DANMARKS TEKNISKE UNIVERSITET

02450 – Introduction to Machine Learning and Data Mining

PROJECT 1

Martin Mikšík s212075

Lukáš Málek s212074

[1]

Table of Contents

1 The goal of the project 1
1.1 Data set reference 1
1.2 Previous work on the data set 1
1.3 The problem of interest of this project 1
2 Description of the data set 2
2.1 Missing data 2
3 Visualisation 3
3.1 Box plot 3
3.2 Column Histogram 4
3.3 Pearson correlation 4
3.4 Visualisation 5
4 PCA analysis 6
4.1 Variance without standardisation 6
4.2 Variance with Standardisation 6
4.3 Accuracy 7
4.4 Component Coefficients 7
4.5 PCA Visualization 7
5 Summary 8
6 Exam Problems 9
6.1 Question 1 9
6.2 Question 2 9
6.3 Question 3 9
6.4 Question 4 9
6.5 Question 5 9
6.6 Question 6 9
7 References 10

[1] Art work by Allison Horst [9]

1 The goal of the project[2]

The main goal of this project is to perform further analysis of the dataset in Python with the help of 02250_Toolbox [1], exercises and lectures from Week1 till Week4 [2].
The project is expected to deepen our understanding of data mining, data visualisation and ML algorithms in Python.

This project is a cooperation of two students who directly collaborate. The percentage contributions of the students are referred to in a footnote for each task.

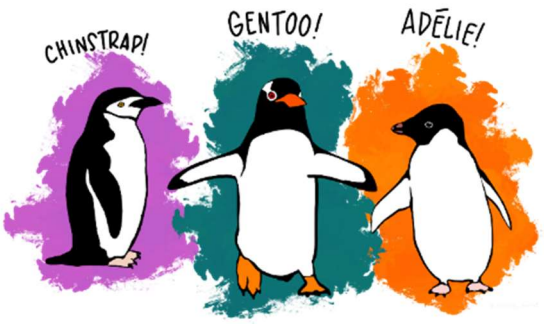
1.1 Data set reference

We have chosen Palmer Penguins Dataset [3] formatted as CSV with data provided by Palmer Station Antarctica LTER [4]. [5] The data set contains seven attributes of three penguin species observed in Antarctica. To be concrete, island and year of observation, bill length and bill depth [mm], body mass [g] and sex.

1.2 Previous work on the data set

For example, an article by Alison Horst [6] introduces and describes the chosen dataset. The report visualises the data on three different graphs. The first graph is a linear regression of bill length and flipper length differentiated by species. The second observation is a scatter plot of flipper length and body mass, and the last one is the column chart of flipper length and frequency for each species.

From the plots, we can summarise that the flipper length is the most suitable for differentiating the species using only one attribute. When using two attributes, the flipper and bill length are a reasonable choice since the first graph has only a few values that would be wrongly qualified when using, for example, the nearest-neighbour algorithm.



1.3 The problem of interest of this project

Our goal for this project is to analyse further the data set compared to the reports currently available. The goal is to achieve a complex understanding of the data and find the relationships between attributes.

Using linear regression, we will predict the depth of the bill for one of the species using all available ratio attributes.

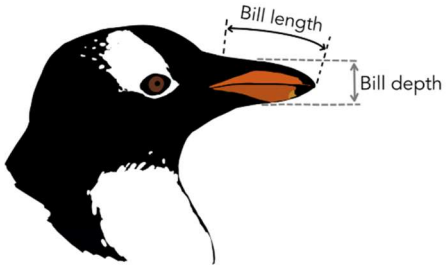
[2] Martin 60%, Lukáš 40%

For the classification problem, we will try to estimate missing sex variables based on all other variables except year and island, as these variables do not carry meaningful information for this task.

No data transformation will be used for both problems, except standardisation.

To complete these tasks, we analyse and visualise the data to conclude suitable algorithms and methods. The classification tree with the Hunts algorithm adjusted for continuous variables could be one of the options.

2 Description of the data set[3]



[4]

Figure 1: Picture describing the difference between bill length and bill depth

No outliers are expected in the dataset, as the range of values is reasonable around the median value for all ratio attributes.

	Row ID	Species[5]	Island[6]	Bill Length [mm]	Bill Depth [mm]	Flipper length [mm]	Body mass [g]	Sex	Year
Data	Discrete	Discrete	Discrete	Continuous	Continuous	Continuous	Continuous	Discrete	Discrete
Attribute	Nominal	Nominal	Nominal	Ratio	Ratio	Ratio	Ratio	Nominal	Interval
Missing Values	No	No	No	2x N/A	2x N/A	2x N/A	2x N/A	11x N/A	No
Mean	-	0	Biscoe	43.922	17.1512	200.9152	4201.7544	-	2008
STD[7]	-	-	-	5.4596	1.9748	14.0617	801.9545	-	-
Median	-	-	-	44.4500	17.3000	187.0000	4050.000	-	-
Range	1-345	0-2	1-3	32.1-59.6	13.1-21.5	172-231	2700-6300	M/F	2007-2009

Table1: Description of the attributes and basic statistics of the data set.

2.1 Missing data

There are two instances in the data set with multiple missing attributes crucial in further data analysis. Therefore, these two instances were removed entirely from the data set analytics, resulting in 342 penguin instances in total.

[3] Martin 60%, Lukáš 40%

[4] Art by Allison Horst [9]

[5] Corresponding statistical values reffers to a dictionary: {'Adelie': 0, 'Chinstrap': 1, 'Gentoo': 2}

[6] Corresponds to three islands: Biscoe, Dream and Torgersen

[7] Standard Deviation

In the remaining 342 instances, there are Nan values present nine times in the sex attribute. All unknown data will be predicted during the classification algorithm. Nevertheless, these instances are pretermitted for visualisation of the dataset, leading us to the final number of 333 penguin instances used for visualising with no missing values.

3 Visualisation [8]

This section will analyse the data set using a box plot, a histogram, and a Pearson correlation. [7] to find the correlation between attributes.

3.1 Box plot

The box plots divide the data using five points: 'minimum', first quartile (Q1), median, third quartile (Q3) and 'maximum' as shown below. [8]

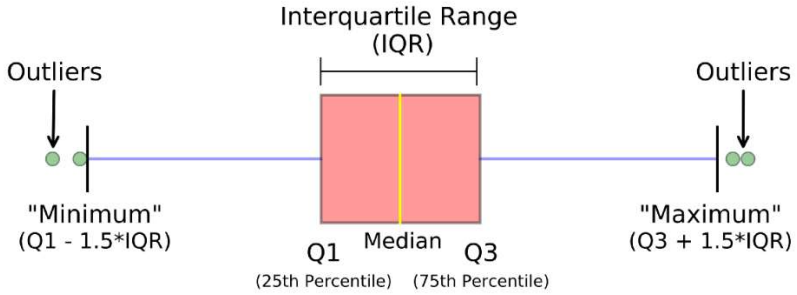


Figure 2: Graphic showing the box plot points.

Looking at our data's interquartile (IQR), we can say that bill length is the densest around the median, and the body mass is most asymmetrical. As previously mentioned, there are no outliers in the dataset after analysing the values within the table.

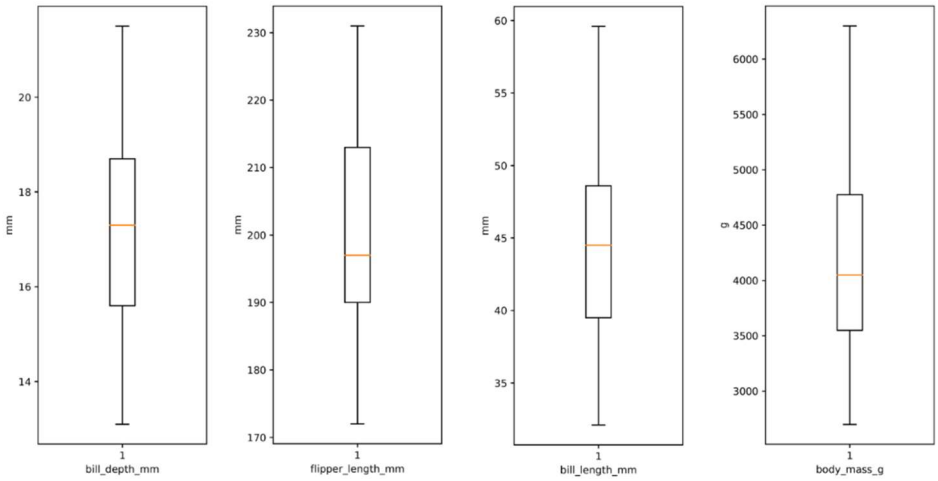


Figure 3: Box plots of ratio attributes

[8] Martin 70%, Lukas 30%

3.2 Column Histogram

Column histograms represent distributed numerical data intervals on the X-axis and the frequency of data in Y-Axis. Bill depth and body mass are the closest to the Normal (Gaussian) distribution. Nevertheless, we are working with a small amount of data and measurements. Using the statistical hypothesis test, we receive a very low certainty. Although the ratio attributes seem to be normally distributed, we cannot make a conclusive statement unless we have more instances in the dataset.

The bimodal distribution of flipper length shows us that the three species are tall or short, probably based on their sex. Body mass is skewed to the right, where the tail might contain extreme values. While the log or square root transformation might lead to a better result, the data skewness is not significant enough to apply the transformation. Hence, no data transformation is applied to this dataset.

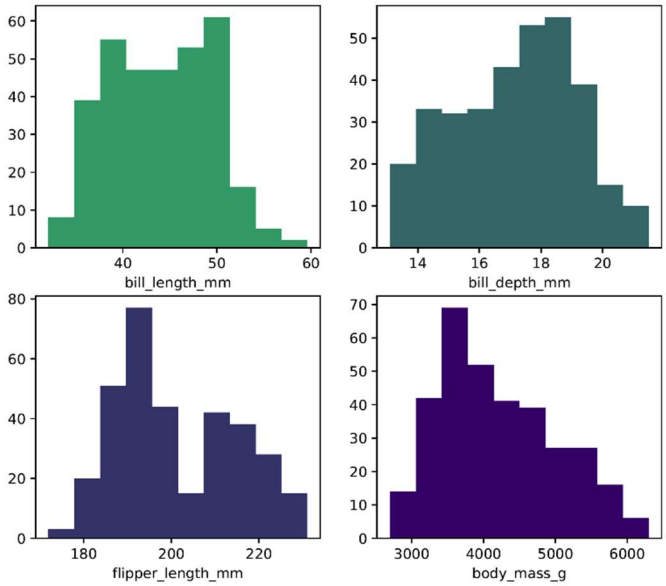


Figure 4: Column histogram of ratio attributes

3.3 Pearson correlation

Pearson's correlation measures the strength of the linear relationship between two attributes, where -1 means a total negative linear correlation, 0 being no correlation, and + 1 represents a total positive correlation. The Pearson correlation table shows us that the Bill depth and Bill length are not correlated, while other attributes are moderately correlated with a strong positive correlation between body mass and Flipper length.

Pearson Correlation	**Bill length**	**Bill depth**	**Flipper length**	**Body mass**
Bill length	-	-0.23	0.65	0.59
Bill depth	-0.23	-	-0.58	-0.47
Flipper length	0.65	-0.58	-	0.87
Body mass	0.59	-0.47	0.87	-

Table 2: Pearson correlation of ratio attributes

3.4 Visualisation

Looking at the sex of the penguins, which we want to predict using classification in a future report, the possible conclusion from the graphs below is that the male penguins have higher values for all ratio attributes. They are heavier and have a more prominent flipper and bill. This fact is supported by observing the pictures of real penguins, proving that the female penguins have a different bill. The initially chosen goal to cluster missing sex instances based on the body mass and bill depth was the best decision for the two attributes as the difference between males and females is the most significant. Nevertheless, using all attributes for classification would give us better results, and for this reason, we changed our initial idea from using only two attributes to using all of them.

The bill depth correlation, which we use for the linear regression, is more challenging to predict, as we are discussing ratio attribute and not nominal one as in the last paragraph. However, we can predict the bill depth interval by knowing the species, as each has different bills. Also, penguins with high body mass and flipper length would likely have a smaller bill depth.

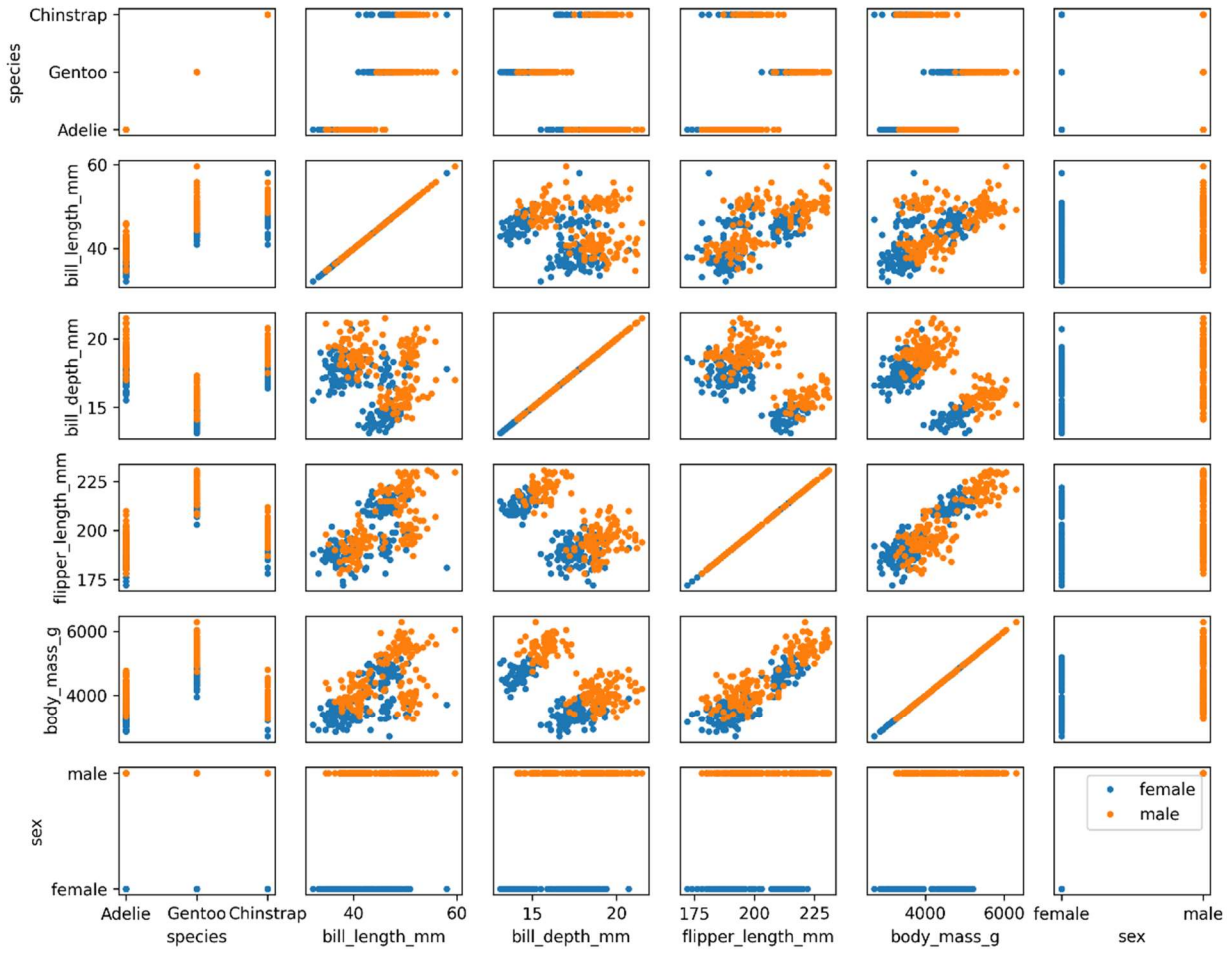


Figure 5: Visualising data in 2D using two attributes

4 PCA analysis[9]

We want to accomplish dimension reduction by this task while perceiving most of the data accuracy. We will use the Principal Component Analysis for the ratio attributes to visualise the penguin sex. Running PCA for nominal attributes is not recommended since finding the position of nominal value in space might be complicated. Moreover, Multiple Correspondence analysis might also be used when working with binary data.

4.1 Variance without standardisation

We have tried explaining the variance on a number of principal components with data standardised only by subtraction of mean and have encountered a phenomenon where the first principal component has described the data with 100% accuracy.

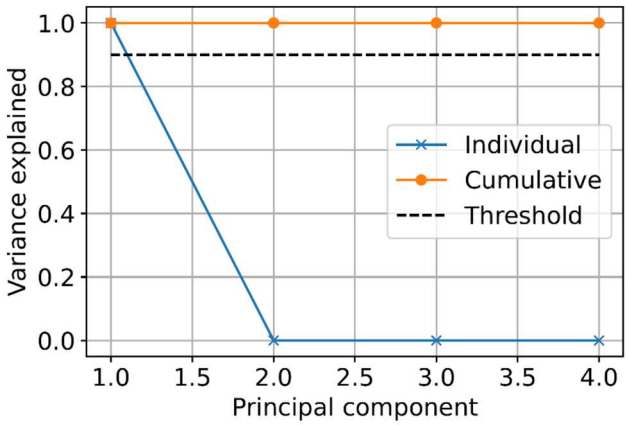


Figure 6: Variance explained by principal components without standardisation

This means that all of the data set attributes can be described as a linear transformation of only one of them, which would mean an extreme case of linear dependence and perfect variable collinearity. As this is clearly not the case for our data set, we will need to transform the data for further analysis accordingly.

4.2 Variance with Standardisation

We will further standardise the data by division of the standard deviation of each attribute, resulting in meaningful results.

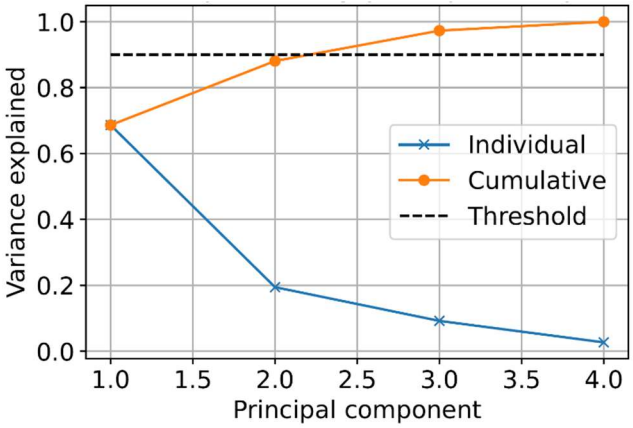


Figure 7: Variance explained by principal components with standardisation

[9] Martin 60%, Lukas 40%

4.3 Accuracy

With the above-applied standardisation, we have computed that the first two principal components achieved only 88.09 % accuracy. Three components were needed to perceive at least 97.3% accuracy of the data.

4.4 Component Coefficients

We can now use the principal directions[10] to see which of the principal component's original attributes mainly captures the variation. For example, the third principal component mainly captures the fourth attribute (body mass). The original data would have to be sign-flipped for the projection to be positive. As we have already transformed the data accordingly, there is no need for further standardisation of the graph.

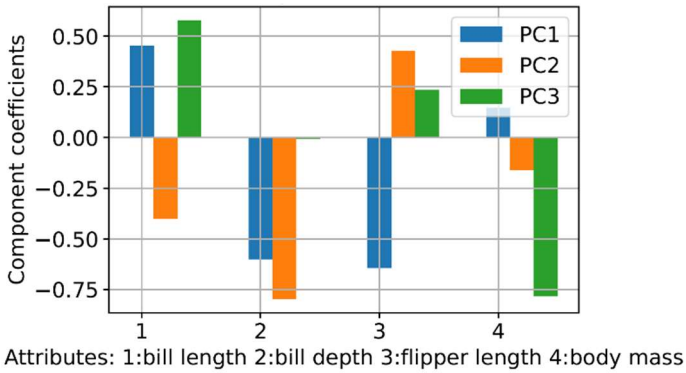
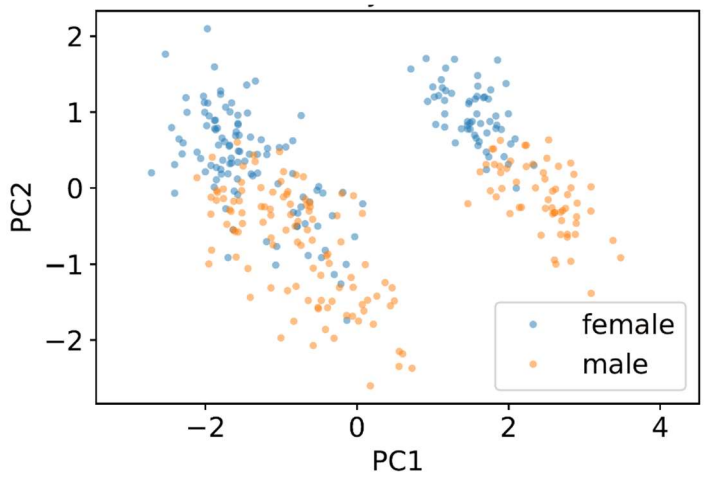


Figure 8: PCA component coefficients

4.5 PCA Visualization

Plotting the standardised projection $Z = U \cdot S$[11] with the first two principal components, we perceived almost 90% of the data and successfully reduced the dimension to 2D. Now we can apply this method for linear regression and classification tasks, which would result in a less computational demanding task and more suitable for graphical representation and visual analysis of non-correlated data.



[10] V is matrix decomposed by SVD

[11] U, S are decomposed by SVD

Figure 9: Zero-mean and unit variance projection in 2D

Applying the same method but with the first three principal components. We perceived 97.3% accuracy and reduced one dimension from the original data.

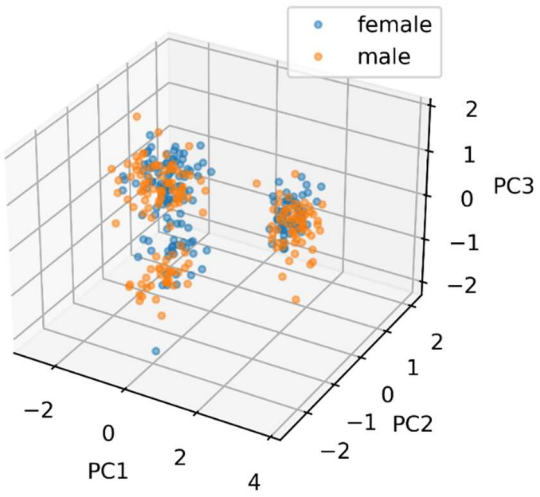


Figure 10: Zero-mean and unit variance projection in 3D

5 Summary

The attributes do not contain any outliers and only a few missing values. The missing sex will be predicted using binary classification in the following report. The classification problem is achievable, as there are significant differences between males and females mentioned throughout the report. The classification problem can be run either in the multidimensional space with original attributes or using the PCA, reducing a dimension and resulting in more straightforward computations while perceiving 88% of the data for first two PCs and more than 97% for the first three PCs.

In linear regression, we predict the depth of the bill. Looking at Figure 5, we can find patterns of how each attribute influences the final bill depth. Therefore this method is also suitable for the correct prediction. By creating the linear regression model, we can predict the unknown depth of the bill based on the input attributes. The decision tree could also be used as one of the methods for predicting unknown attributes.

Lastly, no data transformation is needed, as all of the ratio attributes are distributed close enough to the normal distribution.

6 Exam Problems[12]

6.1 Question 1

Correct answer: D

Explanation: x_1 is the interval, $x_{2\text{-}6}$ are all rations, and x_7 is ordinal; therefore, the only correct option is D.

6.2 Question 2

Correct Answer: A

Explanation: The general equation for this problem is $\left(x_1^p + x_2^p + x_3^p + x_4^p + x_5^p + x_6^p + x_7^p\right)^{(1/p)}$, for our vector x_{14}-x_{18} it results in a correct answer A

6.3 Question 3

Correct Answer: A

Explanation: Using equation $\frac{13.9^2+12.47^2+11.48^2+1\ .03^2}{13.9^2+12.47^2+1\ .48^2+10.03^2+9.45^2}$ for the first four components, we get the result 0.8668, which is greater than 0.8, therefore A) is the correct answer.

6.4 Question 4

Correct Answer: D

Explanation: We are using PCA2, the second column of matrix V. The value of time is -0.5, and it is the only negative number since we know that this value is low, and other positive values are high, it would result in positive projection.

6.5 Question 5

Correct Answer: A

Explanation: Jaccard similarity is similar to SMC, but SMC has the term M_{00} in its numerator and denominator, whereas the Jaccard index does not. Therefore, we divide the number of matches by the number of total words without duplicates, resulting in $2/13 = 0.153845$

6.6 Question 6

Correct Answer: B

Explanation: We calculate the probability of x_2, assuming that y_2 is two. Therefore, we look at the column y_2 of Table 2 and find the probability that x_2 is zero, 0.81 + 0.03.

[12] Lukas 60%, Martin 40%

7 References

[1] DTU Compute, 2022. [Online]. Available: https://www.compute.dtu.dk/.

[2] B. S. Jensen, DTU Compute, 2022. [Online]. Available: http://compute.dtu.dk/courses/02450/.

[3] S. Lopp, 2021. [Online]. Available: https://gist.github.com/slopp/ce3b90b9168f2f921784de84fa445651.

[4] L. a K. Gorman, Environmental Data Initiative, 2007-2009. [Online]. Available: https://environmentaldatainitiative.org/dataset-design/.

[5] K. B. Gorman, "Ecological Sexual Dimorphism and Environmental Variability within a Community of Antarctic Penguins, "https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0090081, 2014.

[6] A. Horst, 2022. [Online]. Available: https://github.com/allisonhorst/palmerpenguins/.

[7] Freedman, Pisani, Purves a Rog, "Statistics (international student edition), "2007. [Online].

[8] M. Galarnyk, "Understanding Box Plots, "https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51, 2018.

[9] A. Horst, "Art work, "2021. [Online]. Available: https://github.com/allisonhorst/stats-illustrations.