SOCS0100 Computational Tools for Reproducible Social Science

2025-01-14

## Set up R for Document Preparation and Data Manipulation

### Part I-A Automated Data Collection (30 points)

The first step that a researcher has to make in the process of data collection is choosing the website of interest and the site address scheme. Specifically, the scraping object was selected as the search results of iPhone in Amazon India. This aspect is basic for web scraping because the URL structure will define how we will be scrolling through pages of products. The base URL is “<https://www.amazon.in/s?k=iphone&qid=1730566584&ref=sr_pg_>”; which is then change dynamically to scrap few more pages of the results. The second part of the data extraction shows the main information that should to be scraped from each of the Web page. In this case the product data in the analysis entails the name, rating, rating from people, price, and number of purchases made on the item. These variables were chosen because they give some important information about the sales of the product, its quality, and its price, which are all useful for analysis.  The last aspect of the collection of data is to automate that scraping process and then join the data from different pages. The function scrape\_page is created to scrape all the pages one by one, extract the meaningful information and present it as data frame. To achieve data from multiple pages the author has used the lapply function that iterates the scrape\_page function for the number of pages 1 to 20

## Part I-B Data Exploration and Contextualisation (10 points)

### Rationale Behind the Data Selected

The reason for using the data from Amazon India for this study stems from the fact that it is one of the most popular platforms for selling consumer products including iPhones. Due to coverage and dominance of the virtual marketplace, especially in India, using Amazon as a platform provides rich scope and understanding of consumers’ purchasing behavior, preferences and product evaluations, albeit with several key socio-economic indicators more relevant and useful in the social sciences, particularly on socio-economic factors influencing technology acceptance and consumer decision making processes.

### Dimension of the Data

[1] 360 5

### View the data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Name | Rating | People | Price | Bought |
| Apple iPhone 15 (128 GB) - Black | 4.5 out of 5 stars | 2,634 | 57,499 | 5K+ bought in past month |
| iPhone 16 128 GB: 5G Mobile Phone with Camera Control, A18 Chip and a Big Boost in Battery Life. Works with AirPods; Teal | 4.4 out of 5 stars | 208 | 74,900 | 400+ bought in past month |
| Apple iPhone 13 (128GB) - Starlight | 4.5 out of 5 stars | 32,149 | 43,999 | 1K+ bought in past month |
| Apple iPhone 15 (128 GB) - Blue | 4.5 out of 5 stars | 2,634 | 57,999 | 5K+ bought in past month |
| Apple iPhone 15 (128 GB) - Black | 4.5 out of 5 stars | 2,634 | 57,499 | 5K+ bought in past month |
| Apple iPhone 15 (128 GB) - Pink | 4.5 out of 5 stars | 2,634 | 57,499 | 5K+ bought in past month |

### Remove commas from the Price and People columns and convert to numeric

### Extract the first component (numeric rating) and convert to numeric

### Check the structure of the data frame to confirm changes

The dataset consists of five variables: Name, Rating, People, Price, and Bought. The Name variable is a character vector containing the product names and descriptions, such as “iPhone 16 128 GB: 5G Mobile Phone with Camera Control, A18 Chip and a Big Boost in Battery Life.” These names provide essential details about each product, including storage capacity, features, and color. The Rating variable is numeric and represents the customer rating for each iPhone model, ranging from 4.4 to 5.0 stars.

## Handling Missing Values

### Check for the Missing Values

Name Rating People Price Bought   
 0 0 0 0 0

### Eliminate the Missing Values for the Model Name and the Number of Pieces Bought

Name Rating People Price Bought   
 0 0 0 0 0

### Descriptive Statistics

Data summary

|  |  |
| --- | --- |
| Name | all\_data |
| Number of rows | 360 |
| Number of columns | 5 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 1 |
| factor | 1 |
| numeric | 3 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: character**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| skim\_variable | n\_missing | complete\_rate | min | max | empty | n\_unique | whitespace |
| Name | 0 | 1 | 30 | 122 | 0 | 17 | 0 |

**Variable type: factor**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| skim\_variable | n\_missing | complete\_rate | ordered | n\_unique | top\_counts |
| Bought | 0 | 1 | FALSE | 8 | 1K+: 80, 5K+: 80, 400: 60, 3K+: 40 |

**Variable type: numeric**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Skim variable | N missing | Complete rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| Rating | 0 | 1 | 4.49 | 0.03 | 4.4 | 4.5 | 4.5 | 4.5 | 4.5 | ▁▁▁▁▇ |
| People | 0 | 1 | 10913.94 | 13254.23 | 208.0 | 2634.0 | 2634.0 | 32149.0 | 32149.0 | ▇▁▁▁▃ |
| Price | 0 | 1 | 58337.61 | 11613.38 | 43499.0 | 45490.0 | 57499.0 | 67999.0 | 84900.0 | ▆▇▃▂▁ |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | vars | n | mean | sd | median | trimmed | mad | min | max | range | skew | kurtosis | se |
| Name\* | 1 | 360 | 8.9 | 4.8 | 8.5 | 8.9 | 5.9 | 1.0 | 17.0 | 16.0 | 0.0 | -1.1 | 0.3 |
| Rating | 2 | 360 | 4.5 | 0.0 | 4.5 | 4.5 | 0.0 | 4.4 | 4.5 | 0.1 | -2.5 | 4.1 | 0.0 |
| People | 3 | 360 | 10913.9 | 13254.2 | 2634.0 | 9597.8 | 1798.4 | 208.0 | 32149.0 | 31941.0 | 1.0 | -1.0 | 698.6 |
| Price | 4 | 360 | 58337.6 | 11613.4 | 57499.0 | 57427.7 | 15567.3 | 43499.0 | 84900.0 | 41401.0 | 0.4 | -0.5 | 612.1 |
| Bought\* | 5 | 360 | 4.8 | 2.3 | 5.0 | 4.8 | 3.7 | 1.0 | 8.0 | 7.0 | 0.0 | -1.3 | 0.1 |

Rating (M = 4.5, SD = 0.0), which represents the average customer rating, showed very little variation, as evidenced by the zero-standard deviation. The median rating was 4.5, and the range was from 4.4 to 4.5, with a skewness of 0.0 and a kurtosis of -1.2, indicating a nearly normal distribution.

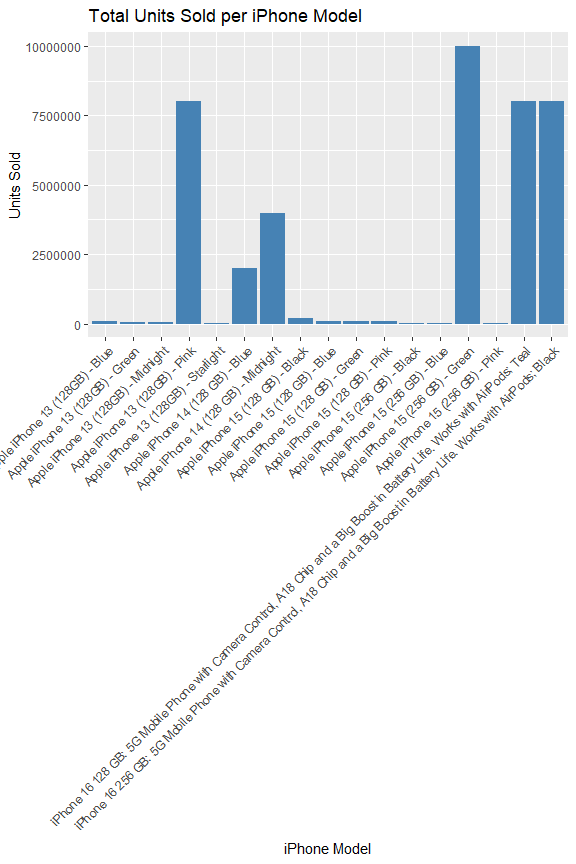
People (M = 12,453.9, SD = 14,042.7), representing the number of people who reviewed each product, exhibited high variability, with values ranging from 202 to 32,138. The median number of reviews was 2,625. The distribution was positively skewed (skew = 0.7) and showed platykurtic behavior (kurtosis = -1.5), suggesting that the distribution was not normal and that fewer products had exceptionally high review counts.

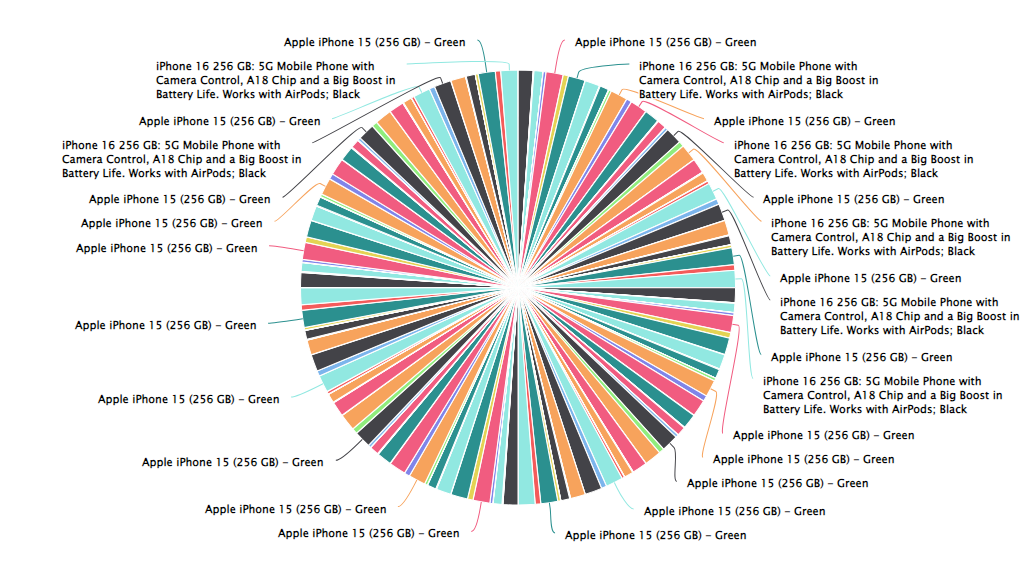
Price (M = 58,054.5, SD = 13,340.7) reflected the price of the iPhones, with values ranging from 43,499 to 144,900. The median price was 56,999. The price data were positively skewed (skew = 1.1) and leptokurtic (kurtosis = 3.7), suggesting a concentration of lower-priced products with a few products at higher prices.

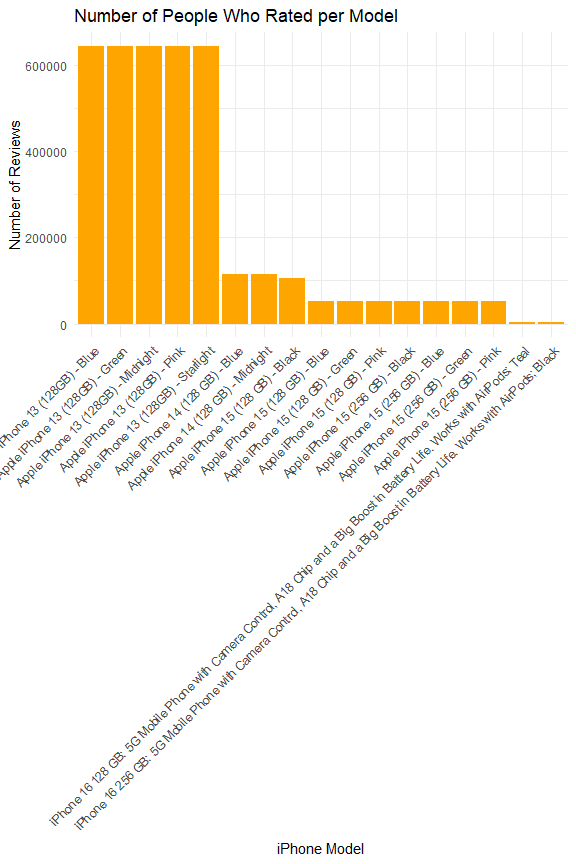
## Assessment Part II

### Part II-A Building an Interactive Dashboard with R Shiny (30 points)

### Bar Graph of Units per iPhone Model







## Justification of the Visuals Selected

### Bar Graph of Units per iPhone Model

Among the visualization used in this report is the “Total Units Sold per iPhone Model” for it is a reliable means of identifying the performance of a product in a specific period as well as the popularity of certain models. It can be said that this bar chart conveys a simple idea of the total units sold of all iPhones, and will help to compare the popularity and the sales of the specific iPhone models. This presentation of information makes it easy to work out which models delivered the most sales, which models did not, and perhaps which areas could benefit from increased attention due to shifting customer preferences.

### Revenue Distribution by Model

The “Revenue Distribution by Model” method is one of the tools for the analysis of the model’s contribution to revenue. This chart proves helpful in showing figures derived per model, making it easier to determine which product contributes to the company’s overall revenue. Through revenue/ sale share of each model, other stakeholders would easily see which particular product is more profitable to the organization/ business. Of particular note about this visual is the ability to evaluate the degree of coherence at the price level with demand. For instance, a model revealing moderate volume but high dollar sales shows that superiority through superior pricinguracy has been realized.

### Comparative Rating for Items Bought

Comparatively, ‘Comparative Rating for Items Bought’ is one of the most important enablers of defining customer satisfaction based on the extent of purchase made. This chart allows analyzing the correlation of the item’s rating (for example, 4.5 against 4.4) and the amount of its purchases (500+, 1k+, 5k+) to reveal subtleties of the connection between product popularity and its quality in the eyes of clients. As with all of the graphs presented here, each of these two is an important tool for analyzing trends in customer preference and satisfaction.

### Number of People Who Rated per Model

In particular, the “Number of People Who Rated per Model” chart is a very effective method to measure customer interest and product models’ appeal. With the help of the presented number of ratings for each model, this chart gives an idea of the quantity of requests and interest on the part of customers in definite goods. This kind of visualization is especially helpful for deciding which models are the most popular since the greater the number of ratings, the more often they are used as well as the higher demand for them on the market.

## Part II-C Critical Engagement with AI: ChatGPT (20 points)

### Reflection on ChatGPT’s Contributions to the Computational Process

ChatGPT was incredibly useful as a collaboration aid during the creation of the interactive dashboard. The initial competence of Signal to polish code and produce fresh strategies enriched the algorithm computationally to a large extent. For instance, while solving a problem where it was required to plot an interactive comparative bar graph for ratings based on the purchase volume, ChatGPT came up with accurate, technique-oriented and explorable solutions using ggplot2 and plotly. This was particularly helpful in avoiding the time and effort that would otherwise be spent in developing the best methods towards debugging and visualization.

### Evaluation of Challenges and Learning Experience

Thus, although using ChatGPT provided significant advantages, some issues arose which suggest critical estimation of the results. For example, although while coming up all the solutions it had correct syntax and code snippets some of them did not understand the data structure for what it was meant for and they had to be modified. This brought out the fact that the tool relies on clear specific instructions most of the time if the instructions given are ambiguous the resultant solutions are generalized or irrelevant.