Final Exam Project : BANA 680

Introduction:

Purpose of the project is to implement various data managemnet approach like data cleaning,manipulation, EDA with pandas. We used 2 datasets 1st NCHS dataset on leading causes of death in USA from 1999 to 2016 & 2nd nst-est showing annual estimates of REsident population for USA regios, states & Perto Rico: April 12010 to July 2018. I decided to do some further research on factors affecting heart stroke in USA and used dataset from Kaggle.

In [55]:

```python
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as px
import plotly.graph_objects as go
from plotly.subplots import make_subplots
import os
import warnings
warnings.filterwarnings("ignore")
```

In [56]:

```python
df1 = pd.read_csv('C:/Users/Palam/Downloads/NCHS_-_Leading_Causes_of_Death__United_States.
df2 = pd.read_excel('C:/Users/Palam/Downloads/nst-est2018-01.xlsx')
df3= pd.read_csv('C:/Users/Palam/Downloads/healthcare-dataset-stroke-data.csv')
```

Q1. What are the leading causes of Death ?

In [57]:

```python
df_UScountry= df1[df1['State'].isin(['United States'])]
```

In [58]:

```python
TTDCUS = df_UScountry.groupby('Cause Name').agg({'Deaths':sum})['Deaths'].nlargest(4).rese
TTDCUS
```

Out[58]:

	Cause Name	Deaths
0	All causes	44915066
1	Heart disease	11575183
2	Cancer	10244536
3	Stroke	2580140

New Additional Question: What are the factors resposible for heart stroke?

About the columns in the dataset:

id: unique identifier
gender: "Male", "Female" or "Other"
age: age of the patient
hypertension: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension
heart_disease: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
ever_married: "No" or "Yes"
work_type: "children", "Govt_jov", "Never_worked", "Private" or "Self-employed"
Loading [MathJax]/extensions/Safe.js "Rural" or "Urban"

avg_glucose_level: average glucose level in blood

bmi: body mass index

smoking_status: "formerly smoked", "never smoked", "smokes" or "Unknown"

stroke: 1 if the patient had a stroke or 0 if not Note: "Unknown" in smoking_status means that the information is unavailable for this patient

In [59]:

```
for col in ['hypertension', 'heart_disease', 'stroke']:
    df3[col] = df3[col].apply(lambda x: {0:'No', 1:'Yes'}[x])
```

In [60]:

```
df3 = df3[df3['bmi'].notna()]
```

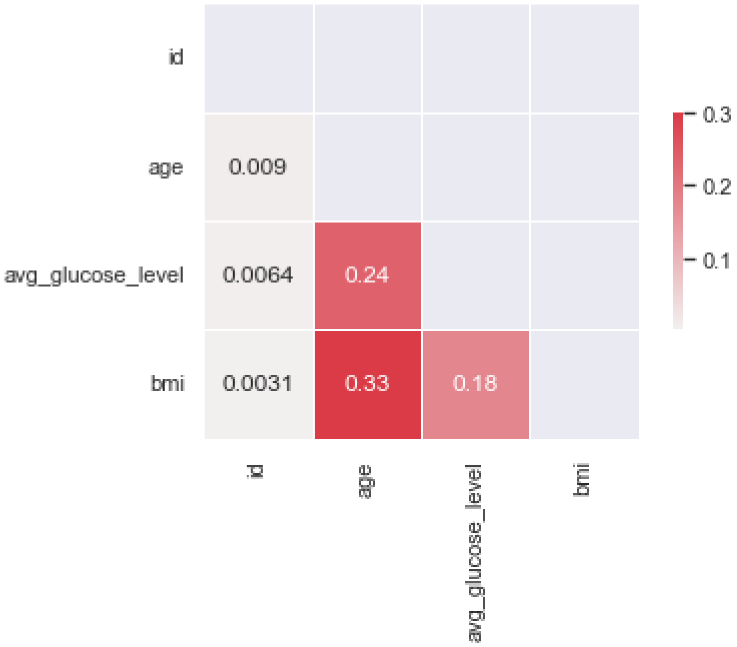
In [62]:

```
correl = df3.corr()
cmap = sns.diverging_palette(220, 10, as_cmap = True)

mask = np.zeros_like(correl, dtype = np.bool)
mask[np.triu_indices_from(mask)] = True

f, ax = plt.subplots(figsize = (6,4))

sns.heatmap(correl, mask = mask, cmap = cmap, vmax = 0.3, center = 0,annot = True, square
```



In [63]:

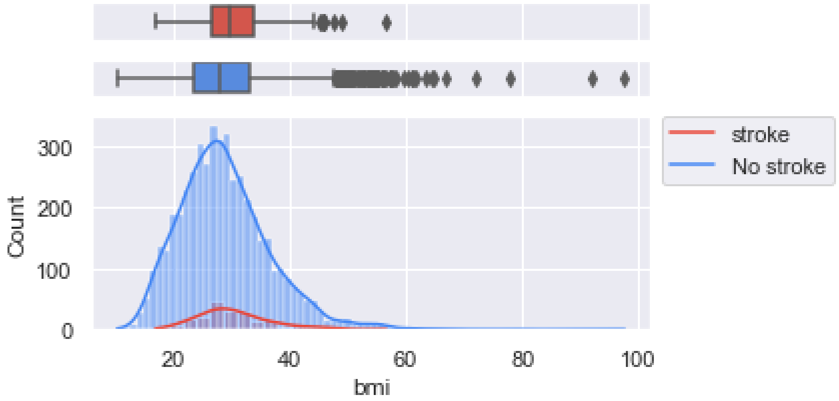
```
sns.set(style="darkgrid")
sns.set(rc={'figure.figsize':(5,3)})

# creating a figure composed of 3 matplotlib.Axes objects
f, (ax_box1, ax_box2, ax_hist) = plt.subplots(3, sharex=True, gridspec_kw={"height_ratios'
# assigning a graph to each ax
sns.boxplot(x=df3[df3['stroke']=='Yes']["bmi"], ax=ax_box1, color="#ea4335")
sns.histplot(df3[df3['stroke']=='Yes'], x="bmi", ax=ax_hist, kde=True, color="#ea4335")

sns.boxplot(x=df3[df3['stroke']=='No']["bmi"], ax=ax_box2, color='#4285f4')
sns.histplot(df3[df3['stroke']=='No'], x="bmi", ax=ax_hist, kde=True, color='#4285f4')

# Remove x axis name for the boxplots
ax_box1.set(xlabel='')
ax_box2.set(xlabel='')
```

```
plt.legend(title='', loc=2, labels=['stroke', 'No stroke'],bbox_to_anchor=(1.02, 1), borde
plt.show()
```



From the above visualization we can People who had a stroke have a slightly higher bmi than those who never had a stroke. According to a research AHAjournal the risk of stroke increases by 5% for every 1 unit increase in BMI (7 pounds for a human of average height), and the risk appears to be practically linear starting at a still-normal BMI of 20 kg/m2.

In [64]:

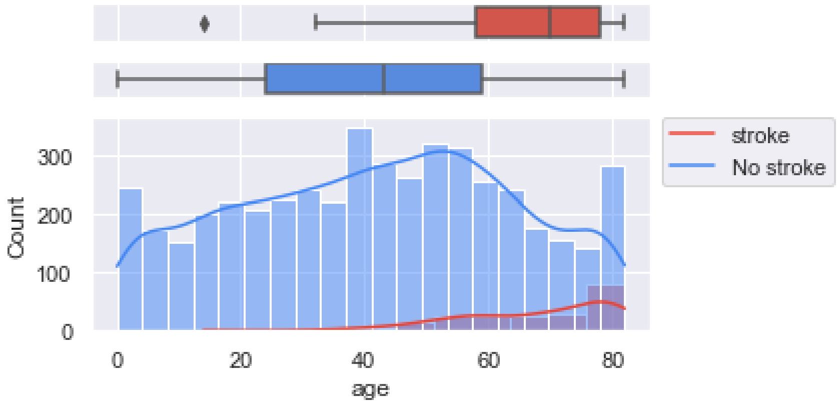
```
sns.set(style="darkgrid")
sns.set(rc={'figure.figsize':(5,3)})

# creating a figure composed of 3 matplotlib.Axes objects
f, (ax_box1, ax_box2, ax_hist) = plt.subplots(3, sharex=True, gridspec_kw={"height_ratios'
# assigning a graph to each ax
sns.boxplot(x=df3[df3['stroke']=='Yes']["age"], ax=ax_box1, color="#ea4335")
sns.histplot(df3[df3['stroke']=='Yes'], x="age", ax=ax_hist, kde=True, color="#ea4335")

sns.boxplot(x=df3[df3['stroke']=='No']["age"], ax=ax_box2, color='#4285f4')
sns.histplot(df3[df3['stroke']=='No'], x="age", ax=ax_hist, kde=True, color='#4285f4')

# Remove x axis name for the boxplots
ax_box1.set(xlabel='')
ax_box2.set(xlabel='')

plt.legend(title='', loc=2, labels=['stroke', 'No stroke'],bbox_to_anchor=(1.02, 1), borde
plt.show()
```



From the above visualization we can say that people having a stroke are mostly old or elderly people

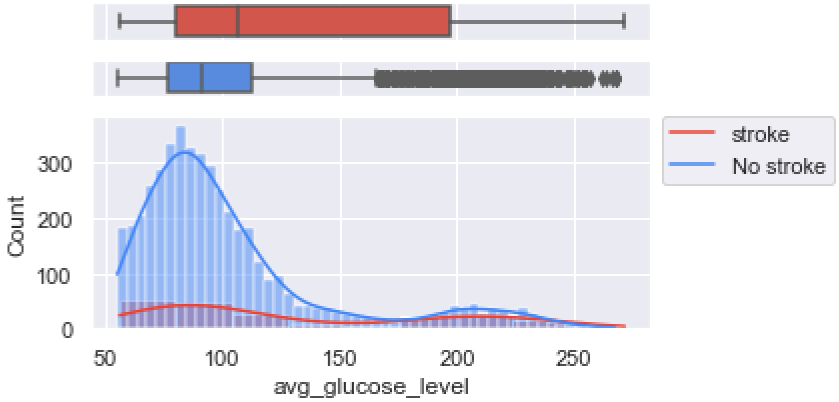
In [65]:

```
sns.set(style="darkgrid")
sns.set(rc={'figure.figsize':(5,3)})

# creating a figure composed of 3 matplotlib.Axes objects
f, (ax_box1, ax_box2, ax_hist) = plt.subplots(3, sharex=True, gridspec_kw={"height_ratios'
# assigning a graph to each ax
sns.boxplot(x=df3[df3['stroke']=='Yes']["avg_glucose_level"], ax=ax_box1, color="#ea4335"
sns.histplot(df3[df3['stroke']=='Yes'], x="avg_glucose_level", ax=ax_hist, kde=True, color

sns.boxplot(x=df3[df3['stroke']=='No']["avg_glucose_level"], ax=ax_box2, color='#4285f4')
sns.histplot(df3[df3['stroke']=='No'], x="avg_glucose_level", ax=ax_hist, kde=True, color=
ax_box1.set(xlabel='')
ax_box2.set(xlabel='')

plt.legend(title='', loc=2, labels=['stroke', 'No stroke'],bbox_to_anchor=(1.02, 1), borde
plt.show()
```



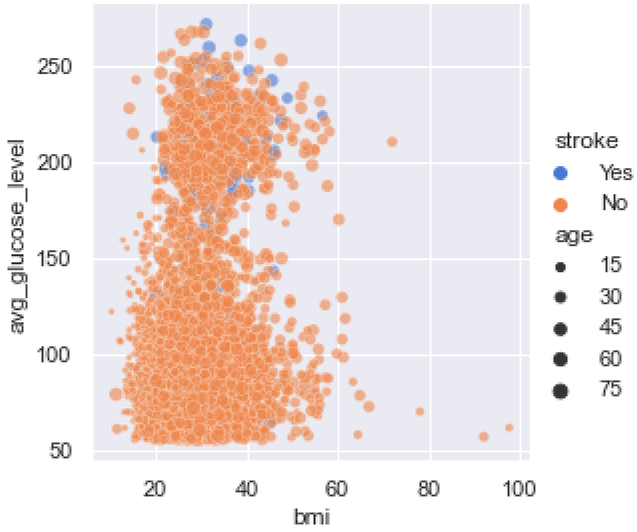
So we can say that The majority of persons who have a stroke have a higher average glucose level than people who have never had a stroke.

In [66]:

```
sns.relplot(x="bmi", y="avg_glucose_level", hue="stroke", size="age",
            sizes=(10, 50), alpha=0.6, palette="muted",
            height=4, data=df3)
```

Out[66]:

```
<seaborn.axisgrid.FacetGrid at 0x218e0d7d040>
```

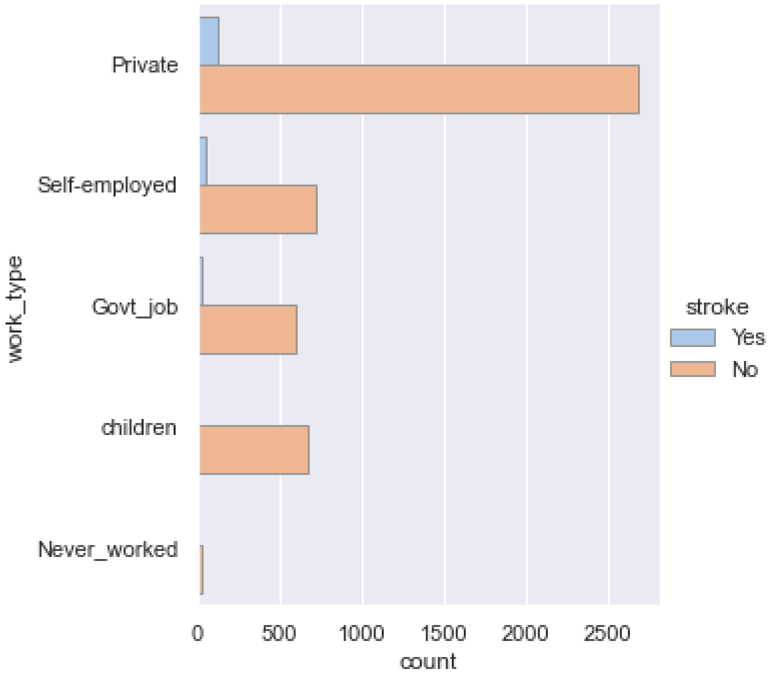


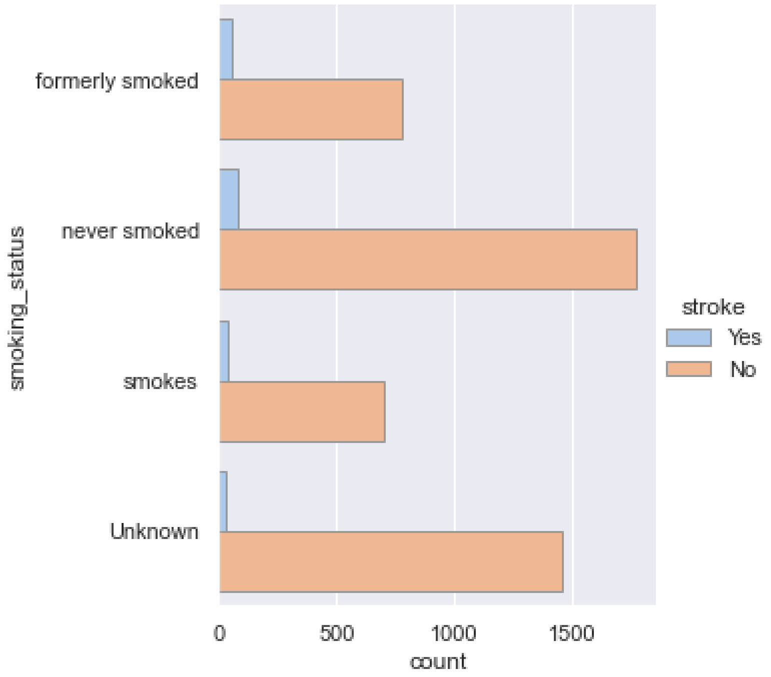
From the above visualization strokes appear to be more common in those with a low BMI and a high blood glucose level.

In [67]:

```
sns.catplot(y="work_type", hue="stroke", kind="count",
            palette="pastel", edgecolor=".6",
            data=df3)
sns.catplot(y="smoking_status", hue="stroke", kind="count",
            palette="pastel", edgecolor=".6",
            data=df3)
```

```
Out[67]: <seaborn.axisgrid.FacetGrid at 0x218e0f5a070>
```





The proportion of people who suffer a stroke is similar in the private and self-employed sectors. However, as compared to both the first and second gategories, persons in the government are less likely to have a stroke, and children are also less likely to have a stroke. Perhaps this is due to the amount of strain workers are under.Surprisingly, it appears that stroke is not strongly linked to smoking, as the proportion of people who have a stroke is almost the same regardless of smoking status.

Conclusion : So from all the above visualization we can clearly say that glucose level, age ecpsepicilly elderly group people, BMI are some of the significant factors that can cause heart stroke among people. So we cna say to that to avoid chances of having stroke we need to maintain a good BMI, idulge in more physical activities and also if we are old try to do some sorts of exercise & have healthy diet.