Data Science & Data Engineering
Code.Learn Upskilling Program
4 April 2018

1. **Download** the following datasets: RawConstructionData.csv; Schedule.csv; Building.csv; Floor.csv;Trade.csv
2. Perform the necessary **data sources exploration** activities and **load** the data to **Microsoft SQL Server** by using SSMS capabilities. Create the appropriate **OLTP system** with at least primary and foreign keys. Perform necessary data transformations for the completeness and validity of this stage.
3. **Design** the corresponding **data warehouse**, produce the necessary **schema** and identify the appropriate **fact tables, dimension tables and measures**. The data warehouse must be able to accommodate at least the following company's needs:
    a. Analysis of the company's situation with respect to Scope, BOQ, Activity, Construction Element and Time.
    b. Analysis of the Scope with respect to Project, Level and Trade.
    c. Analysis of the BOQ with respect to its category.
    d. Analysis of the Construction Element with respect to its Part, Type and Family.
    e. Report based on the quantity and cost of the construction progress.
4. **Load** the data to your **Data Warehouse** by performing the necessary **ETL processes by using python.** At least the below mentioned points should be addressed:

    a. Assign appropriate data types to all attributes (e.g. quantities should be appropriate number data types).
    b. Handle appropriately all the missing/unknown/wrong values and state clearly your assumptions.
    c. Transform the dates in Start and End Date columns to the below format and also add the missing 0s from the days/months where necessary. DD-MM-YYYY

   d. From the Month extract the starting quarter for each activity. (Q1,Q2,Q3,Q4)
   e. From the Activity Desc remove the level indication and keep the rest of the Activity description
   f. Set the Thickness, the Length and the Height in two decimals where it is not.
   g. Make the descriptions consistent. For example check the BOQ Description for “Polysterine Insulation”

5. Write the following **SQL statements by using SSMS**:

   a. Find the total cost of each BOQ Category.
   b. Find the average quantity and cost with respect to each construction element type and floor.
   c. Find the total quantity and cost with respect to each construction element family, BOQ Category and Project.

6. Use **NumPy, SciPy and Pandas** python libraries in order to perform the following data analysis tasks:

   a. **Create** a Pandas DataFrame from *schedule* and *RawConstruction data*.
   b. Inspect the DataFrame, describing its **shape** and main **statistical attributes**.
   c. Perform an **EDA** on the two DataFrames:
      - For each categorical column find the number of **unique** values and the number of times each of these values appears.
      - For numeric columns find the **range** of the values.
      - For date columns find the **range** of the dates.
   d. Wherever you deem necessary:
      - Handle the **missing data** in the dataframe.
      - Correct any **spelling** or data entry **inconsistencies**.
   e. From *schedule* figure out what building materials (*ConstructionElementType*) are used for building walls (keyword 'walls' in *BOQ*).
   f. In *RawConstructionData* **split** the variable *Quantity* into two groups: 'low' and 'high', depending on whether or not *Quantity* is higher or lower than the value 10. Store this into a new column named *BinCuantity*.

Note: any assumption taken should be documented (e.g during the process of filling null values, cleaning the data etc).

7. Create necessary and meaningful **data visualizations** that can reply at least to the aforementioned inquiries by using **Matplotlib** and additionally to the following points:

   a. Visualize the distribution of *BOQCategory* in *RawConstructionData* through a **pie chart**.
   b. Visualize the distribution of *Lengths* in *RawConstructionData*, while first having removed all NaN values.

c. Visualize the relationship between *Quantity* and *TotalCost* in *RawConstructionData* (tip: remove outliers*).

d. Visualize the results from subtask 6.e with whichever graph you see fit.

e. Visualize the relationship between *BOQCategory* and *TotalCost* for all costs under 1000.

f. Same as above (7.e.), depending on whether or not we have low/high quantity (variable we created for 6.f.).

*outliers: unreasonably high quantities/costs in this case.

Notes:

- All graphs should be as aesthetically pleasing as possible and should be accompanied with a title, labels for x, y axes, etc.
- You may present any other visualization you feel is meaningful, especially in case it helps you towards better understanding of the data.