Assessment Submission of Data Analyst Role

Name: - Manav Kumar Chhetri

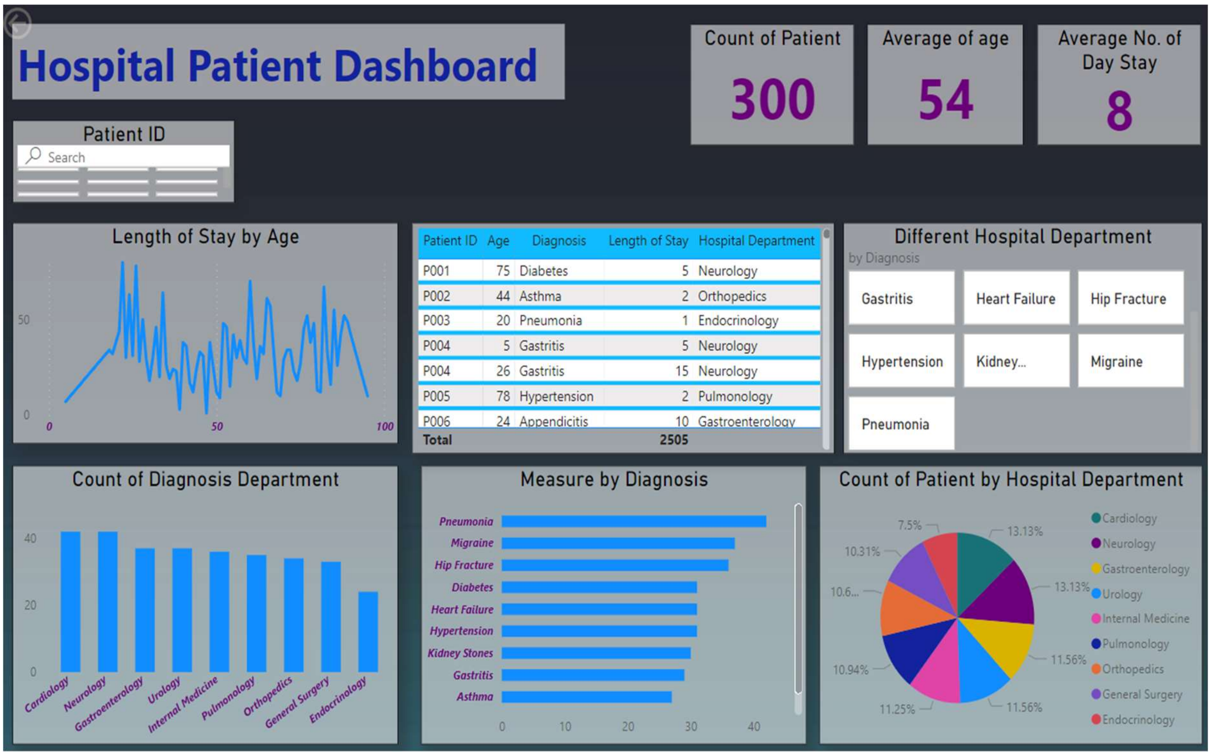
Email: - manavkumarchhetri.dcmc@gmail.com

Phone: - 7837951230

Task1

Screenshot

The Below dashboard is created for assessment task assign by the company to showcase my skill in the field of Data Analysis.



Patient ID	Age	Diagnosis	Length of Stay	Hospital Department
P001	75	Diabetes	5	Neurology
P002	44	Asthma	2	Orthopedics
P003	20	Pneumonia	1	Endocrinology
P004	5	Gastritis	5	Neurology
P004	26	Gastritis	15	Neurology
P005	78	Hypertension	2	Pulmonology
P006	24	Appendicitis	10	Gastroenterology
Total			2505	

Dashboard Creation

Dashboard Summary

The **Hospital Patient Dashboard** provides a comprehensive overview of key metrics and insights related to hospital patients' demographics, diagnoses, and length of stay. It enables stakeholders to monitor performance and make data-driven decisions effectively. Below are the key components of the dashboard:

1. Key Performance Indicators (KPIs):

- **Total Count of Patients**: **300** patients are included in the dataset.
- **Average Age of Patients**: **54** years, reflecting the demographic distribution of patients.
- **Average Length of Stay**: **8 days**, indicating the average duration patients remain in the hospital.

These KPIs provide a quick snapshot of the hospital's patient statistics.

2. Search and Filter Functionality:

- A **search bar** allows users to quickly locate data by entering specific **Patient IDs**. This enhances interactivity and usability, helping users focus on individual patient records.

3. Visualizations:

a. Length of Stay by Age:

- A **line chart** shows the relationship between patient age and length of stay. Patterns in the chart suggest varying hospital stays across different age groups, providing insights into which age groups tend to have longer stays.

b. Count of Diagnosis by Department:

- A **bar chart** represents the count of diagnoses handled by each hospital department. Departments like **Cardiology, Neurology, Gastroenterology**, and **Urology** are among the top contributors, highlighting the hospital's areas of expertise or high patient volume.

c. Measure by Diagnosis:

- A **horizontal bar chart** illustrates the frequency of various diagnoses (e.g., **Pneumonia, Asthma, Hypertension**). This helps identify the most common medical conditions treated.

d. Count of Patients by Hospital Department:

- A **pie chart** displays the percentage of patients treated in each department. The chart shows a fairly even distribution, with departments like **Cardiology, Neurology**, and **Gastroenterology** having the largest shares.

---

4. Detailed Data Table:

- A **data table** lists all patients with columns for:
  - **Patient ID**
  - **Age**
  - **Diagnosis**
  - **Length of Stay**
  - **Hospital Department**

This table enables a granular view of patient data for detailed analysis and reference.

---

5. Additional Filter Options:

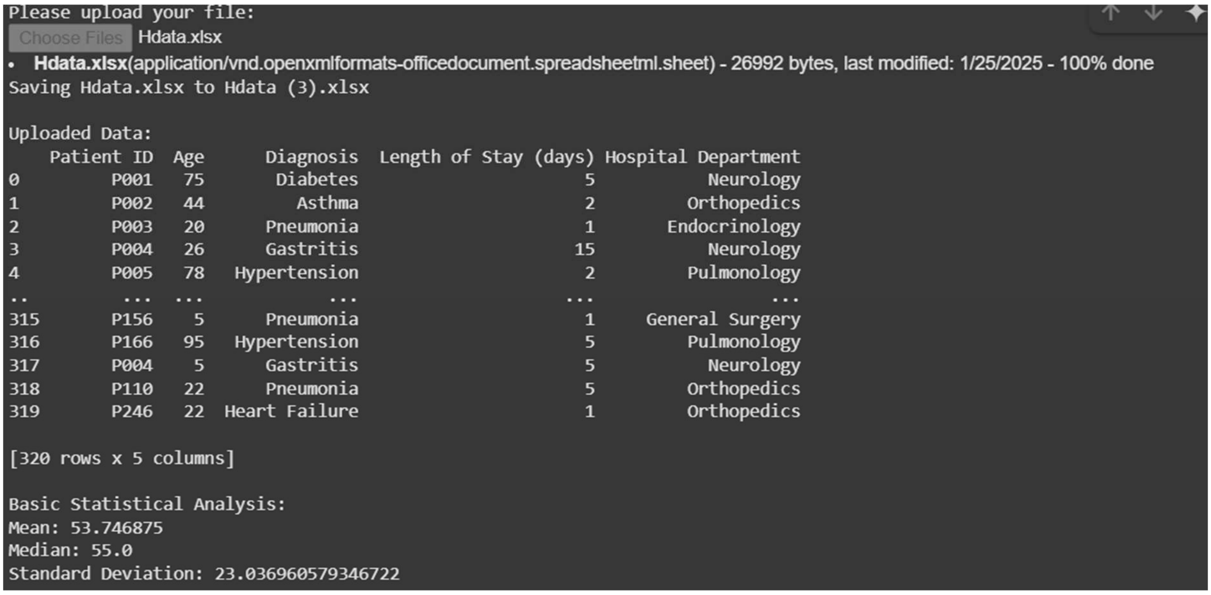
- **Diagnosis Tags**: Clickable diagnosis options (e.g., **Gastritis, Heart Failure, Migraine**) allow users to filter and analyze data based on specific diagnoses.

Insights and Utility:

- **Trends in Length of Stay**: The chart reveals how patient age affects hospital stays, which could inform staffing, resource allocation, and discharge planning.
- **Top Diagnoses and Departments**: Identifying the most common diagnoses and busiest departments helps allocate resources efficiently and improve patient care.
- **Patient Data Drill-Down**: The search and filtering capabilities allow for a deep dive into individual cases for targeted analysis.

Codebase Environment

I use Google Collab for my Data Analysis and EDA below is the screenshot of that and do some basic Statistical Analysis like Mean, Median and Standard Deviation.

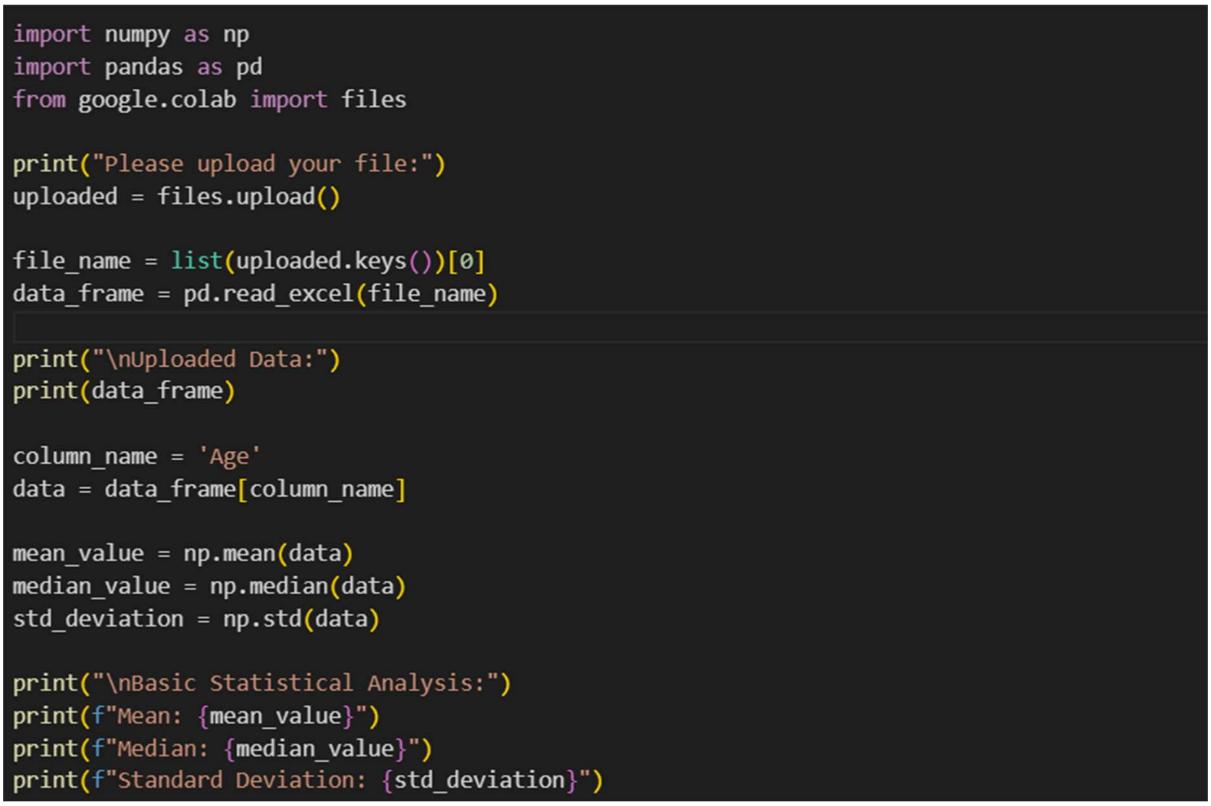


```
Please upload your file:
Choose Files  Hdata.xlsx
• Hdata.xlsx(application/vnd.openxmlformats-officedocument.spreadsheetml.sheet) - 26992 bytes, last modified: 1/25/2025 - 100% done
Saving Hdata.xlsx to Hdata (3).xlsx

Uploaded Data:
    Patient ID  Age     Diagnosis  Length of Stay (days) Hospital Department
0         P001   75      Diabetes                      5           Neurology
1         P002   44        Asthma                      2         Orthopedics
2         P003   20     Pneumonia                      1       Endocrinology
3         P004   26     Gastritis                     15           Neurology
4         P005   78  Hypertension                      2         Pulmonology
..         ...  ...           ...                    ...                 ...
315       P156    5     Pneumonia                      1     General Surgery
316       P166   95  Hypertension                      5         Pulmonology
317       P004    5     Gastritis                      5           Neurology
318       P110   22     Pneumonia                      5         Orthopedics
319       P246   22 Heart Failure                      1         Orthopedics

[320 rows x 5 columns]

Basic Statistical Analysis:
Mean: 53.746875
Median: 55.0
Standard Deviation: 23.036960579346722
```



```python
import numpy as np
import pandas as pd
from google.colab import files

print("Please upload your file:")
uploaded = files.upload()

file_name = list(uploaded.keys())[0]
data_frame = pd.read_excel(file_name)

print("\nUploaded Data:")
print(data_frame)

column_name = 'Age'
data = data_frame[column_name]

mean_value = np.mean(data)
median_value = np.median(data)
std_deviation = np.std(data)

print("\nBasic Statistical Analysis:")
print(f"Mean: {mean_value}")
print(f"Median: {median_value}")
print(f"Standard Deviation: {std_deviation}")
```

Challenges

1.Creation of the dataset that contain incorrect data or some inconsistence.

2. Then I clean the data and make it suitable for further process, for this I use power query in excel.

Task 2: Problem-Solving Scenario

● **Scenario**: Imagine you are tasked with analysing a dataset of patient records but find that a large portion of the data is missing or inaccurate. What steps would you take to clean and handle this data?

● **Deliverables**: A brief written response (250-300 words) explaining your approach.

Answer: - We can use the following approach in this scenario to solve the problem of the missing and incorrect dataset. In this I use power query for the same as it is simple, easy and more powerful.

Understanding the Dataset

Start loading the dataset and doing exploratory data analysis (EDA) to see how many missing or inaccurate values there are. You should focus on columns that are necessary for your analysis (such as age, diagnosis), while others can have missing values.

Handling Missing Data

Missingness: categorize it so that you know if your data is missing at random or systematically. On which, implement suitable approaches: Imputation: For numerical data, mean, median or predictive models (Neymar or KNN). Using the mode to fill missing values for categorical data or using machine learning modes for prediction.

Correcting Inaccurate Data

Checking the data against the known medical standard (age < 150 years, blood pressure in an acceptable range, etc_) Verify conflicting records (e.g., patient IDs that do not match) against other fields or subsequent external sources. For important errors, use domain experts for corrections.

Documentation

Keep an extensive record of transformations, including the imputation methods used, thresholds for any data purge or filtration, and corrections made; This is to maintain transparency and reproducibility.

Iterative Review

Clean Data – Run quality assurance checks by reviewing summary statistics and conducting sample data checks. Work with stakeholders (e.g., clinicians) to confirm that the dataset is now ready for analysis.

Task 3: Multiple-Choice Questions (MCQs)

1.Which of the following is NOT a typical step in data cleaning?

a) Removing duplicate rows

b) Filling missing data with random values

c) Standardizing format

d) Identifying outliers

Answer: -b) Filling missing data with random values.

Explanation: - Filling missing data with random values is not a recommended or typical step in data cleaning because it can introduce noise and inaccuracies.

2.What is the purpose of normalization in data analysis?

a) To reduce the size of the data

b) To ensure all variables are on a similar scale

c) To remove duplicates from the data

d)To convert data into categorical variables

Answer: - b) To ensure all variables are on a similar scale.

Explanation: - Normalization is used to transform features to be on a similar scale, especially for algorithms that rely on distance measurements