

Business Analytics

Chapter 2 – Descriptive Statistics



Overview of Using Data: Definitions and Goals

* **Data**:

The facts and figures collected, analyzed, and summarized for presentation and interpretation.



Overview of Using Data: Definitions and Goals

* **Variable**:

A characteristic or a quantity of interest that can take on different values.

* **Observation**:

A set of values corresponding to a set of variables.



Overview of Using Data: Definitions and Goals

* **Variation**:

The difference in a variable measured over observations.

72.41

-246.16

235.61

* **Random variable/uncertain variable**:

A quantity whose values are not known with certainty.



Types of Data – Population and Sample Data

**Population**:

Are students in a classroom a population or sample?

All elements of interest.

**Sample**:

Subset of the population.

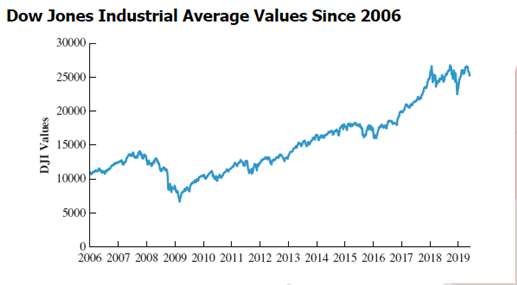


Types of Data

* **Random sampling**:

A sampling method to gather a representative sample of the population data.

* Is it truly a random sample?
* Does it really represent the population?
  + Where is your data coming from?
  + Who was polled?
  + How many people were polled?



Types of Data (Slide 1 of 5)

C

Q

Q

* **Quantitative data**:

Data on which numeric and arithmetic operations, such as addition, subtraction, multiplication, and division, can be performed.

* **Categorical data**:

Data on which arithmetic operations cannot be performed.



Types of Data

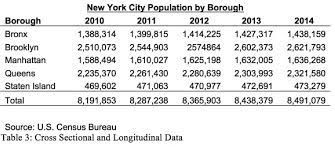
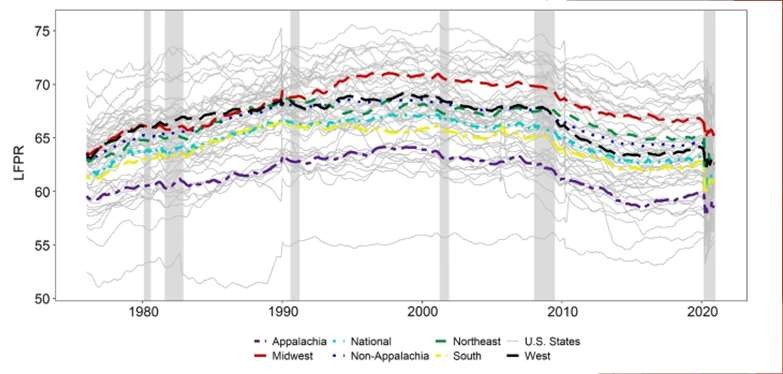
* **Cross-sectional data**:

Data collected from several entities at the same, or approximately the same point in time.

* **Time series data**:

Data collected over several time periods.

Types of Data – Data Sources



Other Examples

T.S.

C.S.

C.S. + T.S.

T.S.

1,450,000

1,440,000

1,430,000

1,420,000

1,410,000

1,400,000

1,390,000

1,380,000

1,370,000

1,360,000

Bronx

2010 2011

2012

Year

2013 2014

Population

# Experimental study:

* + A variable of interest is first identified.
  + Then one or more other variables are identified and controlled or manipulated
  + so that data can be obtained about how they influence the variable of interest.

# Example: Drug Study

* + Variables: Blood Pressure, Dosage
  + Random Sample
  + Controls: Placebo, dosage
  + Data collected before and after
  + Statistical Analysis



Types of Data – Data Sources

Customer Opinion Questionnaire Used by Chops City Grill Restaurant

Nonexperimental study or observational study:

* Makes no attempt to control the variables of interest.
* A survey is perhaps the most common type of observational study.



Modifying Data in Excel

Sorting and Filtering Data in Excel Conditional Formatting of Data in Excel



Modifying Data in Excel

Table 2.2: 20 Top-Selling Automobiles in United States in February 2019



Modifying Data in Excel

Table 2.2: 20 Top-Selling Automobiles in United States in February 2019 (cont.)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Rank (by February 2019 Sales)** | **Manufacturer** | **Model** | **Sales (February 2019)** | **Sales (February 2018)** |
| 1 | Toyota | Corolla | 29,016 | 25,021 |
| 2 | Toyota | Camry | 24,267 | 30,865 |
| 3 | Honda | Civic | 22,979 | 25,816 |
| 4 | Honda | Accord | 20,254 | 19,753 |
| 5 | Nissan | Sentra | 17,072 | 17,148 |
| 6 | Nissan | Altima | 16,216 | 19,703 |
| 7 | Ford | Fusion | 13,163 | 16,721 |
| 8 | Chevrolet | Malibu | 10,799 | 11,890 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Rank (by February 2019 Sales)** | **Manufacturer** | **Model** | **Sales (February 2019)** | **Sales (February 2018)** |
| 9 | Hyundai | Elantra | 10,304 | 15,724 |
| 10 | Kia | Soul | 8,592 | 6,631 |
| 11 | Chevrolet | Cruze | 7,361 | 12,875 |
| 12 | Nissan | Versa | 7,410 | 7,196 |
| 13 | Volkswagen | Jetta | 7,109 | 4,592 |
| 14 | Kia | Optima | 7,212 | 6,402 |
| 15 | Kia | Forte | 6,953 | 7,662 |
| 16 | Hyundai | Sonata | 6,481 | 6,700 |



Modifying Data in Excel

Table 2.2: 20 Top-Selling Automobiles in United States in February 2019 (cont.)

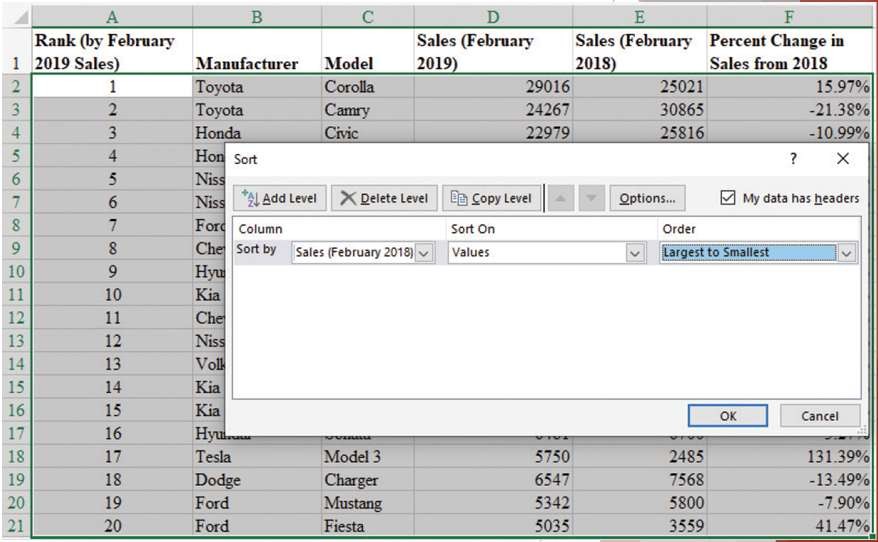


|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Rank (by February 2019 Sales)** | **Manufacturer** | **Model** | **Sales (February 2019)** | **Sales (February 2018)** |
| 17 | Tesla | Model 3 | 5,750 | 2,485 |
| 18 | Dodge | Charger | 6,547 | 7,568 |
| 19 | Ford | Mustang | 5,342 | 5,800 |
| 20 | Ford | Fiesta | 5,035 | 3,559 |



Modifying Data in Excel

Figure 2.3: Data for 20 Top-Selling Automobiles Entered into Excel with Percent Change in Sales from 2018



Modifying Data in Excel (Slide 5 of 14)

Sorting and Filtering Data in Excel:

* To sort the automobiles by February 2018 sales:
  + Step 1: Select cells A1:F21.
  + Step 2: Click the **Data** tab in the Ribbon.
  + Step 3: Click **Sort** in the **Sort & Filter** group.
  + Step 4: Select the check box for **My data has headers.**
  + Step 5: In the first **Sort by** dropdown menu, select **Sales (February 2018).**
  + Step 6: In the **Order** dropdown menu, select **Largest to Smallest.**
  + Step 7: Click **OK.**



Modifying Data in Excel

Figure 2.4: Using Excel’s Sort Function to Sort the Top-Selling Automobiles Data



Modifying Data in Excel

Figure 2.5: Top-Selling Automobiles Data Sorted by Sales in February 2018 Sales



Modifying Data in Excel

Sorting and Filtering Data in Excel (cont.):

* Using Excel’s Filter function to see the sales of models made by Nissan:
  + Step 1: Select cells A1:F21.
  + Step 2: Click the **Data** tab in the Ribbon.
  + Step 3: Click **Filter** in the **Sort & Filter** group.
  + Step 4: Click on the **Filter Arrow** in column B, next to **Manufacturer.**
  + Step 5: If all choices are checked, you can easily deselect all choices by unchecking (**Select All**). Then select only the check box for **Nissan**.
  + Step 6. Click **OK.**



Modifying Data in Excel (Slide 9 of 14)

Figure 2.6: Top-Selling Automobiles Data Filtered to Show Only Automobiles Manufactured by Nissan



Modifying Data in Excel (Slide 10 of 14)

Conditional Formatting of Data in Excel:

* Makes it easy to identify data that satisfy certain conditions in a data set.
* To identify the automobile models in Table 2.2 for which sales had decreased from February 2018 to February 2019:
  + Step 1: Starting with the original data shown in Figure 2.3, select cells F1:F21.
  + Step 2: Click on the **Home** tab in the Ribbon.
  + Step 3: Click **Conditional Formatting** in the **Styles** group.
  + Step 4: Select **Highlight Cells Rules**, and click **Less Than…** from the dropdown menu.
  + Step 5: Enter *0%* in the **Format cells that are LESS THAN**: box.
  + Step 6: Click **OK.**



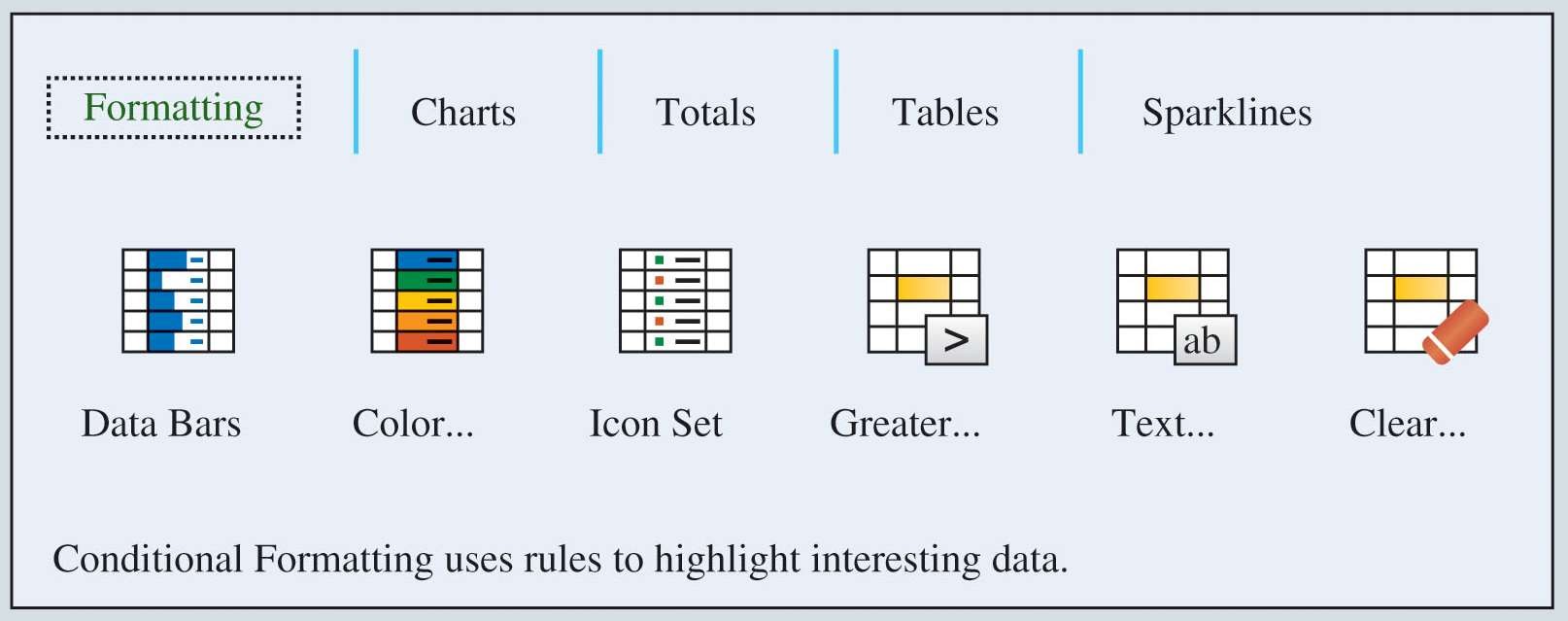
Modifying Data in Excel (Slide 11 of 14)

Figure 2.7: Using Conditional Formatting in Excel to Highlight Automobiles with Declining Sales from February 2018



Modifying Data in Excel (Slide 12 of 14)

Figure 2.8: Using Conditional Formatting in Excel to Generate Data Bars for the Top-Selling Automobiles Data



Modifying Data in Excel (Slide 13 of 14)

Conditional Formatting of Data in Excel (cont.):

* **Quick Analysis** button appears just outside the bottom-right corner of a group of selected cells.
* It provides shortcuts for Conditional Formatting, adding Data Bars, and other operations.



Modifying Data in Excel (Slide 14 of 14)

Figure 2.9 Excel Quick Analysis Button Formatting Options



Creating Distributions from Data

Frequency Distributions for Categorical Data

Relative Frequency and Percent Frequency Distributions Frequency Distributions for Quantitative Data Histograms

Cumulative Distributions



Creating Distributions from Data

Frequency Distributions for Categorical Data:

* **Frequency distribution**: A summary of data that shows the number (frequency) of observations in each of several nonoverlapping classes, typically referred to as **bins.**



Creating Distributions from Data

Table 2.4: Frequency Distribution of Soft Drink Purchases

* The frequency distribution summarizes information about the popularity of the five soft drinks:
  + Coca-Cola is the leader.
  + Pepsi is second.
  + Diet Coke is third.
  + Sprite and Dr. Pepper are tied for fourth.



Creating Distributions from Data

Table 2.3: Data from a Sample of 50 Soft Drink Purchases

Coca-Cola Diet Coke Pepsi

Diet Coke Coca-Cola Coca-Cola Dr. Pepper Diet Coke Pepsi Pepsi Coca-Cola Dr. Pepper Sprite Coca-Cola Diet Coke Coca-Cola Coca-Cola

Sprite Coca-Cola Diet Coke Coca-Cola Diet Coke Coca-Cola Sprite Pepsi Coca-Cola Coca-Cola Coca-Cola Pepsi Coca-Cola Sprite

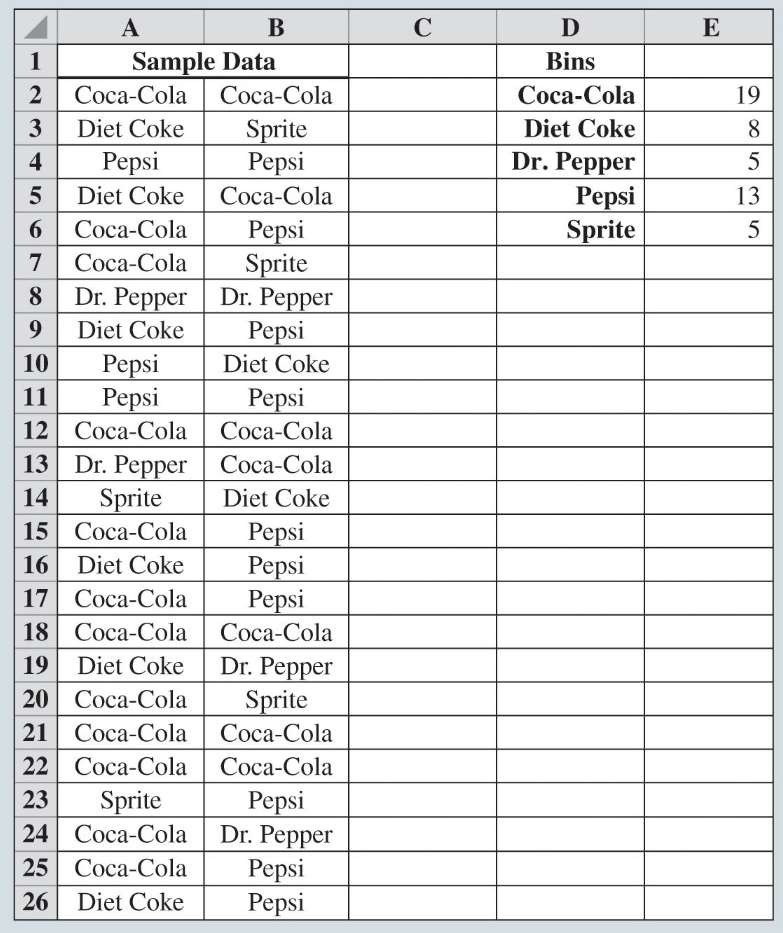
Dr. Pepper Pepsi

Diet Coke

Pepsi Coca-Cola Coca-Cola Coca-Cola Pepsi

Dr. Pepper Coca-Cola Diet Coke Pepsi Pepsi Pepsi Pepsi Coca-Cola Dr. Pepper Pepsi Sprite

|  |  |
| --- | --- |
| **Soft Drink** | **Frequency** |
| Coca-Cola | 19 |
| Diet Coke | 8 |
| Dr. Pepper | 5 |
| Pepsi | 13 |
| Sprite | 5 |
| Total | 50 |



Creating Distributions from Data

Figure 2.10: Creating a Frequency Distribution for Soft Drinks Data in Excel



Creating Distributions from Data

Relative Frequency and Percent Frequency Distributions:

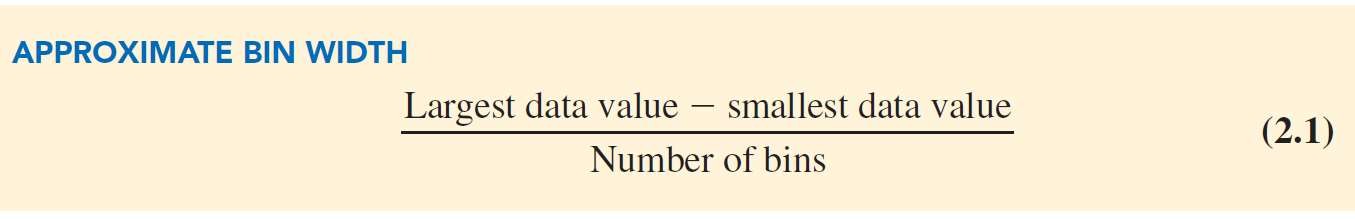
* **Relative frequency distribution**: A tabular summary of data showing the relative frequency for each bin.
* **Percent frequency distribution**: Summarizes the percent frequency of the data for each bin.

Used to provide estimates of the relative likelihoods of different values of a random variable.



Creating Distributions from Data

Table 2.5: Relative Frequency and Percent Frequency Distributions of Soft Drink Purchases



|  |  |  |
| --- | --- | --- |
| **Soft Drink** | **Relative**  **Frequency** | **Percent**  **Frequency (%)** |
| Coca-Cola | 0.38 | 38 |
| Diet Coke | 0.16 | 16 |
| Dr. Pepper | 0.10 | 10 |
| Pepsi | 0.26 | 26 |
| Sprite | 0.10 | 10 |
| Total | 1.00 | 100 |



Creating Distributions from Data

Frequency Distributions for Quantitative Data:

* Three steps necessary to define the classes for a frequency distribution with quantitative data:
  1. Determine the number of nonoverlapping bins.
  2. Determine the width of each bin.
  3. Determine the bin limits.



Creating Distributions from Data

Table 2.6: Year-End Audit Times (Days)

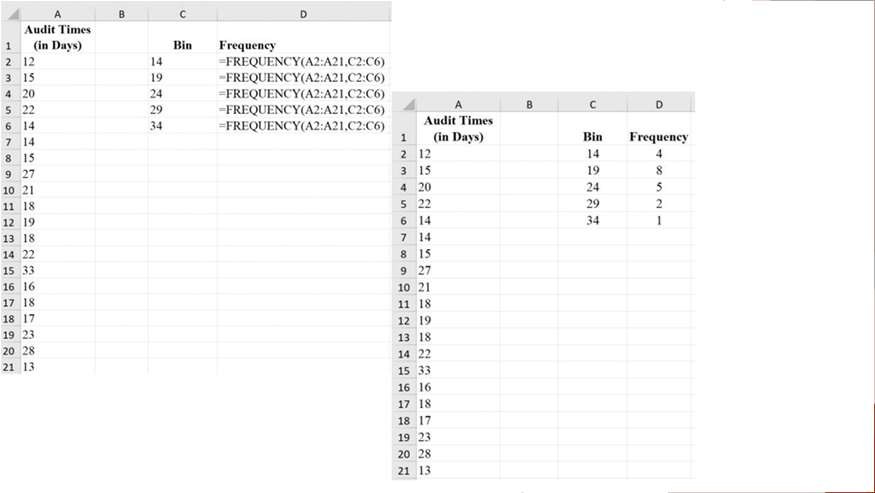


Creating Distributions from Data

Table 2.7: Frequency, Relative Frequency, and Percent Frequency Distributions for the Audit Time Data

|  |  |  |  |
| --- | --- | --- | --- |
| 12 | 14 | 19 | 18 |
| 15 | 15 | 18 | 17 |
| 20 | 27 | 22 | 23 |
| 22 | 21 | 33 | 28 |
| 14 | 18 | 16 | 13 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Audit Times**  **(days)** | **Frequency** | **Relative**  **Frequency** | **Percent**  **Frequency** |
| 10–14 | 4 | 0.20 | 20 |
| 15–19 | 8 | 0.40 | 40 |
| 20–24 | 5 | 0.25 | 25 |
| 25–29 | 2 | 0.10 | 10 |
| 30–34 | 1 | 0.05 | 5 |



Creating Distributions from Data

Figure 2.11: Using Excel to Generate a Frequency Distribution for Audit Times Data