## Datasets

The two datasets used for this assignment are about heart disease. They both come from the American Centres for Disease Control and Prevention (CDC) and they are both about heart disease.

### Dataset #1

The first one is the National Health Interview Survey (NHIS) National Cardiovascular Disease Surveillance Data. [1] This tracks the rate of cardiovascular disease and any risk factors for cardiovascular diseases, and the data is organised based on location such as state, county etc. it has 4,455 rows and 27 columns. The table below contains a description of the 27 columns.

| Column | Description | Data Type |
|---|---|---|
| Year | - | Temporal |
| Locationabbr | Location Abbreviation | Location |
| Locationdesc | Location Description | Location |
| Datasource | Where the information came from | Categorical |
| priorityarea1 | Priority Area (Million Hearts® or None) | Categorical |
| priorityarea2 | Priority Area (ABCS or None) | Categorical |
| priorityarea3 | Priority Area (Healthy People 2020 or None) | Categorical |
| priorityarea4 | Priority Area (AHA 2020 Goals: Cardiovascular Health Metrics or None) | Categorical |
| Category | - | Categorical |
| Topic | - | Categorical |
| Indicator | - | Categorical |
| break_out_category | How is the data being divided up (e.g. race, age) | Categorical |
| break_out | Which group from the break_out_category (e.g. if break_out_category is race, break_out could be 'indian') | Categorical |
| data_value_type | The data type e.g. percentage | Categorical |
| data_value_unit | Symbol for the datatype e.g. % | Categorical |
| data_value | Value for the data type | Continuous |
| data_value_footnote_symbol | Symbol that would be used for the flag footnotes | Categorical |
| data_value_footnote | Footnote description | Categorical |
| confidence_limit_low | 95% confidence interval lower bound | Continuous |
| confidence_limit_high | 95% confidence interval upper bound | Continuous |
| Categoryid | - | Categorical |
| Topicid | - | Categorical |
| Indicatorid | - | Categorical |
| Breakoutcategoryid | - | Categorical |
| Breakoutid | - | Categorical |
| Locationid | - | Location |
| Geolocation | Coordinates of location | Location |

### Dataset #2

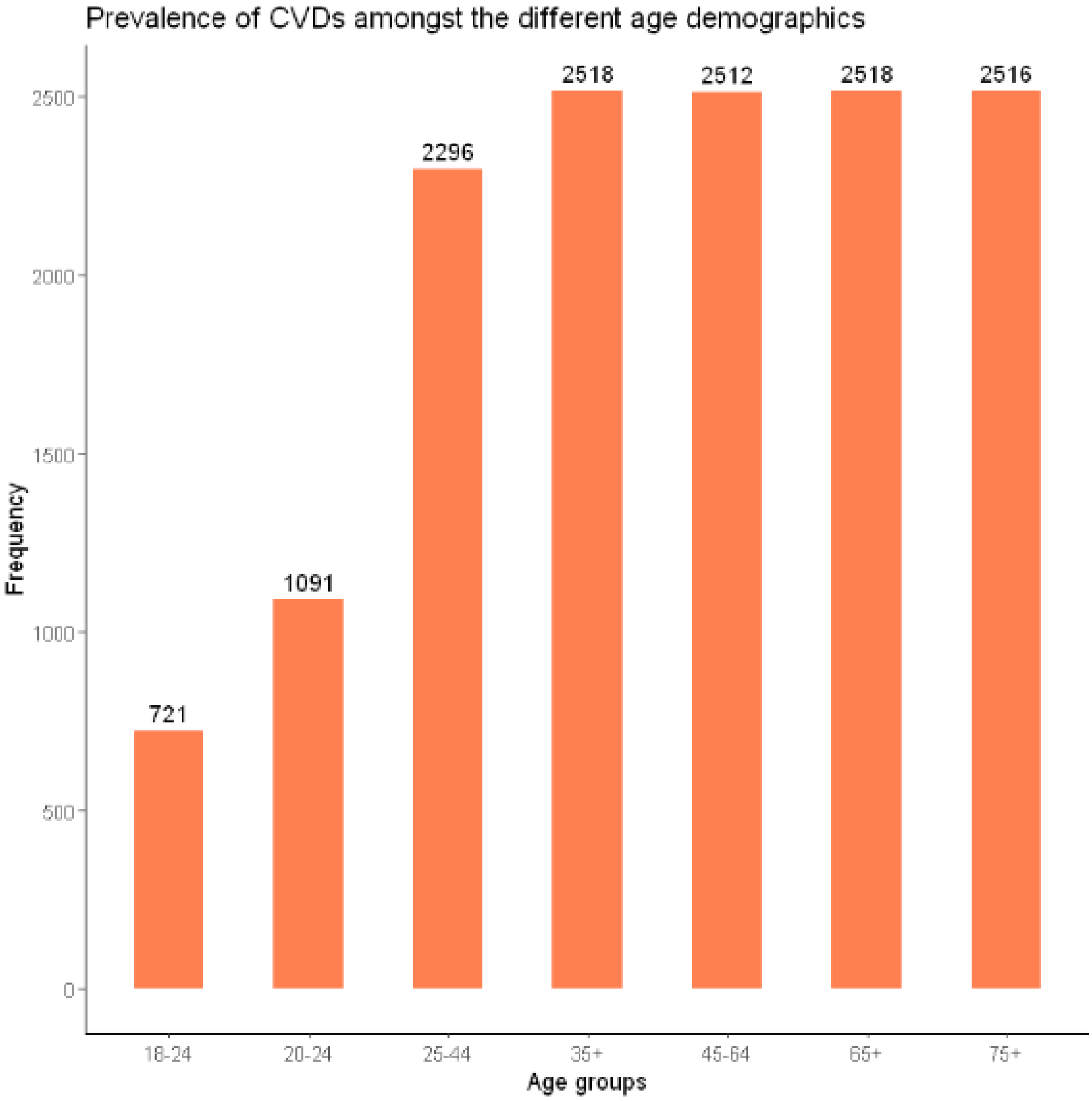
The second dataset also comes from CDC and it contains all of the same columns as the first, so the table above also describes the columns in this dataset. [2] However, the difference is that the first one tracks the rate of cardiovascular disease from 2000 - 2014, this one tracks the rate of heart disease from 2011 – 2014. The other difference is that this dataset contains 35,004 rows, which is a lot more than the first dataset.

These datasets can be binded with rbind() function, since they share the same columns.

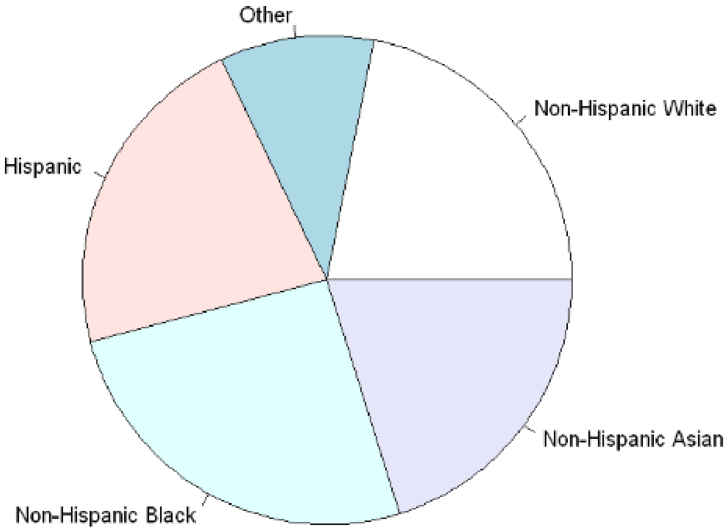
## Form Big Idea

- Demonstrate true understanding of what the data says using visualisation.

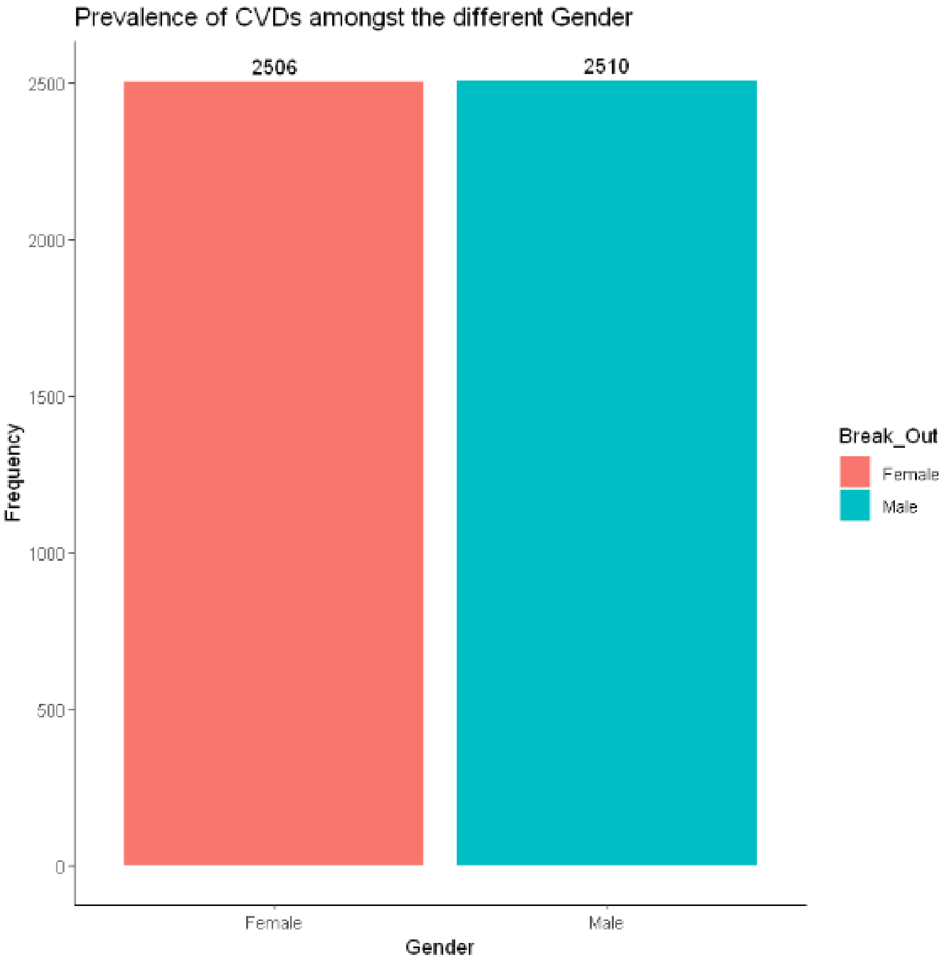
The data sets contain information about people who have cardiovascular disease, but it also contains information about those who are risk for cardiovascular diseases. The data can be plotted by multiple different categories (which is tracked by the break_out column). For example if we plot this column based in the age groups we get a graph like this:



We can see they increase as the age increases. Plotting by race reveals that there is not much of a difference in terms of ethnicity.

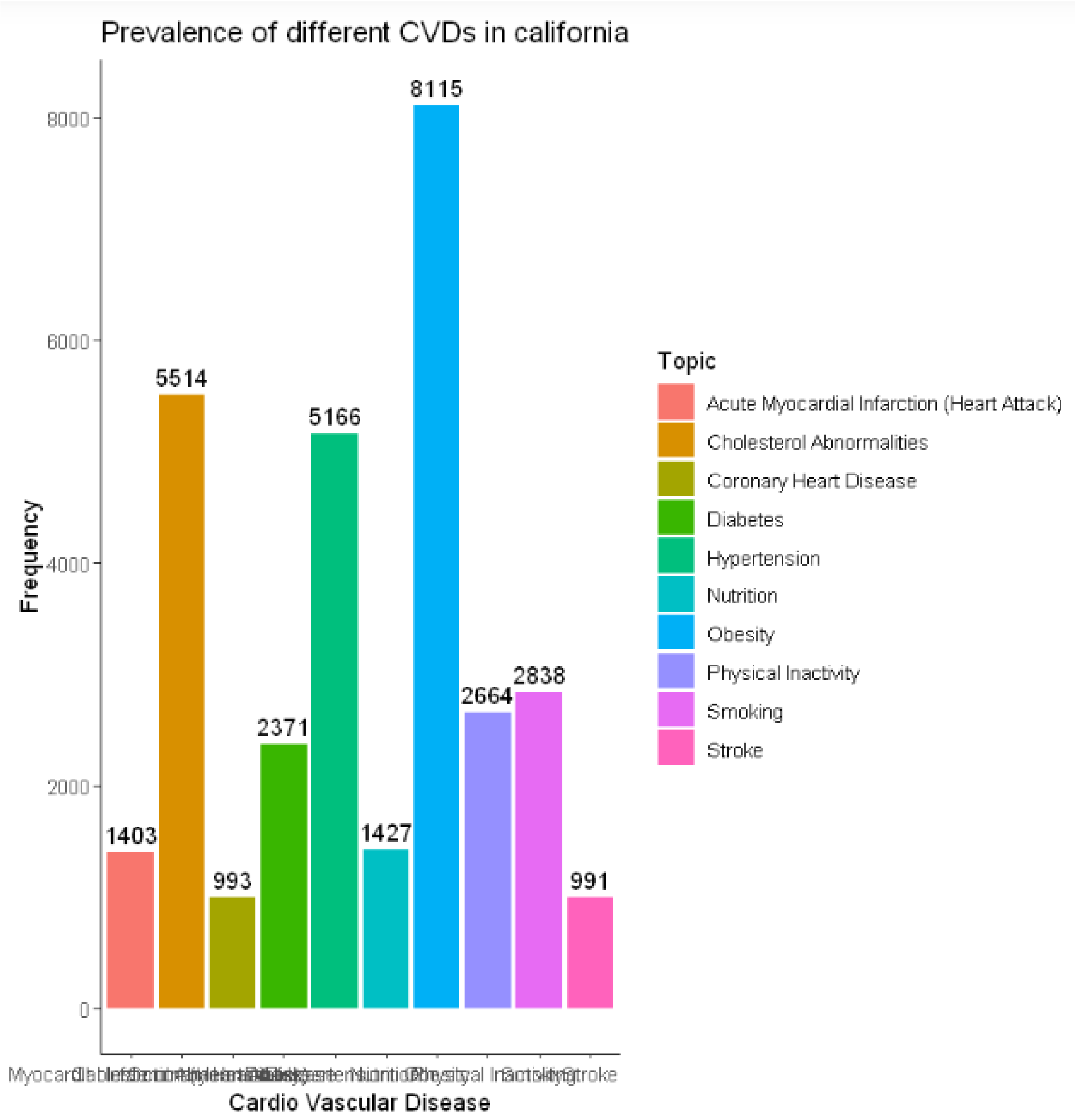


Lastly, theres also not that much of a difference between genders



`

One final note on this is that obesity appears to be the most common CVD by far.



- Identify a strong story context.

According to the data, the ethnicity someone is doesn't make too much of a difference, the biggest thing appears to be age. The odds of getting heart disease increases with age rather than race or gender. Obesity is the most common type of risk people face in America.

- Identify a valid audience and an individual or group to which the story will be delivered.

The audience could be the people from not only the locations mentioned in the Location columns, but also everyone in the USA. More specifically, the specific individual this targeted towards could be an elected politician within these locations. For example, if they see that there is a greater prevalence of heart disease for a certain age group, they can put methods in place to try minimise their risk.

For example, if there is a rise of the risk for cardiovascular disease amongst 18-24 year olds in a certain state over the past few years, that may indicate that they need to bring in a certain policy saying that colleges in that area should supply healthier food.

- Identify risks and opportunities

There are many opportunities by getting this information to the right target audience. For one thing, it allows health officials to track which demographics are at the most risk for cardiovascular disease. This may help them better organise hospitals so it can cater to these demographics. The same goes for location. A politician may be able to see which locations need better health care facilities.

This also outlines the risk of not paying attention to the data. If they are not paying attention to the areas that are at higher risk, then the healthcare in those high-risk areas may suffer.

- Story in a sentence

CVDs are one of the most common causes of deaths annually, it affects people from all different ethnicities and genders almost equally and the chance of getting one increases as you age, with obesity being the most common.

## Datasets

[1] Centres for Disease Control and Prevention, (2016), National Health Interview Survey (NHIS) - National Cardiovascular Disease Surveillance Data, https://data.world/cdc/nhis-national-cardiovascular, ***Date Accessed:*** 3rd April 2020

[2] Centres for Disease Control and Prevention, (2016), Behavioral Risk Factor Heart, https://data.world/cdc/behavioral-risk-factor-heart, ***Date Accessed:*** 3rd April 2020