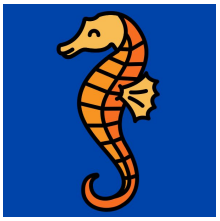
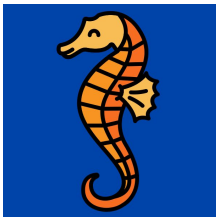


EDA Checklist

There are no real hard rules on how to do Exploratory Data Analysis thus there is yet no software to do this to the quality level humans would. To get good at EDA requires a lot of practice. There are some things to look at and keep in mind.

1. Understanding the problem and the data.. description and domain
2. Looking at the data information, check the values, check the summaries
3. Make some assumptions (this should be done before you look at the data in depth so that you are not biased)
4. Check for missing values... if there are then deal with the missing values
5. Check for distributions, boxplots/violin plots, histograms and scatterplots
6. Check for extreme values, are there outliers or do they make sense from a domain perspective
7. After removing outliers... replot the data...
8. Do not forget what are the 5 things you can look at in a data feature:
    a. appearance of groups
    b. skewness
    c. appearance of unexpected values
    d. where are the data values centered
    e. how widely are the values separated (does this fit with your domain knowledge)
9. Check for correlations/ heatmap
    a. do the correlations make sense? or could they be due to cofounding factors? what could though factors be (domain knowledge helps)
10. At all times your aim is to try to verify your assumptions
11. Clean up... feel free to remove plots that are redundant/ non-informative
12. Add explanations and your thought process... write it all down... first on paper and then in notebook

The next step after EDA would be predicting the target variable.



EDA Checklist

1. For this value and the business problem you are solving, fix the dataset (in your case you have the data) and find the right evaluation metric

2. Holding out some of the data

Here before doing much one should split the data in train / test.. depending on the amount of data a 90(train)/10(test) or 80/20 split should be fine. It is important to have enough train data.

3. Feature engineering:

On the train data you can now do the analysis you need to figure out what would be the good features for your model.

4. Train

There are two ways to train.. just train or do cross validation if you have time.

5. Evaluate

Evaluation happens then on the holdout test set. And results are also compared to the the train/CV results.