Air plane Crash

Air travel has become an integral part of our lives. The air traffic growth is estimated to increase at an annual rate of 5.5%. With such a large system and the number of people opting for this faster mode of transportation, safety concerns become paramount. Most people think of only the aircraft when it comes to travel. But indeed it involves a synergistic and synchronized working of the air operations, air traffic control, crews, airports and weather and security services. To have a closer look at the tragic reality of airplane crashes, my project consists of various different statistics on airplane crashes between 1908 – 2009.

These statistics are helpful for the people who are interested in analysing various reasons for air plane crashes and take necessary steps to prevent it. With the necessary steps taken to prevent the air crash, the crashes might reduce in the future and make the air travel a safe place.

Data Analysis

Air travel has become an integral part of our lives. The air traffic growth is estimated to increase at an annual rate of 5.5%. With such a large system and the number of people opting for this faster mode of transportation, safety concerns become paramount. Most people think of only the aircraft when it comes to travel. But indeed it involves a synergistic and synchronized working of the air operations, air traffic control, crews, airports and weather and security services. To have a closer look at the tragic reality of airplane crashes, my project consists of various different statistics on airplane crashes between 1908 – 2009.

Metadata:
The metadata of the dataset is as per the picture here.
Crash calculations are made with respect to important attributes from this dataset.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5268 entries, 0 to 5267
Data columns (total 13 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Date          5268 non-null   object
 1   Time          3049 non-null   object
 2   Location      5248 non-null   object
 3   Operator      5250 non-null   object
 4   Flight #      1069 non-null   object
 5   Route         3562 non-null   object
 6   Type          5241 non-null   object
 7   Registration  4933 non-null   object
 8   cn/In         4040 non-null   object
 9   Aboard        5246 non-null   float64
 10  Fatalities    5256 non-null   float64
 11  Ground        5246 non-null   float64
 12  Summary       4878 non-null   object
dtypes: float64(3), object(10)
memory usage: 535.2+ KB
```

Sample Data:

The sample data looks like this

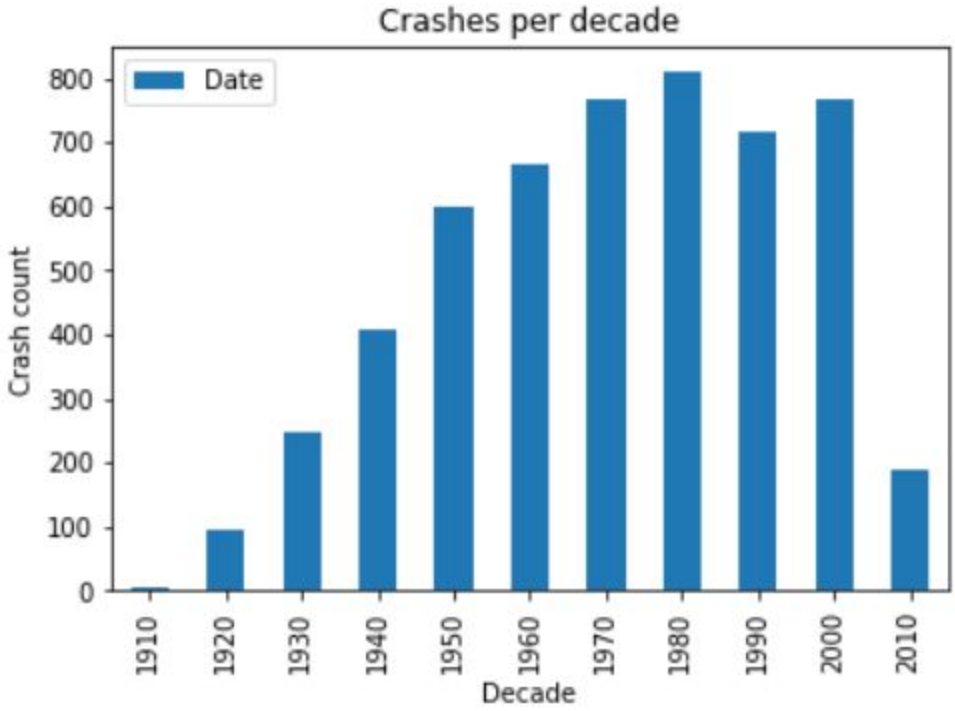
	Date	Time	Location	Operator	Flight #	Route	Type	Registration	cn/In	Aboard	Fatalities	Ground	Summary
0	09/17/1908	17:18	Fort Myer, Virginia	Military - U.S. Army	NaN	Demonstration	Wright Flyer III	NaN	1	2.0	1.0	0.0	During a demonstration flight, a U.S. Army fly...
1	07/12/1912	06:30	AtlantiCity, New Jersey	Military - U.S. Navy	NaN	Test flight	Dirigible	NaN	NaN	5.0	5.0	0.0	First U.S. dirigible Akron exploded just offsh...
2	08/06/1913	NaN	Victoria, British Columbia, Canada	Private	-	NaN	Curtiss seaplane	NaN	NaN	1.0	1.0	0.0	The first fatal airplane accident in Canada oc...
3	09/09/1913	18:30	Over the North Sea	Military - German Navy	NaN	NaN	Zeppelin L-1 (airship)	NaN	NaN	20.0	14.0	0.0	The airship flew into a thunderstorm and encou...
4	10/17/1913	10:30	Near Johannisthal, Germany	Military - German Navy	NaN	NaN	Zeppelin L-2 (airship)	NaN	NaN	30.0	30.0	0.0	Hydrogen gas which was being vented was sucked...

Crash v Decade:

The graph represents the number of crashes occured per decade from 1908-2009.

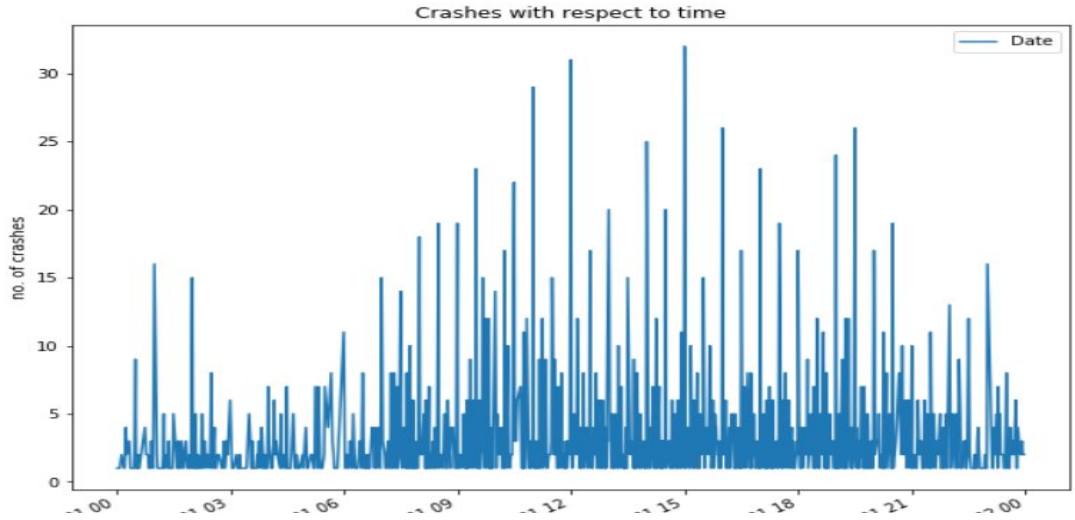
It is observed that crashes increased eventually from the 1910's to 1980’s and showed a drastic reduction in 2010’s. It is observed that most crashes have occurred during the 1980’s.

Significantly less number of crashes occured in the 2010’s. It could be due to the improvement of the technology.



Crashes v Time:

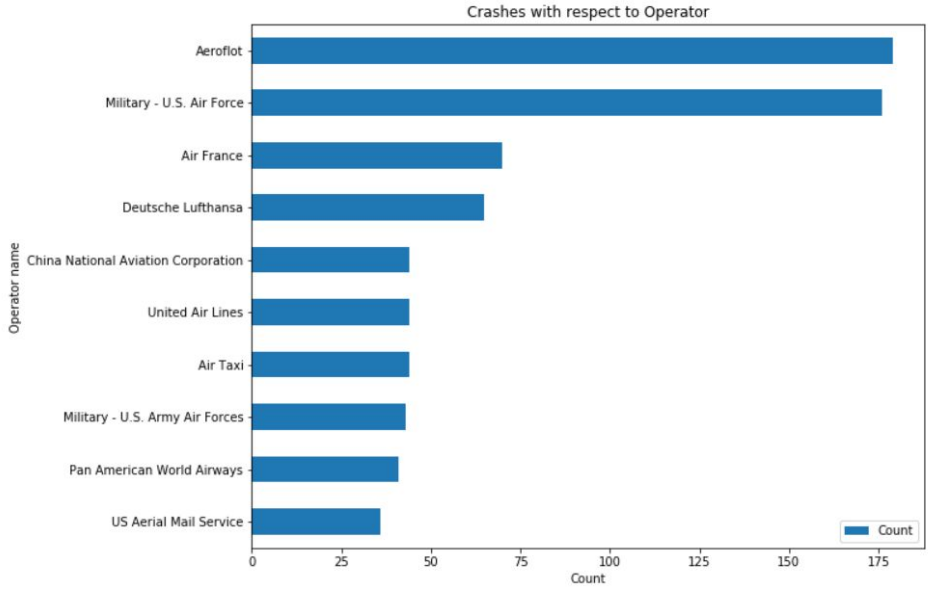
It is observed that most of the crashes happened during the mid day in the sunlight. Around 12PM and 15PM. Such an irony.



Crashes v Operator:

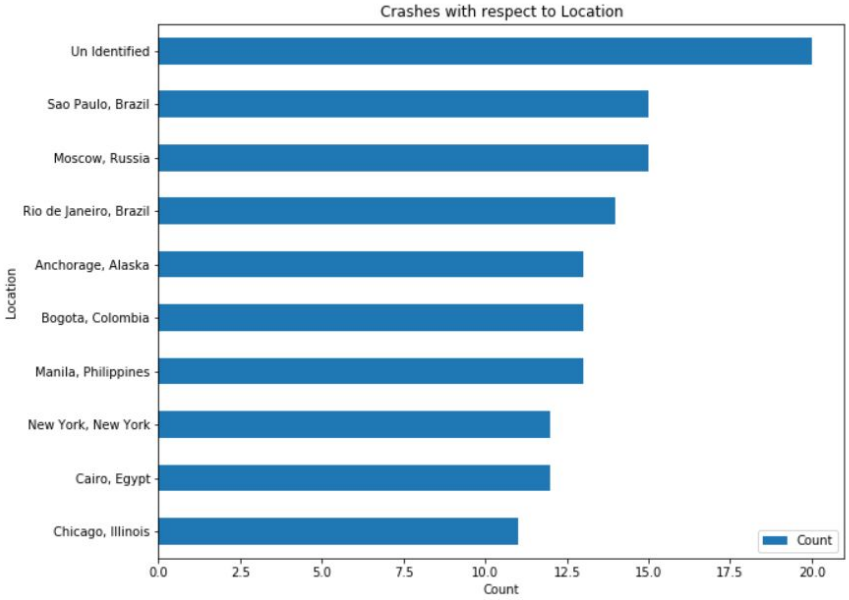
When a graph is plotted for Crash counts v Operator, it is observed that most of the crashes happened with the operator Aeroflot and Military - U.S Air Force as expected with around 175 crashes each. This could be due to the wars.

The graph here is displaying the top 10 Operators with the highest number of crashes. There are other private operators involved in the Crashes.



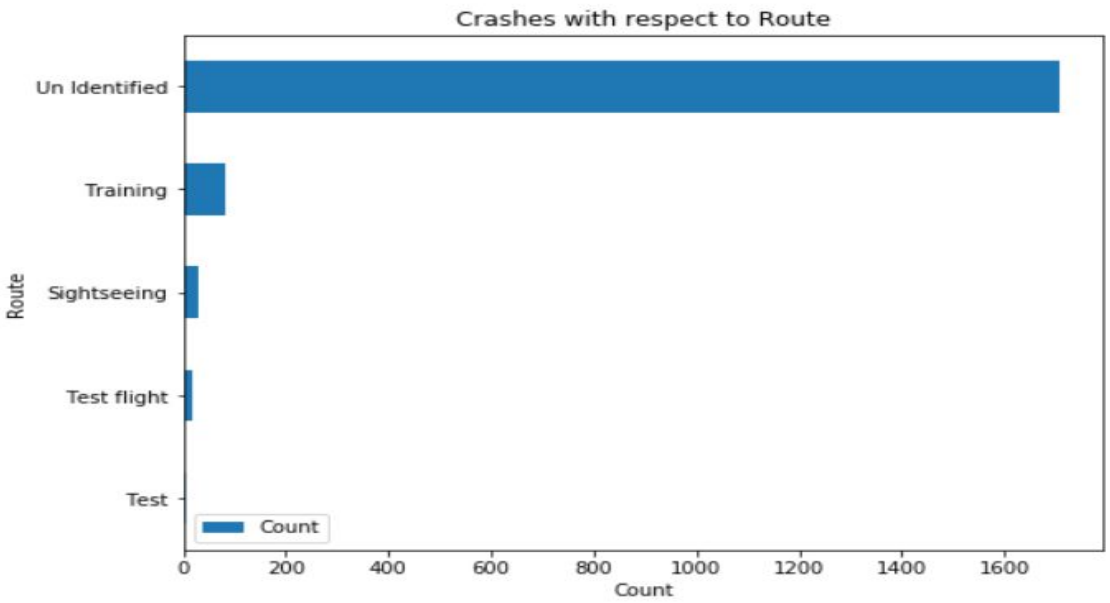
Crashes v Location:

From the graph plotted between the number of crashes and Location, it is observed that most of the Locations are unidentified. Next goes to Brazil and Russia. Again, this plot is just the top 10 Locations. The Dataset has many more locations.



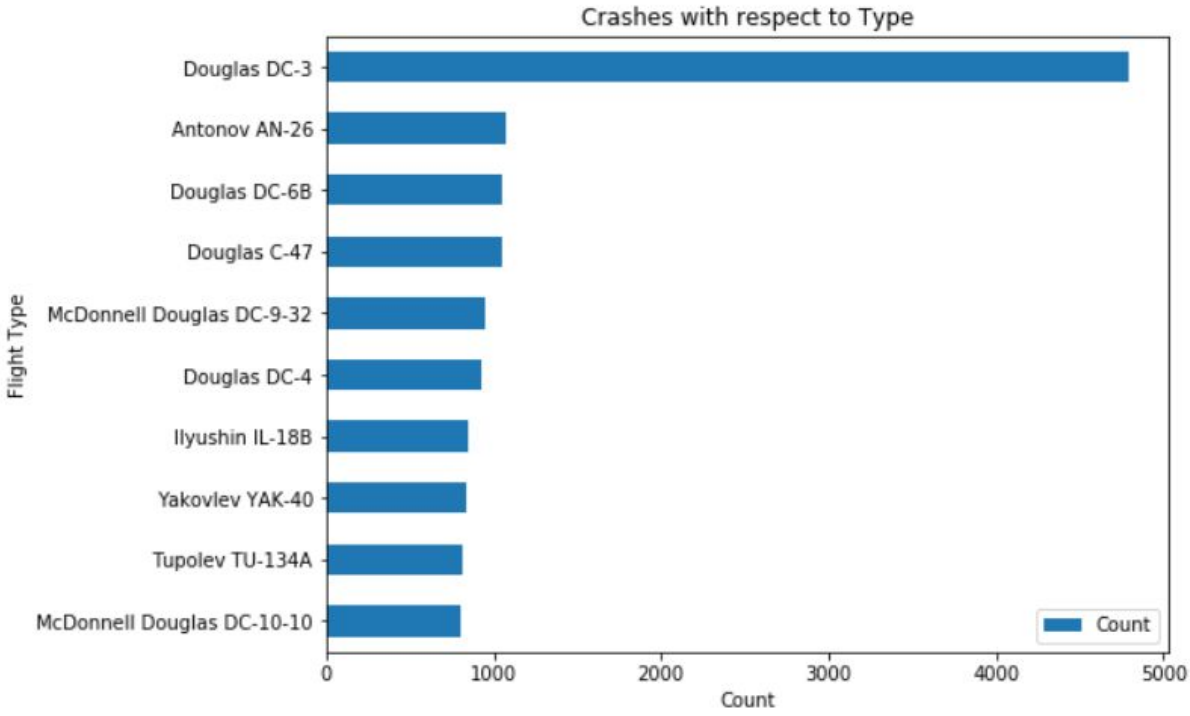
Crash v Route:

Again, it is observed from the graph plotted between the crash counts and route that the majority of the crashes occurred in an unidentified route. Second place goes to Training.



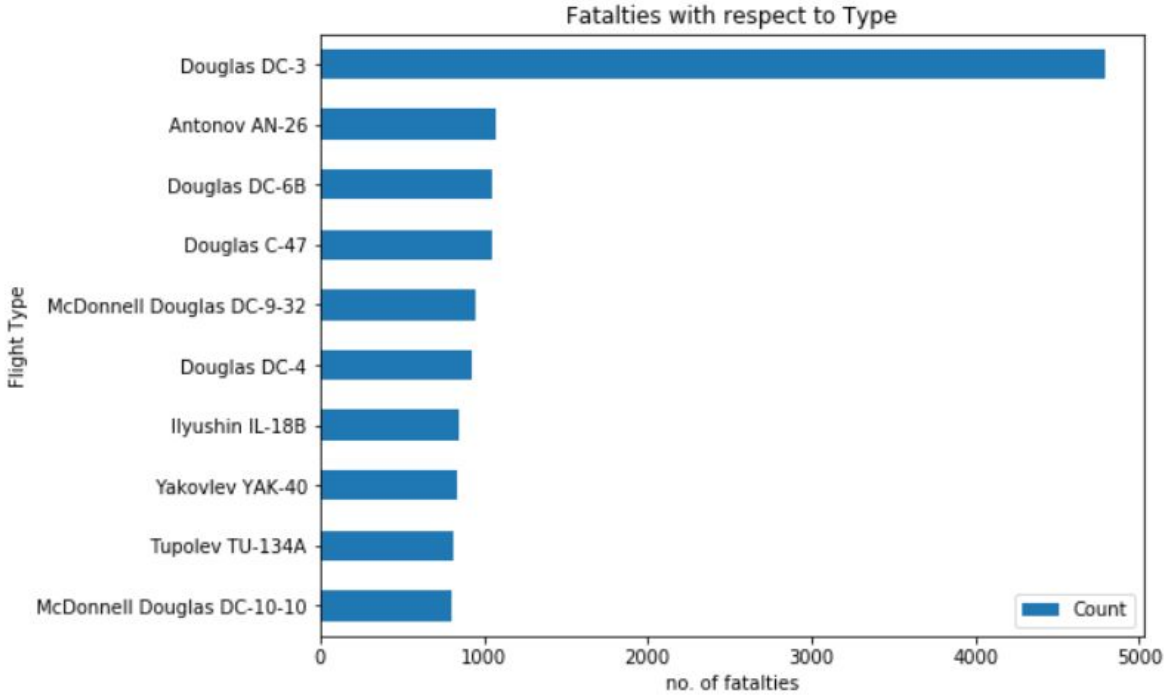
Crash v Flight Type:

From the plot generated between Crash count and Flight type, it is observed that most of the crashes occurred with the Flight type *Douglas DC - 3*. There is a possibility that mostly this flight is used in wars considering the previous graphs.



Fatalties v Flight Type:

The graph plotted between the number of fatalities occurred v Flight type is similar to the Crashes v Flight type. Most of the fatalities occurred during the *Douglas DC - 3* crash.

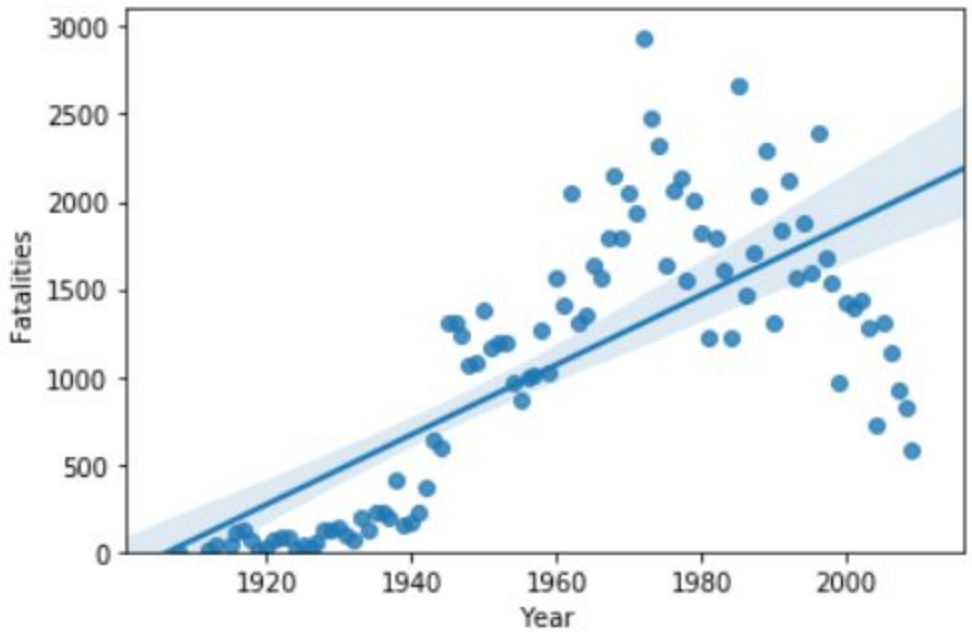


Methodology:

Fatalities v Year

Regression Plot:

Plotted a graph between aggregate of Fatalities and Years between 1908 and 2009.
It is observed that the rise in fatalities started around 1960s and hiked around mid 1970's and then eventually the number of fatalities has decreased.



P-Value:

The Pearson Coefficient value for the number of fatalities vs year is 1.0843863908885843e-17

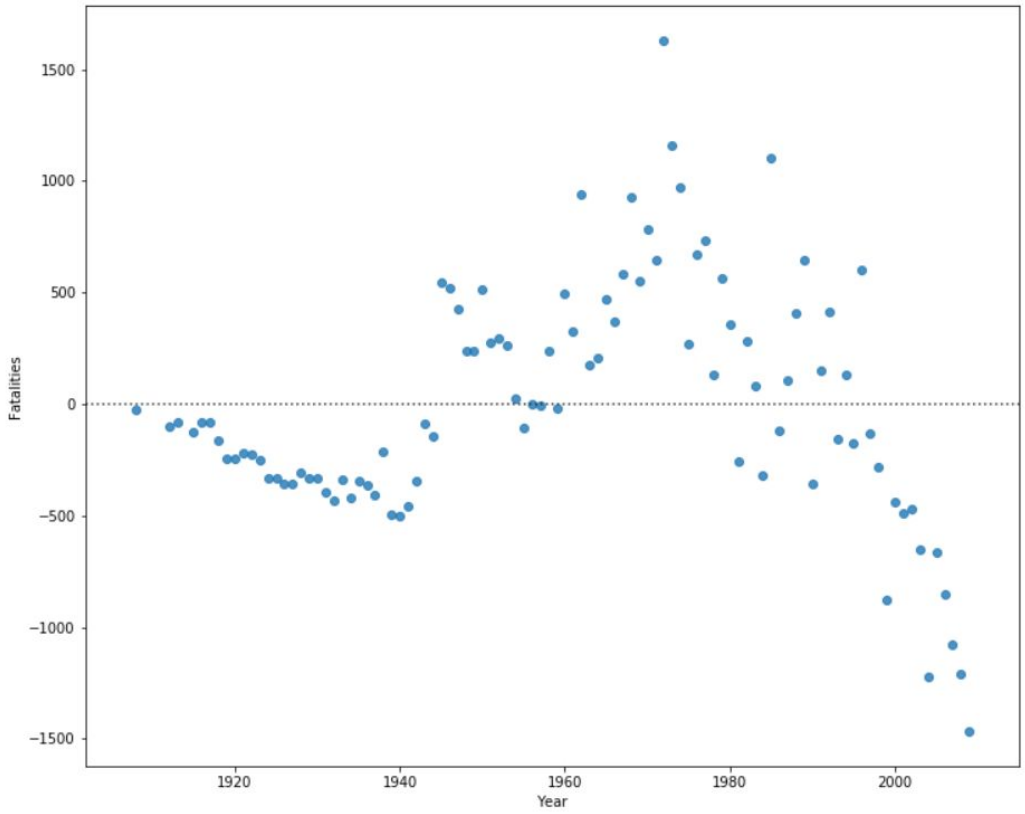
```
pearson_coef, p_value = stats.pearsonr(Year_Fatalities_dataset['Year'], Year_Fatalities_dataset['Fatalities'])
print("The Pearson Correlation Coefficient is", pearson_coef, " with a P-value of P =", p_value)
```

```
The Pearson Correlation Coefficient is 0.7321053374292882  with a P-value of P = 1.0843863908885843e-17
```

Since the p-value is < 0.001, the correlation between Fatalities and Year is statistically significant, and the linear relationship is quite strong (~0.732, close to 1)

Residual Plot:

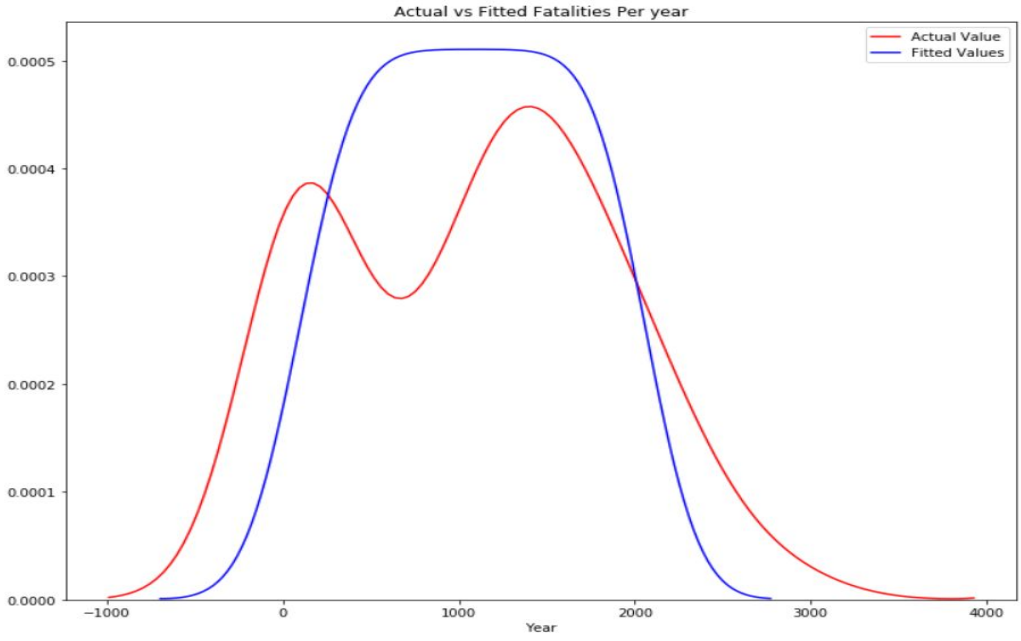
The residual plot between the number of fatalities and the years is here.



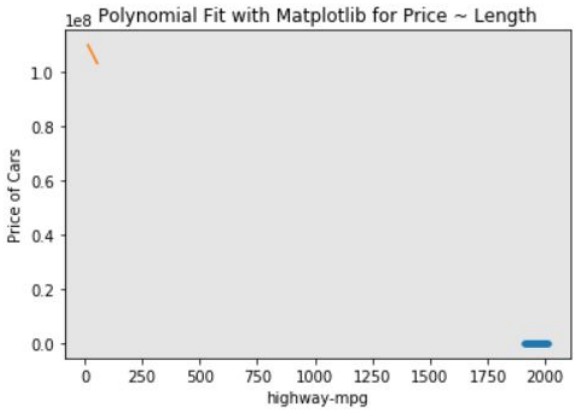
Multiple Linear Regression:

The equation used in this regression is

$$Yhat = a + b\_1\ X\_1 + b\_2\ X\_2 + b\_3\ X\_3 + b\_4\ X\_4$$



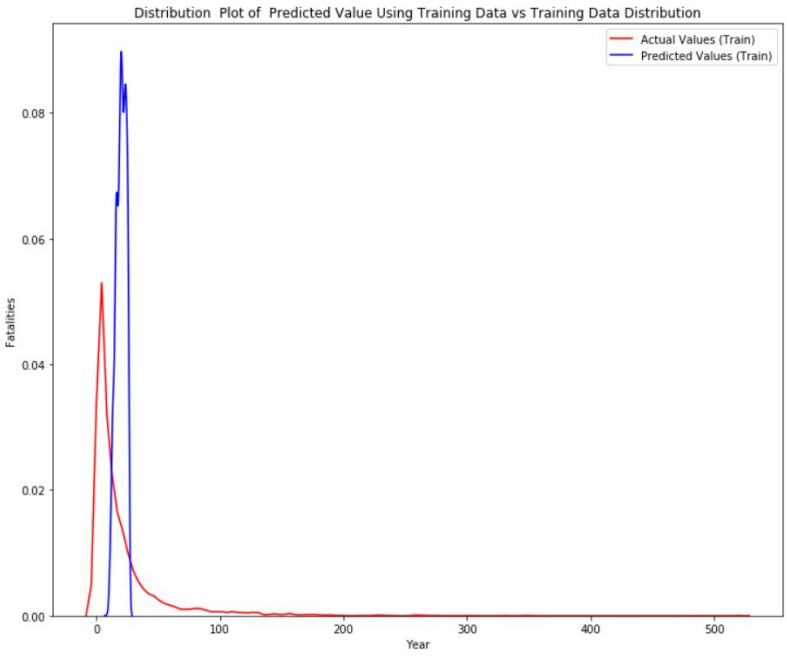
Polynomial Regression:



Model Evaluation:

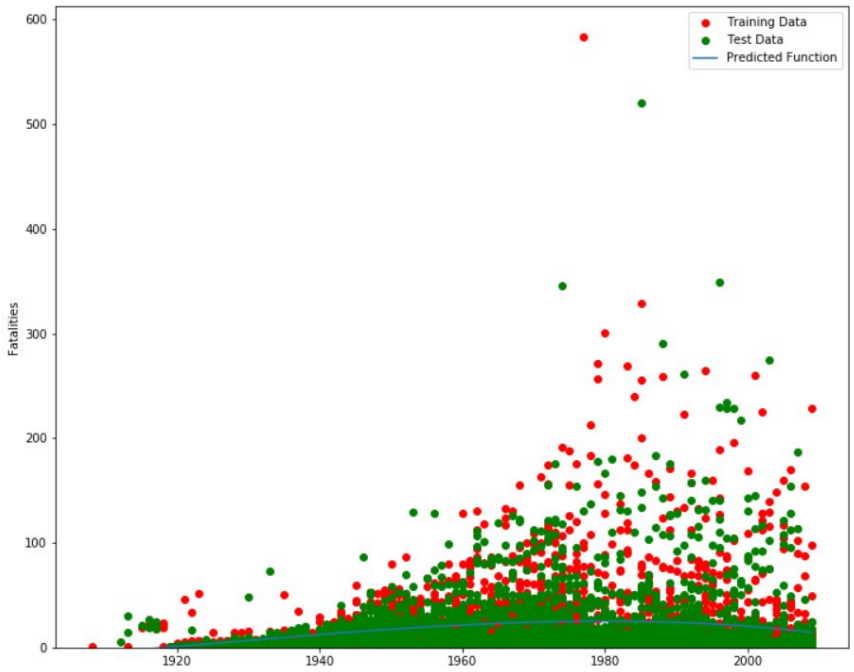
By doing testing and training, Evaluated the R2 data values for test and train data.
Here are the values:
number of test samples : 10
number of training samples: 88
R2 value for train data: 0.557856432337242
R2 value for test data: 0.329951606386753

Here is the Plot of predicted values using the training data compared to the training data. This model seems to be more sensible than other models with less error value



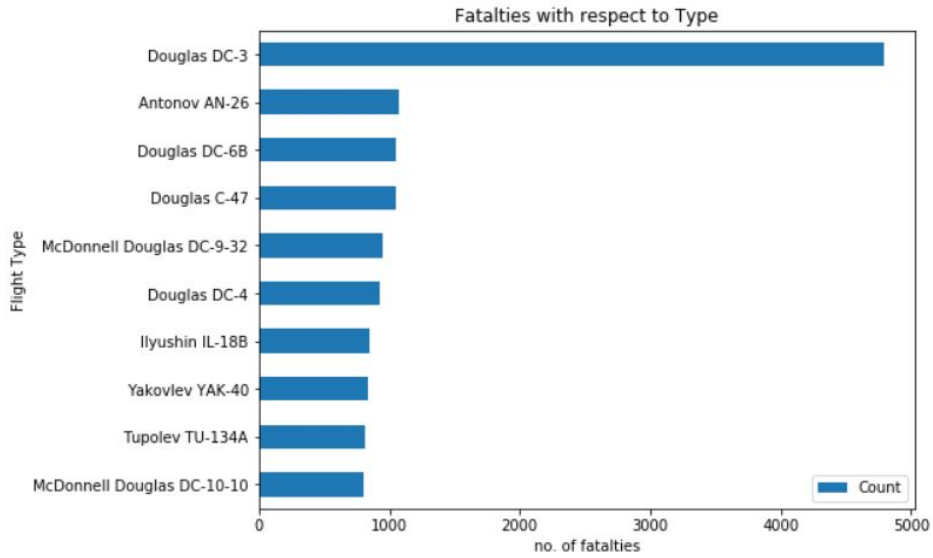
Overfitting:

Overfitting occurs when the model fits the noise, not the underlying process. Therefore when testing your model using the test-set, your model does not perform as well as it is modelling noise, not the underlying process that generated the relationship. Let's create a degree 5 polynomial model.



Fatalities v Flight Type:

Plotted a graph between aggregate of Fatalities and Flight type between 1908 and 2009.
It is observed that the most number of fatalities occurred with the Flight type *'Douglas DC-3'*



Results

By working on the dataset which consists of the details about airplane crashes from the year 1908 to 2009, it is observed that Multiple Linear Regression method works well with this predicting and working with minimum error difference.

Discussion

After analysing the data, I observed that most of the airplane crashes occured between the 1960s and 1980s with the Flight type *'Douglas DC-3'*.

Conclusion

With this, I would like to conclude that Polynomial Linear Regression is one of the effective ways in evaluating the model.