Homework Assignment 1

Your Name Here

Assigned: Oct 24, 2020, Due Sun Nov 01, 2020 11:59PM

HW-1 Instructions

> Our class emphasizes clear communication of data analysis results to non-technical audiences. I expect your HTML output documents to be readable and well formatted. I expect you to put ample comments in your R code to make the code understandable. Along with accuracy of results, I am looking for quality of presentation as well. This homework is due by **11:59PM on Nov 01st**. To complete this assignment, follow these steps:

1. Create a new RStudio Project for this HW. Sync the project directory with a GitHub Repository (see instructions on Canvas on how to do that).
2. Download the `HW1.Rmd` file from Canvas. Save the file in the RStudio Project Directory.
3. Open `HW1.Rmd` in RStudio. Replace the "Your Name Here" text in the `author:` field with your name.
4. Supply your solutions to the homework by editing `HW1.Rmd`.
5. Run your code in the Console and Knit HTML frequently to check for errors.
6. You may find it easier to solve a problem by interacting only with the Console at first.
7. When you have completed the homework please check that your knits correctly when you click `Knit HTML`.
8. "Push" all your local changes to the GitHub Repo for the project.
9. Submit your RMD file, the HTML output and a link to your GitHub Repo on Canvas.

Data frame basics

We will continue working with the nycflights dataset we looked at in class. **Please be sure to keep the data file in the same directory as the RMD file - your project directory.**

First - make sure that you are reading the data correctly and doing all the data cleaning steps that we did in class. Feel free to go beyond our class work in cleaning the data if you feel the need.

I now would like you to answer the following - all in their own separate R code blocks.

Data Exploration

Let's first do some simple exploration of this data.

- How many airlines are there? (Hint: `levels` and `length` can be useful here)

- How many flights there were by the airline with code `OO`? (Hint: `nrow` can be useful here along with logical indexing)
- How long is the shortest flight out of any NYC airport? (Hint: `min` can be useful, remember to handle `NA` values)
- How many flights where there by United Airlines (code: UA) on Jan 12th 2013?

Arrival Delay

Lets focus on Arrival Delay.

- What was the average arrival delay for all airports and all airlines combined in Jan 2013?
- Whats was the median arrival delay for all airports and all airlines combined in Jan 2013?

Based on your answers to the two questions above, what can you say about the distribution of arrival delays? Provide your answer in a text paragraph form.

Airline Performance

Lets see if all airlines are equally terrible as far as flight arrival delays are concerned. For this question you will have to make sure that airline column is coded as a factor.

- Calculate average arrival delays by airline (Hint: look up the command `tapply`)
- Draw a Bar Plot of Average Arrival Delays for all the Airlines (Hint: command for making a Bar Plot is simply `barplot`)
- Which airline has the highest average arrival delay? Which airline has the smallest average arrival delay? Are there airlines that actually have negative average delay? Provide answer to this question in a text paragraph form using **inline R code**.

Air Gain

Create a new column named airgain such that airgain = (departure delay - arrival delay) : this is the amount of delay a flight made up while in air.

a) Explore airgain data - calculate suitable descriptive statistics and appropriate graphics to better understand this data. This part is open ended - you do what you feel works best for you.
b) Answer the questions:

- do airlines actually gain any time when in air on average?
- Calculate average airgain for different airlines - which airlines do a better job, which do a worse job?

Merging Data Frames

> This section and the next is new compared to the class exercise. As you had an opportunity to work together in your breakout rooms for previous questions, this and the next section will carry a higher weight in grading for this HW.

You can get detailed information about the physical planes in our dataset in this file: `planes.csv`. Download and save this file in your project directory.

a) Read the `planes.csv` file using `read.csv` command. Do any data cleaning necessary.

b) Merge the flights data and the planes data using the `merge` command. You should do the merge on the common column named `tailnum`. *getting this right may need some trial and error and getting some help.*

c) Now that you have a merged dataset, think of what interesting questions that you can ask that can be answered using the merged dataset. You are asked to pose five interesting questions and answer them. (For example: who are the top 10 manufacturers of planes that fly out of NYC airports?) **Be creative. Be bold. Ask questions that you would want to know answers to even if you were not doing this for a HW.**

Making Your HTML Look Nice

We want our report to be good looking, professional documents. To that end, I am asking you to do the following:

- Have a floating table of contents
- Include code folding in your output. You can find more about code folding here: https://bookdown.org/yihui/rmarkdown/html-document.html#code-folding

That's it. Once you are done, make sure everything works and knits well and then you can push your changes to the GitHub repo and uplaod the RMD flile and the html output to Canvas.

Have Fun!

Sanjeev