DH302: **Midsem** (and Endsem) project

Chirag (BS Mathematics and IDDDP Healthcare informatics) and Ameya (BS Mathematics)

Dataset: https://catalog.data.gov/dataset/u-s-chronic-disease-indicators-cdi-d1f3d

Name: "U.S._Chronic_Disease_Indicators__CDI_"

Dimensions: 956638 x 34

In this crude data set, unorganised "facts" or "data" have been presented. We will try to extract case time series (yearwise) of four conditions:

- Coronary artery disease
- Congestive heart failure
- Cerebro-vascular disease (Stroke)
- Other Diseases of the heart

We will try to answer the following questions using the above data after spreadsheet analysis in R. We will consider the states of. We are leaving out Texas because of the lack of diversity. We are leaving other states due to many NA values in the data set due to inconsistent measurements. We will consider the following races.

i. White
ii. Black
iii. Asian
iv. Latin/Hispanic
v. Asia/Pacific

We then consider the following questions:

1. How common is cholesterol screening among adults (18+) in these states? And what is the high cholesterol prevalence in these states?
2. What is the prevalence of high blood pressure (hypertension) in the adult population in USA?
3. What is the hospitalization rate for stroke and acute myocardial infarction?

And the main topic:

Race-wise and Gender-wise mortality rates due to the above 4 mentioned conditions as a measure out of 100,000 population:

Which race has a higher mortality rate due to the above 4 diseases? What is the SMR(Standardized mortality ratio) in the case of each race, when compared to the overall population? How does the mortality look like when we compare men and women?

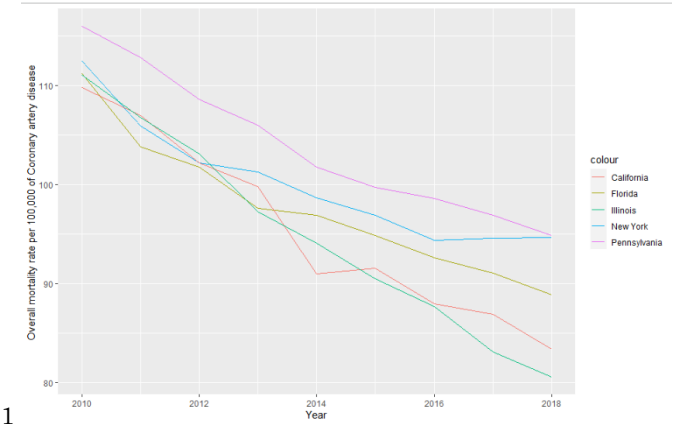
We will also see time series of these conditions over the years 2010-2018

Let us start with a few definitions:

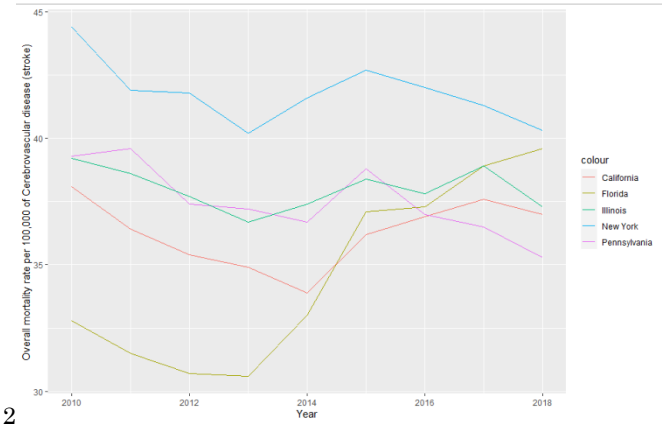
- **Age-adjusted mortality rate**: The difference in the ages of between any two population is normalized. In short, we assume equal age distribution in both the prospective populations. Otherwise, say a community A has 90 % of people aged 65+ while community B has only 10% of people aged 65+. It would not be insightful to compare these two populations directly, and hence age adjustment is needed.
- **Standardized Mortality Ratio**: How likely is a certain section of the population to die from a particular disease when compared to the general population? This likelihood is expressed in terms of a factor n which is calculated using a standard formula (using indirect age adjustment)
- **Prevalence:** It is a fraction that indicates the percentage of the population at risk living with a certain condition or ability. It can also be adjusted for age.

We will start with the main question in bold:

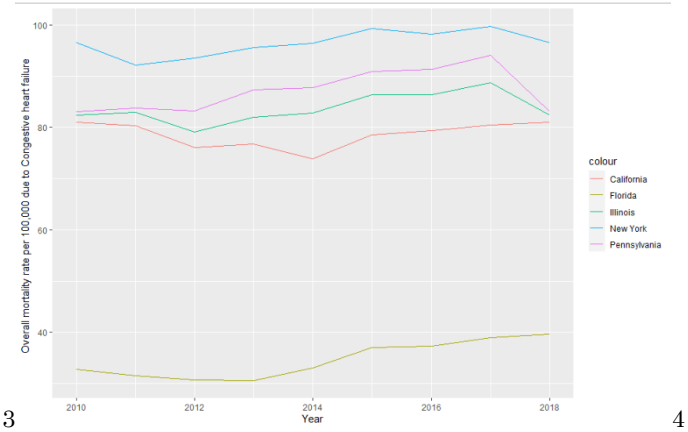
First, we will see the time series for the 5 states for the overall population for all the 4 conditions. We will construct time series only for the years 2010 through 2018 because of inconsistencies in measurements in the last two years.



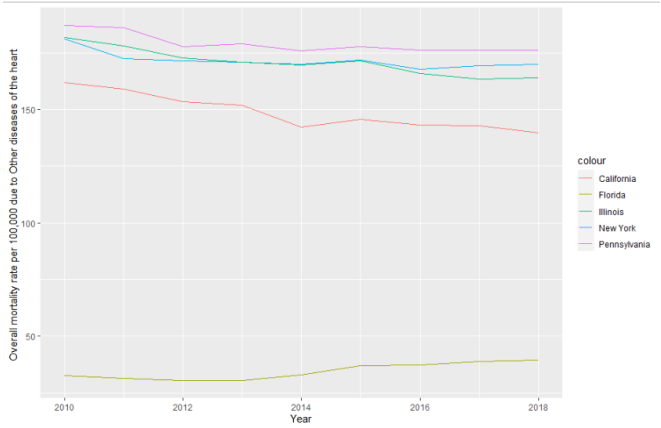
1



2

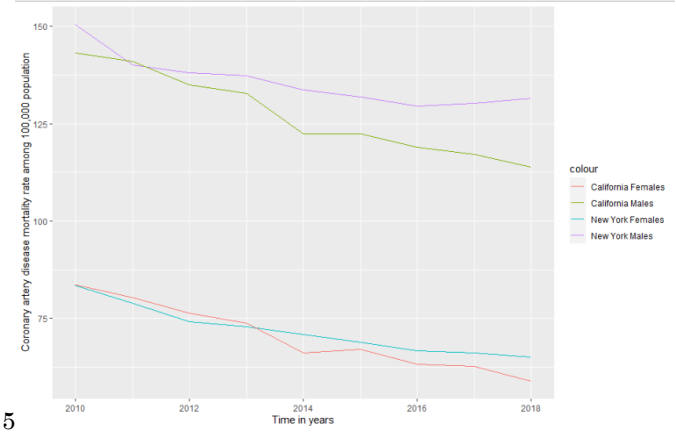


3

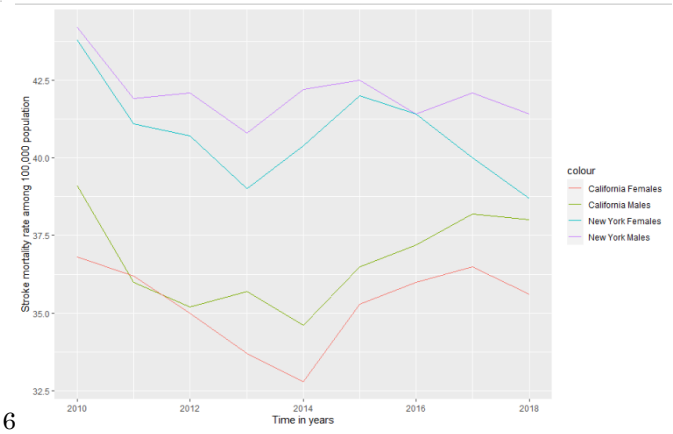


4

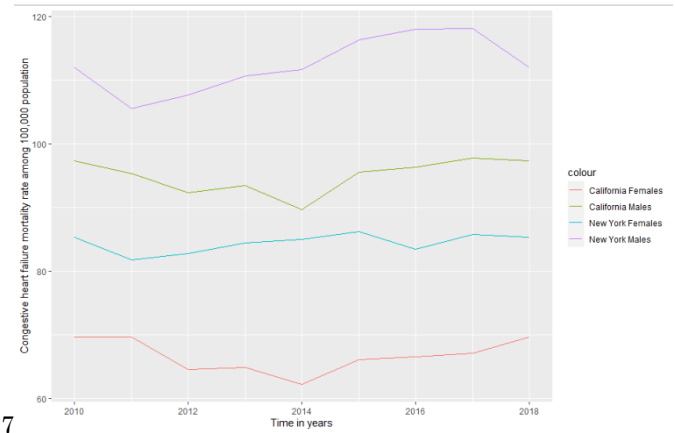
Now, we will fix two states (say New York and California) and look at the distribution between MALE and FEMALE subjects of each of these conditions



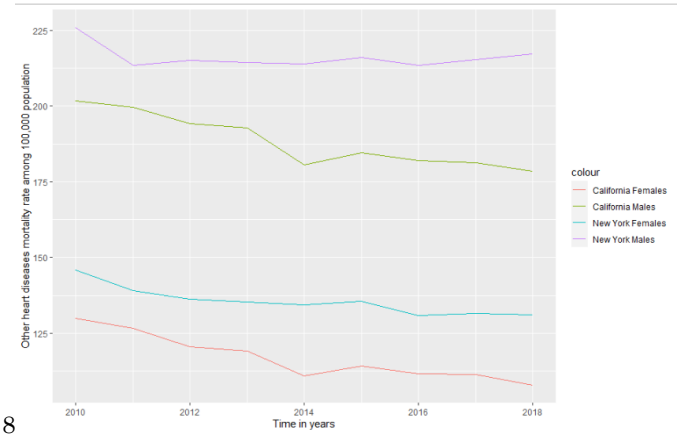
5



6



7



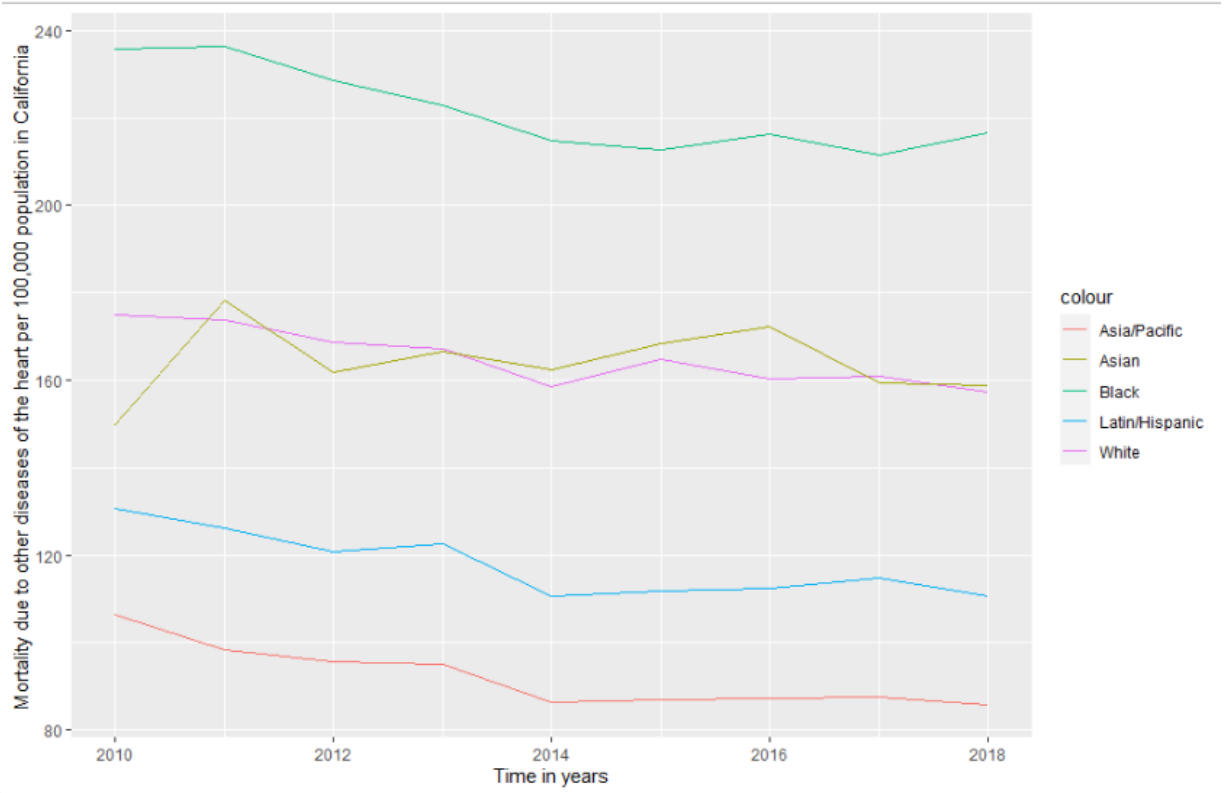
8

From the above graphs, excluding the anomaly figure 6, we can see that males are at higher risk of dying from any of the four mentioned conditions than females.

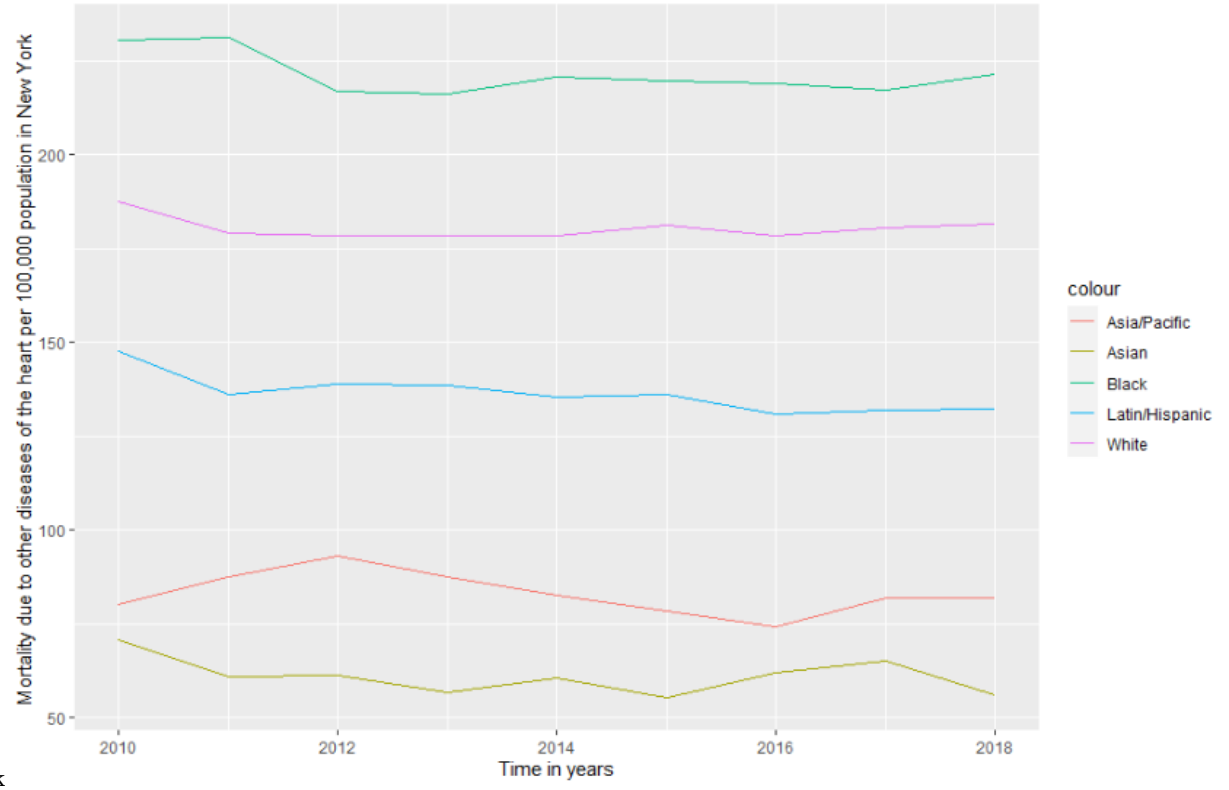
Let us choose the state with the highest population (California) and the most diverse population (New York) and compare the mortalities in different racial populations for **Congestive heart failure** and **Other diseases of the heart**. A similar type of analysis can be done for the whole of USA and if the data is given, for states in India and for India as a whole.

We will also calculate the SMR.

Other diseases of the heart:



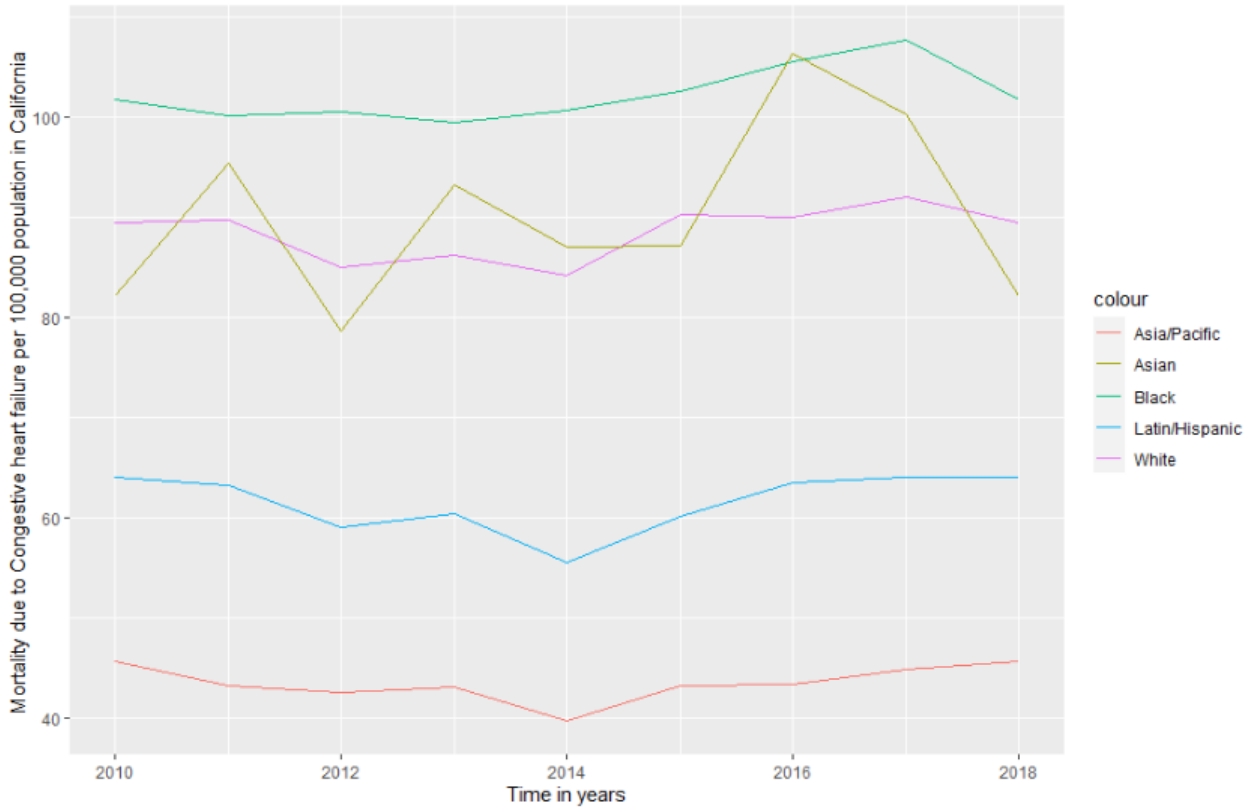
9 California



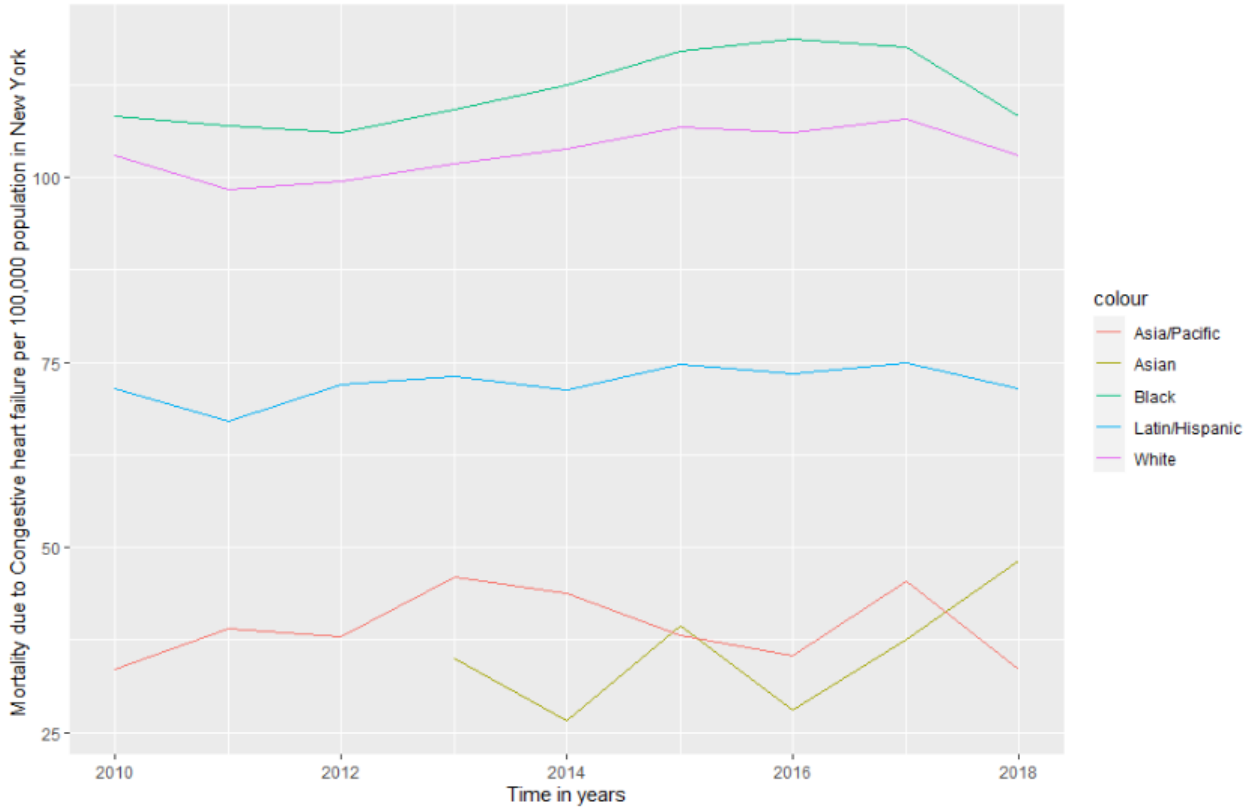
10 New York

Congestive heart failure: It is the weakening of the heart muscles owing to pathological reasons or age-related reasons. Symptoms include but are not limited to oedema (fluid collection in the lower extremities), fatigue and shortness of breath. Often, a differential diagnosis is severe anaemia.

11 California



12 New York



Interestingly, we can see an anomaly in the Asian population in New York and California. It is noteworthy that Indians have been included in the category of Asians, and American-Indians have been shown to be much more at risk for heart failure when compared to the white population.

However, the black population take the #1 spot at risk for all kinds of heart diseases.

The SMR vector in California for Congestive heart failure is calculated for all the races to be:

SMR (averaged for all the years is calculated)

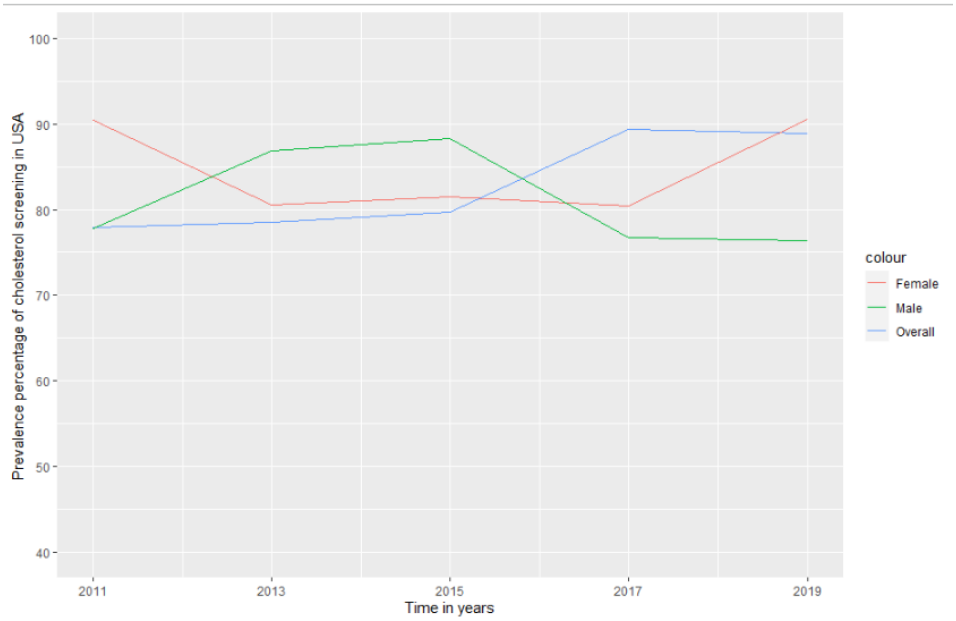
= (HispanicMortality, WhiteMortality, **BlackMortality**, AsianMortality, PacificMortality)÷Overall Mortality

= (0.78 , 1.13 , **1. 30** , 1.15, 0.56)

We will now turn our attention to the three questions we asked.

1) How common is cholesterol screening among adults (18+)? And what is the high cholesterol prevalence ?

We will consider first the cholesterol screening in the all of the states combined among only males and females. (because separate racewise data is not given)

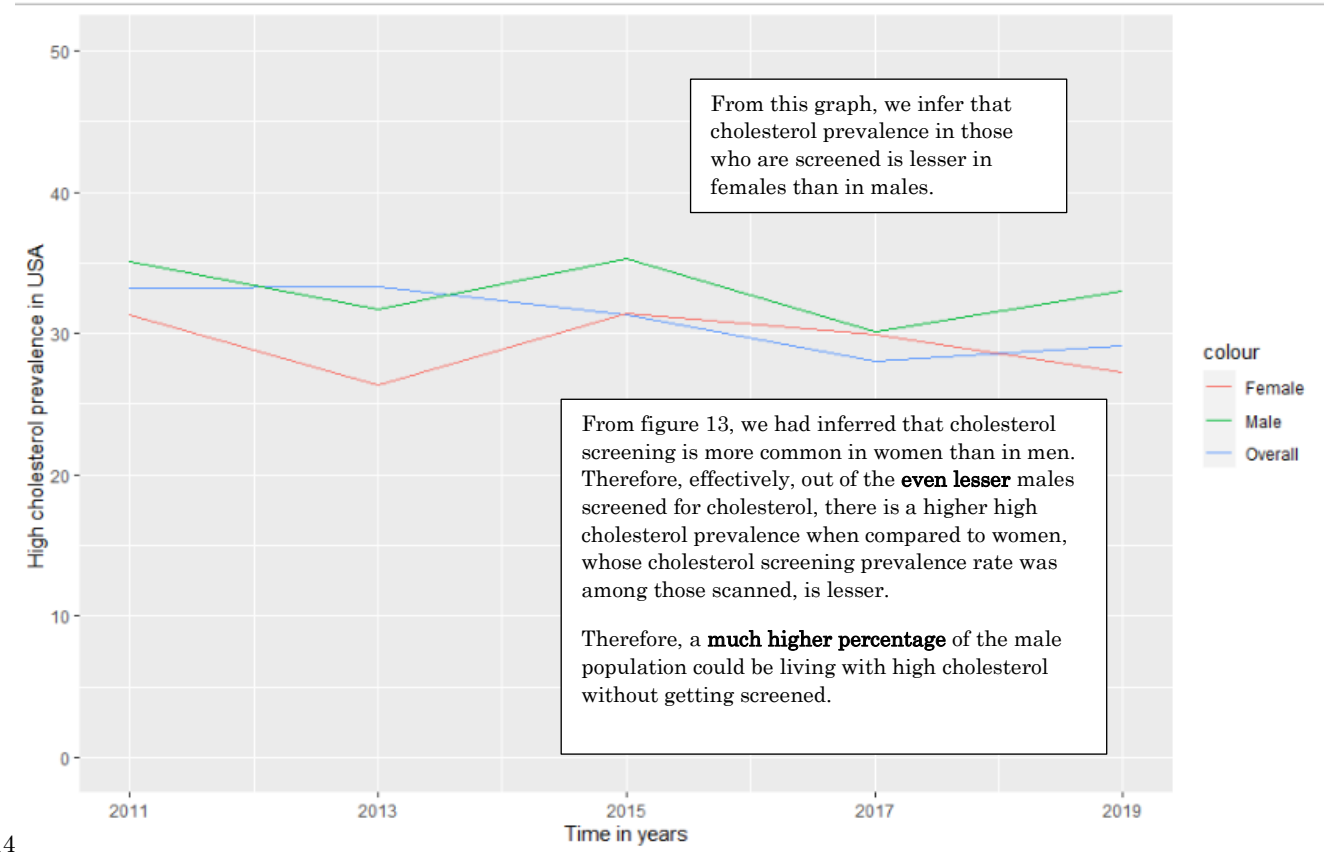


It is clear that the overall screening rate for cholesterol has been increasing possibly because of awareness of the dangers of high cholesterol in adults (18+).

An interesting observation is that male age-adjusted cholesterol prevalence seems to be decreasing while the corresponding numbers for females is increasing. Overall numbers are increasing.

13

Now, we consider the **gender specific** high cholesterol prevalence = $\frac{\text{The number of people at risk screened positive for high cholesterol}}{\text{Total number of people at risk screened}}$



From this graph, we infer that cholesterol prevalence in those who are screened is lesser in females than in males.

From figure 13, we had inferred that cholesterol screening is more common in women than in men. Therefore, effectively, out of the **even lesser** males screened for cholesterol, there is a higher high cholesterol prevalence when compared to women, whose cholesterol screening prevalence rate was among those scanned, is lesser.

Therefore, a **much higher percentage** of the male population could be living with high cholesterol without getting screened.

14

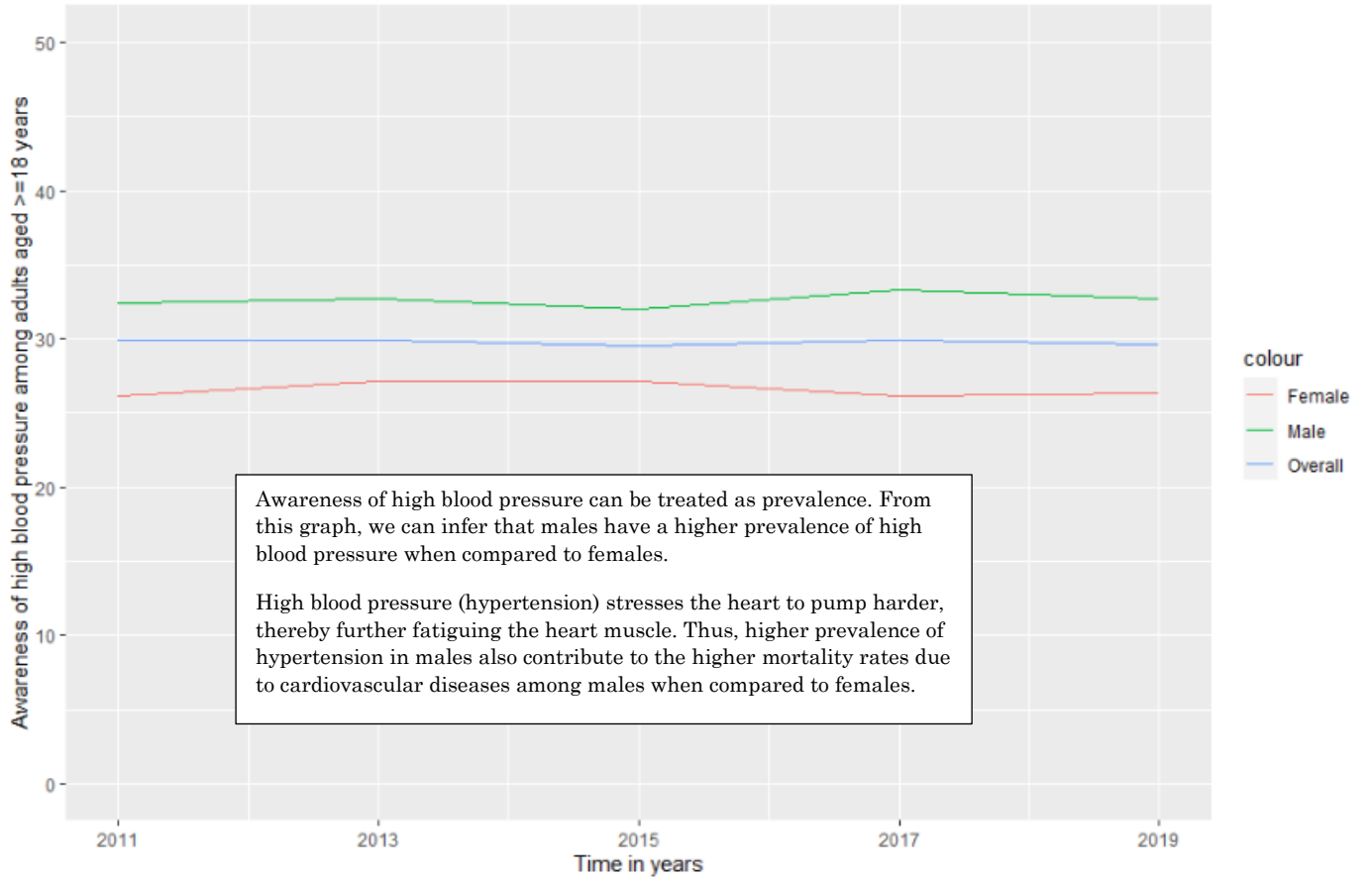
Furthermore, the male population is slightly **higher** in the USA as compared to the female population. This difference is still significant considering the enormous population of the USA.

Earlier, from figure 5,7, and 8, we had seen that males had a significantly higher mortality rate due to **coronary artery disease**, **congestive heart failure** and **other diseases of the heart.** High cholesterol levels are the **basic causing factor** of coronary artery disease. There are "plaques" of cholesterol that are formed along vital arteries which block blood flow, thereby requiring the heart to pump much harder. These plaques continue to build up over time and constrict arteries. This is called coronary artery disease. If an artery which pumps blood to the heart is blocked, it will lead to insufficient blood being circulated with the heart muscle, and therefore goes on to cause **congestive heart failure** and **ischemic heart disease**, which falls under the other diseases of the heart.

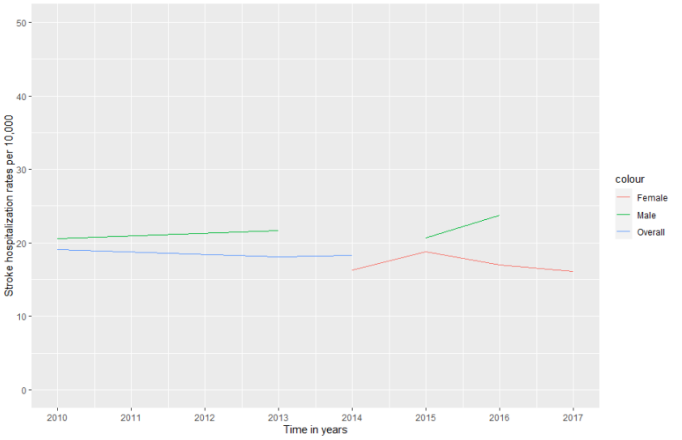
The disparity in cholesterol screening may as well be an important factor that is causing higher mortality rates in males than in females due to cardiovascular diseases.

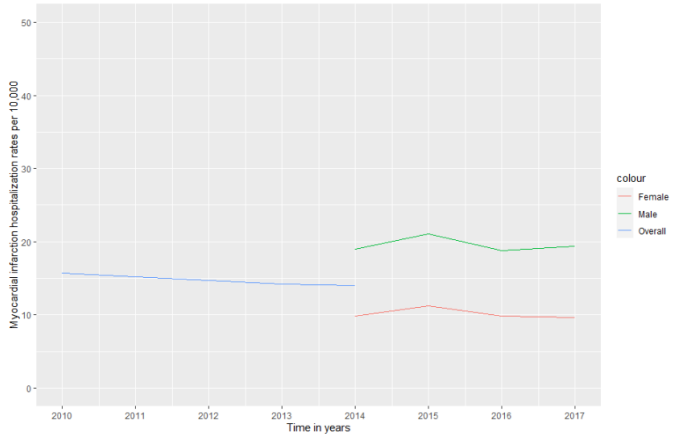
Now, let us consider another risk factor: **High blood pressure**.

15



Finally, we will consider hospitalization rates due to cerebrovascular disease(stroke) and myocardial infarction, and analyse their risk factors among males and females in the state of **New York** per 10,000 population.





Here too, available data suggest that males have higher hospitalization rates when compared to females. We tried analysing race-wise hospitalization rates but we didn't find much difference.

Our methods:

- All coding and analysis in R, codes attached here: (we varied one variable to get answers for the rest)
- Code has been attached in a separate .txt file.

Noteworthy points:

- The alpha numeric coding of the US states aided us a lot to create time series for all the states of mortality due to all the four conditions
- The coding of the categories as: : {"OVR" "GENF" "GENM" "HIS" "WHT" "BLK" "AIAN" "API" } helped us analyze the data easier by making accessibility better
- There were a lot of "NA" values which led me to restrict the data sets significantly, and there were still further "NA"s in the hospitalization due to stroke and MI.

In principle, I got 11 lists with element with each element representing the statistics for a given state each for a given year.
This enabled me to access any statistic of any state of any race at any given year. Then, I varied the year to construct a time series.
This can be done similarly for all the states, and for all conditions. I have only considered the **Age adjusted** rates to eliminated the differences in the age compositions of the states, and therefore is the more accurate measure when compared to crude rates.

However, we had to consider a US dataset due to the unavailability of an Indian dataset. India should start making it's hospital data more transparent by developing newer E-H-R standards and phasing out the outdated. This is so that information and knowledge can be more freely harvested by the general population and so that insightful inferences can be made. While doing so, a standard format (like the aforementioned for alphanumeric coding, coding of categories) should be framed so that the data because readily accessible.

The rate of cardiovascular disease has been slowly creeping up and we must find methods to mitigate it. Awareness of better diet options, **frequent screening,** more accessible health centres with advanced equipment, and natural primary prevention measures like exercising, reducing salt intake should be popularized.

Studies have shown that Indians are more susceptible to cardiovascular disease than Caucasians, who are the main topic of this analysis. So, it becomes all the more important for India to step up cardiovascular care facilities, and along with that, better information systems and management.

Thanks,

Chirag, 18B090003
Ameya, 19B090002

For the endsem project, we plan to analyse statewise cancer statistics which is a huge subset of the already big data, and will try to compare cancer prevalence rates and mortality rates of different states. We will also consider the geographical aspect (using the GPS location given in the data) and try to apply a primitive learning algorithm to determine the "clustering areas" of high cancer rates , and will try to draw insights and explain them.