Assignment 2 BUSA8000

Koala Conservation Research

Student Name: Manuel Cabeza

Student Number: 48622605

Word count = 3290

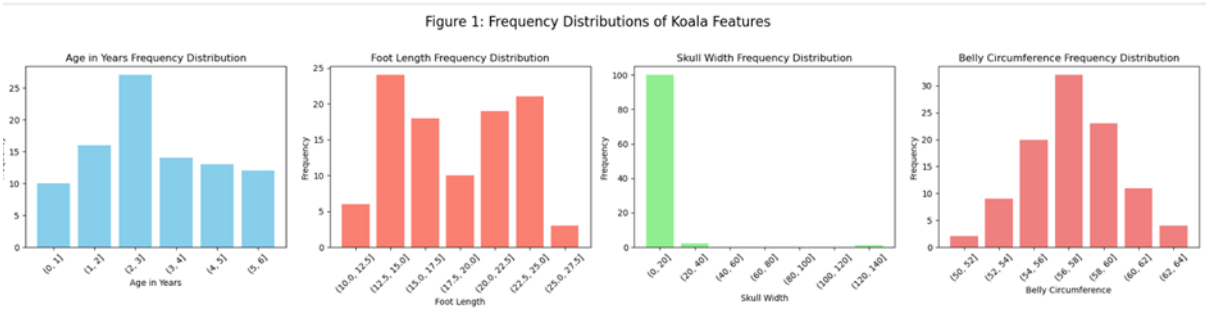
I acknowledge that I have **only used GAITs (e.g., ChatGPT) in drafting and proofreading this assignment, which is permitted in the assignment instructions.**

Contents

Section 1: Exploratory Data Analysis (30 Marks)

1.1 Find NAs and clean them.

We can notice that the following columns have missing values: **age_in_years:** 2, **foot_length:** 2, **skull_width:** 1, **belly_circumference:** 1. To understand how to replace each of the specific values we are going to look at the data (figure 1) to decide between mean and median to replace the nulls.



- **age_in_years** (2 nulls): We will use the mean to replace the null values since it has a good central tendency and is effective when the data is relatively symmetrically distributed.
- **foot_length** (2 nulls): Has considerate variability, in this case the best option is to use the median since it isn´t impacted by the possible outliers
- **skull_width** (1 null): Clearly has 120 as outlier, so we are using the median instead of the mean to replace the null values.
- **belly_circumference** (1 null): Has a small standard deviation and central tendency to its mean, for that reason we are going to use the mean.

1.2 Find wrong entries and clean them:

Let´s check now the wrong entries for each column, we do it finding the unique entries in each column

- In **habitat** I corrected entries to have a standard habitat with only 3 capital letters, specifically for the option "q" I checked to which region belonged that entry which was **QLD**. Also, Queenstown was a wrong entry that belongs to the state of **QLD**
- In **gender** I standardized the entries to use **male** and **female** as the only values.
- In head_length I found 110.5 as a wrong entry (wrong units, mm instead of cm) I replaced it with 11.05.

- In ear_size It looks like 900.8 is a wrong entry that was supposed to be 9.8 and 110.4 was supposed to be 11.04
- In skull_width It looks like 120.6 is a wrong entry that was supposed to be 12.06
- In eye_ diameter It looks like 400.5 is a wrong entry that was supposed to be 4.5

1.3 Find duplicates and clean them:

The duplicates were searched in terms of the column koala_id, I didn't´ find any duplicates

1.4 Find outliers and clean if any

I checked if we had any outliers printing boxplots for the data that is numeric:

Figure 2: Boxplots numeric data

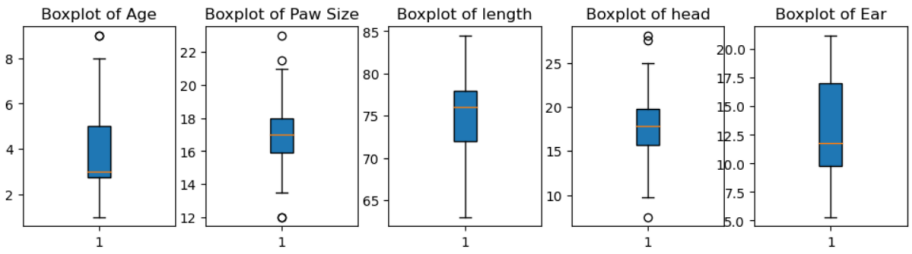
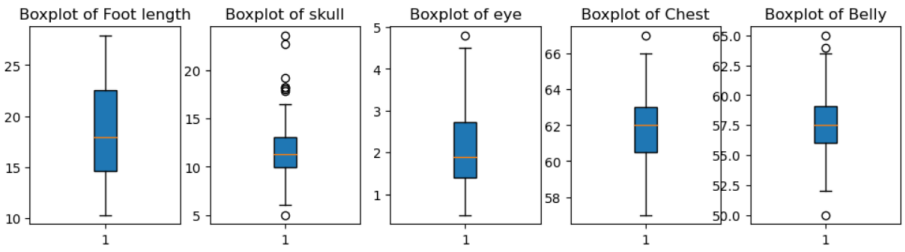
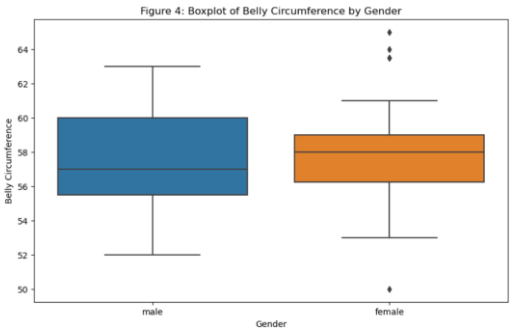


Figure 3: Boxplots numeric data



I noticed from the boxplots in Figure 2 and Figure 3 that skull, eye, chest, belly circumference, age, paw size and head could potentially have some outliers.

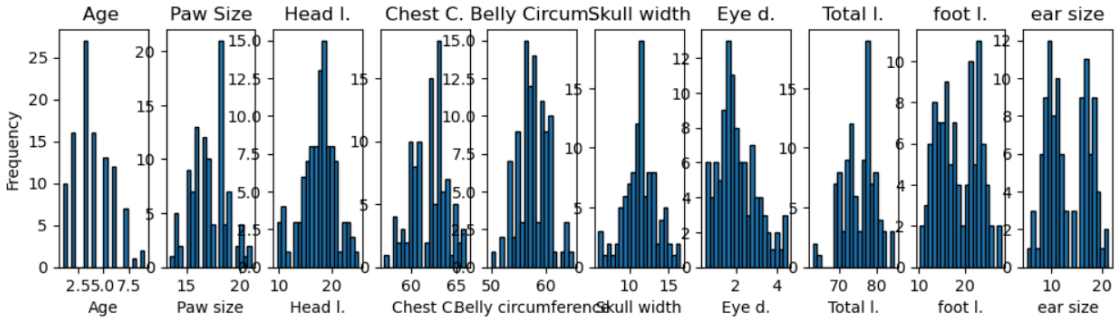
- The **Age "**outlier" wasn´t replaced as it could be a 1-year older koala that could give us important information.
- For **belly_circumference** as all the "outliers" values are of females it could be due to them being pregnant, I didn't replace them (see figure 4).
- For the rest of the features I looked for values that fall 1.5 times below or above our interquartile range and replace them with the median.



1.5 Find skewness and correct it if any

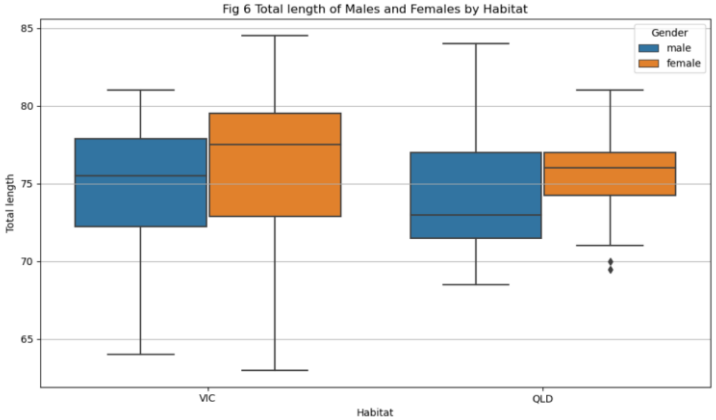
Even if some features didn't have noticeable outliers, they could still be skewed. Therefore, I checked the skewness of all variables to identify potential issues that may not have been immediately obvious. By examining the frequency distributions (Figure 5), I specifically checked the skewness of the numerical variables and concluded that there was no significant skewness in the data.

Figure 5: Frequency distribution



Section 2: Data visualisation (10 Marks)

1. **Visualization 1: Total length of Males and Females by Habitat.**



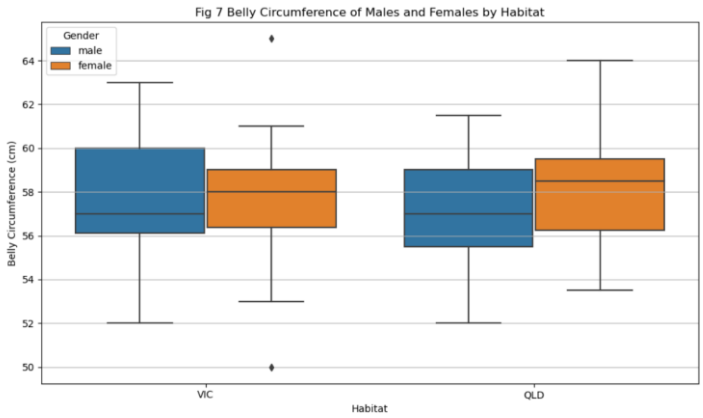
Habitat Comparison: In Fig 6 we can see as for both genders, Koalas in VIC show a wider distribution of total lengths, with some outliers present. This could imply that there is a greater variety of total lengths in this region compared to QLD. In QLD, the box for female Koalas is

shorter than that for males, indicating less variability in their total length compared to the males in the same habitat.

While both males and females have longer total lengths in VIC compared to QLD, the lengths seem to cluster more tightly in QLD. This suggests that the environmental conditions in VIC may support a broader range of sizes for Koalas.

Overall, this box plot provides valuable insights into the physical characteristics of male and female Koalas based on their habitat. In this case, the significant differences in total length and variability suggest that gender and habitat play crucial roles in the physical development of Koalas. Further analysis could investigate the underlying factors contributing to these differences, such as food availability, environmental conditions, or genetic factors.

2. Visualization 2: Belly Circumference of Males and Females by Habitat.



I decided to do this visualization (fig 7) to further investigate the belly circumference of koalas after finding some “outliers” during the EDA.

Central Tendencies: The median belly circumference for males appears to be higher than for females in all the habitats VIC and QLD. Males generally have a wider interquartile range (IQR) compared to females, indicating more variability in belly circumference among males.

Variability: The spread (IQR) for males is greater than for females in VIC, which suggests that male belly circumference measurements are more dispersed. In QLD, the distributions for males and females seem to have similar spreads.

Habitat Differences: The overall belly circumference seems to be higher in the VIC habitat compared to QLD for both genders, indicating possible environmental or biological factors influencing these measurements. The boxplots indicate notable differences in belly circumference based on gender and habitat, suggesting a possible link between these factors. Further analysis may be needed to explore potential biological, environmental, or lifestyle factors contributing to these differences.

Section 3: Analysis (50 Marks)

3.1. Is the mean head length of the Koalas significantly different from 92.0 mm?

I conducted hypothesis testing to understand if the head length of koala is significantly different from 92.00 mm following the next steps:

1. State the hypothesis:

Ho: μ = 92.00 mm

Ha: μ ≠ 92.00 mm

2. Define a significance level α = 0.05

3. Define type of test: This is a two-tailed one sample test because we are testing for any difference (either greater than or less than 92.0 mm) and we don´t know the population details (standard deviation)

I obtained the next results: T-statistic: -242.62, P-value: 7.37e-144

T-statistic: A large negative value of the t-statistic suggests that the sample mean of the koalas' head length is much smaller than the hypothesized value of 92.0 mm. In fact, the **mean** head length is **17.5 mm**

P-value: A p-value of **7.37e-144** is practically zero, which is far below my significance level 0.05 this is overwhelming evidence to reject the null hypothesis Ho.

Conclusion: I can confidently **reject the null hypothesis** that the mean head length of koalas is 92.0 mm. There is a statistically significant difference, and the data suggests that the mean head length is likely much lower than 92.0 mm.

3.2. Do male and female Koalas have significantly different mean head lengths?

I conducted hypothesis testing to understand if the head length of koalas' males and females is significantly different the steps followed were:

1. State the hypothesis

Ho: μmale = μfemale

Ha: μmale ≠ μfemale

2. Define significance level α as 0.05

3. This is a two-sample t-test, because I am comparing two independent groups (males and females).

I obtained the next results:

- **T-statistic:** 0.97, This t-statistic close to 0 suggests that the difference between the sample means of male and female koalas is small.
- **P-value:** 0.33, Since this p-value is greater than my significance level 0.05, I **fail to reject the null hypothesis**.

Conclusion: There is no significant difference in the mean head lengths between male and female koalas based on the data. The data suggests that the head lengths of male and female koalas are statistically similar, and I do not have enough evidence to say that male and female koalas have different head lengths. In fact, the **mean head length** for males is **17.76** and for females is **17.15**

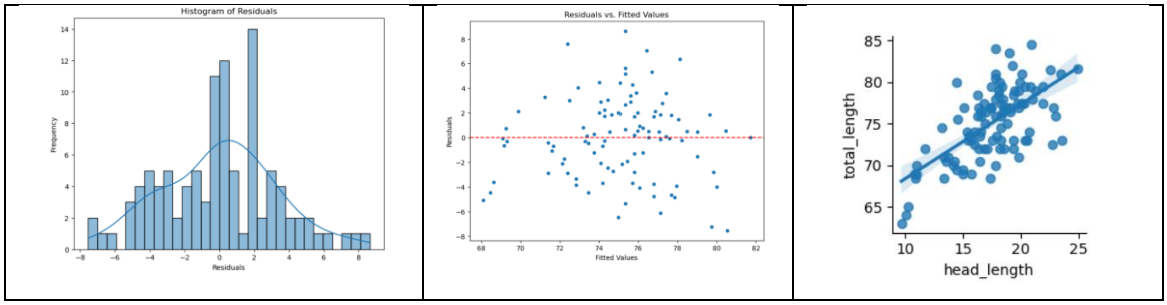
3.3. Can we predict the total length of a Koala based on its head length?

To understand if I could predict the total length of a Koala based on its head length I followed the next steps:

1. Build an OLS model
2. Check the significance of the coefficients.
3. Chek the model assumptions

Results:

- The **p-value** for the **head_length** coefficient is **0.000**, which is well below the commonly accepted significance of 0.05. This indicates that there is a **statistically significant** relationship between head length and total length, therefore I can reject the null hypothesis that there is no relationship between these two variables.
- The **coefficient** for head_length is **0.89**, meaning that for every additional cm increase in head length, the total length is expected to increase by approximately 0.89 cm, holding all else constant. This positive relationship supports the idea that as head length increases, total length also increases.
- The **R-squared** value is **0.423**, indicating that approximately 42.3% of the variability in total length can be explained by head length. While this shows that the model has some explanatory power, it is **small** and suggests there are other factors influencing total length that are not included in the model.
- The model assumptions were fulfilled as seen in the graph. I assume the samples were taken using a strategy that make them independent.



Conclusion: In summary, the analysis shows that head length is a statistically significant predictor of total length, with a positive relationship between the two variables. While head length provides some predictive power, the model could benefit from additional variables to improve its explanatory capability.

3.4. Can we predict the total length of a Koala based on multiple factors such as head length, skull width, and foot length?

Similarly to the previous questions, the following steps were followed. As this is a multilinear regression, I needed to check that there was not multicollinearity present between the variables.

1. Check multicollinearity
2. Build an OLS model
3. Check the significance of the coefficients.
4. Chek the model assumptions

Results:

Multicollinearity: After checking the VIF I conclude there is not multicollinearity as all the variables have factors smaller than 2

	Feature	VIF
0	const	41.822457
1	head_length	1.997762
2	skull_width	1.904364
3	foot_length	1.159728

Statistical significance: Head Length: The p-value for the head_length coefficient is 0.000, indicating that it is highly statistically significant. We can confidently reject the null hypothesis that there is no relationship between head length and total length.

Skull Width: The p-value for skull_width is 0.006, also below the significance threshold of 0.05, indicating a statistically significant relationship between skull width and total length.

Foot Length: The p-value for foot_length is 0.004, demonstrating a significant relationship between foot length and total length.

The model assumptions were fulfilled, normality of residuals, homoscedasticity and linear relationship between the dependent and independent variables is fulfilled. I assume the samples were taken in a strategy that make them independent.

Predictive Relationship: The coefficient for head_length is 0.5293, meaning that for every additional cm of head length, total length is expected to increase by approximately 0.5293 cm, holding other variables constant.

The coefficient for skull_width is 0.5481, indicating that for every additional cm of skull width, total length is expected to increase by 0.5481 cm, holding other variables constant.

The coefficient for foot_length is 0.2193, meaning that for every cm increase in foot length, total length is expected to increase by approximately 0.2193 cm, holding other variables constant.

Model fit (R2): The R2 value of 0.511 indicates that approximately 51.1% of the variability in total length is explained by the combined predictors of head length, skull width, and foot length. This is a better fit than the model using only head length, which explains 42.3% of the variability. The adjusted R-squared of 0.496 accounts for the number of predictors and suggests a good model fit with room for improvement by considering additional factors.

Conclusion: In summary, the analysis demonstrates that head length, skull width, and foot length are significant predictors of total length. The positive relationships between these variables and total length support their inclusion in predictive models. While the model explains a significant portion of the variability in total length, further improvements could be made by considering other factors.

3.5. Do environmental factors such as state affect Koala's physical characteristics?

In this question I created models for each of the physical characteristics of koalas using the state as the independent variable and the respective physical characteristic as the dependent variable. In this question it was also necessary to convert the categorical variable habitat using one hot encoding. The steps followed were:

1. Convert habitat using one hot encoding
2. Perform OLS for each feature
3. Interpret the p value of the coefficient of the state in each OLS

Since using one hot encoding it was necessary to drop one of the states (QLD) to avoid multicollinearity, the dropped category served as the baseline group and the coefficient of the remaining states (VIC) were interpreted relative to this reference state.

Conclusion:

Non-significant Features: The physical characteristics of total length, head length, skull width, eye diameter, and belly circumference showed no statistically significant differences between Victoria and Queensland ($p > 0.05$). This suggests that these features are relatively uniform across the different habitats.

Significant Features: On the other hand, the following features showed statistically significant differences ($p < 0.05$) between Victoria and Queensland, indicating environmental factors may influence them:

Paw Size: Koalas in Victoria have paw sizes that are 1.53 cm smaller than those in Queensland.

Foot Length: Koalas in Victoria have feet that are 6.90 cm larger than those in Queensland.

Ear Size: Koalas in Victoria have ears that are 7.28 cm larger than those in Queensland.

Chest Circumference: Koalas in Victoria have chests that are 0.88 cm larger than those in Queensland.

3.6. What factors are correlated with the total length of a Koala?

In this question I followed the multiple regression framework explained in classes, these steps were repeated and not necessarily followed in the strict order described here.

1. **Multiple Regression Approach:** Used to assesses the relationship between a continuous dependent variable (in this case, the total length of a Koala) and several independent variables (factors).
2. **Assumptions for Multiple Regression**:
   - Linearity: Each predictor variable (like foot length or paw size) should have a linear relationship with Koala length. A scatter plot was used to check this
   - Normality: The residuals (errors) should follow a normal distribution.
   - Independence of observations: Each data point (Koala measurement) should be independent.
   - Homoscedasticity: The variance of errors should be consistent across the model
3. **Handling Multicollinearity:** Used The Variance Inflation Factor (VIF) to check for multicollinearity, and if needed, variables can be transformed or removed.
4. Ran multiple regression again with the selected features.

Conclusion:

In summary, the selected features **paw size**, **head length**, and **foot length** are the factors significantly correlated with the total length of Koalas. After employing the **multiple regression framework** to determine these factors were significantly correlated with the total length of a Koala I followed these steps:

1. **Multiple Regression Approach**:
   - The dependent variable was **total length** (a continuous variable), and several independent variables (like paw size, foot length, head length, and gender) were used as predictors.
2. **Addressing Multicollinearity**:
   - I checked for multicollinearity using the **Variance Inflation Factor (VIF)**. Variables with a VIF greater than 5 indicated potential multicollinearity issues. The initial model had strong multicollinearity, as indicated by a large condition number of **4.03e+03**.

- Variables such as **region, ear size**, and **habitat_VIC** had VIF values exceeding 5, with **ear size** having a VIF of 5.88. I initially kept ear size and dropped **region** and **habitat_VIC** to see how the multicollinearity changed.

3. **Refined Model**:

    - After removing **region** and **habitat_VIC**, I recalculated the VIF and found that all remaining variables had acceptable VIF values (less than 5). The refined model was then built using **paw size, head length, foot length**, and the constant term which were the features statistically significant ($p < 0.05$).

4. **Significant Features**:

    - As mentioned before, in the final model (as seen in the next figure) the following variables were statistically significant ($p < 0.05$) and correlated with the total length of a Koala:

        - **Paw Size**
        - **Head Length**
        - **Foot Length**
        - The **constant** (intercept)

    - These variables explained a substantial portion of the variation in total length, with an **R2 value of 0.709**, indicating that 70.9% of the variance in total length can be explained by the model.

OLS Regression Results

Dep. Variable:	total_length	R-squared:	0.709
Model:	OLS	Adj. R-squared:	0.701
Method:	Least Squares	F-statistic:	81.35
Date:	Mon, 14 Oct 2024	Prob (F-statistic):	1.00e-26
Time:	12:36:51	Log-Likelihood:	-234.76
No. Observations:	104	AIC:	477.5
Df Residuals:	100	BIC:	488.1
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	34.9861	2.842	12.310	0.000	29.347	40.625
Paw Size	1.3066	0.145	9.026	0.000	1.019	1.594
head_length	0.6250	0.081	7.673	0.000	0.463	0.787
foot_length	0.3763	0.059	6.360	0.000	0.259	0.494

5. **Model Evaluation**:

    - The residuals of the final model were **normally distributed**, satisfying the normality assumption.
    - The model showed **homoscedasticity**, with consistent variance of residuals across all levels of the predictors.
    - There was a clear **linear relationship** between **paw size, head length, foot length**, and the total length of Koalas.
    - I assume the samples were taken with a strategy that makes them independent.

Section 4: Analysis (10 Marks)

Final Analysis of Koala's Physical Characteristics Based on Environmental Factors and Other Variables: This study explored the relationship between koalas' physical characteristics and environmental factors, such as habitat, using multiple regression analysis and hypothesis testing. Significant findings revealed that paw size, foot length, and ear size vary by habitat, with koalas in Victoria having smaller paws and larger foot lengths and ears compared to those in Queensland. However, traits like total length, head length, and belly circumference did not show significant variation across habitats. A key result was that paw size, head length, and foot length were strongly correlated with total length, explaining 70.9% of its variability. Additionally, no significant differences in head length were found between male and female koalas. The analysis involved addressing multicollinearity and cleaning the data to refine the models. Visualizations further highlighted regional differences in koala sizes and gender-based variability in belly circumference, underscoring the influence of both environmental and biological factors on koala development.