THE

ULTIMATE GUIDE

— TO —

basic DATA CLEANING

A free resource by

socialcops

About SocialCops

SocialCops is a data intelligence company on a mission to confront the world's most critical problems with data. We work with over 150 companies in 7 countries. One of India's fastest growing startups, we were featured on the Forbes Asia 30 Under 30 and Fortune India 40 Under 40 lists two years in a row, and were recognized as one of India's top 10 startups by NASSCOM.

Our platform brings the entire decision-making process — collecting primary data, accessing external data, linking internal data, cleaning and transforming data, and visualizing data — to one place. This makes it faster, more efficient, and easier to make an important decision through data.

Read more about our work here.

Table of Contents

INTRODUCTION

Getting Started with Data Cleaning

You have spent time and energy designing a questionnaire that focuses on important topics. Your surveyors have worked with respondents to ensure quality data collection. Now, it is up to you to turn all this hard work and preparation into meaningful insights. What do you do first?

Before you begin analyzing your data, it is crucial to ensure that your data set is complete and correct. The data cleaning process helps you achieve this by detecting and removing errors and inconsistencies in your data.

But what if you can't wait to get your hands on the data and start analyzing it? Can you skip the data cleaning process?

The answer is NO. In fact, the more you care about data analysis, the more you should care about data cleaning. After all, you don't want to be making important decisions based on wrong insights simply because your data had errors!

Since data collection is prone to human error — during the collection and data entry stages (in the case of paper-based surveys) — it becomes really important to double-check the sanctity of your data set before you begin analysis.

TOOL TIP

Data cleaning (also called data cleansing) is the process of finding and dealing with problematic data points within a data set. Data cleaning can involve fixing or removing incomplete data, cross checking data against a validated data set, standardizing inconsistent data, and more. Data scientists usually spend 50% to 80% of their time cleaning data before they can start analyzing it.

In this ebook, we describe simple, yet crucial, techniques to help you clean your data effectively. In case you wish to apply this learning in the future, we have incorporated a number of examples and exercises to give you hands-on learning. All you need is **Microsoft Excel** (versions 2007 or above). Don't worry if you are unfamiliar with Excel — each chapter will teach you everything you need to know.

The examples and exercises will use a sample data set for sales transactions of a dummy e-commerce portal, XYZshopping.com. **You can download the** sample data set **here**. The Excel spreadsheet has two sheets — Sales Data and Demographic Data. We recommend that you use this data set for getting hands-on learning on the topics as you come across 'how-to' exercises during your reading. **The solutions to these exercises have been included at the end of the book**.

There are eight chapters in this ebook, each covering a distinct aspect of data cleaning — defining data structure using rows, columns, and unique identifiers; basic sanity checks; defining question types; identifying and dealing with missing values and outliers; conditional questions; and a case study highlighting the importance of data cleaning from one of SocialCops' actual deployments.

Excited? Great, let's get started!

CHAPTER 1

Rows and Columns: The Building Blocks of Data Cleaning

Understanding how to structure your data into rows and columns is a crucial first step in cleaning any data set, especially in survey data cleaning. All manipulation and analysis that you perform on the data depends on the framework of rows and columns. Learn about rows and columns and how they set the stage for data cleaning and analysis.

Chapter 1 will help you learn the following techniques:

- Organize survey questions into **rows and columns**
- **Build a unique ID** into your survey
- How to **identify a unique ID** in your data

Understanding the structure of your data is a crucial first step in cleaning any data set. The way a data set is laid out is just as important as the actual values, because the structure determines how the data can be manipulated and interpreted.

The main components of a data set that we are concerned with are columns, rows, and unique IDs. Together, these three features of a data set make up the structure that later helps in data cleaning and analysis.

Columns

Columns are vertical and store information related to one particular variable. A variable is the name for a data point whose value changes in different situations.

In Excel, columns are labelled using letters of the English alphabet. Once the labels pass the letter Z, the number of letters increases to two and then three, with letters combined to create a unique column-labeling pattern such as AA, AB, AC, and so on until you reach XFD at the maximum of 16,384 columns.

To select an entire column, click on the letter at the top of the column and a grey shading will appear over the entire column.

In the case of primary data (data that you collected), variables correspond to questions in a survey. For example, the variable "Number of Males in a Household" contains values related to the number of male members in a particular household. You could use an alias titled "Household_Male" that might contain all responses to the survey question, "How many members of the household are male?" The value recorded in the corresponding column changes based on the specific characteristics of each household.

TOOL TIP

In Collect, SocialCops' mobile data collection tool, you can add an alias for every question while building your survey!

Because columns store information for one variable, data in each column should always be similar. In Chapter 3, we will learn functions in Excel that help us clean data over an entire column or variable.

EXERCISE 1

Imagine that you provide customers with a short survey after they purchase an item on XYZshopping.com. Try putting the survey questions into columns of an Excel spreadsheet.

1. What is your name?
2. Which city do you live in?
3. Are you male or female?
4. How old are you?
5. How did you come to know about our online shopping site?
    a. Recommendation from someone
    b. Advertisement
    c. Online search

TOOL TIP

While you could include the full question in each column, questions are often long and can be difficult to view in an Excel column. Therefore, you can assign aliases — abbreviated versions of column headings. For example, the question, "How many 4-wheeler vehicles does the household own?" may be assigned the alias "4_wheeler_number".

It is a good practice to keep a dictionary — an Excel sheet with the two columns containing information on "Survey Question" and "Alias" — which clearly links each alias to the original survey question or variable. This will help people who are viewing a data set of survey responses for the first time.

Rows

Rows are the horizontal entities that give you information about one particular unit. A unit is the object of evaluation and changes based on the type of analysis performed. In a household survey, for example, the unit of analysis would be a household. In a survey about industries, however, the unit of analysis would be an industry.

Each row contains data on several different variables, all of which relate to the unit of analysis specific to that row. While values within a row can be of different data types, they are always related to the unit of analysis and should therefore make sense in relationship to one another.

Rows in Excel are labelled with numbers from 1 to the maximum number of 1,048,576 rows. Similar to columns, clicking on a row number highlights the entire row and all the data within it. This allows you to make changes to an entire row without affecting any of the other rows. Rows can also be selected by clicking on a specific cell once to highlight it or twice to edit the value in the cell.

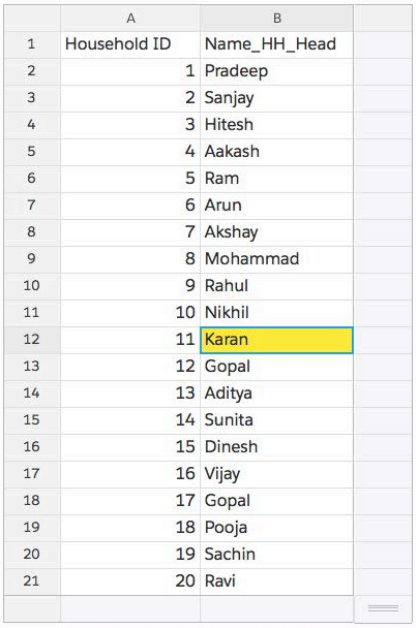
EXERCISE 2

Consider the sales data for XYZshopping.com.

1. Identify the unit of analysis in this data set.
2. How many of these units of analysis are present in this data set?

Cells

Every intersection of a row and column is called a cell. A cell contains a single data point that defines the value of the variable of the column for the unit of analysis in the row.



	A	B
1	Household ID	Name_HH_Head
2	1	Pradeep
3	2	Sanjay
4	3	Hitesh
5	4	Aakash
6	5	Ram
7	6	Arun
8	7	Akshay
9	8	Mohammad
10	9	Rahul
11	10	Nikhil
12	11	Karan
13	12	Gopal
14	13	Aditya
15	14	Sunita
16	15	Dinesh
17	16	Vijay
18	17	Gopal
19	18	Pooja
20	19	Sachin
21	20	Ravi

For example, the spreadsheet to the left has rows with households numbered 1 to 20 and a column titled "Name_HH_Head". Each cell gives us the name of the household head for the particular household ID of the reference row.

In Excel, each cell is referenced by the column label (letter) combined with the row label (number). In column B and row 12 of this example, the name of the head of household for household 11 is Karan. We would reference this cell as "B12".

Unique Identifier

The final element in survey data structure is the unique identifier (unique ID or **UID**). A unique ID allows us to distinguish units from each other by assigning a custom ID to each unit. In the case of surveys, we need UIDs to keep track of which responses belong to which units. Without the ability to distinguish between units, we run the risk of including duplicates in our analysis.

UIDs are frequently used in our everyday lives. For example, in the United States, almost every citizen and permanent resident has a 9-digit Social Security Number. This number gives the United States a unique ID for every person, which is used for taxation, income

tracking, banking, and more. Other examples of UIDs are driver's license numbers, passport numbers, Aadhar Card numbers in India, Resident Identity Card numbers in China, and National Health Service numbers in the United Kingdom.

A unique ID not only helps distinguish units from each other within a single worksheet, but can also be used to link separate worksheets that contain information about the same units. For example, imagine a questionnaire divided into three sections — Household Member Details, Assets and Amenities, and Health. If data for each section of this survey is entered into separate Excel workbooks, the UID will enable linking all three data sets together so that there are full details for each household in one master file. (Excel lets you do this through a function called VLOOKUP. Chapter 7 talks about how to join values based on a unique ID using the VLOOKUP() function.)

Building a Unique ID into Your Survey

Survey tools like SocialCops' Collect have made it simple to automatically assign unique IDs to units while conducting a survey. Collect automatically creates a unique ID for each survey response. These responses can be used to collect and monitor data about a unit. For example, if you want to survey a set of households, you can collect data on each household (such as its location and members) and Collect will generate a UID for each household. Then, with Collect's monitoring feature, you can search the available households, select the correct one, and add data about that household, all without having to remember its UID.

Building a unique ID into a paper-based survey, however, can be more challenging. Unique IDs should ideally be intuitive to both the survey administrator and to survey enumerators.

For example, a unique ID may combine digits representing various levels of geographic information to create a unique ID that represents the geographic location of a unit. At the top of each survey page (or booklet in the case of longer surveys) there should be boxes which enumerators can fill out with the unique ID.

The following is an example of a unique ID format that may be used in a household survey:

In this example, the unique ID is made up of 14 numbers which represent the geographic location of the household at the state, district, and village level. This information is easily available to the enumerator and also ensures that each household has a unique ID. Moreover, this format also lets you cluster responses at each level of geography.

What if you are cleaning survey data that does not have a unique ID? While this situation is never ideal, you can form a unique ID once the data is collected, based on a combination of variables that can uniquely identify every unit within the survey sample. This method involves concatenating (or combining) two or more fields until every ID in the data set is unique. In other words, you can combine responses from several questions into one string of letters and numbers in a way that the combination is unique for each unit of analysis.

The post-enumeration method of assigning UIDs is imperfect because: 1) it assumes

that there can be a combination of responses that uniquely distinguishes every unit, and 2) because of data cleanliness issues, it may be difficult to find fields that have consistent, reliable responses to include in the concatenated unique ID. If missing or incorrect values in a field are used to create the unique ID, the validity of the ID could be jeopardized.

EXERCISE 3

Consider the sales data for XYZshopping.com. What unique ID can you create for this data set?

CHAPTER 2

Run Quick Sanity Checks

Errors can creep into your data in multiple ways. Here's a handy checklist of basic data checks to help you rule out some obvious errors in your data. Performing these easy data checks is a good starting point in your data cleaning process.

Chapter 2 will help you learn the following techniques:

- Data checks on the **number of respondents** in your data
- Data checks on the **number of questions** in your data
- Data checks on **geocoded variables**
- Data checks on **timestamps**

Conducting a survey is a fantastic way to receive feedback. Before we get down to analyzing the data collected from the survey, however, it is important to ensure that our data is error-free and clean. This requires a good understanding of the behavior of the various variables in the data.

This chapter talks about some basic sanity checks that you should conduct on your data before it is ready for analysis.

Paper vs. Digital Surveys

Before we begin with the checklist, it is important to understand the possible sources of errors in data. Data collection has traditionally been done using paper and pen. However, technology has made digital data collection a lot easier and more scalable.

Paper surveys do not allow organizations to make incremental changes. Since the data collection process in paper surveys occurs in batches, analysis of survey data can take place only after the completion of the entire surveying process. In addition, before any data analysis can be done, data has to be typically entered into spreadsheets, which is cumbersome. Because of the time lag between data collection and data analysis, there is the risk that insights will become dated. The less recent the analysis, the less immediate the feedback becomes — reducing the scope for evolving or adapting the survey. On the whole, digital surveys are easier to create, distribute, complete, and analyze.

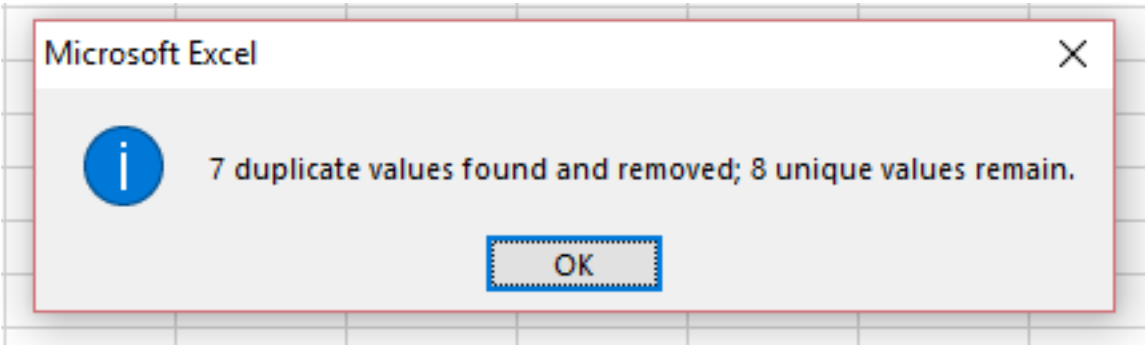
For data collected through both paper and digital surveys, you should conduct some basic sanity checks before thorough data cleaning. These are some general rules that you can tailor to the specific nature of your survey.

Data Checks

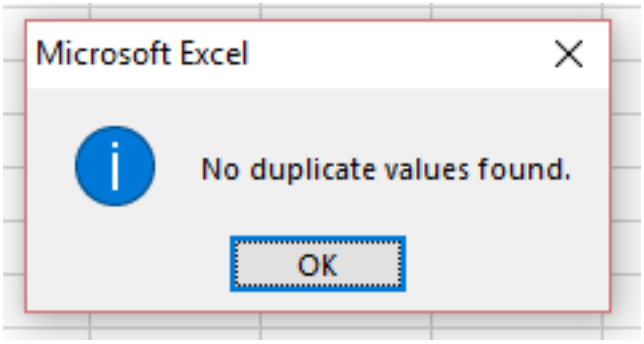
Number of Respondents vs. Rows

For any kind of survey, you should always match the number of rows in your data to the number of respondents surveyed to ensure data completeness. For example, for a survey involving 500 households, the first step will be to ensure the number of rows in your data set equals 500. This is a good first check. However, it doesn't discount instances of simultaneous exclusion and duplication of a certain number of households.

You can use a simple hack to check if all your UIDs are unique. Copy the values in the column to a different sheet in Excel. Select the entire column, and then click on the "Remove Duplicates" option under the "Data" tab of the Excel ribbon. A pop-up like this will appear:



You know you're in the clear if you receive a pop-up like this:



Number of Questions vs. Columns

Make sure to be well-versed with the structure of your survey — have a thorough understanding of all question and response types. It will be of immense help when you validate the responses received for these questions. You can then quickly match the total number of columns in your spreadsheet with the total number of questions. Similar to the rows check done earlier, this will help you identify missed or duplicate values.

Geocoded Variables

Often, surveys require recording the location of the respondent. The common practice is to add the location in the form of a latitude-longitude (lat-long) combination. Make sure that all your lat-long combinations are within the desired range of geo-coordinates.

Sometimes, especially in a digital survey, there could be outliers in the location due to issues with the collection device's accuracy. For example, while conducting a survey in a certain country, you should be aware of the north, south, east, and west boundaries of the country and should also run a basic check on whether the collected geo-coordinates are within this predefined range.

EXERCISE 1

The rectangular area whose vertices are defined by the geo-coordinates
(1.644745, 49.582651)
(1.644745, 91.492333)
(-9.773677, 91.492333)
(-9.773677, 49.582651)
corresponds to an area in the middle of the Indian Ocean.

Consider the sales data for XYZshopping.com. Columns J and K contain the lat-long of the location where each transaction was made. Since we know that no one made a transaction from the middle of the Indian Ocean, find the transactions that have incorrect lat-longs from this area.

Hint: Use this formula in an empty column to find transactions with lat-longs in the Indian Ocean:

```
=IF(AND(lat>-9.773677,lat<1.644745,long>49.582651,long<91.492333),
"In ocean","Not in ocean")
```

In this formula:

- IF: formula to evaluate the flag to be assigned based on whether a condition is true or false
- AND: function to check if all four conditions in the parenthesis are true
- lat: reference to the cell containing the latitude for the corresponding row
- long: reference to the cell containing the longitude for the corresponding row
- "In ocean": what will appear in Excel if the AND function is true (meaning that the lat-long is in the Indian Ocean)
- "Not in ocean": what will appear in Excel if the AND function is not true

Time Stamps

Time stamps can be in different formats. Some common formats are listed below:

1. YYYY-MM-DD HH:MM:SS (Year-Month-Day Hour:Minutes:Seconds)
2. MM-DD-YY
3. DD-MM-YY
4. MM-YYYY

No matter what format you choose, it is important to keep it consistent throughout the data set. Also ensure that the second, minute, hour, day, month, and year are valid. For instance, in the DD-MM-YY format, the date 35-13-16 would be incorrect because the day and month cannot go beyond 31 and 12 respectively.

In case you need to convert from one format to another, you should be very careful, since this is a highly error-prone step. Make sure to check the date ranges before and after conversion to check if the minimum and maximum dates match.

EXERCISE 2

Consider the sales data for XYZshopping.com. Column B contains the transaction time for each transaction. What format are these timestamps in?

CHAPTER 3

Check Different Question Types

Question types help you set expectations for the range and format of responses that you receive in your survey. Assigning specific question types to your questions will help you identify and deal with values outside the anticipated range of responses. Learn about common question types and how to ensure data consistency for your survey responses.

Chapter 3 will help you learn the following techniques:

- Identify **correct question types** for each question
- Check data for **question type consistency**
- Using **data validation tools** to improve the data entry process

If you asked 50 people their age, what types of responses would you expect to receive? You might expect the response to be a number. You could also expect it to be greater than 0 and less than, say, 120. However, if someone answered with a name or phrase, you would likely record this as a bad response.

Setting such expectations for response types — called data types — and quality of responses is extremely helpful in data cleaning. They help us identify values that fall outside the type and range of responses anticipated. It is important to identify values that do not match the predetermined data type and range, because these may have to be removed to preserve data quality.

Checking Response Types During the Cleaning Process

Step 1: Identify the Data Type for Each Question

The first step is to go through each survey question in the questionnaire and mark it with the data type of the response. In surveys, questions from a questionnaire correspond to columns within a data set, so we can infer the expected data type of each column from the data type of the relevant question. The questionnaire now becomes the guide to evaluate survey responses within each column of the Excel spreadsheet.

When designing a survey, it is important to assign a response data type to each question. Response types can be broadly categorized as categorical, integers, dates, or strings (non-numeric text responses). While there are more specific and complex data types, these broad question types form the basis of a typical survey.

Categorical

Categorical question types — otherwise known as multiple-choice questions — elicit an answer from a predefined set of categories. In the case of a categorical question,

responses should only fall within the categories provided.

For example, an enumerator may be asked to assess whether a property is “Residential”, “Commercial”, or “Empty”.

Many surveys also provide an “Other” option with a comments section for responses that do not fall within the category options.

You can code responses using letters to reduce data entry efforts. For example, the options in the question above could be coded as “R” (Residential), “C” (Commercial), and “E” (Empty). Then the responses expected should lie within the range of letter codes provided — in this example, there shouldn’t be letters other than R, C, and E.

A special type of categorical questions is “Boolean”. Boolean question types have only two possible responses: Yes and No. Similar to multiple-choice questions, Yes and No can be assigned values — 1 and 0, or Y and N respectively — to make data entry, cleaning, and analysis easier.

Numeric

Numeric data covers numbers that can be either positive or negative (except for the integer zero, which is neither positive nor negative). Numeric data can be an integer (whole number), such as in questions like “How many cows does the household own?” It can also be decimal numbers, such as in questions like “How many acres of land does the household own?”

Dates

Dates comprise of integers but are stored in their own special format to convey a particular point in time. Formats for dates change between countries and industries,

so the format of a date should be clearly standardized prior to beginning a survey. This ensures consistency across survey responses.

Strings

String responses comprise of a sequence of characters and are often a word or group of words. For example, responses to the question, "What is the name of the head of household?" will be sequences of letters that form a name. Strings are different from categorical responses because the number of responses are not limited by a predefined set of options.

EXERCISE 1

Consider the sales data for XYZshopping.com. Identify the data type for Columns A through K.

TOOL TIP

In Excel, right-click on a cell or range of cells, and select the "Format Cells" option to help you identify/ modify the data type of the selected values.

Step 2: Check for Consistency

Once you have identified the desired data types for each column, there are several ways to check for data type consistency within columns. When done manually, this is a time-intensive process prone to human error. Fortunately, there are preset tools available in Excel to automate the data type analysis process. You can also write simple functions in Excel to determine the spread of data types for each column and thus check for consistency within that column.

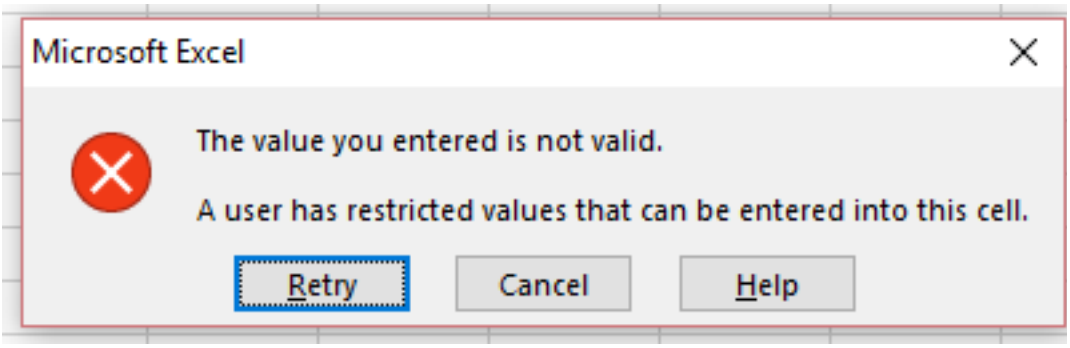
One useful tool in Excel is Data Validation. The Data Validation tool is located on the Data tab of the ribbon at the top of the Excel workspace. In this tool, you can customize settings to run automatic checks on data within a column.

For example, if you are cleaning a column called "Age of Household Head," you know that the responses should be integers. By setting the validation criteria to allow only whole numbers, you can run a check over the column to flag values that do not meet this criteria.

Similarly, you can limit string data types to a length of 100 characters and flag all responses above this limit for review. Additional criteria such as the range of acceptable values (for example, ages between 0 and 120 years) can also be applied using the Data Validation tool. This makes it a powerful method of flagging quality issues in your data set without you having to manually check each data point.

Using the Data Validation Tool to Improve Data Entry

Excel's Data Validation tool is not only useful in identifying incorrect data types and values; it can also be used during the data entry process itself to improve the quality of manually entered data. The Data Validation tool allows you to flag incorrect responses by setting data entry rules on each column individually. As someone enters survey data, any values outside the specified parameters can either be rejected altogether or flagged for review, as shown below.



Setting up data validation rules prior to data entry can significantly increase data quality and make it easier to clean data later on.

Using the Sort and Filter Tools

The Sort tool is a simple way to change the order of your data. It makes it easy to spot outliers or data with different data types. For example, if one column contains a text string where there should only be integers, sorting the column will put the string value at the beginning or end of the list where it is easy to spot.

To use the Sort tool, select a column and click “Sort” in the top menu.

The Filter tool is an easy way to view the range of values lying within a column. To use it, select a column by clicking the letter label or the column header and then select “Filter”. A small box with a down arrow will now appear in the topmost cell. Clicking on the arrow opens a menu that displays all unique values that lie within the column.

The menu also makes it possible to select or deselect data based on the unique values. For example, you can filter for blank or non-blank cells, highlighted or non-highlighted cells, or more complex queries.

EXERCISE 2

Consider the demographic data for XYZshopping.com. Column E contains the age, in integers, for each customer. Use the Sort tool to find any string values in the age column.

Specific Validations for Strings

What if you want to know the specific length of string responses? An easy way to measure the length of responses is by using the LEN function in Excel. In a blank column, type the function =LEN() in the cell adjacent to the row you are measuring.

To apply the function to the rest of the column, click on the cell and double click, and a green square will appear on the bottom right corner of the cell. The resulting column will provide integer values for character lengths. These can be sorted in increasing or decreasing order to help find responses that need validation.

EXERCISE 3

Consider the demographic data for XYZshopping.com. Column G contains the source of recommendation through which the customer got to know about XYZshopping.com. Try the LEN() formula and Filter tool to identify responses longer than 12 characters.

CHAPTER 4

Deal with Missing Data

Most data sets have instances of missing data. Incomplete or missing data points, no matter how few, can reduce your sample size (the number of people or entities in your data set) considerably. Learn how to appropriately deal with missing data to ensure the best balance between data accuracy and minimal loss in sample size.

Chapter 4 will help you learn the following techniques:

- Identify the **source** of missing data
- Difference between **independent and dependent variables**
- **Techniques** to deal with missing data

Data is almost never complete or perfect. In most practical cases, some amount of variables would be missing. Missing or partial information, no matter how few in number, can reduce the sample size considerably.

Why Do Missing Values Arise?

Data could have missing values due to a number of reasons. Some of them are avoidable, while some are not, depending on the nature of your questionnaire, audience, and collection method. It is important to plug as many of these gaps as possible during the surveying process.

Incomplete Responses

The individual number of observations (i.e. the sample size) is crucial in data analysis. Before you can start applying statistical techniques to analyze your data, you should ideally have a sample that is large enough to accommodate the diversity of your population while being in the limits of your available survey resources.

Non-responses reduce the sample of responses available for analysis. For example, if you receive only 25 completed surveys out of the 100 that were given out, your sample size automatically reduces to one-fourth of what was intended.

Ambiguous Response Choices

In categorical questions, sometimes choices such as “Prefer not to say” or “Don’t know” are added due to the sensitive nature of questions. Questions related to income, expenditure, ethnicity, among others, often need to include such options. However, while such responses are helpful to respondents, they end up adding little or no value during data analysis.

A way around this problem is to incorporate questions that can extract the required information without making respondents uncomfortable. For example, in order to estimate the respondent's income, they can be questioned about their property, income tax paid, and so on.

Dropouts in Longitudinal Data Analysis

Longitudinal data analysis — research involving the study of subjects over a period of time, ranging from a few months to a few years — often suffers from "dropouts". Participants tend to withdraw from the project at different stages due to various external factors. The final longitudinal data analysis can thus be performed only on a small fraction of the intended sample size you started out with. In such studies, you should make sure to plan appropriately in advance to ensure minimum exits during the research process or have a large enough initial sample size to account for the exits.

Independent and Dependent Variables

Data has two kinds of variables — predictor (or independent) and response (or dependent) variables. Predictor variables influence or affect the values of the response variables.

Consider the following table containing a few examples of independent and dependent variables in different contexts:

	Subject	Independent Variable	Dependent Variable
1	Movie industry	Genre, cast, production house, movie distribution, and marketing	Box office earnings
2	Making cookies	Baking time, oven temperature, inclusion of chocolate chips	Thickness, taste
3	Plant cultivation	Frequency of watering, exposure to light/sun, soil pH level	Plant height and yield

The objective of a study is usually to learn more about the behavior of, or establish patterns in, the values of dependent variables, and then use independent variables to explain their behavior.

EXERCISE 1

Consider the sales data for XYZshopping.com. Assume that we are looking to analyze country-wise consumption patterns — products bought and expenditure entailed. What are the dependent and independent variables in this case?

Dealing with Missing Values

Listwise Deletion

In the listwise deletion method, all rows that have one or more column values missing are deleted.

Missing values in dependent variables would often require you to delete the entire record, since it cannot contribute to the research. Alternatively, for a particular dependent variable, too many missing independent variables can also result in no meaningful insights, which would also require you to delete the entire record.

Pro: Easy to apply, does not tamper with the data
Con: Can greatly reduce your sample size

	A	B	C	D	E	F	G
1	Original Data set				Data set after Listwise deletion		
2	Name	Age	Gender		Name	Age	Gender
3	Robin	28	Male		Robin	28	Male
4	Heather	29	Female		Heather	29	Female
5	Jamie	22			Carl	32	Male
6	Carl	32	Male		Sarah	26	Female
7		35	Male				
8	Sarah	26	Female				
9							
10							

TOOL TIP

Use this method when — in a particular row — either the dependent variable being studied or too many independent variables are missing.

EXERCISE 2

Consider the sales data for XYZshopping.com. Assume that we are looking to analyze country-wise consumption patterns (products bought and expenditures). Look at missing values for Product and Price, then identify rows that you should delete for this analysis.

Mean/Median/Mode Imputation

In the mean/median/mode imputation method, all missing values in a particular column are substituted with the mean/median/mode, which is calculated using all the values available in that column. You can use appropriate functions in Excel to compute the mean/median/mode by simply plugging in the range of the column into the input of the function.

- **Mean**: Mean — commonly known as average — is equal to the sum of all values in the column divided by the number of values present in the column. In Excel, use the AVERAGE() function to directly compute the mean.
- **Median**: Median is the "middle" value amongst the range of values. To compute the median of a range containing 'n' number of values, you need to sort these 'n' values in ascending order.
    - For an odd number of observations, the median is the ((n+1)/2)th value. For example, the median for a sorted list of 13 observations is the 7th value.
    - For an even number of observations, the median is the average of the (n/2)th and ((n+2)/2)th values. For example, the median for a sorted list of 12 observations is the average of the 6th and 7th values.
    - In Excel, use the MEDIAN() function to directly compute the median.
- **Mode**: Mode is the value that occurs the most often in the range of values. In Excel, use the MODE() function to directly compute the mode.

Pro: No loss in sample size, no skewing of data

Con: Cannot be applied on categorical variables, i.e. non-numerical/qualitative data

TOOL TIP

Use this method for missing values only when the potential loss in sample size from listwise deletion is significant and unaffordable.

EXERCISE 3

Consider the sales data for XYZshopping.com. Column D contains the price of the item purchased. Compute the mean, median, and mode for this column.

Last Observation Carried Forward (LOCF)

LOCF is a technique specific to longitudinal data analysis. This is a crude method where a missing value for a particular row is filled in with a value available from the previous stages.

Pro: Ensures no sample size loss from dropouts

Con: Can only be applied to longitudinal data analysis

	A	B	C	D	E	F	G	H	I
1	**Original Data Set**					**Data After LOCF**			
2	Name	Visit	Month	Weight		Name	Visit	Month	Weight
3	Robin	1	January	65		Robin	1	January	65
4	Robin	2	February	68		Robin	2	February	68
5	Robin	3	March			Robin	3	March	68
6	Robin	4	April			Robin	4	April	68
7	Robin	5	May	72		Robin	5	May	72
8	Robin	6	June	71		Robin	6	June	71
9	Heather	1	January	52		Heather	1	January	52
10	Heather	2	February	51		Heather	2	February	51
11	Heather	3	March	56		Heather	3	March	56
12	Heather	4	April	52		Heather	4	April	52
13	Heather	5	May			Heather	5	May	52
14	Heather	6	June			Heather	6	June	52
15	Jamie	1	January			Jamie	1	January	-
16	Jamie	2	February	78		Jamie	2	February	78
17	Jamie	3	March	81		Jamie	3	March	81
18	Jamie	4	April			Jamie	4	April	81
19	Jamie	5	May			Jamie	5	May	81
20	Jamie	6	June	75		Jamie	6	June	75
21									

TOOL TIP

Try not to use this method for more than 3 continuous stages, in which case it is better to opt for listwise deletion.

CHAPTER 5

Handle Outlier Detection

Data points that are much bigger or smaller than most other data points are called outliers. Outlier detection is crucial since these can skew the interpretation of data. Outliers can be accurate or inaccurate. Learn different ways to detect outliers and deal with inaccurate data better.

Chapter 5 will help you learn the following techniques:

- **Importance** of outlier detection
- Outlier detection using **visualization**
- Outlier detection using **statistical techniques**
- **Dealing** with outliers

Imagine that you generally keep spare change and small bills in your pocket. If you reach in your pocket and find a $1 bill, a quarter, a dime, and 3 pennies, you won't be surprised. If you find a $100 bill, you will certainly be surprised.

That $100 bill is an outlier — a data point that is much bigger or smaller than other data points.

Outliers can represent accurate or inaccurate data. For example, if you reported finding a $200 bill in your pocket, people would rightly ignore your story. That outlier would be inaccurate, since $200 bills do not exist. This is likely to be misreporting for a $20 bill. However, a report of finding a $100 bill could be an accurate outlier. While that data point is abnormal, it is possible. Perhaps you had just withdrawn $100 from an ATM with no small bills.

It is important to find and deal with outliers, since they can skew interpretation of the data. For example, imagine that you want to know how much money you keep in your pocket each day. At the end of each day, you empty your pockets, count the money, and record the total. The results after 12 days are in the table below.

	Day	Total money
1	Day 1	1.38
2	Day 2	0.45
3	Day 3	4.23
4	Day 4	101.2
5	Day 5	2.5
6	Day 6	1.77
7	Day 7	0.25
8	Day 8	0.68
9	Day 9	3.32
10	Day 10	1
11	Day 11	9.04
12	Day 12	0.1

Day 4 is clearly an outlier. If you exclude Day 4 from your calculations, you would conclude that you keep an average of $2.25 in your pocket. However, if you don't exclude Day 4, the average money in your pocket would be $10.49. These are vastly different results.

Outliers are inevitable, especially for large data sets. On their own, they are not problematic. However, in the context of the larger data set, it is essential to identify, verify, and accordingly deal with outliers to ensure that your data interpretation is as accurate as possible.

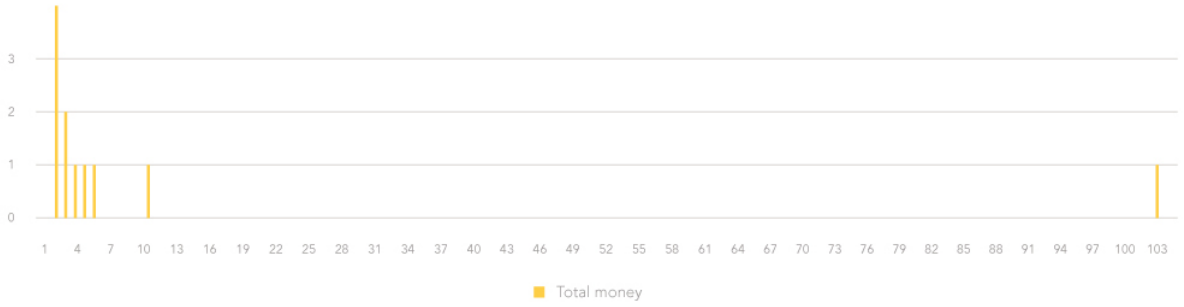
Identifying Outliers

The first step in dealing with outliers is finding them. There are two ways to approach this.

Visualize the Data

Depending on your data set, you can use some simple tools to visualize your data and spot outliers visually.

Histogram: A histogram is the best way to check univariate data — data containing a single variable — for outliers. A histogram divides the range of values into various groups (or buckets), and then shows the frequency — how many times the data falls into each group — through a bar graph. Assuming that these buckets are arranged in increasing order, you should be able to spot outliers easily at the far left (very small values) or at the far right (very large values).



This histogram of our pocket change example shows an outlier on the far right for Day 4 ($101.2).

EXERCISE 1

Consider the sales data for XYZshopping.com. Column D contains the price of the product purchased in each transaction. Create a histogram for this data.

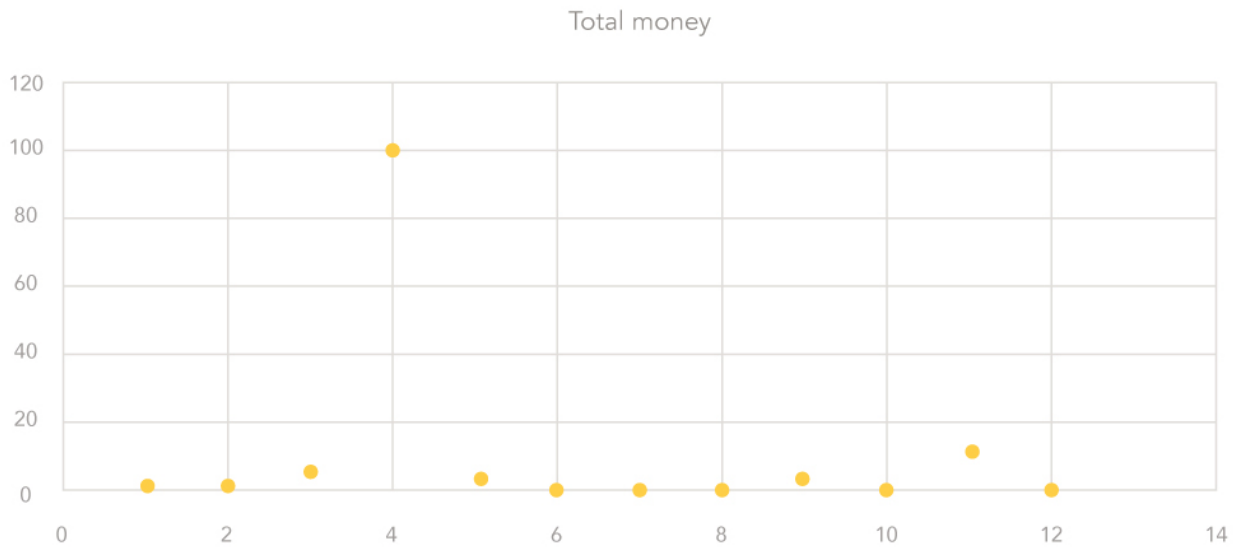
Hint: You can create a histogram in Excel using the "Data Analysis" tool by following these steps:

1. Add the buckets/bins, for which you want to count the frequencies, in a separate column.
2. Select the "Data Analysis" option inside the "Data" tab of the Excel ribbon.
3. Provide the input range and the bin range in the pop-up dialog box and click OK.
4. The data for your histogram (Bin and Frequency) will get populated in a fresh Excel sheet. You can now use a bar chart to visualize this.

Scatter Plot: A scatter plot (also called a scatter diagram or scatter graph) shows a collection of points on an x-y coordinate axis, where the x-axis (horizontal axis) represents the independent variable and the y-axis (vertical axis) represents the dependent variable.

A scatter plot is useful to find outliers in bivariate data (data with two variables). You can easily spot the outliers because they will be far away from the majority of points on the scatter plot.

This scatter plot of our pocket change example shows an outlier — far away from all the other points — for Day 4 ($101.2).



EXERCISE 2

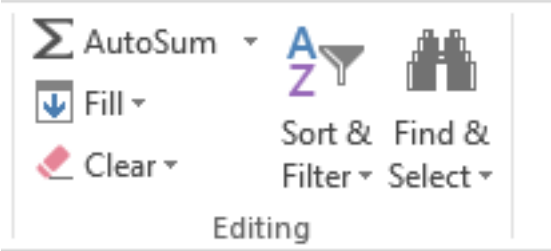
Consider the sales data for XYZshopping.com. Column D contains the price of the product purchased in each transaction, whereas Column B contains the transaction date. Create a scatterplot containing transaction date on the x-axis and price on the y-axis.

The Statistical Way

Using statistical techniques is a more thorough approach to identifying outliers. There are several advanced statistical tools and packages that you could use to identify outliers.

Here we'll talk about a simple, widely used, and proven technique to identify outliers.

Step 1: Sort the Data



Sort the data in the column in ascending order (smallest to largest). You can do this in Excel by selecting the "Sort & Filter" option in the top right in the home toolbar.

Sorting the data helps you spot outliers at the very top or bottom of the column. Here is the data sorted for the pocket change example we mentioned earlier.

	A	B
1	Day	Total money ($)
2	Day 12	0.1
3	Day 7	0.25
4	Day 2	0.45
5	Day 8	0.68
6	Day 10	1
7	Day 1	1.38
8	Day 6	1.77
9	Day 5	2.5
10	Day 9	3.32
11	Day 3	4.23
12	Day 11	9.04
13	Day 4	101.2

You can easily spot the $101.2 value right at the bottom of the column. However, there could be more outliers that might be difficult to catch.

TOOL TIP

It is always good to copy the data to be sorted in a different location so that you don't mess up the actual order of the data.

EXERCISE 3a

Consider the sales data for XYZshopping.com. Column D contains the price of the product purchased in each transaction. Copy the data into a separate column and sort it in ascending order.

Step 2: Quartiles

In any ordered range of values, there are three quartiles that divide the range into four equal groups. The second quartile (Q2) is nothing but the median, since it divides the ordered range into two equal groups. For an odd number of observations, the median is equal to the middle value of the sorted range.

	A	B	C
1	Day	Total money ($)	
2	Day 12	0.1	
3	Day 7	0.25	
4	Day 2	0.45	0.565
5	Day 8	0.68	
6	Day 10	1	
7	Day 1	1.38	1.575
8	Day 6	1.77	
9	Day 5	2.5	
10	Day 9	3.32	3.775
11	Day 3	4.23	
12	Day 11	9.04	
13	Day 4	101.2	

In this example, since we have an even number of observations (12), we need to calculate the average of the sixth and seventh-position values in the ordered range — that is, the average of 1.38 and 1.77. The median of the range works out to be 1.575.

To calculate the first (Q1) and third quartiles (Q3), you need to simply calculate the medians of the first half and second half respectively. In this case, Q1 is 0.565 and Q3 is 3.775.

You can directly use the MEDIAN() and QUARTILE() functions in Excel to calculate the median and quartiles. The syntax for the median function is:

```
=median(comma-separated list of values or array containing values)
```

The syntax for the quartile function is:

```
=quartile(array containing values, 1/2/3 for Q1/Q2/Q3
respectively)
```

Bear in mind that Excel calculates quartiles differently from the method shown above, although both methods are correct.

EXERCISE 3b

Find the median, Q1, and Q3 for the sorted data from Exercise 3a.

Step 3: Inner and Outer Fences

The inner and outer fences are ranges that you can calculate using the Q1 and Q3. To do this, you need to first calculate the interquartile range — the difference between Q1 and Q3. In this case, Q3-Q1 = 3.21.

A data point that falls outside the inner fence is called a minor outlier.

```
Inner Fence
Lower bound = Q1 - (1.5 * (Q3-Q1))
Upper bound = Q3 + (1.5 * (Q3-Q1))
```

In our example, the bounds for the inner fence are:

Lower Bound = 0.565 - (1.5*3.21) = -4.25
Upper Bound = 3.775 + (1.5*3.21) = 8.59

The data points for Day 11 and Day 4, that is 9.04 and 101.20 respectively, qualify as

minor outliers.

A data point that falls outside the outer fence is called a major outlier.

```
Outer Fence
Lower bound = Q1 - (3 * (Q3-Q1))
Upper bound = Q3 + (3 * (Q3-Q1))
```

In our example, the bounds for the outer fence are:

Lower Bound = 0.565 - (3*3.21) = -9.07
Upper Bound = 3.775 + (3*3.21) = 13.41

The data point for Day 11 (which is $101.20) qualifies as a major outlier.

EXERCISE 3c
Find the minor and major outliers for the data from Exercise 3b.

Dealing With Outliers

Now that you have identified all your outliers, you should look at each outlier in the context of the other data points in the range, as well as the whole data set. This requires prior knowledge on the nature of the data set, data validation protocols (as detailed in Chapter 3), and the behavior of the variable you are analyzing.

For example, you have the following data points as peak temperature of Delhi (in

Celsius) over the past two weeks: 30°, 31°, 28°, 30°, 31°, 33°, 32°, 31°, 300°, 30°, 29°, 28°, 30°, 31°. Day 9 had a peak temperature of 300°C, which is clearly unrealistic. On the other hand, when you look at the pocket change example, it is not unrealistic to have $101.20 in your pocket. It is possible that you just withdrew $100 from the ATM right before you recorded the data point.

To handle such situations, it is a good practice to have protocols in place to verify each outlier.

If your outlier is verified to be correct, you should leave it untouched. Such an outlier, in fact, emerges as a useful insight from your data set — and is worth looking into.

If the outlier is incorrect, there are two ways to deal with it:

1. **Resurvey the data point:** This is the most foolproof way of dealing with incorrect outliers. Resurveying becomes easier while using mobile-based data collection applications like Collect.

2. **Delete the outlier data point:** Resurveying may not be feasible in all cases due to resource constraints. In such situations, you should delete the outlier data point such that it becomes a missing value. You can now deal with this missing value using the techniques outlined in Chapter 4.

CHAPTER 6

Tackle Conditional Questions

In a questionnaire, conditional questions are asked based on the response to a previous question. Conditional questions help you gain more information on the response to a previous question. While adding speed and clarity to a questionnaire, conditional questions also add complexity. Learn about dealing with conditional questions in your data cleaning processes.

Chapter 6 will help you learn the following techniques:

- **Building** conditional questions into your survey
- **Representing** conditional questions in your data set
- **Data cleaning protocol** for conditional questions

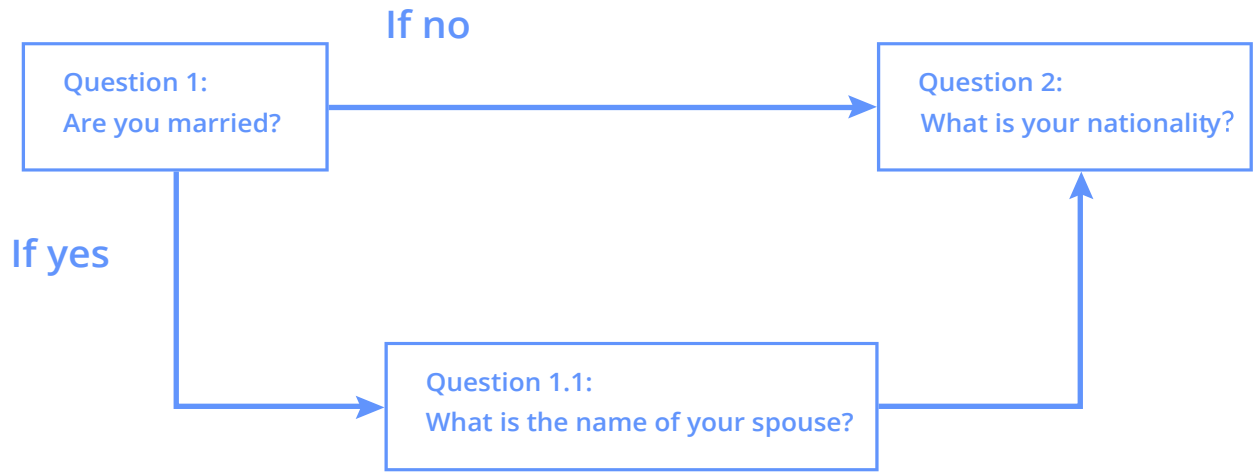
Conditionality allows you to direct a respondent to a specific question based on his/her answer to a prior question. For example, the question, "Are you pregnant?" should be asked to a respondent in a survey only if the answer to the question "What is your gender?" is "Female".

Conditional branching takes this a step further and directs the user to a different questionnaire based on their answer to a prior question. For example, you may have a separate set of questions for people in different age groups. The more complex the survey — a country's census, for instance — the higher the need for conditional branching in your survey. Conditionality helps optimize the time and effort spent on conducting a survey.

In large-scale surveys, conditional branching helps break data sets into fragments that can be tackled individually in the cleaning process later. Consider a survey that has a few common questions before it branches out separately for males and females at the question, "What is your gender?" In a data collection tool that generates separate reports, for instance, an analysis on female participants would become quick and easy since you will only have to clean data in the branched report generated for female participants.

Building Conditional Questions into the Survey

It is important to visualize the flow of your survey while building conditionality into it. A flowchart is an easy and effective tool to help you do this.



In this example, Question 1.1 should be asked to the respondent only if the answer to Question 1 is "Yes". If the answer is "No", the user should be sent to Question 2. The survey should explicitly specify which questions are conditional.

In such cases, planning the survey in a flowchart and using a clear protocol to assign serial numbers to questions becomes crucial. Numbering your questions in a format similar to the one shown above is also useful in the data cleaning process.

Representing Conditionality in the Data Set

You can choose to represent conditionality in your data set in various ways.

Responses to questions in a conditional branch can be represented in a separate (secondary) table mapped to the main (primary) table. This mapping can be done in an Excel spreadsheet by using the VLOOKUP function. (We shall focus on the VLOOKUP function in detail in Chapter 7.)

You may also choose to keep the conditional branches within the main data set by creating separate variables that indicate whether a particular question is conditional or not.

Data Cleaning Protocol for Conditional Questions

Crosschecking with the Survey

The first step to cleaning a data set containing conditional questions is to crosscheck all questions — conditional and non-conditional — against the original survey. Make sure that your data set follows the flow that you outlined in your flowchart earlier.

All questions involved in conditionality need to be checked carefully. For example, if the conditional question "What is the name of your spouse?" has a string response while the question "Are you married?" draws the value "No", you know something is wrong.

Consistency Check Using IF/ELSE

You can check for consistency in your data set using IF/ELSE options in platforms like Excel. The syntax for the IF() statement in Excel is pretty straightforward.

```
Syntax: =IF(condition, true result, false result)
```

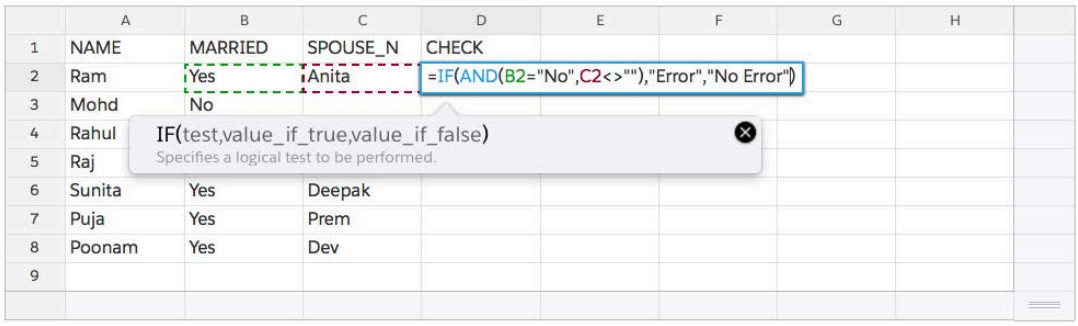
where

- condition: the condition you want to test
  (for example: gender = "male" and pregnancy_status = non-blank)
- true result: output if condition is true (for example: "error")
- false result: output if condition is false (for example: "good data")

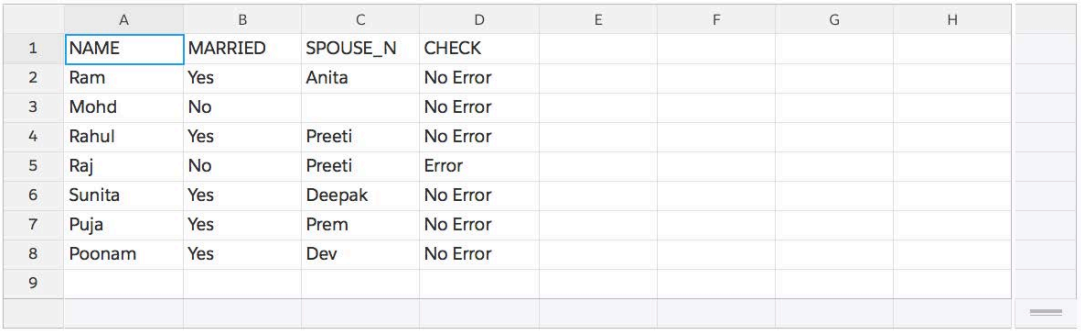
It is always a good a practice to quickly add another column in the Excel sheet and use the IF() formula to check for consistency.

Consider the example where we need to check for data consistency in marital status and spouse name. The image shows the formula that we need to use to check this.

The formula returns “Error” in Column D if it finds a non-blank spouse name for an unmarried person.



Row 5 contains an error because the person in this record is not married (Cell B5 = “No”), but has a spouse (Cell C5 = Non blank)



You can now quickly filter on these “Error” rows using the “Filter” command on Column D, then clean the data in these rows.

EXERCISE 1

In the demographic survey mentioned in Chapter 1, Question 4 has conditional branches:

1. Q4: How did you come to know about our online shopping site?
    a. Recommendation from someone
    b. Advertisement
    c. Online search
2. Q4a: Who recommended our site to you?
    a. Friend
    b. Colleague
    c. Family
    d. Other (please specify)
3. Q4b: Where did you first see our advertisement?
    a. Facebook
    b. TV

Question 4a should only be answered by people who chose "Recommendation from someone" for Question 4, and Question 4b should only be answered by people who chose "Advertisement". There isn't a follow-up conditional question for "Online search".

Consider the demographic data for XYZshopping.com. The rows that do not have "Advertisement" in Column F should have "NA" under Column H (since Column H is a conditional question which gets asked only after the respondent has gone with the "Discovery" = "Advertisement" option). Identify the rows where this does not hold.

CHAPTER 7

Join, Delimit, or Concatenate Data

Your data is now clean and structured. Before you go ahead and analyze it, learn these basic data functions to join, delimit, and concatenate your data. These functions will help you make better sense of your data and draw better insights.

Chapter 7 will help you learn the following techniques:

- Join different data sets using **unique IDs**
- Delimit data to help create more **granular fields**
- Concatenate data for **advanced analysis**

Once your survey data is structured and cleaned at a basic level, you may need to carry out a series of functions across your data set. In this chapter, we cover frequently used operations that will make your data set ready for analysis.

Joining Data

In Chapter 1, we discussed the importance of creating unique IDs in surveys. One reason why UIDs are important is because separate data sets can be combined based on the UID. This process is called joining data. It involves using the UID to match data in one data set with that in another.

In Excel, the VLOOKUP formula is one way to join data sets on the basis of a common UID. This is how the VLOOKUP formula looks in Excel:

```
=VLOOKUP(lookup_value, table_array, col_index_num, [range_lookup])
```

This formula is a powerful tool to save hours of manual effort spent on matching data manually. The table below explains each aspect of the formula:

	Formula	B
1	=	All formulas in Excel start with the = symbol. This is the user's way of telling Excel that they will be typing in a formula
2	VLOOKUP	VLOOKUP is the name of the function. This tells Excel how to treat the values you enter next.
3	lookup_value	This is the shared UID you'll use to join the data sets.
4	table_array	The table array is the range of cells within which Excel looks up the corresponding value for your lookup value. The table array must contain the common UIDs in its first column
5	col_index_num	This is the column number to be looked up in the table_array. This number should a whole number ranging from 1 to the total number of columns in the table_array
6	range_lookup (optional, default TRUE)	This tells Excel whether to look up approximate or exact matches. If you specify TRUE, Excel finds the closest value to complete the match. If you specify FALSE, Excel only returns matches for the exact value found in the first column.

Now let's look at an example of how to use the VLOOKUP formula to combine two sheets in Excel.

In this example, there are two sheets:

- The "**Description**" sheet includes an ID for each piece of produce (Produce_ID), the color of each produce (Color), and a description of the produce (Description).
- The "**Cost**" sheet includes the same produce ID as the first sheet (Produce_ID) and cost information for one kilogram of each product (Cost_Per_Kg).

Let us assume that a store owner wants to view both the produce description and cost information on the same page. We can use the VLOOKUP function to combine data from both sheets.

```
=VLOOKUP(Description!A2, Cost!A2:B11, 2, FALSE)
```

Here's a breakdown of how the function will operate:

	A	Formula	Explanation
1	=	=	
2	VLOOKUP	VLOOKUP	
3	lookup_value	Description!A2	The lookup value is Description!A2. "Description!" tells Excel that the lookup value is present in cell A2 in sheet "Description". This means we are matching information with this product ID. **Remember, we are using the Produce_ID for the lookup because it is present in both sheets.**
4	table_array	Cost!A2:B11	The table array is Cost!A2:B11. "Cost!" tells Excel that the table_array is present in the Cost sheet. The range A2:B11 is the specific table_array range where the match needs to be found.
5	col_index_num	2	The column index number is 2, because we are extracting the Cost_Per_Kg (which is in column 2 of the lookup_array) for a given Produce_ID.
6	range_lookup (default TRUE)	FALSE	We specified FALSE because we do not want approximate matches. Only exact matches will be returned now.

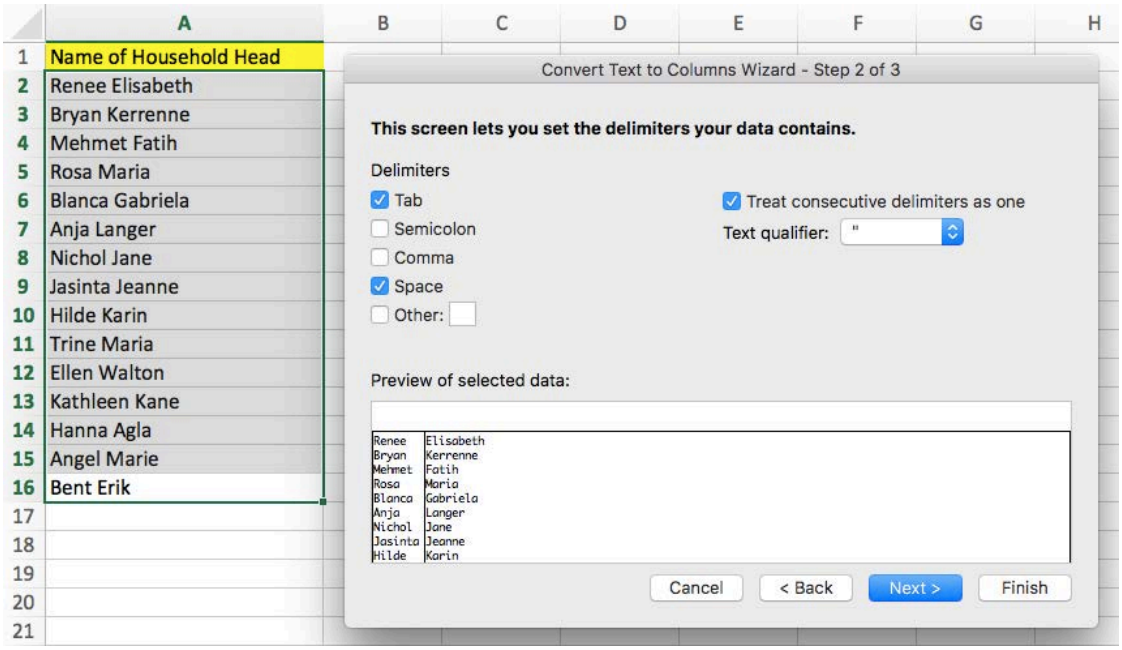
Delimiting Data

Delimiting data is helpful to automate data segregation. Imagine you are analyzing responses to the survey question, "What is the name of the household head?" and the

responses include both the first and last name of the household head separated by a space. If you wish to separate this information into two columns — first_name and last_name — you can delimit the values using the space as the delimiting character.

In Excel, click on "Text to Columns" in the "Data" tab of the Excel ribbon. A dialogue box will pop up that says "Convert Text to Columns Wizard". Select the "Delimited" option. Now choose the delimiting character to split the values in the column. You can see a preview below. Click on 'Next' for additional options and then click 'Finish' once you are satisfied with the preview.

This example shows how you can separate the first and last names in this column by using the space character as a delimiter.



Once you click on "Finish", you will have two columns — containing the first and last names.

Concatenating Data

Concatenating data achieves the opposite of delimiting data — it combines strings from separate fields into one text string. The concatenate function in Excel allows us to carry out this function across all cells. Here's how the concatenate function looks in Excel:

```
=CONCATENATE(comma-separated list of strings to be concatenated)
```

Let's again use the example of names; assume that we want to combine the first and last names of each household head into one cell with an underscore in between the names. We can concatenate this by applying the formula:

```
=CONCATENATE(first_name,"_",last_name)
```

EXERCISE 1

Consider the sales data for XYZshopping.com. Try concatenating the "City", "State", and "Country" by using commas to separate each field, under a new column titled "Address" for every customer. For example: In the row with serial number 1, the address will be "Basildon, England, United Kingdom".

CHAPTER 8

Case Study: Cleaning Data from a Paper-Based Survey

At **SocialCops**, our partners have collected data for hundreds of surveys on our mobile data collection tool — Collect. However, we've learned from experience that cleaning data from a paper-based survey comes with its unique challenges. SocialCops has worked on cleaning data in one such survey. This case study will help you understand the challenges involved and the basic steps followed in cleaning data from this paper-based survey.

Now that we have learned the concepts behind basic data cleaning, let us delve deeper into a case study. This is a real-life example from one of the projects SocialCops has worked on.

In December 2014, under the Sansad Adarsh Gram Yojana (SAGY) — a rural development program in India focused on tracking and improving various village-level socio-economic indicators — a baseline household survey was conducted in Gollamandala village in Krishna district in the state of Andhra Pradesh. The survey was done to create a socio-economic and sectoral (education, health, employment, etc.) profile of the village, based on which decisions could be made to build it into a model village.

This was followed by a series of data cleaning exercises to ensure that the resulting data set was clean and credible for further analysis.

1. Data Structure

The first raw data set had 1,060 rows and 182 columns. First things first, we carried out three important checks on the data structure.

Identifying the Unique ID (UID)

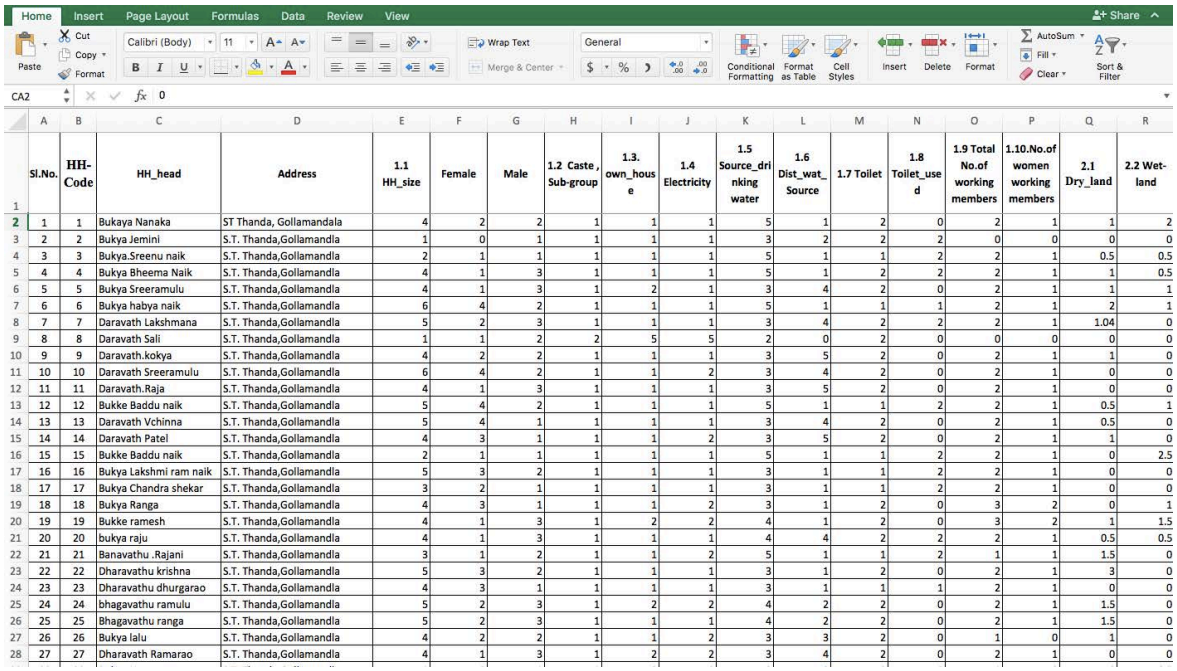
In this data set, the Unique ID was the Household ID. Before finalizing this as the UID, we checked for duplicates. We identified the duplicate IDs using the "Conditional Formatting" tool in Excel and selected the option "Duplicate Values" under the "Highlight Color Rules" option. There were 7 duplicate UIDs in all.

- Six of these had duplicate UIDs, but their combinations of household head, address, and household composition were unique. We assigned new UIDs to

these, since they were essentially different households, but with incorrect UIDs.

- One of these had a repeated combination of household head, address, and household composition. Since this was clearly a case of data duplication, we deleted the entire record.

Our data set now had 1,059 rows.



Sl.No.	HH-Code	HH_head	Address	1.1 HH_size	Female	Male	1.2 Caste , Sub-group	1.3. own_house	1.4 Electricity	1.5 Source_drinking water	1.6 Dist_wat_ Source	1.7 Toilet	1.8 Toilet_used	1.9 Total No.of working members	1.10.No.of women working members	2.1 Dry_land	2.2 Wet-land
1	1	Bukaya Nanaka	ST Thanda, Gollamandala	4	2	2	1	1	1	5	1	2	0	2	1	1	2
2	2	Bukya Jemini	S.T. Thanda,Gollamandla	1	0	1	1	1	1	3	2	2	2	0	0	0	0
3	3	Bukya.Sreenu naik	S.T. Thanda,Gollamandla	2	1	1	1	1	1	5	1	1	2	2	1	0.5	0.5
4	4	Bukya Bheema Naik	S.T. Thanda,Gollamandla	4	1	3	1	1	1	5	1	2	2	2	1	1	0.5
5	5	Bukya Sreeramulu	S.T. Thanda,Gollamandla	4	1	3	1	2	1	3	4	2	0	2	1	1	1
6	6	Bukya habya naik	S.T. Thanda,Gollamandla	6	4	2	1	1	1	5	1	1	1	2	1	2	1
7	7	Daravath Lakshmana	S.T. Thanda,Gollamandla	5	2	3	1	1	1	3	4	2	2	2	1	1.04	0
8	8	Daravath Sali	S.T. Thanda,Gollamandla	1	1	2	2	5	5	2	0	2	0	0	0	0	0
9	9	Daravath.kokya	S.T. Thanda,Gollamandla	4	2	2	1	1	1	3	5	2	0	2	1	1	0
10	10	Daravath Sreeramulu	S.T. Thanda,Gollamandla	6	4	2	1	1	2	3	4	2	0	2	1	0	0
11	11	Daravath.Raja	S.T. Thanda,Gollamandla	4	1	3	1	1	1	3	5	2	0	2	1	0	0
12	12	Bukke Baddu naik	S.T. Thanda,Gollamandla	5	4	2	1	1	1	5	1	1	2	2	1	0.5	1
13	13	Daravath Vchinna	S.T. Thanda,Gollamandla	5	4	1	1	1	1	3	4	2	0	2	1	0.5	0
14	14	Daravath Patel	S.T. Thanda,Gollamandla	4	3	1	1	1	2	3	5	2	0	2	1	1	0
15	15	Bukke Baddu naik	S.T. Thanda,Gollamandla	2	1	1	1	1	1	5	1	1	2	2	1	0	2.5
16	16	Bukya Lakshmi ram naik	S.T. Thanda,Gollamandla	5	3	2	1	1	1	3	1	1	2	2	1	0	0
17	17	Bukya Chandra shekar	S.T. Thanda,Gollamandla	3	2	1	1	1	1	3	1	1	2	2	1	0	0
18	18	Bukya Ranga	S.T. Thanda,Gollamandla	4	3	1	1	1	2	3	1	2	0	3	2	0	1
19	19	Bukke ramesh	S.T. Thanda,Gollamandla	4	1	3	1	2	2	4	1	2	0	3	2	1	1.5
20	20	bukya raju	S.T. Thanda,Gollamandla	4	1	3	1	1	1	4	4	2	2	2	1	0.5	0.5
21	21	Banavathu .Rajani	S.T. Thanda,Gollamandla	3	1	2	1	1	2	5	1	1	2	1	1	1.5	0
22	22	Dharavathu krishna	S.T. Thanda,Gollamandla	5	3	2	1	1	1	3	1	2	0	2	1	3	0
23	23	Dharavathu dhurgarao	S.T. Thanda,Gollamandla	4	3	1	1	1	1	3	1	1	1	2	1	0	0
24	24	bhagavathu ramulu	S.T. Thanda,Gollamandla	5	2	3	1	2	2	4	2	2	0	2	1	1.5	0
25	25	Bhagavathu ranga	S.T. Thanda,Gollamandla	5	2	3	1	1	1	4	2	2	0	2	1	1.5	0
26	26	Bukya lalu	S.T. Thanda,Gollamandla	4	2	2	1	1	2	3	3	2	0	1	0	1	0
27	27	Dharavath Ramarao	S.T. Thanda,Gollamandla	4	1	3	1	2	2	3	4	2	0	2	1	0	0

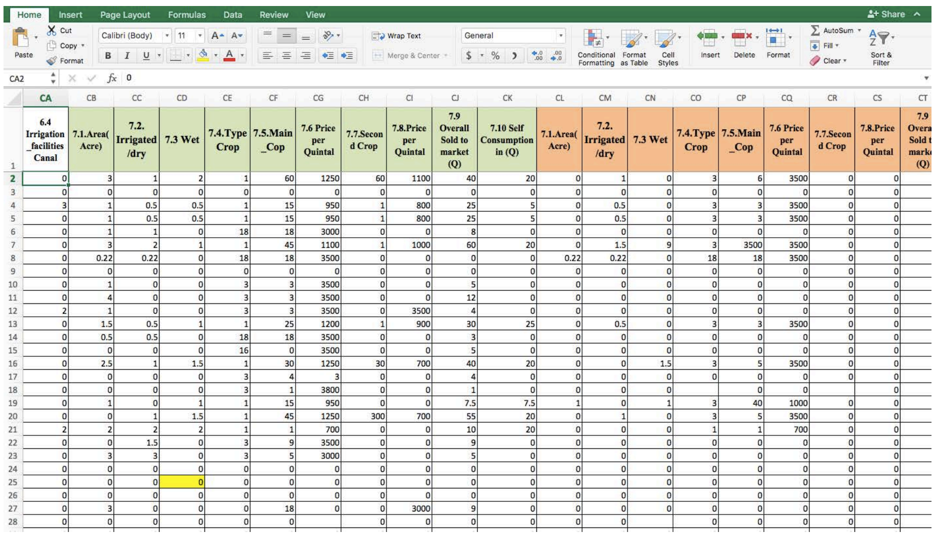
Comparing the Columns to the Questionnaire

We then matched the 182 columns (variables) with the questions in the questionnaire. We found missing column headers, repetitive column headers, and columns with repetitive header and data.

- Two columns — one categorical and one with decimal values on land area — had been merged into one. Here's where matching the data types with those specified in the questionnaire came handy in separating it into two columns.
- 2 sets of 10 columns each had repetitive headers across the set, making all 20 column headers error-prone. But, when compared with questionnaire, we found

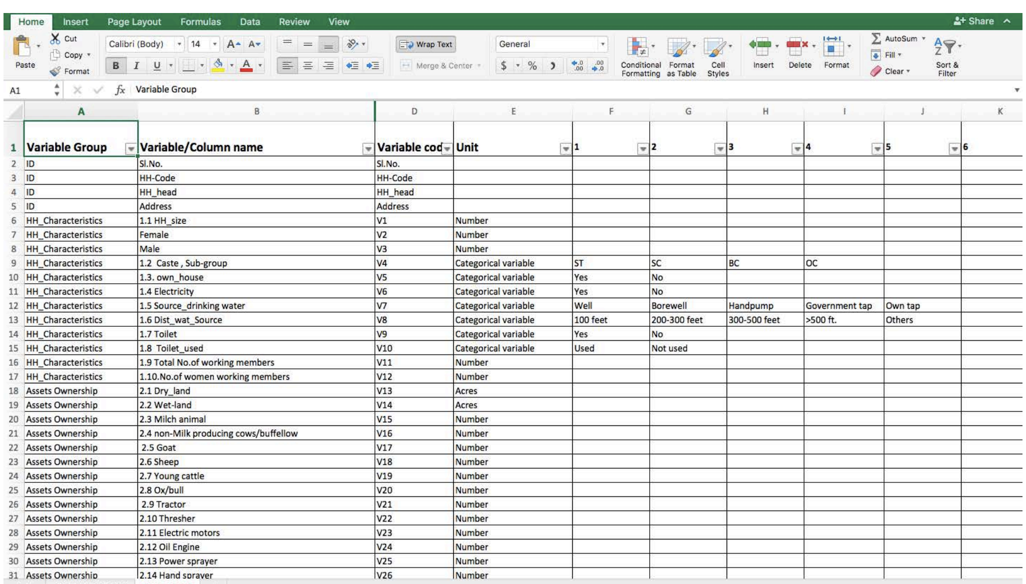
that these two sets represented land and crop characteristics across 2 different plots — Plot 1 and Plot 2. The idea was to compare these characteristics for two different plots.

Improper data entry leads to errors like these. A good prior understanding of survey questions helps tackle such issues better, both after and during the data entry process (in offline data collection methods).



	CA	CB	CC	CD	CE	CF	CG	CH	CI	CJ	CK	CL	CM	CN	CO	CP	CQ	CR	CS	CT
1	6.4 Irrigation _facilities Canal	7.1.Area(Acre)	7.2. Irrigated /dry	7.3 Wet	7.4.Type Crop	7.5.Main _Cop	7.6 Price per Quintal	7.7.Secon d Crop	7.8.Price per Quintal	7.9 Overall Sold to market (Q)	7.10 Self Consumption in (Q)	7.1.Area(Acre)	7.2. Irrigated /dry	7.3 Wet	7.4.Type Crop	7.5.Main _Cop	7.6 Price per Quintal	7.7.Secon d Crop	7.8.Price per Quintal	7.9 Overa Sold t mark (Q)
2	0	3	1	2	1	60	1250	60	1100	40	20	0	1	0	3	6	3500	0	0	
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
4	3	1	0.5	0.5	1	15	950	1	800	25	5	0	0.5	0	3	3	3500	0	0	
5	0	1	0.5	0.5	1	15	950	1	800	25	5	0	0.5	0	3	3	3500	0	0	
6	0	1	1	0	18	18	3000	0	0	8	0	0	0	0	0	0	0	0	0	
7	0	3	2	1	1	45	1100	1	1000	60	20	0	1.5	9	3	3500	3500	0	0	
8	0	0.22	0.22	0	18	18	3500	0	0	0	0	0.22	0.22	0	18	18	3500	0	0	
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
10	0	1	0	0	3	3	3500	0	0	5	0	0	0	0	0	0	0	0	0	
11	0	4	0	0	3	3	3500	0	0	12	0	0	0	0	0	0	0	0	0	
12	2	1	0	0	3	3	3500	0	3500	4	0	0	0	0	0	0	0	0	0	
13	0	1.5	0.5	1	1	25	1200	1	900	30	25	0	0.5	0	3	3	3500	0	0	
14	0	0.5	0.5	0	18	18	3500	0	0	3	0	0	0	0	0	0	0	0	0	
15	0	0	0	0	16	0	3500	0	0	5	0	0	0	0	0	0	0	0	0	
16	0	2.5	1	1.5	1	30	1250	30	700	40	20	0	0	1.5	3	5	3500	0	0	
17	0	0	0	0	3	4	3	0	0	4	0	0	0	0	0	0	0	0	0	
18	0	0	0	0	3	1	3800	0	0	1	0	0	0	0		0	0		0	
19	0	1	0	1	1	15	950	0	0	7.5	7.5	1	0	1	3	40	1000	0	0	
20	0	0	1	1.5	1	45	1250	300	700	55	20	0	1	0	3	5	3500	0	0	
21	2	2	2	2	1	1	700	0	0	10	20	0	0	0	1	1	700	0	0	
22	0	0	1.5	0	3	9	3500	0	0	9	0	0	0	0	0	0	0	0	0	
23	0	3	3	0	3	5	3000	0	0	5	0	0	0	0	0	0	0	0	0	
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
27	0	3	0	0	0	18	0	0	3000	9	0	0	0	0	0	0	0	0	0	
28	0	0	0	0	0	0		0	0	0	0	0	0		0	0	0	0	0	

We created a variable directory with variable category/group, name, renamed code, and unit. This was useful in uploading data to different software and understanding the inter-dependence of variables.



	Variable Group	Variable/Column name	Variable cod	Unit	1	2	3	4	5	6
2	ID	Sl.No.	Sl.No.							
3	ID	HH-Code	HH-Code							
4	ID	HH_head	HH_head							
5	ID	Address	Address							
6	HH_Characteristics	1.1 HH_size	V1	Number						
7	HH_Characteristics	Female	V2	Number						
8	HH_Characteristics	Male	V3	Number						
9	HH_Characteristics	1.2 Caste , Sub-group	V4	Categorical variable	ST	SC	BC	OC		
10	HH_Characteristics	1.3. own_house	V5	Categorical variable	Yes	No				
11	HH_Characteristics	1.4 Electricity	V6	Categorical variable	Yes	No				
12	HH_Characteristics	1.5 Source_drinking water	V7	Categorical variable	Well	Borewell	Handpump	Government tap	Own tap	
13	HH_Characteristics	1.6 Dist_wat_Source	V8	Categorical variable	100 feet	200-300 feet	300-500 feet	>500 ft.	Others	
14	HH_Characteristics	1.7 Toilet	V9	Categorical variable	Yes	No				
15	HH_Characteristics	1.8 Toilet_used	V10	Categorical variable	Used	Not used				
16	HH_Characteristics	1.9 Total No.of working members	V11	Number						
17	HH_Characteristics	1.10.No.of women working members	V12	Number						
18	Assets Ownership	2.1 Dry_land	V13	Acres						
19	Assets Ownership	2.2 Wet-land	V14	Acres						
20	Assets Ownership	2.3 Milch animal	V15	Number						
21	Assets Ownership	2.4 non-Milk producing cows/buffellow	V16	Number						
22	Assets Ownership	2.5 Goat	V17	Number						
23	Assets Ownership	2.6 Sheep	V18	Number						
24	Assets Ownership	2.7 Young cattle	V19	Number						
25	Assets Ownership	2.8 Ox/bull	V20	Number						
26	Assets Ownership	2.9 Tractor	V21	Number						
27	Assets Ownership	2.10 Thresher	V22	Number						
28	Assets Ownership	2.11 Electric motors	V23	Number						
29	Assets Ownership	2.12 Oil Engine	V24	Number						
30	Assets Ownership	2.13 Power sprayer	V25	Number						
31	Assets Ownership	2.14 Hand sprayer	V26	Number						

2. Data Checks

Households Surveyed vs. Unique Rows in Data

The number of households surveyed was 1,060. This did not match the 1,059 unique household IDs in the data. There was a repetitive observation, leading to an incorrect count of household IDs.

To correct this anomaly, the duplicated observation was dropped. As a result, data related to one household was lost, which could have been missed during data collection or data entry.

Questions in the Survey vs. Columns

The number of questions in the survey (at 176) also did not match the number of variables in the data (at 178). There were two repeated columns. The repeated instances were dropped from the data set.

Then, before moving forward, we did a quick sanity check. A quick calculation for average household size (=population count/number of households) yielded 3.36 as the average, which seemed reasonable in the context. This allowed us to proceed further.

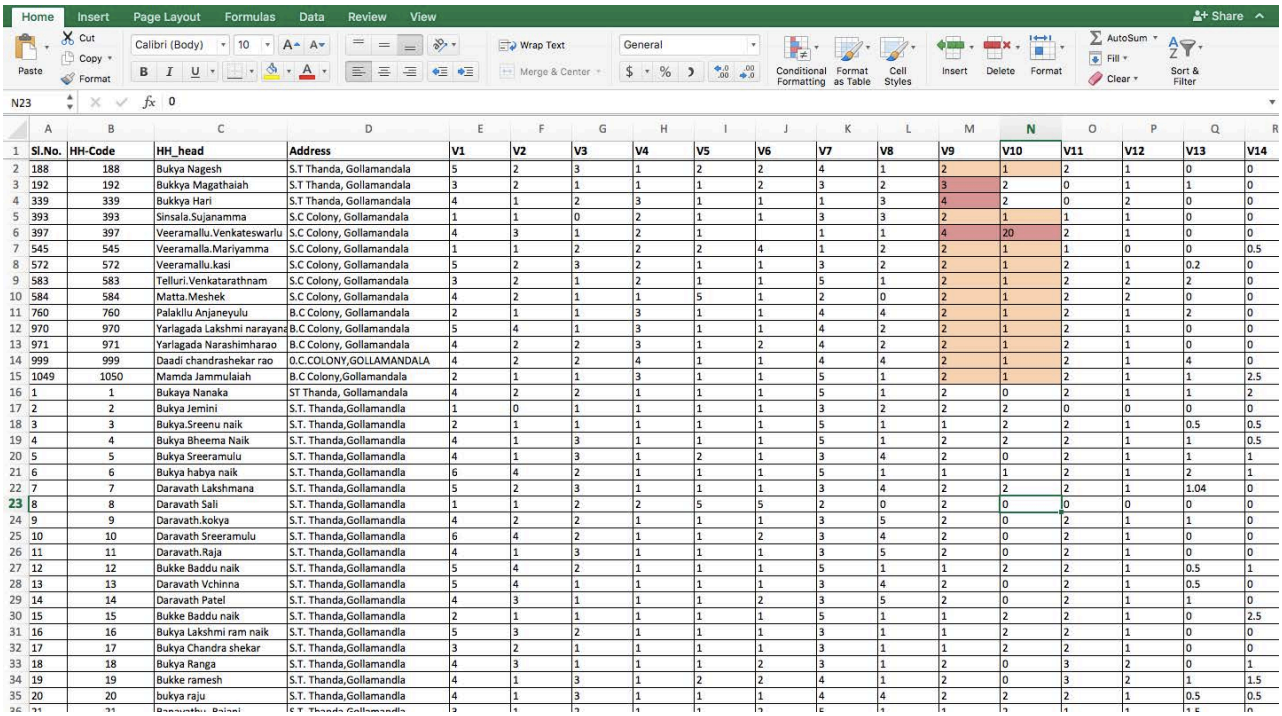
3. Question-Type Checks

The 176 columns in the data set were now checked for the data types in the responses received. Checking for conditional variables was also a part of this exercise.

- Categorical variables like “household with a toilet in the premises” could only have two possible responses — “Yes” (1) or “No” (2) — according to the questionnaire. We observed, however, that there were values such as 4, 20, “p” and so on. These

data points were deleted and treated as missing values instead.

- Integer variables like “number of cows owned by the household” had values like 1.5, 9.75, and so on. Once again, these data points were deleted and instead treated as missing values.
- Anomalies in conditional variables were identified and deleted. For example, “usage of toilet” (Question 10) — whose value depended upon the previous question, “household with a toilet within the premises” (Question 9) — can’t be “Yes” if the response to Q9 is “No”.



Sl.No.	HH-Code	HH_head	Address	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14
188	188	Bukya Nagesh	S.T Thanda, Gollamandala	5	2	3	1	2	2	4	1	2	1	2	1	0	0
192	192	Bukkya Magathaiah	S.T Thanda, Gollamandala	3	2	1	1	1	2	3	2	3	2	0	1	1	0
339	339	Bukkya Hari	S.T Thanda, Gollamandala	4	1	2	3	1	1	1	3	4	2	0	2	0	0
393	393	Sinsala.Sujanamma	S.C Colony, Gollamandala	1	1	0	2	1	1	3	3	2	1	1	1	0	0
397	397	Veeramallu.Venkateswarlu	S.C Colony, Gollamandala	4	3	1	2	1		1	1	4	20	2	1	0	0
545	545	Veeramalla.Mariyamma	S.C Colony, Gollamandala	1	1	2	2	2	4	1	2	2	1	1	0	0	0.5
572	572	Veeramallu.kasi	S.C Colony, Gollamandala	5	2	3	2	1	1	3	2	2	1	2	1	0.2	0
583	583	Telluri.Venkatarathnam	S.C Colony, Gollamandala	3	2	1	2	1	1	5	1	2	1	2	2	2	0
584	584	Matta.Meshek	S.C Colony, Gollamandala	4	2	1	1	5	1	2	0	2	1	2	2	0	0
760	760	Palaklu Anjaneyulu	B.C Colony, Gollamandala	2	1	1	3	1	1	4	4	2	1	2	1	2	0
970	970	Yarlagada Lakshmi narayana	B.C Colony, Gollamandala	5	4	1	3	1	1	4	2	2	1	2	1	0	0
971	971	Yarlagada Narashimharao	B.C Colony, Gollamandala	4	2	2	3	1	2	4	2	2	1	2	1	0	0
999	999	Daadi chandrashekar rao	O.C.COLONY,GOLLAMANDALA	4	2	2	4	1	1	4	4	2	1	2	1	4	0
1049	1050	Mamda Jammulaiah	B.C Colony,Gollamandala	2	1	1	3	1	1	5	1	2	1	2	1	1	2.5
1	1	Bukaya Nanaka	ST Thanda, Gollamandala	4	2	2	1	1	1	5	1	2	0	2	1	1	2
2	2	Bukya Jemini	S.T. Thanda,Gollamandla	1	0	1	1	1	1	3	2	2	2	0	0	0	0
3	3	Bukya.Sreenu naik	S.T. Thanda,Gollamandla	2	1	1	1	1	1	5	1	1	2	2	1	0.5	0.5
4	4	Bukya Bheema Naik	S.T. Thanda,Gollamandla	4	1	3	1	1	1	5	1	2	2	2	1	1	0.5
5	5	Bukya Sreeramulu	S.T. Thanda,Gollamandla	4	1	3	1	2	1	3	4	2	0	2	1	1	1
6	6	Bukya habya naik	S.T. Thanda,Gollamandla	6	4	2	1	1	1	5	1	1	1	2	1	2	1
7	7	Daravath Lakshmana	S.T. Thanda,Gollamandla	5	2	3	1	1	1	3	4	2	2	2	1	1.04	0
8	8	Daravath Sali	S.T. Thanda,Gollamandla	1	1	2	2	5	5	2	0	2	0	0	0	0	0
9	9	Daravath.kokya	S.T. Thanda,Gollamandla	4	2	2	1	1	1	3	5	2	0	2	1	1	0
10	10	Daravath Sreeramulu	S.T. Thanda,Gollamandla	6	4	2	1	1	2	3	4	2	0	2	1	0	0
11	11	Daravath.Raja	S.T. Thanda,Gollamandla	4	1	3	1	1	1	3	5	2	0	2	1	0	0
12	12	Bukke Baddu naik	S.T. Thanda,Gollamandla	5	4	2	1	1	1	5	1	1	2	2	1	0.5	1
13	13	Daravath Vchinna	S.T. Thanda,Gollamandla	5	4	1	1	1	1	3	4	2	0	2	1	0.5	0
14	14	Daravath Patel	S.T. Thanda,Gollamandla	4	3	1	1	1	2	3	5	2	0	2	1	1	0
15	15	Bukke Baddu naik	S.T. Thanda,Gollamandla	2	1	1	1	1	1	5	1	1	2	2	1	0	2.5
16	16	Bukya Lakshmi ram naik	S.T. Thanda,Gollamandla	5	3	2	1	1	1	3	1	1	2	2	1	0	0
17	17	Bukya Chandra shekar	S.T. Thanda,Gollamandla	3	2	1	1	1	1	3	1	1	2	2	1	0	0
18	18	Bukya Ranga	S.T. Thanda,Gollamandla	4	3	1	1	1	2	3	1	2	0	3	2	0	1
19	19	Bukke ramesh	S.T. Thanda,Gollamandla	4	1	3	1	2	2	4	1	2	0	3	2	1	1.5
20	20	bukya raju	S.T. Thanda,Gollamandla	4	1	3	1	1	1	4	4	2	2	2	1	0.5	0.5

4. Outlier Detection

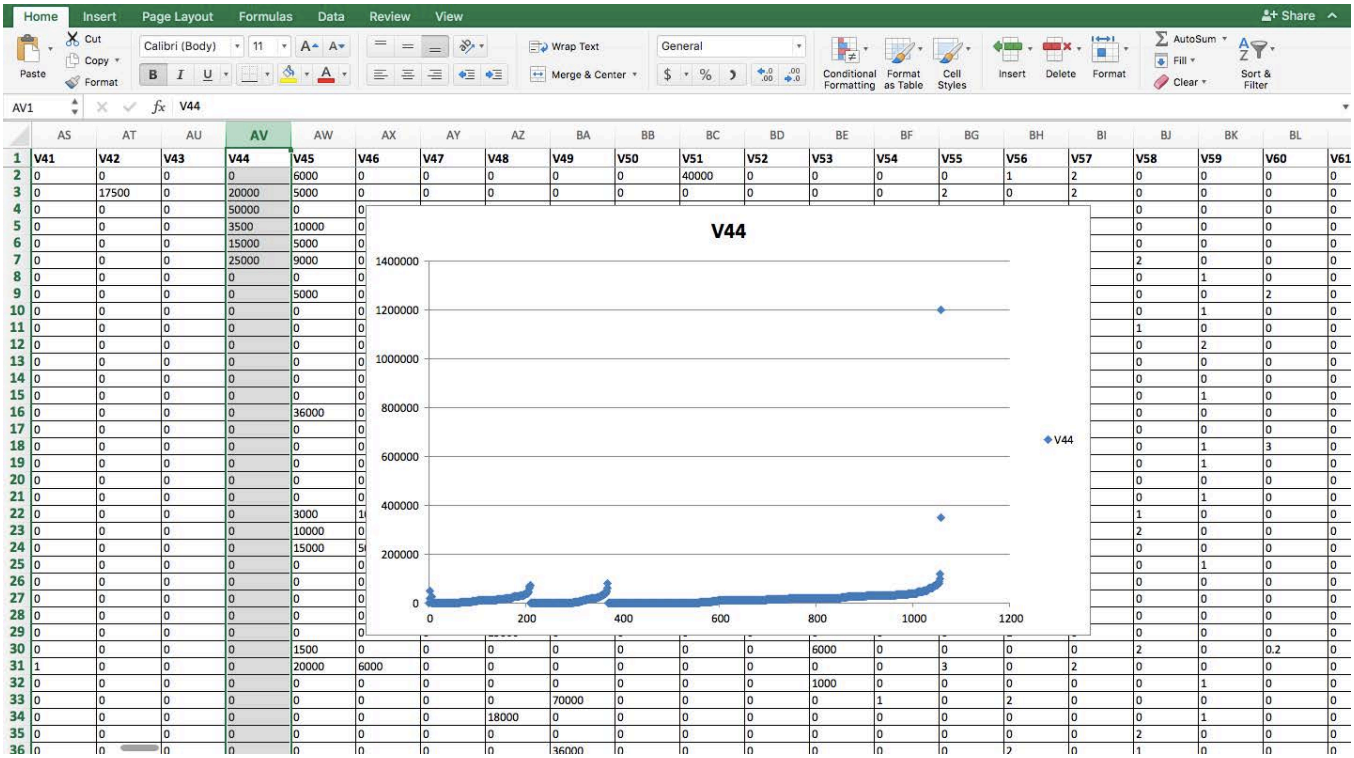
We used scatter plots and z-score tests to detect outliers in non-categorical variables such as wages and expenditure.

Scatter Plots

Scatter plots provide an easy, quick visualization of outliers. On plotting a scatter plot

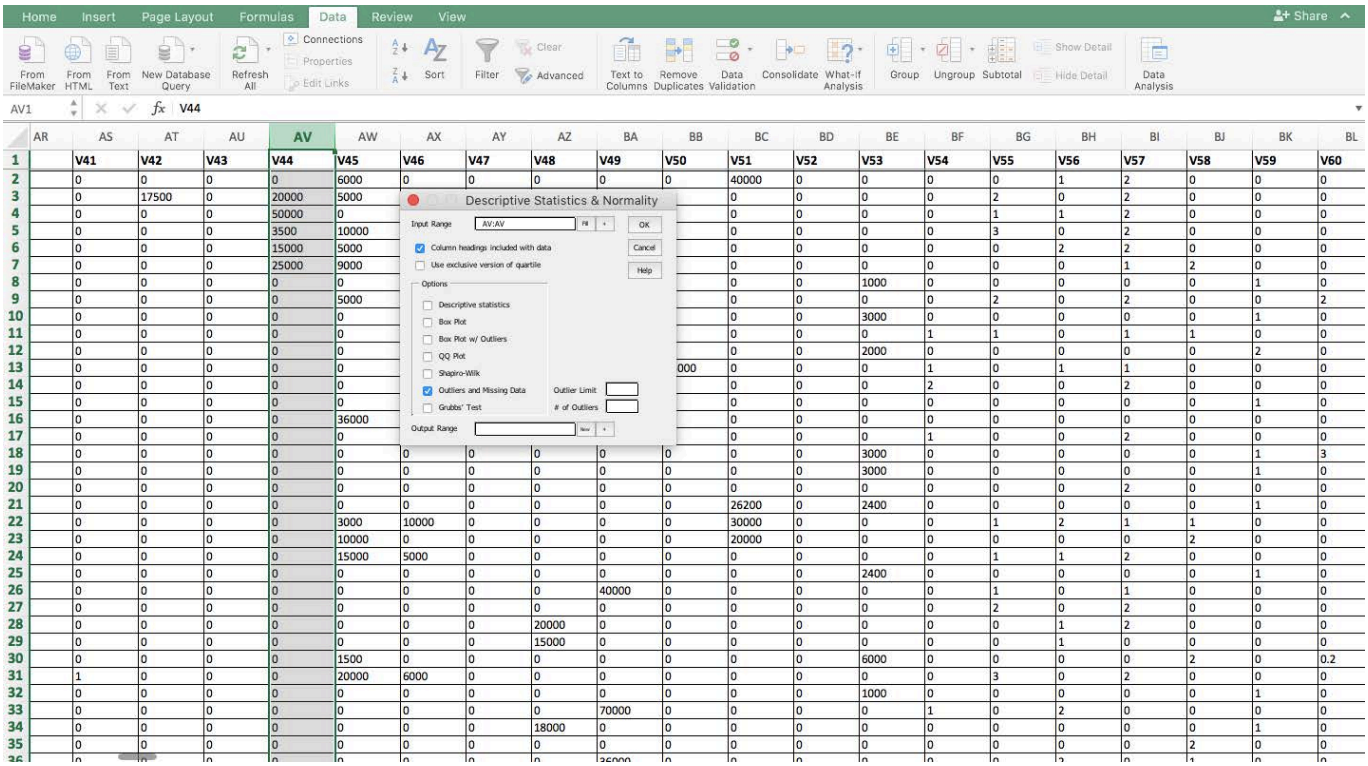
for wages, an outlier of 1,200,000 was detected. By including this one outlier, the average wage rose from 22,200 to 23,800. It was therefore crucial for us to validate the outlier to decide whether or not to include it.

To do this, we looked at associated variables such as asset ownership, toilet in the premises, education qualification and so on. We found this household lagging in these parameters, which made the high wage clearly unrealistic. We leveraged our understanding of the data set to take a rational call and delete this data point and treat it as a missing value.



Real Statistics Data Analysis Tool

The Real Statistics tool assumes the variable to be normally distributed and computes the z-score. The default z-score limit is set between -2.5 to +2.5. This can be customized too. All values beyond this z-score range are highlighted as outliers.



When applied on the wage data, the Real Statistics tool highlighted 3 cases with a z-score beyond +2.5. Once again, these statistically-proven outliers were validated or deleted by looking at other associated income and lifestyle parameters of the household.

5. Missing Data

Instances of missing data — arising from no response or deletion of bad data — were converted into one standard notation of a dot "." to avoid misinformation and inconsistency.

Takeaways

Since this was an paper-based survey, there were several gaps in data collection, leading to a fairly long data cleaning process. Finally, when this was done, we ended up with a clean and credible data set that we then plugged into our data visualization engine to equip the officers in Gollamandala with data-driven decision-making.

Exercise Solutions

Chapter 1

1. Columns should be: Name, City, Gender, Age, Discovery (or similar words).
2. The solutions are:
    a. The unit of analysis in this data set is the sales transaction. Each row corresponds to one online sales transaction.
    b. There are 998 sales transactions — hence, units of analysis — in this data set.
3. One unique ID for this data set is the combination of Name and City. For example, in this data set, there are 10 women named Lisa. But each of the women named Lisa hails from a different city. Hence the combination of Name and City would yield a Unique ID in this data set.

Chapter 2

1. The rows with serial numbers 304, 395, 692, and 814 have geo-coordinates inside the Indian Ocean.
2. The timestamps are in the format 'dd-mm-yyyy hh:mm:ss' with
    a. dd: Day
    b. mm: Month
    c. yyyy: Year
    d. hh: Hour (in 24-hour format)
    e. mm: Minute
    f. ss: Second

Chapter 3

1. The data types in each column are:
    a. Column A: numeric
    a. Column B: date (in the format 'dd-mm-yyyy hh:mm:ss')
    b. Column C: string
    c. Column D: numeric
    d. Column E: categorical (with 4 categories: Amex, Diners, Mastercard, Visa)
    e. Columns F-I: string
    f. Columns J-K: numeric
2. In the rows with serial numbers 11, 206, and 627, the ages are string values.
3. The rows with serial numbers 13 and 599 have responses longer than 12 characters.

Chapter 4

1. Since we are looking to analyze how consumption changes depending on the location of the customer:
    a. Dependent variables: Product, Price
    b. Independent variables: City, State, Country
2. 11 rows — serial numbers 29, 38, 397, 517, 624, 662, 682, 742, 878, 964, and 994 — have missing values for Product and/or Price. These rows will unable to contribute to the analysis of country-wise consumption patterns, and hence it is best to delete these rows for this analysis.
3. The mean, median, and mode are:
    a. Mean = 2,444
    b. Median = 1,800
    c. Mode = 250

Chapter 5

3b. Using Excel functions, we get:

- Median = 1,800
- Q1 = 800
- Q3 = 3,600

3c. Interquartile range = Q3-Q1 = 2,800

- The outer fence ranges from -7,600 to 12,000. Therefore, 13,000 and 47,000 are major outliers.
- The inner fence ranges from -3,400 to 7,800. Therefore, there are no minor outliers.

Chapter 6

1. 13 rows — serial numbers 75, 105, 203, 233, 271, 441, 471, 476, 494, 560, 578, 721, and 980 — have Column F populated with values other than “Advertisement”, but have a value other than “NA” in Column H. In these rows, Column H should be changed to “NA” since this conditional question should be unanswered if the “Discovery” is not equal to “Advertisement”.

Written, Edited, and Designed By:

Aarti Gupta

Apoorv Anand

Christine Garcia

Lilianna Bagnoli

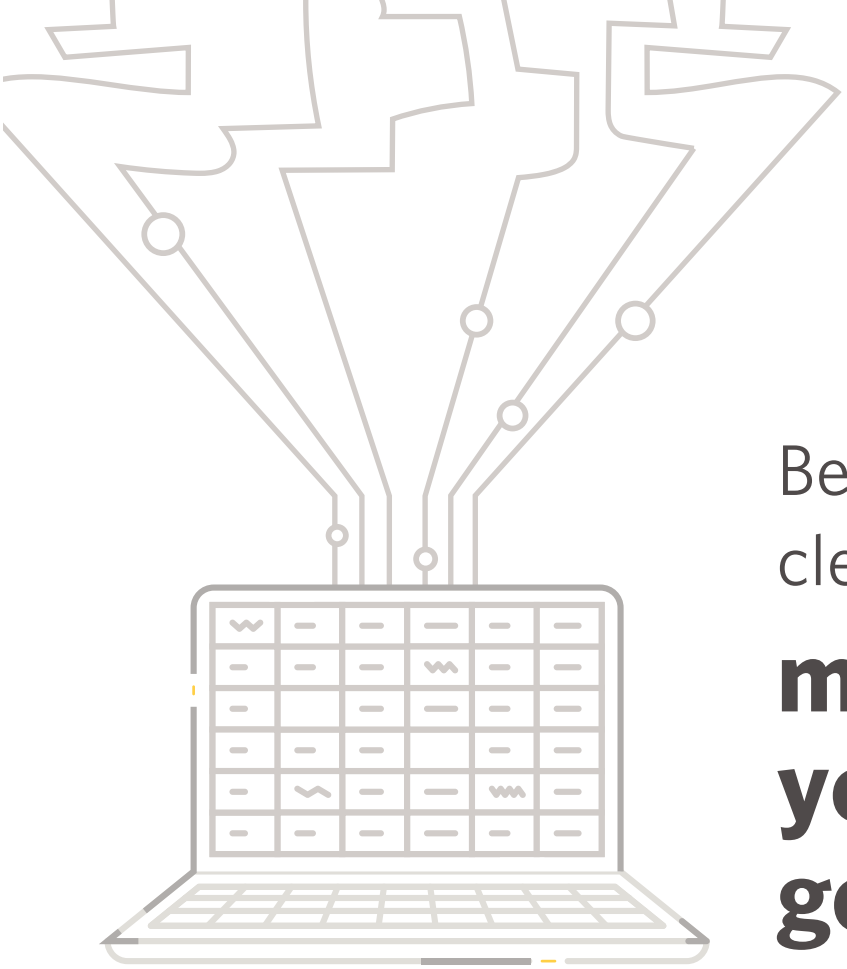
Sahaj Talwar

Shilpa Arora

Udit Poddar

Uttara Rajasekar

Vasavi Ayasomayajula



Before you even start cleaning your data,

make sure you're collecting good data

collect

The world's most **resilient** data collection tool

Try Collect Now!

Collect — our mobile data collection tool — has been used by **Google**, **Tata Trusts**, **Azim Premji Foundation** and others to collect over **20 million survey responses**.

From intuitive **skip logic and data validations** to automatic **UID creation**, Collect has everything you need to create robust surveys and collect high quality data.

About SocialCops

"SocialCops is taking big data in a direction that very few companies have been able to do: providing data and insights that can help solve real problems for most of the planet."

– ***Pankaj Jain***, Venture Partner at 500 Startups

Our world's most important decisions are crippled by an astonishing lack of data. No map in the world can tell women the safest route home or tell the police the best route to patrol. No survey tracks parameters like teacher attendance or school infrastructure in real time. National-level healthcare decisions affecting millions are made based on a sample survey of 100 people.

SocialCops is a data intelligence company that empowers organizations to make tough decisions and solve the world's most critical problems. Our platform brings the entire decision-making process — collecting primary data, accessing external data, linking internal data, cleaning and transforming data, and visualizing data — to one place. This makes it faster, more efficient, and easier to make an important decision through data.

Our goal is to take the big data revolution to where it matters the most — to use in decisions that affect human health, well being, safety and livelihoods.

Partners

We work with over 150 partners from 7 different countries to confront the world's most critical problems through data intelligence. Our partners include the Government of India, United Nations, Bill & Melinda Gates Foundation, Tata Trusts, Oxfam India, Unilever, and BASF.

Recognition

One of India's fastest growing startups, SocialCops was featured on Fortune India 40 Under 40 and Forbes Asia 30 Under 30 two years in a row. Our work has won us accolades globally — including recognition from Nascomm as one of India's top 10 startups, the United Nations World Youth Summit Award, Global Social Entrepreneurship Competition, IBM/IEEE Smart Planet Challenge and grants from Microsoft, IBM and Salesforce.

Thank You for Reading

What Did You Think?

Give us feedback and help us improve.

(It takes less than a minute.)

CLICK HERE