Case Study:

Predicting Severity of Car Accidents

Gabija Vaisvilaite

2020

# Introduction

This is a project for IBM course Applied Data Science Capstone on Coursera. The aim of the project is to work on a case study which is to predict the severity of an accident using machine learning models and data science techniques learned on previous courses. To do that we will try to gain some insight from the data: under what weather, conditions of the road and lighting the incidents resulting in injuries occur more often? Are they more likely to happen in during the day or during the night? Do accidents occurring at an intersection tend to be more severe? Does speeding on average increase the severity of the accident? We will try to answer such questions using data. Specifically, we will try to find out which factors have the most impact on severity of the accident, create machine learning algorithms and train them to predict the severity of a possible accident.

The given problem arises due to the lack of warnings and information about the weather and road conditions for the drivers, having those people would drive more carefully or, if possible, even change their travel. Such warnings with predictions how severe the car accident could be based on current circumstances, may reduce the number of accidents, casualties and injuries, which in turn reduce cost of damage. In order to develop methods reducing the damage of potential accidents it is crucial to determine factors leading to said accidents. To efficiently use limited resources the priority is to create strategies aimed at minimizing the risk of severe accidents, hence we particularly try to find factors leading to more severe accidents. As the project's objective is to determine what causes more severe incidents, this information could be most useful for insurance companies and governments (so they could, for example, improve lighting conditions at certain intersections/roads), although it may possibly be useful for car manufacturers and road construction companies.

# Data Understanding

The given data set *Data-Collisions.csv* contains all collisions provided by Seattle Police Department from 2004 to present. The attributes in the data set include:

1. Time and Location
   1. Coordinates of the collision **X** and **Y**
   2. Description of the general location of the collision **LOCATION**
   3. Type of the location **ADDRTYPE** 
      1. Alley
      2. Block
      3. Intersection
   4. Category of junction at which collision took place **JUNCTIONTYPE**
   5. Key that corresponds to the intersection associated with a collision **INTKEY**
   6. Key for the lane segment in which the collision occurred **SEGLANEKEY**
   7. Key for the crosswalk at which the collision occurred **CROSSWALKKEY**
   8. Date of the incident **INCDATE**
   9. Date and time of the incident **INCDTTM**
2. Conditions
   1. Description of the weather conditions during the time of the collision **WEATHER**
   2. Condition of the road during the collision **ROADCOND**
   3. Light conditions during the collision **LIGHTCOND**
3. Involved parties
   1. Number of people **PERSONCOUNT**
   2. Number of pedestrians **PEDCOUNT**
   3. Number of pedal cyclists **PEDCYLCOUNT**
   4. Number of vehicles **VEHCOUNT**
4. Behaviour of involved parties
   1. Whether or not speeding was a factor in the collision (Y/N) **SPEEDING**
   2. Whether or not the collision involved hitting a parked car (Y/N) **HITPARKEDCAR**
   3. Whether or not the pedestrian right of way was not granted (Y/N) **PEDROWNOTGRNT**
   4. Whether or not collision was due to inattention (Y/N) **INATTENTIONIND**
   5. Whether or not a driver involved was under the influence of drugs or alcohol **UNDERINFL**
5. Details of the incident
   1. Code given to the collision by SDOT **SDOT\_COLCODE** (for more information see the [State Collision Code Dictionary](https://d.docs.live.net/1dc3c47e602e5ea2/Attachments/Documents/State%20Collision%20Code%20Dictionary))
   2. Description of the collision corresponding to the collision code **SDOT\_COLDESC**
   3. Code provided by the state that describes the collision **ST\_COLCODE**
   4. Description that corresponds to the state’s coding designation **ST\_COLDESC**
   5. Number given to the collision by SDOT **SDOTCOLNUM**
   6. Collision type **COLLISIONTYPE**
6. Severity of the incident
   1. Code that corresponds to the severity **SEVERITYCODE**
      1. **3**—fatality
      2. **2b**—serious injury
      3. **2**—injury
      4. **1**—property damage
      5. **0**—unknown
   2. Detailed description of the severity **SEVERITYDESC**

The data set attributes also include a unique number for each incident **OBJECTID**, a report number **REPORTNO**, the column **SEVERITYCODE** copy **SEVERITYCODE.1**, **INCKEY** and **COLDETKEY**, whether **INCKEY** matches **COLDETKEY** column **STATUS**, **EXCEPTRSNCODE** and **EXCEPTRSNDESC**.

# Data Preparation

* We have changed columns **INCDTTM** and **INCDATE** to type datetime, realized that column **INCDATE** has less information than **INCDTTM**, thus we deleted this column.
* Due to many missing entries we have also deleted columns **EXCEPTRSNCODE**, **EXCEPTRSNDESC.**,
* As columns **SEVERITYCODE** and **SEVERITYCODE.1** are equal, **SEVERITYCODE** takes only values 1 and 2 and **SEVERITYDESC** values are *'Injury Collision'* if **SEVERITYCODE** is equal to 2 *and 'Property Damage Only Collision'* if **SEVERITYCODE** is equal to 1, we deleted column **SEVERITYDESC** and **SEVERITYCODE.1** and changed values of **SEVERITYCODE** to numeric values so that '1' corresponds to *'Minor Injury'* and '0' corresponds to *'Property Damage Only'*.
* We also changed values of columns **HITPARKEDCAR**, **SPEEDING**, **PEDROWNOTGRNT**, **UNDERINFL**, **INATTENTIONIND** from 0, N to 0 and from Y, 1 to 1. Also, as **SPEEDING**, **PEDROWNOTGRNT** and **INATTENTIONIND** have only values Y and NaN, we changed Nan to 0 (assuming that if there was no speeding, pedestrian was granted their right of way and the collision did not happen due to inattention, it was simply not recorded). Column **UNDERINFL** has null values as well, and as it has not been recorded, similarly as before we assume that parties involved where not under influence of drugs or alcohol and set null values to 0.
* Column **STATUS** represents whether **INCKEY** and **COLDETKEY** are equal. **INCKEY** and **COLDETKEY** are certain keys, that is, some identification numbers which do not provide us any relevant information; hence we deleted these columns. Similarly, **OBJECTID**, **REPORTNO** and **SDOTCOLNUM** are certain identification numbers which do not seem to provide us any information regarding factors of the incidents, hence we also dropped these attributes.
* Attribute **ST\_COLCODE** had some null values, as 31 corresponds to *'Not Stated'* in **ST\_COLCODE** we changed all null values to 31.
* **ST\_COLDESC** also had some null values, we changed those to *'Not Stated'*.
* We arranged weather, road and lighting conditions (**WEATHER**, **ROADCOND** and **LIGHTCOND**) into fewer categories so that it is easier to plot and model later:
  + Replaced *‘Overcast’* and *‘Partly* *Cloudy’* to *‘Cloudy’* in **WEATHER**.
  + Replaced *‘Standing* *Water’* and *‘Wet’* to *‘Wet’* in **ROADCOND**.
  + Replaced *'Dark* *-* *No* *Street* *Lights'* and *'Dark* *-* *Street* *Lights* *Off’* to *'Dark* *-* *Street* *Lights* *Off'* in **LIGHTCOND**.
  + Replaced *'Dusk'* and *'Dawn'* to *‘Dusk/Dawn’* in **LIGHTCOND**.
  + Changed values *‘Unknown’* and *‘Other’* to null values as they provide us no information.
* Similarly, we changed *‘Other’* and *‘Unknown’* in **COLLISIONTYPE** and **JUNCTIONTYPE** to missing values.
* Noticed that **X**, **Y**, **SEGLANEKEY**, **CROSSLANEKEY**, **INTKEY** and **LOCATION** attributes provide good location information. One could use these attributes to identify locations where most incidents occur. However, they do not provide any information about the characteristics of the location. Further, **SEGLANEKEY**, **CROSSLANEKEY** and **INTKEY** have most of the entries as zeroes i.e. most collisions happened in lanes and crosswalks that do not have their unique identifier, thus they can be considered as null values, so they are not very useful. Using these attributes might result in overfitting and generally they do not provide insightful results. So, we also dropped these columns.
* Changed types of attributes **ST\_COLCODE**, **INATTENTIONIND**, **UNDERINFL**, **PEDROWNOTGRNT** and **SPEEDING** to integer.
* As **JUNCTIONTYPE** is closely related to **ADDRTYPE**, using values in **JUNCTIONTYPE** we filled in some missing values in **ADDRTYPE**.
* Using descriptions **ST\_COLDESC** and **SDOT\_COLDESC** for **ST\_COLCODE** and **SDOT\_COLCODE** we have extracted some of the information to other attributes before dropping **SDOT\_COLDESC** and **ST\_COLDESC**:
  + As there are many codes in **ST\_COLCODE** that correspond to hitting an object or getting struck by one, namely codes 17, 18, 40, 41, 48, 49, 50, 51, 60, 64, 65, 66, 67, 85 and 87, we included *'Object'* as a collision type. Also, as code 42 involves a pedestrian and code 43 involves a pedal cyclist, we assigned corresponding types of collision.
  + Noticed that codes 21 and 22 in **ST\_COLCODE** correspond to accidents happening in a driveway access. Thus, we changed null values in **JUNCTIONTYPE** to *'Driveway Junction'* if the collision has recorded **ST\_COLCODE** 21 or 22.
  + Based on **SDOT\_COLCODE** and **SDOT\_COLDESC** we filled in some of null values in **COLLISIONTYPE** with *'Object'* if the code is one of 25, 26, 28 and 48; *'Pedestrian'* if the code is 24, and *'Cycles'* if the code is one of 18, 21, 23, 51, 54, 56, 66 and 69.

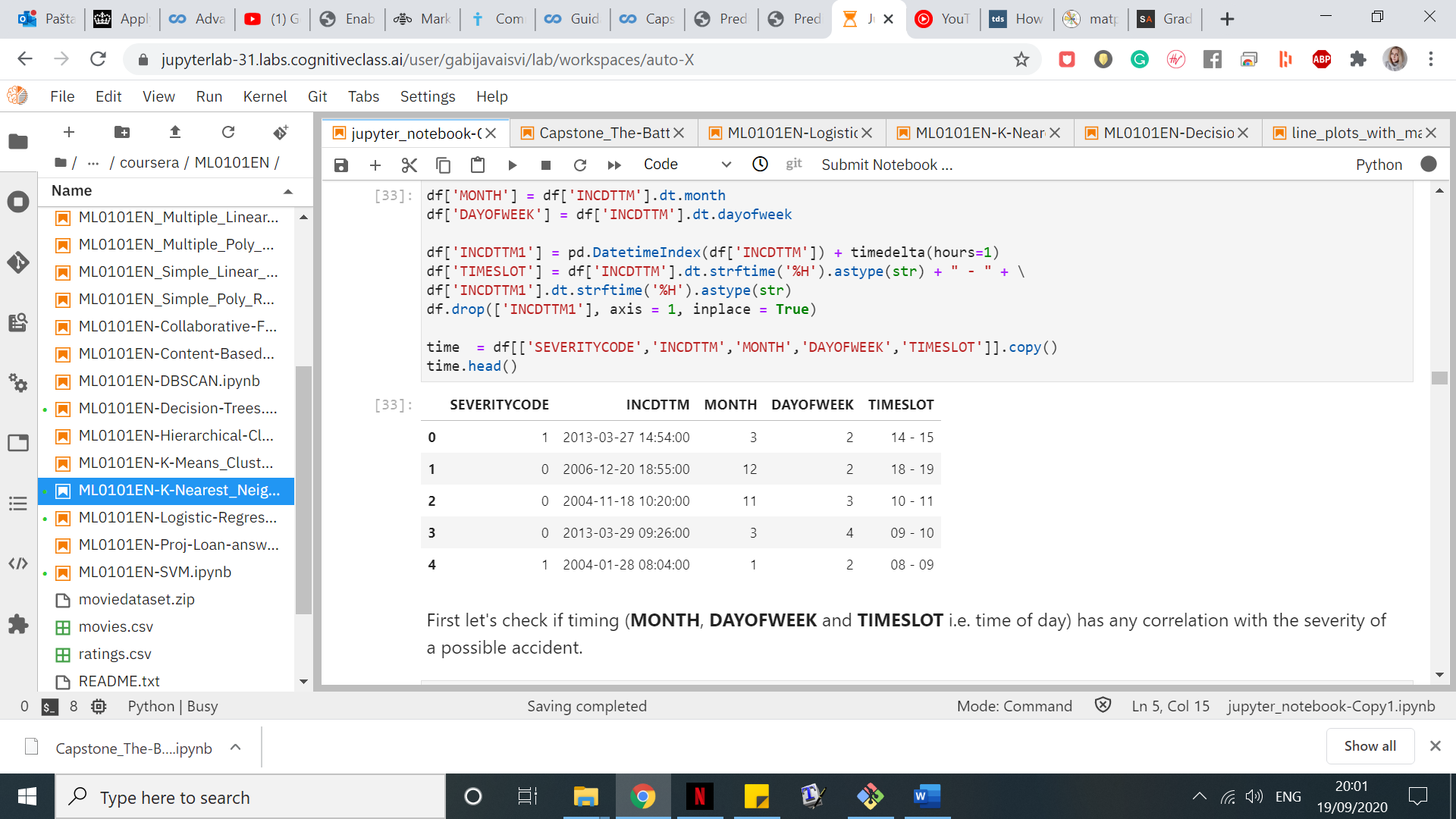
At the end we kept features:

* **SEVERITYCODE** (aim of the case study)
* **ADDRTYPE**, **JUNCTIONTYPE** (location information)
* **INCDTTM** (date and time information)
* **COLLISIONTYPE**, **SDOT\_COLCODE**, **ST\_COLCODE**
* **PERSONCOUNT**, **PEDCOUNT**, **PEDCYLCOUNT**, **VEHCOUNT**
* **UNDERINFL**, **PEDROWNOTGRNT**, **SPEEDING**, **INATTENTIONIND**, **HITPARKEDCAR**
* **WEATHER**, **ROADCOND**, **LIGHTCOND**

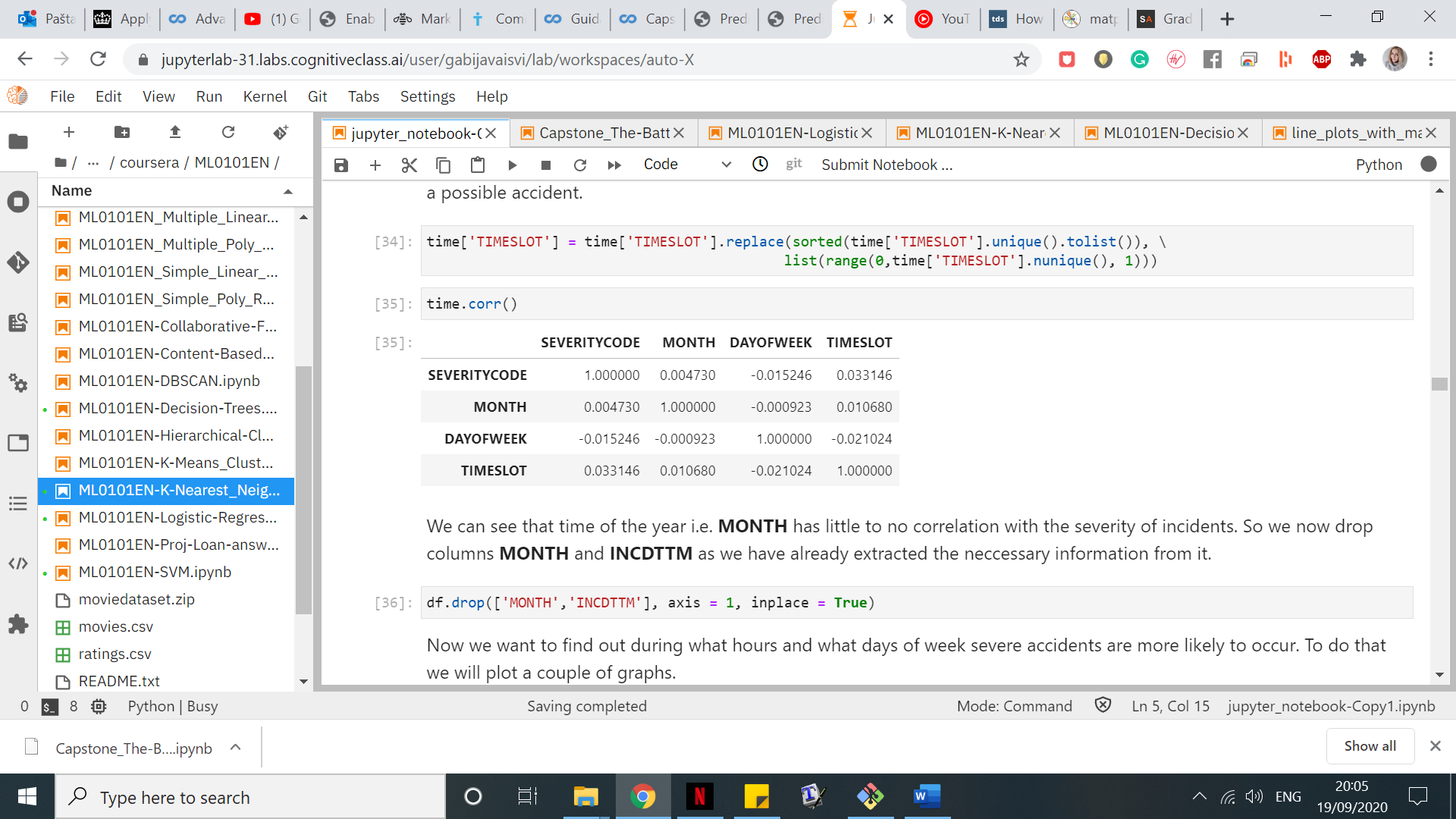
# Explanatory Data Analysis

## Date and time

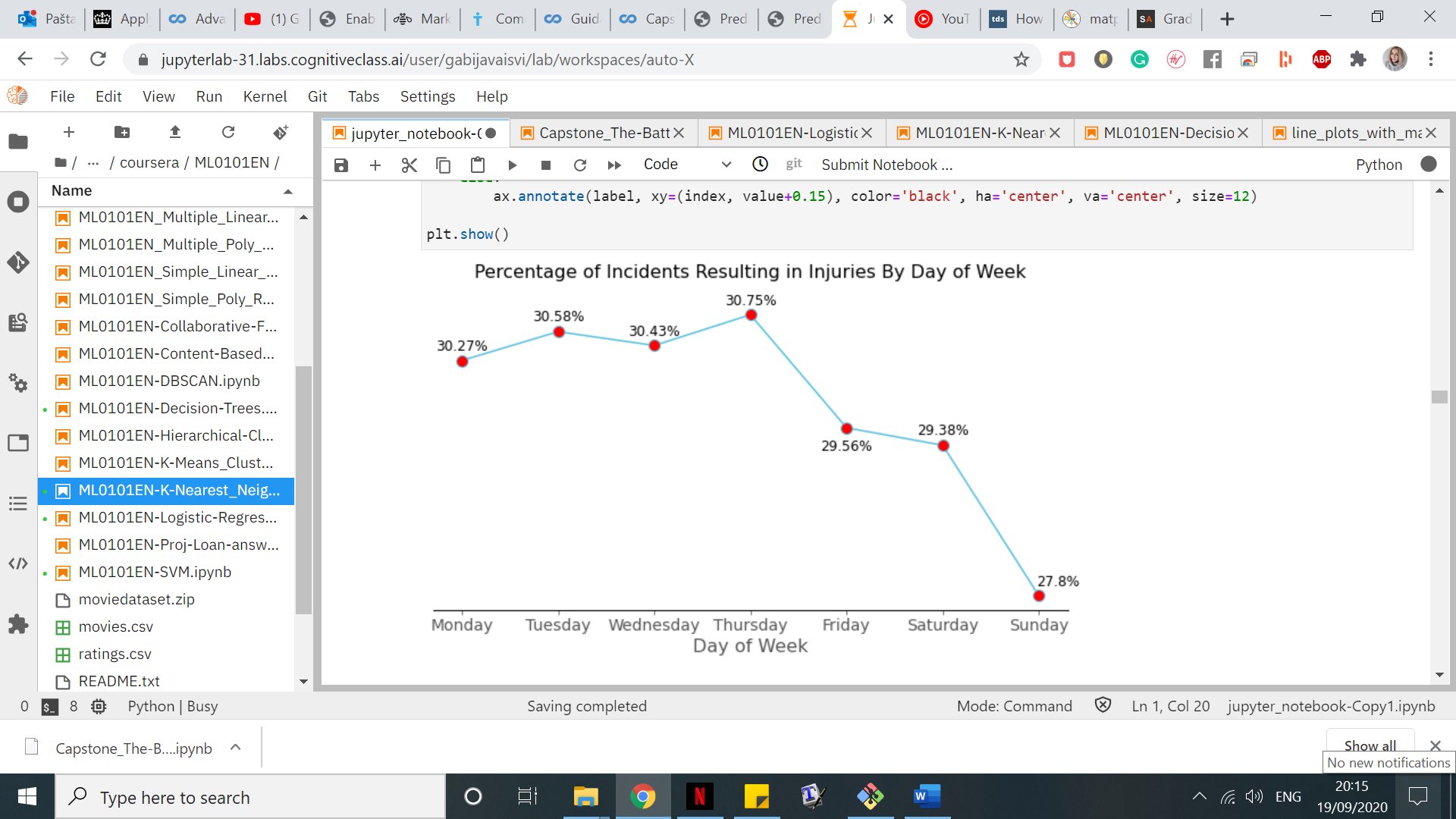
We analysed whether timing of the accident (month, day of the week, hour) had any impact on the severity of the accident. This is the first few entries of the data we used for this analysis.

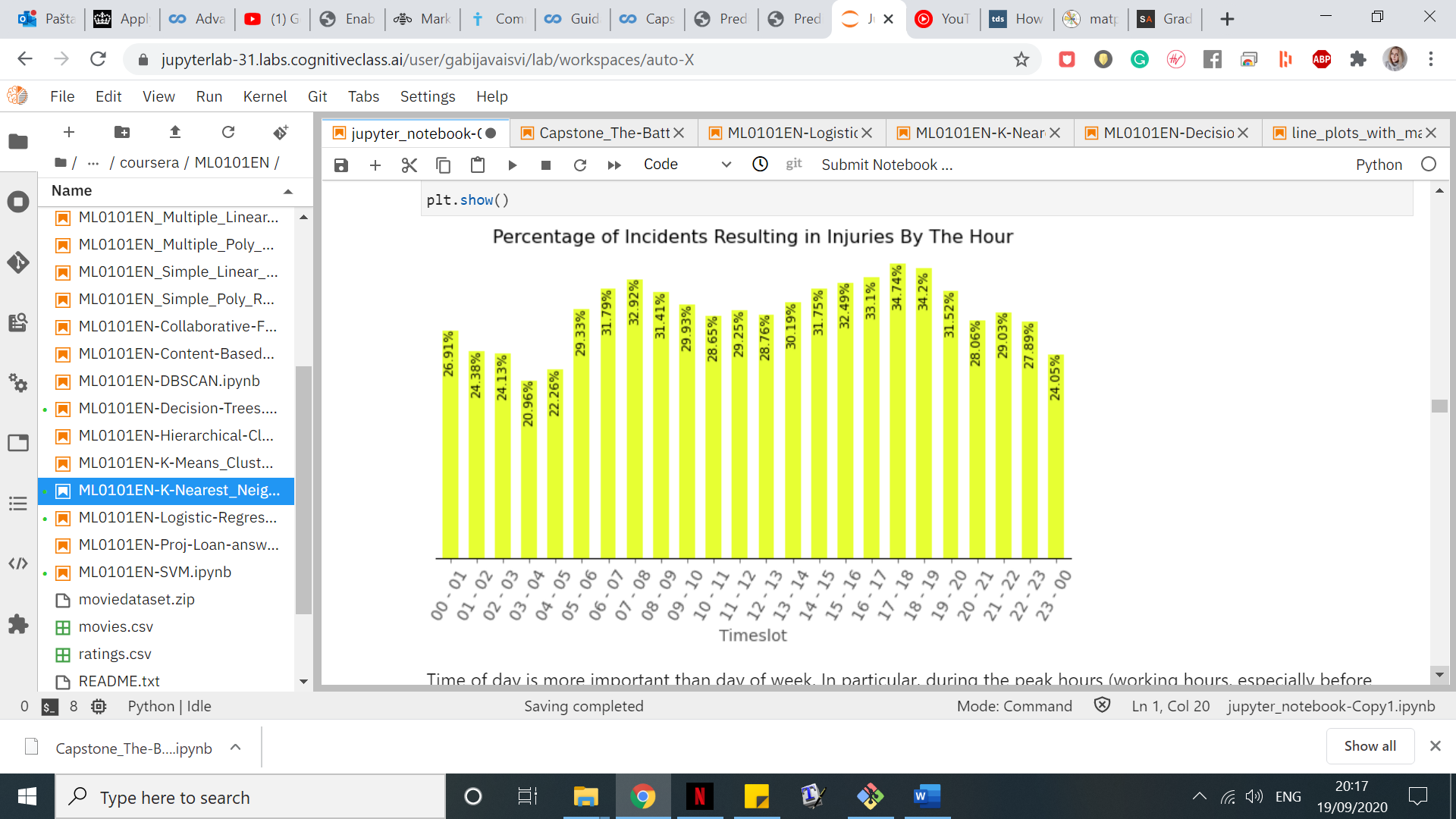


We checked the correlation between severity, month, day of the week and hour of the accident.



Seeing that month has almost zero correlation with the severity of the accident we only plotted graphs for day of the week and hour of the accident.

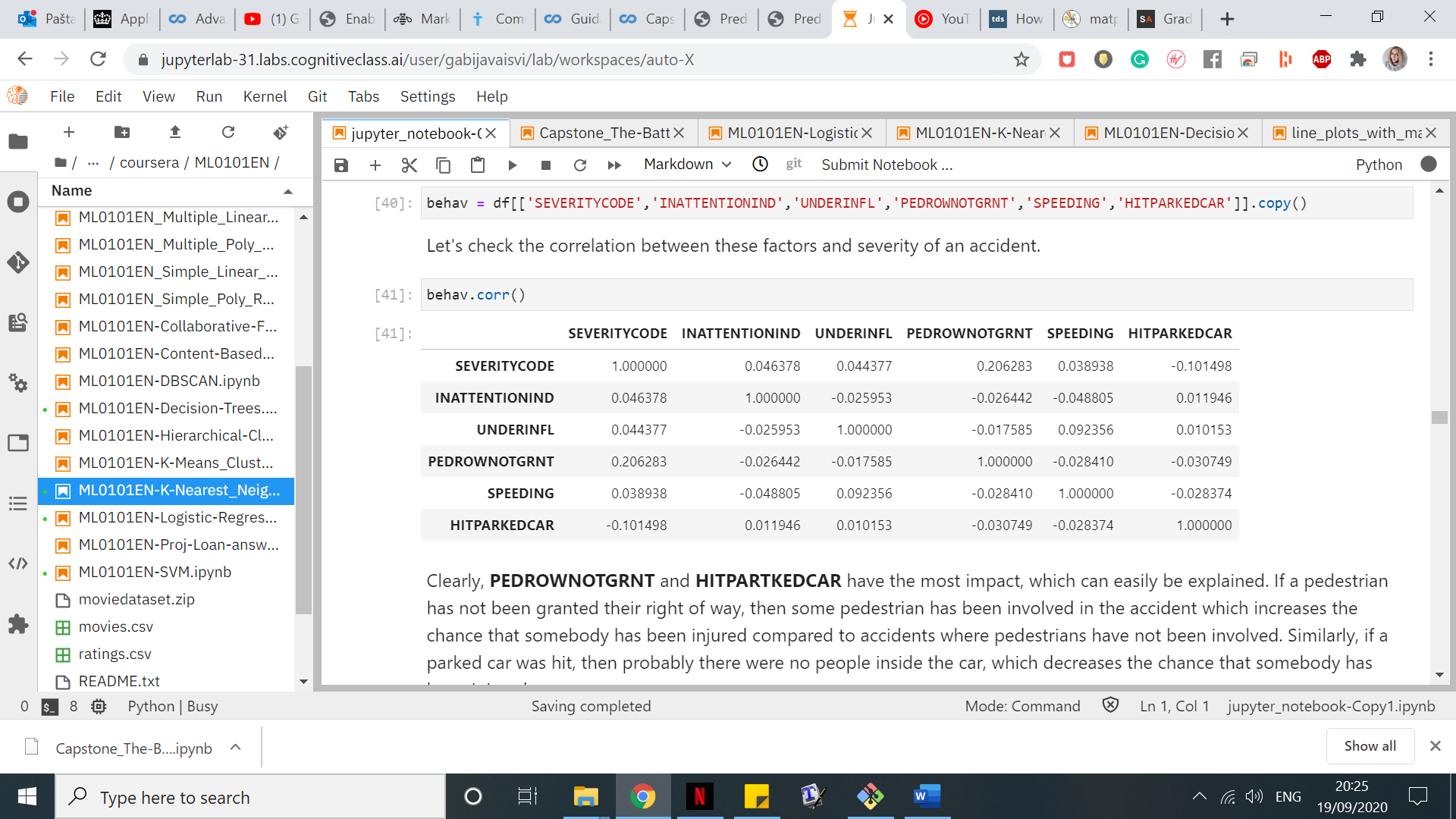




From the line graph we can deduce that day of the week is not a very relevant factor in our study, but we can still see that on Sundays there are around 2% fewer incidents resulting in injuries. Time of day is more important than day of week. In particular, during the working hours, especially before and after work i.e. around 8A.M. and after 5P.M. the accidents resulting in injuries are more common. In fact, they are 10%-15% more likely than during the calmest hours, such as 3A.M. Beforing continuing with our study we updated our dataframe by deleting column INCDATE (as we have already extracted any relevant information from it) and adding columns TIMESLOT and DAYOFWEEK.

## Behaviour

In this part we will analyse the impact of driver’s actions. We will examine how the attention of the driver **INATTENTIONIND**, being under influence while driving **UNDERINFL**, not granting the pedestrian their right of way **PEDROWNOTGRNT**, speeding **SPEEDING** and hitting a parked car **HITPARKEDCAR** affects the severity of an accident. We started with looking at correlation.



Clearly, **PEDROWNOTGRNT** and **HITPARTKEDCAR** have the most impact, which can easily be explained. If a pedestrian has not been granted their right of way, then some pedestrian has been involved in the accident which increases the chance that somebody has been injured compared to accidents where pedestrians have not been involved. Similarly, if a parked car was hit, then probably there were no people inside the car, which decreases the chance that somebody has been injured. We create a plot to gain better insight.

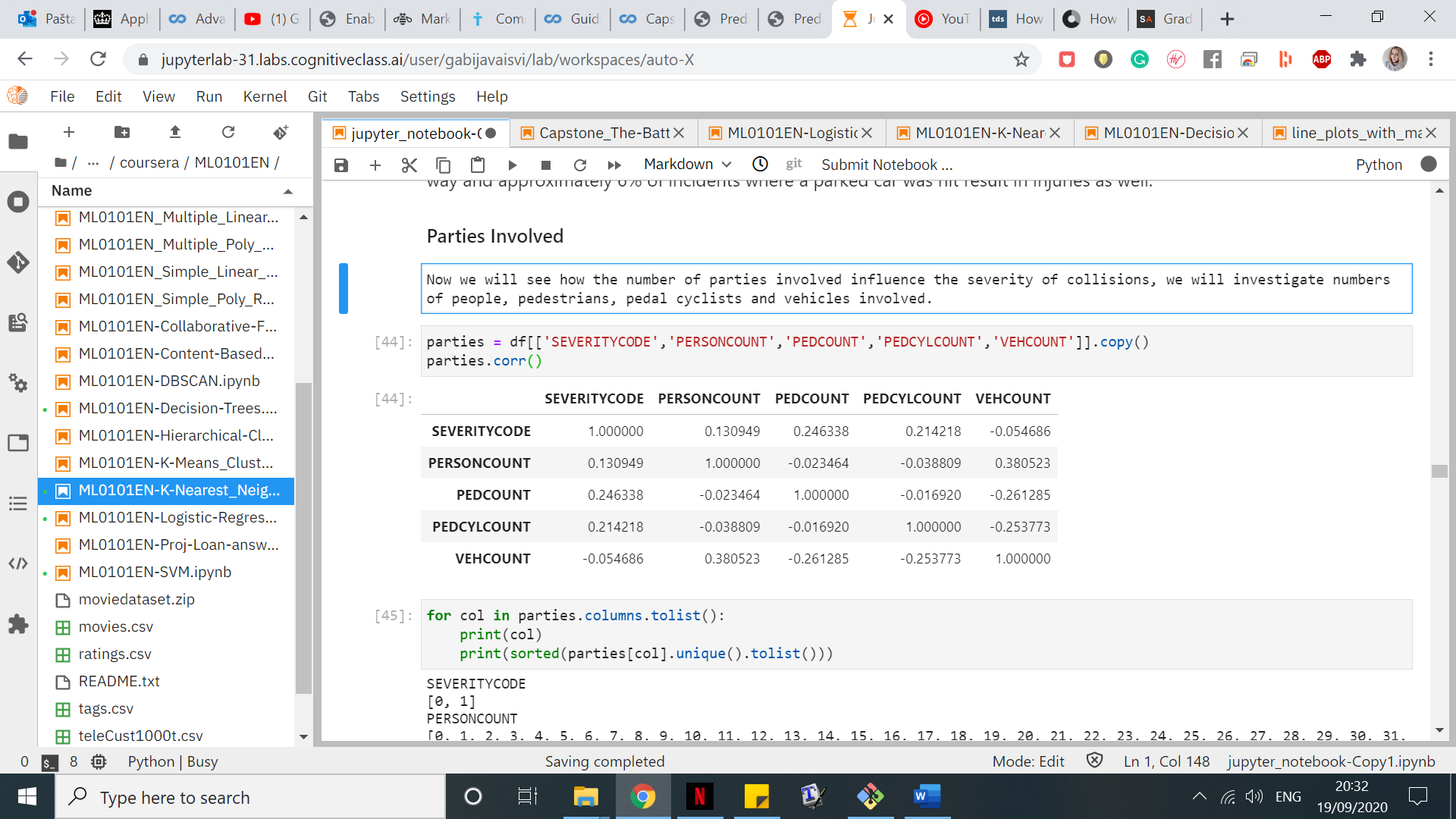
A screenshot of a cell phone

Description automatically generated

From around 35% to 40% incidents involving inattention, being under influence of alcohol or drugs and speeding result in injuries. We also know that around 90% of incidents where pedestrian was not granted their right of way and approximately 6% of incidents where a parked car was hit result in injuries as well.

## Parties involved

Now we will see how the number of parties involved influence the severity of collisions, we will investigate numbers of people, pedestrians, pedal cyclists and vehicles involved. First, we look at correlation. We can see that there is comparatively high correlation between the severity of an incident and the number of pedestrians/pedal cyclists involved, and a small negative correlation between the severity of an incident and the number of vehicles involved.



As incidents involving high numbers of people, pedestrians, pedal cyclists or vehicles are rare, we do not have a lot of data with them thus group such incidents as follows. We will also use it for modelling, as otherwise our model might be overfitting. Thus, if **PERSONCOUNT** is greater than 7 we change it to ‘8+’, if **PEDCOUNT** is greater than 1 we change it to ‘2+’ and if **VEHCOUNT** is greater than 4 we change it to ‘5+’ and plot some graphs.

A screenshot of a cell phone

Description automatically generated

The plots tell us that whenever there are more than 2 people involved, the likelihood of a collision resulting in injuries is increasing as the number of people increases. A collision involving any number of pedestrians or pedalcyclists has at least 87% chance to result in injuries. The last plot tells us that an incident involving one vehicle is more than 50% likely to result in injury (which makes sense, if there was only one vehicle involved, then there was also either an object or a person involved). Incidents involving none or two vehicles are the least likely to result in injuries, however, the chance of an incident involving more than 2 vehicles resulting in injuries increases the number of vehicles involved goes up.

As information from **PEDCYLCOUNT** and **PEDCOUNT** attributes coincide, instead of having both attributes we will keep only **PEDCOUNT** where each entry will be equal to the sum of pedal cyclists and pedestrians involved.

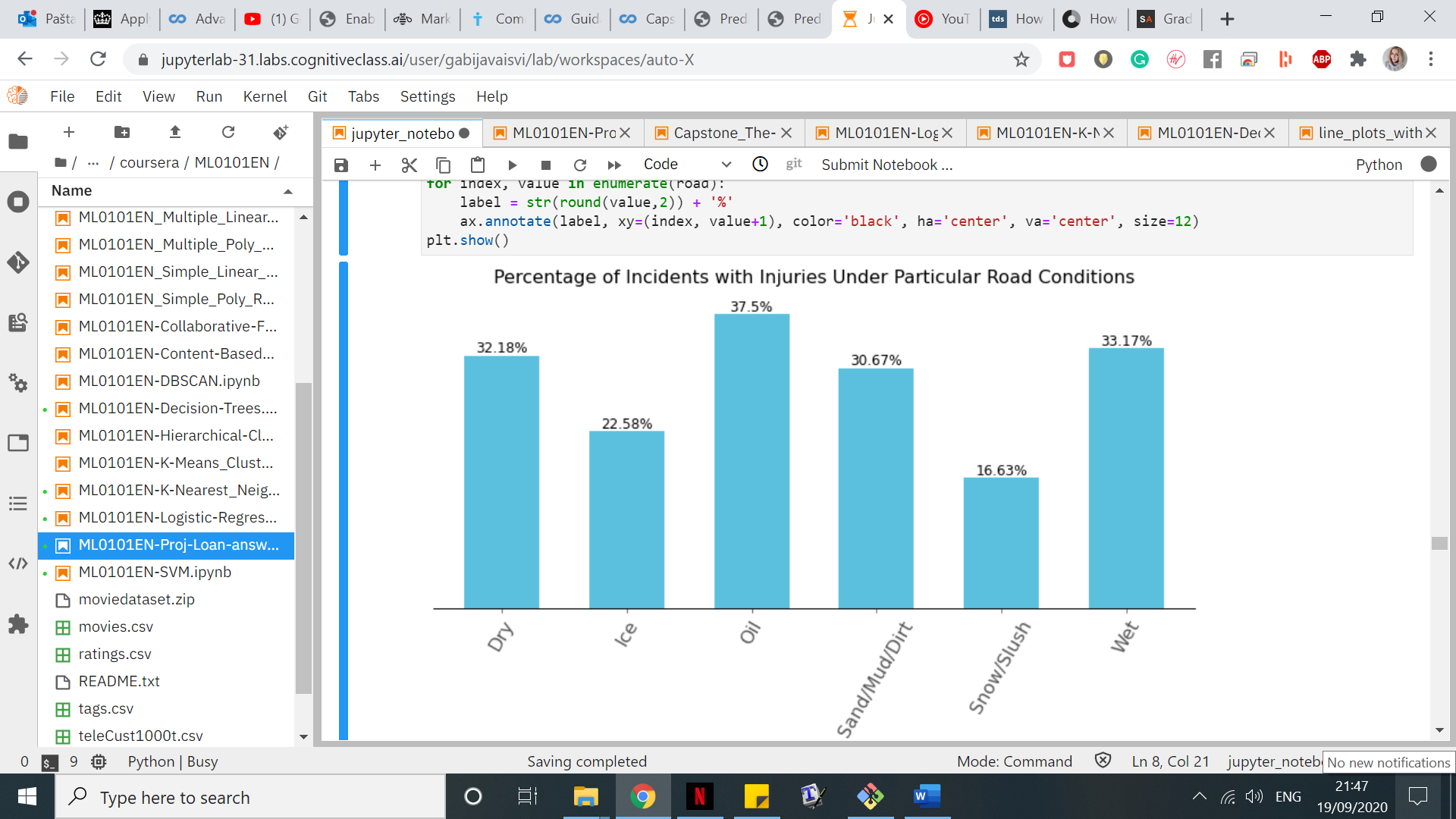
## Weather, Road and Light Conditions

To analyse weather, road and light conditions we plot some graphs as well.

A screenshot of a cell phone

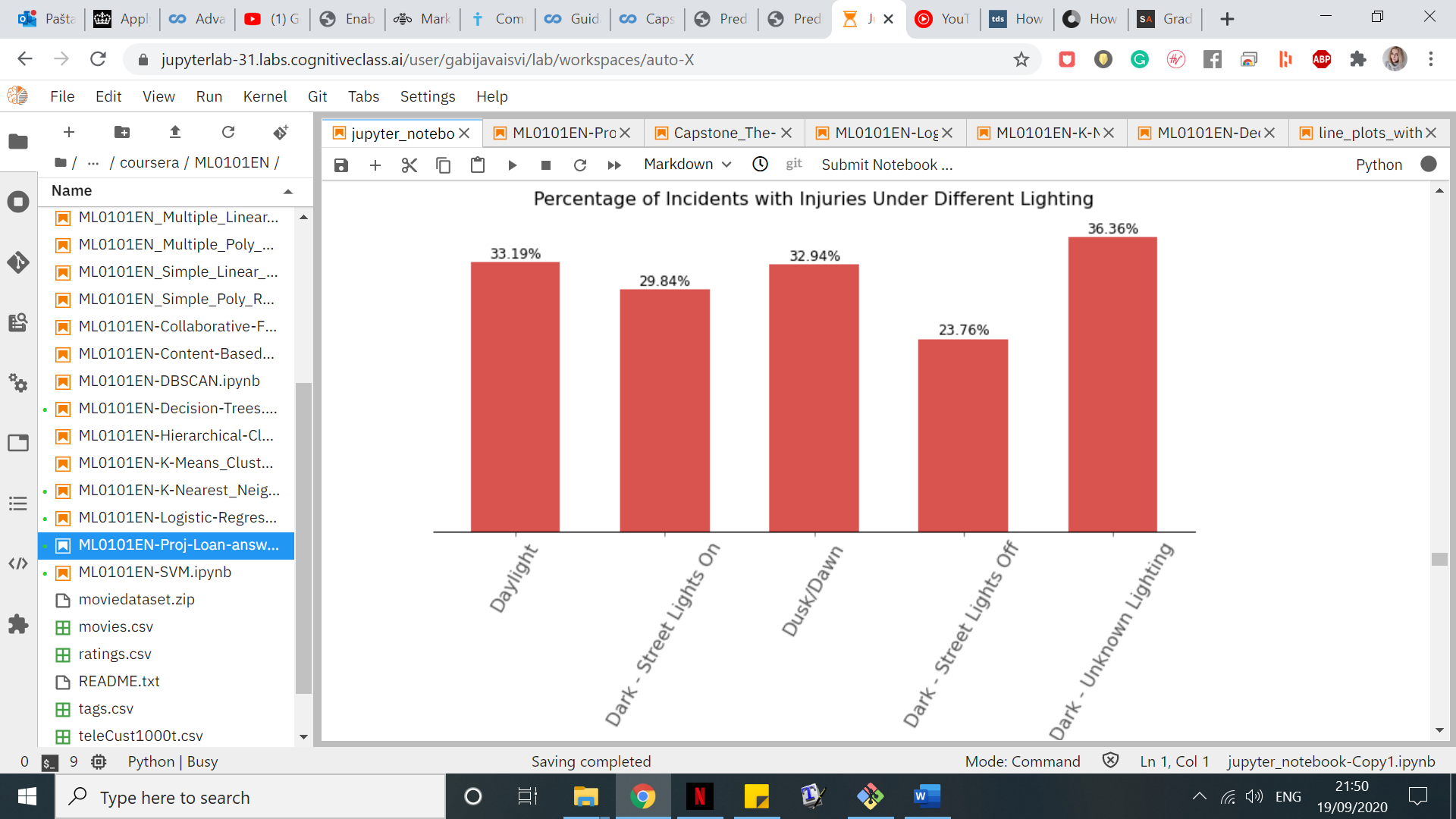
Description automatically generated

We can deduce that more incidents result in injuries when the weather is clear, cloudy, there is fog, smog or smoke or it is raining, whereas when it is blowing sand or dirt, there is severe crosswind, sleet, hail, freezing rain or it is snowing we have much fewer incidents resulting in injuries.



Regarding the road conditions oil on the road highly increases the possibility of getting injured if a car accident is to happen. Injuries are also often when the road is dry, wet or there is sand, dirt or mud on the road. However, incidents result in injuries much more rarely when it is snow, slush or ice on the road. This supports our earlier results from investigating weather.

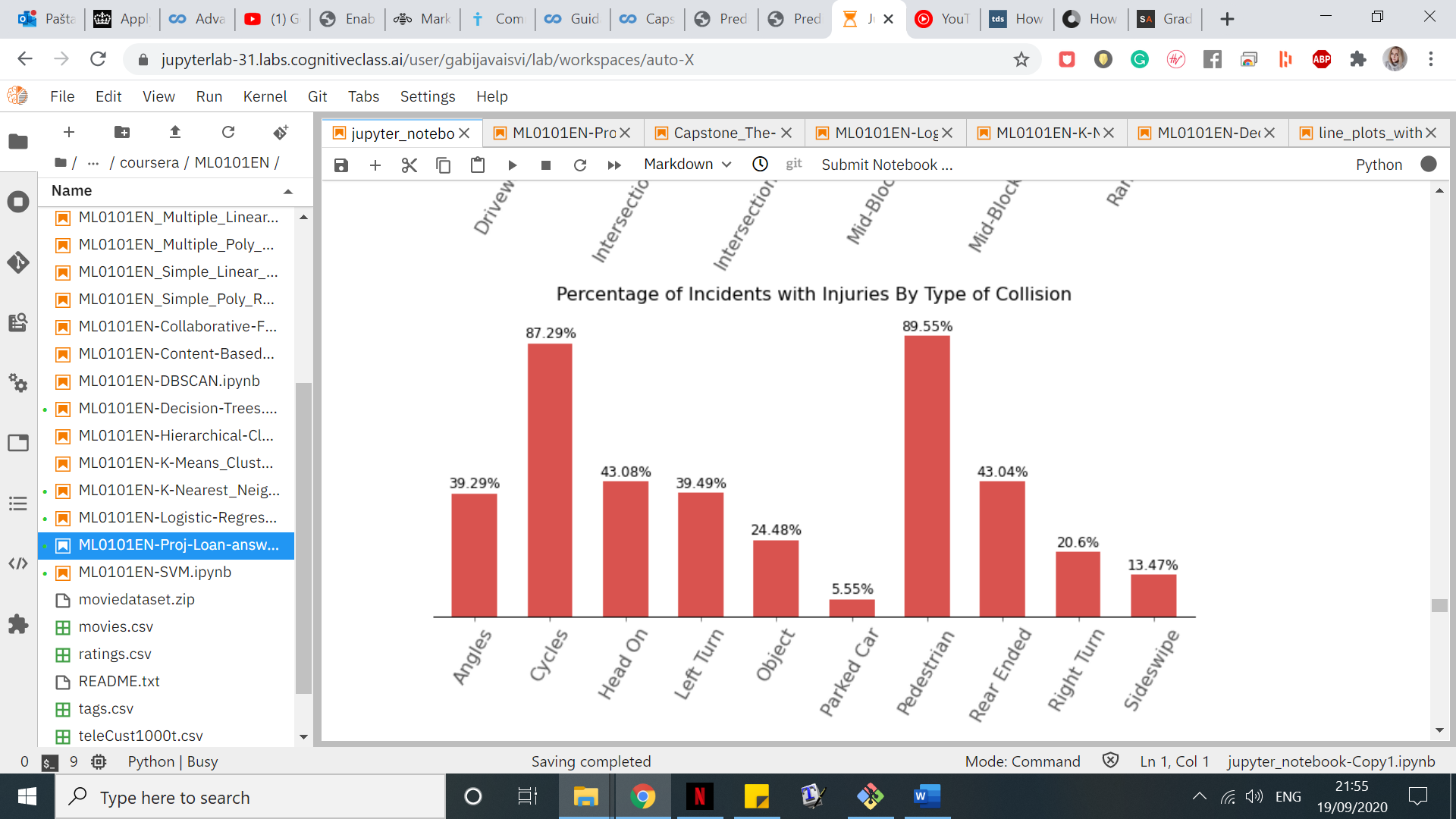
Generally, except for oil on the road, we can see that fewer incidents result in injuries under extreme conditions, such as blowing sand or dirt, severe crosswind, sleet, hail, freezing rain or snowing; ice, snow or slush on the road. It is possible that under such (more extreme and dangerous) conditions people are simply more concentrated and cautious, pay more attention and drive more slowly than they normally would.

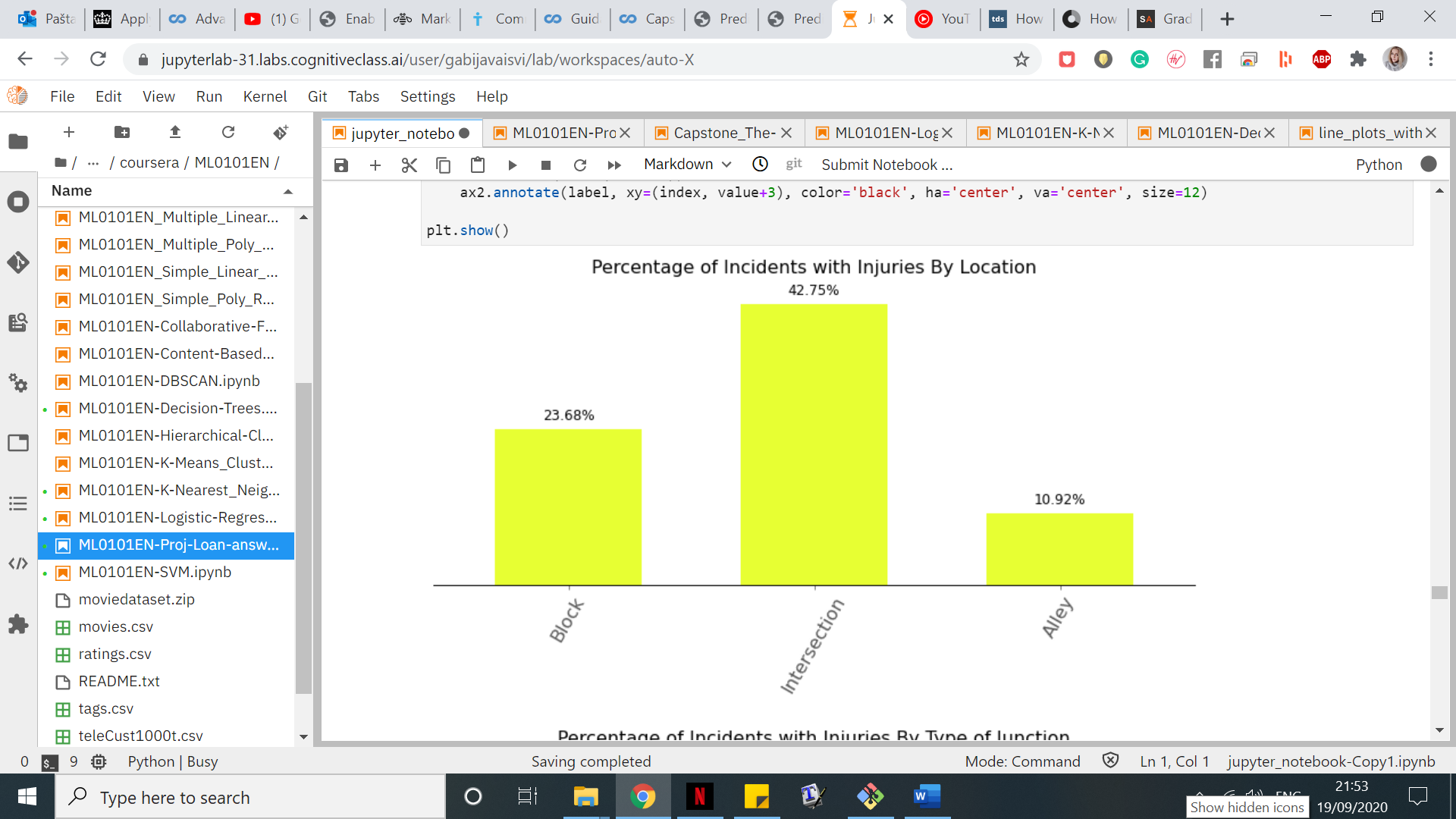


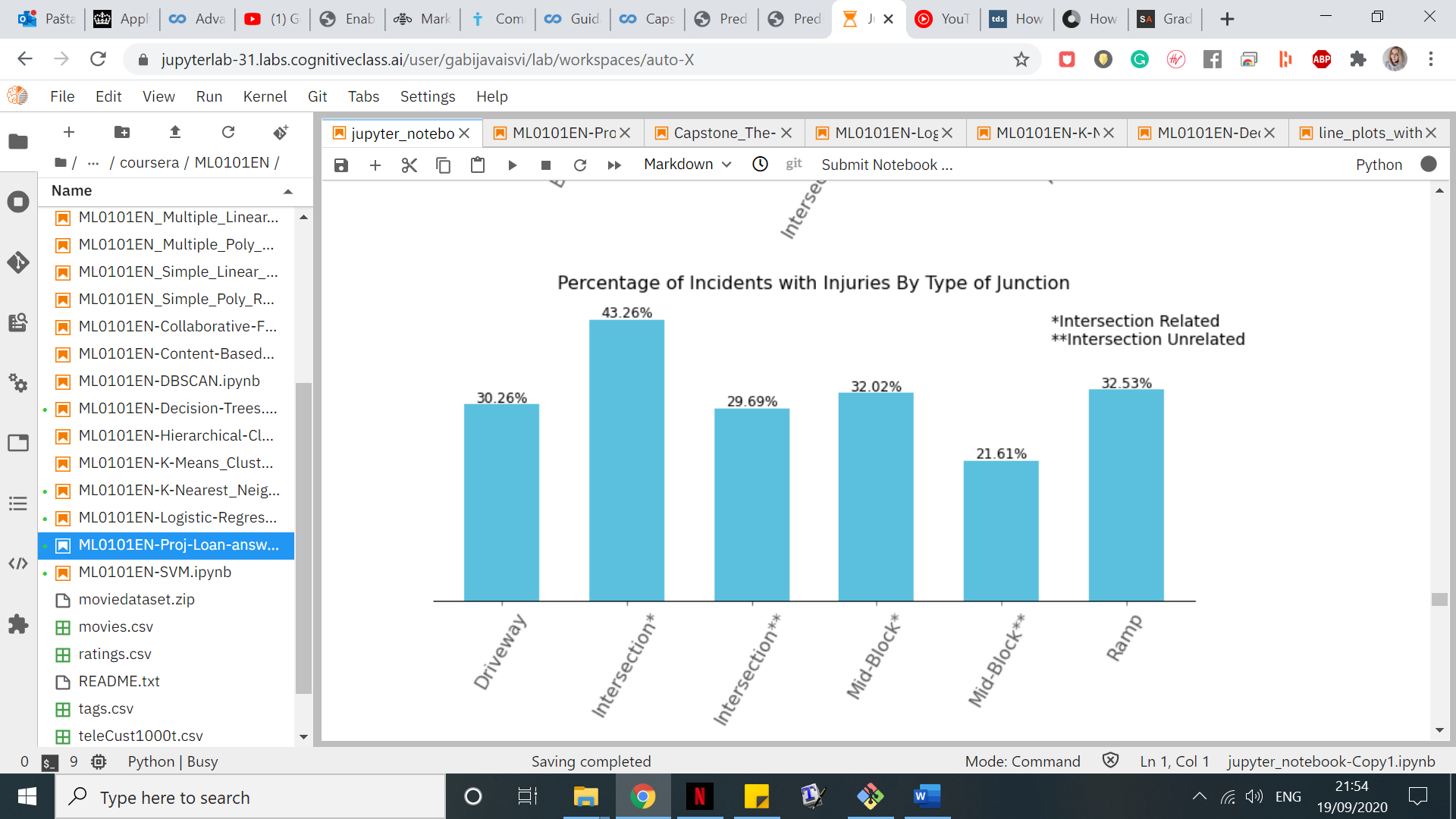
Data about light conditions do not tell us much: it seems like there are fewer severe incidents during the dark regardless whether the street lights are on or off, however, the last bar shows the exact opposite - the highest percentage of severe incidents is during the dark when the lighting is unknown. Observe that we can state with confidence that the percentages of severe incidents during daylight, dusk and dawn do not differ much from the percentage of all severe incidents where the light conditions are recorded (i.e. 32.11%). This means that the percentage of severe incidents when it is dark (regardless whether lighting is unknown, streetlights are on or off) should be similar. Thus, we can deduce that lighting is not important to our case study and drop this feature from our data frame.

## Location

Finally, using the help of graphs we investigate how important location is to our study.







Using these graphs, we found out that incidents happening at intersection and their cause being intersection related are most likely to end up in injuries. The intersection unrelated incidents happening in alleys or blocks are have the slimmest chance of resulting in injuries.  
Predictably, as we have seen when investigating the effect of pedestrians and pedal cyclists involved and whether a parked car was hit or a pedestrian involved, the last plot also shows the same results: incidents involving pedestrians or/and pedal cyclists are extremely likely to result in injuries, whereas incidents where a parked car was hit have a very low chance of resulting in injuries. We have also found out that if the collision happened with an object or it was a sideswipe, then it is more likely to end up just with property damage. Interestingly, collision happening when turning left instead of right have a twice higher chance that somebody gets injured!

# Modeling

For better accuracy we keep a lot of features, namely: **SEVERITYCODE** (aim of the case study), **ADDRTYPE**, **JUNCTIONTYPE**, **COLLISIONTYPE**, **SDOT\_COLCODE**, **ST\_COLCODE**, **PERSONCOUNT**, **PEDCOUNT**, **VEHCOUNT, UNDERINFL**, **PEDROWNOTGRNT**, **SPEEDING**, **INATTENTIONIND**, **HITPARKEDCAR**, **WEATHER**, **ROADCOND**, **DAYOFWEEK**, **TIMESLOTS**. We replace all remaining null values with ‘Unknown’ and convert all data to numerical values.

We pre-process data in two ways: we use MinMaxScaler for Gradient Boosting Classifier and XGBoost and StandardScaler for Decision Tree and Logistic Regression. We also split our data into training and testing sets in order to train our models better. Because our data is unbalanced i.e., we have 29.89% (not 50%) of all collisions resulting in injuries, we will have to scale our models. For Gradient Boosting Classifier, Decision Tree and Logistic Regression we will use sample\_weight function and for XGBoost we will use scale\_pos\_weight function. It is important to understand that without this scaling we would get higher overall accuracy scores, but then the percentage of collisions resulting in injuries predicted correctly would be much lower than we want. We need to predict collisions resulting in injuries as accurately as possible, as our goal is to prevent them i.e. it is better for us to have a lower percentage of accurately predicted incidents resulting in property damage only, but a higher percentage of accurately predicted collisions resulting in injuries. We create all our models and make predictions with them.

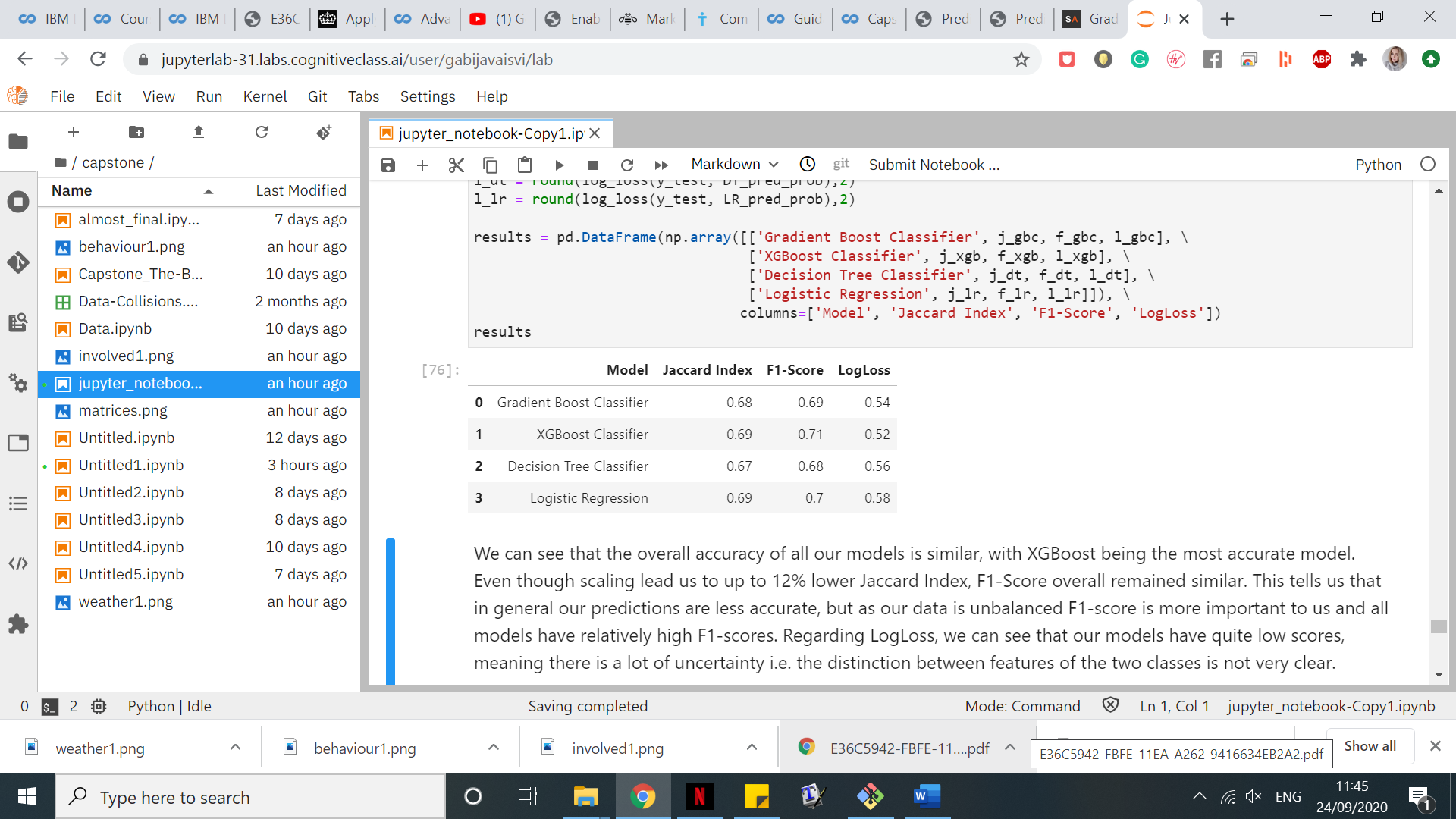
# Results

We will compute accuracies for all machine learning models that we trained and analyze our models using confusion matrices. A confusion matrix consists of percentages of accidents with injuries predicted correctly (true positive TP), accidents with injuries predicted incorrectly (false negative FN), accidents without injuries predicted correctly (true negative TN) and accidents without injuries predicted incorrectly (false positive FP). For accuracy evaluation we will also use:

* *Jaccard Index*, which corresponds to the size of the intersection of the set of our predicted values and the set of actual values divided by the size of the union of these two sets. We use function *accuracy\_score* as in binary and multiclass classification this function is equal to the *jaccard\_score* function.
* *F1-score*, which is the harmonic average of the precision and recall. Precision is a measure of accuracy given that a class label has been predicted, whilst recall is true positive rate.
* *Logarithmic loss*, which measures the performance of a classifier where the predicted output is a probability value between 0 and 1, as our models can return the probability of the class label predicted correctly.

*Jaccard index* and *F1-score* take values between 0 and 1, with 0 being the worst value and 1 being the best, meanwhile *LogLoss* takes values between 0 and 1 with 0 being the best value and 1 being the worst.

And these are our results:



We can see that the overall accuracy of all our models is similar, with XGBoost being the most accurate model. Even though scaling lead us to up to 12% lower Jaccard Index, F1-Score overall remained similar. This tells us that in general our predictions are less accurate, but as our data is unbalanced F1-score is more important to us and all models have relatively high F1-scores. Regarding LogLoss, we can see that our models have quite low scores, meaning there is a lot of uncertainty i.e. the distinction between features of the two classes is not very clear.

Interestingly, even though Logistic Regression is mostly used for predicting continuous values, according to accuracy measures we can see that in our classification problem it has performed quite well. The worst performing model seems to be the Decision Tree. Now we can take a look at confusion matrices for all our models and examine them to understand better how well our models perform.

A screenshot of a cell phone

Description automatically generated

We can see a clearer picture of what is going on. As we assumed in the beginning, the Logistic Regression performs the worst in our problem i.e. it has the lowest rate of correctly predicting collisions resulting in injuries. In fact, 33% of the time it fails to predict which collisions will result in injuries. Not only that, 30% of the time it predicts that a collision will result in injury, even though it will not! In comparison, Decision Tree is slightly better and yet has similar flaws. Using these models could lead to too many resources spent on preventing severe collisions when in fact a high percentage of them would result in property damage only, and quite high percentage of them would be misclassified as property damage.

Gradient Boosting and XGBoost classifiers are the best for our problem, they both relatively accurately predict which collisions will result in injuries and have a comparatively small chance of failing to predict that a collision will result in property damage only. Out of these two, XGBoost is slightly more accurate. Thus, we conclude that XGBoost is the best classifier amongst the ones we tried.

# Discussion & Conclusions

In this project we have analysed the given data of collisions in Seattle between 2014 and now. We were tasked to find out which factors are the most influential on severity of a collision and create a machine learning model which predicts the severity of a collision. We have found out that more severe collisions tend to happen during peak hours when people go to/from work, that is, around 8A.M. and 5P.M. Collisions occurring at intersections or mid-block, but intersection related are also likely to be severe. If there are pedestrians and/or pedal cyclists involved, the collision is highly likely to result in injuries. We have created several machine learning models which predict whether the collision will result in injuries or property damage only, evaluated them and concluded that XGBoost was the best classifier in our case study.

Our study results could helpful in preventing more severe accidents, which potentially could be done by placing warning signs in certain intersections and roads, as well as reminders to be careful of pedestrians and cyclists and to not rush to work or home after it (better late, than hurt), educating pedestrians and cyclists to behave more carefully on the road, and providing information regarding weather and road conditions.