بسم الله الرحمن الرحيم

IBM Data Analysis Professional Certificate

Course No. 2

**Excel Basics for Data Analysis**

**(Final Assignment)**

**Content:**

* Overview.
* Part 1: Clean and Prepare Data for Analysis.
* Part 2: Analyze the Data.

**Written by:**

**Abobaker Ahmed Khidir Hassan**

Overview

About the Course

This course was so useful and helpful using Excel in the basics task of Data Analysis Process, such as:

Dealing with Excel workbooks and translating on it.

Extracting data from different sources.

Basics of Data Privacy.

Transforming data from the row data to a ready data for analysis including cleaning data using the basic functions and Excel feature for preparing data.

Filtering and Sorting Data in Excel.

Basic Functions in Excel useful in Data Analysis.

Basic analysis focusing in the Descriptive analysis using basic functions that useful in Data Analysis, and using Pivot Tables.

Assignment Scenario

In this final assigment, you will be following the scenario of a recently hired Junior Data Analyst in a local government office, who has been tasked with importing some data from another department which relates to inventory information about their fleet of vehicles. The data is in comma-separated value (CSV) format and the data also needs cleaning up before you can start to run any kind of analysis on it, and I will explain each step in this process.

Data

The original data in this assignment consist of 4 columns and 63 rows/records.

Columns are:

1. Department (String / General).
2. Department (String / General).
3. Equipment Class (String / General).
4. Equipment Count (Discrete Numbers).

We note that the first two columns should be in a single column, and there are blanks and duplicated rows. And that what we will deal with in the first part of the assignment.

Cleaning and Prepare Date

Extracting Phase

Task 1: Importing Data

In this step, I extract the data from a (.CSV) file that is become with the material into an Excel Workbook to get started dealing with it.

There are many ways to do that. I opened a new Excel Workbook and go to Data -> Get Data -> From File -> From CSV/Text file.

I have late the delimiter as it (Comma), and choose to set the data types manually for three reasons:

1. To make sure there is no any wrong setting these types for more quality cleaning.
2. The dataset is very small.
3. To make good practice dealing with different formats.

I choose “Use first row as headers”, because the data has a header row.

Transformation Phase

Task 2: Column Widths

In this step, I resized the column widths to have a good overview of the content and to start dealing with the data.

I do it using selecting the column A, then press (Ctrl + Shift + the Right erro), thin double click any column’s separator.

Task 3: Empty Rows

In this step, I removed any empty rows (blanks) to make sure these rows will not be counted as zeroes in the analysis phase.

The easier way to do that, is just to filter the data by any primary field that must hold a content in each row, then if any non-empty rows become out, you can filter it again up to have only the empty rows, then select all of them and delete them and clear the filter.

Task 4: Duplicate Records

In this step, I removed any duplicated rows, to ensure the analysis phase will be performed only on the real data.

There are two ways to do that, and the easier way is to go to the Data menu, in the Data Tools field press “Remove Duplicates”, and select all columns.

Task 5: Spelling

In this step, I cleaned any syntax error to make sure the content of any categorical field will not be affected by these errors, and the final result will be well spelling.

To perform that go to first cell in the worksheet (ctrl + home), and go to Review menu and select Spelling, you have to review any change because sometimes you Excel will not hold your words as it, (especially if you have your own shorts).

Task 6: Whitespaces

In this step, I removed any whitespaces that is the wrong double spaces between words, and the space before or after the content of any cell, especially string (General) field format.

To do that, go to the first cell in the worksheet as in the previous task, then go to the home menu and select “Find and Replace” from the Editing field, or just use the shortcut (ctrl + h), and replace double space with one space.

And to remove spaces after and before the real content, I used TRIM function in a temporary column and then paste it in the real column using “Paste Value” option.

Task 7: Department names

The scenario assumes that when the data was converted from its data source, the department names (see correct list below) didn’t import correctly and they are now split over two columns in the data and I should deal with this problem.

There is more than option to do that like using any formula but I simply used Flash Fill to reduce the department names to just one column, and then remove unnecessary columns.

Loading Phase

Task 8: Uploading the Ready Data in a Repository

After finishing cleaning and preparing data, it is ready now to get insights from it.

I uploaded this work into a GitHub repository following these steps:

1. Opening my CMD and creating a rsa public key to translate the data securely in the internet.
2. Opening my GitHub account and defined that key.
3. From the command line I goes to the path of the folder that holds the project and set it as a git repository.
4. Creating a new repository in my GitHub account.
5. Uploading my work in that repository using the SSH link.

**The End**