1. The following dataset concerns a government experiment administering a serum to give people superpowers. Please answer the following questions. Submit an ipython notebook and a PDF that answers the questions and shows your work.
    a. How many people got super powers with the serum, vs how many people got them naturally?

       Number of people who got superpowers with the serum: 44902
       Number of people who got superpowers naturally: 44145

       first the naturally means who didn't use serum but got superpowers so [df['SuperSerum'] == 0] could find who got superpowers naturally, otherwise it's got super powers with the serum. The sum could count the times how many people, because all the number are 1 or 0 so could use sum to calculate.

    b. Does the serum cause superpowers? How confident are you?

       To test if the serum causes superpowers, we can use a hypothesis test where the null hypothesis is that there is no difference in the proportion of people who get superpowers with the serum and without the serum. The alternative hypothesis is that the proportion of people who get superpowers with the serum is higher than the proportion without the serum.

       Based on the statistical analysis conducted, the serum appears to cause superpowers. The z-score of 10.154 and the p-value of 0.000 suggest that the difference in the proportion of people who got superpowers with the serum versus those who got it naturally is statistically significant. The confidence interval of (0.016, 0.024) also supports this conclusion. Therefore, with a high degree of confidence, we can say that the serum causes superpowers.

    c. The serum sometimes causes horrible mutations. Who should avoid the serum?

       indicating that the serum was administered. It then groups the resulting dataframe by Gender and Age, where Age is binned into 10-year intervals. Finally, it calculates the proportion of individuals in each group who got a mutation, and sorts the groups in descending order of mutation proportion.

       This code can help to identify which age and gender groups are most likely to experience mutations after receiving the serum. The output of this code will show the mutation proportion for each group, which can be used to determine which groups are most adversely affected by the serum.

From the data, the female is easy to get horrible mutations, and the age is 20 to 30.

2. The following datasets are based on popular media. Each has something obviously strange about it. Identify the way in which the dataset is unusual. For each, please submit an ipython notebook, explaining what is unusual about the dataset, with supporting Pandas code and at least one visual (for example, a histogram or a box and whisker plot). Please also submit a PDF.

    a. A dataset of crime statistics from a particular region.

        From this data, and searching, it could find when the Weapon_Involved is true always is theft, this is the unusual paet and should not happen like that.

    b. A dataset of demographic info

        from the graph in this data, we could find most of the people are below 7.5 feet but some of them are over it, such as 2 years old but have 16 feet, it’s not possible, the data is incorrect for human.

    c. A dataset of information about people's job satisfaction

        The unusual aspect of this dataset is that all of the individuals are unemployed, yet they have reported incomes and job satisfaction levels. This may indicate that the dataset is not accurate or that there was a mistake in recording the data. Additionally, the reported incomes seem to vary greatly, with some individuals reporting zero income while others report incomes in the thousands of dollars.