REPORT ON MINI PROJECT

Name	: Kalaivanan
Batch	: TN_DA_FNB07
Contact No.	: +966 - 509143765 +91- 9901710159
Email ID	: kalaivanan102@gmail.com
Project Title	: Patients Health Record
Project Domain	: Healthcare
Submission Date	: 10-Dec-25
Raw Dataset Link	: Github Link
Cleaned Dataset Link	:Github Link

Introduction:

This project involves analysing a healthcare dataset containing patient demographics, medical readings, lifestyle indicators, and diagnostic information. The objective is to clean raw data, prepare a structured dataset, and build a Power BI dashboard to identify key health insights.

Purpose Of Project

The purpose of this project is to analyze healthcare patient records by comparing a *raw dataset* and a *cleaned dataset*, and to generate meaningful insights that support better decision-making in healthcare management. To clean, transform, and analyze healthcare patient data to generate accurate insights and build an interactive Power BI dashboard for informed healthcare decision-making.

Objectives Of The Project:

To clean and standardize the raw healthcare dataset

- Remove missing, duplicate, and inconsistent records
- Fix formatting issues in date, gender, diabetic status, BP, etc.
- Convert noisy text and emojis into meaningful, usable data

To prepare a reliable, analysis-ready cleaned dataset

- Ensure correct data types
- Normalize categorical values
- Maintain accuracy across all patient health fields

- Improve dataset consistency for analytics

To analyze key health indicators in the patient population

- Study patterns in BMI, blood pressure, cholesterol, diabetes, smoking, and medications
- Identify high-risk patient groups
- Understand correlations between variables

To build interactive Power BI visualizations

- Create dashboards for patient health distribution
- Display KPIs like high-risk cases, disease counts, average metrics
- Enable filtering by age, gender, disease, diabetes, smoking, etc.

To assist healthcare decision-making through insights

- Help doctors and administrators identify trends
- Support early detection of risks
- Provide interpretable visuals for quick diagnosis planning

To compare the raw and cleaned datasets for quality improvement

- Show difference in data quality
- Highlight cleaning steps and impact
- Ensure transparency in the data transformation process

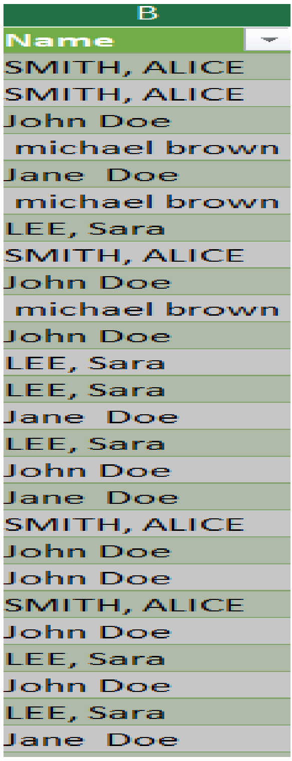
DATA CLEANING USING EXCEL (COLUMN WISE):

1. Patient ID

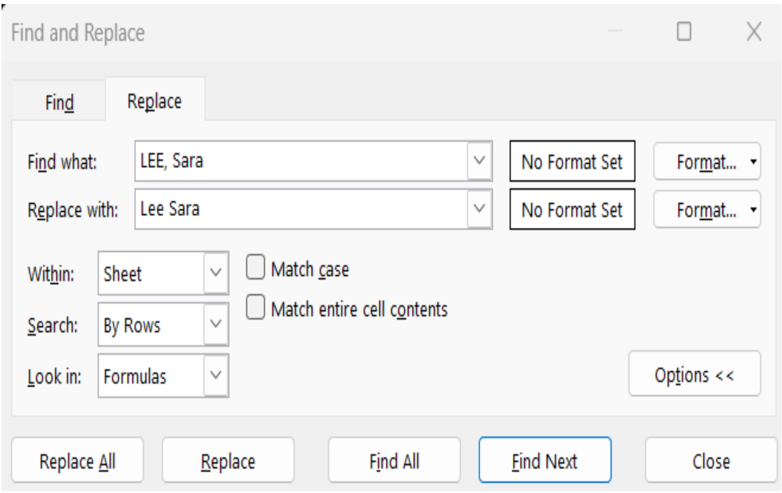
- Missing / blank IDs checked and removed
- Duplicates removed
- Converted to consistent format

2. Name

- Trimmed extra spaces
- Removed special characters and Proper Case



B
Name
SMITH, ALICE
SMITH, ALICE
John Doe
michael brown
Jane Doe
michael brown
LEE, Sara
SMITH, ALICE
John Doe
michael brown
John Doe
LEE, Sara
LEE, Sara
Jane Doe
LEE, Sara
John Doe
Jane Doe
SMITH, ALICE
John Doe
John Doe
SMITH, ALICE
John Doe
LEE, Sara
John Doe
LEE, Sara
Jane Doe



Find and Replace

Find what: LEE, Sara — No Format Set

Replace with: Lee Sara — No Format Set

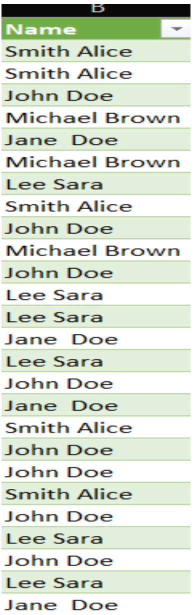
Within: Sheet

Search: By Rows

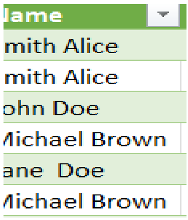
Look in: Formulas

□ Match case

□ Match entire cell contents



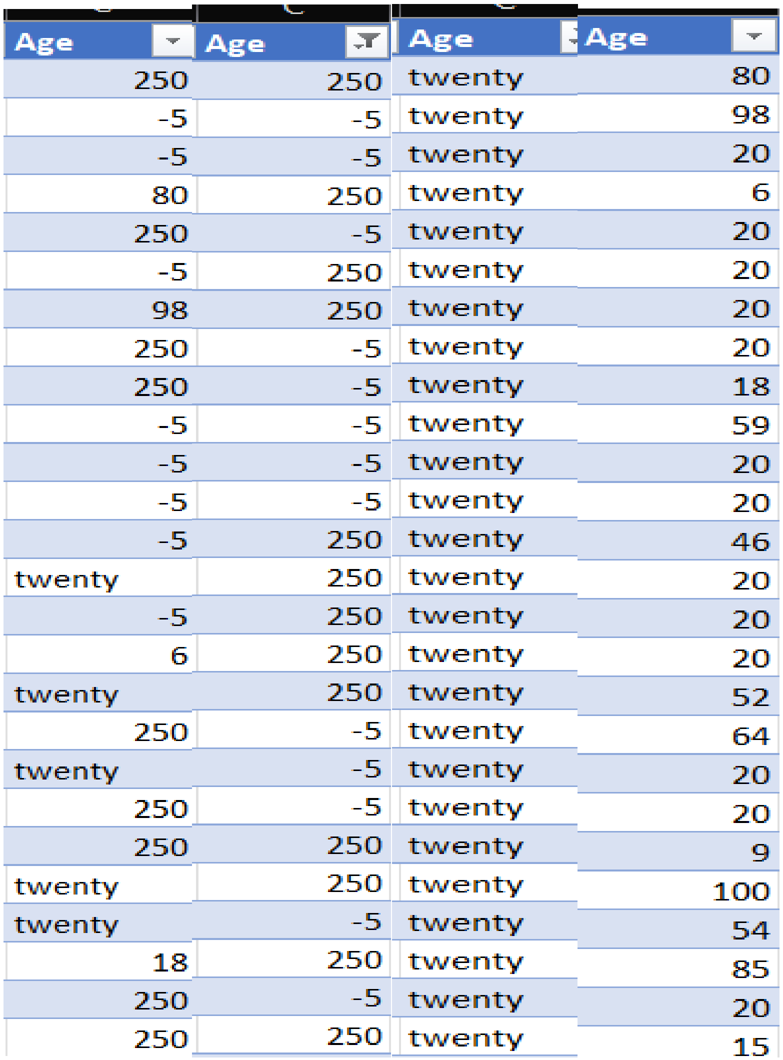
B
Name
Smith Alice
Smith Alice
John Doe
Michael Brown
Jane Doe
Michael Brown
Lee Sara
Smith Alice
John Doe
Michael Brown
John Doe
Lee Sara
Lee Sara
Jane Doe
Lee Sara
John Doe
Jane Doe
Smith Alice
John Doe
John Doe
Smith Alice
John Doe
Lee Sara
John Doe
Lee Sara
Jane Doe



lame
mith Alice
mith Alice
ohn Doe
ichael Brown
ane Doe
ichael Brown

3. Age

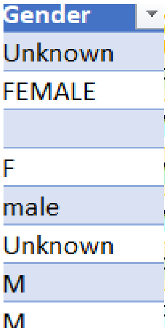
- Converted to numeric
- Fixed invalid ages
- Filled missing values using median or logical estimation
- Outliers detected & validated



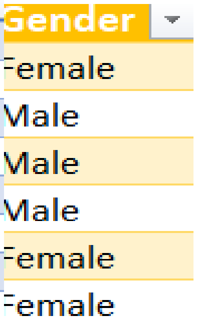
Age	Age	Age	Age
250	250	twenty	80
-5	-5	twenty	98
-5	-5	twenty	20
80	250	twenty	6
250	-5	twenty	20
-5	250	twenty	20
98	250	twenty	20
250	-5	twenty	20
250	-5	twenty	18
-5	-5	twenty	59
-5	-5	twenty	20
-5	-5	twenty	20
-5	250	twenty	46
twenty	250	twenty	20
-5	250	twenty	20
6	250	twenty	20
twenty	250	twenty	52
250	-5	twenty	64
twenty	-5	twenty	20
250	-5	twenty	20
250	250	twenty	9
twenty	250	twenty	100
twenty	-5	twenty	54
18	250	twenty	85
250	-5	twenty	20
250	250	twenty	15

4. Gender

- Standardized to Male / Female / Other
- Fixed inconsistent entries:
    - “M”, “male”, “m” → Male
    - “F”, “female”, “f” → Female



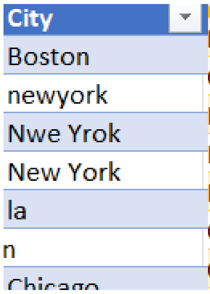
Gender
Unknown
FEMALE
F
male
Unknown
M
M



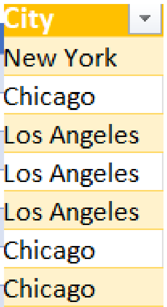
Gender
Female
Male
Male
Male
Female
Female

5. City

- Capitalization standardized
- Spelling mistakes corrected



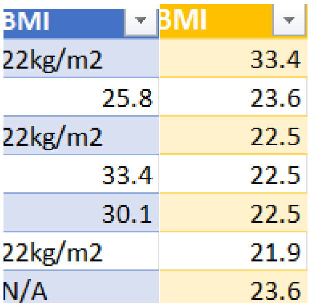
City
Boston
newyork
Nwe Yrok
New York
la
n
Chicago



City
New York
Chicago
Los Angeles
Los Angeles
Los Angeles
Chicago
Chicago

6. BMI

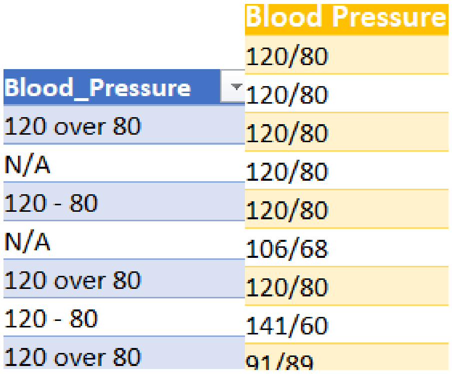
- Converted datatype to decimal
- Removed outliers
- Filled missing values using median BMI



BMI	BMI
22kg/m2	33.4
25.8	23.6
22kg/m2	22.5
33.4	22.5
30.1	22.5
22kg/m2	21.9
N/A	23.6

7. Blood Pressure

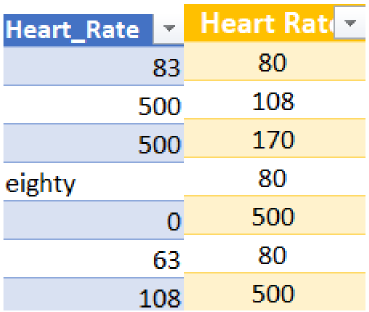
- Removed text noise (“120/80 mmHg” → “120/80”)
- Converted to numeric
- Outlier handling



	Blood Pressure
	120/80
Blood_Pressure	120/80
120 over 80	120/80
N/A	120/80
120 - 80	120/80
N/A	106/68
120 over 80	120/80
120 - 80	141/60
120 over 80	91/89

8. Heart Rate

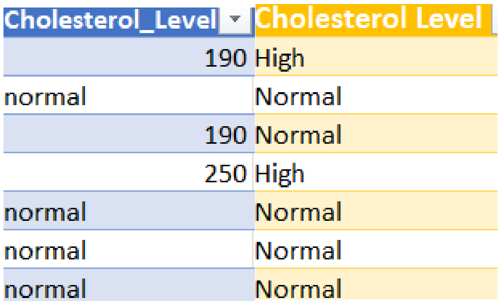
- Converted to numeric
- Invalid values corrected
- Missing values filled with mean Heart Rate



Heart_Rate	Heart Rat
83	80
500	108
500	170
eighty	80
0	500
63	80
108	500

9. Cholesterol Level

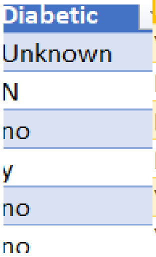
- Converted to numeric
- Standardize Number Fomat into valid appropriate Values
- Missing values replaced with mode
- Consider Above 200 as High ,Below 200 Normal



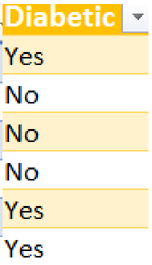
Cholesterol_Level	Cholesterol Level
190	High
normal	Normal
190	Normal
250	High
normal	Normal
normal	Normal
normal	Normal

10. Diabetic

- Standardized to **Yes / No**
- Fixed entries:
  - “Y”, “yes” → Yes
  - “N”, “no” → No



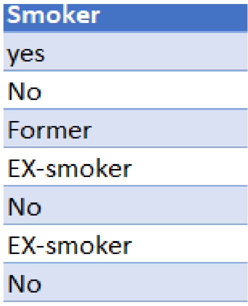
Diabetic
Unknown
N
no
y
no
no



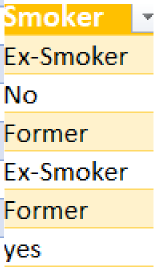
Diabetic
Yes
No
No
No
Yes
Yes

11. Smoker

- Standardized to **Smoker / Non-Smoker**
- “yes”, “Y” → Smoker
- “no”, “N” → Non-Smoker



Smoker
yes
No
Former
EX-smoker
No
EX-smoker
No



Smoker
Ex-Smoker
No
Former
Ex-Smoker
Former
yes

Final Cleaned Dataset:

✓ All values standardized

✓ Missing values handled

✓ Outliers corrected

✓ Categories normalized

✓ Text cleaned

✓ Dates and numeric fields fixed

✓ Dataset converted into fully analysis-ready format

Data Visualization Using Powerbi

1.Cards

◈ Average BMI – 23.61

Shows the overall average Body Mass Index of patients.

- **Indicates that the population is mostly within the *normal BMI range*.**
- **Useful for understanding general health profile.**

◈ Average Heart Rate – 210.91

Displays the mean heart rate across patients.

- **High value suggests either outliers or a large number of patients with abnormal heart rate.**
- **Helps identify cardiovascular risk groups.**

◈ Diabetic Count – 2170

The total number of patients marked as diabetic.

- **Useful for resource planning, medication allocation, and targeted interventions.**

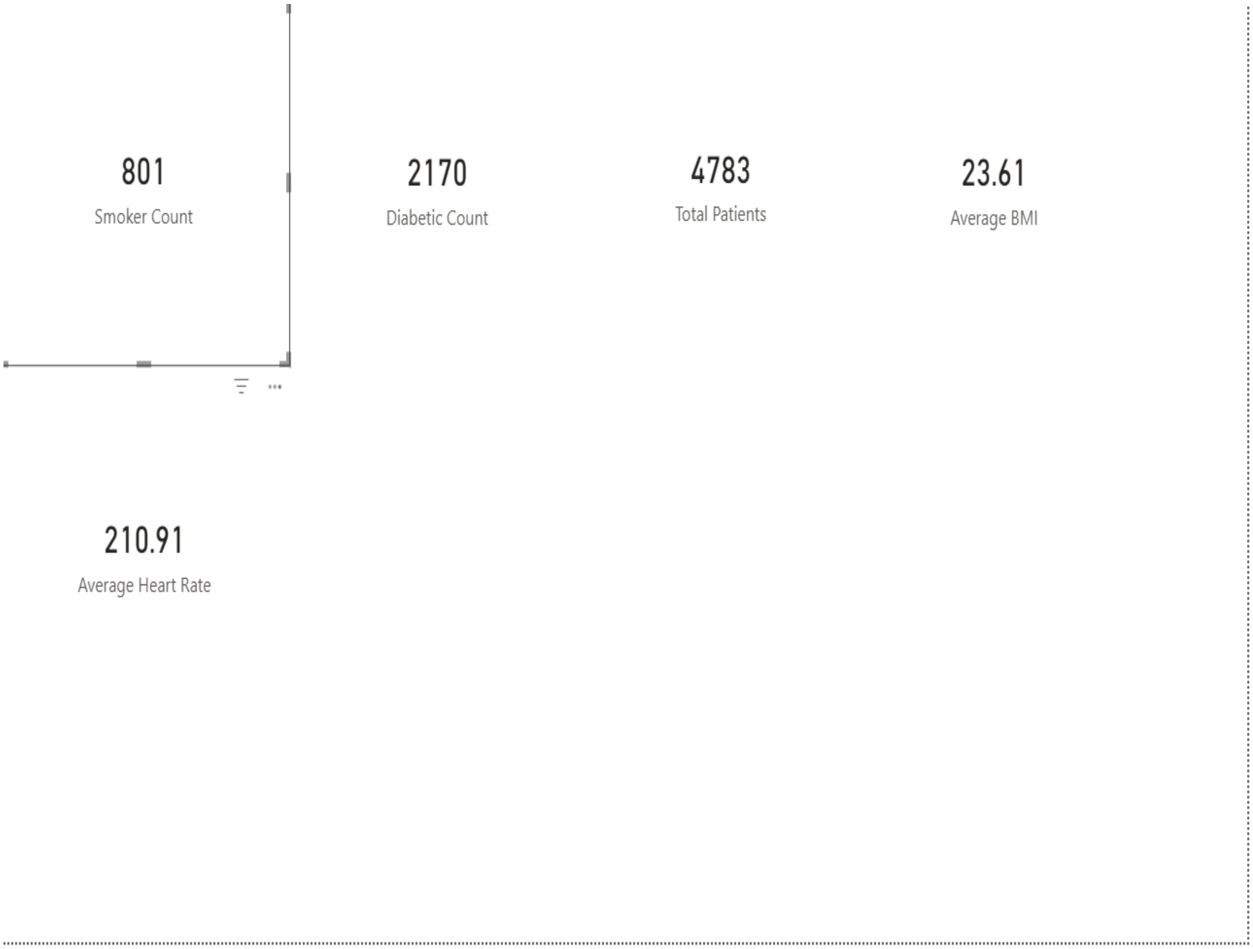
◈ Smoker Count – 801

Total number of current smokers.

- **Important for risk analysis related to lung issues, cardiovascular diseases, and BMI variations.**

◈ Total Patients – 4783

Displays the dataset size and total number of patients analyzed.



801
Smoker Count

2170
Diabetic Count

4783
Total Patients

23.61
Average BMI

210.91
Average Heart Rate

2. Pie Chart — Patients by City and Smoker Status

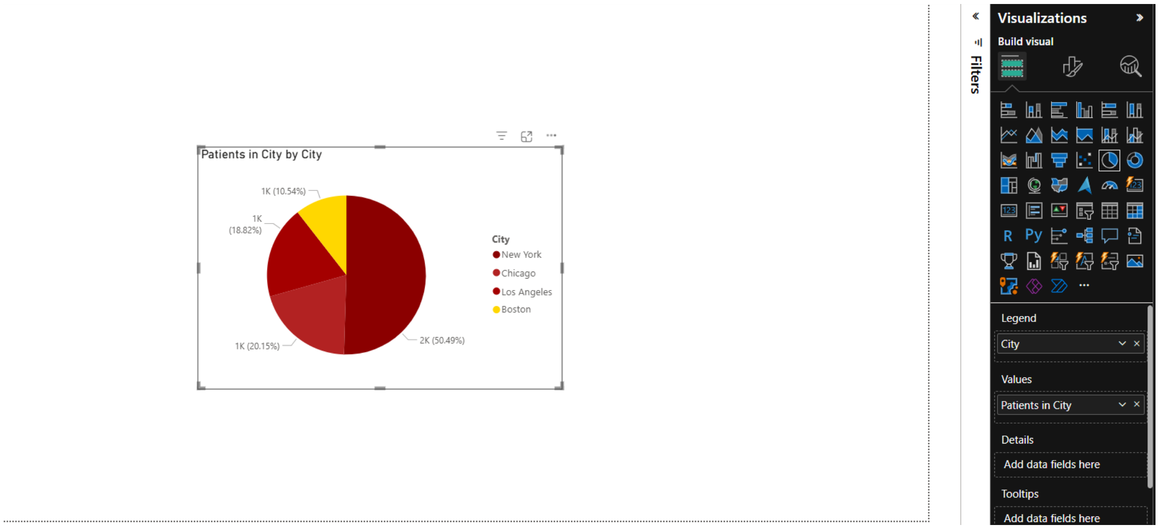
Represents the proportion of smokers across different cities:

- New York
- Chicago
- Los Angeles
- Boston

What it shows:

- Major chunk of patients are from New York.
- Chicago, LA, and Boston have smaller segments.
- Sectors are color-coded for clarity.

This helps identify city-level smoking patterns.



Legend
City

Values
Patients in City

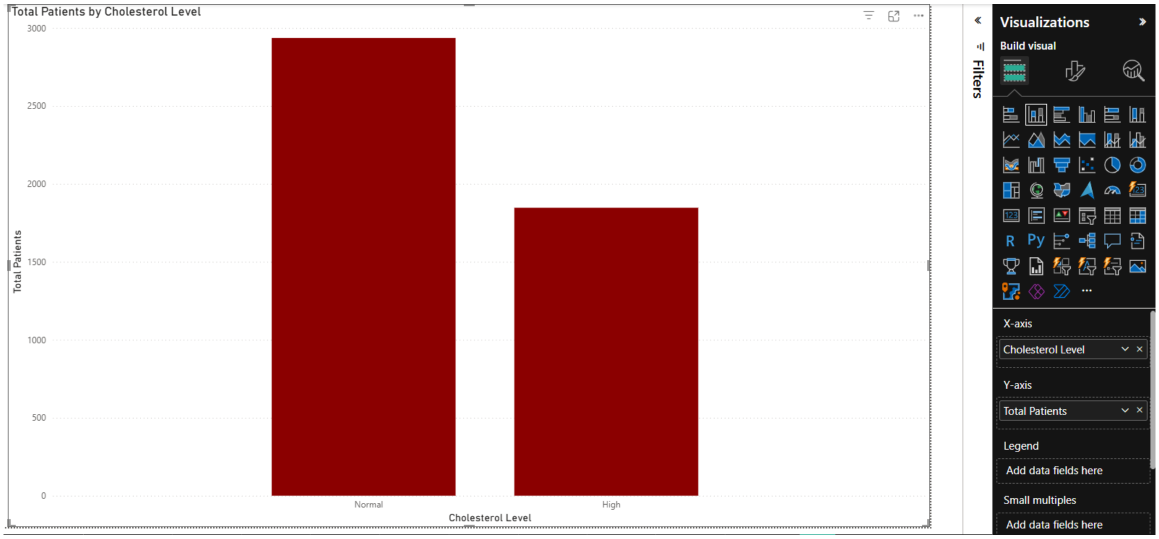
3. Bar Chart — Total Patients by Cholesterol Level

Visualizes patient distribution across Cholesterol categories (Normal vs High).

Insights:

- Normal cholesterol group is larger.
- High cholesterol group is significant → indicates risk.

Useful for analyzing metabolic health across the population.



X-axis

Cholesterol Level

Y-axis

Total Patients

Legend

Add data fields here

Small multiples

Add data fields here

4. Clustered Bar Chart — Average Heart Rate by Smoker Type

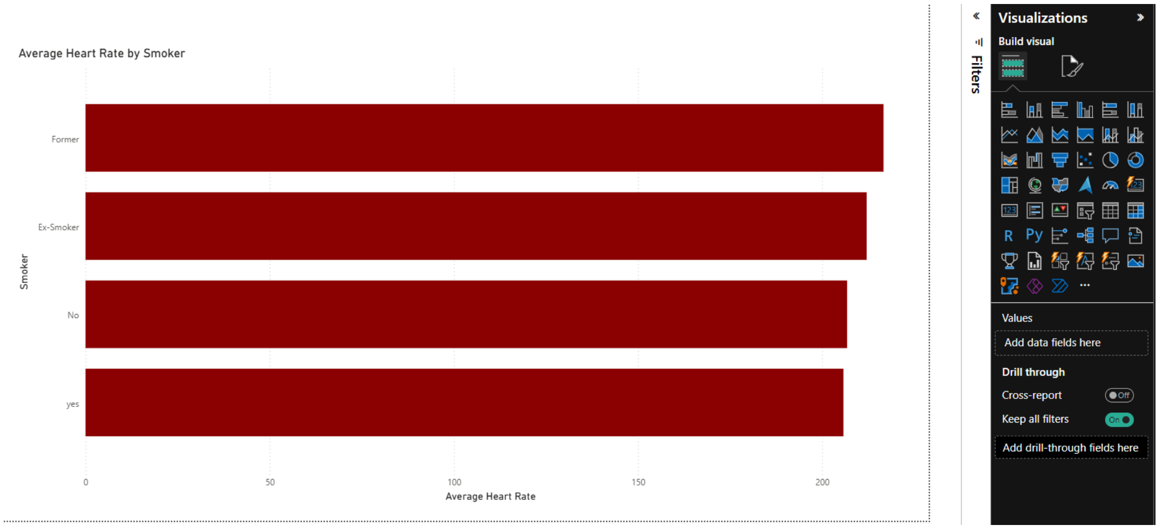
Compares heart rate across smoking categories:

- **Yes**
- **No**
- **Ex-Smoker**
- **Former**

Insights:

- **Heart rate is highest among active smokers.**
- **Former and ex-smokers show slightly lower average heart rate.**
- **Non-smokers show better stability.**

This supports the correlation between smoking and elevated heart rate.



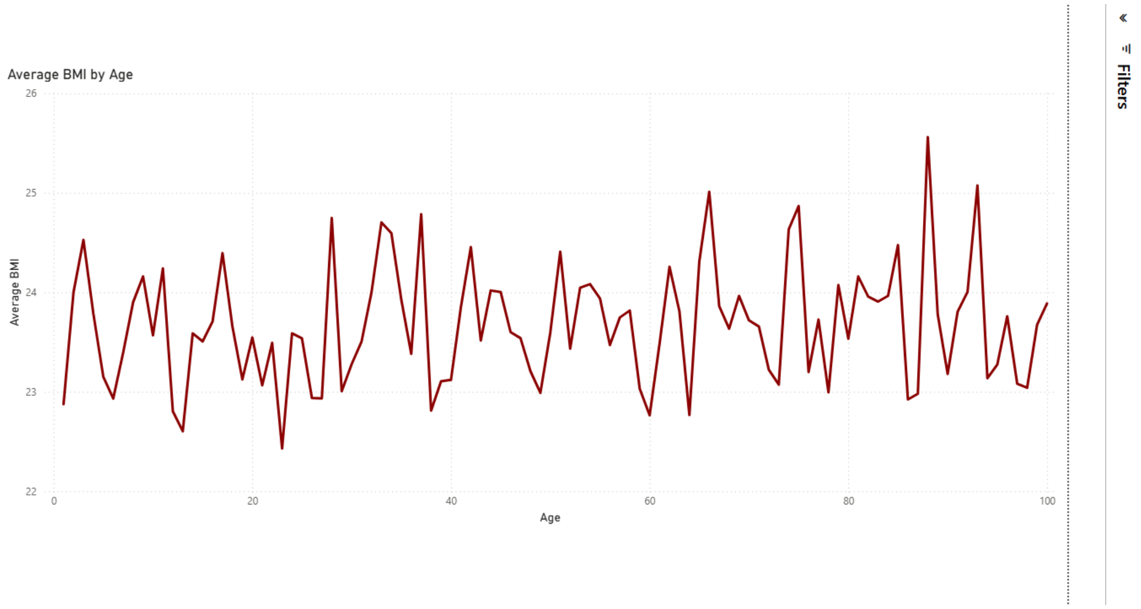
5. Line Chart — Average BMI by Age

Plots Age (X-axis) against Average BMI (Y-axis).

Insights:

- **BMI fluctuates across age groups but mostly stays between 23–25.**
- **Slight peaks can be seen around mid-age (40–60).**
- **Indicates that BMI rises slightly with age.**

Useful for age-based risk segmentation.

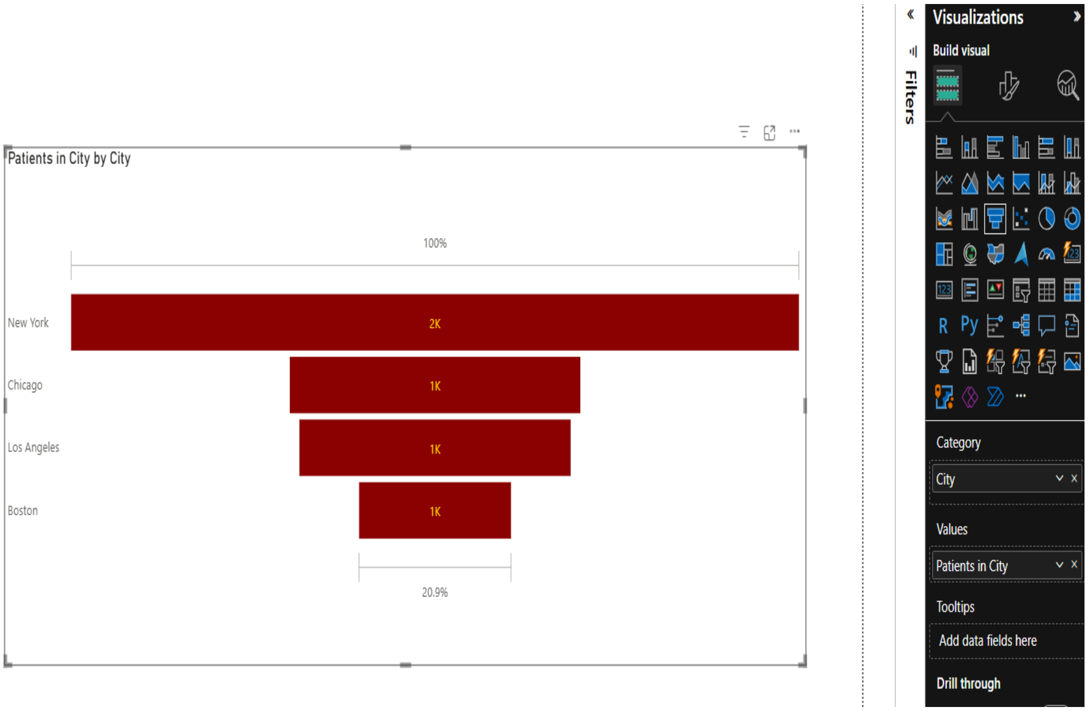


6. Funnel Chart — Patients in City by City

Shows total patient count across cities.

Insights:

- **New York has the highest patient count (~2K).**
- **Chicago, Los Angeles, and Boston are nearly equal (1K each).**
- **Helps prioritize city-wise healthcare planning.**



Category: City

Values: Patients in City

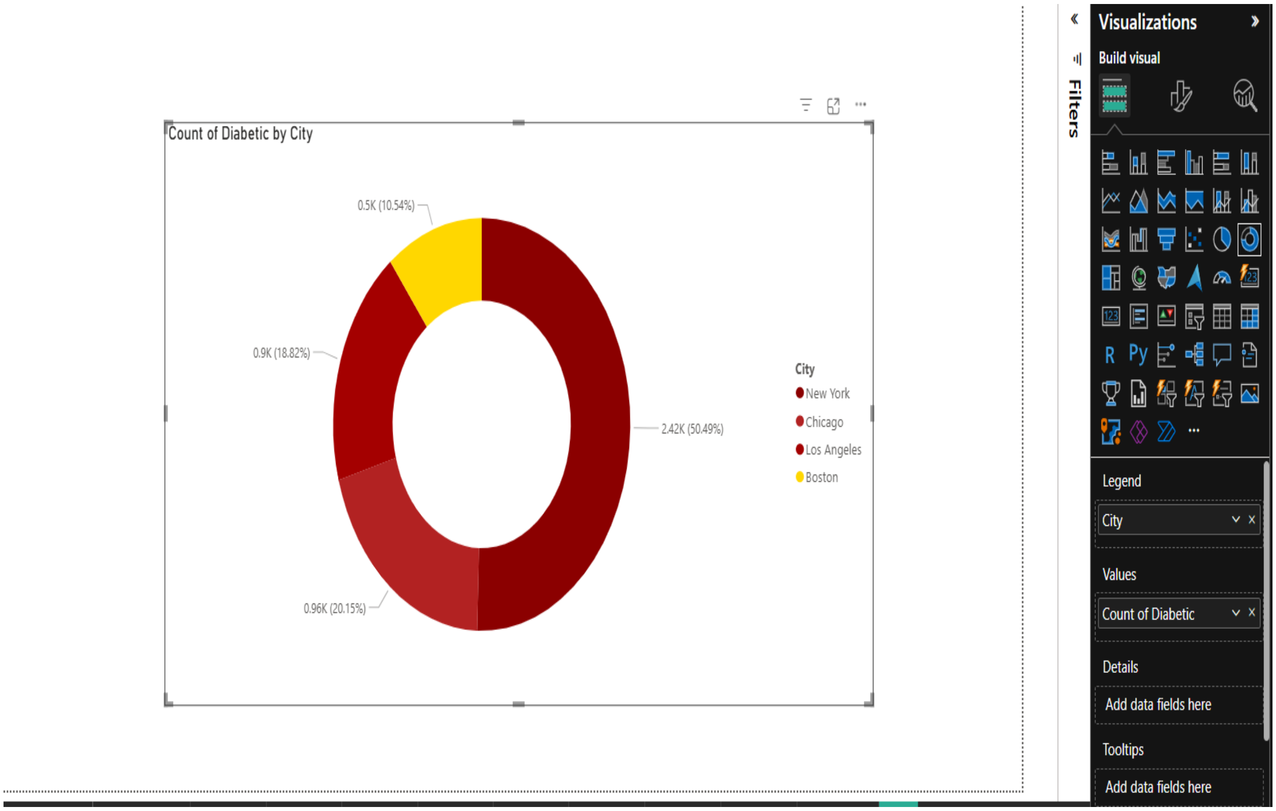
Tooltips: Add data fields here

7. Donut Chart — Count of Diabetic by City

Displays how diabetic patients are spread across cities.

Insights:

- New York again has the highest diabetic population.
- Other cities contribute smaller shares.
- Indicates regional diabetic hotspots.



Legend: City

Values: Count of Diabetic

8. Combined Column + Line Chart — Total Patients vs Avg Has Disease by Diabetic & Smoker

This is a dual-axis visualization.

Bars:

- Show total patients grouped by diabetic condition (Yes/No) and smoker types.

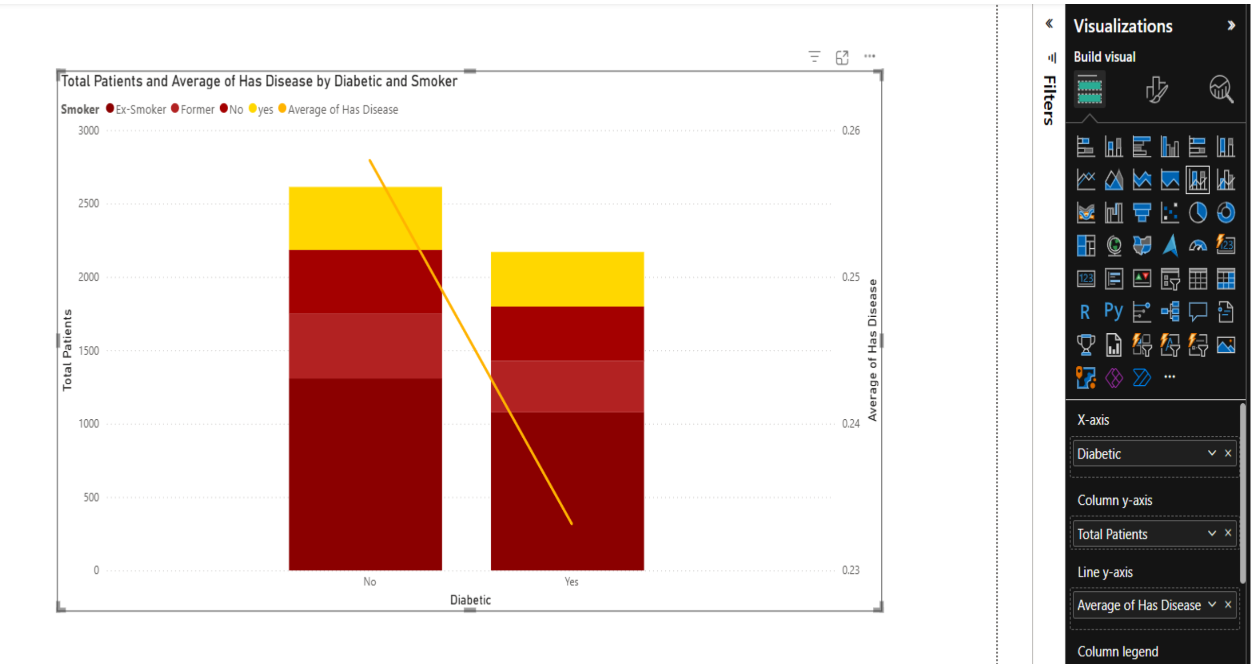
Yellow Line:

- Shows the *average of Has Disease* metric for each group.

Insights:

- Diabetic patients have higher disease rates.
- Smokers in the diabetic group show higher disease risk.
- Non-diabetics have fewer disease cases comparatively.

This is a key highlight of the dashboard → Diabetes + Smoking = Highest Disease Risk.



X-axis

Diabetic

Column y-axis

Total Patients

Line y-axis

Average of Has Disease

Column legend

9. Slicers (Filters Used in Dashboard)

✓ Diabetic (Yes/No)

Filters visuals based on diabetic status.

✓ Gender (Male/Female)

Enable gender-based analysis.

✓ Smoker Type (Yes / No / Former / Ex-smoker)

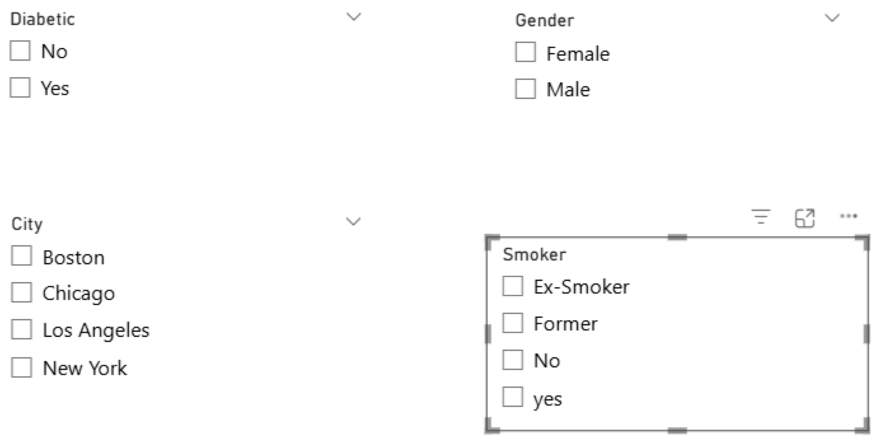
Segment heart rate and disease risk.

✓ City

Allows viewing metrics for specific locations:

- New York
- Chicago
- Los Angeles
- Boston

Slicers make the dashboard fully interactive and dynamic for exploratory analysis.



Diabetic

□ No

□ Yes

Gender

□ Female

□ Male

City

□ Boston

□ Chicago

□ Los Angeles

□ New York

Smoker

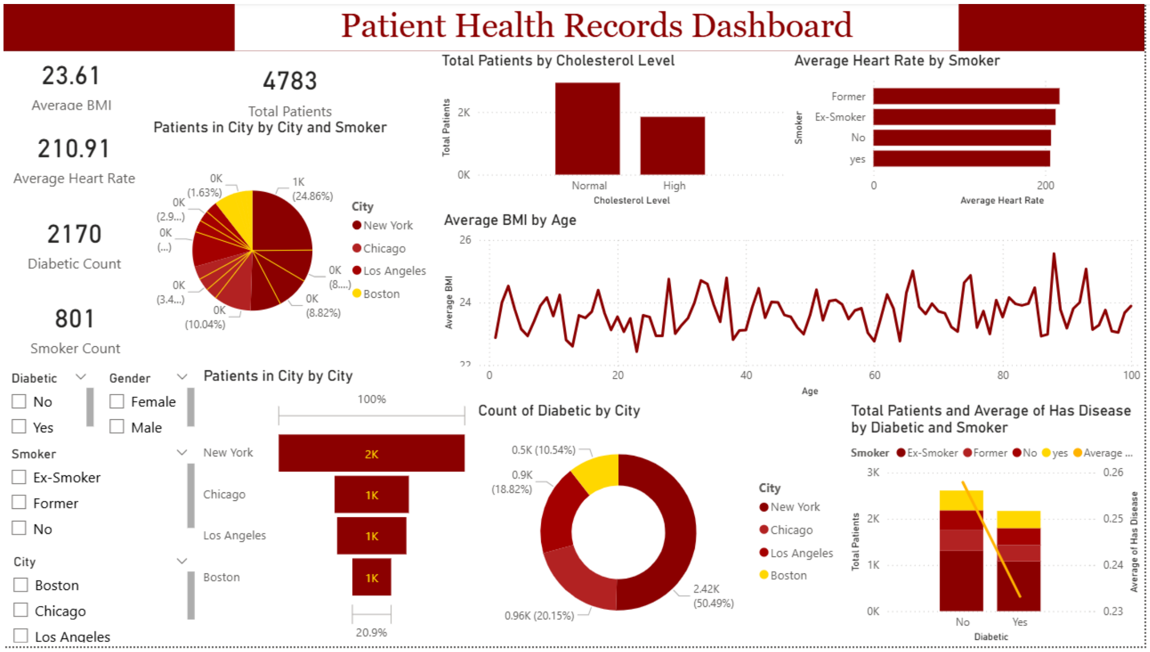
□ Ex-Smoker

□ Former

□ No

□ yes

PATIENTS HEALTH RECORD DASHBOARD



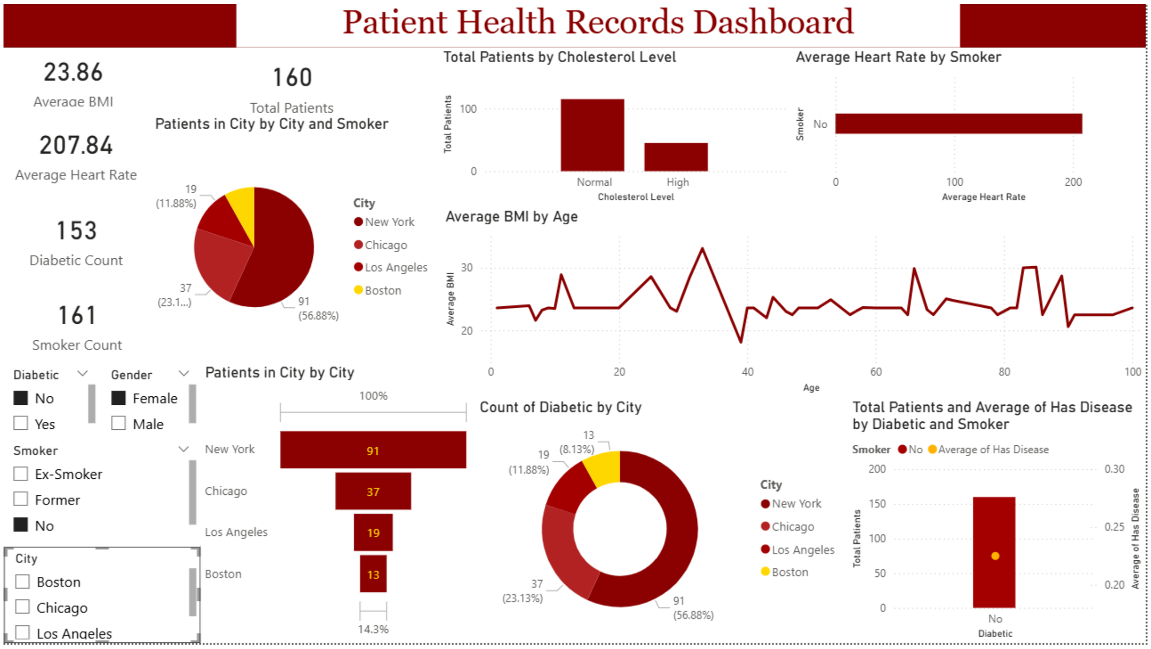
Diabetic
☐ No
☐ Yes

Gender
☐ Female
☐ Male

Smoker
☐ Ex-Smoker
☐ Former
☐ No

City
☐ Boston
☐ Chicago
☐ Los Angeles

DASHBOARD (AFTER APPLYING SLICERS)



Diabetic
■ No
☐ Yes

Gender
■ Female
☐ Male

Smoker
☐ Ex-Smoker
☐ Former
■ No

City
☐ Boston
☐ Chicago
☐ Los Angeles

Conclusion:

The combined dataset and dashboard analysis reveal important health patterns:

- **New York** has the highest patient and diabetic population.
- **Smoking status** has a strong correlation with increased heart rate and disease severity.
- **Diabetic patients** consistently show higher risk indicators.
- **BMI remains relatively stable** across age groups, showing lifestyle influence.
- **Cities with higher populations** show higher disease prevalence.

The dashboard enables healthcare professionals to identify high-risk groups and design targeted health interventions.