**PREDICTING CAR ACCIDENT SEVERITY**

Andrés Mora

22.SEP.2020

1. **INTRODUCTION**

Local government institutions are accountable for organizing traffic circulation and guaranteeing traffic safety in their respective jurisdictions. This is of special relevance when dealing with the most unfortunate traffic-related occurrence: car accidents, which pose at stake human, social and economic losses. Hence, government institutions could benefit greatly from information regarding the traffic behavior of their constituents, especially that related to car accidents. A long held intuition supposes that car accident severity can worsen given certain environmental conditions, like those related to weather, lighting, or the road conditions. Were this true, government institutions could use such information in their advantage and be ready to design and implement traffic safety regulations with environmental conditions in regard. So: can environmental conditions predict the severity of a car accident?

1. **DATA**

Car accident data has been registered by the Seattle, WA jurisdiction since 2004 until the present time. Such database is administered by the SDOT (Seattle Department of Transportation) and the organization Traffic Records Group, and it is open for the public. The variable to be predicted in this report is ‘Car Accident Severity’, recorded in the database as ‘SEVERITYCODE’: a nominal variable which classifies with a numerical code the severity of the registered collision. The selected independent variables are those considered as environmental: that is, variables which indicate external conditions; excluding then, those related to the driver, the passengers, and the car itself. From the given database, four environmental variables were chosen to be analyzed: Weather, Lighting Conditions, Road Conditions, and Date&Time. From these, Date&Time was discarded, as the registered format was not constant and the data was too specific, generating much noise. Thus, only three independent variables remained; all of them categorical, just as the dependent variable. Once both target and independent variables were established, data cleaning was ensued: missing data, and data labeled as ‘unknown’ or ‘other’, were dropped in order to reduce uncertainty. From a total of 194,673 car accidents registered, the pre-processed data for this study ended up with 169,949 rows.

1. **METHODOLOGY**

Once data was pre-processed, as described in the heading above, an analysis study was determined. Since the selected variables, both dependent and independent, are categorical variables, this report required a categorical analysis. Such approach implies the exploration of frequency, therefore distribution, of the data across the categories involved. In this case, it meant exploring how the car accident severity labels were distributed across the labels of each of the environmental variables. In order to do so, a three-step method was applied:

1. Visualizing frequencies:

It involved generating a count-plot in order to visually recognize how the severity labels were distributed across the labels belonging to each of the environmental variables.

1. Obtaining distribution percentages:

A contingency table was generated to showcase the distribution percentages, as to demonstrate numerically the frequency of each severity label with the labels of each environmental variable.

1. Applying categorical correlation metrics

A Chi-square test and a Cramer’s V test were passed in order to numerically understand the relationship between categorical values. A Chi-square test can reflect association if the Pearson’s Chi-square is high and the p-value low. However, one should notice that such values can happen indistinctively for large datasets, as the one used in this project. A more reliable metric is the Cramer’s V test, that reflects the association strength on a scale from 0 to 1, in which the higher the score the stronger the relationship.

Through this method, the relationship between each of the environmental variables with the target variable was established; thus, the suitability of such variables for building a predictive model.

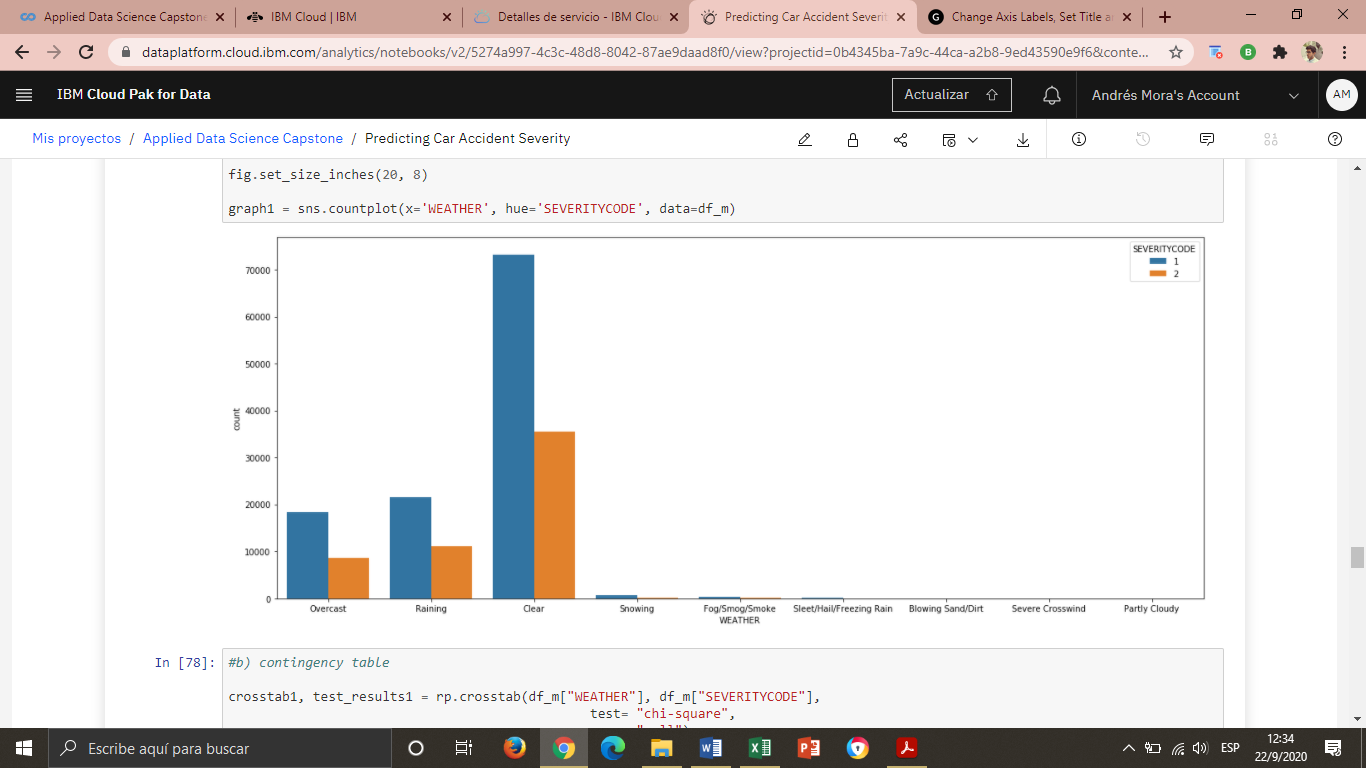
The ‘car accident severity’ variable only showcased two labels: 1-property damage, and 2-injury. As for the environmental labels, they displayed respectively: ‘weather’ (9), ‘lighting condition’ (6), and ‘road condition’ (7).

1. **RESULTS**

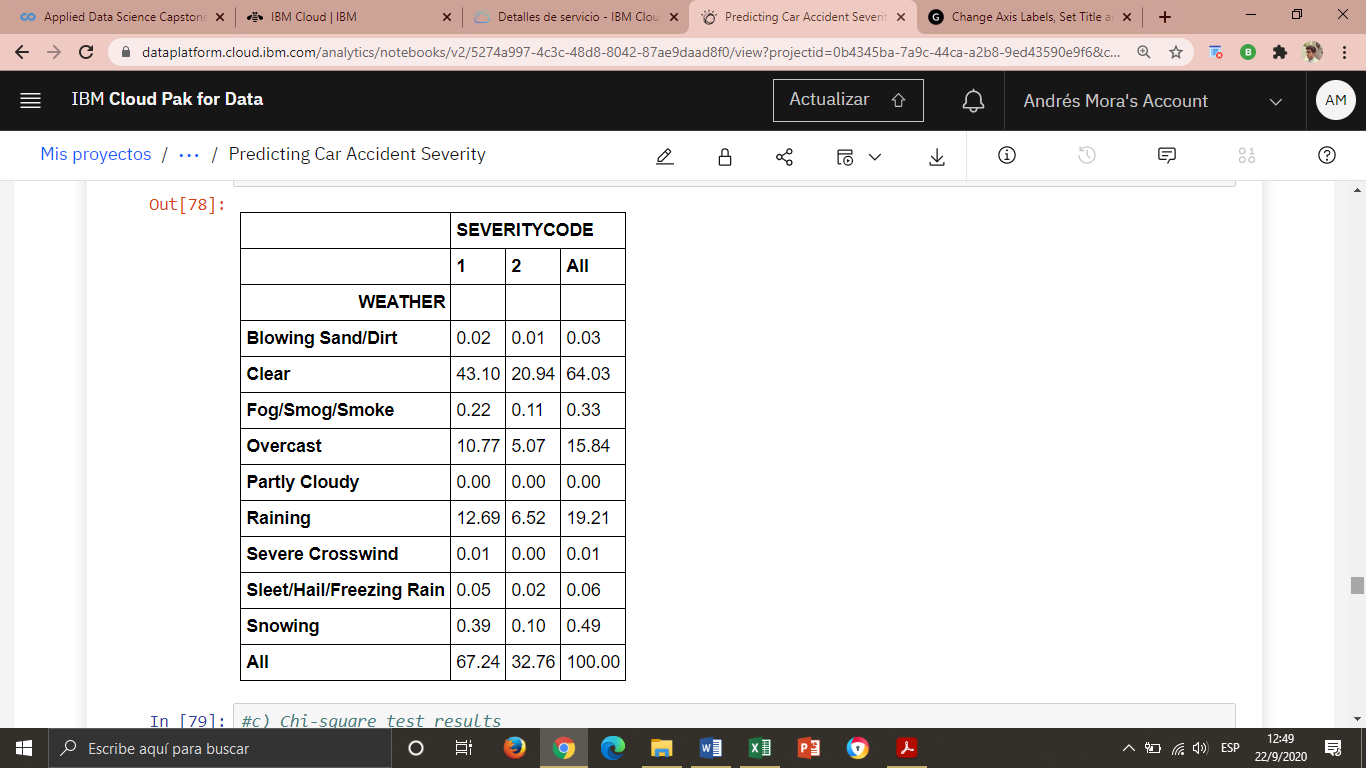
Here is displayed the proposed method applied to each of the independent variables: ‘weather’, ‘lighting condition’, and ‘road condition’.

* Relationship between CAR ACCIDENT SEVERITY and WEATHER

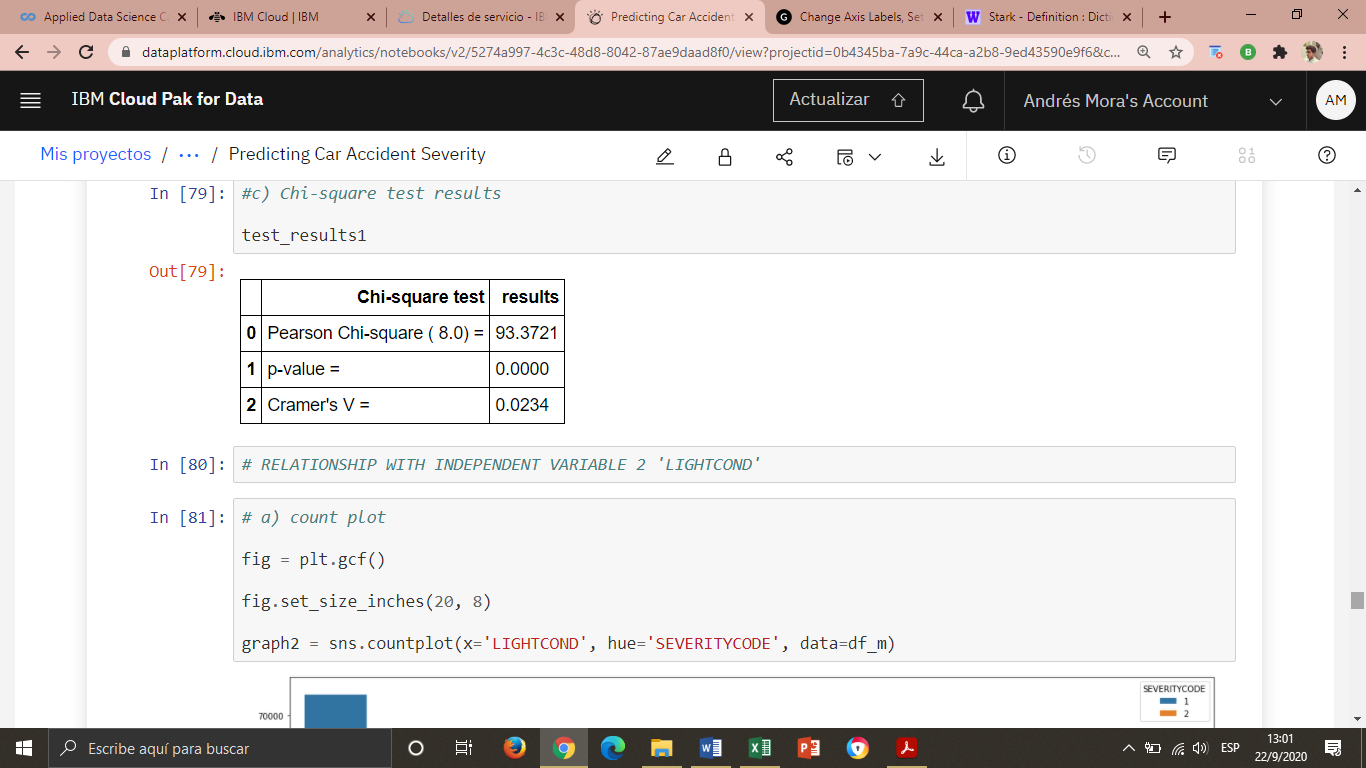
1. The count-plot shows that almost all data is concentrated in three labels: the great majority of car accidents happen under ‘Clear’ weather, another noticeable part happen under ‘Raining’ and ‘Overcast’ weather, and almost none happen under the remaining six labels. Differences of severity only happen in the three top labels, with the majority of accidents pointing a low score in severity: only property damage.



1. The table confirms even more the previous visualization. 64,03% of car accidents occur under ‘Clear’ weather: of these, two thirds correspond to property damage severity, and one third to injury severity. The same distribution of severity occurs under the other top weather labels ‘Raining’ and ‘Overcast’. The remaining labels do not present a significant difference in accident severity.

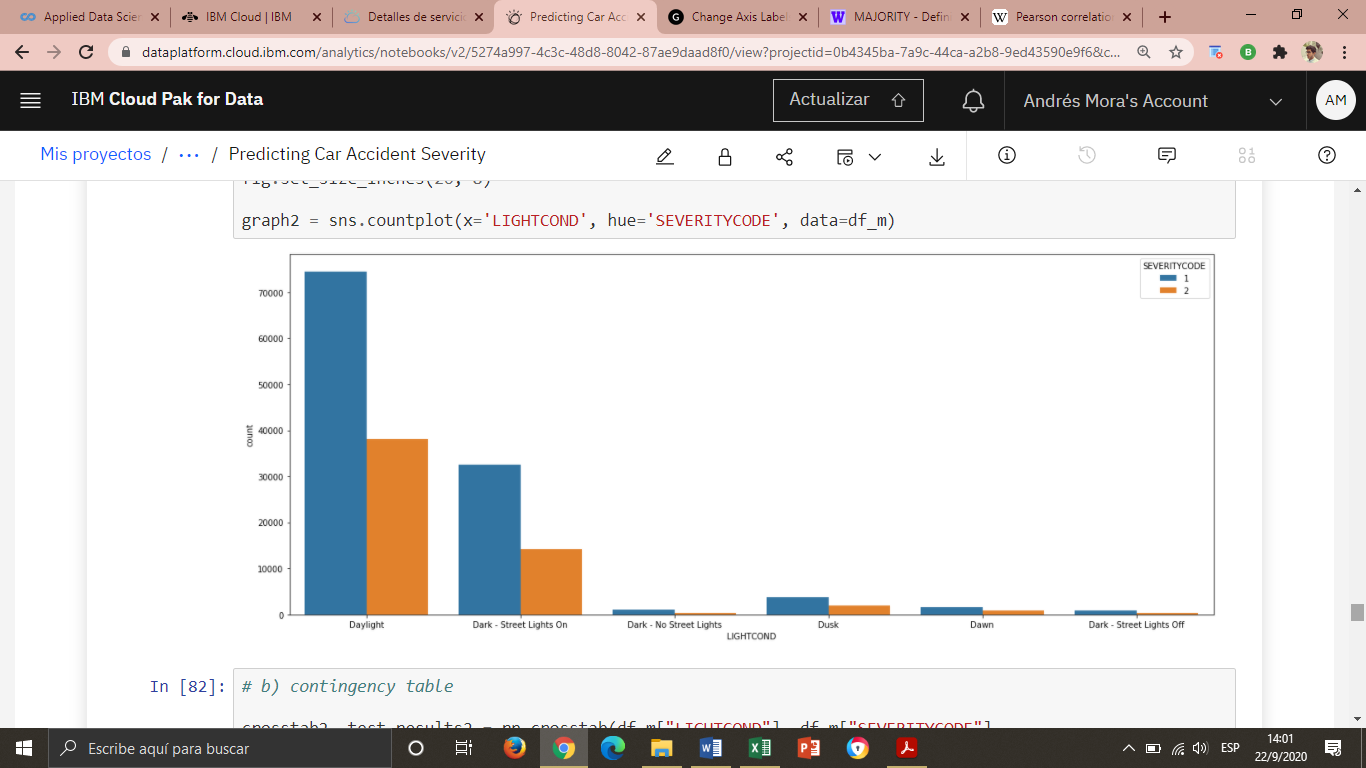


1. The test results indicate that, though there is association according the Pearson Chi-square and the p-value, such association is very weak according to the Cramer’s V. Such association could even be attributed to the large dataset size.



* Relationship between CAR ACCIDENT SEVERITY and LIGHTING CONDITIONS

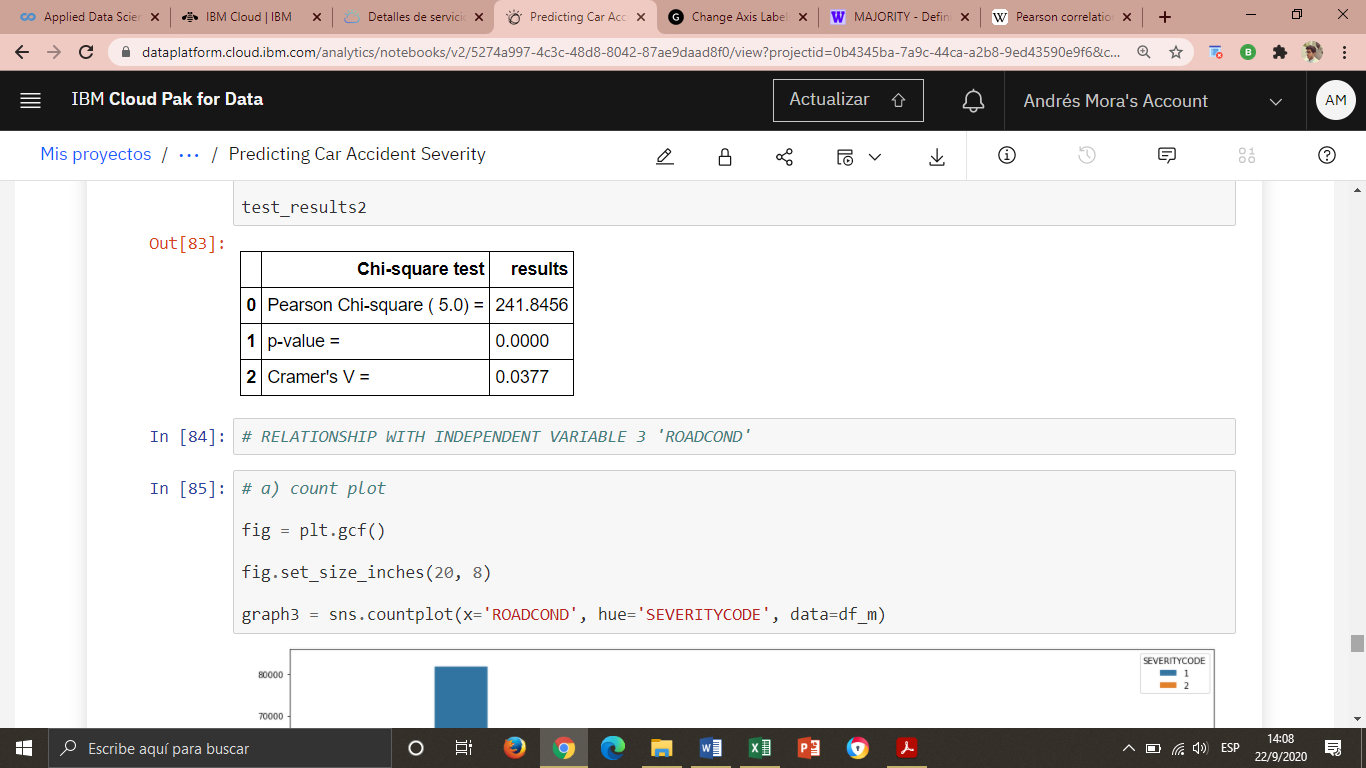
1. The count-plot shows that almost all data in concentrated in two lables: the great majority of car accidents happen under ‘Daylight’ conditions, and another noticeable part happens under ‘Dark – Street Lights On’. The rest of lables resgister almost no car accidents. Differences of severity only happen in the two top labels, with the majority of accidents pointing a low score in severity: only property damage.



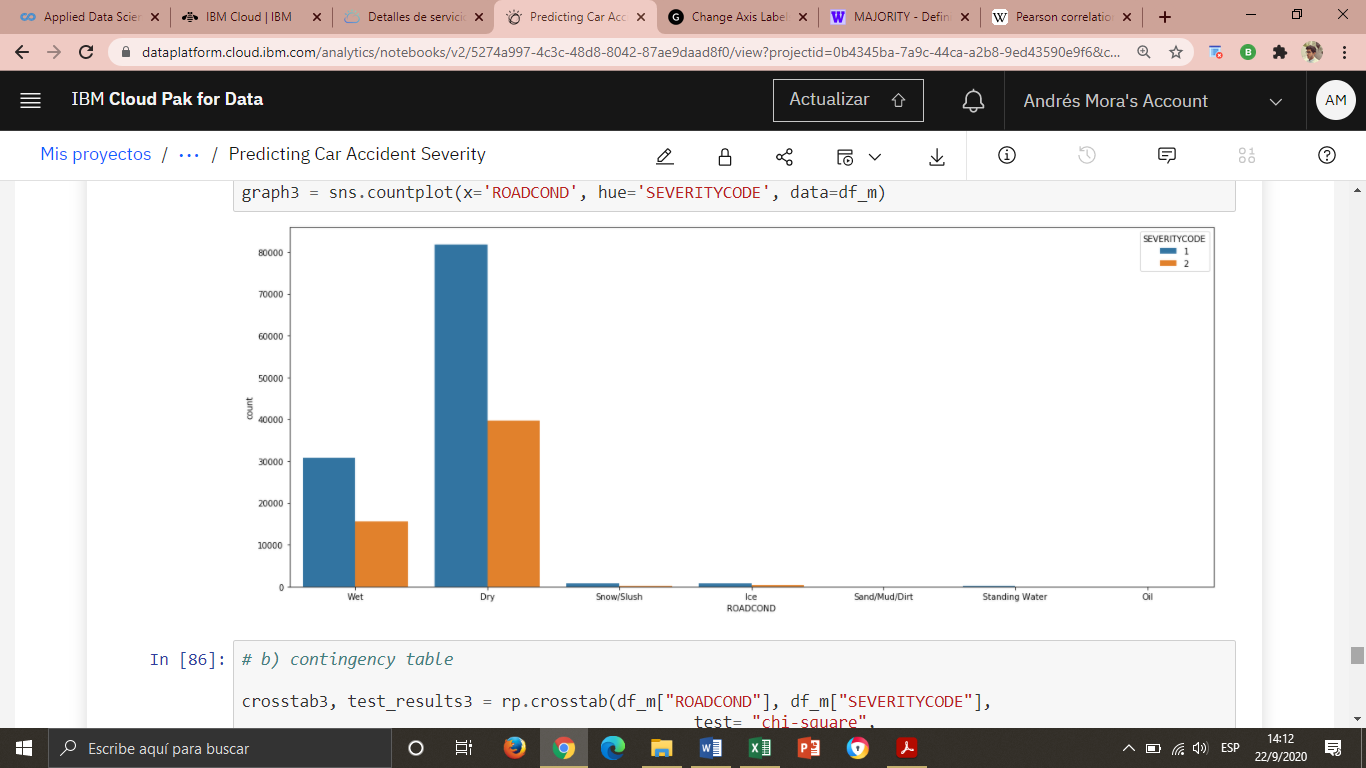
1. This table also confirms the previous visualization. 66,27% of car accidents occur under ‘Daylight’ conditions: of these, two thirds correspond to property damage severity, and one third to injury severity. The same distribution of severity occurs under the other significant label ‘Dark – Street Lights On’. The remaining labels do not present a significant difference in accident severity.

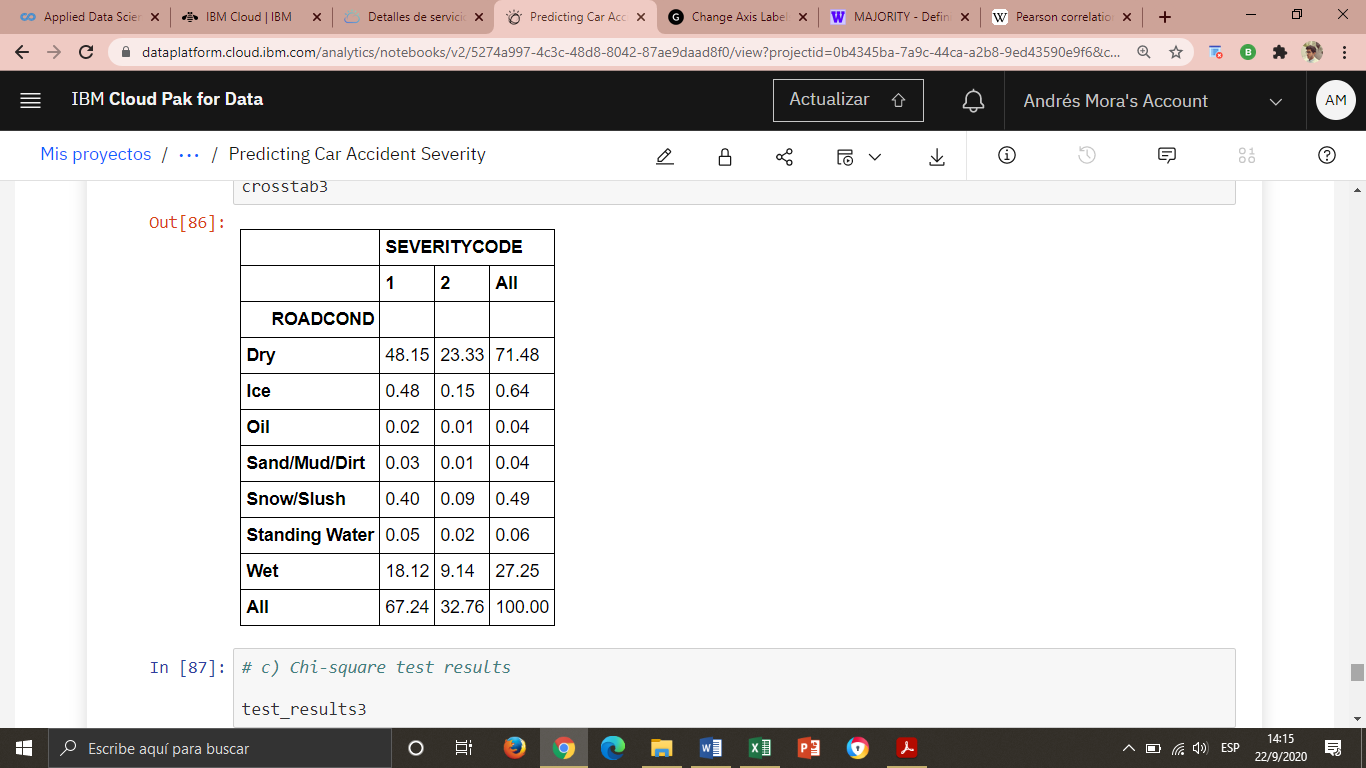


1. The test results indicate association according the Pearson Chi-square and the p-value; but such association is very weak according to the Cramer’s V. The association could even be attributed to the large dataset size.

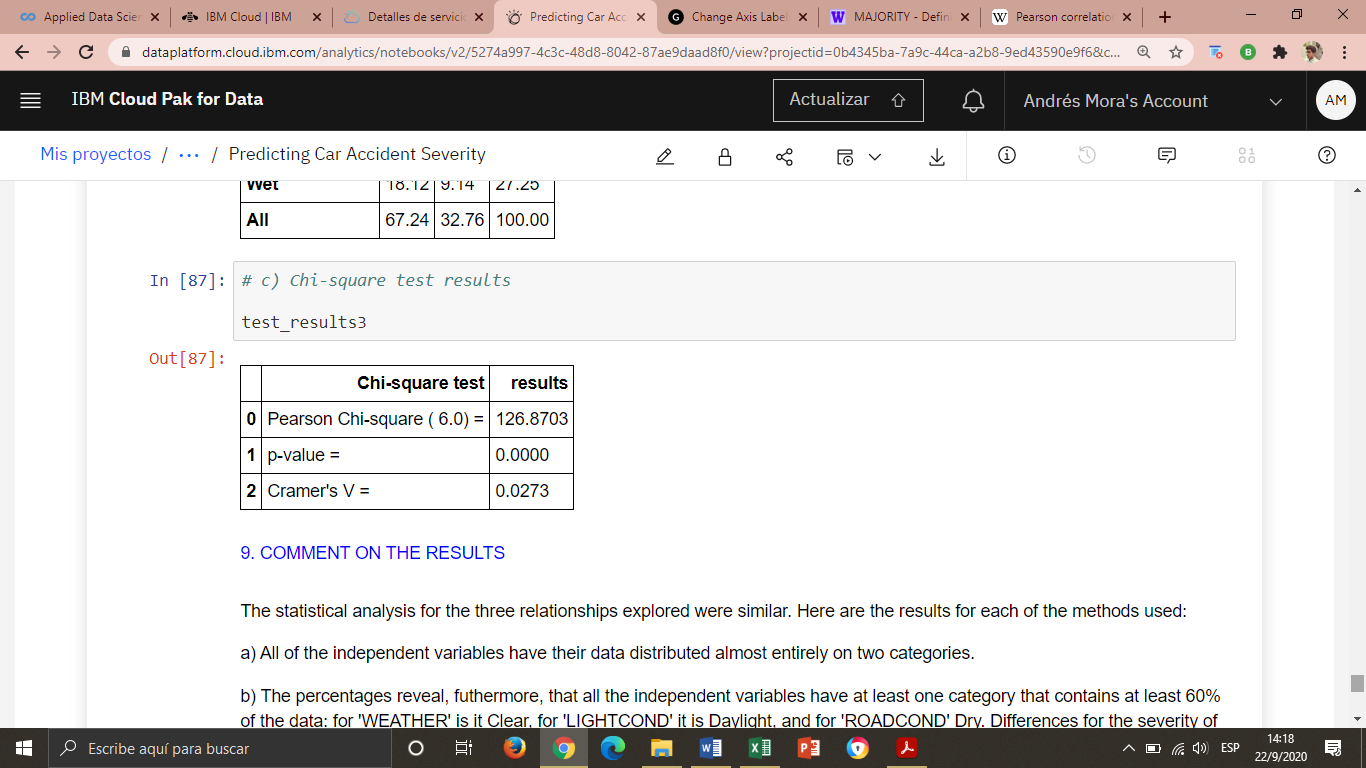


* Relationship between CAR ACCIDENT SEVERITY and ROAD CONDITIONS

1. The count-plot shows that almost all data in concentrated in two lables: the great majority of car accidents happen under ‘Dry’ road conditions, and another noticeable part happens under ‘Wet’ road conditions. The rest of lables resgister almost no car accidents. Differences of severity only happen in the two top labels, with the majority of accidents pointing a low score in severity: only property damage.
2. This table confirms as well the previous visualization. 71,48% of car accidents occur under ‘Dry’ road conditions: of these, two thirds correspond to property damage severity, and one third to injury severity label. The same distribution of severity occurs under the other significant label ‘Wet’. The remaining labels do not present a significant difference in accident severity.



1. The test results indicate association according the Pearson Chi-square and the p-value; but such association is very weak according to the Cramer’s V. The association could even be attributed to the large dataset size.



1. **DISUSSION**

The statistical analysis for the three relationships explored were similar. Here is a summary of the main observations for each of the steps of the method employed:

a) All of the independent variables have their data distributed almost entirely under two labels.

b) The percentages reveal, furthermore, that all the environmental variables have at least one category that contains at least 60% of the data: for 'Weather' it is Clear, for 'Light Conditions' it is Daylight, and for 'Road Conditions' Dry. Differences for the severity of car accident occur in these categories, and the second highest frequent category in each variable, but are rendered almost imperceptible for the rest of the labels.

c) The Chi-square test revealed some association: Pearson Chi-square was high and the p-value quite small for the three relationships; however, this tends to happen in big datasets such as the one used. The most reliable metric is the Cramer's V, which measures the correlation strength between categorical variables from 0 to 1. All of the relationships reported Cramer's V below 0.05, signaling a very weak relationship.

1. **CONCLUSION**

None of the environmental variables displayed a significant relationship predicting the severity of a car accident. Thus, environmental conditions are not good predictors of such severity and no predictive model can be obtained out of them. The great majority of car accidents happen under “normal” environmental conditions: clear weather, daylight, dry road. The Seattle, WA jurisdiction registered no car accident severity above “injury”, the medium label in a scale of five. This suggests that car accidents happening in this jurisdiction are not severe (they never involve serious injuries or fatalities; mostly property damage and injuries) and that such severity is not connected to environmental conditions. Therefore, the severity of car accident might have to be studied regarding human and/or machine related variables (those having to do with the driver, the passengers, or the car). Local government institutions’ effort towards a safer traffic management should aim at a statistical study of such variables.