Data Challenge - Corrupted Data

Young Explorer, welcome to the first game of the Data Challenge! As you are starting your Data Science journey, in this game, you are supposed to identify and correct different data anomalies and irregularities in a dataset. These anomalies and irregularities are keys that you need to collect in order to leave the room. If you are ready for the challenge, please enter the magic chamber!

Are you ready to enter the magic chamber? *

- (●) Already entering...

Okay, we see the dataset here! Hmmm, wait... It looks like we have two datasets, not one. Well, let's explore them!

What is the most popular genre of the first dataset? *

Childrens

How many reviews are there in the second dataset? *

920

Review

There is a mistake in your answer to the first question. Try again!

First, you can create a list of all genres with the following code:
genre_list = df['genre'].str.split(', ').explode().tolist()

Then, you can count mentions of each genre:

genre_counts = pd.Series(genre_list).value_counts()

Finally, you can print your results and see the most popular genre:
print(genre_counts)

What is the most popular genre of the first dataset? *

Choose

Data Challenge - Corrupted Data

Good job!

Do you think that the dataset looks suspicious? *

- (●) Seems to be fine so far
- ( ) I can clearly see some irregularities

Are you sure?

Are you sure?

Try to print the names of ten most popular book genres. Can you see some misspellings? *

- ○ Still cannot see anything...
- ◉ True, I can clearly see some irregularities now.

Please print ten most popular book genres!

Please print ten most popular book genres! You can do it with the following function:

genre_counts.head(10)

Can you see something wrong there?

Wait a second... *

- Owww, Gotcha now!
- Still do not see anything :(

Data Challenge - Corrupted Data

Exactly! There are some genres which are supposed to be the same, but they are written differently. For this reason, we are getting incorrect counts.

These two datasets are corrupted, so we need to find all irregularities and correct them. *

- (●) Let's Do the Job
- ( ) Looks scary, I am out!

Looks scary, I am out!

Very sad that you decided to give up so early. Come back once you are ready for the challenge!

Let's Do the Job

Great attitude!

Usually, there are several general issues you can encounter when working with datasets:

- Missing Values
- Outliers
- Duplicated rows/values
- Misspellings/wrong inputs
- Inconsistent units of measurement

Here, you will have to check each one of these issues in our datasets and correct those that are present.

Choose where do you want to start :) *

- (●) Missing Values
- ( ) Outliers
- ( ) Duplicated rows/values
- ( ) Misspellings/wrong inputs
- ( ) Inconsistent units of measurement

Let's Do the Job: Continue

Choose where do you want to continue :)

- ( ) Missing Values
- ( ) Outliers
- ( ) Duplicated rows/values
- ( ) Misspellings/wrong inputs
- (●) Inconsistent units of measurement

Choose this option once you finished all previous exercises.

- (●) I'm done, so let's proceed!

I'm done!

Let's check if you get it correct ;)

What is the shape of the first and the second datasets after all the transformations? *

Put 4 numbers: rows in df1, columns in df1, rows in df2, columns in df2.

Please pay attention to the order! You reply should be something like: 110, 10, 930, 12

100, 8, 920, 10

Missing Values

Check if the first dataset has missing values!

Which columns contain missing values? *

- ○ There are no missing values
- ○ book title, book price, rating, author
- ○ rating, year of publication, genre, url
- ◉ book price, rating, genre, url
- ○ book title, book price, year of publication, url

Missing Values: Correct!

Perfect! The key is collected. You can also check and see that there are no missing values in the second dataset.

Good practice is to infer missing values (NAs). There are different techniques for this, such as taking the average value or taking values from the most similar observations (in our case - books).

Let's follow the following strategy to deal with NAs:
1. Drop all observations with ranking over 100 (these are fake observations which contain NAs across all 4 columns)
2. For other NAs (specifically - in a 'rating' column), please use mean values

Once you used this strategy to deal with NAs, are there any columns with NAs left? If no, write 'NO', otherwise you will see a hint. *

You can check it with df.isna().sum()

NO

Missing Values: Incorrect!

Not quite correct! Try again.

You can use the following function to see which columns have the missing values:

df.isna().sum()

Which columns contain missing values? *

- There are no missing values
- book title, book price, rating, author
- rating, year of publication, genre, url
- book price, rating, genre, url
- book title, book price, year of publication, url

Outliers

Outliers are data points that deviate too much from the main data. Here, try to see if there are some outliers in quantitative columns, for instance, rating in the first dataset.

How many data points in column 'rating' of the first dataset can be considered as outliers? (Please enter a number) *

2

Outliers: Correct!

Yes! Boxplot shows that 2 data points in the 'rating' column can be considered as outliers: they have 4.2 and 4.1 start, while the median rate is around 4.7 euros. You can similarly try to check different columns for outliers. Since we are not modelling this data, outliers are not so important for us, but generally, they can cause serious problems by shifting the distribution of values!

The key is collected!

Duplicated rows/values

Check if the data has duplicates -- rows which are completely the same.

How many rows are duplicates in the first and the second dataset? Please enter the sum of these two numbers. *

- ○ 4
- ◉ 5
- ○ 6
- ○ 7
- ○ 8
- ○ 9
- ○ 10

Duplicated rows/values: Correct!

Great job! The key is collected.

Please remove all duplicates from the datasets.

You can do this with the following function:
df.drop_duplicates()

You can also add '.reset_index()', to make sure that the indexing is consistent.

Duplicated rows/values: Incorrect!

Something is wrong! Please try again.

You can use the following function to check the number of duplicates:

df.duplicated().sum()

How many rows are duplicates in the first and the second dataset? Please enter the SUM of these two numbers. *

- ○ 4
- ○ 5
- ○ 6
- ○ 7
- ○ 8
- ◉ 9
- ○ 10

Misspellings/wrong inputs

You have already noticed before that genre names have some irregularities: for example, the same words start either with a capital letter or they are fully lowercase. Now, you should correct these irregularities in the first dataset.

One solution can be to make all genre names lowercase, so that these differences can be removed.

Once correcting the errors in genre names, what are the top-5 most popular genres across the whole dataset? *

- ○ childrens, nonfiction, picture books, fiction, fantasy
- ◉ childrens, fiction, nonfiction, picture books, fantasy
- ○ childrens, picture books, fantasy, fiction, nonfiction
- ○ childrens, fiction, picture books, fantasy, nonfiction
- ○ childrens, fantasy, nonfiction, fiction, picture books

Misspellings/wrong inputs: Correct!

Correct, it seems you have corrected all misspellings! Now you can also see real counts of different genre. The key is collected!

Misspellings/wrong inputs: Incorrect!

Not quite. Please check genre names again!

For simplicity, you can just lowercase all genre names. You can use the following command for this:

df['genre'].str.lower()

As for counting the values and seeing top-5 genres, you have already did it in the very beginning! You can use the same code.

Once correcting the errors in genre names, what are the top-5 most popular genres across the whole dataset? *

- childrens, nonfiction, picture books, fiction, fantasy
- childrens, fiction, nonfiction, fantasy, picture books
- childrens, picture books, fantasy, fiction, nonfiction
- childrens, fiction, picture books, fantasy, nonfiction
- childrens, fantasy, nonfiction, fiction, picture books

Inconsistent units of measurement

Sometimes it is possible to notice errors in the units of measurement. For example, prices of books can be measured not in USD but in EUR, or numeric values can be of different types. There are no measurement inconsistencies, but there are some related to data types.

Try to complete the following tasks!

What is the mean price of books from the first dataset in USD?
Please type the number up to 2 digits after decimal point (for instance. 22.33)

12.71

What is the MEDIAN rate of all reviews from the second dataset? (with one decimal number)

5.0

Click here only if you have difficulties with completing any of these two tasks.

- Too difficult!

Inconsistent units of measurement: Correct!

Exactly! The key is collected.

Inconsistent units of measurement: Incorrect!

No worries, let's see if I can help you!

You need to transform the column type from string to a numeric. For this, you can use the following function:

df.loc[:, '*column name*'] = pd.to_numeric(df.loc[:, '*column name*'])

For taking the mean, you can use the following function:

df['*column name*'].mean().round(2)

The last part ensures that you take 2 digits after the decimal point.

For taking the median, you can use the following function:

df['*column name*'].median()

Please try again now:

What is the mean price of books from the first dataset in USD? *
Please type the number up to 2 digits after decimal point (for instance. 22.33)

What is the median rate of all reviews from the second dataset? (with one decimal number) *

Data is Corrected!

Bravo, you successfully managed to correct all issues in these datasets and collect all 5 keys! The last thing left: let's merge the datasets (and keys), so that we could have one big dataset of books and their reviews (open the magic chest with data :)).

Which function from pandas library will you use to combine these two datasets? Write its name in lowercase. *

merge

The Chest is Open!

The chest is open, and you have the merged dataset. Congrats, the Corrupted Data challenge is completed now, and you can leave the room!

Google Forms