Pet Box Subscription Analysis

By: AJANAKU AYOMIDE

Unsplash

Background Information

PetMind is a retailer of products for pets. They are based in the United States. PetMind sells products that are a mix of luxury items and everyday items. Luxury items include toys, everyday items include food. The company wants to increase sales by selling more everyday products repeatedly.They have been testing this approach for the last year. They now want a report on how repeat purchases impact sales.

Executive Summary

This project sheds light on the distribution of purchases, sales, and the relationship between repeat purchases and sales across different product categories. It provides essential insights for understanding the market dynamics and making informed decisions to optimize sales and business performance.

Methodology & Tool Used

The analysis of the PetMind dataset will be done using R. Necessary visualizations will also be done using the ggplot2 package contained in R. Each phase of the analysis will have some objectives/tasks that all need to be achieved before moving on to the next step of the analysis.

Data Pre-Processing

Import the dataset and view the first 10 rows.

```r
pet_supplies <- read.csv("D:/Users/User/Documents/R practice/scripts/pet_supplies_2212.csv")
```

```r
head(pet_supplies, 10) # glimpse of the first 5 rows of the dataset to understand what's contained in i
```

```
##    product_id  category animal   size price   sales rating repeat_purchase
## 1           1      Food   Bird  large  51.1 1860.62      7               1
## 2           2   Housing   Bird MEDIUM 35.98  963.60      6               0
## 3           3      Food    Dog medium 31.23  898.30      5               1
## 4           4  Medicine    Cat  small 24.95  982.15      6               1
## 5           5   Housing    Cat  Small 26.18  832.63      7               1
## 6           6   Housing    Dog  Small 30.77  874.58      7               0
## 7           7   Housing    Dog  Small 31.04  875.07      5               0
## 8           8      Toys    Cat medium  28.9 1074.31      4               0
## 9           9 Equipment   Fish MEDIUM 17.82  503.67      5               0
## 10         10  Medicine    Dog medium 24.93  838.88      8               0
```

The pet supplies dataset contains 1500 rows and 8 columns of data.

Objective 1

For every column in the data: a. Check whether the values match the description given in the data table. b. State the number of missing values in the column. c. Describe what you did to make values match the description if they did not match.

1. *product_id*: has an integer datatype and the numbers are serially arranged from 1-1500, therefore no missing values.
2. *category*: it is of character datatype and has 1500 rows. 25 of those values were not specified/missing and were represented by '-', which was then changed to 'Unknown'.
3. *animal*: the animal category contained 0 missing values and all values correctly match-up with those in the data table.
4. *size*: the size column had no missing values. However, the formatting was inconsistent with that specified in the data table, so all values were converted to lower case thereby solving the problem of inconsistent formatting.
5. *price*: the price column was initially in character datatype, but this is not appropriate for the kind of data it contains so it was converted to a numeric datatype. This introduced nulls into column, and the null values were replaced with the overall median value, all then rounded to 2 dp. There were no negative whole numbers in the price column.
6. *sales*: the sales column is of numeric datatype and it contains no missing values.
7. *rating*: The values in this column range from 1-9 which is consistent with the description. There were 150 missing values. All values were replaced with 0.
8. *repeat_purchase*: this column has values between 0 and 1. There were no missing values, hence it complies with the description.

Data Cleaning

The steps taken to clean the data to arrive at the description given above are highlighted in the chunks of code in the code section at the end. - Import the tidyverse and dplyr libraries that will be used for data cleaning and manipulation

```r
#import the necessary libraries
library(tidyverse)
library(ggplot2)
```

- The size column has values with different cases, so convert all of them to lowercase to ensure consistent formatting.

```r
# change the values of the size column to lowercase to ensure consistent formatting
pet_supplies$size <- tolower(pet_supplies$size)
unique(pet_supplies$size)
```

```
## [1] "large"  "medium" "small"
```

- According to the data table, there should be 6 unique categories contained in the category column. The unknown ones in this case are represented with "-" which will then be changed to "Unknown".

```r
# some rows in the category column contain '-' which have to be changed to 'Unknown'
pet_supplies <- pet_supplies %>%
  mutate(category = ifelse(category == "-", "Unknown", category))

unique(pet_supplies$category)
```

```
## [1] "Food"      "Housing"   "Medicine"  "Toys"      "Equipment" "Accessory"
## [7] "Unknown"
```

- To confirm the total number of rows that were changed from "-" to "Unknown"

```r
# 25 values have been changed
pet_supplies %>%
  filter(category == 'Unknown') %>%
  summarise(n())
```

```
##   n()
## 1  25
```

- Inspect the animal column for unique values and then check if any are missing.

```r
# check for missing data in the animal column
unique(pet_supplies$animal)
```

```
## [1] "Bird" "Dog"  "Cat"  "Fish"
```

```r
pet_supplies %>%
  summarise(sum(is.na(animal)))
```

```
##   sum(is.na(animal))
## 1                  0
```

- Inspect the size column for unique values and then check if any are missing

```r
# check for missing data in the size column
unique(pet_supplies$size)
```

```
## [1] "large"  "medium" "small"
```

```r
pet_supplies %>%
  summarise(sum(is.na(size)))
```

```
##   sum(is.na(size))
## 1                0
```

- the price column has to be converted from character type to numeric type

```r
# the price column has been converted from character datatype to numeric.
pet_supplies$price <- as.numeric(pet_supplies$price)
```

```
## Warning: NAs introduced by coercion
```

```
median <- median(pet_supplies$price, na.rm = TRUE)
r_med <- round(median, 2)
```

```
# this introduced some nulls which will be replaced by the median value calculated
pet_supplies$price <- ifelse(is.na(pet_supplies$price)|!is.numeric(pet_supplies$price), r_med, pet_suppl
```

```
class(pet_supplies$price)
```

```
## [1] "numeric"
```

```
pet_supplies$price <- round(pet_supplies$price,2)
```

- check for missing values in the sales column

```
# check for missing values in the sales column
pet_supplies %>%
  filter(is.na(sales)) %>%
  summarise(n())
```

```
##   n()
## 1   0
```

```
# assign 0 to missing values in the rating column
pet_supplies$rating <- ifelse(is.na(pet_supplies$rating), 0, pet_supplies$rating)
```

```
pet_supplies$rating[is.na(pet_supplies$rating)] <- 0
```

- check for missing values in the repeat purchase column

```
# check for missing values in the repeat purchase column
pet_supplies %>%
  filter(is.na(repeat_purchase)) %>%
  summarise(n())
```

```
##   n()
## 1   0
```

- In the rating column, missing values are to be replaced with 0

Analysis Phase

The data cleaning has been properly handled and it's time to delve into the data to try to understand the volume of repeat purchases across different products. This analysis will be done using a visualization which I will then draw inferences from. There are 3 objectives for this phase.

Objective 1

Create a visualization that shows how many products are repeat purchases. Use the visualization to:

- State which category of the variable repeat purchases has the most observations
- Explain whether the observations are balanced across categories of the variable repeat purchases

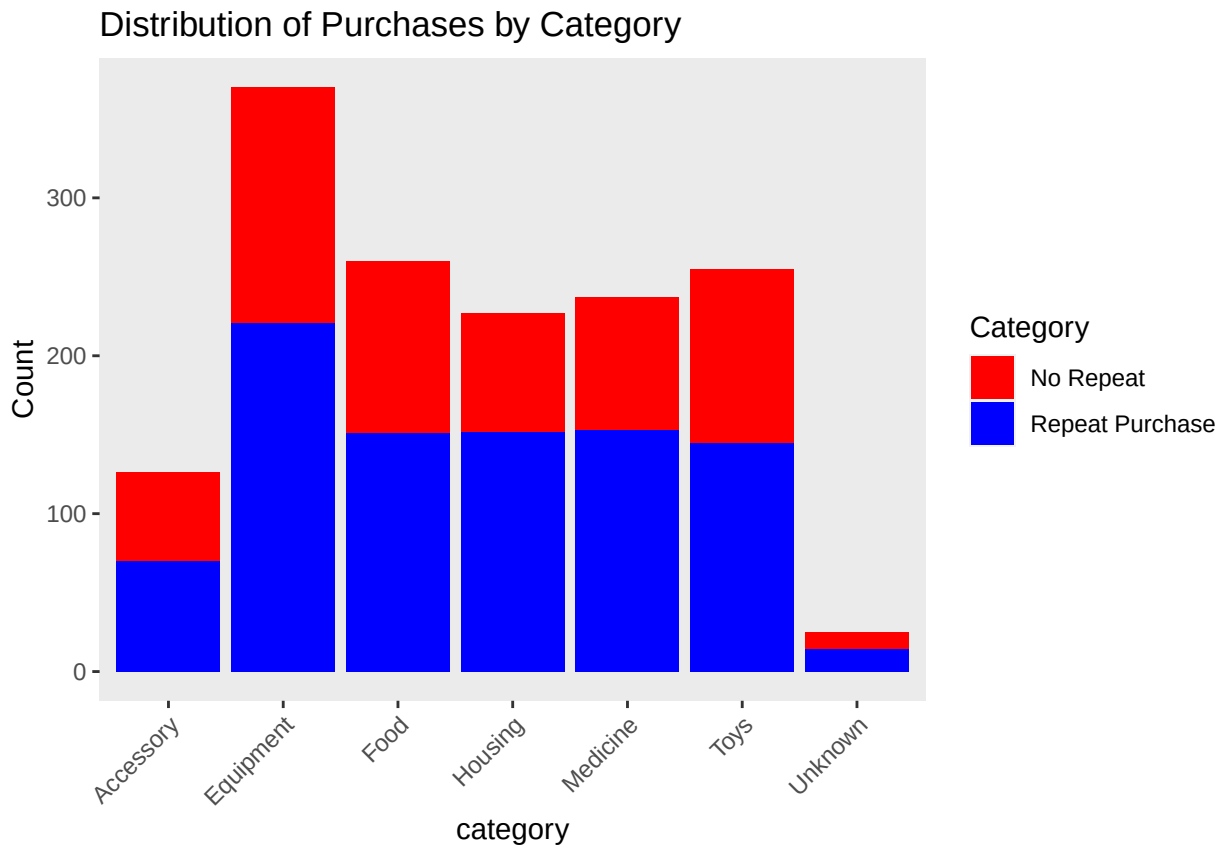
To create this visualization, a table which shows repeat purchases across different categories is needed

```
# create a dataframe for repeat purchases and category
viz <- table(pet_supplies$category, pet_supplies$repeat_purchase)
viz_df <- data.frame(viz)
colnames(viz_df) <- c("category", "repeat_purchase", "frequency")
head(viz_df)
```

```
##    category repeat_purchase frequency
## 1 Accessory               0        56
## 2 Equipment               0       149
## 3      Food               0       109
## 4   Housing               0        75
## 5  Medicine               0        84
## 6      Toys               0       110
```

```
# visualize this dataframe
ggplot(viz_df, aes(category, frequency, fill = repeat_purchase)) + geom_bar(stat = "identity") + labs(x
  theme(axis.text.x = element_text(angle = 45, hjust = 1), panel.grid = element_blank()) +
  scale_fill_brewer(palette = "Set1") + scale_fill_manual(values = c('0' = "red", '1' = "blue"), labels
```

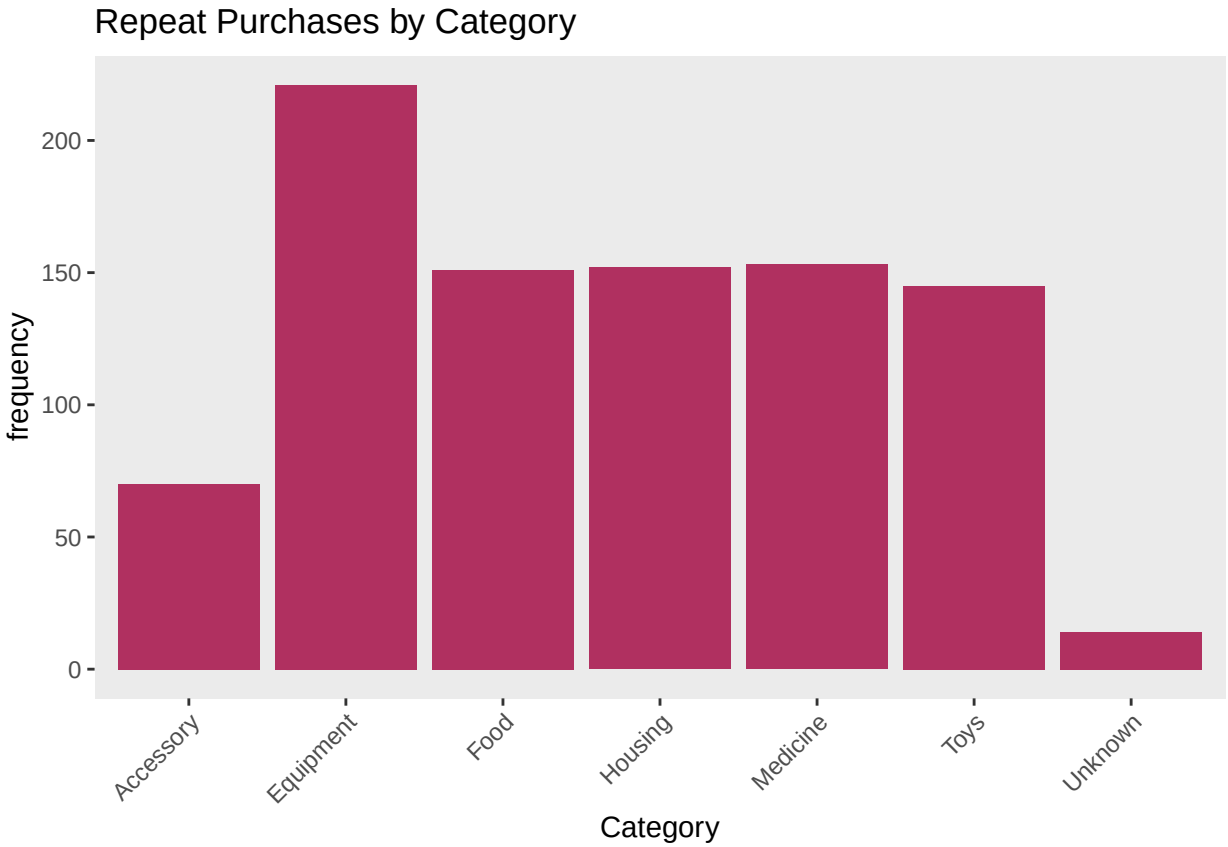
```
## Scale for fill is already present.
## Adding another scale for fill, which will replace the existing scale.
```



This visualization shows the overall distribution of purchases between the categories. It is immediately obvious that products in the *Equipment* category have had the most purchases, closely followed by *Food* and *Toy*. Further analysis which is beyond the scope of this project will be needed to determine why the *Accessory* category has very few purchases.

```
rpur <- pet_supplies %>%
  filter(repeat_purchase == 1) %>%
 group_by(category) %>%
  summarise(frequency = n())

ggplot(rpur, aes(category, frequency)) + geom_bar(stat = "identity", fill = "maroon") + labs(title = "R
```

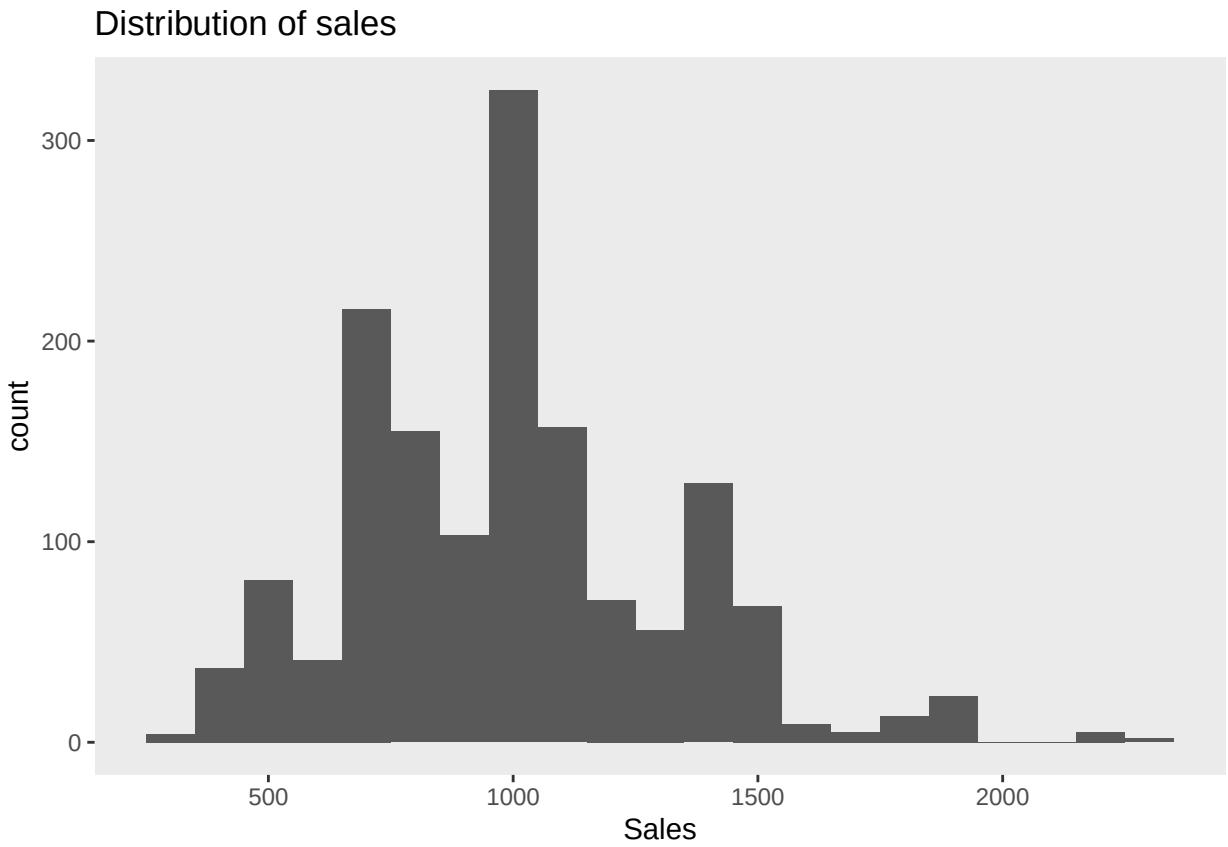


Of the 6 possible categories, repeat purchases were mostly observed in the Equipment category. Similar amount of observations were made across the other categories except Accessories. Since the repeat purchases are skewed towards the Equipment category, the observations are fairly unbalanced.

Objective 2

- Describe the distribution of all of the sales. This should include a visualization that shows the distribution.

```
ggplot(pet_supplies, aes(x = sales)) + geom_histogram(binwidth = 100) + theme(panel.grid = element_blank
```

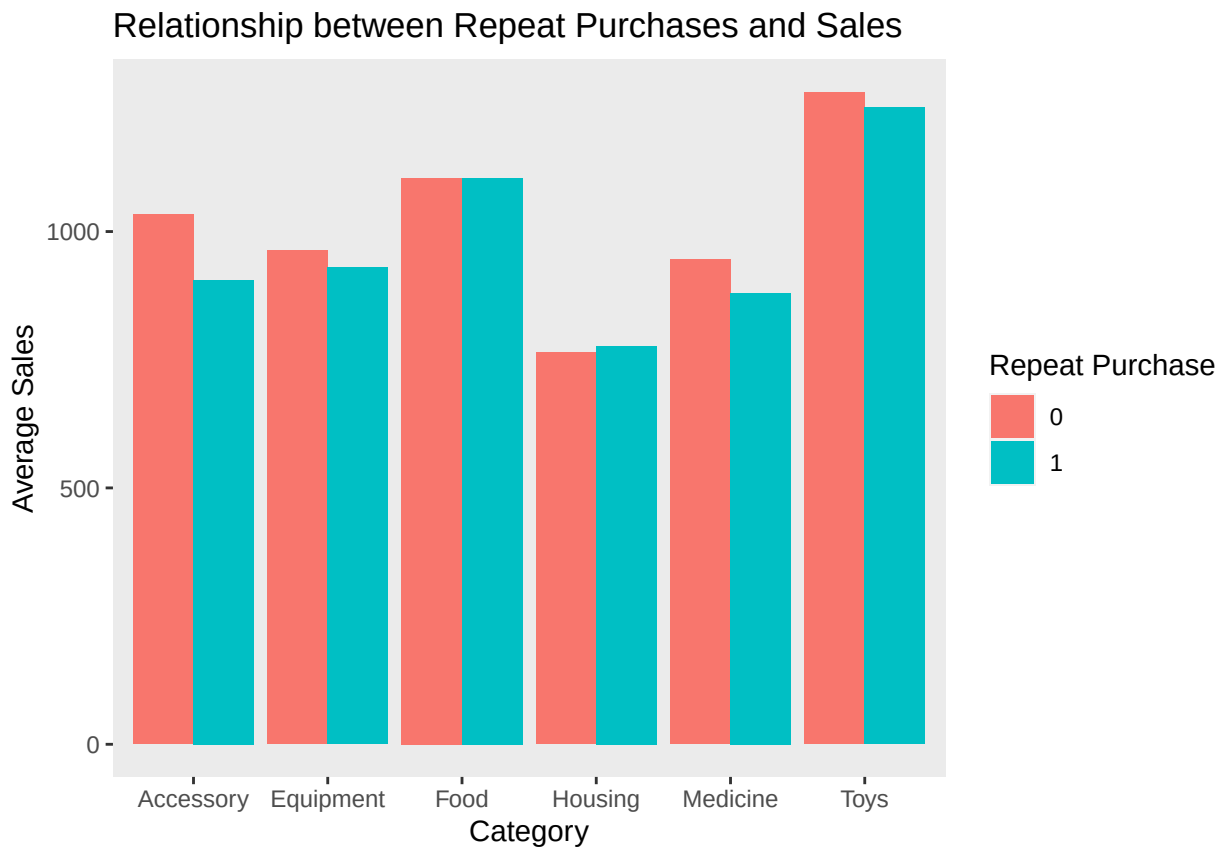


The distribution shows that a high proportion of sales range from about 600 to 1300, with the peak for any product being at the 1000 mark. Most products had less than 2000 sales, while very few had more than 2000. It is important to note this benchmark, so the success or failure of new products can be judged using this.

Objective 3

Describe the relationship between repeat purchases and sales. This must also include a visualization to demonstrate the relationship.

```
p <- pet_supplies %>%
  group_by(category, repeat_purchase) %>%
  filter(category != 'Unknown') %>%
  summarise(avg = mean(sales))
ggplot(p, aes(category, avg, fill=factor(repeat_purchase))) + geom_bar(stat = "summary", fun = "median"
```



For most categories, the average number of sales is higher when there are no repeated purchases.

Insights

- The "Equipment" category has the most purchases, followed by "Food" and "Toy." However, the "Accessory" category has very few purchases, which may require further investigation to understand why this is the case. The observations are fairly unbalanced since repeat purchases are skewed towards the "Equipment" category.
- Most products have less than 2000 sales, while only a few have more than 2000. Knowing this benchmark is valuable for evaluating the success or failure of new products.
- Interestingly, for most categories, products without repeat purchases tend to have higher average sales. This suggests that repeat purchases might not be the primary driver of sales in these categories.

Business Recommendation

- This project's findings have practical implications for business decisions. For instance, to boost sales in the "Accessory" category, targeted marketing campaigns, product improvements, or promotions might be necessary. Additionally, focusing on customer acquisition strategies could help maximize sales in product categories where repeat purchases have less impact.