Technical paper: Predicting House Sales
By: Sarita Patel

Abstract:

In this paper, we explore the key components that affect the price of houses in the King County, Washington state housing market. Nearly 22,000 observations with 20 features on houses were selected for the study. Various multivariate methods were applied to the data to find statistical relationships between the variables that affect price. A linear regression model was applied to predict the housing prices as well as to find the most important features which results in fetching a good price for the houses. The model determined that the highest correlated variable is living space in square footage of a house. If sq ft living space increases by 20%, the price of the house increases by 6%. We also applied Principal Component Analysis and found that the number of floors, bedrooms, bathrooms are all important features in predicting the price. Using an exploratory factor analysis, we tried to test the underlying theory that the price of a house is greatly influenced by the size of the living room. Factor analysis found that the square footage of the basement, number of bathrooms, bedrooms, and floors all play a crucial role for predicting house price. Cluster analysis was used to determine if there were groups of houses with similar characteristics that could be used to differentiate distinct price ranges. LDA was applied to look at the different categories of houses. From LDA we can see that there are several categories among houses such as (bad, ok, good, awesome and remarkable). Having an idea about the current condition of the house the buyer/seller can make good decisions about their investment plans. As per LDA analysis we found that the houses which were built in year 1930, had old construction design and small area overall fall under bad category and for sure needed renovation in future. So, this categorization can be helpful user to predict the correct amount for the house.

Introduction:

Data analytics can be used to benefit various groups of people by breaking down a dataset to gain insight into what the data is telling us. In terms of business, data analytics can be used to help protect and empower consumers of various goods and services. For the purpose of this study, we will focus on the real estate market. In the United States, the real estate market is a large facet of the economy and on average, nearly 6 million houses are sold every year. In this research, house price data is analyzed to help consumers buy or sell a house. Selling a house can be stressful and knowing that one is selling their house at an appropriate price can give consumers confidence to sell. Conversely for buyers, knowing one is paying a fair price can alleviate the stress associated with not knowing if a house is appropriately priced. The House Sales in King County, USA dataset contains data on nearly 22,000 houses and contains 20 descriptive features. The data was collected between May 2014 and May 2015. King County is a county that contains Seattle, Washington, a large metropolitan city in the United States. Since this is a small region within the US, it should not be inferred that the results from this research can be generalized to the entire US. However, the methods and techniques used can be a platform for other researchers to begin further investigations.

Literature review:

We find some, but relatively few relevant studies on performing Principal component analysis, Latent Discriminant Analysis and Cluster Analysis. Gordana Ivosev, Lyle Burton, and Ron Bonner [1] explained about the application of Principal Component Analysis on Megavariate Data Set where the applied Pareto Scaling to scale the data and performed Principal Component Analysis on this scaled data and visualised the loading and scores using two- and three-dimensional plots. P.Anupriya and S.Karpagavalli [2] researched about Topic models which conceive latent topics in text using hidden random variables, and discovered the structure with posterior inference. A dataset with 200 abstracts which fall under four topics are collected from two different domain journals for tagging journal abstracts and Then the document models were built using LDA (Latent Dirichlet Allocation) with Bayes and Gibbs sampling which is used to extract appropriate tags for abstracts. Sepulveda, et al. [3] researched how the understanding of delirium and dementia has evolved. They use cluster analysis to compare dementia tests throughout the years to determine which test would prove the most accurate with current patients.

Data and Methodology:

This study utilizes data that serves as a representation of metropolitan cities in the United States of America. We are trying to predict the price of the data based on 20 independent features. The dataset is a mix of numerical and categorical variables. Number of bedrooms, Number of bathrooms, where .5 accounts for a room with a toilet but no shower, Number of floors, whether the house has a waterfront view, basement, a grade index from 1 to 13, where 1-5 falls short of building construction and design, 6-10 has an average level of construction and design, and 11-13 have a high quality level of construction and design. The measurements of the houses like square footage of the living space, land space, above and below basement, and square footage of 15 nearest neighbors are provided as well. Variables specific to the location such as zip code, Latitude and longitude are also given. The dataset utilized has no null values.

A simple linear regression model was applied on the data to create a model which predicts the price of a house using 20 independent features to understand which features are strongly associated in determining the price of the house. Additionally, a new column was created to calculate the age of the house from the year built and date posted columns. This column turned out to be significant in predicting the sale price of the house. As linear regression will accept only numeric columns, categorical variables were treated with the categorical encoding method. All the numerical features were not normal and were skewed, hence log transformation was performed on those variables.

Principal component analysis was performed on the dataset to identify the most important features for predicting price. Before Principal Component Analysis was performed, multicollinearity was checked. Square footage of living space is strongly correlated with square footage of living space in 2015, and square footage above the ground is highly correlated with the number of bathrooms. Square footage of living space was removed from the PCA analysis. Also, square footage of lot space is strongly correlated with square footage of Lot space in 2015, therefore square footage of Lot space in 2015 was removed. After these multicollinearities were checked, performed PCA with covariance matrix, which returned the cumulative variance and proportion variance of 99.9% for first component. This is due to the standard deviation of the first component is much higher when compared with other components. Next, PCA with correlation Matrix was performed which scaled the data and returned 8 components where the 100 % cumulative variance is explained at the end of 8th component. The knee method and scree plot indicated that only two components had Eigenvalues greater than 1. This led to performing rotation using two components. A varimax rotation was utilized because the variables are independent of each other. The threshold was set to 0.6. After rotating the components, the cumulative variance of 60.3% at the end of the second component. Exploratory factor analysis was also applied to the dataset utilizing many of the same aspects as PCA. For the factor analysis, the threshold was set to 0.4, and the variables were separated into factors to gain insight into the variable covariances resulting in 2 factors.

Cluster analysis is used to determine the types of underlying groups that exist within a dataset. For example, this analysis could be used to determine certain species of flowers have distinct flower dimensions which could differentiate one species from another. In terms of this dataset, cluster analysis was used to determine if there were underlying groups of homes based on price groups, which could be possibly differentiated by square footage of living space, square footage of lot space, or number of bedrooms.

To perform this analysis, the dataset was broken into two subsets. The first subset contained the numerical variables: price, number of bedrooms, number of bathrooms, square feet of living area, square feet of lot, number of floors, number of views, square feet above ground, square feet basement, year built, latitude and longitude of house, square feet of living space in 2015, and square feet of lot space in 2015. The second dataset contains all the above variables without the price, latitude and longitude of the house. Since the determination is around the price variable, it is not included in the second clustering process, but used to compare the two subsets of data. Additionally, since some of the variables were not normal, all variables were scaled. This allowed for further analysis.

To determine the best number of clusters, the average silhouette width of clusters on a range of clusters from one to ten cluster. The average silhouette calculation is the average distance between clusters. The best number of clusters will have the largest average silhouette value, meaning the clusters are spread apart. The silhouette graph for both data subsets showed the same number of optimal clusters: two.

Next, the k-medoids method was performed on both datasets. The k-medoids method determines the cluster by finding a medoid or middle point and minimizes the distance between medoids and other points of the cluster. Additionally, it maximizes the distance between other medoids, meaning that the different clusters are more distinct the larger the distance.

Linear discriminant analysis (LDA) helps to identify the accuracy of the model from configuration of the independent variables that can help determine the value of the dependent variable. LDA discovers a grouping pattern among different categories of houses and test the accuracy of model that was created. The purpose of running LDA on the kc_house_price dataset was to identify the accuracy of the model from configuration of the IV's that can help us predict the value of the dependent variable. One factor field was created called "Category", which was based on the condition field which is ordinal variable from the dataset. So here the values are changing where condition = (1,2,3,4,5) to category = ("Bad", "OK", "Good", "Awesome", "Remarkable"). The count of houses in different categories differ in the dataset such as the count for "bad" are 23 and count for “good” are 10211. Which indicates that the dataset is unbalanced. Before applying the LDA on dataset, it was split into training and testing set following a 75%- 25% ratio. The train set had 15719 observations and test set had 5894 observations. After the multicollinearity check removed several variables (sqft_living, sqft_lot, condition, grade, sqft_above) then performed the further LDA analysis to avoid misclassification of the data.

Discussions and Results:

To check for multicollinearity simple correlation analysis is done between the dependent and independent variables, it revealed that square footage of the living area in the house, grade of the house, number of bathrooms have high positive correlation with the sale price of a house. Some intercorrelation among the independent variables was also clearly visible hence checked for VIF and removed highly intercorrelated variables. A standard multiple regression was run on cleaned and transformed data. The entire model was significant with f statistic 5.16e+03 and p value less than 0.01. As the p-value is much less than 0.05, we reject the null hypothesis that β = 0. Hence there is a significant relationship between the variables in the linear regression model. The model was able to explain 77% of the variation in the data. From the coefficients from the linear model, Living Square footage of a house increases by 20% then the selling price of that house increases by 6%. All other factors remaining constant an average house would sell at $540,000 and with a 10% increase in total square footage, the remodeled house will sell at $572,4000. Similarly, for every unit increase in the grade of the house it results in 18% increase in sale price. How the other variables affect the sales price of the house can be seen at Appendix-(1). To test for overfitting and to score new data as they come in real time, the available data was split into 70% training on which the model was developed and tested with 30% remaining data. The model was able to explain 69% of the variability in the new data.

First, factorability conditions where checked to know whether perform Principal component Analysis could be performed. KMO Sampling adequacy returned as 0.69 and Barlett test of sphericity which returned the p-value was less than 0.001, indicating that this number was very small. Additionally, Cronbach's Alpha which returned as 0.415. These results indicate we can perform PCA. PCA was applied with a covariance matrix as well as PCA with correlation matrix to know which explains the data better. PCA with covariance matrix returned the cumulative variance and proportion variance of 99.9% for the first component which is not good. The standard deviation of first component is much higher when compared with other components and scree plot also gives only one component. This indicates that PCA with covariance matrix should not be considered because one component is dominating other components, and the variables need to be scaled to explain the data better. After running PCA with correlation Matrix, 8 PCA components were returned, i.e. the total variance is explained at the end of the eighth component. However, the scree plot indicated that there are 2 components whose eigenvalue is greater than 1 and the Knee method indicates only 2 components could be used. 2 components were chosen because the cumulative variance after the end of the 2th component is 60.3%. Additionally, rotation was performed with 2 components with the threshold of 0.6 to get the equation of those two PCA components and after rotation the cumulative variance after 2 components is 60.3%.(See Appendix-(2))The sqft_basement variable appeared in two components, which is cross loading limitation and sqft_basement has negative effect on PC2 where as a positive effect on PC1.

PC1:Housedynamics:
0.775Bathrooms+0.758bedrooms+0.658sqft_above+0.623sqft_basement+0.735sqft_living15

For every unit increase in Bathroom PC1 increases by 0.775 units and

For every unit increase in bedrooms PC1 increases by 0.758 units

For every unit increase in sqft_above PC1 increases by 0.658 units

For every unit increase in sqft_living15 PC1 increases by 0.735 units

For every unit increase in sqft_basement PC1 increases by 0.623 units

PC2: Condition of House:0.811floors-0.604sqft_basement+0.714yr_built

For every unit increase in floors PC2 increases by 0.811 units

For every unit increase in sqft_basement PC2 decreases by 0.604 units

For every unit increase in yr_built PC2 increases by 0.714 units

Exploratory Factor Analysis was used to try and reduce the number of variables. relationships were established between the variables and grouping them into components. The dependent variable is the price and the remaining features are independent variables. Conducting the factor analysis with threshold 0.4 shows that Factor1 explains the variance best, and factor2 explains the second best. Factor1 contains the following variables: bedrooms, bathrooms, floors, sqft_above, yr_built, sqft_living15. Factor 2 Is comprised of sqft_basement. In total using our two factors we are able to explain 52.4% of the variance in our data (See Appendix-(3))

After completing data cleaning and using a silhouette plot to determine the appropriate number of clusters, the k-medoids algorithm was performed on both subsets of the data. First, the k-medoids technique was applied to the second cluster. The average silhouette width is .23, which shows that the two clusters are not very distinct. From the graph that shows the clusters, it is apparent that the two clusters overlap and border each other closely. Additionally, the clusters are unbalanced; the first cluster has 12276 data points while the second has only 9337points. The medoid for the first cluster has a price of $440,000, 3 bedrooms, 1.75 bathrooms, was built in 1947, has a basement sized at 440 square feet, and had a living area square footage of 1740 in 2015. The medoid for the second cluster has a price of 619420, 4 bedrooms, 2.75 bathrooms, was built in 1988, no basement (basement square footage of 0), and a living area square footage of 2330 in 2015 (See Appendix-(5) for cluster visualization).

Next, k-medoids was applied to the second dataset. The average width of the silhouette width is .28, again showing that the two clusters are not very distinct. Similar to the first dataset, the clusters are unbalanced. The first cluster contains 12436 data points while the second has only 9177 points. The medoid for the first cluster was priced at $365,000, had 3 bedrooms, 1.5 bathrooms, was built in 1958, had a basement of 350 square feet, and a living area square footage of 1640. The second cluster had a medoid with house price of $485,000, 4 bedrooms, 2.5 bathrooms, was built in 1995, no basement (basement square footage of 0), and living area was measured at 2390 square feet in 2015.

The cluster analysis results showed that the clusters performed very poorly. The average silhouette width was below .3. Typically, those closest to 1.00 are the best performing. Additionally, the lack of balance between the cluster size is another reason to give pause. It is interesting that the medoids seem similar in some aspects, such as lack of basement, and the year difference between both pairs of medoids. However, the initial research question seems unanswered through cluster analysis.

As LDA deals with categorical variable, with multiple levels and try to avoid misclassification. The purpose of running LDA on the kc_house_price dataset is to identify the accuracy of the model from configuration of the IV's that can help us predict the value of the dependent variable. We ran LDA on the dataset and found that the prior probability shows that the count of the houses in bad & OK category are less than 1%, Good category houses are 64% and houses in Awesome & Remarkable category are less than 27%. The Group from the LDA run shows that houses which had a small number of bedrooms, built in year 1930 and have old construction are in bad category and secondly, houses with bigger sq_ft living and built in 1948 are in OK category and then houses that are built in 1957 and had bigger sq_ft living and greater number of bedroom and bathroom falls in Good category and finally the houses that had magnificent architecture with some waterfronts falls under awesome and remarkable category.

These groups that are formed after running LDA also gives a fair idea about the Price points such as bad category house price range is around 329273, OK category house price range is around 359673, Good category house price range is around 541887 and lastly awesome and remarkable category house price range is around 500000 to 640000(See Appendix-(4)).

Basically, the dataset shows that the better the condition of the overall house the chances of it to fall in good, awesome and remarkable category is more and can result in higher cost. And this analysis with the appropriate price estimate and category split can help assist buyer/seller to look and understand the condition of the house and make an appropriate investment.

As we ran the prediction model on the train dataset on the test dataset (small subset of the original dataset). The accuracy for the model to classify the different categories in 72% with LDA (See Appendix-(4)). To validate this, we performed cross validation and accuracy of the LDA seems improved 71% with cross validation technique (See Appendix-(4)). We Can see the overall accuracy is quite good as small size of test set can justify the overall accuracy in prediction.

LDA Equation: -4.63*price - 7.61*bedrooms - 1.37*bathrooms + 4.35*floor - 2.83*waterfront - 5.74*view - 4.51*sqft_basement + 3.57*yr_built + 9.82*yr_renovated + 6.41*zip_code + 2.6* sqft_living15 - 1.1*sqft_lot15

Future work:

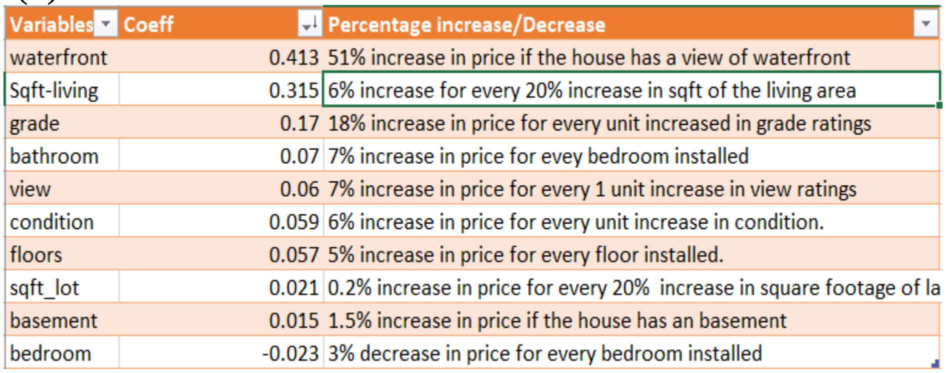
Future work could be expanded to Four different projects. First, the dataset could have expanded features. It could include features such as type of heating and cooling, possible taxes, and distances to public amenities such as public schools, community centers and public parks. This type of data could make a house more or attractive to possible buyers. Next, it would be interesting to research the first listing price of the house and expand the dataset to contain other houses that were on the market at the same time but did not sell. This would help future researchers better understand what houses have, and possibly what buyers are looking for while shopping for a home.A third possible project would entail collecting data on all houses which sell during a decade. Looking at data from a particular market over an expanded period would allow researchers to track houses that sell multiple times throughout a decade, which could indicate turbulent economic times. Additionally, it could indicate a fluctuation in a county's housing value. Finally, we can apply a canonical correlation analysis on the dataset as it has many continuous variables.

Conclusion:

The King County dataset carries multiple independent variables which are good predictors in estimating the selling price of a house. Additionally, the results from different regression and clustering techniques can help individuals in making confident decisions for buying and selling the property. This helps identifying the important features to maintain in a house, which can result in good investment. While house prices could be viewed as important to real estate agents or consumers, others such as architects, builders, or bankers may find this analysis of importance. When working on designing, building, or financing a new house, it may be important to understand what the house will be valued at, or the possible selling price of the house.

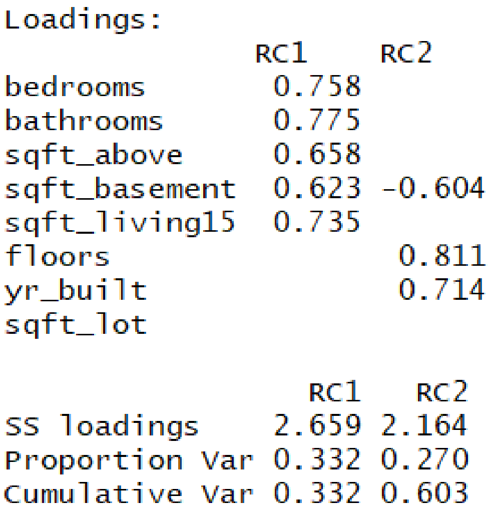
Appendix:

-(1)



Variables	Coeff	Percentage increase/Decrease
waterfront	0.413	51% increase in price if the house has a view of waterfront
Sqft-living	0.315	6% increase for every 20% increase in sqft of the living area
grade	0.17	18% increase in price for every unit increased in grade ratings
bathroom	0.07	7% increase in price for evey bedroom installed
view	0.06	7% increase in price for every 1 unit increase in view ratings
condition	0.059	6% increase in price for every unit increase in condition.
floors	0.057	5% increase in price for every floor installed.
sqft_lot	0.021	0.2% increase in price for every 20% increase in square footage of la
basement	0.015	1.5% increase in price if the house has an basement
bedroom	-0.023	3% decrease in price for every bedroom installed

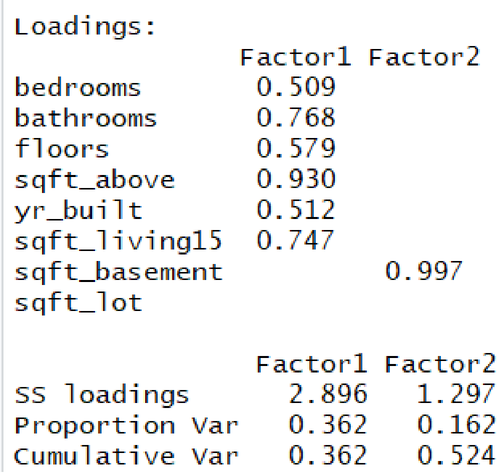
-(2)



```
Loadings:
              RC1    RC2   
bedrooms       0.758       
bathrooms      0.775       
sqft_above     0.658       
sqft_basement  0.623 -0.604
sqft_living15  0.735       
floors                0.811
yr_built              0.714
sqft_lot                   

                 RC1   RC2
SS loadings    2.659 2.164
Proportion Var 0.332 0.270
Cumulative Var 0.332 0.603
```

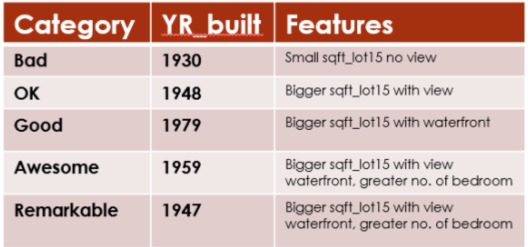
-(3)



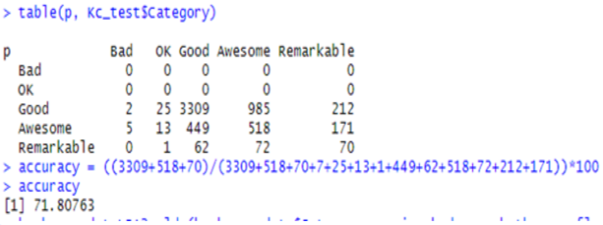
```
Loadings:
              Factor1 Factor2
bedrooms       0.509         
bathrooms      0.768         
floors         0.579         
sqft_above     0.930         
yr_built       0.512         
sqft_living15  0.747         
sqft_basement          0.997 
sqft_lot                     

               Factor1 Factor2
SS loadings      2.896   1.297
Proportion Var   0.362   0.162
Cumulative Var   0.362   0.524
```

-(4)

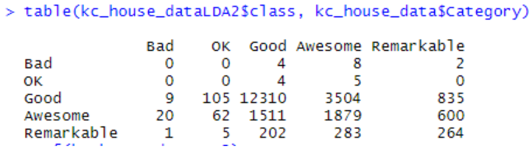


Category	YR_built	Features
Bad	1930	Small sqft_lot15 no view
OK	1948	Bigger sqft_lot15 with view
Good	1979	Bigger sqft_lot15 with waterfront
Awesome	1959	Bigger sqft_lot15 with view waterfront, greater no. of bedroom
Remarkable	1947	Bigger sqft_lot15 with view waterfront, greater no. of bedroom



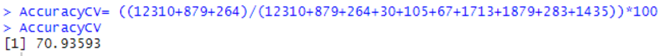
```
> table(p, Kc_test$Category)

p              Bad   OK Good Awesome Remarkable
  Bad            0    0    0       0          0
  OK             0    0    0       0          0
  Good           2   25 3309     985        212
  Awesome        5   13  449     518        171
  Remarkable     0    1   62      72         70
> accuracy = ((3309+518+70)/(3309+518+70+7+25+13+1+449+62+518+72+212+171))*100
> accuracy
[1] 71.80763
```



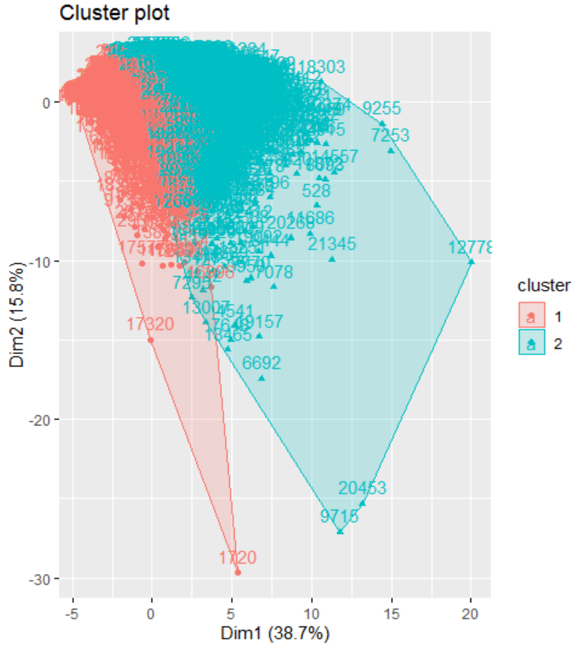
```
> table(kc_house_dataLDA2$class, kc_house_data$Category)

              Bad    OK  Good Awesome Remarkable
  Bad           0     0     4       8          2
  OK            0     0     4       5          0
  Good          9   105 12310    3504        835
  Awesome      20    62  1511    1879        600
  Remarkable    1     5   202     283        264
```



```
> AccuracyCV= ((12310+879+264)/(12310+879+264+30+105+67+1713+1879+283+1435))*100
> AccuracyCV
[1] 70.93593
```

-(5)



Code:

Code for linear regression in Python and R:

```
import numpy as np
import pandas as pd
data = pd.read_csv(r'C:\Users\\Mani\Downloads\kc_house_data.csv')
data.head(3)
data['price'].describe()
data = data.drop(['id'],axis=1)
data.head(3)
correlation = data.corr(method='pearson')
columns = correlation.nlargest(15, 'price').index
columns
correlation_map = np.corrcoef(data[columns].values.T)
```

```
sns.set(font_scale=0.7)
heatmap = sns.heatmap(correlation_map, cbar=True, annot=True, square=True, fmt='.2f',
yticklabels=columns.values, xticklabels=columns.values)

plt.show()
data['date'] = pd.to_datetime(data['date'])
data = data.drop(['yr_built','date'],axis=1)
data.head(3)
data['price'] = np.log(data['price'])
data['sqft_living'] = np.log(data['sqft_living'])
data['sqft_lot'] = np.log(data['sqft_lot'])
data['sqft_living15'] = np.log(data['sqft_living15'])
data['sqft_lot15'] = np.log(data['sqft_lot15'])
data.head()
data = data.drop(['yr_renovated'],axis=1)
data.head(3)
data['basement'] = data['sqft_basement'].apply(lambda yr: 0 if yr == 0 else 1)
data = data.drop(['sqft_basement','sqft_above'],axis=1)
data.head()
hh<- read.csv(file="kp.csv", header=TRUE, sep=",")
#Creating Automatic Models
hh$zipcode=NULL
hh$long =NULL
null = lm(price ~ 1, data=hh)
null
full = lm(price ~ ., data=hh)
full

#Forward Regression
train_Forward = step(null, scope = list(lower=null, upper=full), direction="forward")
summary(train_Forward)

#Backward Regression
train_Backward = step(full, direction="backward")
summary(train_Backward)

#Stepwise Regression
train_Step = step(null, scope = list(upper=full), direction="both")
summary(train_Step)

model1 <- lm(price ~ ., data=hh)
summary(model1)
library(DescTools)
#Check VIF
VIF(model1)
```

Code for LDA:

```
library(MASS)
head(kc_house_data)
is.factor(kc_house_data$condition)
is.numeric(kc_house_data$condition)
kc_house_data$Category <- factor(kc_house_data$condition, labels = c("Bad", "OK", "Good", "Awesome", "Remarkable"))

install.packages("caTools")
require(caTools)
set.seed(123)
kc_house_data_sample = sample.split(kc_house_data,SplitRatio = 0.75) # splits the data in the ratio mentioned in SplitRatio. After splitting marks these rows as logical TRUE and the remaining are marked as logical FALSE
Kc_train =subset(kc_house_data,kc_house_data_sample ==TRUE) # creates a training dataset named train1 with rows which are marked as TRUE
Kc_test=subset(kc_house_data, kc_house_data_sample==FALSE)
head(Kc_train)
head(Kc_test)

install.packages("corrplot")
kc_house_data_num <- kc_house_data[,c(3:21)]
M<-cor(kc_house_data_num)
library(corrplot)
corrplot(M, method="number")

# The dependent variable must be categorical
kc_house_dataLDA=lda(Kc_train$Category~price+bedrooms+bathrooms+floors+waterfront+view+sqft_basement+yr_built+yr_renovated+zipcode+lat+long+sqft_living15+sqft_lot15, data = Kc_train)
kc_house_dataLDA
table (Kc_train$Category)
table (Kc_test$Category)
plot(kc_house_dataLDA)

kc_house_data_validation1 <- kc_val[,c(3:10,12:22)]
head(kc_house_data_validation1)
# Try to predict the class from the original data
# Note ... this is JUST a test to see how this works
# In practice you will want to use cross-validation!
p = predict(kc_house_dataLDA,newdata=Kc_test)$class
p
# Compare the results of the prediction
table(p,kc_house_data_validation1$Category)
accuracy = ((3309+518+70)/(3309+518+70+7+25+13+1+449+62+518+72+212+171))*100
accuracy
# Setting "CV = T" will have the lda function perform
# "Leave-one-out" cross-validation
```

```
kc_house_dataLDA2=lda(kc_house_data$Category~price+bedrooms+bathrooms+floors+waterfront+vie
w+sqft_basement+yr_built+yr_renovated+zipcode+lat+long+sqft_living15+sqft_lot15, data =
kc_house_data, CV=T)
kc_house_dataLDA2
table(kc_house_dataLDA2$class, kc_house_data$Category)
AccuracyCV= ((12310+879+264)/(12310+879+264+30+105+67+1713+1879+283+1435))*100
AccuracyCV
```

Code for Cluster Analysis:

```
library(cluster)
library(factoextra)
house_data <- read.csv("kc_house_data.csv")
summary(house_data)
data1 = house_data[,c(3,4,5,6,7,8,10,13,14,15,18,19,20,21)]
summary(data1)
data2 = house_data[,c(4,5,6,7,8,10,13,14,15,20,21)]
summary(data2)
num_data_scale1 <- scale(data1)
num_data_scale2 <- scale(data2)

#calculating clusters and printing time to calculate
ptm <- proc.time()
fviz_nbclust(num_data_scale1, cluster::pam,k.max = 10, method = "silhouette", verbose = True)
proc.time() - ptm

ptm <- proc.time()
fviz_nbclust(num_data_scale2, cluster::pam,k.max = 10, method = "silhouette", verbose = True)
proc.time() - ptm

#data1 run
data1_kmed_2 <- pam(num_data_scale1, k=2, diss=F)
data1_kmed_2
#visualization
fviz_cluster(data1_kmed_2)

#info on medoids
data1_kmed_2$medoids

#rows of medoids are 16348 and 3502
medoid1_1 = house_data[16348,]
medoid1_1
medoid2_1 = house_data[3502,]
medoid2_1

#info on sill width (this is an important value)
data1_kmed_2$silinfo$avg.width
#since this is .28, this is very low, meaning bad clustering
data1_kmed_2$clustering  # printing the "clustering vector"
```

```
data1_kmed_2$silinfo$avg.width
data1_kmed_2_clust <- lapply(1:2, function(nc)
row.names(house_data)[data1_kmed_2$clustering==nc])
data1_kmed_2_clust

#counting sizes of the clusters
lengths(data1_kmed_2_clust)
#unbalanced dataset counts of clusters
data2_kmed_2 <- pam(num_data_scale2, k=2, diss=F)
data2_kmed_2
#visualization
fviz_cluster(data2_kmed_2)

#info on medoids
data2_kmed_2$medoids

#rows of medoids are 1708 and 10806
medoid1_2 = house_data[1708,]
medoid1_2
medoid2_2 = house_data[10806,]
medoid2_2

#info on sil width
data2_kmed_2$silinfo$avg.width
#since this is .23, this is very low, meaning bad clustering
data2_kmed_2$clustering  # printing the "clustering vector"
data2_kmed_2$silinfo$avg.width
data2_kmed_2_clust <- lapply(1:2, function(nc)
row.names(house_data)[data2_kmed_2$clustering==nc])
data2_kmed_2_clust

#counting sizes of the clusters
lengths(data2_kmed_2_clust)
```

Code for Principal Component Analysis and Factor Analysis:

```
#Libraries
library(Hmisc) #Describe Function
library(psych) #Multiple Functions for Statistics and Multivariate Analysis
library(GGally) #ggpairs Function
library(ggplot2) #ggplot2 Functions
library(vioplot) #Violin Plot Function
library(corrplot) #Plot Correlations
library(REdaS) #Bartlett's Test of Sphericity
library(psych) #PCA/FA functions
library(factoextra) #PCA Visualizations
library("FactoMineR") #PCA functions
library(ade4) #PCA Visualizations
#Set Working Directory
```

```
setwd('C:/Users/smartron/Desktop')
#Read in Datasets
Housedata <- read.csv(file="kc_house_data.csv", header=TRUE, sep=",")
Housedata
#Check for Missing Values (i.e. NAs)
#For All Variables
sum(is.na(Housedata))
#no total missing values
#Show Structure of Dataset
str(Housedata, list.len=ncol(Housedata))
#Show column Numbers
names(Housedata)
#Check Sample Size and Number of Variables
dim(Housedata)
head(Housedata)
names(Housedata)
Housedata1 <- Housedata[,c(4:8,13:15,20:21)]
names(Housedata1)
names(Housedata1)
head(Housedata1)
names(Housedata1)
#Check for Multicollinearity with Correlations no multicollinearity.
M<-cor(Housedata1, method="pearson")
M
names(Housedata1)
Housedata3 <- Housedata1[,c(1:2,4:9)]
names(Housedata3)
M<-cor(Housedata3, method="pearson")
M
options("scipen"=100, "digits"=5)
round(cor(Housedata3), 2)
MCorrTest = corr.test(Housedata3, adjust="none")
MCorrTest
M = MCorrTest$p
M
# Now, for each element, see if it is < .01 (or whatever significance) and set the entry to
# true = significant or false
MTest = ifelse(M < .01, T, F)
MTest
# Now lets see how many significant correlations there are for each variable.  We can do
# this by summing the columns of the matrix
colSums(MTest) - 1  # We have to subtract 1 for the diagonal elements (self-correlation)
#Test KMO Sampling Adequacy
library(psych)
KMO(Housedata3)
#Overall MSA =  0.69
#Test Bartlett's Test of Sphericity
library(REdaS)
```

```
bart_spher(Housedata3)
#p-value < 2.22e-16 (Very Small Number)
#Test for Reliability Analysis using Cronbach's Alpha
library(psych)
alpha(Housedata3,check.keys=TRUE)
#raw_alpha = 0.015
#Create PCA
p = prcomp(Housedata3)
#Check Scree Plot
plot(p)
abline(1, 0)
#Check PCA Summary Information
summary(p)
print(p)
p = prcomp(Housedata3, center=T, scale=T)
#Check Scree Plot
plot(p)
abline(1, 0)
summary(p)
print(p)
p2 = psych::principal(Housedata3, rotate="varimax", nfactors=2, scores=TRUE)
p2
print(p2$loadings, cutoff=.6, sort=T)
#PCAs Other Available Information
ls(p2)
p2$values
p2$communality
p2$rot.mat
#Conducting Factor Analysis
fit = factanal(Housedata3, 2)
print(fit$loadings, cutoff=.4, sort=T)
summary(fit)
```

References:

1. Ivosev, G., Burton, L., & Bonner, R. (2008). Dimensionality Reduction and Visualization in Principal Component Analysis. Analytical Chemistry, 80(13), 4933–4944. doi: 10.1021/ac800110w

2. Anupriya, P. (1), & Karpagavalli, S. (2). (n.d.). LDA based topic modeling of journal abstracts. ICACCS 2015 - Proceedings of the 2nd International Conference on Advanced Computing and Communication Systems. https://doi.org/10.1109/ICACCS.2015.7324058

3. Sepulveda, E., Franco, J. G., Trzepacz, P. T., Gaviria, A. M., Meagher, D. J., Palma, J., … de Pablo, J. (2016). Delirium diagnosis defined by cluster analysis of symptoms versus diagnosis by DSM and ICD criteria: diagnostic accuracy study. BMC Public Health, 16(1), 1–14. https://doi.org/10.1186/s12888-016-0878-6