Chapter 5: Use Case Notebook for Instructors

Ram Gopal, Dan Philps, and Tillman Weyde

Summer 2022

Contents

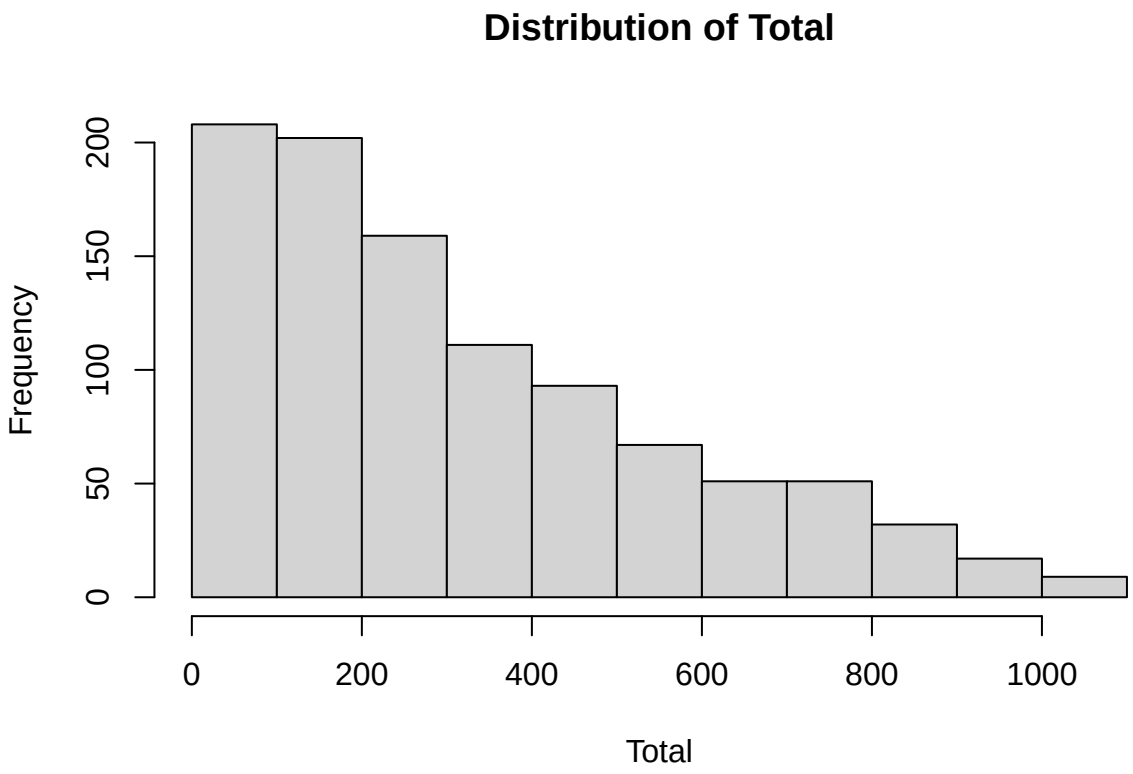
Use Case: Probability and Distributions in Sales Data

In this use case we use a small database that reports supermarket sales of various products across three stores. The data includes a customer rating (rating), as well as other key sales data items. We will study if there are any important patterns that we can exploit.

```
library(readxl)
supermarket_sales <- read_excel("../../data/supermarket_sales.xlsx")
```

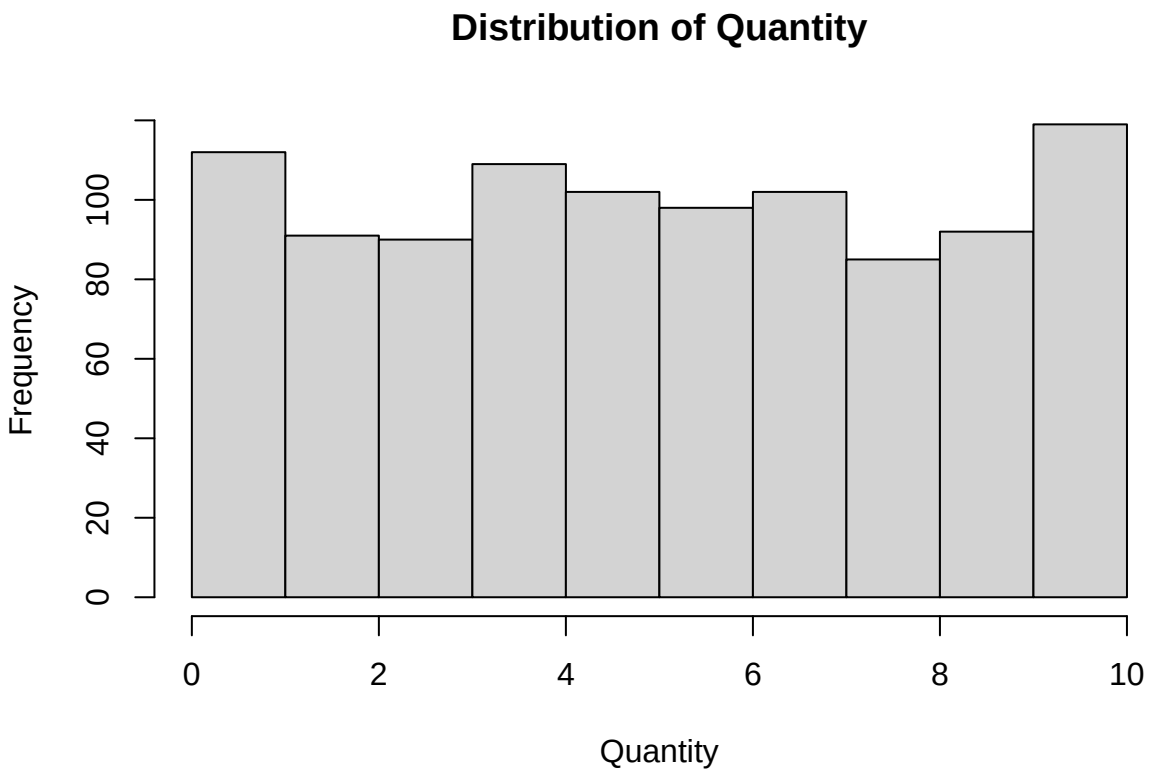
We can examine the distribution of key data items such as "Total", the cash value of a customer transaction. When we examine Total, the frequency of transactions looks similar to the exponential distribution we explored previously in this chapter.

```
hist(supermarket_sales$Total,main = "Distribution of Total",
     xlab = "Total")
```



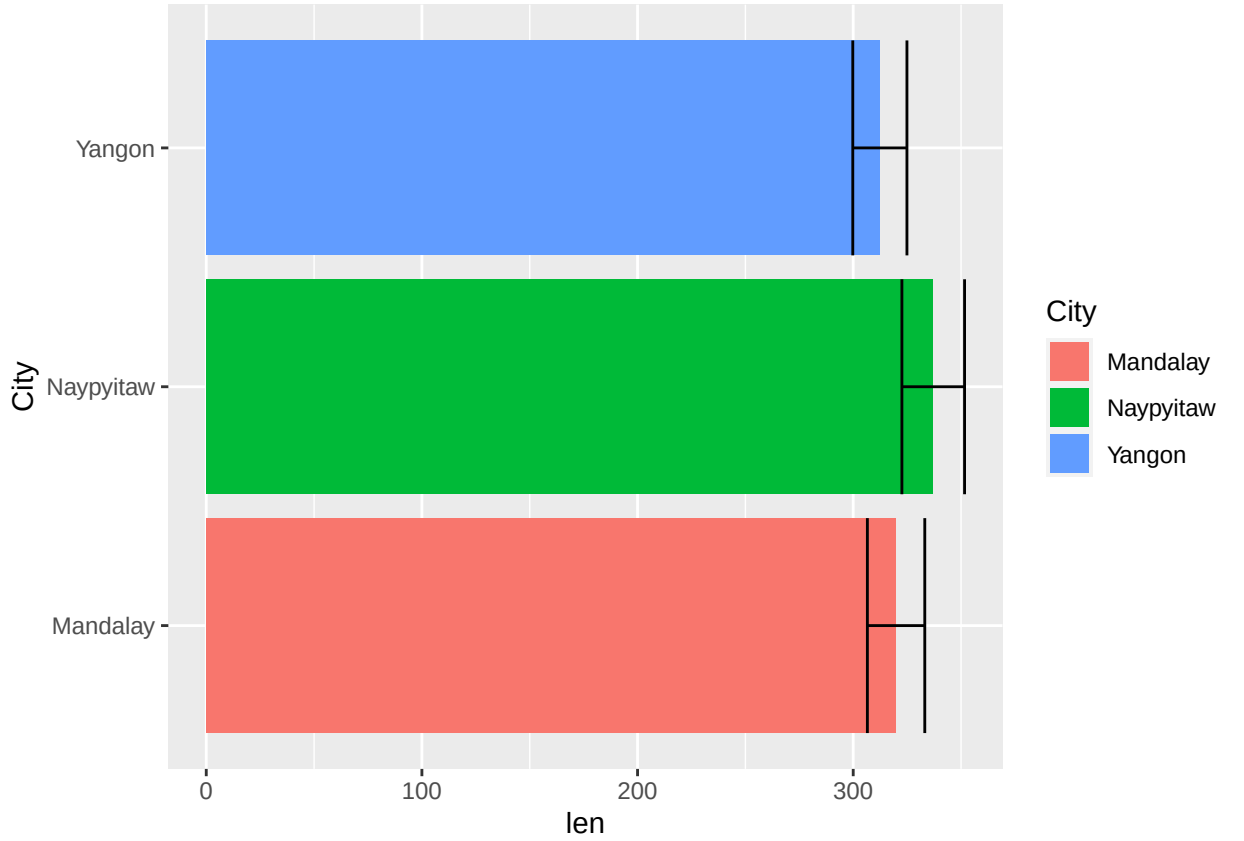
We can also examine the distribution of Quantity in our transactions, and see that this is a very different distribution. We notice that all the transactions tend to be of random size and evenly distributed with exception for large quantity sales that are significantly more frequent. In other words, one in ten transactions were of a significantly higher Quantity. It could be useful to understand what drives this:

```
hist(supermarket_sales$Quantity,main = "Distribution of Quantity",
     xlab = "Quantity",breaks = seq(0,10))
```



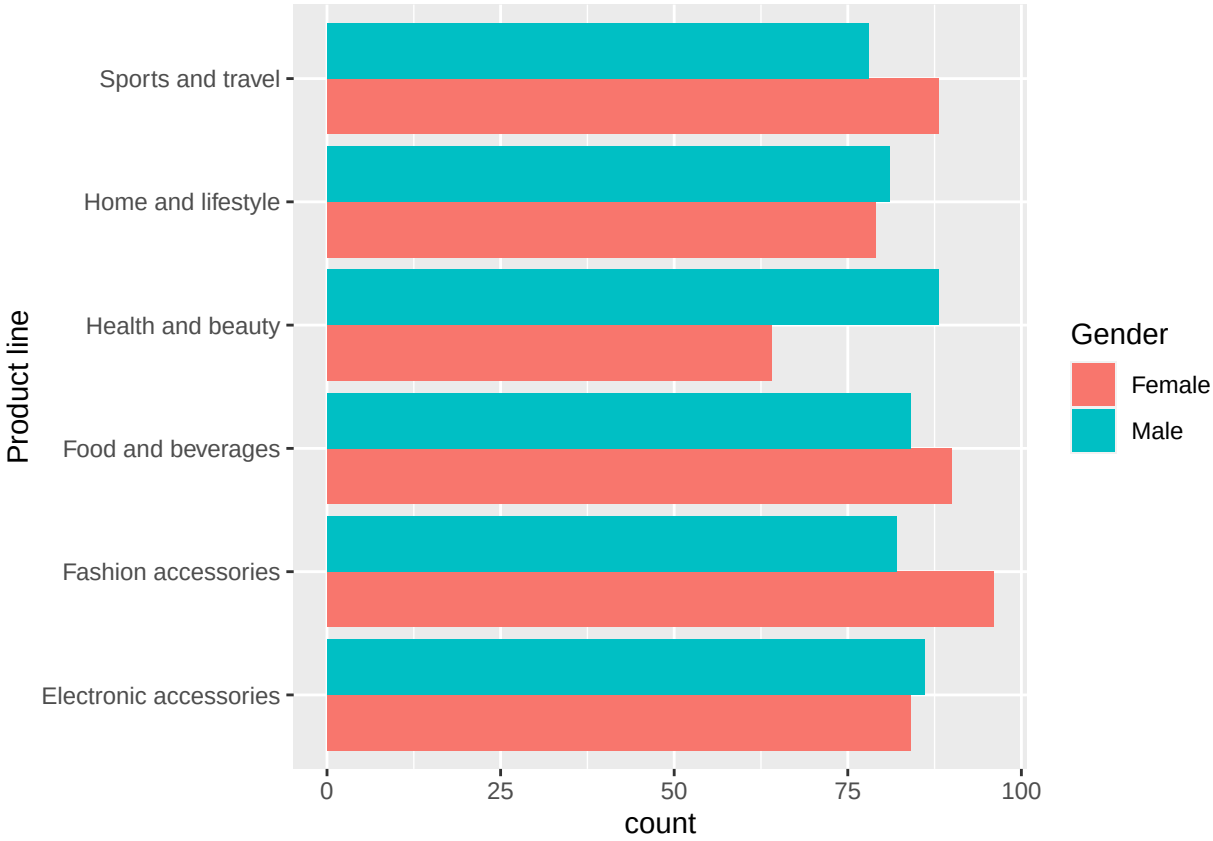
If we look at the distribution of sales using the categorical data item "City", we can understand how sales are distributed across the three cities where the supermarkets are. In addition, we can show a measure of confidence for each category, as a line plotted over the horizontal bars to represent the degree of uncertainty around that estimate (i.e., error bars). The larger the line the greater the degree of uncertainty:

```
library(dplyr)
library(ggplot2)
df = supermarket_sales %>% group_by(City)%>%summarise(
    sd = sd(Total),
    len = mean(Total),
    num = n()
  )
ggplot(df,aes(x =City, y = len,fill=City)) +
  geom_col(size = 1)+
  geom_errorbar(data = df, aes(x=City, ymin = len-sd/sqrt(num),
                               ymax = len+sd/sqrt(num)))+
  coord_flip()
```



We can also drill down into categories within the dataset to examine the distribution of sales. Here we show how sales vary by gender in different product areas, which could be used to use to optimize our promotions and marketing, for example:

```
ggplot(supermarket_sales,aes(x = `Product line`)) +
  geom_bar(aes(fill = Gender),position = "dodge") +
  coord_flip()
```


Sports and travel
Home and lifestyle
Health and beauty
Food and beverages
Fashion accessories
Electronic accessories
Product line
0
25
50
75
100
count
Gender
Female
Male