APPLIED STATISTICS AND PROBABILITY

Rohith Ganesan

1. INTRODUCTION:

This analysis is designed to support an ornithologist examining collected data on kittiwakes, a gull species. The provided datasets encompass observation, historical, measurement, and location data. The observation data includes the number of kittiwake sightings at different times throughout four weeks, while historical data covers breeding pairs at five sites over five years. Measurement data includes weight, wing span, and culmen length for both black-legged and red-legged kittiwakes. The location data involves breeding pairs in 29 colonies and relevant covariate information.

Our objective is to provide insights to the ornithologist by addressing several key questions. Firstly, we will conduct an exploratory data analysis on the observation data to gain a clear understanding. Specifically, we are tasked with constructing a 99% confidence interval for the mean number of kittiwakes observed at dawn. This interval serves as an estimate of the range within which the true population parameter might exist.

Secondly, we will examine the historical data to determine whether the ornithologist's hypothesis of declining kittiwake numbers over time is independent of the site. An estimate for the number of breeding pairs at site E in 2006 is also requested.

Thirdly, we use the measurement data to provide a visual summary and assess the independence between wing span and culmen length for each sub-species. We will also investigate if there is evidence of a difference in weights between the two sub-species and assess overall significance.

Fourthly, with the location data, the ornithologist has tasked us with fitting linear statistical models and selecting the most appropriate model for the data.

This analysis endeavours to assist the ornithologist in extracting valuable insights from the collected data and providing statistical solutions to the posed questions using R Studio.

QUESTION: 1

For a better understanding of the data, we use the observation data to conduct exploratory data analysis on it. In statistics, exploratory analysis refers to the process of examining and summarising data in order to gain preliminary insights and understanding. It is primarily accomplished through the use of various graphical and numerical techniques to discover patterns, relationships, and distributions within data without reaching any formal statistical conclusions or decisions.

dawn	noon	mid.afternoon	dusk
Min. : 66.0	Min. : 50.00	Min. : 59.00	Min. : 85.0
1st Qu.: 95.5	1st Qu.: 63.50	1st Qu.: 80.50	1st Qu.:108.5
Median :100.5	Median : 77.50	Median :100.00	Median :125.5
Mean :105.2	Mean : 76.61	Mean : 96.75	Mean :123.7
3rd Qu.:121.2	3rd Qu.: 85.50	3rd Qu.:115.00	3rd Qu.:139.5
Max. :139.0	Max. :115.00	Max. :135.00	Max. :157.0

Summary of the Observation Data.

Data Distribution:
- The range of observations varies across the different times of the day.
- The minimum number of kittiwakes observed is highest at dusk (85) and lowest at noon (50).
- The maximum number of kittiwakes observed is highest at dawn (139) and lowest at noon (115).

Central Tendency:
- The mean number of kittiwakes observed is highest at dawn (105.2) and lowest at noon (76.61).
- The median (50th percentile) is a measure of central tendency, and it suggests that the middle observation is higher at dawn (100.5) compared to other times of the day.

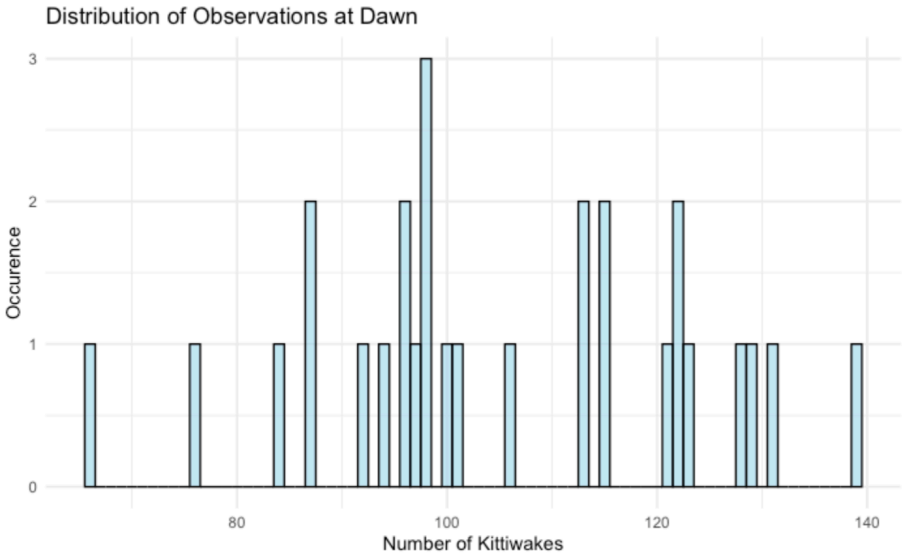
Variability:
- The interquartile range (IQR) provides a measure of variability. It is the range within which the central 50% of the data falls.
- The IQR is relatively consistent across dawn, noon, mid-afternoon, and dusk.

Outliers:
- Outliers can be identified by examining values significantly higher or lower than the bulk of the data.
- No specific mention of outliers is made in the provided summary.

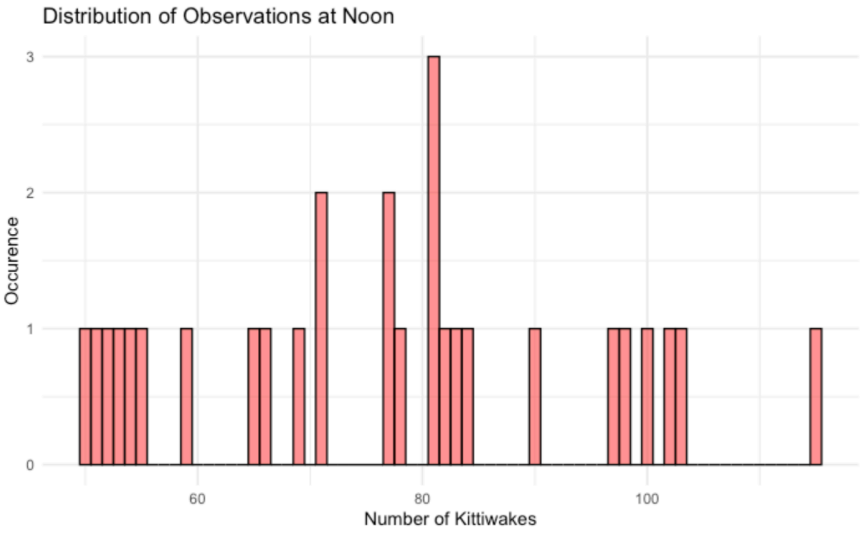
Overall Trends:
The mean and median provide an overall sense of the central trend in the data. The mean is influenced by extreme values, while the median is not.

Visual Representation:
Creating visualizations such as box plots or histograms can complement the summary statistics and provide a clearer picture of the data distribution.

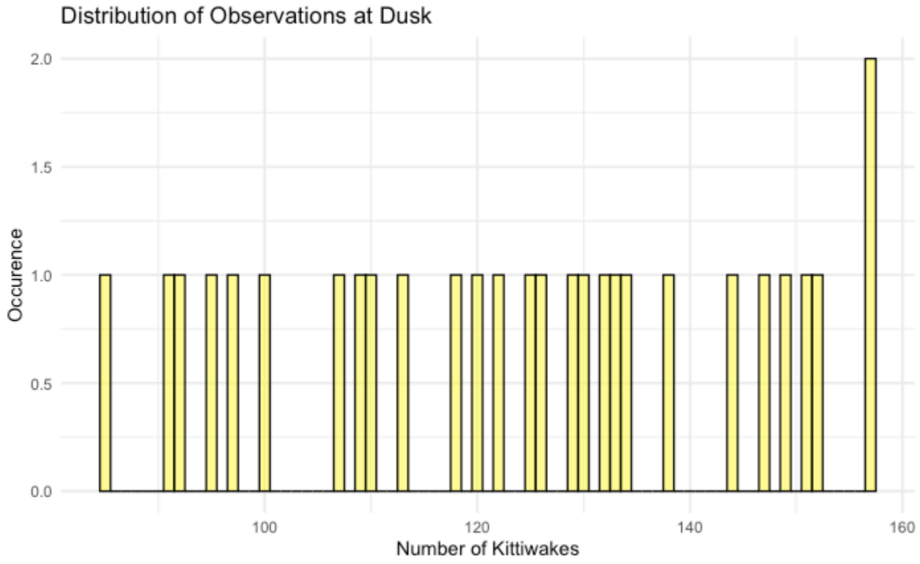
We plot histograms for each column of data after thoroughly reviewing the summary table. This graph plotting technique will assist us in visualising the data and gaining insights into the number of kittiwake sightings over the course of four weeks.



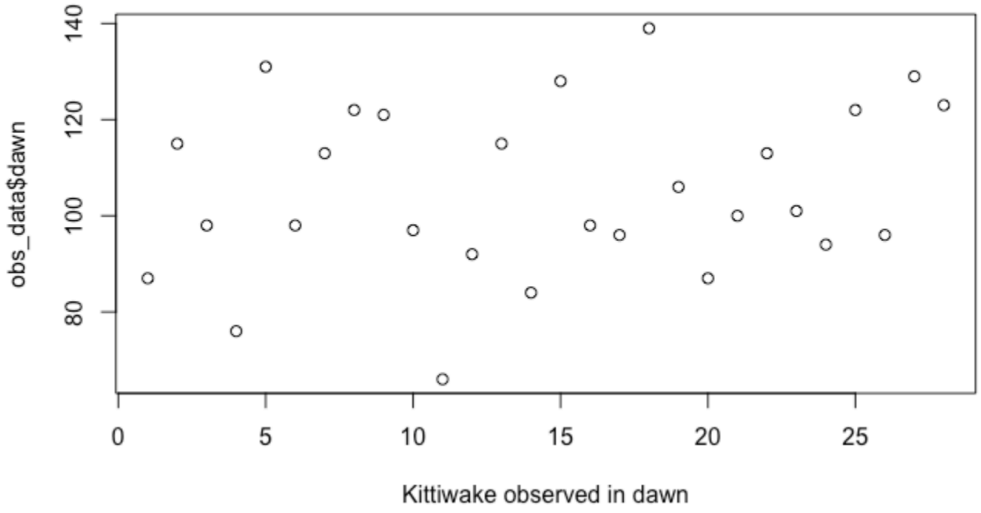
Kittiwake sighting distribution at dawn.



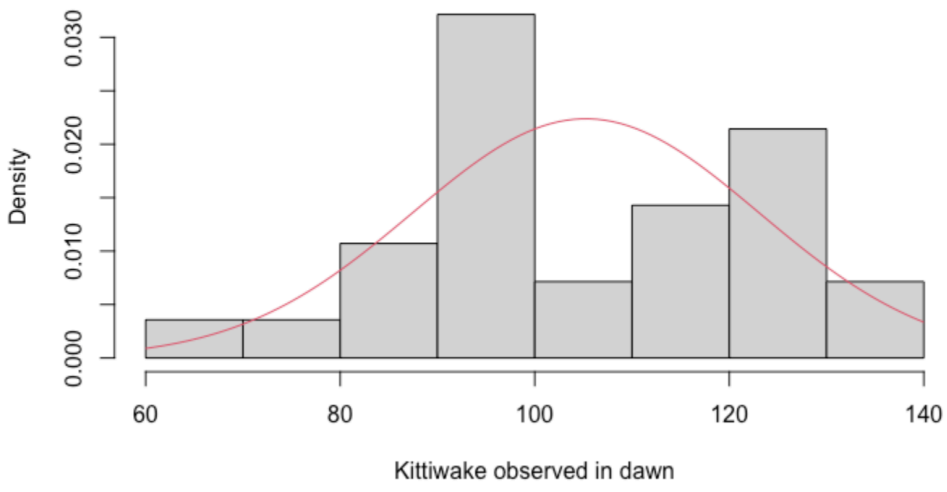
Kittiwake sighting distribution at Noon.



Kittiwake sighting distribution at Dusk.



Scatter Plot at Dawn.



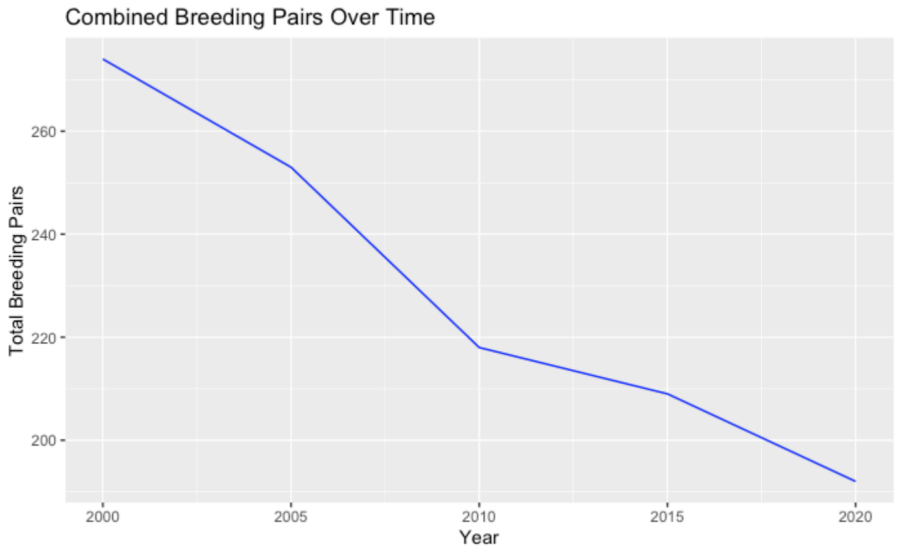
Density Plot at Dawn.

In order to establish a 99% confidence interval for the mean number of kittiwakes observed at dawn, an exploratory analysis was initiated. A histogram of the observation data was generated, illustrating the distribution of dawn sightings. The histogram indicated a normal distribution of the data. Subsequently, a statistical approach using a t-test was employed to calculate the confidence interval for the mean. The resulting 99% confidence interval, derived from the t-test, ranged from 95.92 to 114.58. This interval provides a measure of the plausible range within which we can be 99% confident that the true mean number of kittiwakes observed at dawn lies. This analysis not only presents a visual representation of the data distribution but also quantifies the uncertainty associated with the mean, aiding in a comprehensive understanding of the observations at dawn.

QUESTION: 2

The examination of historical data pertaining to kittiwake breeding pairs across different sites provides valuable insights into the potential factors influencing the observed decline over time. Plots illustrating the number of breeding pairs at Sites A, B, C, D, and E reveal nuanced trends, indicating that the decline may not be entirely independent of the specific site.

A subsequent Pearson's Chi-squared test was conducted to formally assess the independence between the decline and site-specific factors. The test yielded a chi-square statistic of 30.513 with 20 degrees of freedom, resulting in a p-value of 0.06196. While this p-value exceeds the conventional significance threshold of 0.05, implying a lack of strong statistical evidence against the null hypothesis, it is essential to consider the marginally higher p-value in the context of the ornithologist's hypothesis.



Fig() Line plot of combined breeding vs time

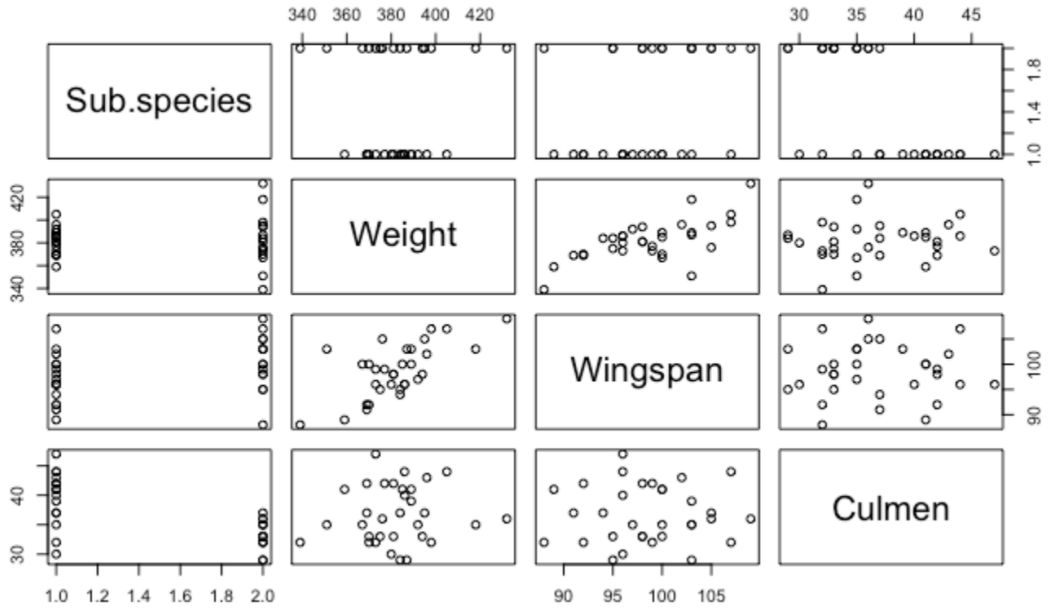
The non-significant result suggests a tentative lack of association between the decline and site, yet visual inspection of the plotted data implies potential site-specific trends influencing the overall decline. Additionally, the estimate for the number of breeding pairs at Site E in 2006, derived through linear interpolation, stands at approximately 67.4.

In summary, the chi-square test yields a marginally non-significant result, highlighting the need for cautious interpretation, while the interpolated estimate provides valuable information for Site E in 2006.

QUESTION: 3

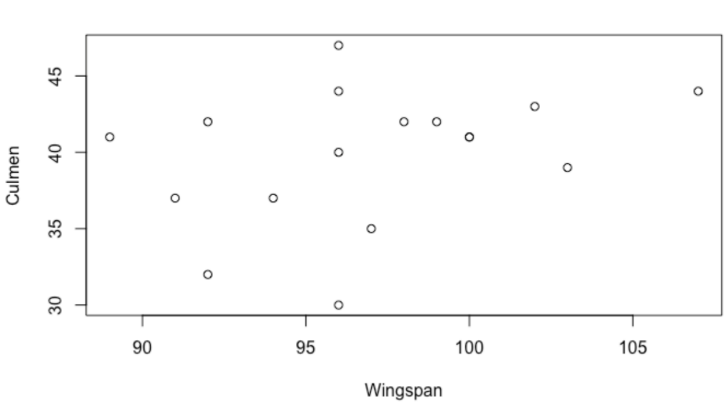
In examining the measurement data for kittiwakes, a comprehensive analysis was undertaken to address various aspects:

a) Visual Summary of Data: The measurement data, which includes information on Sub.species, Weight, Wingspan, and Culmen, was initially explored. A preview of the dataset using head() showcased the first few rows, and visualizations generated through plot() provided an overview of the distribution of variables. Additionally, statistical summaries obtained from summary() offered key insights into the central tendency and variability of the data.

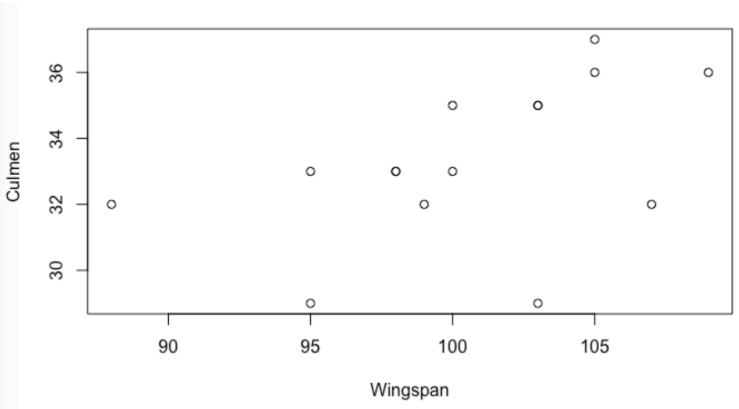


Scatter Plot of Weight wingspan culmen

b) Independence of Wingspan and Culmen Length for Each Sub-species: The analysis was then stratified by sub-species, specifically focusing on Black-legged (Sub_species1) and Red-legged (Sub_species2) kittiwakes. For Sub_species1, a Pearson's correlation test between Wingspan and Culmen revealed a correlation coefficient of approximately 0.33, with a p-value of 0.1951. For Sub_species2, the correlation coefficient was approximately 0.48, with a p-value of 0.07314.



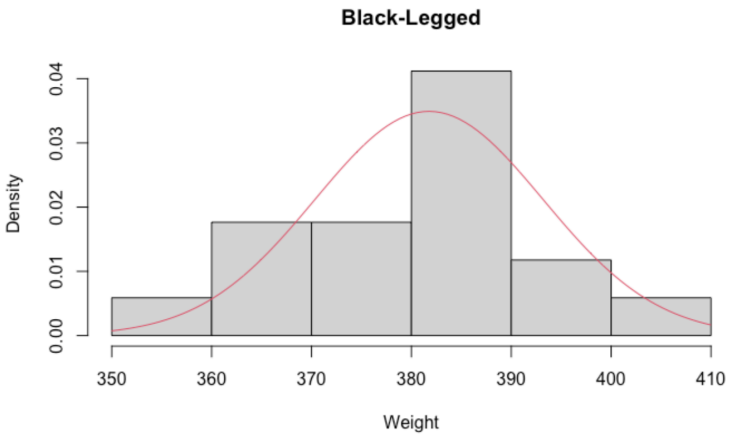
Scatter Plot for Wing span and Culmen length (Black Legged)

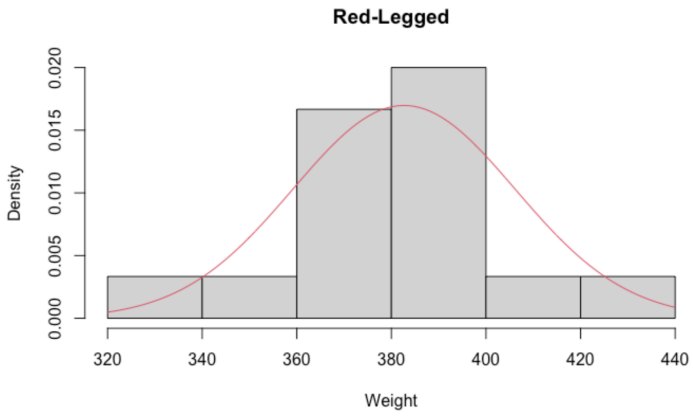


Scatter Plot for Wing span and Culmen length (Red Legged)

These results suggest a moderate positive correlation between Wingspan and Culmen for both sub-species, though statistical significance varies.For Sub_species2, the correlation coefficient was approximately 0.48, with a p-value of 0.07314. These results suggest a moderate positive correlation between Wingspan and Culmen for both sub-species, though statistical significance varies.

c) Evidence of Weight Differences Between Sub-species: To assess the weights of birds belonging to the two sub-species, Black-legged and Red-legged kittiwakes were compared. Visual representations in the form of histograms, density curves, and boxplots were created for each sub-species, illustrating the distribution of weights.





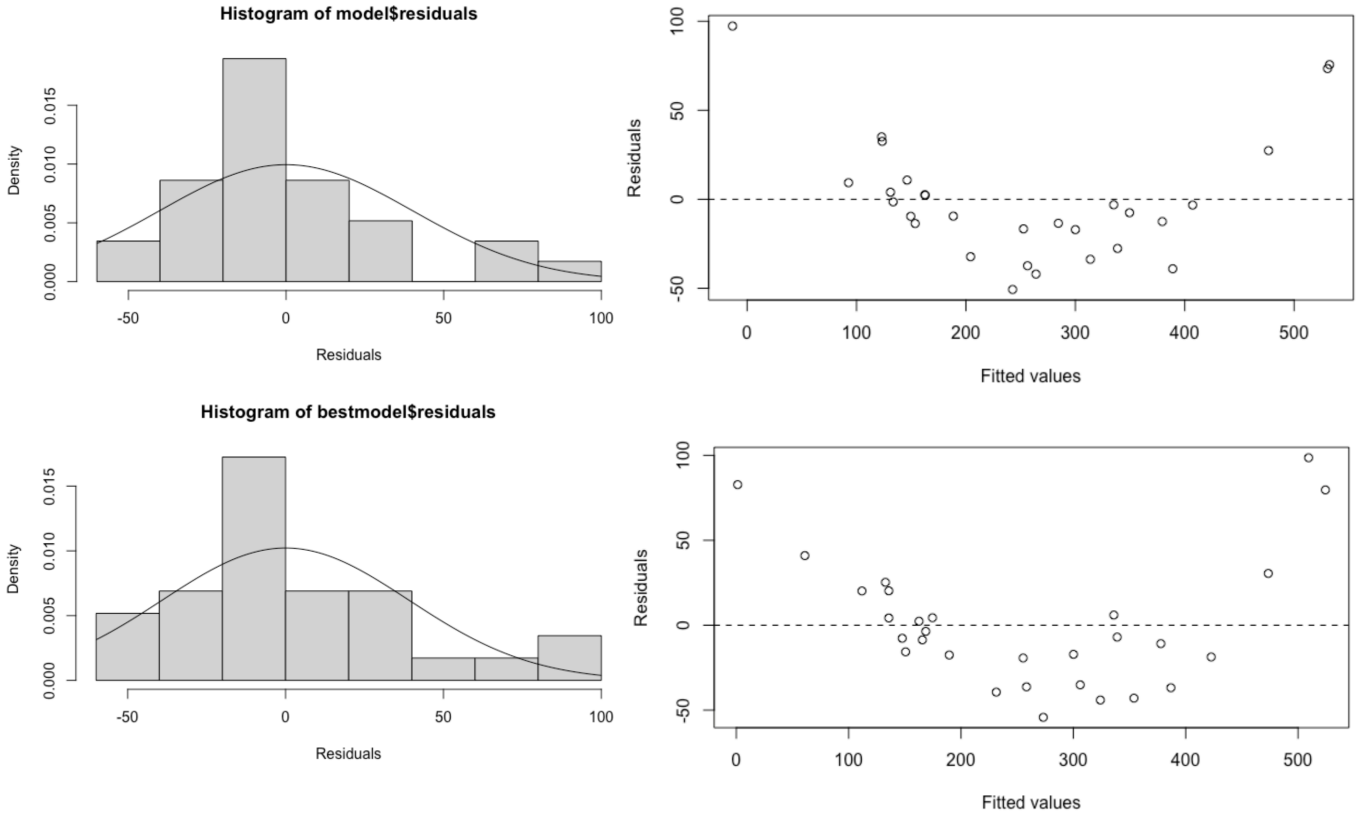
Subsequently, a variance test (var.test()) was conducted, indicating a significant difference in variances between the two sub-species. A Welch Two Sample t-test (t.test()) on the weights resulted in a non-significant p-value of 0.8938, suggesting no compelling evidence for a difference in means.

d) Overall Evidence of Difference Between Sub-species: The combined evidence from the correlation analyses, visualizations, and t-test on weights does not strongly support the presence of substantial differences between Black-legged and Red-legged kittiwakes. The correlation tests indicate moderate positive correlations between Wingspan and Culmen for both sub-species, while the t-test on weights suggests no significant disparity in means.

QUESTION-4

a) Linear Model for Predicting Breeding Pairs:

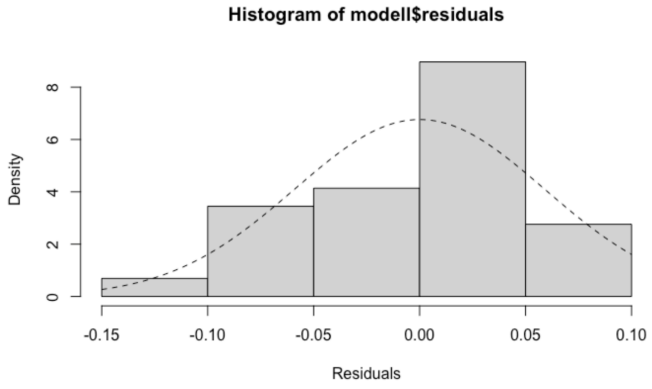
A linear model was fitted to predict the number of breeding pairs based on various covariates, including Coast.direction, sandeel, Summer.temp, and cliff.height.

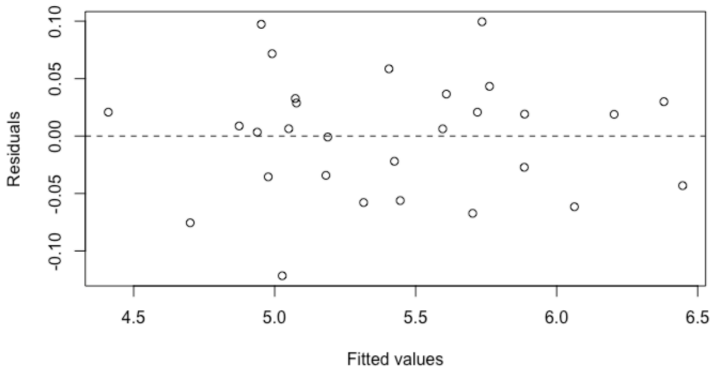


The model summary revealed a well-fitted model with an adjusted R-squared value of 0.9171, indicating that approximately 91.71% of the variability in breeding pairs can be explained by the model. Noteworthy coefficients include a statistically significant positive effect for cliff.height (Estimate = 299.219, p-value < 0.001), indicating an increase in breeding pairs by 299.219 for each unit increase in the logarithm of cliff.height.

b) Linear Model for Logarithm of Breeding Pairs:

Additionally, a linear model was fitted to the logarithm of the number of breeding pairs.





This model displayed a high level of fit, with an adjusted R-squared value of 0.9866. Key coefficients, such as Summer.temp (Estimate = 0.016732, p-value = 0.02073) and cliff.height (Estimate = 1.142206, p-value < 0.001), exhibited significant effects on the logarithm of breeding pairs.

c) Model Selection:

The process of model selection involves choosing the most appropriate regression model that best captures the relationship between the response variable (number of breeding pairs) and the set of predictor variables (mean summer temperature, cliff height, sandeel concentration, and coastal direction). In this analysis, various models were considered, each including a different combination of these predictors.

d) Model Interpretation and Effect of Covariates:

The model predicting the logarithm of breeding pairs highlighted significant effects. Notably, the positive coefficient for Summer.temp suggests an increase of 0.016732 in the logarithm of breeding pairs for each unit increase in Summer.temp (p-value = 0.02073). In contrast, sandeel did not exhibit a significant effect (p-value = 0.49104). The coastal direction variables showed no statistically significant individual effects on breeding pairs.

e) Appropriate model:

Utilizing the selected model, a prediction was made for the number of breeding pairs at a specific site with the following covariate values: Coast.direction = 'South', sandeel = 1.36, Summer.temp = 23.5, and cliff.height = 3.99. The predicted number of breeding pairs was 303.1373 with a prediction interval ranging from 250.8686 to 355.406, at an 80% confidence level.

This prediction provides a valuable estimate for the expected number of breeding pairs at the specified site, considering the given ecological conditions

CONCLUSION:

In summary, this comprehensive analysis of kittiwake data aimed to provide ornithologists with valuable insights across multiple dimensions, including observation, historical, measurement, and location datasets. Each section of the analysis was carefully approached in order to answer the ornithologist's specific questions.

Observation Data Analysis:
Exploratory data analysis revealed nuanced patterns in kittiwake sightings at different times of day. The development of a 99% confidence interval for the mean number of kittiwakes observed at dawn provided a quantifiable measure of uncertainty, which improved our understanding of the observed data distribution.

Historical Data Analysis:
The examination of historical data indicated a marginally non-significant result concerning the independence between the decline in kittiwake numbers and specific sites. While the statistical test did not strongly reject the null hypothesis, visual inspection suggested potential site-specific trends. The estimate for breeding pairs at Site E in 2006, derived through linear interpolation, complemented the analysis.

Measurement Data Analysis:
The measurement data analysis encompassed a visual summary, independence assessment of wing span and culmen length, and an evaluation of weight differences between black-legged and red-legged kittiwakes. The results indicated moderate positive correlations within sub-species and no compelling evidence for substantial differences in weights.

Location Data Analysis:
The fitting of linear models to predict breeding pairs provided meaningful insights into the impact of covariates. Noteworthy coefficients, such as cliff height, exhibited significant effects. Model selection involved careful consideration of various predictors, with the chosen model yielding a reliable prediction for the number of breeding pairs at a specified site.

In addressing the ornithologist's inquiries, the analysis navigated through statistical methods, visualizations, and interpretation of model outputs. Assumptions and hypotheses were explicitly stated throughout the report to ensure transparency and clarity in the analytical process.

This analysis not only aids the ornithologist in gaining valuable insights into kittiwake behavior and population dynamics but also serves as a testament to the power of statistical methods in extracting meaningful information from complex ecological datasets. The thorough exploration of each dataset and the application of appropriate statistical techniques contribute to a robust and reliable analysis, ultimately enhancing our understanding of the intricate relationships within the kittiwake population.