Team 1: Danting Huang, Aparna Raman, Kai Sun, Yanyue Fu, Jinyan Yu
BA888 - Capstone Project
Faculty Guide: Mohammad Soltanieh-ha
Github Link: https://github.com/Jinyan-Yu/Capstone_project_BA888
5/1/2020

How would Uber Consumer Behavior Change if we Knew More about Car Crashes?

Abstract:

Car sharing has become an integral part of our lives in the modern age. From the time that Uber and Lyft were founded in 2009 and 2012 respectively, they have changed consumer behavior like never before. Hopping into a strangers vehicle was redefined by these companies and consumers started treating this as a way of life. In today's world *ordering an uber* is treated as a verb more than a service. But, would this consumer behavior change if customers start losing faith in the brand or if the safety of the service provided by these brands is compromised. Our analysis tries to answer some of these pressing questions that will help the ride sharing industry in getting, keeping and growing their customer base by making their service safer for their everyday customers. Our project will also help companies like Uber and Lyft in making their driving partners feel safer by avoiding some of these fatal and nonfatal crashes by having complete information about the road types, weather conditions and manner of collision that most affects these accidents that are life and property threatening.

Our ultimate goal during this project is to be able to make a recommender system that allows drivers to have an app feature that would allow them to learn about the most preferred or safest route to travel to a particular destination that would avoid any kind of crash at all times. We have taken into consideration multiple attributes to determine these *safe routes.* Some of our attributes contain how weather affects car crashes, how different location and destination pairs differ depending on how safe they are and how fatal a crash could be based on the manner that the vehicles collided.

Understanding Our Data

To enable our analysis, we decided to explore 2 sets of data. One that gave us better understanding and knowledge of our service provider in question, *Uber* and another dataset that re-instilled our knowledge about crashes in the city of Boston and what are the factors that are closely associated with these fatal and non-fatal accidents that have often resulted in loss of life and property.

Post the measures we took for cleaning our *Uber dataset* consists of 57 variables after which has been divided into five categories: time, location, car type, price and weather situation. Further from this categorization, half of our variables are concerned with weather and give us a better understanding of the weather condition associated with that particular ride. Some weather conditions that were associated with our rides were, *temperature and pressure* which has been combined with the data and route information, to analyze price and location.
Link: (https://www.kaggle.com/brllrb/uber-and-lyft-dataset-boston-ma)

Further to understand how crashes take place in the metropolis of Boston, we looked closely into the *Boston crash dataset*.This dataset contains road crash data for cars that have met with accidents in the Boston area in the period 2010- 2019. The dataset contains 25 variables, and can be further divided into four categories: crash time, crash location, types and severity of car crashes. This data is being utilized to gain an in-depth understanding of the crashes in the Boston area and map these based on various criterias such as weather, type of car, fatality based on manner of collisions etc. Further, this data set is being utilized to give recommendations to likes of Uber in the future to develop an app feature that would notify drivers with the best possible routes as well as which routes have a higher likelihood of accidents.
Link:(https://drive.google.com/drive/folders/1gcJFDgBqYQFqVsed2epWGXH4W-9ncMhpsp=sharing)

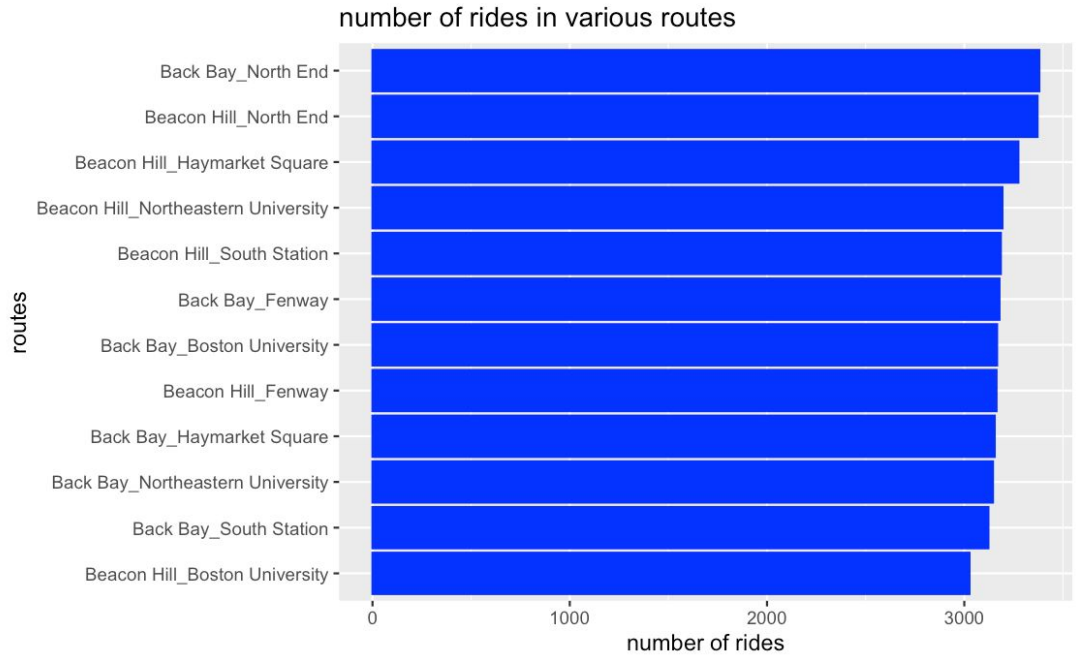
Key variables:

1. ***Crash Number*** – Unique number used by Registry of Motor Vehicles to identify each crash.
2. ***Crash Severity*** –Type of Crash, such as Fatal injury, Non-Fatal injury.
3. ***Number of Vehicles*** – Total number of vehicles involved in the crash
4. ***Total Nonfatal Injuries*** - Number of persons injured in the crash excluding fatalities
5. ***Total Fatal Injuries*** - Number of persons killed in the crash
6. ***Manner of Collision*** - Manner of Collision or Collision Type
7. ***Vehicle Action Prior to Crash –*** The action that each vehicle was taking prior to the crash; V1 = Vehicle 1, V2 = Vehicle 2, etc.
8. ***Vehicle Configuration*** – The type of each vehicle involved in the crash
9. ***Road Surface Condition –***The condition of the road's surface at the time of the crash
10. ***Ambient Light –*** Light conditions
11. ***Weather Condition*** – A maximum of two weather conditions may be reported

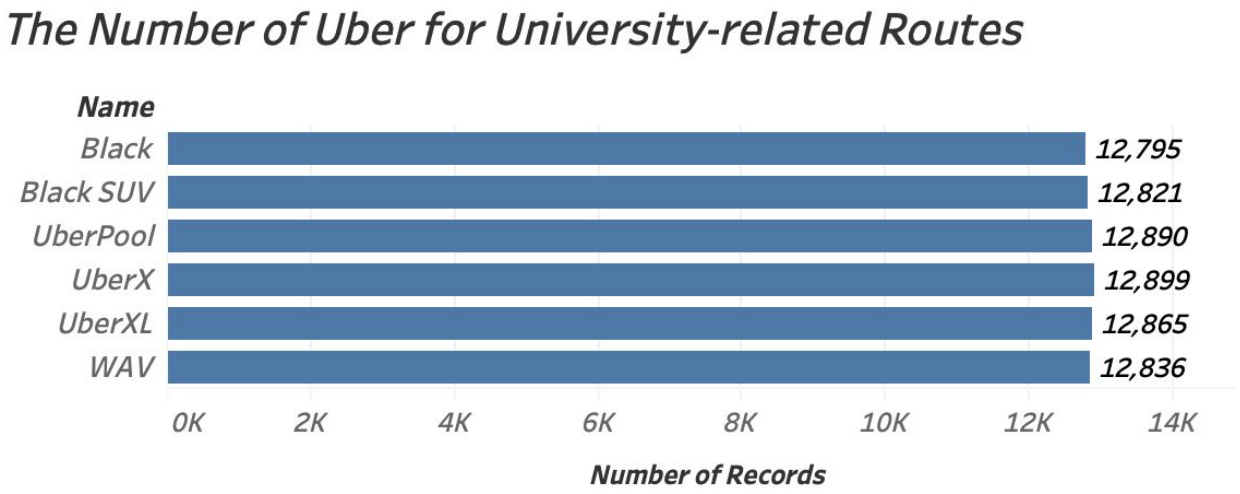
Exploratory Data Analysis

Uber Route Analysis: Top destinations and Average price

The route of a cab service gives us in-depth understanding of many aspects of the service: from the most popular places in a city or town to the busiest and costliest times to travel. In our dataset, there are 12 starting points for the journeys and 12 destinations in the downtown part of Boston, which include locations such as Boston University, Northeastern University and the Fenway Park area. We assessed these routes by charting the connections between the starting and end points of these routes and using our knowledge of data visualization to exhibit the busiest routes in all of downtown Boston. Based on our analysis, we have found 72 such routes. The below graph exhibits the 12 busiest routes in all of Boston downtown according to our data analysis of the Uber travel records.



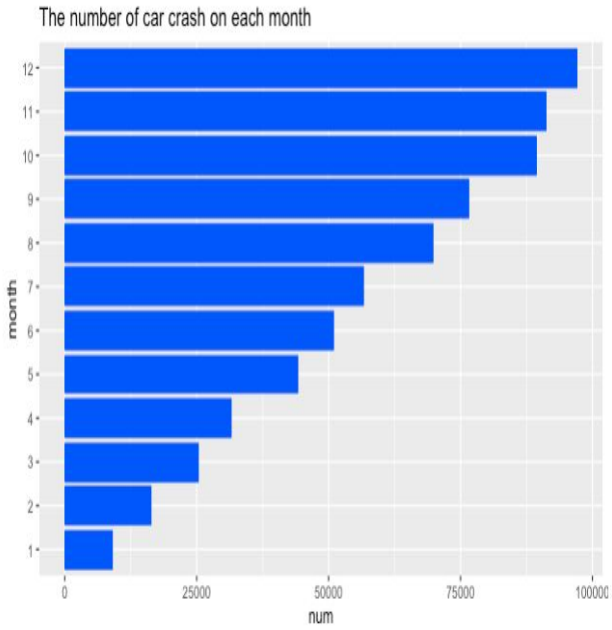
Our main assumption was that, since Boston is a huge college town most of these rides would be between 2 of the largest Universities in Boston downtown: Boston University and Northeastern University. Further, we also assumed that, majority of the rider occupancy must be of students which should lead to a large number of rides being called for the UberPool service. The UberPool service is a low cost service option offered by Uber, where customers with a similar source and destination location can split the cost of the ride. This assumption of ours was challenged by our continued analysis of the data set.



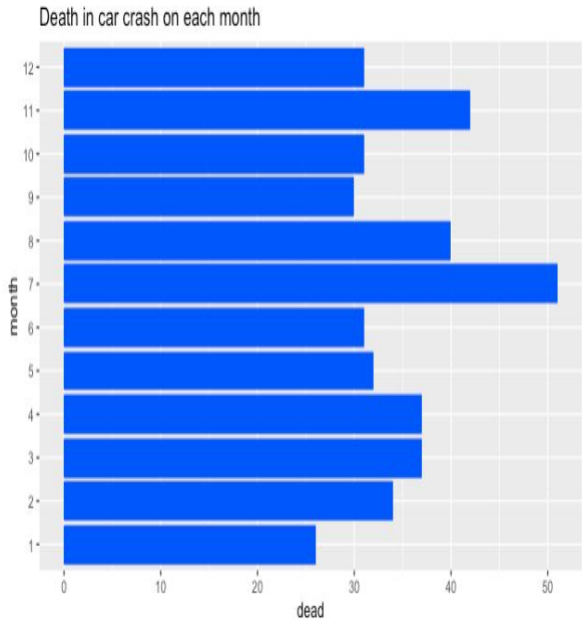
Further analysis of the dataset, led to the calculation of mean price per route, to understand our service provider further. To our surprise, the university-related routes have an average price of $17.3 while the other routes have an average price of $15. It will not be an exaggeration to conclude from this analysis that students against popular belief do not avoid the use of high price services such as UberX and Uber Black. The average price of the university routes to much of our surprise is fifteen percent higher than that of other routes.
Furthermore, based on our analysis of average price, we were able to analyze various routes. Using which, we observed that the most expensive route is the one from Boston University to the financial district. The price of this route is more expensive than that of others, so it would therefore be safe to say that consumers traveling from Boston University to the financial district are willing to pay a higher price for the uber service.

Weather analysis

In our initial analysis, we explore the number of car crashes that occured in the period between 2010 - 2019, to assess the number of deaths resulting from accidents in this period. The interesting thing that we found is that the total number of car accidents were increasing month by month.



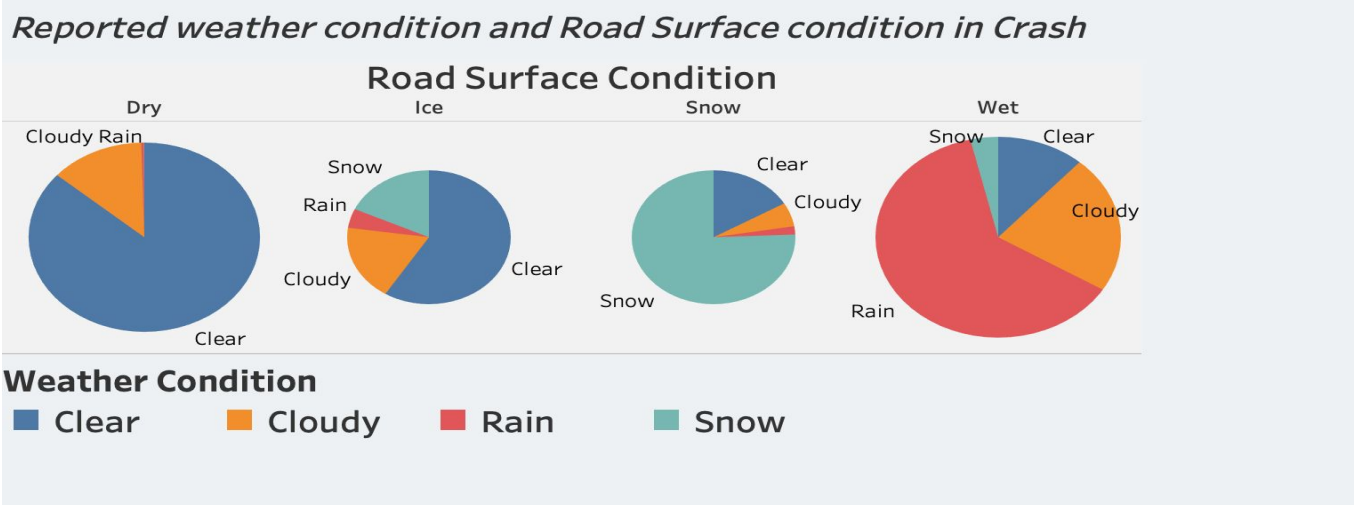
Even though the number of car crashes are extremely high, we were able to observe that the death rate is actually miniscule, nearing 38 deaths for each month in a period of 9 years. The chart on the left, describes how the number of crashes in the Boston area have increased steadily over the course of time, making the month of December the most fatal month with the most number of crashes. It is interesting to observe how different months have different weather conditions and that might be largely affecting the number of crashes. It is also essential to note that some months are notably more busy in the Boston area such as the academic year between september and May, but surprisingly this was not captured by our data.



Our number of deaths on the other hand did not go in accordance with the increased crashes in the city of Boston. Our above graph analysis depicts that the month of December has the highest number of crashes, where our graph on the left has led us to conclude that the month of July has the highest number of deaths, which is even more than some of the highly fatal months according to our above graph. Looking at this initial analysis, we had deducted many theories including the role of dew, most and fog adding to low visibility in the winter months, causing the drivers to meet with

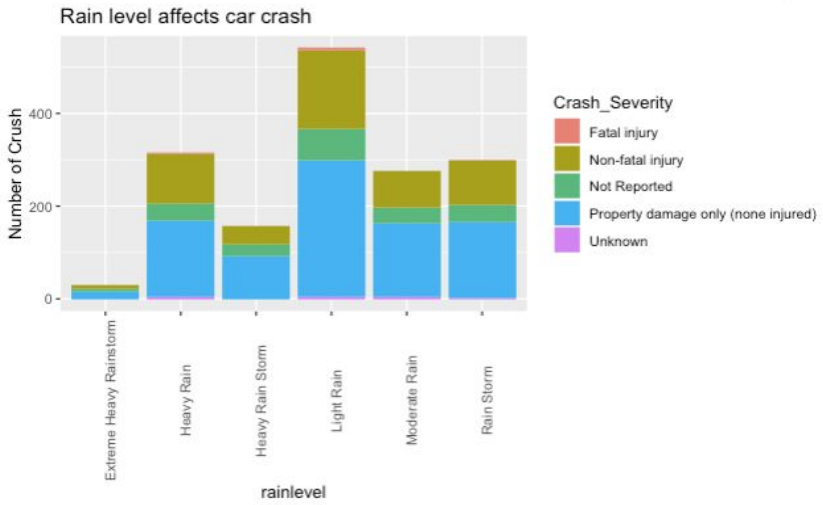
accidents. This assumption was certainly challenged in our further analysis.

After learning about the high number of car crashes that took place in the Boston area, our initial assumption was that the weather situation strongly affects car accidents. For further analysis, we joined our 2 datasets: the car crash dataset and the weather dataset to find relationships between weather and car crashes. Certainly, different weather conditions affect the road surfaces differently. In our analysis we inferred that, snow causes the road surface to freeze and the car to skid, which in turn increases the probability of road accidents. A close look at the below chart will help in understanding the different road conditions and the climatic variance that affects them differently. In this graph, the varied sizes of the pie charts represent the number of car crashes, which in turn gives us a better understanding of the varied road surfaces that cause the highest car crashes. Using this analysis, we were able to conclude that both dry and wet road surfaces pose a high risk of car accidents. Since our goal is to assess the influence of weather on car accidents, we chose to take into account the main weather condition: *rain* that causes the road surface to turn, and this composed of our further discussion.



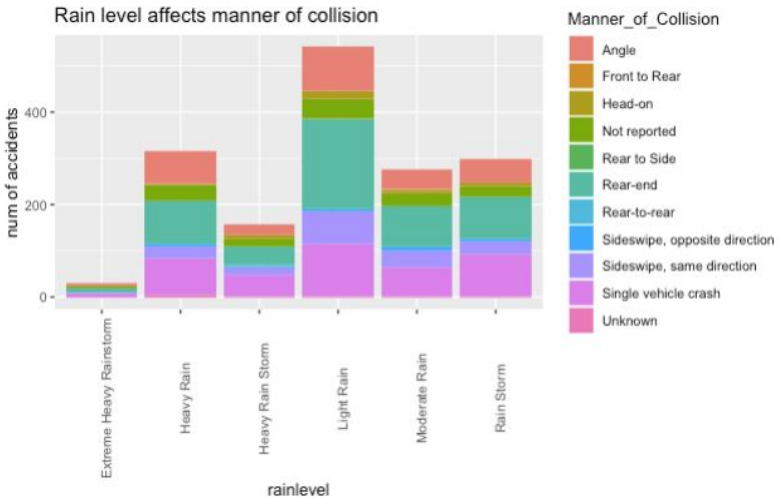
We often in our daily lives assess the intensity of the rain. Similarly for this project, we realized that differently intensity of rains has a high likelihood of affecting crashes differently. To understand the correlation between rain and crashes we used the precipitation variable to define rain level for 7 situations:

- No Rain: precip=0
- Drizzle: 0<precip<0.01
- Light Rain: 0.01<=precip<0.1
- Moderate Rain: 0.1<=precip<0.25



- Heavy Rain: 0.25<=precip<0.5
- Rain Storm: 0.5<=precip<1
- Heavy Rain Storm: 1<=precip<2
- Extreme Heavy Rainstorm: precip>=2

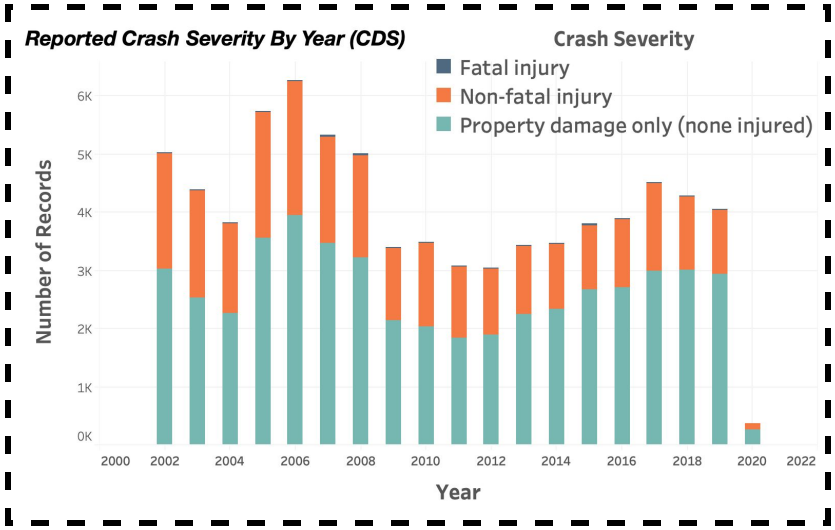
This interrelatedness analysis helped us understand that except for situations where there was *no rain*. A *light rain* weather condition causes a high risk of car crashes, but it is interesting to note that all of these crashes are non-fatal injuries and only lead to the damage of property. Our further analysis led us to conclude that different conditions of rain most certainly cause accidents to occur, but the fatality associated with most of these conditions is low.



From our analysis that helped us assess the effect of rain on car crashes and also how the different intensities of rains determined the falatily factor, we posed an even interesting correlation. Our assumption this time strongly inclined us towards believing that the different parts of a vehicle involved in a collision, lead to different intensities of fatality. In this particular analysis we are trying to also help insurance companies that provide services to companies such as Uber to price their insurance products accordingly. Further, the analysis of our assumption gave us a strong understanding of how collisions on different body parts of the car, in different rain intensities, lead to different levels of fatality. The graph on the right depicts how the angle, rear-end and sideswipe are three most likely spots for collision during a car accident. It is therefore highly advisable for drivers to be aware of these spots and weather conditions and furthermore, maintain a safe distance to avoid these collisions.

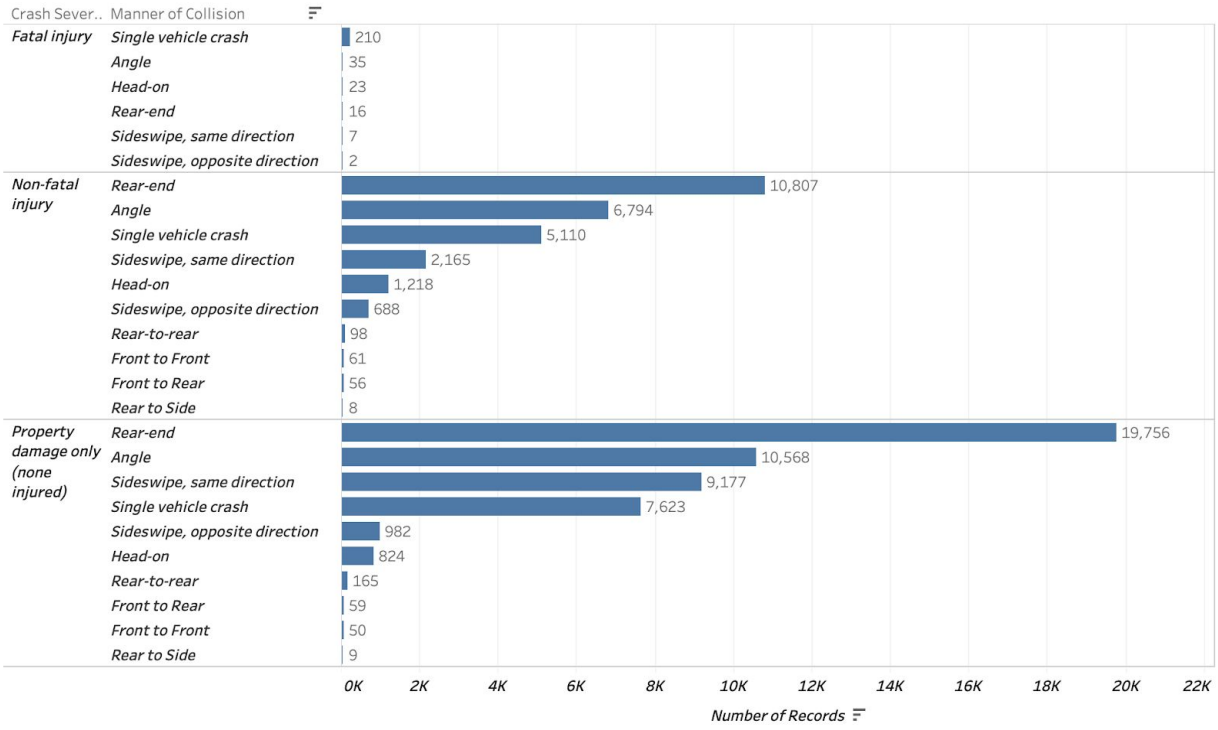
Collision analysis

In the period between 2002 and 2020, 64.3 percent of all car crashes that occurred were mild, and caused a mere loss in property. Whereas, 35.3 percent of those accidents caused non-fatal injuries and a miniscule number of 0.4 percent of these accidents led to fatal injuries. To further enhance our understanding of the intensity of collisions we parsed our crash severity into three categories: fatal injuries, non-fatal injuries and a mere property damage. As the figure (“Reported Crash Severity By Year”) shows, the amount of crashes has increased between 2010 and 2019. The only significant increase that we observed was in the number of crashes that were causing a mere property damage, whereas the crashes that were causing non-fatal injuries were on a decline in the period.



We further believed that there exists a strong relation between the manner of collision and crash severity. Figure below (The Number of Each Collision Manner based on Crash Severity) shows the crash severity based on different manners of collision. For fatal injuries, 83.6 percent of people died because of ‘single vehicle crash’ and ‘angle crash’. On the other hand, for non-fatal injuries, over 40% of people got hurt as a result of rear-end crash and 25% caused by angle crash. Compared with situations where fatal injuries are involved, single-vehicle-crash makes up almost 19% of the total in the non-fatal scenario and therefore should not be completely ignored. For situations where there is a mere property damage, the top three collision manners are rear-end, angle and sideswipe (with same direction), accounting for 40.1%, 21.5% and 18.6% respectively. It can be strongly inferred that, no matter whether people died, got injured or there was merely property damage in a car accident, rear-end crash and angle crash are situations we cannot underestimate and have the potential to cause larger damage.

The Number of Each Collision Manner Based on Crash Severity



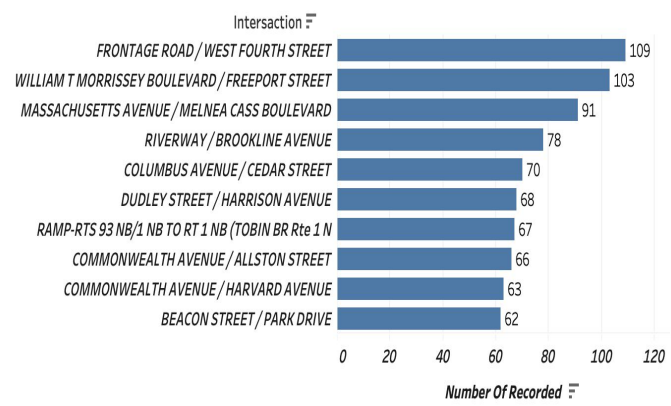
Further, our assumptions took us to the analysis of which manner of collision is most dangerous. To understand this better, we designed a severity score system based on "KABCO Injury Classification Scale and Definitions". We set ratings for each class. For example, 6 for fatal injuries, 4 for non-fatal injuries and 2 for mere property damage and 0 for 'unknown' and 'not reported' situations. Based on this weight system, we were able to know that *Head-on* is the most severe collision and the second most severe is a single vehicle crash. It is therefore safe to conclude that we need to pay more attention to people's driving behaviors and situations of drunken driving or drugged driving should be severely prohibited.

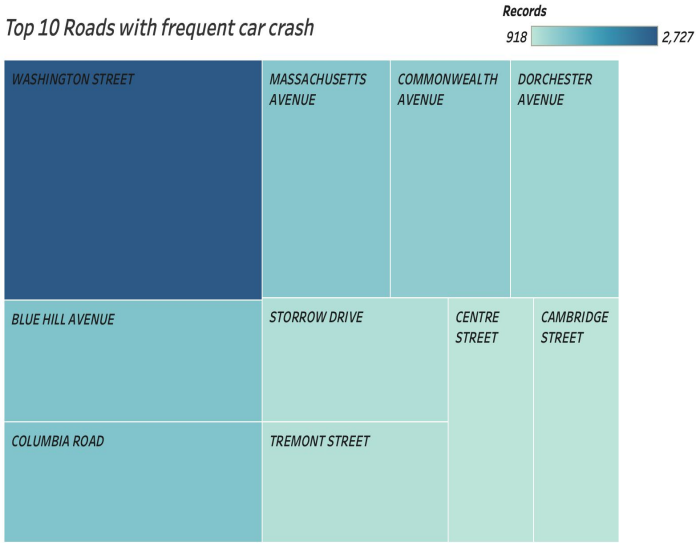
Average Severity Score for Manner of Collision

Manner of Collision	
Head-on	2.5800
Single vehicle crash	2.5120
Angle	2.4110
Rear-end	2.4020
Front to Front	2.1500
Sideswipe, opposite direc..	2.0910
Front to Rear	2.0730
Rear-to-rear	1.9620
Sideswipe, same direction	1.9540
Rear to Side	1.9230

We strongly believed that the intersections with frequent traffic accidents and a high volume of car movement versus intersections that are primarily low traffic seeking experience car crash situations at a different volume. While assessing this we inferred that the most frequent crash intersection is between Frontage Rd and West Fourth St, but it is also important to notice that there are only 109 records between 18 years because of the missing values. Thus we tried to find the top 10 main streets with frequent crashes. There were 2727 crashes that occured on Washington St, 1375 crashes occurred on Blue Hill Ave and 1368 crashes on Columbia Rd. It is

noteworthy that the number of accidents on Washington St is twice as many as the ones that occurred on the second most street intensity high location: *Blue Hill Ave*. Further analysis points out that this is because Washington St is the longest street in Boston, and it remains one of the longest streets in the state of Massachusetts and therefore gets the most amount of vehicular traffic.



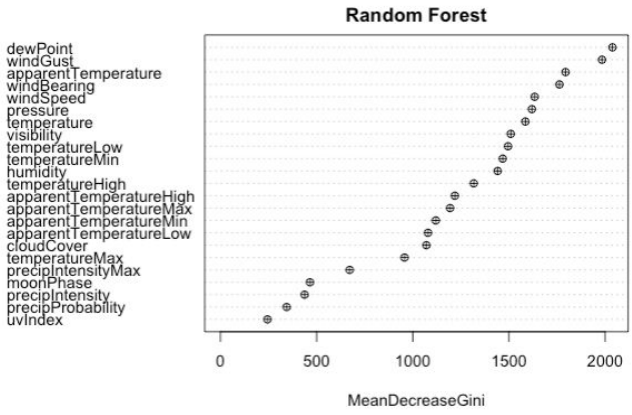


Results

When we started this analysis, we wanted to make uber rides safer by first understanding the routes these service cars traveled on. Our understanding then made us conclude that in Boston, the route from Back Bay to North End is the busiest during our research period. In this situation, it is advisable to pay a higher degree of attention to this kind of route to keep most people safe on the busy streets. We are further suggesting that our service provider in question, *Uber* should try to implement a system where they set an alarm functionality for the drivers based on the current Uber rides condition. The government entities can also set an alert system for vehicles based on the current traffic situation that would inform drivers of the severity of weather, car count and road conditions on the said route. On the other hand, some routes which cost the most based on our analysis of the average price of Uber's taken, the route from Boston University to the Financial District, also needs Uber's attention, since the passengers spend more on this route for a higher level of hospitality and precision. We observed that of the 200,000 rides, about 70,000 rides involve a route associated with universities, not to mention the higher price of university-related routes, to maintain the highest possible safety measure on these routes during rush hours in particular we recommend having an alert system.

Furthermore, based on the pricing models we were able to infer that when prices for different uber services are similar to each other or when safety and time are a matter of key importance, consumers are likely to choose more expensive Uber service offerings, such as UberX and Uber Luxury instead of Uber Pool. In order to improve the company's service quality, the system should arrange more UberX and Uber luxury cars on the busiest and high risk routes. The drivers of the cars on these routes should also be highly experienced and reminded of the risks that these routes entail.

We also want to make sure Uber drivers and customers feel safer in different weather situations by knowing which weather condition causes car crashes to occur the most. Since we have an Uber data set for a 3 weeks period from November to December 2018, and our weather data set is from 2010 to 2019, we combined the 2 datasets to explore the impact of weather on car accidents during these 3 weeks from November to December 2018. We selected all the weather variables in the dataset such as temperature, humidity, visibility, and wind to predict whether or not these conditions would cause a car crash to occur. We used the random forest method to predict the influence of weather factors on car accidents, and finally found 4 important factors: dew point, wind gust, apparent temperature and wind bearing. We were able to infer from this analysis that Uber's dataset in the winter, dew point, wind conditions and temperature has a greater impact on the crash. We considered a situation where the dew point is related to an icing situation on the road, and the low dew point leads to the road frosting or icing of the roads, further leading to the roads becoming extremely slippery, and thus increasing the probability of accidents. The weather in different seasons or months can have different effects on car accidents. Although the weather factor is one of the main factors that lead to road accidents, we continue to predict the impact on road accidents caused by other variables of different types, so as to make safer suggestions for Uber.



In order to see the big picture of which factors impact the car crash severity the most, we used OLS regression to show clear correlations for different variables with the crash severity. Below is a summary page of the regression result:

OLS Regression Results

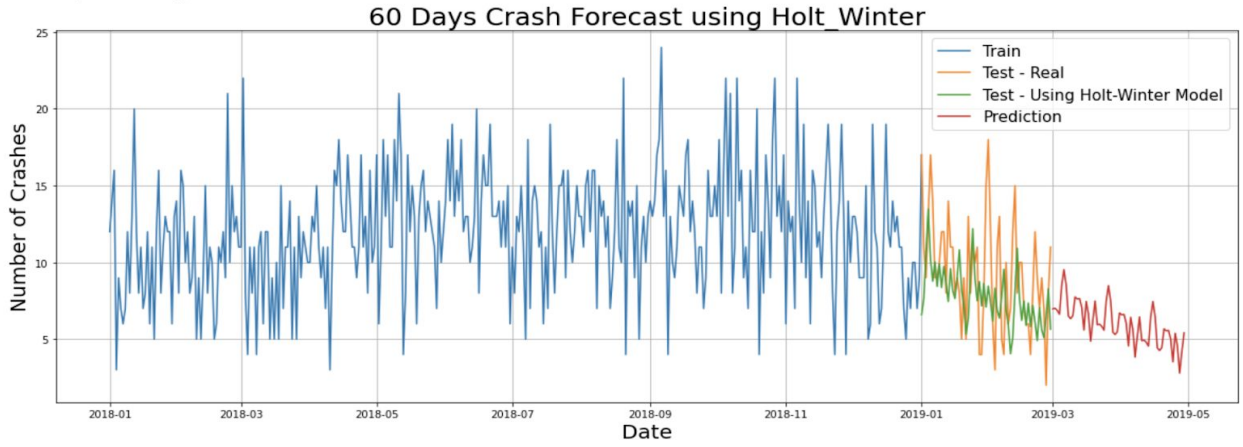
Dep. Variable:	Crash_Severity	**R-squared (uncentered):**	0.933
Model:	OLS	**Adj. R-squared (uncentered):**	0.932
Method:	Least Squares	**F-statistic:**	3718.
Date:	Wed, 15 Apr 2020	**Prob (F-statistic):**	0.00
Time:	15:22:37	**Log-Likelihood:**	-4374.7
No. Observations:	4582	**AIC:**	8783.
Df Residuals:	4565	**BIC:**	8893.
Df Model:	17		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Manner_of_Collision	-0.0019	0.003	-0.679	0.497	-0.007	0.004
Most_Harmful_Events	-0.0018	0.000	-8.073	0.000	-0.002	-0.001
Vehicle_Configuration	0.0001	0.000	0.680	0.496	-0.000	0.001
Road_Surface_Condition	0.0055	0.004	1.502	0.133	-0.002	0.013
Ambient_Light	-0.0054	0.005	-1.109	0.267	-0.015	0.004
Weather_Condition	-0.0022	0.001	-1.735	0.083	-0.005	0.000
At_Roadway_Intersection	7.286e-05	2.52e-05	2.894	0.004	2.35e-05	0.000
Distance_From_Nearest_Exit	6.448e-05	0.000	0.285	0.775	-0.000	0.001
Distance_From_Nearest_Landmark	-4.879e-06	4.67e-05	-0.105	0.917	-9.64e-05	8.66e-05
Number_of_Vehicles	0.1185	0.014	8.670	0.000	0.092	0.145
Total_Nonfatal_Injuries	-0.8423	0.012	-72.865	0.000	-0.865	-0.820
avg_temp_(f)	0.0357	0.003	13.773	0.000	0.031	0.041
avg_dew_point_(f)	-0.0394	0.003	-14.275	0.000	-0.045	-0.034
avg_humidity_(%)	0.0245	0.001	23.725	0.000	0.023	0.027
avg_visibility_(mi)	0.0461	0.006	8.205	0.000	0.035	0.057
avg_wind_(mph)	0.0065	0.004	1.686	0.092	-0.001	0.014
high_wind_gust_(mph)	0.0016	0.002	0.993	0.321	-0.002	0.005

Omnibus:	673.249	**Durbin-Watson:**	1.925
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	4999.168
Skew:	0.484	**Prob(JB):**	0.00
Kurtosis:	8.025	**Cond. No.**	2.04e+03

From the regression results above we are able to infer that the number of vehicles involved in a crash has a significant impact on the crash severity of the crash. Besides that, various weather factors like average temperature, humidity and visibility also greatly affect the severity of crashes.

We further did an analysis where we ran a time-series prediction based on Holt-Winters model to see the results of future 60 days prediction of the number of crashes. First of all, we seperated the data into training and testing. The entire data for the year of 2018 was used as a training set while the first quarter of 2019 was used as testing. After testing, we are able to make future 60 days car crash predictions. Below is the graph showing the training, testing and prediction result.



Conclusion and Room for Future Implementation

In order to achieve our ultimate goal to make a recommender system that would let drivers know what a good (safer) route for them or what is a preferred route at a certain time, we have conducted exploratory analysis on both Uber and crash datasets, as well as regression and prediction analysis. As the results shown above, we are able to draw the conclusion that several weather and crash factors could affect the risk of crash.

Based on our observation, the recommender system we suggested should include the following functions:

- Drivers will receive notifications when passing through areas with high crash risk.
- Uber will reroute upon drivers' and passengers' requests based on the volume and price.
- As weather could affect road safety, Uber could increase price accordingly based on weather and road surface conditions.

However, we still find room for improvement as the recommender system takes time and effort to build and maintain. The accuracy of recommendation still needs to be tested.

Other Sources Used:

1) https://safety.fhwa.dot.gov/hsip/spm/conversion_tbl/pdfs/kabco_ctable_by_state.pdf