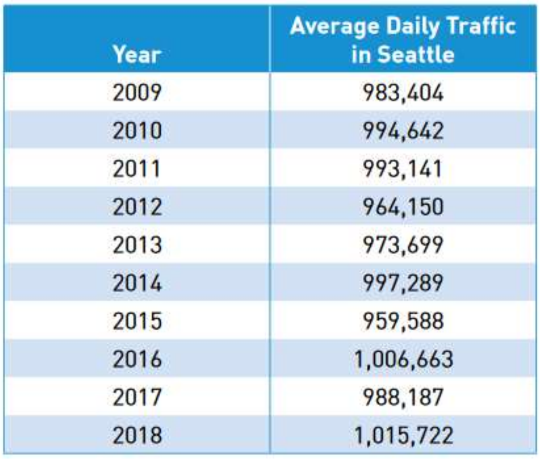
Predicting the Severity of Car Accidents

Elias Jockheck

October 2nd, 2020

1. Background

Seattle is one of the fastest-growing major cities in the United States. In the year 2019 Seattle's population increased by 1.5% and thus was sixth among the 50 largest U.S. cities. As population grows, the amount of vehicles on the road also rises. In 2017, the number of cars in Seattle reached 435,000. Per square mile we are dealing with a number of 5,185 cars. With so many cars being around in Seattle the probability of car accidents also increases. The following table shows the average daily traffic volumes since 2009 until 2018.



Year	Average Daily Traffic in Seattle
2009	983,404
2010	994,642
2011	993,141
2012	964,150
2013	973,699
2014	997,289
2015	959,588
2016	1,006,663
2017	988,187
2018	1,015,722

Table 1 Average Daily Traffic Volumes

Having such a high and constant traffic volume over the year major car accidents can lead to huge traffic jams in and around the city. As such it could be advantageous for drivers and for the Seattle Department of Transportation to further analyze the main causes for car accidents, so that they can take concrete actions to help drivers avoid potential traffic jams or even accidents itself.

If we take a closer look on the police reported collisions on Seattle streets there is already a positive downward trend (see figure 1)

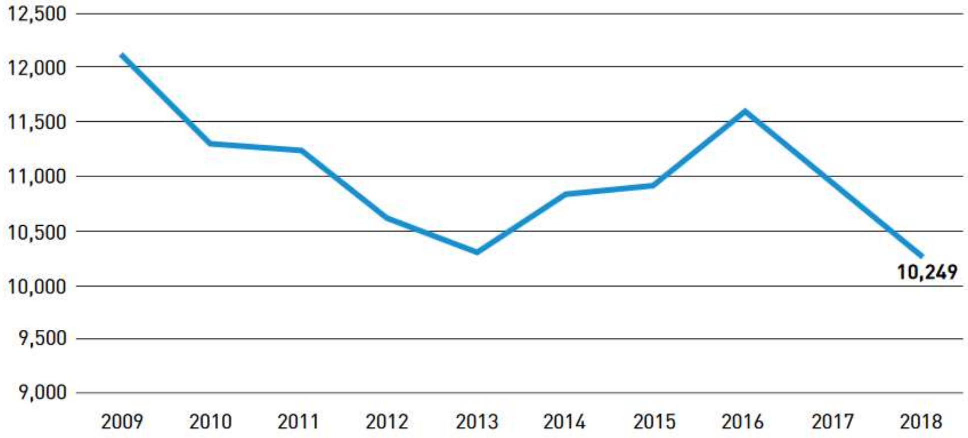


Figure 1 Police reported collisions on seattle streets

But the numbers are still high and the goal should be to achieve even lesser numbers in the future.

2. Business Understanding

Say you are driving to another city for work or to visit some friends. It is rainy and windy, and on the way, you come across a terrible traffic jam on the other side of the highway. Long lines of cars barely moving. As you keep driving, police car start appearing from afar shutting down the highway. Oh, it is an accident and there's a helicopter transporting the ones involved in the crash to the nearest hospital. They must be in critical condition for all of this to be happening.

Imagine we were able to find a way of predicting the severity of a car accident. Or in other words predict the possibility of you getting into a car accident and predict the possible severity of the accident on top. In order to make such predictions it is necessary to take different attributes like weather or road conditions into account. In the end, given the prediction drivers can move more carefully or if possible even change their actual travel route to avoid major traffic jams.

Our Target audience is the SDOT Traffic Management Division. They could use the data they are already collecting and analyze which are the main causes for accidents happening in certain areas. On top it is possible to calculate the influence of the main factors that lead to accidents. Knowing the main factors and their influence the objective is to define measurements that improve for example road conditions or light conditions. Let’s say our analysis shows that severe accidents happen because of bad light conditions in a certain area the SDOT can take actions to improve these conditions and hopefully lower the future number of accidents.

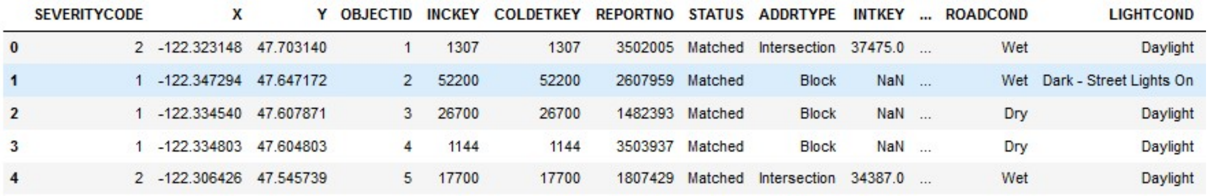
3. Analytical Approach

In the first step we will graphically analyze (e.g. using bar charts) the data and visualize how different features relate to the number and severity of accidents. In the second step we will use a decision tree and a logistic regression to address the classification problem in this use case.

In general classification is a form of data analysis where a model or classifier is constructed to predict categorical labels (in our case: the severity of accident). The objective of the predicting using classification technology is to accurately forecast the target class variable (severity of accident) for each new data point. In our case the goal of the classifier would be to predict accident severity for a new accident which is missing in the data.

4. Data Understanding

To address the issue described above we will use a data set provided by the SDOT Traffic Management Division. This Dataset includes all types of collisions provided by SPA and recorded by Traffic Records from 2004 until present. This data gets weekly updated. The following figure shows an excerpt of the data we will be working with.



	SEVERITYCODE	X	Y	OBJECTID	INCKEY	COLDETKEY	REPORTNO	STATUS	ADDRTYPE	INTKEY	...	ROADCOND	LIGHTCOND
0	2	-122.323148	47.703140	1	1307	1307	3502005	Matched	Intersection	37475.0	...	Wet	Daylight
1	1	-122.347294	47.647172	2	52200	52200	2607959	Matched	Block	NaN	...	Wet	Dark - Street Lights On
2	1	-122.334540	47.607871	3	26700	26700	1482393	Matched	Block	NaN	...	Dry	Daylight
3	1	-122.334803	47.604803	4	1144	1144	3503937	Matched	Block	NaN	...	Dry	Daylight
4	2	-122.306426	47.545739	5	17700	17700	1807429	Matched	Intersection	34387.0	...	Wet	Daylight

Figure 2 Collisions all years provided by SPD

Our imported dataset consist of the labeled data column 'severity code' and 37 attributes that might be useful for the later prediction. Later on we will decide what to keep and what to drop. In total we have 194,673 rows which seems to be a good amount of data.

Each row represents an accident/collisions reported, and for each accident/collision, the corresponding severity code is documented. The other attributes contain information on i.e. weather/road/light conditions, the location, speeding etc.

Let's take a look a specifiy row (objext id 3). From the Data we can see that an accident took place during daylight (see attribute - LIGHTCOND) and the road was dry (see - attribute ROADCOND). The severitycode (see attribute SEVERITYCODE) for this accident was labeled 1—prop damage, which in this case means : Proporty Damage Only Collision (see attribute SEVERITYDESC).

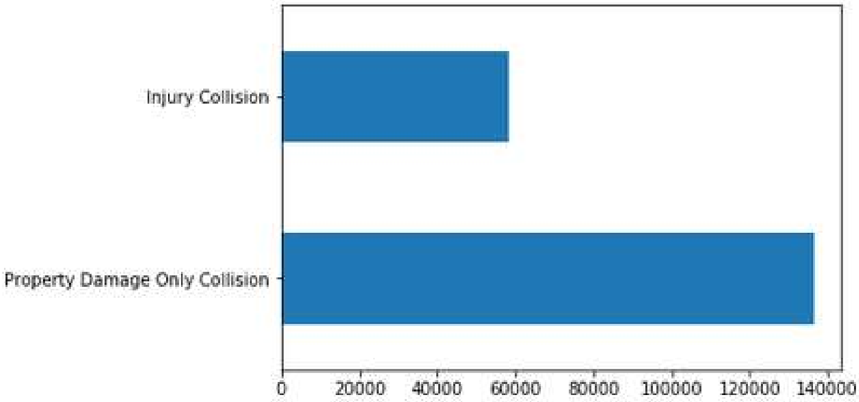
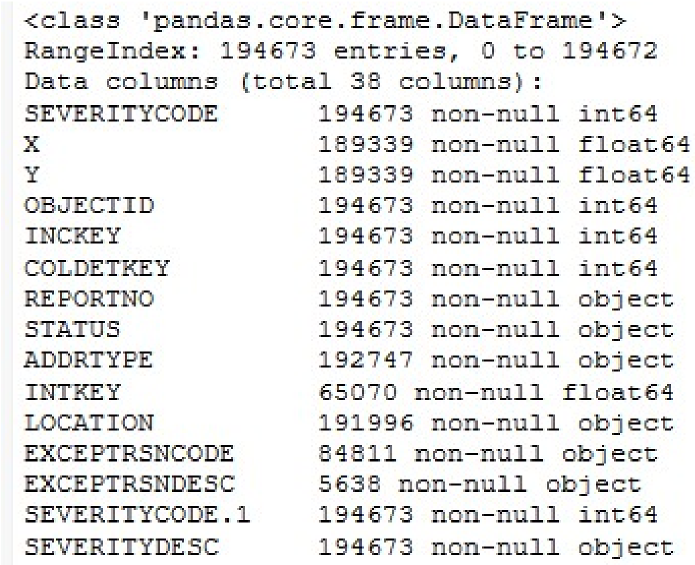


Figure 3 Value count of severity code

As we can see our labeled data (SEVERITYCODE) is slightly unbalanced. We have the double amount of data with a severity code 1 (136485) in comparison to severity code 2 (58188). This might cause problems for the Modeling and the prediciton later on.

Concerning the datatypes we can see that our dataset has numerical (float64 or int64 types) or categorical (object type) data types (see figure 4). In Addition some attributes have missing data (NaN) as well.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 194673 entries, 0 to 194672
Data columns (total 38 columns):
SEVERITYCODE       194673 non-null int64
X                  189339 non-null float64
Y                  189339 non-null float64
OBJECTID           194673 non-null int64
INCKEY             194673 non-null int64
COLDETKEY          194673 non-null int64
REPORTNO           194673 non-null object
STATUS             194673 non-null object
ADDRTYPE           192747 non-null object
INTKEY             65070 non-null float64
LOCATION           191996 non-null object
EXCEPTRSNCODE      84811 non-null object
EXCEPTRSNDESC      5638 non-null object
SEVERITYCODE.1     194673 non-null int64
SEVERITYDESC       194673 non-null object
```

Figure 4 Abstract of information about the dataset

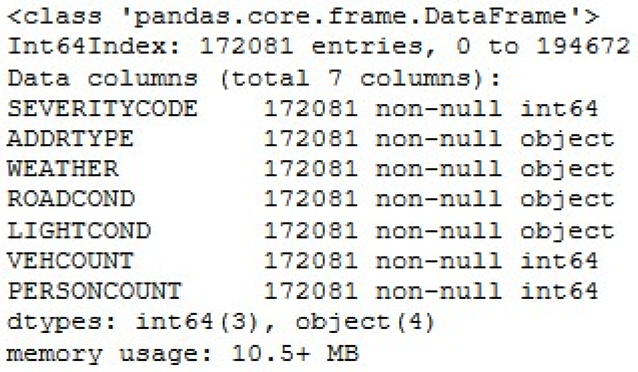
5. Data Preparation

The data preparation includes all the required activities to construct the final dataset which will be fed into the modeling tools. Data preparation can be performed multiple times and it includes balancing the labeled data, transformation, filling missing data, and cleaning the dataset.

The following figure shows the main information of our data set. Some features have missing values. Especially the features INTKEY, EXCEPTRSNCODE, EXCEPTRSNDESC, INATTENTIONIND, PEDROWNOTGRNT and SPEEDING have a high number of missing values. The attribute SPEEDING, which would be interesting four analyzing the main causes of collisions only has 9333 values so it is missing over 180,000 values and will therefore not be interesting for further analysis.

Some features had ‘unkown’ values as well which do not hold much significance in modelling, due to them not giving information about the circumstances of the accident. As a consequence, entries holding such values are dropped from our data set.

Having completed the cleansing process we will take a look at the following features for further analysis:



```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 172081 entries, 0 to 194672
Data columns (total 7 columns):
SEVERITYCODE    172081 non-null int64
ADDRTYPE        172081 non-null object
WEATHER         172081 non-null object
ROADCOND        172081 non-null object
LIGHTCOND       172081 non-null object
VEHCOUNT        172081 non-null int64
PERSONCOUNT     172081 non-null int64
dtypes: int64(3), object(4)
memory usage: 10.5+ MB
```

Figure 5 Final features selected

6. Exploratory Data Analysis

In this project the target or dependent variable is the feature 'SEVERITYCODE'. As for the independent input variables shown in the figure above we will graphically analyze the correlation of selected features to our target variable.

Relationship between vehicles involved and number of Accidents

By visualizing the number of vehicles involved we want so find out if there is a total number of vehicles involved where the number of accidents is extremely high. As figure 5 shows in our seattle data set most of the accidents happened with 2-3 cars involved.

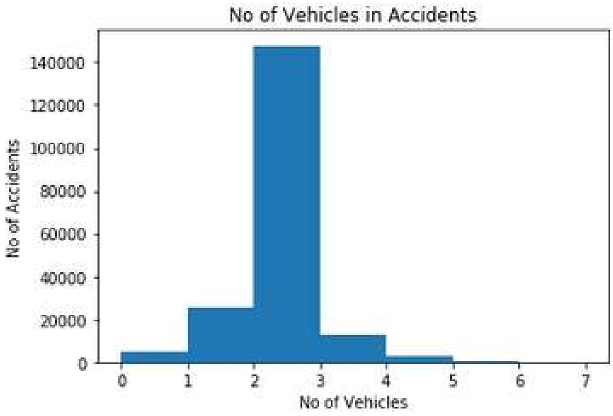


Figure 6 Number of vehicles against accident counts

Relationship between persons involved and number of Accidents

Maybe more people in the car lead to a higher number of accidents because more people in one car might chatter and thus distract the driver. We all might have experienced such situation ourselves but have been lucky not to get into accident. Let's see what our data is showing us.

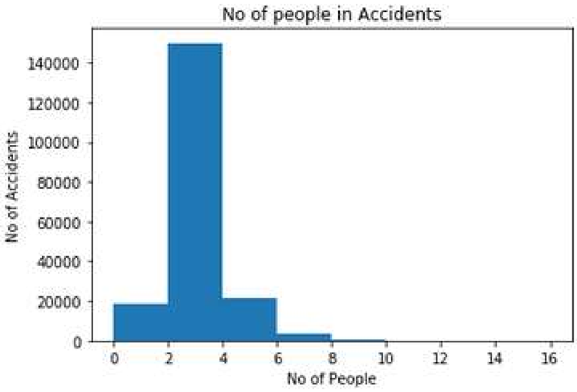
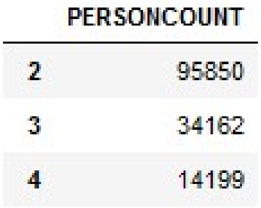


Figure 7 Bar graph No of persons against accident counts

Figure 7 clearly showdthat most accidents happened within the bin of 2-4 people involved. To get a better picture let's get the exact values and rank them.



	PERSONCOUNT
2	95850
3	34162
4	14199

Figure 8 Top 3 accident count

Most accidents happened with two people involved (95850). So solo driving in this case is a real cause for high numbers of accidents.

Relationship between collision address type and number of Accidents

In our dataset the feature 'ADRESSTYPE' (Intersection, Block, Alley) tells us on what type of road address a collision took place. Looking at the plot below (see figure 9) it can be inferred that most accident occur at blocks, followed by intersections and the alleys. It can also be inferred that a higher proportion of injury collisions (severity code 2) occurred at intersections compared to property damage only collisions (severity code 1).

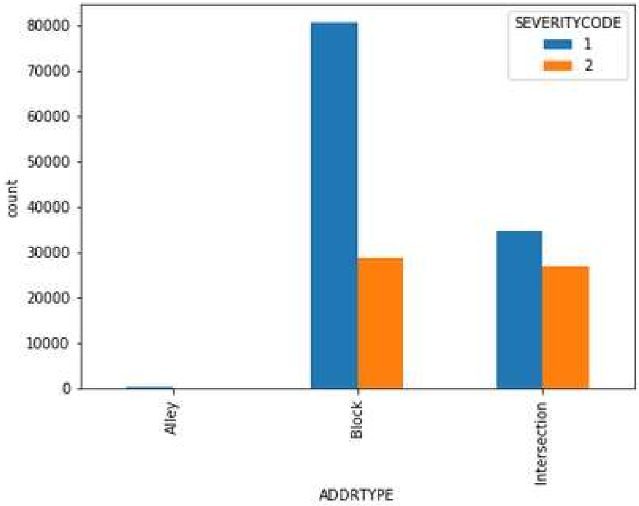


Figure 9 Adress type against number of accidents

Relationship between weather condition and number of accidents

During bad weather conditions it more risky to drive and we have to be very cautious with our surroundings while being on the road. So in general one might think that during bad weather also occur more accidents. For example during heavy rain or snow drivers face slippery roads or reduced visibility. However our data (see figure 10) seems to disagree as it is observed that the majority of accidents occur during clear weather. There also seems to be no clear indication which weather condition will result in more accidents.

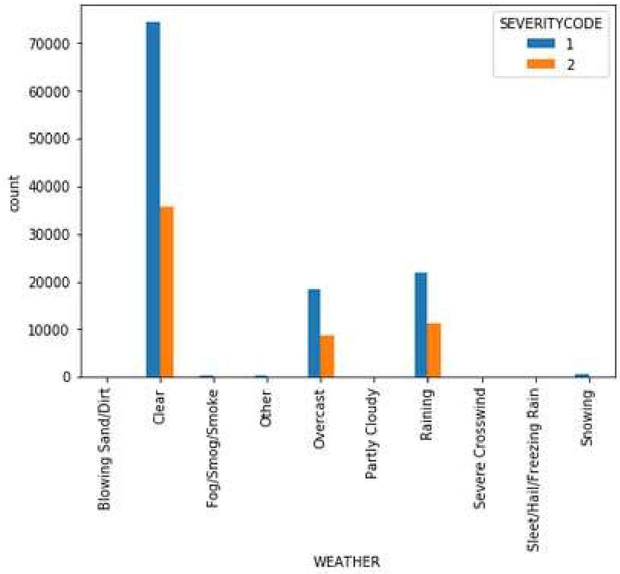


Figure 10 Weather conditions against number of accidents

Relationship between road condition and number of accidents

When it is raining and roads get slippery, drivers tend to drive slower and more carefully as the braking distance of vehicles increases in such conditions. In general it is widely accepted that condition of the road is closely related to the risk of an accident. In the Seattle case the data shows (see figure 11) to our surprise that most accidents occur when the road is dry compared to other road conditions.

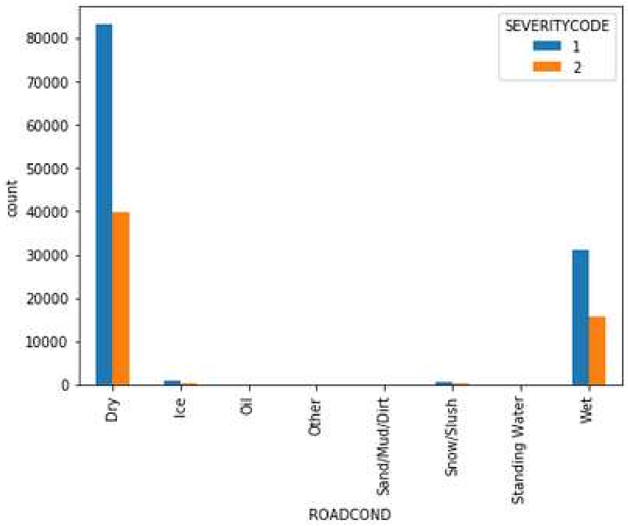


Figure 11 Road conditions against number of accidents

Relationship between light conditions and number of accidents

It is also widely agreed that poor light conditions can easily result in higher risks of accidents as the drivers may not be able to see surrounding vehicles due to poor visibility. Similar to weather conditions, the data (see figure 12) also disagrees with the above statement. It is observed that a lot more accidents occur in environments of good lighting – such as –daylight- and -dark with street lights on. It is also observed to be difficult to determine which light condition results in more severe accidents.

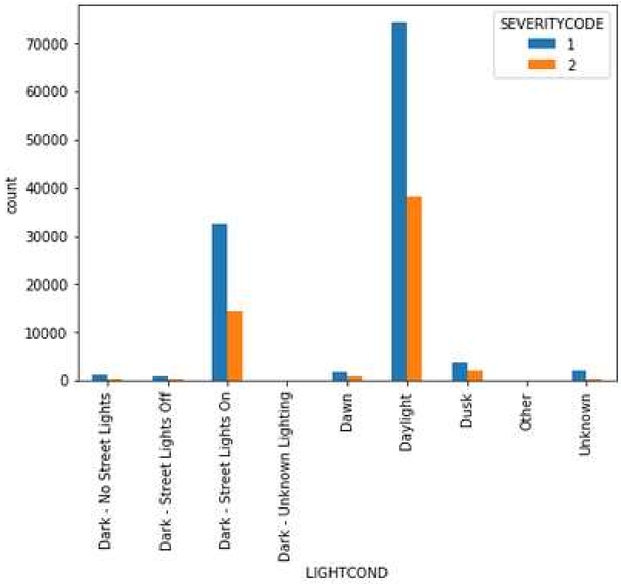


Figure 12 Light conditions against number of accidents

7. Modeling

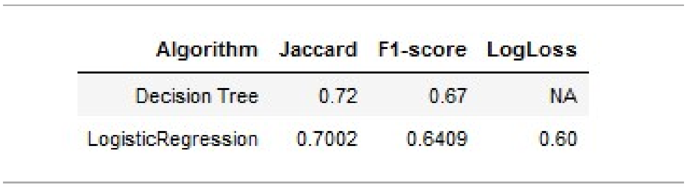
As already describe in section 3 classification models will be used in order to categorize accidents into the likely severity codes depending on the characteristics of the accident.

Logistic Regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable (in our case: SEVERITYCODE). The Decision Tree Analysis breaks down a data set into smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes.

The applications of classification models were quite straightforward. After having done the necessary preprocessing I simply needed to split the original data set into training and test sets and run them through the models.

8. Evaluation

Generally, the metrics used to evaluate the models are Jaccard Index, F1 Score and Logarithmic Loss (Log Loss). In the following you can see how the models performed (see figure 13).



Algorithm	Jaccard	F1-score	LogLoss
Decision Tree	0.72	0.67	NA
LogisticRegression	0.7002	0.6409	0.60

Figure 13 Evaluation results

By comparing the Jaccard and the F1 score we can conclude that decision tree performed slightly better than the logistic regression.

9. Conclusion

In this study, I analyzed the relationship between the severities of vehicle accidents and the circumstances surrounding the accidents. I identified addresstype, road, light and weather conditions to be some of the most important features that affect the severity of an accident. Classification models were built to predict the severity of accidents, so that drivers can use these predictions to gauge the impact of the accident on traffic. This could be useful to drivers as they can change their route of advancements early so avoid being stuck in traffic jams should an accident occur. For SDOT Traffic Management Division we can conclude that they are already doing a great job by collection all the data. By using the data with machine learning algorithms they can benefit even more from it.

10. Future Directions

Despite the classification models achieving about 72% accuracy, the circumstances around accidents are very dependent on the drivers themselves and how careful and alert they are while driving. People nowadays use smartphone almost everywhere. A lot of people event connect them to their cars by using car play or android auto. Using the GPS data of the driver the SDOT Traffic Management Division could develop an application that is warning the driver when he is an around an area of high accident risk. In Addition the application could give suggestion (based on the main causes for accidents in this area) how to act while driving or how to avoid a specific area (e.g. current traffic jam).