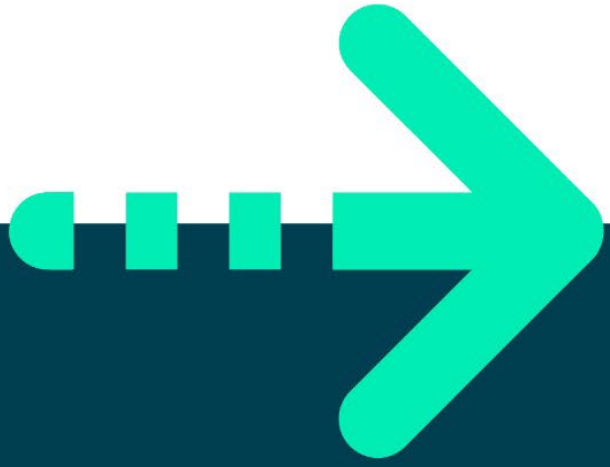
Activity 5 – Data cleaning challenge

Learner guide



CONTENTS

How to Use This Workbook

Activity

Alongside this icon you'll find details of the group/individual activity or a point for everyone to discuss.

Useful Tool

This icon indicates a technique that will help you put what you learn into practice.

Important Idea or Concept

Generally, this icon is used to draw your attention to ideas that you need to understand by this point in the course. Let your trainer know if you do not understand or see the relevance of this idea or concept.

Helpful Hint

This icon guides you to tips or hints that will help you avoid the standard pitfalls that await the unwary practitioner or to show you how you might increase your effectiveness or efficiency in practising what you have learnt.

Key Point

This icon is used to indicate something that practitioners in this field should know. It's likely to be one of the major things to remember from the course, so check you do understand these key points.

Reference Material

When we have only touched briefly on a topic this icon highlights where to look for additional information on the subject. It may also be used to draw your attention to International or National Standards or Web addresses that have interesting collections of information.

Definition

Where a word with a very specific definition (or one that could be described as jargon) is introduced this will highlight that a definition is provided. (These words will also be found in the Glossary at the back of the workbook.)

Warning

This icon is used to point out important information that may affect you and your use of the product or service in question.

Introduction

In this activity, you will transform text data from one format to another. You will begin by separating a list of people's names into separate Titles, Forenames and Surname Fields. The data set consists of 1309 passenger names from the well-known Titanic dataset.

Prerequisite knowledge

Before you begin, ensure you have read and completed the activities in the following sections located in "Guide to Microsoft Excel v#.#.#.pdf".
Italics: optional.

- Appendix 1: Text functions
    - The LEFT, RIGHT, and MID functions
    - The TRIM function
    - The LEN and FIND functions
    - The SUBSTITUTE function
    - *The TEXTJOIN function*
- Logical Functions:
    - IF and IFERROR functions on pages 66 – 70.

Resources

You will need the following files to complete this activity:

- Guide – Passenger Names (this document)
- Data Cleaning Challenge – Passenger Names.xlsx
    - Passenger Names (Worksheet 1)
    - Extra details (Worksheet 2)
    - Formula Help (Worksheet 3)
- Guide to Microsoft Excel v#.#.#.pdf".
- A MS word document to enter your answers.

Learning outcomes

- Gain additional experience with Excel formula functions. In particular: IF(), AND(), OR(), NOT(), LEFT(), RIGHT(), MID(), LEN(), TRIM(), SUBSTITUTE(), FIND() and IFERROR().
- Use formulas in a table.
- Parse and manipulate strings using Excel formulas.

Challenge details

- You will process a list of names of 1,309 passengers of the Titanic. These are in the “Name” column of the table “tblPassengerNames” on the “Data” tab.
- Each name is written in a standard format.
- You will split each name into its component parts—Title, Forenames and Surname—and place these in the corresponding output columns (orange).
- You will do this by entering Excel formulas into the “PassengerNames” worksheet, which is located on the "Data" tab. Familiarise yourself with the table before you continue."
- Table 1 below demonstrates and example of what is expected.

Sample input		Expected output		
Name	**Gender**	**Title**	**Forenames**	**Surname**
Allen, Miss. Elisabeth Walton	female	Miss.	Elisabeth Walton	Allen
Allison, Master. Hudson Trevor	male	Master.	Hudson Trevor	Allison
Allison, Miss. Helen Loraine	female	Miss.	Helen Loraine	Allison
Allison, Mr. Hudson Joshua Creighton	male	Mr.	Hudson Joshua Creighton	Allison
Allison, Mrs. Hudson J C (Bessie Waldo Daniels)	female	Mrs.	Bessie Waldo	Allison
Bradley, Mr. George ('George Arthur Brayton')	male	Mr.	George	Bradley

Table 1: Example sample input and expected output.

Task 1: Surname and title

In this section, you will complete the surname and title of the Passenger Name table.

Guided activity:

FIND Function

1. Open the **Data Cleaning Challenge – Passenger Names.xlsx**
2. Navigate to the **“Passenger Names”** worksheet.
3. To get started, find the first row of the **“CommaPos”** (comma position) column of the **Passenger Name** table and enter the formula **=FIND(“,”, [@Name])** . Click enter. The formula should automatically copy itself downwards to all cells in that column. If not, double click the bottom right corner of the cell to populate the rest of the cells:

Independent activity

FIND, LEFT, and MID functions

Reference material

Refer to the ‘Formula Help’ Worksheet and ‘Guide to Microsoft Excel’ pdf for help with the FIND(), LEFT(), and MID() functions.

1. Explain what the **FIND()**formula function does.
2. We can now use our “CommaPos” column to parse out the “Surname” field.
3. In the first row of the “Surname” column (cell V2), enter a formula that returns a substring of the “Name” column , starting the first character and ending one character before the first comma.
4. You will need to use the **LEFT()** function.
5. In the first row of the “DotPos” column (cell E2), enter a formula that returns the position of the first “.” Character in the “Name” column.
6. Explain what the **MID()** formula function does.
7. We have now made sufficient pre-calculations to parse out the “Title” field.
8. In the “Title” column (cell T2), enter a formula that returns a substring of the “Name” column, starting with the first character of the title, after the first comma and white space and ending at the first full stop. (The full stop should be included). You will need to use the MID() function.

Task 2: Conventional forenames

Independent activity:

LEN, IF, and IFERROR Functions

1. In the "OpenParPos" (Open Parenthesis Position) and "CloseParPos" (Close Parenthesis Position) columns (columns F and G respectively), use the FIND function to enter a formula that returns the position of the first "(" character and first ")" character, respectively.

Helpful Hint

Many of the names do not contain parenthesis characters; in these cases, the FIND () function will return **#VALUE!.** This is expected.

2. Explain what the **LEN()** formula function does.
3. Explain what the **IF()** formula function does.
4. Explain what the **IFERROR()** formula function does.

According to antiquated convention of etiquette, a married woman was sometimes referred to by her husband's name. For example, the wife of Mr. Hudson J C Allison might be formally referred to as 'Mrs. Hudson J C Allison', even though her given forename is Bessie.

What are conventional forenames?

- Over a hundred years ago, when a woman was married, she would take her conventional forenames, which were her **husband's forenames**. For examples, *Bessie Waldo Allison's* conventional forenames are *'Hudson J C'*.
- For everybody else, that person's conventional forenames are simply their **given forenames.** For example, *Miss. Elisabeth Walton Allen's* conventional forenames are *'Elisabeth Walton'*.

5. In the 'EndOfConvFNPos' (end of conventional forenames position) column, enter a formula that returns the position of the last character of each passenger's conventional forenames. For example, the 'EndOfConvFNPos' value for 'Allison, Mrs. Hudson J(Bessie Waldo Daniels)' should be 24, corresponding to the position two places before the open parenthesis character.

Helpful Hint

You may find it helpful to use the IF() function or IFERROR() function.

Helpful Hint

If a person's name contains parentheses, their conventional forenames end just before the opening parenthesis.

If a person's name does not contain parentheses, where do their conventional forenames end?

6. In the **“ConvForenames”** (cell I2) column, enter a formula that returns the conventional forenames of the passenger. You will need to use the MID() function. For example, the “ConvForenames” value for “Appleton, Mrs. Edward Dale (Charlotte Lamson)” should be “Edward Dale”.

Helpful Hint

Make use of the "EndOfConvFNPos" column, which you computed in step 5.

Reference Material

Refer to the “Extra details” for guidance on the names.

Task 3: Extracting the contents of the parentheses

Independent activity:

LEFT, AND, and RIGHT Functions

1. In the ‘InParentheses’ column (cell J2), enter a formula that returns everything between the parentheses in the passenger’s name, or #VALUE! If the passenger name does not contain parentheses. Do not include the parentheses themselves – just the text between them. For example, the ‘InParentheses’ value for ‘Appleton, Mrs. Edward Dale (Charlotte Lamson)’ should be ‘Charlotte Lamson’.

Helpful Hint

You have already found and recorded the positions of the opening and closing parenthesis characters.

You will need to use the MID() function.

Helpful Hint

For most formula functions, if one of the inputs is an error value (such as #VALUE!), the output will be an error value of the same type.

2. In the “InParenthesesCleaned” column (cell K2), enter a formula that returns the same as “InParentheses”, except that it returns #VALUE! if the string in parentheses is enclosed in single quotes. This step helps remove some aliases and nicknames. For example, the “InParenthesesCleaned” value for "Appleton, Mrs. Edward Dale (Charlotte Lamson)" should be "Charlotte Lamson", and the “InParenthesesCleaned" value for "'Bradley, Mr. George ('George Arthur Brayton')" should be #VALUE!.

Helpful Hint

Consider the IF(), AND(), LEFT() and RIGHT() functions.

Reference Material

Refer to the ‘Formula Help’ worksheet and ‘Guide to Microsoft Excel’ pdf’ for guidance.

Task 4: Extracting the forenames of married women

Independent activity

SUBSTITUTE Function

1. Explain what the **SUBSTITUTE()** formula function does.
2. We have seen that the forenames and maiden name of a married woman are enclosed in parentheses, and typically adhere to the following format: **(<Forename 1> <Forename 2> … <Forename N> <Maiden name>)**
3. We would like to extract the forenames and **discard the maiden name**. For married women, the maiden name may be assumed to be the **last name in the parentheses**. To extract everything in the parentheses except the last name, we are going to implement this technique step by step:
    a. For the first step, in the 'P_NSpaces' column (cell L2), enter a formula that counts the number of spaces in the value of the 'InParenthesesCleaned' column. (Here, 'P' is short for 'parentheses').
    b. If the 'InParenthesesCleaned' value is #VALUE!, your 'P_NSpaces' formula should return #VALUE! also. For example, the 'P_NSpaces' value for "Bessie Waldo Daniels" should be 2, since there are two spaces in that string. The 'P_NSpaces' value for 'Farnham' should be 0, since there are no spaces in that string.

Helpful Hint

You will need to use the **SUBSTITUTE()** function.
You will need to use the **LEN()** function twice.
Use SUBSTITUTE() to replace " " with "" (space with empty string) in the 'InParenthesesCleaned' column. How many characters does this shorten the string by?

Reference Material

Refer to this weblink to help you: https://trumpexcel.com/find-characters-last-position/ as well as the 'Guide to Microsoft Excel' pdf and 'Extra details' tab of the worksheet.

4. In the 'P_FinalSpacePos' column (cell M2), enter a formula that returns the position of the final space character in the value of the 'InParenthesesCleaned' column. If the 'InParenthesesCleaned' value is #VALUE!, your 'P_FinalSpacePos' formula should return #VALUE! also.

Helpful Hint

You will need to use **FIND()** and **SUBSTITUTE()** functions.

Use **SUBSTITUTE()** to replace the final space character with some uncommon character, such as '@'. You will need to make use of the optional third parameter, **'instance_num' or 'Instance'**.

Use **FIND()** to return the position of this uncommon character.

Helpful Hint

The **'instance_num'** parameter of the **SUBSTITUTE()** function allows us to substitute a **single instance** of a found substring.
We need to set **'instance_num'** to the number of spaces in the "InParenthesesCleaned" value.

Reference Material

Refer to section 'The SUBSTITUTE function', page 208 - 209 in the 'Guide to Microsoft Excel' pdf for guidance.

5. In the 'P_Forenames' column, enter a formula that returns all the text in the value of the 'InParenthesesCleaned' column, except the last word. For example, if the 'InParenthesesCleaned' column contains 'Bessie Waldo Daniels', 'P_Forenames' should return 'Bessie Waldo'.

Helpful Hint

In your formula, you will need to use the **IF()** and **LEFT()** function and reference the 'P_FinalSpacePos' column.

6. You are now ready to populate the final column, "Forenames". In the "Forenames" column, compose a formula that returns a passenger's forenames.

Helpful Hint

You will need to use the **IFERROR()** function and reference the 'ConvForenames' and 'P_Forenames' columns.