STA141A - Project

Effects on Ozone - Group 1

6/1/2021

	Contribution	E-mail
Shih-Chi Chen	Set up for the Model, Visualizations and Statistical Analysis, Game function	scichen@ucdavis.edu
Hien Nguyen	Conclusion , Editing / Quality Overseer	- hqnnguyen@ucdavis.edu
Elizabeth Jiang	Data Cleanup, Editing / Quality Overseer	- lizjiang@ucdavis.edu
Andre Martinez	Introduction, Data background and Questions of Interest, RMD editor	- amartinezr@ucdavis.edu
Cecilia Nguyen	Interpretation/Reporting Analysis	- cecnguyen@ucdavis.edu

Emanuela Furfaro Instructor

STA 141A - Fundamentals of Statistical Data Science

University of California, Davis

June 1, 2021

A. Introduction

The following paper will analyze New York daily air quality measurements from the months of May to September 1973. This data set specifically measures the Ozone in parts per billion during the daily periods of 1 P.M. – 4 P.M. at Roosevelt Island, Solar radiation from 8 A.M– 12 P.M. at central park, average wind speeds in mph, and the maximum daily temperatures in Fahrenheit at LaGuardia Airport (approximately 2 miles away from Roosevelt Island). Although this data is reasonably outdated, it still provides a lot of useful information about the ozone, as air pollution is still a major problem we face all over the world. Monitoring the ozone layer is extremely valuable as it protects public health. Many agencies around the world have air quality monitors that use past data to confirm that ozone concentrations meet the standards of protecting both public health and the environment [4]. Therefore, we will analyze the **airquality** data set (found within base R) from 1973 and assess whether ozone can be explained by certain predictors, or if certain characteristics are present when dealing with a high or low concentration reading.

B. Data background and Questions of Interest

As previously mentioned, this data set measures ozone in parts per billion (ppb) during the daily periods of 1 P.M. – 4 P.M at Roosevelt Island, Solar radiation from 8 A.M – 12 P.M at Central Park, average wind speeds in mph, and the maximum daily temperatures in Fahrenheit at LaGuardia Airport (approximately 2 miles away from Roosevelt Island).

Ozone is a gas that is made up of three oxygen atoms (chemical formula O_3) and it is produced naturally in small amounts in the upper atmosphere [5]. This Ozone layer protects all life on Earth from the sun's ultraviolet radiation [5]. However, in the lower atmosphere, harmful ozone is produced by a combination of air pollutants that humans make. While upper atmosphere ozone is important to our survival, high concentrations of lower atmosphere ozone are very toxic to people and the environment [5]. That is why even in the year 1973 up to the present day, monitoring ozone concentration is more important than ever.

Another variable that was also measured during this five month period was solar radiation, which is the radiation that is emitted directly from the sun and is measured in wavelengths. In our data specifically, it is the Solar radiation in Langley's in the frequency band 4000–7700 Angstroms from 0800 to 1200 hours at Central Park. Mean wind speeds in miles per hour and daily temperature in Fahrenheit were measured at LaGuardia Airport. Overall, we have a total of 6 variables including ozone concentration, solar radiation, wind speeds, average temps, months 5-9 (corresponding to May to September), and the days of the specific months.

Our goal, along with the combination of these variables is to test whether there is a statistically significant relationship between these potential predictors and the ozone concentrations. Given this thought process, we have some questions we would like to answer:

1. Does Solar radiation, Average wind speed, Temperature, Month, Day have significant effects on Ozone?
2. Which of these factors has the greatest influence on Ozone?
3. When the Ozone is high? Are there some characteristics to make high Ozone?
4. Is there a relationship between Ozone and Solar radiation?

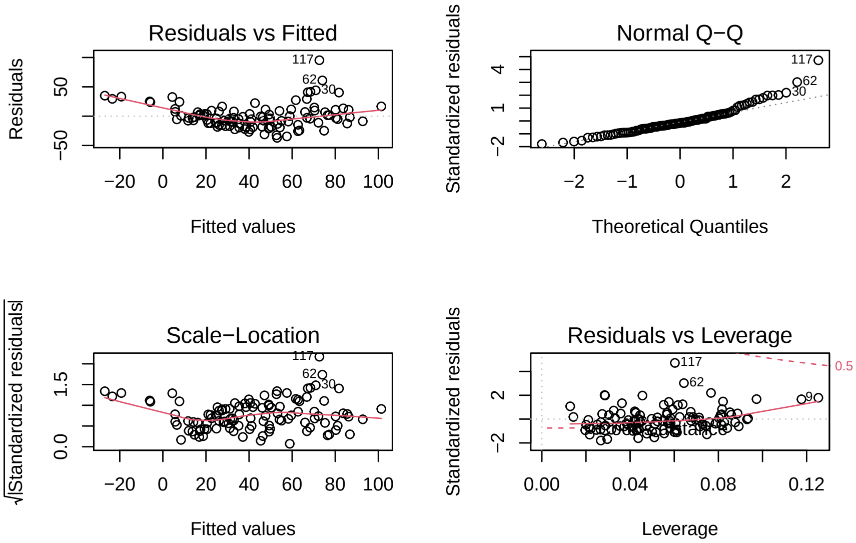
Through statistical analysis, we will be able to find answers to back up our claims and make decisions based on our findings. Having answered these questions will allow us to learn more about what influences ozone concentration.

C. Set up for the Model

1. Analysis of Full Model

(i) Assumption of Full Model

`Full model:Ozone~Solar.R+Wind+Temp+Month+Day`



From observing the Normal Q-Q Plot, the assumption of normality holds because most dots are close to or on the regression line; though, some outliers exist (#62, #117).

From observing the Residuals vs Fitted Plot, some dots do not bounce randomly around the 0 line and some dots stand out from the basic random pattern of residuals. In addition, the residuals do not roughly form a horizontal band around the 0 line. Therefore, the assumption of equal variance may be violated.

From the Residuals vs Leverage plot, since our leverage statistic $=\frac{(5+1)}{153} = 0.03921569$, there are many dots with high leverage, and #62, #117 are also outliers.

(ii) Summary of Full Model

The full model has $R^2 = 62.49$, which means these predictor variables can explain about 62.49% of the changes in the response variable y (mean ozone). Besides, the residual standard error is high in this model (20.86). When running an F test for a multiple regression model with an intercept, the small p-value($<$ 2.2e-16) is obtained. In other words, there is a relationship between these predictor variables and response variable Y (mean ozone concentration). Therefore, analyses on in-depth individual predictor variables are needed.

(iii) Model Descion

`Reduced model:Ozone~Solar.R+Wind+Temp+Month`

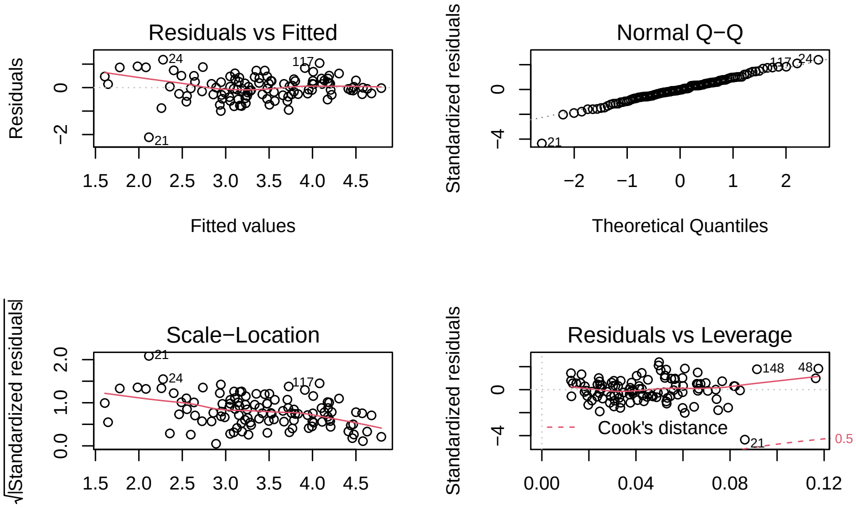
The stepwise selection shows that the final model should not include the Day variable. It can also be found that the Day variable has the smallest AIC statistic, which suggests that the Day variable can be dropped. Therefore, we would use the reduced model in the following analysis.

2. Transformations and diagnostics

Since there are some outliers and the assumption of equal variance is violated, transformations are needed.

(i) Log transformation

After using the log transformation, $R^2 = 66.76$, which is higher than before. It suggests that approximately 66.76% of the observed variation can be explained by the model's inputs. Besides, the residual standard error greatly reduces to 0.5085. Therefore, this log transformation model is better than the original model.



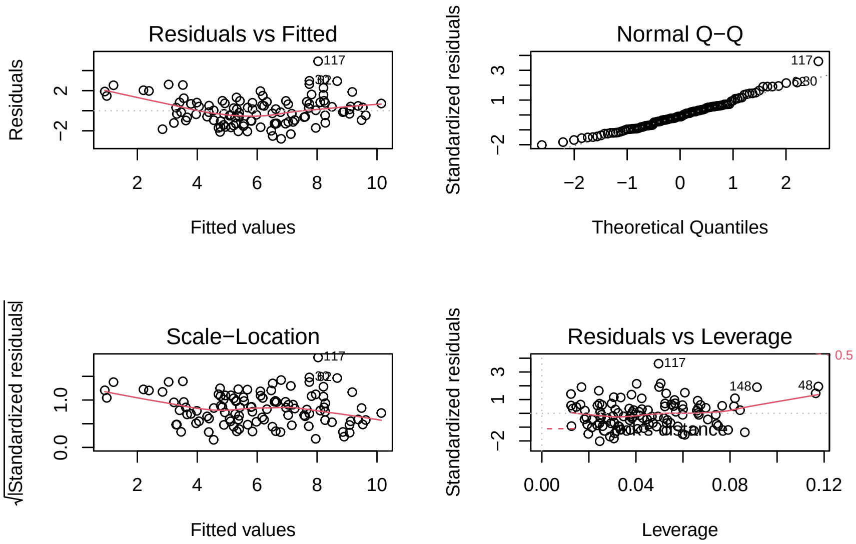
From the Normal Q-Q Plot, the assumption of normality holds because most dots are close to the line or on the line. However, there is an outlier (#21) that still exists.

From the Residuals vs Fitted Plot, many dots bounce randomly around the 0 line but there are some dots #21, #24 and #117 observations standing out from the basic random pattern of residuals. In addition, the residuals roughly form a horizontal band around the 0 line. Since some dots are still far away from the 0 line, the assumption of equal variance may not hold.

From the Residuals vs Leverage plot, since our leverage statistic $= \frac{(4+1)}{153} = 0.03267974$, there are still many dots with high leverage including the outlier #21.

(ii) Square-root transformation

After using the square-root transformation, $R^2 = 68.17$, which is higher than before. It suggests that approximately 68.17% of the observed variation can be explained by the model's predictors. Besides, the residual standard error greatly reduces to 1.403. Therefore, this square-root transformation model is better than the original model.



From the Normal Q-Q Plot, the assumption of normality holds because most dots are close to the line or on the line, but there is an outlier #117 far away from the line.

From the Residuals vs Fitted Plot, many dots bounce randomly around the 0 line and only #117 observation stands out from the basic random pattern of residuals. In addition, the residuals roughly form a horizontal band around the 0 line. Therefore, the assumption of equal variance may hold.

From the Residuals vs Leverage plot, since the leverage statistic $=\frac{(4+1)}{153}$=0.03267974, there are still many dots with high leverage including the outlier #117.

(iii) Transformation Decision

Since the square-root transformation holds the assumptions of normality/equal variance and has a higher R-square value than the log transformation, we will use the square-root transformation in the following analysis. However, it is worth noting that the transformed data still have roughly heteroscedastic residuals, but it is better than that of the original model.

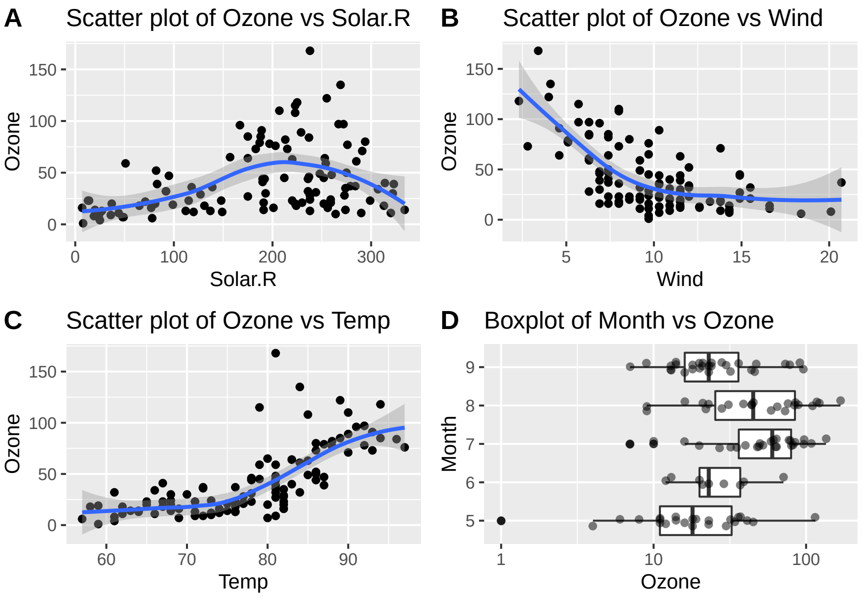
3. Outliers Analysis

According to Residuals vs Fitted and Normal Q-Q plots of the square-root transformation, we see that #117 is an outlier. This observation has the highest ozone concentration, which is almost 4 standard deviations from the mean ozone concentration. This observation has the highest ozone concentration at the end of August with high temp. Therefore, further analyses of month, temperature and ozone are necessary. Since we analyze various possible causes of high ozone and only one outlier is found, this outlier will not be removed in the following analysis.

```
    Ozone Solar.R Wind Temp Month Day
117   168     238  3.4   81     8  25
```

D. Visualizations and Statistical Analysis

1. Scatter plots and Boxplot



Scatter Plot	Correlation value
Ozone vs Solar.R	0.3483
Ozone vs Wind	-0.6125
Ozone vs Temp	0.6985

A - From the scatter plot (Ozone vs Solar.R), it shows that Solar.R and Ozone do not have a linear relationship because these two variables have low correlation (r = 0.3483417).

B - From the scatter plot (Ozone vs Wind), it shows that Wind and Ozone have a negative linear relationship, which means that mean ozone decreases as average wind speed increases. Therefore, they have a moderately high negative correlation (r = -0.6124966).

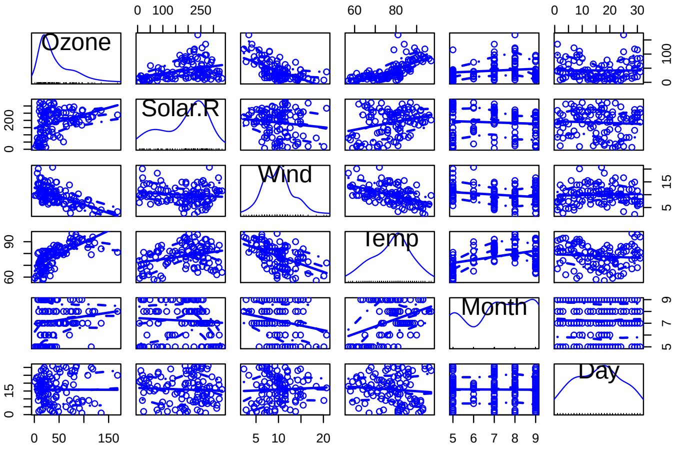
C - From the scatter plot (Ozone vs Temp), it shows that Temp and Ozone have a positive linear relationship, which means that mean ozone increases as maximum daily temperature increases. Therefore, they have a high positive correlation (r = 0.6985414).

Since Month is a categories variable, let's look at boxplot **- D**.

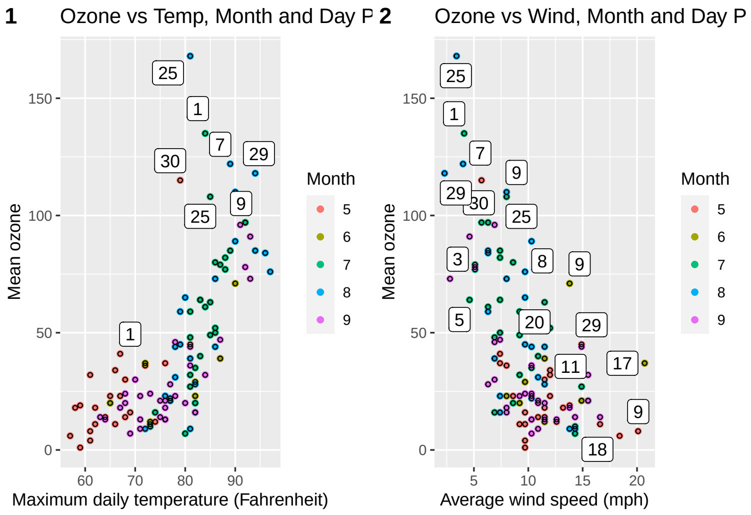
From the boxplot (Ozone vs Month), observations in July and August have higher ozone concentrations on average. We can also see the maximum ozone level is observed in August. In contrast, the minimal ozone level is observed in May. These two months also have the largest IQR (interquartile range), which means that data of these two months are more scattered. In addition, there is a high right-skewed distribution and the smallest IQR (interquartile range) in June.

2. Three and more varaiables of Two dimensional plots

From all variables of two dimensional plots, it shows that Ozone has a high positive relationship with both Temperature and Month because Temperature is associated with Month. On the other hand, Ozone has a high negative relationship with Wind Speed. However, there is no significant relationship associated with Wind Speed and any other predictor variable.



After finding the relationships between Ozone and other variables, the variables worth analyzing are Temperature and Wind Speed associated with Month and Day. Therefore, some more variables of two dimensional plots can be investigated.



From more variables of two dimensional plots, high mean Ozone seems to occur from early July to the end of August because of high temperature. However, high mean ozone is also found on May 30th, and this observation is also associated with the highest temperature in May.

From more variables of two dimensional plots, the high mean Ozone almost occurs from early July to the end of August because of low average Wind Speed. However, high mean Ozone is also found on May 30th, and this observation is also associated with low average wind speed in May. On the other hand, the low mean Ozone almost occurs in May because of high average Wind Speed.

3. Statistical Analysis

(i) F Test of Multiple Regression Analysis

State the hypothesis:

$H_0 : \beta_0 = \beta_1 = ... = \beta_4 = 0$

H_a : At least one of $\beta_k \neq 0$

Since the F-stat's p-value is small (< 2.2e-16), we would reject the null hypothesis and conclude that these predictor variables and mean ozone have significant relationship.

(ii) T Test of Standardized Regression (Important factor analysis by coefficients)

In order to find which factor has the most significant influence on Ozone, standardized regression analysis is needed because each predictor variable has different units.

After the standardized regression is conducted, R^2 and p-values do not change, and the difference exists on the estimator of each variable. According to the t-test p-values, the same results are those obtained from previous linear regression. The results include that Temperature and Wind have statistically significant effects on mean Ozone because of small p-values. For the Temperature, it results in the largest absolute value of the estimator (0.5808), which means that the mean ozone increases 0.5808 as temperature increases one unit. For the Wind variable, it results in the second largest absolute value of estimator (0.3194). Since it is a negative estimator, it means that the mean ozone decreases 0.3194 as average wind speed increases one unit. To sum up, according to these standardized regression coefficients, temperature has the most significant influence on ozone by increasing ozone while wind has the second-highest significant influence on ozone by decreasing it.

(iii) Two sample T Test for two different months.

In order to analyze whether there is a significant difference between these two months or not, two sample T tests are applied.

	5 vs 6	5 vs 7	5 vs 8	5 vs 9	6 vs 7
P - value	0.4962	4.601e-05	0.000927	0.2634	0.002182
C.I.	(-21.40528, 10.76639)	(-50.62806, -19.35271)	(-55.96393, -15.78607)	(-20.327858, 5.681306)	(-47.55383, -11.78805)
C.I. width	-	31.27535	40.17786	-	35.76578
Decision	No significant difference	Significant Difference	Significant Difference	No Significant Difference	Significant Difference

	6 vs 8	6 vs 9	7 vs 8	7 vs 9	8 vs 9
P - value	0.00737	0.7937	0.9345	0.0007361	0.006314
C.I.	(-52.251922, -8.859189)	(-17.88127, 13.87360)	(-22.48373, 20.71450)	(12.26467, 43.06955)	(8.631324, 48.472124)
C.I. width	43.39273	-	-	30.80488	39.8408
Decision	Significant difference	No Significant Difference	No Significant Difference	Significant Difference	Significant Difference

From the table, it shows that June and August have the largest difference of mean ozone because of the widest confidence interval obtained. And since this confidence interval is negative, it suggests that the mean ozone of June is less than that of August by between 8.859189 and 52.251922 ppb (parts per billion).

On the other hand, July and September have the smallest difference of mean ozone because of the narrowest confidence interval obtained. And since this confidence interval is positive, it suggests that the mean ozone of July is more than that of September by between 12.26467 and 43.06955ppb (parts per billion).

And according to the large p-values, there are some months without significant difference of mean ozone, which is May vs June, May vs September, June vs September, and July vs August. It can also be observed on the confidence intervals, which all contain 0.

E. Interpretation and Reporting

Does Solar radiation, Average wind speed, Temperature, Month, Day have significant effects on Ozone?

In the final linear regression model, we removed the predictor variable of Day using stepwise selection, which solves the multicollinearity problem; the predictor variables affecting Ozone are Solar, Wind, Temperature, and Month. With the F test of multiple linear regression performed, the terms' coefficients are not all equal to zero, meaning that there is sufficient evidence to support a significant association between the predictor variables and the response variable, Ozone. An individual t-test is then applied for each predictor variable with the null hypothesis that there is a significant relationship (correlation) between the majority of each predictor variable and response variable (Ozone). The p-values obtained from the t-test for each predictor variable shows that solar radiation, average wind speed and temperature have significant effects on ozone because of small p-values obtained. However, the conclusion of p-value (0.05863) of month depends on which alpha value will be used. If the alpha is 0.05, then month has no significant association with mean ozone. If the alpha is 0.10, then month has significant association with mean ozone.

Which of these factors has the greatest influence on Ozone?

With the t-test of standardized regression coefficients performed on our final linear regression model, the factor with the greatest influence on ozone would be temperature, where it has the largest absolute value of the estimator compared to other predictor variables. An increase in temperature is correlated with an increase in ozone.

When the Ozone is high? Are there some characteristics to make high Ozone?

From the scatter plot for Ozone and Wind, according to the correlation value (-0.6125) obtained, it shows that there is a negative correlation, which means that there is a negative relationship between wind speeds and ozone, that is, the lower the wind speed, the higher the ozone. This conclusion is also the same as that from the scatter plot for Ozone and Wind.

From the boxplot for Ozone vs Month, it shows that the months of July and August have higher ozone compared to other months. Based on the information from the two dimensional plots, it shows that ozone has a high positive relationship with temperature. Meanwhile, temperature has a positive relationship with month since temperature is associated with month; from the variables of the two dimensional plots, high mean ozone almost occurs from early July to the end of August because of high temperature. However, high mean ozone is also found on May 30th, where this was just associated with the highest temperature in May.

We possibly can see that during the months where temperature is mainly high, such as in the summer months, there is a correlation that the ozone is also high. There is a relationship between temperature and the months and ozone. However, the high ozone is directly and significantly affected by temperature rather than month. For example, May 30th is a typical observation since the highest mean ozone occurred on this day that had the highest temperature.

Is there a relationship between Ozone and Solar radiation?

From the scatter plot between Ozone and Solar Radiation, it shows that Solar Radiation and Ozone do not have a significant relationship due to the low correlation between the two variables. However, solar radiation is used in our final linear regression model since it still has a relationship with ozone and the standardized regression coefficient is 0.1827, which means that solar radiation is more influential on ozone than month (-0.1174).

F. Conclusion

Based on our research on New York’s air quality data, we come to the conclusion that ozone increases as temperature increases, and ozone decreases as average wind speed increases. These conclusions can be also proved by the principles of natural phenomena, that is, ozone generates heat in the stratosphere [6], which is the reason why high ozone occurs in high temperature. On the other hand, high wind speed tends to disperse pollutants and can dilute ozone concentrations [7], which is the reason why low ozone occurs in high average wind speed.

In order to find the factors affecting ozone, the reduced model: $Ozone \sim Solar.R + Wind + Temp + Month$ is used. Since some assumptions are violated, the square-root transformation is needed. The square root transformation shows less residuals than the original model, and the assumptions are also held. Therefore, it was suitable for use in our analysis. Furthermore, the application of some scatter plots help us determine the correlation among different factors. From the plots, temperature is positively correlated with ozone; wind is negatively correlated with ozone. Then, based on the t-test of standardized regression, it shows that temperature has the most significant influence on ozone by increasing ozone while wind has the second significant influence on ozone by decreasing ozone. These results are the same as the previous scatter plots.

Moreover, in order to find which two months tend to have a significant difference on ozone, in-depth analysis of two sample tests and multiple variables plots are used. From the variables of the two dimensional plots, high mean ozone almost occurs from early July to the end of August because of high temperature. However, high mean ozone is also found on May 30th, where this was just associated with the highest temperature in May. In contrast, the low mean ozone almost occurs in May because of high average wind speed. In addition, from the two sample tests for two different months, the result shows that June and August have the largest difference of mean ozone, that is, mean ozone of June is less than that of August. On the other hand, May and July have the smallest difference of mean ozone. That is, the mean ozone of May is less than that of July.

However, some limitations exist in this research such as the time the data was taken. The measurement of the air quality was done in 1973. In about 48 years the weather has been affected by global warming that causes higher temperatures everywhere. Another limitation in this research is the specification of locations, such as New York in this research. Different locations around the world may have different ozone levels and may produce different results. Therefore, future research that includes more locations and more updated data can ensure a more accurate and persuasive analysis.

G. Air Quality game function

Now, let’s play a game with an air quality robot. Type the value in parts per billion (ppb) of ozone level for today (only numbers are accepted) and see what the robot will tell you about air quality of today.

```r
#Air Quality game function
fungame <-  function(x) {
  if (class(x) != "numeric")
    stop(’This operation can only be performed on number’)
  if (x < 0)
    stop(’Please type positive number’)
  if (x <= 50) {
    return("Good Air Quality! What a Wonderful Day!")
  } else if (x <= 70) {
    return("Moderate Air Quality")
  } else if (x <= 85) {
    return("Unhealthy Air Quality for Sensitive Groups")
  } else if (x <= 105) {
    return("Unhealthy Air Quality! Wear mask!")
  } else if (x <= 200) {
    return("Very Unhealthy Air Quality! Wear N95 mask!")
  } else {
    return("Hazardous Air Quality! Stay home!")
  }
}
```

H. References

(1) kassambara, “Linear Regression Assumptions and Diagnostics in R: Essentials”, Nov. 2018 http://www.sthda.com/english/articles/39-regression-model-diagnostics/161-linear-regression-assumptions-and-diagnostics-in-r-essentials/

(2) “Impact of Medical Intervention on Birth Rates- STA 141A Final Project (example1)”, Dec. 2019.

(3) “CODING SYSTEMS FOR CATEGORICAL VARIABLES IN REGRESSION ANALYSIS (UCLA)” https://stats.idre.ucla.edu/spss/faq/coding-systems-for-categorical-variables-in-regression-analysis-2/

(4) Environmental Protection Agency. (2021, May 21). Ozone Trends. EPA. https://www.epa.gov/air-trends/ozone-trends#:~:text=Air%20quality%20monitors%20measure%20concentrations,public%20health%20and%20t

(5) NASA. (n.d.). Nasa Ozone Watch: Ozone facts. NASA. https://ozonewatch.gsfc.nasa.gov/facts/.

(6) Allen, Jeannie, “Tango in the Atmosphere: Ozone and Climate Change”, Feb. 2004 https://www.giss.nasa.gov/research/features/200402_tango/

(7) Ozone Concentrations. EPA. https://www.epa.gov/report-environment

I. Appendix Code

```
#plot for full model
the.data<-airquality
full_model<-lm(Ozone~Solar.R+Wind+Temp+Month+Day,data=the.data)
par(mfrow=c(2,2))
plot(full_model)
high_leverage<-which(hatvalues(full_model)>0.03921569)
estimate = c(-58.05384,0.04960,-3.31651,1.87087,-2.99163)
se = c(22.97114,0.02346,0.64579,0.27363,1.51592)
tval = c(-2.527,2.114,-5.136,6.837,-1.973)
p = c("0.0130*","0.0368*","1.29e-06***","5.34e-10***","0.0510 .")
AIC = c(680.21,682.98,703.2,719.94,682.4)
lmsummary = data.frame(A = estimate, B = se, C = tval, D = p, E = AIC)
rownames(lmsummary) = c("(intercept)","Solar.R","Wind","Temp","Month")
colnames(lmsummary) = c("Estimate","Std. Error","T-Value","Pr(>|t|)","AIC")
panderOptions(’table.split.table’, Inf)
pander(pandoc.table(lmsummary),style = ’rmarkdown’)
```

```r
#summary for full model
summary(full_model)
#Stepwise Regression
library(MASS)
step.model <- stepAIC(full_model, direction = "both",
                      trace = FALSE)
summary(step.model)

# Check AIC
drop1(full_model, test="F")
#log Transformation
red_logmodel<-lm(log(Ozone)~Solar.R+Wind+Temp+Month,data=the.data)
summary(red_logmodel)
#plot for log reduced model
par(mfrow=c(2,2))
plot(red_logmodel)
high_leverage<-which(hatvalues(red_logmodel)>0.03267974)

#Square Transformation
red_sqmodel<-lm(sqrt(Ozone)~Solar.R+Wind+Temp+Month,data=the.data)
summary(red_sqmodel)
#plot for Square root reduced model
par(mfrow=c(2,2))
plot(red_sqmodel)
#outlier
the.data[117,]
library(ggplot2)
library(dplyr)
newdata<-na.omit(the.data)
#scatter plot (Ozone vs Solar.R)
ggplot(data = newdata, mapping = aes(x =Solar.R, y =Ozone)) +
  geom_point()+
  geom_smooth()+
  labs(title = "Scatter plot of Ozone vs Solar.R",x = "Solar.R",y = "Ozone")
#correlation (Ozone vs Solar.R)
cor(newdata$Ozone,newdata$Solar.R)

#scatter plot (Ozone vs Wind)
ggplot(data = newdata, mapping = aes(x =Wind, y =Ozone)) +
  geom_point()+
  geom_smooth() +
  labs(title = "Scatter plot of Ozone vs Wind",x = "Wind",y = "Ozone")
#correlation (Ozone vs Wind)
cor(newdata$Ozone,newdata$Wind)

#scatter plot (Ozone vs Temp)
ggplot(data = newdata, mapping = aes(x =Temp, y =Ozone)) +
  geom_point()+
  geom_smooth() +
  labs(title = "Scatter plot of Ozone vs Temp",x = "Temp",y = "Ozone")
#correlation (Ozone vs Temp)
cor(newdata$Ozone,newdata$Temp)
```

```r
#boxplot (Ozone vs Month)
may <-newdata %>% filter(Month == 5)
jun <-newdata %>% filter(Month == 6)
jul <-newdata %>% filter(Month == 7)
aug <-newdata %>% filter(Month == 8)
sep <-newdata %>% filter(Month == 9)

ggplot(newdata, aes(x=factor(Month), y=Ozone))+
  geom_boxplot(data=may)+
  geom_boxplot(data=jun)+
  geom_boxplot(data=jul)+
  geom_boxplot(data=aug)+
  geom_boxplot(data=sep) +
  geom_jitter(width=0.15, alpha=0.5)+
  scale_y_log10()+
  theme(legend.position="none")+
  labs(title = "Boxplot of Month vs Ozone",x = "Month",y = "Ozone") +
  coord_flip()
figure <- ggarrange(A, B, C,D,
                    labels = c("A", "B", "C","D"),
                    ncol = 2, nrow = 2)
figure
graphnames = c("Ozone vs Solar.R","Ozone vs Wind","Ozone vs Temp")
corvalues = c(cor(newdata$Ozone,newdata$Solar.R),
              cor(newdata$Ozone,newdata$Wind),cor(newdata$Ozone,newdata$Temp))
P2 = data.frame(A = graphnames, B = corvalues)
colnames(P2) = c("Scatter Plot","Correlation value")
panderOptions(’table.split.table’, Inf)
pander(pandoc.table(P2),style = ’rmarkdown’)
#Three and more varaiables of Two dimensional plot
library(car)
scatterplotMatrix(formula = ~ Ozone+Solar.R+Wind+Temp+Month+Day ,
                  data = newdata, reg.line = lm, smooth = TRUE, span = 0.5,
                  diagonal = "histogram")
which(hatvalues(red_sqmodel)>0.03267974)

#Ozone vs Temp, Month and Day Plot
ggplot(newdata, aes(Temp, Ozone)) +
geom_point(aes(color = as.factor(Month))) +
  geom_point(size = 1, shape = 1, data = newdata) +
  ggrepel::geom_label_repel(aes(label = Day), data = newdata)+
  labs(title = "Ozone vs Temp, Month and Day Plot",
       x = "Maximum daily temperature (Fahrenheit)",
       y = "Mean ozone",
       colour = "Month")

#Ozone vs Wind, Month and Day Plot
ggplot(newdata, aes(Wind, Ozone)) +
  geom_point(aes(color = as.factor(Month))) +
  geom_point(size = 1, shape = 1, data = newdata) +
  ggrepel::geom_label_repel(aes(label = Day), data = newdata)+
  labs(title = "Ozone vs Wind, Month and Day Plot",
       x = "Average wind speed (mph)",
       y = "Mean ozone",
       colour = "Month")
v56 = c("0.4962","(-21.40528, 10.76639)"," - ","No significant difference")
v57 = c("4.601e-05","(-50.62806, -19.35271)","31.27535","Significant Difference")
v58 = c("0.000927","(-55.96393, -15.78607)","40.17786","Significant Difference")
v59 = c("0.2634","(-20.327858, 5.681306)"," - ", "No Significant Difference")
v67 = c("0.002182","(-47.55383, -11.78805)","35.76578","Significant Difference")
ttest1 = data.frame(A = v56,B = v57, C = v58, D = v59, E = v67)
colnames(ttest1) = c("5 vs 6","5 vs 7","5 vs 8","5 vs 9","6 vs 7")
rownames(ttest1) = c("P - value","C.I.","C.I. width","Decision")
panderOptions(’table.split.table’, Inf)
pander(pandoc.table(ttest1),style = ’rmarkdown’)

v68 = c("0.00737","(-52.251922, -8.859189)"," 43.39273 ","Significant difference")
v69 = c("0.7937","(-17.88127, 13.87360)"," - ","No Significant Difference")
v78 = c("0.9345","(-22.48373, 20.71450)"," - ","No Significant Difference")
v79 = c("0.0007361","(12.26467, 43.06955)"," 30.80488 ", "Significant Difference")
v89 = c("0.006314","(8.631324, 48.472124)","39.8408","Significant Difference")
ttest2 = data.frame(A = v68,B = v69, C = v78, D = v79, E = v89)
colnames(ttest2) = c("6 vs 8","6 vs 9","7 vs 8","7 vs 9","8 vs 9")
rownames(ttest2) = c("P - value","C.I.","C.I. width","Decision")
panderOptions(’table.split.table’, Inf)
pander(pandoc.table(ttest2),style = ’rmarkdown’)

#summary square root reduced model
summary(red_sqmodel)

#Standardized Regression coefficients
red_stand<-lm(scale(sqrt(Ozone))~scale(Solar.R)+scale(Wind)+
                scale(Temp)+scale(Month),newdata)
summary(red_stand)


#(5,6) 0.4962 CI:(-21.40528, 10.76639) no significant diff.
t.test(may$Ozone,jun$Ozone, alternative = "two.sided")

#(5,7) 4.601e-05 CI:(-50.62806, -19.35271) CI width:31.27535
t.test(may$Ozone,jul$Ozone, alternative = "two.sided")

#(5,8) 0.000927 CI:(-55.96393, -15.78607) CI width: 40.17786
t.test(may$Ozone,aug$Ozone, alternative = "two.sided")

#(5,9) 0.2634 CI:(-20.327858, 5.681306) no significant diff.
t.test(may$Ozone,sep$Ozone, alternative = "two.sided")

#(6,7)  0.002182 CI:(-47.55383, -11.78805) CI width: 35.76578
t.test(jun$Ozone,jul$Ozone, alternative = "two.sided")

#(6,8)  0.00737 CI:(-52.251922, -8.859189) CI width:43.39273
t.test(jun$Ozone,aug$Ozone, alternative = "two.sided")

#(6,9)  0.7937 CI:( -17.88127 , 13.87360) no significant diff.
t.test(jun$Ozone,sep$Ozone, alternative = "two.sided")
```

```
#(7,8)  0.9345 CI:(-22.48373,  20.71450)  no significant diff.
t.test(jul$Ozone,aug$Ozone, alternative = "two.sided")

#(7,9)  0.0007361 CI:(12.26467, 43.06955)  CI width:30.80488
t.test(jul$Ozone,sep$Ozone, alternative = "two.sided")

#(8,9)  0.006314 CI:(8.631324, 48.472124) CI width: 39.8408
t.test(aug$Ozone,sep$Ozone, alternative = "two.sided")

#Air Quality game function
fungame <-  function(x) {
  if (class(x) != "numeric")
    stop(’This operation can only be performed on number’)
  if (x < 0)
    stop(’Please type positive number’)
  if (x <= 50) {
    return("Good Air Quality! What a Wonderful Day!")
  } else if (x <= 70) {
    return("Moderate Air Quality")
  } else if (x <= 85) {
    return("Unhealthy Air Quality for Sensitive Groups")
  } else if (x <= 105) {
    return("Unhealthy Air Quality! Wear mask!")
  } else if (x <= 200) {
    return("Very Unhealthy Air Quality! Wear N95 mask!")
  } else {
    return("Hazardous Air Quality! Stay home!")
  }
}
```