Module 5

Week 5: R Practice

ALY6010

Instructor: Prof. Amin Karimpour

Date: 11 Dec 2021

Submitted by: -

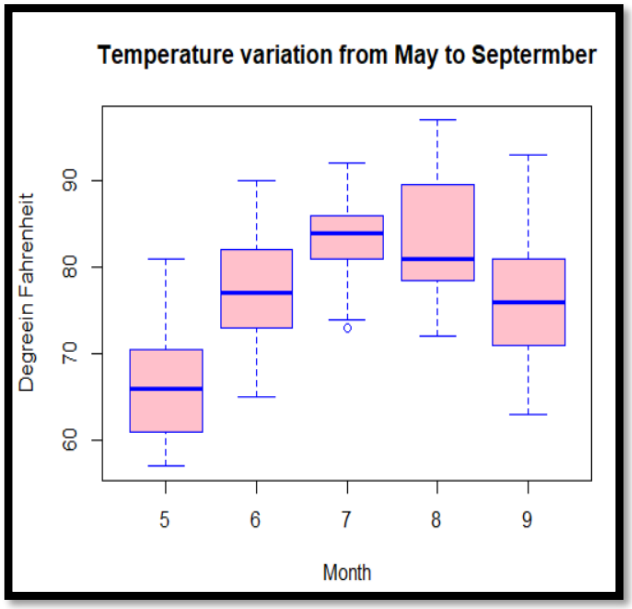
Jeseeka Shah

Nuid – 002134289

Introduction

The dataset is taken from the inbuilt dataset present in R called the “air quality”. The air quality is the dataset that has the values recorded from May to September 1973. In total it contains 153 rows and 6 variables. After data cleaning and removing the null values, it has 111 rows and 6 variables. This dataset was obtained from the New York State Department of Conservation and the National Weather Service (meteorological data).

Exploratory Data Analysis



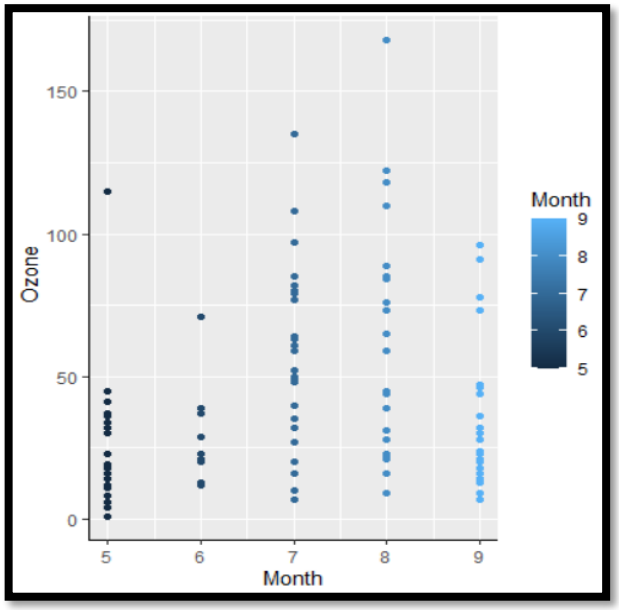
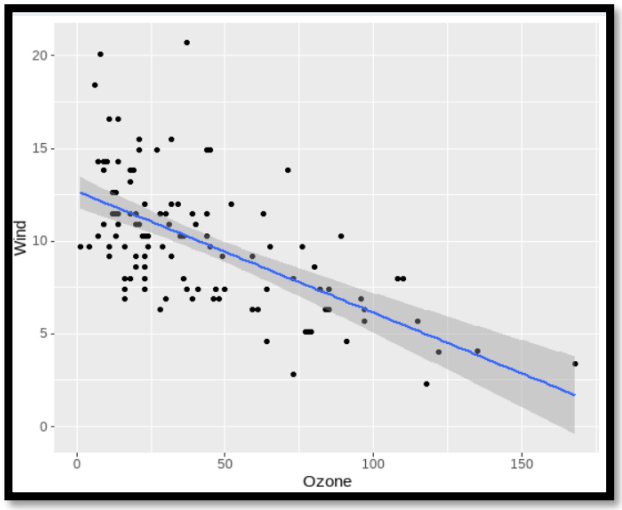


Fig 1: Temperature variation from May to Sept. **Fig 2: Ozone depletion from May to Sept.**

The **figure 1** shows the temperature from month of May to September. Its clearly seen that maximum temperature was observed during month of August. But the median of the boxplot shows that month of July had the most adverse temperature. The **figure 2** shows ozone layer depletion per month. The month of august has an outlier which show the highest ozone layer depletion. The month of july has the most occurring ozone layer depletion as the distribution has more density though out compared to just an single outlier of month of august.



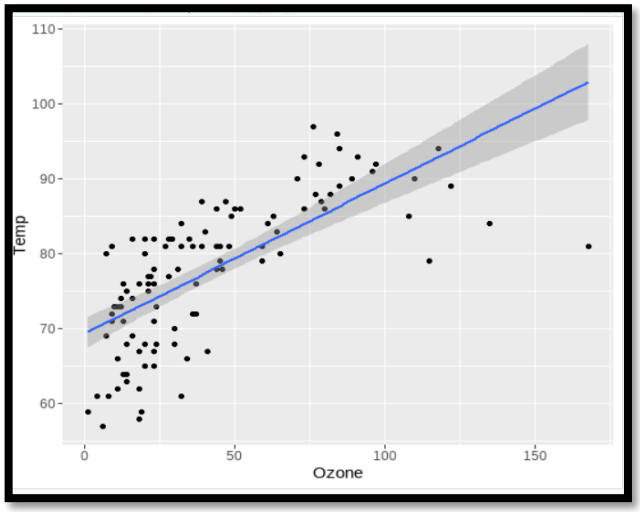


Fig 3: Scatter plot of Ozone and Wind **Fig 4: Scatter plot of Ozone and Temp**

The **fig 3** and **fig 4** shows the relationship between wind-ozone which is clearly negative correlation and temperature-ozone is seen as positive correlation. To understand correlation more , I have conducted correlation analysis in Part A and in Part B used linear regression to understand the association between two variables better.

Part A: Correlation Analysis

The correlation coefficient is measured from the range of -1 to 1. Understand the degree of association, following depicts the range for the respective correlation: -

Degree of association	Range
Perfect	+1 (Positive) or -1(Negative)
High – Strong correlation	± 0.50 and ± 1
Moderate – Moderate Correlation	± 0.5
Low degree- Weak Correlation	Below and ± 0.30 to ± 0.49

Note: when the correlation is Zero there is no correlation involved.

In this case, the temperature and ozone as positive correlation, and negative correlation is between wind and temperature which was seen in the above scatter plot too**.** The matrix shows the correlation between the variables of the dataset airquality. The **fig 6** shows values of correlation with X on it which means those are non-significant variables considering $p < 0.05$. Looking at the table above we can see that solar.r -ozone has a weak correlation, wind has non-significant correlation with solar.r, temperature has weak correlation with solar.r., month has no significant correlation with other variables except temperature where it shows negative correlation, and day has not a single significant correlation at $p < 0.05$. Similarly, **Fig 7** depicts the same with different format. The scale is given on the side and the significance is shown with the color as well as the size of the circle.

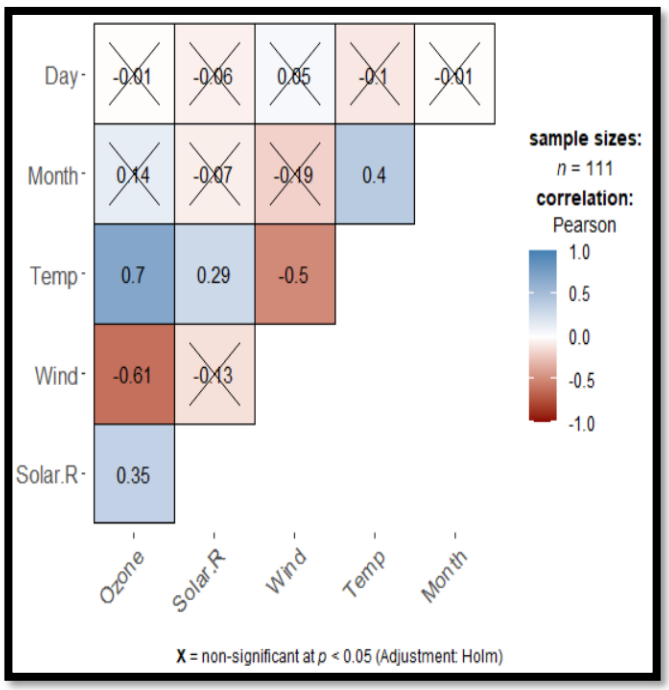


Fig6: Correlation Matrix using ggstatsplot

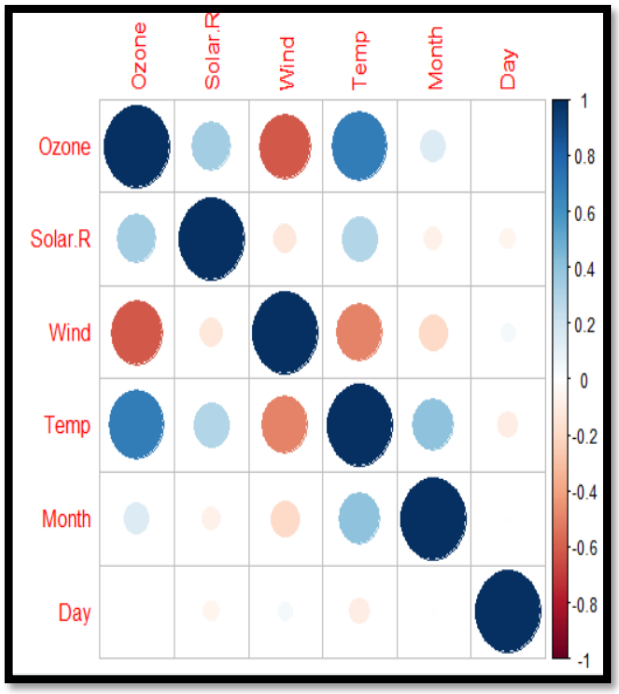


Fig 7 : Correlation Matrix using corrplot

```
> corr_aq_tab <- rcorr(as.matrix(airquality))
> corr_aq_tab
         Ozone Solar.R  Wind  Temp Month   Day
Ozone     1.00    0.35 -0.61  0.70  0.14 -0.01
Solar.R   0.35    1.00 -0.13  0.29 -0.07 -0.06
Wind     -0.61   -0.13  1.00 -0.50 -0.19  0.05
Temp      0.70    0.29 -0.50  1.00  0.40 -0.10
Month     0.14   -0.07 -0.19  0.40  1.00 -0.01
Day      -0.01   -0.06  0.05 -0.10 -0.01  1.00

n= 111


P
         Ozone  Solar.R Wind   Temp   Month  Day
Ozone           0.0002  0.0000 0.0000 0.1346 0.9569
Solar.R 0.0002          0.1835 0.0017 0.4398 0.5471
Wind    0.0000 0.1835          0.0000 0.0408 0.6032
Temp    0.0000 0.0017  0.0000         0.0000 0.3134
Month   0.1346 0.4398  0.0408 0.0000         0.9253
Day     0.9569 0.5471  0.6032 0.3134 0.9253
```

Fig 8 : Correlation Table

The rcorr() used here is used to compute the significance level of variables for the Pearson and Spearman correlations. The **figure 8** shows correlation coefficients along with p-values for all the possible variables of the dataset. P values should be less than 0.05 and the null hypothesis should be rejected to have correlation else failing to reject the null hypothesis means that there is zero correlation between the variables.

PART B: Linear Regression

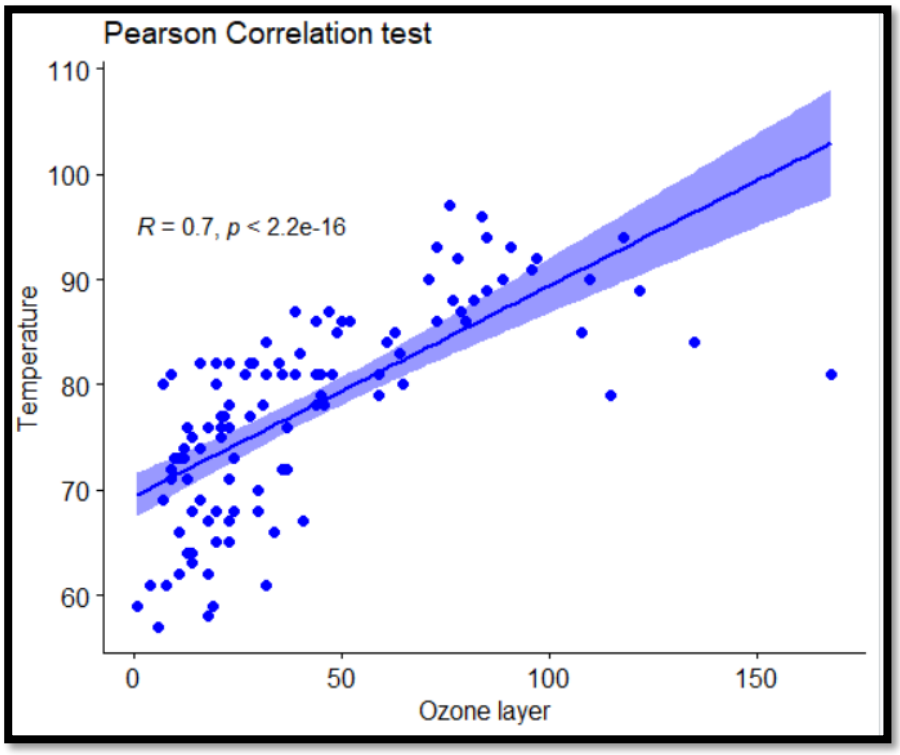


Fig 9 : Correlation test on temperature and Ozone layer

The Pearson correlation test shows a ***positive relation*** between Ozone layer and Temperature.

```
        Pearson's product-moment
        correlation

data:  airquality$Ozone and airquality$Temp
t = 10.192, df = 109, p-value <
2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5888139 0.7829869
sample estimates:
      cor
0.6985414
```

```
> cor(airquality$Ozone,airquality$Temp)
[1] 0.6985414
```

There is a positive correlation between ozone and temperature. As the correlation by the Pearson correlation test in scatter plot matches the result of Cor.test() function. Which stated that correlation is 0.7 which is close to 1.

The null hypothesis here will be **H0: ρ=0, H1: ρ≠0** where ρ is denotes the correlation. The test depends on number of observation and correlation coefficients. As the observation increases the correlation between two variables also increases which increases the chances of rejecting the null hypothesis that states that there is no correlation between the two variables. Here, p-value is less than the significance level which is 0.05 which points towards **rejecting the null hypothesis**.

Regression Table

```
> summary(Regression_tb)

Call:
lm(formula = Ozone ~ Solar.R + Wind + Temp + Month + Day, data = airquality)

Residuals:
    Min      1Q  Median      3Q     Max
-37.014 -12.284  -3.302   8.454  95.348

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -64.11632   23.48249  -2.730  0.00742 **
Solar.R       0.05027    0.02342   2.147  0.03411 *
Wind         -3.31844    0.64451  -5.149 1.23e-06 ***
Temp          1.89579    0.27389   6.922 3.66e-10 ***
Month        -3.03996    1.51346  -2.009  0.04714 *
Day           0.27388    0.22967   1.192  0.23576
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20.86 on 105 degrees of freedom
Multiple R-squared:  0.6249,    Adjusted R-squared:  0.6071
F-statistic: 34.99 on 5 and 105 DF,  p-value: < 2.2e-16
```

```
tab_model(Regression_tb)
```

	Ozone		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	-64.12	-110.68 – -17.55	**0.007**
Solar R	0.05	0.00 – 0.10	**0.034**
Wind	-3.32	-4.60 – -2.04	**<0.001**
Temp	1.90	1.35 – 2.44	**<0.001**
Month	-3.04	-6.04 – -0.04	**0.047**
Day	0.27	-0.18 – 0.73	0.236
Observations	111		
R^2 / R^2 adjusted	0.625 / 0.607		

Fig 10: Regression Table

The regression table shows the relation of one variable with other variables. Here, we choose to check the relation of ozone with all other variables present. The linear regression model depends on the number of independent variables, shape of regression line and the type of dependent variable. The regression table has statistical values such as **residuals** which shows the difference between actual and predictable values. Then **coefficients** which is depicted by the generalized

formula of $y=mx+b$, Where b is a coefficient and m is a slope . This gives us a quick idea about variables and its significance. After which **residual standard error** comes which must be small only then our actual and predicted line will be very close. Adjusted R-square considered and test different independent variables against the model. In this case almost 60 percent values will fall on the best fit line. F-statistics or F values are used along with p-values to come to conclusion whether to reject the null hypothesis or not. Here, as the p value is less than 0.05 (alpha) .Looking at p-value have **rejected the null hypothesis that correlation of ozone with all variables is zero**.

Summary

- The dataset airquality measures the air quality from May to September of 1973. Here, the worst temperature was seen in the month of august and july. And the ozone layer depletion was maximum in the month of august, but intensity was more during the month of july . As august has a outlier as the maximum temperature .
- Correlation is a unit free; both variables does not have to be of the same scale. Correlation is having a range of -1 to 1. Relationship which is close to 0 has weak correlation. The signs provide the direction of the association of the variables only.
- In the dataset airquality, the temperature -ozone has a strong positive correlation , wind - ozone has strong negative corelation, and solar.r – ozone has weak corelation.
- Correlation matrix clearly showed that the variable “day “ has no significance in determining other variables. The variable “month” having only significant correlation with the variable temperature.
- Linear regression helped in showing the relationship between the variables between temperature -ozone and wind -ozone.
- I have conducted the Pearson’s correlation test having alpha as 0.05 and confidence interval as 0.95 with temperature and ozone. The slope is positive and p -value was less than alpha(0.05) which means the correlation exist. This result was also shown by the correlation matrix where the p value was 0.70.
- To find best-fit-line points by checking the regression table. R-squared, should be carefully considered as it cannot show the if the model is bias or not and cannot take into consideration the different independent variables against the model. Hence, adjusted R is considered more often .
- Adjusted R gave the value as around 60 % which mean that sixty percentage of point will be on the best-fit line.
- All the above factors proved that there is a strong correlation between the variable’s ozone and temperature. As the ozone layer depletion increases the temperature also increases.
- The regression table is different from the correlation table. As the regression shows how the independent variable is numerically related to the dependent variable. Whereas correlation table shows the association of two variables. Furthermore, regression line shows per unit change of known variable with respect to predicted variable. And correlation coefficient shows at what extend do the variables move together.

.

Reference

- *Correlation matrix : A quick start guide to analyze, format and visualize a correlation matrix using R software - Easy Guides - Wiki - STHDA*. (n.d.). Correlation Matrix. http://www.sthda.com/english/wiki/correlation-matrix-a-quick-start-guide-to-analyze-format-and-visualize-a-correlation-matrix-using-r-software
- *F Statistic / F Value: Definition and How to Run an F-Test*. (2021, November 20). Statistics How To. https://www.statisticshowto.com/probability-and-statistics/f-statistic-value-test/
- F. (2021a, May 14). *Correlation Analysis Different Types of Plots in R*. Finnstats. https://finnstats.com/index.php/2021/05/13/correlation-analysis-plot/
- *Correlation tests, correlation matrix, and corresponding visualization methods in R*. (n.d.). Correaltion Rstudio. https://rstudio-pubs-static.s3.amazonaws.com/240657_5157ff98e8204c358b2118fa69162e18.html
- S, S. (2021, February 26). *Difference Between Correlation and Regression (with Comparison Chart)*. Key Differences. https://keydifferences.com/difference-between-correlation-and-regression.html