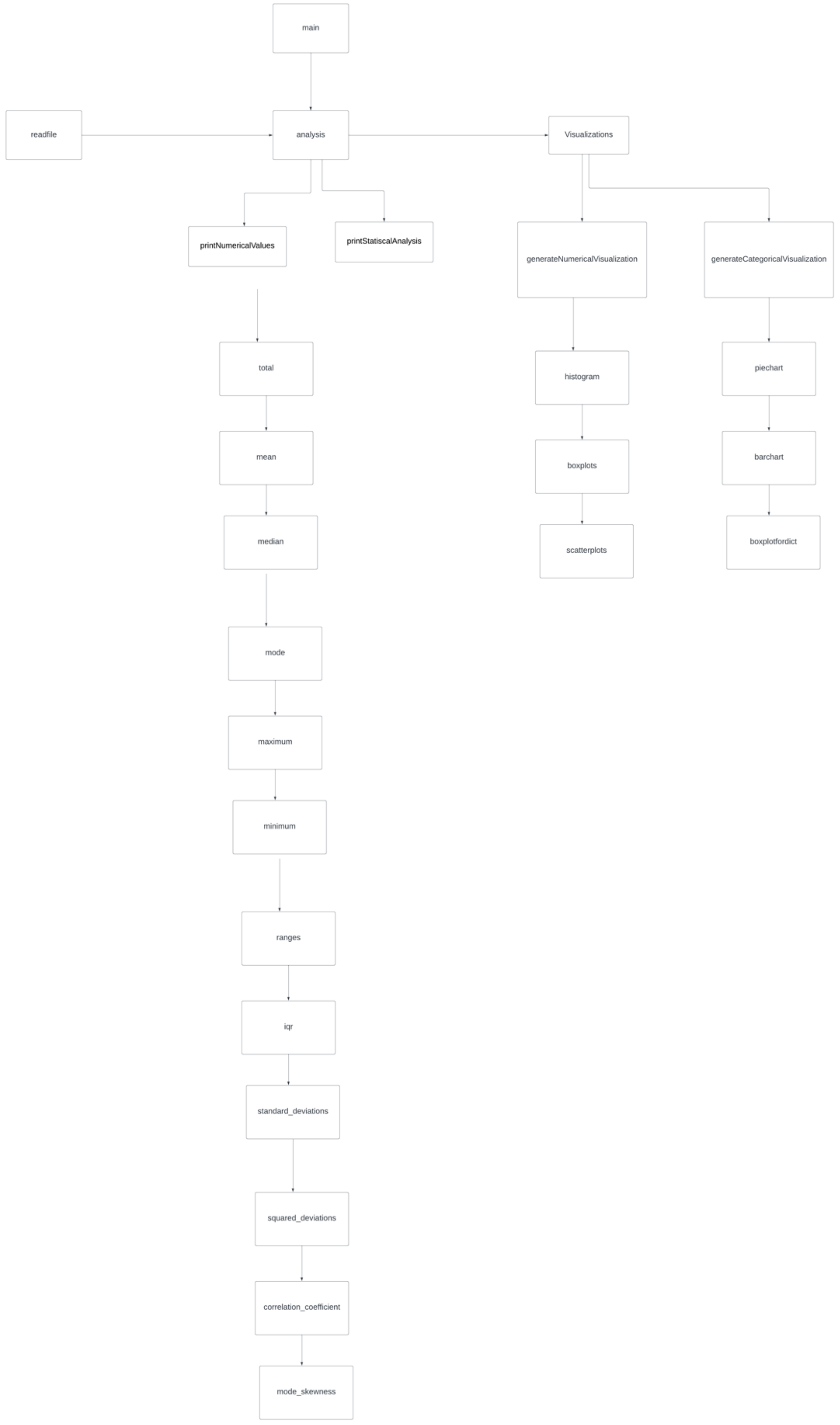
Data Analysis Report

Contents

Program Design: Structure Chart

A Structure Chart showing the Top-Down Modular Design of your program.



Unit Testing

Include a Test Case for each User-Defined Statistics Function

Test Case ID	1.
Function Tested	total()
Test Case Description	Determine the mean (average) of a list of numbers.
Test Data (Inputs)	[5,5,4,3,1,5,6]
Expected Results	29
Actual Results	29
Pass/Fail	Pass

Test Case ID	2.
Function Tested	mean()
Test Case Description	Determine the mean (average) of a list of numbers.
Test Data (Inputs)	[5,5,4,3,1,5,6]
Expected Results	4.14
Actual Results	4.14
Pass/Fail	Pass

Test Case ID	3.
Function Tested	median()
Test Case Description	Determine the median of a list of numbers
Test Data (Inputs)	[5,5,4,3,1,5,6]
Expected Results	5
Actual Results	5
Pass/Fail	Pass

Test Case ID	4.
Function Tested	mode()
Test Case Description	Determine the mode of a list of numbers.
Test Data (Inputs)	[5,5,4,3,1,5,6]
Expected Results	5
Actual Results	5
Pass/Fail	Pass

Test Case ID	5.
Function Tested	ranges()
Test Case Description	Determine the range of the lists of numbers.
Test Data (Inputs)	[5,5,4,3,1,5,6]
Expected Results	5
Actual Results	5
Pass/Fail	Pass

Test Case ID	6.
Function Tested	Iqr()
Test Case Description	Determine the interquartile range (IQR) of a list of numbers.
Test Data (Inputs)	[5,5,4,3,1,5,6]
Expected Results	2
Actual Results	2
Pass/Fail	Pass

Test Case ID	7.
Function Tested	standard_dev()
Test Case Description	Determine the standard deviation of a list of numbers.
Test Data (Inputs)	[5,5,4,3,1,5,6]
Expected Results	1.68
Actual Results	1.68
Pass/Fail	Pass

Test Case ID	8.
Function Tested	squared_dev()
Test Case Description	Determine the squared deviation of a list of numbers.
Test Data (Inputs)	[5,5,4,3,1,5,6]
Expected Results	0.7396000000000006, 0.7396000000000006, 0.01959999999999991, 1.2995999999999992, 9.859599999999999, 0.7396000000000006, 3.4596000000000013]

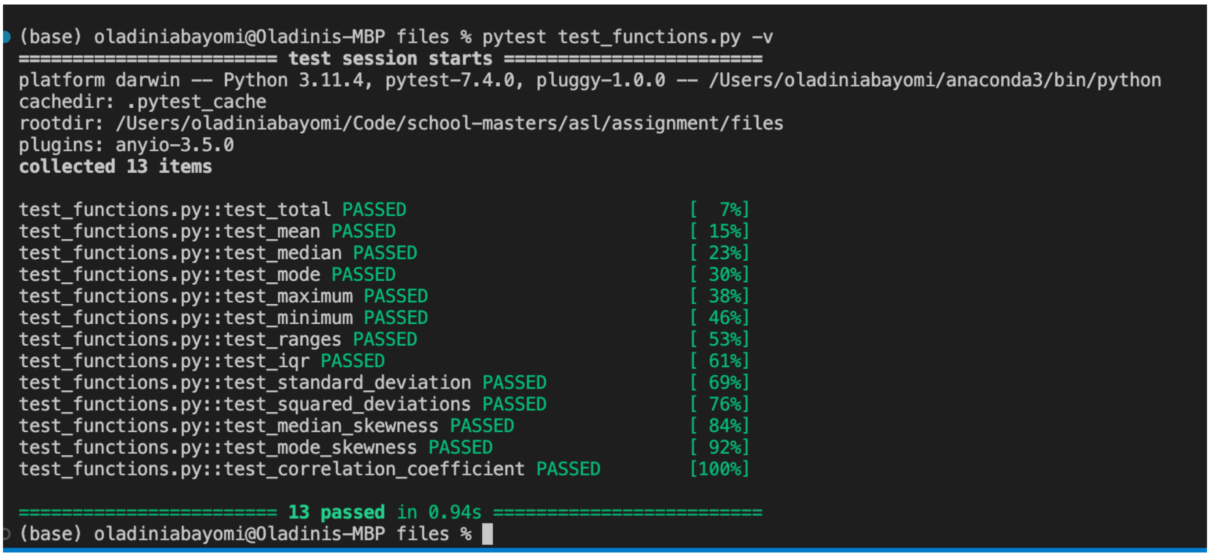
Actual Results	0.7396000000000006, 0.7396000000000006, 0.019599999999999991, 1.2995999999999992, 9.859599999999999, 0.7396000000000006, 3.4596000000000013]
Pass/Fail	Pass

Test Case ID	9.
Function Tested	median_skewness()
Test Case Description	Determine the median skewness of a list of numbers.
Test Data (Inputs)	[5,5,4,3,1,5,6]
Expected Results	-1.54
Actual Results	-1.54
Pass/Fail	Pass

Test Case ID	10.
Function Tested	mode_skewness()
Test Case Description	Determine the mode skewness of a list of numbers.
Test Data (Inputs)	[5,5,4,3,1,5,6]
Expected Results	0.51
Actual Results	0.51
Pass/Fail	Pass

Test Case ID	11.
Function Tested	correlation_coefficient()
Test Case Description	Determine the standard deviation of a list of numbers.
Test Data (Inputs)	[5,5,4,3,1,5,6] [1,3,2,1,1,2,1]
Expected Results	0.3
Actual Results	0.3
Pass/Fail	Pass

b) Screenshot of the PyTest output in Verbose Mode



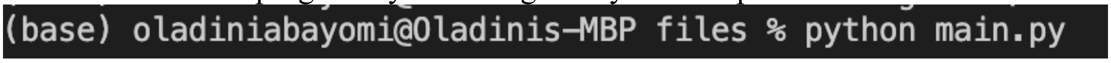
```
(base) oladiniabayomi@Oladinis-MBP files % pytest test_functions.py -v
======================== test session starts ========================
platform darwin -- Python 3.11.4, pytest-7.4.0, pluggy-1.0.0 -- /Users/oladiniabayomi/anaconda3/bin/python
cachedir: .pytest_cache
rootdir: /Users/oladiniabayomi/Code/school-masters/asl/assignment/files
plugins: anyio-3.5.0
collected 13 items

test_functions.py::test_total PASSED                               [  7%]
test_functions.py::test_mean PASSED                                [ 15%]
test_functions.py::test_median PASSED                              [ 23%]
test_functions.py::test_mode PASSED                                [ 30%]
test_functions.py::test_maximum PASSED                             [ 38%]
test_functions.py::test_minimum PASSED                             [ 46%]
test_functions.py::test_ranges PASSED                              [ 53%]
test_functions.py::test_iqr PASSED                                 [ 61%]
test_functions.py::test_standard_deviation PASSED                  [ 69%]
test_functions.py::test_squared_deviations PASSED                  [ 76%]
test_functions.py::test_median_skewness PASSED                     [ 84%]
test_functions.py::test_mode_skewness PASSED                       [ 92%]
test_functions.py::test_correlation_coefficient PASSED             [100%]

======================== 13 passed in 0.94s ========================
(base) oladiniabayomi@Oladinis-MBP files %
```

User Manual

Describe in detail the program's *menu system* using output screenshots with appropriate descriptions.

User Manual for Exploratory Data Analysis
Introduction
This program is built to analyse and visualise datasets.
To Install
Install Python 3.
To Execute

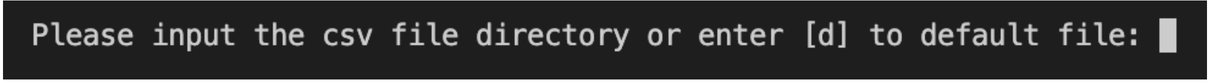
1. Open the Program
    a. Run the program by executing the Python script.



```
(base) oladiniabayomi@Oladinis-MBP files % python main.py
```

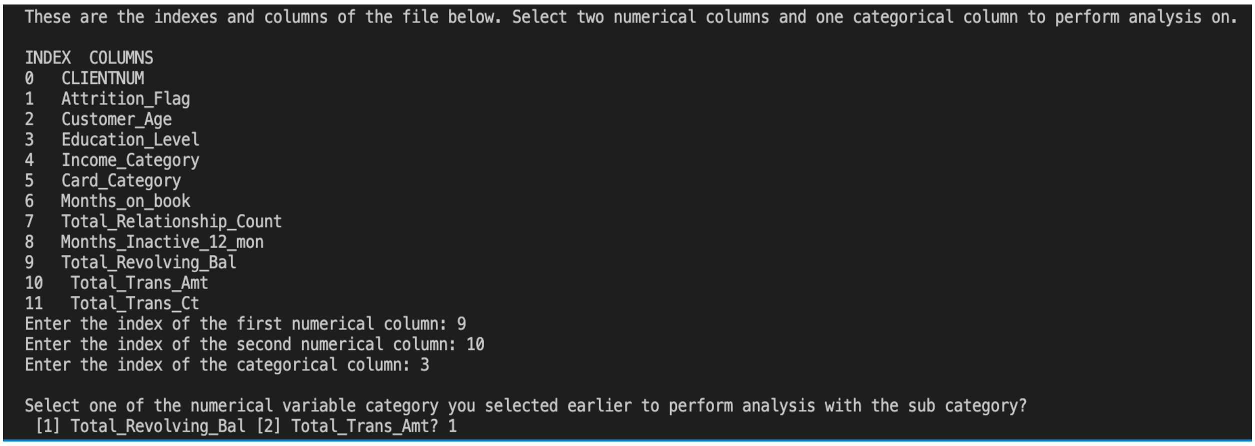
    b. Keep the files (including CSV data) are in the same directory.

2. Input
    a. The program request for a csv. You can select the default file by entering 'd'.
    b. If using a different file, enter the path. The program validates if the input is a CSV file.
    c. Enter the separator you are using



```
Please input the csv file directory or enter [d] to default file:
```

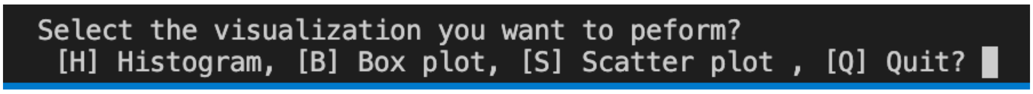
3. Column Selection
    a. The program displays a list of column codes and headers from the CSV.
    b. Choose two columns for statistical analysis by entering their respective code numbers.



```
These are the indexes and columns of the file below. Select two numerical columns and one categorical column to perform analysis on.

INDEX  COLUMNS
0   CLIENTNUM
1   Attrition_Flag
2   Customer_Age
3   Education_Level
4   Income_Category
5   Card_Category
6   Months_on_book
7   Total_Relationship_Count
8   Months_Inactive_12_mon
9   Total_Revolving_Bal
10   Total_Trans_Amt
11   Total_Trans_Ct
Enter the index of the first numerical column: 9
Enter the index of the second numerical column: 10
Enter the index of the categorical column: 3

Select one of the numerical variable category you selected earlier to perform analysis with the sub category?
 [1] Total_Revolving_Bal [2] Total_Trans_Amt? 1
```

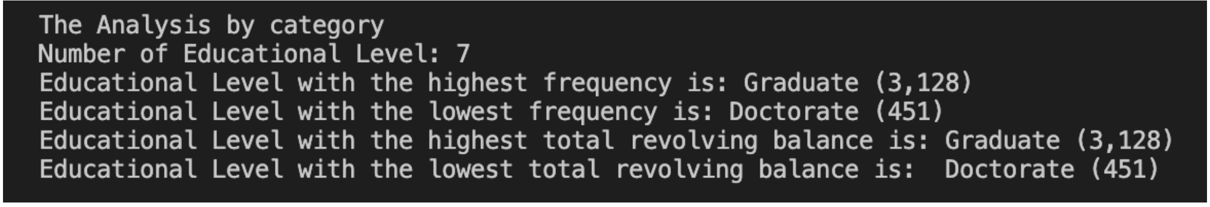
4. Statistical Analysis
    a. Optionally, perform statistical calculations on the chosen columns. • You'll be asked if you want to proceed with these calculations

5. Visualization •
    a. Select the type of visualization you want for your data:
        i. Histogram (h)
        ii. Box plot (b)
        iii. Scatter plot (s)
        iv. The program generates the chosen visualization based on your selection.



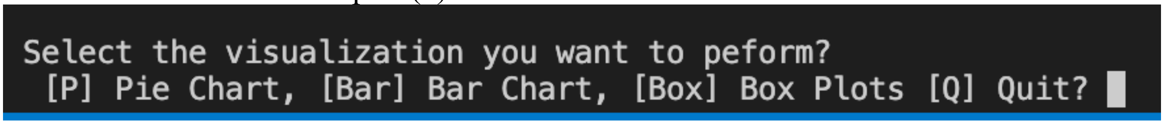
```
Select the visualization you want to peform?
 [H] Histogram, [B] Box plot, [S] Scatter plot , [Q] Quit?
```

6. Sub-Category Analysis
    a. Once the column analysis is completed, the program offers further analysis on sub-categories within the dataset.
    b. You will be prompted to select a sub-category column and a numeric column to visualize with the sub-category.
7. Sub-Category Statistics and Visualization
    a. The program calculates frequency and average values for each subcategory.



```
The Analysis by category
Number of Educational Level: 7
Educational Level with the highest frequency is: Graduate (3,128)
Educational Level with the lowest frequency is: Doctorate (451)
Educational Level with the highest total revolving balance is: Graduate (3,128)
Educational Level with the lowest total revolving balance is:  Doctorate (451)
```

    b. Choose a visualization type for the sub-categories:
        i. Pie chart (p)
        ii. Bar chart (r)
        iii. Box plot (b)



```
Select the visualization you want to peform?
 [P] Pie Chart, [Bar] Bar Chart, [Box] Box Plots [Q] Quit?
```

8. Quitting
    a. At any point, you can choose to exit the program by entering 'q' when prompte

The program is designed to provide a comprehensive suite of data analysis and visualization tools. Users have the flexibility to terminate the program at their convenience by inputting 'q' during any prompt.

Key Features:

- Statistical Analysis: The program computes essential statistical metrics such as the mean, median, mode, maximum, minimum, range, interquartile range, standard deviation, and skewness (both median and mode skewness). It also includes functionality to assess the correlation between chosen data columns.
- Visualizations: A range of graphical representations is available to better interpret data patterns:
    - Histograms: Offer a visual interpretation of variable frequencies.
    - Box Plots: Provide insights into data spread and highlight outliers, with options to include or exclude these outliers.
    - Scatter Plots: Map out the association between two distinct variables.
    - Pie Charts: Display the proportional makeup of sub-categories within the data.
    - Bar Charts: Demonstrate averages across different sub-categories.
    - Box Plots for Sub-Categories: Depict the distribution across various sub-categories in a single, cohesive illustration.

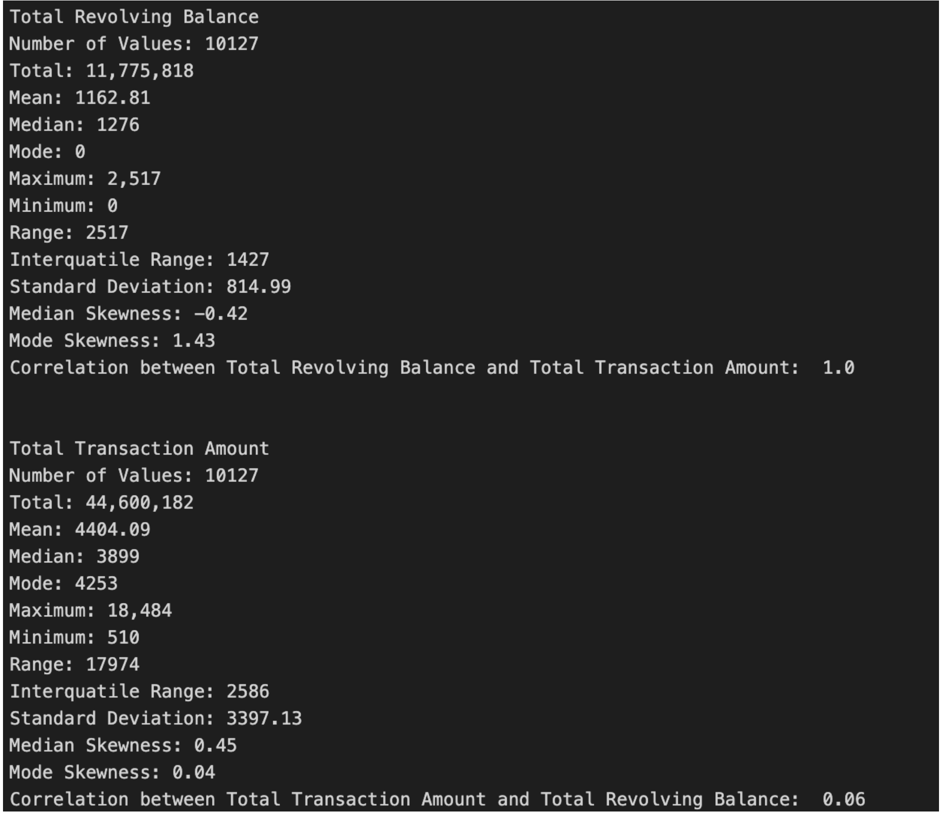
In Summary: The program is an efficient tool for in-depth data analysis and visualization, designed to enhance the comprehension of complex data through statistical detailing and varied graphical formats.

Analysis, Visualisation, Results and Conclusions

a): Analyse the Two Numeric Columns using a List

	Total Revolving Balance	***Total Transaction Amount***
Number of values	10127	10127
Total	11,775,818	44,600,182
Mean	1162.81	4404.09
Median	1276	3899
Mode	0	4253
Maximum	2,517	18,484
Minimum	0	510
Range	2517	17974
Inter-Quartile Range	1427	2586
Standard Deviation	814.99	3397.13
Median Skewness	0.42	0.45
Mode Skewness	1.43	0.04
Correlation	0.06	

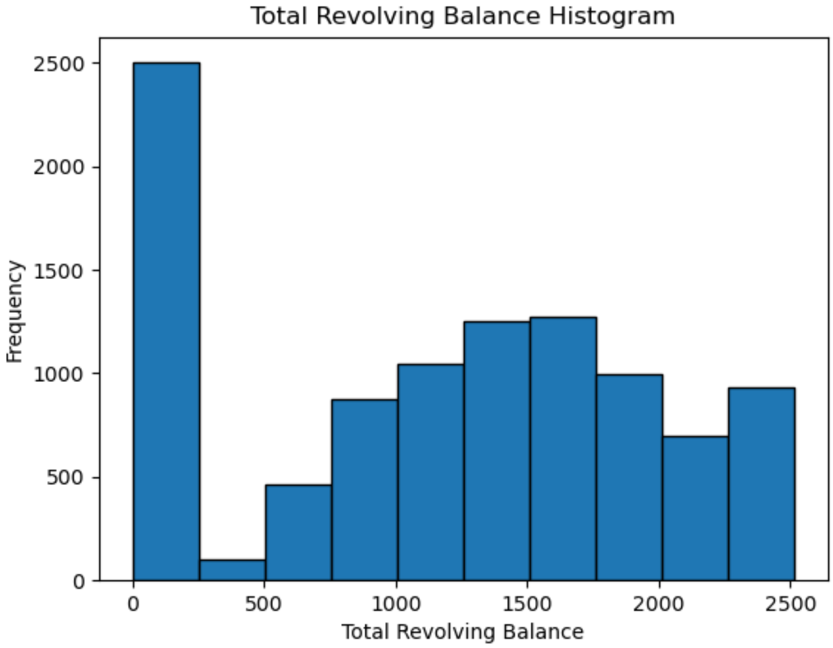
Output Screenshot(s) showing the above results

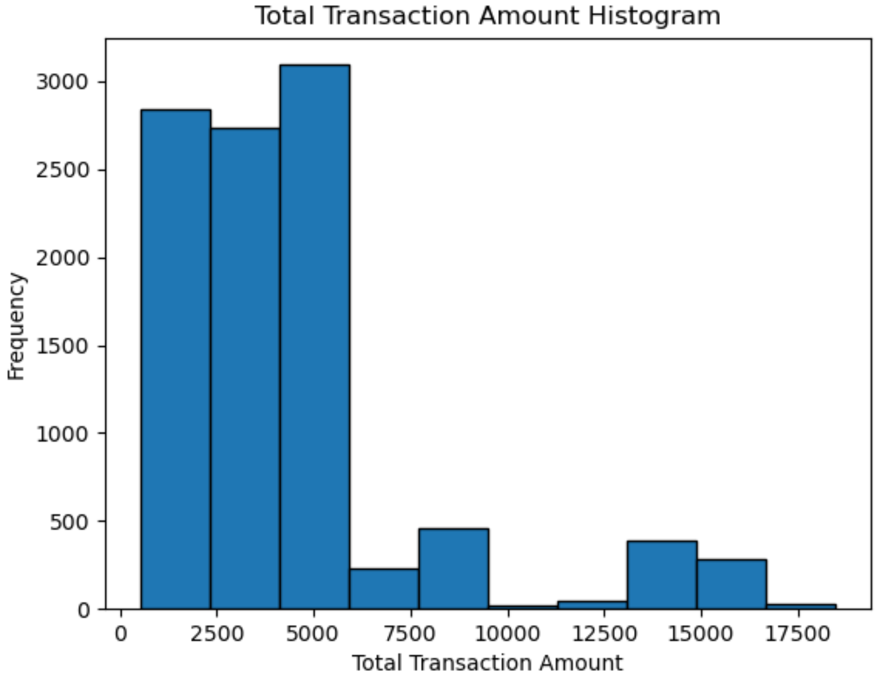


```
Total Revolving Balance
Number of Values: 10127
Total: 11,775,818
Mean: 1162.81
Median: 1276
Mode: 0
Maximum: 2,517
Minimum: 0
Range: 2517
Interquatile Range: 1427
Standard Deviation: 814.99
Median Skewness: -0.42
Mode Skewness: 1.43
Correlation between Total Revolving Balance and Total Transaction Amount:  1.0


Total Transaction Amount
Number of Values: 10127
Total: 44,600,182
Mean: 4404.09
Median: 3899
Mode: 4253
Maximum: 18,484
Minimum: 510
Range: 17974
Interquatile Range: 2586
Standard Deviation: 3397.13
Median Skewness: 0.45
Mode Skewness: 0.04
Correlation between Total Transaction Amount and Total Revolving Balance:  0.06
```

Visualisations

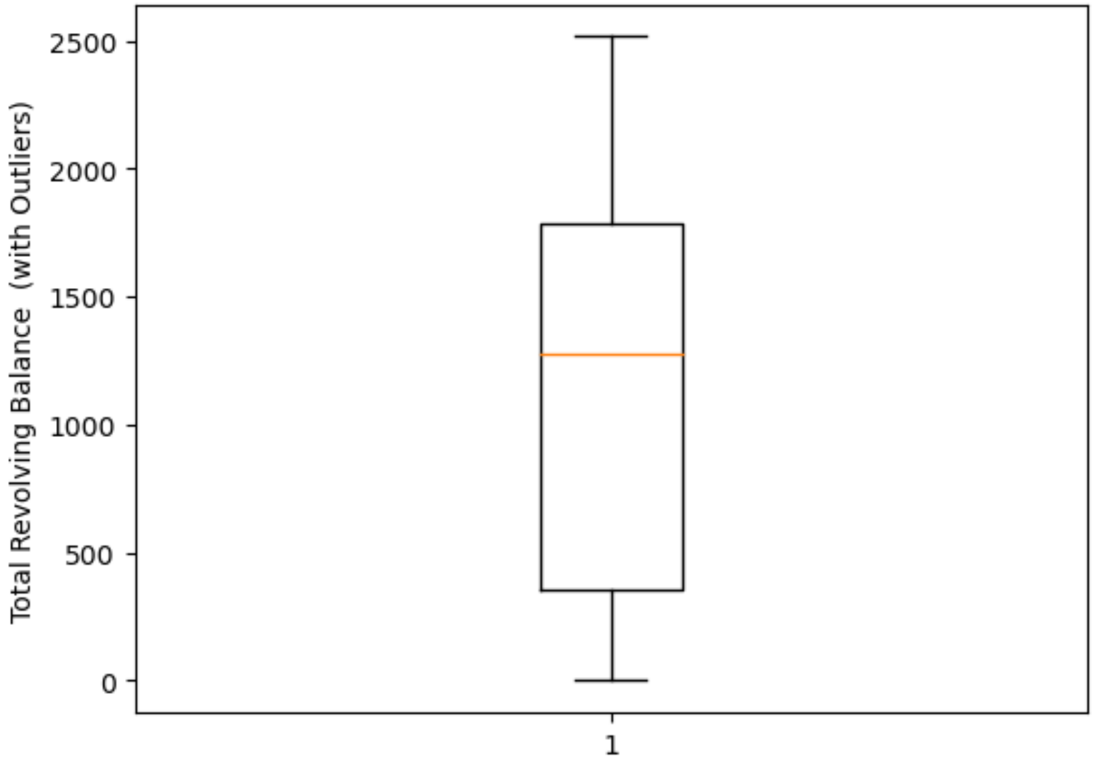
Histograms



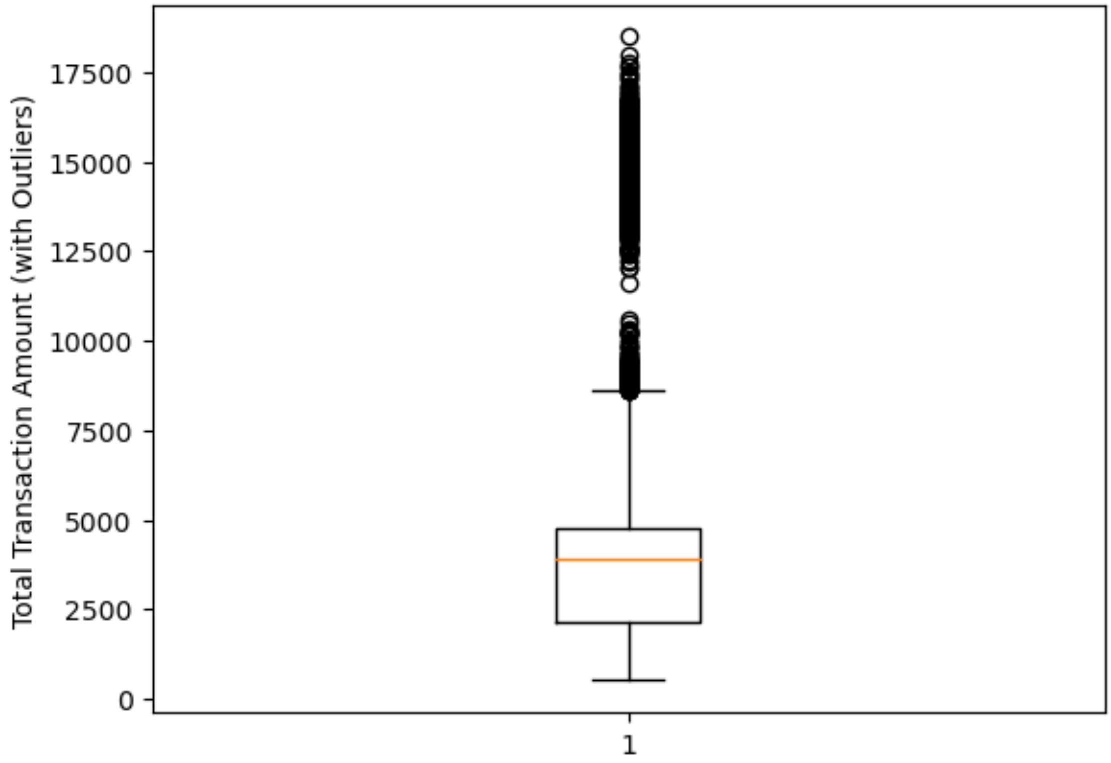


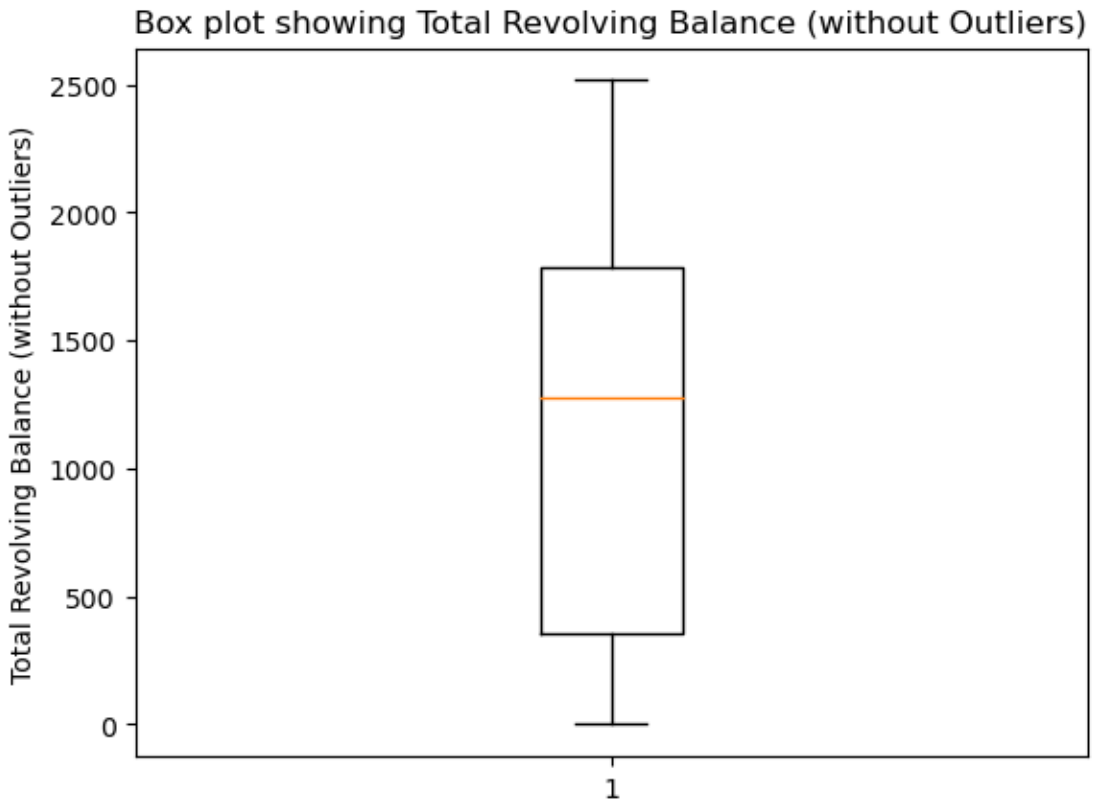
Box Plots

Box plot showing Total Revolving Balance (with Outliers)

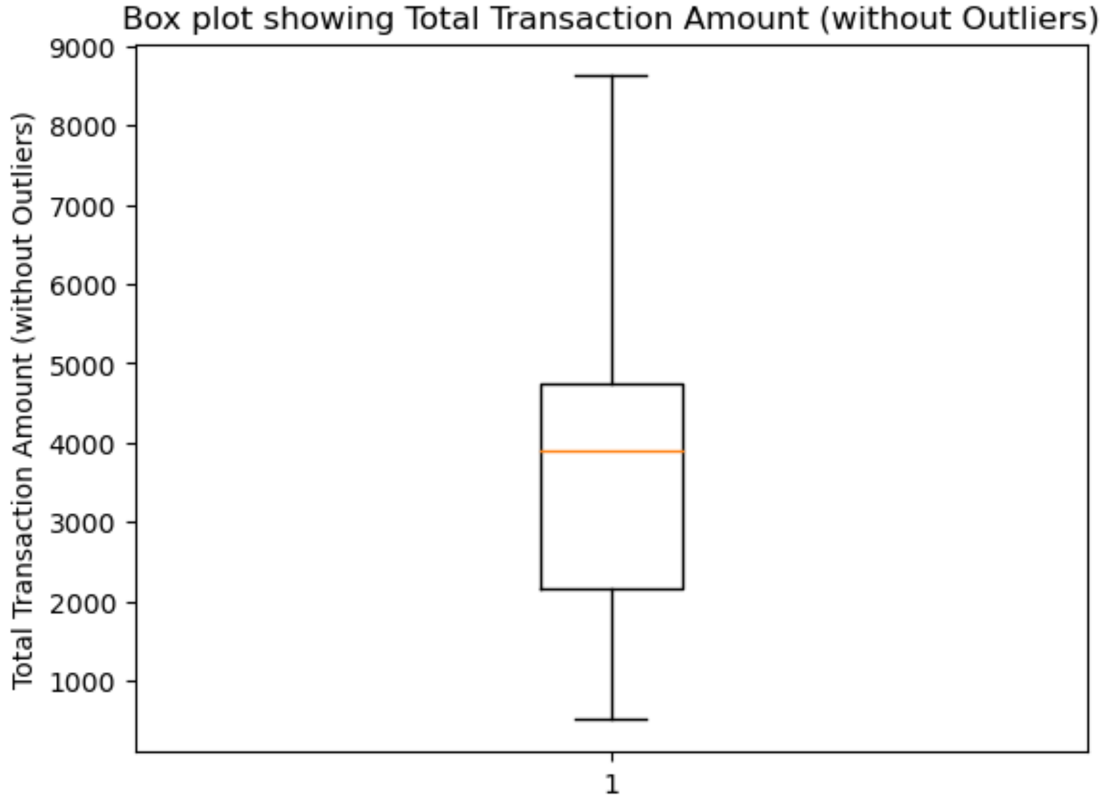


Box plot showing Total Transaction Amount (with Outliers)



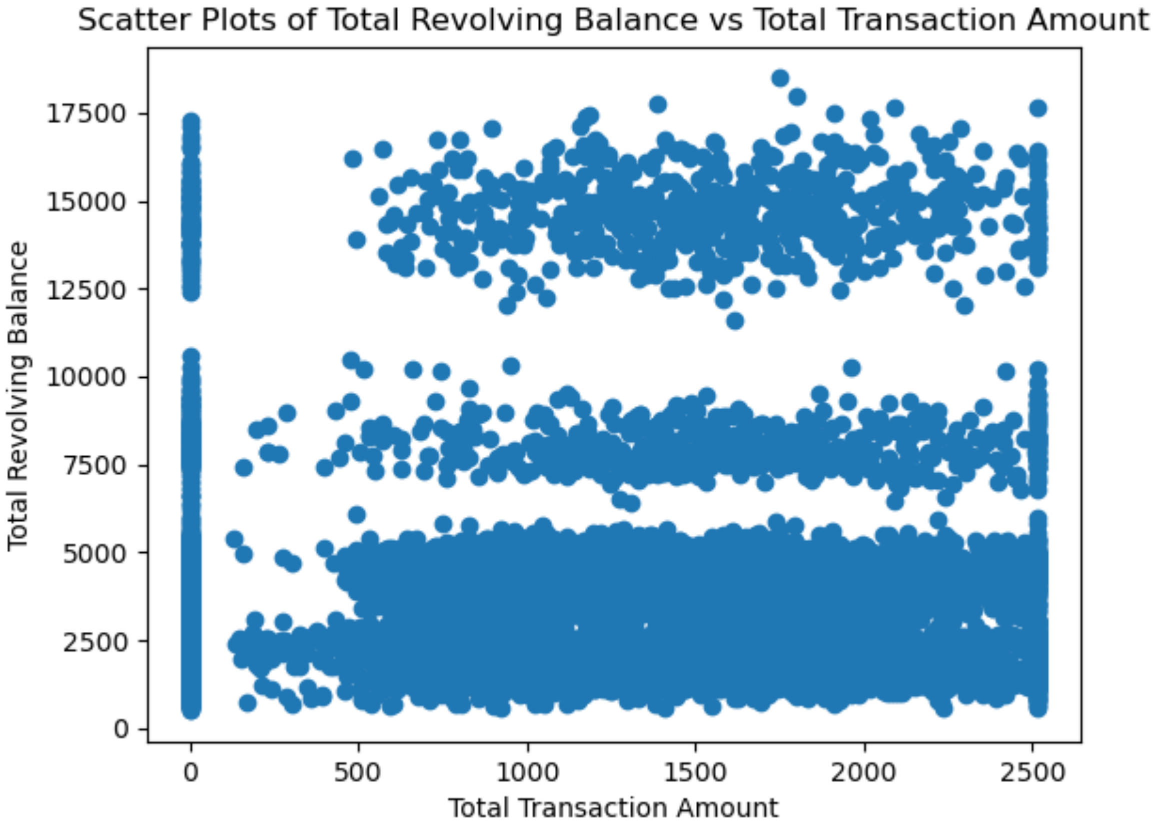


Box plot showing Total Revolving Balance (without Outliers)



Box plot showing Total Transaction Amount (without Outliers)

Scatter Plot

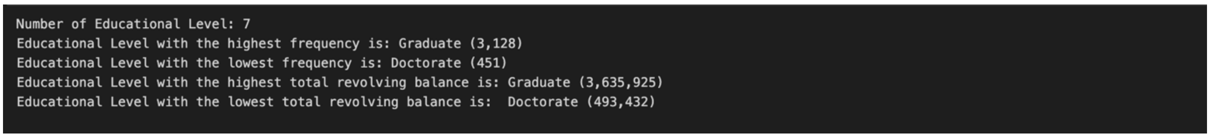


b) Analyse by Category using a Dictionary

Category Name	***Educational Level***
Number of Subcategories	*7*
Subcategory with highest frequency	*Graduate* (*3,128*)
Subcategory with lowest frequency	*Doctorate* (*451*)

Analysis by Category	***Total Educational Level***
Subcategory with highest total	*Graduate* (3,635,925)
Subcategory with lowest total	*Doctorate* (*493,432*)

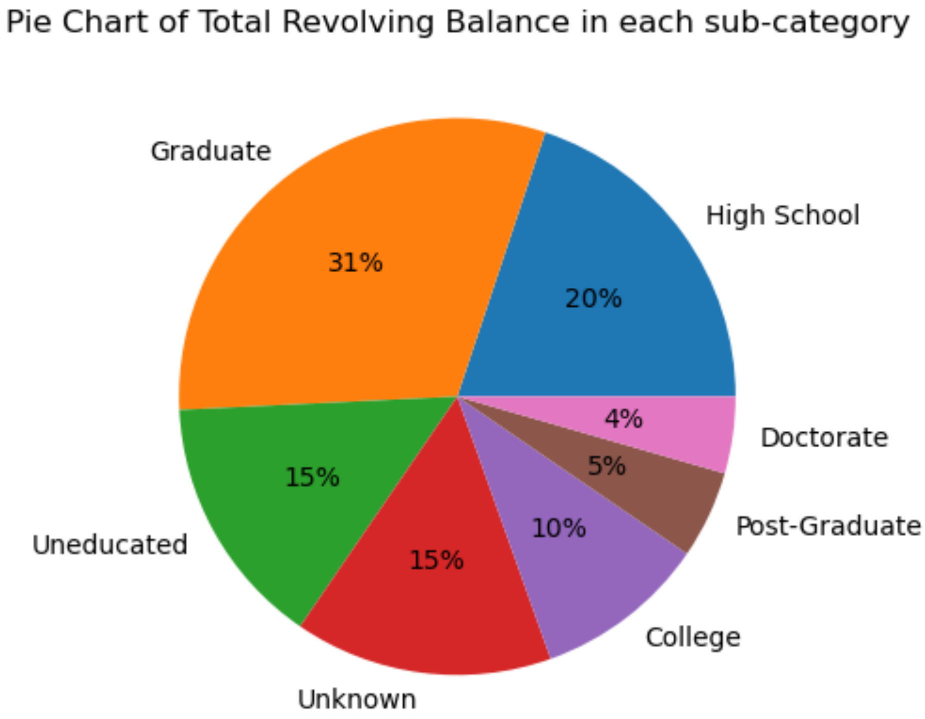
Output Screenshot(s) showing the above result.



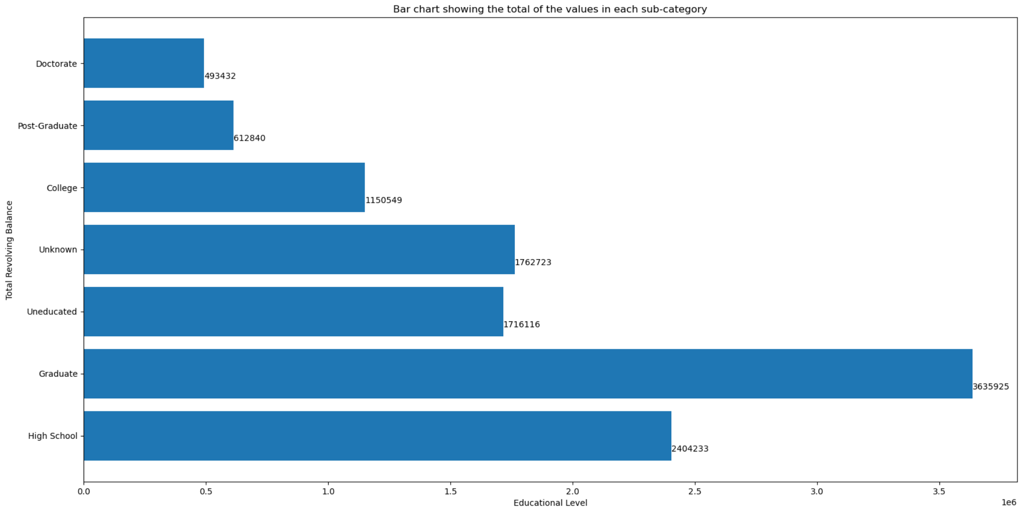
```
Number of Educational Level: 7
Educational Level with the highest frequency is: Graduate (3,128)
Educational Level with the lowest frequency is: Doctorate (451)
Educational Level with the highest total revolving balance is: Graduate (3,635,925)
Educational Level with the lowest total revolving balance is:  Doctorate (493,432)
```

Visualisations

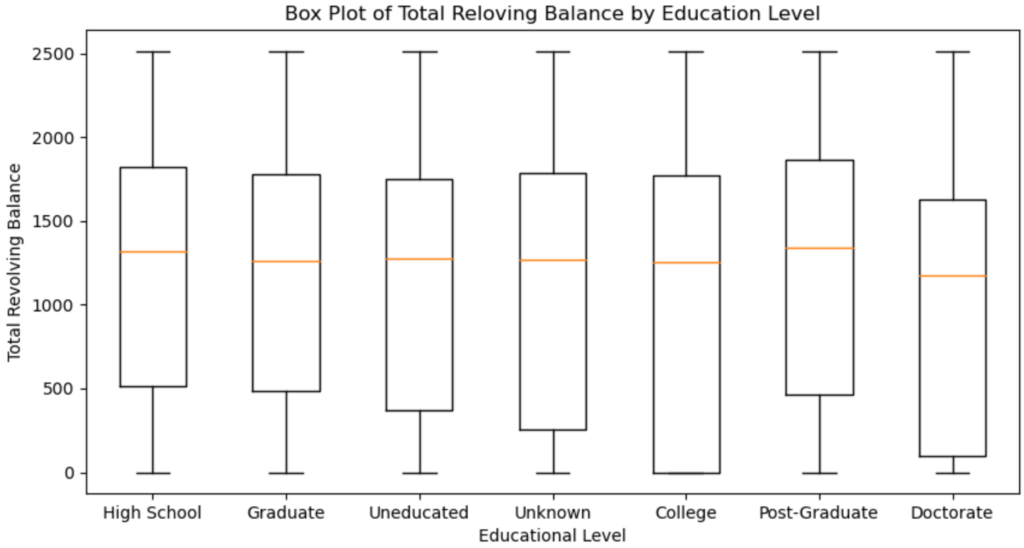
Pie chart showing the percentage number of values in each category.



Bar chart showing the total of the values in each sub-category



Box plots of the values for each sub-category, done in a single visualisation



Conclusions

1. a) The box plots shows that each category has similar revolving balance which is indicated by their median
   b) This is important to credit companies because they can design programs to improve financial literacy since higher educational level does not improve financial literacy

2. a) The scatter plots indicates that there is no direct relationship between Total Revolving Balance and Total Transaction Amount.
   b) This data is important as it allows the credit companies to conclude that spending more or less does not increase or decrease revolving balance.

3. a) The bar chart shows that the highest revolving balance belongs to the graduates
   b) It provides insight into the credit usage patterns of customers based on their education level. For example, it might indicate that customers with a high school education level are more reliant on revolving credit compared to those with higher education.

4. a) High school education has the second largest revolving balance, making up 20%.
   b) The surprisingly high percentage of revolving balance among the 'Uneducated' suggests a potential need for financial education and debt management services targeted at this group.

5. a) The majority of transactions occur in the lower range of amounts between 0 to 5000.
   b) This histogram can provide insights into customer spending behaviour, showing that customers are more likely to make smaller transactions than larger ones.