Flights Project

AUTHOR
Benjamin Fuentes

```
library(tidyverse)
```

```
── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
✓ dplyr     1.1.4     ✓ readr     2.1.5
✓ forcats   1.0.0     ✓ stringr   1.5.1
✓ ggplot2   3.4.4     ✓ tibble    3.2.1
✓ lubridate 1.9.3     ✓ tidyr     1.3.1
✓ purrr     1.0.2
── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become
errors
```

```
filename <- file.choose()
flights <- readRDS(filename)
```

Introduction

This is an analysis of current airlines and their performance over the course of one year between January 2015 and January 2016. The analysis will take into consideration many factors that will help determine which airline companies were the best, which were the worst, and shed light on geographical differences between the companies. As a business journalist, all companies will be evaluated on the same factors and using the same data sources. A business analyst report for United Airlines makes up the second part of this report, where United Airlines is evaluated when compared to other airlines, along with pointing out key weaknesses and strengths.

Data Overview

The data that will be used in this analysis was obtained from the website Kaggle and consists of 3 data frames: “flights”, “airlines” and “airports”. A quick glimpse of the data and its content can be seen below, it’s important to note that most of the data is found in the “flights” table, thus a major portion of analysis will be performed on this table. As a business journalist, this analysis will be done with the intent to determine which airlines are the best and which are the worst by using several analysis lenses.

https://www.kaggle.com/.

Flights table:

```
glimpse(flights$flights)
```

```
Rows: 5,819,079
Columns: 16
$ date                <date> 2015-01-01, 2015-01-01, 2015-01-01, 2015-01-01, 2…
$ day_of_week         <dbl> 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4,…
$ airline             <chr> "AS", "AA", "US", "AA", "AS", "DL", "NK", "US", "A…
$ flight_number       <dbl> 98, 2336, 840, 258, 135, 806, 612, 2013, 1112, 117…
$ origin_airport      <chr> "ANC", "LAX", "SFO", "LAX", "SEA", "SFO", "LAS", "…
$ destination_airport <chr> "SEA", "PBI", "CLT", "MIA", "ANC", "MSP", "MSP", "…
$ scheduled_departure <chr> "0005", "0010", "0020", "0020", "0025", "0025", "0…
$ departure_time      <chr> "2354", "0002", "0018", "0015", "0024", "0020", "0…
$ scheduled_time      <dbl> 205, 280, 286, 285, 235, 217, 181, 273, 195, 221, …
$ elapsed_time        <dbl> 194, 279, 293, 281, 215, 230, 170, 249, 193, 203, …
$ distance            <dbl> 1448, 2330, 2296, 2342, 1448, 1589, 1299, 2125, 14…
$ arrival_delay       <dbl> -22, -9, 5, -9, -21, 8, -17, -10, -13, -15, -30, -…
$ diverted            <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, F…
$ cancelled           <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, F…
$ flight_nubmer       <chr> "98", "2336", "840", "258", "135", "806", "612", "…
$ delayed             <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, F…
```

Airlines table:

```
glimpse(flights$airlines)
```

```
Rows: 14
Columns: 2
$ iata_code <chr> "UA", "AA", "US", "F9", "B6", "OO", "AS", "NK", "WN", "DL", …
$ airline   <chr> "United Air Lines Inc.", "American Airlines Inc.", "US Airwa…
```

Airports table:

```
glimpse(flights$airports)
```

```
Rows: 322
Columns: 7
$ iata_code <chr> "ABE", "ABI", "ABQ", "ABR", "ABY", "ACK", "ACT", "ACV", "ACY…
$ airport   <chr> "Lehigh Valley International Airport", "Abilene Regional Air…
$ city      <chr> "Allentown", "Abilene", "Albuquerque", "Aberdeen", "Albany",…
$ state     <chr> "PA", "TX", "NM", "SD", "GA", "MA", "TX", "CA", "NJ", "AK", …
$ country   <chr> "USA", "USA", "USA", "USA", "USA", "USA", "USA", "USA", "USA…
$ latitude  <dbl> 40.65236, 32.41132, 35.04022, 45.44906, 31.53552, 41.25305, …
$ longitude <dbl> -75.44040, -99.68190, -106.60919, -98.42183, -84.19447, -70.…
```

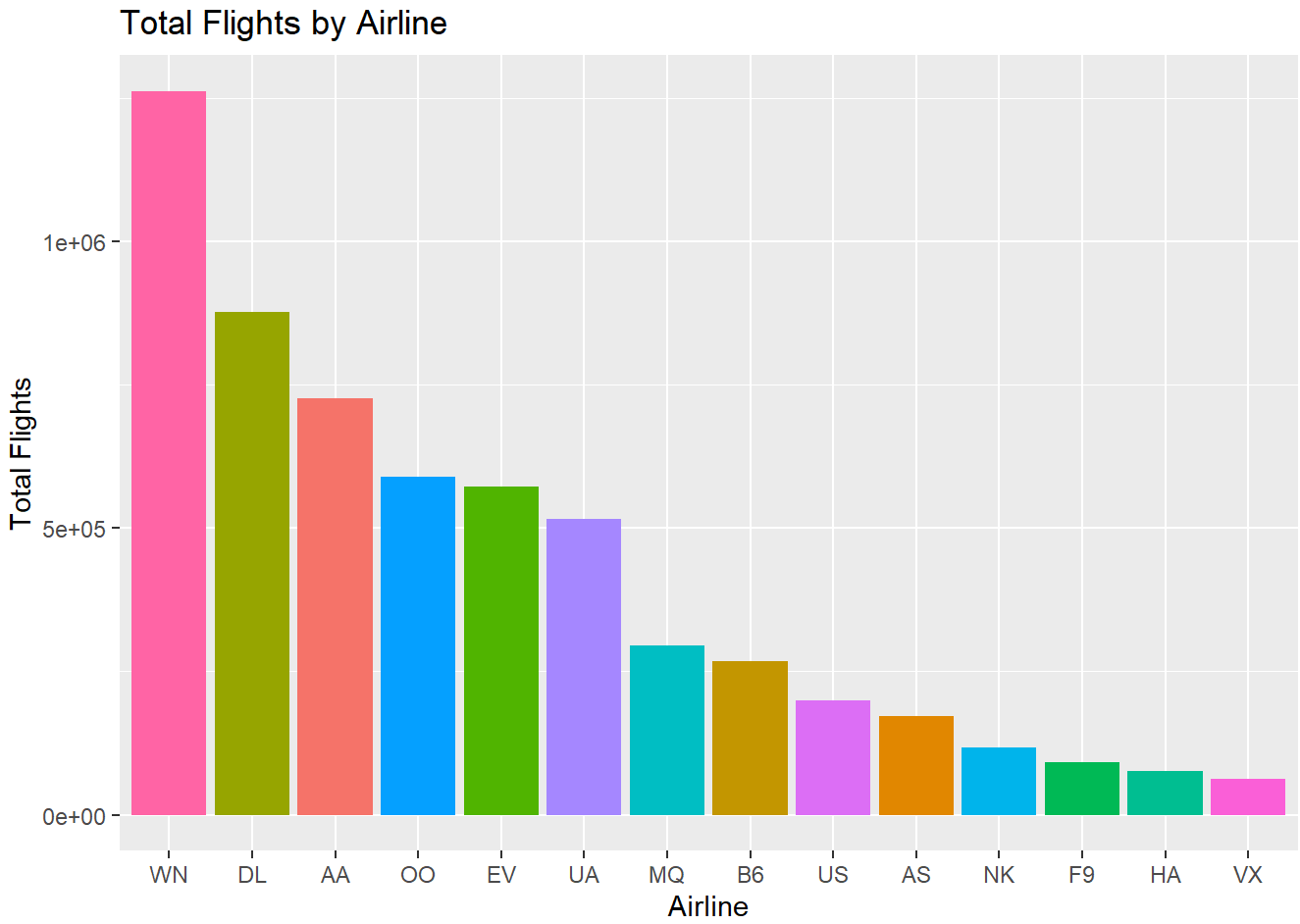
Flight Volume

The first point of analysis will be flight volume for all airlines. Flight volume essentially refers to the total number of flights recorded for each airline. In this case, "total_flights_summary" will be the variable used to reference flight volume:

```
total_flights_summary <- flights$flights %>%
  group_by(airline) %>% # grouping by airline
  summarize(total_flights = n()) %>% # assigning total observations to total_flights
  arrange(desc(total_flights)) # rearranging in descending order by total flights
```

Taking a look at flight volume visualized below, we can identify the airlines with the most flights and the least flights. Southwest Airlines Co. had the most amount of flights during this time frame with 1,261,855 flights followed by Delta Air Lines Inc. with 875,881. Virgin America had the least amount of flights at 61,903 flights with Hawaiian Airlines Inc. following after. It's worth noting that Virgin America later merges with Alaska Airlines in 2018, 2 years after the recorded data, we can then consider flight volume a possible factor that might have contributed to this merge. Further information and analysis could help provide more possible factors to aid this discussion.

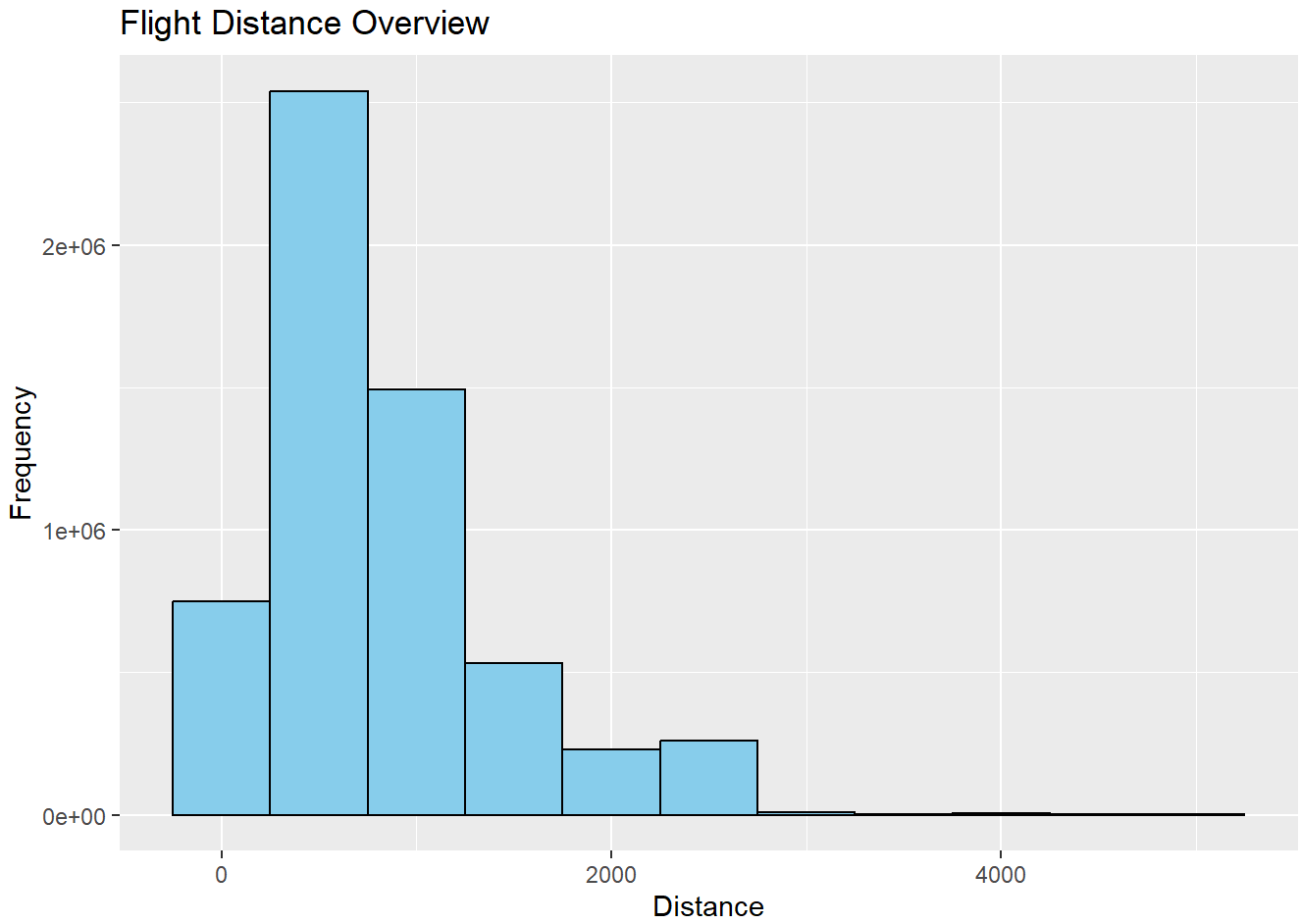
```
total_flights_summary %>%
  ggplot(aes(x = reorder(airline,
                         rev(total_flights)),
             y = total_flights,
             fill = airline)) +
  geom_col() + # setting bar chart
  labs(title = "Total Flights by Airline", x = "Airline", y = "Total Flights") +
  theme(legend.position = "none") # removing legend
```



Flight Duration and Distance

We've examined the airlines with the most and least amount flights, let's take a look at the distribution of flight travel distance overall determine if there are any possible connections.

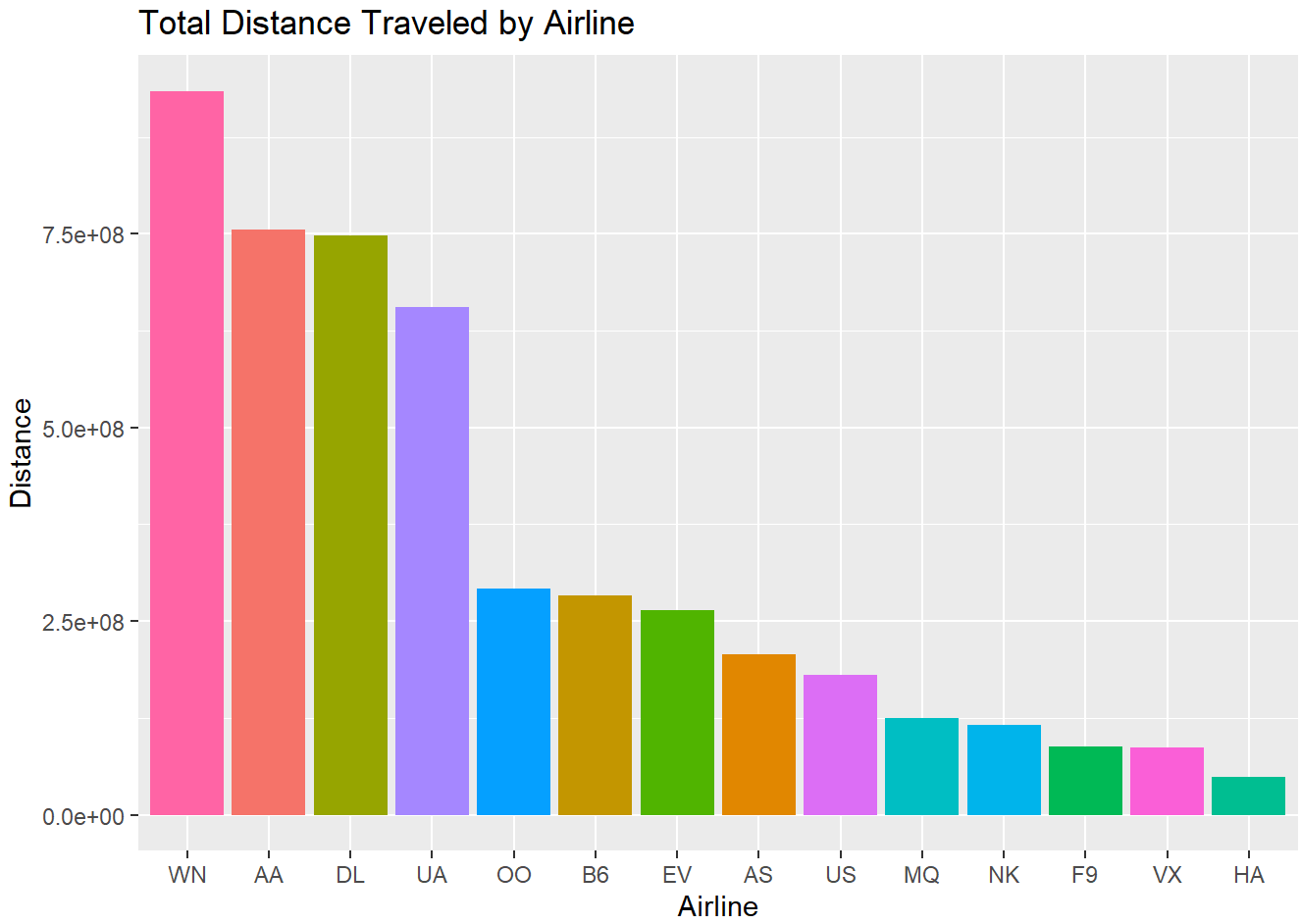
```
flights$flights %>%
  ggplot(aes(x = distance)) +
  geom_histogram(binwidth = 500,
                 fill = "skyblue",
                 color = "black") +
  labs(title = "Flight Distance Overview", x = "Distance", y = "Frequency")
```



We can see here that overall flight distances were more frequently on the shorter end of the spectrum comapred to longer flights which were much less frequent. Let's a take a deeper look then and see which airlines accounted for these flight distances. In this case, "total_distance" will be the variable used to reference flight duration:

```
total_distance <- flights$flights %>%
  group_by(airline) %>% # grouping by airline
  summarize(distance = sum(distance, na.rm = TRUE)) # new column calculating sum of distance, rem

total_distance %>%
  ggplot(aes(x = reorder(airline, -distance), y = distance, fill = airline)) + # arranging in des
  geom_col() +
  theme(legend.position = "none") + # removing legend from fill
  labs(title = "Total Distance Traveled by Airline", x = "Airline", y = "Distance")
```



The airline with the most traveled distance is observed to be Southwest Airlines Inc. with 934,670,301 miles and the airline having traveled the least distance is Hawaiian Airlines Inc. with 4,824,9045 miles. There seems to be a pattern forming, where Southwest Airlines is once again at the top of the table, having the most amount of flights and also the most amount of traveled distance. On the other hand, Hawaiian Airlines is once again in the bottom two airlines when it comes to these same two factors. Geographical presence is definitely a factor that comes into play when it comes to traveled distance, this analysis will go over that soon, but it's possible customer base might differ greatly between airlines. Perhaps certain customers of one airline are more likely to travel further, while other airlines are more popular and used for short travel distance.

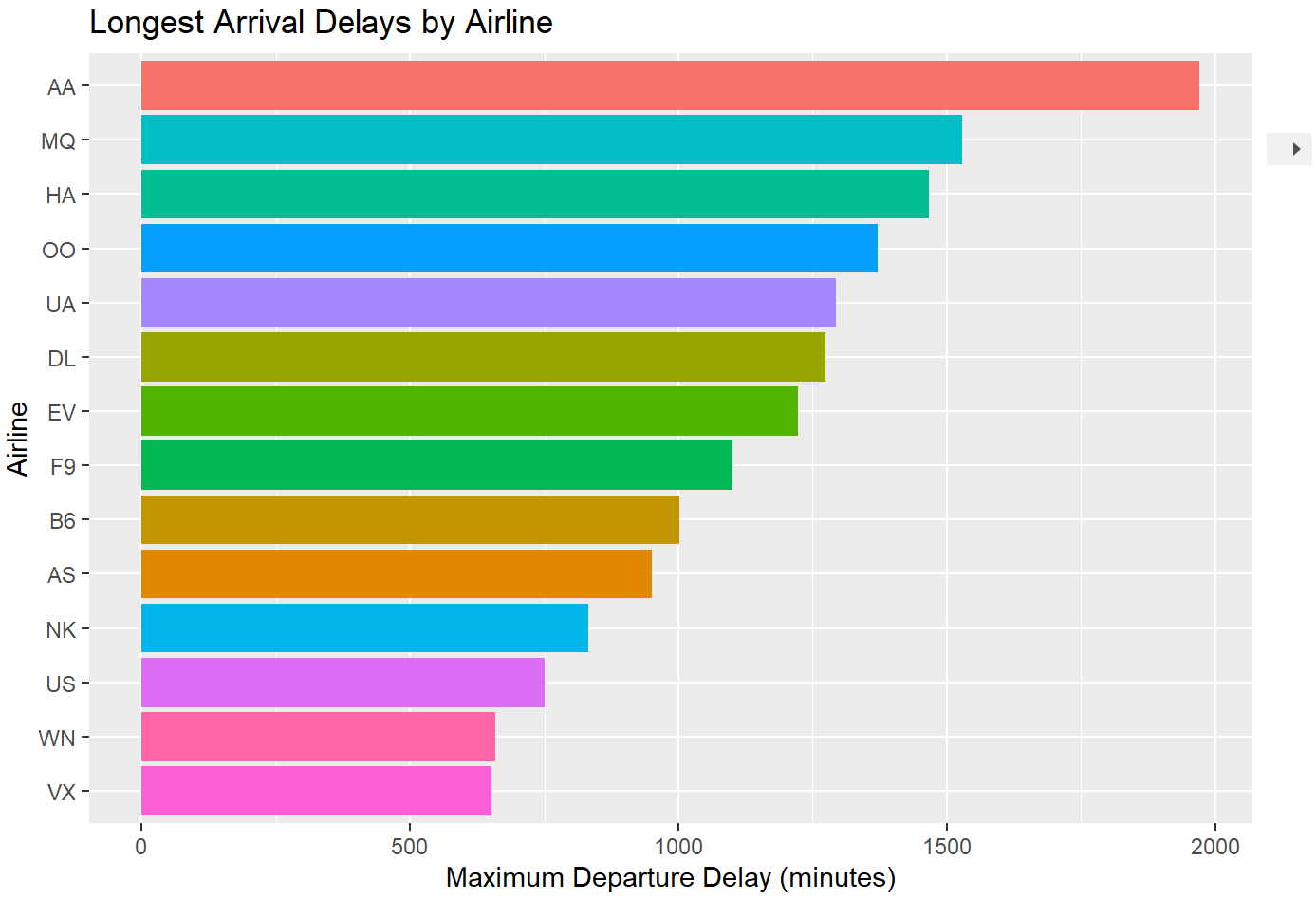
Flight Delays

Up to now distance and flight volume have been examined, and while these are factors that help determine a superior airline, customer satisfaction is a key influencor in a successful airline. When it comes to airlines, customer satisfaction often relies on timing, more specifically, delays. Below is a visualization of the longest arrival delays per airline:

```
max_arrival_delay <- flights$flights %>%
  group_by(airline) %>% # grouping by airline
  summarise(max_arrival_delay = max(arrival_delay, na.rm = TRUE)) # new column calculating max ar

max_arrival_delay %>%
```

```
  ggplot(aes(x = reorder(airline, max_arrival_delay), y = max_arrival_delay, fill = airline)) +
  geom_col() + # setting bar chart
  coord_flip() + # flip x and y coordinates so airlines is on y
  labs(title = "Longest Arrival Delays by Airline", x = "Airline", y = "Maximum Departure Delay (
  theme(legend.position = "none") # removing theme
```

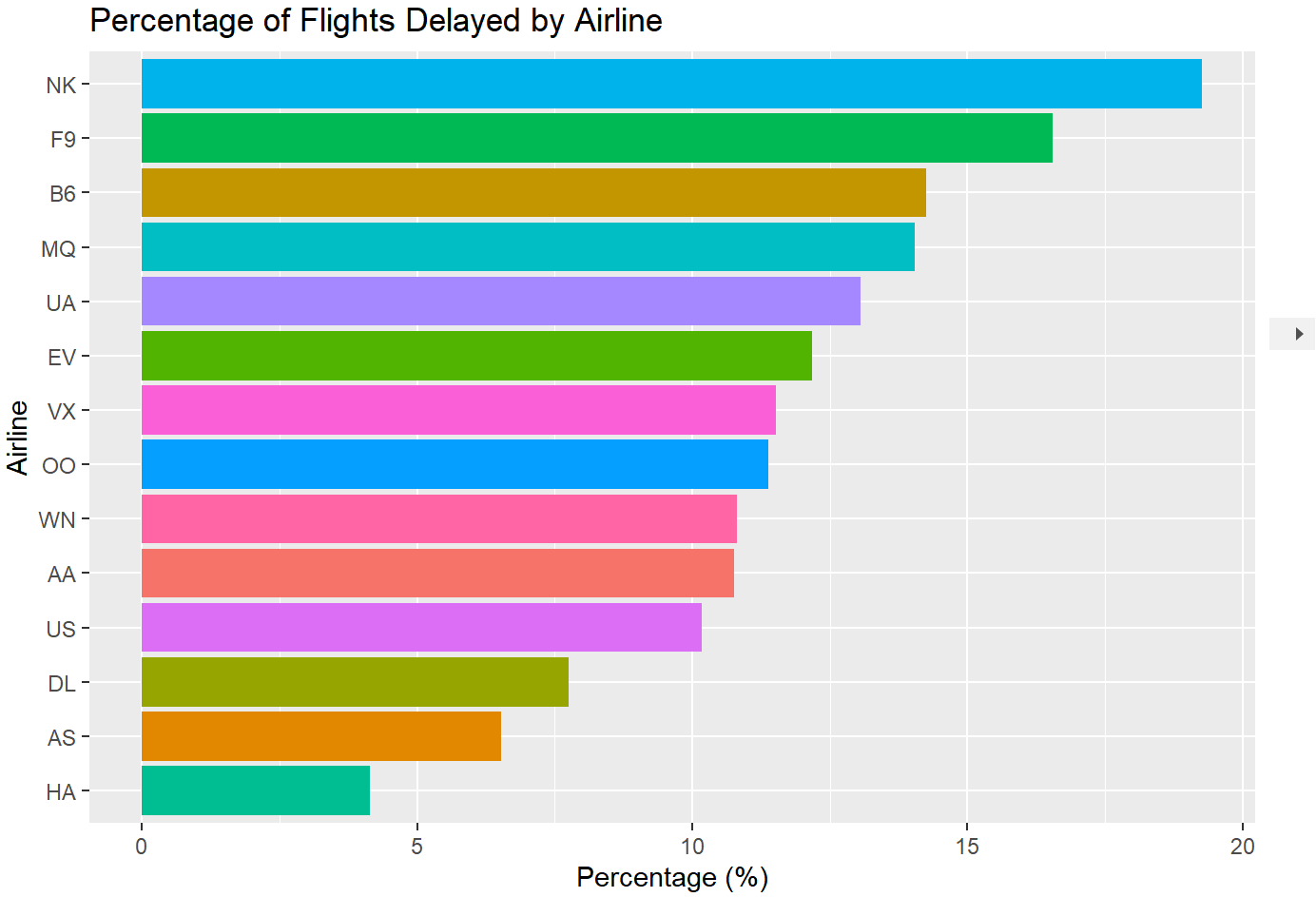


American Airlines Inc. is found to have the maximum recorded departure delay with 1,971 minutes or 32.85 hours. On the other end, Virgin America had the least maximum departure delay at 651 minutes or 10.85 hours. It's worth noting Southwest Airlines Co. can once again be found in the top airline data, having the second least amount of departure delay, despite also having the most recorded total flights. Let's take a further look at flight delays and see what percentage of total flights were delayed for each airline when compared to one another.

```
flights_delayed_airline <- flights$flights %>%
  mutate(delayed = as.numeric(delayed)) %>% # converting delayed column to be treated as numeric
  group_by(airline) %>% # group by airline
  summarise(percent_flights_delayed = mean(delayed, na.rm = TRUE) * 100) # new column calculating

flights_delayed_airline %>%
  ggplot(aes(x = reorder(airline, percent_flights_delayed), y = percent_flights_delayed, fill = a
  geom_bar(stat = "identity") + # displaying values in data frame as they are
  coord_flip() + # flip x and y coordinates so airlines is on y
```

```
  labs(title = "Percentage of Flights Delayed by Airline", x = "Airline", y = "Percentage (%)") +
  theme(legend.position = "none") # removing legend
```



Here we see Hawaiian Airlines Inc. had the least percentage of delayed flights, with Spirit Air Lines having the highest percentage. This is the first time Spirit Air Lines is brought to our attention, so we'll keep a close watch on it moving forward. Taking a step back, these results could be influenced by a few possible factors, mainly, flight volume. Virgin America for example had the least amount flights, thus it's possible there's less room for error in delays, however, this would be a factor needing further analysis. Additionally, despite Hawaiian Airlines Inc. having the highest recorded departure delay, it can be considered to the best airline in terms of percentage of flights delayed.

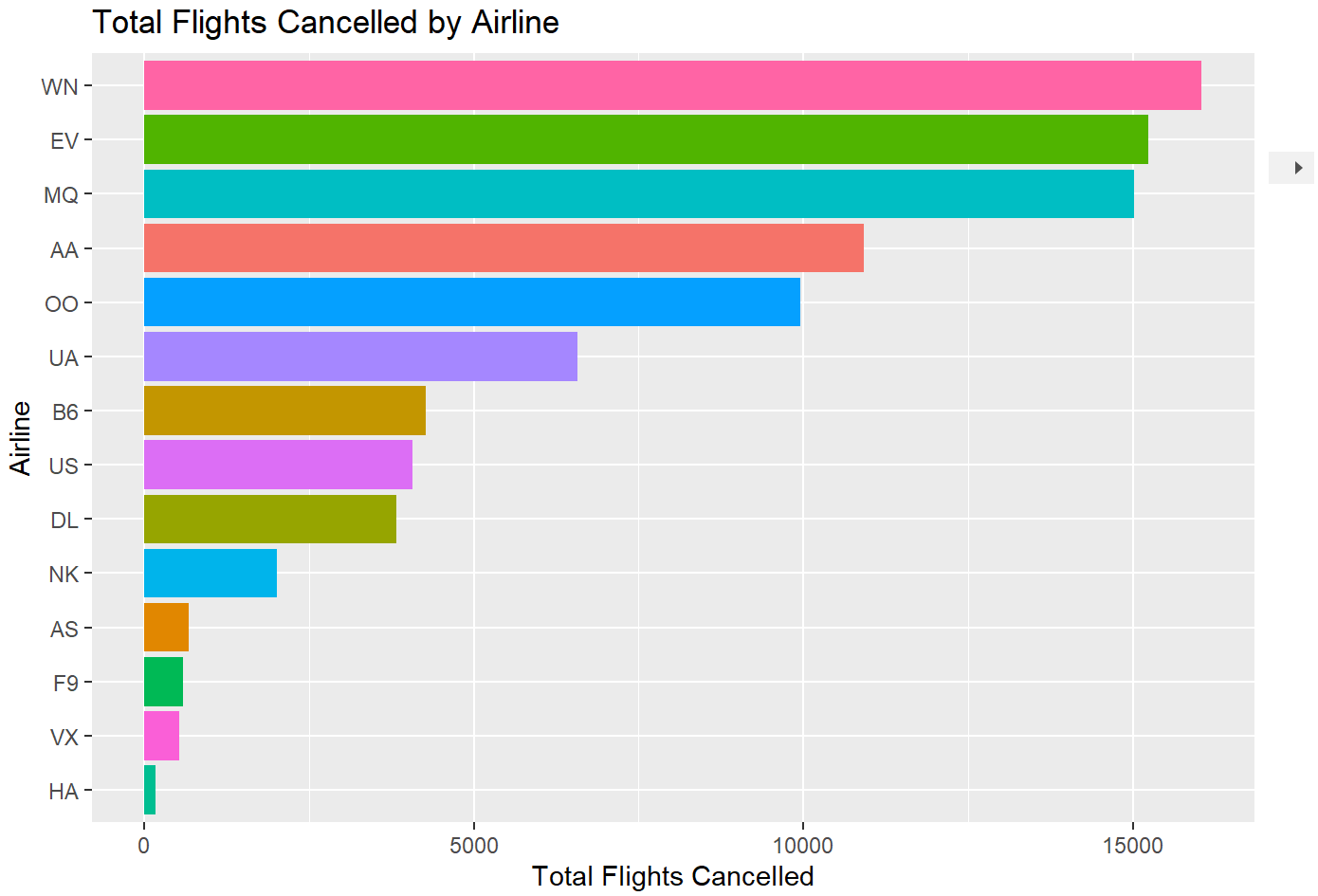
Flight Cancellation

Worse than having a flight delayed, is having a flight that is cancelled. Next we look into flight cancellation, more specifically, the total flights cancelled by each airline.

```
flights_cancelled_summary <- flights$flights %>%
  group_by(airline) %>% # grouping total count of delays by each airline
  summarize(total_cancelled = sum(cancelled, na.rm =TRUE)) # summing the count of delays and remo

flights_cancelled_summary %>%
```

```
ggplot(aes(x = reorder(airline, total_cancelled), y = total_cancelled, fill = airline)) + # reo
geom_col() +
coord_flip() + # flip x and y coordinates so airlines is on y
theme(legend.position = "none") + # removing legend
labs(title = "Total Flights Cancelled by Airline", x = "Airline", y = "Total Flights Cancelled"
```



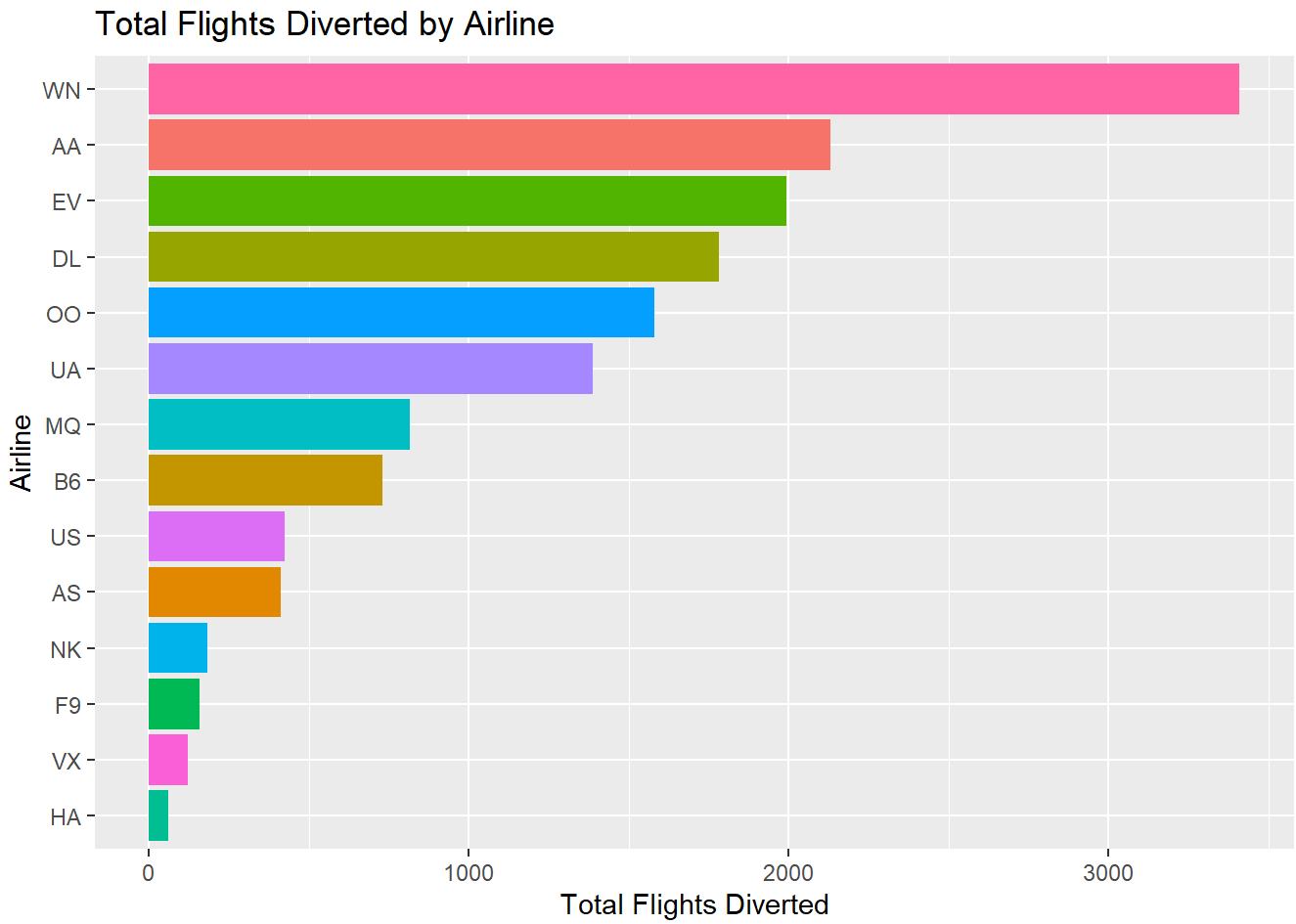
Up until now we have a few airlines that have been considered to be the best and worst for certain factors, so flight cancellation totals raise some interesting points. At the start, it was discovered that Southwest Airlines Co. had the most amount of flights, then it was seen how it also had the most flight distance and least maximum departure delay, however, this visualization shows it also has the most cancelled flights. Flights are cancelled for many reasons and is certainly a point of analysis that can be further explored, however, it might be possible that having the most amount of flights provides a greater chance of cancellation. This too is a point that can be further explored and lead to more details, for now, we'll note that Hawaiian Airlines Inc. had the least amount of flight cancelled, a clear winner when it comes to flight cancelled.

Flights Diverted

We analyzed cancellation of flights, however, not all flights are cancelled from the get go, sometimes flights are diverted and then potentially cancelled altogether. Let's take a look flights diverted by each airline and determine if there are any connections.

```
diverted_flights_summary <- flights$flights %>%
  group_by(airline) %>% # grouping by airline
  summarize(total_diverted = sum(diverted, na.rm = TRUE)) # new column with sum of diverted fligh

diverted_flights_summary %>%
  ggplot(aes(x = reorder(airline, total_diverted), y = total_diverted, fill = airline)) + # setti
  geom_col() +
  coord_flip() + # flipping coordinates to have Airline on y
  theme(legend.position = "none") + # removing legend
  labs(title = "Total Flights Diverted by Airline", x = "Airline", y = "Total Flights Diverted")
```



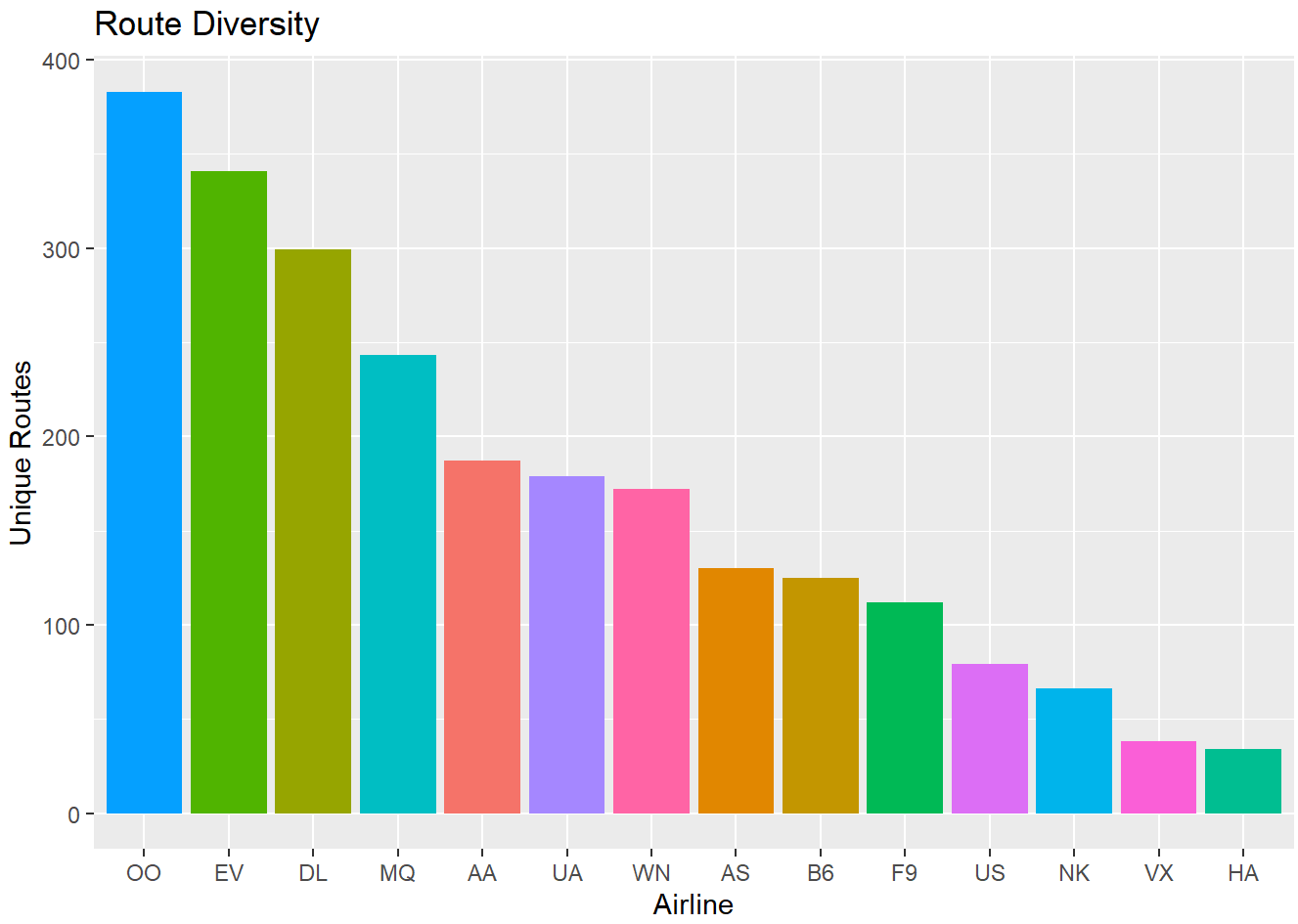
As expected, we can see similarities with this bar graph of total flights diverted and the total flights cancelled graph. Once again we have Southwest Airlines Co. in a negative spot, having the most flights diverted at 3,409 flights and Hawaiian Airlines Inc. with least amount of diverted flights at only 60 flights. It's important to note that Hawaiian Airlines Inc. also had the least amount of total flight travel distance, it's possible having less distance to travel and can minimize the chances of flights being diverted, but what factors would allow distance to affect diverting possibility? This is a point of analysis that can be further analyzed, given there are many factors causing flights to be diverted, such as weather, which can also be tied to a geographic region, including climate, more points that can lead to more factors.

Flight Geography

Speaking of regions, it's critical in our analysis of best and worst airlines to determine which airlines had the most diverse flights.

```
unique_route <- flights$flights %>%
  group_by(airline) %>% # grouping total distinct destinations by airline
  summarize(unique_route = n_distinct(destination_airport)) %>% # new column counting all distinc
  arrange(desc(unique_route)) # arranged in descending order

unique_route %>%
  ggplot(aes(x = reorder(airline, -unique_route), y = unique_route, fill = airline)) + # reorderi
  geom_col() +
  labs(title = "Route Diversity", x = "Airline", y = "Unique Routes") +
  theme(legend.position = "none") # removing legend
```



Let's first understand what this graph is showing, in determining route diversity, the approach taken was to consider the distinct destination airports each airline went to. Understandably, each new unique destination airport marks a different geographic location, and that is what this bar graph is demonstrating. We have a new airline entering the scene, Skywest Airlines Inc. had the most diverse routes, visiting 383 unique locations. Overall Skywest Airlines throughout this analysis has maintained itself in the top middle position when it comes to factors analyzed, such as flight volume, and on the lower end of the spectrum for factors such as flights cancelled. It can be safe to consider Skywest Airlines one of the more well-rounded airlines

then from the list of airlines analyzed, though, there are certainly airlines that were more notable on the factors analyzed such as Hawaiian Airlines and Southwest Airlines.

Hawaiian Airlines, A Closer Look

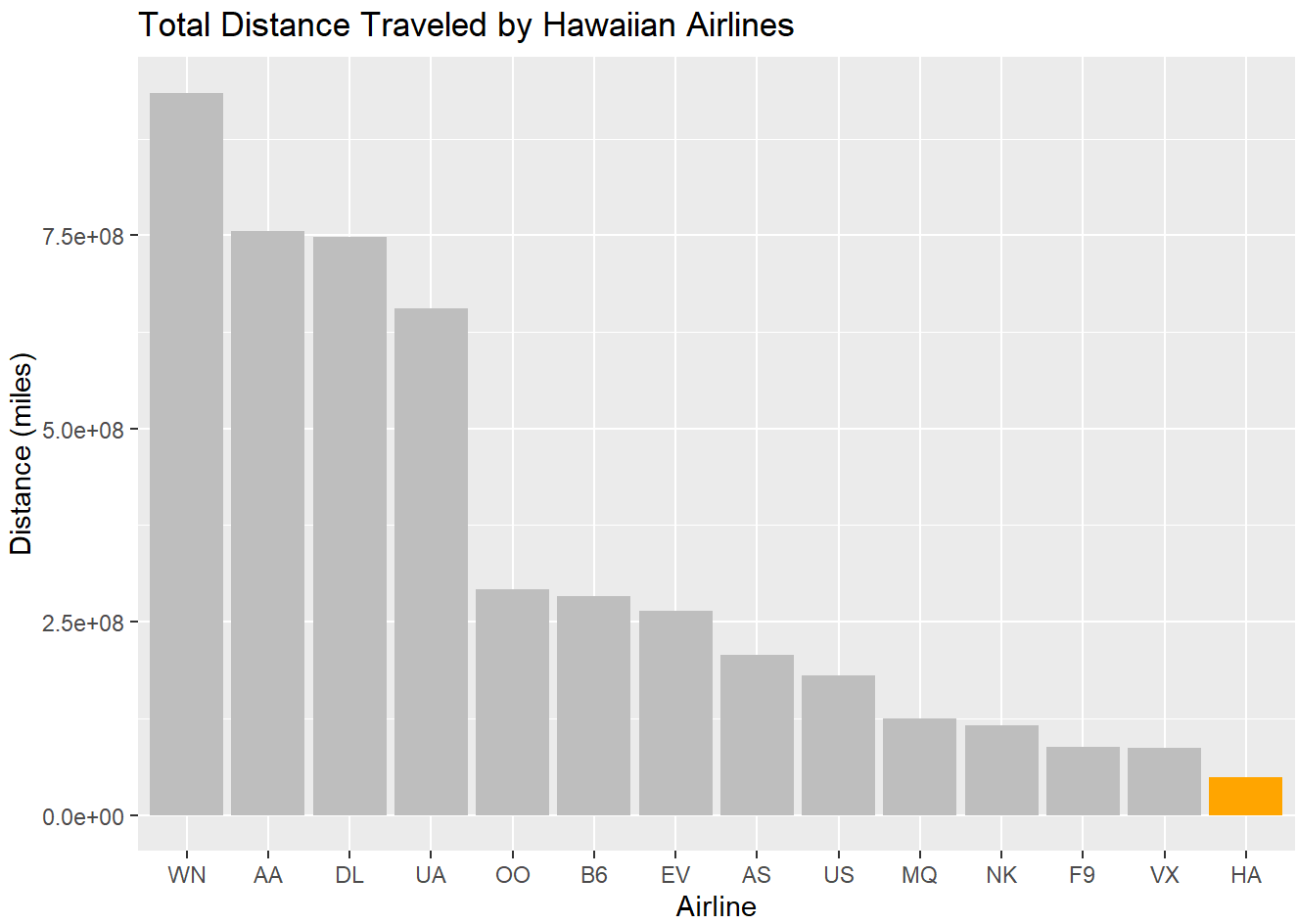
The following analysis will be done from a business analyst perspective when it comes to the performance and standing of Hawaiian Airlines. It's important to note this analysis will cover many of the same factors in the previous business journalist analysis, with an emphasis on how Hawaiian Airlines fairs against the other airlines.

HA Flight Travel Distance

```
total_distance <- flights$flights %>%
  group_by(airline) %>% # grouping by airline
  summarize(distance = sum(distance, na.rm = TRUE)) # creating new column storing sum of distance

ua_total_distance <- total_distance %>% # creating new variable for UA to distinct its values
  mutate(ToHighlight = ifelse(airline == "HA", "yes", "no")) # creating new column to assign 'yes

ua_total_distance %>%
  ggplot(aes(x = reorder(airline, -distance), y = distance, fill = ToHighlight)) + # fill is base
  geom_col() + # bar graph
  theme(legend.position = "none") + # removing legend
  labs(title = "Total Distance Traveled by Hawaiian Airlines", x = "Airline", y = "Distance (mile
  scale_fill_manual(values = c("yes" = "orange", "no" = "grey")) # only highlight airlines with '
```



When it comes to distance traveled by Hawaiian Airlines, a total of 48,249,045 miles were traveled. As can be seen in the graph, this puts the airline at the bottom of the spectrum of airlines. Compared to the other airlines, Hawaiian Airlines fairs poorly when it comes to travel distance, it is the airline with least traveled miles making it flight travel distance a major weakness.

HA Flights Cancelled

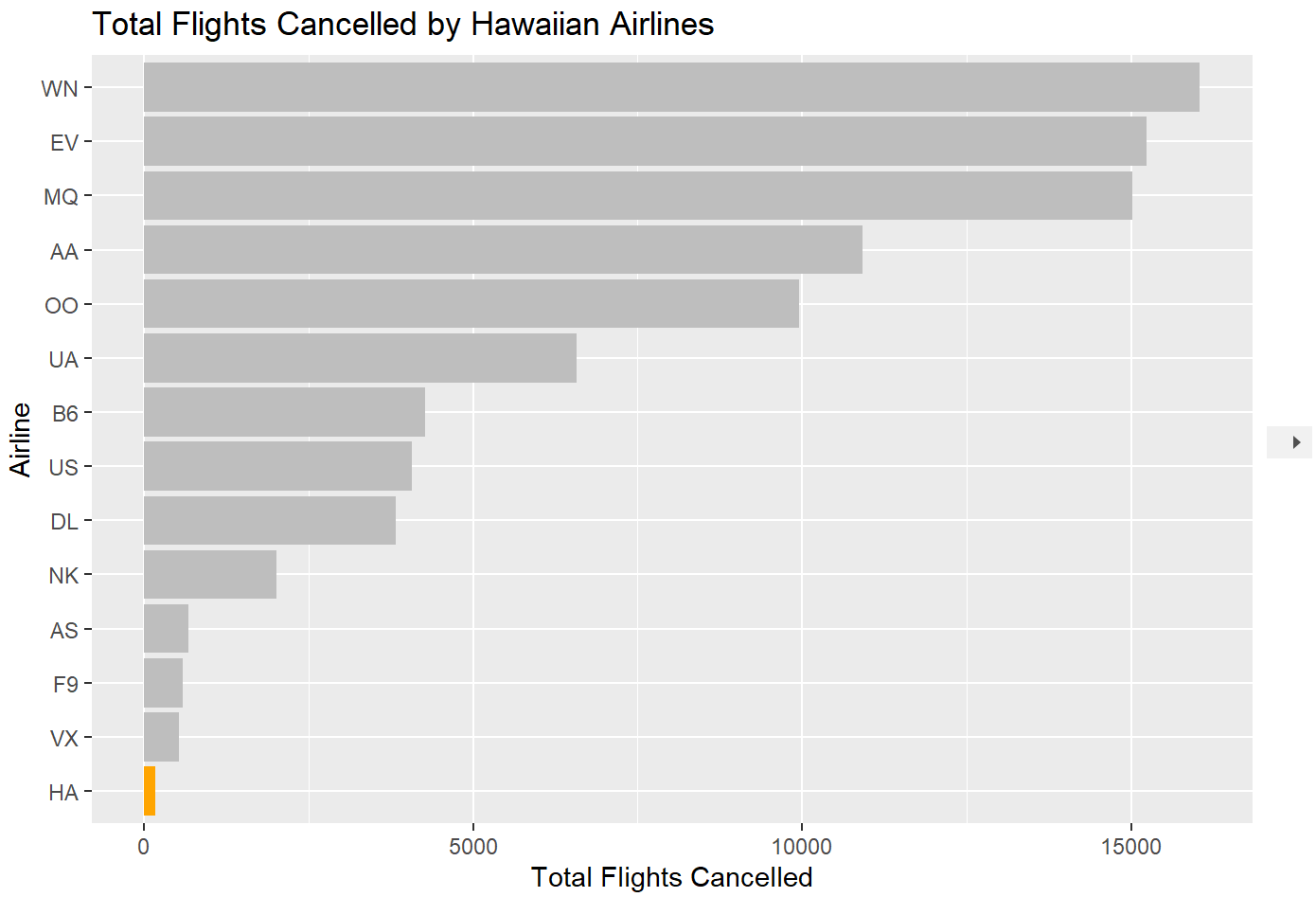
United Airlines is in a competitive range when it comes to traveled distance, let's see how that relates to the total cancelled flights.

```
flights_cancelled_summary <- flights$flights %>%
  group_by(airline) %>% # grouping total count of delays by each airline
  summarize(total_cancelled = sum(cancelled, na.rm =TRUE)) # summing the count of delays and remov

ua_flights_cancelled_summary <- flights_cancelled_summary %>% # creating new variable for UA to d
  mutate(ToHighlight = ifelse(airline == "HA", "yes", "no")) # creating new column to assign 'yes

ua_flights_cancelled_summary %>%
  ggplot(aes(x = reorder(airline, total_cancelled), y = total_cancelled, fill = ToHighlight)) + #
  geom_col() + # assigning bar chart
  coord_flip() + # flip x and y coordinates so airlines is on y
  labs(title = "Total Flights Cancelled by Hawaiian Airlines", x = "Airline", y = "Total Flights (
```

```
  theme(legend.position = "none") + # removing legend
  scale_fill_manual(values = c("yes" = "orange", "no" = "grey")) # only highlight airlines with '
```



With a total flight cancellation amount of 171, Hawaiian Airlines finds itself with the least number of cancelled flights compared to the other airlines. We can see flight cancellation being a strength for the airline when stacked against the rest of the airline. However, it is worth noting the airline had the least amount of traveled distance, a potential factor when it comes to flights cancelled.

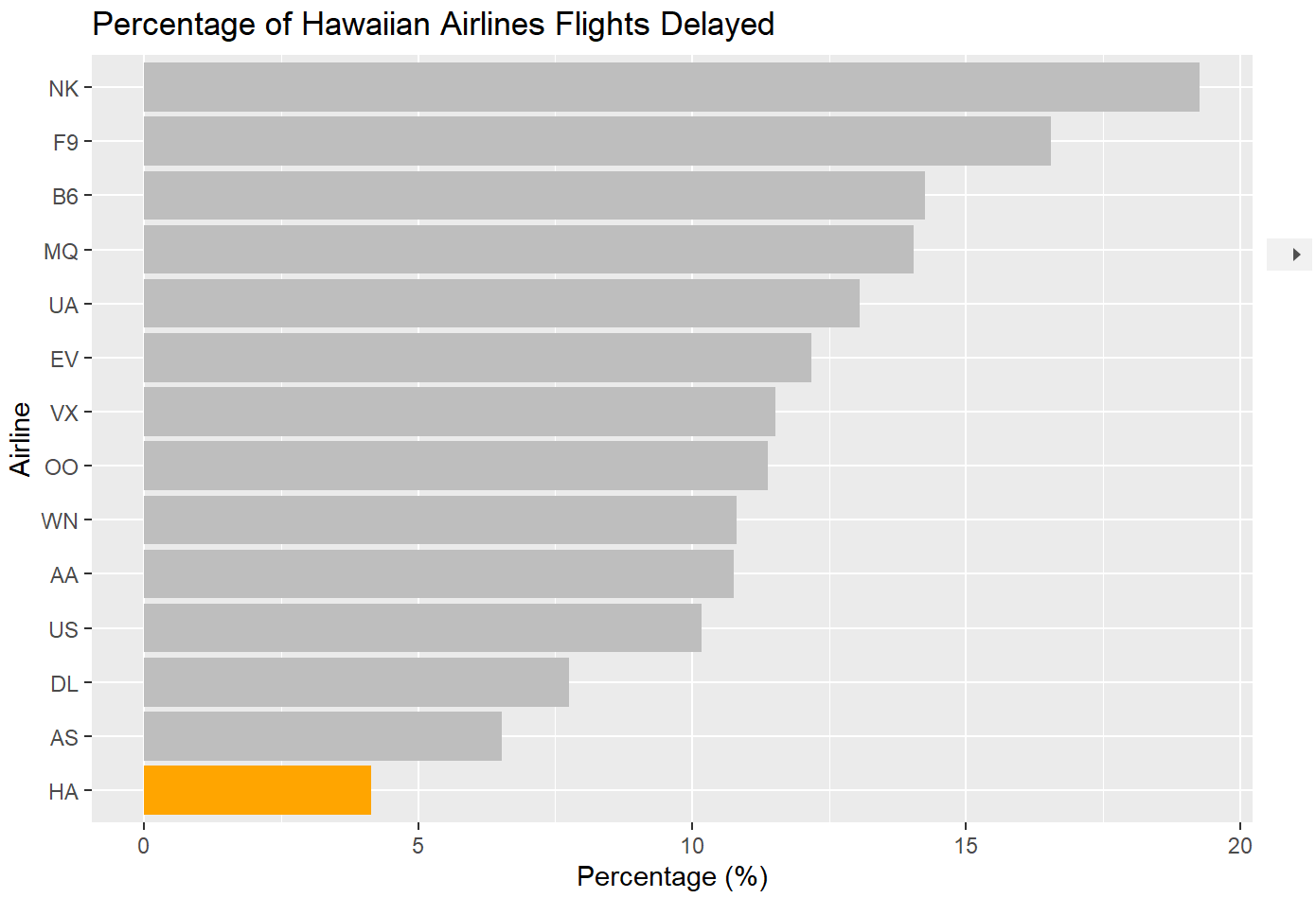
HA Flights Delayed

As mentioned before, flights cancelled and flights delayed go hand in hand in terms of one likely causing the others. Below we observe the percentage of United Airline flights cancelled, and how it compares with the other airlines.

```
percent_delayed <- flights$flights %>%
  mutate(delayed = as.numeric(delayed)) %>% # converting delayed column to be treated as numeric
  group_by(airline) %>% # grouping by airline
  summarise(percent_flights_delayed = mean(delayed, na.rm = TRUE) * 100) # creating new column st

ua_percent_delayed <- percent_delayed %>% # creating new variable for UA to distinct its values
  mutate(ToHighlight = ifelse(airline == "HA", "yes", "no")) # creating new column to assign 'yes
```

```
ua_percent_delayed %>%
  ggplot(aes(x = reorder(airline, percent_flights_delayed), y = percent_flights_delayed, fill = T
  geom_bar(stat = "identity") +
  coord_flip() +
  labs(title = "Percentage of Hawaiian Airlines Flights Delayed ", x = "Airline", y = "Percentage
  theme(legend.position = "none") +
  scale_fill_manual(values = c("yes" = "orange", "no" = "grey")) # only highlight airlines with '
```



4.142502% of Hawaiian Airlines' flights were cancelled, this puts the company at the lower end of the spectrum, with the lowest percentage of cancelled flights. Once again, it appears delay of flights is a strength for the company when compared to other airlines. So far, Hawaiian Airlines has maintained with notable strengths that are more likely to influence customer satisfaction.

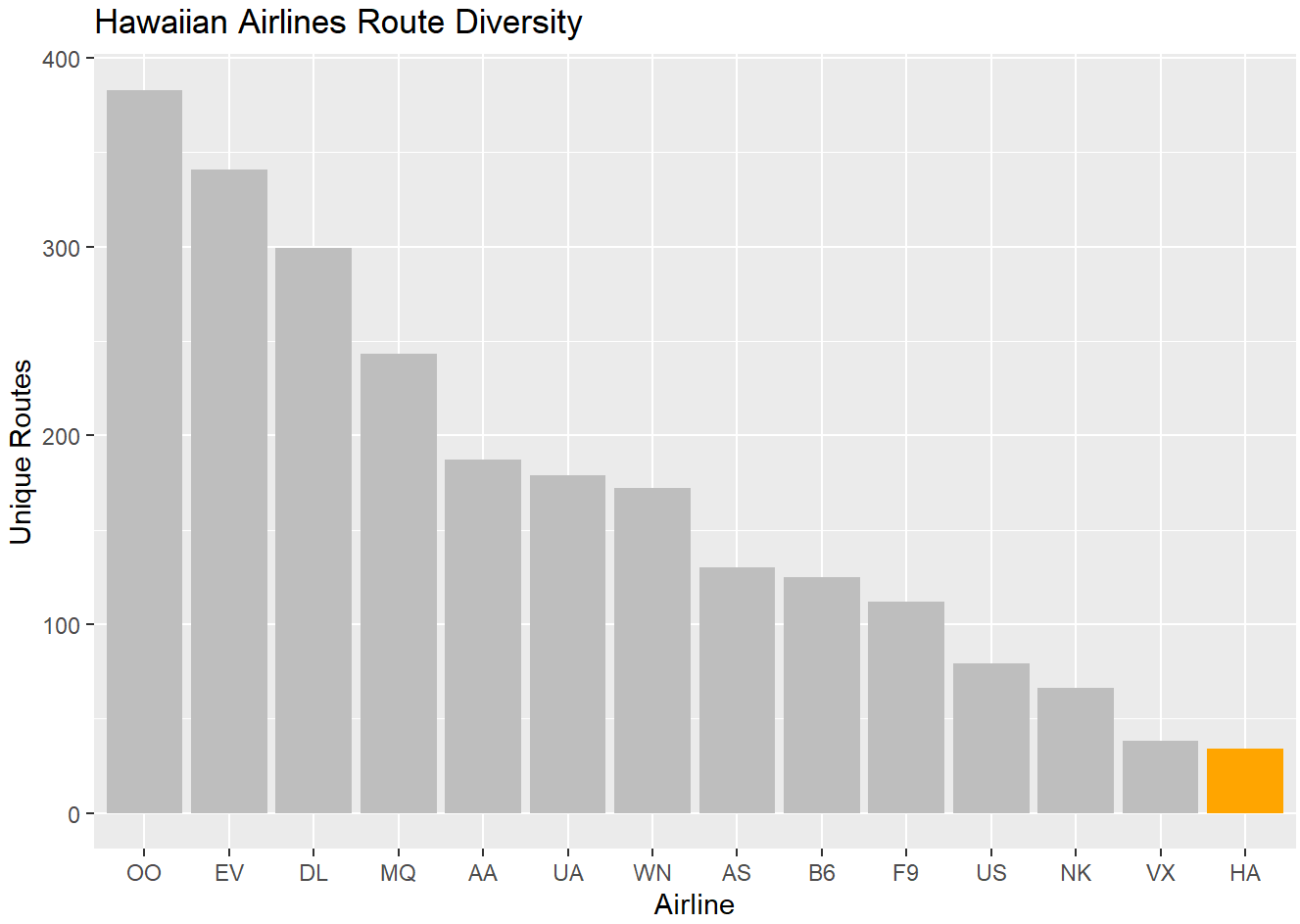
HA Flight Geography

Flight diversity is critical when it comes to customer satisfaction, more airports visited allows the airline to appeal to a broader audience and more likely to have an available route for any given customer. Let's take a look at Hawaiian Airline's route diversity.

```
unique_route <- flights$flights %>%
  group_by(airline) %>% # grouping total distinct destinations by airline
  summarize(unique_route = n_distinct(destination_airport)) %>% # counting all distinct destinati
  arrange(desc(unique_route)) # arranged in descending order

ua_unique_route <- unique_route %>% # creating new variable for UA to distinct its values
  mutate(ToHighlight = ifelse(airline == "HA", "yes", "no")) # creating new column to assign 'yes

ua_unique_route %>%
  ggplot(aes(x = reorder(airline, -unique_route), y = unique_route, fill = ToHighlight)) + # reor
  geom_col() +
  labs(title = "Hawaiian Airlines Route Diversity", x = "Airline", y = "Unique Routes") +
  theme(legend.position = "none") +
  scale_fill_manual(values = c("yes" = "orange", "no" = "grey")) # only highlight airlines with '
```



Here we can see the airline has a rather unique factor, it has the least amount of visited unique locations, totaling 34 locations. This is can be considered to be Hawaiian Airlines' biggest weakness, there's simply not enough locations to allow the company to compete with the rest of the airlines, and thus, a smaller population of customers. Customer satisfaction will depend on factors like this, however, strengths such as low flight cancellations and delays allow the company to have notable traits keeping it in the air.