**Capstone Project I – Data Story**

Data is obtained from United States Cancer Statistics, 1999-2013 provided at the Centers for Disease Control and Prevention.

<https://wonder.cdc.gov/CancerMort-v2013.html>

The data included various demographic features like cancer sites, states, year, sex, age, race and represents the population, deaths and crude rate.

Various features are:

Cancer Site: It is the type of cancer that person had.

State: It represents one of the 51 states of USA.

Year: It represents the year in which cancer was reported.

Sex: It represents the sex of the person.

Age: It represents the age of the person.

Race: It represents the race of the person.

Population: It represents the total population of the category.

Deaths: It represents the number of deaths for a category.

Crude Rate: It represents the number of cancer cases that were reported to the hospitals for a particular category.

The data was cleaned to remove various NaN values like None, Not Applicable, or Suppressed values.

In order to explore the data further various questions were asked and visualizations were made to get insights into the data.

1. **A. Is Deaths, Crude Rates or Population related to each other**: In order to understand whether Deaths, Crude Rates and Population were related to each other, I started visualizing the distribution of Deaths, Crude Rates and Population over different features.

**Results**:

* + It was found that for most of the States Population follow similar trend as Deaths, however Crude Rates were different from both population and Deaths for most of the states.
  + It was found that for most of the Cancer Sites Deaths follow similar trend as Crude rates, however Population was very different.
  + It was very interesting to see the Age group and its relation with Population, Deaths and Crude Rates. Deaths and Crude rate showed inverse relation to the Population in most of the cases.
  + Also looking at the race, it was particularly interesting to look at Black and African Americans, because they do not comprise a huge percent of of population however their crude rate was pretty high and their deaths was much lower. It kind of suggests that lot of black and African Americans who report for cancer are mostly cured and do not face death as compared to other races.
  + Also it was interesting to look at male and female distribution of deaths, crude rates and population. It was observed that though crude rate of females is much lower than males, their deaths were similar to males, suggesting that most of females that die due to cancer are not diagnosed or treated as cancer patients. Indicating that detection rate of cancer is much lower in females as compared to males.

**B. Relation between Deaths/Crude Rate with population:** To further explore the relation between Deaths and Crude rate with Population regression plots were visualized.

**Results:**

* + It was found that for States, there was a positive correlation of Deaths/Crude rate with Population.
  + It was found that for Cancer Sites, there was no correlation of Population with Crude rate or Death. However, Deaths was correlated with Crude rates.
  + Similar to Cancer sites, Age group also showed no correlation of Population with Crude rate or Death . However, Deaths was correlated with Crude rates.
  + Also looking at the race, they did not show very strong correlation of population with crude rate or Death.
  + Interestingly for Sex, there was no correlation between any of the two variables.

1. **Which are the States that have highest deaths due to cancer:** Next, I wanted to see geographical distribution of deaths due to cancer and find out which state results in maximum deaths due to cancer.

**Results:**

* + It was found that California resulted in maximum number of deaths due to cancer.
  + Florida was found to result in second highest in deaths.
  + New York was the third highest in number of deaths due to cancer

Not surprisingly, the states with highest population were found to have highest number of deaths due to cancer. Therefore, these results could be biased due to population aggregation and thus needs to be analysed further.

1. **What type of cancers result in highest deaths:** Next, I wanted to explore the types of cancers and find out which of these cancers result in highest deaths.

**Results:**

* + It was found that Respiratory System related cancers resulted in maximum number of deaths.
  + Digestive system resulted in second highest in number of deaths due to cancer.
  + Male and Female Breasts related cancer was found to result in third highest deaths.

This is an interesting result and will be explored in future to see if cancer site is indeed a true estimate of deaths due to cancer or if it is just a bias due to population or crude rate.

1. **What category results in highest deaths:** Next, I wanted to explore that what category, as in what state, which cancer, what age group, what race and sex, results in highest deaths.

**Results:**

* + It was found that Digestive system cancer in California among 85+ years old white females resulted in maximum number of deaths due to cancer in 2011.
  + Next was Digestive system cancer in New York among 85+ years old white females resulted in second maximum number of deaths due to cancer in 2011.
  + Next was Digestive system cancer in Florida among 85+ years old white females resulted in third maximum number of deaths due to cancer in 2011.

1. **How does the distribution of deaths due to different factors like Cancer Sites, States, Race, Age Group change over time**? To see the change in distribution of deaths due to various features, the distributions were visualized over time.

**Results**:

* + It was found that distribution of deaths did not change over the years due to any of the features particularly. Although, a slight noticeable increase was observed for females deaths for all the features particularly in the year of 2011.

1. **How does the mortality change over time for each cancer site:** To see the change in mortality over time due to cancer site, the date was grouped by year and cancer site and visualized.

**Results**:

* + It was very surprising to see that number of deaths didn’t really that much over time, however there was a sudden jump in the number of deaths by certain certain cancers in the year of 2011. This needs to be explored further.

**Conclusions and Future Data Analysis:**

The data shows some very interesting trends in the number of deaths due to cancer.

* + It shows that the number of deaths due to cancer increases with the increase in age, however population shows the opposite trend. This suggests that age could be true indicator of the deaths due to cancer and is not biased by population.
  + It also shows that there is no correlation between deaths and crude rate for States, race and sex, suggesting that these factors influence deaths independent of the crude rate.
  + It was also interesting to see that though Respiratory system related cancer results in highest number of deaths, if other factors like location, sex, and age group are considered then digestive system results in highest mortality due to cancer.
  + The sudden jump in the mortality due to certain cancers was observed in the year of 2011, which needs to be explored further and could be due to some bug in the data.

The interesting insights that we found for the data need to be explored further to determine which features are true indicators of the mortality and crude rate. And to use those indicators to predict the mortality for a cancer patient.