**Data Analysis Report**

Contents

[Contents 1](#_Toc149298429)

[**Program Design: Structure Chart** 2](#_Toc149298430)

[**Unit Testing** 3](#_Toc149298431)

[**User Manual** 7](#_Toc149298432)

[**Analysis, Visualisation, Results and Conclusions** 9](#_Toc149298433)

[*a): Analyse the Two Numeric Columns using a List* 9](#_Toc149298434)

[*b) Analyse by Category using a Dictionary* 14](#_Toc149298435)

[Conclusions **Error! Bookmark not defined.**](#_Toc149298436)

[**References** **Error! Bookmark not defined.**](#_Toc149298437)

# **Program Design: Structure Chart**

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

A diagram of a company

Description automatically generated

# **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,  0.01959999999999991,  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*

A computer screen with green text

Description automatically generated

# **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.

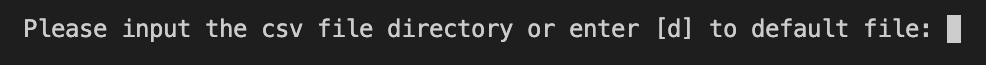
**To Execute**

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



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

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

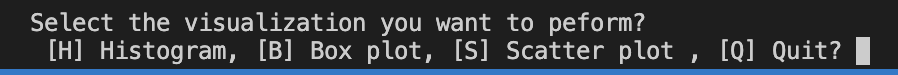


1. Column Selection
   1. The program displays a list of column codes and headers from the CSV.
   2. Choose two columns for statistical analysis by entering their respective code numbers.

A screenshot of a computer

Description automatically generated

1. Statistical Analysis
   1. Optionally, perform statistical calculations on the chosen columns. • You'll be asked if you want to proceed with these calculations
2. Visualization •
   1. Select the type of visualization you want for your data:
      1. Histogram (h)
      2. Box plot (b)
      3. Scatter plot (s)
      4. The program generates the chosen visualization based on your selection.

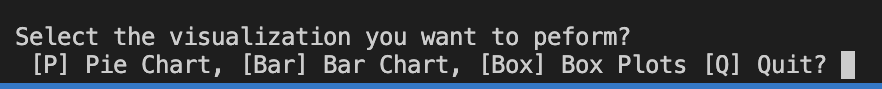


1. Sub-Category Analysis
   1. Once the column analysis is completed, the program offers further analysis on sub-categories within the dataset.
   2. You will be prompted to select a sub-category column and a numeric column to visualize with the sub-category.
2. Sub-Category Statistics and Visualization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

A screen shot of a computer

Description automatically generated

* 1. Choose a visualization type for the sub-categories:
     1. Pie chart (p)
     2. Bar chart (r)
     3. Box plot (b)



1. Quitting
   1. 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*

A screenshot of a computer

Description automatically generated

*Visualisations*

Histograms

A graph of a bar graph

Description automatically generated with medium confidence

A graph of a diagram

Description automatically generated with medium confidence

Box Plots

A graph with a line and a rectangular object

Description automatically generated with medium confidence

A diagram of a box with a line and a square

Description automatically generated with medium confidence

A graph with a line and a rectangular object

Description automatically generated with medium confidence

A graph with a box and a line

Description automatically generated with medium confidence

Scatter Plot

A diagram of blue dots

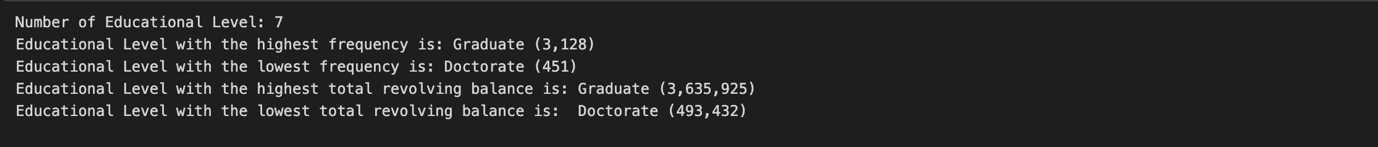
Description automatically generated

## *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.*



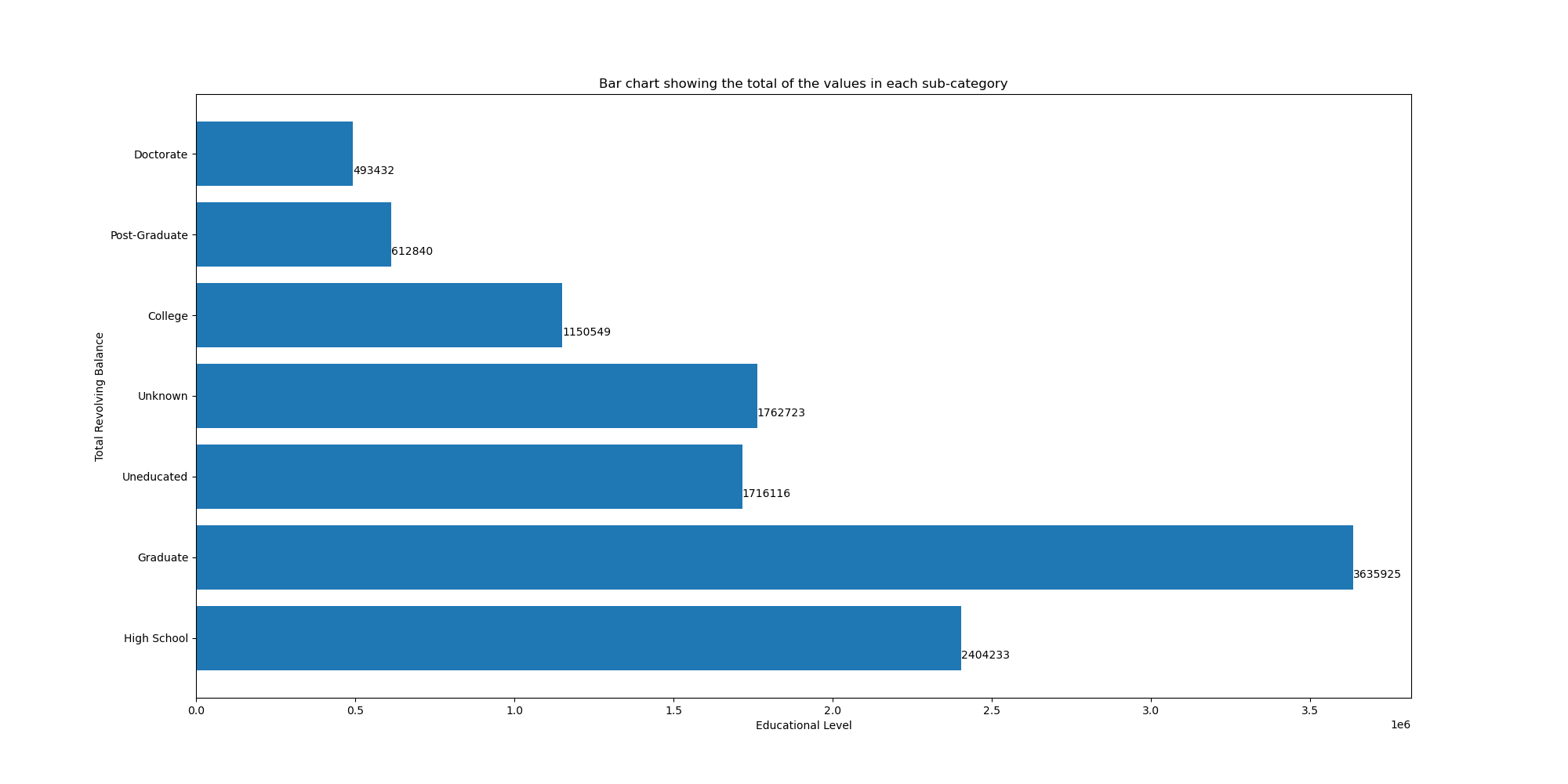
*Visualisations*

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

A pie chart of a number of people

Description automatically generated

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

A diagram of a box diagram

Description automatically generated with medium confidence

## 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

1. 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.

1. 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.

1. 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.

1. 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.