Exploratory data analysis (EDA) task overview

- A quick explanation of an EDA task: https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15
- A free course from CognitiveClass.AI: https://cognitiveclass.ai/learn/data-science-with-python

Data and domain introduction

When you only start to work with some data, you probably are not very familiar with the domain this data came from.

- What kind of relations might be meaningful to search for?
- What statistical tests are applicable here?
- What kind of distances make sense in this domain?

And so, when you do a report for your exploratory data analysis it is very good practice to do an introduction of the data you are working with.

Examples of the notebook with good domain introduction:

- https://www.kaggle.com/tanulsingh077/prostate-cancer-in-depth-understanding-eda-model
- https://www.kaggle.com/pmarcelino/comprehensive-data-exploration-with-python

Please pay attention here, that these notebooks are given only for demonstration of good data and domain introduction and description.

No need to waste too much time on them.

Data types: continuous, discrete, ordinal, nominal

Data usually contain some number of features, something we know about objects in our domain.

These features might be different in terms of their quality/quantity properties.

There is a difference between, for example, **education** and **body weight**, because the first feature might have only a few options whereas the second one might have hundreds or even thousands of unique values, depending on the precision of measurements.

This is a very important topic that needs to be taken seriously because it directly affects the choice of pre-processing method, statistical tests, visualizations, etc,

so the section is about different feature types.

The page with data types description:

- https://towardsdatascience.com/data-types-in-statistics-347e152e8bee

Videos from Coursera:

- https://www.coursera.org/lecture/clinical-research/data-types-wP55H
- https://www.coursera.org/lecture/clinical-research/arbitary-classification-nominal-categorical-data-ffRYP
- https://www.coursera.org/lecture/clinical-research/natural-ordering-of-attributes-ordinal-categorical-data-DstjM
- https://www.coursera.org/lecture/clinical-research/measurements-and-numbers-numerical-data-types-lDOdW
- https://www.coursera.org/lecture/clinical-research/how-to-tell-the-difference-discrete-and-continuous-variables-0VkAn

Video from Udacity course:

- Types of Data Quiz 1 - Intro to Machine Learning

Some explanation of cardinality and binning:

- https://www.youtube.com/watch?v=NguHPzUqvBc
- https://www.youtube.com/watch?v=iv_ec0EfXcE

Plots: histogram, scatterplot, boxplot, heatmap, pie chart

One of the most significant parts of EDA is data visualization.

Lots of different kinds of plot exist out there and it is not always clear which one suits better for your particular case.

This section gives some useful links to review.

- The page with an overview of various plot types: https://python-graph-gallery.com/
- Data visualization course on Kaggle: https://www.kaggle.com/learn/data-visualization
- Storytelling Through Data Visualization course on Dataquest: https://app.dataquest.io/course/storytelling-data-visualization

Univariate analysis

Anomalies in data

This section only covers some basic (or most frequently used) methods to determine outliers based on separate features.

- Some basic ways to detect anomalies in data: https://towardsdatascience.com/5-ways-to-detect-outliers-that-every-data-scientist-should-know-python-code-70a54335a623
- Median absolute deviation (MAD): https://medium.com/james-blogs/outliers-make-us-go-mad-univariate-outlier-detection-b3a72f1ea8c7
- An article about adjusted Boxplot rule: https://wis.kuleuven.be/stat/robust/papers/2008/adjboxplot-revision.pdf

Gaps in data

It is often the case that the data is incomplete, it contains gaps.

There might be a lot of different reasons for it, both technical and business-related.

Some algorithms (actually the majority of them) cannot work with data that contains gaps.

This section gives some overview of how this issue might be handled.

Some good examples of code with Pandas:

- https://www.youtube.com/watch?v=EaGbS7eWSs0
- https://www.youtube.com/watch?v=XOxABiMhG2U

An article with a description of various handling missing data ways:

- http://ceur-ws.org/Vol-2136/10000108.pdf

Multivariate analysis

Continuous features: correlation coefficients

Of course, there was already some coverage of these topics in the Basic Statistics module.

Nevertheless, it is worth mentioning it also here because searching for highly correlated feature pairs is a very important part of EDA.

One reason is the so-called multicollinearity problem:

https://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/

Another reason is that there might be possible features that are highly correlated with the target variable, so it is also very important to find them before actual modeling.

So, this section gives several links with some correlation coefficient explanation.

If you are already familiar with this topic, good, do not waste much time on it. Just remember to take this topic into account when doing your EDA.

- Correlation and dependence: [pdf - 8 slides];
- Pearson's correlation simply explained with Python code examples [article];
- Pearson's, Cohen's, Spearman's rank, Kendall's rank correlations [article].

Discrete features: co-occurrences, crosstabs

As it was shown in one of the previous sections, there are several kinds of features.

This section covers some topics about the analysis of discrete features.

The main topic covered here is the so-called crosstab.

It allows us to take a look at various combinations of certain categories from different discrete features and thus get some insights about their nature and gather some interesting statistics.

https://www.youtube.com/watch?v=jeNnL-Cj20w

Crosstab explained:

- https://pbpython.com/pandas-crosstab.html
- https://medium.com/@yangdustin5/quick-guide-to-pandas-pivot-table-crosstab-40798b33e367
- https://www.w3resource.com/pandas/crosstab.php