MECH481A6: Engineering Data Analysis in R

Chapter 5 Homework: Exploring Univariate Data

Student Name

25 September, 2024

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.

```
## tibble [16,914 x 2] (S3: tbl_df/tbl/data.frame)
##  $ ozone_ppm: num [1:16914] 0.017 0.017 0.017 0.017 0.015 0.017 0.028 0.03 0.036 0.036 ...
##  $ datetime : POSIXct[1:16914], format: "2019-01-01 07:00:00" "2019-01-01 08:00:00" ...
##  - attr(*, "na.action")= 'omit' Named int [1:606] 417 848 1183 1185 1208 1569 1834 1835 1836 1856 ..
##   ..- attr(*, "names")= chr [1:606] "417" "848" "1183" "1185" ...
```

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.

For the location of the data I am guessing that the central tendency will by around the middle of the year. For dispersion, I expect there to be a peak in the middle of the year and lower values in the beginning and tail end of the year. For shape, I think it will be a fairly uniform distribution with higher values on the latter 6 months of the year than on the front 6 months.

Question 2: Quartiles

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

```
##    0%   25%   50%   75%  100% 
## 0.000 0.023 0.033 0.043 0.096
```

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

```
## [1] 0.096
```

```
## [1] 0.033
```

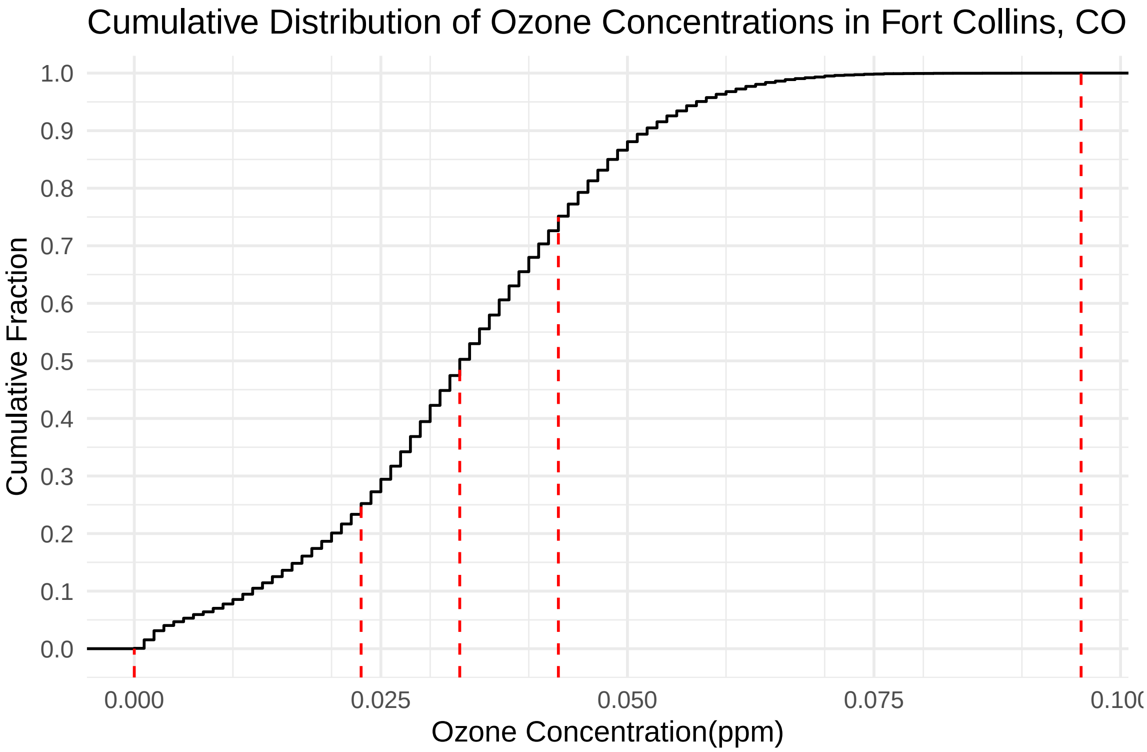
Extra Credit

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

Quantile	**Descriptor**	**Value**
0	Minimum	0.000
0.25	1st Quartile	0.023
0.5	Median	0.033
0.75	3rd Quartile	0.043
1	Maximum	0.096
IQR	Inter Quartile Range	0.020

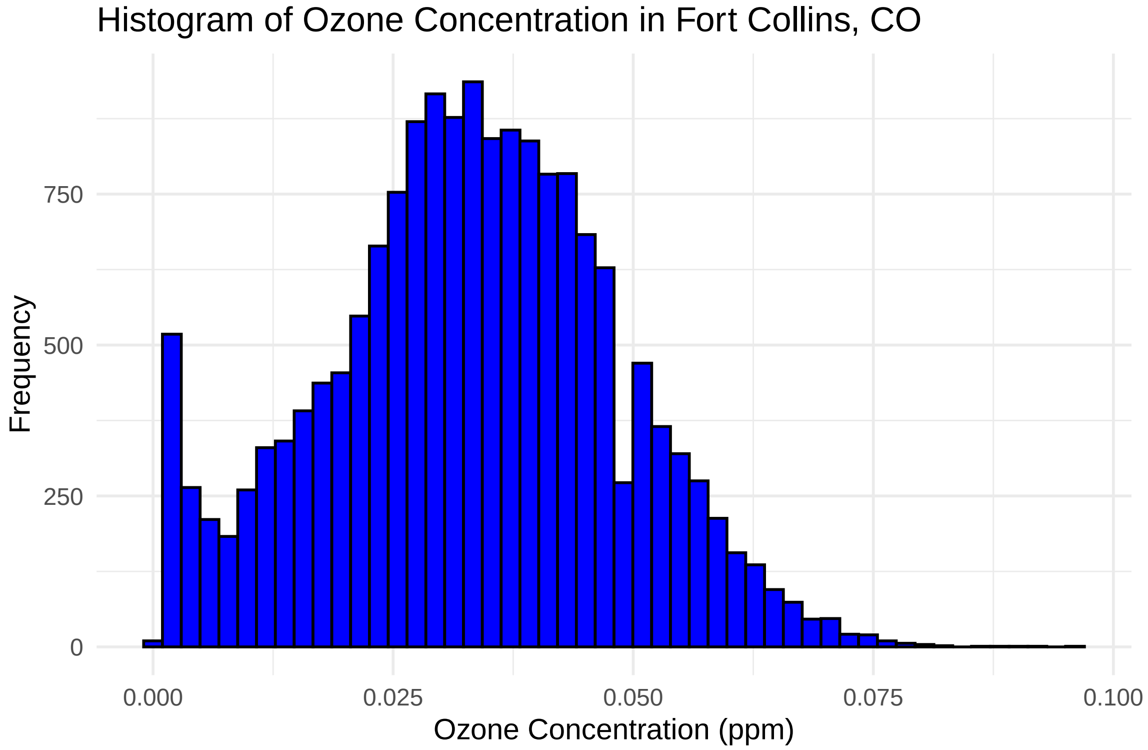
Question 3: Cumulative Distribution Plot

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?

Histograms are useful for showing the central tendency, range, dispersion, and general distribution of a data set.

Question 6: Distribution

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

The ozone concentration does not appear to be normally distributed.

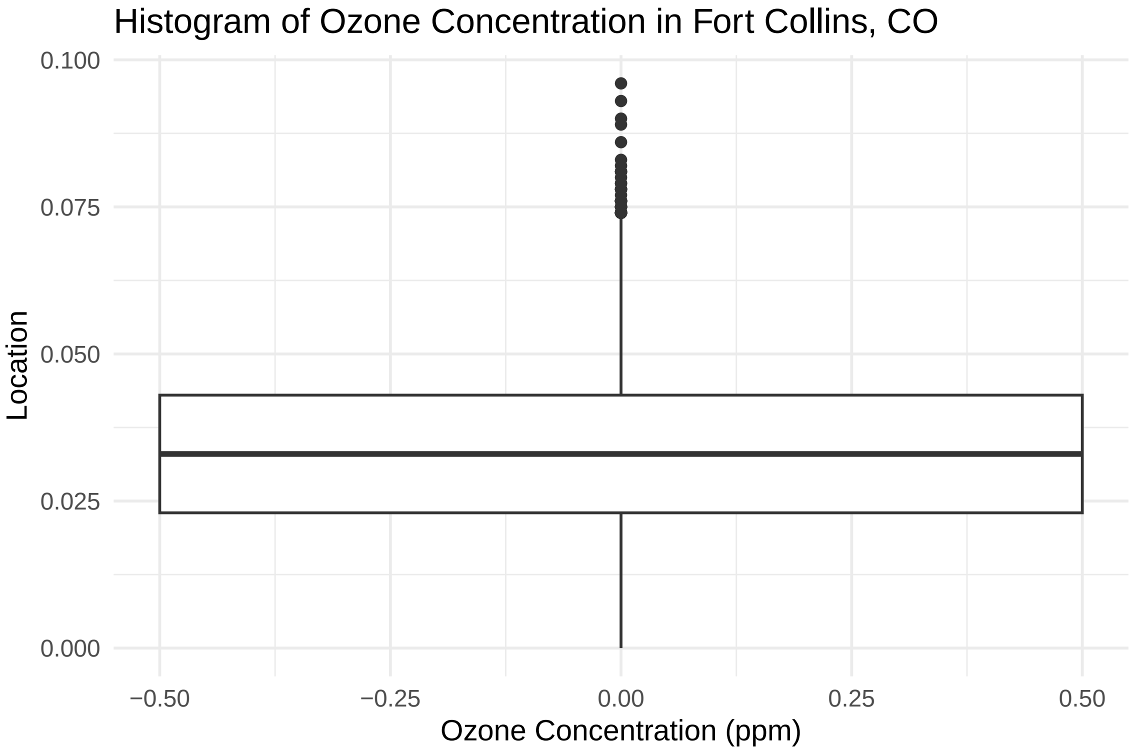
Question 7: Outliers

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

In the histogram you can see that there is skewness towards the lower side of the data. There is also some outliers on the higher end of the data. This would affect the spread by causing changes to the standard deviation. It would affect the central tendency because the mean will be shifted.

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.



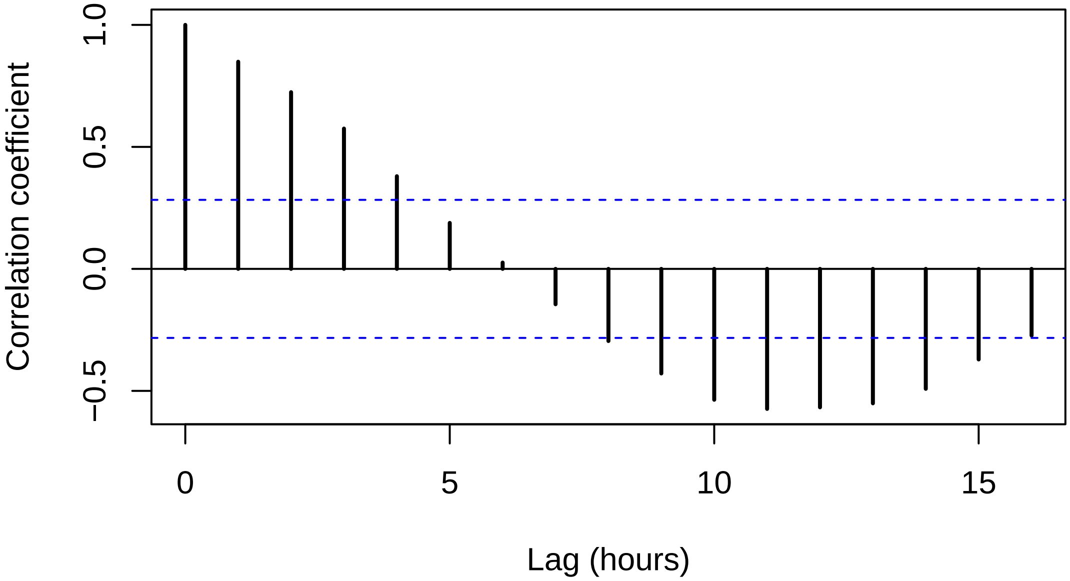
Subset Data

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.

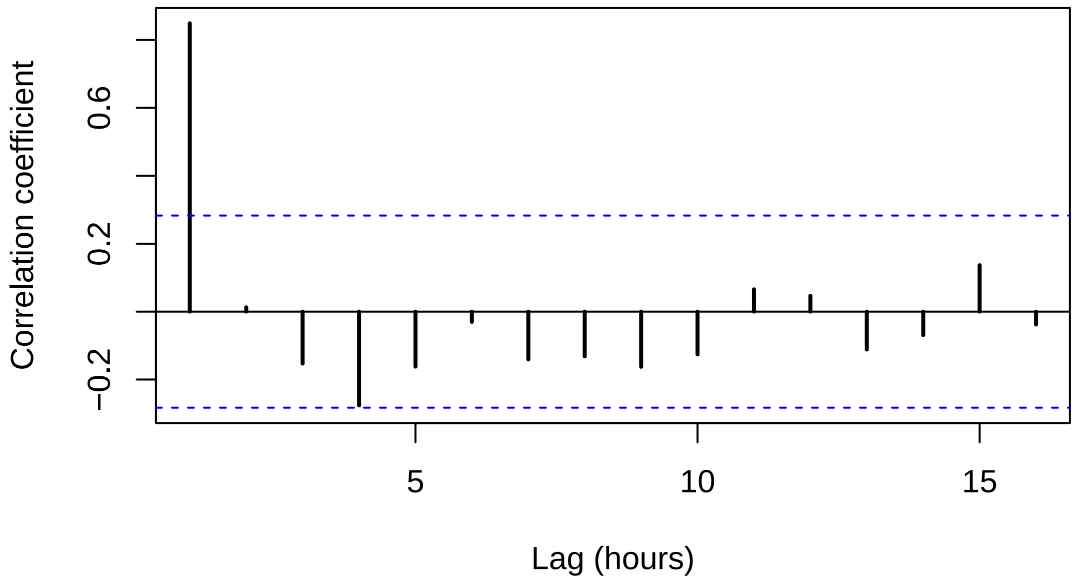
Create an autocorrelation plot of ozone concentration, using `stats::acf()` and include axis labels and title. Describe what you see based on the features of interest outlined in the coursebook.



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

```r
# set global options for figures, code, warnings, and messages
knitr::opts_chunk$set(fig.width = 6, fig.height = 4, fig.path = "../figs/",
                      echo = FALSE, warning = FALSE, message = FALSE)
# load packages
library(readr)
library(dplyr)
library(ggplot2)
library(ggpubr)
# ozone: import, select, drop missing observations, rename
ozone <- read_csv("./ftc_o3.csv") %>%
  select(sample_measurement, datetime) %>%
  na.omit() %>%
  rename(ozone_ppm = sample_measurement)
ozone_tibble <- as_tibble(ozone)
# examine dataframe object
str(ozone_tibble)
# calculate quantiles of ozone concentration
quantile(ozone_tibble$ozone_ppm, seq(0, 1, 0.25))
min(ozone_tibble$ozone_ppm)
max(ozone_tibble$ozone_ppm)
median(ozone_tibble$ozone_ppm)
values <- quantile(ozone_tibble$ozone_ppm, seq(0,1,0.25))
iqr <- IQR(ozone_tibble$ozone_ppm)
descriptors <- c("Minimum", "1st Quartile", "Median", "3rd Quartile", "Maximum", "Inter Quartile Range"

table <- data.frame(
  Quantile = c(0, 0.25, 0.5, 0.75, 1, "IQR"),
  Descriptor = descriptors,
  Value = c(values, iqr)
)

ggtexttable(table, rows = NULL, theme = ttheme("minimal"))
ordered_ozone <- ozone_tibble %>%
  dplyr::select(ozone_ppm) %>%
  dplyr::arrange(ozone_ppm) %>%
  dplyr::mutate(cum_pct = seq.int(from = 1/length(ozone_ppm),
                                  to = 1,
                                  by = 1/length(ozone_ppm)))

ordered_ozone %>%
  ggplot2::ggplot(mapping = aes(x = ozone_ppm)) +
  geom_step(stat = "ecdf") +
  labs(x = "Ozone Concentration(ppm)", y = "Cumulative Fraction", title = "Cumulative Distribution of O
  scale_y_continuous(limits = c(-0.05, 1.03),
                     expand = c(0,0),
                     breaks = seq(from = 0,
                                  to = 1,
                                  by = 0.1)) +
  scale_x_continuous(minor_breaks = seq(from = 0,
                                        to = 0.1,
                                        by = 0.01))+
  geom_segment(data = data.frame(x = quantile(ordered_ozone$ozone_ppm),
                                 y = rep.int(-.05, 5),
                                 xend = quantile(ordered_ozone$ozone_ppm),
                                 yend = seq(from = 0, to = 1, by = 0.25)),
               aes(x = x, y = y, xend = xend, yend = yend),
               color = "red",
               linetype = "dashed") +
  theme_minimal()
# create histogram of ozone concentration
ggplot(ordered_ozone, aes(x = ozone_ppm)) +
  geom_histogram(bins = 50, fill = "blue", color = "black") +
  labs(title = "Histogram of Ozone Concentration in Fort Collins, CO", x = "Ozone Concentration (ppm)",
  theme_minimal()
# create ozone boxplot
ggplot(data = ordered_ozone,
       mapping = aes(y = ozone_ppm)) +
  geom_boxplot(width = 1) +
  labs(y = "Location",
       x = "Ozone Concentration (ppm)",
       title = ("Histogram of Ozone Concentration in Fort Collins, CO")) +
  theme_minimal()
# 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_tibble %>%
   dplyr::filter(stringr::str_detect(string = datetime,
                                     pattern = "2019-07-04"))
# create autocorrelation plot with ozone_day df
stats::acf(ozone_day$ozone_ppm,
           main = " " ,
           xlab = "Lag (hours)",
           ylab = "Correlation coefficient",
           lwd = 2)
# create partial autocorrelation plot
stats::pacf(ozone_day$ozone_ppm,
           main = " " ,
           xlab = "Lag (hours)",
           ylab = "Correlation coefficient",
           lwd = 2)
```