Identify working file

```
file = 'aviationData/AviationData.csv'
```

Import Necessary libraries

```
import csv
import pandas as pd
```

Transforming the data

```
with open(file) as f:
    data = pd.read_csv(file, sep=',' ,encoding='cp1251')
```

```
C:\Users\Mulwa\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,
```

```
data.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',
       'Schedule', 'Purpose.of.flight', 'Air.carrier',
'Total.Fatal.Injuries',
       'Total.Serious.Injuries', 'Total.Minor.Injuries',
'Total.Uninjured',
       'Weather.Condition', 'Broad.phase.of.flight', 'Report.Status',
       'Publication.Date'],
      dtype='object')
```

Isolating analytically relevant columns

```
relevant_columns = [
    'Investigation.Type',
    'Aircraft.Category',
    'Make',
    'Model',
    'Injury.Severity',
    'Event.Date',
    'Country',
    'Location',
    'Purpose.of.flight',
    'Weather.Condition',
```

```
    'Broad.phase.of.flight',
    'Aircraft.damage',
]
```

Creating the dataframe

```
df = data[relevant_columns]
```

Dropping rows of missing values for relevant fields

```
df = df.dropna(subset=['Injury.Severity', 'Country', 'Location',
'Aircraft.Category', 'Purpose.of.flight', 'Weather.Condition',
'Broad.phase.of.flight', 'Aircraft.damage', 'Make', 'Model'])
```

Confirm null or missing values

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 7080 entries, 7 to 63911
Data columns (total 12 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   Investigation.Type     7080 non-null   object
 1   Aircraft.Category      7080 non-null   object
 2   Make                   7080 non-null   object
 3   Model                  7080 non-null   object
 4   Injury.Severity        7080 non-null   object
 5   Event.Date             7080 non-null   object
 6   Country                7080 non-null   object
 7   Location               7080 non-null   object
 8   Purpose.of.flight      7080 non-null   object
 9   Weather.Condition      7080 non-null   object
 10  Broad.phase.of.flight  7080 non-null   object
 11  Aircraft.damage        7080 non-null   object
dtypes: object(12)
memory usage: 719.1+ KB
```

Isolate data from the United States

```
us = df.loc[df['Country'] == 'United States']
```

Normalize location data into states

```
bas = us['Location'].to_list()
cleaned = []
for index in bas:
```

```
    cleaned.append(index.split(',')[1].lstrip())
states = pd.Series(cleaned)
```

Reseting Index

```
us.reset_index(inplace=True)

us = us.drop('index', axis=1)
```

Change location column to states and rename

```
us['Location'] = states
us.rename(columns={'Location':'State'}, inplace=True)
```

Export cleaned dataframe to csv format

```
us.to_csv('cleaned_NTSB.csv')
```