Final Assignment Introduction to Modelling semester 2 2023-2024

As a final assignment, you will have to write a report in which you show that you are able to analyze data, present the results, and critically evaluate existing quantitative research. The assessment consists of a few questions that you have to answer based on data that is given to you.

The deadline for the report is Friday 5 April at 23:59. The report must be uploaded to hand-in as a word or pdf file. The python scripts should also be uploaded. Upload them as separate files and not as a zip file.

The data

The file pet_shop.csv contains data on the sales results of different items from a company that manages pet shops. The company has both a web shop and brick-and-mortar shops. The data also contains various characteristics of the products the company sells. The managers have hired you as an outside data analyst to make sense of this data. You will have to use your modelling skills to show what the patterns in the data are. Below you can find a list of what each variable means.

1. **Products_sold:** The total number of items sold for this specific product.
2. **Product_category:** the type of product, split up into four categories: food, health, toys, and other.
3. **Quality:** whether the product is of premium quality or off-brand.
4. **Satisfaction:** average rating given to this product on the web shop
5. **Discount:** The numbers of weeks this item has been on discount in 2023
6. **Retail_price:** the retail price of this product (rounded to a whole number).
7. **Perc_physical**: the percentage of this item that was sold in physical stores. For example, a score of 40 means that 40% of this product was sold in physical stored and 60% was sold in the webshop.
8. **Market_size**: an estimation of what the market size of this product is (in thousands). The estimation is done by an algorithm that the company bought.

Based on this data, carry out the following assignments:

Part 1: Data preparation

Assignment 1: (15%)
First, turn the categorical variables into dummy variables and explain which category you chose as the reference category. Second, run a regression model where all independent variables are included in a single model. Use Cook's D to find out if there are any outliers. Note: you will first have to remove missing values first in order to get Cooks D to work.

After you identified the relevant outliers, go back to the original data and turn these outliers into missing values.

Assignment 2: (15%)
The original data contained missing values, and if you did assignment 1 correctly some more should be added. Use the correct imputation techniques for dealing with both the categorical and continuous missing values. Explain what you did. After this, check if there are potential issues with multicollinearity, and if there are, explain how you dealt with it.

Assignment 3: (20%)
There might non-linear relationships in the data. Investigate if this is the case and if you find any show it with a scatterplot and a lowess-curve (remember: the dependent variable should be on the y-axis). If you found any, make the correct transformation and test whether this improved the model fit.

Part 2 starts on the next page.

Part 2: Data modelling
After you have completed assignment 1, 2, and 3, you should now have a final dataset fit for analysis, and it is now time to actually make the final model. If you were not able to do one of the steps in the previous assignment, just work with the data that you have. You can still get some points for the model even if the data is not entirely correct.

Assignment 4: (30%)
First, create a model where all independent variables are included and clearly explain what the outcome of each variable in the model means for how many products are sold.

Second, the management wants you to settle a debate that is going on among the staff. Some people say that the price matters the most for how much a product sells. After all, products that are cheaper will sell more. A second group claims that the market size matters the most. After all, the more potential buyers there are, the more products you can sell. Use the correct regression techniques to figure out who is correct. Clearly explain how you got to your conclusion.

Assignment 5: (20%)
Finally, the management is interested in stocking a new product and wants to know if you can use your regression model to predict how many items it would sell. Make your prediction using your regression model, keeping in mind the principle of parsimony, and report how accurate you think this prediction is.

In the table below you can find the characteristics that the product has (or at least that it will likely have based on what they plan for the product and independent research they did):

Variable	Value
Product Category	Toy
Quality	Premium
Satisfaction	4.6 stars
Discount	20 weeks
Retail price	10 euros
Percent physical	55%
Market size	1000

THE END