Math 300 Lesson 4 Notes

Boxplots and Barcharts

YOUR NAME HERE

June, 2022

Contents

Objectives

1. Create and interpret boxplots using `ggplot()`; identify and interpret quartiles, interquartile range, whiskers, and outliers.
2. Compare and contrast boxplots, barcharts, histograms, and pie charts.
3. Create and interpret barcharts for one or two categorical variables using `ggplot()`.
4. Create tables to summarize one or two categorical variables.

Reading

Chapter 2.7 - 2.9

Lesson

Work through the learning checks LC2.22 - LC2.37. Complete code when required.

- Plots change based on the nature of our data. The boxplot was developed prior to widespread use of computers. It is comparable to the use of a histogram and/or density plot. It is meant for a single quantitative variable. It a second categorical variable is added, we can generate side-by-side boxplots.
- The barchart is for a categorical, qualitative variables. It can be argued that a table is just as informative. If a second categorical variable is added, we must consider how to visually represent it. A table is often better than a barchart especially for a single variable. We will `tidyverse` and base code to create a table in our solution.
- Note the difference in the use of `fill` and `color`. The former is the fill color and the latter the color of the bounding box.

- See the following for more on the principles of visualizing data and technical details.
- Reading 2.9.2 will help you understand `R` code. Remember, what do we want `R` to do? What does `R` need to do this?

Setup

```
library(nycflights13)
library(ggplot2)
library(dplyr)
```

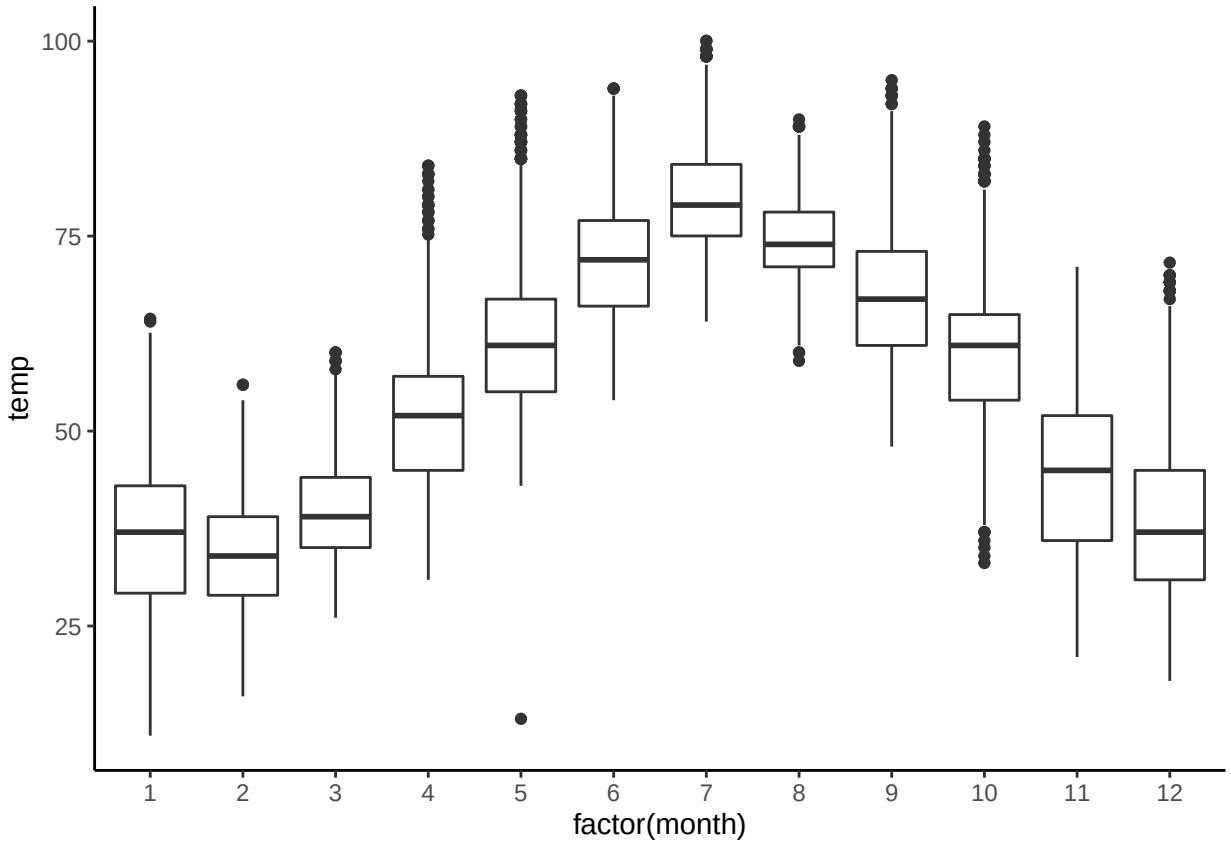
Create the side-by-side boxplots from the book.

Why do we have to use factor in `aes()` function?

Be able to explain the elements of the boxplot.

```
# Boxplots in the reading of section 2.7, adding a different theme.

ggplot(data = weather, mapping = aes(x = factor(month), y = temp)) +
  geom_boxplot() +
  theme_classic()
```



LC 2.22 (Objective 1)

(LC 2.22) What does the dot at the bottom of the plot for May correspond to? Explain what might have occurred in May to produce this point.

```
#Explore the outlier. Complete the code and remove comment labels
#weather %>%
#  filter(month == _____________ & temp < ___________)
```

Solution:

LC 2.23 (Objective 1)

(LC 2.23) Which months have the highest variability in temperature? What reasons do you think this is?

Solution:

Here's how we compute the exact IQR values for each month (we'll see this more in depth when we learn more about data wrangling later in the course):

- `group` the observations by `month` then
- for each `group`, i.e. `month`, `summarize` it by applying the summary statistic function `IQR()`, while making sure to skip over missing data via `na.rm=TRUE` then
- `arrange` the table in `desc`ending order of `IQR`

```
# Which month has the most variability in temperature?
weather %>%
  group_by(month) %>%
  summarize(IQR = IQR(temp, na.rm = TRUE)) %>%
  arrange(desc(IQR))
```

```
## # A tibble: 12 x 2
##    month   IQR
##    <int> <dbl>
##  1    11 16.0
##  2    12 14.0
##  3     1 13.8
##  4     9 12.1
##  5     4 12.1
##  6     5 11.9
##  7     6 11.0
##  8    10 11.0
##  9     2 10.1
## 10     7  9.18
## 11     3  9
## 12     8  7.02
```

LC 2.24 (Objective 1)

(LC 2.24) We looked at the distribution of the numerical variable `temp` split by the numerical variable `month` that we converted to a categorical variable using the `factor()` function. Why would a boxplot of `temp` split by the numerical variable `pressure` similarly converted to a categorical variable using the `factor()` not be informative?

Solution:

```
# Complete the code and remove comments symbols
# ggplot(data = weather, mapping = aes(x = factor(_____________), y = _____)) +
#   geom_boxplot()
```

LC 2.25 (Objective 2)

(LC2.25) Boxplots provide a simple way to identify outliers. Why may outliers be easier to identify when looking at a boxplot instead of a faceted histogram?

Solution:

LC 2.26 (Objective 2)

(LC 2.26) Why are histograms inappropriate for visualizing categorical variables?

Solution:

LC 2.27 (Objective 2)

(LC 2.27) What is the difference between histograms and barplots?

Solution:

LC 2.28 (Objective 3, 4)

(LC 2.28) How many Envoy Air flights departed NYC in 2013?

Solution:

Plots that will help you if you complete the code. Remember to remove # from the code chunks.

```
# Complete the code and remove comments symbols
#ggplot(data = flights, mapping = aes(x = _________________)) +
#  geom_bar()
```

```
# Complete the code and remove comments symbols
#flights %>%
#  count(_____________) %>%
#  arrange(desc(n))
```

```
#Base code
table(flights$carrier)
```

```
##
##    9E    AA    AS    B6    DL    EV    F9    FL    HA    MQ    OO    UA    US
## 18460 32729   714 54635 48110 54173   685  3260   342 26397    32 58665 20536
##    VX    WN    YV
##  5162 12275   601
```

LC 2.29 (Objective 3, 4)

(LC 2.29) What was the seventh highest airline in terms of departed flights from NYC in 2013? How could we better present the table to get this answer quickly?

Solution:

```
# Complete the code and remove comments symbols
# flights %>%
#   count(_______) %>%
#   arrange(desc(____))
```

LC 2.30 (Objective 2)

(LC 2.30) Why should pie charts be avoided and replaced by barplots?

Solution:

LC 2.31 (Objective 2)

(LC 2.31) What is your opinion as to why pie charts continue to be used?

Solution:

LC 2.32 (Objective 3)

Complete code needed for this question.

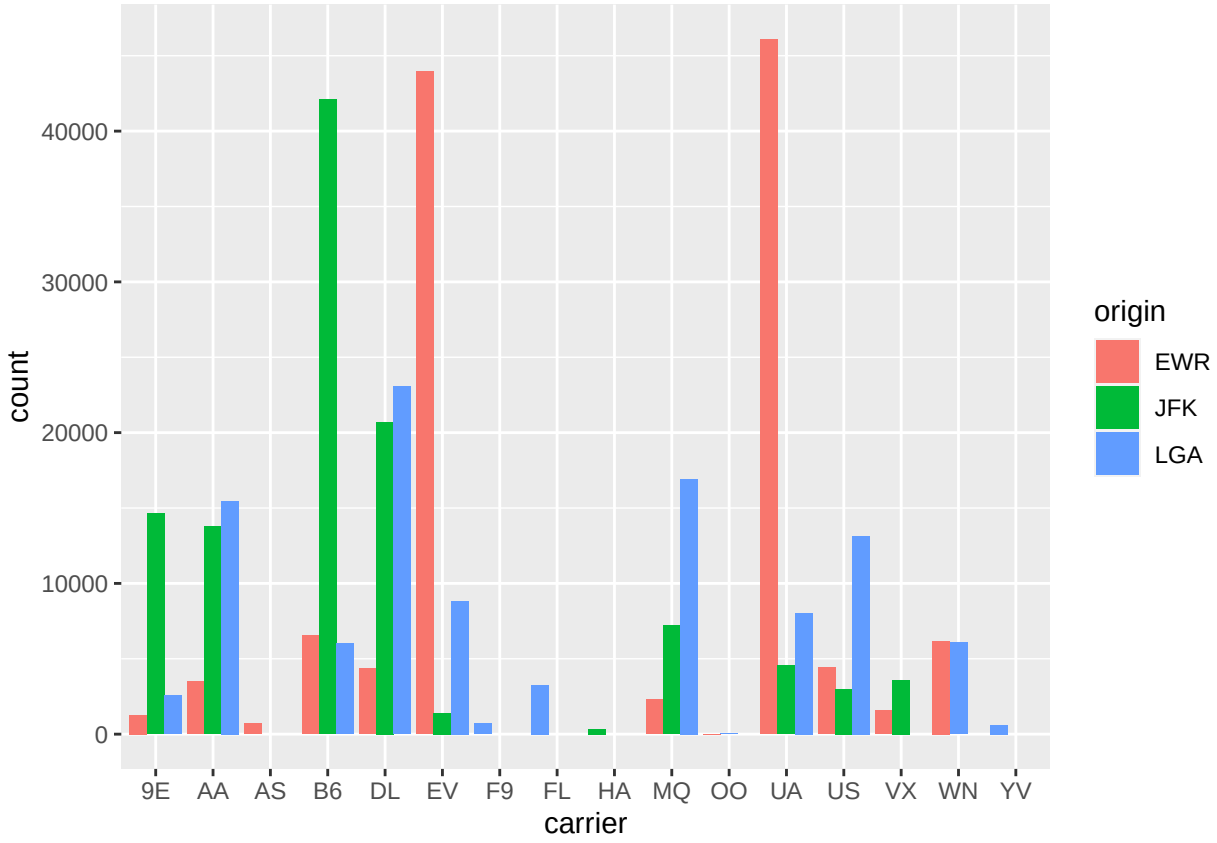
```
# Code needed for this problem
#ggplot(data = flights, mapping = aes(x = _____________, fill = _________)) +
#  geom_bar()
```

(LC 2.32)‘ What kinds of questions are not easily answered by looking at the above figure?

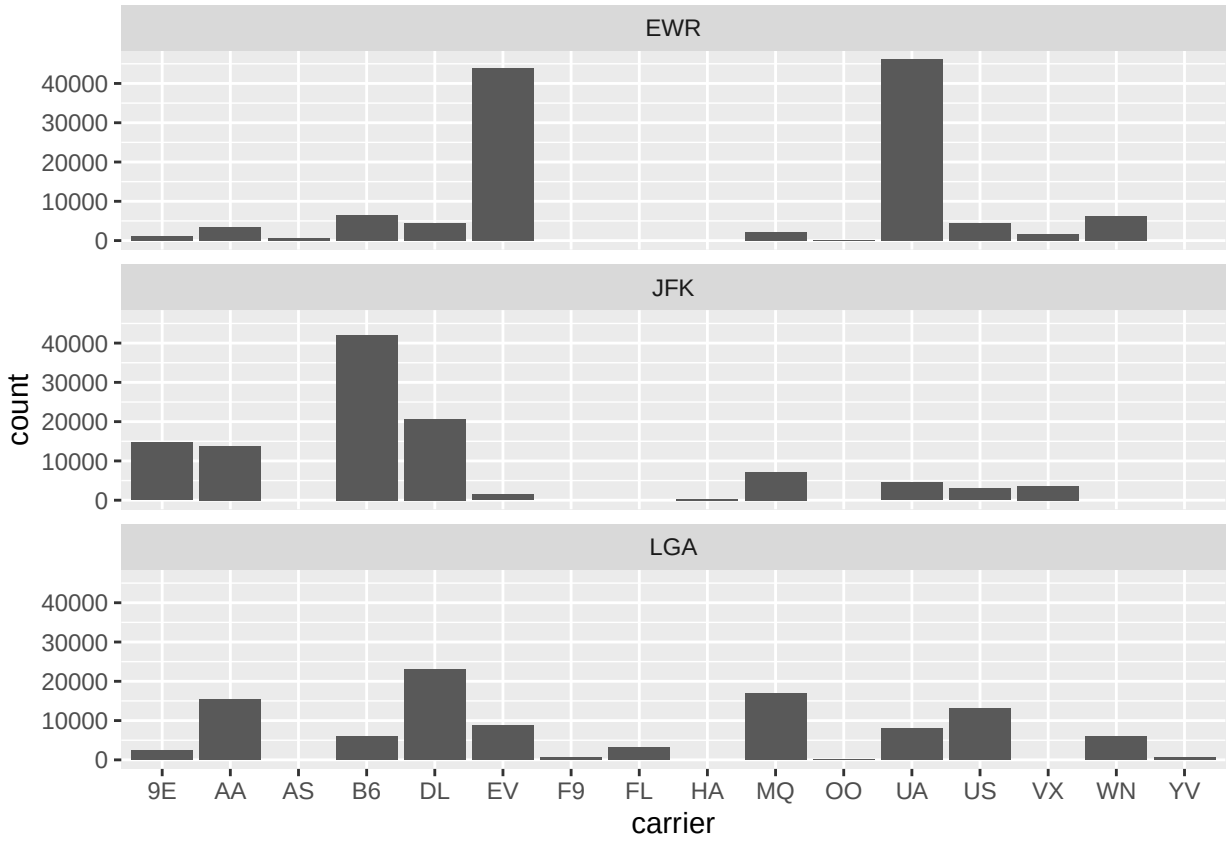
Solution:

Here are some other plots that may help.

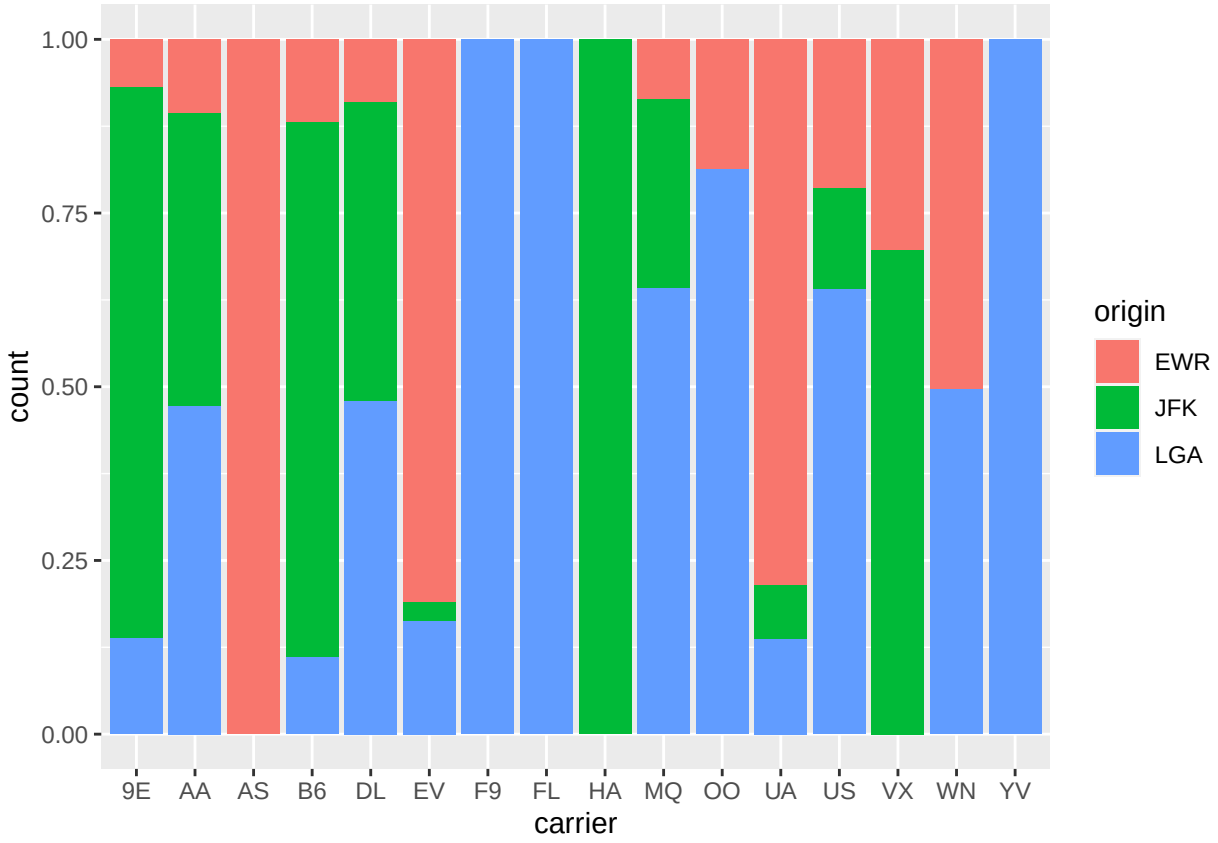
```
ggplot(data = flights, mapping = aes(x = carrier, fill = origin)) +
  geom_bar(position = position_dodge(preserve = "single"))
```




```
ggplot(data = flights, mapping = aes(x = carrier)) +
  geom_bar() +
  facet_wrap(~ origin, ncol = 1)
```




```
# Percentage
ggplot(data = flights, mapping = aes(x = carrier, fill = origin)) +
  geom_bar(position='fill')
```



LC 2.33 (Objective 3)

(LC 2.33) What can you say, if anything, about the relationship between airline and airport in NYC in 2013 in regards to the number of departing flights?

Solution:

LC 2.34 (Objective 2)

(LC 2.34) Why might the side-by-side (AKA dodged) barplot be preferable to a stacked barplot in this case?

Solution:

LC 2.35 (Objective 2)

(LC 2.35) What are the disadvantages of using a side-by-side (AKA dodged) barplot, in general?

Solution:

LC 2.36 (Objective 2)

(LC 2.38) Why is the faceted barplot preferred to the side-by-side and stacked barplots in this case?

Solution:

LC 2.37 (Objective 2)

(LC 2.37) What information about the different carriers at different airports is more easily seen in the faceted barplot?

Solution:

Documenting software

- File creation date: 2022-06-15
- R version 4.1.3 (2022-03-10)
- `ggplot2` package version: 3.3.6
- `dplyr` package version: 1.0.9
- `nycflights13` package version: 1.0.2