nobel_prize_laureates

June 24, 2024

[ ]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

[ ]:
```
data = pd.read_excel(io='nobel-prize-laureates.xlsx')
```

[ ]:
```
#Qn.1 How many Nobel Prize laureates are included in the dataset?
data.info()
#They are 1000 Nobel prize laureates
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 21 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Id                    1000 non-null   int64
 1   Firstname             1000 non-null   object
 2   Surname               968 non-null    object
 3   Born                  956 non-null    object
 4   Died                  670 non-null    object
 5   Born country          969 non-null    object
 6   Born country code     969 non-null    object
 7   Born city             967 non-null    object
 8   Died country          657 non-null    object
 9   Died country code     657 non-null    object
 10  Died city             651 non-null    object
 11  Gender                1000 non-null   object
 12  Year                  1000 non-null   int64
 13  Category              1000 non-null   object
 14  Overall motivation    23 non-null     object
 15  Motivation            1000 non-null   object
 16  Organization name     736 non-null    object
 17  Organization city     731 non-null    object
 18  Organization country  733 non-null    object
 19  Geo Shape             640 non-null    object
 20  Geo Point 2D          640 non-null    object
dtypes: int64(2), object(19)
memory usage: 164.2+ KB
```

[ ]:
```
data
```

[ ]:
```
      Id   Firstname     Surname        Born        Died  \
0    109     John H.    Van Vleck  1899-03-13  1980-10-27
1    154  William D.     Phillips  1948-11-05         NaN
2    164       Adolf   von Baeyer  1835-10-31  1917-08-20
3    840   George E.        Smith  1930-05-10         NaN
4    883    Lloyd S.      Shapley  1923-06-02  2016-03-12
..   …         …          …          …          …
995  808       Orhan       Pamuk  1952-06-07         NaN
996  220     Nikolay     Semenov  1896-04-03  1986-09-25
997  272     Hartmut      Michel  1948-07-18         NaN
998  419       Baruj  Benacerraf  1920-10-29  2011-08-02
999  423    David H.       Hubel  1926-02-27  2013-09-22

                  Born country Born country code        Born city  \
0                          USA                US    Middletown CT
1                          USA                US  Wilkes-Barre PA
2         Prussia (now Germany)                DE           Berlin
3                          USA                US  White Plains NY
4                          USA                US     Cambridge MA
..                         …                …              …
995                     Turkey                TR         Istanbul
996                     Russia                RU          Saratov
997  West Germany (now Germany)                DE      Ludwigsburg
998                  Venezuela                VE          Caracas
999                     Canada                CA       Windsor ON

          Died country Died country code  … Gender  Year    Category  \
0                  USA                US  …   male  1977     Physics
1                  NaN               NaN  …   male  1997     Physics
2              Germany                DE  …   male  1905   Chemistry
3                  NaN               NaN  …   male  2009     Physics
4                  USA                US  …   male  2012   Economics
..                 …                …  …  …    …   …         …
995                NaN               NaN  …   male  2006  Literature
996  USSR (now Russia)                RU  …   male  1956   Chemistry
997                NaN               NaN  …   male  1988   Chemistry
998                USA                US  …   male  1980    Medicine
999                USA                US  …   male  1981    Medicine

    Overall motivation                                        Motivation  \
0                  NaN  "for their fundamental theoretical investigati…
1                  NaN  "for development of methods to cool and trap a…
2                  NaN  "in recognition of his services in the advance…
3                  NaN  "for the invention of an imaging semiconductor…
4                  NaN  "for the theory of stable allocations and the …
```

```
..                  …                                                  …
995                 NaN  "who in the quest for the melancholic soul of …
996                 NaN  "for their researches into the mechanism of ch…
997                 NaN  "for the determination of the three-dimensiona…
998                 NaN  "for their discoveries concerning genetically …
999                 NaN  "for their discoveries concerning information …

                                    Organization name      Organization city  \
0                                     Harvard University           Cambridge MA
1       National Institute of Standards and Technology         Gaithersburg MD
2                                     Munich University                  Munich
3                                     Bell Laboratories          Murray Hill NJ
4                               University of California          Los Angeles CA
..                                                   …                     …
995                                                NaN                   NaN
996  Institute for Chemical Physics of the Academy …                    Moscow
997                   Max-Planck-Institut für Biophysik  Frankfurt-on-the-Main
998                               Harvard Medical School              Boston MA
999                               Harvard Medical School              Boston MA

    Organization country                                          Geo Shape  \
0                    USA  {"coordinates":[[[[-155.60651897,20.137955566]…
1                    USA  {"coordinates":[[[[-155.60651897,20.137955566]…
2                Germany  {"coordinates":[[[[6.742198113,53.57835521],[6…
3                    USA  {"coordinates":[[[[-155.60651897,20.137955566]…
4                    USA  {"coordinates":[[[[-155.60651897,20.137955566]…
..                   …                                                …
995                  NaN                                             NaN
996   USSR (now Russia)  {"coordinates":[[[[132.448985222,42.845404364]…
997              Germany  {"coordinates":[[[[6.742198113,53.57835521],[6…
998                  USA  {"coordinates":[[[[-155.60651897,20.137955566]…
999                  USA  {"coordinates":[[[[-155.60651897,20.137955566]…

                              Geo Point 2D
0    45.68753333949257, -112.49433391594509
1    45.68753333949257, -112.49433391594509
2     51.10627343575876, 10.381710872747147
3    45.68753333949257, -112.49433391594509
4    45.68753333949257, -112.49433391594509
..                                       …
995                                    NaN
996    61.98434173753343, 96.69345576745796
997   51.10627343575876, 10.381710872747147
998  45.68753333949257, -112.49433391594509
999  45.68753333949257, -112.49433391594509

[1000 rows x 21 columns]
```

```
[ ]: data.rename(columns={'Id':'id','Firstname':'first_name','Surname':
      ↪'sur_name','Born':'born','Died':'died','Born country':
                          'born_country','Born country code':
      ↪'born_country_code','Born city':'born_city','Died country':'died_country',
                          'Died country code':'died_country_code','Died city':
      ↪'died_city','Gender':'gender','Year':'year','Category':'category',
                          'Overall motivation':'overall_motivation','Motivation':
      ↪'motivation','Organization name':'organization_name','Organization city':
                          'organization_country','Geo Shape':'geo_shape','Geo Point 2D':'geo_point_2d'}
                          ,inplace=True)
```

Qn.2 Which country has the highest number of Nobel laureates?

USA

```
[ ]: data.born_country.value_counts()
```

```
[ ]: born_country
     USA                                                 292
     United Kingdom                                       90
     Germany                                              67
     France                                               58
     Sweden                                               30
                                                        ...
     Ethiopia                                              1
     French protectorate of Tunisia (now Tunisia)          1
     Madagascar                                            1
     Belgian Congo (now Democratic Republic of the Congo)  1
     Venezuela                                             1
     Name: count, Length: 127, dtype: int64
```

#Qn.3 What is the distribution of Nobel laureates across different prize categories?

```
[ ]: data.category.value_counts()
```

```
[ ]: category
     Medicine      227
     Physics       225
     Chemistry     194
     Peace         141
     Literature    120
     Economics      93
     Name: count, dtype: int64
```

Qn.4 How many Nobel laureates were awarded in each decade?

```
[ ]: group_year=pd.qcut(data['year'],q=10)
     group_year.value_counts()
```

```
[ ]: year
     (1969.0, 1979.0]      104
     (1989.0, 1999.0]      104
     (1900.999, 1920.0]    102
     (1956.7, 1969.0]      101
     (1920.0, 1938.0]      100
     (2015.1, 2023.0]      100
     (1938.0, 1956.7]       98
     (1999.0, 2007.0]       98
     (1979.0, 1989.0]       97
     (2007.0, 2015.1]       96
     Name: count, dtype: int64
```

```
[ ]: grouped_year=pd.cut(data['year'],bins=np.arange(data['year'].min(),data['year'].
       ↪max()+10,10))
```

```
[ ]: grouped_year.value_counts()
```

```
[ ]: year
     (2011, 2021]    122
     (2001, 2011]    119
     (1991, 2001]    114
     (1971, 1981]    111
     (1981, 1991]     93
     (1961, 1971]     82
     (1951, 1961]     69
     (1901, 1911]     62
     (1941, 1951]     58
     (1921, 1931]     55
     (1931, 1941]     45
     (1911, 1921]     39
     (2021, 2031]     25
     Name: count, dtype: int64
```

Qn.5 Are there any missing values in the dataset? If so, in which columns?

Yes they are there sur_name
born
died
born_country
born_country_code
born_city
died_country
died_country_code
died_city

overall_motivation
organization_name
organization_city
organization_country
geo_shape
geo_point_2d

[ ]:
```
data.isna().sum()
```

[ ]:
```
id                      0
first_name              0
sur_name               32
born                   44
died                  330
born_country           31
born_country_code      31
born_city              33
died_country          343
died_country_code     343
died_city             349
gender                  0
year                    0
category                0
overall_motivation    977
motivation              0
organization_name     264
organization_city     269
organization_country  267
geo_shape             360
geo_point_2d          360
dtype: int64
```

Qn.6 Perform data cleaning by handling missing values appropriately. Describe your approach.

[ ]:
```
#I am goint to leave all died columns because the Laureates have not yet died
#I am going to leave the sur_name column because from my research the two␣
 ↪individuals do not have surnames and the remaining are organizations
# which do not have surnames
data.
 ↪drop(columns=['overall_motivation','organization_name','organization_city','organization_co
```

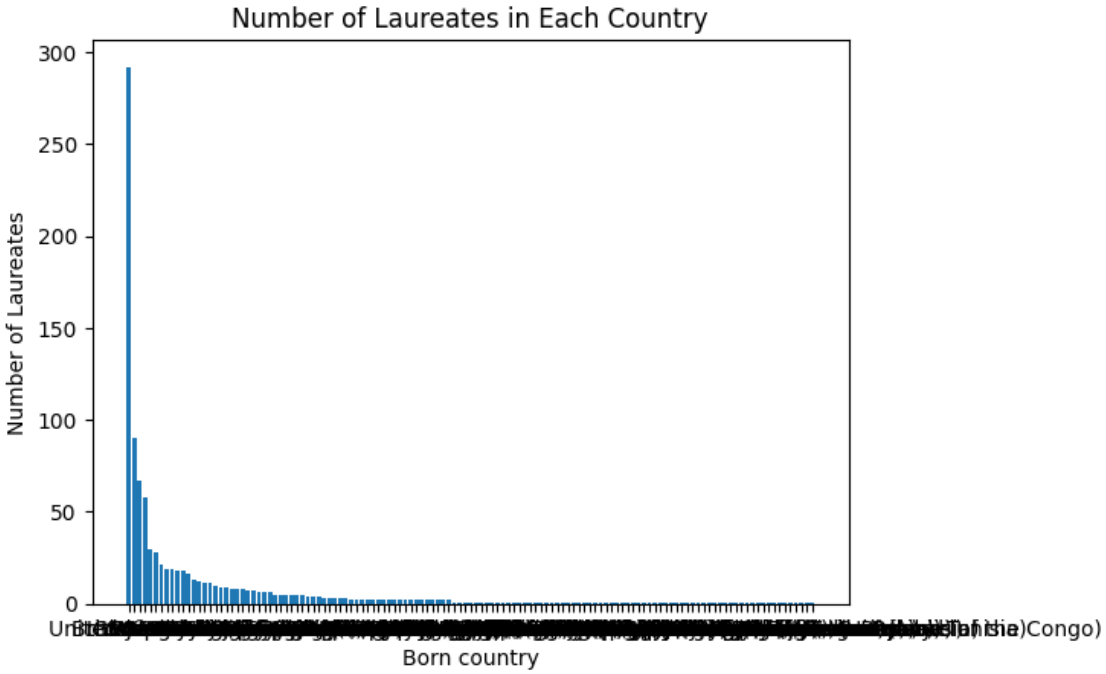
[ ]:
```
      id  first_name    sur_name        born        died  \
0    109     John H.   Van Vleck  1899-03-13  1980-10-27
1    154  William D.    Phillips  1948-11-05         NaN
2    164       Adolf  von Baeyer  1835-10-31  1917-08-20
3    840   George E.       Smith  1930-05-10         NaN
4    883    Lloyd S.     Shapley  1923-06-02  2016-03-12
..   …          …          …          …          …
```

```
995  808       Orhan       Pamuk  1952-06-07         NaN
996  220     Nikolay     Semenov  1896-04-03  1986-09-25
997  272     Hartmut      Michel  1948-07-18         NaN
998  419       Baruj  Benacerraf  1920-10-29  2011-08-02
999  423    David H.       Hubel  1926-02-27  2013-09-22

                  born_country born_country_code        born_city  \
0                          USA                US    Middletown CT
1                          USA                US  Wilkes-Barre PA
2         Prussia (now Germany)                DE           Berlin
3                          USA                US  White Plains NY
4                          USA                US     Cambridge MA
..                         …                 …               …
995                     Turkey                TR         Istanbul
996                     Russia                RU          Saratov
997  West Germany (now Germany)                DE      Ludwigsburg
998                  Venezuela                VE          Caracas
999                     Canada                CA       Windsor ON

          died_country died_country_code     died_city gender  year  \
0                  USA                US  Cambridge MA   male  1977
1                  NaN               NaN           NaN   male  1997
2              Germany                DE     Starnberg   male  1905
3                  NaN               NaN           NaN   male  2009
4                  USA                US     Tucson AZ   male  2012
..                 …                 …            …    …   …
995                NaN               NaN           NaN   male  2006
996  USSR (now Russia)                RU        Moscow   male  1956
997                NaN               NaN           NaN   male  1988
998                USA                US     Boston MA   male  1980
999                USA                US    Lincoln MA   male  1981

       category                                          motivation
0       Physics  "for their fundamental theoretical investigati…
1       Physics  "for development of methods to cool and trap a…
2     Chemistry  "in recognition of his services in the advance…
3       Physics  "for the invention of an imaging semiconductor…
4     Economics  "for the theory of stable allocations and the …
..          …                                                 …
995  Literature  "who in the quest for the melancholic soul of …
996   Chemistry  "for their researches into the mechanism of ch…
997   Chemistry  "for the determination of the three-dimensiona…
998    Medicine  "for their discoveries concerning genetically …
999    Medicine  "for their discoveries concerning information …

[1000 rows x 15 columns]
```

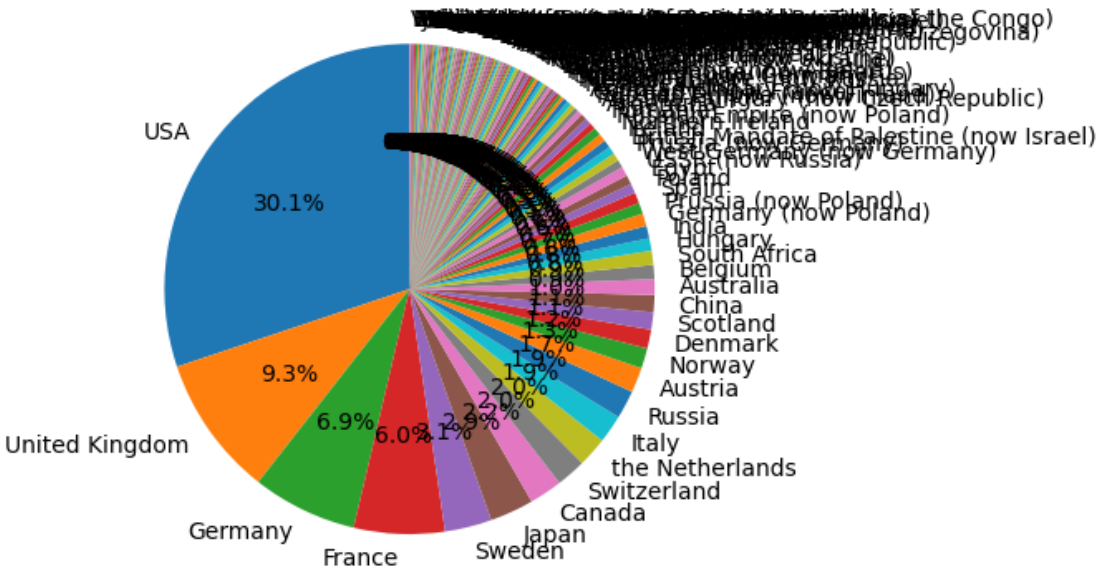
Qn.7 Visualize the distribution of Nobel laureates' birth countries on a world map.

[ ]:
```
born_counts=data.born_country.value_counts()
```

[ ]:
```
plt.bar(born_counts.index,born_counts.values)
plt.xlabel('Born country')
plt.ylabel('Number of Laureates')
plt.title('Number of Laureates in Each Country')
plt.show()
```



[ ]:
```
plt.pie(born_counts, labels=born_counts.index, autopct='%1.1f%%', startangle=90)
plt.show()
```



Qn.8 Is there any correlation between a laureate's birth year and the year they were awarded the Nobel Prize? Visualize if there's any relationship.

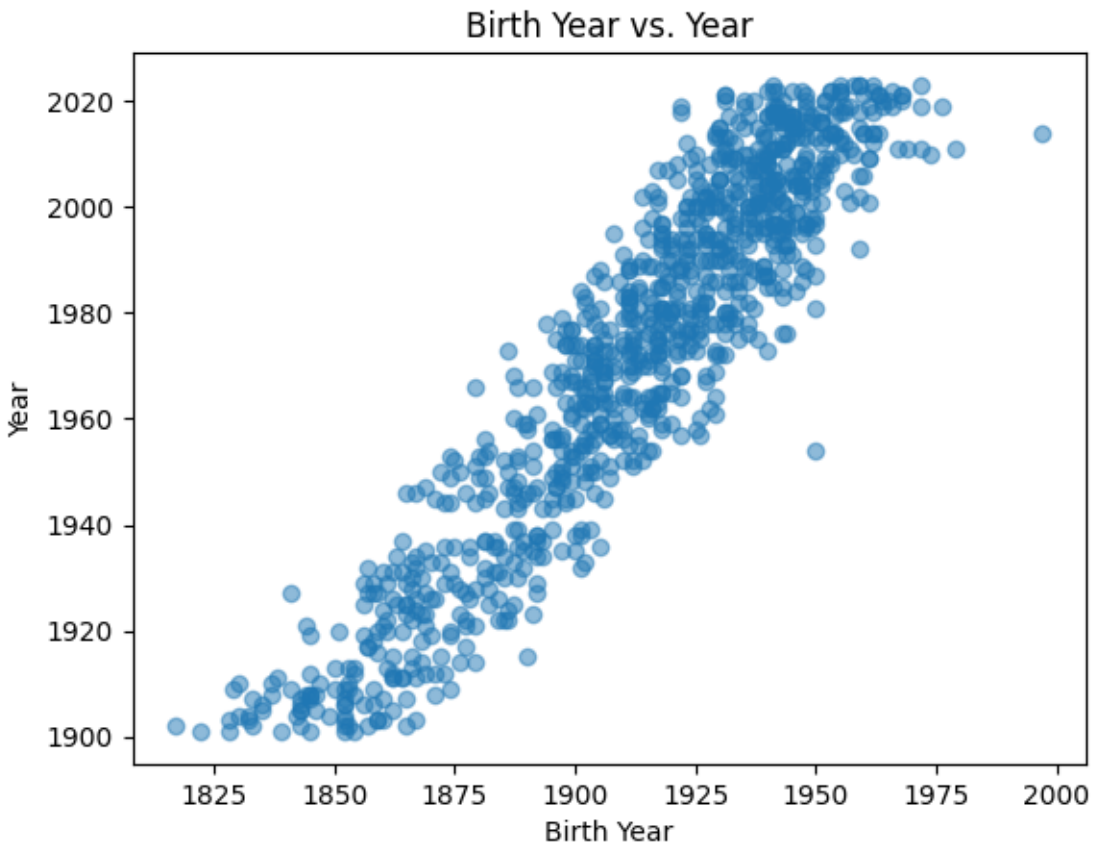
[ ]:
```
data['birth_year']=pd.to_datetime(data['born']).dt.year
data.head()
```

[ ]:

	id	first_name	sur_name	born	died	born_country	born_country_code	born_city	died_country	died_country_code	…	year	category	overall_motivation
0	109	John H.	Van Vleck	1899-03-13	1980-10-27	USA	US	Middletown CT	USA	US	…	1977	Physics	NaN
1	154	William D.	Phillips	1948-11-05	NaN	USA	US	Wilkes-Barre PA	NaN	NaN	…	1997	Physics	NaN
2	164	Adolf	von Baeyer	1835-10-31	1917-08-20	Prussia (now Germany)	DE	Berlin	Germany	DE	…	1905	Chemistry	NaN
3	840	George E.	Smith	1930-05-10	NaN	USA	US	White Plains NY	NaN	NaN	…	2009	Physics	NaN
4	883	Lloyd S.	Shapley	1923-06-02	2016-03-12	USA	US	Cambridge MA	USA	US	…	2012	Economics	NaN

```
                                          motivation  \
0  "for their fundamental theoretical investigati…
1  "for development of methods to cool and trap a…
2  "in recognition of his services in the advance…
3  "for the invention of an imaging semiconductor…
4  "for the theory of stable allocations and the …

                              organization_name organization_city  \
0                            Harvard University      Cambridge MA
1  National Institute of Standards and Technology   Gaithersburg MD
2                              Munich University            Munich
3                              Bell Laboratories    Murray Hill NJ
4                         University of California    Los Angeles CA

  organization_country                                          geo_shape  \
0                  USA  {"coordinates":[[[[-155.60651897,20.137955566]…
1                  USA  {"coordinates":[[[[-155.60651897,20.137955566]…
2              Germany  {"coordinates":[[[[6.742198113,53.57835521],[6…
3                  USA  {"coordinates":[[[[-155.60651897,20.137955566]…
4                  USA  {"coordinates":[[[[-155.60651897,20.137955566]…

                            geo_point_2d birth_year
0  45.68753333949257, -112.49433391594509     1899.0
1  45.68753333949257, -112.49433391594509     1948.0
2   51.10627343575876, 10.381710872747147     1835.0
3  45.68753333949257, -112.49433391594509     1930.0
4  45.68753333949257, -112.49433391594509     1923.0

[5 rows x 22 columns]
```

[ ]:
```
plt.scatter(data['birth_year'], data['year'], alpha=0.5)
plt.xlabel('Birth Year')
plt.ylabel('Year')
plt.title('Birth Year vs. Year')
plt.show()
```



There is a linear relation between the laureates' birth year and the year they were awarded the prize

Qn.9 Perform anomaly detection on the birth years of laureates. Identify and explain any outliers.

[ ]:

Qn.10 Based on the dataset, can you identify any interesting trends or patterns regarding Nobel laureates' demographics or the fields in which they were awarded?

[ ]: