Stats 21 - HW 8 - Due 6/3/2023 by 11:59PM

Blair Lee - 005721089

Homework is generally an opportunity to practice coding and to train your problem solving and critical thinking skills. Putting Python to use is where learning happens.

Copying and pasting another's solutions takes away your learning opportunities. It is also academic dishonesty.

ChatGPT is always allowed in this class, but do remember, it is not foolproof and if your solution looks too much like another submission, I am required to file a report

Please use this document as your homework template and submit both the modified .ipynb file and a PDF OR HTML export.

Introduction

The data were derived from the US Centers for Disease Control 2010. It can also be found in Tableau. https://www.cdc.gov/obesity/data/index.html

Description of the Data

- County: Name of location
- Region: Region of the US
- State: State Name
- State_ABB: State Abbrevation (e.g., CA)
- Adult Obesity: Percentage Obese (BMI > 30)
- Adult Smokers: Percentage Smokers
- Children in Poverty: Percentage in Poverty (under age 18)
- Diabetic: Percentage Diabetic (all ages)
- Food Insecure: Percentage reporting difficulty having enough food to eat (all ages)
- Physically Inactive: Percentage Physically Inactive (all ages)

Problem 1: read the data

Correctly read the data from the CSV file named "Obesity.csv" using pandas and provide evidence (e.g., dimensions, data summary) that it was correctly read. Some of the data values may require cleaning/correction before they can be used properly.

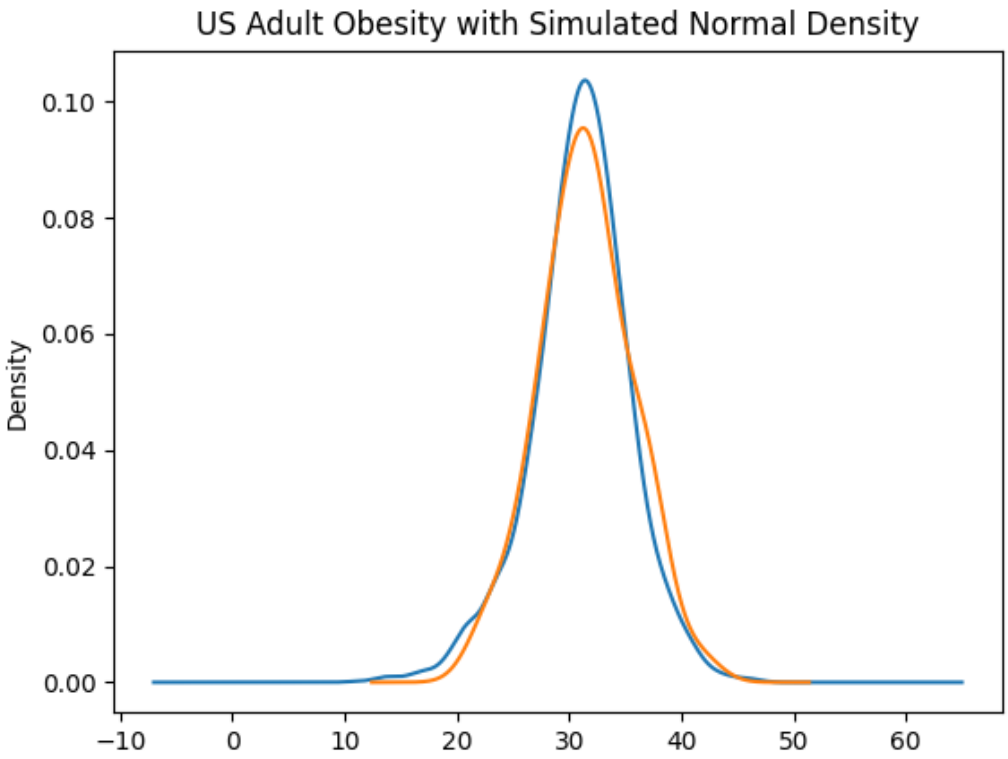
In [ ]:
```
## reserved for your answer
## read the data, remove the % and convert columns to numeric
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

obesity = pd.read_csv("Obesity.csv")
columns_with_percent = ["Adult Obesity", "Adult Smokers", "Children in Pover
for col in columns_with_percent:
    obesity[col] = obesity[col].str.replace("%", "").astype(float)

obesity.head()
```

Out[ ]:

	County	Region	State	State_ABB	Adult Obesity	Adult Smokers	Children in Poverty	Diabetic	F Insec
0	Adams	Midwest	Illinois	IL	35.0	16.0	20.0	10.0	
1	Alexander	Midwest	Illinois	IL	32.0	24.0	52.0	16.0	2
2	Bond	Midwest	Illinois	IL	31.0	17.0	21.0	10.0	
3	Boone	Midwest	Illinois	IL	34.0	16.0	15.0	10.0	
4	Brown	Midwest	Illinois	IL	32.0	16.0	15.0	8.0	

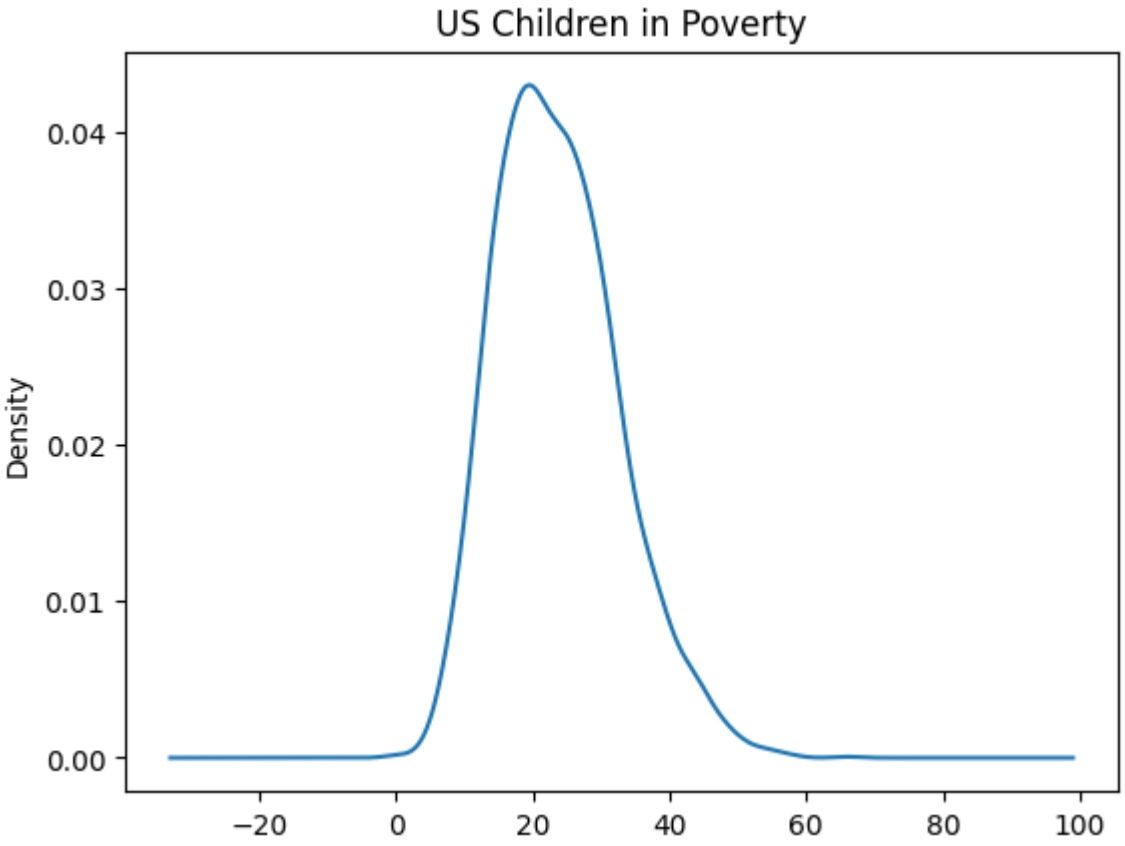
Problem 2: Choose any one numeric variable and create a simple graphic

In a comment, please offer an interpretation of what you are seeing in the data that is illustrated/made visible with your graphic. For example here is a graphic I made:



It appears that the distribution of obesity rates by state is close to normally distributed around a mean of 31% Obese

In [ ]:
```
## reserved for your answer
data = obesity["Children in Poverty"]
graph = data.plot.kde(title = "US Children in Poverty")
```

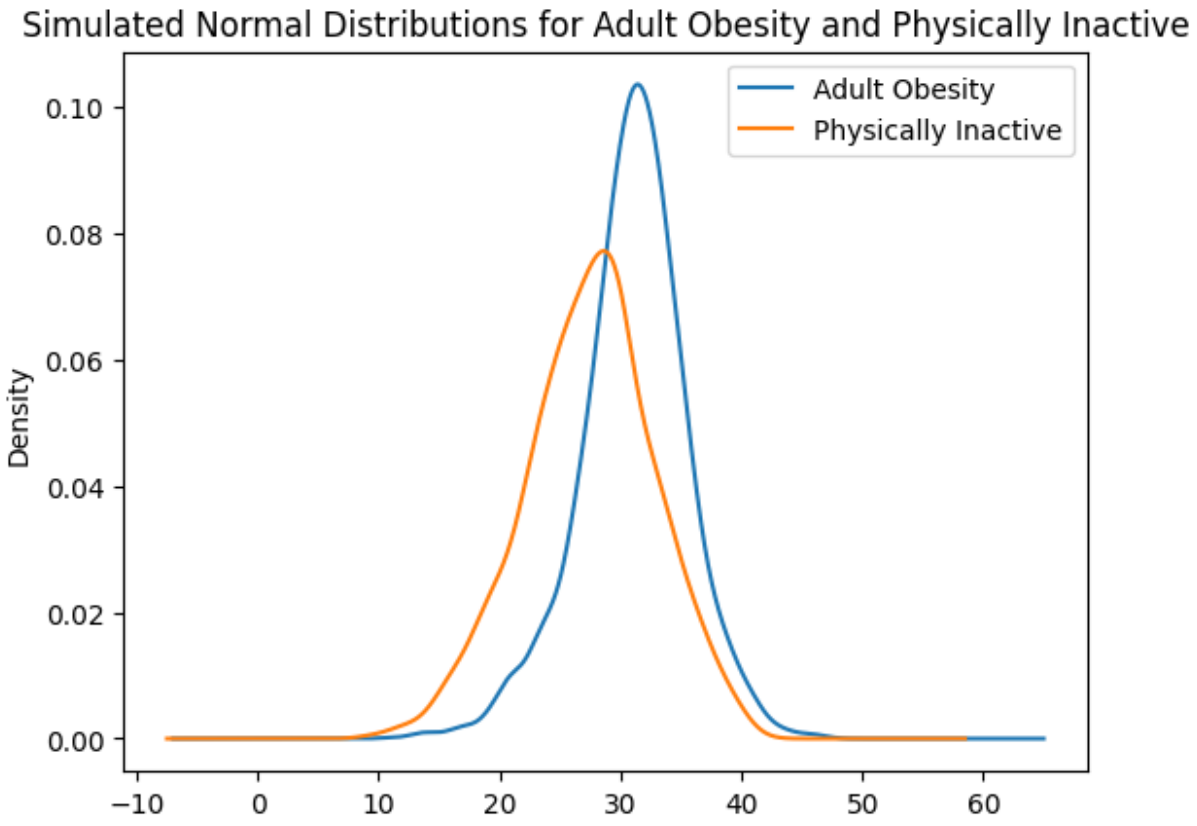


It appears the distribution is about normally distributed around a mean of 20% US children living in poverty.

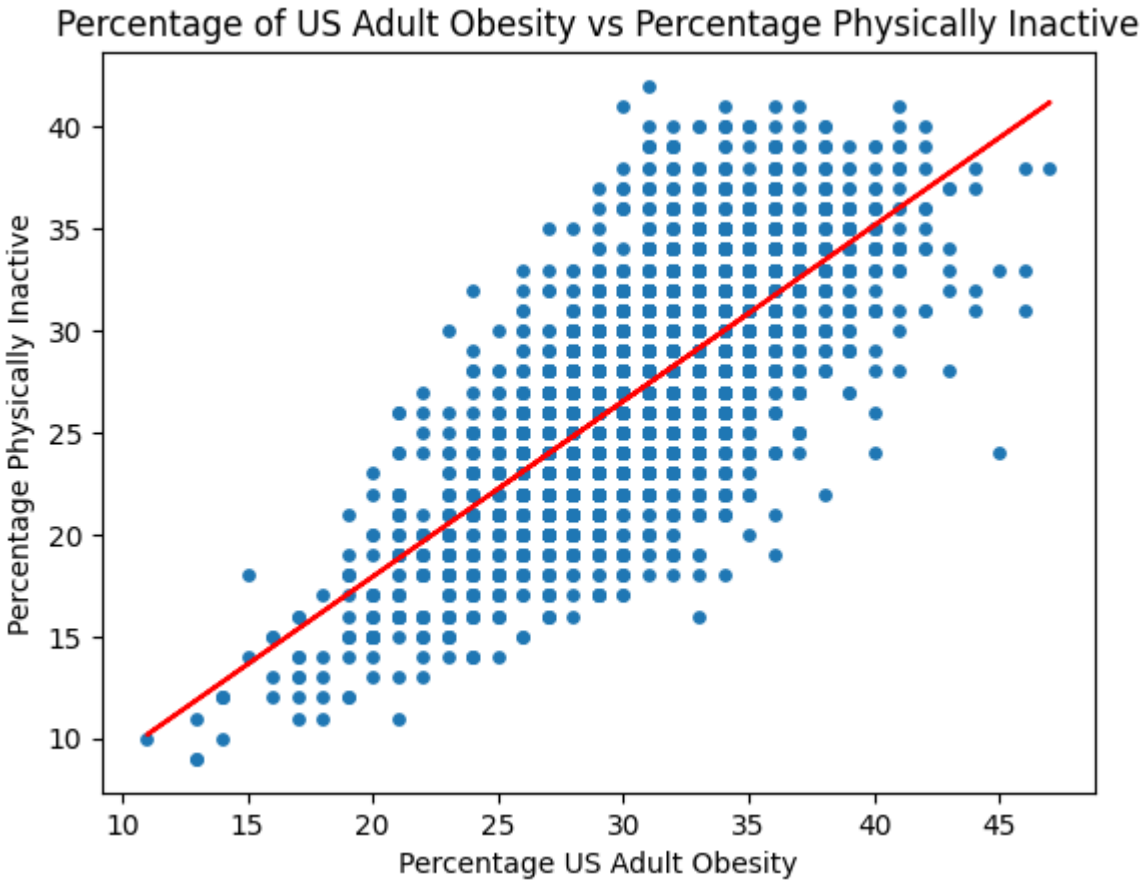
Problem 3: Choose any two numeric variables and create a graphic

Similar to #2, in a comment, please offer an interpretation of what you are seeing in the data that is illustrated/made visible with your graphic.

In [ ]:
```
## reserved for your answer
two_var = pd.DataFrame([obesity["Adult Obesity"], obesity["Physically Inacti
two_var_graph = two_var.plot.kde(title = "Simulated Normal Distributions for
```



In [ ]:
```
plt.scatter(obesity["Adult Obesity"], obesity["Physically Inactive"], s = 15
plt.xlabel(xlabel = "Percentage US Adult Obesity")
plt.ylabel(ylabel = "Percentage Physically Inactive")
plt.title("Percentage of US Adult Obesity vs Percentage Physically Inactive"

x = obesity["Adult Obesity"].values
y = obesity["Physically Inactive"].values
coeff = np.polyfit(x, y, 1)
regression_line = coeff[0] * x + coeff[1]
plt.plot(x, regression_line, color = "red")
plt.show()
```



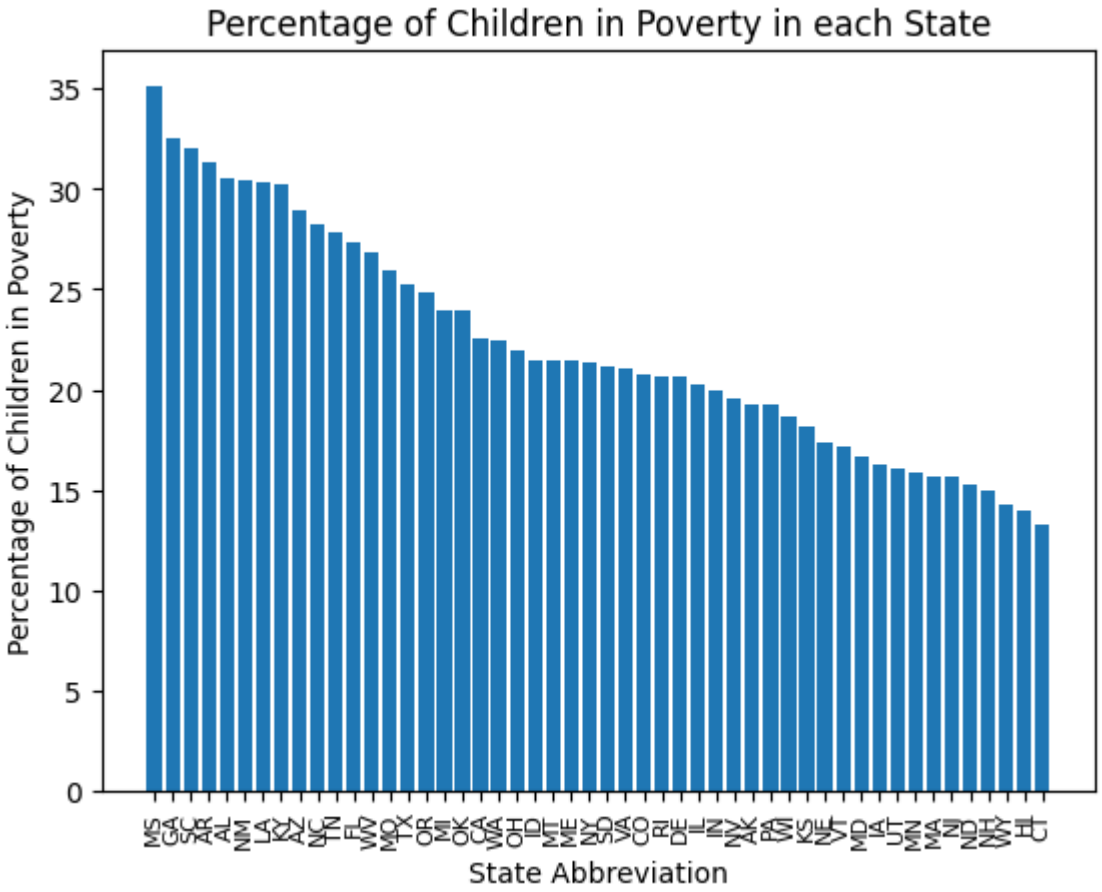
Looking at the density graph and the scatterplot, there appears to be a very positive, strong, linear relationship between the percentage of adults that is obese and the percentage of physically inactive people.

Problem 4: Choose one non-numeric variable and one numeric variable a create a graphic

Similar to #2, in a comment, please offer an interpretation of what you are seeing in the data that is illustrated/made visible with your graphic.

In [ ]:
```
df = pd.DataFrame([obesity["State_ABB"], obesity["Children in Poverty"]]).tr
sorted_df = df.groupby(by = "State_ABB").agg(np.mean)
sorted_df = sorted_df.sort_values("Children in Poverty", ascending=False)
sorted_df = sorted_df.reset_index()
```

In [ ]:
```
## reserved for your answer
plt.bar(sorted_df["State_ABB"], sorted_df["Children in Poverty"])
plt.xlabel("State Abbreviation")
plt.ylabel("Percentage of Children in Poverty")
plt.xticks(fontsize = 8, rotation = 90)
plt.title("Percentage of Children in Poverty in each State")

plt.show()
```



In looking at the percentage of children in poverty in each state, I took the mean percentage for each state and then plotted them in a bar graph. With limited resources, some states clearly need more help with impoverished children than other states. Missouri has almost 35% of children living in poverty.

Problem 5: Your choice of a custom plot

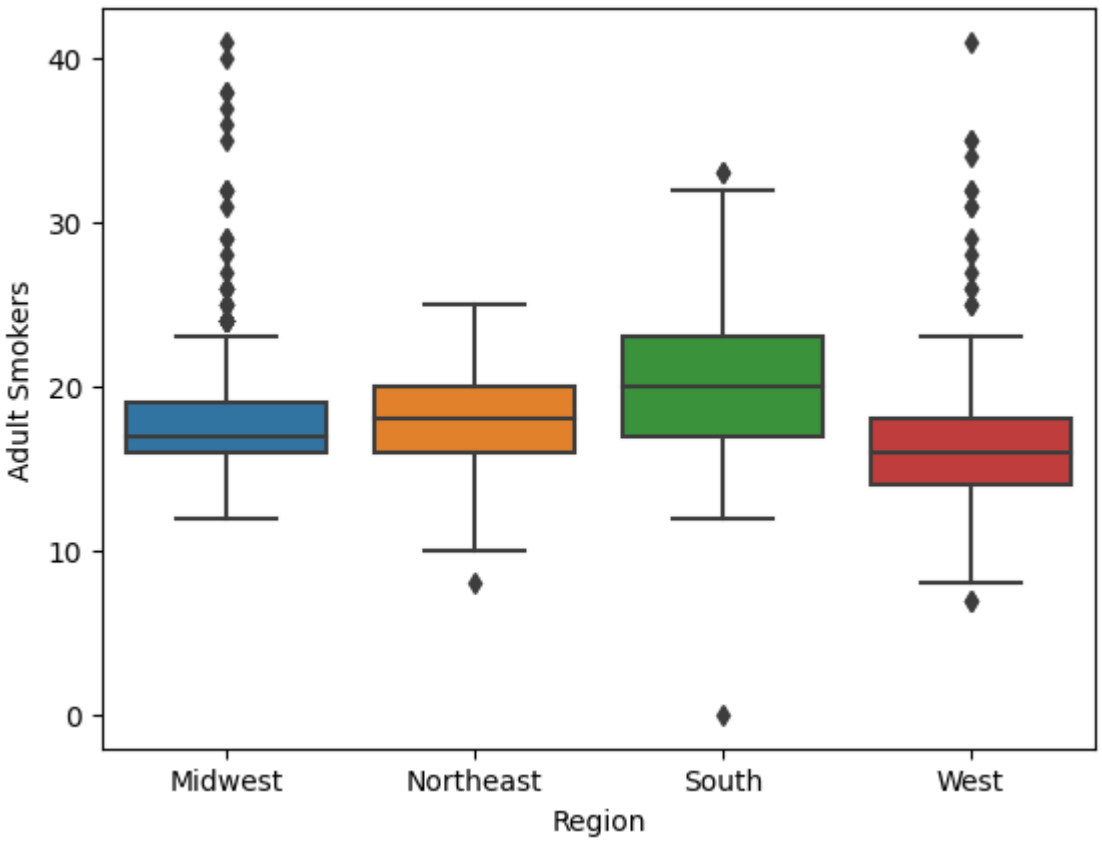
This is your opportunity to be creative. Use whatever module/package/library you want.

Make sure you label your axes with descriptive names and give a title to the graphic. Make sure your graph displays in your PDF or HTML submission

Please write a few sentences telling us about your decision of graphic type. For example, suppose you decide to create a map, we would like to know your justification, like "Oh, I thought it would be easy for anyone to understand because..."

In [ ]:
```
## reserved for your answer
map_df = pd.DataFrame([obesity["Region"], obesity["Adult Smokers"]]).transpo
map_df = map_df.drop(labels = "index", axis = 1)
```

In [ ]:
```
boxplot = sns.boxplot(x = map_df["Region"], y = map_df["Adult Smokers"])
```



The graphic shows each region and the corresponding boxplot for the percentage of adult smokers. The South appears to have the highest median while the midwest and west have a lot of outliers for high percentage of adult smokers.