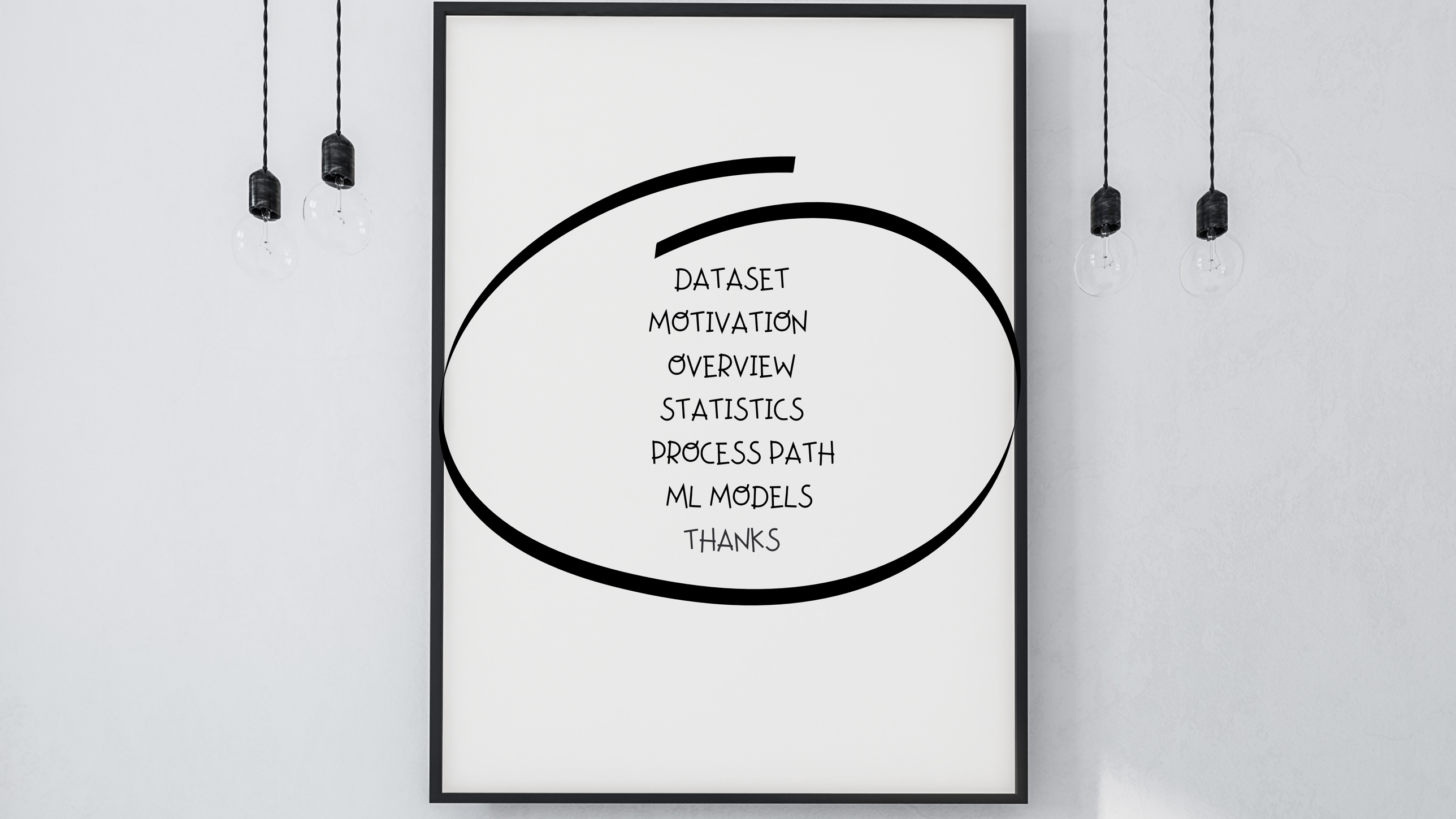
KODLUYORUZ APPLIED DATA SCIENCE 101 BOOTCAMP
CAPSTONE PROJECT

TEAM I-5NORTH

METİN KEREM ÜRKMEZ
ELİF YÜZER
FURKAN BALIKÇI
BORA ÇOLAKOĞLU

NOVEMBER 28, 2020


DATASET
MOTIVATION
OVERVIEW
STATISTICS
PROCESS PATH
ML MODELS
THANKS

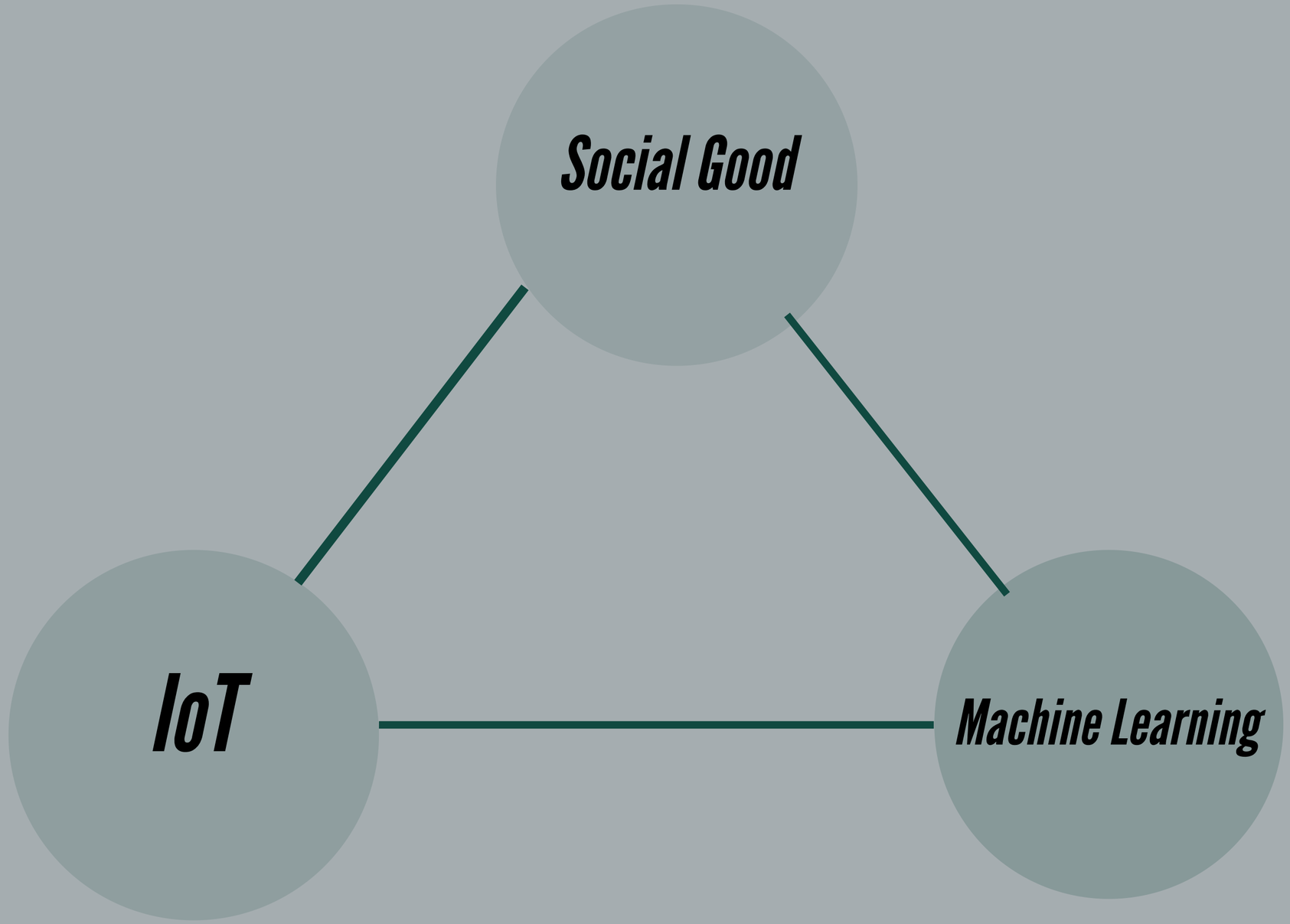
SEVERITY PREDICTION OF USA ACCIDENTS

US Accidents (3.5 million records)

This is a countrywide car accident dataset, which covers 49 states of the USA. The accident data are collected from February 2016 to June 2020, using two APIs that provide streaming traffic incident (or event) data. These APIs broadcast traffic data captured by a variety of entities, such as the US and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road-networks. Currently, there are about 3.5 million accident records in this dataset. Check here to learn more about this dataset.

This dataset has been collected in real-time, using multiple Traffic APIs. Currently, it contains accident data that are collected from February 2016 to June 2020 for the Contiguous United States. Check here to learn more about this dataset.

US-Accidents can be used for numerous applications such as real-time car accident prediction, studying car accidents hotspot locations, casualty analysis and extracting cause and effect rules to predict car accidents, and studying the impact of precipitation or other environmental stimuli on accident occurrence. The most recent release of the dataset can also be useful to study the impact of COVID-19 on traffic behavior and accidents.


Social Good
IoT
Machine Learning

OVERVIEW

✔ 3.5 MILLION RECORD-50 COLUMNS

✔ 1.24 GB

✔ MANY OUTLIER POINTS

✔ MISSING VALUES

✔ TEXT COLUMN

✔ NOMINAL-ORDINAL CATEGORICAL FEATURES

✔ INTERVAL-RATIO NUMERICAL FEATURES

✔ BOOLEAN FEATURES

✔ DATE FEATURES

Traffic Attributes(12)

Address Attributes(9)

Weather Attributes (11)

POI Attributes (13)

Period of Day (4)

STATE WITH THE MOST ACCIDENTS

816825
CA

329284
TX

COUNTY WITH THE MOST ACCIDENTS

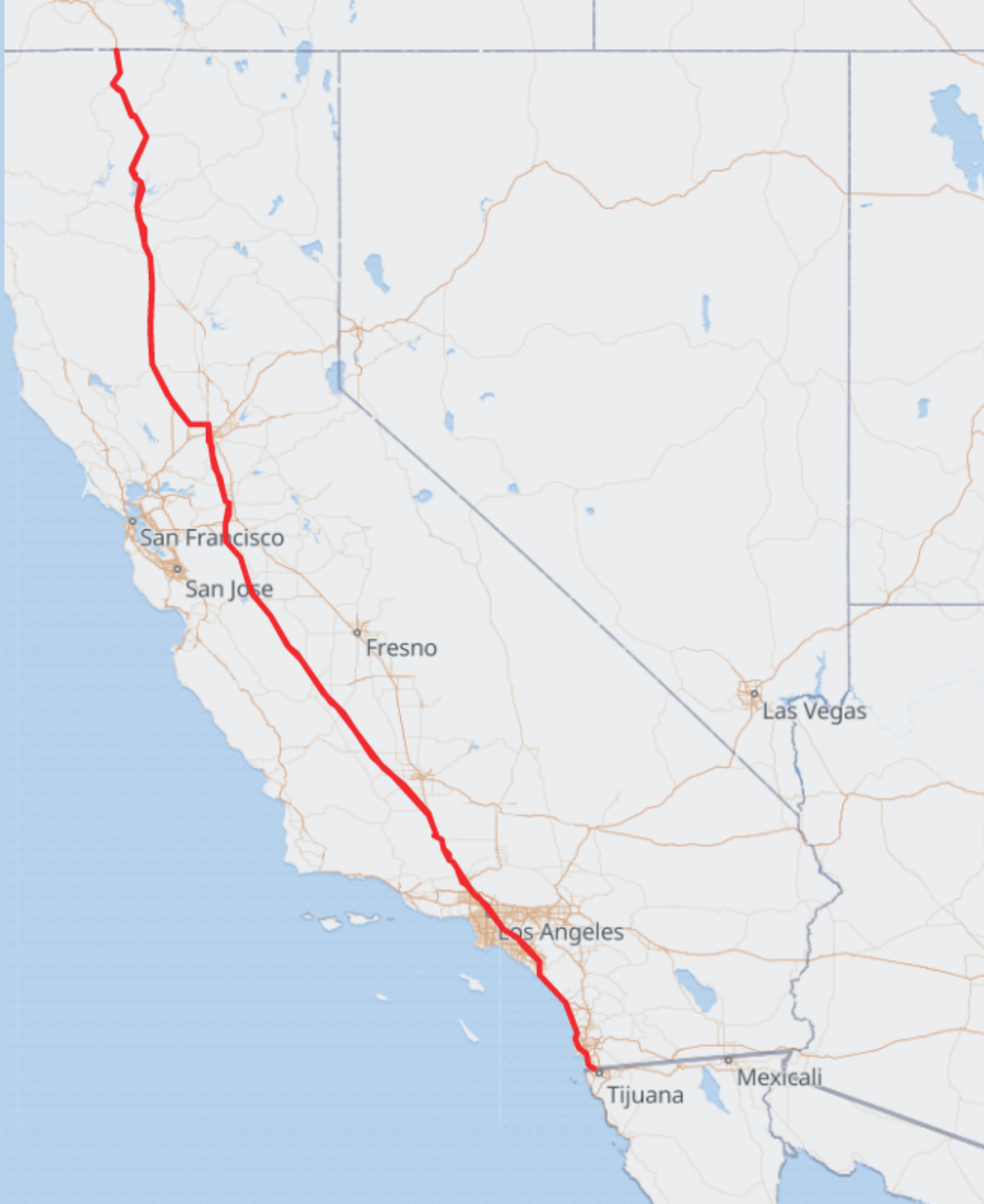
101240

HOUSTON

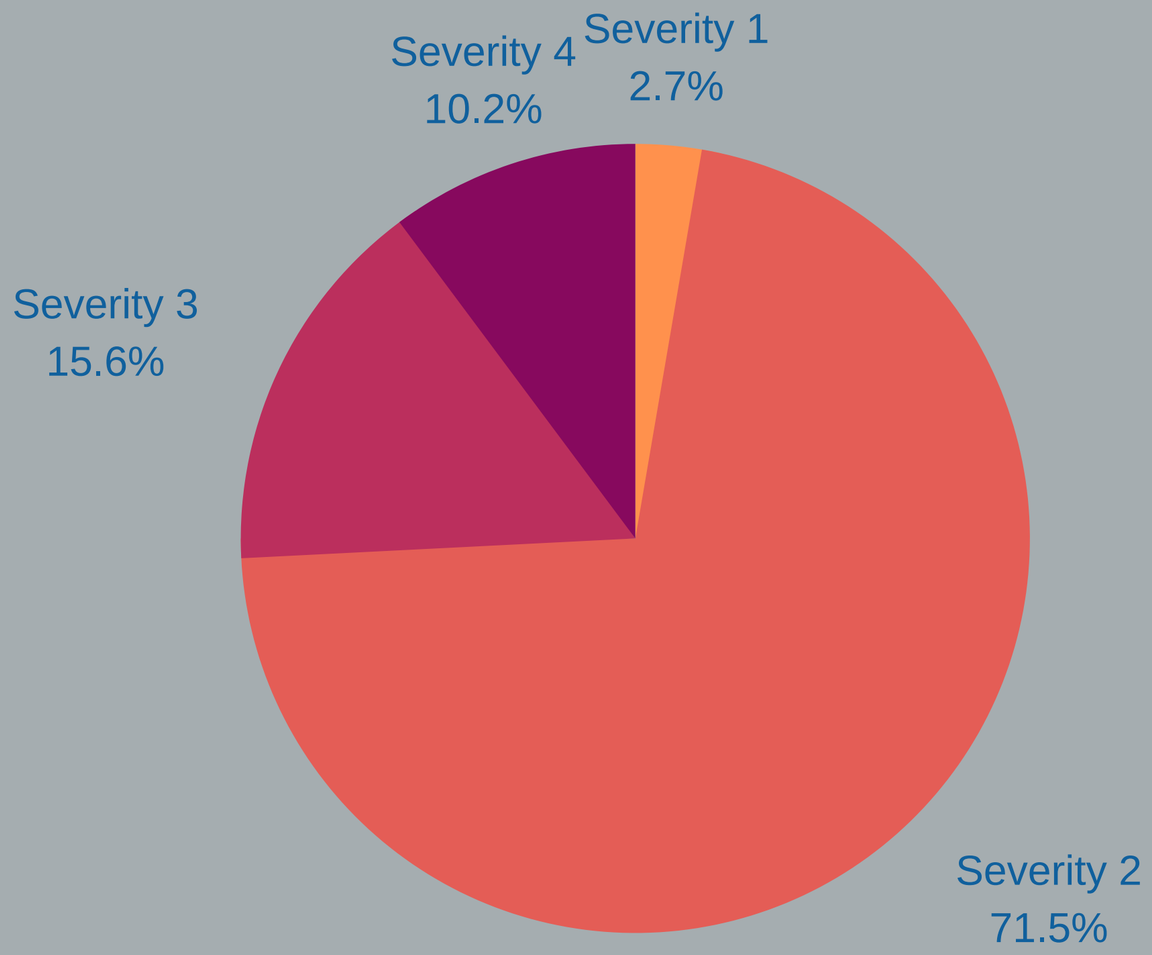
I-5 N

41264

The road with the most accidents

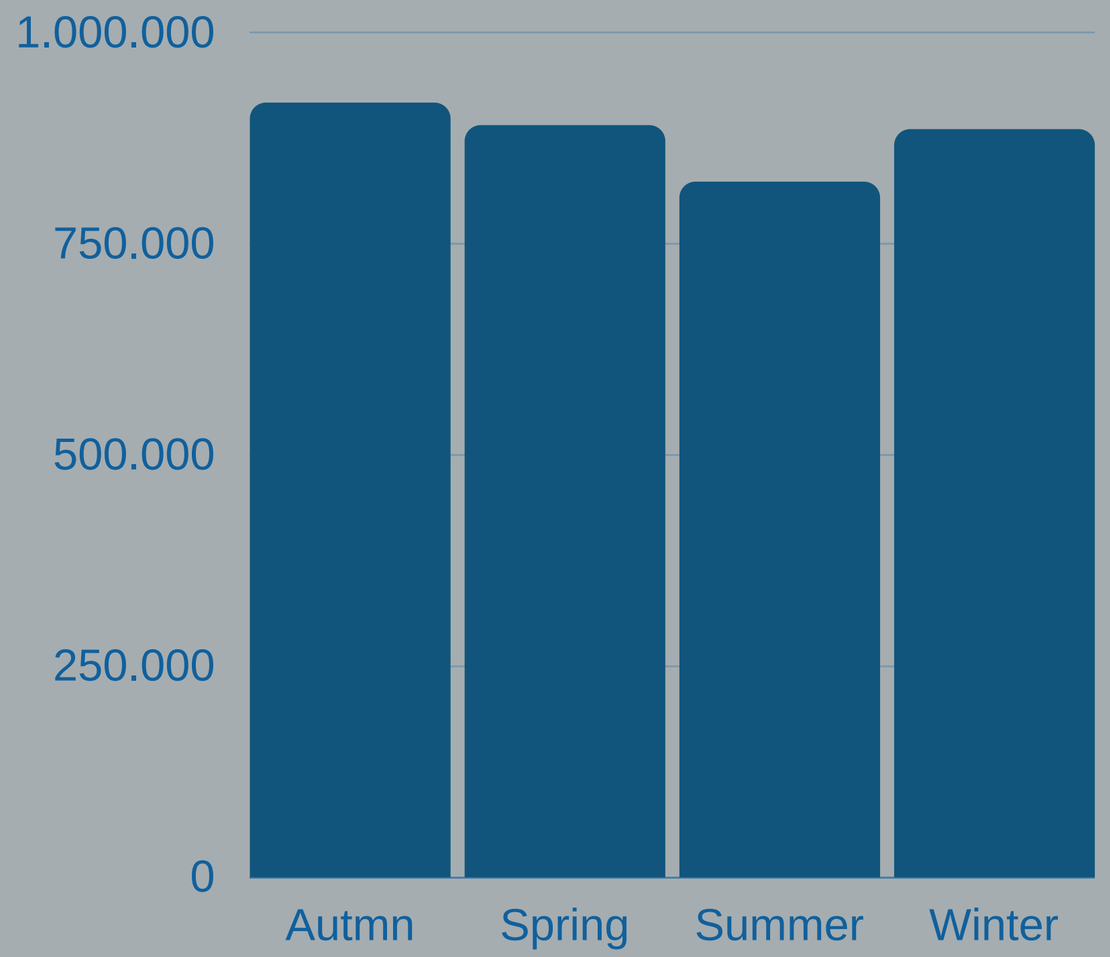


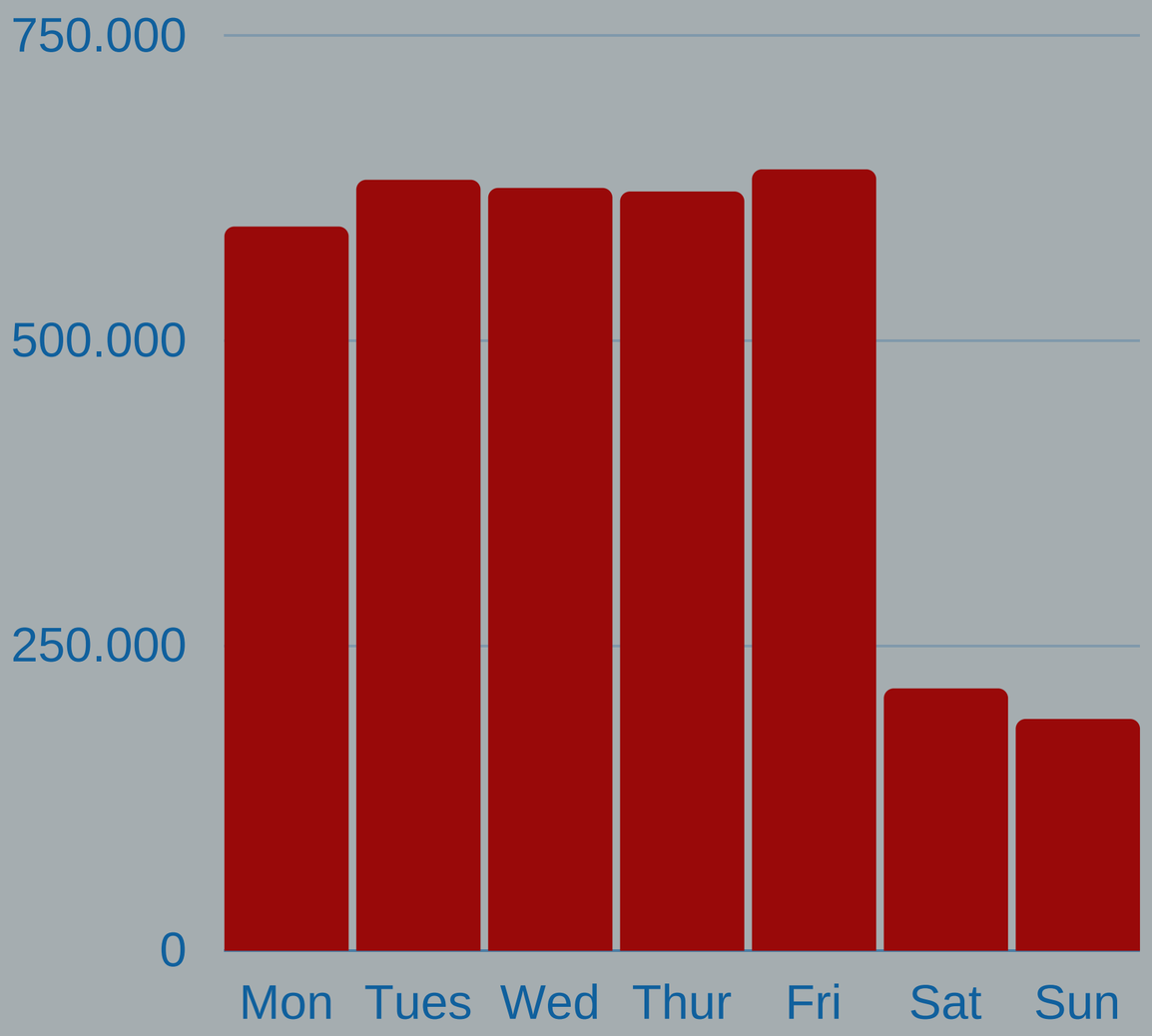
STATISTICS



SHOWS THE SEVERITY OF THE ACCIDENT, A NUMBER BETWEEN 1 AND 4

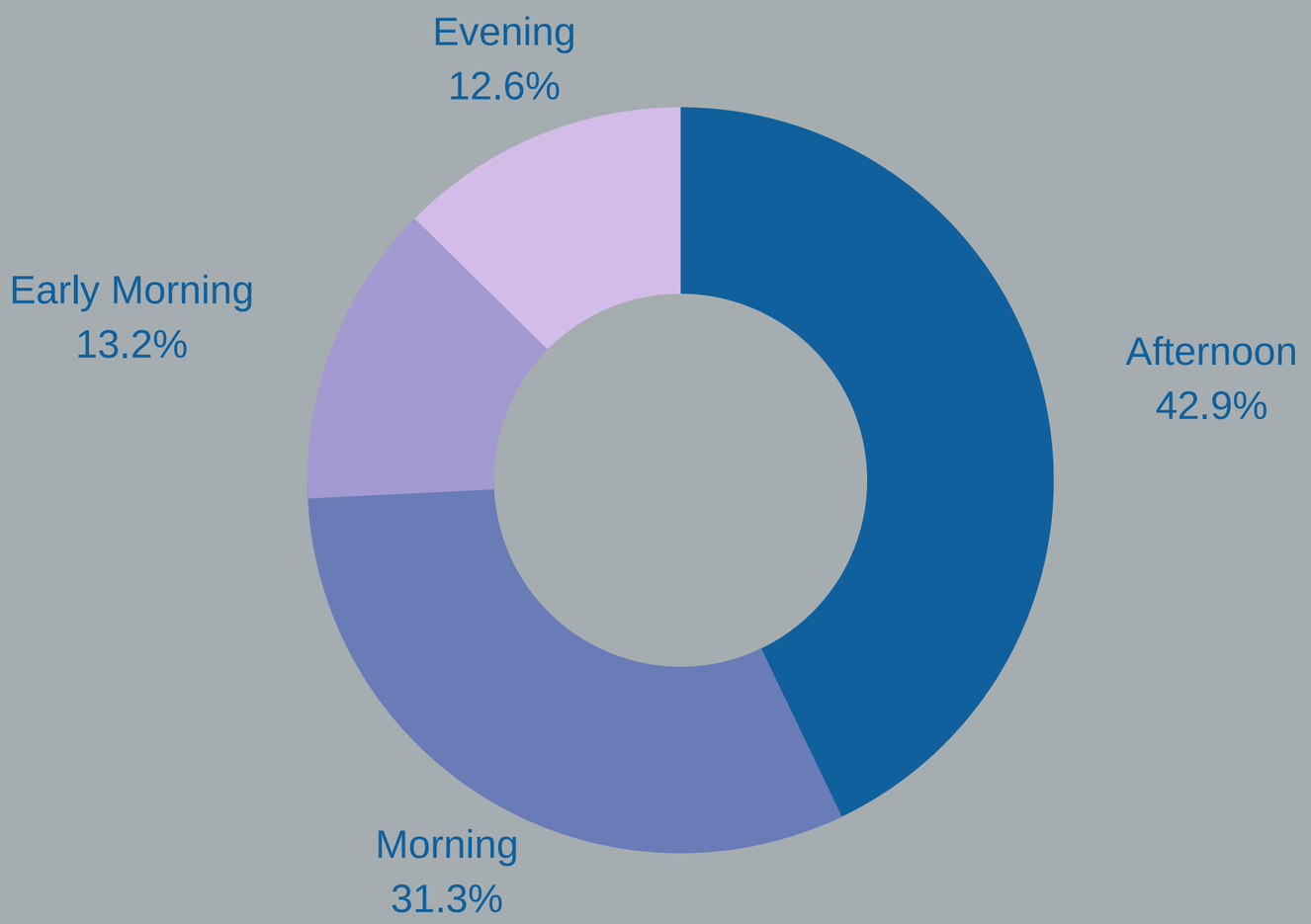
DISTRIBUTION OF ACCIDENT SEVERITY BY SEASONS





SEVERITY RATES BY WEEKENDS AND WEEKDAYS

ACCIDENTS RATES BY TIME OF DAY



81.96%	82.25%	0.005%	39.49%
The rate of accidents on the left side	The rate of no traffic light	The rate of accidents at a roundabout	Accidents that have a lower duration than 30 minutes

Interesting Facts

RECORDED PEAK TEMPERATURE (C)

77

The lowest
-67.2222222

RECORDED PEAK HUMİDİTY (%)

100

The lowest
1

MOST ACCİDENT DATE

15-05-2017
09:22:55

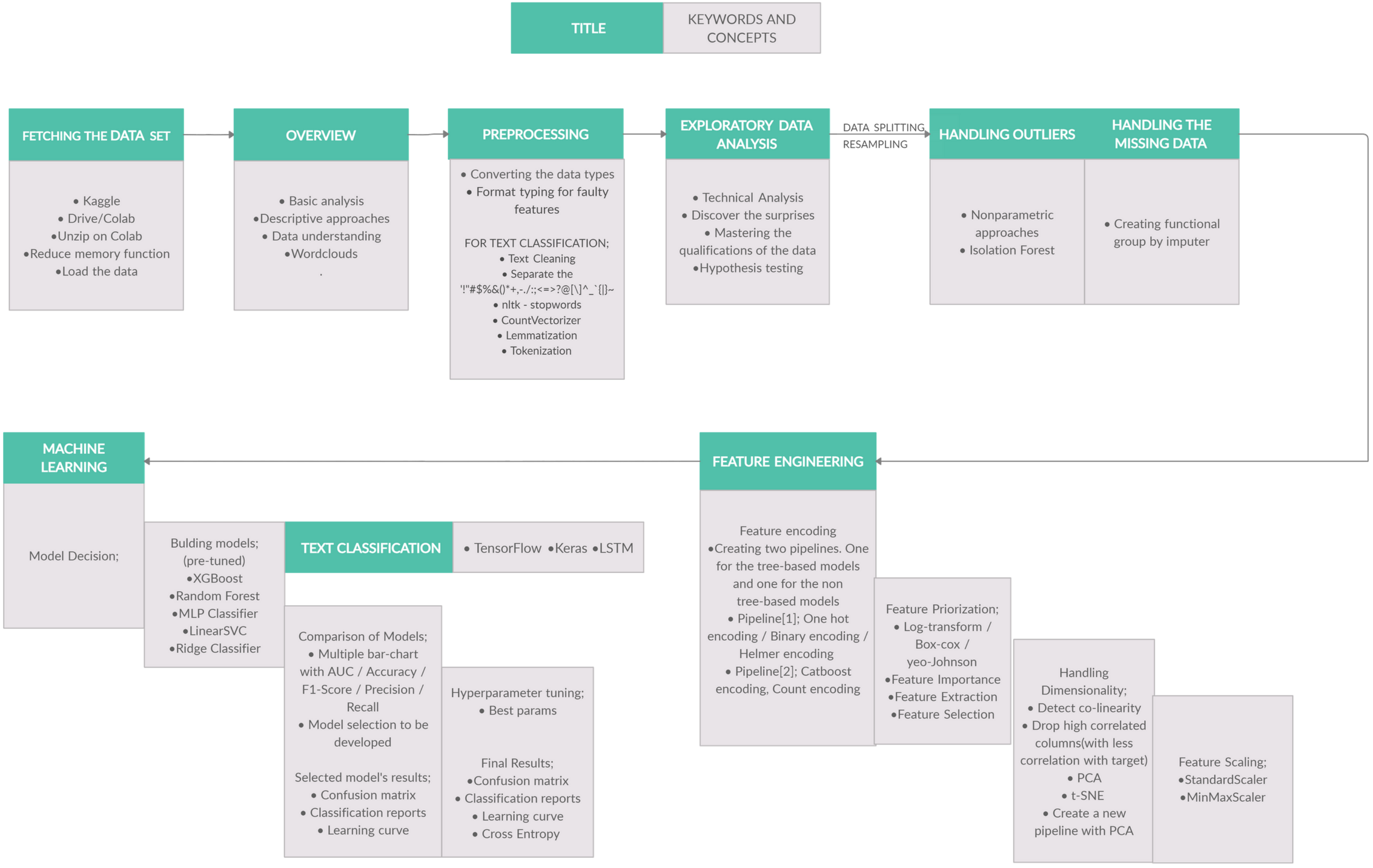
(74 crashes)

THE LARGEST DISTANCE OF ACCIDENT(KM)

249.6

The Smallest
0

PROCESS PATH


TITLE
KEYWORDS AND CONCEPTS
FETCHING THE DATA SET
• Kaggle
• Drive/Colab
•Unzip on Colab
•Reduce memory function
•Load the data
OVERVIEW
• Basic analysis
•Descriptive approaches
• Data understanding
•Wordclouds
.
PREPROCESSING
• Converting the data types
• Format typing for faulty features
FOR TEXT CLASSIFICATION;
• Text Cleaning
• Separate the '!"#$%&()*+,-./:;<=>?@[\]^_`{|}~
• nltk - stopwords
• CountVectorizer
• Lemmatization
• Tokenization
EXPLORATORY DATA ANALYSIS
• Technical Analysis
• Discover the surprises
• Mastering the qualifications of the data
•Hypothesis testing
DATA SPLITTING
RESAMPLING
HANDLING OUTLIERS
• Nonparametric approaches
• Isolation Forest
HANDLING THE MISSING DATA
• Creating functional group by imputer
FEATURE ENGINEERING
Feature encoding
•Creating two pipelines. One for the tree-based models and one for the non tree-based models
• Pipeline[1]; One hot encoding / Binary encoding / Helmer encoding
• Pipeline[2]; Catboost encoding, Count encoding
Feature Priorization;
• Log-transform / Box-cox / yeo-Johnson
•Feature Importance
•Feature Extraction
•Feature Selection
Handling Dimensionality;
• Detect co-linearity
• Drop high correlated columns(with less correlation with target)
• PCA
• t-SNE
• Create a new pipeline with PCA
Feature Scaling;
•StandardScaler
•MinMaxScaler
MACHINE LEARNING
Model Decision;
Bulding models; (pre-tuned)
•XGBoost
•Random Forest
•MLP Classifier
•LinearSVC
•Ridge Classifier
TEXT CLASSIFICATION
• TensorFlow •Keras •LSTM
Comparison of Models;
• Multiple bar-chart with AUC / Accuracy / F1-Score / Precision / Recall
• Model selection to be developed
Selected model's results;
• Confusion matrix
• Classification reports
• Learning curve
Hyperparameter tuning;
• Best params
Final Results;
•Confusion matrix
• Classification reports
• Learning curve
• Cross Entropy

	Traffic Message Channel (TMC) Description	Percentage of having severity4 records for each tmc code
343	(Q) earlier accident(s)	0.081612
341	(Q) jackknifed articulated lorr(y/ies)	0.052811
339	(Q) jackknifed trailer(s)	0.042810
248	accident. Delays (Q) expected	0.027317
246	\t(Q) accident(s). Three lanes blocked	0.017712
200	multi vehicle pile up. Delays (Q)	0.015152
202	(Q) serious accident(s)	0.013348
406	(Q th) entry slip road closed	0.012703
336	(Q) oil spillage accident(s)	0.011236
245	(Q) accident(s). Two lanes blocked	0.008679
222	(Q) accident(s). Queuing traffic	0.008136
236	(Q) accident(s). Heavy traffic	0.006601
203	multi-vehicle accident	0.006520
206	\t(Q) fuel spillage accident(s)	0.005495
247	accident. Delays (Q)	0.005026
241	Accident(s). Right lane blocked	0.003965
244	(Q) accident(s). Hard shoulder blocked	0.003942
201	Accident(s)	0.001980
229	(Q) accident(s). Slow traffic	0.001308
239	(Q) accident(s). Traffic building up	NaN
351	(Q) accident(s) in roadworks area	NaN

OVERVIEW

- Basic analysis
- Descriptive approaches
- Data understanding
- Wordclouds

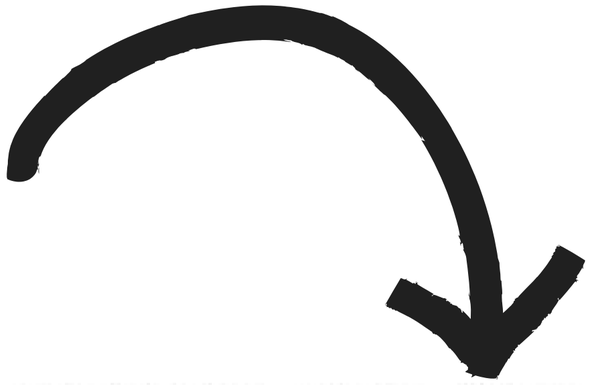
```
array(['Light Rain', 'Overcast', 'Mostly Cloudy', 'Rain', 'Light Snow',
       'Haze', 'Scattered Clouds', 'Partly Cloudy', 'Clear', 'Snow',
       'Light Freezing Drizzle', 'Light Drizzle', 'Fog', 'Shallow Fog',
       'Heavy Rain', 'Light Freezing Rain', 'Cloudy', 'Drizzle', nan,
       'Light Rain Showers', 'Mist', 'Smoke', 'Patches of Fog',
       'Light Freezing Fog', 'Light Haze', 'Light Thunderstorms and Rain',
       'Thunderstorms and Rain', 'Fair', 'Volcanic Ash', 'Blowing Sand',
       'Blowing Dust / Windy', 'Widespread Dust', 'Fair / Windy',
       'Rain Showers', 'Mostly Cloudy / Windy', 'Light Rain / Windy',
       'Hail', 'Heavy Drizzle', 'Showers in the Vicinity', 'Thunderstorm',
       'Light Rain Shower', 'Light Rain with Thunder',
       'Partly Cloudy / Windy', 'Thunder in the Vicinity', 'T-Storm',
       'Heavy Thunderstorms and Rain', 'Thunder', 'Heavy T-Storm',
       'Funnel Cloud', 'Heavy T-Storm / Windy', 'Blowing Snow',
       'Light Thunderstorms and Snow', 'Heavy Snow', 'Low Drifting Snow',
       'Light Ice Pellets', 'Ice Pellets', 'Squalls', 'N/A Precipitation',
       'Cloudy / Windy', 'Light Fog', 'Sand', 'Snow Grains',
       'Snow Showers', 'Heavy Thunderstorms and Snow', 'Rain / Windy',
       'Heavy Rain / Windy', 'Heavy Ice Pellets', 'Light Snow / Windy',
       'Heavy Freezing Rain', 'Small Hail', 'Heavy Rain Showers',
       'T-Storm / Windy', 'Patches of Fog / Windy', 'Drizzle / Windy',
       'Thunder / Windy', 'Wintry Mix', 'Squalls / Windy', 'Rain Shower',
       'Drizzle and Fog', 'Haze / Windy', 'Sand / Dust Whirlwinds',
       'Blowing Dust', 'Fog / Windy', 'Smoke / Windy',
       'Wintry Mix / Windy', 'Snow / Windy', 'Light Rain Shower / Windy',
       'Heavy Snow / Windy', 'Snow and Sleet',
       'Light Freezing Rain / Windy', 'Light Drizzle / Windy',
       'Light Snow and Sleet', 'Partial Fog', 'Light Snow Shower',
       'Light Snow and Sleet / Windy', 'Freezing Rain',
       'Blowing Snow / Windy', 'Freezing Drizzle', 'Sleet', 'Light Sleet',
       'Rain and Sleet', 'Heavy Sleet', 'Light Snow Grains',
       'Partial Fog / Windy', 'Light Snow with Thunder',
       'Widespread Dust / Windy', 'Sand / Dust Whirlwinds / Windy',
       'Tornado', 'Snow and Thunder', 'Snow and Sleet / Windy',
       'Heavy Snow with Thunder', 'Thunder / Wintry Mix / Windy',
       'Light Snow Showers', 'Heavy Blowing Snow', 'Light Hail',
       'Heavy Smoke', 'Heavy Thunderstorms with Small Hail',
       'Light Thunderstorm', 'Heavy Freezing Drizzle',
       'Light Blowing Snow', 'Thunderstorms and Snow',
       'Freezing Rain / Windy', 'Dust Whirls',
       'Sand / Dust Whirls Nearby', 'Heavy Rain Shower',
       'Thunder and Hail', 'Drifting Snow', 'Thunder and Hail / Windy'],
      dtype=object)
```

PREPROCESSING

- Converting the data types
- Format typing for faulty features

FOR TEXT CLASSIFICATION;

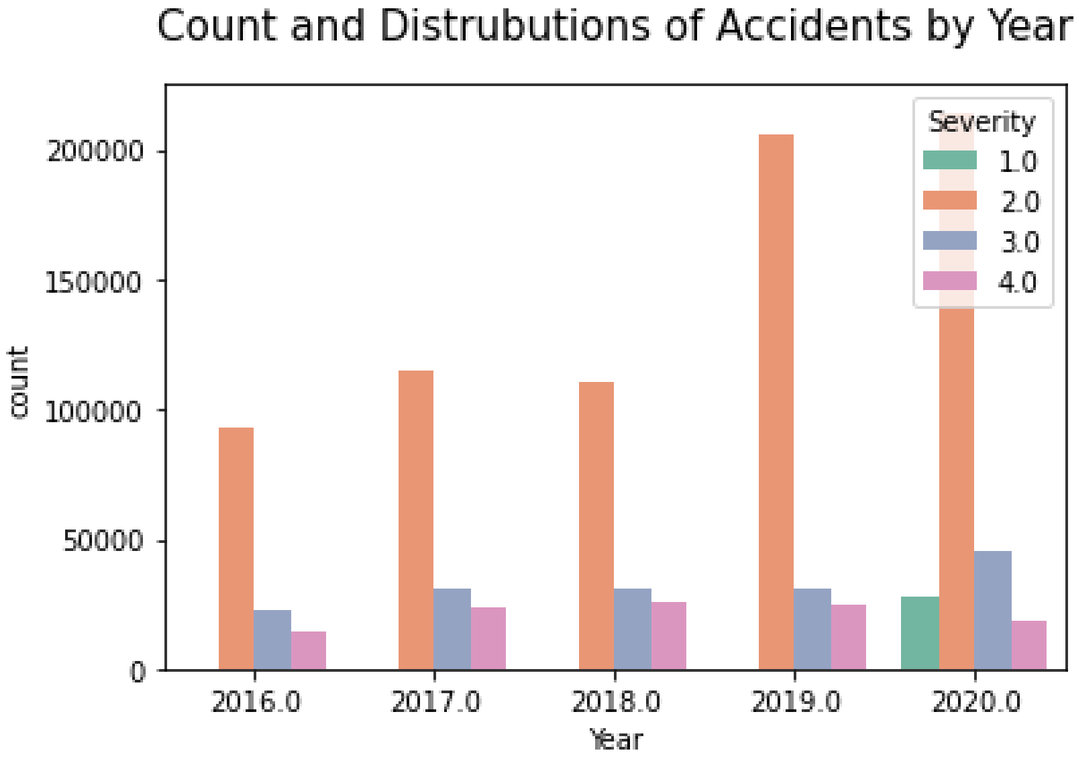
- Text Cleaning
- Separate the '!"#$%&()*+,-./:;<=>?@[\]^_`{|}~
- nltk - stopwords
- CountVectorizer
- Lemmatization
- Tokenization

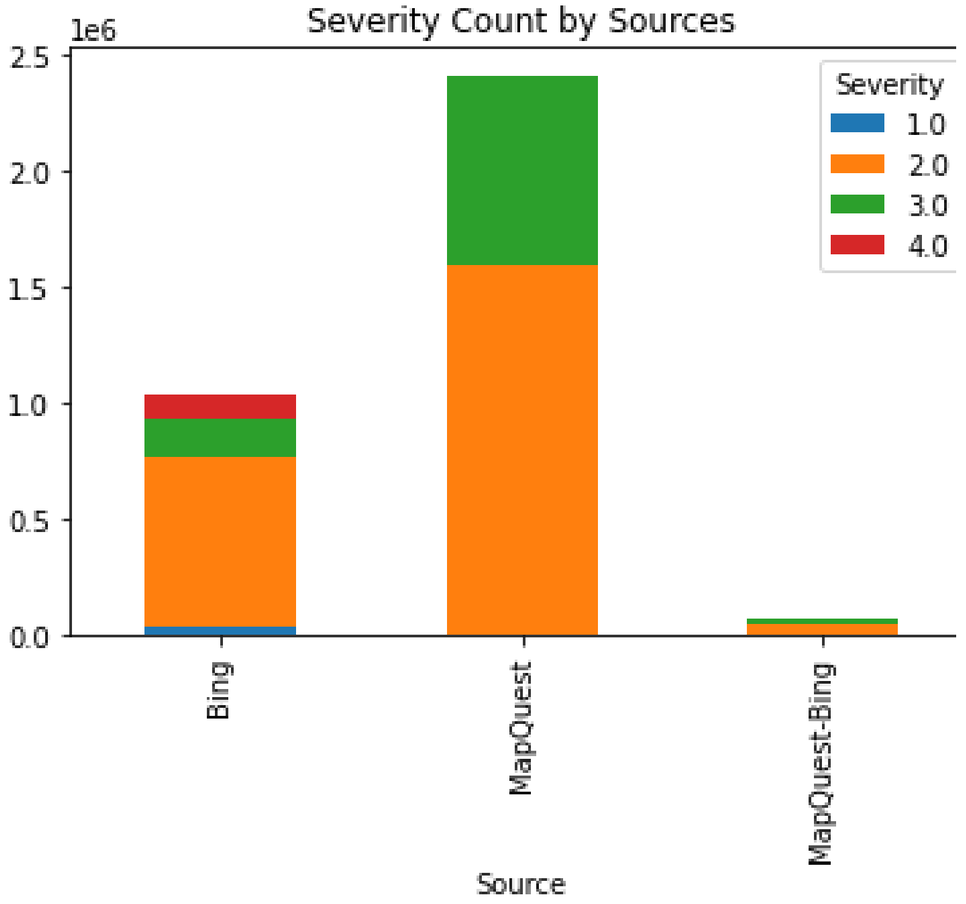


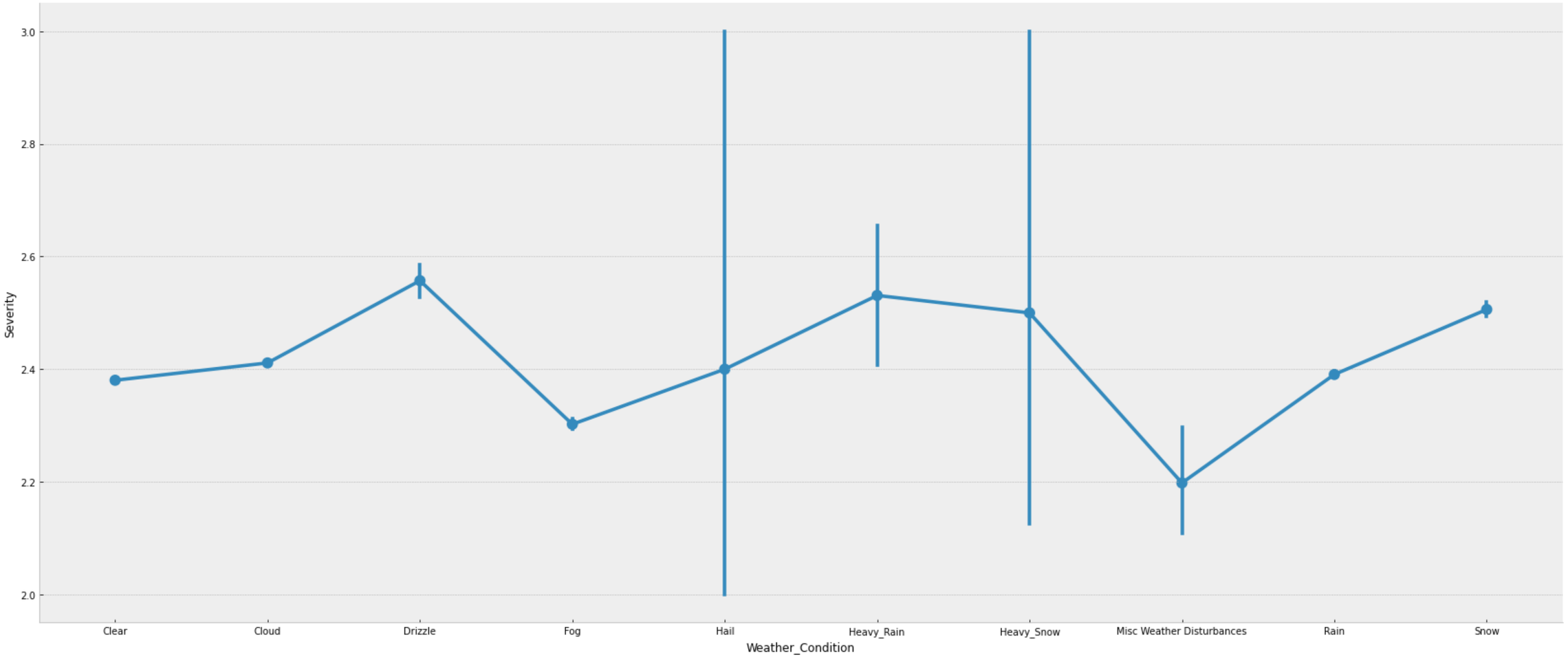
```
array(['Rain', 'Cloud', 'Snow', 'Fog', 'Clear', 'Drizzle', nan,
       'Misc Weather Disturbances', 'Hail', 'Heavy_Rain', 'Heavy_Snow'],
      dtype=object)
```

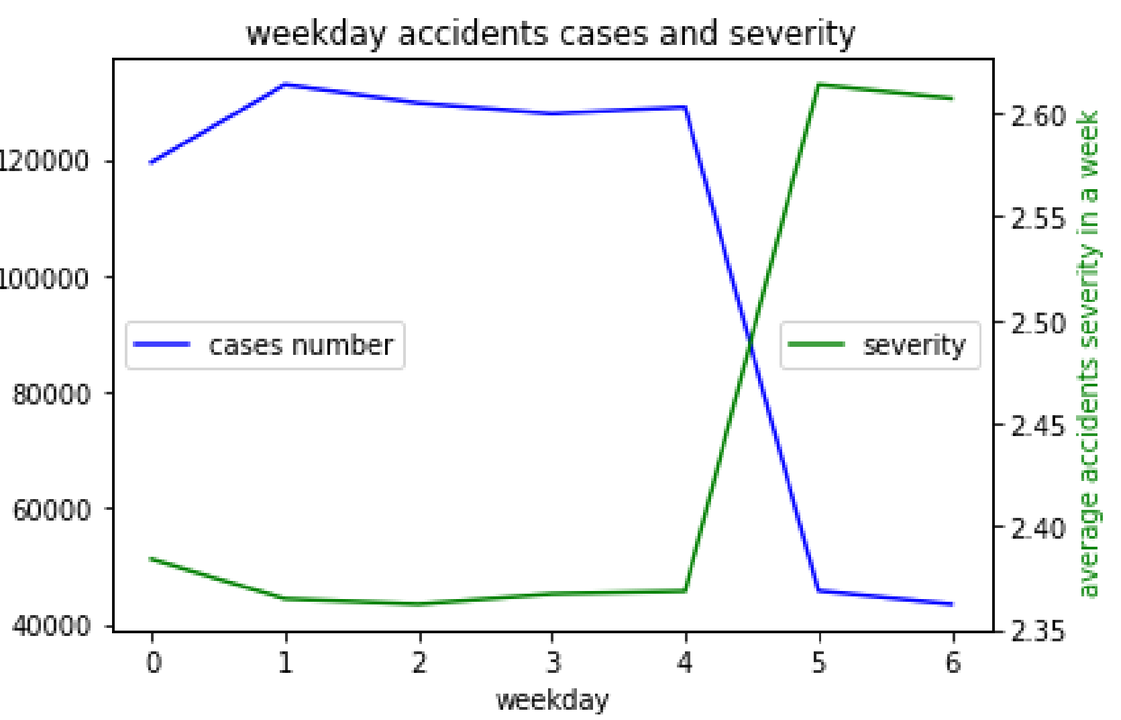
EXPLORATORY DATA ANALYSIS

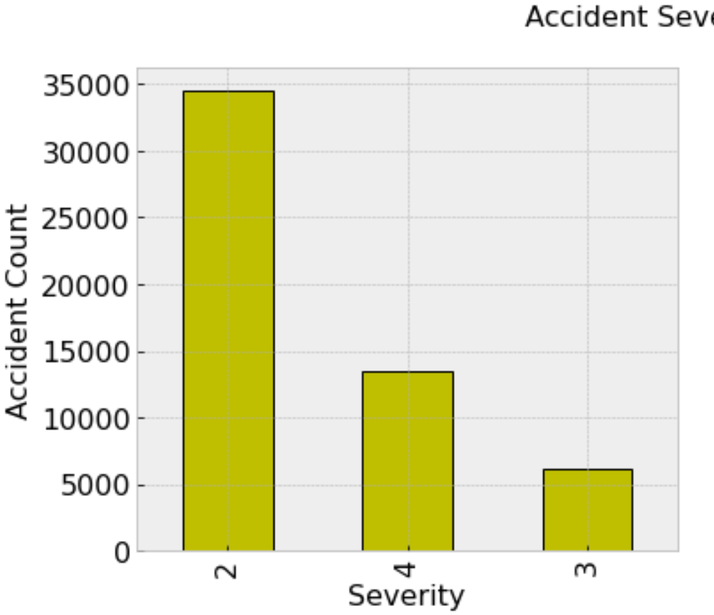
- Technical Analysis
- Discover the surprises
- Mastering the qualifications of the data
- Hypothesis testing

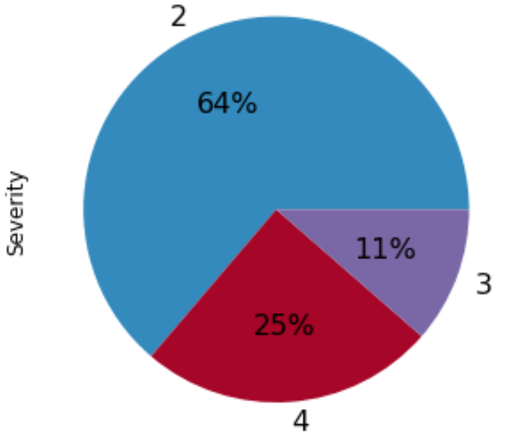


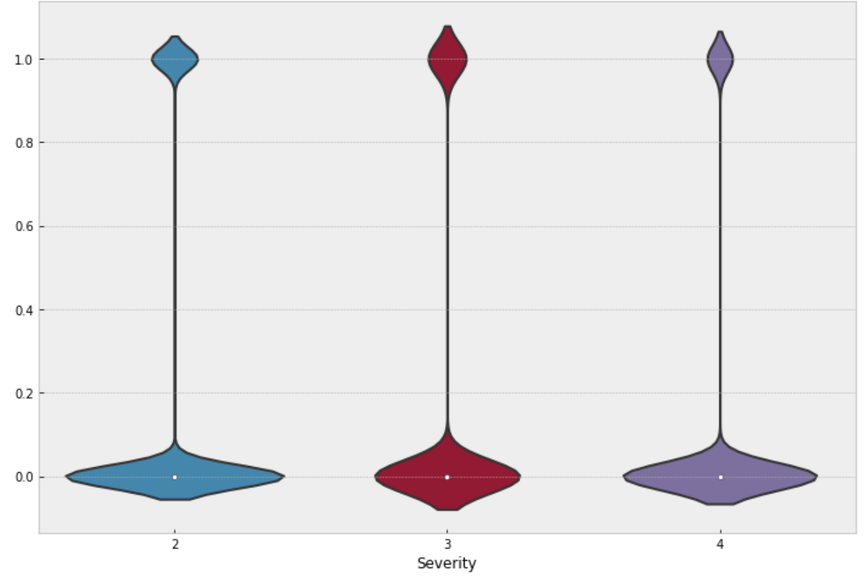



3.0
2.8
2.6
2.4
2.2
2.0
Severity
Clear
Cloud
Drizzle
Fog
Hail
Heavy_Rain
Heavy_Snow
Misc Weather Disturbances
Rain
Snow
Weather_Condition

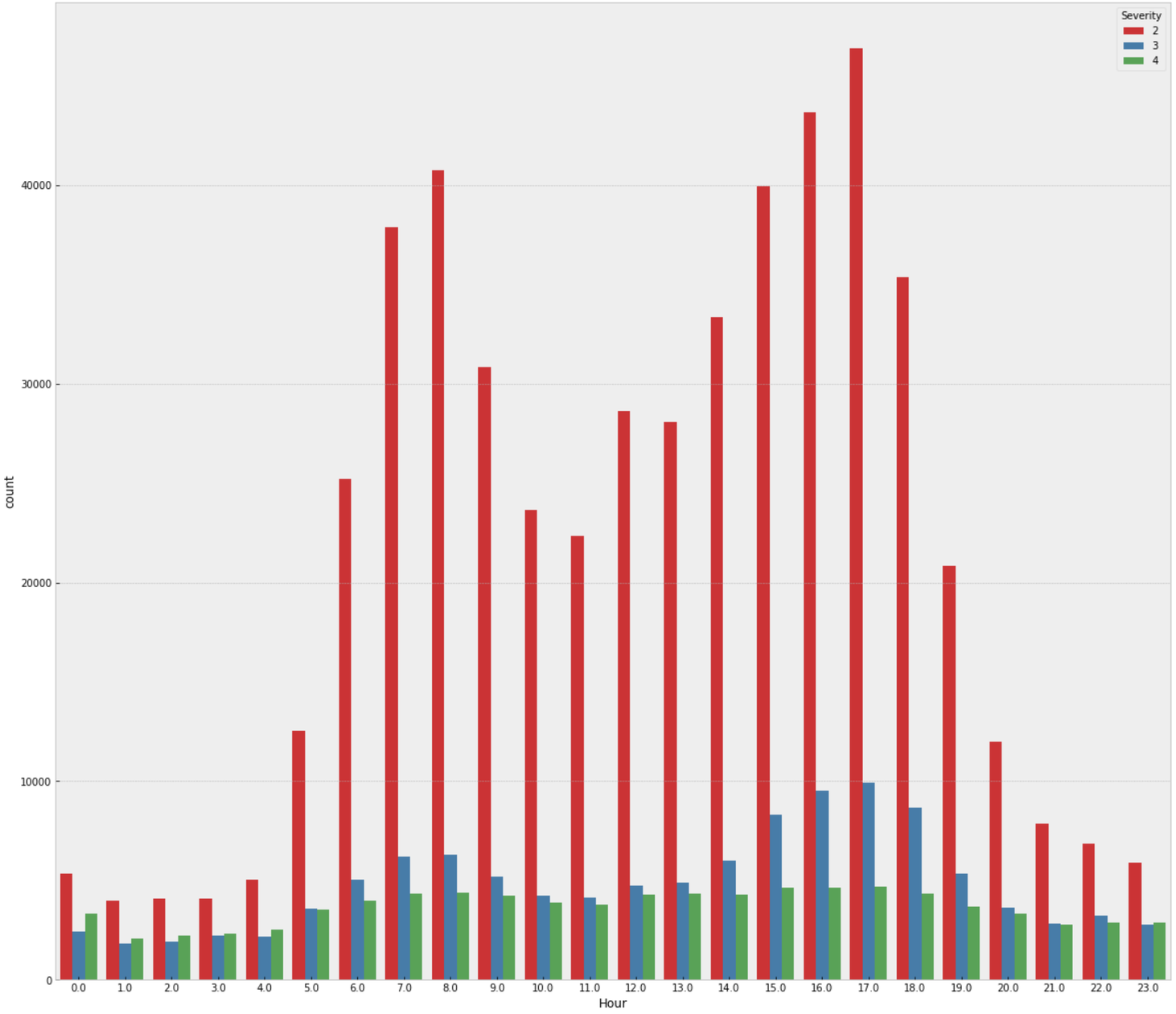

weekday accidents cases and severity
cases number
severity
average accidents severity in a week
weekday
120000
100000
80000
60000
40000
2.60
2.55
2.50
2.45
2.40
2.35
0
1
2
3
4
5
6


Accident Severity Near Rd
Accident Count
Severity
35000
30000
25000
20000
15000
10000
5000
0
2
4
3

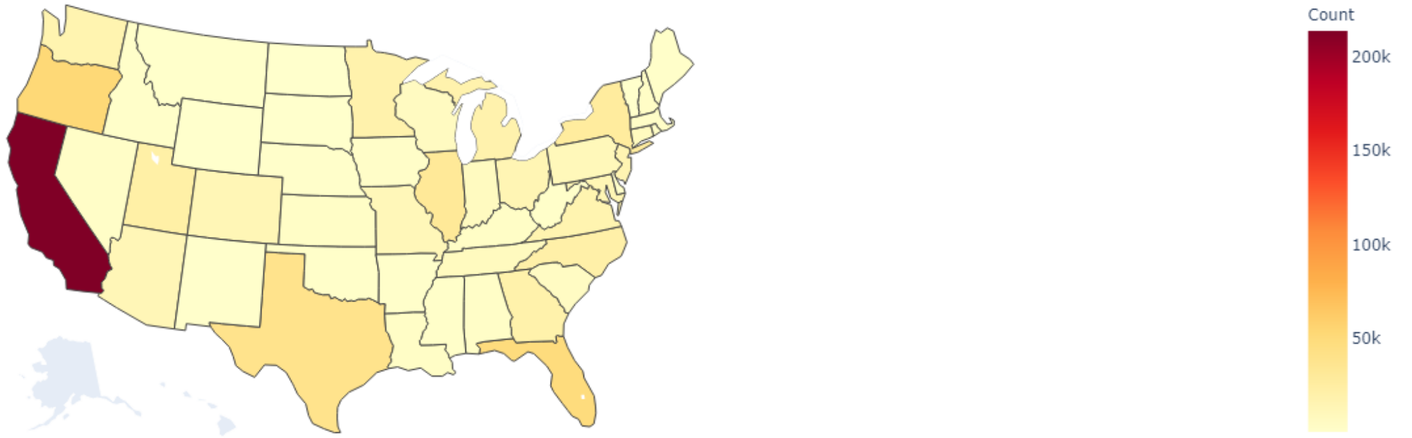

Severity
2
64%
11%
3
25%
4


1.0
0.8
0.6
0.4
0.2
0.0
2
3
4
Severity

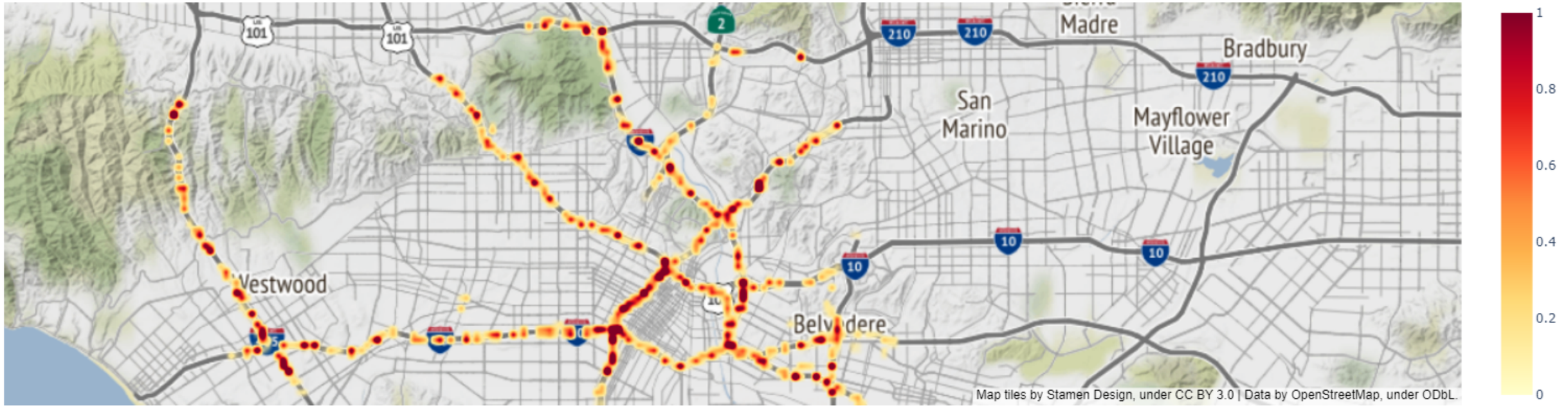
Count and Distrubutions of Accidents by Hour

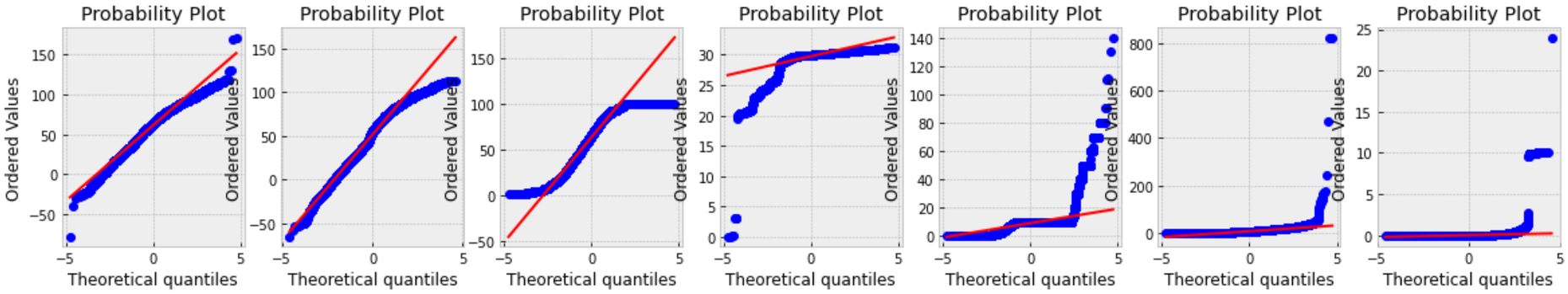


US Accidents by State



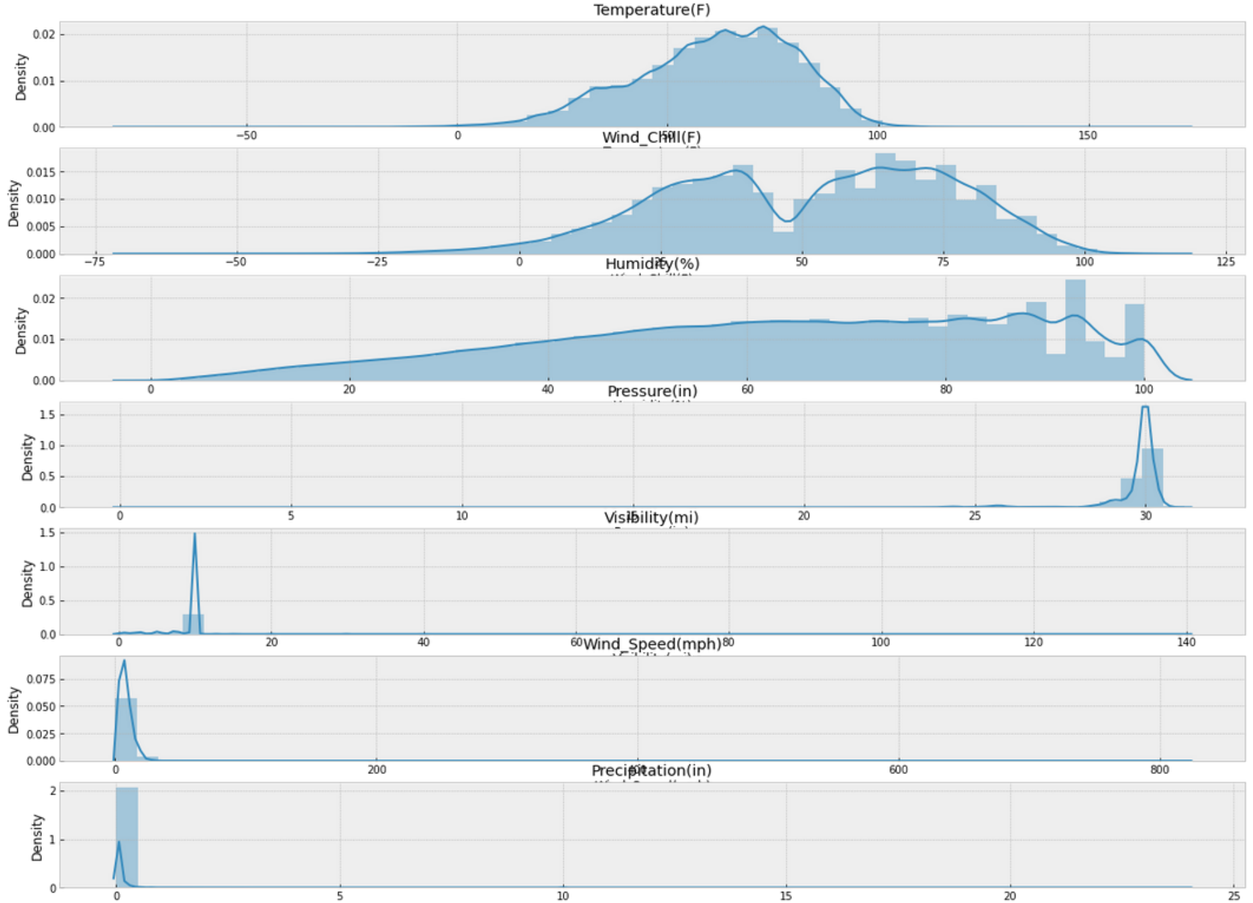
Los Angeles - Accidents

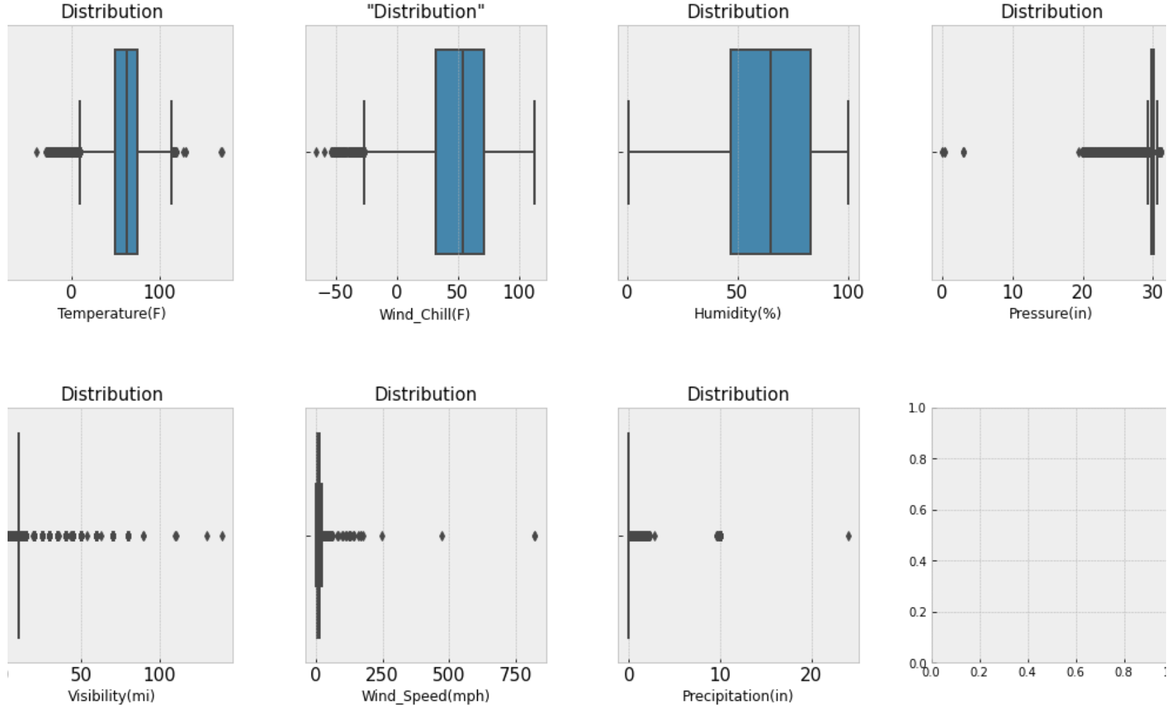




HANDLING OUTLIERS

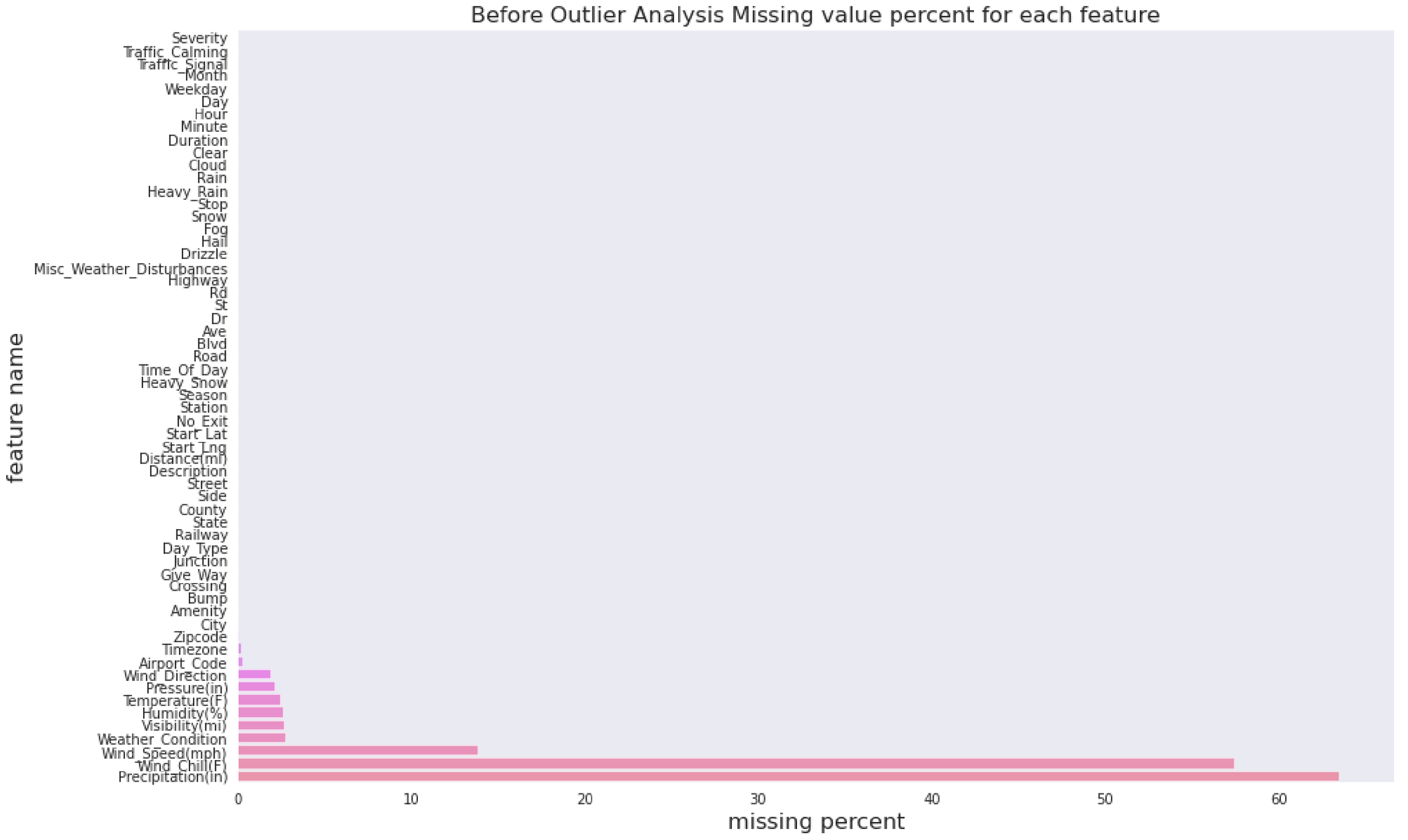
- Nonparametric approaches
- Isolation Forest

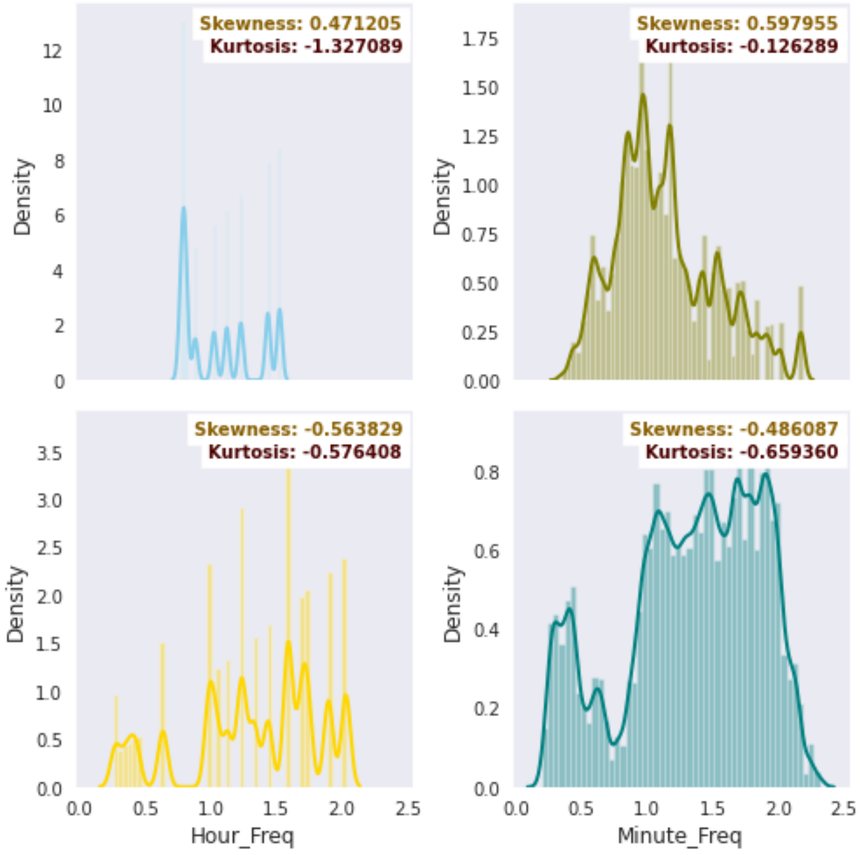


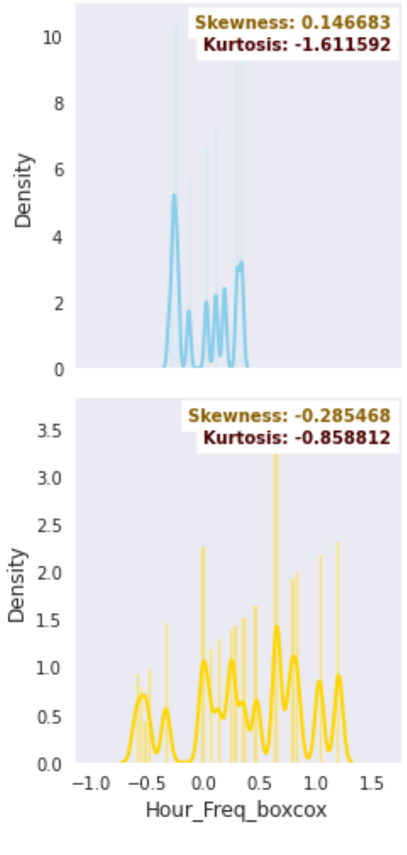


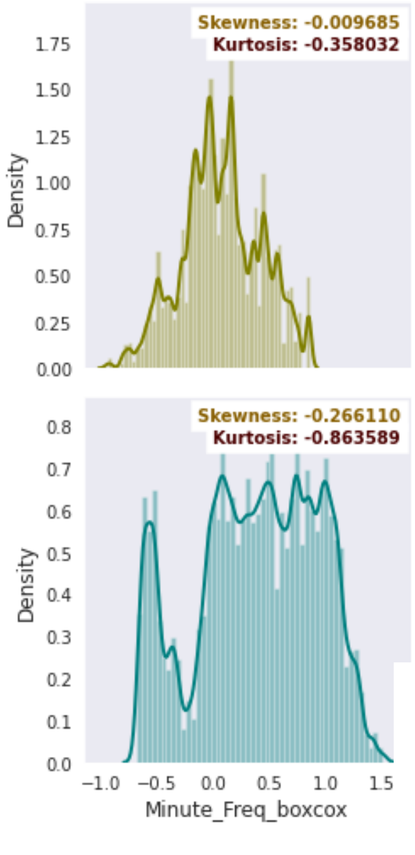
HANDLING THE MISSING DATA

- Creating functional group by imputer





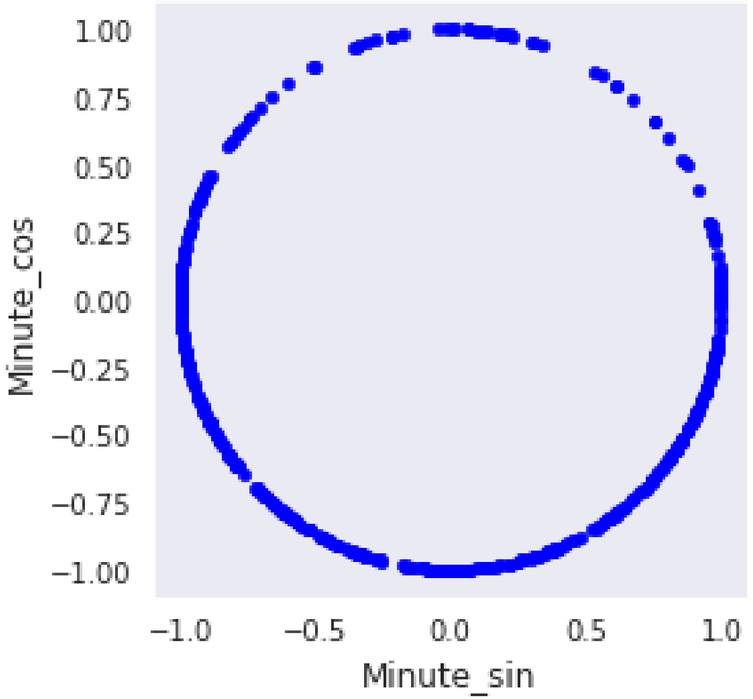


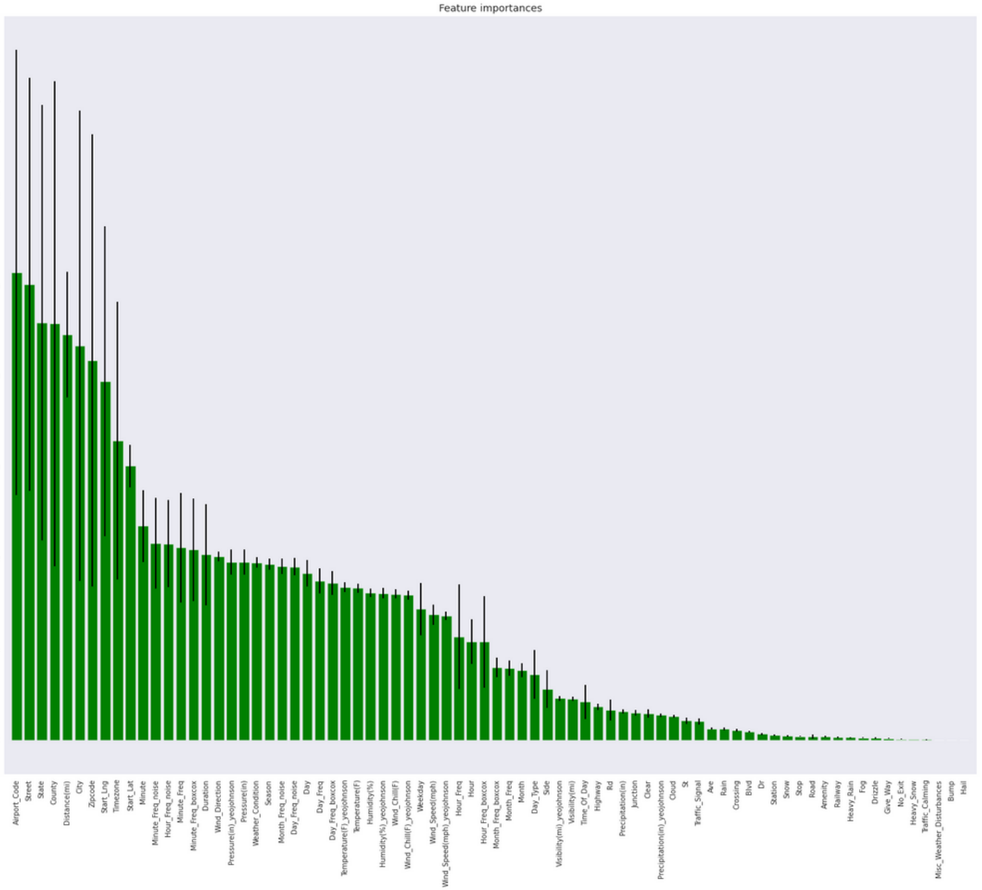


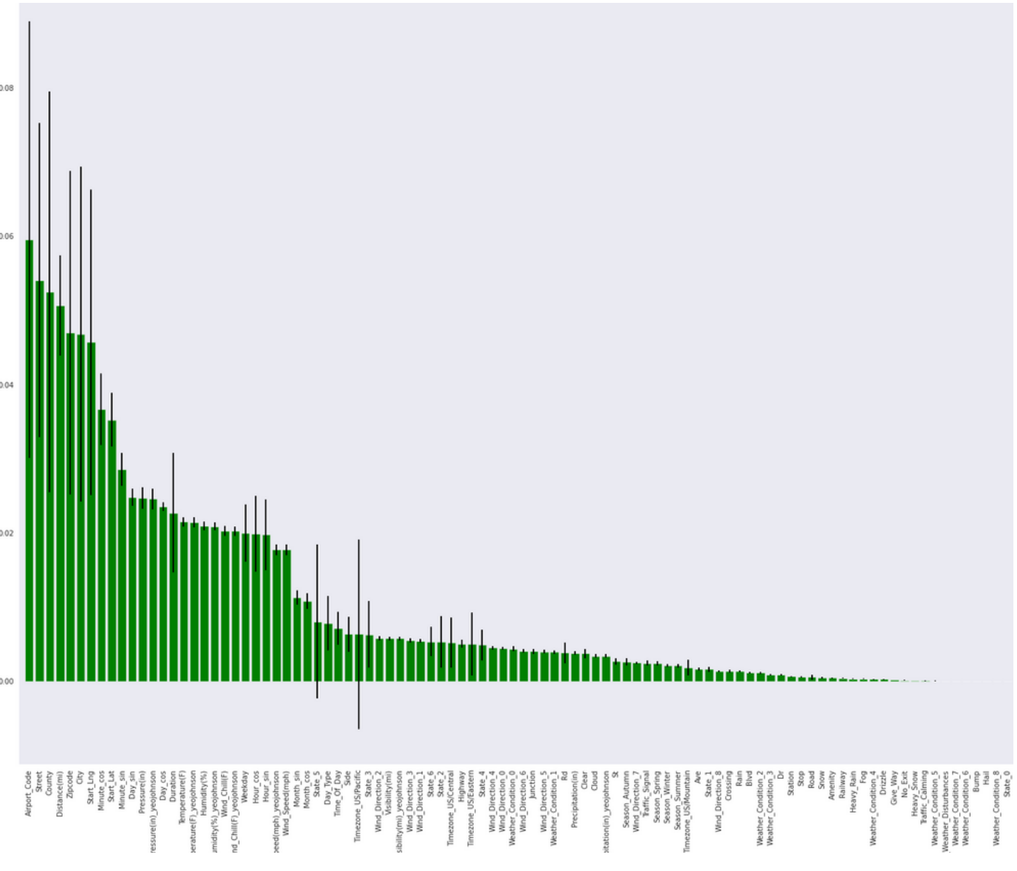
FEATURE ENGINEERING

Feature encoding

- Creating two pipelines. One for the tree-based models and one for the non tree-based models
- Pipeline[1]; One hot encoding / Binary encoding / Helmer encoding
- Pipeline[2]; Catboost encoding, Count encoding

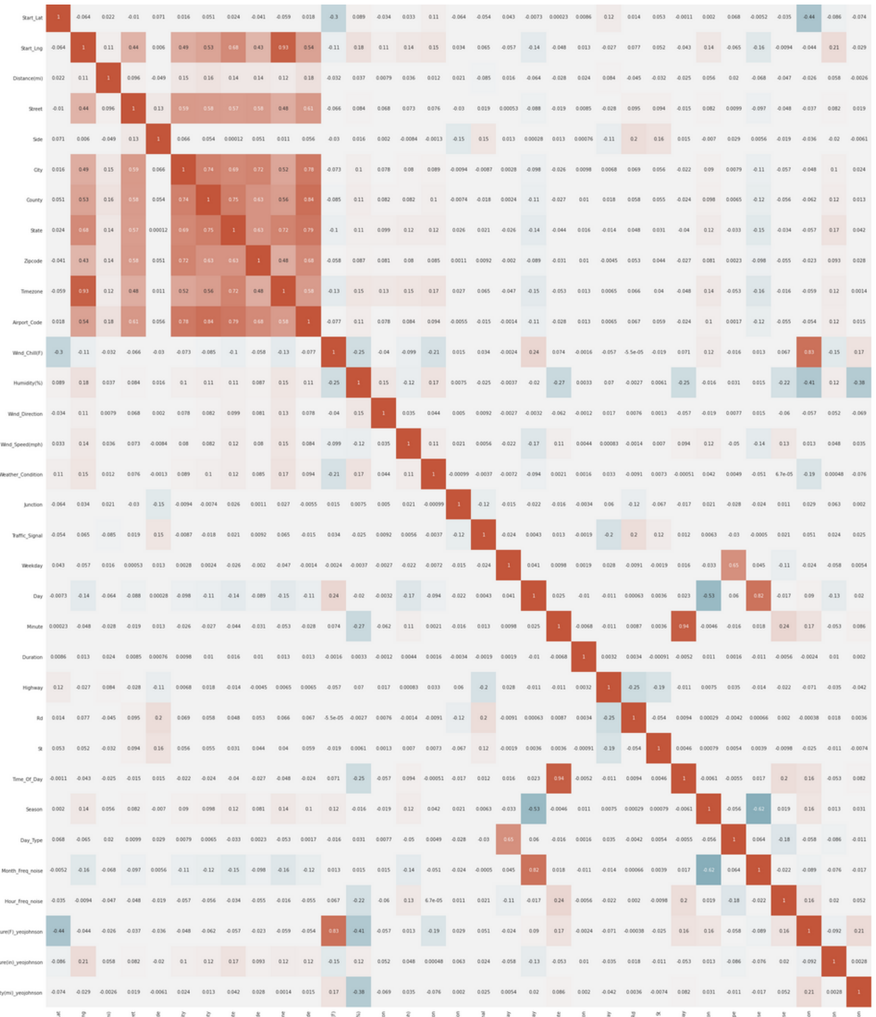


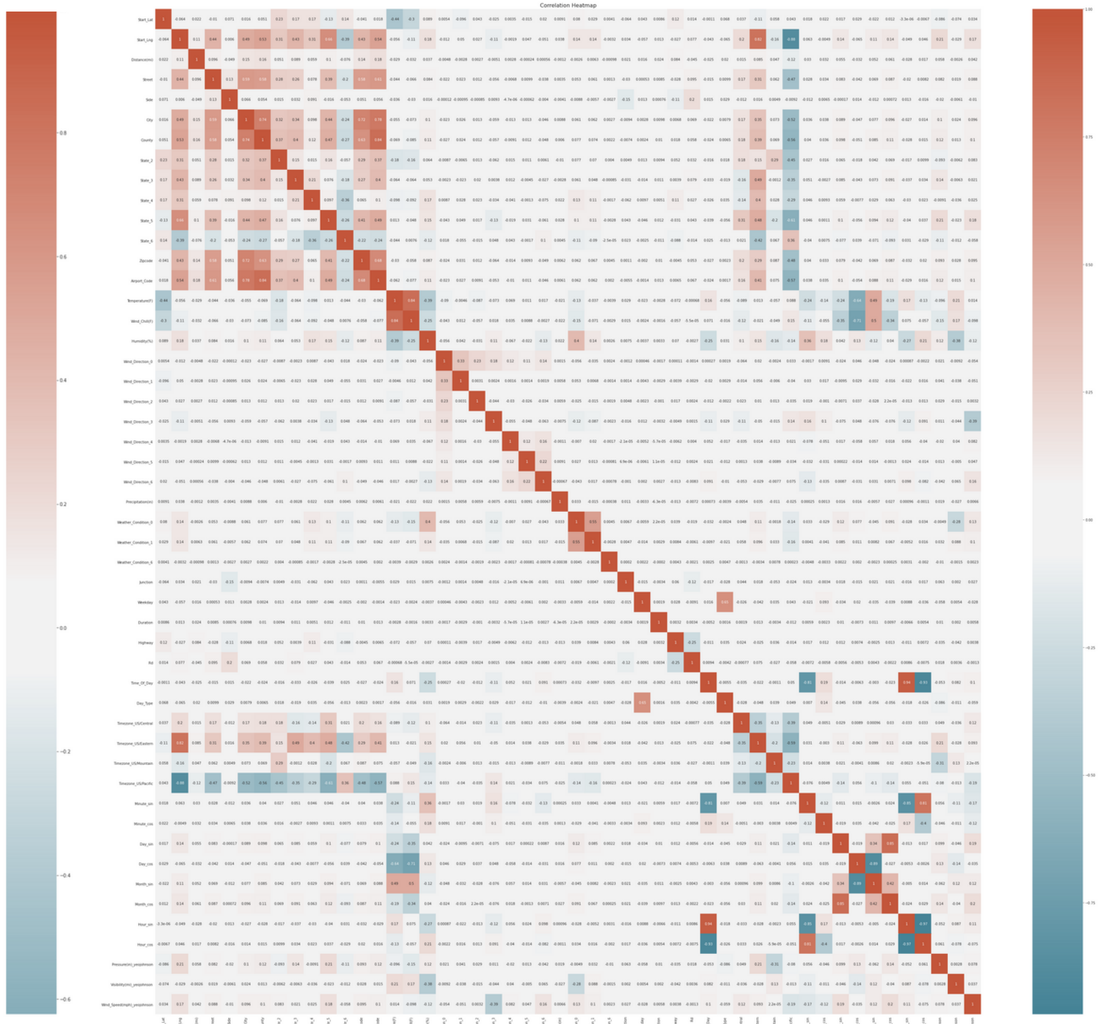




Feature Priorization;

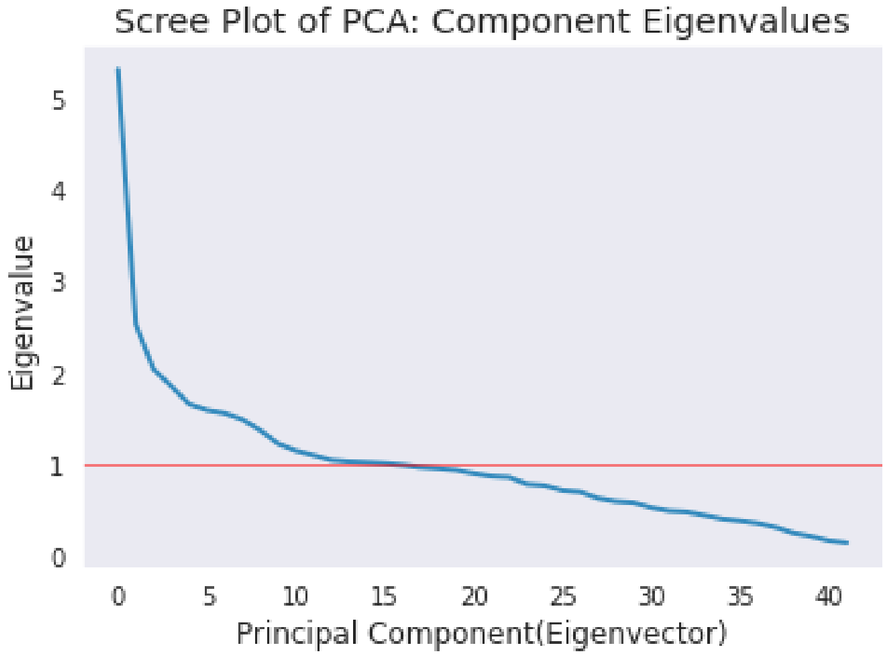
- Log-transform / Box-cox / yeo-Johnson
- Feature Importance
- Feature Extraction
- Feature Selection

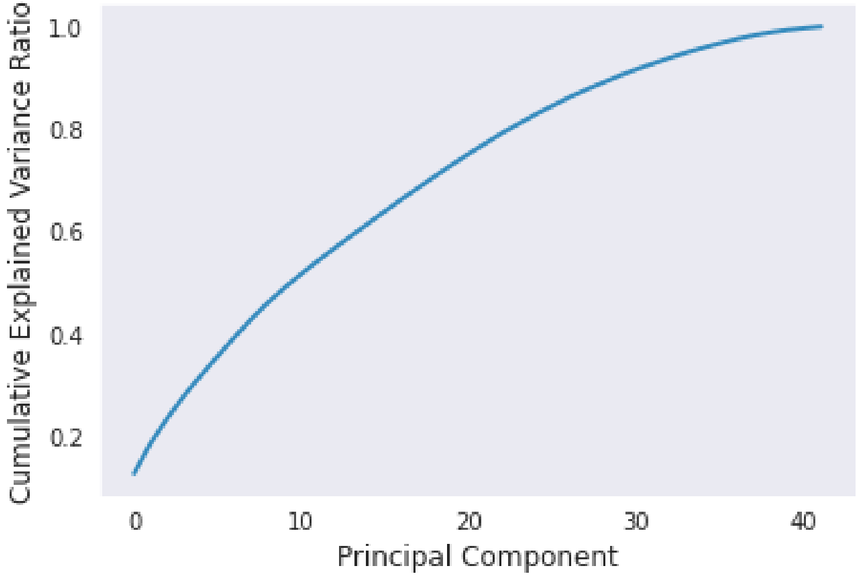




Handling Dimensionality;

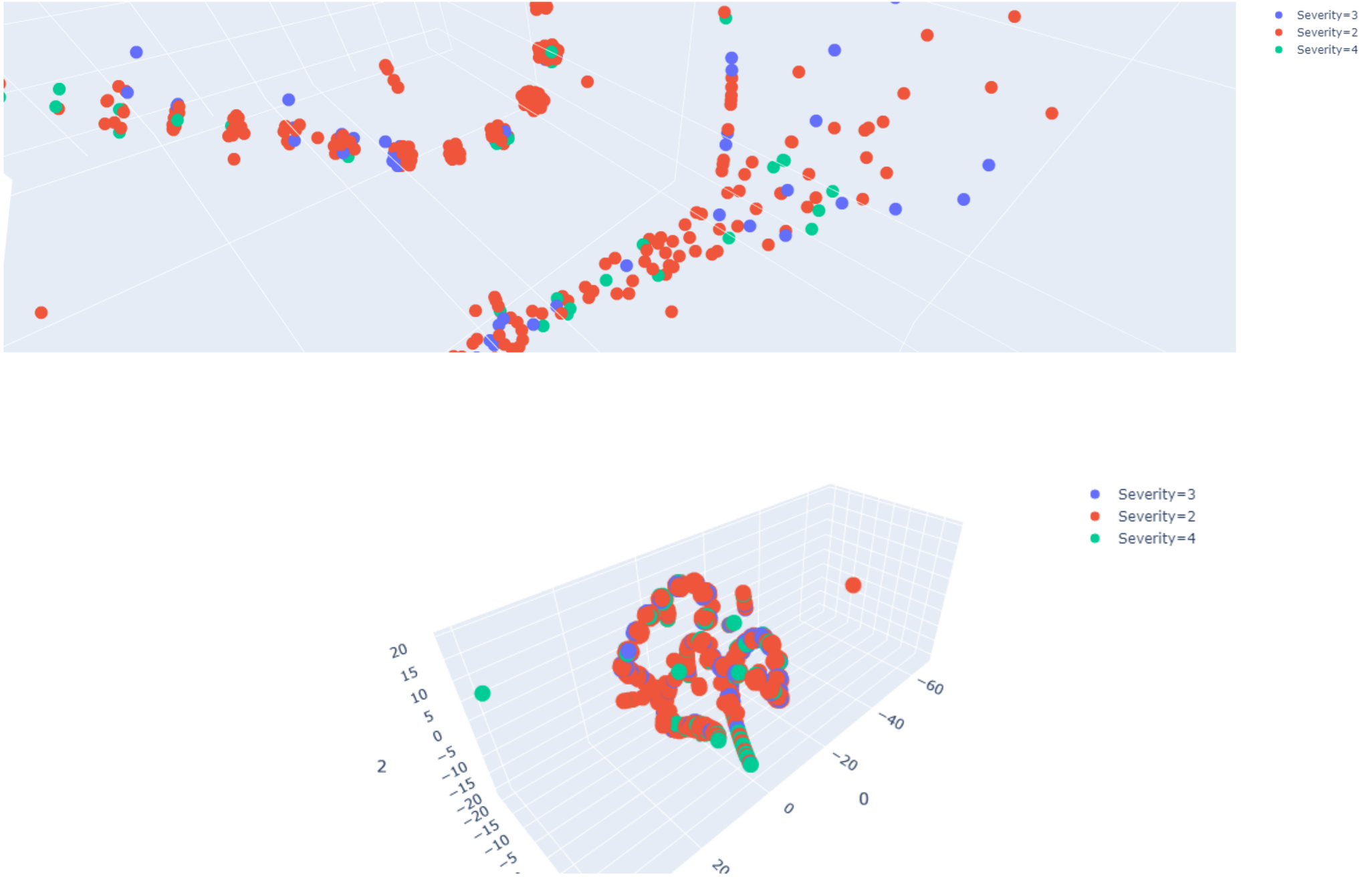
- Detect co-linearity
- Drop high correlated columns(with less correlation with target)
- PCA
- t-SNE
- Create a new pipeline with PCA





	VIF_value
Start_Lat	0.298720
Start_Lng	0.395768
Distance(mi)	1.003610
Street	0.319468
Side	0.992280
City	0.339383
County	0.334377
State_2	0.970163
State_3	1.019652
State_4	0.971017
State_5	1.042308
State_6	0.905838
Zipcode	0.301572
Airport_Code	0.371329
Temperature(F)	0.657045
Humidity(%)	0.857698
Wind_Direction_0	0.995263
Wind_Direction_1	1.000018
Wind_Direction_2	1.000585
Wind_Direction_3	0.999995

	VIF_value
Start_Lat	0.354291
Start_Lng	0.473015
Distance(mi)	1.003830
Street	0.377490
Side	0.995111
City	0.397727
County	0.392307
State	0.343565
Zipcode	0.356443
Timezone	0.196622
Airport_Code	0.431079
Wind_Chill(F)	0.813128
Humidity(%)	0.858612
Wind_Direction	0.006488
Wind_Speed(mph)	0.976578
Weather_Condition	0.004546
Junction	1.003347
Traffic_Signal	1.000375

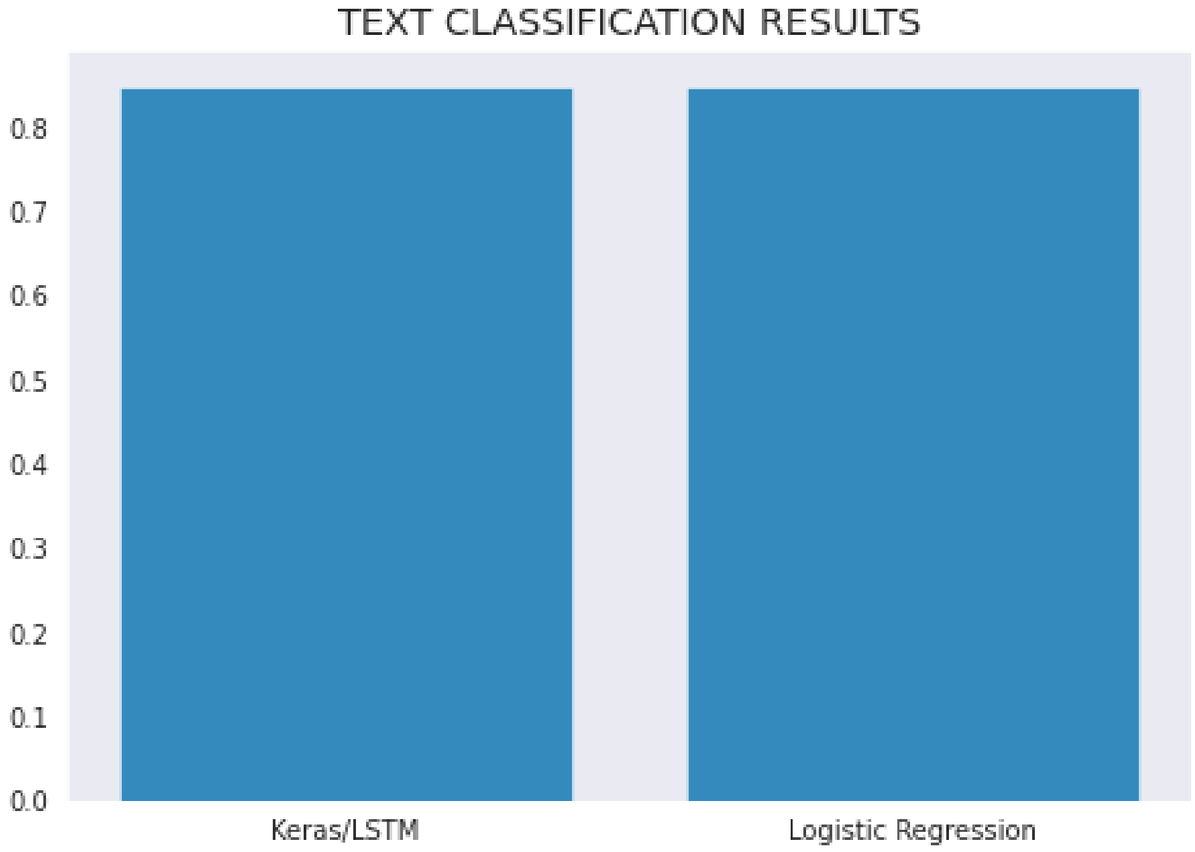

Severity=3
Severity=2
Severity=4
Severity=3
Severity=2
Severity=4
20
15
10
5
0
−5
−10
−15
−20
−20
−15
−10
−5
2
0
−20
−40
−60
20

MACHINE LEARNING

LOGISTIC REGRESSION
accuracy:
0.8492016857911026

KERAS
loss: 0.2637 - accuracy:
0.8474 - val_loss: 0.2581 -
val_accuracy: 0.8489

TEXT CLASSIFICATION




```
Right lane blocked due to accident on I-70 Eastbound at Exit 41 OH-235 State Route 4.
Accident on Brice Rd at Tussing Rd. Expect delays.
Accident on OH-32 State Route 32 Westbound at Dela Palma Rd. Expect delays.
Accident on I-75 Southbound at Exits 52 52B US-35. Expect delays.
Accident on McEwen Rd at OH-725 Miamisburg Centerville Rd. Expect delays.
Accident on I-270 Outerbelt Northbound near Exit 29 OH-3 State St. Expect delays.
Accident on Oakridge Dr at Woodward Ave. Expect delays.
Accident on I-75 Southbound at Exit 54B Grand Ave. Expect delays.
Accident on Notre Dame Ave at Warner Ave. Expect delays.
Right hand shoulder blocked due to accident on I-270 Outerbelt Westbound at Exit 29 OH-3 State St.
Accident on I-270 Outerbelt Northbound at Exits 7 7A 7B US-40 Broad St. Expect delays.
One lane blocked due to accident on I-70 Westbound at Exits 110 110A 110B Brice Rd. Expect delays.
Accident on Revere Ave at Watervliet Ave. Expect delays.
Accident on Salem Ave at Hillcrest Ave / Kensington Dr. Expect delays.
Accident on OH-16 Broad St at James Rd. Expect delays.
Accident on Wayne Ave at Glencoe Ave. Expect delays.
Accident on James H McGee Blvd at US-35. Expect delays.
Accident on Delphos Ave at Brooklyn Ave. Expect delays.
Accident on Stewart St near Rubicon St. Expect delays.
Accident on Hillcrest Ave at Piccadilly Ave. Expect delays.
Accident on Brookhill Dr at Glenhurst Ct.
Accident on Princeton Dr at Catalpa Dr. Expect delays.
Accident on OH-741 Springboro Pike at Lyons Rd. Expect delays.
Accident on I-71 Northbound at Exit 113 Silver Dr. On exit ramp.
Accident on Hoover Ave at Elmhurst Rd.
```

	Description	Severity
0	Right lane blocked due to accident on I-70 Eas...	3
1	Accident on Brice Rd at Tussing Rd. Expect del...	2
2	Accident on OH-32 State Route 32 Westbound at ...	2
3	Accident on I-75 Southbound at Exits 52 52B US...	3
4	Accident on McEwen Rd at OH-725 Miamisburg Cen...	2
...	...	...
3513612	At Market St - Accident.	2
3513613	At Camino Del Rio/Mission Center Rd - Accident.	2
3513614	At Glassell St/Grand Ave - Accident. in the ri...	2
3513615	At CA-90/Marina Fwy/Jefferson Blvd - Accident.	2
3513616	At Highland Ave/Arden Ave - Accident.	2

MACHINE LEARNING

Model Decision;

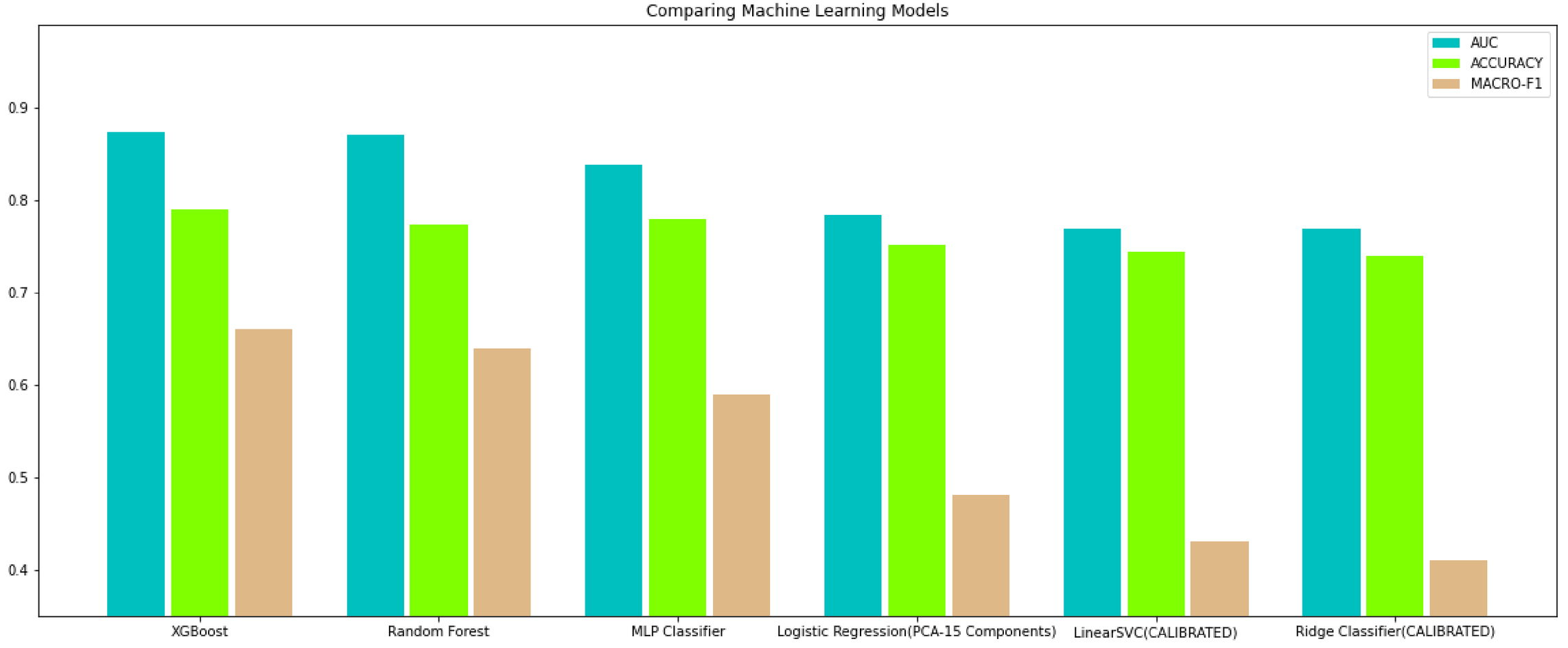
Bulding models;
(pre-tuned)
- XGBoost
- Random Forest
- MLP Classifier
- LinearSVC
- Ridge Classifier

Comparison of Models;
- Multiple bar-chart with AUC / Accuracy / F1-Score / Precision / Recall
- Model selection to be developed

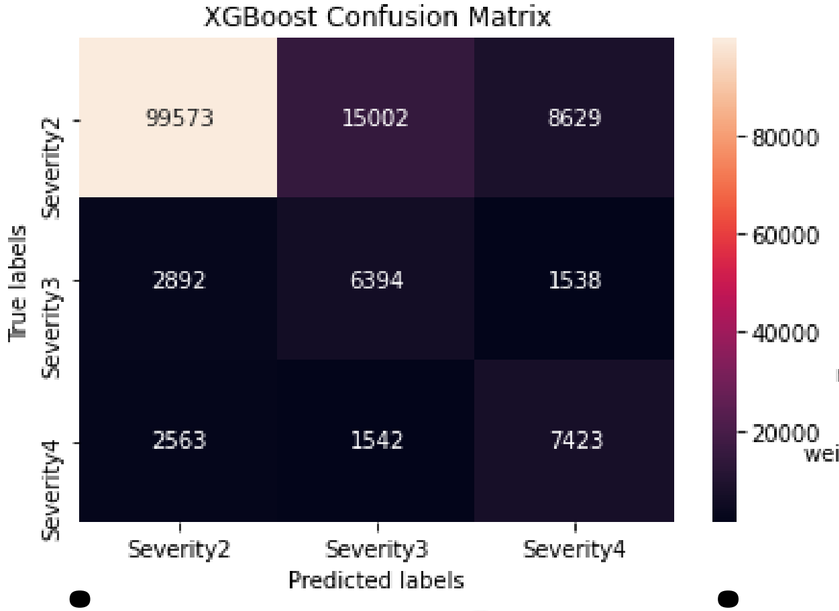
Selected model's results;
- Confusion matrix
- Classification reports
- Learning curve

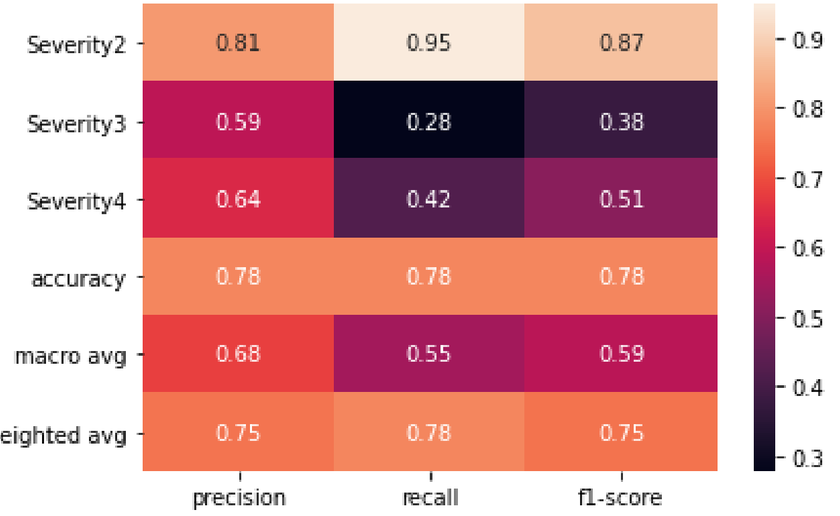
Hyperparameter tuning;
- Best params

Final Results;
- Confusion matrix
- Classification reports
- Learning curve
- Cross Entropy

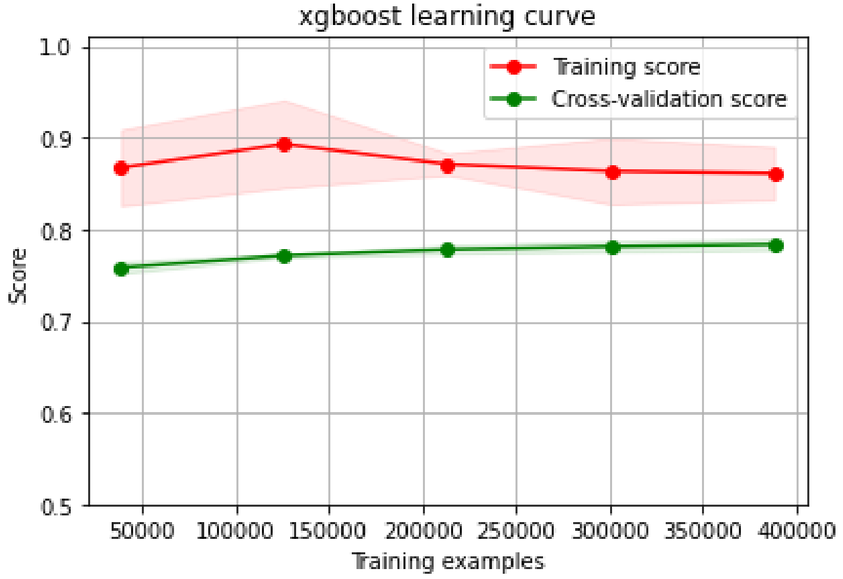

Comparing Machine Learning Models
AUC
ACCURACY
MACRO-F1
0.9
0.8
0.7
0.6
0.5
0.4
XGBoost
Random Forest
MLP Classifier
Logistic Regression(PCA-15 Components)
LinearSVC(CALIBRATED)
Ridge Classifier(CALIBRATED)

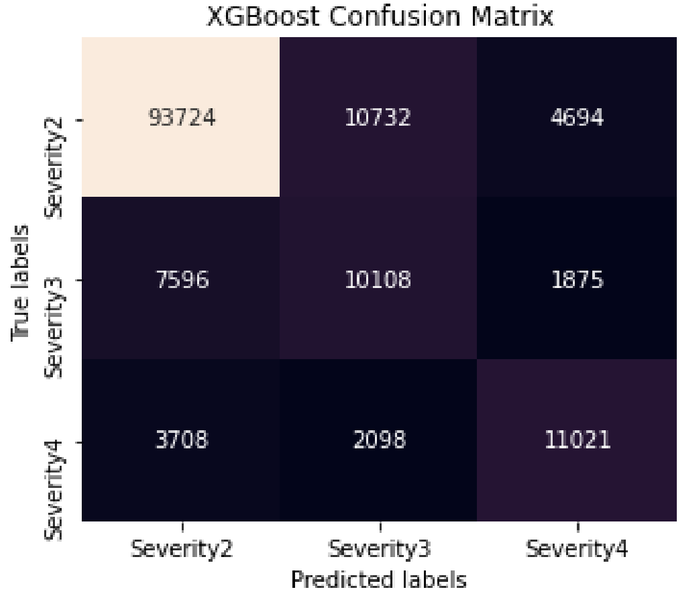
İLK DENEMELER

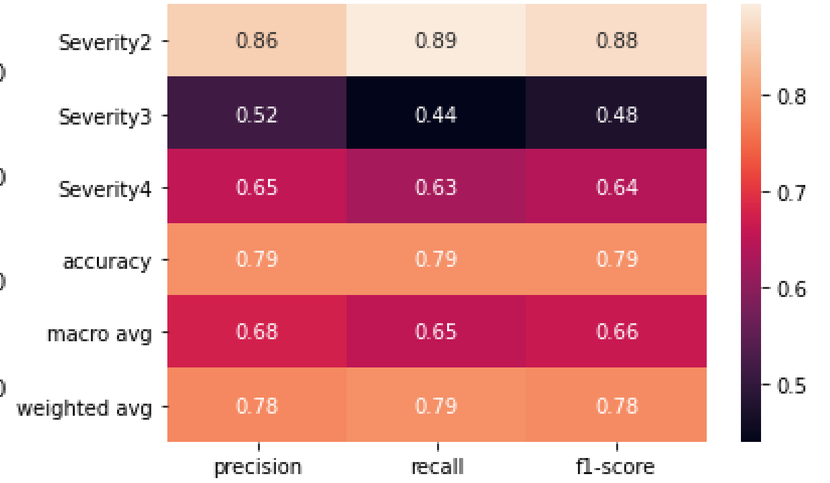


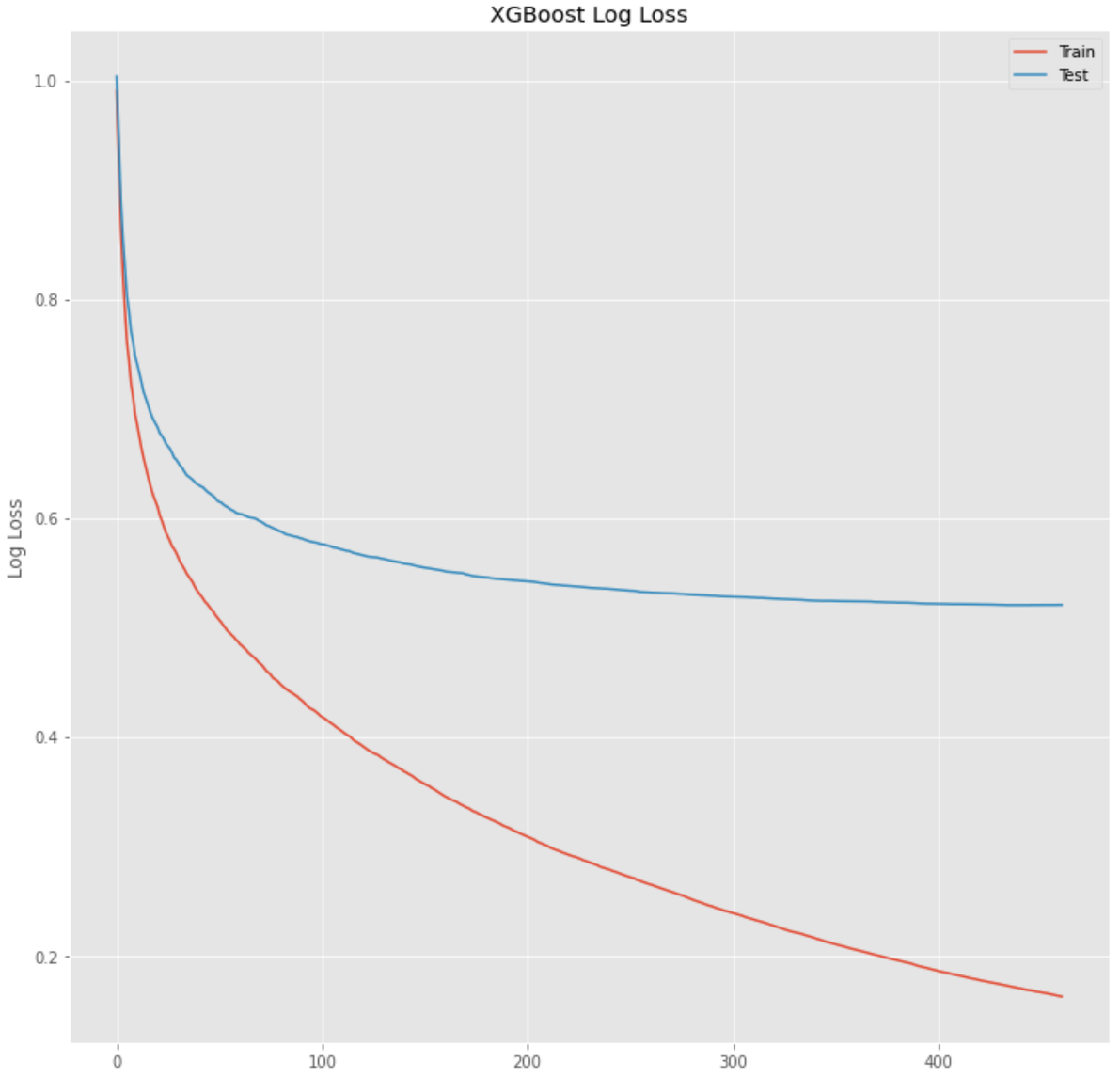


FİNAL MODELİ








XGBoost Log Loss
Train
Test
Log Loss
1.0
0.8
0.6
0.4
0.2
0
100
200
300
400


pandas
scikit
learn
Seaborn
NumPy
SciPy
matplotlib
jupyter
colab
kaggle
TextBlob
K

Thank You

KODLUYORUZ

geleceği kodluyoruz >_

Metin Kerem Ürkmez : metinkeremurkmez@gmail.com
Elif Yüzer : elifyuzer07@gmail.com
Furkan Balıkçı : balikcifurkann@gmail.com
Bora Çolakoğlu : colakoglu.bora35@gmail.com