STAT 385 - Homework 6

Jia Lin Mei - jmei43

Due 11:59 PM, 4/01/2020

Contents

Dataset: diamonds from the ggplot2 package 1
Dataset: 2015 Flight Delays and Cancellations Data 10

Dataset: diamonds from the ggplot2 package

The ggplot2 package include many interesting datasets. We looked at `mpg` in lecture this week. Now, we will explore the `diamonds` dataset. This data contains the prices and other attributes of almost 54,000 diamonds.

```
library(tidyverse)
str(diamonds)
```

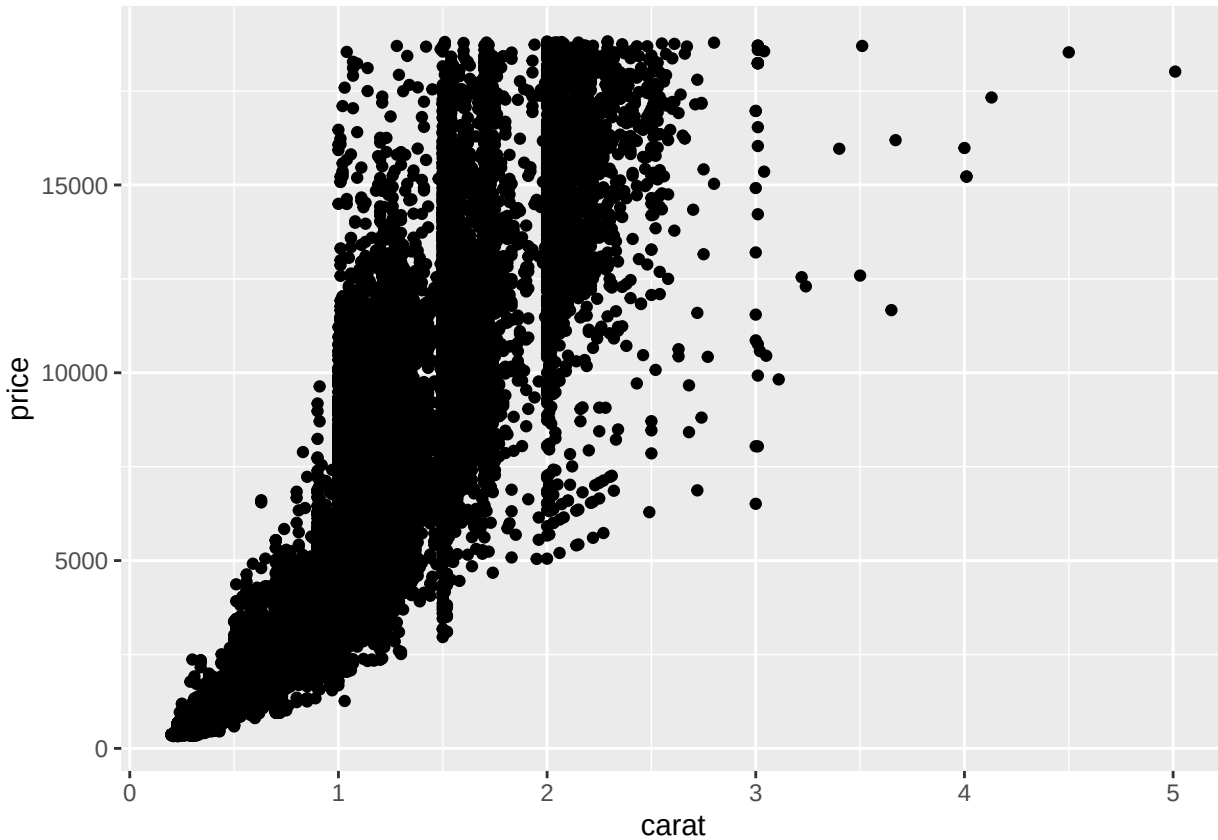
```
## tibble [53,940 x 10] (S3: tbl_df/tbl/data.frame)
##  $ carat  : num [1:53940] 0.23 0.21 0.23 0.29 0.31 0.24 0.24 0.26 0.22 0.23 ...
##  $ cut    : Ord.factor w/ 5 levels "Fair"<"Good"<..: 5 4 2 4 2 3 3 3 1 3 ...
##  $ color  : Ord.factor w/ 7 levels "D"<"E"<"F"<"G"<..: 2 2 2 6 7 7 6 5 2 5 ...
##  $ clarity: Ord.factor w/ 8 levels "I1"<"SI2"<"SI1"<..: 2 3 5 4 2 6 7 3 4 5 ...
##  $ depth  : num [1:53940] 61.5 59.8 56.9 62.4 63.3 62.8 62.3 61.9 65.1 59.4 ...
##  $ table  : num [1:53940] 55 61 65 58 58 57 57 55 61 61 ...
##  $ price  : int [1:53940] 326 326 327 334 335 336 336 337 337 338 ...
##  $ x      : num [1:53940] 3.95 3.89 4.05 4.2 4.34 3.94 3.95 4.07 3.87 4 ...
##  $ y      : num [1:53940] 3.98 3.84 4.07 4.23 4.35 3.96 3.98 4.11 3.78 4.05 ...
##  $ z      : num [1:53940] 2.43 2.31 2.31 2.63 2.75 2.48 2.47 2.53 2.49 2.39 ...
```

- Run `?diamonds` to know what each variable presents.

Scatterplot

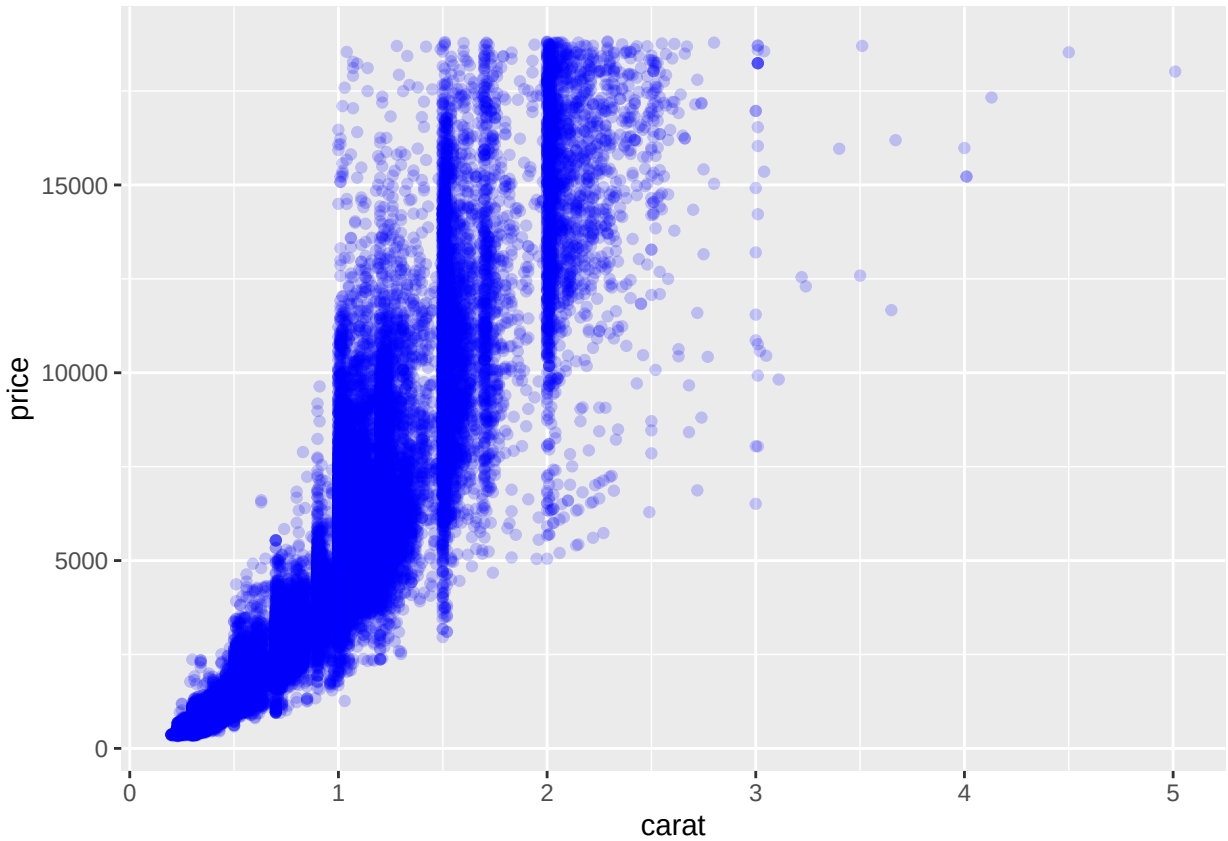
- There are 4 parts to this exercise: each part worths 5 points.
- **(a)** Graph a scatterplot of `price` vs `carat`.
- `price` as the y-axis and `carat` as the x-axis.

```
ggplot(data=diamonds) +
  geom_point(mapping = aes(x = carat, y = price))
```



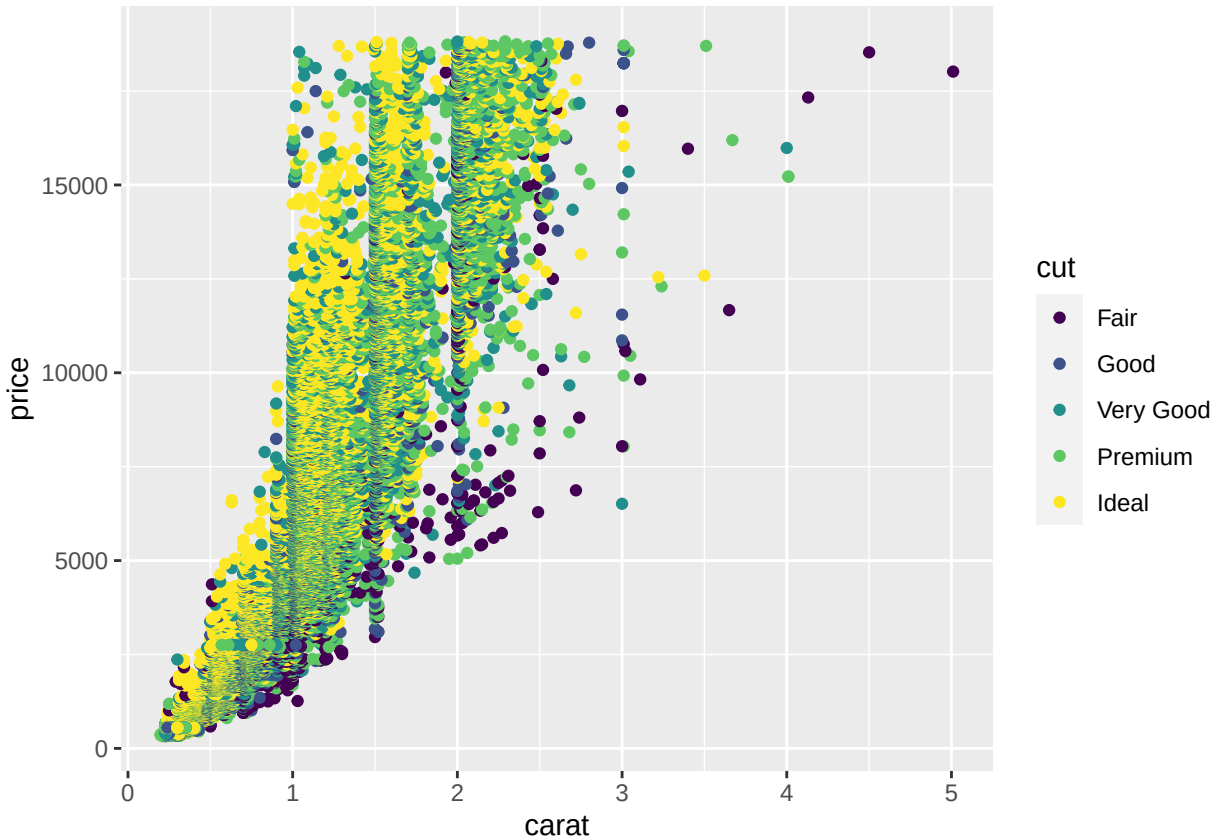
- **(b) With the above graph**, adjust the transparency level of ALL the points so we can see clearly where the clusters are:

```
ggplot(data=diamonds) +
  geom_point(mapping = aes(x = carat, y = price), alpha = 0.2, color = "blue")
```



- **(c) Now create a new plot**, graph a scatterplot of `price` vs `carat`, and color the points BY the quality of the cut (`cut`)!
- Don't change the transparency level of the points.

```
ggplot(data=diamonds) +
  geom_point(mapping = aes(x = carat, y = price, color = cut))
```

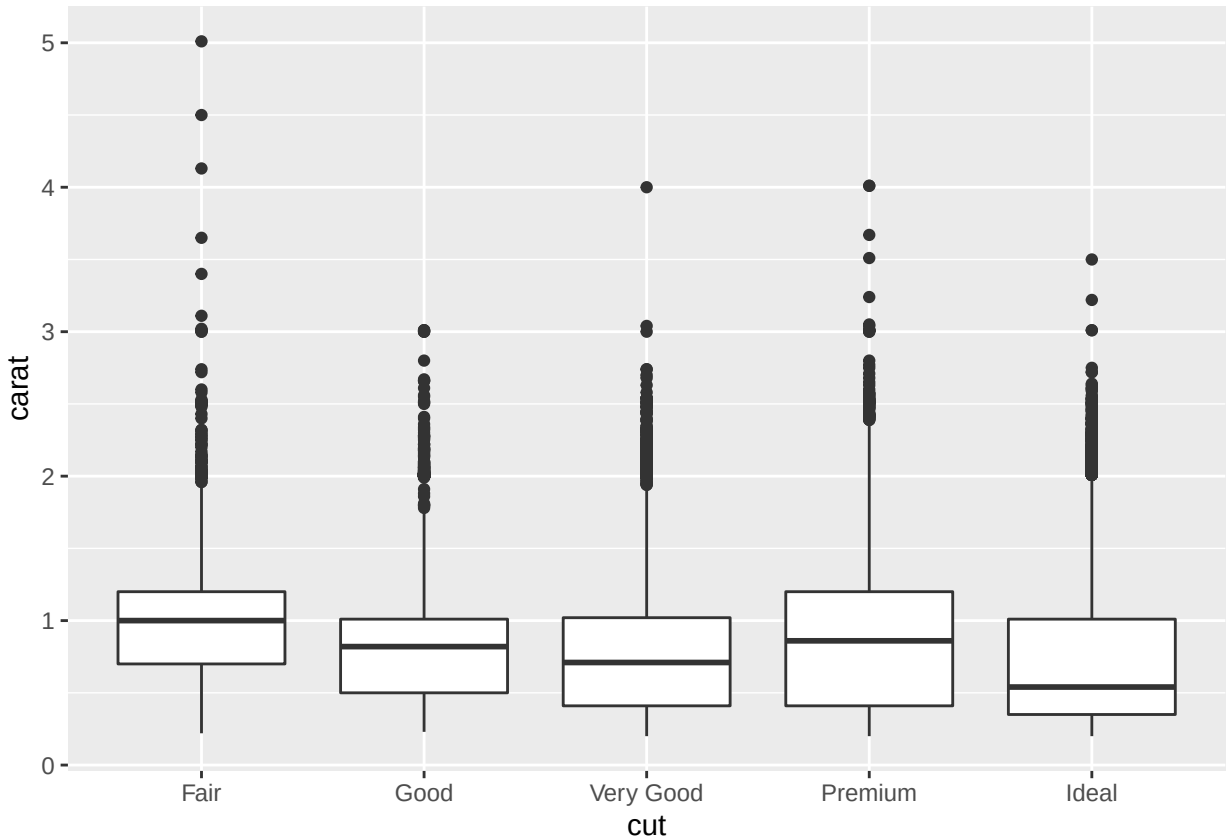


- **(d) What can you conclude based on the above graph?**
- **ANSWER: Cut isn't a good indicator of price. Diamonds of all different qualites of cuts exist at different price ranges, and you can get a diamond with a good cut for a low price. Instead, carats seems to be a better indicator of price - the more carats a diamond weighs, the more expensive it is.**

Boxplot

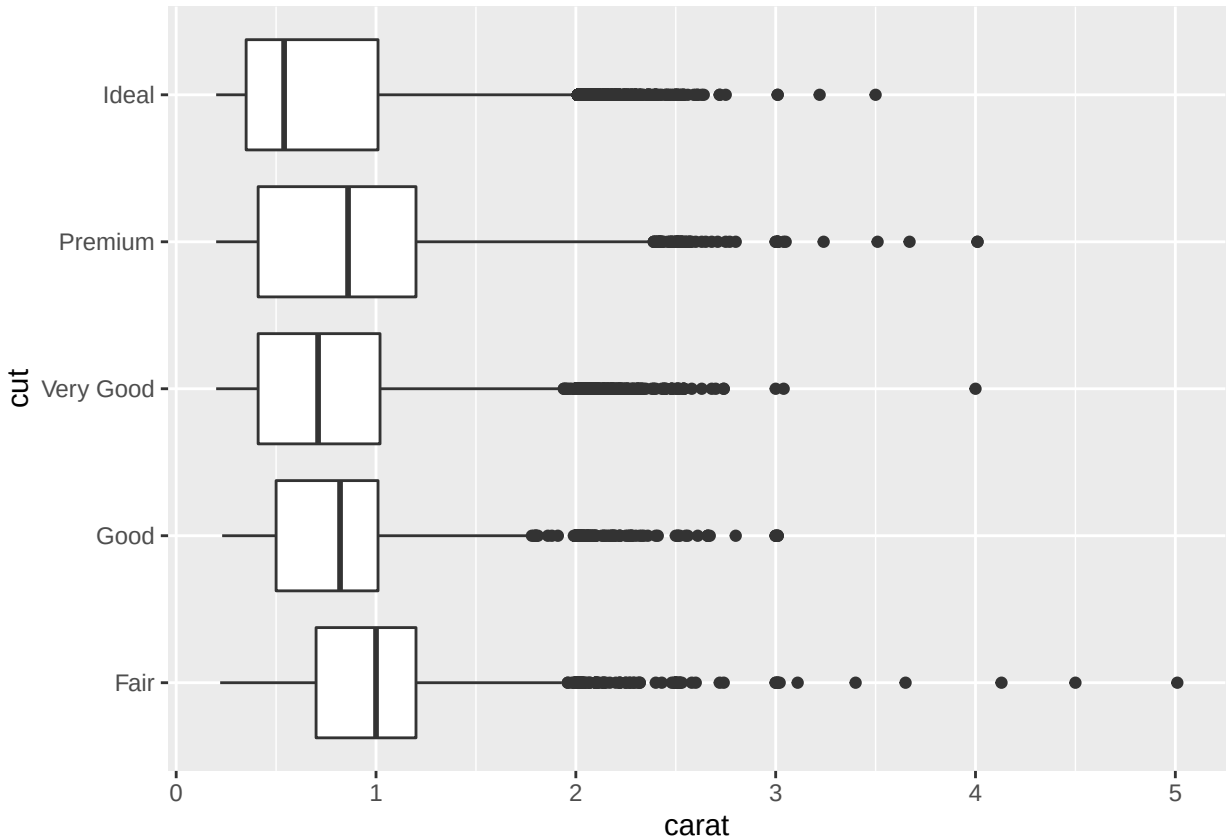
- There are 3 parts to this exercise: each part worths 5 points.
- **(a)** Graph a boxplot of `carat` vs `cut`.
- `carat` as the y-axis and `cut` as the x-axis.

```
ggplot(data = diamonds, mapping = aes(x=cut, y=carat)) +
  geom_boxplot()
```



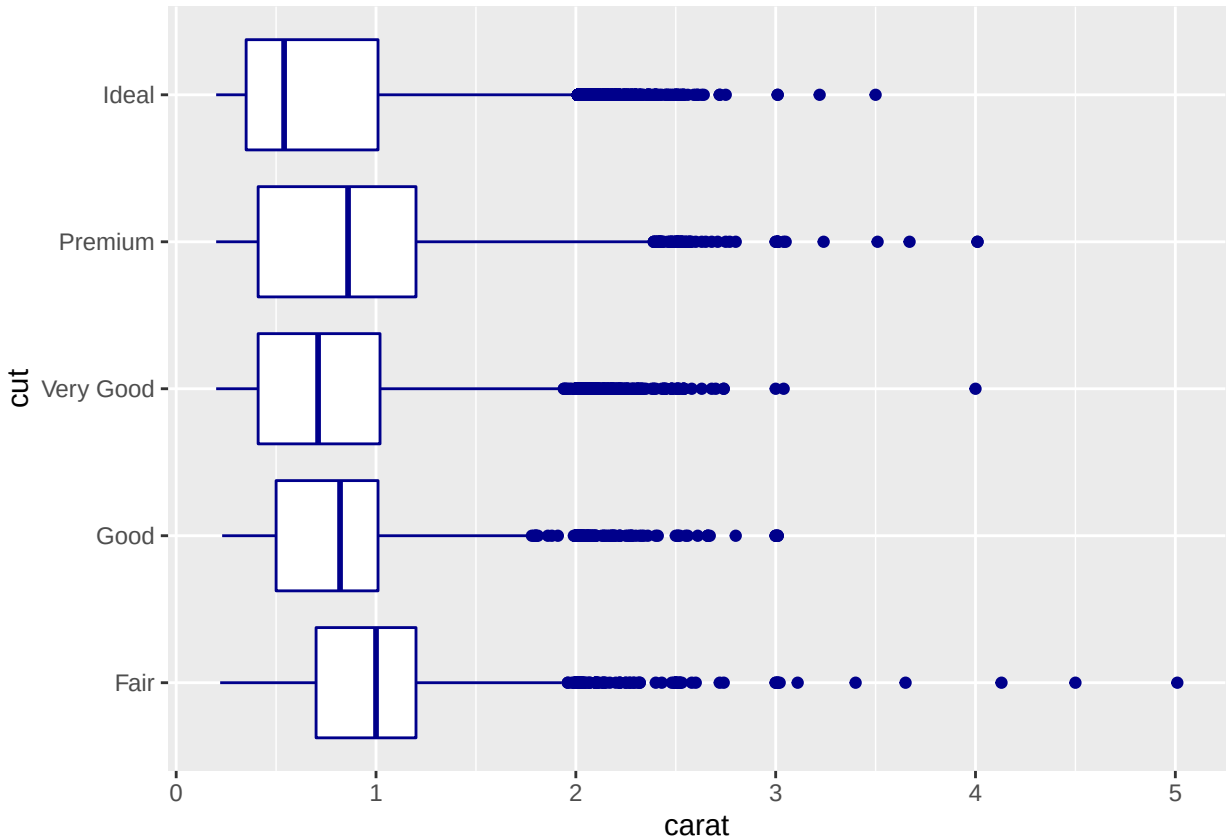
- **(b) With the plot above**, rotate it so `cut` is on the y-axis and `carat` is on the x-axis.
- Make sure the plot is still a boxplot.

```
p <- ggplot(data = diamonds, mapping = aes(x=cut, y=carat)) +
  geom_boxplot()

p + coord_flip()
```



- **(c) With the plot from (b)**, color the boxplot dark blue.

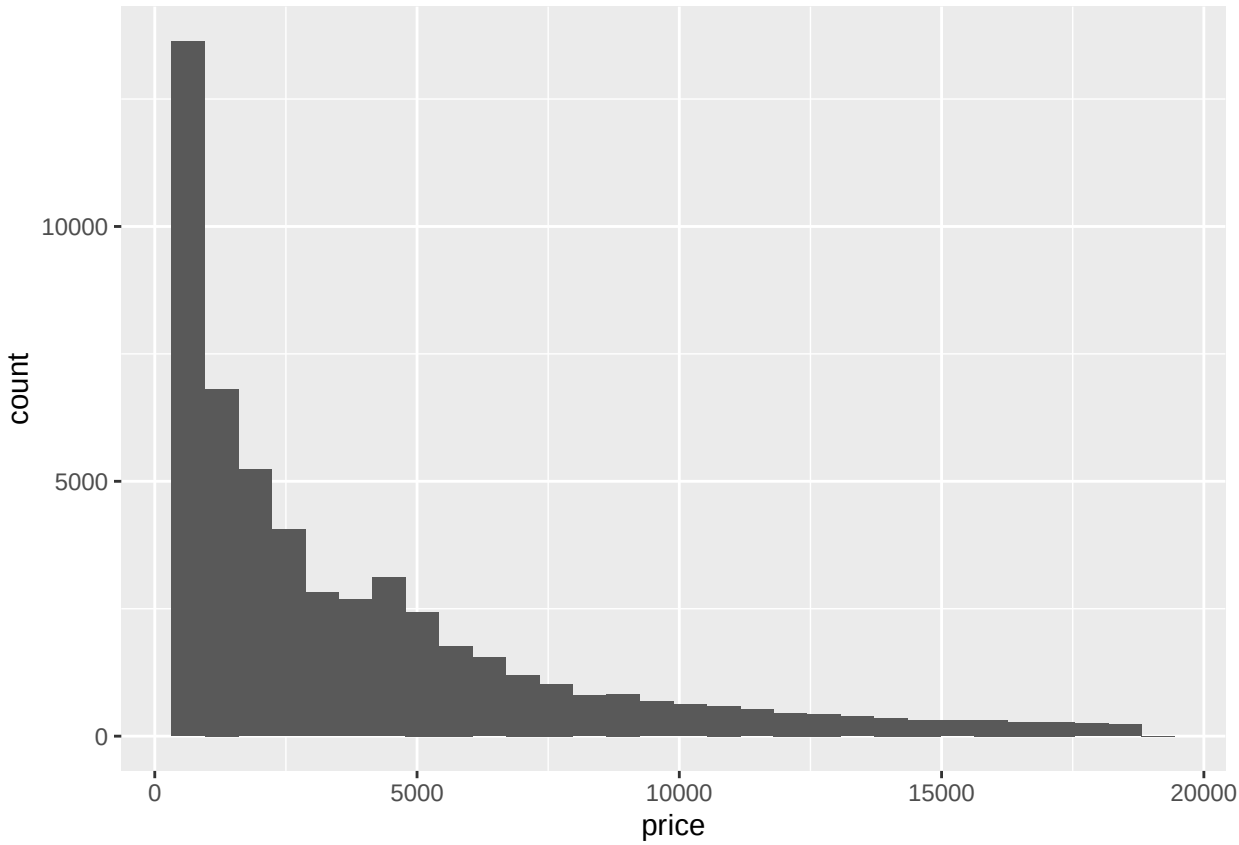
```
p <- ggplot(data = diamonds, mapping = aes(x=cut, y=carat)) +
  geom_boxplot(color = "darkblue")

p + coord_flip()
```



Histogram

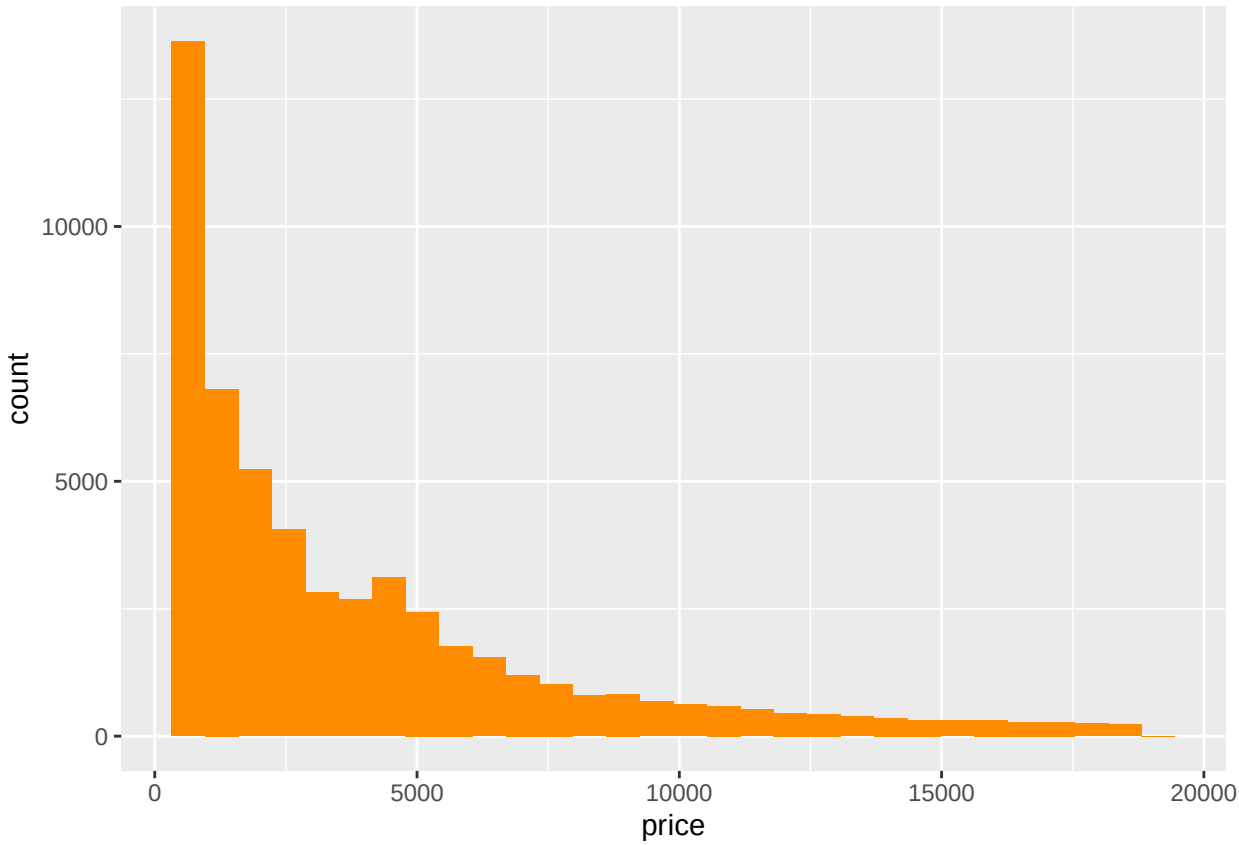
- There are 3 parts to this exercise: each part worths 5 points.
- **(a)** Graph a histogram of `price`.

```
ggplot(data = diamonds, mapping = aes(x=price)) + geom_histogram()
```



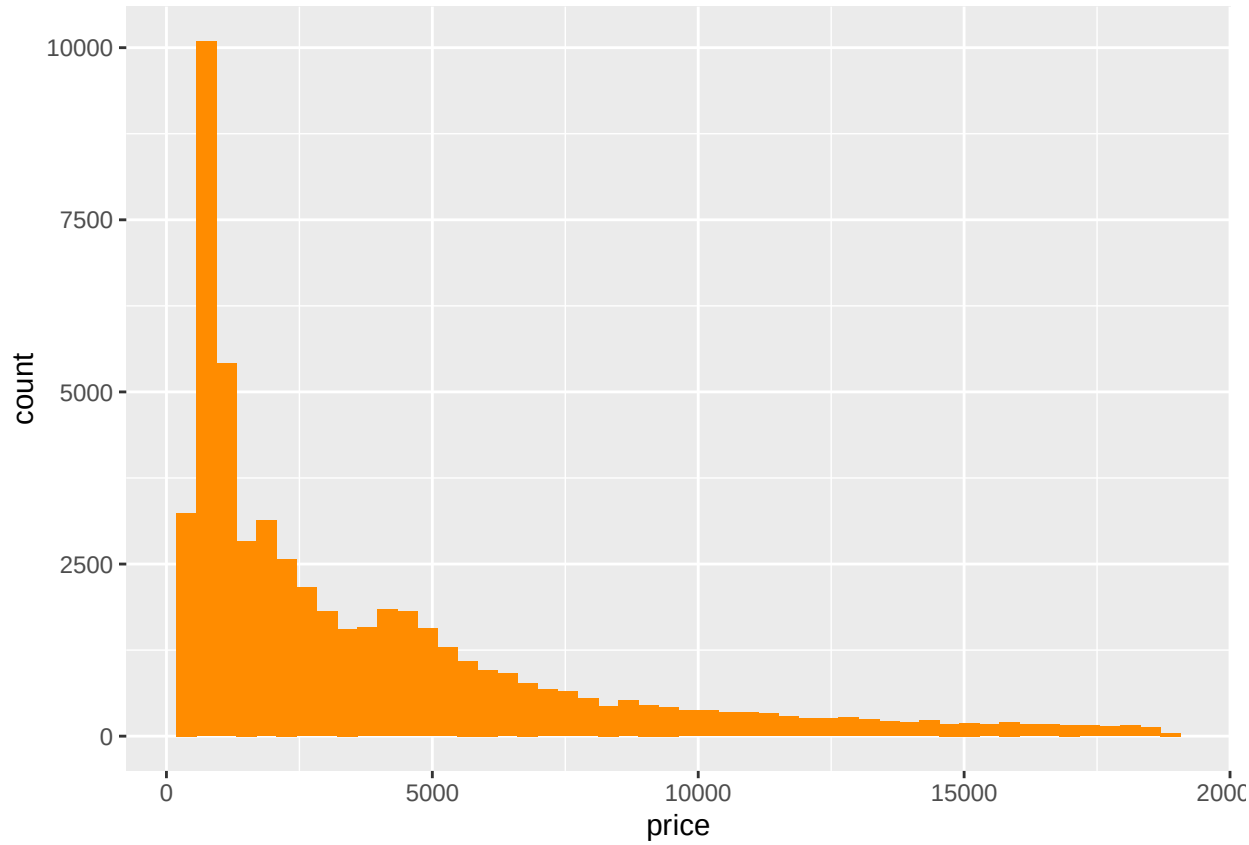
- **(b) With the plot above**, color the bars dark orange.

```
ggplot(data = diamonds, mapping = aes(x=price)) + geom_histogram(fill="darkorange")
```



- **(c) With the plot from (b)**, change the number of bins to 50.

```
ggplot(data = diamonds, mapping = aes(x=price)) +
  geom_histogram(bins = 50, fill="darkorange")
```



Dataset: 2015 Flight Delays and Cancellations Data

In this exercise, we continue to work with the 2015 Flight Delays and Cancellations Data provided by the U.S. Department of Transportation. This is a huge dataset availalbe on Kaggle. But for us, we will only take a look at flights **flying out** from O'Hare International Airport (ORD) in January, 2015.

Load data

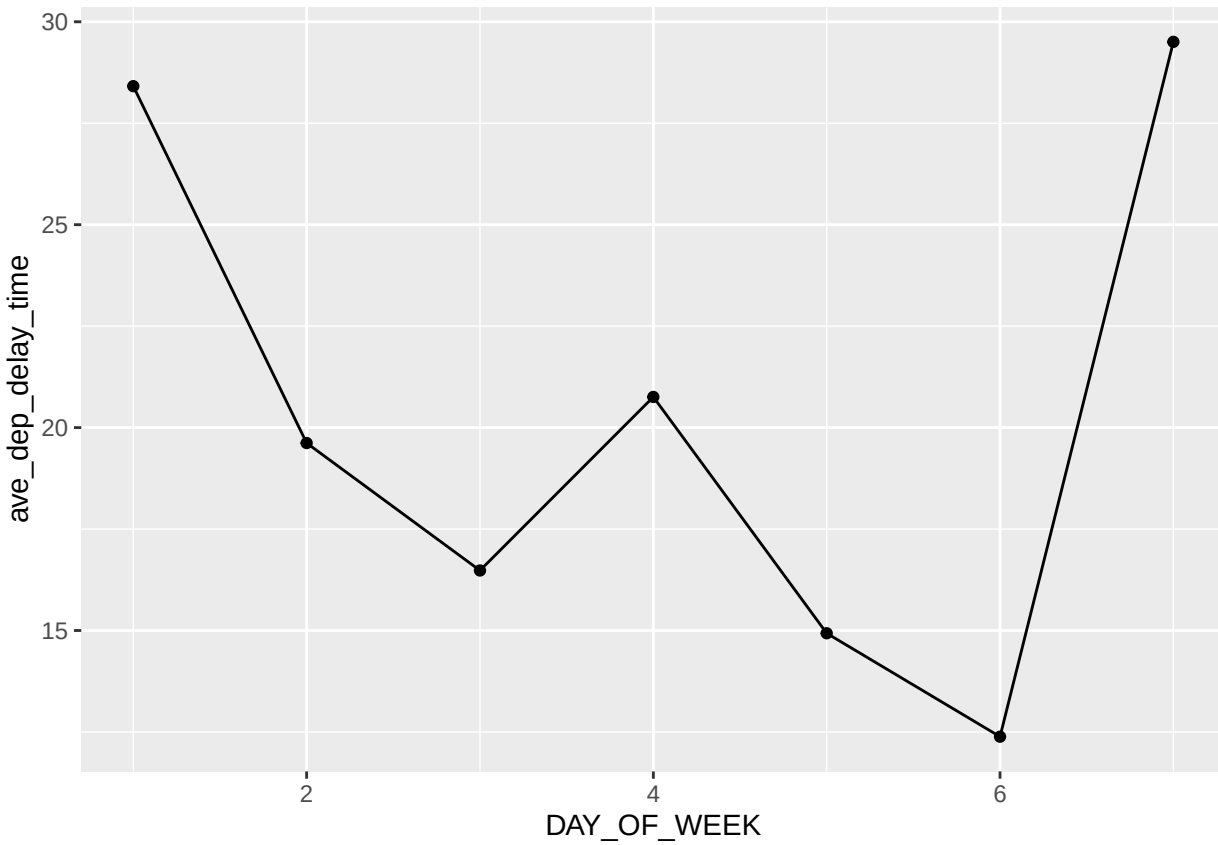
- I have filtered out the data specific to O'Hare and stored it in `ohare_jan.csv`. This filtered data is available at the URL: https://nkha149.github.io/stat385-sp2020/files/data/ohare_jan.csv.

```
flights <- read_csv(file = "https://nkha149.github.io/stat385-sp2020/files/data/ohare_jan.csv")
```

Using `dplyr` and `ggplot2` together

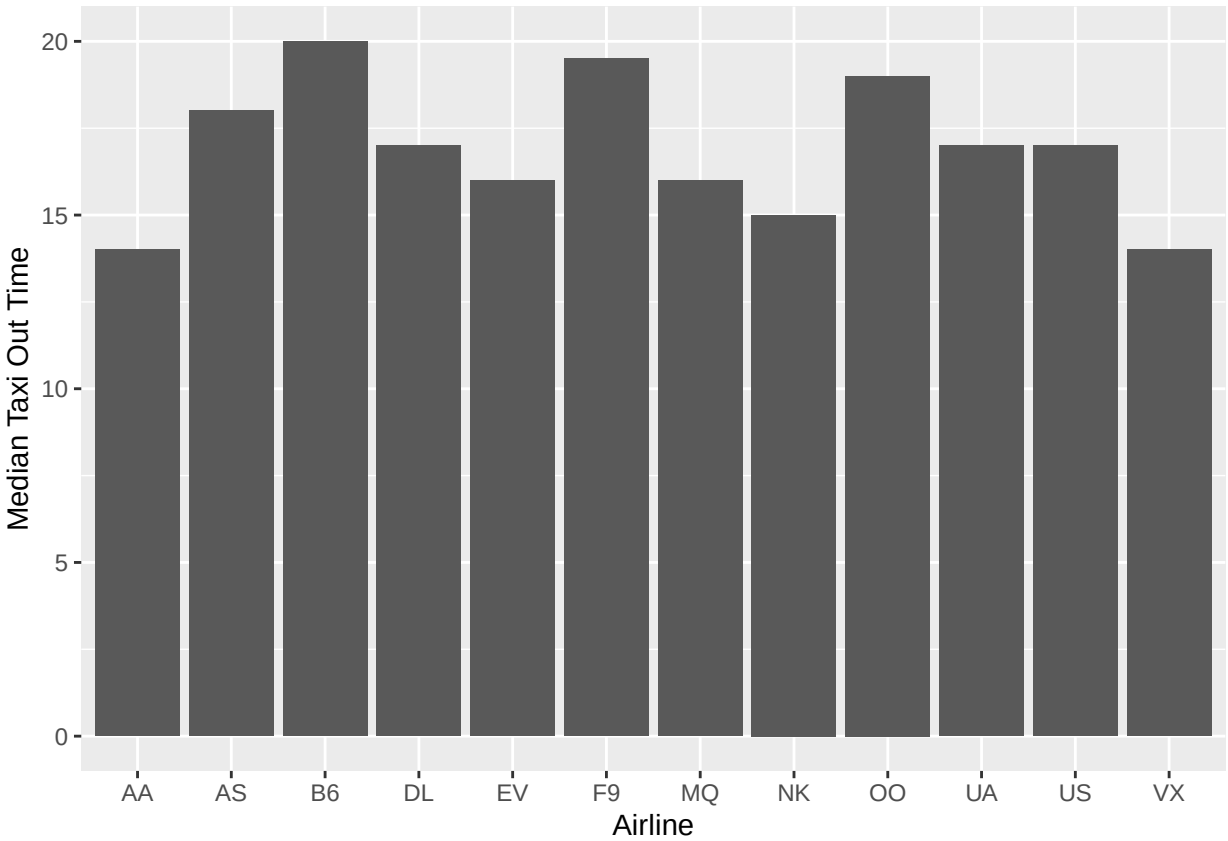
- For this exercise, you will need to:
- Graph a plot that can be used to answer the question.
- Determine the type of plot that is needed.
- (You might need to) Manipulate the data to get the specific information that you want.
- There are 2 parts to this exercise: each part worths 10 points.
- **(a)** How does the **average** departure delay time differ among days of the week?

```r
dep_delay <- flights %>%
  group_by(DAY_OF_WEEK) %>%
  summarize(ave_dep_delay_time = mean(DEPARTURE_DELAY, na.rm = TRUE))

ggplot(data = dep_delay, aes(x = DAY_OF_WEEK, y = ave_dep_delay_time)) +
  geom_point() + geom_line()
```



- **(b)** Airlines are grouped into different terminals. There is a belief that certain terminals have better taxing out time. How does **median** taxi out time differ among diferrent airlines?

```r
taxi_time <- flights %>%
  group_by(AIRLINE) %>%
  summarise(taxi_out = median(TAXI_OUT, na.rm = TRUE))

ggplot(data = taxi_time, mapping = aes(x = AIRLINE, y = taxi_out)) +
geom_bar(stat = "identity") +
labs(x = "Airline", y = "Median Taxi Out Time")
```


Median Taxi Out Time
20
15
10
5
0
AA
AS
B6
DL
EV
F9
MQ
NK
OO
UA
US
VX
Airline