WEEK 6 - HOMEWORK 1
Introduction to Kaggle

By: Anugrah Yazid Ghani

Data Description & Preparation

DigitalSkola

Week 6 - HW 1

Notebook | Data | Logs | Comments (0) | Settings

0 | Edit

Table of Contents

WEEK 6 - HOMEWORK 1 - INTRODUCTION TO KAGGLE

By: Anugrah Yazid Ghani

Contents:

1. Load Data
2. Brief information about the data type (.info())
3. Brief information about the basic statistics (.describe())

Data Description

Data about countries who participate and wins the medals (Gold, Silver, Bronze) in Winter Olympic from year 1924 - 2018.

Data Preparation

In [2]:

```
# Import Library
import pandas as pd
import numpy as np

# Import Visualization Library
import matplotlib.pyplot as plt
import seaborn as sns
```

1. Load Data

1. Load Data

In [3]:

```
# Read Data
data = pd.read_csv("../input/winter-olympic-medals-1924-2018/Winter_Olympic_Medals.csv")
data
```

Hide output

Out[3]:

	Year	Host_country	Host_city	Country_Name	Country_Code	Gold	Silver	Bronze
0	1924	France	Chamonix	United States	USA	1	2	1
1	1924	France	Chamonix	Great Britain	GBR	1	1	2
2	1924	France	Chamonix	Austria	AUT	2	1	0
3	1924	France	Chamonix	Norway	NOR	4	7	6
4	1924	France	Chamonix	Finland	FIN	4	4	3
...	...	...	...	...	...	...	...	...
404	2018	South Korea	Pyeongchang	Slovakia	SVK	1	2	0
405	2018	South Korea	Pyeongchang	China	CHN	1	6	2
406	2018	South Korea	Pyeongchang	Hungary	HUN	1	0	0
407	2018	South Korea	Pyeongchang	Poland	POL	1	0	1
408	2018	South Korea	Pyeongchang	Great Britain	GBR	1	0	4

409 rows × 8 columns

There are 409 rows x 8 columns in this data. The columns are:

1. Year (Year when the Winter Olympic held)
2. Host_country (Country who held the Winter Olympic)
3. Host_city (City who held the Winter Olympic)
4. Country_Name (Countries who participates in Winter Olympic)
5. Country_Code (Code of each countries who participate at Winter Olympic)
6. Gold (The number of gold medals earned by each Countries at Winter Olympic)
7. Silver (The number of silver medals earned by each Countries at Winter Olympic)
8. Bronze (The number of bronze medals earned by each Countries at Winter Olympic)

2. Data Types

2. Data Types

In [4]:

```
# Chechk Data Type and Number of Null Values
data_type = pd.DataFrame(data.dtypes).T.rename({0 : "Data Type"})
data_type
```

Out[4]:

	Year	Host_country	Host_city	Country_Name	Country_Code	Gold	Silver	Bronze
Data Type	int64	object	object	object	object	int64	int64	int64

There are 2 data types in the Data.

- Integer : Year, Gold, Silver, Bronze
- Object : Host_country, Host_city, Country_Name, Country_Code

In [5]:

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 409 entries, 0 to 408
Data columns (total 8 columns):
 #   Column        Non-Null Count  Dtype 
---  ------        --------------  ----- 
 0   Year          409 non-null    int64 
 1   Host_country  409 non-null    object
 2   Host_city     409 non-null    object
 3   Country_Name  409 non-null    object
 4   Country_Code  409 non-null    object
 5   Gold          409 non-null    int64 
 6   Silver        409 non-null    int64 
 7   Bronze        409 non-null    int64 
dtypes: int64(4), object(4)
memory usage: 25.7+ KB
```

3. Basic Statistics

3. Basic Statistics

In [6]:
```
data.describe()
```

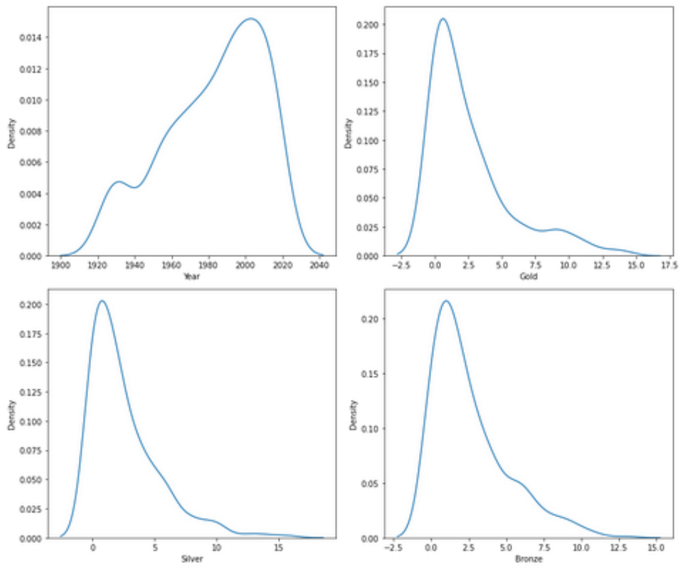
Out[6]:

	Year	Gold	Silver	Bronze
count	409.000000	409.000000	409.000000	409.000000
mean	1982.611247	2.596577	2.589242	2.567237
std	26.721658	3.085559	2.821049	2.491358
min	1924.000000	0.000000	0.000000	0.000000
25%	1964.000000	0.000000	1.000000	1.000000
50%	1988.000000	1.000000	2.000000	2.000000
75%	2006.000000	4.000000	4.000000	4.000000
max	2018.000000	14.000000	16.000000	13.000000

In [7]:
```
# Create function to show the density of the data in distribution
columns = ['Year', 'Gold', 'Silver', 'Bronze']

fig, axes = plt.subplots(2, 2, figsize=(12,10))
axes = [ax for axes_row in axes for ax in axes_row]

for i, col in enumerate(columns):
    plot = sns.kdeplot(data=data, x=col, ax=axes[i])
plt.tight_layout()
```



DigitalSkola

Kaggle Link

https://www.kaggle.com/anugrahyazidghani/week-6-hw-1

THANK YOU!