# Final Project

**Course: Stat 3690**

**Name: Nirmit Patel**

**Student Number: 7793925**

**Abstract:** The title of the dataset is “Real estate valuation data set”. This dataset contains 414 observations. Each observation contains 7 attributes (6 explanatory variables and 1 response variable) of a house property. The purpose of the study is to predict the price of the houses. The problem of interest here is: How accurately can we predict the price of the house from the given features of the house? This problem will be answered using Principal component analysis and Principal component regression.

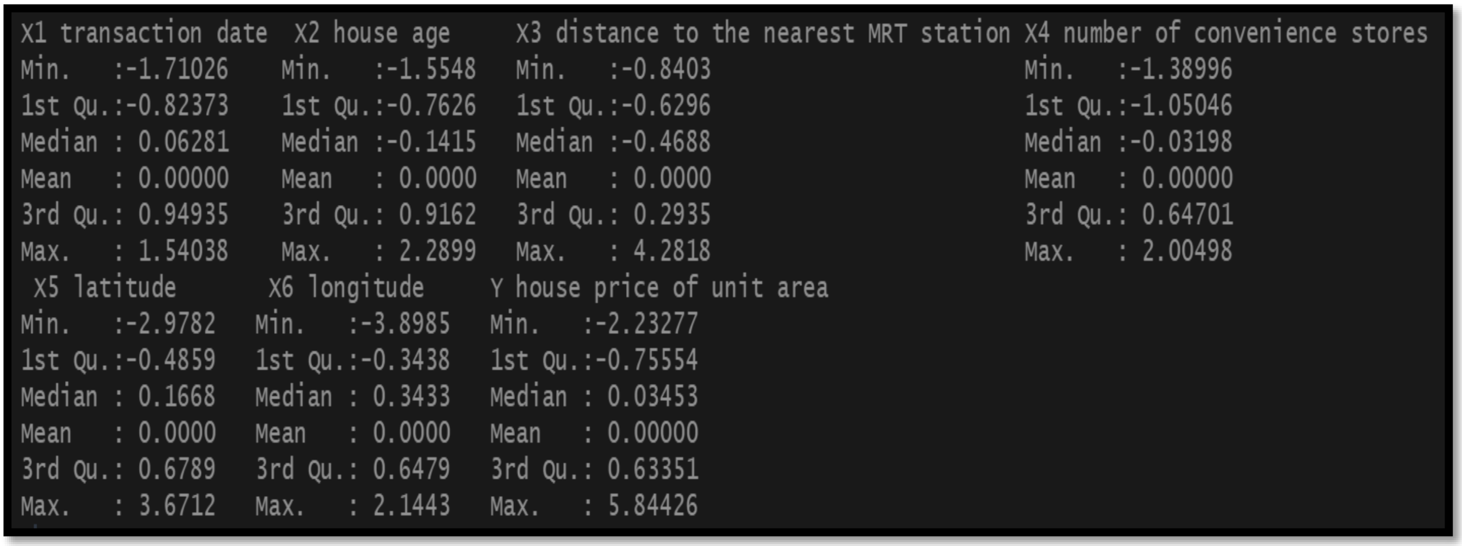
**Introduction:** The real estate industry is one of the biggest and one of the fastest growing industries in the world. Which makes it very interesting to analyze the dataset “Real estate valuation data set”, which is collected from Sindian Dist, New Taipei City, Taiwan. The dataset was taken from the UCI Machine Learning Repository. The dataset contains 7 attributes which are: The transaction date of the house (for example, 2013.250 = 2013 march), age of the house (unit: year), the distance to the nearest MRT station (unit: meter), the number of convenience stores in the living circle on foot (integer), the geographic coordinate latitude (unit: degree), the geographic coordinate longitude (unit: degree), and the price of the house per unit area (10000 New Taiwan Dollar/Ping, where Ping is a local unit, 1 Ping = 3.3 meter squared). In the real estate industry, it can be very helpful to be able to roughly predict the price of a property given information about certain features of the property, which leads us to the question of interest: How successfully can the price of the houses in New Taipei City, Taiwan., be predicted from the 6 features given in the dataset?

**Methods:** There are two methods which will be used for the analysis of this dataset, Principal component analysis and Principal component regression.

1. Principal component analysis: A dimension reduction method, used to reduce large set of variables into a smaller one without losing much information about the dataset. If there are p variables in a dataset, principal component analysis will produce p principal components which will account for total system variability. Often most of this variability can be accounted by q principal component, where q < p. For instance, for the “Real estate valuation data set” there are 6 explanatory variables, Principal component analysis will produce 6 principal components out of which the first 2 or 3 might contain most of the variability of the dataset. The only assumption before performing principal component analysis is linearity, that is, the dataset is a linear combination of the variables. The objective of using principal component analysis for this dataset, is to find the principal components which will be used to perform principal component regression.
2. Principal component regression: It is a regression technique similar to multiple linear regression, but it uses the principal components instead of the actual variables to carry out the regression. The reason for using principal component is to prevent over-fitting of the model, to reduce the noise in the dataset and to prevent multicollinearity. Since first few principal components carry most of the information about the dataset, there is no need to fit the full model using multivariate regression. There are few assumptions of performing regression which are linearity, no multicollinearity, homoscedasticity, and residuals should be approximately normally distributed. Principal component regression will be used to predict the price of the houses. Comparing the predicted values with the actual values (residuals) will help us answer the research question.

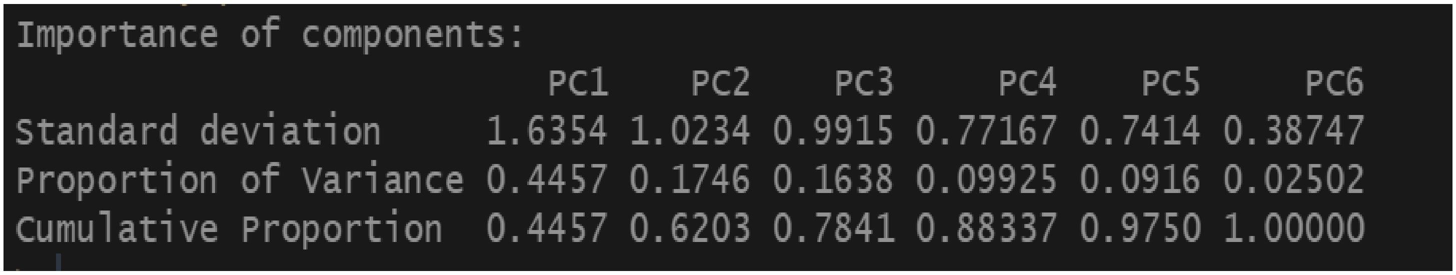
**Data Analysis/Results:**

- Summary of the dataset after scaling (each variable is standardized by subtracting its mean and dividing the result by its standard deviation):



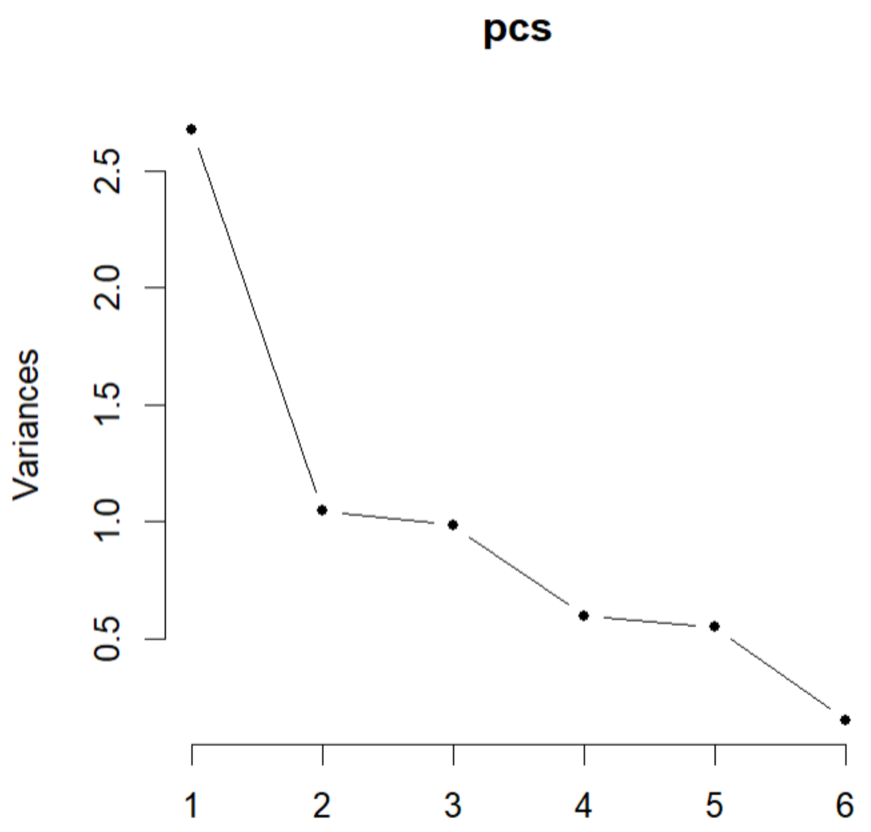
```
X1 transaction date  X2 house age     X3 distance to the nearest MRT station X4 number of convenience stores
Min.   :-1.71026    Min.   :-1.5548   Min.   :-0.8403                        Min.   :-1.38996
1st Qu.:-0.82373    1st Qu.:-0.7626   1st Qu.:-0.6296                        1st Qu.:-1.05046
Median : 0.06281    Median :-0.1415   Median :-0.4688                        Median :-0.03198
Mean   : 0.00000    Mean   : 0.0000   Mean   : 0.0000                        Mean   : 0.00000
3rd Qu.: 0.94935    3rd Qu.: 0.9162   3rd Qu.: 0.2935                        3rd Qu.: 0.64701
Max.   : 1.54038    Max.   : 2.2899   Max.   : 4.2818                        Max.   : 2.00498
 X5 latitude       X6 longitude     Y house price of unit area
Min.   :-2.9782   Min.   :-3.8985   Min.   :-2.23277
1st Qu.:-0.4859   1st Qu.:-0.3438   1st Qu.:-0.75554
Median : 0.1668   Median : 0.3433   Median : 0.03453
Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.00000
3rd Qu.: 0.6789   3rd Qu.: 0.6479   3rd Qu.: 0.63351
Max.   : 3.6712   Max.   : 2.1443   Max.   : 5.84426
```

- Principal Component Analysis:

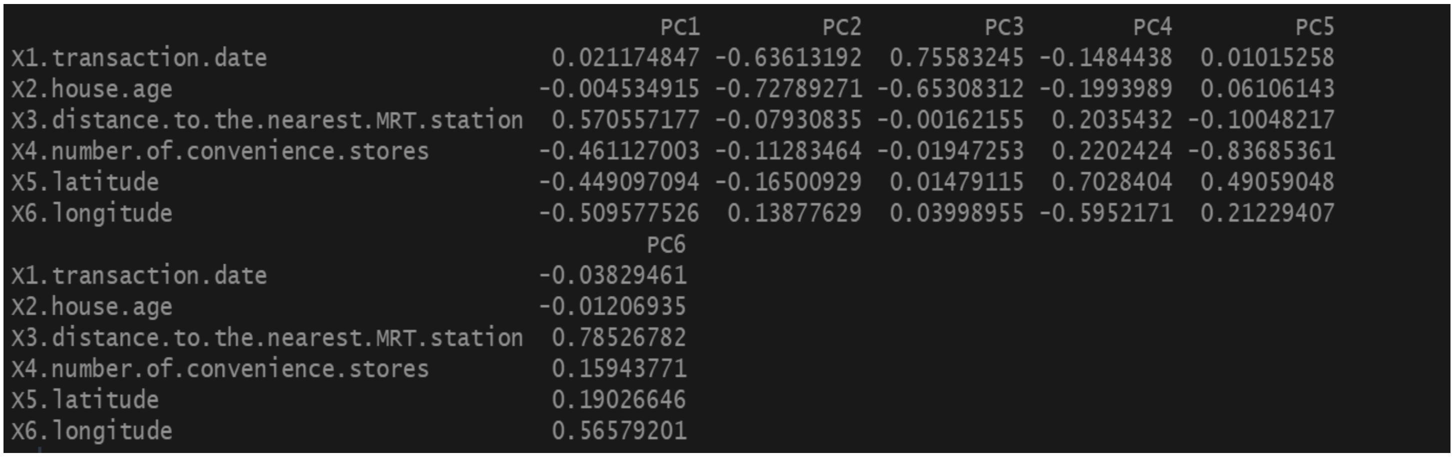


```
Importance of components:
                          PC1    PC2    PC3     PC4    PC5     PC6
Standard deviation     1.6354 1.0234 0.9915 0.77167 0.7414 0.38747
Proportion of Variance 0.4457 0.1746 0.1638 0.09925 0.0916 0.02502
Cumulative Proportion  0.4457 0.6203 0.7841 0.88337 0.9750 1.00000
```

Here we can see proportion of variance explained by each principal component and also the cumulative proportion explained by each component. These components are ordered from highest variability to lowest variability. We can see the first 4 principal components explain 88% of the variation in the data. Thus, using the first 4 principal components for the regression will be feasible.



This is a scree plot of the Principal Components that tell us how much variation each principal component captures from the data. By looking at this plot we can say the first 4 principal components capture the most variation from the data.



```
                                               PC1         PC2         PC3        PC4         PC5
X1.transaction.date                    0.021174847 -0.63613192  0.75583245 -0.1484438  0.01015258
X2.house.age                          -0.004534915 -0.72789271 -0.65308312 -0.1993989  0.06106143
X3.distance.to.the.nearest.MRT.station  0.570557177 -0.07930835 -0.00162155  0.2035432 -0.10048217
X4.number.of.convenience.stores       -0.461127003 -0.11283464 -0.01947253  0.2202424 -0.83685361
X5.latitude                           -0.449097094 -0.16500929  0.01479115  0.7028404  0.49059048
X6.longitude                          -0.509577526  0.13877629  0.03998955 -0.5952171  0.21229407
                                              PC6
X1.transaction.date                   -0.03829461
X2.house.age                          -0.01206935
X3.distance.to.the.nearest.MRT.station  0.78526782
X4.number.of.convenience.stores        0.15943771
X5.latitude                            0.19026646
X6.longitude                           0.56579201
```

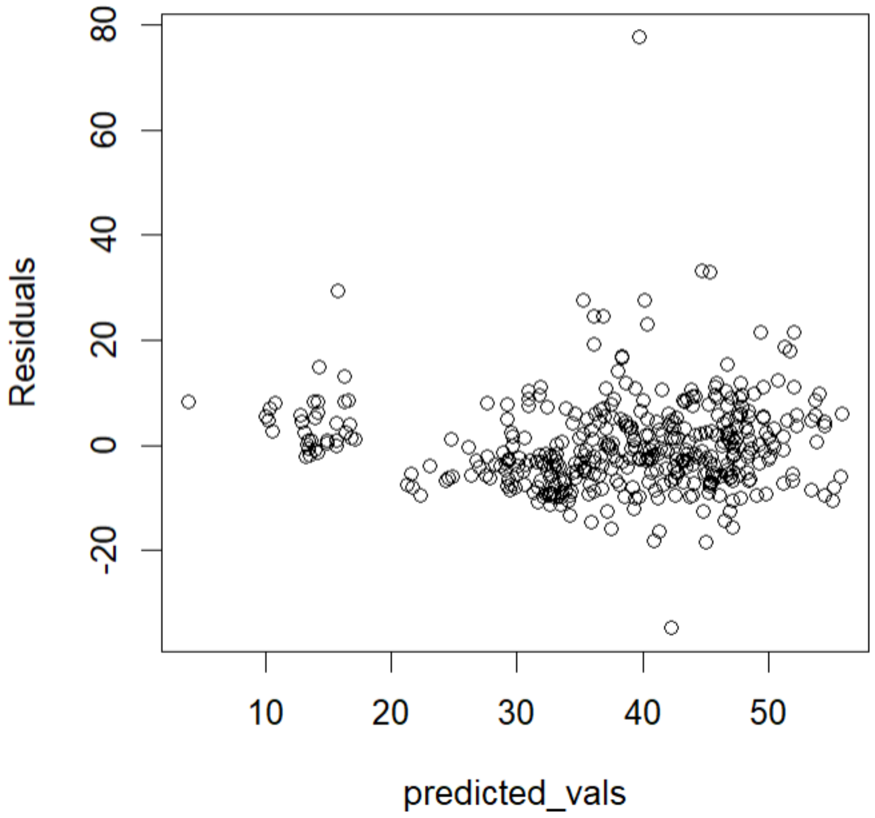
Here we have the loadings of the PCA, loadings tell us how much each variable contributes to each principal component, since each principal component is just a linear combination of the original variables. For instance, looking at the loadings we can tell the first loading vector places approximately equal weight on variables X3, X4, X5 and X6.

- Principal Component Regression:

Before performing the principal component regression, lets check if the assumptions of multicollinearity, homoscedasticity and normality of residuals are met:

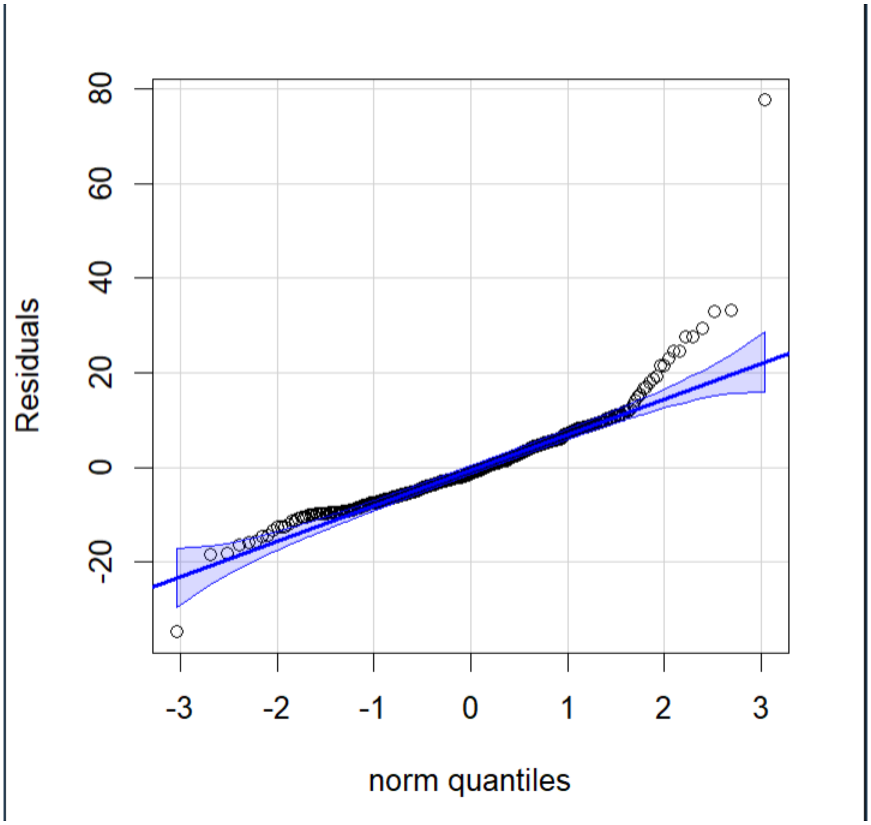
Multicollinearity: Since Principal Component Analysis removes any multicollinearity from the data, we can conclude assumption of multicollinearity is met.

Homoscedasticity:



Here we have the residuals on the y-axis and the predicted values on the x-axis. Since the residuals are approximately constant as the predicted values change, we can say the assumption of homoscedasticity is met.

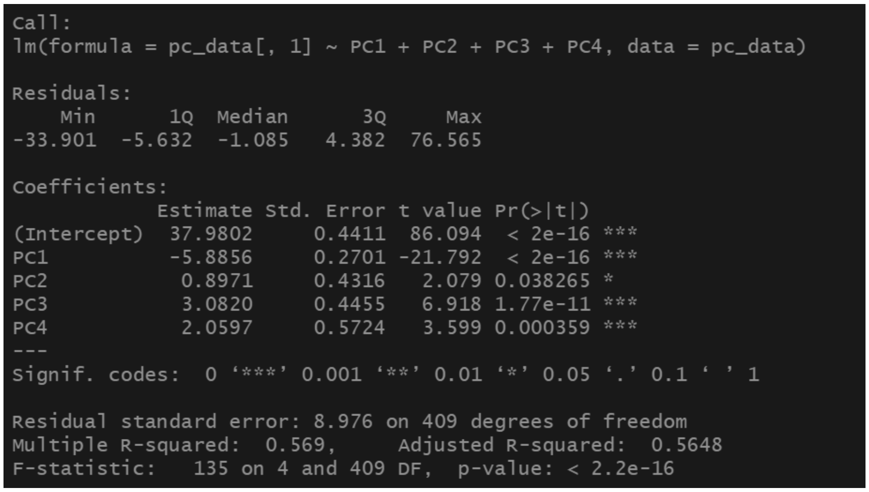
Normality of the Residuals:



From this qqplot of the residuals we can tell the residuals follow approximately normal distribution. Thus, the assumption of normality is met.

Since all the assumptions seem to be met, we can carry on with the regression analysis. We will be fitting the model using the first 4 Principal components.

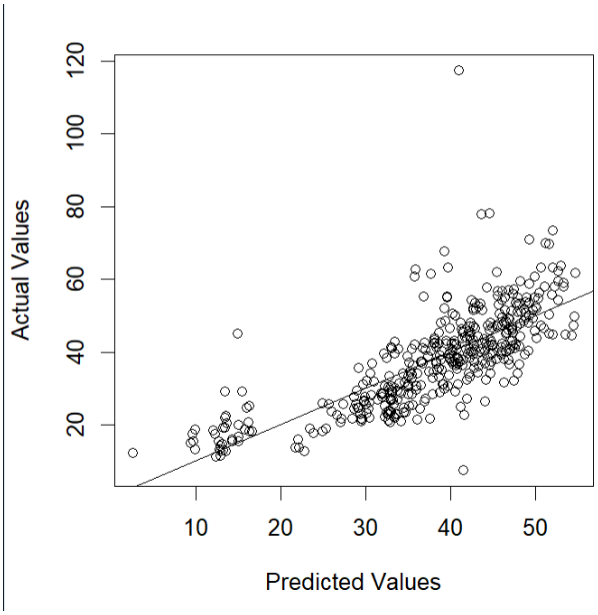
Output of Principal Component Regression:



```
Call:
lm(formula = pc_data[, 1] ~ PC1 + PC2 + PC3 + PC4, data = pc_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-33.901  -5.632  -1.085   4.382  76.565 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  37.9802     0.4411  86.094  < 2e-16 ***
PC1          -5.8856     0.2701 -21.792  < 2e-16 ***
PC2           0.8971     0.4316   2.079 0.038265 *  
PC3           3.0820     0.4455   6.918 1.77e-11 ***
PC4           2.0597     0.5724   3.599 0.000359 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.976 on 409 degrees of freedom
Multiple R-squared:  0.569,	Adjusted R-squared:  0.5648 
F-statistic:   135 on 4 and 409 DF,  p-value: < 2.2e-16
```

From the coefficients we get the equation of the predicted values to be:

E(Y) = -5.8856(PC1) + 0.8971(PC2) + 3.0820(PC3) + 2.0597(PC4) + 37.9802. The value of R^2 is around 0.56 which means the model is not a great fit but is a decent fit.



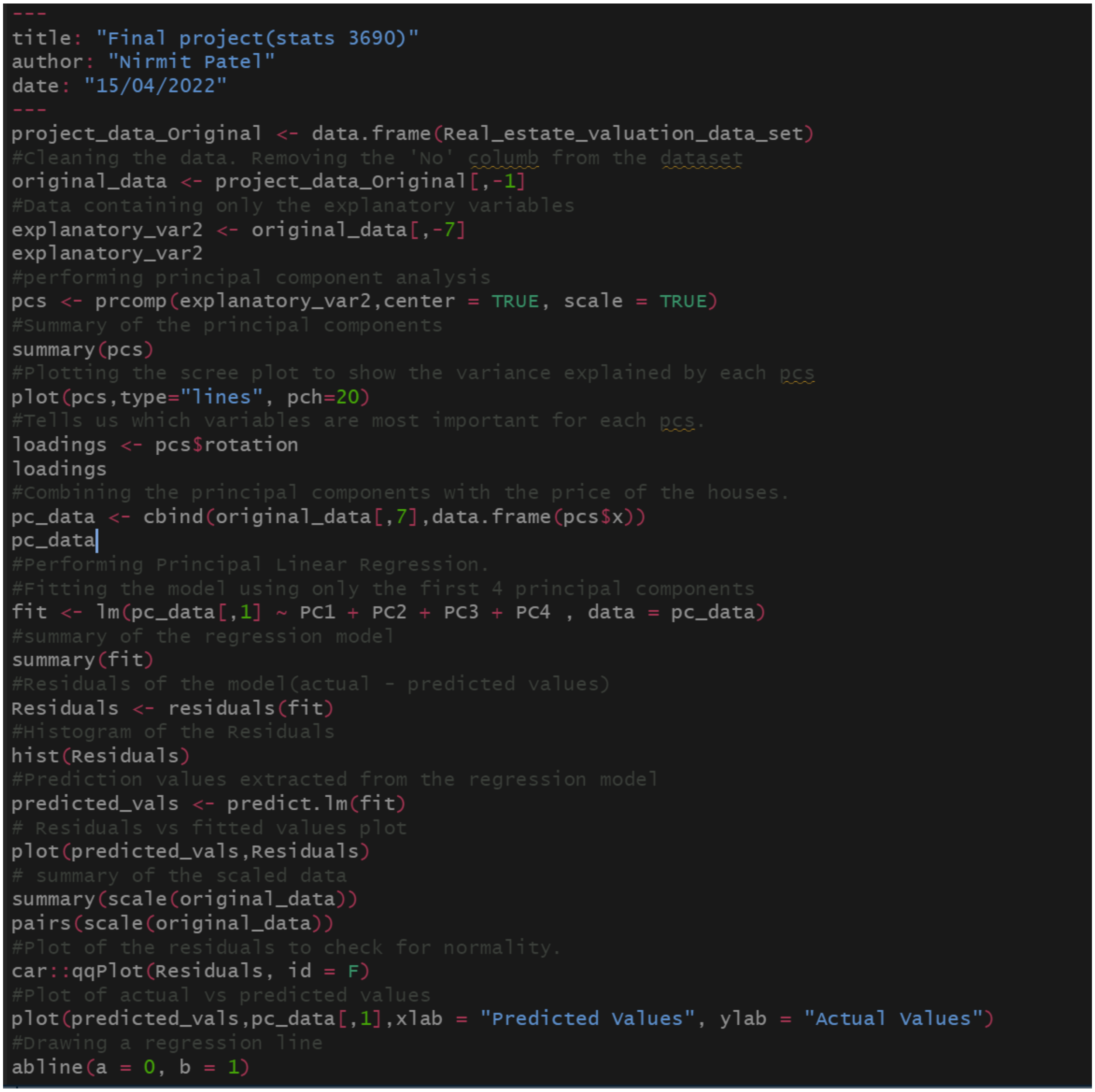
This is a plot of the actual values and the predicted values. It can be said from the plot, that the regression model doesn’t do a great job of predicting the price of the houses, it does a decent job.

**Research Question Answer:** Since the regression model isn't a great fit, but a decent fit. The regression analysis can roughly predict the price of the house, but the predictions are not very reliable. One way to improve this model might be to include more Principal Components in the model or maybe more information about the house property is needed. It can be concluded that we can get a decent prediction of the price of the houses in New Taipei City, Taiwan., given the 6 features of the house, which are the transaction date, house age, distance to the nearest MRT station, number of convenience stores in the area, latitude and longitude of the house.

**Conclusion:** The objective of this analysis was to predict the price of the houses in New Taipei City, Taiwan., given information about certain features of the houses. Using Principal Component Analysis, we were able to reduce the dimension of the dataset without losing much information about the dataset. Principal Component Regression was carried out using the first 4 Principal Components which contained about 89% of the variation in the data. From the regression analysis, we were able to do a decent job of predicting the price of the house property, using more principal components or having more information about the property could have improved the regression model resulting in better predictions. Being able to predict property prices given some information about the property is very beneficial in the real estate industry, having a rough idea about a price of a property can prevent us from overspending on a property.

**Appendix**:

R Code:



```
---
title: "Final project(stats 3690)"
author: "Nirmit Patel"
date: "15/04/2022"
---
project_data_Original <- data.frame(Real_estate_valuation_data_set)
#Cleaning the data. Removing the 'No' columb from the dataset
original_data <- project_data_Original[,-1]
#Data containing only the explanatory variables
explanatory_var2 <- original_data[,-7]
explanatory_var2
#performing principal component analysis
pcs <- prcomp(explanatory_var2,center = TRUE, scale = TRUE)
#Summary of the principal components
summary(pcs)
#Plotting the scree plot to show the variance explained by each pcs
plot(pcs,type="lines", pch=20)
#Tells us which variables are most important for each pcs.
loadings <- pcs$rotation
loadings
#Combining the principal components with the price of the houses.
pc_data <- cbind(original_data[,7],data.frame(pcs$x))
pc_data
#Performing Principal Linear Regression.
#Fitting the model using only the first 4 principal components
fit <- lm(pc_data[,1] ~ PC1 + PC2 + PC3 + PC4 , data = pc_data)
#summary of the regression model
summary(fit)
#Residuals of the model(actual - predicted values)
Residuals <- residuals(fit)
#Histogram of the Residuals
hist(Residuals)
#Prediction values extracted from the regression model
predicted_vals <- predict.lm(fit)
# Residuals vs fitted values plot
plot(predicted_vals,Residuals)
# summary of the scaled data
summary(scale(original_data))
pairs(scale(original_data))
#Plot of the residuals to check for normality.
car::qqPlot(Residuals, id = F)
#Plot of actual vs predicted values
plot(predicted_vals,pc_data[,1],xlab = "Predicted Values", ylab = "Actual Values")
#Drawing a regression line
abline(a = 0, b = 1)
```

## Bibliography:

http://archive.ics.uci.edu/ml/datasets/Real+estate+valuation+data+set