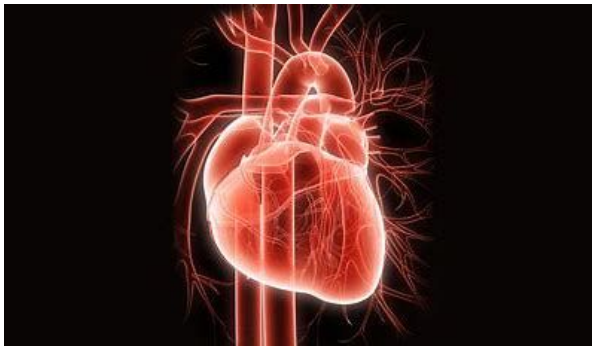
Cardiovascular Analysis

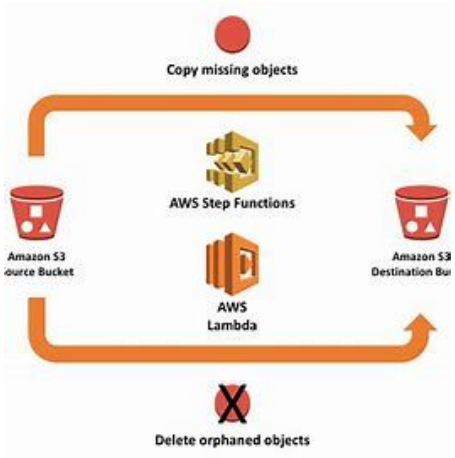
Chi - Scrum Master / Developer
Gabby - Developer
Alejandro - Developer



Introduction

We aim to establish a streamlined data pipeline in a cloud environment

- Integrating Amazon S3 buckets
- Lambda functions
- RDS
- Potentially EC2



This is crucial for organizations aiming to derive actionable insights from their data assets effectively, enabling informed decisions.

Objectives & Issues

Use dataset (Cardiovascular Disease) to run analysis on our local machine:

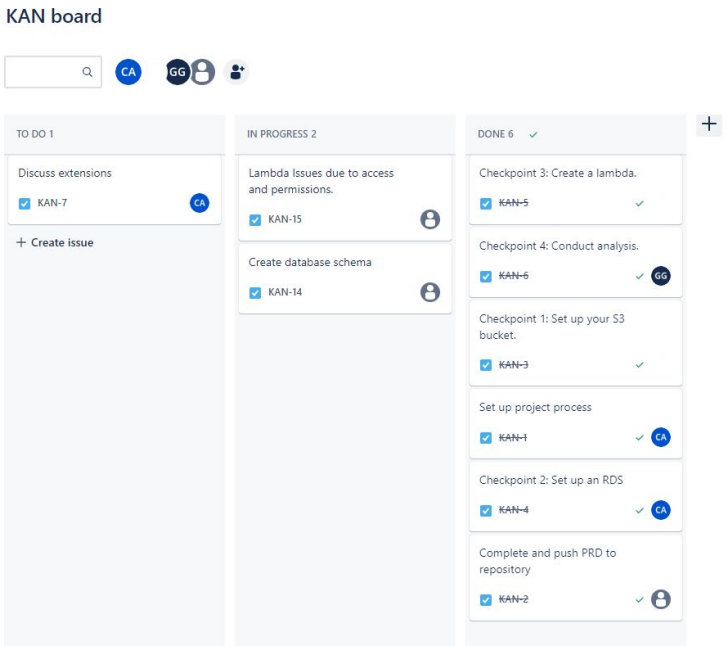
- Creating an S3 bucket
- Creating an RDS
- Creating a triggering Lambda function
- Pull data from the cloud environment and run analysis

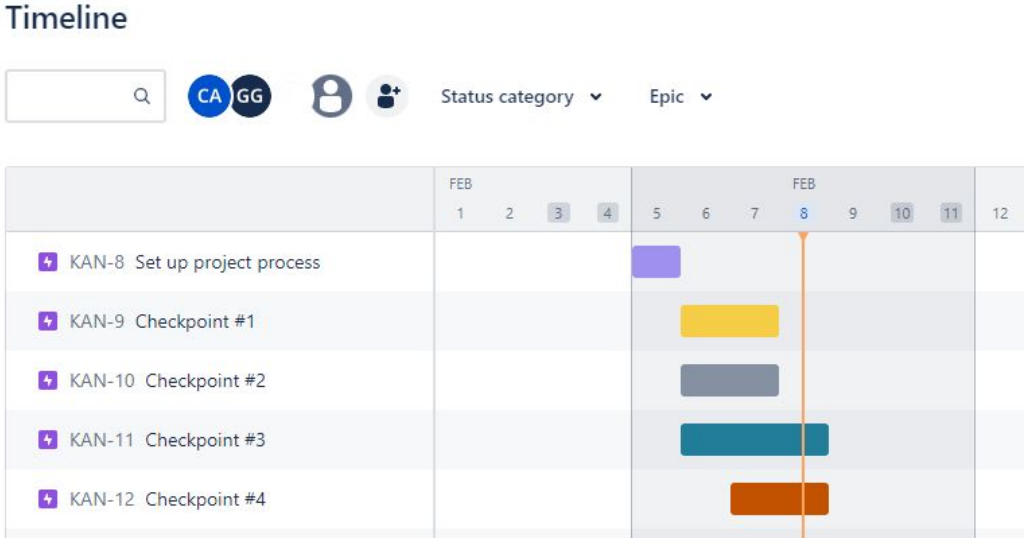
Issues: Although we were able to successfully build this infrastructure, we ran into some issues

- Incomplete data (only about a ⅓ of our data was displayed
- Permissions (Access denied)

Project Tracking

- We used Jira to track our progress

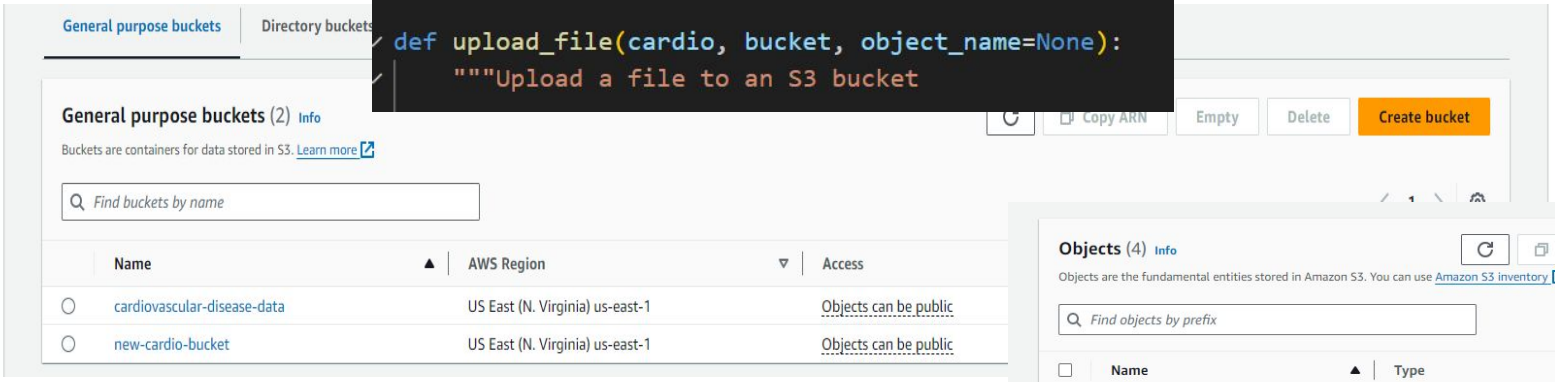


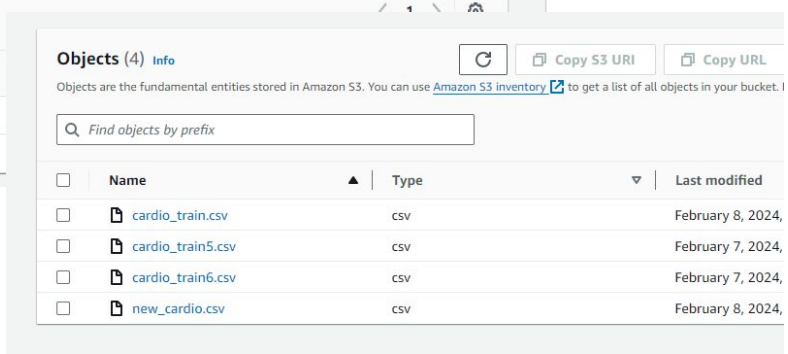


Checkpoint 1: Create S3 bucket

- After specifying the bucket name and additional configuration, we were able to successfully create a bucket.
- IAM users were created and given permission to access the bucket.

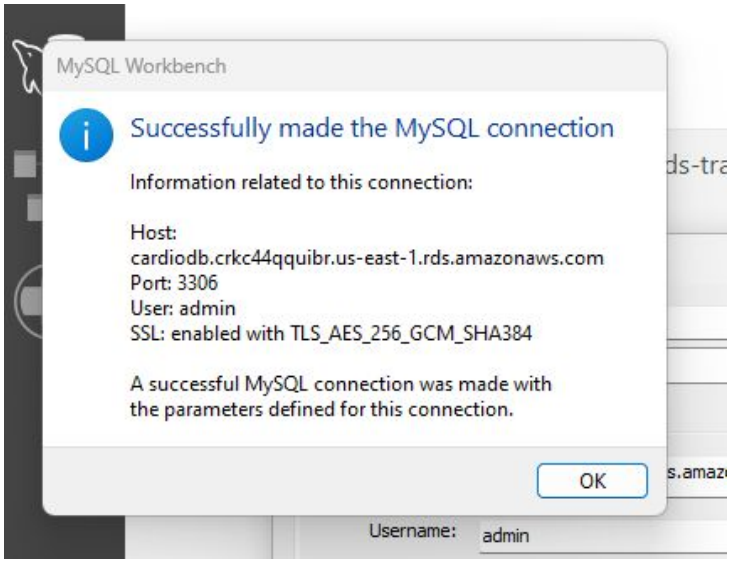
```
cardio = 'cardio_train.csv'
bucket = 'cardiovascular-disease-data'

def upload_file(cardio, bucket, object_name=None):
    """Upload a file to an S3 bucket
```

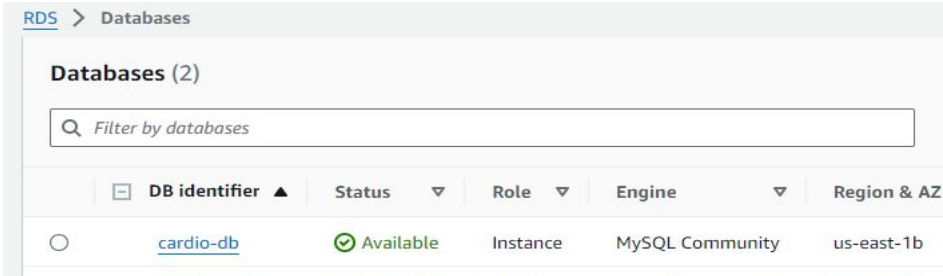




Checkpoint 2: Create RDS

Used AWS console to create and configured the RDS.



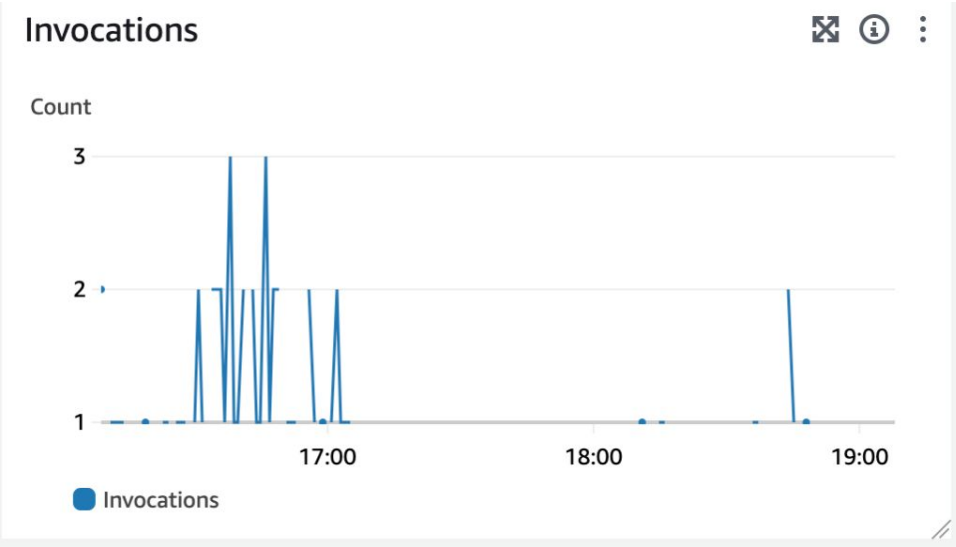


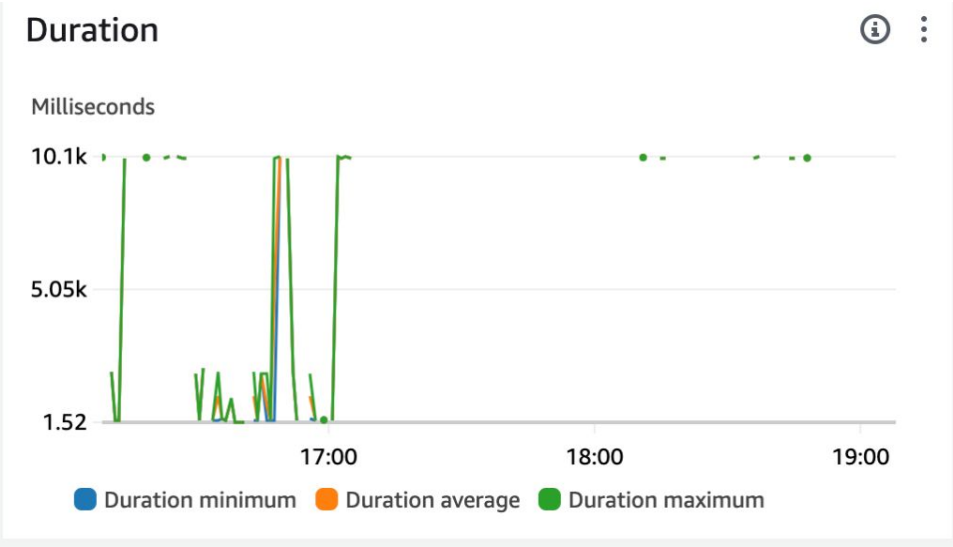

```
# RDS configuration
rds_endpoint = 'cardio-db.crkc44qquibr.us-east-1.rds.amazonaws.com'
rds_port = 3306
rds_username = 'admin'
rds_password = 'rootroot'
```

```
def connect_to_rds():
    try:
        connection = mysql.connector.connect(
            host=rds_endpoint,
            port=rds_port,
            user=rds_username,
            password=rds_password,
        )
        print("Connected to RDS successfully.")
        return connection

    except Exception as e:
        print("An error occurred while connecting to RDS:", e)
        return None
```

Checkpoint 3: Create a Lambda function

- This function will act as a triggering event
  - Listening for any activity in the s3 bucket
  - Cleaning the data and registering it to the rds database.
- Creating the Lambda from the rds.
  - Creating a direct connection to the rds proxy





Issues

Lambda Issues:

- Permissions
    - Getting the correct access for the correct actions
- Timeouts
    - Understanding where and why the data was not being processed successfully
- Troubleshooting
    - Using standard debugging technique and pair programing to pinpoint a solution

```
2024-02-08T13:44:46.600-05:00	START RequestId: 80923608-6bf8-4215-9142-34953122feaa Version: $LATEST
2024-02-08T13:44:48.191-05:00	succesfully pulled data from bucket
2024-02-08T13:44:48.191-05:00	bucket: cardiovascular-disease-data
2024-02-08T13:44:48.191-05:00	key: cardio_train.csv
2024-02-08T13:44:56.652-05:00	2024-02-08T18:44:56.652Z 80923608-6bf8-4215-9142-34953122feaa Task timed out after 10.05 seconds

2024-02-08T18:44:56.652Z 80923608-6bf8-4215-9142-34953122feaa Task timed out after 10.05 seconds
```

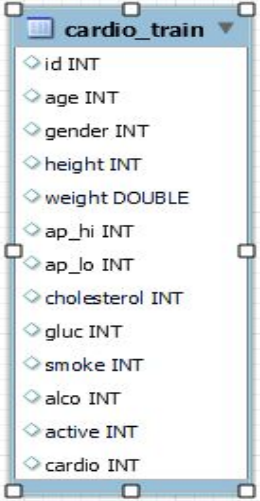
Database

- All the information needed (ex. patient information and details) - was stored under the same table.

```
1 •  use cardio_train;
2
3 •  select * from cardio_train order by id asc;
4
5
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
0	18393	2	168	62	110	80	1	1	0	0	1	0
1	20228	1	156	85	140	90	3	1	0	0	1	1
2	18857	1	165	64	130	70	3	1	0	0	0	1
3	17623	2	169	82	150	100	1	1	0	0	1	1
4	17474	1	156	56	100	60	1	1	0	0	0	0
8	21914	1	151	67	120	80	2	2	0	0	0	0
9	22113	1	157	93	130	80	3	1	0	0	1	0
12	22584	2	178	95	130	90	3	3	0	0	1	1
13	17668	1	158	71	110	70	1	1	0	0	1	0



Checkpoint 4: Data Analysis

Checkpoint 4- Pull Data from RDS to create meaningful insight

To uncover patterns, correlations and potential risk factors related to cardiovascular disease we used:

- Pandas
- Maplotlib.pyplot
- Sqlalchemy to create engine

```
from sqlalchemy import create_engine
from config import rds_port, rds_username, rds_password, rds_endpoint

hostname = rds_endpoint
port = rds_port
username =rds_username
password = rds_password
database_name = 'cardio_train'

engine = create_engine(f'mysql+mysqlconnector://{username}:{password}@{hostname}:{port}/{database_name}')
connection = engine.connect()
```

✓ 0.7s

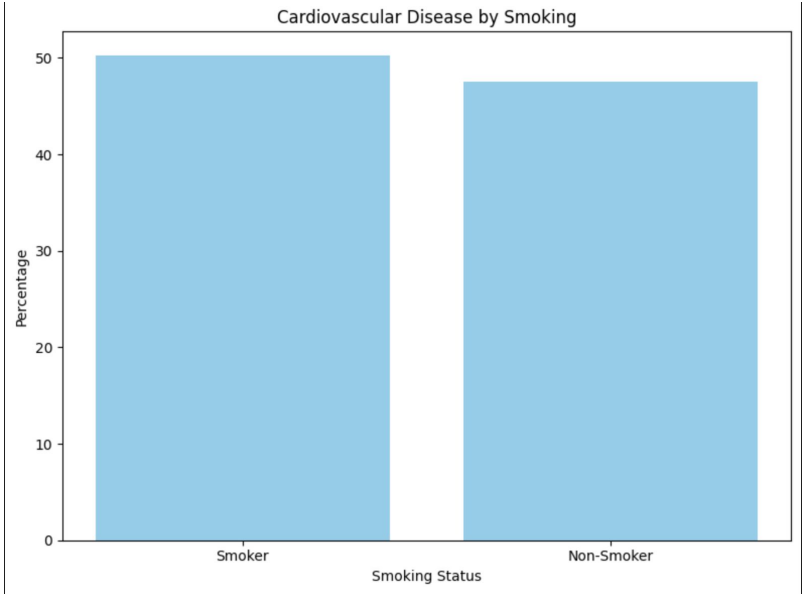
```
import pandas as pd
import matplotlib.pyplot as plt
from sqlalchemy import create_engine

cardio_csv = 'cardio_train2.csv'

cardio_df = pd.read_csv(cardio_csv)
```

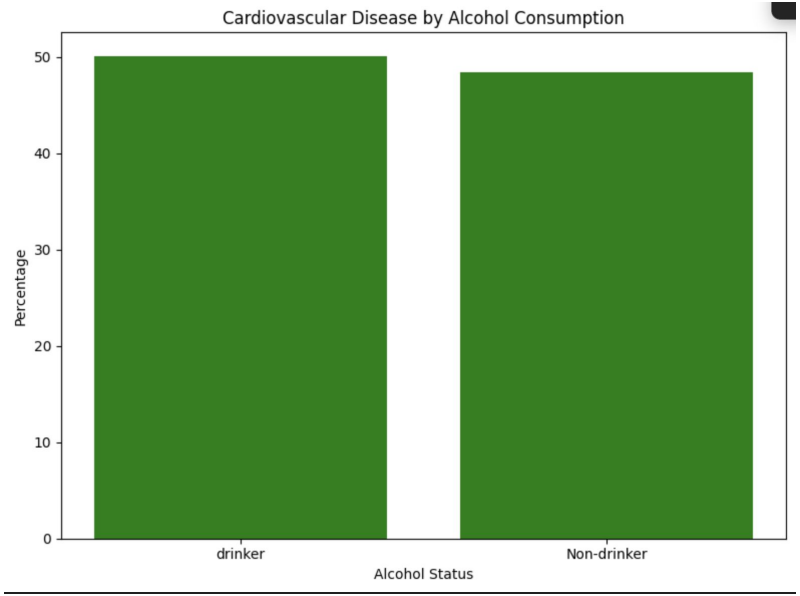
✓ 0.0s

```
engine = create_engine('sqlite:///cardio.db', echo=False)
cardio_df.to_sql('cardio_table', engine, if_exists='replace', index=False)

print("Database created successfully.")
```

Impact to Lifestyle

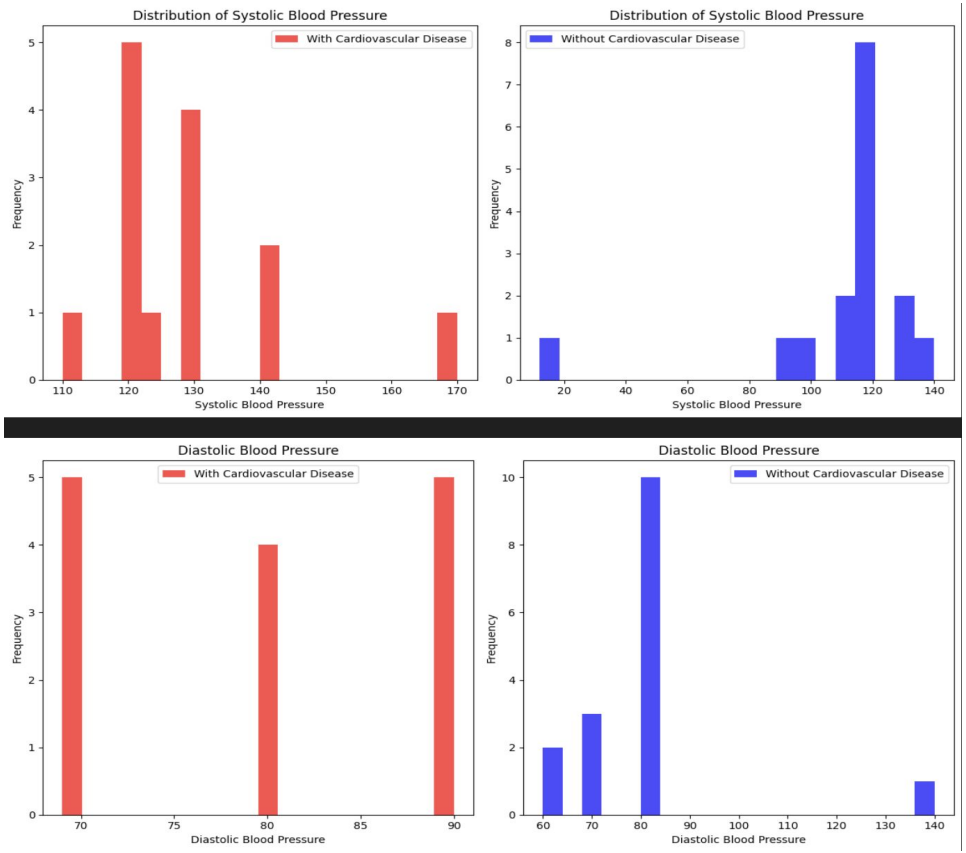
Both Alcohol consumption and Smoking should increase the risk of cardiovascular disease. However from the data set everything was almost split evenly.





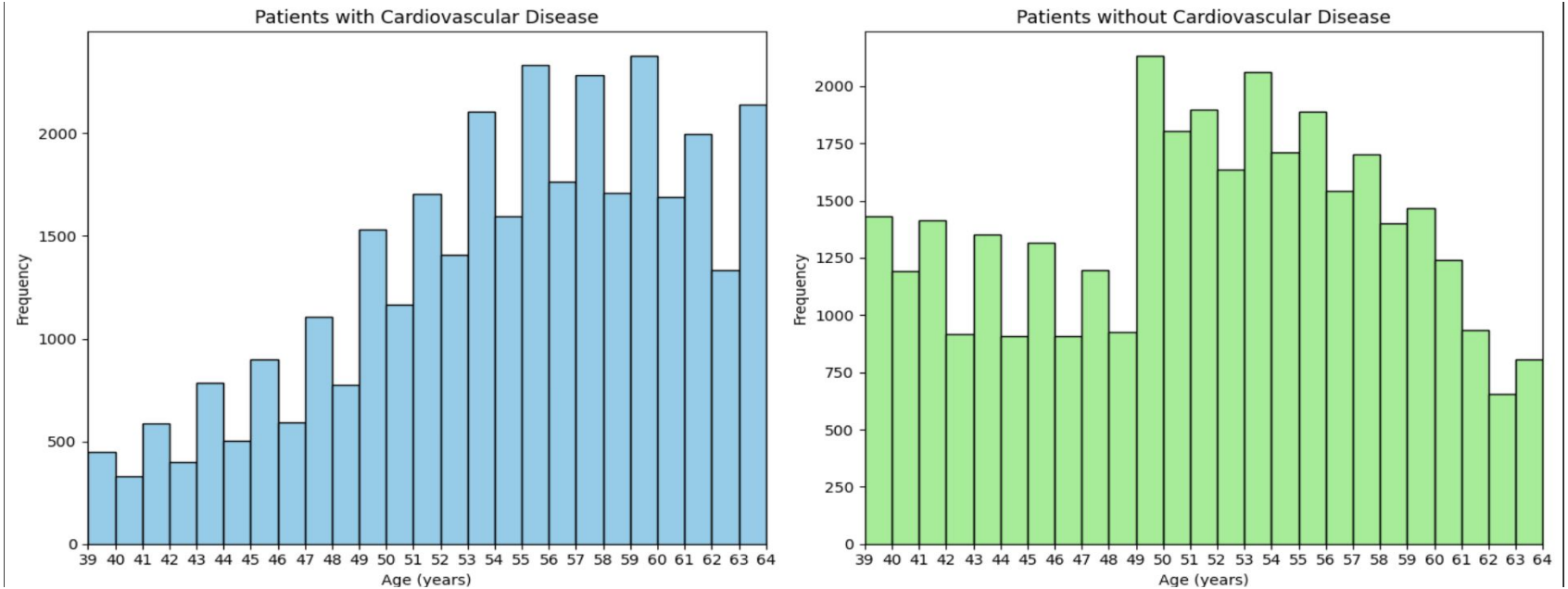
Blood Pressure and Cardiovascular Disease

Normal blood pressure for most adults is defined as a systolic pressure of less than 120 and a diastolic pressure of less than 80. Elevated blood pressure is defined as a systolic pressure between 120 and 129 with a diastolic pressure of less than 80.

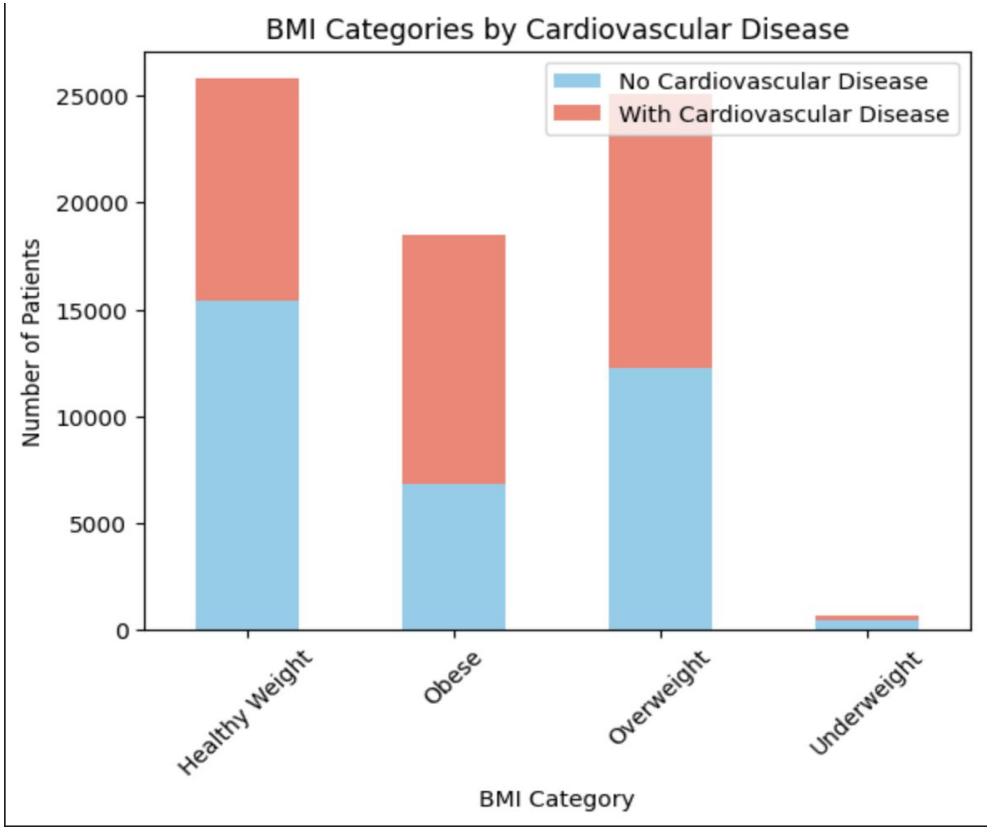


Age Distribution

The first graph is the main one used for analysis. We can see with an increase of age the risk of cardiovascular disease.



BMI



Finding the BMI gave clearer insight to which patients were more at risk for cardiovascular disease.

```
print("Numerical values for each BMI category with and without cardiovascular disease:")
print(bmi_cardio_grouped)
```

✓ 0.0s

```
Numerical values for each BMI category with and without cardiovascular disease:
cardio              0      1
bmi_category
Healthy Weight  15440  10350
Obese            6823  11651
Overweight      12290  12796
Underweight       468    182
```

BMI is calculated as BMI = weight(kg)/height(m)^2 the dataset had height in cm so we did a conversion.

Cardiovascular Disease by Gender

We wanted to see what the correlation between Gender and cardiovascular disease was.