ECO 395 Homework 1:

Annie Nguyen

2023-01-24

1) Data Visualization: Flights at ABIA

Introduction

In this section, we are looking at a 2008 dataset that contains information on every commercial flight coming in and out of the Austin-Bergstrom Internationial Airport. We are interested in seeing which cities in the dataset have the highest average departure delays.

Methods

To explore the data and create visualizations, we first need to import the `tidyverse` and `ggplot2` packages and read the `ABIA.csv` file which contains the data. Since we are also interested in visualizing the data on a geographical map, we will also import the `ggmap` package and read the `airport-codes` files, which contain the latitude and longitude reading of each airport.

```
library(ggplot2)
library(tidyverse)
library(ggmap)
abia <- read.csv("ABIA.csv")
airport_codes <- read.csv("airport-codes.csv")
```

First, the `airport_codes` dataframe is cleaned to exclude closed airports and drop rows with NA's or no available IATA-codes. Then, the `airport_codes` dataframe is joined with `abia` dataframe, so each US airport in our targeted dataframe has a respective value for longitude and latitude.

Then, we calculate the average departure delay from each Origin airport. Airports with the highest departure delays include: Knoxville, TN; Birmingham, AL; Washington, D.C.; Raleigh/Durham, NC; San Antonio, TX; Philadelphia, PA.

```
airport_codes <- airport_codes %>%
  filter(type != "closed", iata_code != "") %>%
  separate(col = coordinates, into= c("lat", "long"), sep = ",") %>%
  transform(lat = as.numeric(lat), long = as.numeric(long)) %>%
  select(iata_code, lat, long) %>% drop_na %>% distinct

avg_delay <- abia %>% group_by(Origin) %>%
  summarise(avg_delay_time = mean(DepDelay, na.rm = TRUE)) %>%
  arrange(-avg_delay_time) %>%
  left_join(airport_codes, by= c("Origin" = "iata_code"))

avg_delay %>% select(Origin, avg_delay_time) %>% head %>%
  knitr::kable(caption = "Top 6 Cities with Highest Departure Delays")
```

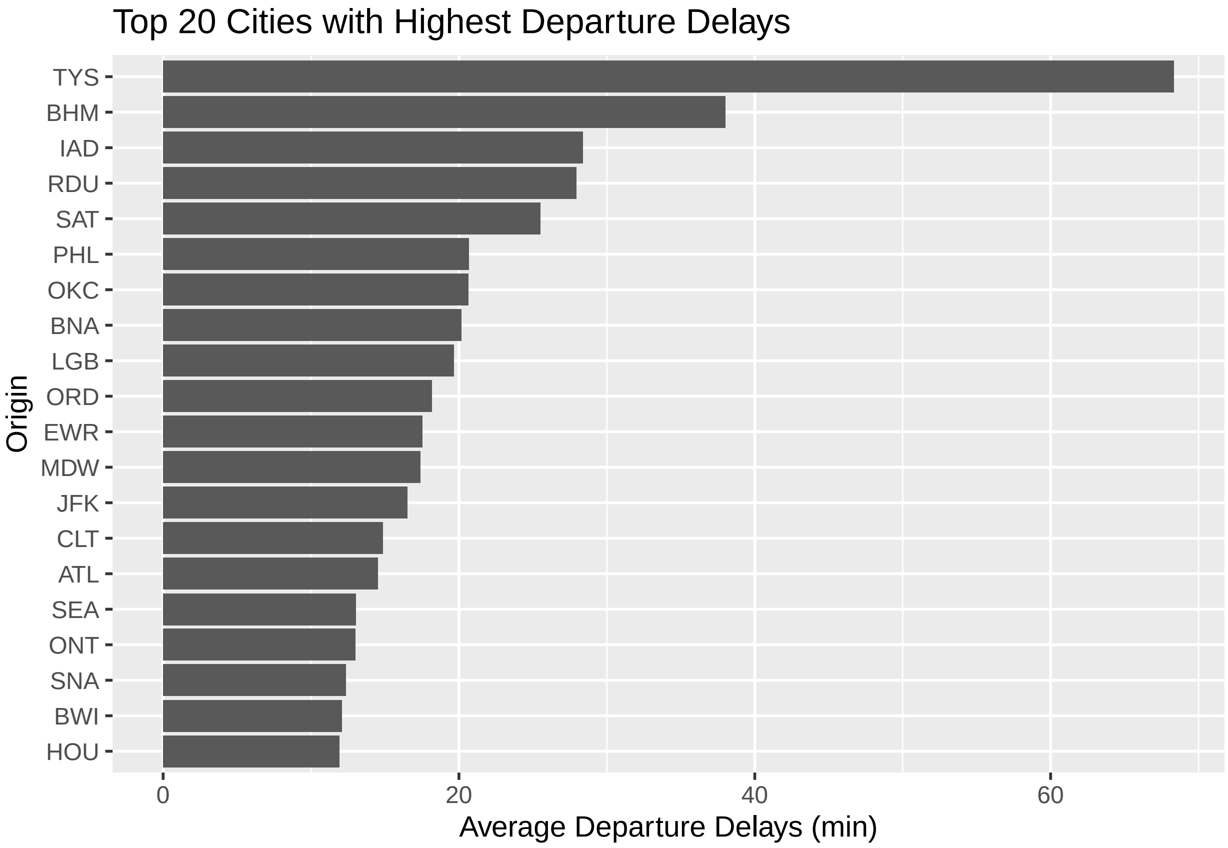
Table 1: Top 6 Cities with Highest Departure Delays

Origin	avg_delay_time
TYS	68.33333
BHM	38.00000
IAD	28.39550
RDU	27.93450
SAT	25.50000
PHL	20.67241

We will use a barplot to visualize a larger portion of the dataset, looking at the top 20 cities with the highest departure delays.

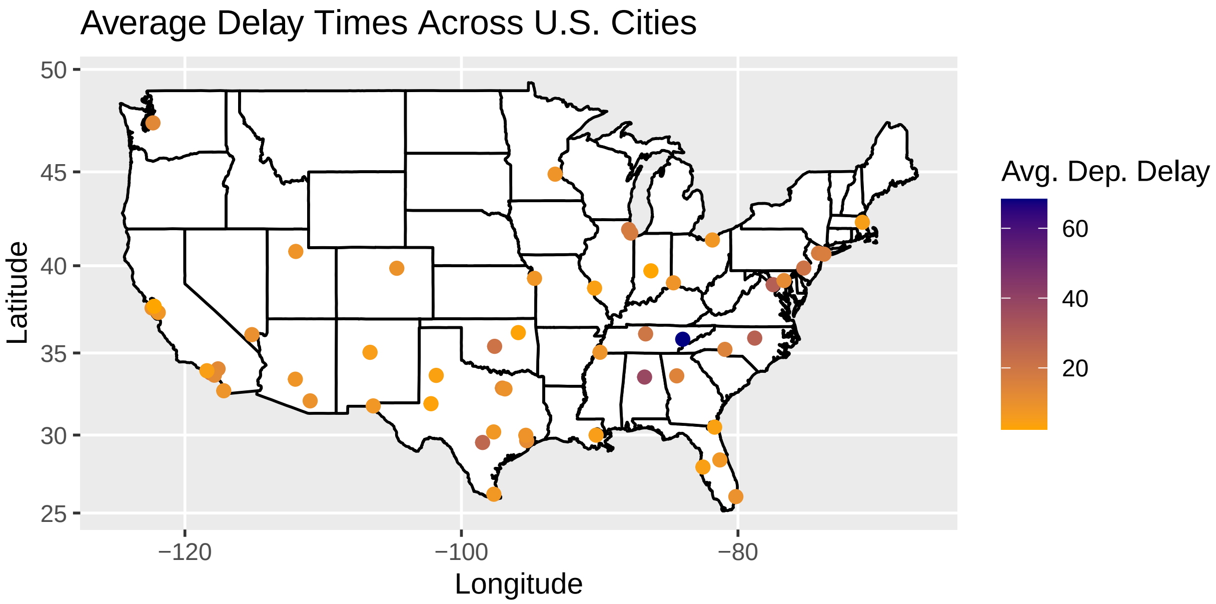
From the plot, we see that Knoxville has, on average, significantly higher delays than Birmingham (with departure delays close to 70 minutes for Knoxville and near 40 minutes for Birmingham). In comparison, the average delays for the remaining cities stay below 30 minutes.

```
avg_delay %>% head(20) %>%
  ggplot(aes(reorder(Origin, avg_delay_time), avg_delay_time)) +
  geom_bar(stat = "identity") +   coord_flip() +
  ggtitle("Top 20 Cities with Highest Departure Delays") +
  xlab("Origin") + ylab("Average Departure Delays (min)")
```



We will visualize the data for all 53 cities included on the dataset by plotting the points on the US map. This will not only show us average airport delays but also give us a brief look at all of the US cities that flyers can travel to from Austin-Bergstrom Internationial Airport.

```
us_map <- map_data(map = "state")
avg_delay %>% ggplot() +
  geom_polygon(aes(long, lat, group = group), fill= "white", color="black", data= us_map) +
  coord_map() +
  geom_point(aes(long, lat, color = avg_delay_time), size = 2) +
  ggtitle("Average Delay Times Across U.S. Cities")+
    scale_color_gradient(low = "orange", high = "navy", name = "Avg. Dep. Delay") +
  xlab("Longitude") + ylab("Latitude")
```



Results and Conclusion

Based on the results of the plots, we see that Knoxville and Birmingham experience significantly higher departure delays compared to other cities.

From the data visualization generated in this section, we were able to identify airports that experienced the highest average departure delays. This information can help future flyers know what to expect in terms of departure delays when traveling to Austin.

Of course, this brief data exploration and visualization report comes with its own set of limitations. Departure delays are influenced by holiday rushes and weather conditions, so outliers from unusual weather or from other potential problems may skew results. Furthermore, the dataset only includes observations from 2008 and delays may affect each airport differently on a yearly basis.

2) Wrangling the Olympics

```
olympics <- read.csv('olympics_top20.csv')
```

3) K-Nearest Neighbors: Cars

```
sclass <- read.csv('sclass.csv')
```