ECO395M Final project: Exploring about car crashes in Austin

Yuri Lee, Seunghoon Choi

2021 5 12

A. Abstract

We analyzed the car crash data and weather data for two years(2018, 2019) in Austin in order to predict the number of daily car crashes, the daily percentage of severe car crash injuries, and if one crash will have injuries or not under specific external circumstances: these are outcomes of this research. After analyzing the relations between all possible x variables in the data sets and our target outcomes and examining the relations, several variables were chosen for the prediction models: month, minute of day, day of week, road category, precipitation, etc. Next, we fitted the prediction model to predict outcomes. The linear/stepwise, Random Forest models were applied for prediction of the number of daily car crashes and the daily percentage of severe car crash injuries, while two logit models were used for prediction of injury-inducing car crashes. Comparing the out-of-sample RMSEs of the models gave us the best model with the highest accuracy. We could find correlation between three outcome variables and several x variables such as time, day of week, roads' category and precipitation, and predict the daily car crashes, the daily percentage of severe car crashes and injury-inducing car crashes through the research.

B. Introduction

Have you ever heard that every year, more than 38,000 people die in car crashes on U.S. roadways? If you add seriously injured people who need considerable amounts of medical treatments, the number of victims from car crashes might be more than double of the total deaths. However, nobody knows exactly when and where car accidents often occur, and which external features except very individual driving circumstances affect car accidents. Of course, we have vague thoughts that we need to pay more attention to driving under the bad weather such as rain or snow, or in a heavy traffic but we don't know exactly if car accidents are related to features such as weather, roads, time, etc and how much the effects would be, if related features exist. So, we did research about characteristics of car crashes in the Austin area and which features are closely related to the quantity of car crashes and severity of them, and how to predict the number of car crashes and human damage from them.

C. Methods

We obtained the car crash data from 01/01/2018 to 12/31/2019 in Austin from the Texas Department of Transportation. A reported motor vehicle traffic crash is defined as: "Any crash involving a motor vehicle in transport that occurs or originates on a traffic way, results in injury to or death of any person, or damage to the property of any one perso to the apparent extent of $1,000".1) The data contains 15 variables 2) including crash date, crash time, severity of injury, crash location of all the 45814 car crash vehicles. The weather data came from the National Climatic Data Center 3) in the U.S. which contains daily weather information such as precipitation, snow depth, fog, etc. during the same two-year period(2018-2019). These weather data were collected from several weather stations located in Austin, and it is daily-based-summarized weather data.

We analyzed both the car crash data and the merged data of the car crash and weather respectively. First, we looked at which factors are related to the amount of daily car crashes, daily severe car crashes, percentage of severe car crashes and whether each accident is injury-inducing or not(binary outcome):they were set up as the outcomes of our analysis. We created new variables such as minute of a day, month, day of a week, category of roads and severity of crashes through feature engineering to fit and run regression models more efficiently and accurately. In Particular, we set up an outcome variable, "injury" , which represents the level of severity of each crash by accumulating other variables. If one crash causes fatal or any kind of injury, the injury variable has "1", while if it is "0", it means there is no injury in the car crash.

Next, we checked the relation between these three outcomes and weather features(precipitation, snow, fog) with the merged data of car crashes and weather.4) We merged these two data by the "DATE" variable. The number of days of rain, snow and fog were very limited, so we only used data which had specific conditions such as rain, not all data in some analysis.

After looking across every candidate x features' relations to outcome variables, we fit the prediction model for these three outcome variables, applying linear/stepwise, Random Forest models and logit model. We applied the former models in order to predict daily total crashes and the percentage of severe car crashes, and the last, logit model to predict level of injury from car crashes.

D. Results

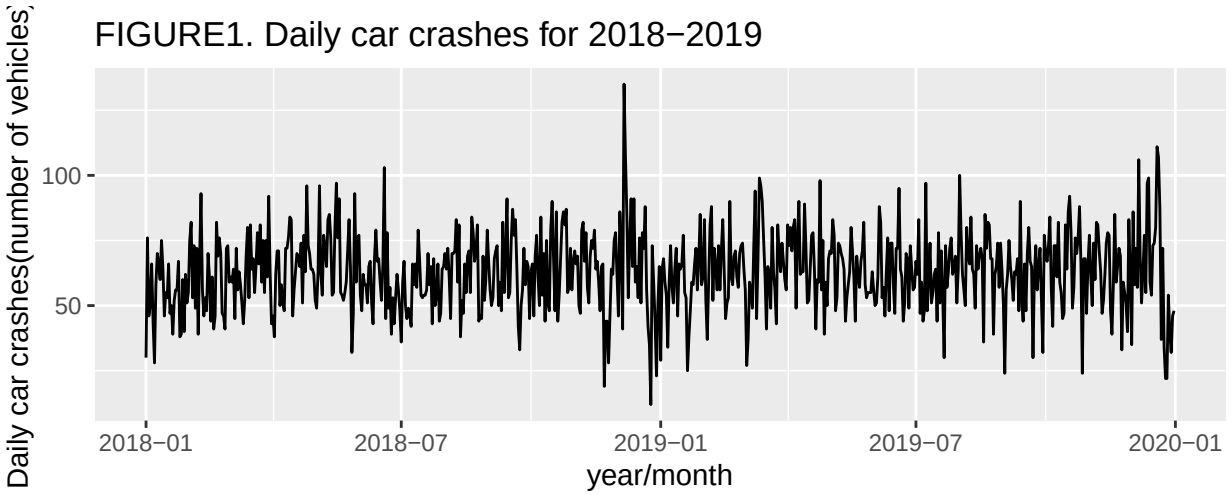
1. Characteristics of car crashes in Austin

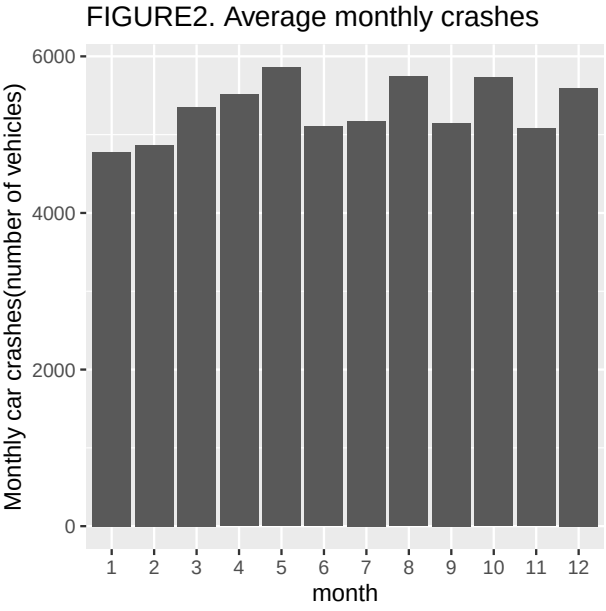
1-1. Car accidents by time

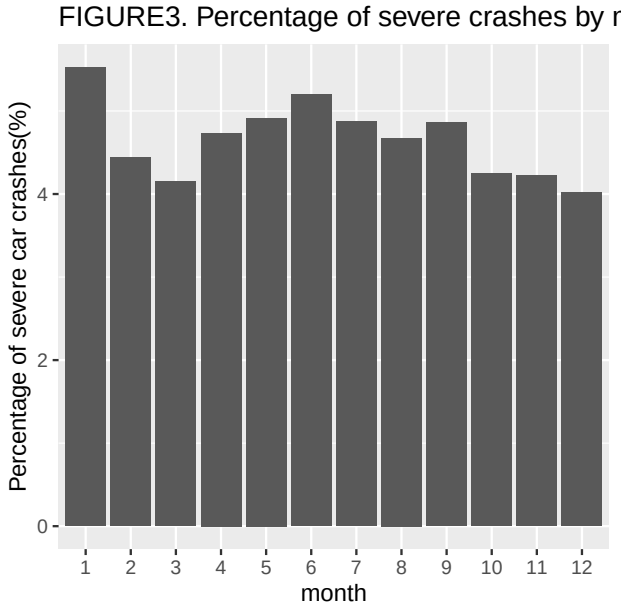
First of all, we worked through the car crashes data alone. It contains detailed information such as date, time, location and severity of every car crash, and among those variables we chose date, time, roads as x variables, and analyzed effects on y variables.

(1) Yearly and monthly trend

To begin with, we scanned the changes of car crash trends over years and months. The number of daily car crashes in Austin shows a slightly upward trend over two years in 2018 and 2019. We can find peaks in the number of daily total crashes in December from the line graph, but there is no stark difference between December and other months in car crashes by month, comparing monthly total number of crashes and severe crashes.







(2) Weekly and daily changes

We also looked into more microscopic changes, by week and day. The graph below says that Weekdays have more car crashes than weekends on average, and we can find that Friday has the highest level both in total crashes and severe crashes.

FIGURE4. Average weekly total crashes

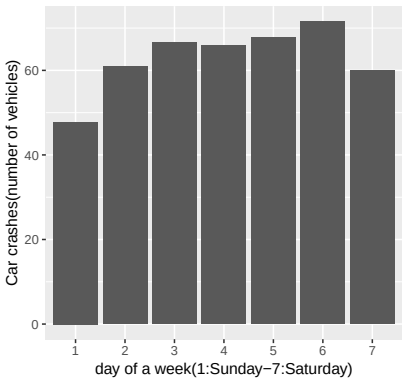


FIGURE5. Average weekly severe crashes

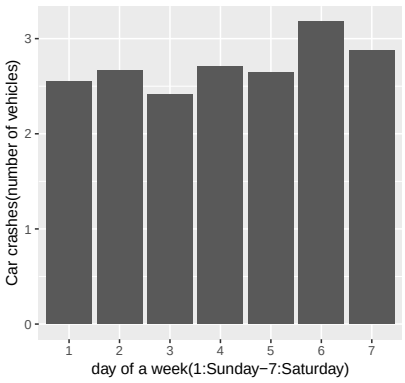
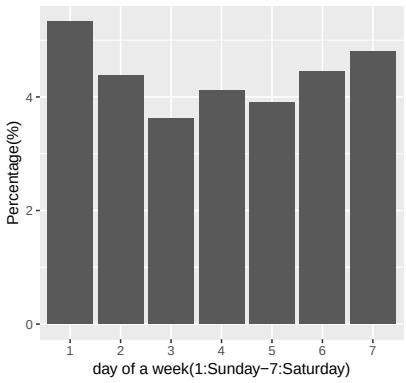


FIGURE6. Percentage of severe crashes



When we see changes in a day, the results show accidents are most common between 3pm and 7pm, showing its peak around 5pm(1000 minutes of day) which is the time for leaving for home from work or school. We can capture these results from the graph and table below. What is interesting is that severe car crashes have different patterns. It does not have its peak around 5pm. Instead, late night(12:00) or early morning(7:00) show higher levels of cases than other time zones of a day.

```
## # A tibble: 6 x 2
##   Time    count
##   <chr>   <int>
## 1 5:00 PM   144
## 2 7:00 PM   116
## 3 3:00 PM   112
## 4 4:00 PM   111
## 5 5:30 PM   108
## 6 6:00 PM   103
```

FIGURE7. Total crashes by every minute of

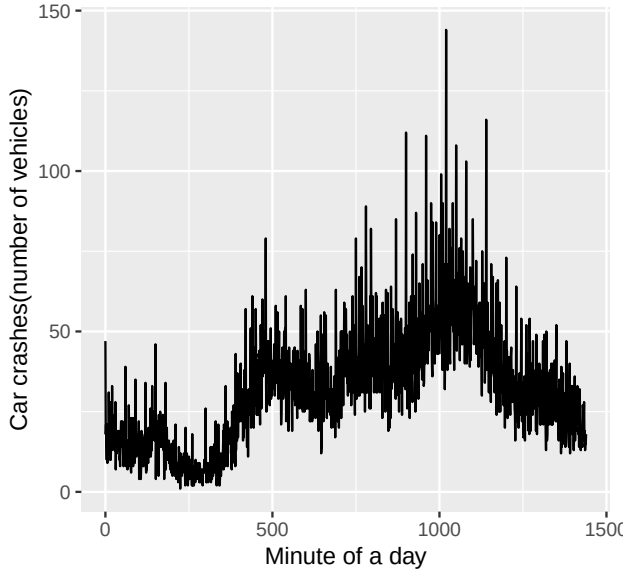
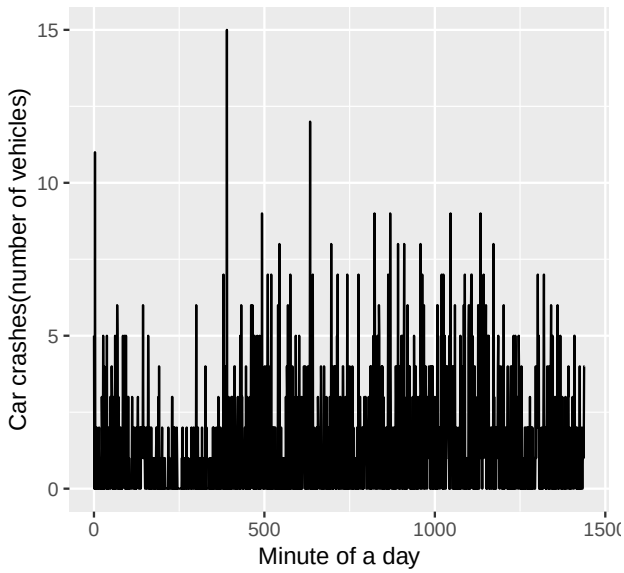


FIGURE8. Severe crashes by every minute o



1-2. Car crashes by category of roads

Austin has three main highways(#35, #183, #290), so we divided all the roads into two categories, "highway" and "else". The results show that the percentage of severe accidents is much higher on highways than out-of-highway roads, while out-of-highway roads have a higher number of total crashes and severe crashes.

FIGURE9. Total crashes by road categorie

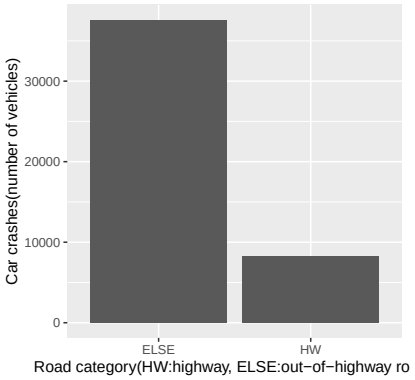


FIGURE10. Severe crashes by road categ

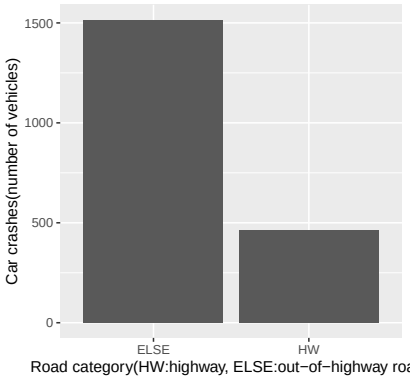
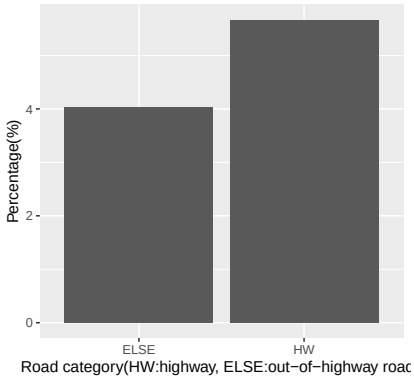


FIGURE11. Percentage of severe crashes



2. Car crashes and weather

We looked through the relation between weather and car accidents(16 cases), and specific variables in each data is as below.

a) weather: PRCP(precipitation), day of snow or not, day of rain or not, day of fog or not

b) car accidents: daily total crashes, daily fatal crashes, daily serious injury, daily percentage of severe car crashes

From the result of analyzing the 16(4X4) cases, we could find a slight correlation between the amount of PRCP(precipitation) and daily total amount of accidents, except for the case of little rain(PRCP<0.5). So, we did analyze, narrowing down to more specifically limited data, which had bad weather conditions.

FIGURE12. Daily crashes by Precipitation

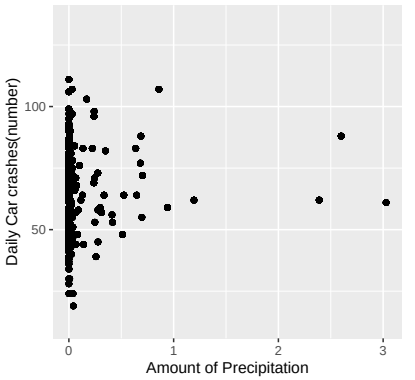


FIGURE13. Daily serious injury by Precipitat

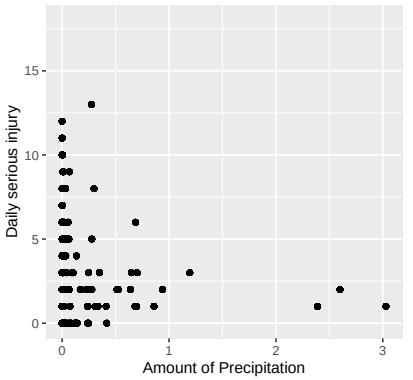
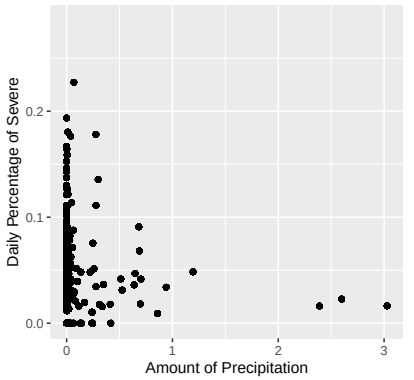


FIGURE14. Daily Percentage of Severe



2-1. Analysis on rainy days

The results of analyzing only on rainy days show there exists a slight correlation between the amount of PRCP(precipitation) and daily total amount of accidents, except for the case of

little rain(PRCP<0.5).

FIGURE15. Total Crashes on rainy days

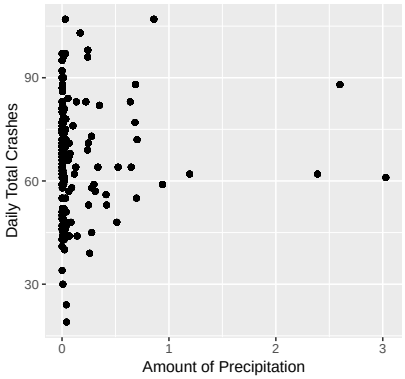


FIGURE16. Serious injury on rainy days

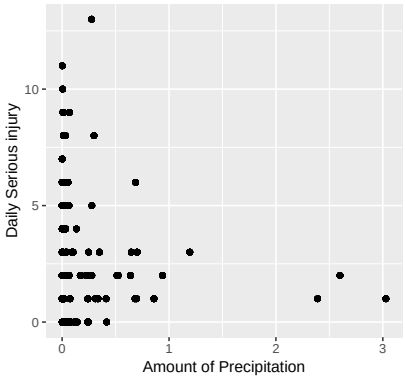
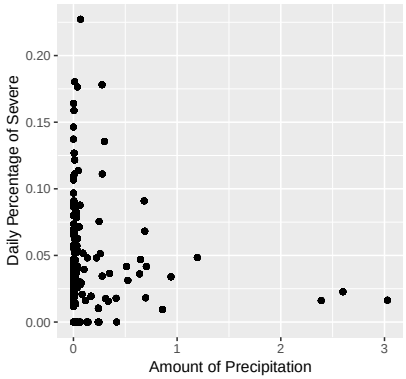


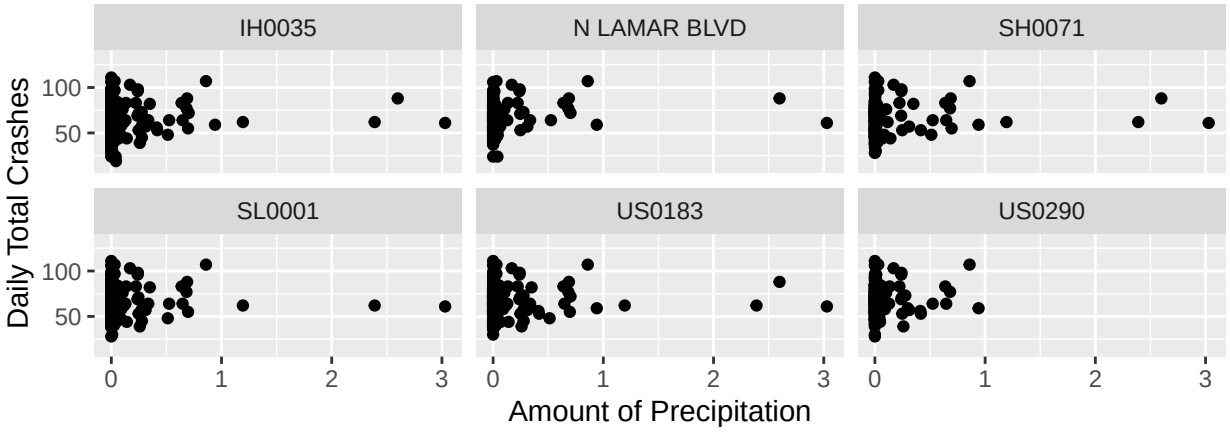
FIGURE17. Percentage of Severe on rainy



2-2. Analysis on rain and roads with high amount of accidents

Next, we analyzed more narrowly the roads which have relatively higher amounts of car crashes. Our targets were set as IH0035, US0183, IH0035, SL0001, N LAMAR BLVD, US0290 and SH0071. In addition, we limited our analysis to days with rain whose precipitation is equals or higher than 0.5. We can find there is a slight correlation between the amount of PRCP(precipitation) and daily total number of accidents.

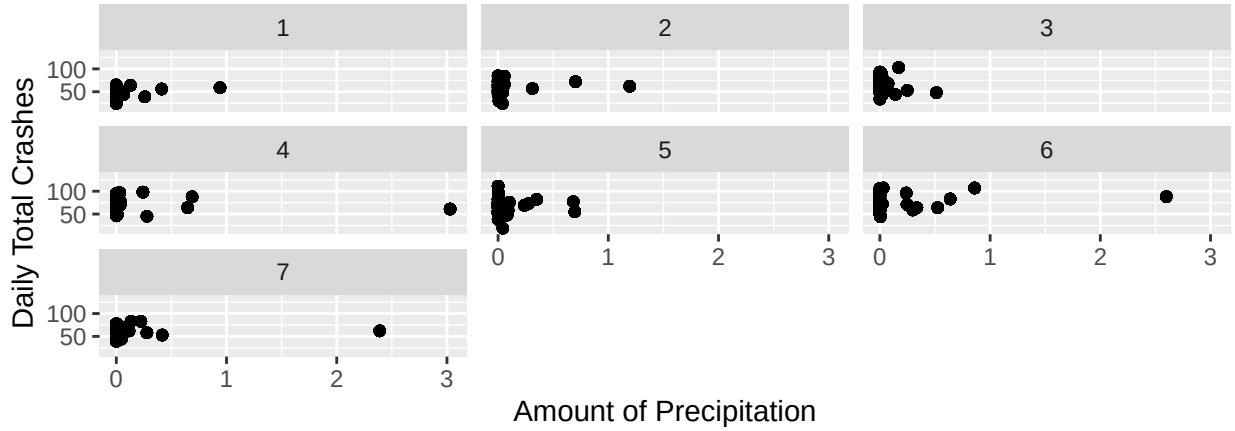
FIGURE18. Daily Total Crashes on the roads with the most accidents



2-3. Analysis on rain and day of a week

As a result of analyzing interaction between precipitation and day of week, On Friday, the correlation between PRCP(precipitation) and the amount of accidents is higher than others, except for the case of little rain.

FIGURE19. Daily Total Crashes by day of the week(2018−2019)



2-4. Analysis on rain and roads' category

The graph below shows that there is no difference between Highway and Else(from 16 cases), but on each road, it can be said that a slight correlation exists between precipitation and the amount of accidents, except for the case of low rainfall.

FIGURE20. Total Crashes by the Road cate

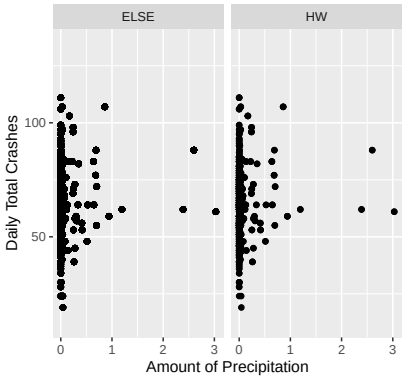


FIGURE21. Serious injury by Road category

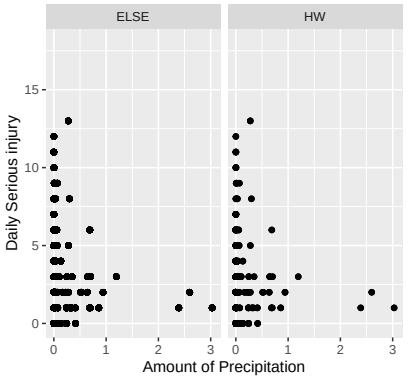
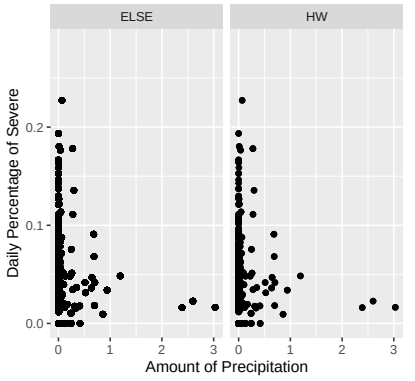


FIGURE22. Percentage of Severe by Road



3. Prediction model

We predicted three outcomes: daily total crashes, percentage of severe car crashes, whether injury or not. Linear, stepwise, Random Forest models were applied to predict the former two outcomes, while a logit model was adapted for prediction of the last outcome.

3-1. Prediction of Daily total crashes

First, in order to predict daily total crashes we used 'day of week', 'road category', 'month' and 'precipitation(PRCP)' variables as x variables. We fitted linear/stepwise, Random Forest models, and we got the out-of-sample RMSE from these models. The RMSE of Random Forest model is lowest of these models. So, we concluded that it has the highest prediction accuracy.

RMSE of Linear

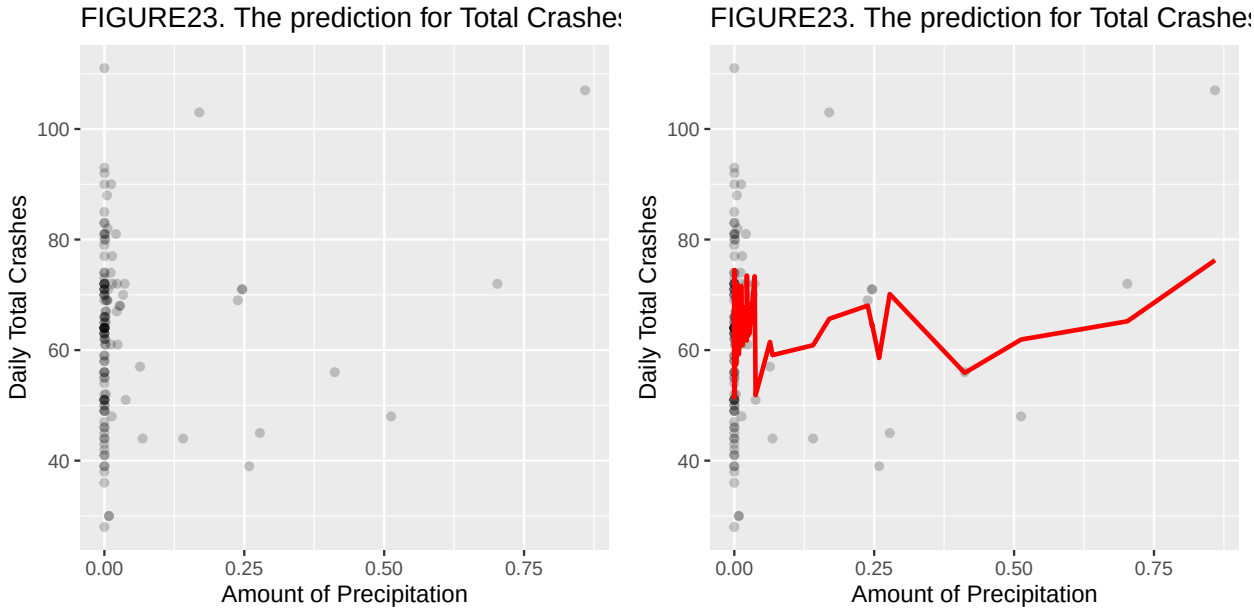
```
## [1] 11.95189
```

RMSE of Stepwise

```
## [1] 11.98067
```

RMSE of Random Forest

```
## [1] 12.34524
```



3-2. Prediction of daily percentage of severe car crash

Secondly, we predicted the 'daily percentage of severe car crash injuries' by 'day of week', 'road category', 'month', 'precipitation(PRCP)' variables. Comparison of the results of linear/stepwise, Random Forest prediction models show us the out-of-sample RMSE. The RMSE of Random Forest model is lowest of these models. So, we concluded that it has the highest prediction accuracy.

RMSE of Linear

```
## [1] 0.04288775
```

RMSE of Stepwise

```
## [1] 0.04346
```

RMSE of Random Forest

```
## [1] 0.04168013
```

FIGURE24. The prediction for Percentage S

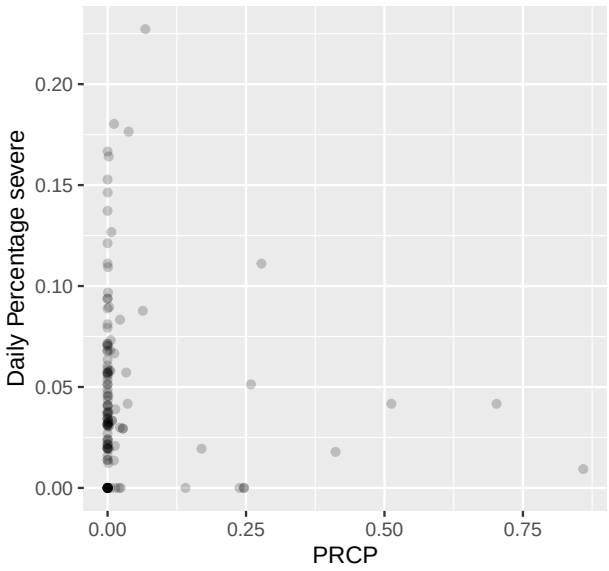
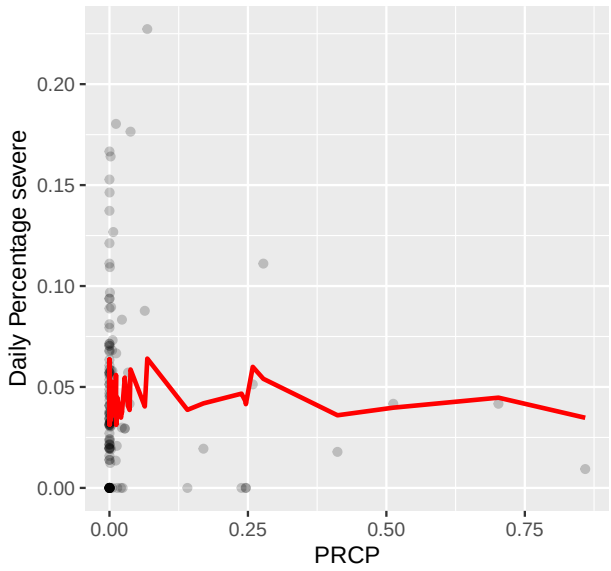


FIGURE24. The prediction for Percentage S



3-3. Prediction of injury by crashes: logit model

We fitted the prediction model to verify if each car crash will have injury or not. We applied for a logit model because the outcome variable is binary. Using the previous research results, we input possibly related x variables in the model. We fit the first logit model with x variables from the merged data, while the second model has only x variables from car crash data. The fitted two models are as below.

logit_injury1 model: injury ~ day of week, road category, month, snowy or not, precipitation

logit_injury2 model: injury ~ road category minute of day, day of week, month

We can see that the cases which have injury in the test set have higher probability to be classified into "injury" from the two logit prediction models. However, the prediction model's accuracy rates represented 55~56%, which means there might be a lot of unpredictable intrinsic factors in car crashes, which are said to be individualized crash circumstances and can affect the severity of crashes.

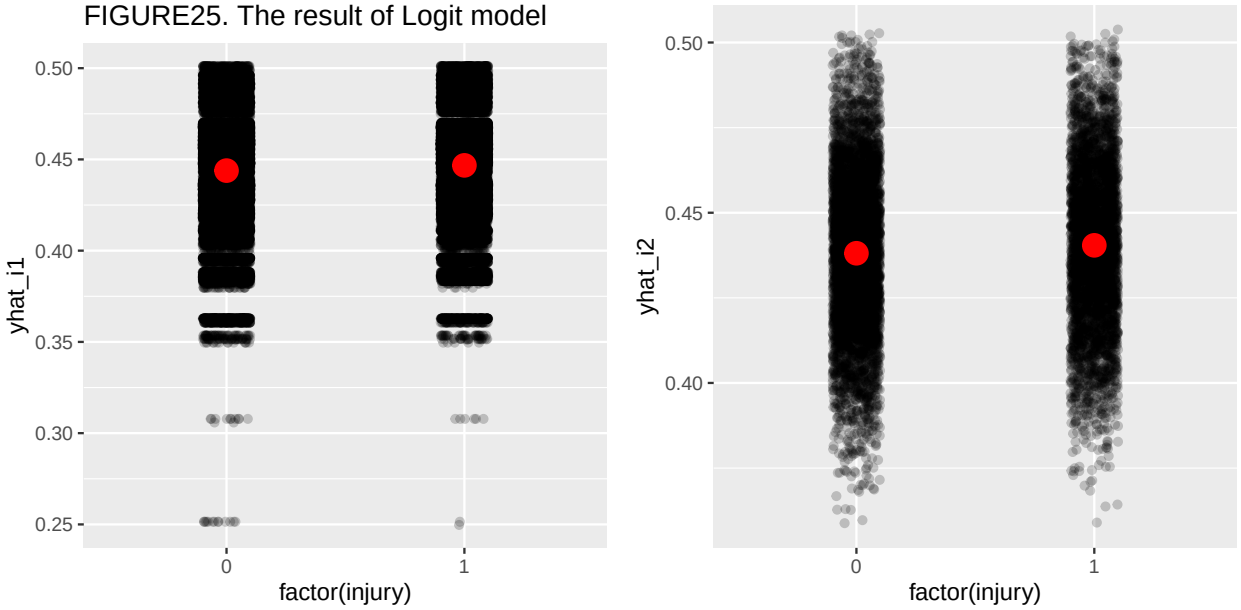
```
##    yhat
## y      0     1
##   0 66817   136
##   1 53575   199
```

```
## [1] 0.5551037
```

```
##    yhat
## y     0    1
##   0 5081   11
##   1 4066    5
```

```
## [1] 0.5550584
```

FIGURE25. The result of Logit model



E. Conclusion

We analyzed the relation between car crashes and external features such as weather, time, roads category, etc. using car crash and weather data for 2018-2019 in Austin. We could find the amount of car crashes tend to be higher in late afternoon on weekdays(especially Friday), under rain, and crashes are likely to be more severe late at night or early in the morning, on highways rather than out-of-highway roads. It is not certain that a little amount of rain does affect the occurrence of car crashes, but if the amount of rainfalls goes up, car crashes are more related to precipitation. We could predict the number of daily car crashes and the percentage of severe car accidents which induce fatal or serious injuries using the linear/stepwise, Random Forest models. We could also predict if each crash will have injury or not with the logit model. These prediction results can be used for allocation of traffic control forces under specific time or weather conditions. They would be meaningful in terms of preparing local emergency hospitals for car crash injuries under certain conditions.

One thing we need to discuss is that prediction of a single case was relatively more difficult to have high prediction accuracy than that of daily average outcomes. It can be seen that the individuality that cannot be correctly predicted is largely attributable to traffic accidents. I expect if more location- and - time specified weather data 5) or each driver's characteristic data is available, the prediction models' accuracy can be improved, and it can be said to be the next research subject.

F. Appendix

Fit the model for prediction of each crash's severity

We fit the model to predict if every single car crash is severe or not with logit model. We found that severe cases in test set has higher yhat value, which means higher probability of prediction as "severe crashes" through the plot. However, the confusion matrix spitted out all prediction value "0". We guessed it might be caused by extremely rare frequency of severe

crashes. So, we created another variable, "injury" which had more cases than severe crashes, and analyzed about this new output variable.

```
##    yhat
## y       0
##   0 116586
##   1   4141
```

Footnotes and references

1) http:///www.txdot.gov

2) Variables in the car crash data: Crash, Crash Date, Crash Time "Crash Latitude, Crash Longitude, Reported Road,Reported Intersecting Road,Vehicle Direction of Travel, Total Crashes, Total Vehicles - Fatal Crashes, Total Vehicles - Suspected Serious Injury Crashes, Total Vehicles - Non-Incapacitating Injury Crashes, Total Vehicles - Possible Injury Crashes,Total Vehicles - Non-Injury Crashes, Total Vehicles - Unknown Injury Crashes.

3) http://www.ncdc.noaa.gov

4) The weather data has 29 variables, but the majority of them are only collected at very limited numbers of weather stations, so we could not include them into the prediction model. The precipitation data were used by the form of the average precipitation of the weather stations.

5) The weather data we could get from NOAA was the only daily summarized information, so it was hard to know the exact weather condition when each car crash occurred. Moreover, it was hard to match all the weather stations' locations and all the car crash points as well.