Lab 8

Enter the names of the group members here: Austine Do, Graceanne Becker, Catherine Zhong

This assignment is due by the end of the lab. Only one student in the group submits a pdf file on Gradescope.

For all questions, include the R commands/functions that you used to find your answer (show R chunk). Answers without supporting code will not receive credit. Write full sentences to describe your findings.

In this lab, you will explore data that were originally collected by researchers at the Johns Hopkins Bloomberg School of Public Health. Let's first load the appropriate packages for today:

```
library(tidyverse)
library(ggmap)
library(plotROC)
```

```
## Warning: package 'plotROC' was built under R version 4.3.1
```

Let's upload the data from Github and take a quick look:

```
pollution <- read_csv("https://raw.githubusercontent.com/laylaguyot/datasets/main//pm25.csv")

# Take a quick look!
head(pollution)
```

```
## # A tibble: 6 x 11
##      id state   county  city         value  zcta   lat   lon   pov  CMAQ zcta_pop
##   <dbl> <chr>   <chr>   <chr>        <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>    <dbl>
## 1 1003. Alabama Baldwin Fairhope      9.60 36532  30.5 -87.9   6.1  8.10    27829
## 2 1027. Alabama Clay    Ashland      10.8  36251  33.3 -85.8  19.5  9.77     5103
## 3 1033. Alabama Colbert Muscle Sho~ 11.2  35660  34.8 -87.7  19    9.40     9042
## 4 1049. Alabama DeKalb  Crossville  11.7  35962  34.3 -86.0  13.8  8.53     8300
## 5 1055. Alabama Etowah  Gadsden     12.4  35901  34.0 -86.0   8.8  9.24    20045
## 6 1069. Alabama Houston Dothan      10.5  36303  31.2 -85.4  15.6  9.12    30217
```

It contains the following variables:

Variable Name	Description
state, county, city	Name of the state, county, city where monitor is located
value	Annual level of PM2.5 in $\mu g/m^3$
zcta	ZIP code where monitor is located
lat	Latitude coordinate of monitor location
lon	Longitude coordinate of monitor location

Variable Name	Description
pov	Percentage of ZIP code population (where monitor is located) living in poverty
zcta_pop	Population of ZIP code where monitor is located (based on 2010 Census)
CMAQ	Computer model estimate of PM2.5 levels

The goal of the lab is to make predictions for the PM2.5 levels with two different approaches.

Question 1 (6 pts)

Let's start exploring the dataset! Which state has the largest number of PM2.5 monitors within the state?

```
# Finding the state with the largest number PM2.5 monitors
pollution |>
    group_by(state) |>
    summarize(count = n()) |>
    arrange(desc(count))
```

```
## # A tibble: 49 x 2
##    state          count
##    <chr>          <int>
##  1 California        85
##  2 Ohio              44
##  3 Illinois          38
##  4 Indiana           36
##  5 North Carolina    35
##  6 Pennsylvania      32
##  7 Michigan          30
##  8 Florida           29
##  9 Georgia           28
## 10 Texas             27
## # i 39 more rows
```

California has the largest number of PM2.5 monitors with 85 PM2.5 monitors in the state.

Find the mean of the PM2.5 values within each state. Which state in the U.S. has the highest mean PM2.5 value? Which state has the lowest mean PM2.5 value?

```
# Finding the states with the highest and lowest mean PM2.5 value
pollution |>
    group_by(state) |>
    summarize(mean_PM2.5_value = mean(value)) |>
    arrange(desc(mean_PM2.5_value))
```

```
## # A tibble: 49 x 2
##    state                mean_PM2.5_value
##    <chr>                           <dbl>
##  1 West Virginia                    13.4
```

```
##   2 Ohio                              12.9
##   3 Pennsylvania                      12.8
##   4 Georgia                           12.5
##   5 Indiana                           12.4
##   6 Maryland                          12.3
##   7 California                        12.2
##   8 Delaware                          12.2
##   9 District Of Columbia              12.1
##  10 Kentucky                          12.1
## # i 39 more rows
```

```
pollution |>
    group_by(state) |>
    summarize(mean_PM2.5_value = mean(value)) |>
    arrange(mean_PM2.5_value)
```

```
## # A tibble: 49 x 2
##    state        mean_PM2.5_value
##    <chr>                   <dbl>
##  1 Maine                    5.58
##  2 Wyoming                  5.85
##  3 New Mexico               6.24
##  4 North Dakota             6.57
##  5 Colorado                 7.34
##  6 South Dakota             7.57
##  7 Vermont                  7.81
##  8 Florida                  7.90
##  9 Montana                  8.02
## 10 Idaho                    8.21
## # i 39 more rows
```

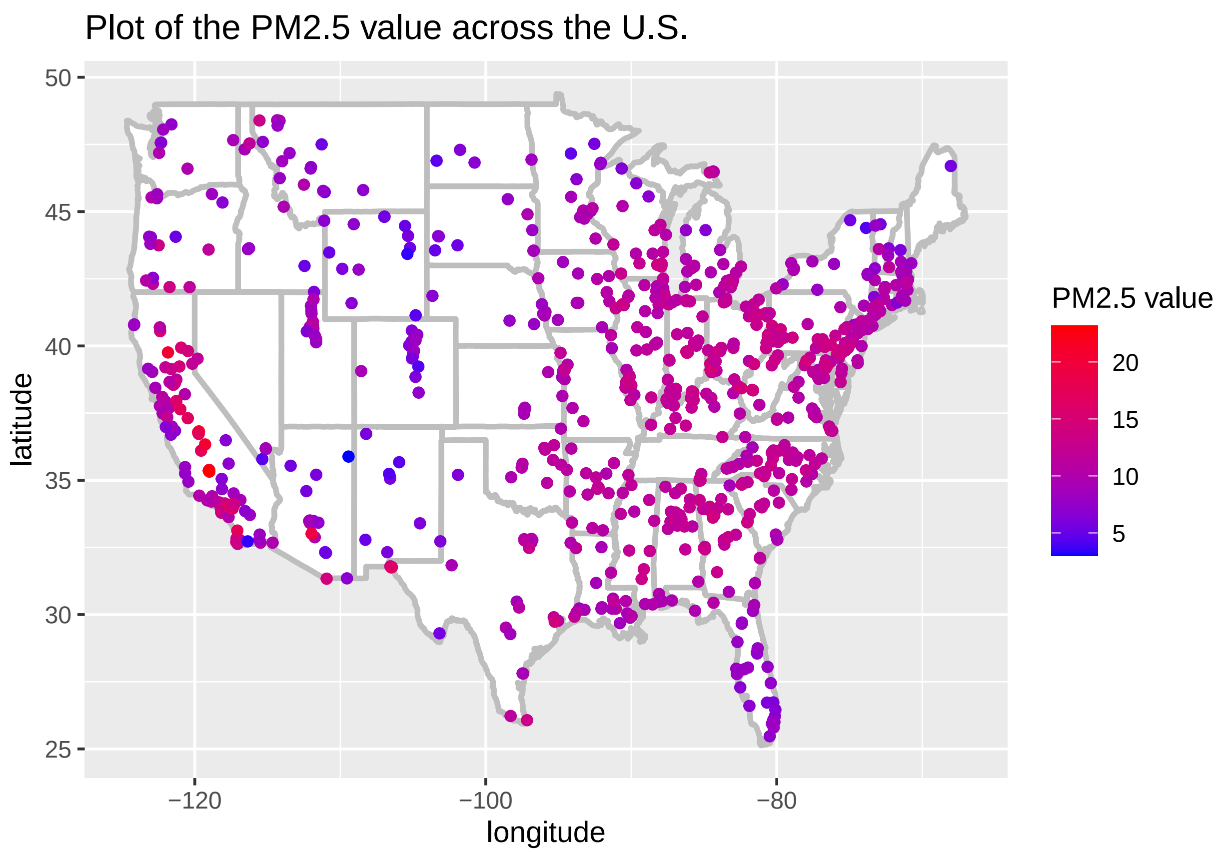
The state with the highest mean PM2.5 value is West Virginia and the state with the lowest mean PM2.5 value is Maine

We can represent the values of PM2.5 on a map! Consider the code below that creates a map of the United States divided by states. Remember that `ggplot` works in layers: add a layer to the code below to represent the PM2.5 values from the `pollution` dataset across the states. Make sure to add colors to distinguish between lower vs higher values.

```
# Create data for a map of the United States divided by states
state_data <- map_data("state")

# Create a map with `ggplot`
ggplot() +
    geom_polygon(data = state_data, aes(x = long, y = lat, group = group),
                 fill = "white", color = "grey", size = 1) +
    # Add a layer with data from `pollution`
    geom_point(data = pollution, aes(x = lon, y = lat, color = value)) +
    labs(title = 'Plot of the PM2.5 value across the U.S.',
         x = 'longitude',
         y = 'latitude',
         color = 'PM2.5 value') +
    scale_color_gradient(low = "blue", high = "red")
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



Where are the maximum values of PM2.5 located?

The maximum values of PM2.5 are located mostly in California.

Question 2 (3 pts)

Create a new variable called `violation` that takes a value of 1 if the location has a value of PM2.5 that is in violation of the national standards (greater than 12 $\mu\text{g/m}^3$) and is 0 if that location is not in violation. Add this new variable to the `pollution` dataset.

```r
# Adding the 'violation' variable to the pollution dataset
pollution <- pollution |>
                  mutate(violation = ifelse(value > 12, 1, 0))
```

Using your newly created `violation` variable, what percentage of all of the locations in the dataset are in violation of the national PM2.5 standards?

```r
# Calculating the percentage of locations in the dataset that violate PM2.5 standards
sum(pollution$violation)/nrow(pollution)
```

```
## [1] 0.3219178
```

About 32% of all locations in the dataset are in violation of the national PM2.5 standards

---

Question 3 (2 pts)

Next, we will build two different models to predict the PM2.5 levels, using some other variables:

- A linear regression model to predict the PM2.5 values at a given location.
- A logistic regression model to predict whether a given location is in violation of the national ambient air quality standards.

What is the outcome variable for each model?

The outcome variable is PM2.5 value for the linear regression model and the outcome variable is violation status for the logistic regression model.

To do so, we will split the `pollution` dataset into two parts, a `train_data` set and a `test_data` set:

- The train set will be all of the locations outside the state of Texas.
- The test set will be all of the locations inside the state of Texas.

Create the `train_data` set and the `test_data` set as described above:

```r
# Creating the training and test dataset for our models
train_data <- pollution |>
                    filter(state != 'Texas')

test_data <- pollution |>
                    filter(state == 'Texas')
```

Question 4 (6 pts)

Let's build a linear regression model called `train_lin` to predict the `value` variable in the `train_data` set. Only use the following predictors: `lat`, `lon`, `pov`, and `zcta_pop`. Which predictors seem to be the most useful in predicting the PM2.5 values?

```r
# Building the linear regression model
train_lin <- lm(value ~ lat + lon + pov + zcta_pop, data = train_data)
summary(train_lin)
```

```
##
## Call:
## lm(formula = value ~ lat + lon + pov + zcta_pop, data = train_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.737 -1.482  0.161  1.339 12.244
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.462e+01  9.340e-01  15.653  < 2e-16 ***
## lat         -4.880e-02  1.934e-02  -2.523   0.0118 *
## lon          3.248e-02  5.682e-03   5.715 1.52e-08 ***
## pov          3.169e-02  7.720e-03   4.105 4.43e-05 ***
## zcta_pop     2.431e-05  4.902e-06   4.959 8.56e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.481 on 844 degrees of freedom
## Multiple R-squared:  0.09135,    Adjusted R-squared:  0.08704
## F-statistic: 21.21 on 4 and 844 DF,  p-value: < 2.2e-16
```

The variables that seems the most significant/useful in predicting the PM2.5 value are `lon`, `pov`, and `zcta_pop` for the linear regression model.

Use the linear model to make predictions for the violation in the `train_data` set and compute the corresponding RMSE, as shown below. Then compute the value of RMSE when applying the linear model to the `test_data` set.

Get rid if `eval = FALSE` below before knitting your lab report.

```r
# Calculate RMSE for predictions in train data
sqrt(mean((train_data$value - predict(train_lin, newdata = train_data))^2))
```

```
## [1] 2.473524
```

```r
# Calculate RMSE for predictions in test data
sqrt(mean((test_data$value - predict(train_lin, newdata = test_data))^2))
```

```
## [1] 1.973338
```

How well does our model predict the values of PM2.5 for the train set vs for the test set?

The test set performs better since the RSME is lower than the train set by about 0.5.

Question 5 (6 pts)

Let's build a logistic regression model called `train_log` to predict the `violation` variable in the `train_data` set. Only use the following predictors: `lat`, `lon`, `pov`, and `zcta_pop`. Which predictors seem to be the most useful in predicting the violation?

```r
# Building the logistic regression model
train_log <- glm(violation ~ lat + lon + pov + zcta_pop, data = train_data, family = 'binomial')
summary(train_log)
```

```
##
## Call:
## glm(formula = violation ~ lat + lon + pov + zcta_pop, family = "binomial",
##     data = train_data)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.654e+00  8.433e-01   1.961 0.049904 *
## lat         -3.186e-02  1.704e-02  -1.869 0.061605 .
## lon          1.957e-02  5.283e-03   3.705 0.000211 ***
## pov          2.419e-02  6.572e-03   3.680 0.000233 ***
## zcta_pop     1.044e-05  4.266e-06   2.448 0.014348 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1072.3  on 848  degrees of freedom
## Residual deviance: 1032.8  on 844  degrees of freedom
## AIC: 1042.8
##
## Number of Fisher Scoring iterations: 4
```

The variables that seems significant/useful in predicting the violation status are `lon`, `pov`, and `zcta_pop`.

Use the logistic model to make predictions for the violation in the `train_data` set and compute the corresponding AUC, as shown below. Then compute the value of AUC when applying the logistic model to the `test_data` set.

Get rid if `eval = FALSE` below before knitting your lab report.

```r
# Calculate AUC for predictions in train data
calc_auc(ggplot(train_data) +
           geom_roc(aes(d = violation,
                        m = predict(train_log, type = "response"))))$AUC

# Calculate AUC for predictions in test data
calc_auc(ggplot(test_data) +
           geom_roc(aes(d = violation,
                        m = predict(train_log, type = "response", newdata = test_data))))$AUC
```

How well does our logistic model indicate whether a given location is in violation of the national ambient air quality standards for the train set vs for the test set?

Logistic model performs better for the train set than for the test set in predicting whether the given location is in violation of the national ambient air quality standards.

Question 6 (1 pt)

After investigating what features of a location seem to affect the PM2.5 levels, did the data match your expectations or not? If the data differed from your expectation, provide a possible explanation for why the data differed from what you expected.

The data did seem to match our expectations as we expected California to have some of the highest PM2.5 and some state in the Northern parts of the U.S. to have some of the lowest PM2.5 levels (that being Maine, etc.).

Formatting: (1 pt)

Make sure the names of all group members are included at the beginning of the document.

Knit your file! You can knit into pdf directly or into html. Once it knits in html, click on `Open in Browser` at the top left of the window pops out. Print your html file into pdf from your browser.

Any issue? Ask other classmates or TA!

Finally, remember to select pages for each question when submitting your pdf to Gradescope and to identify your group members.