Project 3

SGD

Department of Informatics Engineering

Delivery date: see Inforestudante/Submissao de Trabalhos

Objectives

- Learn how to do data analysis.

Final Delivery

- You must submit your project in a zip file using Inforestudante. Do not forget to associate your work colleague during the submission process.
- The submission contents are:
  - All source code.
  - Report with setup, step-by-step how you did it

The REPORT is expected to be complete in the sense that it needs to contain all necessary and sufficient information for the teacher to give the score to the evaluation item. For that you can include descriptions, screenshots, code extracts, whatever is needed for a complete and thorough evaluation. If the report is absent the score is 0, and if it is incomplete the score is significantly affected. This is to make sure you do have a complete report.

Before the main body, the REPORT starts with the complete identification of the students and group, then a table of contents, then the following:

a. Lists of what the group succeeded to do and what is missing
b. self-evaluation of the group (0-100%)
c. List of what each student contributed (no repetitions)
d. self-evaluation of each student in the group (0-100%)
e. hours of effort by each student separately

These items a,b,c are important for the teacher to check whether his evaluation coincides more or less with what the group and student thinks.

Grading

- REPORT (confirmed later by defense)
  - setup, step-by-step how you did it
  - how you did each analysis and charts with results and conclusions
  - Code

Resources

- Jupyter notebook :
- Book: **[0] Mastering Python for Data Science, Packt-Book, Samir Madhavan**
- SSB analysis data in URL to be given:_____

Fundamental

Start early, and you need to discuss your task with the teacher to better define its details. Always keep in touch with the teacher, DO NOT TRY TO DO THE WORK ONLY WHEN THE DEADLINE IS APPROACHING, YOU WILL FAIL AND GET MAD ABOUT IT. ALSO, DO THE WORK IN LAB CLASSES, YOU NEED TO UNDERSTAND THNGS, DO NOT TRY TO DO IT ON YOUR OWN ONLY AND LATE.

Project Description

Files with new SSB data will have to be uploaded to the database. The aim will be to use techniques given in class to analyze new sales data and to compare with the last year already existing in SSB, all using python. You should be able to discover the relevant patterns of that data and the variations between the two years, showing the results and demonstrating how you got them (the steps). In addition to discovering patterns, such as variations in sales (brands, seasons, months) when comparing the two years given, you should describe how you did it. You must include at least the following EDA components in your analysis: histograms, statistics, charts, hypothesis testing, degree of significance (p-values) and confidence intervals.

You should do an investigative work to find data patterns. You should discover at least relative to the following details:

Some of the obvious questions:
How do sales compare across countries and regions?
How do sales vary between the two years, considering different categories of attributes of products, time, supplier and customer?
What are the maximum, mean, minimum sales considering categories of attributes of products, time, supplier and customers
Histograms considering different categories of attributes of products, time, supplier and customers
Statistics considering different categories of attributes of products, time, supplier and customers
Charts describing the performance of sales taking into consideration categories of attributes of products, time, supplier and customer.
Hypothesis testing regarding sales between the two years over categories of attributes
Regression analysis, with degree of significance (p-values)
Confidence intervals for average sales considering different categories of attributes of products, time, supplier and customers
Clustering into groups.

For this, of course you should know what categories of attributes are: just look at the columns. Things such as sellingseason, p_brand and a lot others are of course such categories of attributes.