GSBO city collision data analysis
Report 2

Report - 20 pts

- R1 Work done towards the tasks with links to notebooks, documents, etc. (10 pts)
- R2 Number of hours spent on each task. (5 pts)
- R3 Description of the results obtained from the tasks. (4-5 lines for each task) (5 pts)

Task 1 - Distribution modeling and hypothesis testing.(Vanteru Sahithi)
Number of hours spent: 45.

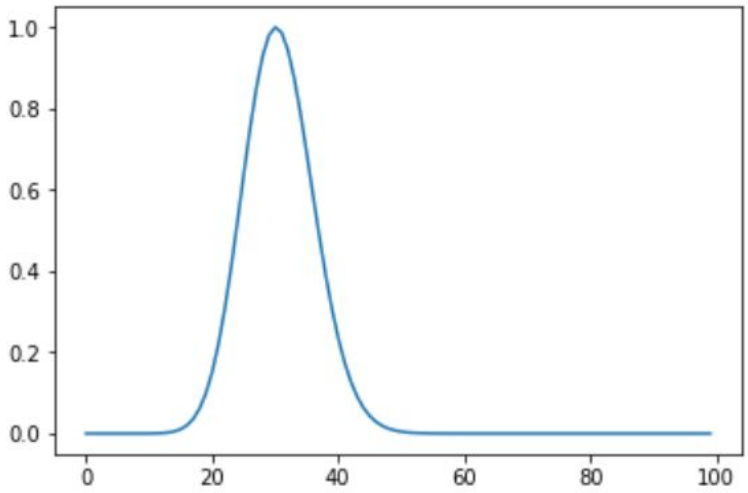
Task 1.1 Made a poisson distribution model for 2 sets of data:

1. Distribution of frequency of accidents during weekdays against the frequency of accidents during weekends.
2. Distribution of frequency of accidents in different time intervals.

➢ **Result :**

1. **Data set 1 :**
   Distribution for weekdays:
   Total_weekday_accidents: 39398
   Lambda_weekday = 30.612276612276613

```
Out[76]: [<matplotlib.lines.Line2D at 0x23cca1dba90>]
```



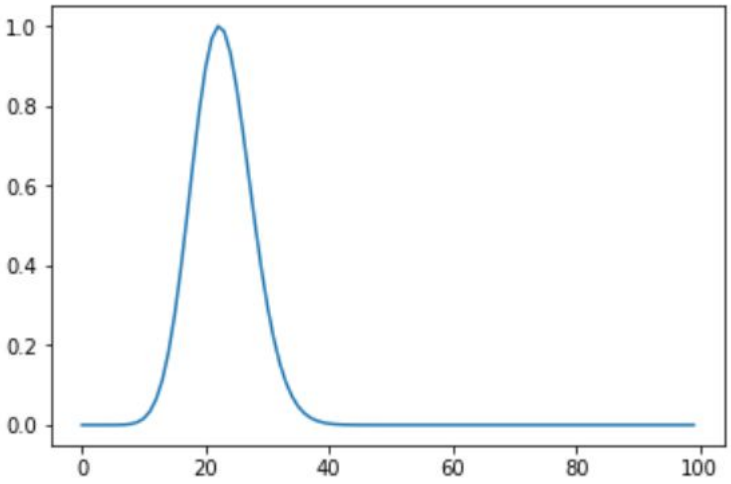
- For the above plot, to normalize probability to 0 -1 range, I did normalized distribution.
- From above plot, we can observe that the probability of occurrence of accidents in weekdays is around 30.6.

Distribution for weekends :

Total_weekend_accidents : 518

Lambda_weekend: 22.698841698841697

```
Out[78]: [<matplotlib.lines.Line2D at 0x23ccb6506d8>]
```



- From above plot, we can observe that the probability of occurrence of accidents in weekends is around 22.69 .
- From above both results, we can conclude that the probability of occurance of accidents in weekdays is greater.

T-test:

```
Out[124]: Ttest_indResult(statistic=-0.7819889926273857, pvalue=0.4342574560691521)
```

- The result above shows hypothesis testing of both the probability of weekdays and weekends distributions, the pvalue = 0.4 , which is greater than 5% significant value .i.e. 0.05,hence, null hypothesis is false.This tells us that the number of accidents are effected by whether they are occuring on weekdays and weekends.

2.Dataset 2 :

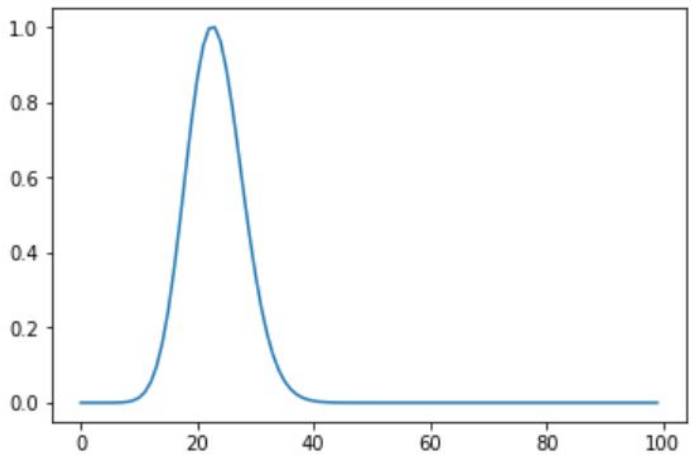
Distribution for time intervals.

- **Morning - 6am - 8pm**

  Number of total accidents in this time : 41586

  Lambda_daytiem = 23.077691453940066

```
Out[145]: [<matplotlib.lines.Line2D at 0x23cd20a86a0>]
```

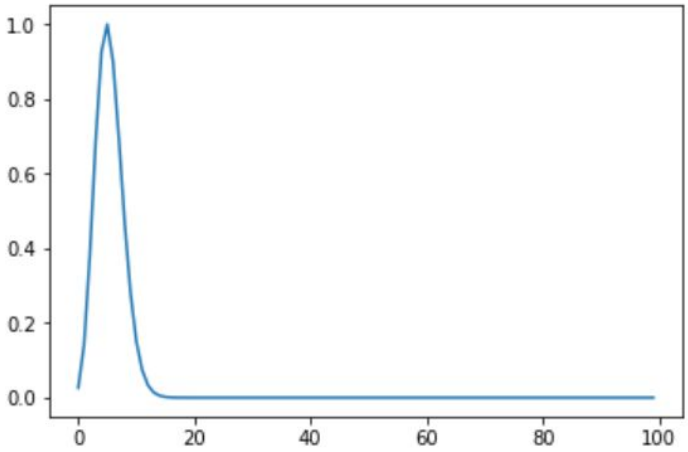


- From above plot, we can observe that the probability of occurance of accidents in the time interval 6 am - 8pm in a day is around 23.07.
- **Night - 8pm - 6am**

  Total number of accidents : 9570

  Lambda_night : 5.394588500563698

```
Out[146]: [<matplotlib.lines.Line2D at 0x23cd630ceb8>]
```



- From above plot, we can observe that the probability of occurance of accidents in the time interval 8pm - 6am in a day is around 5.39 and we can conclude that the probability for an accident to occur between 6 am to 8pm is greater.
-

T-Test :

```
Out[151]: Ttest_indResult(statistic=26.495338664053563, pvalue=2.194043308556357e-151)
```

- The above result shows us the T-test result where p value is less than 5% significant value i.e. 0.05, which means null hypothesis can be true, where it says that the frequency of accidents is not effected by whether it occurs in day or night.

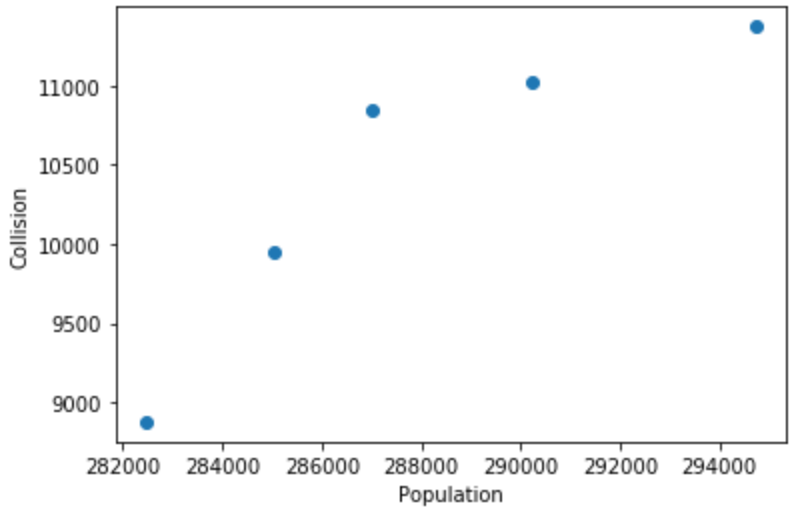
LINK TO NOTEBOOK :
https://github.com/UNCG-CSE/GSBO_City_Collision_Data_Analysis/blob/master/src/PoissonDistrVanteruSahithi.ipynb

Task 2 - Correlation between GSBO population and collisions by year and block (Jin Kang)
Number of hours spent: 50.

1. Data set 1

The number of collisions per 1000 people in Greensboro keeps increasing every year.

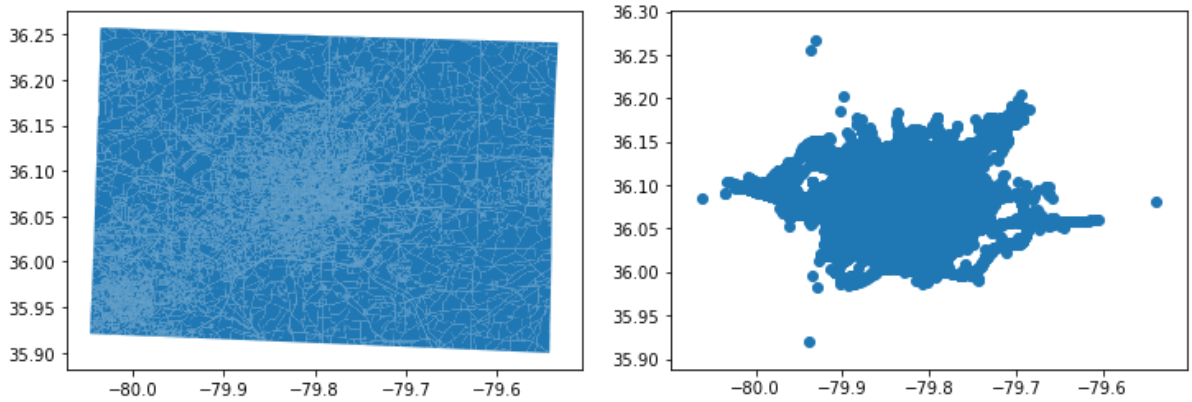
Year	Collisions per 1000 people
2014	31.42
2015	34.89
2016	37.77
2017	37.98
2018	38.59



The correlation between Greensboro population and collision looks quadratic rather than linear. But I should check R-squared value later so which regression is more fit for the data.

2. Data set 2

The left one is from a shapefile of Guilford county. And the right one is from x, y coordinate from collision data.



Once I figure out how to code to count the number of collisions in each block, I will use Chi-squared test for the null hypothesis: There is no difference between blocks.

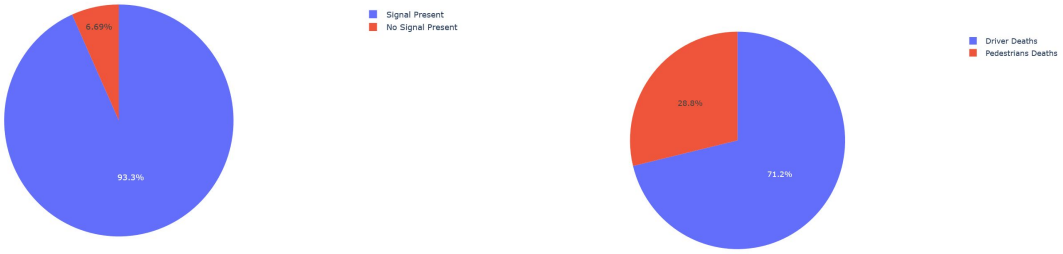
LINK TO NOTEBOOK :

https://github.com/UNCG-CSE/GSBO_City_Collision_Data_Analysis/blob/master/src/Project%20Review%202%20by%20Jin%20Kang.ipynb

Task 3 - Correlation between intersections by all collisions and collisions that a fatality occurred. The long term goal of this map is to be an interactive map that can show the collision hotspots in greensboro and then compare that to the population per square mile.
https://github.com/UNCG-CSE/GSBO_City_Collision_Data_Analysis/tree/Daniel

Number of hours spent: 45 hours

Extra: - pedestrian deaths vs driver deaths, and pedestrians hit with vs without traffic controller present.



Task 4 - Analysing effects of different types of collisions. Initially started with Pedestrian vs Vehicle collisions, but later added Cyclist collisions as some data were available.

4.1 - Data Cleaning

Hours spent: 8-10

There were many recording and other errors in the data. So when analysing categorical data, these values were throwing off numbers. Therefore, all the errors and mislabeled data were corrected.

Example:

Data such as the following were corrected individually.

```
DARK-RAODWAY NOT LIGHTED                                    1
SNOW                                                        1
RAIN                                                        1
DARK-ROADWAY NOT LIGHTEDDARK - ROADWAY NOT LIGHTED\r        1
DAYLGHT                                                     1
DARK-LIGHTED ORADWAY                                        1
DARK ROADWAY-NOT LIGHTED                                    1
BLOWING SAND, DIRT, SNOW                                    1
DARK - ROADWAY NOT LIGHTED\r                                1
DARK UNKNOWN LIGHTING                                       1
```

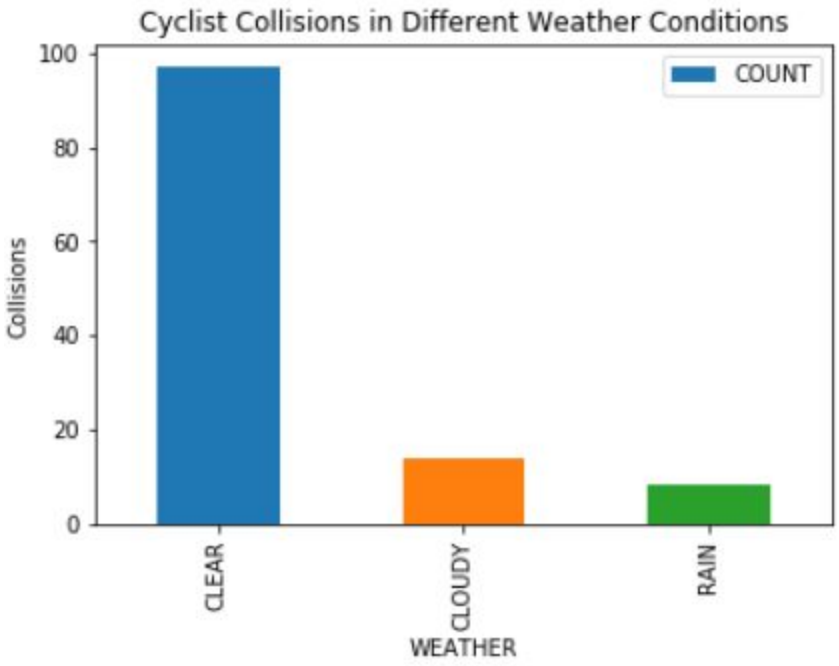
Result

```
DAYLIGHT                    36744
DARK-LIGHTED ROADWAY        10773
DARK - ROADWAY NOT LIGHTED   2172
DUSK                         1387
DAWN                          523
UNKNOWN                       194
DARK-UNKNOWN LIGHTING         154
OTHER                          23
CLOUDY                          4
SEVERE CROSSWINDS               1
BLOWING SAND, DIRT, SNOW        1
RAIN                            1
SNOW                            1
```

4.2 Statistical Analysis

Hours spent: ~20

Main consideration was given to categorical variables as most of the variables a categorial. Different hypotheses, correlations and basic statistics were tested. Did not conduct many correlation tests as most of the variables are categorical. Example: Hypothesis testing for independence between collision type and weather




```
df11 = df1[df1['WEATHER'].isin(['RAIN', 'CLOUDY', 'CLEAR'])]
df11_p = df11.pivot(index='BIPED', columns='WEATHER')
stats.chi2_contingency(df11_p)
```

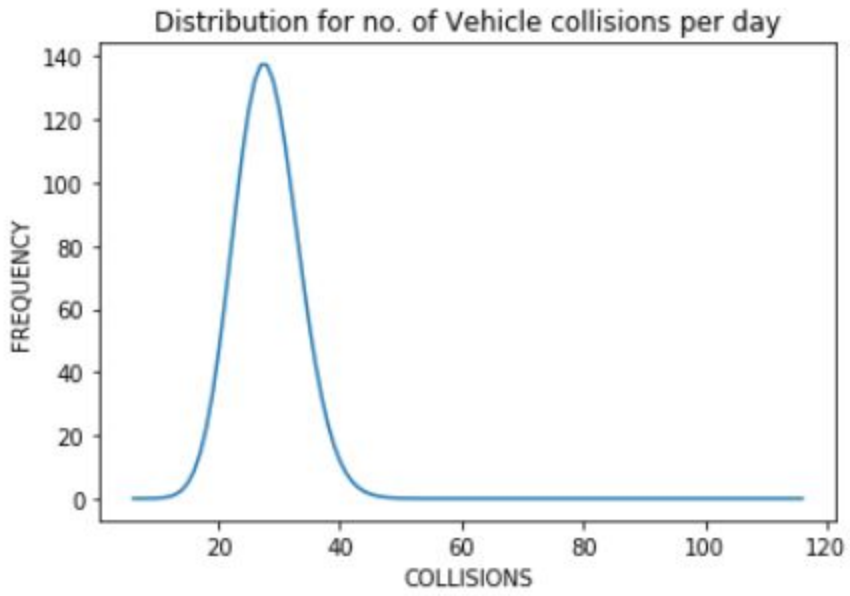
```
(7.103750029797511,
 0.13050603923391968,
 4,
 array([[8.78752983e+01, 1.75689577e+01, 1.35557440e+01],
        [3.67872583e+04, 7.35489719e+03, 5.67484453e+03],
        [5.67866423e+02, 1.13533853e+02, 8.75997239e+01]]))
```

Chi-Square test with a p-value 0.13 indicate that it is not possible to reject the null hypothesis of variable independence.

4.3 Distribution Analysis (still continuing)

Hours spent: ~10

Used the poisson distribution to analyse the distributions of collision types in a specific time period. Maximum likelihood method was used to get the lambda values and ks test is being used to test the goodness of fit.




```
lamB = ds[ds.BIPED=='B'].COUNT.mean()
lamP = ds[ds.BIPED=='P'].COUNT.mean()
lamM = ds[ds.BIPED=='M'].COUNT.mean()

print('Distribution for no. of Vehicle collisions per day is: Poisson(',lamM,')')
print('Distribution for no. of Pedestrian collisions per day is: Poisson(',lamP,')')
print('Distribution for no. of Cyclist collisions per day is: Poisson(',lamB,')')
```

```
Distribution for no. of Vehicle collisions per day is: Poisson( 27.99069003285871 )
Distribution for no. of Pedestrian collisions per day is: Poisson( 1.270935960591133 )
Distribution for no. of Cyclist collisions per day is: Poisson( 1.0818181818181818 )
```

Link to Notebook:

https://github.com/UNCG-CSE/GSBO_City_Collision_Data_Analysis/blob/master/src/Pedestrian.ipynb

All data can be found in my branch of the repository:
https://github.com/UNCG-CSE/GSBO_City_Collision_Data_Analysis/tree/jorge/src

2 hours looking for weather data

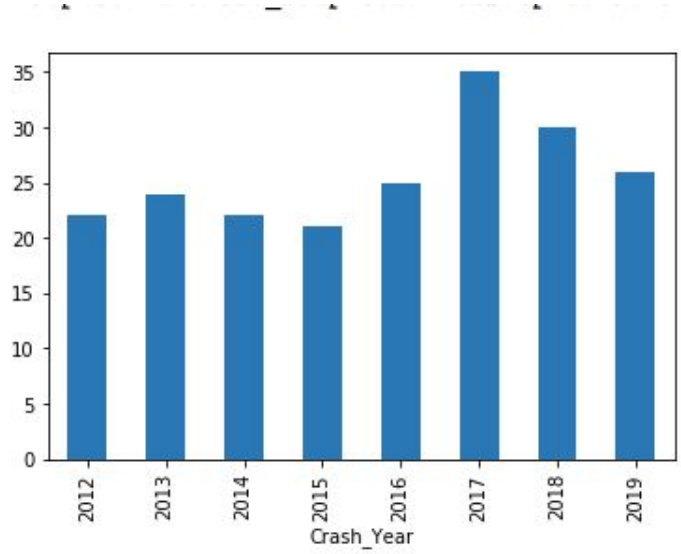
```
day_tables["2012"]["10"]["10"]
```

	city_id	main	wind	clouds	weather	dt	dt_iso	rain	snow
203	4169510	{'temp': 295.8, 'temp_min': 294.82, 'temp_max'...	{'speed': 1, 'deg': 40}	{'all': 1}	[{'id': 800, 'main': 'Clear', 'description': '...	1349827200	2012-10-10 00:00:00 +0000 UTC	NaN	NaN
204	4169510	{'temp': 294.63, 'temp_min': 293.71, 'temp_max...	{'speed': 0, 'deg': 0}	{'all': 1}	[{'id': 800, 'main': 'Clear', 'description': '...	1349830800	2012-10-10 01:00:00 +0000 UTC	NaN	NaN
205	4169510	{'temp': 294.37, 'temp_min': 293.71, 'temp_max...	{'speed': 0, 'deg': 0}	{'all': 1}	[{'id': 800, 'main': 'Clear', 'description': '...	1349834400	2012-10-10 02:00:00 +0000 UTC	NaN	NaN
206	4169510	{'temp': 293.95, 'temp_min': 292.59, 'temp_max...	{'speed': 0, 'deg': 0}	{'all': 1}	[{'id': 800, 'main': 'Clear', 'description': '...	1349838000	2012-10-10 03:00:00 +0000 UTC	NaN	NaN
207	4169510	{'temp': 293.79, 'temp_min': 292.59, 'temp_max...	{'speed': 0, 'deg': 0}	{'all': 75}	[{'id': 803, 'main': 'Clouds', 'description': ...	1349841600	2012-10-10 04:00:00 +0000 UTC	NaN	NaN
...	...	...	...	...	...	...	...	...	...
247946	4815207	{'temp': 287.21, 'temp_min': 287.15, 'temp_max...	{'speed': 0, 'deg': 271}	{'all': 0}	[{'id': 800, 'main': 'Clear', 'description': '...	1349895600	2012-10-10 19:00:00 +0000 UTC	{'1h': 0, 'today': 0}	NaN
247947	4815207	{'temp': 287.14, 'temp_min': 287.04, 'temp_max...	{'speed': 0, 'deg': 235}	{'all': 0}	[{'id': 800, 'main': 'Clear', 'description': '...	1349899200	2012-10-10 20:00:00 +0000 UTC	{'1h': 0, 'today': 0}	NaN
247948	4815207	{'temp': 287.59, 'temp_min': 287.59, 'temp_max...	{'speed': 0, 'deg': 130}	{'all': 0}	[{'id': 800, 'main': 'Clear', 'description': '...	1349902800	2012-10-10 21:00:00 +0000 UTC	{'1h': 0, 'today': 0}	NaN
247949	4815207	{'temp': 287.845, 'temp_min': 287.845, 'temp_m...	{'speed': 0, 'deg': 191}	{'all': 0}	[{'id': 800, 'main': 'Clear', 'description': '...	1349906400	2012-10-10 22:00:00 +0000 UTC	{'1h': 0, 'today': 0}	NaN
247950	4815207	{'temp': 288.1, 'temp_min': 288.1, 'temp_max':...	{'speed': 1, 'deg': 252}	{'all': 0}	[{'id': 800, 'main': 'Clear', 'description': '...	1349910000	2012-10-10 23:00:00 +0000 UTC	NaN	NaN

120 rows × 9 columns

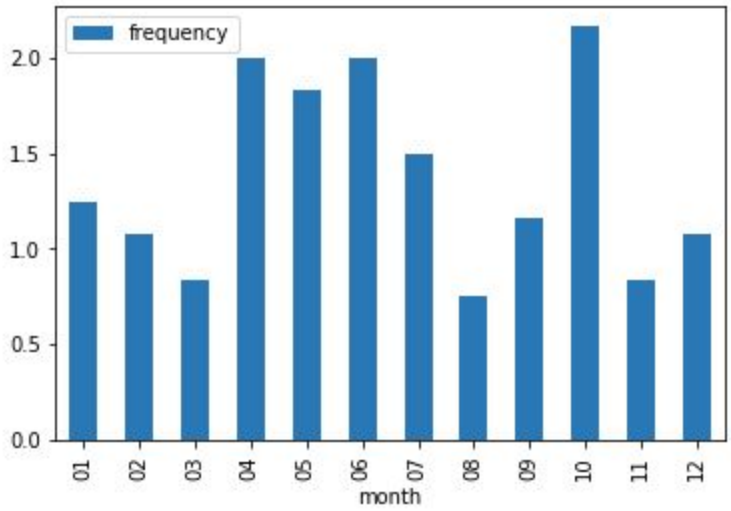
Bought weather data and analyzed fatality statistics.

1 hour analyzing total number of fatalities per year



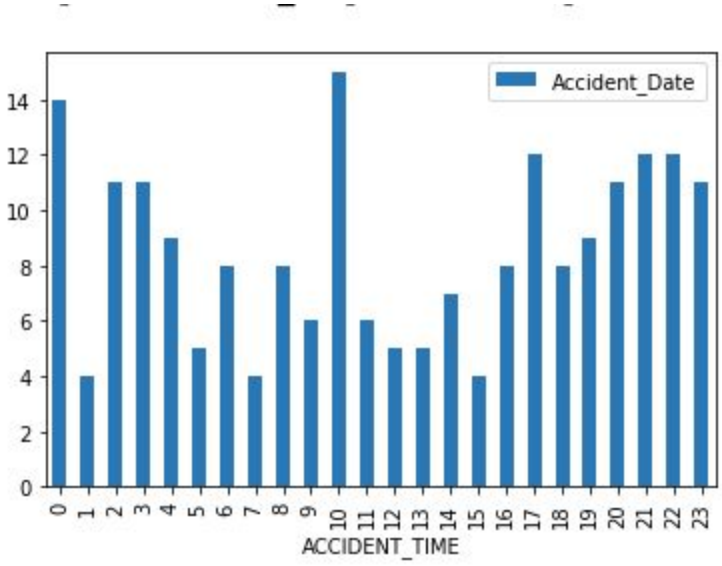
2017 had the maximum number of deaths give the interval while 2015 the lowest. Something unique might have happened in those months.

4 Hours analyzing average fatalities from 2012-2019 by month



2 hours
So some months have double the fatalities but I think the data needs a second look in terms of how the fatalities in a month change over the years.

4 hours analyzing fatalities frequency by time of data



Most deaths occur at 10 in the morning. While the minimum happens at 1 am in te morning. More needs to be looked at.

There is still a lot of work to be done. The data is very vague. Given other responsibilities its difficult to make data analysis full time.