INFO370 Problem Set 1/2: Data manipulations

Your name:

Deadline: Wed, Oct 16th/Tue Oct 23th 2pm

Instructions

This problem set is about exploring and manipulating datasets in python/-pandas. Before beginning this assignment, ensure you have python installed and working. Also, read background material about pandas, I recommend McKinney (2017), chapters 4,5 (numpy and pandas), 7 (data ccleaning), 10 (grouped operations). You can download chapters 3-5 from the first edition on canvas, this broadly corresponds to the same topics.

This problem set is to be submitted twice (marked as two separate problem sets on canvas). We intend to give some feedback between the two submissions. In the first round attempt to solve everything but it is fine if you cannot figure out all solutions. You can still receive full points if some solutions are incomplete and others are missing altogether. In the second round the full points require all answers done.

I recommend you to do this problem set in jupyter notebooks, or in rmarkdown (knitr can rather well include and run python code in rmarkdown, and mix it with R). If neither of it will work for you, you can also write a code file that outputs the question numbers, and write a separate explanatory text.

1. Be sure to include well-documented (i.e. commented) code chucks, figures, tables, and clearly written text explanations as necessary. All figures should be clearly labeled and appropriately referenced within the text. Be sure that each visualization (graph or table) adds value to your written explanation; avoid redundancy – you do no need four different visualizations of the same pattern.

2. Don't output irrelevant, or too much of relevant information. A few figures is helpful. A few thousand figures is only noise.

3. All materials and resources that you use (with the exception of lecture slides) must be appropriately referenced within your assignment. In particular, note that Stack Overflow is licensed as Creative Commons (CC-BY-SA). This means you have to attribute any code you pick from SO (a link to the question/answer webpage will normally do).

4. Remember partial credit will be awarded for each question for which a serious attempt at finding an answer has been shown. Students are *strongly* encouraged to attempt each question and to document their reasoning process even if they cannot find the correct answer. If you would like to include code to show this process, but it does not run withouth errors you can just comment it out and explain what do you attempt to do.

   If you go the knitr/rmarkdown way, you can also set the chunk option `eval` to `FALSE`:

   ````
   ```{python, eval=FALSE}
   a + b # these object don't exist
   # if you run this on its own it with give an error
   ```
   ````

5. When you have completed the assignment and have *checked* that your code both runs and builds correctly, convert the notebook/markdown to HTML, name it 'ps1-YourLastName-YourFirstName.html', and submit the html file on Canvas. Please submit also your original files (original notebook, or code files, or rmarkdown or whatever else do you have).

   html is much easier and quicker to check, but in case of questions, we also want to see your original code.

6. Working together is fun and useful but you have to submit your own work. Discussing the solutions and problems with your classmates is all right but do not copy-paste their solution! First understand it, and thereafter create your own solution. Please list all your collaborators on the solution.

1 Work with NYC flights data (35/65 pts)

Setup

In this problem set you will work with NYC flights data. The data is copied from the corresponding R package, you can read the documentation e.g. at RDocumentation.

1. Load the data
2. Ensure you know the variables in the data. Keep the documentation nearby.
3. Make sure you have read the background readings about pandas (see above).

1.1 Explore the data

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

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

   Hint: I recommend to check out string pattern matching and regular expressions. There are separate versions for base python and for pandas, here you may want to use the pandas versions.
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?

1.2 Flights are delayed...

Flights are often delayed. Let's look closer at the delays.

Try to use the pandas' grouped operations (`groupy`) and aggregation functions when appropriate.

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 typical delay?
4. Delays may be partly related to weather. We do not have weather information in this dataset but let's analyze how it is related to season. Do it in two ways: one graphical and one table. (Feel free to add more if you like).

   I recommend to use matplotlib for plots, but you can opt for something else if you prefer.

   Hint: you may want to create a date variable
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? Or perhaps late night when all the daily delays accumulate? Create a visualization (graph or table) using different tools than what you did above.
6. Do you see any problems with these questions (and answers)?

1.3 Let's fly to Orlando!

Now let's see how is it to fly from NYC to Orlando (airport code MCO).

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

4. How many unique planes fly from NYC to Orlando?

   Hint: airplane tail number is a unique identifier for the plane, similar to car license plate.

5. How many different airplanes arrived from each of the three NYC airports to Orlando?

6. What percentage of flights to Portland were delayed at arrival for more than 15 minutes?

7. And finally answer the question above for each origin airport separately. Is one of the airports noticeably worse than others?

1.4 How big are these planes?

Your final data analysis task is to analyze the size of planes. You need to load the *planes.csv* dataset and merge with the flights data.

1. Load the planes data. What are the variables? How many planes do we have?

2. What would be the *merge key*, the variable that can connect a flight in the flights data with a plane in the planes data?

3. Merge the two datasets.

   How do you want to merge in order to be able to answer the next question?

4. How many flights to Orlando do we have where we don't have the data about number seats in the plane?

5. What was the largest plane (number of seats, manufacturer, and model) that flow to Orlando from NYC?

6. What is the median number of seats to Orlando, grouped by each origin airport?

1.5 Think about all this

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? Can these answers be used for anything useful?

References

McKinney, W. (2017) *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*, O'Reilly Media, 2nd edn.