Pet Box Subscription Analysis

S A Nawash Akhtar

0.0 Introduction

PetMind is a fictional 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.

1.0 Data Validation

Initially, the dataset has 8 column and 1500 rows. Now we are going to validate dataset against the description provided and will perform cleaning process if necessary.

```
library(tidyverse)
library(data.table)
library(stringr)

my_path <- "E:\\Nawash\\Works\\Datacamp\\Data Analyst in SQL\\Practical exam\\Dataset"
```

```
setwd(my_path)
dir()
```

```
## [1] "Original dataset"      "pet_supplies_2212.csv"
```

```
petmind_data <- fread("pet_supplies_2212.csv")
petmind_data$repeat_purchase <- as.character(petmind_data$repeat_purchase)
glimpse(petmind_data)
```

```
## Rows: 1,500
## Columns: 8
## $ product_id      <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16,~
## $ category        <chr> "Food", "Housing", "Food", "Medicine", "Housing", "Hou~
## $ animal          <chr> "Bird", "Bird", "Dog", "Cat", "Cat", "Dog", "Dog", "Ca~
## $ size            <chr> "large", "MEDIUM", "medium", "small", "Small", "Small"~
## $ price           <chr> "51.1", "35.98", "31.23", "24.95", "26.18", "30.77", "~
## $ sales           <dbl> 1860.62, 963.60, 898.30, 982.15, 832.63, 874.58, 875.0~
## $ rating          <int> 7, 6, 5, 6, 7, 7, 5, 4, 5, 8, 7, 5, 4, 6, 4, 1, 5, 3, ~
## $ repeat_purchase <chr> "1", "0", "1", "1", "1", "0", "0", "0", "0", "0", "1",~
```

1.1 product_id column validation

product_id : This column has nominal values and do not have any duplicate or missing values.

```
sum(is.na(petmind_data$product_id))
```

```
## [1] 0
```

```
petmind_data$product_id[duplicated(petmind_data$product_id)]
```

```
## integer(0)
```

1.2 category column validation

category : This column contains 6 values as expected and also I found 25 rows of missing values. Replaced those missing values with the “Unknown” value.

```
sum(is.na(petmind_data$category))
```

```
## [1] 0
```

```
petmind_data$category <- str_replace(petmind_data$category, "-", "Unknown")
table(petmind_data$category)
```

```
##
## Accessory Equipment      Food   Housing  Medicine      Toys   Unknown
##       126       370       260       227       237       255        25
```

1.3 animal column validation

animal : This column contains 4 unique values and none missing values were found here.

```
sum(is.na(petmind_data$animal))
```

```
## [1] 0
```

```
table(petmind_data$animal)
```

```
##
## Bird  Cat  Dog Fish
##  197  567  367  369
```

1.4 size column validation

size : This column should be ordinal and only “Small”, “Medium” and “Large” values should be found here. In order to get that I had to change cases of the character data.

```
petmind_data <- petmind_data %>%
  mutate(size = tolower(size))
```

```
sum(is.na(petmind_data$size))
```

```
## [1] 0
```

```
table(petmind_data$size)
```

```
##
##  large medium  small
##    254    492    754
```

1.5 price column validation

price : I converted the data type from character to numeric and missing values were replaced with overall median price.

```
sum(is.na(petmind_data$price))
```

```
## [1] 0
```

```
petmind_data$price <- as.numeric(petmind_data$price)
med <- round(median(petmind_data$price, na.rm = TRUE), 2)
med
```

```
## [1] 28.06
```

```
petmind_data$price <- replace_na(petmind_data$price, med)
summary(petmind_data$price)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12.85   25.00   28.06   29.29   33.14   54.16
```

```
view(petmind_data$price)
```

1.6 sales column validation

sales : None missing values were found here and values followed the description provided.

```
sum(is.na(petmind_data$sales))
```

```
## [1] 0
```

```
summary(petmind_data$sales)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   286.9   747.4  1000.8   996.6  1158.2  2256.0
```

```
view(petmind_data$sales)
```

1.7 rating column validation

rating : Here I found 150 missing values and replaced those values with 0.

```
petmind_data$rating <- replace_na(petmind_data$rating, 0)
sum(is.na(petmind_data$rating))
```

```
## [1] 0
```

```
table(petmind_data$rating)
```

```
##
##   0   1   2   3   4   5   6   7   8   9
## 150  12  43 190 283 304 299 143  61  15
```

1.8 repeat_purchase column validation

repeat_purchase : This column has only two values 0 and 1 as expected and there are none missing values.

```
sum(is.na(petmind_data$repeat_purchase))
```

```
## [1] 0
```

```
table(petmind_data$repeat_purchase)
```

```
##
##   0   1
## 594 906
```

2.0 Data analysis and visualization

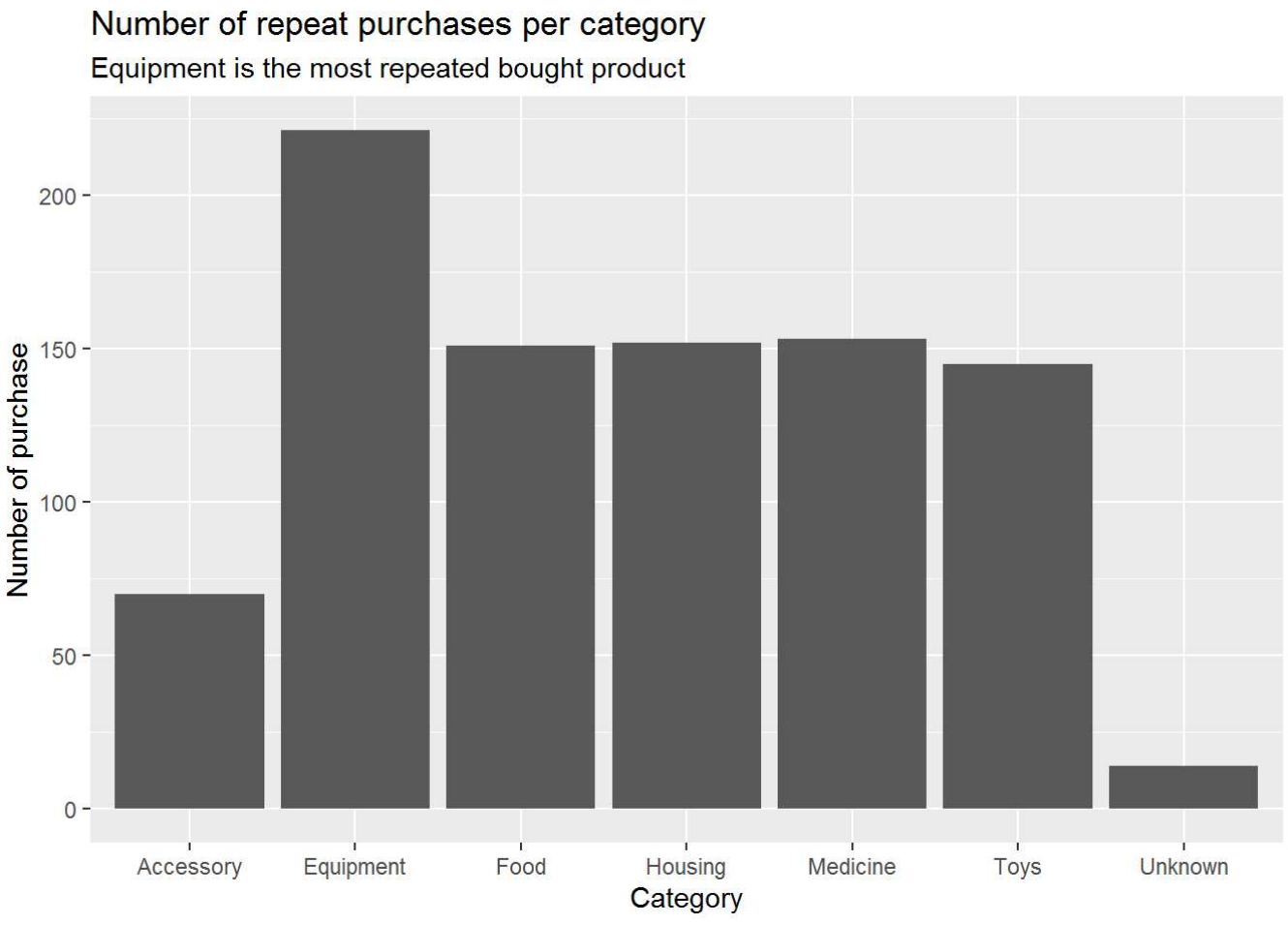
Data analysis process is explained with necessary visualization.

2.1 How many product are repeat purchase?

There are 7 type of category available for analysis including the “Unknown” category. The most repeated bought products is in the equipment category. The rest of the categories food, housing, medicine and toys has a balanced observation apart from the accessory category which has the least repeat purchase. This would suggest that, based on the sale of the previous year the company should put more focus on the equipment category for more repeat sells in future.

```
df1 <- petmind_data %>%
  filter(repeat_purchase == "1") %>%
  group_by(category) %>%
  summarise(purchased_num = n())  %>%
  ggplot(aes(x = category, y = purchased_num)) +
  geom_bar(position = "dodge", stat = "identity") +
  labs(title = "Number of repeat purchases per category",
       subtitle = "Equipment is the most repeated bought product",
       x = "Category", y = "Number of purchase")

df1
```




```
ggsave("001.jpg")
```

```
## Saving 7 x 5 in image
```

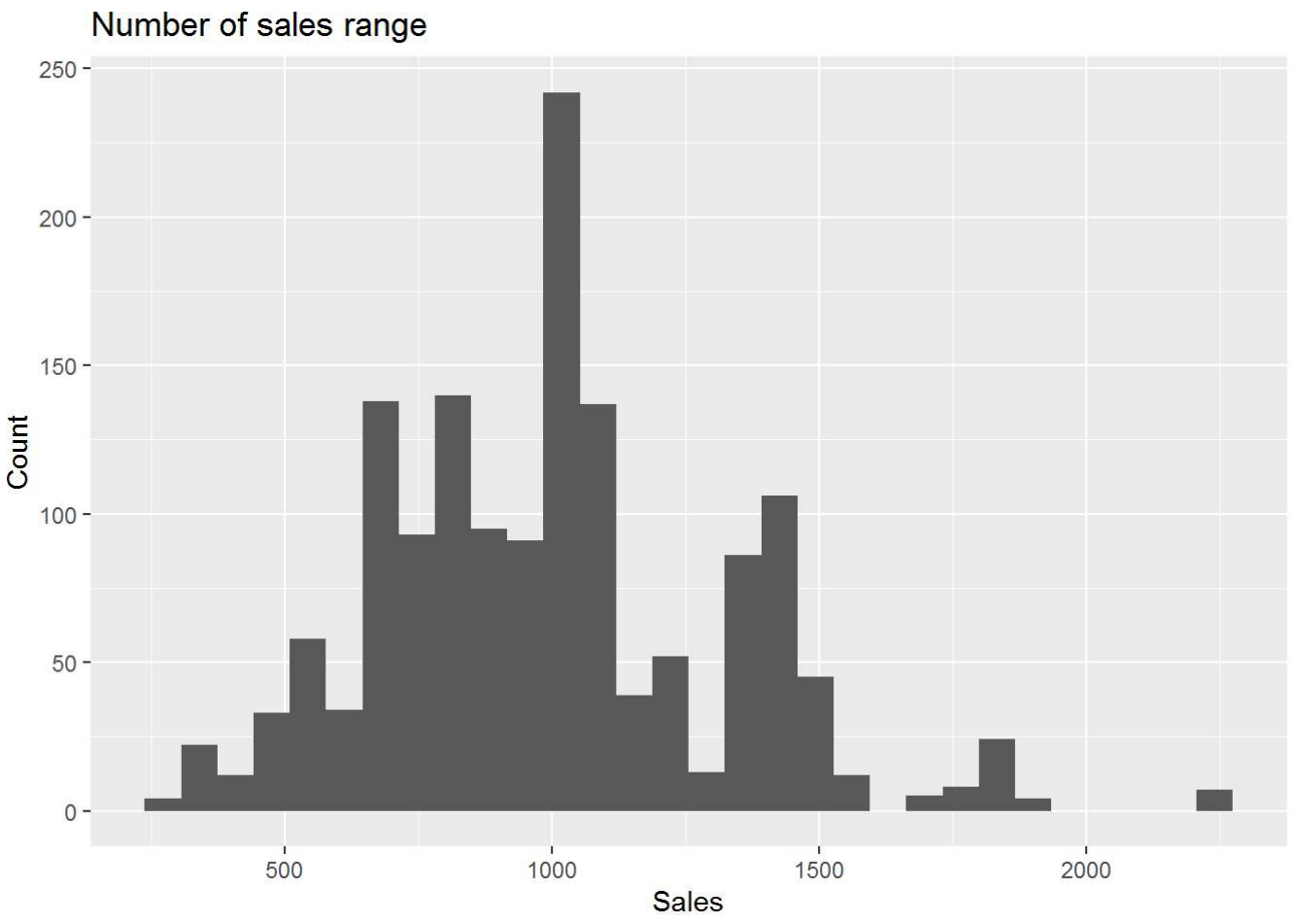
2.2 Distribution of number of sales

We need to observe this distribution to see how the company approach was achieved. Looking at the distribution we can see that most of sold product has sales between 700-1100 range. There are some products which sold over 1600 but they are very uncommon. When looking to increase sales over time the company must focus to the products which has a sale around 1000.

```
df2 <- petmind_data %>%
  ggplot(aes(x = sales)) +
  geom_histogram() +
  labs(title = "Number of sales range",
       x = "Sales", y = "Count")

df2
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
ggsave("002.jpg")
```

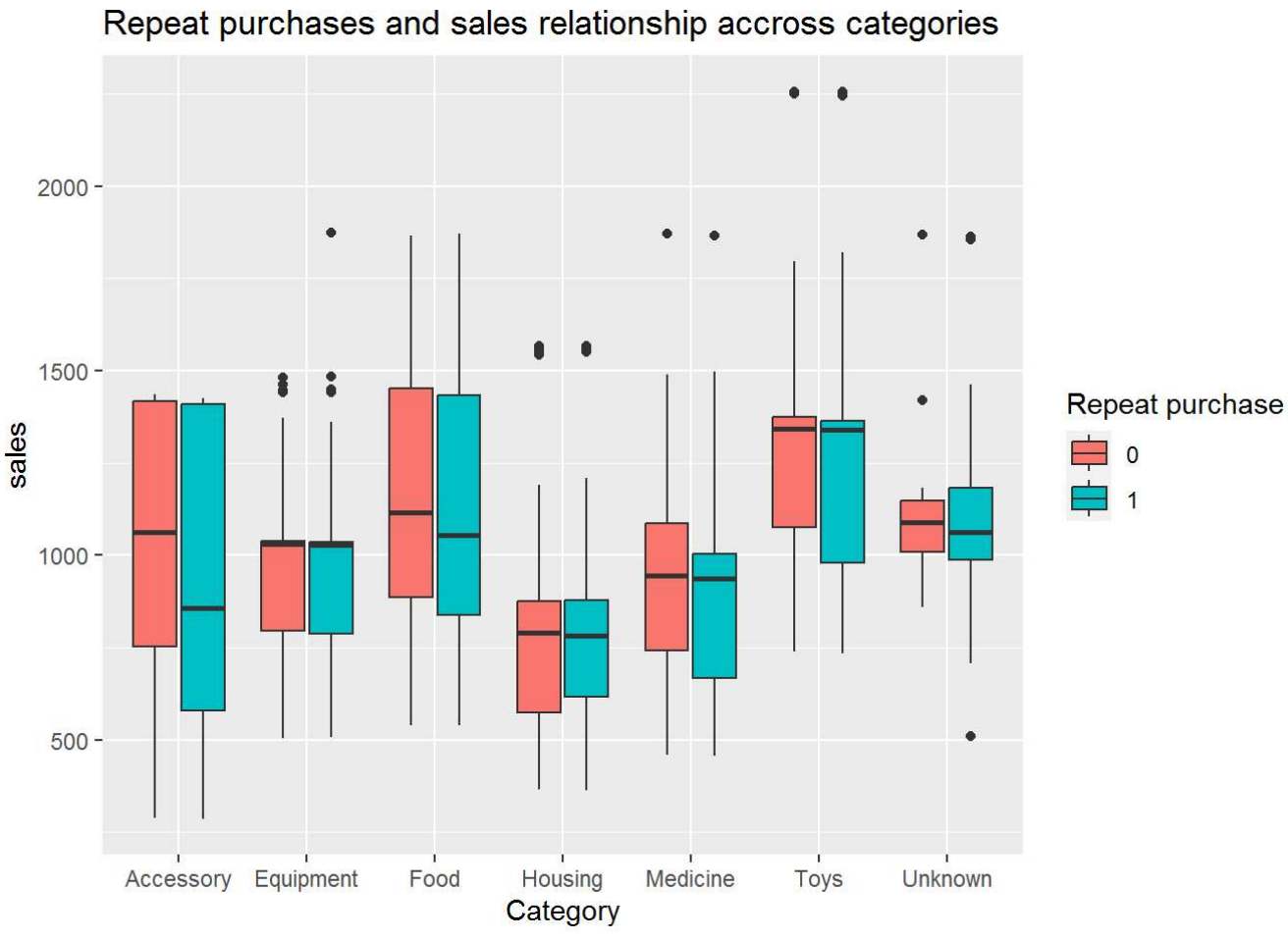
```
## Saving 7 x 5 in image
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

2.3 Relationship between repeat purchase and sales

Now we are going to explore the relationship between repeat purchase and sale range for all the categories. Although from previous plot we saw that the equipment category was most repeated bought product but from this plot we can see that the everyday item food has the most interquartile range than other categories. The luxury item toys has the highest outlier as expected. But comparing to other categories, food category has wide range of sales revenues.

```
df3 <- petmind_data %>%
  group_by(category) %>%
  ggplot(aes(x = category, y = sales, fill = repeat_purchase)) +
  geom_boxplot() +
  labs(title = "Repeat purchases and sales relationship accross categories",
       x = "Category", y = "sales", fill = "Repeat purchase")

df3
```




```
ggsave("003.jpg")
```

```
## Saving 7 x 5 in image
```

3.0 Recommendation

Based on above plots we can recommend that the company can put their focus on the food category as it has wide range of sale revenue. Also the company can keep an eye on equipment category as it has the highest repeat amount of selling but has a shortened sales revenue range. Further analysis can be done on the accessory category as it has the lowest repeat selling amount but has good range in sales revenue.
