INFO370 Problem Set 2: Data manipulations

Your name:

Deadline: Wed, Jan 22th 10:30am

Instructions

This problem set is about exploring and manipulating datasets in python/-pandas. Make sure you are familiar with the pandas basics, such as McKinney (2018), chapters 4,5 (numpy and pandas), 7 (data cleaning), 10 (grouped operations).

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, check out how in a separate article. 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. 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. Your results will only count if accompanied with sufficiently and clear explanatory text. Just plain output, with no explanation, will not count.

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

4. 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).

5. Partial credit will be awarded for questions for which a serious attempt at finding an answer has been shown. Attempt each question and to document your reasoning process even if they cannot find the correct answer.

6. As the final submission, you should submit a) code; b) output; and c) explanations. If you do notebooks/rmarkdown, all this will be included automatically but you still have to submit both your original file (which can be run) and a html/pdf version of it (which is much easier to check).

7. 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

1.1 Setup (0 pt)

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).

1.2 Explore the data (15pt)

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. (1pt) How many flights out of NYC where there in 2013?

2. (2pt) How many NYC airports are included in this data? Which airports are these?

3. (2pt) Into how many airports did the airlines fly from NYC in 2013?

4. (2pt) How many flights were there from NYC to Seattle (airport code *SEA*)?

5. (2pt) Were the any flights from NYC to Spokane (GEG)?

6. (3pt) 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 with `Series.str.match()`.

7. (3pt) 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?

1.3 Flights are delayed... (40pt)

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

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

1. (4pt) What is the typical delay of the flights in this data?

2. (4pt) 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 if you haven't done it already.

3. (4pt) Now compute the delay by destinations. Which ones are the worst three destinations in terms of the longest typical delay?

4. (15pt) 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 do more if you consider it useful).

   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. (7pt) 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. (6pt) Do you see any problems with these questions (and answers)?

1.4 Let's fly to San Diego! (20pt)

Now let's see how is it to fly from NYC to San Diego (airport code *SAN*).

1. (1pt) How many flights were there from NYC airports to San Diego in 2013?

2. (1pt) How many airlines fly from NYC to San Diego?

3. (2pt) Which are these airlines (find the 2-letter abbreviations)? How many times did each of these go to San Diego?

4. (2pt) How many unique planes fly from NYC to San Diego?

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

5. (2pt) How many different airplanes arrived from each of the three NYC airports to San Diego?

6. (5pt) What was the average flight duration to San Diego (arrival minus departure time)? How long was the slowest as the fastest flight?

   Compare your results with those you find in flight schedules. What do you find?

7. (4pt) What percentage of flights to San Diego were delayed at arrival by more than 15 minutes?

8. (3pt) And finally answer the question above for each origin airport separately. Is one of the airports noticeably worse than others?

1.5 What are these planes? (20pt)

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

1. (1pt) Load the planes data. What are the variables? How many planes do we have?
2. (2pt) 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. (3pt) Merge the two datasets.

   How do you want to merge in order to be able to answer the next question?
4. (6pt) Were there more Airbus 320-series or more Boing 737 planes flying from NYC to San Diego?

   Note: while Boeing names it's 737-series planes like "737-xxx", Airbus calls these "A318", "A319", "A320" and "A321". (A300, A310, A330 and others are different planes). You have to find a way to count these different naming schemas.
5. (3pt) How many flights to San Diego do we have where we don't have the data about number seats in the plane?
6. (3pt) What was the largest plane in terms of seats that flow to San Diego from NYC? Tell us it's number of seats, manufacturer, and model.
7. (2pt) What is the median number of seats to San Diego, grouped by each origin airport?

1.6 Think about all this (5pt)

Finally, think about the questions and the analysis.

1. (1pt) Do you see any issues with data?
2. (2pt) Ethical concerns?
3. (2pt) Can these questions be answered? Can these answers be used for anything useful?

References

McKinney, W. (2018) *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*, O'Reilly Media, 2nd edn.