Exploratory Data Analysis (EDA) for Data Science projects

Template

Please use the Jupyter notebook available on Canvas. You can use insights from the Exploratory Data Analysis as input for the subsection data description under the methodology section in the final thesis. There is no need to create a *story line*. You just show your understanding of the data.

Why?

In your thesis you will work with data. In most cases you will *predict* the value on one variable based on the values of (lots of) other variables. The aim of your EDA is to *understand* your data. Because with this understanding you can better analyze, debug, explain, improve, ... your system. Everything you do in your EDA comes from this desire to *understand*. You focus on those aspects of your data which are central to your research questions, to your methods, and to your desired outcomes.

Data cleaning

EDA and data cleaning go hand in hand. While doing the EDA you discover anomalies, missing values, outliers, mistakes (typos/..) that can be repaired.

Provenance

Of course, you clean data and repair mistakes, but make sure it is totally transparent. You should log everything *what* you do, *why* you do it, and *what effect* it has on your corpus.

What?

You want to know your data at several levels:

- Corpus
- Variables by themselves
- Interaction between variables

Corpus

- How many instances, how many variables (df.shape)
- Number of missing values for each variable
- Type of each variable
- If applicable: clusters
- ...

Pre variable

- Sometimes you want to describe your corpus also before you actually have variables that you can give to scikit learn
- Time
- If data is text or video: lengths, all kind of counts (voc size eg, hapax count, et cetera)

Univariate analysis

- How does the population look when you focus on just one aspect (variable)
- df.describe()
- sns.boxplot
- sns.displot , histograms, kde (possibly factored by values of the to be predicted variable)
- priors

Baseline
Based on this analysis you should be able to determine baseline scores for your predictor. If we know, for example, that male/female ratio is .8 in our corpus, it is easy to make a classifier with 80% accuracy without any thinking, machine learning or anything. You will need to beat that figure, of course. You want to know and understand that number long before you start

Multivariate analysis

- How do variables interact? Often that means, how do they *correlate*?
    - pd.crosstabs , groupby, pivot tables for categorical variables
    - correlation for numerical ones
- sns.pairplot
- Should you remove variables or not?

How?

EDA can be done brilliantly in pandas combined with seaborn. Read section 4.14 of the Data Science Handbook for ideas. Create one or a few EDA/cleaning/normalization notebooks with good markdown explanations, simple nicely refactored code, and great looking visuals

Normalization
Normalization is something that may be needed. It should be based on a good understanding (=EDA) of your data.

And my thesis?
The EDA provides the necessary input for the data description part of the methodology section in your master thesis. Besides the need of understanding, it is good to keep the advantages below in mind:

- The EDA is often the most useful and used part of a thesis from the perspective of the company
- It is easy and nice to write part of your thesis with cool (why not even interactive) visuals
    - Always nice to do when you have that writers block
    - Impressive visuals for your thesis cover and possible presentations
    - Without a deep understand your data, you can never satisfactorily justify that you made the right choices in your thesis and this affects your final grade