Adeline Odunjo

Python Code & Interpretations

Adeline Odunjo

1.

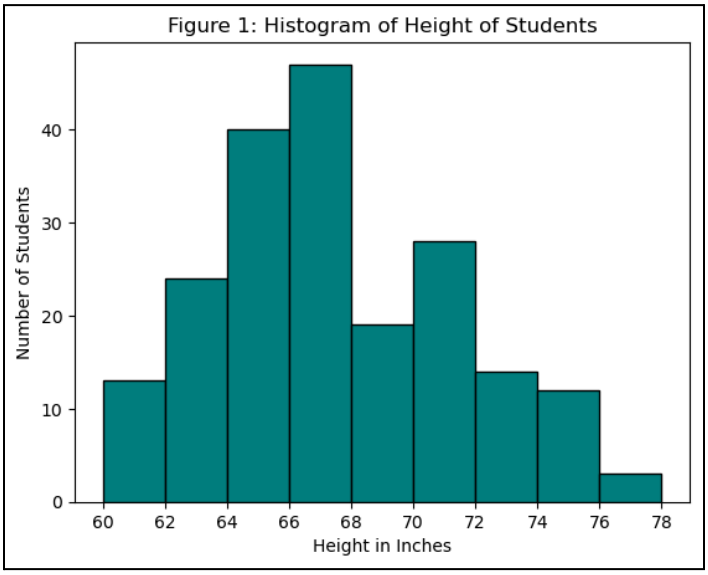
Table 1 : Data Dictionary				
Variable Name	Label	General Type	Specific Type	Measurement Units
Gender	Male of Female	Categorical	Nominal	N/A
Tattoo	Does the student have a tatoo - Yes or No	Categorical	Nominal	N/A
CDs	Categorized number of the student's estimate of how many music CDs he/she owns	Categorical	Ordinal	BelowAvg = 0 to 40, Average = 41 to 80, AboveAvg = more than 80
Height	Self reported height	Quantitative	Discrete	Inches
EarPrc	Number of ear piercings	Quantitative	Discrete	Number of ear piercings

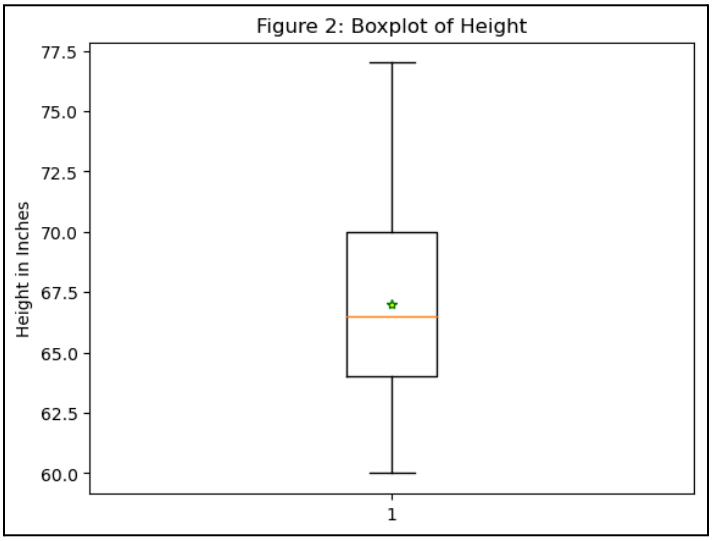
Table 1 provides an overview of all variables observed in this study.

2.

Table 2: Descriptive Statistics for Quantitative Variables		
	Height	EarPrc
Count	200.000	200.000
Mean	67.029	2.795
Standard Deviation	3.823	2.438
Minimum	60.000	0.000
Quartile 1	64.000	0.000
Median	66.500	2.000
Quartile 3	70.000	4.000
Maximum	77.000	13.000
IQR	6.000	4.000

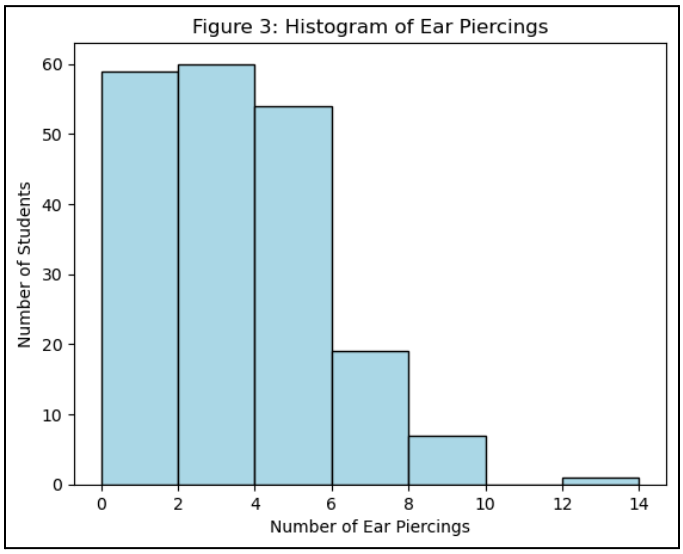
Table 2 depicts the descriptive statistics for each of the quantitative variables in this study: Height and EarPrc.

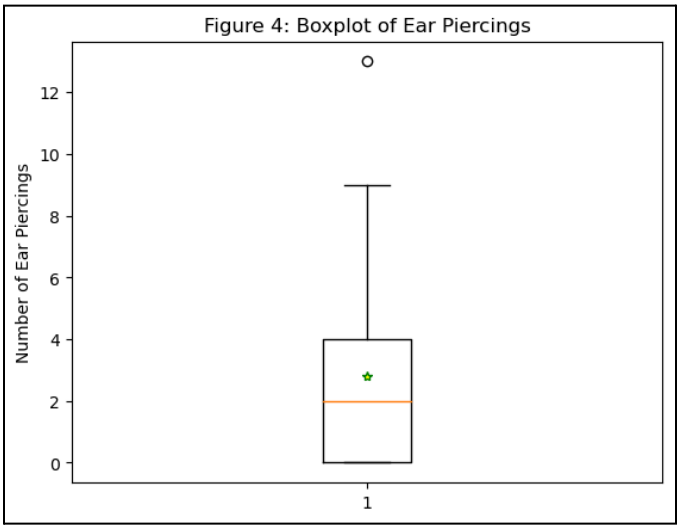




As shown in Figure 1, the distribution of the variable Height is unimodal and slightly right skewed. I can tell that the distribution is slightly right skewed because as seen in Figure 2, the median line on the boxplot is slightly lower than the center of the box, while the positive tail is longer than the negative tail.

Due to the skewness of the distribution, the appropriate measures of center and dispersion are median and IQR, respectively. The median for this variable is 66.500 inches, and the IQR for this variable is 6.000 inches. These values can be found in Table 2.





As shown in Figure 1, the distribution of the variable EarPrc is unimodal and right skewed. I can tell that the distribution is right skewed because as seen in Figure 2, the mean point is higher than the median line in the center of the box, there is a long positive tail with no apparent negative tail, and a positive outlier present.

Due to the skewness of the distribution, the appropriate measures of center and dispersion are median and IQR, respectively. The median for this variable is 2 ear piercings, and the IQR for this variable is 4 ear piercings. These values can be found in Table 2.

Table 3: Frequency Table of Gender	
Female	135
Male	65

The variable Gender is unevenly distributed. I know this because as shown in the frequency table for Gender in Table 3, out of 200 students, 135 are female and 65 are male. The mode of the variable Gender is Female.

Table 4: Frequency Table of Tattoo	
No	170
Yes	30

The variable Tattoo is unevenly distributed. I know this because as shown in the frequency table for Tattoo in Table 4, out of 200 students, 170 do not have Tattoos and 30 have tattoos. The mode of the variable Tattoo is No.

Table 5: Frequency Table of CDs	
Below Average	73
Average	69
Above Average	58

In this study, the categorizations for CDs are listed as 'BelowAvg', 'Average,' and 'AboveAvg' in Table 1. For the sake of clear understanding, these categories have been written out in full in Table 5 and may be referred to as such.

The variable CDs is unevenly distributed. I know this because as shown in the frequency table for CDs in Table 5, out of 200 students, 73 own a below average amount of CDs, 69 own an average amount of CDs, and 58 own an above average amount of CDs. The mode of the variable CDs is Below Average.

3.

Table 6: Stratified Means of Height by Tattoo	
No	66.869
Yes	67.933

Table 7: Stratified Means of EarPrc by Tattoo	
No	2.576
Yes	4.033

Table 6 depicts the mean height of students who do not have tattoos compared to those who do have tattoos. The mean height of students who do not have tattoos is 66.869 inches, which is close to the mean height of students who do have tattoos, which is 67.933. The difference between these two means is 1.064 inches.

Table 7 depicts the mean number of ear piercings for students who do not have tattoos compared to those who do have tattoos. The mean number of ear piercings of students who do not have tattoos is 2.576 piercings, and the mean number of ear piercings of students who do have tattoos is 4.033. The difference between these two means is 1.457 ear piercings.

The stratified means for Height by Tattoos are much closer comparatively to the means for Ear Piercings by Tattoo. According to Table 2, the variable EarPrc has a minimum of 0 piercings and a maximum of 13 piercings, while the variable Height has a minimum of 60 inches and a maximum of 77 inches. The means for the EarPrc variable have both a greater numerical difference, and a greater significance attributed to that difference compared to the Height variable.

4.

Table 8: Total Number of Piercings per CD Group	
Below Average	136
Average	201
Above Average	222

Table 9: Average Number of Pierıcings per CD group	
Below Average	2.345
Average	2.913
Above Average	3.041

As shown in Table 8, the total number of piercings from students who own a below average amount of CDs is 136, with a total of 201 piercings from those who own an average amount of CDs, and 222 piercings from those who own an above average amount of CDs.

As shown in Table 9, the mean number of piercings from students who own a below average amount of CDs is 2.345, with a mean of 2.913 piercings from students who own an average amount of CDs, and 23.041 from all students who own an above average amount of CDs.

The totals found in Table 8 are the more useful values in analyzing how many ear piercings are present within the different groups of CDs because they are not affected by outliers. The average values in Table 9 may be skewed by an upper outlier like 13, the maximum value found in the EarPrc column in Table 2. By looking at only the totals, you can analyze if the distribution of piercings between CD groups is even or not based on frequency alone.

5.

Table 10: Contingency Table of Count for Tattoo by CDs

		CDs		
		Below Average	Average	Above Average
Tattoo	No	65	57	48
	Yes	8	12	10

Table 10 depicts a contingency table of count for the variable Tattoo by the variable CDs, meaning it shows the counts of students identified as intersecting both categories of variables.

In this context, Table 10 shows that out of the 200 students, 48 own an above average amount of CDs and do not have tattoos.

Table 11: Contingency Table of Percent of Row for Tattoo by CDs

		CDs		
		Below Average	Average	Above Average
Tattoo	No	38.2	33.5	28.2
	Yes	26.7	40	33.3

A contingency table of percent of row means that each of the values of count found in Table 10 were divided by the total number of students observed for each row. These values are multiplied by 100, to depict what percentage of students own an above average, average, or below average amount of CDs, given that they are in a certain tattoo category.

In this context, Table 11 depicts a contingency table of percent of row for the variable Tattoo by the variable CDs. This table shows that 40% of students who have tattoos own an average amount of CDs.

Table 12: Contingency Table of Percent of Column for Tattoo by CDs				
		CDs		
		Below Average	Average	Above Average
Tattoo	No	89.041	82.609	82.759
	Yes	10.959	17.391	17.241

A contingency table of percent of column means that each of the values of count found in Table 10 were divided by the total number of students observed for each column. These values are multiplied by 100, to depict what percentage of students do or do not have tattoos, given that they are in a certain CD category.

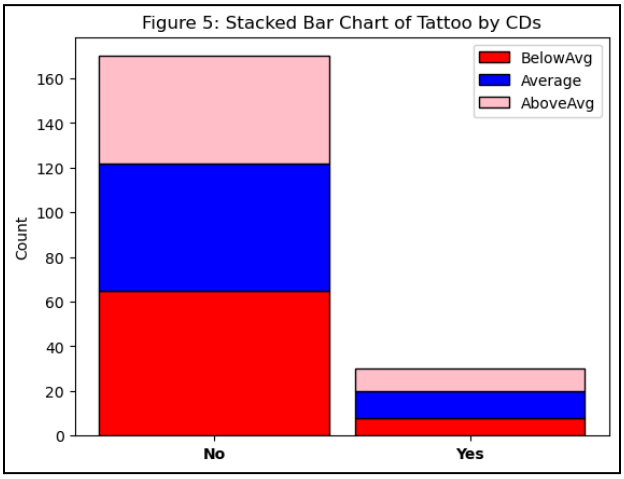
In this context, Table 12 depicts a contingency table of percent of column for the variable Tattoo by the variable CDs. This table shows that 17.391% of students who own an average amount of CDs have tattoos.

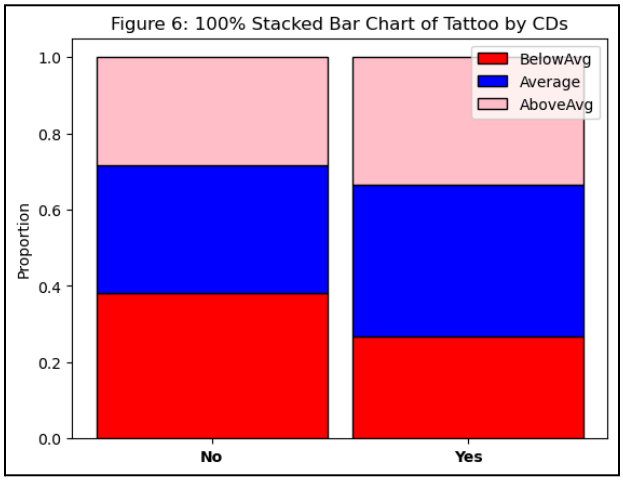
Table 13: Contingency Table of Percent of Total for Tattoo by CDs				
		CDs		
		Below Average	Average	Above Average
Tattoo	No	32.50	28.50	24.00
	Yes	4.00	6.00	5.00

A contingency table of percent of total means that each of the values of count found in Table 10 were divided by 200 (the total number of observed students) and multiplied by 100, to depict what percentage of students represent the intersection of both categories of variables.

In this context, Table 13 depicts a contingency table of percent of total for the variable Tattoo by the variable CDs. This table shows that 6% of all 200 students observed for this study both have tattoos and own an average number of CDs.

6.

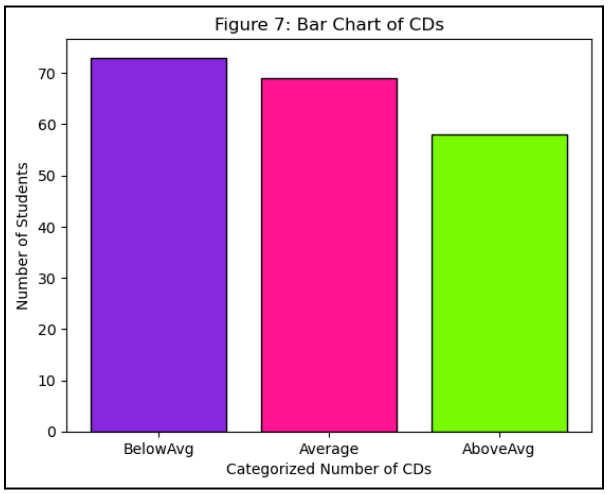


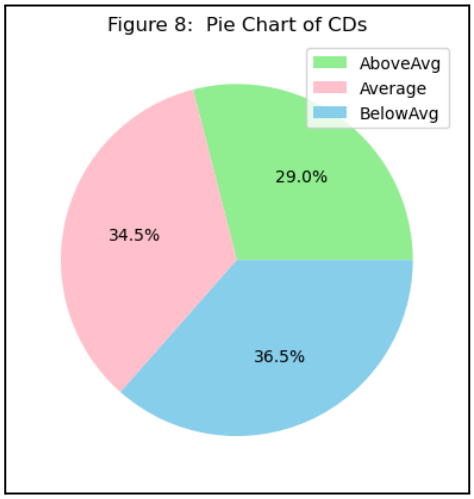


The stacked bar chart in Figure 5 shows the proportion of students owning different amounts of CDs separated by if they do or do not have tattoos. In this stacked bar chart, there is a significant difference between the number of students who do not have tattoos compared to those who do, but the proportions of CD categories seem evenly distributed in both the No and Yes categories. According to Table 4, out of 200 students, 170 of them have tattoos and 30 of them do not. This difference in count makes it more difficult to interpret the relationship between the variable Tattoo and the variable CDs.

The number of students in each category of the Tattoo variable is different, so the use of a 100% stacked bar chart as shown in Figure 6 is a better visualization of the relationship between the variable Tattoo and the variable CDs. This way, both populations of the Yes and No categories for the variable Tattoo can be evaluated compared to a total of 100%. By referring to Figure 6, it is clearer to see that the No category has a slightly smaller proportion of students who own an above average number of CDs compared to an average or below average number of CDs. It is also clear to see that the Yes category has a slighlty smaller proportion of students who own a below average number of CDs compared to an average or above average number of CDs.

7.





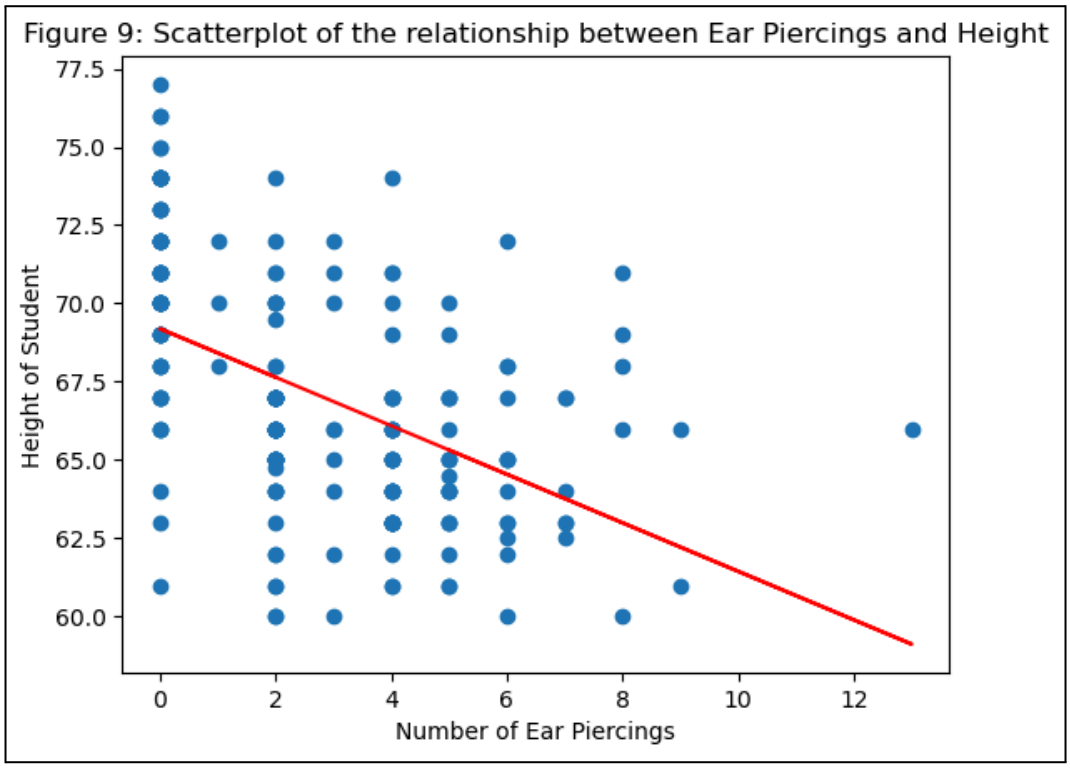
The variable CDs is ordinal as shown in Table 1, so the bar chart shown in Figure 7 is a better visualization of this variable because it can show progression. The pie chart shown in Figure 8 has no definitive beginning or end, making it more difficult to interpret progression. As shown in Figure 7, it is visually clearer that as the category amount of CD's increases, the number of students in that category decreases.

8.

Table 14: 90%, 95%, and 99% Confidence Interval for Height Variable					
n	Confidence Level	Mean	Lower	Upper	Margin of Error
60	90	67.496	66.719	68.273	0.777
60	95	67.496	66.565	68.427	0.931
60	99	67.496	66.258	68.734	1.238

Table 14 depicts the results of performing 90%, 95%, and 99% confidence intervals on the variable Height. The results show that when using the 95% interval, based on a representative sample of 60 observations, I am 95% sure that the mean Height in inches among all 200 participants is estimated to be between 66.565 inches and 68.427 inches. The margin of error was 0.931 and the sample mean was 67.496 inches.

9.



Based on the line of best fit included in the scatterplot shown in Figure 9, as the number of ear piercings a student has increases, the height of the student is expected to decrease.

Appendix

```
import pandas as pd
pen = pd.read_excel("C:/Users/adeli/OneDrive - Kennesaw State University/Comp Sci/pennstate2 Fall23.xlsx", sheet_name="Sheet1")
print(pen.head())

#Quantitative Variable: Height

#Height Descriptive Statistics
print("Summary Statistics of the variable Height")
print(round(pen['Height'].describe(),3))
print("The median of the variable Height is:",(round(pen['Height'].median(),3)))
import scipy.stats as scipy
print("The IQR of the variable Height is",round(scipy.iqr(pen['Height']),3))

#Height Histogram
Height_data = pen['Height']
binwidth=2
plt.hist(Height_data, color="teal",edgecolor="black",bins=range(60,80,binwidth))
plt.xlabel("Height in Inches")
plt.ylabel("Number of Students")
plt.title("Figure 1: Histogram of Height of Students")
plt.xticks(range(60,80,binwidth))
plt.show()

#Height Vertical Boxplot
import matplotlib.pyplot as plt
```

```
plt.boxplot(pen['Height'], showmeans=True, vert=True,
                    meanprops={"marker":"*","markerfacecolor":"yellow",
                   "markeredgecolor":"green"})
plt.ylabel("Height in Inches")
plt.title("Figure 2: Boxplot of Height")
plt.show()

#Quantitative Variable: EarPrc

#EarPrc Descriptive Statistics
print("Summary Statistics of the variable EarPrc")
print(round(pen['EarPrc'].describe(),3))
print("The median of the variable EarPrc is:",(round(pen['EarPrc'].median(),3)))
import scipy.stats as scipy
print("The IQR of the variable EarPrc is",round(scipy.iqr(pen['EarPrc']),3))

#EarPrc Histogram
Height_data = pen['EarPrc']
binwidth=2
plt.hist(Height_data, color="lightblue",edgecolor="black",bins=range(0,15,binwidth))
plt.xlabel("Number of Ear Piercings")
plt.ylabel("Number of Students")
plt.title("Figure 3: Histogram of Ear Piercings")
plt.xticks(range(0,15,binwidth))
plt.show()

#EarPrc Boxplot
```

```
import matplotlib.pyplot as plt
plt.boxplot(pen['EarPrc'], showmeans=True, vert=True,
                meanprops={"marker":"*","markerfacecolor":"yellow",
               "markeredgecolor":"green"})
plt.ylabel("Number of Ear Piercings")
plt.title("Figure 4: Boxplot of Ear Piercings")
plt.show()

#Question 2 - Frequency Tables

#Frequency Table for Gender
print(pen['Gender'].value_counts())
print()

#Frequency Table for Tattoo
print(pen['Tattoo'].value_counts())
print()

#Frequency Table for CDs
print(pen['CDs'].value_counts())

#Question 3: Perform a stratified analysis of the mean of the two quantitative variables by
“Tattoo”
#status. Compare and contrast the mean number of piercings and mean height between
#these two groups.

#Height by Tattoo
print(round(pen.groupby(['Tattoo'])['Height'].mean(),3))
```

```
print()

#Piercings by Tatoo
print(round(pen.groupby(['Tattoo'])['EarPrc'].mean(),3))

#Question 4
#Perform a stratified analysis of the “EarPrc” variable by the “CDs” variable. Include the
#total number of piercings per “CD” group and the average number of piercings per “CD” group in the table

#Total number of piercings per CD group
print(round(pen.groupby(['CDs'])['EarPrc'].sum(),3))

#Average number of piercings per cd group
print(round(pen.groupby(['CDs'])['EarPrc'].mean(),3))

#Question 5 - Contingency Table of Count for Tattoo by CDs
import pandas as pd
pd.crosstab(pen["Tattoo"],pen['CDs'])

#Question 5 - Contingency Table of  Percent of Row for Tattoo by CDs
import pandas as pd
round(pd.crosstab(pen["Tattoo"],pen['CDs'],normalize=0),3)*100

#Question 5 - Contingency Table of  Percent of Column for Tattoo by CDs
import pandas as pd
round(pd.crosstab(pen["Tattoo"],pen['CDs'],normalize=1)*100,3)
```

```
#Question 5 - Contingency Table of  Percent of Total for Tattoo by CDs
import pandas as pd
round(pd.crosstab(pen["Tattoo"],pen['CDs'],normalize=True),3)*100

#Question 6 - Create a Stacked Bar Chart based on the Frequency Table from Question 5
import matplotlib.pyplot as plt
import numpy as np

Tattoo = ["No","Yes"]
AboveAvg=np.array([48,10])
Average=np.array([57,12])
BelowAvg=np.array([65,8])
r = [0,1]  #Based on the number of bars
barWidth = .9
plt.bar(r, BelowAvg, color="red", edgecolor="black", width=barWidth)
plt.bar(r, Average, bottom = BelowAvg, color="blue", edgecolor="black",
        width=barWidth)
plt.bar(r, AboveAvg, bottom = Average + BelowAvg, color="pink", edgecolor="black",
        width=barWidth)
plt.xticks(r, Tattoo, fontweight = "bold")
plt.ylabel("Count")
plt.legend(['BelowAvg','Average','AboveAvg'])
plt.title("Figure 5: Stacked Bar Chart of Tattoo by CDs")
plt.show()

#Question 6 - Create a 100% Stacked Bar Chart based on the Frequency Table from Question 5
```

```
import matplotlib.pyplot as plt
import numpy as np

Tattoo = ["No","Yes"]
AboveAvg=np.array([48/170,10/30])
Average=np.array([57/170,12/30])
BelowAvg=np.array([65/170,8/30])

r = [0,1]  #Based on the number of bars
barWidth = .9
plt.bar(r, BelowAvg, color="red", edgecolor="black", width=barWidth)
plt.bar(r, Average, bottom = BelowAvg, color="blue", edgecolor="black",
        width=barWidth)
plt.bar(r, AboveAvg, bottom = BelowAvg + Average, color="pink", edgecolor="black",
        width=barWidth)
plt.xticks(r, Tattoo, fontweight = "bold")
plt.legend(['BelowAvg','Average','AboveAvg'])
plt.title("Figure 6: 100% Stacked Bar Chart of Tattoo by CDs")
plt.ylabel("Proportion")
plt.show()

#Question 7 - Bar chart of CDs

import matplotlib.pyplot as plt
CDname = ['BelowAvg','Average','AboveAvg']
CDfreq =[73,69,58]
plt.bar(CDname, CDfreq, color=("blueviolet", "deeppink", "lawngreen"),
```

```
        edgecolor="black")
plt.xlabel("Categorized Number of CDs")
plt.ylabel("Number of Students")
plt.title("Figure 7: Bar Chart of CDs", loc="center")
plt.show()

#Question 7 - Pie chart chart of CDs

import matplotlib.pyplot as plt
import pandas as pd

my_tab = pd.crosstab(index = pen["CDs"],columns="Count")
print(my_tab)
colors = ('lightgreen', 'pink', 'skyblue')
plt.pie(my_tab.values.reshape(-1), colors=colors, autopct='%1.1f%%')
plt.legend(['AboveAvg','Average','BelowAvg'])
plt.title("Figure 8:  Pie Chart of CDs")
plt.show()

#Question 8
import random as rn
import numpy as np

rn.seed(55673)
pen['Random'] = np.random.rand(200)
pen.head()
```

```
#Question 8
pensort = pen.sort_values(by=['Random'])
print(pensort.head())
pensample = pensort[:60]

#Question 8
from scipy import stats as sp
import pandas as pd
import numpy as np

def CI(var,c_lev,units,digits):
    alpha = 1 - c_lev
    n = len(var)
    mean = round(var.mean(),digits)
    sd = np.std(var, ddof=1)
    tstar = sp.t.ppf(1-(alpha/2), n-1)
    me = round(tstar*sd/np.sqrt(n),digits)
    lower = round(mean-me,digits)
    upper = round(mean+me,digits)
    print("In a sample of",n, "observations, the sample mean was",mean, units)
    print("The margin of error was", me, units)
    print("The", c_lev*100, "% confidence interval for the population mean")
    print("is between", lower, "and", upper, units)

#Question 8
CI(pensample['Height'],0.90,"inches",3)
print()
```

```
CI(pensample['Height'],0.95,"inches",3)
print()
CI(pensample['Height'],0.99,"inches",3)


#Question 9: Create a scatterplot for the two quantitative variables, predicting the height of a student
#based solely on the number of ear piercings the student has. Include the line of best fit


import matplotlib.pyplot as plt
import numpy as np


x = pen["EarPrc"]
y = pen["Height"]
m, b = np.polyfit(x, y, 1)
plt.figure()
plt.scatter(x, y)
plt.plot(x, m*x+b, color = 'red')
plt.xlabel("Number of Ear Piercings")
plt.ylabel("Height of Student")
plt.title("Figure 9: Scatterplot of the relationship between Ear Piercings and Height")
plt.show()
```