|  |  |
| --- | --- |
| Introduction to Data Science  FCUP  **Practical Class 2**  **Exercises** |  |

1. Assign the following skills to the two roles:
   1. Data Engineer
   2. Machine Learning Scientist

Skills:

* Strong Java Skills(a.)
* Proficient in using Python for prediction and modelling(b.)
* Expert at building and maintaining SQL databases(a.)
* Advanced knowledge of statistics(b.)
* Experience in using Tensorflow and implementing Deep Learning architectures(b.)

1. Assign each task to the correct job title
   1. Data Engineer
   2. Data Analyst
   3. Data Scientist

Skills:

* Run correlation analysis between weather and ice cream sales(c.)
* Train an anomaly detection algorithm(b.)
* Create a dashboard for the marketing team(b.)
* Create a new table in the SQL database(a.)
* Update Excel spreadsheet with new graphics(b.)
* Give a new team member access to database(a)

1. Plan your data science project.

|  |
| --- |
| *Imagine that you work for a bank that feels that it is losing too much money and wants to reduce its losses. You were hired to develop a tool to help loan officers to more accurately detect risky loans.*  *Example adapted from Practical Data Science with R, 2nd Edition Manning* |

Define the following:

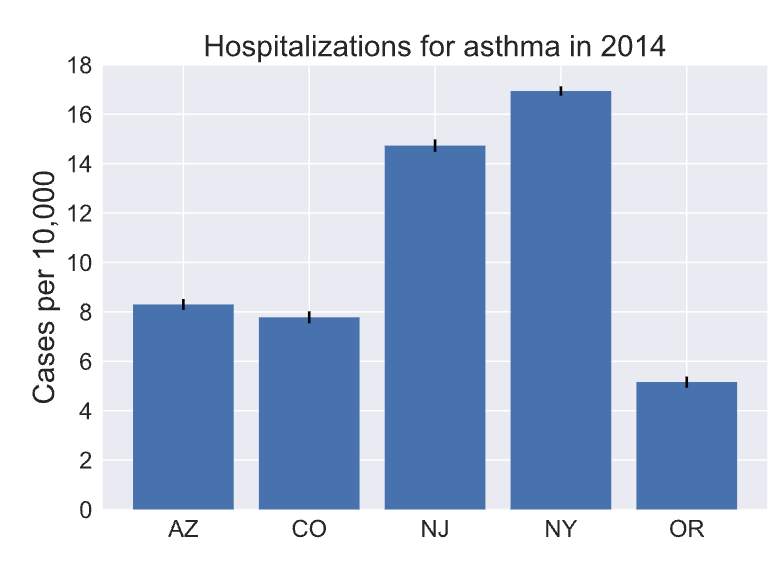
* Who are the players in this project? Which roles do they have?

Project sponsor -> Bank

Client -> loan officers

* What would be a main goal with this project?
* Which kind of data will you need and that could be relevant?
* What would be the main task of your model?
* How could you determine that you are using a satisfactory model?
* How would you present the results of your model?
* How could the model be used by the institution?

1. Consider the following visualization. What source has the data scientist most likely used to collect this data?



Example adapted from DataCamp

* APIs
* Public records
* Web events

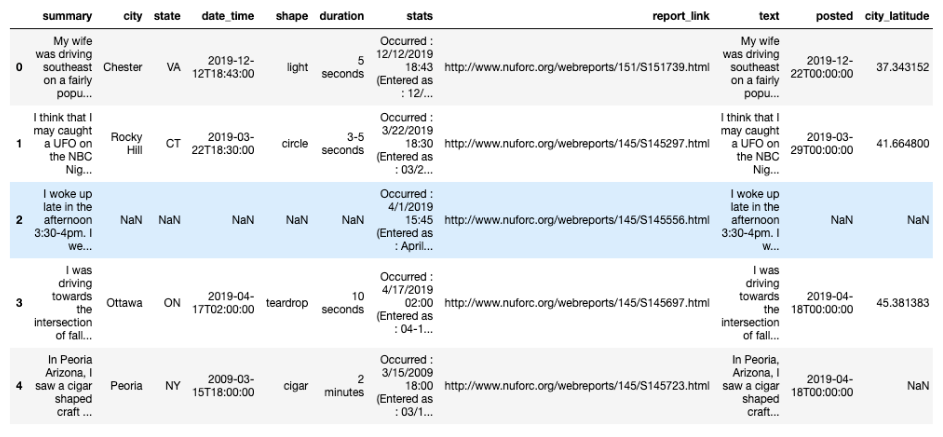
1. Some data is tabular and belongs in a Relational Database. Some is unstructured and belongs in a Document Database. Sort each dataset into the correct type of database.

* Relational Database
* Document Database
* Images of different traffic events, including metadata about image’s content.
* Text from various emails sent and received by you.
* The dates, times, subjects and recipient’s address for all emails sent by you.
* Customer information for all students of a university, such as name, phone number and location.

1. Data pipeline characteristics. Which of the following statements is true?

* A data pipeline is essential for every data science project.
* In the transform phase of ETL, data analysts perform exploratory data analysis.
* Data engineers design and build custom data pipelines for projects.
* Data pipelines do not require automation.

1. Consider a dataset on UFO reporting with the following entries:



Example adapted from DataCamp

* The last row is missing a value for city\_latitude, you should drop it.
* The third row is full of NaN values, which is unusual. You should try to understand why that is the case before taking the decision to drop or keep it.
* The third row is full of NaN values. A UFO sighting event has to have a place, a date, and a description (the shape of the object). This row does not have this information, you should drop it.
* These five observations report seeing aliens, so what you and your friend saw was probably aliens too.