final_project_draft

Omer Shasman

5/3/2022

Introduction

On Time Performance analysis of an airline network - This is an important metric for the airline that is calculated as the percentage of flights which are delayed by more than 14 minutes while the aircraft arrives at the gate. There are multiple reasons which contribute to the variation in OTP. An analysis of the OTP metric breaking it down into its individual components namely different delay and historical delays can provide insights into how the OTP for an airline can be managed by operational/process changes. The Department of Transport releases the flight level, On Time Performance data. This dataset also has various other factors which affect the Arrival Delay of a flight. An exploratory analysis of this data with the Arrival Delay as the response variable analyzed against different dimensions provided in the dataset can reveal several insights to improve the OTP of an Airline.

What

As part of my Final Project, I am planning to use a subset of OTP data to perform analysis of delays in actual file arrivals focusing on one particular Station and Airline. Since airline operations are very complex, the arrival delays itself can be due to varying factors, like weather delay, carrier delays, security delays, Late aircraft delay... etc or any combinations of any of these in general. My focus is only on 3 types of delays so that I can minimize the complexities in data structures and limit any repeating processes or steps, and rather focus on how to manipulate and do analysis/inference with few variables. Hence I will be considering only 5 years data ranging from year 2014 till 2019 two types of delays "Weather Delays" and "Carrier Delays"

Why

I thought airline is an interesting business with lot of complex operation/data and business itself is most of us are familiar with. Also, with the time constraint we have, there are few site which helped me to get on board and which had the proof of concept directly. Few of the reference links are below : Kaggle
DOT On-Time_performance

This data is presented as yearly file in csv format, I have to merge different years data using dply bind command to append rows at the end and build one file.

How

Use RMarkdown and explore Rfunctions that can integrate some of the topics we learned in the class for flight arrival delay analysis. The following are steps which will be followed as part of the project

- Load data into R using R chunks

- Merge Data using ddplyr
- Filter/melt/massage data using Tidy Data approach
- Use sampling strategy for identifying sample observations.
- Use Stats function to determine mean, median, IQR,...
- Regression Analysis

Body

This project perform analysis of flight arrival delays focusing on one particular Station and Airline. Since airline delays are unavoidable there is always a chance that a particualr flight will be delayed. I think this analysis can be used to further study on why a particular delay happens and if the process/schedule/operations can be enhanced or refined to minimize the delay risk in future flights.

Packages Required

```
library(knitr) ##for printing tables in R Markdown
library(dplyr) ##for data munging
```

```
##
## 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(ggplot2) ## for charts
```

```
library(readr)
X2014 <- read_csv("Data/2014.csv")
```

```
## Rows: 5819811 Columns: 28
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr   (4): OP_CARRIER, ORIGIN, DEST, CANCELLATION_CODE
## dbl  (22): OP_CARRIER_FL_NUM, CRS_DEP_TIME, DEP_TIME, DEP_DELAY, TAXI_OUT, W...
## lgl   (1): Unnamed: 27
## date  (1): FL_DATE
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(X2014)
```

```
## # A tibble: 6 x 28
##   FL_DATE    OP_CARRIER OP_CARRIER_FL_NUM ORIGIN DEST  CRS_DEP_TIME DEP_TIME
##   <date>     <chr>                  <dbl> <chr>  <chr>        <dbl>    <dbl>
## 1 2014-01-01 AA                      2377 ICT    DFW           1135     1144
## 2 2014-01-01 AA                      2378 MIA    TPA           2225     2220
## 3 2014-01-01 EV                      2500 DFW    HOU           2105       NA
## 4 2014-01-01 EV                      2502 CRW    DFW           1655     1805
## 5 2014-01-01 EV                      2502 DFW    CRW           1320     1440
## 6 2014-01-01 EV                      2503 AMA    DFW           1925     1909
## # ... with 21 more variables: DEP_DELAY <dbl>, TAXI_OUT <dbl>,
## #   WHEELS_OFF <dbl>, WHEELS_ON <dbl>, TAXI_IN <dbl>, CRS_ARR_TIME <dbl>,
## #   ARR_TIME <dbl>, ARR_DELAY <dbl>, CANCELLED <dbl>, CANCELLATION_CODE <chr>,
## #   DIVERTED <dbl>, CRS_ELAPSED_TIME <dbl>, ACTUAL_ELAPSED_TIME <dbl>,
## #   AIR_TIME <dbl>, DISTANCE <dbl>, CARRIER_DELAY <dbl>, WEATHER_DELAY <dbl>,
## #   NAS_DELAY <dbl>, SECURITY_DELAY <dbl>, LATE_AIRCRAFT_DELAY <dbl>,
## #   `Unnamed: 27` <lgl>
```

```
#
#X2015 <- read_csv("Data/2015.csv")
#head(X2015)
#
#X2016 <- read_csv("Data/2016.csv")
#head(X2016)
#
#X2017 <- read_csv("Data/2017.csv")
#head(X2017)
#
#X2018 <- read_csv("Data/2018.csv")
#head(X2018)
```

```
# Combine the 5 vectors to a single vector
#all_years <- dplyr::bind_rows(X2014, X2015, X2016, X2017, X2018)
#summary(all_years)
```

Topics From Class

Topic 1:

R Markdown - I will be presenting the project in R Markdown and knit the file to a pdf document. Will be using R chunks extensively to demonstrate and build the project components.

Topic 2:

GitHub - Will host the project in github as source control repository for others to view my project components.

Topic 3:

Sampling strategies for an Observational study - Will be using startified sampling strategy to group the data together(maybe by flight# range) and then use simple random sampling method to create sample data for Topics #4 and #5.

```
# s_data <- dplyr::sample_n(all_years, 10000, replace=FALSE)
# View(s_data)
```

Topic 4:

Detailing Summary statistics (Min. , 1st Qu., Median, Mean, 3rd Qu., Max.) of a variable and plotting graphs using ggplot2

Topic 5:

Regression analysis(on if a particular type of delay has decreased over the time or not) or alternately will be pereforming ANOVA(Analysis of Variance) to identify any mean differences

Conclusion

I designed this project as a way to review some of the topics we learned in the class/homework/assignments to reinforce and refer back some of the materials we learned already. Hence I thought of picking a variety of topics like sampling strategies, summary statistics, ANOVA and regressions will be the best approach and most I can get given the time constraints. If I have more time, I would have included some more topics (like binom, dbinom, geom. . . etc distributions) and see if my dataset have variables that can fit these distributions. Given only a academic background in statistics almost two decades ago, I think this subject has given me much learning experience in statistics and I appreciate how these topics are applicable to find solutions to real world problems.