In [39]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from scipy import stats
from scipy.stats import zscore
```

Section 1

1. A statistics test was conducted for 10 learners in a class. The mean of their score is 85 and the variance of the score is zero. What can you interpret about the score obtained by all learners?

Answer ---> Mean score is 85 and the variance is zero. Variance zero means the values are not distrubuted. That means every student has same value.

—

2. In a residential locality, the mean size of the house is 2224 square feet and the median value of the house is 1500 square feet. What can you interpret about the skewness in the distribution of house size? Are there bigger or smaller houses in the residential locality?

Answer ---> Thats indication of skewness in data or we can say it is possible that data has outliers. by respect to mean and median relationship, Mean is higher than median thats because locality has bigger houses and data is Right skewed.

—

3. The following table shows the mean and variance of the expenditure for two groups of people.You want to compare the variability in expenditure for both groups with respect to their mean. Which statistical measure would you use to evaluate the variability in expenditure? Please provide an explanation for your answer.

Answer --> To compare the variability of two data sets we have to compare there Coefficient of Variation.

To compare the variability in expenditure for both groups with respect to their mean, you should use the Coefficient of Variation (CV). The Coefficient of Variation is a standardized measure of dispersion of a probability distribution or frequency distribution. It is useful because it allows comparison of the degree of variation from one data series to another, even if the means are drastically different.

Coefficient of Variation (CV) The Coefficient of Variation is a standardized measure of dispersion/distribution. The Coefficient of Variation is calculated using the following formula:

CV = σ/μ

Group I -: CV = 125000/500000 = 0.25 Group II -: CV = 40000/10000 = 0.25

Conclusion -

1. This means that the relative variability in expenditure for both groups with resp to their mean is the same.
2. Fluctuation in both the data sets are the same resp to each other.

—

4. During the survey, the ages of 80 patients infected by COVID and admitted to one of the city hospitals were recorded and the collected data is represented in the less than cumulative frequency distribution table.

a. Which class interval has the highest frequency? Answer --> The class interval 35 to 45 has the highest frequency with 23 patients.

b. Which age was affected the least? Answer--> The class interval 55 to 65 has the least frequency with 5 patients.

c. How many patients aged 45 years and above were admitted?

Answer --> The total number of patients aged 45 and above is 14+5=19.

d. Which is the modal class interval in the above dataset Answer --> Modal Class means - Class which has highest occurance hence 35 - 45 with 23 patients is Modal Class of the data.

e. What is the median class interval of age?

Answer - Median=L+((N/2 - CF)/f)×h:

L = Lower boundary of the median class N = Total number of observations CF = Cumulative frequency of the class preceding the median class f = Frequency of the median class h = Class width (interval size):

L=35, CF=38, N=80, f=23, h=10: **The median age of the patients is approximately 35.25 years. Therefore, the median lies in the interval 35 to 45.**

5. Assume you are the trader and you have invested over the years, and you are worried about the average return on investment. What average method would you use to compute the average return for the data given below?

In [31]:
```
Table = { "Year" :[2015,2016,2017,2018,2019,2020],
          "Return" :[36,23,-48,-30,15,31],
          "Asset Price":[5000,6400,7890,9023,4567,3890]
                  }
df = pd.DataFrame(Table)

df['Adjusted Return'] = 1 + df['Return'] / 100

product = df['Adjusted Return'].prod()

n = len(df)

gmean = product ** (1/n)

average_return = (gmean -1) *100

average_return.round(2) #in%
```

Out[31]: -1.43

6. Suppose you have been told to measure the average height of all the males on the earth. What would be your strategy for the same? Would the average height be a parameter or a statistic? Justify your answer.

Answer -->

1. **Sampling**: Due to the impracticality of measuring the height of every male on Earth, you would need to select a representative sample. A representative sample should ideally cover different geographic regions, age groups, ethnicities, socioeconomic statuses, etc. This helps ensure that the sample reflects the diversity within the global male population.
2. **Data Collection**: Measure the height of each male in your sample using standardized methods. Height measurements are typically taken with individuals standing barefoot on a flat surface against a vertical measuring tool.
3. **Calculate the Sample Mean**: Once you have collected height measurements from your sample, calculate the average (mean) height. This is done by summing all the heights and dividing by the number of measurements
4. **Statistical Analysis**: Use statistical methods to analyze your data. Calculate measures of central tendency (mean, median) and measures of dispersion (standard deviation) to describe the average height and variability within your sample.
5. **Inference to Population**: Recognize that the average height obtained from your sample is an estimate of the average height of all males on Earth. The accuracy of this estimate depends on the representativeness and size of your sample.

7. Calculate the z score of the following numbers: X = [4.5,6.2,7.3,9.1,10.4,11]

Answer -->

Z Score indicates the deviation of data point is from the mean, positive value shows how many standard deviations a data point is above the mean & negative value shows how many standard deviations a data point is below the mean. **Z-score = x - mean / std.dev**

In [35]:
```
X = [4.5,6.2,7.3,9.1,10.4,11]
z_scores = stats.zscore(X)
print("Z-scores:", z_scores)
```

```
Z-scores: [-1.55385602 -0.81667781 -0.33968015  0.44086148  1.00458598  1.26
476653]
```

Section 2

In [3]:
```
df = pd.read_csv(r"C:\Users\aksha\OneDrive\Desktop\Bank Personal Loan Modellir
df
```

Out[3]:

	ID	Age	Experience	Income	ZIP Code	Family	CCAvg	Education	Mortgage	Personal Loan	S
0	1	25	1	49	91107	4	1.6	1	0	0	
1	2	45	19	34	90089	3	1.5	1	0	0	
2	3	39	15	11	94720	1	1.0	1	0	0	
3	4	35	9	100	94112	1	2.7	2	0	0	
4	5	35	8	45	91330	4	1.0	2	0	0	
...	...	...	...	...	...	...	...	...	...	...	
4995	4996	29	3	40	92697	1	1.9	3	0	0	
4996	4997	30	4	15	92037	4	0.4	1	85	0	
4997	4998	63	39	24	93023	2	0.3	3	0	0	
4998	4999	65	40	49	90034	3	0.5	2	0	0	
4999	5000	28	4	83	92612	3	0.8	1	0	0	

5000 rows × 14 columns

8. Give us the statistical summary for all the variables in the dataset.

In [4]:
```
df.describe(include='all').round(2)
```

Out[4]:

	ID	Age	Experience	Income	ZIP Code	Family	CCAvg	Education	Mortgage
count	5000.00	5000.00	5000.00	5000.00	5000.00	5000.00	5000.00	5000.00	5000.00
mean	2500.50	45.34	20.10	73.77	93152.50	2.40	1.94	1.88	56.50
std	1443.52	11.46	11.47	46.03	2121.85	1.15	1.75	0.84	101.71
min	1.00	23.00	-3.00	8.00	9307.00	1.00	0.00	1.00	0.00
25%	1250.75	35.00	10.00	39.00	91911.00	1.00	0.70	1.00	0.00
50%	2500.50	45.00	20.00	64.00	93437.00	2.00	1.50	2.00	0.00
75%	3750.25	55.00	30.00	98.00	94608.00	3.00	2.50	3.00	101.00
max	5000.00	67.00	43.00	224.00	96651.00	4.00	10.00	3.00	635.00

9. Evaluate the measures of central tendency and measures of dispersion for all the quantitative variables in the dataset.

Measures of Central Tendency

1. Mean
2. Median
3. Mode

Measures of Dispersion

1. Range
2. Interquartile Range (IQR)
3. Variance
4. Standard Deviation

In [33]:
```
df1 = df[["Age","Experience","Income","CCAvg"]]
```

In [110]:
```
# **Measures of Central Tendency**
# 1) Mean
Mean = df1.mean()
df_mean = pd.DataFrame(Mean,columns=['Mean'])
```

In [111]:
```
# 2) Median
Median = df1.median()
df_median = pd.DataFrame(Median,columns=['Median'])
```

In [113]:
```
# 3) Mode
mode = df1.mode().iloc[0]
df_mode = pd.DataFrame(mode).rename(columns={0: 'Mode'})
```

In [121]:
```
Central_Tendency = pd.concat([df_mean,df_median,df_mode],axis=1)
Central_Tendency
```

Out[121]:

	Mean	Median	Mode
Age	45.338400	45.0	35.0
Experience	20.104600	20.0	32.0
Income	73.774200	64.0	44.0
CCAvg	1.937938	1.5	0.3

In [120]:
```
# **Measures of Dispersion**
# 1) Range
range = df1.max() - df1.min()
RANGE = pd.DataFrame(range,columns=["Range"])
```

In [117]:
```
# 2) IQR
iqr = df1.quantile(0.75) - df1.quantile(0.25)
IQR = pd.DataFrame(iqr,columns=["Iqr"])
```

In [118]:
```
# 3) Variance
var = df1.var().round(2)
VAR = pd.DataFrame(VAR, columns=["Var"])
```

In [119]:
```
# 4) Std.Dev
std_Dv= df1.std().round(2)
STD_DV = pd.DataFrame(std_Dv,columns=["Std_DeV"])
```

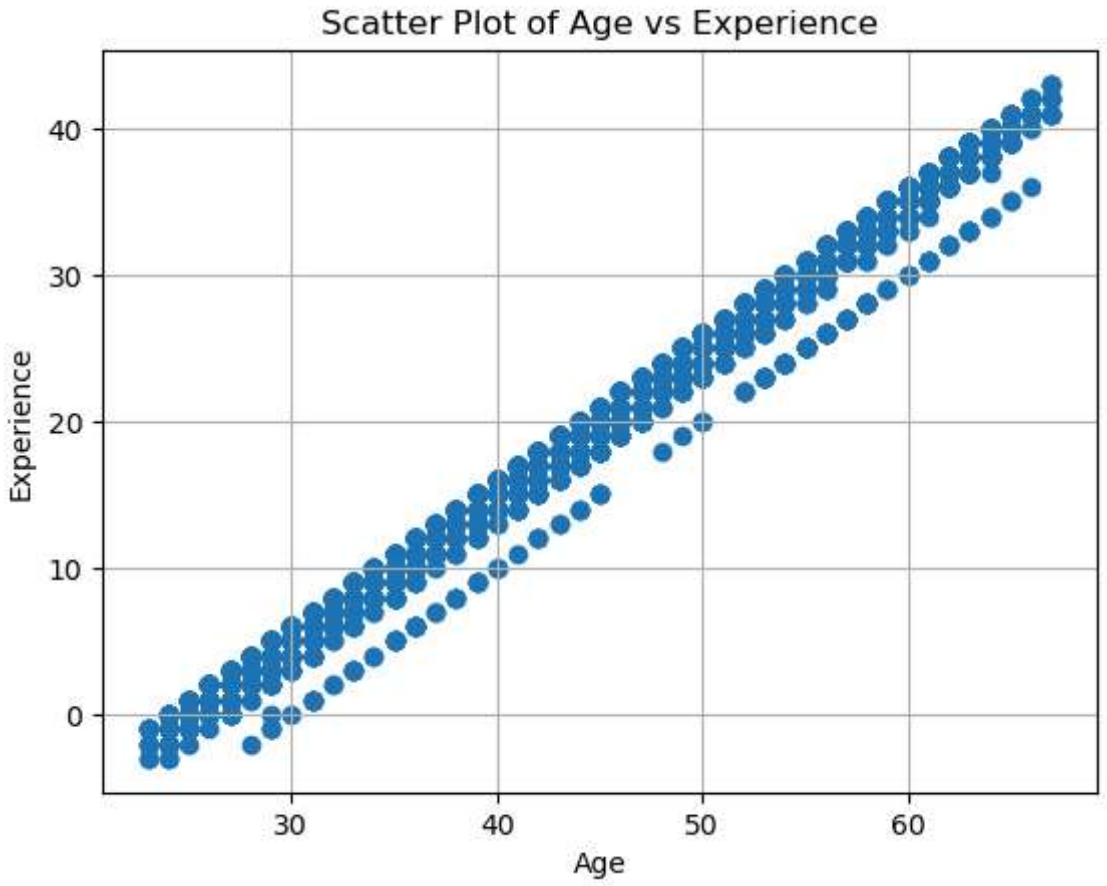
In [124]:
```
Dispersion = pd.concat([RANGE,IQR,VAR,STD_DV],axis = 1)
Dispersion
```

Out[124]:

	RANGE	IQR	VAR	STD_DV
Age	44.0	20.0	131.40	11.46
Experience	46.0	20.0	131.51	11.47
Income	216.0	59.0	2119.10	46.03
CCAvg	10.0	1.8	3.05	1.75

10. What statistical method will you use to examine the presence of a linear relationship between age and experience variables? Also, create a plot to illustrate this relationship.

In [139]:
```
df1 = df[["Age","Experience"]]
plt.scatter(df1['Age'], df1['Experience'],)
plt.title('Scatter Plot of Age vs Experience')
plt.xlabel('Age')
plt.ylabel('Experience')
plt.grid(True)
plt.show()
```



Plot Observation : **There is a POSITIVE LINEAR RELATIONSHIP between Age and Experience. As Age increases, Experience generally increases as well.**

11. What is the most frequent family size observed in this dataset?

In [182]:
```
family_mode = df["Family"].mode()
print("Most_Frequent_family_size : ",family_mode.iloc[0])
```

```
Most_Frequent_family_size :  1
```

12. What is the percentage of variation you can observe in the 'Income' variable?

In [183]:
```
mean = df["Income"].mean()
print("Mean:",mean)

std_dev = df["Income"].std()
print("Std_dev",std_dev)
CV = (std_dev/mean)*100
print("Percentage of variation",CV.round(2),"%")
```

```
Mean: 73.7742
Std_dev 46.033729321086334
Percentage of variation 62.4 %
```

13. The 'Mortgage' variable has a lot of zeroes. Impute with some business logical value that you feel fit for the data.

In [216]:
```
df2 = df
df2_median = df2[df2["Mortgage"]>0]["Mortgage"].median()
df2_median
```

Out[216]:
```
153.0
```

In [223]:
```
df2["Mortgage"]=df2["Mortgage"].replace(0,"No Mortgage")
df2
```

Out[223]:

	ID	Age	Experience	Income	ZIP Code	Family	CCAvg	Education	Mortgage	Personal Loan	S
0	1	25	1	49	91107	4	1.6	1	NO Mortgage	0	
1	2	45	19	34	90089	3	1.5	1	NO Mortgage	0	
2	3	39	15	11	94720	1	1.0	1	NO Mortgage	0	
3	4	35	9	100	94112	1	2.7	2	NO Mortgage	0	
4	5	35	8	45	91330	4	1.0	2	NO Mortgage	0	
...	...	...	...	...	...	...	...	...	...	...	
4995	4996	29	3	40	92697	1	1.9	3	NO Mortgage	0	
4996	4997	30	4	15	92037	4	0.4	1	85	0	
4997	4998	63	39	24	93023	2	0.3	3	NO Mortgage	0	
4998	4999	65	40	49	90034	3	0.5	2	NO Mortgage	0	
4999	5000	28	4	83	92612	3	0.8	1	NO Mortgage	0	

5000 rows × 14 columns

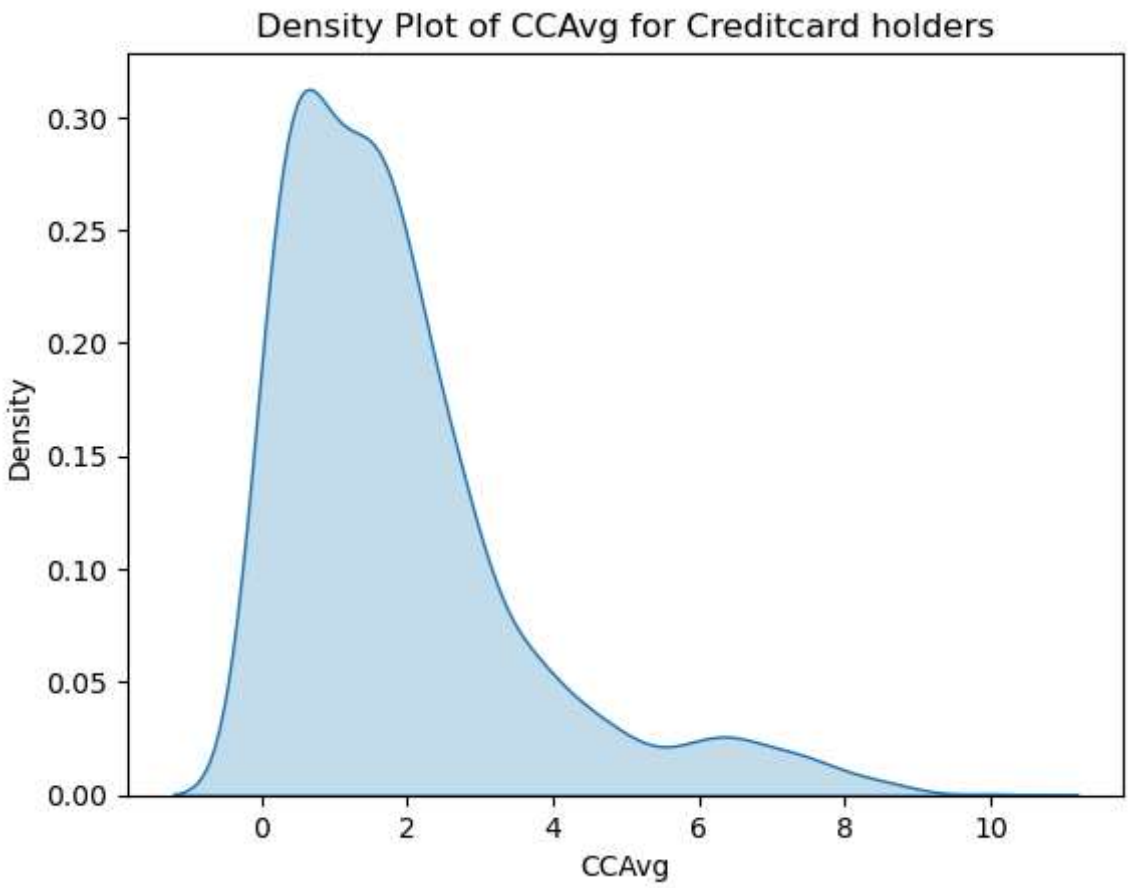
14. Plot a density curve of the CCAvg variable for the customers who possess credit cards and write an interpretation about its distribution.

In [245]:
```
df3 = df[df["CreditCard"]>0]
sns.kdeplot(df3['CCAvg'],shade=True)
plt.title('Density Plot of CCAvg for Creditcard holders')
plt.xlabel('CCAvg')
plt.ylabel('Density')
plt.show()
```

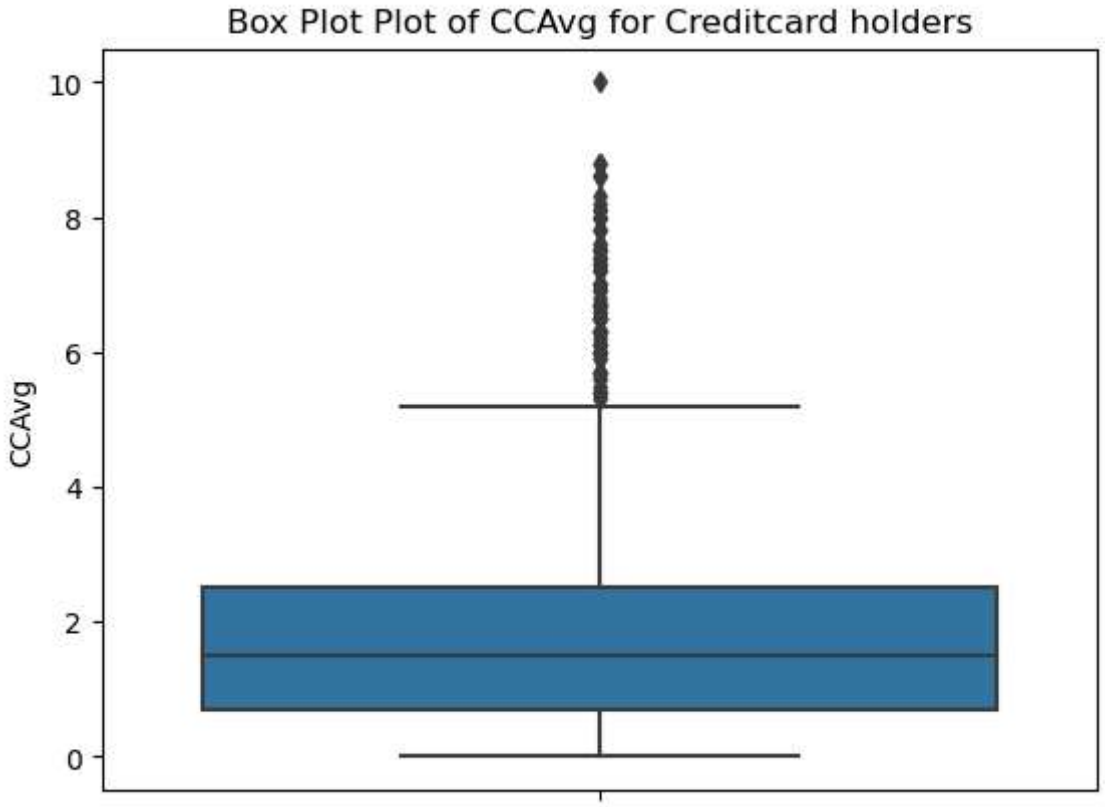
```
C:\Users\aksha\AppData\Local\Temp\ipykernel_20524\2777207839.py:2: FutureWar
ning:

`shade` is now deprecated in favor of `fill`; setting `fill=True`.
This will become an error in seaborn v0.14.0; please update your code.

  sns.kdeplot(df3['CCAvg'],shade=True)
```



In [277]:
```
sns.boxplot(y=df3['CCAvg'])
plt.title('Box Plot Plot of CCAvg for Creditcard holders')
plt.ylabel('CCAvg')
plt.show()
```



Right Skewed Distribution:

1. Mean spending high than median spending that means Majority customers spend less than average spendings. but few of them spends exceptionally high that gives outlers to the data. and pull mean towords right.

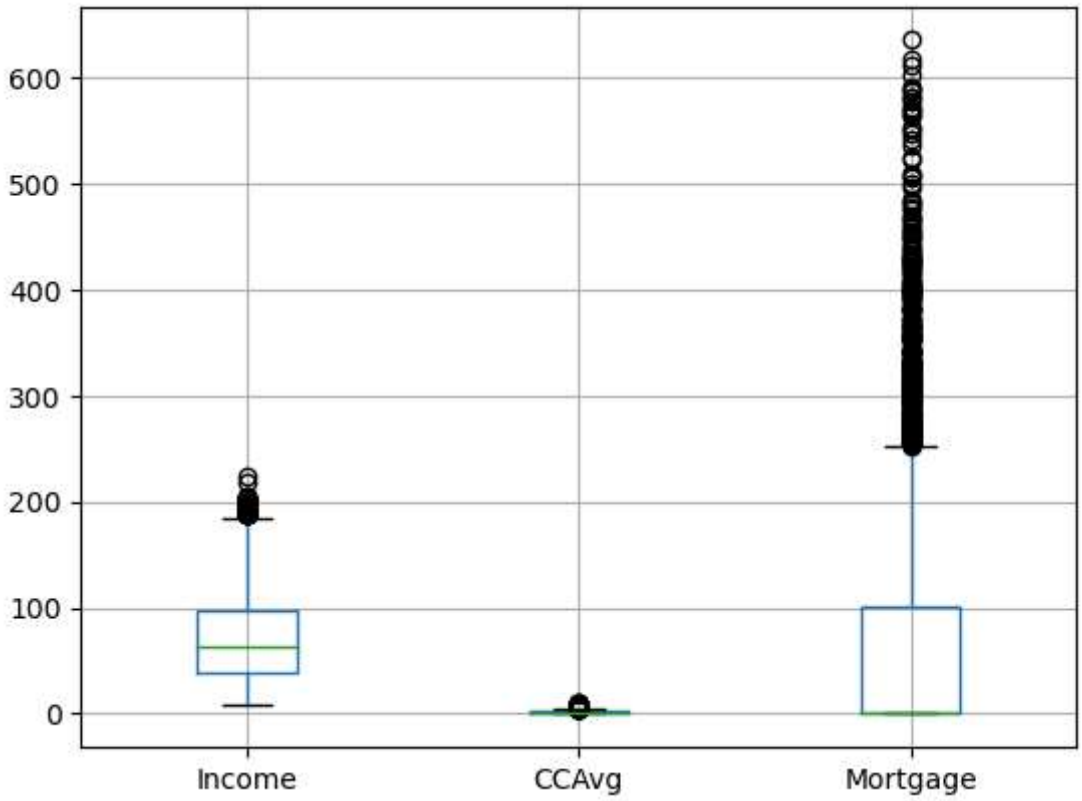
Conclusion - This indicates different catogories of customers having different lifestyle , spending behaviour and purchasing capabilities. so in this data outliers also very important, because they are potential buyers.

15. Do you see any outliers in the dataset? If yes, what plot you would think will be suitable to showcase to the stakeholders?

Answer - Yes, I have seen outliers in this dataset and i would be prefer to use boxplot for showcase to stakeholders.

In [37]:
```
df.boxplot(column=["Income","CCAvg","Mortgage"])
```

Out[37]: <Axes: >



16. Give us the decile values of the variable 'Income' in the dataset.

In [279]:
```
Deciles = [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9,1]
decile_values = df['Income'].quantile(deciles)
print(decile_values)
```

```
0.1     22.0
0.2     33.0
0.3     42.0
0.4     52.0
0.5     64.0
0.6     78.0
0.7     88.3
0.8    113.0
0.9    145.0
1.0    224.0
Name: Income, dtype: float64
```

17. Give the IQR of all the variables which are quantitative and continuous.

In [298]:
```
# IQR
df4 = df[["Age","Experience","Income","CCAvg","Mortgage"]]
iqr = df4.quantile(0.75)-df4.quantile(0.25)
IQR = pd.DataFrame(iqr,columns=["IQR"])
IQR
```

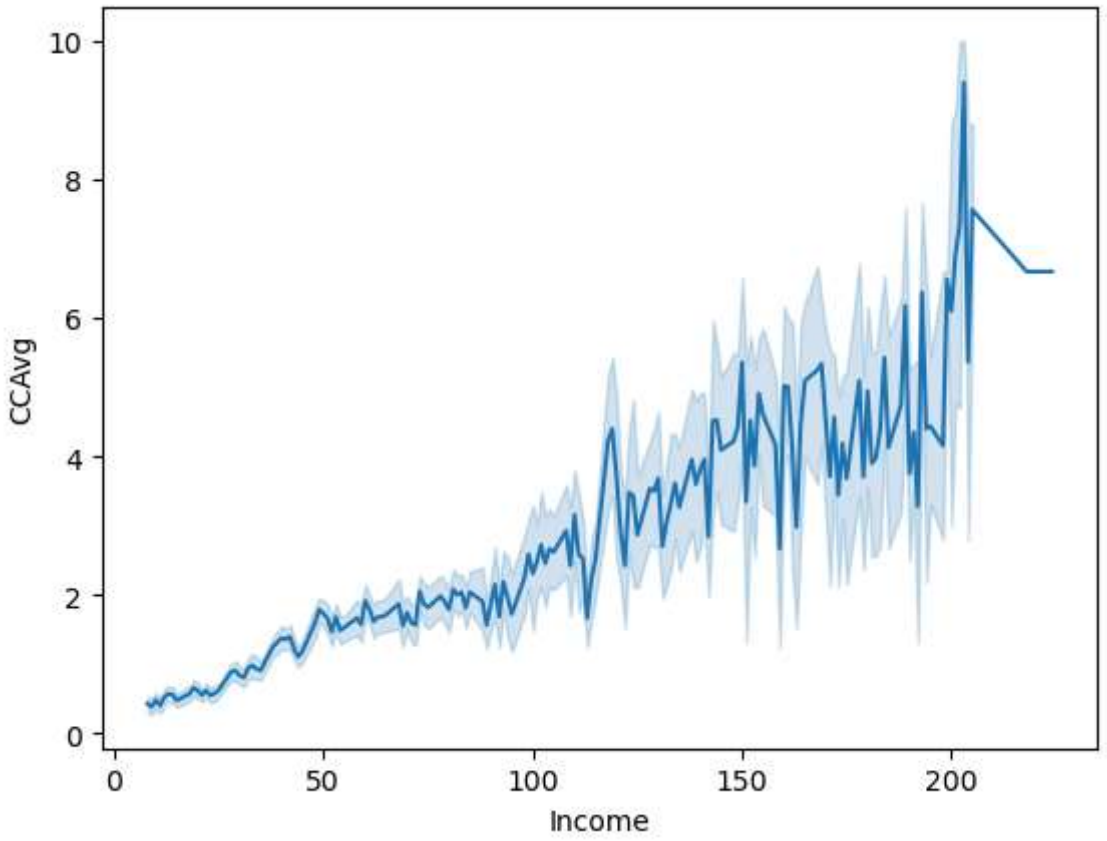
Out[298]:

	IQR
Age	20.0
Experience	20.0
Income	59.0
CCAvg	1.8
Mortgage	101.0

18. Do the higher-income holders spend more on credit cards?

In [314]:
```
df5= df[["Income","CCAvg"]]
sns.lineplot(x='Income', y='CCAvg', data=df5)
```

Out[314]:
```
<Axes: xlabel='Income', ylabel='CCAvg'>
```



The line suggests a positive relationship between income and credit card average spending. Higher-income individuals tend to have higher average credit card spending

19. How many customers use online banking? Do customers using bank internet facilities have higher incomes?

In [33]:
```
df6= df[["Income","Online"]]
df6 = df6[df6["Online"]>0]
print("No of customers uses internet banking:",df6["Online"].value_counts())
```

```
No of customers uses internet banking: Online
1    2984
Name: count, dtype: int64
```

In [29]:
```
df_online = df[df['Online'] == 1]
print("Average_income of online customers :-",round(df_online["Income"].mean()

df_offline = df[df['Online'] == 0]
print("Average_income of offline customers :-", round(df_offline["Income"].mea
```

```
Average_income of online customers :- 74
Average_income of offline customers :- 73
```

Hence, customers using bank internet facilities have slightly higher incomes, but its not that much conclusive as the difference between the means are very least.

20. Using the z-score of the income variable, find out the number of observations outside the +-3σ.

In [40]:
```
df["Zscore"] = zscore(df['Income'])
```

In [62]:
```
print((df["Zscore"]<-3).value_counts())
print((df["Zscore"]>+3).value_counts())
```

```
Zscore
False    5000
Name: count, dtype: int64
Zscore
False    4998
True        2
Name: count, dtype: int64
```

Conclusion --> In a normal distribution, approximately 99.7% of the data falls within ±3 standard deviations from the mean (68% within ±1σ, 95% within ±2σ, and 99.7% within ±3σ). Therefore, a z-score greater than +3 indicates that, The 2 data point lies in the extreme right tail of the distribution

In [ ]:
```
```