Homework 5

Jack Wright

9/24/2020

Load Packages

```
library(tidyverse)
library(RCurl)
```

Get data from github using RCurl package

```
url<-"https://raw.githubusercontent.com/JackJosephWright/data/master/hw5.csv"



dat<-read.csv(url)
```

Tidying data

lets take a look, I will note in the comments what actions I will take to tidy the data.

```
#rename a couple columns for ease of use
names(dat)[1]<-"carrier"
names(dat)[2]<-"status"




#remove blank row
dat1<-dat%>%
  rowwise()%>%
  filter(!is.na(Los.Angeles))

#fill carrier column, fill() is not working I will dig deeper if i have time

dat1$carrier[2]<-"ALASKA"
dat1$carrier[4]<-"AM WEST"
```

In order for data to be tidy, every column must be a variable and every row must be an observation. On first glance, the destinations look like values of a “destination” column and their values look like “count”.

the “X column looks like values of a”carrier" variable

```
#converting "destinations"  to destination column and their values to "count"
carrier_status<-dat1%>%
  pivot_longer(c("Los.Angeles":"Seattle"),names_to="destination",values_to="count")%>%
  arrange(destination)
head(carrier_status)
```

```
## # A tibble: 6 x 4
##   carrier status  destination count
##   <chr>   <chr>   <chr>       <int>
## 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
```

Data Analysis

I want to see which carrier supplied more on time flights.

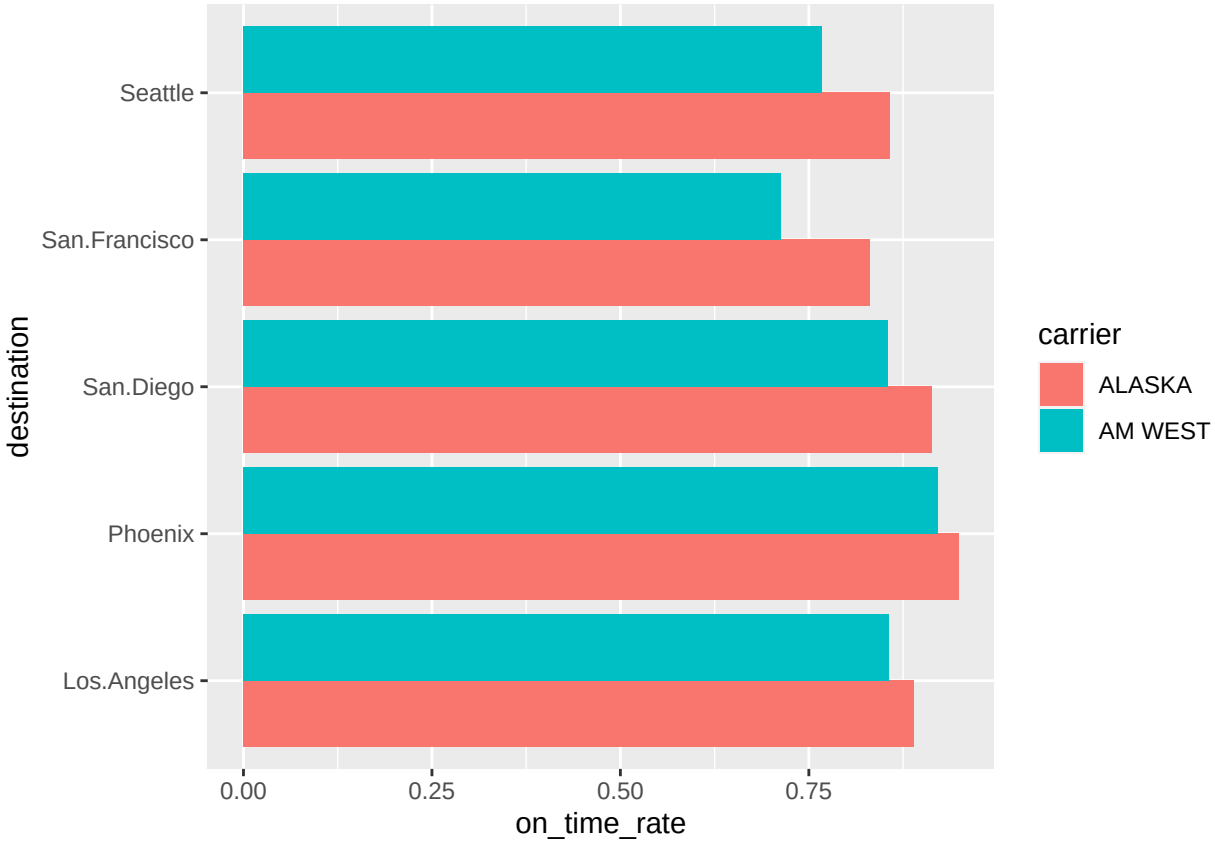
I will turn the counts into percentages, only display the on time results, because the delayed is redundant, and drop the “status and”count" columns.

```
carrier_stats<-carrier_status%>%
  group_by(carrier,destination)%>%
  mutate(on_time_rate=count/sum(count))%>%
  filter(status=="on time")%>%
  select(-status, -count)
head(carrier_stats)
```

```
## # A tibble: 6 x 3
## # Groups:   carrier, destination [6]
##   carrier destination on_time_rate
##   <chr>   <chr>              <dbl>
## 1 ALASKA  Los.Angeles        0.889
## 2 AM WEST Los.Angeles        0.856
## 3 ALASKA  Phoenix            0.948
## 4 AM WEST Phoenix            0.921
## 5 ALASKA  San.Diego          0.914
## 6 AM WEST San.Diego          0.855
```

I want to see how each carrier did flying to each city. I will make a “dodge” bar plot allowing easy comparison.

```
carrier_stats%>%
  ggplot(aes(fill=carrier,y=on_time_rate,x=destination))+
  geom_bar(position="dodge",stat="identity")+coord_flip()
```



From this plot, it is easy to see Alaska Airlines is better across the board. You can also see that Phoenix has the most on time flights. Further analysis could be done on other datasets to see why Phoenix is so effective.