VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

ABHINAV VIJAYAKUMAR

19BCE1311

CSE3506 – ESSENTIALS OF DATA ANALYTICS
LAB-3

DR. LAKSHMI PATHI JAKKAMPUTI (L21 + L22)

Tasks for Week-3: Regression and Forecasting on Weather Data

Perform multi-regression and forecasting on weather related dataset "weatherHistory2016.csv"

Aim: To forecast the dependent variable temperature, based multiple independent variables.

Algorithm:

1. Attach library forecast, dplyr, corrplot, tseries.

2. Set working directory and read data.

3. Check correlation of variable.

4. Make multiple linear regression models.

5. Choose the best fit model.

6. Make a new dataset using the correlated variables only.

7. Formulate time series data.

8. Plot the time series data.

9. Plot the acf and pacf graph.

10. Perform the adf test, to determine the p value.

11. Check for stationary values.

12. Use auto ARIMA function to get the best fit model.

13. Perform forecasting with 95% level confidence.

14. Plot the forecasted data.

Statistics:

❖ **Multivariate Regression:**

Correlation Test:

a. Apparent Temperature

```
> cor.test(a$Temperature..C.,a$Apparent.Temperature..C.)

        Pearson's product-moment correlation

data:  a$Temperature..C. and a$Apparent.Temperature..C.
t = 148.4, df = 198, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9941002 0.9966209
sample estimates:
      cor
0.9955346
```

b. Humidity

```
> cor.test(a$Temperature..C.,a$Humidity)

        Pearson's product-moment correlation

data:  a$Temperature..C. and a$Humidity
t = -15.171, df = 198, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.7914113 -0.6617462
sample estimates:
       cor
-0.7331736
```

c. Wind Speed

```
> cor.test(a$Temperature..C.,a$Wind.Speed..km.h.)

        Pearson's product-moment correlation

data:  a$Temperature..C. and a$Wind.Speed..km.h.
t = 0.82873, df = 198, p-value = 0.4083
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.08060538  0.19593590
sample estimates:
       cor
0.05879314
```

d. Wind Bearing

```
> cor.test(a$Temperature..C.,a$Wind.Bearing..degrees.)

        Pearson's product-moment correlation

data:  a$Temperature..C. and a$Wind.Bearing..degrees.
t = 0.29656, df = 198, p-value = 0.7671
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.1180149  0.1593463
sample estimates:
       cor
0.02107109
```

e. Visibility

```
> cor.test(a$Temperature..C.,a$Visibility..km.)

        Pearson's product-moment correlation

data:  a$Temperature..C. and a$Visibility..km.
t = 7.2632, df = 198, p-value = 8.473e-12
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3416761 0.5616718
sample estimates:
      cor
0.4586739
```

Model Statistics:

```
Call:
lm(formula = a$Temperature..C. ~ a$Apparent.Temperature..C. +
    a$Humidity + a$Visibility..km.)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7773 -0.4679  0.1833  0.4463  2.8864

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                 4.48121    0.45802   9.784  < 2e-16 ***
a$Apparent.Temperature..C.  0.86288    0.00846 101.993  < 2e-16 ***
a$Humidity                 -2.16871    0.46174  -4.697 4.96e-06 ***
a$Visibility..km.          -0.01307    0.01541  -0.848    0.397
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8756 on 196 degrees of freedom
Multiple R-squared:  0.992,	Adjusted R-squared:  0.9919
F-statistic:  8094 on 3 and 196 DF,  p-value: < 2.2e-16
```

❖ **Forecasting:**

```
                Augmented Dickey-Fuller Test

data:  data
Dickey-Fuller = -6.287, Lag order = 12, p-value = 0.01
alternative hypothesis: stationary
```

Best ARIMA Model:

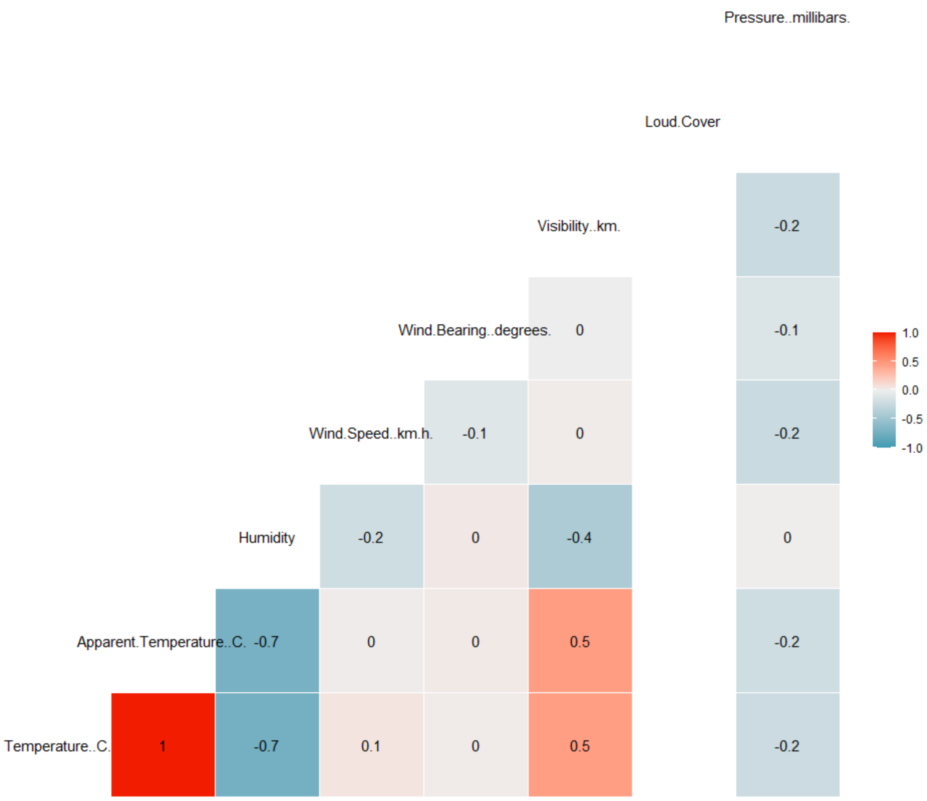
```
Fitting models using approximations to speed things up...

ARIMA(2,0,2)(1,1,1)[24] with drift        : Inf
ARIMA(0,0,0)(0,1,0)[24] with drift        : 11466.33
ARIMA(1,0,0)(1,1,0)[24] with drift        : 5676.337
ARIMA(0,0,1)(0,1,1)[24] with drift        : 8977.075
ARIMA(0,0,0)(0,1,0)[24]                   : 11473.89
ARIMA(1,0,0)(0,1,0)[24] with drift        : 6252.305
ARIMA(1,0,0)(2,1,0)[24] with drift        : 5438.054
ARIMA(1,0,0)(2,1,1)[24] with drift        : Inf
ARIMA(1,0,0)(1,1,1)[24] with drift        : Inf
ARIMA(0,0,0)(2,1,0)[24] with drift        : 11281.55
ARIMA(2,0,0)(2,1,0)[24] with drift        : 5374.887
ARIMA(2,0,0)(1,1,0)[24] with drift        : 5600.859
ARIMA(2,0,0)(2,1,1)[24] with drift        : Inf
ARIMA(2,0,0)(1,1,1)[24] with drift        : Inf
ARIMA(3,0,0)(2,1,0)[24] with drift        : 5331.394
ARIMA(3,0,0)(1,1,0)[24] with drift        : 5559.53
ARIMA(3,0,0)(2,1,1)[24] with drift        : Inf
ARIMA(3,0,0)(1,1,1)[24] with drift        : Inf
ARIMA(4,0,0)(2,1,0)[24] with drift        : 5332.032
ARIMA(3,0,1)(2,1,0)[24] with drift        : 5331.313
ARIMA(3,0,1)(1,1,0)[24] with drift        : 5558.243
ARIMA(3,0,1)(2,1,1)[24] with drift        : Inf
ARIMA(3,0,1)(1,1,1)[24] with drift        : Inf
```

```
ARIMA(2,0,1)(2,1,0)[24] with drift         : 5340.401

ARIMA(4,0,1)(2,1,0)[24] with drift         : 5334.033

ARIMA(3,0,2)(2,1,0)[24] with drift         : 5332.077

ARIMA(2,0,2)(2,1,0)[24] with drift         : 5330.361

ARIMA(2,0,2)(1,1,0)[24] with drift         : 5556.545

ARIMA(2,0,2)(2,1,1)[24] with drift         : Inf

ARIMA(1,0,2)(2,1,0)[24] with drift         : 5343.612

ARIMA(2,0,3)(2,1,0)[24] with drift         : 5331.938

ARIMA(1,0,1)(2,1,0)[24] with drift         : 5390.12

ARIMA(1,0,3)(2,1,0)[24] with drift         : 5332.634

ARIMA(3,0,3)(2,1,0)[24] with drift         : 5334.228

ARIMA(2,0,2)(2,1,0)[24]                    : 5329.467

ARIMA(2,0,2)(1,1,0)[24]                    : 5555.177

ARIMA(2,0,2)(2,1,1)[24]                    : Inf

ARIMA(2,0,2)(1,1,1)[24]                    : Inf

ARIMA(1,0,2)(2,1,0)[24]                    : 5342.563

ARIMA(2,0,1)(2,1,0)[24]                    : 5339.546

ARIMA(3,0,2)(2,1,0)[24]                    : 5331.22

ARIMA(2,0,3)(2,1,0)[24]                    : 5331.029

ARIMA(1,0,1)(2,1,0)[24]                    : 5388.923

ARIMA(1,0,3)(2,1,0)[24]                    : 5331.69

ARIMA(3,0,1)(2,1,0)[24]                    : 5330.489

ARIMA(3,0,3)(2,1,0)[24]                    : Inf


Now re-fitting the best model(s) without approximations...


ARIMA(2,0,2)(2,1,0)[24]                    : 5384.374


Best model: ARIMA(2,0,2)(2,1,0)[24]
```

Inference:

❖ Multivariate Regression:

The best model was made after considering 3 variable which were highly correlated to the dependent variable and those variables were Apparent.Temperature (0.9955), Humidity (-0.733) and Visibility(0.458).




```
Call:
lm(formula = a$Temperature..C. ~ a$Apparent.Temperature..C. +
    a$Humidity + a$Visibility..km.)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7773 -0.4679  0.1833  0.4463  2.8864

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                 4.48121    0.45802   9.784  < 2e-16 ***
a$Apparent.Temperature..C.  0.86288    0.00846 101.993  < 2e-16 ***
a$Humidity                 -2.16871    0.46174  -4.697 4.96e-06 ***
a$Visibility..km.          -0.01307    0.01541  -0.848    0.397
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8756 on 196 degrees of freedom
Multiple R-squared:  0.992,	Adjusted R-squared:  0.9919
F-statistic:  8094 on 3 and 196 DF,  p-value: < 2.2e-16
```

❖ **Forecasting:**

Best ARIMA Model:

```
ARIMA(2,0,2)(2,1,0)[24]                    : 5384.374


 Best model: ARIMA(2,0,2)(2,1,0)[24]
```

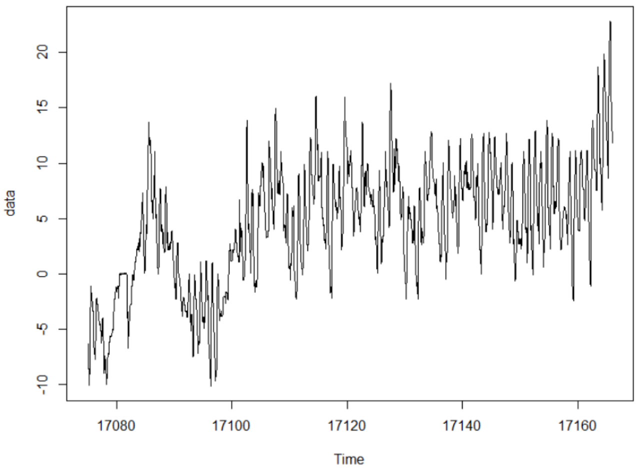
Accuracy of the Model:

```
> accuracy(model)
                     ME      RMSE       MAE MPE MAPE      MASE         ACF1
Training set 0.01742321 0.8309363 0.6157592 NaN  Inf 0.2180051 0.0008098431
```

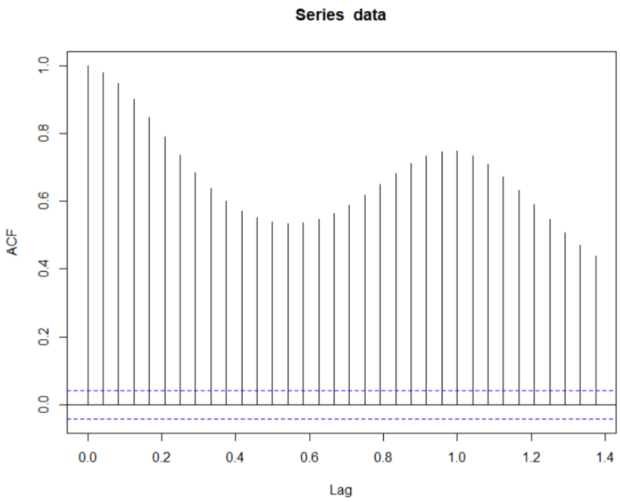
Forecast for 1 day:

Point	Forecast	Lo 95	Hi 95
17166.04	11.248803	9.60890098	12.88871
17166.08	10.731641	8.28507274	13.17821
17166.12	10.136301	6.97240528	13.30020
17166.17	9.559795	5.80095929	13.31863
17166.21	8.281384	4.03549680	12.52727
17166.25	8.287412	3.64185058	12.93297
17166.29	10.885030	5.90918055	15.86088
17166.33	13.721924	8.47093588	18.97291
17166.38	15.721102	10.23917779	21.20303
17166.42	17.608383	11.93132980	23.28544
17166.46	19.327787	13.48492337	25.17065
17166.50	19.884300	13.89986870	25.86873
17166.54	20.491771	14.38599322	26.59755
17166.58	20.578257	14.36812422	26.78839
17166.62	19.558943	13.25882055	25.85906
17166.67	19.521777	13.14387780	25.89968
17166.71	18.510491	12.06524140	24.95574
17166.75	15.073873	8.57020882	21.57754
17166.79	13.978204	7.42380488	20.53260
17166.83	13.325428	6.72691417	19.92394

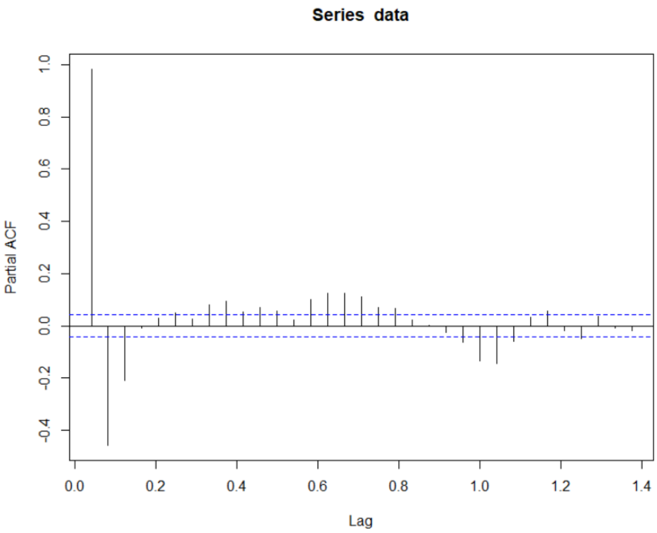
17166.88	13.118927	6.48201677	19.75584
17166.92	12.649515	5.97915833	19.31987
17166.96	11.722233	5.02272037	18.42175
17167.00	11.328230	4.60328548	18.05318



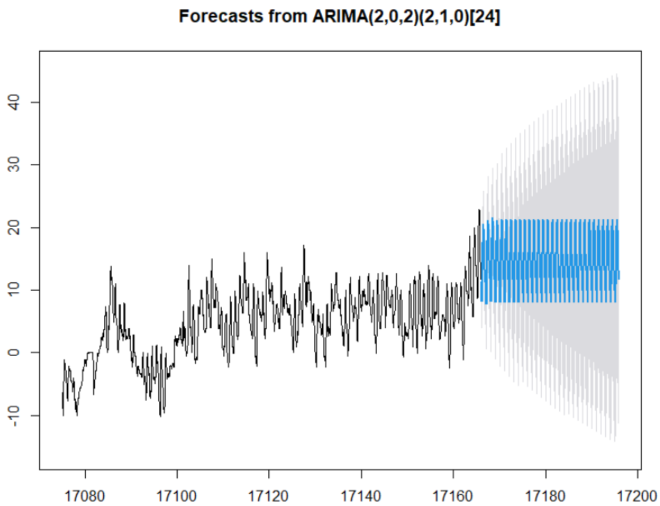
Plotting the data time series



Autocorrelation(acf)



Partial acf



Plotting the forecast

Program:

i) Multivariate Regression:

```
setwd("C:/Users/Abhinav Vijayakumar/Desktop/VIT Academics/Sem 6/Essentials of Data Analytics/LAB/LAB 3")

dff=read.csv("weatherHistory2016.csv")

head(dff)

library(dplyr)

library(GGally)

a=sample_n(dff,200)

head(a)

cor.test(a$Temperature..C.,a$Apparent.Temperature..C.)

cor.test(a$Temperature..C.,a$Humidity)

cor.test(a$Temperature..C.,a$Wind.Speed..km.h.)

cor.test(a$Temperature..C.,a$Wind.Bearing..degrees.)

cor.test(a$Temperature..C.,a$Pressure..millibars.)

cor.test(a$Temperature..C.,a$Visibility..km.)

cor.test(a$Temperature..C.,a$Loud.Cover)

ggcorr(a %>% mutate_if(is.factor, as.numeric), label = TRUE)

lmodel=lm(a$Temperature..C.~a$Apparent.Temperature..C.+a$Humidity+a$Visibility..km.)

summary(lmodel)

plot(lmodel)
```

ii) Forecast:

```
setwd("C:/Users/Abhinav Vijayakumar/Desktop/VIT Academics/Sem 6/Essentials of Data Analytics/LAB/LAB 3")

dff=read.csv("weatherHistory2016.csv")


library(forecast)

library(tseries)


data<-ts(dff$Temperature..C.,start = as.Date("2016-10-01"),end =
           as.Date("2016-12-31"),frequency = 24)
```

```
plot(data)

acf(data)

pacf(data)

adf.test(data)

model=auto.arima(data,ic="aic",trace=TRUE)

f=forecast(model,level=c(95),h=720)

f

plot(f)

accuracy(model)
```