**Analysis for the project – Q&A’s as per code:**

Summary:

Methods used: Descriptive Statistics, Data Manipulation, Missing Data Handling, Data Scaling, Outlier Detection, Data Visualization, Exploratory Data Analysis (EDA)

1. What is the median income of the dataset? Enter values round up to two decimal points.

* 65,331.50

2. How many unique job types are present in the dataset? Enter values round up to two decimal points.

* 4.00

3. Calculate the average number of dependents per customer. Enter values round up to two decimal points.

* 2.20

4. What is the maximum credit score recorded in the dataset? Enter values round up to two decimal points.

* 758.55

5. Count the number of customers who have a car, enter values round up to two decimal points.

* 104.00

6. What is the total number of missing values in the 'Years in Current Job' column? Enter values round up to two decimal points.

* 20.00

7. Find the standard deviation of the 'Annual Spending' for the dataset. Enter values round up to two decimal points.

* 9,833.23

8. How many customers are classified as 'High' in customer satisfaction?

* 63

9. What is the mode of the 'Education Level' in the dataset?

* Master

10. Calculate the average age of customers who do not have a car. Enter integer value

* 40

11. Which column has the highest number of outliers?

* Income

12. What is the most common type of job among the customers?

* Part-time

13. Which variable is most likely to have a normal distribution?

* Cred Score

14. What percentage of the dataset is missing the 'Income' data?

* 10

15. Which age group shows the highest median income?

* 61-75

16. Explain why the 'Credit Score' might (or might not) have a normal distribution.

* Reasons 'Credit Score' Could Seem Normally Distributed:  
  1. Visual examinations of the 'Credit Score' graph suggest a bell-shaped curve, indicating a normal distribution. Also, skewness is close to 0 (0.19)  
  2. The Central Limit Theorem implies that averaging many samples from any data set will generally produce a bell-shaped curve, regardless of the original data's shape.  
  3. Credit scores are determined from various financial behaviors. Combining these diverse factors, each affected by many small, random variables, might result in a normal-like distribution.  
    
  Reasons 'Credit Score' May Not Be Normally Distributed:  
  1. Credit scores are confined within a specified range, preventing the distribution from extending indefinitely, which is a characteristic of a true normal distribution.  
  2. The behavior of the credit-holding population might not be evenly distributed around the mean.   
  3. There could be more individuals with either exceptionally high or low scores.

17. Discuss the potential impact of outliers in the 'Income' column on data analysis.

* Min and Max of Income are: 30,093, 249,363. The difference: 219,270. Mean of the Income column is 69,852.  
  There seems to be a large amount of difference in the highest and the lowest incomes, the provided income values may or may not be real. In our data, the highest income holder is a part-time employee who does not have a college degree and earns 249,363 at the age of 44. The lowest income holder is a also a part-time employee, has a Masters degree and earns but earns 30,093 at the age of 58. And the annual spendings are 34,922 and 10,488 respectively.  
    
  Impact of outliers on the 'Income' Column:  
  1. Outliers can throw off average income calculations and make it seem like everyone earns more or less than they actually do.  
  2. These extreme values can lead to incorrect predictions or conclusions about trends in our data also in regression models, outliers can leverage points that disproportionately influence the model’s fit and predictions.  
  3. Outliers can skew the mean and inflate the standard deviation, providing a misleading representation of the central tendency and dispersion in the data.

18. Why would it be important to fill in missing values in the 'Years in Current Job' column before performing predictive modeling?

* There are 24 missing entries in the 'Years in Current Job' column out of a total of 200. These missing entries appear to be randomly distributed across various categories like age groups, job types, and educational levels, suggesting no evident pattern. As such, these missing values are unlikely to introduce significant bias.

Completing the 'Years in Current Job' data before conducting predictive modeling enhances the model’s effectiveness and ensures the reliability and accuracy of the insights derived from our analysis

19. Reason why one-hot encoding is beneficial for handling the 'Job Type' categorical data in machine learning models.

* One-hot encoding is particularly beneficial for handling categorical data like 'Job Type' in machine learning models by converting catergorical values into numericals. Which when perfromed analysis on would bring out relations that did not exist while the data was in categorical form. Also most machine learning algorithms require numerical inputs. In a general scenario categorical values must be converted into a numerical format and then analysis should be performed, which results in usage or more memory  and wastage of time. Hence, one-hot coding is a critical preprocessing step for achieving optimal model performance when dealing with categorical inputs.

20. Suggest why using the median, rather than the mean, might be more appropriate for imputing missing values in 'Income'.

* As discussed earlier, Min and Max of Income are: 30,093, 249,363. The difference: 219,270. Mean of the Income column is 69,852 and the median: 65,331.5.  
  In our dataset, using the median, rather than the mean, for imputing missing values in 'Income' wouldn't have shown much difference since the difference in the mean and median is nto significant(4520).  
  However, in general scenarios, using median would be a better choice as it is not usually affected by the outliers as much as mean would be. The median, being a positional measure, does not assume any distribution and thus is safer for skewed data like income when compared to mean which assumes data symmetry around the central point.

21. Replace the missing values in income with median, education level with mode, years in current job with mode, what is the standard deviation of income? Enter values round up to two decimal points.

* 34,128.87

22. Using minmax scaler scale the dataset and what is the new standard deviation of income? Enter values round up to two decimal points.

* 0.16