**THIS IS A DATA PROCESS DIARY ABOUT THE DATA I COLLECT FOR LAST MONTH ON SPENDINGS**

First of all, to be able to create my project I needed to be sure if the data I collected is proper for the case projects. So, currently I am in a stage where I am trying to find a better way to show data in excel. And I want to say that I am using Chat-GPT to go deeper in understanding my own data and understanding in how to make it better.

**CLEANING THE DATA AND MAKING IT BETTER**

I have currently these categories: Income, Food&Drinks, Entertainment, Groceries, Education, Housing, Utilities, Transportation, Personal Care, Healthcare, Miscellaneous, Debt Payments. The first problem about this is there are some categories that are unnecessary, or two categories can be combined in one. For example, "Utilities" and "Housing" might be combined into a broader category like "Living Expenses".

Some of the categories can be renamed/removed. For example, “Debt Payments” is not used properly in the data, because I had no debts to pay. That’s why in the data its used for loaning. So, it can be renamed for understandable category. And “Healthcare” category can be removed because it never been used in the data.

Other problem I have with the data is I used description column as the name for each product/spending. But some of the descriptions miss typed or wrongly typed. Which I mean for example I have the data with description “Hot-dog”. The next entry which is indeed a “Hot-dog” is named “Hot-dog Hary”. This caused by the misunderstanding that I had where I wanted to specify the place where I bought the “Hot-dog”. This can be fixed by adding a new column with the place where the spending bought for each entry. But the problems with this is every entry needs to be entered and renamed again. And redoing all of these can create false entries because I need to enter the place for each entry, and which means I need to remember the place where it’s bought for each entry. Probably I’m going to add column “Place” for each entry. But before that I need to be sure if I really need the buy place for each entry, or how it can really help to understand the reason why I collected all of these data in the first place.

**THE GOAL OF THE DATASET**

The goal of the dataset is really important topic. Because without any purpose the data became nothing more than texts and numbers.

---\*(this place’s purpose is to be able to understand the goal of the data) \*---

One of the goals was to determine where my money goes. In the first semester I spent much more money than I thought. But I was not sure if my money spent necessarily or unnecessarily. So, then I created some categories that I thought it would fit into my future entries and create an excel data table.

1 – How did I spend my money?

2 – Which products do I spend my money on?

3 – Usually where/which places do I spend my money?

4 – Getting some insights about my life? (Do I eat healthy, usually where do I spent my day etc.)

---\*(this place’s purpose is to be able to understand the goal of the data) \*---

**In that manner,** first I need to fix my category set and then check and fix the data accordingly.

-spendings\_final is the final version of the dataset

**UNDERSTANDING OF DATA’S PO**TENTIAL

Categories are clear, data is there and clear. First thing is which headers I have. Date, Place, Description, Category, Amount. First thing first we can group the data by its categories.

This is a pause for given instructions.

1. Get the data and briefly explain it.
2. Provide answers to questions like:
3. Where did you find the dataset, and what format is it in?
4. How many observations and variables are there in the dataset?
5. **Explore the data with these steps:**
6. check the data type for each variable contained in the data set. X
7. check if the null values exist in the data set. X
8. check the number of unique values for each column in the data set, if they exist X
9. explore whether the data set contains outliers for each variable. X
10. if outliers are detected in the previous step, remove them from the data set for the purpose of making further analysis. (cannot proceed, data is not proper for removing data.)
11. **Visualize at least two variables.**
12. **Calculate basic descriptive statistics measures of central tendency and dispersion. Explain the results.**
13. **Develop linear regression model by with one independent and one dependent data.**
14. **Explain the slope and intercept from regression model.**
15. **Predict the future value of used dependent variable Y by using some value of independent variable X not contained in the data set.**
16. **For archive purposes: Use the word document to paste the entire Python code. Save the code in notebook.**
17. **Prepare a presentation with max 10 slides.**
18. **Due date is 27th May 2024 at 23:59 PM.**
19. **Sent all of these 3 files to the mail:** [fkanlic@ius.edu.ba](mailto:fkanlic@ius.edu.ba)
20. **Presentations will be on Thursday 30th May.** The duration is max 10 min.

For the 3rd instructions “d-“ section. Finding if there is any outlier is not proper for this data set. The reason is in these data set the amount can be everything whether its categories same or not. It could only indicate if there is any obvious miss inputs such as hamburger being 105 km instead of 10,5 km. But in this data set there is not such as miss inputs. For solution, we can change the meaning of the word “outlier” to “too expensive” or “unnecessary high expense”. With this solution we can make outliers meaningful.

I designed function to find possible outliers or in the other words high expenses compared to the other data. Func name is find\_outliers()

I don’t know when I started to write this but today’s date is 16.05.2024.

Let’s go back what were I doing.

I was trying to create a regression line. But the problem was I have only one data. I mean one numerical data. Then, I asked chatgpt if there is any other way that we can create a regression line from one numerical and nominal. Then I got it I could use date data with amount and might create a regression line. Lastly I was trying that but it Is hard to find a data that is correlated. And after that week 13th lecture, professor mentioned his data project on regression line and how he did create a regression line with multiple independent variables. And of course in that lecture I learnt that there are a lot of different regression lines for example, regression line with multiple independent variable. 2nd thing is logical regression line. And I don’t know that yet. But basically logical regression line is the correlation between the numerical and categorical or nominal data. Which is what exactly I am looking for.

// date 22.05.2024

Today I noticed that “Income” category useless for my project. So I removed and made all expenses positive again.