Champions
Dec 28 · 20 min read · Listen

Save

Analyzing United States traffic accidents and comparing them to Saudi Arabia cases

Analyzing traffic accidents in the United States and comparing them to Saudi Arabia cases with the objectives of Vision 2030

Bolg Content

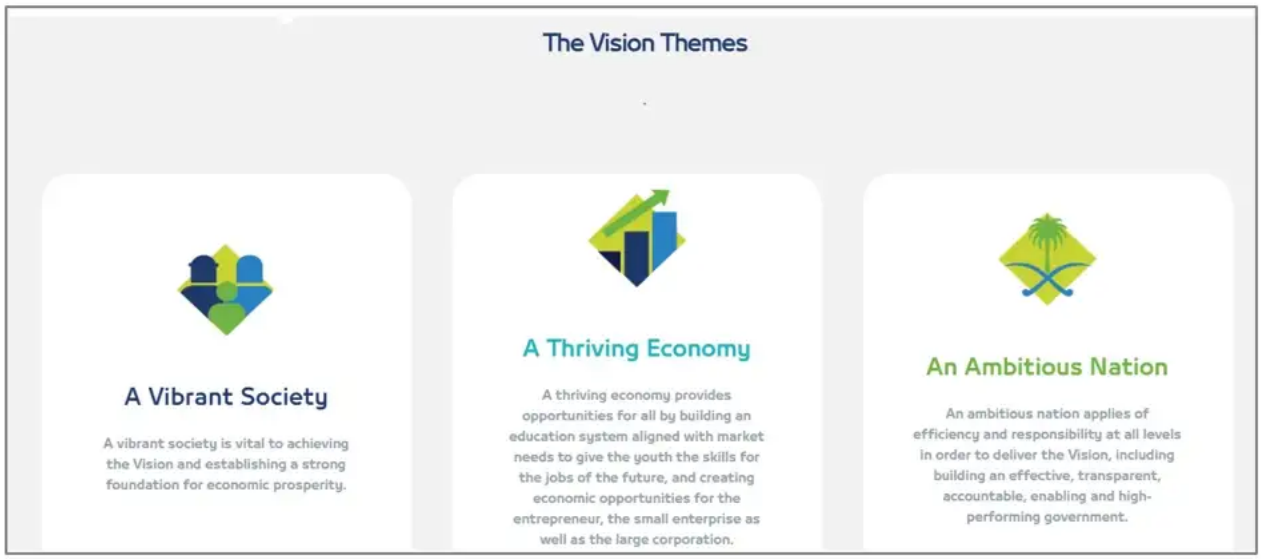
1. *Introduction*
2. *How our project is related to Saudi Arabia 2030 vision*
3. *Dataset description*
4. **Exploratory Data Analysis — EDA**
5. *Dashboards*
6. *Pig Latin Exploratory Data Analysis*
7. *Necessary dataset preprocessing steps and results*
8. *Apache PySpark Machine Learning models — Classification*
9. *Comparison between the Apache PySpark ML models*
10. *Best Apache PySpark ML model improvement*
11. *Apache Kafka weather-related tweets sentiment analysis application*
12. *Future work*
13. *Conclusion*

::::::

2. How our project is related to Saudi Arabia 2030 vision

Under the leadership of the Custodian of the Two Holy Mosques, Vision 2030 was launched, and a roadmap was drawn up by His Royal Highness the Crown Prince. The first details were announced on 25 April 2016.

One of the vision themes is called a Vibrant Society, the goal of it is to create a society in which every citizen enjoys a happy fulfilling lifestyle complemented by a standard of living that provides a safe and secure environment for families, and access to world-class health care and education.



As a part of this theme, Saudi Arabia has put effort and investment into reviewing previous studies of developed countries in this field and following in their footsteps in improving traffic discipline such as the United States, as there are many similarities between the both of them like large countries and right-driving seat. From here, raising traffic efficiency aims to improve its performance and raise its ability to strictly apply traffic regulations on violators of road users inside and outside cities, which requires providing the necessary financial resources to support it humanly and capitally without being fully borne by the government, as traffic can finance a large part of its financial needs. Relying on self-financing from the implementation of traffic regulations on violators.

A new automated citation system was implemented in 2018 to improve compliance with seatbelt and mobile phone laws. Also, Increased the implementation of automated speed detection systems proved to be useful in other countries, the Saudi traffic police adopted and introduced a similar system, known as Saher. The Kingdom has invested heavily in the construction and maintenance of road infrastructure. Saudi Arabia sees a significant drop in traffic-related fatalities. The number of road accident deaths in Saudi Arabia has decreased by 55% in 5 years.

::

3. Dataset description

> *This is a countrywide traffic accident dataset, which covers 49 states of the United States. The data is continuously being collected from February 2016 to Dec 2021, using several data providers, including multiple APIs that provide streaming traffic event data. These APIs broadcast traffic events 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 **1.5 million** accident records in this dataset. Check the below descriptions for more detailed information.*

#	Attribute	Description	Nullable
1	ID	This is a unique identifier of the accident record.	No
2	Severity	Shows the severity of the accident, a number between 1 and 4, where 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).	No
3	Start_Time	Shows start time of the accident in local time zone.	No
4	End_Time	Shows end time of the accident in local time zone. End time here refers to when the impact of accident on traffic flow was dismissed.	No
5	Start_Lat	Shows latitude in GPS coordinate of the start point.	No
6	Start_Lng	Shows longitude in GPS coordinate of the start point.	No
7	End_Lat	Shows latitude in GPS coordinate of the end point.	Yes
8	End_Lng	Shows longitude in GPS coordinate of the end point.	Yes
9	Distance(mi)	The length of the road extent affected by the accident.	No
10	Description	Shows natural language description of the accident.	No
11	Number	Shows the street number in address field.	Yes
12	Street	Shows the street name in address field.	Yes
13	Side	Shows the relative side of the street (Right/Left) in address field.	Yes
14	City	Shows the city in address field.	Yes
15	County	Shows the county in address field.	Yes
16	State	Shows the state in address field.	Yes
17	Zipcode	Shows the zipcode in address field.	Yes
18	Country	Shows the country in address field.	Yes
19	Timezone	Shows timezone based on the location of the accident (eastern, central, etc.).	Yes
20	Airport_Code	Denotes an airport-based weather station which is the closest one to location of the accident.	Yes
21	Weather_Timestamp	Shows the time-stamp of weather observation record (in local time).	Yes
22	Temperature(F)	Shows the temperature (in Fahrenheit).	Yes
23	Wind_Chill(F)	Shows the wind chill (in Fahrenheit).	Yes
24	Humidity(%)	Shows the humidity (in percentage).	Yes
25	Pressure(in)	Shows the air pressure (in inches).	Yes
26	Visibility(mi)	Shows visibility (in miles).	Yes
27	Wind_Direction	Shows wind direction.	Yes
28	Wind_Speed(mph)	Shows wind speed (in miles per hour).	Yes
29	Precipitation(in)	Shows precipitation amount in inches, if there is any.	Yes
30	Weather_Condition	Shows the weather condition (rain, snow, thunderstorm, fog, etc.)	Yes
31	Amenity	A POI annotation which indicates presence of amenity in a nearby location.	No
32	Bump	A POI annotation which indicates presence of speed bump or hump in a nearby location.	No
33	Crossing	A POI annotation which indicates presence of crossing in a nearby location.	No
34	Give_Way	A POI annotation which indicates presence of give_way in a nearby location.	No
35	Junction	A POI annotation which indicates presence of junction in a nearby location.	No
36	No_Exit	A POI annotation which indicates presence of no_exit in a nearby location.	No

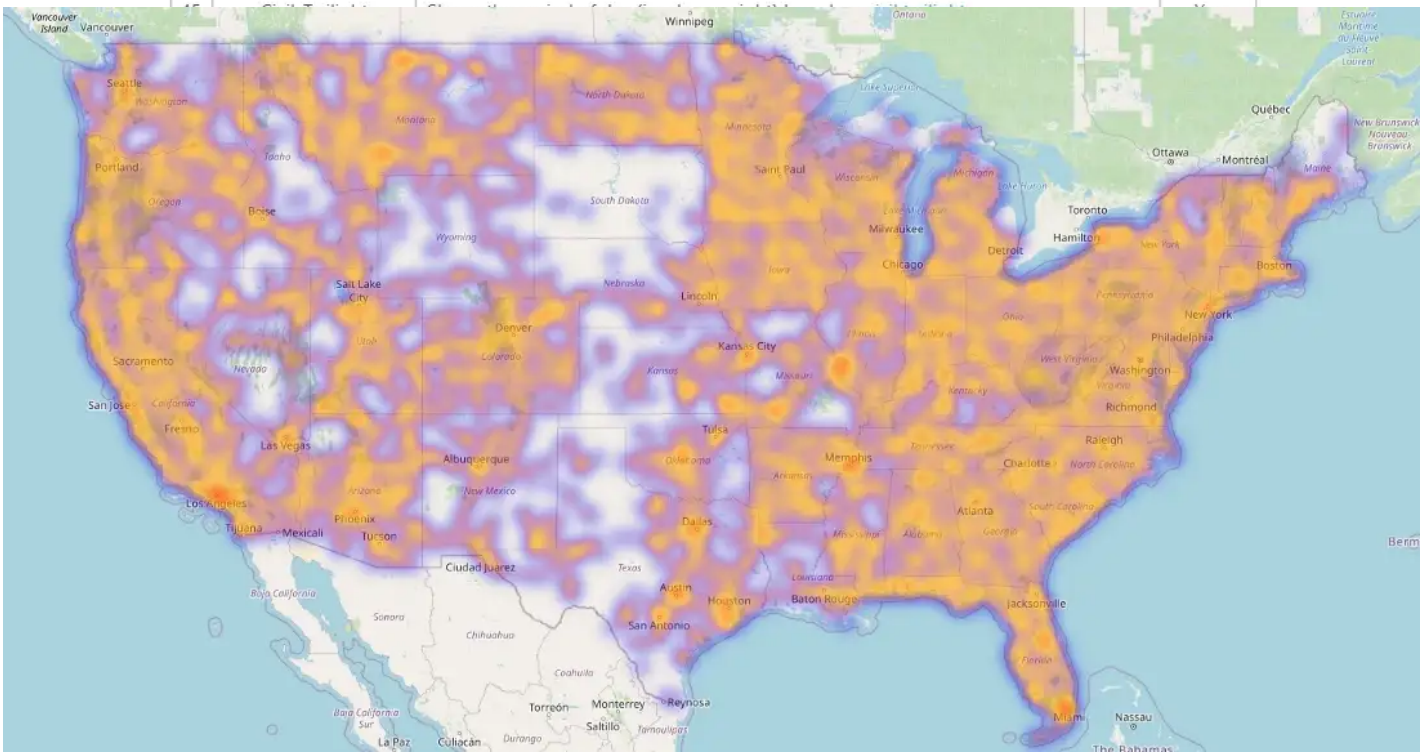
36	No_Exit	A POI annotation which indicates presence of no_exit in a nearby location.	No
37	Railway	A POI annotation which indicates presence of railway in a nearby location.	No
38	Roundabout	A POI annotation which indicates presence of roundabout in a nearby location.	No
39	Station	A POI annotation which indicates presence of station in a nearby location.	No
40	Stop	A POI annotation which indicates presence of stop in a nearby location.	No
41	Traffic_Calming	A POI annotation which indicates presence of traffic_calming in a nearby location.	No
42	Traffic_Signal	A POI annotation which indicates presence of traffic_signal in a nearby loction.	No
43	Turning_Loop	A POI annotation which indicates presence of turning_loop in a nearby location.	No
44	Sunrise_Sunset	Shows the period of day (i.e. day or night) based on sunrise/sunset.	Yes

:::::::::::::::::::::

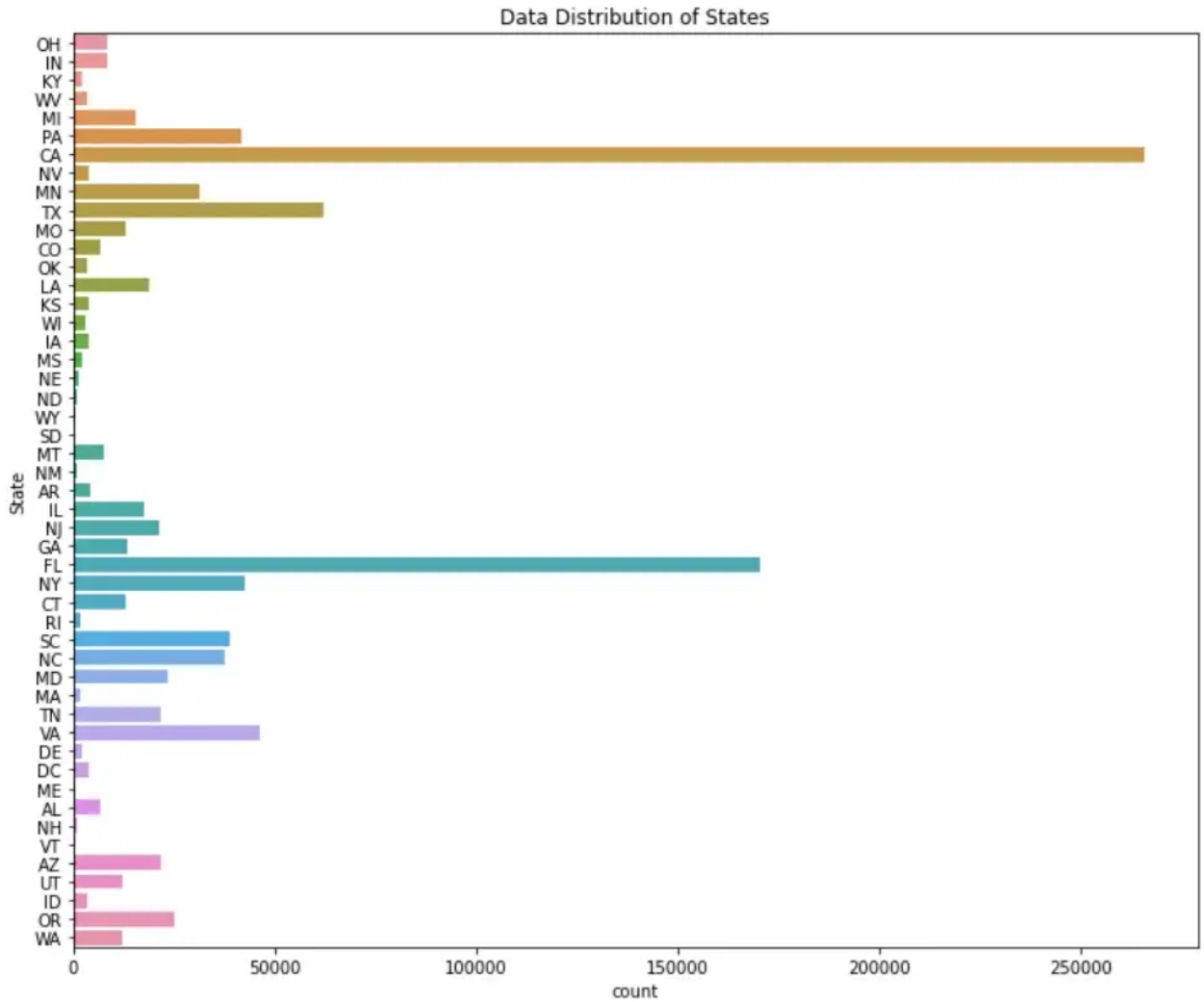
4. Exploratory

4.1 US Acciden

This heatmap s visualize the locations. we c a really low rate of acciden s Angeles have greater accider

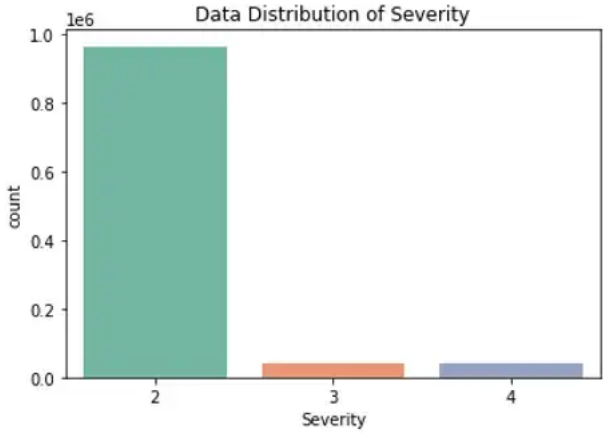


Here in the chart below we explored the state column, we can notice that most of the data are in the states of California and Florida.



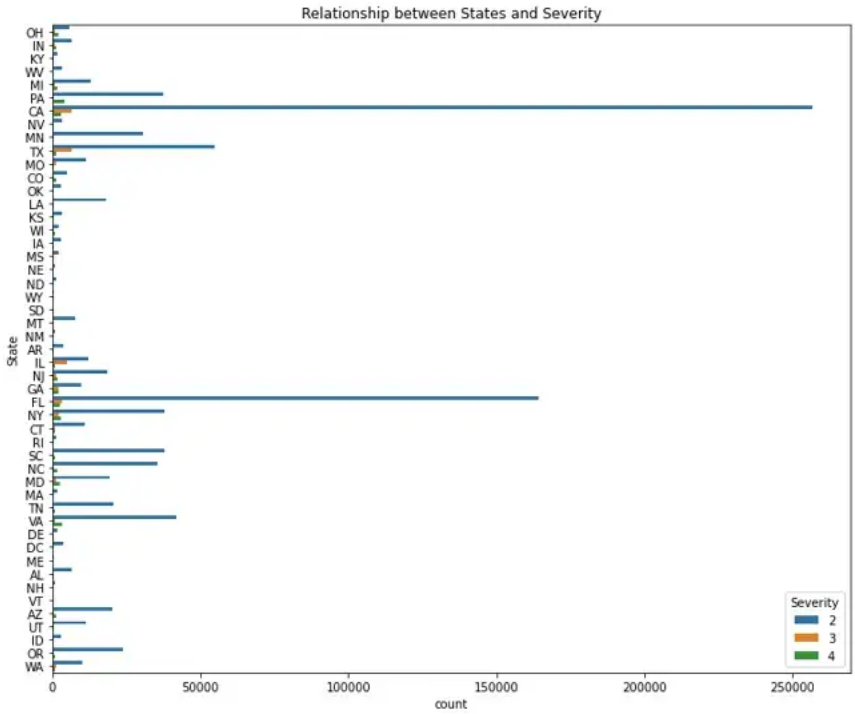
4.2 The Distribution of Severity Values

The chart shows the distribution of the Severity column in the dataset. The Severity of accidents ranges from 2 to 4. We can notice here that most of the accidents presented in the dataset are in the range of 2. This indicates that most accidents have the least impact on traffic.



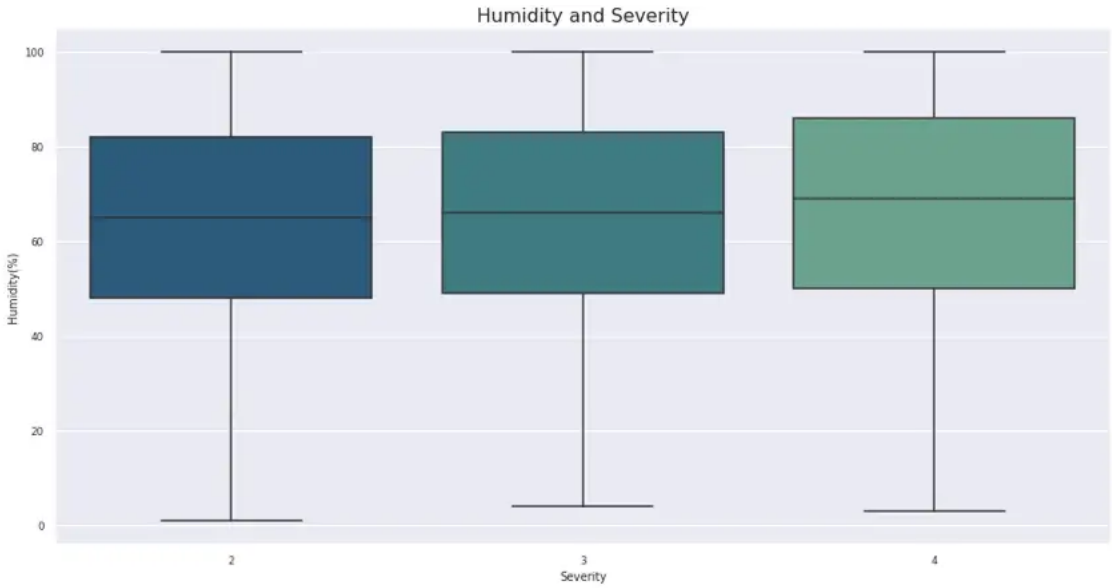
4.3 Severity by States

In this chart, we explored the relationship between the two columns' state and severity. As we can see that the state of Pennsylvania has the most severe accidents compared with other states presented in the dataset.



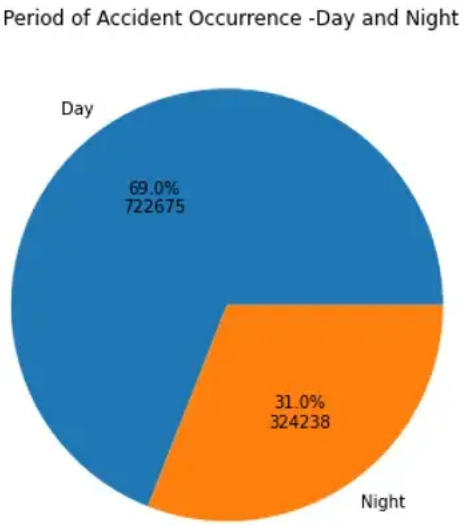
4.4 The link between Humidity and the Severity of an accident

This box plot shows the link between humidity and the severity of an accident. The levels of humidity range from 2 to 100. However, the majority are in the 50–80 range. and we noticed that there is some sort of linear relationship between the severity and the humidity. We believe this is related to the accident's location. There are more accidents in the coastal states, and they have higher humidity rates than the middle states in the US.



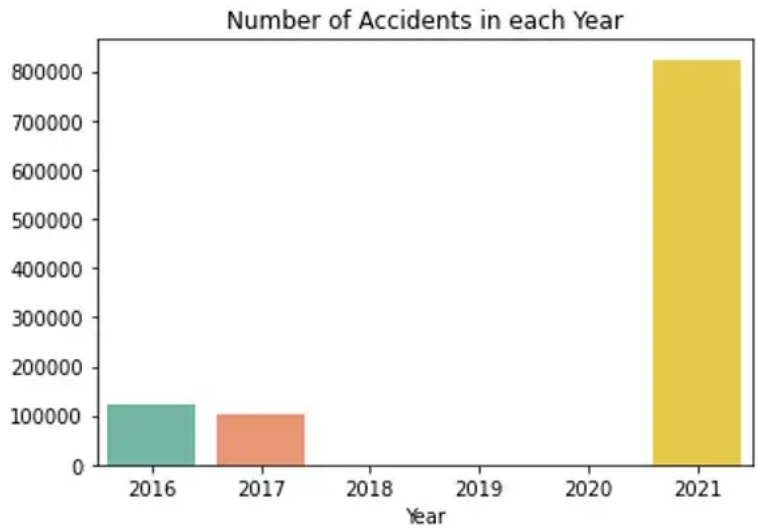
4.5 Accident occurrence by Day and Night

The pie chart shows that 69% of accidents occurred during the day.



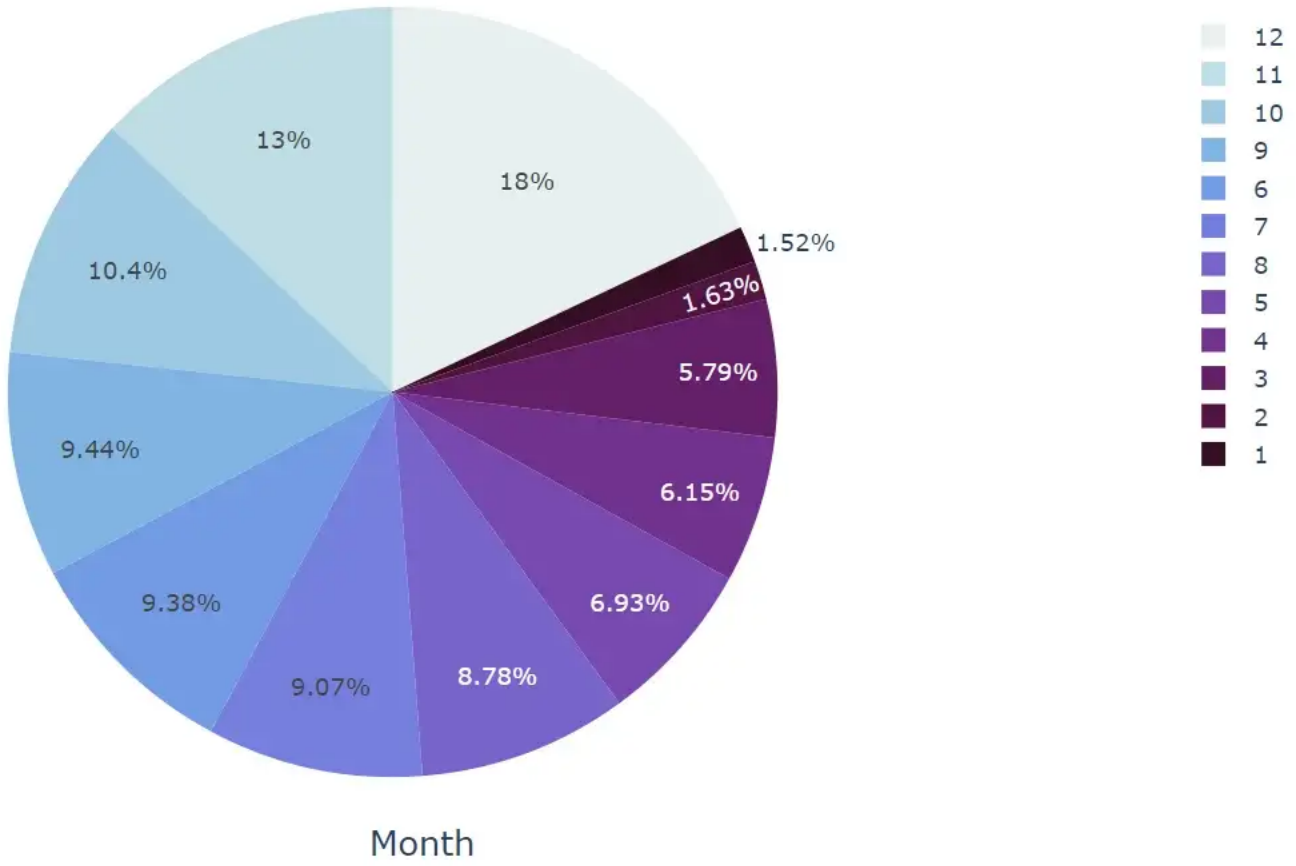
4.6 Number of Accidents in each Year

In this chart, we explored the number of accidents each year. we can notice that 2021 has the highest number of accidents compared with the past five years.



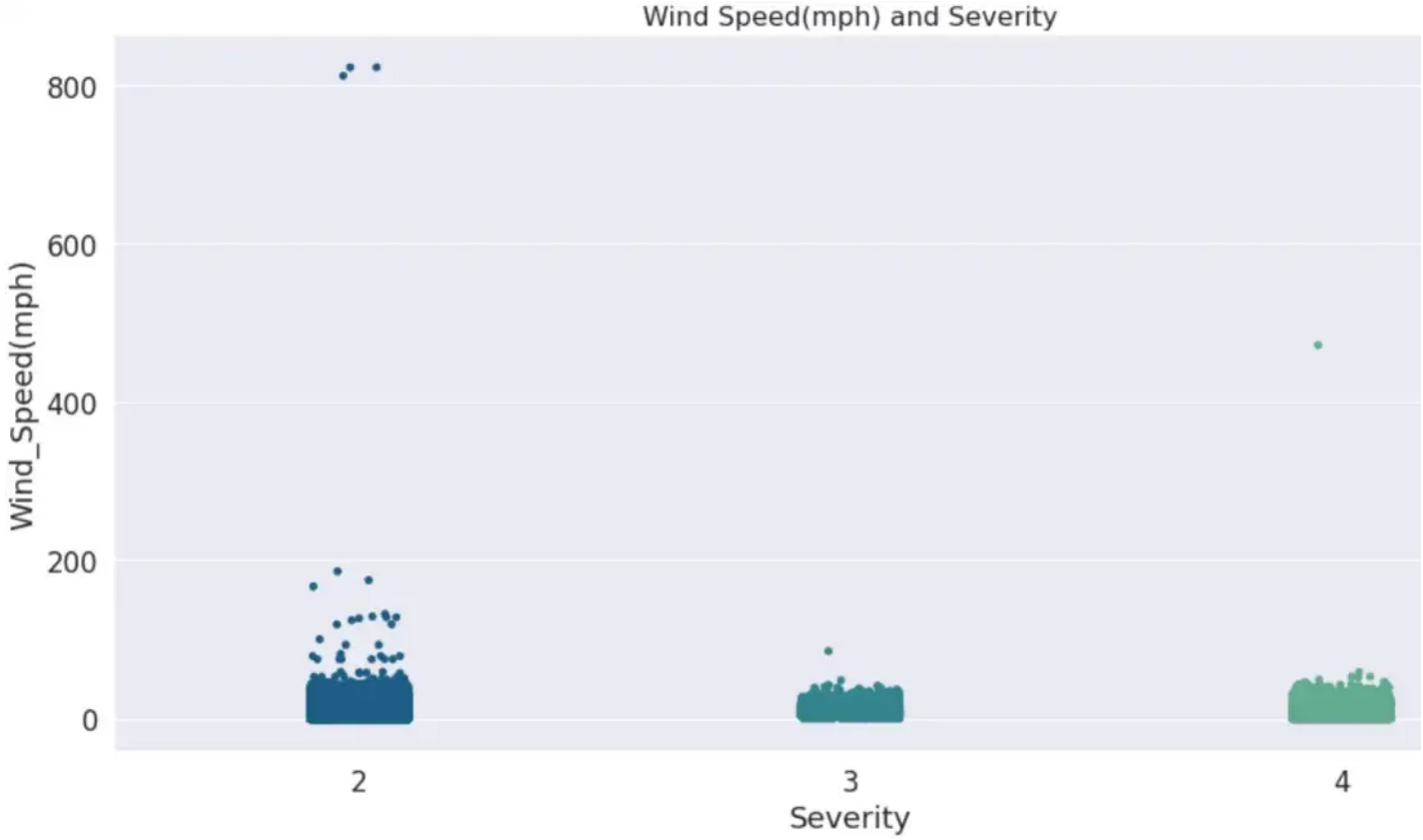
4.7 Accidents Distribution by Month

In this pie chart, we want to know the accident distribution for the months. Since December is a holiday season and usually snowy, it's not surprising that there are more accidents in December, accounting for 18% of all incidents.



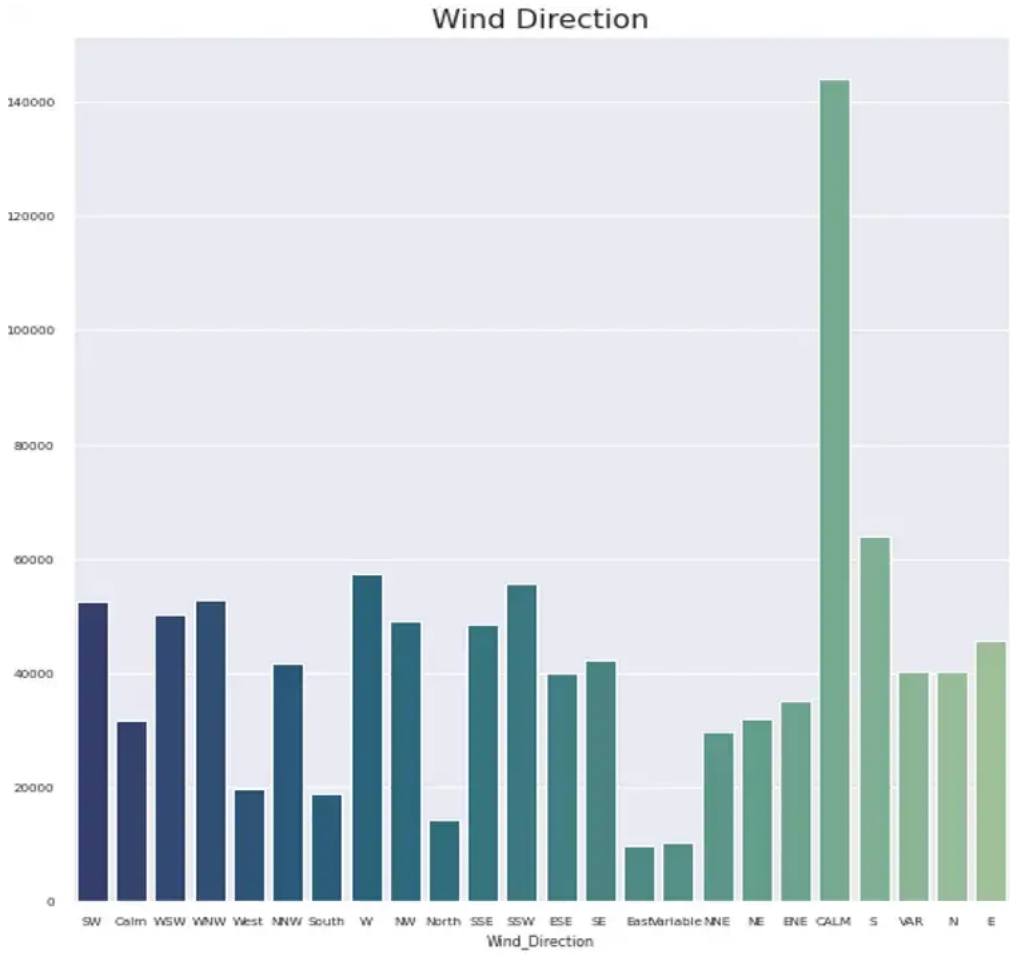
4.8 Wind Speed(mph) and Severity

In this plot, we want to study the relationships between wind speed and the severity of an accident. And as we can see, there is no clear connection between wind speed and severity. However, we discovered that the wind speed contains outliers, and we must eliminate them during the data preprocessing stage.



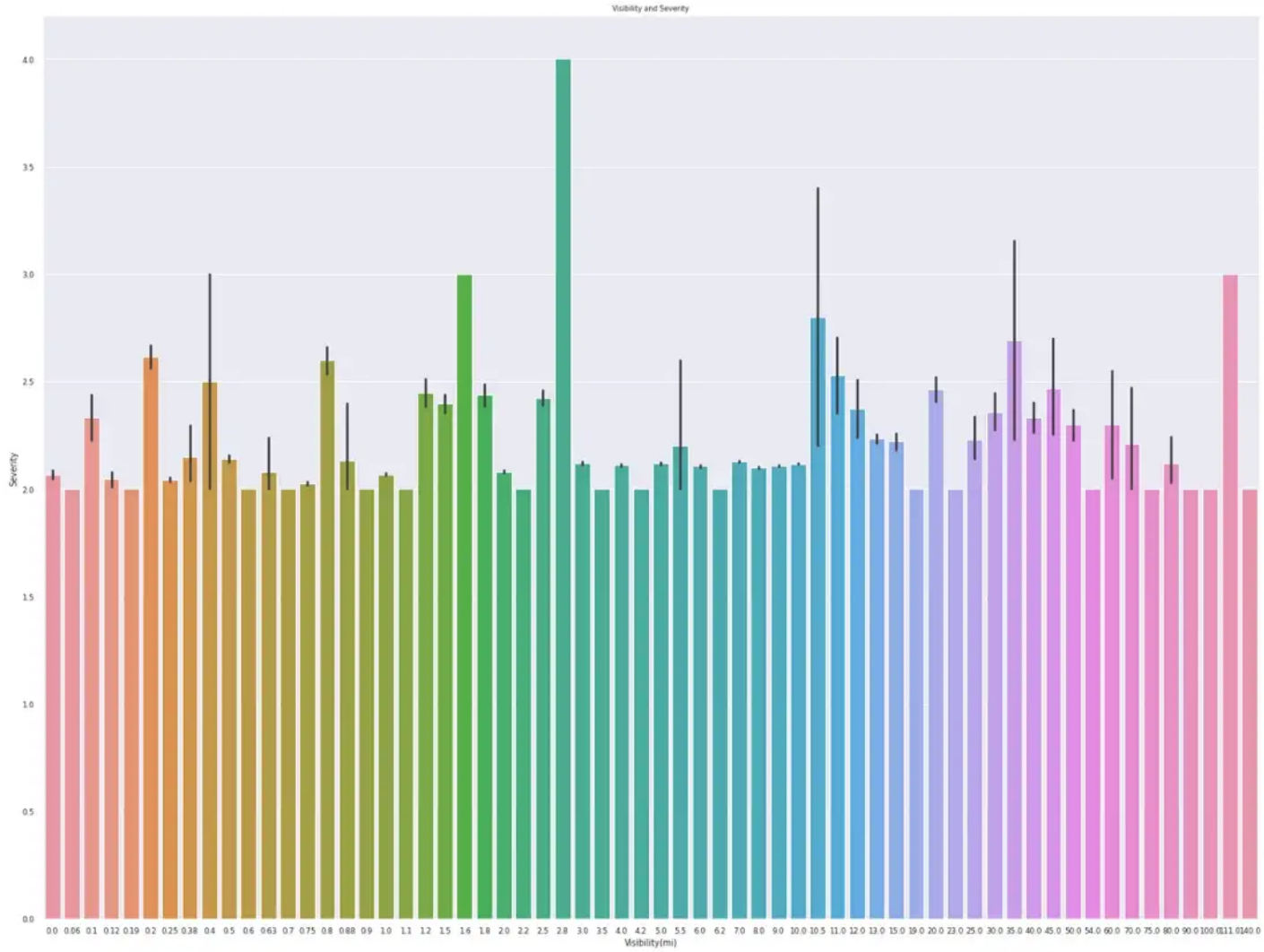
4.9 Wind Direction

We investigated the number of accidents and the wind directions in this graph. We can conclude that there is no clear relation between wind direction and accidents because the majority of incidents occurred while the wind was calm.



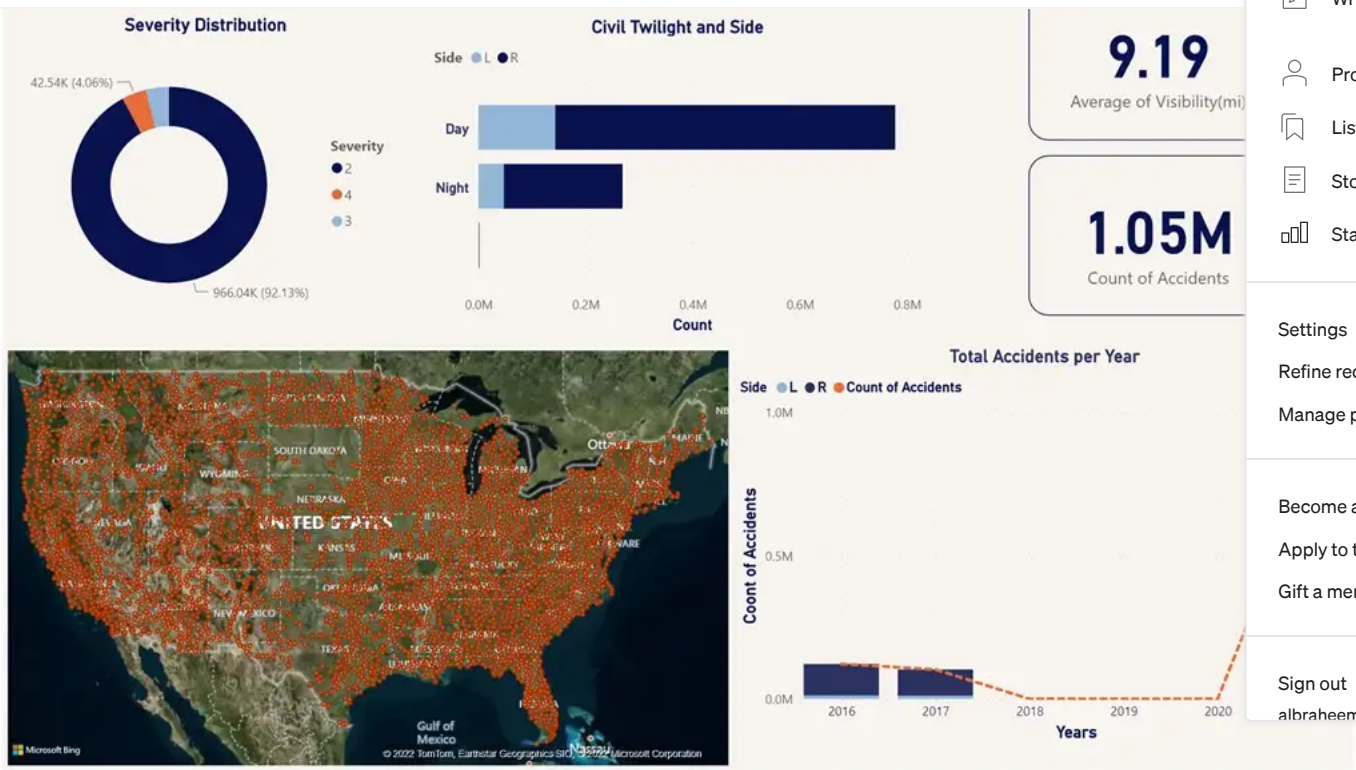
4.10 Visibility and Severity

This graphic shows the relationship between visibility and accident severity, and it can be concluded that there is no correlation directly relating visibility to accident severity. However, we have a high rate of accidents with a severity of 4 when the visibility is 3 miles, and a high rate of accidents with a severity of 3 when the visibility is 1.6 miles.

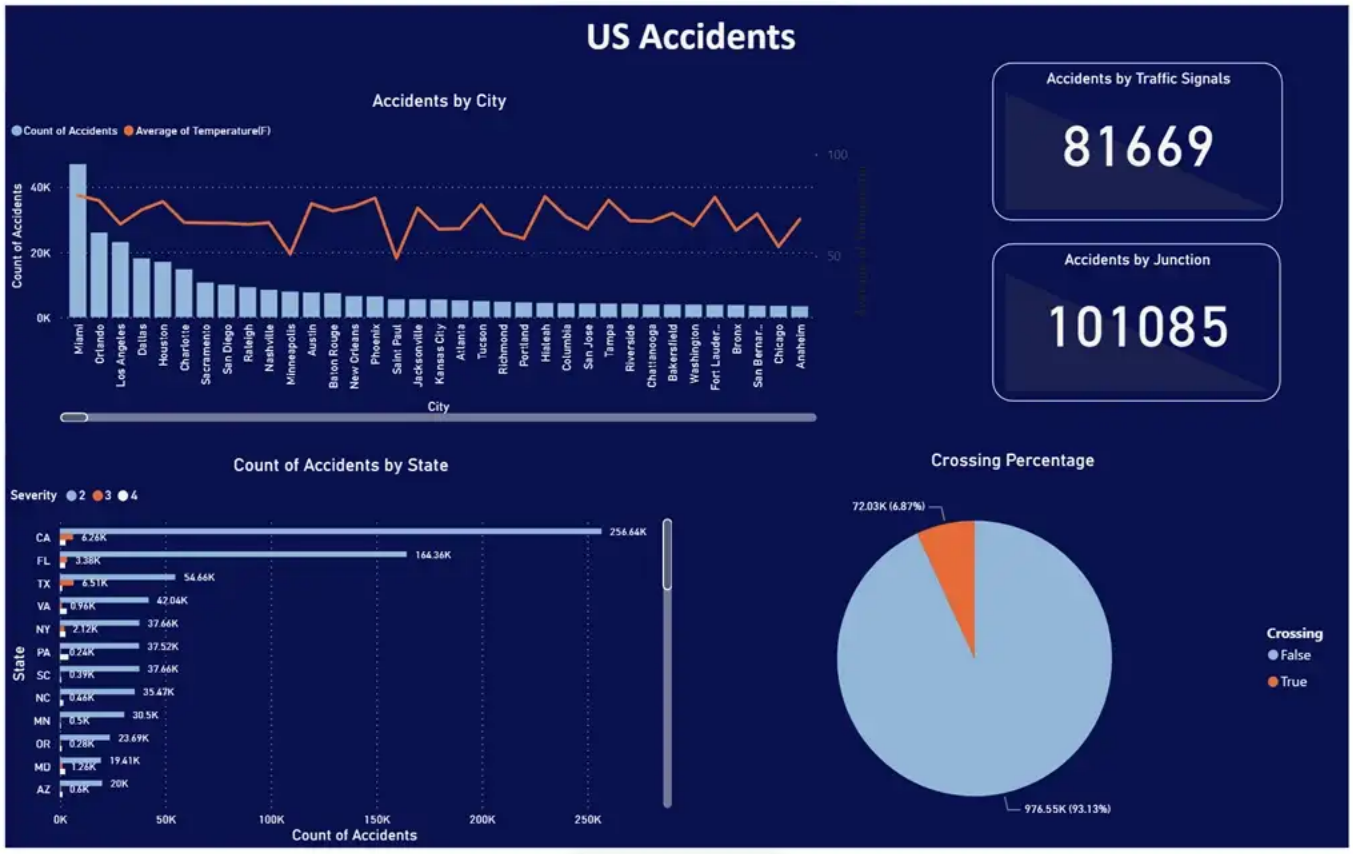

Visibility and Severity
Severity
Visibility(mi)

::::::

Open in app
Get unlimited access
Search Medium
Write
Profile
Lists
Stories
Stats
Settings
Refine recommei
Manage publicat
Become a memb
Apply to the Part
Gift a membersh
Sign out
albraheem3m@gr



Dashboard number one



6. Pig Latin Exploratory Data Analysis

We used Pig Latin to answer 20 interesting questions about the US accidents dataset that helped us find directions for our project and shape our investigation. The questions and results are as follows:

6.1 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.

Observations: The results show that California, which is the highest populated State, has the highest count of accidents in all months, especially in December probably for Christmas and new year celebrations with family. The overall count results indicate the higher the population, the higher the probability of accidents.

```
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)
```

Pig Latin question (6.1) result

If we compare this result with Saudi Arabia, we will find similar situations. For 2020, the total population in Makkah was 9.26 million persons which accounts for 26.45% of Saudi Arabia's total population. As a result, it is considered the top region in population. The others are Riyadh, Eastern Region, Aseer, and Al Madinah account for 80.10% of the total population. During holidays and the months of Hajj and Ramadan, the possibility of traffic accidents increases. Here is the illustration from the Saudi General Authority for statistics about traffic accidents site by Region:

Region	percent	*No. of Accidents*
Riyadh	33.78%	*138974*
Makkah	24.94%	*102607*
Madinah	2.95%	*12137*
Al-Qasim	3.47%	*14295*
Eastern	16.60%	*68280*
Aseer	4.57%	*18806*
Tabouk	4.11%	*16924*
Hail	1.31%	*5378*
Northern Boarders	2.74%	*11260*
Jazan	2.36%	*9700*
Najran	0.48%	*1989*
Al-Baaha	0.75%	*3079*
Al-Jowf	1.94%	*7968*
Total	**100%**	***411397***

The statistics of the Saudi General Authority for traffic accidents site by Region

6.2 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.

Observations: we can see from this result that most accidents on the right side without the means of amenity occur when the weather is fair. In other words, the accident may not have been caused primarily by the weather in general.

```
(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)
```

Pig Latin question (6.2) result

This can be linked with Saudi Arabia, where the weather is fair on most days of the year, yet the accident rate is high.

6.3 The highest temperature recorded in KSA was between 45(113) C(F) — 53(127) C(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.

Observations: the result shows that most of the accidents occur during temperatures from 37(100) — 44(111.4) C(F)degrees Fahrenheit. This temperature range is in most areas where traffic accidents occur.

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

Pig Latin question (6.3) result

With the exception of the mountain highlands in the southwest of the Kingdom, which witness moderate temperatures, summer marks the beginning of a rise in temperatures in most regions to about 45(113 F) degrees. As shown in the table below:

1	K.f.i.a. (King Fahad Int. Airport) Dammam (Saudi Arabia)	48.0 °C
2	Al Qaysumah (Saudi Arabia)	47.4 °C
3	Rafha (Saudi Arabia)	46.5 °C
4	Al Ahsa (Saudi Arabia)	46.3 °C
5	Wadi Al-Dawasser Airport (Saudi Arabia)	46.0 °C
6	King Khaled International Airport (Saudi Arabia)	45.7 °C
7	Sharurah (Saudi Arabia)	45.4 °C
8	Gassim (Saudi Arabia)	44.2 °C
9	Madinah (Saudi Arabia)	43.5 °C
10	Al-Dawadami (Saudi Arabia)	43.0 °C
11	Arafat (Saudi Arabia)	43.0 °C
12	Arar (Saudi Arabia)	42.1 °C
13	Al-Jouf (Saudi Arabia)	41.8 °C
14	Alula (Saudi Arabia)	41.0 °C
15	Najran (Saudi Arabia)	41.0 °C
16	Hail (Saudi Arabia)	40.4 °C
17	Bisha (Saudi Arabia)	40.1 °C
18	Guriat (Saudi Arabia)	39.0 °C
19	Gizan (Saudi Arabia)	38.3 °C
20	Tabuk (Saudi Arabia)	38.0 °C
21	Al Baha (Saudi Arabia)	37.7 °C
22	Turaif (Saudi Arabia)	37.0 °C
23	Yenbo (Saudi Arabia)	37.0 °C
24	Taif (Saudi Arabia)	36.0 °C
25	Jeddah King Abdul Aziz International Airport (Saudi Arabia)	35.2 °C
26	Khamis Mushait (Saudi Arabia)	34.3 °C
27	Abha (Saudi Arabia)	33.8 °C
28	Wejh (Saudi Arabia)	33.6 °C

The maximum temperature of 28 locations in Saudi Arabia

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

Observations: the result shows maximum and minimum temperatures only for four states. Among the four states, Alabama has the highest temperature of 196 °F and Texas has the lowest temperature of 14 °F.

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

Pig Latin question (6.4) result

Among the ten hottest US states, these four were selected to demonstrate a relationship with Saudi Arabia. That is due to the fact that Saudi Arabia has a number of hot cities like Makkah, Dammam, and Al-Hasa.

6.5 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%.

Observations: The query result shows the majority of accidents recorded when speed(mph) is 20 or above and humidity is less than 30% are in the 2nd severity level. We think it is imbalanced data as we have more records for severity level 2, and fewer for 3 and 4. Because the dust wave obscures the level of vision, and in turn causes obstruction of road traffic as well.

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

Pig Latin question (6.5) result

6.6 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?

Observations: according to this result, the top three counties with the highest average wind speed are Natrona, Clear Greek, and Allegany. As the west wind blows and the temperature is 25 degrees or below.

```
(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)
```

Pig Latin question (6.6) part of the result

It might be related to the low temperatures and moving winds in some parts of southern and northern Saudi Arabia. Besides the heights and mountains, accidents may also occur there.

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

Observations: the total number of accidents that occurred at night with humidity between 85 and 100 is 85683. This is less than a quarter of the data but still a big number.

```
(85683)
```

Pig Latin question (6.7) result

Regions near the sea in Saudi Arabia suffer from a high percentage of humidity that hinders vision and can result in accidents.

6.8 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?

Observations: from this result, we can see that there are 5188 accidents that have a distance of more than 3 miles and have the highest severity so that might correlate with the distance being quite large.

```
(5188)
```

Pig Latin question (6.8) result

Saudi Arabia is large in terms of area and most of the transportation is by car, which makes the number of drivers greater and the possibility of accidents higher.

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

Observations: from this result, we can see that there are more accidents that happen on the right side of the street rather than the left side. In Saudi Arabia, the driving laws for vehicles are similar as both drivers drive on the left side of the vehicle so it's more likely that accidents happen on the side that is furthest than the driver.

```
(L, 27288)
(R, 392535)
```

Pig Latin question (6.9) result

6.10 What is the average distance for accidents for each side?

Observations: from this result, we can see that accidents that occur on the right side are way higher in distance than those that happen on the left side. So they are way more dangerous as they are from the furthest side of the driver.

```
(L,0.31062299907573987)
(R,0.9159718676994905)
```

Pig Latin question (6.10) result

6.11 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.

Observations: from this result, we can see that accidents involving sand have a severity of 2. And a lot of those accidents have weather type as Sand / Dust whirlwinds which are common weather types in Saudi Arabia.

```
(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)
```

Pig Latin question (6.11) result

6.12 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.

Observations: The number of accidents in these cities due to fog that makes visibility less than 0.5 mi in 2016 is 2 accidents, while in 2017 it is only one accident, and the highest number of accidents is in 2021, where 11 accidents were recorded this year due to fog.

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

Pig Latin question (6.12) result

Separate regions of the north and east of the Kingdom, and the foggy atmosphere included all of the southern region and the mountainous highlands in Asir and Taif, and parts of Hail from Arar and Sakaka, where thick fog formed. It may sometimes be a source of danger for road users due to the lack of horizontal visibility that may reach less than two meters and result in traffic

accidents. The dark, foggy atmosphere paralyzes traffic, which sometimes forces motorists to stop and not drive in bad weather for safety reasons.

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

Observations: The result shows the maximum and minimum visibility values for each year. For those accidents that happened at night, we can notice that the maximum visibility value of 140 was recorded in 2021. In the other hand, the minimum visibility values were 0 for most of the years.

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

Pig Latin question (6.13) result

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

Observations: the result shows the maximum and minimum visibility values for each year. For those accidents that happened during the day, we can notice that the maximum visibility value of 110 was recorded in 2016. And the minimum visibility values were zero for most of the years.

```
(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)
```

Pig Latin question (6.14) result

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

Observations: We can notice that the wind speed was highest from January to April compared to other months of the year. Then the wind speed decreases slightly for the rest of the months.

```
(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)
```

Pig Latin question (6.15) result

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

Observations: We can notice that the number of accidents in 2021 is more than in 2016 and 2017. We think it is imbalanced data as we have more records for 2021.

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

Pig Latin question (6.16) result

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

Observations: The average precipitation values were the highest in both January and July compared with other months of the year presented in the dataset.

```
(1,0.07783146376025413)
(2,0.036029100425975034)
(3,0.007879681612952693)
(4,0.00606344259951312)
(5,0.006773309463624175)
(6,0.010632766632945056)
(7,0.010865065153074855)
(8,0.009269080263029672)
(9,0.00807935677003483)
(10,0.006601139358833231)
(11,0.004944148444712884)
(12,0.0074871104426414165)
```

Pig Latin question (6.17) result

Rainfall continued in a number of regions of the Kingdom, ranging between medium and heavy, causing the pooling of water, cars, and broken signals, in addition to the presence of obstacles on the road such as trees, stones, and others. As a result, the traffic statistics showed a high incidence of traffic accidents and reports of injuries during rainy days.

6.18 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.

Observations: we filtered the visibility to 6 miles, which is the average in Saudi Arabia, and the result shows us the total number of accidents in different weather conditions. We can see that light rain, cloudy weather, and haze are the weather conditions with the highest number of accidents, which makes sense.

```
(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)
```

Pig Latin question (6.18) result

6.19 How many accidents occurred where there was a roundabout?

Observations: The number of roundabouts in the US is 5000 roundabouts and the number of accidents when there is a roundabout is 47 accidents which means it's rare to have an accident at a roundabout in the US. However, in Saudi, there are a lot of roundabouts, and one of the top causes of accidents since that Saudi drivers fail to yield at the roundabout.

```
(47)
```

Pig Latin question (6.19) result

6.20 Number of accidents where the Stop sign is True.

Observations: The number of accidents due to failure to stop at a stop sign is 19075 accidents.

```
(19075)
```

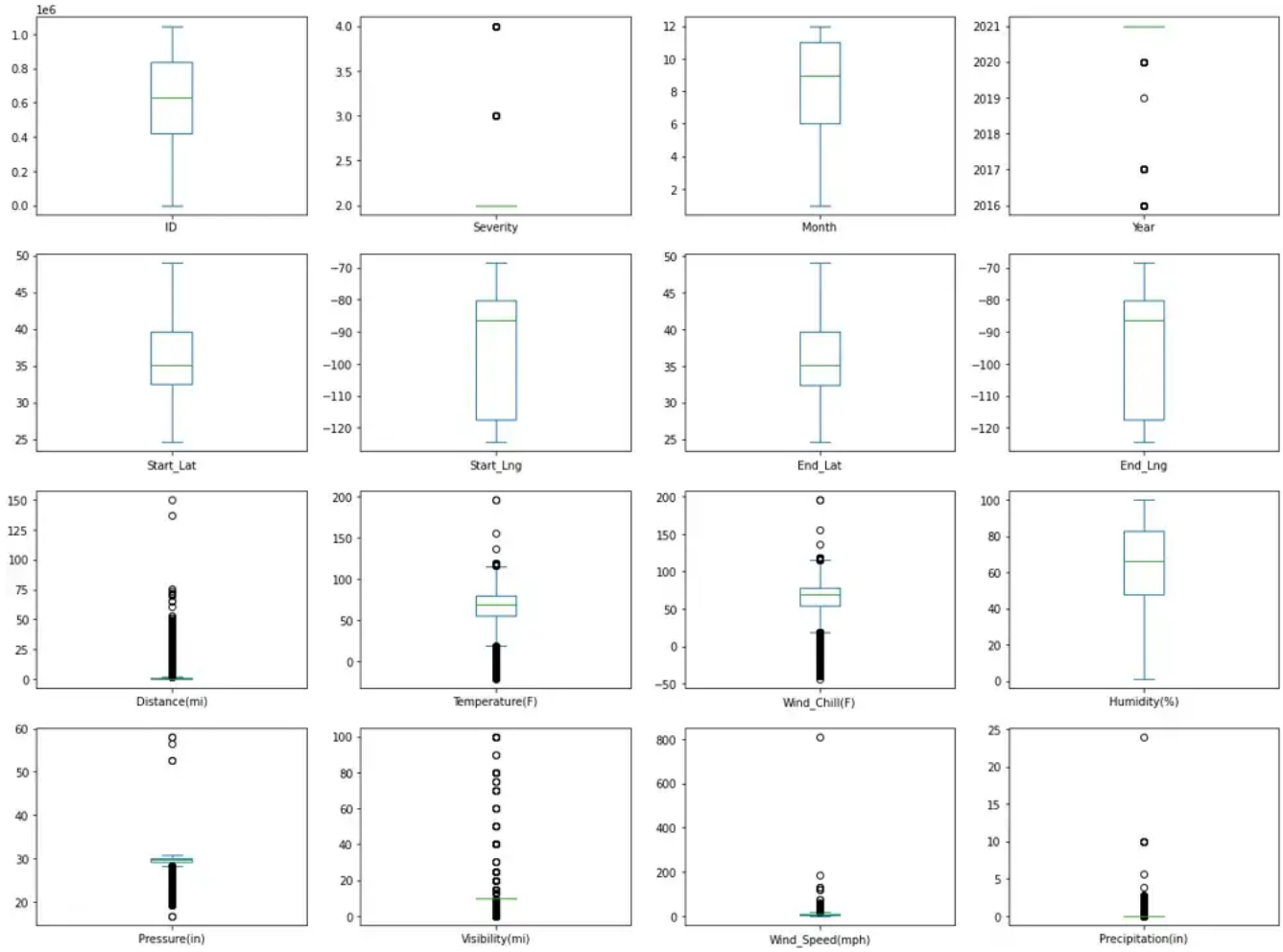
Pig Latin question (6.20) result

::

7. Necessary dataset preprocessing steps and results

7.1 Handling outliers

The first step we did is detecting outliers in the columns using a box plot.



Box plot for all dataset features

Then, created a list with columns that contain outlier values.

```
cols_with_outliers = ['Distance(mi)','Wind_Chill(F)','Pressure(in)','Visibility(mi)','Wind_Speed(mph)',
                      'Precipitation(in)']
```

After that, defined a function called ‘outliers’ which returns a list of index 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
```

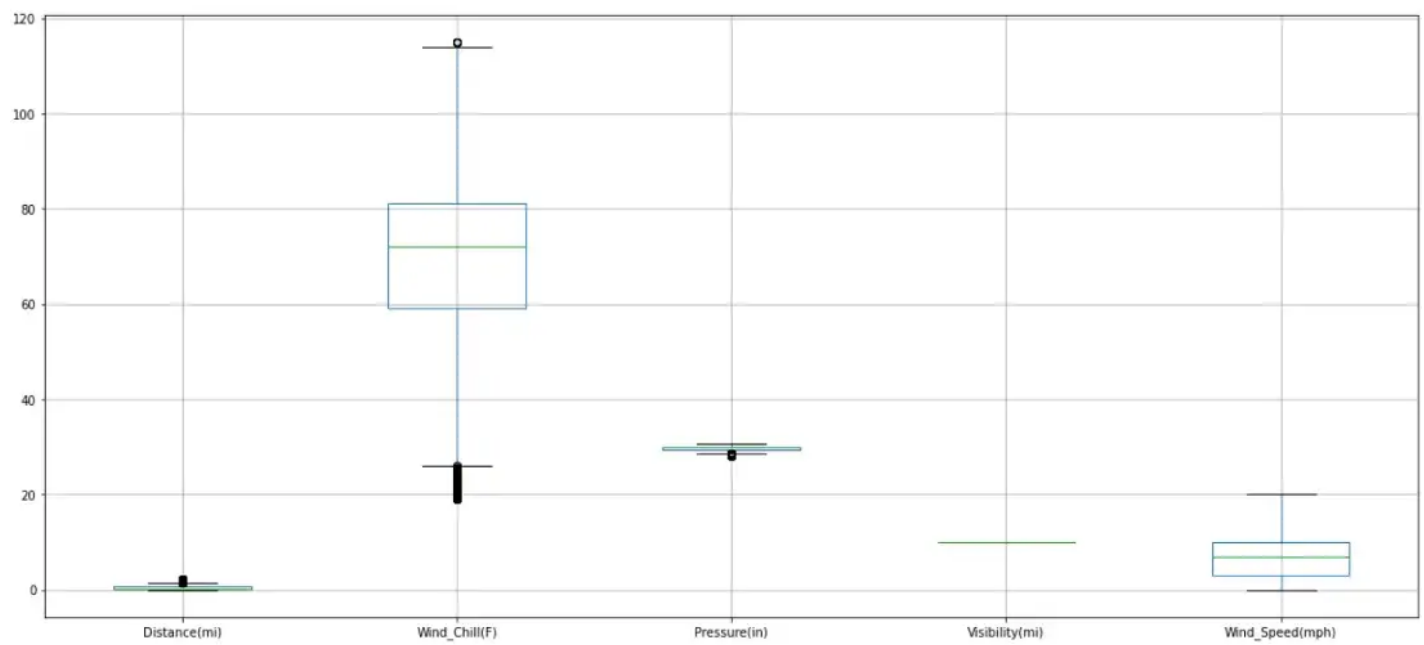
Creating an empty list to store the output indices from multiple columns

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

Lastly, define a function called ‘remove’ which returns a cleaned data frame without outliers.

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

Here is the final result after dealing with outliers.



Box plot for all dataset features without outliers

7.2 Implementing oversampling

We found that we have imbalanced data in the target column “ Severity”.

```
+--------+------+
|Severity| count|
+--------+------+
|       3|   329|
|       2|506721|
|       4|  9363|
+--------+------+
```

Count of records for each value in the “severity” column

Since we have three categories in the Severity column we will oversample the first category (which is 3) then we will report the same steps to oversample the second category (which is 4).

We handle it as follows:

1- Isolate all the data where the Severity is 2 and then the data where the Severity is 3.

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

2- Calculate the ratio to determine the difference between the number of each severity.

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

```
ratio: 1540
```

ratio result

3- Create a range on this ratio and store it in variable ‘a’.

```
a = range(ratio)

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

We can now see that the severity3_df has a new size which is quite similar to severity2_df

```
oversampled_df.count()
```

```
506660
```

Oversampled data frame count

4- Combine both oversampled minority rows and previous majority rows using .unionAll()

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

5- Checking the ratio after oversampling

```
#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 ratio after oversampling

Here is the oversampling result. The count of records for each value in the ‘Severity’ column.

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

values count after oversampling

::::::

```
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)
```

Multiclass Classifier Evaluator for ML evaluation

The accuracy that we got from the model was actually a bit less than the decision tree classifier model so far it was the highest in accuracy.

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

Confusion Matrix for **Random forest classifier**

8.3 Logistic regression model

For the third model we decided to use the Logistic regression model from the Pyspark ML classification library and we decided to start with the maximum iteration of 10.

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

Logistic regression model implementation

We used the multiclassClassifierEvaluator available in the Pyspark ML evaluation library to get the accuracy for this model.

```
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
```

Multiclass Classifier Evaluator for ML evaluation

We can see that the accuracy of this model is even less than the random forest model.

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

Confusion Matrix for **Logistic regression model**

8.4 Naive Bayes model

For the fourth model, we decided to use the Logistic regression model from the Pyspark ML classification library with the model type being multinomial as we have three categories in our target.

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

Naive Bayes model implementation

We used the multiclassClassifierEvaluator available in the Pypark ML evaluation library to get the accuracy for this model.

```
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
```

Multiclass Classifier Evaluator for ML evaluation

We can see that the accuracy of this model is the worst among all four models.

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

Confusion Matrix for **Naive Bayes model**

::::::

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

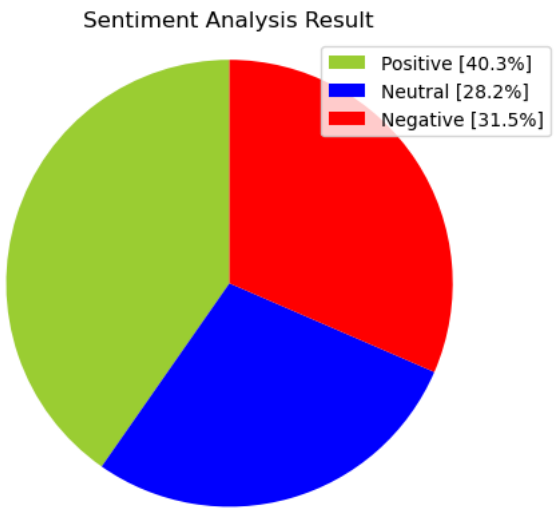
Confusion Matrix for the decision tree classifier after improvement

::

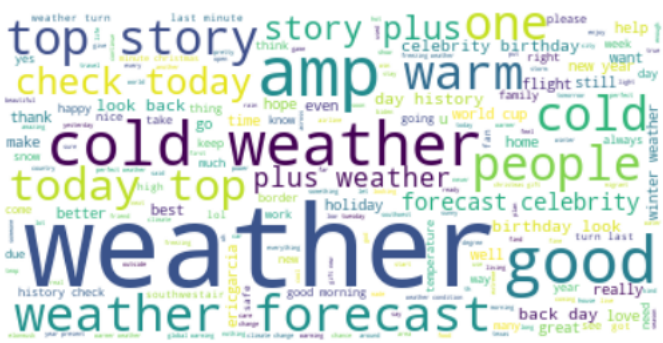
11. Apache Kafka weather-related tweets sentiment analysis application

Many factors contribute to the overall severity of an accident. The US accident dataset strongly relays on the weather factor. So, we implemented a sentiment analysis app for weather-related tweets using Kafka. We basically used Twitter stream and connected it to our Producer, to send the tweets to our Topic. The number of tweets collected is 5000 tweets.

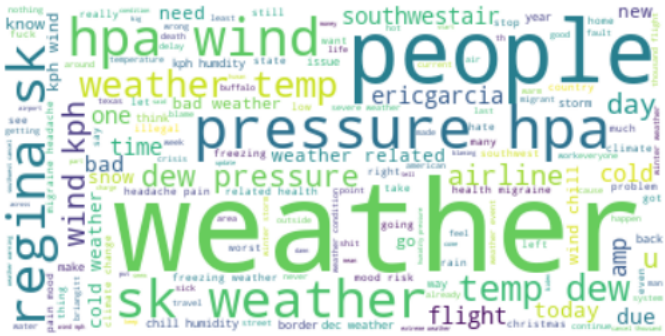
Pie Chart for representing polarity as proportions for the collected tweets



Word cloud for positive sentiments



word cloud for negative sentiments


weather
people
pressure
hpa
wind
sk
regina
weather
temp
ericgarcia
southwestair
airline
dew
pressure
weather related
flight
temp
dew
cold
day
new
need
know
one
time
bad
snow
today
due
u
amp
wind
kph
cold
weather
chill
humidity
dec
weather
border
christmas
weather
bad weather
kph humidity
freezing weather
mood risk
headache pain
related health
health migraine
storm
country
year
stop
home
fault
good
still
least
really
nothing
fuck
wind
kph
headache
migraine
mood
pain
thing
make
climate change
worst
going
go
left
rain
already
may
travel
stick
street
continue
system
man
even
water
outside
area
condition
shit
feel
case
weather
pressure
weather
system
system weather
weather
weather