

Question 0

Question 1

Question 2

Question 3

Submission Instructions

Homework 1

Neal Kar (ink2105)

2/4/2020

Question 0

a)

Add your name and the date to the header of the rmd.

b)

Add a floating table of contents to the HTML document.

Question 1

The US Measles dataset contains yearly reports of measles prevalences in the US. The variables are: year, state, and prevalence.

a)

Read in the US Measles dataset.

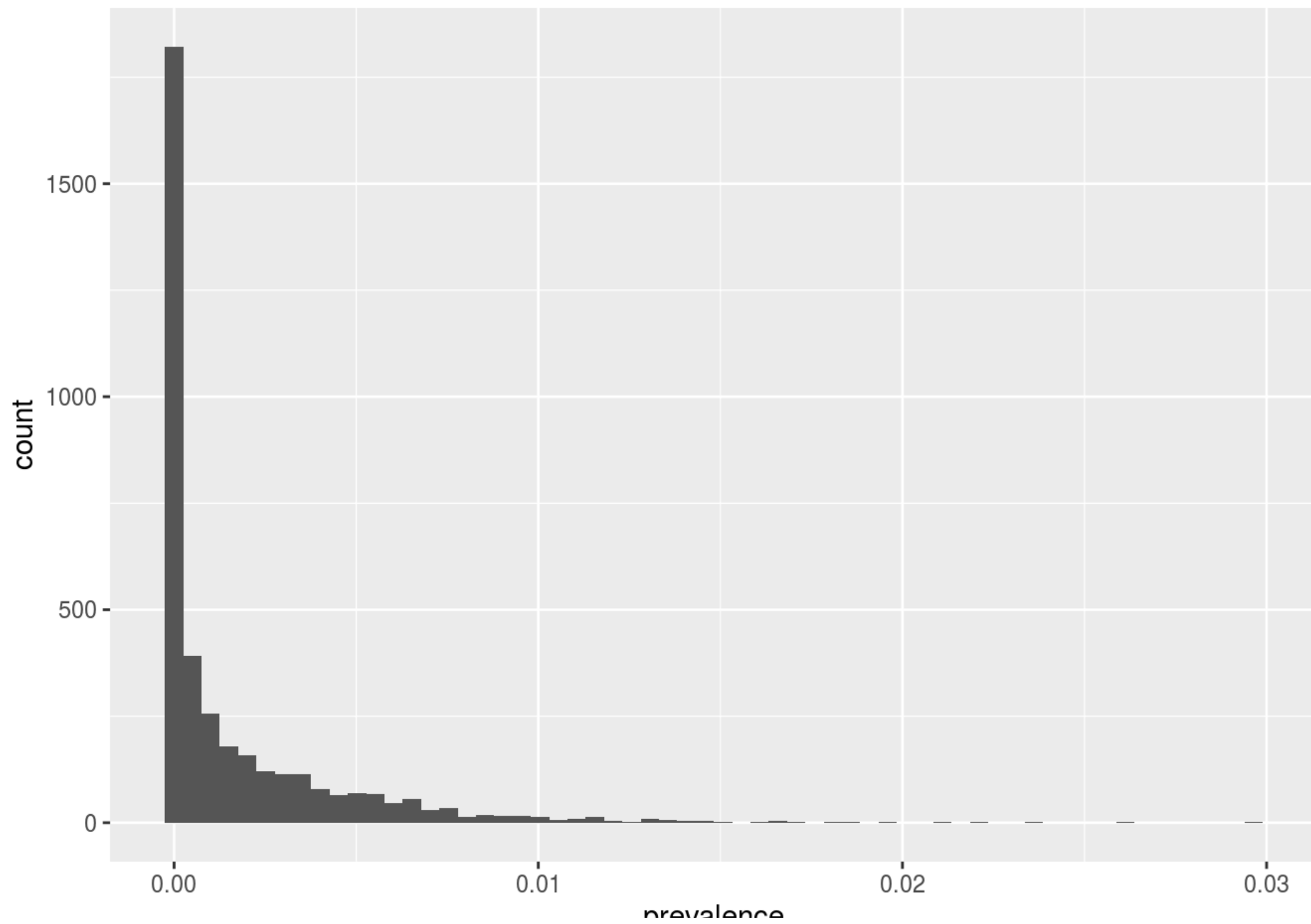
```
measles_data <- read_csv("data/us_measles.csv")
```

```
## Parsed with column specification:  
## cols(  
##   state = col_character(),  
##   year = col_double(),  
##   prevalence = col_double()  
## )
```

b)

Create a histogram of measles prevalence and comment on the shape of the distribution.

```
ggplot(data = measles_data) +  
  geom_histogram(aes(x = prevalence), bins = 60)
```



The distribution is right-skewed.

c)

Calculate the mean and standard deviation of measles prevalence for years 1940 and 1990. Interpret these values in context.

```
#Year = 1940
filter_1940 <- measles_data %>%
  filter(year == 1940) %>%
    pull(prevalence)
  mean_prev_1940 <- mean(filter_1940)
  sd_prev_1940 <- sd(filter_1940)

#Year = 1990
filter_1990 <- measles_data %>%
  filter(year == 1990) %>%
    pull(prevalence)
  mean_prev_1990 <- mean(filter_1990)
  sd_prev_1990 <- sd(filter_1990)
```

The mean prevalence of measles in 1940 was 0.0026519. with a standard deviation of 0.0026536. This means that in 1940, the average prevalence of measles per state in the U.S. was about 2.6519139 per 1000 people. It also means, 68.2% of

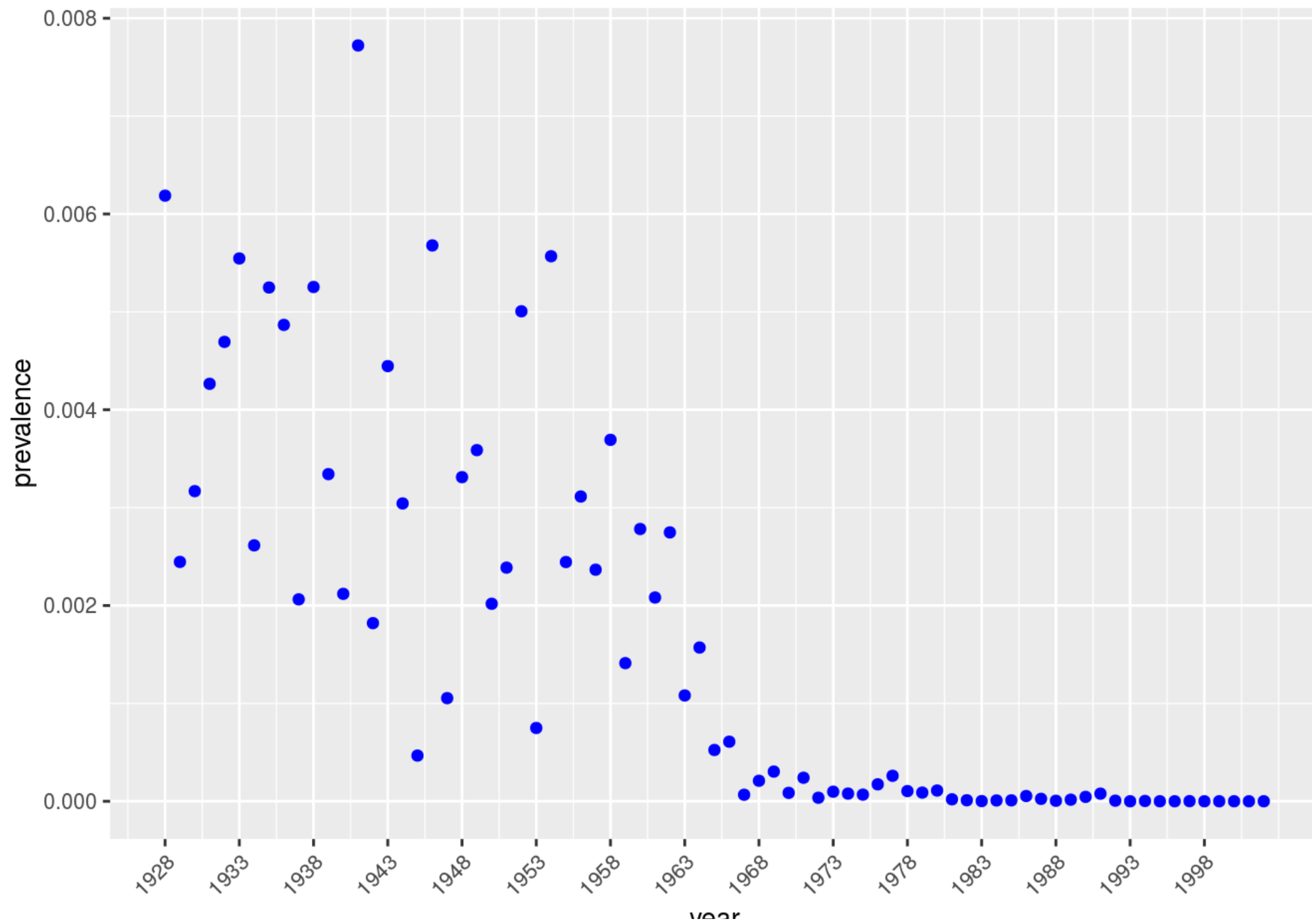
U.S. states in 1940 had measles prevalence between 0 and 5.3055393 per 1000 people.

The mean prevalence of measles in 1990 was 4.6190804×10^{-5} with a standard deviation of 6.0316885×10^{-5} . This means that in 1990, the average prevalence of measles per state in the U.S. was about 0.0461908 per 1000 people, which was about 50x lower than in 1940. It also means, 68.2% of U.S. states in 1990 had measles prevalence between 0 and 0.1065077 per 1000 people. So overall, the prevalence of measles in the U.S. was much lower in 1990, compared to 1940.

d)

Create a scatterplot for measles prevalence by year in New York, and set the points to the color of your choice (black isn't allowed). What does this graph demonstrate?

```
filter_NY <- measles_data %>%  
  filter(state == "New York")  
  ggplot(filter_NY) +  
    geom_point(aes(x = year, y = prevalence), color = "blue")  
+  
  scale_x_continuous(breaks = seq(1928, 2002, by = 5)) +  
  theme(text = element_text(size=10),  
        axis.text.x = element_text(angle=45, hjust=1))
```



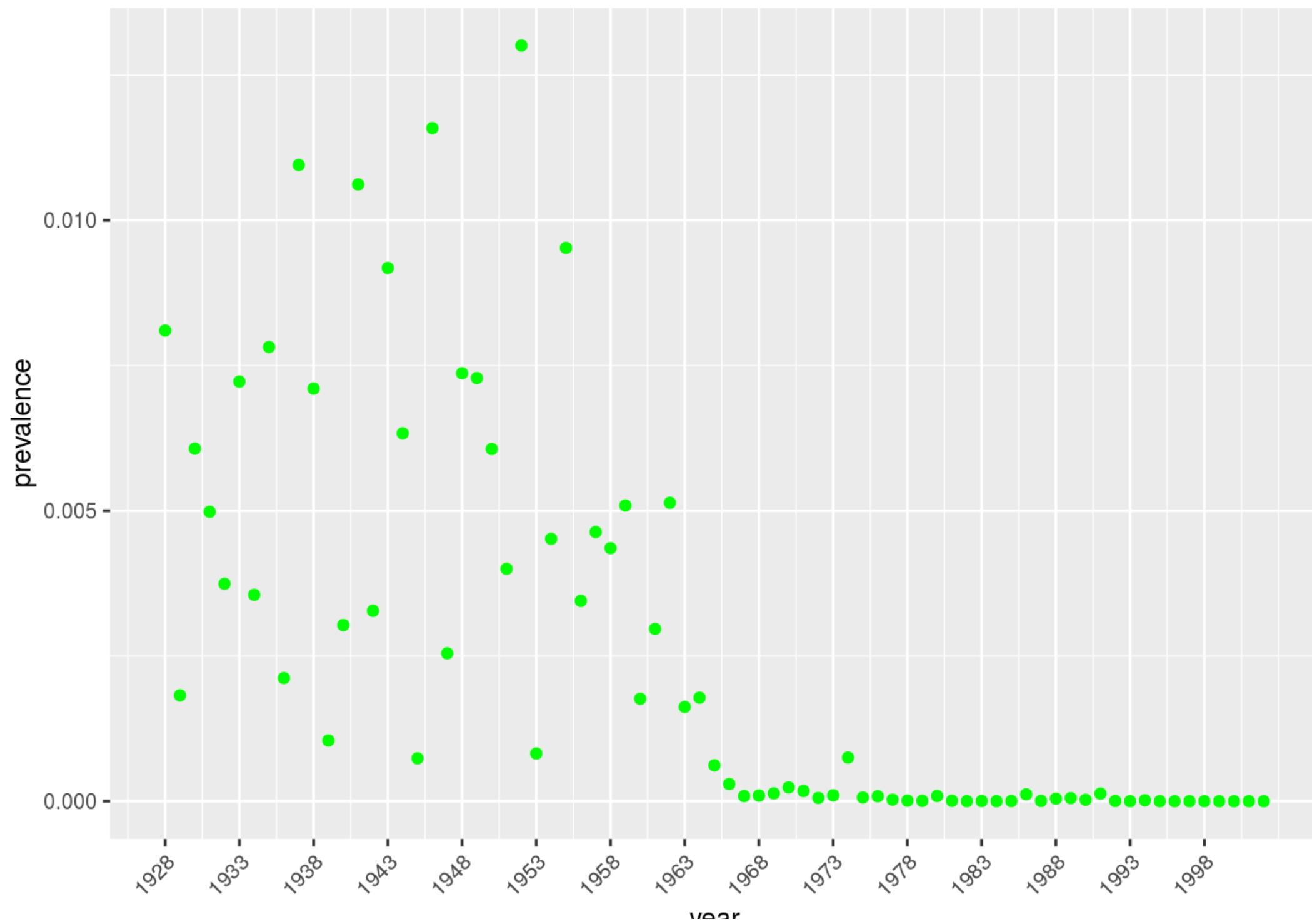
year

This graph demonstrates that measles prevalence in New York was relatively high until about 1968. There is a noticeable drop in prevalence from 1964-1968. From 1968 onwards, the drop in prevalence was sustained, and the prevalence remained at a much lower rate than it had in the past. This suggests some sort of scientific breakthrough or policy intervention occurred sometime around 1964-1968. It turns out that the first measles vaccine was licensed in the U.S. in 1963, and an improved measles vaccine was licensed in the U.S. in 1968. The availability of the vaccine may explain the drop in measles prevalence in New York.

e)

Create a scatterplot for measles prevalence by year in another state of your choice, and set the points to the color of your choice (black isn't allowed). How does this graph compare to the graph of New York's prevalence by year?

```
filter_CT <- measles_data %>%  
  filter(state == "New Jersey")  
  ggplot(filter_CT) +  
    geom_point(aes(x = year, y = prevalence), color = "green")  
+  
  scale_x_continuous(breaks = seq(1928, 2002, by = 5)) +  
  theme(text = element_text(size=10),  
        axis.text.x = element_text(angle=45, hjust=1))
```



year

This graph of measles prevalence in New Jersey is very similar to that of New York. There is a noticeable drop in prevalence from 1964-1968, and from 1968 onwards, the drop in prevalence is sustained (with the exception of 1974).

Question 2

The Tooth Growth dataset contains the results of an experiment conducted on 60 Guinea Pigs to evaluate the effect of vitamin C supplements on tooth growth. The variables are: Length (tooth length in cm), Supplement (supplement type, either VC-ascorbic acid or OJ-orange juice), and Dose (in milligrams/day).

a)

Read in Tooth Growth data. Check the data carefully...

```
tooth_growth <- read_csv("data/ToothGrowth.csv", skip = 2)
```

```
## Parsed with column specification:  
## cols(  
##   Length = col_double(),  
##   Supplement = col_character(),  
##   Dose = col_double()  
## )
```

b)

How many variables and observations are in this dataset? What is each variable's type?

Note: This can be determined with or without code.

There are 60 observations and 3 variables in this dataset. The “length” variable is a numeric variable, the “supplement” variable is a character variable, and the “dose” variable is a numeric variable.

c)

Calculate the mean and standard deviation of tooth length for each dosage and report them.

```
#Dose = 0.5 mg/day
filter_0.5 <- tooth_growth %>%
  filter(Dose == 0.5) %>%
  pull(Length)
mean_length_0.5 <- mean(filter_0.5)
sd_length_0.5 <- sd(filter_0.5)
```

```
#Dose = 1.0 mg/day
filter_1.0 <- tooth_growth %>%
  filter(Dose == 1.0) %>%
  pull(Length)
mean_length_1.0 <- mean(filter_1.0)
sd_length_1.0 <- sd(filter_1.0)
```

```
#Dose = 2.0 mg/day
filter_2.0 <- tooth_growth %>%
  filter(Dose == 2.0) %>%
  pull(Length)
```

```
mean_length_2.0 <- mean(filter_2.0)
sd_length_2.0 <- sd(filter_2.0)
```

Among guinea pigs receiving 0.5 mg/day of supplement, the mean tooth length was 10.605 centimeters with a standard deviation of 4.4997632 cm.

Among guinea pigs receiving 1.0 mg/day of supplement, the mean tooth length was 19.735 centimeters with a standard deviation of 4.4154364 cm.

Among guinea pigs receiving 2.0 mg/day of supplement, the mean tooth length was 26.1 centimeters with a standard deviation of 3.7741503 cm.

d)

Use R as a calculator to calculate the 95% confidence intervals for tooth length for each dosage.

```
#Compute standard error (SE)
# SE = SD / sqrt(n), where SD = standard deviation and n = sample size (60)

#For dose = 0.5 mg/dat
se_length_0.5 <- sd_length_0.5 / sqrt(60)

#For dose = 1.0 mg/dat
se_length_1.0 <- sd_length_1.0 / sqrt(60)

#For dose = 2.0 mg/dat
se_length_2.0 <- sd_length_2.0 / sqrt(60)

#95% CI = Mean +/- t(SE)
#t = 1.960

#For dose = 0.5 mg/day
lower_bound_length_0.5 <- mean_length_0.5 - (1.960 * (se_length_0.5))
```

```
upper_bound_length_0.5 <- mean_length_0.5 + (1.960 * (se_length_0.5))

#For dose = 1.0 mg/day
lower_bound_length_1.0 <- mean_length_1.0 - (1.960 * (se_length_1.0))
upper_bound_length_1.0 <- mean_length_1.0 + (1.960 * (se_length_1.0))

#For dose = 2.0 mg/day
lower_bound_length_2.0 <- mean_length_2.0 - (1.960 * (se_length_2.0))
upper_bound_length_2.0 <- mean_length_2.0 + (1.960 * (se_length_2.0))
```

Among guinea pigs receiving 0.5 mg/day of supplement, the 95% confidence interval of tooth length was (9.4664028 cm, 11.7435972 cm).

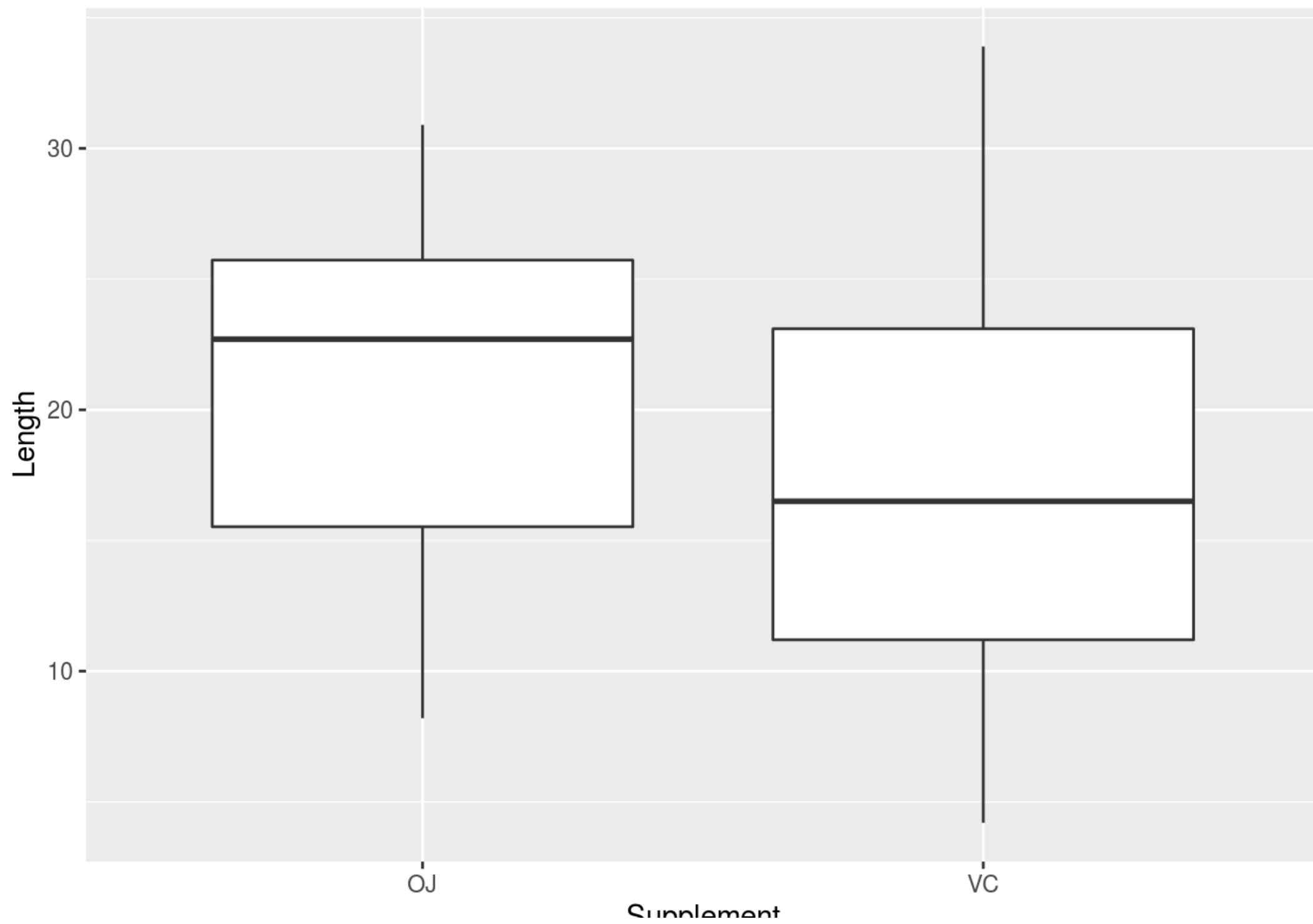
Among guinea pigs receiving 1.0 mg/day of supplement, the 95% confidence interval of tooth length was (18.6177404 cm, 20.8522596 cm).

Among guinea pigs receiving 2.0 mg/day of supplement, the 95% confidence interval of tooth length was (25.1450082 cm, 27.0549918 cm).

e)

Make a boxplot for tooth length based on supplement. Comment on the distribution, and any differences between OJ and VC supplement groups.

```
ggplot(data = tooth_growth) +  
  geom_boxplot(aes(x = Supplement, y = Length))
```



The distribution for the OJ group is right-skewed whereas the distribution for the VC group is left-skewed. The median tooth length of the OJ group was longer than the median tooth length of the VC group. There was a greater range of values for tooth length among the VC group.

Question 3

The Murders dataset contains information on murder rates in the US in 2012. The variables are: state, region, population (number of residents in the region), and total_murders (number of murders in the region).

a)

The code below attempts to read in the murders dataset but requires additional options to read in the data correctly. Take a look at the data file and check the data carefully after you read it in...

```
murders <- read_excel("data/murders.xlsx", sheet = "murders", range = "E5:H56")
```

b)

How many variables and observations are in this dataset? What is each variable's type?

There are 51 observations and 4 variables in this dataset. The “state” variable is a character variable, the “region” variable is a character variable, the “population” variable is a numeric variable, and the “total_murders” variable is a numeric variable.

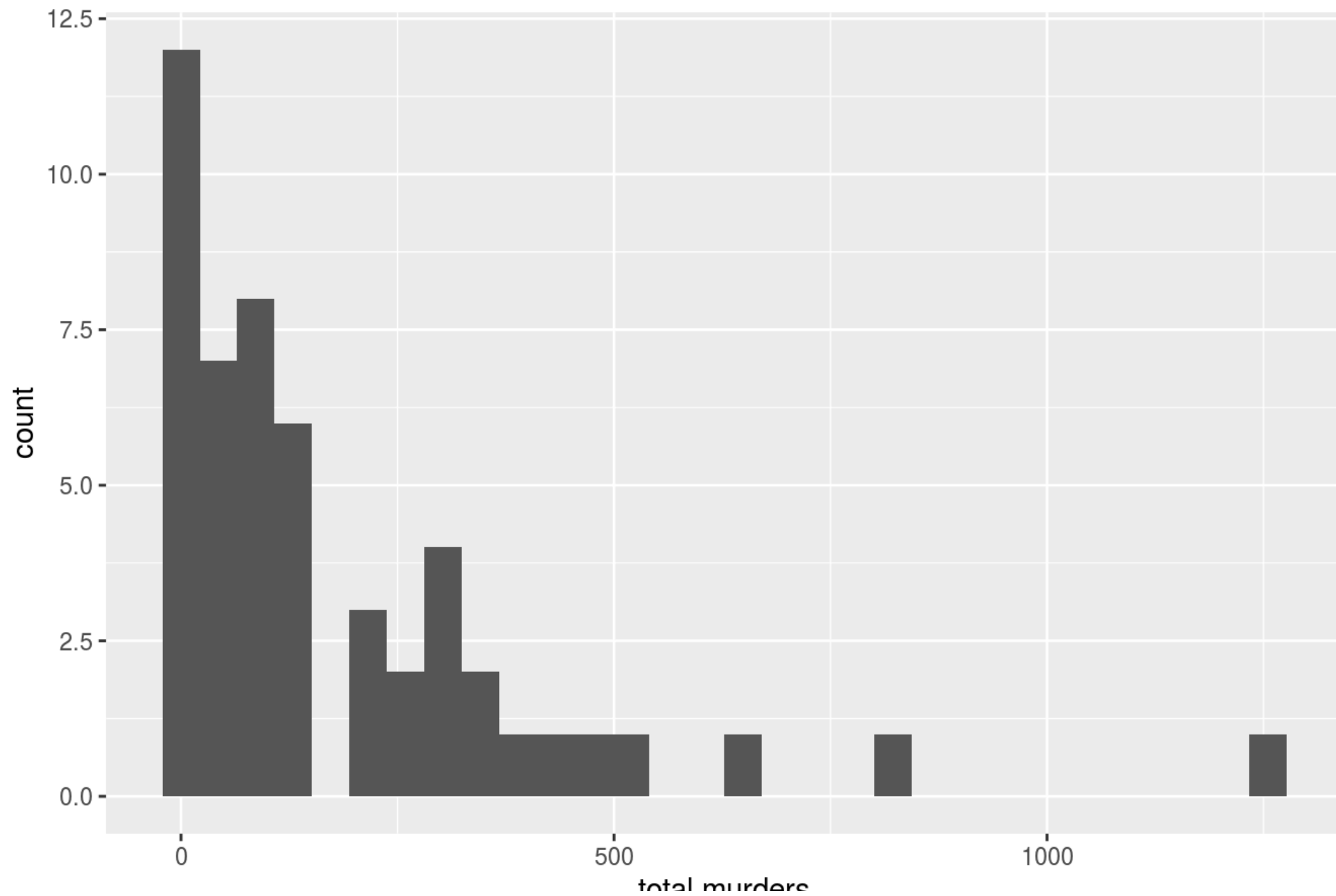
c)

The following code is creating a histogram of the total murders, yet contains four errors. Identify and correct each error, and describe what was wrong below the graph. Once you have fixed all errors, be sure to remove `eval = FALSE` from the code chunk options.

```
ggplot(data = murders) +  
  geom_histogram(aes(x = total_murders)) +  
  labs(title = "Histogram of Murders", x = "total murders")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

Histogram of Murders



Error 1: The name of the dataframe in line 255 was originally “murder”, which was incorrect. I fixed it so that the name of the dataframe is now “murders”.

Error 2: Originally, the code had “%>%” instead of “+” at the end of lines 255 and 256. I replaced “%>%” with “+” at the end of lines 255 and 256.

Error 3: Originally, line 256 was missing the aes() function. I added the aes() function to line 256.

Error 4: The “total_murders” variable in line 256 was originally denoted as the y-variable. I fixed it so that it’s now denoted as the x-variable.

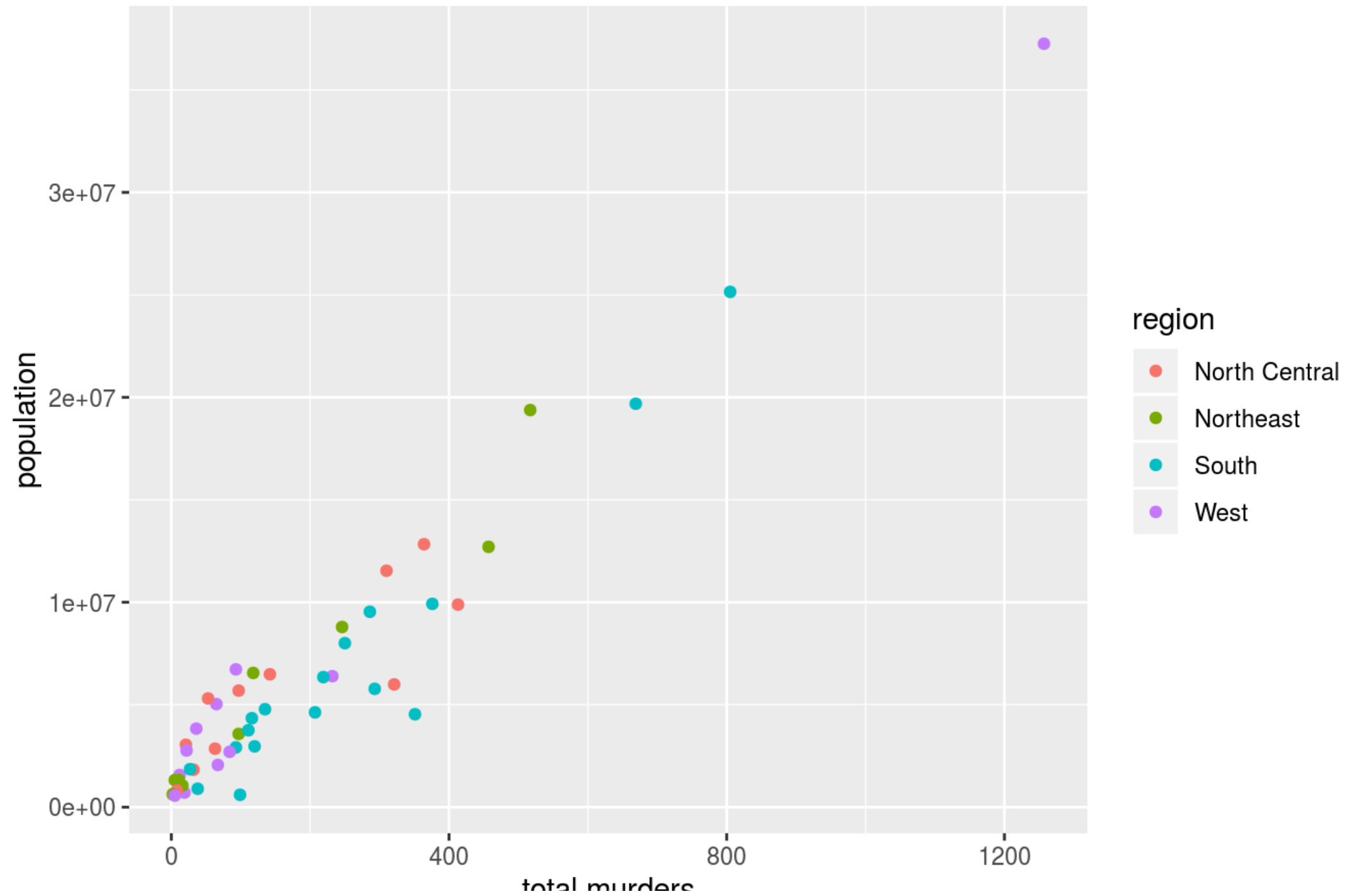
Error 5: The x-axis was originally labeled as “cities.” I fixed it so that the x-axis is now labeled “total murders.”

d)

The following code attempts to visualize total murders by population and region, but there are four errors. Find and correct each, and describe what was wrong below the graph. Once you have fixed all errors, be sure to remove `eval = FALSE` from the code chunk options.

```
ggplot(data = murders) +  
  geom_point(aes(y = population, x = total_murders, color = region)) +  
  labs(title = "Murders by Population and Region", x = "total murders", y = "population")
```

Murders by Population and Region



Error 1: Originally the code said “geom_scatter” in line 277. I fixed it so that it now says “geom_point.”

Error 2: Originally in line 277, the x-variable was written as “totalmurders”, which was incorrect. I fixed it so that it’s now written correctly as “total_murders”.

Error 3: Originally in line 277, “color = region” was not inside the same set of parentheses as the y-variable and x-variable. I corrected it so that it’s inside the same set of parentheses as the y-variable and x-variable.

Error 4: The x-axis and y-axis were originally mis-labeled as “population” and “region”, respectively. I fixed it so that the x-axis is now labeled as “total murders” and the y-axis is now labeled as “population”.

e)

Without coding, think of a research question you could answer with the Murders dataset. Describe this question and how you might answer it.

Research question: How did murder rates differ by region in the United States in 2012? Murder rate would be defined as the total number of murders in a given area divided by the total population of that area. I would compute the murder rate for a given region by first adding up the population of each state in the region to get a value for regional population. Then, I would add up the total number of murders in each state in the region to get a value for total regional murders. Then, I would take the total regional murders and divide it by the regional population to obtain a regional murder rate. I would convert this rate to a rate per 1000 people by multiplying the quotient by 1000. I would obtain murder rate per 1000 in each region: Northeast, South, North Central, and West. I would compare which region had the lowest murder rate and which had the highest murder rate.

Submission Instructions

When you are finished with this homework, Knit it to HTML and then submit both your .rmd and .html files to courseworks.