Koala's Physical Characteristics Analysis

Name: Pradyot Jain

Student Number: 48479985

Word Count: 1912

Section 1: Exploratory Data Analysis

Step 1: Check for Missing Values

We can see missing values in 4 columns. To ensure data completeness, missing values need to be addressed, before proceeding with further analysis.

Step 2: Handle Missing Values

We replaced missing values with the median depending on gender, habitat, and region, we kept all data while properly portraying each variable's central tendency. Dropping columns would reduce the dataset size, which is significant given its modest size. After that, we verified, and we got zero null values.

Step 3: Check for Duplicate Rows

There are no duplicate rows as the "koala_id" is unique for every row, which ensures the integrity of the data.

Step 4: Similar Records

Only non-numerical columns were checked for similar records to ensure consistency. Variations in the habitat and gender columns were standardized as part of the data-cleaning procedure to maintain uniformity. Multiple state names were condensed into QLD and VIC, while gender data were standardized to male and female. This guarantees a consistent dataset, improving the analysis's accuracy and reliability.

Step 5: Checking Data Types

I checked the data type of the non-numerical columns. The "habitat" and "gender" columns were found to have the "object" data type.

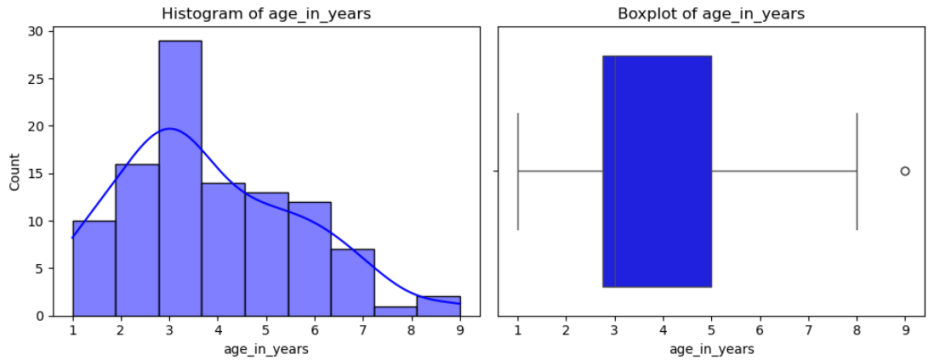
Step 6: Converting the Data Types

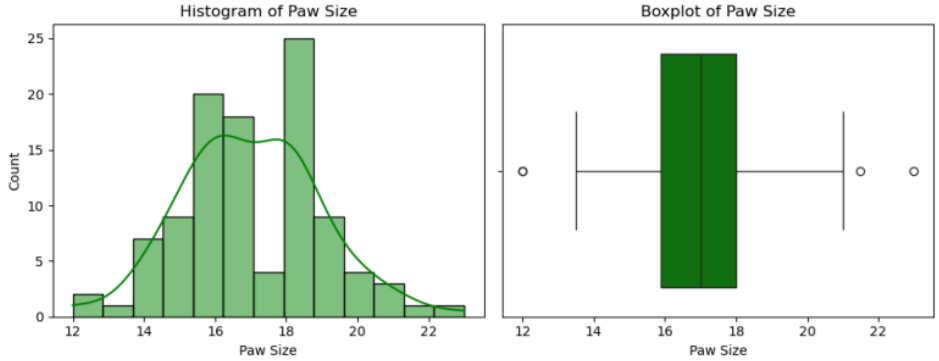
To improve efficiency and performance in the analysis, Data types of “habitat” and “gender” were converted to “category” data types.

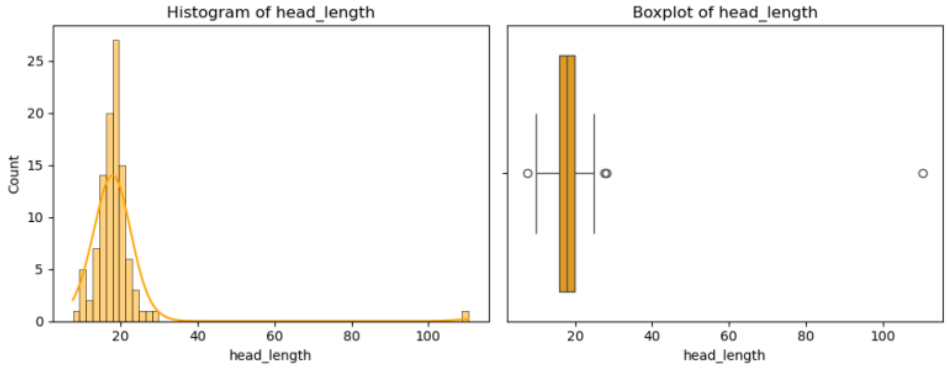
Step 7: Outliers

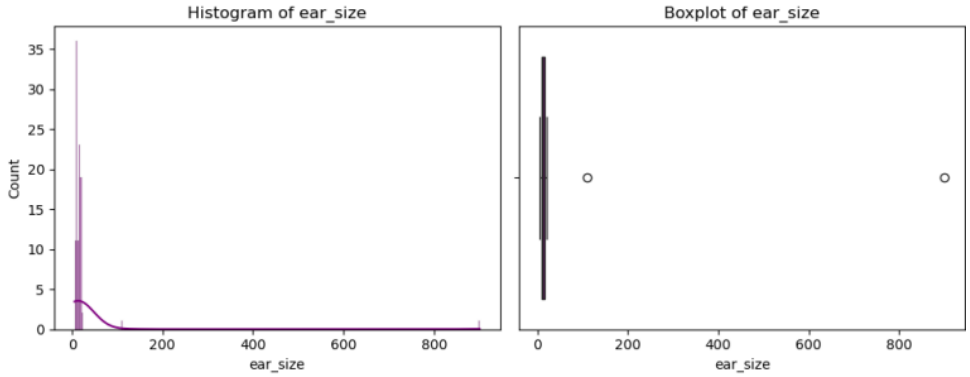
First, we checked how many outliers are there in the data set with the help of IQR. We got:

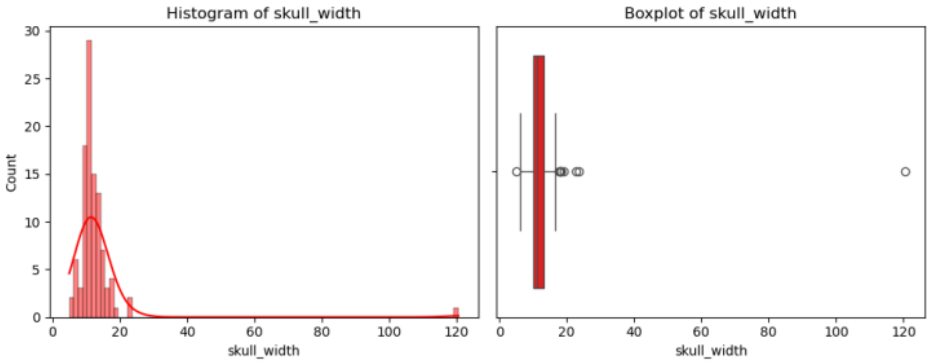
I have plotted the variable distributions with outliers using a boxplot or histogram, which enables us to determine whether the outliers deviate significantly from the main distribution yet remain within a reasonable range.

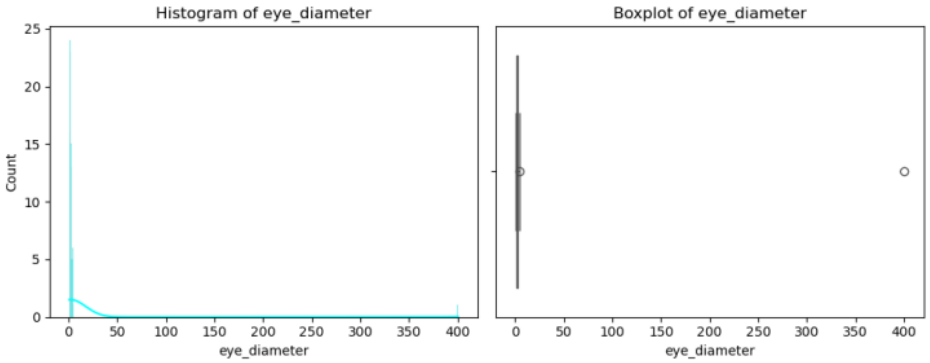

Histogram of age_in_years
Count
age_in_years
Boxplot of age_in_years
age_in_years

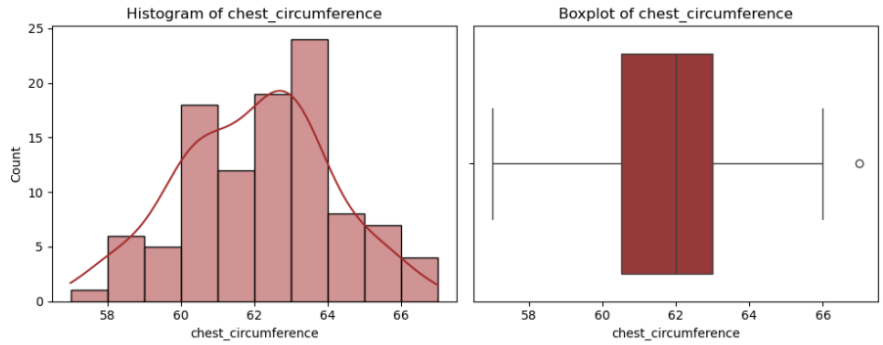

Histogram of Paw Size
Count
Paw Size
Boxplot of Paw Size
Paw Size

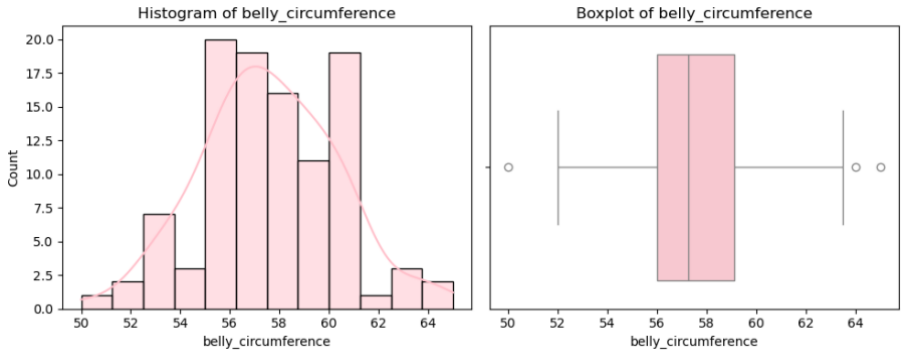

Histogram of head_length
Count
head_length
Boxplot of head_length
head_length


Histogram of ear_size
Count
ear_size
Boxplot of ear_size
ear_size


Histogram of skull_width
Count
skull_width
Boxplot of skull_width
skull_width


Histogram of eye_diameter
Count
eye_diameter
Boxplot of eye_diameter
eye_diameter


Histogram of chest_circumference
Count
chest_circumference
Boxplot of chest_circumference
chest_circumference


Histogram of belly_circumference
Count
belly_circumference
Boxplot of belly_circumference
belly_circumference

Reason for doing this: Even if the outliers are far from the main data cluster, they may still be legitimate if they match the distribution structure. They could require more research if they significantly depart from the overall pattern.

Step 8: Transformation of Outliers

1. Age in years:

- **Histogram**: The data is right-skewed, with koalas aged between two and five years.
- **Boxplot:** There is one outlier of about 9 years, which may be unusual considering the average age distribution.
- **Conclusion:** The 9-year-old outlier might be true but extreme age, indicating an older koala. Since the age is biologically possible we will retain them.

2. Paw Size:

- **Histogram:** The distribution is roughly normal, with a few values on the high end.
- **Boxplot**: There are a few outliers between 20 and 22 cm.
- **Conclusion:** Paw sizes of 20-22 cm are unusual compared to the average range, but they are biologically possible, so these outliers are retained.

3. Head Length:

- **Histogram:** This distribution is right-skewed, with a concentration of around 15-20 cm, although some values exceed 40 cm.
- **Boxplot**: Several extreme values exceed 40 cm and differ substantially from the median.
- **Conclusion:** Values greater than 40 cm are considered possible outliers because they are far beyond the normal range. I have removed these outliers as they are not biologically plausible. This approach ensures my analysis remains accurate and unaffected by extreme, unrepresentative data.

4. Ear size:

- **Histogram**: Highly skewed, with an odd spike at very high values, indicating potential data input mistakes.

• **Boxplot:** Extreme values appear well above the main cluster, indicating mistakes or highly exceptional situations.

• **Conclusion:** Extremely large results for ear size may represent data problems. Values that are much over the normal range are removed as not removing them will affect our analysis.

5. Skull Width:

• **Histogram:** The histogram shows a right-skewed distribution with most values below 30 cm and a few extreme values up to 120 cm.

• **Boxplot:** Several values are outliers, stretching well outside the main distribution, indicating either data input mistakes or extremely exceptional situations.

• **Conclusion:** Skull widths more than 30 cm, particularly those around 120 cm, are most likely mistakes or rare anomalies. The IQR approach was used to filter them out since it successfully controls skewed distributions and eliminates extreme values, resulting in only reasonable observations for trustworthy analysis.

6. Eye diameter:

• **Histogram:** The distribution is skewed, with most values on the low side and some quite high.

• **Boxplot:** Significant outliers extend well outside the main cluster, indicating probable data input mistakes.

• **Conclusion:** Values beyond 50-100 cm are likely above biological standards. The IQR approach reduces these severe outliers because it successfully manages skewed data and guarantees that only realistic eye diameter values are maintained for reliable analysis.

7. Chest circumference:

- **Histogram:** The distribution is significantly skewed, with most values ranging between 58 and 64 cm. The distribution indicates a peak at 60-62 cm.

- **Boxplot:** There is one outlier above the upper whisker, measuring slightly over 66 cm and deviating from the main cluster.

- **Conclusion:** The outlier at roughly 66 cm is valid because higher chest circumferences are common in some koalas. Since it is biologically possible, I have winsorize these outliers as this

method ensures that biologically plausible values are retained while reducing the influence of any extremely high values.

8. Belly Circumference:

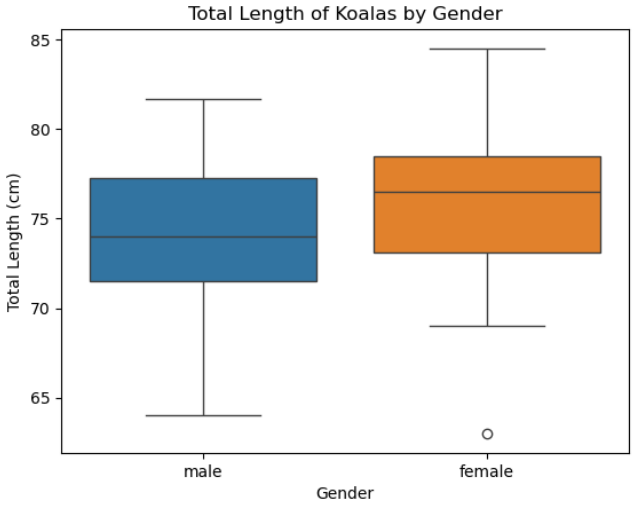
- **Histogram:** The distribution is more symmetrical, with a relatively typical distribution centred about 56-58 cm.

- **Boxplot:** There are some outliers above 62 cm. These numbers are beyond the upper whisker.

- **Conclusion:** Since belly circumferences in the 62-64 cm range are on the high side. After checking for the outliers I got that the outliers are female koalas so we can assume they are pregnant. Since these outliers are of female koalas, we can assume that are pregnant so I will not remove them but will retain them and cap them with winsorize method, which ensures that biologically plausible values are retained while reducing the influence of any extremely high values.

Section 2: Data Visualization

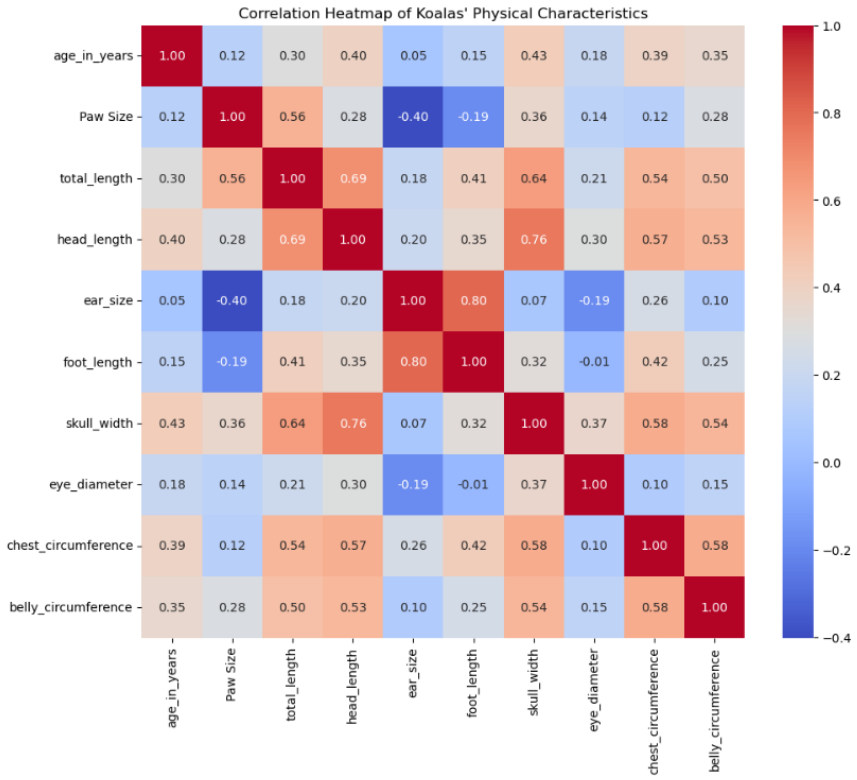
Visualization 1: Total Length of Koalas by Gender (boxplot)



Explanation: This boxplot compares the overall length of koalas by gender. The visualization was selected to show potential sexual dimorphism in the size of male and female koalas. Boxplots effectively display the distribution, median, quartiles, and outliers in the data.

Insight: According to the boxplot, male koalas generally have a longer total body length than female koalas. This supports the idea of sexual dimorphism in koalas. Although there are some outliers, the difference between the genders is still noticeable.

Visualization 2: Correlation of Koala Physical Characteristics (Heatmap)



Explanation: A heatmap was used to show the correlation between various physical attributes of koalas. Correlation matrices are useful for determining which variables are related to each other.

Insight: The heatmap reveals strong positive correlations between variables such as total length, foot length, and head length. These relationships suggest that koalas with longer bodies also tend to have larger feet and heads, showing proportional growth among physical characteristics.

Section 3: Analysis

Q1. Is the mean head length of the Koalas significantly different from 92.0 mm?

Ans. I have used one sample t-test. The one-sample t-test shows that the mean head length of the koalas is significantly different from 92.0 mm (p-value < 0.05).

Q2. Do male and female Koalas have significantly different mean head lengths?

Ans. I have used a two-sample t-test. Since the mean head lengths of male and female koalas are not significantly different (p-value ≥ 0.05), I can conclude that gender does not significantly impact head length in this population.

Q3. Can we predict the total length of a Koala based on its head length?

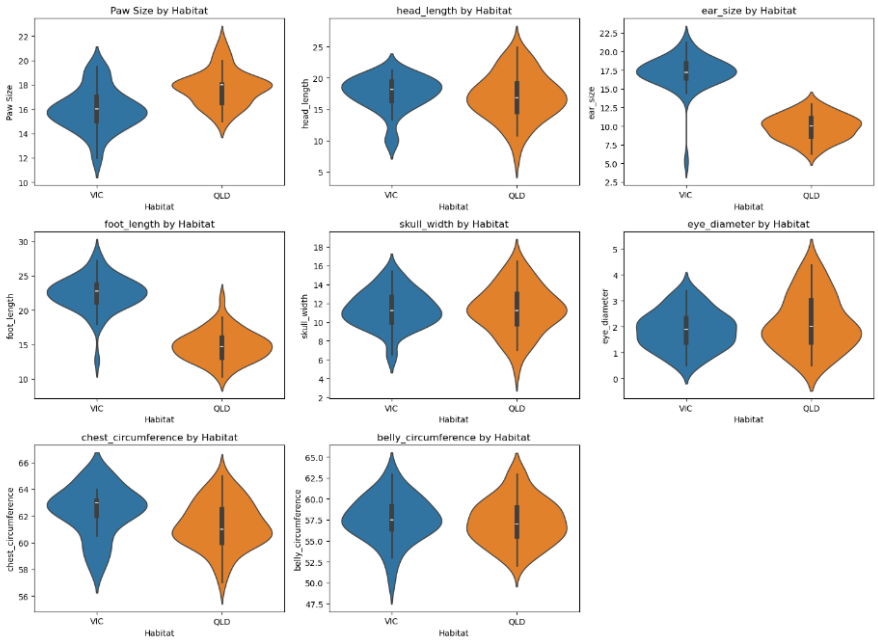
Ans. For this, I have used **Ordinary Least Square (OLS)** to find the R2 score. R-squared value to see how well head length explains the variability in total length. The model suggests that head length can be used to predict the total length of a koala to some extent, as head length is a statistically significant predictor. However, because the R-squared value **(0.471)** is moderate, head length alone cannot entirely explain the overall length and other physical or biological properties may also be involved. As a result, while head length is useful for predicting total length, adding more predictors could improve the model's accuracy.

Q4. Can we predict the total length of a Koala based on multiple factors such as head length, skull width, and foot length?

Ans. I used **Ordinary Least Squares (OLS)** to assess how well **head length**, **skull width**, and **foot length** predict koala total length, achieving an R-squared value of **0.531**. This indicates a reasonable predictive power but suggests other factors also contribute to total length. **Testing all physical traits** raised the R-squared to **0.732**, showing that although additional traits improve the model's fit, not all are significant predictors. Focusing on key predictors or exploring other variables could enhance accuracy and provide a clearer understanding of what influences koala size.

Q5. Do environmental factors such as state affect Koala's physical characteristics?

Ans. For this, I have used group statistics of all the physical features based on "Habitat" and I also created a violin plot for better understanding.



I can observe some patterns in koala physical characteristics based on habitat from the data and visualizations.

Observation:

- **Paw Size, Foot Length, and Chest Circumference**: Koalas in Victoria have bigger paws, foot lengths, and chest circumferences than those in Queensland, indicating probable adaptations to local vegetation or climate.
- **Head Length and Skull Width:** Average head lengths vary throughout environments, but VIC koalas have a higher variety. Habitat appears to have less of an influence on skull width.
- **Ear Size and Eye Diameter:** These features vary little between habitats, suggesting that they are more genetically determined.
- **Belly and chest circumference:** VIC koalas have a somewhat greater variation in these measurements.

Conclusion: The data suggest that koalas in various habitat have varied physical characteristics, with paw size and chest circumference changing significantly between VIC and QLD. These variances are most likely the result of adaptations to unique environmental conditions such as climate and food availability. Traits like ear size and eye diameter, on the

other hand, stay consistent throughout habitats, suggesting that they are genetically fixed rather than modified by environmental factors.

Q6. What factors are correlated with the total length of a Koala?

Ans. The factors correlated with the total length of Koalas are:

- **Head Length, Skull width, Paw Size, and Chest Circumference:** These characteristics have the strongest correlations with total length, making them reliable predictors of a koala's overall body size.
- **Belly Circumference, Foot Length, and Eye Diameter:** These traits show weaker correlations with total length, indicating they are less effective in predicting a koala's overall length.
- **Ear Size:** This feature has no significant association with total length, suggesting it does not contribute to predicting body size.

Overall: The dimensions related to the chest, belly, paws, and feet are the most influential factors in determining a koala's total body length.

Section 4: Recommendation

The study of koala's physical characteristics such as head length, skull width, and chest circumference give significant information on their health and adaptability in various habitats. Notably, koalas in Victoria and Queensland have unique physical traits, most likely due to adaptations to local environments. This indicates that conservation methods should be tailored to the unique needs of koalas in different settings. Conservation activities that focus on maintaining and restoring natural ecosystems, guaranteeing enough food supplies, and avoiding habitat fragmentation can considerably improve koalas' chances of prospering. Furthermore, monitoring these physical characteristics provides a realistic method for assessing the health of koala populations over time. Such focused tactics allow conservationists to adapt to the environmental problems encountered by koalas, eventually helping their long-term survival and well-being.