

Scientific Programming with Python: Assignment 3

Problem Description

Global grocery stores play a vital role in connecting communities and fostering cultural exchange through food. By sourcing products from around the world, these stores provide access to diverse ingredients and flavors that enrich diets and culinary experiences. They bridge geographical boundaries, supporting local farmers and producers while meeting the demands of international consumers. Beyond convenience, global grocery stores contribute to food security by offering a reliable supply of essential goods and promoting sustainable practices in sourcing and distribution. Their importance lies not just in feeding people but also in nurturing global connections and understanding through the universal language of food.

For this assignment, we will explore the global activity of a fictional Global Superstore, the MTU groces LTDA. The dataset we will use is comprised of 51289 rows and the following 27 columns:

- **category:** The category of products sold in the superstore.
- **city:** The city where the order was placed.
- **country:** The country in which the superstore is located.
- **customer_id:** A unique identifier for each customer.
- **customer_name:** The name of the customer who placed the order.
- **discount:** The discount applied to the order.
- **market:** The market or region where the superstore operates.
- **ji_lu_shu:** An unknown or unspecified column.
- **order_date:** The date when the order was placed.
- **order_id:** A unique identifier for each order.
- **order_priority:** The priority level of the order.
- **product_id:** A unique identifier for each product.
- **product_name:** The name of the product.
- **profit:** The profit generated from the order.
- **quantity:** The quantity of products ordered.
- **region:** The region where the order was placed.
- **row_id:** A unique identifier for each row in the dataset.
- **sales:** The total sales amount for the order.
- **segment:** The customer segment (e.g., consumer, corporate, or home office).
- **ship_date:** The date when the order was shipped.
- **ship_mode:** The shipping mode used for the order.
- **shipping_cost:** The cost of shipping for the order.
- **state:** The state or region within the country.
- **sub_category:** The sub-category of products within the main category.
- **year:** The year in which the order was placed.
- **market2:** Another column related to market information.
- **weeknum:** The week number when the order was placed.

Based on these datasets, you will be required to explore the dataset using the methodologies we've learned throughout the semester, meaning that the group is free to explore the dataset the way you see fit, however you will need to meet the following requirements:

1. Load the dataset using Pandas methods.
2. Identify and deal with NaNs. You are required to justify your reasoning.
3. Explore the distribution of data and relationship between different columns.
4. Group the data into multi-dimensional tables and generate plots out of it
5. Formulate a Hypothesis and test it by using data exploration.
- 1- Build a Power point presentation to present your project at the last lab of the term.

The methods used to explore, process, and generate visualisations must be those introduced within this module. The answers to this project must be provided in a well organised jupyter notebook, containing the rationale, visualisations and a small discussion of your findings (Use the markdown for cells containing text).

In other words, for every notebook cell containing code, such as functions, data being processed and analysis being performed, you need to add a subsequent markdown cell containing the explanation and discussion for that particular piece of your work. If you don't know how to use markdown, please refer to this [link](#).

The jupyter notebook will stand for **50%** of this assessment marks, so document everything, don't be shy about it. The presentation will stand for **25%** of your marks and the Q&A will stand for another **25%** of your marks.

Submission

You are required to submit:

- A Jupyter notebook (.tar.gz or .zip archive).
- Please put the student names and numbers at the first cell of your notebook.
- It is your responsibility to make sure you upload the correct file.

All files can be submitted with Canvas before December 9th 2023 (23:59). Please ensure to include your group name on the Jupyter notebook document.

Code

- All code should be completed using Python as the programming language.

Late submissions

- If you don't get the assignments done to your satisfaction and don't meet the minimum requirements by the deadline, you have the option (as with any assignment at MTU) of submitting up to 1 week late for a penalty of 10%.

- This penalty is subtractive. Work that would have earned 55% if on time would get 45% (not 49.5%) if late.
- The penalty is applied weekly. So, one day late costs the same as 6.
- If you have a specific reason for submitting a late assignment (sickness, etc.), please submit directly a medical certificate to the department secretary.

Plagiarism

Please read and strictly adhere to the MTU Honesty, Plagiarism and Infringements Policy Related to Examinations and Assessments. Note that reports are checked against each other and external web sources for plagiarism. Any suspected plagiarism will be treated seriously and may result in penalties and a zero grade.

Grading

The assignment is worth 40% of the overall mark for the module. Marks will be awarded based on the quality of the code and the results. In particular, I will be checking to see if you are handling data correctly, carrying out exploratory analysis to gain insights, correctly performing visualisations, and critically documenting everything in a clear and concise way. The submitted code will also be checked to ensure that the work is your own.

Presentation

The presentation will add an individual layer to your grade. I will evaluate your presentation skills as well as your understanding during the Q&A session. The grade, ranging from 1 to 5 (2.5 for the presentation and 2.5 for the Q&A), so, if your group was able to secure a project grade of 50 points but you got a Presentation score of 2, you will be entitled to 70 points. Make no mistake, this is a team effort, and this is a way to ensure that you will all be important players for the team goal