MECH481A6: Engineering Data Analysis in R

Chapter 5 Homework: Exploring Univariate Data

Ethan Rutledge

04 October, 2023

Grading

We will grade the **knitted** PDF or HTML document from within your private GitHub repository. Remember to make regular, small commits (e.g., at least one commit per question) to save your work. We will grade the latest knit, as long as it occurs *before* the start of the class in which we advance to the next chapter. As always, reach out with questions via GitHub Issues or during office hours.

Data

You are probably sick of seeing the ozone data, but there's still more to do with the file. Ozone concentration measurement is considered univariate, thus we can use basic exploratory data analysis approaches to examine the data.

Preparation

Load the necessary R packages into your R session.

Recreate the pipe of `dplyr` functions that you used to import the data, select and rename the variables listed below, drop missing observations, and assign the output with a good name.

- `sample_measurement` renamed as `ozone_ppm` (ozone measurement in ppm)
- `datetime` (date in YYYY-MM-DD format and time of measurement in HH:MM:SS)

Check that the data imported correctly.

```
##    ozone_ppm          datetime
##  Min.   :0.00000   Min.   :2019-01-01 07:00:00.00
##  1st Qu.:0.02300   1st Qu.:2019-03-30 21:00:00.00
##  Median :0.03300   Median :2019-06-27 12:30:00.00
##  Mean   :0.03305   Mean   :2019-07-01 06:37:00.78
##  3rd Qu.:0.04300   3rd Qu.:2019-10-03 22:45:00.00
##  Max.   :0.09600   Max.   :2020-01-01 06:00:00.00
```

```
## [1] 0
```

Chapter 5 Homework: Exploring Univariate Data

Through Question 5, you will use all of the available ozone measurements from January 2019 through January 2020. Starting in Question 6, you will use a subset of the dataset: ozone concentration measurements on July 4, 2019.

Question 1: Definitions

Guess the location, dispersion, and shape of ozone concentration data, based on the definitions of each described in the coursebook. No code needed; just use your intuition. For shape, take a look at the coursebook appendix on reference distributions.

```
Based on Homework 4, I would expect the data to be mostly located around 0.025 ozone concentration. The
```

Question 2: Quartiles

Calculate the quartiles of `ozone_ppm`. What is the minimum? Maximum? Median?

```
## [1] "Minimum: 0"
```

```
## [1] "Maximum: 0.096"
```

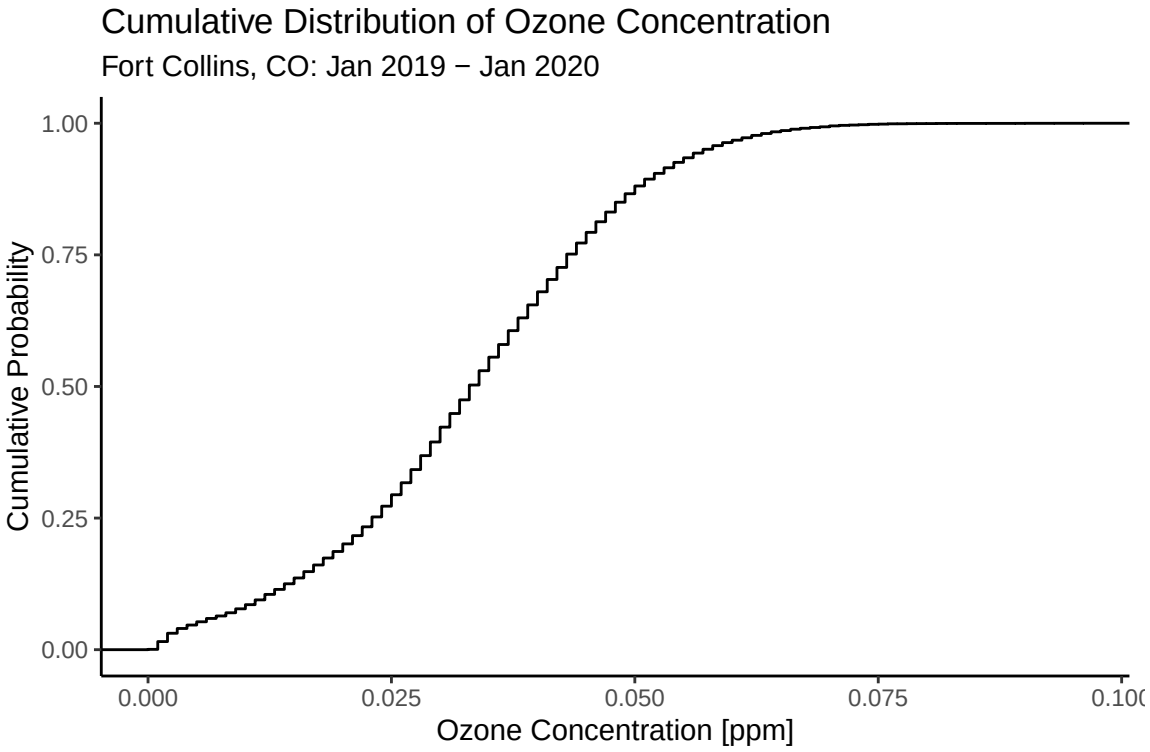
```
## [1] "Median: 0.033"
```

Extra Credit

Create a similar table for `ozone_ppm`. Hint: You will need to investigate table options in the `knitr` package.

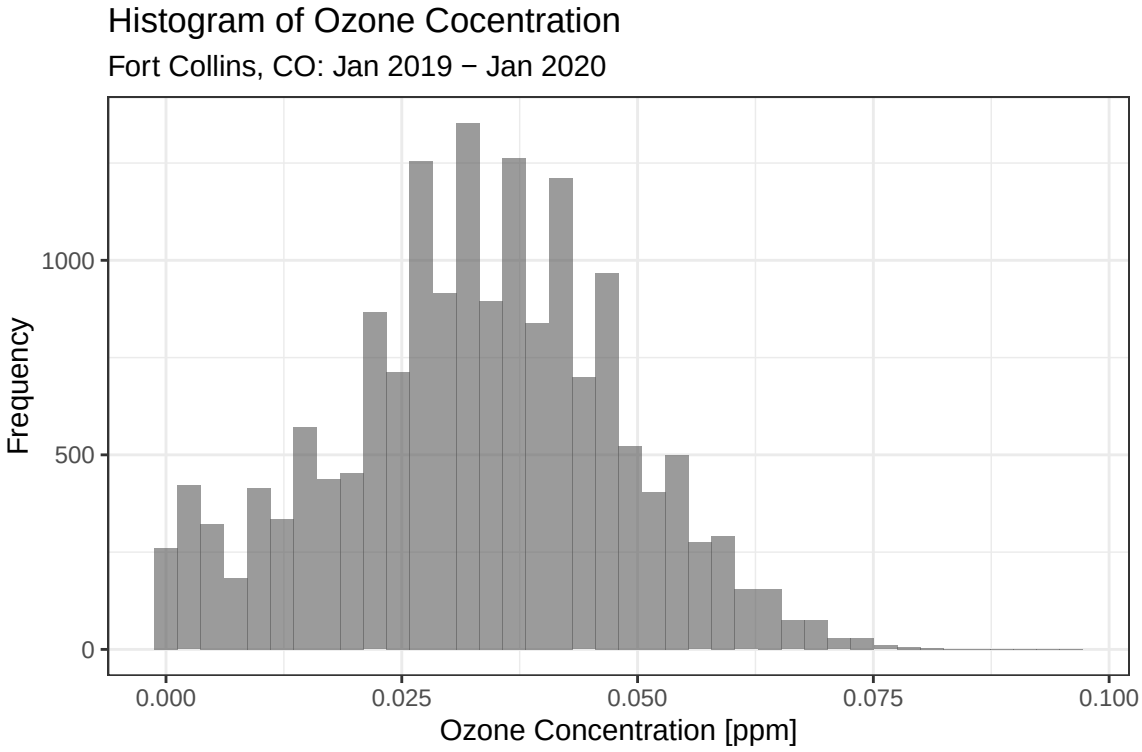
Question 3: Cumulative Distribution Plot

Using either relevant `ggplot2 geom` option, create a cumulative distribution plot of `ozone_ppm`. Tweak the axis ranges for optimal data representation, using `scale_*_continuous()` with `breaks =` and `minor_breaks =` arguments. Add axis labels, title, subtitle, and theme.



Question 4: Histogram

Create a histogram of `ozone_ppm`. Within the `geom`, mess with the number of bins (e.g., 20, 50, 75, 100, 200) to explore the true shape and granularity of the data. Match the plot style (e.g., title, subtitle, axis labels, theme) you chose in Question 3, with the relevant adjustments such as "Histogram" instead of "Cumulative Distribution Plot".



Question 5: Concept

What mathematical concept is a histogram (Q4) attempting to visualize?

```
It shows the how often a certain range of values shows up in a dataset giving you a visualized overview
```

Question 6: Distribution

Based on the histogram (Q4), does ozone concentration appear to be normally distributed?

```
It is not perfectly normal as it is shifted slightly to the right but it does resemble a normal distribu
```

Question 7: Outliers

Based on the histogram (Q4), do you see any possible outliers? Skewness? How might this affect the spread and central tendency?

```
Yes, it appears to have some outliers around 0.00 concentration, as well as along the right tail of the
```

Question 8: Boxplot

Generate a boxplot of ozone concentration on the y-axis with a title, subtitle, y-axis label, and theme consistent with the style of the previous two plots. Use quotes (`""`) as the `x` arguments within the calls to the aesthetic and labels to remove the x-axis scale and label.

Subset Data

Use the following code to create a dataframe for use in the remaining questions. These ozone concentration measurements were taken on July 4, 2019 in Fort Collins, CO. This code detects certain characters with the `datetime` object and filters to observations containing those characters. There are other ways this could have been done (e.g., `dplyr::filter()` with `%in%` operator).

Question 9: Autocorrelation Plot

Define autocorrelation as it relates to ozone concentration measurement.

Create an autocorrelation plot of ozone concentration, using `stats::acf()` and include axis labels and title. Describe what you see based on the features of interest outlined in the coursebook.

Question 10: Parial Autocorrelation Plot

Define partial autocorrelation as it relates to ozone concentration measurement.

Now create a partial autocorrelation plot of day ozone concentration with axis labels. Describe what you see. How does this compare to the autocorrelation plot in the previous question?

Appendix

```
# set global options for figures, code, warnings, and messages
knitr::opts_chunk$set(fig.width = 6, fig.height = 4, fig.path = "../figs/",
                      echo = FALSE, warning = FALSE, message = FALSE)
# load packages
library(tidyverse)
library(ggplot2)
library(dplyr)
# ozone: import, select, drop missing observations, rename
# use relative pathname
ozone_data <- read_csv("../data/ftc_o3.csv")%>%
 # select needed variables
  select(sample_measurement, datetime)%>%
 # drop missing observations
  na.omit()%>%
 # rename main variable
  rename(ozone_ppm = sample_measurement)
# examine dataframe object
summary(ozone_data)
sum(is.na(ozone_data$ozone_ppm))
# calculate quantiles of ozone concentration
seq <- seq(0,1,0.25)
quan <- quantile(ozone_data$ozone_ppm, prob = seq)

minimum <- quan[1]
maximum <- quan[5]
median <- quan[3]

print(paste0("Minimum: ",minimum))
print(paste0("Maximum: ",maximum))
print(paste0("Median: ",median))

# plot cumulative distribution of ozone concentration
ggplot(data=ozone_data,
        aes(x=ozone_ppm))+
  geom_step(stat = "ecdf")+
  labs(title = "Cumulative Distribution of Ozone Concentration",
       subtitle = "Fort Collins, CO: Jan 2019 - Jan 2020",
       x = "Ozone Concentration [ppm]",
       y = "Cumulative Probability") +
  theme_classic()

# create histogram of ozone concentration
ggplot(data = ozone_data, aes(x = ozone_ppm))+
  geom_histogram(bins = 40, alpha = 0.6) +
  labs(title = "Histogram of Ozone Cocentration",
       subtitle = "Fort Collins, CO: Jan 2019 - Jan 2020",
       x = "Ozone Concentration [ppm]",
       y = "Frequency") +
  theme_bw()
# create ozone boxplot
# create subset of data with only one day to examine daily pattern
```

```
# I did not ask you to code this because we have not discussed dates or stringr
# You need to uncomment the below three lines and run it; check object names
# ozone_day <- ozone_data %>%
#   dplyr::filter(stringr::str_detect(string = datetime,
#                                     pattern = "2019-07-04"))
# create autocorrelation plot with ozone_day df
# create partial autocorrelation plot
```