MECH481A6: Engineering Data Analysis in R

Chapter 5 Homework: Exploring Univariate Data

Michael Thill

11 December, 2025

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.

```
## # A tibble: 6 x 2
##   datetime            ozone_ppm
##   <dttm>                  <dbl>
## 1 2019-01-01 07:00:00     0.017
## 2 2019-01-01 08:00:00     0.017
## 3 2019-01-01 09:00:00     0.017
## 4 2019-01-01 10:00:00     0.017
## 5 2019-01-01 11:00:00     0.015
## 6 2019-01-01 12:00:00     0.017
```

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.

Question 2: Quartiles

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

```
## # A tibble: 5 x 2
##   stat   value
##   <chr>  <dbl>
## 1 Min    0
## 2 Q1     0.023
## 3 Median 0.033
## 4 Q3     0.043
## 5 Max    0.096
```

Extra Credit

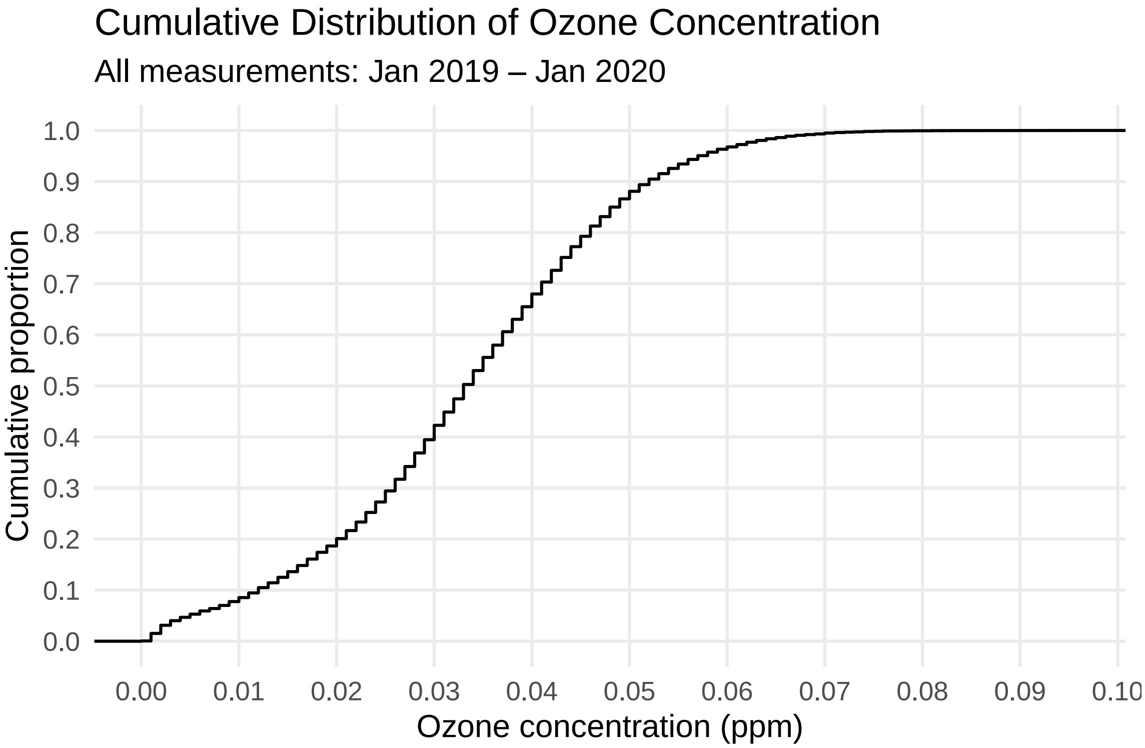
Create a similar table for `ozone_ppm`. Hint: You will need to investigate table options in the `knitr` package.

Table 1: Quantile Descriptors and Values for Ozone Concentration (ppm)

Quantile	Descriptor	Values
0	minimum	0.000
1	maximum	0.096
0.5	median	0.033
0.25	25th%	0.023
0.75	75th%	0.043
0.75 - 0.25	IQR	0.020

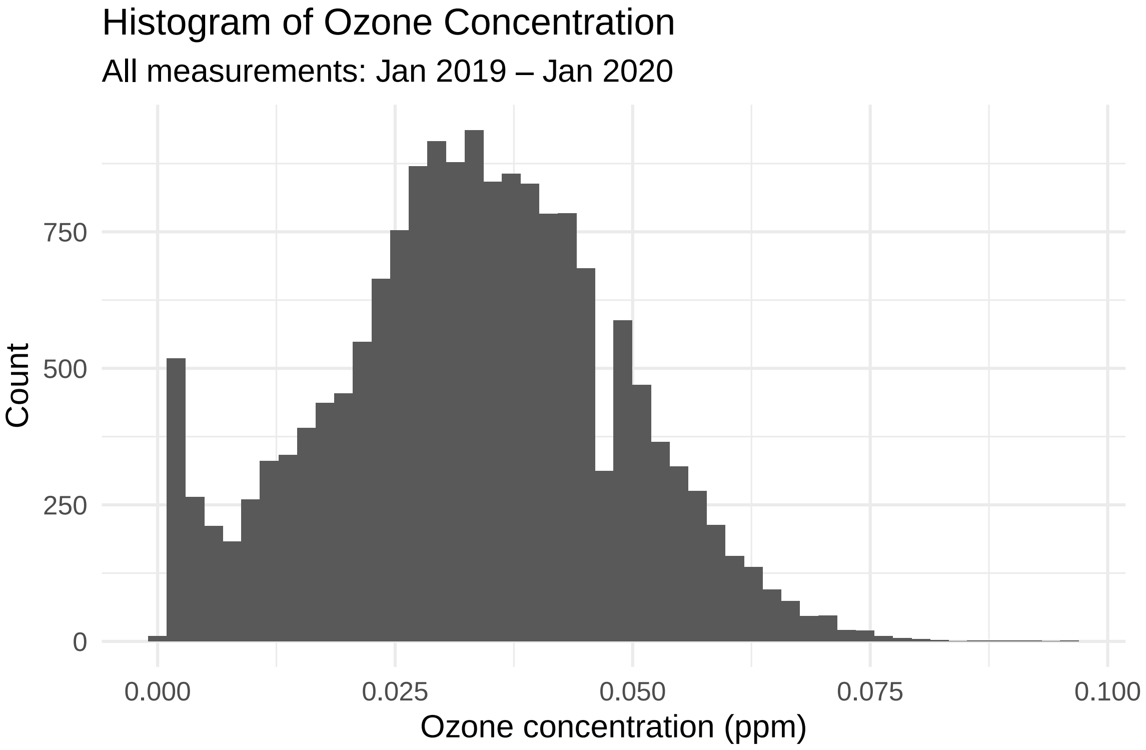
Question 3: Cumulative Distribution Plot

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?

*** The Histogram is attempting to visualize the distrobution of a quantitative variable of Ozone concentrations for this specific example *** ### Question 6: Distribution

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

*** based on the histogram above there is not a normalized distribution instead there is a right skewed uni modal distribution indicating ozone concentrations are typically low or moderate with occasionally high values. ***

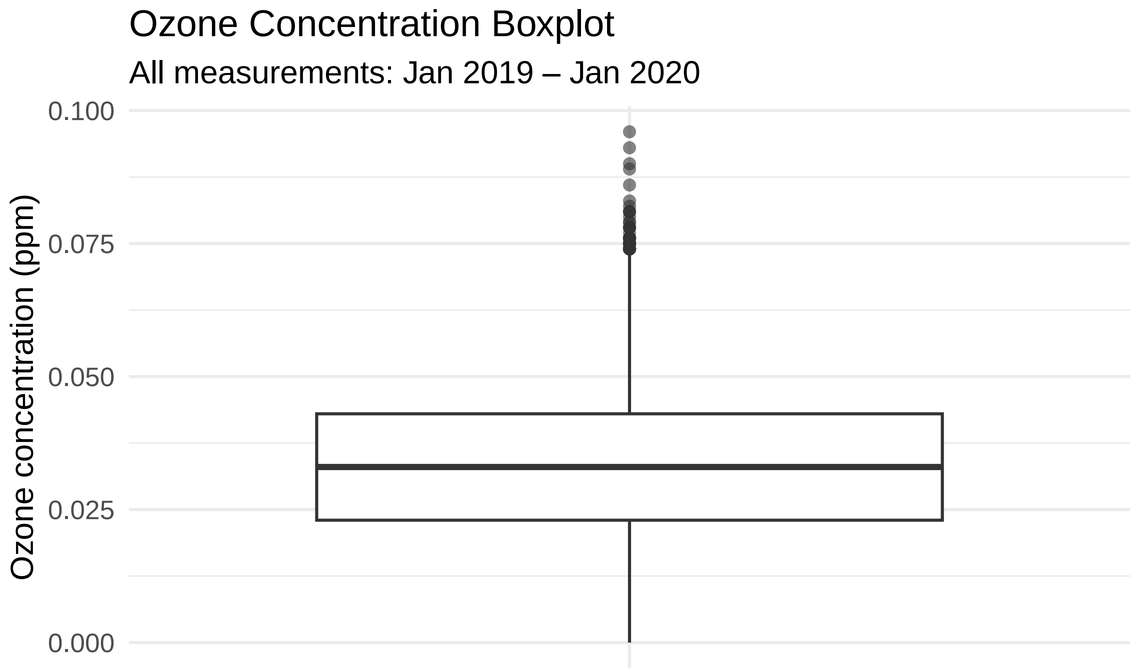
Question 7: Outliers

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

*** based on the histogram we can see a spike in values outside of the standard .025 - .045 that is closer to .06 and a few more following closely after this. this cluster of larger values does p[l[ay a rolling skewing the center of the distribution to the right ***

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.



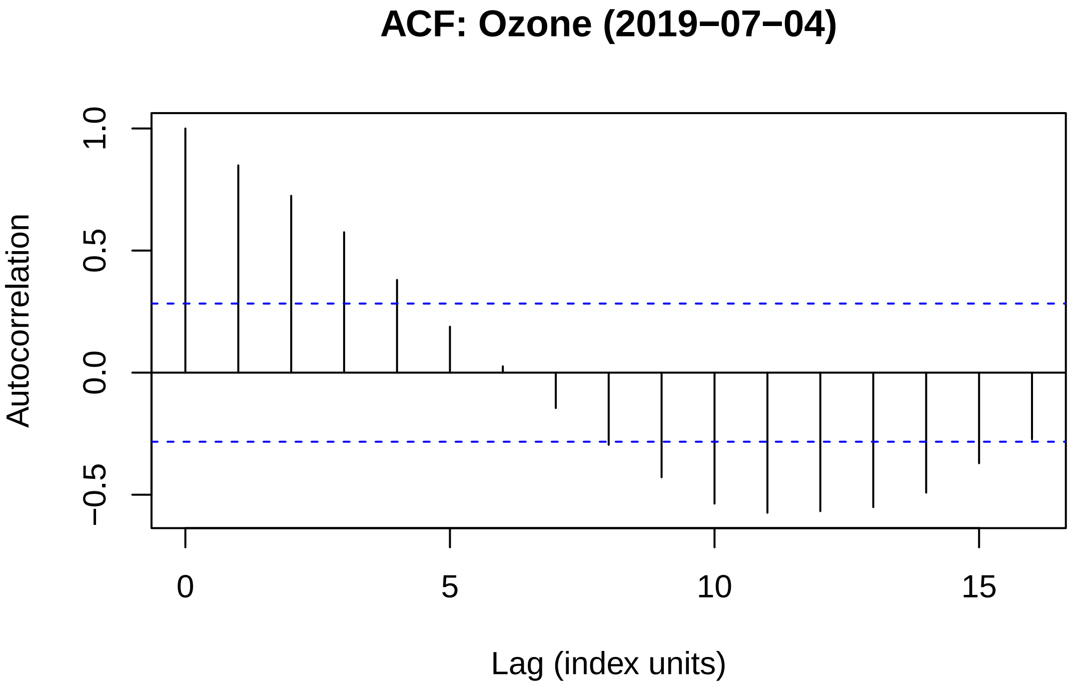
Subset Data

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.

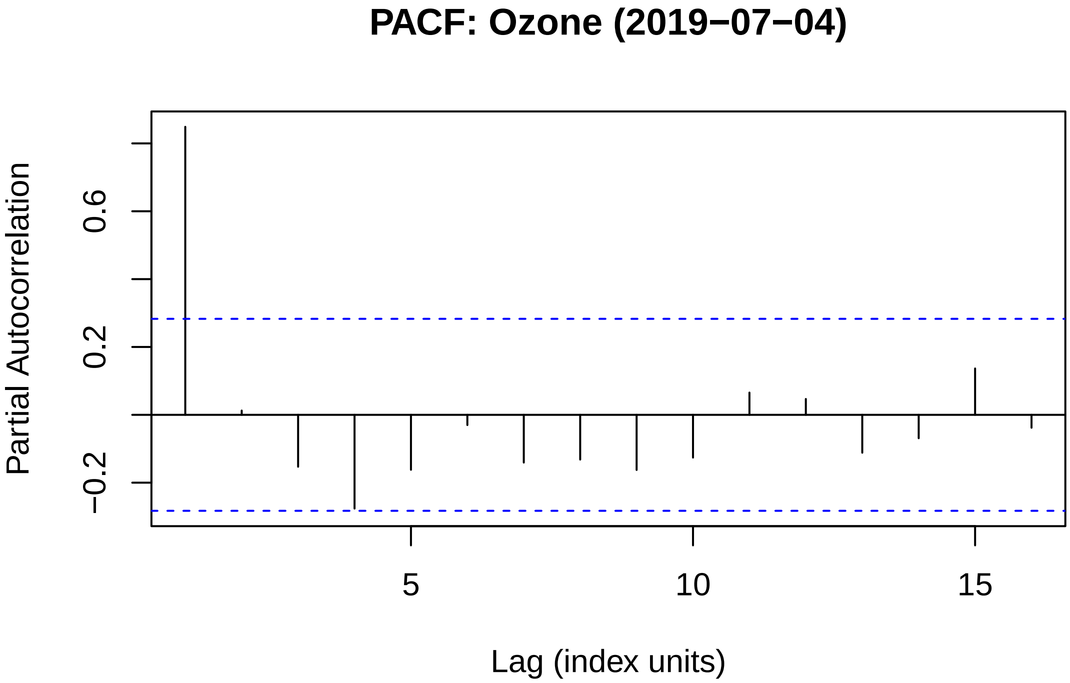


*** the authorization graph shows a gradual drop across the lags, this signifies ozone concentration is persistent across levels in short time intervals. this has a positive corrilation with other local measurements ***

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?



*** this graph indicates significant correlations at the initial lags that quickly drops off. this indicates that ozone concentrations are dependent on most recent measurements ***

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("tidyverse")

# ozone: import, select, drop missing observations, rename

ozone_data <- readr::read_csv("../Data/ftc_o3.csv") %>%
 select(datetime, sample_measurement) %>%
 drop_na() %>%
 rename(ozone_ppm = sample_measurement)

# examine dataframe object

head(ozone_data)

# calculate quantiles of ozone concentration
ozone_q <- quantile(ozone_data$ozone_ppm, probs = c(0, .25, .5, .75, 1), na.rm = TRUE)
tibble(
  stat = c("Min", "Q1", "Median", "Q3", "Max"),
  value = as.numeric(ozone_q)
)

# Compute quartiles
ozone_q <- quantile(ozone_data$ozone_ppm, probs = c(0, .25, .5, .75, 1), na.rm = TRUE)
ozone_iqr <- IQR(ozone_data$ozone_ppm, na.rm = TRUE)

# Create labeled table matching the example
quartile_table <- tibble(
  Quantile = c("0", "1", "0.5", "0.25", "0.75", "0.75 - 0.25"),
  Descriptor = c("minimum", "maximum", "median", "25th%", "75th%", "IQR"),
  `Values` = c(
    round(ozone_q[1], 3),
    round(ozone_q[5], 3),
    round(ozone_q[3], 3),
    round(ozone_q[2], 3),
    round(ozone_q[4], 3),
    round(ozone_iqr, 3)
  )
)

knitr::kable(
  quartile_table,
  caption = "Quantile Descriptors and Values for Ozone Concentration (ppm)"
)

# plot cumulative distribution of ozone concentration
ggplot(ozone_data, aes(x = ozone_ppm)) +
  stat_ecdf(geom = "step") +
  scale_x_continuous(
    name = "Ozone concentration (ppm)",
    breaks = scales::pretty_breaks(8),
    minor_breaks = NULL
  ) +
  scale_y_continuous(
    name = "Cumulative proportion",
    breaks = seq(0, 1, by = 0.1),
    minor_breaks = NULL,
    limits = c(0, 1)
  ) +
  ggtitle("Cumulative Distribution of Ozone Concentration",
          subtitle = "All measurements: Jan 2019 - Jan 2020") +
  theme_minimal(base_size = 12)

# create histogram of ozone concentration
ggplot(ozone_data, aes(x = ozone_ppm)) +
  geom_histogram(bins = 50, boundary = 0, closed = "left") +
  labs(
    x = "Ozone concentration (ppm)",
    y = "Count",
    title = "Histogram of Ozone Concentration",
    subtitle = "All measurements: Jan 2019 - Jan 2020"
  ) +
  theme_minimal(base_size = 12)


# create ozone boxplot
ggplot(ozone_data, aes(x = "", y = ozone_ppm)) +
  geom_boxplot(outlier.alpha = 0.6) +
  labs(
    x = "",
    y = "Ozone concentration (ppm)",
    title = "Ozone Concentration Boxplot",
    subtitle = "All measurements: Jan 2019 - Jan 2020"
  ) +
  theme_minimal(base_size = 12)

# 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
stats::acf(ozone_day$ozone_ppm, na.action = na.omit,
           main = "ACF: Ozone (2019-07-04)",
           xlab = "Lag (index units)", ylab = "Autocorrelation")

# create partial autocorrelation plot
stats::pacf(ozone_day$ozone_ppm, na.action = na.omit,
            main = "PACF: Ozone (2019-07-04)",
            xlab = "Lag (index units)", ylab = "Partial Autocorrelation")
```