Submission [10 Points]

For replication purposes, upload your data using the Box link that I will email to you following these **specific instructions**:

1. Create a folder with your first and last name with a brief description.
   Ex) "Julian Oolman - An Example"
2. Upload downloaded data set(s). That is, the one(s) used at the beginning of your notebook.
3. Upload your final cleaned data set using the **pickle format**.

In your Jupyter Notebook, `Restart Kernel and Run All...` Then on Compass submit both

- the Jupyter notebook
- a PDF export

Content	Full Points		No Points	
Uploaded raw data	2	yes	0	no
Uploaded final data as .pkl	2	yes	0	no
Restarted kernel and ran all	2	yes	0	no
Submitted Jupyter notebook	2	yes	0	no
Submitted PDF export	2	yes	0	no

Format [15 Points]

Content	Full Points		No Points	
Removed all non-essential code	3	yes	0	no
Figure Setup				
Increase size of figures, axis labels, title, and ticks with `sns.set()`	4	1 per item	0	none
Typos/grammatical errors	3	none to some	0	many
Section headers for: Introduction, Data Wrangling and Cleaning, Label Figure, Feature Transformations, and Feature vs Label Figures	5	1 per header	0	none

Introduction [5 Points]

Provide a paragraph containing your prediction problem, motivation, and context in whichever order suits you best. Improve upon what you wrote for the proposal. Do not copy and paste from proposal; we can always improve our writing. You will need to have compelling writing in your career. This is good practice.

Data Wrangling and Cleaning [20 Points]

1. Provide the name(s) and source(s) from where you retrieved your data
2. Wrangle and clean your data into its form for analysis.
   - Use an excessive amount of markdown cells and/or comments to describe what you are doing.
   - Ensure variables are correct data type.
   - Use hierarchical indexing when appropriate.
   - Deal with any codes for missing data.
   - Remove observations when necessary.
   - Ensure data columns and rows are correctly formatted for analysis.

3. On your final data set, output the head, shape, info, and describe.
4. Use `.to_pickle()` to save your data once it is in a form appropriate for analysis.

Content	Full Points		No Points	
Name(s) & Source(s)	2	1 point per	0	missing
Wrangling & Cleaning	12	2 points per bullet	0	missing
head, shape, info, describe	4	1 point per item	0	missing
Save as pickle	2	yes	0	np

Label Figure [15 Points]

If your label is...

- Discrete with...
  - Two classes
    - **Plot(s)** - produce a histogram
    - **Response** - Do you care more about sensitivity, specificity, or accuracy? Why?
  - Three or more classes
    - **Plot(s)** - produce a histogram
    - **Response** - Is there a class that is more important to predict accurately? If so, what is it and why? If not, why not?
- Continuous with a structure of...
  - Cross-sectional or panel
    - **Plot(s)** - *In one figure*, produce a histogram of your label **and** log transformed label
    - **Response** - Is a log transformation appropriate?
  - Time series
    - **Plot(s)** - *In one figure*, you will have **up to three plots**: (1) plot your label vs time, (2) if you have an exponential trend, plot log(label) vs time, (2 or 3) if you have a trend, plot difference of the (log if needed) of your label.
    - **Response** - Do you need a difference of your label or a difference of a log transformation of your label?

Content	Full Points		No Points	
Plot(s)				
All plots in one figure	5	yes	0	no
Appropriate titles	2	yes	0	no
Labelled axis	3	yes	0	no
Response	5	reasonable	0	not

If a transformation is necessary, then add the transformed label to your data frame and `.drop` the original label.
If you have a time series, your lagged features should be from your transformed label.

Feature Transformations [15 Points]

In one figure with a grid layout and dimension of your choosing, produce histograms of your continuous features. Do any of these features need transformations?

Hints:

1. You may need to adjust `figsize`
2. Monetary (ex. income) or count (ex. population) features are more likely to need transforming

Content	Full Points		No Points	
Plot(s)				
All plots in one figure	5	yes	0	no
Appropriate titles	2	yes	0	no
Labelled axis	3	yes	0	no
Transformations	5	reasonable	0	not

Feature vs Label Plots [20 Points]

You will produce two figures using your label (transformed if applicable):

Figure 1. Regardless of your type of label, produce a hist kind of pair-plot with the plot and diagonal bins set to 10.

Response: Briefly describe any relationships you see between your features (ex. linear, quadratic, logarithmic, etc.).

Figure 2. Using at least two plots of your label versus a feature from the list below, produce a figure.

- **Response:** Briefly describe what you see.

Options:

1. **Discrete vs. Discrete Variables**: Use a bar plot with the x-axis set to your label and the hue set to the discrete feature.
2. **Continuous vs. Discrete Variables**: Use a violin plot with the x-axis set to the discrete variable.
3. **Continuous vs. Continuous Variables**: If you are uncertain about a plot from the pairplot, use a binned scatter plot with 100 bins.

★ You may use an alternative plot if it better illustrates the relationship.

Content	Full Points		No Points	
Figure 1				
Hist pairplot with 10 bins	6	yes	0	no
Response	4	reasonable	0	not
Figure 2				
$\geq$ 2 listed plots in one figure	2	yes	0	no
Appropriate titles	2	yes	0	no
Labelled axis	2	yes	0	no
Response	4	reasonable	0	not