Playing with domestic airline perfomance data

Kirill

September 04, 2019

Introduction

I found this rather interesting data set at data.gov.au, Domestic Airlines - On Time Performance and I decided to investigate it a bit closer.

First thing first is to download it. Note that `read_csv` from `readr` package can "read" directly from url, but I wasn't sure if everytime I compile html it would re-download the file or use chached version. The data is *Creative Commons Attribution 3.0 Australia* and so no problem in downloading and using the data.

Loading libraries

We are going to use

```
library(tidyverse)
library(knitr)
```

Downloading the data

We are going to use `tidyverse` library that includes several other useful libraries, such as:

- `readr`
- `tidyr`
- `dplyr`
- `ggplot2`

to name a few

Note that we are doing conditional download here, obviously don't want to re-download if we already have the file.

```
fn_data <- "domestic_airline_performance.csv"
fn_notes <- "domestic_airline_performance_notes.txt"
if(!file.exists(fn_data)) {
  url_data <- "https://data.gov.au/data/dataset/29128ebd-dbaa-4ff5-8b86-d9f30de56452/resource/cf663ed1-0
  url_notes <- "https://data.gov.au/data/dataset/29128ebd-dbaa-4ff5-8b86-d9f30de56452/resource/69e214b9-
  download.file(url_data, fn_data)
  download.file(url_notes, fn_notes)
}
df <- read_csv(fn_data, quote = "")
df
```

```
## # A tibble: 80,615 x 14
##    Route Departing_Port Arriving_Port Airline Month Sectors_Schedul~
##    <chr> <chr>          <chr>         <chr>   <dbl>            <dbl>
##  1 Adel~ Adelaide       Brisbane      All Ai~ 37987              155
##  2 Adel~ Adelaide       Canberra      All Ai~ 37987               75
##  3 Adel~ Adelaide       Gold Coast    All Ai~ 37987               40
##  4 Adel~ Adelaide       Melbourne     All Ai~ 37987              550
##  5 Adel~ Adelaide       Perth         All Ai~ 37987              191
```

```
##  6 Adel~ Adelaide       Sydney        All Ai~ 37987                486
##  7 Albu~ Albury         Sydney        All Ai~ 37987                168
##  8 Alic~ Alice Springs  Sydney        All Ai~ 37987                 63
##  9 All ~ All Ports      All Ports     All Ai~ 37987              31913
## 10 Bris~ Brisbane       Adelaide      All Ai~ 37987                155
## # ... with 80,605 more rows, and 8 more variables: Sectors_Flown <dbl>,
## #   Cancellations <dbl>, Departures_On_Time <dbl>, Arrivals_On_Time <dbl>,
## #   Departures_Delayed <dbl>, Arrivals_Delayed <dbl>, Year <dbl>,
## #   Month_Num <dbl>
```

Exploring the data

Now that we've got the data lets explore it. It always helps if we can find more information about the data set, particular what information each column might have.

Working with data

Great, the information above gives us some starting material. However it wasn't that explicit what each column meant and how man columns are there. Let's quickly take a pick

```
d <- df %>% dim
```

total number of observation 80615 and total number of variables 14

There are many ways you can explore this data, but i just want to have a look at the types of Airlines there are.

```
df %>%
  select(Airline) %>%
  distinct() %>%
  arrange(Airline)
```

```
## # A tibble: 13 x 1
##    Airline
##    <chr>
##  1 All Airlines
##  2 Jetstar
##  3 Macair
##  4 MacAir
##  5 Ozjet
##  6 Qantas
##  7 QantasLink
##  8 Regional Express
##  9 Skywest
## 10 Tigerair Australia
## 11 Virgin Australia
## 12 Virgin Australia - ATR/F100 Operations
## 13 Virgin Australia Regional Airlines
```

Cleaning up

I've noticed that there "All Airlines" name in the `Airline` column that appears to have the most number of occurrences in the data

```
df %>%
  group_by(Airline) %>%
  summarise(n = n()) %>%
  arrange(-n)
```

```
## # A tibble: 13 x 2
##    Airline                                   n
##    <chr>                                 <int>
##  1 All Airlines                          21010
##  2 Virgin Australia                      18252
##  3 Jetstar                               11294
##  4 Qantas                                11107
##  5 QantasLink                             9622
##  6 Tigerair Australia                     3982
##  7 Regional Express                       2599
##  8 Virgin Australia Regional Airlines     1655
##  9 Skywest                                 752
## 10 Virgin Australia - ATR/F100 Operations   290
## 11 Macair                                   40
## 12 Ozjet                                     9
## 13 MacAir                                    3
```

Also there one of the routes is `All Ports-All Ports`. Googling for that name didn't reveal any places in Australia by that name.

```
df %>%
  group_by(Route) %>%
  summarise(n = n()) %>%
  arrange(-n)
```

```
## # A tibble: 149 x 2
##    Route                   n
##    <chr>               <int>
##  1 All Ports-All Ports  1505
##  2 Cairns-Brisbane       908
##  3 Brisbane-Cairns       907
##  4 Broome-Perth          907
##  5 Perth-Broome          907
##  6 Hobart-Melbourne      900
##  7 Melbourne-Hobart      900
##  8 Adelaide-Sydney       872
##  9 Sydney-Adelaide       872
## 10 Adelaide-Melbourne    871
## # ... with 139 more rows
```

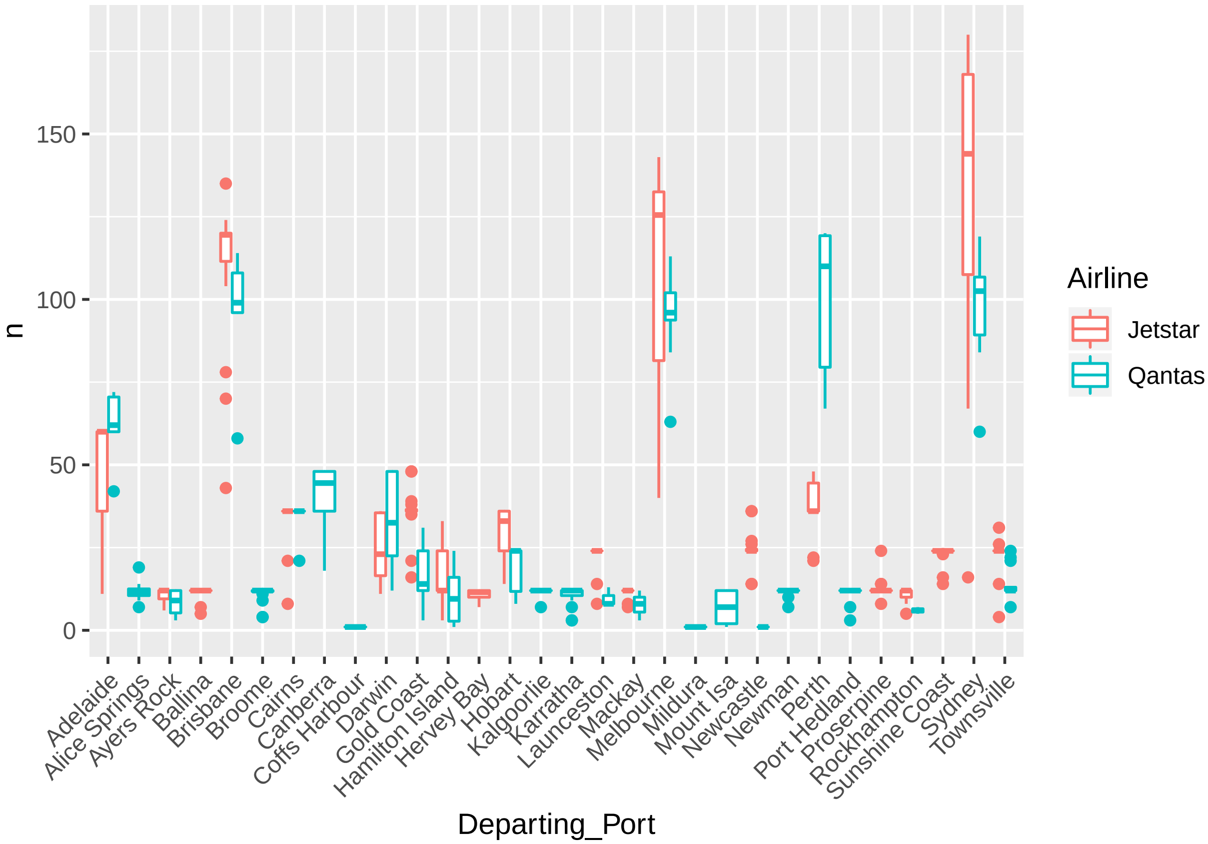
I decide going forward to drop those data points.

```
df2 <- df %>%
        filter(Airline != "All Airlines",
               Route != "All Ports-All Ports")
```

Visualising the data

Here we are summarising so that we have an idea of how many times a particular location had be use per airline per year and we are only going to look at two airlines, Jetstar and Qantas.

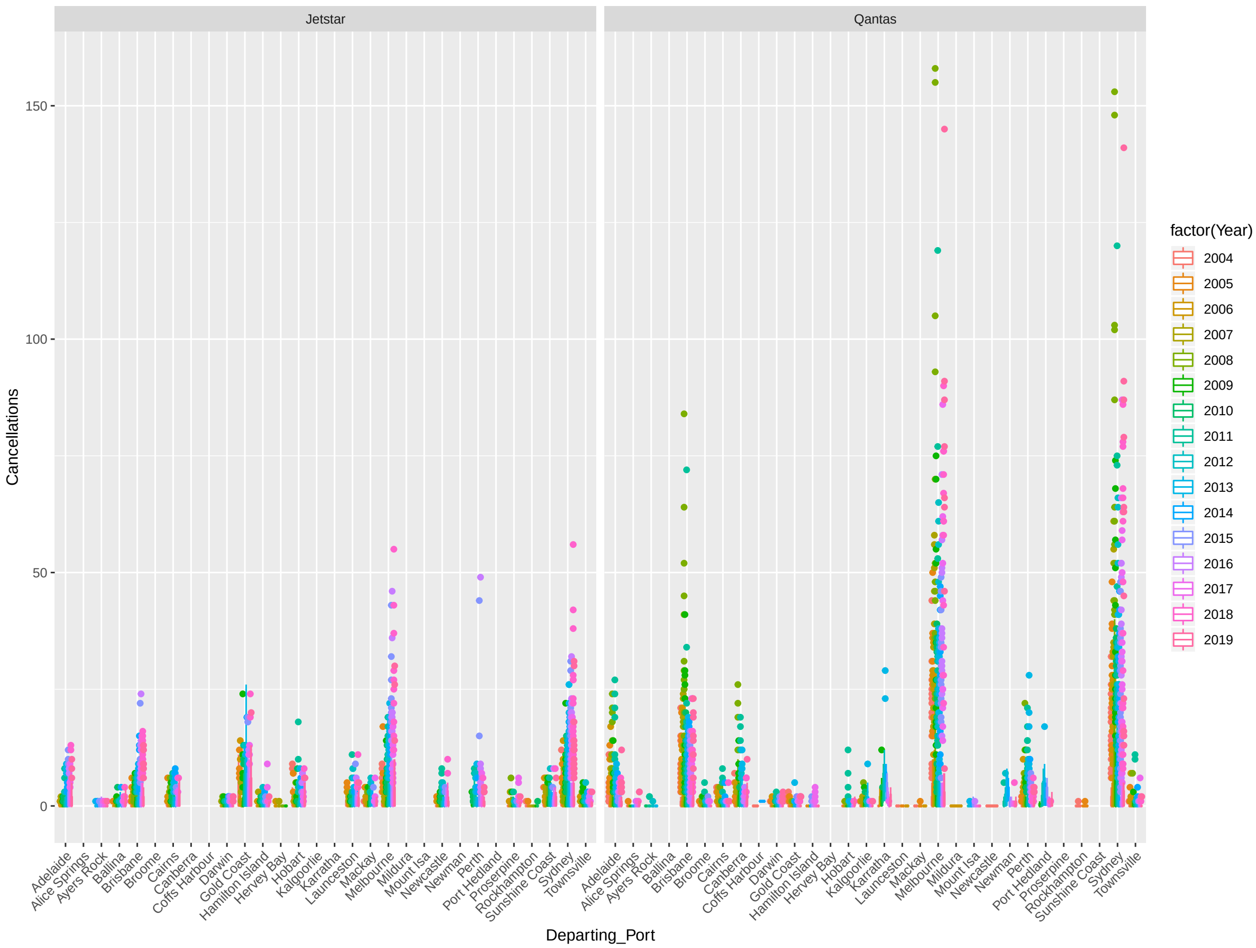
```r
p2 <- df2 %>%
  group_by(Airline, Year, Departing_Port) %>%
  summarise(n = n()) %>%
  ungroup %>%
  filter(Airline == "Jetstar" | Airline == "Qantas") %>%
  ggplot(aes(Departing_Port, n, color = Airline)) +
    geom_boxplot() +
    theme(axis.text.x=element_text(angle=45, hjust=1))
p2
```



In any given year what is the distribution of cancellation

```r
p3 <- df2 %>%
        filter(Airline == "Jetstar" | Airline == "Qantas") %>%
        select(Airline, Departing_Port, Cancellations, Year) %>%
        ggplot(aes(Departing_Port, Cancellations, color = factor(Year))) +
          geom_boxplot() +
          facet_wrap(~Airline) +
          theme(axis.text.x=element_text(angle= 45, hjust=1))
p3
```

```
## Warning: Removed 219 rows containing non-finite values (stat_boxplot).
```



References

- themes