DATA607 Assignment 5

Warner Alexis

2023-10-04

Introduction

We were given an data set of Arrival Airport delays between some of the big cities. The data set has a wide structure that needs some manipulation to transform it into long form. We load the csv file then rename the empty column name. We deleted all the empty row and reprocess the data inton a long structure.

```
# read data csv file
raw <- read.csv("https://raw.githubusercontent.com/joewarner89/CUNY-607/main/homeworks/Assignment%205/a
# Delete all Empty rows
raw <- raw %>% na.omit(raw)

raw <- as.data.frame(raw)

# rename row

raw <- raw %>% rename(Airlines = 1, Arrival_Status = 2)
head(raw)
```

```
##   Airlines Arrival_Status Los.Angeles Phoenix San.Diego San.Francisco Seattle
## 1   Alaska        on time         497     221       212           503    1841
## 2                 delayed          62      12        20           102     305
## 4                 on time         694    4840       383           320     201
## 5  AM WEST        delayed         117     415        65           129      61
```

Data Analysis

The layout of this data set require data manipulation to transform this data set in the long form.

```
## 'data.frame':    20 obs. of  4 variables:
##  $ Airlines      : chr  "Alaska" "" "" "AM WEST" ...
##  $ Arrival_Status: chr  "on time" "delayed" "on time" "delayed" ...
##  $ City          : chr  "Los.Angeles" "Los.Angeles" "Los.Angeles" "Los.Angeles" ...
##  $ Arrival_Delays: int  497 62 694 117 221 12 4840 415 212 20 ...
```

```
##   Airlines Arrival_Status        City Arrival_Delays
## 1   Alaska        on time Los.Angeles            497
## 2                 delayed Los.Angeles             62
## 3                 on time Los.Angeles            694
## 4  AM WEST        delayed Los.Angeles            117
## 5   Alaska        on time     Phoenix            221
## 6                 delayed     Phoenix             12
```

We are going to replace the dot(.) in the City Column with a space.

```
airline_data$City <- str_replace(airline_data$City, "\\.", " ")
head(airline_data)
```

```
##   Airlines Arrival_Status        City Arrival_Delays
## 1   Alaska        on time Los Angeles            497
## 2                 delayed Los Angeles             62
## 3                 on time Los Angeles            694
## 4  AM WEST        delayed Los Angeles            117
## 5   Alaska        on time     Phoenix            221
## 6                 delayed     Phoenix             12
```

We will fill the value of NA with the most recent value that above the empty strings.

```
final <- airline_data %>% mutate(Airlines = as.character(na_if(Airlines,""))) %>%  fill(Airlines,.direc
head(final)
```

```
##   Airlines Arrival_Status        City Arrival_Delays
## 1   Alaska        on time Los Angeles            497
## 2   Alaska        delayed Los Angeles             62
## 3   Alaska        on time Los Angeles            694
## 4  AM WEST        delayed Los Angeles            117
## 5   Alaska        on time     Phoenix            221
## 6   Alaska        delayed     Phoenix             12
```

```
head(final)
```

```
##   Airlines Arrival_Status        City Arrival_Delays
## 1   Alaska        on time Los Angeles            497
## 2   Alaska        delayed Los Angeles             62
## 3   Alaska        on time Los Angeles            694
## 4  AM WEST        delayed Los Angeles            117
## 5   Alaska        on time     Phoenix            221
## 6   Alaska        delayed     Phoenix             12
```

```
final %>% group_by(Airlines,City,Arrival_Status) %>% summarise(Delay_total = sum(Arrival_Delays))
```

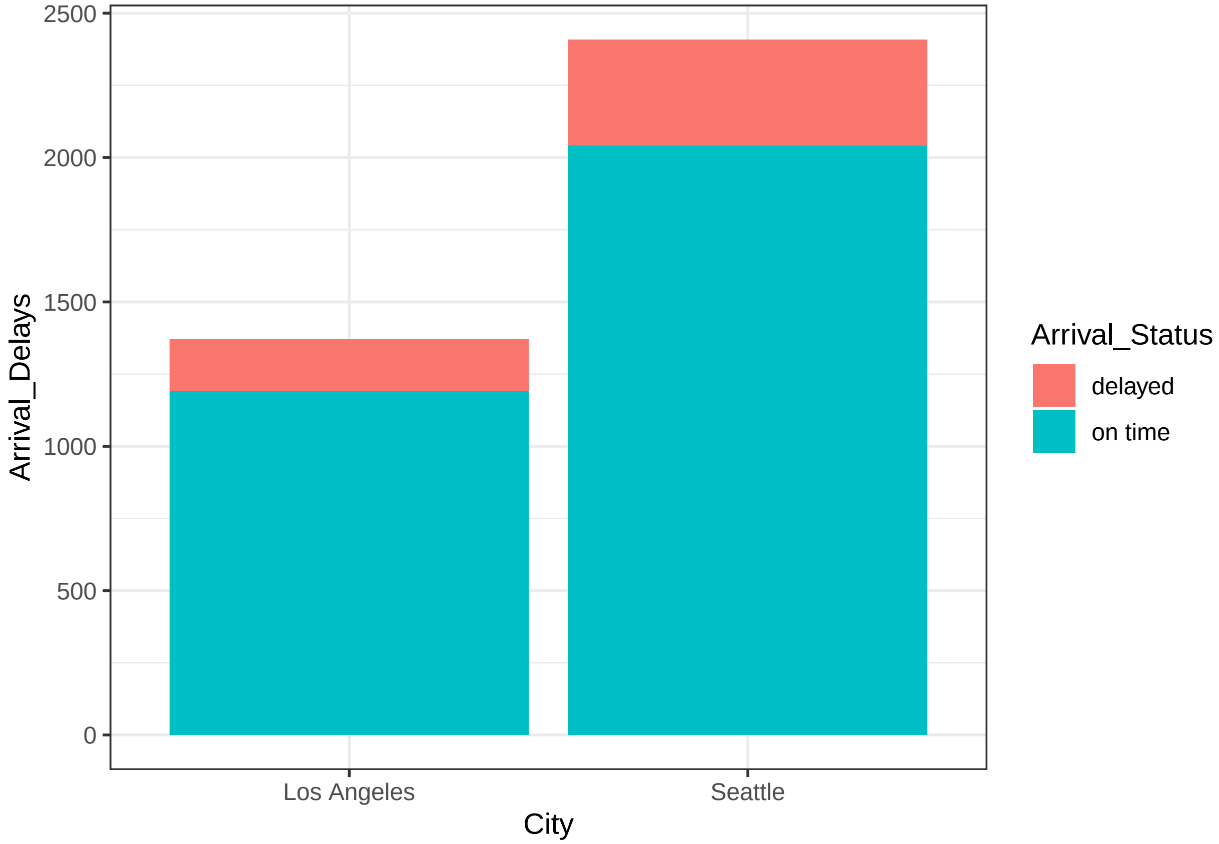
```
## `summarise()` has grouped output by 'Airlines', 'City'. You can override using
## the `.groups` argument.
```

```
## # A tibble: 15 x 4
## # Groups:   Airlines, City [10]
##    Airlines City          Arrival_Status Delay_total
##    <chr>    <chr>         <chr>                <int>
##  1 AM WEST  Los Angeles   delayed                117
##  2 AM WEST  Phoenix       delayed                415
##  3 AM WEST  San Diego     delayed                 65
##  4 AM WEST  San Francisco delayed                129
##  5 AM WEST  Seattle       delayed                 61
##  6 Alaska   Los Angeles   delayed                 62
##  7 Alaska   Los Angeles   on time               1191
##  8 Alaska   Phoenix       delayed                 12
##  9 Alaska   Phoenix       on time               5061
## 10 Alaska   San Diego     delayed                 20
## 11 Alaska   San Diego     on time                595
## 12 Alaska   San Francisco delayed                102
## 13 Alaska   San Francisco on time                823
## 14 Alaska   Seattle       delayed                305
## 15 Alaska   Seattle       on time               2042
```

```
air <- final %>% filter(City %in% c("Los Angeles","Seattle") )
# Only 2 Airport
air <- final %>% filter(City %in% c("Los Angeles","Seattle") )
head(air)
```

```
##   Airlines Arrival_Status        City Arrival_Delays
## 1   Alaska        on time Los Angeles            497
## 2   Alaska        delayed Los Angeles             62
## 3   Alaska        on time Los Angeles            694
## 4  AM WEST        delayed Los Angeles            117
## 5   Alaska        on time     Seattle           1841
## 6   Alaska        delayed     Seattle            305
```

```
ggplot(air, aes(x=City, y=Arrival_Delays, fill=Arrival_Status)) +
  geom_bar(stat='identity') +
  theme_bw()
```


Arrival_Delays
Los Angeles
Seattle
City
Arrival_Status
delayed
on time