# Lab 2 - Introduction to Data

Shamecca Marshall

2023-09-05

## Load packages

## The data

```
data(nycflights)
names(nycflights)
```

```
##  [1] "year"      "month"     "day"       "dep_time"  "dep_delay" "arr_time"
##  [7] "arr_delay" "carrier"   "tailnum"   "flight"    "origin"    "dest"
## [13] "air_time"  "distance"  "hour"      "minute"
```

```
?nycflights
```

## Taking a glimpse at the data

```
glimpse(nycflights)
```

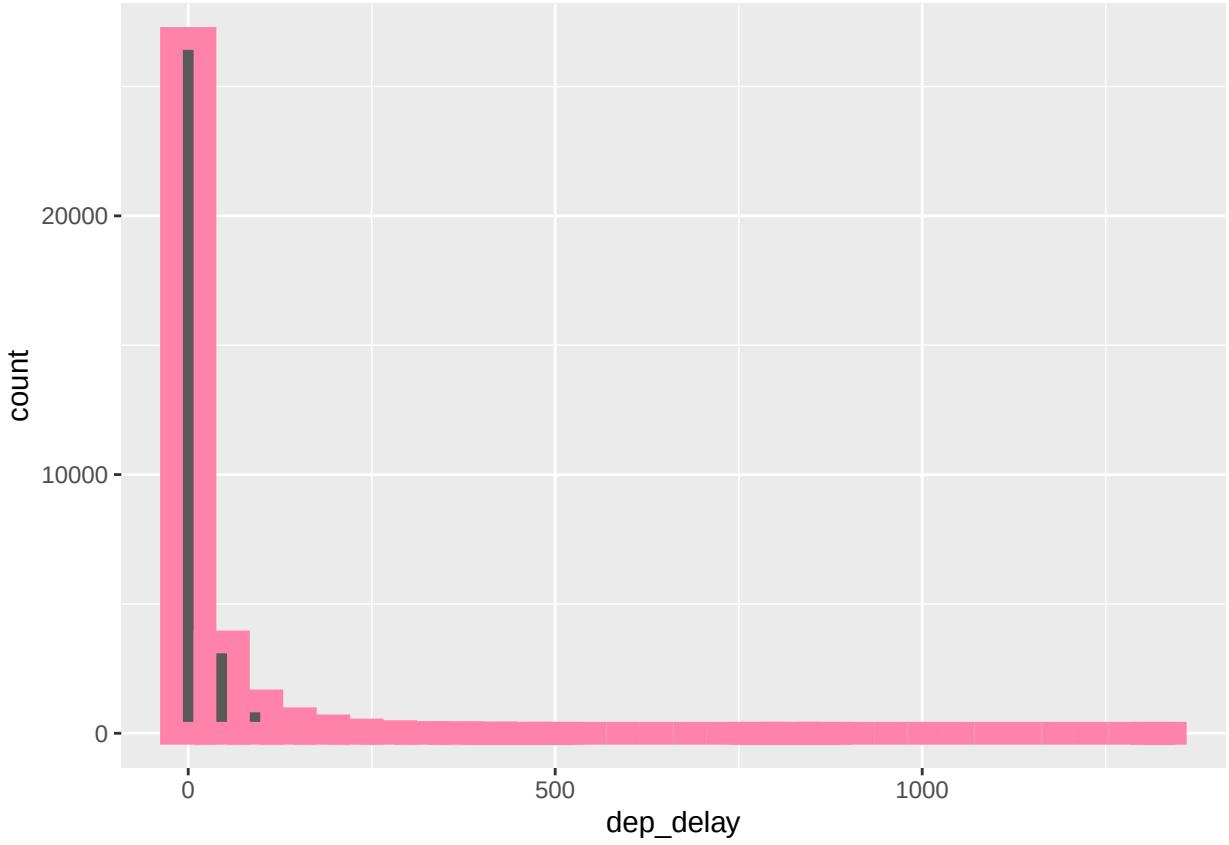
```
## Rows: 32,735
## Columns: 16
## $ year      <int> 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, ~
## $ month     <int> 6, 5, 12, 5, 7, 1, 12, 8, 9, 4, 6, 11, 4, 3, 10, 1, 2, 8, 10~
## $ day       <int> 30, 7, 8, 14, 21, 1, 9, 13, 26, 30, 17, 22, 26, 25, 21, 23, ~
## $ dep_time  <int> 940, 1657, 859, 1841, 1102, 1817, 1259, 1920, 725, 1323, 940~
## $ dep_delay <dbl> 15, -3, -1, -4, -3, -3, 14, 85, -10, 62, 5, 5, -2, 115, -4, ~
## $ arr_time  <int> 1216, 2104, 1238, 2122, 1230, 2008, 1617, 2032, 1027, 1549, ~
## $ arr_delay <dbl> -4, 10, 11, -34, -8, 3, 22, 71, -8, 60, -4, -2, 22, 91, -6, ~
## $ carrier   <chr> "VX", "DL", "DL", "DL", "9E", "AA", "WN", "B6", "AA", "EV", ~
## $ tailnum   <chr> "N626VA", "N3760C", "N712TW", "N914DL", "N823AY", "N3AXAA", ~
## $ flight    <int> 407, 329, 422, 2391, 3652, 353, 1428, 1407, 2279, 4162, 20, ~
## $ origin    <chr> "JFK", "JFK", "JFK", "JFK", "LGA", "LGA", "EWR", "JFK", "LGA~
## $ dest      <chr> "LAX", "SJU", "LAX", "TPA", "ORF", "ORD", "HOU", "IAD", "MIA~
## $ air_time  <dbl> 313, 216, 376, 135, 50, 138, 240, 48, 148, 110, 50, 161, 87,~
## $ distance  <dbl> 2475, 1598, 2475, 1005, 296, 733, 1411, 228, 1096, 820, 264,~
## $ hour      <dbl> 9, 16, 8, 18, 11, 18, 12, 19, 7, 13, 9, 13, 8, 20, 12, 20, 6~
## $ minute    <dbl> 40, 57, 59, 41, 2, 17, 59, 20, 25, 23, 40, 20, 9, 54, 17, 24~
```

## Analyzing depature delays with a histogram

```
ggplot(data = nycflights, aes(x = dep_delay))+
  geom_histogram(colour = "palevioletred1", size = 4)
```

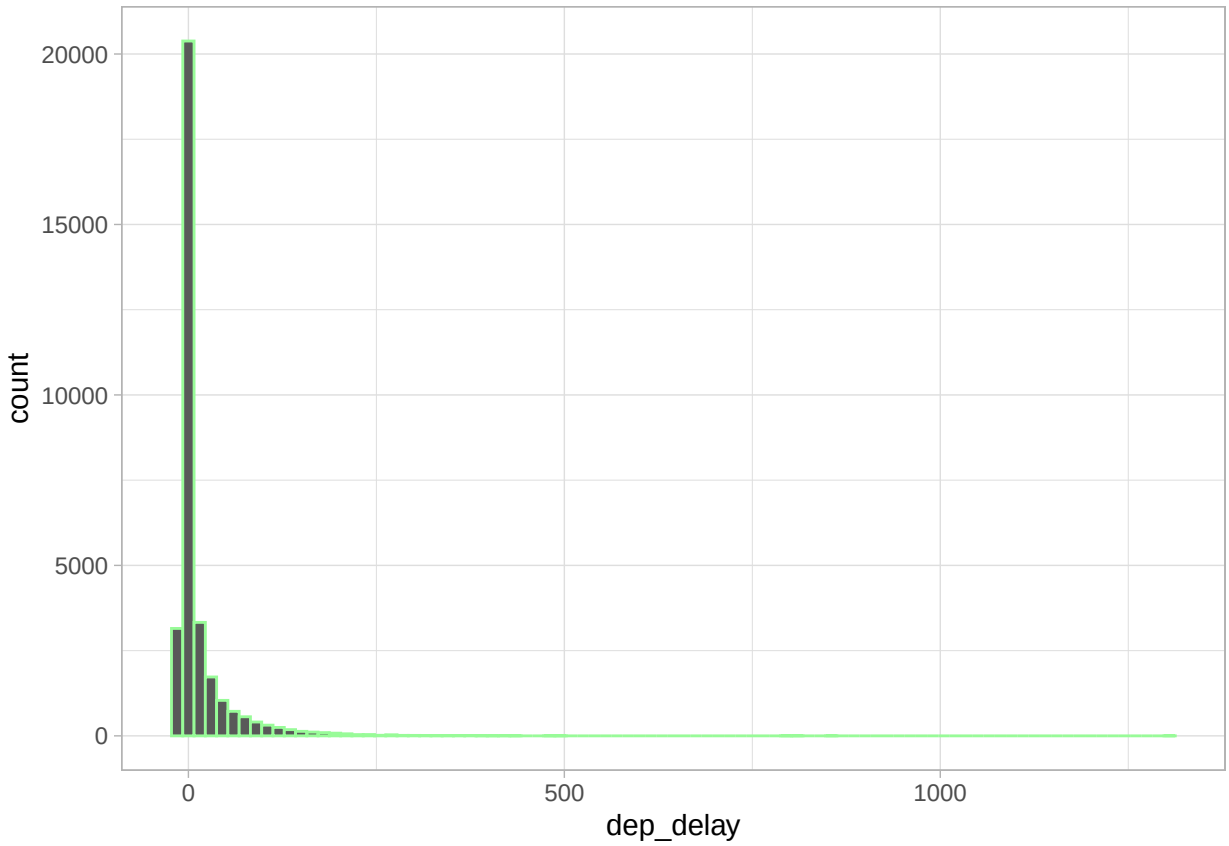
```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

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

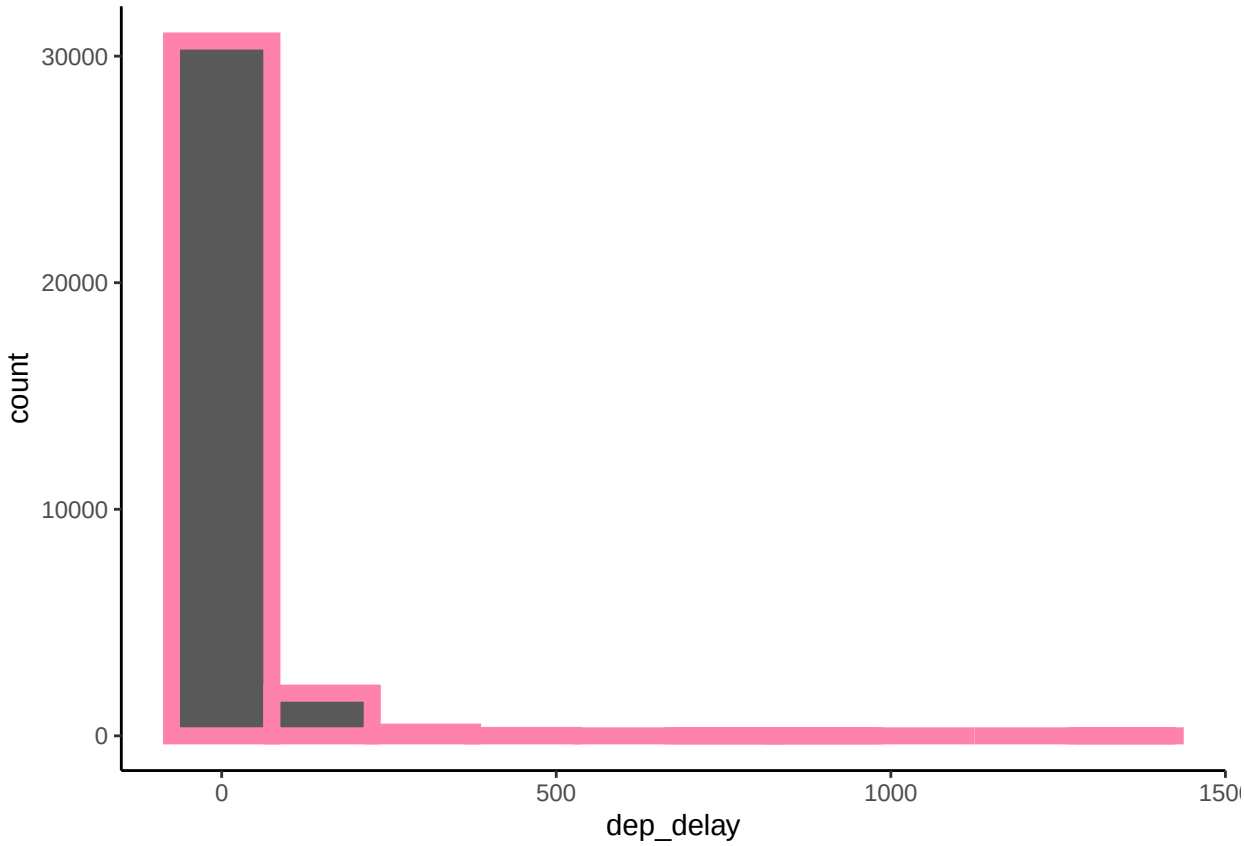



## Definining the binwidth on depature delays histogram

```
ggplot(data = nycflights, aes(x = dep_delay))+
  geom_histogram(colour = "palegreen", binwidth = 15)+
  theme_light()
```




```
ggplot(data = nycflights, aes(x = dep_delay))+
  geom_histogram(colour = "palevioletred1", size = 3, binwidth = 150)+
theme_classic()
```




## Exercise 1

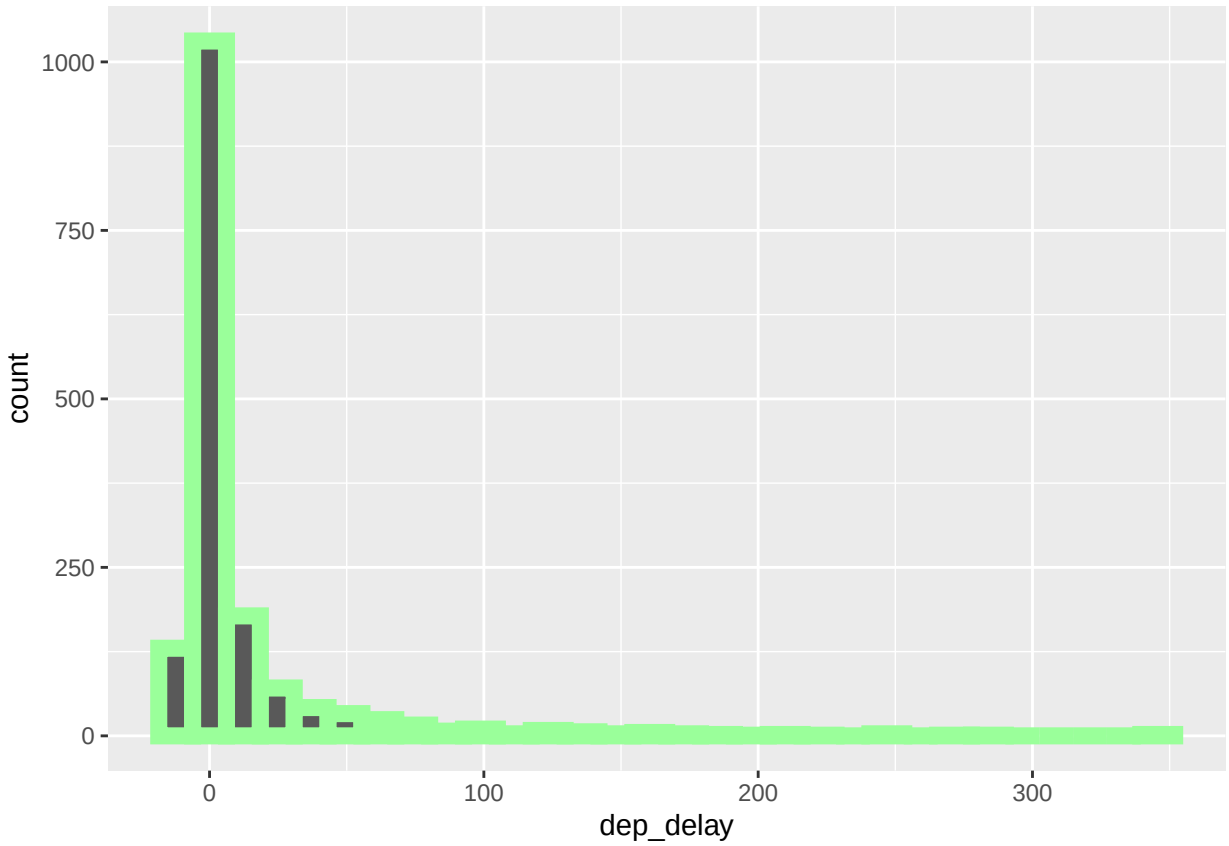
**Look carefully at these three histograms. How do they compare? Are features revealed in one that are obscured in another?**

**Answer: The smaller the binwidth is, the finer the details are.You are able to see a chunk of data that shows that most flights left with a delay of 15 minutes or less.**

## Delays of flights headed to LAX

```
lax_flights <- nycflights %>%
dplyr:: filter(dest == "LAX")
ggplot(data = lax_flights, aes(x = dep_delay))+
geom_histogram(colour = "palegreen1", size = 3)
```

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




## Numeric summaries for the delayed of flights headed to LAX

```
lax_flights %>%
  summarise(mean_dd   = mean(dep_delay),
            median_dd = median(dep_delay),
            n         = n())
```

```
## # A tibble: 1 x 3
##   mean_dd median_dd     n
##     <dbl>     <dbl> <int>
## 1    9.78        -1  1583
```

## Flights headed to San Francisco in February

```
sfo_feb_flights <- nycflights %>%
dplyr::  filter(dest == "SFO", month == 2)
```

## Exercies 2

**Create a new data frame that includes flights headed to SFO in February, and save this data frame as sfo_feb_flights. How many flights meet these criteria?**

**68 flights met the criteria**

```
sfo_feb_flights %>%
  group_by(origin) %>%
  summarise(median_dd = median(dep_delay), iqr_dd = IQR(dep_delay), n_flights = n())
```

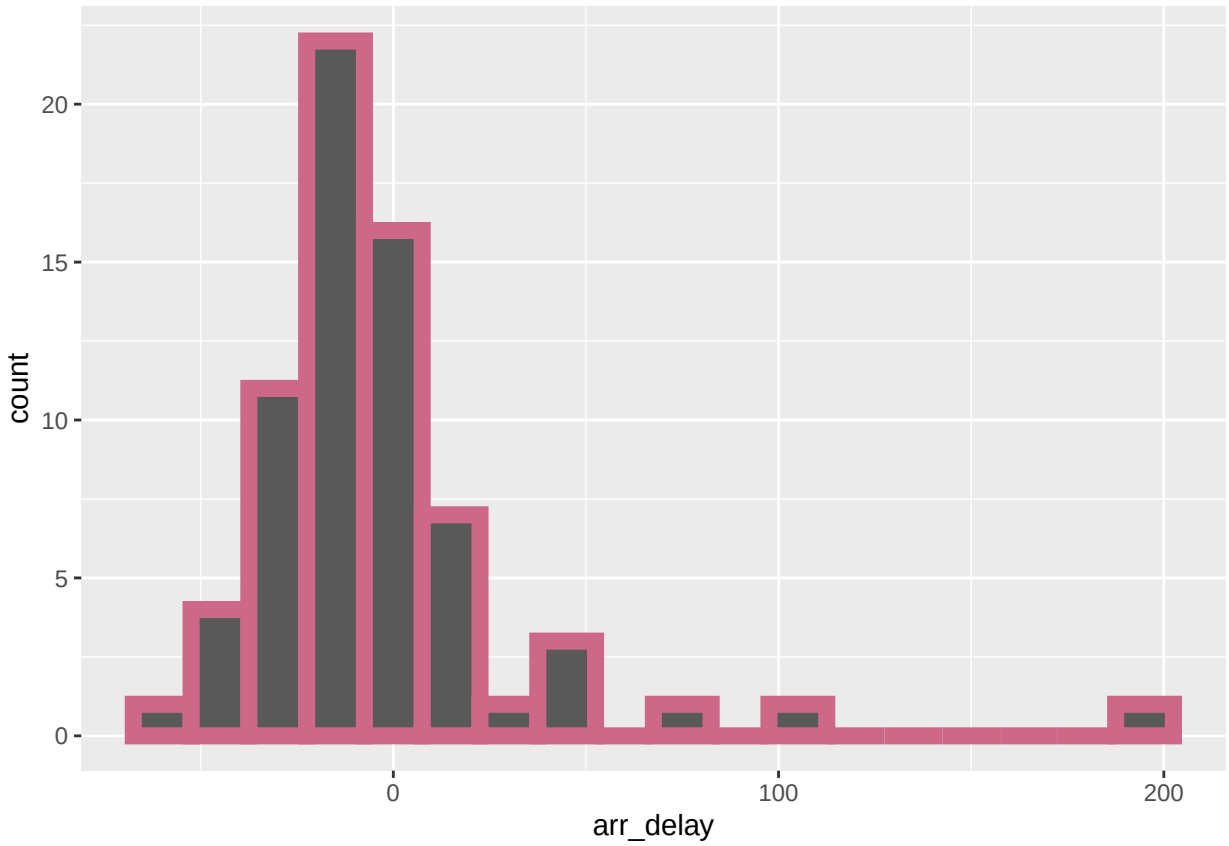
```
## # A tibble: 2 x 4
##   origin median_dd iqr_dd n_flights
##   <chr>      <dbl>  <dbl>     <int>
## 1 EWR          0.5   5.75         8
## 2 JFK         -2.5  15.2         60
```

## Exercise 3

**Describe the distribution of the arrival delays of these flights using a histogram and appropriate summary statistics. Hint: The summary statistics you use should depend on the shape of the distribution.**

**Answer: This group is distributed monomodally and skewed right. Most of the flights arrived early.**

```
ggplot(sfo_feb_flights, aes(x = arr_delay)) + geom_histogram(binwidth=15, colour = "palevioletred3", si
```




```
sfo_feb_flights %>%
  summarise(mean_ad = mean(arr_delay), median_ad = median(arr_delay), iqr_ad = IQR(arr_delay), n_flights
```

```
## # A tibble: 1 x 4
##   mean_ad median_ad iqr_ad n_flights
##     <dbl>     <dbl>  <dbl>     <int>
## 1    -4.5       -11   23.2        68
```

## Exercise 4

**Calculate the median and interquartile range for arr_delays of flights in in the sfo_feb_flights data frame, grouped by carrier. Which carrier has the most variable arrival delays?**

**Answer: United has the most variable arrival delays**

```
sfo_feb_flights %>%
  group_by(carrier) %>%
  summarize(var_arr_delay = mean(var(arr_delay))) %>%
  arrange(desc(var_arr_delay))
```

```
## # A tibble: 5 x 2
##   carrier var_arr_delay
##   <chr>           <dbl>
## 1 UA              2335.
## 2 VX              1669.
## 3 AA               868.
```

```
## 4 DL               485.
## 5 B6               121.
```

## Departure delays by month

```
nycflights %>%
  group_by(month) %>%
  summarise(mean_dd = mean(dep_delay)) %>%
  arrange(desc(mean_dd))
```

```
## # A tibble: 12 x 2
##    month mean_dd
##    <int>   <dbl>
##  1     7   20.8
##  2     6   20.4
##  3    12   17.4
##  4     4   14.6
##  5     3   13.5
##  6     5   13.3
##  7     8   12.6
##  8     2   10.7
##  9     1   10.2
## 10     9    6.87
## 11    11    6.10
## 12    10    5.88
```

## Exercise 5

**Suppose you really dislike departure delays and you want to schedule your travel in a month that minimizes your potential departure delay leaving NYC. One option is to choose the month with the lowest mean departure delay. Another option is to choose the month with the lowest median departure delay. What are the pros and cons of these two choices?**

**The median can tell you more about how likely it is for a flight to be delayed for a given amount of time.**

## On time departure rate for NYC airports

```
nycflights <- nycflights %>%
  mutate(dep_type = ifelse(dep_delay < 5, "on time", "delayed"))
```

```
nycflights %>%
  group_by(origin) %>%
  summarise(ot_dep_rate = sum(dep_type == "on time") / n()) %>%
  arrange(desc(ot_dep_rate))
```

```
## # A tibble: 3 x 2
##   origin ot_dep_rate
##   <chr>        <dbl>
## 1 LGA          0.728
## 2 JFK          0.694
```

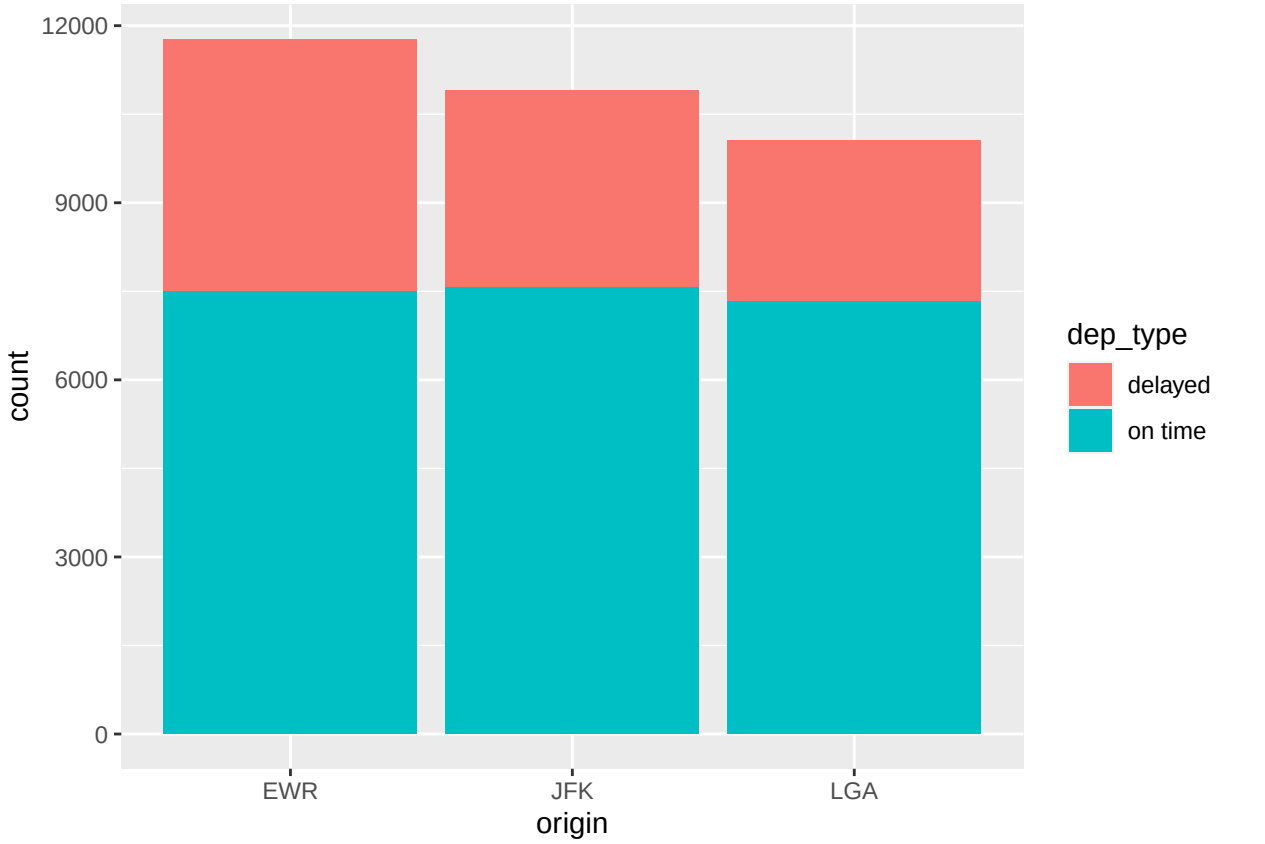
```
## 3 EWR          0.637
```

## Exercies 6

**If you were selecting an airport simply based on on time departure percentage, which NYC airport would you choose to fly out of?**

**Answer: I would select LGA**

```
ggplot(data = nycflights, aes(x = origin, fill = dep_type))+
  geom_bar()
```




## Exercise 7

**Mutate the data frame so that it includes a new variable that contains the average speed, avg_speed traveled by the plane for each flight (in mph). Hint: Average speed can be calculated as distance divided by number of hours of travel, and note that air_time is given in minutes.**

```
nycflights <- nycflights %>%
  mutate(nycflights , avg_speed = distance / air_time)
```

## Exercise 8

**Make a scatterplot of avg_speed vs. distance. Describe the relationship between average speed and distance. Hint: Use geom_point().**

```
nycflights %>%
  group_by(tailnum) %>%
  summarise( avg_speed = mean(avg_speed) ) %>%
  arrange(desc(avg_speed))
```

```
## # A tibble: 3,490 x 2
##    tailnum avg_speed
##    <chr>       <dbl>
##  1 N526AS       8.49
##  2 N637DL       8.43
##  3 N66051       8.41
##  4 N907JB       8.41
##  5 N522VA       8.38
##  6 N5BTAA       8.32
##  7 N654UA       8.31
##  8 N382HA       8.25
##  9 N75861       8.25
## 10 N5DRAA       8.22
## # i 3,480 more rows
```

```
nycflights %>% ggplot() +
  geom_point(aes(x = avg_speed, y = distance, color = carrier))
```