**A Closer Look: Analysis of the Real Estate Markets of San Francisco, Philadelphia and Dallas**

**Introduction**

The focus of this project is an analysis of the housing market of three cities: San Francisco, Philadelphia and Dallas. The housing market in the United States has been a concern for many since the crash of 2008. However, the housing market has regained all value lost in the crises. According to research by Zillow the total housing market value has never been worth more than it is right now. Zillow estimates the U.S housing stock grew to a total of $29.6 trillion in 2016. A massive market complemented by a heavy focus on analysis conveys a great deal of attention being placed on leading indicators that seem to affect the housing market and the economy. Our research emphasizes the significance of a select few of these indicators, in which led our team to model a forecast of the average price per square feet of all three of these cities. In this project, we consider both the linear model and ARIMA model as we believe it is of use to look at more than one model in predicting house prices. Ultimately, we decided to use the ARIMA model to forecast the data, as well as using various tests to take a closer look at how our chosen indicators effect the property prices within these select cities.

**Dataset**

We collected our data through government data stores which contain details of macroeconomic factors within San Francisco, Philadelphia and Dallas. In terms of our response variable, we gained this dataset through the Zillow datastore. We took data ranging from 2012 to 2017 in a monthly format. Throughout our research we have identified six leading indicators that we strongly believe greatly effect housing prices. Table 1.1 highlights these variables as well as the response variable.

|  |  |
| --- | --- |
| **Independent Variables** | **Dependent Variables** |
| Consumer Price Index (Monthly)  Average Age on Market (Monthly)  Unemployment Rate (Monthly)  Population Growth (Monthly)  Mortgage Rate (Monthly)  Housing Price Index (Monthly) | Price Per Square Feet (Monthly) |

1.1

**References of dataset**

https://www.zillow.com/research/data/

https://datasf.org/

<http://metadata.phila.gov/#home/datasetdetails/59a87e1295e23b332961a8e8/>

<https://fred.stlouisfed.org/source?soid=89>

**Part I**

**Regression Methods – Linear Regression San Francisco dataset**

lm(formula = Price.Per.Sqft ~ CPI + Average.Age.on.Market + Unemployment\_Rate +

Population\_growth + Mortgage.rate + HPI, data = dataSan)

Residuals:

Min 1Q Median 3Q Max

-36.038 -10.998 -0.246 12.569 37.111

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -3.818e+02 1.143e+02 -3.339 0.00143 \*\*

CPI 4.025e+00 4.359e-01 9.232 2.98e-13 \*\*\*

Average.Age.on.Market -1.942e-03 2.267e-01 -0.009 0.99319

Unemployment\_Rate -6.151e+00 7.333e+00 -0.839 0.40484

Population\_growth 3.676e+00 2.090e+01 0.176 0.86098

Mortgage.rate -1.553e+01 8.051e+00 -1.929 0.05825 .

HPI 6.243e-01 5.168e-01 1.208 0.23160

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.31 on 62 degrees of freedom

Multiple R-squared: 0.9788, Adjusted R-squared: 0.9768

F-statistic: 477.1 on 6 and 62 DF, p-value: < 2.2e-16

The output above highlights a multiple regression model we performed on the data. A clearer representation: y=β0+β1x1+β2x2+β3x3+β4x4+ β5+ β6+e. x1=CPI, x2=Average Age on Market, x3=Unemployment rate, x4=population growth, x5=mortgage rate, x6=HPI. The summary conveys the significance of each indicator variable. With a high coefficient of determination of 0.9768, this model can predict a large amount of variation in the price per square feet can be predicted by the regression model. Looking at the p – values it seems only CPI seems to be significant. Next, we performed backward stepwise regression to further determine what indicator variables are significant to use for a good model fit.

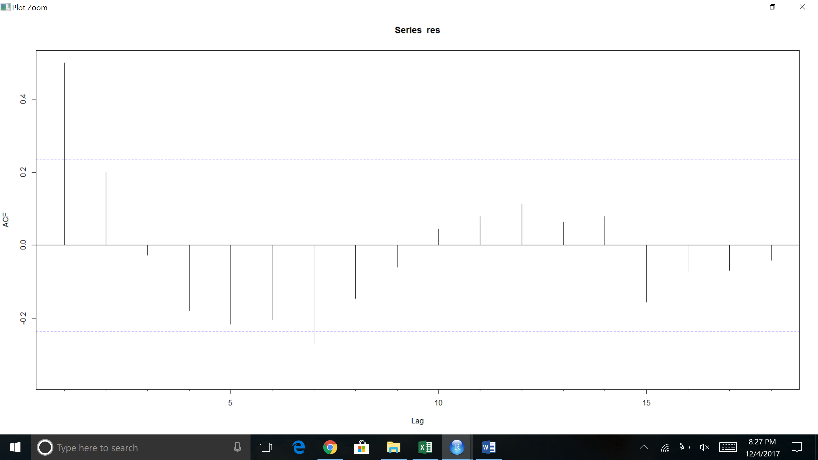
lm(formula = Price.Per.Sqft ~ CPI + Average.Age.on.Market + Unemployment\_Rate +

Mortgage.rate + HPI, data = data)

After performing stepwise regression, the following indicator variables remained:

* CPI
* Average Age on Market
* Unemployment Rate
* Mortgage Rate
* HPI

**Checking Residuals**



Durbin-Watson test

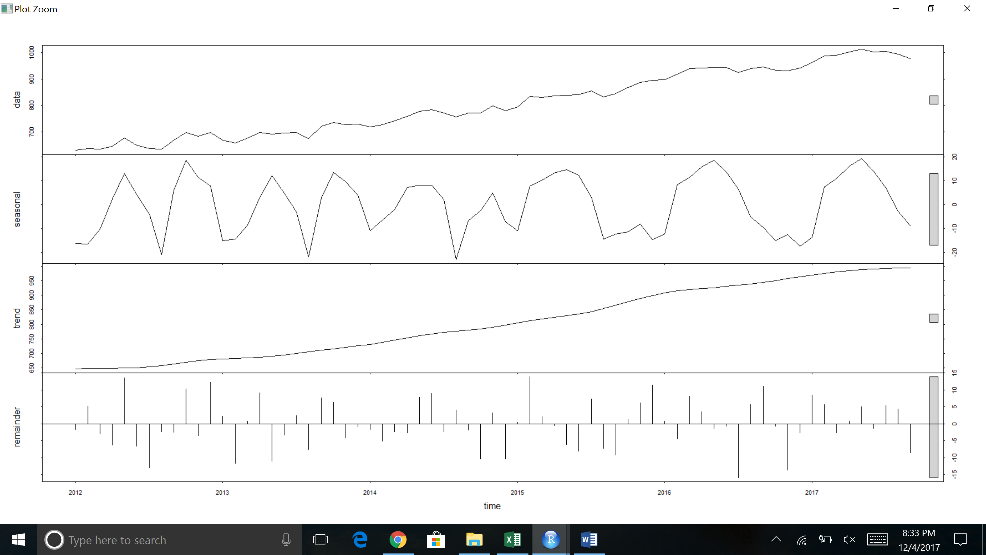
data: model1

DW = 0.93693, p-value = 4.226e-08

alternative hypothesis: true autocorrelation is not 0

Looking at the residuals we can see some spikes in the Acf which shows some correlation and more information to be extracted from the data. This is backed up by a Durbin Watson test in which also shows significant correlation in the residuals. We will leave the forecasting to the Arima section of the report.

**Arima Section – Checking for Seasonality and Trend**



There is no real concern for the impact of seasonality from looking at the stl plot. We will continue onto Arima forecasting.

**Stationarity**

Augmented Dickey-Fuller Test

data: dataSan[, 2]

Dickey-Fuller = -3.0382, Lag order = 4, p-value = 0.1532

alternative hypothesis: stationary

Augmented Dickey-Fuller Test

data: dataSan2

Dickey-Fuller = -4.2024, Lag order = 4, p-value = 0.01

alternative hypothesis: stationary

After performing the Augmented Dickey-Fuller test, we determined to difference the data once. After the first differencing the test resulted in stationarity.

**Choosing the Model**

ARIMA(2,1,2)(0,0,1)[12] with drift

Coefficients:

ar1 ar2 ma1 ma2 sma1 drift

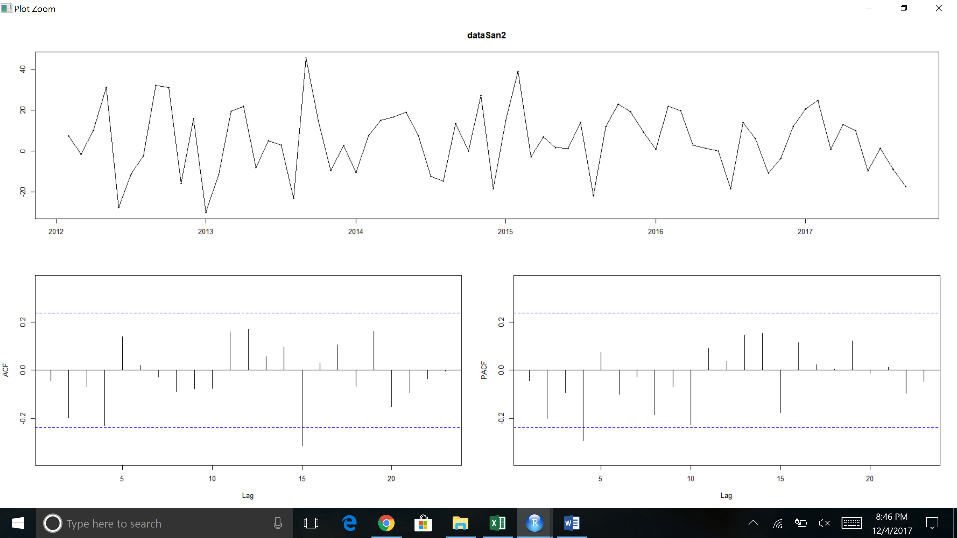
-0.2361 0.3554 0.0826 -0.8266 0.2872 5.4044

s.e. 0.2046 0.2088 0.1421 0.1444 0.1402 0.7202

sigma^2 estimated as 216.3: log likelihood=-277.1

AIC=568.2 AICc=570.07 BIC=583.74

An auto.arima coveys a model of p=2,d=1,q=2. (2,1,2). Ts display below of Acf and Pacf also can support this model. Thus, we chose to use this model.



arima(x = dataSan[, 2], order = c(2, 1, 2))

Coefficients:

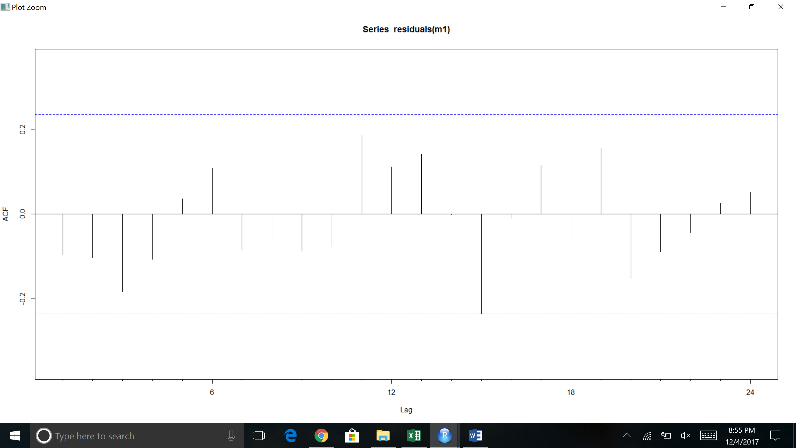
ar1 ar2 ma1 ma2

-1.7041 -0.8365 1.9263 0.9997

s.e. 0.0714 0.0841 0.0825 0.0829

sigma^2 estimated as 235.2: log likelihood = -284.21, aic = 578.42

**Residuals**

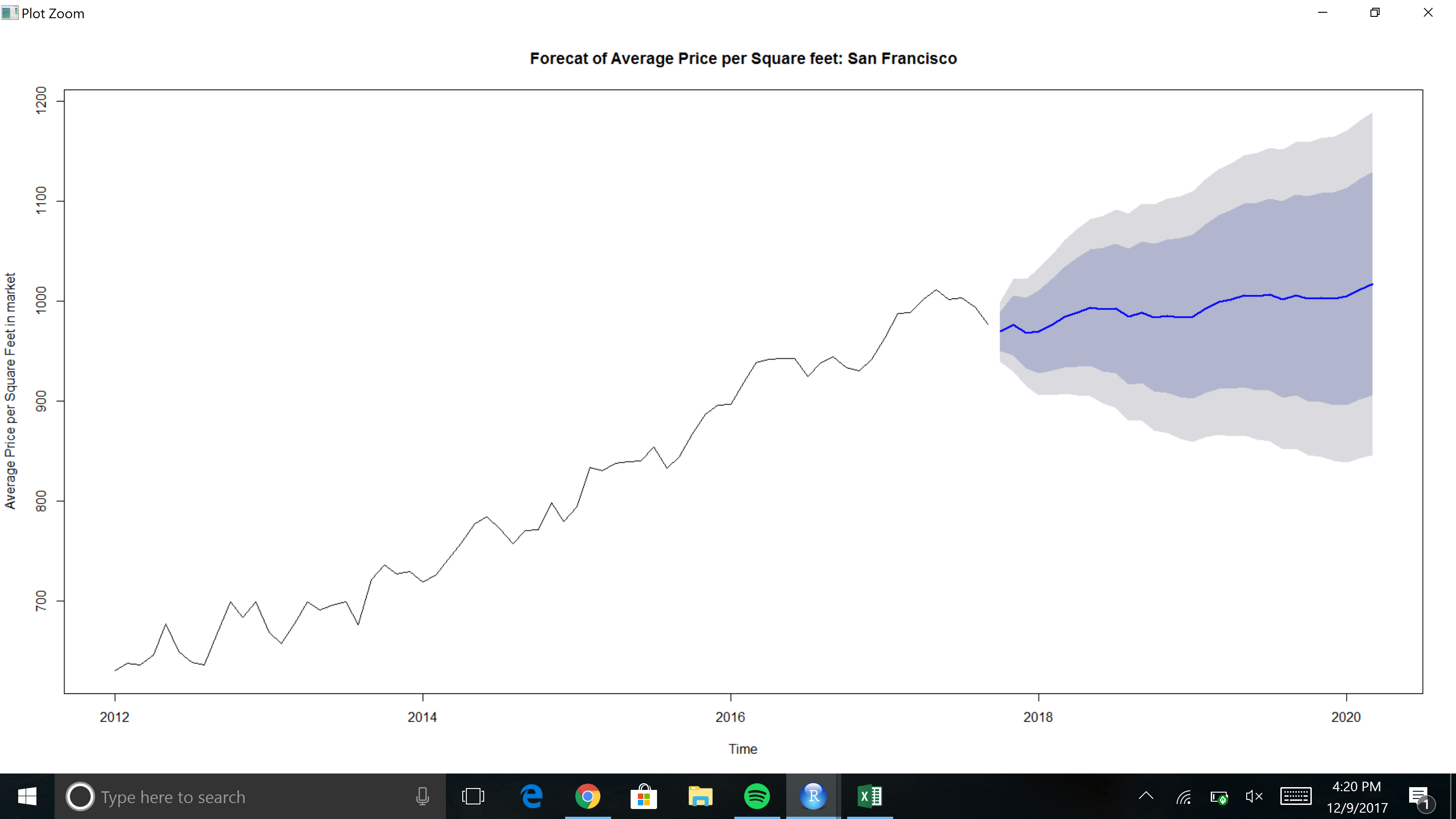


Box-Ljung test

data: residuals(m1)

X-squared = 7.7532, df = 6, p-value = 0.2567

Looking over the Acf plot of the residuals and performing a box – Ljung test indicated that this model was adequate to use in forecasting the response variable.



Above highlights our forecast for the average price per square feet of housing in the greater San Francisco area from 2017 and above (30 periods). We used a dynamic regression model with ARIMA errors. It is important to remember that the prediction intervals from the regression models (with or without ARIMA Errors) do not consider the uncertainty in the forecast predictors.

**Part II**

**Regression Methods – Linear Regression Philadelphia dataset**

lm(formula = Price.Per.Sqft ~ CPI + Average.Age.on.Market + Unemployment\_Rate +

Population\_growth + Mortgage.rate + HPI, data = dataPhilly)

Residuals:

Min 1Q Median 3Q Max

-6.490 -2.917 -0.894 2.431 12.718

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -53.58558 112.04060 -0.478 0.6341

CPI 0.21097 0.50279 0.420 0.6762

Average.Age.on.Market -0.24919 0.03813 -6.535 1.36e-08 \*\*\*

Unemployment\_Rate 2.08633 1.16768 1.787 0.0789 .

Population\_growth -11.22741 5.95781 -1.884 0.0642 .

Mortgage.rate 3.51747 1.85598 1.895 0.0627 .

HPI 0.55323 0.12695 4.358 5.03e-05 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.288 on 62 degrees of freedom

Multiple R-squared: 0.8608, Adjusted R-squared: 0.8473

F-statistic: 63.9 on 6 and 62 DF, p-value: < 2.2e-16

The output above highlights a multiple regression model we performed on the data. A clearer representation: y=β0+β1x1+β2x2+β3x3+β4x4+ β5+ β6+e. x1=CPI, x2=Average Age on Market, x3=Unemployment rate, x4=population growth, x5=mortgage rate, x6=HPI. The summary conveys the significance of each indicator variable. With a high coefficient of determination of 0.8473, this model can predict a large amount of variation in the price per square feet can be predicted by the regression model. Looking at the p – values it seems average age on market and HPI seems to be significant. Next, we performed backward stepwise regression to further determine what indicator variables are significant to use for a good model fit.

Call:

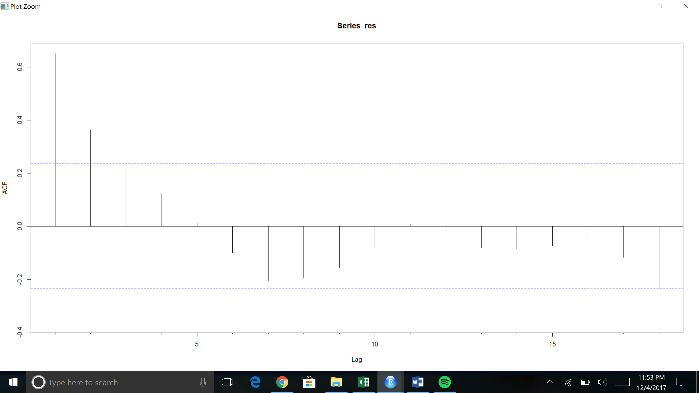
lm(formula = Price.Per.Sqft ~ CPI + Unemployment\_Rate + Mortgage.rate +

HPI, data = dataP)

After performing stepwise regression, the following indicator variables remained:

* CPI
* Unemployment Rate
* Mortgage Rate
* HPI

**Checking Residuals**

Durbin-Watson test

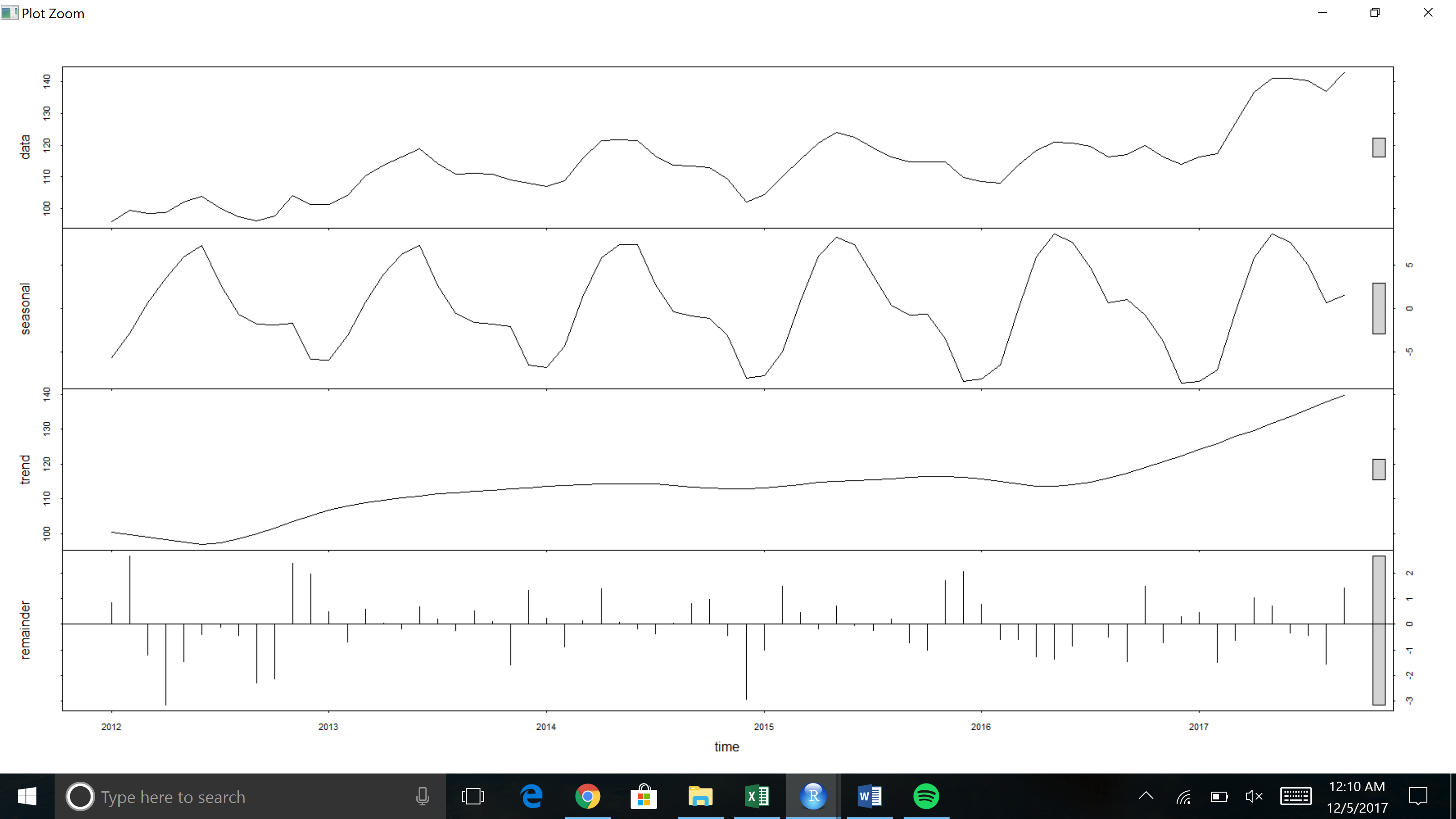
data: model1

DW = 0.53977, p-value = 1.031e-15

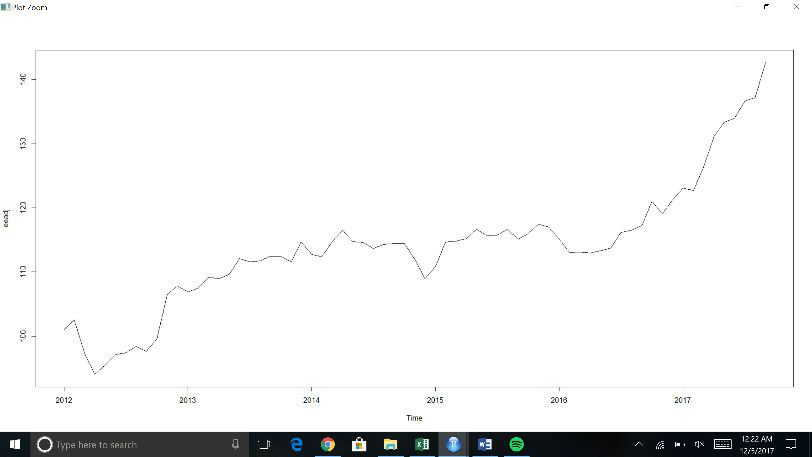
alternative hypothesis: true autocorrelation is not 0

Looking at the residuals we can see some spikes in the Acf which shows some correlation and more information to be extracted from the data. This is backed up by a Durbin Watson test in which also shows significant correlation in the residuals. We will leave the forecasting to the Arima section of the report of part II.

**Arima Section Part II – Checking for Seasonality and Trend**



There is a concern for the impact of seasonality from looking at the stl plot. To overcome this, we used seasonally adjusted data when modelling Arima. This is shown below.



Seasonally adjusted data

**Stationarity Part II**

Augmented Dickey-Fuller Test

data: dataPhilly[, 2]

Dickey-Fuller = -1.6422, Lag order = 4, p-value = 0.7206

alternative hypothesis: stationary

Augmented Dickey-Fuller Test

data: dataPhilly2

Dickey-Fuller = -3.7985, Lag order = 4, p-value = 0.02401

alternative hypothesis: stationary

After performing the Augmented Dickey-Fuller test, we determined differencing the data once. After the first differencing the test resulted in stationarity.

ARIMA(2,1,2)(0,1,1)[12]

Coefficients:

ar1 ar2 ma1 ma2 sma1

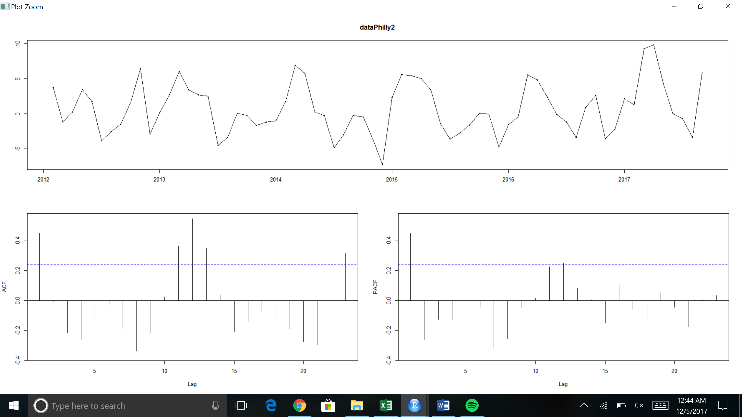
-1.1202 -0.7129 1.5278 0.9714 -0.3477

s.e. 0.1798 0.1252 0.3432 0.3282 0.1750

sigma^2 estimated as 5.812: log likelihood=-128.35

AIC=268.7 AICc=270.41 BIC=280.85

An auto.arima coveys a model of p=2,d=1,q=2. (2,1,2). Ts display below of Acf and Pacf also can support this model. Thus, we chose to use this model.



arima(x = dataPhilly[, 2], order = c(2, 1, 2))

Coefficients:

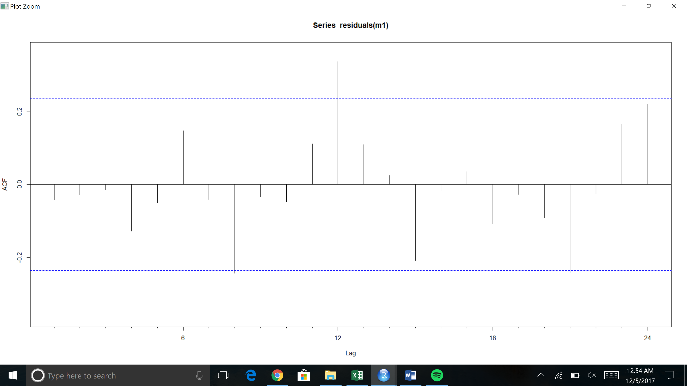
ar1 ar2 ma1 ma2

1.0503 -0.3810 -0.4569 -0.1748

s.e. 0.2552 0.2081 0.2879 0.2394

sigma^2 estimated as 9.325: log likelihood = -172.65, aic = 355.3

**Residuals**

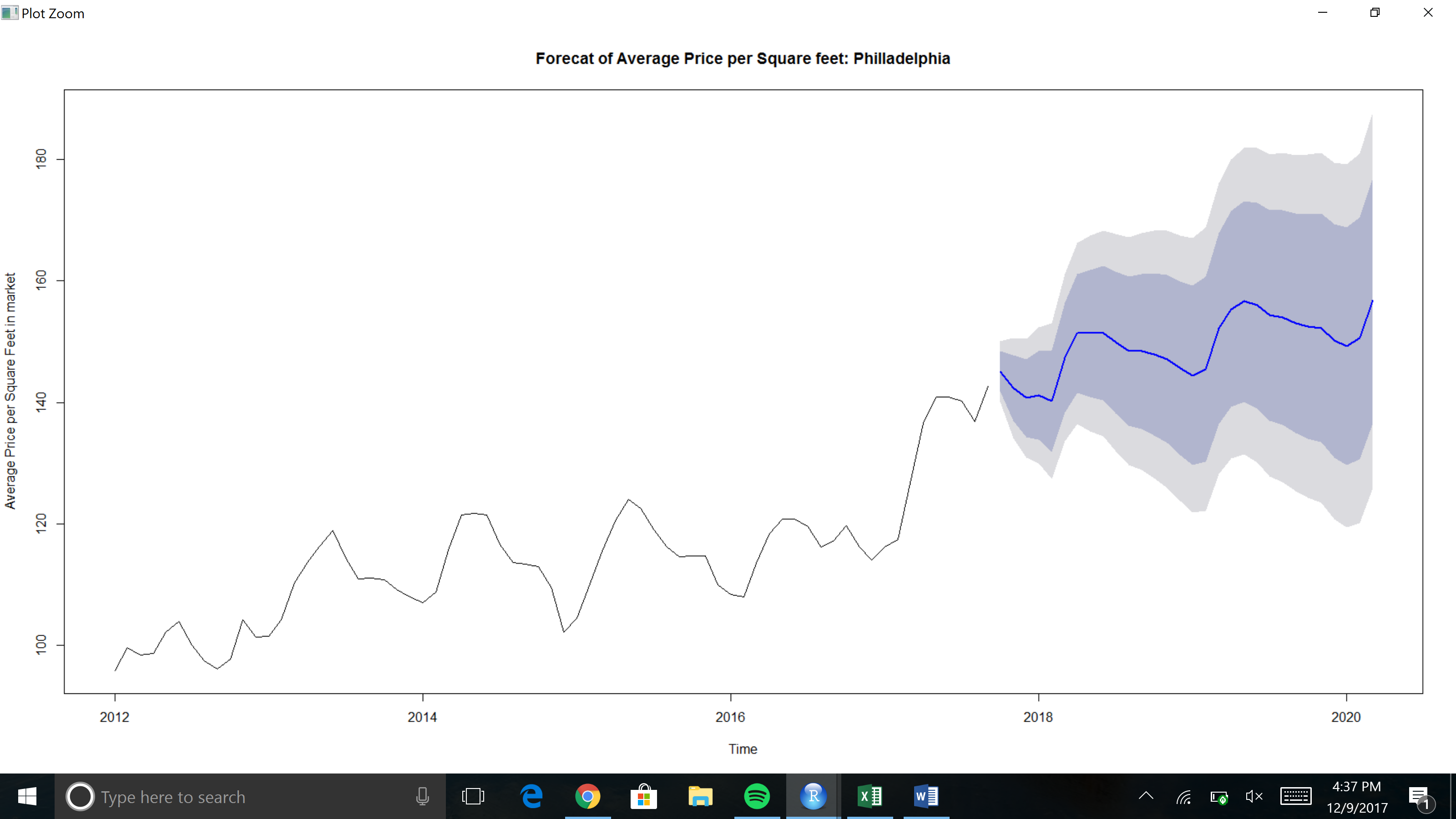


Box-Ljung test

data: residuals(m1)

X-squared = 8.4952, df = 6, p-value = 0.204

Looking over the Acf plot of the residuals there seems to be some spikes. However, performing a box – Ljung test it seems this model was adequate to use in forecasting the response variable.



Above highlights our forecast for the average price per square feet of housing in the greater Philadelphia area from 2017 and above (30 periods). We used a dynamic regression model with ARIMA errors. It is important to remember that the prediction intervals from the regression models (with or without ARIMA Errors) do not consider the uncertainty in the forecast predictors.

**Part III**

**Regression Methods – Linear Regression Dallas dataset**

lm(formula = Price.Per.Sqft ~ CPI + Average.Age.on.Market + Unemployment\_Rate +

Population\_growth + Mortgage.rate + HPI, data = dataDallas)

Residuals:

Min 1Q Median 3Q Max

-9.7840 -2.5408 0.1828 3.1915 8.2484

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -92.96793 79.23243 -1.173 0.245141

CPI 0.05235 0.42877 0.122 0.903220

Average.Age.on.Market -0.01918 0.05593 -0.343 0.732805

Unemployment\_Rate 3.70486 1.37422 2.696 0.009023 \*\*

Population\_growth -15.44227 4.17527 -3.699 0.000462 \*\*\*

Mortgage.rate -6.59551 2.17087 -3.038 0.003481 \*\*

HPI 1.80233 0.12354 14.589 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.004 on 62 degrees of freedom

Multiple R-squared: 0.979, Adjusted R-squared: 0.977

F-statistic: 482.2 on 6 and 62 DF, p-value: < 2.2e-16

The output above highlights a multiple regression model we performed on the data. A clearer representation: y=β0+β1x1+β2x2+β3x3+β4x4+ β5+ β6+e. x1=CPI, x2=Average Age on Market, x3=Unemployment rate, x4=population growth, x5=mortgage rate, x6=HPI. The summary conveys the significance of each indicator variable. With a high coefficient of determination of 0.977, this model can predict a large amount of variation in the price per square feet can be predicted by the regression model. Looking at the p – values it seems unemployment rate, population growth, mortgage rate and HPI seems to be significant. Next, we performed backward stepwise regression to further determine what indicator variables are significant to use for a good model fit.

lm(formula = Price.Per.Sqft ~ Average.Age.on.Market + Unemployment\_Rate +

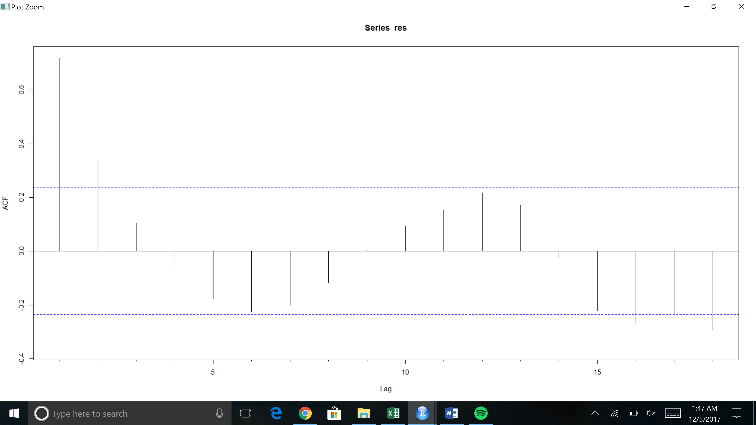
Mortgage.rate + HPI, data = dataD)

After performing stepwise regression, the following indicator variables remained:

* Average Age on Market
* Unemployment Rate
* Mortgage Rate
* HPI

**Checking Residuals**

Looking at the residuals we can see some spikes in the Acf which shows some correlation and more information to be extracted from the data. This is backed up by a Durbin Watson test in which also shows significant correlation in the residuals. We will leave the forecasting to the Arima section of the report of part III.

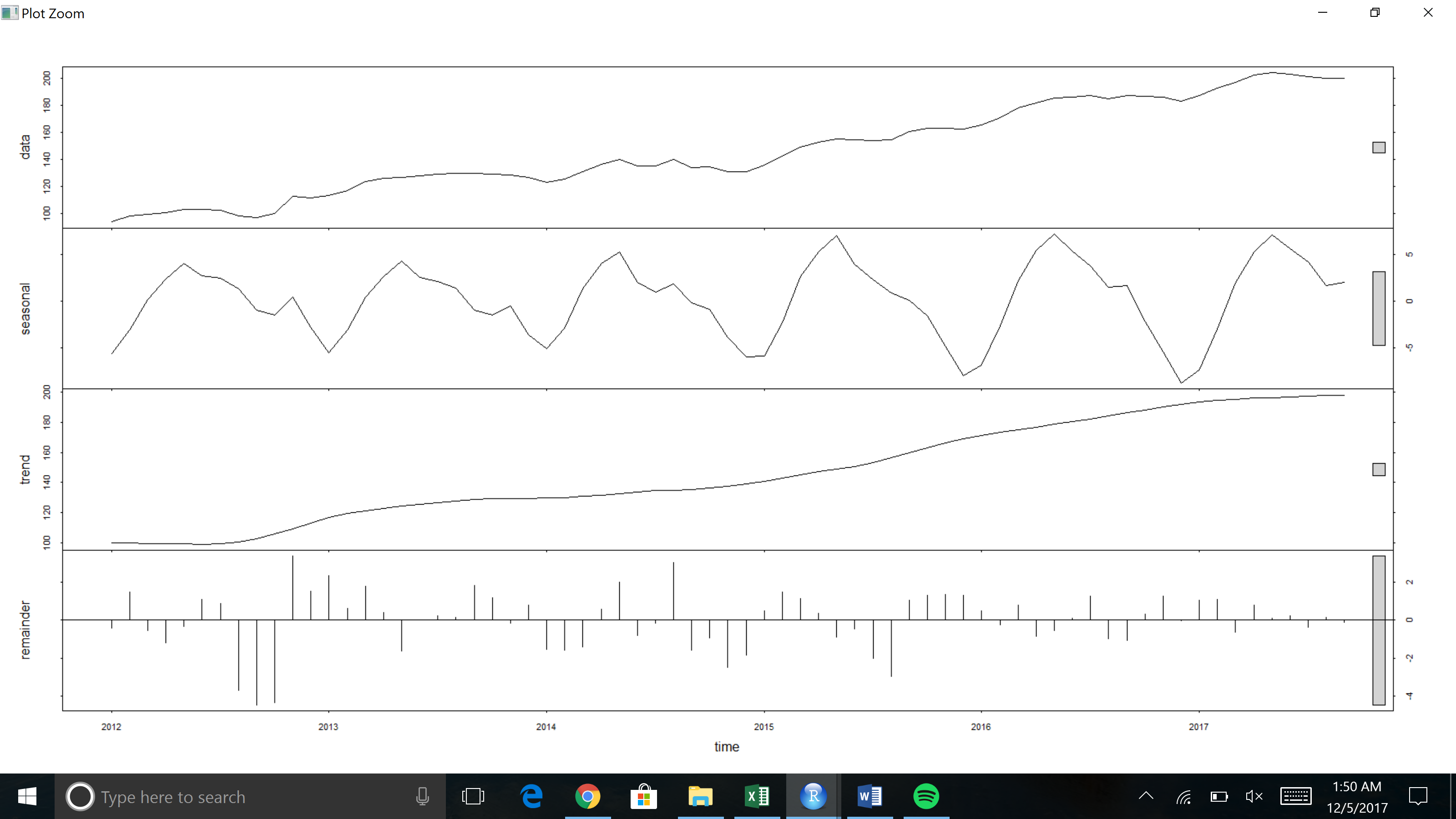
Durbin-Watson test

data: model1

DW = 0.54436, p-value = 1.385e-15

alternative hypothesis: true autocorrelation is not 0

**Arima Section Part III – Checking for Seasonality and Trend**



There is a not much concern for the impact of seasonality from looking at the stl plot. We decided to not de-seasonalize the data and continued to model using Arima.

**Stationarity Part III**

Augmented Dickey-Fuller Test

data: dataDallas1[, 2]

Dickey-Fuller = -2.5064, Lag order = 4, p-value = 0.3694

alternative hypothesis: stationary

Augmented Dickey-Fuller Test

data: dataDallas2

Dickey-Fuller = -4.1032, Lag order = 4, p-value = 0.01042

alternative hypothesis: stationary

After performing the Augmented Dickey-Fuller test, we determined differencing the data once. After the first differencing the test resulted in stationarity.

Series: dataDallas1[, 2]

ARIMA(1,1,0)(1,0,0)[12] with drift

Coefficients:

ar1 sar1 drift

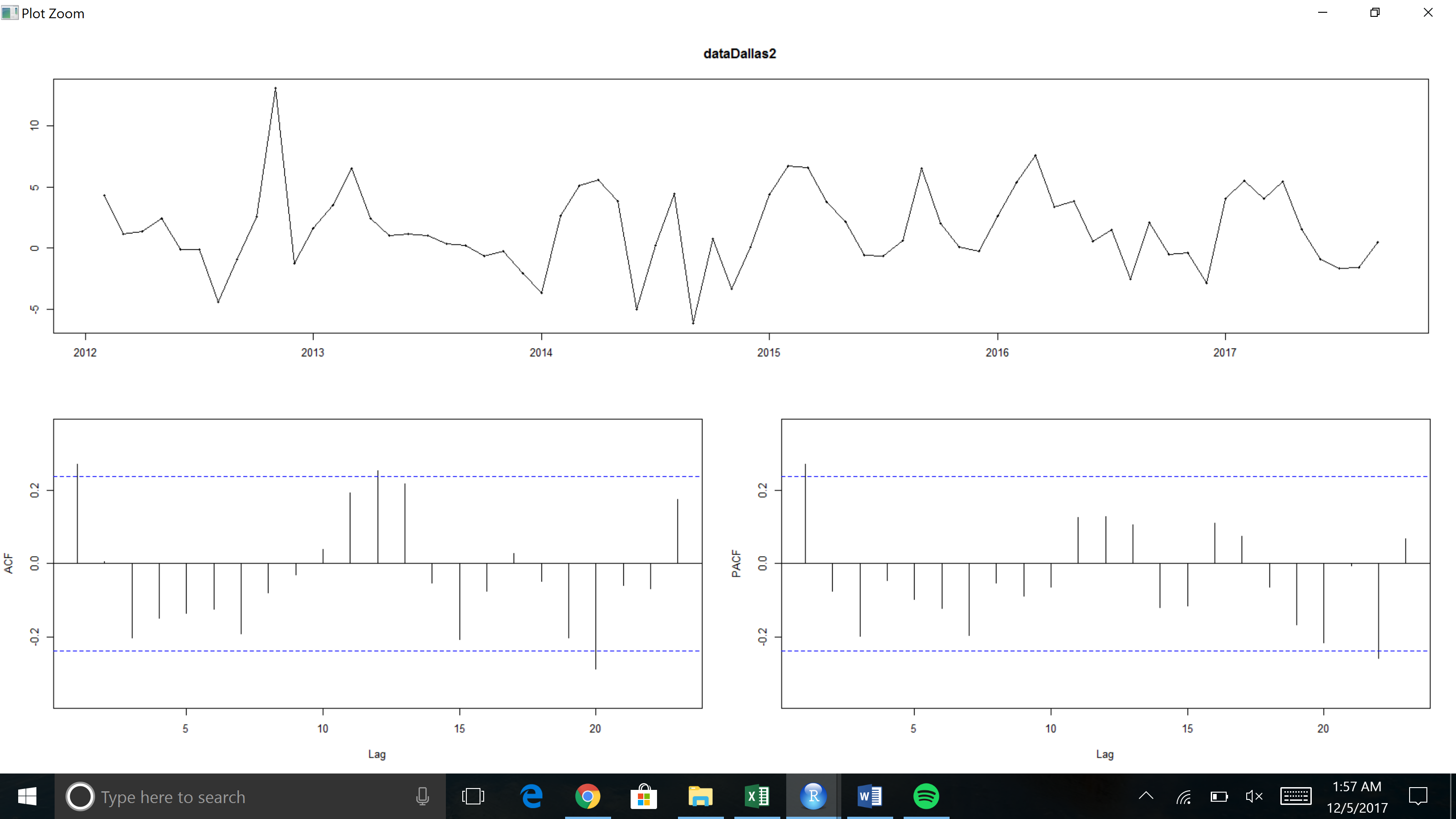
0.1955 0.2594 1.5476

s.e. 0.1253 0.1404 0.5898

sigma^2 estimated as 9.925: log likelihood=-173.42

AIC=354.85 AICc=355.48 BIC=363.72

An auto.arima coveys a model of p=1,d=1,q=0. (1,1,0). Ts display below of Acf and Pacf also can support this model. Thus, we chose to use this model.



arima(x = dataDallas1[, 2], order = c(1, 1, 0))

Coefficients:

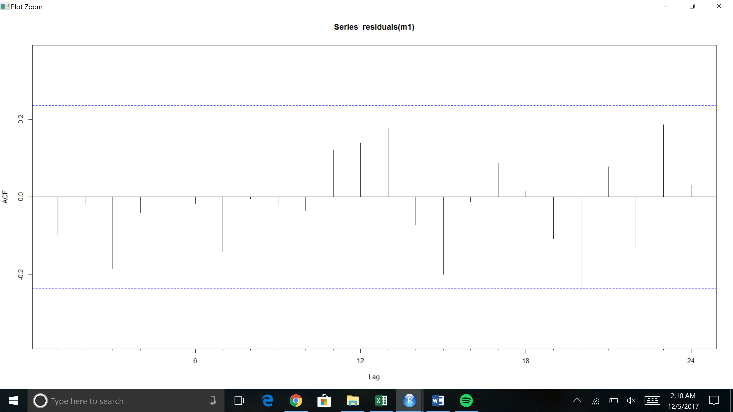
ar1

0.4013

s.e. 0.1110

sigma^2 estimated as 11.15: log likelihood = -178.56, aic = 361.12

**Residuals**

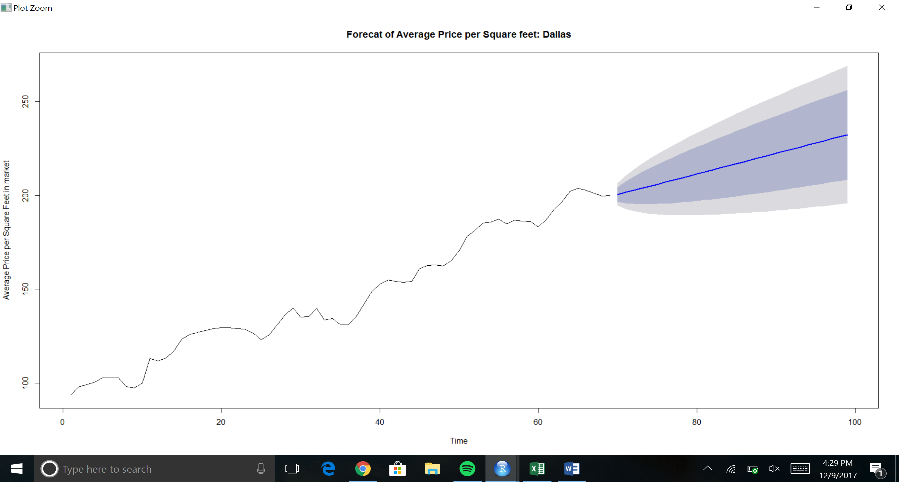


Box-Ljung test

data: residuals(m1)

X-squared = 5.2723, df = 6, p-value = 0.5094

Looking over the Acf plot of the residuals the lags seem to stay in the boundaries. Additionally, performing a box – Ljung test it seems this model was adequate to use in forecasting the response variable.

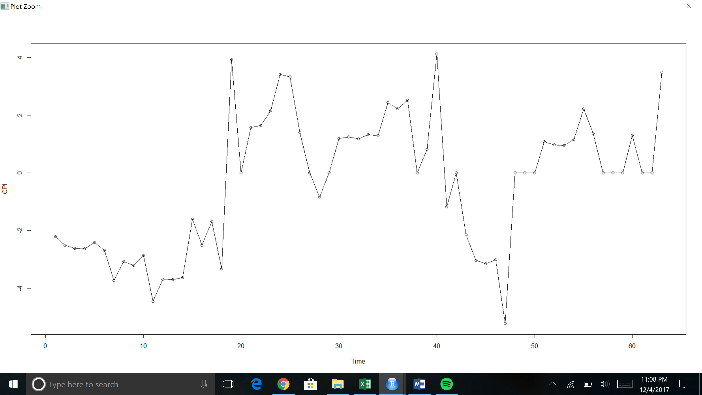


Above highlights our forecast for the average price per square feet of housing in the greater Dallas area from 2017 and above (30 periods). We used a dynamic regression model with ARIMA errors. It is important to remember that the prediction intervals from the regression models (with or without ARIMA Errors) do not consider the uncertainty in the forecast predictors.

**Regression plotting Co-efficient – Dynamically checking every 10 months by shifting over 1 month each time**

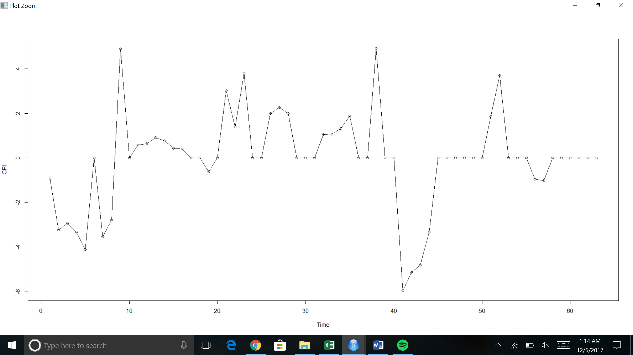
**CPI**

**San Francisco Philadelphia**

****

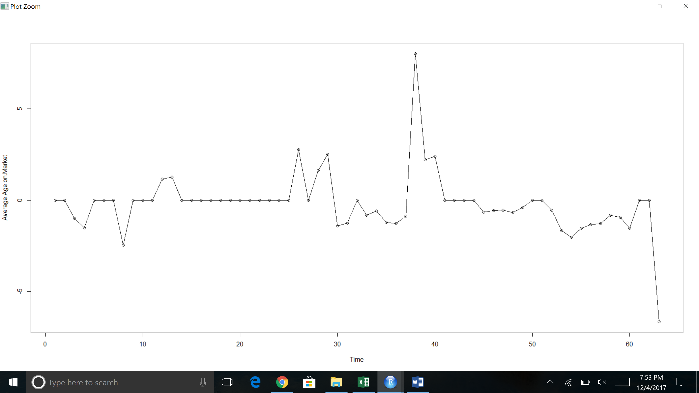
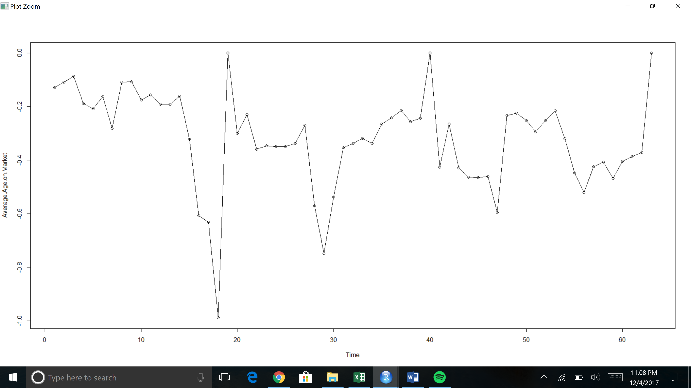
**Dallas**

All the plots show strong cyclicality. The plots for San Francisco and Dallas indicate no apparent trend. There seems to be a slight trend observed in Philadelphia. There is little relation between the peaks and troughs and when they occur. Philadelphia and Dallas experience a large trough between time 0 and 60.

****

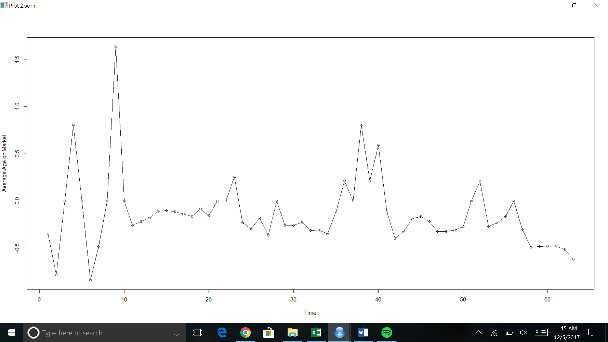
**Average Age on Market**

**San Francisco Philadelphia**

****

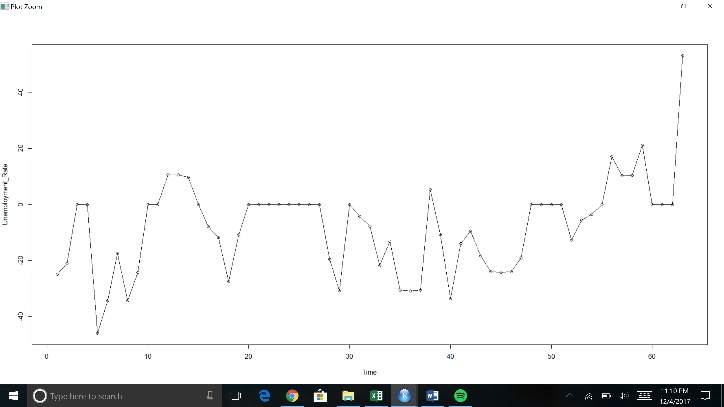
**Dallas**

All the plots show strong cyclicality. There is no apparent trend in San Francisco but there is a slight trend in both Philadelphia and Dallas. Dallas starts off with strong peaks, followed by a slight downwards trend and peaks again transitions into another downward trend.

****

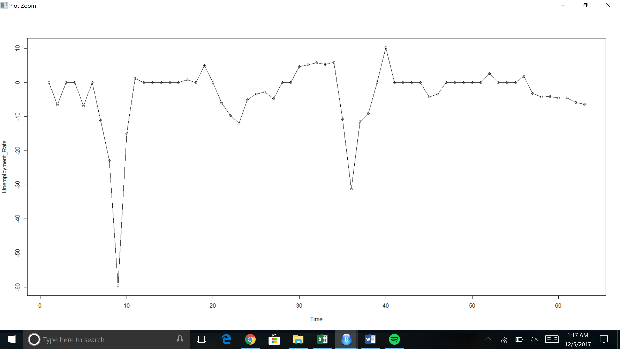
**Unemployment Rate**

**San Francisco Philadelphia**

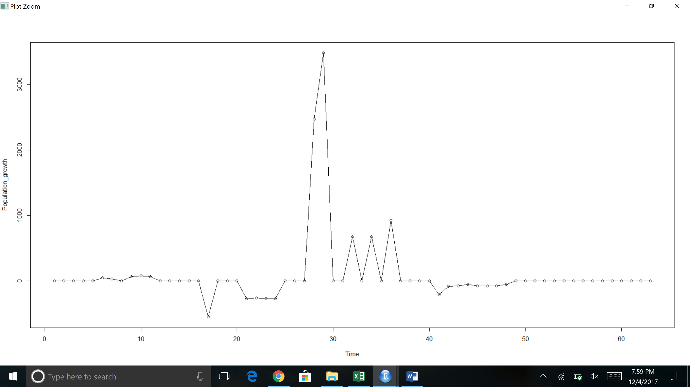
****

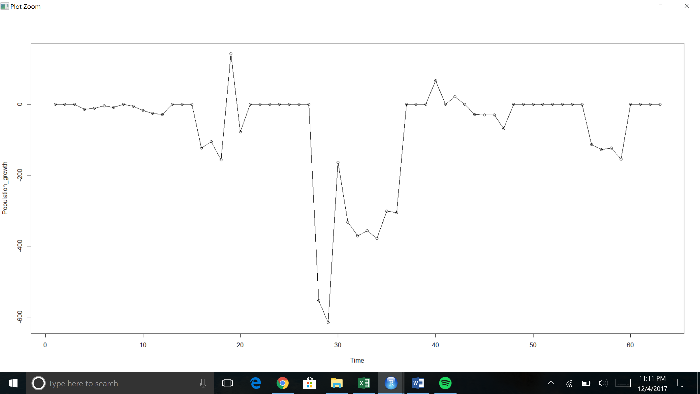
**Dallas**

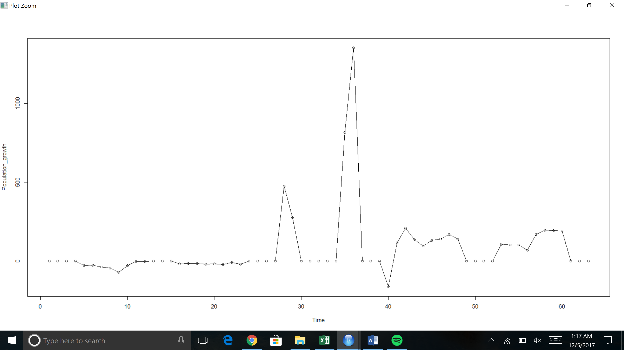
No apparent trend observed in San Francisco. A slight upward trend observed in Philadelphia which peaks towards the end. Dallas follows along a series of troughs with a large apparent trough observed at time 9. There seems to be similarity between San Francisco and Philadelphia between time 20 and 30.

****

**Population Growth**

** San Francisco Philadelphia**

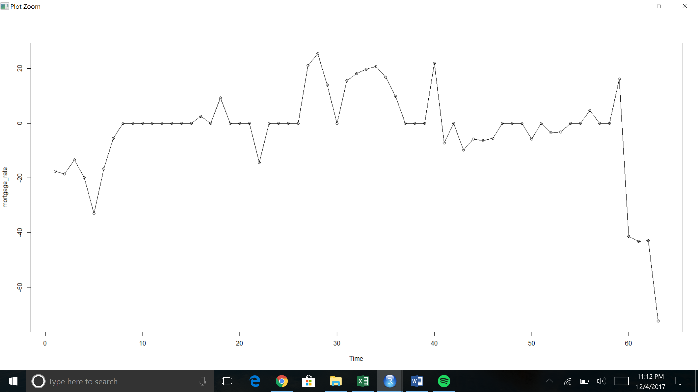
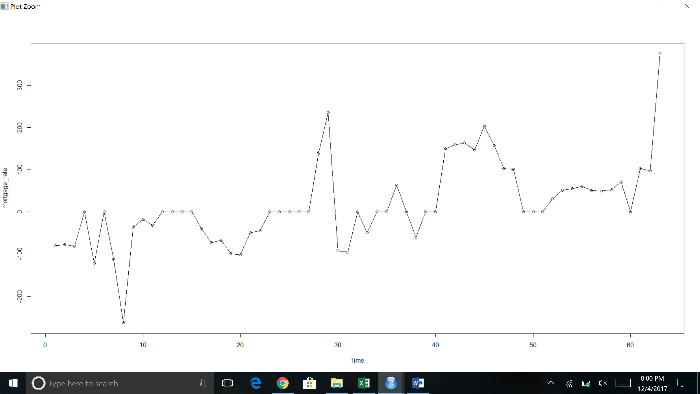
****

**Dallas**

There are large peaks observed in San Francisco around time 30, and around time 35 in Dallas. On the other hand, a large trough is also observed around time 30 in Philadelphia. All the plots show strong cyclicality but there is no apparent trend.

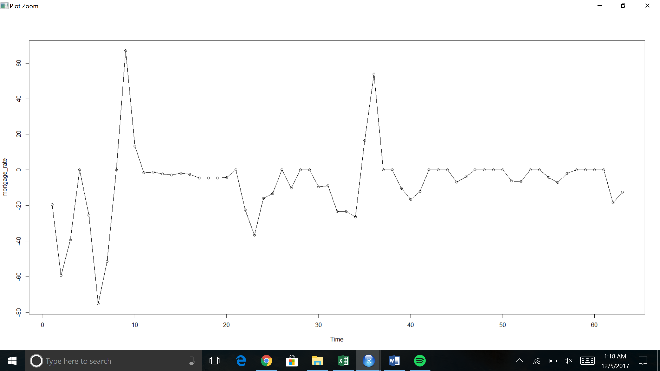
**Mortgage Rate**

**San Francisco Philadelphia**

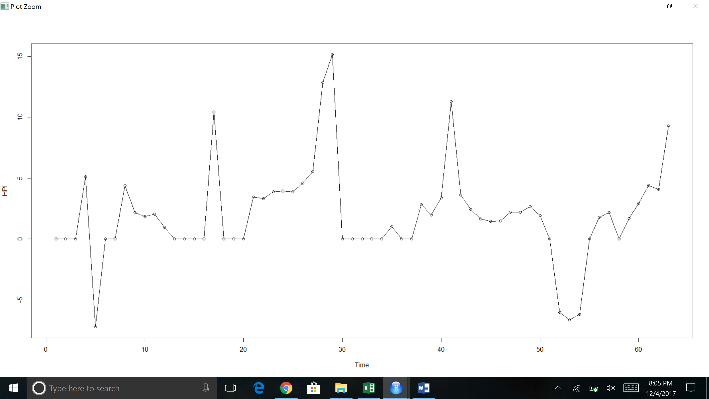
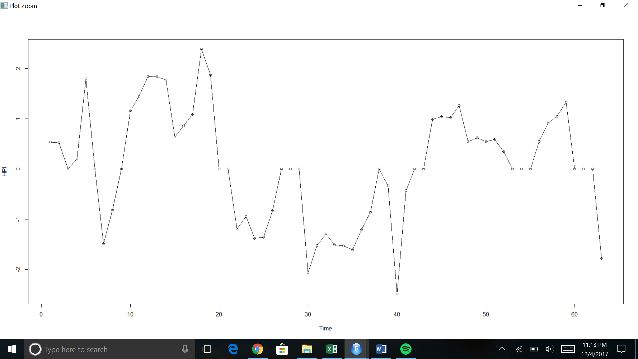
****

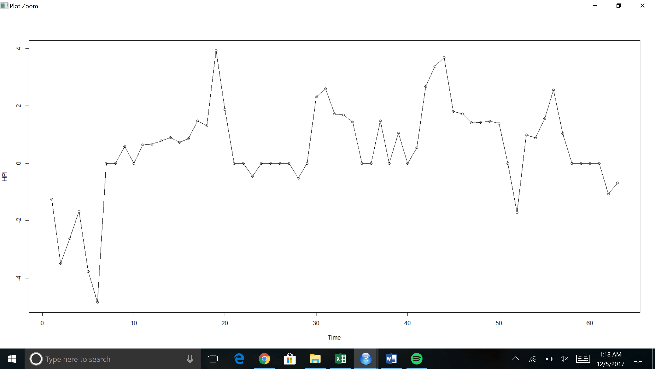
**Dallas**

There seems to be no apparent trend in San Francisco and Dallas, but a slight trend is observed in Philadelphia. Around time 60, there is an increase in the mortgage rates in San Francisco and a decrease in Philadelphia indicating an inverse relation around that time. A large trough following a large peak is observed in Dallas around time 7 and 10 and another peak around time 35. All plots show strong cyclicality.

****

**HPI**

**Philadelphia San Francisco**

**Dallas**

There is a slight trend in Philadelphia, but no apparent trend is observed in either San Francisco and Dallas. All the plots show series of peaks and troughs indicating strong cyclicality.