In [1]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
```

In [29]:
```
AC = pd.read_csv('E:/data/Airplane_Crashes_and_Fatalities.csv')
AC.sample(5)
```

Out[29]:

	Date	Time	Location	Operator	Flight #	Route	AC Type	Registration	cn/ln
470	08/10/1937	04:40	Daytona Beach, Florida	Eastern Air Lines	7	Chicago - Daytona	Douglas DC-2	NC-13739	1289
3837	10/04/1993	NaN	Near Svanetia, Georgia	Military - Georgian Air Force	NaN	NaN	Mil Mi-8 (helicopter)	NaN	NaN
3004	04/18/1979	18:23	Newark, New Jersey	New York Airways	NaN	Newark, NJ - Flushing, NY	Sikorsky S61-L	N618PA	61426
418	04/30/1936	NaN	Basel, Switzerland	Swissair	NaN	NaN	General Aviation GA-43	HB-ITU	2204
4705	04/09/2009	07:00	Near Wamena, Indonesia	Aviastar Mandiri	NaN	Jayapura - Wamena	British Aerospace BAe-146-300	PK-BRD	E3189

In [3]:
```
AC.sample(2)
```

Out[3]:

	Date	Time	Location	Operator	Flight #	Route	AC Type	Registration	cn/ln	Ab
2324	05/03/1968	15:48	Near Dawson, Texas	Braniff Airlines	352	Houston - Dallas-Fort Worth	Lockheed L188A Electra	N9707C	1099	
3972	10/04/1995	NaN	Barskoon, Kirghizia	Kirghizia Aba Zaoldoru	NaN	NaN	Mil Mi-8MTV-1	EX-25179	95489	

In [4]:
```
AC['Route'].value_counts().head(15)
```

Out[4]:
```
Training                      92
Sightseeing                   31
Test flight                   22
Sao Paulo - Rio de Janeiro     7
Test                           6
Rio de Janeiro - Sao Paulo     5
Tallinn - Helsinki             4
Villavicencio - Mitu           4
Barranquilla - Bogota          4
Croydon - Paris                4
Paris - London                 4
Bogota - Barranquilla          4
Huambo - Luanda                4
Sao Paulo - Porto Alegre       4
Demonstration                  4
Name: Route, dtype: int64
```

In [5]:
```
AC['Route'].replace('Test','Test flight', inplace = True)
```

In [6]:
```
AC['Operator'].value_counts().head(15)
```

Out[6]:
```
Aeroflot                                255
Military - U.S. Air Force               140
Air France                               72
Deutsche Lufthansa                       63
United Air Lines                         44
China National Aviation Corporation      43
Military - U.S. Army Air Forces          43
Pan American World Airways               41
American Airlines                        37
US Aerial Mail Service                   35
Indian Airlines                          34
KLM Royal Dutch Airlines                 34
Philippine Air Lines                     33
Military - Royal Air Force               32
Private                                  30
Name: Operator, dtype: int64
```

In [7]:
```
AC.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4967 entries, 0 to 4966
Data columns (total 17 columns):
Date                    4967 non-null object
Time                    3457 non-null object
Location                4963 non-null object
Operator                4957 non-null object
Flight #                1315 non-null object
Route                   4193 non-null object
AC Type                 4952 non-null object
Registration            4694 non-null object
cn/ln                   4299 non-null object
Aboard                  4949 non-null float64
Aboard Passangers       4738 non-null float64
Aboard Crew             4741 non-null float64
Fatalities              4959 non-null float64
Fatalities Passangers   4725 non-null float64
Fatalities Crew         4726 non-null float64
Ground                  4926 non-null float64
Summary                 4903 non-null object
dtypes: float64(7), object(10)
memory usage: 465.7+ KB
```

In [8]:
```
AC['Location'] = AC['Location'].astype(str)
```

In [9]:
```
AC['Location'] = AC['Location'].apply(lambda x: x.split(',')[-1])
```

In [10]:
```
AC['Year'] = AC['Date'].str[6:]
```

In [11]:
```
AC['Month'] = AC['Date'].str[:2]
```

In [12]:
```
AC.sample(2)
```

Out[12]:

	Date	Time	Location	Operator	Flight #	Route	AC Type	Registration	cn/
4132	07/13/1998	23:15	United Arab Emirates	ATI Aircompany	NaN	Rasal - Nikolaev, Ukraine	Ilyushin 76MD	UR-76424	00634700
883	09/05/1946	02:00	Nevada	Trans-Luxury Airlines	850	New York City - San Francisco	Douglas DC-3 (C-47-A5-DL)	NC57850	92

In [13]:
```
AC['Year'].value_counts().head(5)
```

Out[13]:
```
1946    88
1989    83
1947    82
1948    77
1972    77
Name: Year, dtype: int64
```

In [14]:
```
AC['AC Type'].value_counts()
```

Out[14]:

Douglas DC-3	333
de Havilland Canada DHC-6 Twin Otter 300	81
Douglas C-47A	70
Douglas C-47	64
Douglas DC-4	38
Yakovlev YAK-40	35
Antonov AN-26	34
Junkers JU-52/3m	31
De Havilland DH-4	27
Douglas C-47B	27
Douglas DC-6B	27
Breguet 14	22
Douglas DC-6	20
Antonov AN-24	19
Curtiss C-46A	19
Antonov AN-12	18
Curtiss C-46	18
Douglas C-47-DL	18
McDonnell Douglas DC-9-32	18
Fokker F-27 Friendship 600	17
Junkers F-13	17
Tupolev TU-134A	16
Fokker F-27 Friendship 200	16
Embraer 110P1 Bandeirante	16
Douglas DC-3C	16
de Havilland Canada DHC-6 Twin Otter 100	15
Lockheed 18 Lodestar	15
Ilyushin IL-18B	15
Lockheed C-130H	15
Ilyushin IL-12	13
...	...
ATR 72-201	1
Fiat G.212CP	1
MD Douglas DC-9-15 / Beechcraft Baron-55	1
Lockheed EC-130E Hercules	1
Curtiss Carrier Pigeon	1
Boeing B-737-2A8 Advanced	1
Fokker F-VIIA	1
Fokker F-VIIB	1
Fokker F-VIII	1
Rockwell Sabreliner 60	1
Aerospatiale ATR-42-512	1
Boeing B-737-2P5	1
Tupolev TU-134A / Antonov An-26	1
Boeing B-247-D	1
Ford Tri-motor F-VIII	1
\tBoeing 737-8KN (WL)	1
British Aerospace Nimrod MR-2P	1
Boeing B-707-331	1
Rockwell CT-39A Sabreliner	1
British Aerospace BAe-146-300	1
Shorts SC-7 Skyvan	1
Rockwell 500S Shrike Commander	1
Boeing B-727-228	1
Douglas R6D-1 (DC-6)	1
Sikorsky S-43 (flying boat)	1
Airspeed Ambassador AS-57	1

```
Arado V1                                    1
Shaanxi Y-8F-200W                           1
Airbus A-330-243                            1
deHavilland DH-86                           1
Name: AC Type, Length: 2447, dtype: int64
```

In [15]:
```
AC['Survived'] = AC['Aboard']-AC['Fatalities']
```

Include Fatalities on ground in Total Fatalities:-

In [16]:
```
AC['Fatalities'] = AC['Fatalities'] + AC['Ground']
```

In [17]:
```
AC['Year'] = AC['Year'].astype(int)
AC['Location'] = AC['Location'].astype(str)
```

In [18]:
```
len(AC[AC['Survived']== 0])
```

Out[18]: 3164

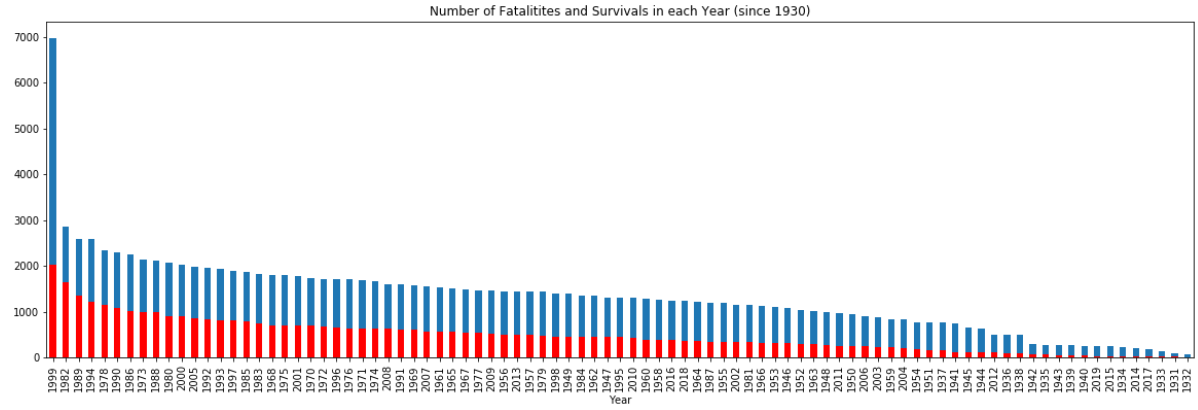
In [19]:
```
AC.drop(['Time','Flight #','cn/ln','Summary','Aboard','Aboard Passangers','Abo
ard Crew','Fatalities Passangers','Ground','Fatalities Crew','Date','Registrat
ion'],axis=1,inplace=True)
```

In [20]:
```
AC.sample(5)
```

Out[20]:

	Location	Operator	Route	AC Type	Fatalities	Year	Month	Survived
372	Michigan	Knowles Flying Service of Detroit	Sightseeing	Ford model 4-AT-E	3.0	1935	05	9.0
1380	Washington	Miami Airlines	Spokane - Ellenburgh	Douglas DC-3 (C-47-DL)	7.0	1953	04	18.0
3525	Canada	Air Ontario	Dryden - Winnipeg	Fokker F-28 Fellowship 1000	24.0	1989	03	45.0
932	England	Spencer Airways / Ceskoslovenske Aerolinie	Croydon - Rome - Salisbury	Douglas C-47 / Douglas DC-3	12.0	1947	01	11.0
4204	Cape Verde Islands	TACV-Cabo Verde Airlines	Sao Vicente - Santo Antao	Dornier 228-201	18.0	1999	08	0.0

In [21]:
```
plt.figure(figsize=(20,6))
AC[AC['Year']>1930].groupby(['Year'])['Fatalities'].sum().sort_values(ascending=False).head(109).plot(kind='bar');
AC[AC['Year']>1930].groupby(['Year'])['Survived'].sum().sort_values(ascending=False).head(109).plot(kind='bar',color='red');
plt.title('Number of Fatalitites and Survivals in each Year (since 1930)');
```



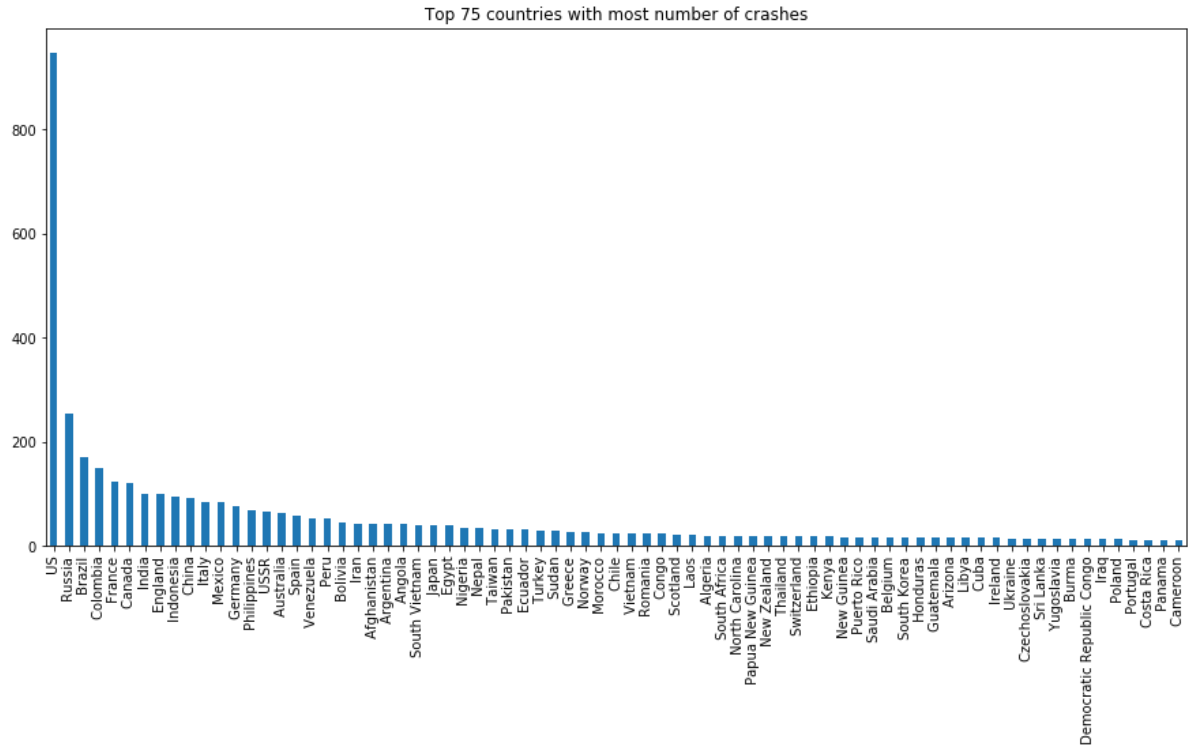
In [22]:
```
AC['Location'].value_counts().head(10)
```

Out[22]:
```
Russia        253
Brazil        170
Colombia      149
France        123
Canada        121
California    115
India         101
Alaska        101
England       100
Indonesia      94
Name: Location, dtype: int64
```

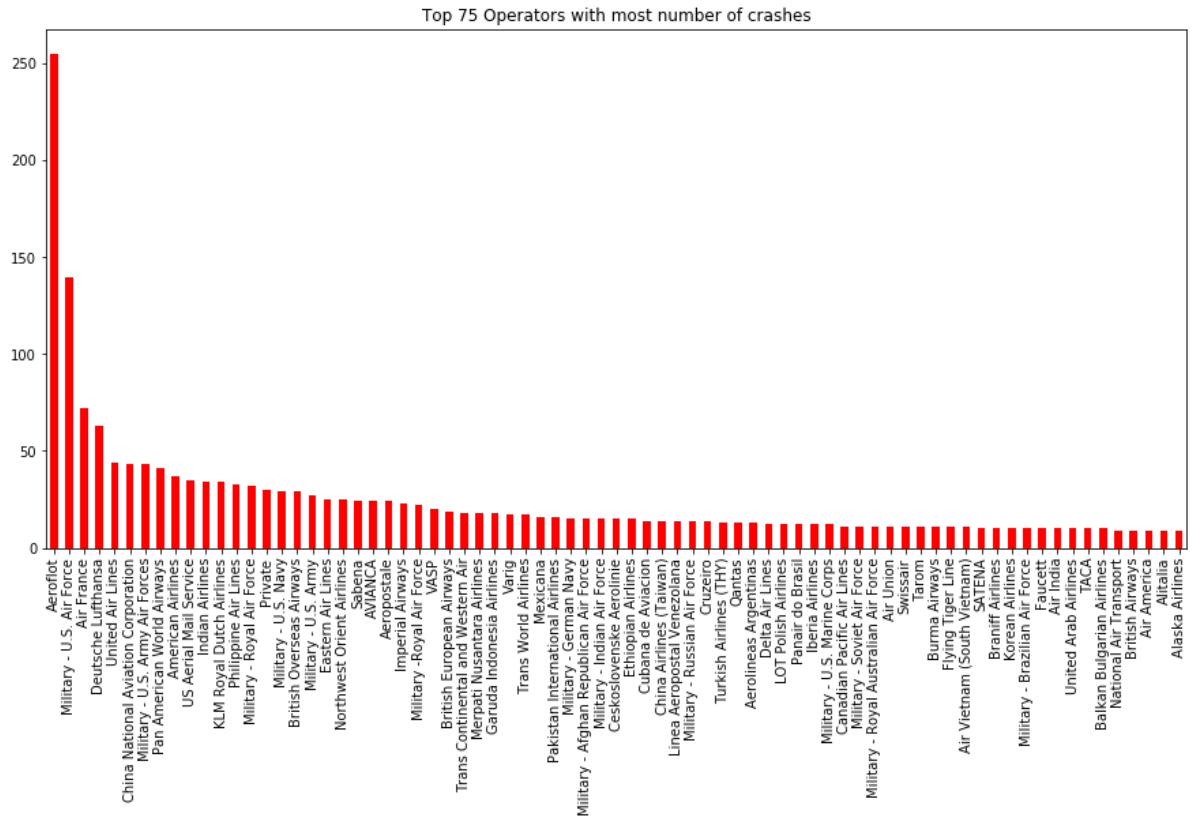
In [23]:
```
AC['Location'] = AC['Location'].str.replace('Ohio','US').str.replace('Nevada',
'US').str.replace('Texas','US').str.replace('Alabama','US').str.replace('Alask
a','US').str.replace('Ariona','US').str.replace('Arkansas','US').str.replace(
'California','US').str.replace('Colorado','US').str.replace('Connecticut','US'
).str.replace('Delaware','US').str.replace('Florida','US').str.replace('Hawai
i','US').str.replace('Georgia','US').str.replace('Idaho','US').str.replace('Il
linois','US').str.replace('Indiana','US').str.replace('Iowa','US').str.replace
('Kansas','US').str.replace('Kentucky','US').str.replace('Louisiana','US').str
.replace('Maine','US').str.replace('Maryland','US').str.replace('Massachusett
s','US').str.replace('Michigan','US').str.replace('Minnesota','US').str.replac
e('Mississippi','US').str.replace('Missouri','US').str.replace('Montana','US')
.str.replace('Nebraska','US').str.replace('New Hampshire','US').str.replace('N
ew Jersey','US').str.replace('New Mexico','US').str.replace('New York','US').s
tr.replace('North Caolina','US').str.replace('North Dakota','US').str.replace(
'Oklahoma','US').str.replace('Oregon','US').str.replace('Pennsylvania','US').s
tr.replace('Rhode Island','US').str.replace('South Dakota','US').str.replace(
'Tennessee','US').str.replace('Utah','US').str.replace('Vermont','US').str.rep
lace('Virginia','US').str.replace('Washington','US').str.replace('West Virgini
a','US').str.replace('Wisconsin','US').str.replace('Wyoming','US').str.replace
('United States','US').str.replace('United States of America','US').str.replac
e('USA','US').str.replace('The United States of America','US')
AC['Location'].value_counts().head(10)
```

Out[23]:
```
US           947
Russia       253
Brazil       170
Colombia     149
France       123
Canada       121
India        101
England      100
Indonesia     94
China         91
Name: Location, dtype: int64
```

In [28]:
```
plt.figure(figsize=(15,7))
AC['Location'].value_counts().head(75).plot(kind='bar');
plt.title('Top 75 countries with most number of crashes');
```



In [25]:
```
plt.figure(figsize=(15,7))
AC['Operator'].value_counts().head(75).plot(kind='bar', color = 'red');
plt.title('Top 75 Operators with most number of crashes');
```



In [26]:
```
plt.figure(figsize=(15,7))
AC['AC Type'].value_counts().head(75).plot(kind='bar', color = 'grey');
plt.title('Top 75 Aircrafts with most number of crashes');
```