**HOUSING PRICES AND CRIME**

**Data Mining**

**Information Systems Department**

**Table of Contents**

[**1. ABSTRACT**](#_8gfbvm9l453f) **3**

[**2. INTRODUCTION**](#_craovse6w7wj) **3**

[2.1. Background](#_gfijv6yg13qa) 3

[2.2. Technology](#_t46ivfq6w6jl) 4

[**3. PROBLEM STATEMENT**](#_norim3s69fs0) **4**

[**4. METHODOLOGY**](#_oe4ryxkc7vtf) **4**

[4.1. Feature Selection](#_qp2bp7g30xg7) 5

[4.2. Data Preprocessing](#_ksp9weyptag6) 6

[4.3. Model Building](#_hqi8sm68o9r3) 6

[4.3.1. Linear Regression](#_ykupw4ryyd9j) 6

[4.3.2. Polynomial Regression](#_owf6m4w4oo0u) 8

[4.4. Model Evaluation](#_lhfuj2h5xuyu) 10

[4.4.1. Evaluation Plots](#_wg2q2zriz1cx) 10

[4.4.2. AIC Analysis](#_dm2x4vf0siny) 12

[**5. CONCLUSION**](#_glghw1vub5r4) **13**

[**6. FUTURE WORK**](#_r1awxgrpnbni) **13**

[**7. REFERENCES**](#_9uj0f1610t5u) **13**

# 

# **1. ABSTRACT**

With the crimes happening in the U.S. and with increasing number of people considering crime rate near by, the concern that the housing prices being affected is put forth by the stakeholders. Purchasing a house is considered the most used and beneficial venture to the people who know the advantage of owning a house. Every year, on an average, 200,000 houses are sold in the United States. Therefore, an accurate house price prediction based on crime rate is more important to investors, house owners, buyers and also to other real estate shareholders.

The historical sales data set which we got to predict the house price contains major US metro area sales records from 2014 - 2017 includes several predicting factors such as location, house type, square feet, number of bedrooms, year built, etc [1].

The correlations between the dependent and independent variables are found out using ‘Pearson Correlation of features’ and the most significant and least significant variables are obtained. The ‘crime data’ variable is then merged with the housing price data set and both linear regression model and polynomial regression model are built for comparison to get the most accurate model. Using polynomial regression model, we find that crime rate has a large impact on the housing prices around the cities. Also, important to the explanation of variations of housing prices are variables derived from linear regression and polynomial regression model, such as square feet, beds and baths. This article concludes with the effort to apply the model to predict housing prices for different cities present in the dataset based on the crime rate.

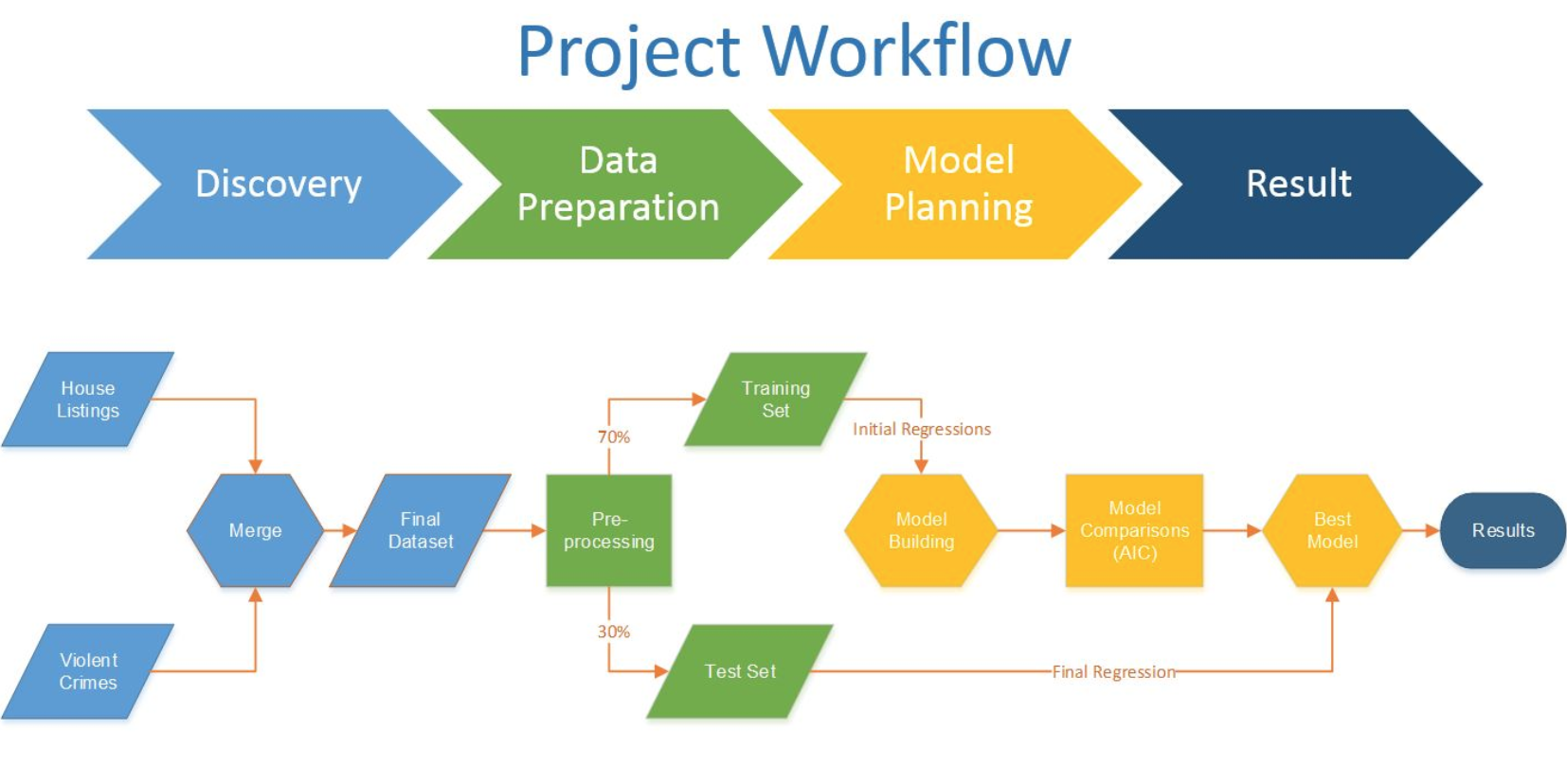
# **2. INTRODUCTION**

## **2.1. Background**

Houses are the most common means of accumulating personal wealth in the U.S. It becomes very important to predict the housing prices accurately since the owners change and it can be used for mortgages and home owners’ insurance. There are a lot of factors that can influence the selling price of the house including square feet, number of beds, number of baths. The other factor that can influence is the crime rate in that particular city.

The project consists of four phases namely, discovery of data, data preparation, model planning and results. Discovery of data includes obtaining the data and converting it into quality data. Data preparation includes the preprocessing of the raw data where least significant variables are removed. One method of predicting the housing prices is to use the data and apply linear regression on it. Firstly, Pearson correlation factor is applied to the historical data obtained and the correlations between the independent and dependent variables are found out. The least and the most significant factors are obtained. The correlations are then used to predict the housing prices. However, the housing prices are predicted without the crime data.

The crime data is then merged to the housing price data and then the model is run again. For comparison and to get a better model, polynomial regression model is run for degree 2 and degree 3. The workflow of our model goes as below,



## **2.2. Technology**

This project utilizes different technologies from data processing to model building. The data preprocessing is done with Python scripting where the pearson correlation factor is obtained for different variables and a heat map is drawn distinguishing the most significant and the least significant variables. The same was confirmed when used with WEKA tool.

Once the most significant variables were obtained, the model building was the next phase. Both linear regression model and polynomial regression models (Degree 2 and Degree 3) were built in RStudio using R language. Building models in R not only gave an appropriate and an accurate model, also visualization of the model through graphs.

# **3. PROBLEM STATEMENT**

Develop a machine learning model which can predict house price in the US based house features and crime rate within a particular zone. The resulting pricing model will be used to determine initial asking prices for houses.

# **4. METHODOLOGY**

The project is developed by using two different datasets. The first data set is fully focused to predict the house price and the second dataset contains the crime’s information based on the cities. The Historical data set to predict house price is obtained from the website called redfin a full-service real estate brokerage who helps people to buy and sell the homes. This dataset contains 25235 instances and 13 attributes which is more about the house and not about the surrounding features of the house. For example, a number of Beds and Baths. Initially, the project’s objective is to predict the house price and crime based on the Zipcode or Latitude and longitude. But after facing so many challenges in finding the crime dataset, we ended up with selecting aggregated Crime dataset from the website called data.gov were we can get government data from across the Federal Government. And the dataset contains 11 instances and 15 attributes. For example,Violent crime and City. After finalizing the data set, the next process is feature selection.

## **4.1. Feature Selection**

Both, Home price and crime dataset contain the mixture of attributes which are relevant to perform the prediction and some are not relevant. So the Feature selection is an important process of eliminating the unwanted attributes and selecting the relevant attributes without changing anything in the dataset for use in model construction. In this project, the dataset house price contains 13 attributes and crime contains 15 attributes. Some attributes might have a strong relationship and would be relevant to the predictive modeling. Basically, the feature selection process is divided into two parts. That is, attribute evaluator and search methods. In this project, we used a popular method called **Pearson Correlation method** for selecting the most relevant attributes in a dataset[4].

The correlation method is a process of calculating the correlation between each attribute and the output variable and selecting only those attributes that have a moderate-to-high positive or negative correlation, which is close to -1 or 1 and drop the irrelevant attributes with a low correlation, which is close to zero.

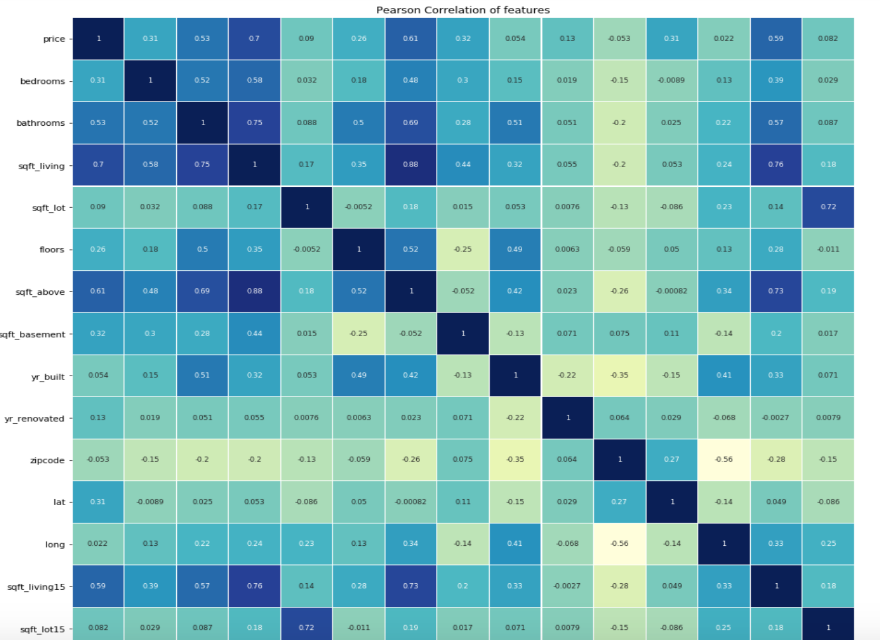


Fig.4.1 Pearson Correlation Method

Figure 4.1 is an output of the Pearson correlation method which has applied to house price dataset. In this project, the correlation is performed between price, the output variable and other attributes. Based on the output, we chose four attributes from the house dataset which are relevant (close to -1 or 1) to this prediction model. The attributes are, Square feet, Beds, Baths, Year Built. But for Crime dataset, we did not apply any correlation method, we chose only one attribute that is a Violent crime which is more related to the house price dataset. Now the feature selection process is done by selecting the attributes which we needed for this predictive modeling.

The real world data are generally noisy, inconsistent and incomplete. So the whole data mining process cannot be completed only by feature selection. A quality decision is based on how quality our data is. To get a quality data, we need some data preprocessing or a data preparation steps.

## **4.2. Data Preprocessing**

Data preprocessing is basically any type of process applied to a raw data to prepare the dataset for another processing procedure. The goal of data preprocessing is to change the dataset into the easy and adequate format. The data preprocessing steps are cleaning, integration, reduction, transformation, and discretization. As we are using historical and real-time dataset, we do not want to perform all the data processing steps. In this project, the required steps are Data cleaning and Data Integration[5].

**Data Cleaning**

The use of data cleaning step is to fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies. In this project, the dataset house price only has null values which can easily be identified and eliminated from the dataset. The house price dataset contains 25235 instances. We did a cleaning process on this dataset and finalized the dataset with **7214** instances.

**Data Integration**

Data Integration is a process of combining data from multiple data source into a coherent data store. After cleaning step, we have two datasets, that is house price and crime dataset which has relevant attributes and reduced volume of data without compromising the integrity of the original data. In integration step, we merged our two datasets into one and the attributes we considered are, Square feet living, Beds, Baths, Year Built and Violent Crime.

## **4.3. Model Building**

Our goal is to provide an accurate prediction of a house price in a given US metro area, this type of problem can be addressed by different regression methods, we considered different regression options.

### **4.3.1. Linear Regression**

In this project our goal is develop a prediction model, whenever it comes to regression problem, Linear regression is the simplest and most widely used statistical technique for predictive modeling. Linear Regression establishes a relationship between dependent variable (**Y**) and one or more independent variables (**x**) using a best fit straight line (also known as regression line)[2].

It is represented by an equation

|  |
| --- |
| **Y= β0+ β1x + e** |

where **β0** is intercept, **β1** is slope of the line and e is error term. This equation can be used to predict the value of House Price. For more than one independent variable, the process is called multiple linear regression[2].

Linear regression is one of the core methods that can be used to develop a model to predict a

continuous response from multiple predictor variables. In the midterm report, we generated our linear model lm with R2 = 0.58, which indicated that the model didn’t fit the input data well enough as linear prediction. From the previous model, we selected significant features of square feet of house, the number of bedrooms, the number of bathrooms, square feet of lot and the year built. Besides, we included a new attribute, the number of violent crimes of the city where the house located. After fitting the new linear regression model, we found that all the features we selected had significant correlations for predicting the price of a house, where the P‐values were significantly small (less than 0.01). So, we are confident that all the attributes are significantly important to predict the house pricing data. Also, the adjusted R2 of the new linear regression model is 0.9605, which is a huge improve (increased R2 by 65.6%) compared to the previous R2, and it is very close to 1. We believe that our decision of adding crime data into the price prediction model works well in this domain.

Here is the model summary and plots generated by R for the house price predicting linear regression model:

|  |  |
| --- | --- |
| Call:  lm(formula = formula, data = mhp.train)  Residuals:  Min 1Q Median 3Q Max  -1977840 -129665 -30154 78763 6105371  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 7.337e+06 3.210e+05 22.86 <2e-16 \*\*\*  SQUARE.FEET 2.669e+02 1.260e+00 211.91 <2e-16 \*\*\*  BEDS -7.624e+04 5.112e+03 -14.91 <2e-16 \*\*\*  BATHS 1.032e+05 5.994e+03 17.21 <2e-16 \*\*\*  YEAR.BUILT -3.688e+03 1.623e+02 -22.73 <2e-16 \*\*\*  Violent.Crimes -1.330e+01 7.209e-01 -18.46 <2e-16 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 279600 on 5043 degrees of freedom  Multiple R-squared: 0.9605, Adjusted R-squared: 0.9604  F-statistic: 2.451e+04 on 5 and 5043 DF, p-value: < 2.2e-16 |  |

### 

### **4.3.2. Polynomial Regression**

Polynomial regression is a form of regression analysis in which the relationship between the independent variable x and the dependent variable y is modeled as an nth degree polynomial in x[2]. Polynomial regression fits a nonlinear relationship between the value of x and the corresponding conditional mean of y, denoted E(y |x), and has been used to describe nonlinear phenomena. Although polynomial regression fits a nonlinear model to the data, as a statistical estimation problem it is linear, in the sense that the regression function E(y | x) is linear in the unknown parameters that are estimated from the data. For this reason, polynomial regression is considered to be a special case of multiple linear regression. Polynomial regression extends the linear model by adding extra predictors, obtained by raising each of the original predictors to a power. Following equation represents generic polynomial regression of degree **n**.

|  |
| --- |
| **y =β0+ β1x + β2x2 + β3x3+ … + βnxn + ε** |

For our problem we used Polynomial regression of degree 2 and 3, following are the results

**Polynomial Regression of Degree 2:**

In polynomial regression of degree 2 the power of independent variable is raise to the power of 2.

|  |  |
| --- | --- |
| Call:  lm(formula = mhp.train$PRICE ~ poly(mhp.train$SQUARE.FEET, degree = 2,  raw = TRUE) + poly(mhp.train$BEDS, degree = 2, raw = TRUE) +  poly(mhp.train$BATHS, degree = 2, raw = TRUE) + poly(mhp.train$YEAR.BUILT,  degree = 2, raw = TRUE) + poly(mhp.train$Violent.Crimes,  degree = 2, raw = TRUE))  Residuals:  Min 1Q Median 3Q Max  -2128723 -113726 -22887 75429 6076579  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 2.594e+08 1.872e+07 13.855 < 2e-16 \*\*\*  2.234e+02 7.017e+00 31.842 < 2e-16\*\*\*  2.497e-04 5.333e-05 4.682 2.92e-06\*\*\*  -2.436e+04 1.085e+04 -2.245 0.0248\*  -5.236e+03 1.205e+03 -4.346 1.41e-05\*\*\*  5.068e+04 1.018e+04 4.976 6.70e-07\*\*\*  7.451e+03 1.134e+03 6.569 5.58e-11\*\*\*  -2.599e+05 1.900e+04 -13.681 < 2e-16\*\*\*  6.523e+01 4.819e+00 13.537 < 2e-16\*\*\*  -1.036e+02 4.144e+00 -25.010 < 2e-16\*\*\*  3.157e-03 1.442e-04 21.898 < 2e-16\*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 261400 on 5038 degrees of freedom  Multiple R-squared: 0.9655, Adjusted R-squared: 0.9654  F-statistic: 1.409e+04 on 10 and 5038 DF, p-value: < 2.2e-16 |  |

**Polynomial Regression of order 3:**

In polynomial regression of degree 3 the power of independent variable is raise to the power of 3.

|  |  |
| --- | --- |
| Call:  lm(formula = mhp.train$PRICE ~ poly(mhp.train$SQUARE.FEET, degree = 3,  raw = TRUE) + poly(mhp.train$BEDS, degree = 3, raw = TRUE) +  poly(mhp.train$BATHS, degree = 3, raw = TRUE) + poly(mhp.train$YEAR.BUILT,  degree = 3, raw = TRUE) + poly(mhp.train$Violent.Crimes,  degree = 3, raw = TRUE))  Residuals:  Min 1Q Median 3Q Max  -2539173 -82370 -9032 61173 5912583  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) -8.397e+09 9.866e+08 -8.511 < 2e-16\*\*\*  2.323e+01 1.355e+01 1.714 0.08651 .  3.738e-02 1.896e-03 19.715 < 2e-16\*\*\*  -1.019e-07 5.494e-09 -18.545 < 2e-16\*\*\*  7.694e+03 1.742e+04 0.442 0.65885  -9.449e+03 3.148e+03 -3.002 0.00270\*\*  2.669e+02 1.447e+02 1.844 0.06522 .  4.347e+04 1.601e+04 2.715 0.00664\*\*\*  8.188e+03 2.735e+03 2.994 0.00277\*\*  -4.255e+02 1.066e+02 -3.991 6.68e-05\*\*\*  1.290e+07 1.505e+06 8.571 < 2e-16\*\*\*  -6.606e+03 7.657e+02 -8.627 < 2e-16\*\*\*  1.127e+00 1.298e-01 8.682 < 2e-16\*\*\*  -6.166e+02 1.564e+01 -39.430 < 2e-16\*\*\*  4.658e-02 1.287e-03 36.209 < 2e-16\*\*\*  -1.085e-06 3.193e-08 -33.971 < 2e-16\*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 225600 on 5033 degrees of freedom  Multiple R-squared: 0.9743, Adjusted R-squared: 0.9742  F-statistic: 1.273e+04 on 15 and 5033 DF, p-value: < 2.2e-16 |  |

## 

## **4.4. Model Evaluation**

When we run a regression analysis, our first task might be check if the model works well for the data and we can check this in many ways based on a test set. The regression results such as slope coefficients, p-values or R-squared (Goodness - of - fit) tells us how well a model represents our house pricing data. And the residual shows how perfectly a model represents data. The diagnostic plots show residuals in four different ways and it can not check only the linear regression assumptions but also improve the model in an empirical way. This is also known as model evaluation for regression.

**Goodness-of-fit:**

A common way to summarize how well a linear regression model fits the data is via the coefficient of determination or R-squared. This can be calculated as **the square of the correlation between the observed values and the predicted values**. Alternatively, it can also be calculated as:

=

Where, SSE is the error sum of squares and SST is the total sum of squares.

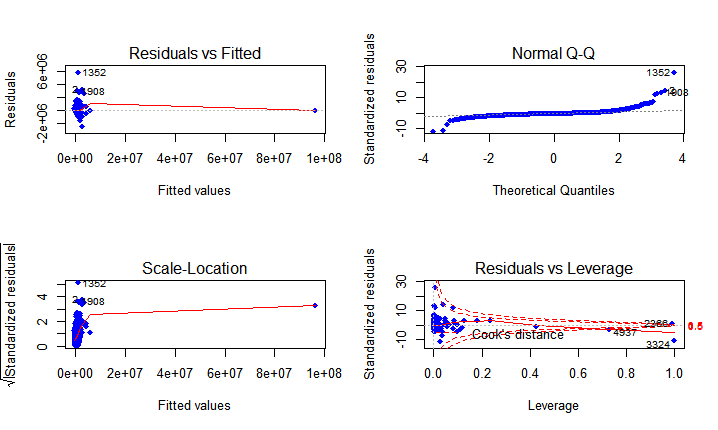
Thus, it is also **the proportion of variation in the forecast variable that is accounted for (or explained) by the regression model**. So, from our model we get R-squared= 0.9605 that is 96% of the variation in the house price is captured by the model.

### **4.4.1. Evaluation Plots**

In this section we will evaluate the plots generated from Polynomial regression of degree 3.

**Residuals vs Fitted:**

From this plot we can find that the residuals have no distinctive patterns that means residuals are equally spread around a horizontal line without distinct patterns. Which indicate that there is a linear relationship between house pricing and the predictors such as square feet, number of beds, number of baths, lot size, year of built and violent crimes. So, from the residuals plot we can say that the house pricing data meets the regression assumptions very well and it’s the indication of good model.



**Normal Q-Q plot:**

This plot shows that the residuals are normally distributed. The residuals follow a straight line well except for the observation numbered 1352,12 and 908. That might not be a problem for normality assumptions.

**Scale-Location:**

This is called the Spread-Location plot. This plot shows if the residuals are spread equally along the ranges of predictors. It’s the assumption of homoscedasticity (i.e. the equal variance). It will be good if we see a horizontal line with equally (randomly) spread point. We can see from the graph that the residuals are not spread much except for one point and the red smooth line is not horizontal and shows a steep angle . So it violate the assumption of regression.

**Residuals vs Leverage:**

|  |  |
| --- | --- |
|  |  |

The Residuals vs Leverage plot is used for model diagnosis, which helps find influential cases and ensure that not all outliers are influential in regression analysis. Even though data have extreme values, but they might not be influential to determine a regression line. When cases are outside of the Cook’s distance (e.g. they have high Cook’s distance scores), they are influential to the regression model. Thus, the regression results will be altered if we exclude those cases. These two plots are generated from polynomial degree 3 regression. The left plot shows that number 2366 & 3324 observations are beyond the Cook’s distance line as the influential observations. The right plot shows that the influential outliers are excluded from the model.

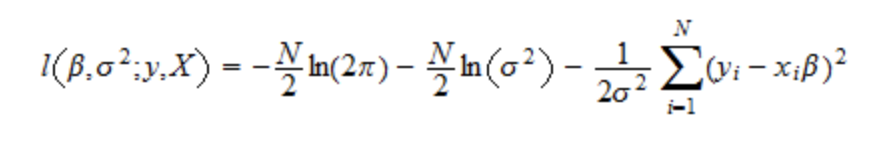
### **4.4.2. AIC Analysis**

**Akaike information criterion** (**AIC**) is an estimator of the relative quality of statistical models for a given set of data. Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. Thus, AIC provides a means for [model selection](https://en.wikipedia.org/wiki/Model_selection). The formula to calculate the AIC value is:

AIC = 2k - 2 ln()

Where, k is the number of estimated [parameters](https://en.wikipedia.org/wiki/Statistical_parameter) in the model and be the maximum value of the [likelihood function](https://en.wikipedia.org/wiki/Likelihood_function) for the model.

The log likelihood functions



The log-likelihood for Model 1 : -70481.86 with 7 degrees of freedom

The log-likelihood for Model 2 : -70139.69 with 12 degrees of freedom

The log-likelihood for Model 3 : -69392.65 with 17 degrees of freedom

**The Adjusted R-squared Value: (Better if Close to 1)**

Model 1 (Linear): 0.9604

Model 2 (Poly 2): 0.9654

Model 3 (Poly 3): 0.9742

**The AIC Value: (Better if Smaller)**

Model 1 (Linear): 140975.4

Model 2 (Poly 2): 140303.4

Model 3 (Poly 3): 138819.3

# 

# 

# **5. CONCLUSION**

We develop a prediction model which can predict house price in US metro area. This model will evaluate and weigh several relational attributes of real estate property for predictions. The is model capable of predicting house prices by considering crime rate as one of the major factors. Several other factors were also considered for predicting house prices. In order to find the solution for our goal, we have identified that this problem falls under the regression problem. We researched on different regression algorithms, and applied appropriate algorithm based on performance measure matrix. We analysed 3 models and concluded that Polynomial Degree 3 model is the best amongst 3 models.

# 

# **6. FUTURE WORK**

Our model has certain limitations and has good scope for enhancements. In current study, we generated a prediction model which uses house attributes and relative information like crime rate. This can be enhanced further by including more location-based information like employment rate, school rating, grocery stores, walking score, medical facilities, shopping centers etc.

# 

# 

# 

# **7. REFERENCES**

1. Data Source: REDFIN ([http://redfin.com](http://redfin.com/)) and U.S. Government’s Open Data ([http://data.gov](http://data.gov/))
2. John O. Rawlings, Sastry G. Pantula, David A. Dickey.Applied Regression Analysis: A Research Tool, Second Edition, 1989.
3. Alexander Ly, Maarten Marsman, Eric-Jan Wagenmakers.Analytic posteriors for Pearson’s correlation coefficient, 2017.
4. Brownlee. J. (2016). How to Perform Feature Selection With Machine Learning Data in Weka. Retrieved from <https://machinelearningmastery.com/perform-feature-selection-machine-learning-data-weka/>
5. Rouse. M. (2005). data preprocessing. Retrieved from <https://searchsqlserver.techtarget.com/definition/data-preprocessing>
6. David A. Freedman (2009). Statistical Models: Theory and Practice. Cambridge University Press. p. 26. A simple regression equation has on the right hand side an intercept and an explanatory variable with a slope coefficient. A multiple regression equation has two or more explanatory variables on the right hand side, each with its own slope coefficient.

**8. APPENDIX**

R program we developed to generate our prediction model:

|  |
| --- |
| # Purpose : This code will load and merge house price data, build, train and compare multiple regression models  #Loading data  library(stats4)  library(bbmle)  library(Matrix)  hp <- read.csv("metro\_sales\_updated.csv")  vc <-read.csv("ViolentCrimeRates.csv")  # Removing unwanted columns from Metro  hp$SOLD.DATE=NULL  hp$ZIP=NULL  hp$LATITUDE=NULL  hp$LONGITUDE=NULL  # Removing unwanted columns from Violent Crimepar(mfrow=c(1, 2))  vc$Aggravated.Assault.Rate.Per.100.000.People=NULL  vc$Robbery.Rate.Per.100.000.People=NULL  vc$Rape.Rate.Per.100.000.People=NULL  vc$Murder.Rate.Per.100.000.People=NULL  vc$Violent.Crime.Rate.Per.100.000.People=NULL  vc$Population=NULL  #Renaming Column to match on both the dataset  names(vc)[1] <-"CITY"  # Merging house price and crime datasets and storing into new one  mhp <- merge(hp,vc,by=c("CITY"))  #Omitting NA records  mhp <-na.omit(mhp)  #Factoring property type  mhp$PROPERTY.TYPE <- factor(mhp$PROPERTY.TYPE)  mhp$Latitude=NULL  mhp$Longitude=NULL  #setting seed  set.seed(1234)  #splitting data set into training 70% and test 30%  mhp\_train <- sample(nrow(mhp), 0.7\*nrow(mhp))  mhp.train <- mhp[mhp\_train,]  mhp.test <- mhp[-mhp\_train,]  # Setting up variables for Linear Regression  y <- "PRICE"  x <- c("SQUARE.FEET","BEDS","BATHS","YEAR.BUILT","Violent.Crimes")  formula <- paste(y,paste(x,collapse = "+"),sep = "~")  #building model based on the formula  ln\_mod <- lm(formula,data=mhp.train)  #summary ln\_mod  summary(ln\_mod)  coefficients(ln\_mod)  #Setting plot parameters  par(mfrow=c(2,2))  #plotting ln\_mod  plot(ln\_mod)  #Log likelihood of linear reg model  logLik(ln\_mod)  #AIC of linear regression model  AIC(ln\_mod)  # building Polynomial model of Degree 2  poly2\_mod <- lm(mhp.train$PRICE ~ poly(mhp.train$SQUARE.FEET, degree=2, raw=TRUE) +  poly(mhp.train$BEDS, degree=2, raw=TRUE) +  poly(mhp.train$BATHS, degree=2, raw=TRUE) +  poly(mhp.train$YEAR.BUILT, degree=2, raw=TRUE) +  poly(mhp.train$Violent.Crimes, degree=2, raw=TRUE))  #summary poly2\_mod  summary(poly2\_mod)  #Log likelihood of polynomial regression model of degree 2  logLik(poly2\_mod)  #AIC of polynomial regression model of degree 2  AIC(poly2\_mod)  #plotting poly2\_mod  plot(poly2\_mod,pch=18,col="green")  # building Polynomial model of Degree 3  poly3\_mod <- lm(mhp.train$PRICE ~ poly(mhp.train$SQUARE.FEET, degree=3, raw=TRUE) +  poly(mhp.train$BEDS, degree=3, raw=TRUE) +  poly(mhp.train$BATHS, degree=3, raw=TRUE) +  poly(mhp.train$YEAR.BUILT, degree=3, raw=TRUE) +  poly(mhp.train$Violent.Crimes, degree=3, raw=TRUE))  #summary poly3\_mod  summary(poly3\_mod)  #Log likelihood of polynomial regression model of degree 3  logLik(poly3\_mod)  #AIC of polynomial regression model of degree 3  AIC(poly3\_mod)  #plotting poly3\_mod  plot(poly3\_mod,pch=18,col="blue") |