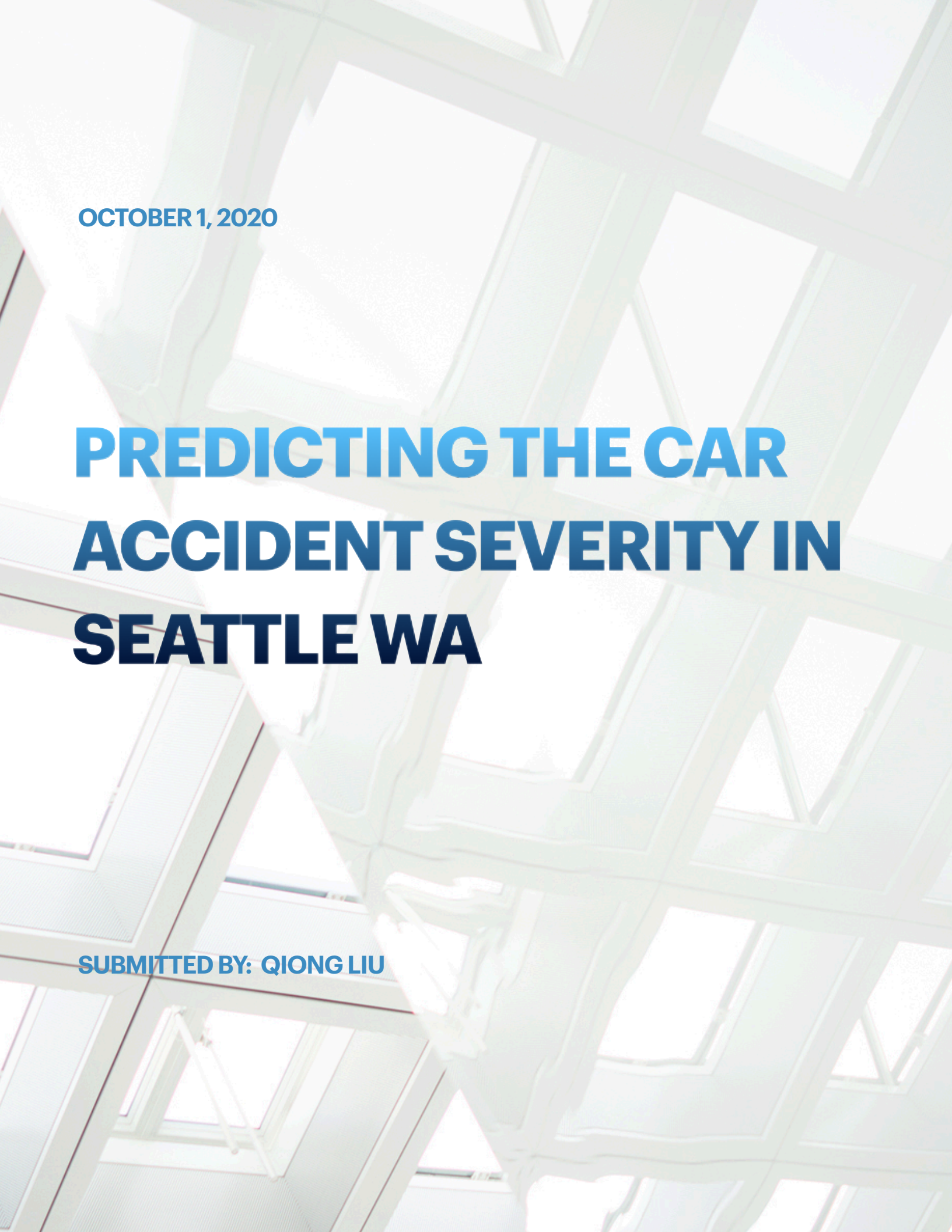


OCTOBER 1, 2020

PREDICTING THE CAR ACCIDENT SEVERITY IN SEATTLE WA

SUBMITTED BY: QIONG LIU

Contents:

IBM DATA SCIENCE PROJECT

1. INTRODUCTION

According to U.S. Census data released in 2019, the Seattle metropolitan area's population stands at 3.98 million, making it the 15th-largest in the United States. As the headquarters of Boeing, Amazon, Microsoft, Seattle attracted widespread attention as home to these many companies and their employees. The city has found itself "bursting at the seams", and with the country's sixth-worst rush hour traffic.Driving to Seattle city to visit some friends or enjoy some seafood is almost a routine for people living in Vancouver.

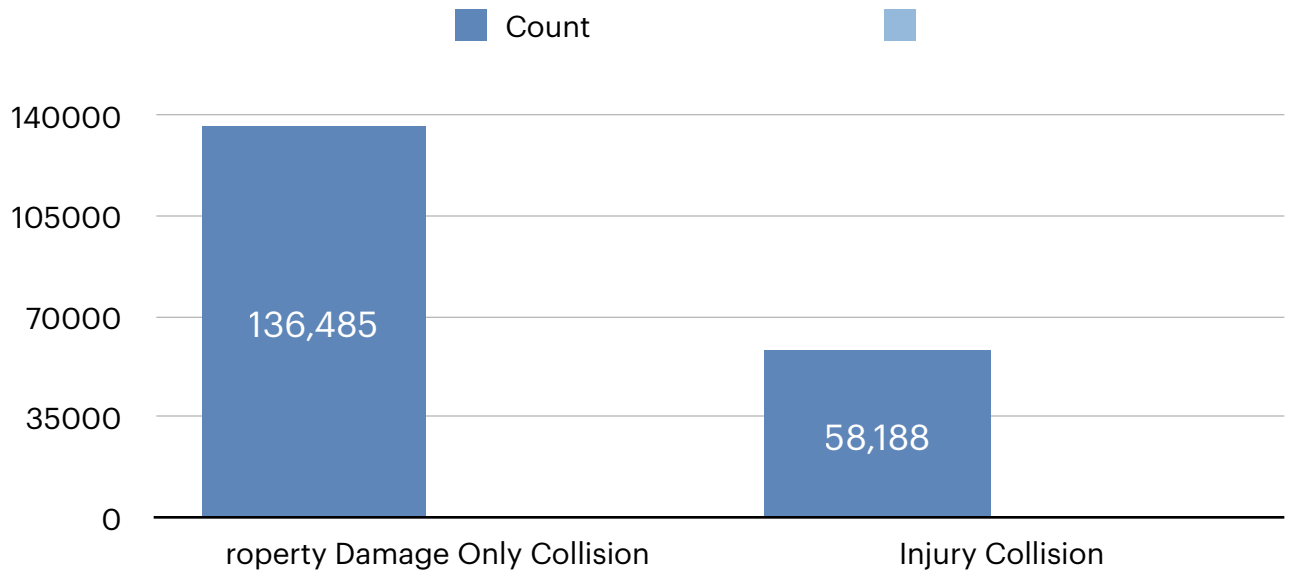
However, it is always rainy and windy in this area, and on the way, you always come across a terrible traffic jam on the other side of the highway, with long lines of cars barely moving. It would be great if there is something in place that could warn you, given the weather and the road conditions about the possibility of you getting into a car accident and how severe it would be, so that you would drive more carefully or even change your travel if you are able to.

Luckily enough, The Seattle Police Department (SPD) has recorded all car collision accident from 2004 to present. Basing on those historical data (194,673 records), we can create a map and information chart to help us understand the high-risk areas, understand car injury factors to avoid accident, and plan our next trip to Seattle better.

2. UNDERSTANDING DATA

2.1 DATA SOURCES

1. Collisions data from 2004 to present. Those data are provided by the Traffic Records Group in the SDOT Traffic Management Division from Seattle, WA. It includes all collisions (194,673 records) provided by the Seattle Police Department and recorded by the Traffic Record, displayed at the intersection or mid-block of a segment from 2004 to the present. Each record has 38 variables/attributes which contains information such as severity code, address type, location, collision type, weather, road condition, speeding, among others. Of the 194,673 records, only 58,188 records are injury collision, so this is **an unbalanced dataset for our research**.

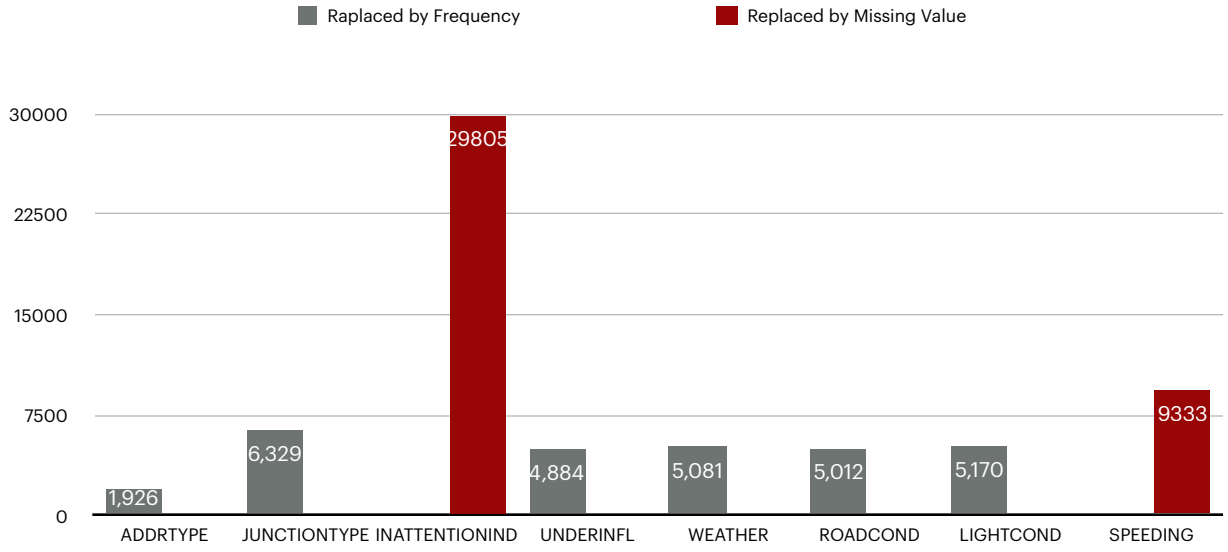


2. Seattle Map data . This can be found at Github searching for "seattle-boundaries-data". And can also be accessed via a JSON API by using boundaries-api.seattle.io.

2.2 DATA CLEANING

2.2.1 Drop unrelated features

As we want to analyze what factors will probably cause a car collision and the severity of the accident, we would drop those unrelated features and useless information, only keep



15 features/attributes of the original dataset. Those 23 features we dropped/deleted are: 'OBJECTID', 'INCKEY', 'COLDETKEY', 'REPORTNO', 'STATUS', 'INTKEY', 'LOCATION', 'EXCEPTRSNCODE', 'EXCEPTRSNDESC','PERSONCOUNT', 'PEDCOUNT', 'PEDCYLCOUNT', 'VEHCOUNT', 'SDOT_COLCODE', 'SDOT_COLDESC', 'PEDROWNOTGRNT', 'ST_COLCODE', 'ST_COLDESC', 'SEGLANEKEY', 'CROSSWALKKEY', 'HITPARKEDCAR'.

2.2.2 Handle missing values

- Convert "?" to NaN.

We replace "?" with NaN (Not a Number), which is Python's default missing value marker, for reasons of computational speed and convenience.

- Replace missing value by frequency/

Whole columns should be dropped only if most entries in the column are empty. In our dataset, none of the columns are empty enough to drop entirely. Basing on the character of the data features we choose, we mainly replace the missing value with the most frequent values.

However, feature "INATTENTIONIND" and "SPEEDING" only have "Y" value, thus we replace the missing value in those two columns with "Y".

2.2.3 Correct data format

Two features ("INCDTTM" and "INCDATE") should NOT be object type, thus we change those two columns with "datetime64" type.

And then use those data to calculate which hour ("hourofday") and which weekday ("dayofweek") those accidents happened.

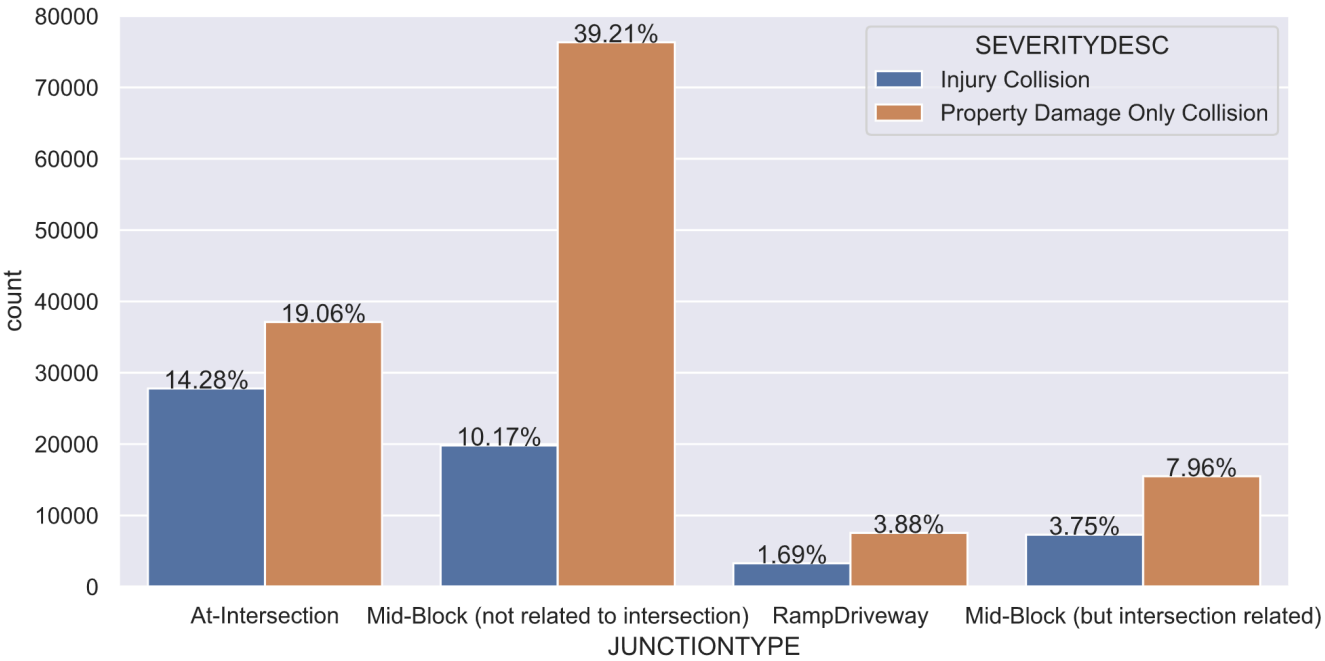
2.2.4 Delete/Drop some rows

We want to use location data to map the accident, so have to drop more than 5000 records who don't have X and Y data.

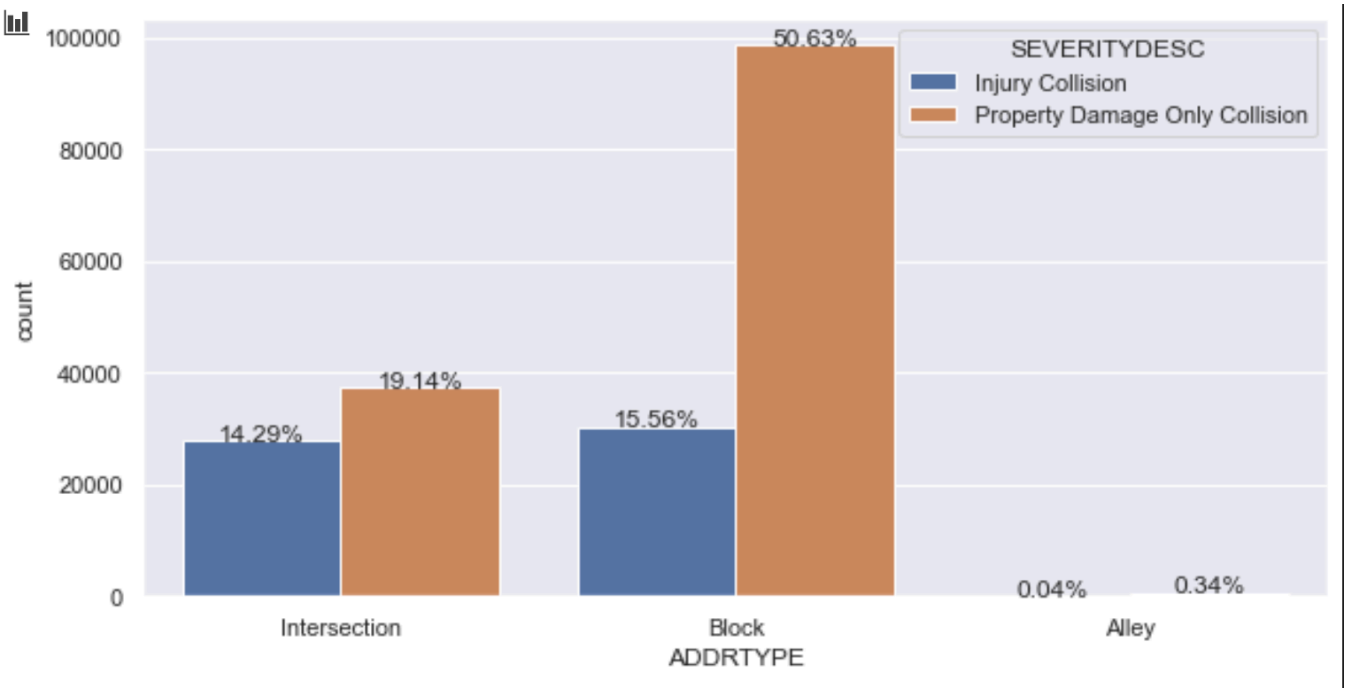
2.3 DATA ANALYSIS AND DESCRIPTION (ANALYZING INDIVIDUAL FEATURE PATTERNS USING VISUALIZATION)

2.3.1 Junction Type have big impact on severity

From the plot above, we can see "At Intersection" is an important factor, where the accidents are more likely to involve injury, only a little less than the chances of property damage.

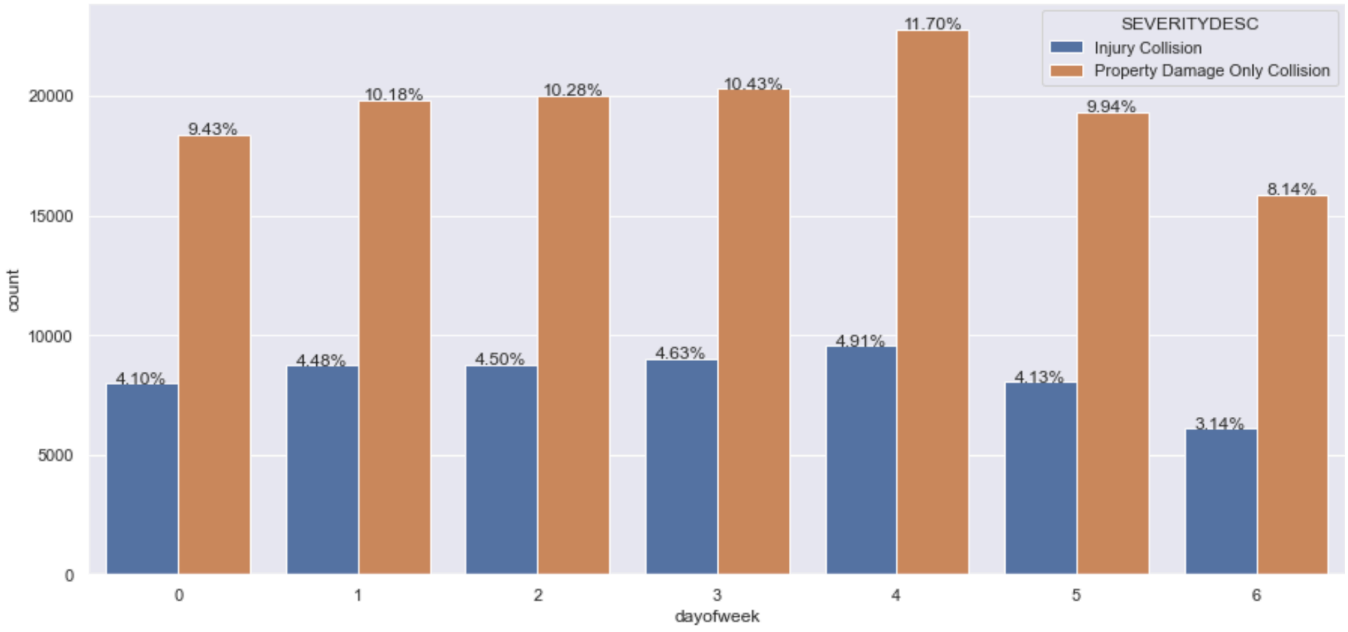


2.3.2 Address Type data also show that Intersection greatly influence the severity

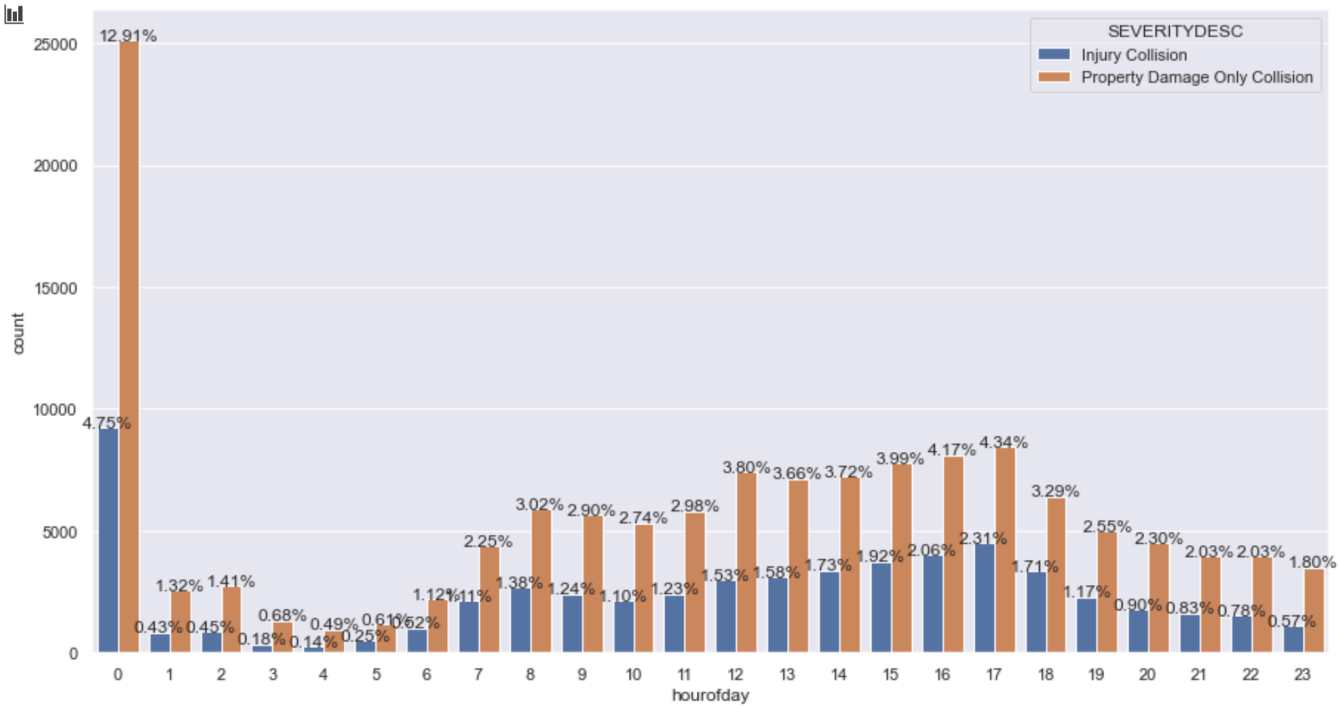


2.3.3 No evidence show that weekday plays a role in influencing severity

Previous I thought weekend will be more likely to have injury collision, but data shows that there is no clear evidence.

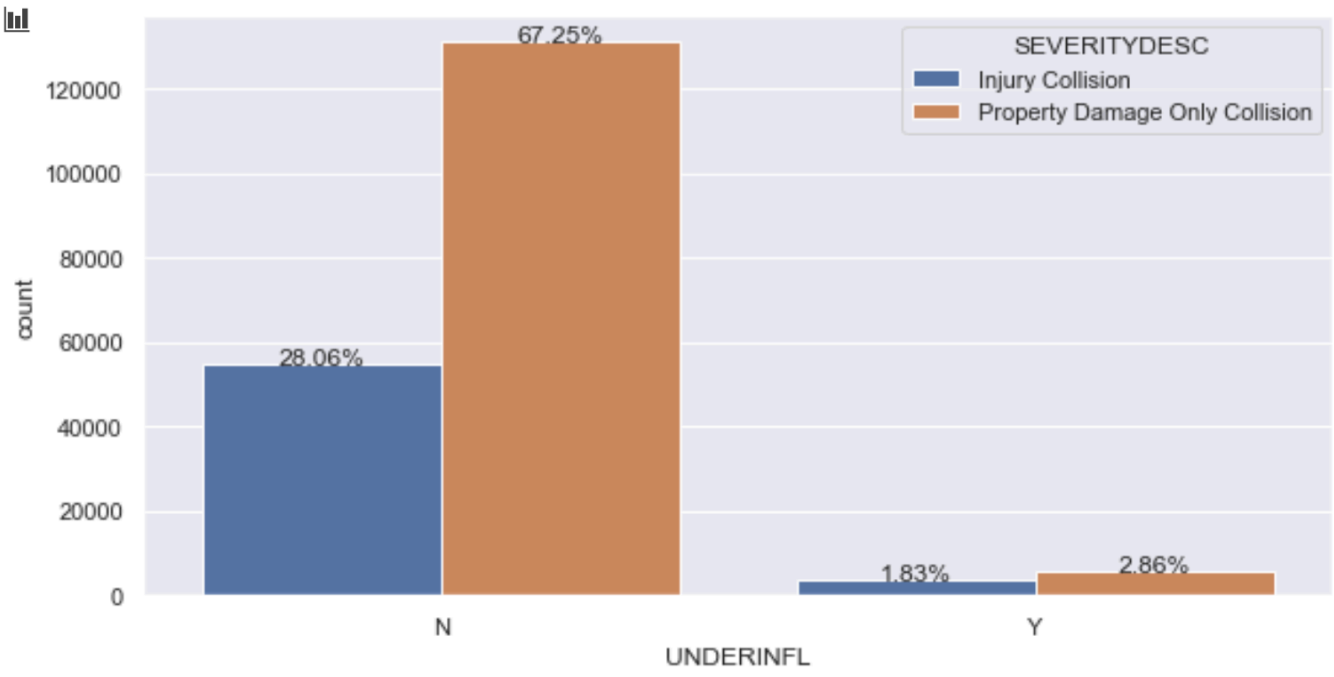


2.3.4 Hour of Day seems plays a role.

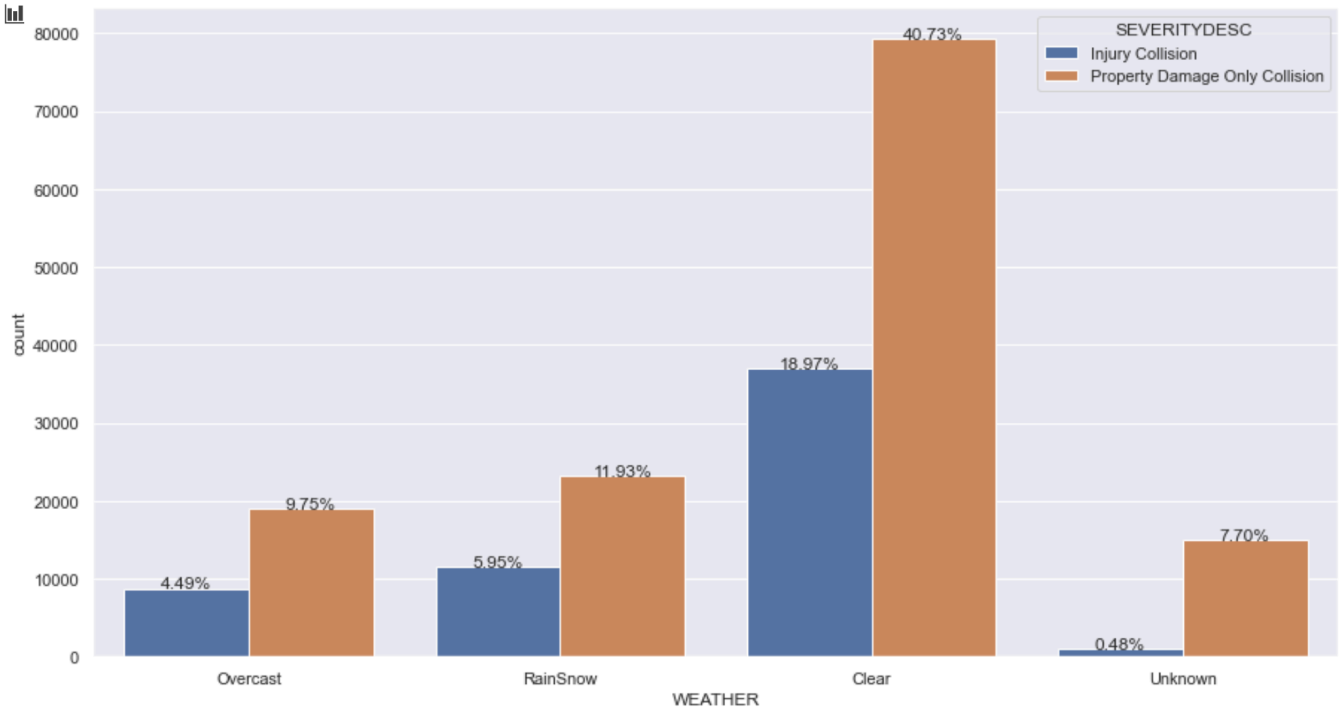


We can see that early morning will be less risky to drive and are less likely to have injury, and there are differences between the trend of day and night.

2.3.5 Drug and alcohol will greatly increase the possibility of injury collision

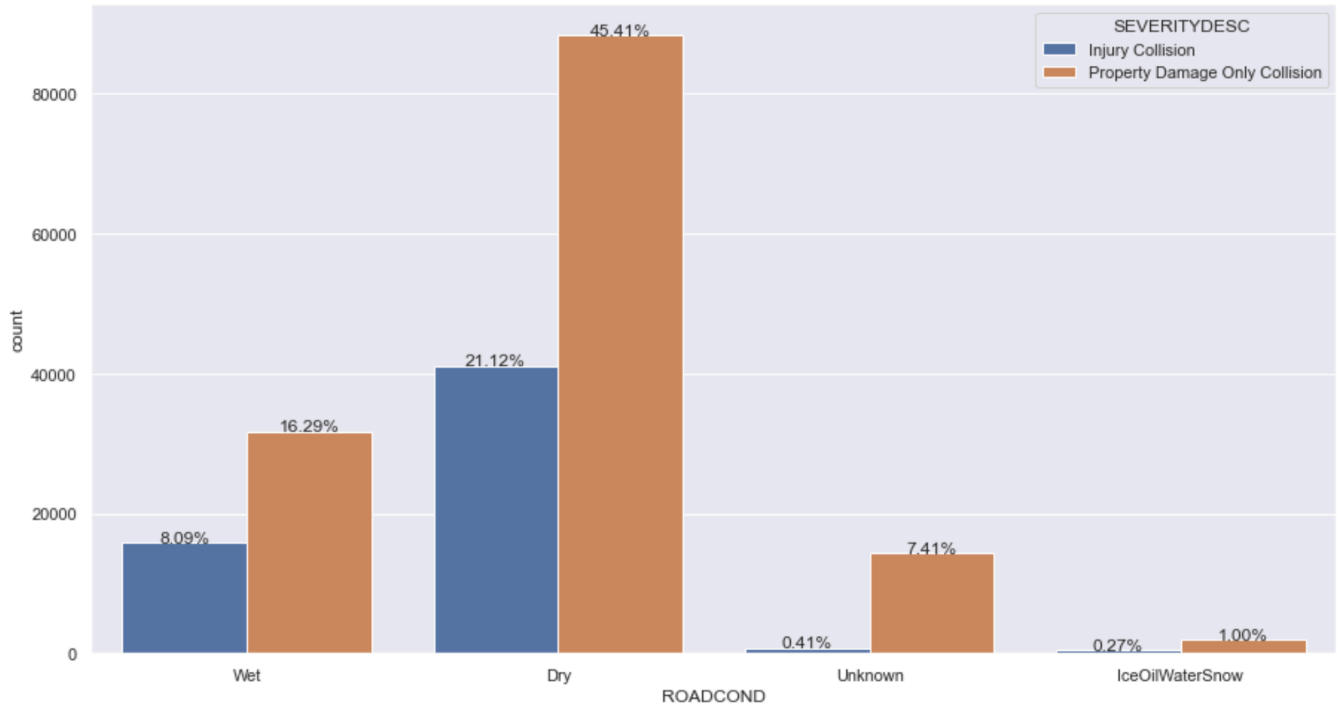


2.3.6 "Unknown" weather condition plays an important role



However, we can see the "unknown" weather condition differs hugely with other weather, we should get more about the "unknown" data to dig into and find out the reason.

2.3.7 "Unknown" Road condition has the same pattern with weather



2.3.8 Light condition is an important factor