MECH476: Engineering Data Analysis in R

Chapter 3 Homework: Fort Collins Ozone

Ethan Rutledge

26 September, 2023

Note: Homework can be submitted as *either* .html or .pdf documents. If you haven't installed *LaTex*, change the output mode in the above YAML to `html_document` for ease of knitting and homework submission.

This R Markdown (.Rmd) file is a template for your Chapter 3 Homework. Do everything within this file. Make it your own, but be careful not to change the code-figure-text integration I set up with the code appendix and the global options. If you have used R Markdown before and are comfortable with the extra options, feel free to customize to your heart's desire. In the end, 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.

Ozone Data

The corresponding data file (.csv) for Homework 3 contains *hourly* ozone data from two sites in Fort Collins.

Background

Incidentally, the ozone standard is set to 0.07 parts per million (ppm). Outdoor ozone levels are measured every hour, but the Environmental Protection Agency states that the limit should be judged against an eight-hour rolling average (a transformation that is possible in R, but outside the purview of this chapter). Fort Collins, and most of the Front Range, is in non-attainment for this standard, which is why you are required to get the emissions checked on your car every year.

Question 0: Load R Packages

Question 1: Preparation

Import, Select, and Clean Data

Using the pipe (`%>%`) to connect **three** lines of code, *import* the file with the appropriate `readr` function and a relative pathname, *select* the below variables, and *drop* missing observations. Remember to assign the output to a `tibble` object with an informative name on the left side of the `gets arrow`.

Retain the following variables in Step 2 of the "pipe":

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

`sample_measurement` is a vague variable name; what is being measured? It is also a little long. Add a fourth line of code to your pipe that renames this variable as `ozone_ppm`, indicating an ozone concentration measurement in parts per million (ppm). *FYI*: If the dataset had multiple ozone measurements on different time scales, it would be important to include that information in the variable name (e.g., `ozone_ppm_hourly`).

Examine Data

Examine the structure and contents of the dataframe to confirm the file imported and was manipulated properly. How many observations were dropped due to missing values? *Hint*: Only consider the `ozone_ppm` variable; there were no missing values for `datetime`.

```
##    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
```

Question 2: Extract and Compare

Using a variant of the `dplyr::slice()` function with **two** arguments (one to specify number of observations to extract and one to specify by which variable R should the observations in the output), extract the top ten ozone values and assign them to a separate object.

Now, complete the same process for the bottom ten ozone values.

Do the highest and lowest values tend to occur at certain times of the day?

```
## # A tibble: 10 x 2
##    ozone_ppm datetime
##        <dbl> <dttm>
##  1     0.096 2019-07-03 20:00:00
##  2     0.093 2019-07-03 21:00:00
##  3     0.09  2019-07-03 20:00:00
##  4     0.089 2019-07-03 19:00:00
##  5     0.086 2019-07-24 20:00:00
##  6     0.083 2019-06-29 20:00:00
##  7     0.082 2019-09-10 23:00:00
##  8     0.081 2019-07-03 19:00:00
##  9     0.081 2019-06-07 19:00:00
## 10     0.081 2019-08-03 19:00:00
```

```
## # A tibble: 10 x 2
##    ozone_ppm datetime
##        <dbl> <dttm>
```

```
##  1         0 2019-01-05 07:00:00
##  2         0 2019-02-02 07:00:00
##  3         0 2019-02-04 07:00:00
##  4         0 2019-02-28 11:00:00
##  5         0 2019-03-01 07:00:00
##  6         0 2019-03-08 07:00:00
##  7         0 2019-05-11 11:00:00
##  8         0 2019-11-24 10:00:00
##  9         0 2019-12-20 00:00:00
## 10         0 2019-12-20 01:00:00
```

```
The Highest values all seem to occur around 20:00:00 (evening), and the lowest values all seem to
```

Question 3: Maximum and Minimum

Using the output from the previous question, on what day does the highest value occur? The lowest?

```
## [1] "2019-07-03 20:00:00 UTC"
```

```
## [1] "2019-01-05 07:00:00 UTC"
```

Question 4: Mutate

Create a new variable (`ozone_ugm3`) that provides ozone concentration in micrograms per cubic meter (ug/m3) instead of parts per million (ppm), as in `ozone_ppm`. Because we do not have access to crucial measurements such as atmospheric pressure or temperature, the following exercise will not give the true ozone concentration values (ug/m3) but will give you practice using the appropriate `dplyr` verb.

Although the real data are not available in this dataset, you will use the following information to complete the conversion. The hard-coded values are illustrative estimates, not accurate readings.

- ozone concentration in parts per million (`ozone_ppm`)
- molecular weight of ozone (47.998 g/mol)
- Celsius to Kelvin temperature conversion (K = 273.15 + C) (will be used twice but in different ways in the numerator and denominator)
- estimated atmospheric pressure in Fort Collins, CO (637 mmHg)
- universal gas constant (22.4136 L/mol)
- estimated temperature in Fort Collins, CO (10° Celsius)

Then, you will need to embed the following ideal gas law equation into a function call that creates a new variable based on the values from `ozone_ppm`. *Hint*: Based on this approach, an ozone concentration of 0.001 ppm converts to approximately 2.066 ug/m3.

$$\texttt{ozoneugm3} = 1000 * \frac{\Big((\texttt{ozone\_ppm})(\text{molecular weight of ozone})(273.15)(\text{atmospheric pressure})\Big)}{\Big((\text{universal gas constant})(\text{Celsius to Kelvin value} + \text{temperature})(\text{atmospheric pressure})\Big)}$$

```
## # A tibble: 16,914 x 3
##    ozone_ppm datetime            ozoneugm3
##        <dbl> <dttm>                  <dbl>
##  1     0.017 2019-01-01 07:00:00      35.1
##  2     0.017 2019-01-01 08:00:00      35.1
##  3     0.017 2019-01-01 09:00:00      35.1
##  4     0.017 2019-01-01 10:00:00      35.1
##  5     0.015 2019-01-01 11:00:00      31.0
##  6     0.017 2019-01-01 12:00:00      35.1
##  7     0.028 2019-01-01 13:00:00      57.8
##  8     0.03  2019-01-01 14:00:00      62.0
##  9     0.036 2019-01-01 15:00:00      74.3
## 10     0.036 2019-01-01 16:00:00      74.3
## # i 16,904 more rows
```

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)

# do you need to install each R package?
# load packages for current R session
library(tidyverse)
# 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 tibble
summary(ozone_data)
# calculate number of missing observations
sum(is.na(ozone_data$ozone_ppm))
# extract top ten ozone values and save them to df
df_max <- slice_max(ozone_data, ozone_ppm, n=10)
# extract bottom ten ozone values and save them to df
df_min <- slice_min(ozone_data, ozone_ppm, n=10)
print(df_max)
print(df_min)
# extract date/time of highest ozone concentration
max_ozone_datetime <- df_max$datetime[1]
print(max_ozone_datetime)
# extract date/time of lowest ozone concentration
min_ozone_datetime <- df_min$datetime[1]
print(min_ozone_datetime)
# create new variable of ug/m3 from ppm and overwrite dataset
molecular_weight_ozone <- 47.98   # g/mol
C_to_K = 273.15                   # K
atm_P_foco = 637                  # mmHg
R_universal = 22.4136             # L/mol
temp_foco_c = 10                  # degC

mutate(.data = ozone_data, ozoneugm3 = 1000 * (ozone_ppm * molecular_weight_ozone * C_to_K * atm_P_foco
```