Amazon Sales Data Analysis 2019

Jomaica Lei

2024-08-03

Amazon

In this assignment, you will be given a dataset of Amazon sales of technology products placed over several months in 2019 in a select number of urban ZIP codes.

Data

sales_data.Rdata

`sales_data.Rdata` can be loaded in using the `import()` function from the **rio** package. Put the file in your working directory (generally, the same folder you've saved your .RMD file in; you can then go Session →→ Set Working Directory to be sure), and load it in:

Let's take a quick look:

```
                     Product Quantity PriceEach            DateTime       Date
1       USB-C Charging Cable        2     11.95 2019-04-19 08:46:00 2019-04-19
2 Bose SoundSport Headphones        1     99.99 2019-04-07 22:30:00 2019-04-07
3               Google Phone        1       600 2019-04-12 14:38:00 2019-04-12
4           Wired Headphones        1     11.99 2019-04-12 14:38:00 2019-04-12
5           Wired Headphones        1     11.99 2019-04-30 09:27:00 2019-04-30
6       USB-C Charging Cable        1     11.95 2019-04-29 13:03:00 2019-04-29
    ZIP State          City
1 75001    TX        Dallas
2 02215    MA        Boston
3 90001    CA   Los Angeles
4 90001    CA   Los Angeles
5 90001    CA   Los Angeles
6 94016    CA San Francisco
```

Table 1: sales

Name	Class	Values
Product	character	’20in Monitor’ ’27in 4K Gaming Monitor’ ’27in FHD Monitor’ ’34in Ultrawide Monit
Quantity	character	’1’ ’2’ ’3’ ’4’ ’5’ ’6’ ’7’ ’8’ ’9’
PriceEach	character	’109.99’ ’11.95’ ’11.99’ ’14.95’ ’149.99’ ’150’ ’150.0’ ’1700’ ’1700.0’ and 14 more
DateTime	POSIXct	Time: 2019-01-01 03:07:00 to 2020-01-01 05:13:00
Date	Date	Time: 2019-01-01 to 2020-01-01
ZIP	character	’02215’ ’04101’ ’10001’ ’30301’ ’73301’ ’75001’ ’90001’ ’94016’ ’97035’ and 1 more
State	character	’CA’ ’GA’ ’MA’ ’ME’ ’NY’ ’OR’ ’TX’ ’WA’
City	character	’Atlanta’ ’Austin’ ’Boston’ ’Dallas’ ’Los Angeles’ ’New York City’ ’Portland’ ’San Fra

This data set contains eight variables:

- `Product`, the product that has been ordered
- `Quantity`, how many of the product was ordered (note this is a string, you’ll want to fix that!)
- `PriceEach`, the price of each item (note this is a string, you’ll want to fix that!)
- `DateTime` and `Date`, when the order was placed. `DateTime` includes both day and time-of-day when the order was placed, while `Date` is just the date
- `ZIP`, the ZIP code where the order was sent to
- `State` and `City`, the city and state where the order was sent to

Let’s look at some descriptive statistics:

- We can first see that `Quantity` and `PriceEach` are both stored as character variables, we’ll have to do something about that (hint: `as.numeric()`).
- We can also see that the data covers from January 2019 through December 2019
- We also see that a limited set of only 9 ZIP codes are covered, and all of them are in urban areas (Atlanta, Austin, etc.).
- There is only one ZIP per city, and cities generally have many ZIP codes, so we are only covering a single part of each city. For example, that Seattle ZIP code is for a part of downtown Seattle. You can put the ZIP code into Google Maps if you want to see specifically where each ZIP is
- There’s a Portland in Oregon and a Portland in Maine

- If you want to work with the time element of that `DateTime` variable, note that you can pull out information like the hour and minute with **lubridate** functions like `hour()` and `minute()`. Also, as always, keep other **lubridate** functions in mind that might be handy, like `floor_date()` to "round" dates to the first day in that week/month/etc.

Finally, we see in the `Product` variable that we are tracking 19 different products, not the entire lineup of everything Amazon sells. If you look at that list (perhaps with `table()` or `unique()`) you'll see that this covers a few different kinds of products, including monitors, laptops, smartphones, and batteries. If you want to analyze some things as a group you may need to use some of the **stringr** functions we used (or perhaps `case_when()`) to pull information from the product names.

zip_info.csv

`zip_info.csv` can be read in using the `import()` function. Put the file in your working directory (generally, the same folder you've saved your .QMD file in; you can then go Session →→ Set Working Directory to be sure), and load it in:

Let's take a quick look. This file contains information on the population of the ZIP codes included in our data. Numbers come from the 2018 American Community Survey (ACS) estimates, i.e. they use five years of ACS data from 2014-2018 to estimate the 2018 numbers:

```
      ZIP        TotalPopulation MedianHHIncome      PCIncome     
 Min.   : 2215   Min.   :12792   Min.   : 46309   Min.   : 14814  
 1st Qu.:15076   1st Qu.:18670   1st Qu.: 59614   1st Qu.: 39644  
 Median :74151   Median :24656   Median : 84555   Median : 53114  
 Mean   :57407   Mean   :26052   Mean   : 81151   Mean   : 57085  
 3rd Qu.:93012   3rd Qu.:27922   3rd Qu.:100026   3rd Qu.: 79410  
 Max.   :98101   Max.   :58975   Max.   :119370   Max.   :100364  
   MedianAge       Race_White      Race_Black   Race_American_Indian
 Min.   :21.60   Min.   : 9231   Min.   : 459   Min.   :148.0       
 1st Qu.:30.38   1st Qu.:13021   1st Qu.:1502   1st Qu.:268.8       
 Median :35.25   Median :15981   Median :2038   Median :315.0       
 Mean   :34.02   Mean   :16141   Mean   :2180   Mean   :426.3       
 3rd Qu.:37.30   3rd Qu.:19960   3rd Qu.:2571   3rd Qu.:606.2       
 Max.   :44.30   Max.   :22921   Max.   :5483   Max.   :802.0       
   Race_Asian      Race_Pacific_Islander   Race_Other      Ethnicity_Hispanic
 Min.   :  173.0   Min.   :  0.0         Min.   :  181.0   Min.   :  609     
 1st Qu.:  744.8   1st Qu.: 12.0         1st Qu.:  295.2   1st Qu.: 1216     
 Median : 2198.5   Median : 42.5         Median :  854.0   Median : 3077     
 Mean   : 3238.9   Mean   : 71.3         Mean   : 5003.8   Mean   : 9223     
 3rd Qu.: 5303.5   3rd Qu.: 90.0         3rd Qu.: 1980.8   3rd Qu.: 3548     
 Max.   :10134.0   Max.   :237.0         Max.   :30491.0   Max.   :53085  
    Citizens    
 Min.   :10432  
 1st Qu.:13910  
 Median :18034  
 Mean   :17172  
 3rd Qu.:19745  
 Max.   :24069  
```

What we see here are:

- `ZIP`, which is a ZIP code we can use to join this data set with the `sales` data
- `TotalPopulation`, which is the population in that ZIP code
- `MedianHHIncome`, which is the median annual household income in that ZIP. Household income calculates the total income from everyone in a given household, and then finds the median household (Income statistics use 2020 ACS instead of 2018)
- `PCIncome`, which is the annual per-capita (i.e. per-person) income in that ZIP. Per-capita income sums up all the income earned by everyone in the ZIP, and then divides it by the number of people in that ZIP (which may include a lot of non-earners, or children) (Income statistics use 2020 ACS instead of 2018)
- `MedianAge`, the median age of people in the ZIP code
- `Race_*` variables, the number of people of each broad-category race in that ZIP code. Note that races are not mutually exclusive. Someone who is, for example, both White and Asian will be counted once as White and once as Asian
- `Ethnicity_Hispanic`, which is the number of people who are Hispanic in the ZIP code. Ethnicity can overlap with any race, so someone who is, for example, both Hispanic and Black will be counted once as Hispanic and once as Black
- `Citizens`, which is the number of US citizens living in the ZIP code

Some notes about this data:

- You can get the proportion of the ZIP code that is White/Black/Hispanic/Citizen/etc. by dividing that value by the `TotalPopulation`
- After you do your join with the sales data, check to make sure the join works correctly! Some of those ZIP codes have leading 0s which can sometimes be a problem (tip: convert everything to numeric, or use `str_pad()` in **stringr** to make the ZIPs five-digit-long strings, with leading 0s)
- Using this file is not required; you could do everything with the `sales` data alone and ignore this if you want

Exploratory Data Analysis

- Look at summary statistics of key variables.
- Distribution of sales across different ZIP codes.
- Popular products and their sales patterns.
- Temporal analysis of sales (e.g., sales trends over time).
- Relationship between sales and demographic factors.

```
     ZIP            TotalPopulation MedianHHIncome      PCIncome     
 Length:10          Min.   :12792   Min.   : 46309   Min.   : 14814  
 Class :character   1st Qu.:18670   1st Qu.: 59614   1st Qu.: 39644  
 Mode  :character   Median :24656   Median : 84555   Median : 53114  
                    Mean   :26052   Mean   : 81151   Mean   : 57085  
                    3rd Qu.:27922   3rd Qu.:100026   3rd Qu.: 79410  
                    Max.   :58975   Max.   :119370   Max.   :100364  
   MedianAge       Race_White      Race_Black   Race_American_Indian
 Min.   :21.60   Min.   : 9231   Min.   : 459   Min.   :148.0       
 1st Qu.:30.38   1st Qu.:13021   1st Qu.:1502   1st Qu.:268.8       
 Median :35.25   Median :15981   Median :2038   Median :315.0       
 Mean   :34.02   Mean   :16141   Mean   :2180   Mean   :426.3       
 3rd Qu.:37.30   3rd Qu.:19960   3rd Qu.:2571   3rd Qu.:606.2       
 Max.   :44.30   Max.   :22921   Max.   :5483   Max.   :802.0       
   Race_Asian      Race_Pacific_Islander   Race_Other      Ethnicity_Hispanic
 Min.   :  173.0   Min.   :  0.0         Min.   :  181.0   Min.   :  609     
 1st Qu.:  744.8   1st Qu.: 12.0         1st Qu.:  295.2   1st Qu.: 1216     
 Median : 2198.5   Median : 42.5         Median :  854.0   Median : 3077     
 Mean   : 3238.9   Mean   : 71.3         Mean   : 5003.8   Mean   : 9223     
 3rd Qu.: 5303.5   3rd Qu.: 90.0         3rd Qu.: 1980.8   3rd Qu.: 3548     
 Max.   :10134.0   Max.   :237.0         Max.   :30491.0   Max.   :53085     
    Citizens    
 Min.   :10432  
 1st Qu.:13910  
 Median :18034  
 Mean   :17172  
 3rd Qu.:19745  
 Max.   :24069  
```

```
   Product             Quantity       PriceEach      
 Length:185950      Min.   :1.000   Min.   :   2.99  
 Class :character   1st Qu.:1.000   1st Qu.:  11.95  
 Mode  :character   Median :1.000   Median :  14.95  
                    Mean   :1.124   Mean   : 184.40  
                    3rd Qu.:1.000   3rd Qu.: 150.00  
                    Max.   :9.000   Max.   :1700.00  
                                                     
    DateTime                             Date                 ZIP           
 Min.   :2019-01-01 03:07:00.00   Min.   :2019-01-01   Length:185950     
 1st Qu.:2019-04-16 21:08:45.00   1st Qu.:2019-04-16   Class :character  
 Median :2019-07-17 20:36:30.00   Median :2019-07-17   Mode  :character  
 Mean   :2019-07-18 21:54:36.96   Mean   :2019-07-18                     
 3rd Qu.:2019-10-26 08:14:00.00   3rd Qu.:2019-10-26                     
 Max.   :2020-01-01 05:13:00.00   Max.   :2020-01-01                     
 NA's   :82                       NA's   :82                             
    State               City             MonthYear                     
 Length:185950      Length:185950      Min.   :2019-01-01 00:00:00.00  
 Class :character   Class :character   1st Qu.:2019-04-01 00:00:00.00  
 Mode  :character   Mode  :character   Median :2019-07-01 00:00:00.00  
                                       Mean   :2019-07-03 12:45:27.72  
                                       3rd Qu.:2019-10-01 00:00:00.00  
                                       Max.   :2020-01-01 00:00:00.00  
                                       NA's   :82                      
```

```
# A tibble: 6 x 21
  Product    Quantity PriceEach DateTime            Date       ZIP   State City 
  <chr>         <dbl>     <dbl> <dttm>              <date>     <chr> <chr> <chr>
1 USB-C Cha~        2      12.0 2019-04-19 08:46:00 2019-04-19 75001 TX    Dall~
2 Bose Soun~        1     100.  2019-04-07 22:30:00 2019-04-07 02215 MA    Bost~
3 Google Ph~        1     600   2019-04-12 14:38:00 2019-04-12 90001 CA    Los ~
4 Wired Hea~        1      12.0 2019-04-12 14:38:00 2019-04-12 90001 CA    Los ~
5 Wired Hea~        1      12.0 2019-04-30 09:27:00 2019-04-30 90001 CA    Los ~
6 USB-C Cha~        1      12.0 2019-04-29 13:03:00 2019-04-29 94016 CA    San ~
# i 13 more variables: MonthYear <dttm>, TotalPopulation <dbl>,
#   MedianHHIncome <dbl>, PCIncome <dbl>, MedianAge <dbl>, Race_White <dbl>,
#   Race_Black <dbl>, Race_American_Indian <dbl>, Race_Asian <dbl>,
#   Race_Pacific_Islander <dbl>, Race_Other <dbl>, Ethnicity_Hispanic <dbl>,
#   Citizens <dbl>
```

- **ZIP Info Summary: `zip_info_summary`**
  - Provides a detailed overview of the demographic data, including population, income, age, and racial composition for each ZIP code.
- **Sales Summary:** `sales_summary`

  - Offers a glimpse into the overall sales data, including the number of products sold and price details.

- **Sales Distribution by ZIP Code: `sales_by_zip`**
  - Shows the number of sales transactions for each ZIP code, which can help identify areas with higher sales activity.

- **Merged Data: `merged_data`**
  - Combines sales data with demographic information, enabling deeper analysis of how demographics might influence sales patterns.

Visualization

- **Sales Trends Over Time:**
  - Analyze and visualize sales trends over different time periods (e.g., daily, monthly).

- **Popular Products Analysis:**
  - Identify and visualize the most popular products and their sales distribution across different ZIP codes.

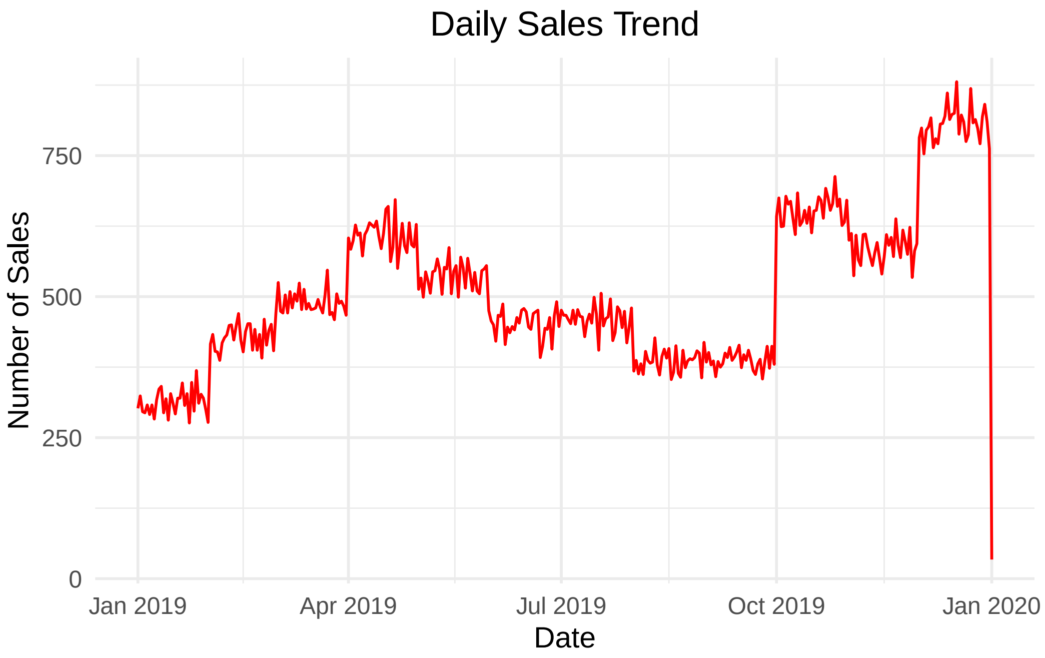
- **Demographic Influence:**
  - Explore and visualize the relationship between sales and demographic factors such as income, age, and population.

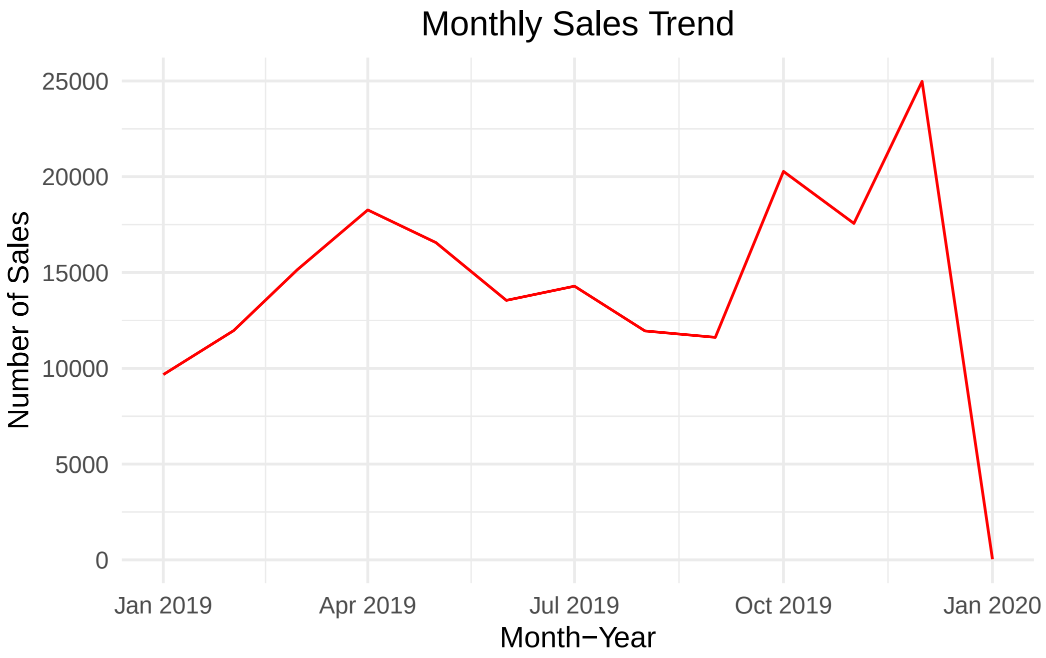
- **Sales by City and State:**
  - Visualize sales distribution across different cities and states to identify regional trends.

- **Product Sales by Time**:
  - Analyze how the sales of top products vary over time to identify any seasonal or periodic trends.

Sales Trend Over Time

Visualize sales trends over time to identify any patterns or spikes in sales activity. (Daily, Monthly)

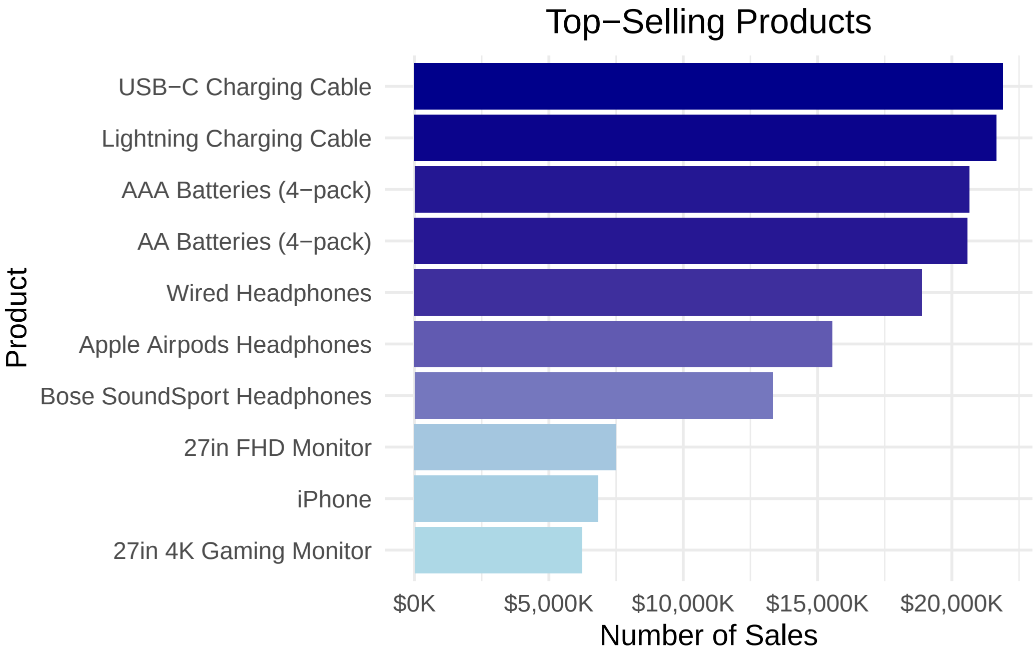




- **Daily Sales Trend:**
  - The line plot shows the fluctuation of sales on a daily basis. There may be visible peaks or troughs which could correlate with specific events, holidays, or promotional periods.
- **Monthly Sales Trend:**
  - The monthly trend provides a broader view of sales activity over time, helping to identify any long-term trends or seasonal patterns.
  - A noticeable increase in sales during the holiday season in December.

Popular Product Analysis

Identify the top-selling products and visualize their sales distribution.

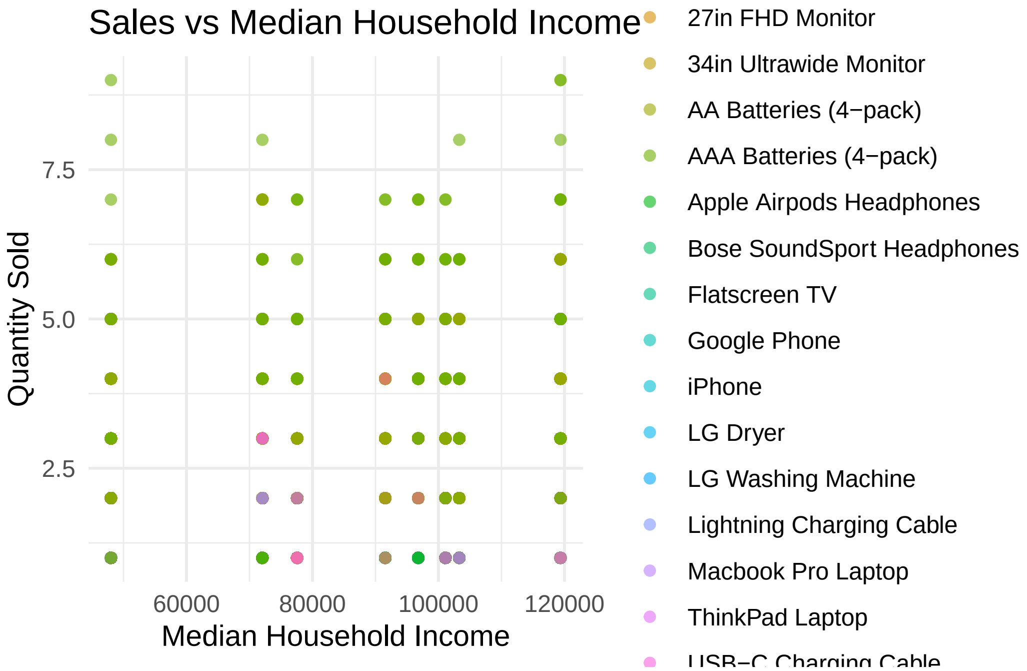


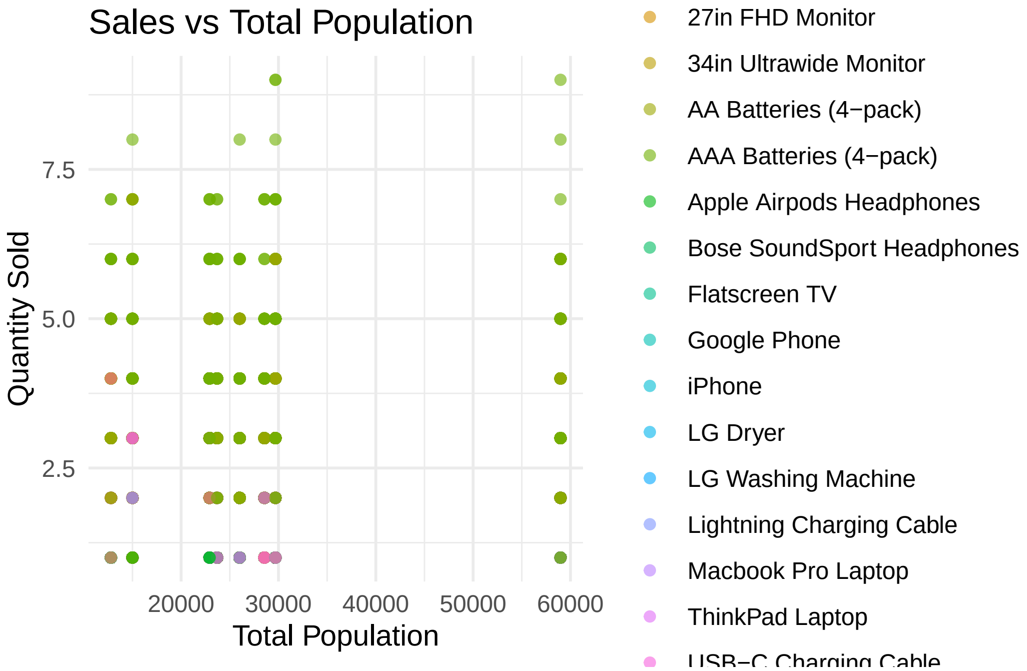
Top-Selling Products:

- The bar chart highlights the most popular products based on the number of sales from darkest to lightest color emphasizing sales. This can help identify key products driving revenue. (Focus on high-demand items)
- The top products by sales volume are identified, e.g., “USB-C Charging Cable” and “Bose SoundSport Headphones”.

Demographic Influence

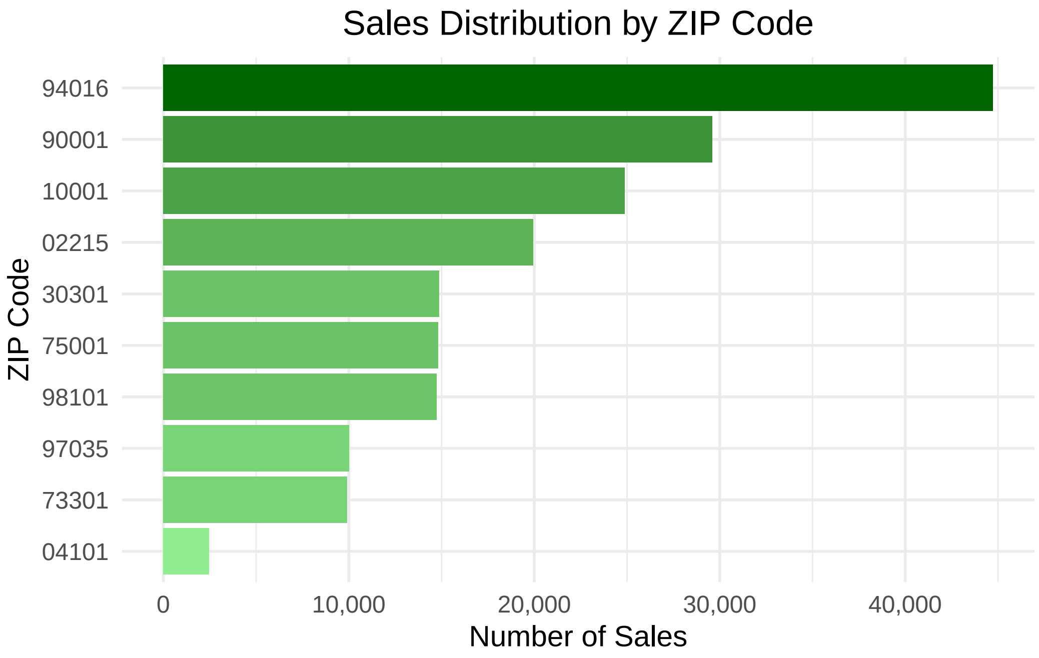
Explore and visualize the relationship between sales and demographic factors such as median household income and total population.

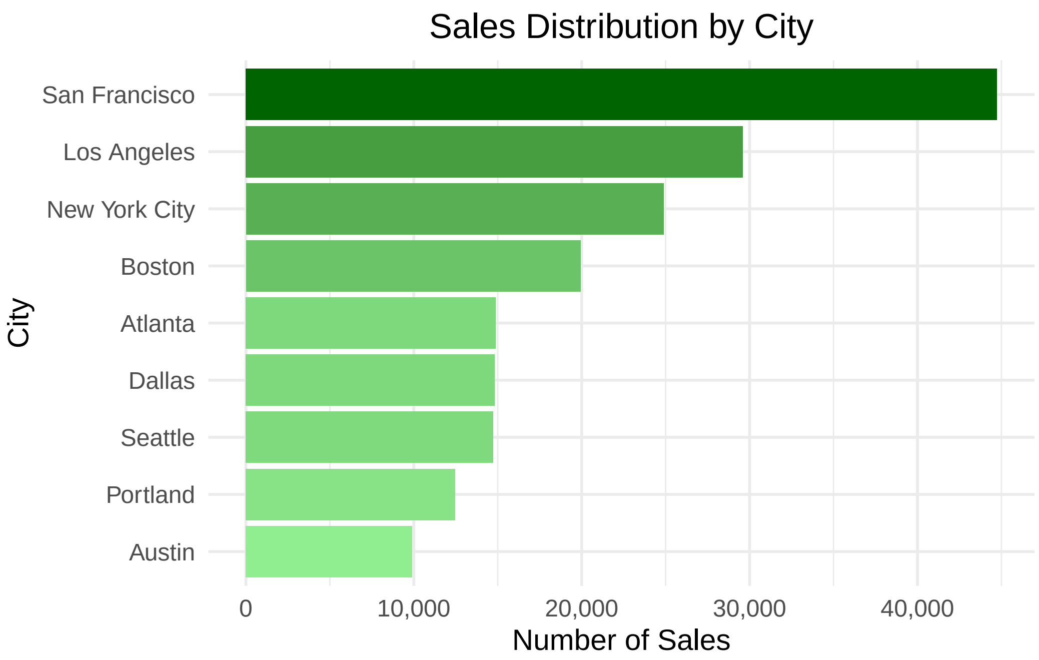




- **Sales versus Median Household Income:**
  - The scatter plot shows the relationship between the quantity of products sold and the median household income in different ZIP codes. Different colors represent different products. (This can reveal purchasing power and target markets)
  - Higher sales in ZIP codes with higher median household income.
- **Sales versus Total Population:**
  - This scatter plot illustrates how the quantity of products sold correlates with the total population in various ZIP codes which helps us understand market size and potential.
  - Densely populated areas showing higher sales volumes.5.4 Sales Distribution by ZIP Code and City

Visualize how sales are distributed across different ZIP codes and cities.

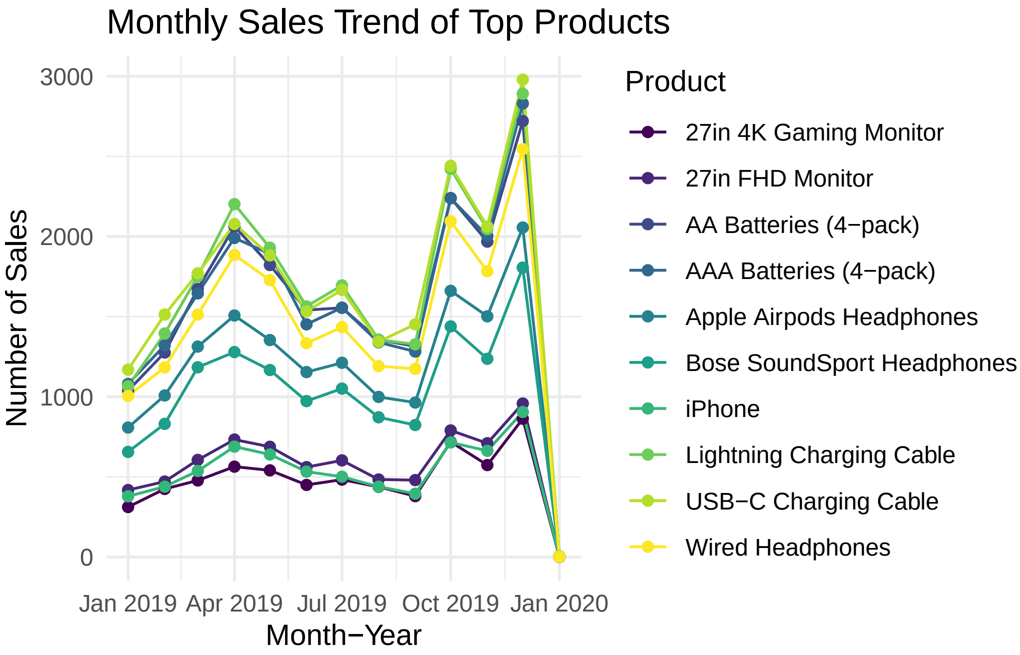




- **Sales Distribution by ZIP Code:**
  - The count plot shows how sales are distributed across different ZIP codes, highlighting areas with higher or lower sales activity.
  - You can see that ZIP Code like 94016 and 90001 shows significant sales activity.
- **Sales Distribution by City:**
  - The count plot illustrates the distribution of sales across different cities, providing insights into regional sales performance.
  - Cities like SF and LA shows as being major sales hubs.

Product Sales by Time

Trends in the sales of top products over time highlight any seasonal patterns, aiding in inventory and marketing planning.



Monthly Sales Trend of Top Products:

- The line plot shows how the sales of the top products vary over time, highlighting any seasonal patterns or trends for specific products.
- Monthly sales trends of top products highlight seasonal preferences.
- Increased sales of "Bose SoundSport Headphones" during summer months, possibly due to outdoor activities.

Conclusion

In this analysis, we explore the sales trends and patterns of Amazon's technology products across various urban ZIP codes in 2019. Our dataset includes product names, quantities sold, prices, dates, and locations of the sales, as well as demographic information for the ZIP codes. Our goal is to uncover key insights about consumer behavior and market dynamics in different urban areas.

After cleaning and processing the data, we begin with a high-level overview and proceed with analyzing daily and monthly sales trends. We observe fluctuations in daily sales with noticeable peaks during certain periods, likely corresponding to events, holidays, or promotions. Monthly trends show a significant spike during the holiday season in December, indicating increased consumer spending during that time.

Next, we identify the top-selling products, such as “USB-C Charging Cable” and “Bose SoundSport Headphones,” which are key drivers of revenue. We then explore the relationship between sales and demographic factors, finding that higher sales volumes are associated with ZIP codes that have higher median household incomes and larger populations. This suggests that income levels and population density play significant roles in purchasing behavior.

We also visualize sales distribution across different ZIP codes and cities, highlighting areas with higher sales activity. Notably, San Francisco and Los Angeles emerge as major sales hubs. Finally, we analyze the sales trends of top products over time, uncovering seasonal preferences such as increased sales of “Bose SoundSport Headphones” during the summer months, likely due to outdoor activities.

In conclusion, our analysis reveals that sales patterns are influenced by income levels, population density, and seasonal trends. These insights can help Amazon optimize inventory management and tailor marketing strategies to different urban areas, ultimately enhancing sales performance and customer satisfaction.