Final Research Reports:

First Importing necessary libraries

Pandas : Used for data manipulation and analysis.

import pandas as pd

Numpy: A fundamental package for scientific computing in Python.

import numpy as np

Matplotlib: A comprehensive library for creating static, animated, and interactive visualizations in Python.

import matplotlib.pyplot as plt

Seaborn: A library for creating aesthetically pleasing statistical graphics in Python.

Import seaborn as sns

Loading the datasets

Then reading a csv file and import it into a pandas **df using df = pd.read_csv(filepath)**

Then perform some Statistical Analysis:

- By using df.head() we get to know quick view of datasets, columns name and which type of data it will contains like character, numerical or binary.
- By using df.info() get information about Total No. of Columns and Rows, their datatypes and any null values they contains.
  Also view of all columns name and their memory used by them.
  Here 253680 rows and 22 columns as present.
- df.describe(): Give brief statistical summary of datasets including mean, median , standard deviations, min & max value and more.
- df.isnull().sum() : get to know if any columns contains any null values, if contain then show their total counts associated with each columns.

Uni-Variate Analysis

Examine Distributions of Individual Variables

1. **Histogram chart**: This histogram shows that the frequency of individuals in the dataset with in age groups.
   The bars on the left, representing younger age groups, are shorter. As the age group number increases, the height of the bars also increases significantly, with the tallest bars at the right end of the plot (age groups 10, 12, and 13).

This pattern indicates that the dataset is **positively skewed toward older individuals**, meaning there are more people in the older age groups than in the younger ones

2. **Bar chart**: Plotting the BMI columns to know the frequency based on BMI

The distribution is **positively skewed**, meaning the long tail of the distribution extends to the right.

This shows that while most people have a BMI in the normal or overweight range, there are a few individuals with very high BMI values, representing a smaller percentage of the population.

3. **Histogram plot**: Plot of **MentHlth** columns to know through visualizaton 'how many

people have not good their mental health from last past days.

Here we can see **175680** people have not good their mental health form last 0 days and similar to form 2-30 days onwards.

This indicates that for the majority of the population surveyed, mental health was stable and good.

4. **Histogram plot**: Plot of **PhysHlth** columns to know through visualizaton 'how many people have not good their physical health from last past days’.

Here we can see 160052 people have not good their Physical health form last 0 days and similar to form 2-30 days onwards.

This indicates that for the majority of the population surveyed, Physical health was stable and good.

Investigate Prevalence of Health Conditions:

5. **Bar plot**: showing the distribution of HighBP among people through visualization.

   Here we find that 144851 people have No HighBP and 108829 people have HighBP.

6. **Pie chart**: Showing the distribution of HighBP & their percentage among people through visualization.
   Here we find that 57.10% people have No HighBP and 42.90% have HighBP.

7. **Bar plot**: Showing the distribution of High Cholesterol among people through visualization.
   Here we find that 146089 people have No HighCohl and 107591 people have Highcohl.

8. **Pie chart**: showing the distribution of Highcohl & their percentage among people through visualization.
   Here we find that 57.59% people have No Highcohl and 42.41% have Highcohl.

9. **Bar plot**: Showing the distribution of smokers among individuals through visualization.

Here we find that 141257 people have No smoked and 112423 people have smoked at least 5 packs of cigarettes.

- The **green bar** for "Nosmoked" is taller, indicating there are **more non-smokers** in the dataset. The number is around 140,000.

- The **orange bar** for "Smoked" is shorter, indicating there are **fewer smokers** in the dataset. The number is around 110,000.

Analyze Distribution of Heart Disease (Target Variable)

10. **Pie chart**: Showing the distribution of Healthdisease or Attack & their percentage among people through visualization.

Here we find that **90.58%** people have No heartdisease of Attack and **9.42%** have Heartdisease or Attack.

11. **Bar chart**: showing the parts of genders among datasets.
    Here we find that **141974** female and **111706** male are present in give datasets.

Bi-variate Analysis

12. **Determine relationship between High BP & HeartDiseaseorAttack**

Crosstabulation: Make the cross table to determine the relationship.

The resulting table shows the percentage of people in each high blood pressure group who have or don't have heart disease.

Bar charts: Show the relationship between HighBP status with Heart Diseases.
here we find that there are :
60.44 % of people have NO heart disease & have No highBP
24.69% of people have Heart disease with NO highBP
39.55% of people have HighBP with no Heart disease.
75.05% of people have HighBP and have Heart disease.
Means, here people with highBP have high chance of Heart disease and people with No highBP have less chance of any heart disease or Attack.

13.**Determine relationship between High Cholesterol & HeartDiseaseorAttack**

Crosstabulation: Make the cross table to determine the relationship.

The resulting table shows the percentage of people in each Cholesterol group who have or don't have heart disease.

Bar charts: Show the relationship between HighChol groups with Heart Diseases.
here we find that there are:
60.46 % of people have NO heart disease & have No highchol.
29.88% of people have Heart disease with NO highchol.
39.53% of people have NO heart disease with highchol.
70.11% of people have Heardisease and have Highchol.
Means, here people have highchol have high chance of Heart disease.
and no highchol have less change of Heart disease.

14.**Plotting boxplot to analyse BMI distrubution by heart disease status.**

Median BMI:
No Heart Disease: The median BMI is around 27.
Heart Disease: The median BMI is slightly higher, around 29.
This suggests that, on average, individuals with heart disease in this dataset tend to have a slightly higher BMI than those without it.

BMI Distribution (The Box):
The middle 50% of people without heart disease have a BMI between approximately 25 (Q1) and 31 (Q3).
The middle 50% of people with heart disease have a BMI between approximately 26 (Q1) and 32 (Q3).
The box for the "Heart Disease" group is positioned slightly higher on the y-axis, indicating a generally higher range of BMI values for this group compared to the "No Heart Disease" group.

15. **Visualizing Correlations Between Variables.**

A correlation matrix shows how different variables relate to each other. The values in the matrix, called correlation coefficients, range from -1 to +1.
Key Components:
The Numbers: The values indicate the strength and direction of the relationship.
+1.00: A perfect positive correlation. This means when one variable increases, the other increases in perfect proportion.
-1.00: A perfect negative correlation. This means when one variable increases, the other decreases in perfect proportion.
0: No linear relationship.
Values closer to +1 or -1 show a stronger relationship, while values closer to 0 show a weaker relationship.

Here, in correlation heatmap of Income and BMI shows that:
Diagonal Cells: The cells where a variable is **correlated with itself** (e.g., "Income" with "Income") always have a value of **1.00.**
This is expected, as a variable is perfectly correlated with itself.

Off-Diagonal Cells: The cells comparing two different variables (e.g., "Income" and "BMI") contain the meaningful information.
The value is **-0.10**. This is a **weak negative correlation** between income and BMI.
In context, this means that as an individual's income increases, their BMI tends to slightly decrease, but the relationship is not strong.

16. **Relationship between Age and BMI using scatter plot**

Scatter plot: It displays the relationship b/w two values for typically two variables for a set of data points.

Here, we find that there is no clear upward or downward trend across the age groups

indicating there is no linear correlation between BMI and age group.

17. **Identify the correlation between Mental health & their smoking status.**

Here we try to identify the correlation between mental health & their smoking status.
we find that
Diagonal Cells: The cells where a variable is **correlated with itself** (e.g., "menthlth" with "menthlth" & 'smoker' with 'smoker') always have a value of **1.00**.
This is expected, as a variable is perfectly correlated with itself.

Off-Diagonal Cells: The cells comparing two different variables 'smoker' and 'menthlth' contain the meaningful information.
The value is **0.09**, this is a **weak positive correlation** between menthlth and smoker.
In context, this means that as an individual's smoking increases, their menthlth tends to **slightly increases**, but the relationship is not strong positive.

18. **Identify the correlation between HeartdiseaseorAttack & their smoking status.**

we find that:

Diagonal Cells: The cells where a variable is **correlated with itself** (e.g., "HeartDiseaseorAttack" with "HeartDiseaseorAttack" & 'smoker' with 'smoker') always have a value of **1.00**.

This is expected, as a variable is perfectly correlated with itself.

Off-Diagonal Cells: The cells comparing two different variables 'smoker' and 'HeartDiseaseorAttack' contain the meaningful information.

The value is **0.11**, This is a weak positive correlation between HeartDiseaseorAttack and smoker.

In context, this means that as an individual's smoking increases, their HeartDiseaseorAttack tends to slightly increases, but the relationship is not strong positive.

Compare Heart Disease Across Demographic Groups

19. **Bar graph:** This show the percentage of HeartDiseaseorAttack with different age groups:

First calculates the prevalence of heart disease for each age group and stores it in a new table called `age_prevalence`.

By using df.pivot_table() function , it will contain two columns Age and HeartdiseaseorAttack .

Use `aggfunc='mean'`: It calculates the **mean** (average) of the `HeartDiseaseorAttack` values for each age group.

`*100`: This multiplies the proportion by 100 to convert it into a **percentage**, making the result easier to interpret.

here we find the **clear upward trend**: as the age group number increases, the percentage of individuals with heart disease consistently and **significantly rises**.

This demonstrates a **strong positive correlation** between age and the prevalence of heart disease.

20. **Bar graph**: This show the percentage of HHeartDiseaseorAttack with different Education level:

Again first calculates the prevalence of heart disease for each Education level and stores it in a new table called `Education_prevalence`.

By using df.pivot_table() function , it will contain two columns Education level and HeartdiseaseorAttack.

Use `aggfunc='mean'`: It calculates the **mean** (average) of the `HeartDiseaseorAttack` values for each Education level.

`*100`: This multiplies the proportion by 100 to convert it into a **percentage**, making the result easier to interpret.

Here find as an individual's education level **increases**, the prevalence of heart disease or a heart attack **decreases**.

This graph gives a **negative correlation** between education level and the prevalence of heart disease, meaning higher education levels are associated with a lower risk of heart disease.