* **The provenance of the data:**

<https://data.cityofnewyork.us/d/jb7j-dtam?category=Health&view_name=New-York-City-Leading-Causes-of-Death>

The original dataset used for this research was from the New York City Open Data hub provided by the City of New York as a public source of large datasets about the city. From this we were able to access an exhaustive list of the causes of death in New York City from 2007 until 2014. Since the data comes from the New York State OpenSource data hub which is provided by the government in New York City, it is reasonable to argue that the provenance of this data is reputable. However, there could, perhaps be potential biases. Because the data comes from the department of health, there could be reason to believe that when accounting for causes of death, there are more accounts of natural deaths/deaths from disease than unnatural/accidental deaths. The data is likely mostly from hospital records and mortician records which are more likely to record deaths from natural causes explicitly, while deaths recorded from less reputable sources are more likely to be identified as “unknown cause” or “other cause”.

* **The data model we used:**

We chose to represent our data using bar graphs to explicitly show the differences in the death rates between men and women based on natural/unnatural causes of death. We used bar graphs because in this way we are able to show the difference between the sexes in our data collection as well as some change over time by placing the bar graphs next to one another chronologically. By using the bar graphs to model our data, one can explicitly see not only the explicit difference in death rates in men and women by our categorization, but also the effect of time, should there be any.

* **The curation of the data you carried out and the choices made:**

The goal of our research was mainly to see if there is a significant difference in the death rate by natural and unnatural causes (i.e. accident, injury, self-harm, etc.) in men and women in New York City given our timeframe (note: we used these years due to the limitations of our dataset only providing death rates from the years 2007-2014). Gender disparity in mortality rates as a point of research are much greater than expected. Moreover, as a result of initial research, we learned that not only are men more likely than women to die of injury, but that due to women being more likely predisposed to health conditions, they are also more likely to die from natural causes than men.

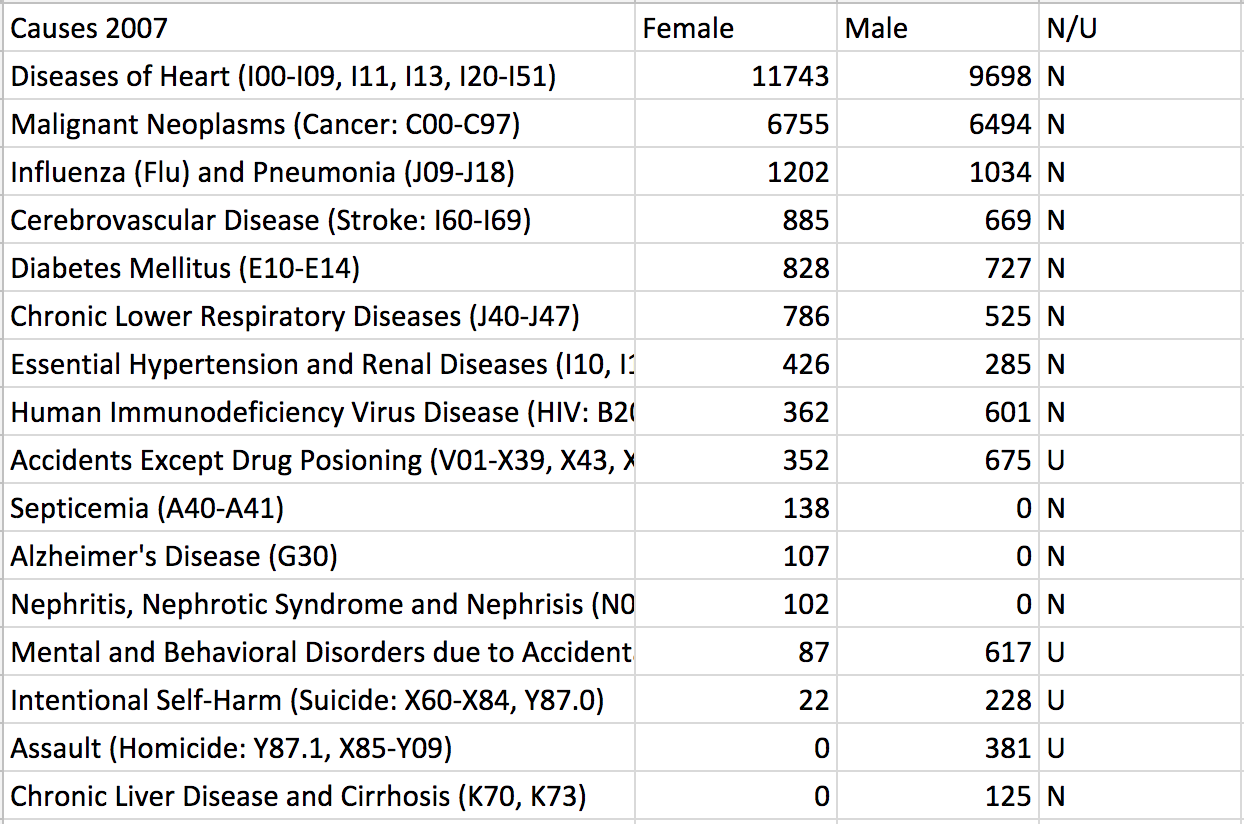
From the original dataset the steps taken to curate the dataset for our needs are as follows:

1. Using Microsoft excel, the data for the years 2007, 2010, and 2014 were isolated respectively.

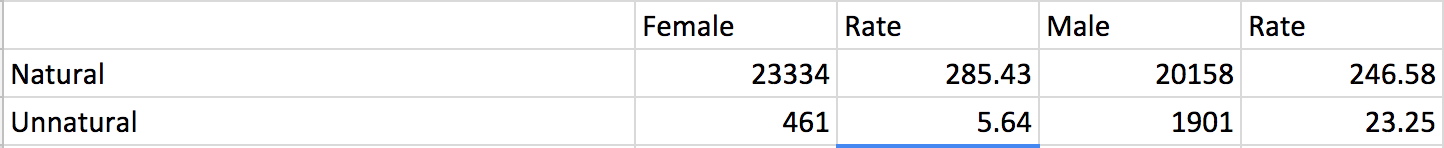
2. In the original dataset, the causes of death and total deaths were separated by racial background, making it difficult to see the total number of deaths by gender only, as we were aiming to achieve. So, in order to get total deaths for each cause of death, all of the racial groups per cause of death were grouped together to get a total number of deaths per sex for each cause rather than a total number of deaths per race for each cause.  


This was done using the SUMIFS function in excel. After removing the “Race Ethnicity” column, SUMIFS was called per leading cause to sum the total number of deaths based on if the sex was “M” or “F”. If there was a “.” in place of a number for “Deaths”, it was replaced with a 0, assuming there was no significant data for the number of deaths for this ethnicity.

The resulting curated data from this step is three charts depicting the cause of death for that year, the total number of deaths by cause per gender, and a column denoting if the cause of death was natural or unnatural, denoted by a “N” or “U” respectively.

A similar chart as depicted below is given for 2007, 2010, and 2014.

3. From this point, the SUMIFS function was used again to total the deaths from the chart based on if they were natural (“N”) or unnatural (“U”) for each respective gender in each year.

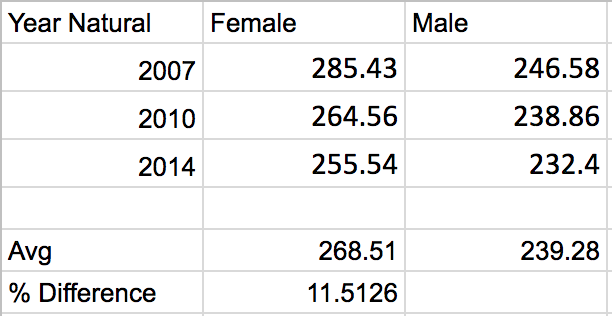


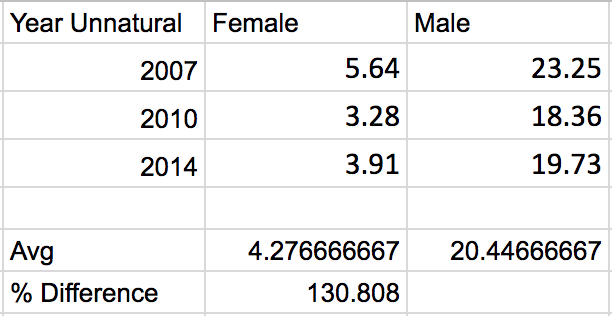
Since our origional dataset did not provide us with reasoable death rates, death rates per year were calulated using this formula:

***(D/P)\*100,000***

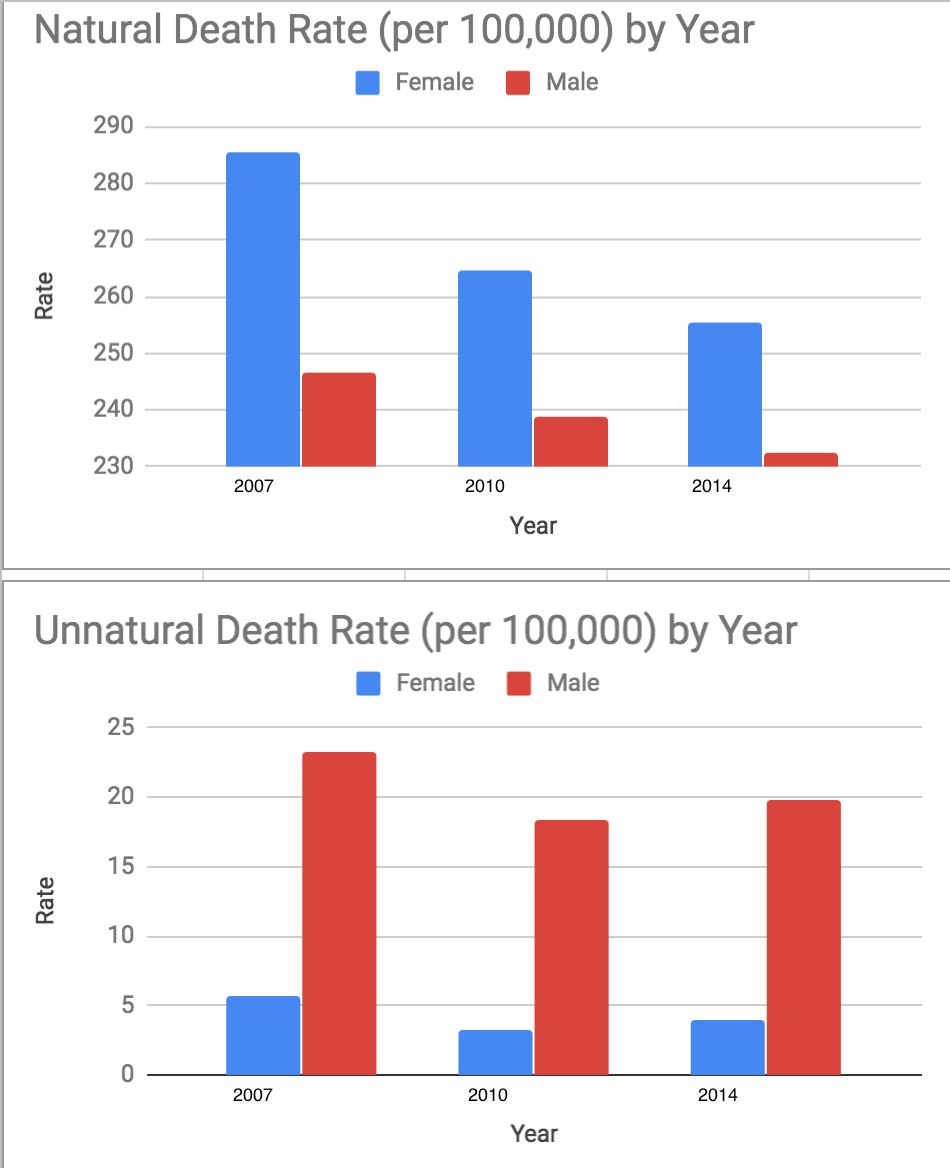
Where D is the total number of deaths given gender and nature of death, P is the total population of New York City (taken from the new York census of 2010), and then multiplying this by 100,000 as a standardized number to record mortality rates. E.g. if the death rate is 285.43, this can be read as 285.43 deaths per 100,000 people in the given population.

4. Finally, the results from the three charts were consolidated into two charts to show, the death rates of men and women from natural and unnatural deaths by year. Additionally, the average death rate of the three years by sex were calculated as well as a % Difference in the death rates in order to fully show the disparity between the sexes.





These two charts were used in order to depict the discrepancy between the sexes in unnatural deaths and natural deaths using bar charts.



* **The steps you took to annotate or enrich the data:**

The original dataset is comprised of data for the number of deaths in New York City from 2007 to 2014 categorized by sex and race. It provides raw numbers of deaths as well as death rates based on these categorizations. This is problematic of our dataset in respect to the results we are looking for (death rates based on sex *only*). Additionally, because the dataset did not provide a total population number, or any information on how the death rates were derived/which standards were used, it was necessary for us to derive our own death rates based on our categorization of deaths. Moreover, data from the New York City census (<https://www1.nyc.gov/assets/planning/download/pdf/data-maps/nyc-population/census2010/pgrhc.pdf>) for the population of the city from 2010 was used to derive death rates from our raw data. We did this because 2010 is the closest census year to all of the years we use in our data. We enriched the data in order to obtain the results we needed for our research by categorizing the types of deaths as natural and unnatural and deriving the rates of death from these totals.

* **The tools used:**

Microsoft Excel:

* 1. The dataset was sorted by year
  2. Data per year was cut and pasted into separate spreadsheets within the same project
  3. SUMIF function was used to sum the total deaths of each cause based on sex
  4. Charts function (based on selected data i.e “Recommended Charts”) used to create bar graph data model.