جامعة الطائف
TAIF UNIVERSITY

PRINCIPLES OF DATA SCIENCE

2021 - 2022

Assoc. Prof. Dr. Salha Alzahrani

Department of Computer Science

جامعة الطائف
TAIF UNIVERSITY

CHAPTER 3: THE FIVE STEPS OF DATA SCIENCE

- OVERVIEW OF THE FIVE STEPS
  - Ask an interesting question
  - Obtain the data
  - Explore the data
  - Model the data
  - Communicate and visualize the results
- EXPLORE THE DATA

OVERVIEW OF THE FIVE STEPS

Overview of the five steps

Data science follows *a structured, step-by-step process that, when followed, preserves the integrity of the results*.

The five essential steps to perform data science are as follows:

1. Asking an interesting question
2. Obtaining the data
3. Exploring the data
4. Modeling the data
5. Communicating and visualizing the results

1. Ask an interesting question

- Treat this step as you would treat a brainstorming session.
- Start writing down questions regardless of whether or not you think the data to answer these questions even exists.
- The reason for this is twofold:
  - Firstly, don't be bias before searching for data.
  - Secondly, obtaining data might involve searching in both public and private locations and, therefore, might not be very straightforward. You might ask a question and immediately tell yourself "Oh, but I bet there's no data out there that can help me," and cross it off your list. Don't do that! Leave it on your list.

2. Obtain the data

- Once you have selected the question you want to focus on, it is time to scour the world for the data that might be able to answer that question.
- As mentioned before, the data can come from a variety of sources; so, this step can be very creative!

3. Explore the data

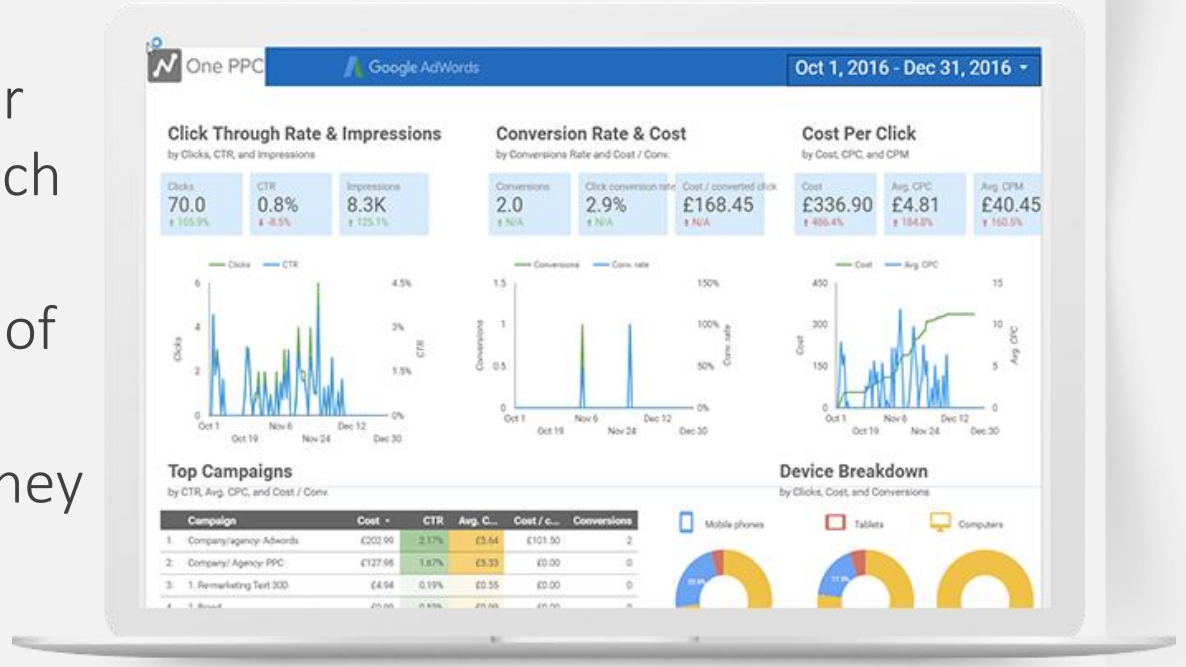
- Once we have the data, we use the lessons learned in Chapter 2, Types of Data, and begin to break down the types of data that we are dealing with.
- This is a pivotal step in the process.
- Once this step is completed, the analyst generally has spent several hours learning about the domain, using code or other tools to manipulate and explore the data, and has a very good sense of what the data might be trying to tell them.

4. Model the data

- This step involves the use of statistical and machine learning models.
- In this step, we are not only fitting and choosing models, but we are also implanting mathematical validation metrics in order to quantify the models and their effectiveness.

5. Communicate and visualize the results

- This is arguably the most important step.
- While it might seem obvious and simple, the ability to conclude your results in a digestible format is much more difficult than it seems.
- We will look at different examples of cases when results were communicated poorly and when they were displayed very well.



Note

In this book, we will focus mainly on steps 3, 4 and 5.

Why are we skipping steps 1 and 2 in this book?

While the first two steps are undoubtedly imperative to the process, they generally precede statistical and programmatic systems. Later in this book, we will touch upon the different ways to obtain data, however, for the purpose of focusing on the more scientific aspects of the process, we will begin with exploration right away.

EXPLORE THE DATA

Explore the data

- The process of exploring data is not defined simply.
- It involves the ability to recognize the different types of data, transform data types, and use code to systemically improve the quality of the entire dataset to prepare it for the modeling stage.
- In order to best represent and teach the art of exploration, I will present several different datasets and use the python package pandas to explore the data. Along the way, we will run into different tips and tricks for how to handle data.

Basic questions for data exploration

Q #1 : Is the data organized or not?

- We are checking for whether or not the data is presented in a row/column structure.
- In this book, over 90% of our examples will begin with organized data. Nevertheless, this is the most basic question that we can answer before diving any deeper into our analysis.
- A general rule of thumb is that if we have unorganized data, we want to transform it into a row/column structure. For example, earlier in Ch2., we looked at ways to transform text into a row/column structure by counting the number of words/phrases.

Basic questions for data exploration

Q #2 : What does each row represent?

- Once we have an answer to how the data is organized and are now looking at a nice row/column based dataset, we should identify what each row actually represents.
- This step is usually very quick, and can help put things in perspective much more quickly.

Basic questions for data exploration

Q #3 : What does each column represent?

- We should identify each column by the level of data and whether or not it is quantitative/qualitative, and so on.
- This categorization might change as our analysis progresses, but it is important to begin this step as early as possible.

Basic questions for data exploration

Q #4 : Are there any missing data points?

- Data isn't perfect.
- Sometimes we might be missing data because of human or mechanical error.
- When this happens, we, as data scientists, must make decisions about how to deal with these discrepancies.

Basic questions for data exploration

Q #5 : Do we need to perform any transformations on the columns?

- Depending on what level/type of data each column is at, we might need to perform certain types of transformations.
- For example, generally speaking, for the sake of statistical modeling and machine learning, we would like each column to be numerical.
- Of course, we will use Python to make any and all transformations.

Pandas

DATASETS

Dataset 1 – Yelp

- The first dataset we will look at is a public dataset made available by the restaurant review site, Yelp.
- All personally identifiable information has been removed.

Dataset 2 – titanic

- The titanic dataset contains a sample of people who were on the Titanic when it
- struck an iceberg in 1912.

s.zahrani@tu.edu.sa