DATA ANALYSIS AND VISUALISATION OF PALMER PENGUINS DATASET

IMPORTING NECESSARY LIBRARIES

In [2]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

In [39]:
```
sns.set_context("notebook", rc = {"grid.linewidth": 1})
sns.set_style('darkgrid')
```

READING THE DATASET

In [4]:
```
df = pd.read_csv('penguins.csv')
```

DISPLAYING THE FIRST AND LAST FIVE ROWS

In [5]:
```
df.head()
```

Out[5]:

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
0	Adelie	Torgersen	39.1	18.7	181.0	3750.0	MALE
1	Adelie	Torgersen	39.5	17.4	186.0	3800.0	FEMALE
2	Adelie	Torgersen	40.3	18.0	195.0	3250.0	FEMALE
3	Adelie	Torgersen	NaN	NaN	NaN	NaN	NaN
4	Adelie	Torgersen	36.7	19.3	193.0	3450.0	FEMALE

In [6]:
```
df.tail()
```

Out[6]:

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
339	Gentoo	Biscoe	NaN	NaN	NaN	NaN	NaN
340	Gentoo	Biscoe	46.8	14.3	215.0	4850.0	FEMALE
341	Gentoo	Biscoe	50.4	15.7	222.0	5750.0	MALE
342	Gentoo	Biscoe	45.2	14.8	212.0	5200.0	FEMALE
343	Gentoo	Biscoe	49.9	16.1	213.0	5400.0	MALE

NUMBER OF ROWS AND COLUMNS IN THE DATASET

In [7]:
```
df.shape
```

Out[7]: (344, 7)

GETTING THE NUMBER OF NULL VALUES FOR EACH COLUMN

In [8]:
```
df.isnull().sum()
```

Out[8]:
```
species              0
island               0
bill_length_mm       2
bill_depth_mm        2
flipper_length_mm    2
body_mass_g          2
sex                 11
dtype: int64
```

QUESTION 1.

FILL THE NULL VALUES FOR CATEGORICAL DATA AS `UNKNOWN`

In [9]:
```
df['sex'].fillna(value = 'UNKNOWN', inplace = True)
```

In [10]:
```
df.isnull().sum()
```

Out[10]:
```
species              0
island               0
bill_length_mm       2
bill_depth_mm        2
flipper_length_mm    2
body_mass_g          2
sex                  0
dtype: int64
```

QUESTION 2.

FILL THE NULL VALUES FOR NUMERICAL DATA WITH THEIR `MEDIAN`.

In [14]:
```
print(sp.median(df['bill_length_mm']))
print(sp.median(df['bill_depth_mm']))
print(sp.median(df['flipper_length_mm']))
print(sp.median(df['body_mass_g']))
```

Out[14]:
```
44.45
17.3
197.0
4050.0
```

In [15]:
```
df['bill_length_mm'].fillna(value = 44.45 , inplace = True)
df['bill_depth_mm'].fillna(value = 17.30 , inplace = True)
df['flipper_length_mm'].fillna(value = 197.00 , inplace = True)
df['body_mass_g'].fillna(value = 4050.00 , inplace = True)
```

In [16]:
```
df.isnull().sum()
```

Out[16]:
```
species              0
island               0
bill_length_mm       0
bill_depth_mm        0
flipper_length_mm    0
body_mass_g          0
sex                  0
dtype: int64
```

QUESTION 3.

CALCULATE THE TOTAL NUMBER OF OBSERVATIONS FOR EACH SPECIES , SEX AND ISLAND.

In [17]:
```
df['species'].value_counts()
```

Out[17]:
```
Adelie       152
Gentoo       124
Chinstrap     68
Name: species, dtype: int64
```

In [18]:
```
df['sex'].value_counts()
```

Out[18]:
```
MALE       168
FEMALE     165
UNKNOWN     11
Name: sex, dtype: int64
```

In [19]:
```
df['island'].value_counts()
```

Out[19]:
```
Biscoe       168
Dream        124
Torgersen     52
Name: island, dtype: int64
```

QUESTION 4.

WHICH SPECIES HAVE THE `HEAVIEST` & `HIGHEST` NUMBER OF MALES ?

In [20]:
```
df2 = df[df['sex'] == 'MALE']  # A DATASET CONTAINS ONLY MALE PENGUINS.
df2.head(11)
```

Out[20]:

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
0	Adelie	Torgersen	39.1	18.7	181.0	3750.0	MALE
4	Adelie	Torgersen	36.7	19.3	193.0	3450.0	MALE
7	Adelie	Torgersen	39.2	19.6	195.0	4675.0	MALE
10	Adelie	Torgersen	39.3	20.6	190.0	3650.0	MALE
14	Adelie	Torgersen	34.6	21.1	198.0	4400.0	MALE
16	Adelie	Torgersen	42.5	20.7	197.0	4500.0	MALE
18	Adelie	Torgersen	46.0	21.5	194.0	4200.0	MALE
20	Adelie	Biscoe	37.7	18.7	180.0	3600.0	MALE
23	Adelie	Biscoe	39.2	19.1	195.0	3900.0	MALE
25	Adelie	Biscoe	38.6	17.2	199.0	3800.0	MALE

In [21]:
```
df2['species'].value_counts()
```

Out[21]:
```
Adelie       73
Gentoo       61
Chinstrap    34
Name: species, dtype: int64
```

- `ADELIE` HAVE THE MOST NUMBER OF MALES.
- `CHINSTRAP` HAVE THE LEAST NUMBER OF MALES.

QUESTION 5.

WHICH SPECIES HAVE THE `HEAVIEST` & `HIGHEST` NUMBER OF FEMALES ?

In [22]:
```
df3 = df[df['sex'] == 'FEMALE']  # DATASET CONTAINS ONLY FEMALE PENGUINS.
df3.head(11)
```

Out[22]:

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
1	Adelie	Torgersen	39.5	17.4	186.0	3800.0	FEMALE
2	Adelie	Torgersen	40.3	18.0	195.0	3250.0	FEMALE
5	Adelie	Torgersen	38.9	17.8	181.0	3625.0	FEMALE
6	Adelie	Torgersen	39.8	17.8	191.0	3625.0	FEMALE
9	Adelie	Torgersen	41.1	17.6	182.0	3200.0	FEMALE
12	Adelie	Torgersen	36.6	17.8	185.0	3700.0	FEMALE
15	Adelie	Torgersen	38.7	19.0	195.0	3450.0	FEMALE
19	Adelie	Torgersen	35.9	19.2	189.0	3800.0	FEMALE
22	Adelie	Biscoe	37.8	18.3	174.0	3400.0	FEMALE
24	Adelie	Biscoe	37.3	19.2	196.0	3450.0	FEMALE

In [23]:
```
df3['species'].value_counts()
```

Out[23]:
```
Adelie       73
Gentoo       61
Chinstrap    34
Name: species, dtype: int64
```

- `ADELIE` HAVE THE MOST NUMBER OF FEMALES.
- `CHINSTRAP` HAVE THE LEAST NUMBER OF FEMALES.

QUESTION 6.

HOW MANY FEMALES ARE HEAVIER THAN THE LIGHTEST MALE ?

In [24]:
```
df2.loc[df2['body_mass_g'] == min(df2['body_mass_g'])]  # GETTING THE WEIGHT OF THE LIGHTEST MALE
```

Out[24]:

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
3236	Chinstrap	Dream	51.5	18.7	187.0	3250.0	MALE

In [25]:
```
df4 = df3.loc[df3['body_mass_g'] > 3250.0]
df4.head(11)  # DATASET OF FEMALE PENGUINS THAT ARE HEAVIER THAN THE LIGHTEST MALE.
```

Out[25]:

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
1	Adelie	Torgersen	39.5	17.4	186.0	3800.0	FEMALE
4	Adelie	Torgersen	36.7	19.3	193.0	3450.0	FEMALE
6	Adelie	Torgersen	39.8	17.8	191.0	3625.0	FEMALE
15	Adelie	Torgersen	38.7	19.0	195.0	3450.0	FEMALE
19	Adelie	Torgersen	35.9	19.2	189.0	3800.0	FEMALE
20	Adelie	Biscoe	37.8	18.3	174.0	3400.0	FEMALE
23	Adelie	Biscoe	39.2	19.1	195.0	3900.0	FEMALE
24	Adelie	Biscoe	37.3	19.2	196.0	3450.0	FEMALE
25	Adelie	Biscoe	35.3	18.9	187.0	3800.0	FEMALE
26	Adelie	Biscoe	39.5	17.8	188.0	3300.0	FEMALE

In [26]:
```
df4.count()
```

Out[26]:
```
species              134
island               134
bill_length_mm       134
bill_depth_mm        134
flipper_length_mm    134
body_mass_g          134
sex                  134
dtype: int64
```

- 134 `FEMALES` ARE HEAVIER THAN THE LIGHEST MALE

QUESTION 7.

NAME THE SPECIES OF MALE PENGUINS WITH THE MAXIMUM `BILL LENGTH , BILL DEPTH AND FLIPPER LENGTH`

In [27]:
```
df2.head()  # DATASET OF MALE PENGUINS.
```

Out[27]:

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
0	Adelie	Torgersen	39.1	18.7	181.0	3750.0	MALE
4	Adelie	Torgersen	36.7	19.3	193.0	3450.0	MALE
7	Adelie	Torgersen	39.2	19.6	195.0	4675.0	MALE
10	Adelie	Torgersen	39.3	20.6	190.0	3650.0	MALE
14	Adelie	Torgersen	34.6	21.1	198.0	4400.0	MALE

In [28]:
```
df2['species'].loc[df2['bill_length_mm'] == max(df2['bill_length_mm'])]
```

Out[28]:
```
213    Gentoo
Name: species, dtype: object
```

In [29]:
```
df2['species'].loc[df2['bill_depth_mm'] == max(df2['bill_depth_mm'])]
```

Out[29]:
```
3236    Chinstrap
Name: species, dtype: object
```

In [30]:
```
df2['species'].loc[df2['flipper_length_mm'] == max(df2['flipper_length_mm'])]
```

Out[30]:
```
213    Gentoo
Name: species, dtype: object
```

- THE SPECIES `GENTOO` HAVE THE MAXIMUM BILL LENGTH AND FLIPPER LENGTH.
- THE SPECIES `ADELIE` HAVE THE MAXIMUM BILL DEPTH.

QUESTION 8.

VISUALIZE THE RELATION BETWEEN BILL LENGTH, BILL DEPTH ,SPECIES AND SEX.

In [39]:
```
sex_explain = 'bill_length_mm', y = 'bill_depth_mm', data = df ,hue = 'sex', col = 'species',col_wrap = 2, height = 5)
```

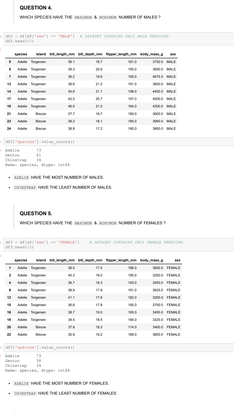
Out[39]:
```
<seaborn.axisgrid.FacetGrid at 0x1944d736550>
```

QUESTION 9.

VISUALIZE THE RELATION BETWEEN FLIPPER LENGTH AND BODY MASS OVER SPECIES. DO HEAVIER PENGUINS TEND TO HAVE LONGER FLIPPERS ?

In [40]:
```
sex_explain = 'flipper_length_mm', y = 'body_mass_g',hue = 'species', data = df, height = 7)
```

Out[40]:
```
<seaborn.axisgrid.FacetGrid at 0x1944d872260>
```



- WE OBSERVE THAT WE HAVE A LINEAR RELATIONSHIP BETWEEN THE FLIPPER LENGTH AND THE BODY MASS. THE LONGER THE FLIPPER OF THE PENGUIN, THE HEAVIER THE PENGUIN.

QUESTION 10.

VISUALIZE AND TELL WHICH ISLAND HAS THE HEAVIEST PENGUINS ? WHICH SPECIES DO THEY BELONG TO ? WHAT ARE THE OTHER INFORMATION YOU CAN GRASP FROM THE PLOT ?

In [41]:
```
sex_explain = 'body_mass_g', y = 'island', data = df, hue = 'species', kind = 'bar', height = 7)
```

Out[41]:
```
<seaborn.axisgrid.FacetGrid at 0x1945cfe7e60>
```

- `BISCOE` ISLAND HAS THE HEAVIEST PENGUINS AND THE SPECIES THOSE PENGUINS BELONG TO IS `GENTOO`.
- OTHER INFORMATION THAT WE CAN GET HERE IS THAT NONE OF THE ISLANDS HAVE ALL THE THREE SPECIES AND `TORGERSEN` ISLAND HAS THE LEAST NUMBER OF SPECIES.