BUSA8000:

Techniques in Business Analytics

Assignment 2

Koala Conservation Research

Quoc Phong Nguyen

Student ID: 48203971

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

Word count: 2964

Table of Contents

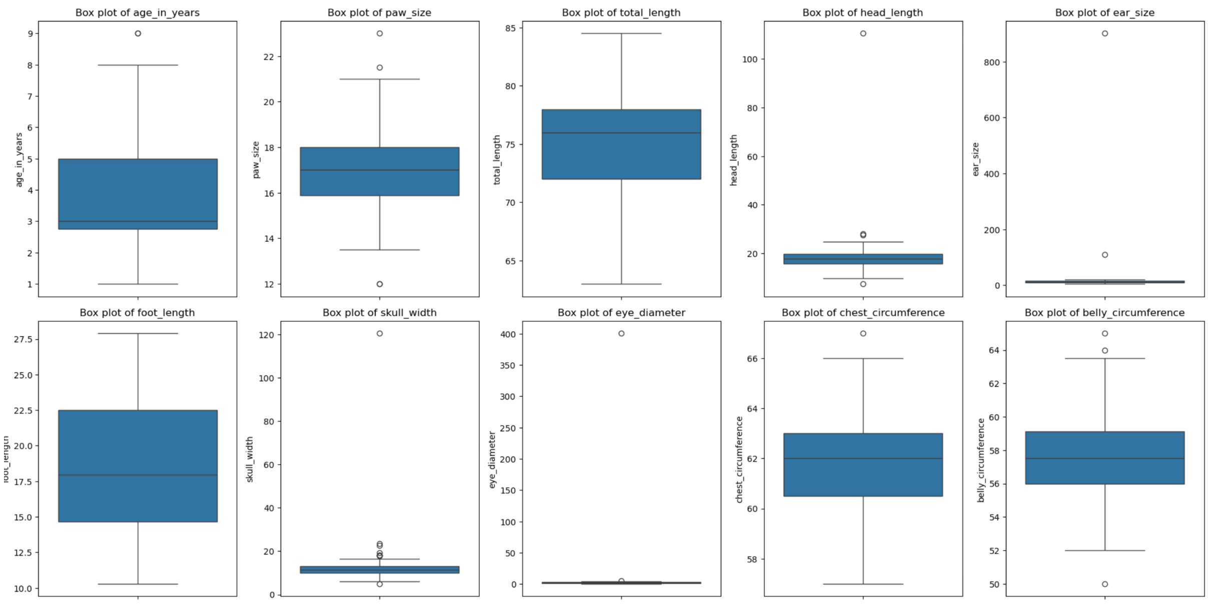
SECTION 1: Exploratory Data Analysis

In Section 1, we undertook a comprehensive data cleaning process to ensure the quality and consistency of our koala dataset. We began by loading the dataset and conducting an initial exploration, examining its structure, unique values, data types, and summary statistics. This initial review revealed issues needed addressing such as missing values or incorrect data entry.

Our first step in cleaning the data was to handle format inconsistencies and incorrect entries. We renamed the 'Paw Size' column to 'paw_size' for consistency with other column names. We then corrected inconsistent entries in the 'habitat' column, mapping various spellings and abbreviations to standardized values of 'QLD' and 'VIC'. Similarly, we addressed inconsistencies in the 'gender' column, ensuring all entries were uniformly recorded as either 'male' or 'female'.

Next, we tackled the issue of missing values. We identified four columns with missing data: 'age_in_years', 'foot_length', 'skull_width', and 'belly_circumference'. We note that there are extreme values in the dataset, which may be outliers. To address null values, we employed median imputation, filling in the missing values with the median value of each respective column. This method was chosen as it is less sensitive to outliers compared to mean imputation. We then checked for duplicate records based on 'koala_id', 'region', and 'habitat', but found no duplicates in our dataset.

A significant part of our cleaning process involved examining the distribution of numerical variables and addressing outliers. We created visualizations such as histograms and box plots to better understand the data distribution.



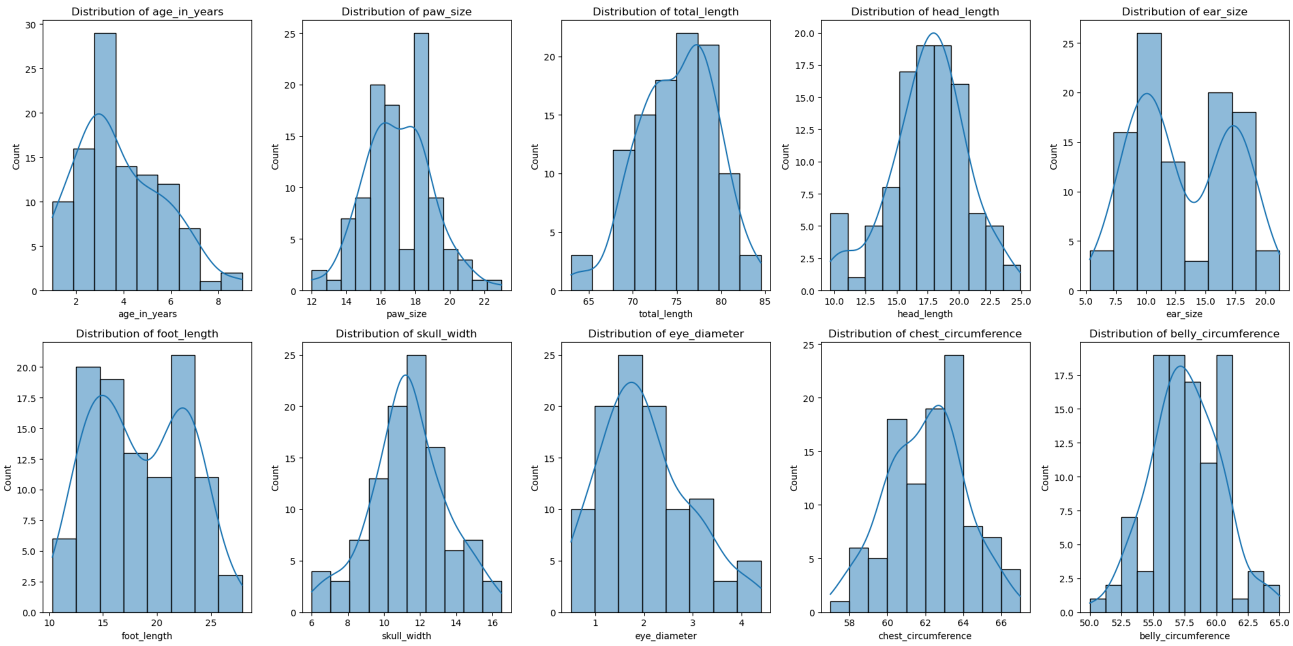
We also calculated skewness for numerical columns to assess distribution symmetry. This analysis revealed extreme outliers in variables like head_length, ear_size, skull_width, and eye_diameter.

- The preliminary calculation of skewness indicates that there are some factors with highly skewed distribution, which is less than -1 or greater than 1:
    - head_length: 8.11
    - ear_size: 9.85
    - skull_width: 8.89
    - eye_diameter: 10.04

For these variables, we implemented a function to replace outliers with the median value of the respective column. This approach helped to mitigate the impact of extreme values that were likely due to measurement errors or data entry mistakes.

It is worth noting that we chose to retain milder outliers in variables such as age_in_years, paw_size, chest_circumference, and belly_circumference. We made this decision because these outliers likely represent natural variation in the koala population and could provide valuable insights in our analysis. For instance, outliers in belly_circumference can be pregnant koalas, while variations in paw size might represent natural differences among koalas.

Finally, we conducted a re-examination of the distributions and skewness of numerical variables after our outlier treatment to confirm that our cleaning process had indeed improved the quality of the data.




```
Skewness after outlier treatment:
Skewness of age_in_years: 0.56
Skewness of paw_size: 0.17
Skewness of total_length: -0.28
Skewness of head_length: -0.35
Skewness of ear_size: 0.14
Skewness of foot_length: 0.1
Skewness of skull_width: -0.06
Skewness of eye_diameter: 0.57
Skewness of chest_circumference: -0.05
Skewness of belly_circumference: 0.08
```

- After handling outliers, the skewness of most variables has significantly improved. These values are now ranging from around -0.5 to 0.5.
- The previously problematic variables (`head_length`, `ear_size`, `skull_width`, and `eye_diameter`) are successfully addressed, with their skewness values now within acceptable ranges.
- Only `age_in_years` (0.56) and `eye_diameter` (0.57) show mild positive skewness, which is generally acceptable and may not require further transformation.

SECTION 2: Data Visualisation

Insight 1: Koala Body Proportions and Gender Differences

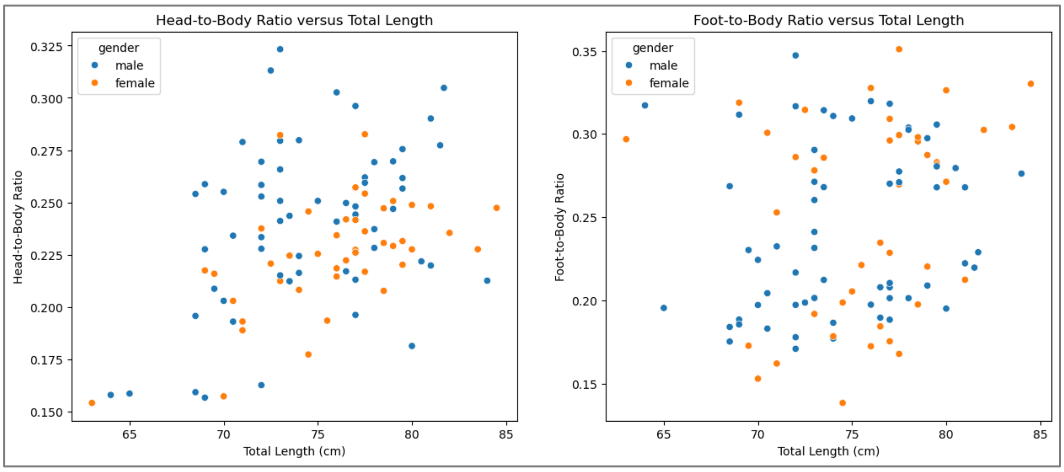
In studying koalas, understanding their physical characteristics is crucial for assessing their health, adaptations, and overall well-being. One particularly interesting aspect of the koala study is how the sizes of different body parts, such as the head and feet, compared to their overall body length. These relationships can vary as koalas grow and may differ between males and females. To explore this, we focused on two key measurements: the ratio of head length to total body length, and the ratio of foot length to total body length. To calculate the body proportion ratios, we create two new variables:

1. Head-to-body ratio = Head length / Total length

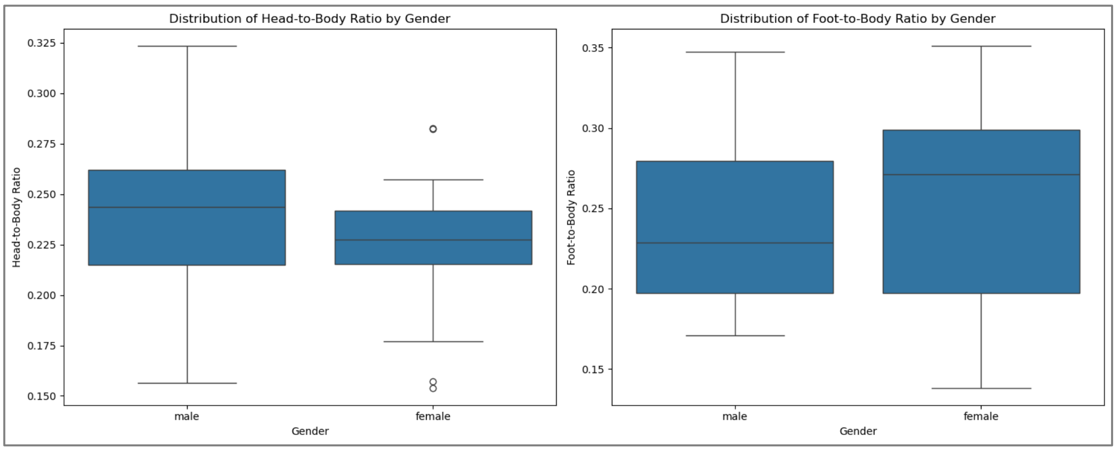
2. Foot-to-body ratio = Foot length / Total length

These ratios can provide valuable insights into the koalas' adaptations to their tree-dwelling lifestyle. For instance, the head-to-body ratio might reflect adaptations for feeding on tough eucalyptus leaves, while the foot-to-body ratio could indicate climbing efficiency. By examining how these ratios change with overall body length and comparing them between genders, we can uncover patterns that might not be apparent from individual measurements alone.

To visualize this data effectively, we employed a combination of scatter plots and box plots. The scatter plots show how the body proportion ratios change as koalas increase in total length, with points coloured to distinguish between males and females. This allows us to see trends across the full range of koala sizes.



The box plots, on the other hand, display the distribution of these ratios for each gender, providing a clear statistical summary of any differences.



The scatter plots and box plots provide valuable insights into the body proportions of koalas and how they differ between males and females. There are some findings we can gain from them.

When it comes to the head-to-body ratio, we can observe a slight tendency for this ratio to increase as the total length of the koala increases, suggesting that larger koalas may have proportionally larger heads. Interestingly, male koalas appear to have a slightly higher head-to-body ratio on average than females, as shown by the higher median in the box plot. This could potentially be related to differences in feeding habits or behaviours between genders.

When we look at the foot-to-body ratio, we notice a different pattern. There seems to be more variation in this ratio across all body sizes, with no clear trend as total length increases. However, the box plot shows that female koalas generally have a slightly higher foot-to-body ratio compared to males. This subtle difference could suggest an adaptation that gives female koalas better climbing abilities, which may be especially helpful when they are caring for their young.

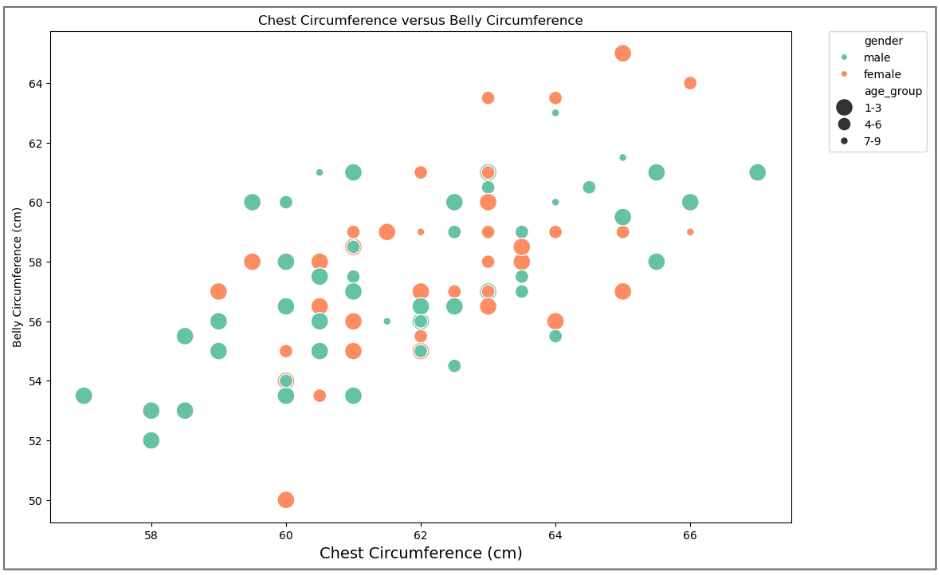
It is important to note that while gender differences are observable, there is considerable overlap in the distributions for both genders. This suggests that while there are average differences between males and females, individual variation plays a significant role in koala body proportions. The scatter plots also reveal that koalas with similar total lengths can have quite different body proportions, highlighting the diversity within the species.

Insight 2: Relationship between Chest Circumference and Belly Circumference

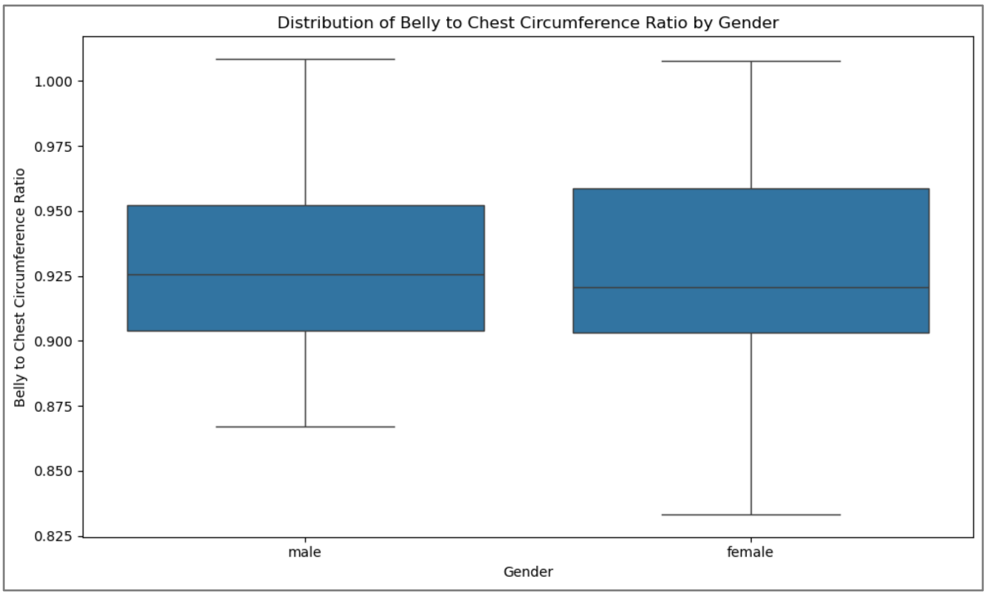
The next aspect of the koala study that we want to visualize is the relationship between chest and belly circumferences. This comparison can provide valuable information about koala body shape, potential gender differences, and how body proportions might change with age.

First, we divided the 'age_in_years' variable into three categories: 1-3 years, 4-6 years, and 7-9 years to create age groups. This helps us study how body shape changes as koalas transition from young adults to mature individuals. We selected these age ranges to create evenly spaced groups in terms of years, while still capturing meaningful differences between life stages. However, we should also note that the number of observations in each group may not be evenly distributed due to limited sample sizes.

We chose to use a scatter plot as our main visualization, showing chest circumference against belly circumference. To make the plot more informative, we used different colored markers to distinguish between genders and varied the size of the points to represent various age groups. This allows us to see how chest and belly circumferences are related, how this relationship might differ between males and females, and how it might change as koalas age.



To provide a more focused view of gender differences, we calculated the belly-to-chest circumference ratio for each koala and visualized the distribution using a box plot, separated by gender. This additional plot helps us identify any significant differences in body shape between male and female koalas, regardless of overall size.



As we can see from the scatter plot, it reveals a clear positive correlation between chest and belly circumference, which is expected as larger koalas tend to have larger measurements overall. However, the relationship is not perfectly linear, indicating some variation in body shapes among individuals.

The scatter plot indicates that female koalas, represented by orange dots, tend to have slightly larger belly circumferences relative to their chest circumferences compared to males, especially in the middle and upper ranges of chest circumference. However, the box plot shows that males actually have a slightly higher median belly-to-chest circumference ratio than females, although the difference is minimal. The box plot also clearly demonstrates that females have greater variability in this ratio, as shown by the larger overall range and longer whiskers, especially on the lower end. This variability in females could indicate more diverse body shapes, possibly related to different reproductive stages.

In the scatter plot, dot size indicates koala age groups. Larger dots represent younger koalas (1-3 years), medium dots represent middle-aged koalas (4-6 years), and smaller dots represent older koalas (7-9 years). Generally, older koalas tend to have slightly larger chest and belly circumferences, aligning with expected growth patterns. However, the relationship between age and body proportions is not straightforward, as some younger koalas have larger measurements than older ones. This suggests significant individual variation regardless of age. It is important to note that the unequal sample sizes across age groups may impact these observed patterns.

The greater variability in female measurements, particularly in belly circumference, could be attributed to physiological differences related to reproduction. Female koalas carry their young in a pouch located on their belly, which may lead to more fluctuation in belly circumference depending on their reproductive status. This could explain why some female koalas have notably larger belly circumferences relative to their chest size.

SECTION 3: Analysis

Question 1. Is the mean head length of the Koalas significantly different from 92.0 mm?

We employed a one-sample t-test for this analysis as it is appropriate for comparing a sample mean to a specific value, given our sample size and data distribution. This test allows us to determine if the mean head length of our koala sample differs significantly from 92.0 mm (9.2 cm). We have the hypotheses as: H0: $\mu = 9.2$ and H1: $\mu \neq 9.2$.

```
t-statistic: 27.852
p-value: 0.000
--------------------------------------------------------------------------------
Reject the null hypothesis. The mean head length is significantly different from 9.2 cm.
Sample mean: 17.5721
Sample standard deviation: 3.0654
```

Based on the above test result, we can confirm that the mean head length of the Koalas is significantly different from 92.0 mm or 9.2 cm at the significance level of 5%.

Question 2. Do male and female Koalas have significantly different mean head lengths?

We used a two-sample t-test to compare the mean head lengths of male and female koalas. This test is appropriate for comparing means between two independent groups. Our null hypothesis (H0) stated that the mean head lengths of male and female koalas are equal, while the alternative hypothesis (H1) proposed a significant difference.

```
t-statistic: 1.1846
p-value: 0.239
--------------------------------------------------------------------------------
Fail to reject the null hypothesis.
The data does not provide sufficient evidence that the mean head lengths of male and female koalas are significantly different.

Male koalas:
Sample mean: 17.8705 cm
Sample standard deviation: 3.3573 cm

Female koalas:
Sample mean: 17.1488 cm
Sample standard deviation: 2.5749 cm
```

Based on the above test result, we can believe that male and female Koalas do not have significantly different mean head lengths at the significance level of 5%.

Question 3. Can we predict the total length of a Koala based on its head length?

To answer the question whether we can predict the total length of a Koala based on its head length, we conducted a simple linear regression analysis. First, we visualized the relationship between head length and total length using a scatter plot with a regression line, which revealed a positive linear relationship between these two variables.

Our analysis shows that head length is indeed a significant predictor of total length in koalas. The regression model indicates that for every 1 cm increase in head length, we can expect the total length to increase by approximately 0.89 cm. The model explains about 40.3% of the variance in total length, suggesting that while head length is a useful predictor, other factors also influence a koala's total length.

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	59.4088	1.919	30.955	0.000	55.602	63.216
head_length	0.8923	0.108	8.292	0.000	0.679	1.106

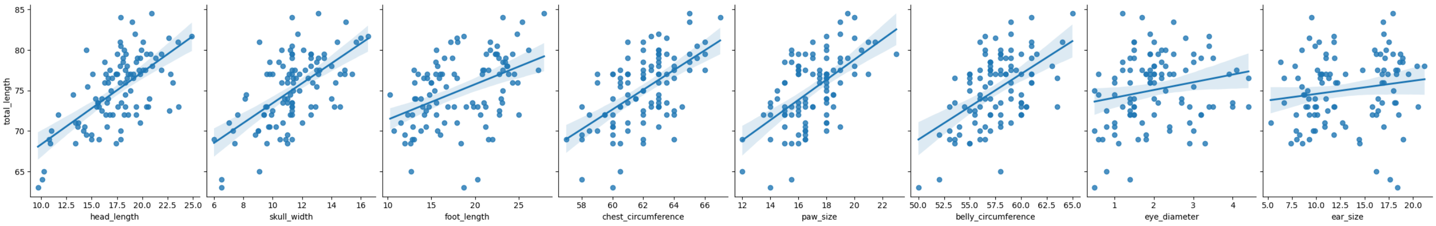
Omnibus:	0.043	Durbin-Watson:	1.580
Prob(Omnibus):	0.979	Jarque-Bera (JB):	0.190
Skew:	0.017	Prob(JB):	0.909
Kurtosis:	2.793	Cond. No.	105.

R-squared:	0.403
Adj. R-squared:	0.397
F-statistic:	68.76
Prob (F-statistic):	4.73e-13
Log-Likelihood:	-272.22
AIC:	548.4
BIC:	553.7

We verified also that the assumptions of linear regression were met by examining residual plots, which showed that the residuals were approximately normally distributed and homoscedastic. Therefore, our simple linear regression model is valid and we can use it to predict the total length of a Koala based on its head length.

Question 4. Can we predict the total length of a Koala based on multiple factors such as head length, skull width, and foot length?

To investigate whether we can predict the total length of a Koala based on multiple factors, we conducted a multiple linear regression analysis. We initially considered several body measurements as potential predictors, including head length, skull width, foot length, chest circumference, paw size, belly circumference, eye diameter, and ear size.



The visualization revealed strong correlations between total length and several predictors. However, it also indicated potential multicollinearity issues among some predictors. To address multicollinearity, we calculated VIF to refine our model by excluding variables with high VIF values and those showing strong correlations with other predictors. This process led us to a final model that includes **head length, skull width, foot length, and paw size** as predictors of total length.

Our final multiple regression model is statistically significant ($p < 0.05$) and explains 71% of the variance in total length (R-squared = 0.710). All four predictors in the model are statistically significant ($p < 0.05$), indicating they contribute meaningfully to predicting total length.

	coef	std err	t	P>|t|	[0.025	0.975]
const	39.6437	2.421	16.376	0.000	34.840	44.447
head_length	0.3963	0.103	3.850	0.000	0.192	0.601
skull_width	0.3201	0.149	2.141	0.035	0.023	0.617
foot_length	0.3555	0.059	5.995	0.000	0.238	0.473
paw_size	1.0752	0.131	8.237	0.000	0.816	1.334

Omnibus:	0.232	Durbin-Watson:	1.945
Prob(Omnibus):	0.890	Jarque-Bera (JB):	0.188
Skew:	-0.100	Prob(JB):	0.910
Kurtosis:	2.945	Cond. No.	344.

R-squared:	0.710
Adj. R-squared:	0.698
F-statistic:	60.61
Prob (F-statistic):	8.76e-26
Log-Likelihood:	-234.64
AIC:	479.3
BIC:	492.5

Our examination of diagnostic plots also confirmed that the model met the key assumptions of multiple linear regression, including normality of residuals and constant variance across predicted values. Overall, we can conclude that we can predict the total length of a Koala based on multiple factors such as head length, skull width, foot length, and paw size.

Question 5. Do environmental factors such as state affect Koala's physical characteristics?

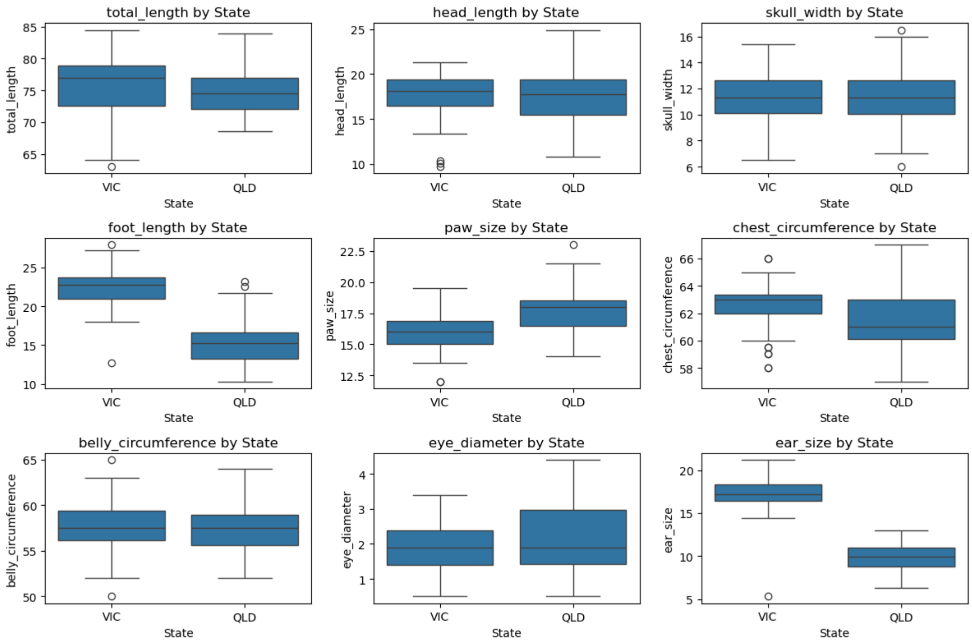
To investigate the influence of environmental factors on koala physical characteristics, we compared various measurements between koalas from Queensland (QLD) and Victoria (VIC) using independent t-tests. This approach allowed us to determine if there were statistically significant differences in physical traits between the two states.

We conducted t-tests for several physical characteristics, including total length, head length, skull width, foot length, paw size, chest circumference, belly circumference, eye diameter, and ear size. The results showed significant differences in some, but not all, of these traits.

Independent T-Test results:

	Characteristic	T-statistic	P-value	Statistically significant
0	total_length	-0.797	0.427	No
1	head_length	-0.076	0.940	No
2	skull_width	-0.129	0.898	No
3	foot_length	-12.902	0.000	Yes
4	paw_size	5.690	0.000	Yes
5	chest_circumference	-2.007	0.047	Yes
6	belly_circumference	-0.266	0.791	No
7	eye_diameter	1.281	0.203	No
8	ear_size	-18.775	0.000	Yes

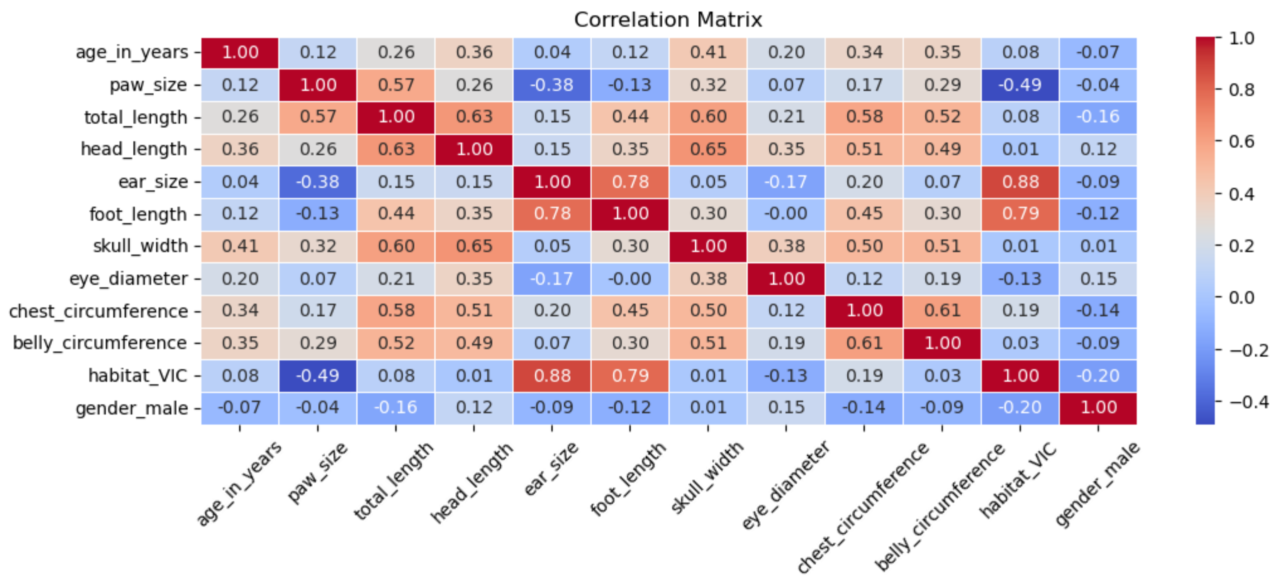
Notably, we found statistically significant differences in foot length, paw size, chest circumference, and ear size between QLD and VIC koalas. These variations are clearly visible in the box plots comparing the two groups.



Our analysis shows that VIC koalas tend to have longer feet and larger ears, while QLD koalas have larger paws. Chest circumference also differs significantly between the two groups, although there is some overlap in the distributions. These findings suggest that environmental factors associated with different habitats might influence certain physical characteristics of koalas. However, it is important to note that other factors such as genetic variation, local diets, and human interventions could also contribute to these variations.

Question 6. What factors are correlated with the total length of a Koala?

To identify factors correlated with the total length of koalas, we conducted a comprehensive correlation analysis. We examined relationships between total length and various physical characteristics, as well as categorical variables such as habitat and gender (encoded using one-hot encoding and dropped the first item). We began by building a correlation matrix to visualize the relationships between all variables:



This heatmap indicated several moderate to strong positive correlations between total length and other physical characteristics. Notably, head length (0.63), skull width (0.60), and chest circumference (0.58) showed the strongest relationships with total length. To validate these correlations and assess their statistical significance, we performed **Pearson correlation tests** for each factor against total length. The results are summarized in the following table and chart:

	Factor	Correlation	P-value	Statistically significant
0	head_length	0.634559	0.000	Yes
1	skull_width	0.602056	0.000	Yes
2	chest_circumference	0.577890	0.000	Yes
3	paw_size	0.565646	0.000	Yes
4	belly_circumference	0.518741	0.000	Yes
5	foot_length	0.443706	0.000	Yes
6	age_in_years	0.264556	0.007	Yes
7	eye_diameter	0.205051	0.037	Yes
8	ear_size	0.153644	0.119	No
9	habitat_VIC	0.078666	0.427	No
10	gender_male	−0.160200	0.104	No

Our tests and analysis confirmed that **head length, skull width, chest circumference, paw size, and belly circumference** are all strongly and significantly correlated with total length ($p < 0.05$). Foot length also has a significant moderate correlation with total length. Other factors such as ear size, habitat, and gender do not show statistically significant correlations with total length. However, we acknowledge that while these correlations and tests are valid in this case, they might do not imply causation because total length is likely also impacted by other factors that are not captured in this analysis.

SECTION 4: Recommendations

Our thorough analysis of koala physical traits has uncovered important findings for conservation efforts. We have identified strong correlations between the total length and measurements such as head length, skull width, and chest circumference, which serve as reliable indicators of the overall size of koalas. It is worth noting that we have observed significant regional differences between Queensland and Victorian koalas, especially in foot length, paw size, and ear size, indicating adaptations to local environments.

Gender differences in koalas are subtle but noticeable in body proportions. Female koalas have slightly higher foot-to-body ratios and greater variability in belly circumference, which may be linked to reproductive stages. It is necessary to note that certain traits like foot length and ear size vary based on habitat, but these traits do not strongly correlate with total body length. This suggests that there are complex interactions between the koalas' environment and their physiology.

These findings have important implications for conservation strategies. We suggest the development of standardized health assessment tools that are based on the strong correlations between body measurements. For instance, wildlife researchers could create a standardized index based on head length, skull width, and chest circumference to estimate a koala's overall body condition without invasive procedures. This would enable more efficient and less stressful health monitoring of wild populations.

Given the state variations, we recommend implementing conservation plans tailored to specific habitats. These plans should take into consideration local adaptations and could be used to guide focused habitat restoration efforts. To preserve genetic diversity, we propose establishing interconnected wildlife reserves or so-called "green bridges" between distant populations, with a particular focus on linking Queensland and Victoria. This could include creating protected areas that connect isolated koala habitats.

Moreover, gender differences and variations in female measurements might provide an opportunity to develop gender-specific approaches. For instance, studying the variability in female belly circumference could lead to non-invasive methods for assessing reproductive status, which would be valuable for population management. Lastly, we recommend initiating long-term studies investigating how these physical traits might impact koalas' resilience to climate change. This research could provide crucial insights for developing adaptive conservation strategies for koalas.

By implementing these data-driven approaches, we can enhance the effectiveness of koala conservation efforts and ensure the long-term survival of this iconic species in the Australian ecosystem.