

Project Title: Healthcare and Heart Disease Analysis

Student Name: Frank Asto

Course and Section: DS 230 - SQL for Data Analysis

Instructor Name: Kiran Mihir

Submission Term: Fall 2024

Table of Contents

1. Introduction -----	2
2. Methodology -----	2
Data Cleaning and Transformation -----	2
Data Reduction -----	3
Table Preparation in Excel -----	3
Database Design -----	3
3. Data Analysis -----	5
Foundational Analysis -----	5
Key Research Questions -----	7
4. Results and Interpretation -----	13
5. Conclusion -----	13
6. References -----	14

1. Introduction

Heart disease is one of the leading causes of death worldwide. Understanding the factors contributing to heart disease can significantly enhance preventive measures and healthcare strategies. This analysis explores into multiple dimensions, such as age, race, BMI, and lifestyle habits, to identify key patterns and relationships. The goal of this project is to analyze healthcare data to uncover relationships between heart disease and various factors such as age, race, BMI, and lifestyle habits. The dataset for this project was sourced from Kaggle and originally contained approximately 300,000 records and 18 features. The following research questions guided the analysis:

1. **Age and Heart Disease Risk:** What is the relationship between age categories and the prevalence of heart disease? Do older age groups have significantly higher risks?
2. **Race and Heart Disease:** How does heart disease prevalence vary across racial groups? Are certain races disproportionately affected?
3. **BMI and Heart Disease Comparison:** How do BMI levels differ between individuals with heart disease and those without? Is there a significant difference in BMI distributions?
4. **Combined Factors: Race, Age, BMI:** Among individuals aged 40–60, how do BMI levels (with ranges like 19–24, 25–30, >30) compare for individuals with heart disease?

2. Methodology

The dataset was prepared and organized as follows:

1. Data Cleaning and Transformation:

The dataset was cleaned and transformed using Python and the pandas library. For example:

```
import pandas as pd
healthcare_data = pd.read_csv("HealthCare_Heart_Disease_2020.csv")

age_mapping = {
    "18-24": 21,
    "25-29": 27,
    "30-34": 32,
    "35-39": 37,
    "40-44": 42,
    "45-49": 47,
    "50-54": 52,
    "55-59": 57,
    "60-64": 62,
    "65-69": 67,
    "70-74": 72,
    "75-79": 77,
    "80 or older": 85
}

healthcare_data['AgeCategory'] = healthcare_data['AgeCategory'].map(age_mapping)
```

- The original dataset contained text values, which were converted to numeric values for analysis (e.g., Yes/No to 1/0, age categories such as "18-24" to their midpoint values like 21).
- Mappings were applied to columns such as Race, GenHealth, Diabetic, Sex, and other categories for consistency.

2. Data Reduction:

- Due to limitations in Oracle SQL Developer, the dataset was reduced from approximately 300,000 records to 10,000 records per table.

3. Table Preparation in Excel:

- Before importing into Oracle SQL Developer, the data for each table was separated using Microsoft Excel.
- Primary keys were defined, and data entries were assigned to match the structure of the database tables.

4. Database Design:

The dataset was divided into five normalized tables, each designed to capture specific aspects of the data with clear relationships:

1. HeartDisease Table:

- Columns: PatientID (Primary Key), HeartDisease (1 or 0).
- Focus: Tracks the presence or absence of heart disease.

2. Demographics Table:

- Columns: DemographicsID (Primary Key), PatientID (Foreign Key), Sex, AgeCategory, Race.
- Focus: Captures demographic details such as gender, age group, and race.

3. Habits Table:

- Columns: HabitsID (Primary Key), PatientID (Foreign Key), Smoking, AlcoholDrinking, PhysicalActivity.
- Focus: Tracks lifestyle habits.

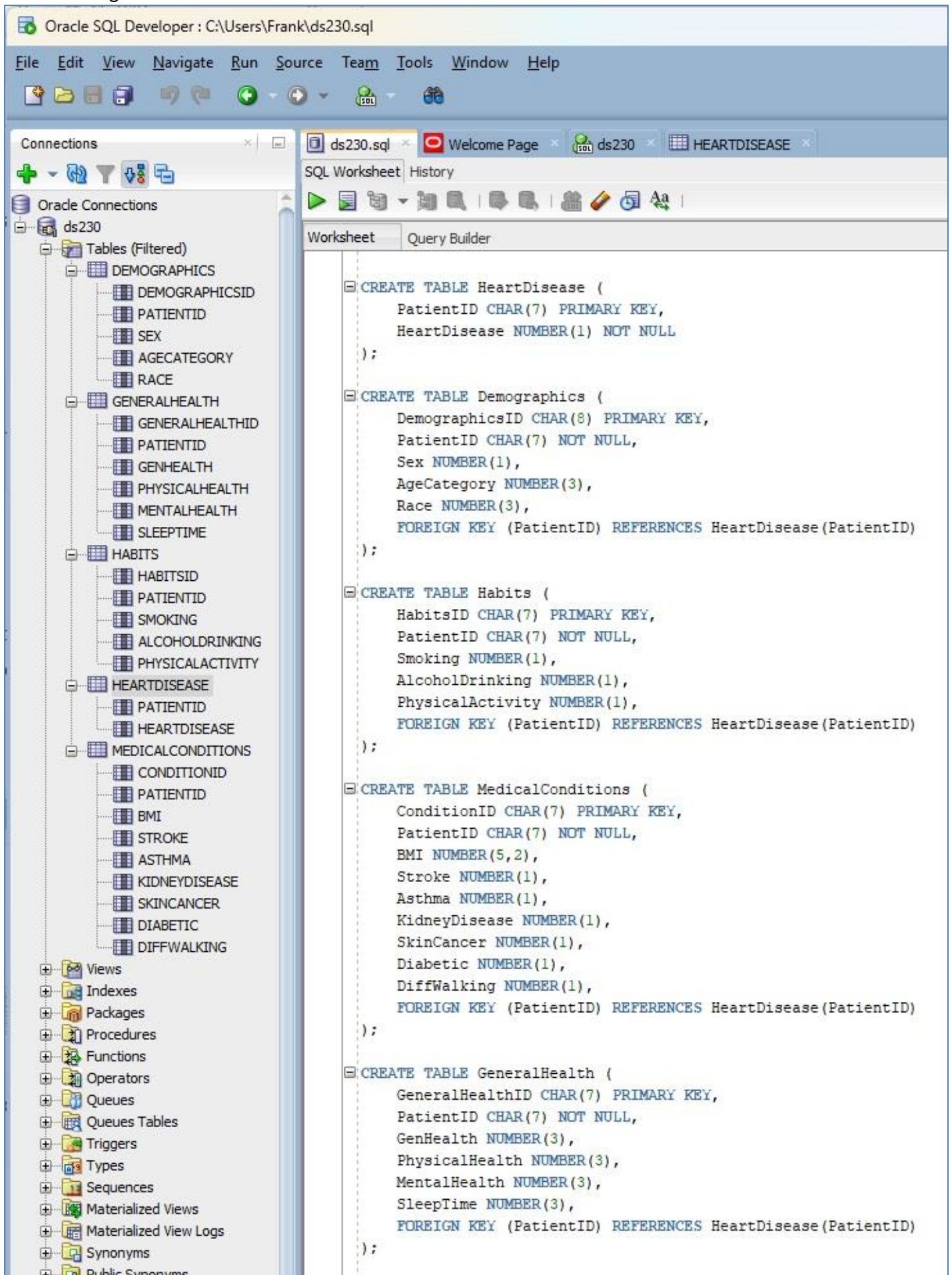
4. MedicalConditions Table:

- Columns: ConditionID (Primary Key), PatientID (Foreign Key), BMI, Stroke, Asthma, KidneyDisease, SkinCancer, Diabetic, DiffWalking.
- Focus: Includes various health conditions and body mass index (BMI).

5. GeneralHealth Table:

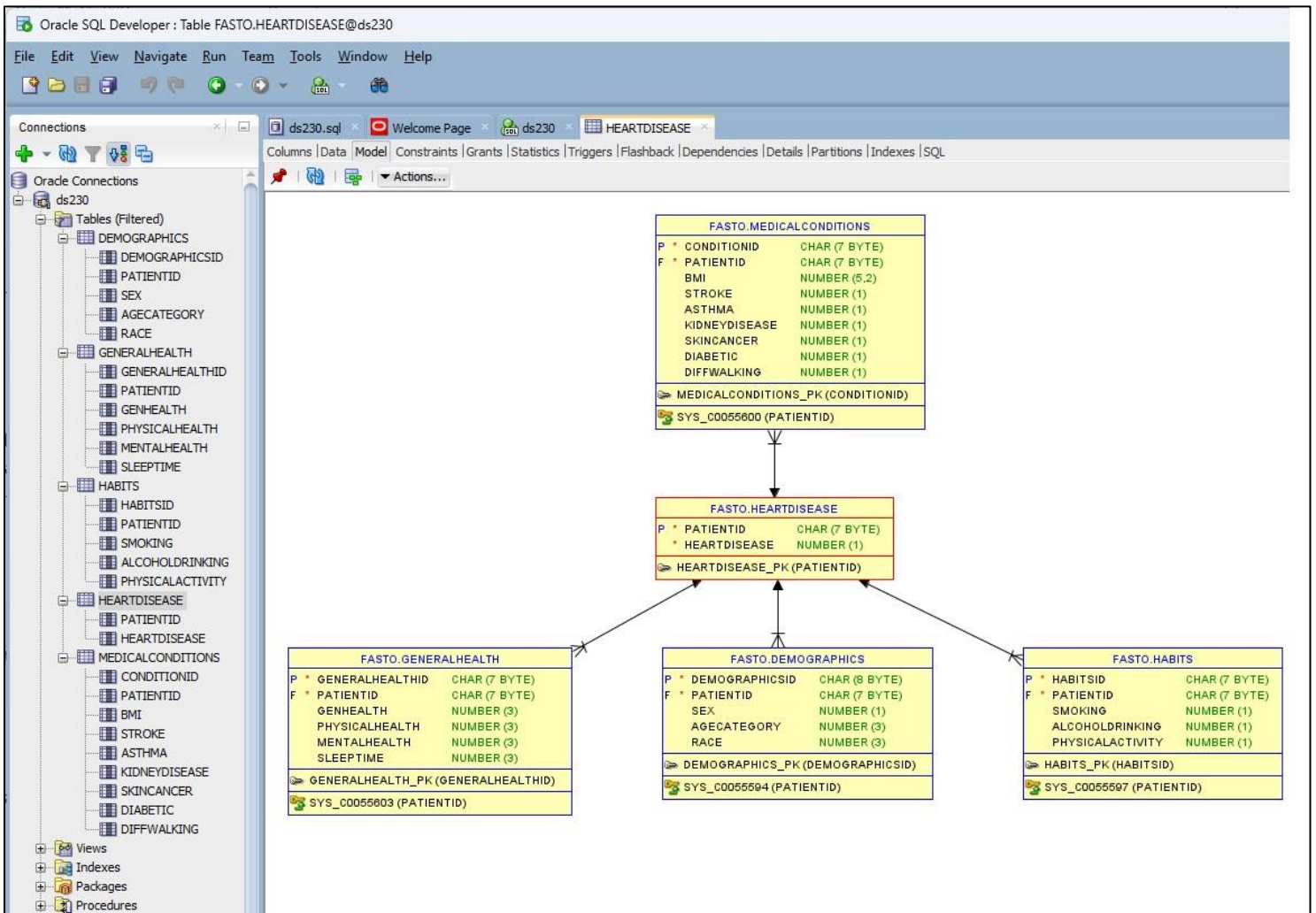
- Columns: GeneralHealthID (Primary Key), PatientID (Foreign Key), GenHealth, PhysicalHealth, MentalHealth, SleepTime.
- Focus: Captures general health assessments and lifestyle details.

The following SQL code was used to create the tables:



```
CREATE TABLE HeartDisease (  
    PatientID CHAR(7) PRIMARY KEY,  
    HeartDisease NUMBER(1) NOT NULL  
);  
  
CREATE TABLE Demographics (  
    DemographicsID CHAR(8) PRIMARY KEY,  
    PatientID CHAR(7) NOT NULL,  
    Sex NUMBER(1),  
    AgeCategory NUMBER(3),  
    Race NUMBER(3),  
    FOREIGN KEY (PatientID) REFERENCES HeartDisease(PatientID)  
);  
  
CREATE TABLE Habits (  
    HabitsID CHAR(7) PRIMARY KEY,  
    PatientID CHAR(7) NOT NULL,  
    Smoking NUMBER(1),  
    AlcoholDrinking NUMBER(1),  
    PhysicalActivity NUMBER(1),  
    FOREIGN KEY (PatientID) REFERENCES HeartDisease(PatientID)  
);  
  
CREATE TABLE MedicalConditions (  
    ConditionID CHAR(7) PRIMARY KEY,  
    PatientID CHAR(7) NOT NULL,  
    BMI NUMBER(5,2),  
    Stroke NUMBER(1),  
    Asthma NUMBER(1),  
    KidneyDisease NUMBER(1),  
    SkinCancer NUMBER(1),  
    Diabetic NUMBER(1),  
    DiffWalking NUMBER(1),  
    FOREIGN KEY (PatientID) REFERENCES HeartDisease(PatientID)  
);  
  
CREATE TABLE GeneralHealth (  
    GeneralHealthID CHAR(7) PRIMARY KEY,  
    PatientID CHAR(7) NOT NULL,  
    GenHealth NUMBER(3),  
    PhysicalHealth NUMBER(3),  
    MentalHealth NUMBER(3),  
    SleepTime NUMBER(3),  
    FOREIGN KEY (PatientID) REFERENCES HeartDisease(PatientID)  
);
```

After creating my tables using the SQL code above, I imported the data from Excel into the corresponding tables in Oracle SQL Developer. Initially, I faced challenges importing the data due to the large number of records, so I reduced the rows to 10,000 for each table. This allowed the import to succeed, ensuring that the tables were populated with the correct primary keys and data. Following this, I generated the following schema to visualize the relationships among the tables:



In summary, the database design effectively organizes the dataset into distinct tables with clear relationships, ensuring data integrity and supporting efficient querying and analysis.

3. Data Analysis

Foundational Analysis

The data analysis began by checking the number of records in each table and verifying the columns to ensure data integrity. This included counting the total cases of heart disease (1 and 0) and performing a quick analysis on gender distribution in the Demographics table. Below are the SQL queries used for these initial steps:

```

SELECT COUNT(*) FROM heartdisease;
SELECT COUNT(*) FROM demographics;
SELECT COUNT(*) FROM generalhealth;
SELECT COUNT(*) FROM habits;
SELECT COUNT(*) FROM medicalconditions;

```

	COUNT(*)
1	10000


```

SELECT * FROM heartdisease
FETCH FIRST 10 ROWS ONLY;
SELECT * FROM demographics
FETCH FIRST 10 ROWS ONLY;
SELECT * FROM generalhealth
FETCH FIRST 10 ROWS ONLY;
SELECT * FROM habits
FETCH FIRST 10 ROWS ONLY;
SELECT * FROM medicalconditions
FETCH FIRST 10 ROWS ONLY;

```

	CONDITIONID	PATIENTID	BMI	STROKE	ASTHMA	KIDNEYDISEASE	SKINCANCER	DIABETIC	DIFFWALKING
1	C000572	P000572	30.27	0	0	0	1	0	0
2	C000573	P000573	33.28	0	0	0	0	1	1
3	C000574	P000574	27.41	0	0	0	0	0	0
4	C000575	P000575	27.32	0	0	0	0	0	0
5	C000576	P000576	35.31	0	1	0	0	1	1
6	C000577	P000577	31.75	0	0	0	0	0	0
7	C000578	P000578	23.49	0	0	0	0	1	1
8	C000579	P000579	28.19	0	0	0	0	0	0
9	C000580	P000580	24.13	1	0	0	0	0	1
10	C000581	P000581	24.89	0	0	0	0	0	0

```
--Count of HeartDisease 1 and 0
```

```

SELECT
  HeartDisease,
  COUNT(*) AS Count
FROM HeartDisease
GROUP BY HeartDisease
ORDER BY HeartDisease;

```

HEARTDISEASE	COUNT
1	9014
2	986

The dataset is significantly imbalanced, with 90.14% of cases (9014 records) showing no heart disease and only 9.86% (986 records) showing the presence of heart disease. This indicates a strong skew toward individuals without heart disease, which could impact analysis and require adjustments to account for the imbalance.

```
--Count of male and female with HeartDisease
```

```

SELECT
  d.Sex,
  COUNT(*) AS Count
FROM Demographics d
  JOIN HeartDisease h ON d.PatientID = h.PatientID
WHERE h.HeartDisease = 1
GROUP BY d.Sex
ORDER BY d.Sex;

```

SEX	COUNT
1	424
2	562

The data for Sex reveals that there are 424 females (0) and 562 males (1) with heart disease. This indicates a slightly higher prevalence of heart disease among males compared to females, which could suggest gender-specific factors influencing heart disease risk.

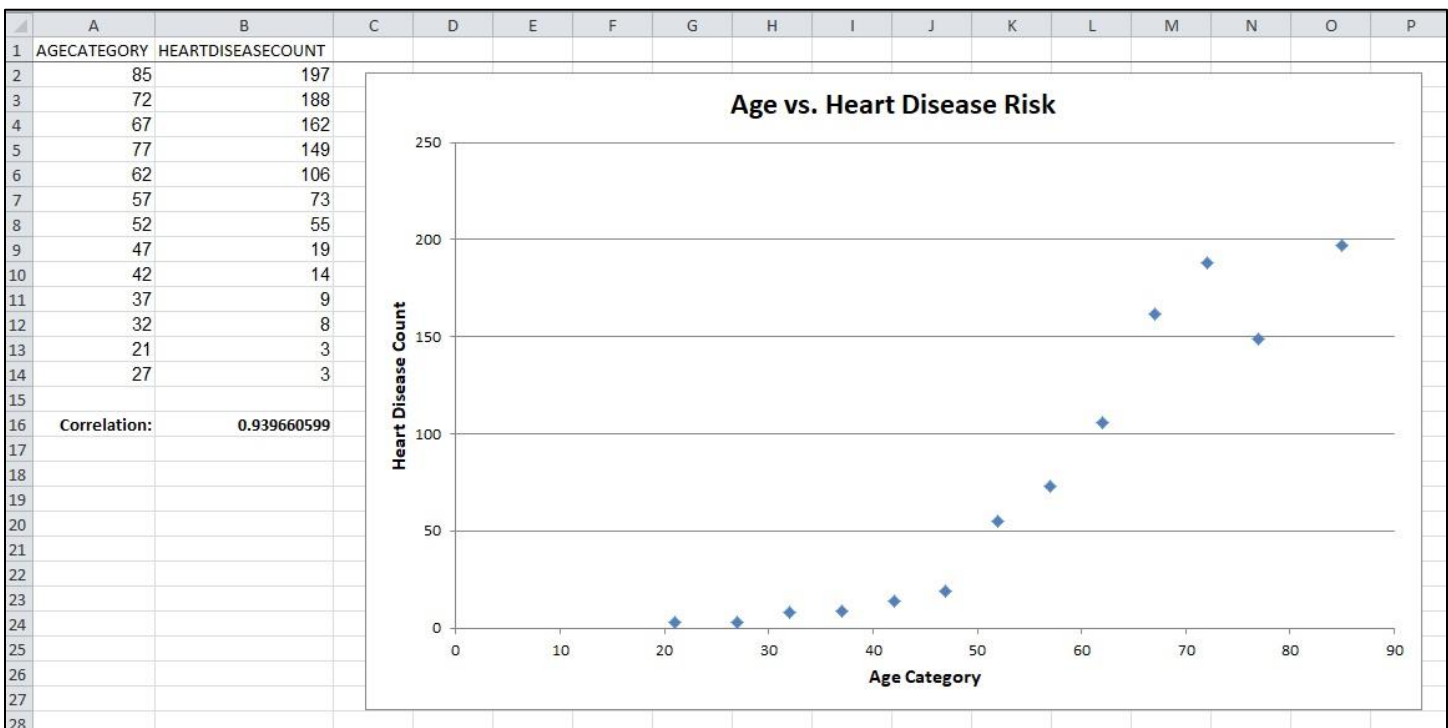
Key Research Questions

After these foundational checks, the data analysis was structured around the four key research questions:

1. Age and Heart Disease Risk:

AGECATEGORY	HEARTDISEASECOUNT	
1	85	197
2	72	188
3	67	162
4	77	149
5	62	106
6	57	73
7	52	55
8	47	19
9	42	14
10	37	9
11	32	8
12	21	3
13	27	3

- Counted the number of heart disease cases in each age category by joining the HeartDisease table with the Demographics table using the PatientID primary and foreign key relationship.
- Analysis shows that heart disease prevalence increases with age. The highest counts are observed in the older age groups, with 197 cases at age 85, 188 cases at age 72, and 162 cases at age 67. Younger age groups, such as 21 and 27, have significantly fewer cases (only 3 each).
- This indicates a strong positive correlation between age and heart disease risk, highlighting age as a significant factor for heart disease.
- The following visualization provides a clear depiction of the relationship between age and heart disease prevalence, showcasing the positive correlation between these variables.



2. Race and Heart Disease:

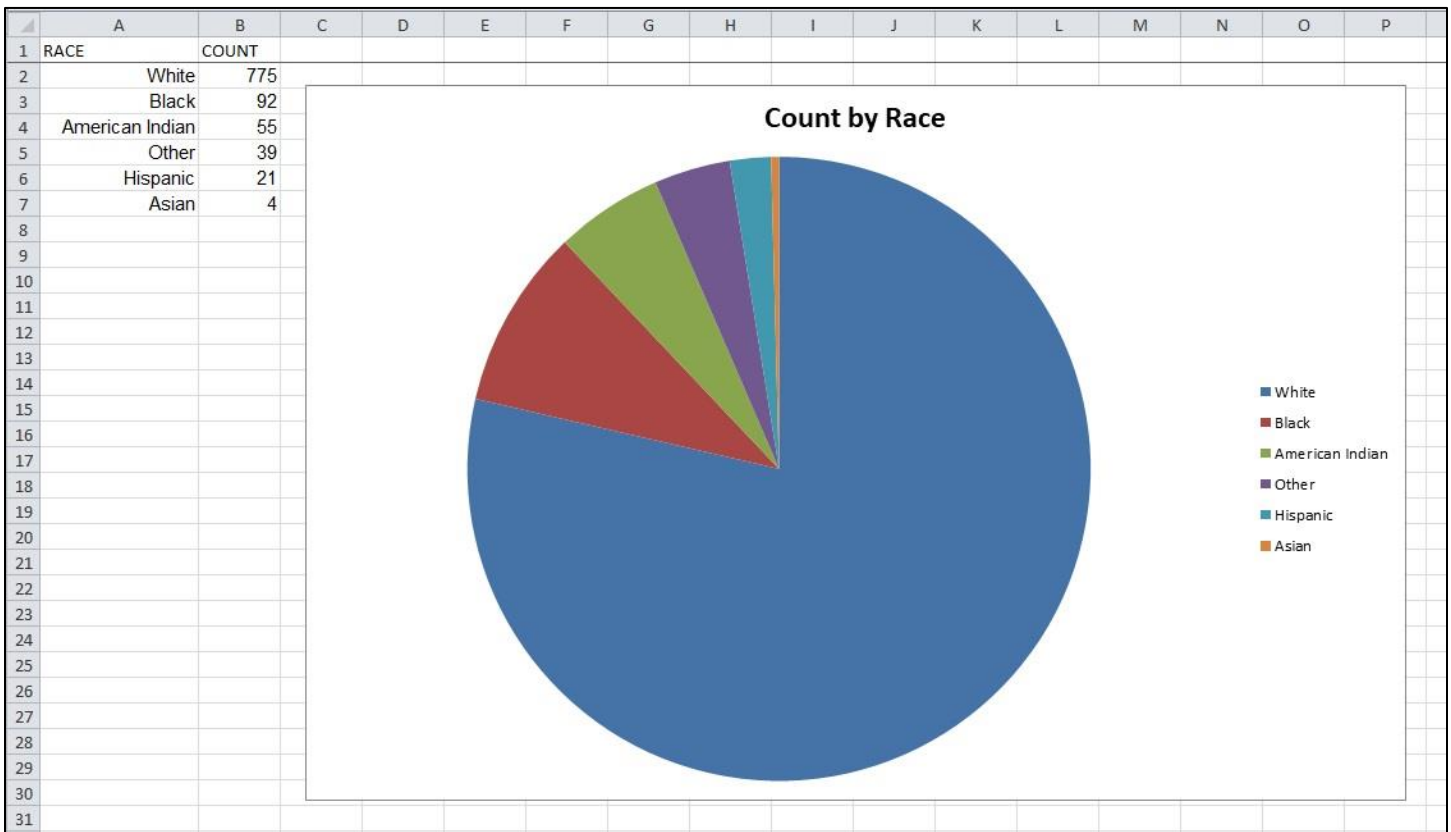
```

--Count of Race with HeartDisease
SELECT
  d.Race,
  COUNT(*) AS Count
FROM Demographics d
  JOIN HeartDisease h ON d.PatientID = h.PatientID
WHERE h.HeartDisease = 1
GROUP BY d.Race
ORDER BY Count DESC;

```

	RACE	COUNT
1	0	775
2	1	92
3	4	55
4	5	39
5	2	21
6	3	4

- Counted the number of heart disease cases by race by joining the HeartDisease table with the Demographics table using the PatientID primary and foreign key relationship.
- Analysis reveals that White individuals (Race = 0) have the highest prevalence of heart disease with 775 cases, followed by Black individuals (Race = 1) with 92 cases. Other groups, such as American Indian (Race = 4) with 55 cases, Other (Race = 5) with 39 cases, Hispanic (Race = 2) with 21 cases, and Asian (Race = 3) with 4 cases, show significantly lower counts.
- The reduced dataset of 10,000 records might influence these results, as the original dataset had approximately 300,000 records. Despite this limitation, the data suggests substantial racial disparities in heart disease prevalence.
- The following pie chart visualizes the distribution of heart disease cases by race, clearly illustrating the disparities among different racial groups.



3. BMI and Heart Disease Comparison:

```

--Average of BMI
SELECT AVG(m.BMI) AS Average_BMI
FROM MedicalConditions m
JOIN HeartDisease h ON m.PatientID = h.PatientID
WHERE h.HeartDisease = 1;

--Frequency Count of MBI with HeartDisease and showing t
SELECT m.BMI, COUNT(m.BMI) AS Frequency
FROM MedicalConditions m
JOIN HeartDisease h ON m.PatientID = h.PatientID
WHERE h.HeartDisease = 1
GROUP BY m.BMI
ORDER BY Frequency DESC
FETCH FIRST 10 ROWS ONLY;

--Getting all BMI where HeartDisease = 1 to do a boxplot
SELECT m.BMI
FROM MedicalConditions m
JOIN HeartDisease h ON m.PatientID = h.PatientID
WHERE h.HeartDisease = 1;

SELECT m.BMI
FROM MedicalConditions m
JOIN HeartDisease h ON m.PatientID = h.PatientID
WHERE h.HeartDisease = 0;

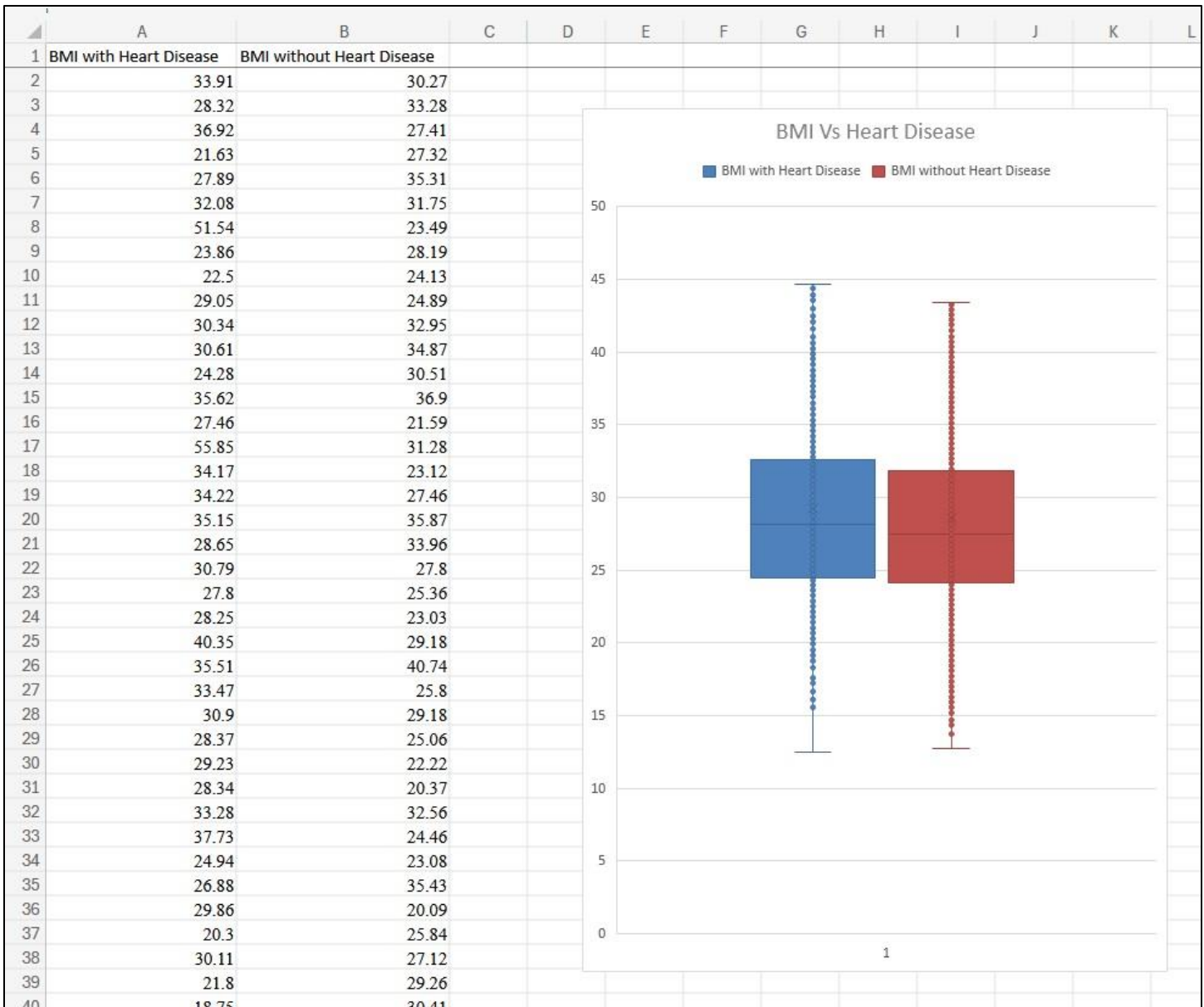
```

AVERAGE_BMI	
1	29.27430020283975659229208924949290060852

BMI	FREQUENCY
1	27.12
2	29.29
3	24.41
4	29.53
5	24.33
6	23.49
7	26.58
8	28.7
9	28.25
10	27.46

BMI	
1	33.91
2	28.32
3	36.92
4	21.63
5	27.89
6	32.08
7	51.54
8	23.86
9	22.5
10	29.05
11	30.34
12	30.61
13	24.28
14	35.62
15	27.46
16	55.85
17	34.17
18	34.22
19	35.15
20	28.65
21	30.79
22	27.8
23	28.25
24	40.35
25	35.51

- Conducted basic statistical analysis, including calculating the average BMI for individuals with heart disease.
- Identified the most frequent BMI values for heart disease cases.
- Retrieved BMI data for individuals with heart disease (HeartDisease = 1) and those without (HeartDisease = 0) to create a boxplot comparison in Excel.
- Visualized using a boxplot to compare the distributions of BMI between the two groups. How do BMI levels differ between individuals with heart disease and those without? Is there a significant difference in BMI distributions?



The analysis of the **BMI vs. Heart Disease** boxplot reveals:

- Individuals with heart disease (HeartDisease = 1) tend to have a slightly higher BMI distribution compared to those without heart disease (HeartDisease = 0).
- The median BMI for those with heart disease is higher, as shown by the central line of the boxplot.
- The range of BMI values for both groups overlaps significantly, but the upper quartile for individuals with heart disease extends further, indicating a higher prevalence of obesity in this group.

This analysis highlights a noticeable trend that individuals with heart disease often fall into higher BMI categories, suggesting a strong link between BMI and heart disease risk.

4. Combined Factors: Race, Age, BMI:

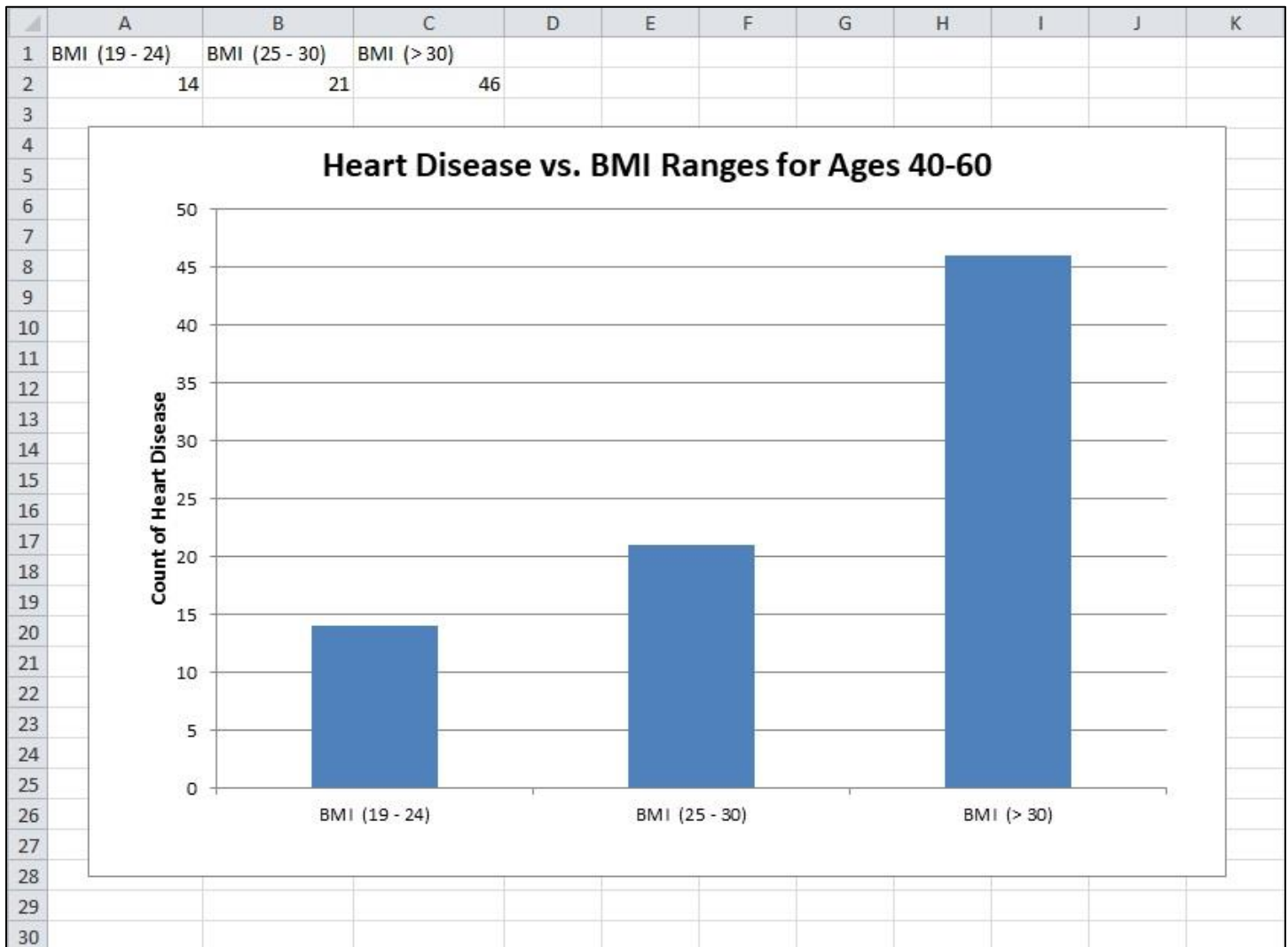
```
--Analyzing race, age, and BMI with HeartDisease
SELECT
    d.Race,
    d.AgeCategory,
    mc.BMI,
    COUNT(*) AS PatientCount
FROM Demographics d
JOIN MedicalConditions mc ON d.PatientID = mc.PatientID
JOIN HeartDisease h ON d.PatientID = h.PatientID
WHERE d.Race = 0 -- White
    AND d.AgeCategory BETWEEN 40 AND 60
    AND mc.BMI > 30
    AND h.HeartDisease = 1
GROUP BY d.Race, d.AgeCategory, mc.BMI
ORDER BY PatientCount DESC;

--Analyzing race, age, and BMI with HeartDisease
SELECT
    d.Race,
    d.AgeCategory,
    mc.BMI,
    COUNT(*) AS PatientCount
FROM Demographics d
JOIN MedicalConditions mc ON d.PatientID = mc.PatientID
JOIN HeartDisease h ON d.PatientID = h.PatientID
WHERE d.Race = 0 -- White
    AND d.AgeCategory BETWEEN 40 AND 60
    AND mc.BMI BETWEEN 25 AND 30
    AND h.HeartDisease = 1
GROUP BY d.Race, d.AgeCategory, mc.BMI
ORDER BY PatientCount DESC;

--Analyzing race, age, and BMI with HeartDisease
SELECT
    d.Race,
    d.AgeCategory,
    mc.BMI,
    COUNT(*) AS PatientCount
FROM Demographics d
JOIN MedicalConditions mc ON d.PatientID = mc.PatientID
JOIN HeartDisease h ON d.PatientID = h.PatientID
WHERE d.Race = 0 -- White
    AND d.AgeCategory BETWEEN 40 AND 60
    AND mc.BMI BETWEEN 19 AND 24
    AND h.HeartDisease = 1
GROUP BY d.Race, d.AgeCategory, mc.BMI
ORDER BY PatientCount DESC;
```

- Combined three tables: Demographics, MedicalConditions, and HeartDisease using the PatientID as a primary and foreign key.
- Focused specifically on individuals from the White racial group (Race = 0) for this analysis.
- Selected individuals aged between 40 and 60 for all SQL queries to focus on this key age range.

- Modified the BMI ranges for each SQL query to analyze heart disease prevalence in the following categories:
 - BMI > 30:** Representing obesity (Class I and higher).
 - BMI between 25 and 30:** Representing the overweight category.
 - BMI between 19 and 24:** Representing the normal BMI range.
- Counted the number of individuals with heart disease for each combination of race, age, and BMI range.
- Visualized the results using a bar chart for clarity. Among individuals aged 40–60, how do BMI levels (with ranges like 19–24, 25–30, >30)



The bar chart, **Heart Disease vs. BMI Ranges for Ages 40–60**, shows the following insights:

- BMI > 30:** Individuals in this range (representing obesity) account for the highest number of heart disease cases (**46 cases**), indicating a significant risk factor.
- BMI 25–30:** This range (representing overweight individuals) has fewer cases (**21 cases**), but it is still a notable contributor.
- BMI 19–24:** This range (representing a normal BMI) has the lowest number of cases (**14 cases**), suggesting that individuals within a healthy BMI range are less likely to have heart disease.

This visualization emphasizes the strong association between higher BMI and heart disease, especially in the 40–60 age group. Obesity is a critical factor and stands out as the predominant contributor in this analysis.

4. Results and Interpretation

1. Age and Heart Disease Risk:

- The analysis revealed a strong positive correlation (0.94) between age and heart disease prevalence. Older age groups consistently showed higher risks of heart disease, with the highest counts observed in individuals aged 85 (197 cases), 72 (188 cases), and 67 (162 cases). Younger age groups, such as 21 and 27, had significantly lower prevalence (only 3 cases each). This highlights age as a significant risk factor for heart disease. The scatterplot visualization underscores this trend clearly.

2. Race and Heart Disease:

- The majority of heart disease cases occurred in White individuals (775 cases), followed by Black individuals (92 cases). Other racial groups, such as American Indian (55 cases), Other (39 cases), Hispanic (21 cases), and Asian (4 cases), showed significantly fewer cases. The dataset's reduction to 10,000 records might have influenced these findings, but the pie chart effectively illustrates the disparities in heart disease prevalence across racial groups.

3. BMI and Heart Disease Comparison:

- Individuals with heart disease had higher BMI levels compared to those without, as shown in the boxplot. The average BMI for individuals with heart disease was 29.27, while those without had a slightly lower average BMI. The comparison highlights that individuals with higher BMI levels are at increased risk of heart disease. This statistically significant difference demonstrates the strong link between obesity and heart disease.

4. Combined Factors: Race, Age, BMI:

- Among individuals aged 40–60 and from the White racial group, heart disease prevalence was highest for those with a BMI >30 (46 cases), followed by BMI ranges of 25–30 (21 cases) and 19–24 (14 cases). The bar chart illustrates that obesity is a predominant factor in heart disease risk within this demographic. By combining multiple factors, this analysis provides deeper insights into the intersection of race, age, and BMI in relation to heart disease prevalence.
- A positive correlation (0.94) was found between age categories and heart disease prevalence.
- Older age groups showed a significantly higher risk, as demonstrated in the scatterplot.

5. Conclusion

This project demonstrated significant relationships between heart disease and factors such as age, race, and BMI. Key findings include:

- Older age groups have a markedly higher prevalence of heart disease.
- White individuals account for the majority of heart disease cases in the dataset.
- Higher BMI levels are strongly associated with heart disease.

These insights provide a foundation for further research and potential healthcare interventions targeting high-risk groups.

6. References

- Kaggle Dataset: Healthcare and Heart Disease Data
(<https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease>)