```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

#Loading the data
df = pd.read_csv(r"C:\\Users\\HP\\Downloads\\phase 1 project dataset\\
AviationData.csv" , encoding= "latin-1")
```

```
C:\Users\HP\anaconda3\envs\learn-env\lib\site-packages\IPython\core\
interactiveshell.py:3145: DtypeWarning: Columns (6,7,28) have mixed
types.Specify dtype option on import or set low_memory=False.
  has_raised = await self.run_ast_nodes(code_ast.body, cell_name,
```

```
#check the first few rows
df.head()
```

```
         Event.Id Investigation.Type Accident.Number  Event.Date  \
0  20001218X45444           Accident      SEA87LA080  1948-10-24
1  20001218X45447           Accident      LAX94LA336  1962-07-19
2  20061025X01555           Accident      NYC07LA005  1974-08-30
3  20001218X45448           Accident      LAX96LA321  1977-06-19
4  20041105X01764           Accident      CHI79FA064  1979-08-02

           Location        Country Latitude Longitude Airport.Code  \
0  MOOSE CREEK, ID  United States      NaN       NaN          NaN
1   BRIDGEPORT, CA  United States      NaN       NaN          NaN
2    Saltville, VA  United States  36.9222  -81.8781          NaN
3       EUREKA, CA  United States      NaN       NaN          NaN
4       Canton, OH  United States      NaN       NaN          NaN

  Airport.Name  ... Purpose.of.flight Air.carrier Total.Fatal.Injuries
\
0          NaN  ...          Personal         NaN                  2.0

1          NaN  ...          Personal         NaN                  4.0

2          NaN  ...          Personal         NaN                  3.0

3          NaN  ...          Personal         NaN                  2.0

4          NaN  ...          Personal         NaN                  1.0


  Total.Serious.Injuries Total.Minor.Injuries Total.Uninjured  \
0                    0.0                   0.0             0.0
1                    0.0                   0.0             0.0
2                    NaN                   NaN             NaN
3                    0.0                   0.0             0.0
4                    2.0                   NaN             0.0

  Weather.Condition  Broad.phase.of.flight   Report.Status
```

```
Publication.Date
0               UNK                  Cruise  Probable Cause
NaN
1               UNK                 Unknown  Probable Cause       19-
09-1996
2               IMC                  Cruise  Probable Cause       26-
02-2007
3               IMC                  Cruise  Probable Cause       12-
09-2000
4               VMC                Approach  Probable Cause       16-
04-1980

[5 rows x 31 columns]

#check on the data info
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 88889 entries, 0 to 88888
Data columns (total 31 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Event.Id                88889 non-null  object
 1   Investigation.Type      88889 non-null  object
 2   Accident.Number         88889 non-null  object
 3   Event.Date              88889 non-null  object
 4   Location                88837 non-null  object
 5   Country                 88663 non-null  object
 6   Latitude                34382 non-null  object
 7   Longitude               34373 non-null  object
 8   Airport.Code            50249 non-null  object
 9   Airport.Name            52790 non-null  object
 10  Injury.Severity         87889 non-null  object
 11  Aircraft.damage         85695 non-null  object
 12  Aircraft.Category       32287 non-null  object
 13  Registration.Number     87572 non-null  object
 14  Make                    88826 non-null  object
 15  Model                   88797 non-null  object
 16  Amateur.Built           88787 non-null  object
 17  Number.of.Engines       82805 non-null  float64
 18  Engine.Type             81812 non-null  object
 19  FAR.Description         32023 non-null  object
 20  Schedule                12582 non-null  object
 21  Purpose.of.flight       82697 non-null  object
 22  Air.carrier             16648 non-null  object
 23  Total.Fatal.Injuries    77488 non-null  float64
 24  Total.Serious.Injuries  76379 non-null  float64
 25  Total.Minor.Injuries    76956 non-null  float64
 26  Total.Uninjured         82977 non-null  float64
 27  Weather.Condition       84397 non-null  object
```

```
 28  Broad.phase.of.flight   61724 non-null  object
 29  Report.Status           82508 non-null  object
 30  Publication.Date        75118 non-null  object
dtypes: float64(5), object(26)
memory usage: 21.0+ MB
```

```
#check on the descriptive statisitcs
df.describe()
```

```
       Number.of.Engines  Total.Fatal.Injuries  Total.Serious.Injuries
\
count       82805.000000          77488.000000            76379.000000

mean            1.146585              0.647855                0.279881

std             0.446510              5.485960                1.544084

min             0.000000              0.000000                0.000000

25%             1.000000              0.000000                0.000000

50%             1.000000              0.000000                0.000000

75%             1.000000              0.000000                0.000000

max             8.000000            349.000000              161.000000


       Total.Minor.Injuries  Total.Uninjured
count          76956.000000     82977.000000
mean               0.357061         5.325440
std                2.235625        27.913634
min                0.000000         0.000000
25%                0.000000         0.000000
50%                0.000000         1.000000
75%                0.000000         2.000000
max              380.000000       699.000000
```

```
#Replace ' .' with ' _ ' in the columns(standardize columns)
df.columns = df.columns.str.replace('.', '_')

#Make a copy of the original dataset
df_copy = df.copy()

#check on the sum of null values per column
df_copy.isna().sum()
```

```
Event_Id                    0
Investigation_Type          0
Accident_Number             0
Event_Date                  0
```

```
Location                    52
Country                    226
Latitude                 54507
Longitude                54516
Airport_Code             38640
Airport_Name             36099
Injury_Severity           1000
Aircraft_damage           3194
Aircraft_Category        56602
Registration_Number       1317
Make                        63
Model                       92
Amateur_Built              102
Number_of_Engines         6084
Engine_Type               7077
FAR_Description          56866
Schedule                 76307
Purpose_of_flight         6192
Air_carrier              72241
Total_Fatal_Injuries     11401
Total_Serious_Injuries   12510
Total_Minor_Injuries     11933
Total_Uninjured           5912
Weather_Condition         4492
Broad_phase_of_flight    27165
Report_Status             6381
Publication_Date         13771
dtype: int64
```

```
#check for duplicates
duplicates = df_copy.duplicated()
num_duplicates = duplicates.sum()
num_duplicates
```

```
0
```

```
#Handling missing values

#Drop the columns with more than 30% null values
df_copy = df_copy.dropna(axis=1 ,thresh = int(0.3 * len(df_copy)))

#Fill the numerical values with the median and data types== object
with mode
for col in df_copy.select_dtypes(include=['number']).columns:
    df_copy[col].fillna(df_copy[col].median(), inplace=True)

for col in df_copy.select_dtypes(include=['object']).columns:
        df_copy[col].fillna(df_copy[col].mode()[0], inplace=True)

df_copy.columns
```

```
Index(['Event_Id', 'Investigation_Type', 'Accident_Number',
'Event_Date',
       'Location', 'Country', 'Latitude', 'Longitude', 'Airport_Code',
       'Airport_Name', 'Injury_Severity', 'Aircraft_damage',
       'Aircraft_Category', 'Registration_Number', 'Make', 'Model',
       'Amateur_Built', 'Number_of_Engines', 'Engine_Type',
'FAR_Description',
       'Purpose_of_flight', 'Total_Fatal_Injuries',
'Total_Serious_Injuries',
       'Total_Minor_Injuries', 'Total_Uninjured', 'Weather_Condition',
       'Broad_phase_of_flight', 'Report_Status', 'Publication_Date'],
      dtype='object')
```

```
# Data Cleaning: Remove whitespace from strings
df_copy = df_copy.apply(lambda x: x.str.strip() if x.dtype == "object"
else x)

df_copy.isna().sum()
```

```
Event_Id                0
Investigation_Type      0
Accident_Number         0
Event_Date              0
Location                0
Country                 0
Latitude                9
Longitude               9
Airport_Code            0
Airport_Name            0
Injury_Severity         0
Aircraft_damage         0
Aircraft_Category       0
Registration_Number     0
Make                    0
Model                   0
Amateur_Built           0
Number_of_Engines       0
Engine_Type             0
FAR_Description         0
Purpose_of_flight       0
Total_Fatal_Injuries    0
Total_Serious_Injuries  0
Total_Minor_Injuries    0
Total_Uninjured         0
Weather_Condition       0
Broad_phase_of_flight   0
Report_Status           0
Publication_Date        0
dtype: int64
```

```
df_copy['Total_Fatal_Injuries'] =
df_copy['Total_Fatal_Injuries'].astype(int)
df_copy['Total_Serious_Injuries'] =
df_copy['Total_Serious_Injuries'].astype(int)
df_copy['Total_Minor_Injuries'] =
df_copy['Total_Minor_Injuries'].astype(int)
df_copy['Total_Uninjured'] = df_copy['Total_Uninjured'] .astype(int)

df_copy
```

	Event_Id	Investigation_Type	Accident_Number	Event_Date
0	20001218X45444	Accident	SEA87LA080	1948-10-24
1	20001218X45447	Accident	LAX94LA336	1962-07-19
2	20061025X01555	Accident	NYC07LA005	1974-08-30
3	20001218X45448	Accident	LAX96LA321	1977-06-19
4	20041105X01764	Accident	CHI79FA064	1979-08-02
...	...	...	...	...
88884	20221227106491	Accident	ERA23LA093	2022-12-26
88885	20221227106494	Accident	ERA23LA095	2022-12-26
88886	20221227106497	Accident	WPR23LA075	2022-12-26
88887	20221227106498	Accident	WPR23LA076	2022-12-26
88888	20221230106513	Accident	ERA23LA097	2022-12-29

	Location	Country	Latitude	Longitude	Airport_Code
0	MOOSE CREEK, ID	United States	332739N	0112457W	NONE
1	BRIDGEPORT, CA	United States	332739N	0112457W	NONE
2	Saltville, VA	United States	NaN	NaN	NONE
3	EUREKA, CA	United States	332739N	0112457W	NONE
4	Canton, OH	United States	332739N	0112457W	NONE
...	...	...	...	...	...
88884	Annapolis, MD	United States	332739N	0112457W	NONE

```
88885      Hampton, NH  United States  332739N  0112457W         NONE

88886       Payson, AZ  United States  341525N  1112021W          PAN

88887       Morgan, UT  United States  332739N  0112457W         NONE

88888       Athens, GA  United States  332739N  0112457W         NONE


      Airport_Name  ... FAR_Description Purpose_of_flight  \
0          Private  ...             091          Personal
1          Private  ...             091          Personal
2          Private  ...             091          Personal
3          Private  ...             091          Personal
4          Private  ...             091          Personal
...            ...  ...             ...               ...
88884      Private  ...             091          Personal
88885      Private  ...             091          Personal
88886       PAYSON  ...             091          Personal
88887      Private  ...             091          Personal
88888      Private  ...             091          Personal

      Total_Fatal_Injuries Total_Serious_Injuries Total_Minor_Injuries
\
0                        2                       0                      0

1                        4                       0                      0

2                        3                       0                      0

3                        2                       0                      0

4                        1                       2                      0

...                    ...                     ...                    ...

88884                    0                       1                      0

88885                    0                       0                      0

88886                    0                       0                      0

88887                    0                       0                      0

88888                    0                       1                      0


      Total_Uninjured Weather_Condition  Broad_phase_of_flight  \
0                   0               UNK                 Cruise
1                   0               UNK                Unknown
2                   1               IMC                 Cruise
```

```
3                    0                IMC                    Cruise
4                    0                VMC                 Approach
...                ...                ...                     ...
88884                0                VMC                  Landing
88885                0                VMC                  Landing
88886                1                VMC                  Landing
88887                0                VMC                  Landing
88888                1                VMC                  Landing

        Report_Status Publication_Date
0      Probable Cause       25-09-2020
1      Probable Cause       19-09-1996
2      Probable Cause       26-02-2007
3      Probable Cause       12-09-2000
4      Probable Cause       16-04-1980
...                ...              ...
88884  Probable Cause       29-12-2022
88885  Probable Cause       25-09-2020
88886  Probable Cause       27-12-2022
88887  Probable Cause       25-09-2020
88888  Probable Cause       30-12-2022

[88889 rows x 29 columns]
```

```
df_copy.to_csv('cleaned aviation_data.csv', index=False)
```