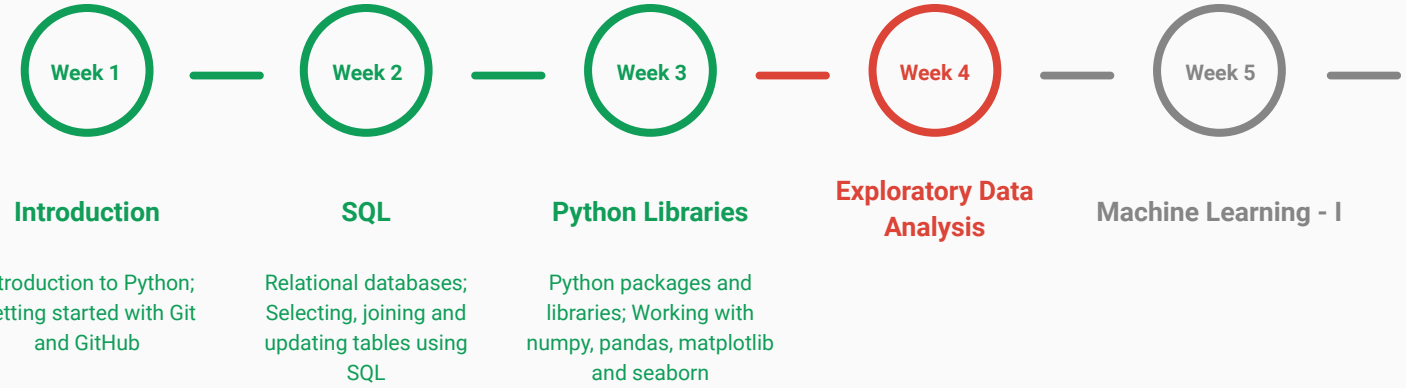
Week 04: Exploratory Data Analysis

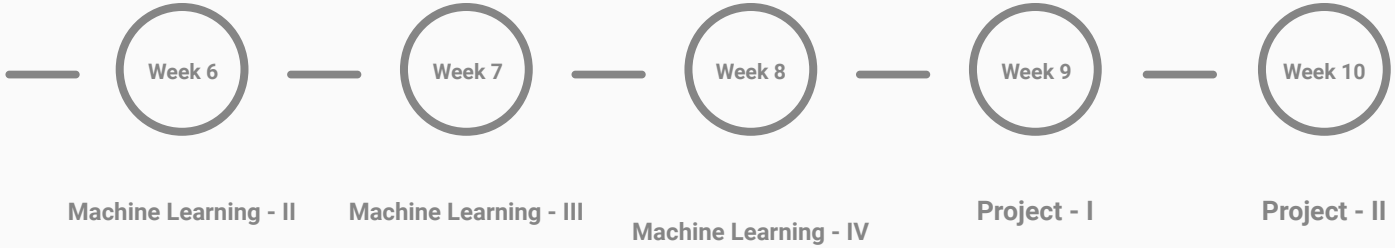
Data Science Bootcamp
Fall, 2021

Instructor: Sagar Patel

Where are we?


Week 1
Week 2
Week 3
Week 4
Week 5
Introduction
Introduction to Python; Getting started with Git and GitHub
SQL
Relational databases; Selecting, joining and updating tables using SQL
Python Libraries
Python packages and libraries; Working with numpy, pandas, matplotlib and seaborn
Exploratory Data Analysis
Machine Learning - I

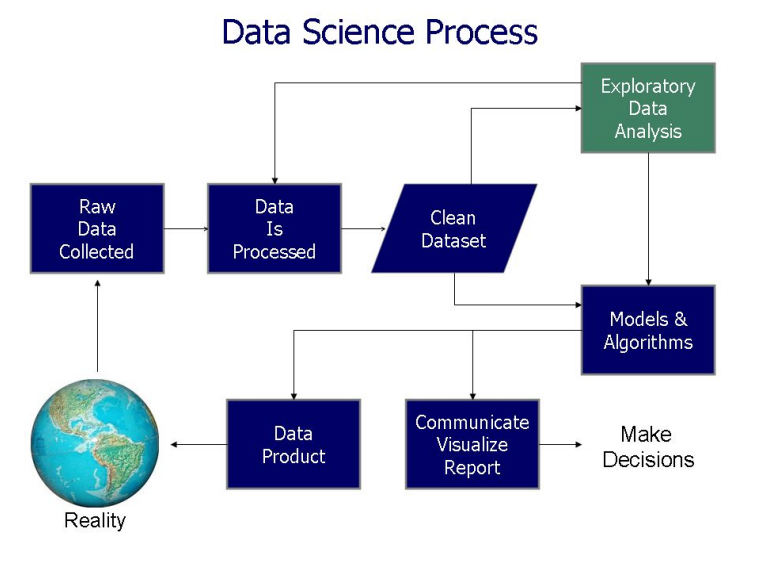
Where are we?


Week 6
Week 7
Week 8
Week 9
Week 10
Machine Learning - II
Machine Learning - III
Machine Learning - IV
Project - I
Project - II

Agenda

- What makes data “good”?
- Purpose of Exploratory Data Analysis
- A short “recipe” for EDA

Data Science Process Flowchart



Getting started with EDA

What makes data “good”?

- **Size**
  - The **more** samples are in the data set, the **more** examples your machine learning model will be able to learn from
  - Often, a simple machine learning model trained on a large data set will **outperform** a “fancy” models on a small data set
- **Quality**
  - Are there **predictive features** in the data?
  - Are no values (or very few values) **missing, noisy, or incorrect?**
  - Is the scenario in which the data collected similar to the scenario in which your model will be used?

Why do we need to perform EDA?

- Once we have identified one or more candidate data sets for a particular problem, we perform some exploratory data analysis. This process helps us
  - **Detect** and possibly correct mistakes in the data
  - Check our **assumptions** about the data
  - Determine **relationships** between features
  - Assess the **direction** and rough size of **relationships** between **features** and the **target variable**

Why do we need to perform EDA?

- Once we have identified one or more candidate data sets for a particular problem, we perform some exploratory data analysis. This process helps us
  - **Detect** and possibly correct mistakes in the data
  - Check our **assumptions** about the data
  - Determine **relationships** between features
  - Assess the **direction** and rough size of **relationships** between **features** and the **target variable**

EDA is important for understanding whether this data set is appropriate for the machine learning task at hand, and if any extra cleaning or processing steps are required before we use the data

The “Recipe”

- There are **no specific rules** as to “how” one performs EDA

- However, there are a few **general** steps which need to be followed in order to make sense of the data
  - **Learn** about your data
  - **Load** data and check that it is loaded correctly
  - Visually **inspect** the data
  - Compute **summary** statistics
  - Explore the data further and look for potential issues

Summary

- EDA is the stepping stone to designing a Machine Learning model
- It is pivotal as performing wrong analysis can lead to an incorrect (business) decision

That’s all Folks!

See you in the next session :)

Give us a feedback: https://bit.ly/3EX8MYh