Contributing Factors To Car Accidents In The US

Aaron Chen, Rita Chen, Tiffany Chen

Significance/Goal

- ➢ Determine relationship among factors and how they affect the number of car crashes, severity (scale from 1- 4 on its impact on traffic), and distance (length of the road extent affected)
- ➢ Create multiple predictive models for car crashes based on significant factors
- ➢ Create visualizations of these models through graphs and charts

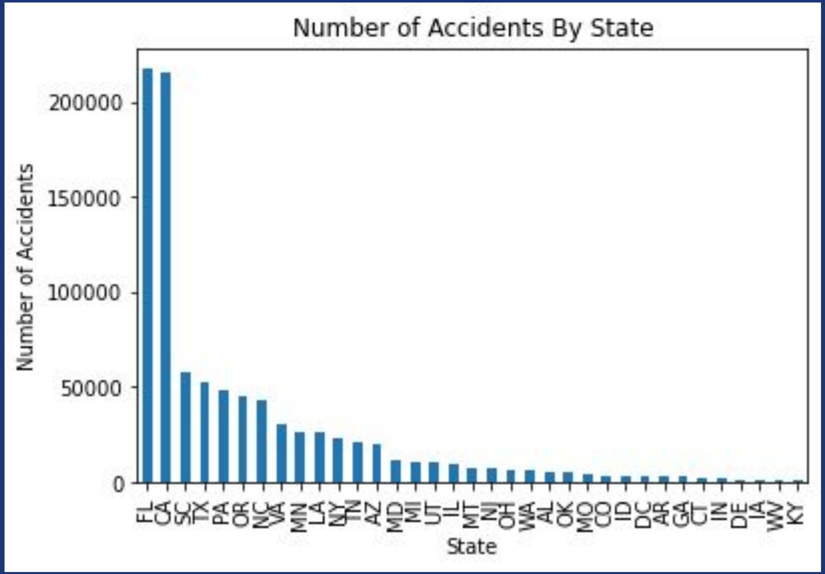
Dataset (~930,000 sample size)

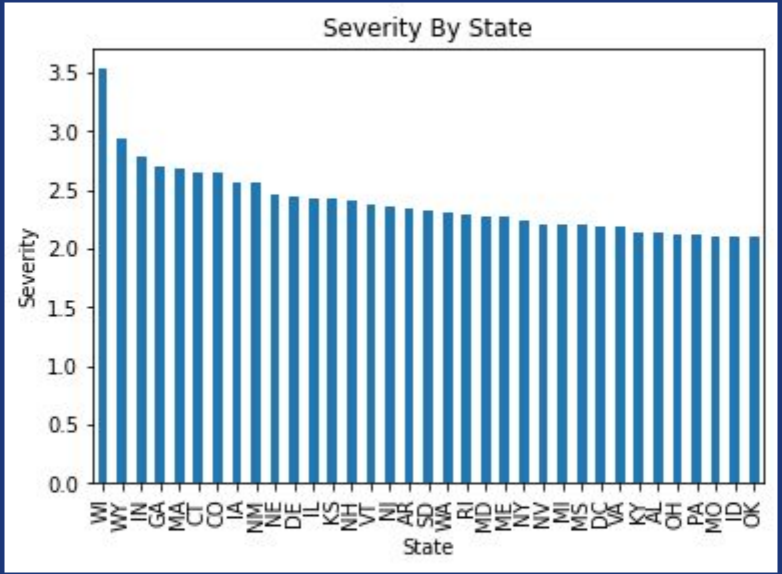
A countrywide accident dataset, with data from Feb 2016 to the end of Dec 2021.

# Severity	Start_Time	End_Time	# Start_Lat	# Start_Lng	# End_Lat	# End_Lng	# Distance(mi)
Shows the severity of the accident, a number between 1 and 4, where 1 indicates the least impact on traffic (i.e., short delay	Shows start time of the accident in local time zone.	Shows end time of the accident in local time zone. End time here refers to when the impact of accident on traffic flow	Shows latitude in GPS coordinate of the start point.	Shows longitude in GPS coordinate of the start point.	Shows latitude in GPS coordinate of the end point.	Shows longitude in GPS coordinate of the end point.	The length of the road extent affected by the accident.
1 – 4	14Jan16 – 31Dec21	8Feb16 – 31Dec21	24.6 – 49	-125 – -67.1	24.6 – 49.1	-125 – -67.1	0 – 155
3	2016-02-08 00:37:08	2016-02-08 06:37:08	40.108909999999995	-83.09286	40.11206	-83.03187	3.23
2	2016-02-08 05:56:20	2016-02-08 11:56:20	39.86542	-84.0628	39.86501	-84.04873	0.747
2	2016-02-08 06:15:39	2016-02-08 12:15:39	39.10266	-84.52468	39.102090000000004	-84.52396	0.055
2	2016-02-08 06:51:45	2016-02-08 12:51:45	41.062129999999996	-81.53784	41.06217	-81.53546999999998	0.12300000000000001
3	2016-02-08 07:53:43	2016-02-08 13:53:43	39.172393	-84.49279200000002	39.170476	-84.501798	0.5

Variables: Severity, Start_Time, Distance(mi), State, Bump, Crossing, Sunrise_Sunset, etc.

Exploratory Visualizations

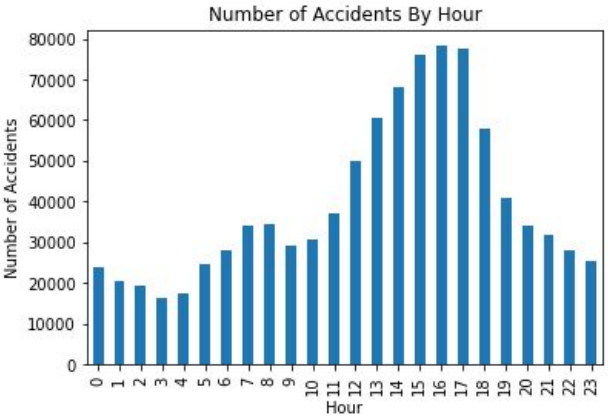


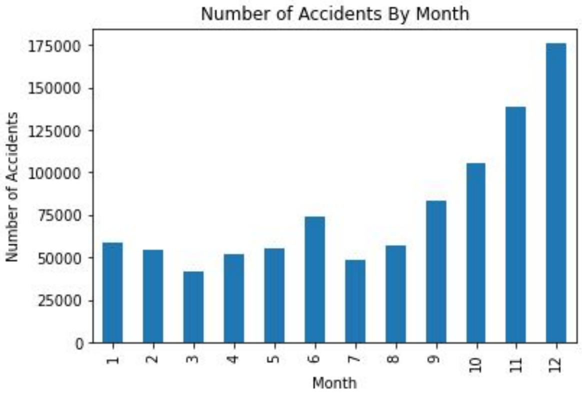


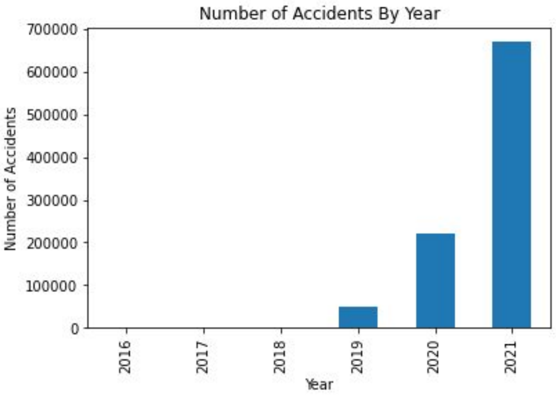
Top 35 states: large gap between top two states and the rest ~150,000

Top 35 states: severity ranges between 2 and 3.5

Time vs Number of Accidents

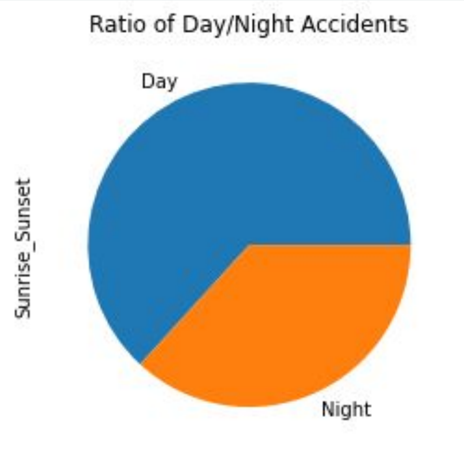






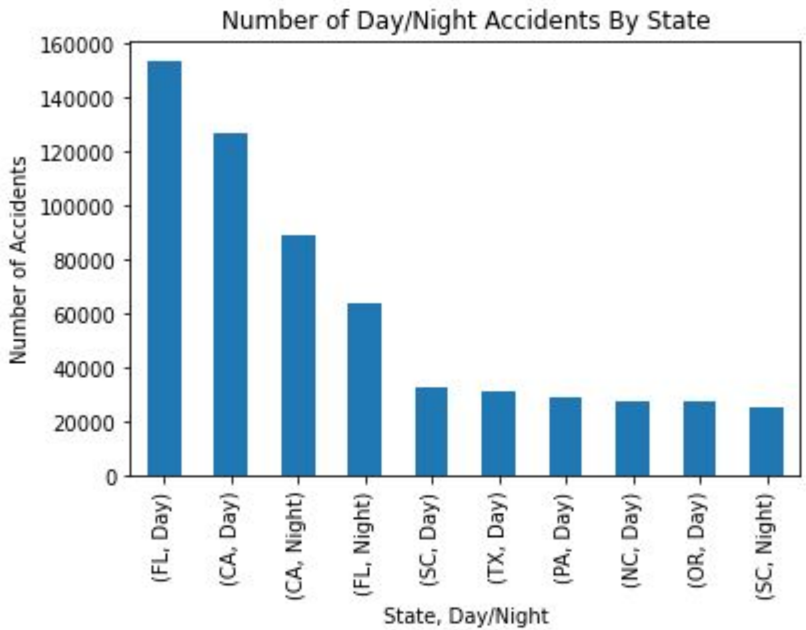
- ➢ Most accidents appear to occur from noon to 6 PM and between October and December
- ➢ Visible increase in the number of accidents per year (may be due to differences in number of samples/records)

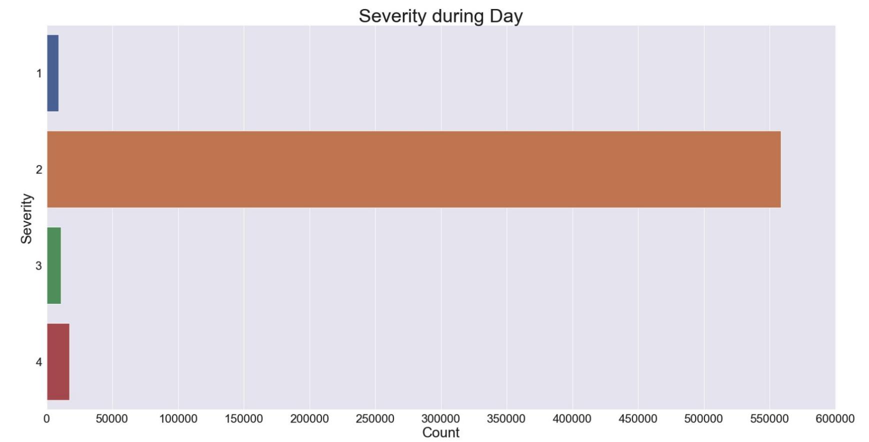
Day vs Night

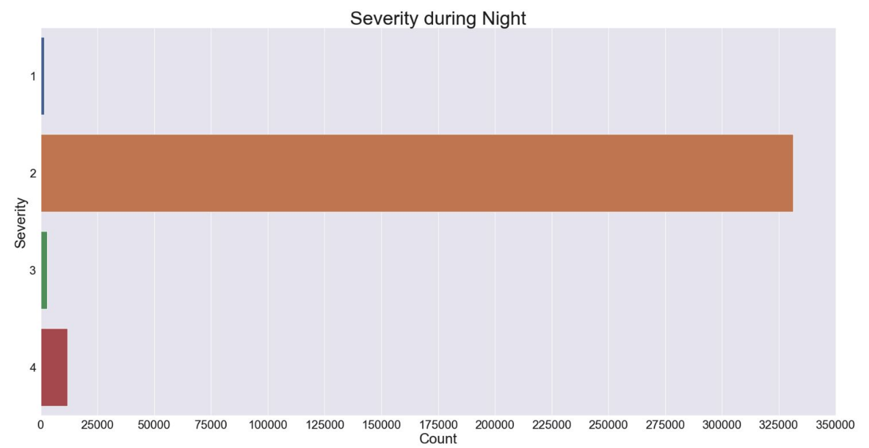


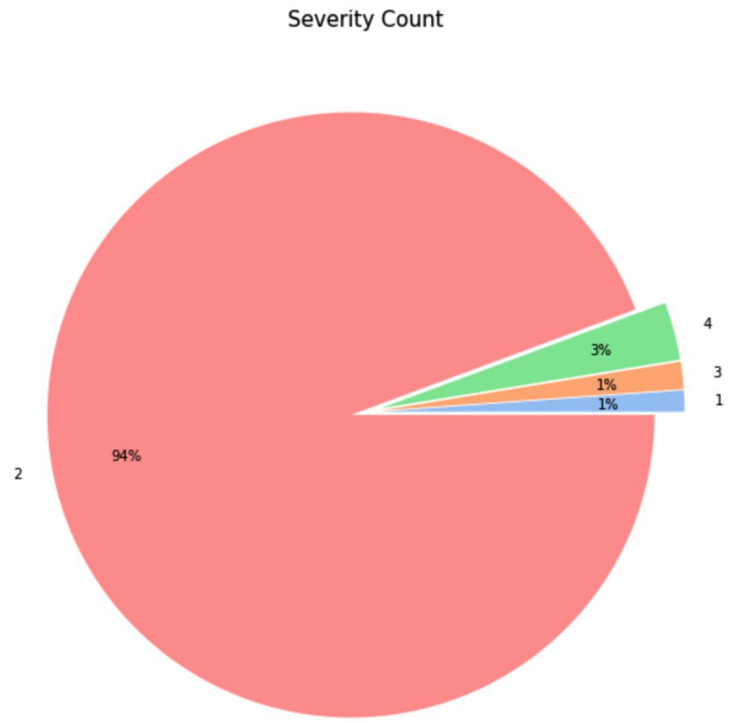
Large proportion of accidents are during the day

Florida and California have the most accidents during the day and night, as expected. Majority of other states are also during the day.



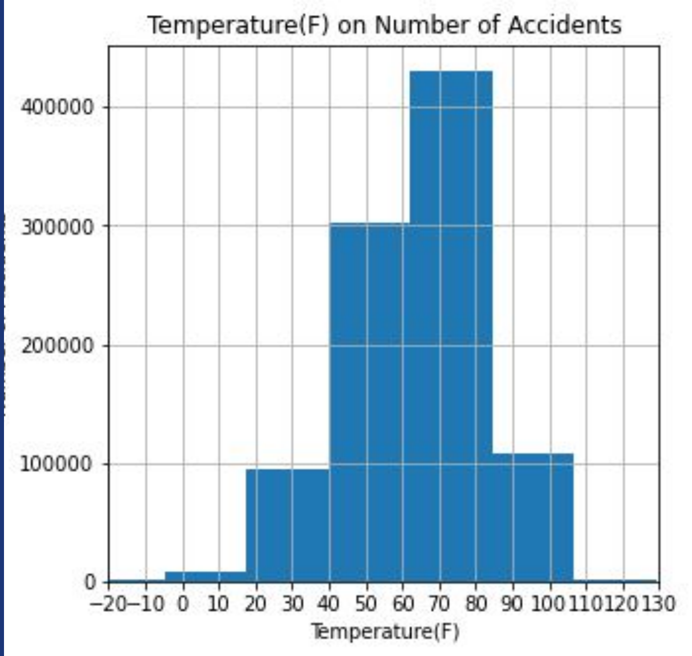






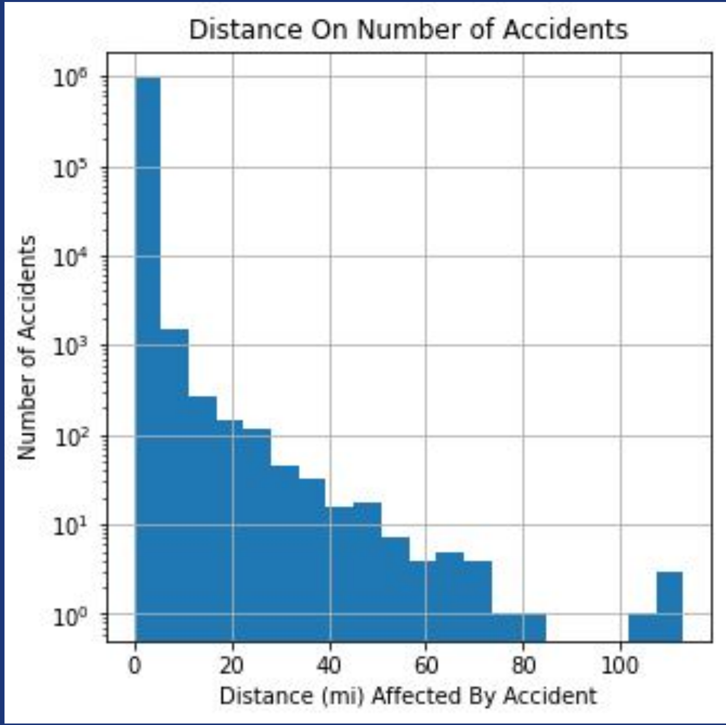
Severity of 2 was extremely common

Histogram of Temperature(F) On Number of Accidents



Majority of the number of accidents occurred around ~40F to ~80F

Distance(mi) Affected by Accident



Most accidents affect up to 20 miles of road

Attempting To Predict Categorical Values

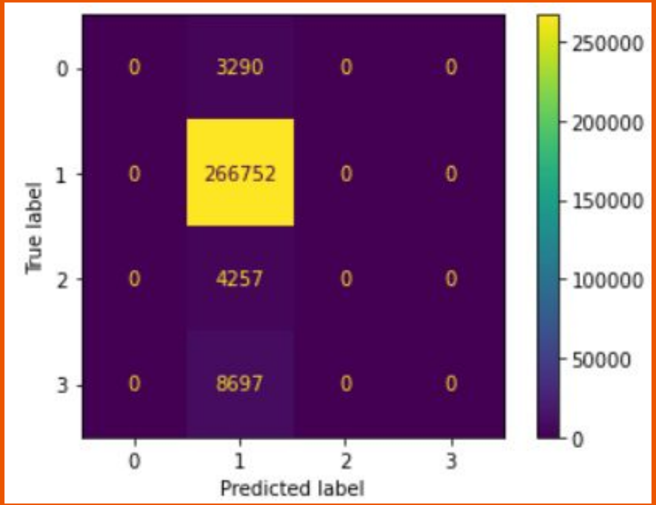
```
                            OLS Regression Results
==============================================================================
Dep. Variable:               Severity   R-squared:                       0.002
Model:                            OLS   Adj. R-squared:                  0.002
Method:                 Least Squares   F-statistic:                     1085.
Date:                Sat, 03 Dec 2022   Prob (F-statistic):               0.00
Time:                        16:36:20   Log-Likelihood:            -4.2622e+05
No. Observations:              943318   AIC:                         8.524e+05
Df Residuals:                  943315   BIC:                         8.525e+05
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          2.1063      0.001   1675.889      0.000       2.104       2.109
Hour       -9.644e-05   6.73e-05     -1.433      0.152      -0.000    3.55e-05
Month         -0.0051      0.000    -46.374      0.000      -0.005      -0.005
==============================================================================
Omnibus:                   770363.124   Durbin-Watson:                   1.380
Prob(Omnibus):                  0.000   Jarque-Bera (JB):         16247753.636
Skew:                           3.975   Prob(JB):                         0.00
Kurtosis:                      21.713   Cond. No.                         52.4
==============================================================================
```

```python
x = df[['Hour','Month']]
y = df[["Severity"]]

regr = linear_model.LinearRegression()
regr.fit(x, y)

print('Intercept: \n', regr.intercept_)
print('Coefficients: \n', regr.coef_)

# with statsmodels
x = sm.add_constant(x) # adding a constant

model = sm.OLS(y, x).fit()
predictions = model.predict(x)

print_model = model.summary()
print(print_model)
```

Linear regression doesn't work well with predicting categorical values (In this case, we tried to predict Severity, a scale from 1-4 using the hour and month of which the car accident happened)

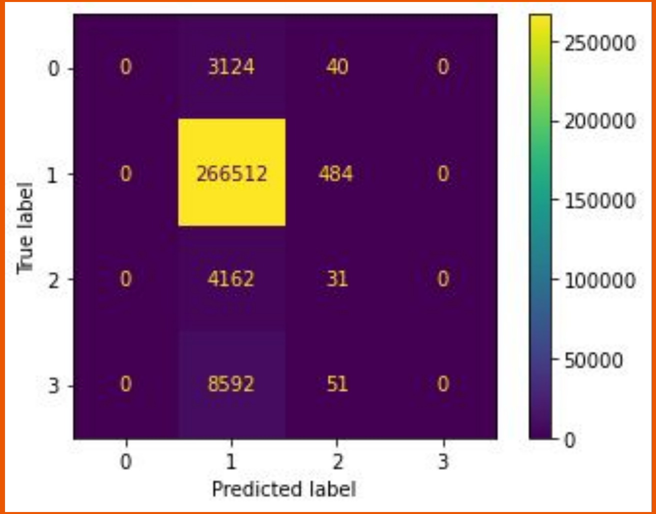
Classifying Severity Using Hour and Month

Logistic Regression:
Accuracy Score = 0.9425998954048821



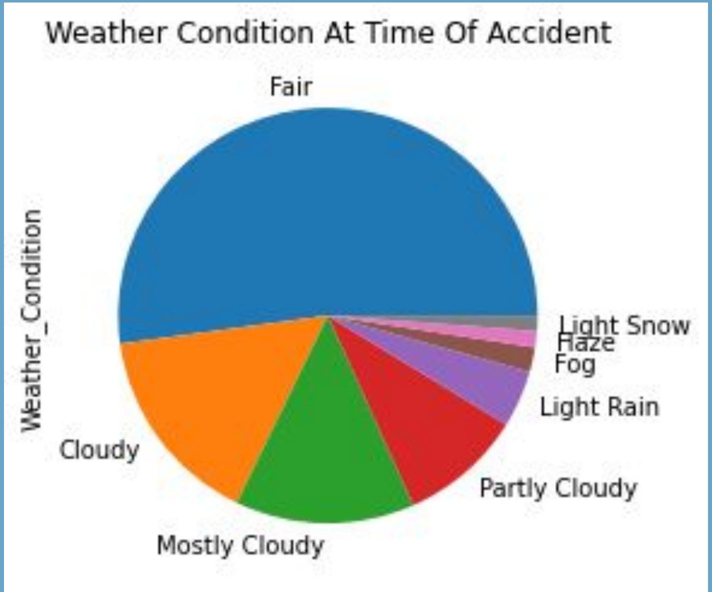
Extreme Problem of overfitting: Almost all of accidents were labeled as a severity of 2

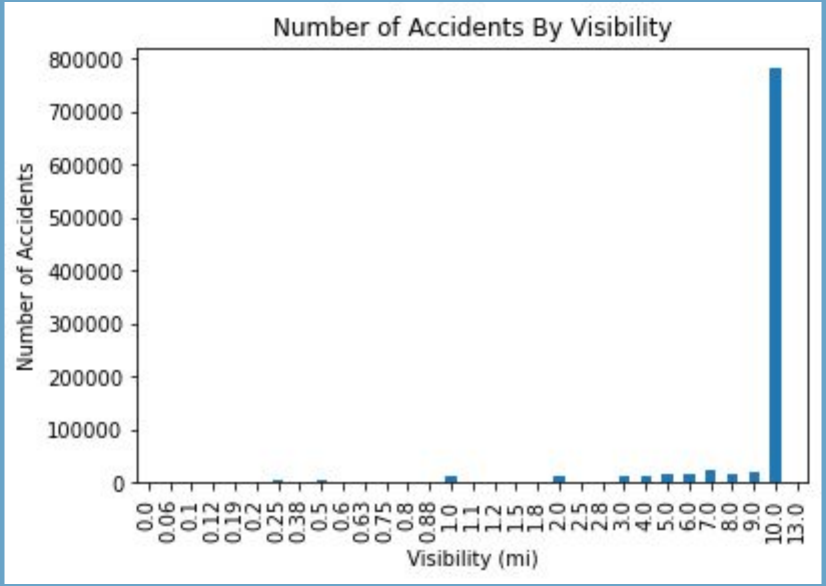
KNeighbors:
Accuracy score = 0.9418613690652872



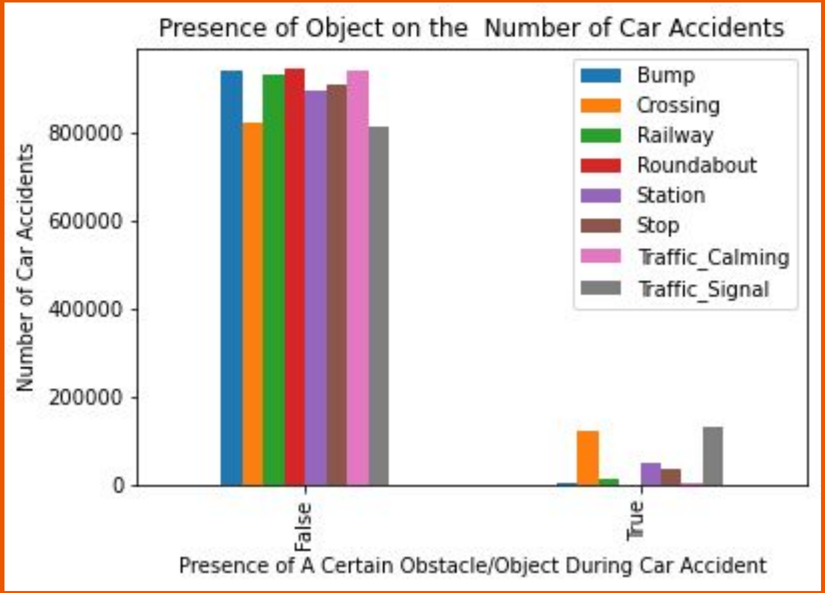
Note: True Label 0 = Severity 1; True Label 1 = Severity 2; True Label 2 = Severity 3; True Label 3 = Severity 4

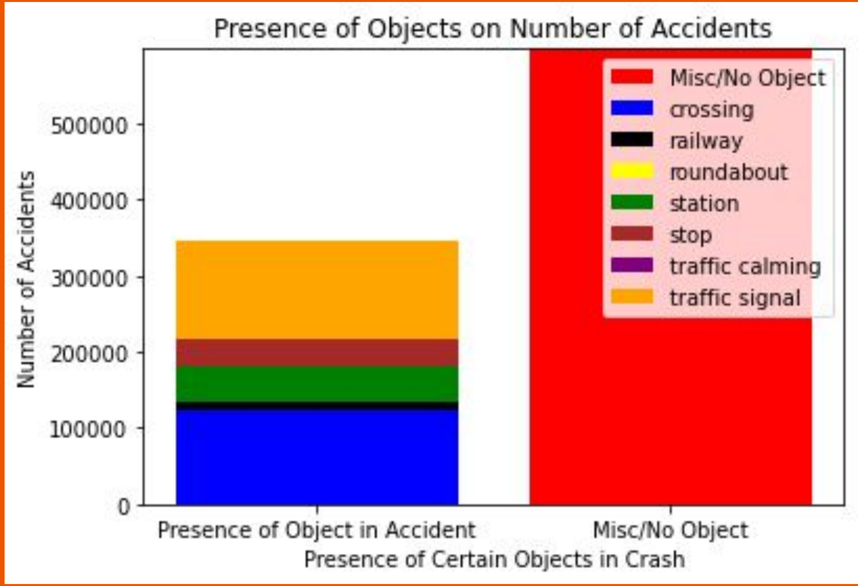
Majority of accidents occur during “good weather” and when visibility(mi) is high





Presence of Object vs Number of Accidents

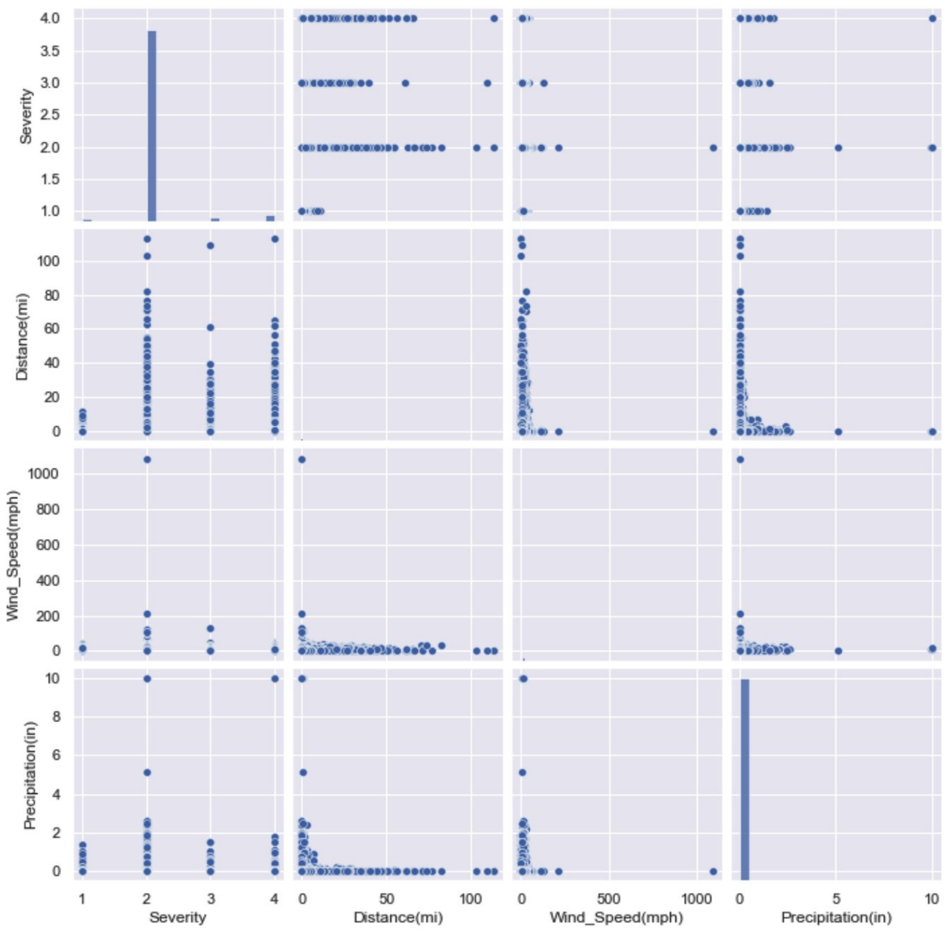




Note: The Presence of one object and another is mutually exclusive

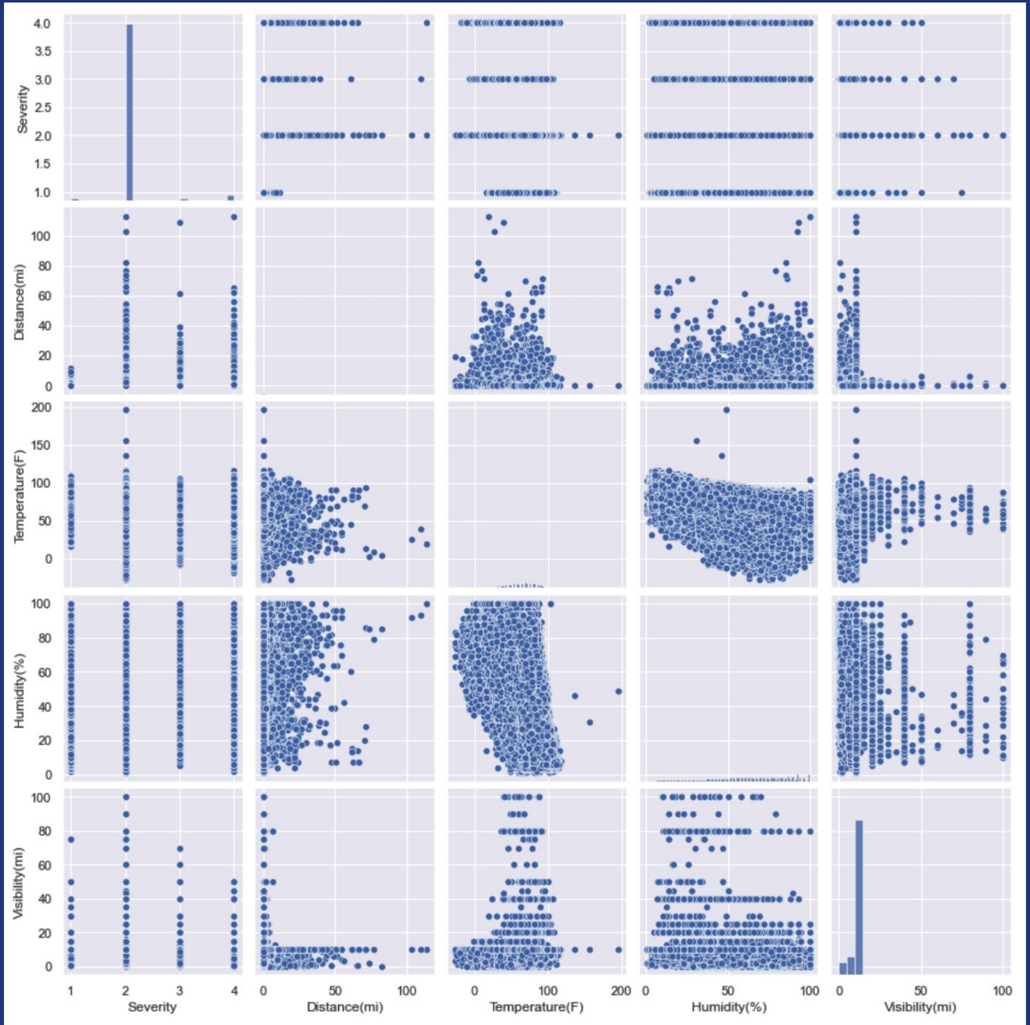
Pairwise scatterplots

- Severity
- Distance (mi)
- Wind Speed (mph)
- Precipitation (in)



Pairwise scatter plots

- Severity
- Distance(mi)
- Temperature(F)
- Humidity(%)
- Visibility(mi)



Linear Regression & Classification Modeling

- Train: 70%
- Test: 30%
- Want to predict Distance and Severity
- Regression based on Weather features
    - Temperature
    - Humidity
    - Visibility
    - Wind Speed
    - Precipitation

Distance Regression

Continuous y variable

Regression Models:

- Linear
- Ridge
- Lasso
- Elastic

All models had a very low score

- About 0.005 for all

```
model = linear_model.LinearRegression()
model.fit(X_train, y_train)
model.score(X_train, y_train), model.score(X_test, y_test)
```

```
(0.005237966826281859, 0.005580352350720741)
```

```
Ridge_reg = linear_model.RidgeCV(cv=2)
Ridge_reg.fit(X_train, y_train)
Ridge_reg.score(X_train, y_train), Ridge_reg.score(X_test, y_test)
```

```
(0.005237966825868l9, 0.005580351053086963)
```

```
Lasso_reg = linear_model.LassoCV(cv=2)
Lasso_reg.fit(X_train, y_train)
Lasso_reg.score(X_train, y_train), Lasso_reg.score(X_test, y_test)
```

```
(0.005237624820073172, 0.005579344253808682)
```

```
Elastic_reg = linear_model.ElasticNetCV(cv=2)
Elastic_reg.fit(X_train, y_train)
Elastic_reg.score(X_train, y_train), Elastic_reg.score(X_test, y_test)
```

```
(0.005237620613651184, 0.0055793294625275935)
```

Severity Classification

- Categorical y variable
- Classification Models
  - Logistic
  - Decision Tree
  - KNeighbors
  - Random Forest
- Regression based on Weather features
  - Temperature
  - Humidity
  - Visibility
  - Wind Speed
  - Precipitation

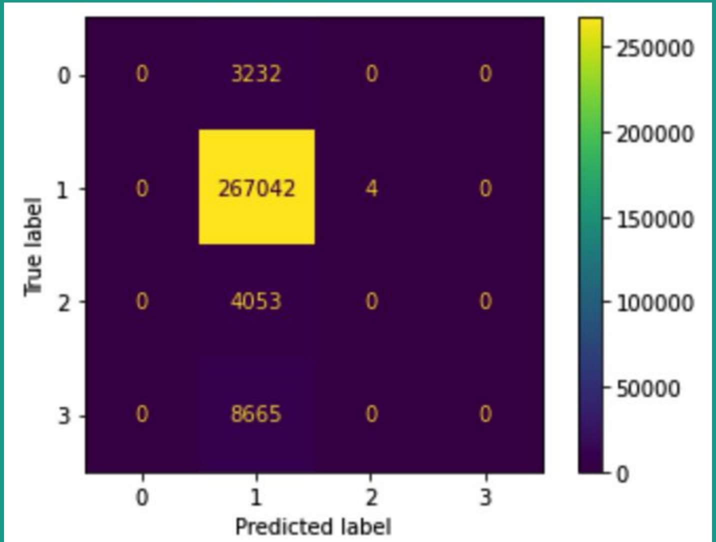
Logistic Classification

Classification Report

	precision	recall	f1-score	support
1	0.00	0.00	0.00	3232
2	0.94	1.00	0.97	267046
3	0.00	0.00	0.00	4053
4	0.00	0.00	0.00	8665
accuracy			0.94	282996
macro avg	0.24	0.25	0.24	282996
weighted avg	0.89	0.94	0.92	282996

Accuracy score = 0.9436246448713056

Confusion Matrix



- Model had good accuracy, but overfitted
- Severity 2 was predicted majority of the time
- Only 4 samples were not classified as Severity 2

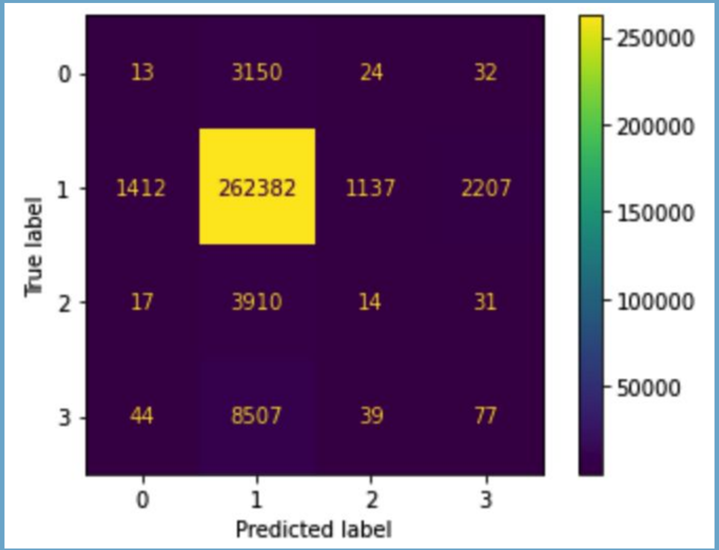
Decision Tree Classification

Classification Report

	precision	recall	f1-score	support
1	0.01	0.00	0.01	3219
2	0.94	0.98	0.96	267138
3	0.01	0.00	0.01	3972
4	0.03	0.01	0.01	8667
accuracy			0.93	282996
macro avg	0.25	0.25	0.25	282996
weighted avg	0.89	0.93	0.91	282996

Accuracy score = 0.9275254773919066

Confusion Matrix



- Model had slightly less accuracy, still overfitted
- Severity 2 still predicted majority of the time, but somewhat less
- Other severities now have precision and recall greater than 0

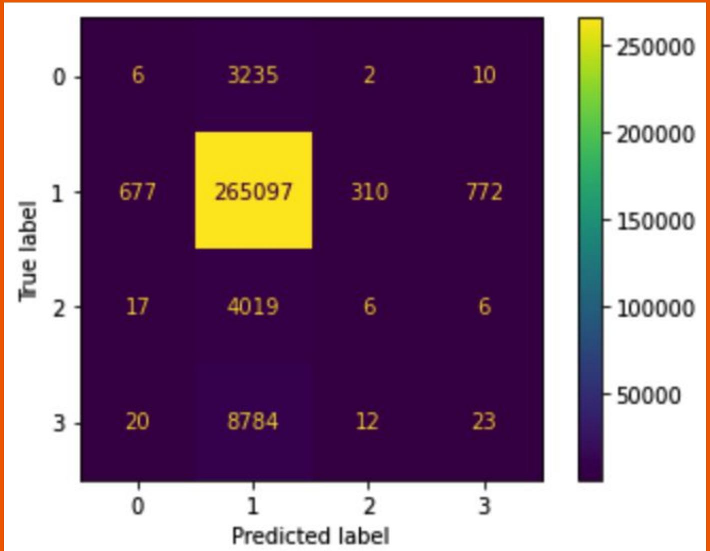
KNeighbors Classification

Classification Report

	precision	recall	f1-score	support
1	0.01	0.00	0.00	3253
2	0.94	0.99	0.97	266856
3	0.02	0.00	0.00	4048
4	0.03	0.00	0.00	8839
accuracy			0.94	282996
macro avg	0.25	0.25	0.24	282996
weighted avg	0.89	0.94	0.91	282996

Accuracy score = 0.9368754328683091

Confusion Matrix



- Model had slightly better accuracy, overfitting increased
- Severity 2 was predicted more often again
- Other severities decreased in their precision and recall

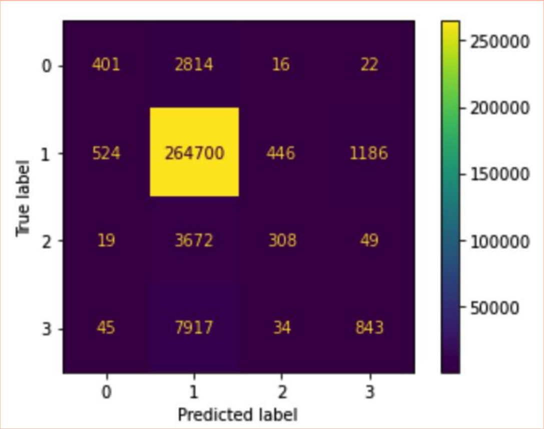
Random Forest Classification

n_estimators = 10

	precision	recall	f1-score	support
1	0.41	0.12	0.19	3253
2	0.95	0.99	0.97	266856
3	0.38	0.08	0.13	4048
4	0.40	0.10	0.15	8839
accuracy			0.94	282996
macro avg	0.53	0.32	0.36	282996
weighted avg	0.92	0.94	0.92	282996

Accuracy score = 0.9408330859800138

Classification Report



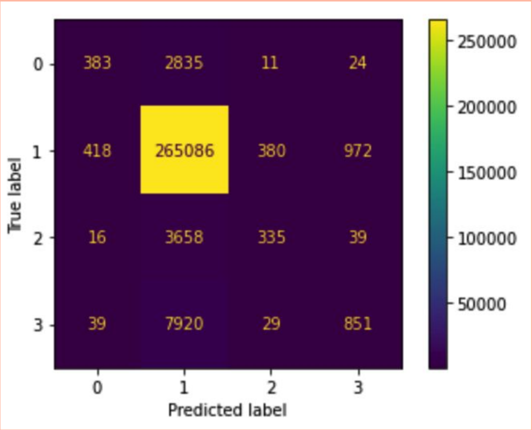
Confusion Matrix

n_estimators = 100

	precision	recall	f1-score	support
1	0.45	0.12	0.19	3253
2	0.95	0.99	0.97	266856
3	0.44	0.08	0.14	4048
4	0.45	0.10	0.16	8839
accuracy			0.94	282996
macro avg	0.57	0.32	0.36	282996
weighted avg	0.92	0.94	0.92	282996

Accuracy score = 0.9422571343764576

Classification Report



Confusion Matrix

- Using 100 trees was only slightly better than 10
- Difference was 0.0.00142404839
- Take into account computational cost
- Other severities increased in their precision and recall

Conclusions

- <u>**Many accidents occurred…**</u>
  - In **FL and CA** both during the **day and night** (in general, majority of accidents were during the day)
  - From **noon to 6 PM** and between **October and December**
  - During "**good weather" with high visibility** (10 miles)
  - Between **40℉ and 80℉**
  - With a **severity of 2**

Remarks

- Our classification data could be susceptible to overfitting/bias when classifying the expected severity
- There also could've been unequal sampling from all the states in the US (ex: a few hundred samples from in some states and tens of thousands of samples from other states)
- Some of the descriptions of our data were vague such as the scale of severity being unclear, 1 indicates the least impact on traffic (i.e., short delay as a result of the accident) and 4 indicates a significant impact on traffic (i.e., long delay) but short/long isn't measurable
- Variables like population size, traffic/road density, rush hours, holidays, state proximity, etc. are not included

References

Data:

Moosavi, S. (no date) US Accidents : A Countrywide Traffic Accident Dataset (2016 - 2021), Kaggle. Kaggle. Available at: https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents (Accessed: November 18, 2022).