Final Project

With a partner, select a substantial data set, one that will allow multiple avenues of exploration.

Here are some suggested sites for finding data:

1. www.kaggle.com
2. https://zenodo.org/
3. www.data.world
4. data.nasa.gov
5. www.datahub.io
6. cloud.google.com/datasets
7. catalog.data.gov/dataset/
8. archive.ics.uci.edu/ml/datasets.php

Scoring

- Presentation = 30 points Either 6th or 11th Dec
- Write-up = 70 points Due by noon on Monday, 18th Dec.

The analysis

Do all necessary data cleaning

- Any funky rows?
- Duplicate rows:
    - How many duplicate rows?
    - Do a drop w/ keep = first
- Outliers:
    - What method are you using?
    - How many outliers and in which columns (number per column)?
- Missing data
    - How many and which columns (number per column)
    - How did you deal with them?

Looking at the columns of data, develop some initial ideas of what you might learn from the data.

Do the univariate analyses to understand the properties of the individual columns

- Graph each column vs the index
    - What can you learn from each graph?

- Do histograms for each numeric column
    - What can you learn from each graph?
- Do bar charts showing the counts of each value for each categorical column (.value_counts() will help with this)
    - What can you learn from each graph?
- Any other graphs that are informative
    - What can you learn from each graph?
- Descriptive stats
    - Which columns are normally distributed (or close to being normally distributed)?
    - Which are not normally distributed?
    - How much skew and kurtosis in the columns?
    - What can you learn from them?
- If your data is time-series data
    - What aggregations can you do?
        - What do you learn from them?
    - What do you learn from looking at the data on different time scales?

Do bi-variate analyses

- Do pair-wise plots for all numeric columns (the pair- plot in the seaborn package will help with this)
    - What can you learn from them?
- Do correlations between all numeric columns
    - Pearson
    - And either Spearman or Kendall
    - What can you learn from them?
        - Which columns are highly correlated?
- Do crosstabs w/ and w/o seaborn's visualization on them.
    - What can you learn from them?
- Do pivot tables w/ and w/o seaborn's visualization
    - What can you learn from them?

Do the multivariate analyses to understand how columns relate to each other

- Regression
    - Which columns or sets of columns show a linear relationship?
- Clustering
    - Do you see any patterns?
    - What can you learn from this?
- Can you apply desirability functions to learn anything?
- Any other analyses that can used? What do you learn from them?

The presentation

Give a 15 minute presentation to the class on either 6th or 11th Dec. Everyone in the group should be involved in the presentation.

The presentation should include:

1) Describe the data: what is it? where did you get it from?

2) Are there any distinguishing features about the data: are there distributions that are skewed towards something interesting?

3) Were there any significant problems encountered? If so, how did you solve them?

4) Were there any critical decisions to be made about the data? If so, what and why did you make the decisions you did?

5) What story did the data tell? What were the most informative analyses and their results? What did you learn from doing the analysis?

The write-up

Individually, write a complete report due at noon, Monday, 18th Dec. At a minimum, the write-up must include:

- An abstract summarizing what you did and what you found
- Data
  - The file name
  - Where you got it from
  - A description of the data (can be from the site you got it from)
- Results
  - Everything that you did (see above) including the graphs and tables
- Discussion
  - What you learned from the analysis
- Summary
  - Include what you learned from doing the analysis

The write-up should be as long as needed. If it ends up being 20 - 30 pages, that's fine, it's all electronic (no paper) and will mostly be graphs with descriptions. **Completeness is a virtue when it comes to the write-up.**