Penguins Analysis

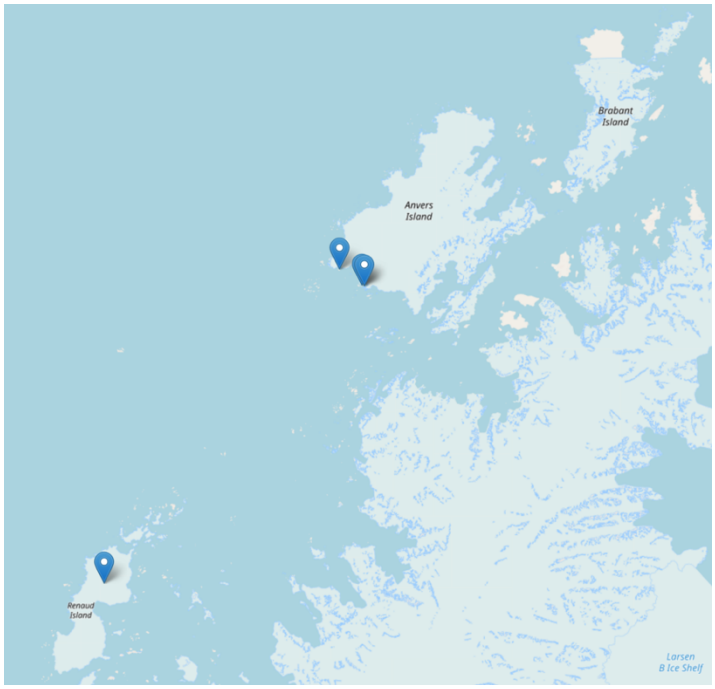
All code with description can be found at: https://github.com/izsolnay/Penguins_Python

Objectives

- To build a potent classifier model which can predict penguin future expectations (body mass) for sustainability requirements, specifically feeding.
- To perform statistical analyses on penguin data from three penguin species and build a clustering model to confirm partitions of data.

Data

This project uses the Palmer Penguins data set collected and made available by Dr. Kristen Gorman and Palmer Station, Antarctica, LTER (https://allisonhorst.github.io/palmerpenguins). The penguins included are from the Antarctica islands close to Terra del Fuego known as the Biscoe Islands.



Deliverables

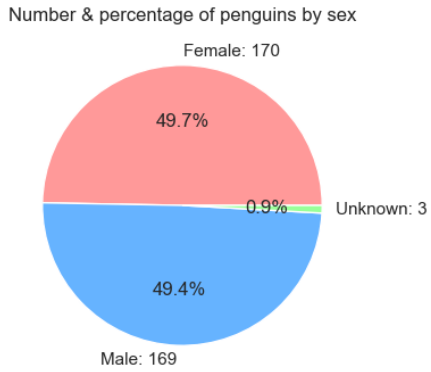
- In depth statistical analyses and visualizations of penguin data set
- Simple linear regression model to test whether bill length can significantly predict body mass
- Multiple linear regression model to test whether bill length, sex, and species can significantly predict body mass with two-way ANOVA and Tukey's Honestly Significant Difference (HSD) tests
- K-means partitioning model to provide insights into the underlying structure within the dataset and confirm if a clustering by sex and species is most expedient

Statistical analyses

The Penguin data set contains 344 observations and 7 variables:

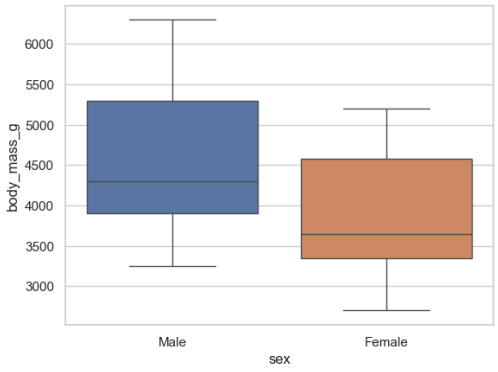
- 3 *object*: species, island, and sex
    - species: Adelie, Chinstrap, and Gentoo
    - island: Biscoe, Dream, and Torgersen
    - sex: Female and Male
        - The variable sex is missing 11 values
- 4 *continuous float*: bill_length_mm, bill_depth_mm, flipper_length_mm, and body_mass_g
    - These are all missing 2 values each

Once a thorough EDA was performed, the division of the 342 penguins contained in the data set by sex is:

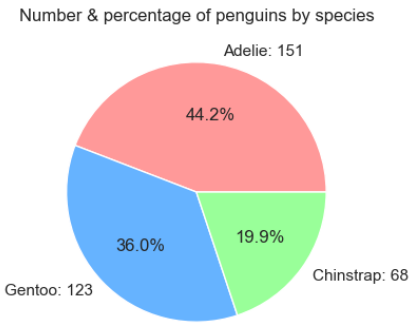


- 170 females (49.7%)
- 169 males (49.4%)
- with 3 observations (0.9%) missing a sex label

On the whole, the body mass between these male and female penguins is evident, with female penguins weighing on average 3.87 kilograms and male penguins 4.54 kilograms. However, as can be seen in the boxplot, body mass skews heavily to the left indicating more variability amongst the heavier birds.



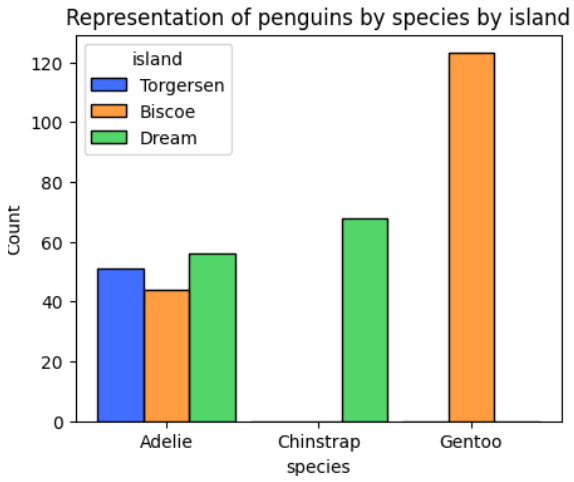
The division between the three species penguins in the data set is:



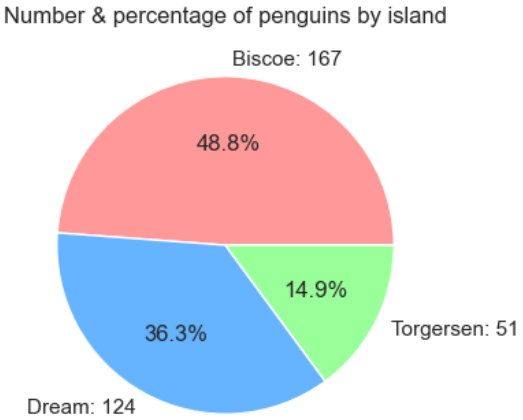
- 151 Adelie (44.2%)
- 123 Gentoo (36.0%)
- 68 Chinstrap (19.9

The distribution of each species of penguins between three islands in the data set is:

- Adelie penguins
  - Torgersen 51 (33.8%)
  - Biscoe 44 (29.1%)
  - Dream 56 (37.1%)
- Chinstrap penguins
  - Dream 68 (100%)
- Gentoo penguins
  - Biscoe 123 (100%)



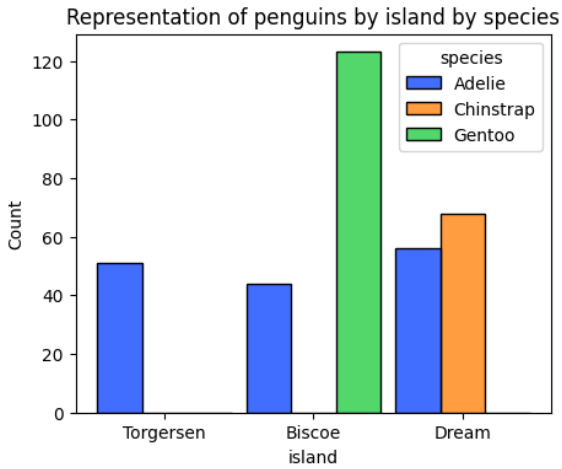
The distribution of penguins on each of the three islands in the data set



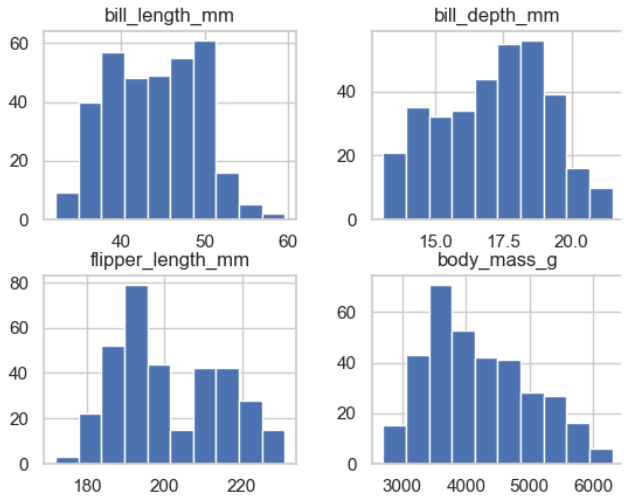
- 168 from Biscoe (48.8%)
- 124 from Dream (36.3%)
- 52 from Torgersen (14.9%)

This breaks down to:

- Torgersen island
  - Adelie 51 (100%)
- Biscoe island
  - Gentoo 123 (73.7%)
  - Adelie 44 (26.3%)
- Dream island
  - Chinstrap 68 (54.8%)
  - Adelie 56 (45.2%)



Body measurements for all penguins, regardless of species or sex



bill_length_mm has a mostly normal distribution

- 32.10 mm > 59.60 mm
- median: 44.45 mm

bill_depth_mm has a mostly normal distribution, though it skews right

- 13.10 mm > 21.50 mm
- median: 17.30 mm

flipper_length_mm has a curious distribution

- 172.00 mm > 231.00 mm
- median: 197.00 mm

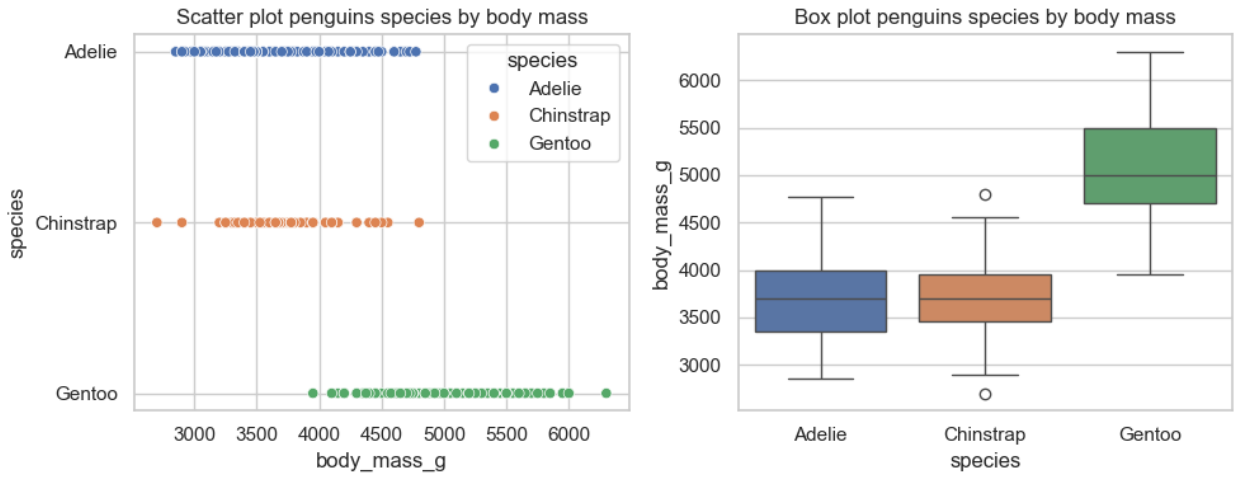
body_mass_g has a left skewed distribution

- 2700 g > 6300 g
- median: 4050

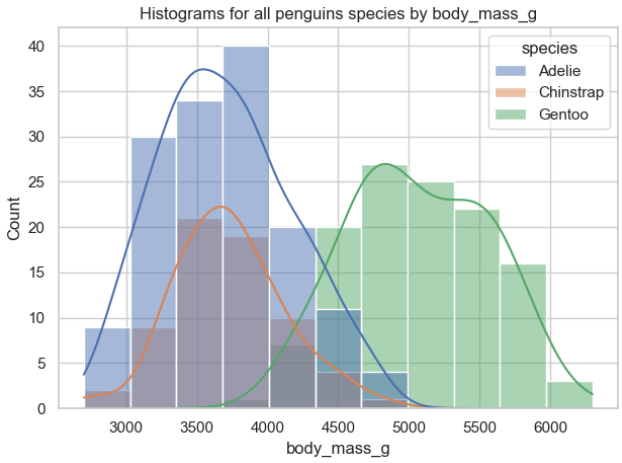
Measurement distributions by species by sex

To measure the penguin's data set by sex using the most observations, sex was assigned to 6 of the 9 penguin observations which had no assigned sex: 5 female and 1 male. These assignments were based on the minimum body masses of Adelie and Gentoo male penguins and, one by the maximum mass of a female Adelie penguin. This then provided 339 observations.

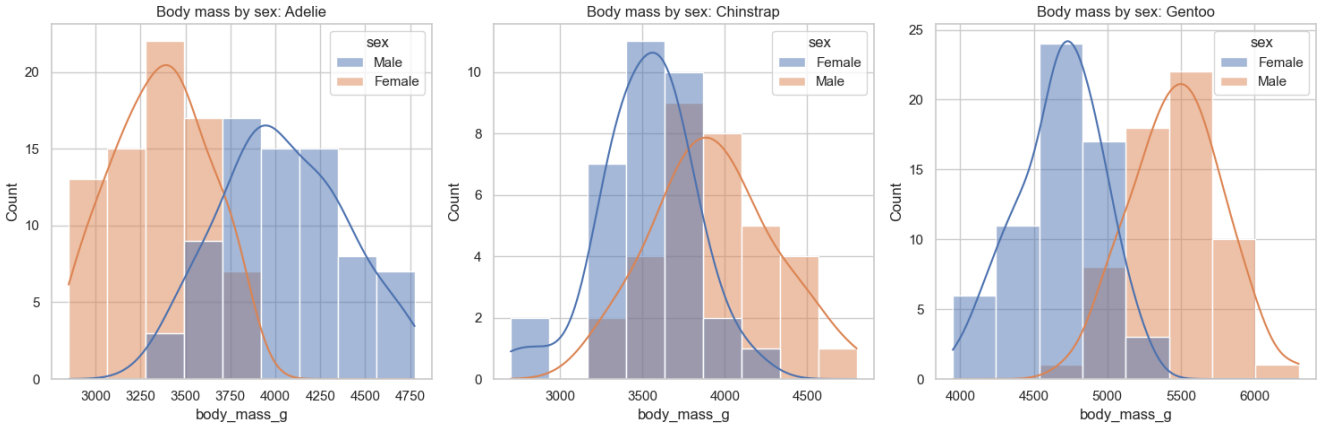
body_mass_g



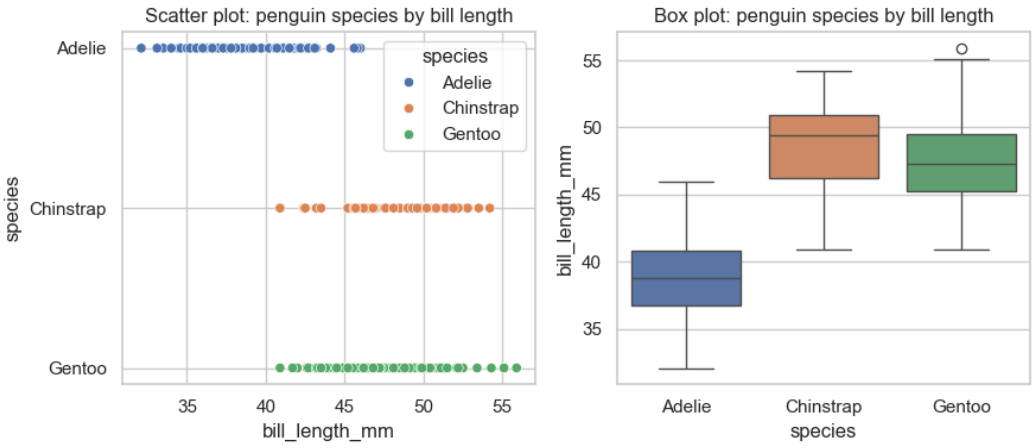
The Chinstrap group has some of the smallest of penguins by mass in the data set, and Gentoos the largest. The smallest Gentoos are of similar mass to the larger Adelie and Chinstraps.



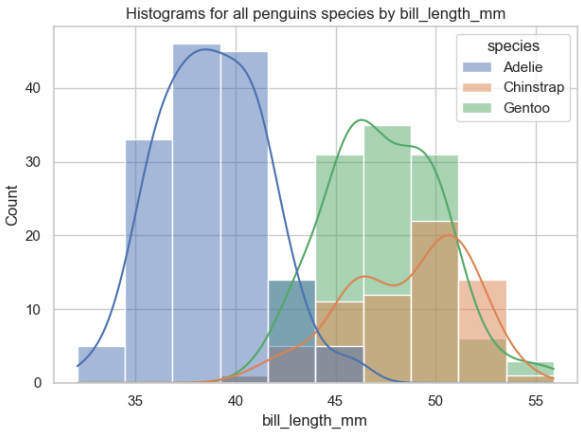
By sex and species, it can easily be seen that male penguins of all species have significantly more body mass:



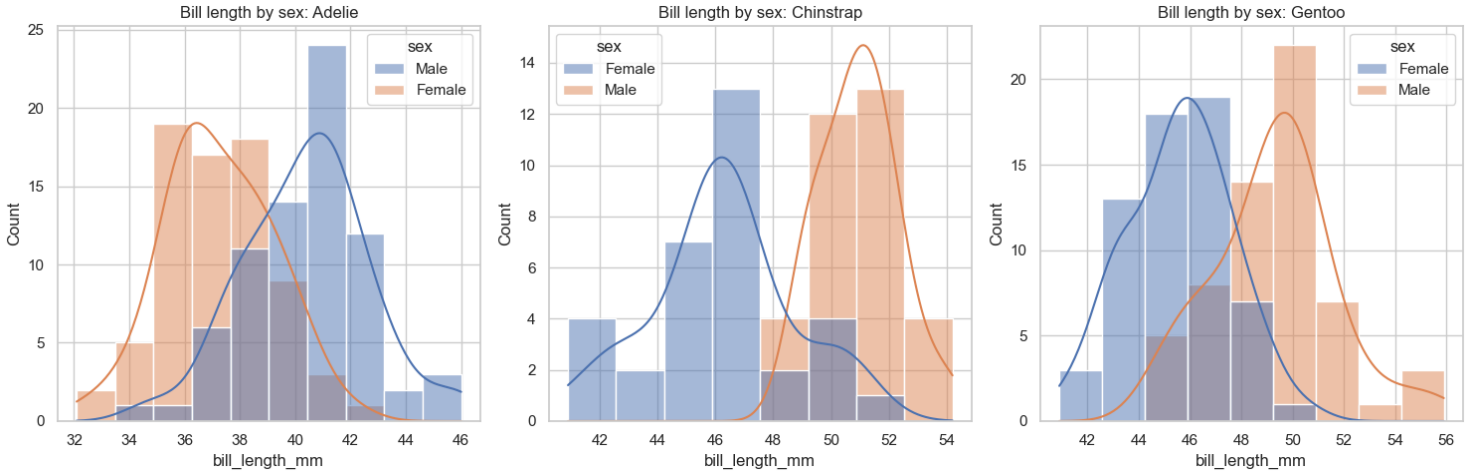
bill_length_mm



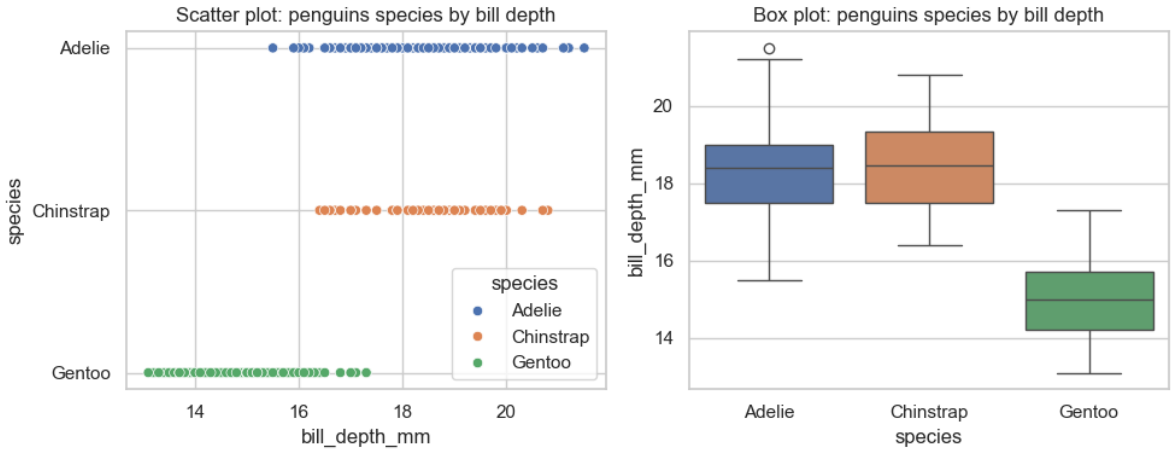
Adelies have the smallest bill lengths. Somewhat surprisingly, Chinstraps and Gentoo have comparable bill lengths. This is surprising because of their great differences in body mass. All the more so, since the median Chinstraps have the longest bills of all species.



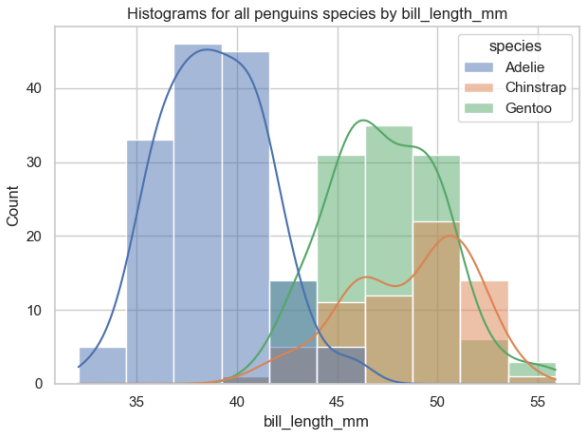
Although, when accounting for sex, by species the median male penguin has a longer bill length, female penguin's, particularly the Chinstraps, can have bills as long as their male counterparts.



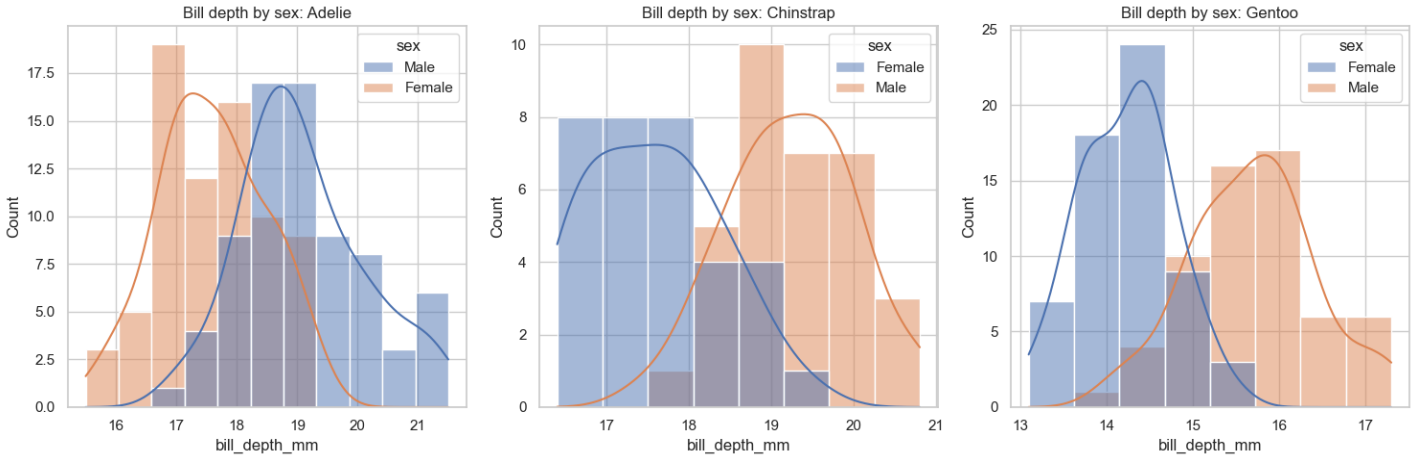
bill_depth_mm



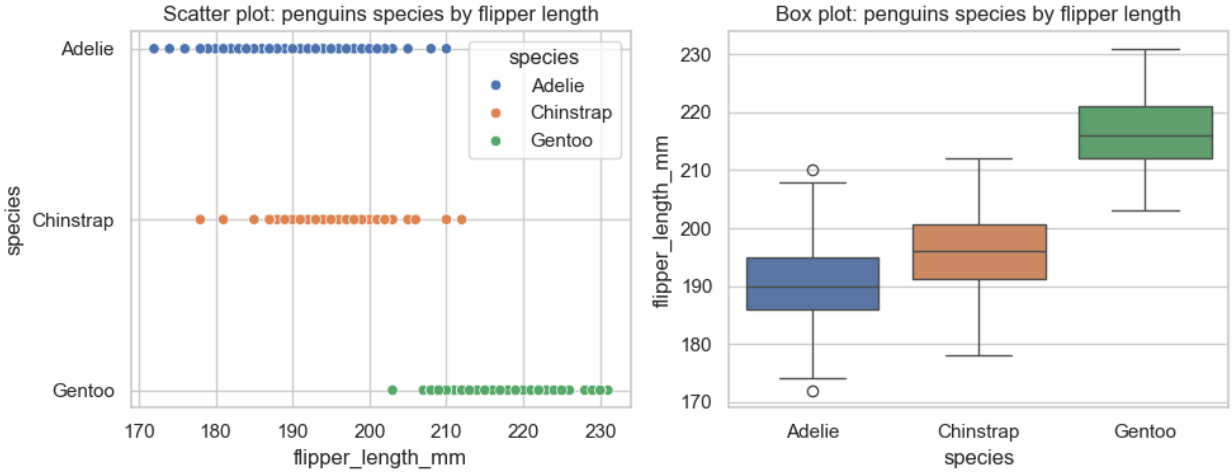
Also surprising, even though they have the most mass, Gentoos have the smallest bill depth by far, with Adelies having slightly more bill depth than Chinstraps.



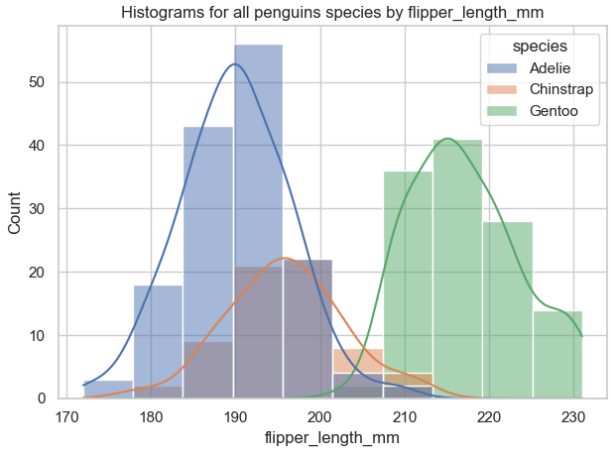
By sex and species, it can easily be seen that male penguins of all species have significantly more bill depth:



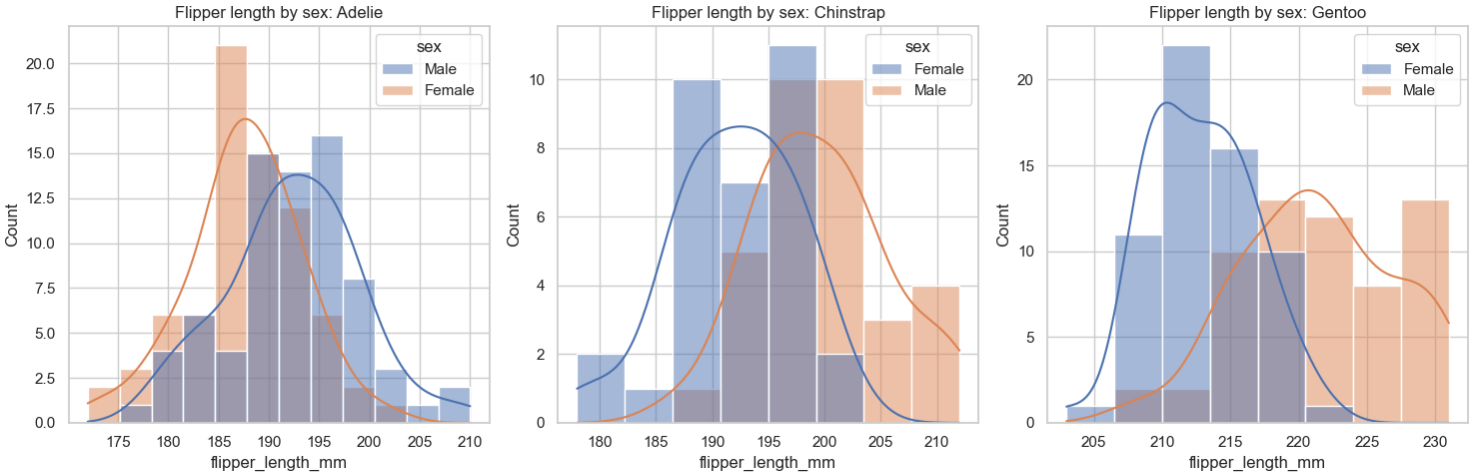
flipper_length_mm



Not surprising, Gentoo, the penguins with the largest body mass, have the longest flippers. Adelies have examples of penguins with some of the smallest flippers, but were essentially matched in flipper length with Chinstraps.



When accounting for sex by species, there is a bit of variety in flipper length. Male Gentoo have flippers significantly longer than female Gentoos, however, female Chinstraps and Adelies can have flippers almost as long as their male counterparts:



Simple linear regression

Data

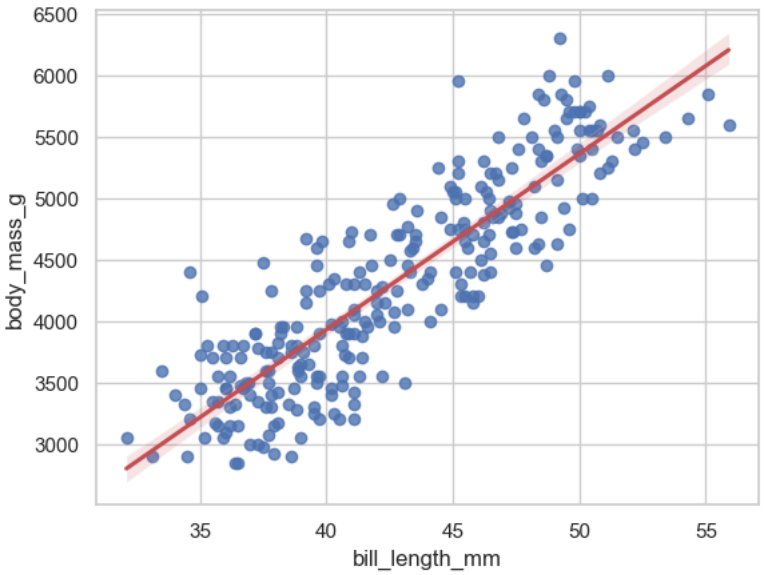
The cleaned Penguin data set contains 335 observations and 7 variables (see above for distributions):

- 3 *object*: species, island, and sex
    - species: Adelie, Chinstrap, Gentoo
    - island: Biscoe, Dream, Torgersen
    - sex: Female, Male
- 4 *continuous float*: bill_length_mm, bill_depth_mm, flipper_length_mm, and body_mass_g
- No missing variables

Results of simple linear regression

A simple linear regression model was built to test whether a penguin's bill length could significantly predict body mass. The OLS model was built using Adelie and Gentoo penguins where the variable body_mass served as the dependent y and bill_length_mm served as the independent X. Data on the Chinstrap penguins was removed for the model because far fewer Chinstraps are present in the data set, and their inclusion could skew the model. 269 observations in total were used to train the model.

The coefficients arrived at by the model indicate that there is a positive correlation between bill length and body mass with the best fit line having an intercept of -1793.48 and a slope of 143.10. Thus, there is a 95% chance that the interval 133.60 to 152.61 contains the true parameter value of the estimated slope (143.10) – see this confidence band visualized around the best-fit line below:



Furthermore, the overall regression was statistically significant with an R^2 of 0.767, and adjusted R^2 of 0.766, and a p-value of <0.05. These reading indicate that there is a correlation between bill length and body mass and that:

- ~77% of body mass can be explained by (correlated to) a variance in bill length between penguins
- 23% of that variance is left unexplained

Conclusion

This model estimates that:

- On average, a penguin of either the Gentoo or Adelie penguin species, regardless of sex, with 1 mm longer bill length is expected to have 143 grams higher body mass
- ~77% of this body mass is explained by bill length
- ~23% of the Gentoo or Adelie penguin species body mass must be explained by other factors

Multiple linear regression

Data

The cleaned Penguin data set contains 335 observations and 7 variables (see above for distributions):

- 3 *object*: species, island, and sex
    - species: Adelie, Chinstrap, Gentoo
    - island: Biscoe, Dream, Torgersen
    - sex: Female, Male
- 4 *continuous float*: bill_length_mm, bill_depth_mm, flipper_length_mm, and body_mass_g
- No missing variables

Results of multiple linear regression

A multiple linear regression model was built to test whether bill length, sex, and species could significantly predict body mass. The OLS model was built using the data of all three penguin species: Adelie, Chinstrap, and Gentoo. The variable body_mass served as the dependent y and bill_length_mm, sex, and species served as the independent X variables. 234 observations were used to train the model and 101 to test it.

The model uncovered valuable information that sheds light on the relationships between various variables and body mass within different penguin species:

1. Male penguins are predicted to have a body mass approximately 515g greater than female penguins of the same species and bill length. This difference is statistically significant with a p-value of 0.000 and a confidence interval of [412.60g, 617.33g].

2. Chinstrap penguins are predicted to have a body mass approximately 260g less than Adelie penguins with the same bill length and sex. This difference is statistically significant with a p-value of 0.019 and a confidence interval of [-478.64g, -42.30g].

3. Gentoo penguins are predicted to have a body mass approximately 1090g greater than Adelie penguins with the same bill length and sex. This difference is statistically significant with a p-value of 0.000 and a confidence interval of [900.89g, 1278.40g].

4. For each 1mm increase in bill length, a penguin is predicted to have a body mass approximately 33g greater, assuming the penguins are of the same sex and species. This relationship is statistically significant with a p-value of 0.001 and a confidence interval of [14.29g, 51.86g].

Furthermore, the adjusted R^2 value of 0.842 and an adjusted R^2 of 0.840 validates the robustness of the model in explaining the variance in body mass across penguin species.

The model also has a high VIF, which indicates a high multicollinearity between the variables. Multicollinearity is expected, since it is reasonable to assume, for example, the larger a penguin's body mass, the larger it's bill length would be expected to be.

Further tests

In addition to the regression analysis, a two-way ANOVA and post hoc analyses were conducted.

Two-way ANOVA test

The results of the two-way ANOVA found that all p-values were statistically significant. This indicates that all predictors included in the model have a significant effect on penguin body mass.

The null and alternative hypotheses for the variable sex:

- H_0: μfemale = μmale
- H_1: μfemale ≠ μmale

The null and alternative hypotheses for the variable species:

- H_0: μAdelie = μChinstrap = μGentoo
- H_1: μAdelie ≠ μChinstrap ≠ μGentoo

The two-way ANOVA also observed a high F-test statistic of 98.25 for the sex group *means*, suggesting that the variability between the sexes is larger than the variability within each group. Similarly, a high F-test statistic of 340.86 was found for the species group *means*, indicating that the variability between the species is much larger than the variability within each group. For example, Gentoos are much larger birds than both Chinstraps and Adelies. The F-test statistic is a measure that compares the variability between group *means* (the effect of the predictors) to the variability within each group (residual variability).

These results suggest that at least one predictor in the model has a significant effect on the dependent variable, penguin body mass.

HSD Post hoc test

The results of the post hoc analysis using Tukey's Honestly Significant Difference (HSD) test with a Family-Wise Error Rate (FWER) of 0.05 indicated that there is no statistically significant difference in body mass between Adelie and Chinstrap penguins, with a *mean* difference of 28.70 and a p-value of 0.908. This finding suggests that, based on the data analyzed, the body mass of Adelie and Chinstrap penguins does not differ significantly.

The HSD post hoc test further indicated that there is a statistically significant difference in body mass between Adelie and Gentoo penguins and Chinstrap and Adelie Penguins:

- There is a significant difference in body mass between Adelie and Gentoo penguins, with a *mean* difference of approximately 1369 and a p-value of 0.00.
- There is also a significant difference in body mass Chinstrap Adelie and Gentoo penguins, with a *mean* difference of approximately 1340 and a p-value of 0.00.

A second post hoc analysis using Tukey's HSD test indicated that there is a statistically significant *mean* difference of ~672.55 in body mass between female and male penguin groups with a p-adj value of 0.0.

These HSD post hoc tests confirm the results of the two-way ANOVA.

Further HSD Post hoc test

Finally, a comprehensive HSD post hoc test was performed on all groups.
Key findings are:

- Adelie_Female vs. Adelie_Male:
  - A significant mean difference of 690.88 was identified
- Adelie_Female vs. Chinstrap_Female:
  - No statistically significant difference was found
- Adelie_Female vs. Chinstrap_Male:
  - A significant mean difference of 581.72 was identified
- Adelie_Female vs. Gentoo_Female:
  - A significant mean difference of 1315.09 was identified
- Adelie_Female vs. Gentoo_Male:
  - A significant mean difference of 2120.01 was identified
- Adelie_Male vs. Chinstrap_Female:
  - A significant mean difference of -524.31 was identified
- Adelie_Male vs. Chinstrap_Male:
  - No statistically significant difference was found
- Adelie_Male vs. Gentoo_Female:
  - A significant mean difference of 624.21 was identified
- Adelie_Male vs. Gentoo_Male:
  - A significant mean difference of 1429.13 was identified
- Chinstrap_Female vs. Chinstrap_Male:
  - A significant mean difference of 415.15 was identified
- Chinstrap_Female vs. Gentoo_Female:
  - A significant mean difference of 1148.52 was identified
- Chinstrap_Female vs. Gentoo_Male:
  - A significant mean difference of 1953.48 was identified
- Chinstrap_Male vs. Gentoo_Female:
  - A significant mean difference of 733.37 was identified
- Chinstrap_Male vs. Gentoo_Male:
  - A significant mean difference of 1538.30 was identified
- Gentoo_Female vs. Gentoo_Male:
  - A significant mean difference of 804.92 was identified

A note on seeming discrepancies

While the findings from the regression analysis and post hoc testing provide valuable insights, it is crucial to acknowledge and explain the differences in the conclusions drawn from these analyses.

Regression Analysis vs. Post Hoc Testing

In the regression analysis, specific relationships were identified between variables such as sex and species with body mass, indicating that male penguins tend to have a higher body mass compared to females, and Adelie penguins exhibit a higher body mass than Chinstrap penguins. Although the post hoc analysis using Tukey's HSD test revealed confirmed that there is a statistically significant body mass differential between female and male penguins, it also concluded that there was no statistically significant difference in body mass between Adelie and Chinstrap penguins.

The disparities in the results can be attributed to the distinct methodologies employed in each analysis. The regression analysis focuses on estimating the effects of individual variables on body mass within a model, while the post hoc testing compares *mean* differences between groups. These differing approaches can lead to nuanced interpretations of the data and may result in varying conclusions.

K-means Clustering model

A K-means clustering model was built to discover if a clustering model would partition by sex and species, offering further confidence that this would be the best manner by which to plan penguin care. The six cluster K-means model was built using the cleaned Penguin data set. The cleaned Penguin data set contains

Data

335 observations and 7 variables (see above for distributions):

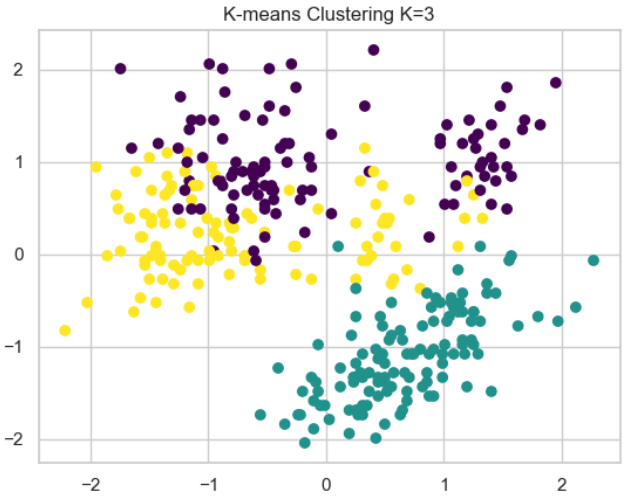
- 3 *object*: species, island, and sex
    - species: Adelie, Chinstrap, Gentoo
    - island: Biscoe, Dream, Torgersen
    - sex: Female, Male
- 4 *continuous float*: bill_length_mm, bill_depth_mm, flipper_length_mm, and body_mass_g
- No missing variables

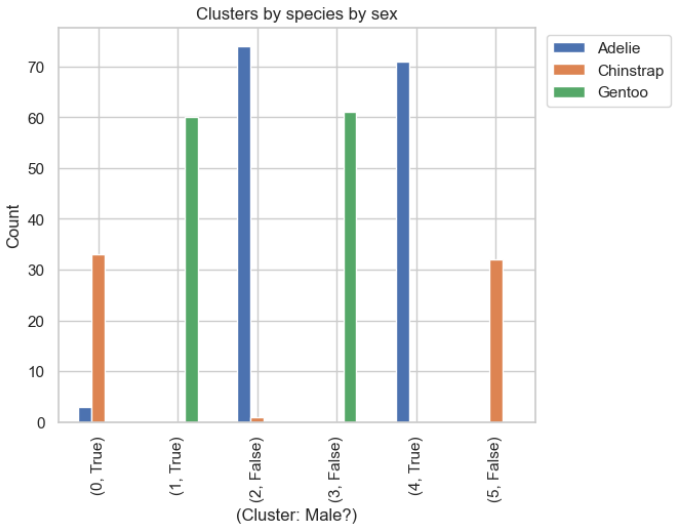
For the model, the unnecessary categorical variable island was dropped in addition to species. The variable island offered no information specific to penguin body measurements and the variable species could have had the potential to bias the model.

Results of K-means model with 6 clusters

The K-mean model with 6 clusters correctly assigned sex and species to 331 out of 335 observations. The model correctly assigned all sex designations. It made 4 species errors. The model misplaced 3 male Adelies

clustering them with the male Chinstraps and misplaced 1 female Chinstrap placing her with the female Adelies:

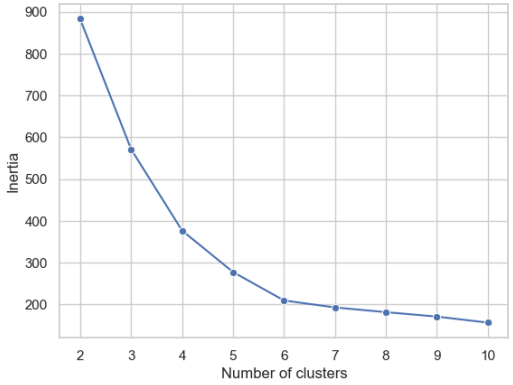




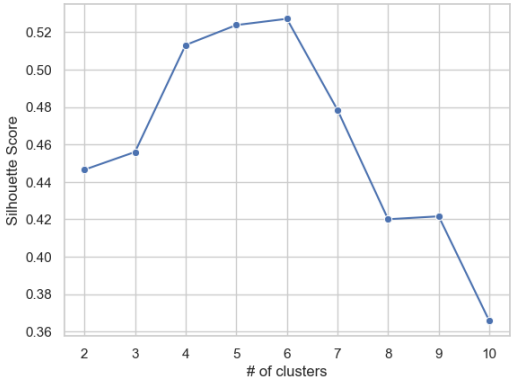
Several tests were performed to evaluate the 6-cluster model:

1. Inertia Analysis:

Inertia, or within-cluster sum of squares, was used to evaluate the effectiveness of the clustering. An elbow plot was created to identify the optimal number of clusters based on the point where inertia starts to decrease at a slower rate. The plot show confirmation at 6 clusters with a score of 209.65983912770113.

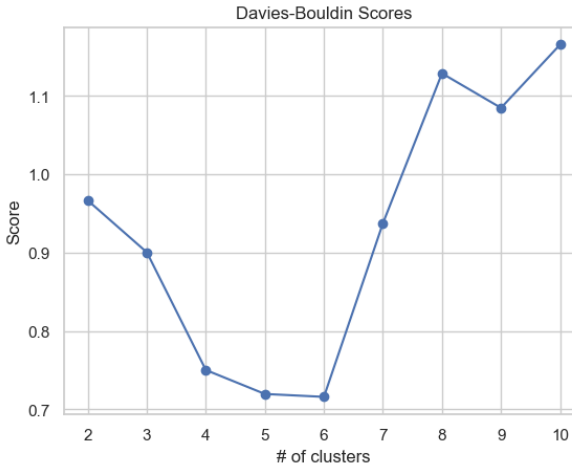


2. Silhouette Score:



The silhouette score was calculated to assess the quality of the clusters formed by the model. A higher silhouette score indicates that the samples are well matched to their own clusters and poorly matched to neighboring clusters. The plot confirms 6 clusters with a score of 0.5271813454904477.

3. Davies-Bouldin Index Score:



Davies-Bouldin Index score was used to evaluate the average similarity between each cluster and its most similar cluster. A lower Davies-Bouldin Index score indicates better clustering. The plot confirms 6 clusters with a score of 0.7159389140028006.

4. Gap Statistics: Gap statistics were employed to compare the clustering quality against a null reference distribution of the data. It helps in determining the optimal number of clusters. The optimal Gap statistic arrived at was variable, either 6 or 8 or 9.

Conclusion

The analysis delivered:

- An in-depth statistical analyses and visualizations of the penguin data set
- A simple linear regression model to test whether bill length can significantly predict body mass
- A multiple linear regression model to test whether bill length, sex, and species can significantly predict body mass with two-way ANOVA and Tukey's Honestly Significant Difference (HSD) tests
- A K-means partitioning model to provide insights into the underlying structure within the dataset and confirm if a clustering by sex and species is most expedient

Both the simple and multiple linear models uncovered valuable information that sheds light on the relationships between various variables and body mass within different penguin species which can help in their future caretaking.

The simple linear regression model ascertained that on average, a penguin of either the Gentoo or Adelie penguin species, regardless of sex, with 1 mm longer bill length is expected to have 143 grams higher body mass. This conclusion was refined to 33g (with a confidence interval of [14.29g, 51.86g]) by the multiple linear regression which allowed for a deeper dive into the data permitting the inclusion of the Chinstrap observations.

The multiple linear regression model also established that male penguins exhibit, on average, a higher body mass compared to their female counterparts, as indicated by a positive coefficient of 515 for the variable

male. This finding underscores the importance of considering sex differences when studying penguin biology and ecology.

The species of penguins also plays a significant role in determining body mass. Specifically, Chinstrap penguins exhibit a lower body mass compared to Adelie penguins, as evidenced by a coefficient of -260 for the variable C(species)[T.Chinstrap]. However, the post hoc test comparing the means of the two species resulted in a ~29g difference. This suggests that either more data is needed, other factors might be in play such as the skew in weight distribution. Conversely, Gentoo penguins tend to have a higher body mass than Adelie penguins, with a coefficient of 1090 for the variable C(species)[T.Gentoo]. These insights into species-specific body mass variations can inform conservation efforts and research initiatives focused on penguin populations.

Finally, the K-means6 clustering model portioned the dataset by sex and species indicating that in allocating resources, a one size fits all approach may not be the best tactic. Expectations as laid out by the regression models must be taken into account with any increase in penguin population.

These results provide valuable insights into the factors indicating body mass in penguins, offering a deeper understanding of penguin biology and ecology. As we continue to explore and analyze data in this field, we remain committed to leveraging these findings to drive informed decision-making and strategic initiatives.