What is our guiding question?

What are the most dangerous roads and months to drive in the state of Virginia.

What is an observation in our study?

An observation in our study is the occurrence of an accident. This includes information like the location, date, weather, location type, and severity.

Supervised or unsupervised?

We are planning on using a supervised model.

Classification or regression?

We are planning on using a KNN classification model. The model will make predictions on how dangerous a certain road and date are to drive on, classifying by the interplay of month, count, and road.

The classifications will divide the roads into thirds of the highest, middle and lowest accidents per month. For example:

- “Highest Risk”
- “Medium Risk”
- “Lowest Risk”

What models or algorithms do you plan to use in your analysis? How?

We are using a KNN classification model. The model will take into account the interplay of month, road and previous accident counts to predict the classification of what the road will be on specific months in the future.

How will you know if your approach "works"? What does success mean?

Metrics we are using to measure success:

1. Accuracy:
    a. The model should predict the classification of roads based upon month with 80% - 90% accuracy. We are using this percentage as our “works” benchmark because that means that the model performs correctly a vast majority of the time.
2. Logic
    a. The model should make sense in its predictions. For instance, say one random month, a usually dangerous road had far less accidents than usual. The model should still predict this month and road to be “high risk” despite this once instance of it being lower because of the other months where it was high. (this is a matter of selecting the ideal K)
    b. Essentially, there shouldn’t be cases where I-95 is assessed as a “low risk” and likewise there shouldn't be cases where Wertland Street in Charlottesville is predicted as “high risk”.

What are weaknesses that you anticipate being an issue? How will you deal with them if they come up? If your approach fails, what might you learn from this unfortunate outcome?

We anticipate difficulty in developing a hard definition/algorithm to determine what is a “high risk” road. This is largely due to the population density– if 100,000 people travel on a road on a daily basis and there are 20 crashes, that is less risky than a road with 4 crashes with 10,000 daily commuters. This may limit the applications of our model, or may force us to make multiple models depending on the type of road. For instance, there may be a model that predicts highways, local major roads, and then smaller roads. That being said, feature engineering is key for our project, using the right factors to calculate the risk scores is pivotal for the success of our model/project without necessarily relying on the number of accidents based on the number of cars/people traveling.

Assuming that this calculation is as troublesome as we are intending, we plan to find a database/dataset that includes population distribution. This would enable us to utilize the data to calculate a value that accurately represents the number of people that, on average, use that road at certain times throughout the year– accounting for common holiday and seasonal travel. This would provide us with an accurate “road density” value to utilize in our algorithm that will be used to appropriately calculate the risk of the road.

While we are confident in our ability to solve this problem, utilizing the approach listed above, there is a possibility that our approach is not successful. In this situation, we would first address the calculation of the predictions, particularly the type of algorithm we are using. If this is not the route of the problem, we will then reevaluate the calculation of the “road density” value and see if it is proficient in contributing to the prediction of a dangerous road. If we cannot attain a “road density” value that properly represents the true values and contributes to accuracy, we will have to create a simpler model that only assesses based upon total count. This is fine, but the model will only predict well on roads that are high traveled + high accidents / low traveled low accidents. The middle ground roads, where they may have been traveled a medium amount, will not be as well predicted.

One thing that we would learn from this experience is that attaining the pre-calculated, direct data is sufficient in a lot of situations. While it may not be as readily available as the population distribution due to the census among other commonly performed data gathering surveys, the direct data has ensured accuracy.

Another thing we would learn is that testing is necessary in any project. If KNN does not yield the desired results, it might highlight the limitations of using a distance-based classifier for a task that has complex interactions between variables. This experience could reinforce the importance of testing different algorithms early in the modeling process and of not overcommitting to a particular method. We could also revisit the classifications and discuss if that is the optimal way to approach our goal.

Feature Engineering:

We will need to hot encode the road variable and the month as these are categorical variables. We will use this feature in accordance with the count of accidents over the past years to predict the road risk.

Results:

We plan to measure the accuracy of our model through a confusion matrix. For instance, we will compare the number of predicted “high risk” travel plans to what the actual classification of those models are. This will easily allow us to assess our models' performance.