Statistical analysis of weight and physical differences between cockatoos

STAT6170 Statistical Report

Do Nam Phong Phung - 47828013

I. Introduction

Cockatoo is one of the most famous bird species in Australia. There are many types of cockatoos and they are usually categorized by their personalities, living habits and appearances. The analysis aim to answer two key research questions by learning a dataset of 218 sample cockatoos, in which, we will consider different statistics of them, such as wing span, weight and body length.

In this report, the main objective is to further understand the physicality differences between black and white cockatoos in additional to how physical characteristics such as weight, wing span and body length supports to differentiate between the two types. Specifically, the analysis aims to study if there is a signifcant difference of the average weight between black and white cockatoos and to explore the relationship between their wing span and body length (whether if a cockatoo with a large wing span also have a long body).

The knowledge from this report could be important in nature science and environment preservation. The insights allow experts to not only differentiate black and white cockatoos, but also better assess cockatoos health status and their adaptability to different environment.

II. Methods

Taking into consideration that the data was collected and pre-processed in advance, the primary task of the research is to conduct statistical tests with the given data, provide in-depth analysis and suggestions according to the insights. The analysis will be conducted using R programming language in Markdown format for reproducibility and convenience in information presentations. There are two R packages are ultilzied in the report including `tidyverse` for data manipulation, visualization and `gridExtra`, `knitr` and `kableExtra` for visualization supports. This report focuses on two key questions:
1. Is there any difference in the average weight of black and white cockatoos?
2. What is the relation between the wingspan of cockatoos and the body length?

To answer the first key question, which requires to explore the difference between the average weight of black and white cockatoos, a two sample t-test will be conducted to understand the relationship between different `weight` at different `colour`. Considering the two weight samples of white and black cockatoos are two independent samples as well as we are comparing two numerical sets of data, a two sample t-test is the most suitable method for the analysis. Additionally, it is required to check the data assumptions before conducting the test, which includes examining the variance equality and the normality of the distribution.

In order to determine the relationship between the wingspan of cockatoos and their body length, we will look at the impact of body length on the size of the wing span. A linear regression model will be fitted, with the wing span is defined as dependent and the body length to be considered as the independent variable. After constructing the model, validation steps should be taken to make sure that the model is reliable and meets required assumptions. Moreover, as the `colour` of the cockatoos is considered to be a significant interaction terms that may affect the model prediction, it is important to consider this matter when training the Linear models. The first two models, which are trained with the default data, will consider the usage of `colour` as the interaction term. While the two final models will investigate the effect of training with two colour-separated subsets.

III. Approaches

As the main approach to achieve the answer the ultimates questions are by using statistical test, it is essential to set up the hypotheses for each test. Knowing each question hypotheses allows not only the test but also the analysis to follow the right direction.

1. The difference in the average weight of black and white cockatoos

To determine whether the average weight of black cockatoos differs from white cockatoos, we test the hypotheses associated with with this relationship, which is formulated as:

- The null hypothesis states that the mean weight for both cockatoo groups are equal

$$H_0 : \mu_1 = \mu_2$$

- The alternative hypothesis states that the mean weight for both cockatoo groups are different

$$H_1 : \mu_1 \neq \mu_2$$

- Where:
  - μ_1 is the mean weight of black cockatoos
  - μ_2 is the mean weight of white cockatoos

2. The relationship between the wingspan and the body length of cockatoos

In order to study the relationship between the wing span and the body length of the cockatoos, we formulate the hypotheses that describe the relationship, which is:

- The null hypothesis states that there is no significant linear relationship between wing span and body length

$$H_0 : \beta_1 = 0$$

- The alternative hypothesis states that there is a significant linear relationship between wing span and body length

$$H_1 : \beta_1 \neq 0$$

- Where:
  - β_1 is the population slope

IV. Results

1. The difference in the average weight of black and white cockatoos

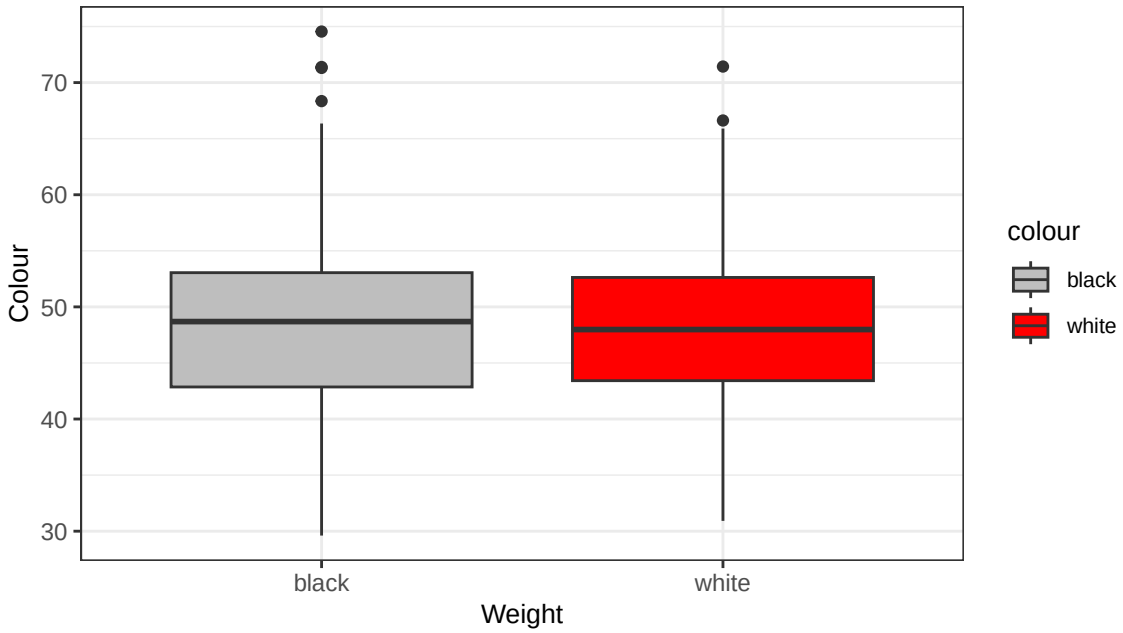
In the dataset, there are 218 random cockatoos subjects belongs to two colour group of black and white. Table 1 demonstrates that there are 54.6% of the studied subjects are black. It is visible that the mean and std of weight of each group are similar.

Table 1: Statistics between black and white cockatoos

Colour	Count	Mean of weight	STD of weight
black	119	48.63487	9.022529
white	99	48.64960	8.175473
Total	218	48.64156	8.628557

To test the hypothesis, it is important to investigate if the data matches all the required assumptions. To check the variance equality assumption, we will use a boxplot to observe the weight spread.

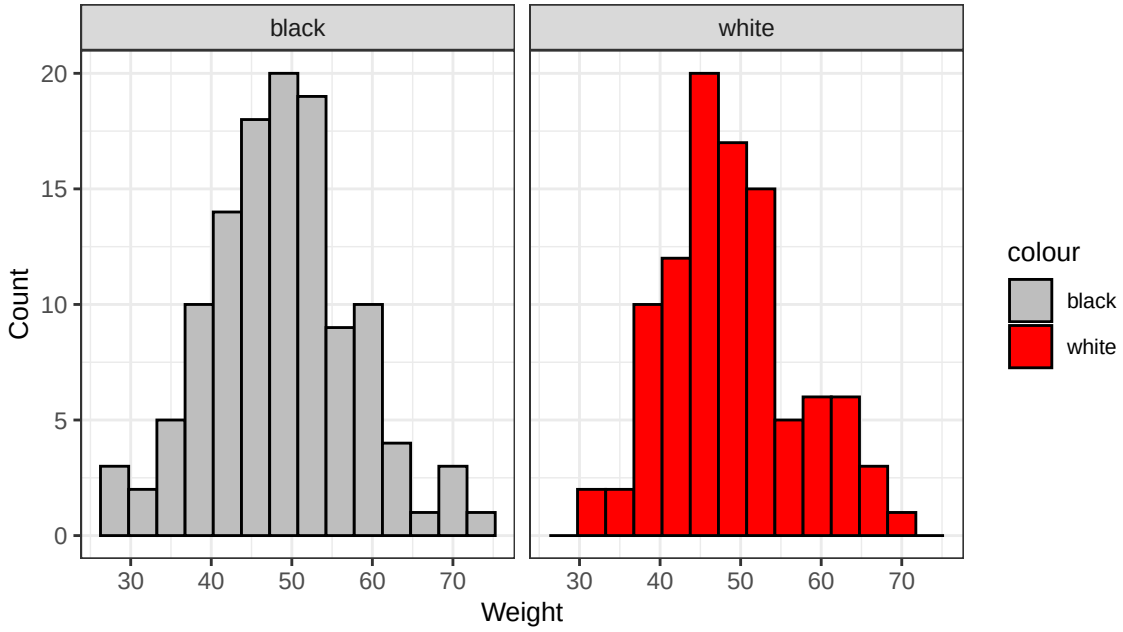
Figure 1: Box plot of black and white cockatoos weight



Observation shows that both black and white cockatoos have an equal variance in their weight. The similar conclusion can be found when calculating the standard deviation of each colour's weight in table 1. The standard deviation of the two groups are close to each other.

To explore the normality of each colour's weight, histograms are used to describe the distribution of weights between black and white cockatoos. Based on Figure 2, both histograms illustrate a bell-shaped distribution of the two groups' weight, which proves that both samples are drawn from normal distributions.

Figure 2: The weight distribution of black and white cockatoos



As both of the assumptions are valid, it is reasonable to apply a two way t-test to test the hypotheses. The test results are visualized in table 2, by applying `weight` as the target variable and `colour` is the predictor. Moreover, `var.equal` is set to `TRUE` to present equal variance and the confidence level is 95%.

Table 2: Two way t-test results

	t.statistic	p.value	CI.Lower	CI.Upper	Mean.of.black	Mean.of.white
t	-0.0125138	0.9900272	-2.333539	2.304095	48.63487	48.6496

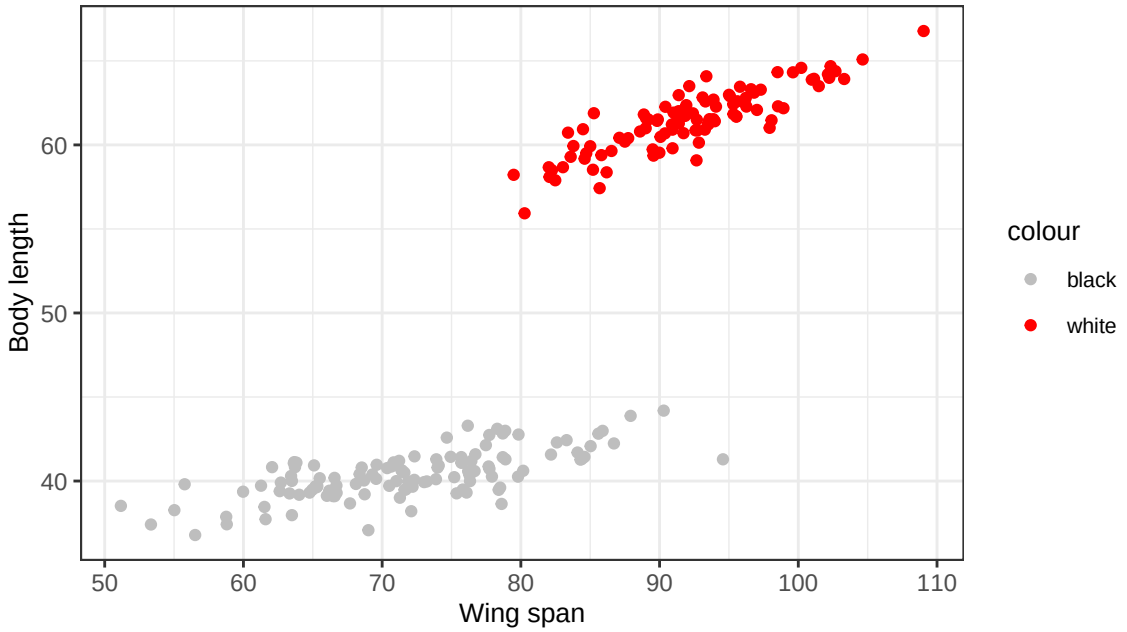
We do not reject the null hypothesis H_0 as the t-test yields a p-value of 0.99, which is larger than 0.05. There is insufficient evidence to suggest that the mean weight between black and white cockatoos is different to each other. This means that the average weight of the two types are likely to be similar. The confidence interval suggests that there is a 95% to observe the weight of black cockatoos to be between 2.33 grams less and 2.30 grams more compared to the white cockatoos.

2. The relationship between the wing span and the body length of cockatoos

a. The default relationship of `wingspan` and `bodylength`

To test the formed hypothesis, a linear regression model is fitted with `wingspan` as the target variable and `bodylength` as the predictor. The results of the model are presented in Table 3. After fitting the model, it is important to investigate if the data matches all the required assumptions. A scatter plot is plotted using `wingspan` and `bodylength` variables to explore the relationship between `wingspan` and `bodylength` (Figure 3).

Figure 3: The relationship between wing span and body length



By observing, we can see a general non-linear pattern of the data. Taking `colour` into consideration, there are two separated areas where the data points concentrates on based on the cockatoo color, the black subjects have lower wing span and body length with their data points lies on the bottom left of the plot. The white cockatoos, in contrast, are likely to have higher body length and wing span. It is noticable that there are (splitted) linear trends in the plot, where cockatoos with bigger wing span tend to have longer body. The insights suggest a sign that the relationship between `wingspan` and `bodylength` could be affected by an interaction term `colour`. However, as we are considering the whole dataset together, this pattern suggests a violation in linearity.

Using validation plots from the linear model (Figure 4), we can examine the model residuals from a normal distribution and have an equal variance. The Q-Q plots of the model yields a linear pattern, which proves the normality of residuals. On the other hand, while observing the `body length vs residuals` graph, unconstant variance can be seen. The distribution of the residuals and body length does not form a horizontal pattern across the x-axis, thus violates the homoscedasticity assumption. Therefore, the model is considered to be invalid and it is reasonable to discontinue analyzing.

Figure 4: Validation plots of the default linear model

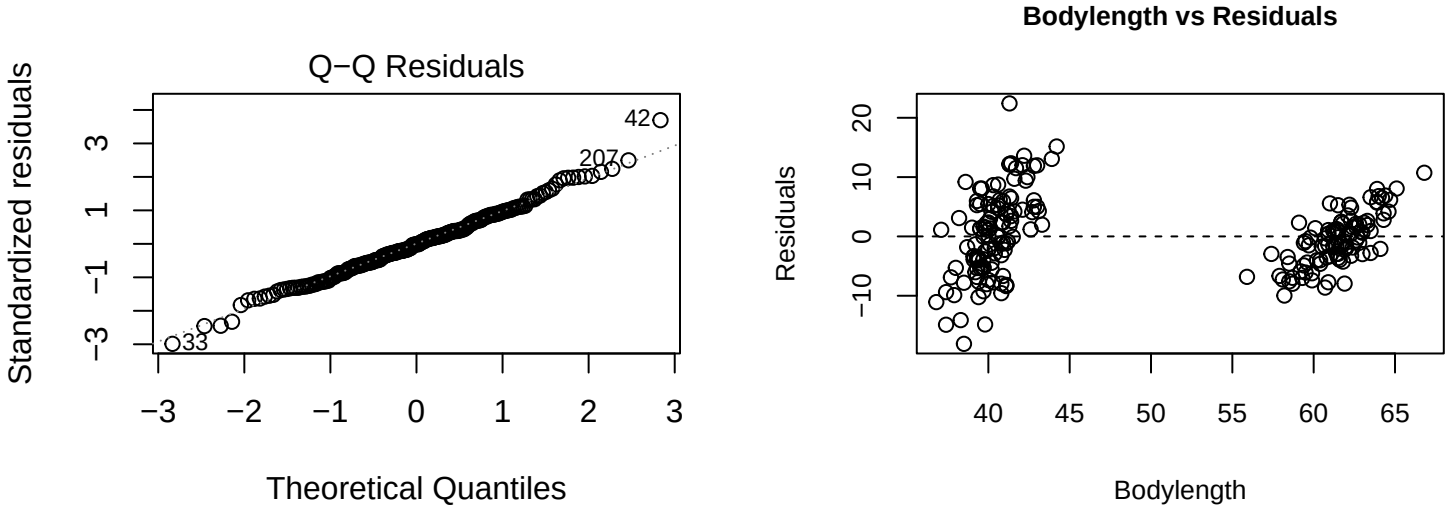


Table 3: Linear Regression model

	Estimate	Std..Error	t.value	Pr...t..
(Intercept)	29.918352	1.9804327	15.10698	0
bodylength	1.023184	0.0387714	26.39015	0

Statistic	Value
R-squared	0.7632721
Adjusted R-squared	0.7621761

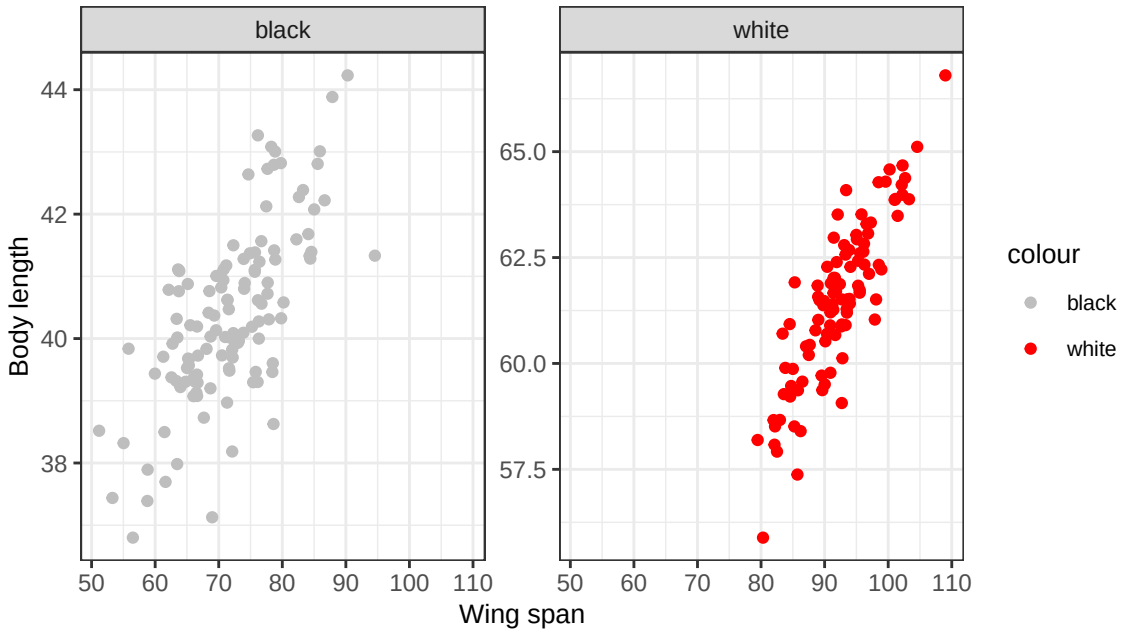
b. The relationship of `wingspan` and `bodylength` when splitting the dataset based on colour

In figure 3, a seperated but linear pattern was found, while in figure 4, the residuals versus predictor `bodylength` plot also yields two data group jitters. Observing table 4, the white cockatoos are found to have a 30% larger wingspan and 50% longer body in average, shows a clear difference between the two colour in `wingspan` and `bodylength`. Hence, it is sensible to attemp training linear regression models with seperated data by colour, in order to not only achieve a linearity assumption, but also to understand if `colour` has an impact on the relationship of `wingspan` and `bodylength`.

Table 4: Wingspan and Bodylength statistics of each colour group

colour	Mean of wingspan	STD of wingspan	Mean bodylength	STD of bodylength	Count
black	71.74034	8.132018	40.36807	1.415707	119
white	92.20808	5.909862	61.48687	1.893525	99

Figure 5: The relationship between wing span and body length



Two linear models are fitted with two seperated data based on their colour. The results can be found on Table 5 and 6. After fitting the model, the three linear assumptions are checked for both models. Figure 5 illustrates the relationship of `wingspan` and `bodylength` between black and white cockatoos. Generally, linear relationship can be observed in both data, while black's pattern seems to have a bigger variance compared to a more uniform pattern of white subjects. Observing the Q-Q plot (Figure 6-7), a straight pattern in both Q-Q residuals plots suggests that the data follows a normal distribution. All of the Bodylength vs. Residuals plots (Figure 6-7) have a horizontal linear pattern, indicates homoscedasticity - constant variance of residuals. Even though both of the two plots show signs of outliers, they can be neglected due to insignificance. As all assumptions are met, it is reasonable to further analyze the fitted models.

Figure 6: Validation plots for the linear model of black cockatoos

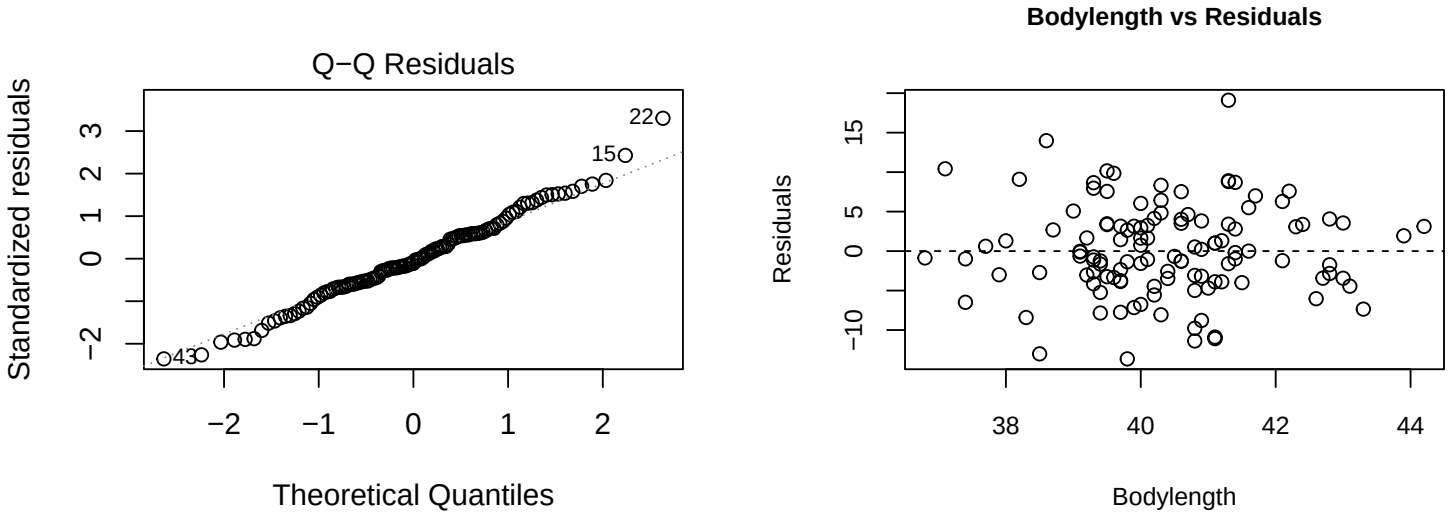
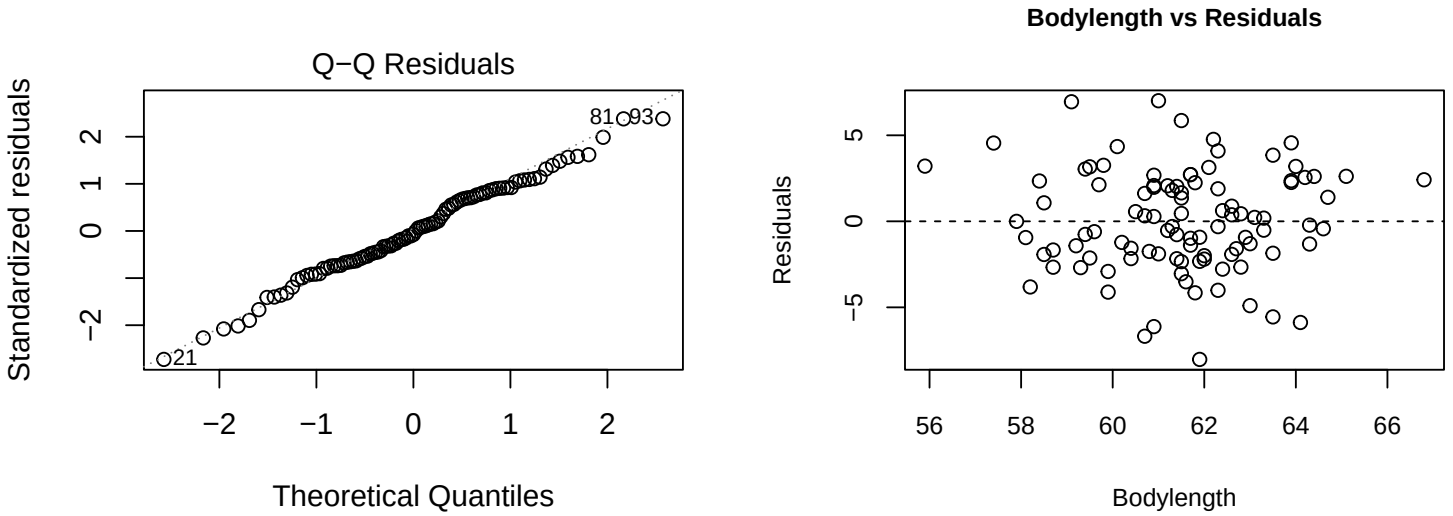


Figure 7: Validation plots for the linear model of white cockatoos



Both models (Table 5-6) yield p-value < 0.05 at each coefficient, therefore null hypothesis H_0 is rejected. There is a significant linear relationship between `wingspan` and `bodylength` in each model. With black cockatoos, 1 unit increase in body length equals to 4.02 unit increases in wing span. Each additional 1 unit increase in body length of white cockatoos results in 2.70 unit increases in wing span. With the model of black cockatoos, the R-squared value suggests that 49.2% of the variation in wing span can be explained by the linear relationship with body length. In contrast, the R-squared of the white model indicates that 75.2% of the variation in wing span can be explained by the linear relationship with body length, which is interpreted that body length is a better predictor of wing span in white cockatoos compared to black cockatoos. In Figure 8, the relability of the model on white subjects can also be studied, as the variance of the white model is visibly smaller compared to its counterpart.

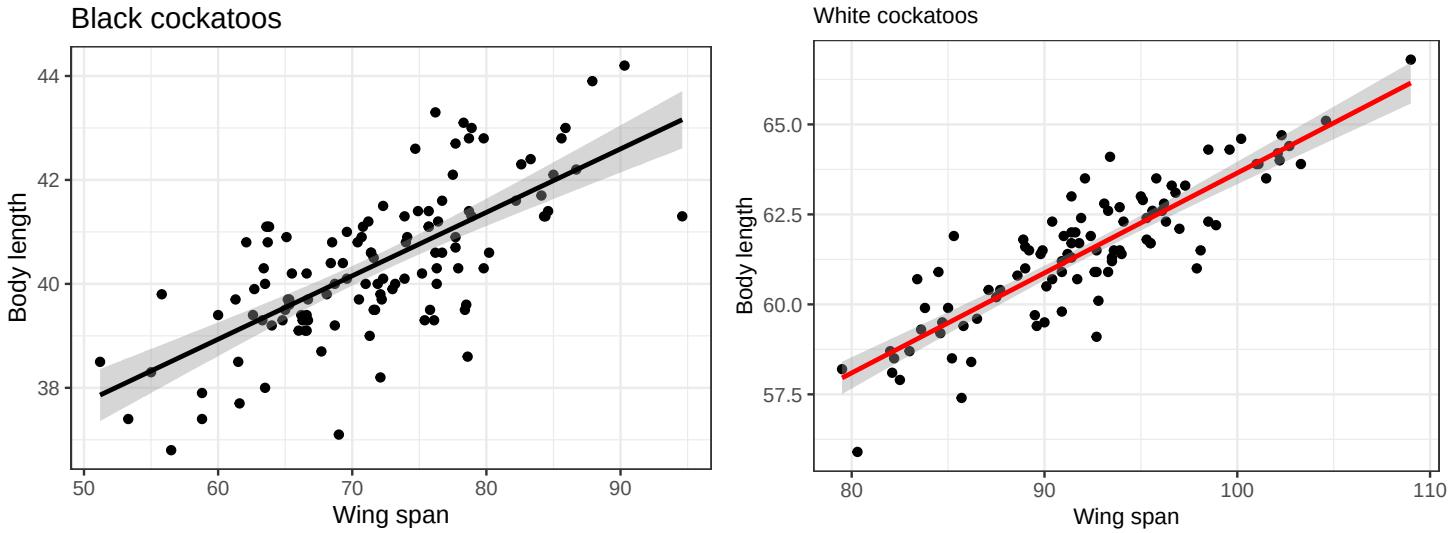
Table 5: Linear Regression model of black cockatoos data

	Estimate	Std..Error	t.value	Pr...t..	Statistic	Value
(Intercept)	-90.854847	15.2932553	-5.940844	0	R-squared	0.4916878
bodylength	4.027817	0.3786146	10.638304	0	Adjusted R-squared	0.4873433

Table 6: Linear Regression model of white cockatoos data

	Estimate	Std..Error	t.value	Pr...t..	Statistic	Value
(Intercept)	-74.167574	9.7153539	-7.634058	0	R-squared	0.7516270
bodylength	2.705873	0.1579329	17.133059	0	Adjusted R-squared	0.7490665

Figure 8: Model fit to each data



V. Conclusions

Studying on cockatoos can provide experts and scientist useful knowledge to support assessing their health status, habits and adaptability in different environment. The report focuses on studying different physical characteristics of cockatoos and how different colours affect they physicality. In order to determine these relationships, different statistical tests were conducted.

Firstly, it was revealed that there is no difference between weight of black and white cockatoos. A two way t-test between weight and colour yielded a insignificant relationship between the target and the predictor, thus declared that the two types of cockatoo samples have similar weight. By splitting the data into two subsets, we were able to met the linearity assumptions, thus allowed a valid linear regression model. Using the model, a linear relationship between wing span and body length of cockatoos was found. The linear regression models highlighted that an increase in wing span associated with an increase in body length. In additional, the difference in colour of cockatoos was discovered to have distinct contribution to the relationship between wing span and body length. Even though the coefficient shows a smaller positive linear relationship on white cockatoos, the study indicated that body length is a better predictor of wing span in white cockatoos compared to black cockatoos.

Taking into the consideration that we used a single Linear Regression model, it is more suitable to train the the model with splitted data subsets based on colour. The original distribution of `wingspan` and `bodylength` suggested a non-linear relationship in general, which violated the linearity assumptions, thus made the linear model invalid. However, a single linear pattern of this relationship was found in each spliited data, hence confirmed a valid linearity assumption and allowed the linear regression model to work. In future works, to overcome this problem, a multiple linear regression model could be trained using `colour` as the interaction term. This method will not only increase the speed of analyzing, but also make it possible to directly observe the impact of `colour` on the relationship.