# DELIVERABLE 2

## Exploratory Data Analysis:

Our health disease dataset contains 18 columns and 319795 rows. Here the dependent variable is **HeartDisease** and the remaining are the independent variables.

The 18 columns are:

```
In [5]: data_df.info()
```

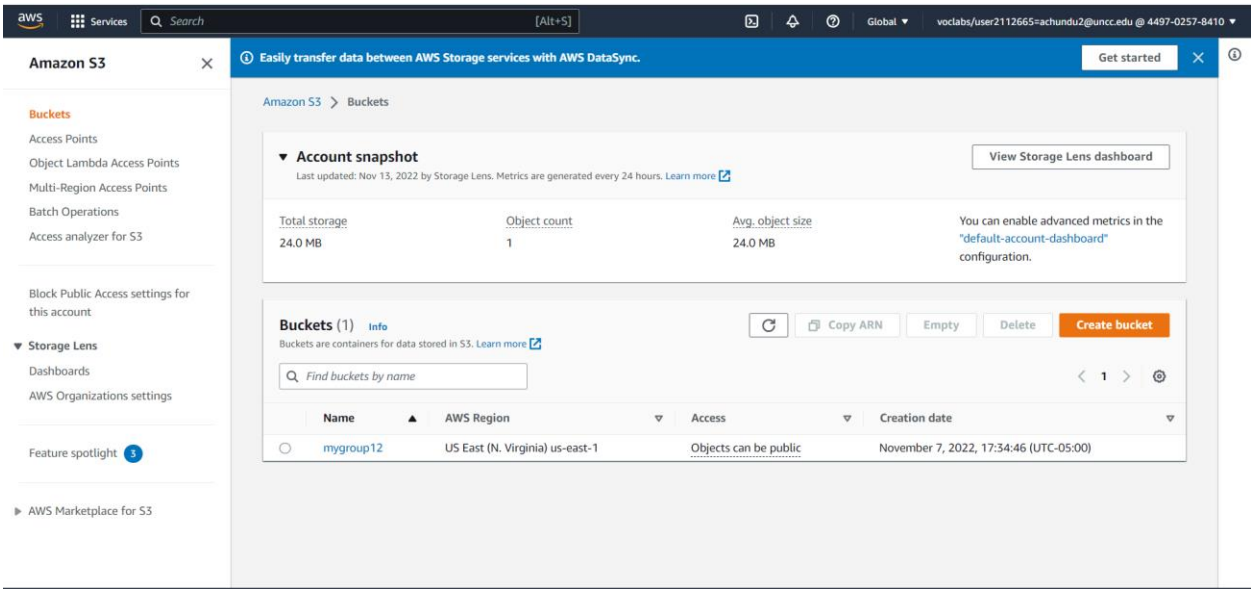
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 319795 entries, 0 to 319794
Data columns (total 18 columns):
 #   Column            Non-Null Count   Dtype
---  ------            --------------   -----
 0   HeartDisease      319795 non-null  object
 1   BMI               319795 non-null  float64
 2   Smoking           319795 non-null  object
 3   AlcoholDrinking   319795 non-null  object
 4   Stroke            319795 non-null  object
 5   PhysicalHealth    319795 non-null  float64
 6   MentalHealth      319795 non-null  float64
 7   DiffWalking       319795 non-null  object
 8   Sex               319795 non-null  object
 9   AgeCategory       319795 non-null  object
 10  Race              319795 non-null  object
 11  Diabetic          319795 non-null  object
 12  PhysicalActivity  319795 non-null  object
 13  GenHealth         319795 non-null  object
 14  SleepTime         319795 non-null  float64
 15  Asthma            319795 non-null  object
 16  KidneyDisease     319795 non-null  object
 17  SkinCancer        319795 non-null  object
```

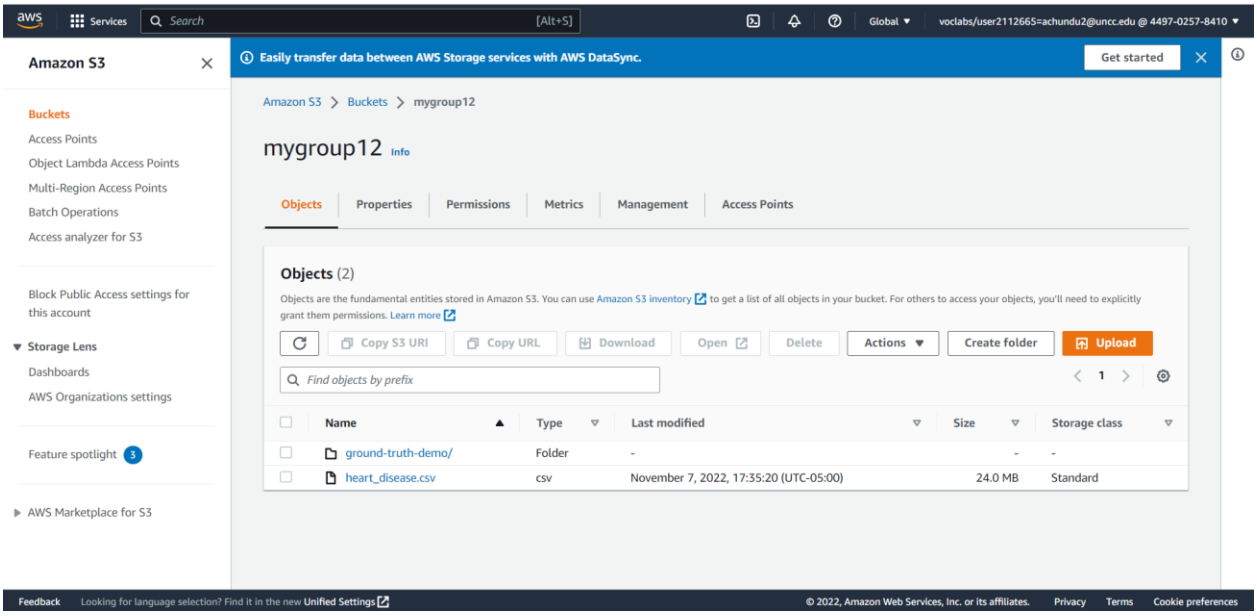
We have Independent variables in the form of categorical and numerical values.

**BMI, Physical health, Mental health, Age, and Sleeptime** are numerical values, and the rest are categorical variables.

We have performed some data analysis on our dataset using **Amazon Quicksight**. Amazon quick sight is used to create the visualizations for the data and create the dashboard.

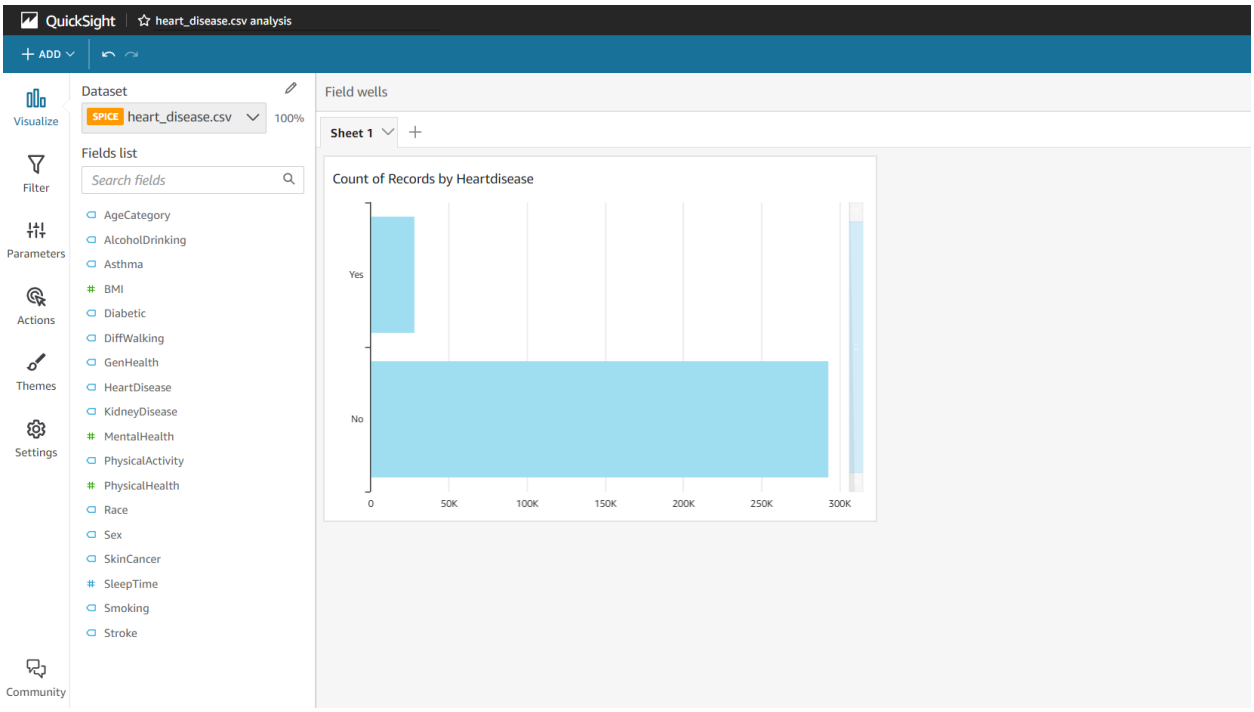
First, we created a bucket in **Amazon s3** and uploaded the **health_disease.csv** dataset into the bucket.



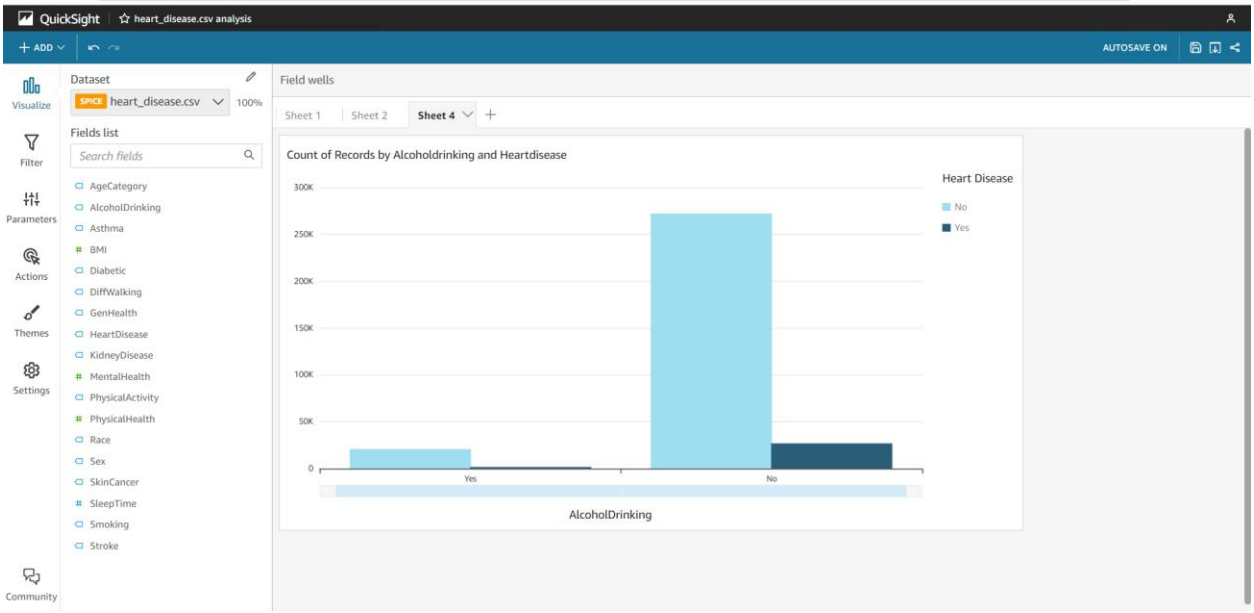



**Amazon Quicksight** is used for the below data analysis.

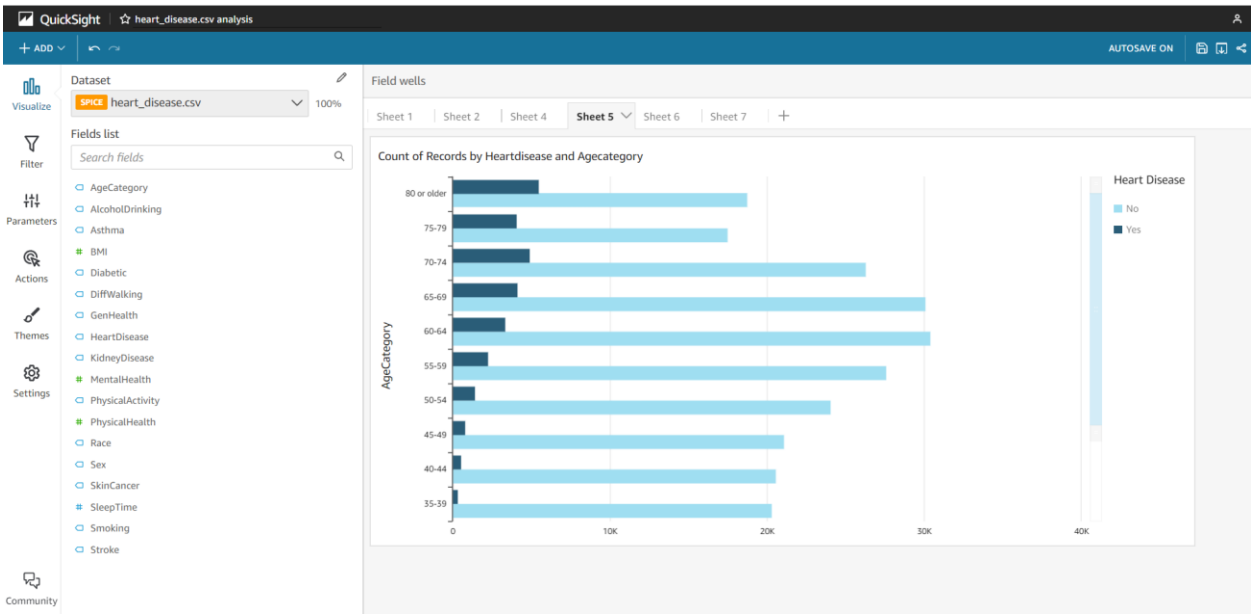
The below visualization represents the number of people having heart disease and those of not have it.



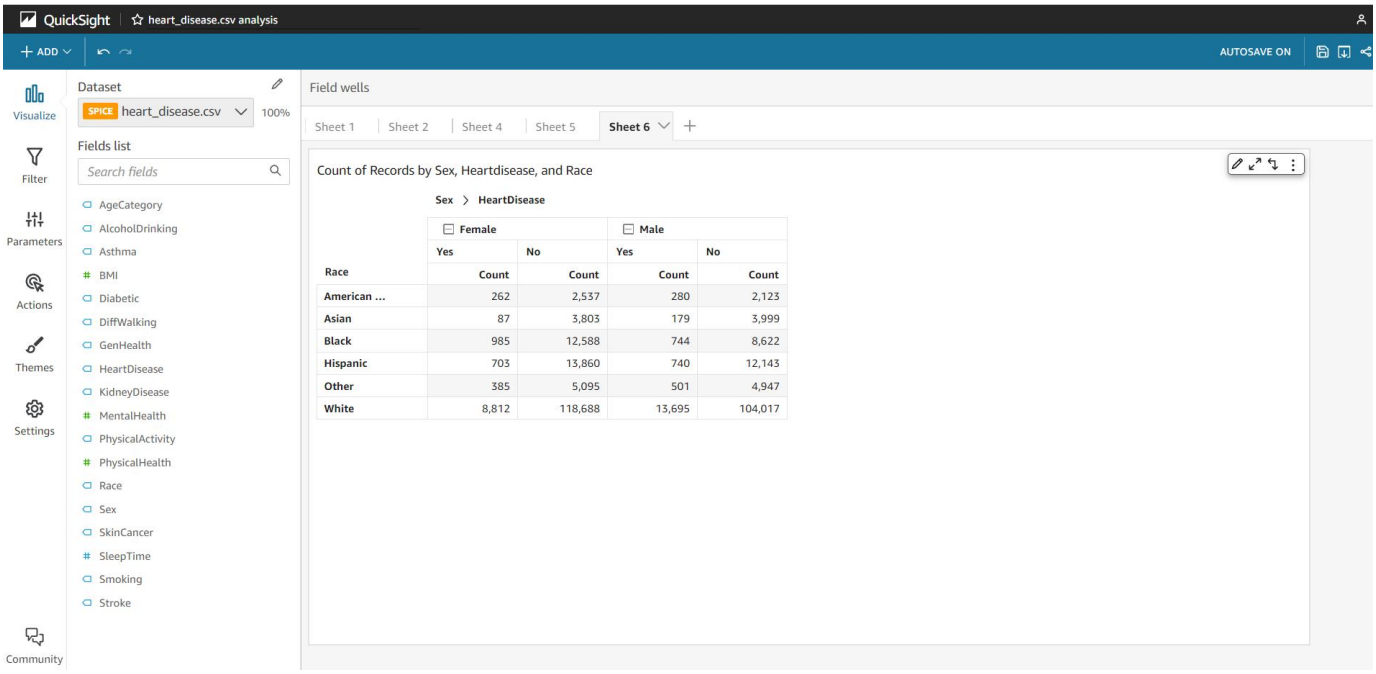
The below visualization is between the alcohol variable and the heart disease variable. It gives the impact of alcohol on heart disease.



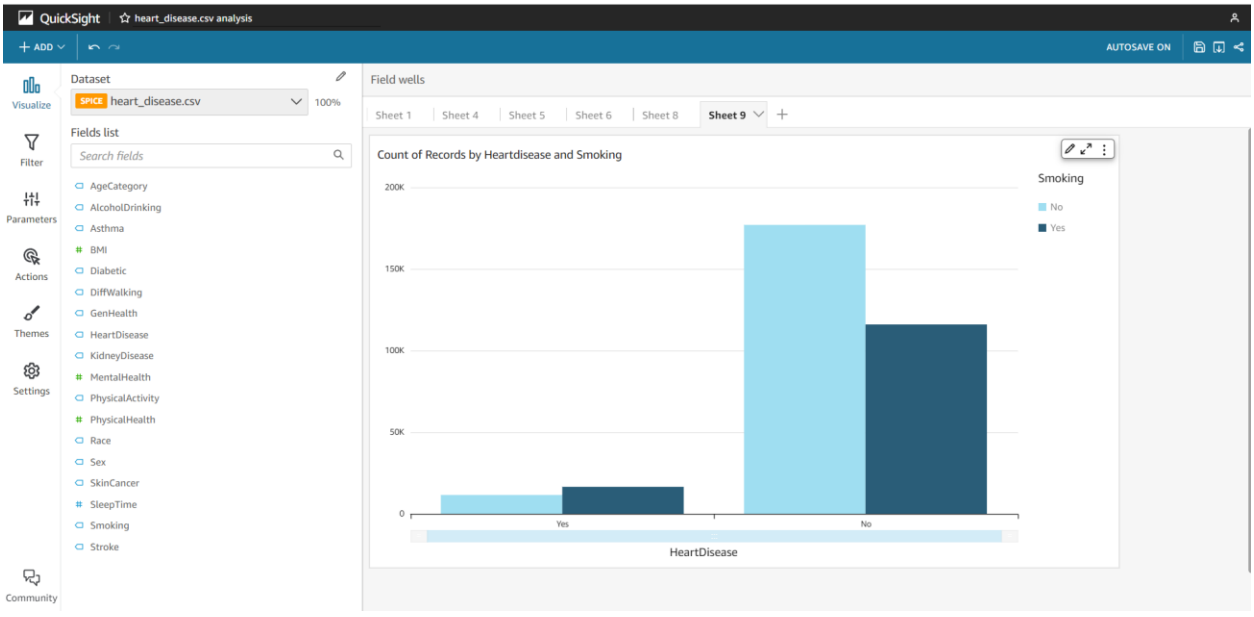

The below visualization represents the relation between age category and heart disease. It gives an analysis of how many people belonging to a particular age have been exposed to heart disease or not.



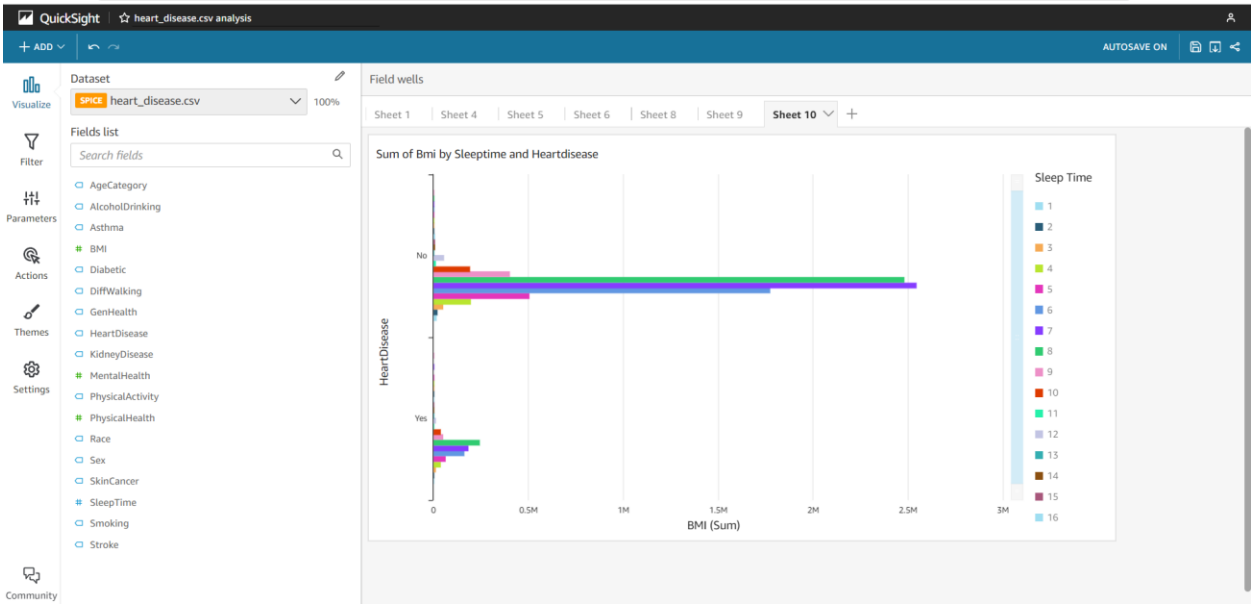
The below visualization is for Sex/Race against heart disease. It gives us an analysis of which race people and gender people are more effected by heart disease.



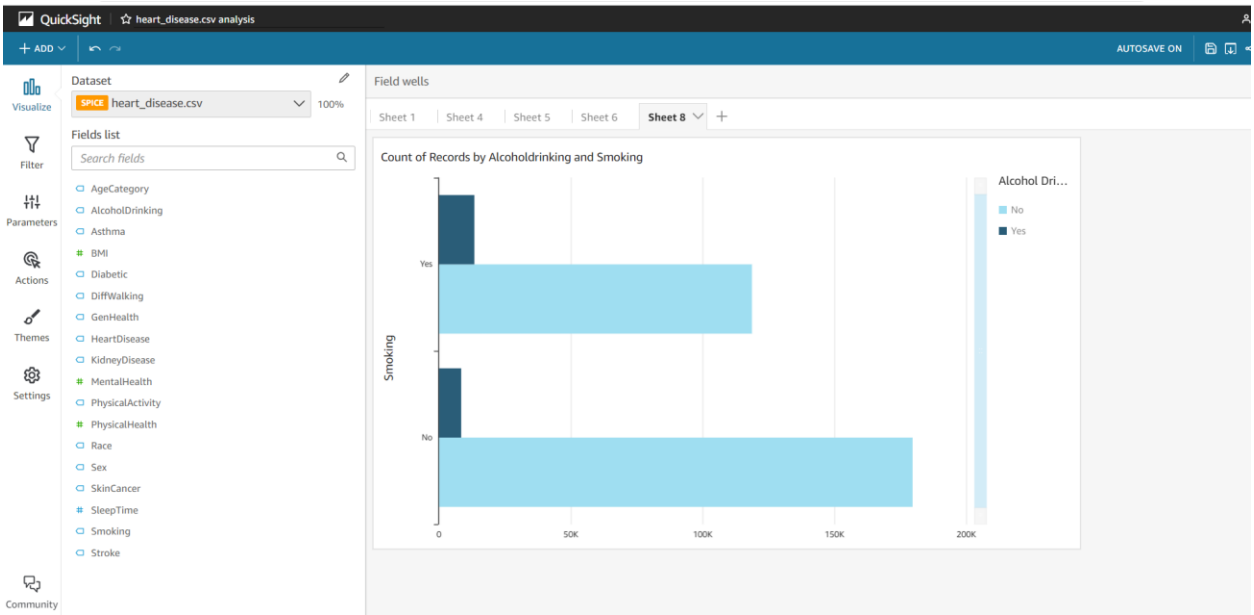

It represents the relationship between smoking and heart disease.



The below chart explains the relationship between BMI, sleep time, and heart disease.




Below dashboard gives the relationship between the alcohol and smoking variables of the dataset.



## Data Preparation:

- There are no missing values in the dataset in any columns.

```
In [4]: print("The existence of missing values in each column:")
        data_df.isnull().any()
```

```
The existence of missing values in each column:

Out[4]: HeartDisease        False
        BMI                 False
        Smoking             False
        AlcoholDrinking     False
        Stroke              False
        PhysicalHealth      False
        MentalHealth        False
        DiffWalking         False
        Sex                 False
        AgeCategory         False
        Race                False
        Diabetic            False
        PhysicalActivity    False
        GenHealth           False
        SleepTime           False
        Asthma              False
        KidneyDisease       False
        SkinCancer          False
        dtype: bool
```

We have multiple categorical values in each column. We need to convert all of them into numerical values to train the model.

**HeartDisease**, replace values Yes to 1 and No to 0

**Smoking**, replace values Yes to 3 and No to 2

**AlcoholDrinking**, replace values Yes to 3 and No to 2

**Stroke**, replace values Yes to 3 and No to 2

**DiffWalking**, replace values Yes to 2 and No to 3

**Sex**, replace values Male to 2 and Female to 3

**Diabetic**, replace the values Yes to 5,No to 4,No, borderline diabetes to 3,Yes (during pregnancy) to 2

**Asthma**: replace values Yes:2,No:3

**PhysicalActivity**, replace values Yes to 2 and No to 3

**KidneyDisease**, replace values Yes to 2 and No to 3

**SkinCancer**, replace values Yes to 2 and No to 3

**AgeCategory**, replace values '18-24':14,'25-29':13,'30-34':12,'35-39':11,'40-44':10,'45-49':9,'50-54':8,'55-59':7,'60-64':6,'65-69':5,'70-74':4,'75-79':3,'80 or older':2

**Race**, replace values 'White':7,'Black':6,'Asian':5,'American Indian/Alaskan Native':4,'Other':3,'Hispanic':2

**GenHealth**, replace values 'Very good':6,'Fair':5,'Good':4,'Poor':3,'Excellent':2

In [21]:
```
data_df['HeartDisease'].replace({'Yes':1,'No':0},inplace=True)
data_df['Smoking'].replace({'Yes':3,'No':2},inplace=True)
data_df['AlcoholDrinking'].replace({'Yes':3,'No':2},inplace=True)
data_df['Stroke'].replace({'Yes':3,'No':2},inplace=True)
data_df['DiffWalking'].replace({'Yes':3,'No':2},inplace=True)
data_df['Sex'].replace({'Male':3,'Female':2},inplace=True)
data_df['Diabetic'].replace({'Yes':3,'No':2,'No, borderline diabetes':4,'Yes (during pregnancy)':5},inplace=True)
data_df['PhysicalActivity'].replace({'Yes':3,'No':2},inplace=True)
data_df['Asthma'].replace({'Yes':3,'No':2},inplace=True)
data_df['KidneyDisease'].replace({'Yes':3,'No':2},inplace=True)
data_df['SkinCancer'].replace({'Yes':3,'No':2},inplace=True)
data_df['AgeCategory'].replace({'18-24':2,'25-29':3,'30-34':4,'35-39':5,'40-44':6,'45-49':7,'50-54':8,'55-59':9,'60-64':10,'65-6
data_df['Race'].replace({'White':2,'Black':3,'Asian':4,'American Indian/Alaskan Native':5,'Other':6,'Hispanic':7},inplace=True)
data_df['GenHealth'].replace({'Very good':2,'Fair':3,'Good':4,'Poor':5,'Excellent':6},inplace=True)
```

After the changes, the data is converted into the below form

Out[22]:

| | HeartDisease | BMI | Smoking | AlcoholDrinking | Stroke | PhysicalHealth | MentalHealth | DiffWalking | Sex | AgeCategory | Race | Diabetic | PhysicalActivity |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 0 | 16.60 | 3 | 2 | 2 | 3.0 | 30.0 | 2 | 2 | 9 | 2 | 3 | 3 |
| **1** | 0 | 20.34 | 2 | 2 | 3 | 0.0 | 0.0 | 2 | 2 | 14 | 2 | 2 | 3 |
| **2** | 0 | 26.58 | 3 | 2 | 2 | 20.0 | 30.0 | 2 | 3 | 11 | 2 | 3 | 3 |
| **3** | 0 | 24.21 | 2 | 2 | 2 | 0.0 | 0.0 | 2 | 2 | 13 | 2 | 2 | 2 |
| **4** | 0 | 23.71 | 2 | 2 | 2 | 28.0 | 0.0 | 3 | 2 | 6 | 2 | 2 | 3 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **319790** | 1 | 27.41 | 3 | 2 | 2 | 7.0 | 0.0 | 3 | 3 | 10 | 7 | 3 | 2 |
| **319791** | 0 | 29.84 | 3 | 2 | 2 | 0.0 | 0.0 | 2 | 3 | 5 | 7 | 2 | 3 |
| **319792** | 0 | 24.24 | 2 | 2 | 2 | 0.0 | 0.0 | 2 | 2 | 7 | 7 | 2 | 3 |
| **319793** | 0 | 32.81 | 2 | 2 | 2 | 0.0 | 0.0 | 2 | 2 | 3 | 7 | 2 | 2 |
| **319794** | 0 | 46.56 | 2 | 2 | 2 | 0.0 | 0.0 | 2 | 2 | 14 | 7 | 2 | 3 |

319795 rows × 18 columns

These are the steps taken in the Data preparation phase which makes the data ready to train using the machine learning model.