Synopsis

According to the analysis from this article below based on the National Highway Traffic Safety Administration (NHTSA) crash data. New York city placed number 7 among the states as one of the deadliest states for driving. New York City has over 1000 fatalities each year relating to accidents on the roads, which is 1.6 times the national average. However, the number of deaths decreased by approximately 22% from 2010 to 2019. The article also noted that the highest risk age group involved in car accidents in New York is 21- to 25-year-olds, and men make up a total of 71% of New York's car accident deaths. Most of New York's car accident deaths are pedestrians, which is about one third of accidents. About 23% of all car accident deaths happen there each year. Speeding and driving while under the influence are some of the top driver mistakes that lead to fatal accidents in New York, making up 27% of all fatal car accidents. Motorcycles were involved in about 15% of New York's fatal vehicle crashes in 2019, and single-vehicle crashes cause 64% of all road deaths, which is twice as many as multi-car crashes.

Website link
https://www.finder.com/new-york-car-accident-statistics

This project's purpose is to investigate and analyze the impact of motor collisions In New York City on several factors such as the time during the day the collisions are happening, the type of vehicles that cause the most crashes, and the contributing factors. As several safety concerns arise about motor collisions, this project seeks to identify the contributing factors and ultimately come up with a solution.

This data concerns people who live in New York, especially people who drive cars, motorcycles, ride bikes, and pedestrians who are walking in the streets of New York in different boroughs. This project addresses the problem involving motor collisions in traffic in the city of New York such as vehicles that are likely to be involved in a collision and the contributing factors. The potential sources for this project were data.gov, and the Department of Transportation. By analyzing this data, we can find the number of collisions per year, the most common types of vehicles involved in a crash, and gain insight into the different characteristics and contributing factors of these collisions in New York City.

After analyzing the data, passenger vehicles were most likely to be involved in a collision. I find several ways we can reduce collisions involving passenger vehicles. One of the ways we can reduce collisions is by launching an education and awareness campaign aimed at drivers and pedestrians to learn about safe driving and give them tips to avoid collisions. The campaigns could also be targeted specifically at midday drivers and pedestrians, highlighting the risks associated with this time of day since most collisions happen during this time. And collision avoidance systems could be installed in vehicles to warn drivers of potential collisions which can help them avoid accidents. And finally, enforcement measures could be put in place to help reduce collisions involving passenger vehicles. That could include increased police presence on the roads during midday, as well as stricter penalties for drivers who violate traffic laws.

Data
https://www.finder.com/new-york-car-accident-statistics

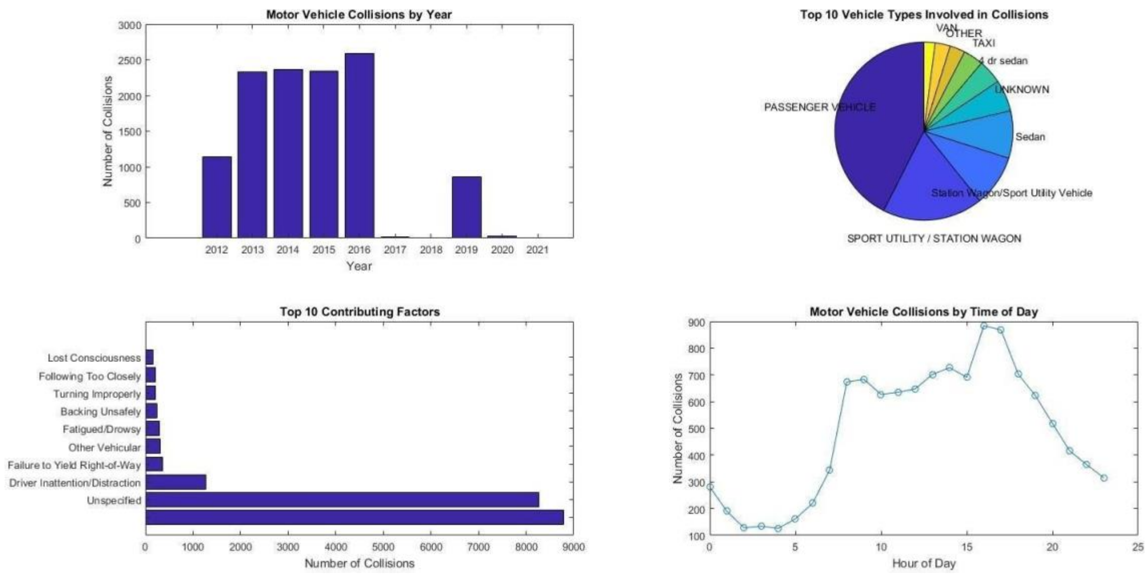
```matlab
% Load CSV file into a table
data = readtable('motor_data.csv');

% Plot number of collisions by year
years = unique(year(data.CRASH_DATE));
num_collisions = zeros(size(years));
for i = 1:length(years)
    num_collisions(i) = sum(year(data.CRASH_DATE) == years(i));
end
subplot(2, 2, 1)
bar(years, num_collisions)
xlabel('Year')
ylabel('Number of Collisions')
title('Motor Vehicle Collisions by Year')

% Identify most common vehicle types
vehicle_types = unique(data.VEHICLE_TYPE);
num_vehicles = zeros(size(vehicle_types));
for i = 1:length(vehicle_types)
    num_vehicles(i) = sum(strcmp(data.VEHICLE_TYPE, vehicle_types(i)));
end
[~, idx] = sort(num_vehicles, 'descend');
top_vehicle_types = vehicle_types(idx(1:10));
top_num_vehicles = num_vehicles(idx(1:10));
subplot(2, 2, 2)
pie(top_num_vehicles, top_vehicle_types)
title('Top 10 Vehicle Types Involved in Collisions')

% Identify contributing factors
factors = unique([data.CONTRIBUTING_FACTOR_1; data.CONTRIBUTING_FACTOR_2]);
factor_counts = zeros(size(factors));
for i = 1:length(factors)
    factor_counts(i) = sum(strcmp(data.CONTRIBUTING_FACTOR_1, factors(i)) |
strcmp(data.CONTRIBUTING_FACTOR_2, factors(i)));
end
[~, idx] = sort(factor_counts, 'descend');
top_factors = factors(idx(1:10));
top_factor_counts = factor_counts(idx(1:10));
subplot(2, 2, 3)
barh(top_factor_counts)
yticklabels(top_factors)
xlabel('Number of Collisions')
title('Top 10 Contributing Factors')

% Plot number of collisions by time of day
hours = unique(hour(data.CRASH_TIME));
num_collisions = zeros(size(hours));
for i = 1:length(hours)
    num_collisions(i) = sum(hour(data.CRASH_TIME) == hours(i));
end
subplot(2, 2, 4)
```

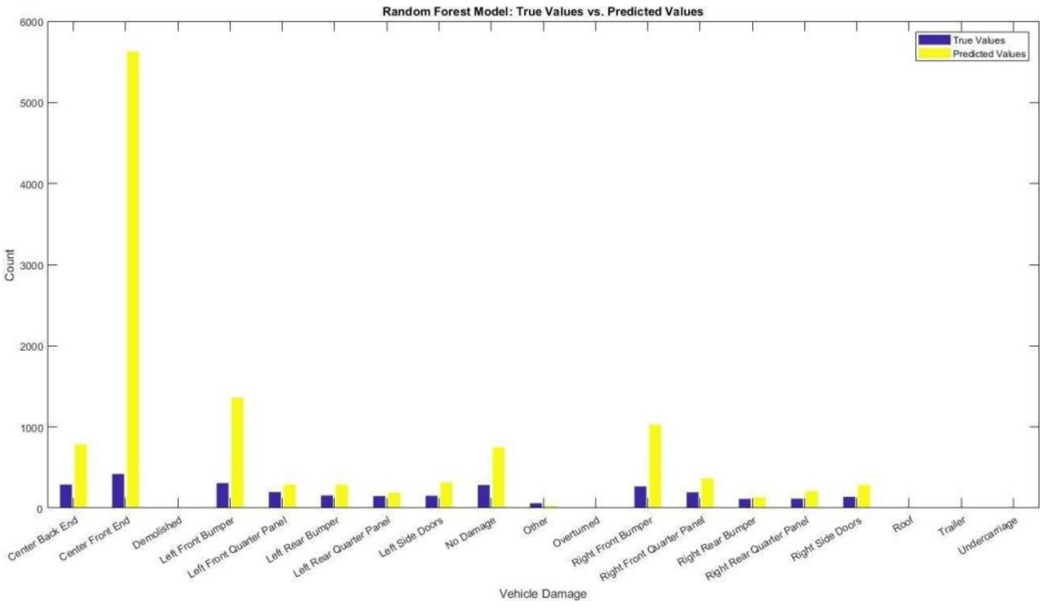
```
plot(hours, num_collisions, '-o')
xlabel('Hour of Day')
ylabel('Number of Collisions')
title('Motor Vehicle Collisions by Time of Day')


% Load the data from CSV file
data = readtable('motor_data.csv');
% Select the relevant columns
X = data.VEHICLE_YEAR;
y = data.VEHICLE_OCCUPANTS;

% Split the data into training and testing sets
train_ratio = 0.8; % Use 80% of the data for training
train_size = round(train_ratio * height(data));
train_idx = randperm(height(data), train_size);
test_idx = setdiff(1:height(data), train_idx);
X_train = X(train_idx);
y_train = y(train_idx);
X_test = X(test_idx);
y_test = y(test_idx);

% Fit the linear regression model
mdl = fitlm(X_train, y_train);

% Print the model summary
disp(mdl)

% Visualize the data and the regression line
figure;
subplot(2,2,1)
scatter(X_train, y_train, 'filled')
xlabel('Vehicle Year')
ylabel('Number of Occupants')
title('Training Data')
hold on
plot(X_train, predict(mdl, X_train), '-r')

subplot(2,2,2)
scatter(X_test, y_test, 'filled')
xlabel('Vehicle Year')
ylabel('Number of Occupants')
title('Testing Data')
hold on
plot(X_test, predict(mdl, X_test), '-r')

subplot(2,2,[3,4])
scatter(X_train, y_train, 'filled')
hold on
scatter(X_test, y_test, 'filled')
xlabel('Vehicle Year')
ylabel('Number of Occupants')
title('All Data')
hold on
plot(X_train, predict(mdl, X_train), '-r')
```

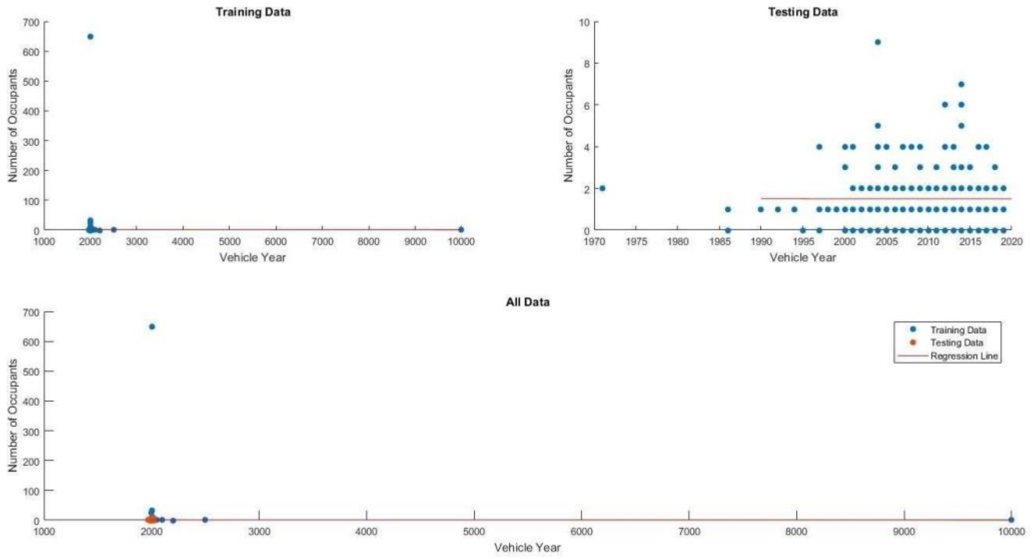
```matlab
plot(X_test, predict(mdl, X_test), '-r')
legend('Training Data', 'Testing Data', 'Regression Line')

% Convert categorical variables if necessary
data.CRASH_TIME = categorical(data.CRASH_TIME);
data.VEHICLE_TYPE = categorical(data.VEHICLE_TYPE);
data.VEHICLE_DAMAGE = categorical(data.VEHICLE_DAMAGE);

% Predictor and response names
predictorNames = {'CRASH_TIME', 'VEHICLE_TYPE'};
responseName = 'VEHICLE_DAMAGE';

% Fit decision tree model
treeModel = fitctree(data, responseName, 'PredictorNames', predictorNames);

% Fit random forest model
rfModel = TreeBagger(100, data(:, predictorNames), data.(responseName),
'Method', 'classification', 'OOBPrediction', 'on');

% Predict the vehicle damage based on the decision tree model
predDamageTree = predict(treeModel, data);

% Predict the vehicle damage based on the random forest model
predDamageRF = predict(rfModel, data);

% Display confusion matrix for decision tree model
confusionMatrix = confusionmat(data.(responseName), predDamageTree);
disp('Confusion Matrix for Decision Tree Model:');
disp(confusionMatrix);

% Get unique categories and their counts for true and predicted values
[uniqueCategoriesTrue, ~, idTrue] = unique(data.(responseName));
trueCounts = accumarray(idTrue, 1);
[uniqueCategoriesPred, ~, idPred] = unique(categorical(predDamageRF));
predCounts = accumarray(idPred, 1);

% Combine true and predicted unique categories and counts
allCategories = union(uniqueCategoriesTrue, uniqueCategoriesPred);
countsTrue = zeros(length(allCategories), 1);
countsPred = zeros(length(allCategories), 1);

for i = 1:length(allCategories)
    indexTrue = find(uniqueCategoriesTrue == allCategories(i));
    if ~isempty(indexTrue)
        countsTrue(i) = trueCounts(indexTrue);
    end

    indexPred = find(uniqueCategoriesPred == allCategories(i));
    if ~isempty(indexPred)
        countsPred(i) = predCounts(indexPred);
    end
end

counts = [countsTrue countsPred];
```

```
% Plot a grouped bar chart comparing true values and predicted values for the
random forest model
figure;
bar(categorical(allCategories), counts, 'grouped');
legend('True Values', 'Predicted Values');
xlabel('Vehicle Damage');
ylabel('Count');
title('Random Forest Model: True Values vs. Predicted Values');
```

Outputs graph.







I found that passenger vehicles are more likely to be involved in collisions than any other type of vehicle. I also found data where the vehicles are more likely to hit in a collision which has true value versus predicted values. I was looking for some data in the data sets but there wasn't any such thing, such as the weather conditions at the time of the collisions and the demographics of the drivers who were involved in a crash, that could help with a more effective solution.

Dear Professor, Attaway, or Grader

Mohamed Camara did not contribute anything to this project. As a result, the burden of the whole project fell entirely on me. I want to make it clear that Mohamed Camara did not contribute anything to this project and should not receive any credit for it. Thank you for your understanding.