PES UNIVERSITY
(Established under Karnataka Act No. 16 of 2013)
100 Feet Ring Road, BSK III Stage, Bengaluru-560 085
Department of Computer Science and Engineering
Session : Aug-Dec 2019

2019 - IDS Phase-2 Evaluation

Team Name:

Title: US Accidents

Member Name	USN	Section	#hrs effort	Team No

Date of Evaluation: **Evaluator:**

Note: Evaluate against the Evaluation scheme

Signature

US ACCIDENTS

This submission explores traffic accidents across 49 states within the United States over the past 3 years (from February 2016 to March 2019).

The dataset has 2.25 millon rows and 49 columns

It has following Columns:

ID:This is a unique identifier of the accident record.
Source:Indicates source of the accident report (i.e. the API which reported the accident.).
TMC:A traffic accident may have a Traffic Message Channel (TMC) code which provides more detailed description of the event.
Severity:Shows severity of the accident (a number between 1 and 4).
Start_Time:Shows start time of the accident in local time zone.
End_Time:Shows end time of the accident in local time zone.
Start_Lat:Shows latitude in GPS coordinate of the start point.
Start_Lng:Shows longitude in GPS coordinate of the start point.
End_Lat:Shows latitude in GPS coordinate of the end point.
End_Lng:Shows longitude in GPS coordinate of the end point.
Distance(mi):The length of the road extent affected by the accident.
Description:Shows natural language description of the accident.
Number:Shows the street number in address field.
Street:Shows the street name in address field.
Side:Shows the relative side of the street (Right/Left) in address field.
City:Shows the city in address field.
County

Shows the county in address field.

State

Shows the state in address field.

Zipcode

Shows the zipcode in address field.

Country

Shows the country in address field.

Timezone

Shows timezone based on the location of the accident (eastern, central, etc.).

Airport_Code

Denotes an airport-based weather station which is the closest one to location of the accident.

Weather_Timestamp
Shows the time-stamp of weather observation record (in local time).

Temperature(F)
Shows the temperature (in Fahrenheit).

Wind_Chill(F)
Shows the wind chill (in Fahrenheit).

Humidity(%)
Shows the humidity (in percentage).

Pressure(in)
Shows the air pressure (in inches).

Visibility(mi)
Shows visibility (in miles).

Wind_Direction
Shows wind direction.

Wind_Speed(mph)
Shows wind speed (in miles per hour).

Precipitation(in)
Shows precipitation amount in inches, if there is any.

Weather_Condition
Shows the weather condition (rain, snow, thunderstorm, fog, etc.)

Amenity
A POI annotation which indicates presence of amenity in a nearby location.

Bump
A POI annotation which indicates presence of speed bump or hump in a nearby location.

Crossing
A POI annotation which indicates presence of crossing in a nearby location.

Give_Way
A POI annotation which indicates presence of give_way in a nearby location.

Junction
A POI annotation which indicates presence of junction in a nearby location.

No_Exit
A POI annotation which indicates presence of no_exit in a nearby location.

Railway

A POI annotation which indicates presence of railway in a nearby location.

Roundabout
A POI annotation which indicates presence of roundabout in a nearby location.

Station
A POI annotation which indicates presence of station in a nearby location.

Stop
A POI annotation which indicates presence of stop in a nearby location.

Traffic_Calming
A POI annotation which indicates presence of traffic_calming in a nearby location.

Traffic_Signal
A POI annotation which indicates presence of traffic_signal in a nearby location.

Turning_Loop
A POI annotation which indicates presence of turning_loop in a nearby location.

Sunrise_Sunset
Shows the period of day (i.e. day or night) based on sunrise/sunset.

Civil_Twilight
Shows the period of day (i.e. day or night) based on civil twilight.

Nautical_Twilight
Shows the period of day (i.e. day or night) based on nautical twilight.

Astronomical_Twilight
Shows the period of day (i.e. day or night) based on astronomical twilight.

1.Data Cleaning:

The data set contains both categorical and numerical with missing values. So in order to do any sort of testing it is necessary to clean the data before. All the NAN's for categorical column is replaced with its previous values and all the NAN's for numerical column is replaced with average of the column. This was possible with help of python library pandas.

Temperat	Wind_Chi	Humidity(	Pressure(i	Visibility(	Wind_Dir
36.9		91	29.68	10	Calm
37.9		100	29.65	10	Calm
36	33.3	100	29.67	10	SW
35.1	31	96	29.64	9	SW
36	33.3	89	29.65	6	SW
37.9	35.5	97	29.63	7	SSW
34	31	100	29.66	7	WSW
34	31	100	29.66	7	WSW
33.3		99	29.67	5	SW
37.4	33.8	100	29.62	3	SSW
35.6	30.7	93	29.64	5	WNW
37.4	33.8	100	29.62	3	SSW
33.8		100	29.63	3	SW
36	31.1	89	29.65	10	NW
37.4	33.8	100	29.62	3	SSW
33.8		100	29.63	3	SW
35.6		99	29.65	7	WSW
36	31.1	89	29.65	10	NW
37.4	32.1	93	29.63	10	WSW
36	30.3	89	29.65	10	West
33.8	29.6	100	29.62	2	NNW
36	30.3	89	29.65	10	West
35.1	28.6	89	29.65	6	WSW

Temperat	Wind_Chi	Humidity(	Pressure(i	Visibility(	Wind_Dir
36.9	33.09473	91	29.68	10	Calm
37.9	33.09473	100	29.65	10	Calm
36	33.3	100	29.67	10	SW
35.1	31	96	29.64	9	SW
36	33.3	89	29.65	6	SW
37.9	35.5	97	29.63	7	SSW
34	31	100	29.66	7	WSW
34	31	100	29.66	7	WSW
33.3	33.09473	99	29.67	5	SW
37.4	33.8	100	29.62	3	SSW
35.6	30.7	93	29.64	5	WNW
37.4	33.8	100	29.62	3	SSW
33.8	33.09473	100	29.63	3	SW
36	31.1	89	29.65	10	NW
37.4	33.8	100	29.62	3	SSW
33.8	33.09473	100	29.63	3	SW
35.6	33.09473	99	29.65	7	WSW
36	31.1	89	29.65	10	NW
37.4	32.1	93	29.63	10	WSW

Normalization and Standardization:

Database normalization is the process of organizing the attributes of database to reduce or eliminate Data Redundancy (having same data but at different places) . We can design database without the normalization process, that is, without any organized way to design database. Such designs lack the standards and hence we cannot evaluate the database design. Normalization provides a systematic approach to determining a table's structure through a set of simple techniques through which we can achieve the desired table structure. Normalization ensures that attributes are grouped together in such a way that there is no redundancy or at least controlled redundancy.If we have data redundancies, then data in the database can create problems when we update the data. Such problems can lead to data inconsistencies, data integrity issues, and data update (add, insert, delete) anomalies. The bottom line is "each table should have only one source of data" if it is in the normal form. We aim at good database design. This means good table structures. Data stored in tables with controlled redundancy reduces data anomalies.

Standardized Dataset

Severity	Start_Lat	Start_Lng	Distance(r	Temperat	Wind_Chi	Humidity(	Pressure(i	Visibility(	Wind_Spe	Precipitat	Side_code	Amenity_	Bump_coc	Crossing_	Give_Way J
1.2322	2.316767	3.511128	-0.02862	-1.31515	-1.84E-15	1.106991	-1.05402	0.424601	1.83E-14	-0.50424	0.40782	-0.09741	-0.01	-0.19157	-0.0283
-0.80348	2.400703	3.634335	-0.02862	-1.25732	-1.84E-15	1.47707	-1.14906	0.424601	1.83E-14	-1.80216	-2.45206	-0.09741	-0.01	-0.19157	-0.0283
-0.80348	1.246753	3.51375	-0.02862	-1.36719	0.053021	1.47707	-1.0857	0.424601	-1.28724	6.75E-16	0.40782	-0.09741	-0.01	-0.19157	-0.0283
1.2322	2.160142	3.496388	-0.02862	-1.41923	-0.54107	1.312591	-1.18074	0.033596	-1.02311	6.75E-16	0.40782	-0.09741	-0.01	-0.19157	-0.0283
-0.80348	2.000077	3.498118	-0.02862	-1.36719	0.053021	1.024752	-1.14906	-1.13942	-1.28724	6.75E-16	0.40782	-0.09741	-0.01	-0.19157	-0.0283
1.2322	2.630891	3.624899	-0.02862	-1.25732	0.621285	1.35371	-1.21241	-0.74841	-1.28724	0.144729	0.40782	-0.09741	-0.01	-0.19157	-0.0283
-0.80348	2.174179	3.493887	-0.05376	-1.48284	-0.54107	1.47707	-1.11738	-0.74841	-1.28724	6.75E-16	0.40782	-0.09741	-0.01	-0.19157	-0.0283
1.2322	2.190333	3.49746	-0.02862	-1.48284	-0.54107	1.47707	-1.11738	-0.74841	-1.28724	6.75E-16	0.40782	-0.09741	-0.01	-0.19157	-0.0283
-0.80348	2.200578	3.499758	-0.05376	-1.52331	-1.84E-15	1.43595	-1.0857	-1.53042	-1.83951	6.75E-16	-2.45206	-0.09741	-0.01	-0.19157	-0.0283
1.2322	2.630891	3.624899	-0.02862	-1.28624	0.182172	1.47707	-1.24409	-2.31243	-1.02311	-0.50424	0.40782	-0.09741	-0.01	-0.19157	-0.0283
1.2322	2.433728	3.605418	-0.02862	-1.39032	-0.61856	1.189231	-1.18074	-1.53042	-0.73497	6.75E-16	0.40782	-0.09741	-0.01	-0.19157	-0.0283
1.2322	2.406907	3.634362	-0.02862	-1.28624	0.182172	1.47707	-1.24409	-2.31243	-1.02311	-0.50424	0.40782	-0.09741	-0.01	-0.19157	-0.0283
-0.80348	2.14664	3.501974	-0.05376	-1.4944	-1.84E-15	1.47707	-1.21241	-2.31243	-1.57538	6.75E-16	0.40782	-0.09741	-0.01	-0.19157	-0.0283
-0.80348	2.217521	3.492779	-0.02862	-1.36719	-0.51524	1.024752	-1.14906	0.424601	-0.73497	6.75E-16	-2.45206	-0.09741	-0.01	-0.19157	-0.0283
-0.80348	2.459379	3.626071	-0.02862	-1.28624	0.182172	1.47707	-1.24409	-2.31243	-1.02311	-0.50424	-2.45206	-0.09741	-0.01	-0.19157	-0.0283
-0.80348	2.157653	3.499919	-0.02862	-1.4944	-1.84E-15	1.47707	-1.21241	-2.31243	-1.57538	6.75E-16	0.40782	-0.09741	-0.01	-0.19157	-0.0283
-0.80348	2.16091	3.494539	-0.02862	-1.39032	-1.84E-15	1.43595	-1.14906	-0.74841	-1.57538	6.75E-16	0.40782	-0.09741	-0.01	-0.19157	-0.0283
-0.80348	2.16604	3.492939	-0.05376	-1.36719	-0.51524	1.024752	-1.14906	0.424601	-0.73497	6.75E-16	0.40782	-0.09741	-0.01	-0.19157	-0.0283
-0.80348	2.15069	3.498541	-0.02862	-1.28624	-0.25694	1.189231	-1.21241	0.424601	-0.47084	6.75E-16	-2.45206	-0.09741	-0.01	5.220018	-0.0283
-0.80348	2.217445	3.492486	-0.02862	-1.36719	-0.72188	1.024752	-1.14906	0.424601	-0.47084	6.75E-16	0.40782	-0.09741	-0.01	-0.19157	-0.0283
-0.80348	2.566742	3.629201	-0.05376	-1.4944	-0.9027	1.47707	-1.24409	-2.70344	-1.02311	-1.1532	0.40782	-0.09741	-0.01	-0.19157	-0.0283
-0.80348	2.194287	3.494471	-0.05376	-1.36719	-0.72188	1.024752	-1.14906	0.424601	-0.47084	6.75E-16	0.40782	-0.09741	-0.01	-0.19157	-0.0283
-0.80348	2.000753	3.494324	-0.02862	-1.41923	-1.161	1.024752	-1.14906	-1.13942	-0.1827	-0.50424	0.40782	-0.09741	-0.01	-0.19157	-0.0283
1.2322	2.528022	3.617904	-0.02862	-1.30937	-0.17945	1.312591	-1.21241	-0.35741	-0.73497	-1.80216	0.40782	-0.09741	-0.01	-0.19157	-0.0283

Normalized Dataset

Severity	Start_Lat	Start_Lng	Distance(mi)	Temperature(F)	Wind_Chill(F)	Humidity(%)	Pressure(in)	Visibility(mi)	Wind_Speed(mph)	Precipitation(in
0.666667	0.691971	0.940653	0.000416	0.323213	0.774155	0.902174	0.965954	0.122807	0.237915	0.04081
0.333333	0.704364	0.969701	0.000416	0.333007	0.774155	1.000000	0.964868	0.122807	0.237915	0.00000
0.333333	0.533978	0.941271	0.000416	0.314398	0.777985	1.000000	0.965592	0.122807	0.071429	0.05667
0.666667	0.668844	0.937178	0.000416	0.305583	0.735075	0.956522	0.964506	0.110276	0.105590	0.05667
0.333333	0.645210	0.937586	0.000416	0.314398	0.777985	0.880435	0.964868	0.072682	0.071429	0.05667
0.666667	0.738353	0.967476	0.000416	0.333007	0.819030	0.967391	0.964143	0.085213	0.071429	0.06122
0.333333	0.670917	0.936588	0.000000	0.294809	0.735075	1.000000	0.965230	0.085213	0.071429	0.05667
0.666667	0.673302	0.937431	0.000416	0.294809	0.735075	1.000000	0.965230	0.085213	0.071429	0.05667
0.333333	0.674815	0.937973	0.000000	0.287953	0.774155	0.989130	0.965592	0.060150	0.000000	0.05667
0.666667	0.738353	0.967476	0.000416	0.328110	0.787313	1.000000	0.963781	0.035088	0.105590	0.04081
0.666667	0.709241	0.962883	0.000416	0.310480	0.729478	0.923913	0.964506	0.060150	0.142857	0.05667
0.666667	0.705280	0.969707	0.000416	0.328110	0.787313	1.000000	0.963781	0.035088	0.105590	0.04081
0.333333	0.666851	0.938495	0.000000	0.292850	0.774155	1.000000	0.964143	0.035088	0.034161	0.05667
0.333333	0.677317	0.936327	0.000416	0.314398	0.736940	0.880435	0.964868	0.122807	0.142857	0.05667
0.333333	0.713028	0.967752	0.000416	0.328110	0.787313	1.000000	0.963781	0.035088	0.105590	0.04081
0.333333	0.668477	0.938010	0.000416	0.292850	0.774155	1.000000	0.964143	0.035088	0.034161	0.05667
0.333333	0.668958	0.936742	0.000416	0.310480	0.774155	0.989130	0.964868	0.085213	0.034161	0.05667
0.333333	0.669715	0.936365	0.000000	0.314398	0.736940	0.880435	0.964868	0.122807	0.142857	0.05667
0.333333	0.667449	0.937686	0.000416	0.328110	0.755597	0.923913	0.964143	0.122807	0.177019	0.05667

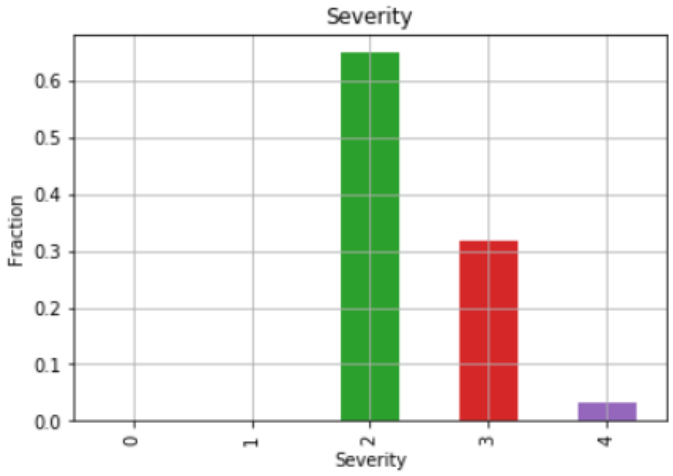
3.Graph Visualization.

The use of visual representations (i.e., pie charts,grouped graph,histogram) has been part of data science, and their use makes it possible for data scientists to interact with and represent complex phenomena, and come to conclusions easily.

Plotting graphs and making inferences

```
import datetime
print('There are {} accidents in the data'.format(len(df)))
df.Severity.value_counts(normalize=True).sort_index().plot.bar()
plt.grid()
plt.title('Severity')
plt.xlabel('Severity')
plt.ylabel('Fraction');
```

```
There are 2243939 accidents in the data
```

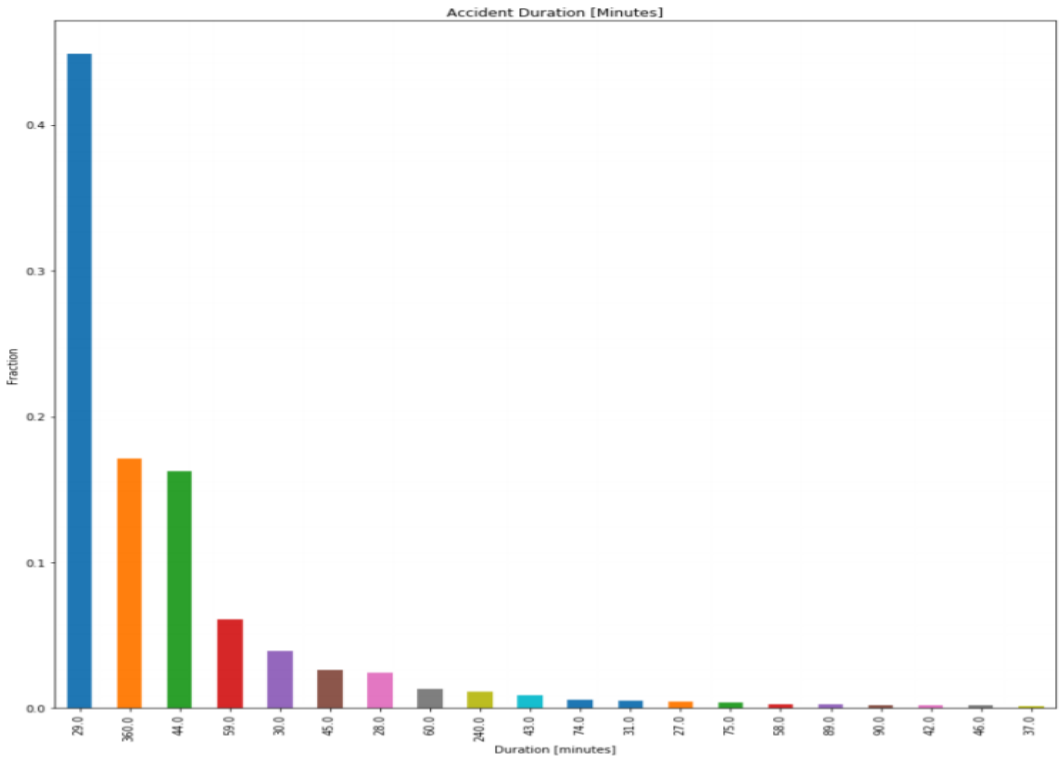



```
bool_cols = [col for col in df.columns if df[col].dtype ==np.dtype('bool')]
booldf = df[bool_cols]
not_one_hot = booldf[booldf.sum(axis=1) > 1]
print("There are {} non one hot metadata rows, which are {:.1f}% of the data".format(len
```

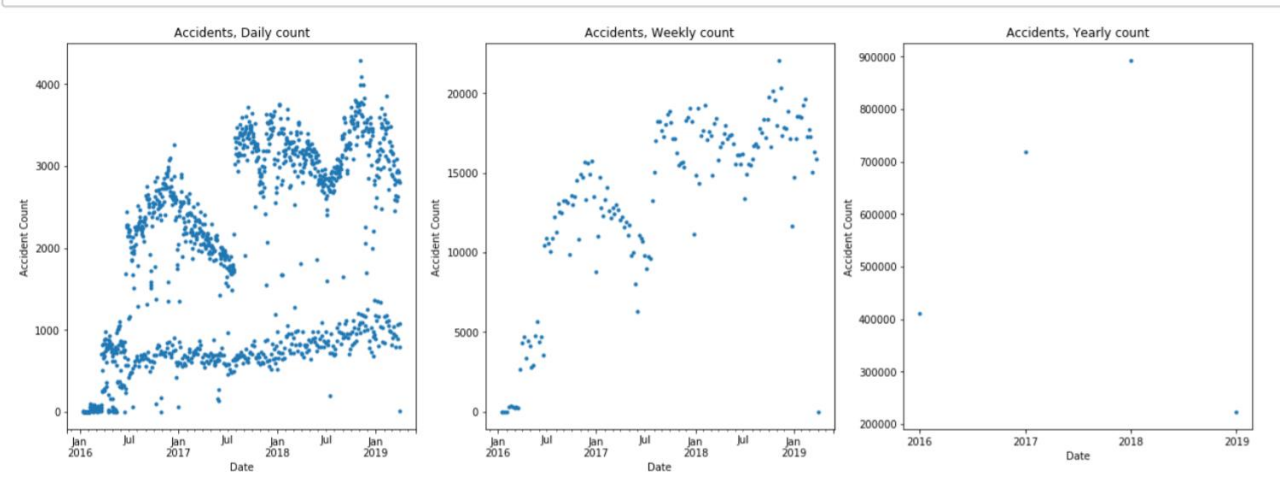
```
There are 134798 non one hot metadata rows, which are 6.0% of the data
```

The severity level can have only 3 values i.e., 2,3 and 4. And around 6% of the dataset are non one hot metadata rows.

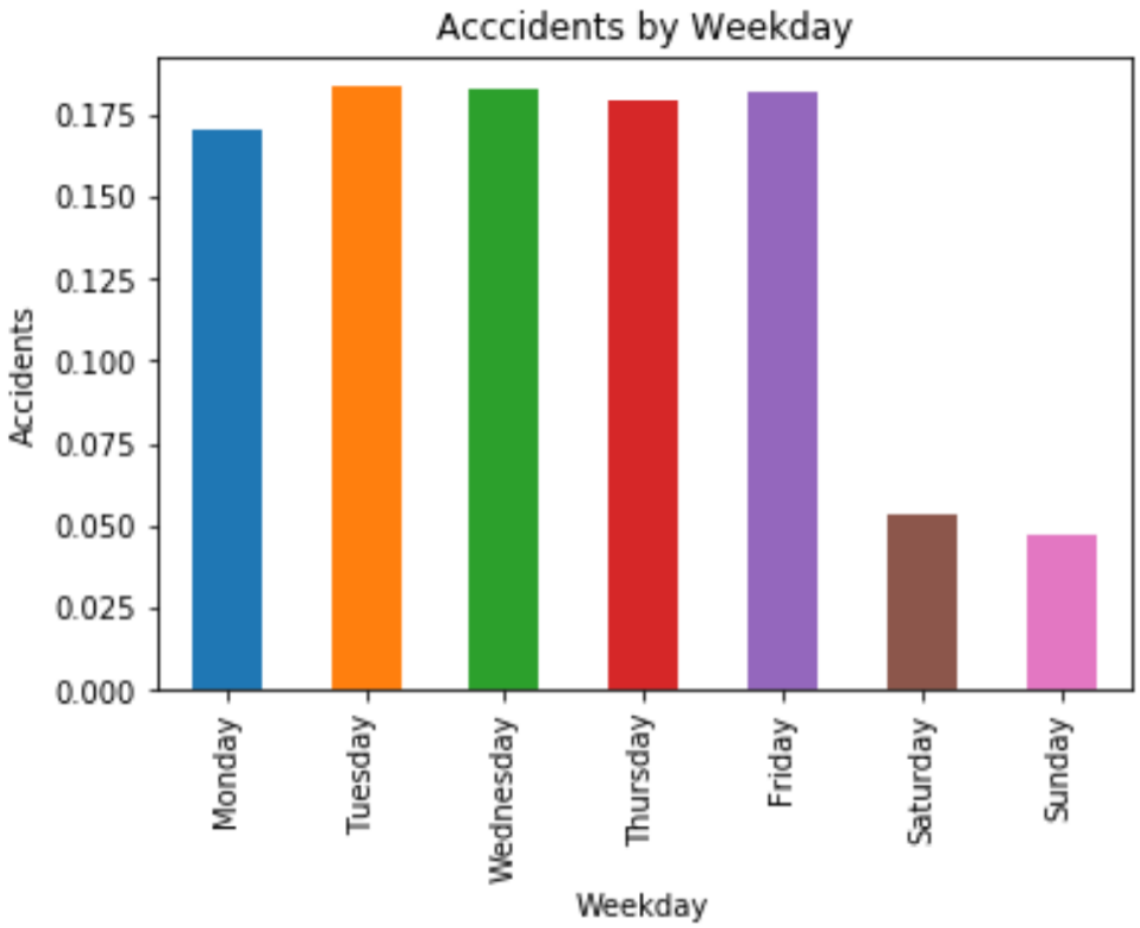
```
top 20 accident durations correspond to 96.1% of the data
Out[13]: Text(0, 0.5, 'Fraction')
```



Most common durations are 'Below half an hour', 'exactly 1h', 'Below 3/4 of an hour', 'Below 1h' respectively. Hence they are probably approximate, and probably correspond to the time it took to resolve the accident rather than the accident itself



The are two populations for days (prob. weekday and weekend) as they merge when we look at weeks. There is some seasonal pattern, and the amount increases each year.



From the graph it is clear that the number of accidents occur in weekdays is more compared to weekends.

5.Hypothesis testing:

Assuming

H0:Temperature and Pressure are dependent.

H1:Temperature and Pressure are independent.

```
df1=df.sample(n=100)
tables=pd.crosstab(df1['Temperature(F)'],df1['Pressure(in)'])
chi2, p, dof, expected=chi2_contingency(tables.values)
print('Chi-square statistic %0.3f p_value %0.3f' %(chi2,p))
```

```
Chi-square statistic 3726.111 p_value 0.130
```

Here p>0.1 so we accept the H0 i.e., Pressure and Temperature are related.

Assuming

H0:Temperature and Humidity are dependent.

H1:Temperature and Humidity are independent.

```
df1=df.sample(n=100)
tables=pd.crosstab(df1['Temperature(F)'],df1['Humidity(%)'])
chi2, p, dof, expected=chi2_contingency(tables.values)
print('Chi-square statistic %0.3f p_value %0.3f' %(chi2,p))
```

```
Chi-square statistic 3674.167 p_value 0.034
```

Here p<0.1 so we reject the H0 i.e., Humidity and Temperature are not related.

Assuming

H0:Severity and Visibility are dependent.

H1:Severity and Visibility are independent.

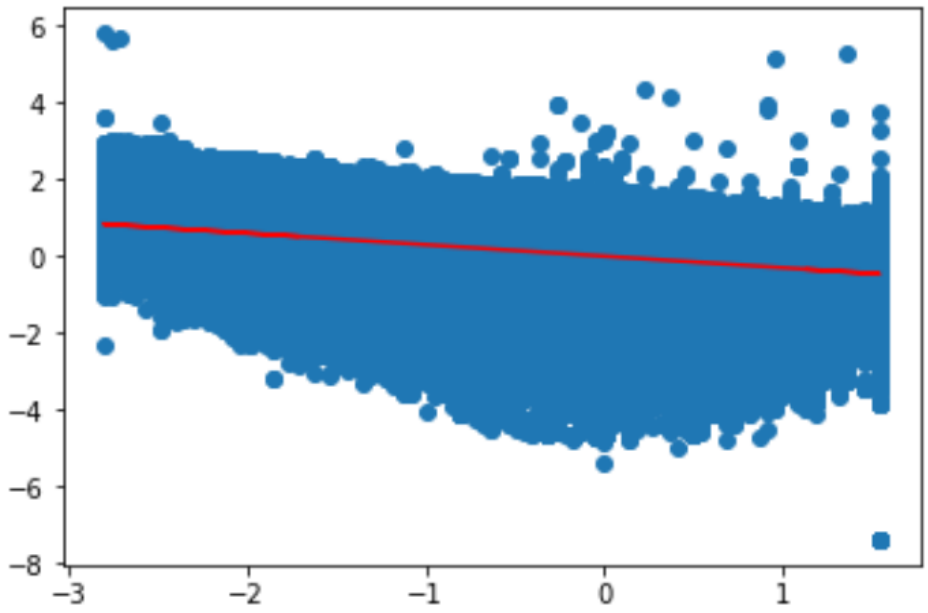
```
df1=df.sample(n=100)
tables=pd.crosstab(df1['Severity'],df1['Visibility(mi)'])
chi2, p, dof, expected=chi2_contingency(tables.values)
print('Chi-square statistic %0.3f p_value %0.3f' %(chi2,p))
```

```
Chi-square statistic 17.913 p_value 0.329
```

Here p>0.1 so we accept the H0 i.e., Severity and Visibility are related.

6.Correlation:

```
from sklearn.linear_model import LinearRegression
X = da.iloc[:, 6].values.reshape(-1, 1)  # values converts it in
Y = da.iloc[:, 4].values.reshape(-1, 1)  # -1 means that calcula
olumn
linear_regressor = LinearRegression()  # create object for the c
linear_regressor.fit(X, Y)  # perform linear regression
Y_pred = linear_regressor.predict(X)  # make predictions
plt.scatter(X, Y,)
plt.plot(X, Y_pred, color='red')
plt.show()
```



Only Humidity vs Temperature graph has negative correlation