SUMMER ANALYTICS

MODULE 03: EXPLORING DATA

Hello everyone!
We hope you are having fun in the course.

This week let's learn how to gather data, visualize it, and make sense of it so that we can run machine learning algorithms on the same.

TOPICS :

1. Introduction to EDA and Data Visualisation
   EDA, or exploratory data analysis, is the first step in your mastery over data.

TASK 01	Introduction to EDA	A short blog post that will introduce you to the topic
TASK 02	Various types of plots	Knowledge of various graphs and where they are used is essential

2. Introduction to Matplotlib
   Matplotlib is the most popular plotting python library and is a prerequisite for learning data science because what good is data if you can't plot it?

TASK 01	Overview of Matplotlib	A short video to give you the basic grounding in code
TASK 02	A detailed look	This playlist will be helpful in case you need a refresher on some aspects of the library
TASK 03	Reference resource (please navigate ahead till end)	Series of articles based on the Data Science Handbook - helpful for brushing up on any concept
TASK 04	Self-check assignment Self check assignment solutions	Do try to solve the questions given here and in case you get stuck, please refer to the solutions.

3. Introduction to Seaborn
   Does matplotlib seem boring and drab? Well, you may want to have a look at Seaborn then. Another plotting library, this is built on top of matplotlib and is more versatile.

TASK 01	Introduction to Seaborn	A great blog for an introduction to the library: feat. Pokemon
TASK 02	Seaborn Tutorial Part 01 Seaborn Tutorial Part 02	Distribution Plots tutorial Categorical Plots tutorial
TASK 03	Reference Resource	Another one from the Handbook, reference article for Seaborn
TASK 04	Self-check assignment Self check assignment solutions	Do try to solve the questions given here and in case you get stuck, please refer to the solutions.

4. Data cleaning - dealing with outliers and missing values
   Let's talk about data cleaning. Data cleaning and preparation takes up around 80% of a data scientist's time - this has been estimated by thorough data analysis conducted by the data scientists in question.
   Nevertheless, this is an extremely important part of the job. Remember, bad data in - bad predictions out.

TASK 01	Basic stats concepts (please navigate ahead till end)	Here is a 5 part article that'll give you a grounding in the basics of statistics.
TASK 02	Finding Outliers	Outliers can mess up predictions big-time. Learn how to find them and deal with them.
TASK 03	Handling Missing Values	Without proper data in place, you can't predict anything anyway. Learn how to handle them here.

5. Exploratory Data Analysis
   Now that we've talked about what EDA is, why it is required, and had a look at the various tools and steps involved, let's get started with the same. These two tutorials will show you how to apply all the steps we talked about.

TASK 01	Tutorial 01	Applying EDA on the famous Titanic Dataset
TASK 02	Tutorial 02	Applying EDA on a real-life example of Sales Dataset

6. Web Scraping
   You might be wondering - we mentioned gathering data but haven't covered that anywhere. Well, here it is, a section on web scraping. Web scraping is the act of collecting data from websites and processing it into a form that can be manipulated.

TASK 01	Building a web scraper	An hour-long video which will introduce you to the BeautifulSoup library
TASK 02	Web scraping tutorial	For those of you who prefer blogs, this concise article gives a step-by-step walkthrough in web scraping
TASK 03	Web scraping using scrapy (Optional)	Scrapy is a more powerful web scraping framework (you'll need to download a code editor for this)

Congratulations, you've made it through another module!
Now that you've worked through all this, you may have another look at the previous module's PyCon video - you may feel more at ease with the content taught there!
Now just one last step is left - assignments.

TASK 01	Self-check assignment (folder with qs, soln & dataset)	This self-check is different from the above two in that it's not based on one single library.
TASK 02	Assignment on EDA	This is a graded component, the quiz for this week will be based on this assignment.
TASK 03	Assignment on web scraping	This is a graded component, the quiz for this week will be based on this assignment.

Click here to attempt the test.