Title: Seattle Traffic Accidents

Group Members: Scott Kinder, Siying Li, Wei-shao Chen

**Abstract:**

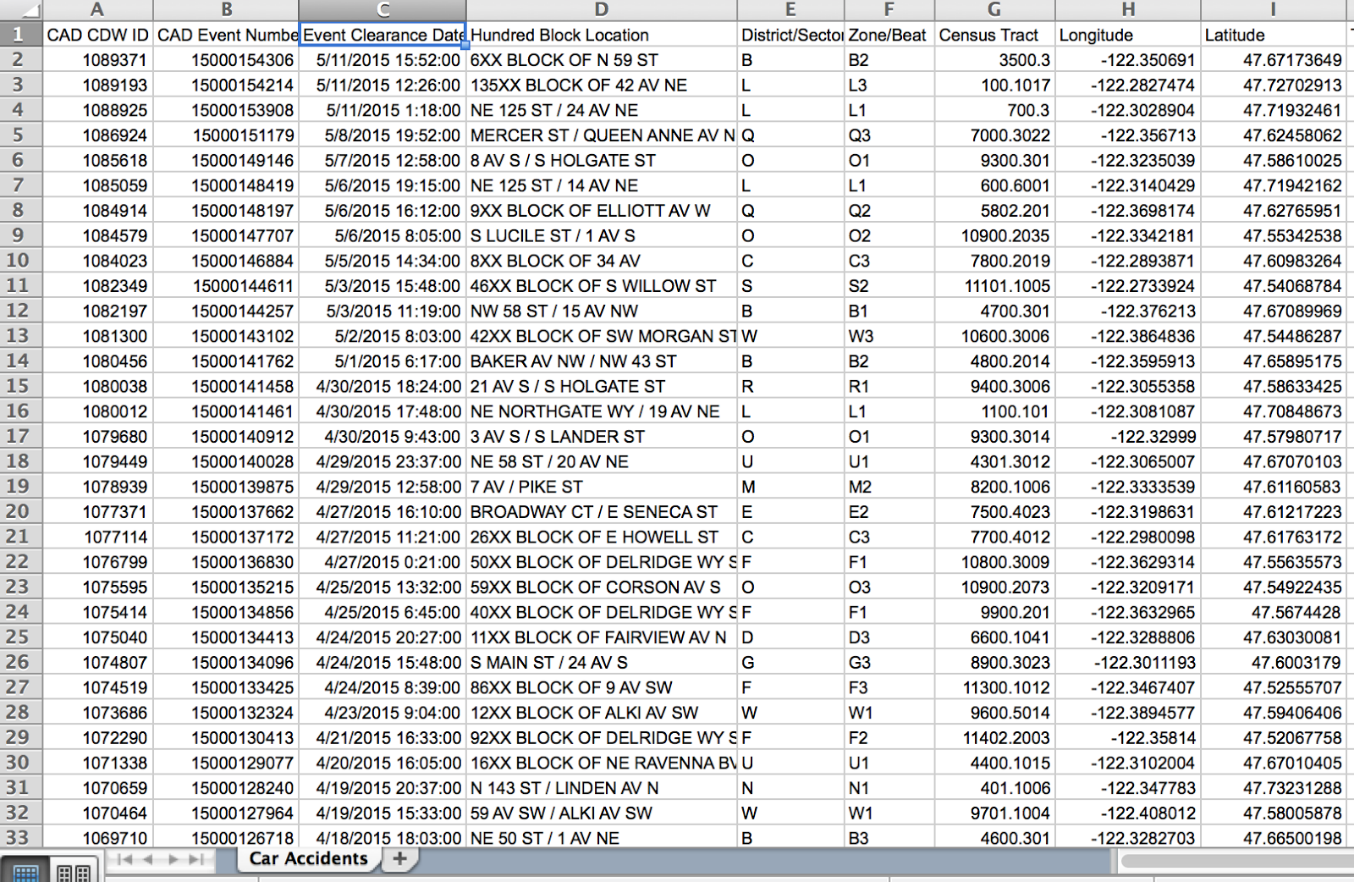
In effort to help find a solution to the 1.3 million deaths from car crashes every year, our group wanted to find out what are all the factors that affect the chances of car crashes, and see if we would be able to predict the chances of car crashes in a given condition. Our findings could help raise people’s awareness of serious reality of car crashes, as well as save people’s lives if they become more aware. By studying datasets of all car accidents in Seattle from 2010 to 2015 in regards to how much rain, visibility and time would affect car accidents, we repeatedly conducted student t-test on all three variables and were able to find out lower-than-alpha-value p-values for all of them. Thus we concluded that all the selected variables are significant and have an impact on the total number of car accidents.

From this research, we were able to predict the chances of car accidents with given time, precipitation and visibility. By standardizing the traffic data and car accidents under certain condition, we generated a formula that could predict the percentage of getting in a car accident under the given conditions. Overall, our prediction model generates probabilities of getting in car accidents from a range of 0.04% to 1.4%. The data, combined with our study, shows that the highest total number of accidents is between 11-12 P.M, while the highest chance of getting in a car accident is at 3 A.M.

**Datasets used:**

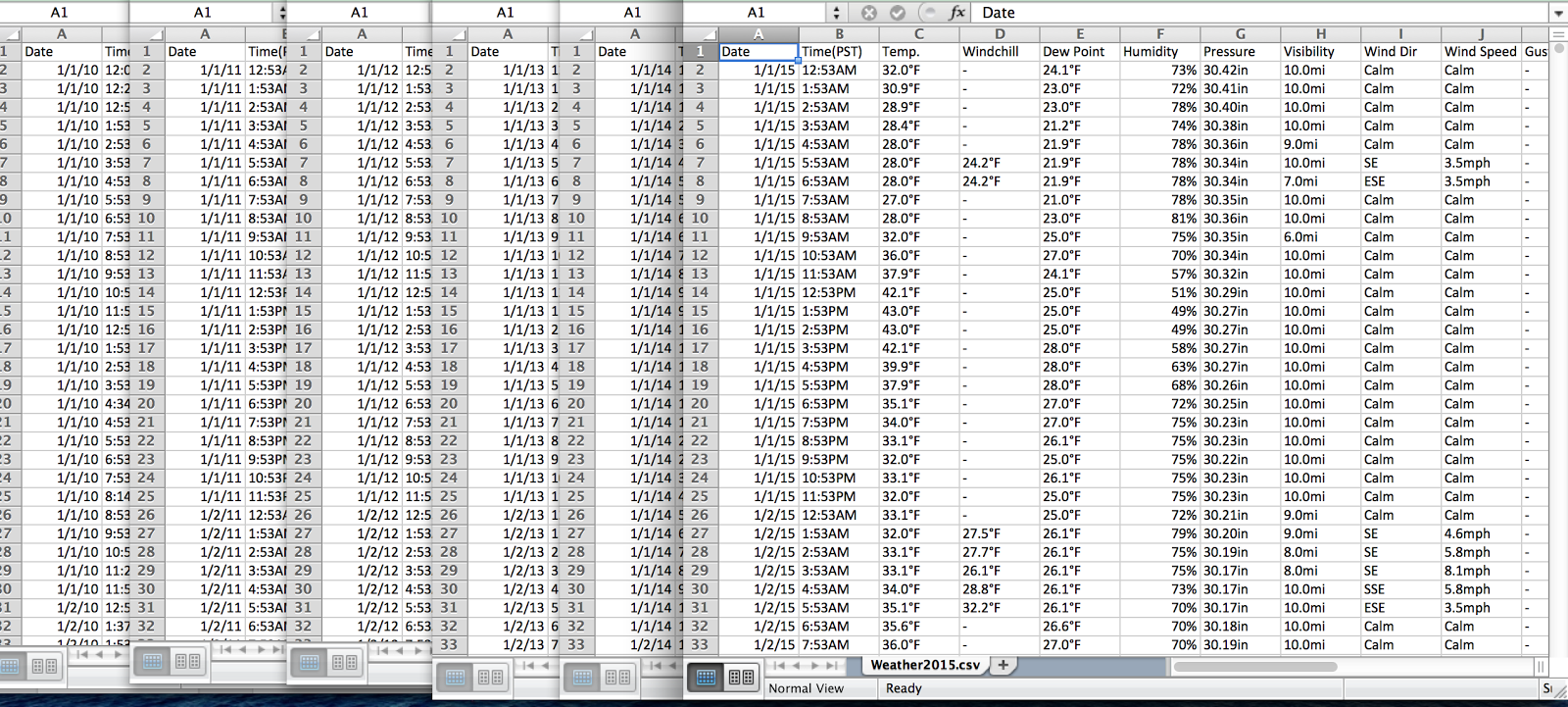
Since our group’s objective was to study how visibility, rain, highway/local roads and time affect the probability of car crashes, we collected four main datasets which included comprehensive and accurate information on all of the four factors, as well as information on car crashes.

We found our core dataset from data.seattle.gov, which is a highly reliable data source, which has vast amounts of useful information on all the car crashes from June 2010 to May of 2015 based on Seattle Police Department 911 Incident Response. There is information on the CAD Event Number, event clearance, hundred block location, longitude, latitude and others, for all of the 54006 car crashes in the dataset. In order to have a clear visualization of how car crashes grew in different areas in Seattle, we converted our main datasets to JSON files, (which could be found here: <http://students.washington.edu/kinders/i370/csvtojson.json>), and generated a heat map based on the JSON files (link to the heat map: <http://students.washington.edu/kinders/i370/index.html>). Below is an image of our core dataset.



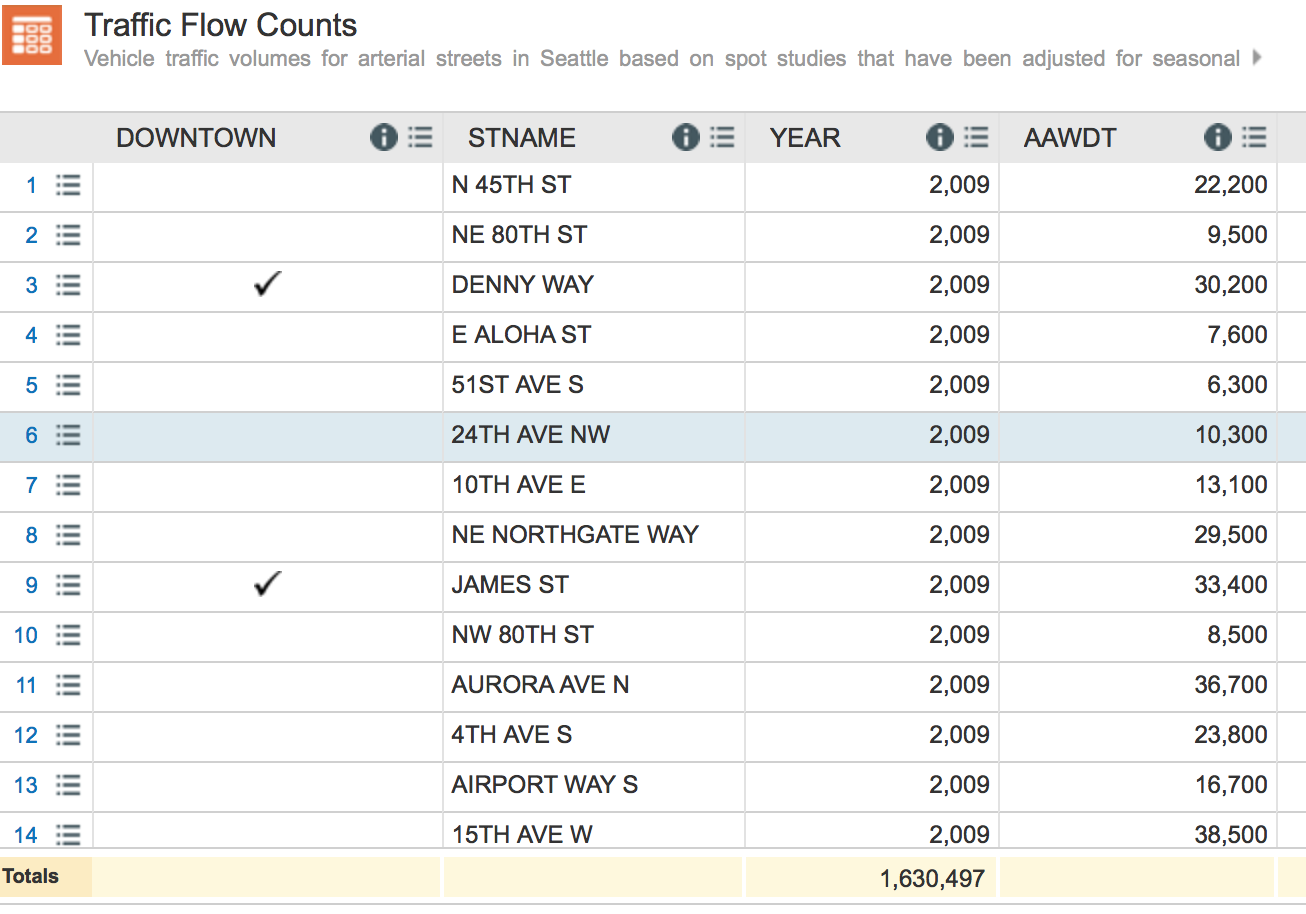
However, we had to work around with two problems in order to format the datasets to meet our research questions. The first was that the dataset only gave us the address of each accident *without* a clear expression of whether it is on highway or not. We processed each address of each car crash seeking an indication of highway name and labeled those with 1; the rest were labeled with 0 indicating the accident happened on a local road. The other main problem we had with this dataset was that it only provided the event clearance time when the incident was considered safe to close out, even though what we needed was the actual event occurrence time. Fortunately, on city-data.com, we were able to find all the fatal accidents information from 1994 to 2013, which included event occurrence time and location for all 1,492 accidents. We managed to manually match all of those accidents with the corresponding ones from core dataset and compared their time difference. By subtracting the fatal accidents’ event occurrence time from the corresponding car accidents’ event clearance time, we were able to calculate the mean of that difference and estimate the event occurrence time of all the car accidents from our core data.

To get the weather data that we needed for the study, we crawled all the hourly weather information from the Weather Underground website from 2010 to 2015, which provided us precise visibility distance and precipitation every hour. It also included various other information, but due to scope as well as relevance, we only choose those two variables to look at closely. Below is an image of the dataset.

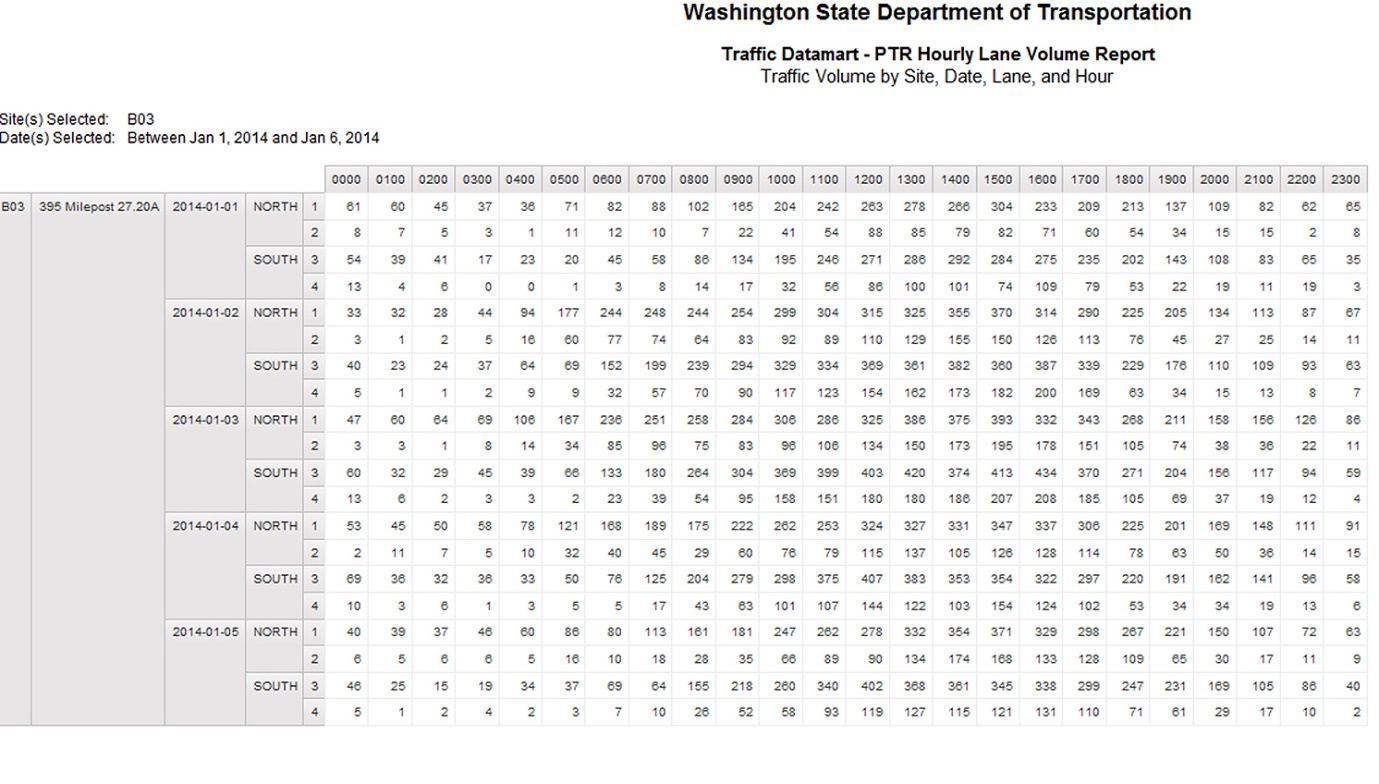


One problem we ran into with this dataset was that the time format did not match those in our core data set, especially when we had to round times to the hour to represent a categorical variable. What we did to solve this was also round the times on the weather dataset the same way we rounded the times on the accident dataset. Besides changing the format of the time, we also decided to use visibility 0-5 miles as one categorical variable, and the rest of 6-10 miles each represents an individual categorical variable. This wasn’t extremely desirable to do, since we hypothesized that very low visibility might be interesting to look at, however due to the extremely small amount of car crashes that happened when visibility distances are between 0 to 5 miles, we were forced to go this route.

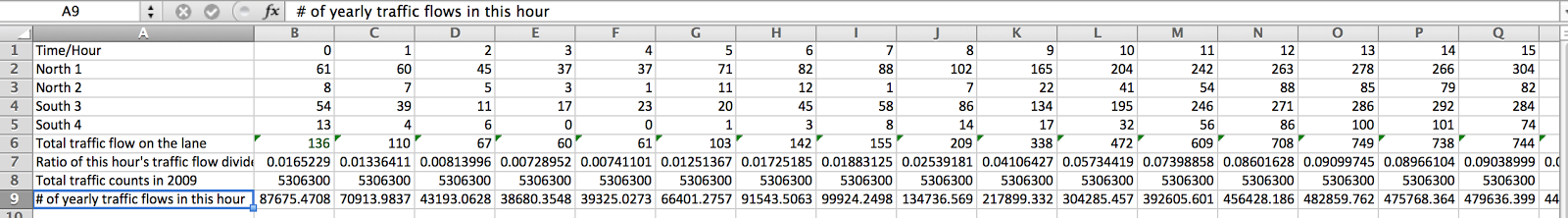
For estimating the percent chance of getting in an accident at a given hour, we utilized two relatively related datasets and standardized them. The first was a dataset of average weekday daily total on all arterial streets in Seattle from data.seattle.gov.



The second was a dataset of hourly traffic flows on selected streets from Washington State Department of Transportation.



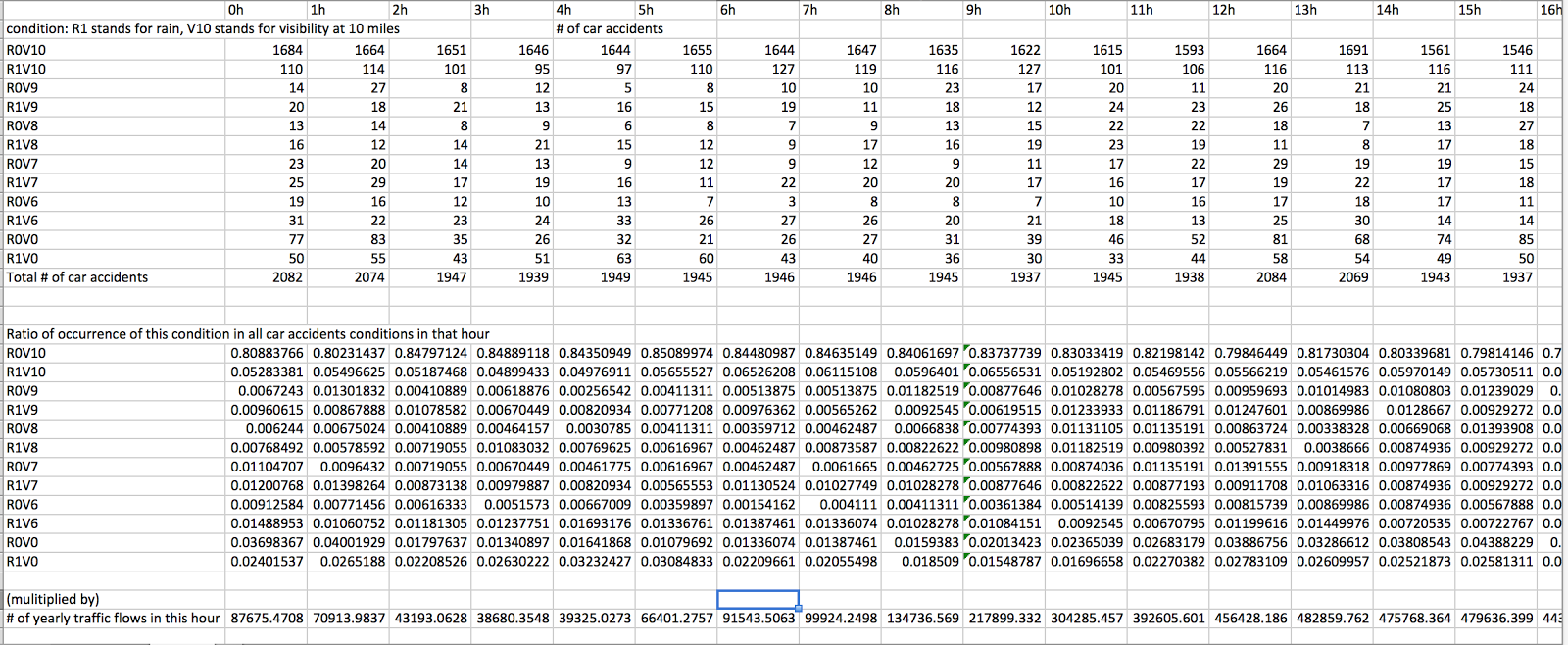
We connected these two datasets by first getting a ratio of how much of the traffic flow for each hour in a day in relation to the total amount of traffic flow of the day, and then multiplied this by number of accidents at that hour for the given condition. Below is the data generated by the first step of calculation and estimation.

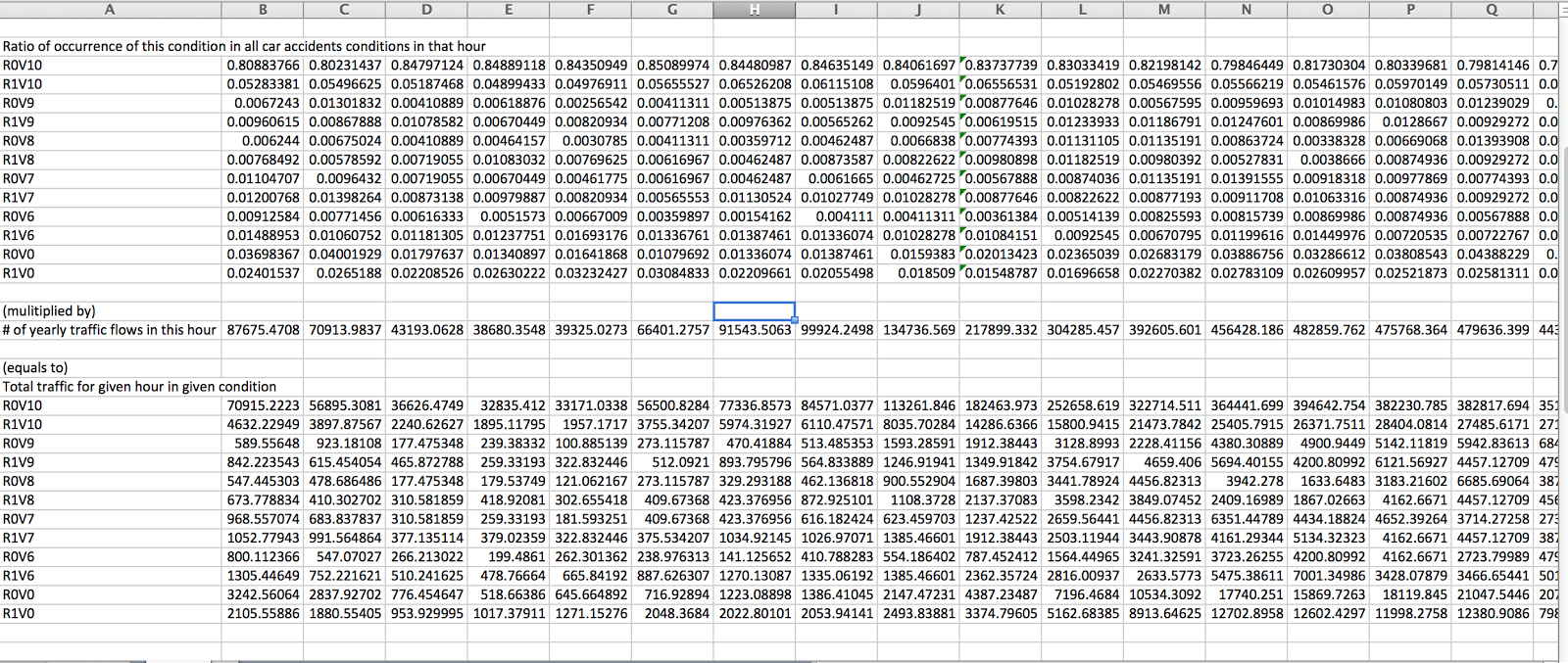


For example, we would find what ratio of cars were on that road during the given hours:

Ex: Ratio of Cars on the road at 3 A.M = 60 (traffic flow at 3 A.M) / 8231 (total traffic flow of the entire day) = 0.00728952

Below are images how we estimated the total traffic for given hour in every condition.





Then we multiplied the total number of cars at that hour by the ratio of that hour.

Ex: total traffic for 3 A.M with no Rain and visibility at 10 miles = 0.849118 (ratio of R0V10 of all conditions) \* 38680.3548 (# of yearly traffic flows at 3 A.M) = 32835.412

You can see how problems might arise here from two main things. First, the ratio of cars on the road per hour might be inaccurate for all roads, see we generalize them as being all the same, which is probably false.  Second, the total number of cars on the road might also be inaccurate, since it was labeled the “average weekday daily total”. We made a bold assumption that this is for a whole 24 span. Also we made an assumption it is equal every day of the week, including weekends. We only made these assumptions because there were no other possible workarounds.

**Primary methods implemented:**

After deciding that the experiments should include hours, visibility, rain and highway location, then the next step would be checking whether or not these variables truly has an effect on the car accidents.

To set up the test, the variables must be identified in terms of categorical or numerical. The relationship that the statistical analysis is trying to show is whether or not the factor has an impact on the car accidents. In other words, it can also be tested with the total number of car accidents over a period of time.

As identified in the previous section, hours, visibility, rain and highway location are all in categorical forms while the number of total car accidents is a numerical variable.

The suitable analytic test for a categorical and a numerical variable is a student t-test. By running the t-test for every individual variable and setting the confidence interval to be 95%, the results show that every variables have a p value less than 0.05, which proves the null hypothesis wrong and accept the alternative hypothesis: the variable has an impact on total number of car accidents (refer in results section figure 1.1)

By repeating the same process, setting three other categorical variables (visibility, rain, and highway location) as the x variable and keeping the total number of car accidents as the y variable that correspond to the particular chosen independent variable, the p values generated by the student t-test are all below alpha level. In other words, the three other categorical variables are significant in terms of influencing the dependent variable.

After confirming the selected variables are all significant and have an impact on the total number of car accidents, it is sufficient to proceed to the next step of analysis: finding the chances of having a car under certain condition. To do this, the traffic data for every single day and every single hour is needed. However, the data available on Internet only includes the total number of cars on the roads of non-highways of a week. In order to make the data usable, it is necessarily to standardize the data, which is already explained in the datasets used section.

By standardizing the traffic data, the total number of car accidents must also be standardized in terms of many times a specific variable appears compared to the other. In other words, because there are significantly more hours where it doesn’t rained compared to hours where it rained, it is crucial to standardize by a ratio of total number of hours that rained and total number of hours that didn’t rained. This way, both the total number of car accidents and the total number of traffic on a given hour, visibility, rain condition, and highway location, can be used to calculate the percentage.

**Results:**

Figure 1.1: Student t-test on hours and total number of car accidents

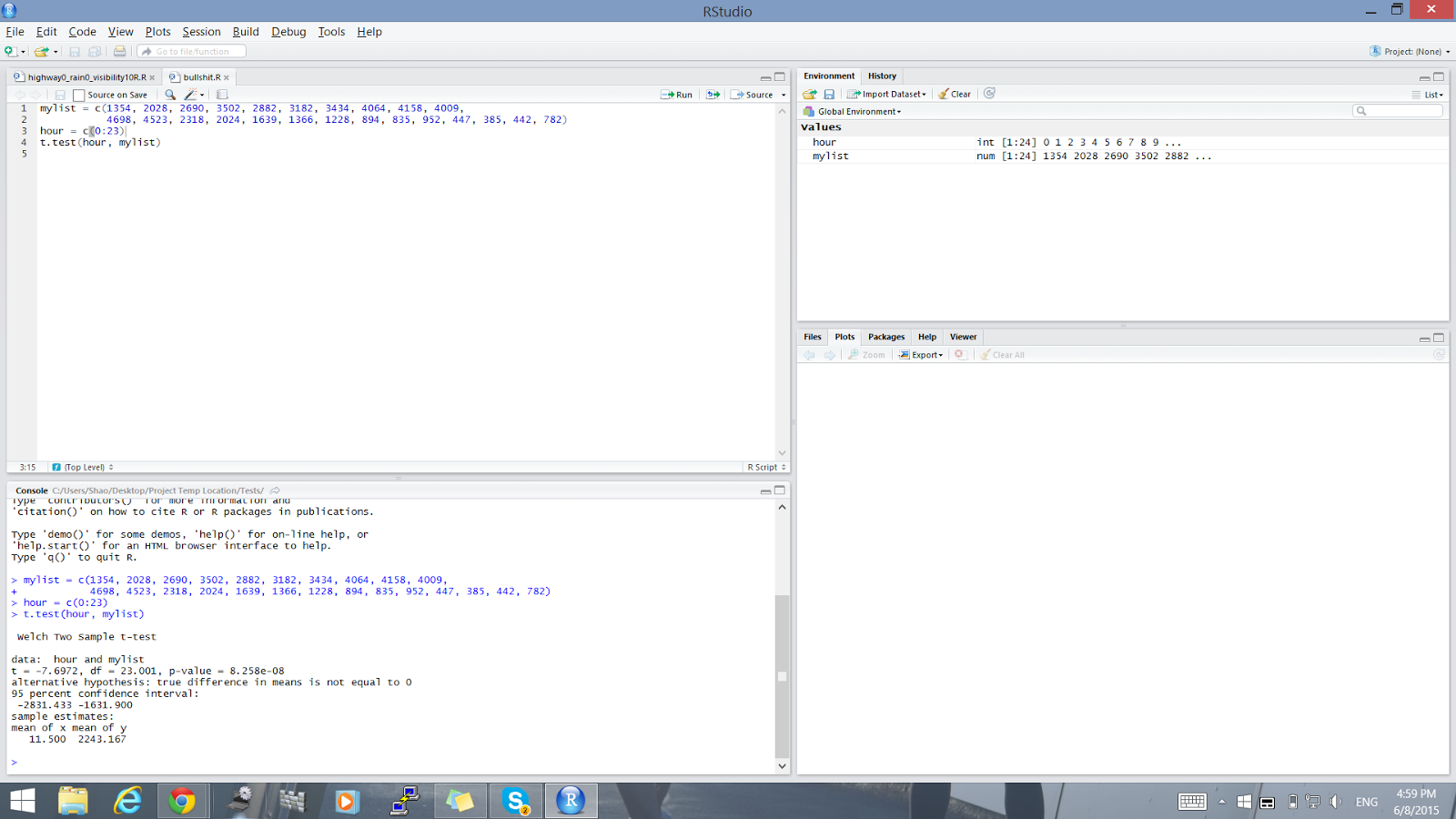
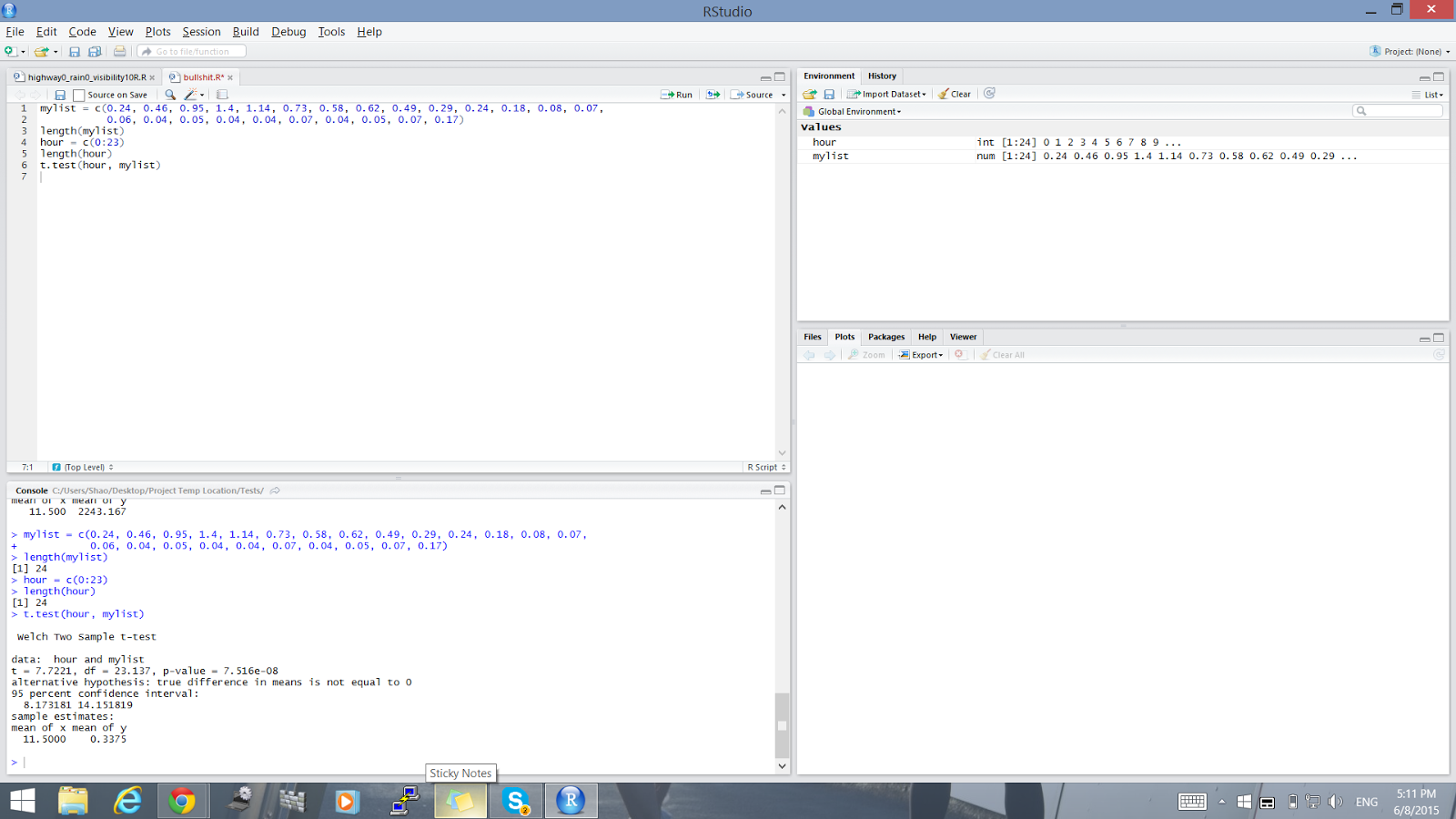


Figure 1.2: Student t-test on hours and the chances of car accidents under certain condition



In a similar fashion, a two sample student t-test should be conducted to test whether or not hour and the chances of car accidents are significant under 95% confidence. In figure 1.2, the p value is 7.516e-08 which is way below the alpha level, 0.05. This concludes the fact that the results are significant at 95% confidence. Furthermore, because the p value is less than 0.05, it also rejects the null hypothesis: the hour doesn’t have impacts on the chances of getting into a car accident. Because the null hypothesis was rejected, the alternative hypothesis is said to be accepted, which states that the given hour of the day will influence the chances of getting into a car accident.

Not only that, in order to present the final results in a fashionable and an understable way, a model is generated. The graph for calculating the chances of car accidents can be found in this following link: <http://students.washington.edu/kinders/i370/graphs.html> . It asks the user to choose some variables/conditions before outputting a graph. The graph that is being generated has an x-axis of hour and an y-axis as total number of car accidents. Then in order to calculate the actual percentage under a given hour, the text box below is where the hours should be inputted. With previously calculated standardized total car accidents and standardized traffic, a simple formula of accident/traffic will calculate the probability of encountering an unfortunate accident. Overall, the calculated chances of car accidents ranges from 0.04% to 1.4%, which can be outputted by entering different hours.

**Discussion:**

As it is clear from our results, we have come away with some very interesting takeaways. We determined that not only is the total number car accidents at a given hour statistically significant, but also that the percent chance of getting in a car accident at a given hour is also statistically significant. This was really huge for us, because going in we weren’t positive that this was the case. Not that our research would have been any less significant if we had not rejected our null hypothesis, but it is satisfying to find results that could be reported back for a good cause.

To back up to the beginning of the project, our initial assumptions were that car accidents were slightly random, due to our human observations we make every day. Sometimes, while walking down the road, a car accident might be seen and it is hard to determine why; it seems almost random. Of course the problem with that is the fact that a single human cannot possibly witness all of the accidents, the only way this is possibly is by having all the data. Yet, we knew that probably driving late at night would be risky, due to the increase in drunk drivers, and poorer conditions on the road.

What we found, and what our results prove, is that actually driving at different times does increase both your chances of getting in a car accident, as well as the total number of car accidents. The ladder is not extremely interesting, seeing as it is probably likely that the more cars are on the road, the higher number of total accidents would occur. What is more interesting, in our opinion, is that the percentage of getting in a car accident is also highly related to the time you drive at. It also different from the total number of accidents, in the sense that the highest total number of accidents is between 11-12 P.M., and the highest chance of getting in a car accident is at 3 A.M.

Along with our great results, we did encounter some hardships and perhaps slight miscalculations might arise from possible errors in the data set. As mentioned earlier, we were left to infer average traffic on the streets. Not only that, but we speculated that the datasets themselves might have possible errors, as the fatal car accidents did not match up perfectly with the seattle.gov data sets. However, we are fairly confident that it is very close, and therefore extremely relevant. Another hardship that we faced during this project was working with the variables in certain ways, such as grouping what could be numerical to categorical variables. This was essentially necessary, however, due to the datasets we could find.

All that said, would you ever do anything if you knew you had a 1% chance of getting injured? Sure, there are many adrenaline junkies out there who love that thrill, but driving should not be giving you an adrenaline rush! Ironically enough, all of us are in favor of autonomous (self-driving) cars, because it is easy to see the extremely high risk of driving. I mean, not only is it statistically significant what hour you drive at that determines the total number, but it is also statistically significant your percent chance of getting in a car accident as well! It’s scary because it plots humans as data points, puts percentages over their head every time they drive. We know it’s not always the drivers’ fault, sometimes it’s the other driver, or a whole plethora of reasons that we would have dug into if we had the time. We hope that as we progress into the future, more people will research this topic that is one of the leading causes of death in America that certainly can be avoided. We also hope that people will heed our results and be more cautious on the road.

**Conclusions:**

Throughout this project, we have worked extremely hard to put everything together into digestible pieces which people can really take important facts away from. It was both challenging and rewarding to work in a group in environment on a data science project, but definitely worth all the effort.

In general, we have been extremely pleased with our results and believe that it is more than enough to hopefully make an impact on people's’ lives. In the hopes that various groups of people heed our results, which others have also done similarly in the industry level, we hope that these numbers will go down soon. However, as the total number of cars on the road goes up every year (cite traffic data), we fear that the numbers will only keep rising in terms of total accidents.

Going forward in the future, perhaps if we had more time and data, we could make a much bigger report which could utilized by other outside groups. This could include but not be limited to: The State of Washington, groups that advocate for safe driving, and other students and organizations doing their own studies on topics related to this matter. We will keep our two information graphics up for the next year and a half until we graduate and the UW revokes our hosting space, yet our code will be posted on the group repository, and can always be accessed later by contacting a group member.

Thank you for reading, and please drive safe!

**References:**

1. Car accidents | Data.Seattle.Gov | Seattle's Data Site. (n.d.). Retrieved June 9, 2015, from https://data.seattle.gov/Public-Safety/car-accidents/htah-4jvq

2. Fatal car crashes and road traffic accidents in Seattle, Washington. (n.d.). Retrieved June 9, 2015, from http://www.city-data.com/accidents/acc-Seattle-Washington.html

3. Traffic Flow Counts | Data.Seattle.Gov | Seattle's Data Site. (n.d.). Retrieved June 9, 2015, from https://data.seattle.gov/Transportation/Traffic-Flow-Counts/7svg-ds5z

4. Weather History for Seattle, WA | Weather Underground. (n.d.). Retrieved June 9, 2015, from http://www.wunderground.com/history/airport/KSEA/2006/10/4/MonthlyHistory.html?MR=1

5. WSDOT - Traffic Reports by Request. (n.d.). Retrieved June 9, 2015, from http://www.wsdot.wa.gov/mapsdata/travel/traffic\_reports.htm