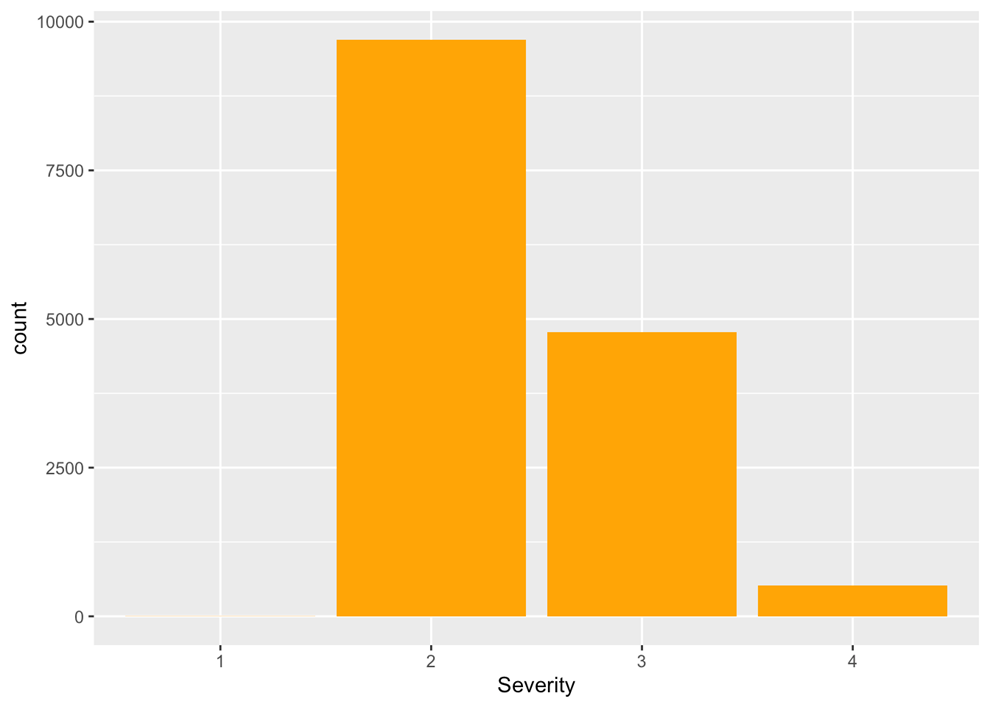
|  |  |
| --- | --- |
|  | 4/29/2020 |

|  |  |
| --- | --- |
| to: | Proffesor Bushong |
| from: | Dylan PatelAver MckayJianqi Liu |
| subject: | SSC442 fINAL Assignment |
|  |  |

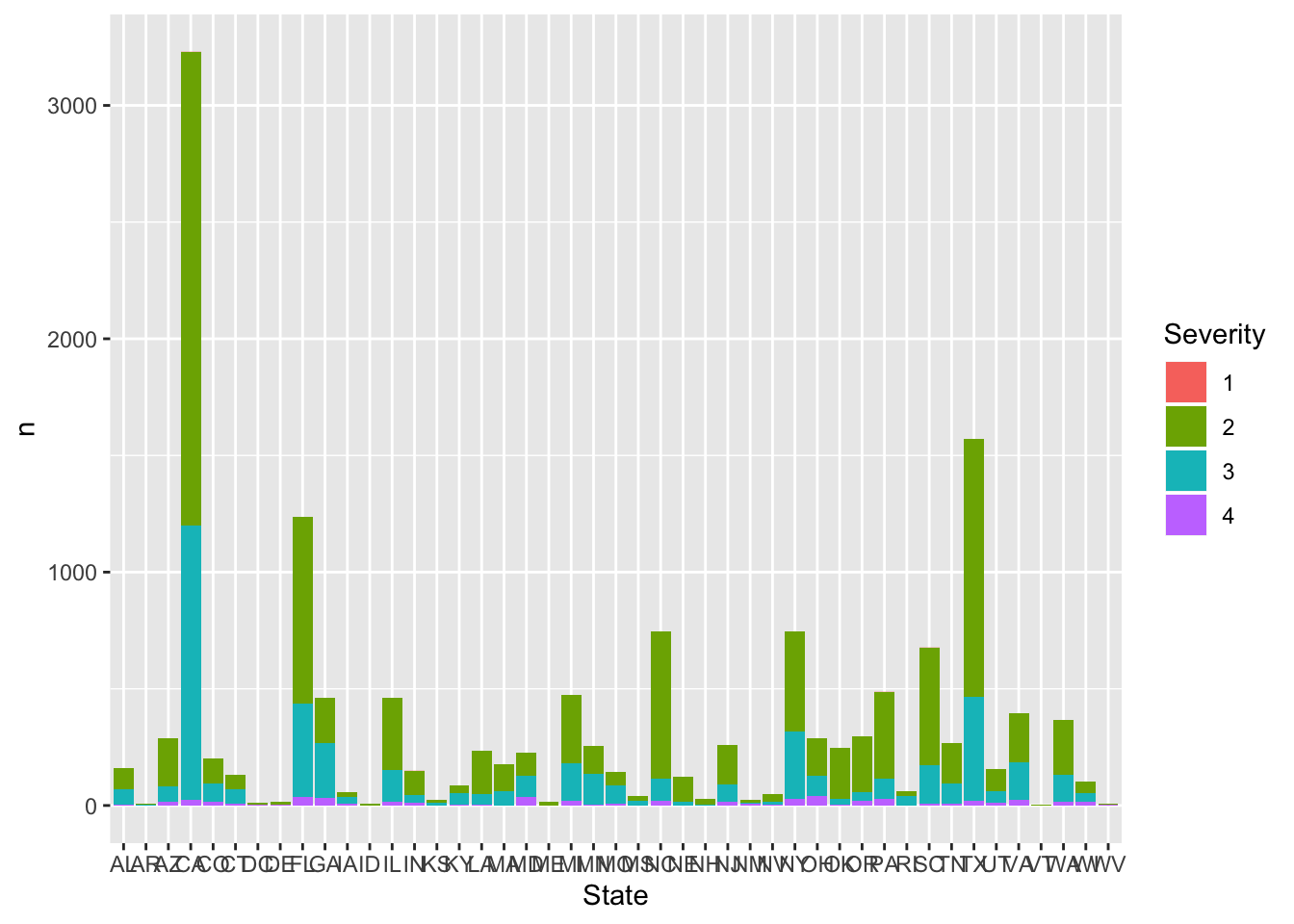
According to the World Health Organization vehicle accidents were the number 8 cause of death in the year 2016. With the human population increasing every year, only leading to more cars and people on the road, our group decided to analyze traffic accidents in the United States. The data set that we analyzed was provided to us by Sobhan Moosavi, via Kaggle.com. This data set contains over 3 million accident records dating from February of 2016 to December of 2019. The data provides many details about each accident, such as the date of the accident, the time of the accident, the zip code of where the accident occurred, and even dwells in deeper details on whether there was any participation, the wind speed, humidity, and much more. The most important column of the data we wanted to analyze for our predictive model though is going to be the column labeled severity. This column gives a numerical ranking based on how the accident effected the traffic jam caused afterwards. A 1 would indicate a very low impact on the traffic jam caused because of the accident, while a number 4 would result in a large traffic jam following the accident. In order to do this, we had to filter out the some of the data if they had N/A’s in categories that we were using such as the wind speed or humidity.

The predictive model we tried doing was KNN model. The main point of this is to use some of our data to predict the rest of our data. This is important because if we can do this accurately, we can try to predict future car crash outcomes to the best of our ability and makes the projections a lot more accurate. We set up our training/ predictive data as follows. The earliest date we chose was February 2nd of 2016, the first quartile March 24th of 2017, our mean date was December 5th of 2017, the third quartile date was August 21st of 2018, and finally the max date for the predictive data we chose was March 31st of 2019. We used this setup to predict our data for the period of April 2019 to December 2019. The variables we were trying to predict for this time period were the severity, the state the accident occurs in, and whether the data happened at morning or night. We entered all of the other data in the period to our model to try to predict these statistics and the results were varied.

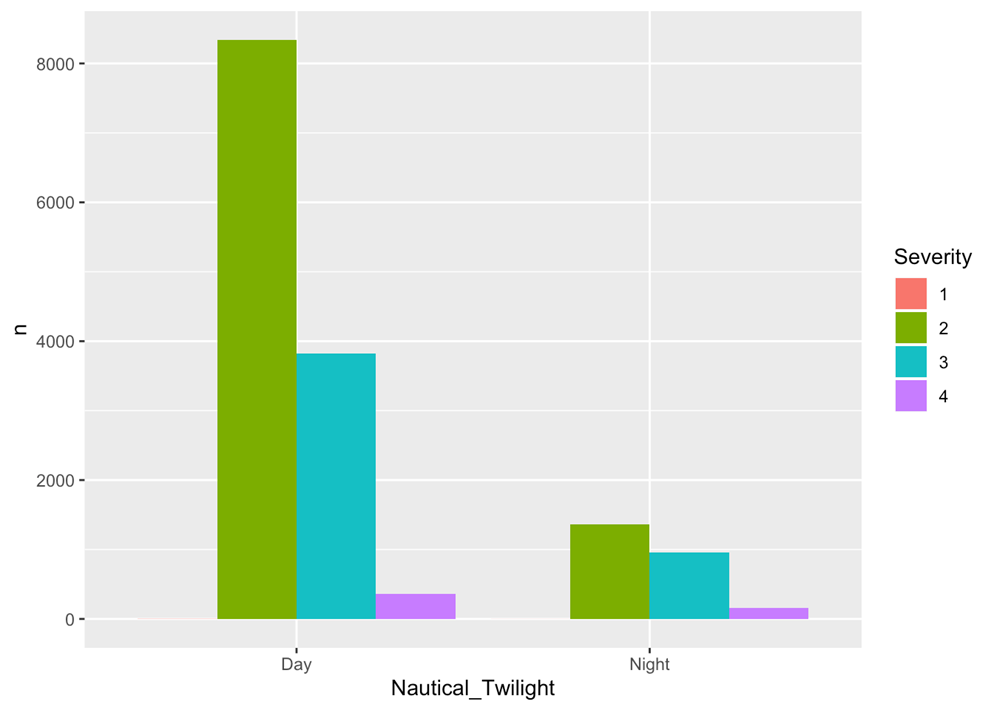
After filling in for missing data we created many different figures of the predicted data based on our KNN model. The first figure we created was a simple count of how many predicted accidents there would be between April 2019 and December 2019 broken down into the projected severity. As you can see in the model below the most anticipated severity of the car accidents seem to be at a 2. This means that most of the anticipated crashes in our predicted period will not cause too long of traffic jams. A severity level of 2 seems to make sense with the rest of the data too for 2 seems to be the mode severity amount. The fact that our predicted model showed no car crashes with a severity of 1 was very surprising to us. No matter how much we played around with the data and put different variables from this period into our training data, the fact that there was no severity 1 crashes for this period seemed to be consistent. After taking a deeper look into the actual data however we found that there were actually only 153 crashes with a severity 1 label in this period, therefore this projection seemed to be pretty spot on with that hypothesis.



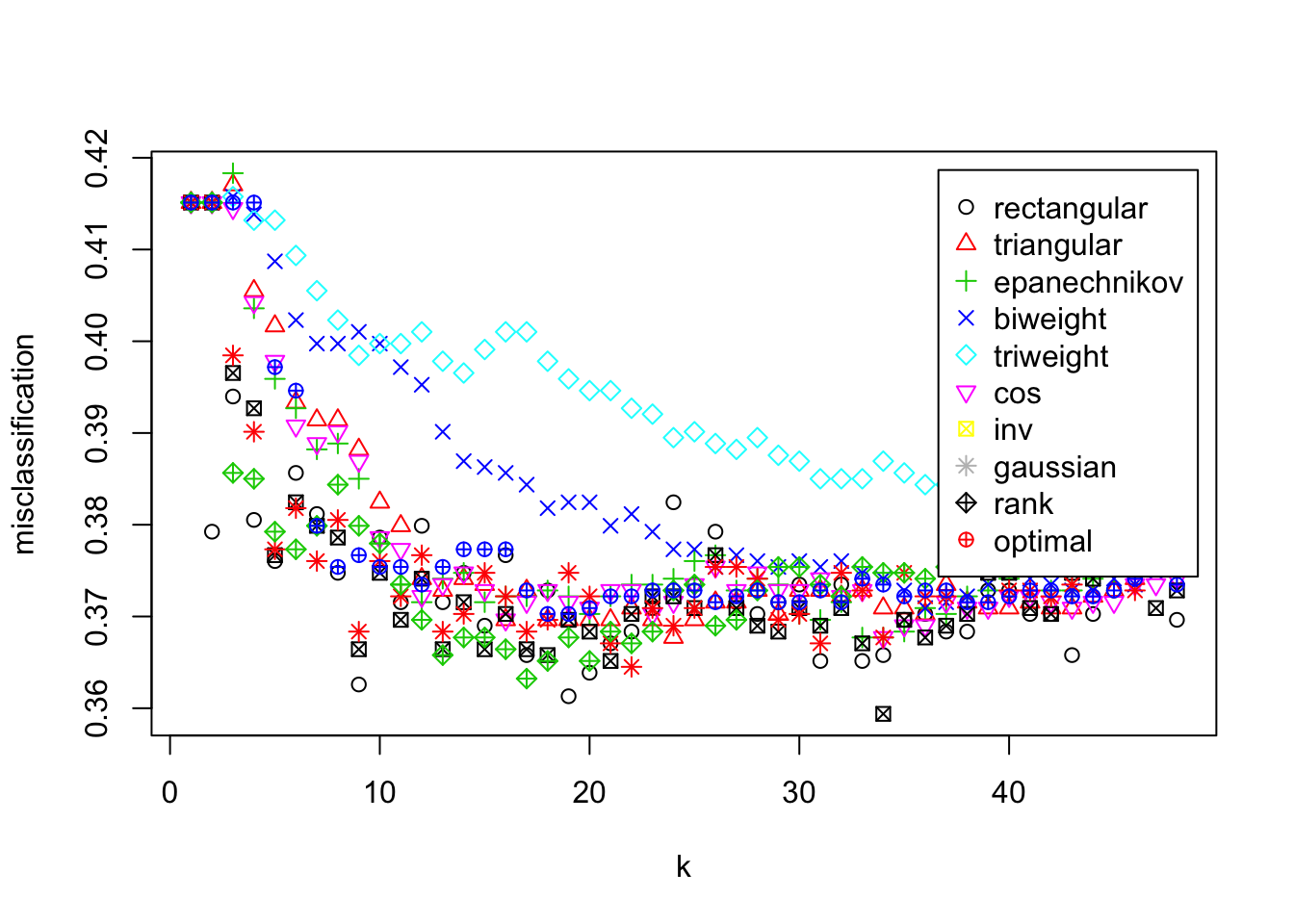
The next analytical model we created was trying to project the number of crashes by state. This graph is also broken into the severity of the accident. As we stated above, accidents with a severity of 2 are the most prominent therefore you will see a lot of green on the bar graph. It is notable to point out the two states with the largest amounts of total accidents on our model. These two states of course being California and Texas. This would seem to make sense just due to the population of those two states alone. In third place is Florida. Again just intuitively this would make a lot of sense based on population. This is another thing our model has seemed to gotten correct.



A third model we created was the severity based on the time of day for the projected period. We wanted to see whether the day or the night impacted the severity of the crashes between the periods of April and December 2019. To do this we used the column entitled Nautical Twilight which gave us whether it was the day or night. Not only were more crashes projected to happen during the day, but we also noticed that the ratio between severe 2 crashes to the rest of the severity’s dramatically increased during the days. For the nighttime there were more severe 2 crashes than severe 3 crashes however not by that much.



The last picture included is due to the fact that knn has many kinds of kernels, such as rectangular, and triangular (calculated distance). Because there is also a parameter for us to choose, that is, the number of k clusters, this picture includes the different kernels for the separate k values. In all this was a helpful tool to help us make our parameter selections.



In conclusion the model that we built ended up being a very reliable one for us when forecasting for the severity of car crashes for the period of April 2019 to December of 2019. Based on external factors of the car crash including time of day, wind speed, humidity, etc, the severities of the crash ended up matching up with the actual data 73.45 of the time. Our model could have maybe included other external forces that might have helped out accuracy a little more, however based on the data we were given, 73% seemed to be as accurate as we could get it. Overall based the parameters that we tested for we believe that this would be a good indicator to predict future crash data for other years such as 2020 or so on.

**Credits**

“Home.” *World Health Organization*, World Health Organization, www.who.int/.

“US-Accidents: A Countrywide Traffic Accident Dataset.” *Sobhan Moosavi*, 31 Dec. 2019, smoosavi.org/datasets/us\_accidents.

Moosavi, et al. “Accident Risk Prediction Based on Heterogeneous Sparse Data: New Dataset and Insights.” *ArXiv.org*, 19 Sept. 2019, arxiv.org/abs/1909.09638.