Hyperiondev

TASK

Exploratory Analysis

Visit our website

Introduction

WELCOME TO THE EXPLORATORY ANALYSIS TASK!

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.

Get in touch

Connect for support

Remember that with our courses, you're not alone! You can contact an expert code reviewer to get support on any aspect of your course.

The best way to get help is to login to Discord at **https://discord.com/invite/hyperdev** where our specialist team is ready to support you.

Our team is happy to offer you support that is tailored to your individual career or education needs. Do not hesitate to ask a question or for additional support!

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.

NIST also highlights what we try to accomplish when conducting EDA. “The primary goal of EDA is to maximise the analyst's insight into a data set and into the underlying structure of a data set, while providing all of the specific items that an analyst would want to extract from a data set, such as:

- a good-fitting, parsimonious model
- a list of outliers
- a sense of robustness of conclusions
- estimates for parameters
- uncertainties for those estimates
- a ranked list of important factors
- conclusions as to whether individual factors are statistically significant
- optimal settings ” (NIST, n.d).

Simply then, EDA is what one does when you explore data, find patterns and extract insights. It allows you to understand your data and gain a general idea of the length, depth, range and scope of your data.

Photo on **Unsplash**

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 basically 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, a Handbook by the National Institute of Standards and Technology (NIST), provides excellent further information about EDA. In Section 1.3.2 of the additional reading, the following 13 questions which can be asked during EDA are listed:

1. *"What is a typical value?*
2. *What is the uncertainty for a typical value?*
3. *What is a good distributional fit for a set of numbers?*
4. *What is a percentile?*
5. *Does an engineering modification have an effect?*
6. *Does a factor have an effect?*

7. *What are the most important factors?*
8. *Are measurements coming from different laboratories equivalent?*
9. *What is the best function for relating a response variable to a set of factor variables?*
10. *What are the best settings for factors?*
11. *Can we separate signal from noise in time-dependent data?*
12. *Can we extract any structure from multivariate data?*
13. *Does the data have outliers?"*

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

Compulsory Task 1

Create a jupyter notebook called **titanic.ipynb**. Use this Notebook to create an in-depth EDA on the Titanic dataset provided in this task. Your EDA should contain descriptions of your EDA and appropriate visualisations.

Use the following guiding questions for your EDA:

- What is the most important factor in determining survival of the Titanic incident?
- 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.

If you are having any difficulties, please feel free to contact our specialist team **on Discord** for support.

Completed the task(s)?

Ask an expert to review your work!

Review work

Rate us

Share your thoughts

HyperionDev strives to provide internationally-excellent course content that helps you achieve your learning outcomes.

Think that the content of this task, or this course as a whole, can be improved? Do you think we've done a good job?

Click here to share your thoughts anonymously.

REFERENCES

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**