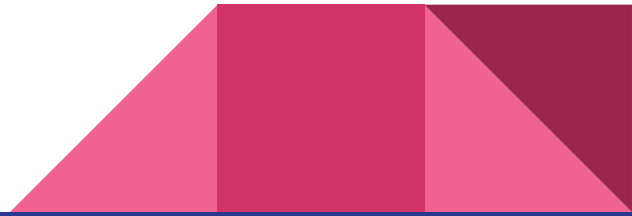
Statistics of Airplane Crashes

Statistics of airplane crashes is valuable for the airlines

By understanding the reasons behind the air plane crashes, it will be really helpful for the airlines to reduce the crashes and improve safety standards



Metadata of the dataset

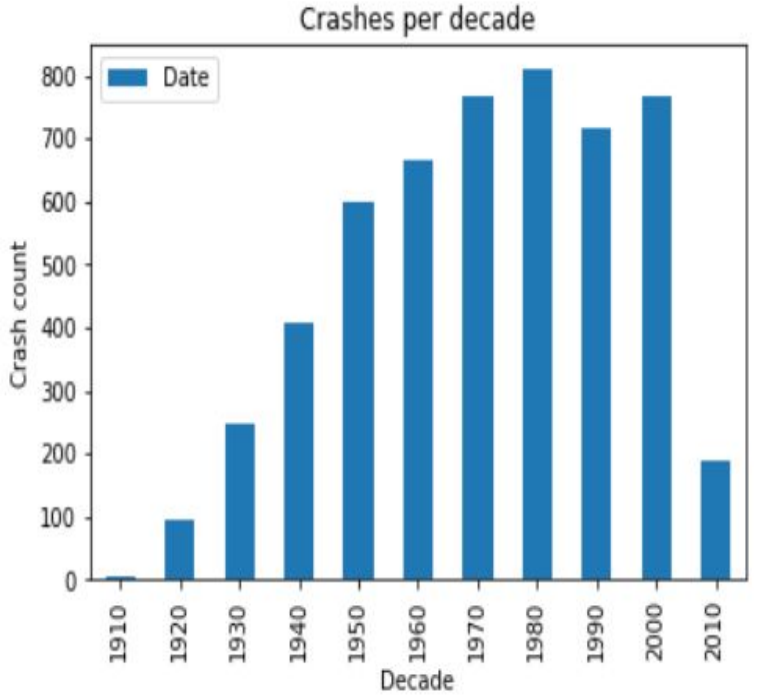
Here is the metadata of the dataset used to

Understand the reasons behind aircrashes.

```
<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
```

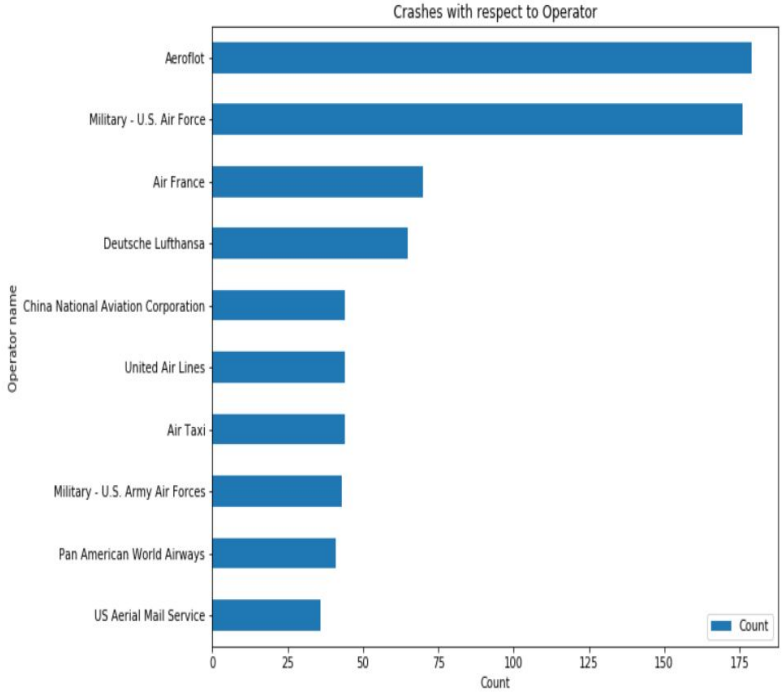
Crashes per Decade

- The picture here shows the number of fatalities occurred during the decades 1910s to 2010s.
- It is observed that most of the fatalities happened during 1980s.



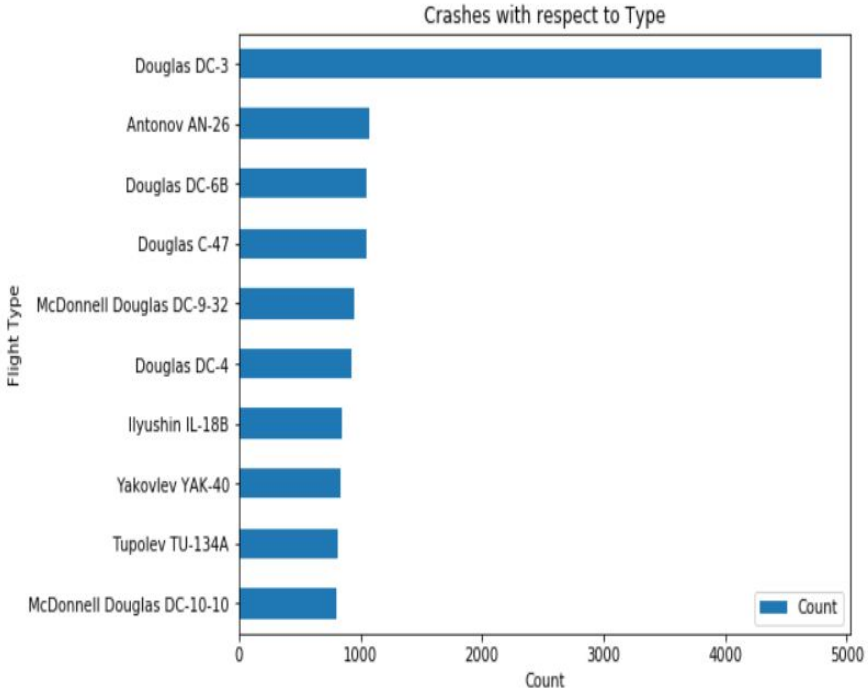
Crash v Operator

From the image attached here, it is observed that most of the fatalities happened with the operators Aeroflot and Military - U.S Air Force with around 175 crashes each



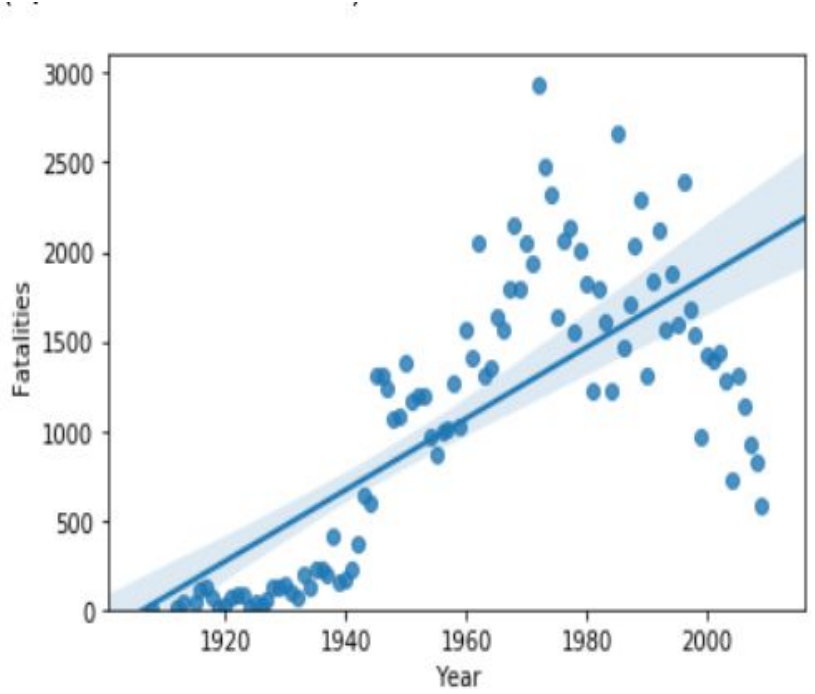
Crash v Flight Type

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



Regression Plot

Here is the regression plot between no. of fatalities and the years.



P Value

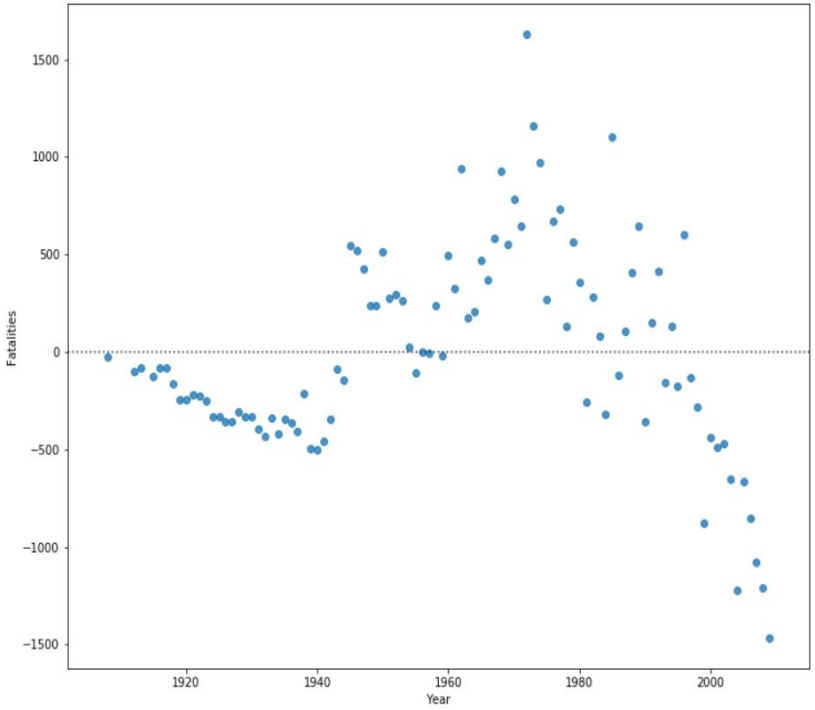
```
pearson_coef, p_value = stats.pearsonr(Year_Fatalities_dataset['Year'], Year_Fatalities_dataset['Fatalities'])
print("The Pearson Correlation Coefficient is", pearson_coef, " with a P-value of P =", p_value)
```

```
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

Here is the residual Plot between the no. of years and the Fatalities. It is observed that there is a sudden hype in the 1970s and reduced drastically.



Model Evaluation

By working on 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