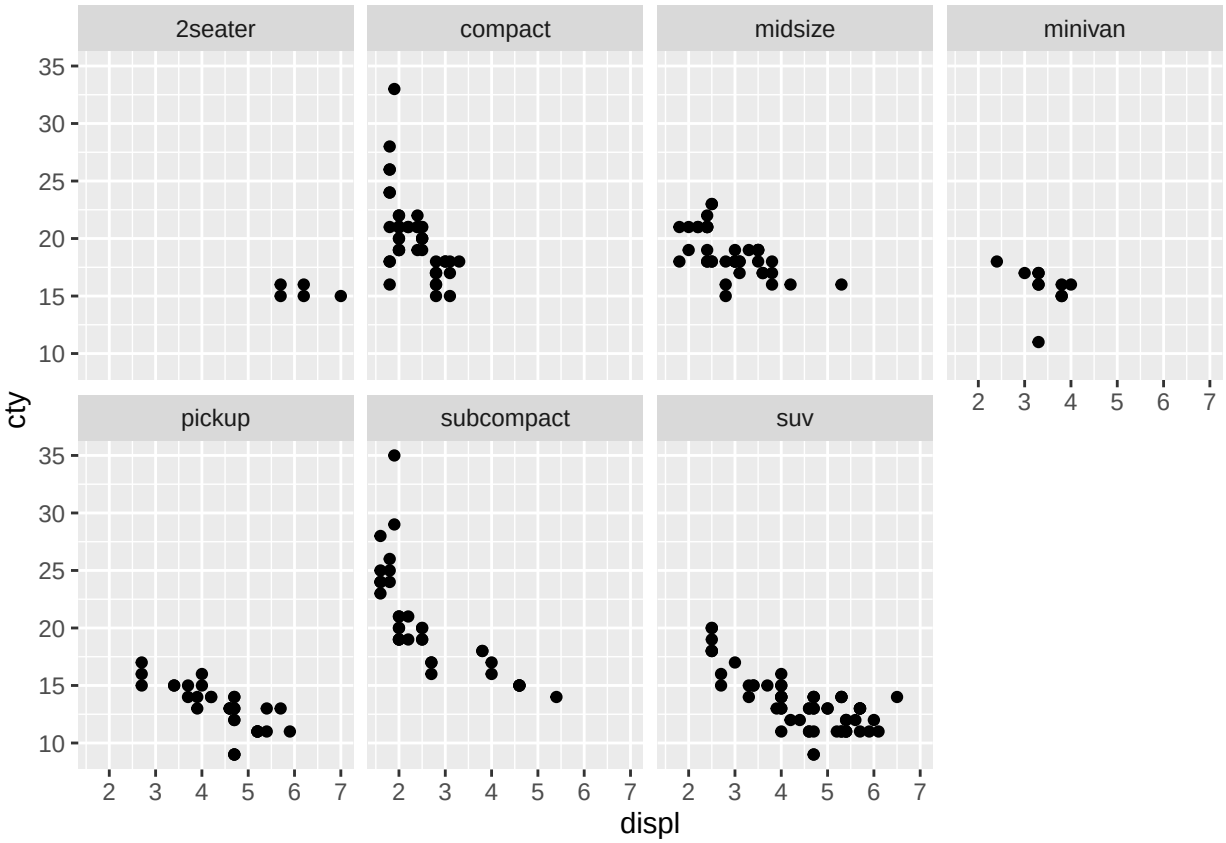
STAT3622 A1

Havin Chung (3035729772)

2025-02-14

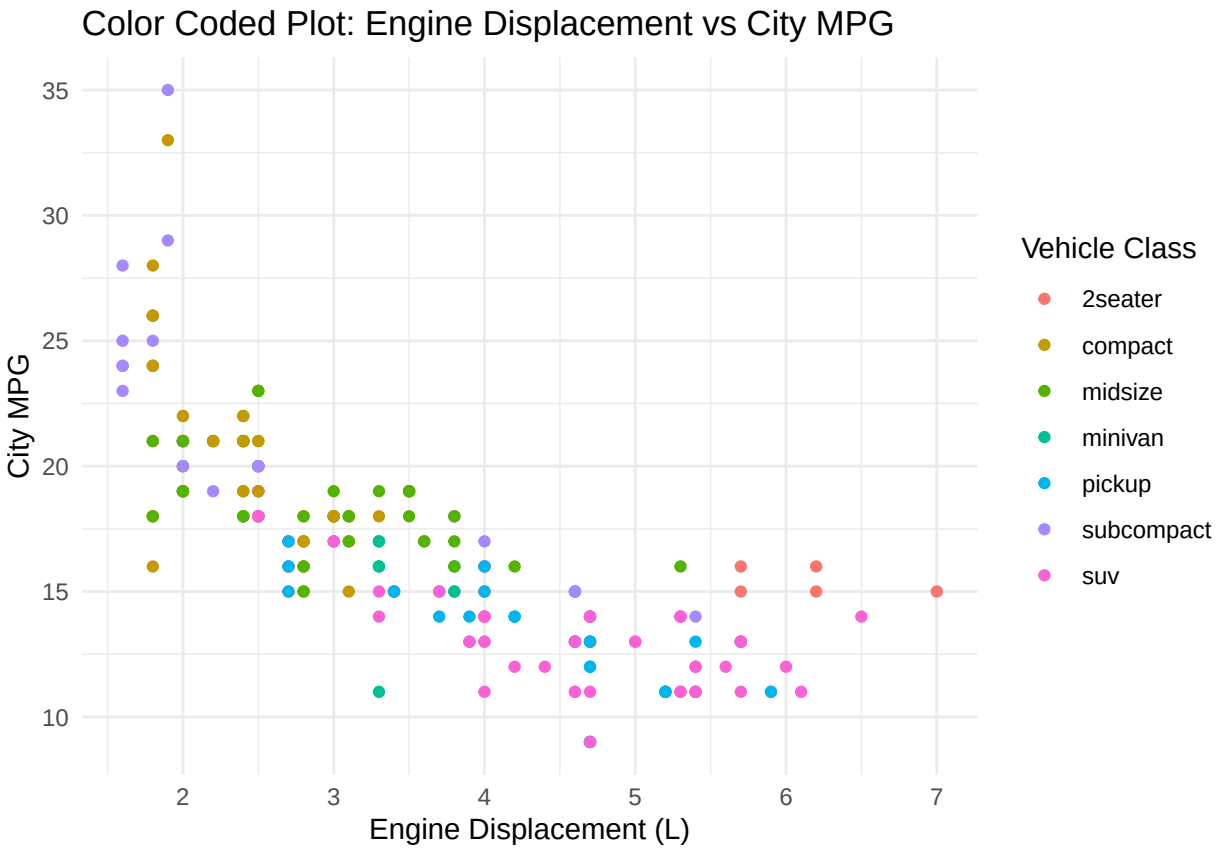
Q1

```
ggplot(mpg, aes(x = displ, y = cty)) +
  geom_point() +
  facet_wrap(~class, nrow = 2)
```



The above is the plot faceted on *class*

```
ggplot(mpg, aes(x = displ, y = cty, color = class)) +
  geom_point() +
  labs(title = "Color Coded Plot: Engine Displacement vs City MPG",
       x = "Engine Displacement (L)",
       y = "City MPG",
       color = "Vehicle Class") +
  theme_minimal()
```



Advantages of Faceting:

1. It provides single compact visualization with single panel which makes easier to interpret
2. All points are together which allows to directly compare values between different classes

Disadvantages:

1. With large data, the overlapping points and many colors could make it hard to distinguish and confusing
2. Not color blind friendly

If the dataset were much larger, faceting would still help in separating data, but too many categories could make the visualization cluttered. Using color might not be effective because points would overlap heavily, making it difficult to distinguish categories. In such cases, a combination of faceting and color might be needed.

Q2

(1)

```
mean_waiting <- mean(faithful$waiting)
mean_waiting
```

```
## [1] 70.89706
```

The average waiting time between eruptions of the *Old Faithful* geyser in the dataset *faithful* is 70.89706 minutes.

(2)

```
faithful$eruptions[1:4]
```

```
## [1] 3.600 1.800 3.333 2.283
```

This code extract first four eruption duration from dataset.

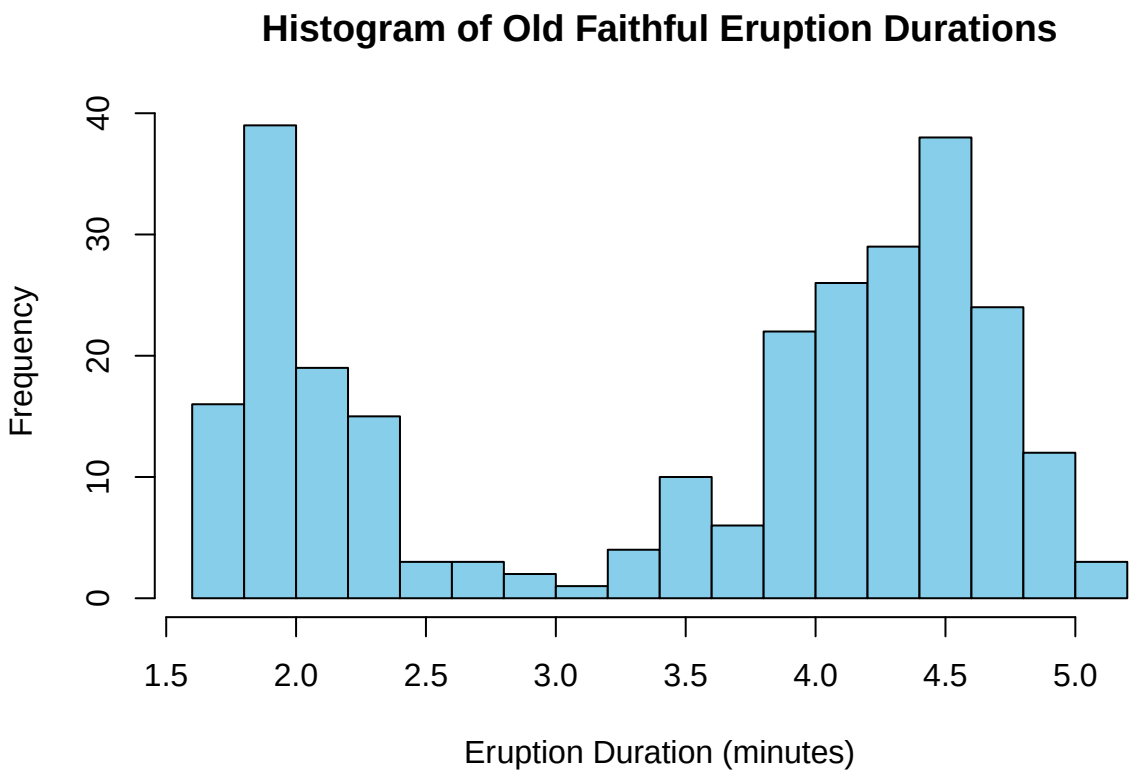
(3)

eruptions	waiting
3.600	79
1.800	54
3.333	74
2.283	62
4.533	85

The shown table above indicates the first five rows of the *faithful* data frame.

(4)

```
hist(faithful$eruptions,
     main = "Histogram of Old Faithful Eruption Durations",
     xlab = "Eruption Duration (minutes)",
     col = "skyblue",
     border = "black",
     breaks = 20)
```



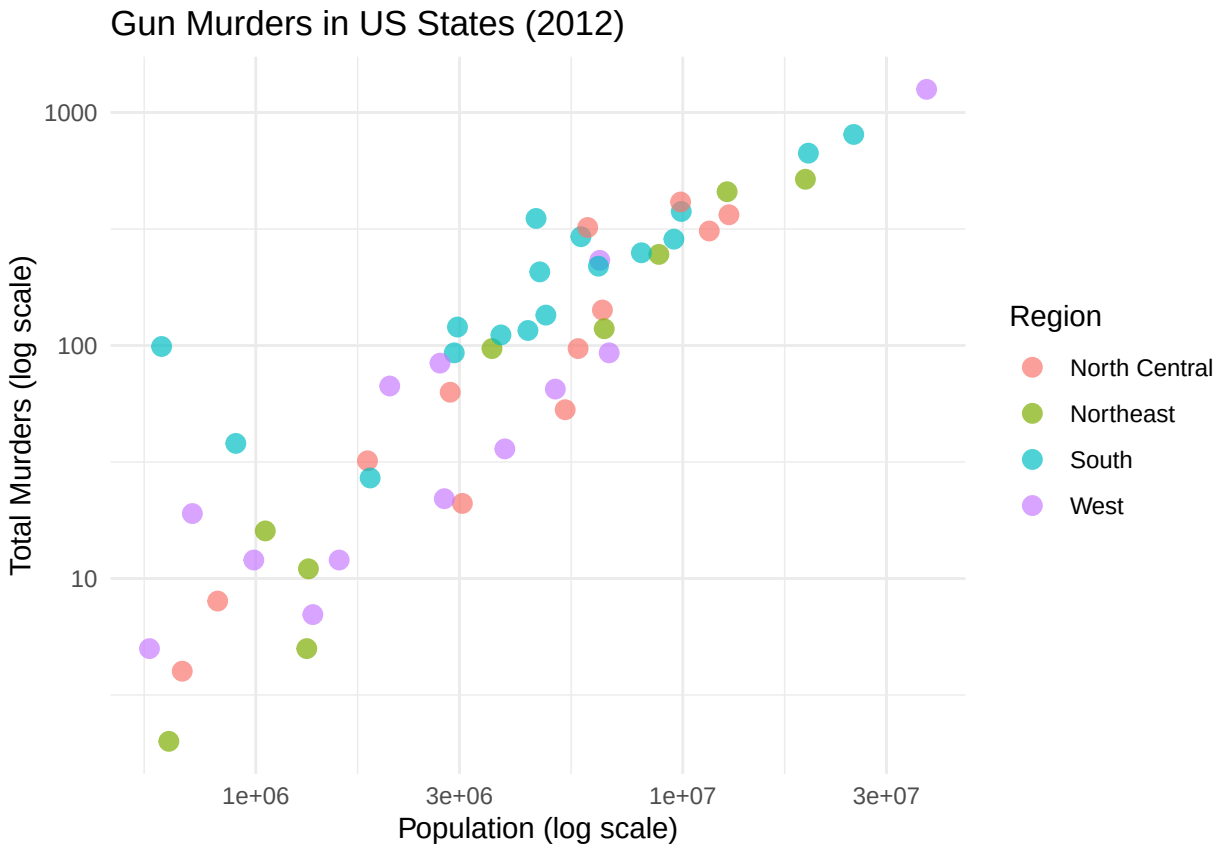
This is the histogram of the eruption durations in the *faithful* data set.

Q3

```
murders <- read_csv("/Users/havinchung/Desktop/STAT3622/A1/murders.csv", show_col_types = FALSE)
head(murders)
```

```
## # A tibble: 6 x 5
##   state      abb   region population total
##   <chr>      <chr> <chr>       <dbl> <dbl>
## 1 Alabama    AL    South     4779736   135
## 2 Alaska     AK    West       710231    19
## 3 Arizona    AZ    West      6392017   232
## 4 Arkansas   AR    South     2915918    93
## 5 California CA    West     37253956  1257
## 6 Colorado   CO    West      5029196    65
```

```
ggplot(murders, aes(x = population, y = total, color = region)) +
  geom_point(size = 3, alpha = 0.7) +
  scale_x_log10() +  # Log scale for x-axis
  scale_y_log10() +  # Log scale for y-axis
  labs(title = "Gun Murders in US States (2012)",
       x = "Population (log scale)",
       y = "Total Murders (log scale)",
       color = "Region") +
  theme_minimal()
```



The scatter plot shows a strong correlation between state population and total murders. It also shows some regional differences where Southern states generally showing higher murder rates, while the Northeast has lower values.

Q4

```
unique(flights$origin)
```

```
## [1] "EWR" "LGA" "JFK"
```

(1)

```
delay_data <- flights %>%
  drop_na()%>%
  group_by(origin) %>%
  summarise(avg_delay = mean(dep_delay), median_delay = median(dep_delay)) %>%
  arrange(origin)

kbl <- knitr::kable(delay_data, format = "pipe")
kableExtra::kable_styling(kbl, full_width = FALSE)
```

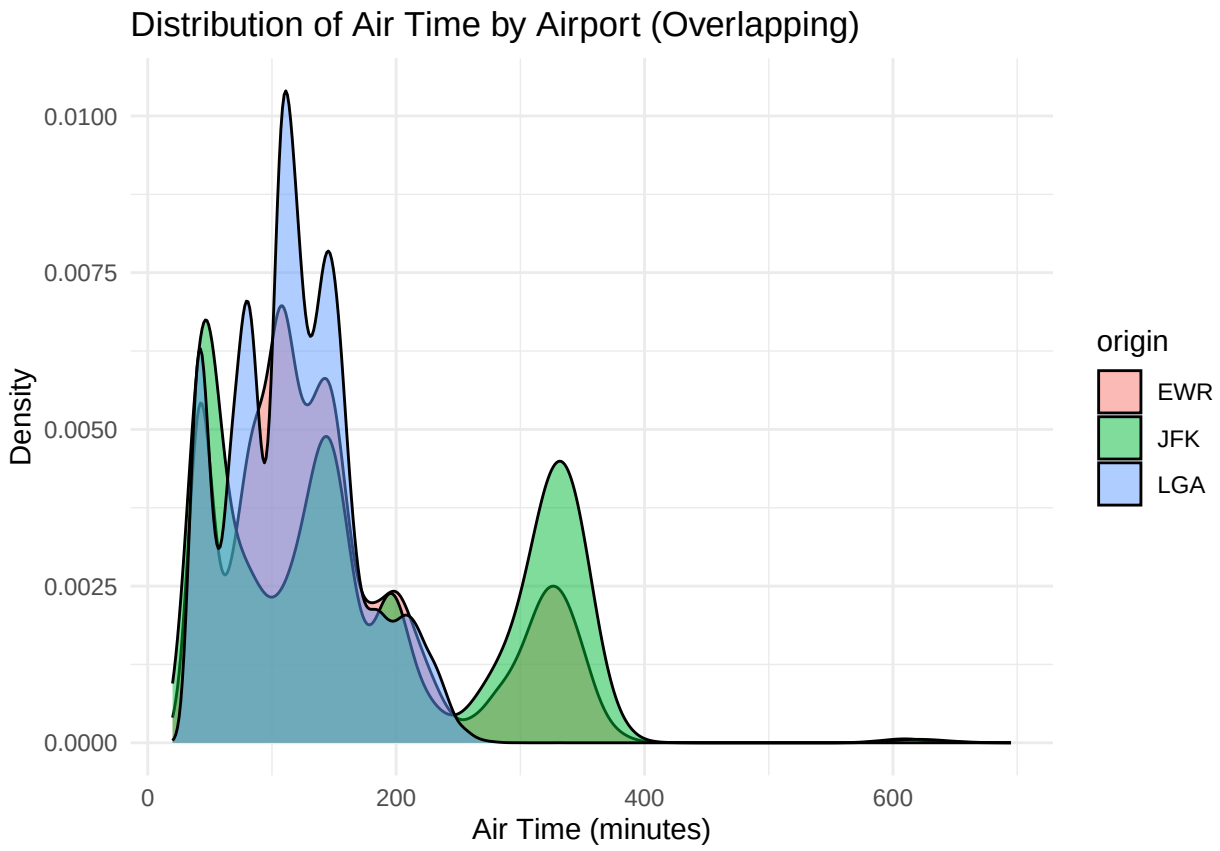
origin	avg_delay	median_delay
EWR	15.00911	-1
JFK	12.02361	-1
LGA	10.28658	-3

The average and median departure delays for the three New York City airports (JFK, LGA, and EWR) show interesting patterns. EWR exhibits the longest average departure delays, while LGA has the shortest. A notable observation is the significant difference between the mean and median values for all three airports. The mean values being considerably larger than the median values suggests that the distribution of departure delays is right-skewed. This indicates that while most flights experience delays closer to the median, there are some flights with exceptionally long delays that pull the average higher.

(2)

First Approach: Single Plot

```
ggplot(flights %>% drop_na(), aes(x = air_time, fill = origin)) +
  geom_density(alpha = 0.5) +
  theme_minimal() +
  labs(title = "Distribution of Air Time by Airport (Overlapping)",
       x = "Air Time (minutes)",
       y = "Density")
```



Advantages:

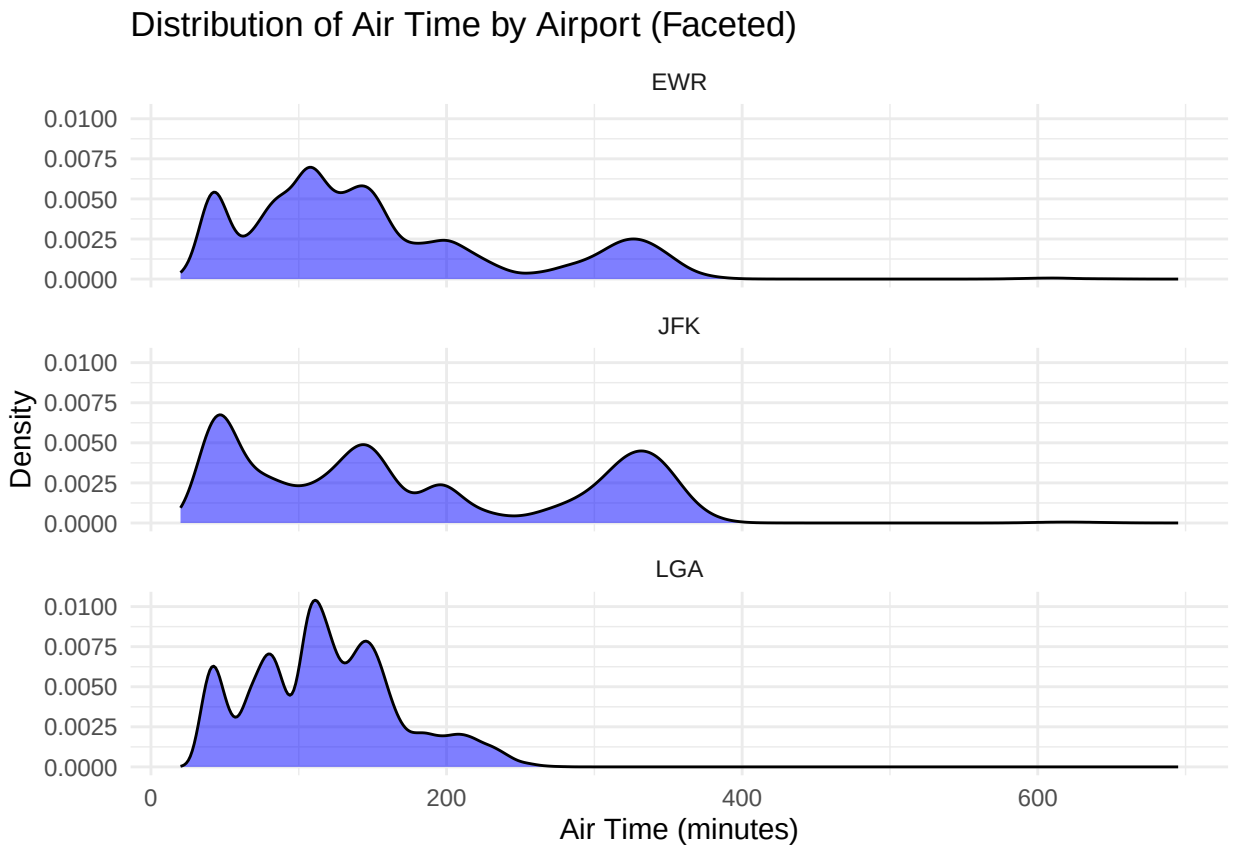
- Direct comparison of distributions is easy
- Useful for a small groups like the example
- Easy to recognize overall patterns and differences

Disadvantages:

- Can become cluttered and hard to read with many groups
- May be challenging to distinguish overlapping areas, especially with similar distributions

Second Approach: Faceted Plot

```
ggplot(flights %>% drop_na(), aes(x = air_time)) +
  geom_density(fill = "blue", alpha = 0.5) +
  facet_wrap(~origin, ncol = 1) +
  theme_minimal() +
  labs(title = "Distribution of Air Time by Airport (Faceted)",
       x = "Air Time (minutes)",
       y = "Density")
```



Advantages:

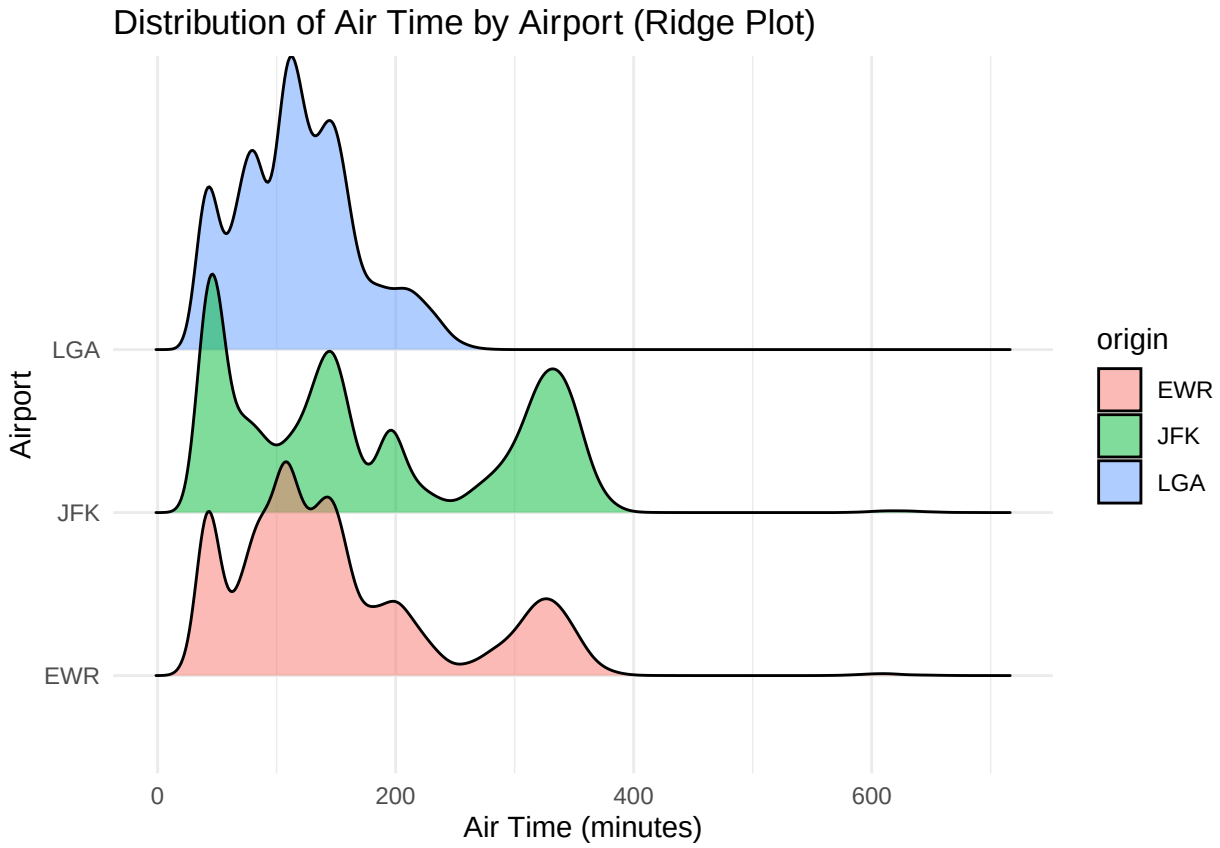
- Clear, unobstructed view of each individual distribution
- Easier to examine details of each group without interference
- Scales well to a larger number of groups

Disadvantages:

- Direct comparison between groups may be more difficult
- Takes up more space, especially with many groups
- May make it harder to see overall patterns across all groups at once

Third Approach: Ridge Line Plot

```
ggplot(flights %>% drop_na() , aes(x = air_time, y = origin, fill = origin)) +
  geom_density_ridges(alpha = 0.5) +
  theme_minimal() +
  labs(title = "Distribution of Air Time by Airport (Ridge Plot)",
       x = "Air Time (minutes)",
       y = "Airport")
```



Advantages:

- Combines benefits of single plot and faceted approaches
- Allows for easy comparison of distribution shapes and peaks

Disadvantages:

- Can be less familiar to some audiences
- Precise comparisons of density values may be challenging
- Lower distributions may be partially obscured by those above them

Observed Difference

The distribution of air times reflects the different roles and types of flights each airport predominantly handles. LGA's shorter flight times suggest a focus on shorter-haul domestic routes, while JFK's longer durations indicate a significant number of long-haul international flights. EWR's intermediate position in the distribution implies a mix of both domestic and international services, showcasing its versatility as an airport serving diverse flight types