Individual Delivery

22/12/2020 to 10/01/2021

—

Ariadna Puigventós Mendoza
Data Science
The Bridge

General Vision

If I don't find any dataset about my original idea about Smart City.

I wanted to confirm if the African Athletes won all running proves in over the world. I have started with Best World Marathons.

I could extend this project about if the marathons increase the temperatures in the cities where performance these races. Or for example, I would like to know if brand shoes (Nike or Adidas) have incresed their sales thanks to the winners.

Goals

My goals it's getting Option A.

If I have time it would be to get to create a pull request in the option A+.

Specifications

I tried to do all my functions and bucles in modules and I have imported each operation.

I have detected some outliers in my variable of the dataframe.

I have checked which values are most repeated in my dataset.

Software

- Code Visual Studio
- Google Search
- Python Tutor
- Numbers (another version excel)
- Pages (another version word)
- Photoshop
- Canvas

Hardware

- MacBook Pro 8GB

Requirements

Before starting to project I need to organize my folders and files, because I needed a space to work my notes and my fail and test code.

Set Up:

1. I worked with different tests files in notebook/ folder where there are different operations, correct and fail code.
2. When the code returned correctly, it was transferred and copied in each module appropriate with its function in utils/ folders.
3. Imported every function to main.ipynb file in src/ folder.

Steps

I. Research the context

I was searching some datasets about Smart City but I didn't find 2 datasets for 2 variables about my hypothesis. In addition, I lost 3 days about this task, then my second idea was about running (I like this hobbie).

I found some articles about one and unique idea that a lot of people have got in their minds: "The Kenyan runners always win the marathons".

There are a lot of reasons because they win the marathons or athletic proves.

II. Get Data

Main dataset is about 6 Best Marathons Majors by Countries from Kaggle and I download a csv format in local file.

So my hypothesis is:

"The Best World Marathons have been won by African Athletes."

I got another datasets like 120 Olympic Games and Altitude by Country in Kaggle y data.world, respectively. But I don't have much time to do a data wrangling to get more information.

III. Data Wrangling

In the main dataset there were not any duplicates but it detected some outliers in variable “Time”. It was necessary to change of Time Column from object type to float64 with method “to_timedelta64[s]”.

It applied different methods like **head(), tail() and most repeated values** to know the dimension.

```
   year              winner gender   country      time marathon
0  2014      Dennis Kimetto   Male     Kenya  02:02:57   Berlin
1  2011      Geoffrey Mutai   Male     Kenya  02:03:02   Boston
2  2016     Kenenisa Bekele   Male  Ethiopia  02:03:03   Berlin
3  2016      Eliud Kipchoge   Male     Kenya  02:03:05   London
4  2013      Wilson Kipsang   Male     Kenya  02:03:23   Berlin
5  2017      Eliud Kipchoge   Male     Kenya  02:03:32   Berlin
6  2011     Patrick Musyoki   Male     Kenya  02:03:38   Berlin
7  2013      Dennis Kimetto   Male     Kenya  02:03:45  Chicago
8  2017      Wilson Kipsang   Male     Kenya  02:03:58    Tokyo
9  2008  Haile Gebrselassie   Male  Ethiopia  02:03:59   Berlin
......
     year           winner  gender        country      time marathon
531  1966       Bobbi Gibb  Female  United States  03:21:40   Boston
532  1974  Jutta von Haase  Female        Germany  03:22:01   Berlin
533  1969  Sara Mae Berman  Female  United States  03:22:46   Boston
534  1967       Bobbi Gibb  Female  United States  03:27:17   Boston
535  1968       Bobbi Gibb  Female  United States  03:30:00   Boston
```

And I want to know whick winner has been the most winner, which winner name is most repeat in the dataset:

```
Grete Waitz           11
Bill Rodgers           8
Ingrid Kristiansen     8
Paula Radcliffe        7
Uta Pippig             7
Clarence DeMar         7
Eliud Kipchoge         6
Mary Keitany           6
Rosa Mota              6
Catherine Ndereba      6
Martin Lel             5
Joyce Chepchumba       5
Khalid Khannouchi      5
Wilson Kipsang         5
Steve Jones            4
```

IV. Data Mining / Clean Data

The main dataset was cleaned and changed type of majority columns to astype category or float64, even in one case did an encoding.

```
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   year      536 non-null    int64
 1   winner    536 non-null    object
 2   gender    536 non-null    object
 3   country   536 non-null    object
 4   time      536 non-null    object
 5   marathon  536 non-null    object
dtypes: int64(1), object(5)
```

```
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   year              536 non-null    int64
 1   winner            536 non-null    object
 2   gender            536 non-null    category
 3   country           536 non-null    object
 4   time              536 non-null    float64
 5   marathon          536 non-null    object
```

V. Encoding

TIn the case Correlation Matrix it has needed to tell different sub-steps:

1. The columns in the DF were object type, so it needed to change the type to integer to show correlation matrix.
2. But, it was not possible from object to category or object to integer, so it did an Encode each column:
    1. Encoding the gender
    2. Encoding the country to 37 codes
    3. Encoding the marathon city to 6 codes.
3. I created 3 new columns with encoding values, respectively.

	year	winner	gender	country	time	marathon	gender_1	encoded_country	encoded_marathon
0	2014	Dennis Kimetto	Male	Kenya	7377.0	Berlin	0	17	0
1	2011	Geoffrey Mutai	Male	Kenya	7382.0	Boston	0	17	1
2	2016	Kenenisa Bekele	Male	Ethiopia	7383.0	Berlin	0	8	0
3	2016	Eliud Kipchoge	Male	Kenya	7385.0	London	0	17	3
4	2013	Wilson Kipsang	Male	Kenya	7403.0	Berlin	0	17	0
...	...	...	...	...	...	...	...	...	...
531	1966	Bobbi Gibb	Female	United States	12100.0	Boston	1	35	1
532	1974	Jutta von Haase	Female	Germany	12121.0	Berlin	1	10	0
533	1969	Sara Mae Berman	Female	United States	12166.0	Boston	1	35	1
534	1967	Bobbi Gibb	Female	United States	12437.0	Boston	1	35	1
535	1968	Bobbi Gibb	Female	United States	12600.0	Boston	1	35	1