AIRPLANE CRASH DATA REPORT

Statistics for Data Science (UE19CS203)

Name	Susheen Kanungo	Shweta Patki	Sravya Yepuri	Sri Ramya Priya Vedula
SRN	PES1UG19CS527	PES1UG19CS477	PES1UG19CS502	PES1UG19CS504
Section	H	H	H	H

Abstract

This report documents and presents a statistical analysis of Airplane crashes since 1908. Although the chances of dying on a commercial airline flight are as low as 9 million to 1, a lot could go wrong at 33,000 feet above the ground, and the difference between survival and death could come down to the choices the passenger/crew makes.
This dissertation aims to better understand the Airplane Crashes Dataset and then hypothesize on the basis of this extended understanding.

Introduction

Through our analyses, some of the questions we wanted to take a look at and answer were:

- Where in the world most airline crashes occur?
- Which is the safest/most dangerous airline to travel in?
- Which year/decade had the highest/lowest number of air crashes?
- Which time is the safest to fly?
- Interesting information on the number of fatalities and survivors.

Dataset Description

This is an analysis of the Airplane Crash dataset, obtained from Kaggle. It is a dataset on aviation accidents throughout the world, from 1908 to 2019. This data has been scrapped from planecrashinfo.com.

The dataset column descriptions:

- Date: Date the crash had taken place (format- mm/dd/yyyy)
- Time: The time the crash had taken place (24-hour format)
- Location: Location of the crash
- Operator: It is the operator or airline of the aircraft
- Flight #: Flight number assigned by the aircraft operator
- Route: The complete/partial route flown prior to the accident
- AC Type: Aircraft Type
- Registration: The International Civil Aviation Organization (ICAO) registrations of the aircraft.
- cn/ln: //Construction/line number
- Aboard: Number of people aboard (Passengers and Crew)
- Fatalities: Total fatalities (Passengers and Crew)
- Ground: Total killed on ground
- Summary: Brief summary of the case

Data Cleaning

Data cleaning for our data set was divided into 2 sub tasks:

a. Working with missing data

This can be done in a few ways, dropping the observations, imputing the missing data or replacing the missing data. In our data set, we first impute the time column (which had a high percentage of missing values) with the mode of the values. The missing values in columns Route, Location, AC Type and Summary are replaced with "Unknown". Next, the values in columns Aboard and Fatalities are replaced with mode and mean respectively. The columns for Registration, cn/ln, Ground and Flight # are irrelevant to our study and hence have been dropped. The rest are kept as is as they provide us with valuable insights.

Here is what the data looked like before and after cleaning the missing values:

```
#Calculating missing values in rows
Data.isnull().sum()
```

```
Date                      0
Time                   1510
Location                  4
Operator                 10
Flight #               3652
Route                   774
AC Type                  15
Registration            273
cn/ln                   668
Aboard                   18
Aboard Passangers       229
Aboard Crew             226
Fatalities                8
Fatalities Passangers   242
Fatalities Crew         241
Ground                   41
Summary                  64
dtype: int64
```

```
#Finding the number of missing values
#in each column after cleaning the dataset.
Data.isnull().sum()
```

```
Decade          0
Year            0
Date            0
Time            0
Location        0
Operator        0
Op_Code         0
Route           0
AC Type         0
Make_Type       0
Aboard          0
Fatalities      0
Accident_Code   0
Summary         0
dtype: int64
```

b. Working with inconsistent data

The only column in the data set with inconsistent formatting was Date which has been converted into a single format as mm/dd/yyyy

Date before cleaning:

```
0    09/17/1908
1    09-07-1909
2    07-12-1912
3    08-06-1913
4    09-09-1913
Name: Date, dtype: object
```

After cleaning:

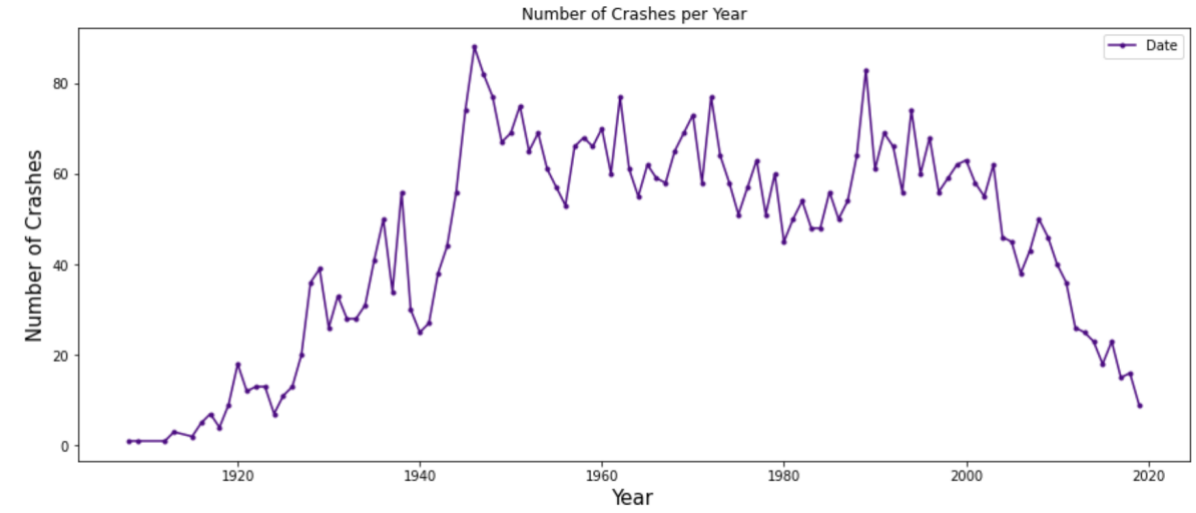
```
0       09/17/1908
1       07/09/1909
2       12/07/1912
3       06/08/1913
4       09/09/1913
           ...
4962    04/16/2019
4963    05/05/2019
4964    05/05/2019
4965    03/06/2019
4966    07/30/2019
```

Exploratory Data Analysis

a) **Graphical Analysis**

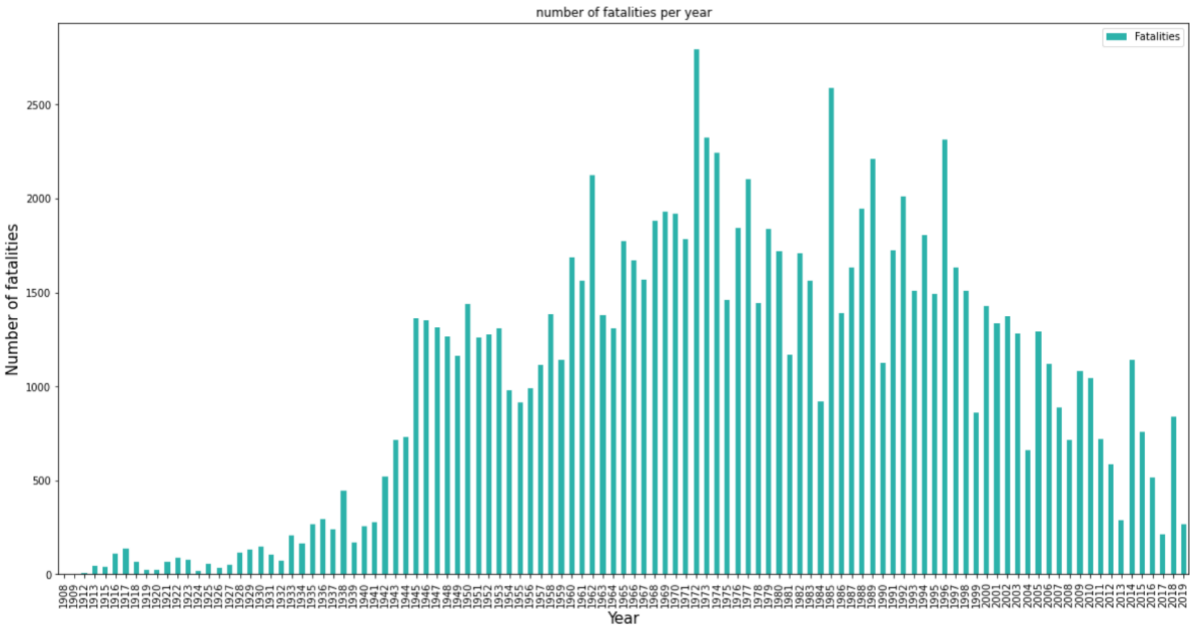
While researching, we learnt that the two parameters used in any graph should be selected in such a way that the graph yields meaningful information and the interpreter can take meaningful action based on the plot. With this idea in mind, we decided to plot graphs for the following:

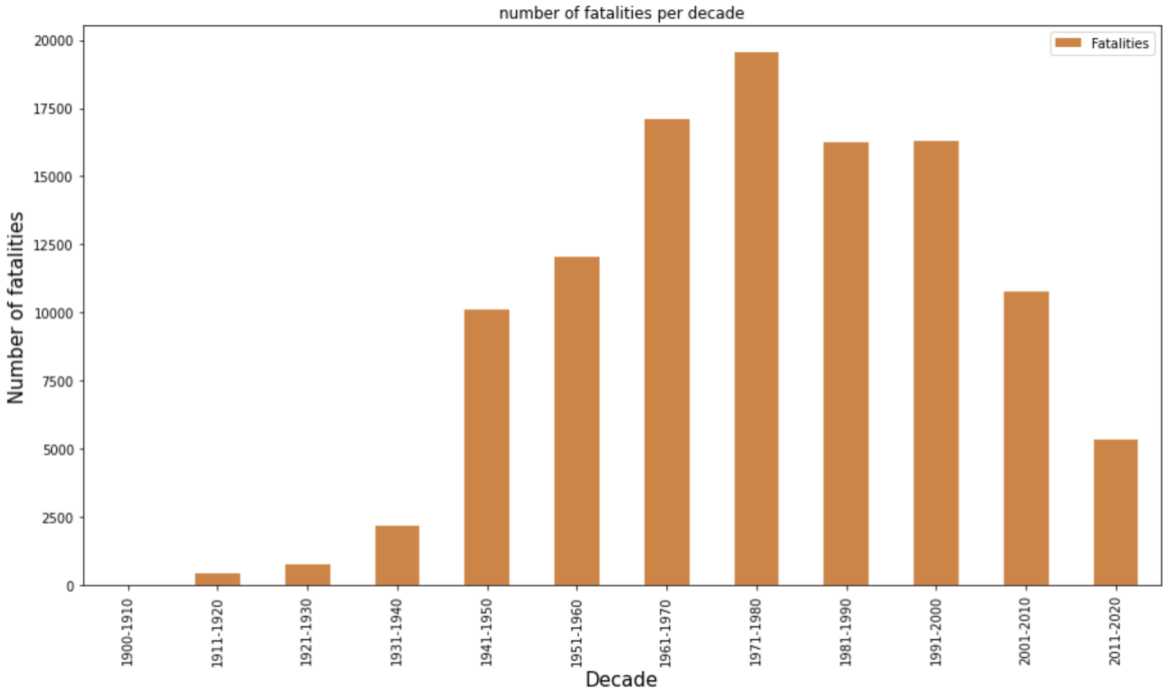
- **Number of crashes per year**



It started with one air crash in 1908 and the number kept increasing thereafter. There is a steep rise in the number of crashes between 1940 and 1946 with the year 1946 witnessing the highest number of crashes. As the data includes military planes also, this huge number of crashes during the period ranging from 1940 to 1946 can be attributed to the second world war. This fact can also be validated by the steep decrease in the number of crashes after 1946. Subsequently the number kept increasing and decreasing between 1946 and 1982 with another peak in the year 1990. However, after 1990 the number more or less kept decreasing with the least number of crashes in the recent past being in the year 2019. This steep decrease after 1990 can be attributed to technological advancement.

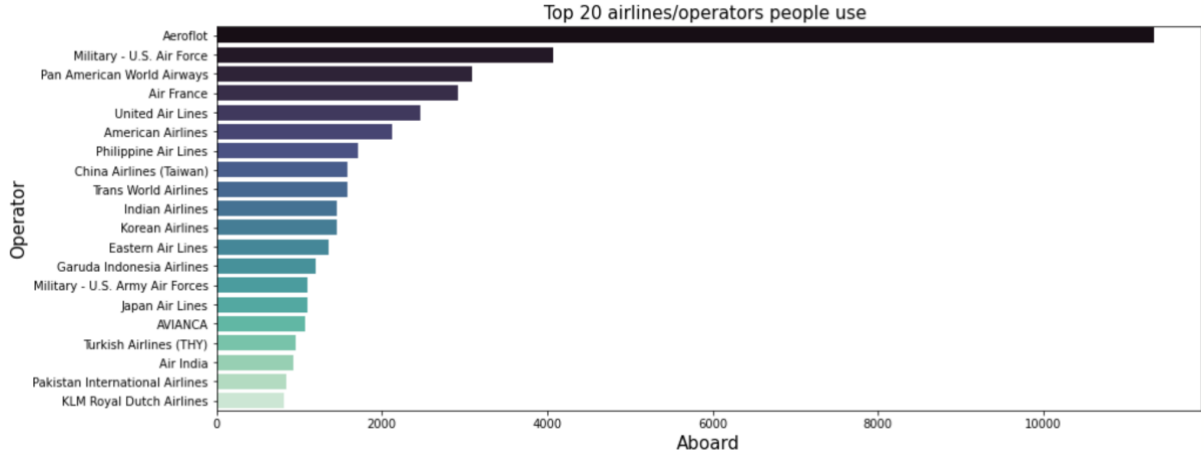
- **Number of fatalities per year and decade**

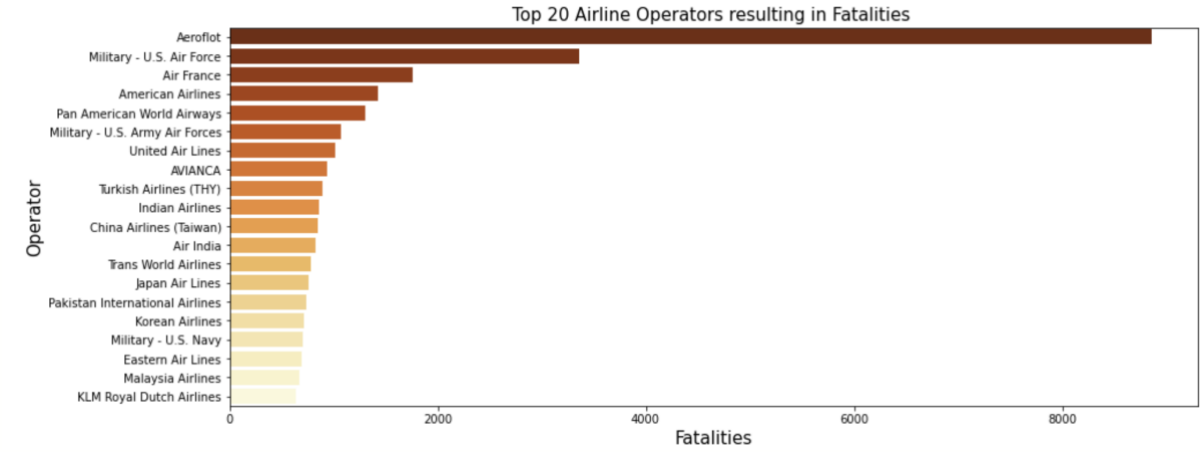




A graph has been plotted between the number of fatalities and the year. However, as it was difficult to interpret the same, another graph has been plotted between the number of fatalities decade wise. This graph gives meaningful insights and we can clearly identify that the decade 1971 to 1990 saw the maximum number of fatalities. The decades 1980-1990 and 1990-2000 had the same amount of fatalities and thereafter the numbers started decreasing.

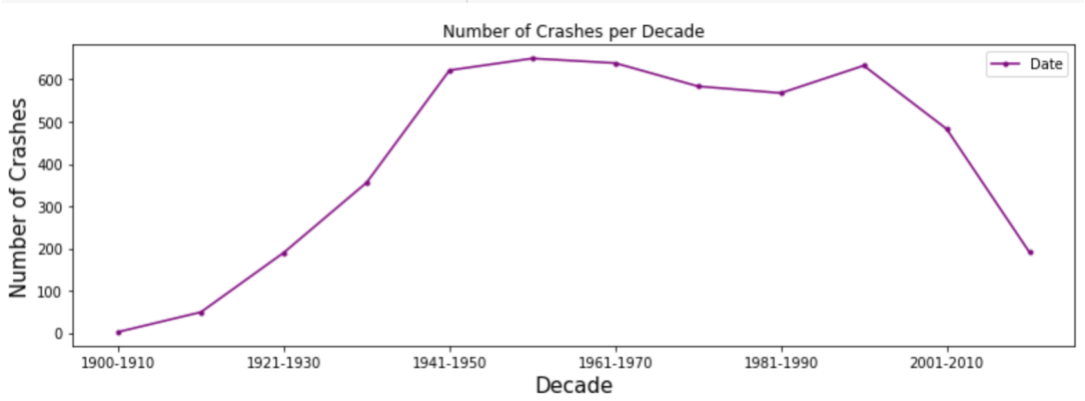
- **The top 20 Airlines people use and the top 20 Airlines resulting in fatalities**

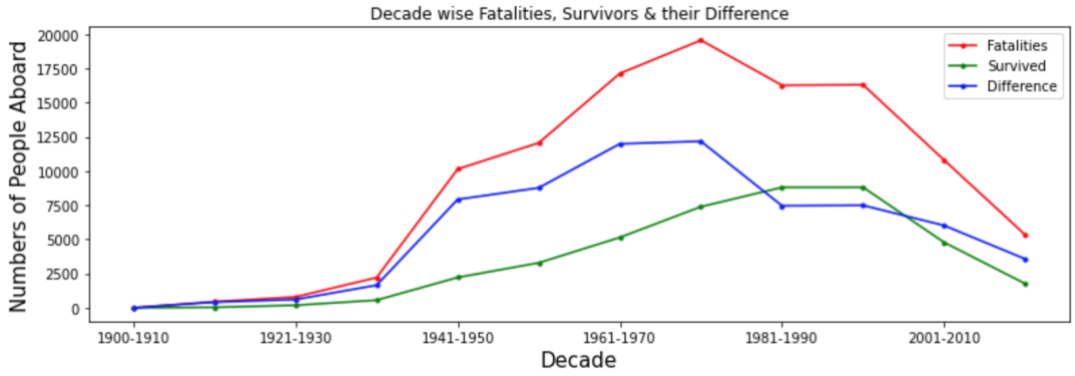




Aeroflot is airlines most people travel in. It also results in the greatest number of fatalities. Comparing the above two graphs, Pan American World Airways is third in the list of preferred airways but fifth in the list of fatalities. United Airlines is fifth in the list of preferred airways but seventh in the list of fatalities. This means that they are safer compared to other airlines.
Air France and American airlines have moved up in the list of fatalities as compared to the list of preferred airways. This indicates that their safety record is not that good.
Apart from Aeroflot, one of the worst airlines to travel in is AVIANCA as it is the top 16th airlines people prefer to travel in but the number of fatalities, it is ranking 8th.

- **Number of Crashes per decade and decade wise fatalities, survivals and their difference**

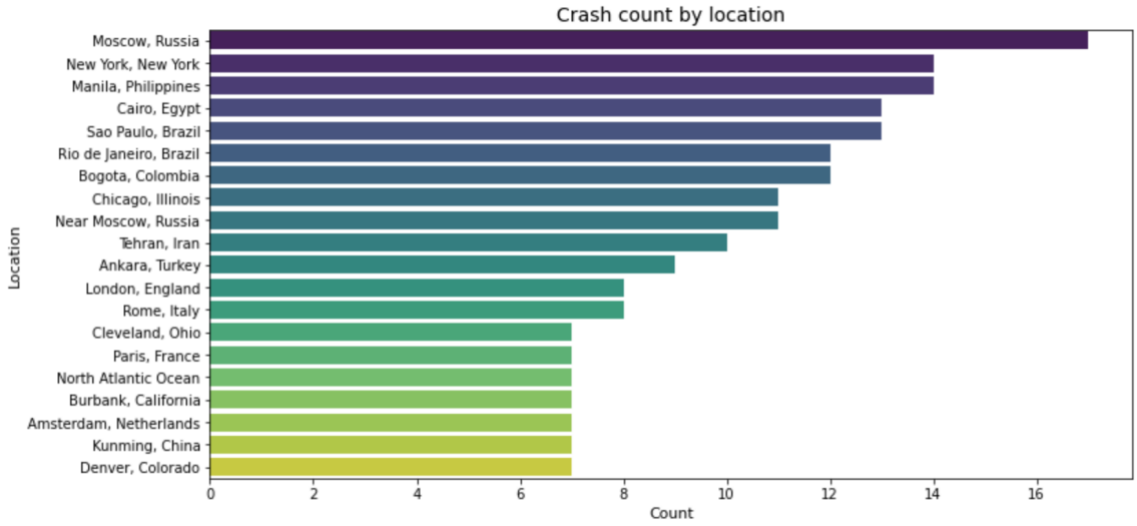




The rate of Fatalities kept increasing until 1970 and then it kept decreasing till 1980. For the next 10 years, it was constant. After that there has been a sharp decline till 2019 in spite of huge increase in the number of flights and number of people aboard.
The rate of Survivors kept increasing until 1980 and it was constant for the next 10 years until 1990 and then there was a decrease in the number of survivors as the number of crashes also decreased.
As just the numbers of fatalities and survivors do not give the correct picture, their difference, which reveals the true information, has also been plotted. The difference between the fatalities and survivors indicates that, till 1980, the rate of increase of fatalities was more than the rate of increase of survivors resulting in increase in the rate of the difference. From the year 1980 to the year 1990, This rate was constant. There after the rate kept decreasing till 2019.The decrease of the rate of differences from 1980 till 2019 is a positive sign and indicates the technological advancement in this field.

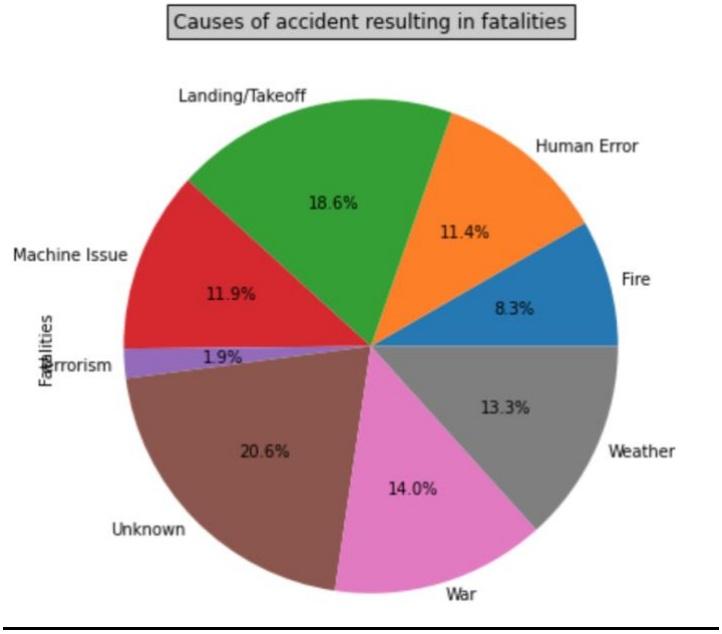
- **Top 20 places where crashes take place**

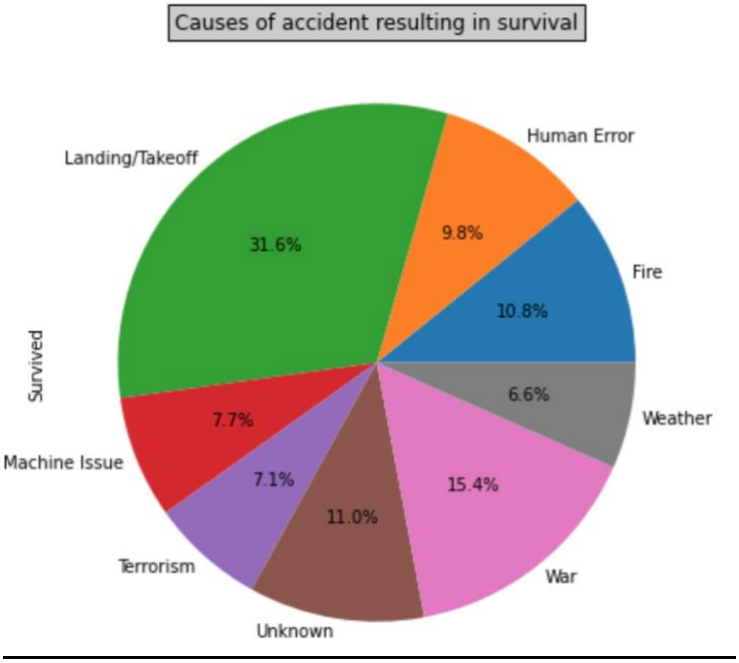
	No. of crashes
Moscow, Russia	17
New York, New York	14
Manila, Philippines	14
Cairo, Egypt	13
Sao Paulo, Brazil	13
Rio de Janeiro, Brazil	12
Bogota, Colombia	12
Chicago, Illinois	11
Near Moscow, Russia	11
Tehran, Iran	10
Ankara, Turkey	9
London, England	8
Rome, Italy	8
Cleveland, Ohio	7
Paris, France	7
North Atlantic Ocean	7
Burbank, California	7
Amsterdam, Netherlands	7
Kunming, China	7
Denver, Colorado	7



Most of the crashes took place in Moscow, Russia. The locations where crashes took place are spread over the entire world From Russia in the East to the Western states of the United States of America and from Atlantic Ocean in the North to Brazil in the South. Hence this graph indicates that no meaningful inferences can be made about particular locations being more vulnerable to crashes.

- **Causes of accidents resulting in fatalities & survival**



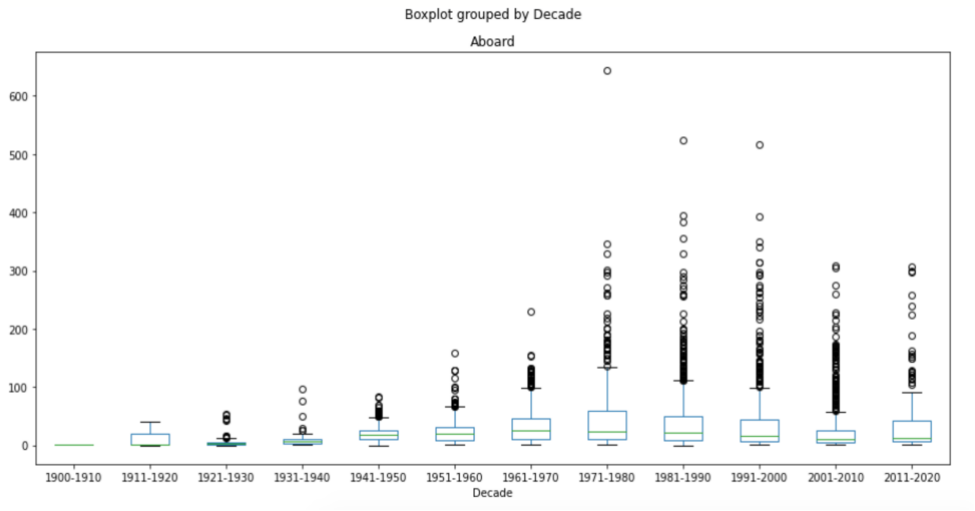


Leaving the unknown factors, we can say that the number of fatalities occurring due to landing/take-off are more than the crashes due to any other causes. The next cause of fatalities is war. If we ignore that, the next cause is weather. Ignoring war, the main causes of fatalities are landing/ take off and weather. If we take care of these two factors, about 32% of the facilities can be avoided. Comparing the above two graphs, we can also find out that though landing/ take off result in 18.6% fatalities, they also result in 31.6% of the survivals. This means that, during the crashes because of landing/take off, many people survive as compared to the other causes when the survivors are comparatively less. we can conclude that, the number of fatalities occurring due to landing/take-off is less, as its percentage is comparatively higher in survival than fatalities. Whereas, it is the opposite in case of weather and machine issue.

Filtering unwanted outliers

A. Decade wise passengers aboard using box plot:

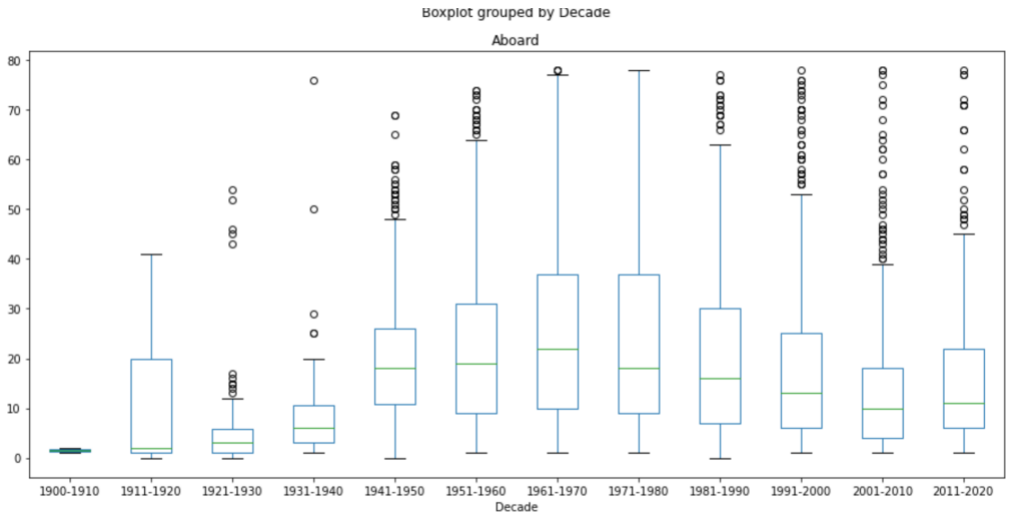
Before unwanted outliers' removal:



From the above boxplot we understand that there is a lot of variation in the number of passengers aboard the crashed flights during various decades.

After unwanted outliers' removal:

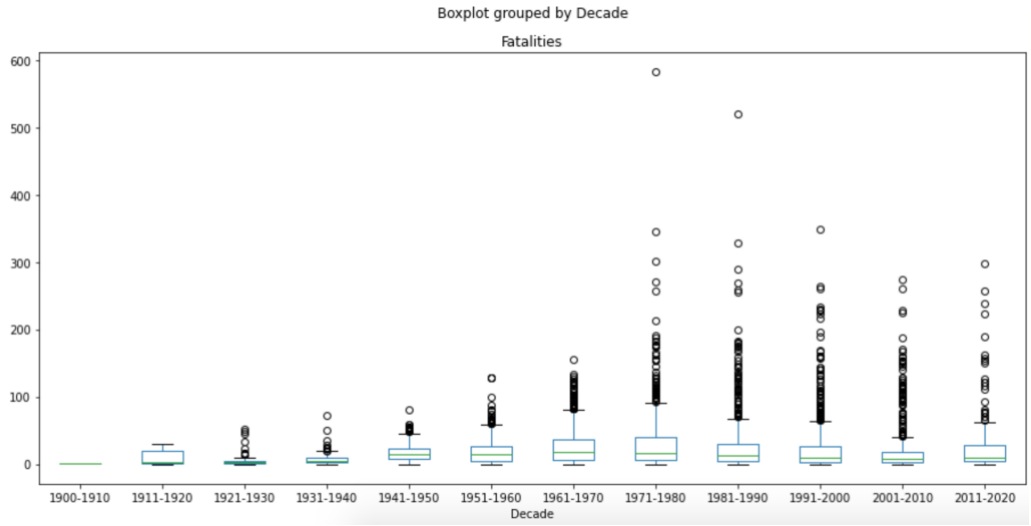
```
count    4967.000000
mean       30.986511
std        45.387259
min         0.000000
25%         6.000000
50%        16.000000
75%        35.000000
max       644.000000
Name: Aboard, dtype: float64

interquartile range: 29.0
upper_fence: 78.5
lower_fence: -37.5
percentage of Aboard out of lower fence: 0.00 %
percentage of Aboard out of upper fence: 9.74 %
length of input dataframe: 4967
length of new dataframe after Unwanted Outlier removal: 4483
```



For unwanted outliers we have fixed a limit of 1.5 IQR above the upper whisker and below the lower whisker. This limit is arbitrary and is fixed based on the data. This resulted in 9.74% unwanted outliers. The original data consisted of 4967 values and after removal of outliers as per the above criteria, it decreased to 4483.

B. Decade wise Fatalities using Box Plot:

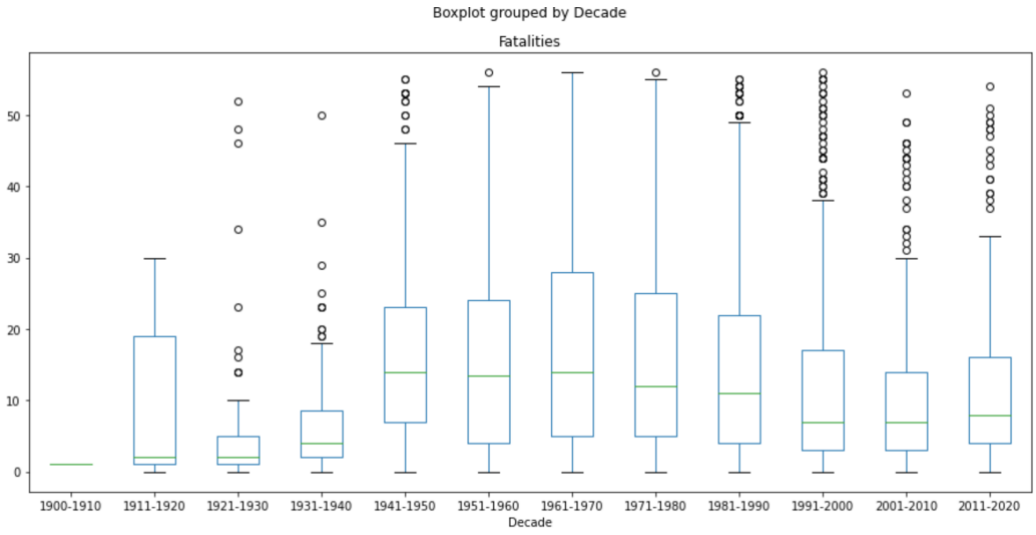
Before unwanted outliers' removal:



From the above boxplot we understand that there is a lot of variation in the number of fatalities during various decades.

After unwanted outliers' removal:

```
count    4967.000000
mean       22.339239
std        34.997961
min         0.000000
25%         4.000000
50%        11.000000
75%        25.000000
max       583.000000
Name: Fatalities, dtype: float64

interquartile range: 21.0
upper_fence: 56.5
lower_fence: -27.5
percentage of Fatalities out of lower fence: 0.00 %
percentage of Fatalities out of upper fence: 9.06 %
length of input dataframe: 4967
length of new dataframe after Unwanted Outlier removal: 4517
```



For unwanted outliers we have fixed a limit of 1.5 IQR above the upper whisker and below the lower whisker. This limit is arbitrary and is fixed based on the data. This resulted in 9.06% unwanted outliers. The original data consisted of 4967 values and after removal of outliers as per the above criteria, it decreased to 4517.

C. Normalization

Normalization, also known as feature scaling, is a method to make values of different scales easier to compare and to visualize. There are many methods of feature scaling, some of which are:

- rescaling, also known as min-max normalisation
- mean normalisation
- standardization or z-score normalisation

- scaling to unit length.

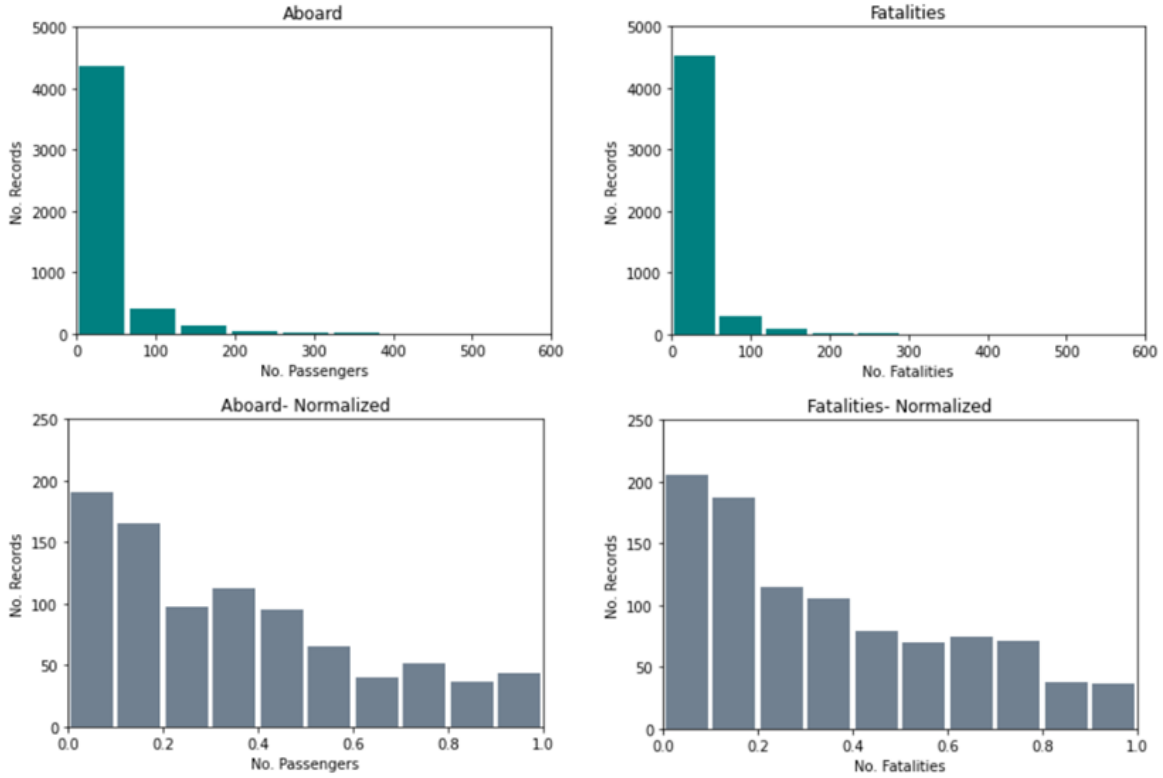
The method used throughout this section is standardisation, which involves scaling the values such that the mean is 0 and the variance is 1.

$$x' = x - x$$

Looking into the columns for Aboard and Fatalities in the dataset, we notice that the values range widely. The first bin in the graph, which includes the limit of 1 to 60 passengers, has the highest number of entries by a large margin, and thus overwhelms the graph, rendering the bins for . After normalisation by the above formula, we get the following summary:

```
Mean of total passengers:  31
Variance of total passengers:  2060
Mean of normalized passengers:  0
Variance of normalized passengers:  1

Mean of total deaths:  22
Variance of total deaths:  1225
Mean of normalized deaths:  0
Variance of normalized deaths:  1
```

Graphical analysis makes the effect of normalisation more obvious; as visible ahead:



While the number of passengers on a flight is maximum in the bin 1 to 60, it is large to the extent of hiding the other bins. Normalization changes the values such that they lie between 0 to 1. Now, although the first bin is still the largest, it does not skew the graph to extremes.

Hypothesis Testing

The purpose of hypothesis testing is to determine whether there is enough statistical evidence in favour of a certain belief, or hypothesis, about a parameter. Here, we assume different hypotheses for some of the columns of our dataset. Since our dataset is a large dataset, we use z-statistic for our hypothesis's tests below. We have used population proportion and p-value approach for all of our hypotheses test calculations below.
To calculate the z-statistic, we use the following formula:

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

We also check if the conditions $n.p_0 > 10$ and $n. (1 - p_0) > 10$ are satisfied as we are using proportion for our hypotheses.
We know that the number of values in each column of our dataset are 4967, so n = 4967.

- **Time:**
  We assume that the null hypothesis (H_0), more than or equal to 75% of the aircrafts crash before 12PM, is true.
  So, our alternate hypothesis (H_1) will be less than 75% of the aircrafts crash before 12PM.
  Null Hypothesis: $p \geq p_0 \rightarrow p \geq 0.75$
  Alternate Hypothesis: $p > p_0 \rightarrow p < 0.75$
  The conditions $n.p_0 > 10$ and $n. (1 - p_0) > 10$ are satisfied from the information we know.
  This is a one tailed test and it is left tailed. So, we take the standard significance level, $\alpha = 0.05$.
  Value of p when hypothesis is true is $p_0 = 0.75$
  We have the null distribution as $p \sim N (p_0, p_0.(1 - p_0))$.
  Number of successes, X = 1483
  p= X/n = 0.2985705657.
  After the calculations using the formulae we know, we get z = -73.4744454073 and p = 0.0000000000.
  We know that $\alpha = 0.05$, and since value $< \alpha$, our null hypothesis is rejected and the alternate hypothesis is accepted.
  Therefore, less than 75% of the aircrafts crash before 12PM.

- **Fatalities:**
  We take the null hypothesis (H_0) as more than or equal to 75% of the aircrafts fatalities are greater than 50% of the maximum fatalities.
  Then our alternate hypothesis (H_1) will become less than 75% of the aircraft's fatalities are greater than 50% of the maximum fatalities.
  We assume that the null hypothesis is true.
  Null Hypothesis: $p \geq p_0 \rightarrow p \geq 0.75$
  Alternate Hypothesis: $p > p_0 \rightarrow p < 0.75$
  The conditions $n.p_0 > 10$ and $n. (1 - p_0) > 10$ are satisfied from the information we know.
  This hypothesis test also is a one tailed test and it is left tailed. Hence, we take the standard significance level, $\alpha = 0.05$.
  So, the value of p when the hypothesis is true is $p_0 = 0.75$.

Here, the number of successes, X = 4960.
We know that the null distribution is $p \sim N(p_0, p_0.(1 - p_0))$.
$p = X/n = 0.9985906986$
After the calculations, we get z = 40.4605068421 and respectively, p_value as p = 1.0000000000.
We know that $\alpha = 0.05$, and since p_value > α, we fail to reject the null hypothesis. So, the null hypothesis will become plausible, and so the alternate hypothesis will also become plausible.

- **Operators:**
  We assume that the null hypothesis (H_0), less than or equal to 50% of the aircrafts are of military or passenger operated, is true.
  So, our alternate hypothesis (H_1) will be that more than 50% of the aircrafts are military or passenger operated.
  So, null Hypothesis: $p \leq p_0 \rightarrow p \leq 0.50$
  Alternate Hypothesis: $p > p_0 \rightarrow p > 0.50$
  The conditions $n.p_0 > 10$ and $n.(1 - p_0) > 10$ are satisfied from the information we know.
  This is a one tailed test and it is left tailed. So, we take the standard significance level, α as 0.05.
  Value of p when hypothesis is true is $p_0 = 0.50$
  We have the null distribution as $p \sim N(p_0, p_0.(1 - p_0))$.
  Number of successes, X = 4880
  $p = X/n = 0.9824844$.
  After the calculations using the formulae we know, we get z = 68.0080541542 and p = 0.0000000000.
  We know that $\alpha = 0.05$, and since p_value < α , our null hypothesis is rejected and the alternate hypothesis is accepted.
  Therefore, more than 50% of the aircrafts are military or passenger operated.

Correlation

Correlation is used to test relationships between quantitative variables or categorical variables. It is a measure of how things are related.

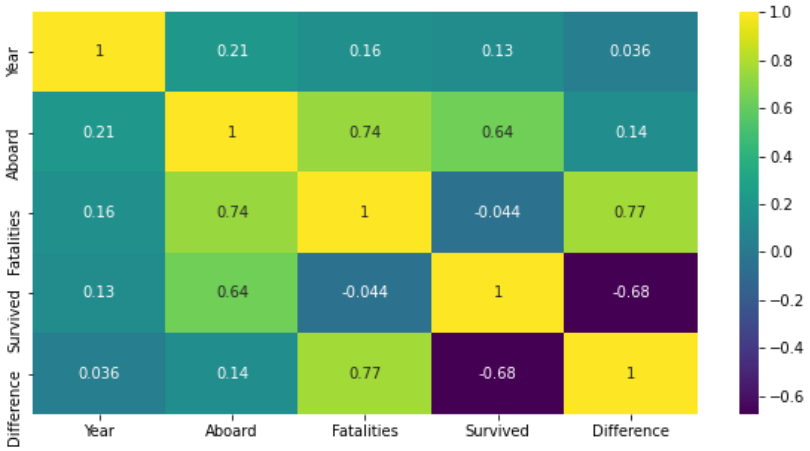
A. A correlation analysis is performed between the Passengers Aboard that have Survived and between those that have not i.e., the Fatalities.

	Aboard	Survived	Fatalities
Aboard	1.000000	0.345533	0.744119
Survived	0.345533	1.000000	-0.173281
Fatalities	0.744119	-0.173281	1.000000

There is a high correlation of 0.74 between the passengers aboard and the number of Fatalities. This is in line with the logic that if more passengers travel, the fatalities, too, will be more.

B. A heatmap for Correlation matrix is created which includes only the numerical values.

	Year	Aboard	Fatalities	Survived	Difference
Year	1.000000	0.211712	0.161567	0.129724	0.035854
Aboard	0.211712	1.000000	0.741470	0.637626	0.137351
Fatalities	0.161567	0.741470	1.000000	-0.044111	0.766469
Survived	0.129724	0.637626	-0.044111	1.000000	-0.675467
Difference	0.035854	0.137351	0.766469	-0.675467	1.000000



From this heatmap, the inferences we can draw are:
There is a high **positive** correlation between:
Passengers Abroad and Fatalities
Fatalities and Difference (Fatalities-Survived)
There is a **negative** correlation between:
Passengers Survived and Difference

Results and Conclusion

On the analysis of this data, we were able to draw some meaningful insights as shown below:

- The greatest number of crashes were in Moscow, Russia as the worst operator Aeroflot is a Russian based aircraft.
- Number of fatalities occurring due to landing and takeoff is less as its percentage is comparatively higher in survivals than deaths.
- The decrease in plane crashes from 1980 till 2019 indicates the technological advancement in aeronautics.
- 1946 had the highest number of plane crashes; however, the highest fatalities were in 1972 as passenger planes were not widely used before the 1960's.
- Less than 75% of the aircrafts crash before 12PM.
- More than 50% of the aircrafts are military or passenger operated.