# IS 607: Assignment 6

*MUSA T. GANIYU*

*March 2, 2016*

```
# load the require packages.
library(stringr);
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(tidyr);
library(ggplot2);
require(knitr);
```

```
## Loading required package: knitr
```

```
# load data file from url.

data1 <- read.csv("https://raw.githubusercontent.com/mascotinme/MSDA-IS607/master/ontime_delayed.csv", s

kable(head(data1));
```

| | | Los Angeles | Phoenix | San Diego | San Franscisco | Seatle |
|---|---|---|---|---|---|---|
| ALASKA | on time | 497 | 221 | 212 | 503 | 1,841 |
| NA | delayed | 62 | 12 | 20 | 102 | 305 |
| NA | NA | NA | NA | NA | NA | NA |
| AM WEST | on time | 694 | 4,840 | 383 | 320 | 201 |
| NA | delayed | 117 | 415 | 65 | 129 | 61 |

```
# It revealed from the data above that the third role has no data at all (empty), we will therefore rem
which(is.na(data1));
```

```
## [1]  2  3  5  8 13 18 23 28 33
```

```
data <- data.frame(data1[-3, ]);

kable(head(data))
```

| | Var.1 | Var.2 | Los.Angeles | Phoenix | San.Diego | San.Franscisco | Seatle |
|---|---|---|---|---|---|---|---|
| 1 | ALASKA | on time | 497 | 221 | 212 | 503 | 1,841 |
| 2 | NA | delayed | 62 | 12 | 20 | 102 | 305 |
| 4 | AM WEST | on time | 694 | 4,840 | 383 | 320 | 201 |
| 5 | NA | delayed | 117 | 415 | 65 | 129 | 61 |

```
# From the glimpes of the data, we can see that the 1st & 2nd columns has no name, we therefore assign
colnames(data)[1] = "Airline"
colnames(data)[2] = "Status"


# after the removal of the row and naming the columns, we saw that there are still empty rows, we there

data[2,1] = "ALASKA"
data[4,1] = "AM WEST"
data[1,7] = 1841
data[3,4] = 4840

kable(head(data));
```

| | Airline | Status | Los.Angeles | Phoenix | San.Diego | San.Franscisco | Seatle |
|---|---|---|---|---|---|---|---|
| 1 | ALASKA | on time | 497 | 221 | 212 | 503 | 1841 |
| 2 | ALASKA | delayed | 62 | 12 | 20 | 102 | 305 |
| 4 | AM WEST | on time | 694 | 4840 | 383 | 320 | 201 |
| 5 | AM WEST | delayed | 117 | 415 | 65 | 129 | 61 |

```
# We now use tidyr to gather the respective rows and columns together in a reasonabe manner.

tidy_data <- gather(data, "Destination", "Number_of_time", 3:7, na.rm = TRUE);
kable(head(tidy_data));
```

| Airline | Status | Destination | Number_of_time |
|---|---|---|---|
| ALASKA | on time | Los.Angeles | 497 |
| ALASKA | delayed | Los.Angeles | 62 |
| AM WEST | on time | Los.Angeles | 694 |
| AM WEST | delayed | Los.Angeles | 117 |
| ALASKA | on time | Phoenix | 221 |
| ALASKA | delayed | Phoenix | 12 |

```
tidy_data1 <- spread(tidy_data, key = Status, value= Number_of_time )
colnames(tidy_data1)[4] = "ontime"
kable(head(tidy_data1));
```

| Airline | Destination | delayed | ontime |
|---|---|---|---|
| ALASKA | Los.Angeles | 62 | 497 |
| ALASKA | Phoenix | 12 | 221 |
| ALASKA | San.Diego | 20 | 212 |

| Airline | Destination | delayed | ontime |
|---|---|---|---|
| ALASKA | San.Franscisco | 102 | 503 |
| ALASKA | Seatle | 305 | 1841 |
| AM WEST | Los.Angeles | 117 | 694 |

```
str(tidy_data1)
```

```
## 'data.frame':    10 obs. of  4 variables:
##  $ Airline    : chr  "ALASKA" "ALASKA" "ALASKA" "ALASKA" ...
##  $ Destination: chr  "Los.Angeles" "Phoenix" "San.Diego" "San.Franscisco" ...
##  $ delayed    : chr  "62" "12" "20" "102" ...
##  $ ontime     : chr  "497" "221" "212" "503" ...
```

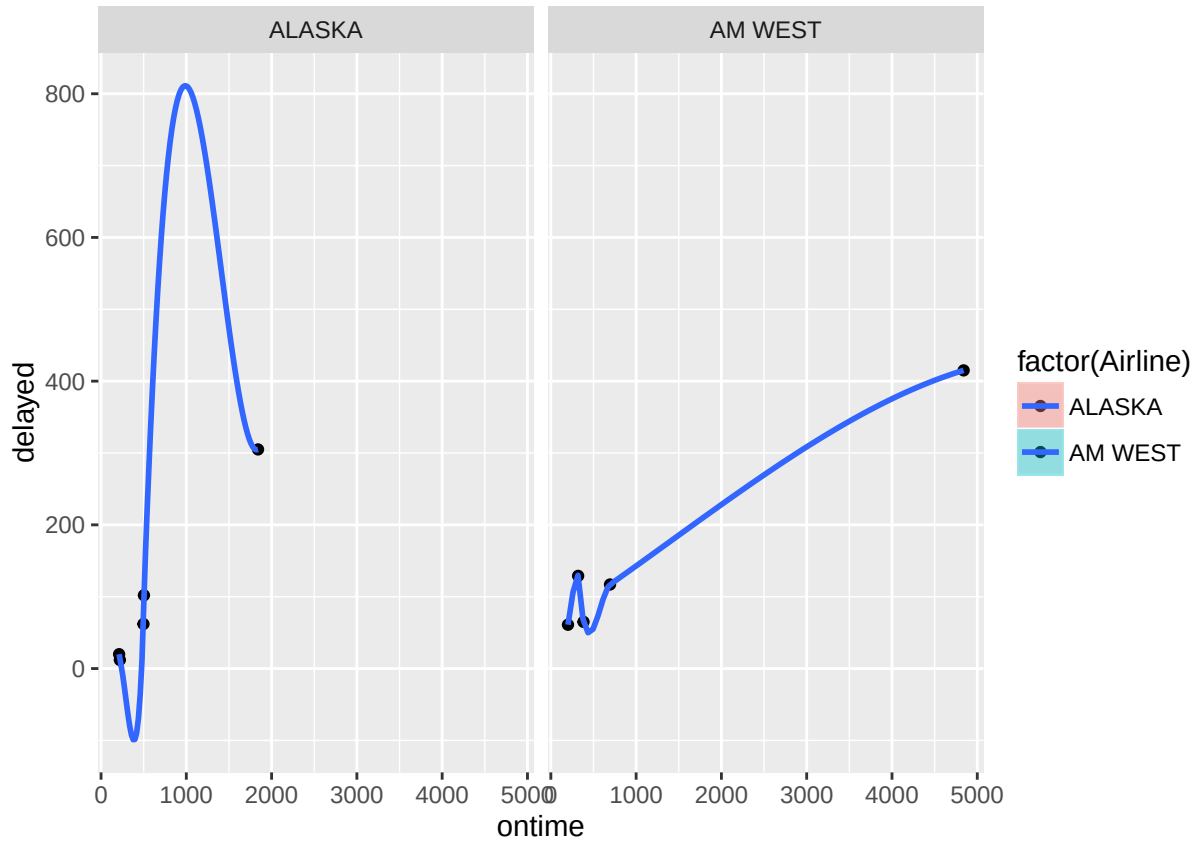
```
# Note that the data type for ontime & delayed is character format, we now change them to numeric for e

tidy_data1<- within(tidy_data1, {
delayed<- as.numeric(as.character(delayed))
ontime<- as.numeric(as.character(ontime))})

# Summarize the average mean for the ontime
data_avg <- tidy_data1 %>% group_by(Airline, Destination) %>%
summarise(Avg=mean(ontime))
data_avg
```

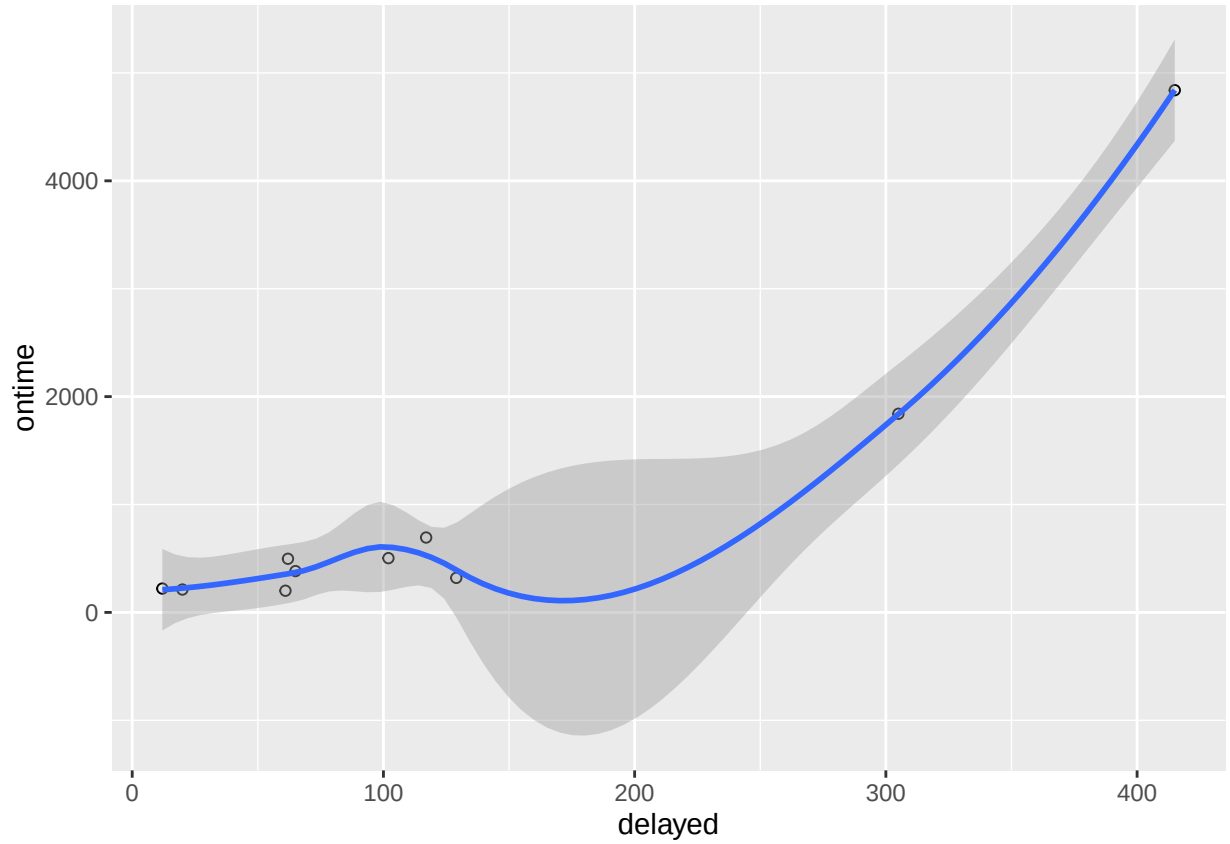
```
## Source: local data frame [10 x 3]
## Groups: Airline [?]
##
##    Airline    Destination   Avg
##      (chr)          (chr) (dbl)
## 1   ALASKA    Los.Angeles   497
## 2   ALASKA        Phoenix   221
## 3   ALASKA      San.Diego   212
## 4   ALASKA San.Franscisco   503
## 5   ALASKA         Seatle  1841
## 6  AM WEST    Los.Angeles   694
## 7  AM WEST        Phoenix  4840
## 8  AM WEST      San.Diego   383
## 9  AM WEST San.Franscisco   320
## 10 AM WEST         Seatle   201
```

```
# Plotting the delayed with ontime
options(warn=-1)
plot <- ggplot(data= tidy_data1, aes(y=delayed, x=ontime, fill=factor(Airline))) + geom_point()
plot + geom_smooth() + facet_wrap(~Airline);
```




```
# plotting graph to visually represent the conclusion.

data_plot <- ggplot(tidy_data1, aes(y=ontime, x=delayed)) + geom_point(shape=1)
data_plot + geom_smooth()
```




```
# some inferences.
kable(tidy_data1 %>% select(Destination,Airline, ontime, delayed) %>% filter(ontime == max(tidy_data1$o
```

| Destination | Airline | ontime | delayed |
|---|---|---|---|
| Phoenix | AM WEST | 4840 | 415 |

```
kable(tidy_data1 %>% select(Destination,Airline, ontime, delayed) %>% filter(delayed == min(tidy_data1$
```

| Destination | Airline | ontime | delayed |
|---|---|---|---|
| Phoenix | ALASKA | 221 | 12 |

```
summary(c(tidy_data1$delayed, tidy_data1$ontime));
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12.00   92.75  216.50  550.00  435.50 4840.00
```

**THANKS FOR YOUR TIME**