Flight Data Analysis

Frederick Jones

October 1, 2023.

Contents

Load the required libraries

```
library(tidyr)
library(dplyr)
library(ggplot2)
library(flextable)
library(rstatix)
```

GitHub URL for the CSV file

```
url <- "https://raw.githubusercontent.com/jewelercart/Data607/main/flight_data.csv"
```

Read the CSV file into R, specifying header and line termination

```
flight_data <- read.csv(url, header = TRUE, sep = ",", quote = "", fill = TRUE)
```

Specify the column names explicitly

```
colnames(flight_data) <- c("Time_Zone", "Cities", "on_time", "delayed")
```

Tidy the data (convert to long format)

```
flight_data_long <- flight_data %>%
  gather(key = "Status", value = "Count", -Time_Zone, -Cities)
```

Removing the quotes from the names of Time Zones and Cities.

```
flight_data_long$Time_Zone<- gsub("\"", "", flight_data_long$Time_Zone)
flight_data_long$Cities<- gsub("\"", "", flight_data_long$Cities)
table <- knitr::kable(flight_data_long)
table
```

Time_Zone	Cities	Status	Count
ALASKA	Los Angeles	on_time	497
ALASKA	Phoenix	on_time	221
ALASKA	San Diego	on_time	212
ALASKA	San Francisco	on_time	503
ALASKA	Seattle	on_time	1841
AM WEST	Los Angeles	on_time	694
AM WEST	Phoenix	on_time	4840
AM WEST	San Diego	on_time	383
AM WEST	San Francisco	on_time	320
AM WEST	Seattle	on_time	201
ALASKA	Los Angeles	delayed	62
ALASKA	Phoenix	delayed	12
ALASKA	San Diego	delayed	20
ALASKA	San Francisco	delayed	102
ALASKA	Seattle	delayed	305

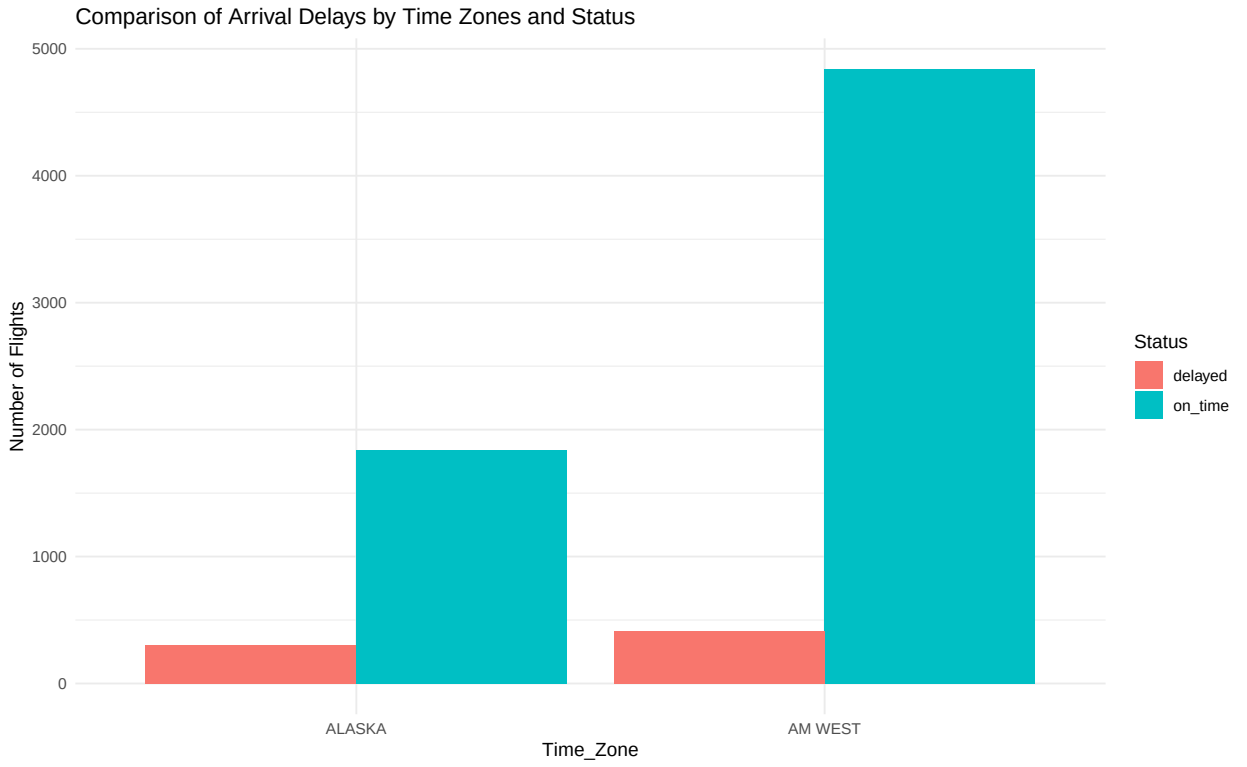
Time_Zone	Cities	Status	Count
AM WEST	Los Angeles	delayed	117
AM WEST	Phoenix	delayed	415
AM WEST	San Diego	delayed	65
AM WEST	San Francisco	delayed	129
AM WEST	Seattle	delayed	61

Calculate summary statistics, handling missing values

```
summary_stats =
flight_data_long  %>%
  group_by(Time_Zone, Status) %>%
   get_summary_stats(Count, show = c("mean", "median", "max","min"))
```

Create bar plots to compare arrival delays

```
ggplot(flight_data_long, aes(x = Time_Zone, y = Count, fill = Status)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Comparison of Arrival Delays by Time Zones and Status", y = "Number of Flights") +
  theme_minimal()
```

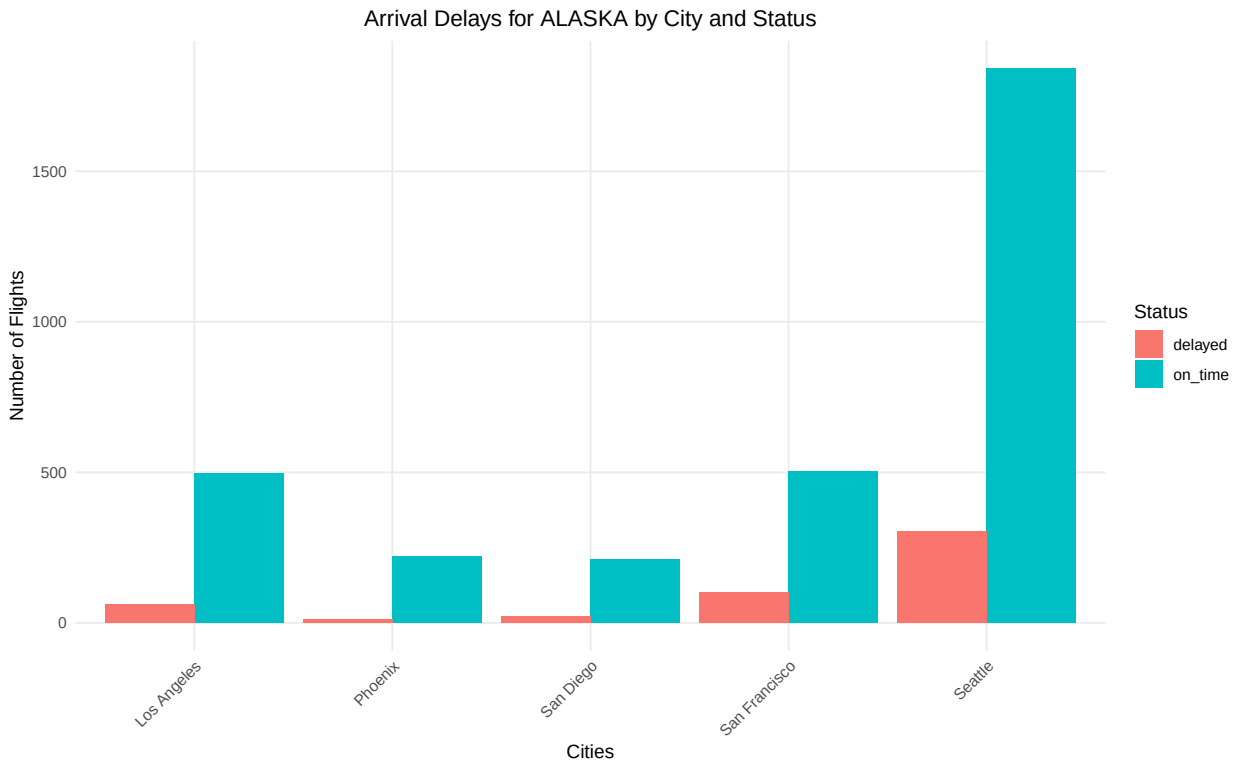


Filter the data for ALASKA and AM WEST separately:

```
alaska_data <- flight_data_long %>% filter(Time_Zone == "ALASKA")
am_west_data <- flight_data_long %>% filter(Time_Zone == "AM WEST")
```

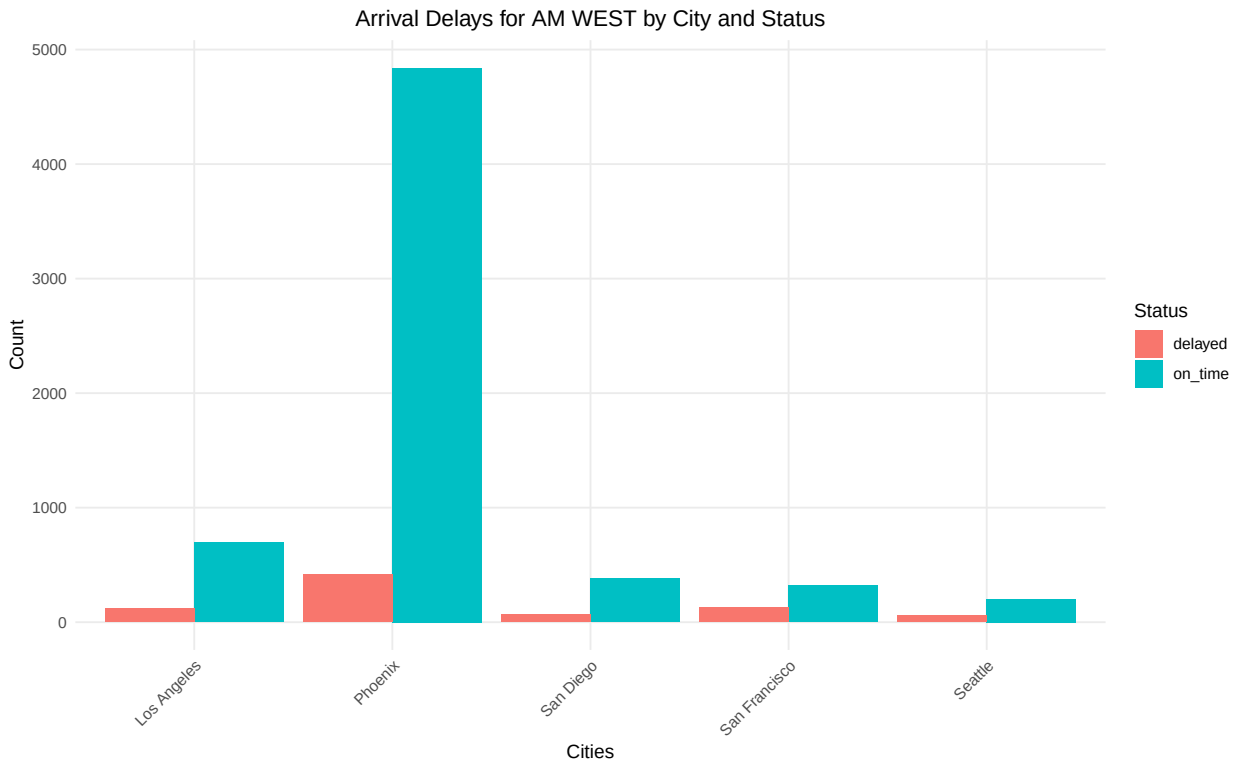
Create bar plots for ALASKA and AM WEST:

```
plot_alaska <- ggplot(alaska_data, aes(x = Cities, y = Count, fill = Status)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Arrival Delays for ALASKA by City and Status", y = "Number of Flights") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        plot.title = element_text(hjust=0.5),
        panel.grid.minor= element_blank())

plot_alaska
```




```
plot_am_west <- ggplot(am_west_data, aes(x = Cities, y = Count, fill = Status)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Arrival Delays for AM WEST by City and Status", y = "Count") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        plot.title = element_text(hjust=0.5),
```

```
        panel.grid.minor= element_blank())

plot_am_west
```



Summary Statistics:

```
summary_stats
```

```
## # A tibble: 4 x 8
##   Time_Zone Status  variable     n  mean median   max   min
##   <chr>     <chr>   <fct>    <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1 ALASKA    delayed Count        5  100.     62   305    12
## 2 ALASKA    on_time Count        5  655.    497  1841   212
## 3 AM WEST   delayed Count        5  157.    117   415    61
## 4 AM WEST   on_time Count        5 1288.    383  4840   201
```

A Chi-square test will compare whether the delay difference of a flight varies by comparing two time zones.

```
tbl <- xtabs(~ Time_Zone + Status, data = flight_data_long)
summary(tbl)
```

```
## Call: xtabs(formula = ~Time_Zone + Status, data = flight_data_long)
## Number of cases in table: 20
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 0, df = 1, p-value = 1
```

```
proportions(tbl, "Status")
```

```
##          Status
## Time_Zone delayed on_time
##   ALASKA      0.5     0.5
##   AM WEST     0.5     0.5
```

Conclusion

There is no difference between the two time zones regarding flight delays and punctuality. This was confirmed with a chi-squared test p-value = 1.