A logo for college computing

Description automatically generated

**Assessment Cover Page**

|  |  |
| --- | --- |
| *Student Full Name* | Letícia Kohler Barreto |
| *Student Number* | 2024118 |
| *Module Title* | Statistical Techniques for Data Analysis |
|  |  |
| *Assessment Due Date* | 21/03/2024 |
| *Date of Submission* | 21/04/2024 |

**Declaration**

By submitting this assessment, I confirm that I have read the CCT policy on academic misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source.

I declare it to be my own work and that all material from third parties has been appropriately referenced.

I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

Contents

[**Introduction** 1](#_Toc164633234)

[**1. Data Analysis** 1](#_Toc164633235)

[**Importing libraries and loading the dataset** 1](#_Toc164633236)

[**Verifying the first 5 dataset rows** 2](#_Toc164633237)

[**Calling a descriptive analysis of our dataset to verify some main features of the numerical variables** 3](#_Toc164633238)

[**Checking duplicates and missing values** 3](#_Toc164633239)

[**Getting the unique values present in the categorical columns.** 4](#_Toc164633240)

[**Getting insights into the timeframe of the dataset** 4](#_Toc164633241)

[**Analysing Product Category x Sales in 2023** 5](#_Toc164633242)

[**Creating a heatmap to observe the Product Category Sales over the months** 5](#_Toc164633243)

[**Checking the transactions frequency** 6](#_Toc164633244)

[**2.** **Probability - Discrete** 7](#_Toc164633245)

[**3.** **Probability - Continuous** 8](#_Toc164633246)

[**References** 11](#_Toc164633247)

# **Introduction**

The Student Performance dataset that will be explored in this project refers to data collected from students, capturing essential attributes that drive how the student’s performance is affected by other variables. It includes details such as Gender, Ethnicity, Parental level of education, Lunch, Test preparation course and the score achieved in three subjects.

Data preparation, data cleaning and data visualization tools will be applied in order to verify the data types and values, giving the opportunity to understand about the dimension of the dataset, handle missing values and inconsistencies, discover patterns, identify relationships between variables, show what to expect from the data.

# **1. Data Analysis**

## **Importing libraries and loading the dataset**

We import all the libraries that are important for the analysis and the dataset that will be explored.

A screenshot of a computer

Description automatically generated

## **Verifying the first 5 dataset rows**

This syntax is recommended to explore the data through observing the first few records of the dataset that will allows to verify the data types and values, showing what to expect from the data. A Data dictionary is also very useful to give a context about the data we are dealing with.

A screenshot of a computer

Description automatically generated

**Verifying the dataset shape and the type of data that this dataset contains**

This syntax helps in understanding the dataset's dimensions by resuming the number of rows and columns and checking the dataset's size. The Retail Sales dataset has 1000 rows and 8 features, segregated into 5 categorical variables and 3 numerical variables. Initially, no variable with null values can be identified.

A screenshot of a computer

Description automatically generated

## **Calling a descriptive analysis of our dataset to verify some main features of the numerical variables**

The syntax bellow is used to describe statistical insights about the dataset.

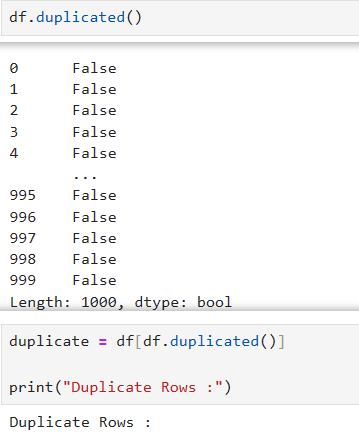
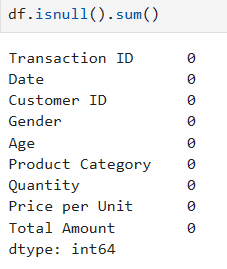
It can be observed that the Student Performance dataset has 1.000 values in all the numerical columns, meaning that there is no missing value, but it will be confirmed on the next steps.

A screenshot of a graph

Description automatically generated

## **Checking duplicates and missing values**

Data cleaning techniques can change depending on the problem context, so it is important understand the problem question to apply the correct techniques and remove, replace or correct the data helping the future model to understand better the data as well as increase the data accuracy. It can be observed the Retail Sales dataset does not have duplicates and missing values.



## **Getting the unique values present in the categorical columns.**

This syntax is also a tool to visualize what type of data the feature contains. It was used in order to understand what are the Product Categories, it can be observed that is divided into 3 categories: Beauty, Clothing and Electronics.

A screenshot of a computer code

Description automatically generated

## **Getting insights into the timeframe of the dataset**

The Date column was transformed into two new columns: Year and Month to help the understanding of which period of time the dataset contains.

A screenshot of a computer

Description automatically generated

The Year and Month columns unique values shows that the data relates to the years 2023 and 2024 which enables the analysis to explore the seasonality of sales.

A screenshot of a computer code

Description automatically generated

The transactions completed in 2024 represent only 0.2% of all transactions so they will be removed, and the dataset will be analysed monthly considering the year 2023.

A screenshot of a cell phone

Description automatically generated

## **Analysing Product Category x Sales in 2023**

A pie chart is utilized to obtain a basic visualization of the total of sales for each product category. It can be observed that in 2023, the sales amounts for each category are similar.

A pie chart with text on it

Description automatically generated

## **Creating a heatmap to observe the Product Category Sales over the months**

A heatmap is a graphical representation of data where values are depicted by color that helps to determine the dependent variables or the connection between the characteristics. Referring to this project it was used in order to demonstrate the sales behave over the months. The dark squares represent the lowest sales and the light ones represent the highest sales, so it can be observed that

Electronics category has the highest sales amount of the year that represent approximately 43% of sales in May. A similar result can be observed in December, but they differ from the other months of the category, one of the reasons could be specific marketing campaigns or because of the Christmas seasonality.

A chart of sales

Description automatically generated

## **Checking the transactions frequency**

It can be observed that the highest transaction frequency occurred in May, probably driven by the high number of sales. and the lowest transaction frequency occurred in September.

A graph of a line

Description automatically generated

# **2.** **Probability - Discrete**

**- What is the probability of rolling exactly two 6s in five rolls of a fair die?**

The Binomial formula is recommended to calculate the probability of an event occurring in a specific number of times in a series of independent trials, which means that one attempt has no influence on the outcome of the other and each attempt has only one possible outcome: success or failure.

The binomial function can be found on *spicy.stats* library to calculate the probability needed. First the parameters are defined and the syntax *binom.pmf (k=,n=,p=)* is applied in order to get the probability:

A screen shot of a computer

Description automatically generated

* **The number of industrial injuries on average per working week in a factory is 0.75. Assuming that the distribution of injuries follows a Poisson distribution, find the probability that in a particular week there will be no more than two accidents.**

The Poisson formula is recommended to calculate the probability of how many times an event occurring in an specific period of time or space, knowing the average.

The poisson function can be found on *spicy. Stats* library. First the parameters are defined and the syntax *poisson.pmf (k=, mu=mu)* is applied in order to get the probability of 0, 1 and 2 accidents and the results are summed to obtain the total probability:

A screenshot of a computer

Description automatically generated

# **3.** **Probability - Continuous**

* **The time a person spends at Dublin Zoo is Normally distributed with a mean of 90 minutes and a standard deviation of 10 minutes. Using this distribution, answer the following:**
* **If a visitor is selected at random, find the probability that they will spend at most 85 minutes visiting the zoo?**

The *norm.cdf ()* function is used to calculate the cumulative distribution function (CDF) of a normal distribution with a given mean and standard deviation. It returns the probability that a random variable following a normal distribution is less than or equal to a given value (x).

First the parameters *mean*, *standard deviation* and *x* are defined.

The standard score represented by *z-score* is calculated. This is a statistical measure that shows a data point's distance from the dataset mean in standard deviations. It is used to comprehend and interpret a specific data point's relative location within a distribution.

The syntax *norm.cdf()* is applied to calculate the probability:

A screenshot of a computer

Description automatically generated

* **If a visitor is selected at random, find the probability that they will spend at least 100 minutes visiting the zoo:**

Following the principles of the syntax above, the function *norm.sf* is used to calculate the probability of a random variable following a normal distribution is bigger than a given value (x):

A screen shot of a computer

Description automatically generated

* **Given that you know that a particular visitor has spent longer than average visiting the Zoo, what is the probability that they have spent more than 100 minutes there?**

Following the principles of the previous question, the same syntax is applied considering the additional condition that the visitor has spent longer than the average. The standard score and the probability are calculated for both parameters: *x* and *x\_condition*. Therefore, the formula *conditional\_prob = prob\_x / prob\_x\_condition* gives the conditional probability that a visitor spends more than 100 minutes at the zoo, given that they spent more than 90 minutesA screenshot of a computer code

Description automatically generated

# **References**

‌

www.kaggle.com. (n.d.). Retail Sales Dataset. [online] Available at: http://kaggle.com/datasets/mohammadtalib786/retail-sales-dataset [Accessed 9 Apr. 2024].

GeeksforGeeks. (2018). ML | Overview of Data Cleaning. [online] Available at: <https://www.geeksforgeeks.org/data-cleansing-introduction/?ref=lbp>.

GeeksforGeeks. (2020). Seaborn Heatmap - A comprehensive guide. [online] Available at: https://www.geeksforgeeks.org/seaborn-heatmap-a-comprehensive-guide/?ref=header\_search [Accessed 10 Apr. 2024].

www.inf.ufsc.br. (n.d.). Distribuição Normal (Gaussiana). [online] Available at: https://www.inf.ufsc.br/~andre.zibetti/probabilidade/normal.html.

docs.scipy.org. (n.d.). scipy.stats.binom — SciPy v1.10.1 Manual. [online] Available at: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.binom.html.

‌

‌

‌

‌

‌