**Assignment # 1**

I take the customer data dataset to apply all the operations.

**Question 1: Take your respective dataset and Do the following in a report form.**

1. **Data Exploration: explore your data with some python instruction that we saw in the class.**

**Answer:**

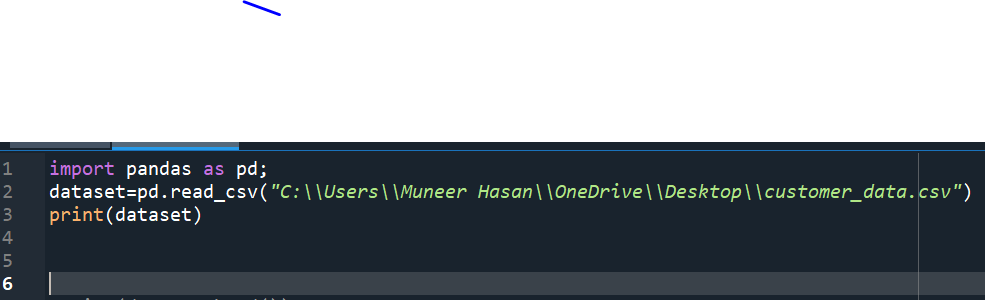
**Data Exploration:**

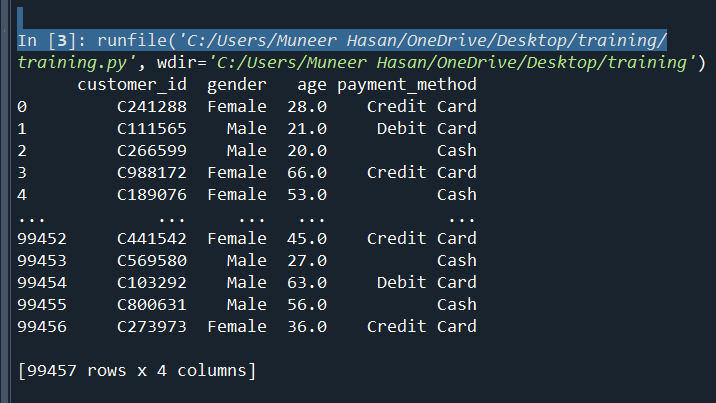
Data exploration is a critical step in the data analysis process that involves examining and understanding your dataset. It helps you gain insights, identify patterns, and make informed decisions about how to preprocess, model, or analyze your data. Here are some common techniques and methods for data exploration.

**Loading Data:**

Read your data into a suitable data structure, like a Pandas Data Frame in Python.

I First import pandas library in my program then give the path where my dataset files lies.





**Viewing Data:**

Use methods like head() and tail() to view the first and last rows of the dataset.

A close up of a sign

Description automatically generated

A screen shot of a computer

Description automatically generated

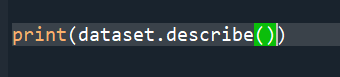


A screen shot of a computer screen

Description automatically generated

**Data Summary:**

Check basic statistics and summary information of the data.



A screenshot of a computer

Description automatically generated

A white text on a black background

Description automatically generated

A screenshot of a computer

Description automatically generated

1. **Describe the dataset: What does it contain? What are the variables? What are the data types?**

**Answer 2:**

**Dataset Description:**

A screenshot of a computer

Description automatically generated

**Name Of dataset:**

The dataset name is customer\_data which is provided by our teacher.

**Number of Rows and Columns:**

here in this dataset, there are 4 number of columns and 99457 are number of rows.

**Variables:**

There are four number of variables(columns).

**Variable Names:**

customer\_id. Gender, age, payment\_method.

**Description:**

customer\_id store the id of the customer, gender show the gender of customer, age store the age of customer and payment method store the payment type.

**Data Types:**

customer\_id (object), Gender(object), age (float) and payment\_method(object)

1. **Clean the data: Handle missing values, outliers, and any inconsistencies.**

**Answer:**

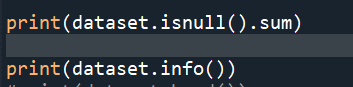
Cleaning the data is an essential step in the data preprocessing process to ensure the dataset is reliable and suitable for analysis. Here are common tasks for cleaning data, including handling missing values, outliers, and inconsistencies:

**Handling Missing Values:**

**Identify Missing Values:**

Use data. isnull (). sum () to check the number of missing values in each column.

Use data.info () to see which columns have missing values.



A screenshot of a computer program

Description automatically generated

Here in this dataset, there is no missing values in columns.

**Handling Outliers:**

**Identify Outliers:**

Visualize data with box plots or scatter plots to identify outliers.

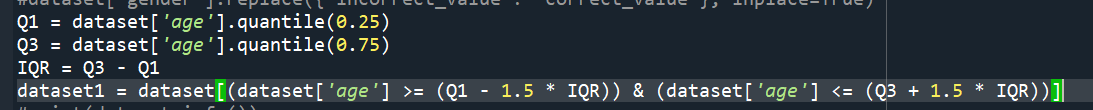
Calculate the Z-score or IQR (Interquartile Range) to detect outliers statistically.

**How to Handle Outliers:**

Outliers are extreme values that stand out greatly from the overall pattern of values in a dataset.

There is no outliers in my numerical column. Son if there in any data set and you want to remove there is a method.

**Example:** Removing outliers in a numerical column using the IQR method:

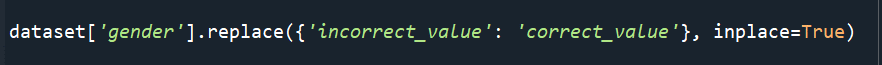


**Handling Inconsistencies:**

Check for data inconsistencies and errors, such as typos or incorrect values, and correct them.

**Example:** Replacing inconsistent values in a categorical column:

There is no Inconsistencies in my datasets but if found anywhere in dataset there is syntax to collect.

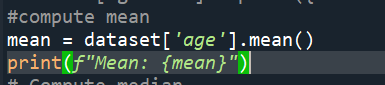


1. **Summarize the data: Compute basic statistics, such as mean, median, mode, standard deviation,**

**Ans:**

**Mean:**

The mean (average) of the 'numerical column’ in my dataset there is only one numerical columns “age”

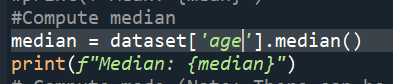


A screen shot of a computer code

Description automatically generated

**Median:**

The median (middle value) of the 'numerical column’ in my dataset there is only one numerical columns “age”.

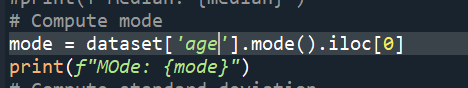


A computer screen shot of text

Description automatically generated

**Mode:**

The mode (most frequently occurring value) of the 'numerical column'. Note that it's possible to have multiple modes, but we typically report the first mode using iloc [0].



A screen shot of a computer code

Description automatically generated

**Standard deviation:**

The standard deviation of the 'numerical column', which measures the spread or dispersion of the data.

A screen shot of a computer code

Description automatically generated

A screenshot of a computer code

Description automatically generated

**Question 2: Fill in the table below with the require nature of data asked:**

1. Number of instances
2. Number of attributes
3. Categorical/Numerical
4. Nominal/ordinal
5. Classify as binary/discrete/continuous
6. Category names if categorical
7. Qualitative/quantitative
8. Missing values
9. Simple statistics i.e. Mean, median, mode, min, max etc.
10. Units if applicable
11. Description

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Attribute | a | b | c | d | E | F | g | h | i | j | k |
| Attrib 1 |  |  |  |  |  |  |  |  |  |  |  |
| Attrib 2 |  |  |  |  |  |  |  |  |  |  |  |
| Attrib 3 |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |

**Question 3:** Visualize the data: Create meaningful plots and visualizations to gain insights (e.g., histograms, scatter plots, bar charts, etc.) using matplotlib.pyplot and seaborn libraries and explain the plots briefly.

**Note:** The class section are divided into groups of 5/6 students each in the order of roll numbers and depending on the number of students that should not exceed 6 students in each group.

|  |  |
| --- | --- |
| Datasets | Groups |
| 1. **Adult Census dataset** | Group 1 |
| 1. **Abalone seashells** | Group 2 |
| 1. **Supermarket sales & customers dataset** | Group 3 |
| 1. **Heart Disease ucimlrepo (ID=45)** | Group 4 |
| 1. **Automobile dataset ucimlrepo (ID=9)** | Group 5 |