HyperionDev

Exploratory Data Analysis

Task

Visit our website

Introduction

This task introduces key components of exploratory data analysis along with a few examples to get you started on analysing your own data. By the end of this task, you will be conducting an exploratory data analysis on a data set.

What is exploratory data analysis?

As the term suggests, exploratory data analysis (EDA) has to do with exploring a dataset. It has been defined as “a philosophy as to how we dissect a data set; what we look for; how we look; and how we interpret” (NIST, n.d.).

In simpler terms, EDA is a method used to gain a good level of understanding of the data. This typically means normalising and/or standardising the data, and generating some graphs to show certain statistical attributes. Recall that Machine Learning methods sometimes make assumptions about the underlying statistical properties of data: EDA is your way to determine what these properties are.

Before the use of EDA, critical information and insights that could be drawn from data were being overlooked. John W. Tukey, the developer of EDA, believed that more emphasis needed to be placed on using data to suggest hypotheses to test. Tukey based his method on the following principle, “It is important to understand what you can do before you learn to measure how well you seem to have done it” (Tukey, 1977). With this in mind, some of the main objectives of EDA are:

- Identify and understand patterns, anomalies and relationships in the data
- Identify outliers, errors and inconsistencies in the data
- Prepare data for further analysis

Since EDA has to do with exploring a data set, there isn’t a prescribed method or set of steps for doing EDA. It’s an iterative process. However, there are a number of questions that you should generally consider during EDA. These include but are not limited to (Bourke, 2019):

- What kind of data do you have? Is your data numerical, categorical or something else? How do you deal with each kind?
- What’s missing from the data and how do you deal with it?
- Where are the outliers and should we pay attention to them? An outlier is a piece of data that doesn’t seem to fit with the other data in the dataset. You need to know where the outliers in your data are. Do you need them or are they damaging your model? Are they a representation of a real-world scenario, or just a problem with data entry or data collection?
- How can you add, change or remove features to get more out of your data?

Extra resource

The additional reading for this task is a page in the Causal Analysis/Diagnosis Decision Information System (CADDIS) developed by the United States Environmental Protection Agency. All the sections of this page are relevant except for the final two (Multivariate Data and Mapping Data). This resource contains excellent further information about EDA and illustrates how important EDA is as a first step in data analysis. Pay close attention to how simple visualisation of data can lead to important insights being drawn before any other analysis is done. This resource also provides some additional software tools which can be used to experiment with the different techniques highlighted.

Access the additional reading at: **Exploratory Data Analysis | US EPA**

There are many techniques that you can use during an EDA. However, it is often much easier to explore data when using visualisations. Therefore, various visualisations are often created during EDA. For example, what visualisation could you use to identify outliers? What visualisations could you use to detect patterns in the data?

Ultimately, there's no limit to the number of experiments one can perform in the EDA process – it completely depends on the data you're analysing, as well as your knowledge of packages such as pandas and matplotlib.

Instructions

Before attempting the task, read the Jupyter Notebook (**Exploratory Data Analysis.ipynb**) that accompanies this task.

Practical task

Create a Jupyter notebook called **titanic.ipynb**. Use this notebook to create an in-depth EDA on the Titanic dataset provided. Your EDA should contain descriptions of each step and appropriate visualisations. Use the **metadata** for the Titanic dataset to familiarise yourself with the data.

Use the following guiding questions for your EDA:

- What is the most important factor in determining whether a person was likely to survive the sinking of the Titanic?
- In the movie, the upper-class passengers were given preference on lifeboats. Does this show in the data?
- “Women and children first”. Was this the case?
- Add one other observation that you have noted in the dataset.

Important: Be sure to upload all files required for the task submission inside your task folder and then click "Request review" on your dashboard.

Share your thoughts

Please take some time to complete this short feedback **form** to help us ensure we provide you with the best possible learning experience.

Reference list

Bourke, D. A Gentle Introduction to Exploratory Data Analysis. Retrieved January 13, 2019, from towardsdatascience.com: **https://towardsdatascience.com/a-gentle-introduction-to-exploratory-data-analysis-f11d843b8184**

NIST. (n.d.). What is EDA? Retrieved May 6, 2019, from Engineering Statistics Handbook: **https://www.itl.nist.gov/div898/handbook/eda/section1/eda11.htm**

Wikipedia. Exploratory data analysis. Retrieved April 27, 2019, from Wikipedia: **https://en.wikipedia.org/wiki/Exploratory_data_analysis**