Practise Data Cleansing in Excel (Answers)

Worksheet section	Contents
1	Renaming and duplicates
2	Handling missing and outlying values

Version: 1.0

This lesson has been created by Effini in partnership with Data Education in Schools, The Data Lab and Data Skills for Work, with funding from the Scottish Government.

If you require this document in an alternative format, such as large print or a coloured background, please contact

hello@effini.com

or

4th Floor, The Bayes Centre
47 Potterrow
Edinburgh
EH8 9BT

1. Renaming and duplicates

Section 1.1

In this workbook we are going to use a dataset called music_messy.

In the lesson 'Practise Data Understanding in Excel' we discovered the following about 'music_messy' dataset,

Number of rows:	647
Number of columns:	10
Format of the dataset:	Wide
Outliers:	BPM has outliers of -20 and 2969
Number of empty rows:	3
Variables with high % of missing values:	decade (65%) and loudness (94%)

column	data type	definition
title	string	The title of the song
Artist	string	Singer or Band
genre	string	Genre of the song
TheYearTheSongWasReleased	integer	Release year of the song
BPM	integer	Beats per minute (tempo)
energy	float	Value between 0 and 1 - the higher the value, the more energetic the song
danceability	float	Value between 0 and 1 - the higher the value, the easier it is to dance to
duration	integer	Duration of the song in seconds
decade	integer	Decade in which this song was published
loudness	float	The higher the value, the louder the song is.

1. Renaming and duplicates

1) In this stage of data cleansing you need to rename the variables to make them easier to use. Follow the steps below to rename the variables in **music_messy** dataset.

a) What naming convention are you going to use? snake case

b) Fill in the boxes below with the new column name you are going to use. (you might not need to change them all)

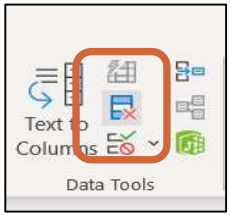
c) Change the column names in the **music_messy** dataset to match the new column names below.

column	new_column
title	title
Artist	artist
genre	genre
TheYearTheSongWasReleased	release_year
BPM	bpm
energy	energy
danceability	danceability
duration	duration
decade	decade
loudness	loudness

2) The next stage is to remove any duplicates from the dataset.

If you have any problems completing question 1, you can use the dataset in music_messy_rename in this question.

Before you run the deduplicating, how rows do you have in your dataset? 647



Using the deduplicating button, remove the duplicates from the dataset.

How many rows have been removed? 6

How rows does your dataset have now? 641

2. Handling missing/outlying values

Now we have renamed and remove duplicates, we are going to cleanse the missing/outlying values in the **music_messy** dataset.

If you have had any issues with the questions in section 1, you can use the dataset in tab ***music_messy_dedup*** *for these questions.*

Section 2.1

First we are going to look at the rows that contain the outliers in the **bpm** variable.

1) Turn on the filters in your dataset (reminder: crtl+shift+L)

2) Filter your dataset to find the row that contains bpm = -20

3) What do you think you should do with this outlier? Replace, remove or leave as it is

Remove, as you have no way to replace these values with accurate values, the best option is to drop the rows where bpm is -20 or 2969

4) If you think it should be removed, right click on the row and select 'Delete Row'

5) Now, filter the dataset for bpm = 2969 and delete the row.

Section 2.2

Now we are going to cleanse the variable **loudness** which has 94% missing values.

6) Given the amount of missing values in the variable **loudness** do you think you should replace, remove or leave the column as it is?

The best option for this variable is to remove it as:
1) there is no way to replace these missing values, and
2) with only a small number of non-missing values, the variable doesn't provide enough useful information.

7) If you think you should remove the variable. Right click on any cell in the **loudness** column and select Delete.

2. Handling missing/outlying values

Section 2.3

Next we are going to look at the variable **decade**, which is 65% missing. However we have the release year. So we can replace the missing values with a calculated value.

8) You need to create a new column of data next to **decade** variable. Type '**decade_calc**' in cell J1 of the tab with your dataset you are using.

9) Type the formula below into cell J2. This formula will calculate the decade a year is in.

FLOOR(D2/10,1)*10

10) Copy the formula you have just typed into cell J2 into all the other cells in that column.

If you want to know more about the FLOOR function, please see,
https://support.microsoft.com/en-us/office/floor-function

Section 2.4

The last task is to remove the rows that only contain missing values. When you ran the deduplicating function, Excel has left just one row with all missing values.

If you have had any issues with the questions up to this point, you can use the dataset in tab ***music_messy_missing*** *for these questions.*

11) If not already on, turn on the filters your dataset (reminder: crtl+shift+L)

12) Filter your dataset to find the row that contains all missing values (hint: try filtering energy = NA)

13) Right click on the row and select 'Delete Row'

14) Remove the filter from your dataset to see all the data.

It is best practise to check the number of rows matches what you expect when you delete/drop rows.

2. Handling missing/outlying values

15) Does the number of rows in your dataset match the figures below. Use counta() to check.

Number of rows:	638
Number of columns:	10

Congratulations! You have now cleansed a dataset.