CODING DOJO

UNITED STATES TRAFFIC ACCIDENTS VS SAUDI ARABIA CASES

Presented by Champions

Champions Team

ZAINAB

THURAIA

ESRAA

LINA

SHATHA

ROBA

AESHA

TABLE OF CONTENT

Introduction

Traffic accidents issues

Related 2030 vision goals

01

Dataset description

US Accidents Dataset

This is a countrywide traffic accident dataset, which covers 49 states

Collected from February 2016 to Dec 2021

There are about 1.5 million accident records and has 47 features

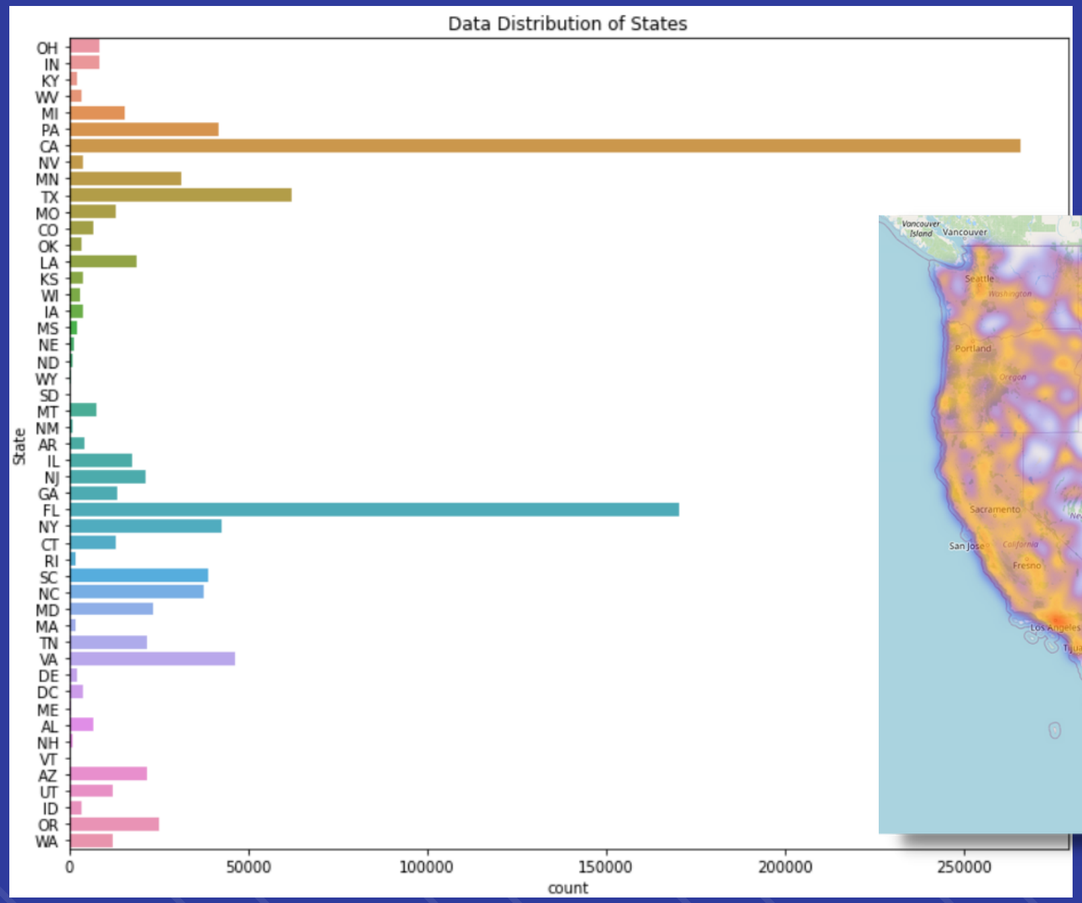
1	ID
2	Severity
3	Start_Time
4	End_Time
5	Start_Lat
6	Start_Lng
7	End_Lat
8	End_Lng
9	Distance(mi)
10	Description
11	Number
12	Street
13	Side
14	City
15	County
16	State
17	Zipcode
18	Country
19	Timezone
20	Airport_Code
21	Weather_Timestamp
22	Temperature(F)

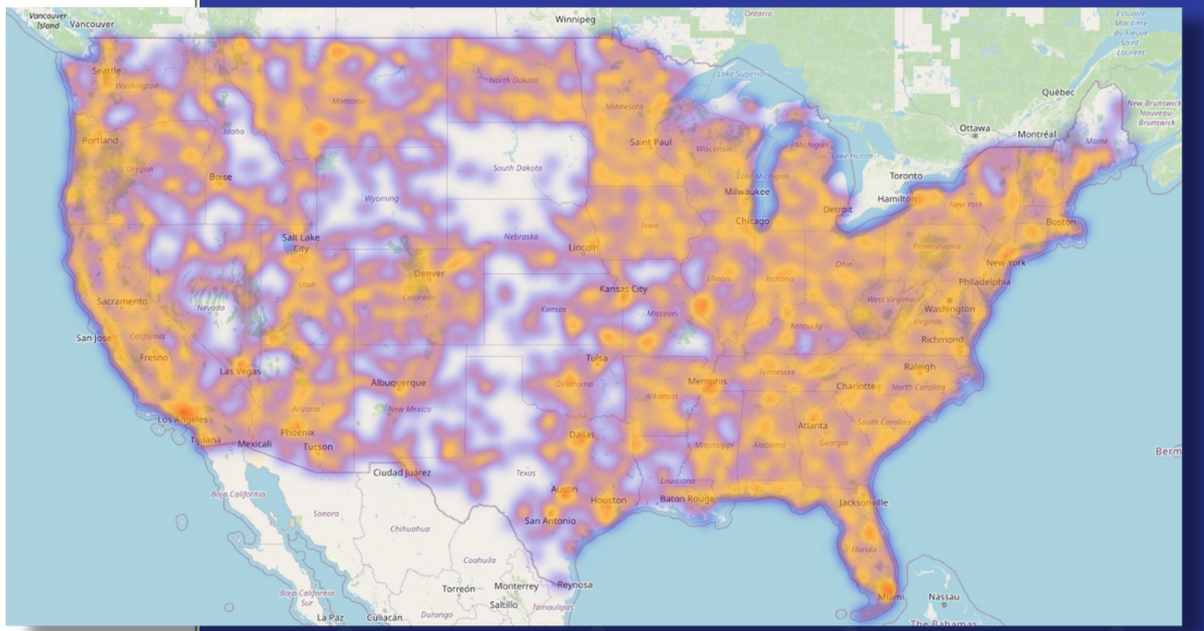
23	Wind_Chill(F)
24	Humidity(%)
25	Pressure(in)
26	Visibility(mi)
27	Wind_Direction
28	Wind_Speed(mph)
29	Precipitation(in)
30	Weather_Condition
31	Amenity
32	Bump
33	Crossing
34	Give_Way
35	Junction
36	No_Exit
37	Railway
38	Roundabout
39	Station
40	Stop
41	Traffic_Calming
42	Traffic_Signal
43	Turning_Loop
44	Sunrise_Sunset
45	Civil_Twilight
46	Nautical_Twilight
47	Astronomical_Twilight

03

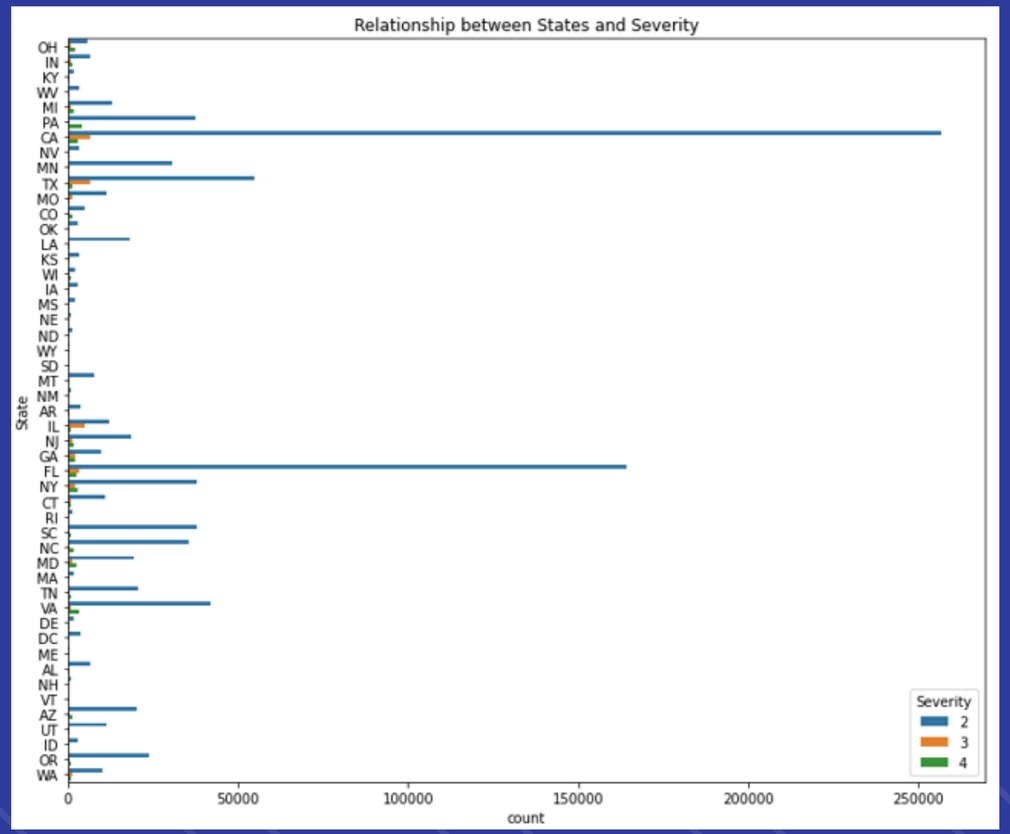
Exploratory Data Analysis EDA

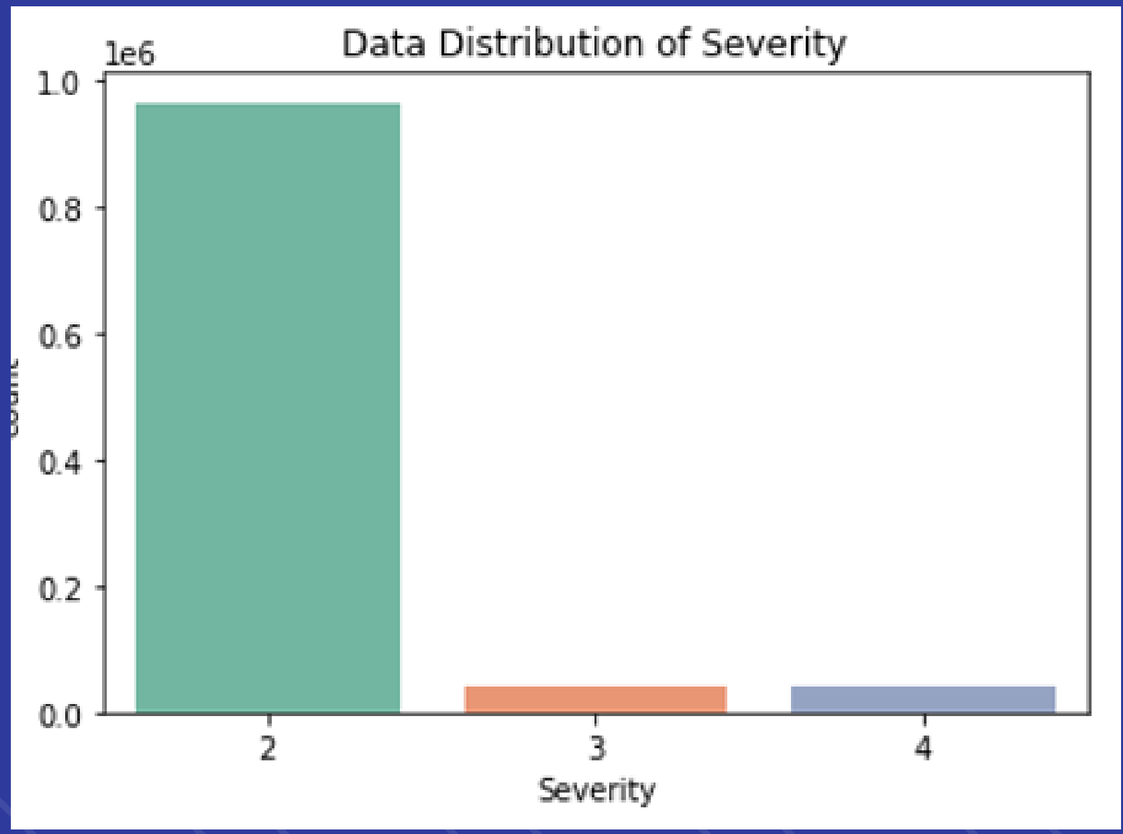
US Accidents Locations



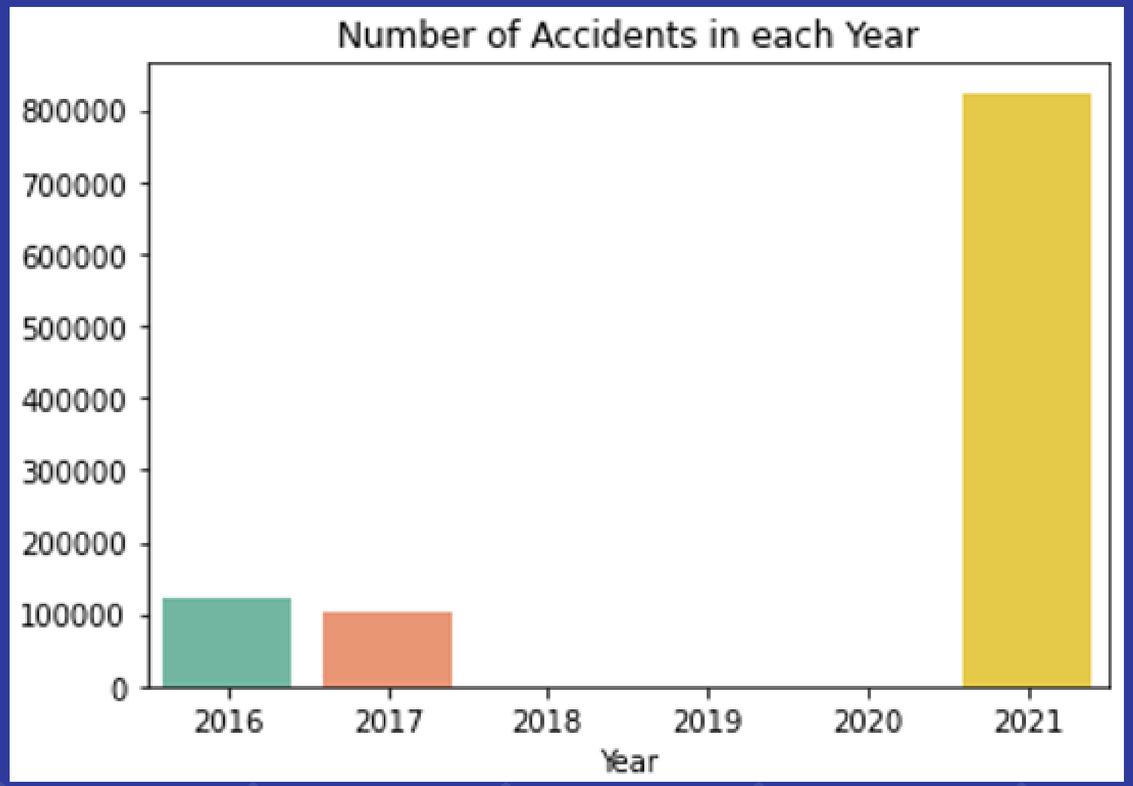


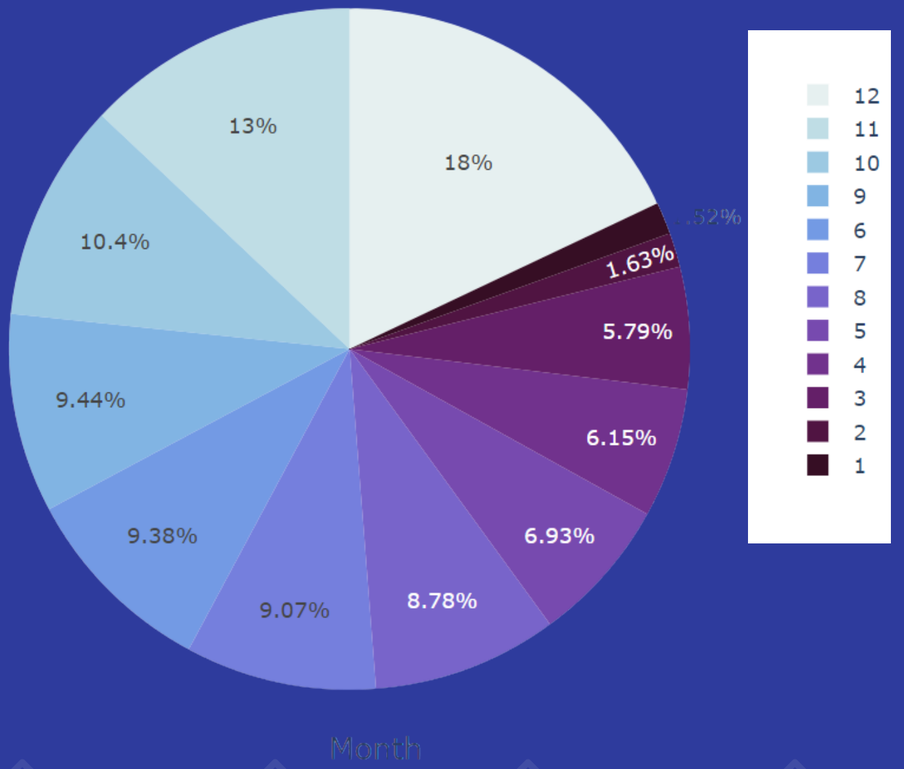
The Distribution of Severity Values





Accidents in each Year & month



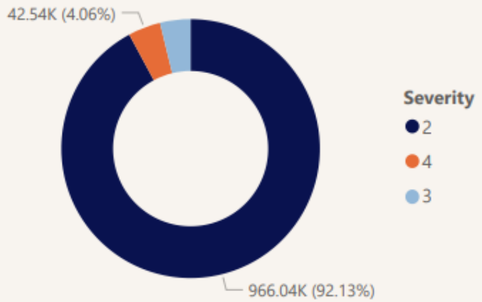


02

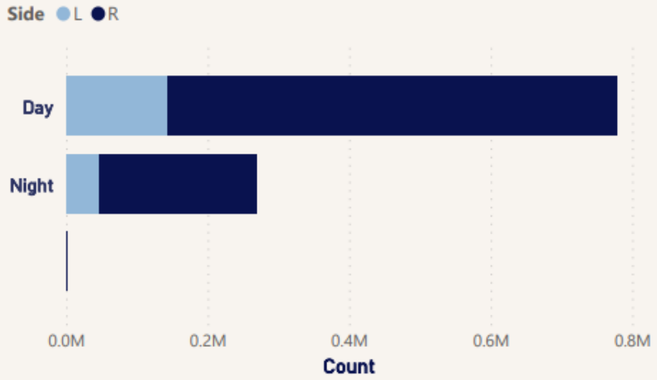
Dashboard

US Accidents

Severity Distribution



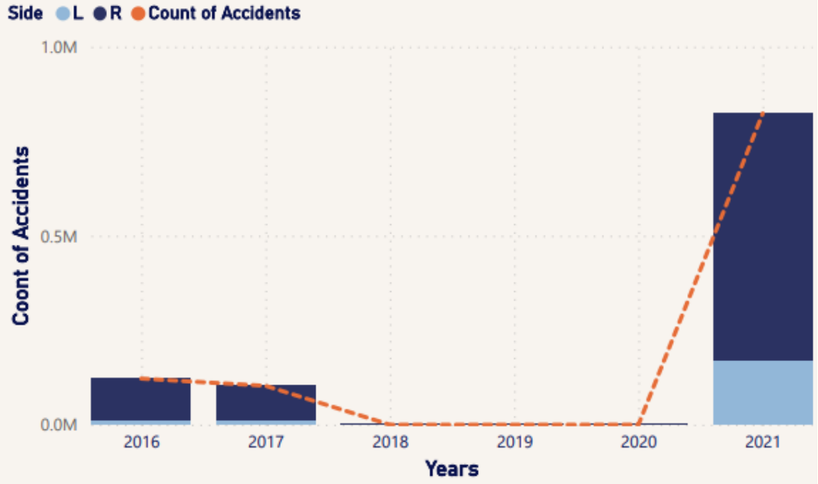
Civil Twilight and Side

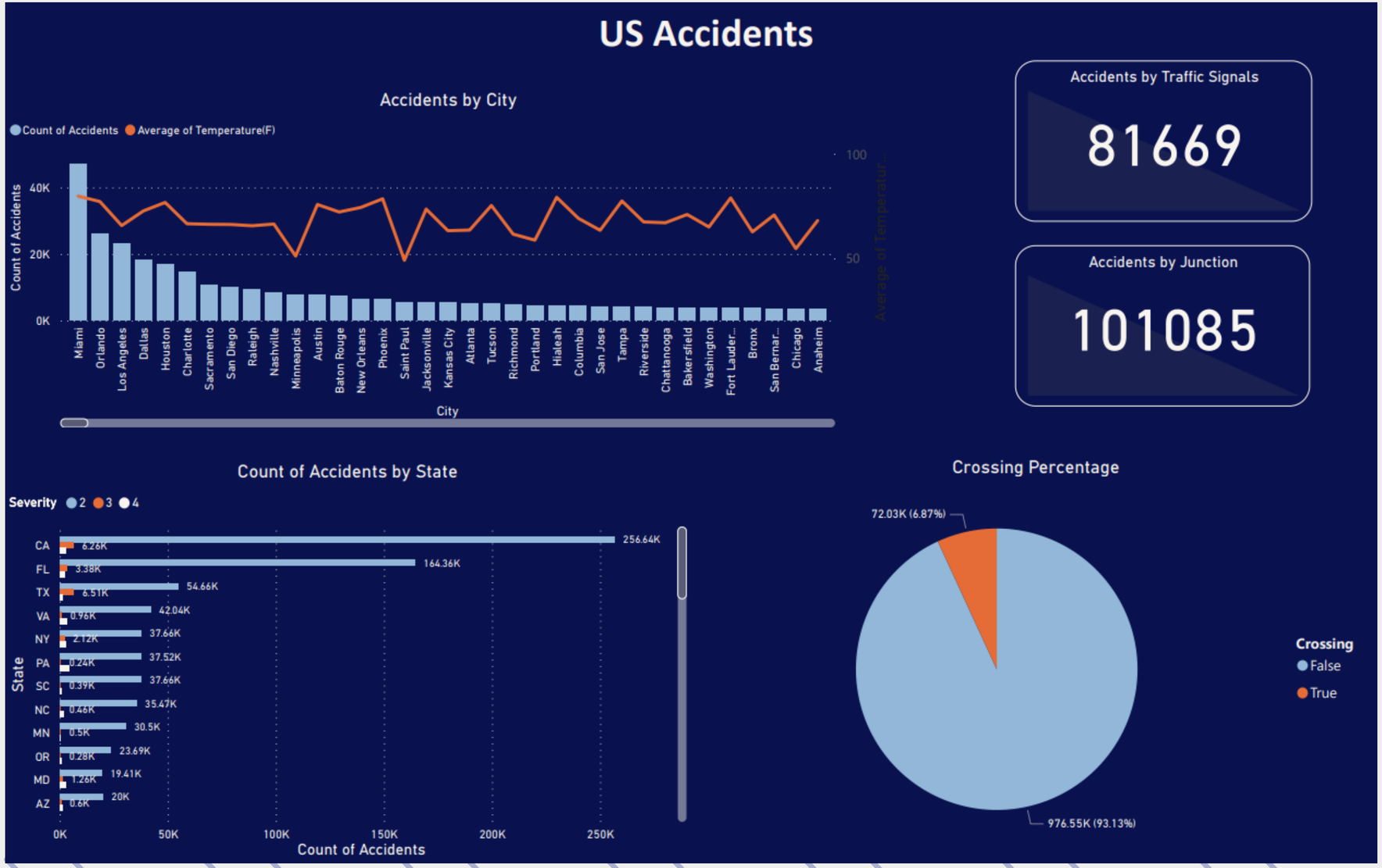


9.19
Average of Visibility(mi)

1.05M
Count of Accidents

Total Accidents per Year




US Accidents
Accidents by City
Count of Accidents
Average of Temperature(F)
Count of Accidents
40K
20K
0K
100
50
Average of Temperatur...
Miami
Orlando
Los Angeles
Dallas
Houston
Charlotte
Sacramento
San Diego
Raleigh
Nashville
Minneapolis
Austin
Baton Rouge
New Orleans
Phoenix
Saint Paul
Jacksonville
Kansas City
Atlanta
Tucson
Richmond
Portland
Hialeah
Columbia
San Jose
Tampa
Riverside
Chattanooga
Bakersfield
Washington
Fort Lauder...
Bronx
San Bernar...
Chicago
Anaheim
City
Accidents by Traffic Signals
81669
Accidents by Junction
101085
Count of Accidents by State
Severity 2 3 4
State
CA 6.26K 256.64K
FL 3.38K 164.36K
TX 6.51K 54.66K
VA 0.96K 42.04K
NY 2.12K 37.66K
PA 0.24K 37.52K
SC 0.39K 37.66K
NC 0.46K 35.47K
MN 0.5K 30.5K
OR 0.28K 23.69K
MD 1.26K 19.41K
AZ 0.6K 20K
0K
50K
100K
150K
200K
250K
Count of Accidents
Crossing Percentage
72.03K (6.87%)
976.55K (93.13%)
Crossing
False
True

04

Pig Latin Exploratory Data Analysis

Q 1

Find the total number of accidents by month of 2019, 2020, and 2021. For the highest populated US states, which are California CA, Texas TX, Florida FL, New York NY, and Pennsylvania PA. Then compare the results with the rate of accidents in the highest populated KSA regions

Accidents_Count:

(2020,8,PA,1)
(2020,10,CA,1)
(2020,10,FL,1)
(2020,11,FL,1)
(2021,2,CA,367)
(2021,2,FL,198)
(2021,2,NY,15)
(2021,2,PA,88)
(2021,2,TX,85)
(2021,3,CA,11392)
(2021,3,FL,7641)
(2021,3,NY,1433)
(2021,3,PA,2031)
(2021,3,TX,2242)
(2021,4,CA,12140)
(2021,4,FL,7680)
(2021,4,NY,1925)
(2021,4,PA,2093)
(2021,4,TX,2627)
(2021,5,CA,12343)
(2021,5,FL,8183)
(2021,5,NY,2430)
(2021,5,PA,2471)
(2021,5,TX,3239)
(2021,6,CA,18773)
(2021,6,FL,12772)
(2021,6,NY,3167)
(2021,6,PA,3480)
(2021,6,TX,4848)
(2021,7,CA,17841)
(2021,7,FL,11520)
(2021,7,NY,2886)
(2021,7,PA,2982)
(2021,7,TX,4519)
(2021,8,CA,18872)
(2021,8,FL,12673)
(2021,8,NY,2927)
(2021,8,PA,3491)
(2021,8,TX,5247)
(2021,9,CA,21230)
(2021,9,FL,16220)
(2021,9,NY,2931)
(2021,9,PA,3818)
(2021,9,TX,4957)
(2021,10,CA,23397)
(2021,10,FL,18424)
(2021,10,NY,3238)
(2021,10,PA,3837)
(2021,10,TX,4679)
(2021,11,CA,27481)
(2021,11,FL,24470)
(2021,11,NY,4482)
(2021,11,PA,5178)
(2021,11,TX,5849)
(2021,12,CA,44850)
(2021,12,FL,32838)
(2021,12,NY,5468)
(2021,12,PA,5869)
(2021,12,TX,6333)

Q 1

Then compare the results with the rate of accidents in the highest populated KSA regions

Region	No. of Accidents
Riyadh	147568
Makkah	126537
Madinah	19058
Al-Qasim	24273
Eastern	88065
Aseer	32163
Tabouk	20638
Hael	8415
Northern Boarders	13076
Jazan	22229
Najran	3220
Al-Baaha	4166
Al-Jowf	9387
Total	**518795**

THE STATISTICS OF THE SAUDI GENERAL AUTHORITY FOR TRAFFIC ACCIDENTS SITE BY REGION

Q 2

Find the number of accidents in each state where the accident side is “R”, and the Amenity is labeled as “FALSE”. Group By the Weather_Condition.

(Fair,315139)
(Mostly Cloudy,113455)
(Cloudy,85749)
(Partly Cloudy,78035)
(Clear,72156)
(Light Rain,37884)
(Overcast,31221)
(Scattered Clouds,20139)
(,19485)
(Haze,11263)
(Fog,10318)
(Light Snow,10253)
(Rain,9794)
(Heavy Rain,3963)
(Fair / Windy,3790)
(Thunder in the Vicinity,2589)
(T-Storm,2503)
(Thunder,2255)
(Smoke,2210)
(Light Rain with Thunder,2074)
(Light Drizzle,1818)
(Cloudy / Windy,1602)
(Heavy T-Storm,1447)
(Snow,1419)
(Mostly Cloudy / Windy,1359)
(Partly Cloudy / Windy,856)
(Light Rain / Windy,834)
(Light Thunderstorms and Rain,577)
(Light Snow / Windy,481)
(Sand / Dust Whirlwinds,6)
(Rain Shower,6)
(Light Sleet,6)
(Blowing Snow / Windy,5)
(Light Snow Showers,4)
(Squalls / Windy,4)
(Volcanic Ash,4)
(Light Sleet / Windy,3)
(Small Hail,3)
(Light Thunderstorms and Snow,3)
(Heavy Thunderstorms and Snow,3)
(Snow Grains,3)
(Heavy Ice Pellets,2)
(Drizzle / Windy,2)
(Snow and Sleet,2)
(Partial Fog,1)
(Thunder / Wintry Mix,1)
(Duststorm,1)
(Light Snow with Thunder,1)
(Dust Whirls,1)
(Hail,1)
(Funnel Cloud,1)
(Snow and Thunder / Windy,1)
(Heavy Rain Shower / Windy,1)
(Light Freezing Rain / Windy,1)
(Blowing Snow Nearby,1)
(Sand,1)
(Light Fog,1)
(Freezing Rain / Windy,1)

Q 3

The highest temperature recorded in KSA was between 113 – 127 (F) degrees. Count the number of accidents when the temperature is in the range between 100 to 127 degrees Fahrenheit, and when the traffic calm is 'FALSE'. Then, conclude if there is a relation between the temperature and the increased probability of accidents.

Accidents_Count:

(100.0,1186)	(106.9,3)
(100.2,3)	(107.0,182)
(100.4,109)	(107.1,40)
(100.6,1)	(107.2,1)
(100.8,2)	(107.6,22)
(100.9,118)	(108.0,202)
(101.0,587)	(109.0,163)
(101.1,1)	(109.4,17)
(101.3,1)	(109.6,1)
(101.5,2)	(109.8,1)
(101.7,1)	(109.9,26)
(101.8,2)	(110.0,90)
(102.0,897)	(111.0,115)
(102.2,76)	(111.2,10)
(102.4,1)	(111.9,12)
(102.6,1)	(112.0,55)
(102.7,1)	(113.0,40)
(102.9,110)	(114.0,28)
(103.0,446)	(114.1,8)
(103.5,5)	(114.8,4)
(104.0,595)	(115.0,19)
(104.4,1)	(116.0,4)
(104.7,1)	(116.1,1)
(104.9,1)	(116.6,2)
(105.0,246)	(117.0,12)
(105.1,68)	(118.0,3)
(105.4,2)	(118.4,3)
(105.8,34)	(119.0,4)
(106.0,365)	
(106.3,1)	
(106.5,1)	

Q 4

Find the minimum and maximum temperatures when accidents occurred in the following states: TX, GA, AL, and FL. Group by the state

```
(AL,196.0,25.0)
(FL,156.0,27.0)
(GA,97.0,18.0)
(TX,129.2,14.0)
```

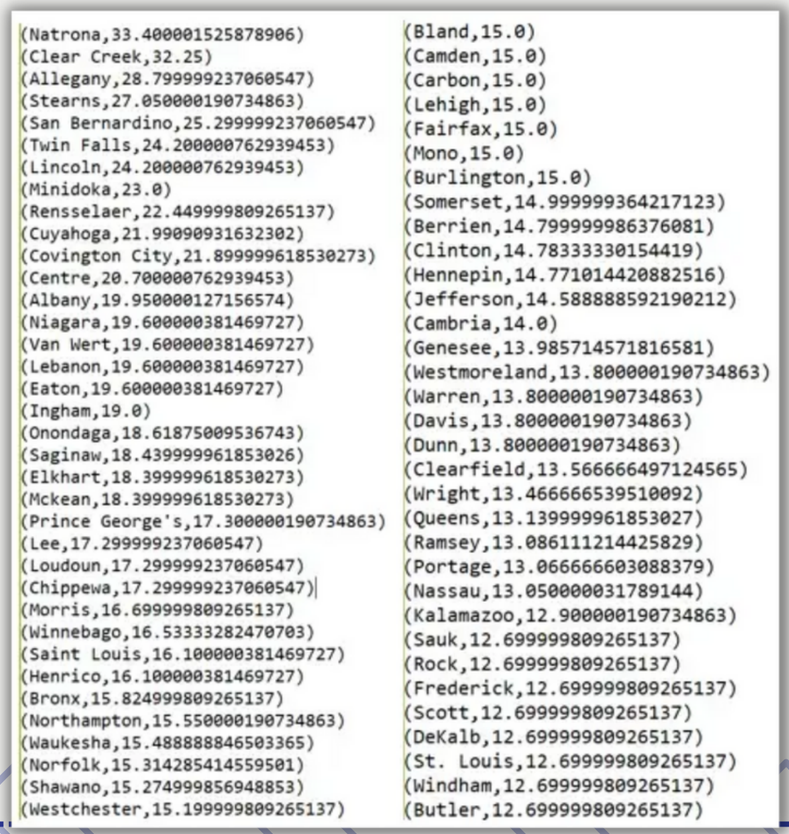
Q 5

The sand cover in the Kingdom of Saudi Arabia constitutes 34% of the total area of the Kingdom, which increases the activity and speed of sandstorms. Wind speed varies during sandstorms in Saudi Arabia, and increases to range between 54-63 km/h and low humidity(%). For each severity level, find how many accidents happened when the wind speed(mph) is 20 or above and humidity is less than 30%

Accidents_Count:
(2,3158)
(3,114)
(4,157)

Q 6

What is the average wind speed for accidents in each county when the wind is blowing from the west and the temperature is 25 degrees or lower?



```
(Natrona,33.400001525878906)
(Clear Creek,32.25)
(Allegany,28.799999237060547)
(Stearns,27.050000190734863)
(San Bernardino,25.299999237060547)
(Twin Falls,24.200000762939453)
(Lincoln,24.200000762939453)
(Minidoka,23.0)
(Rensselaer,22.449999809265137)
(Cuyahoga,21.99090931632302)
(Covington City,21.899999618530273)
(Centre,20.700000762939453)
(Albany,19.950000127156574)
(Niagara,19.600000381469727)
(Van Wert,19.600000381469727)
(Lebanon,19.600000381469727)
(Eaton,19.600000381469727)
(Ingham,19.0)
(Onondaga,18.61875009536743)
(Saginaw,18.439999961853026)
(Elkhart,18.399999618530273)
(Mckean,18.399999618530273)
(Prince George's,17.300000190734863)
(Lee,17.299999237060547)
(Loudoun,17.299999237060547)
(Chippewa,17.299999237060547)
(Morris,16.699999809265137)
(Winnebago,16.53333282470703)
(Saint Louis,16.100000381469727)
(Henrico,16.100000381469727)
(Bronx,15.824999809265137)
(Northampton,15.550000190734863)
(Waukesha,15.488888846503365)
(Norfolk,15.314285414559501)
(Shawano,15.274999856948853)
(Westchester,15.199999809265137)
(Bland,15.0)
(Camden,15.0)
(Carbon,15.0)
(Lehigh,15.0)
(Fairfax,15.0)
(Mono,15.0)
(Burlington,15.0)
(Somerset,14.999999364217123)
(Berrien,14.799999986376081)
(Clinton,14.78333330154419)
(Hennepin,14.771014420882516)
(Jefferson,14.588888592190212)
(Cambria,14.0)
(Genesee,13.985714571816581)
(Westmoreland,13.800000190734863)
(Warren,13.800000190734863)
(Davis,13.800000190734863)
(Dunn,13.800000190734863)
(Clearfield,13.566666497124565)
(Wright,13.466666539510092)
(Queens,13.139999961853027)
(Ramsey,13.086111214425829)
(Portage,13.066666603088379)
(Nassau,13.050000031789144)
(Kalamazoo,12.900000190734863)
(Sauk,12.699999809265137)
(Rock,12.699999809265137)
(Frederick,12.699999809265137)
(Scott,12.699999809265137)
(DeKalb,12.699999809265137)
(St. Louis,12.699999809265137)
(Windham,12.699999809265137)
(Butler,12.699999809265137)
```

Q 7

How many accidents occurred during the sunset when the humidity percentage was between 85 and 100?

```
(85683)
```

Q 8

How many accidents with a distance of the car drifting to a stop of more than 3 miles and with the highest severity of 4?

(5188)

Q 9

How many accidents happened that had a distance of 0.5 miles or more and with a severity of 2 or more? Split the results for each accident side and save each of them in a separate dataset.

(L, 27288)
(R, 392535)

Q10

What is the average distance for accidents for each side?

```
(L,0.3106229907573987)
(R,0.9159718676994905)
```

Q 1 1

Show accidents that occurred with their severity, month and year that have the weather condition as ‘Sand’ or ‘Sand / Dust Whirlwinds’ or ‘Sand / Windy’ in ascending order.

```
(2,4,2021,Sand / Dust Whirlwinds)
(2,4,2021,Sand / Dust Whirlwinds)
(2,6,2021,Sand / Dust Whirlwinds)
(2,7,2016,Sand)
(2,9,2021,Sand / Dust Whirlwinds)
(2,9,2021,Sand / Dust Whirlwinds)
(2,9,2021,Sand / Dust Whirlwinds)
(2,10,2021,Sand / Windy)
(2,10,2021,Sand / Windy)
```

Q 1 2

Find the total number of accidents in these cities Cleveland, Westerville, Fairdale, Youngstown, and Lake Forest when the weather is Fog , Fog / Windy, and the Visibility is less or equal to 0.5 mi group by year.

```
(2016,2)
(2017,1)
(2021,11)
```

Q 1 3

What is the Maximum and Minimum visibility value for accidents that occurred at 'night'. Group the result by year.

```
(2016,60.0,0.0)
(2017,60.0,0.0)
(2020,10.0,0.75)
(2021,140.0,0.0)
```

Q 1 4

What is the Maximum and Minimum visibility value for accidents that occurred during the 'day'? Group the result by year.

```
(2016,111.0,0.0)
(2017,80.0,0.0)
(2018,10.0,10.0)
(2019,10.0,10.0)
(2020,10.0,3.0)
(2021,100.0,0.0)
```

Q 1 5

What is the Average wind speed where the severity is 2 and Junction is True? Group by month.

```
(1,9.246437009520664)
(2,9.758325313236762)
(3,9.188787361927856)
(4,9.007891312005821)
(5,8.796491994434396)
(6,8.275506244828025)
(7,7.768450184093378)
(8,7.304615847548962)
(9,7.301613157509611)
(10,7.4693058304777376)
(11,6.938817196750652)
(12,7.33798948111484)
```

Q 1 6

Find the total number of accidents where the weather condition is cloudy and the state: TX, PX, and FL sort the result by year.

```
(2016,13)
(2017,7)
(2021,18514)
```

Q 1 7

What is the Average precipitation for the occurred accidents grouped by month?

```
(1,0.077831463760254 13)
(2,0.036029100425975034)
(3,0.007879681612952693)
(4,0.006063442599513 12)
(5,0.006773309463624175)
(6,0.010632766632945056)
(7,0.010865065153074855)
(8,0.009269080263029672)
(9,0.00807935677003483)
(10,0.006601139358833231)
(11,0.004944148444712884)
(12,0.007487110442641 4165)
```

Q 1 8

Find the total number of accidents where the state is Texas or Florida or Alabama and the visibility is 6. Sort the result by the weather condition

```
(Light Rain,473)
(Cloudy,303)
(Haze,260)
(Rain,241)
(Fair,148)
(T-Storm,128)
(Mostly Cloudy,119)
(Overcast,118)
(Light Rain with Thunder,114)
(Partly Cloudy,69)
(Clear,50)
(Heavy Rain,30)
(Thunder,24)
(Light Thunderstorms and Rain,21)
(Scattered Clouds,17)
(Light Drizzle,17)
(Thunderstorms and Rain,12)
(Heavy T-Storm,11)
(Light Rain / Windy,8)
(Shallow Fog,7)
(Thunder in the Vicinity,5)
(Heavy Thunderstorms and Rain,5)
(Haze / Windy,4)
(Widespread Dust,4)
(Rain / Windy,3)
(Light Snow,2)
(Smoke,2)
(,2)
(T-Storm / Windy,1)
```

Q 1 9

How many accidents occurred where there was a roundabout?

(47)

Q 2 0

Number of accidents where the Stop sign is True

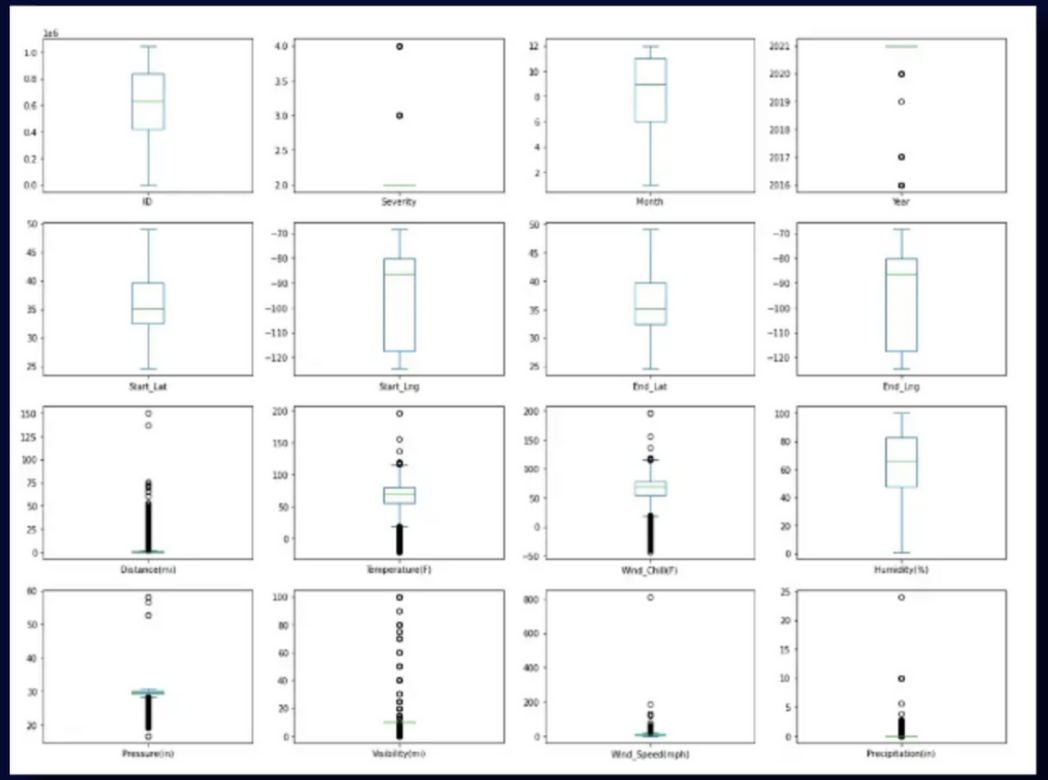
```
(19075)
```

06

Dataset Preprocessing

Handling outliers

THE FIRST STEP WE DID IS DETECTING OUTLIERS IN THE COLUMNS USING A BOX PLOT.



Handling outliers

DEFINED A FUNCTION CALLED 'OUTLIERS'

```
def outliers(df,ft):
  Q1 =df[ft].quantile(0.25)
  Q3 =df[ft].quantile(0.75)
  IQR = Q3 - Q1

  lower_bound = Q1 - 1.5 * IQR
  upper_bound = Q3 + 1.5 * IQR

  ls = df.index[ (df[ft] < lower_bound) | (df[ft] > upper_bound) ]

  return ls
```

Handling outliers

CREATING AN EMPTY LIST

```
index_list = []
for feature in ['Distance(mi)','Wind_Chill(F)','Pressure(in)','Visibility(mi)',
  index_list.extend(outliers(df, feature))
```

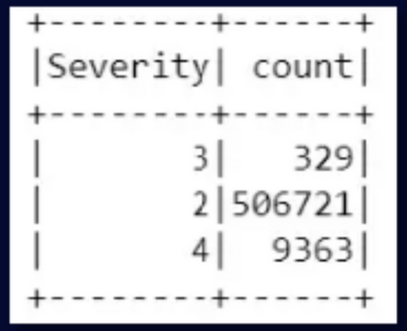
DEFINE A FUNCTION CALLED 'REMOVE'

```
def remove(df, ls):
  ls = sorted(set(ls))
  df = df.drop(ls)
  return df
```

Implementing oversampling

WE WILL OVERSAMPLE THE FIRST AND SECOND CATEGORY.

ISOLATE ALL THE DATA WHERE THE SEVERITY IS 2 AND THEN THE DATA WHERE THE SEVERITY IS 3



Severity	count
3	329
2	506721
4	9363

```
severity2_df = sparkDF.filter(col("Severity") == 2)
severity3_df = sparkDF.filter(col("Severity") == 3)
```

Implementing oversampling

CALCULATE THE RATIO TO DETERMINE THE DIFFERENCE

```
ratio = int(severity2_df.count()/severity3_df.count())
print("ratio: {}".format(ratio))
```

```
ratio: 1540
```

WE CAN NOW SEE THAT THE SEVERITY3_DF HAS A NEW SIZE WHICH IS QUITE SIMILAR TO SEVERITY2_DF

```
#duplicate the minority rows
oversampled_df = severity3_df.withColumn("dummy", explode(array([lit(x) for x in a]
```

```
3  oversampled_df.count()
```

```
506660
```

Implementing oversampling

COMBINE BOTH ROWS.

```
combined_df = severity2_df.unionAll(oversampled_df)
```

CHECKING THE RATIO.

```
#Lets isolate all the data where the Severity are 2 and then the data where the Severity are 4.

severity2_df = sparkDF.filter(col("Severity") == 2)
severity4_df = sparkDF.filter(col("Severity") == 4)


#we will calculate the ratio to determine the difference between the number of each severity.
ratio = int(severity2_df.count()/severity4_df.count())
print("ratio: {}".format(ratio))
```

```
ratio: 54
```

THE COUNT OF RECORDS FOR EACH VALUE IN THE ‘SEVERITY’ COLUMN

```
+--------+------+
|Severity| count|
+--------+------+
|       2|506721|
|       3|506660|
|       4|505602|
+--------+------+
```

07

Apache PySpark Machine Learning models

- Classification -

COMBINE FEATURE COLUMNS

```
cols = final_df.columns #extract the column names from the dataframe
cols.remove('Severity') #remove severity -> we need this to be our label

#vector assembler will take all the columns and convert them into one column called features
assembler = VectorAssembler(inputCols=cols, outputCol='features', handleInvalid = "skip")

#the .transform will apply the changes here
final_df = assembler.transform(final_df)
```

SPLIT DATA

```
# We will now create a new dataframe only consisting of the features column and the label column
df_data = final_df.select(col('features'), col('severity').alias('label'))

#simple data splitting
df_train, df_test = df_data.randomSplit([0.8, 0.2])
```

DECISION TREE CLASSIFIER MODEL

INITIALIZE & FIT THE MODEL

```
# Decision Tree
dt = DecisionTreeClassifier(labelCol="label", featuresCol="features")
model_dt = dt.fit(df_train)
```

MODEL'S ACCURACY

```
from pyspark.ml.evaluation import *

dt_eval= MulticlassClassificationEvaluator(predictionCol="prediction", labelCol="label")
dt_ACC  = dt_eval.evaluate(pred_dt, {dt_eval.metricName:"accuracy"})
print("Decision Tree Performance Measure")
print("Accuracy = %0.2f" % dt_ACC)
```

```
Decision Tree Performance Measure
Accuracy = 0.80
```

MODEL'S CONFUSION MATRIX

prediction_label	2	3	4
0	2.0	64504	0	19824
1	3.0	227	100360	2432
2	4.0	36294	0	77314

RANDOM FOREST CLASSIFIER MODEL

INITIALIZE & FIT THE MODEL

```
# Random Forest
rf = RandomForestClassifier(labelCol="label", featuresCol="features", numTrees=10)
model_rf = rf.fit(df_train)
```

MODEL'S ACCURACY

```
rf_eval= MulticlassClassificationEvaluator(predictionCol="prediction", labelCol="label")
rf_ACC  = rf_eval.evaluate(pred_rf, {rf_eval.metricName:"accuracy"})
print("Random Forest Performance Measure")
print("Accuracy = %0.2f" % rf_ACC)
```

```
Random Forest Performance Measure
Accuracy = 0.77
```

MODEL'S CONFUSION MATRIX

	prediction_label	2	3	4
0	2.0	58905	927	22745
1	3.0	1181	99433	2633
2	4.0	40939	0	74192

LOGISTIC REGRESSION MODEL

INITIALIZE & FIT THE MODEL

```
# Logistic Regression
lr = LogisticRegression(maxIter=10, labelCol="label", featuresCol="features")
model_lr = lr.fit(df_train)
```

MODEL'S ACCURACY

```
lr_eval= MulticlassClassificationEvaluator(predictionCol="prediction", labelCol="label")
lr_ACC  = lr_eval.evaluate(pred_lr, {lr_eval.metricName:"accuracy"})
print("Logistic Regression Performance Measure")
print("Accuracy = %0.2f" % lr_ACC)
```

```
Logistic Regression Performance Measure
Accuracy = 0.75
```

MODEL'S CONFUSION MATRIX

prediction_label	2	3	4
0	2.0	65317	0	36479
1	3.0	234	100360	3135
2	4.0	35474	0	59956

NAIVE BAYES MODEL

INITIALIZE & FIT THE MODEL

```
# Naïve Bayes
nb = NaiveBayes(labelCol="label", featuresCol="features", modelType='multinomial')
model_nb = nb.fit(df_train)
```

MODEL'S ACCURACY

```
nb_eval= MulticlassClassificationEvaluator(predictionCol="prediction", labelCol="label")
nb_ACC  = nb_eval.evaluate(pred_nb, {nb_eval.metricName:"accuracy"})
print("Naïve Bayes Performance Measure")
print("Accuracy = %0.2f" % nb_ACC)
```

```
Naïve Bayes Performance Measure
Accuracy = 0.13
```

MODEL'S CONFUSION MATRIX

	prediction_label	2	3	4
2	0.0	12292	3406	6923
1	1.0	49422	83100	45762
0	2.0	39311	13854	46885

08

Apache PySpark ML models Comparison

ML Evaluation as Dataframe

COMPARING ALL OF OUR MODELS USING DIFFERENT METRICS

	Accuracy	F1-Score	Weighted Precision	Weighted Recall
Decision Tree	0.804698	0.802729	0.806786	0.804698
Random Forest	0.772640	0.769755	0.773808	0.772640
Logistic Regression	0.749723	0.747623	0.745891	0.749723
Naïve Bayes	0.130621	0.131254	0.131894	0.130621

09

Best Apache PySpark ML model improvement

Model Optimization

THE DECISION TREE CLASSIFIER IS OPTIMIZED WITH A MAXIMUM DEPTH OF 19

```
# Decision Tree - with the best params -> max depth = 19
dt = DecisionTreeClassifier(labelCol="label", featuresCol="features", maxDepth=19)
model_dt = dt.fit(df_train)
```

```
Decision Tree Performance Measure
Accuracy = 0.94
```

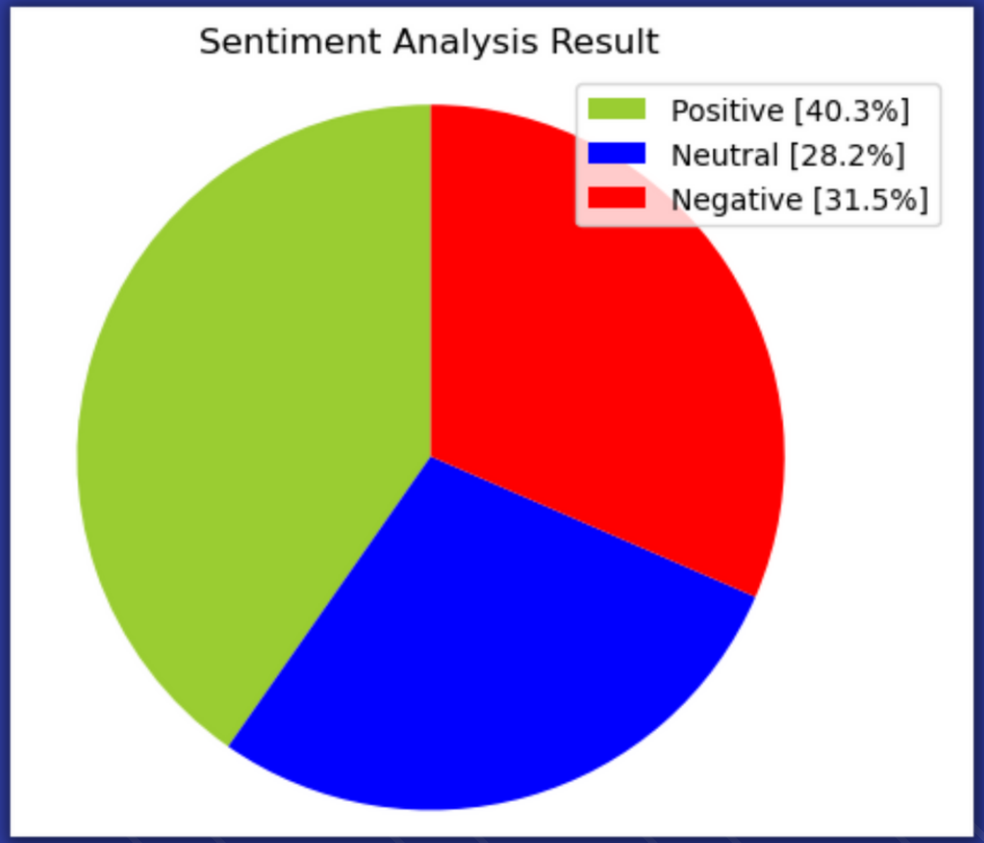
05

Apache Kafka Application

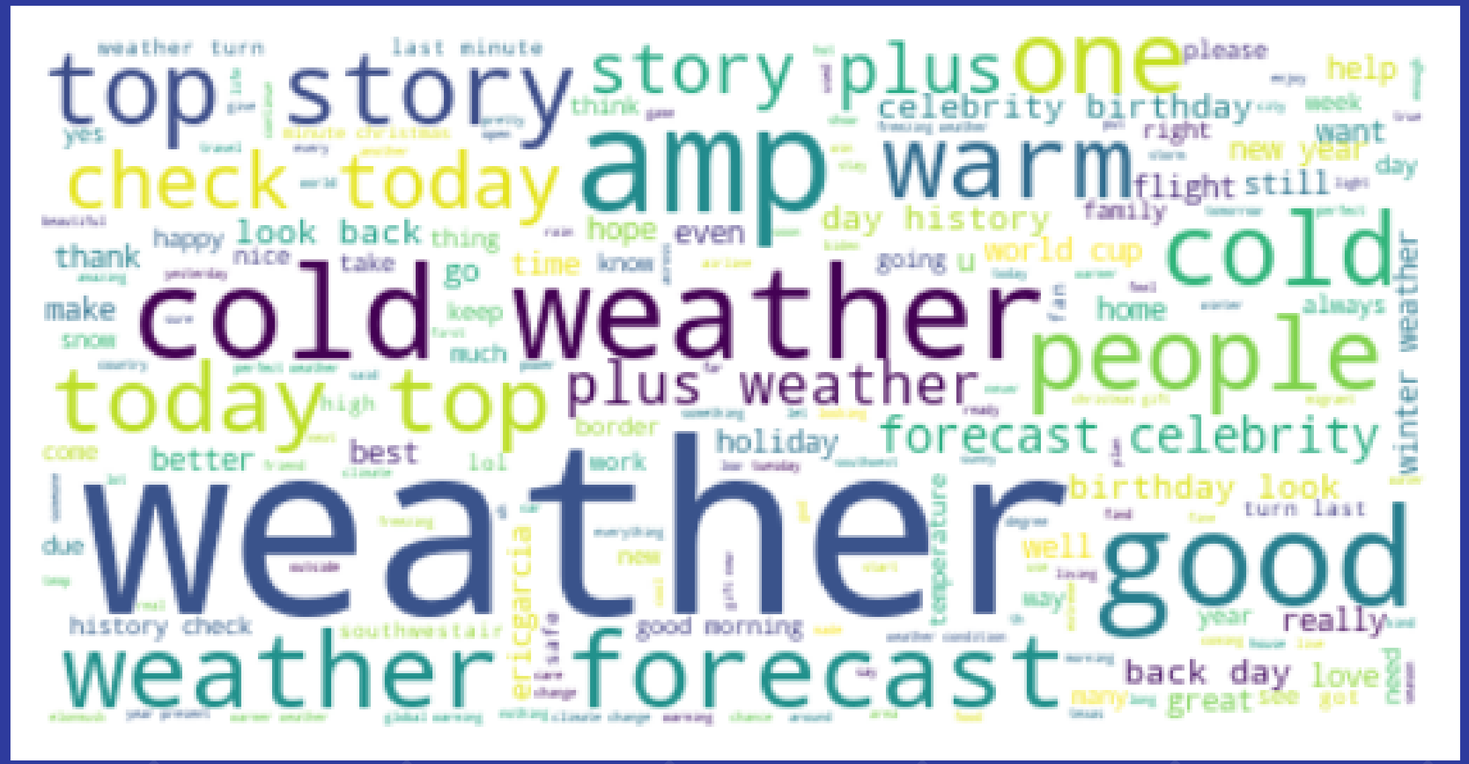
Apache Kafka weather-related tweets sentiment analysis application

- The US accident dataset strongly relays on the weather factor.
- we implemented a sentiment analysis app for weather-related tweets using Kafka
- The number of tweets collected is 5000 tweets.

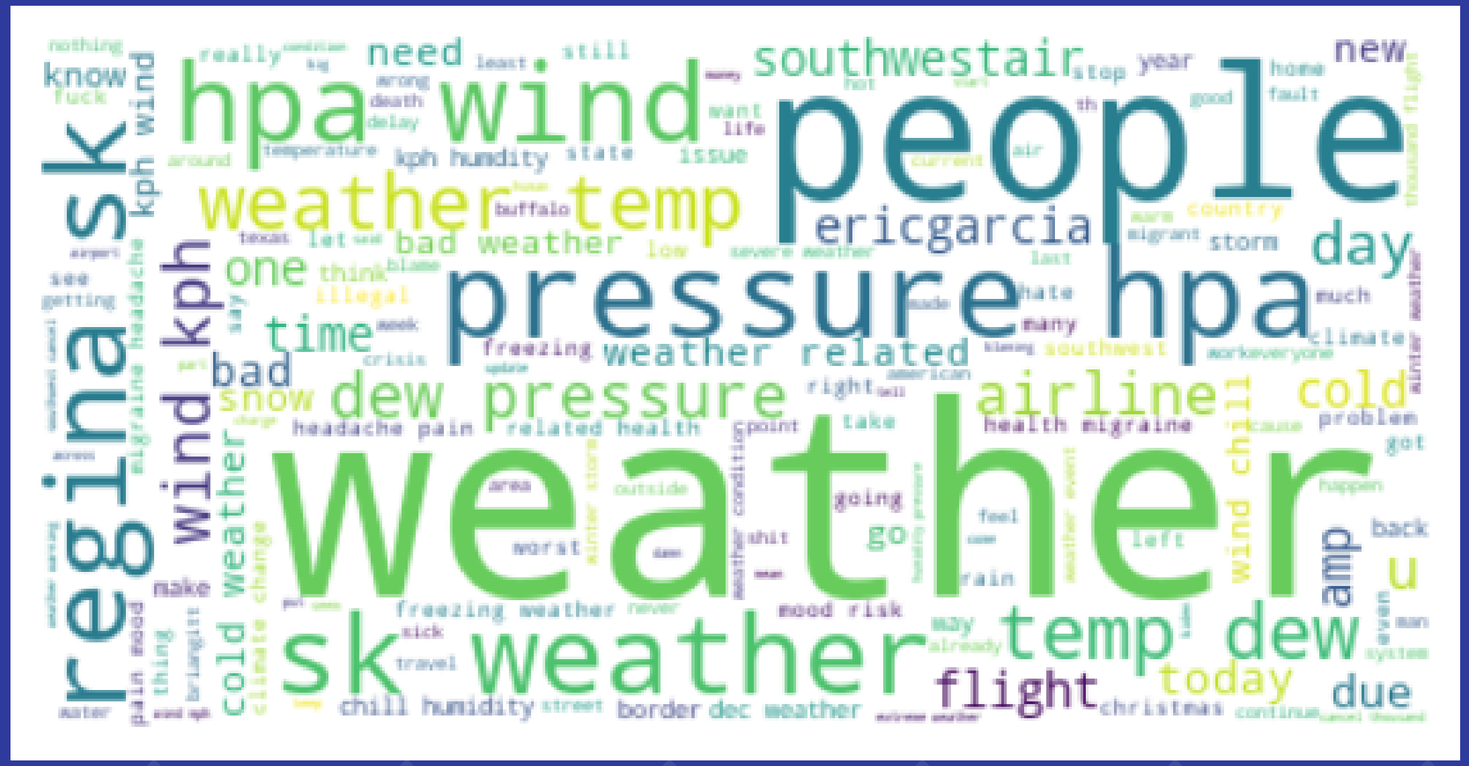
Pie Chart for representing polarity as proportions for the collected tweets



Word cloud for positive sentiments



Word cloud for negative sentiments



10

Conclusion and Future Work

Thank you

Champions Team