STA130H1F – Winter 2020

Week 2 Practice Problems

N. Moon and L. Bolton [ADD YOUR NAME HERE]

Instructions

How do I hand in these problems for the January 16 deadline?

Your complete .Rmd file that you create for these practice problems and the resulting pdf (i.e., the one you 'Knit to PDF' from your .Rmd file) must be uploaded into a Quercus assignment (link: https://q.utoronto.ca/courses/138992/assignments/278855) by 11:59PM, on January 16. Late problem sets or problems submitted another way (e.g., by email) are *not* accepted.

What should I bring to tutorial on January 17?

R output (e.g., plots) for Questions 1 AND 2. You can either bring a hardcopy or bring your laptop/tablet with the output.

Tutorial Grading

Tutorial grades will be assigned according to the following marking scheme.

	Mark
Completion of required problems (due on Quercus the day before your tutorial)	1
Attendance for the entire tutorial	1
In-class exercises	4
Total	6

Practice Problems

You may have seen in the news that Australia is currently suffering from one of the worst bushfire seasons it has ever seen. High temperatures and low rainfall have contributed to this crisis. In these practice problems you will explore historical data on rainfall and maximum daily temperatures for the 17th of January. Question 1 focusses on the temperature data, Question 2 the rainfall data and you will briefly summarise what you've learned in Question 3.

It is summer in Australia right now. You can view the most recent daytime maximum temperatures *here*.

[Question 1]

The `temp` data has the following variables for January 17 over the past several decades.

variable	description
city_name	City name
temperature	Highest daily temperature (Celsius)
decade	Decade the data is from

```
temp <- read_csv("temp_Jan17.csv")

glimpse(temp)
```

```
## Observations: 721
## Variables: 3
## $ city_name   <chr> "PERTH", "PERTH", "PERTH", "PERTH", "PERTH", "PERTH", "...
## $ temperature <dbl> 25.7, 37.5, 32.9, 25.7, 32.6, 31.3, 35.5, 27.5, 37.4, 2...
## $ decade      <chr> "1910s", "1910s", "1910s", "1910s", "1910s", "1910s", "...
```

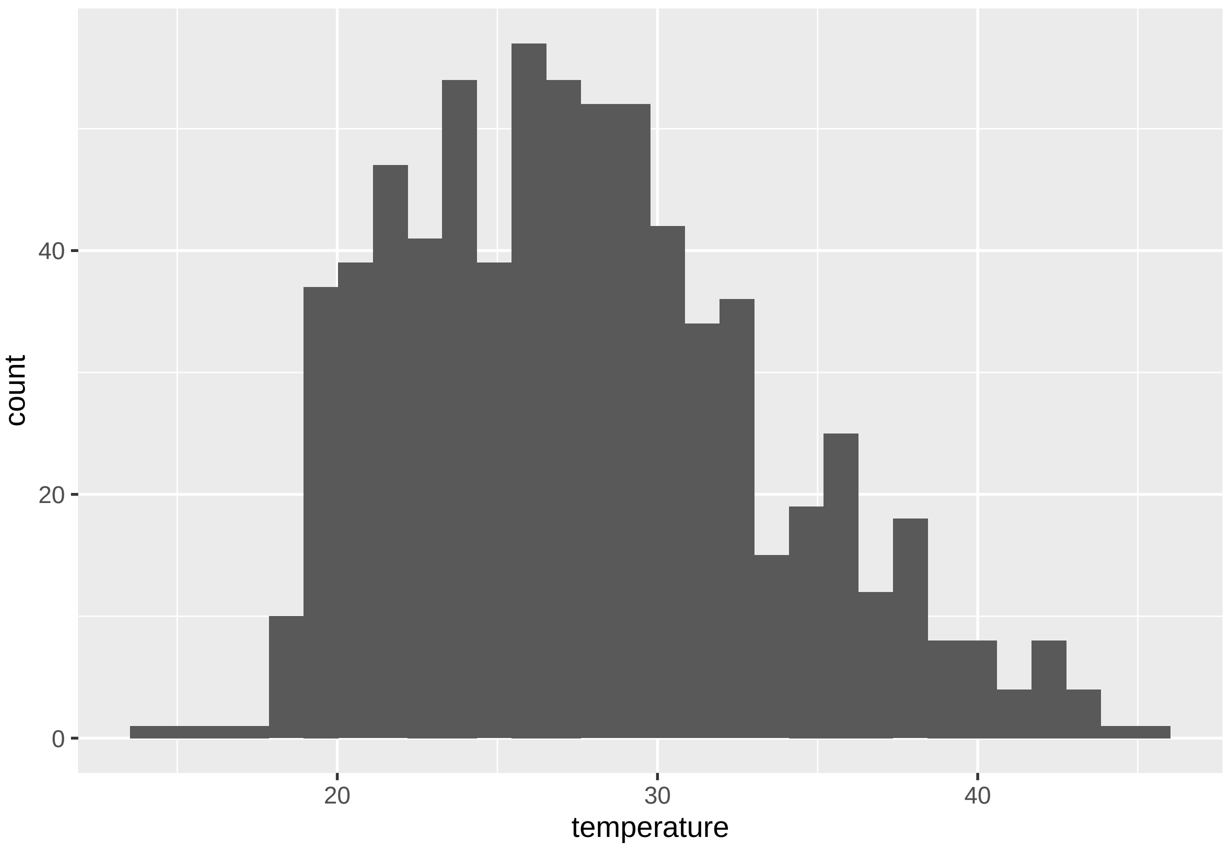
(a) What R data type describes the `temperature` variable?

The values (i.e., temperature in Celcius) in the variable `temperature` are quantiative. The data stored in `temperature` is in double format, a numeric R data type.

(b) We are interested in the distribution of the highest temperature on January 17. Consider each of the following graphical summaries. If it will show the distribution of temperatures, use R to produce the graph (note: you can click on the Insert button above and choose R to add an R chunk where you can produce your graph). If the graph will not show the distribution of temperatures, *do not produce a graph* and instead explain why it would not be appropriate.

Histogram

```
ggplot(temp, aes(x = temperature)) +
  geom_histogram()
```



Bar Plot

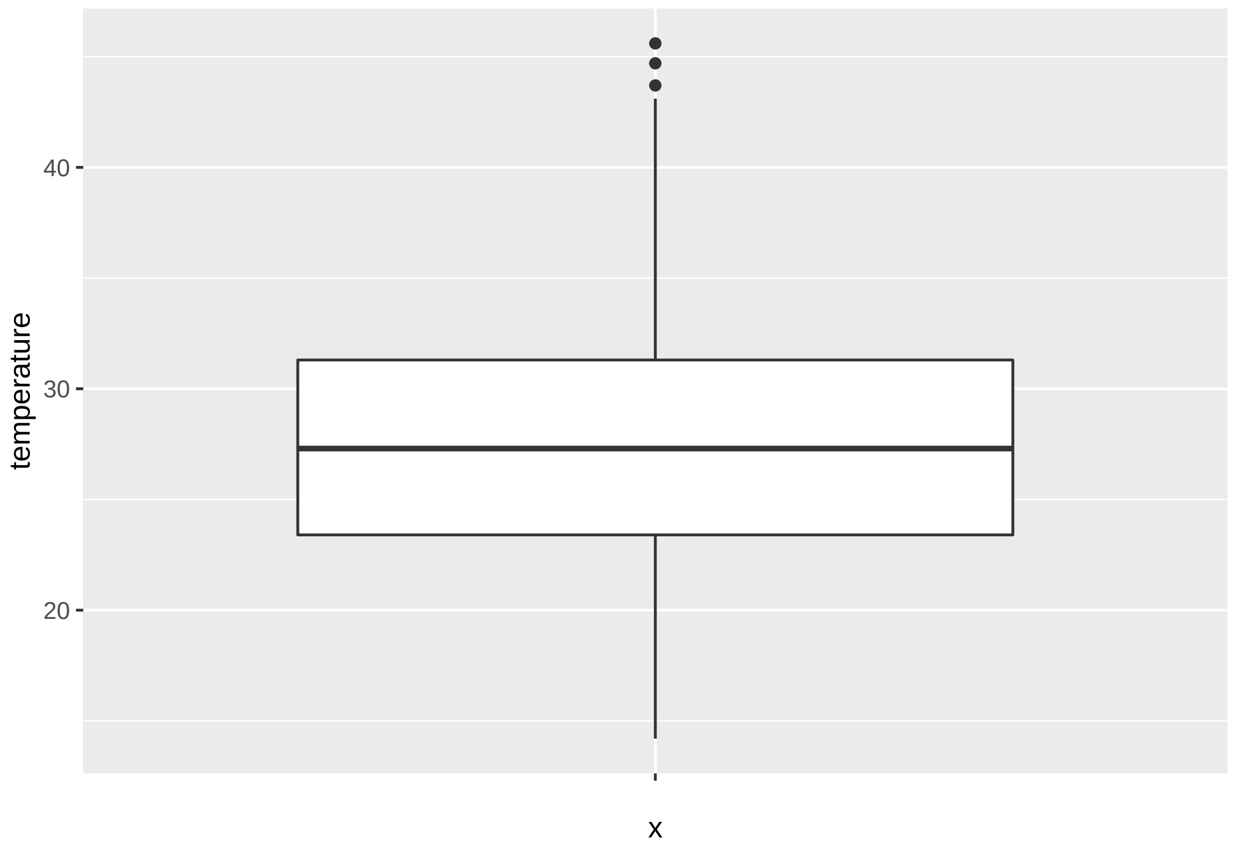
This is not an appropriate visualisation of the distribution of temperature data because the temperature is quantitative, not categorical. Each distinct temperature isn't interesting as a category; it respresents an amount on a numerical scale.

Scatterplot

This is not an appropriate visualisation of the distrbution of temperature data. Scatterplots summarise the association between two quantitative variables. Here we just have one quantative variable (i.e., `temperature`) and we are interested in its distribution; not its association with another quantitative variable.

Boxplot

```
ggplot(data = temp, aes(x = "", y = temperature)) +
  geom_boxplot()
```

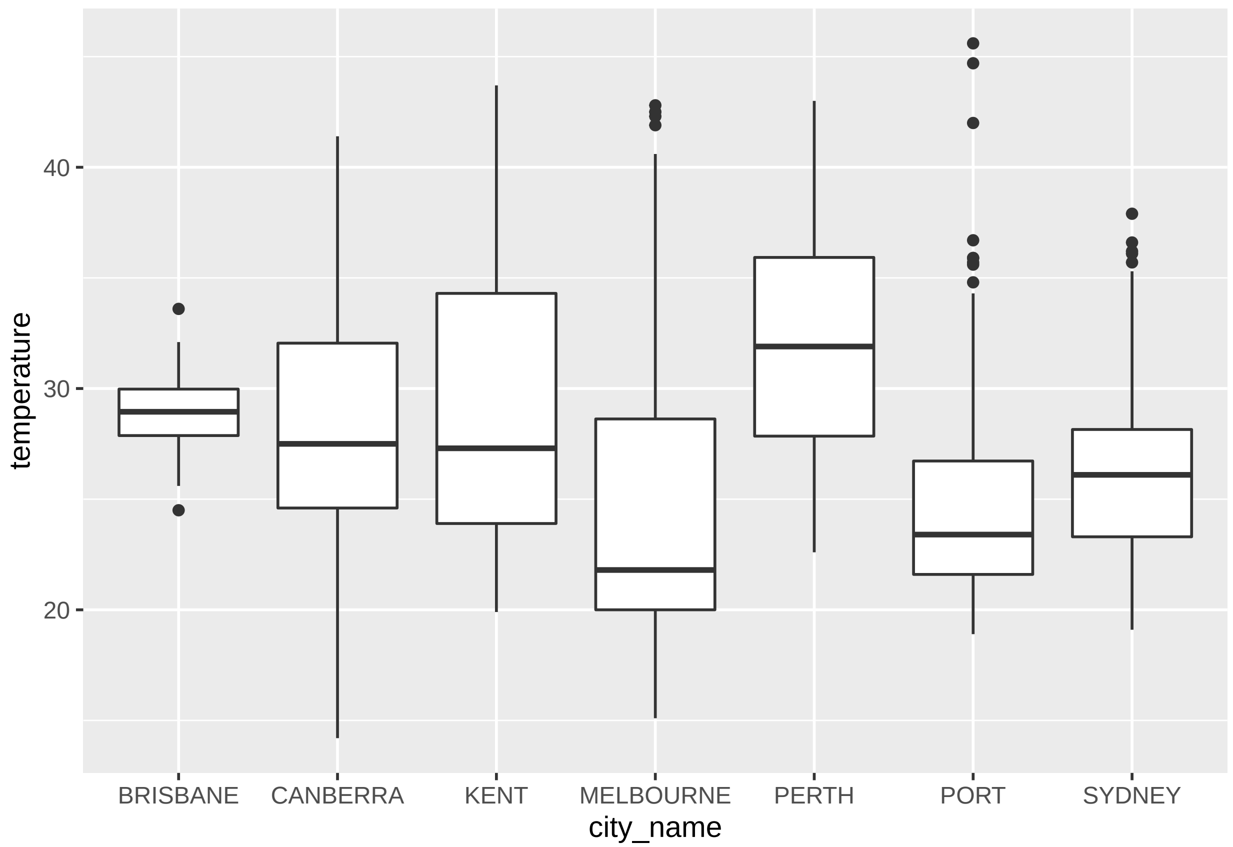


(c) Using *only* the graph(s) you produced in part (b), make a prediction about how the mean and median of the temperatures compare. Refer to appropriate graph(s) in part (b) to justify your answer.

I predict that the mean temperature will be slightly more than the median. This is because the distribution of is somewhat right skewed (we can see this from the longer right tail in the histogram and boxplot) so the mean will be pulled up from the middle value of the temperature data.

(d) Create side-by-side box plots to compare the distribution of highest daily temperatures by city.

```
ggplot(data = temp, aes(x = city_name, y = temperature)) +
  geom_boxplot()
```



(e) Based *only* on your plot in (d), answer the following questions:

Which city had the smallest inter-quartile range?

Brisbane

Which city had the highest upper quartile?

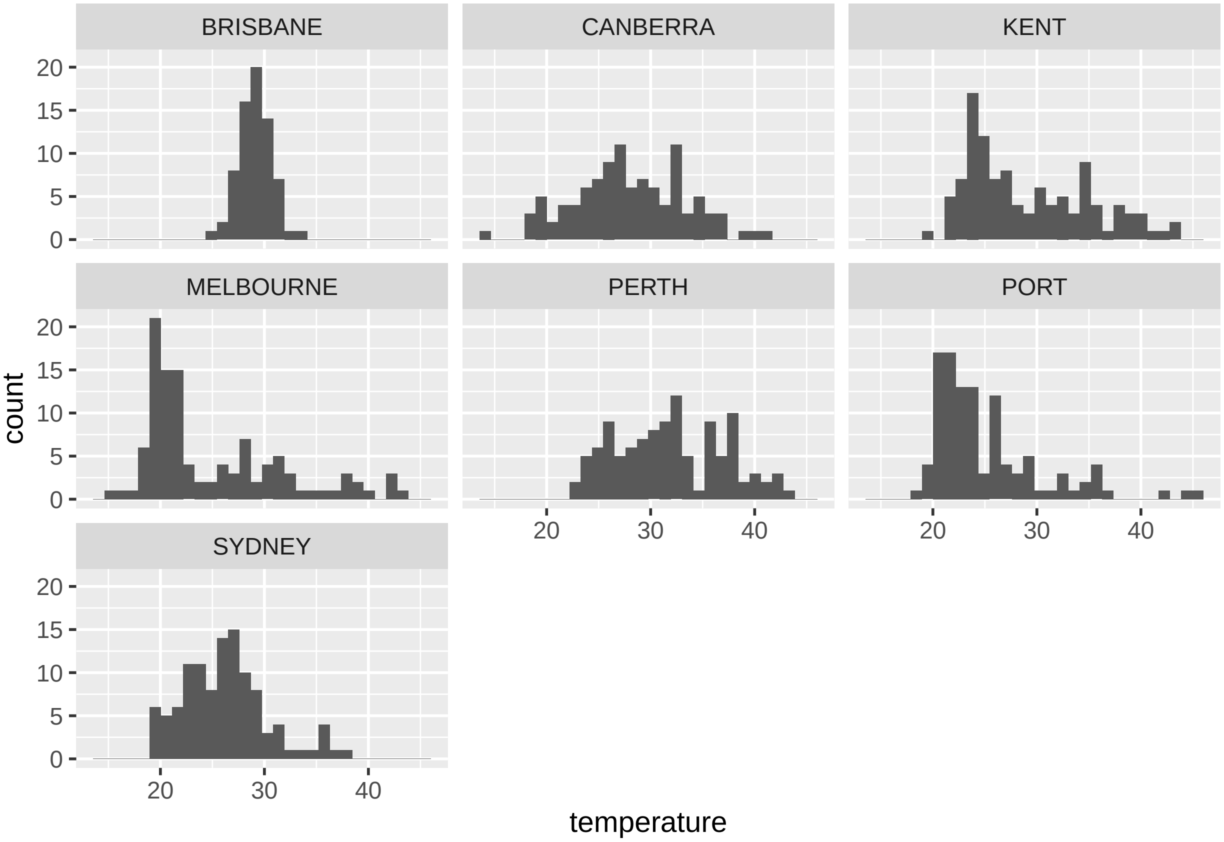
Perth

Which city had the highest temperature value?

Port

(d) Create a new plot with histograms of temperature for each city. Hint: You will find the `facet_wrap()` function seen in Week 1 lectures useful.

```
ggplot(data = temp, aes(x = temperature)) +
  geom_histogram() +
  facet_wrap(~city_name)
```



(e) Based on the plot in (d), which city historically has the most variable highest temperatures on January 17ths? I.e., for which city do you think the standard deviation of highest temperatures would be the biggest?

Melbourne.

[Question 2] In this question, you'll use the `rain_Jan17.csv` data. It contains historical rainfall data for January 17 in Australia.

variable	description
city_name	City name
rainfall	rainfall in millimetres
decade	Decade the data is from

(a) Load the data and report how many observations there are.

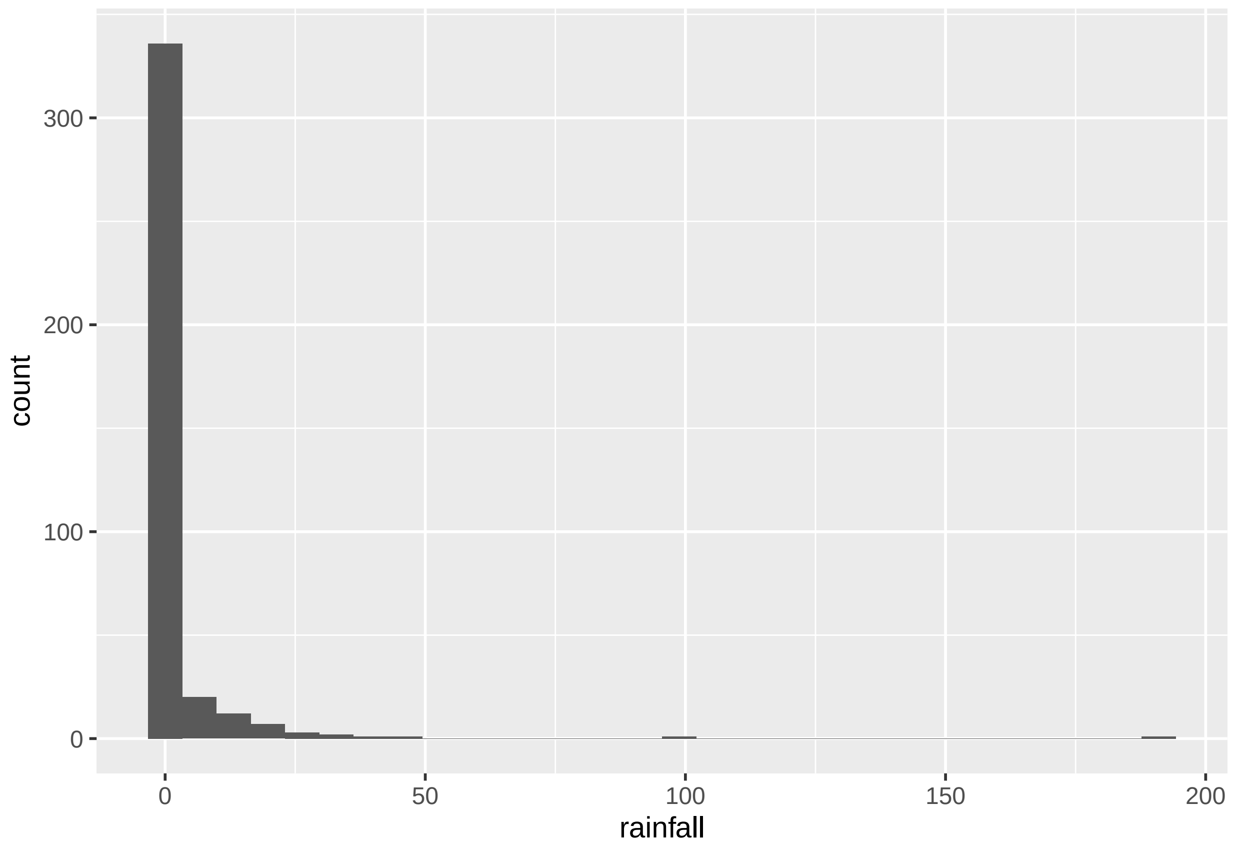
```
rain <- read_csv("rain_Jan17.csv")

glimpse(rain)
```

```
## Observations: 384
## Variables: 3
## $ city_name <chr> "Perth", "Perth", "Perth", "Perth", "Perth", "Perth", "Pe...
## $ rainfall  <dbl> 0.0, 1.5, 0.0, 1.8, 0.0, 2.8, 0.0, 0.0, 0.0, 0.0, 0.0, 0....
## $ decade    <chr> "1960s", "1960s", "1970s", "1970s", "1970s", "1970s", "19...
```

There are 384 observations.

(b) Create a histogram of the `rainfall` variable.

```
ggplot(data=rain, aes(x = rainfall)) +
  geom_histogram()
```

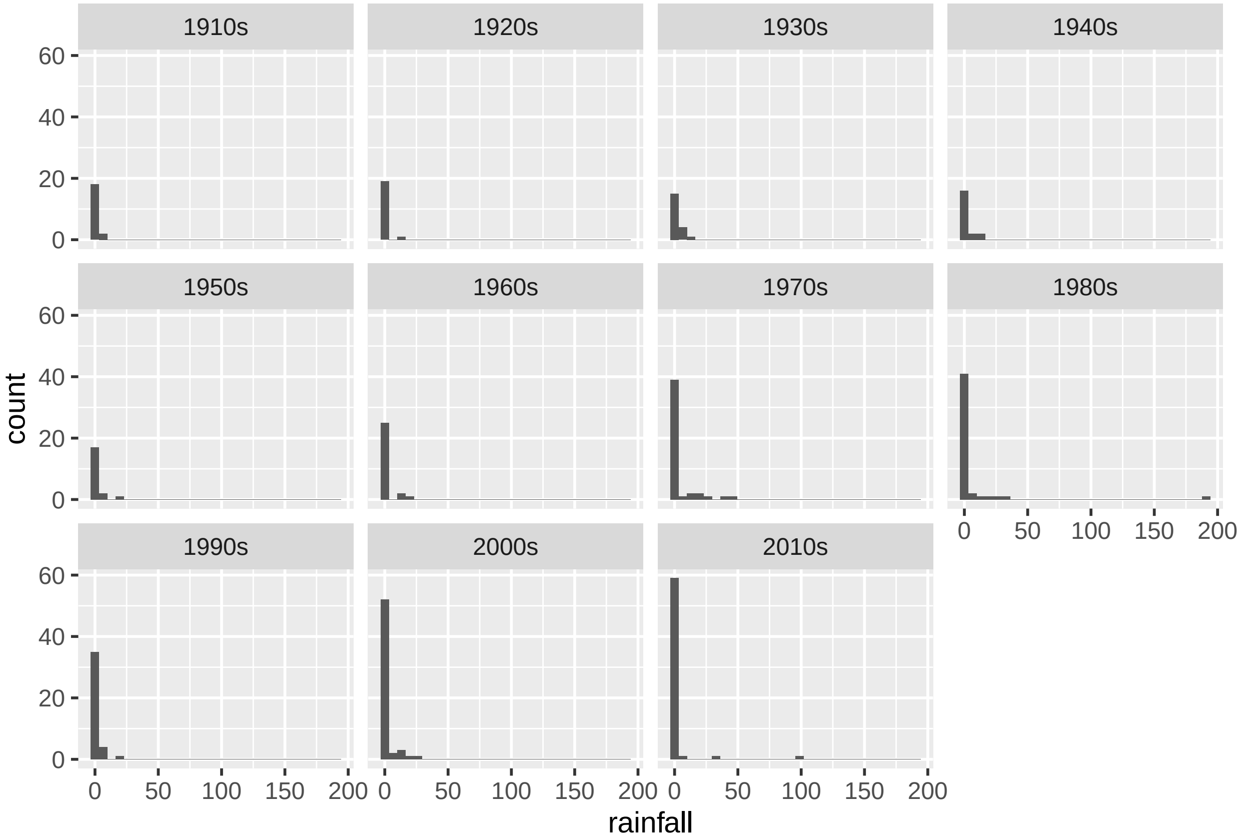


(c) Based on your histogram in (b), would the standard deviation or interquartile range be a more appropriate measure of spread (or variation) in rainfall? Justify your answer.

The interquartile range will likely be more appropriate than the standard deviation as a measure of variation in temperature because the interquartile range focuses in on just the middle half of temperatures and so is not affected by the extremely high observed temperatures. All temperature observations are used to calculate the standard deviation.

(d) Create a series of histograms for rainfall by decade.

```
ggplot(data = rain, aes(x = rainfall)) +
  geom_histogram() +
  facet_wrap(~decade)
```



(e) Based on your plot in (d), can you determine if the same number of rainfall observations (on January 17) were recorded each decade? Why or why not?

In a histogram, the height of each bar corresponds to the number of observations which fall within a particular bin (i.e., a certain range of rainfall in millimetres). In this case, when we look at the histogram for a given decade, the height of each bar is the number of observations made during that decade which are within a specific rainfall range. If we add up the heights of all the bars in one histogram, we get the total number of observations in that decade. We can clearly see that the total height of the bars for some of the early decade histograms is smaller than the total height of the bars in the later decade histograms. This implies there were more observations in some decades, i.e., that we are missing more observations for the earlier decades.

(f) The code below calculates the minimum, first quartile, median, mean, third quartile, and maximum rainfall amounts for our January 17th data for the past several decades.

```
summarise(rain, min=min(rainfall),
          Q1 = quantile(rainfall, 0.25),
          median = median(rainfall),
          mean = mean(rainfall),
          Q3 = quantile(rainfall, 0.75),
          max=max(rainfall))
```

```
## # A tibble: 1 x 6
##     min    Q1 median  mean    Q3   max
##   <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl>
## 1     0     0      0  2.51     0   191
```

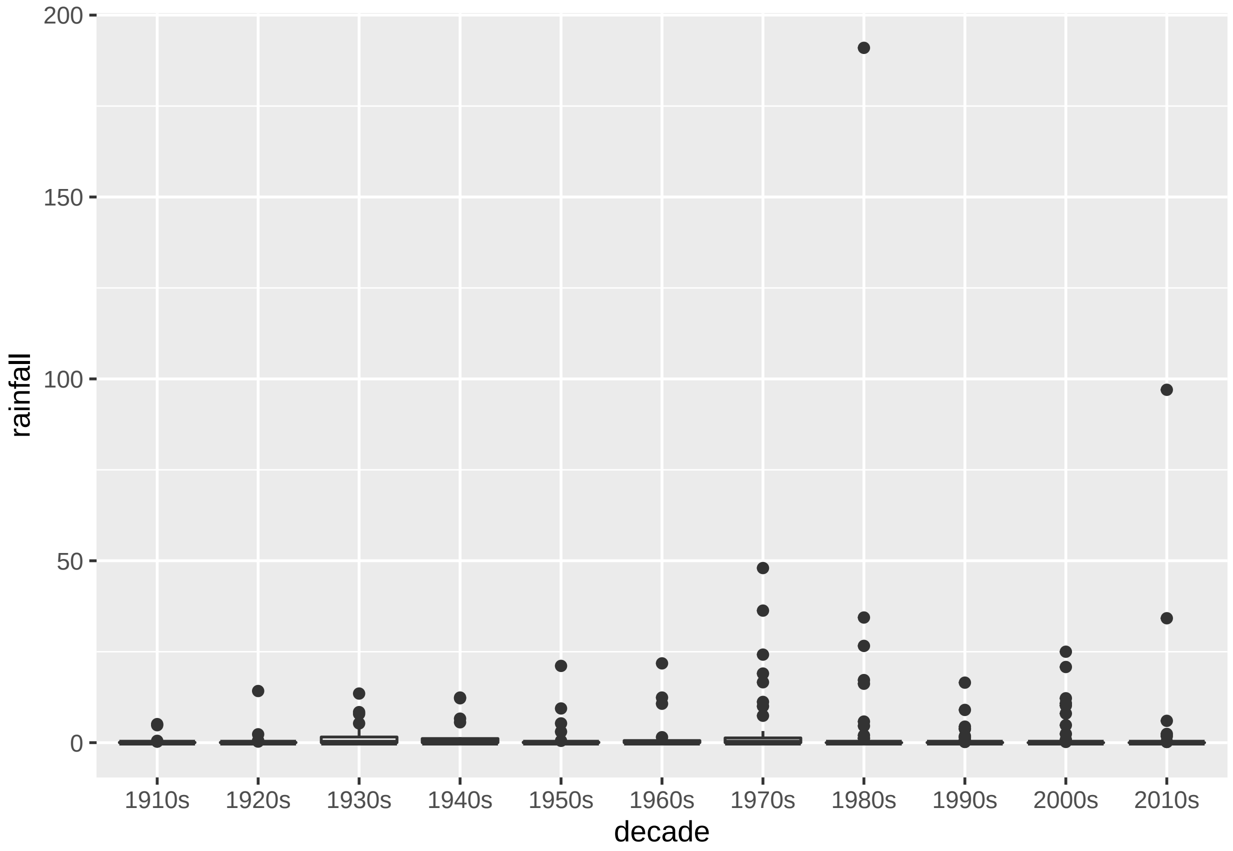
Modify the code above to calculate these statistics (i.e. minimum, Q1, median, mean, Q2, and maximum) for each decade separately. Hint: You'll need to create a new data frame using the `group_by()` function. You may want to call this new data frame `rain_grouped_by_decade` so you remember that it is grouped by decade.

```
rain_grouped_by_decade <- group_by(rain, decade)
summarise(rain_grouped_by_decade, min=min(rainfall),
          Q1 = quantile(rainfall, 0.25),
          median = median(rainfall),
          mean = mean(rainfall),
          Q3 = quantile(rainfall, 0.75),
          max=max(rainfall))
```

```
## # A tibble: 11 x 7
##    decade   min    Q1 median  mean    Q3   max
##    <chr>  <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl>
##  1 1910s      0     0      0 0.550 0.075   5.1
##  2 1920s      0     0      0 0.890 0.075  14.2
##  3 1930s      0     0      0 2.02  1.55   13.5
##  4 1940s      0     0      0 2.04  1.12   12.4
##  5 1950s      0     0      0 1.96  0.125  21.1
##  6 1960s      0     0      0 1.81  0.575  21.8
##  7 1970s      0     0      0 3.91  1.3    48
##  8 1980s      0     0      0 6.25  0     191
##  9 1990s      0     0      0 1.05  0.05   16.5
## 10 2000s      0     0      0 1.62  0      25
## 11 2010s      0     0      0 2.29  0      97
```

(g) Create side-by-side boxplots for rainfall by decade. Comment briefly on why these boxplots may look unusual.

```
ggplot(data = rain, aes(x = decade, y = rainfall)) +
  geom_boxplot()
```



This boxplot looks unusual because of how skewed this data is. You can't see the box with the upper and lower quartiles and median because the minimum, lower quartile, median and upper quartile are so similar compared to how big some of the high values are.

(h) Create a summary table calculating the minimum, mean, median, and maximum rainfall for each city, as well as how many observations with more than 10mm of rain there were. Save your summary table by giving it a name so that you can click on it in the top right panel and view it as a spreadsheet.

```r
rain_grouped_by_city <- group_by(rain, city_name)

rain_summary <- summarise(rain_grouped_by_city,
          n = n(),
          mean = mean(rainfall),
          median = median(rainfall),
          sd = sd(rainfall),
          min = min(rainfall),
          max = max(rainfall),
          x1 = sum(rainfall > 10)
          )
```

(i) View the summary table you created in (f) as a spreadsheet (by clicking on its name in the top right panel of your RStudio window) and answer the following questions:

How much rain fell on the day with the most rain?

191.0mm (in Sydney)

In which city has the mean rainfall on January 17 been the lowest?

Canberra (0.16mm on average)

Which city had the most observations with more than 10mm of rain?

Sydney

[Question 3] One of the professors in the Department of Statistical Sciences will be spending some time in Melbourne, Australia for her upcoming sabatical. Based on the the graphs and summaries you did in Questions 1 and 2, briefly describe to her what you've learned about historical temperatures and rainfall on January 17 in Melbourne. (Remember to include the units for the numerical variables your are discussing.)

On January 17, Melbourne generally has lower maximum temperatures than other cities in Australia. It had the lowest median temperature of all the cities we had measurements for, of around 22 degrees Celcius, but had the highest variation in temperature. The highest values were over 40 degrees Celcius.

There was no rain on at least half the January 17ths we had data for in Melbourne (the median was 0mm) and only three days had more than 10mm of rain. The average rainfall was only 1.5mm, this is higher than the median due to the right skew of the rainfall data.

[Question 4] The `Galton` data set in the `mosaic` library contains data from Francis Galton in the 1880s.

```
glimpse(Galton)
```

```
## Observations: 898
## Variables: 6
## $ family <fct> 1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 4, 4, 4, 4, 4, 5, 5, 5, 5, 5, ...
## $ father <dbl> 78.5, 78.5, 78.5, 78.5, 75.5, 75.5, 75.5, 75.5, 75.0, 75.0, ...
## $ mother <dbl> 67.0, 67.0, 67.0, 67.0, 66.5, 66.5, 66.5, 66.5, 64.0, 64.0, ...
## $ sex    <fct> M, F, F, F, M, M, F, F, M, F, M, M, F, F, F, M, M, M, F, F, ...
## $ height <dbl> 73.2, 69.2, 69.0, 69.0, 73.5, 72.5, 65.5, 65.5, 71.0, 68.0, ...
## $ nkids  <int> 4, 4, 4, 4, 4, 4, 4, 4, 2, 2, 5, 5, 5, 5, 5, 6, 6, 6, 6, 6, ...
```

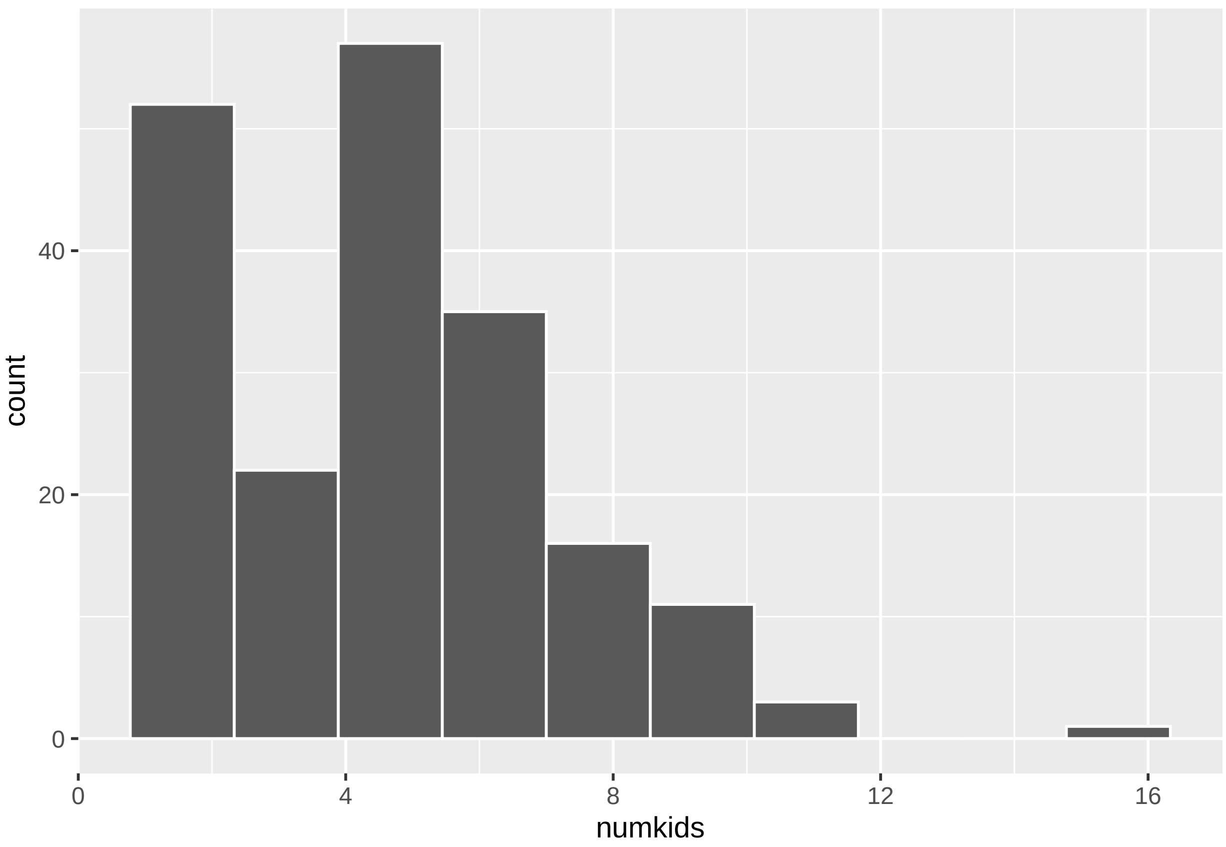
(a) We want to know how many children were there in each family in the `Galton` data set. To answer this, create a data frame called `data` that contains the family id number and the numbers of kids in each family. Note: the number of children is repeated for every member of the family. The data frame you create should not include the repeats.

Hint 1: the `summarise()` function is useful here
Hint 2: the mean of 4, 4, 4 and 4 is.....4

```
data <- summarise(group_by(Galton, family),
  numkids = mean(nkids))

data <- data.frame(data)
```

(b) Graph the distribution of the number of kids in the `Galton` data set families, using the dataframe you created in part (a). Describe (in words) the features of this distribution.

```
ggplot(data, aes(x = numkids)) +
         geom_histogram(bins=10,color="white")
```



The number of kids in the `Galton` data set families follow a positively skewed (or equivalently, right skewed) distribution. There were many more “smaller” families than “larger”" families. The number of kids per family ranged from 1 to around 15. The distribution appears to be bimodal, with quite a few families having 1-2 children, and 4-5 children. The family with around 15 kids appears to be an oulier because it is so much higher than the number of children in the other families. [Remember, when describing features of the distribution of a quantitative variable, you should comment on centre, shape and spread of the distribution.]

(c) Just based on the graph you generated in part (b), how do you think the mean and median would compare? Justify your reasoning.

I would expect the mean to be higher than the median because the distribution of the number of kids is right-skewed and there is a high outlier. The mean would be dragged away from the median in the direction of the long tail and outlier(s).

(d) Compute the mean and median of the number of kids in the `Galton` data set families. Does this match what you expected to see in part (e)?

```
summarise(data, mean = mean(numkids), median = median(numkids))
```

```
##       mean median
## 1 4.563452      4
```

Yes. The mean is higher than the median, as expected.