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 determine whether the serum causes superpowers or not, we can conduct a hypothesis test using the chi-square test for independence. We can set up the following hypotheses:

       Null hypothesis: There is no association between the serum and the superpowers.
       Alternative hypothesis: There is an association between the serum and the superpowers.

       To determine whether the serum causes superpowers or not, we can conduct a hypothesis test using chi-square test for independence. We can set up the following hypotheses: Null hypothesis: There is no association between the serum and the superpowers. Alternative hypothesis: There is an association between the serum and the superpowers.

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

       As we can see from the graph, individuals aged 20- 25 are more likely to experience mutations from the Super Serum, whereas those aged 1-19 and over 26 are less likely to experience mutations. So the people who are 20 to 25 years old need to be careful

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.