DATA 607 Assignment 5

Adam Gersowitz

2/23/2020

Introduction

This assignment is focused on **Tidying and Transforming** data for analysis. Prior to being transformed this data would be difficult to analyze and work with due to its format and inconsistencies.

Importing the Data

I start by bringing in the .csv file from a github repository and making sure "air" is a dataframe. I make sure to convert all blank cells to null or "NA" values. I do this because functions such as fill will only work with "NA" cells.

```
## Loading required package: bitops
```

Reshaping and Cleaning the Data

After the data has been imported, I begin by naming the airline and flight_status columns as they were blank in the original dataset. Next I remove any lines that don't have data in them. Using the melt function I convert the dataframe from a wide format to a long format which makes it much easier to anlayze. Next, I use the fill function to pull the airlines down to the blank cells below them. Finally, I rename the auto-populated variable and value fields and clean up the City field names. Now I am ready to analyze this data set.

```
##
## Attaching package: 'tidyr'

## The following object is masked from 'package:RCurl':
##
##     complete

##
## 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
```

```
##
## Attaching package: 'reshape'

## The following object is masked from 'package:dplyr':
##
##     rename

## The following objects are masked from 'package:tidyr':
##
##     expand, smiths
```

Reshaping Analyzig the Data

After I have transformed the dataset I will analyze it to determine which airlnes are themost frequently on time and if they have any problems being on time for certain destination cities. First I create an aggregate table of flight information. I then perform a chi-square test via the prop.test function and find that the airlines are significantly different in the proportion of times they are ontime with AM WEST being on time 89% of the time vs 86% for ALASKA.I also see that San Francisco is the destination city that most often has delays of flights. The worst Airline and destination combination is AM WEST and San Francisco at 71% on time. This is somewhat suprising because AM WEST is more often on time then ALASKA. This leads me to belive that the difficulty of having San Francisco as destination has caused this on time percentage to be suprisingly low and in turn has dragged AM WEST overall on time percentage down.

```
## Loading required package: gsubfn

## Loading required package: proto

## Loading required package: RSQLite

##
##  2-sample test for equality of proportions with continuity correction
##
## data:  table(air$total_airline_on_time, air$total_airline_flights)
## X-squared = 16.2, df = 1, p-value = 5.699e-05
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.9 1.0
## sample estimates:
## prop 1 prop 2
##      1      0

##    Airline Flight_Status          City Status_Count status_percentage_city
## 1   ALASKA       on time   Los Angeles          497              88.908766
## 2   ALASKA       delayed   Los Angeles           62              11.091234
## 3   ALASKA       on time       Phoenix          221              94.849785
## 4   ALASKA       delayed       Phoenix           12               5.150215
## 5   ALASKA       on time     San Diego          212              91.379310
## 6   ALASKA       delayed     San Diego           20               8.620690
## 7   ALASKA       on time San Francisco          503              83.140496
## 8   ALASKA       delayed San Francisco          102              16.859504
## 9   ALASKA       on time       Seattle         1841              85.787512
## 10  ALASKA       delayed       Seattle          305              14.212488
## 11 AM WEST       on time   Los Angeles          694              85.573366
## 12 AM WEST       delayed   Los Angeles          117              14.426634
## 13 AM WEST       on time       Phoenix         4840              92.102759
## 14 AM WEST       delayed       Phoenix          415               7.897241
## 15 AM WEST       on time     San Diego          383              85.491071
## 16 AM WEST       delayed     San Diego           65              14.508929
## 17 AM WEST       on time San Francisco          320              71.269488
## 18 AM WEST       delayed San Francisco          129              28.730512
## 19 AM WEST       on time       Seattle          201              76.717557
## 20 AM WEST       delayed       Seattle           61              23.282443
##    ontime_percentage_city delayed_percentage_city ontime_airline
## 1                86.93431               13.065693       86.72848
## 2                86.93431               13.065693       86.72848
## 3                92.21939                7.780612       86.72848
## 4                92.21939                7.780612       86.72848
## 5                87.50000               12.500000       86.72848
## 6                87.50000               12.500000       86.72848
## 7                78.08349               21.916509       86.72848
## 8                78.08349               21.916509       86.72848
## 9                84.80066               15.199336       86.72848
## 10               84.80066               15.199336       86.72848
## 11               86.93431               13.065693       89.10727
## 12               86.93431               13.065693       89.10727
## 13               92.21939                7.780612       89.10727
## 14               92.21939                7.780612       89.10727
## 15               87.50000               12.500000       89.10727
## 16               87.50000               12.500000       89.10727
## 17               78.08349               21.916509       89.10727
## 18               78.08349               21.916509       89.10727
## 19               84.80066               15.199336       89.10727
## 20               84.80066               15.199336       89.10727
##    delayed_percentage_airline total_air_city_on_time_perc
## 1                    13.27152                    88.90877
## 2                    13.27152                    88.90877
## 3                    13.27152                    94.84979
## 4                    13.27152                    94.84979
## 5                    13.27152                    91.37931
## 6                    13.27152                    91.37931
## 7                    13.27152                    83.14050
## 8                    13.27152                    83.14050
## 9                    13.27152                    85.78751
## 10                   13.27152                    85.78751
## 11                   10.89273                    85.57337
## 12                   10.89273                    85.57337
## 13                   10.89273                    92.10276
## 14                   10.89273                    92.10276
## 15                   10.89273                    85.49107
## 16                   10.89273                    85.49107
## 17                   10.89273                    71.26949
## 18                   10.89273                    71.26949
## 19                   10.89273                    76.71756
## 20                   10.89273                    76.71756
```

Conclusion

After reshaping and analyzing this data set I have determined that AM WEST is more often on time then ALASKA airlines and that San Francisco is the destination city that most often leads to delays. To expand on this dataset it would be interesting to get the detail of each flight rather than a summary of on time and delayed flights. It would then be interesting to compare this data to external data such as time of year and weather to determine if those are having more of an impact on one ariline over another. Additionally, it would be interesting to get a more robust dataset with more cities and airlines to determine if these are outliers or if in reality they are close to each other in performance when compared with all airlines.