HOUSING PRICE DATA ANALYSIS

MSBA 305 SF1: Business Intelligence

Instructor: Mayank Johri

By

Veena Chintala

Bhavin Bhanushali

Navjot Kaur

Anuradha Patel

Overview

This report gives the detailed view of learnings in the MSBA 305 (tools for business intelligence). The report presents the descriptive, prescriptive and predictive analysis of the "housing price data. The data is collected from Kaggle website. The report represents the full process of our project from collection, cleaning and then analyze with four steps description, prediction and prescription of the data set.

Introduction

The data set for housing price was taken from Kaggle website contains 1598 observation of nine different variables. The variables are year, index_nsa, Housing_Price, City..State, Population, Violent Crimes, homicides, Rapes and Robberies. The housing price is the dependent variable in our project. We are also intrigued that which variable from the various crime variables impact the price of the houses between the year 1979-2015.

We started the analysis by downloading the housing data set in the R environment.

```
data <- read.csv("~/Desktop/housedata.csv")
```

Descriptive Analysis

In descriptive analysis, we will do the descriptive statistics to get the general overview of the dataset. It gives the quantitative information about the various samples of the housing and crimes in the given data. We started by looking at the structure of our dataset.

The structure of our dataset is as below:

```
str(data)
```

```
## 'data.frame':    1598 obs. of  9 variables:
##  $ Year          : int  1979 1979 1979 1979 1979 1979 1979 1979 1979 1979 ...
##  $ index_nsa     : num  49.4 50.8 55.4 46.1 32.6 ...
##  $ Housing_Price : num  4941500 5084000 5544750 4607000 3259000 ...
##  $ City..State   : Factor w/ 44 levels "Albuquerque, NM",..: 1 3 5 6 7 8 9 10 11 12 ...
##  $ Population    : int  302120 423103 337727 790901 599582 383915 300569 3060801 404661 601381 ...
##  $ Violent.Crimes: int  2679 10715 1495 15523 11392 3857 2485 27807 3808 9736 ...
##  $ Homicides     : int  47 231 43 245 92 52 50 856 50 274 ...
##  $ Rapes         : int  215 656 216 564 464 265 115 1655 282 612 ...
##  $ Robberies     : int  815 5189 577 8482 6600 1958 703 14464 1662 5760 ...
```

This gives us the overview of our data by giving the number of observations in the dataset and different variables in the housing data. The variables are year, index_nsa, housing prices and various crimes.

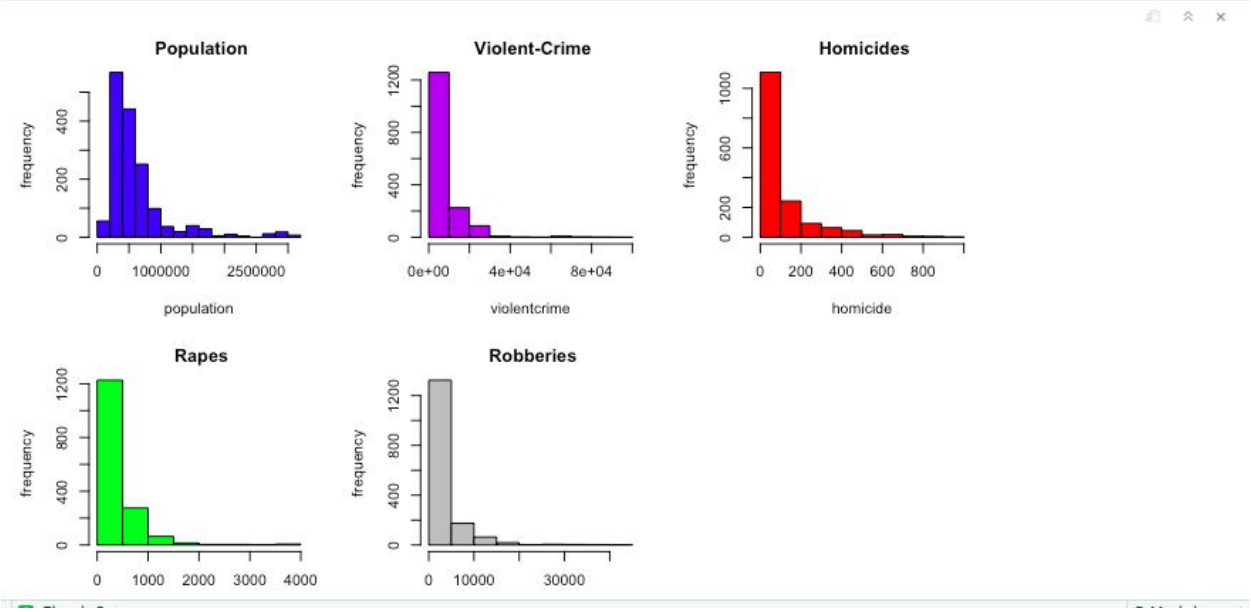
In the next step, the summary of the data is performed to get the idea of the various values of mean, median and mode of all the variables along with minimum and maximum values. In the summary table, it is clear that mean values are greater than median of all the variables in the dataset. This implies that there are fewer larger values in the all of the variables.

```
summary(data)
```

```
##       Year        index_nsa      Housing_Price              City..State
##  Min.   :1979   Min.   : 31.02   Min.   : 3101500   Albuquerque, NM:  37
##  1st Qu.:1988   1st Qu.: 84.51   1st Qu.: 8450938   Arlington, TX  :  37
##  Median :1997   Median :117.73   Median :11772604   Atlanta, GA    :  37
##  Mean   :1997   Mean   :132.05   Mean   :13204671   Aurora, CO     :  37
##  3rd Qu.:2006   3rd Qu.:171.96   3rd Qu.:17196281   Austin, TX     :  37
##  Max.   :2015   Max.   :386.21   Max.   :38621188   Baltimore, MD  :  37
##                                                     (Other)        :1376
##    Population      Violent.Crimes    Homicides         Rapes
##  Min.   : 115498   Min.   :  585   Min.   :  1.0   Min.   :  36.0
##  1st Qu.: 361161   1st Qu.: 3296   1st Qu.: 35.0   1st Qu.: 195.0
##  Median : 471454   Median : 5164   Median : 62.0   Median : 296.0
##  Mean   : 621693   Mean   : 7884   Mean   :113.7   Mean   : 413.1
##  3rd Qu.: 656422   3rd Qu.: 9145   3rd Qu.:122.0   3rd Qu.: 483.8
##  Max.   :3060801   Max.   :90520   Max.   :960.0   Max.   :3754.0
##
##    Robberies
##  Min.   :  127
##  1st Qu.: 1104
##  Median : 2111
##  Mean   : 3403
##  3rd Qu.: 3780
##  Max.   :43783
##
```

The larger mean values in the summary table gives the right-skewness in the histogram of population, homicides, rapes and robberies.

Histogram:



Regression Analysis:

We have more than one independent variables hence we perform multiple linear regression. Firstly, we ran the regression on independent variables including Population, Violent.Crimes, Homicides, Rapes and Robberies against the dependent variable Housing Prices. From the below screenshot we can see that p-value for all the independent variables is less than 0.05 except for Homicides. We remove the variable and rerun the sample.

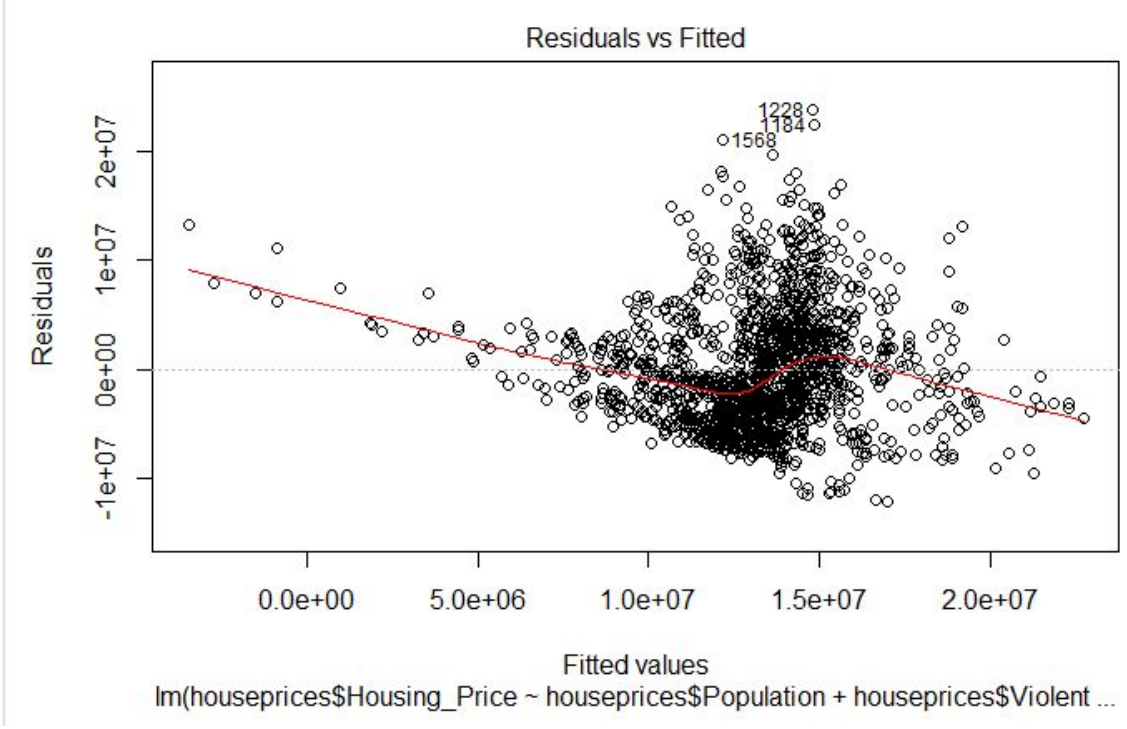
```
Call:
lm(formula = houseprices$Housing_Price ~ houseprices$Population +
    houseprices$Violent.Crimes + houseprices$Homicides + houseprices$Rapes +
    houseprices$Robberies)

Residuals:
      Min        1Q    Median        3Q       Max
-11834051  -4034790   -850761   3161042  23784768

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                1.209e+07  2.317e+05  52.203   <2e-16 ***
houseprices$Population     7.278e+00  5.143e-01  14.151   <2e-16 ***
houseprices$Violent.Crimes 6.845e+02  7.214e+01   9.487   <2e-16 ***
houseprices$Homicides      -4.024e+03  2.503e+03  -1.608    0.108
houseprices$Rapes          -8.408e+03  8.536e+02  -9.850   <2e-16 ***
houseprices$Robberies      -1.434e+03  1.642e+02  -8.730   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5461000 on 1592 degrees of freedom
Multiple R-squared:  0.1895,    Adjusted R-squared:  0.187
F-statistic: 74.45 on 5 and 1592 DF,  p-value: < 2.2e-16
```

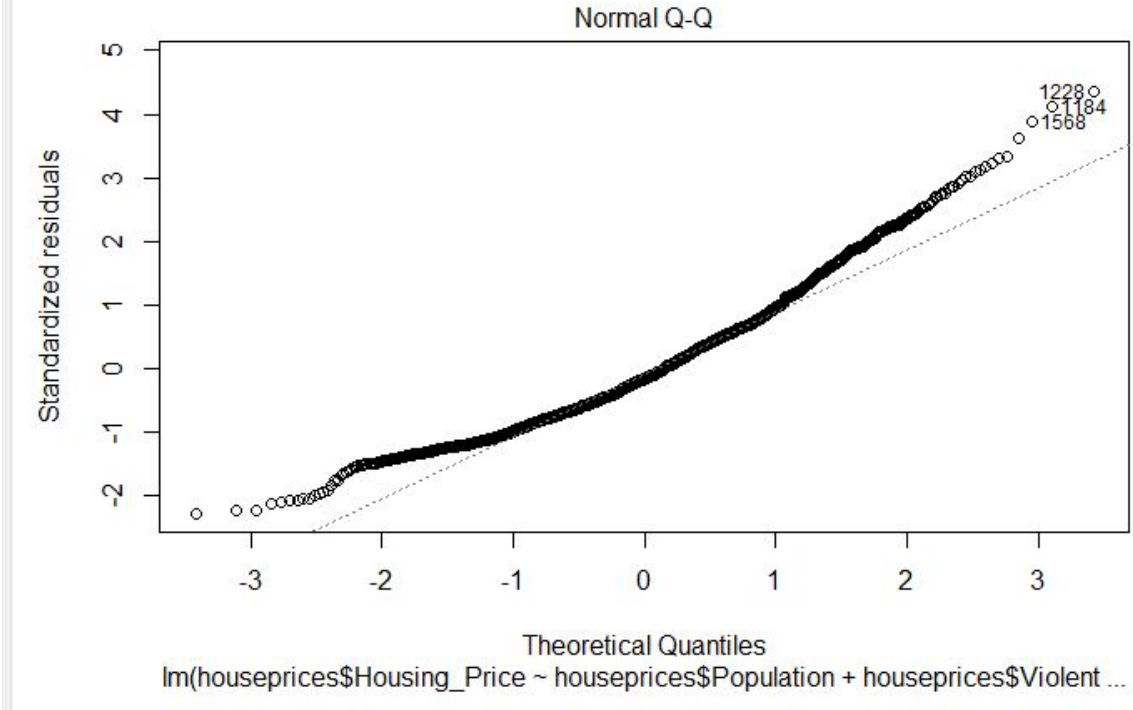
The Second regression model below has all the p-values less than 0.05 hence we consider this model and the equation is 0.0000001.21+ 7.0707*Population + 0.06882*Violent Crimes - 0.008359*Rapes - 0.00155*Robberies

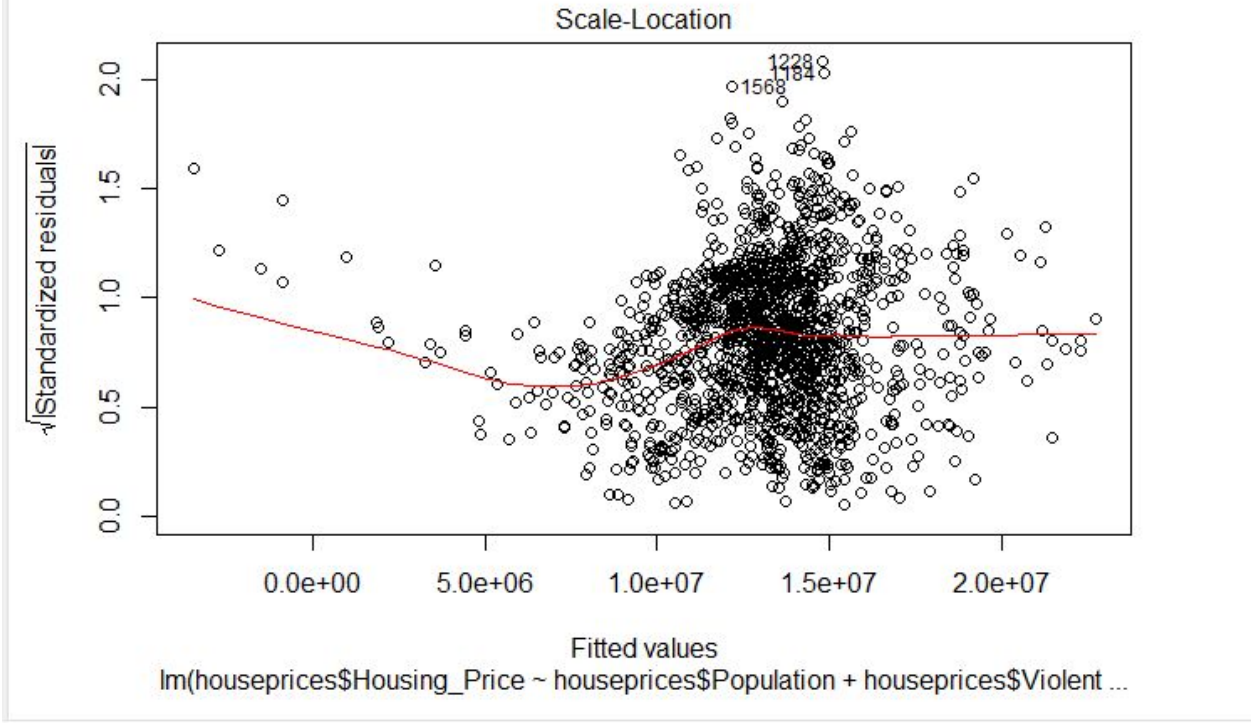
```
Call:
lm(formula = houseprices$Housing_Price ~ houseprices$Population +
    houseprices$Violent.Crimes + houseprices$Rapes + houseprices$Robberies)

Residuals:
      Min        1Q    Median        3Q       Max
-12174921  -4023619   -839569   3154800  23807760

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                1.211e+07  2.315e+05  52.304   <2e-16 ***
houseprices$Population     7.070e+00  4.981e-01  14.196   <2e-16 ***
houseprices$Violent.Crimes 6.882e+02  7.214e+01   9.540   <2e-16 ***
houseprices$Rapes         -8.359e+03  8.535e+02  -9.795   <2e-16 ***
houseprices$Robberies     -1.550e+03  1.476e+02 -10.498   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5463000 on 1593 degrees of freedom
Multiple R-squared:  0.1882,    Adjusted R-squared:  0.1862
F-statistic: 92.33 on 4 and 1593 DF,  p-value: < 2.2e-16
```

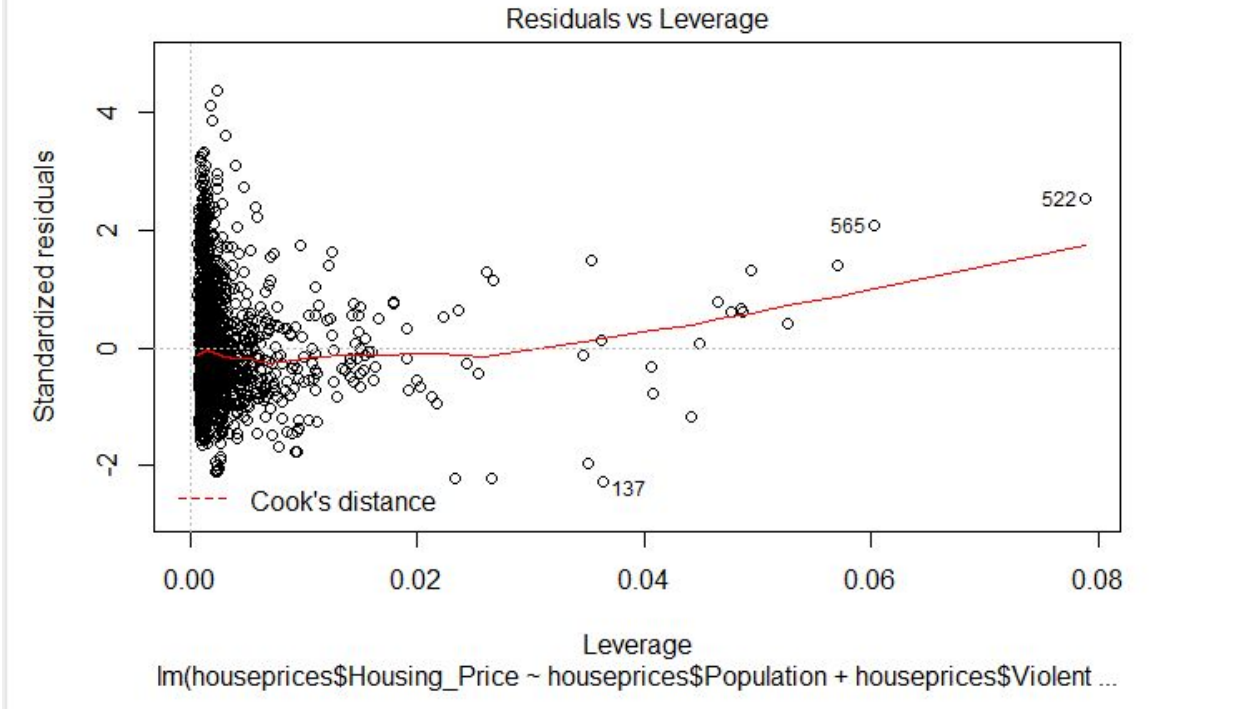
Regression Diagnostic

We can use the *plot* function to verify the assumptions.


Residuals vs Fitted
Residuals
Fitted values
lm(houseprices$Housing_Price ~ houseprices$Population + houseprices$Violent ...
1228
1184
1568
-1e+07
0e+00
1e+07
2e+07
0.0e+00
5.0e+06
1.0e+07
1.5e+07
2.0e+07


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(houseprices$Housing_Price ~ houseprices$Population + houseprices$Violent ...
1228
1184
1568
-2
-1
0
1
2
3
4
5
-3
-2
-1
0
1
2
3


Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
1228
1184
1568
0.0e+00
5.0e+06
1.0e+07
1.5e+07
2.0e+07
Fitted values
lm(houseprices$Housing_Price ~ houseprices$Population + houseprices$Violent ...


Residuals vs Leverage
Standardized residuals
-2
0
2
4
522
565
137
Cook's distance
0.00
0.02
0.04
0.06
0.08
Leverage
lm(houseprices$Housing_Price ~ houseprices$Population + houseprices$Violent ...

from the above graph we can see that the housing price has non-linear relation with the independent variables.

OLS regression assumptions are:

Normality: The Normal Q-Q plot shows that most of the points are on the straight 45- degree line.

Independence: Here we are assuming that data collection of dependent variables- Runs scored are independent.

Linearity: Residual vs. fitted graphs does not show any pattern. The is no symmetric relationship between residual and the predicted value.

Homoscedasticity: The assumption of constant variance is met as residual as numbers are randomly located around a horizontal line.

Correlation:

Correlation coefficient was used to check the relation between the independent variables and dependent variable.

From the below table we can see that population has a weak positive correlation with the housing prices and other independent variables including violent crimes, rapes and robberies

have weak negative correlation with the housing prices.

	houseprices.Housing_Price	houseprices.Population	houseprices.Violent.Crimes	houseprices.Rapes	houseprices.Robberies
houseprices.Housing_Price	1.00000000	0.06873349	-0.08678161	-0.1541507	-0.1387294
houseprices.Population	0.06873349	1.00000000	0.81157151	0.8076039	0.8077311
houseprices.Violent.Crimes	-0.08678161	0.81157151	1.00000000	0.9115993	0.9744025
houseprices.Rapes	-0.15415068	0.80760386	0.91159935	1.0000000	0.9050507
houseprices.Robberies	-0.13872936	0.80773115	0.97440250	0.9050507	1.0000000

ANOVA Test

Anova is used to test to check the effect of each independent variable on dependent variable .

```
                            Df    Sum Sq   Mean Sq F value   Pr(>F)
houseprices$Violent.Crimes   1 4.411e+14 4.411e+14   12.11 0.000515 ***
Residuals                 1596 5.813e+16 3.642e+13
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Above chart shows the results of testing the effect of violent crimes on housing prices.

From the above anova result we see that the p-value is less than 0.05, hence we reject the null hypothesis and have evidence to believe that Violent crimes have significant effect on housing prices.

Anova to check the effect of Population on housing prices.

```
                          Df    Sum Sq   Mean Sq F value  Pr(>F)
houseprices$Population     1 2.767e+14 2.767e+14   7.576 0.00598 **
Residuals               1596 5.830e+16 3.653e+13
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

From the above anova result we can see that p-value is less than 0.05, hence we reject the null hypothesis and have evidence to believe that population has significant effect on housing prices.

Anova to check the effect of Rapes on housing prices.

```
                    Df    Sum Sq   Mean Sq F value   Pr(>F)
houseprices$Rapes    1 1.392e+15 1.392e+15   38.85 5.85e-10 ***
Residuals         1596 5.718e+16 3.583e+13
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

From the above anova result we can see that p-value is less than 0.05, hence we reject the null hypothesis and have evidence to believe that rapes has significant effect on housing prices.

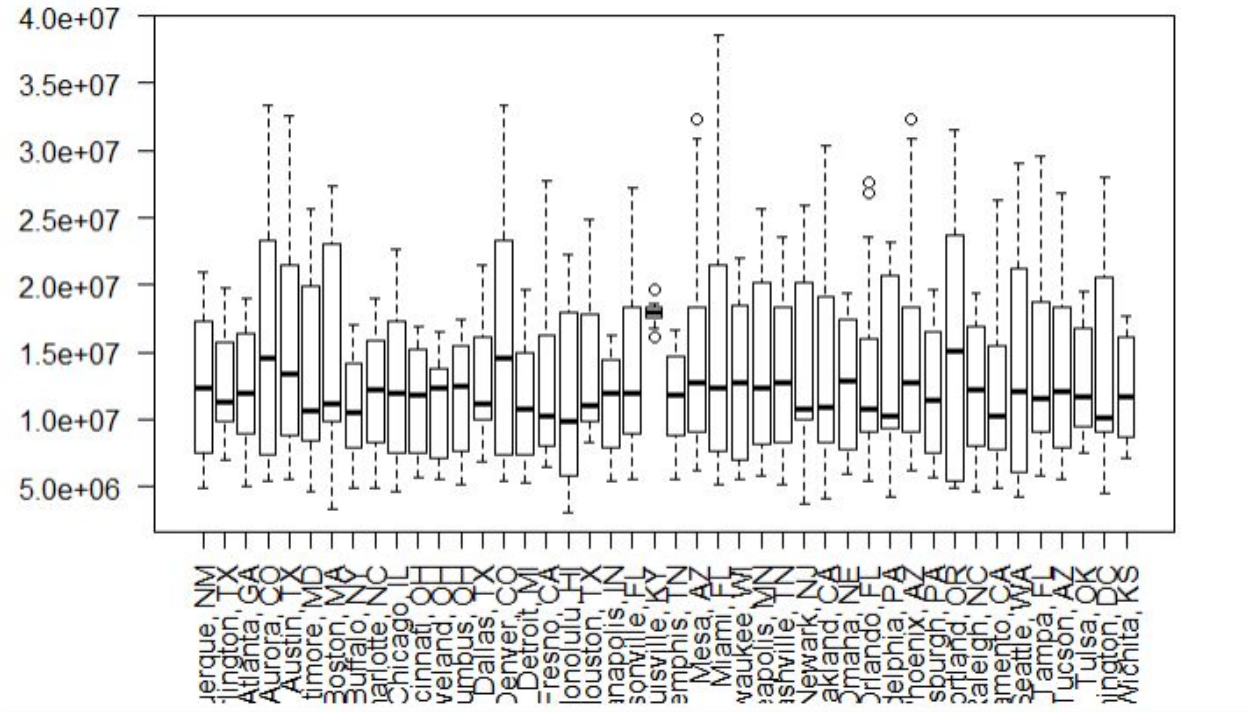
Anova to check the effect of Robberies on housing prices.

```
                      Df    Sum Sq   Mean Sq F value   Pr(>F)
houseprices$Robberies    1 1.127e+15 1.127e+15   31.32 2.57e-08 ***
Residuals             1596 5.745e+16 3.599e+13
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

From the above anova result we can see that p-value is less than 0.05, hence we reject the null hypothesis and have evidence to believe that Robberies have significant effect on housing prices.

Anova on Cities

Firstly, we check the distribution of different cities and housing prices of those cities.



Using Anova we test if the housing prices in different cities is same or different.

```
                          Df    Sum Sq   Mean Sq F value   Pr(>F)
houseprices$City..State   43 3.438e+15 7.996e+13   2.254 7.98e-06 ***
Residuals               1554 5.514e+16 3.548e+13
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

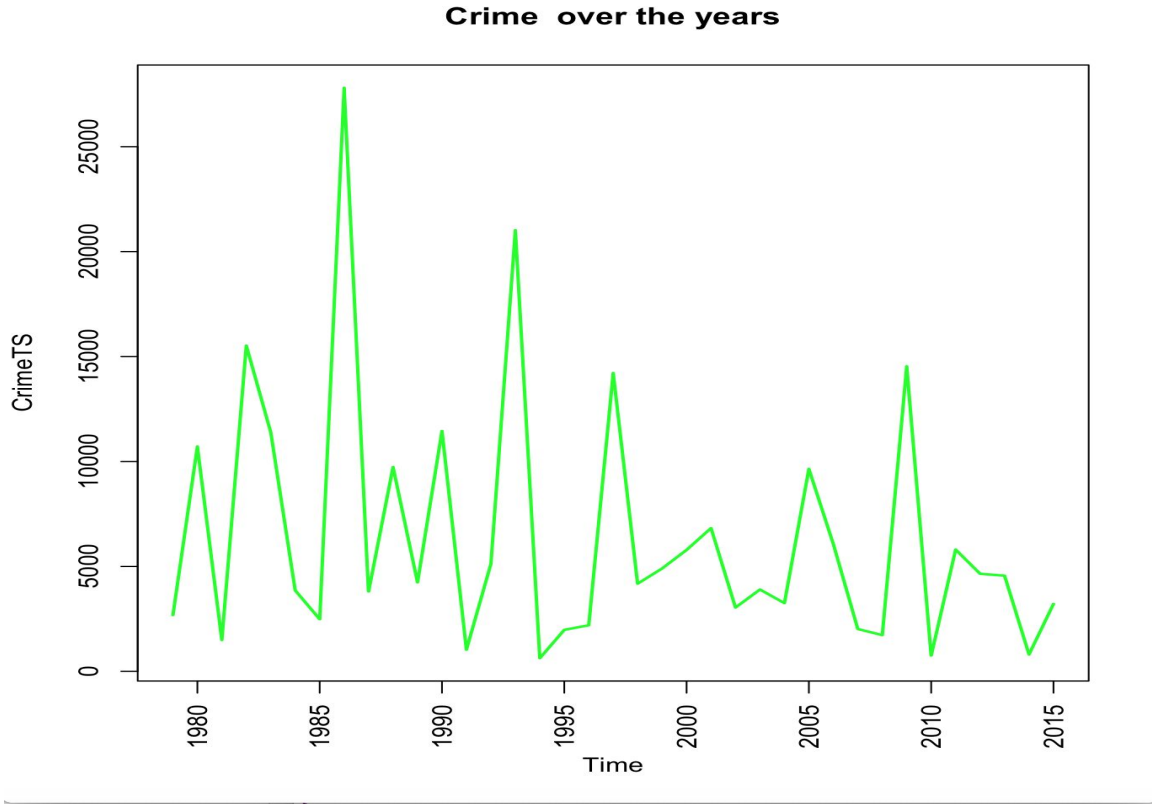
From the above anova result we can see that p-value is less than 0.05 hence we reject the null hypothesis and have evidence to believe that housing prices for different cities is significantly different.

Time Series Analysis:

Time series analysis gives the trend of housing price over these years. In data analysis, it is important to take time factor into account for correct predictions. It is important because there are so many prediction problems that involve a time component.

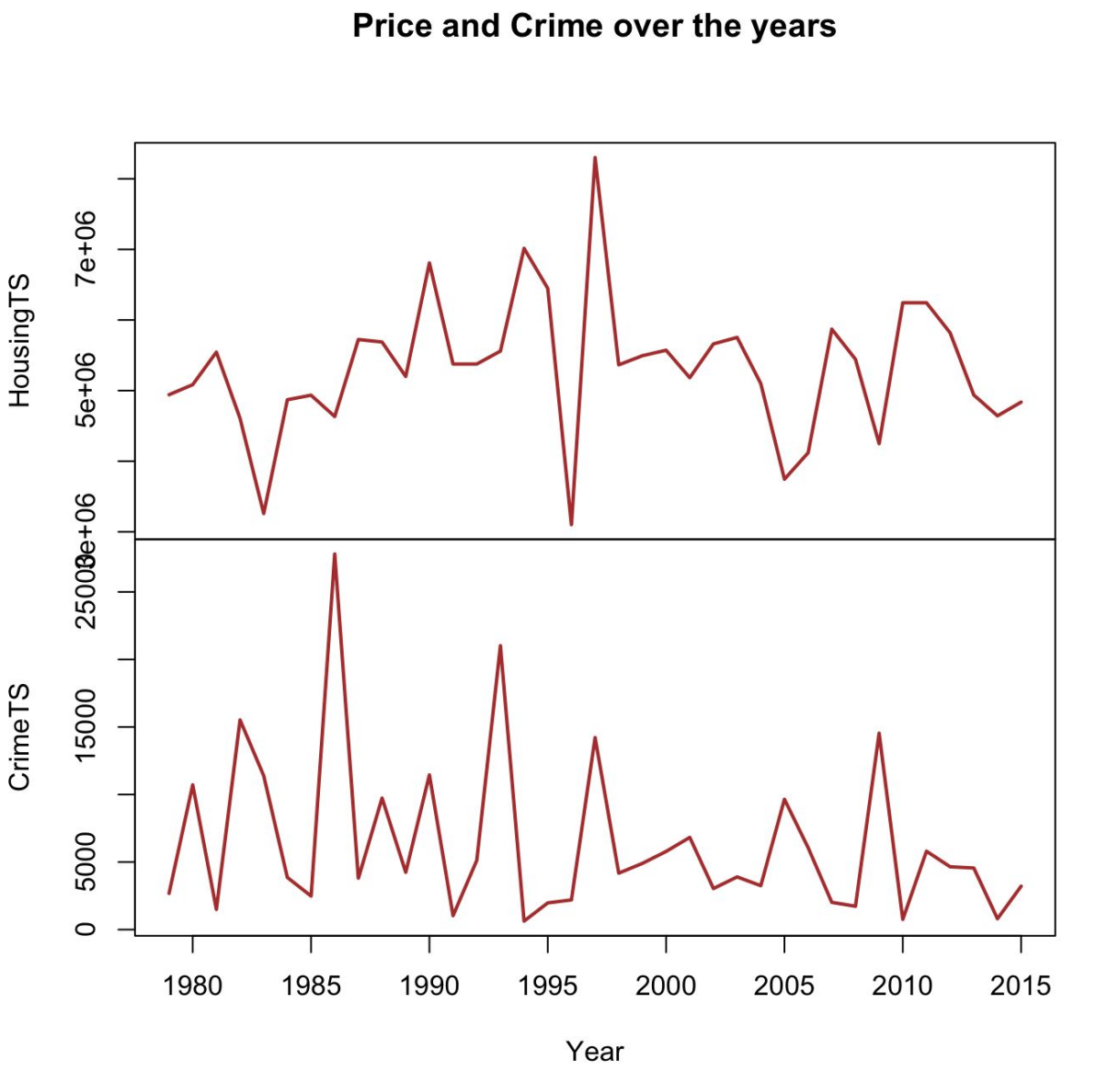
For the team series analysis, we have analyzed the house price from 1979-2015 where Housing Price is our dependant variable and Year is taken as independent variable, and this time series shows how Housing Price have been in these years and how it will be in future.

From the below figure, we see that the Housing Price was increasing and decreasing over the years, and it was highest in the year 1985.



Relation of Crime and Price with Year:

From the below graph, we can see that Crime affects the Housing Price in a region over the years. When the crime is increased, the housing price gets increased.



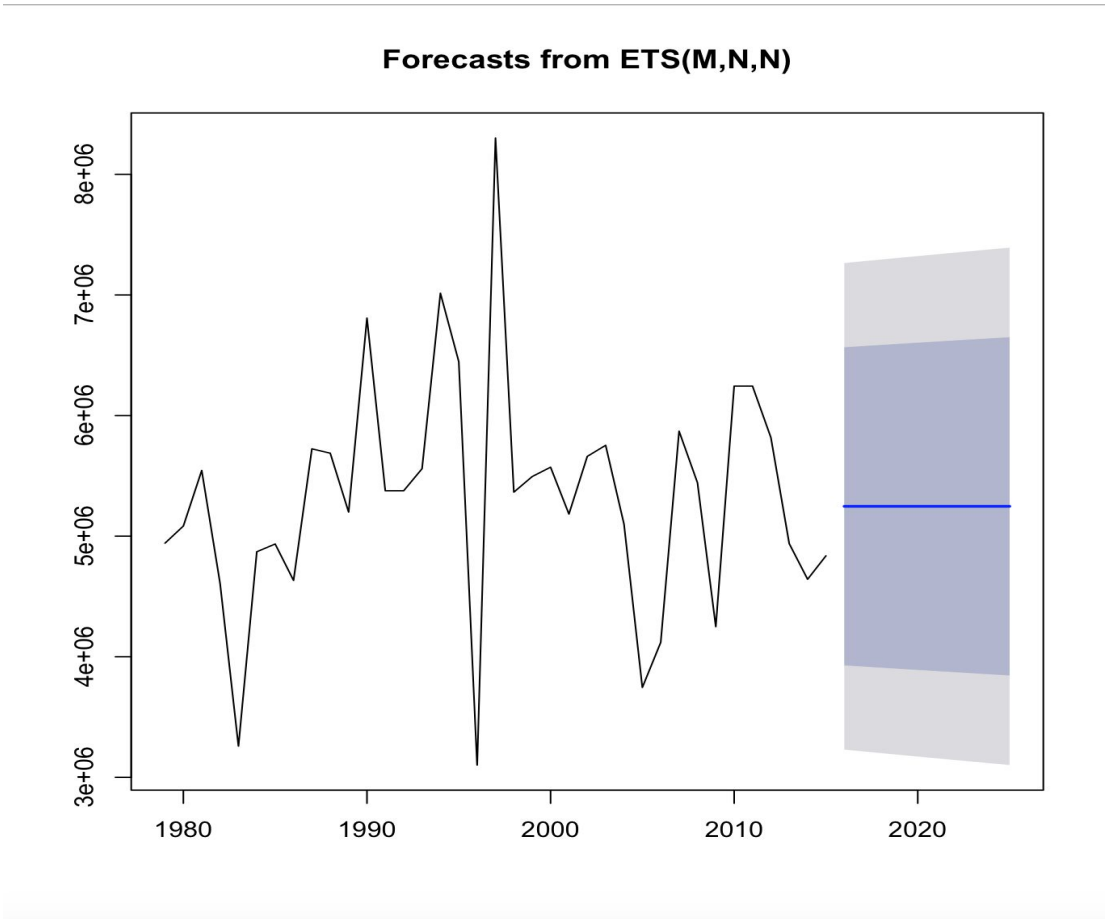
Time Series Forecasting:

```
>
> ETSHousing <- ets(HousingPrice)
>
> print(forecast(ETSHousing, 10))
     Point Forecast   Lo 80   Hi 80   Lo 95   Hi 95
2016        5247187 3928138 6566236 3229875 7264498
2017        5247187 3918625 6575748 3215327 7279046
2018        5247187 3909175 6585198 3200874 7293499
2019        5247187 3899786 6594587 3186515 7307858
2020        5247187 3890457 6603916 3172248 7322125
2021        5247187 3881187 6613186 3158071 7336302
2022        5247187 3871975 6622398 3143982 7350392
2023        5247187 3862819 6631554 3129979 7364395
2024        5247187 3853718 6640655 3116060 7378313
2025        5247187 3844671 6649702 3102225 7392149
>
> plot(forecast(ETSHousing, 10))
>
```

```
> ETSHousing
ETS(M,N,N)

Call:
 ets(y = HousingPrice)

  Smoothing parameters:
    alpha = 0.1181

  Initial states:
    l = 4928600.2367

  sigma:  0.1962

     AIC     AICc      BIC
1161.889 1162.616 1166.722
>
```

From the below graph, we find that we are forecasting house pricing for next 10 years using exponential smoothing with 80% and 90 % prediction interval for the forecast. Also since our alpha value is small (.118),which is close to 0 means that values of houses have not fluctuate much in past .

In the graph, darker areas shows more confidence level where pricing of houses is expected to reside over the period of next 10 years compared to lighter shaded area.



Prescriptive Analysis:

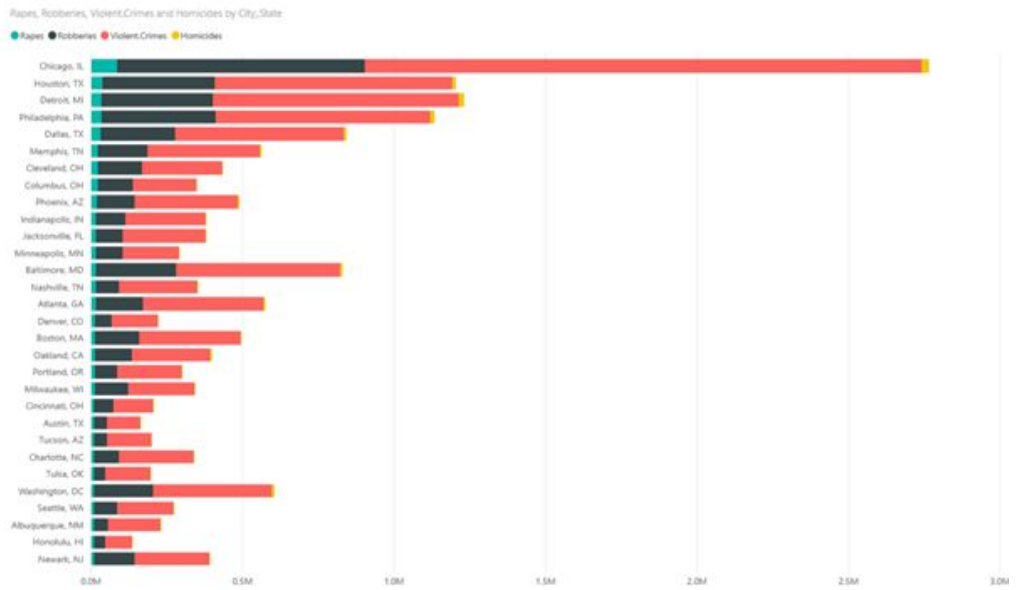
Prescriptive Analysis is the field of business analytics which is mainly to find the best course of action after evaluating any situation. From business perspective, it is an important analysis to quantify the results of future decisions to know the possible outcomes before actual decision is made. It answers the important question to business on What should they do now. After doing our descriptive and predictive analysis on housing price dataset, housing price across the region differs according to the rate of different types of crime occurrence, we have few recommendations to be considered.

1) From the analysis, we find that Chicago has the highest rates of crime. So, people of Chicago has to be made more aware about the upcoming new tools, so that they are prepared
2) The real estate should develop affordable homes, so that common man can have a quality of life, which would eventually decrease in the number of crimes.
3) The dealers should have detailed analysis of a particular region's housing price history before investing in. The customers and company, both should take decision wisely after looking through the analysis.

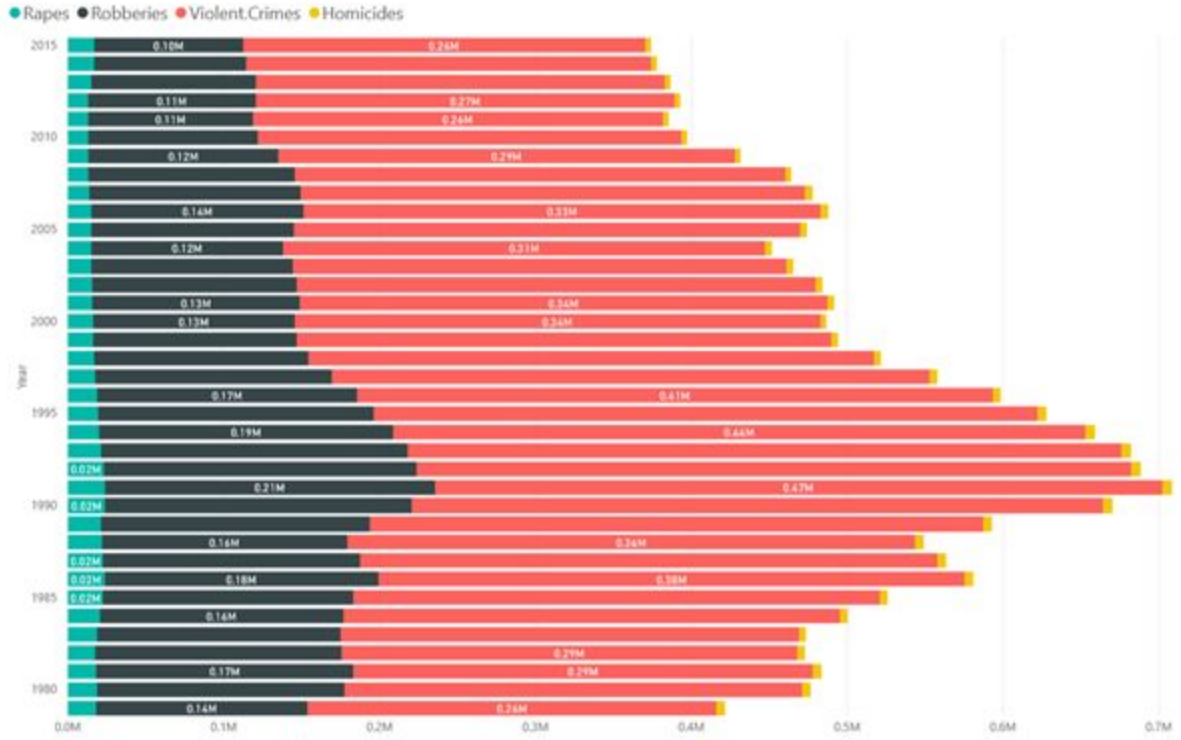
Prescriptive analytics leads to optimization in production, scheduling and inventory in the supply chain to make sure that customers are at the right path. From the analysis, it is clear that rise in housing price will suggest decrease in crime rate over the years and vice versa.

Virtualizations:

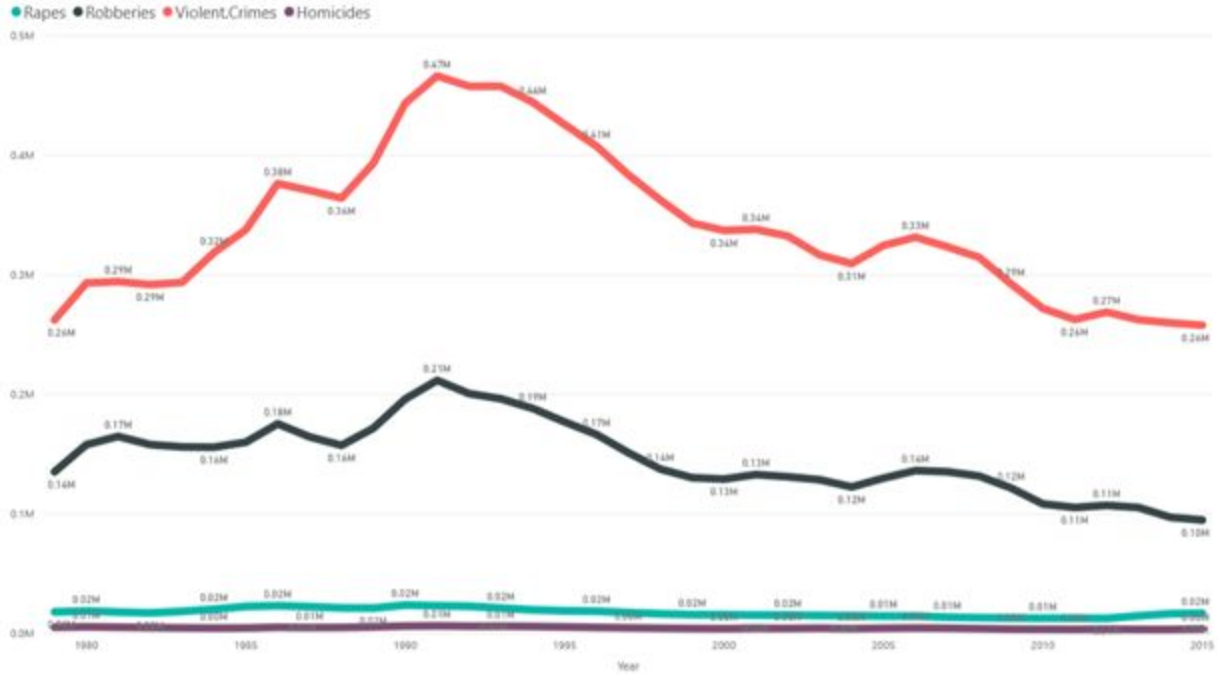
1. Below Graph showing the Rape, Robberies, Violent Crimes and Homicides data by City.



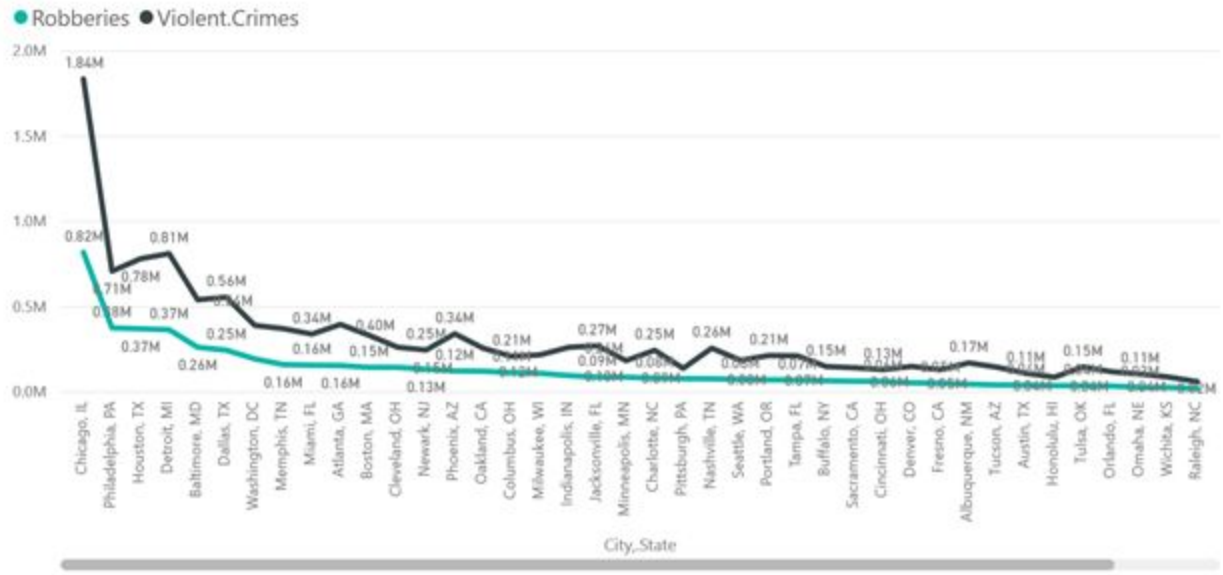
Rapes, Robberies, Violent.Crimes and Homicides by Year



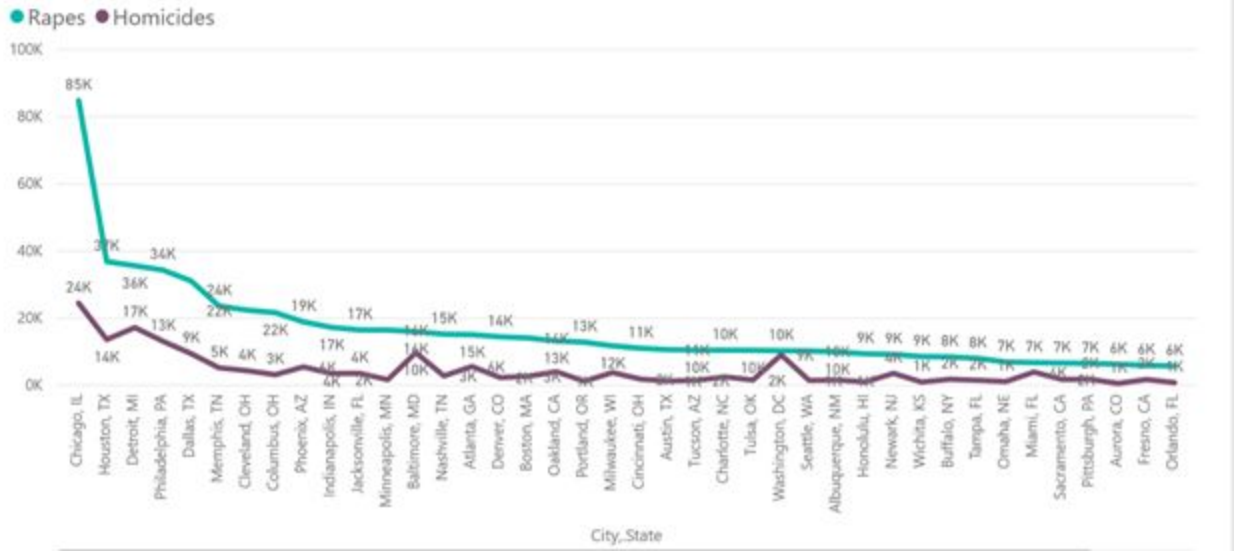
Rapes, Robberies, Violent.Crimes and Homicides by Year

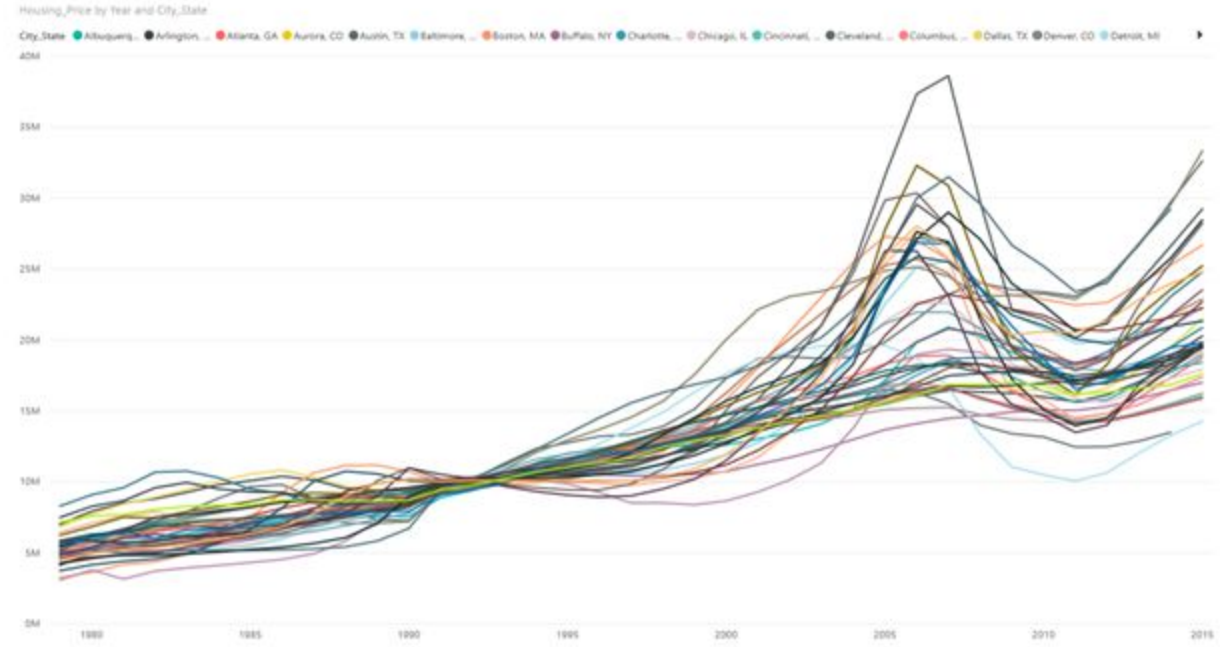


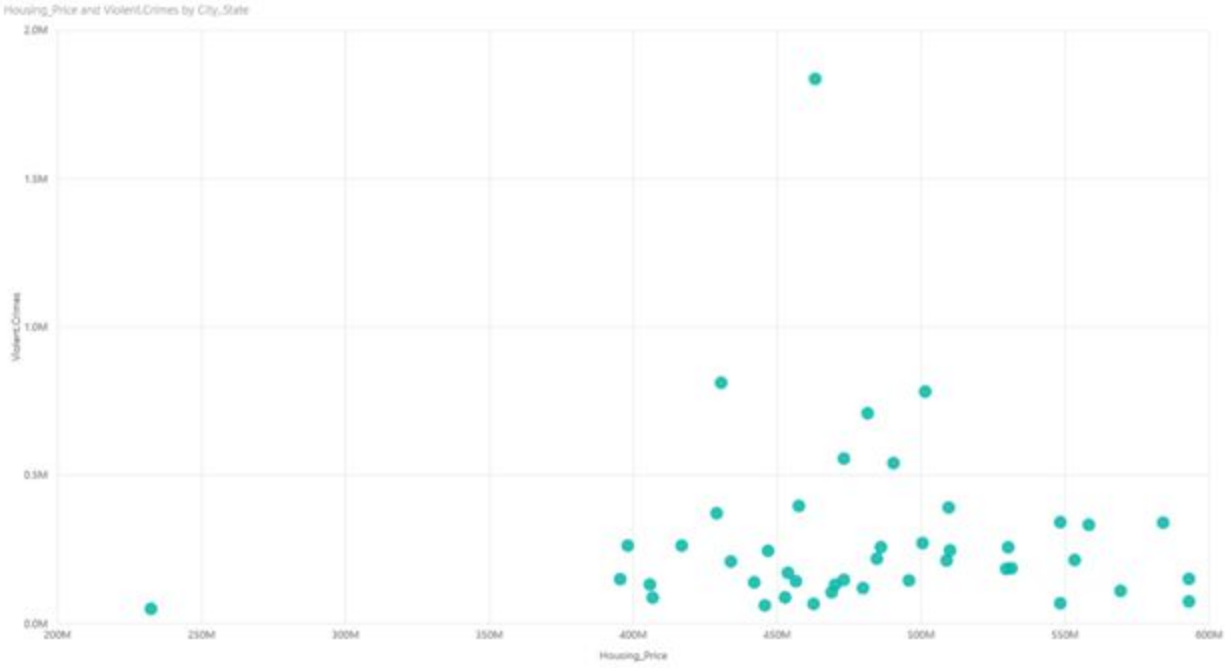
Back to Report | ROBBERIES AND VIOLENT.CRIMES BY CITY_STATE



Back to Report | RAPES AND HOMICIDES BY CITY_STATE

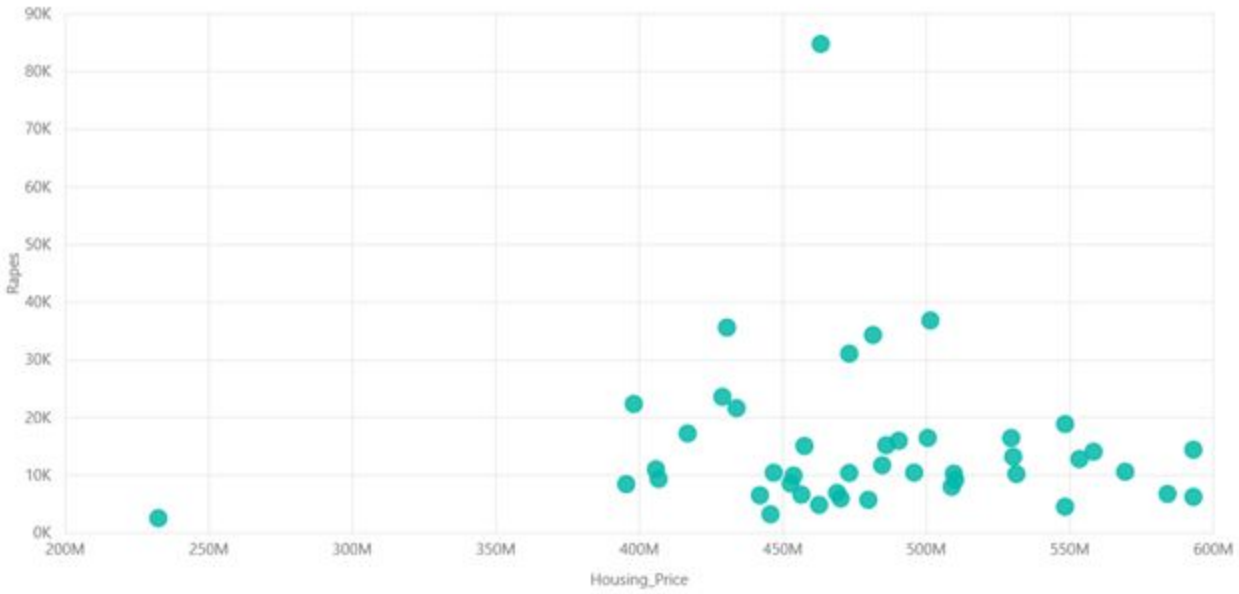



Housing_Price by Year and City_State
City_State
Albuquerq...
Arlington, ...
Atlanta, GA
Aurora, CO
Austin, TX
Baltimore, ...
Boston, MA
Buffalo, NY
Charlotte, ...
Chicago, IL
Cincinnati, ...
Cleveland, ...
Columbus, ...
Dallas, TX
Denver, CO
Detroit, MI
40M
35M
30M
25M
20M
15M
10M
5M
0M
1980
1985
1990
1995
2000
2005
2010
2015

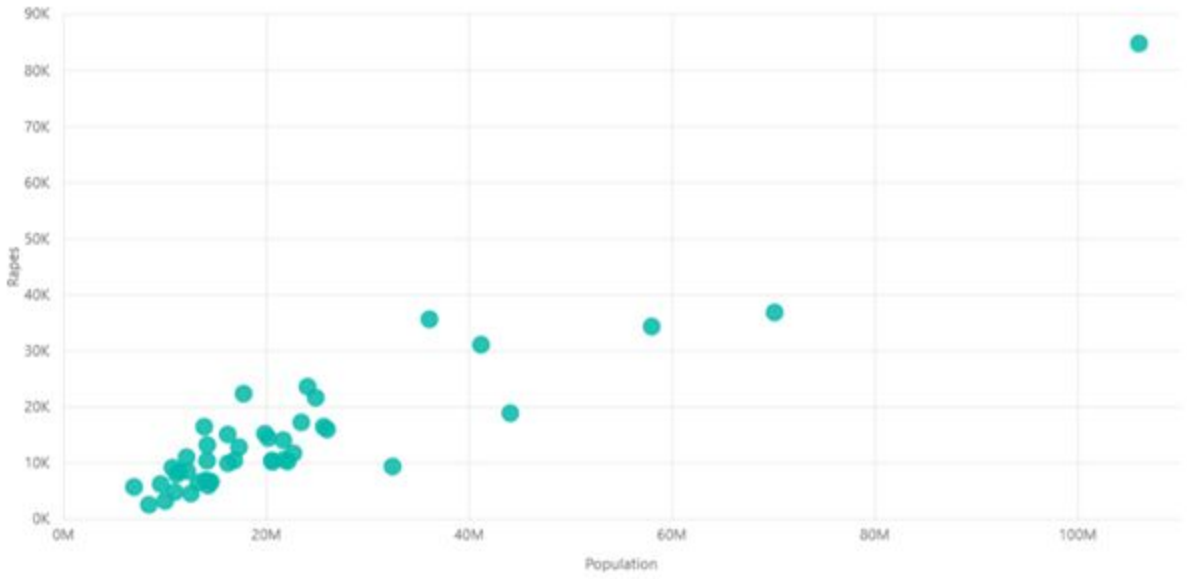

Housing_Price and Violent_Crimes by City_State
2.0M
1.5M
1.0M
0.5M
0.0M
Violent_Crimes
200M
250M
300M
350M
400M
450M
500M
550M
600M
Housing_Price


Back to Report
HOUSING_PRICE AND ROBBERIES
BY CITY_STATE
0.9M
0.8M
0.7M
0.6M
0.5M
0.4M
0.3M
0.2M
0.1M
0.0M
Robberies
200M
250M
300M
350M
400M
450M
500M
550M
600M
Housing_Price

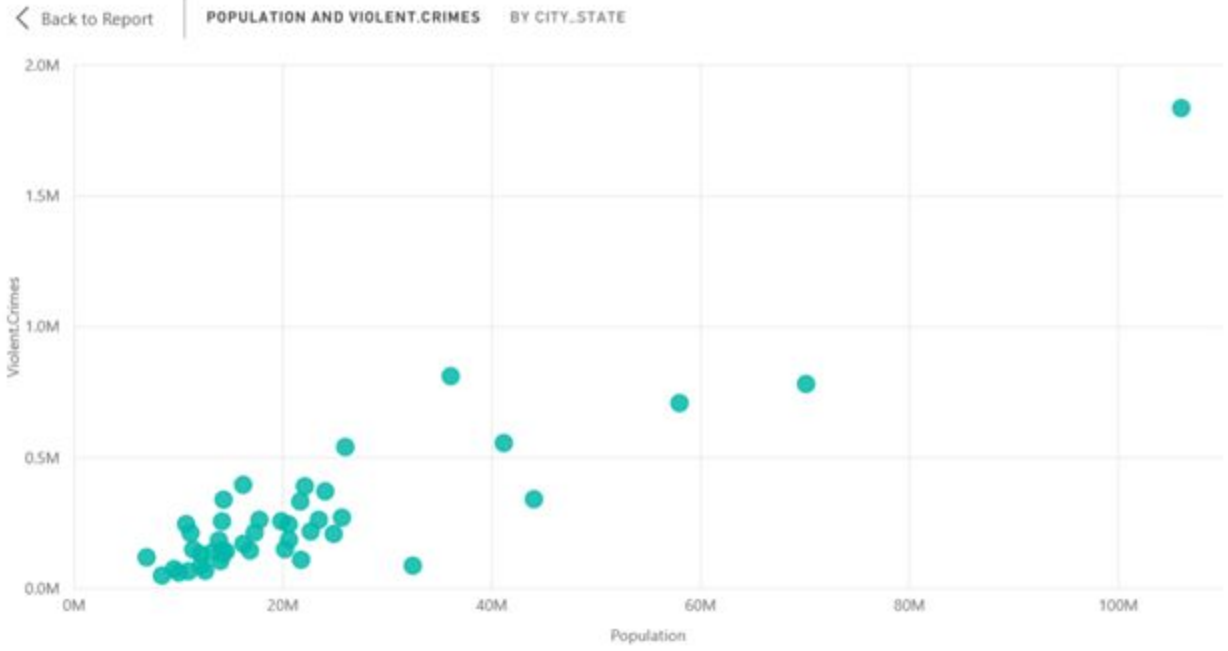
Back to Report | **HOUSING_PRICE AND RAPES** BY CITY_STATE

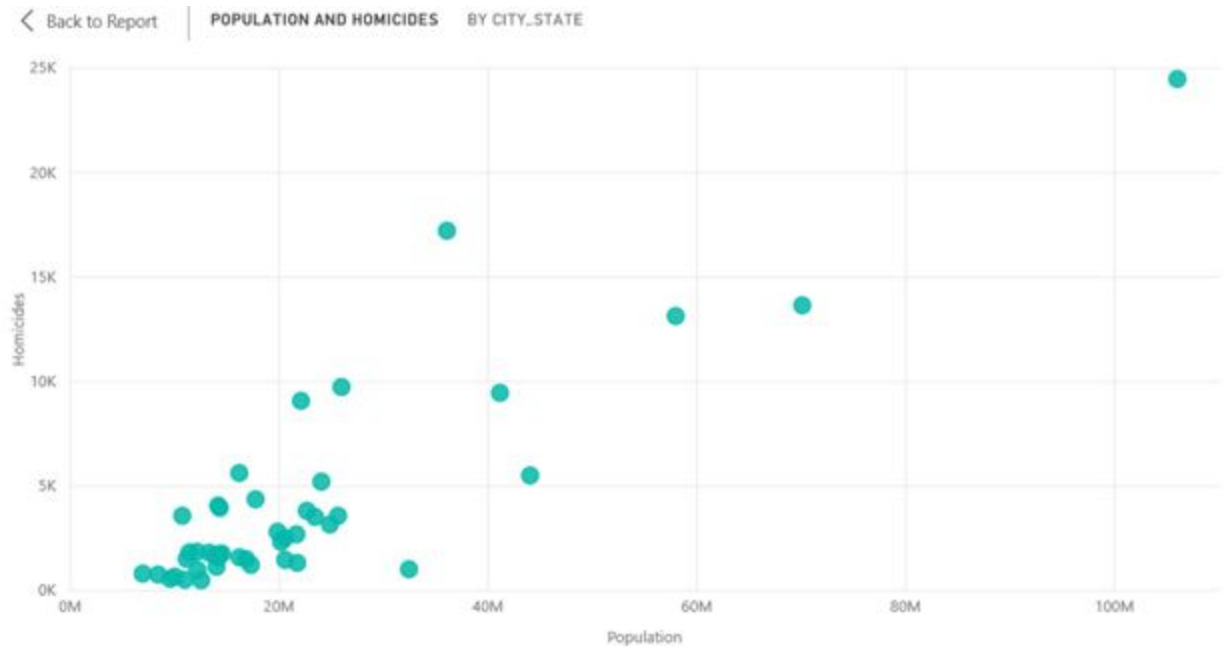


Back to Report | **POPULATION AND RAPES** BY CITY_STATE

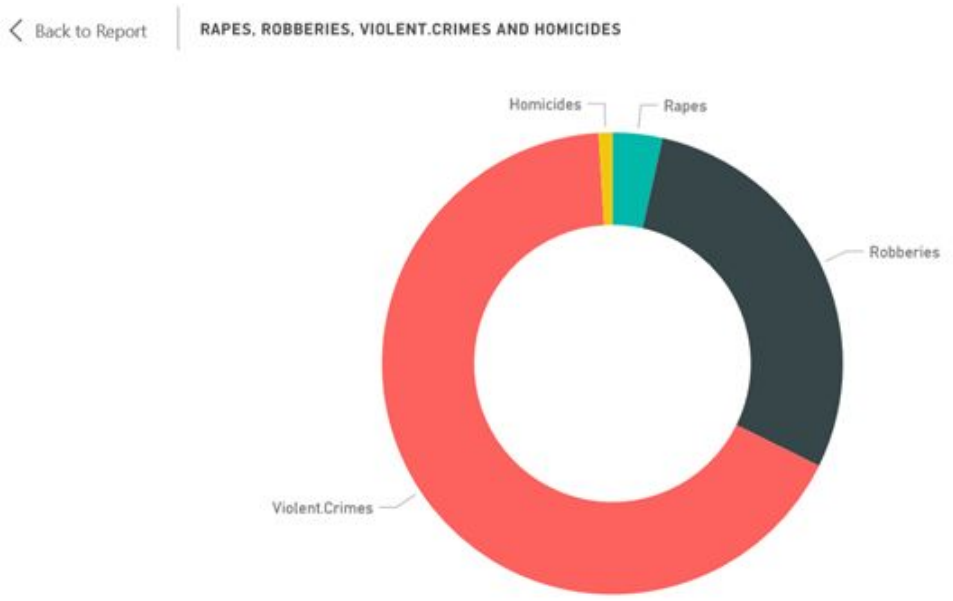



Back to Report
POPULATION AND ROBBERIES
BY CITY_STATE
0.9M
0.8M
0.7M
0.6M
0.5M
0.4M
0.3M
0.2M
0.1M
0.0M
Robberies
0M
20M
40M
60M
80M
100M
Population

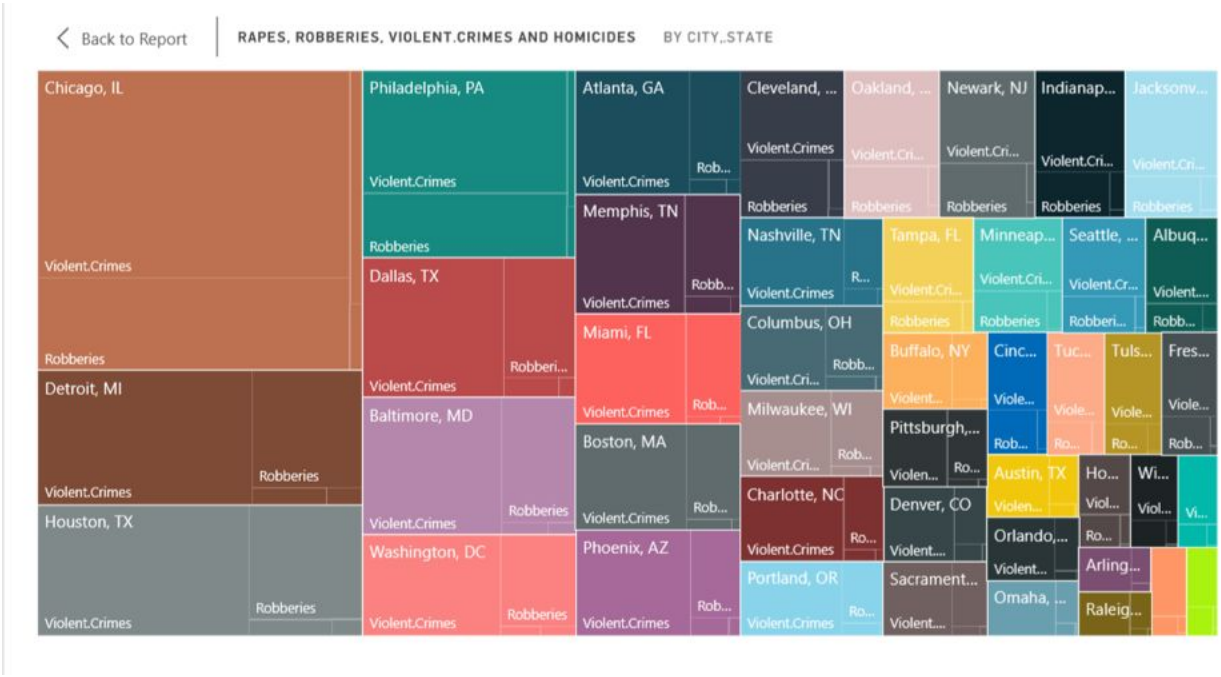

Back to Report
POPULATION AND VIOLENT.CRIMES
BY CITY_STATE
2.0M
1.5M
1.0M
0.5M
0.0M
Violent.Crimes
0M
20M
40M
60M
80M
100M
Population



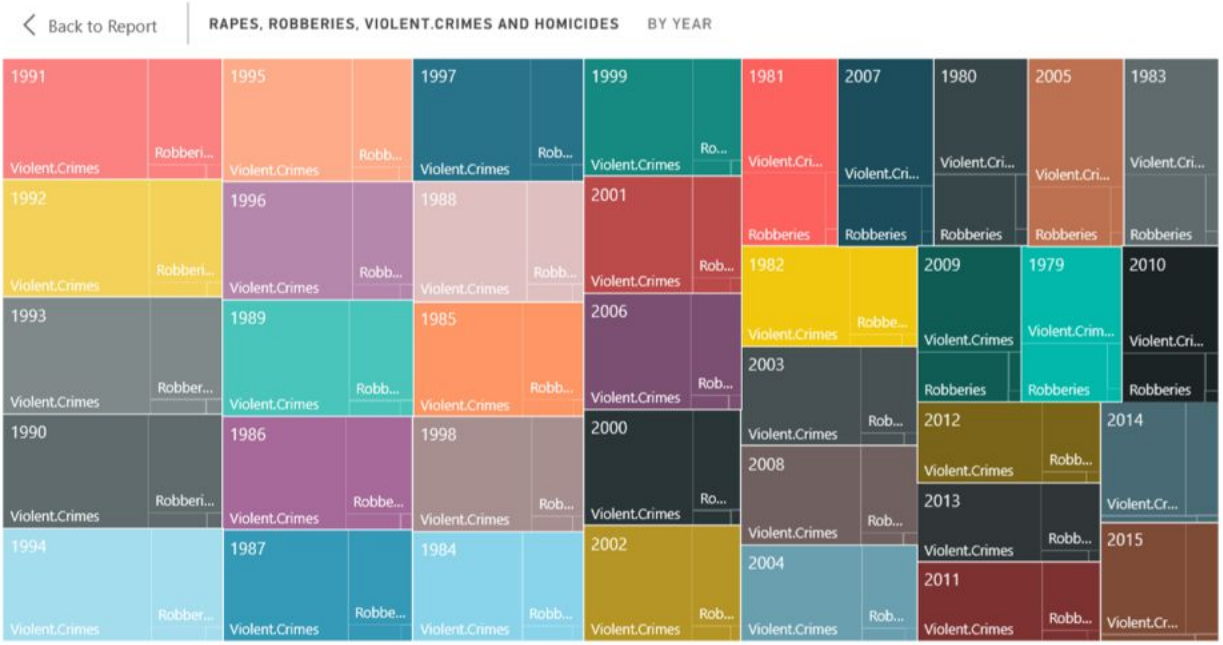
1. Below Graph showing the overall crime rate by Homicides, Rapes, Robberies and Violent Crimes.



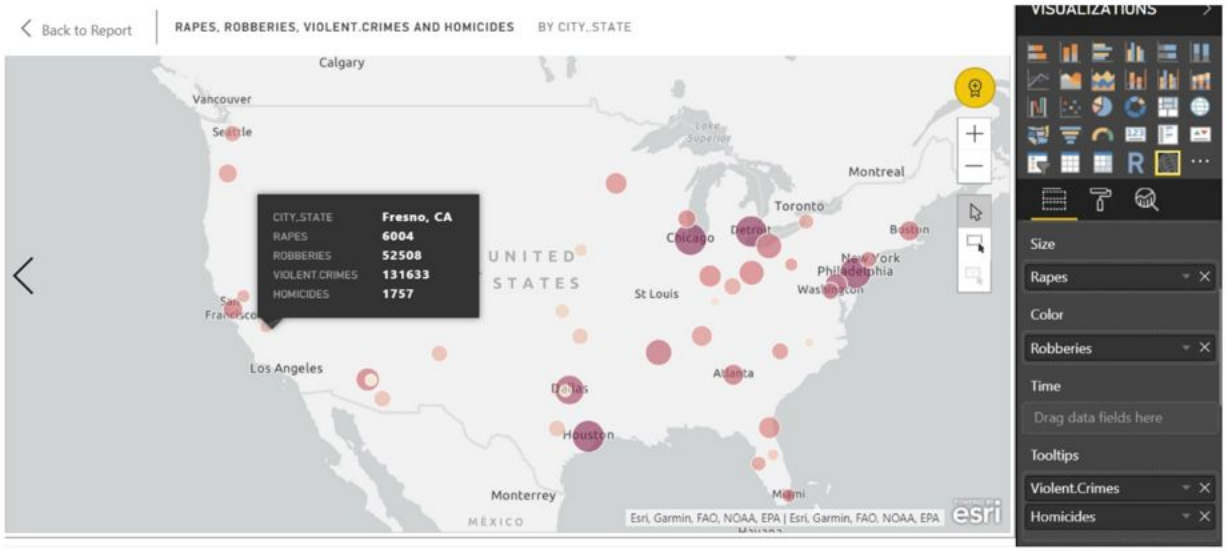
1. Tree map of Rapes, Robberies, Violent crimes and Homicides by city.



1. Tree map of Rapes, Robberies, Violent crimes and Homicides by Year.

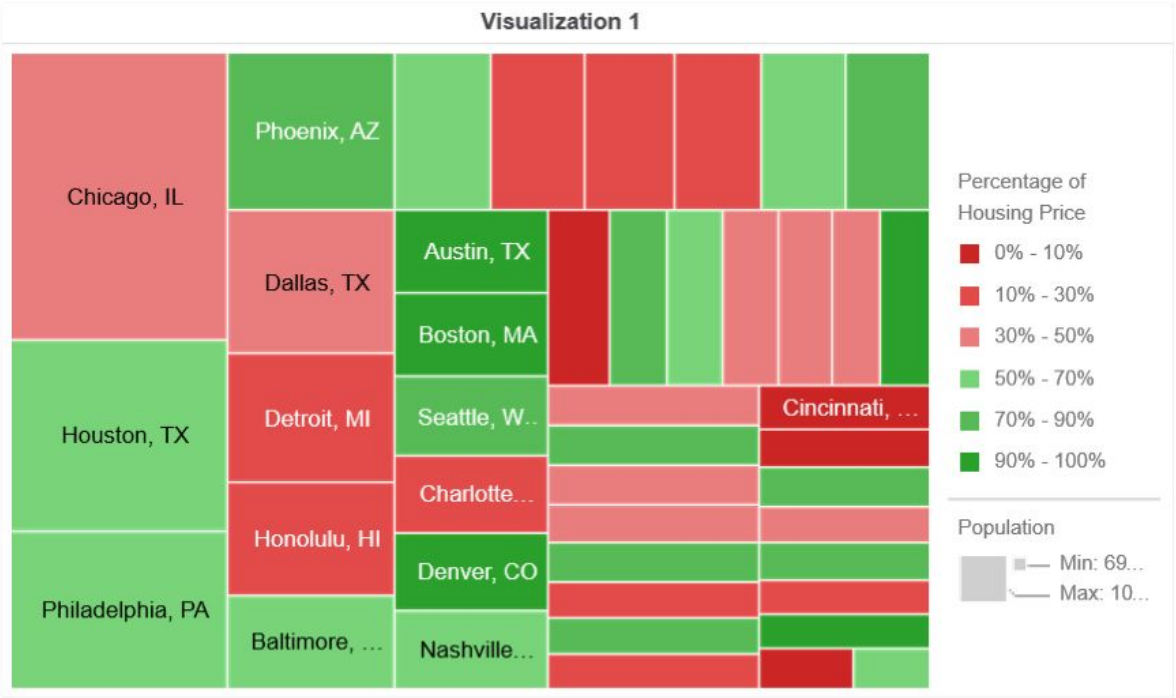


1. Map graph showing the cities impacted by different crimes with Size of circle indicating Rapes and color indicating Robberies.

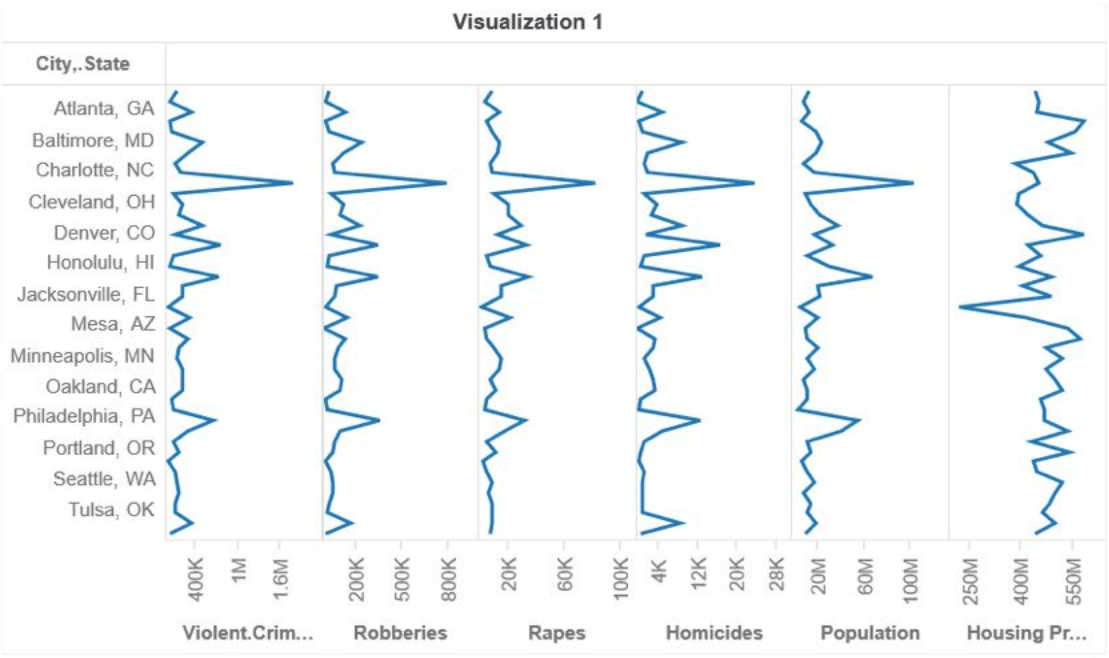


1. Map graph showing the cities impacted by different crimes with Size of circle indicating Homicides and color indicating Violent Crime.

1. Tree map indicating the city by percentage of Housing price.



1. Below graph showing the chart by city indicating violent crime, robberies, rapes, Homicides, Population and Housing price.



Conclusion:

The analysis of Housing Price data set shows that the housing prices for different cities are significantly different. From the correlation we found that population has weak positive correlation on housing prices and violent crimes, rapes and robberies have weak negative correlation with the housing prices. Since, there can be various factors for rise and fall of housing price, so business and consumer should have good information about any region before making an investment. Through this report, we have analysed the behaviour of crime across different

states from year 1985-2015, and we find that whenever there is increase in crime, the housing price would decrease.

As per the analysis, the Chicago has the highest rate of crime, and Detroit being the second. Also, violent crime and robberies have weak positive correlation with the housing price.

References:

1) Housing price index using Crime Rate Data(n.d). Retrieved from https://www.kaggle.com/sandeep04201988/housing-price-index-using-crime-rate-data

2) Sherman, R.(2015). Business Intelligence Guidebook.Waltham,WA: Elsevier