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

**Date: Oct 2020 By: Alex Wong**

**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.**
* **Who would be interested in this project?**
* **The Government official of Seattle, such as Transport department traffic control team, policy station, accident investigation team**
* **The city planning department, the education department (e.g. no speeding / driving after drinking)**
* **Education sectors which investigate mainly the city planning, traffic accident, transportation**
* **Conclusion of this section and also this Capstone Project**

**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 where you describe the data that will be used to solve the problem and the source of the data**

* **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.**
* **'‘JUNCTIONTYPE '‘, 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)**
* **‘UNDERINFL’, Whether 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.)**
* **Source of the data**

**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.**

* **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**

**Methodology section which represents the main component of the report where you discuss and describe any exploratory data analysis that you did, any inferential statistical testing that you performed**

* **Discuss and describe any exploratory data analysis**

**4-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. **Graphic**

**Before, performing any data analysis, I selected the following set of data for analysis.**

1. **'COLLISIONTYPE’, Collision type**
2. **'ADDRTYPE', Collision address type:**
   1. **Alley**
   2. **Block**
   3. **Intersection**
3. **'WEATHER', A description of the weather conditions during the time of the collision**
4. **'INCDATE', The date of the incident.**
5. **With a focus on the date (e.g. Friday / sat / sun)**
6. **'INCDTTM’, The date and time of the incident.**
7. **With a focus on the time (e.g. 12-5am)**
8. **'SPEEDING'**
9. **'ROADCOND', The condition of the road during the collision.**
10. **'LIGHTCOND’, The light conditions during the collision.**
11. **'INATTENTIONIND', Whether or not collision was due to inattention. (Y/N)**
12. **‘UNDERINFL’, Whether or not a driver involved was under the influence of drugs or alcohol.**
13. **'‘JUNCTIONTYPE '‘, The of junction at which collision took place**
14. **'ST\_COLCODE’, A code provided by the state that describes the collision.**
15. **+ 'ST\_COLDESC', A description that corresponds to the state’s coding**
16. **INTKEY+LOCATION, for the black spots**

**With the support of other data in the data set, I tried to enhance the data analysis result**

* **Any inferential statistical testing that you performed**

1. **Using two potentially data to find the correlation**

**I have also selected the following pairs of data to analysis the correlation.**

1. **INCDTTM’, The date and time of the incident.**
   1. **e.g. Friday night 1200 pm.**
2. **UNDERINFL’ VS 'SPEEDING'**
3. **UNDERINFL’vs 'PERSONCOUNT’**
4. **INATTENTIONIND' vs 'HITPARKEDCAR’**
5. **WEATHER' vs 'VEHCOUNT’**
6. **ROADCOND' vs 'PERSONCOUNT’**
7. **'LIGHTCOND vs 'VEHCOUNT’**

**With the support of other data in the data set, I tried to enhance the data analysis result**

* ***What machine learnings were used and why.***

1. ***N/A, since it is the “Applied Data Science Specialization”, Machine learnings, was not included, in this specialization, therefore, we will not use any ML model in this capstone project.***

**Results section where you discuss the results.**

**4-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. **Graphic**
4. **Using two potentially data to find the correlation**

***Remark: Before analyzing the data / test case, we firstly filter the status - “unmatched” data, since the data is suspected as incomplete / inaccurate.***

**Background Data Analysis:**

* + ***Total row x column =*** (189786, 29) after the 1st initial data preparation

**Test Case 1: driver involved was under the influence of drugs or alcohol *(“drugs or alcohol” or UNDERINF)***

**Test Case 1.1: Get insights from UNDERINF**

**# Total case = 189786**

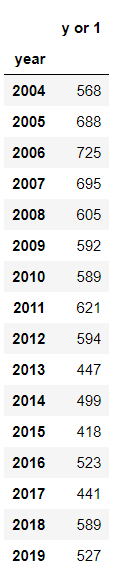
**# Driver involved was under the influence of drugs or alcohol = 9121**

**9121 / 189786**

**# Overall 4.81%**

**Test Case 2.1a: Data Analysis - UNDERINF vs year**

**Accident that the driver involved was under the influence of drugs or alcohol vs. Year**



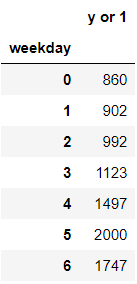
**\* 2020, no any record;**

**\* y or 1, “drugs or alcohol”**

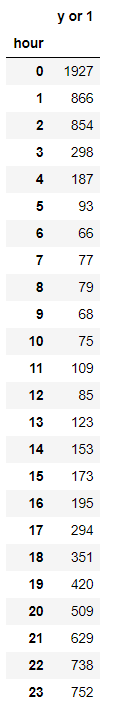
**Test Case 2.1b UNDERINFL vs. month**



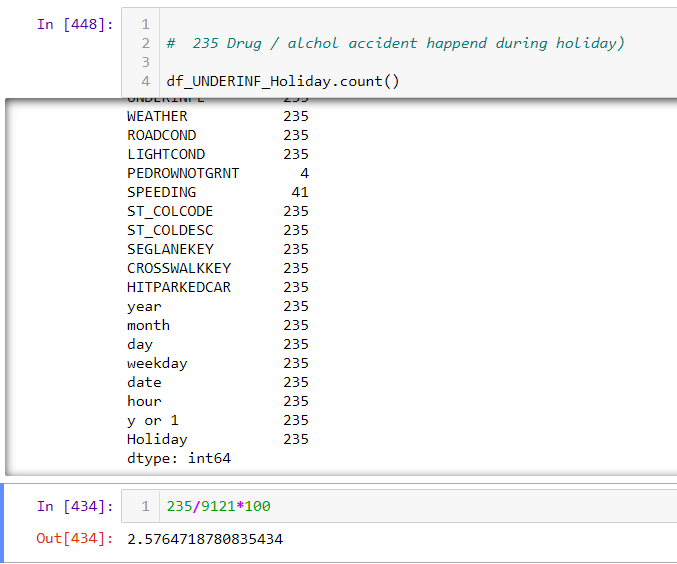
**Test Case 2.1c UNDERINFL vs. weekday (# Monday = 0, Sunday = 6)**



**Test Case 2.1d UNDERINFL vs. hour**



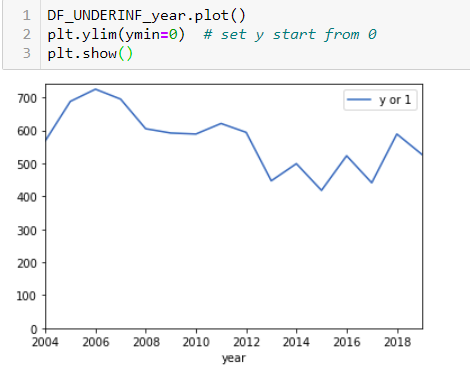
**Test Case 2.1 e UNDERINFL vs. Holiday**

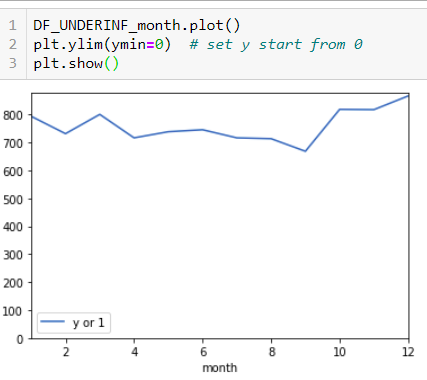


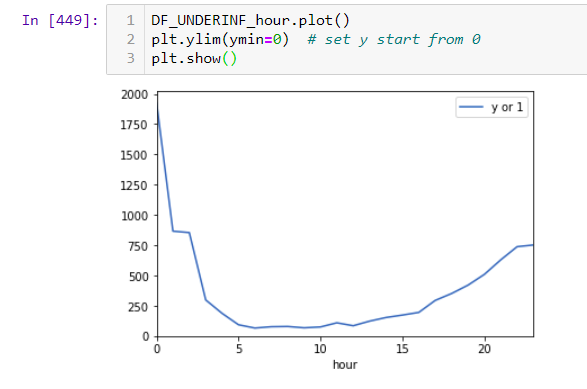
Total - 235

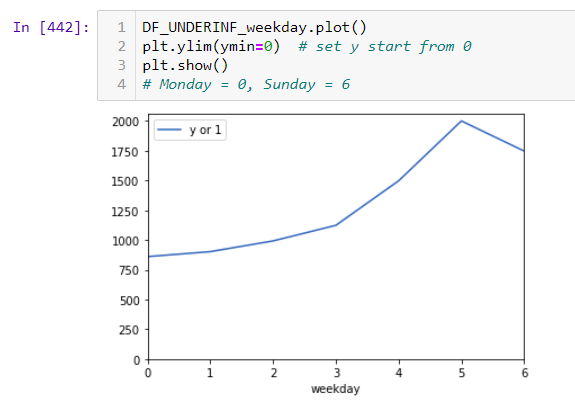
**Total incident caused by drug / alcohol = 9121, from 2004 to 2020.**

**Test Case 3.1: UNDERINF (Graphic to Visualize the result)**









**Test Case 2: SPEEDING**

**Test Case 1.2: Get insights from SPEEDING**

**Test Case 2.2: Data Analysis - SPEEDING**

**Test Case 3.2: Graphic to Visualize the result - SPEEDING**

**Test Case 3: Weather**

**Test Case 1.3: Get insights from Weather**

**Test Case 2.3: Data Analysis - Weather**

**Test Case 3.3: Graphic to Visualize the result - Weather**

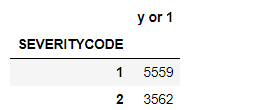
* **Get insights by using correlation and Visualize the result**

1. **Using two potentially data to find the correlation**

**Test Case 4.1: Correlations of Drugs or alcohol vs other factors**

1. **After Drugs or alcohol, 48.7% higher chances getting injured during car accident**

**After Drugs or alcohol**

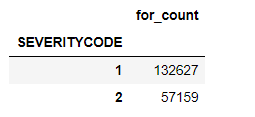


**#• 2—injury**

**#• 1—prop damage**

**Chances getting injured = 64%**

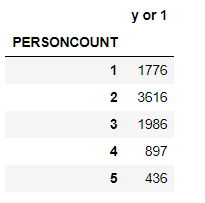
**Total Population** **df2['for\_count']=1**



**Chances getting injured = 44%**

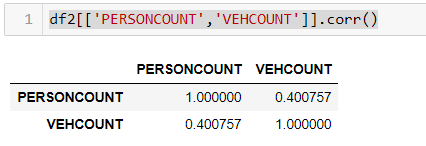
**Therefore, After Drugs or alcohol, 48.7% higher chances getting injured during car accident**

1. **with Drugs or alcohol, always at least 1 PERSON injured**

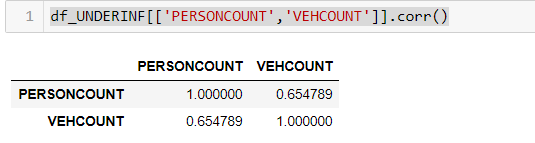


1. **with Drugs or alcohol, the people and vehicles involved has higher correlation. That means, more people injured than normal car accident**

**Normal car accident**



**with Drugs or alcohol**



**Test Case 4.2a: Speed vs Severity Code**

**Test Case 4.2b: Speed vs PERSONCOUNT**

**Test Case 4.2: Visualize the result**

**Test Case 4.3: Weather vs Severity Code**

**Test Case 4.3: Weather vs PERSONCOUNT**

**Test Case 4.3: Visualize the result**

**Discussion section where you discuss any observations you noted and any recommendations you can make based on the results.**

* **Observations & Findings**
* **Recommendations**

**For the Test Case 1: driver involved was under the influence of drugs or alcohol**

**We have the following result**

**1. Background data**

**2. driver involved was under the influence of drugs or alcohol ("drugs or alcohol") vs. Year**

**3. drugs or alcohol vs. Month**

**4. drugs or alcohol vs. hour**

**5. drugs or alcohol vs. weekday**

**6. drugs or alcohol vs. Holiday**

**Also, we have the graphic to show the Trend.**

**Observations and Recommendations for test case 1:**

**1. Background data**

**Observations:**

**# Driver involved was under the influence of drugs or alcohol = 9121**

**# Total case = 189786**

**# Overall %9121/189786 = 4.81%**

**Around 5% of car accidents are caused by driver with drug & Alcohol, the percentage is High. This is a problem.**

**Premilitary recommendation**

**Education and controls are highly recommended and should be strength**

**2. driver involved was under the influence of drugs or alcohol ("drugs or alcohol") vs. Year**

**Observations:**

**Based on the distribution of the result data, from 2004-2019 (2020 should be further investigate whether the data is missing or not entered), no significant down trend was observed. The trend is constantly high.**

**Premilitary recommendation:**

**Education and controls are highly recommended and should be strength**

**3. drugs or alcohol vs. Month**

**Observations:**

**Based on the distribution of the result data, October – December have a slightly higher number of car accident caused by drugs or alcohol. However, the trend is constantly high.**

**The slightly higher number of car accident in Oct – Dec may due to holiday Season (Halloween and Christmas).**

**Premilitary recommendation:**

**More detective controls (e.g. increase number of patrolling police) and preventive controls (TV advertisement, News Paper advertisement, Radio broadcasting, etc.) should be implemented during the holiday season.**

**Warning should be in placed at the roads that have large traffic.**

**4. drugs or alcohol vs. hour**

**Observations:**

**Based on the distribution of the result data, being from 21:00 until 0200, with the peak at 00:00, most car accident with alchol and drug happened during this period of hours.**

**This may due to people gathering after the dinner time.**

**Premilitary recommendation:**

**More detective controls (e.g. increase number of patrolling police) and preventive controls (TV advertisement, News Paper advertisement, Radio broadcasting, etc.) should be implemented during these hours**

**5. drugs or alcohol vs. weekday**

**Observations:**

**Based on the distribution of the result data, being from Friday to Sunday, with the peak on Saturday, most car accident with alcohol and drug happened during the weekend.**

**Premilitary recommendation:**

**More detective controls (e.g. increase number of patrolling police) and preventive controls (TV advertisement, News Paper advertisement, Radio broadcasting, etc.) should be implemented during these hours**

**6. drugs or alcohol vs. Holiday**

**Observations:**

**# Driver involved was under the influence of drugs or alcohol = 9121**

**# happened on USA National Holiday = 235**

**# Overall =2.58%**

**It seems irrelevant.**

**Premilitary recommendation:**

**N/A**

**For the Test Case 2:**

**We have the following result**

**1. Background data**

**2. driver involved was under the influence of drugs or alcohol ("drugs or alcohol") vs. Year**

**3. drugs or alcohol vs. Month**

**4. drugs or alcohol vs. hour**

**5. drugs or alcohol vs. weekday**

**6. drugs or alcohol vs. Holiday**

**Also, we have the graphic to show the Trend.**

**Observations & Findings for test case 2:**

**Recommendations for test case 2:**

**For the Test Case 3:**

**We have the following result**

**1. Background data**

**2. driver involved was under the influence of drugs or alcohol ("drugs or alcohol") vs. Year**

**3. drugs or alcohol vs. Month**

**4. drugs or alcohol vs. hour**

**5. drugs or alcohol vs. weekday**

**6. drugs or alcohol vs. Holiday**

**Also, we have the graphic to show the Trend.**

**Observations & Findings for test case 3:**

**Recommendations for test case 3:**

**Conclusion section where you conclude the report.**

* **Conclude the report**

**Conclusion for test case 1:**

**Around 5% of car accidents are caused by driver with drug & Alcohol, the percentage is High. This is a problem.**

**The trend is constantly high and does not have any significant drop over the years. The following are key factors:**

**- October – December have a slightly higher number of car accident caused by drugs or alcohol.**

**- 21:00 until 0200, with the peak at 00:00, most car accident with alcohol and drug happened during this period of hours.**

**- being from Friday to Sunday, with the peak on Saturday, most car accident with alcohol and drug happened during the weekend**

**More detective controls (e.g. increase number of patrolling police) and preventive controls (TV advertisement, News Paper advertisement, Radio broadcasting, etc.) should be implemented during the date with above factor.**

**Warning Message (e.g. billboard) should be in placed at the roads that have high traffic.**

**Conclusion for test case 2:**

**Conclusion for test case 3:**

**Lastly, the result of this report and the result data analysis may be preliminary, Further study of the results is required, e.g.**

* + **location black spot, vs weather**
  + **location black spot, vs 'PEDCOUNT,**

**in order to get more precise results and root cause.**