US Suicide dataset

This dataset was obtained from Data.gov

In [1]:
```
!date
```

```
Wed 17 May 2023 07:27:17 PM EDT
```

In [2]:
```
from IPython import display
display.Image('/home/harlohutch77/python/python_master/mental_health.jpg')
```

Out[2]:

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

In [2]:
```
S = pd.read_csv('/home/harlohutch77/python/python_master/Suicide__US.csv')
```

In [3]:
```
#6390 entries, 13 columns
S.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6390 entries, 0 to 6389
Data columns (total 13 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   INDICATOR       6390 non-null   object
 1   UNIT            6390 non-null   object
 2   UNIT_NUM        6390 non-null   int64
 3   STUB_NAME       6390 non-null   object
 4   STUB_NAME_NUM   6390 non-null   int64
 5   STUB_LABEL      6390 non-null   object
 6   STUB_LABEL_NUM  6390 non-null   float64
 7   YEAR            6390 non-null   int64
 8   YEAR_NUM        6390 non-null   int64
 9   AGE             6390 non-null   object
 10  AGE_NUM         6390 non-null   float64
 11  ESTIMATE        5484 non-null   float64
 12  FLAG            906 non-null    object
dtypes: float64(3), int64(4), object(6)
memory usage: 649.1+ KB
```

In [4]:
```
S.describe().transpose()
```

Out[4]:

	count	mean	std	min	25%	50%	75%	max
UNIT_NUM	6390.0	1.872926	0.333081	1.0	2.00	2.0000	2.000	2.000
STUB_NAME_NUM	6390.0	4.621909	2.031777	0.0	3.00	5.0000	6.000	11.000
STUB_LABEL_NUM	6390.0	4.686775	1.829624	0.0	3.23	5.1251	6.153	7.235
YEAR	6390.0	1997.525822	14.937451	1950.0	1988.00	1999.0000	2009.000	2018.000
YEAR_NUM	6390.0	22.692019	12.286033	1.0	12.00	23.0000	33.000	42.000
AGE_NUM	6390.0	2.712207	1.932280	0.0	0.00	3.0000	4.100	6.000
ESTIMATE	5484.0	13.709810	11.531805	0.3	5.00	10.5000	19.500	74.800

In [5]:
```
S.fillna(0)
```

Out[5]:

	INDICATOR	UNIT	UNIT_NUM	STUB_NAME	STUB_NAME_NUM	STUB_LABEL	STUB_LABEL_NUM	YEAR	YEAR_NUM	AGE	A
0	Death rates for suicide	Deaths per 100,000 resident population, age-ad...	1	Total	0	All persons	0.000	1950	1	All ages	
1	Death rates for suicide	Deaths per 100,000 resident population, age-ad...	1	Total	0	All persons	0.000	1960	2	All ages	
2	Death rates for suicide	Deaths per 100,000 resident population, age-ad...	1	Total	0	All persons	0.000	1970	3	All ages	
3	Death rates for suicide	Deaths per 100,000 resident population, age-ad...	1	Total	0	All persons	0.000	1980	4	All ages	
4	Death rates for suicide	Deaths per 100,000 resident population, age-ad...	1	Total	0	All persons	0.000	1981	5	All ages	
...	...	...	...	...	...	...	...	...	...	...	
6385	Death rates for suicide	Deaths per 100,000 resident population, crude	2	Sex, age and race and Hispanic origin (Single ...	11	Female: Not Hispanic or Latino: Black or Afric...	7.225	2018	42	65 years and over	
6386	Death rates for suicide	Deaths per 100,000 resident population, crude	2	Sex, age and race and Hispanic origin (Single ...	11	Female: Hispanic or Latino: All races: 15-24 y...	7.232	2018	42	15-24 years	
6387	Death rates for suicide	Deaths per 100,000 resident population, crude	2	Sex, age and race and Hispanic origin (Single ...	11	Female: Hispanic or Latino: All races: 25-44 y...	7.233	2018	42	25-44 years	
6388	Death rates for suicide	Deaths per 100,000 resident population, crude	2	Sex, age and race and Hispanic origin (Single ...	11	Female: Hispanic or Latino: All races: 45-64 y...	7.234	2018	42	45-64 years	
6389	Death rates for suicide	Deaths per 100,000 resident population, crude	2	Sex, age and race and Hispanic origin (Single ...	11	Female: Hispanic or Latino: All races: 65 year...	7.235	2018	42	65 years and over	

6390 rows × 13 columns

In [6]:
```
S.head(4)
```

Out[6]:

	INDICATOR	UNIT	UNIT_NUM	STUB_NAME	STUB_NAME_NUM	STUB_LABEL	STUB_LABEL_NUM	YEAR	YEAR_NUM	AGE	AGE_
0	Death rates for suicide	Deaths per 100,000 resident population, age-ad...	1	Total	0	All persons	0.0	1950	1	All ages	
1	Death rates for suicide	Deaths per 100,000 resident population, age-ad...	1	Total	0	All persons	0.0	1960	2	All ages	
2	Death rates for suicide	Deaths per 100,000 resident population, age-ad...	1	Total	0	All persons	0.0	1970	3	All ages	
3	Death rates for suicide	Deaths per 100,000 resident population, age-ad...	1	Total	0	All persons	0.0	1980	4	All ages	

In [7]:
```
S.tail(4).fillna(0)
```

Out[7]:

	INDICATOR	UNIT	UNIT_NUM	STUB_NAME	STUB_NAME_NUM	STUB_LABEL	STUB_LABEL_NUM	YEAR	YEAR_NUM	AGE	A
6386	Death rates for suicide	Deaths per 100,000 resident population, crude	2	Sex, age and race and Hispanic origin (Single ...	11	Female: Hispanic or Latino: All races: 15-24 y...	7.232	2018	42	15-24 years	
6387	Death rates for suicide	Deaths per 100,000 resident population, crude	2	Sex, age and race and Hispanic origin (Single ...	11	Female: Hispanic or Latino: All races: 25-44 y...	7.233	2018	42	25-44 years	
6388	Death rates for suicide	Deaths per 100,000 resident population, crude	2	Sex, age and race and Hispanic origin (Single ...	11	Female: Hispanic or Latino: All races: 45-64 y...	7.234	2018	42	45-64 years	
6389	Death rates for suicide	Deaths per 100,000 resident population, crude	2	Sex, age and race and Hispanic origin (Single ...	11	Female: Hispanic or Latino: All races: 65 year...	7.235	2018	42	65 years and over	

In [8]:
```
S[S['YEAR_NUM']>40].count()
```

Out[8]:
```
INDICATOR         438
UNIT              438
UNIT_NUM          438
STUB_NAME         438
STUB_NAME_NUM     438
STUB_LABEL        438
STUB_LABEL_NUM    438
YEAR              438
YEAR_NUM          438
AGE               438
AGE_NUM           438
ESTIMATE          418
FLAG               20
dtype: int64
```

In [9]:
```
S[S['YEAR_NUM']>30]['YEAR'].min()
```

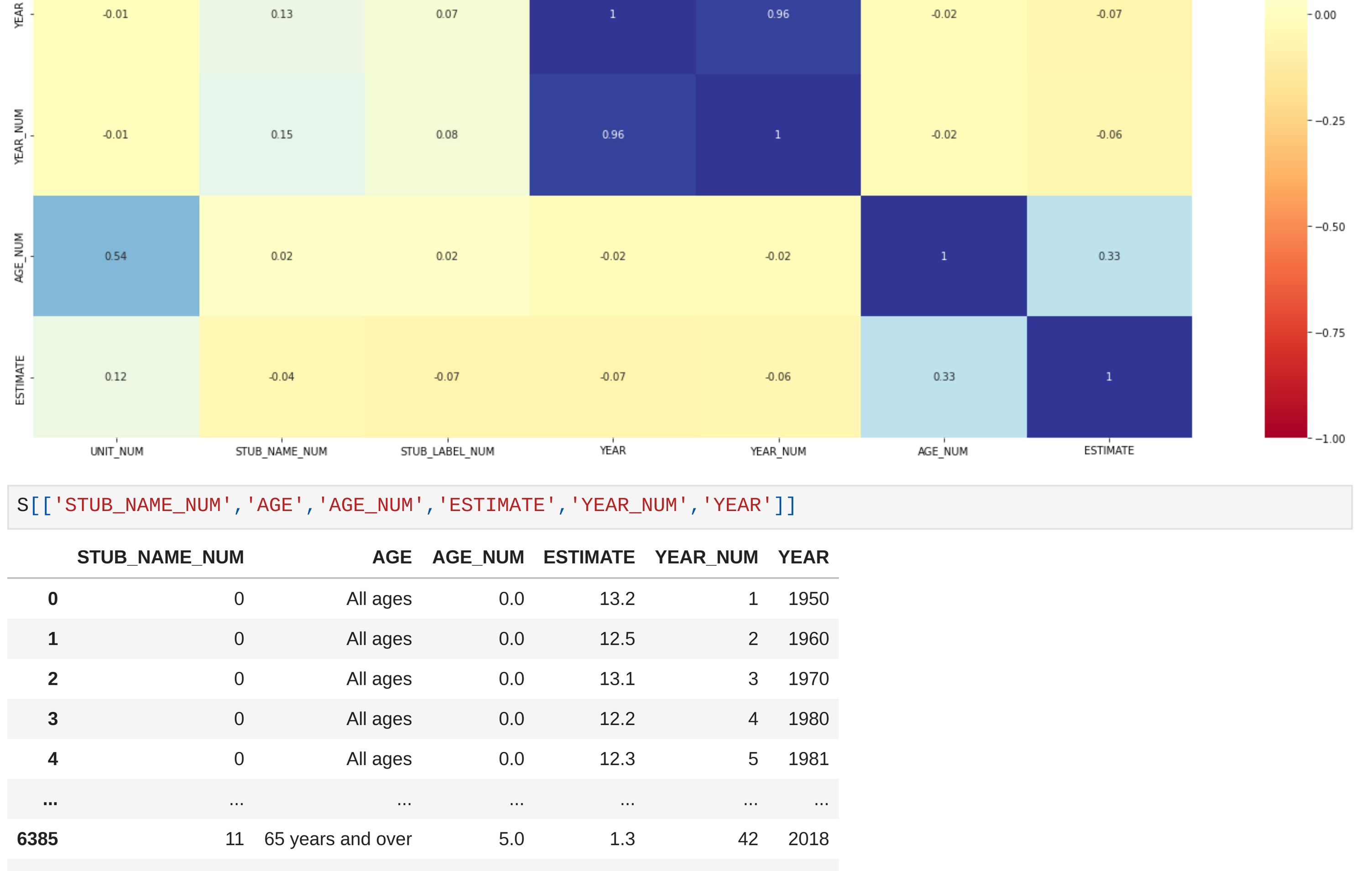
Out[9]:
```
2007
```

In [10]:
```
# Seaborn plt.subplot to use a heatmap visualization
fig, ax = plt.subplots(figsize = (25,15))
corr = S.corr().round(2)
sns.heatmap(corr, annot = True, ax=ax, vmin=-1, center=0, cmap='RdYlBu')
```

```
/tmp/ipykernel_2225/282403844.py:3: FutureWarning:

The default value of numeric_only in DataFrame.corr is deprecated. In a future version, it will default to Fals
e. Select only valid columns or specify the value of numeric_only to silence this warning.
```

Out[10]:
```
<AxesSubplot:>
```



In [14]:
```
S[['STUB_NAME_NUM','AGE','AGE_NUM','ESTIMATE','YEAR_NUM','YEAR']]
```

Out[14]:

	STUB_NAME_NUM	AGE	AGE_NUM	ESTIMATE	YEAR_NUM	YEAR
0	0	All ages	0.0	13.2	1	1950
1	0	All ages	0.0	12.5	2	1960
2	0	All ages	0.0	13.1	3	1970
3	0	All ages	0.0	12.2	4	1980
4	0	All ages	0.0	12.3	5	1981
...	...	...	...	...	...	...
6385	11	65 years and over	5.0	1.3	42	2018
6386	11	15-24 years	2.0	4.1	42	2018
6387	11	25-44 years	3.0	4.4	42	2018
6388	11	45-64 years	4.0	3.2	42	2018
6389	11	65 years and over	5.0	1.6	42	2018

6390 rows × 6 columns

In [11]:
```
S[S['YEAR']==1950]['ESTIMATE'].dropna().head(8)
```

Out[11]:
```
0      13.2
42     21.2
84      5.6
126    22.3
168     7.5
294     6.0
336     1.8
814    11.4
Name: ESTIMATE, dtype: float64
```

In [12]:
```
S[S['YEAR']==2018]['ESTIMATE'].dropna().head(8)
```

Out[12]:
```
41     14.2
83     22.8
125     6.2
167    25.5
209    11.6
251    20.9
293    10.4
335     7.0
Name: ESTIMATE, dtype: float64
```

In [13]:
```
#adding all the esitmates from 1950 to 2018 and dividing them by 6390 to give us an over all estimate
S['ESTIMATE'].sum()/6390
```

Out[13]:
```
11.765978090766824
```

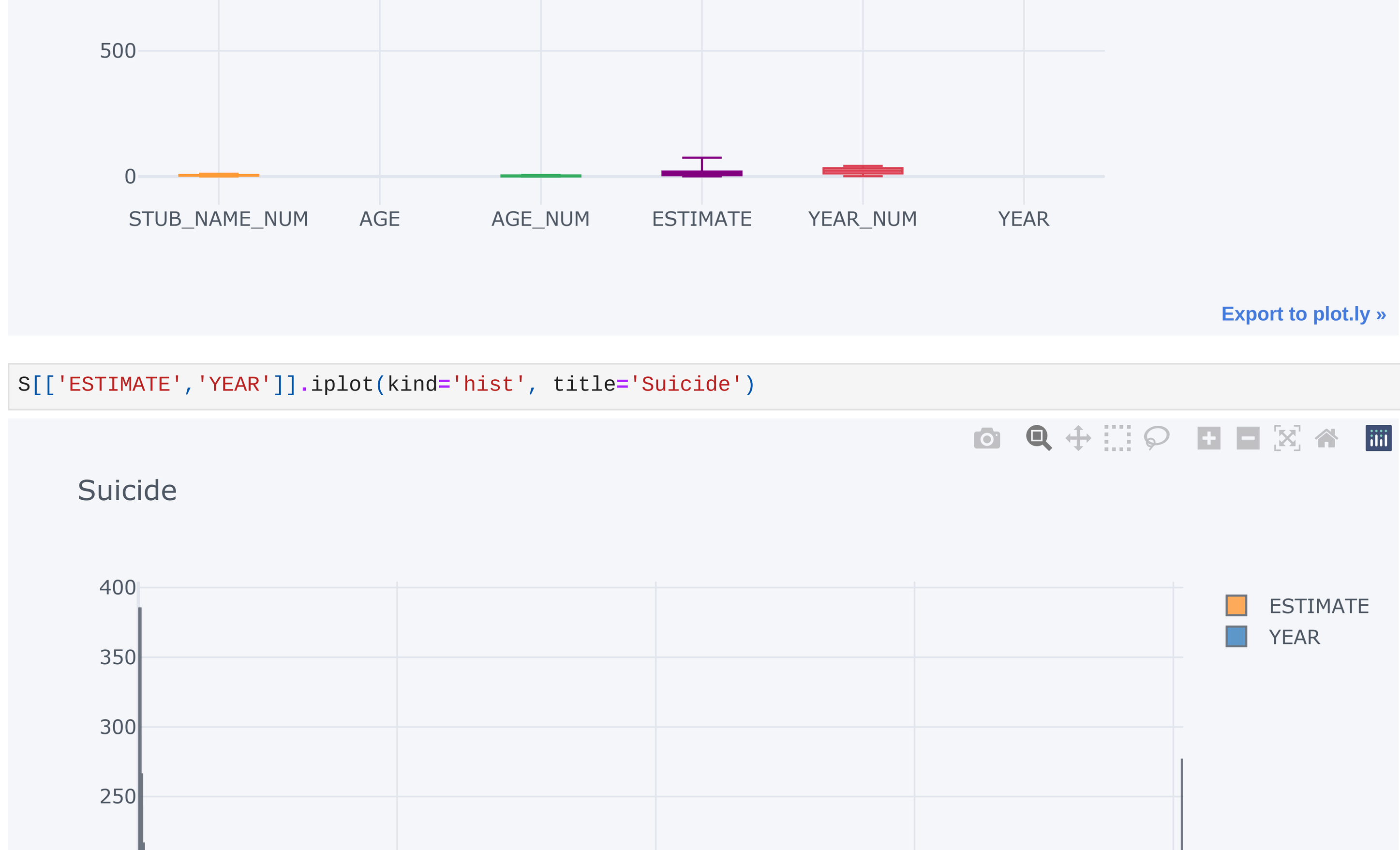
In [17]:
```
#cufflinks visualizations
from plotly import __version__
from plotly.offline import download_plotlyjs, init_notebook_mode, plot, iplot

print(__version__) #requires version >= 1.9.0

init_notebook_mode(connected=True)
cf.go_offline()
```

```
5.7.0
```

In [19]:
```
S[['STUB_NAME_NUM','AGE','AGE_NUM','ESTIMATE','YEAR_NUM','YEAR']].iplot(kind='box', title='Suicide')
```



In [20]:
```
S[['ESTIMATE','YEAR']].iplot(kind='hist', title='Suicide')
```

In [21]:
```
S[['ESTIMATE','YEAR']].iplot(kind='box', title='Suicide')
```