**Capstone Project (Car-accident analysis)**

**Introduction/Business Problem Section**

* **Background:**

**In this Capstone Project, we will use the**

* **Dataset of the All collisions (car-accidents - 2004 to Present.) in Seattle**
* **the Metadata of the dataset**

**provided by SPD and recorded by Traffic Records (seattle.gov) to understand the Car Accident occurred in Seattle.**

**The understanding of the data includes studying the factors such as**

* **the accident severity, the accident type, the details of the accident,**
* **the Transportation (e.g. car / bicycles) and how many people and who involved in the accident**
* **the weather, road condition and environmental factors at the point-in-time of the accident**
* **the location of the accident, the surrounding physical environment of the location accidents, etc.**

**After the understanding, we will start analyzing the accident by using multiple data analytic tools and skills (e.g. pandas) learned in this course,**

**to understand the relation between the factors (e.g. correlation, means, frequency, max, min. etc.)**

**After the data analysis, we will try to**

* **1. get insights from the data**
* **2. make hypothesis by evaluating the relation of the data in the dataset**
* **3. testing the effectiveness / validity of the hypothesis**
* **4. draw conclusion and make forecast**

**And then using the dataset (diving into training data vs sample data) to simulate and test the effectiveness hypothesis (model).**

**Finally, we will write a report and resent the observation in a PPT.**

**During the project, we will use GitHub Account to upload the changes of the Jupyter notebook of The Project to GitHub to ensure all changes are recorded.**

* **A description of the problem**

**As descripted in the background section, we aim at using the dataset & metadata to**

* **1. draw conclusion**
* **2. make prediction**
* **3. represent the observation, and**
* **4. make practical and meaningful suggestions to the settle gov,**

**and hence to reduce the**

* **1. frequency and**
* **2. severity of the accident in future.**

**In conclusion, we hope to give recommendation to the gov. to make remediation / correction / education at the potential problematic factors (black spots) in order to prevent / reduce any car accidents in future. And the most importantly, to SAVE LIFE!**

**Data Section**

* **A description of the data**

**Through studying the dataset and the meta data, we understand that the following data is are useful,**

* **'REPORTNO', Remarks: new data, some incidents have the same report no. therefore, we keep this data to find duplication.**
* **'STATUS', Remarks: new data, the status has two type, matched (~188k), unmatched (4800), it helps us to filter the potential incorrect / inaccurate / incomplete data. Since the total no. of unmatched is less than 1 % of the total data. Filtering out the data is acceptable and it is insignificant to the result / conclusion of the data analysis result**
* **'SEVERITYCODE', A code that corresponds to the severity of the collision:**
  + **3—fatality**
  + **2b—serious injury**
  + **2—injury**
  + **1—prop damage**
  + **0—unknown**
* **'ADDRTYPE', Collision address type:**
  + **Alley**
  + **Block**
  + **Intersection**
* **'INTKEY', Key that corresponds to the intersection associated with a collision**
* **'LOCATION’, Description of the general location of the collision**
* **'SEVERITYDESC’, A detailed description of the severity of the collision**
* **'COLLISIONTYPE’, Collision type**
* **'PERSONCOUNT’, The total number of people involved in the collision**
* **'PEDCOUNT’, The number of pedestrians involved in the collision. This is entered by the state.**
* **'PEDCYLCOUNT', The number of bicycles involved in the collision. This is entered by the state.**
* **'VEHCOUNT’, The number of vehicles involved in the collision. This is entered by the state.**
* **'INCDATE', The date of the incident.**
* **'INCDTTM’, The date and time of the incident.**
* **'‘, The of junction at which collision took place**
* **'SDOT\_COLCODE', A code given to the collision by SDOT.**
* **'SDOT\_COLDESC', A description of the collision corresponding to the collision codes.**
* **'INATTENTIONIND', Whether or not collision was due to inattention. (Y/N)**
* **'codes or not a driver involved was under the influence of drugs or alcohol.**
* **'WEATHER', A description of the weather conditions during the time of the collision**
* **'ROADCOND', The condition of the road during the collision.**
* **'LIGHTCOND’, The light conditions during the collision.**
* **'PEDROWNOTGRNT', Whether or not the pedestrian right of way was not granted. (Y/N)**
* **'SPEEDING', Whether or not speeding was a factor in the collision. (Y/N)**
* **'ST\_COLCODE’, A code provided by the state that describes the collision.**
* **'ST\_COLDESC', A description that corresponds to the state’s coding designation.**
* **'SEGLANEKEY', A key for the lane segment in which the collision occurred.**
* **'CROSSWALKKEY', A key for the crosswalk at which the collision occurred.**
* **'HITPARKEDCAR’, Whether or not the collision involved hitting a parked car. (Y/N)**

**And the data could be categorized into the following groups:**

* **1.the accident severity, the accident type, the details of the accident,**
* **2.the Transportation (e.g. car / bicycles) and how many people and who involved in the accident**
* **3.the weather, road condition and environmental factors at the point-in-time of the accident**
* **4. the location of the accident, the surrounding physical environment of the location accidents, etc.**
* **5. - irrelevant / duplicated for the capstone project (could be drop)**
  + **'X','Y','SEVERITYCODE.1','OBJECTID','INCKEY','COLDETKEY','EXCEPTRSNCODE','EXCEPTRSNDESC'**
  + **(Potential irrelevant data - 'SDOTCOLNUM', A number given to the collision by SDOT.)**
* **How it will be used to solve the problem**

**By using the dataset, data analysis tools and skills, we will try to**

**1. get insights from the data**

**2. make hypothesis by evaluating the relation of the data in the dataset**

**3. testing the effectiveness / validity of the hypothesis**

**4. draw conclusion and make forecast**

**Though the following 3 steps approach to study the data**

**1. get insights from using individual data**

**2. study the trends of the data study the trends of the data over the years**

**3. get insights from using two potentially data to find the correlation**