Question 1:

I was a part of a project which was hosted on Kaggle titled “New York Taxi Fare prediction”. This dataset had a few features like datetime, pickup location, drop-off location both specified in co-ordinates (latitudes and longitudes) and the passenger count. The challenge was to find the most busy and free times of the days with respect to hour of the day. The next challenge was to find the average revenues earned over a month within the given year. We had to calculate the distance of travel using the latitude and longitude details given for the pickup and drop-off locations.

I had to do data cleaning on the dataset in order to remove the null cells as well as to change a few cells which had null location coordinates. This had to be done in order to ensure that I was using only the necessary and relevant portions of the data rather than all the non-essential data which could lead to a serious decrease in the accuracy of the trained model. Since the dataset had just one file, this was a single source problem of data cleaning. Next, I had to do data transformation in order to convert the latitude and longitude coordinates to distance and split the datetime into various fields for date, by year, month, date, hours, and minutes. I used python pandas in order to manipulate and wrangle the data. I stopped wrangling of the dataset when the constraints of the dataset questions were met.

Question 2:

A data quality consists of Completeness, Accuracy, Uniqueness, Uniformity, Timeliness and Security. I found the above data to be unstructured. The biggest quality issue that I could pick out from the dataset was that gender was assumed based on the names. If we consider a case of unisexual names, the software fails to classify the gender assignment properly. When we consider the projects conducted by the scientists and the accolades being presented to them, we can see that a lot of contributors are not recognized properly. Also, another important issue that I could see here is that a lot of letter correspondence have not been noted properly as per todays standards.