Demographic characteristics and customer Behavior insights from a sample Dataset

Ghayasudin Ghayas
National University (NU)
Department of Data Science
Prof. Matthew Vanderbilt
July 14,2024

Statistics is the study of variation. It is the tools and concepts we have invented over time to help us learn from variation. If variation didn't exist, then we wouldn't need statistics. It is the tools and concepts that have been developed, over centuries, to help us understand variation. *Weisberg, Herbert I. (2014). Willful Ignorance: The Mismeasure of Uncertainty. Hoboken, NJ: Wiley.)*

With the exponential growth of data in the digital age, data science has emerged as a crucial field that extracts meaningful insights from vast troves of information, Data science drives innovation and progress across industries, from enhancing business operations to combating disease. Technologies like ML and AI are powered by advanced data analysis, allowing systems to learn and improve autonomously. "R is a free *open-source* coding language commonly used by statisticians. Open-source code is software that is available for anyone to see, modify, and distribute. It is developed as a public, open collaboration, and is made freely available to the public".[1]

Statistics is the study of variation. It is the tools and concepts we have invented over time to help us learn from variation. If variation didn't exist, then we wouldn't need statistics. It is the tools and concepts that have been developed, over centuries, to help us understand variation.

The purpose of this analysis is to provide a comprehensive overview of demographic characteristics and inferential insights from a sample dataset. This statistical report is generated from the simple dataset under name of "Customer_data.csv" through which the researcher wanted to forecast changes in consumer behavior in the United States and better understand consumer confidence about personal finance, employment, price change, *and the perceived state of national business.*

Statistic Summary

The dataset included of 2000 cases (values) and 14 variables or columns including of " household, kids , vehicles , priceExpected, incomeExpected, financialStablity,investments, income, age, employmentsector, region and houseperweek" and the dataset comprises quantitative variables such as income and age, alongside categorical variables including household size and employment sector, the quantitative variables are includes of numerical values and the categorical variables are included non-numerical values.

The mean age of the sample is 35.93 years with a standard deviation of 12.3.

[1] National University,1.3 Doing Statistics with R (instructure.com)

12.3, indicating a moderate dispersion around the mean. Household size, categorized into distinct groups Small, Medum and large categories in consideration of persons number in each household as per below ratio:

breaks <- c(0, 3, 6, Inf)

Small group = (0-3) persons

Medium groups = (3-6) persons

And Large group = (6 above)

The housing grouping is to reflects a diverse distribution within the sample, as per group ranking of the household the small group is included of 860 person, the Medium 1064 persons and large group is 76

```
> summary(wk1Data$household_category)
 Small Medium  Large
   860   1064     76
```

(Ghayas, Ghayasudin. (2024).Ghayas_Week1_Assignment)

Missing data in the whole dataset is identified as whole and then specified per each column which are 156 case and these all are in age category were checked by excluding observations with unspecified age values.

```
> print(total_missing_values)
[1] 156
```

```
> print(missing_values_per_column)
             id         household              kids          vehicles
              0                 0                 0                 0
   priceExpected   incomeExpected businessExpected  financialStability
              0                 0                 0                 0
     investments            income               age   employmentSector
              0                 0               156                 0
          region      hoursPerWeek household_category
              0                 0                 0
```

(Ghayas, Ghayasudin. (2024).Ghayas_Week1_Assignment)

Demographic Frequencies

The sample's regional distribution shows representation as the West and South, the East and the Midwest, comprising the remaining proportions. Employment sector frequencies reveal a significant presence in the Technology, Healthcare, and Retail sectors, with other sectors collectively

Categorical Analysis

Quantitative variables were analyzed across categorical divisions, illustrating distinct profiles. For instance, individuals in higher income brackets tend to have larger household sizes, whereas those

in lower income brackets typically exhibit smaller household sizes. This suggests income influences household composition within the sample.

Impact and Bias Considerations

Factors influencing sample independence include demographic diversity and the randomness of selection methods. The exclusion of missing age data aimed to minimize bias; however, potential biases could arise from non-random exclusion patterns or unobserved characteristics linked to age.

Recommendations for Further Study

Future data collection efforts should prioritize variables like age and household size to enrich demographic insights. Attention to these variables would enhance understanding of their interplay with income and employment sector, providing deeper context for demographic dynamics within diverse populations.

The sample's regional distribution shows representation as the West and South, the East and the Midwest, comprising the remaining proportions. Employment sector frequencies reveal a significant presence in the Technology, Healthcare, and Retail sectors, with other sectors collectively .

the Demographic Frequencies section in the sample should be distribution represent further in percentage of each region (West, South, East and Midwest same for the employment sector frequencies

Conclusion

This analysis underscores the significance of demographic variables in understanding sample characteristics and inferring broader population trends. By focusing on robust sampling methodologies and comprehensive variable inclusion, future research can enhance the accuracy and relevance of demographic analyses in informing strategic decision-making.