**Task 1:Data Analysis & Visualisation**

**A screenshot of a medical report

Description automatically generated**

**Created from Tableau**

**My tableau vizz link:** [**Hospital Patient Trends Insight | Tableau Public**](https://public.tableau.com/app/profile/akhilesh.kumar.singh8760/viz/HospitalPatientTrendsInsight/Dashboard1)

**Step1: Dataset is created from Python Random data generated from code using google colab**

import pandas as pd

import random

def generate\_hospital\_data(rows=500):

columns = ['Patient ID', 'Age', 'Diagnosis', 'Length of Stay (days)', 'Hospital Department']

diagnosis\_options = ['Pneumonia', 'Hip Fracture', 'Appendicitis', 'Migraine', 'Diabetes Management',

'Asthma', 'COVID-19', 'Cancer', 'Heart Disease', 'Stroke']

department\_options = ['Internal Medicine', 'Orthopedics', 'General Surgery', 'Neurology', 'Endocrinology',

'Pulmonology', 'Cardiology', 'Oncology', 'Radiology', 'Emergency']

data = [

[

i + 1,

random.randint(18, 90),

random.choice(diagnosis\_options),

random.randint(1, 15),

random.choice(department\_options)

]

for i in range(rows)

]

return pd.DataFrame(data, columns=columns)

data = generate\_hospital\_data()

data.to\_excel("hospital\_data.xlsx", index=False)

print("dataset with 500 rows created and saved as 'hospital\_data.xlsx'.")

**Step2: Cleaning Dataset or either checking null values to remove by mean.**

**Step3: Performing basic Statistical analysis**

**Step4:Creating visualization.**

**A screenshot of a graph

Description automatically generatedSome portion is analyzed in excel**

**Formulas used to create table:**

* Avg length of stay: =AVERAGE(D2:D501)
* Count of patient by Diagnosis: =COUNTIF(C2:C501,H6)
* diagnosis percent: =COUNTIF(C2:C501, H6)/COUNTA(C2:C501)\*100
* MOST FREQUENT DIAGNOSIS: =INDEX(C2:C501, MODE(MATCH(C2:C501, C2:C501, 0)))

**A chart with numbers and a number of numbers

Description automatically generated with medium confidencethis table is of total patient by diagnosis department.**

**Formulas used to create table:**

* Count: =COUNTIF(E2:E501,L6)
* avg length of stay: =AVERAGEIF(E2:E501,L6,D2:D501)
* total length of stay by department: =SUMIF(E2:E501, L6, D2:D501)

**this table is created to understand the patterns behind that**

* most frequently diagnosis are occurred is Stroke.
* Average length of stay is of 8.17
* Top patient by diagnosis(department) is Endocrinology
* From down patient by diagnosis(department) is orthopedics.

**My tableau vizz link:** [**Hospital Patient Trends Insight | Tableau Public**](https://public.tableau.com/app/profile/akhilesh.kumar.singh8760/viz/HospitalPatientTrendsInsight/Dashboard1)

**Task 2: Problem-Solving Scenario**

1. Understand the Data:

* Learn what each column means and identify the important ones.
* Understand the purpose of the analysis to know which data is most critical.

2. Check Missing/Inaccurate Data:

* Count how much data is missing or wrong and note where the problem is.
* Look for patterns in missing data (e.g., specific rows or columns).
* Use visual tools like heatmaps or charts to highlight missing values.

3. Group Missing Data:

* Some data is missing randomly.
* Some is missing because of other data.
* Some is missing for unknown reasons.
* Identify if missing data is related to certain patient groups or variables.

4. Fix Missing Data:

* Delete: Remove rows or columns if very little data is missing and it’s not important.
* Fill Values: Fill missing data with:
* Average, most common value, or similar data.
* Predictions using other data.
* Domain-specific rules (e.g., filling age based on other demographics).
* Add Flags: Mark the missing data to keep track of it.
* Use advanced methods like machine learning for large, complex datasets.

5. Fix Wrong Data:

* Check if data fits expected values (e.g., valid age range, correct date formats).
* Correct mistakes (e.g., typos, formatting issues) or replace wrong values.
* Standardize data to ensure consistency (e.g., units, date formats).
* Remove duplicate records to avoid errors.

6. Handle Outliers:

* Identify values that are too high or low to be realistic.
* Decide whether to keep, correct, or remove outliers based on context.

7. Test and Validate:

* Run checks to ensure the cleaned data makes sense and matches expectations.
* Test the data in small analyses to confirm it works properly.

8. Document and Share:

* Write down all changes, methods, and assumptions for transparency.
* Share the cleaned dataset and explanation with stakeholders or experts for review.

**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 formats

d) Identifying outliers

**Option b) Filling missing data with random values**

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

**Option b) To ensure all variables are on a similar scale**