HOUWEI CAO
CSCI 436/636

NYC CLIMATE PATTERNS

OZONE VS TEMPERATURE 2013-2018

PRESENTED BY: FERNANDA TOVAR, 1090913
AEMUN AHMAR, 1047508
NEHA BALA,1133176
ARPIT BATTU, 1129990

COLLEGE OF ENGINEERING AND COMPUTING SCEINCES
BIG DATA MANAGEMNET AND ANALYTICS COURSE PROJECT FALL 2019

TABLE OF CONTENTS

NYC CLIMATE PATTERNS

ABSTRACT

Climate change is defined as the average weather in a place over many years. In particular, it relates to the drastic change that shows up from the mid to late 20th century onwards. This is due to the large levels of atmospheric carbon dioxide produced by the use of fossil fuels. Air pollutants contribute to climate change by affecting the amount of incoming sunlight that is reflected or absorbed by the atmosphere. Interestingly enough the global average temperature has risen by about 0.89°C over the period 1901 to 2012.

Therefore, an increase in air pollutants lead to global warming. There is a hypothetical theory with data to prove it. There are many impacts to the Environment and different factors that are changing due to climate change. For example, the temperature is fluctuating and the pollutant ozone as well.

TARGET PROBLEM

In this project, we accumulate data to show the correlation between an increasing temperature and degrading air quality due to the depletion of the ozone. As well, with this we can create a predictable model that predicts the ozone based on a temperature**.**

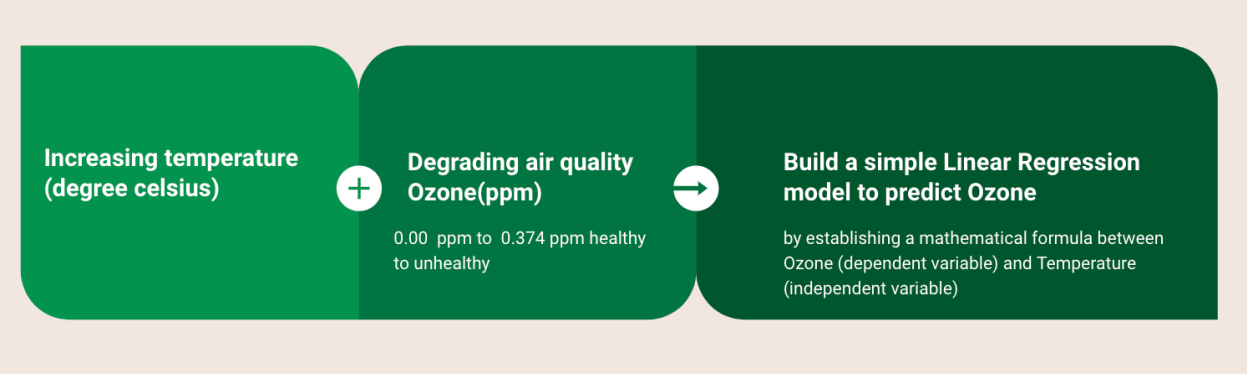


FIGURE 1 GOALS OF THE PROJECT

DEFINTIONS

1. ***Temperature****:* is a degree of hotness or coldness, which can be measured using a thermometer. It's also a measure of how fast the atoms and molecules of a substance are moving. It is measured in degrees on the Fahrenheit, Celsius, and Kelvin scales.
2. ***Ozone:*** is a colorless gas composed of three atoms of oxygen. Ozone forms both in the Earth's upper atmosphere and at the surface. Good ozone naturally forms in a layer about 10 - 30 miles (16 - 48 km) above Earth's surface. This protective layer shields us from the sun's harmful ultraviolet rays. Bad ozone forms near Earth's surface when the ultraviolet light in sunlight triggers a chemical reaction with "precursor pollutants" emitted by cars, power plants, and industrial sources. This creates holes in the ozone and exposes us to more rays from the sun.

3. ***Air Quality Index:*** The Environmental Protection Agency uses this to provide general information to the public about air quality and associated health effects. An Air Quality Index (AQI) of 100 for any pollutant corresponds to the level needed to violate the federal health standard for that pollutant. For ozone, an AQI of 100 corresponds to 0.08 parts per million (ppm) over an 8-hour period -- the current federal standard. Over half of the U.S. population lives in areas where the AQI exceeds 100 and violates the federal health standard at least once per year. Some metropolitan cities have severe air pollution problems and can see ozone AQI values in the 200s or even 300s.

0.00 – 0.060 ppm	Good	No health impacts are expected
0.061 – 0.075 ppm	Moderate	Unusually sensitive people should consider limited prolonged outdoor exertion
0.076 – 0.104 ppm	Unhealthy for Sensitive Groups	Active children and adults, and people with respiratory conditions (e.g., asthma) should limit prolonged outdoor exertion
0.105 – 0.115 ppm	Unhealthy	Active children and adults, and people with respiratory conditions (e.g., asthma) should avoid prolonged outdoor exertion. Everyone else, especially children and elderly, should limit prolonged outdoor exertion
0.116 – 0.374 ppm	Very Unhealthy	Active children and adults, and people with respiratory conditions (e.g., asthma) should avoid all outdoor exertion. Everyone else, especially children and elderly, should limit outdoor exertion

FIGURE 2 AIR QUALITY INDEX

4. ***Linear Regression Model:*** attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable. The overall idea of regression is to examine two things: (1) does a set of predictor variables do a good job in predicting an outcome (dependent) variable? (2) Which variables in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable?
5. ***Scatter Plot:*** show the relationship between two variables.
6. Line of best fit: gives a more exact understanding of the relationship by picking the best slope and y-intercept to fit the data.
7. ***Box plot:*** are a good way to spot any outlier observations in the variable. Having outliers can affect the predictions. It indicates distribution by dividing data into quartiles. Any dots outside of the box and whisker structure entirely are outliers, so no outliers.
8. ***Density Plot:*** visualizes the distribution of data over a continuous interval or time period. The peaks of a Density Plot help display where values are concentrated over the interval. Ideally, a close to normal distribution (a bell-shaped curve), without being skewed to the left or right is preferred.
9. ***Correlation:*** Is statistical measure that shows the degree of linear dependence between two variables. If one variable consistently increases with increasing value of the other, then they have a strong positive correlation (value close to +1).
10. ***P-value:*** is the probability of obtaining the observed results of a test, assuming that the null hypothesis is correct. It is the level of marginal significance within a statistical hypothesis test representing the probability of the occurrence of a given event. The p-value is used as an alternative to rejection points to provide the smallest level of

significance at which the null hypothesis would be rejected. A smaller p-value means that there is stronger evidence in favor of the alternative hypothesis.

EXAMPLES

People don't believe in climate change. They need data and graphs to see the truth in order for change to occur. Just because it's getting hotter or colder, doesn't mean climate change doesn't exist. We can't only look at the temperature changes, but we need to look at other factors such as the ozone to show proof.

MOTIVATION

The motivation behind this project is to learn, but as well shine light on the dangers that this presents. This slight in temperature is causing drastic changes like floods, melting polar caps and hurricanes. The increase in ozone means that poor air quality creates a critical health risk. The contaminated air can cause respiratory illness for all of us.

APPLICATIONS

This can be applied to many areas like predicting the climate change and showing the concerns. As well, collecting more factors of weather like humidity and precipitation, we can do weather forecast.

RELATED WORK

1. **Using Historical Weather Data to Investigate the Climate Change impact on Hurricane Barry By: Sam Helwig**

   Citation: Helwig, S. (2019, October 3). Using Historical Weather Data to Investigate the Climate Change impact on Hurricane Barry. Retrieved December 14, 2019, from https://www.visualcrossing.com/blog/climate-change-impact-hurricane-barry.

   a. This reference discusses how looking at Hurricane Barry, which occurred during the 2019 hurricane season helped discover some interesting findings. Although this specific hurricane was small, the data from this can help understand the impacts of severe weather. Therefore by "predicting the future", the public can be prepared more. As well, they can come up with solutions, so they won't get affected so much by the impacts of the severe storms.
   b. During Hurricane Barry, the Mighty Mississippi was at flood stage with nowhere to go but over the banks and levees. This could possible inundated New Orleans, which is extremely dangerous. A longer-term trend of heavy rainfall due to climate change that could change the habitats of Louisiana's fragile wildlife and change the ecosystem of the Mississippi.
   c. In figure 3 , it shows their platform that aggregated factors of temperature to see the trend line. In the figure, we see temperature in years. Temperature increases could worsen the spread of tropical diseases and lead to more frequent flooding rainfall events. This relates to our project, because we also want to prove climate change, not only with temperature, but with ozone as well.
      i.

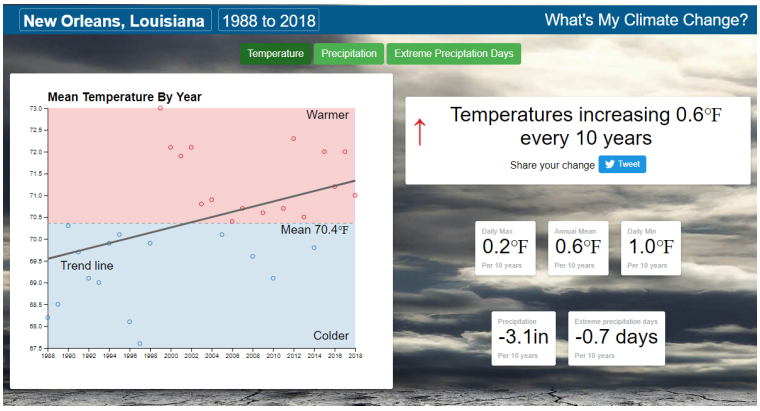


FIGURE 3 TEMPERATURE IN NEW ORLEANS

2. Impact of Climate Change on Ambient Ozone Level and Mortality in Southeastern United States. By: Howard H. Chang, Jingwen Zhou, and Montserrat Fuentes

Citation: Chang, H. H., Zhou, J., & Fuentes, M. (2010, July 14). Impact of climate change on ambient ozone level and mortality in southeastern United States. Retrieved December 14, 2019, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2922733/.

a. This reference discusses the health impacts of climate change. The paper looks at the potential dangers of future ozone levels on death patterns across states in the southeastern United States. They had a modeling framework that consisted of data from climate model outputs, historical meteorology and ozone observations and a health surveillance database.
b. They collected three pieces of vital data. The first was present-day relationships between observed maximum daily 8- hour average ozone concentrations and meteorology measured during the year 2000. Then they continue this by then finding the future ozone concentrations and then projecting those using calibrated climate model output data from the North American Regional Climate Change Assessment Program. The second data had Daily community-level mortality counts for the period 1987 to 2000 were obtained from the National Mortality, Morbidity and Air Pollution Study. The third was temperature, dew-point temperature, and seasonality, relative risks associated with short-term exposure to ambient ozone during the summer months.
c. This relates to our project because we also wanted to find the correlation between the temperature and ozone, but instead of doing it for many states in the southeastern area of the USA, we focused on NYC. They also looked at more variables and retrieve a lot more data from more studies done previously. In the figure 4, you see total cloud-cover explained considerably less variation in ozone concentration compared to temperature and solar radiation. Proving that temperature has an impact on ozone levels.

i.

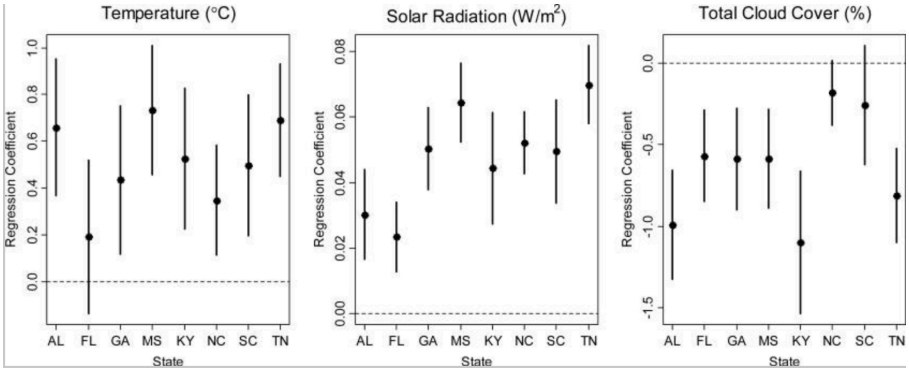


FIGURE 4 PARAMETER ESTIMATES FOR THE MODEL OF DAILY MAXIMUM 8-HOUR OZONE CONCENTRATION IN 2000

3. Trends in Ozone Adjusted for Weather Conditions By: Environmental Protection Agency

Citation: Trends in Ozone Adjusted for Weather Conditions. (2019, July 12). Retrieved December 14, 2019, from https://www.epa.gov/air-trends/trends-ozone-adjusted-weather-conditions.

a. In this reference, EPA uses a statistical model to adjust for the variability in seasonal ozone concentrations due to weather to provide a more accurate assessment of the primary trend in ozone caused by emissions. In figure 5, the dotted red lines show the trend in observed ozone concentrations and the solid blue lines show the underlying ozone trend after removing the effects of the weather. The solid blue lines represent ozone levels anticipated under average weather conditions and serve as a more accurate assessment of the trend in ozone due to changes in precursor emissions.
b. This relates to our project because we also created a statistical model, but in our case, we did the temperate vs ozone instead of just focusing on all weather aspects.

i.

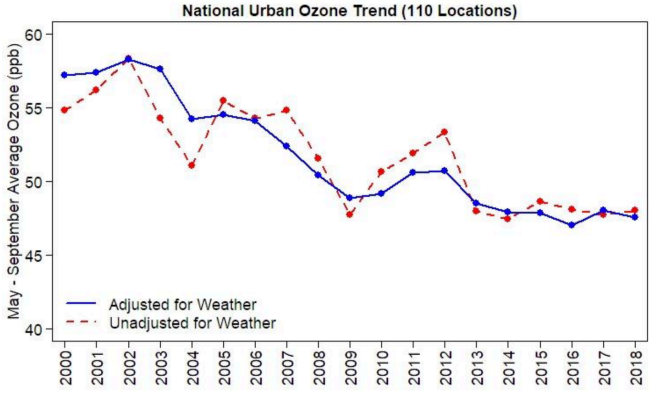


FIGURE 5 NATIONAL TRENDS IN THE MAY - SEPTEMBER AVERAGE OF THE DAILY MAXIMUM 8-HOUR OZONE CONCENTRATIONS FROM 2000 TO 2018 IN URBAN LOCATIONS

4. Stagnant Air on the Rise, Upping Ozone Risk

Citation: Stagnant Air on the Rise, Upping Ozone Risk. (2016, August 17). Retrieved December 14, 2019, from https://www.climatecentral.org/news/stagnation-air-conditions-on-the-rise-20600.

a. In this reference, they show that if air is stagnant and there is little air circulation, hot weather can trigger high levels of air pollution that can have health consequences for millions of Americans. With stagnant air now occurring more frequently in much of the country, and projected to continue increasing, the combination of heat and stagnant air are primed to counteract efforts to reduce ground-level ozone pollution and continue to put thousands of lives at risk every year. This is shown in figure 6.
b. This relates to our project because they also wantt to show that there is a correlation between the ozone and temperature. They specifically looking how high temperature amplifies the stagnate air.

i.

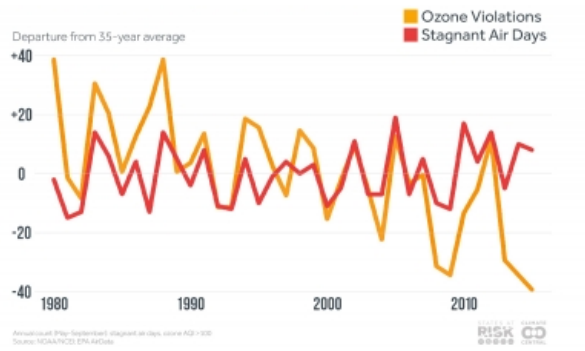


FIGURE 6 ST.LOUIS' STAGNANT AIR DAYS AND OZONE VIOLATIONS

5. Predicted Impacts of Climate Change on Ground Level Ozone in Cities in the Western United States By: Elizabeth Jayne Dresselhaus, Richard Wagner, Scott Landolt

Citation: Dresselhaus, E., Wagner, R., & Landolt, S. (n.d.). Predicted Impacts of Climate Change on Ground Level Ozone in Cities in the Western United States. Retrieved December 14, 2014, from https://opensky.ucar.edu/islandora/object/manuscripts:836/datastream/PDF/view.

a. The results of this study show that ozone levels tend to rise with increasing temperatures and that to expect to see more summer days in the future with dangerous ozone levels in most cities. The study also suggests that the effects of climate change on air quality will vary somewhat between climates and that some geographic areas are more vulnerable to ozone increase with temperature in the future.
b. This relates to our project because of the correlation they find between temperature and ozone. They find that temperature is a statistically significant measurement.

DATASET(S)

DESCRIPTION OF DATASETS

Temperature Dataset: To retrieve 6 years of temperature data in the NYC area, we used the weather query builder page on Visual Crossing website. To collect this data, it wasn't free, so we had to sign up for a free trial to access it. We had to input the year and the location. This was done until all CSV files from year 2013-2018 was downloaded. Each CSV file contained several columns of weather information, but for this project we only were interested in temperature.

Address	Date	Minimum Temperature	Maximum Temperature	Temperature	Dew Point	Relative Humidity	Heat Index	Wind Speed	Wind Gust	Wind Direction	Precipitation	Precipitation Cover	Snow Depth	Visibility	Cloud Cover	Sea Level Pressure	Weather Type	Latitude	Longitude	Resolved Address	Info
10023	01/01/2013	26.9	39	36.5	22.2	55.94		18.1	34.4	283.62	0	0		10	66.1	1012.3	Light Rain	40.775921	-73.982607	10023, USA	
10023	01/02/2013	21.9	32.1	27.2	10.9	49.99		15.9	31.1	308.42	0	0		10	1.7	1017.7		40.775921	-73.982607	10023, USA	
10023	01/03/2013	24.2	32.1	28.6	14.3	55.26		13.5		286.71	0	0		10	23.5	1020.3		40.775921	-73.982607	10023, USA	
10023	01/04/2013	30.9	37.2	34.1	19.4	54.7		16.1	31.1	260.5	0	0		10	40.8	1016.4		40.775921	-73.982607	10023, USA	
10023	01/05/2013	32.1	42.2	36.5	18.9	48.84		13.9	33.3	284.12	0	0		10	16.9	1022		40.775921	-73.982607	10023, USA	

FIGURE 7 TEMPERATURE 2013 (SHOWING ONLY A FEW ROWS)

Ozone Dataset: To retrieve 6 years of Ozone data in the NYC area, we used the Download daily data tool on the United States Environmental Protection Agency website. We had to input the choice of pollutant (ozone), year, geographic location, and sect monitor sites. This was done until all CSV files from year 2013-2018 was downloaded. Each CSV file contained several columns of information, but for this project we only were interested in Daily Max 8-hour Ozone Concentration.

Date	Source	Site ID	POC	Daily Max 8-hour Ozone Concentration	UNITS	DAILY_AQI_VALUE	Site Name	DAILY_OBS_COUNT	PERCENT_COMPLETE	AQS_PARAMETER_CODE	AQS_PARAMETER_DESC	CBSA_CODE	CBSA_NAME	STATE_CODE	STATE	COUNTY_CODE	COUNTY	SITE_LATITUDE	SITE_LONGITUDE
01/01/2013	AQS	360610135	1	0.021	ppm	19	CCNY	17	100	44201	Ozone	35620	New York-Newark-Jersey City, NY-NJ-PA	36	New York	61	New York	40.81976	-73.94825
01/02/2013	AQS	360610135	1	0.022	ppm	20	CCNY	17	100	44201	Ozone	35620	New York-Newark-Jersey City, NY-NJ-PA	36	New York	61	New York	40.81976	-73.94825
01/03/2013	AQS	360610135	1	0.018	ppm	17	CCNY	17	100	44201	Ozone	35620	New York-Newark-Jersey City, NY-NJ-PA	36	New York	61	New York	40.81976	-73.94825

FIGURE 8 OZONE 2013 (SHOWING ONLY A FEW ROWS)

ISSUES COLLECTING DATA

One issue that we encountered is that since we needed to find two different sources because there wasn't one that included both temperature and ozone, there can be a possibility that the recorded data isn't exact. The two different sources could have recorded the measurements differently and at different times of the day. Another issue was having an efficient way of aggregating the data to end up with a clean dataset that will give us the results we wanted too. The first step that was certain was to remove the other unnecessary columns. The next part, we thought it would be better to find the mean of each month for each file (year) for both ozone and temperature. After this then we can use those values to get the linear regression model. Unfortunately, it would have been better to just skip the mean calculation step and use the daily values to get more accurate data for the linear regression model.

REPRESENTATIVE EXAMPLES

Like mentioned before, since we broke up the file with the according months, the rows that included dates that started and ended the month were the representative examples that helped us calculate the mean. By doing this also, we eliminate those interesting values because doing the mean will get an average instead of having some outlier values that can affect the linear regression model.

APPROACH

The general overview of the approach was to first load in data then clean up the data, apply the functions to then visualize data, run analysis and output results. The software that we used was R studio. This platform is ideal for statistical analysis environment to load this weather data. To be more specific, R studio contains lm(), which is a statistical function known as linear Regression Model. This is supervised modeling technique perfect for continuous data. At the start, to get use to and to learn how to work with this software, we apply the mean function to only one csv file. After we found a way to recursively, go through all the files so then steps mentioned previously could be repeated for all corresponding databases. According to the reference, the only slight new contribution is the location that we picked to find the corresponding data.

In the first script called " finalozone7.R" , we focused on chaning the date column and calculating the mean per month for all Ozone csv files in a recursive manner. The first step was to called the libararies "plyr" to split, apply and combine data. The other library "dplyr" was used to work with data frame like objects, both in memory and out of memory.The next step was to set the directory to then list all the files and then read them. After that, we create a fucntion that contains the changes we want to make to apply them to all csv files. In this function, we first select the columns that include data and ozone and then we convert the date column to a date format using the as. Date function. Now the data, is being read as a date, we can manipulate it to format it so we can have a column with just the month in it. In order to compare the different dates, we have to make sure that the date has the correct data type. We need a number and not a character. The next three steps will be to find the mean by splitting the dataframe in correct arreas which is determine by comparing the month column. At the end, we create a new dataframe that consist of the Ozone Mean by Month. These are saved as csv files with the format "ozone_mean_by_month i csv"; the i representing 1- 6. The same steps were applied in a different script called "temp. R" to do the same changes to the temperature files. Th An example of the results of this script is is shown in figure 9.

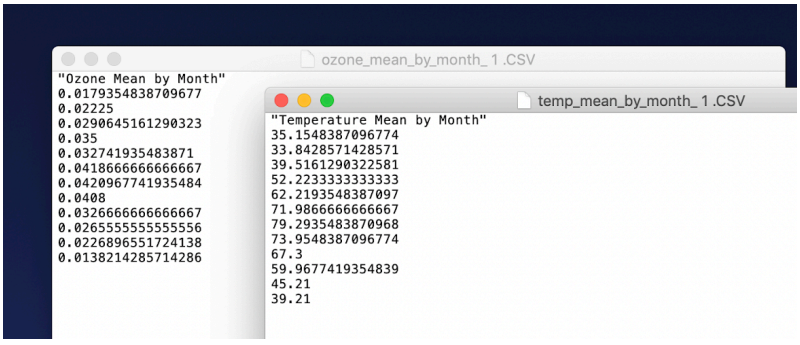


FIGURE 9 OZONE & TEMP MEAN BY MONTH

In the third script called "combine2.R", we focused on combing the according matching temperature and ozone csv files years together. This creates a new csv file with a column for temperature and column for ozone for each year. The first step to make this happen was to set the path to make sure the 6 csv files created go to this place. Now to combine the information for each year we do "cbind()" and the we write this to a csv file. We repeat the steps for each year. The results from this script are shown in figure 10. We discussed the last script in the Experiments and Results section.

combinedOzoneTemp Ozone_vs_Temperature_2013.CSV

```
"Temperature.Mean.by.Month","Ozone.Mean.by.Month"
35.1548387096774,0.0179354838709677
33.8428571428571,0.02225
39.5161290322581,0.0290645161290323
52.2233333333333,0.035
62.2193548387097,0.032741935483871
71.9866666666667,0.0418666666666667
79.2935483870968,0.0420967741935484
73.9548387096774,0.0408
67.3,0.0326666666666667
59.9677419354839,0.0265555555555556
45.21,0.0226896551724138
39.21,0.0138214285714286
```

FIGURE 10 COMBINED OZONE AND TEMP FOR 2013

EXPERIMENTS AND RESULTS

ALGORITHM/ METHODOLOGY

Like mentioned before, we apply the linear Regression Model to predict the ozone by establishing a mathematical formula between ozone (dependent variable) and temperature (independent variable). Before we do this exactly, it's good practice to analyze and understand the variables. The goal is to determine if the variable is statistically significant to be able to perform and see the relationship between the ozone and temperature. The graphical analysis and correlation study below will help with this by visualizing the patterns.

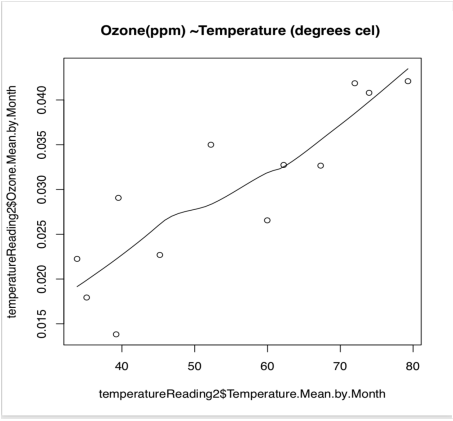


FIGURE 11 SCATTER PLOT

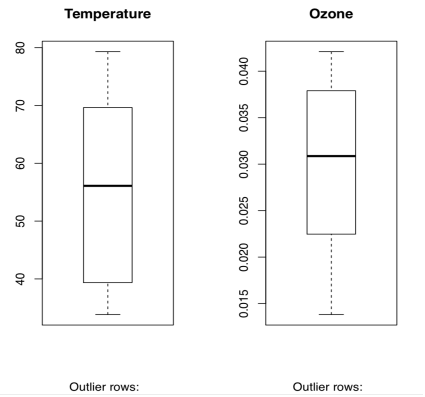


FIGURE 11 BOX PLOT

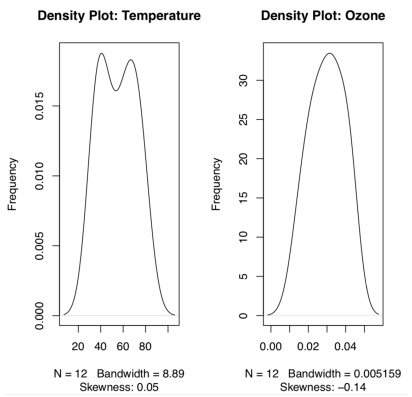


FIGURE 13 DENSITY PLOT

In the last script called “linearRegresion.R”, we broke it up into three different functions. The first function applies the graphical analysis part. With simple RStudio commands we can do the scatter, box, density plots for each csv file.

```
#-------------------------------------------------graphical analyis-----------------------------------------------#
#analyze the filenamee
analyze <- function(filenamee) {
  #read each csv file
  tempoz <- read.csv(file = filenamee, header = TRUE)
  # ----scatterplot----#
  scatt <-scatter.smooth(x=tempoz$Temperature.Mean.by.Month, y=tempoz$ Ozone.Mean.by.Month, main="Ozone(ppm) ~Temperature (degrees cel) ")
  print(scatt)
  # ----box plot---- #
  # divide graph area in 2 columns
  par(mfrow=c(1, 2))
  # box plot for 'temp'
  box1 <- boxplot(tempoz$Temperature.Mean.by.Month, main="Temperature", sub=paste("Outlier rows: ", boxplot.stats(tempoz$Temperature.Mean.by.Month)$out))
  print(box1)
  # box plot for 'ozone'
  box2 <-boxplot(tempoz$ Ozone.Mean.by.Month, main="Ozone", sub=paste("Outlier rows: ", boxplot.stats(tempoz$ Ozone.Mean.by.Month)$out))
  print(box2)
  # ----density plot---- #
  # for skewness function
  library(e1071)
  # divide graph area in 2 columns
  par(mfrow=c(1, 2))
  # density plot for 'temp'
  poly1 <- plot(density(tempoz$Temperature.Mean.by.Month), main="Density Plot: Temperature", ylab="Frequency",
                sub=paste("Skewness:", round(e1071::skewness(tempoz$Temperature.Mean.by.Month), 2)))
  print(poly1)
  # density plot for 'ozone'
  poly2 <-plot(density(tempoz$ Ozone.Mean.by.Month), main="Density Plot: Ozone", ylab="Frequency",
               sub=paste("Skewness:", round(e1071::skewness(tempoz$ Ozone.Mean.by.Month), 2)))
  print(poly2)
}
```

FIGURE 12 CODE FOR GRAPHICAL ANALYSIS

The second function consists of writing csv files with summary statistics and correlation information for each input csv file. The cor() function in RStudio, does the correlation analysis. The lm () functions builds the linear regression model to get the formula and then we do the summary() to get several different statistics.

```
#-------------------------------------------------summarystats/correlation text files-----------------------------------------------#
analyze3 <- function(filenamee){
  tempoz <- read.csv(file = filenamee, header = TRUE)
  #----correlation analysis----#
  print("#---------------Correlation Analysis---------------------#")
  c <- cor(tempoz$Temperature.Mean.by.Month, tempoz$ Ozone.Mean.by.Month)
  print(c)

  #print title
  print("#---------------Building the Linear Regression Model---------------------#")
  #-----linear regression model----#
  linearMod <- lm(Ozone.Mean.by.Month ~ Temperature.Mean.by.Month , data=tempoz)  # build linear regression model on full data
  #print title
  print("Linear Regression model: ")
  print(linearMod)
  #------summary ----------#
  print("This is summary statistics of the linear Regression Model: ")
  summary(linearMod)
}
```

FIGURE 13 CODE FOR SUMMARY STATISTICS AND CORRELATION OF WHOLE DATASET

The last function is applied to the training and test data section. So far at this point, we just built a linear regression model using the whole dataset. If you build it that way, there is no way to tell how the model will perform with new data. So the preferred practice is to split your dataset into a 80:20 sample (training:test), then, build the model on the 80% sample and then use the model thus built to predict the dependent variable on test data. Doing it this way, we will have the model predicted values for the 20% data (test) as well as the actuals (from the original dataset). The first step was to set the seed to reproduce results of random sampling with set.seed() function in RStudio. Then we get the training row index by using sample(). After ,we print out the model training data with tempoz2[trainingRowIndex,] and we print the test data with tempoz2[-trainingRowIndex,]. Noticed the negative in front of traningRowIndex in order to get the test data. The test data is important because it should closely match the predict ozone values.

```
#-------------------------------------------------training and test data-----------------------------------------------#
analyze4 <-function(filenamee){
  tempoz2 <- read.csv(file = filenamee, header = TRUE)
  # ----Create Training and Test data----#
  print("#----------------------Create Training and Test data-----------------------------#")
  # setting seed to reproduce results of random sampling
  set.seed(100)

  #print title
  print("trainingRowIndex:")
  # row indices for training data and predict on test data
  trainingRowIndex <- sample(1:nrow(tempoz2), 0.8*nrow(tempoz2))
  #print trainingRowIndex results
  print(trainingRowIndex)

  #print title
  print("Model Training data:")
  # model training data
  trainingData <- tempoz2[trainingRowIndex, ]
  #print trainingData results
  print(trainingData)

  #print title
  print("Test Data:")
  # test data
  testData  <- tempoz2[-trainingRowIndex, ]
  #print testdata results
  print(testData)
```

FIGURE 14 CODE FOR TRAINING AND TESTING DATA

Now, we fit the model on the training data and predict the ozone on test data. Once again to do this we do lm() and summary(). To predict the ozone, we use predict() function and it takes into consideration the test data and the linear regression model.

```
# ----Fit the model on training data and predict dist on test data----#
print("#-----------Fit the model on training data and predict dist on test data---------#")
# print title
print("Model based on Training Data:")
# Build the model on training data
lmMod <- lm(Ozone.Mean.by.Month ~ Temperature.Mean.by.Month, data=trainingData)
# print lmMod results
print("lmMod results: ")
print(lmMod)
print("Summary staistics of lmMod: ")
#summary
summary(lmMod)

#print title
print("Predict Ozone:")
# predict ozone
ozonePred <- predict(lmMod, testData)
# print ozone predictions
print(ozonePred)
```

FIGURE 15 CODE FOR TRAINING/ TEST DATA AND PREDICTION

```
# ----Calculate prediction accuracy and error rates----#
print("#----------------------Calculate prediction accuracy and error rates----------------------------#")
# make actuals_predicteds dataframe.
actuals_preds <- data.frame(cbind(actuals=testData$Ozone.Mean.by.Month, predicteds=ozonePred))
#print title
print("Actual vs predictions: ")
#print the start values of actuals_preds
head(actuals_preds)

#print title
print("Correlation Accuracy for actual and predictions: ")
correlation_accuracy <- cor(actuals_preds)
#print results
print(correlation_accuracy)


# ----------------Min Max accuracy and MAPE-----------------#
print("#----------------------Min Max accuracy and MAPE----------------------------#")
#print title
print("Min-Max Accuracy: ")
# Min-Max Accuracy Calculation
min_max_accuracy <- mean(apply(actuals_preds, 1, min) / apply(actuals_preds, 1, max))
#print results
print(min_max_accuracy)

#print title
print("MAPE(mean absolute percentage deviation) Calculation: ")
# MAPE (mean absolute percentage deviation) Calculation
mape <- mean(abs((actuals_preds$predicteds - actuals_preds$actuals))/actuals_preds$actuals)
#print results
print(mape)
```

FIGURE 16 CODE FOR EVALUATION

Now we get to the evaluation part, for this we first want to create a table of the actuals values vs the predicted values. Then once again we do cor() to see the correlation of this new table. By calculating accuracy measures (like min_max accuracy) and error rates (MAPE or mean absolute percentage deviation), you can find out the prediction accuracy of the model.

EXPERIMENT

For this project, we did output all necessary graphs and txt files containing the statistics/predictions/evaluation, but for time purposes we just focused on the year 2013. To bring back the graphs shown in figures 11,12,13, we will analyze them to see if the data is optimal for the building of a linear Regression Model. For, the scatter plot along with the smoothing line it suggests a linear and positive relationship. For the box plot, we have no outlier rows according to there being no value following "outlier rows: ". Lastly, we have the density plot, that has close enough to a bell shaped curved and minor skewness values for each variable (ozone = -0.14 and density = 0.05). These graphs show that the values are useful for a prefect model. Now let's continue with the results that included the statistical evaluation in the next section.

PRESENTATIONS/DISCUSSIONS OF RESULTS

In Figure 15 , it shows the results from the snippet of R code from Figure 13. The first value printed is the Correlation Analysis, which is about 0.87 meaning that it's a strong positive correlation. The next result is the is the formula build from the lm(). We get ozone= 0.0019820+0.0005057*temperature. Now we look at the p-values from the summary() to see if its statistically significant. We look at the p-value of the individual predictor variables and also the overall p-value, which is 0.0002212. Since the P-value is less than 0.05, then it is good.

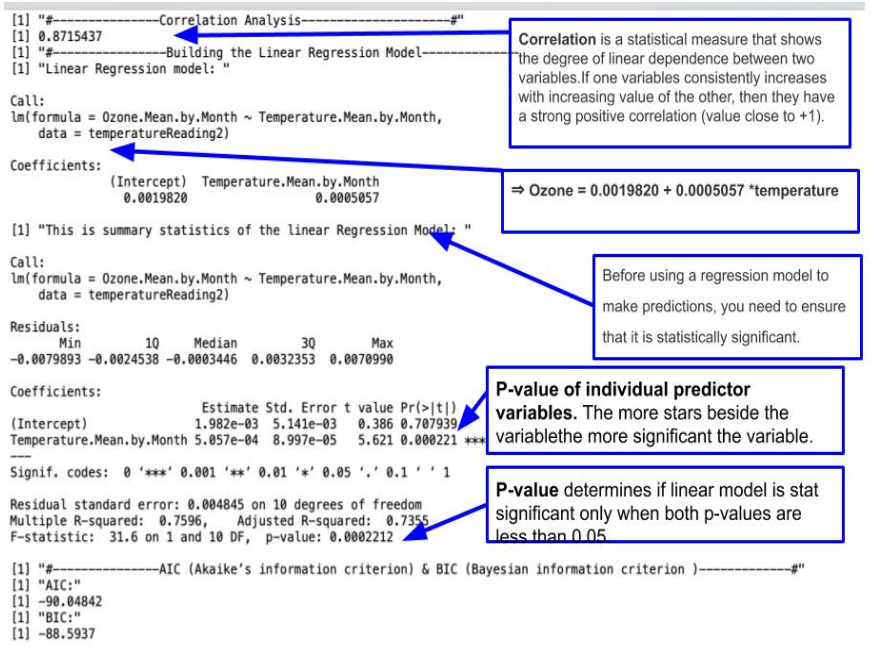
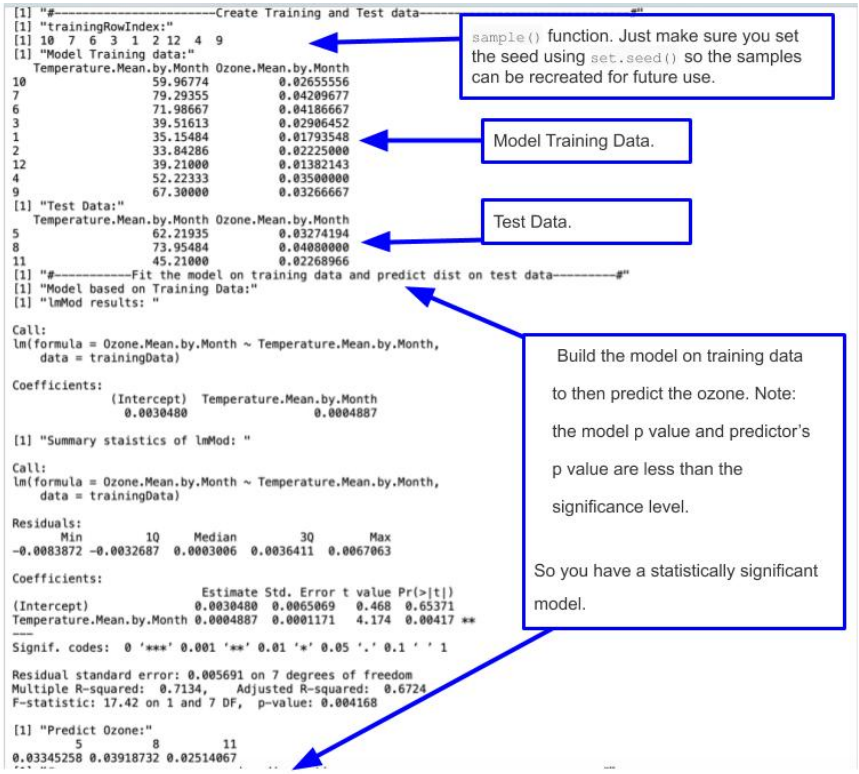


FIGURE 17 RESULTS SUMMARY STATISTICS AND CORRELATION INFORMATION FOR YEAR 2013



For the training and test data portion, we see that for the test data consisted of the months of 5, 8,11 were picked while the training data consisted of the months, 10, 7, 6,3,1,2,12,4,9. If we look at the test data and the predicted ozone values we see that they are quite similar, but not exact proving that the linear regression model worked. For example, for month 11 the actual value was 0.02268966, but the predicted was 0.02514067. Note that the p-values are less than the significance level therefore we have a statistically significant model.

FIGURE 18 RESULTS FOR PREDICTED OZONE

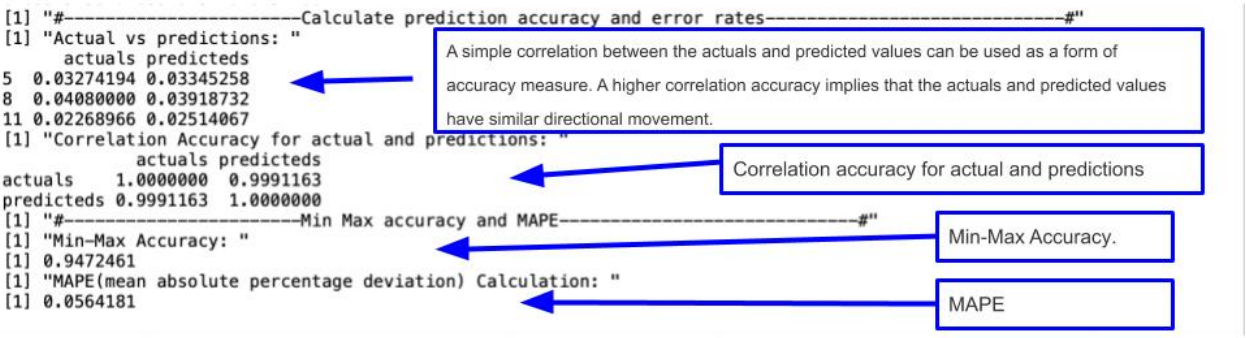


FIGURE 19 RESULTS FOR EVALUATION

Last but not least, we calculate the prediction accuracy and error rates. We can look at the accuracy of the actual versus the prediction values with the correlation values. A higher correlation accuracy implies that the actuals and predicted values have similar directional movement. As well, we have a min-max accuracy of 0.9472461 and a MAPE of 0.0564181. Both pretty high numbers.

CONCLUSIONS

MAIN CONTRIBUTIONS

- Learned to import data as CSV files.
- Learned the R language.
- Learned the statistics behind linear regression to match the data.
- Got each combined pair (ozone and temp) from year 2013-2018.
- Got plots and statistics summaries for each for year 2013-2018.
- Got the linear Regression model to work for Year 2013 and predicted correct values.

WHAT DIDN'T WORKED

As mentioned above, we thought it would be better to find the mean of each month for each file (year) for both ozone and temperature. After this then we can use those values to get the linear regression model. Unfortunately, it would have been better to just skip the mean calculation step and use the daily values to get more accurate data for the linear regression model.

CONSIDERATIONS FOR FUTURE WORK

- User-friendly Platform for user to input temperature and predict an ozone value for health reasons.
- Add more locations instead of just having NYC.
- Add more factors like precipitation, wind, humidity to see how they impact the climate change and can predict more accurate results.
- Use the linear Regression Model to predict future values like in the year 2020.
    - Aggregate in different way: get the mean temperature by adding each month from each year 2013-2018. With this, we can have the estimated temperature at each month for the year 2020
        - ex: January temperature in 2020 is the mean of each January in from each year 2013-2018.

CONTRIBUTIONS OF EACH TEAM MEMBER

- Arpit-collecting the data and cleaning it in RStudio; project proposal presentation
- Neha- collecting the data and cleaning it in RStudio; final report
- Fernanda- build linear regression model in RStudio; demo video
- Aemun – combined the data of the temperature and ozone for each year in RStudio; final project presentation