Dalton Shaver

STAT 3010

November 2, 2021

**Cancer Risk Study in Python & R**

**1)**

Table 1: Descriptive Statistics for Age, Calories, and Fat

|  |  |  |  |
| --- | --- | --- | --- |
| **Statistics** | **Age** | **Calories** | **Fat** |
| Count | 315.00 | 315.00 | 315.00 |
| Mean | 50.15 | 1796.65 | 77.03 |
| Standard Deviation | 14.58 | 680.35 | 33.83 |
| Minimum | 19.00 | 445.20 | 14.40 |
| Q1 | 39.00 | 1338.00 | 53.95 |
| Median | 48.00 | 1666.80 | 72.90 |
| Q3 | 62.50 | 2100.45 | 95.25 |
| Maximum | 83.00 | 6662.20 | 235.90 |
| IQR | 23.50 | 762.45 | 41.30 |

Table 1 displays the descriptive statistics for the three quantitative variables Age, Calories, and Fat. For age, the mean is slightly higher than the median. This is also the case for both calories and fat consumption. This results in all three quantitative variables being right-skewed. When examining calories, there is a wide range from 445.20 to 6662.20. The quartiles are significantly lower, with the interquartile range only being 762.45 calories. We can assume that is at least one significant outlier in the distribution for calories. Fat consumption is also similar, as the first quartile is 53.95 and third quartile is 95.25. It is clear there are outliers as the mean is greater than the median, and the maximum fat consumption is 235.90.

**2)**

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Both Figure 1 and Figure 2 show the distribution is unimodal and right-skewed. Although the graphs display skewness, the mean is still an appropriate measurement of central tendency as there are no outliers. When examining central tendency, the mean (50.15 years) and median (48 years) are very close. Additionally, the best representation for dispersion is the standard deviation, 14.58 years.

Chart, histogram

Description automatically generated

Chart, scatter chart, box and whisker chart

Description automatically generated

Figure 3 is a histogram of calories consumed per day among patients. The distribution is unimodal and skewed to the right. The mean, 1796.65 calories, is fairly higher than the median, 1666.80 calories. In Figure 4, it is clear there are multiple outliers present in the distribution. In this case, the median and interquartile range (762.45 calories) are the appropriate measurements of central tendency and dispersion.

Chart, histogram

Description automatically generated

Chart

Description automatically generated

Similar to the age and calorie consumption of patients, fat consumption is also an unimodal and right-skewed distribution. There are nine outliers skewing the graph, resulting in the mean (77.03g) being higher than the median (72.90g). Therefore, the median is best measure for central tendency. The interquartile range, 41.30g, is the appropriate measurement of dispersion for fat consumption.

**3)**

Table 2: Frequency Table for Age Category

|  |  |  |
| --- | --- | --- |
| **Age Category** | **Frequency** | **Percent** |
| Millennial | 84 | 26.7% |
| Late Gen X | 73 | 23.2% |
| Early Gen X | 74 | 23.5% |
| Baby Boomer | 84 | 26.7% |
| **Total** | **315** | **100.0%** |

Table 2 shows that the number of patients in each age category are fairly similar in size, with the number of Millennials (84) and Baby Boomers (84) being exactly the same.

Chart, pie chart

Description automatically generated

Chart, bar chart

Description automatically generated

The frequency of Millennials and Baby Boomers (shown in Table 2) are equivalent with 84 patients in both categories, making up 53.4% of total patients. The two categories of patients in Generation X are nearly the same, with 73 in Late Gen X and 74 in Early Gen X. Figure 7 gives a visual representation in terms of percentage. The number of Baby Boomers and Millennials are slightly higher than Generation X, with both representing 26.7% of total patients. Figure 8 gives an additional visual graph, displaying the frequency count of each category.

**4)**

Table 3: Contingency Table for Gender by Smoke Status

|  |  |  |  |
| --- | --- | --- | --- |
| **Gender** | **Never** | **Former** | **Current** |
| Male | 13 | 22 | 7 |
| Female | 144 | 93 | 36 |

Table 3 is a contingency table showing the number of males and females that have either never smoked, smoked in the past, or currently smokes. It shows that there are more females than males in this study. Examining the table closer, it also shows that there were more females that had never smoked compared to the number of both female former smokers and female current smokers.

Table 4: Percentage of total for Gender by Smoke Status

|  |  |  |  |
| --- | --- | --- | --- |
| **Gender** | **Never** | **Former** | **Current** |
| Male | 4.13% | 6.98% | 2.22% |
| Female | 45.71% | 29.52% | 11.43% |

Table 4 shows the total percentage of all patients in the study. According to the table, there were more male former smokers (6.98%) than males that had never smoked (4.13%). For women, there was a higher percentage of non-smokers (45.71%) compared to those who were former smokers (29.52%).

Table 5: Percentage of row for Gender by Smoke Status

|  |  |  |  |
| --- | --- | --- | --- |
| **Gender** | **Never** | **Former** | **Current** |
| Male | 30.95% | 52.38% | 16.67% |
| Female | 52.75% | 34.07% | 13.19% |

In Table 5, the percentage of rows is displayed for the variable Gender and Smoke Status. It shows that only 13.19% of the females are current smokers while 16.67% of the males are current smokers.

Table 6: Percentage of column for Gender by Smoke Status

|  |  |  |  |
| --- | --- | --- | --- |
| **Gender** | **Never** | **Former** | **Current** |
| Male | 8.28% | 19.13% | 16.28% |
| Female | 91.72% | 80.87% | 83.72% |

For Table 6, the largest difference of percentages between males and females are patients that have never smoked. Females represents 91.72% of the non-smokers in the study.

**5)**

Chart, bar chart

Description automatically generated

Figure 9 shows the distribution of the smoke statuses by male and female. According to the graph, only 8.28% of patients that have never smoked were male. This is much lower than the 19.13% of former smokers that were male. The graph shows there are higher percentages of males among former and current smokers compared to patients that have never smoked.

**6)**

Table 7: 95% Confidence Interval for Calories Sample

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Sample Size** | **Confidence Level** | **Sample Mean** | **Margin of Error** | **Lower Bound** | **Upper Bound** |
| 80 | 95% | 1758.42 | 134.21 | 1624.21 | 1892.63 |

Table 7 is a confidence interval for the variable Calories. Based on a random sample of 80 patients, we can be 95% confident the average number of calories consumed by patients in the selected study is between 1624.21 and 1892.63 calories.

**7)**

**Research Hypothesis:**

My implied research hypothesis for this study is that the age category Late Generation X (Late Gen X) consumes more calories than other generations. The explanatory variable in the statistical analysis is Age, while the response variable is Calories.

Chart, scatter chart

Description automatically generated

According to the line of regression shown in Figure 10, the amount of calories consumed slowly decreases as age increases. This shows that the variable Calories and Age have a weak negative correlation. The scatterplot also shows a higher number of observations around age 40. The age category Millennial constitutes patients 39 years old and younger, while Late Gen X is composed of patients 40-47 years old.

Chart, box and whisker chart

Description automatically generated

Figure 11 gives a visual representation of each age category, displaying the relative descriptive statistics needed for a deeper analysis. According to the graph, Millennials have the highest median for consumption of calories, while Baby Boomers have the lowest. The interquartile ranges for both Late Gen X and Baby Boomer are smaller than Millennials and Early Gen X. In addition, there are more outliers for Late Gen X, therefore skewing the mean to the right.

Chart, bar chart, treemap chart

Description automatically generated

Figure 12 gives a visual representation of the distribution of each category of calories. Among the four calories categories, the distributions show patterns as the number of calories increase. For Millennials and Late Gen X, the percentage of which they account for each category increases as the number of calories increases. Millennials account for the largest portion of high calorie consumption, at exactly 44.44%. Secondly, Late Gen X is the second largest age group for high calorie consumption, consisting of 33.33%. The Above Average calories category is composed of Early Gen X (30.12%), Millennials (28.92%), Late Gen X (20.48%), and Baby Boomers (20.48%)

Table 8: Stratified Analysis of Calories by Age Category

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Age Category** | **Minimum** | **Quartile 1** | **Median** | **Mean** | **Quartile 3** | **Maximum** |
| Baby Boomer | 647.70 | 1239.73 | 1533.15 | 1649.56 | 1910.20 | 6662.20 |
| Early Gen X | 445.20 | 1328.70 | 1731.80 | 1806.00 | 2181.38 | 4373.60 |
| Late Gen X | 740.60 | 1372.70 | 1660.10 | 1805.96 | 2075.90 | 3711.00 |
| Millennial | 798.20 | 1497.90 | 1832.55 | 1927.43 | 2266.00 | 3511.10 |

Table 8 examines the descriptive statistics of each age category by the number of calories consumed. Millennials have both the highest median (1832.55) and mean (1927.43) of the four age categories, but the highest minimum (798.20) and lowest maximum (3511.10). This shows that younger patients had higher calorie counts on average, but had a lower range. Early Gen X had the second highest median and mean, followed by Late Gen X and Baby Boomers.

Table 9: Contingency Table for Age Category by Calories

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Age Category** | **Low** | **Average** | **Above Average** | **High** |
| Millennial | 11 | 41 | 24 | 8 |
| Late Gen X | 12 | 38 | 17 | 6 |
| Early Gen X | 10 | 37 | 25 | 2 |
| Baby Boomer | 17 | 48 | 17 | 2 |

Table 9 shows the counts of each age category by calories category. The Millennials and Late Gen X age groups had more patients consuming a high amount of calories per day than the age groups Early Gen X and Baby Boomers.

Table 10: Percentage of total for Age Category by Calories

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Age Category** | **Low** | **Average** | **Above Average** | **High** |
| Millennial | 3.49% | 13.02% | 7.62% | 2.54% |
| Late Gen X | 3.81% | 12.06% | 5.40% | 1.90% |
| Early Gen X | 3.17% | 11.75% | 7.94% | 0.63% |
| Baby Boomer | 5.40% | 15.24% | 5.40% | 0.63% |

For each age category, a higher number of patients consumed an average amount of calories per day according to Table 10. A lower percentage of total patients consumed above average, followed by low consumption, and finally high consumption.

Table 11: Percentage of row for Age Category by Calories

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Age Category** | **Low** | **Average** | **Above Average** | **High** |
| Millennial | 13.10% | 48.81% | 28.57% | 9.52% |
| Late Gen X | 16.44% | 52.05% | 23.29% | 8.22% |
| Early Gen X | 13.51% | 50.00% | 33.78% | 2.70% |
| Baby Boomer | 20.24% | 57.14% | 20.24% | 2.38% |

A greater number of patients consumed a high number of calories out of all Millennials (9.52%) compared to other age categories. We also see in Table 11 that roughly half of each age group consumed an average number of calories per day.

Table 12: Percentage of column for Age Category by Calories

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Age Category** | **Low** | **Average** | **Above Average** | **High** |
| Millennial | 22.00% | 25.00% | 28.92% | 44.44% |
| Late Gen X | 24.00% | 23.17% | 20.48% | 33.33% |
| Early Gen X | 20.00% | 22.56% | 30.12% | 11.11% |
| Baby Boomer | 34.00% | 29.27% | 20.48% | 11.11% |

Table 12 displays the distribution of each calories category by age category. This table shows that Baby Boomers account for a larger portion of patients consuming a low or average amount of calories. The table also shows Early Gen X accounts for almost a third of patients that consume an above average amount of calories, while Millennials account for 44.44% of all patients that consume a high amount of calories per day.

**Conclusion:**

The research hypothesis of the age category Late Gen X consuming the most calories is not accurate according to the data, therefore it fails. After further examining the data using the statistical tools of a scatterplot, side-by-side boxplots, 100% stacked bar chart, stratified analysis, and contingency tables, it is accurate to state that the Millennial age group consumes more calories per day. According to the stratified analysis, Millennials possessed the highest median (1832.55) and mean (1927.43). These values are higher than the median (1660.10) and mean (1805.96) of Late Generation X. Furthermore, the contingency tables showed Millennials accounted for a higher percentage of total patients that consumed an above average or high amount of calories. When analyzing the percentage of columns, Early Generation X accounted for 30.12% of the Above Average category while Late Generation X accounted for 20.48%. In addition, Millennials accounted for 44.44% of patients consuming high amounts of calories, compared to Late Generation X patients consuming 33.33%. Because of these results, it is clear that the Millennial age group consumes more calories per day than the other age groups.

**Appendix 1: R Script**

library(ggplot2)

library(ggthemes)

library(tidyverse)

# Question 2

## Figure 1: Histogram of Age

ggplot(data, aes(x=Age))+

geom\_histogram(binwidth = 8, col="black", fill="mediumpurple")+

theme\_clean()+

theme(axis.text.x=element\_text(color="black", size=12),

axis.text.y=element\_text(color="black", size=12),

axis.title.x=element\_text(size=12),

axis.title.y=element\_text(size=12),

plot.title=element\_text(size=15, hjust=0.5))+

labs(title="Figure 1: Histogram of Age", x="Age (years)", y="Frequency")

## Figure 2: Boxplot of Age

ggplot(data, aes(x=Age))+

geom\_boxplot(col="black", fill="mediumpurple")+

theme\_clean()+

theme(axis.text.x=element\_text(color="black", size=12),

axis.text.y=element\_text(color="white"),

axis.title.x=element\_text(size=12),

plot.title=element\_text(size=15, hjust=0.5))+

labs(title="Figure 2: Boxplot of Age", x="Age (years)")

## Figure 3: Histogram of Calories

ggplot(data, aes(x=Calories))+

geom\_histogram(binwidth=450, col="black", fill="salmon")+

theme\_clean()+

theme(axis.text.x=element\_text(color="black", size=12),

axis.text.y=element\_text(color="black", size=12),

axis.title.x=element\_text(size=12),

axis.title.y=element\_text(size=12),

plot.title=element\_text(size=15, hjust=0.5))+

labs(title="Figure 3: Histogram of Calories Consumed Per Day", x="Calories", y="Frequency")

## Figure 4: Boxplot of Calories

ggplot(data, aes(x=Calories))+

geom\_boxplot(col="black", fill="salmon", outlier.size=3)+

theme\_clean()+

theme(axis.text.x=element\_text(color="black", size=12),

axis.text.y=element\_text(color="white"),

axis.title.x=element\_text(size=12),

plot.title=element\_text(size=15, hjust=0.5))+

labs(title="Figure 4: Boxplot of Calories Consumed Per Day", x="Calories")

## Figure 5: Histogram of Fat

ggplot(data, aes(x=Fat))+

geom\_histogram(binwidth = 20, col="black", fill="skyblue3")+

theme\_clean()+

theme(axis.text.x=element\_text(color="black", size=12),

axis.text.y=element\_text(color="black", size=12),

axis.title.x=element\_text(size=12),

axis.title.y=element\_text(size=12),

plot.title=element\_text(size=15, hjust=0.5))+

labs(title="Figure 5: Histogram of Fat Consumed Per Day", x="Fat (grams)", y="Frequency")

## Figure 6: Boxplot of Fat

ggplot(data, aes(x=Fat))+

geom\_boxplot(col="black", fill="skyblue3", outlier.size=3)+

theme\_clean()+

theme(axis.text.x=element\_text(color="black", size=12),

axis.text.y=element\_text(color="white"),

axis.title.x=element\_text(size=12),

plot.title=element\_text(size=15, hjust=0.5))+

labs(title="Figure 6: Boxplot of Fat Consumed Per Day", x="Fat (grams)")

# Question 3

### Categorization of age to create variable 'agecat'

data$agecat[data$Age<40] = "Millennial"

data$agecat[data$Age>=40 & data$Age<48] = "Late Gen X"

data$agecat[data$Age>=48 & data$Age<62] = "Early Gen X"

data$agecat[data$Age>=62] = "Baby Boomer"

## Table 2: Frequency Table for Age Category

table\_agecat = table(data$agecat)

df\_freq = as.data.frame(table(ord\_agecat))

df\_freq$Relative = round(prop.table(table(ord\_agecat))\*100, 1)

names(df\_freq) = c("Age Category", "Frequency", "Percent")

total\_row = c(sum(df\_freq$Frequency), round(sum(df\_freq$Percent), 0))

df\_freq = rbind(df\_freq, total\_row)

df\_freq

## Figure 7: Pie Chart of Age Category

df\_agecat = as.data.frame(table\_agecat)

names(df\_agecat) = c("Age\_Category", "Frequency")

ggplot(df\_agecat, aes(x="", y=Frequency, fill=factor(Age\_Category)))+

geom\_bar(width=1, stat="identity")+

theme\_void()+

geom\_text(aes(x=1.68, label=paste(round(Frequency/ sum(Frequency) \* 100, 1), "%")),

position = position\_stack(vjust=0.5), size=5)+

theme(plot.title=element\_text(hjust=0.5),

axis.line=element\_blank(),

axis.text=element\_blank(),

axis.ticks=element\_blank(),

text = element\_text(size=12))+

labs(fill="Age Category", x=NULL, y=NULL,

title="Figure 7: Pie Chart of Age Category")+

scale\_fill\_manual(values=c("plum2", "plum3", "orchid3", "orchid4"),

labels=c("Millennial", "Late Gen X", "Early Gen X", "Baby Boomer"))+

coord\_polar("y")

## Figure 8: Bar Chart of Age Category

ord\_agecat = ordered(data$agecat, c("Millennial", "Late Gen X", "Early Gen X", "Baby Boomer"))

ggplot(data, aes(x=ord\_agecat))+

geom\_bar(fill=c("plum2", "plum3", "orchid3", "orchid4"))+

theme\_clean()+

theme(axis.text.x=element\_text(color="black", size=12),

axis.text.y=element\_text(color="black", size=12),

axis.title.x=element\_text(size=14),

axis.title.y=element\_text(size=14),

plot.title=element\_text(size=15, hjust=0.5))+

labs(title="Figure 8: Bar Chart of Age Category")+

xlab("Age Categories")+

ylab("Count")

# Question 4

## Table 3: Contingency Table for Gender by Smoke Status

GenSmoke\_table = table(data$Gender, data$SmokeStat)

GenSmoke\_table

## Table 4: Percentage of total for Gender by Smoke Status

round(prop.table(GenSmoke\_table),4)

## Table 5: Percentage of row for Gender by Smoke Status

round(prop.table(GenSmoke\_table, 1),4)

## Table 6: Percentage of column for Gender by Smoke Status

round(prop.table(GenSmoke\_table, 2),4)

# Question 5

## Figure 9: 100% Stacked Bar Chart of Smoke Status by Gender

### Created data frame to place into ggplot

prop\_GenSmoke = round(prop.table(GenSmoke\_table,2)\*100,2)

df\_GenSmoke = as.data.frame(prop\_GenSmoke)

names(df\_GenSmoke) = c("Gender", "Smoke\_Status", "Percentage")

ggplot(df\_GenSmoke, aes(x=Smoke\_Status, y=Percentage, fill=Gender, label=Percentage,))+

geom\_bar(stat = "identity", show.legend = TRUE)+

geom\_text(size=5, position=position\_stack(vjust=0.5))+

theme\_clean()+

theme(legend.position="right",

axis.text.x=element\_text(color="black", size=12),

axis.text.y=element\_text(color="black", size=12),

axis.title.x=element\_text(size=14),

axis.title.y=element\_text(size=14),

plot.title=element\_text(size=15, hjust=0.5, color="black"))+

scale\_fill\_manual(values = c("skyblue3", "plum3"),

labels = c("Male", "Female"))+

scale\_x\_discrete("Smoke Status", labels=c("Never", "Former", "Current Smoker"))+

ggtitle("Figure 9: 100% Stacked Bar Chart of Smoke Status by Gender")

# Question 6

### Defined the seed and gathered 80 observations from the main data set

set.seed(22577)

sample = data[sample(1:315,80,replace=F),]

### Defined the function to output Confidence Intervals

CI = function(x, alpha = .05, dec = 3){

n = sum(!is.na(x))

conf\_level = (1-alpha)\*100

me = qt(1-alpha/2, n-1)\*sd(x, na.rm=T)/sqrt(n)

mean = round(mean(x, na.rn = T), digits = dec)

lower = round(mean(x, na.rn = T) - me, digits = dec)

upper = round(mean(x, na.rn = T) + me, digits = dec)

limits = data.frame(cbind(variable = deparse(substitute(x)), n,

c.level = conf\_level, mean, me = round(me, digits = dec),

lower, upper))

print(limits)

rm(n, conf\_level, lower, upper, mean)}

## Table 7: 95% Confidence Interval for Calories Sample

CI(sample$Calories, 0.05, 2)

# Question 7

## Figure 10: Scatterplot of Calories by Age

ggplot(data, aes(x=Age, y=Calories))+

geom\_point()+

theme\_clean()+

theme(axis.text.x=element\_text(color="black", size=12),

axis.text.y=element\_text(color="black", size=12),

axis.title.x=element\_text(size=14),

axis.title.y=element\_text(size=14),

plot.title=element\_text(size=15, hjust=0.5))+

geom\_smooth(method=lm, se=F, linetype="solid", color="tomato1")+

ggtitle("Figure 10: Scatterplot of Calories by Age")

## Figure 11: Side-by-Side Boxplots of Calories by Age Category

ggplot(data, aes(x=Calories, y=agecat, fill=Calories))+

theme\_clean()+

theme(axis.text.x=element\_text(color="black", size=12),

axis.title.x=element\_text(size=14),

axis.title.y=element\_text(size=14),

plot.title=element\_text(size=15, hjust=0.5))+

geom\_boxplot(fill=c("orchid4", "orchid3", "plum3", "plum2"))+

ggtitle("Figure 11: Side-by-Side Boxplots of Calories by Age Category")+

ylab("Age Category")

### Categorization of calories to create 'calcat'

data$calcat[data$Calories<1200] = "Low"

data$calcat[data$Calories>=1200 & data$Age<2000] = "Average"

data$calcat[data$Calories>=2000 & data$Age<3000] = "Above Average"

data$calcat[data$Calories>=3000] = "High"

### Ordered the categories of calories

ord\_calcat = ordered(data$calcat, c("Low", "Average", "Above Average", "High"))

### Created a contingency table for Age Category by Calories Category

AgeCalories\_table = table(data$agecat, data$calcat)

### Ordered both the age categories and calories categories

col\_order = c("Low", "Average", "Above Average", "High")

row\_order = c("Millennial", "Late Gen X", "Early Gen X", "Baby Boomer")

AgeCalories\_table = AgeCalories\_table[row\_order, col\_order]

### Proportion table for Age Category by Calories Category

prop\_AgeCalories = round(prop.table(AgeCalories\_table,2)\*100,2)

### Created a data frame

df\_AgeCalories = as.data.frame(prop\_AgeCalories)

names(df\_AgeCalories) = c("Age\_Category", "Calories\_Consumption", "Percentage")

## Figure 12: 100% Stacked Bar Chart of Calories by Age

ggplot(df\_AgeCalories, aes(x=factor(Calories\_Consumption,

level=c("Low", "Average", "Above Average", "High")),

y=Percentage, fill=Age\_Category, label=Percentage,))+

geom\_bar(stat = "identity", show.legend = TRUE)+

geom\_text(size=4, position=position\_stack(vjust=0.5))+

theme\_clean()+

theme(axis.text.x=element\_text(color="black", size=12),

axis.text.y=element\_text(color="black", size=12),

axis.title.x=element\_text(size=14),

axis.title.y=element\_text(size=14),

plot.title=element\_text(size=15, hjust=0.5))+

theme(legend.position = "right")+

scale\_fill\_manual(values = c("plum2", "plum3", "orchid3", "orchid4"),

labels = c("Millennial", "Late Gen X", "Early Gen X", "Baby Boomer"))+

labs(fill="Age Category", x="Calories Category", y="Percentage",

title="Figure 12: 100% Stacked Bar Chart of Calories by Age")

## Table 8: Stratified Analysis of Calories by Age Category

strat\_analysis = aggregate(x=data$Calories, by=list(data$agecat), FUN=summary)

names(strat\_analysis) = c("Age Category", "Calories")

strat\_analysis

## Table 9: Contingency Table for Age Category by Calories

AgeCalories\_table

## Table 10: Percentage of total for Age Category by Calories

round(prop.table(AgeCalories\_table),4)

## Table 11: Percentage of row for Age Category by Calories

round(prop.table(AgeCalories\_table, 1),4)

## Table 12: Percentage of column for Age Category by Calories

round(prop.table(AgeCalories\_table, 2),4)

**Appendix 2: Python Script**

import pandas as pd

import numpy as np

data = pd.read\_csv(r'C:\Users\...\STAT 3010 Cancer Risk Dataset.csv)

# Question 1

## Table 1: Descriptive Statistics for Age, Calories, and Fat

print(round(data['Age'].describe(),2))

print(round(data['Calories'].describe(),2))

print(round(data['Fat'].describe(),2))

### Function to receive interquartile range for each variable

def IQR(variable):

q3, q1 = np.percentile(data[variable], [75 ,25])

iqr = q3 - q1

iqr = round(iqr,2)

return (iqr)

### Attached interquartile range to descriptive statistics

IQR('Age')

IQR('Calories')

IQR('Fat')