JN

Project - New_York_City_Leading_Causes_of_Death

Dataset selected: New_York_City_Leading_Causes_of_Death

Dataset found: (https://opendata.cityofnewyork.us/)

The information from New_York_City_Leading_Causes_of_Death dataset provides information on the leading causes of death in new work city between the years of 2007 and 2018 which is provided by the Bureau of Vital Statistics and New York City Department of Health and Mental Hygiene

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

```
#Loading the data as a pandas DataFrame.
nyclcddf = pd.read_csv('data/New_York_City_Leading_Causes_of_Death.csv', sep = ',', in
#displaying the first 5 rows of the dataset using .head(# of rows) function
nyclcddf.head(5)
```

Year	Leading Cause	Sex	Race Ethnicity	Deaths	Death Rate	Age Adjusted Death Rate
2010	Influenza (Flu) and Pneumonia (J09-J18)	F	Hispanic	228	18.7	23.1
2008	Accidents Except Drug Posioning (V01-X39, X43,...	F	Hispanic	68	5.8	6.6
2013	Accidents Except Drug Posioning (V01-X39, X43,...	M	White Non-Hispanic	271	20.1	17.9

```
#renaming column features in nyclcddf dataset to not have space inbetween letters
nyclcddf.rename(columns = {'Leading Cause':'LeadingCause','Race Ethnicity': 'RaceEthn
nyclcddf.head(1)
```

Year	LeadingCause	Sex	RaceEthnicity	Deaths	DeathRate	AgeAdjustedDeathR
2010	Influenza (Flu) and Pneumonia	F	Hispanic	228	18.7	

```
# Getting the number of instances in nyclcddf using .shape that returns 1094 rows and
```

```
nyclcddf.shape
```

```
(1094, 6)
```

```
#NOTE: nyclcddf.shape showed 6 rows excluding the index [year]
# For this project I am interested in using the index [year] in the nyclcddf dataset s
nyclcddf.reset_index(inplace=True)
nyclcddf = nyclcddf.rename(columns = {'index':'Year'})
```

```
# checking if the added colum[Year] is added into the nyclcddf dataframe by using .sha
nyclcddf.shape
```

```
(1094, 7)
```

```
# Getting the number of features in nyclcddf dataset using the info() method
nyclcddf.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1094 entries, 0 to 1093
Data columns (total 7 columns):
 #   Column                Non-Null Count  Dtype 
---  ------                --------------  ----- 
 0   Year                  1094 non-null   int64 
 1   LeadingCause          1094 non-null   object
 2   Sex                   1094 non-null   object
 3   RaceEthnicity         1094 non-null   object
 4   Deaths                1094 non-null   object
 5   DeathRate             1094 non-null   object
 6   AgeAdjustedDeathRate  1094 non-null   object
dtypes: int64(1), object(6)
memory usage: 60.0+ KB
```

```
# Getting the feature names of the nyclcddf datatset in a list by reading the column v
list(nyclcddf.columns.values)
```

```
['Year',
 'LeadingCause',
 'Sex',
 'RaceEthnicity',
 'Deaths',
 'DeathRate',
 'AgeAdjustedDeathRate']
```

```
# Viewing tha data type of each feature in nyclcddf dataset by using .dtypes
nyclcddf.dtypes
```

```
Year                   int64
LeadingCause          object
Sex                   object
RaceEthnicity         object
Deaths                object
```

```
DeathRate               object
AgeAdjustedDeathRate    object
dtype: object
```

```
# Checking for any missing values for each feature in the nyclcddf dataset using isna(
nyclcddf.isna().sum()
```

```
Year                    0
LeadingCause            0
Sex                     0
RaceEthnicity           0
Deaths                  0
DeathRate               0
AgeAdjustedDeathRate    0
dtype: int64
```

```
# NOTE: results show that Deaths, DeathRate and AgeAdjustedDeathRate are objects
# but when displayed are either integers or floats.
# NOTE: although when checking isna() value and add its .sum() returns 0.
# Some columns appear to be empty with '.' I will replace ',' values with np.Nan
nyclcddf = nyclcddf.replace('.', np.NaN)
nyclcddf.head(7)
```

	Year	LeadingCause	Sex	RaceEthnicity	Deaths	DeathRate	AgeAdjustedDea
0	2010	Influenza (Flu) and Pneumonia (J09-J18)	F	Hispanic	228	18.7	
1	2008	Accidents Except Drug Posioning (V01-X39, X43,...	F	Hispanic	68	5.8	
2	2013	Accidents Except Drug Posioning (V01-X39, X43,...	M	White Non-Hispanic	271	20.1	

```
## Getting the percentage of missing values
total = nyclcddf.isnull().sum().sort_values(ascending=False) #adding up isnull values
percent = (nyclcddf.isnull().sum()/len(nyclcddf) * 100).sort_values(ascending=False)#
missing_data = pd.concat([total, percent], axis=1, keys=['total', 'percent'])
missing_data.head(7) #displaying precentage of missing values for each feature
```

	total	percent
AgeAdjustedDeathRate	386	35.283364
DeathRate	386	35.283364
Deaths	138	12.614260

NOTE: I will drop AgeAdjustedDeathRate and DeathRate columns since its missing over 30% of data and does not interfere with our main question of what sex(gender) has been recorded most with the top 5 leading causes of death.

```
# Creating a new dataframe called train_nyclcddf with feature [Year, LeadingCause, Sex
train_nyclcddf = nyclcddf.drop(columns=['DeathRate', 'AgeAdjustedDeathRate'])
train_nyclcddf.head()
```

	Year	LeadingCause	Sex	RaceEthnicity	Deaths
0	2010	Influenza (Flu) and Pneumonia (J09-J18)	F	Hispanic	228
1	2008	Accidents Except Drug Posioning (V01-X39, X43,...	F	Hispanic	68
2	2013	Accidents Except Drug Posioning (V01-X39, X43,...	M	White Non-Hispanic	271
3	2010	Cerebrovascular Disease (Stroke: I60-I69)	M	Hispanic	140

```
#Deaths in the train_nyclcddf dataset still has missing values. I will replace nan wit
train_nyclcddf.isna().sum()
```

```
Year              0
LeadingCause      0
Sex               0
RaceEthnicity     0
Deaths          138
dtype: int64
```

```
#printing the mean of Deaths column
print(train_nyclcddf["Deaths"].astype(float).mean())
```

```
444.55857740585776
```

```
#creating a function to replace column value with input once commanded.
def cat_imputation(column, value):
    train_nyclcddf.loc[train_nyclcddf[column].isnull(),column] = value

cat_imputation('Deaths', 445)
```

```
# converting object to int
train_nyclcddf['Deaths'] = train_nyclcddf.Deaths.astype(int)


#Double checking if any missing values are still present in the train_nyclcddf dataset
#no missing data
print(train_nyclcddf.isna().sum())
print()
#Double checking if data type values updated in the train_nyclcddf dataset.
#correct dtypes for each feature
print(train_nyclcddf.dtypes)
```

```
Year             0
LeadingCause     0
Sex              0
RaceEthnicity    0
Deaths           0
dtype: int64

Year              int64
LeadingCause     object
Sex              object
RaceEthnicity    object
Deaths            int64
dtype: object
```

```
# Descriptive statistics: minimum, mean, median, maximum, standard deviation of year a
print("Year min:", train_nyclcddf['Year'].min())
print("Year mean:",train_nyclcddf['Year'].mean())
print("Year median:", train_nyclcddf['Year'].median())
print("Year max:", train_nyclcddf['Year'].max())
print("Year STD:", train_nyclcddf['Year'].std())

print("")

print("Deaths min:", train_nyclcddf['Deaths'].min())
print("Deaths mean:",train_nyclcddf['Deaths'].mean())
print("Deaths median:", train_nyclcddf['Deaths'].median())
print("Deaths max:", train_nyclcddf['Deaths'].max())
print("Deaths STD:", train_nyclcddf['Deaths'].std())
```

```
Year min: 2007
Year mean: 2010.4771480804388
Year median: 2010.0
Year max: 2014
Year STD: 2.293418863657944

Deaths min: 5
Deaths mean: 444.6142595978062
Deaths median: 186.0
Deaths max: 7050
Deaths STD: 822.6734903625146
```

```
# Extracting the years recorded from the train_nyclcddf dataset
years_recorded = set(train_nyclcddf["Year"])
years_recorded
```

```
{2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014}
```

```
# Couting how many records are collected with that particular year in a sorted order
years_frequencies = train_nyclcddf["Year"].value_counts().sort_values(ascending =False
print(years_frequencies)
```

```
2007    141
2011    141
2010    138
2008    136
2014    136
2009    135
2012    134
2013    133
Name: Year, dtype: int64
```

```
#NOTE: selecting YEAR 2011 with highest recorded data with 141 rows and 5 columns
new_nyclcddf = train_nyclcddf[train_nyclcddf['Year']== 2011]
new_nyclcddf
```

	Year	LeadingCause	Sex	RaceEthnicity	Deaths
11	2011	Cerebrovascular Disease (Stroke: I60-I69)	F	Black Non-Hispanic	281
14	2011	Human Immunodeficiency Virus Disease (HIV: B20...	F	Hispanic	70
15	2011	Accidents Except Drug Posioning (V01-X39, X43,...	F	Other Race/ Ethnicity	445
20	2011	Alzheimer's Disease (G30)	F	Hispanic	90
30	2011	Cerebrovascular Disease (Stroke: I60-I69)	F	White Non-Hispanic	466
...	...	...	...	...	...
1028	2011	Accidents Except Drug Posioning (V01-X39, X43,...	M	Asian and Pacific Islander	61
1056	2011	Intentional Self-Harm (Suicide: X60-X84, Y87.0)	M	Other Race/ Ethnicity	5
1058	2011	Malignant Neoplasms (Cancer: C00-C97)	M	Black Non-Hispanic	1590
1060	2011	Malignant Neoplasms (Cancer: C00-C97)	M	Other Race/ Ethnicity	36

```
# Extracting the LeadingCause recorded values from the new_nyclcddf dataset
LeadingCause_recorded = set(new_nyclcddf["LeadingCause"])
LeadingCause_recorded
```

```
{'Accidents Except Drug Posioning (V01-X39, X43, X45-X59, Y85-Y86)',
```

```
 'All Other Causes',
 "Alzheimer's Disease (G30)",
 'Assault (Homicide: Y87.1, X85-Y09)',
 'Cerebrovascular Disease (Stroke: I60-I69)',
 'Certain Conditions originating in the Perinatal Period (P00-P96)',
 'Chronic Liver Disease and Cirrhosis (K70, K73)',
 'Chronic Lower Respiratory Diseases (J40-J47)',
 'Congenital Malformations, Deformations, and Chromosomal Abnormalities (Q00-Q99
 'Diabetes Mellitus (E10-E14)',
 'Diseases of Heart (I00-I09, I11, I13, I20-I51)',
 'Essential Hypertension and Renal Diseases (I10, I12)',
 'Human Immunodeficiency Virus Disease (HIV: B20-B24)',
 'Influenza (Flu) and Pneumonia (J09-J18)',
 'Intentional Self-Harm (Suicide: X60-X84, Y87.0)',
 'Malignant Neoplasms (Cancer: C00-C97)',
 'Mental and Behavioral Disorders due to Accidental Poisoning and Other Psychoac
 'Nephritis, Nephrotic Syndrome and Nephrisis (N00-N07, N17-N19, N25-N27)',
 "Parkinson's Disease (G20)",
 'Septicemia (A40-A41)',
 'Tuberculosis (A16-A19)',
 'Viral Hepatitis (B15-B19)'}
```

```
# Couting how many LeadingCause values are collected with 2011 data year in a sorted d
LeadingCause_frequencies = new_nyclcddf["LeadingCause"].value_counts().sort_values(asc
print(LeadingCause_frequencies)
```

```
All Other Causes
Influenza (Flu) and Pneumonia (J09-J18)
Chronic Lower Respiratory Diseases (J40-J47)
Diseases of Heart (I00-I09, I11, I13, I20-I51)
Diabetes Mellitus (E10-E14)
Malignant Neoplasms (Cancer: C00-C97)
Essential Hypertension and Renal Diseases (I10, I12)
Cerebrovascular Disease (Stroke: I60-I69)
Accidents Except Drug Posioning (V01-X39, X43, X45-X59, Y85-Y86)
Mental and Behavioral Disorders due to Accidental Poisoning and Other Psychoacti
Human Immunodeficiency Virus Disease (HIV: B20-B24)
Alzheimer's Disease (G30)
Intentional Self-Harm (Suicide: X60-X84, Y87.0)
Certain Conditions originating in the Perinatal Period (P00-P96)
Assault (Homicide: Y87.1, X85-Y09)
Chronic Liver Disease and Cirrhosis (K70, K73)
Nephritis, Nephrotic Syndrome and Nephrisis (N00-N07, N17-N19, N25-N27)
Parkinson's Disease (G20)
Congenital Malformations, Deformations, and Chromosomal Abnormalities (Q00-Q99)
Septicemia (A40-A41)
Viral Hepatitis (B15-B19)
Tuberculosis (A16-A19)
Name: LeadingCause, dtype: int64
```

NOTE: The top 5 leading causes of death in New York City recorded by "**New_York_City_Leading_Causes_of_Death**" dataset in 2011 for both Sex are

[**Malignant Neoplasms** (12),

Diabetes Mellitus(12),

Influenza (Flu) and Pneumonia (J09-J18)(12),

Chronic Lower Respiratory Diseases (J40-J47)(12), and

All Other Causes(12) along side with

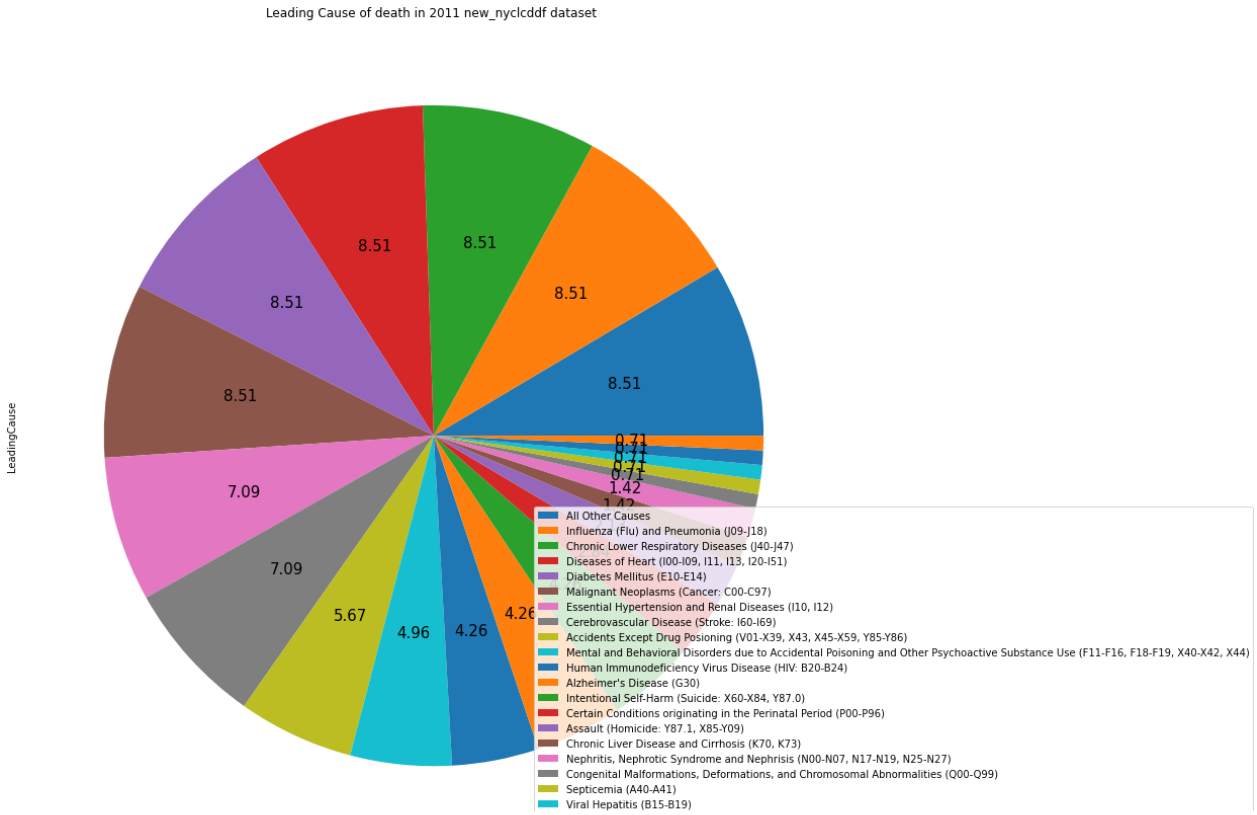
Diseases of Heart (I00-I09, I11, I13, I20-I51)(12)]

```
# Couting how many RaceEthnicity values are collected with 2011 data year in a sorted
RaceEthnicity_frequencies = new_nyclcddf["RaceEthnicity"].value_counts().sort_values(a
print(RaceEthnicity_frequencies)
```

```
Not Stated/Unknown            31
Asian and Pacific Islander    22
Other Race/ Ethnicity         22
Hispanic                      22
Black Non-Hispanic            22
White Non-Hispanic            22
Name: RaceEthnicity, dtype: int64
```

```
#creating a pie chart to visually see which are higher in cause of death
new_nyclcddf["LeadingCause"].value_counts().plot.pie(autopct='%.2f', figsize=(15, 15),
plt.legend(loc="lower right",bbox_to_anchor=(1, 0, 0.5, 1))
```

```
<matplotlib.legend.Legend at 0x7fd45dcc1f10>
```




```
# replacing sex genders from F = 0, and M = 1
new_nyclcddf['Sex'].replace(['F','M'],[0,1],inplace=True)
```

```
/Users/nieves/opt/anaconda3/lib/python3.8/site-packages/pandas/core/generic.py:6
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stab
  self._update_inplace(new_data)
```

```
# Counting the Sex gender collected with that particular year
# Notes more female then male
new_nyclcddf["Sex"].value_counts()
```

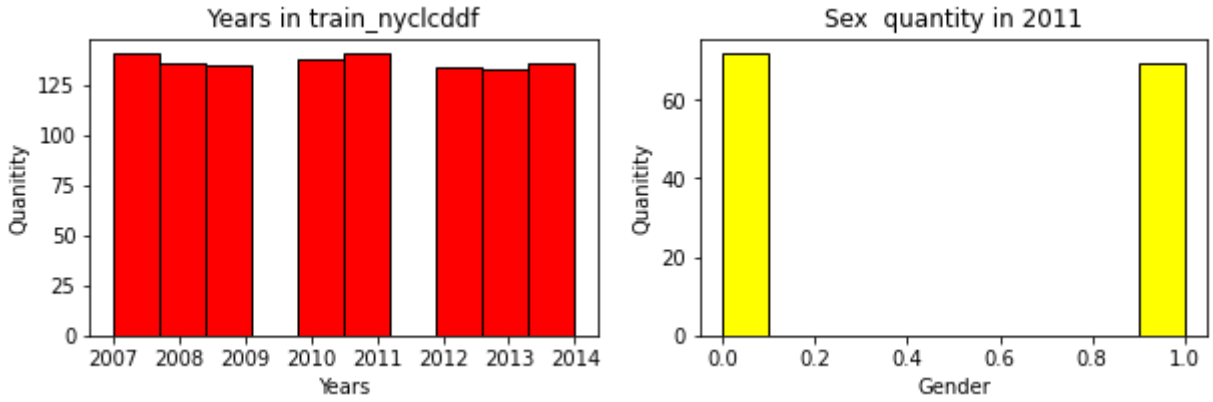
```
0    72
1    69
Name: Sex, dtype: int64
```

```
# Displaying the histogram (setting title, axis labels) on new_nyclcddf dataset year a
fig = plt.figure(figsize=(10, 6))

#histagram of Year with train_nyclcddf dataset
hist_Year = fig.add_subplot(2, 2, 1)
hist_Year.hist(train_nyclcddf['Year'], color = 'red', edgecolor='black')
hist_Year.set_title("Years in train_nyclcddf")
hist_Year.set_xlabel("Years")
hist_Year.set_ylabel("Quanitity")
```

```
#histagram of Sex with train_nyclcddf dataset
hist_Sex = fig.add_subplot(2, 2, 2)
hist_Sex.hist(new_nyclcddf['Sex'], color = 'yellow', edgecolor='black')
hist_Sex.set_title("Sex  quantity in 2011")
hist_Sex.set_xlabel("Gender")
hist_Sex.set_ylabel("Quanitity")
```

```
Text(0, 0.5, 'Quanitity')
```

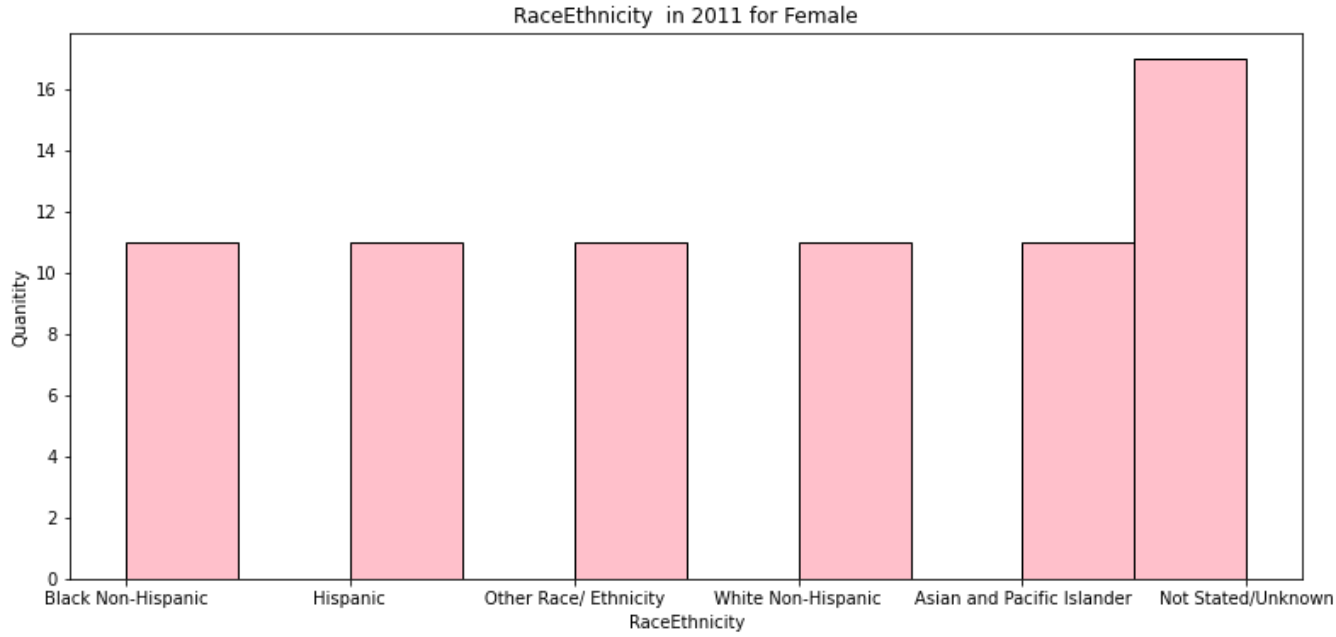


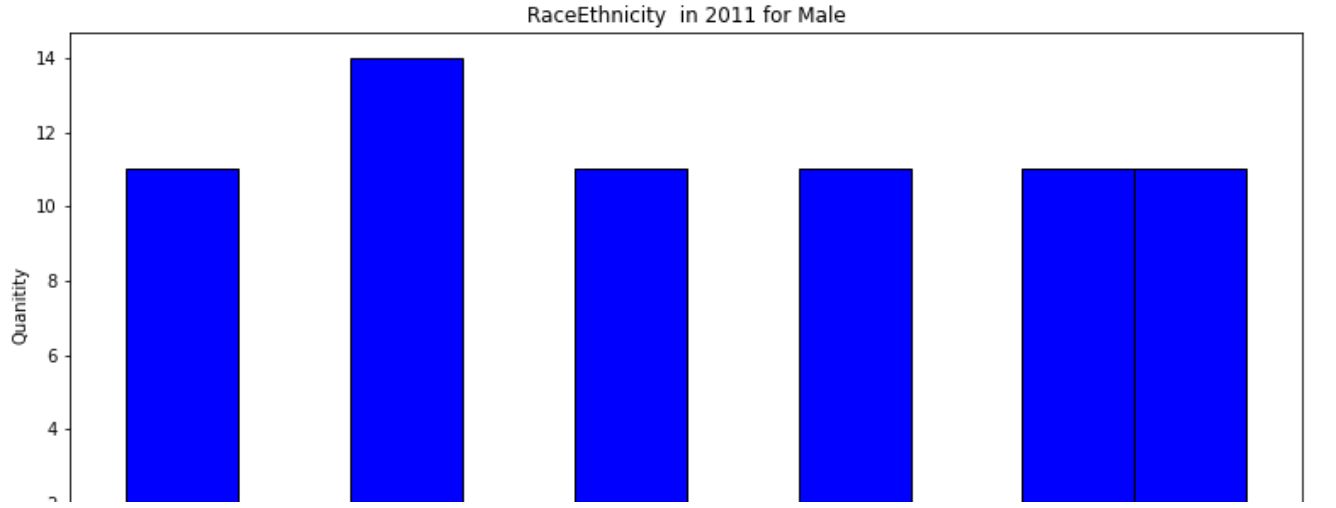

```
# Seperating females and males into a seperate set based on their value
F_nyclcddf = new_nyclcddf[new_nyclcddf['Sex']== 0]
M_nyclcddf = new_nyclcddf[new_nyclcddf['Sex']== 1]
```

```
# Displaying the histogram (setting title, axis labels) on new_nyclcddf dataset year a
fig = plt.figure(figsize=(13, 13))
#histagram of RaceEthnicity for female
hist_RaceEthnicity = fig.add_subplot(2, 1, 1)
hist_RaceEthnicity.hist(F_nyclcddf['RaceEthnicity'], color = 'pink', edgecolor='black
hist_RaceEthnicity.set_title("RaceEthnicity  in 2011 for Female")
hist_RaceEthnicity.set_xlabel("RaceEthnicity")
hist_RaceEthnicity.set_ylabel("Quanitity")

#histagram of RaceEthnicity for male
hist_RaceEthnicity = fig.add_subplot(2, 1, 2)
hist_RaceEthnicity.hist(M_nyclcddf['RaceEthnicity'], color = 'blue', edgecolor='black
hist_RaceEthnicity.set_title("RaceEthnicity  in 2011 for Male")
hist_RaceEthnicity.set_xlabel("RaceEthnicity")
hist_RaceEthnicity.set_ylabel("Quanitity")
```

```
Text(0, 0.5, 'Quanitity')
```

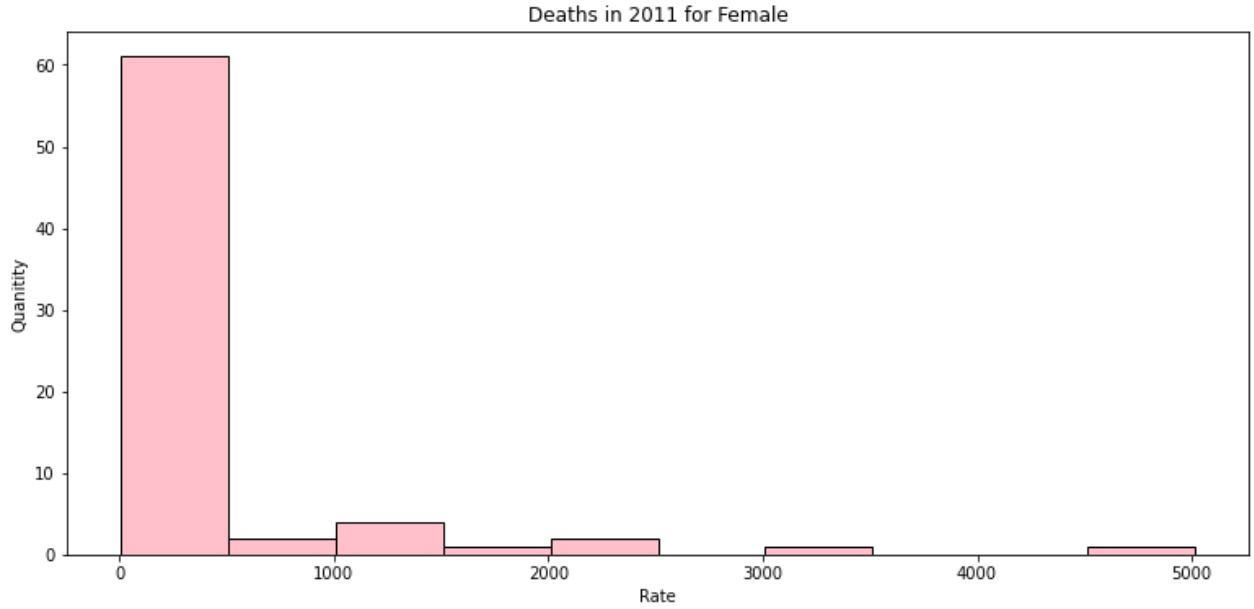


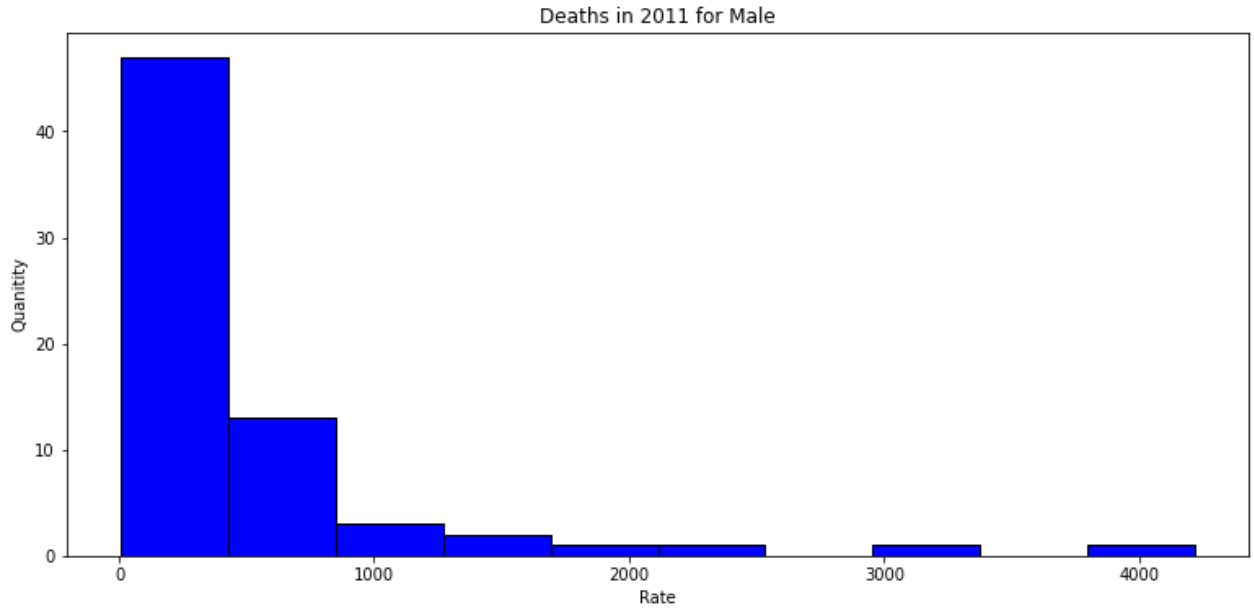



```
fig = plt.figure(figsize=(13, 13))
#histagram of Deaths for female
hist_Deaths = fig.add_subplot(2, 1, 1)
hist_Deaths.hist(F_nyclcddf['Deaths'], color = 'pink', edgecolor='black')
hist_Deaths.set_title("Deaths in 2011 for Female")
hist_Deaths.set_xlabel("Rate")
hist_Deaths.set_ylabel("Quanitity")

#histagram of Deaths for male
hist_Deaths = fig.add_subplot(2, 1, 2)
hist_Deaths.hist(M_nyclcddf['Deaths'], color = 'blue', edgecolor='black')
hist_Deaths.set_title("Deaths in 2011 for Male")
hist_Deaths.set_xlabel("Rate")
hist_Deaths.set_ylabel("Quanitity")
```

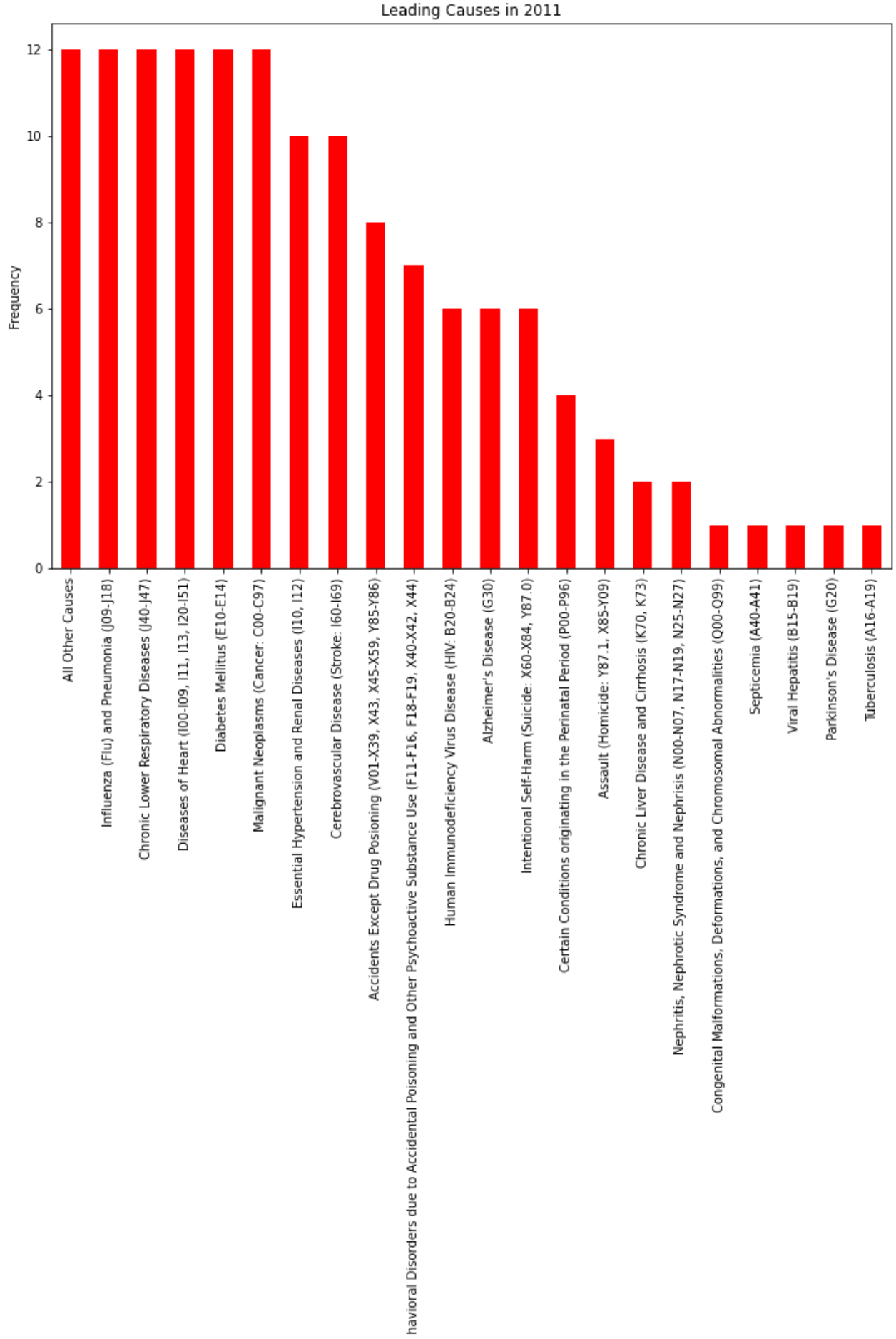
```
Text(0, 0.5, 'Quanitity')
```






```
# Displaying the barchart (setting title, axis labels) on new_nyclcddf dataset
fig = plt.figure(figsize=(15, 30))
plt.figure(figsize=(12, 8))
bar_LeadingCause = new_nyclcddf["LeadingCause"].value_counts().plot(kind='bar',color=
bar_LeadingCause.set_title('Leading Causes in 2011')
bar_LeadingCause.set_xlabel('Cause (s)')
bar_LeadingCause.set_ylabel('Frequency')
```

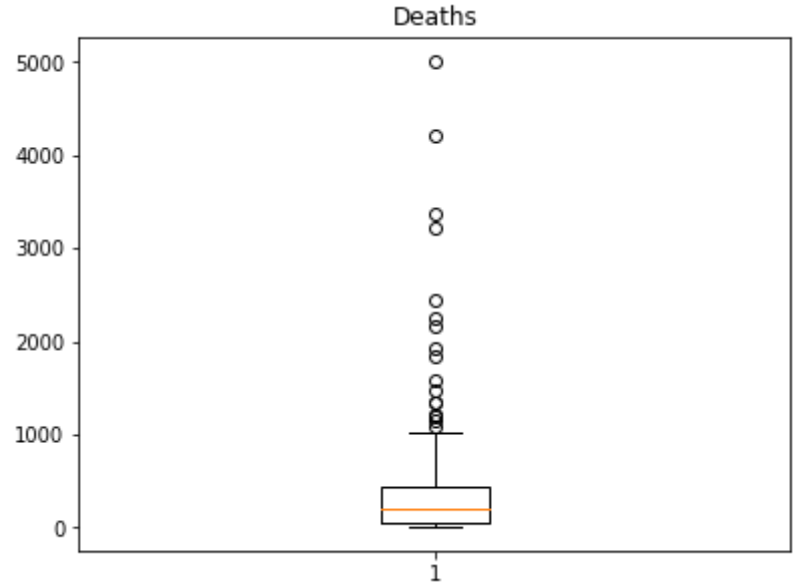
```
Text(0, 0.5, 'Frequency')
<Figure size 1080x2160 with 0 Axes>
```



```
# Using boxplot to see the outliers in a Deaths feature. NOTE: Do not drop more then 1
fig = plt.figure(figsize=(14, 16))

#boxplot of Deaths
boxplot_Deaths = fig.add_subplot(3, 2, 1)
boxplot_Deaths.boxplot(new_nyclcddf['Deaths'])
boxplot_Deaths.set_title("Deaths")
```

```
Text(0.5, 1.0, 'Deaths')
```




```
# Getting the statictis of deaths column to remove 1% of outliers.
new_nyclcddf['Deaths'].describe()
```

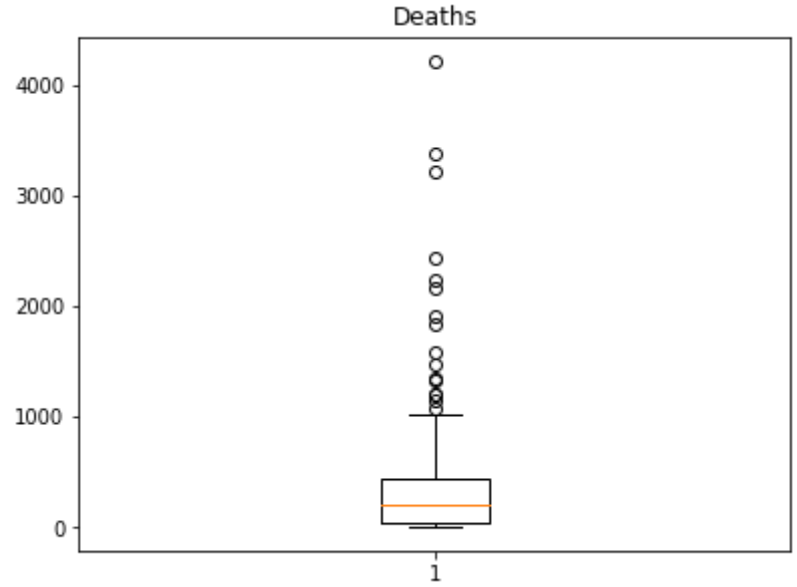
```
count     141.000000
mean      449.687943
std       769.872792
min         5.000000
25%        49.000000
50%       200.000000
75%       445.000000
max      5016.000000
Name: Deaths, dtype: float64
```

```
#GrLivArea has a min of 5.000000 with a 25% of 49.000000
#1% of 5016.000000(max) = 50.16
new_nyclcddf['Deaths'] = np.where(new_nyclcddf['Deaths'] > 4965.84, 5.000000, new_nycl
```

```
# replotting the boxplot to see if drop worked succesfully
fig = plt.figure(figsize=(14, 16))
boxplot_Deaths = fig.add_subplot(3, 2, 1)
boxplot_Deaths.boxplot(new_nyclcddf['Deaths'])
boxplot_Deaths.set_title("Deaths")
```

```
<ipython-input-38-de1cbfe1bb28>:3: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stab
  new_nyclcddf['Deaths'] = np.where(new_nyclcddf['Deaths'] > 4965.84, 5.000000,
Text(0.5, 1.0, 'Deaths')
```



Unsupported Cell Type. Double-Click to inspect/edit the content.

Gathered enough information to evaulate which gender has been more effected by a particular leadingcause of death

```
# Counting how many female deaths were caused by the top 5 LeadingCause
F_LeadingCause_frequencies = F_nyclcddf["LeadingCause"].value_counts().sort_values(asc
F_LeadingCause_frequencies.head(10)
```

```
Diabetes Mellitus (E10-E14)                          6
Chronic Lower Respiratory Diseases (J40-J47)         6
All Other Causes                                     6
Alzheimer's Disease (G30)                            6
Influenza (Flu) and Pneumonia (J09-J18)              6
Essential Hypertension and Renal Diseases (I10, I12) 6
Diseases of Heart (I00-I09, I11, I13, I20-I51)       6
Malignant Neoplasms (Cancer: C00-C97)                6
Cerebrovascular Disease (Stroke: I60-I69)            6
Human Immunodeficiency Virus Disease (HIV: B20-B24)  3
Name: LeadingCause, dtype: int64
```

```
F_LeadingCause_death_by_RE = F_nyclcddf.sort_values(ascending =False, by=['Deaths'])
F_LeadingCause_death_by_RE.head(10)
```

	Year	LeadingCause	Sex	RaceEthnicity	Deaths
170	2011	Diseases of Heart (I00-I09, I11, I13, I20-I51)	0	White Non-Hispanic	5016
201	2011	Malignant Neoplasms (Cancer: C00-C97)	0	White Non-Hispanic	3371
649	2011	All Other Causes	0	White Non-Hispanic	2445
557	2011	Diseases of Heart (I00-I09, I11, I13, I20-I51)	0	Black Non-Hispanic	2243
801	2011	Malignant Neoplasms (Cancer: C00-C97)	0	Black Non-Hispanic	1918
568	2011	All Other Causes	0	Black Non-Hispanic	1473
860	2011	Diseases of Heart (I00-I09, I11, I13, I20-I51)	0	Hispanic	1348
832	2011	Malignant Neoplasms (Cancer: C00-C97)	0	Hispanic	1085
478	2011	All Other Causes	0	Hispanic	1025
588	2011	Influenza (Flu) and Pneumonia (J09-J18)	0	White Non-Hispanic	696

```
# Counting how many male deaths were caused by the top 5 LeadingCause
M_LeadingCause_frequencies = M_nyclcddf["LeadingCause"].value_counts().sort_values(asc
M_LeadingCause_frequencies.head(10)
```

```
Diabetes Mellitus (E10-E14)
All Other Causes
Influenza (Flu) and Pneumonia (J09-J18)
Chronic Lower Respiratory Diseases (J40-J47)
Diseases of Heart (I00-I09, I11, I13, I20-I51)
Malignant Neoplasms (Cancer: C00-C97)
Accidents Except Drug Posioning (V01-X39, X43, X45-X59, Y85-Y86)
Essential Hypertension and Renal Diseases (I10, I12)
Mental and Behavioral Disorders due to Accidental Poisoning and Other Psychoacti
Cerebrovascular Disease (Stroke: I60-I69)
Name: LeadingCause, dtype: int64
```

```
M_LeadingCause_death_by_RE = M_nyclcddf.sort_values(ascending =False, by=['Deaths'])
M_LeadingCause_death_by_RE.head(10)
```

	Year	LeadingCause	Sex	RaceEthnicity	Deaths
439	2011	Diseases of Heart (I00-I09, I11, I13, I20-I51)	1	White Non-Hispanic	4220
33	2011	Malignant Neoplasms (Cancer: C00-C97)	1	White Non-Hispanic	3222
188	2011	All Other Causes	1	White Non-Hispanic	2165
821	2011	Diseases of Heart (I00-I09, I11, I13, I20-I51)	1	Black Non-Hispanic	1840
1058	2011	Malignant Neoplasms (Cancer: C00-C97)	1	Black Non-Hispanic	1590
1001	2011	All Other Causes	1	Hispanic	1207

The aim of this project is to analyze what are the five top causes of death from a particular year and which were the race ethnicity and sex that had deceased from that particular case recorded.

It can be concluded that based on the information from New_York_City_Leading_Causes_of_Death dataset provided by (https://opendata.cityofnewyork.us/) that after selecting the most populated data year 2011, both men and women leading cause of death were similar. There were slightly more women recorded compared to men.

The top five leading cause of death for women were from Malignant Neoplasms, Diseases of Heart, Diabetes Mellitus, and Alzheimer's Disease. Records indicate that more White Non-Hispanic woman had died from Malignant Neoplasms. The second descending are Black Non-Hispanic women who died from Diseases of Heart and Malignant Neoplasms. The third descending are Hispanic with the same death related causes as both Non-Hispanic and Black Non-Hispanic.

The top five leading cause of death for men were Malignant Neoplasms, Influenza (Flu) and Pneumonia, All Other Causes, Diabetes Mellitus and Chronic Lower Respiratory Diseases. Records indicated for the base of men that more deaths are recorded for White Non-Hispanic men who passed from Diseases of Heart, and Malignant Neoplasms. The second descending records Black Non-Hispanic men who died from Diseases of Heart, and Malignant Neoplasms, and the third descending for Hispanic men.