EEMB146 Lab 3 Homework

Written by Rafael Romero

2025-01-27

Contents

Homework 3 - Visualizing Data & Identifying Distributions

This homework will apply your data wrangling (Lab 2) to data visualization (Lab3) skills. Visualizing data is part of Exploratory Data Analysis (EDA) which is a fundamental step to all statistics. If you are having trouble with RStudio or knitting your .Rmd file please speak with a TA before the due date. **All of the information and code needed to answer homework questions can be found in Lab 3 Exercise file.** You will be graded on completeness and correctness.

Note for HW 3: this lab will test your data wrangling and data visualization skills simultaneously. To answer the homework questions, you will need to 1) wrangle data to create the summaries (i.e. total tons of methanol per country), and then 2) pipe (%>%) that data summary into ggplot. If you need more examples of how to do this you can refer to this data wrangling with pipes guide

```
## setting up the style of your knitted document
knitr::opts_chunk$set(echo = TRUE,message=FALSE, warning=FALSE)

################################################################################
## installing packages

library(readr)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v purrr     1.0.2
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(ggplot2)
#install.packages("cowplot")
library(cowplot) # NEW package for combining plots
```

```
##
## Attaching package: 'cowplot'
##
## The following object is masked from 'package:lubridate':
##
##     stamp
```

```r
################################################################################
## upload data
# recommend putting data into a folder called 'data'
air_pollution <- read.csv("data/ejscreen_airpollutants-1.csv")
water_pollution <- read.csv("data/ejscreen_waterpollutants-1.csv")
```

Homework Questions

Homework 3 questions will test your vocabulary knowledge, data wrangling, and visualization skills. Data for this lab comes from EJScreen: Environmental Justice Screening and Mapping Tool hosted by the US Environmental Protection Agency.

Question 1

For question 1, you will just answer vocabulary questions to make sure the core concepts of visualization are mastered. To answer these vocabulary questions you will use the terms: **histogram**, **boxplot**, **scatterplot**, or **facet_wrap**.

a. If you have 1 numeric variable that you want to understand the distribution of, what would be the best type of graph to use for visualization purposes?

Answer: *Histogram* would be the best type of graph to use for visualization purposes.

b. If you have 1 numeric and 1 factor variables, what would be the best type of graph to use for visualization purposes?

Answer: *Boxplot* would be the best type of graph for 1 numeric and 1 factor variables.

c. If you have 2 numeric variables, what would be the best type of graph to use for visualization purposes?

Answer: *Scatterplot* would be the best type of graph for 2 numeric variables.

d. If you have 2 numeric variables, and you think their relationship may be site or species specific, what would be the best type of graph to use for visualization purposes?

Answer: *Facet Wrap* would be the best type of graph to use if we have 2 numeric variables and they may be species specific.

Question 2

For question 2, you will answer **True** or **False** to a series of sub-questions about ways to make plots more informative and visually appealing.

a. True or False, the label "weight" is more informative than "Weight (kg)". **False**

b. True or False, the `theme_classic()` produces a white background whereas `theme_light()` produces a gridded background **True**

c. True or False: The `theme(..., element_line())` function is used to change the font style of axis labels. **False**

d. True or False: The `theme(..., element_blank())` function is used to adjust the thickness of gridlines. **False**

Question 3

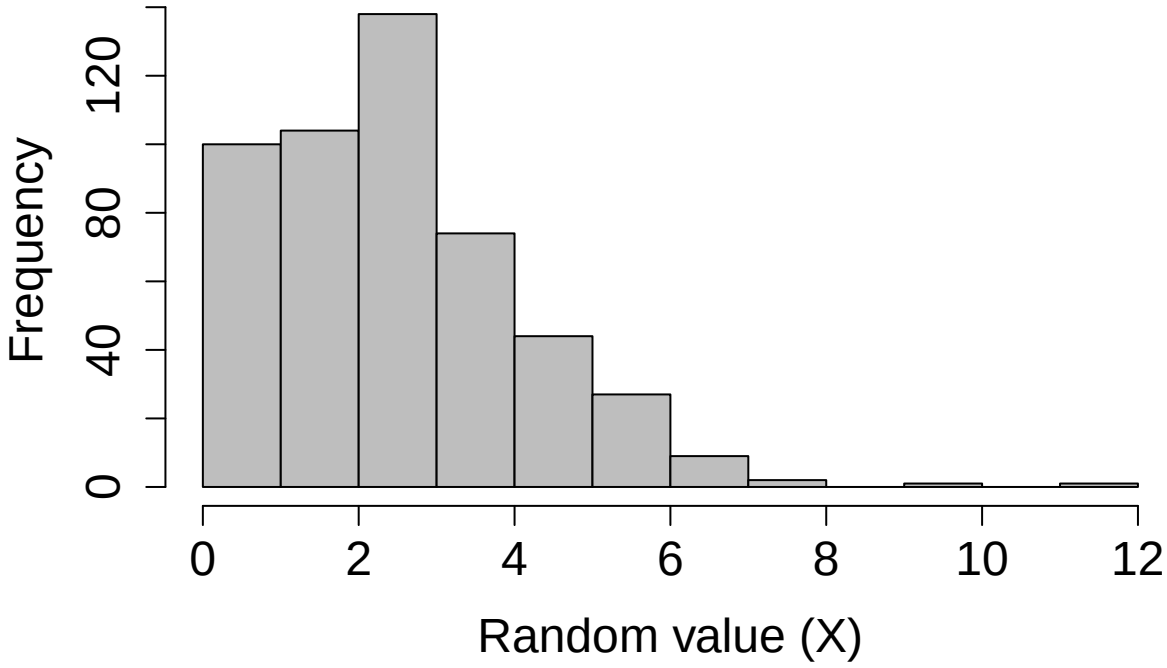
For question 3, you are going to look at simulated data and choose which data distribution (i.e. **normal**, **poisson**, or **uniform**) fits the graph the most accurately.

a. What data distribution is plot A? **poisson**

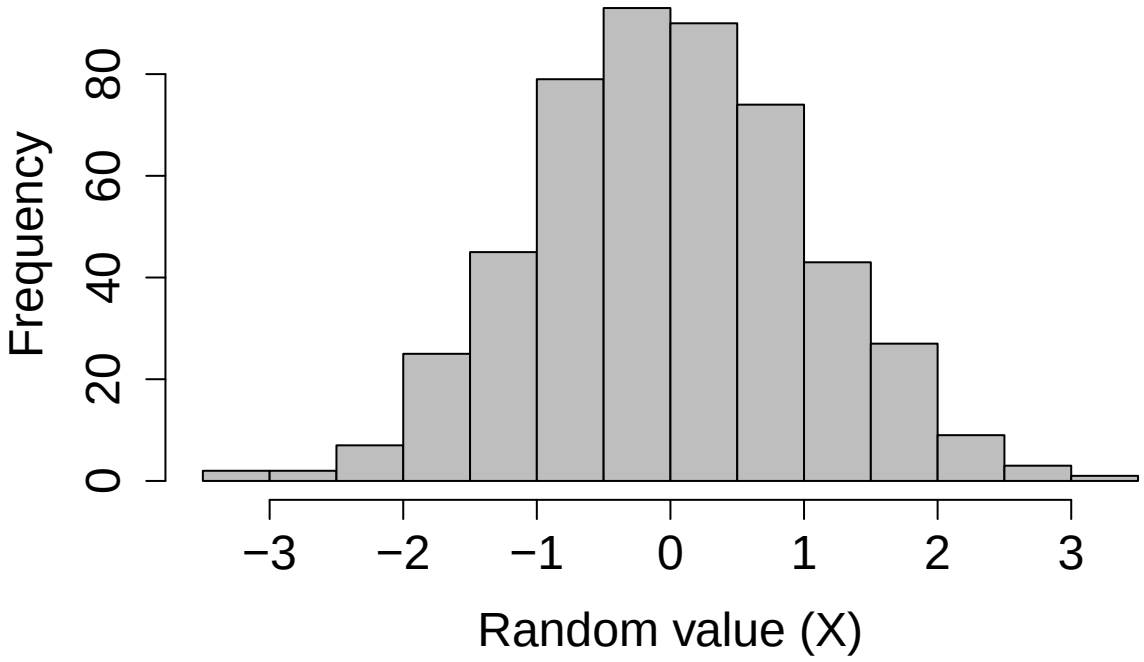
b. What data distribution is plot B? **normal**

c. What data distribution is plot C? **uniform**

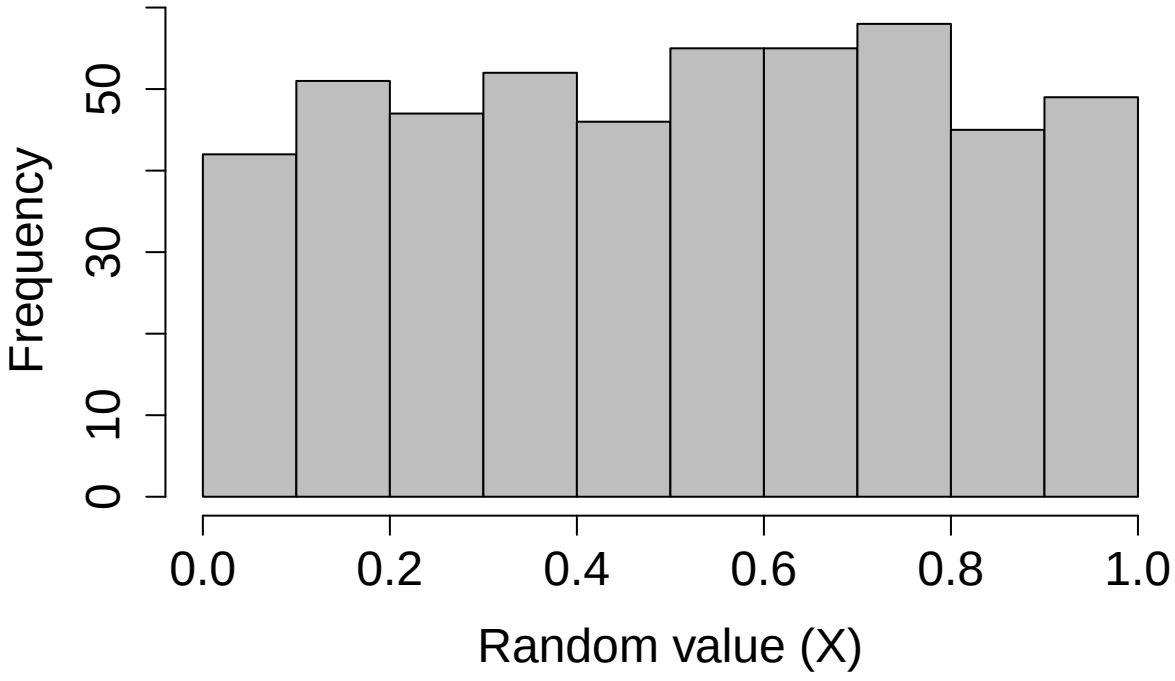
```
# Plot A
plotA <- rpois(n = 500, lambda = 3)
hist(plotA,
     xlab = "Random value (X)", main = "",
     col = "grey", cex.lab = 1.5, cex.axis = 1.5)
```




```
plotB <- rnorm(n = 500, mean = 0, sd = 1)
hist(plotB,
     xlab = "Random value (X)", main = "",
     col = "grey",cex.lab = 1.5, cex.axis = 1.5)
```

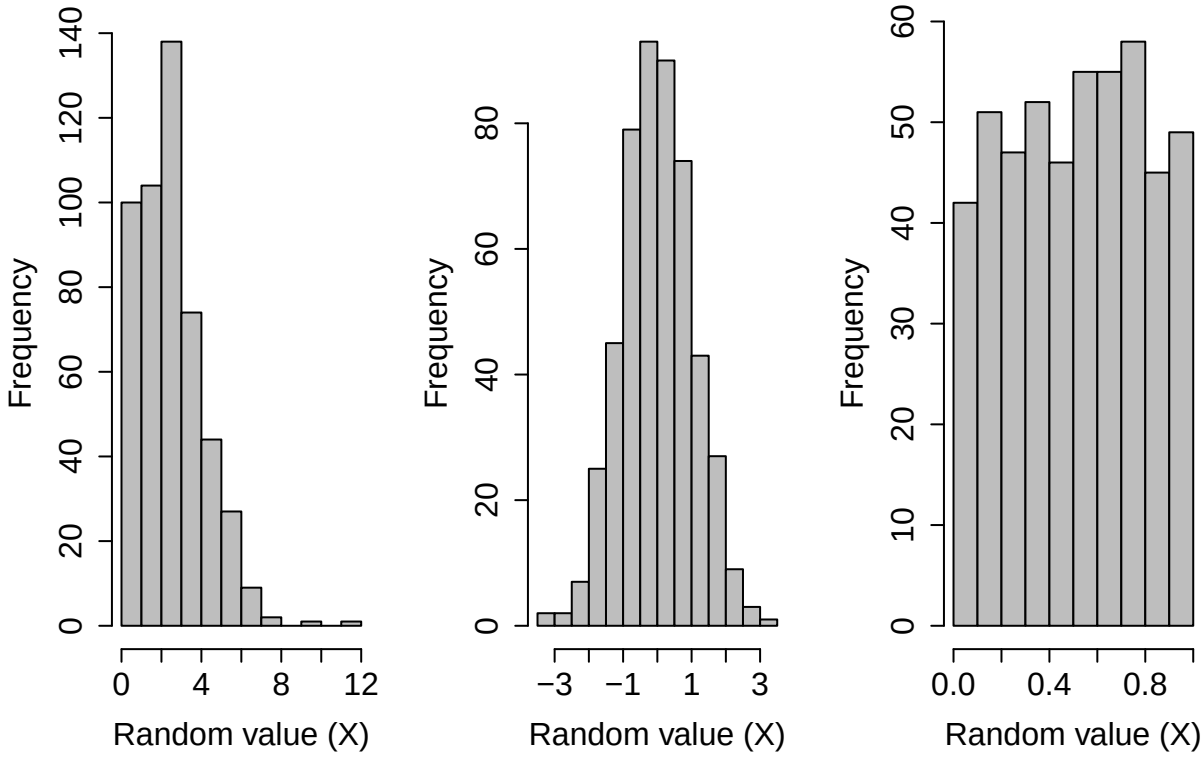



```r
plotC <- runif(n = 500, min = 0, max = 1)
hist(plotC,
     xlab = "Random value (X)", main = "",
     col = "grey", cex.lab = 1.5, cex.axis = 1.5)
```




```r
# plot multiple base R plots together
par(mfrow = c(1,3)) # plot in 1 row, 3 columns
hist(plotA,
     xlab = "Random value (X)", main = "",
     col = "grey", cex.lab = 1.5, cex.axis = 1.5)
hist(plotB,
```

```
     xlab = "Random value (X)", main = "",
     col = "grey",cex.lab = 1.5, cex.axis = 1.5)
hist(plotC,
     xlab = "Random value (X)", main = "",
     col = "grey", cex.lab = 1.5, cex.axis = 1.5)
```



Question 4

For question 4, you will wrangle `air_pollution` data set from the EPA Air Toxics Screening Assessment, where you can also find a glossary of terms regarding `air_pollution` data set.

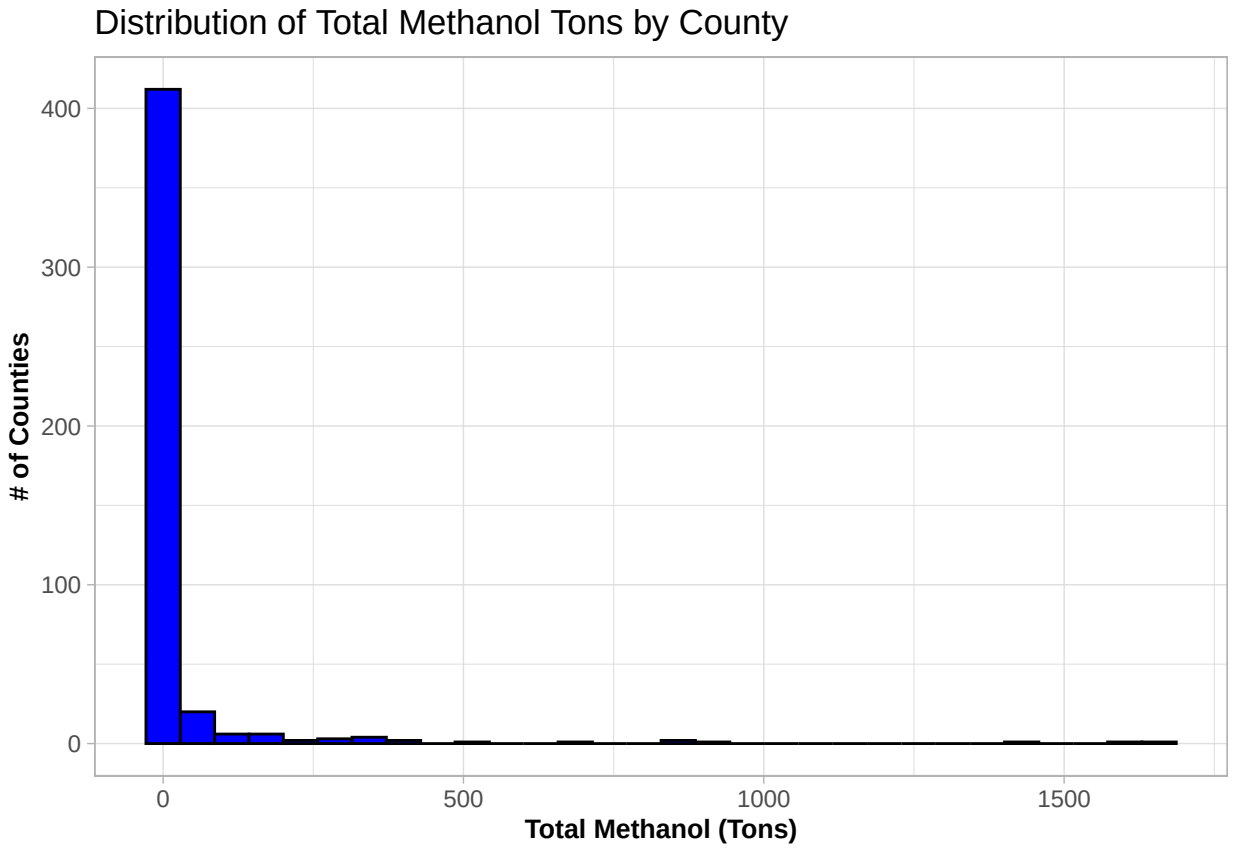
According to US EPA, "Menthanol is released to the environment during industrial uses and naturally from volcanic gases, vegetation, and microbes. Exposure may occur from ambient air and during the use of solvents. Acute (short term) or chornic (long term) exposure of humans to methanol by inhalation or ingestion may result in blurred vision, headache, dizziness, and nausea." For this lab assignment **please report Methanol in the correct units**.

To receive full credit for plots, you must add **informative labels** using `labs()` & `ggtitle()` functions AND `theme_*()` to **change background**.

a. **Calculate the total amount of methanol per county and create a histogram of that county-level data**
   - *hint: first, you will need to wrangle the data using the commands filter(), group_by(county), summarise(total = sum(value_tons)), and then pipe the data into ggplot()*

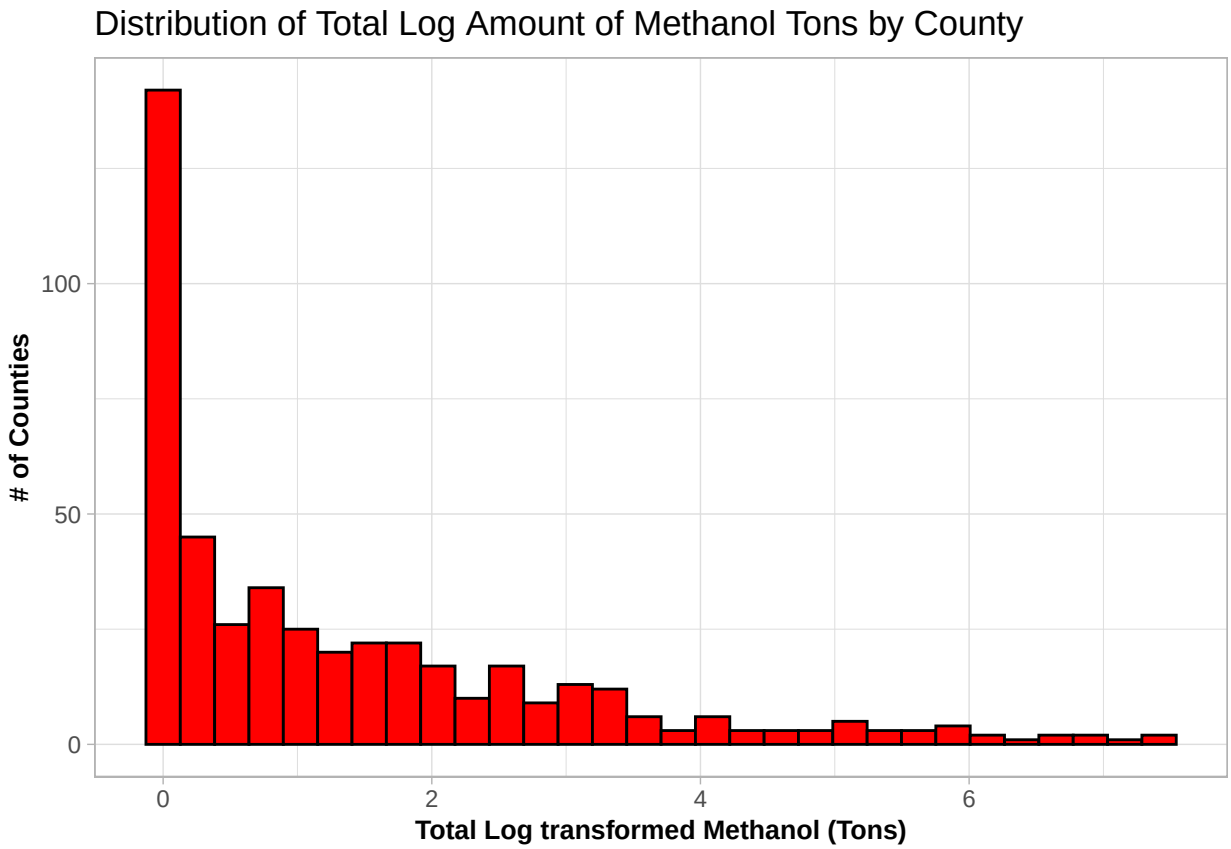
```
total_tons <- air_pollution %>%
  filter(pollutant_name == "Methanol") %>% #Wrangle data so we get only Methanol data and group based o
  group_by(county) %>%
  summarise(total=sum(value_tons)) %>% #sum the tons within each county and attribute it to total so we
  ggplot(aes(x=total))  +
  geom_histogram(color = "black", fill = "blue", position = "identity", alpha = 1) +
```

```
  theme_light()+
  labs(
    title = "Distribution of Total Methanol Tons by County",
    x = "Total Methanol (Tons)",
    y = "# of Counties"
    ) +
   theme(axis.title = element_text(size = 10, face = "bold"))
total_tons
```



b. **Calculate the total log amount of methanol per county and create a histogram of that county-level data**
   - *hint: same data wrangling process as part a*
   - *hint: are there 0 values? how would that change your log transformation equation?*

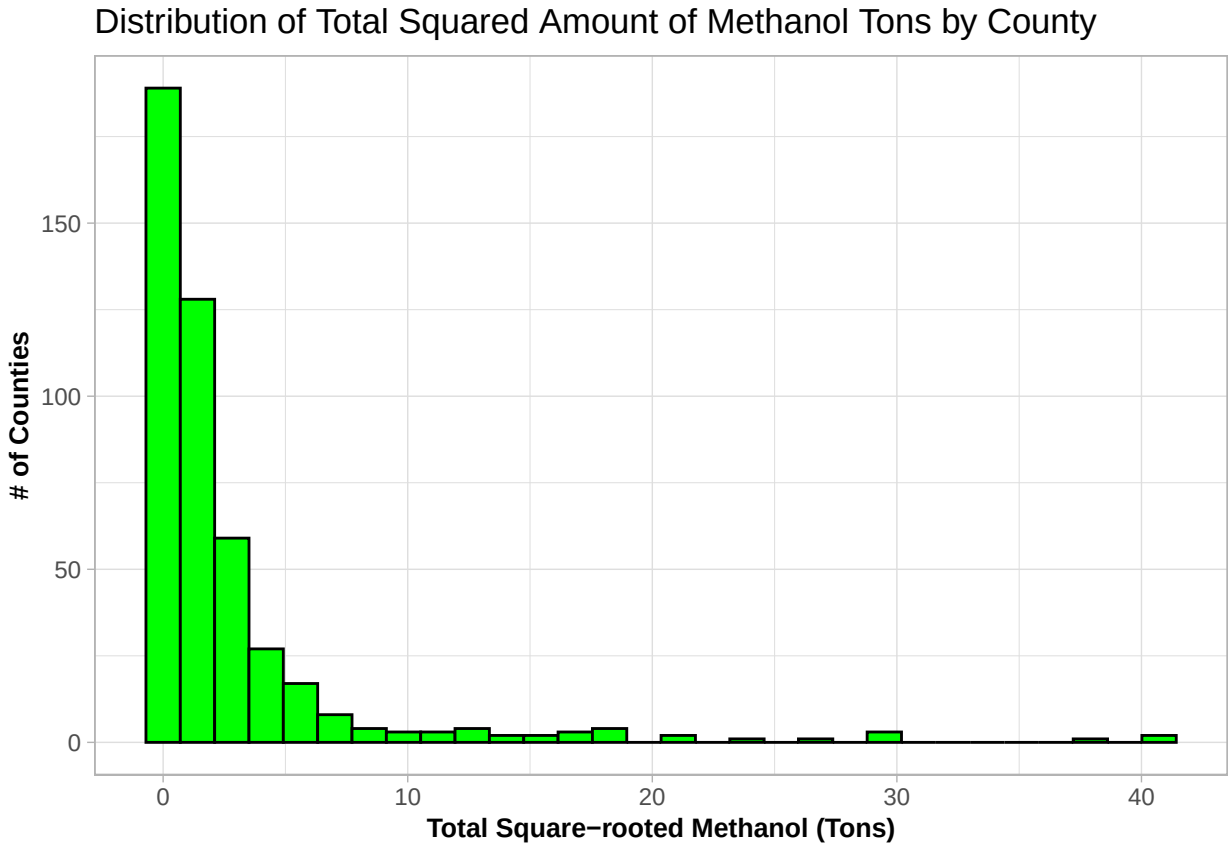
```
log_tons <- air_pollution %>%
 filter(pollutant_name == "Methanol") %>% #wrangle data just as before
 group_by(county) %>%
 summarise(log_total =log(sum(value_tons)+1)) %>%#add +1 because there are counties with 0 Total Metha
 ggplot(aes(x= log_total)) +
 geom_histogram(color = "black", fill = "red", position = "identity") +
 theme_light()+
 labs(
   title = "Distribution of Total Log Amount of Methanol Tons by County",
   x = "Total Log transformed Methanol (Tons)",
   y = "# of Counties")+
  theme(axis.title = element_text(size = 10, face = "bold"))
log_tons
```



c. **Calculate the square root transformed amount of methanol per county and create a histogram of that county-level data**

- *hint: same data wrangling process as part a*

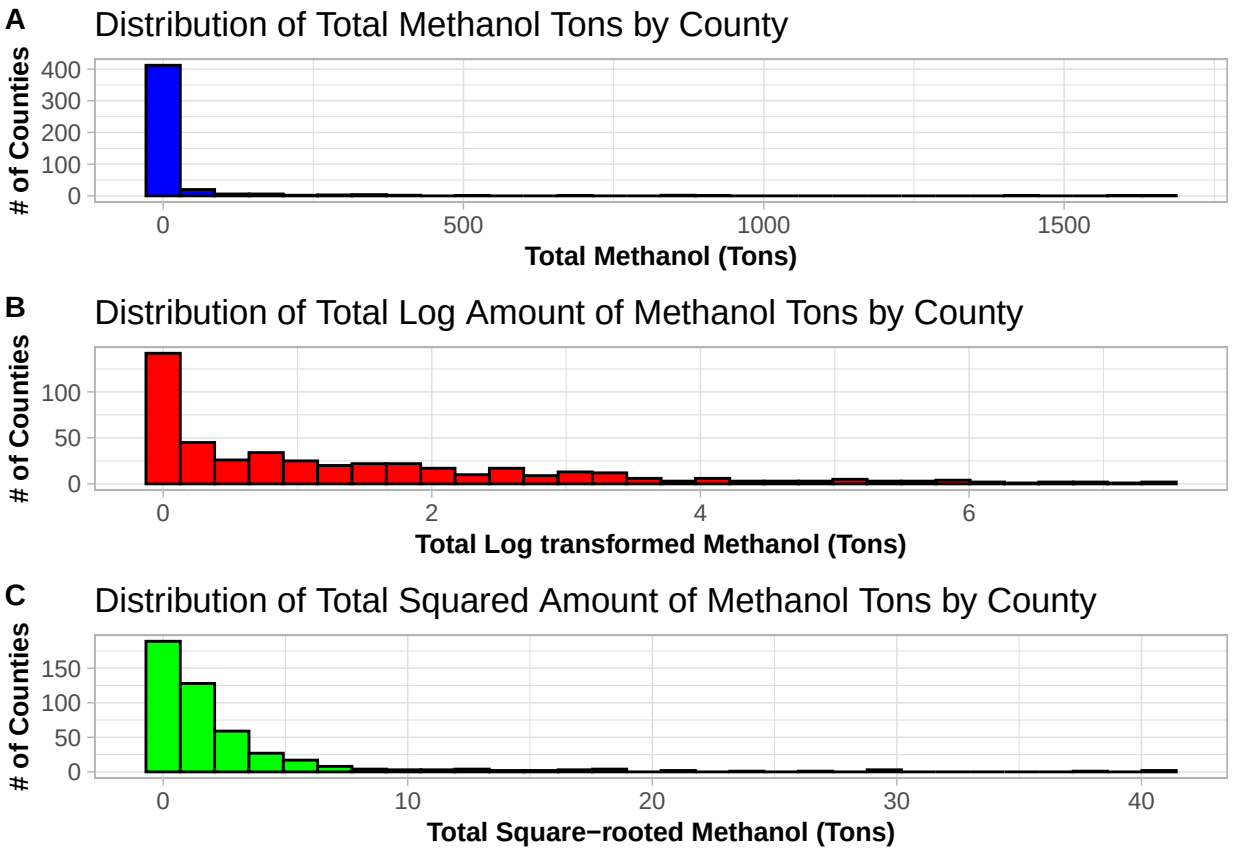
```
sqrt_tons <- air_pollution %>%
  filter(pollutant_name == "Methanol") %>% #Wrangle data just as before
  group_by(county) %>%
  summarise(sqrt_total = sqrt(sum(value_tons)))%>% #calculate the square root transformation by sqrt th
  ggplot(aes(x= sqrt_total)) +
  geom_histogram(color = "black", fill = "green", position = "identity") +
  theme_light()+
  labs(
    title = "Distribution of Total Squared Amount of Methanol Tons by County",
    x = "Total Square-rooted Methanol (Tons)",
    y = "# of Counties")+
   theme(axis.title = element_text(size = 10, face = "bold"))
sqrt_tons
```



d. **Using `plot_grid()` from `cowplot` package, combine all 3 histograms (untransformed, log transformed, square root transformed) into 1 plot. Discuss if the general distribution of 'value_tons' is normal, left skewed or right skewed and the effect of each data transformation (log or sqrt).**

- *hint: look at Lab 3 Exercise File, exercise 1: histogram example*

```
plot_grid(total_tons,log_tons, sqrt_tons,
          ncol = 1, # put all 3 hist in 1 column
          align = "v", # align all 3 hist vertically
          labels = "AUTO", # add the auto label A, B, C
          label_size = 11) # make label font size 11)
```



The general distribution of 'value_tons' is right skewed. The log transformation reduces the large range and allows us to identify and read the smaller values more easily. The sqrt also reduces the large range, but not as heavily as the log transformation and it also tries to minimize the skewdness.

Question 5

For question 5, you will wrangle `water_pollution` data set from the EPA Water Pollution Search, where you can also find a glossary of terms regarding the `water_pollution` data set.

According to US EPA, "The Water Pollution Search allows users to look at the Toxic Release Inventory (TRI) and search for the largest surface water discharges based on total mass and toxicity to help identify discharges that may have the greatest impact on environmental or human health." The discharge data is distinguished between major and non-major watersheds. For this lab assignment **please report discharges in the units lb/year**

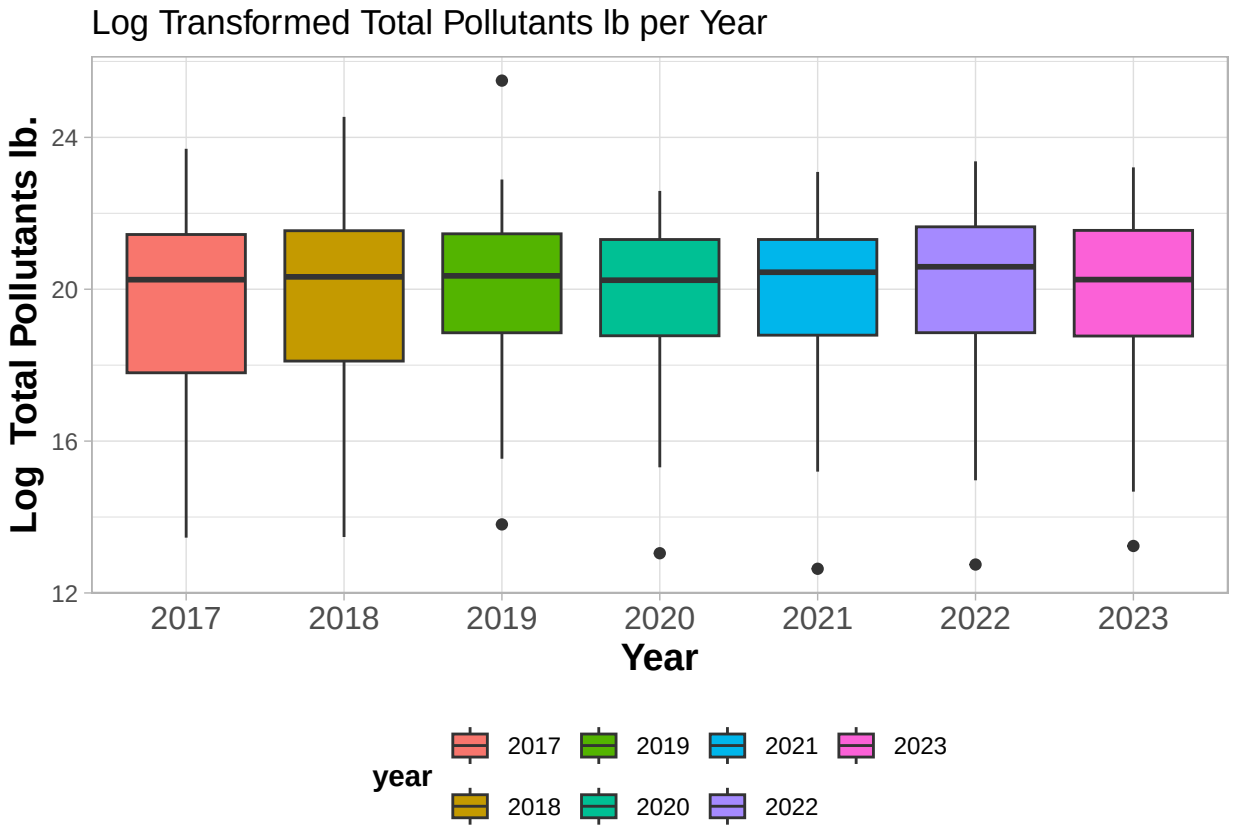
To receive full credit for plots, you must add **informative labels** using `labs()` & `ggtitle()` functions AND `theme_*()` to **change background**. The column names in 'water_pollution' are long and must be changed to titles that are easier to read.

a. **For the 'epa_region' 5, 6, & 9, create a boxplot, illustrating the relationship between the log transformed 'total_pollutants_lb_yr_majorwateshed' and 'year'**

- *hint: use filter(epa_region . . .), mutate(year = as.factor(year)), before piping data into ggplot()*

```
water_pollution %>%
  filter(epa_region %in% c(5,6,9)) %>%
  mutate(year = as.factor(year),log_lbs = log(total_pollutants_lb_yr_majorwateshed)) %>%
    ggplot(aes(x = year, y= log_lbs, fill = year)) +
    geom_boxplot() +
```

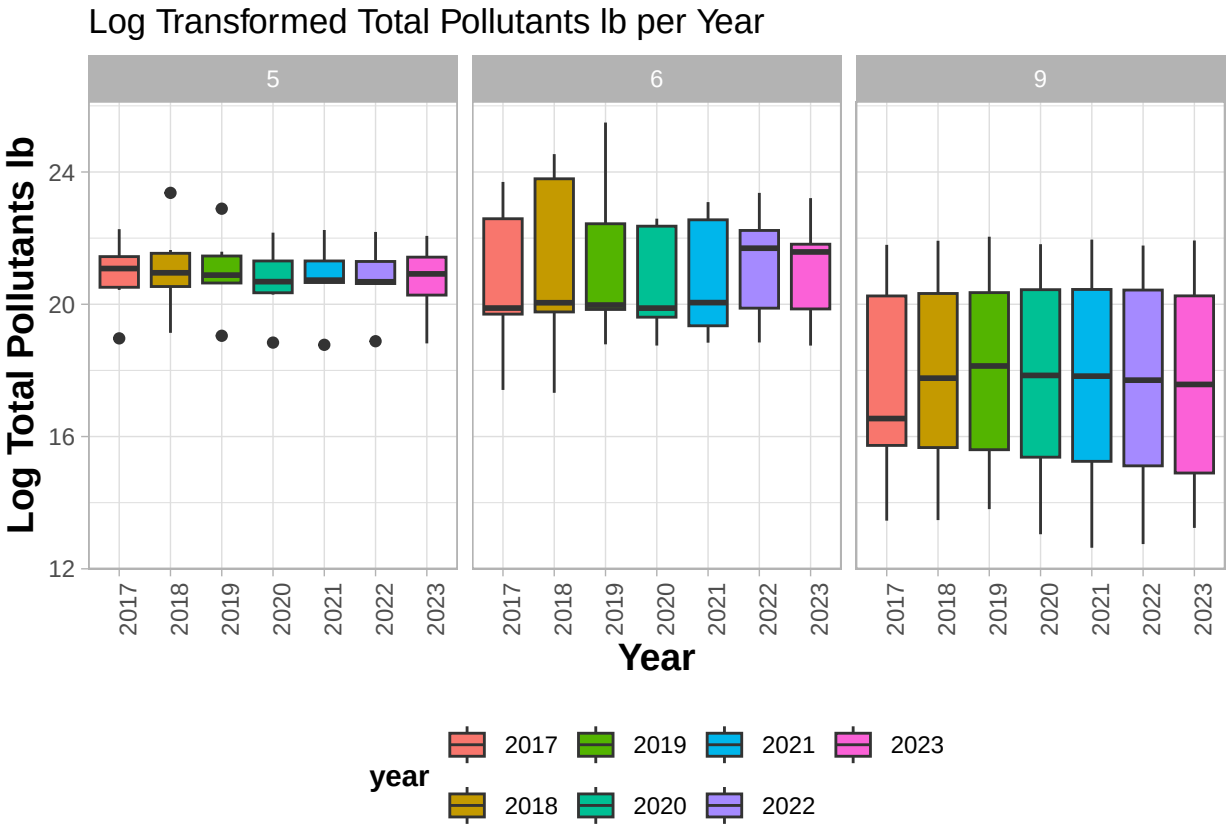
```
  theme_light() +
  labs(x = "Year", y = "Log  Total Pollutants lb.") +
  ggtitle("Log Transformed Total Pollutants lb per Year") +
  theme(axis.title = element_text(size = 14, face = "bold"),
        axis.text.x = element_text(size = 12),
        legend.title = element_text(face = "bold"),
        legend.position = "bottom")
```



b. **With the same code as part a, now add another layer of 'facet_wrap' by 'epa_region' to create a boxplot with a facet wrap**

- *hint: there should only be 3 facet wraps (region 5,6,9)*

```
water_pollution %>%
  filter(epa_region %in% c(5,6,9)) %>%
  mutate(year = as.factor(year),log_lbs = log(total_pollutants_lb_yr_majorwateshed)) %>%
    ggplot(aes(x = year, y= log_lbs, fill = year)) +
    geom_boxplot() +
  facet_wrap(~epa_region)+
  theme_light() +
  labs(x = "Year", y = "Log Total Pollutants lb") +
  ggtitle("Log Transformed Total Pollutants lb per Year") +
  theme(axis.title = element_text(size = 14, face = "bold"),
        axis.text.x = element_text(angle=90),
        legend.title = element_text(face = "bold"),
        legend.position = "bottom")
```



c. **Based on the boxplot with a facet wrap in part b, (i) which 'epa_region' has the most variation in the pollutant distribution (lb/year)? (ii) Which 'epa_region' has the least variation in pollutant distribution (lb/year)? (iii) Which 'epa_region' has on average the highest total pollutants (lb/year)?**

(i) The epa_region with the most variation is *9* as its values for log lbs has the greatest range.

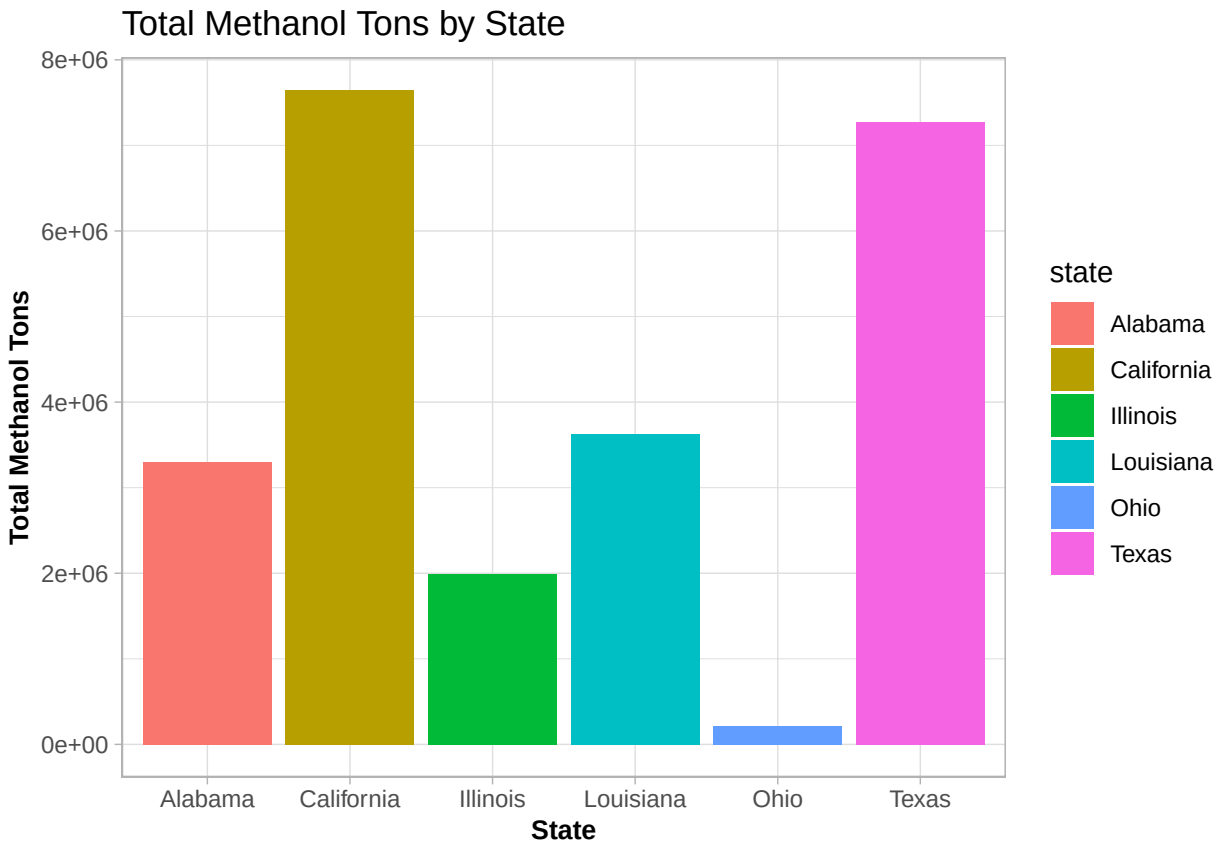
(ii) The epa_region with the least variation is *5*.

(ii) The epa_region that has on average the highest total pollutants is (lb/year) is *5*.

Extra Credit Question 6

a. With either `air_pollution` or `water_pollution` data set, create a barplot or histogram or boxplot of a relationship you're interested in. To receive full credit you must include informative titles, change the background of the plot with theme_*(), and one other ggplot() trick from Lab 3 Exercise file (i.e. alpha, element_text(), element_rec(), color = values(), etc).

```
air_pollution%>%
filter(pollutant_name == "Methanol") %>% #Wrangle data just as before
  group_by(state) %>%
  mutate(total= sum(value_tons)) %>%
  ggplot(aes(x=state, y=total,fill=state)) +
  geom_bar(stat="identity") +
  theme_light()+
  labs(x="State",y="Total Methanol Tons") +
  ggtitle("Total Methanol Tons by State") +
  theme(axis.title = element_text(size = 10, face = "bold"))
```



End of Homework 3