2 Descriptive Statistics I: Visualizing the Data

This is the first in a series of two chapters on descriptive statistics. In this part we cover organizing, presenting and visualizing the data with the help of an example.

2.1 What is Descriptive Statistics?

Descriptive statistics are concerned with "describing" the characteristics of a dataset. This is achieved by organizing, presenting and summarizing the data effectively.

2.2 Organizing and Presenting the Data

When we collect data for any study, we do so in a raw form. Before it can be used for any meaningful analysis, it is a good idea to organize it properly using tables. We usually set all the variables to be studied as columns and all the individual observations as rows of the table.

Let us consider an example. Sophie has just started her trading journey. She has decided to trade in stock XYZ using a simple rule. She looks at the recommendation for stock XYZ given by her favourite analyst each day. When the market opens, she buys one share of XYZ if the recommendation for that day is "buy" and sells one share if the recommendation for that day is "sell". In case the recommendation is "hold", Sophie carries over her existing position that day.

After fifteen days of trading, she decides to take stock of her trades. She realizes that in order to do any analysis, she must first organize the past fifteen days' data into a table. Her dataset now looks like **Table 2.1**.

Table 2.1: Data about Sophie's trading in stock XYZ for the past 15 days

Date	Recommendation	Shares Owned	Closing Price
03-Jan-20	Buy	1	20.50
04-Jan-20	Buy	2	19.80
05-Jan-20	Sell	1	17.90
06-Jan-20	Hold	1	18.70
07-Jan-20	Buy	2	19.10
08-Jan-20	Buy	3	18.80
09-Jan-20	Buy	4	20.00
10-Jan-20	Buy	5	20.50
11-Jan-20	Buy	6	21.00

12-Jan-20	Buy	7	19.10
13-Jan-20	Sell	6	17.90
14-Jan-20	Buy	7	18.10
15-Jan-20	Buy	8	19.50
16-Jan-20	Buy	9	20.50
17-Jan-20	Hold	9	19.70

The above is an example of ‘time series’ data which we often encounter in finance. Put simply, such data studies a list of variables over different points in time.

In Sophie’s case, the three variables being observed on each of the fifteen days are:

- “Recommendation” (analyst’s daily recommendation for stock XYZ),
- “Shares Owned” (the number of shares of XYZ owned by Sophie at the Close) and
- “Closing Price” (the daily closing price for stock XYZ).

Data Types: Quantitative and Qualitative

The data collected on any variable can be classified as qualitative or quantitative. Let’s look at each one more closely.

Qualitative data usually consists of non-numeric values. Qualitative variables are also referred to as categorical variables as each observation can fall in one of the predetermined categories.

For example, in **Table 2.1**, the variable “Recommendation” is qualitative or categorical, as for each observation it can only have one of the three values (“buy”, “sell” or “hold”).

Quantitative data consists of numerical values and can be further divided into following:

1. **Discrete data**: It can take up only integer values and is often used to represent countable items. For instance, the number of tails when we flip a coin fifty times, the number of children in a family etc.
   In **Table 2.1**, the variable ‘Shares Owned’ is a discrete variable, as the number of shares is never fractional.
2. **Continuous data**: It represents variables that cannot be counted but can be measured. Values of continuous data are within the range, but it can take up any numerical value. For example, weight, height etc.
   In **Table 2.1**, the variable ‘Closing Price’ is a continuous variable, as it can be measured in fractional values.

Frequency-Distribution of Data

Frequency distribution of a dataset shows all the possible values that the data can take along with the frequency of occurrence of these values.

While explaining the distribution of categorical data, we look at the percentage or number of observations in each group/category. For example **Table 2.2** represents the distribution of the categorical variable “Recommendation” from **Table 2.1**.

Table 2.2: Distribution of categorical variable “Recommendation”

Recommendation
buy
buy
sell
hold
buy
buy
buy
buy
buy
buy
sell
buy
buy
buy
hold

Category	frequency	Percentage
buy	11	73%
sell	2	13%
hold	2	13%
Total	**15**	**100%**

From the above table, we can see that 73% of the data lies in the “buy” category.

For continuous variables such as ‘Closing Price’ which can take fractional values , we summarize its distribution using ranges or ‘bins’ as follows:

Table 2.3 Distribution of continuous variable “Closing Price”

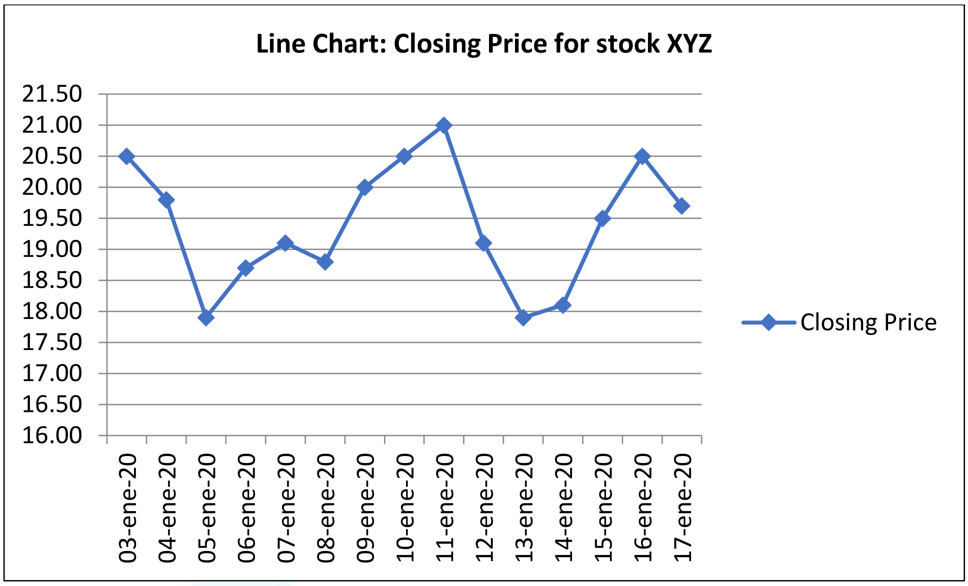
Closing Price
20.50
19.80
17.90
18.70
19.10
18.80
20.00
20.50
21.00
19.10
17.90
18.10
19.50
20.50
19.70

bins	Frequency
0-17	0
17-18	2
18-19	3
19-20	6
20-21	4
More than 21	0

2.3 VISUALIZING THE DATA

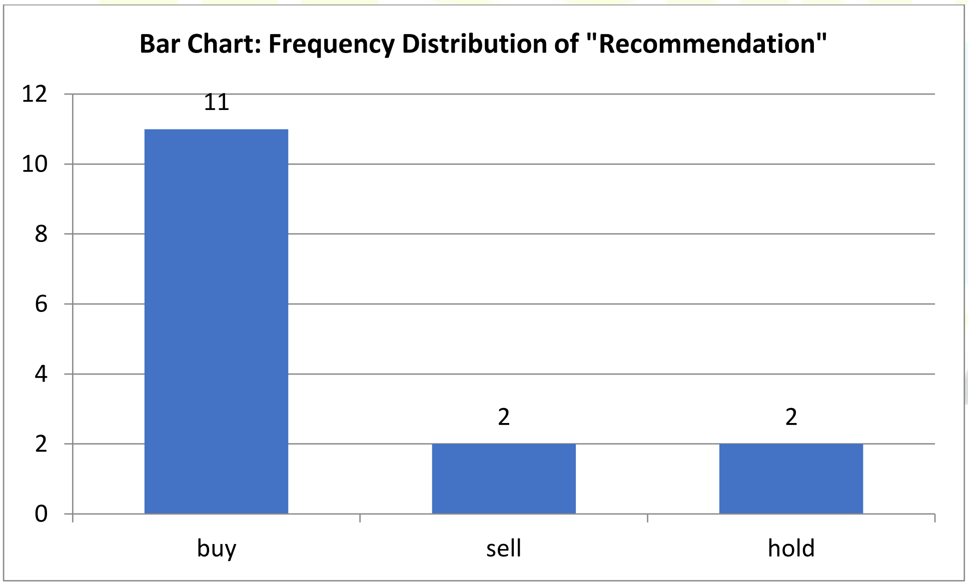
Line Chart

A line chart is used for showing trends over time. It is a simple and intuitive way to get a feel for the underlying data fluctuations. The following line chart depicts the movement of ‘Close Price’ over fifteen days.



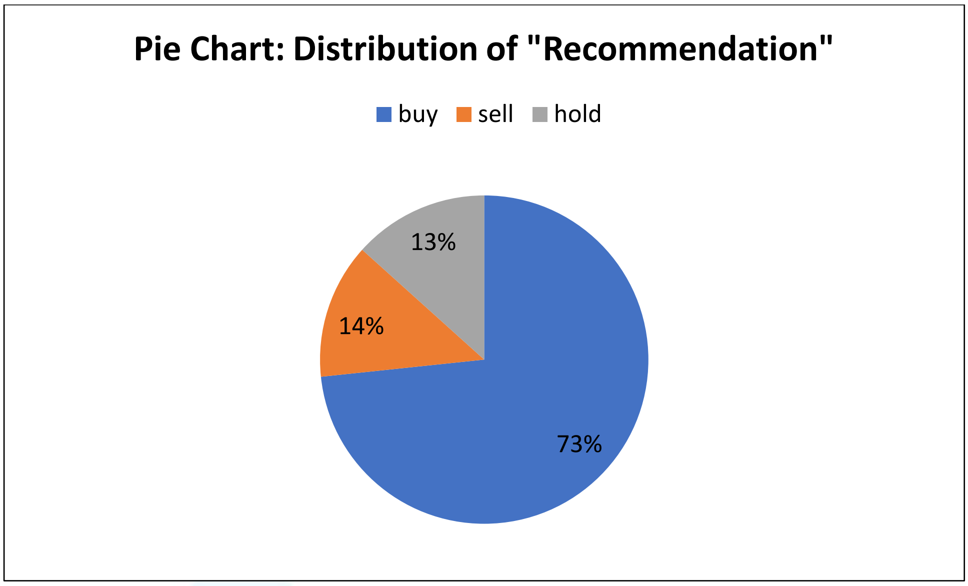
Bar Chart

It is a graphical representation of data using bars of different heights. The bar chart can be horizontal or vertical. In the following bar chart we show the frequencies for the three categories in the "Recommendation" variable:



Pie Chart

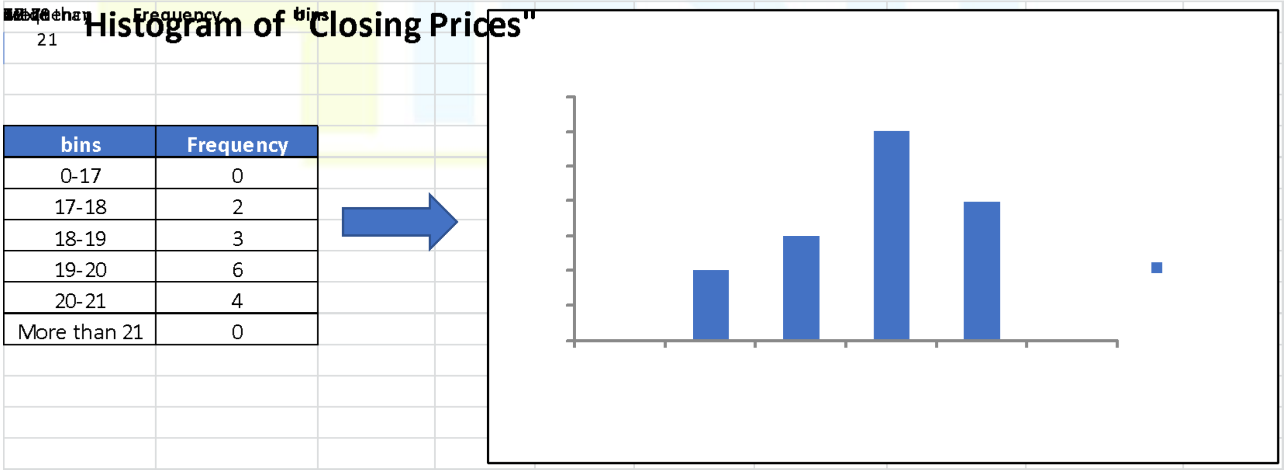
Pie chart is mainly used to represent categorical data. It is used to show percentage and proportions of different categories. It is ideal for understanding the proportional distribution of data. Pie chart is shown in the form of a circle, where the full circle area represents the total data i.e. 100% and each sector represents its respective share of the total.



Histogram

Histogram represents the frequency distribution of a continuous dataset. In a way, they roughly depict the probability distribution of dataset. To construct a histogram, the dataset is first split into consecutive intervals or ranges called ‘bins’. Each bar in the histogram depicts the frequency at associated interval/bin.

In **Table 2.3**, we already have the frequency distribution of the variable “Closing Price”, which can be visually represented as a histogram chart as follows:



bins	Frequency
0-17	0
17-18	2
18-19	3
19-20	6
20-21	4
More than 21	0