Project 3

12-04-2024

Group 5: Neha Karna, Prafulla Shrestha, Aidan Stewart, Josh Lefdal, Shivam Bhardwaj

Predictive Modeling and Vole Species Classification

- Predictive Modeling
  - Predictive modeling is a statistical technique that combines explanatory variables to predict a response variable (IBM, 2024a).
  - Predictive analytics uses historical data to uncover patterns and trends for future predictions (IBM, 2024a).
- Background on Vole Species:
  - Two vole species, Microtus subterraneus and Microtus multiplex, are distinct based on differences in chromosomal counts (Airoldi et al., 1995).
  - Both species have two chromosomal types, but hybrids have not been identified (Airoldi et al., 1995).
- Objective:
  - Use morphometric data (skull length, height, width) from 288 vole specimens to classify species.
  - Chromosomal data:
    - 89 specimens: Definitively classified as subterraneus or multiplex (training set).
    - 199 specimens: Chromosomal data unavailable (test set).
- Goal:
  - Develop 7 predictive models using the classified 89 specimens and apply the model to classify the remaining 199 unclassified specimens into one of the two species.

Generalized Linear Model (GLM): Logistic Regression

- Logistic Regression
  - A type of classification model derived from Generalized Linear Models (GLMs).
  - Used to predict binary outcomes (e.g., subterraneus vs. multiplex).
  - In this study, logistic regression predicts vole species based on skull measurements (length, height, width).
- Leave-One-Out Cross Validation (LOOCV)
  - Each data point is used once as test data and the remaining n-1 observation as training data
  - The model is trained and tested n times, each time excluding a different observation
  - The MSE (test error) is calculated for each iteration and LOOCV test MSE is the average of n test error

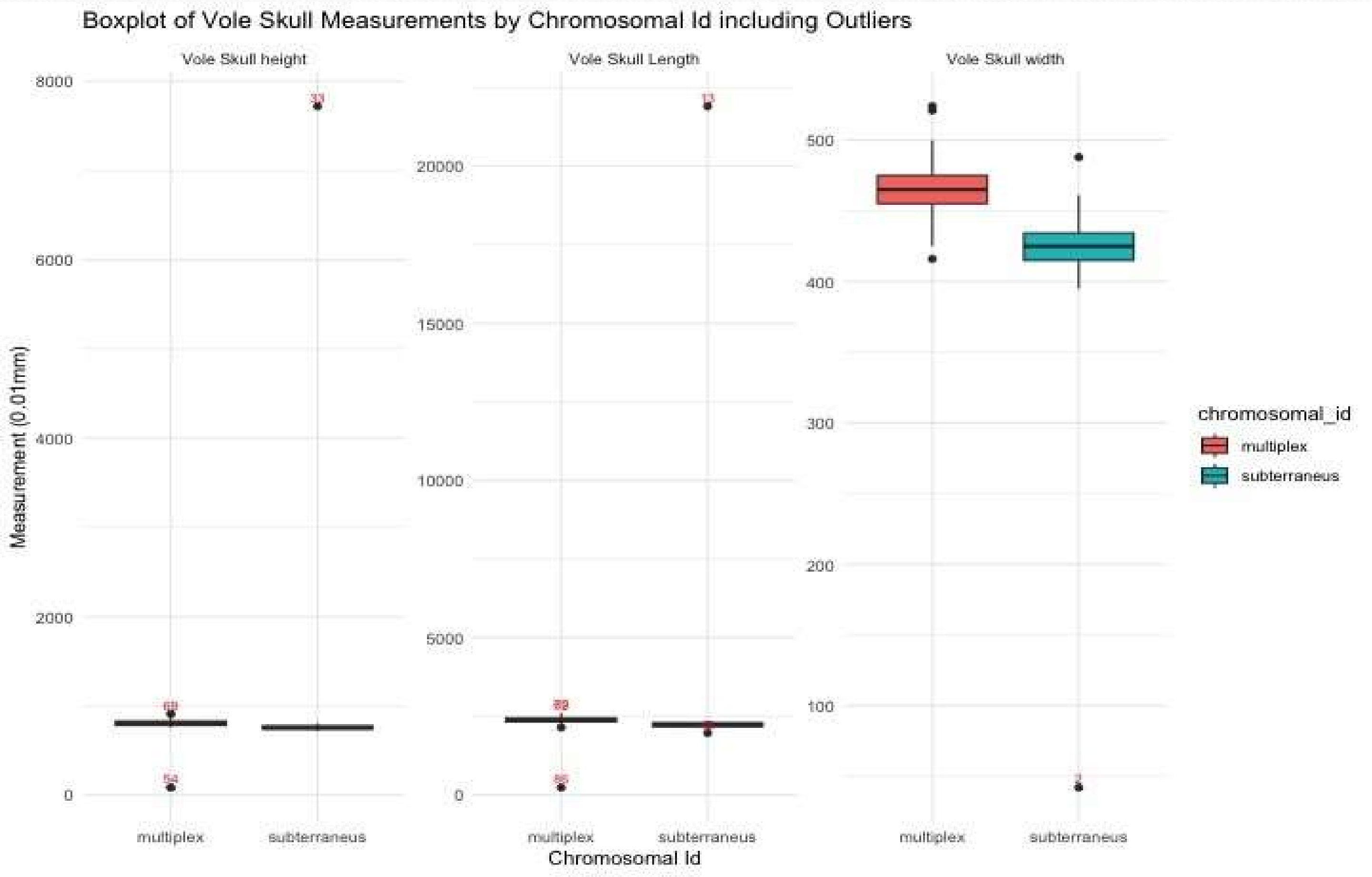
Data Manipulation and Cleaning

- Dataset measurements include **skull length, height, and width** and consists of 288 vole specimens:
  - **89 known specimens:** Classified as subterraneus or multiplex.
  - **199 unknown specimens:** Classification unavailable.
- Data Preparation:
  - Updated the column names of the **known** and **unknown** datasets as "index", "chromosomal_id", "skull_length", "skull_height", "skull_width".
  - Combined subterraneus and multiplex data into a single dataset as **known** dataset.
  - Removed rows with missing values to ensure a complete dataset for model training.
  - Response variable was created where 1 represents subterraneus and 0 represents multiplex
  - For **known** dataset, rows with Outliers, beyond 1.5 × IQR, were removed
  - For **unknown** dataset, replaced extreme outlier with adjusted mean values

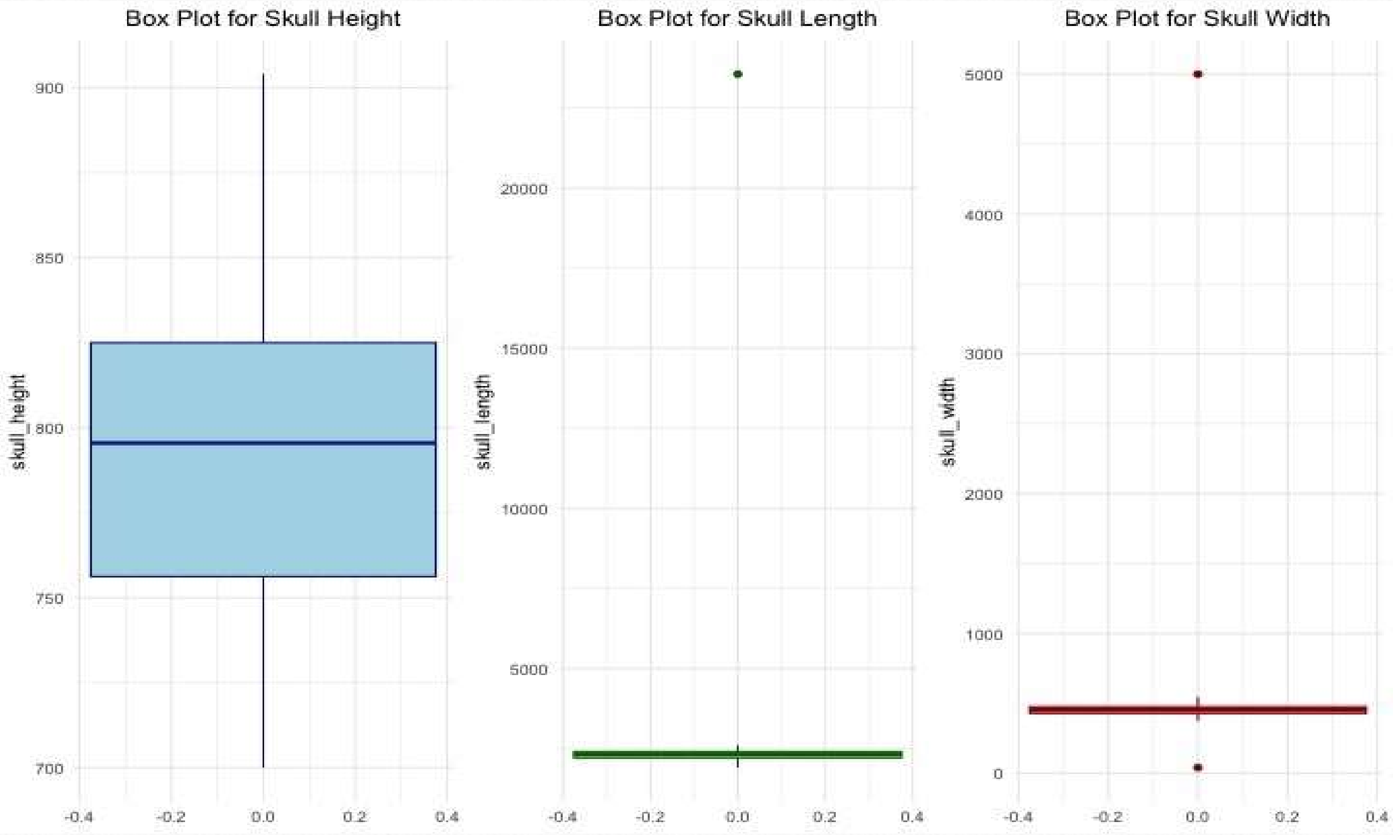
Exploratory Data Analysis (EDA)

- A process used to: Investigate and summarize datasets. Identify errors, patterns, outliers, and relationships. Inform analysis before making assumptions (IBM, 2024b).
- Observation counts of known dataset - M. multiplex: 44 and M. subterraneus: 46 (note: 2 rows had missing values, reducing usable observations to 90).
- Column renaming - Simplified variable names:
  - skull_height
  - skull_length
  - skull_width
- Data Visualization:
  - Created boxplot and histogram
  - Descriptive Summary
  - Pair plot

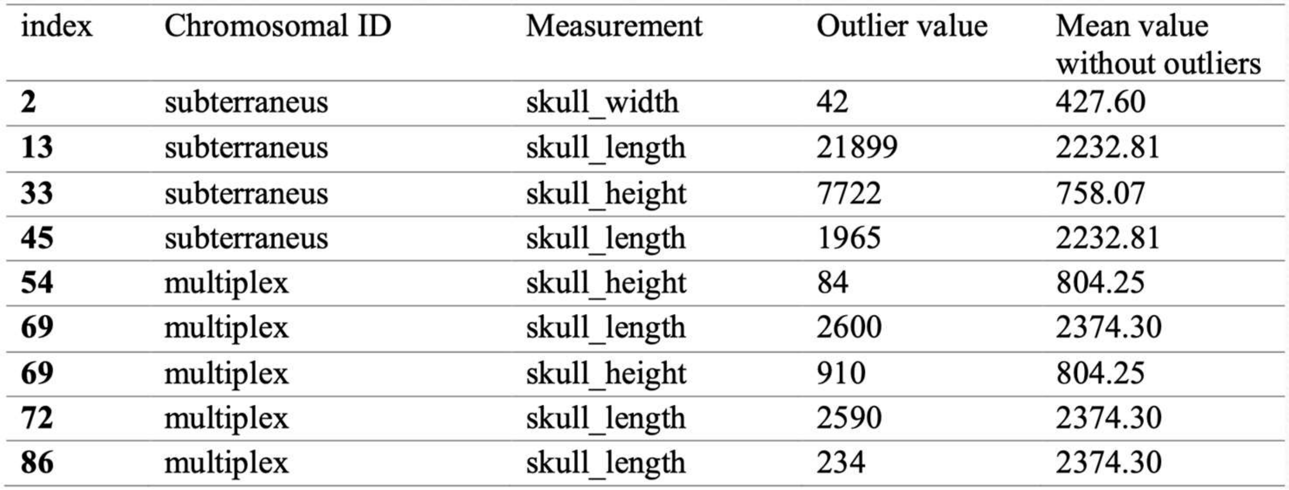
Boxplot for known dataset



Boxplot for unknown dataset

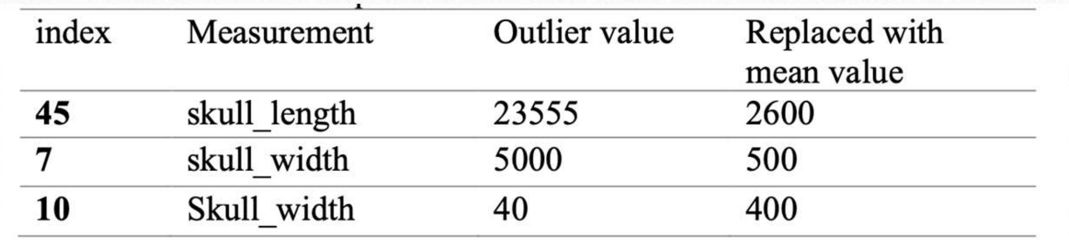


List of outliers removed from known dataset



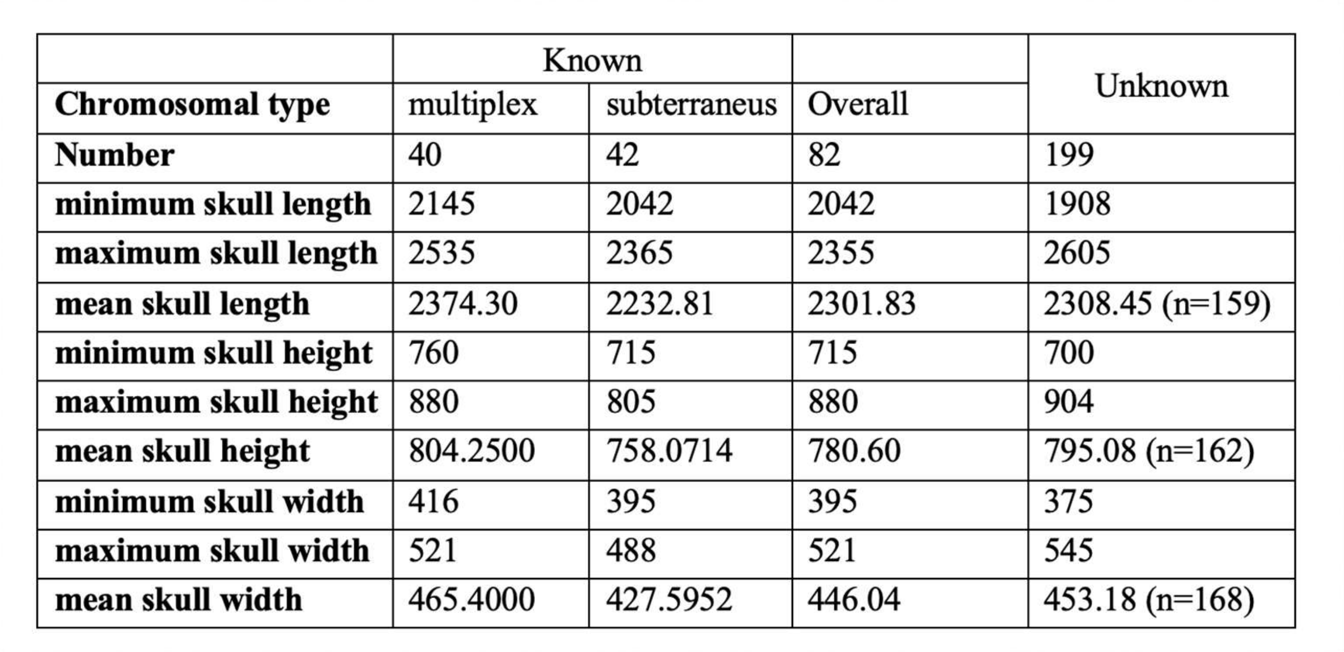
index	Chromosomal ID	Measurement	Outlier value	Mean value without outliers
2	subterraneus	skull_width	42	427.60
13	subterraneus	skull_length	21899	2232.81
33	subterraneus	skull_height	7722	758.07
45	subterraneus	skull_length	1965	2232.81
54	multiplex	skull_height	84	804.25
69	multiplex	skull_length	2600	2374.30
69	multiplex	skull_height	910	804.25
72	multiplex	skull_length	2590	2374.30
86	multiplex	skull_length	234	2374.30

List of extreme outliers (typos) replaced in unknown dataset



index	Measurement	Outlier value	Replaced with mean value
45	skull_length	23555	2600
7	skull_width	5000	500
10	Skull_width	40	400

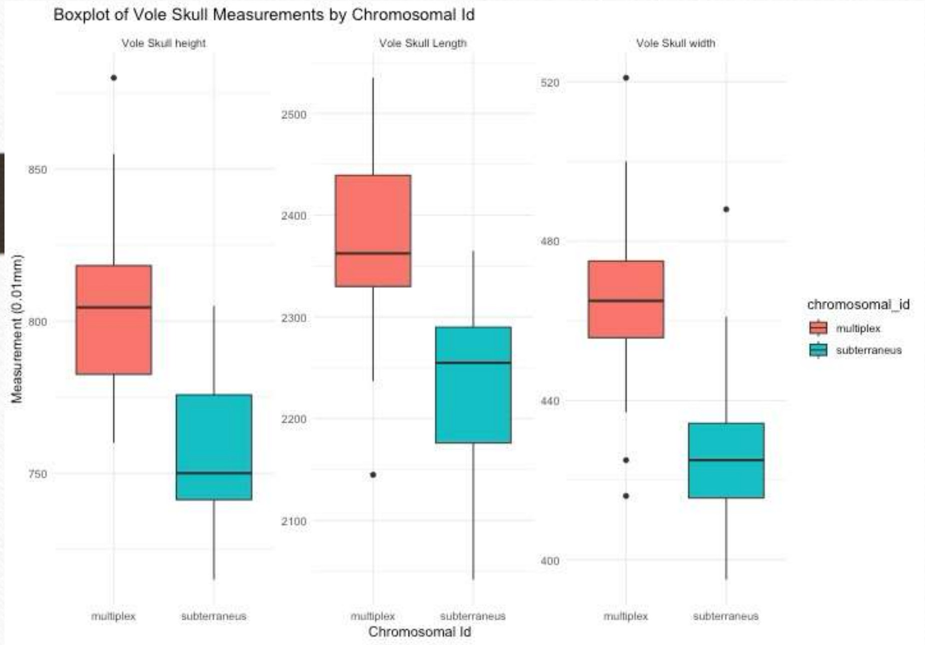
Descriptive Summary after removing outliers

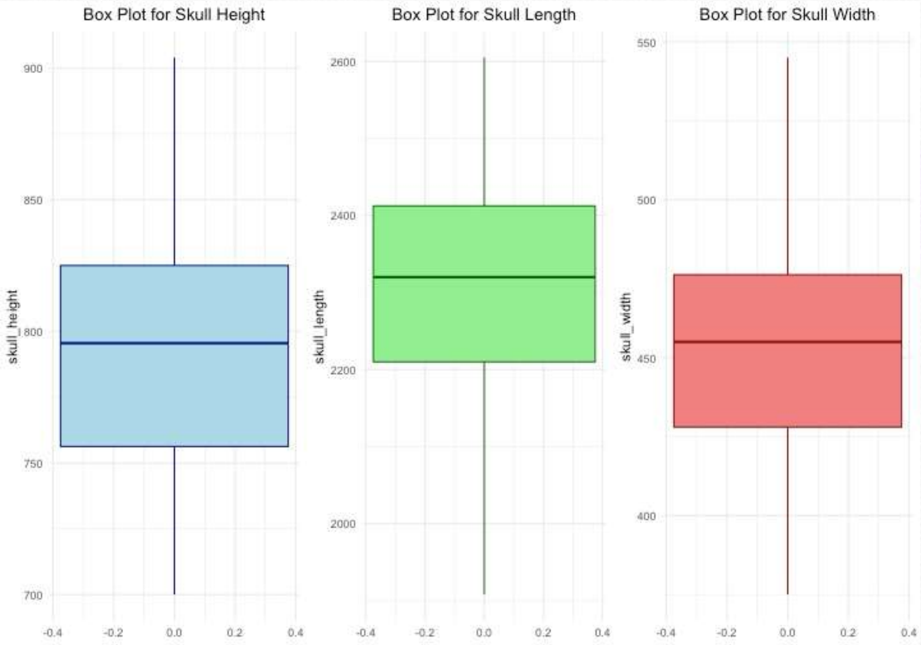


	Known			Unknown
Chromosomal type	multiplex	subterraneus	Overall	
Number	40	42	82	199
minimum skull length	2145	2042	2042	1908
maximum skull length	2535	2365	2355	2605
mean skull length	2374.30	2232.81	2301.83	2308.45 (n=159)
minimum skull height	760	715	715	700
maximum skull height	880	805	880	904
mean skull height	804.2500	758.0714	780.60	795.08 (n=162)
minimum skull width	416	395	395	375
maximum skull width	521	488	521	545
mean skull width	465.4000	427.5952	446.04	453.18 (n=168)

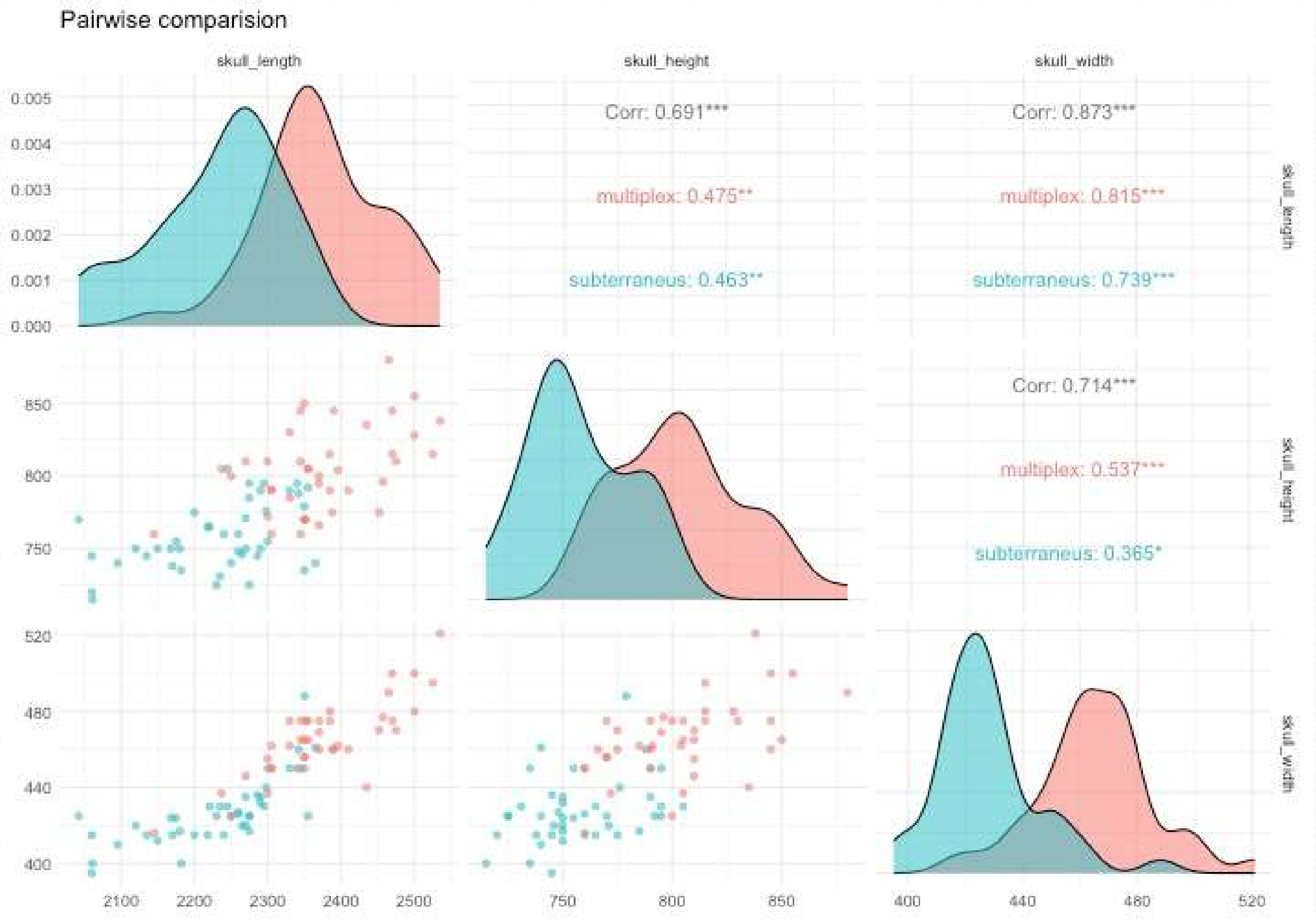
Boxplot for known dataset

Boxplot for unknown dataset

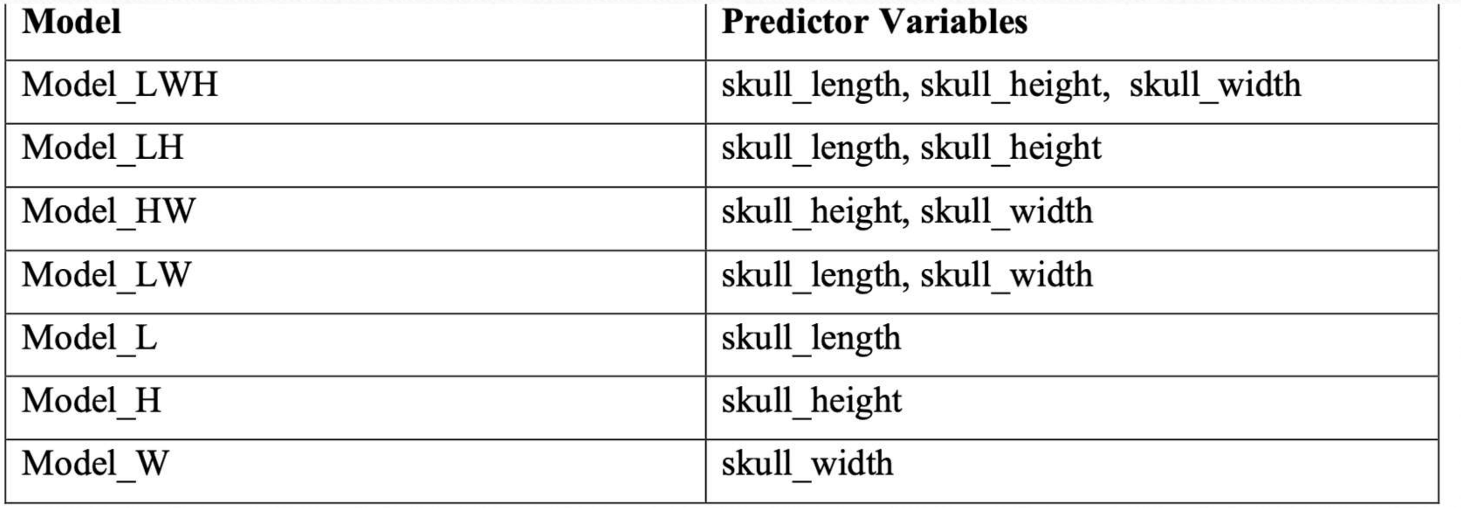




Pair Plot for known dataset



Modeling



Model	Predictor Variables
Model_LWH	skull_length, skull_height, skull_width
Model_LH	skull_length, skull_height
Model_HW	skull_height, skull_width
Model_LW	skull_length, skull_width
Model_L	skull_length
Model_H	skull_height
Model_W	skull_width

Model Assessment

- **Akaike Information Criterion (AIC):** The AIC is the log likelihood adjusted for the number of coefficients and is used to decide input variables for the best fitting model (Zumel and Mount, 2020). The log-likelihood increases with the number of variables. The model with lowest AIC score is best fit.
- **LOOCV MSE:** Mean Squared Error (MSE) is the average of the squared differences between the predicted and actual value. In LOOCV, the MSE is averaged over n iteration, where each data point is used once as the test data. When predicted responses are very close to the true response, the MSE would be small. The best-fit model would have the smallest LOOCV MSE (James et al., 2021).

Results: Model summary

- Model LWH and Model LW are not significant (p-value more than 0.05 for the length variable)

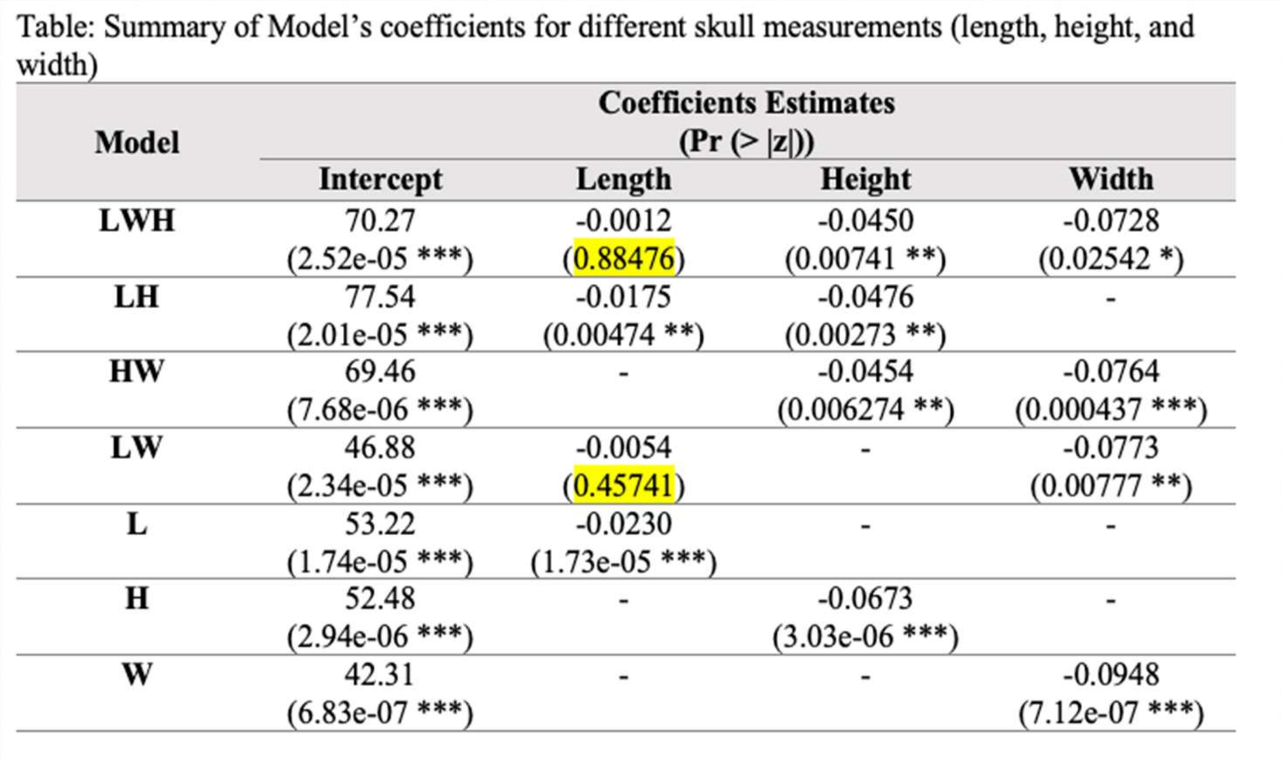


Table: Summary of Model's coefficients for different skull measurements (length, height, and width)

Model	Coefficients Estimates (Pr (> |z|))			
	Intercept	Length	Height	Width
LWH	70.27 (2.52e-05 ***)	-0.0012 (0.88476)	-0.0450 (0.00741 **)	-0.0728 (0.02542 *)
LH	77.54 (2.01e-05 ***)	-0.0175 (0.00474 **)	-0.0476 (0.00273 **)	-
HW	69.46 (7.68e-06 ***)	-	-0.0454 (0.006274 **)	-0.0764 (0.000437 ***)
LW	46.88 (2.34e-05 ***)	-0.0054 (0.45741)	-	-0.0773 (0.00777 **)
L	53.22 (1.74e-05 ***)	-0.0230 (1.73e-05 ***)	-	-
H	52.48 (2.94e-06 ***)	-	-0.0673 (3.03e-06 ***)	-
W	42.31 (6.83e-07 ***)	-	-	-0.0948 (7.12e-07 ***)

Significance (p-values)

***0.001, **0.01 & *0.05

Results: Model summary

Since lower AIC values represent better models, **Model HW** provides the best performance (also evident by the least MSE value of 0.10199 and highest model accuracy of 87.8%); hence is the preferred model.

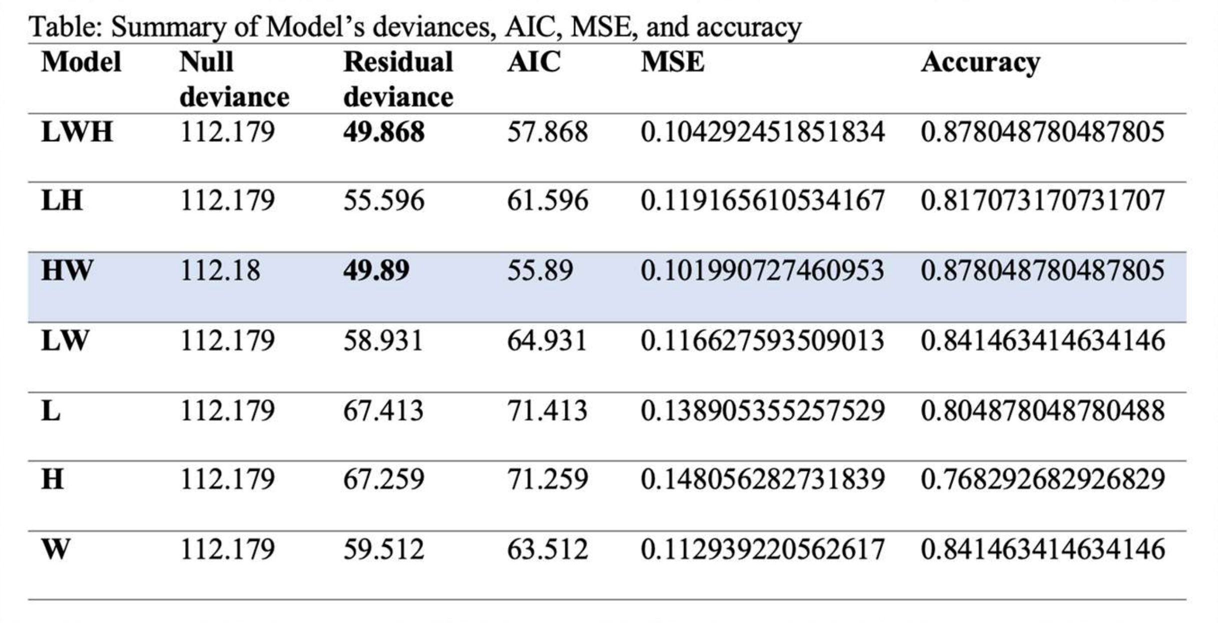


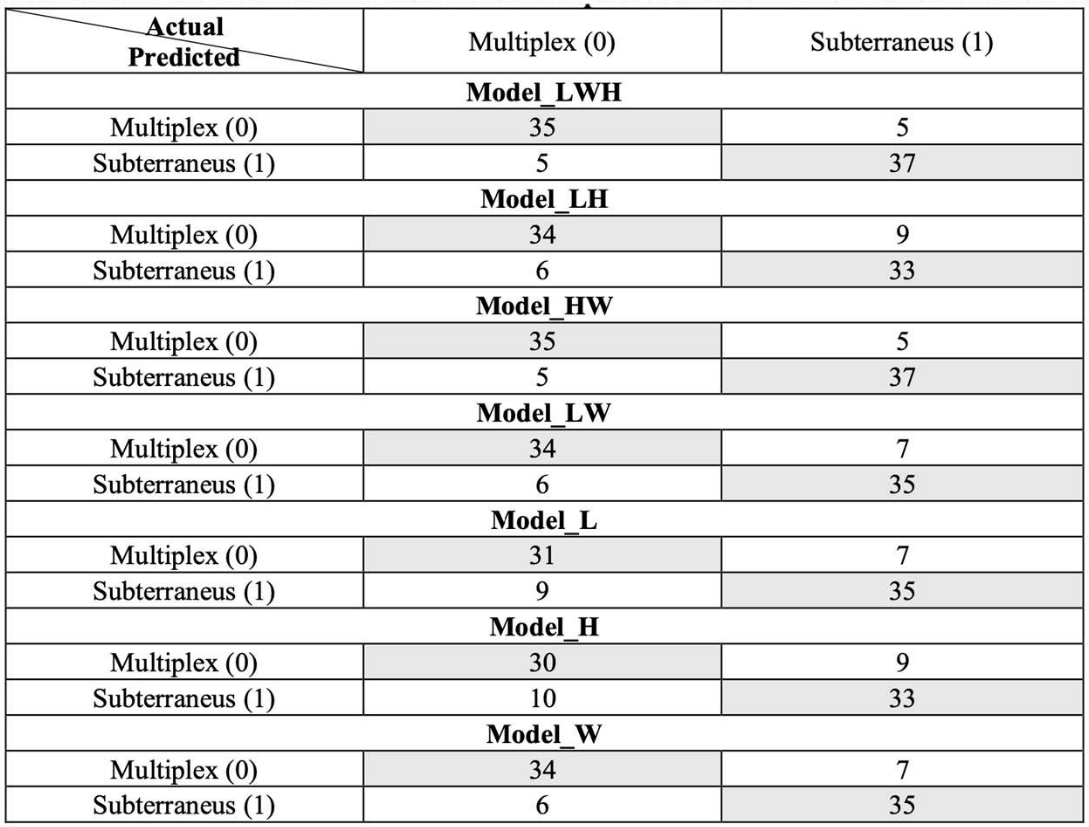
Table: Summary of Model’s deviances, AIC, MSE, and accuracy

Model	**Null deviance**	**Residual deviance**	**AIC**	**MSE**	**Accuracy**
LWH	112.179	**49.868**	57.868	0.104292451851834	0.878048780487805
LH	112.179	55.596	61.596	0.119165610534167	0.817073170731707
HW	112.18	**49.89**	55.89	0.101990727460953	0.878048780487805
LW	112.179	58.931	64.931	0.116627593509013	0.841463414634146
L	112.179	67.413	71.413	0.138905355257529	0.804878048780488
H	112.179	67.259	71.259	0.148056282731839	0.768292682926829
W	112.179	59.512	63.512	0.112939220562617	0.841463414634146

Result

Confusion Matrix (Known)

The diagonals (shaded) represent correct predictions, and off-diagonals (white) represent incorrect predictions. Model_LWH and Model_HW show relatively higher numbers of correct predictions.



Actual Predicted	Multiplex (0)	Subterraneus (1)
	Model_LWH	
Multiplex (0)	35	5
Subterraneus (1)	5	37
	Model_LH	
Multiplex (0)	34	9
Subterraneus (1)	6	33
	Model_HW	
Multiplex (0)	35	5
Subterraneus (1)	5	37
	Model_LW	
Multiplex (0)	34	7
Subterraneus (1)	6	35
	Model_L	
Multiplex (0)	31	7
Subterraneus (1)	9	35
	Model_H	
Multiplex (0)	30	9
Subterraneus (1)	10	33
	Model_W	
Multiplex (0)	34	7
Subterraneus (1)	6	35

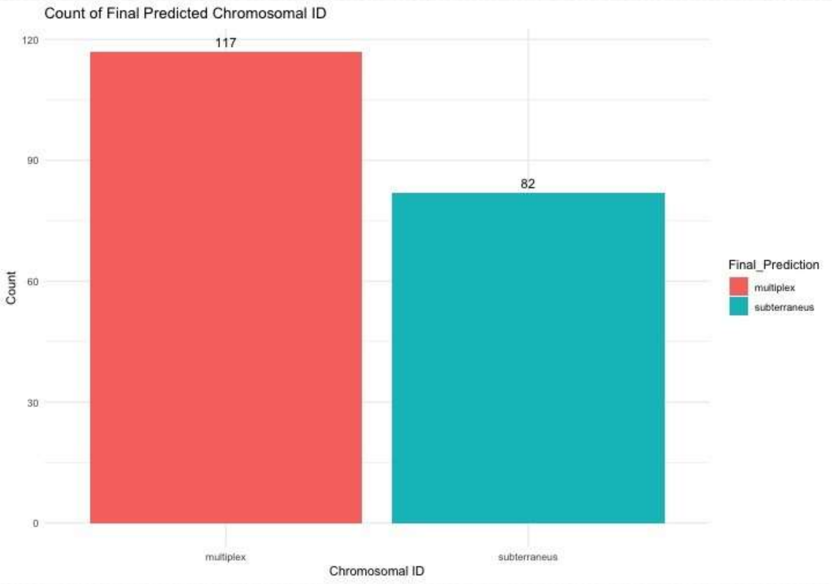
Predictions

- Prediction was made for the unknown dataset using seven subsets.
- Depending on the model being used, a subset of the data is created by removing any rows with missing values. This means that if any of the predictor variables (such as skull measurements) are missing for an observation, that entire observation is discarded.
- For example, in the LWH model, which uses three predictor variables, only the rows where all three variables have valid values will be included. If even one value is missing for any of the predictor variables, the entire row is excluded from the dataset.

	Model_LWH	Model_LH	Model_HW	Model_LW	Model_L	Model_H	Model_W
subterraneus	45	59	55	62	74	69	72
multiplex	72	76	80	75	85	93	96
Total predicted	117	135	135	137	159	162	168

Final Predictions

The final prediction was made based on the frequency of predicted chromosomal IDs by all seven models. Each unknown specimen was assigned a chromosomal type on the basis of how frequently they were predicted to be a certain class on all of our seven models. Using an odd number of models (seven) allowed us to identify the highest frequency, a process that might not have been feasible with an even number of models.



Comment: Out of 199 total unknown specimens, 117 were predicted to be multiplex and remaining 82 were predicted to be subterraneus.

Recommendation

- Based on our analysis, we recommend the use of the **Model HW** (Height and Width) for classification of vole species. This model demonstrated the **lowest AIC and MSE values**, indicating better performance compared to other models.
- While the overall accuracy of all models was above 75%, Model HW stands out as the most efficient for predictive classification given the dataset and its variables.
- If skull height and width are readily available, Model HW is the most efficient choice for prediction.

References

- Airoldi, J., Flury, B., & Salvioni, M. (1995). Discrimination between two species of Microtus using both classified and unclassified observations. Journal of Theoretical Biology, 177(3), 247-262.
- Bobbitt, Z. (2022, February 23). *How to Interpret glm Output in R (With Example)*. Statology. https://www.statology.org/interpret-glm-output-in-r/
- Caughlin, D. E. (n.d.). Chapter 49 applying K-fold cross-validation to logistic regression. R for HR: An Introduction to Human Resource Analytics Using R. https://rforhr.com/kfold.html
- Confusion matrix in machine learning. (2024, July 8). GeeksforGeeks. https://www.geeksforgeeks.org/confusion-matrix-machine-learning/
- Everitt, B. S., & Hothorn, I. (2010). A handbook of statistical analyses using R (SECOND) [Book]. CRC Press. https://www.ehu.eus/ccwintco/uploads/9/93/A_Handbook_of_Statistical_Analyses_Using_R_Second_Edition.pdf
- IBM. (2024, November 1). Predictive Analytics. *IBM*. https://www.ibm.com/topics/predictive-analytics?utm_content=SRCWW&p1=Search&p4=43700075153304567&p5=p&p9=58700008227853819&gclid=EAIaIQobChMIzefRt-LUiQMVFk7_AR01LDiEEAAYASAAEgLVRPD_BwE&gclsrc=aw.ds
- IBM. (2024b, November 7). Exploratory Data Analysis. *IBM*. https://www.ibm.com/topics/exploratory-data-analysis
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An Introduction to Statistical Learning: with Applications in R. https://link.springer.com/content/pdf/10.1007/978-1-0716-1418-1.pdf
- Lecture Notes and Resources (STAT 541, STAT 600, STAT 601)
- Zumel, N., & Mount, J. (2019). Practical Data Science with R, Second Edition. Manning.