Assignment – Tidying and Transforming Data

Mathew Katz

2022-10-01

Read in necessary libraries:

```
library(tidyr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(stringr)
library(ggplot2)
```

Read in csv and look at it:

```
flights <- read.csv(file = 'flight_tidying.csv')
flights
```

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

Look at structure of csv:

```
str(flights)
```

```
## 'data.frame':    5 obs. of  7 variables:
##  $ X            : chr  "Alaska" "" "" "AM WEST" ...
##  $ X.1          : chr  "on time" "delayed" "" "on time" ...
```

```
##  $ Los.Angeles  : int  497 62 NA 694 117
##  $ Phoenix      : int  221 12 NA 4840 415
##  $ San.Diego    : int  212 20 NA 383 65
##  $ San.Francisco: int  503 102 NA 320 129
##  $ Seattle      : int  1841 305 NA 201 61
```

Change columnn names to 'better' names

```
flights <- flights %>%
    rename('Airline' = 1, 'Flight.Status' = 2)
flights
```

```
##   Airline Flight.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
## 3                                NA      NA        NA            NA      NA
## 4 AM WEST       on time         694    4840       383           320     201
## 5               delayed         117     415        65           129      61
```

Remove NAs and empty rows in csv:

```
flights <- flights %>% #
            filter(! is.na(Flight.Status) &
                     str_length(Flight.Status) > 0)
flights
```

```
##   Airline Flight.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
## 3 AM WEST       on time         694    4840       383           320     201
## 4               delayed         117     415        65           129      61
```

Turn csv into dataframe:

```
df <- data.frame(flights)
head(df)
```

```
##   Airline Flight.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
## 3 AM WEST       on time         694    4840       383           320     201
## 4               delayed         117     415        65           129      61
```

Add missing airline information:

```
df[2, "Airline"] <- "Alaska"
df[4, "Airline"] <- "AM WEST"
df
```

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

Lengthen data by increasing the number of rows and decreasing the number of columns:

```
df <- df %>%
        pivot_longer(!c("Airline", "Flight.Status"),
                     names_to = "Destination",
                     values_to = "Count")
df
```

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

Write to CSV:

```
write.csv(df, "./clean_flight_tidying.csv", row.names=FALSE)
```

Create two new dataframes of delayed and non time flights to graph them:

```
delayed_flights <- df %>%
                    filter(df$Flight.Status == "delayed")
delayed_flights
```

```
## # A tibble: 10 x 4
##    Airline Flight.Status Destination   Count
##    <chr>   <chr>         <chr>         <int>
##  1 Alaska  delayed       Los.Angeles      62
##  2 Alaska  delayed       Phoenix          12
##  3 Alaska  delayed       San.Diego        20
##  4 Alaska  delayed       San.Francisco   102
##  5 Alaska  delayed       Seattle         305
##  6 AM WEST delayed       Los.Angeles     117
##  7 AM WEST delayed       Phoenix         415
```

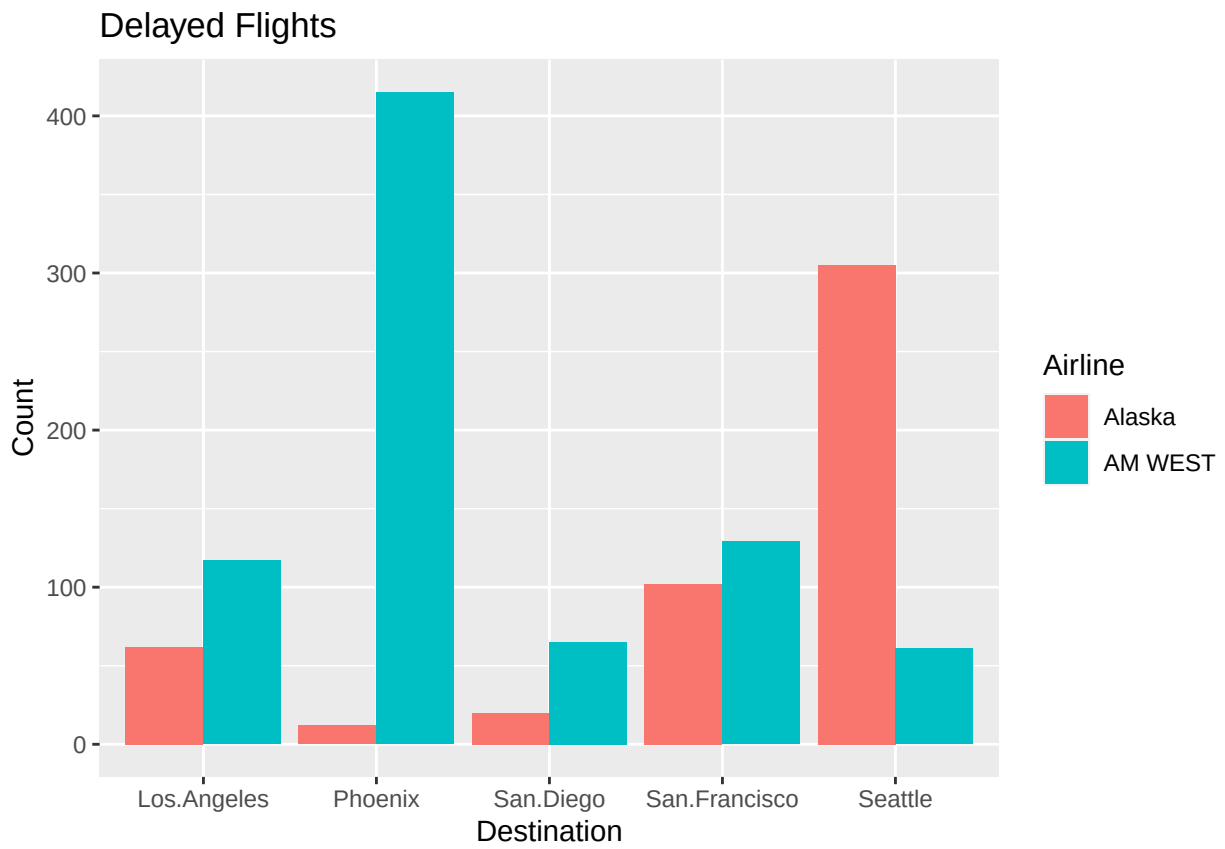
```
##  8 AM WEST delayed       San.Diego        65
##  9 AM WEST delayed       San.Francisco   129
## 10 AM WEST delayed       Seattle          61
```

```
on_time_flights <- df %>%
                   filter(df$Flight.Status == "on time")
on_time_flights
```

```
## # A tibble: 10 x 4
##    Airline Flight.Status Destination   Count
##    <chr>   <chr>         <chr>         <int>
##  1 Alaska  on time       Los.Angeles     497
##  2 Alaska  on time       Phoenix         221
##  3 Alaska  on time       San.Diego       212
##  4 Alaska  on time       San.Francisco   503
##  5 Alaska  on time       Seattle        1841
##  6 AM WEST on time       Los.Angeles     694
##  7 AM WEST on time       Phoenix        4840
##  8 AM WEST on time       San.Diego       383
##  9 AM WEST on time       San.Francisco   320
## 10 AM WEST on time       Seattle         201
```

Graph Delayed Flights:

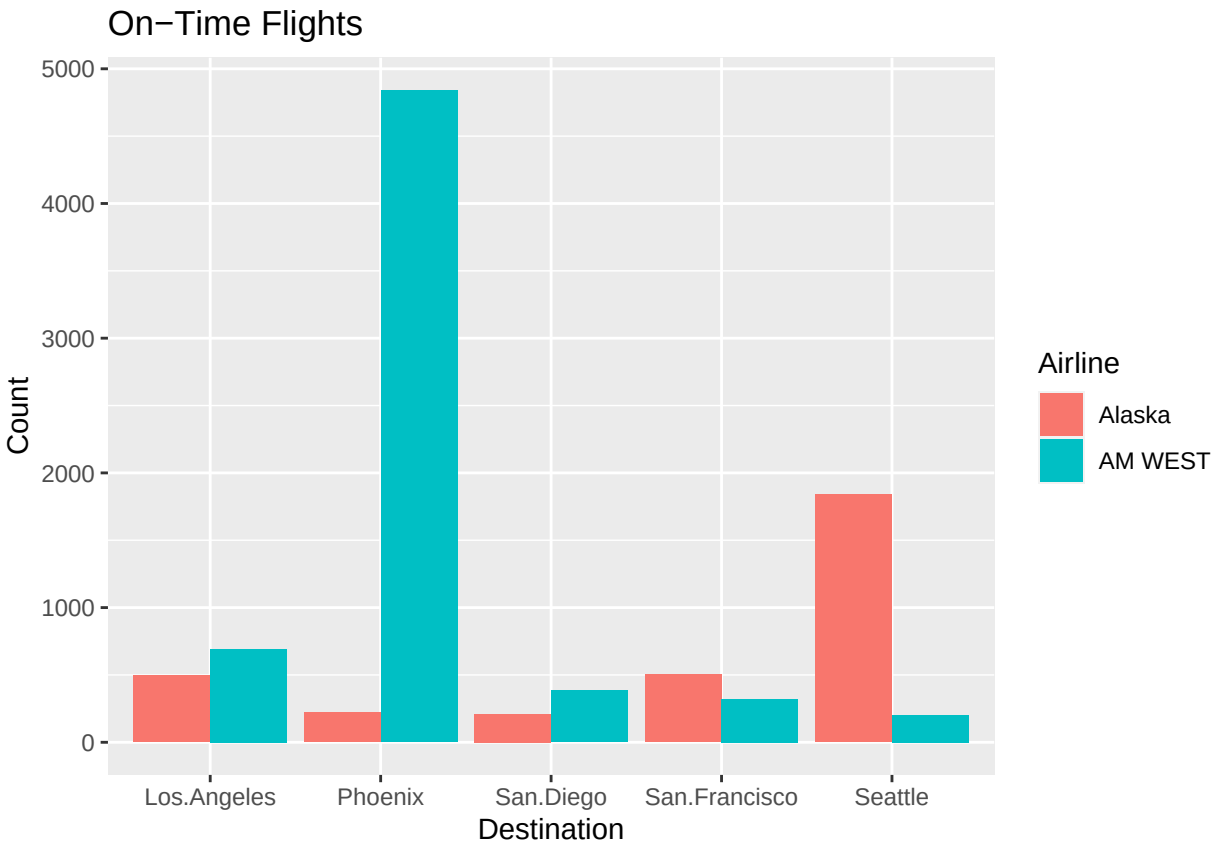
```
 delayed_bar_graph <- ggplot(data=delayed_flights, aes(x=Destination, y=Count, fill=Airline))
    delayed_bar_graph <- delayed_bar_graph +  ggtitle('Delayed Flights') +
            geom_bar(stat="identity", position=position_dodge())
delayed_bar_graph
```



AM WEST has more delayed flights than Alaska in ever city but Seattle.

Graph On-Time Flights:

```
 ontime_bar_graph <- ggplot(data=on_time_flights, aes(x=Destination, y=Count, fill=Airline))
    ontime_bar_graph <- ontime_bar_graph +  ggtitle('On-Time Flights') +
            geom_bar(stat="identity", position=position_dodge())
ontime_bar_graph
```



AM WEST has more on-time flights than Alaska in Los Angeles, Phoenix, and San Diego. Alaska has more on-time flights than AM WEST in San Francisco and Seattle.