**Final Project Report**

**On**

# “Deaths in heart failure and its contributing factors”

*Chirag Sai Shivani , Dwivedi Sudhanshu, Gadde Manusha, Mekarthy Snigdha*

**Group 13**

**School of Informatics and Computing, IUPUI**

**INFO-B556 Biological Database Management**

**Dr. Amir Manzour**

**May 03rd, 2022**

IUPUI
SCHOOL OF INFORMATICS AND COMPUTING

## Abstract:

Kaggle Heart Attack dataset are compared and analyzed using many SQL queries and deep learning. The dataset is made up of 22 key attributes that will be used in the study. We want to discover if the effect of traditional risk studies on the variables that contribute to heart failure later in life changes with age (Andersen, 2019). We wanted to investigate the age-dependent occurrence of early-onset heart failure, heart disease associated with increased BMI, and heart problems associated with preserved changes in cholesterol levels, as well as the differential relationship of risk factors with early-onset heart problems and the relative importance of risk factors to heart failure occurrence based on age strata. Finally, we will review the findings and highlight the most significant aspects among demographics, patient vitals, and environmental factors in causing cardiac problems, which will aid in the development of strategies that can be used to reduce the risks associated.

## Introduction:

Heart failure is currently a frequent illness with a high morbidity and death incidence around the world. It is a failure to meet the total demands of circulation produced by the increased stress of health concerns that either overwork the heart or damage the heart as we age. High blood pressure that is uncontrolled is a primary cause of heart failure. Diabetes increases the risk of heart failure, and obesity causes the heart to work much harder than it would in someone who is not fat. Early detection of cardiovascular disease is critical for saving patients' lives. It is also critical to protect patients against such infections (Andersen, 2019). Various data analytics approaches are used to assist healthcare practitioners in making early diagnoses. Globally, 17.7 million people have died during 2015 as a result of cardiovascular disease. To treat the cardiovascular risk, effective decision and optimal treatment are required (Lawson, 2020).

## Background:

Heart failure has risen to the top of the list of primary causes of death in humans. It is not only the sickness but also the accompanying conditions that contribute to the death of humanity/people (Bui, 2011). Many people are unaware that hazards can be lowered with early identification and therapy of existing disorders. Various data analytics tools are used to assist healthcare practitioners with early diagnosis, accurate decision making, and optimal therapy to manage cardiac risk. The construction of this database may raise understanding of opportunities to help bridge knowledge gaps in cardiac care about cardiac occurrences caused by a variety of risk markers. Furthermore, the database established may be used to collect real-time data inputs, which can then be used to investigate epidemiology and health consequences (Dunlay, 2009).

## Current Problem for the Project topics:

With the increase in the number of variables in the dataset the challenges will also increase. With 37,080 observations, we have a large dataset. There are a few observations that are null or blank. We had a hard time separating the important properties for distinct CSV files. We had some problems importing one of the CSV files we prepared into PhpMyAdmin due to its big size. We also attempted to import the file through the terminal, but were unable to do so. After several unsuccessful attempts, we were eventually able to import the CSV files into the MySQL workbench. Before importing the csv files into PhpMyAdmin, we even compressed them.We even compressed the.csv files before importing them into PhpMyAdmin. We had trouble finding literature for most of the category columns in the data since they had binary values. It took some time for us to research the issue and find relevant literature in order to figure out what the values meant.

**Project Goal:**
The goal of the project is to review the findings and identify the most positive influential factors in the development of cardiac issues, including demographics, patient vitals, and environmental factors. This information can be used to help create ways to reduce the hazards associated with early diagnosis and management of existing disorders.

**Proposed Methods:**

- Data Collection
- Data Extraction, Cleaning, and Storage.
- Data Visualization
- Data Analysis.

**Methods:**

- Data Extraction or Collection:

We have extracted the dataset from Kaggle. With the help of spreadsheets, we decoded the variables and saved our data in CSV file format. Later, we created three separate .csv files based on the relevant attributes.

- Data importing:

We have imported the .csv files created into MySQL workbench.

- Data Cleaning:

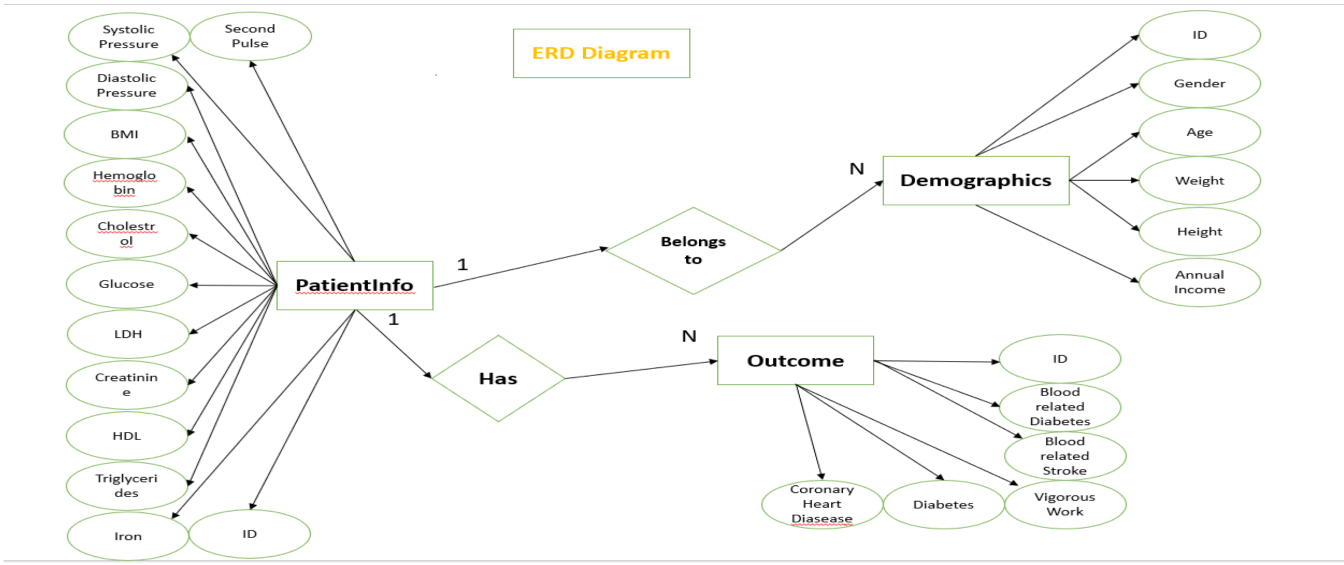
The cleaning and removal of outliers were done in PHPMyAdmin and finally we had 30,000 observations for the data analysis. The data importing was done through import functions.

- Data Modeling:

We have established connections by adding constraints to the tables. Composition of Entity relationship diagram (ERD), logical modeling design, a relational schema of the given entities and their attributes for better understanding of their relationship and structure of the dataset.

**Expected Results:**
We wanted to explore if the impact of traditional risk studies on the variables that contribute to heart failure later in life varied with age. We wanted to investigate the age-related occurrence of early-onset heart failure, heart disease with elevated BMI, and heart conditions with preserved changes in total cholesterol, as well as the differential relationship of risk factors with early-onset heart problems and the relative importance of risk factors to the occurrence of heart failure based on age strata.

## Entity-Relationship Diagram:



## ERD Explanation:

We have used an ERD to design a relational database and to highlight the entities,attributes and relationship for the effective data visualization of the project. There are three tables in the ERD model of this project.They are:

1. Patient information
2. Demographics
3. Outcomes

The first table that we created was "patientinformation" with the "id" attribute as the primary key.

The 2nd table and 3rd tables designed were "demographics" and "outcomes" tables respectively .

The "id" column from the patientinformation table serves as a foreign key in these 2 tables.

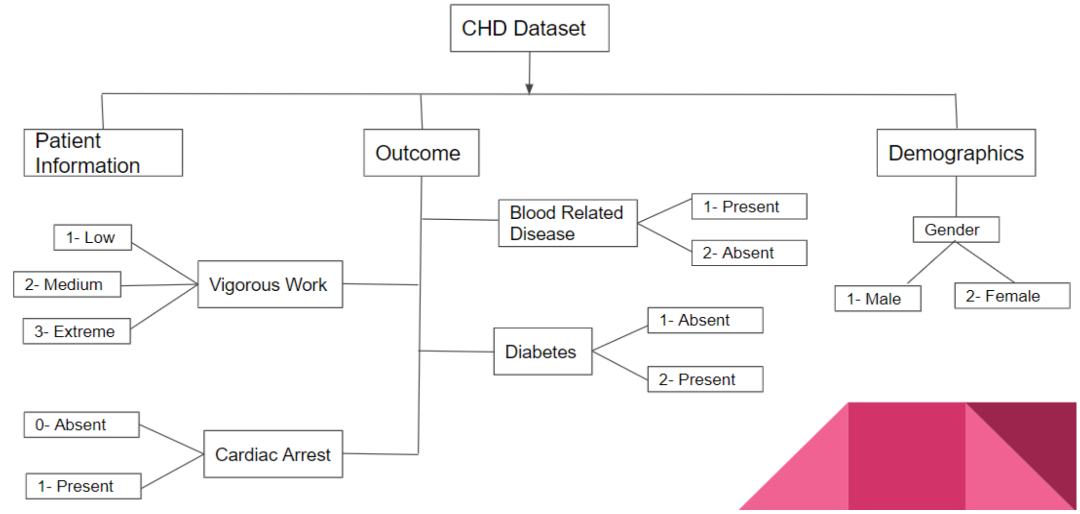
One patient can have many outcomes which indicate one to many relationship between patientinformation and outcomes.

Every id will have individual demographic details that indicate their is a one to one relationship between patientinformation table and demographics table

**Data Analysis**:

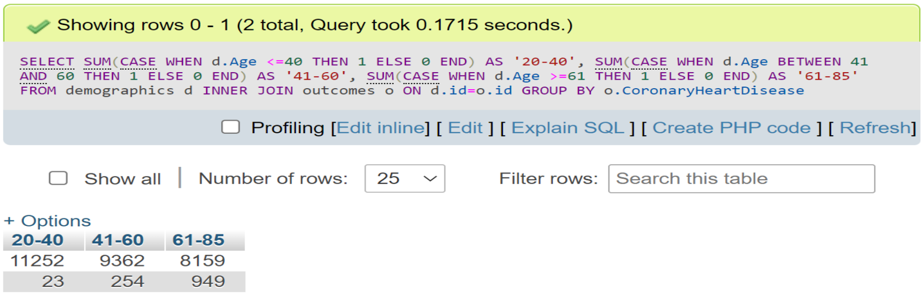
Starting with the data analysis we first checked the literature review of the dataset and got the detailed explanation about the data attributes. The data attributes in outcome were like 1 and 2 values for blood related disease where 1 means the disease is absent and 2 means it's present. We found the same information for a few other attributes.

## Data Attributes



## Research Question:

- Which age group is highly affected with cardiac heart failure?



✓ Showing rows 0 - 1 (2 total, Query took 0.1715 seconds.)

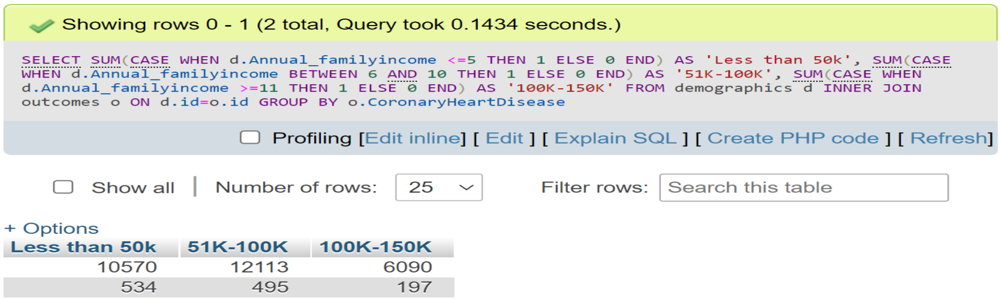
```
SELECT SUM(CASE WHEN d.Age <=40 THEN 1 ELSE 0 END) AS '20-40', SUM(CASE WHEN d.Age BETWEEN 41
AND 60 THEN 1 ELSE 0 END) AS '41-60', SUM(CASE WHEN d.Age >=61 THEN 1 ELSE 0 END) AS '61-85'
FROM demographics d INNER JOIN outcomes o ON d.id=o.id GROUP BY o.CoronaryHeartDisease
```

□ Profiling [Edit inline] [ Edit ] [ Explain SQL ] [ Create PHP code ] [ Refresh]
□ Show all | Number of rows: 25 Filter rows: Search this table
+ Options

| 20-40 | 41-60 | 61-85 |
|---|---|---|
| 11252 | 9362 | 8159 |
| 23 | 254 | 949 |

Age is divided into three cohort groups and according to CDC 41-60 are highly vulnerable however the result shows 61-85 are the most affected age group. So as a result, the chances of occurrence of CHD increases with increase in age.

- How is the annual family income of a person influencing the occurrence of heart failure?



Showing rows 0 - 1 (2 total, Query took 0.1434 seconds.)

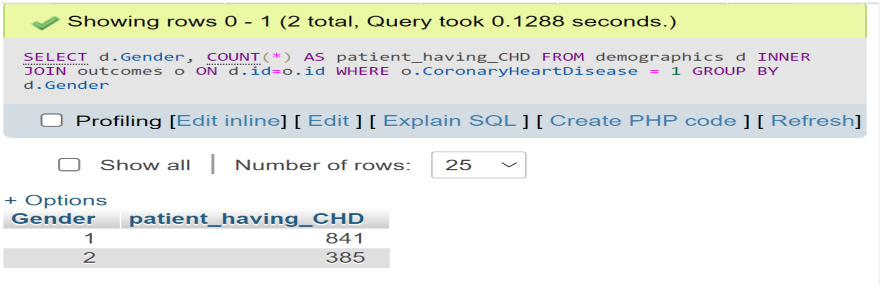
```
SELECT SUM(CASE WHEN d.Annual_familyincome <=5 THEN 1 ELSE 0 END) AS 'Less than 50k', SUM(CASE
WHEN d.Annual_familyincome BETWEEN 6 AND 10 THEN 1 ELSE 0 END) AS '51K-100K', SUM(CASE WHEN
d.Annual_familyincome >=11 THEN 1 ELSE 0 END) AS '100K-150K' FROM demographics d INNER JOIN
outcomes o ON d.id=o.id GROUP BY o.CoronaryHeartDisease
```

□ Profiling [Edit inline] [ Edit ] [ Explain SQL ] [ Create PHP code ] [ Refresh]
□ Show all | Number of rows: 25
Filter rows: Search this table
+ Options

| Less than 50k | 51K-100K | 100K-150K |
|---|---|---|
| 10570 | 12113 | 6090 |
| 534 | 495 | 197 |

The result states that people whose income is less than $50k have higher chances of developing Cardiac Arrest. This could be because if a family is earning less there it means their living habits and lifestyle is not better than someone who is earning higher.

- What is the gender distribution of CHD?



Showing rows 0 - 1 (2 total, Query took 0.1288 seconds.)

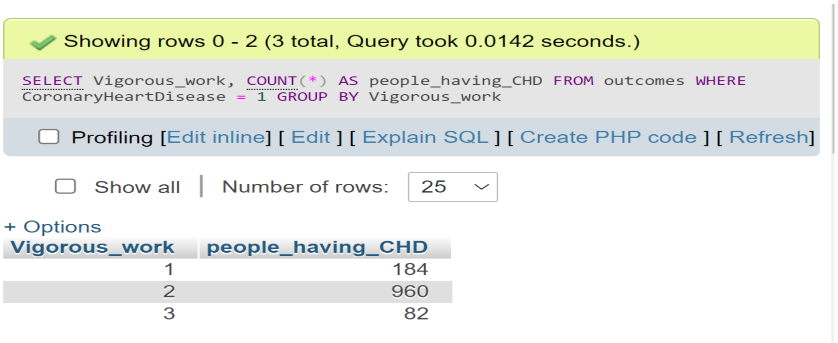
```
SELECT d.Gender, COUNT(*) AS patient_having_CHD FROM demographics d INNER
JOIN outcomes o ON d.id=o.id WHERE o.CoronaryHeartDisease = 1 GROUP BY
d.Gender
```

□ Profiling [Edit inline] [ Edit ] [ Explain SQL ] [ Create PHP code ] [ Refresh]
□ Show all | Number of rows: 25
+ Options

| Gender | patient_having_CHD |
|---|---|
| 1 | 841 |
| 2 | 385 |

Male have higher chances of developing heart arrest than females, which was also accepted by CDC.

- Does the vigorous work done by the people affect CHD?



Showing rows 0 - 2 (3 total, Query took 0.0142 seconds.)

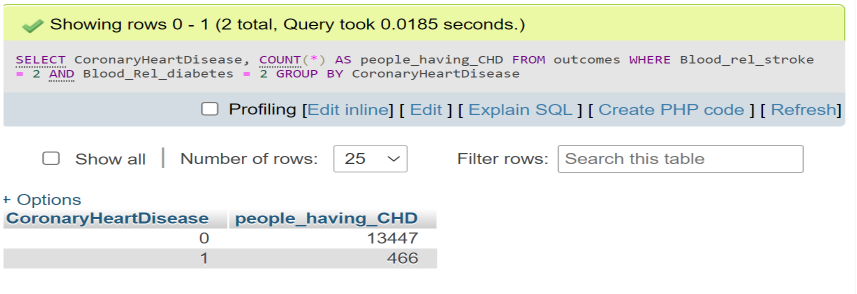
```
SELECT Vigorous_work, COUNT(*) AS people_having_CHD FROM outcomes WHERE
CoronaryHeartDisease = 1 GROUP BY Vigorous_work
```

□ Profiling [Edit inline] [ Edit ] [ Explain SQL ] [ Create PHP code ] [ Refresh]
□ Show all | Number of rows: 25
+ Options

| Vigorous_work | people_having_CHD |
|---|---|
| 1 | 184 |
| 2 | 960 |
| 3 | 82 |

Here, we got a surprising result because according to CDC individuals who do more vigorous work have higher chances of developing CHD but according to the data someone who does mediocre work has higher chances.

- How many people with comorbidities(Blood Related Stroke and Blood Related Diabetes) have CHD?



Showing rows 0 - 1 (2 total, Query took 0.0185 seconds.)

```
SELECT CoronaryHeartDisease, COUNT(*) AS people_having_CHD FROM outcomes WHERE Blood_rel_stroke
= 2 AND Blood_Rel_diabetes = 2 GROUP BY CoronaryHeartDisease
```

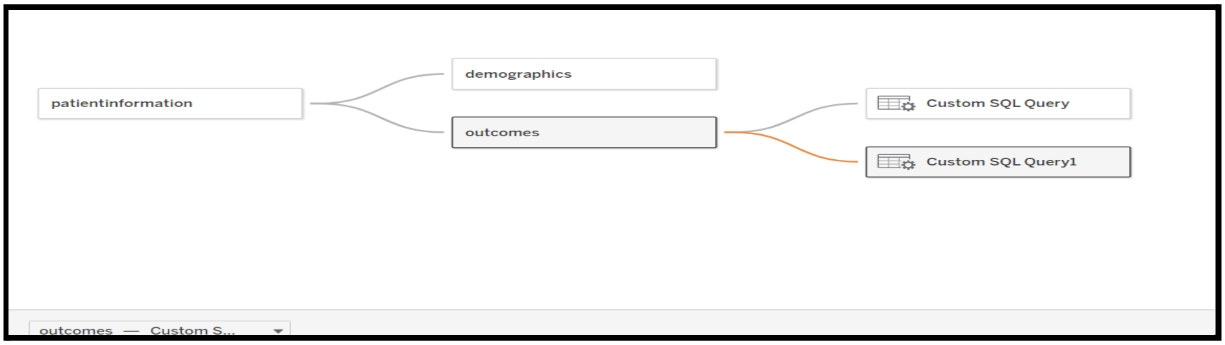
□ Profiling [Edit inline] [ Edit ] [ Explain SQL ] [ Create PHP code ] [ Refresh]
□ Show all | Number of rows: 25 Filter rows: Search this table
+ Options

| CoronaryHeartDisease | people_having_CHD |
|---|---|
| 0 | 13447 |
| 1 | 466 |

Based on this result if an individual have underline or comorbidities such as Blood Related Stroke and Blood Related Diabetes have less chances of developing CHD

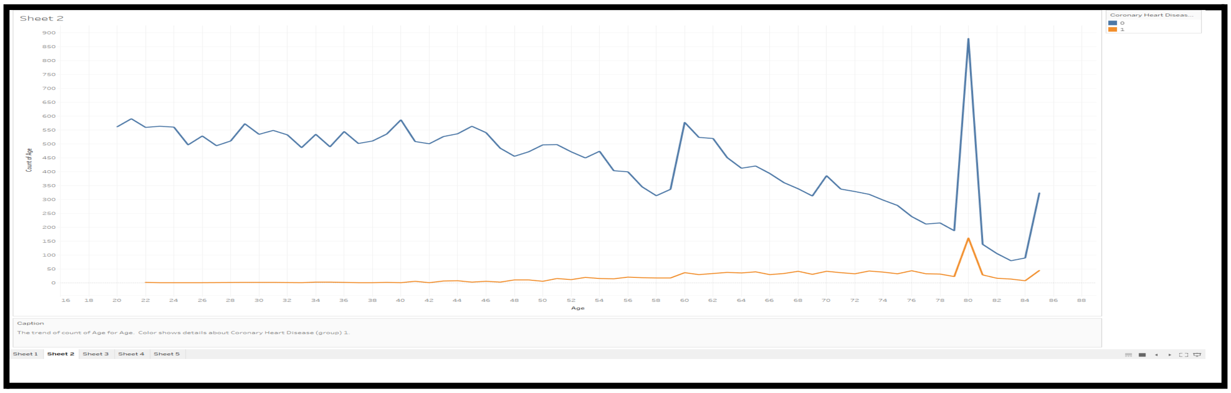
## DATA VISUALIZATION

### Connecting SQL Database to Tableau:



### Visualizations -

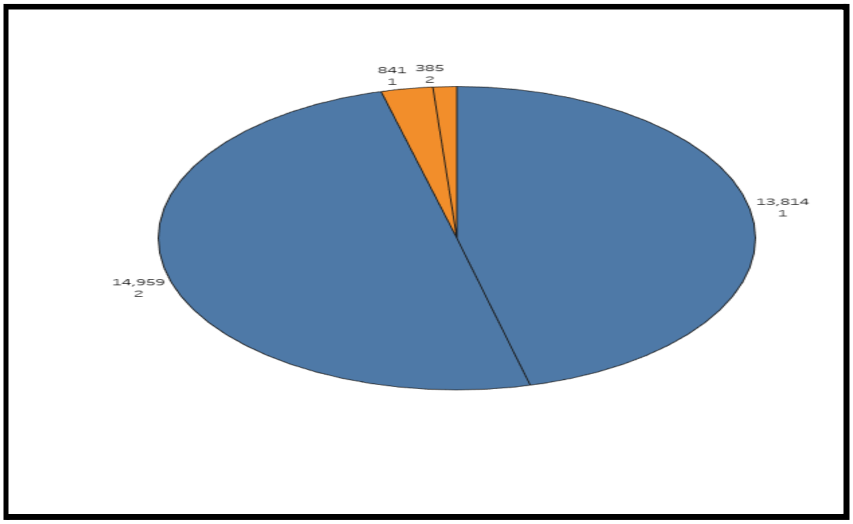
- *Age group and Cardiac Heart Failure relationship*



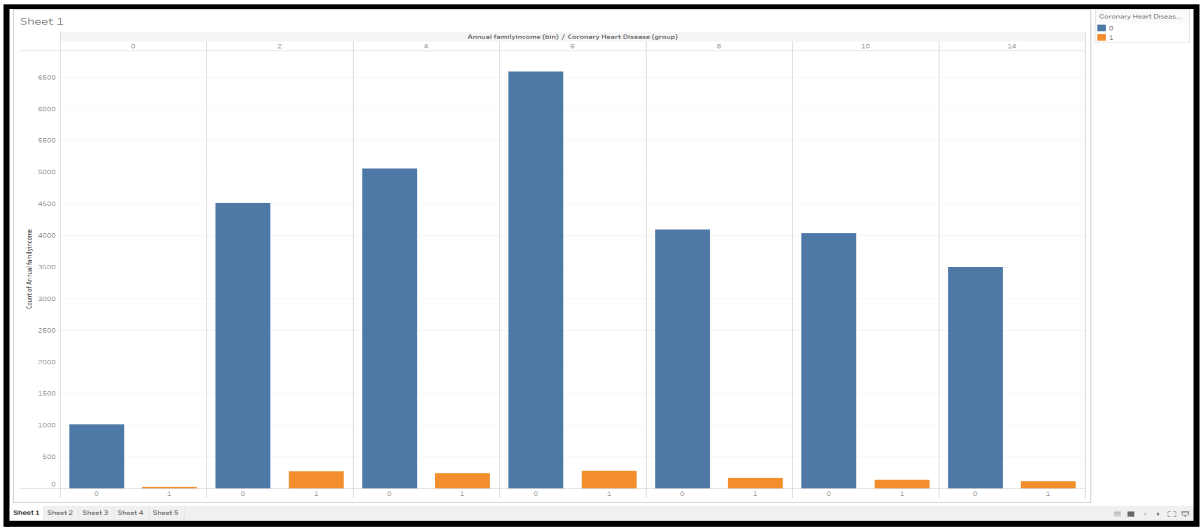
Caption

The trend of count of Age for Age. Color shows details about Coronary Heart Disease (group) 1.

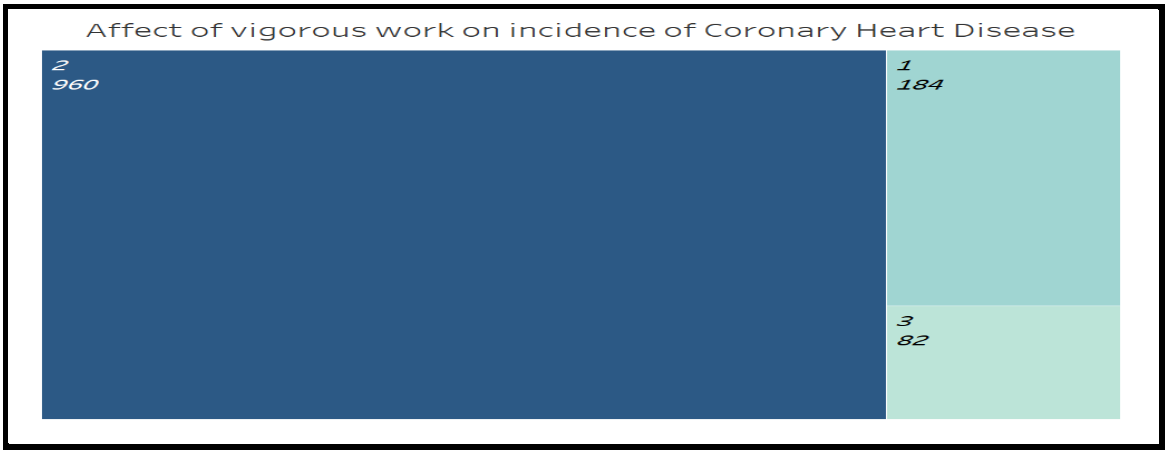
- ***Gender Distribution Of CHD***



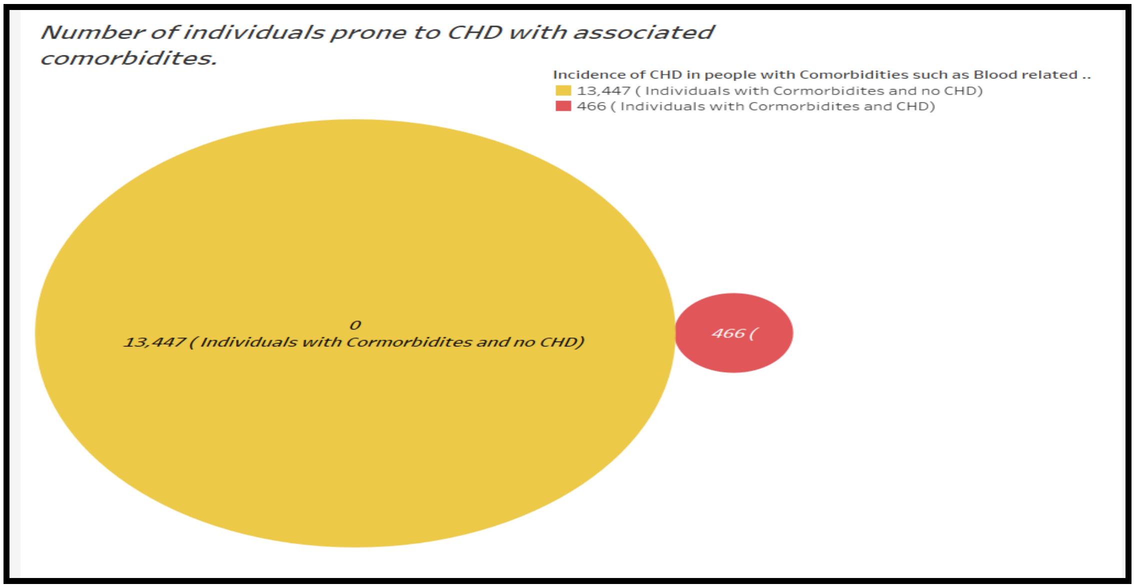
- ***Annual family income of a person influencing the occurrence of heart failure***



- ***Tree Mapping illustrating affect of different levels vigourous work on incidence of Coronary Heart Diseases***



- ***Bubble Chart representing the number of individuals prone to Coronary Heart Diseases with associated comorbities such as Blooc-related stroke , Blood-related diabetes.***



**DISCUSSION:**

Basic patient information such as age and gender, as well as a number of metabolic, biochemical, health-related, and lifestyle variables, are all needed to evaluate risk. Individuals with cardiovascular disease who pose a greater cardiovascular danger (due to the existence of one or more risk factors such as hypertension, diabetes, hyperlipidemia, or pre-existing illness) should be diagnosed and treated as soon as possible. These types of datasets might be used to extract information in real time, supporting us in finding the occurrence of chronic heart disease and, in the end, taking into account all relevant elements. Behavioral modifications such as physical exercise, a good diet, minimizing cigarette exposure, controlling alcohol use, and lowering stress can all help to avoid CVD, despite its high death rate (Hooker, 2013).

## Result:

Based on the research, we carefully searched for variables that contributed to the occurrence of CHD. Demographic characteristics such as age, BMI, and gender distribution have a greater influence on the mortality rate caused by the incidence of CHD. This study also assisted in determining the vitals that increase the death rate among CHD patients, and this clinical dataset will assist various hospitals and public health departments in monitoring the illness and taking necessary preventative measures. Cholesterol builds up in the arteries bloodstream as people age due to hereditary or lifestyle factors. Modifications in the tiny blood arteries of the heart may increase the risk of CHD as you become older.The risk could be higher in older persons, women, and people with diabetes or obesity. Men and women are both affected by coronary heart disease. Men are more likely to have obstructive coronary artery disease than women.The people with the higher annual income are less prone to the Coronary heart disease.

## Conclusion:

To summarize, heart failure develops over a period of time, therefore you must be aware of changes in your body. Keep track of your BP, weight, as well as other vital indicators as directed by your doctor. Get your laboratory tasks completed as directed since it provides important information about your heart health and medication requirements.

Because anxiety and sadness, which can cause you to feel anxious, are typical side effects of congestive heart failure, try to find outlets for your stress.

**References:**

Andersen, L. W., Holmberg, M. J., Berg, K. M., Donnino, M. W., & Granfeldt, A. (2019). In-Hospital Cardiac Arrest: A Review. JAMA, 321(12), 1200–1210. https://doi.org/10.1001/jama.2019.1696

Bui, A. L., Horwich, T. B., & Fonarow, G. C. (2011). Epidemiology and risk profile of heart failure. Nature reviews. Cardiology, 8(1), 30–41. https://doi.org/10.1038/nrcardio.2010.165

Dunlay, S. M., Weston, S. A., Jacobsen, S. J., & Roger, V. L. (2009). Risk factors for heart failure: a population-based case-control study. The American journal of medicine, 122(11), 1023–1028. https://doi.org/10.1016/j.amjmed.2009.04.022

Gardner, A. (n.d.). Heart failure and joint conditions can occur with it. https://www.webmd.com/heartdisease/heart-failure/heartfailure-common-conditions

Lawson, C. A., Zaccardi, F., Squire, I., Okhai, H., Davies, M., Huang, W., Mamas, M., Lam, C. S. P., Khunti, K., & Kadam, U. T. (2020). Risk factors for heart failure. Circulation: Heart Failure, 13(2). https://doi.org/10.1161/circheartfailure.119.006472