a_4: Tidying and Transforming Data

Jie Zou

3/4/2021

Create a .CSV file

The data from provided image is pretty small, therefore I manually enter these data into Excel with exactly the same format showed in the image, and save it as .CSV file and upload it into git first.

Read .CSV file

```
df<-read.csv("https://raw.githubusercontent.com/Sugarcane-svg/R/main/R607/Assignments/a4/a_4.csv")
kable(df)
```

X	X.1	Los.Angeles	Phoenix	San.Diego	San.Francisco	Seattle
ALASKA	on time	497	221	212	503	1841
	delayed	62	12	20	102	305
		NA	NA	NA	NA	NA
AMWEST	on time	694	4840	383	320	201
	delayed	117	415	65	129	61

tidy data

To tidy data, we are going to make up the missing column names, remove na or unnecessary 'white line', and make up some data(if necessary)

```
colnames(df) <- c("airline", "status","LA", "PHX", "SAN", "SFO", "SEA")
head(df)
```

1. modify column names

```
##   airline  status  LA  PHX SAN SFO  SEA
## 1  ALASKA on time 497  221 212 503 1841
## 2         delayed  62   12  20 102  305
## 3                  NA   NA  NA  NA   NA
## 4  AMWEST on time 694 4840 383 320  201
## 5         delayed 117  415  65 129   61
```

```
df <- df[-c(3),]
head(df)
```

2. remove line 3

```
##   airline  status  LA  PHX SAN SFO  SEA
## 1  ALASKA on time 497  221 212 503 1841
## 2         delayed  62   12  20 102  305
## 4  AMWEST on time 694 4840 383 320  201
## 5         delayed 117  415  65 129   61
```

```
df[2,1] <-"ALASKA"
df[4,1] <-"AMWEST"
head(df)
```

3. fill data in blank space

```
##   airline  status  LA  PHX SAN SFO  SEA
## 1  ALASKA on time 497  221 212 503 1841
## 2  ALASKA delayed  62   12  20 102  305
## 4  AMWEST on time 694 4840 383 320  201
## 5  AMWEST delayed 117  415  65 129   61
```

```
alaska.df<-data.frame(airline = character(), dest = character(), on_time = numeric(), delayed = numeric

amwest.df<-data.frame(airline = character(), dest = character(), on_time = numeric(), delayed = numeric

for( i in 1:5){
  alaska.df[i,] <- c("ALASKA",
                     colnames(df)[i+2],
                     (df %>%
                       filter(airline == "ALASKA" & status == "on time")%>%
                       select(-c(airline, status)))[1,i],

                     (df %>%
                       filter(airline == "ALASKA" & status == "delayed")%>%
                       select(-c(airline, status)))[1,i]
                     )
}

alaska.df
```

4(optional). split data into two for the simplicity of following steps

```
##   airline dest on_time delayed
```

```
## 1  ALASKA   LA     497      62
## 2  ALASKA  PHX     221      12
## 3  ALASKA  SAN     212      20
## 4  ALASKA  SFO     503     102
## 5  ALASKA  SEA    1841     305
```

```
for( i in 1:5){
  amwest.df[i,] <- c("AMWEST",
                     colnames(df)[i+2],
                     (df %>%
                       filter(airline == "AMWEST" & status == "on time")%>%
                       select(-c(airline, status)))[1,i],

                     (df %>%
                       filter(airline == "AMWEST" & status == "delayed")%>%
                       select(-c(airline, status)))[1,i]
                     )
}
amwest.df
```

```
##   airline dest on_time delayed
## 1  AMWEST   LA     694     117
## 2  AMWEST  PHX    4840     415
## 3  AMWEST  SAN     383      65
## 4  AMWEST  SFO     320     129
## 5  AMWEST  SEA     201      61
```

Analysis

compare the arrival delays for two airlines

```
df$mean <- rowMeans(df[, 3:7])
df
```

1. lets see the mean of both on time and delayed values from these two airlines

```
##   airline  status  LA  PHX SAN SFO  SEA   mean
## 1  ALASKA on time 497  221 212 503 1841  654.8
## 2  ALASKA delayed  62   12  20 102  305  100.2
## 4  AMWEST on time 694 4840 383 320  201 1287.6
## 5  AMWEST delayed 117  415  65 129   61  157.4
```

as we can see from the mean, it seems like Alaska has better on time and less delayed. However, when we observe the data, we can also see that in some specific destinations, there are some on time flights unexpected longer than others such as flying from Alaska to Seattle and flying from Amwest to Phoenix. The number of both on time are much higher compared to other destinations.

2. check the ratio of arrival delays for both airline The total number of flights for Alaska is 3775 and Amwest is 7225

```
df$total <- rowSums(df[, 3:7])
df %>% select(airline, total) %>%
  group_by(airline) %>%
  summarise(total_number_of_flights = sum(total))
```

```
## # A tibble: 2 x 2
##   airline total_number_of_flights
## * <chr>                     <dbl>
## 1 ALASKA                     3775
## 2 AMWEST                     7225
```

the probability of delay for alaska and amwest flights are 13.27% and 10.89% respectively. compare the ratio, we can see that the delay from alaska is worse than the delay from amwest.

```
# alaska delay ratio
ak_delay <-df %>%
  filter(airline == "ALASKA" & status == "delayed") %>%
  select(total)

ak_delay/3775
```

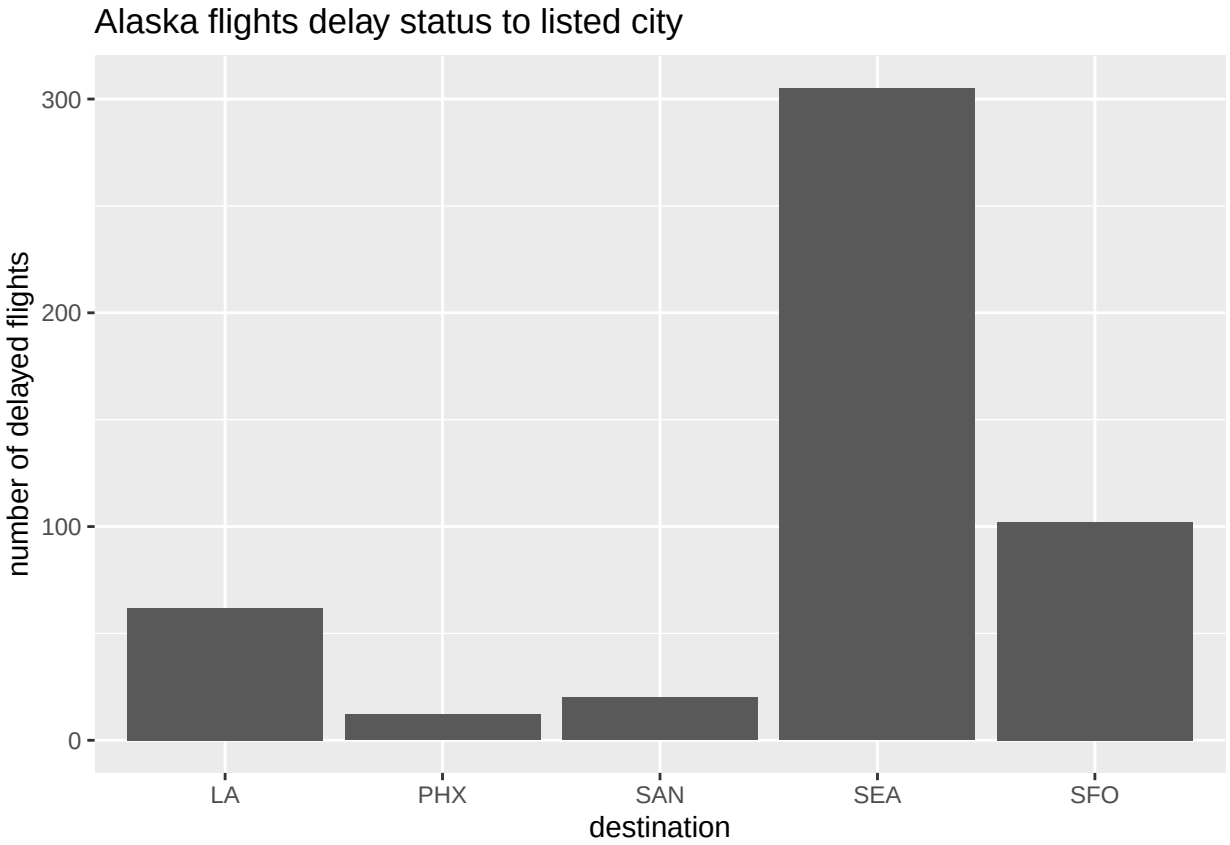
```
##       total
## 1 0.1327152
```

```
# amwest delay ratio

am_delay<- df %>%
  filter(airline == "AMWEST" & status == "delayed") %>%
  select(total)

am_delay/7225
```

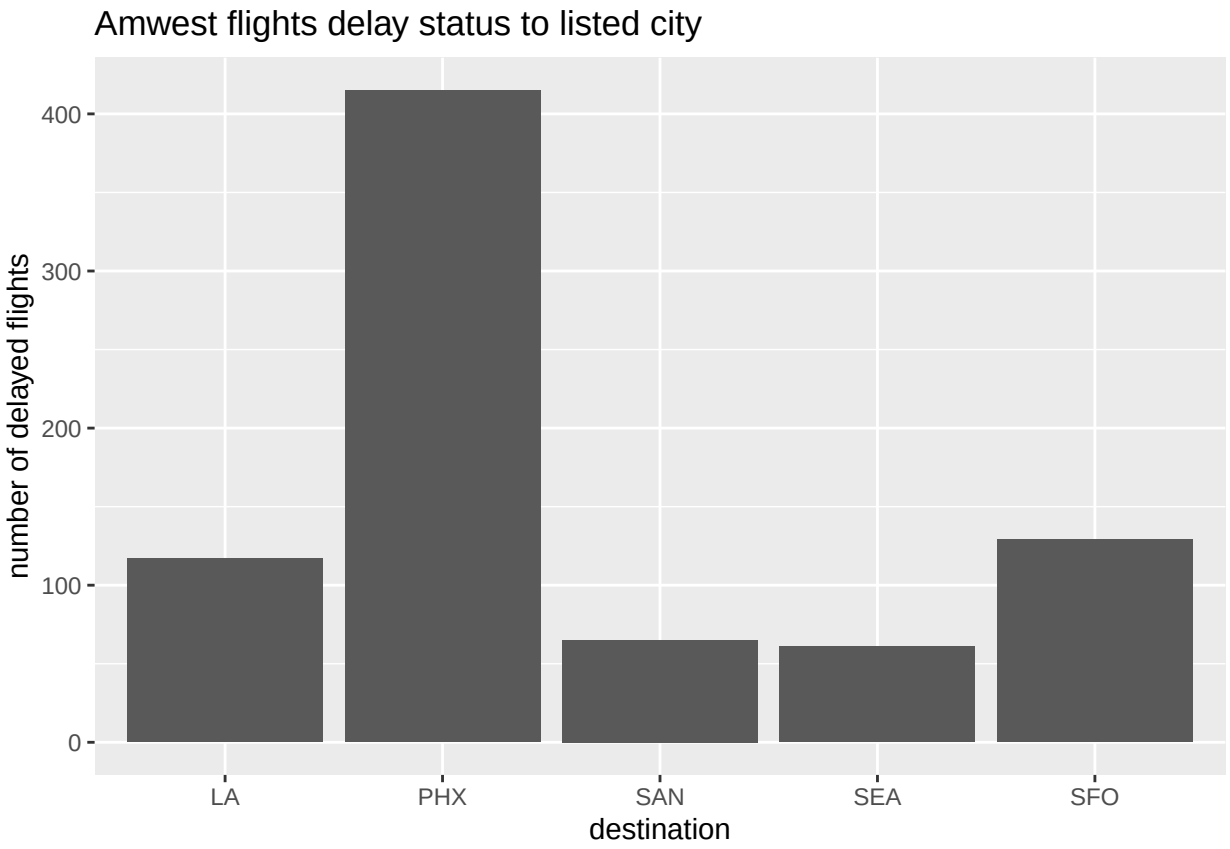
```
##       total
## 1 0.1089273
```

3. see the graph of individual airline delay status of listed city from the graph, it can only tell/confirm that there are more flights delayed from Alaska to Seattle, and there are more flights delayed from Amwest to Phoenix. The flight delayed from Alaska to San Diego is less than the flights from Amwest.

```
ak_delay <- alaska.df %>% select(dest, delayed) %>% arrange(dest)

ggplot(ak_delay, aes(x=dest, y = as.numeric(delayed))) + geom_bar(stat = "identity") + labs(x = "destina
```




```
am_delay <- amwest.df %>% select(dest, delayed)%>% arrange(dest)

ggplot(am_delay, aes(x=dest, y = as.numeric(delayed))) + geom_bar(stat = "identity") + labs(x = "destina
```

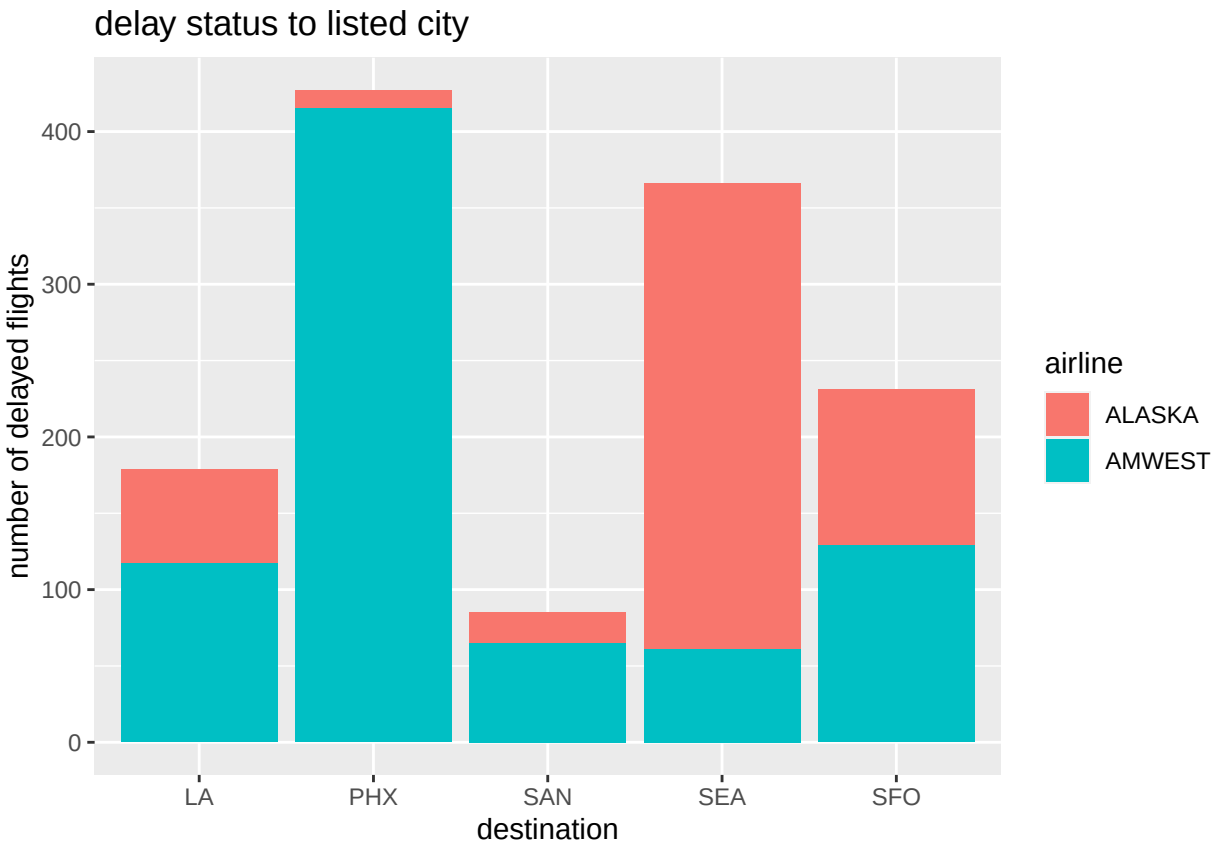


4. see both airline delay status in one graph according to this graph, it shows clear comparison between these two airlines to selected five cities. As we can see that flights from Alaska to LA, PHX, SAN, SFO have less number of delayed. And flying from Amwest to Seattle appears less delayed.

```
df2 <- rbind(alaska.df, amwest.df)
df2
```

```
##    airline dest on_time delayed
## 1   ALASKA   LA     497      62
## 2   ALASKA  PHX     221      12
## 3   ALASKA  SAN     212      20
## 4   ALASKA  SFO     503     102
## 5   ALASKA  SEA    1841     305
## 6   AMWEST   LA     694     117
## 7   AMWEST  PHX    4840     415
## 8   AMWEST  SAN     383      65
## 9   AMWEST  SFO     320     129
## 10  AMWEST  SEA     201      61
```

```
ggplot(df2, aes(x=dest, y = as.numeric(delayed), fill = airline)) +
  geom_bar(stat = "identity")+
  labs(x = "destination", y = "number of delayed flights", title = "delay status to listed city")
```



Conclusion

If a customer is deciding flying from these two airline to San Diego/Phoenix, I will suggest to take Alaska, and if a customer try to find a airline to Seattle, I will suggest to choose Amwest.