Portfolio assignment 1

- What is Data Science?
  - You will be split up into breakrooms with 3-4 students each.
  - 10 min: Do online research to answer the question "What is Data Science?"*
  - 10 min: Create a PowerPoint slide with a summary of your results*

- Optional: Same assignment but the question is "What does a Data Scientist do?"*

*Do not use Avans school materials as a source for this assignment

Portfolio assignment 2

- What are the most popular data science tools?
  - You will be split up into breakrooms with 3-4 students each.
  - 10 min: Do online research to answer the question "What are the most popular data science tools?"*
  - 10 min: Create a PowerPoint slide with a summary of your results

- Optional: Same assignment but the question is "What are the characteristics that make these tools popular for data science tasks?"*

*Do not use Avans school materials as a source for this assignment

Portfolio assignment 3

15 min: Perform a univariate analysis on all the categorical data of the penguins dataset. Commit the notebook to your portfolio when you're finished. Optional: Start working on portfolio assignment 4

In [28]:
```
penguins = sns.load_dataset("penguins")
```

In [29]:
```
penguins.head()
```

Out[29]:

	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

Portfolio assignment 4

15 min: Look online for a datset that you personally find interesting to explore. It can be about any topic that you find interesting: sports, games, software development, etc. Commit the dataset to your portfolio. You will be analysing the dataset in future portfolio assignments.

Required characteristics of the dataset:

- Must be in a tabular format: Contains rows and columns
- Contains at least 100 rows
- Contains at least 2 columns with categorical data and at least 2 columns with numerical data
- Is less than 200 MB

Portfolio assignment 5

20 min:

- Download lifeExpectancyAtBirth.csv from Brightspace (original source).
- Move the file to the same folder as the Notebook that you will be working in.
- Load the dataset in your Notebook with the following code: lifeExpectancy = pd.read_csv('lifeExpectancyAtBirth.csv', sep=',')
- Look at the dataset with the .head() function.
- Filter the dataframe: We only want the life expectancy data about 2019 and 'Both sexes'
- Use this dataframe to perform a univariate analysis on the life expectancy in 2019.
- Which five countries have the highest life expectancy? Which five the lowest?

Commit the notebook and dataset to your portfolio when you're finished.

Portfolio assignment 6

60 min: Perform a univariate analysis on at least 2 columns with categorical data and on at least 2 columns with numerical data in the dataset that you chose in portfolio assignment 4. Commit the Notebook to your portfolio when you're finished.

Portfolio assignment 7

15 min: Look at the histogram of at least 2 columns with numerical data in the dataset that you chose in portfolio assignment 4. Do you recognise the distribution? Does it look like a uniform or normal distribution or something else? If it doesn't look like a uniform or normal distribution, take a quick look here to see if you can find the distribution shape: https://www.itl.nist.gov/div898/handbook/eda/section3/eda366.htm

Portfolio assignment 8

15 min:

- Calculate the 90%, 95%, 99% and 99.99% confidence interval for at least 2 columns with numerical data in the dataset that you chose in portfolio assignment 4. Do you see the impact the confidence has on the interval?
- Now calculate the 95% confidence interval again but use only the first 10% of your rows. Compare this interval to the previous 95% confidence interval you calculated. Do you see the impact of having less data?

Portfolio assignment 9

25 min: Perform a bivariate analysis on the columns with numerical data in the penguins dataset.

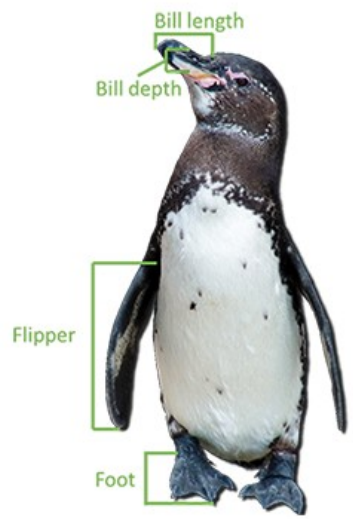
- Use corr() on the DataFrame to calculate all the correlations. Use the code example above to show the correlation table with colors.
- Look at the corrrelations. Do they match your expectations?
- Show a scatter plot for
  - The strongest positive correlation
  - The strongest negative correlation
  - The weakest correlation

In [80]:
```
penguins = sns.load_dataset("penguins")
```

In [84]:
```
penguins.head()
```

Out[84]:

	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



Portfolio assignment 10

15 min: Perform a bivariate analysis (Pearson correlation and scatter plot) on at least 1 combination of 2 columns with numeric data in the dataset that you chose in portfolio assignment 4. Does the correlation and scatter plot match your expectations? Add your answer to your notebook. Commit the Notebook to your portfolio when you're finished.

Portfolio assignment 11

20 min: Do a Numerical VS Categorical bivariate analysis on the penguins dataset.

- Choose one of the categorical columns: species, island or sex
- use .groupby('').mean() too look at the means of the numerical columns. Does it look like there is a difference between categories?
- Use the seaborn barplot to plot the mean and confidence. Create this plot for each of the numerical columns (bill_length_mm bill_depth_mm, flipper_length_mm, body_mass_g)
- For each of the plots, write a conclusion: Is there a statistically significant difference for this numerical column for each category?
- Optional: Repeat this proces for the other two categorical columns

```
penguins.head()
```

	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

Portfolio assignment 12

30 min: Perform a bivariate analysis on at least 3 combinations of a numerical column with a categorical column in the dataset that you chose in portfolio assignment 4. Use *.groupby('columnname').mean()* to calculate the means. Is there a difference between categories? Then use seaborn barplots to check if there is a statistically significant difference.

Portfolio assignment 13 ¶

10 min: Do a bivariate analysis on the penguins dataset for the following combination of columns:

- species VS sex
- island VS sex

For this bivariate analysis, at least perform the following tasks:

- Do you expect their to be a correlation between the two columns?
- Create a contigency table. Do you observe different ratios between categories here?
- Create a bar plot for this contigency table. Do you observe different ratios between categories here?
- Do a chi-squared test. What does the result say? What's the chance of there being a correlation between the two columns?

Portfolio assignment 14

Perform a bivariate analysis on at least 1 combination of 2 columns with categorical data in the dataset that you chose in portfolio assignment 4.

- Do you expect their to be a correlation between the two columns?
- Create a contigency table. Do you observe different ratios between categories here?
- Create a bar plot for this contigency table. Do you observe different ratios between categories here?
- Do a chi-squared test. What does the result say? What's the chance of there being a correlation between the two columns?

Portfolio assignment 15

30 min: Train a decision tree to predict the species of a penguin based on their characteristics.

- Split the penguin dataset into a train (70%) and test (30%) set.
- Use the train set to fit a DecisionTreeClassifier. You are free to to choose which columns you want to use as feature variables and you are also free to choose the max_depth of the tree. **Note**: Some machine learning algorithms can not handle missing values. You will either need to
  - replace missing values (with the mean or most popular value). For replacing missing values you can use .fillna() https://pandas.pydata.org/docs/reference/api/pandas.Series.fillna.html
  - remove rows with missing data. You can remove rows with missing data with .dropna() https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.dropna.html
- Use your decision tree model to make predictions for both the train and test set.
- Calculate the accuracy for both the train set predictions and test set predictions.
- Is the accurracy different? Did you expect this difference?
- Use the plot_tree_classification function above to create a plot of the decision tree. Take a few minutes to analyse the decision tree. Do you understand the tree?

Optional: Perform the same tasks but try to predict the sex of the pinguin based on the other columns

In [19]:
```
penguins = sns.load_dataset("penguins")
penguins.head()
```

Out[19]:

	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

Portfolio assignment 16

30 min: Train a decision tree to predict one of the categorical columns of your own dataset.

- Split your dataset into a train (70%) and test (30%) set.
- Use the train set to fit a DecisionTreeClassifier. You are free to to choose which columns you want to use as feature variables and you are also free to choose the max_depth of the tree.
- Use your decision tree model to make predictions for both the train and test set.
- Calculate the accuracy for both the train set predictions and test set predictions.
- Is the accurracy different? Did you expect this difference?
- Use the plot_tree function above to create a plot of the decision tree. Take a few minutes to analyse the decision tree. Do you understand the tree?

Portfolio assignment 17

30 min: Train a decision tree to predict the body_mass_g of a penguin based on their characteristics.

- Split the penguin dataset into a train (70%) and test (30%) set.
- Use the train set to fit a DecisionTreeRegressor. You are free to to choose which columns you want to use as feature variables and you are also free to choose the max_depth of the tree. **Note**: Some machine learning algorithms can not handle missing values. You will either need to
  - replace missing values (with the mean or most popular value). For replacing missing values you can use .fillna() https://pandas.pydata.org/docs/reference/api/pandas.Series.fillna.html
  - remove rows with missing data. You can remove rows with missing data with .dropna() https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.dropna.html
- Use your decision tree model to make predictions for both the train and test set.
- Calculate the RMSE for both the train set predictions and test set predictions.
- Is the RMSE different? Did you expect this difference?
- Use the plot_tree_regression function above to create a plot of the decision tree. Take a few minutes to analyse the decision tree. Do you understand the tree?

In [178]:
```
penguins = sns.load_dataset("penguins")
penguins.head()
```

Out[178]:

	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

Portfolio assignment 18

30 min: Train a decision tree to predict one of the numerical columns of your own dataset.

- Split your dataset into a train (70%) and test (30%) set.
- Use the train set to fit a DecisionTreeRegressor. You are free to to choose which columns you want to use as feature variables and you are also free to choose the max_depth of the tree.
- Use your decision tree model to make predictions for both the train and test set.
- Calculate the RMSE for both the train set predictions and test set predictions.
- Is the RMSE different? Did you expect this difference?
- Use the plot_tree function above to create a plot of the decision tree. Take a few minutes to analyse the decision tree. Do you understand the tree?

Portfolio assignment 19

30 min: Create a cluster model on the penguins dataset.

- Use the pairplot() function on the penguins dataset. Do you visually notice any clusters? How many clusters do you think there are?
- Use the KMeans algorithm to create a cluster model. Apply this model to the dataset to create an extra column 'cluster' just like we did for the iris dataset above.
  Note: Some machine learning algorithms can not handle missing values. You will either need to
  - replace missing values (with the mean or most popular value). For replacing missing values you can use .fillna() https://pandas.pydata.org/docs/reference/api/pandas.Series.fillna.html
  - remove rows with missing data. You can remove rows with missing data with .dropna() https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.dropna.html
- Calculate the Silhouette Coefficient for your clustering. Play around with the features and n_clusters to search for better results. Keep the cluster model with the highest Silhouette Coefficient.
- Use the pairplot(hue='cluster') function to observe how the model has clustered the data.
- We know the species of each penguin. Use a contingency table to reveal the relation between the cluster results and the species. Is there an exact match? Are there species which ended up in the same cluster? If so, what does it mean that they ended up in the same cluster?

In [48]:
```
penguins = sns.load_dataset("penguins")
penguins.head()
```

Out[48]:

	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

Portfolio assignment 101

- Looking back at 'What is Data Science?'
  - 10 min: Look at your result for portfolio assignment 1 'What is Data Science?'. Is there anything you would like to change? If so then create a second version and add it to your portfolio.
  - 30 min: Look back at all the portfolio assignments you have done. Create a short report in which you:
    - Use the portfolio assignments as examples to explain what Data Science is.
    - Optional: Use the portfolio assignments as examples to explain the relation between Data Science and BI
    - Optional: Use the portfolio assignments as examples to explain the relation between Data Science and AI

Portfolio assignment 102

- 30 min: Look back at all the portfolio assignments you have done. Create a short report in which you use the portfolio assignments as examples to explain the tasks and process of a Data Scientist.

Portfolio assignment 103

- 30 min: Export all your Notebooks to PDF or HTML. For each Jupyter Notebook in your portfolio:
  - Run all code: Cell -> Run all
  - Make sure there are no errors
  - Export to PDF or HTML: File -> Download as -> PDF or HTML
  - Add the PDF or HTML to your portfolio

- Have you finished all your portfolio assignments? Zip your portfolio and hand it in on Blackboard -> Toetsen -> Submission portfolio