Understanding Our Customers: Analysis of Consumer Data for Auto Dealerships

Jason Noble

School of Engineering & Technology, National University

ANA600, Fundamentals of Analytics

Matthew Vanderbilt

July 14, 2024

Understanding Our Customers: Analysis of Consumer Data for Auto Dealerships

Introduction

This report summarizes the analysis of a consumer dataset focusing on demographic characteristics and their implications for the auto dealership industry. The analysis includes 16 descriptive statistics on age, household size, income, number of vehicles, employment sector to name a few. This report aims to provide insights into the sample's demographic structure, makes inferences about the intended population, and addresses missing data concerns. Recommendations for future data collection efforts are also provided.

Methods

The dataset was analyzed to compute descriptive statistics for numbered variables such as age and household size, and categorical variables such as employment sector and region. Missing data was also examined to determine their impact on the analysis. The analysis was conducted using R and involved calculating means, standard deviations, and frequencies for these variables. Tables and figures are provided to illustrate the findings.

Sample Characteristics

Age (Q8):

- Mean: 35.9 years
- Standard Deviation: 9.141 years
- Missing Data: 8% of entries

Household Size (Q7):

- Small (2 members) 11%
- Medium (3-5 members) 74%
- Large (6 or more members) 15%

Income Class (Q14):

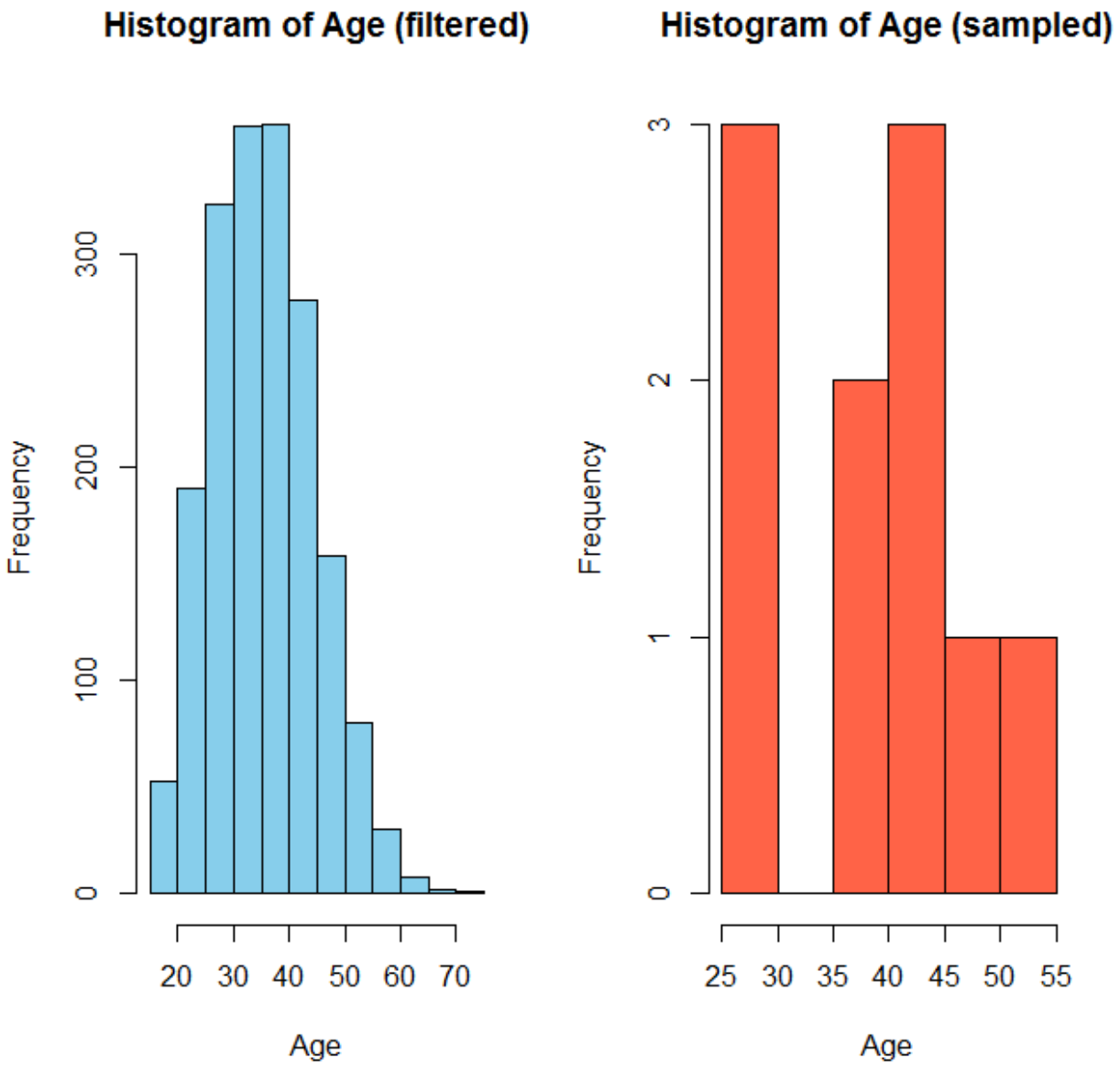
- Lower Class 23%
- Middle Class 54%
- Upper Class 23%

Employment Sector	Frequency (%)		Region	Frequency (%)
Construction	5%		East	25%
Education	4%		Northeast	25%
Financial Services	7%		South	25%
Health Services	10%		West	25%
Leisure & Hospitality	13%			
Manufacturing	7%			
Other	12%			
Prof/Business Services	8%			
Public administration	12%			
Transport & utilities	12%			
Wholesale/retail trade	10%			

Results

```
 --household --      --kids --      --vehicles --       --age --
 Min.   : 2.000   Min.   :1.000   Min.   :1.000   Min.   :18.00
 Median : 4.000   Median :2.000   Median :2.000   Median :35.00
 Mean   : 3.874   Mean   :2.233   Mean   :1.778   Mean   :35.93
 Max.   :10.000   Max.   :8.000   Max.   :4.000   Max.   :75.00
                                                 *NA's   :156

-incomeExpected-   --income --   -priceExpected-  -hoursPerWeek-
 Min.   :-2.000   Min.   : 20.0   Min.  :-4.000   Min.   :26.00
 Median : 1.000   Median : 45.0   Median: 1.000   Median :38.00
 Mean   : 1.222   Mean   : 46.8   Mean  : 1.223   Mean   :38.61
 Max.   : 5.000   Max.   :115.0   Max.  : 7.000   Max.   :69.00

 --investments --      --employmentSector --       --region --
 Bonds         12%     Construction       5%      Midwest   25%
 Mutual Funds  15%     Education          4%      NorthEast 25%
 Retiremt Acct 23%     Finance Services   7%      South     25%
 Savings Acct  33%     Health Services   10%      West      25%
 Stocks        17%     Leisure & Hosp    13%
                       Manufacturing      7%
                       Other             12%
                       Prof/Bus. Services 8%
                       Public admin      12%
                       Trans & utility   12%
                       Wholesale/retail  10%

 -businessExpected-    -financialStability-
 Better Off    28%     Better Off    27%
 Worse Off     72%     Worse Off     73%

 (Q15) --investments household --          --investments household --
                   Bonds  3.670940                    Bonds  1.235495
             Mutual funds  3.748148               Mutual funds  1.308926
      Retirement account  3.903846         Retirement account  1.364042
         Savings account  3.903120            Savings account  1.355595
                  Stocks  3.987302                     Stocks  1.281854
summary(myData_fil$household)
  Min.   1st Qu.  Median  Mean   3rd Qu.  Max.
  2.000   3.000   4.000   3.866   5.000  10.000

      --investments vehicles --          --investments  vehicles --
                   Bonds 1.786325                    Bonds 0.5760463
              Mutual funds 1.714815               Mutual funds 0.6308098
      Retirement account 1.798077         Retirement account 0.6229322
         Savings account 1.753695            Savings account 0.6029726
                 Stocks 1.822222                     Stocks 0.6079213
summary(myData_fil$vehicles)
  Min.   1st Qu.  Median  Mean   3rd Qu.  Max.
  1.000   1.000   2.000   1.774   2.000   4.000

       --investments priceExpected --    --investments priceExpected
                    Bonds     0.9786325                 Bonds      2.507408
               Mutual funds     1.3370370           Mutual funds      2.432990
      Retirement account        1.3870192    Retirement account      2.522520
         Savings account        1.1724138       Savings account      2.495027
                    Stocks     1.0952381                 Stocks      2.539586
summary(myData_fil$priceExpected)
  Min.   1st Qu.  Median  Mean   3rd Qu.  Max.
 -4.000  -1.000   1.000   1.207   3.000   7.000
```

Age Histogram



A chart showing age distributions are mostly between 25-45 years old, with a peak around 35.

Intended Population and Representativeness

Based on the data categories, the intended population appears to be auto dealership salespeople. This hypothesis is supported by the relevance of the questions related to ability to have the finances to buy a vehicle. There is a positive correlation between number in household and number of vehicles.

Representativeness of the Sample

The sample appears representative of a population, with a diverse range of ages, household sizes, income and employment sectors. There are a couple of glaring biases to consider. The hours per week have a range of 26-69 with the mean at 39, so the surveyed population would appear to be mostly full-time employees. Additionally, all surveyed have some level of investments, and if aggregated, appear almost level across all categories, I would hypothesis, that even if their income is low, their investments would assist in making a vehicle purchase.

Factors Affecting Independence and Randomness

- The independence of the sample may be impacted by non-random sampling methods, such as convenience sampling.
- The randomness of the sample is also in question due to potential selection bias and likely voluntary response bias
- The missing data in the age variable is unlikely to introduce significant bias, given the relatively small proportion (8%) and the distribution across other variables.

Recommendations for Future Data Collection

Future data collection should focus on ensuring a more representative sample by using random sampling methods. Efforts should be made to collect complete data, especially for critical variables like age. Additional demographic variables, such as education level, should be included to provide a more comprehensive understanding of the sample.

Conclusion

This report provides an initial summary of the demographic characteristics of a consumer data sample, highlighting the need for careful consideration of sampling methods and data completeness. Future efforts should aim to collect more varied and comprehensive data to support deeper analysis.