Data607HW5

Sean Connin

Overview: Tidying and Transforming Data

In this assignment we are provided an untidy dataset of containing flight counts for two airlines and five cities by flight status (on time vs. delayed). Our task is to create a csv file for this information in wide format and then, using dplyr and tidyr, put it into tidy form. We are also asked to analyze the data and create appropriate graphics related to the following interactions:

1. A comparison of the per-city on time performance of each airline.
2. The overall on time performance of each airline.

We are also asked to explain any discrepancies between the overall and on time performances as well as any paradoxical conclusions.

I divided this work into the following steps:

1. Create an untidy CSV file identical in layout to our homework sheet.
2. Import the csv into RStudio and put it in tidy form with appropriate formatting.
3. Calculate summary statistics for the data
4. Plot data graphics
5. Discuss conclusions
6. Provide URLs for my work on Github & Rpubs

Step 2. Read and Clean Untidy CSV File

I completed Step 1 outside of RPubs. I then imported the resulting untidy csv file (numbersense.csv) into Rstudio and converted it to tidy form using the tidyverse library - which includes dplyr and tidyr. I then saved the resulting dataframe as a separate csv file (hwfivecldata.csv).

In addition, I developed a SQL script to create and populate the latter into my local MYSQL database. I also attempted to upload the dataset to an AWS RMD but was not successful.

```
# Read in csv file

num <- read.csv("numbersense.csv", sep=",")

#Tidy the dataframe

num%<>%mutate(Phoenix=str_remove(Phoenix, ","))%>%mutate(Seattle=str_remove(Seattle, ","))

num%<>%mutate_at(c(4,7), as.numeric)

num%<>%clean_names()%>%pivot_longer(cols=-c(i:x), names_to = "City", values_to = "Flight_Count", values_drop_na=TRUE)

num%<>%rename(Airline = i, Flight_Status=x)
num[c(6:10),1] ="ALASKA"
num[c(16:20),1] ="AMWEST"
num <- num %>% mutate(Id = row_number())%>%relocate(Id, .before = Airline)

# Create table for first five rows

head(num, 5)%>%kbl%>%kable_material(c("striped"))
```

Id	Airline	Flight_Status	City	Flight_Count
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

```
# Save as csv file for import into MYSQL

write.csv(num,"C:\\Users\\seanc\\Documents\\Data_Science\\CUNY\\Data 607 Acquisition and Management\\Assignments\\WK5\\MyDat
a.csv", row.names = FALSE)
```

Step 3. Summary Statistics

First, I calculated on-time performance by airline and city. Then I calculated it for the airlines.

On-time performance was determined by dividing on-time arrival counts by the sum of on-time and delayed counts for each airline.

Finally, I calculated the total number of flights documented for each airline.

```
# Compute performance metric for each airline by city.

(air_city<-num%>%select(Airline, Flight_Status, City, Flight_Count)%>%pivot_wider(names_from = Flight_Status, values_from =
 Flight_Count)%>%clean_names()%>% mutate(performance=round((on_time/(on_time+delayed))*100)))
```

airline	city	on_time	delayed	performance
<chr>	<chr>	<dbl>	<dbl>	<dbl>
ALASKA	los_angeles	497	62	89
ALASKA	phoenix	221	12	95
ALASKA	san_diego	212	20	91
ALASKA	san_francisco	503	102	83
ALASKA	seattle	1841	305	86
AMWEST	los_angeles	694	117	86
AMWEST	phoenix	4840	415	92
AMWEST	san_diego	383	65	85
AMWEST	san_francisco	320	129	71
AMWEST	seattle	201	61	77

1-10 of 10 rows

```
air_city%>%kbl%>%kable_material(c("striped"))
```

airline	city	on_time	delayed	performance
ALASKA	los_angeles	497	62	89
ALASKA	phoenix	221	12	95
ALASKA	san_diego	212	20	91
ALASKA	san_francisco	503	102	83
ALASKA	seattle	1841	305	86
AMWEST	los_angeles	694	117	86
AMWEST	phoenix	4840	415	92
AMWEST	san_diego	383	65	85
AMWEST	san_francisco	320	129	71
AMWEST	seattle	201	61	77

```
# Compute overall on time performance for each airline

air_overall<-num%>%select(Airline, Flight_Status, Flight_Count)%>%group_by(Airline, Flight_Status)%>%summarize(Flight_Total=
sum(Flight_Count))
```

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

```
air_overall%<>%pivot_wider(names_from = Flight_Status, values_from = Flight_Total)%>%clean_names()%>%mutate(performance=roun
d((on_time/(on_time+delayed))*100))

air_overall%>%kbl%>%kable_material(c("striped"))
```

airline	delayed	on_time	performance
ALASKA	501	3274	87
AMWEST	787	6438	89

```
#Compute the total number of flights for each airline

n_total <- num%>%group_by(Airline)%>%summarize(Frequency = sum(Flight_Count))

n_total%>%kbl%>%kable_material(c("striped"))
```

Airline	Frequency
ALASKA	3775
AMWEST	7225

Other Descriptive Stats

Note: I performed addtional analyses but these were not required on the rubric

```
# Descriptive statistics grouped by airline and flight status

stat1sum<-num%>%group_by(Airline, Flight_Status)%>%summarize(Total=sum(Flight_Count), Median=round(median(Flight_Count, na.r
m=TRUE)), Minimum=min(Flight_Count), Maximum=max(Flight_Count),Range=range(Flight_Count),Standard_Deviation=round(sd(Flight_
Count)))
```

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

```
stat1sum%<>%distinct(Airline, .keep_all = TRUE)
stat1sum%>%kbl%>%kable_material(c("striped"))
```

Airline	Flight_Status	Total	Median	Minimum	Maximum	Range	Standard_Deviation
ALASKA	delayed	501	62	12	305	12	120
ALASKA	on time	3274	497	212	1841	212	678
AMWEST	delayed	787	117	61	415	61	147
AMWEST	on time	6438	383	201	4840	201	1994

```
# Descriptive statistics grouped by city and flight status.

(stat2sum<-num%>%group_by(City, Flight_Status)%>%summarize(Median=round(median(Flight_Count, na.rm=TRUE)), Minimum=min(Fligh
t_Count), Maximum=max(Flight_Count),                        Range=range(Flight_Count), Standard_Deviation=round(sd(Flight_Co
unt))))
```

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

City	Flight_Status	Median	Minimum	Maximum	Range	Standard_Deviation
<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
los_angeles	delayed	90	62	117	62	39
los_angeles	delayed	90	62	117	117	39
los_angeles	on time	596	497	694	497	139
los_angeles	on time	596	497	694	694	139
phoenix	delayed	214	12	415	12	285
phoenix	delayed	214	12	415	415	285
phoenix	on time	2530	221	4840	221	3266
phoenix	on time	2530	221	4840	4840	3266
san_diego	delayed	42	20	65	20	32
san_diego	delayed	42	20	65	65	32

1-10 of 20 rows Previous **1** 2 Next

```
stat2sum %<>% distinct(City, Flight_Status, .keep_all = TRUE)
stat2sum%>%kbl%>%kable_material(c("striped"))
```

City	Flight_Status	Median	Minimum	Maximum	Range	Standard_Deviation
los_angeles	delayed	90	62	117	62	39
los_angeles	on time	596	497	694	497	139
phoenix	delayed	214	12	415	12	285

City	Flight_Status	Median	Minimum	Maximum	Range	Standard_Deviation
phoenix	on time	2530	221	4840	221	3266
san_diego	delayed	42	20	65	20	32
san_diego	on time	298	212	383	212	121
san_francisco	delayed	116	102	129	102	19
san_francisco	on time	412	320	503	320	129
seattle	delayed	183	61	305	61	173
seattle	on time	1021	201	1841	201	1160

```
# Descriptive statistics grouped by airline, city, and flight status.

stat3sum<-num%>%group_by(City, Airline, Flight_Status)%>%summarize(median=round(median(Flight_Count)), minimum=min(Flight_Count), maximum=max(Flight_Count), range=range(Flight_Count))
```

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

```
stat3sum %<>% distinct(Airline, City, Flight_Status, .keep_all = TRUE)%>%arrange(Airline)
stat3sum%>%kbl%>%kable_material(c("striped"))
```

City	Airline	Flight_Status	median	minimum	maximum	range

City	Airline	Flight_Status	median	minimum	maximum	range
los_angeles	ALASKA	delayed	62	62	62	62
los_angeles	ALASKA	on time	497	497	497	497
phoenix	ALASKA	delayed	12	12	12	12
phoenix	ALASKA	on time	221	221	221	221
san_diego	ALASKA	delayed	20	20	20	20
san_diego	ALASKA	on time	212	212	212	212
san_francisco	ALASKA	delayed	102	102	102	102
san_francisco	ALASKA	on time	503	503	503	503
seattle	ALASKA	delayed	305	305	305	305
seattle	ALASKA	on time	1841	1841	1841	1841
los_angeles	AMWEST	delayed	117	117	117	117
los_angeles	AMWEST	on time	694	694	694	694
phoenix	AMWEST	delayed	415	415	415	415

City	Airline	Flight_Status	median	minimum	maximum	range
phoenix	AMWEST	on time	4840	4840	4840	4840
san_diego	AMWEST	delayed	65	65	65	65
san_diego	AMWEST	on time	383	383	383	383
san_francisco	AMWEST	delayed	129	129	129	129
san_francisco	AMWEST	on time	320	320	320	320
seattle	AMWEST	delayed	61	61	61	61
seattle	AMWEST	on time	201	201	201	201

Step 4. Data Graphics

I created two bar plots for on time performance. The first compared performance for both airlines by city. The second compared on time performance by airline.

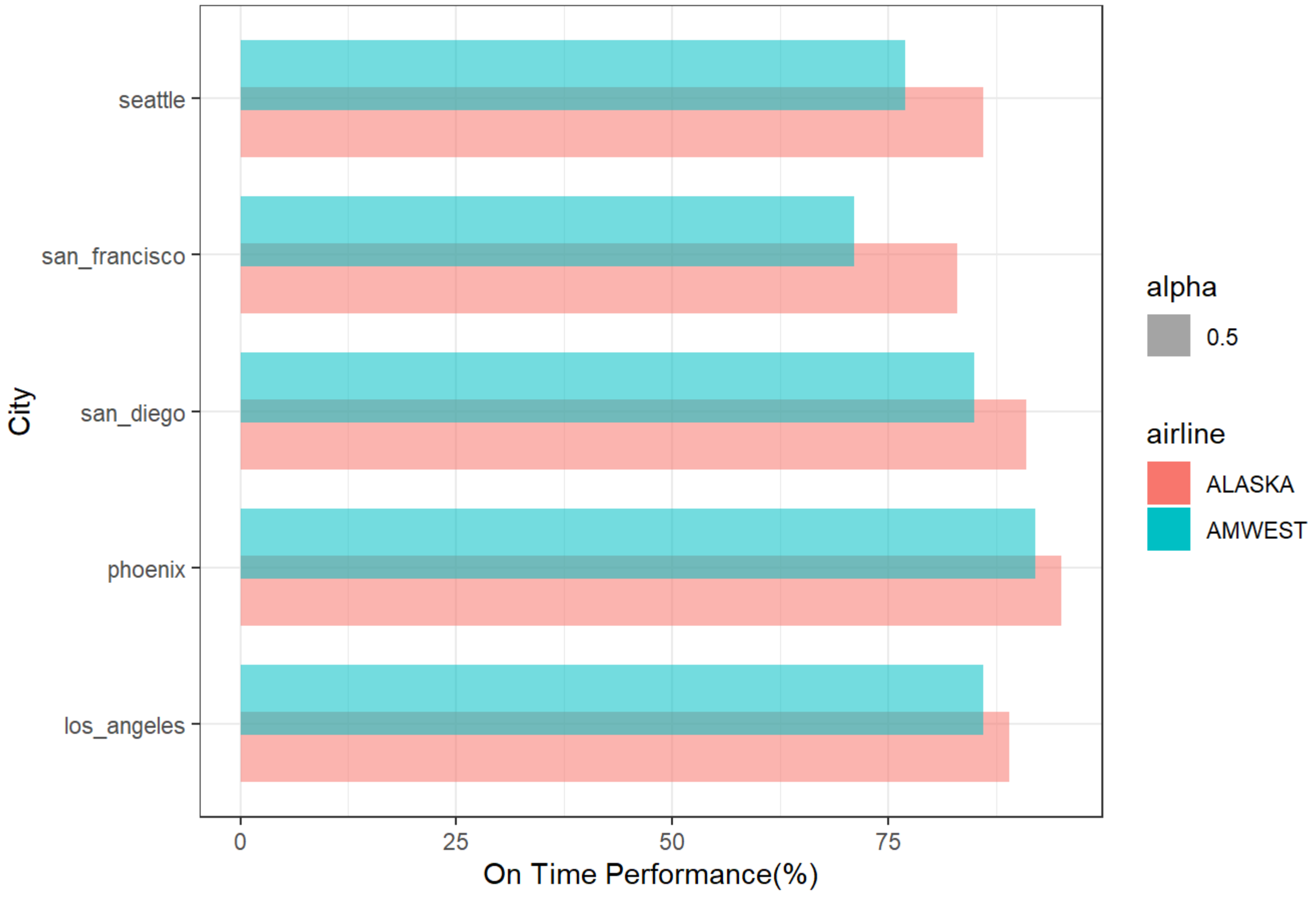
In order to create visually legible boxplots, I removed flight counts > 1000. This applied to airports in Seattle and Phoenix.

I also tried to facet a boxplot graphic highlight counts by flight status, city, and airline. However, the resulting form was not visually legible and, despite much effort, I was unable to improve it.

```
# Bar plot of on time performance by airline and city.

(air_city%>%ggplot(aes(x=city, y=performance, fill=airline, alpha=.5))+geom_bar(stat="identity", position = position_dodge(w
idth = .6))+
theme(axis.text.x= element_text(size=8))+
theme(axis.text.y = element_text(size=8))+
coord_flip()+
theme_bw()+
ggtitle("Figure 1. Performance of Two Airlines by City")+xlab("City")+ylab("On Time Performance(%)"))
```

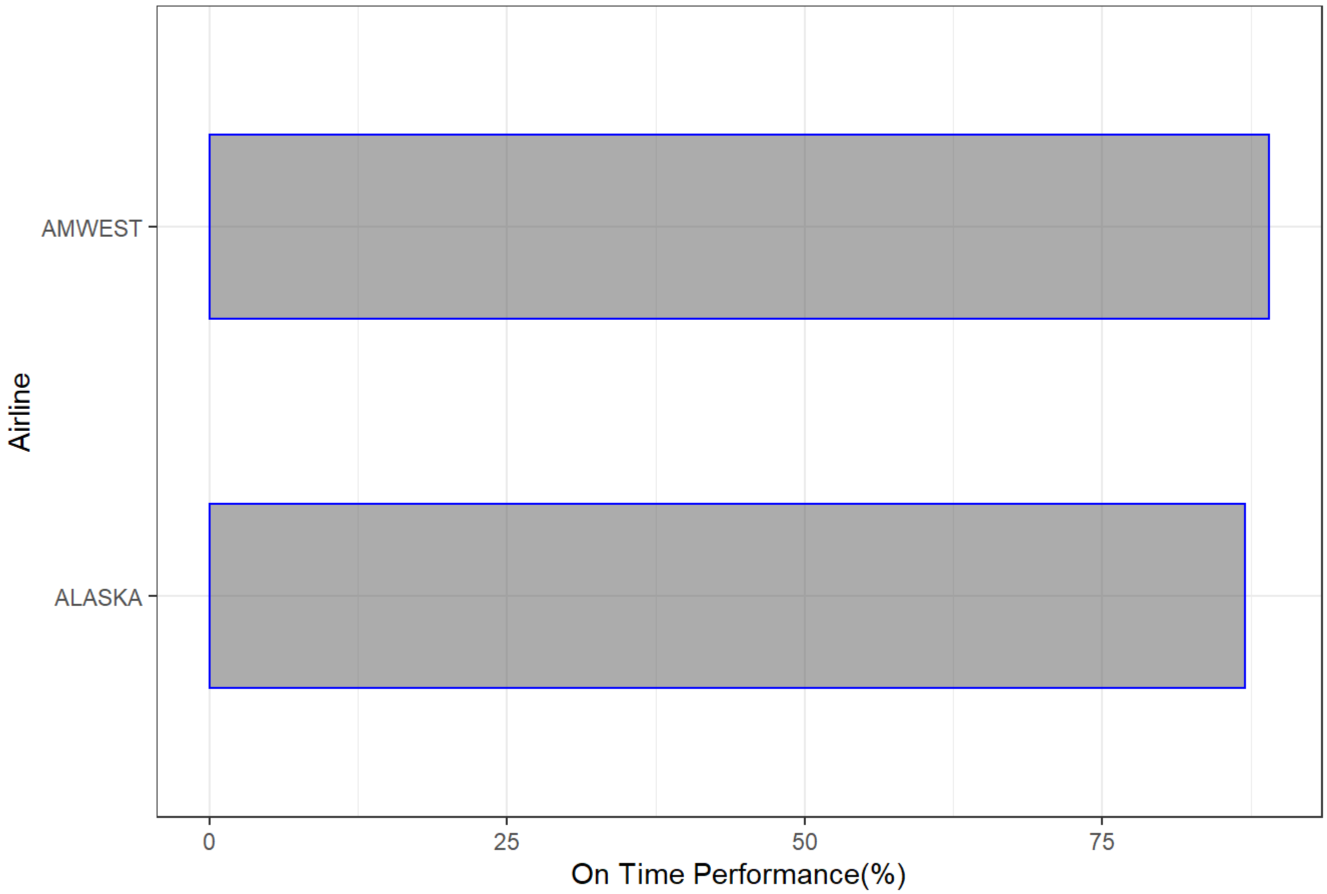
Figure 1. Performance of Two Airlines by City



```
# Bar plot of on time performance by airline.

(air_overall%>%ggplot(aes(x=airline, y=performance))+geom_bar(stat="identity", width=0.5, color="blue", alpha=0.5)+
theme(axis.text.x= element_text(size=8))+
theme(axis.text.y = element_text(size=8))+
coord_flip()+
theme_bw()+
ggtitle("Figure 2. On Time Performance of Two Airlines")+xlab("Airline")+ylab("On Time Performance(%)"))
```

Figure 2. On Time Performance of Two Airlines



```
#Plot boxplots with outliers removed

(plot1 <- num%>%filter(Flight_Count<1000)%>%
    ggplot(aes(x=Airline, y=Flight_Count)) +
    geom_boxplot(alpha=.3, fill="blue")+
    theme(axis.text.x= element_text(size=8))+
    theme(axis.text.y = element_text(size=8))+
    coord_flip()+
    facet_grid(.~Flight_Status, space="free_x")+
    theme_bw()+
    ggtitle("Figure 3. Timeliness of Flights by Airline")+
    xlab("Airline")+
    ylab("Flight Count"))
```

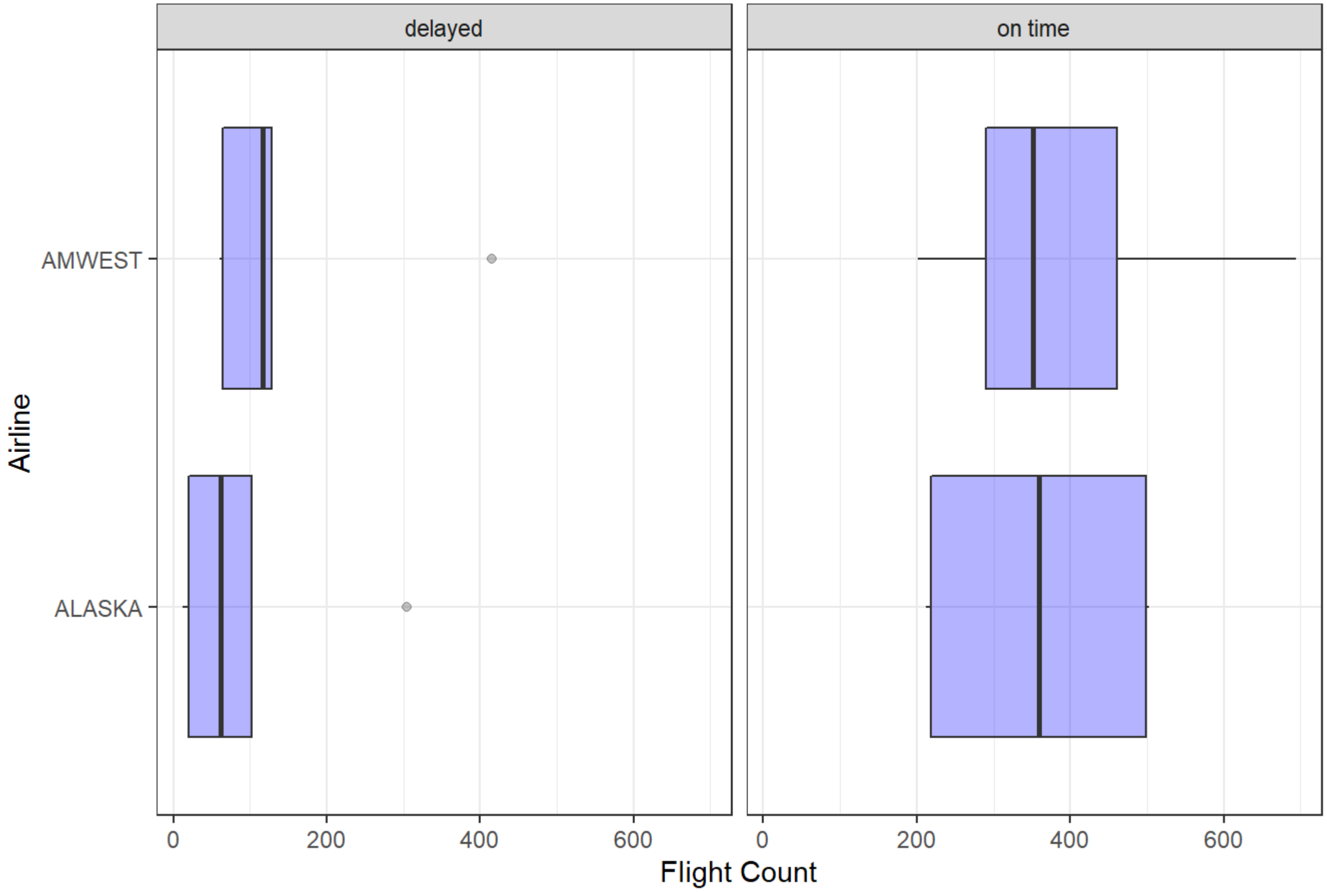
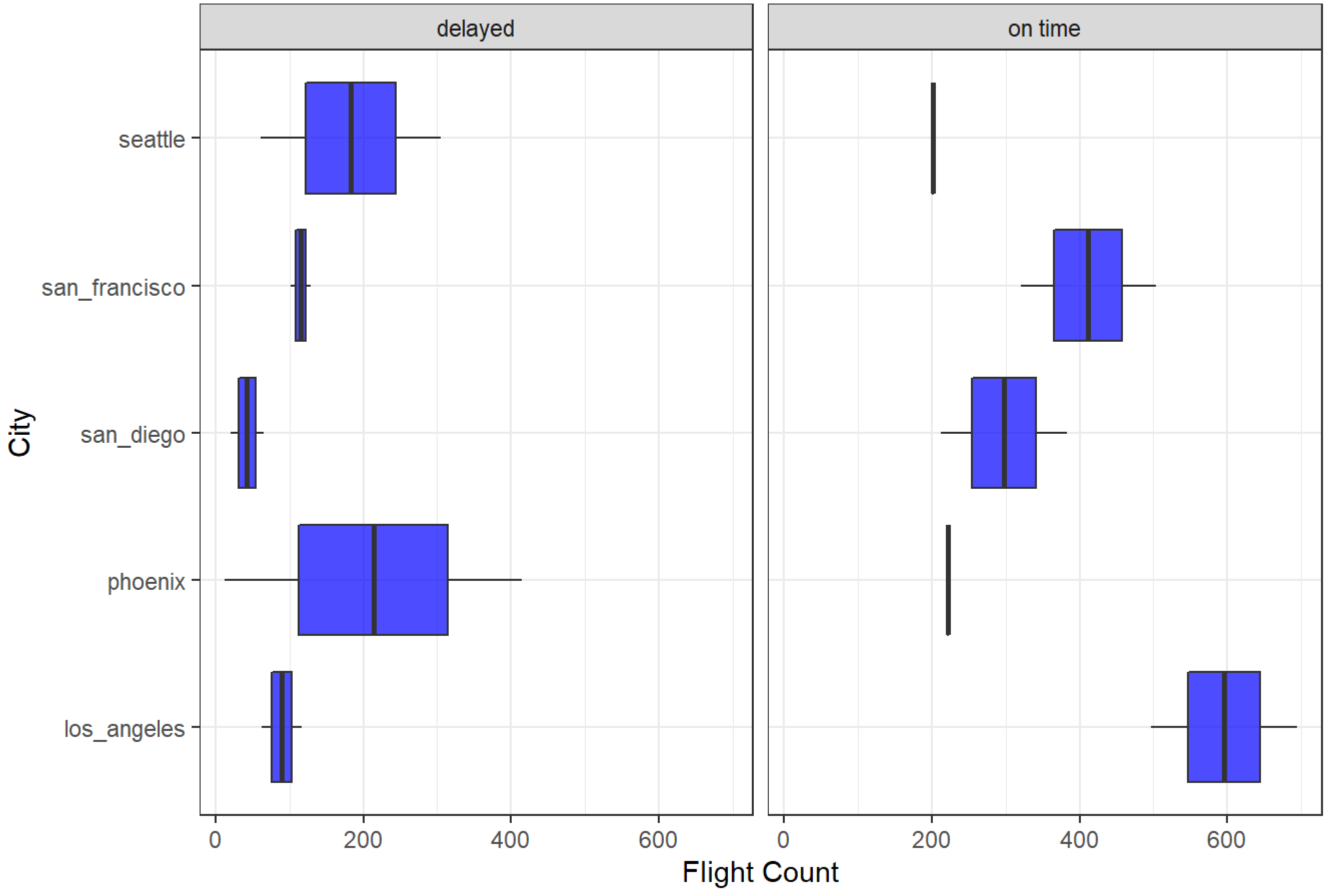


Figure 3. Timeliness of Flights by Airline

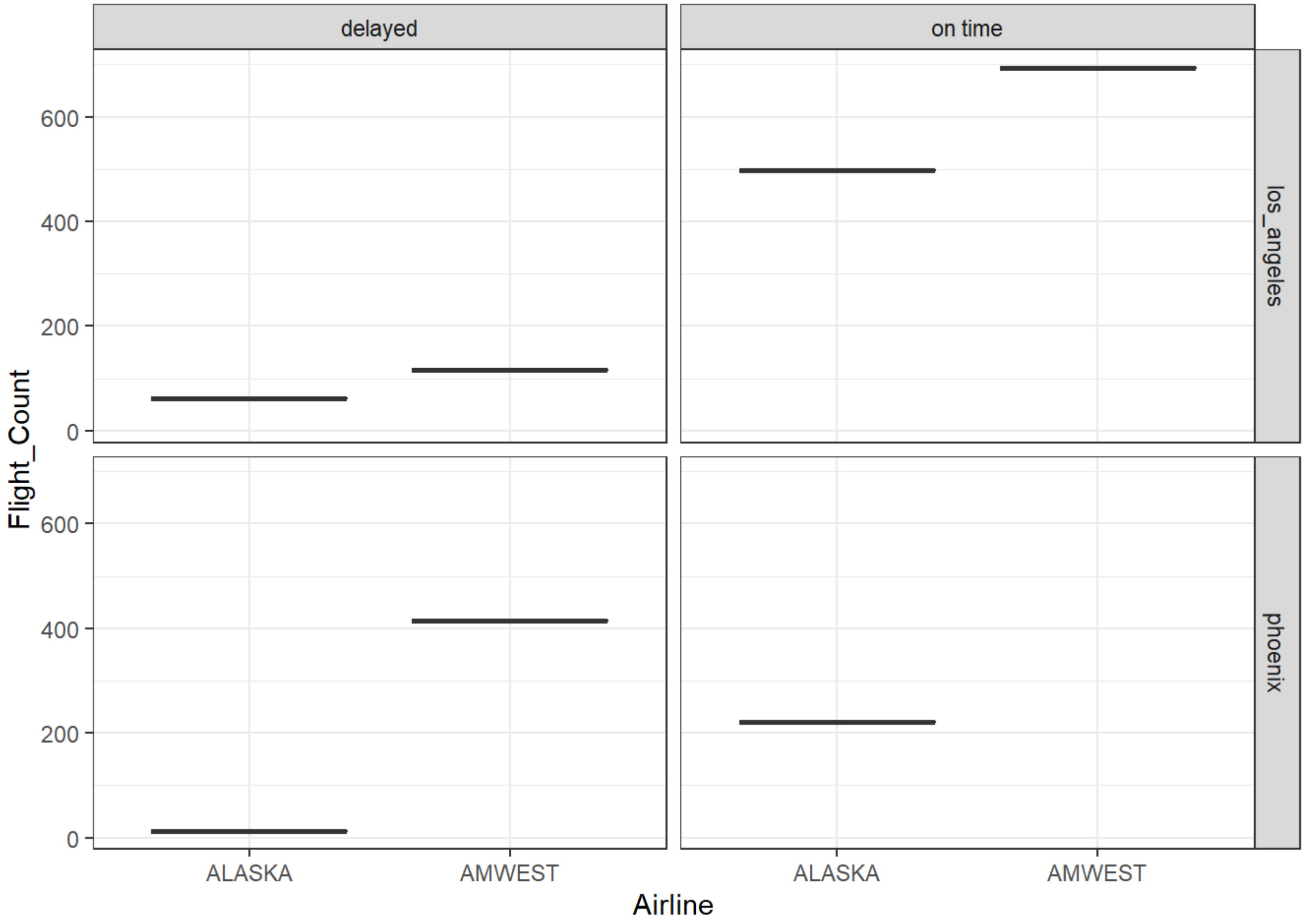
```
(plot2 <- num%>%filter(Flight_Count<1000)%>%
    ggplot(aes(x=City, y=Flight_Count)) +
    geom_boxplot(alpha=.7, fill="blue")+
    theme(axis.text.x= element_text(size=8))+
    theme(axis.text.y = element_text(size=8))+
    coord_flip()+
    facet_grid(.~Flight_Status)+
    theme_bw()+
    ggtitle("Figure 4. Timeliness of Flights by City")+xlab("City")+
    ylab("Flight Count"))
```

Figure 4. Timeliness of Flights by City



```
# Attempt to facet with facet_grid

(plot3 <- num%>%filter(Flight_Count<1000, City=="los_angeles"|City=="phoenix" )%>%
    ggplot(aes(x=Airline, y=Flight_Count))+
    geom_boxplot(alpha=.7, fill="blue")+
    facet_grid_paginate(City~Flight_Status,
    ncol=2, page=1, space="free_y")+
    theme_bw())
```



Step 5. Conclusions

The on-time performance of ALASKA exceeded AMWEST when compared on a city-by-city basis (Figure 1).

However, the overall on-time performance of AMWEST (89%) exceeded ALASKA (87%). See Figure 2.

I ascribe this discrepancy to the fact that the number of flights flown by AMWEST (n=7225) was almost double that of Alaska airlines (n=3775). And the majority of this difference was attributable to on-time flights by AMWEST to Phoenix (n=4840).

It is clear from this dataset that an aggregate measure of performance can lead to a different conclusion (AMWEST outperformed ALASKA) than when distributed measures of performance are analyzed individually (ALASKA outperformed AMWEST).