**An analysis of car accidents in Seattle**

***Coursera Capstone***

**Introduction/Business Problem**

The highest price we pay for car crashes is in the loss of human lives. However, society also bears the brunt of the many costs associated with motor vehicle accidents. According to the National Highway Traffic Safety Administration (NHTSA), U.S. motor vehicle crashes in 2010 cost almost $1 trillion in loss of productivity and loss of life.

Data scientists can contribute to the debate of how to reduce the severity of car accidents by contributing their opinions based on extensive data analysis of traffic incident datasets. Via exploratory analysis, data scientists have the potential to uncover hidden patterns. One human life that is lost in a traffic incident is one too many, hence the importance of making sure the data community contributes to the debate.

Data scientists can employ Machine Learning (ML) techniques to predict the severity of an accident. Machine Learning is, at its most basic, the practice of using algorithms to parse data, learn from it, and then make a determination or prediction about something in the world. Given the existence of multiple datasets that incorporate data about traffic incidents, we can leverage this to apply machine learning and other data analysis techniques to the problem of traffic incidents.

The City of Seattle has a publicly available dataset on traffic incidents in the city, which we will use for our analysis. Our end-goal is to be able to understand the factors that increase the likelihood of a traffic incident in Seattle, and whether we can predict the severity of an incident based on a series of attributes pertaining to the incident.

**The target audience** for this project is the local government of the City of Seattle, in addition to local organisations from different neighbourhoods of Seattle who can understand how to make their neighbourhoods safer.

**Data**[**¶**](https://render.githubusercontent.com/view/ipynb?commit=bb20f6e7547417405db6aabe640bcfe82a8adcc7&enc_url=68747470733a2f2f7261772e67697468756275736572636f6e74656e742e636f6d2f617375616d61742f436f7572736572615f43617073746f6e652f626232306636653735343734313734303564623661616265363430626366653832613861646363372f5765656b2532303225323043617073746f6e652532302831292e6970796e62&nwo=asuamat%2FCoursera_Capstone&path=Week+2+Capstone+%281%29.ipynb&repository_id=291156226&repository_type=Repository#Data)

Our dataset can be accessed via the following link <https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv>, and is made up of 38 columns and 194673 rows.

Therefore, we have 38 attributes to play with in our analysis, and 194673 instances of traffic incidents.

At a quick glance, attributes such as the following might prove useful throughout our analysis:

* **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

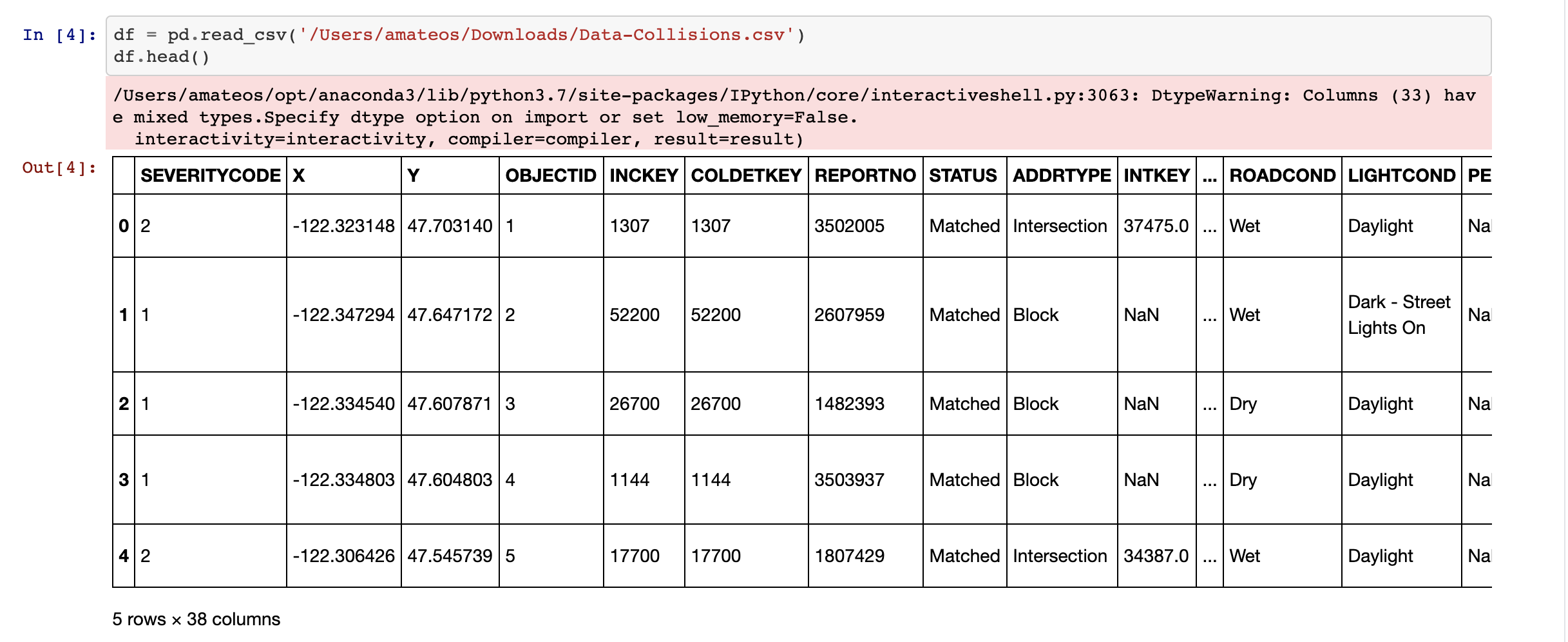
* **LOCATION:** Description of the general location of the collision (street name and number)
* **SEVERITYCODE:** A code that corresponds to the severity of the collision:
* - 3—fatality
* - 2b—serious injury
* - 2—injury
* - 1—prop damage

- 0—unknown

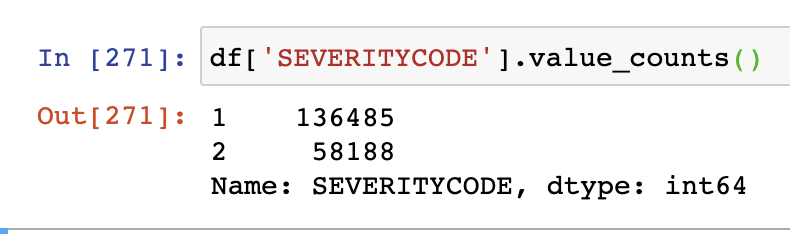
* **PERSONCOUNT:** The total number of people involved in the collision
* **INCDATE:** The date of the incident.

The **label** for the data set is severity, which describes the fatality of an accident and which in our dataset is given by **SEVERITYCODE**.

Upon initial visualisation of the dataframe, we found that many columns contained NaN values:

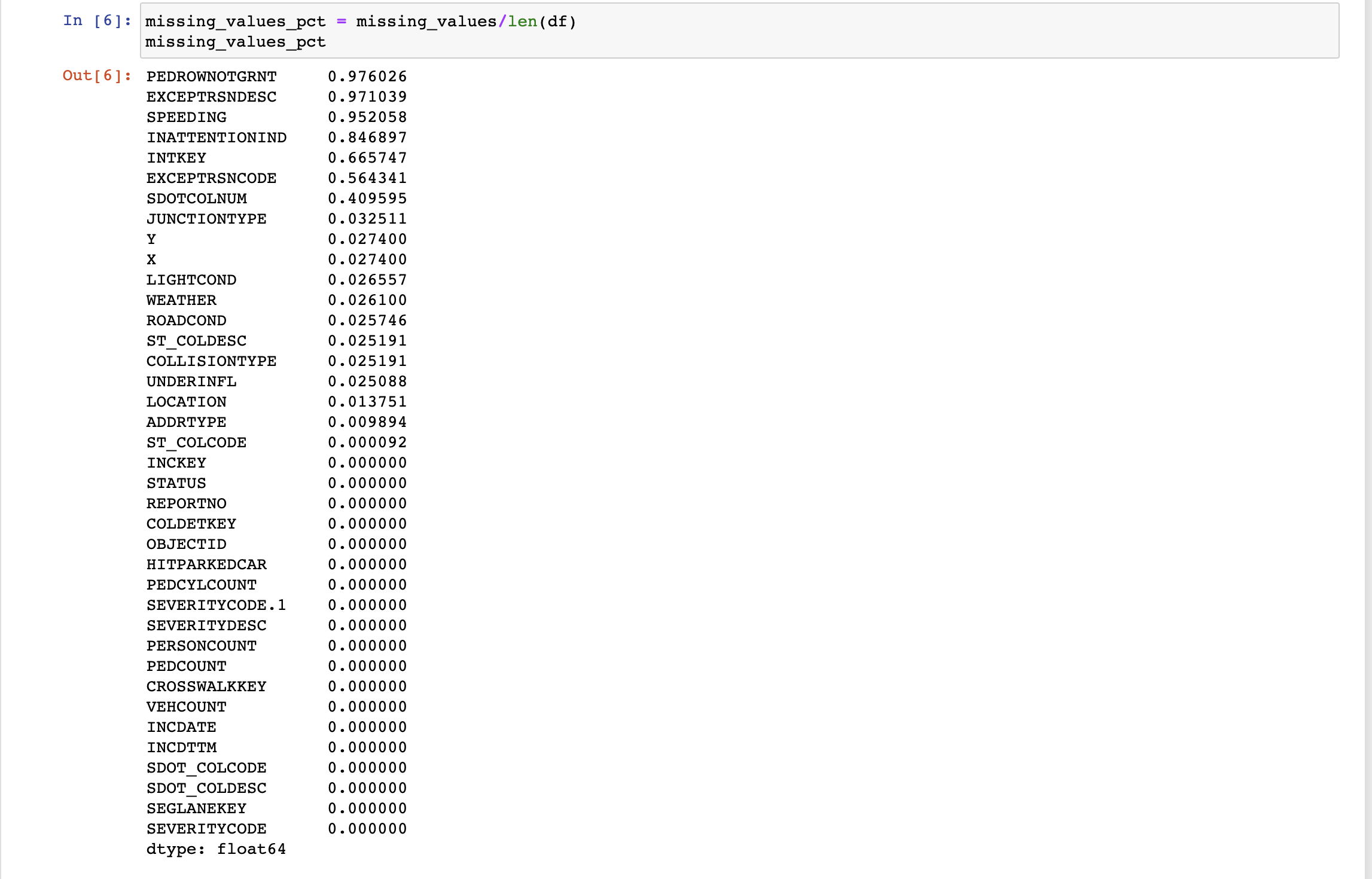


To identify the different possible values for **SEVERITYCODE**, we ran the following piece of code and identified that our dataset only contains accidents of **SEVERITYCODE = 1** and of **SEVERITYCODE = 2.**



An accident of **SEVERITYCODE = 1** is one involving property damage whereas an accident of **SEVERITYCODE = 2** involves injury to one of the implicated persons.

We found that 7 of the columns in our initial dataset had missing values in more than 40% of their rows so we proceeded to drop these.



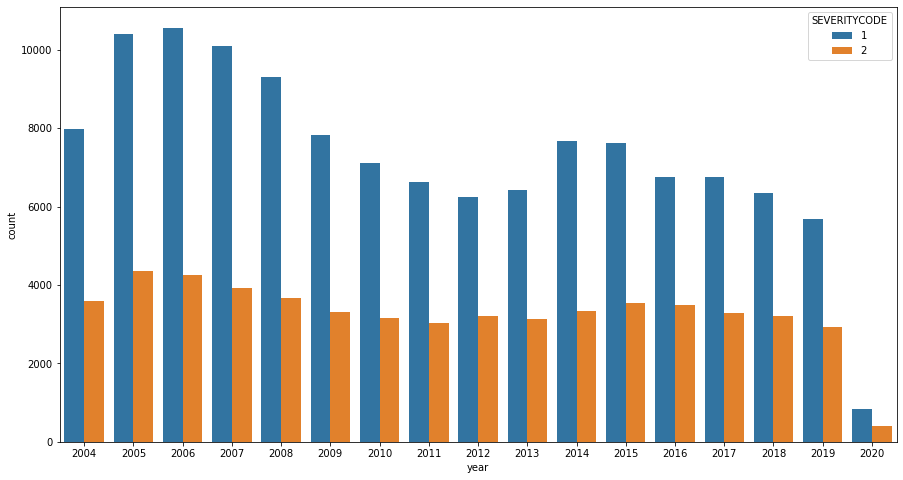
To further simplify our dataframe, we dropped those columns which did not provide useful information about the accident. These were:

* OBJECTID: ESRI unique identifier
* INCKEY: A unique key for the incident
* COLDETKEY: Secondary key for the incident
* REPORTNO: (number of the report? - no description in metadata)
* STATUS: (status? - no description in metadata)
* LOCATION: Description of the general location of the collision - We have latitude and longitude which can give us more information than what LOCATION variable does
* SEVERITYCODE.1: same info as SEVERITYCODE
* SDOT\_COLCODE: A code given to the collision by SDOT.
* SEGLANEKEY: A key for the lane segment in which the collision occurred.
* CROSSWALKKEY: A key for the crosswalk at which the collision occurred.
* JUNCTIONTYPE gives the same information as ADDRTYPE so we can drop it
* SEVERITYDESC gives the same information as SEVERITYCODE so we can drop it
* ST\_COLDESC gives the same information as COLLISIONTYPE so we can drop it
* HITPARKEDCAR gives the same information as COLLISIONTYPE so we can drop it

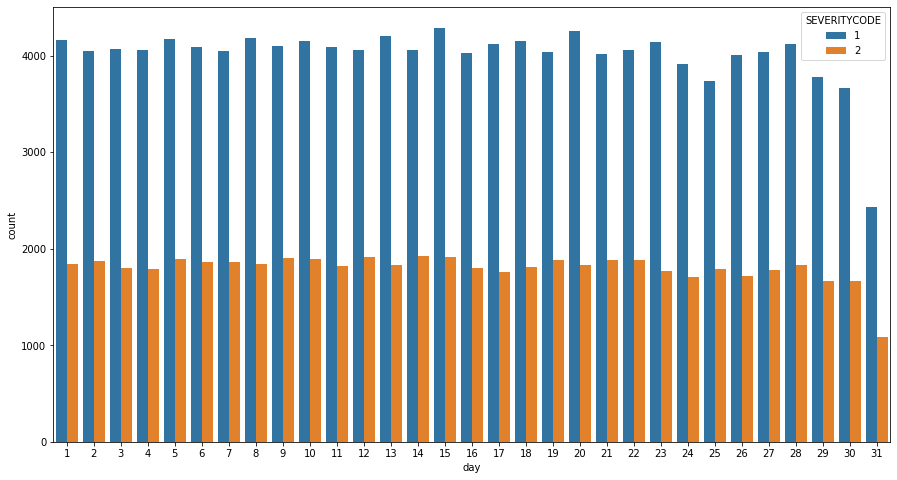
We also proceeded to drop all rows with at least one NaN value.

From our column **INCDTTM** we were able to extract information about the year, month, day, day of the week and hour on which the accident happened using the ***datetime*** library.

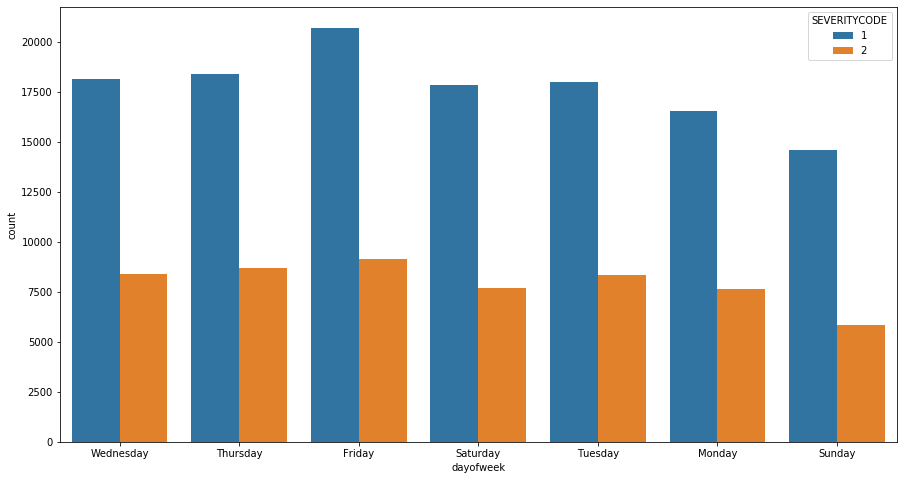
With our dataset ready for Exploratory Data Analysis, we leveraged the ***seaborn*** and ***matplotlib*** libraries to visualise our data.



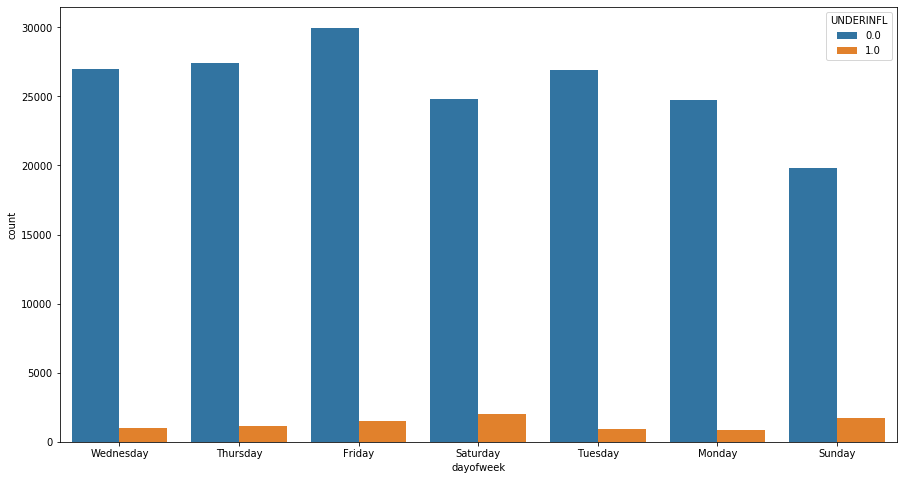
Accidents peaked in 2006 in Seattle and steadily decreased until 2012, after which it rebounded slightly. Throughout the whole timeframe **SEVERITYCODE = 2** accidents remained a minority.



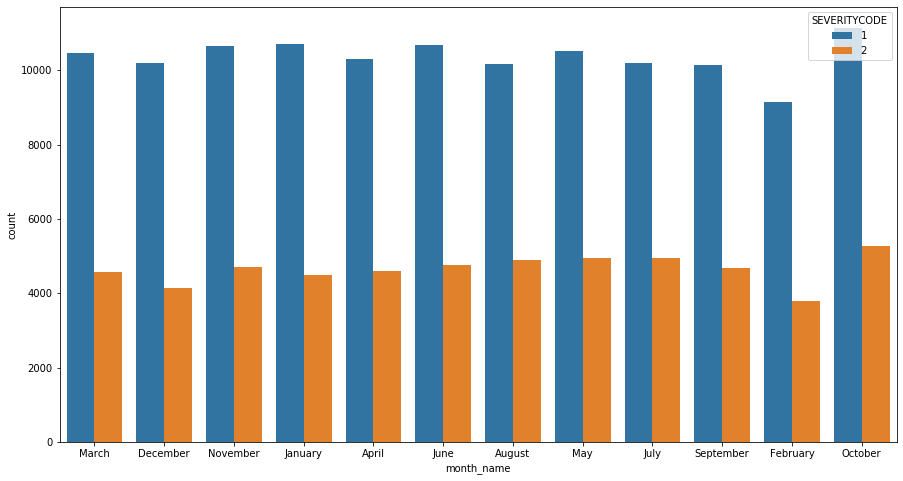
What about the day of the week? Not much variation is found when looking at the day of the month. Only notable is the decrease in the number of accidents that happen on the 29th, 30th and 31st of each month. However, this makes sense given that these days always tally 0 accidents each February and that many months do not have a 31st day.



When we look at the day of the week, we observe that Fridays is the day where more accidents happen and when more **SEVERITYCODE = 2** accidents happen. This is not unsurprising as Fridays signal the start of the weekend and mark the start of a lot of leisure activities - many involving drinking alcohol – and a lot of activity and travel from and towards the city. Also unsurprising is to see that Sundays are the day of the week where least accidents tend to happen as this is a day generally seen as “quiet” in the Western world.



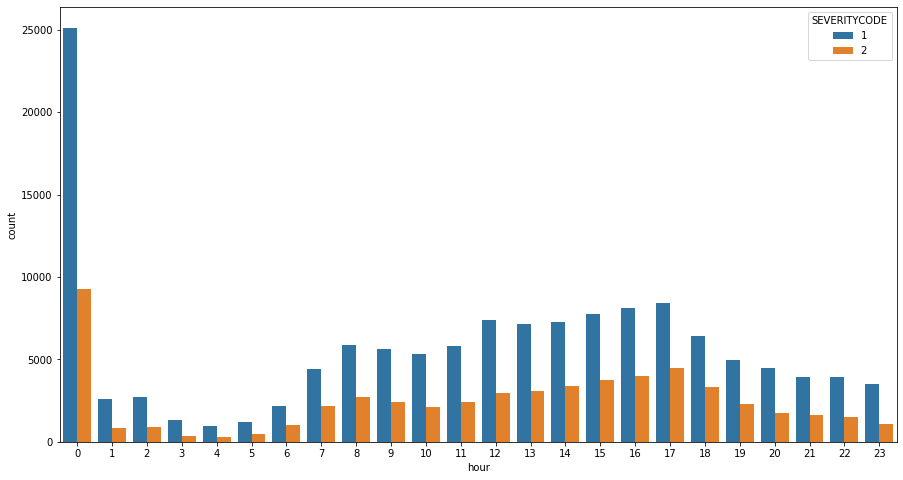
Predictably we can see that Saturdays and Fridays are when most accidents UNDERINFL happen, that is accidents where the driver is under the influence of alcohol.



Finally, we can see that October has the lion’s share of accident both in overall volume and in the volume of **SEVERITYCODE = 2** accidents. It is unclear why but we can hypothesise that this is because October tends to be the first month where it gets markedly colder, there is less light and features the event of Halloween which is one of the most important festivities in the US. As seen by the below image the last days of the month – 28, 30, 31 – are the days when most of the accidents happen and when Halloween parties are most prevalent.

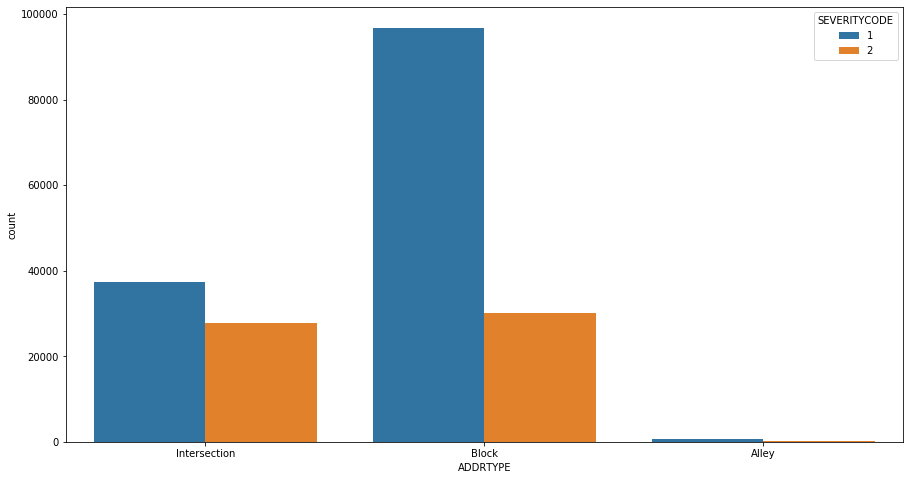


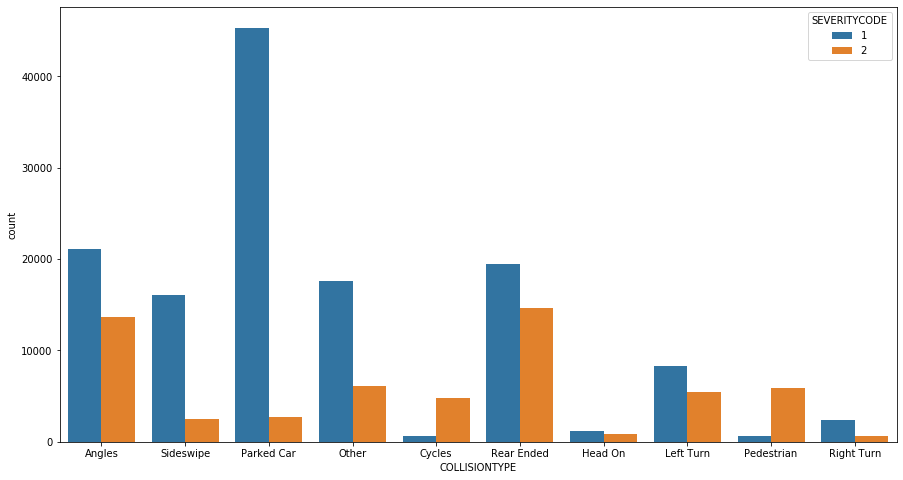
If we extend our analysis to look at the hour of the day when these accidents are most likely to happen, the number of accidents that happen at midnight immediately stands out. Accidents that occur in the 00:00 – 00:59 timeframe are more than 2.5x likelier to occur than the second-most accident “prone” hour of the day, the 17:00 – 17:59 period which is generally the time at which a lot of people end their day of work and return home or go elsewhere.

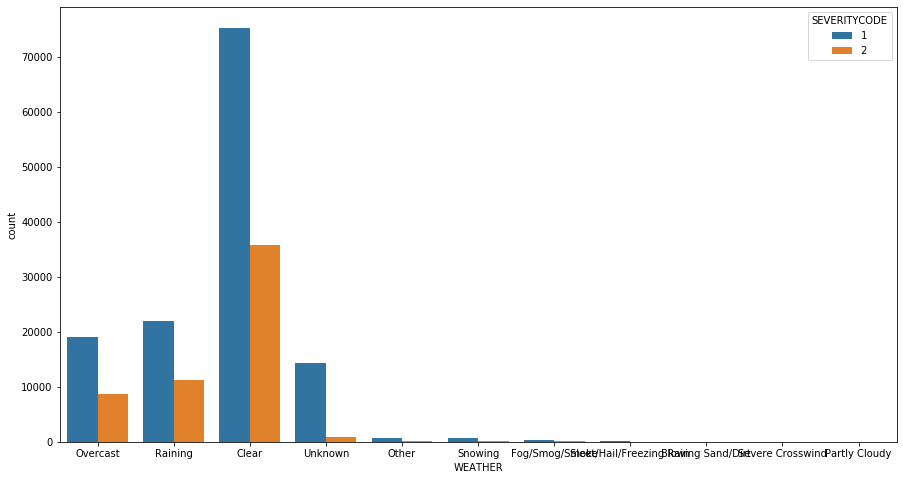


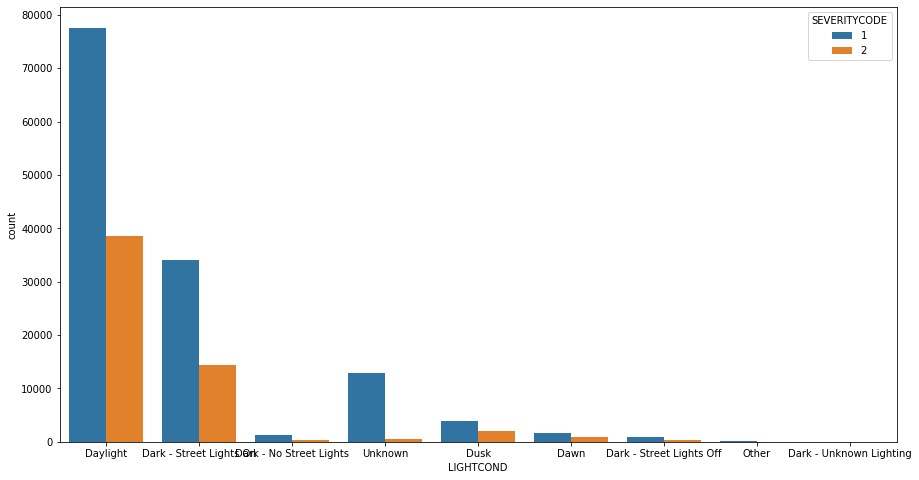
Other interesting visual analysis enabled us to find that:

* Most accidents happen at a block but those that do occur at an intersection tend to be more severe.
* Most accidents involved a parked car. However, the most severe accidents tended to be those involving a pedestrian, those happening on a left turn, on the rear end and at angles.
* Most accidents happened on clear days.
* Most accidents happened in daylight.

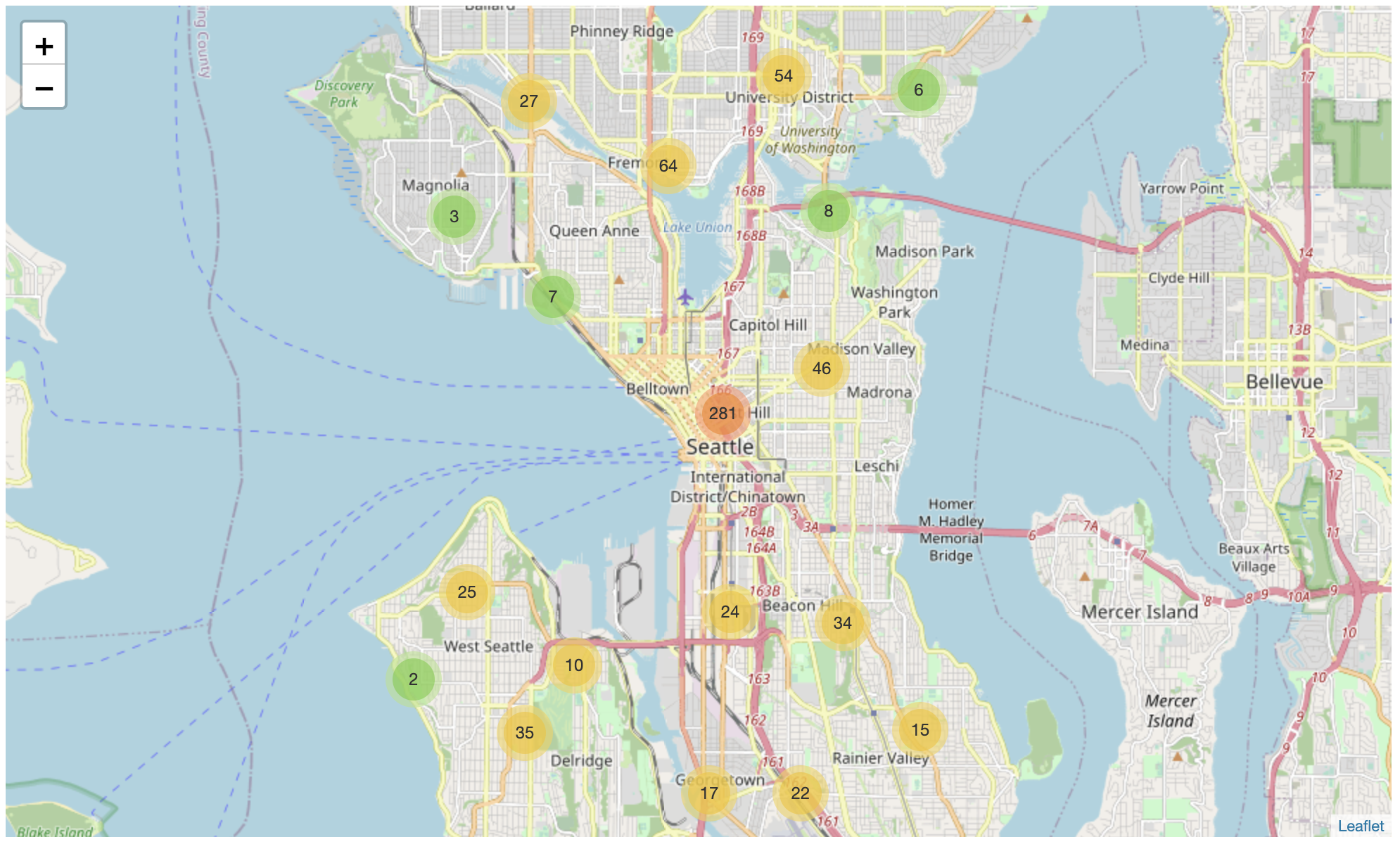




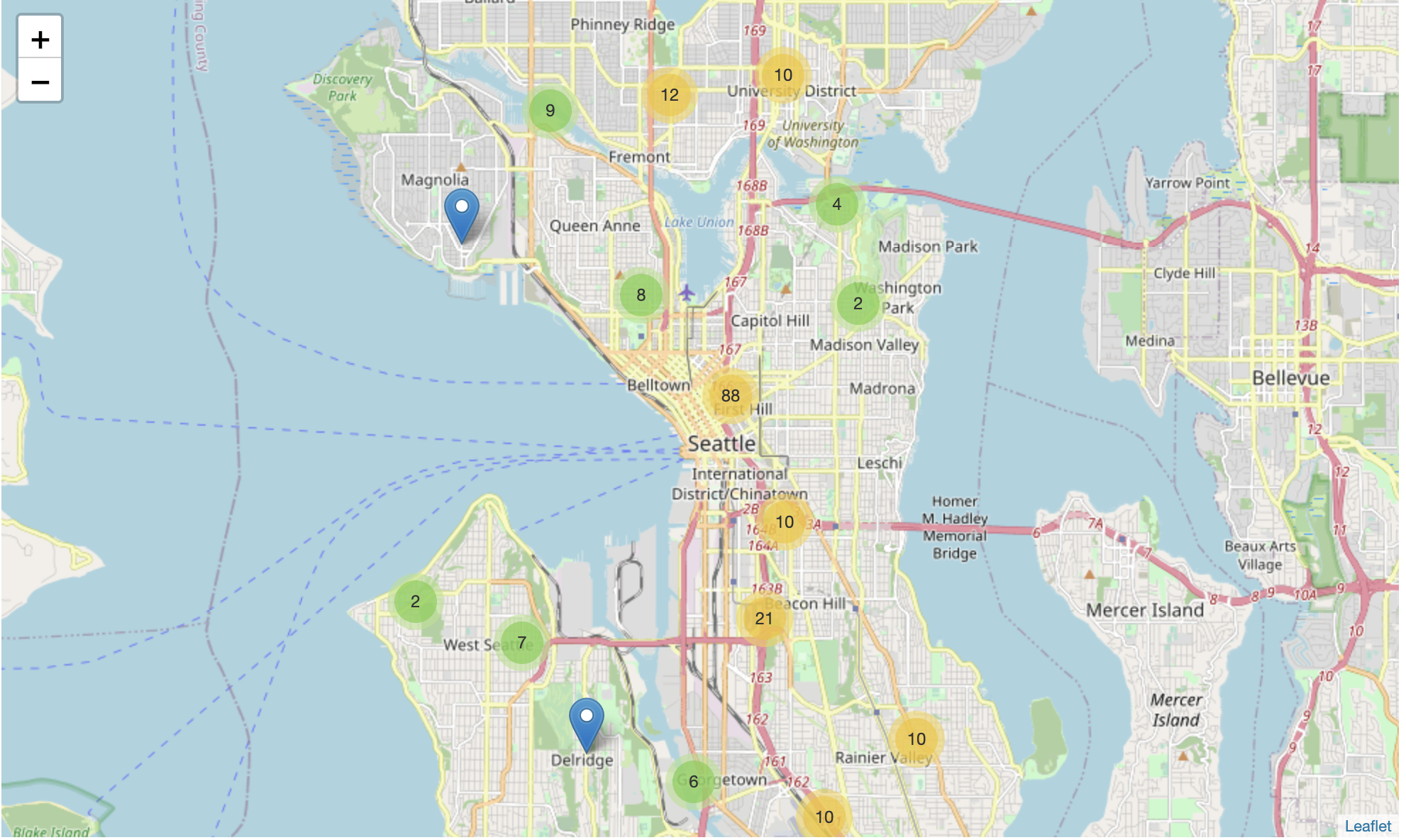




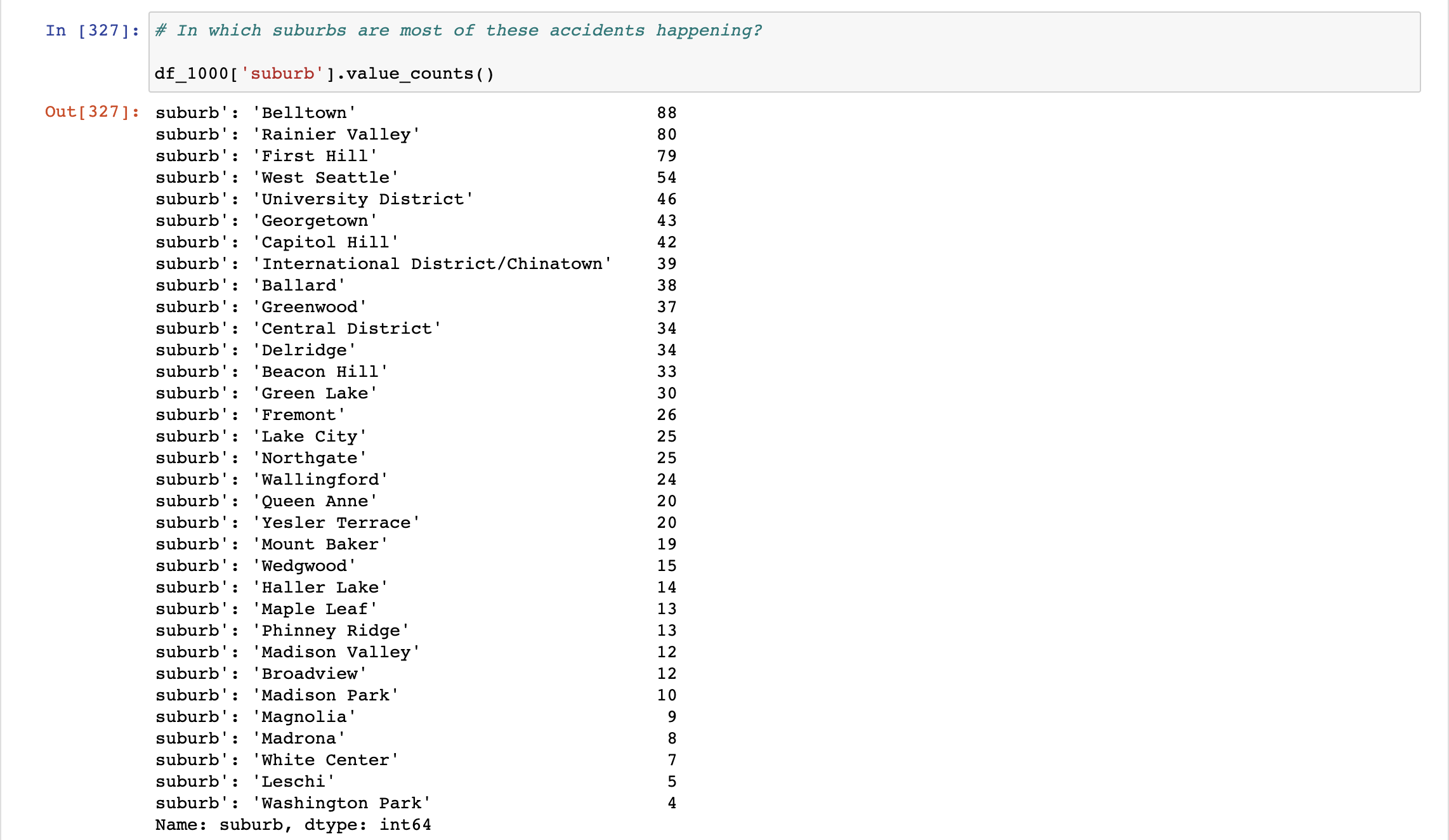
What if we want to see where accidents are happening in Seattle? Visualising accidents by location can tells us where an accident is most likely to happen and also where **SEVERITYCODE = 2** accidents are most likely to happen. However, this tends to be a resource-intensive task and, given that we have over 190 000 rows of data, it is preferable to only visualise a sub-sample of our data. We will visualise where accidents are happening for the first 1000 rows of our dataset.



As we can see most accidents are happening in the Downtown area of Seattle. If we look at where **SEVERITYCODE = 2** accidents are happening, the Downtown area, south east and north east sections of the city are where most of these type of accidents are clustered at. West Seattle and Delridge, despite having a sizable number of accidents (see above pic) have very few **SEVERITYCODE = 2** accidents (see below pic).



By reverse geocoding the latitude and longitude parameters of the accidents in our dataset, we can extract useful location information. To have a more accurate understanding of where these accidents are happening, we can extract the suburb piece of information from the resulting address when we reverse geocode the coordinates. As can be seen below, Belltown and First Hill areas (which previously we considered as being just ‘Downtown’) are where most accidents are happening in the city centre area. Rainier Valley, in the south east of the city also is an accident hub.



**Key takeaways from our Exploratory Data Analysis:**

* Car accidents in Seattle peaked in 2006 and have been falling ever since.
* Most accidents happen on a Friday. Monday and Sunday are the least “accident-prone” days of the week.
* October is the month when most accidents occur. The days just before Halloween tend to be the busiest in terms of collisions.
* Most of the accidents happen at midnight. The 5pm work rush period also leads to a higher number of accidents at that time of the day.
* Most accidents happen at a block but those that do occur at an intersection tend to be more severe.
* Most accidents involved a parked car. However, the most severe accidents tended to be those involving a pedestrian, those happening on a left turn, on the rear end and at angles.
* Most accidents happened on clear days.
* Most accidents happened in daylight.
* The majority of car accidents in Seattle happen in the city centre area, specifically in the Belltown and First Hill suburbs. Rainier Valley is also another accident-prone area.

**Inferential Statistical Testing:**

We did not perform inferential statistical testing on our dataset.

**Machine Learning application: Can we predict the severity of an accident in Seattle?**

We employed a logistic regression model to predict the severity of a car accident in Seattle given information related to the when, where and how of the accident.

More specifically, we tried to gauge whether an accident was of **SEVERITYCODE = 2** or if it was of **SEVERITYCODE = 1, i.e. SEVERITYCODE** was the label we are trying to predict.

A logistic regression model is named for the function used at the core of the method, the logistic function.

The logistic function, is an S-shaped curve that can take any real-valued number and map it into a value between 0 and 1, but never exactly at those limits.

A picture containing object, clock

Description automatically generated

𝑒 is the base of the natural logarithms and 𝑥 is value that you want to transform via the logistic function.

Chart, line chart

Description automatically generated

The logistic regression equation has a very similar representation like linear regression. The difference is that the output value being modelled is binary in nature.

A picture containing diagram

Description automatically generated

Where 𝛽0 is the intercept term

𝛽1 is the coefficient for 𝑥1

𝑦̂ is the predicted output with real value between 0 and 1. To convert this to binary output of 0 or 1, this would either need to be rounded to an integer value or a cut-off point be provided to specify the class segregation point.

The independent variables we used to feed our model were:

* **PERSONCOUNT:** the total number of people involved in the accident
* **PEDCOUNT:** the number of pedestrians involved in the accident
* **PEDCYLCOUNT**: the number of cyclists involved in the accident.
* **VEHCOUNT**: the number of vehicles involved in the accident.
* **UNDERINFL**: whether the driver involved was under the influence of alcohol.
* **HOUR**: the hour of the day when this accident happened.
* **ADDRTYPE**: we performed one-hot encoding on our **ADDRTYPE** column to create dummy variables for each one of the values in that column.
* **COLLISIONTYPE:** we performed one-hot encoding on our **COLLISIONTYPE** column to create dummy variables for each one of the values in that column.
* **WEATHER:** we performed one-hot encoding on our **WEATHER** column to create dummy variables for each one of the values in that column.
* **ROADCOND**: we performed one-hot encoding on our **ROADCOND** column to create dummy variables for each one of the values in that column.
* **LIGHTCOND**: we performed one-hot encoding on our **LIGHTCOND** column to create dummy variables for each one of the values in that column.
* **DAYOFWEEK**: we performed one-hot encoding on our **DAYOFWEEK** column to create dummy variables for each one of the values in that column.
* **MONTH**: we performed one-hot encoding on our **MONTH** column to create dummy variables for each one of the values in that column.
* **SUBURB**: we performed one-hot encoding on our **SUBURB** column to create dummy variables for each one of the values in that column.

**Results summary:**

Because we used reverse geocoding to extract location information, we had to use a reduced sample size as it is a computationally inefficient process. Hence we used 1000 datapoints, which were reduced to 958 once we eliminated the rows with NaN values in their **SUBURB** column.

We split the data into a training (70% of datapoints) and testing set (30%), and trained our model on the training subsample.

From the ***scikitlearn*** library, we imported the necessary techniques and performance reports thus leading to the following performance:

|  |  |
| --- | --- |
| **Confusion Matrix** | |
| 166 | 24 |
| 58 | 40 |

|  |  |
| --- | --- |
| **Accuracy Score** | 0.7153 |

|  |  |
| --- | --- |
| **Classification Report** | |
| **Precision** | 0.7411 |
| **Recall** | 0.8737 |
| **F1\_Score** | 0.8019 |

Therefore our logistic model was 71.5% accurate in predicting if an accident is of SEVERITYCODE = 1 or SEVERITYCODE = 2.

**Discussion:**

Key takeaways for the stakeholders:

* Car accidents in Seattle peaked in 2006 and have been falling ever since.
* Most accidents happen on a Friday. Monday and Sunday are the least “accident-prone” days of the week.
* October is the month when most accidents occur. The days just before Halloween tend to be the busiest in terms of collisions.
* Most of the accidents happen at midnight. The 5pm work rush period also leads to a higher number of accidents at that time of the day.
* Most accidents happen at a block but those that do occur at an intersection tend to be more severe.
* Most accidents involved a parked car. However, the most severe accidents tended to be those involving a pedestrian, those happening on a left turn, on the rear end and at angles.
* Most accidents happened on clear days.
* Most accidents happened in daylight.
* The majority of car accidents in Seattle happen in the city centre area, specifically in the Belltown and First Hill suburbs. Rainier Valley is also another accident-prone area.

How we can improve results going forward:

* Less number of columns: e.g. via dimensionality reduction.
* Bigger dataset than just 958 columns.

**Conclusion:**

Stakeholders can use the findings of this analysis to try and educate the local population about the most common types of accidents and work with local authorities to increase investment to ensuring that roads in the best conditions and that inhabitants are aware of the risks of getting the car at certain points of the year.

For example, campaigns can be launched in the weeks prior to Halloween to increase awareness of the high number of accidents that tend to occur in the days just before the event and on the night of Halloween celebrations.

Furthermore, having singled out the city centre and Rainier Valley areas, local authorities can also launch campaigns at a suburb level to increase awareness of the high number of accidents that occur in that area and to improve the condition of the roads.