Data 607 lab 5

Yina Qiao

2023-02-26

Introduction

This assignment will import uncleaned data from a csv. file. My task is to tidy and transform data as described below. (1) Read the information from a .CSV file into R, and use tidyr and dplyr as needed to tidy and transform the data. (2) Perform analysis to compare the arrival delays for the two airlines

Install and load pakages

```
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──
## ✔ ggplot2 3.4.1     ✔ purrr   0.3.4
## ✔ tibble  3.1.7     ✔ dplyr   1.0.9
## ✔ tidyr   1.2.0     ✔ stringr 1.4.0
## ✔ readr   2.1.2     ✔ forcats 0.5.1
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

Load dataset

```
url = r"(https://raw.githubusercontent.com/yinaS1234/data-607/main/data%20607%20lab%205/lab5data.csv)"
suppressMessages(
df <- read_csv(url, skip_empty_rows = TRUE, show_col_types = FALSE)[-3,]
)
names(df)[1:2] <- c("Airline", "Status")
df
```

```
## # A tibble: 4 × 7
##   Airline Status  `Los Angeles` Phoenix `San Diego` `San Francisco` Seattle
##   <chr>   <chr>           <dbl>   <dbl>       <dbl>           <dbl>   <dbl>
## 1 ALASKA  on time           497     221         212             503    1841
## 2 <NA>    delayed           62      12          20              102     305
## 3 AM WEST on time           694    4840         383             320     201
## 4 <NA>    delayed           117     415          65             129      61
```

Clean and tidy the data

fixing NA transform data, rename

```
for(x in seq(from=2, to=nrow(df), by=2))
{
  df[x, 1] = df[x-1, 1]
}
```

```
library(tidyr)

df_unpivot = df %>%
  gather(key="City", value="Count", c("Los Angeles","Phoenix", "San Diego", "San Francisco", "Seattle"))
df_unpivot
```

```
## # A tibble: 20 × 4
##    Airline Status  City          Count
##    <chr>   <chr>   <chr>         <dbl>
##  1 ALASKA  on time Los Angeles     497
##  2 ALASKA  delayed Los Angeles      62
##  3 AM WEST on time Los Angeles     694
##  4 AM WEST delayed Los Angeles     117
##  5 ALASKA  on time Phoenix         221
##  6 ALASKA  delayed Phoenix          12
##  7 AM WEST on time Phoenix        4840
##  8 AM WEST delayed Phoenix         415
##  9 ALASKA  on time San Diego       212
## 10 ALASKA  delayed San Diego        20
## 11 AM WEST on time San Diego       383
## 12 AM WEST delayed San Diego        65
## 13 ALASKA  on time San Francisco   503
## 14 ALASKA  delayed San Francisco   102
## 15 AM WEST on time San Francisco   320
## 16 AM WEST delayed San Francisco   129
## 17 ALASKA  on time Seattle        1841
## 18 ALASKA  delayed Seattle         305
## 19 AM WEST on time Seattle         201
## 20 AM WEST delayed Seattle          61
```

```
library(dplyr)

df_unpivot_2 = df_unpivot %>%
  spread(key="Status", value="Count")

df_unpivot_2 = df_unpivot_2 %>%
  rename(on_time = `on time`)
```

Data Analysis

On Time Percentage By City by airline

```
df_unpivot_2 = df_unpivot_2 %>%
  mutate(otp = on_time/(on_time + delayed)) %>%
  arrange(desc(otp))

df_unpivot_2
```

```
## # A tibble: 10 × 5
##    Airline City          delayed on_time   otp
##    <chr>   <chr>           <dbl>   <dbl> <dbl>
##  1 ALASKA  Phoenix            12     221 0.948
##  2 AM WEST Phoenix           415    4840 0.921
##  3 ALASKA  San Diego          20     212 0.914
##  4 ALASKA  Los Angeles        62     497 0.889
##  5 ALASKA  Seattle           305    1841 0.858
##  6 AM WEST Los Angeles       117     694 0.856
##  7 AM WEST San Diego          65     383 0.855
##  8 ALASKA  San Francisco     102     503 0.831
##  9 AM WEST Seattle            61     201 0.767
## 10 AM WEST San Francisco     129     320 0.713
```

Overall On Time Percentage By Airline.

- we see that AM West Outperform ALASKA overall. However,could the result skewed due to more flights in certain city?

```
df_unpivot_2 %>%
  select(Airline, delayed, on_time) %>%
  group_by(Airline) %>%
  summarise(delayed = sum(delayed), on_time = sum(on_time), otp = sum(on_time) / (sum(on_time) + sum(delayed))) %
>%
  arrange(desc(otp))
```

```
## # A tibble: 2 × 4
##   Airline delayed on_time   otp
##   <chr>     <dbl>   <dbl> <dbl>
## 1 AM WEST     787    6438 0.891
## 2 ALASKA      501    3274 0.867
```

- it seems like AM WEST has more flights in Phoenix, so lets compute the OTP without phoenix.*

```
df_unpivot_2 %>%
  filter(City != "Phoenix") %>%
  select(Airline, delayed, on_time) %>%
  group_by(Airline) %>%
  summarise(delay = sum(delayed), on_time = sum(on_time), otp = sum(on_time) / (sum(on_time) + sum(delayed))) %>%
  arrange(desc(otp))
```

```
## # A tibble: 2 × 4
##   Airline delay on_time   otp
##   <chr>   <dbl>   <dbl> <dbl>
## 1 ALASKA    489    3053 0.862
## 2 AM WEST   372    1598 0.811
```

Conclusion

Based on the data provided,ALASKA consistently outperforms AM WEST. For flying to Phoenix specifically, I would recommend AM WEST as the on time performance is on par with ALASKA and have more flights available.