Project Proposal

1. Introduction:

The project is meant to educate in the basics of data analysis and Python/Jupyter Notebook and Pandas syntax. Large quantity of data will be imported from a publicly available location and subject to processing using the mentioned technologies. The data acquisition as well as the processing are subject to constraints outlined by the tutor.

1. Objective:

Apart from being in line with the points and constraints outlined in the introduction, the objective of this specific project will be to analyze the correlation of the most commonly committed crime type with the time of the year in which they occur on.

1. Hypothesis / claim:

Outdoor robbery and aggravated assault increase in numbers of occurrence during the spring months, that is April, May and June.

1. Methodology:

The data will be acquired from a publicly available location. It will then be cleaned up programmatically by removing the rows with missing information. This will be achieved using Pandas dropna() method. Further, for the purposes of analysis the inconsistent data formats will be standardized so that it may be used in a programmatic fashion. Data will be subject to analysis by visualizing that will help establish what are the most commonly commited crime types, then what month sees the highest crime rate to finally pointing out the crime composition of a chosen crime type within months of a year.

1. Data gathered for this project was collected from the below website, which was proposed in pdf instruction file: <https://data.baltimorecity.gov/Public-Safety/BPD-Part-1-Victim-Based-Crime-Data/wsfq-mvij/data>.
2. The data for the project will be included in the csv format file. The table will contain about 10000 rows and 12 columns. The volume of the data will be less than 500KB and file will get executed in less than 5 minutes as per constraint.
3. Plan for the project:
   * Data for the analysis concern Victim Based Crime in Baltimore

* Data gathered from the above source will be uploaded and programmatically cleaned up and saved to new file
* The modified data will be next analyzed and visually represented (histogram, line chart, pie). All visualizations will be generated programmatically. Below each visualization summary and conclusions will be written. The analysis will concern:
  + What crime is the most commonly committed
  + Total committed crimes monthly
  + The most commonly committed crime per month
* Lastly project will be summarized

1. The alternative source for the data collection will be <https://wonder.cdc.gov/cancer.html>, which contains United States Cancer Statistics and Data. The source comes from the pdf instruction file.
2. Alternative plan for the project:
   * Data for the analysis concern US and Puerto Rico Cancer statistics

* Data gathered from the above source will be uploaded and programmatically cleaned up and saved to new file
* The modified data will be next analyzed and visually represented (histogram, line chart, pie). All visualizations will be generated programmatically. Below each visualization summary and conclusions will be written. The analysis will concern:
  + Number of cancer deaths by country
  + Number of cancer deaths by type
  + Number of cancer deaths by age
* Lastly project will be summarized

1. Summary:

The project will conclude in a statement that will show a correlation to the most commonly committed crime type and the month / time of a year they’re committed on, showing a correlation between the two. The lessons learned through the process of acquiring, processing and analyzing data will also be summarized within the final Jupyter Notebook.