Data607 : Tidying and Transforming Data

Mehdi Khan

September 29, 2017

load flight data:

```
FlightData <- read.csv("C:\\temp\\FlightInfo.csv", sep = ",", stringsAsFactors = FALSE)
FlightData
```

```
##          X     X.1 Los.Angeles Phoenix San.Diego San.Francisco Seattle
## 1  ALASKA On Time         497     221       212           503    1841
## 2    <NA> Delayed          62      12        20           102     305
## 3 AM WEST On Time         694    4840       383           320     201
## 4    <NA> Delayed         117     415        65           129      61
```

In the original data, values (city names) are being used as variables, the data also have some empty row values and meaningless column names. So in order to make the data tidy the wide format needs to be converted to long format so that all the city names can be arranged under one variable. Empty values in the rows and meaningless column names also need to be replaced with appropriate values and names respectively:

```
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(data.table)
```

```
##
## Attaching package: 'data.table'

## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

```
library(tidyr)
library(zoo)
```

```
##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(ggplot2)
library(gridExtra)
```

```
## Warning: package 'gridExtra' was built under R version 3.4.2

##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine
```

```
FlightInfo <- FlightData %>% mutate(X = na.locf(X, na.rm = F)) %>%
    setnames(old = c("X", "X.1"), new = c("Airline", "Arrival")) %>%
    gather("City", "Flight_Counts", 3:7)

FlightInfo
```

```
##    Airline Arrival          City Flight_Counts
## 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
## 7  AM WEST On Time       Phoenix          4840
## 8  AM WEST Delayed       Phoenix           415
## 9   ALASKA On Time     San.Diego           212
## 10  ALASKA Delayed     San.Diego            20
## 11 AM WEST On Time     San.Diego           383
## 12 AM WEST Delayed     San.Diego            65
## 13  ALASKA On Time San.Francisco           503
## 14  ALASKA Delayed San.Francisco           102
## 15 AM WEST On Time San.Francisco           320
## 16 AM WEST Delayed San.Francisco           129
## 17  ALASKA On Time       Seattle          1841
## 18  ALASKA Delayed       Seattle           305
## 19 AM WEST On Time       Seattle           201
## 20 AM WEST Delayed       Seattle            61
```

The data looks much better now but still there are two rows for each ovseravtion of a city/Airline pair, so more transformation is needed to make it tidy so that each ovservation can be arranged in a single row:

```
FlightData <- spread(FlightInfo, 2, 4)
FlightData
```

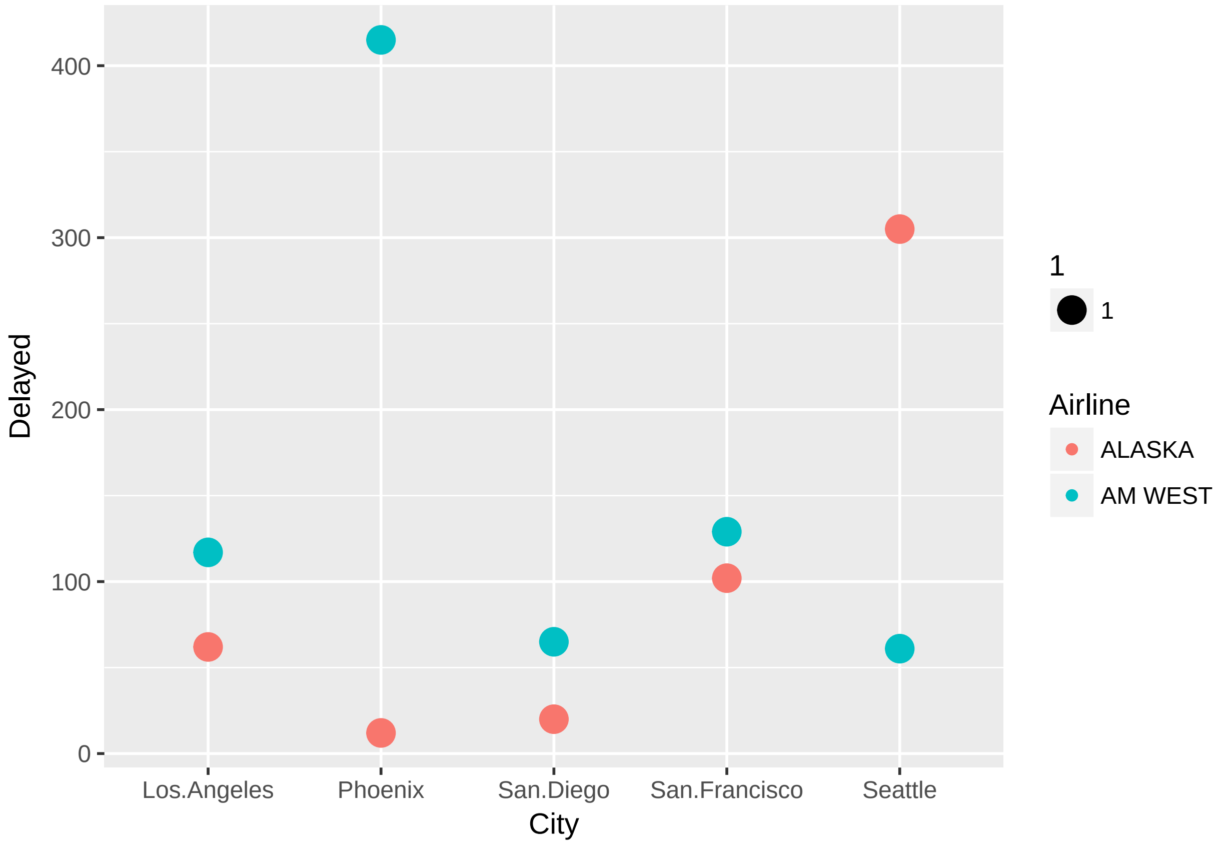
```
##    Airline          City Delayed On Time
## 1   ALASKA   Los.Angeles      62     497
## 2   ALASKA       Phoenix      12     221
## 3   ALASKA     San.Diego      20     212
## 4   ALASKA San.Francisco     102     503
## 5   ALASKA       Seattle     305    1841
## 6  AM WEST   Los.Angeles     117     694
## 7  AM WEST       Phoenix     415    4840
## 8  AM WEST     San.Diego      65     383
## 9  AM WEST San.Francisco     129     320
## 10 AM WEST       Seattle      61     201
```

Analysis and comparison:

The below visualization of the shows comparison of the airlines based on their delayed flights. The plot depicts that in most of the cities AM West Airlines has the larger number of delayed flights except Seattle where Alaska Airline has more delayed flights.

Figure 1:

```
ggplot(FlightData, aes(x = City, y = Delayed)) + geom_point(aes(size = 1,
    color = Airline))
```



Some statistics:

```
Flight_Statistics <- FlightData %>% group_by(Airline) %>% summarise(Avg.Delayed = mean(Delayed),
    `Avg.On Time` = mean(`On Time`), `Total Delayed` = sum(Delayed),
    `Total On Time` = sum(`On Time`), `Total Flights` = sum(Delayed +
        `On Time`), `Percent Delayed` = round((`Total Delayed`/`Total Flights`),
        2), `Percent On Time` = round((`Total On Time`/`Total Flights`),
        2), `Maximum Delay` = max(Delayed), `Minimum Delay` = min(Delayed))

Flight_Statistics
```

```
## # A tibble: 2 x 10
##   Airline Avg.Delayed `Avg.On Time` `Total Delayed` `Total On Time`
##     <chr>       <dbl>         <dbl>           <int>           <int>
## 1  ALASKA       100.2         654.8             501            3274
## 2 AM WEST       157.4        1287.6             787            6438
## # ... with 5 more variables: `Total Flights` <int>, `Percent
## #   Delayed` <dbl>, `Percent On Time` <dbl>, `Maximum Delay` <dbl>,
## #   `Minimum Delay` <dbl>
```

Above data statistics shows that the percentage of delayed flights is higher for Alaska Airlines if all the flights are considered. Therefore if no further analysis is done it is possible to come up with a conclusion that AM West Airline is better since it has lower percentage of delayed flights.

Further Analysis:

Ratio of delayed and on time flights by City:

```
Delyed_Ratio_Cities <- mutate(FlightData, Percent_Delay_City = round(Delayed/(Delayed +
    `On Time`), 2), Percent_ontime_City = round(`On Time`/(Delayed +
    `On Time`), 2))
Delyed_Ratio_Cities
```

```
##    Airline          City Delayed On Time Percent_Delay_City
## 1   ALASKA   Los.Angeles      62     497               0.11
## 2   ALASKA       Phoenix      12     221               0.05
## 3   ALASKA     San.Diego      20     212               0.09
## 4   ALASKA San.Francisco     102     503               0.17
## 5   ALASKA       Seattle     305    1841               0.14
## 6  AM WEST   Los.Angeles     117     694               0.14
## 7  AM WEST       Phoenix     415    4840               0.08
## 8  AM WEST     San.Diego      65     383               0.15
## 9  AM WEST San.Francisco     129     320               0.29
## 10 AM WEST       Seattle      61     201               0.23
##    Percent_ontime_City
## 1                 0.89
## 2                 0.95
## 3                 0.91
## 4                 0.83
## 5                 0.86
## 6                 0.86
## 7                 0.92
## 8                 0.85
## 9                 0.71
## 10                0.77
```

The worst city in terms of delayed flights is San Francisco for both flights, both Airlines have largest delayed flights in San Francisco.Figure 2 and Figure 3 reveal that Alaska Airlines is better in every city compared to AM West Airlines. In every city Alaska Airline has smaller proportion of delayed flights and larger proportion of on time flights.

Figure 2:

```
p1 <- ggplot(Delyed_Ratio_Cities, aes(City, Percent_Delay_City)) +
    geom_bar(aes(fill = Airline), stat = "identity", position = "dodge") +
    labs(title = "Percentage of Delayed Flights by City ", y = "Percentge")

p2 <- ggplot(Delyed_Ratio_Cities, aes(City, Percent_ontime_City)) +
    geom_bar(aes(fill = Airline), stat = "identity", position = "dodge") +
    labs(title = "Percentage of on time Flights by City ", y = "Percentge")

grid.arrange(p1, p2, nrow = 2)
```

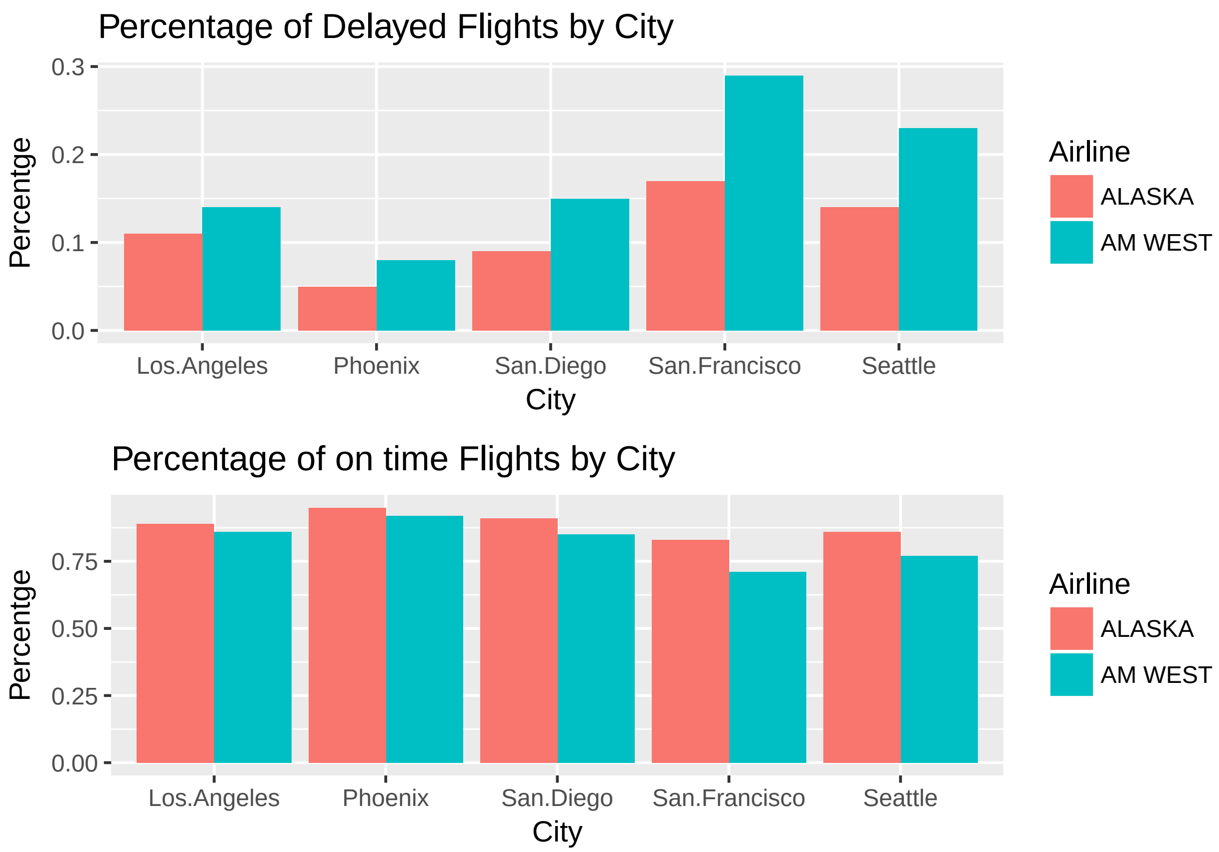
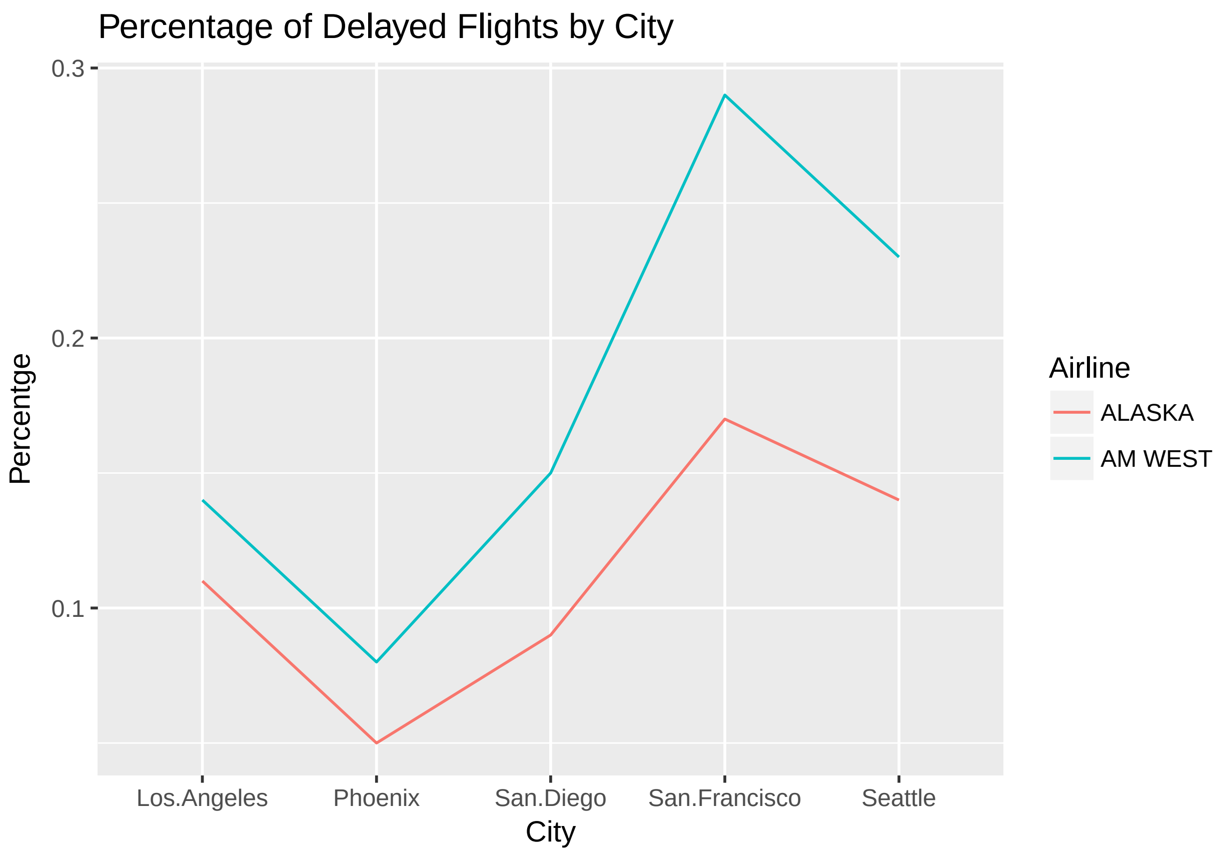


Figure 3:

```
ggplot(Delyed_Ratio_Cities, aes(x = City, y = Percent_Delay_City,
    group = Airline, color = Airline)) + geom_line() + labs(title = "Percentage of Delayed Flights by C
    y = "Percentge")
```



So Alaska Airline is better when the percentage of flights (both delayed and on time) are considered in every city. But AM West appears to be better when all the flights are considered at a time, which suggests that there must be some large values in one or two cities that would explain this discrepancy.

Figure 4 shows that in Phoenix AM West Airline has a huge number of flights compared to what Alaska has in there. Figure 5 shows that Phoenix also has a very large number of on time flights. Since the presence of Alaska Airline in Phoenix is very samll it is obvious that most of those on time flights belong to AM West Airline. Therefore this large number of on time flights in Phoenix affect the overall data in favor of AM West Airline and explains why AM West Airline looks better when the data is seen as a whole.

Figure 4:

```
ggplot(FlightInfo, aes(City, Flight_Counts)) + geom_bar(aes(fill = Airline),
    stat = "identity", position = "dodge") + labs(title = "Flight counts by City ",
    y = "Count")
```

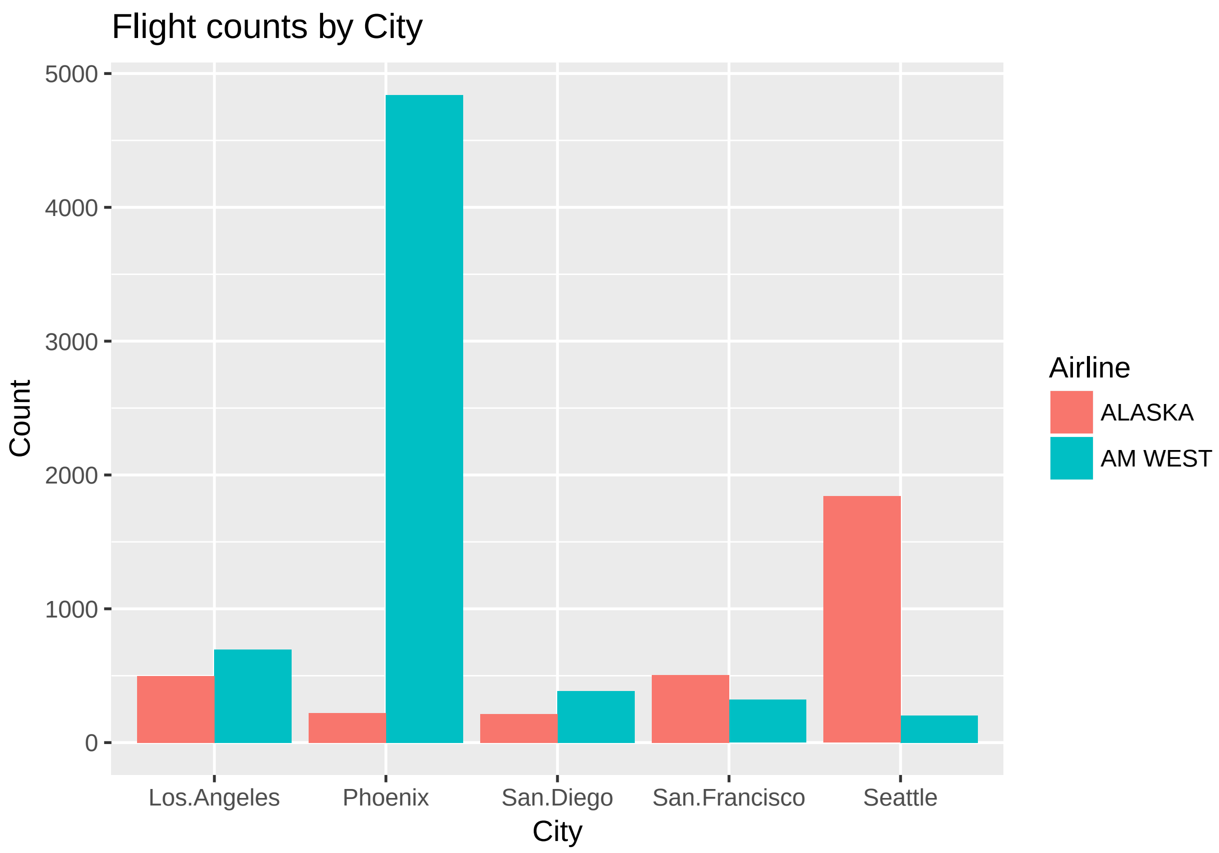
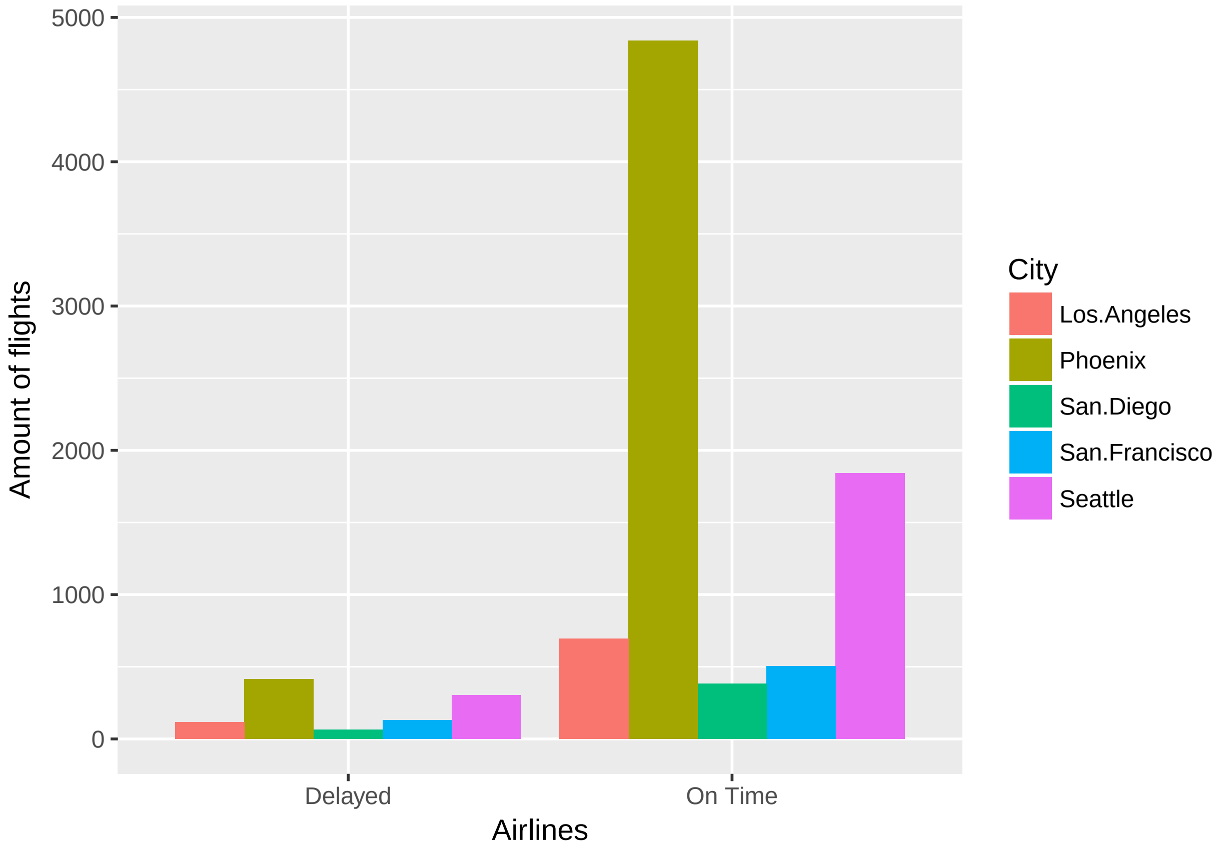


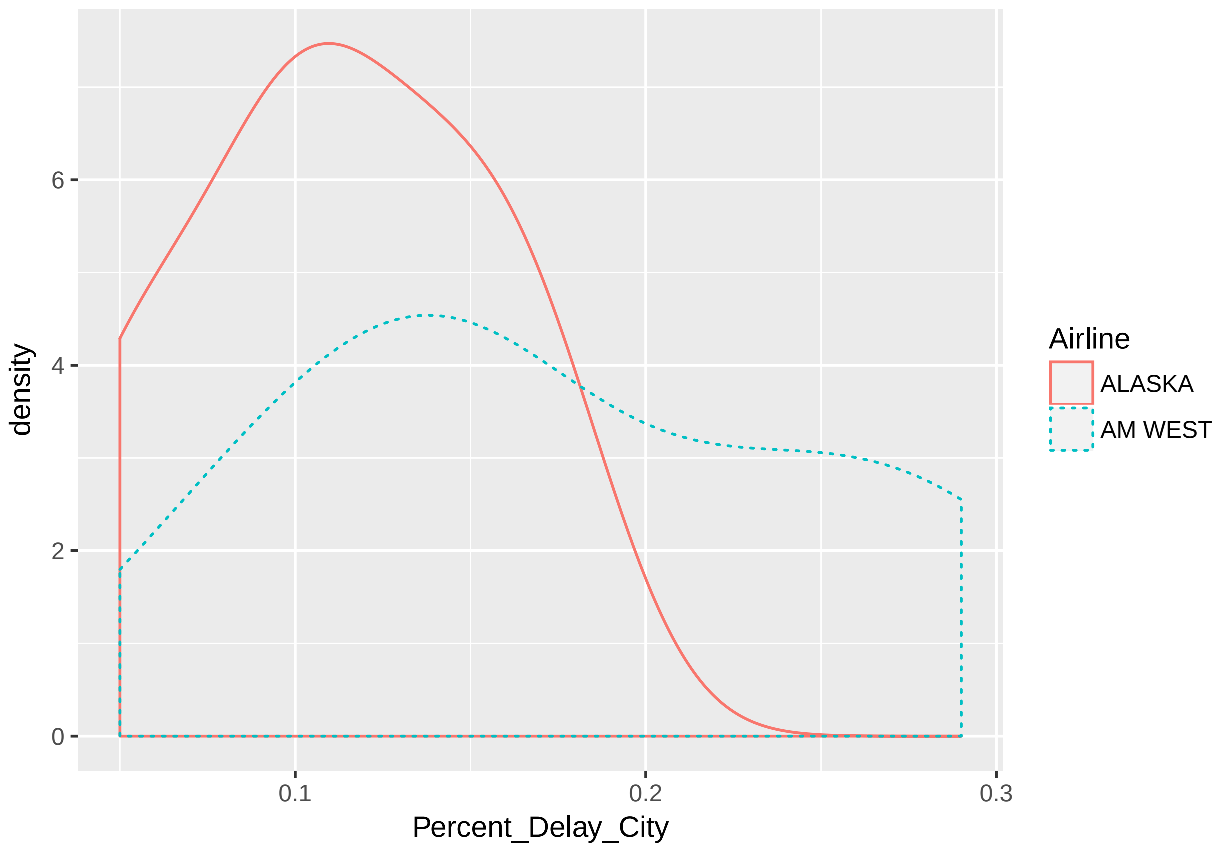
Figure 5:

```
ggplot(FlightInfo, aes(x = Arrival, y = Flight_Counts, fill = City)) +
    geom_bar(stat = "identity", position = "dodge") + xlab("Airlines") +
    ylab("Amount of flights")
```



The density plot below also shows that Alaska Airline is doing better since it has higher density of lower percentage of delayed flights:

```
qplot(Percent_Delay_City, data = Delyed_Ratio_Cities, geom = "density",
    color = Airline, linetype = Airline)
```



Conclusion:

AM West Airline has lower percentage of delayed flights when all the data is considered. But when each city is seperately considered it becomes clear that Alaska Airline performs better and has lower percentage of delayed flights in each city. The huge number of flights of AM West Airline in Phoenix is actually responsible for this false impression that AM West Airline is better (when all the data is considered at a time).