Best Marathons Majors

Xmas Project - Ariadna Puigventós

10th January, 2021

index()

1. Subject & datasets
2. Hypothesis
3. Steps & Pie Time Spent
   a. Data Wrangling
   b. Analyzing
   c. Researching
4. Conclusions
5. Webgraphy

1. Subject & Datasets

The subject is about **where the best marathons majors are from** in athletics category.

1. Main dataset is about 6 Best Marathons Majors by Countries.
2. The second one is about latitude of each Country dataset which compete in marathons.
3. The third dataset is only to reconfirm the global idea with 120 Olympic Games dataset by Countries and categories.

2. Hypothesis

THE BEST WORLD MARATHONS HAVE BEEN WON BY AFRICAN ATHLETES.

3. All Steps

1. **SET UP**
    1. It worked with different tests files in notebook/ folder where there are different operations, correct and fail code.
    2. After, 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.

3. All Steps

2. DATA WRANGLING

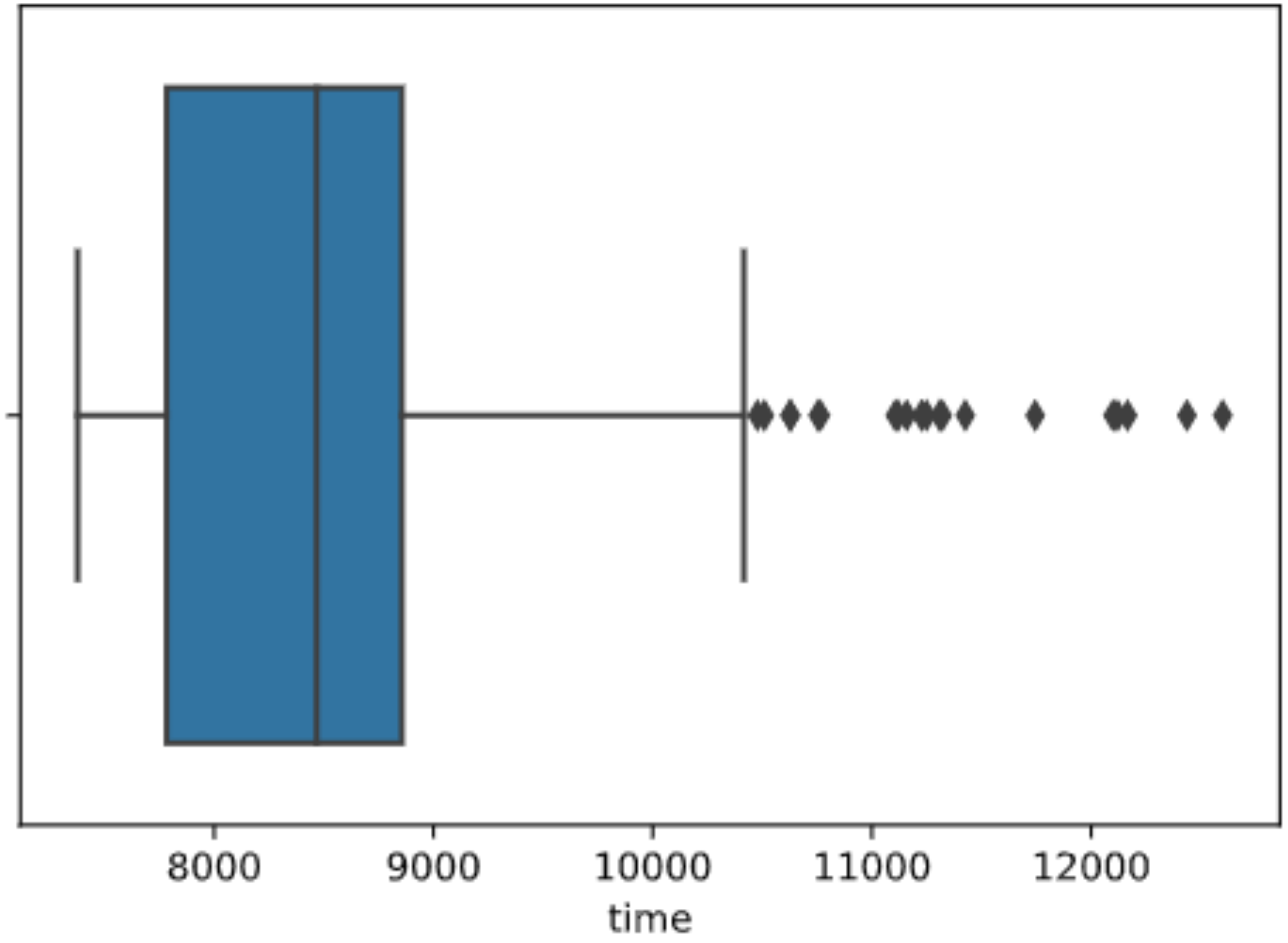
1. The main dataset was cleaned and changed type of majority columns to astype category or float64, even in one case did an encoding.
2. There were not any duplicates but it **detected some outliers** in variable “Time”. It was necessary to change of Time Column from object type **to float64 with method “to_timedelta64[s].**
3. It applied different methods to confirm the hypothesis like **head(), tail() and most repeated values** who won more than one time.

OUTLIERS

2 EXTREMES: THE FIRST TIME WAS 2:02:57 BY KENIAN ATHLETE IN BERLIN MARATHON IN 2014; AND THE LAST TIME WAS 3:30:00 BY UNITED STATES ATHLETE IN BOSTON MARATHON IN 1968.

ALTHOUGH, 25% MARATHON MAJORS GOT A MEDIAN AROUND 2 HOURS AND 16 MINUTS AND THE MOST MAJORS WITH 75% GOT 2 HOURS AND 46 MINUTS.

if you in median_list:
else:
print("welcome to the real world")



3. All Steps

3. **ANALYZING**

    1. A **pie chart** for showing the **top 10 countries and the rest of the countries in a new row "others"** category must be to concat the main dataset with a new DataFrame with the new row.

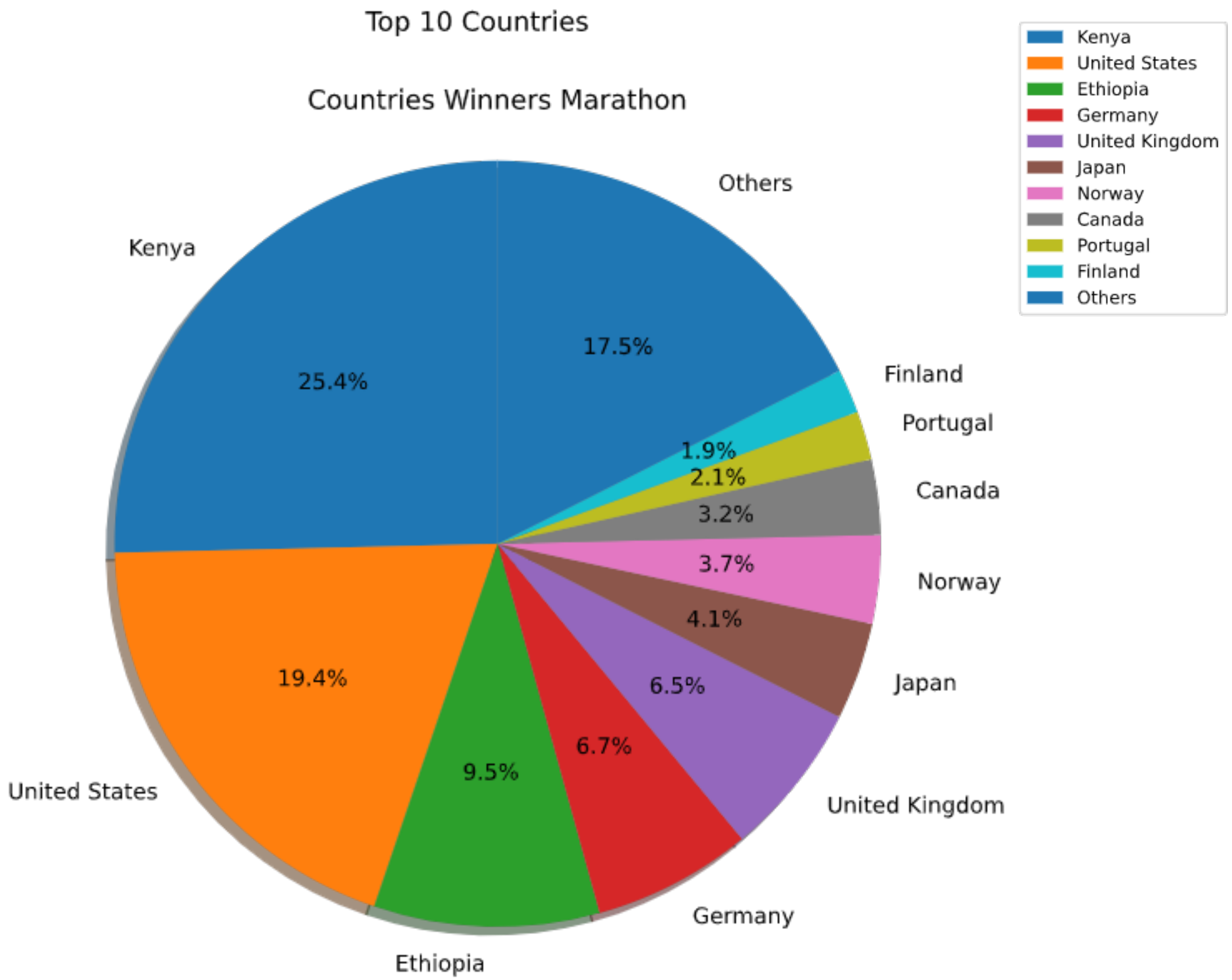
    2. An **histogram to reconfirm** the hypothesis and in this case it must be with 36 bins (not 5) because it was better to demonstrate.

    3. An **histogram** to show the **difference or equality between gender**. It must be to change type from object to category.

TOP 10 COUNTRIES

THANKS TO THIS PIE CHART GRAPHIC IT'S SHOWING THAT KENYA IS THE COUNTRY WINNER WITH 136 MARATHONS, IT'S 25,4%. IN ADDITION, THE THIRD COUNTRY IS ETHIOPIA WITH APROX 10%, SO IF IT'S TALKING ABOUT AFRICAN ATHLETES ARE WINNERS OF THE COMPETITION FOR A APROX. 35% OF THE TOTAL PIE.

FOR CURIOSITY, ONLY THERE WAS 1 SPANISH ATHLETE WHO WON 2 WORLD MARATHONS: BERLÍN 1996 AND LONDON 1998 WITH THE BEST TIME 2:09:15 AND 2:07:57, RESPECTIVELY.

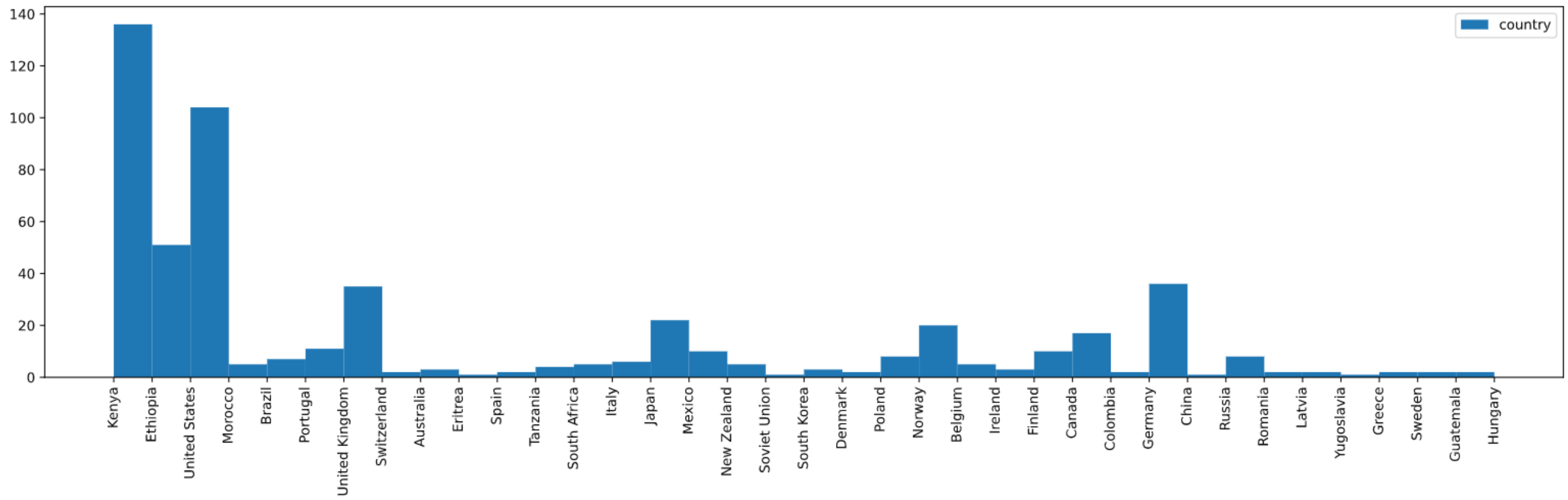

Top 10 Countries

Countries Winners Marathon

RANKING - RE(CONFIRM)

EFFECTIVELY, THERE ARE 2 COUNTRIES IS STANDING OUT (KENIA AND ETHIOPIA) VS. REST OF THE COUNTRIES.
KENYA IS MOST COUNTRY THAT REPEAT AND IS ON THE TOP, BELOW UNITED STATES AND ETHIOPIA.

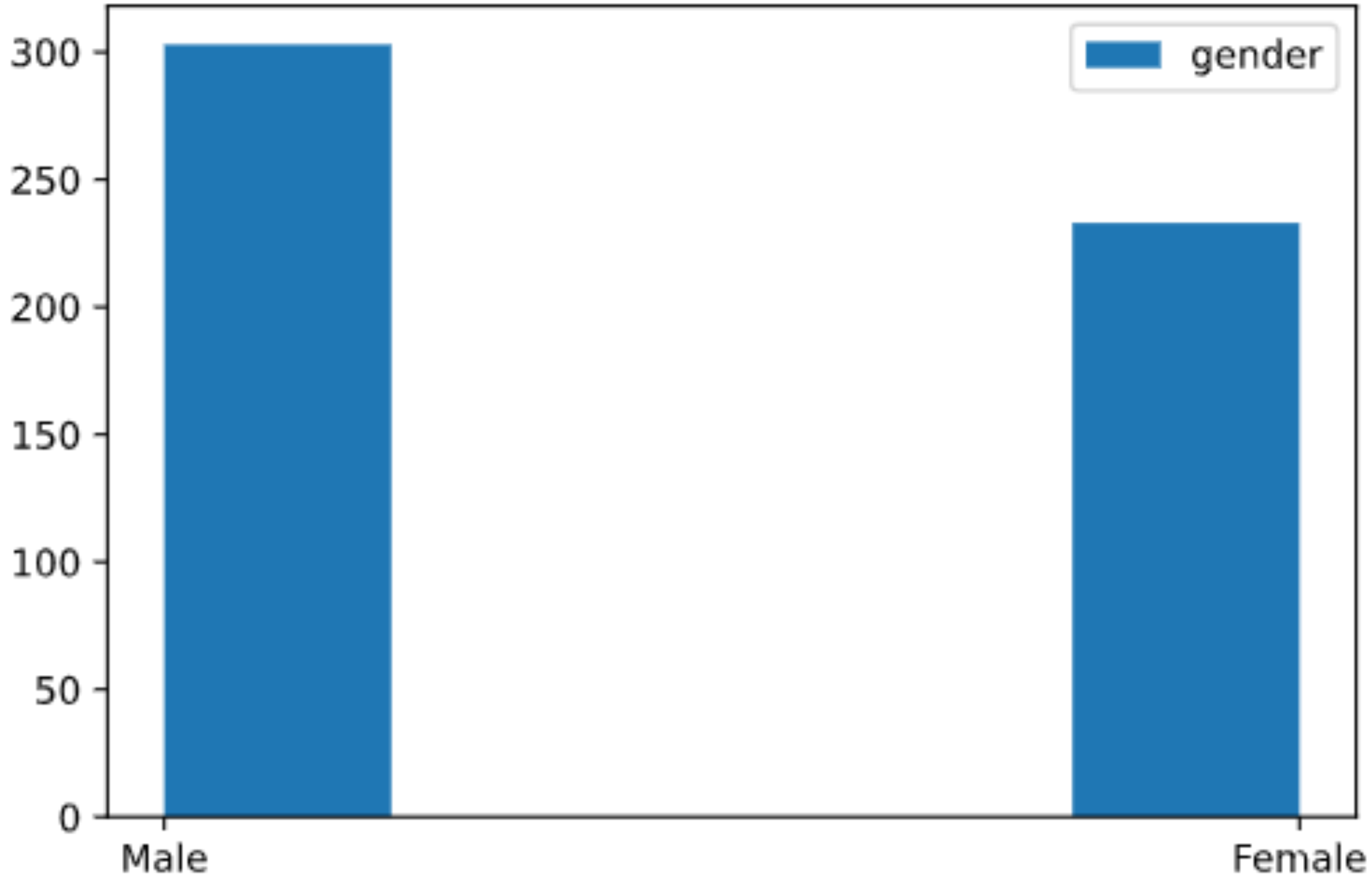
IT'S CONSIDERING THAT KENYA STARTED TO COMPETE IN 1960 AND UNITED STATES SINCE 1896.



A MATTER OF GENDER?

ONLY THERE IS A DIFFERENT OF 13% (70 WORLD MARATHONS) MORE WON THEM BY MALE ATHLETES WITH 56,5% (303 WM) THAN FEMALE WITH 43,4% (233 WM).

IT'S A GREAT APPRECIATION BECAUSE IT NOTICES THAT THE FIRST WOMAN ATHLETE WHO COULD COMPETE WAS 71 YEARS AFTER THAN MEN (IN 1967).

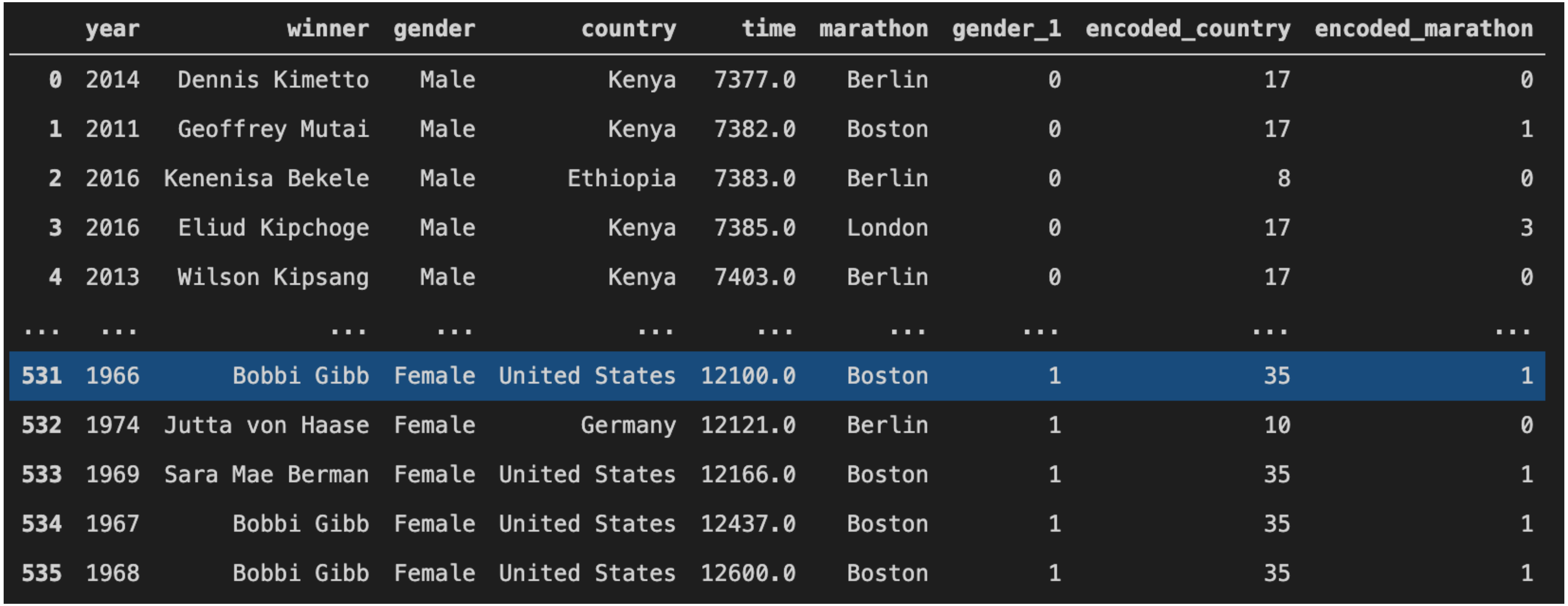


3. All Steps

3. **ANALYZING**
    1. In 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: **encoding the gender (like boolean) to Male is 0 and Female is 1; encoding the country to 37 codes and encoding the marathon city to 6 codes**.
        3. It created 3 new columns with encoding values, respectively.
        4. At the end, the time column has been changed from object to split string and after it was changed to float64 by seconds.

3. ANALYZING

DataFrame with 3 encoded new columns:

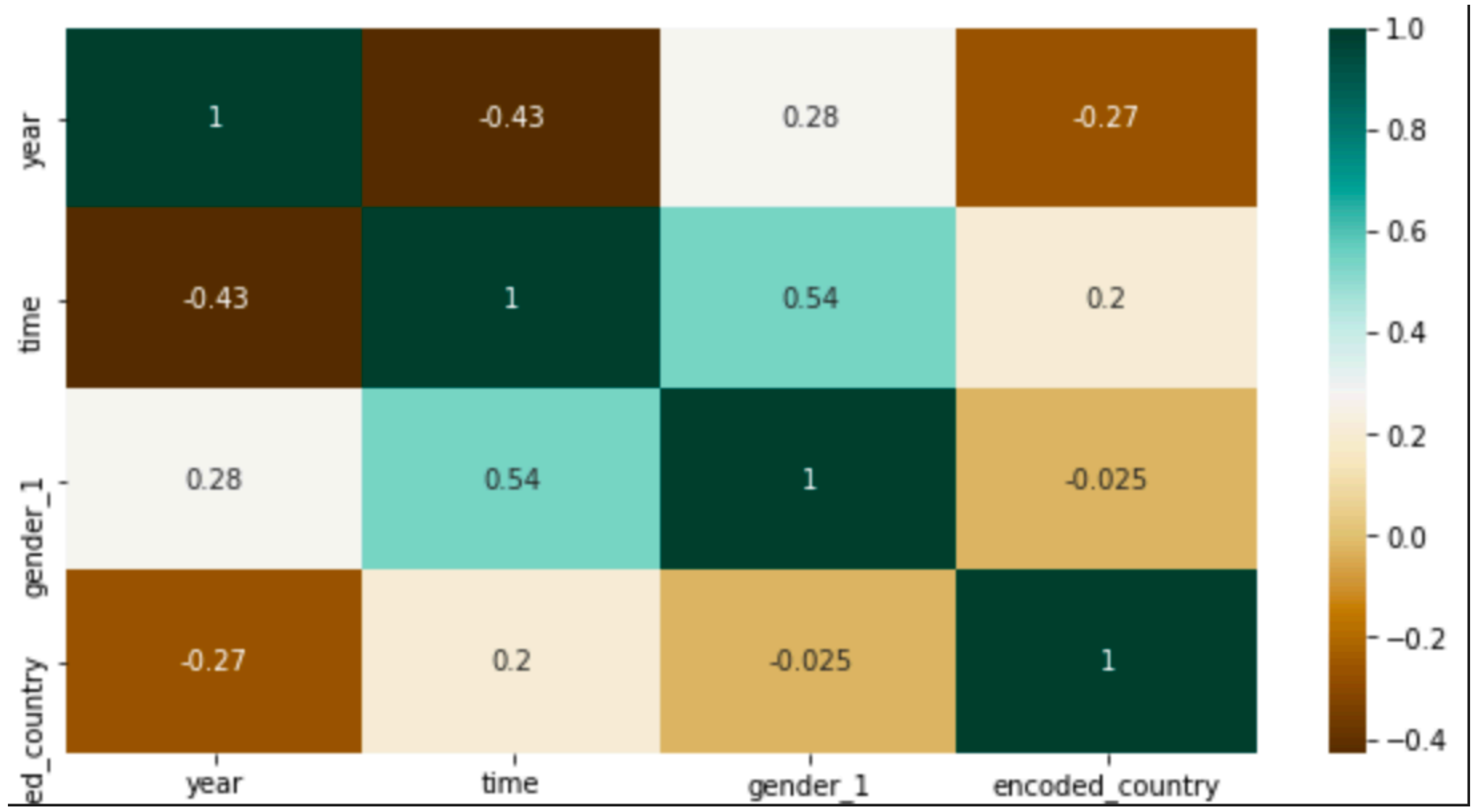


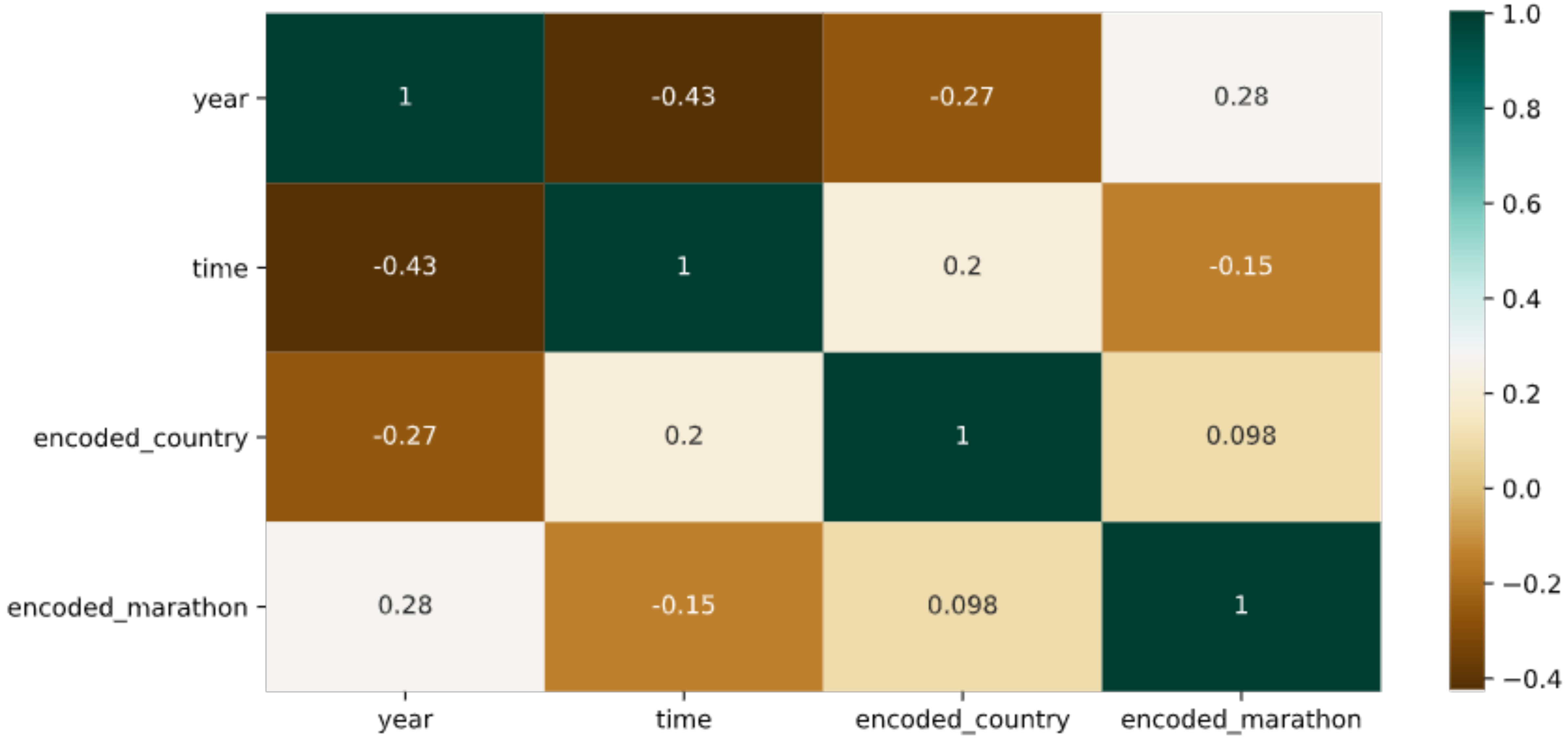
```
     year          winner  gender        country     time  marathon  gender_1  encoded_country  encoded_marathon
0    2014  Dennis Kimetto    Male          Kenya   7377.0    Berlin         0               17                 0
1    2011  Geoffrey Mutai    Male          Kenya   7382.0    Boston         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
```

CORRELATION MATRIX

IN THIS MATRIX, THE VARIABLE BELOW SHOWS THAT THE CORRELATION BETWEEN “GENDER_1” AND “TIME” IS 0.54, WHICH INDICATES THAT THEY’RE STRONGLY POSITIVELY CORRELATED.

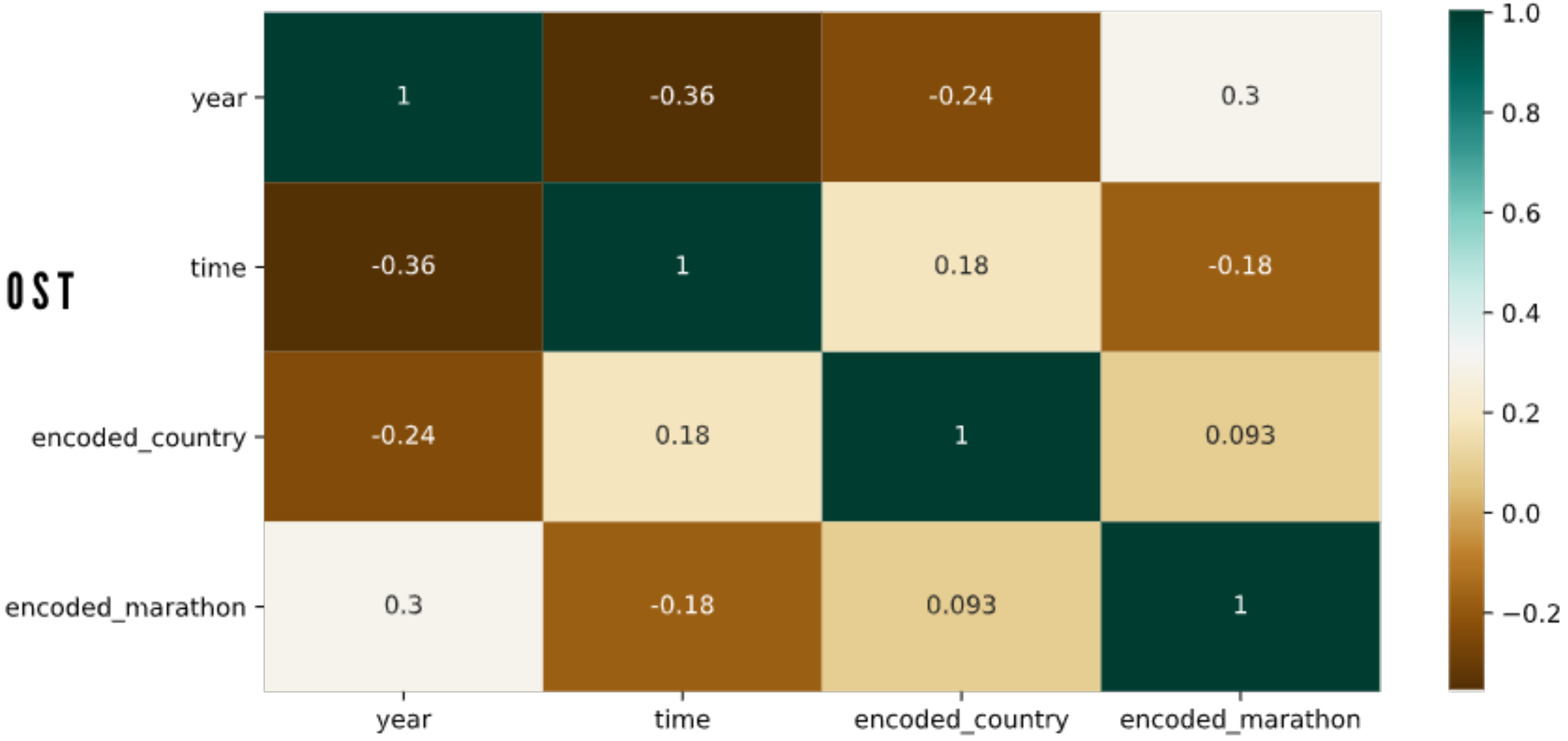
DESPITE OF, THERE ARE POSITIVELY, BUT TINY, BETWEEN "ENCODED_COUNTRY" AND "TIME" IS 0.2.





CORRELATION MATRIX

TO SHOW THE CORRELATION MATRIX FROM 1960 WHEN KENYA WAS THE FIRST TIME COMPETE IN BEST WORLD MARATHON, ALMOST IN 1967 WAS THE FIRST WOMAN WHO STARTED TO COMPETE THERE AS WELL.



3. All Steps

"Male" == O

3. RESEARCHING

ELIUD KIPCHOGE

London, 2019 (2:02:37)

Cleaned, cooked, garden duties and slept in an humiliate way.

DENNIS KIMETTO

Berlin, 2014 (2:02:57)

Grow up in a farm community and trained with runners team.

GEOFFREY MUTAI

Boston, 2011 (2:03:02)

Didn't wear shoes until he was teenager. Applied to a competition and he was accepted as long as he could maintenance the level.

3. All Steps

"Female" == 1

3. RESEARCHING

PAULA RADCLIFFE

London, 2003 (2:15:25)

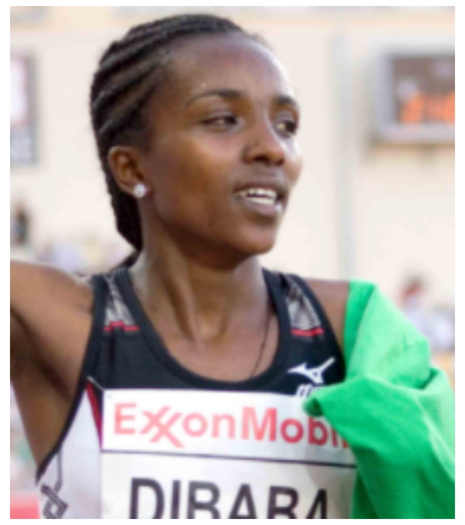
Run since she was 12 years old in an Athletic Club.

MARY KEITANY

London, 2017 (2:17:00)

When she was child had to walk more than 2km without shoes to get some water and 10km to go the school.

adidas RUNNING



TIRUNESH DIBABA

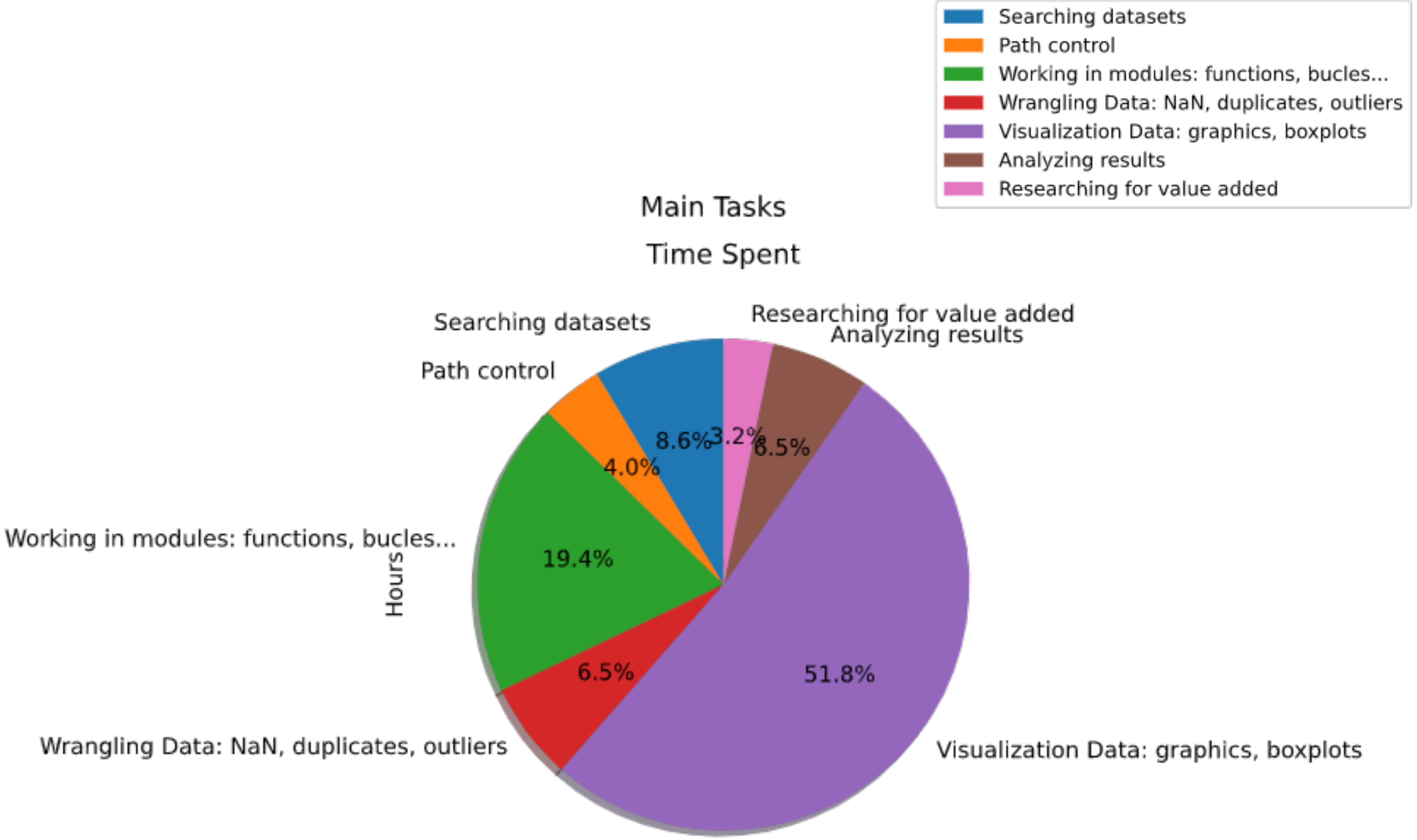
Chicago, 2017 (2:18:31)

From athletic family and she was child she already trained in the highest area in Ethiopia.

TIME SPENT

IN THE BEGGING OF THE XMAS PROJECT, IT SPENT THREE DAYS TO FIND A GOOD DATASET.

WHEN THE DATASET WAS SELECTED, THERE WERE TWO TASKS AS VISUALIZATION WERE MAJORITY TIME SPENT, IN ADDITION IT INCLUDED SEARCHING CODE IN GOOGLE IS APROX 52% AND WORKING IN MODULES IS APROX 19,5%.



4. Conclusions

1. The hypothesis is confirmed because there are more than cases shows that African, overcoat Kenyan, athletes win most best world marathons.
2. There are touchpoint:
   1. **African countries have had a lot of difference to compete** in marathons and another competitions as well (dataset 3).
   2. There are **less variety between African athletes** than north athletes.
   3. The **median time is 2 hours 46 minutes**, it is rare value to get a race with only 2 hours.
   4. **Gender vs Time is strongly positive correlation**, but not much with country.
   5. Two brands more used by **African athletes are Nike or Adidas**.
   6. **Altitude of the countries could be an influence** for the best runners (2 dataset).

4. Conclusions

1. What would you change if you need to do another EDA project?

    1. This subject was so much interesting and funny because it could extend more data about brand shoes sales or if the altitude influences really for African Athletes.
    2. In addition, it is knowing that marathons cause and increase the temperatures of cities.

2. What do you learn doing this project?

    1. First of all, starting the project, set up all functions in the modules.
    2. After, a lot of new methods, functions and searching in Google.
    3. At the end, how the visualization data can show and verify some theories, and learning it's possible to get an idea and extend it to predict the consequences.

5. Webgraphy

General Information to Inspiration

- https://www.mundodeportivo.com/atletismo/20190307/46894079804/por-que-los-atletas-africanos-ganan-siempre-en-maraton.html
- https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0037407
- Impact of Environmental on marathon running performance
- https://www.efdeportes.com/efd148/la-superioridad-de-los-atletas-africanos.htm

Datasets Search

- www.data.world.com
- www.kaggle.com

Recollect Data

- https://github.com/ali-ce/datasets/blob/master/Marathon-Majors/Winners.csv
- https://data.world/newns92/abbott-world-marathon-majors-winners
- https://data.world/johayes13/summer-winter-olympic-games
- https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results?select=noc_regions.csv
- https://developers.google.com/public-data/docs/canonical/countries_csv

Code webpages to work project

- https://stackoverflow.com/questions/38229357/how-to-sum-time-in-a-dataframe
- https://stackoverflow.com/questions/12065885/filter-dataframe-rows-if-value-in-column-is-in-a-set-list-of-values
- https://towardsdatascience.com/exploratory-data-analysis-in-python-c9a77dfa39ce
- https://mode.com/example-gallery/python_histogram/
- https://pbpython.com/pandas_dtypes.html
- https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.hist.html
- https://pandas.pydata.org/pandas-docs/
- https://matplotlib.org/gallery/pie_and_polar_charts/pie_features.html#sphx-glr-gallery-pie-and-polar-charts-pie-features-py
- https://datatofish.com/pie-chart-matplotlib/
- https://www.delftstack.com/es/howto/matplotlib/how-to-change-the-figure-size-in-matplotlib/
- https://stackoverflow.com/questions/57314529/multiple-pie-charts-from-pandas-dataframe
- https://likegeeks.com/es/matrix-correlacion-python/
- **https://seaborn.pydata.org/examples/grouped_barplot.html**
- https://opensource.com/article/19/7/create-pull-request-github
- https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html
- **https://stackoverflow.com/questions/54052471/mapping-values-in-place-for-example-with-gender-from-string-to-int-in-pandas-d**
- https://stackoverflow.com/questions/29432629/plot-correlation-matrix-using-pandas
- **https://towardsdatascience.com/label-encoder-and-onehot-encoder-in-python-83d32288b592**
- https://stackoverflow.com/questions/41463763/merge-2-dataframes-with-same-values-in-a-column
- https://stackoverflow.com/questions/48587997/matplotlib-pie-graph-with-all-other-categories
- **https://realpython.com/pandas-merge-join-and-concat/**
- https://stackoverflow.com/questions/31405860/three-python-modules-calling-one-another
- https://www.statology.org/how-to-read-a-correlation-matrix/
- **https://towardsdatascience.com/pie-charts-in-python-302de204966c**
- **https://markdown.es/sintaxis-markdown/#parrafos**

Thanks ;)