**Predicting the severity of Car Crash**

**1.Introduction**

Car accidents are a huge problem and have been since roughly the beginning of the mass deployment of cars. There are factors which are related to the person at the wheel, such as driving under influence, speeding, stunting or other risky behavior. however then there are also other factors, such as whether conditions, light conditions or even timing (an example is driving at a time where there is a higher percentage of people driving under influence).

Road accidents are social problems- especially devere car accidents which can ruin absolutely ruin a person, their family and as I see it, a critical piece of society. However, severe accidents can probably be reduced significantly, if we were to understand the conditions which make crashing likely more severe. This can help reduce the severity of crashes when they occur, by understanding what conditions make a crash more severe, for example certain weather and road conditions, etc. This information can be given to driving schools, government and private agencies and institutions, which can help develop methods to put what is found into practicality and reduce the rate of severe accidents- so these are key stakeholders in this project. Also another key stakeholder is simply anyone who wants to be safer on the roads and reduce the risk for severe accidents.

So what exactly is the problem? The problem is: there are a lot of crashes in the Seattle area- actually to be more specific, in Washington state, there were over 45,000 car crashes in 2019, severe and over 200 were fatal. That is a lot! Now the question to ask is: how can we most effectively reduce the number of severe crashes? I believe it is to understand how these crashes took place and their conditions, so we can help people avoid being in conditions which have a higher likelihood to end in a severe car crash. This is all aimed at drivers in the Seattle area, who are trying to reduce the risk of ending up in a severe car accident and agencies which can help make driving in Seattle safer- this may include a wide range of people, including engineers, city planners, schools, public transport, etc, who could leverage the machile learning model produced in thsi project to develop safer streets.

To sum it all up, this project is aimed at reducing the amount of severe car crashes, by building a model which will predict whether a car crash will be severe or not given certain conditions and we will also analyze data for additional understanding.

**2. Data**

The data I am using is from the Seattle Police Department and consist of over 194,000 records from 2004 to the present. The data contains 37 columns/attributes. A lot of these attributes are not of concern to us for this project, however there are a few key attributes which have a very strong significance.

Firstly, I dropped all the columns that were of no use to me, which left me with these key rows:

ADDRTYPE - Intersection, block or alley

SEVERITYCODE - Severity of the accident: 1. Not-severe, 2. Severe

VEHCOUNT - number of vehicles involved

INATTENTIONIND - Whether the driver was not paying attention leading to the crash.

UNDERINFL - driver involved under the influence of drugs or alcohol

WEATHER - weather conditions

ROADCOND - Road condtitions

LIGHTCOND -Light Conditions

SPEEDING -Whether a car involved in accident was speeding

HITPARKEDCAR - did the accident involve a parked car.

If the there was no value written for SEVERITYCODE, I immediately dropped the row. Moreon, I changed all boolean columns from Y/N to 1/0 and if there was a missing value in them, I added put in a 0. More on, I also used dummy values for categorical columns, such as ROADCOND. These dummy values created new columns which were boolean in the for of 1/0. The dummy values I created for modeling and testing the models, however for the analysis I did before the models, I kept them in categorical terms, for easier analysis.

For the predictive Modeling I used only 20% of the data for training and testing, as the data was too huge to do modeling with. The data was all selected randomly and still being 20% of the overall data held almost 40,000 records, which is still very good for creating a model.

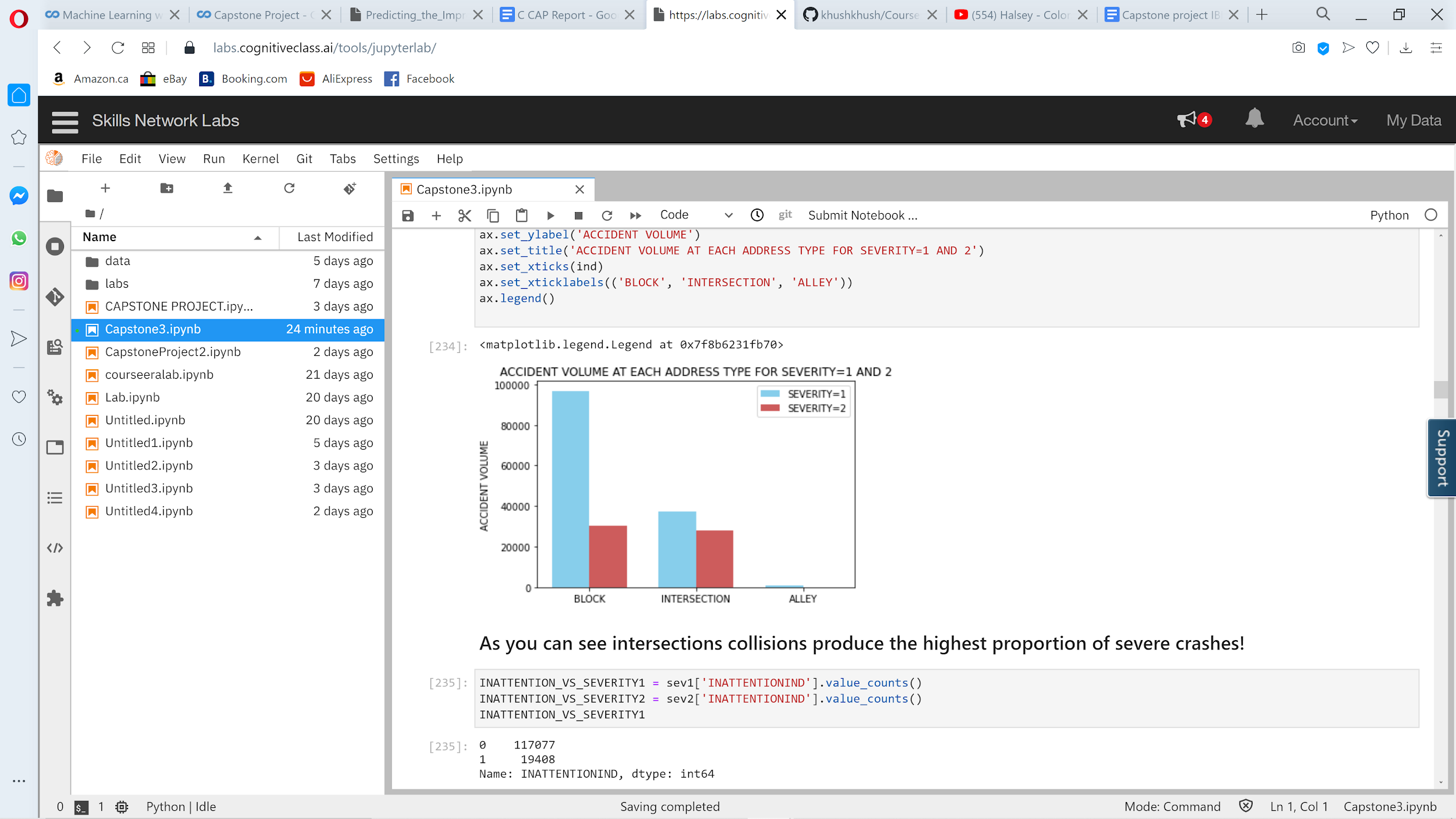
**Methodology:**

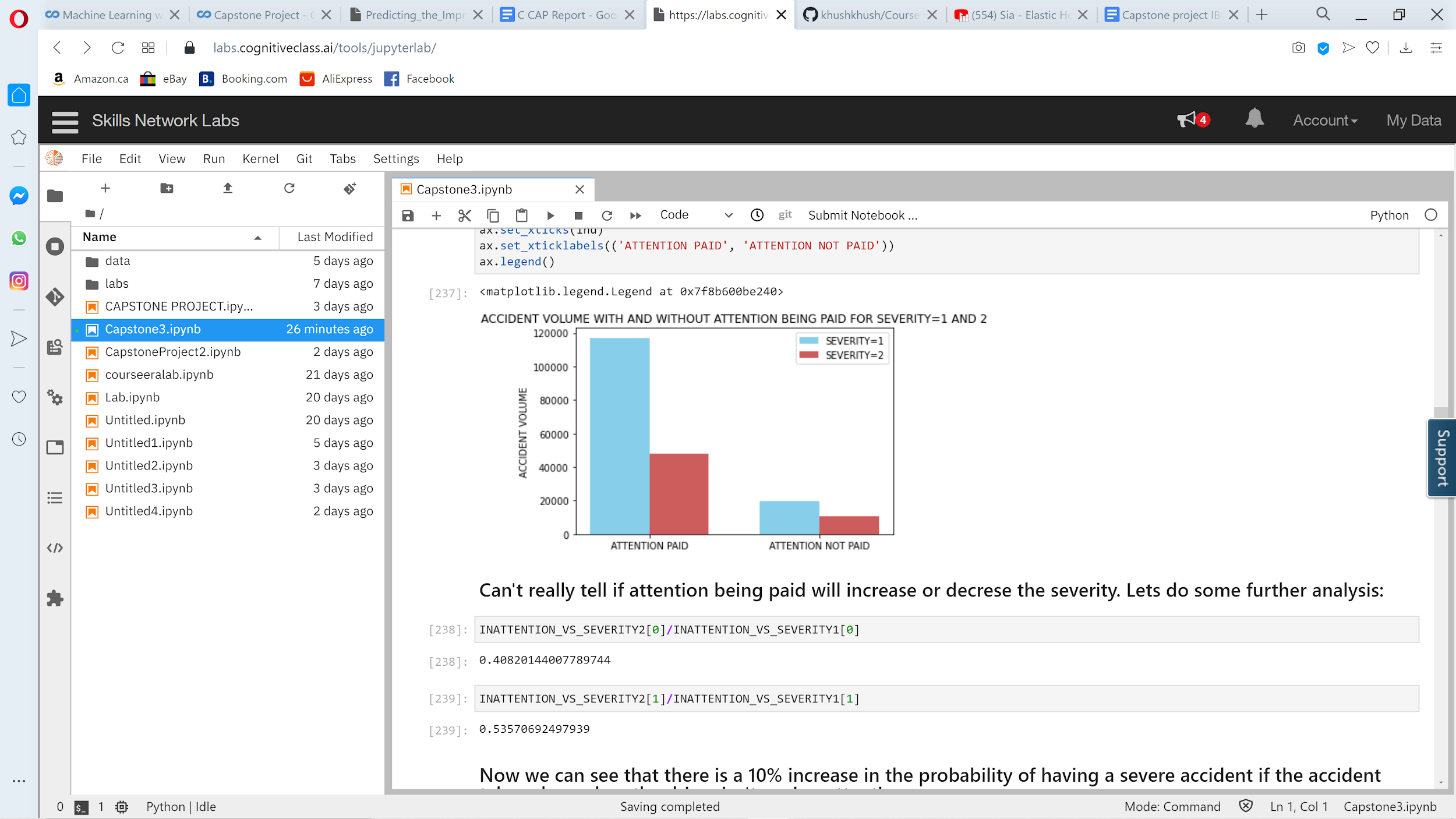
**Data analysis:**

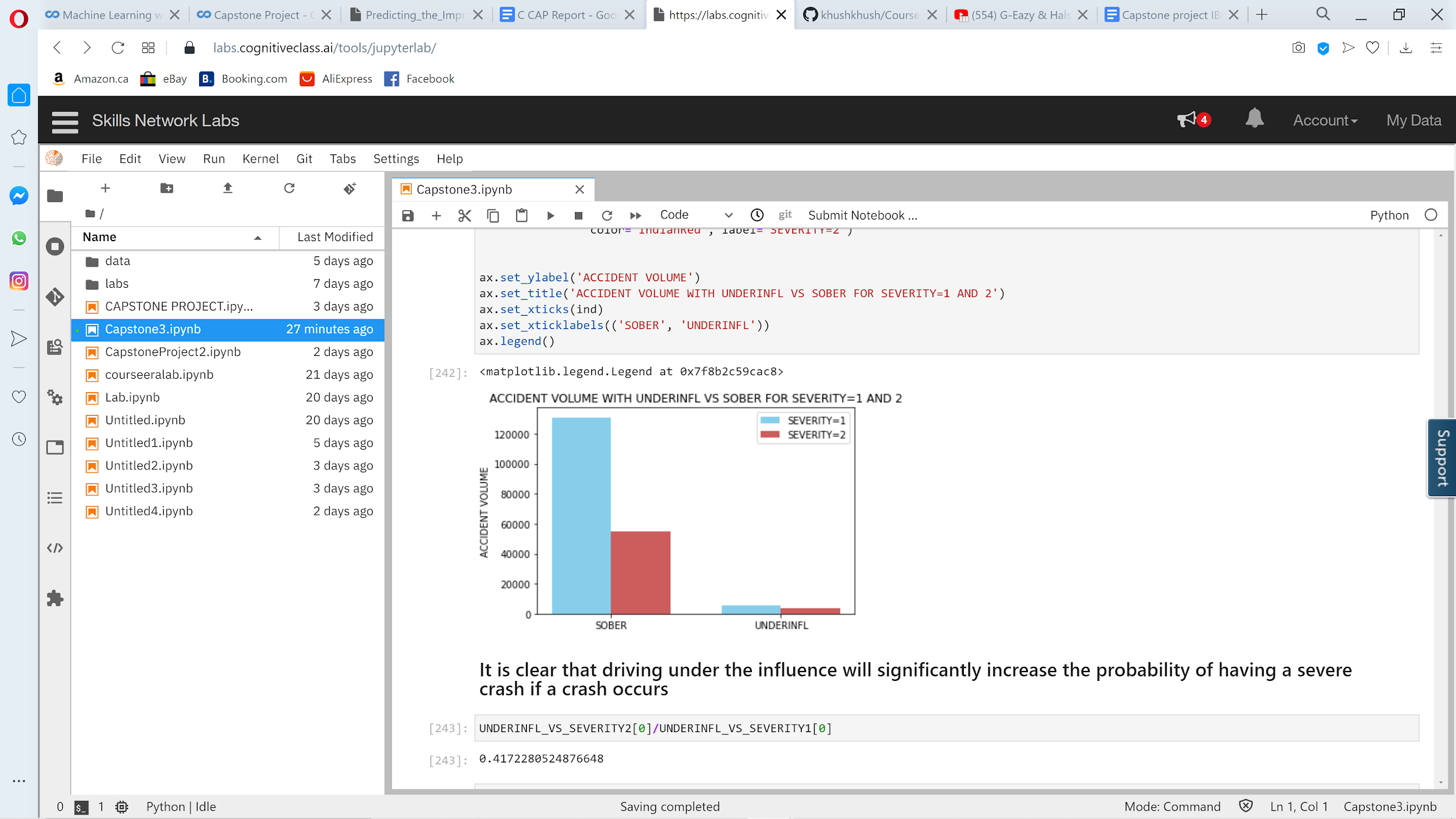
The use of grouped bar charts were used to express the ratio of severe to non severe crashes.

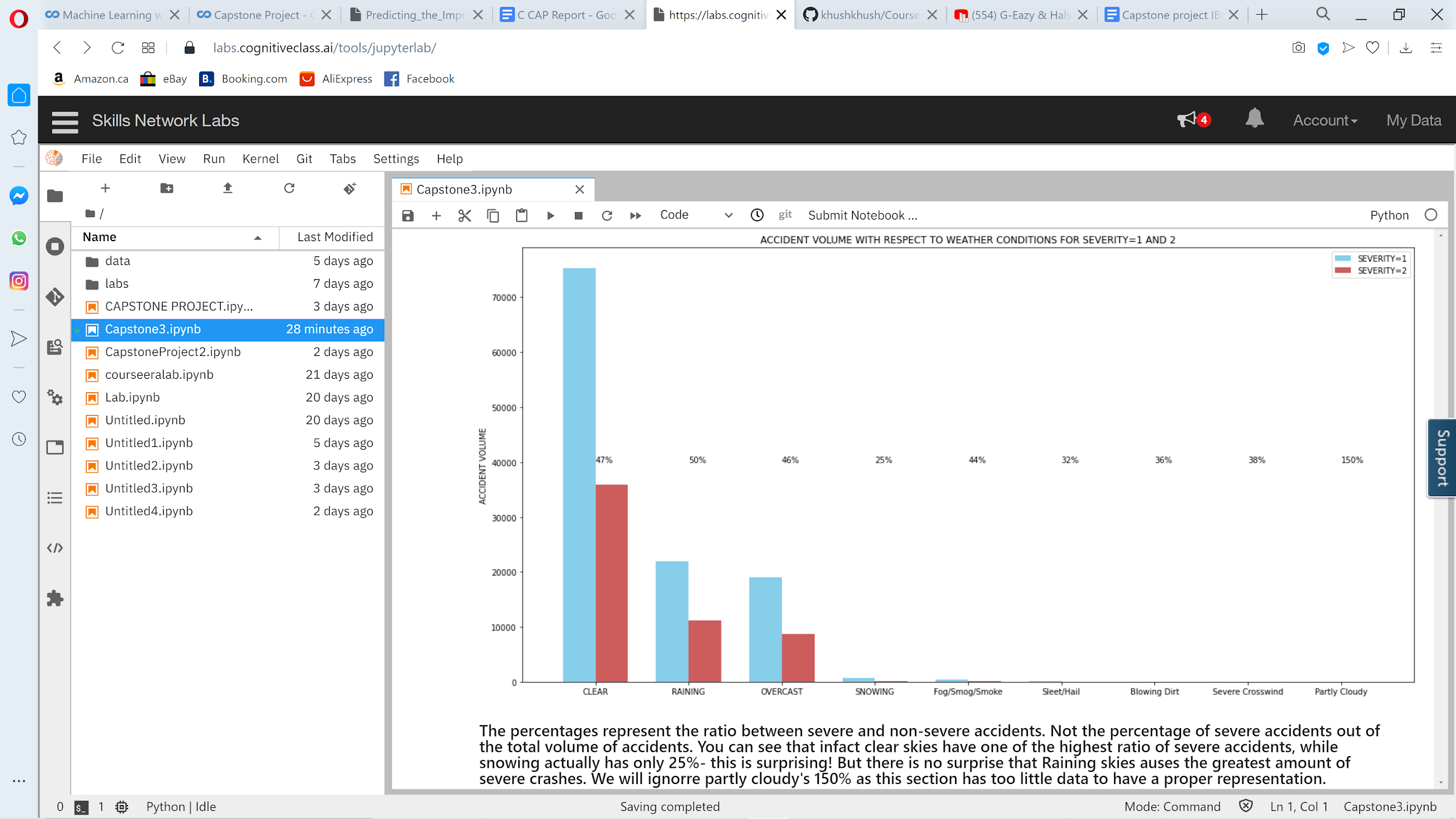
Firstly, I had to create 2 data frames 1 which held all the data for non-severe crashes and 2 for data with severe crashes. Then I created series of datas from each of the frame regarding the attributes(columns) I was interested in observing. Then I was able to produce the grouped bar charts, where the group represents each condition.

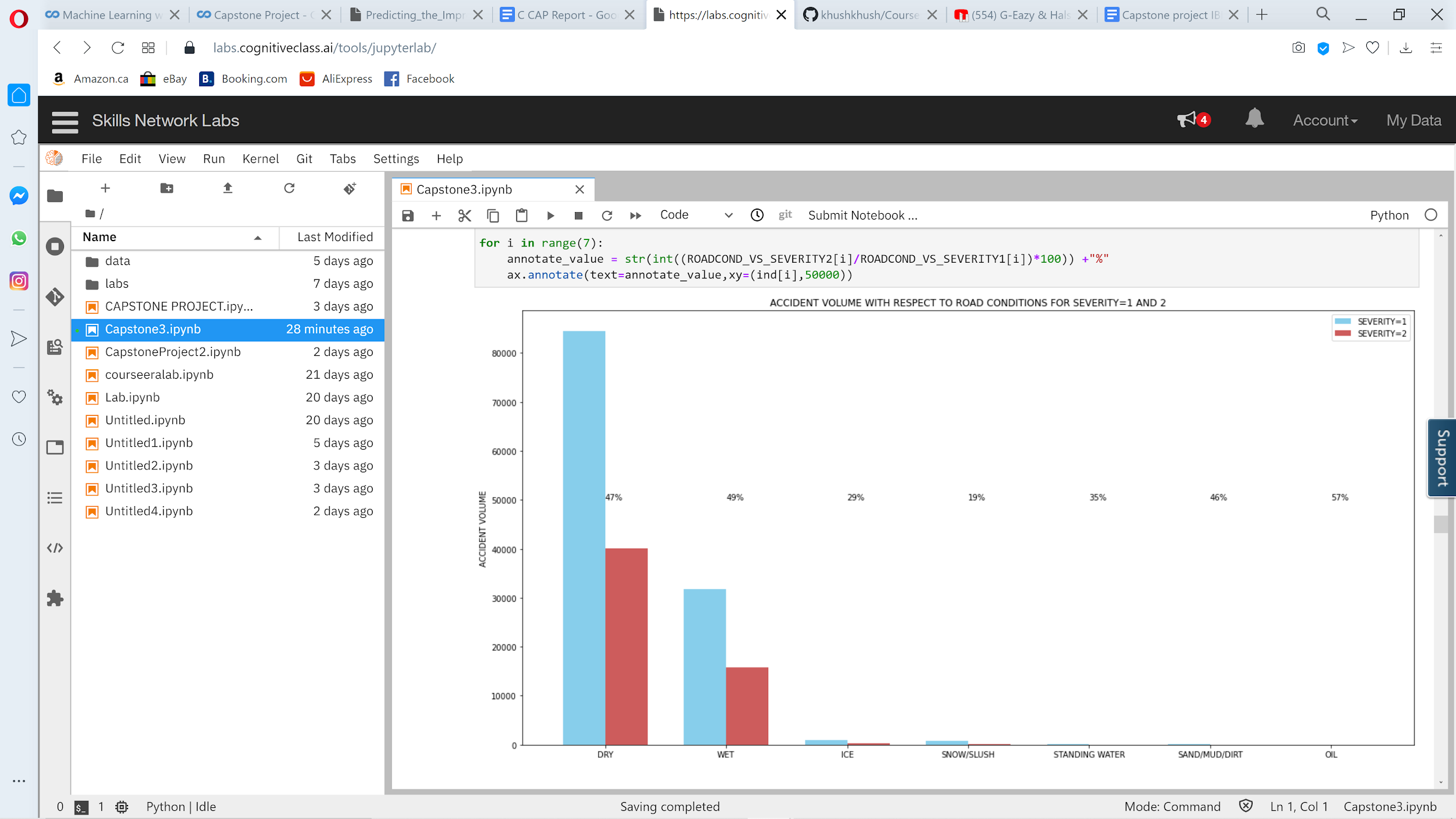
Here are the results below:

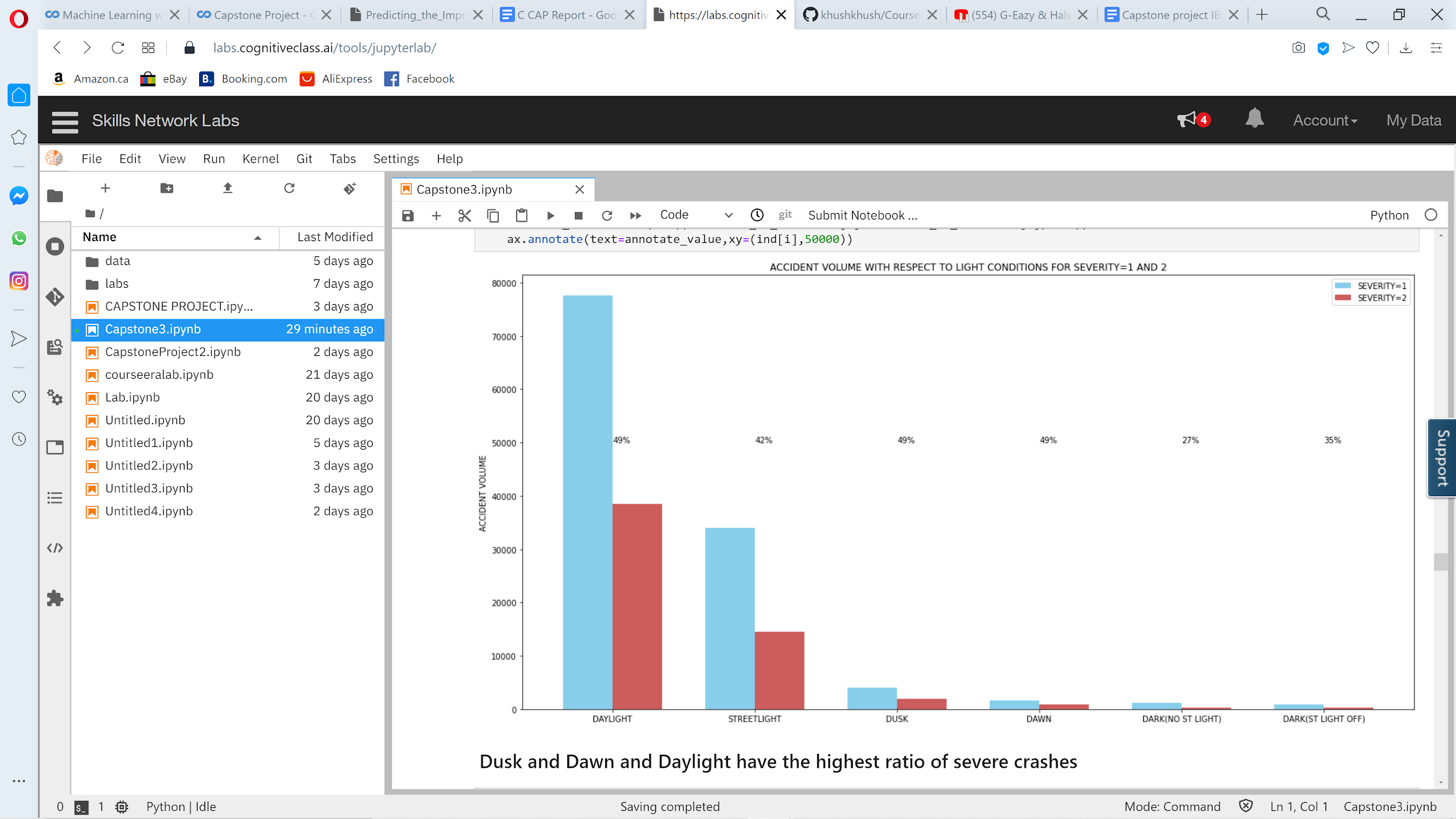


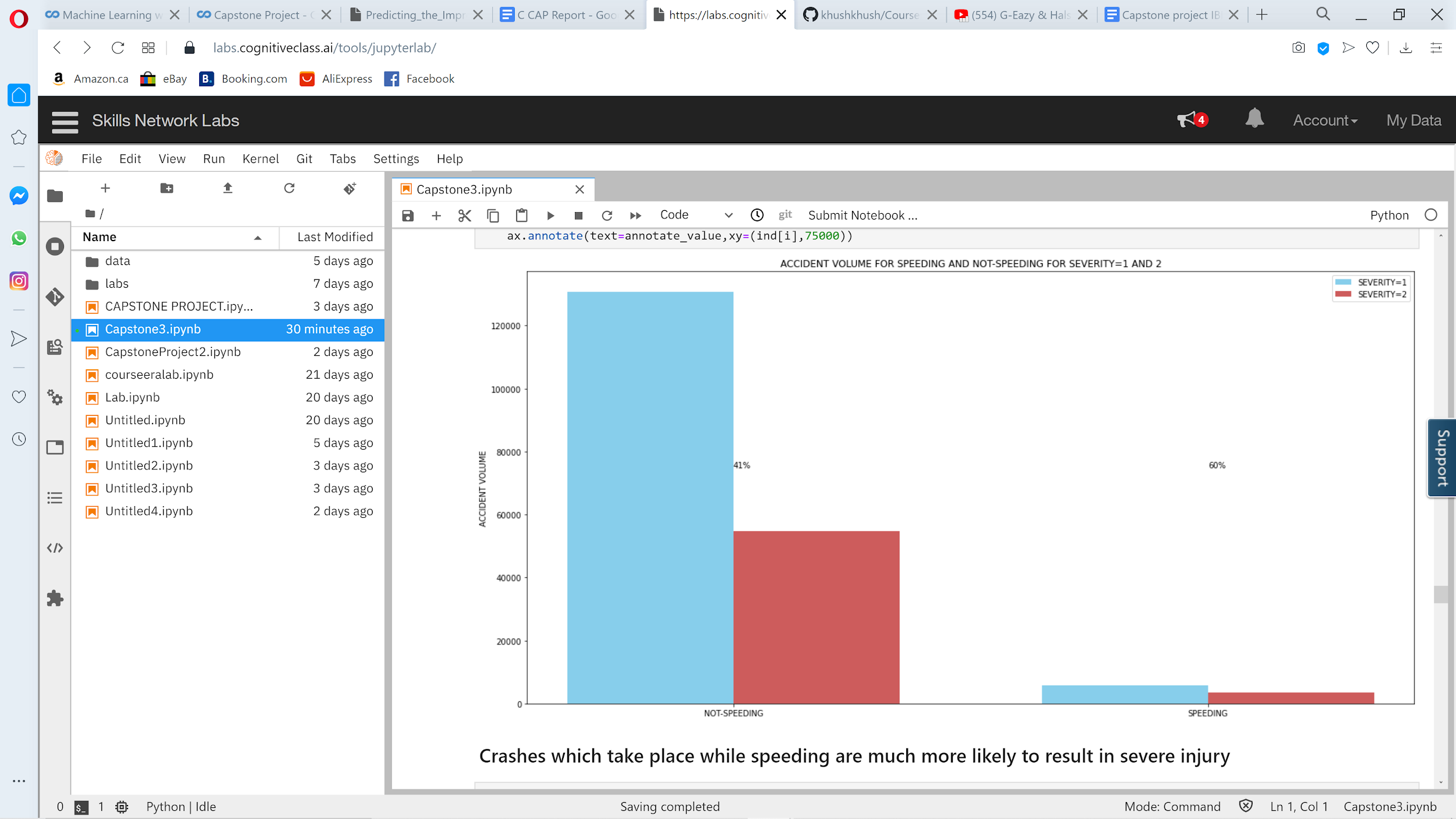
****

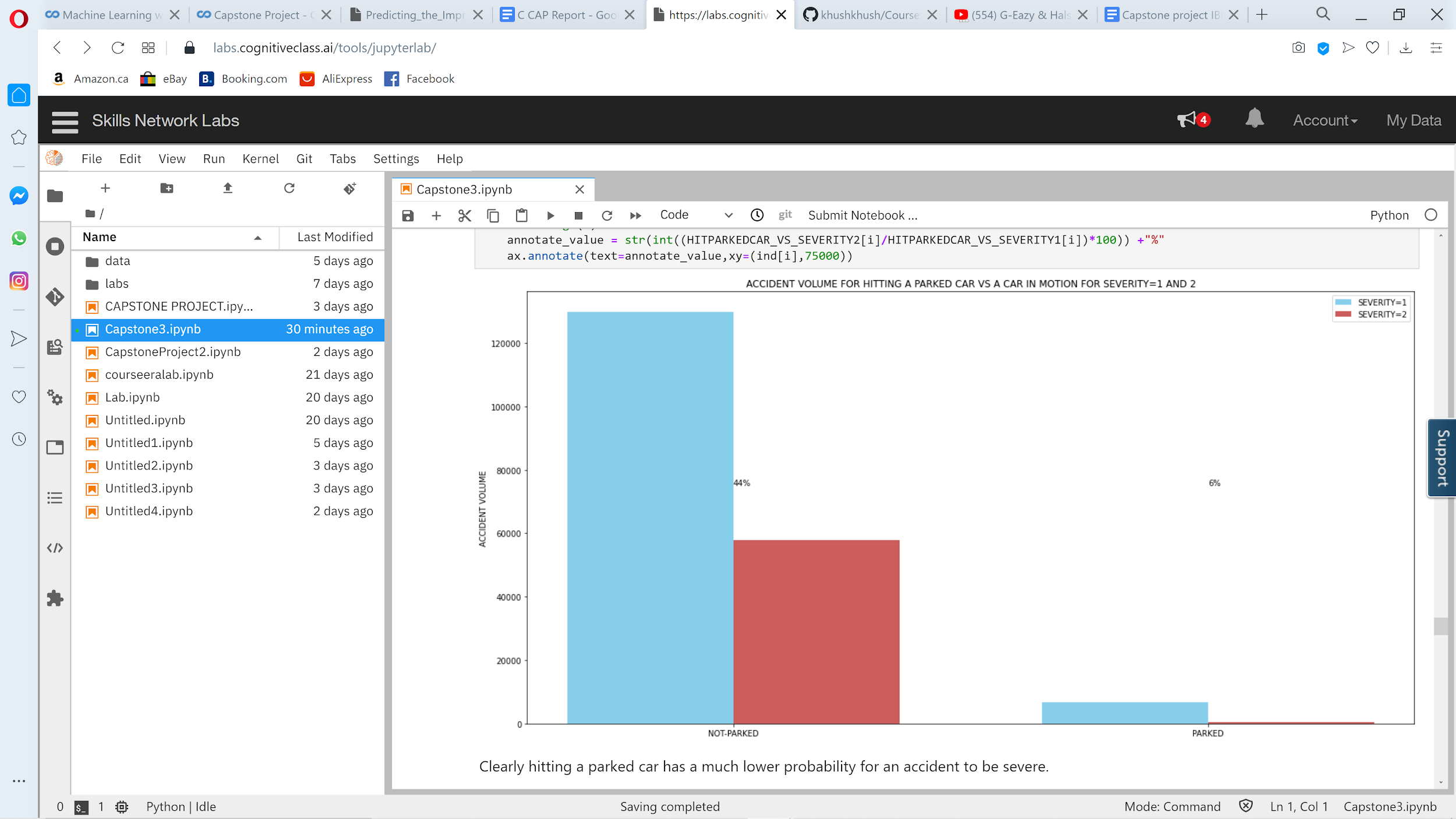
****

****

****

****

****

****

**Model Building:**

Now we will construct a model which will take the factors we have analyzed into it to create a predictive model to predict the severity of a car crash from the input. We will first need to process the data abit in order to make it right for the model to accept and to create a split between training data and test data. Then we will be able to evaluate the data to make sure that we have highly accurate model that we can then deploy and work with to make Seattle a safer city.

We will be using a Classification machine learning model, as we have 2 categories: 1- non-severe 2-Severe. This means that we want a model which will predict whether an accident is severe or non severe, so a classification model is what we are looking for.

Also we need to keep in mind that we are dealing with almost 200,000 data points, and quite a few columns of data, thus, with such huge amounts of data we will need a model which can run relatively fast, as else, it will take a long time computing and this will be rather annoying for anyone working with the model. Furthermore, stakeholders who will be using this model may not always want to input every attribute of the independent variable (and in this case, we will want them to put in a 0, or false value for the component that they are not calculating for), so thus this model needs to be clearly split in a way which doesn't mitigate the values based on a lack of certain test components. For these reason, we will choose a Support Vector Machine Model, as it fits exactly the description of what we need.

More on, we have also built a Decision Tree model just for comparison sake and it had almost identical results after being tested.

Finally a logistic regression model was also built, however it did have somewhat lower accuracy, however the advantage of a logistic regression model is that we are able to predict the probability of a Severe accident occurring given certain points, which is very useful to stakeholders.

We used The Jaccard similarity Score and the f1 score to assess how well the model predicted whether an accident was severe or not. The evaluation types use the models' predictions and compare them to their real values. We have a certain amount of data we allocated only for testing and evaluating and is completely random and real data, which was separated from the data provided by the Seattle Police Department before the model was actually built.

# **Results and Discussion:**

There are quite a few factors which can help us determine whther a car was in a severe car crash or not. Lets look at the figures: 1- We found that car crashes which involve a parked car are highly unlikely to produce severe accidents. 2.Speeding produces a significant increase in likelihood of an accident being severe. 3. Being under the influence also heavily increases the likelihood of having a severe accident. 4.When attention isnt being paid the liklihood of severity also increases. 5.Accidents in intersections have a very high likelihood of being severe, while accidents in alleys have a very low likelihood. The most accidents happen on blocks, which have a moderate likelihood of having a severe accident. 6. Raining weather has the worst probability for severity of accidents, and clear, overcast and foggy conditions are close behind- they all produce roughly a 1/3 probability that an accident will be severe. All other conditions are far lower. 7.Wet then dry road conditions produce the 2 highest probability of severe accidents, once again a roughly 1/3 chance of an accident to be severe, while all other road conditions are much lower. 8. With regards to light conditions, it turns out that dark conditions are actually better in respect to a reduction in severe crashes- when there were no street lights at all we had the lowest proportion of accidents being severe, then when the street lights were turned off. More on, Dusk, Dawn and Daylight are all tied in first place for producing the highest likelihood of an accident to be severe and then streetlights is close behind in second place- in these conditions an accident has roughly a 1/3 chance of being severe. Perhaps if Seattle was a darker city, it would be safer.

# **Conclusion:**

The purpose of this project is to 1.Understand the factors which produce a higher likelihood of a crash being severe and 2. to build a model which can predict the sverity of a crash based on factors of input. We have been able to successfully determine conditions which lead to a higher likelihood of a severe crash and build a model which is ready to deploy to estimate the severity of a crash giving a litany of conditions. The model doesn't have the best accuracy to be honest, however due to the nature of a car crash being based on so many factors and involving alot of luck, this can be expected. However, the model's accuracy is still high enough to help increase safety substantially. There were also some surprises found in the data analysis, which are a good thing, as they differ from what we might have expected prior to the analysis, such as the fact that darker light conditions actually produce a lower probability of an accident being severe. These are valuable insights for constructing safer cities. More on, the model can be deployed for scenario evaluation, to determine which parts of a city or design of a city will produce severe accidents in differnt conditions. The Log loss model, could even be deployed to predict the probability of having a sever crash- these models could be very useful for building safer roads in the future and producing better drivers and allowing stakeholders to find conditions and points of danger so they can be more aware and thus pay more attention to these points of interest.