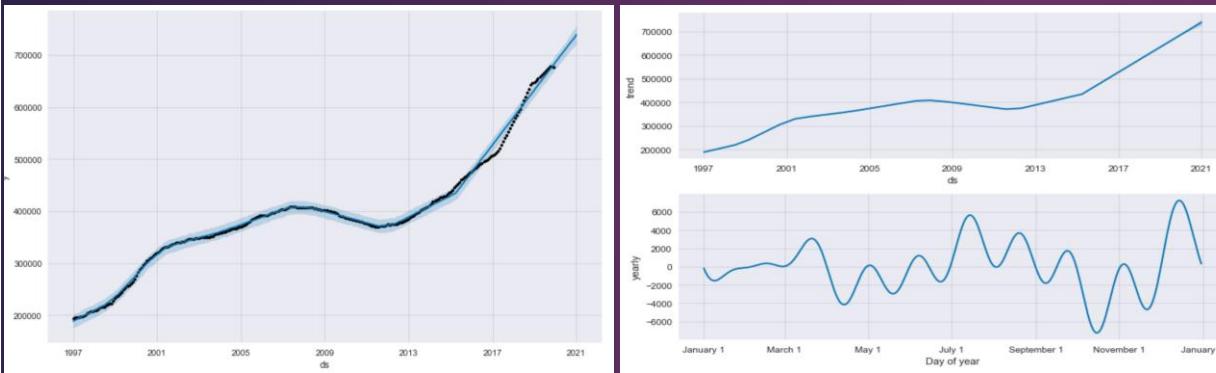
Underwood, WA has the highest growth rate of 274% from Jan 1997 to Dec 2019 with a median house value of \$530K as of Dec 2019.

We used 12 months or 365 days as horizon in predicting performance, and 204 months prior as training data.

Picking on MAPE over 365 days horizon, forecast error range is from 2% to 4% (see graph).

# Models

## Zip Code 80482 – Winter Park, Colorado



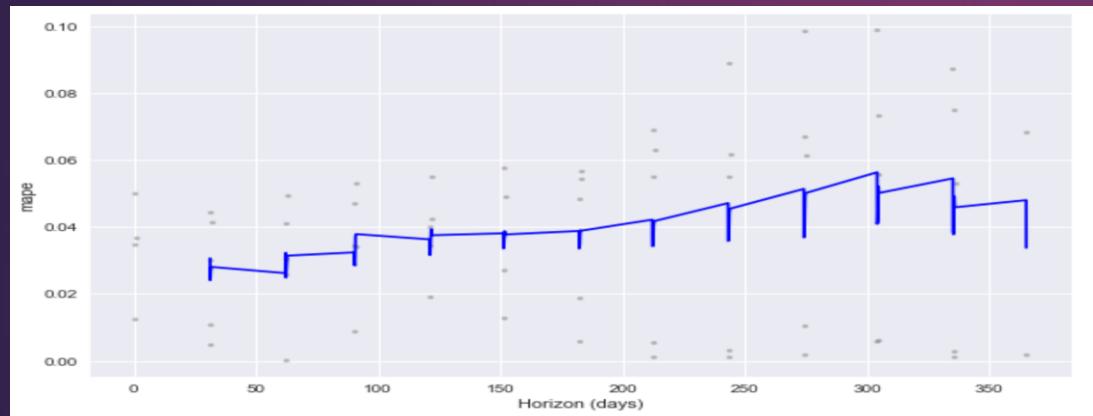
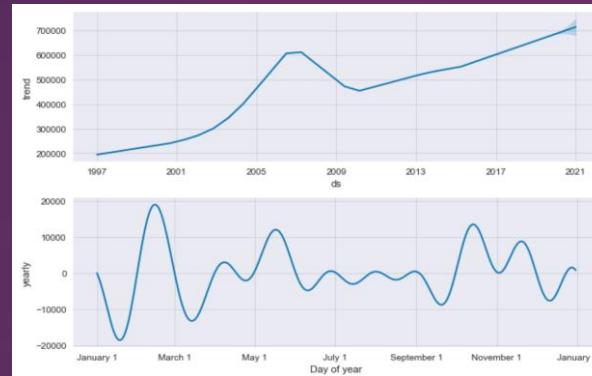
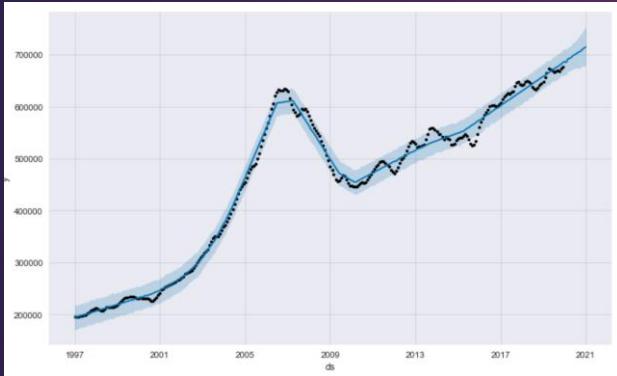
Winter Park, CO has the second highest growth rate of 249% from Jan 1997 to Dec 2019 with a median house value of \$676K as of Dec 2019.

We used 12 months or 365 days as horizon in predicting performance, and 204 months prior as training data.

Picking on MAPE over 365 days horizon, we see that forecast error range is from 2% to 6% (see graph).

# Models

## Zip Code 20144 – Delaplane, Virginia



Delaplane, VA has the third highest growth rate of 247% from Jan 1997 to Dec 2019 with a median house value of \$676K as of Dec 2019.

We used 12 months or 365 days as horizon in predicting performance, and 204 months prior as training data.

Picking on MAPE over 365 days horizon, we see that forecast error range is from 2% to 5% (see graph).

# Models

## 12 Months Forecast – Y2020

|   | RegionID | RegionName | City                | State | Metro                            | CountyName        | SizeRank | 2020-01-31 00:00:00 | 2020-02-29 00:00:00 | 2020-03-31 00:00:00 | 2020-04-30 00:00:00 | 2020-05-31 00:00:00 | 2020-06-30 00:00:00 | 2020-07-31 00:00:00 | 2020-08-31 00:00:00 | 2020-09-30 00:00:00 | 2020-10-31 00:00:00 | 2020-11-30 00:00:00 | 2020-12-31 00:00:00 |
|---|----------|------------|---------------------|-------|----------------------------------|-------------------|----------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|
| 0 | 99945    | 98651      | Underwood           | WA    | Portland-Vancouver-Hillsboro     | Skamania County   | 22693    | 550465.839292       | 554613.452136       | 556657.632906       | 559471.671146       | 562697.056226       | 566112.211199       | 570211.351249       | 573733.514426       | 576668.900097       | 579026.281911       | 581424.974755       | 586556.431486       |
| 1 | 93426    | 80482      | Winter Park         | CO    | Nan                              | Grand County      | 15679    | 689779.459710       | 694135.172726       | 698922.658280       | 702956.013191       | 707318.989494       | 711743.653492       | 716446.360678       | 720876.079826       | 725208.074245       | 729089.177895       | 732956.831768       | 738574.654143       |
| 2 | 66240    | 20144      | Delaplane           | VA    | Washington-Arlington-Alexandria  | Fauquier County   | 21925    | 685138.293561       | 691919.380750       | 692702.053607       | 695173.754188       | 698467.065996       | 700223.813282       | 702630.362966       | 705082.883795       | 706304.316284       | 709871.948399       | 712540.598814       | 714668.616186       |
| 3 | 82086    | 55074      | Shafer              | MN    | Minneapolis-St. Paul-Bloomington | Chisago County    | 18169    | 253644.389618       | 261040.599440       | 259784.187747       | 259824.632097       | 260611.483007       | 262572.080797       | 264362.978651       | 266035.283981       | 266586.858648       | 267875.250199       | 268627.017393       | 269061.847659       |
| 4 | 82081    | 55069      | Rush City           | MN    | Minneapolis-St. Paul-Bloomington | Chisago County    | 13351    | 220876.228446       | 223862.517907       | 223666.274623       | 224345.992793       | 225507.526766       | 226998.445746       | 228521.447644       | 229549.912176       | 230176.347038       | 231217.449718       | 232398.893273       | 233334.415774       |
| 5 | 99045    | 97041      | Mount Hood Parkdale | OR    | Hood River                       | Hood River County | 16248    | 477589.348369       | 480517.481019       | 482175.760356       | 484825.573570       | 487422.636823       | 489699.960108       | 492217.507002       | 494930.222586       | 497925.399060       | 499571.552676       | 501413.068684       | 505025.898556       |
| 6 | 82054    | 55032      | Harris              | MN    | Minneapolis-St. Paul-Bloomington | Chisago County    | 15468    | 281124.910376       | 285782.638594       | 285653.663524       | 286500.297804       | 287871.357950       | 289983.973696       | 291640.722191       | 293396.941744       | 294355.922160       | 296227.402535       | 297371.514815       | 298030.419328       |
| 7 | 94167    | 83442      | Rigby               | ID    | Idaho Falls                      | Jefferson County  | 7026     | 286738.623632       | 291583.112442       | 292880.659279       | 294462.300396       | 296143.176269       | 297874.920680       | 299924.480341       | 301740.035814       | 304034.004847       | 305957.179289       | 307473.471303       | 308678.650782       |
| 8 | 99035    | 97031      | Hood River          | OR    | Hood River                       | Hood River County | 6089     | 503689.527147       | 506212.603894       | 508566.346205       | 511194.249697       | 513880.963397       | 516268.432733       | 519214.594308       | 522103.506542       | 524838.168356       | 526751.249552       | 528414.617633       | 532895.092976       |
| 9 | 93397    | 80447      | Grand Lake          | CO    | Nan                              | Grand County      | 15011    | 464184.261973       | 465433.301542       | 467910.507183       | 470317.422630       | 472697.810458       | 475298.631884       | 477641.467928       | 480002.061732       | 482430.410167       | 484356.383973       | 486151.065435       | 489527.122803       |

# Models

---

| Zip Code | City                | State | Dec 2019<br>House<br>Value | Dec 2020<br>House<br>Value | Difference<br>'20 and '19 | Dec 2020 /<br>Dec 2019 |
|----------|---------------------|-------|----------------------------|----------------------------|---------------------------|------------------------|
| 98651    | Underwood           | WA    | 531,379                    | 586,556                    | 55,177                    | 10.4%                  |
| 97031    | Hood River          | OR    | 485,881                    | 532,895                    | 47,014                    | 9.7%                   |
| 97041    | Mount Hood Parkdale | OR    | 461,224                    | 505,026                    | 43,802                    | 9.5%                   |
| 80482    | Winter Park         | CO    | 676,089                    | 738,575                    | 62,486                    | 9.2%                   |
| 83442    | Rigby               | ID    | 285,927                    | 308,679                    | 22,752                    | 8.0%                   |
| 80447    | Grand Lake          | CO    | 453,762                    | 489,527                    | 35,765                    | 7.9%                   |
| 55069    | Rush City           | MN    | 218,596                    | 233,334                    | 14,738                    | 6.7%                   |
| 55032    | Harris              | MN    | 280,712                    | 298,030                    | 17,318                    | 6.2%                   |
| 20144    | Delaplane           | VA    | 676,899                    | 714,669                    | 37,770                    | 5.6%                   |
| 55074    | Shafer              | MN    | 255,025                    | 269,062                    | 14,037                    | 5.5%                   |

Growth rates from Dec 2019 to Dec 2020 highlighting the initially chosen Top 3 zip codes.

# Conclusion

| Zip Code | City                | State | Growth | STD    | MEAN    | MAPE        | Dec 2019       |                | Dec 2020 |        | Difference<br>'20 and '19 | Dec 2020 /<br>Dec 2019 |
|----------|---------------------|-------|--------|--------|---------|-------------|----------------|----------------|----------|--------|---------------------------|------------------------|
|          |                     |       |        |        |         |             | House<br>Value | House<br>Value | Bet Dec  |        |                           |                        |
| 98651    | Underwood           | WA    | 274%   | 86,549 | 372,801 | 2% - 4%     | 531,379        | 586,556        | 55,177   | 55,177 | 10.4%                     |                        |
| 80482    | Winter Park         | CO    | 249%   | 98,718 | 470,620 | 2% - 6%     | 676,089        | 738,575        | 62,486   | 62,486 | 9.2%                      |                        |
| 20144    | Delaplane           | VA    | 247%   | 68,402 | 558,771 | 2% - 5%     | 676,899        | 714,669        | 37,770   | 37,770 | 5.6%                      |                        |
| 55074    | Shafer              | MN    | 235%   | 31,500 | 193,209 | 1.5% - 3.5% | 255,025        | 269,062        | 14,037   | 14,037 | 5.5%                      |                        |
| 55069    | Rush City           | MN    | 231%   | 28,100 | 163,007 | 1.5% - 3.5% | 218,596        | 233,334        | 14,738   | 14,738 | 6.7%                      |                        |
| 97041    | Mount Hood Parkdale | OR    | 223%   | 68,929 | 338,221 | 2% - 5%     | 461,224        | 505,026        | 43,802   | 43,802 | 9.5%                      |                        |
| 55032    | Harris              | MN    | 221%   | 40,094 | 201,372 | 1.5% - 3.5% | 280,712        | 298,030        | 17,318   | 17,318 | 6.2%                      |                        |
| 83442    | Rigby               | ID    | 218%   | 40,768 | 197,068 | 3% - 5%     | 285,927        | 308,679        | 22,752   | 22,752 | 8.0%                      |                        |
| 97031    | Hood River          | OR    | 218%   | 73,847 | 353,978 | 2% - 4%     | 485,881        | 532,895        | 47,014   | 47,014 | 9.7%                      |                        |
| 80447    | Grand Lake          | CO    | 213%   | 58,313 | 339,919 | 2% - 4%     | 453,762        | 489,527        | 35,765   | 35,765 | 7.9%                      |                        |

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 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.

## Recommendations

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.

# Inbound Crossing at the US-Canada and US-Mexico Border



# Motivation

---

- 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 using Adobe Illustrator 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.

# Business 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?

# The Collection of Data

---

- The 35MB csv data was sourced from Kaggle (<https://www.kaggle.com/akhilv11/border-crossing-entry-data>) with 344,369 rows and 9 columns
- Read the downloaded csv file using R.

| ▲  | Port.Name   | State        | Port.Code | Border           | Date | Measure                     | Value | Lon        | Lat      |
|----|-------------|--------------|-----------|------------------|------|-----------------------------|-------|------------|----------|
| 1  | Van Buren   | Maine        | 108       | US-Canada Border | 2018 | Trucks                      | 1204  | -67.94271  | 47.16207 |
| 2  | Maida       | North Dakota | 3416      | US-Canada Border | 2018 | Trucks                      | 170   | -98.36953  | 48.98568 |
| 3  | Douglas     | Arizona      | 2601      | US-Mexico Border | 2018 | Buses                       | 210   | -109.54472 | 31.34444 |
| 4  | Presidio    | Texas        | 2403      | US-Mexico Border | 2018 | Bus Passengers              | 238   | -104.37167 | 29.56056 |
| 5  | Anacortes   | Washington   | 3010      | US-Canada Border | 2018 | Personal Vehicle Passengers | 5350  | -122.61739 | 48.49988 |
| 6  | Brownsville | Texas        | 2301      | US-Mexico Border | 2018 | Truck Containers Full       | 9871  | -97.49722  | 25.90139 |
| 7  | Maida       | North Dakota | 3416      | US-Canada Border | 2018 | Truck Containers Empty      | 162   | -98.36953  | 48.98568 |
| 8  | Columbus    | New Mexico   | 2406      | US-Mexico Border | 2018 | Pedestrians                 | 24168 | -107.63944 | 31.82750 |
| 9  | Eastport    | Maine        | 103       | US-Canada Border | 2018 | Truck Containers Empty      | 89    | -66.99387  | 44.90357 |
| 10 | Calais      | Maine        | 115       | US-Canada Border | 2018 | Trucks                      | 4504  | -67.27917  | 45.18889 |

# Pre-Processing and Data Cleaning

---

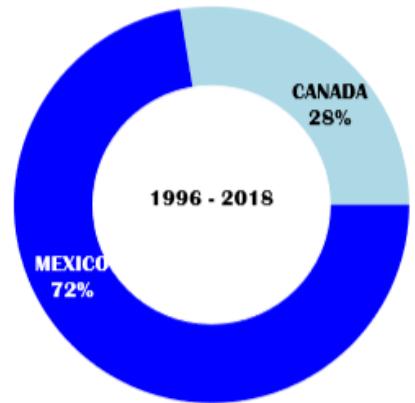
- Abbreviation state names to .
- Renamed “measures” or modes of entry.
- No NaN values were found from the dataset.

# Exploratory Plots

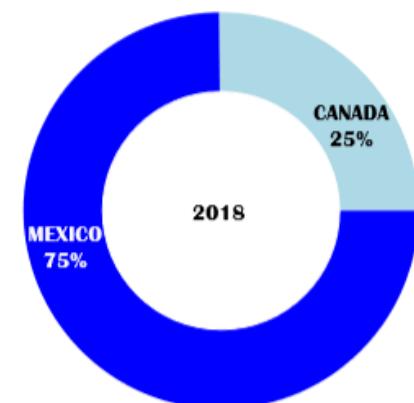
Entries in the United States from Y1996 to Y2018.

**QUESTION #2:** What is the split of border crossing coming from Mexico and Canada?

From 1996 to 2008, 72% and 28% crossed through Mexico and Canada, respectively.

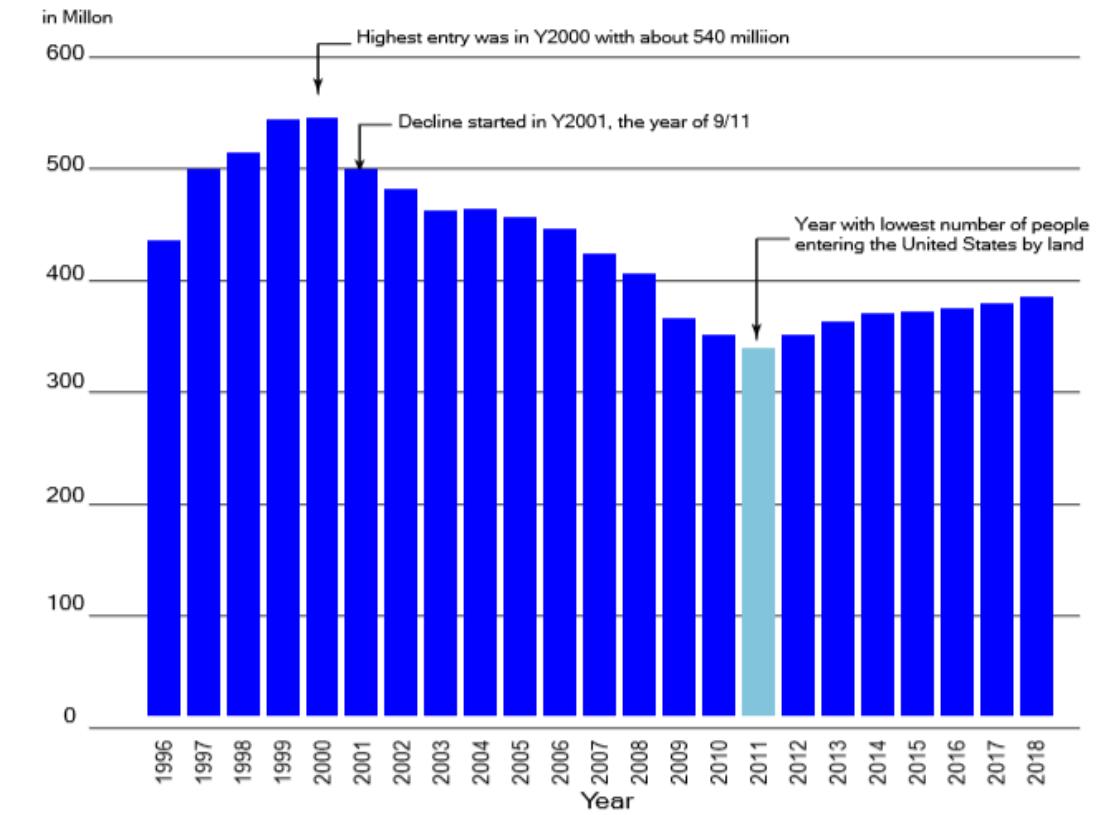


In 2018 alone, the split changed in favor of Mexico now at 75%. The % split started to change to 75% in 2017.



**QUESTION #1:** How many have entered the United States from the Canadian and Mexican ports from Y1996 to Y2018?

**Number of People Entering the US from Y1996 to Y2018**  
Canada to US and Mexico to US



# Exploratory Plots

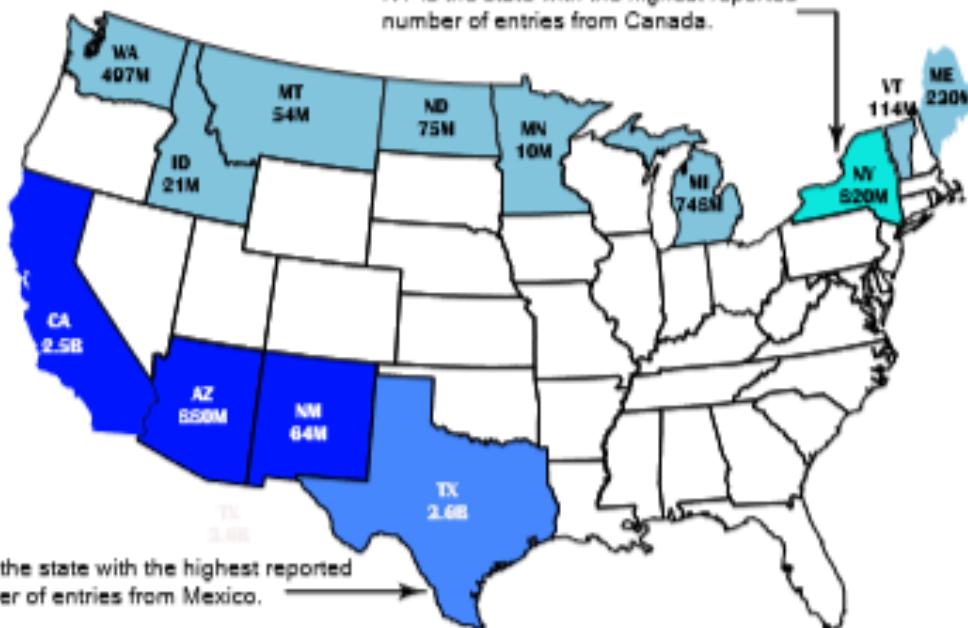
US States where people cross into.

## QUESTION #3: Which states do people cross from?

### Inbound Crossing from Y1996 to Y2018

Canada to US and Mexico to US

Number of Entries for the Past 12 Year



Top 5 States where poeple enter from:

1. Texas (US - Mexico border)
2. California (US - Mexico border)
3. Arizona (US - Mexico border)
4. New York (US - Canada border)
5. Michigan (US - Canada border)

Source: Kaggle-Bureau of Transportation Statistics

# Exploratory Plots

Top 10 port  
of entries.

## QUESTION #4: What are the top 10 port of entries?

### Port of Entry Definition:

It is where one may lawfully enter a country. It has border security staff and facilities to check passports and visas. US has 328 ports of entry.

Top 10 ports of entry haven't changed when comparing combined data from years 1996 to 2018 to year 2018 alone.

El Paso in Texas ranked the highest in terms of entries from years 1996 to 2018.

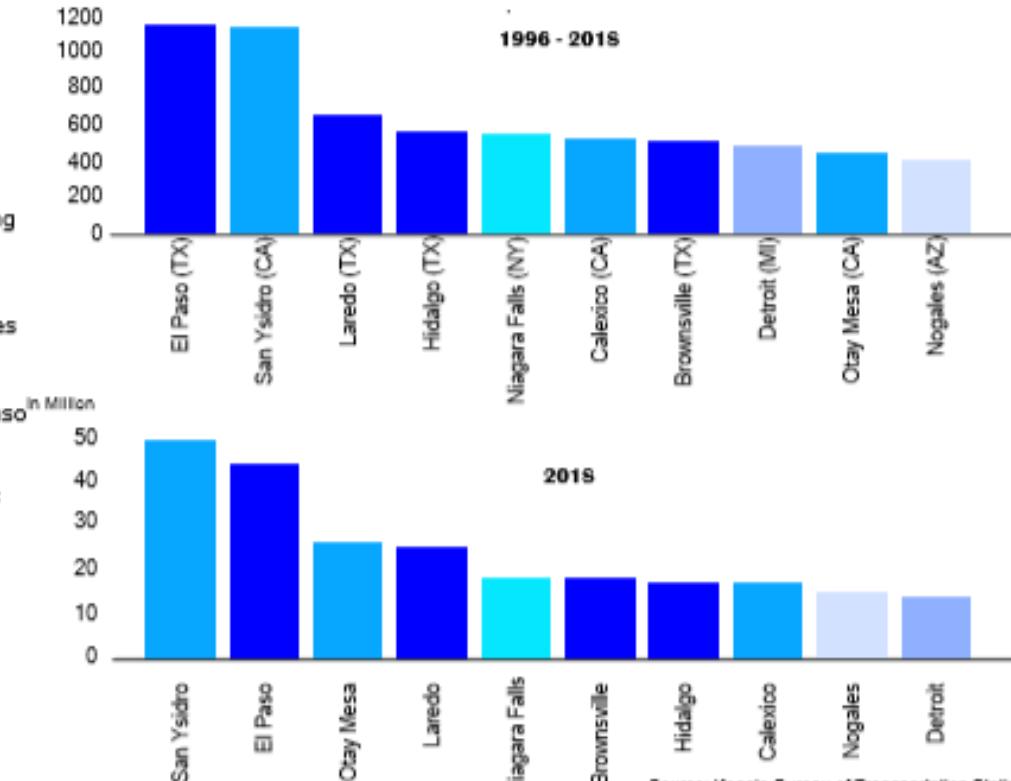
El Paso in Texas is a city and the county seat of El Paso County.

San Ysidro in California ranked the highest number of entries for year 2018 alone.

San Ysidro in California is a district of the City of San Diego.

### Top 10 Port of Entries

In Million



Source: Kaggle-Bureau of Transportation Statistics

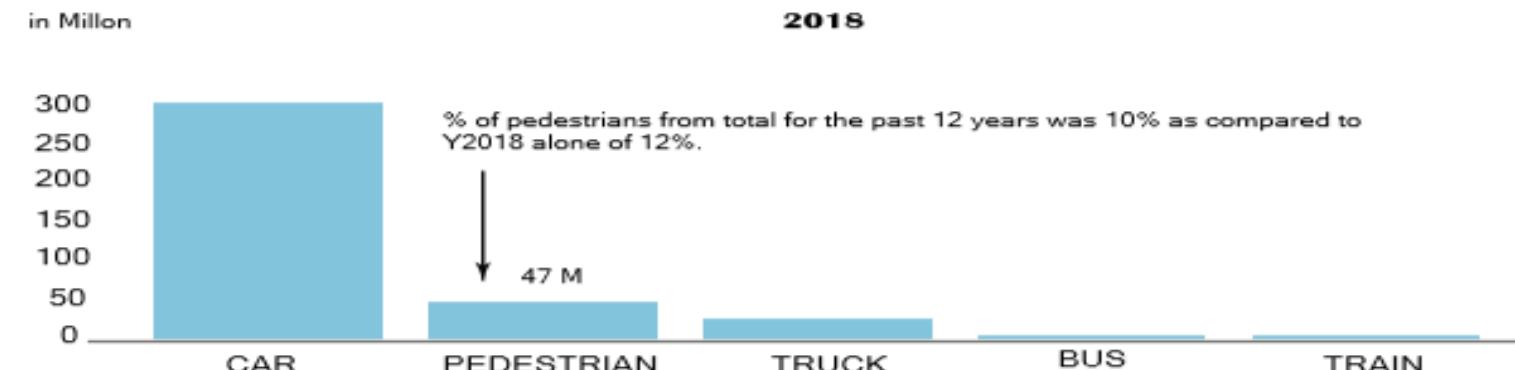
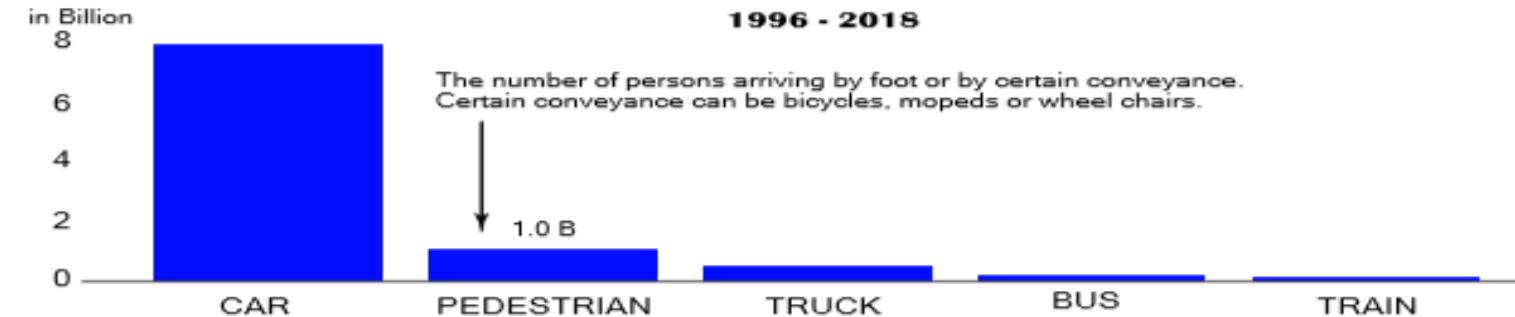
# Exploratory Plots

## Mode of entries

**QUESTION #5:** What are the mode of entries?

### Mode of Entries

Canada to US and Mexico to US



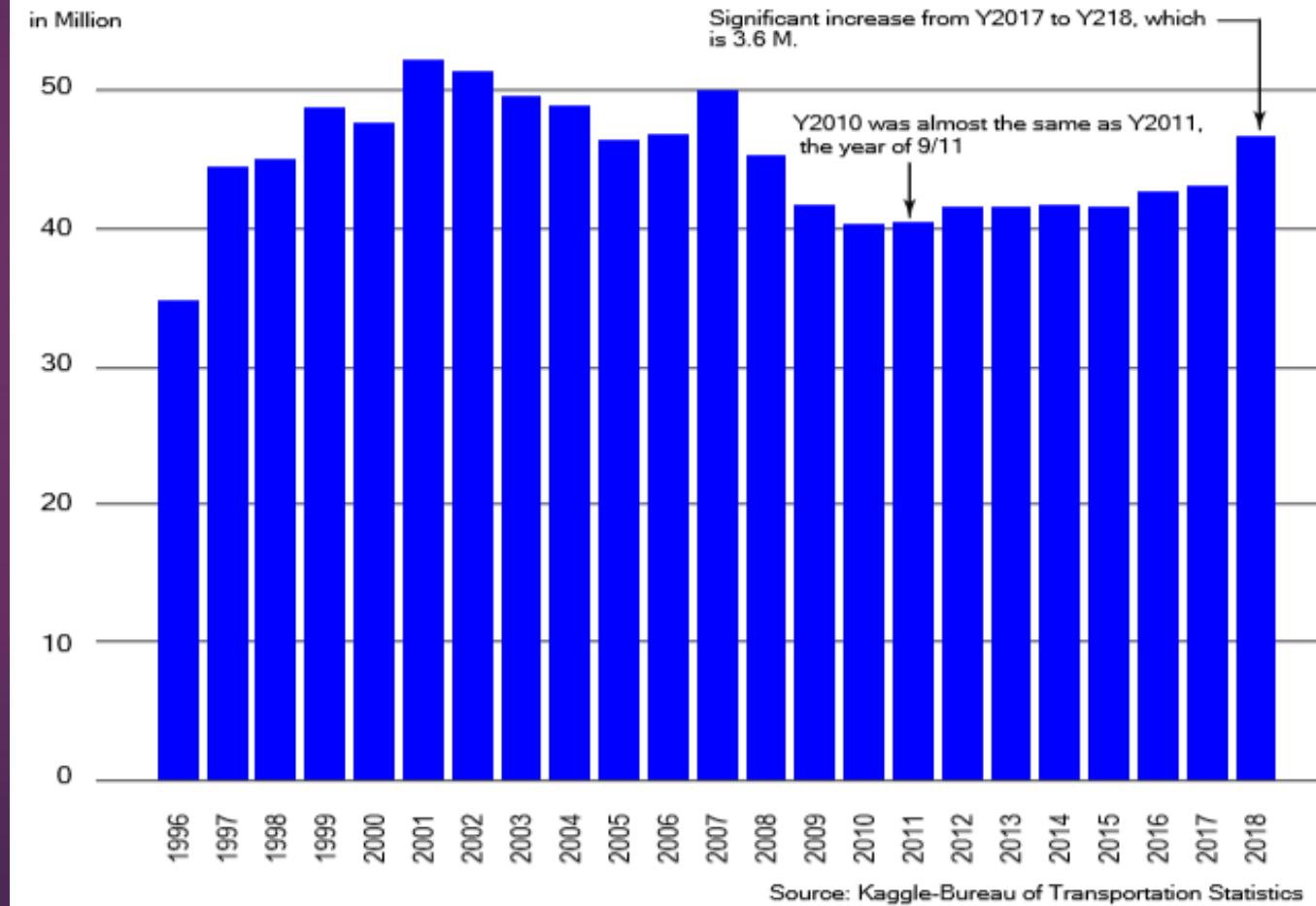
Source: Kaggle-Bureau of Transportation Statistics

# Exploratory Plots

Number of pedestrians crossing the border.

**QUESTION #6:** How many pedestrians have crossed the border? What is the trend?

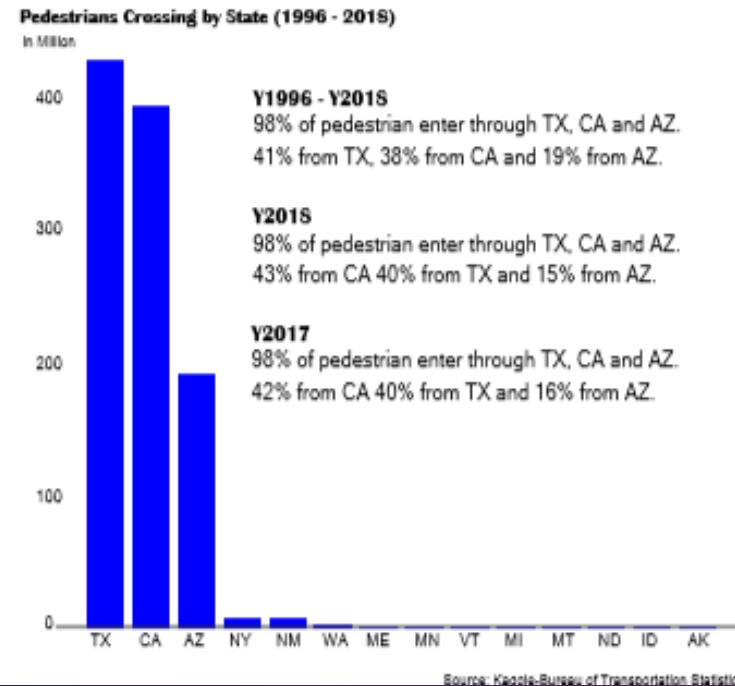
**Pedestrians Crossing the Border Trend**



# Exploratory Plots

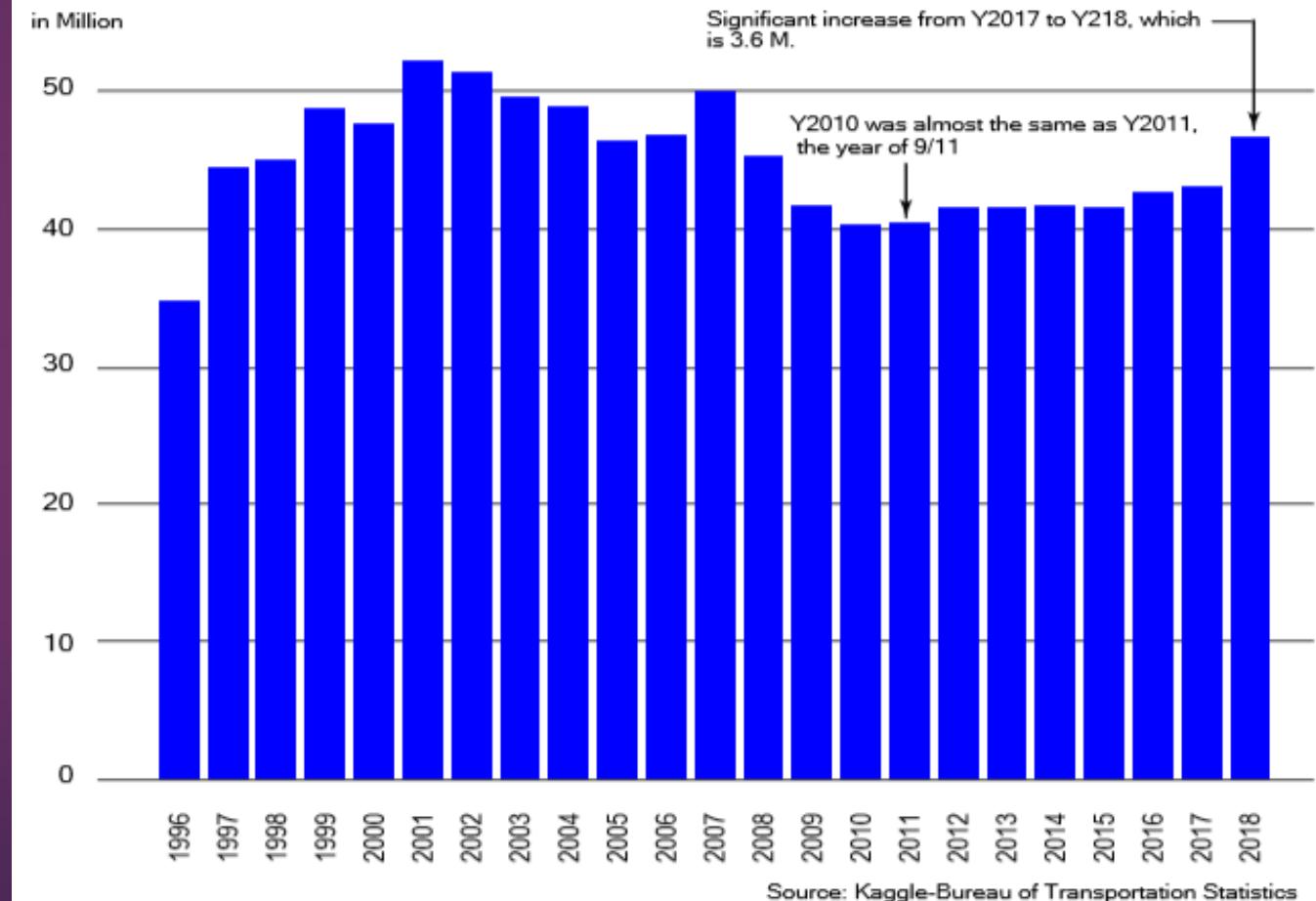
Number of pedestrians crossing the border.

**QUESTION #7:** Which states are pedestrians crossing from?



**QUESTION #6:** How many pedestrians have crossed the border? What is the trend?

## Pedestrians Crossing the Border Trend



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

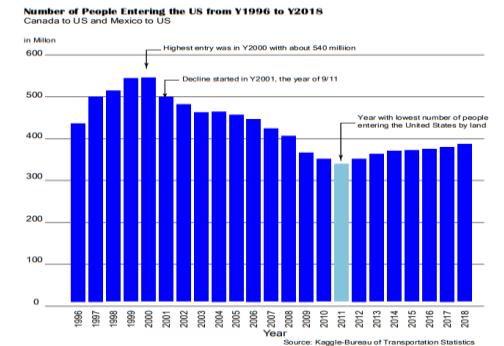
Y1996 to Y2018

Student: Allan Flores  
Course: IST719  
Prof: Gary Krudys

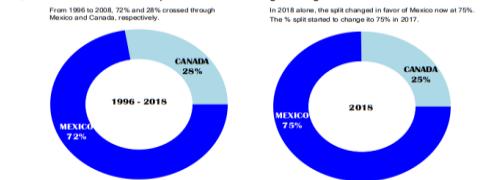
**INTRODUCTION:** Several arguments and exchange of thoughts and ideas have circled around immigration and more so in recent years. Data depicted from this survey is an outlook of people entering the United States from Mexico and Canada by land.

**DATA DESCRIPTION:** Data has around 350K observations and 9 variables sourced from Kaggle. Data covered was from year 1996 to 2018.

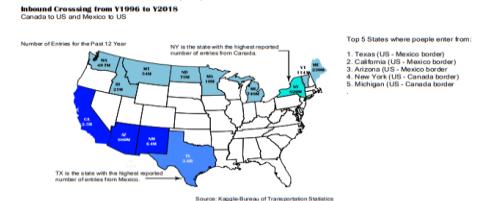
**QUESTION #1:** How many have entered the United States from the Canadian and Mexican ports from Y1996 to Y2018?



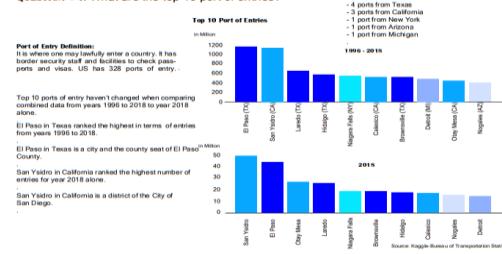
**QUESTION #2:** What is the split of border crossing coming from Mexico and Canada?



**QUESTION #3:** Which states do people cross from?

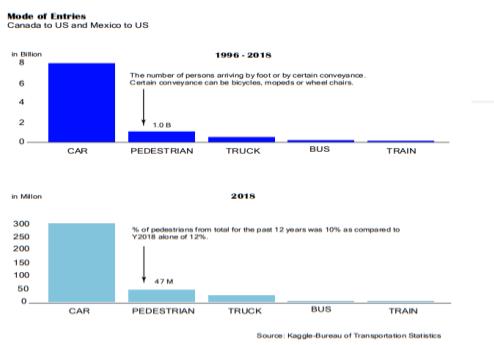


**QUESTION #4:** What are the top 10 port of entries?

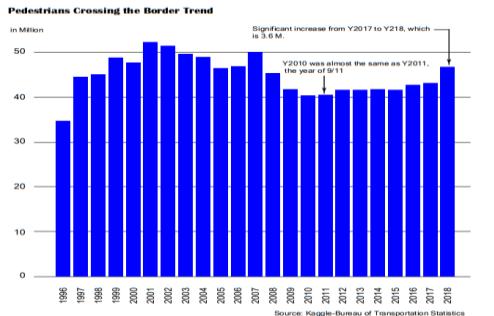


**MOTIVATION:** Understand through data how people enter the United States by land coming from Mexican and Canadian borders. Has there been any decline in the number of entries over time? Moreover, answer some questions as regards to immigration given the questions going around.

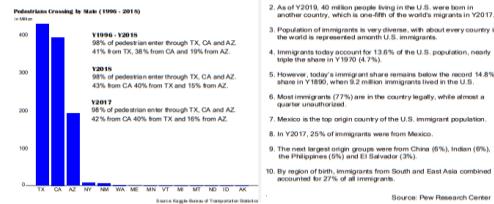
**QUESTION #5:** What are the mode of entries?



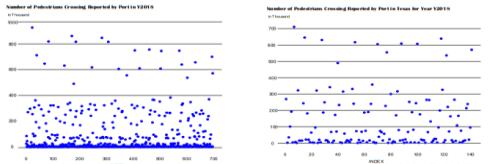
**QUESTION #6:** How many pedestrians have crossed the border? What is the trend?



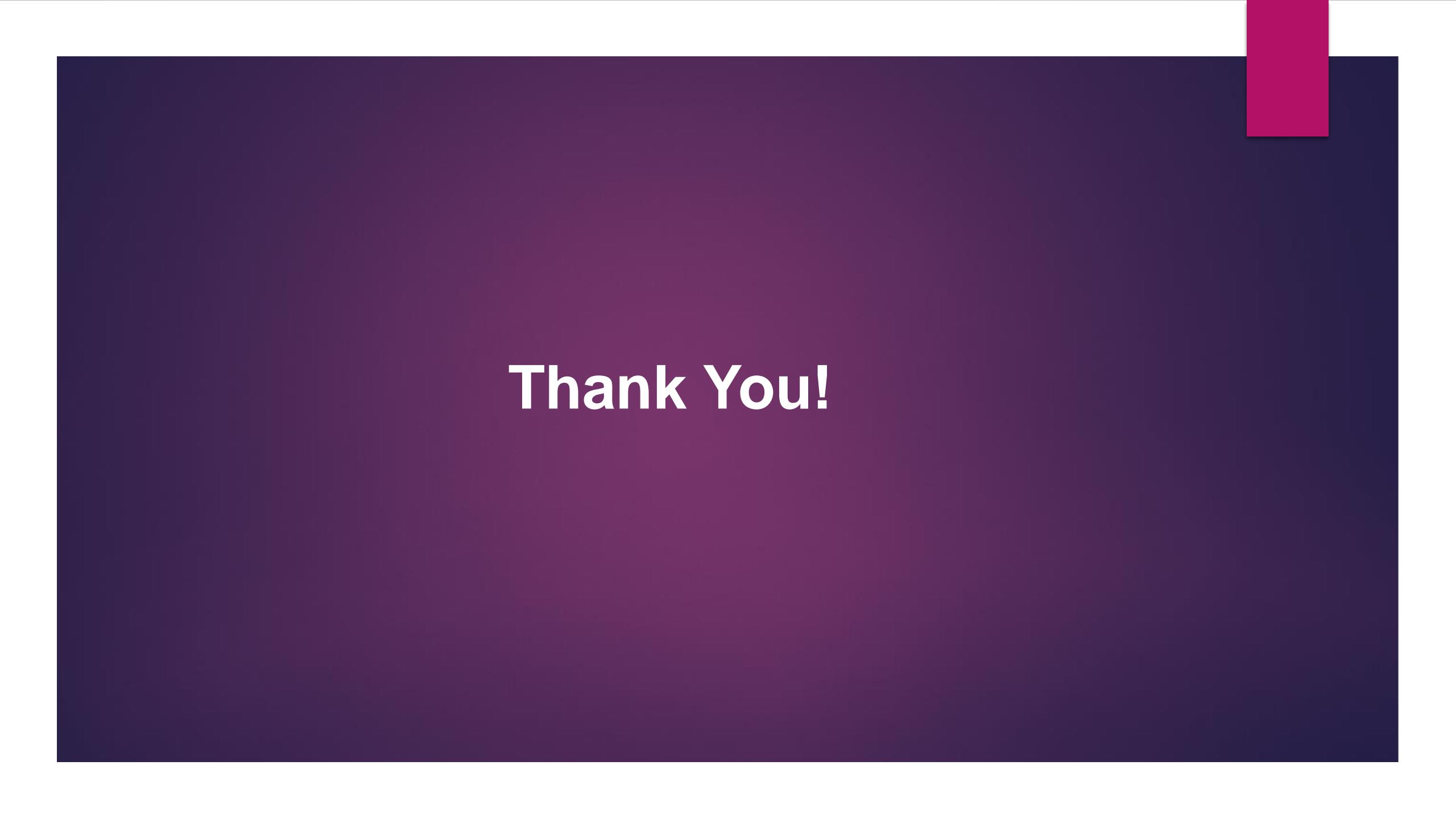
**QUESTION #7:** Which states are pedestrians crossing from?



**QUESTION #8:** What are number of pedestrians crossing by port?



# Final Poster



**Thank You!**