Joseph Annand

DSE5004

Due Diligence Project

Categorizing Columns

Columns 1-5 are demographic-related variables. Columns 6-9 are occupation-related variables. Columns 10-13 and 15 are personal-finance-related variables. Column 14 is blank. Columns 16-22 are household-related variables. Columns 23-27 are car-related variables. Columns 28 and 29 are political variables. Columns 30-33 are credit-card-related variables. Columns 35-37, 39, 40, 43, 44, and 54 are quantitative service history variables. Columns 38, 41, 42, 45-53, and 55-59 are telecommunication-service-related variables.

Recoding Columns

VoiceOverTenure, VoiceOverLastMonth, EquipmentOverTenure, EquipmentLastMonth, DataOverTenure, and DataLastMonth, HHIncome were all initially coded as character vectors with commas and dollar signs. All values in these columns were parsed for numbers and converted to numeric data type. NA values were converted to zero, so that the derived features could be calculated.

Derving New Features

Determining a new feature that is the sum of the total amounts they have spent on phone, equipment, and data services over their tenure with the phone company allows for tracking of the most lucrative customers and find which variables are most related to how much customers spend in telecommunication services. Expanding on the TotalOverTenure feature, the TotalByTenure feature calculates the average amount a customer spends on services per year with the phone company. With this, customers across the range of tenure with the phone company can be compared. TotalLastMonth feature sums the amounts each customer paid for services in the past month. This data gives the most updated look at the total revenue each customer brings in.

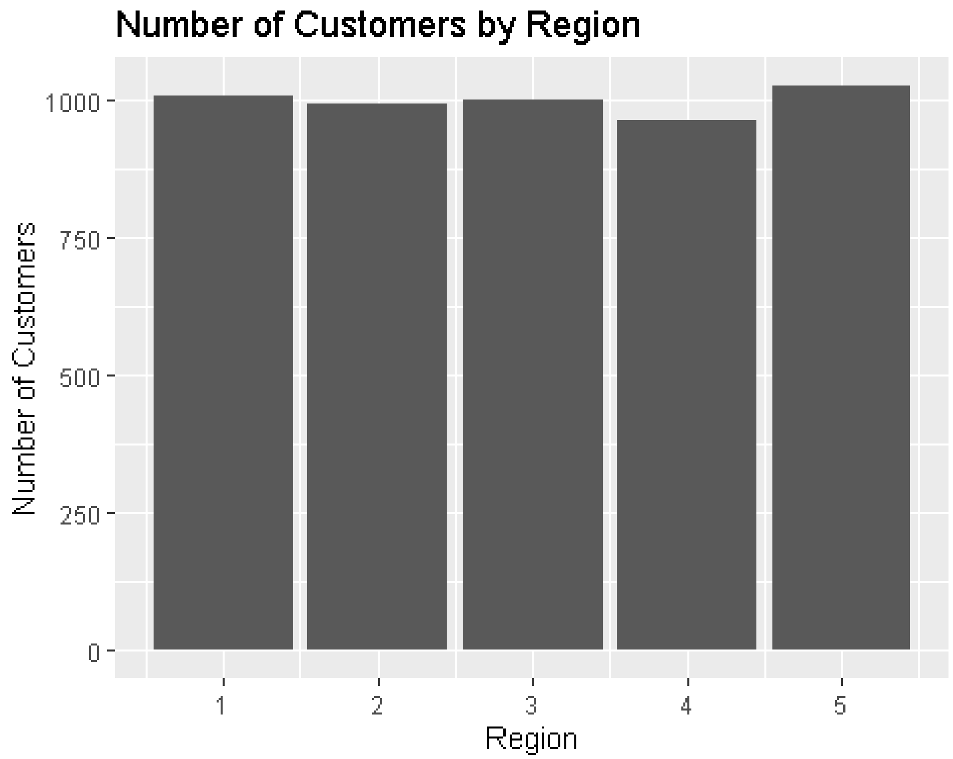
Single Variable Plots

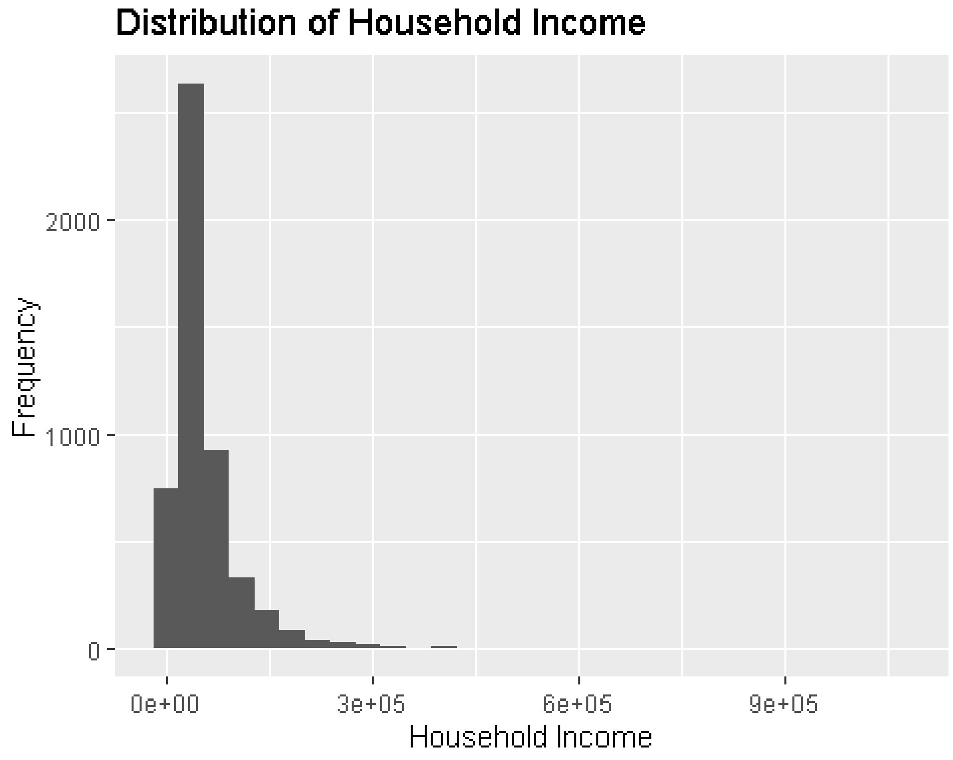
The data from the Region feature was visualized using a bar chart to see the distribution of customers across the five regions. Customers are evenly distributed across the five regions. Household Income and Total Last Month payments were plotted on histograms to visualize the distributions and check for outliers. The histograms showed right-skewed distributions. Consequently, the data was normalized using min-max scaling for both features.

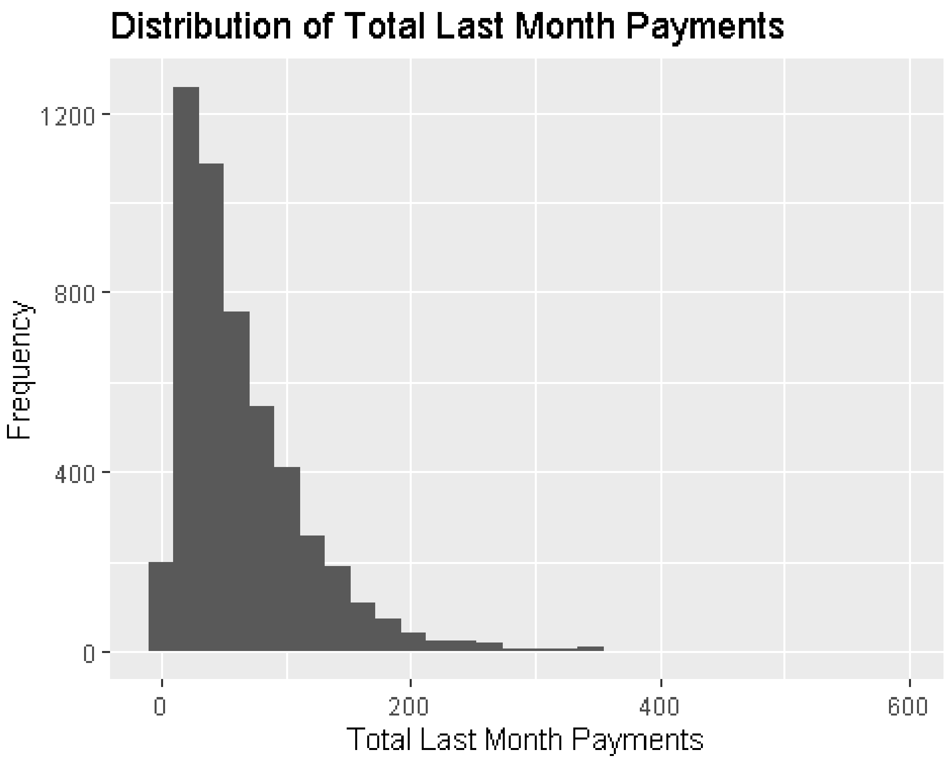
Two Column Analyses

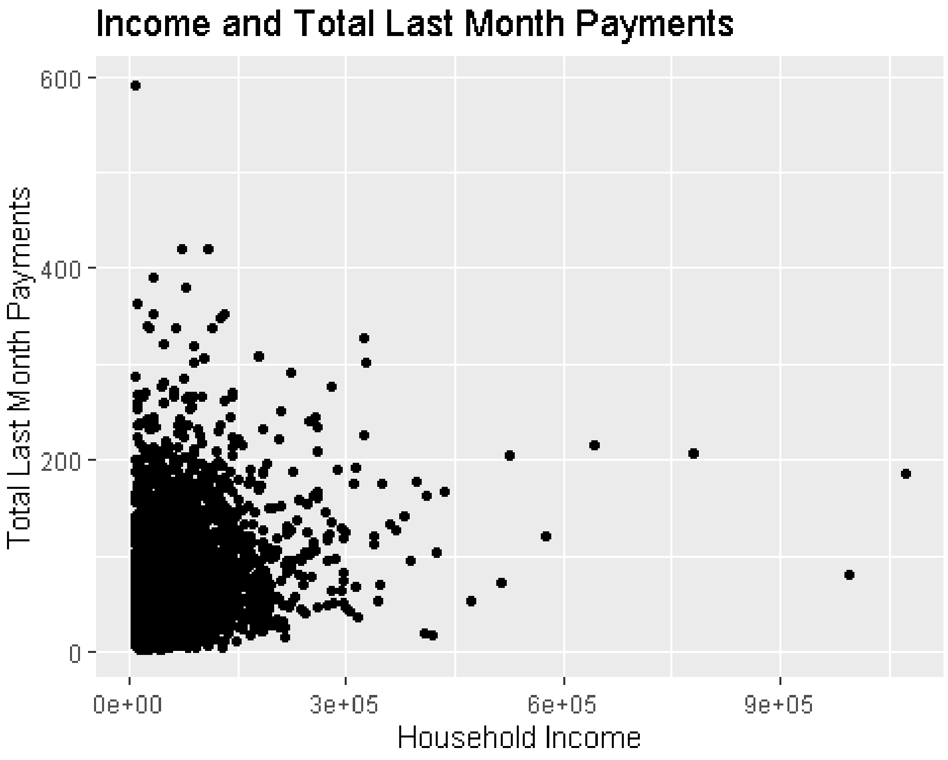
Multiple scatterplots were created to visualize relationships between a variety of continuous, numeric variables. First, total last month payments for telecommunication services was plotted over household income. Both features were normalized using min-max scaling; however, this did not change the scatterplot. In both plots, there may be a weak, positive correlation between the two variables. Additionally, TV Watching Hours was plotted over Age to discover there is no correlation between the two variables. Last Month Voice Payment was plotted over Household Income, and the scales on the axes were reduced to exclude outliers from the visual. No correlation between the two variables was noticeable in the graph. Similarly, there was no correlation between Total Last Month Payment and Debt to Income Ratio.

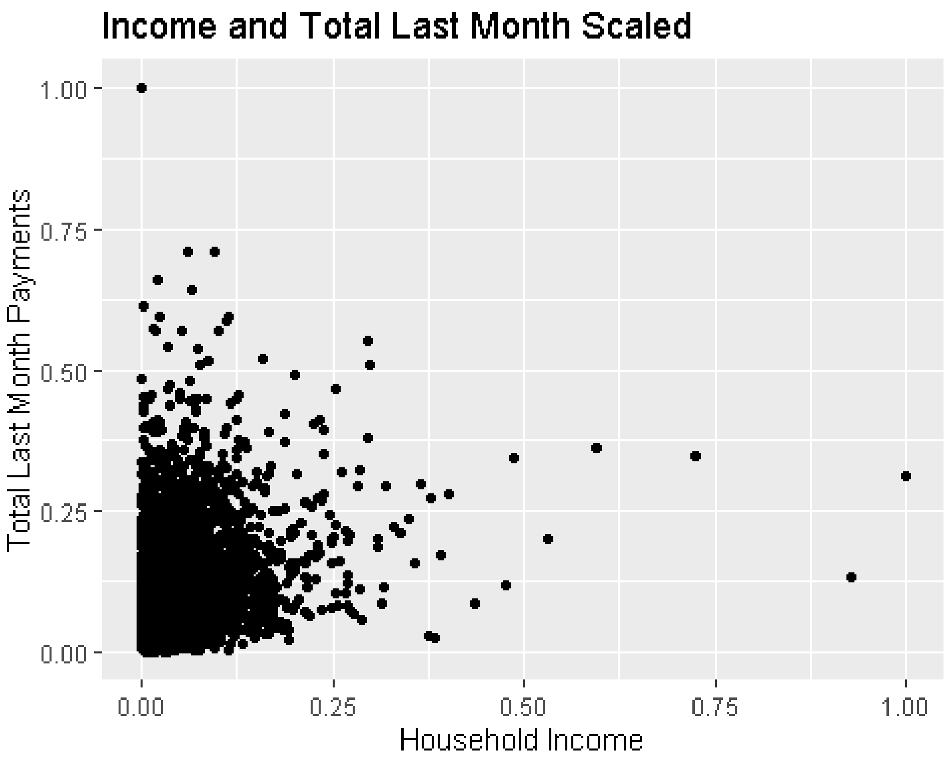
GG Plot Visualizations

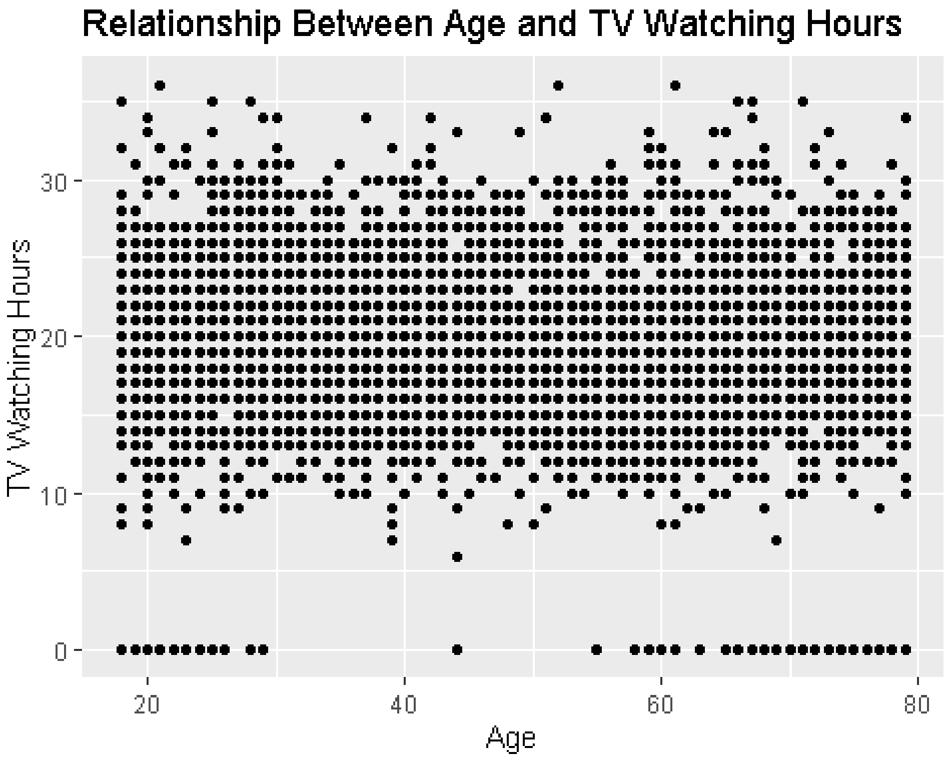




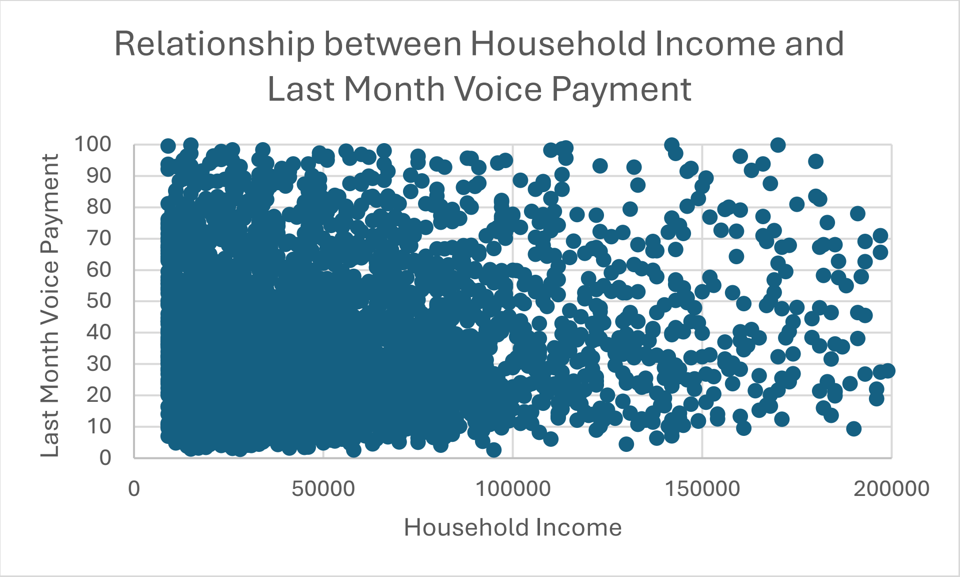

Distribution of Total Last Month Payments
Frequency
1200
800
400
0
0
200
400
600
Total Last Month Payments

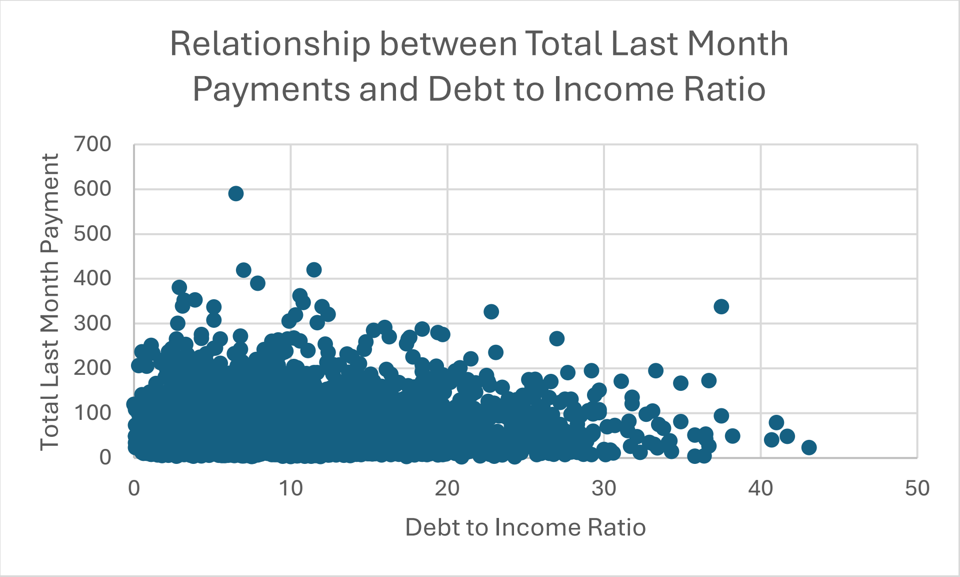

Income and Total Last Month Payments
Total Last Month Payments
600
400
200
0
0e+00
3e+05
6e+05
9e+05
Household Income


Income and Total Last Month Scaled
Total Last Month Payments
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
Household Income


Relationship Between Age and TV Watching Hours
TV Watching Hours
30
20
10
0
20
40
60
80
Age

Excel Visualizations


Relationship between Household Income and Last Month Voice Payment
Last Month Voice Payment
Household Income


Relationship between Total Last Month Payments and Debt to Income Ratio
Total Last Month Payment
Debt to Income Ratio