IMT573 Problem Set 2: Data manipulations and plotting (80pt)

Your name:

Deadline: Wed, Oct 21th 5pm

Instructions

This dataset asks you to manipulate the same flights dataset you were using last time, and answer a number of questions on this data. We expect you to use `dplyr`-framework but you are welcome to use something else.

1. Remember that just numerical answer is not enough. Always comment and explain your results. You can add R code inline as

   ```
   In this data we have `r nrow(flights)` flights...
   ```

   (remember: these are backticks!)

2. Be sure to include well-documented (e.g. commented) code chucks, figures, tables, and clearly written text explanations as necessary. All figures should be clearly labeled and appropriately referenced within the text.

3. Be sure that each visualization adds value to your written explanation; avoid redundancy – you dot not need four different visualizations of the same pattern. Don't output irrelevant, or too much of the relevant information. A few figures is helpful. A few thousand figures is just a noise.

1 Work with NYC flights data (2pt)

1. Load the data
2. Ensure you know the variables in the data. Keep the documentation nearby.

2 Explore the data (20 pts)

First, let's do some data exploration. Answer the following questions: show the code, the computation result, and comment the results in the accompanying text.

1. How many flights out of NYC are there in the data?
2. How many NYC airports are included in this data? Which airports are these?
3. Into how many airports did the airlines fly from NYC in 2013?
4. How many flights were there from NYC to Seattle (airport code *SEA*)?
5. Were the any flights from NYC to Bose, ID (BOI)?

6. What about missing destination codes? Are there any destinations that do not look like valid airport codes (three-letter-all-upper case)?

   Hint: check the function *grepl* to do regular expression matching. You may use `"^[[:upper:]]{3}$"` for a regular expression that matches three upper case letters. See this at work:

```
grepl("^[[:upper:]]{3}$", c("12AB", "ABCD", "UX", "SEA"))

## [1] FALSE FALSE FALSE  TRUE
```

   But there are other options.

7. Comment the questions (and answers) so far. Were you able to answer all of these questions? Are all questions well defined? Is the data good enough to answer all these?

2.1 Flights are delayed... (20pt)

Flights are often delayed. Let's look at closer at that.

1. What is the typical delay of the flights in this data?
2. Did you remember to check how good is the delay variable? Are there missings? Are there any implausible or invalid entries? Go and check this.
3. Now compute the delay by destinations. Which ones are the worst three destinations in terms of the longest delay?
4. Delays may be partly related to weather. We do not have weather information here but let's analyze how it is related to season. Do it in two (or more) ways: one graphical, and one in a table form.
5. We'd also like to know how much do delays depend on the time of day. Are there more delays in foggy morning hours? Late night when all the daily delays accumulate? Create a visualization (graph or table) using a different approach than what you did above.
6. Do you see any problems with these questions (and answers)? If you feel a question is not defined well enough, re-formulate it in a more specific way so you actually can answer this question. And state clearly what exactly is your more precise question.

2.2 Let's fly to Minneapolis! (20pt)

Finally, let's see how well did the flights from NYC to Minneapolis (airport code MSP) go.

1. How many flights were there from NYC airports to Minneapolis in 2013?
2. How many airlines fly from NYC to Minneapolis?
3. Which are these airlines (find the 2-letter abbreviations)? How many times did each of these fly to Minneapolis?
4. How many different airplanes arrived from each of the three NYC airports to Minneapolis?

   Hint: airplane tail number is a unique identifier of an airplane.

5. What percentage of flights to Minneapolis were delayed at arrival by more than 15 minutes?
6. And finally answer the question above for each origin airport separately. Is one of the airports noticeably worse than others?

2.3 Think about all this (18pt)

Finally, think about the questions and the analysis.

1. Do you see any issues with data?
2. Ethical concerns?
3. Can these questions be answered? Are these questions meaningful? Do you see potential applications for these answers?