Car Collision Analysis in WA State for 2022

Group 4: Donald Yin, Anas Fathul, Julien Johnson

Data Sources

We used two sources of for our data

- Primary data set was from Washington State Patrol - Collision Analysis Database
    - Database stores car collision data submitted by law enforcement officers including level of severity, if there were pedestrians involved, which road did the collision happen etc.
    - It does not include level of traffic at each collision.
- Supplementary data was from Washington State Department of Transportation - Traffic Counts (AADT) for 2022
    - For each state road and on parts of each state road, it records the Annual Average Daily Traffic (i.e. the level of traffic), as well as the milepost
- Uploaded them online to have access to them

Research Questions

1. Are there specific areas (county, city, or specific road) that are more prone to collision? Are these characterized by fewer or less severe collisions?
2. What are the key predictors contributing to road collisions in Washington state at various scales (County, City level, weather, etc.)?
3. Can we develop a model to predict high-risk scenarios or high-risk drivers based on historical data and define scenarios that are more likely to involve severe collisions?

Data Cleaning

```
def clean_crash_summary(file_path):
    # Read the CSV file
    crash_summary = pd.read_csv(file_path)

    # Filter and select relevant columns
    crash_summary_clean = crash_summary[(crash_summary['Jurisdiction'] == 'State Road') &
                                        (~crash_summary['Weather Condition'].isna())].drop(columns=['Collision Type'])

    # Create binary columns and convert categorical columns to category type
    crash_summary_clean = crash_summary_clean.assign(
        **{
            'School Zone': crash_summary_clean['School Zone'].apply(lambda x: 1 if x == 'Y' else 0),
            'Intersection Related': crash_summary_clean['Intersection Related'].apply(lambda x: 1 if x == 'Y' else 0),
            'Damage Threshold Met': crash_summary_clean['Damage Threshold Met'].apply(lambda x: 1 if x == 'Y' else 0),
            'Hit and Run': crash_summary_clean['Hit and Run'].apply(lambda x: 1 if x == 'Y' else 0),
            'Passengers Involved': crash_summary_clean['Passengers Involved'].apply(lambda x: 1 if x == 'Y' else 0),
            'Commercial Carrier Involved': crash_summary_clean['Commercial Carrier Involved'].apply(lambda x: 1 if x == 'Y' else 0),
            'School Bus Involved': crash_summary_clean['School Bus Involved'].apply(lambda x: 1 if x == 'Y' else 0),
            'Agency': crash_summary_clean['Agency'].astype('category'),
            'Weather Condition': crash_summary_clean['Weather Condition'].astype('category'),
            'Lighting Condition': crash_summary_clean['Lighting Condition'].astype('category'),
            'Injury Severity': crash_summary_clean['Injury Severity'].astype('category')
        }
    )

    return crash_summary_clean
```

We need cleaned the data for the collisions (primarily changes a lot of variables into 1s and 0s.

Data Structure

Binary (Y,N):

- School Zone
- Intersection Related
- Damage Threshold Met
- Hit and Run
- Passengers Involved
- Commercial Carrier Involved
- School Bus Involved

Numerical

- Motor Vehicles Involved
- Pedestrian Involved
- Pedal cyclists Involved

Factor

- Agency 4
- Weather Condition 11
- Lighting Condition 9
- Injury Severity 5

Date/time

- Collision Date

We end up with 43422 data points with 27 features.

We will use Injury Severity (5 level) to assess the severity of a collision which consists of:

- **Low Severity Collision**
  - Unknown Injury Collision
  - No Injury Collision
  - Minor Injury Collision
- **High Severity Collision**
  - Serious Injury Collision
  - Fatal Collision

Dataset Creation

We combined those data sets into one

- We looked at ONLY the data on collision for state roads and only during 2022
  - To correspond with the 2022 Traffic Count Data
- Found the State Road each collision was on using regex
- Checked if the gotten State Road is a valid state road by checking the state roads on the Traffic Counts data.
- For each Collision
  - if it has a milepost, we would get the AADT of the nearest mile post nearest to it
  - If it didn't, would instead get the median AADT of the state road.
- Removed all collisions that did not have an AADT value.
  - Removed 1184 Data points out of 44606 Data points (roughly 2.5% of our data)

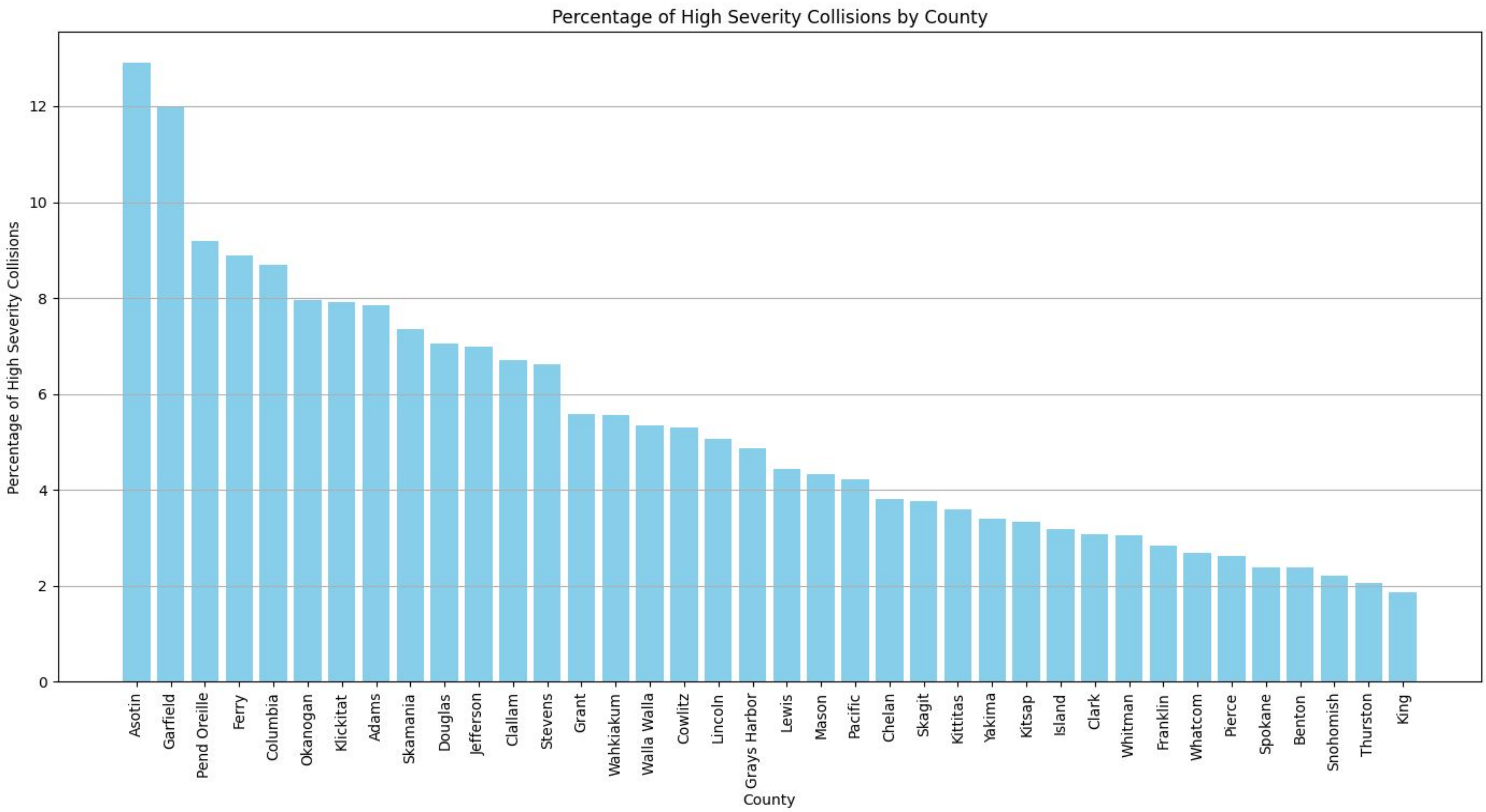
```
validSR = AADT['StateRouteNumber'].unique().tolist()
Trafficway = ["Primary Trafficway","Secondary Trafficway"]
dict = {
  0: [],
  1: [],
}
```

```
for i in np.arange(2):

    for x in Car_Crash[Trafficway[i]]:
        #Gets the state roads numbers from the primary trafficway. Has to match "[String of nondigit text]integer"
        #Does remove some values that don't follow road convenient of state roads. Does account for roads like 123th Highway
        if type(x) == str:
            if (not re.match("\D+\d+", x) == None):
                if int(re.findall(r'\d+', x)[0]) in validSR:
                    dict[i].append(int(re.findall(r'\d+', x)[0]))
                else:
                    dict[i].append(None)
            else:
                dict[i].append(None)
        else:
            dict[i].append(None)

State_Road_Num = []
for x in np.arange(len(dict[0])):
    #From the 2 lists of state road numbers, it will first the state road number of the primary trafficway
    #If that is not available, it takes the state road number of the secondary trafficway
    if not dict[0][x] == None:
        State_Road_Num.append(dict[0][x])
    else:
        State_Road_Num.append(dict[1][x])
```

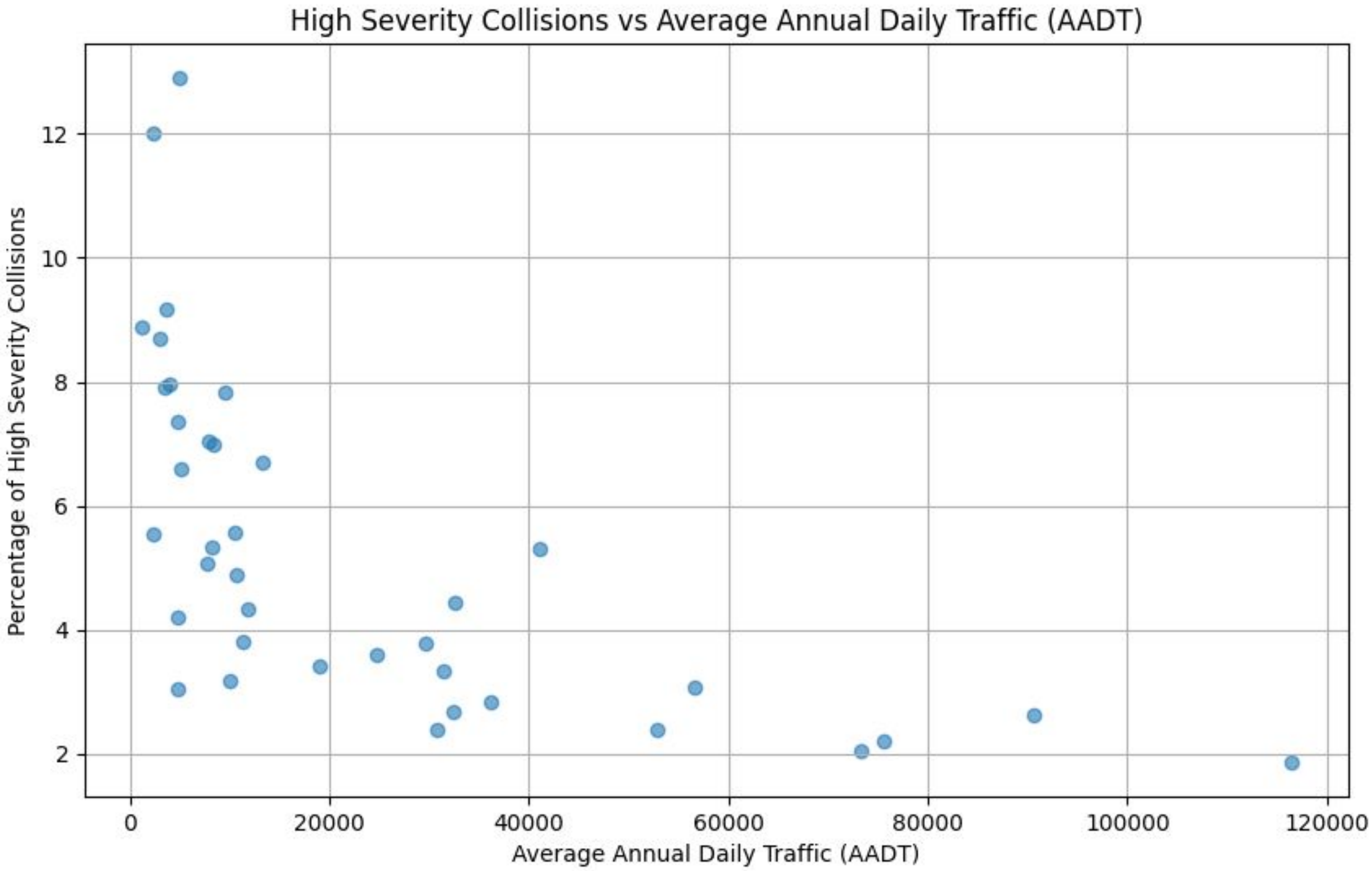
Per County Analysis

```python
# Classify collision severity
high_severity = ["Fatal Collision", "Serious Injury Collision"]

# Create a new column 'Collision Severity'
data['Collision Severity'] = data['Injury Severity'].apply(lambda x: 'High' if x in high_severity else 'Low')

# Summarize the data by county
county_summary = data.groupby('County').agg(
    Total_Collisions=('Collision Report Number', 'count'),
    High_Severity_Collisions=('Collision Severity', lambda x: (x == 'High').sum()),
    Low_Severity_Collisions=('Collision Severity', lambda x: (x == 'Low').sum()),
    Avg_AADT=('AADT', 'mean')
).reset_index()

# Calculate the percentage of high severity collisions
county_summary['High_Severity_Percentage'] = (county_summary['High_Severity_Collisions'] / county_summary['Total_Collisions']) * 100

# Sort the counties by percentage of high severity collisions
county_summary = county_summary.sort_values(by='High_Severity_Percentage', ascending=False)
print(county_summary.head())
```

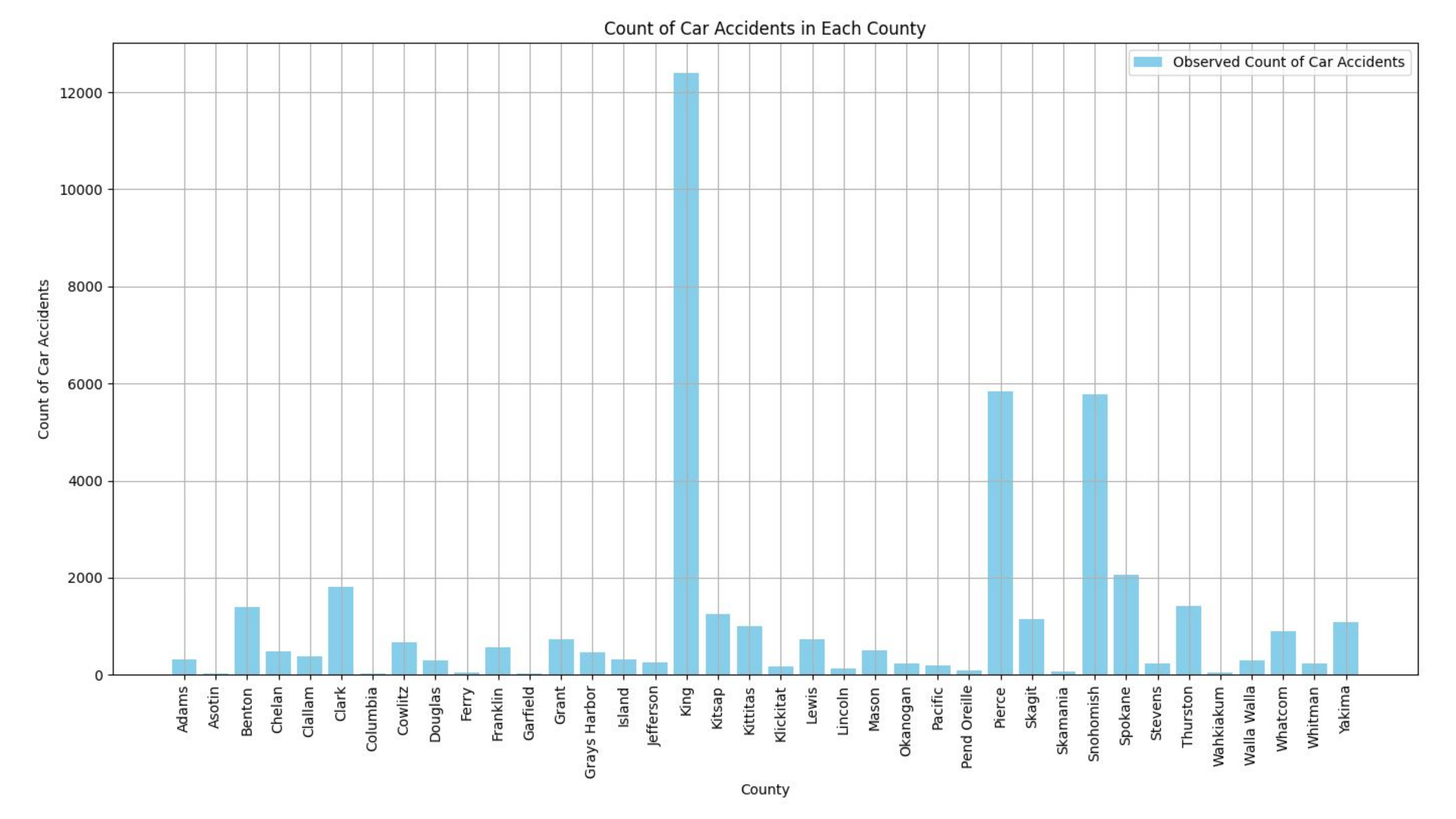
```
          County  Total_Collisions  High_Severity_Collisions  Low_Severity_Collisions     Avg_AADT  High_Severity_Percentage
          Asotin                31                         4                       27  5035.161290                 12.903226
        Garfield                25                         3                       22  2429.200000                 12.000000
    Pend Oreille                98                         9                       89  3588.673469                  9.183673
           Ferry                45                         4                       41  1243.777778                  8.888889
        Columbia                23                         2                       21  3016.521739                  8.695652
```



Percentage of High Severity Collisions by County
Percentage of High Severity Collisions
0
2
4
6
8
10
12
County
Asotin
Garfield
Pend Oreille
Ferry
Columbia
Okanogan
Klickitat
Adams
Skamania
Douglas
Jefferson
Clallam
Stevens
Grant
Wahkiakum
Walla Walla
Cowlitz
Lincoln
Grays Harbor
Lewis
Mason
Pacific
Chelan
Skagit
Kittitas
Yakima
Kitsap
Island
Clark
Whitman
Franklin
Whatcom
Pierce
Spokane
Benton
Snohomish
Thurston
King



High Severity Collisions vs Average Annual Daily Traffic (AADT)
Percentage of High Severity Collisions
Average Annual Daily Traffic (AADT)
0
20000
40000
60000
80000
100000
120000
2
4
6
8
10
12



Count of Car Accidents in Each County
Observed Count of Car Accidents
Count of Car Accidents
0
2000
4000
6000
8000
10000
12000
County
Adams
Asotin
Benton
Chelan
Clallam
Clark
Columbia
Cowlitz
Douglas
Ferry
Franklin
Garfield
Grant
Grays Harbor
Island
Jefferson
King
Kitsap
Kittitas
Klickitat
Lewis
Lincoln
Mason
Okanogan
Pacific
Pend Oreille
Pierce
Skagit
Skamania
Snohomish
Spokane
Stevens
Thurston
Wahkiakum
Walla Walla
Whatcom
Whitman
Yakima

Model Prediction

To predict the severity level of certain collisions we used a Naive Bayesian to classify each data point.

The rationale of using this is that the variables (like weather condition, state road number, AADT, etc.) are assumed to be independent since they seemingly don’t affect one another (a key assumption for Naive Bayesian).

Uses probability to classify the data, classifying the data on which probability is the greatest.

Model Prediction

Classified on 1 of 4 severity levels (No injury, minor injury, serious injury, and fatal) (dropping any data points with unknown injury level)

It would consider the following variables

- State Road Number, Milepost, Intersection Related, Weather Condition, Lighting Condition, AADT, and whether the following were involved:
    - Motor Vehicles
    - Passengers
    - Commercial Carrier
    - School Bus
    - Pedestrians
    - Pedal cyclists

```python
def PredictionReport(X_values: pd.DataFrame, Y_values: pd.DataFrame, classifier: str = "Gaussian") -> None:
    Observation = X_values.to_numpy()
    Results = Y_values.to_numpy().ravel()
    X_train, X_test, y_train, y_test = train_test_split(Observation, Results, test_size=0.2, random_state=0)

    gnb = GaussianNB()
    if classifier == "Multinomial":
        gnb = MultinomialNB()
    elif classifier == "Complement":
        gnb = ComplementNB()
    elif classifier == "Categorical":
        gnb = CategoricalNB()

    y_pred = gnb.fit(X_train, y_train).predict(X_test)
    print("Number of mislabeled points out of a total %d points : %d"
          % (X_test.shape[0], (y_test != y_pred).sum()))
    conf_matrix = confusion_matrix(y_test, y_pred)

    # Generate classification report
    class_report = classification_report(y_test, y_pred)
```

```
# Print classification report
print("\nClassification Report:")
print(class_report)
values_index = Y_values['Injury Severity'].unique().tolist()
values_index.sort()
cm_df = pd.DataFrame(conf_matrix,
                 index = values_index,
                 columns = values_index)
#Plotting the confusion matrix
plt.figure(figsize=(5,4))
sns.heatmap(cm_df, annot=True)
plt.title('Confusion Matrix')
plt.ylabel('Actal Values')
plt.xlabel('Predicted Values')
plt.show()

unique_values, counts = np.unique(y_pred, return_counts=True)
print("Predicted values:")
for value, count in zip(unique_values, counts):
    print(f"{value} occurs {count} times")
unique_values, counts = np.unique(y_test, return_counts=True)

print("\nActual values:")
for value, count in zip(unique_values, counts):
    print(f"{value} occurs {count} times")

return None
```

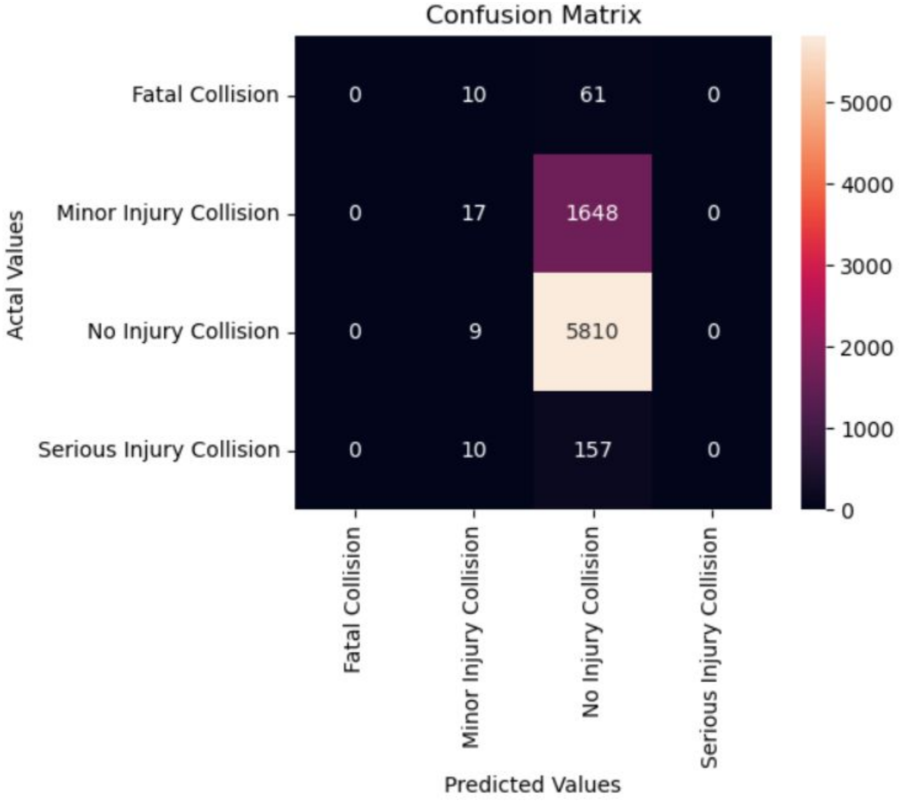
Prediction

1st we predicted using the Categorical Naive Bayesian:

It is described as the following: “The categorical Naive Bayes classifier is suitable for classification with discrete features that are categorically distributed. The categories of each feature are drawn from a categorical distribution.”

Our test size was 20% of the data

```
Number of mislabeled points out of a total 7722 points : 1895

Classification Report:
                          precision    recall  f1-score   support

         Fatal Collision       0.00      0.00      0.00        71
   Minor Injury Collision       0.37      0.01      0.02      1665
      No Injury Collision       0.76      1.00      0.86      5819
 Serious Injury Collision       0.00      0.00      0.00       167

                 accuracy                           0.75      7722
                macro avg       0.28      0.25      0.22      7722
             weighted avg       0.65      0.75      0.65      7722
```




```
Predicted values:
Minor Injury Collision occurs 46 times
No Injury Collision occurs 7676 times

Actual values:
Fatal Collision occurs 71 times
Minor Injury Collision occurs 1665 times
No Injury Collision occurs 5819 times
Serious Injury Collision occurs 167 times
```

Prediction

As you can see it only really predicts 1 of two values as.

It has 100% recall for No Injury but because it was guessing no injury since its the majority of our data.

While it does “well” in term of accuracy, it is presumably being overfit with the larger amount of data primarily consisting of no injury collisions.

We did a second prediction with using Complement Naive Bayesian Classification

“ CNB is an adaptation of the standard multinomial naive Bayes (MNB) algorithm that is particularly suited for imbalanced data sets.”

```
Number of mislabeled points out of a total 7722 points : 3954
0.48795648795648794

Classification Report:
                          precision    recall  f1-score   support

         Fatal Collision       0.50      0.04      0.08        71
   Minor Injury Collision       0.23      0.43      0.30      1665
      No Injury Collision       0.78      0.52      0.62      5819
 Serious Injury Collision       0.04      0.16      0.06       167

                 accuracy                          0.49      7722
                macro avg       0.39      0.29      0.27      7722
             weighted avg       0.64      0.49      0.54      7722
```


```
Predicted values:
Fatal Collision occurs 6 times
Minor Injury Collision occurs 3113 times
No Injury Collision occurs 3870 times
Serious Injury Collision occurs 733 times

Actual values:
Fatal Collision occurs 71 times
Minor Injury Collision occurs 1665 times
No Injury Collision occurs 5819 times
Serious Injury Collision occurs 167 times
```

Model Evaluation

While the first prediction model does better, it generally cannot seem to predict much other collisions besides No Injury.

The second does worse but can actually predict to some extent other collisions and generally has higher precision.

We would suggest that the second model is preferred despite lower accuracy because of it this and it seemingly has a reduced case of overfitting.

1st Model

Number of mislabeled points out of a total 7722 points : 1895

Classification Report:

	precision	recall	f1-score	support
Fatal Collision	0.00	0.00	0.00	71
Minor Injury Collision	0.37	0.01	0.02	1665
No Injury Collision	0.76	1.00	0.86	5819
Serious Injury Collision	0.00	0.00	0.00	167
accuracy			0.75	7722
macro avg	0.28	0.25	0.22	7722
weighted avg	0.65	0.75	0.65	7722

2nd Model

Number of mislabeled points out of a total 7722 points : 3954
0.48795648795648794

Classification Report:

	precision	recall	f1-score	support
Fatal Collision	0.50	0.04	0.08	71
Minor Injury Collision	0.23	0.43	0.30	1665
No Injury Collision	0.78	0.52	0.62	5819
Serious Injury Collision	0.04	0.16	0.06	167
accuracy			0.49	7722
macro avg	0.39	0.29	0.27	7722
weighted avg	0.64	0.49	0.54	7722

Model Evaluation

As you can see it is possible to predict the severity of car crash using naive bayesian. However it is subject to the potential overfitting from the lots of data of no injury. Complement Naive Bayesian seems to address this and give a more versatile predictor that while is less accurate, can be more precise for all collision types.

Classification by Random Forest

We aim to identify the key predictors contributing to road collisions in Washington state.

We will use the Random Forest model to find the features importance across predictors:

- An ensemble classifier that uses multiple decision tree models.
- It improves prediction accuracy and control overfitting by averaging multiple decision trees.
- Can be used for classification or Regression.

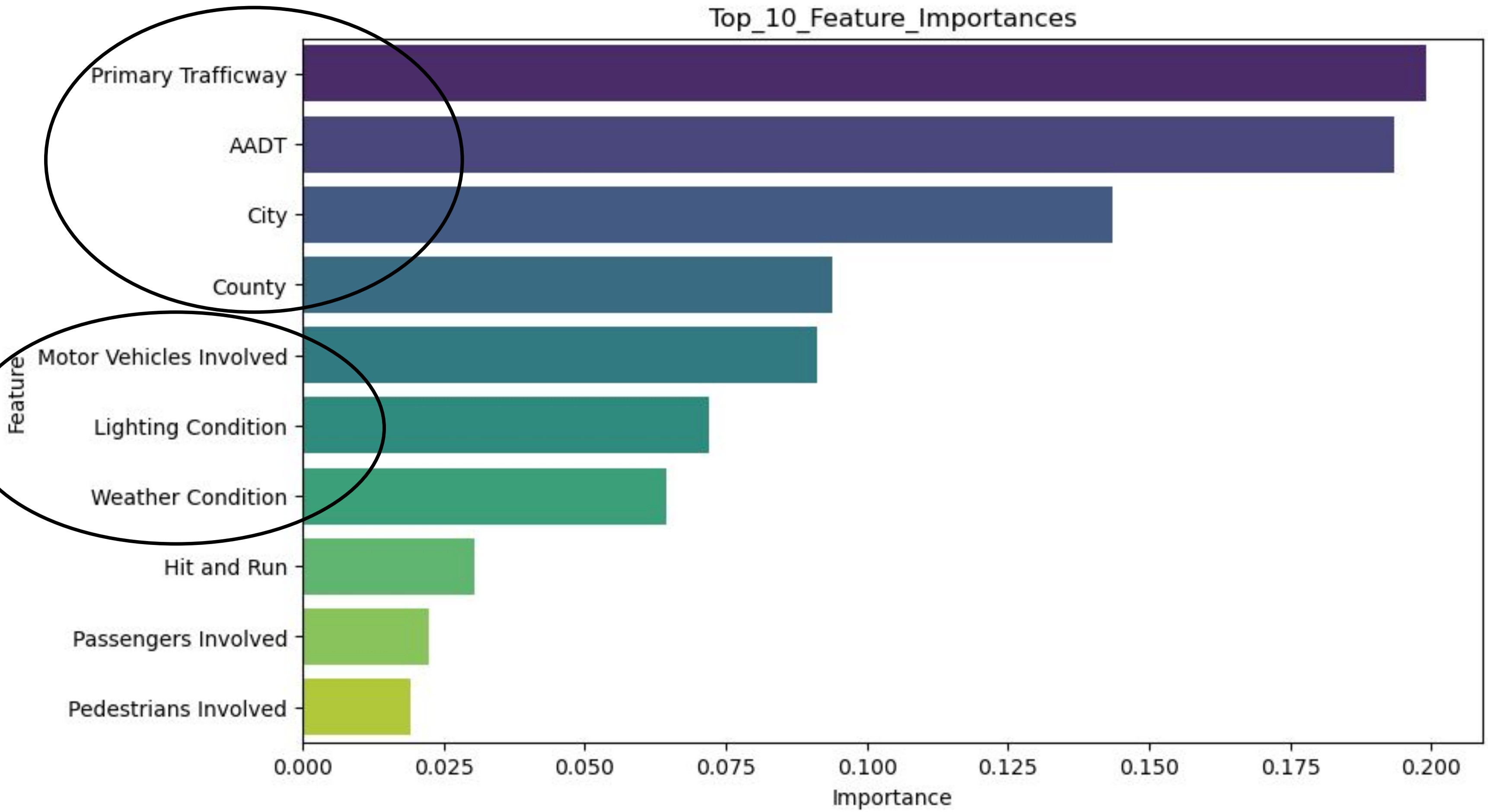
We choose Random Forest as it can handle large datasets with high dimensionality, robust against overfitting, and provide features importance measures.

Data Preparation -> Features Considered -> Data Splitting -> Train the model -> 5 K-fold Cross Validation for tuning parameter -> Test Model

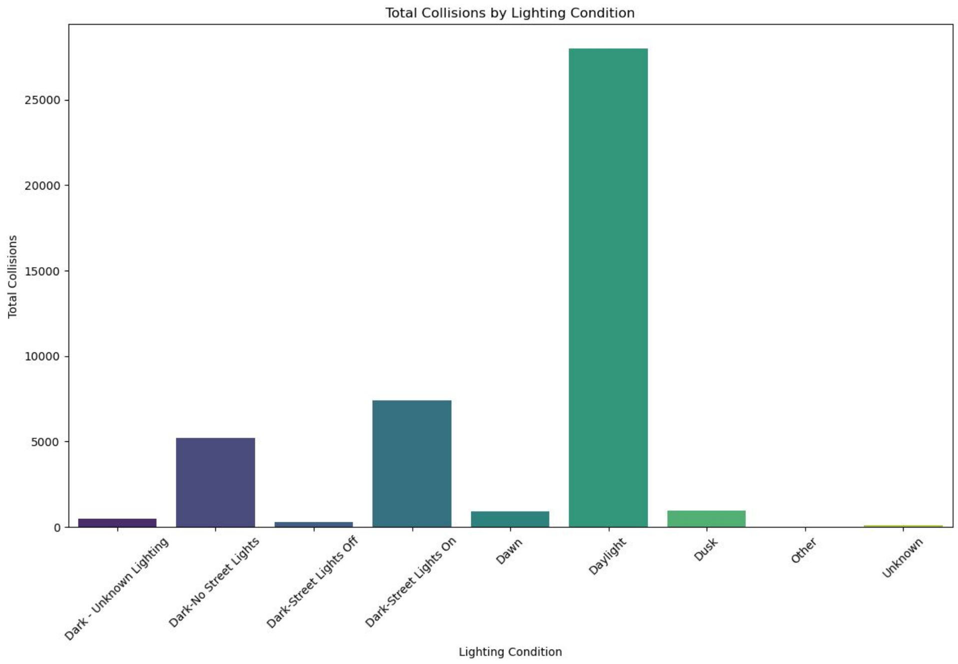
Tuning parameter:

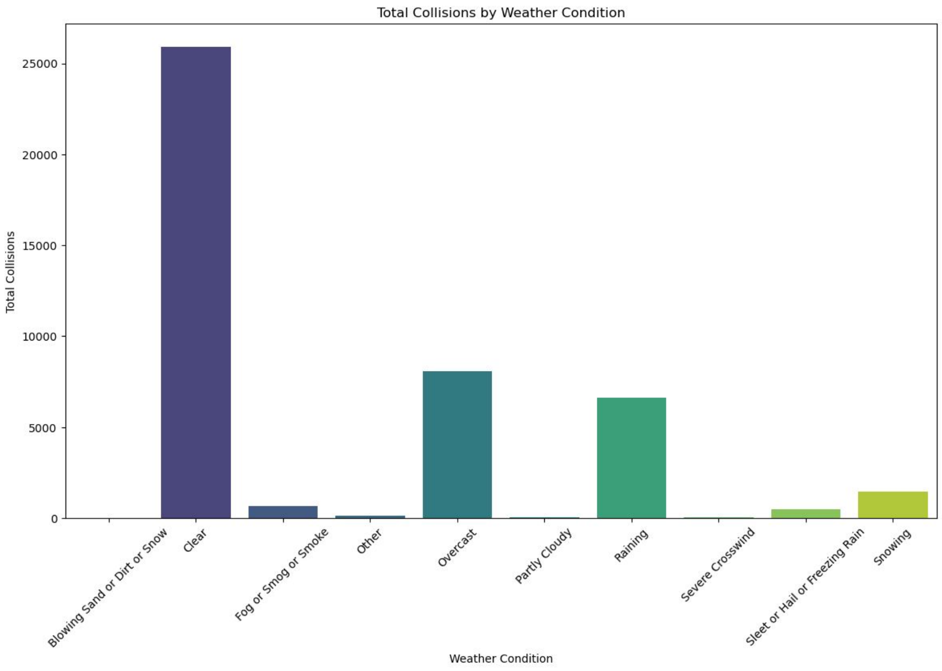
- n_estimator : number of decisions trees
- max_depth: depth of decision trees
- Whole dataset is use instead of bootstrap

Random Forest Result



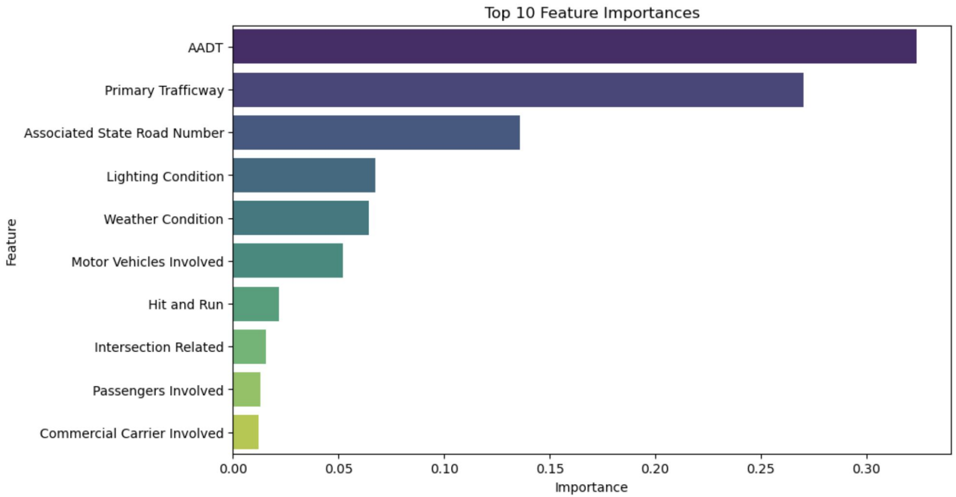
Lighting & Weather Conditions

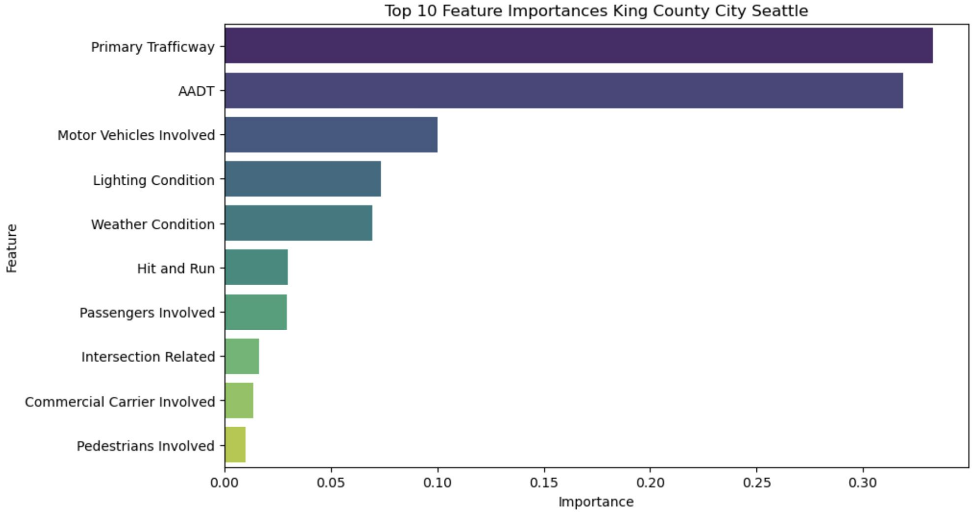




Problems

- Too much data points from higher AADT trafficways/counties
  - Led to skewed results because high traffic regions became the primary feature of collisions.
  - Tried focusing in on the city level but came out with the same results



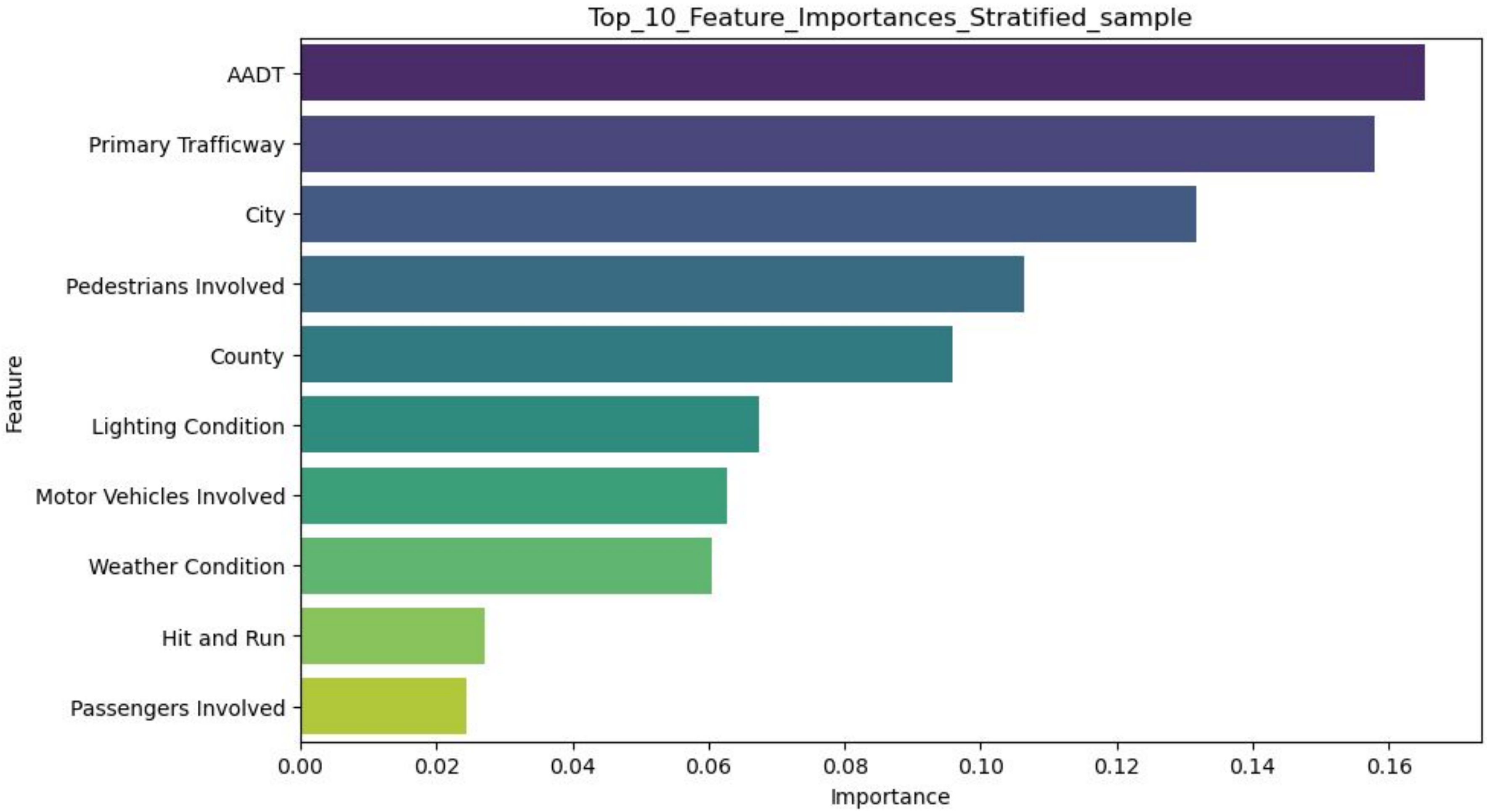


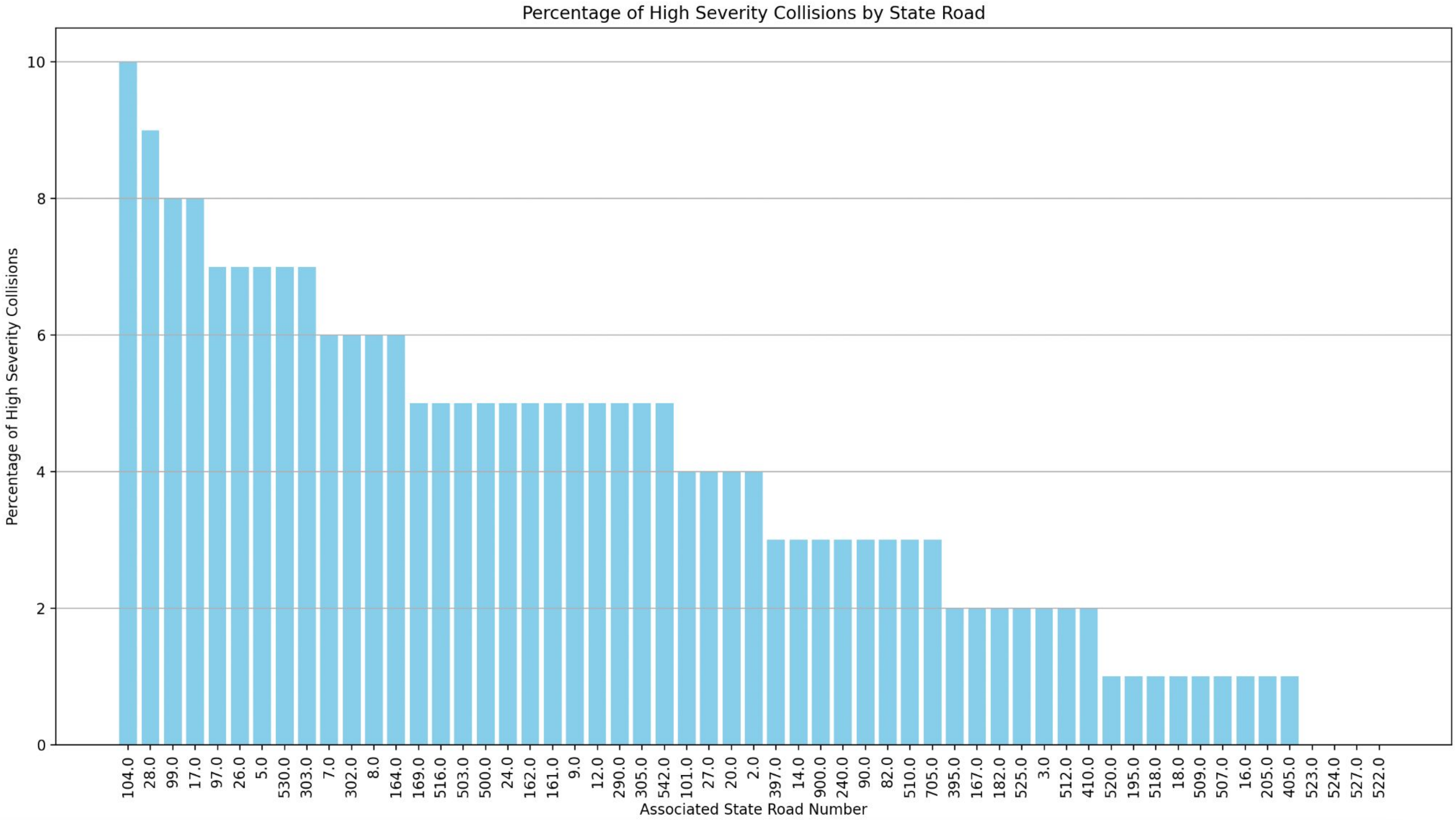
Solution

- Ran a stratified random sample by interstate road.
    - For each interstate road that had >100 reported accidents, we randomly sampled 100.

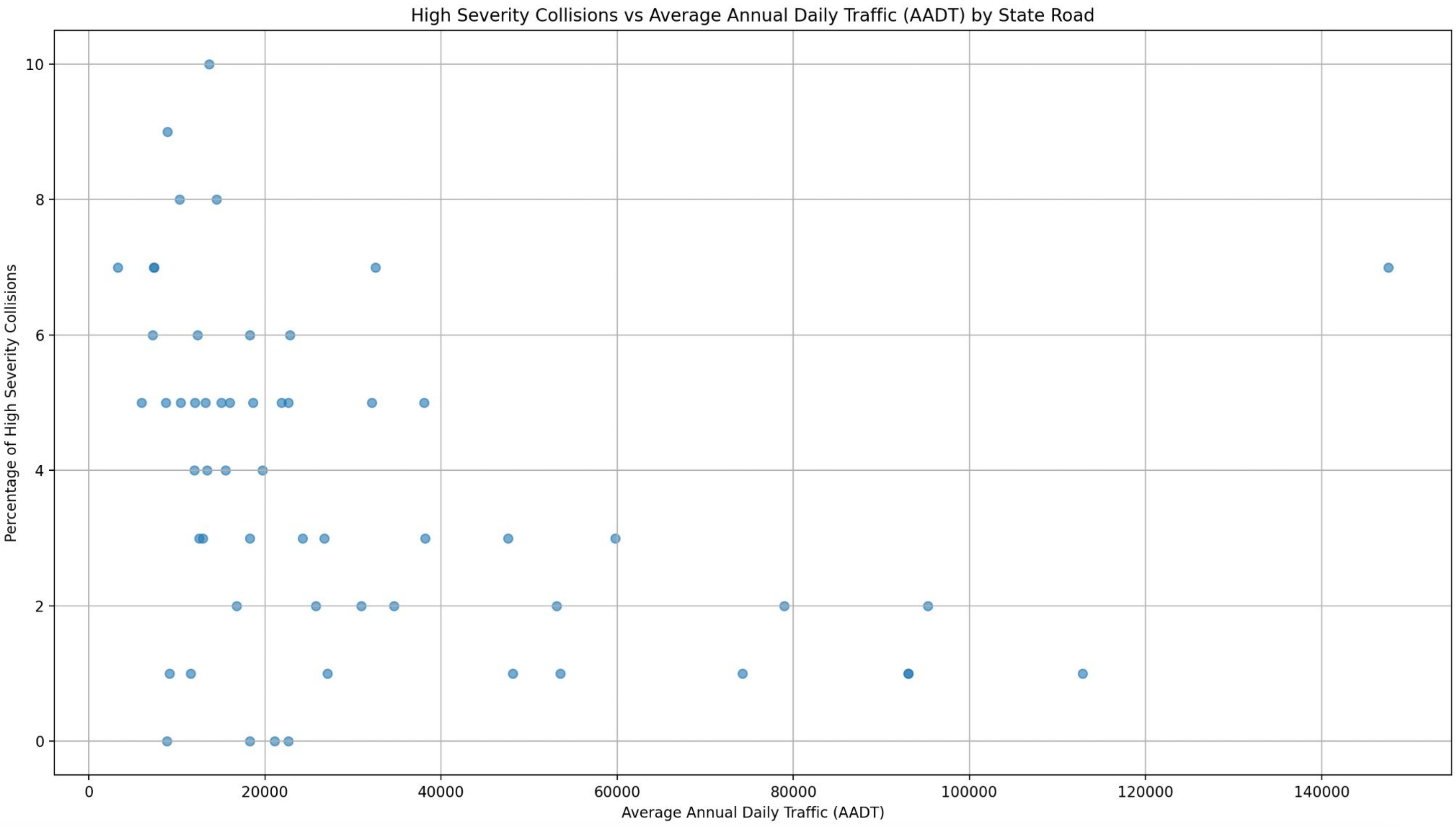
```
filtered_df = df.groupby('Associated State Road Number').filter(lambda x: len(x) > 100)
sampled_df = filtered_df.groupby('Associated State Road Number').apply(lambda x: x.sample(100)).reset_index(drop=True)
```

- This allowed us to have a more standard spread of collision data across major traffic regions.


Top_10_Feature_Importances_Stratified_sample
AADT
Primary Trafficway
City
Pedestrians Involved
County
Lighting Condition
Motor Vehicles Involved
Weather Condition
Hit and Run
Passengers Involved
Feature
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16
Importance



Percentage of High Severity Collisions by State Road
Percentage of High Severity Collisions
Associated State Road Number
0
2
4
6
8
10
104.0
28.0
99.0
17.0
97.0
26.0
5.0
530.0
303.0
7.0
302.0
8.0
164.0
169.0
516.0
503.0
500.0
24.0
162.0
161.0
9.0
12.0
290.0
305.0
542.0
101.0
27.0
20.0
2.0
397.0
14.0
900.0
240.0
90.0
82.0
510.0
705.0
395.0
167.0
182.0
525.0
3.0
512.0
410.0
520.0
195.0
518.0
18.0
509.0
507.0
16.0
205.0
405.0
523.0
524.0
527.0
522.0


High Severity Collisions vs Average Annual Daily Traffic (AADT) by State Road
Percentage of High Severity Collisions
Average Annual Daily Traffic (AADT)
0
2
4
6
8
10
0
20000
40000
60000
80000
100000
120000
140000

LS on the Variables and the Effects

To get an initial idea of the effects of each variable, we set up an Homoskedastic Robust Least Squares model to regress on the Injury Severity (Low or High) (0,1 respectively) by:

- School Zone (was it at a school zone)
- Intersection Related (was it at an intersection)
- Damage Threshold Met (whether the collision cost more than $1,000),
- whether it was a Hit and Run,
- Number of Motor Vehicles Involved,
- Whether if the following were involved (0,1) for each
  - Passengers
  - Commercial Carrier
  - School Bus
  - Pedestrians
  - Pedalcyclists
- AADT (traffic level)
- Dummy Variables for each Lighting Condition, Weather Condition, and County

OLS Regression Results

Dep. Variable:	Injury Severity	R-squared:	0.079
Model:	OLS	Adj. R-squared:	0.078
Method:	Least Squares	F-statistic:	23.22
Date:	Wed, 29 May 2024	Prob (F-statistic):	2.08e-267
Time:	08:56:20	Log-Likelihood:	17712.
No. Observations:	43422	AIC:	-3.529e+04
Df Residuals:	43356	BIC:	-3.472e+04
Df Model:	65		
Covariance Type:	HC1		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.0263	0.007	3.876	0.000	0.013	0.040
School Zone	-0.0320	0.011	-2.834	0.005	-0.054	-0.010
Intersection Related	0.0019	0.003	0.718	0.473	-0.003	0.007
Damage Threshold Met	0.0049	0.002	2.531	0.011	0.001	0.009
Hit and Run	-0.0147	0.002	-8.287	0.000	-0.018	-0.011
Motor Vehicles Involved	0.0037	0.001	2.545	0.011	0.001	0.007
Passengers Involved	0.0126	0.002	6.659	0.000	0.009	0.016
Commercial Carrier Involved	0.0123	0.003	4.011	0.000	0.006	0.018
School Bus Involved	0.0452	0.071	0.637	0.524	-0.094	0.184
Pedestrians Involved	0.4993	0.028	17.534	0.000	0.443	0.555
Pedalcyclists Involved	0.1887	0.055	3.456	0.001	0.082	0.296
AADT	-8.578e-08	1.35e-08	-6.335	0.000	-1.12e-07	-5.92e-08

County_Adams	0.0227	0.015	1.528	0.126	-0.006	0.052
County_Asotin	0.0771	0.058	1.328	0.184	-0.037	0.191
County_Benton	-0.0270	0.005	-5.038	0.000	-0.038	-0.017
County_Chelan	-0.0175	0.009	-1.943	0.052	-0.035	0.000
County_Clallam	0.0115	0.013	0.916	0.359	-0.013	0.036
County_Clark	-0.0193	0.005	-3.572	0.000	-0.030	-0.009
County_Columbia	0.0378	0.057	0.658	0.511	-0.075	0.150
County_Cowlitz	0.0012	0.009	0.130	0.896	-0.016	0.019
County_Douglas	0.0102	0.014	0.751	0.453	-0.017	0.037
County_Ferry	0.0411	0.041	0.999	0.318	-0.040	0.122
County_Franklin	-0.0226	0.008	-2.933	0.003	-0.038	-0.007
County_Garfield	0.0703	0.064	1.100	0.271	-0.055	0.196
County_Grant	-0.0028	0.008	-0.333	0.739	-0.019	0.014
County_Grays Harbor	-0.0010	0.011	-0.092	0.926	-0.022	0.020
County_Island	-0.0244	0.010	-2.529	0.011	-0.043	-0.005
County_Jefferson	0.0164	0.016	1.057	0.291	-0.014	0.047
County_King	-0.0243	0.004	-5.899	0.000	-0.032	-0.016
County_Kitsap	-0.0212	0.006	-3.527	0.000	-0.033	-0.009
County_Kittitas	-0.0136	0.007	-2.009	0.045	-0.027	-0.000
County_Klickitat	0.0284	0.020	1.409	0.159	-0.011	0.068
County_Lewis	-0.0101	0.008	-1.235	0.217	-0.026	0.006

County_Lincoln	-0.0028	0.019	-0.152	0.880	-0.039	0.034
County_Mason	-0.0098	0.009	-1.043	0.297	-0.028	0.009
County_Okanogan	0.0167	0.016	1.027	0.304	-0.015	0.048
County_Pacific	-0.0078	0.015	-0.529	0.597	-0.036	0.021
County_Pend Oreille	0.0313	0.026	1.192	0.233	-0.020	0.083
County_Pierce	-0.0202	0.004	-4.647	0.000	-0.029	-0.012
County_Skagit	-0.0167	0.006	-2.646	0.008	-0.029	-0.004
County_Skamania	0.0249	0.031	0.808	0.419	-0.036	0.085
County_Snohomish	-0.0282	0.004	-6.773	0.000	-0.036	-0.020
County_Spokane	-0.0224	0.005	-4.544	0.000	-0.032	-0.013
County_Stevens	0.0147	0.016	0.895	0.371	-0.018	0.047
County_Thurston	-0.0263	0.005	-4.864	0.000	-0.037	-0.016
County_Wahkiakum	0.0052	0.037	0.139	0.889	-0.068	0.079
County_Walla Walla	-0.0002	0.013	-0.014	0.989	-0.026	0.025
County_Whatcom	-0.0241	0.006	-3.865	0.000	-0.036	-0.012
County_Whitman	-0.0215	0.012	-1.850	0.064	-0.044	0.001
County_Yakima	-0.0198	0.006	-3.134	0.002	-0.032	-0.007

Weather Condition_Blowing Sand or Dirt or Snow	-0.0262	0.008	-3.470	0.001	-0.041	-0.011
Weather Condition_Clear	0.0206	0.002	8.536	0.000	0.016	0.025
Weather Condition_Fog or Smog or Smoke	0.0335	0.008	4.265	0.000	0.018	0.049
Weather Condition_Other	0.0202	0.014	1.419	0.156	-0.008	0.048
Weather Condition_Overcast	0.0129	0.003	4.821	0.000	0.008	0.018
Weather Condition_Partly Cloudy	-0.0133	0.004	-3.223	0.001	-0.021	-0.005
Weather Condition_Raining	0.0072	0.003	2.723	0.006	0.002	0.012
Weather Condition_Severe Crosswind	-0.0266	0.004	-6.150	0.000	-0.035	-0.018
Weather Condition_Sleet or Hail or Freezing Rain	0.0065	0.007	0.910	0.363	-0.007	0.020
Weather Condition_Snowing	-0.0083	0.003	-2.615	0.009	-0.015	-0.002
Lighting Condition_Dark - Unknown Lighting	0.0064	0.009	0.680	0.497	-0.012	0.025
Lighting Condition_Dark-No Street Lights	0.0078	0.007	1.153	0.249	-0.005	0.021
Lighting Condition_Dark-Street Lights Off	-0.0052	0.011	-0.457	0.648	-0.027	0.017
Lighting Condition_Dark-Street Lights On	0.0043	0.007	0.657	0.511	-0.009	0.017
Lighting Condition_Dawn	-0.0124	0.008	-1.626	0.104	-0.027	0.003
Lighting Condition_Daylight	-0.0112	0.006	-1.764	0.078	-0.024	0.001
Lighting Condition_Dusk	-0.0017	0.008	-0.216	0.829	-0.017	0.014
Lighting Condition_Other	0.0485	0.052	0.932	0.351	-0.053	0.150
Lighting Condition_Unknown	-0.0101	0.016	-0.623	0.533	-0.042	0.022

Immediate Analysis

Of the conditions that were most significant

- Pedestrian and Pedalcyclist involvement were the greatest indicators of a collision being severe
- The more the motor vehicles there were, the more dangerous the accident.
- AADT was negatively correlated accident severity with higher AADT leading to less severe accidents
- Most weather conditions were significant
- Generally light did seem to reduce accident severity
- All the counties that were significant had were negatively correlated so less likely to have a severe injury.
    - Might have to do with the large sample of no injury data

LS with Traffic Fixed

We noticed that a large number of collisions had the same AADT value of 5400 (about 3922 counts which is the most of any AADT level).

To see the effects of other conditions without being influenced by the traffic level, we look at only the collisions with the AADT value of 5400 and did a linear regression on the Injury Severity (Low or High) with the same factors (excluding AADT) value

OLS Regression Results

Dep. Variable:	Injury Severity	R-squared:	0.165
Model:	OLS	Adj. R-squared:	0.152
Method:	Least Squares	F-statistic:	2.729
Date:	Wed, 29 May 2024	Prob (F-statistic):	2.96e-11
Time:	09:41:28	Log-Likelihood:	1959.0
No. Observations:	3922	AIC:	-3794.
Df Residuals:	3860	BIC:	-3405.
Df Model:	61		
Covariance Type:	HC1		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.0041	0.014	-0.300	0.764	-0.031	0.023
School Zone	0.0232	0.015	1.553	0.120	-0.006	0.052
Intersection Related	-0.0016	0.005	-0.304	0.761	-0.012	0.009
Damage Threshold Met	0.0123	0.006	2.226	0.026	0.001	0.023
Hit and Run	-0.0206	0.005	-3.877	0.000	-0.031	-0.010
Motor Vehicles Involved	0.0173	0.007	2.359	0.018	0.003	0.032
Passengers Involved	-0.0012	0.005	-0.246	0.805	-0.011	0.009
Commercial Carrier Involved	0.0315	0.014	2.198	0.028	0.003	0.060
School Bus Involved	-0.0752	0.026	-2.874	0.004	-0.126	-0.024
Pedestrians Involved	0.4521	0.058	7.839	0.000	0.339	0.565
Pedalcyclists Involved	0.1075	0.057	1.903	0.057	-0.003	0.218

County_Adams	-0.0336	0.008	-4.278	0.000	-0.049	-0.018
County_Asotin	-0.0333	0.009	-3.692	0.000	-0.051	-0.016
County_Benton	-0.0345	0.008	-4.545	0.000	-0.049	-0.020
County_Chelan	-0.0262	0.033	-0.784	0.433	-0.092	0.039
County_Clallam	0.2083	0.088	2.364	0.018	0.036	0.381
County_Clark	-0.0245	0.015	-1.624	0.104	-0.054	0.005
County_Columbia	-6.643e-17	3.65e-17	-1.822	0.068	-1.38e-16	5.01e-18
County_Cowlitz	0.1169	0.062	1.873	0.061	-0.005	0.239
County_Douglas	-0.0244	0.007	-3.688	0.000	-0.037	-0.011
County_Ferry	3.558e-17	3.33e-17	1.070	0.285	-2.96e-17	1.01e-16
County_Franklin	0.0570	0.057	0.997	0.319	-0.055	0.169
County_Garfield	-7.904e-17	4.09e-17	-1.933	0.053	-1.59e-16	1.09e-18
County_Grant	-0.0106	0.018	-0.598	0.550	-0.046	0.024
County_Grays Harbor	-0.0322	0.010	-3.224	0.001	-0.052	-0.013
County_Island	-0.0118	0.013	-0.937	0.349	-0.037	0.013
County_Jefferson	0.0888	0.067	1.329	0.184	-0.042	0.220
County_King	-0.0076	0.009	-0.848	0.396	-0.025	0.010
County_Kitsap	-0.0048	0.012	-0.389	0.697	-0.029	0.019
County_Kittitas	0.0979	0.090	1.082	0.279	-0.079	0.275
County_Klickitat	-0.0225	0.009	-2.495	0.013	-0.040	-0.005
County_Lewis	0.0409	0.065	0.627	0.530	-0.087	0.169

County_Lincoln	-0.0277	0.015	-1.853	0.064	-0.057	0.002
County_Mason	-0.0292	0.008	-3.630	0.000	-0.045	-0.013
County_Okanogan	-0.0255	0.007	-3.430	0.001	-0.040	-0.011
County_Pacific	-0.0351	0.008	-4.608	0.000	-0.050	-0.020
County_Pend Oreille	-0.0401	0.009	-4.528	0.000	-0.057	-0.023
County_Pierce	-0.0197	0.010	-1.953	0.051	-0.039	6.85e-05
County_Skagit	-0.0239	0.011	-2.101	0.036	-0.046	-0.002
County_Skamania	-0.0272	0.009	-2.894	0.004	-0.046	-0.009
County_Snohomish	-0.0184	0.008	-2.322	0.020	-0.034	-0.003
County_Spokane	-0.0256	0.010	-2.455	0.014	-0.046	-0.005
County_Stevens	0.0519	0.052	0.990	0.322	-0.051	0.155
County_Thurston	-0.0334	0.007	-5.059	0.000	-0.046	-0.020
County_Wahkiakum	-0.0309	0.028	-1.088	0.276	-0.087	0.025
County_Walla Walla	0.0160	0.043	0.368	0.713	-0.069	0.101
County_Whatcom	-0.0303	0.007	-4.294	0.000	-0.044	-0.016
County_Whitman	-0.0349	0.008	-4.297	0.000	-0.051	-0.019
County_Yakima	-0.0139	0.020	-0.711	0.477	-0.052	0.024

Weather Condition_Blowing Sand or Dirt or Snow	-0.0200	0.019	-1.068	0.285	-0.057	0.017
Weather Condition_Clear	-0.0070	0.011	-0.627	0.531	-0.029	0.015
Weather Condition_Fog or Smog or Smoke	-0.0218	0.023	-0.955	0.340	-0.067	0.023
Weather Condition_Other	0.1310	0.096	1.363	0.173	-0.057	0.319
Weather Condition_Overcast	-0.0034	0.012	-0.292	0.771	-0.026	0.020
Weather Condition_Partly Cloudy	-0.0170	0.014	-1.211	0.226	-0.045	0.011
Weather Condition_Raining	-0.0153	0.012	-1.281	0.200	-0.039	0.008
Weather Condition_Severe Crosswind	-0.0114	0.017	-0.649	0.516	-0.046	0.023
Weather Condition_Sleet or Hail or Freezing Rain	-0.0123	0.011	-1.128	0.259	-0.034	0.009
Weather Condition_Snowing	-0.0268	0.016	-1.705	0.088	-0.058	0.004
Lighting Condition_Dark - Unknown Lighting	0.0015	0.023	0.064	0.949	-0.043	0.046
Lighting Condition_Dark-No Street Lights	0.0273	0.014	2.000	0.045	0.001	0.054
Lighting Condition_Dark-Street Lights Off	-0.0102	0.007	-1.521	0.128	-0.023	0.003
Lighting Condition_Dark-Street Lights On	0.0190	0.007	2.545	0.011	0.004	0.034
Lighting Condition_Dawn	-0.0279	0.011	-2.531	0.011	-0.050	-0.006
Lighting Condition_Daylight	-0.0021	0.005	-0.392	0.695	-0.013	0.008
Lighting Condition_Dusk	-0.0067	0.015	-0.439	0.661	-0.037	0.023
Lighting Condition_Other	-0.0185	0.008	-2.329	0.020	-0.034	-0.003
Lighting Condition_Unknown	0.0136	0.013	1.084	0.278	-0.011	0.038

Observations holding AADT constant

- Pedestrians and Pedalcyclist involvement generally lead to more severe accidents
- Passenger involvement is now negative and not significant
- More motor vehicles lead to more severe accidents
- Now lighting conditions seem to be far more significant than before
- Weather conditions are less significant.
- Most significant counties are negatively correlated with accidents (save for Clallam and Cowlitz counties which are positive)

Conclusion

1. Inconclusive results on county level due to higher traffic in specific regions
    a. Random sampling on state road level shows higher severity at lower AADT levels and vice versa
    b. Specific roads do show higher levels of severity
2. Key predictors of car accidents:
    a. AADT and Primary Trafficway w/o sampling due to large numbers
    b. Lighting conditions, weather conditions, and other motor vehicles involved are significant to car accidents
3. It is possible to develop a predictive model however:
    a. It is prone to overfitting due to large amounts of low severity injury
    b. An alternative can be developed with less accuracy but it is able to predict other types of collisions albeit with low recall.

Thank you so much! Questions?

Appendix

We can run this analysis by running this code which will generate the visualization and report for this data:

```
python main.py
```