MAKERERE UNIVERSITY

COLLEGE OF COMPUTING AND INFORMATION SCIENCES.

SCHOOL OF COMPUTING AND INFORMATICS TECHNOLOGY.

DEPARTMENT OF NETWORKS.

BACHELOR OF SCIENCE IN SOFTWARE ENGINEERING.

COURSE UNIT: **BSSE YEAR 2 RECESS TERM.**

GROUP N:

NAME.	REGISTRATION NUMBER.	STUDENT NUMBER.
Mukoza Duncan Mwesigwa	17/U/6502/PS	217001851
Makwasi Chrispus Arnold	17/U/5910/PS	217011259
Ssenono Fransis Xavier	17/U/10247/PS	217005992
Lyazi Marvin	17/U/5850/EVE	217011257

Sales Analysis and Prediction System Design Specification Document.

Summary.

This document contains a detailed description of the components of the sales analysis and prediction data pipeline. In the description, each component of the pipeline is described by explaining what is means, what it involves, how it is done and why it is done. It also contains a diagram to explain the data pipeline.

Below is a description of each of the key components of the data pipeline.

1. Data Loading.

This step involves picking the data from the csv file in which it is stored and loading it into the system for analysis to be carried out on it. This is done to load the data into memory so that it can be cleaned and then analyzed by the system.

Libraries to be used.

There is one python library that are used in this process and these are:

- Pandas.

The process of loading the data from the csv file into the system is done by the use of the function, **read_csv("filename.csv")**, which is found in the **pandas library**. This method will load the data from the csv file in which it is stored into a data frame in the system.

2. Data Cleaning.

This is the component in the data pipeline in which the data that has been loaded into the pipeline is prepared, corrected and put into the right format that can be analyzed by the use of the relevant libraries, methods and algorithms so as to get the right results from the analysis.

This is done because, initially, the data that has been loaded into the pipeline contains a number of errors, missing values and some of the values are in formats and datatypes in which the relevant libraries cannot handle analysis on them. Therefore, these errors and irregularities need to be handled so as to prepare the data for analysis which will yield the correct results.

Methods to be used.

There are three python libraries that are going to be used in this component in the pipeline. These are;

- Pandas.
- Numpy.
- Mode.

Handling Missing values.

This is the section of the data pipeline in which the columns containing missing values will be identified and all the records missing values will be filled up with the appropriate values.

In this component, the columns in the loaded data that are containing missing values will be identified using the function, **sum(DataFrame[‘Column_Name’].isnull())**. This function will return the sum of all the missing values in the specified column in the data frame.

Then all the columns that will have been identified to have more than **zero** rows with missing values will have the data types of their expected values determined so as to determine the right method to be used to fill in their missing values.

In case a row in a numerical column is missing a value for a given product, the average of the values in that column for that specific product will be determined using the method, **DataFrame.pivot_table(values=' Column_name_for_column_with_missing_values ', index='Column_name_for_Unique_Identifier')**. This method will return all the average values of each of the products in the specified column which is containing missing values.

In case a row in a non-numerical column is missing a value for a given product, the mode of the values in that column for that specific product will be determined using the method, **DataFrame.pivot_table(values='Column_name_for_column_with_missing_values', columns='Name_of_column_which_is_the_determinant',aggfunc=(lambda x:mode(x).mode[0]))**. This method will return the modal value of the values of the determinant column in terms of the column which is containing missing values.

Handling misspelled values.

This involves identifying columns which contain values which have been input in different ways (for example, containing more than one value that represent the same thing). In these columns, all the different values that are meaning the same thing will be identified and then replaced by one value which will refer to that specific meaning.

Label Encoding.

Label Encoding refers to converting the labels into numeric form so as to convert it into the machine-readable form. Machine learning algorithms can then decide in a better way on how those labels must be operated.

3. Data Visualization.

This involves the types of graphs that are to be used during the visualization of the data to enable the analyst to identify the different patterns and make the relevant conclusions.

Libraries to be used.

The libraries that are to be used in this part of the pipeline include;

- Matplotlib.
- Pandas.

The different graphs that are to be used to visualize the loaded data during the data analysis process are explained below.

- Histogram.

The histogram is a graphical representation of the distribution of a given set of data.

The histogram will be used to identify the distribution of some of the properties of the products that are being sold in the different store outlets. This will be done to identify the categories under which the different properties fall as far as affecting the volumes of sales of the different products fall.

The histogram will be plotted using the function, **hist()**.

- Line graph.

The line graph is a graphical representation of the loaded data which will show the rate of change and relationship between two properties of a given product.

The line graph will be used to determine how some of the properties change in relation to each other.

The line graph will be plotted using the function, **plot()**.

- Bar graph.

The bar chart is a graphical representation of data which can be used to determine the relationship between two categories of data.

The bar chart will be used to determine the relationship between some of the properties of the products being sold.

This chart will be plotted using the function, **bar()**.

4. Apply Algorithms and Modeling Techniques.

In this section, a set of models will be applied so as to identify the relationships between some of the properties of a product so as to make predictions of the sales of the products.

Libraries to be used.

The libraries that are to be used are listed below:

- Pandas.
- Matplotlib.

- Seaborn.

<u>Models.</u>

These are the algorithms that are to be used so as to analyze the data and come up will a set of conclusions and predictions.

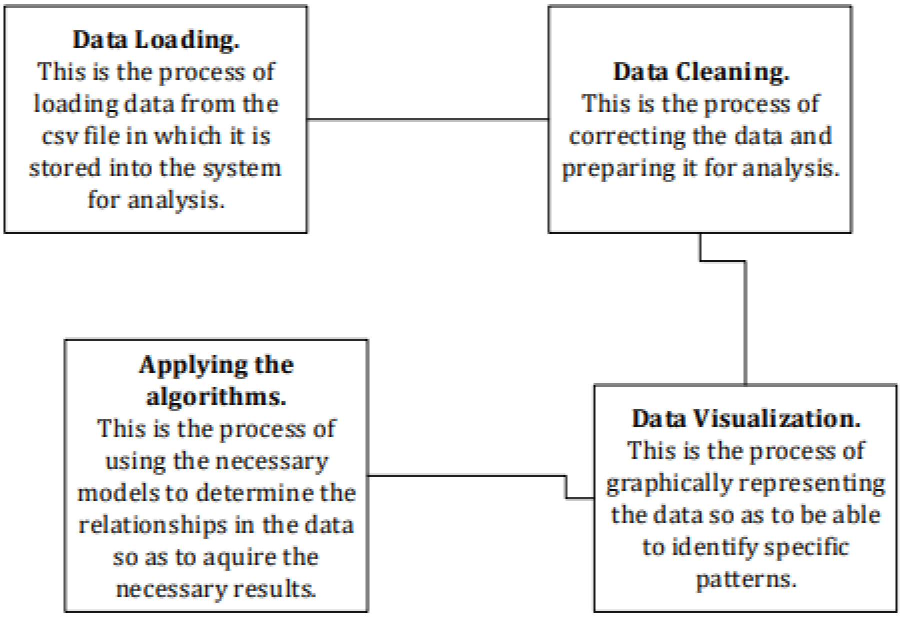
- **<u>Linear Regression.</u>**

  The linear regression model will be used so as to determine the way the different properties affect the volume of sales of a given product and to predict the sales volumes of the product.

- **<u>Correlation.</u>**

  The correlation analysis will be done so as to determine the relationship between some of the product properties.

Diagram of the Data Pipeline.



Data Loading. This is the process of loading data from the csv file in which it is stored into the system for analysis.

Data Cleaning. This is the process of correcting the data and preparing it for analysis.

Data Visualization. This is the process of graphically representing the data so as to be able to identify specific patterns.

Applying the algorithms. This is the process of using the necessary models to determine the relationships in the data so as to aquire the necessary results.