Project 1 - Data Wrangling

Feb 15th, 2020

SUBMISSION DEADLINE: GMT+0 23:59:59 Feb 17th, 2020

OVERVIEW

Apply concepts learned throughout the week to the shared datasets

1. Pew research dataset

This dataset consists of religions and the number of candidates sampled in a research survey,

Click here to download the dataset

Goals

1. Rename the columns
2. Put in this dataset in a tidy format.
3. The final output should have the following **columns** and **datatypes**:
    a. **religion**, **string**(pandas represents strings as 'Objects' with a data type of 'O')
    b. income_range, **string**
    c. frequency, integer

NB: Categorical data types have an order. This order needs to be defined. Pandas can set a datatype to categorical but will use a default order which is usually not correct.

2. Billboard weekly music rank dataset

This dataset is in an untidy format and the goal is to reshape this dataset.

Click here to download the dataset

Goals

1. Prepare this data set for analysis :).
2. Your final dataframe should have the following columns and datatype:
    a. **artiste**, **string**
    b. **track**, **string**
    c. **genre**, **string**
    d. **date.entered**, **datetime**
    e. **date.peaked**, **datetime**
    f. **week**, **integer;** use the following code snippet to extract the week number from the text(assuming your dataframe is called melted):

```
melted['week'] = melted['week'].str.extract('(\d+)').astype(int)
```

    g. **rank**, **float.** Ideally we should be using an int data type for the rank but we have nulls in that column. The Pandas library does not allow you to have an integer column with nulls.
    h. **entered_to_peak, int**
3. Answer the following question using the dataset: **Who and what music ranked 1 in Rock after peaking 35 days from entering into the competition.**

HINT: when 2 dates are subtracted, pandas attaches the days eg: '105 days'.

Assuming your dataframe is called df and the column with the difference of the dates is called 'difference', you can extract the integer number of days as follows:

```
df['difference'].dt.days
```