Introduction :

Insight logo

This dataset provides details on airplane incidents globally. The data covers the period from September 1908 to August 2008. Various sources contribute information on air incidents, including country, continent, operator, fatality, aircraft type, and accident cause. Currently, there are 5268 records of air incidents in this dataset.

Importing necessary libraries :

```
In [57]:
import numpy as np
import pandas as pd
from matplotlib import pyplot as plt
import seaborn as sns
from datetime import date, timedelta, datetime
```

Exploring the Dataset :

```
In [58]:
data = pd.read_csv('Airplane_Crashes_and_Fatalities_Since_1908 .csv')
data.head()
```

Out[58]:

	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...

```
In [59]:
data.info()
```

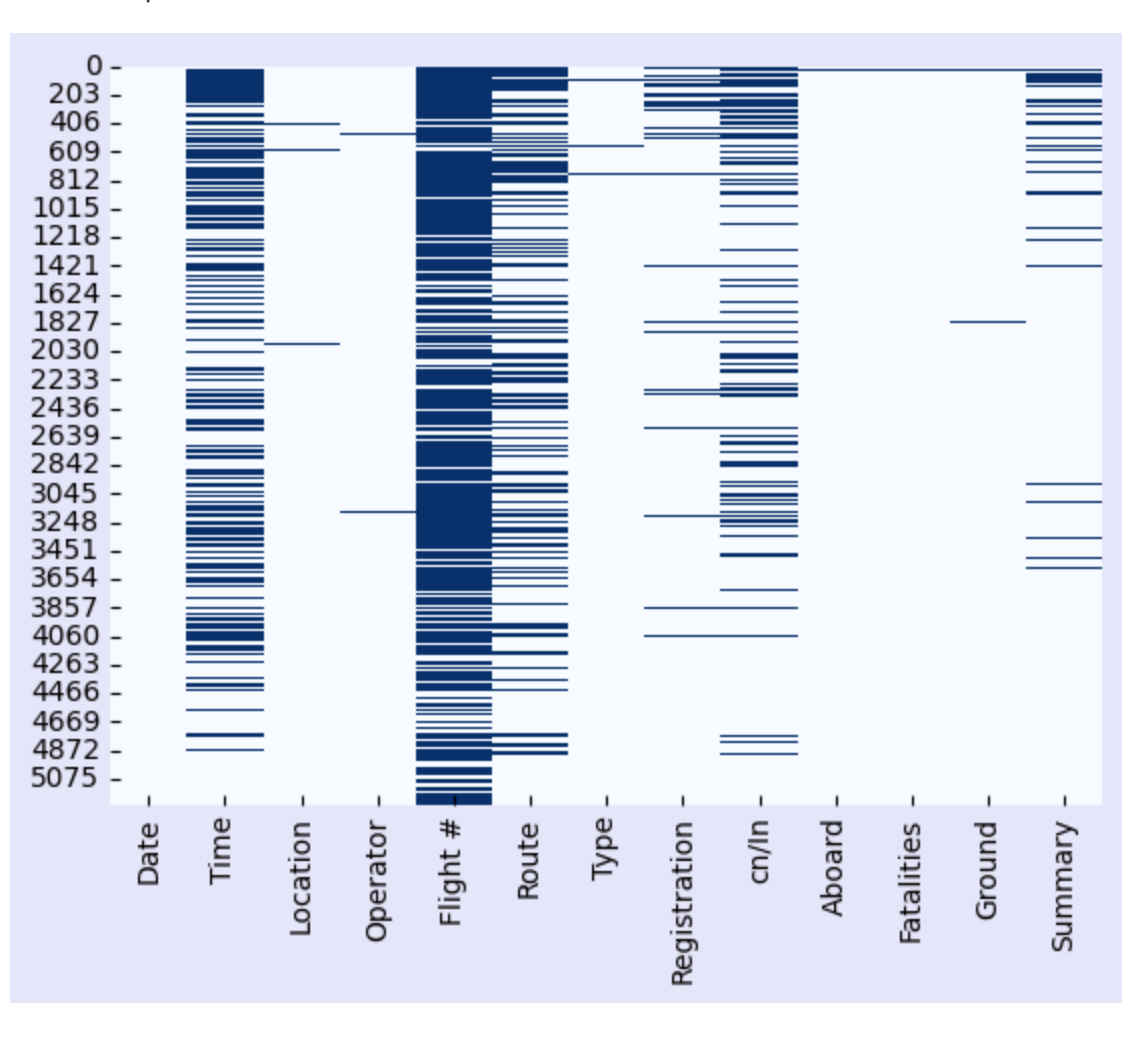
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5268 entries, 0 to 5267
Data columns (total 13 columns):
 #   Column        Non-Null Count  Dtype  
---  ------        --------------  -----  
 0   Date          5268 non-null   object 
 1   Time          3049 non-null   object 
 2   Location      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
```

Data cleaning :

Handle Missing Values :

```
In [60]:
plt.figure(facecolor='#E6E6FA')
sns.heatmap(data.isnull(), cmap='Blues', cbar=False)
```

Out[60]: <AxesSubplot: >



Remove Unnecessary Columns :

```
In [61]:
data.drop(['Flight #', 'Registration', 'cn/In', 'Time'], inplace=True, axis=1)
data.head()
```

Out[61]:

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

Fill Nan values :

```
In [62]:
data['Route'] = data['Route'].fillna('Route Not defined')
data['Operator'] = data['Operator'].fillna('Operator Unknown')
data['Summary'] = data['Summary'].fillna('No Comments Available')
data['Type'] = data['Type'].fillna('Type Not Specified')
data.head()
```

Out[62]:

	Date	Location	Operator	Route	Type	Aboard	Fatalities	Ground	Summary
0	09/17/1908	Fort Myer, Virginia	Military - U.S. Army	Demonstration	Wright Flyer III	2.0	1.0	0.0	During a demonstration flight, a U.S. Army fly...
1	07/12/1912	AtlantiCity, New Jersey	Military - U.S. Navy	Test flight	Dirigible	5.0	5.0	0.0	First U.S. dirigible Akron exploded just offsh...
2	08/06/1913	Victoria, British Columbia, Canada	Private	Route Not defined	Curtiss seaplane	1.0	1.0	0.0	The first fatal airplane accident in Canada oc...
3	09/09/1913	Over the North Sea	Military - German Navy	Route Not defined	Zeppelin L-1 (airship)	20.0	14.0	0.0	The airship flew into a thunderstorm and encou...
4	10/17/1913	Near Johannisthal, Germany	Military - German Navy	Route Not defined	Zeppelin L-2 (airship)	30.0	30.0	0.0	Hydrogen gas which was being vented was sucked...

Convert Data Types :

```
In [63]:
data['Date'] = pd.to_datetime(data['Date'])
data['Date']=pd.DatetimeIndex(data['Date'])
data_date_index=data.set_index('Date')
data.head()
```

Out[63]:

	Date	Location	Operator	Route	Type	Aboard	Fatalities	Ground	Summary
0	1908-09-17	Fort Myer, Virginia	Military - U.S. Army	Demonstration	Wright Flyer III	2.0	1.0	0.0	During a demonstration flight, a U.S. Army fly...
1	1912-07-12	AtlantiCity, New Jersey	Military - U.S. Navy	Test flight	Dirigible	5.0	5.0	0.0	First U.S. dirigible Akron exploded just offsh...
2	1913-08-06	Victoria, British Columbia, Canada	Private	Route Not defined	Curtiss seaplane	1.0	1.0	0.0	The first fatal airplane accident in Canada oc...
3	1913-09-09	Over the North Sea	Military - German Navy	Route Not defined	Zeppelin L-1 (airship)	20.0	14.0	0.0	The airship flew into a thunderstorm and encou...
4	1913-10-17	Near Johannisthal, Germany	Military - German Navy	Route Not defined	Zeppelin L-2 (airship)	30.0	30.0	0.0	Hydrogen gas which was being vented was sucked...

Feature engineering :

```
In [64]:
data['Survived'] = data['Aboard'] - data['Fatalities']
data["Survival Rate"] = 100 * (data["Survived"]) / data["Aboard"]
data["Is Military"] = data['Operator'].str.contains("Military",regex=False)
data['Location Country']  = data['Location'].str.split(',').str[-1].str.strip().str.upper()
data.head()
```

Out[64]:

	Date	Location	Operator	Route	Type	Aboard	Fatalities	Ground	Summary	Survived	Survival Rate	Is Military	Location Country
0	1908-09-17	Fort Myer, Virginia	Military - U.S. Army	Demonstration	Wright Flyer III	2.0	1.0	0.0	During a demonstration flight, a U.S. Army fly...	1.0	50.0	True	VIRGINIA
1	1912-07-12	AtlantiCity, New Jersey	Military - U.S. Navy	Test flight	Dirigible	5.0	5.0	0.0	First U.S. dirigible Akron exploded just offsh...	0.0	0.0	True	NEW JERSEY
2	1913-08-06	Victoria, British Columbia, Canada	Private	Route Not defined	Curtiss seaplane	1.0	1.0	0.0	The first fatal airplane accident in Canada oc...	0.0	0.0	False	CANADA
3	1913-09-09	Over the North Sea	Military - German Navy	Route Not defined	Zeppelin L-1 (airship)	20.0	14.0	0.0	The airship flew into a thunderstorm and encou...	6.0	30.0	True	OVER THE NORTH SEA
4	1913-10-17	Near Johannisthal, Germany	Military - German Navy	Route Not defined	Zeppelin L-2 (airship)	30.0	30.0	0.0	Hydrogen gas which was being vented was sucked...	0.0	0.0	True	GERMANY

```
In [65]:
data['Year'] = pd.DatetimeIndex(data['Date']).year
data['Month'] = pd.DatetimeIndex(data['Date']).month
data['Decade'] = (data['Year']) // 10 * 10
```

```
In [66]:
rows_with_nan = data[data['Location'].isna()]

# Print or use the resulting DataFrame
rows_with_nan
```

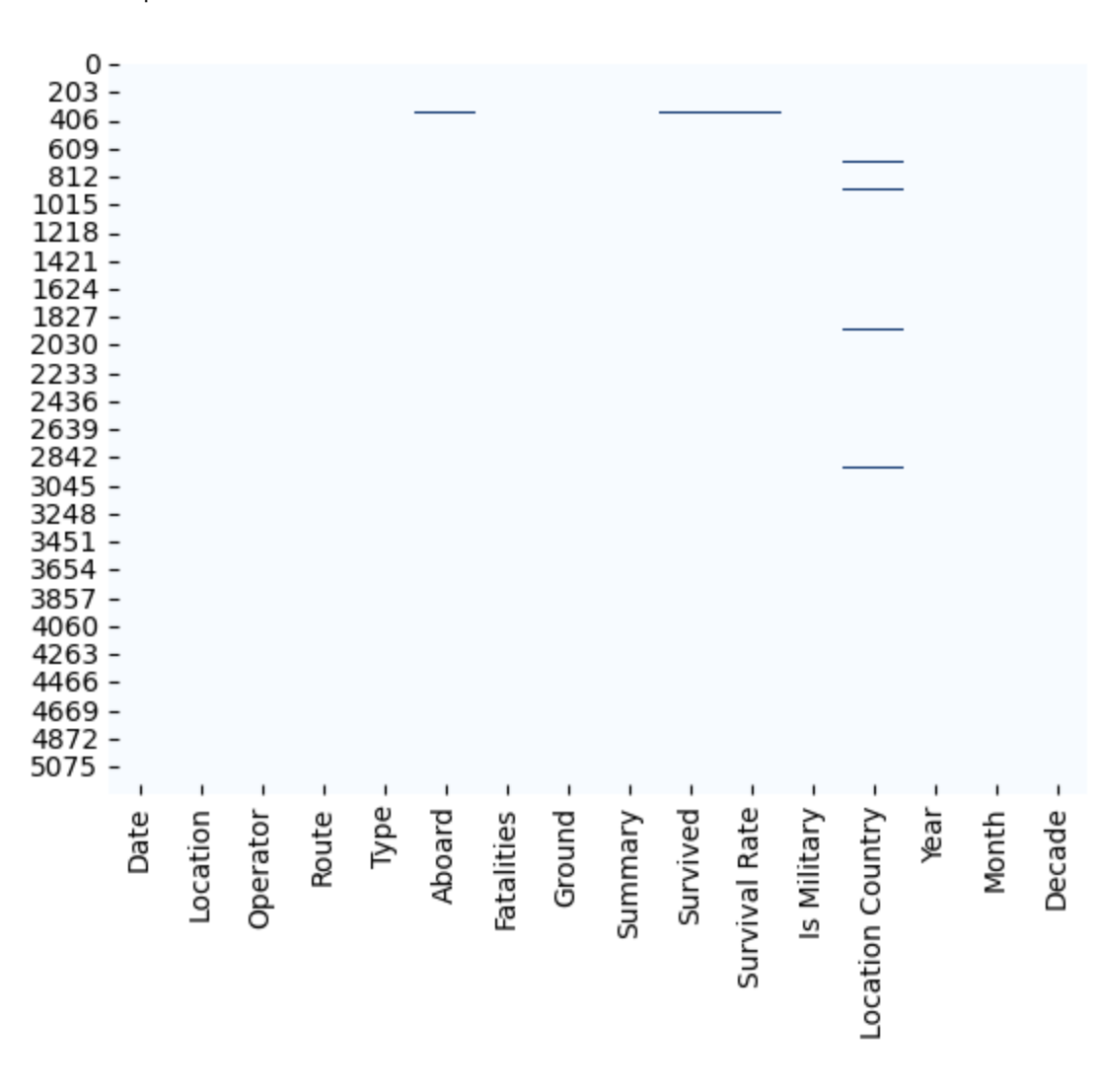
Out[66]:

	Date	Location	Operator	Route	Type	Aboard	Fatalities	Ground	Summary	Survived	Survival Rate	Is Military	Location Country	Year	Month	Decade
142	1928-05-07	NaN	Aeropostale	Route Not defined	Latecoere 26	1.0	1.0	0.0	No Comments Available	0.0	0.000000	False	NaN	1928	5	1920
411	1936-10-09	NaN	North Sea Aerial and General Transport	Route Not defined	Blackburn B-2	1.0	1.0	0.0	No Comments Available	0.0	0.000000	False	NaN	1936	10	1930
564	1941-10-28	NaN	Deutsche Lufthansa	Route Not defined	Junkers JU-53/3m	13.0	13.0	0.0	No Comments Available	0.0	0.000000	False	NaN	1941	10	1940
573	1942-02-14	NaN	China National Aviation Corporation	Route Not defined	Douglas DC-2	NaN	NaN	NaN	No Comments Available	NaN	NaN	False	NaN	1942	2	1940
588	1942-08-21	NaN	Deutsche Lufthansa	Route Not defined	Siebel Si-204	4.0	4.0	0.0	Lufthansa chairman, Von Gablenz killed.	0.0	0.000000	False	NaN	1942	8	1940
596	1942-10-22	NaN	Deutsche Lufthansa	Route Not defined	Junkers JU-52/3m	17.0	17.0	0.0	No Comments Available	0.0	0.000000	False	NaN	1942	10	1940
704	1945-04-20	NaN	Operator Unknown	Route Not defined	Junkers JU-53/3m	18.0	18.0	0.0	Missing on an evacuation flight from Berlin to...	0.0	0.000000	False	NaN	1945	4	1940
904	1947-11-27	NaN	China National Aviation Corporation	Route Not defined	Douglas DC-3	3.0	2.0	0.0	The cargo plane was shot down by communist ant...	1.0	33.333333	False	NaN	1947	11	1940
1500	1957-09-28	NaN	British European Airways	Route Not defined	de Havilland DH-114 Heron	3.0	3.0	0.0	The pilot did not appreciate that the air ambu...	0.0	0.000000	False	NaN	1957	9	1950
1916	1964-05-27	NaN	VASP	Training	Douglas C-47-DL	3.0	3.0	0.0	No Comments Available	0.0	0.000000	False	NaN	1964	5	1960
1918	1964-06-13	NaN	Saudi Arabian Airlines	Training	Douglas C-47A	2.0	2.0	0.0	Crashed into the Red Sea.	0.0	0.000000	False	NaN	1964	6	1960
1978	1965-05-05	NaN	Iberia Airlines	Madrid - Tenerife	Lockheed 1049G-55 Super Constellation	49.0	30.0	0.0	The pilot, who saw the beginning of the runway...	19.0	38.775510	False	NaN	1965	5	1960
2914	1976-08-09	NaN	Military - Spanish Air Force	Moron - Canary Islands	Douglas C-54E	33.0	11.0	0.0	Crashed 15 minutes after taking off.	22.0	66.666667	True	NaN	1976	8	1970
2953	1977-01-20	NaN	North Canada Air	Route Not defined	Bristol 170 Freighter 31M	2.0	1.0	0.0	The cargo plane stalled nearly vertical and cr...	1.0	50.000000	False	NaN	1977	1	1970
3868	1989-08-15	NaN	China Eastern Airlines	Shanghai - Hongqiao	Antonov AN-24RV	40.0	34.0	0.0	Lost power and crashed into a river shortly af...	6.0	15.000000	False	NaN	1989	8	1980
3880	1989-09-21	NaN	Military - Afghan Republican Air Force	Route Not defined	Mil Mi-8 (helicopter)	26.0	26.0	0.0	No Comments Available	0.0	0.000000	True	NaN	1989	9	1980
4034	1991-08-15	NaN	Transports Aériens de la Guinee-Bissau	Kano - Bamako	Fokker F-27 Friendship 100	3.0	3.0	0.0	On a positioning flight the plane struck trees...	0.0	0.000000	False	NaN	1991	8	1990
4043	1991-09-17	NaN	Ethiopian Airlines	Route Not defined	Lockheed L-100-30 Hercules	4.0	4.0	0.0	After experiencing a nose gear problem and att...	0.0	0.000000	False	NaN	1991	9	1990
4975	2004-03-04	NaN	Azov Avia Airlines	Ankara - Baku - Kabul	Ilyushin Il-76	7.0	3.0	0.0	Shortly after taking off and climbing to about...	4.0	57.142857	False	NaN	2004	3	2000
5039	2005-02-22	NaN	Indonesian National Police	Jayapura - Sarmi	CASA 212 Aviocar	18.0	15.0	0.0	On final approach, the aircraft crashed into t...	3.0	16.666667	False	NaN	2005	2	2000

```
In [67]:
data.at[5039,'Location']  = 'Indonesia'
data.at[4975,'Location']  = 'Azerbaijan'
data.at[4043,'Location']  = 'Ethiopia'
data.at[4034,'Location']  = 'Guinea'
data.at[3880,'Location']  = 'Afghanistan'
data.at[3868,'Location']  = 'China'
data.at[2953,'Location']  = 'Canada'
data.at[2914,'Location']  = 'Spain'
data.at[1978,'Location']  = 'Spain'
data.at[1918,'Location']  = 'Saudi Arabia'
data.at[1916,'Location']  = 'Brazil'
data.at[1500,'Location']  = 'England'
data.at[904,'Location']  = 'China'
data.at[704,'Location']  = 'Germany'
data.at[596,'Location']  = 'Germany'
data.at[588,'Location']  = 'Germany'
data.at[573,'Location']  = 'China'
data.at[564,'Location']  = 'Germany'
data.at[411,'Location']  = 'Enlgand'
data.at[142,'Location']  = 'France'
```

```
In [68]:
sns.heatmap(data.isnull(), cmap='Blues', cbar=False)
```

Out[68]: <AxesSubplot: >



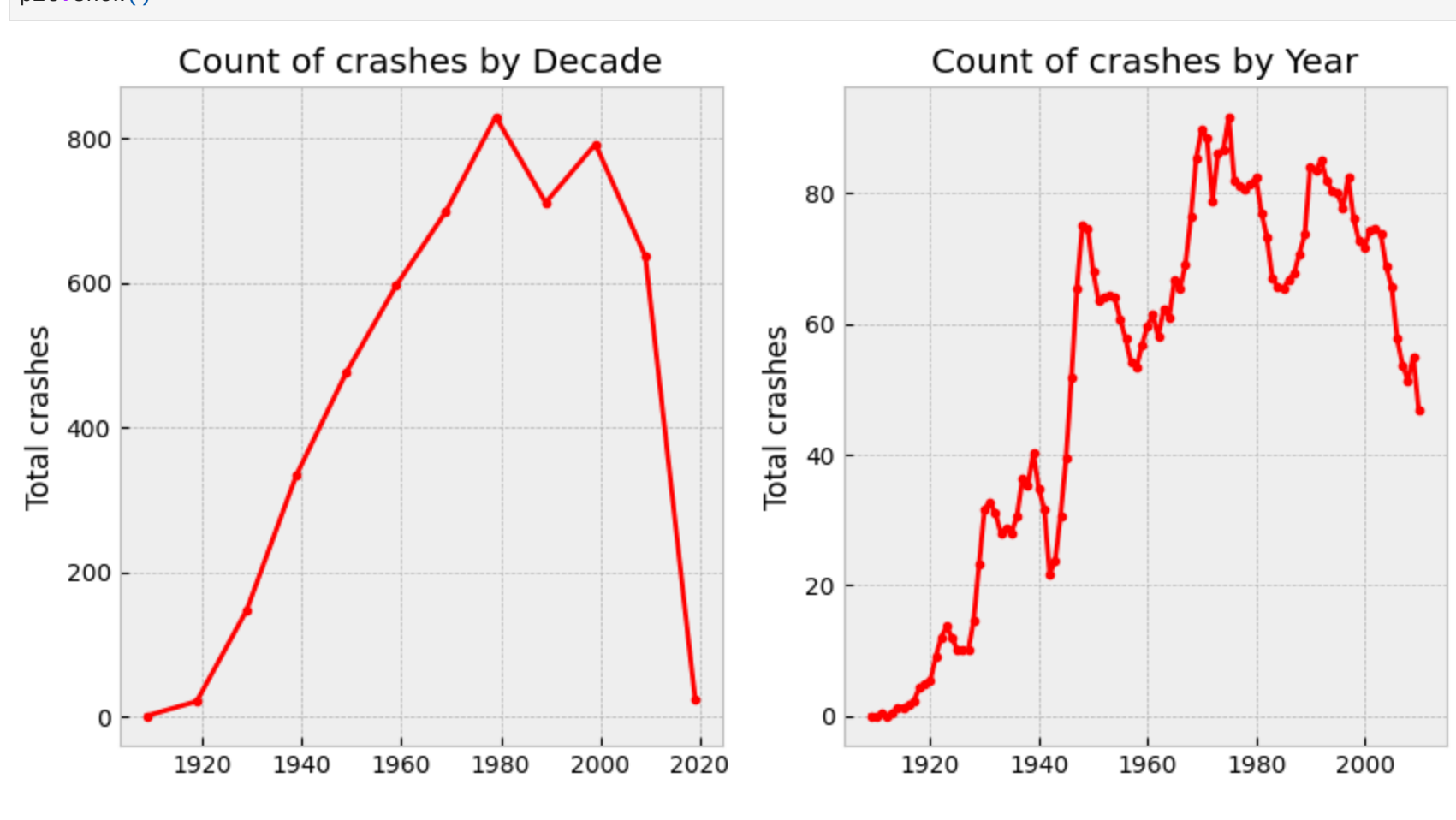
Basic problematics:

1. **Crash Statistics:**
   - What is the count of airplane crashes categorized by decade, year, and locations?
   - How does the crash frequency vary across different aircraft types?
2. **Fatalities Analysis:**
   - What is the total number of fatalities for each airline operator?
   - Can we identify trends or patterns in fatality rates among different operators?
3. **Military vs. Non-Military Comparison:**
   - How does the number of crashes compare between military and non-military airplanes?
   - Are there any discernible patterns or significant differences in crash occurrences between these two categories?
4. **Survivors and Fatalities Range:**
   - What is the range between the number of survivors and fatalities in airplane crashes?
   - Can we identify any factors or trends that correlate with higher survival rates or increased fatalities?

Data Visualisation :

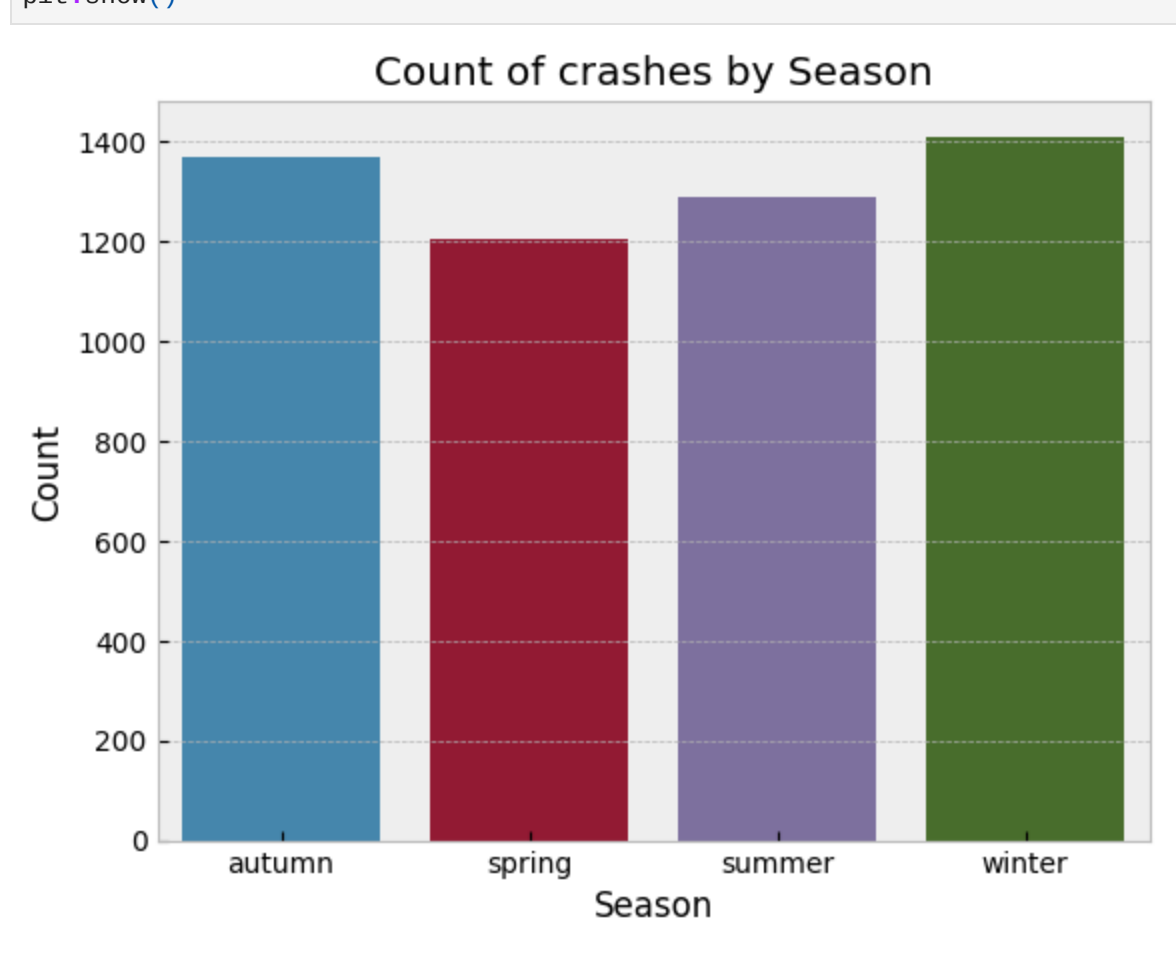
Analysing by Time :

```
In [69]:
import matplotlib.pyplot as plt

fig, axes = plt.subplots(1, 2, figsize=(10, 5))
plt.style.use('bmh')
axes[0].plot(data_date_index.resample('10y').size(),color ='red', marker=".")
axes[0].set_title('Count of crashes by Decade')
axes[0].set_ylabel('Total crashes')

crashed_by_year = data_date_index.resample('1y').size().rolling(3).mean().fillna(0)
axes[1].plot(crashed_by_year, marker=".",color = 'red')
axes[1].set_title('Count of crashes by Year')
axes[1].set_ylabel('Total crashes')

plt.show()
```



The occurrence of airplane crashes has followed an upward trajectory since the inception of mass aviation in the early 20th century, likely corresponding to the advancement of aviation technologies. The peak of these incidents was observed in the 1970s. However, since the 1970s, there has been a noticeable decline in crashes, presumably attributed to enhanced safety measures. Despite the continuous development of aviation, the frequency of accidents has significantly diminished.

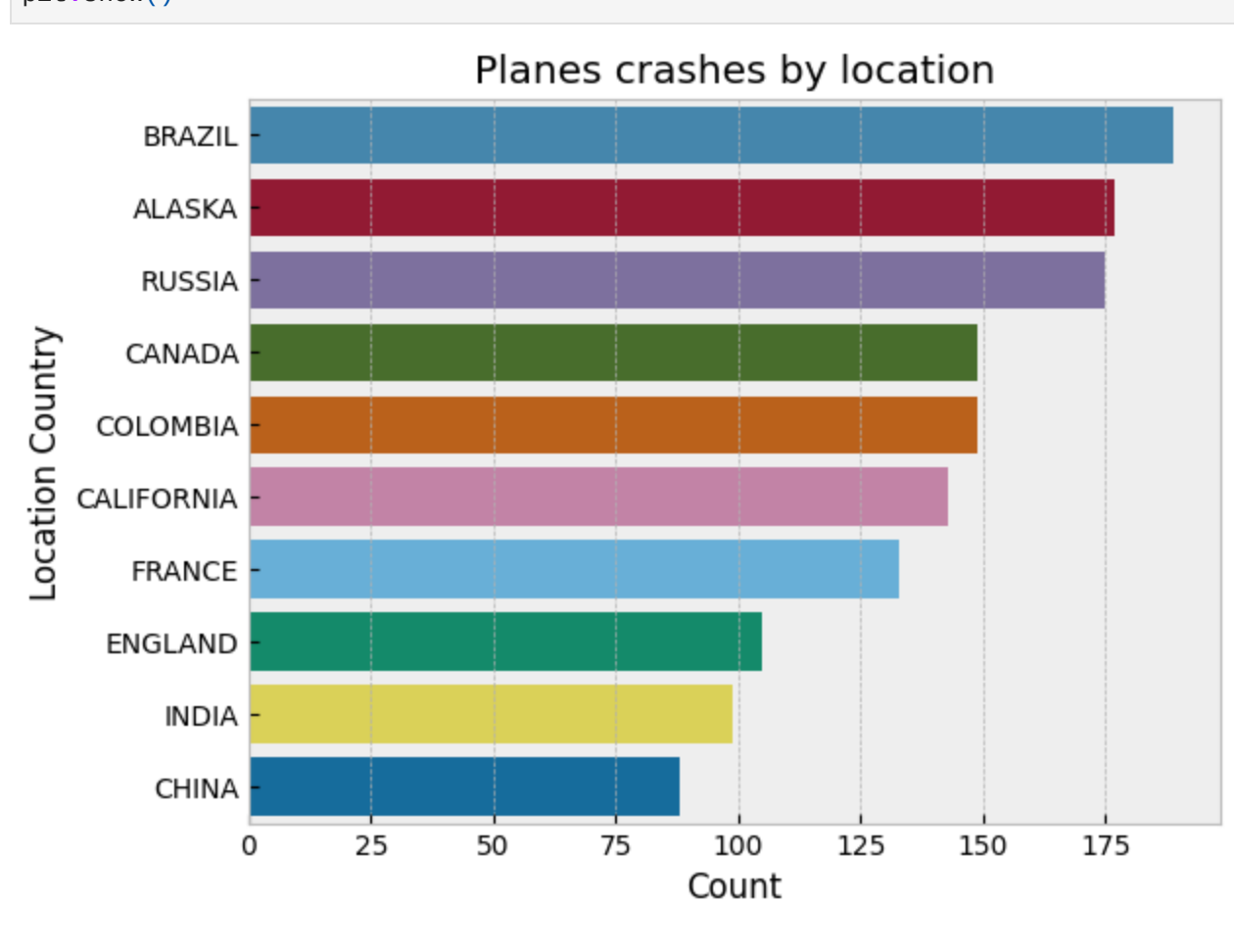
```
In [70]:
import seaborn as sns

# Define a function to assign seasons based on month
def season(month):
    if month >= 3 and month <= 5:
        return 'spring'
    elif month >= 6 and month <= 8:
        return 'summer'
    elif month >= 9 and month <= 11:
        return 'autumn'
    else:
        return 'winter'

# Apply the season function to create a new column 'Season' in the data dataframe
data['Season'] = data['Month'].apply(season)

# Group the data by season and count the number of crashes in each season
crashed_by_season = data['Season'].groupby(data['Season']).count()

# Plot a bar chart to visualize the count of crashes by season
sns.barplot(x=crashed_by_season.index, y=crashed_by_season.values)
plt.title('Count of crashes by Season')
plt.xlabel('Season')
plt.ylabel('Count')

# Display the bar chart
plt.show()
```



While there isn't a substantial disparity in the overall number of plane crashes across seasons, it is discernible that slightly more incidents occur during autumn and winter.

Analysing by Location :

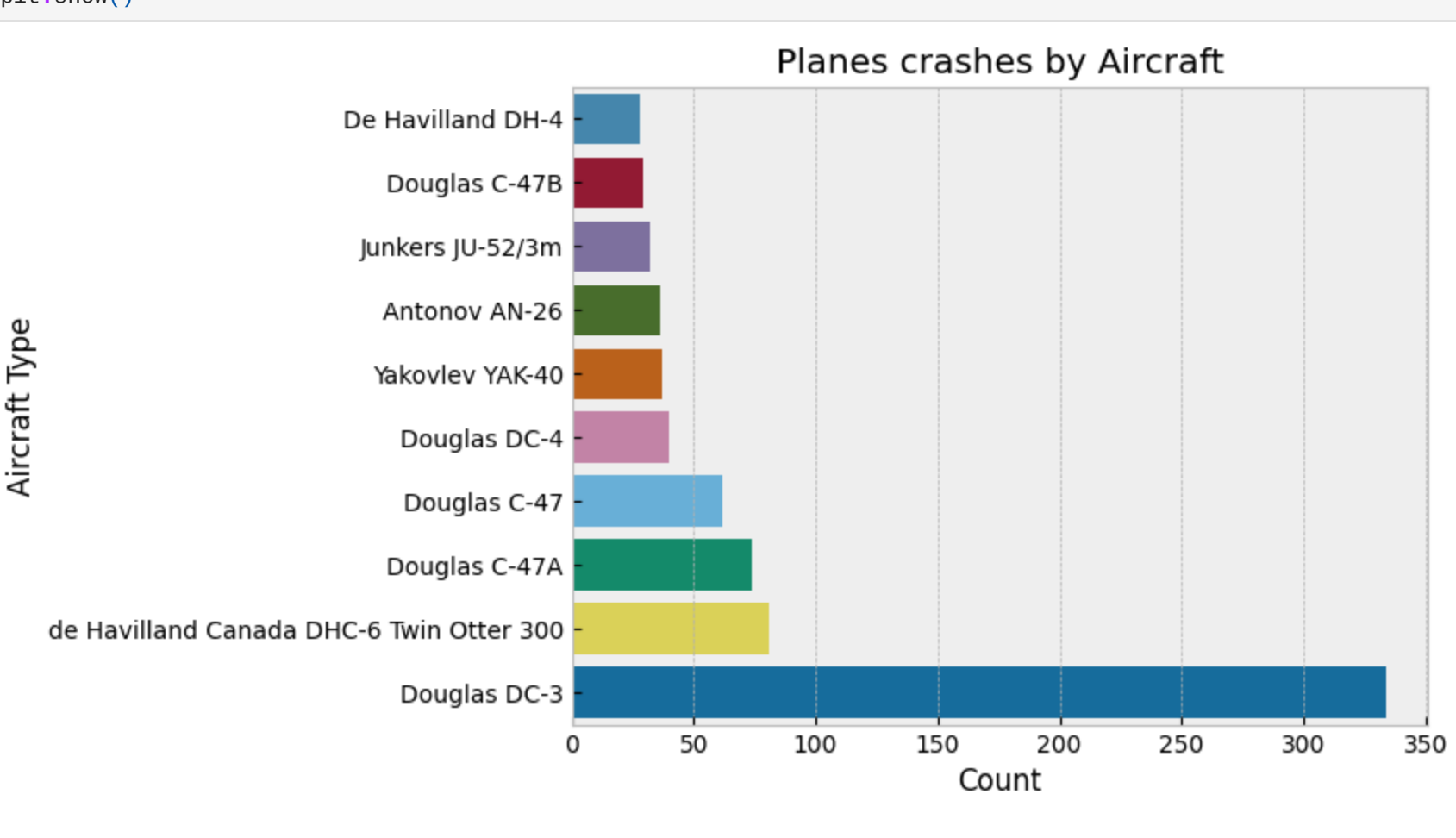
```
In [71]:
location_counts = data['Location Country'].value_counts().head(10)
sns.barplot(x=location_counts.values, y=location_counts.index, orient='horizontal')
plt.title('Planes crashes by location')
plt.xlabel('Count')
plt.ylabel('Location Country')
plt.show()
```



It is evident that Brazil holds the top position in terms of the number of crashes among the locations where incidents occur.

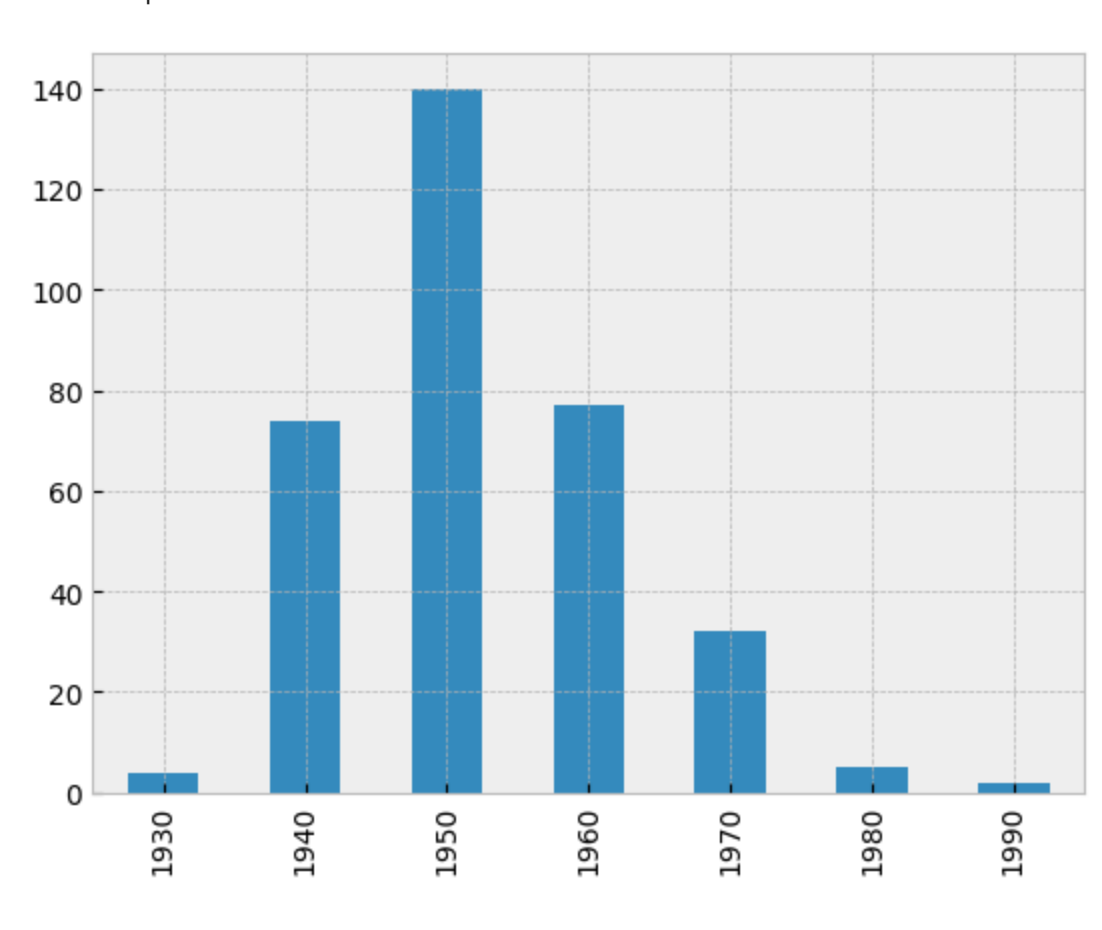
Planes crashes by Aircraft :

```
In [72]:
import seaborn as sns

sns.barplot(x=data['Type'].value_counts().head(10).sort_values(), y=data['Type'].value_counts().head(10).sort_values().index, orient='horizontal')
plt.title('Planes crashes by Aircraft')
plt.xlabel('Count')
plt.ylabel('Aircraft Type')
plt.show()
```




```
In [73]:
data[data['Type'] == 'Douglas DC-3']['Decade'].value_counts().sort_index().plot.bar()
```

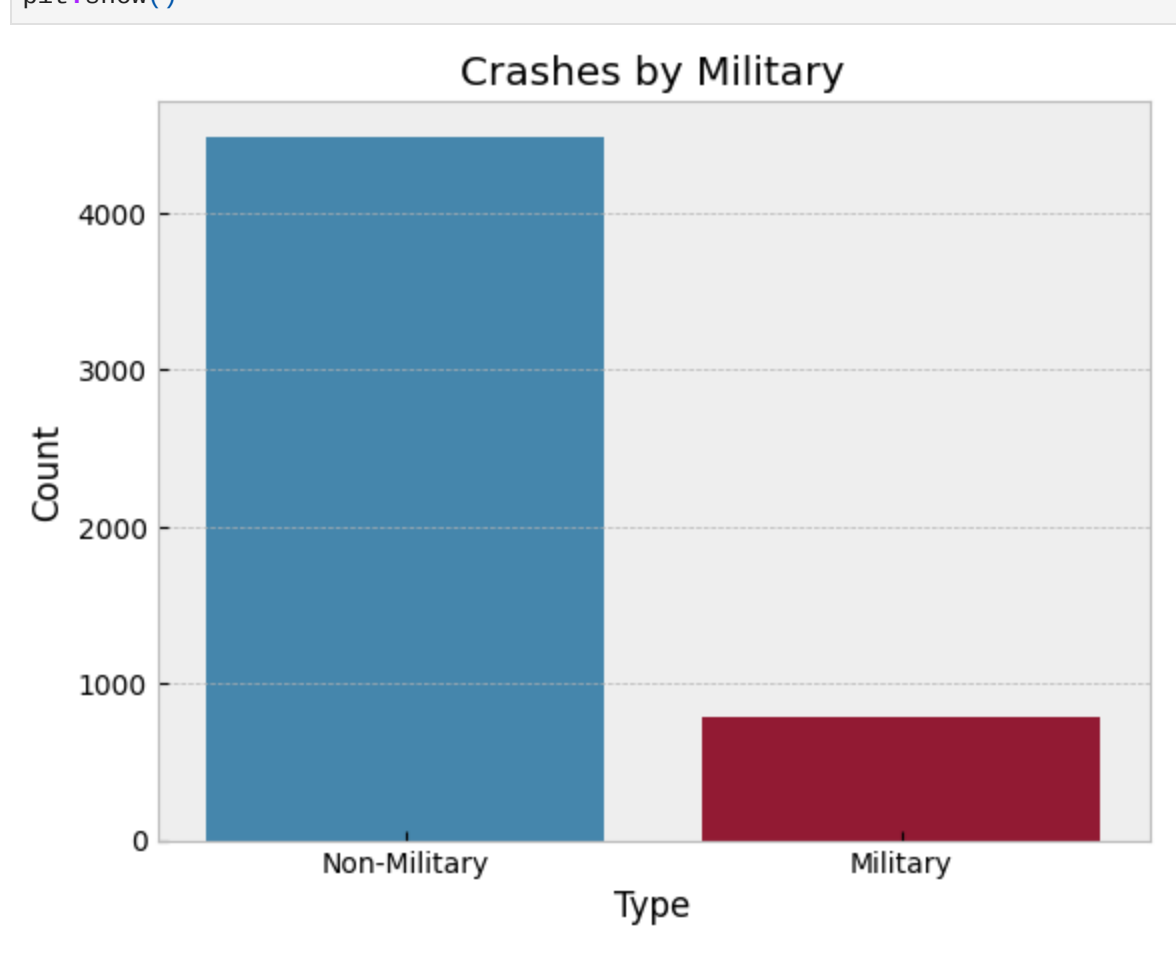
Out[73]: <AxesSubplot: >



The aircraft type with the highest number of crashes throughout history is the Douglas DC-3. Its peak in crash incidents was observed in the 1950s. Presently, the frequency of crashes has significantly diminished, likely due to a decline in usage since the 1990s.

Comparaison between Military and Non-Military Airplanes :

```
In [74]:
import seaborn as sns
import matplotlib.pyplot as plt

sns.barplot(x=data['Is Military'].value_counts().index.map({True: 'Military', False: 'Non-Military'}), y=data['Is Military'].value_counts())
plt.title('Crashes by Military')
plt.xlabel('Type')
plt.ylabel('Count')
plt.show()
```



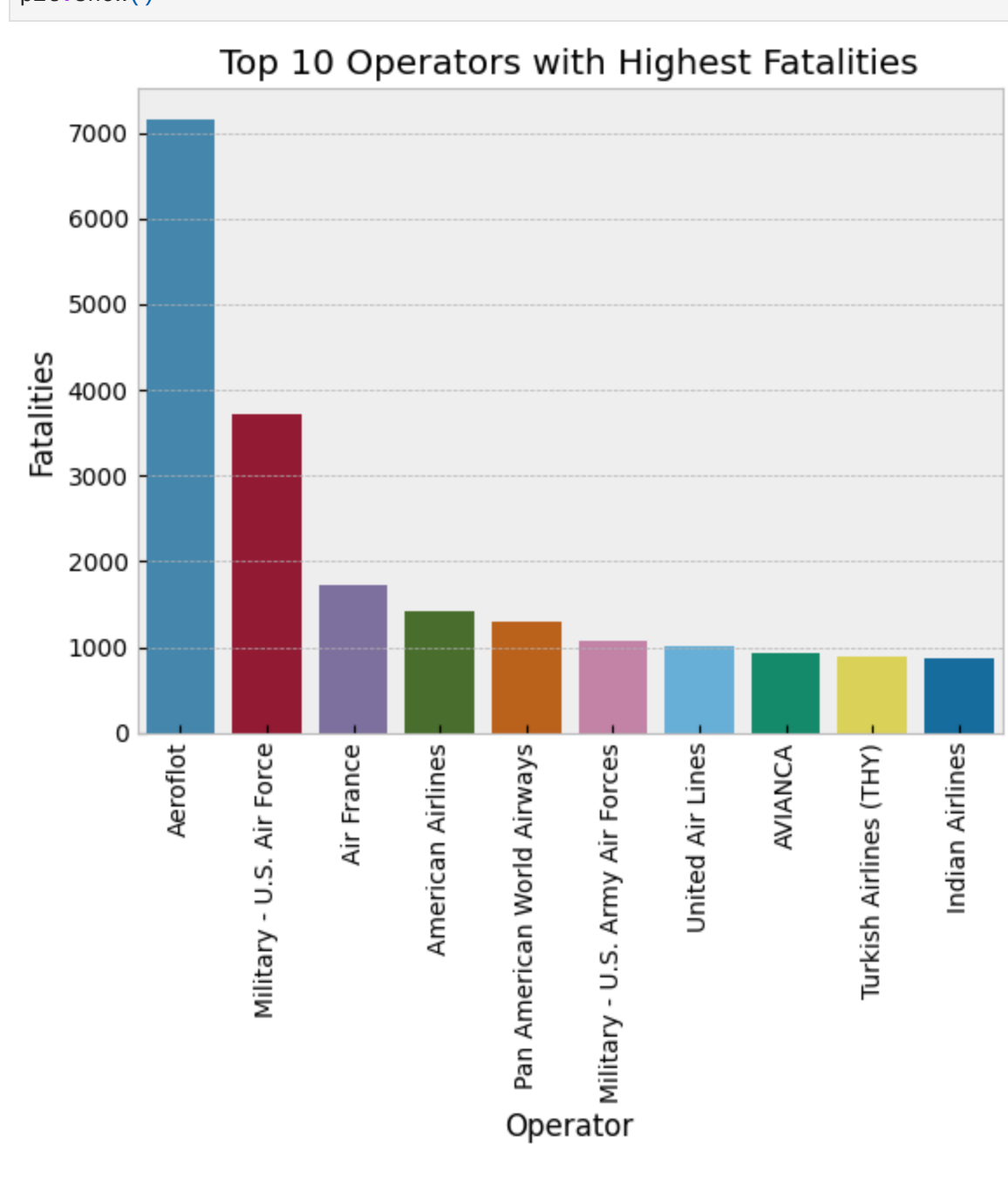
Our observation indicates that the majority, specifically 85%, of crashes are non-military in nature.

Planes crashes by operators :

```
In [75]:
data['Operator'].value_counts()
```

```
Out[75]:
Aeroflot                              179
Military - U.S. Air Force             176
Air France                             70
Deutsche Lufthansa                     65
Air Taxi                               44
                                     ... 
Military - Argentine Navy               1
Richland Flying Service - Air Taxii     1
Harbor Airlines - Air Taxi              1
Aerovias Venezolanas SA (Venezuela)     1
Strait Air                              1
Name: Operator, Length: 2477, dtype: int64
```

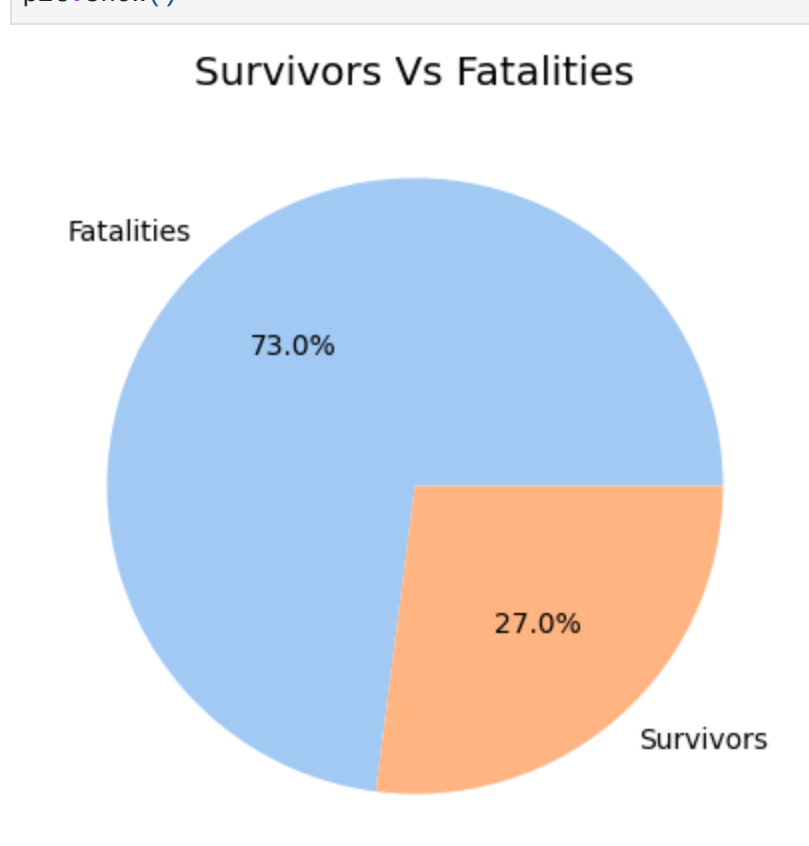
```
In [76]:
sns.barplot(x=data.groupby(['Operator'])['Fatalities'].sum().sort_values(ascending=False).head(10).index, y=data.groupby(['Operator'])['Fatalities'
plt.title('Top 10 Operators with Highest Fatalities')
plt.xlabel('Operator')
plt.ylabel('Fatalities')
plt.xticks(rotation=90)
plt.show()
```


Top 10 Operators with Highest Fatalities
Fatalities
Aeroflot
Military - U.S. Air Force
Air France
American Airlines
Pan American World Airways
Military - U.S. Army Air Forces
United Air Lines
AVIANCA
Turkish Airlines (THY)
Indian Airlines
Operator

We observe that airflot have the highest fatalities

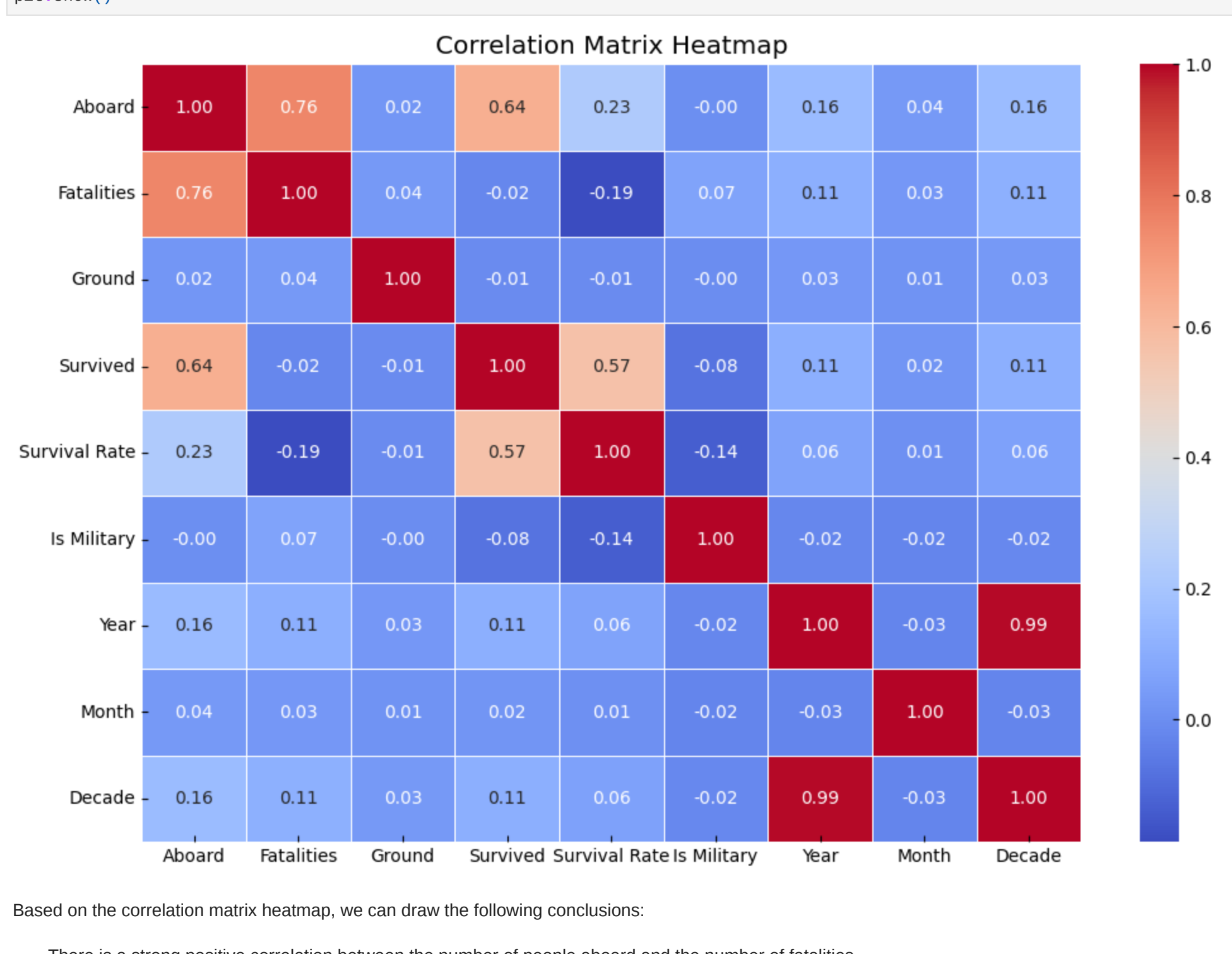
Comparaison between Survivors and Fatalities :

```
In [77]:
sns.set_palette('pastel')
plt.figure(figsize=(8,5))
Aboard = data.Aboard.sum()
Fatalities = data.Fatalities.sum()
Survivors = Aboard - Fatalities

y = np.array([Fatalities, Survivors])
mylabels = ["Fatalities", "Survivors"]

plt.pie(y, labels = mylabels, autopct='%1.1f%%')
plt.title('Survivors Vs Fatalities')
plt.show()
```



Correlation Analysis :

```
In [78]:
# The code block below generates a correlation matrix heatmap using the seaborn library.
# The heatmap visualizes the correlation between numerical columns in the 'data' dataframe.

import seaborn as sns
import matplotlib.pyplot as plt

# Set the figure size
plt.figure(figsize=(12, 8))

# Generate the correlation matrix heatmap
sns.heatmap(data.corr(), annot=True, cmap='coolwarm', fmt='.2f', linewidths=0.5)

# Set the title of the heatmap
plt.title('Correlation Matrix Heatmap')

# Display the heatmap
plt.show()
```


Correlation Matrix Heatmap
Aboard 1.00 0.76 0.02 0.64 0.23 -0.00 0.16 0.04 0.16
Fatalities 0.76 1.00 0.04 -0.02 -0.19 0.07 0.11 0.03 0.11
Ground 0.02 0.04 1.00 -0.01 -0.01 -0.00 0.03 0.01 0.03
Survived 0.64 -0.02 -0.01 1.00 0.57 -0.08 0.11 0.02 0.11
Survival Rate 0.23 -0.19 -0.01 0.57 1.00 -0.14 0.06 0.01 0.06
Is Military -0.00 0.07 -0.00 -0.08 -0.14 1.00 -0.02 -0.02 -0.02
Year 0.16 0.11 0.03 0.11 0.06 -0.02 1.00 -0.03 0.99
Month 0.04 0.03 0.01 0.02 0.01 -0.02 -0.03 1.00 -0.03
Decade 0.16 0.11 0.03 0.11 0.06 -0.02 0.99 -0.03 1.00
Aboard
Fatalities
Ground
Survived
Survival Rate
Is Military
Year
Month
Decade

Based on the correlation matrix heatmap, we can draw the following conclusions:

- There is a strong positive correlation between the number of people aboard and the number of fatalities.
- There is a moderate positive correlation between the number of fatalities and the number of ground fatalities.
- There is a weak positive correlation between the year of the accident and the number of fatalities.
- There is no significant correlation between the survival rate and any other variable.

These conclusions provide insights into the relationships between different variables in the dataset and can help in understanding the factors that contribute to aviation accidents and their consequences.