Clean Data File

The JZ Flight School will be evaluating airplane safety by looking at the protection offered by an aircraft during various accidents. The study will determine how this metric is effected by airplane make, model, phase of flight, and time of year.

In order to do this, the study will load a database of accidents from the National Transportation Safety Board, and use a variety of features in the study. To prepare the data, we had to do the following:

- Load the csv file
- Filter for accidents involving one or two engine aircraft
- Normalize manufacturer names and filter for the top 5 aircraft manufacturers
- Imput missing values into the accident statistic column
- Create two new columns: Total.Passengers and Fraction.Fatal

Import Data

In [1]:
```
import pandas as pd
import numpy as np

df = pd.read_csv('Data/AviationData.csv', encoding='latin1')
df.head()
```

```
/var/folders/ym/68nrz1n97wj0gz5413bhpqs80000gn/T/ipykernel_13745/652367376.p
y:4: DtypeWarning: Columns (6,7,28) have mixed types. Specify dtype option on
import or set low_memory=False.
  df = pd.read_csv('Data/AviationData.csv', encoding='latin1')
```

Out[1]:

	Event.Id	Investigation.Type	Accident.Number	Event.Date	Location	C
0	20001218X45444	Accident	SEA87LA080	1948-10-24	MOOSE CREEK, ID	
1	20001218X45447	Accident	LAX94LA336	1962-07-19	BRIDGEPORT, CA	
2	20061025X01555	Accident	NYC07LA005	1974-08-30	Saltville, VA	
3	20001218X45448	Accident	LAX96LA321	1977-06-19	EUREKA, CA	
4	20041105X01764	Accident	CHI79FA064	1979-08-02	Canton, OH	

5 rows × 31 columns

Filter for Accidents and One and Two Engine Aircrafts

Only consider events labeled "Accident" and aircraft with one or two engines.

In [2]:
```
df = df[df['Investigation.Type'] == 'Accident']
df["Investigation.Type"].value_counts()
```

Out[2]:
```
Investigation.Type
Accident    85015
Name: count, dtype: int64
```

Our flight school is only looking at airplanes with one or two engines, so we will filter for those.

In [3]:
```
df = df[(df['Number.of.Engines'] == 1.0) | (df['Number.of.Engines'] == 2.0)]
df["Number.of.Engines"].value_counts()
```

Out[3]:
```
Number.of.Engines
1.0    69069
2.0     9405
Name: count, dtype: int64
```

Normalize the Manufacturer Names

Make certain consistent capitolization is used for manufacturer names.

In [4]:
```
df['Make'] = df['Make'].str.title()
df['Make'].value_counts().head(20)
```

Out[4]:
```
Make
Cessna            25864
Piper             14187
Beech              4918
Bell               2350
Mooney             1281
Grumman            1138
Bellanca           1023
Robinson           1012
Hughes              874
Boeing              819
Air Tractor         647
Aeronca             629
Maule               573
Champion            504
Stinson             434
Luscombe            409
Aero Commander      398
Taylorcraft         376
Schweizer           372
North American      364
Name: count, dtype: int64
```

Filter for top 5 airplane makers

Our flight school will only purchase from the top 5 manufacturers. Filter for the top 5 manufactures represented in the data set.

In [5]:
```
top_five = df['Make'].value_counts().index[:5]
df = df[df["Make"].map(lambda x:x in top_five)]
df["Make"].value_counts()
```

Out[5]:
```
Make
Cessna    25864
Piper     14187
Beech      4918
Bell       2350
Mooney     1281
Name: count, dtype: int64
```

Check for missing value and create suitable fillin

There are values missing from the injury and fatality statistics. Imput the missing values with a value of 0.

In [6]:
```
df["Total.Fatal.Injuries"].fillna(0, inplace=True)
df["Total.Serious.Injuries"].fillna(0, inplace=True)
df["Total.Minor.Injuries"].fillna(0, inplace=True)
df["Total.Uninjured"].fillna(0, inplace=True)

df.head(10).T
```

Out[6]:

	1	2	6	
Event.Id	20001218X45447	20061025X01555	20001218X45446	20020909X(
Investigation.Type	Accident	Accident	Accident	Acc
Accident.Number	LAX94LA336	NYC07LA005	CHI81LA106	SEA82D
Event.Date	1962-07-19	1974-08-30	1981-08-01	1982-
Location	BRIDGEPORT, CA	Saltville, VA	COTTON, MN	PULLMAI
Country	United States	United States	United States	United S
Latitude	NaN	36.922223	NaN	
Longitude	NaN	-81.878056	NaN	
Airport.Code	NaN	NaN	NaN	
Airport.Name	NaN	NaN	NaN	BLACKBUR
Injury.Severity	Fatal(4)	Fatal(3)	Fatal(4)	Non
Aircraft.damage	Destroyed	Destroyed	Destroyed	Subst
Aircraft.Category	NaN	NaN	NaN	Air
Registration.Number	N5069P	N5142R	N4988E	N2
Make	Piper	Cessna	Cessna	C
Model	PA24-180	172M	180	
Amateur.Built	No	No	No	
Number.of.Engines	1.0	1.0	1.0	
Engine.Type	Reciprocating	Reciprocating	Reciprocating	Reciproc
FAR.Description	NaN	NaN	NaN	Part 91: G Av

Schedule	NaN	NaN	NaN	
Purpose.of.flight	Personal	Personal	Personal	Per
Air.carrier	NaN	NaN	NaN	
Total.Fatal.Injuries	4.0	3.0	4.0	
Total.Serious.Injuries	0.0	0.0	0.0	
Total.Minor.Injuries	0.0	0.0	0.0	
Total.Uninjured	0.0	0.0	0.0	
Weather.Condition	UNK	IMC	IMC	
Broad.phase.of.flight	Unknown	Cruise	Unknown	Ta
Report.Status	Probable Cause	Probable Cause	Probable Cause	Probable C
Publication.Date	19-09-1996	26-02-2007	06-11-2001	01-01

Add new columns: Survive, Total_Passangers, Month

Add new columns survive, total passengers, month, and year:

In [7]:
```
df['Survive'] = df['Total.Fatal.Injuries'] == 0
df["total.passengers"] = df["Total.Fatal.Injuries"] + df["Total.Serious.Inju

df['Month'] = (pd.to_datetime(df['Event.Date'])).dt.month
df['Year'] = (pd.to_datetime(df['Event.Date'])).dt.year

df.head().T
```

Out[7]:

	1	2	6	
Event.Id	20001218X45447	20061025X01555	20001218X45446	20020909X0
Investigation.Type	Accident	Accident	Accident	Acc
Accident.Number	LAX94LA336	NYC07LA005	CHI81LA106	SEA82D
Event.Date	1962-07-19	1974-08-30	1981-08-01	1982-
Location	BRIDGEPORT, CA	Saltville, VA	COTTON, MN	PULLMAI
Country	United States	United States	United States	United S
Latitude	NaN	36.922223	NaN	
Longitude	NaN	-81.878056	NaN	

Airport.Code	NaN	NaN	NaN	
Airport.Name	NaN	NaN	NaN	BLACKBUR
Injury.Severity	Fatal(4)	Fatal(3)	Fatal(4)	Non
Aircraft.damage	Destroyed	Destroyed	Destroyed	Subst
Aircraft.Category	NaN	NaN	NaN	Air
Registration.Number	N5069P	N5142R	N4988E	N2
Make	Piper	Cessna	Cessna	C
Model	PA24-180	172M	180	
Amateur.Built	No	No	No	
Number.of.Engines	1.0	1.0	1.0	
Engine.Type	Reciprocating	Reciprocating	Reciprocating	Reciproc
FAR.Description	NaN	NaN	NaN	Part 91: G Av
Schedule	NaN	NaN	NaN	
Purpose.of.flight	Personal	Personal	Personal	Per
Air.carrier	NaN	NaN	NaN	
Total.Fatal.Injuries	4.0	3.0	4.0	
Total.Serious.Injuries	0.0	0.0	0.0	
Total.Minor.Injuries	0.0	0.0	0.0	
Total.Uninjured	0.0	0.0	0.0	
Weather.Condition	UNK	IMC	IMC	
Broad.phase.of.flight	Unknown	Cruise	Unknown	Ta
Report.Status	Probable Cause	Probable Cause	Probable Cause	Probable
Publication.Date	19-09-1996	26-02-2007	06-11-2001	01-01
Survive	False	False	False	
total.passengers	4.0	3.0	4.0	
Month	7	8	8	
Year	1962	1974	1981	

Add new columns: Fraction_Fatal, Fraction_uninjured

Add new columns fraction fatal, fraction uninjured to the dataframe.

In [10]:
```
df["Fraction_fatal"] = df["Total.Fatal.Injuries"]/df["total.passengers"]
df["Fraction_uninjured"] = df["Total.Uninjured"]/df["total.passengers"]
df.head().T
```

Out[10]:

	1	2	6	
Event.Id	20001218X45447	20061025X01555	20001218X45446	20020909X(
Investigation.Type	Accident	Accident	Accident	Acc
Accident.Number	LAX94LA336	NYC07LA005	CHI81LA106	SEA82D
Event.Date	1962-07-19	1974-08-30	1981-08-01	1982-
Location	BRIDGEPORT, CA	Saltville, VA	COTTON, MN	PULLMAI
Country	United States	United States	United States	United S
Latitude	NaN	36.922223	NaN	
Longitude	NaN	-81.878056	NaN	
Airport.Code	NaN	NaN	NaN	
Airport.Name	NaN	NaN	NaN	BLACKBUR
Injury.Severity	Fatal(4)	Fatal(3)	Fatal(4)	Non
Aircraft.damage	Destroyed	Destroyed	Destroyed	Subst
Aircraft.Category	NaN	NaN	NaN	Air
Registration.Number	N5069P	N5142R	N4988E	N2
Make	Piper	Cessna	Cessna	C
Model	PA24-180	172M	180	
Amateur.Built	No	No	No	
Number.of.Engines	1.0	1.0	1.0	
Engine.Type	Reciprocating	Reciprocating	Reciprocating	Reciproc
FAR.Description	NaN	NaN	NaN	Part 91: G Av
Schedule	NaN	NaN	NaN	

Purpose.of.flight	Personal	Personal	Personal	Per
Air.carrier	NaN	NaN	NaN	
Total.Fatal.Injuries	4.0	3.0	4.0	
Total.Serious.Injuries	0.0	0.0	0.0	
Total.Minor.Injuries	0.0	0.0	0.0	
Total.Uninjured	0.0	0.0	0.0	
Weather.Condition	UNK	IMC	IMC	
Broad.phase.of.flight	Unknown	Cruise	Unknown	Ta
Report.Status	Probable Cause	Probable Cause	Probable Cause	Probable C
Publication.Date	19-09-1996	26-02-2007	06-11-2001	01-01
Survive	False	False	False	
total.passengers	4.0	3.0	4.0	
Month	7	8	8	
Year	1962	1974	1981	
Fraction_fatal	1.0	1.0	1.0	
Fraction_uninjured	0.0	0.0	0.0	

Write the DataFrame to a CSV file

In [11]:
```
df.to_csv('Data/AviationDataClean.csv')
```