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

Project 3: Redo of Project 2-Predictive Modeling for Microtus species

INTRODUCTION

Predictive Modeling: Predictive modeling is a statistical technique used to create a function that describes how various explanatory variables combine to predict a response variable. Predictive analytics, a subset of this field, leverages historical data to uncover patterns and trends that inform future predictions (IBM, 2024).

In this study, our objective is to differentiate between *Microtus subterraneus* and *Microtus multiplex*, two vole species that have been classified as distinct primarily based on differences in their chromosomal counts (Airoldi et al., 1995). Both species exhibit two chromosomal types, but their hybrids have not yet been identified (Airoldi et al., 1995). Additionally, *M. subterraneus* individuals have been found to be smaller than *M. multiplex* in most morphometric measurements.

To achieve this, we analyze morphometric data collected by Salvioni, which includes measurements of skull length, height, and width for 288 vole specimens (Airoldi et al., 1995). Chromosomal analyses have been conducted for 89 specimens, providing definitive identification as either *subterraneus* or *multiplex*. Our goal is to use these 89 specimens as a training set to develop a predictive model. This model will then be applied to classify the remaining 199 specimens, for which chromosomal information is unavailable, into one of the two vole species.

Generalized Linear Model: Logistic regression: Classifying an observation into categories, such as assigning a *vole* specimen to be either *subterraneus* or *multiplex,* involves predicting the probability that an observation belongs to a particular class. Logistic regression, a type of classification model branched from Generalized linear model (GLM) for supervised machine learning, is widely used to predict binary outcomes. Here, we are interested in predicting species type on the basis of the measurements of vole skull length, height and width.
When the response variable is binary, the distribution of y reduces to a single value, the probability $p = P(Y = 1)$, which depends on the linear combination of explanatory variables as follows,

$$\text{p} = \Pr\,(Y_i = 1|x_1 \dots\ x_n) = \pi(x_1 \dots\ x_n)$$

Since the value of p is bounded between 0 and 1, while the linear predictors can vary between $-\infty\ and\ +\infty$, we use logit transformation to get a linear version of our expected value. The probability is linked to the predictor variables using logit function as follows:

$$logit(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_q x_{nq}$$

The logit of a probability is simply the log of the odds of the response taking the value one, so the above equation can be rewritten as,

$$\pi(x_1 \dots\ x_n) = \frac{\exp\,(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_q x_{nq})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_q x_{nq})}$$

where,
i = 1,…,n number of sample
Y_i: ith response (1 represents *subterraneus* and 0 represents *multiplex*)

x_1 ... x_q: q explanatory variables (*skull_length, skull_height or skull_width*)
π: probability (response variable)
β_0: y-intercept,
$\beta_1, \ldots, \beta_q$: regression coefficients for each q response variable

MODEL ASSESSMENT

Leave-One-Out Cross Validation (LOOCV): The cross-validation is used to estimate the test error of a model for model assessment and model selection. **Test error** refers to the difference between the predicted values and the actual values for a model when applied to a new observation (i.e., test set). In LOOCV, each data point is used once as the test set and remaining n-1 observation as the training set. The model is trained and tested n times, each time excluding a different observation. The Mean Squared Error (MSE) is computed for each procedure and the final LOOCV estimate of test MSE is the average of these n test error (James et al., 2021):

$$CV_n = \frac{1}{n}\sum_{i=1}^{n} MSE_i$$

Metrics for model assessment

- **Null Deviance and Residual Deviance:** Deviance is the measure of how well the model fits the data. The null deviance represents the model's fit with intercepts only i.e. without any predictors while residual deviance represents the model's fit after adding the predictors. If the deviance significantly reduces from null deviance to residual deviance, it implies adding predictors help improve the model. A lower residual deviance indicates the better model fit.
- **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).
- **Model Accuracy**: In contrast to test error, Accuracy measures the proportion of correct prediction, where the predicted class matches the actual class. The model with higher accuracy rate is better (further discussed in confusion matrix).
- **Confusion Matrix**
  - The diagonal element of confusion matrix indicates correct predictions whereas off-diagonals represents incorrect predictions.
  - Mean of correct prediction (correct prediction/ total prediction) gives us the model accuracy rate.

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

DATA MANIPULATION AND DATA CLEANING

The dataset consists of skull measurements from 288 vole specimens, of which 89 have been identified as either *subterraneus* or *multiplex* while 199 of the specimens are classified *unknown*. The measurements include skull length, height, and width.

Dataset: From the "*Vole Skull.xlsm*" spreadsheet, which included three sheets "*subterraneus*", "*multiplex*" and "*unknown*", we created two separate data set for our predictive modeling.

1. known (includes data for identified specimen, both s*ubterraneus* and *multiplex*)
2. unknown (includes data for unidentified specimen)

Additionally, we standardized the column names of the known and unknown datasets as "index", "chromosomal_id", "skull_length", "skull_height", "skull_width"

Data Manipulation and Cleaning for Known Dataset

Missing values: We removed any rows with missing values (NAs) as we require a complete dataset to train our model.

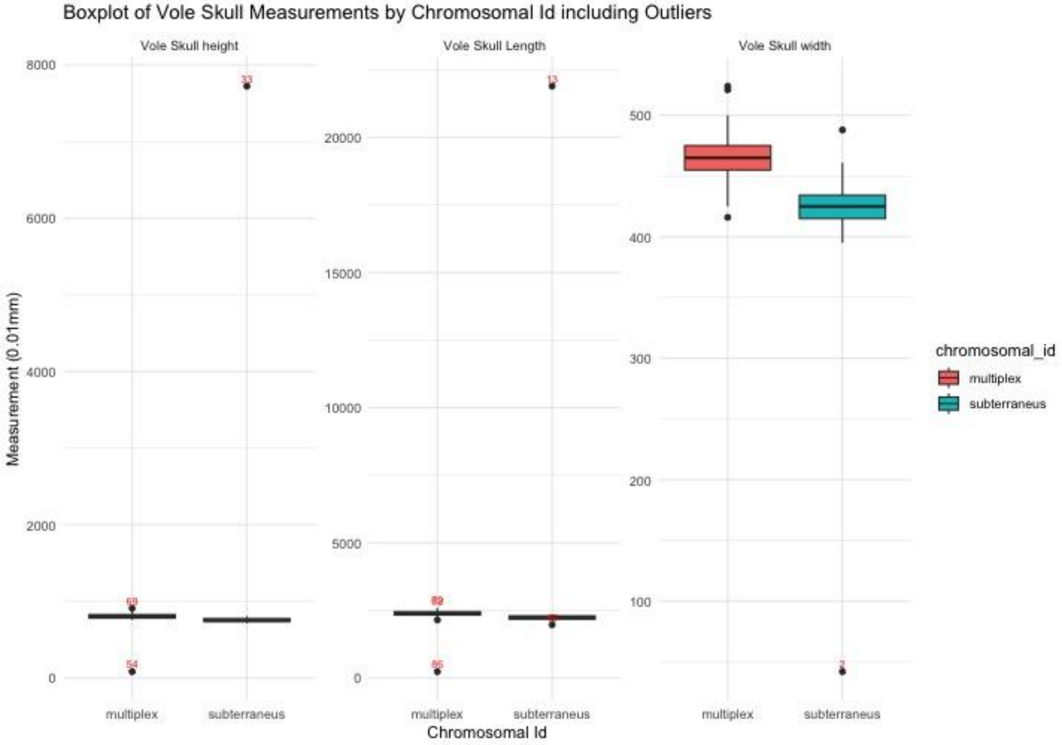
Reset of Index: We reset the index numbers of the known data to be one running list.

Response variable: Add response variable where 1 represents subterraneus and 0 represents multiplex.

Outliers: An outlier is an observation that falls far from the typical range of other observations in a dataset. These outliers can occur due to error in data collection and/or human mistakes (typos). Influential outliers can cause the model to produce inaccurate estimates. The box plot below shows the presence of extreme outliers,

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

Boxplot for known dataset



While there exists various treatment for outliers, we decided to remove the outliers from the known dataset. Outliers that were beyond 1.5XIQR, were removed from known dataset. IQR is the difference between 1st and 3rd quartile (Q1 and Q2). The method set threshold based on typical spread,

- lower bound = Q1 - 1.5 * IQR
- upper bound =Q3 + 1.5 * IQR where IQR is the difference between 1st and 3rd quartile (Q1 and Q2)

Table: 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

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

Data manipulation and cleaning for Unknown Dataset

Missing values: For an unknown dataset, we only removed rows with missing values in every column while keeping other missing values. Seven subsets were created from an unknown dataset according to the combination of known variables.

Outliers: The boxplot below shows the presence of outliers. Outliers in unknown dataset were identified as typo and were manually replaced with mean value excluding the outlier.

Replace outliers with the specified values based on the specified bounds (capping outliers). This was identified based on manual inspection, since the data set was small. We didn't see any typos for skull_height

- For skull length: replace values below 1907 with 1900 and values above 2606 with 2600
- For skull width: replace values below 374 with 400 and values above 546 with 500

Boxplot for unknown dataset

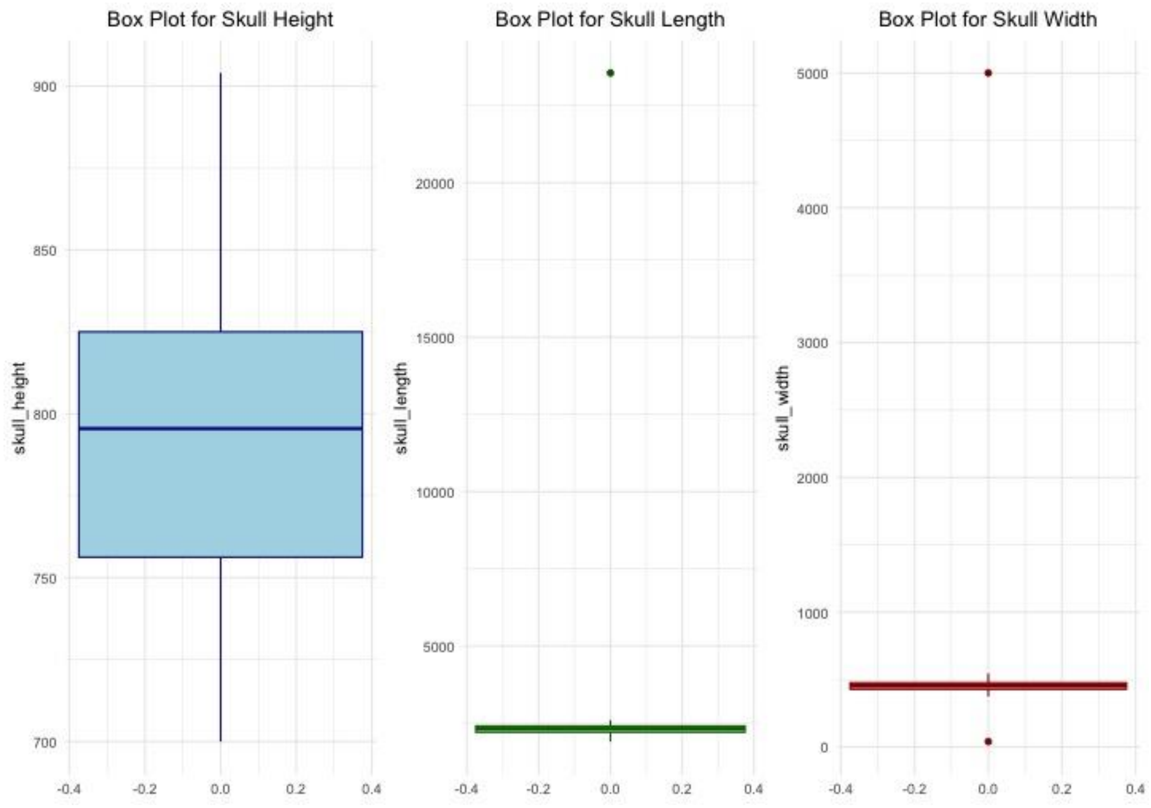


Table: List of outliers replaced in unknown dataset

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

EXPLORATORY DATA ANALYSIS (EDA)

Exploratory data analysis (EDA) is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods. The main purpose of EDA is to help look at data before making any assumptions. It can help identify obvious errors, as well as better understand patterns within the data, detect outliers or anomalous events, find interesting relations among the variables.

EDA for known Dataset

The EDA below is after removing outliers as mentioned above.

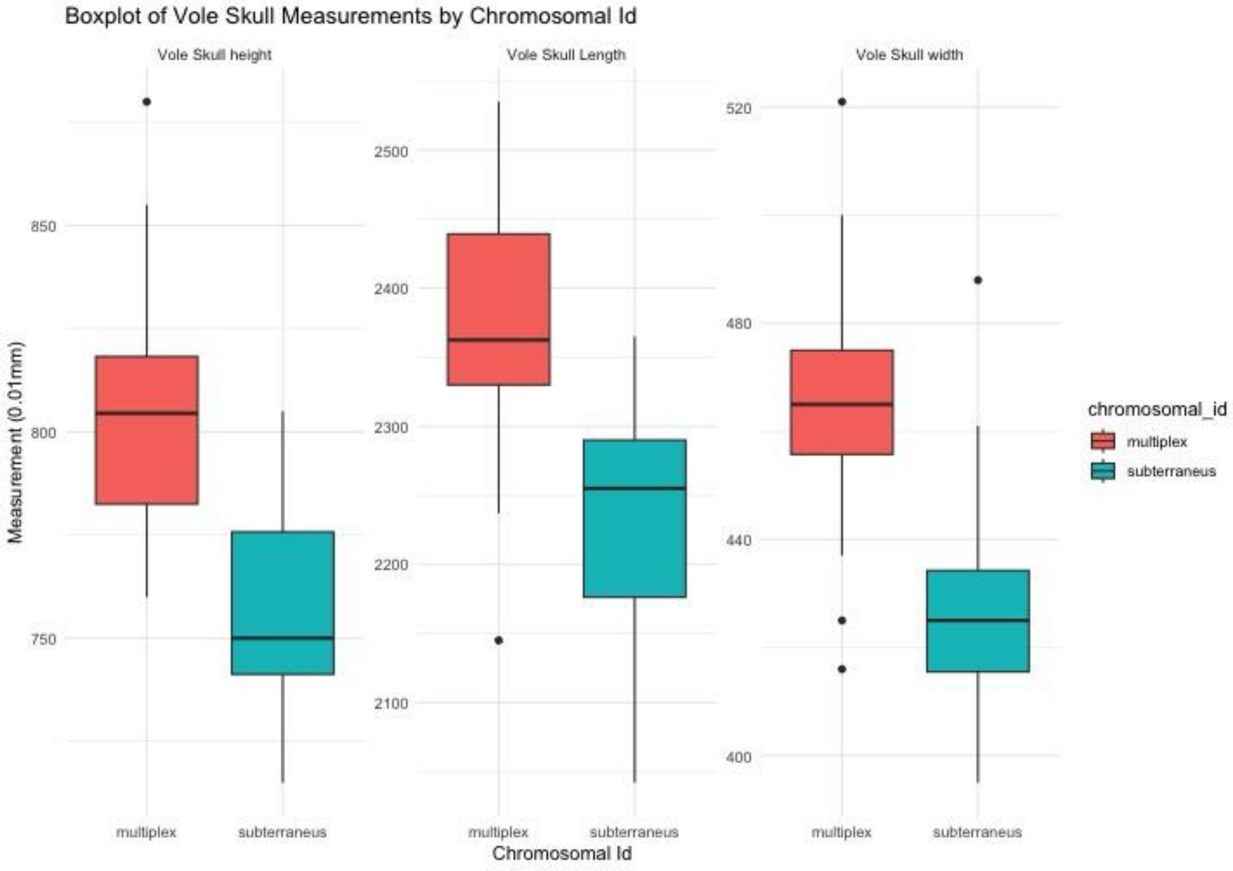
Descriptive Summary

	Known		
Chromosomal type	multiplex	subterraneus	Overall
Number	40	42	82
minimum skull length	2145	2042	2042
maximum skull length	2535	2365	2355
mean skull length	2374.30	2232.81	2301.83
minimum skull height	760	715	715
maximum skull height	880	805	880
mean skull height	804.2500	758.0714	780.60
minimum skull width	416	395	395
maximum skull width	521	488	521
mean skull width	465.4000	427.5952	446.04

Comment on Descriptive Summary: On an average multiplex seems to have higher vole skull measurement as compared to subterraneus

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

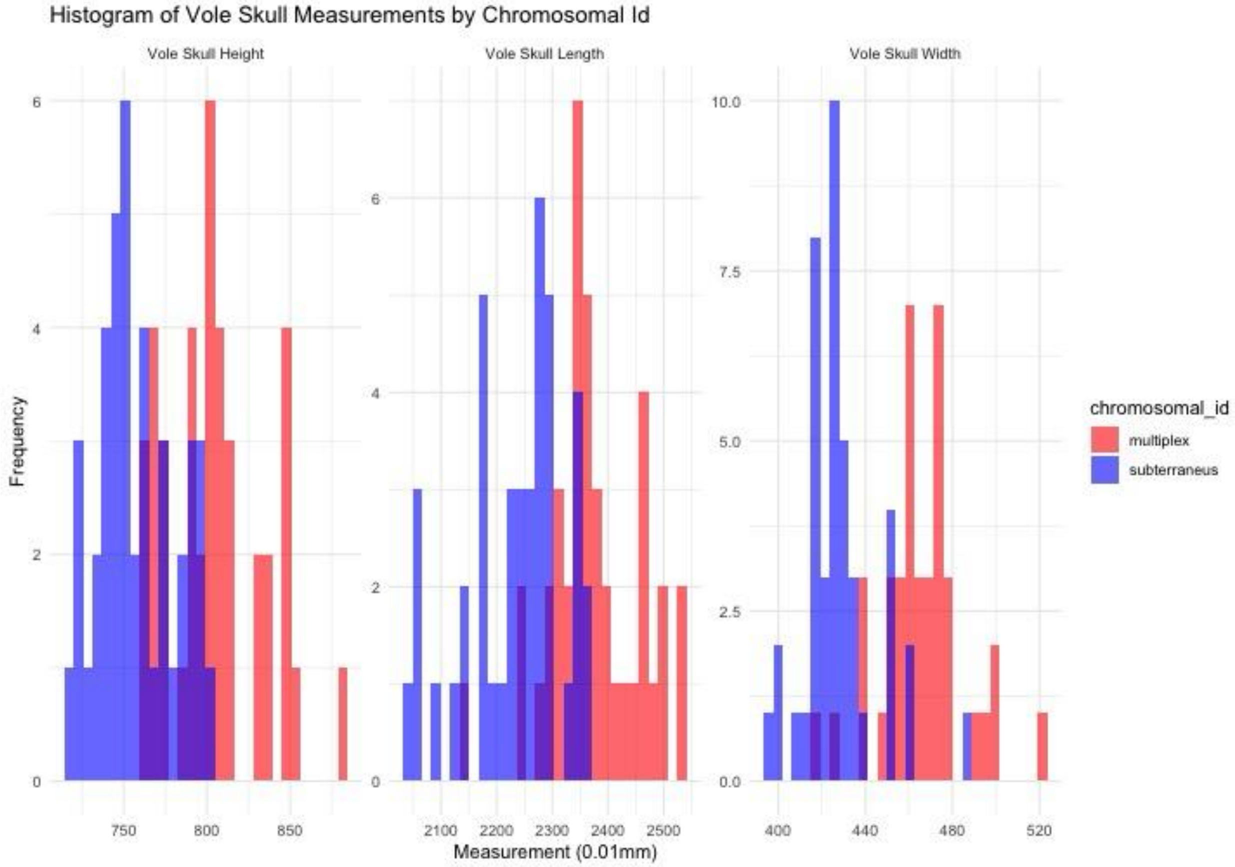
Boxplot of Measurements grouped by Chromosomal ID



Comment on Box plot: The average skull measurements—length, height, and width—are generally higher for multiplex. The skull height distribution for multiplex is left-skewed, indicating that more observations are concentrated on the lower end of the scale. In contrast, the skull height distribution for subterraneus is right-skewed, with more observations concentrated on the higher end. A similar pattern is observed for skull length, where multiplex shows a right-skewed distribution, and subterraneus exhibits a left-skewed distribution. However, the skull width distribution appears to be symmetrical for both groups.

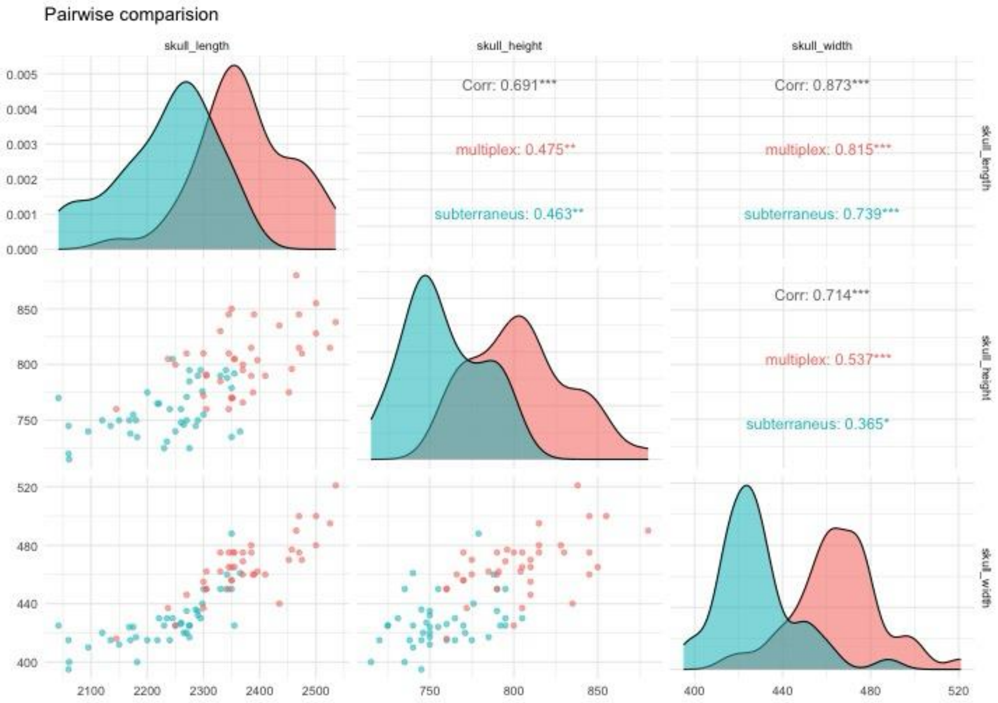
Group 5: Neha Karna, Prafulla Shrestha, Aidan Stewart, Josh Lefdal, Shivam Bhardwaj

Histogram of Vole Skull Measurements by Chromosomal ID



Comment on Histogram: The histograms for skull measurements—length, width, and height—exhibit a similar pattern with multiple peaks and gaps, indicating the presence of outliers. Consistent with the box plot, the distribution of skull height is left-skewed for multiplex and right-skewed for subterraneus. Similarly, skull length shows a left-skewed distribution for multiplex and a right-skewed distribution for subterraneus. In contrast, the skull width distribution for both multiplex and subterraneus appears to be more skewed toward the right.

Pairwise Comparison



Analyzing the correlation between variables in the merged data set reveals a significant correlation between independent variables, which raises concerns about multicollinearity. The scatter plot further represents closely scattered data points.

- Out of all three pairs of independent variables, skull length and skull height show relatively less correlation, around 0.7 for overall data and less than 0.5 for each chromosomal type, which suggests that the model with these two variables may be more reliable to predict the chromosomal identity.
- On the other hand, skull width and skull length show significantly large correlation, more than 80%. This suggests, that the model including skull width and skull length may not be accurate.
- The correlation between skull width and skull height is moderate for two groups individually and around 70% for overall data.

EDA for unknown dataset

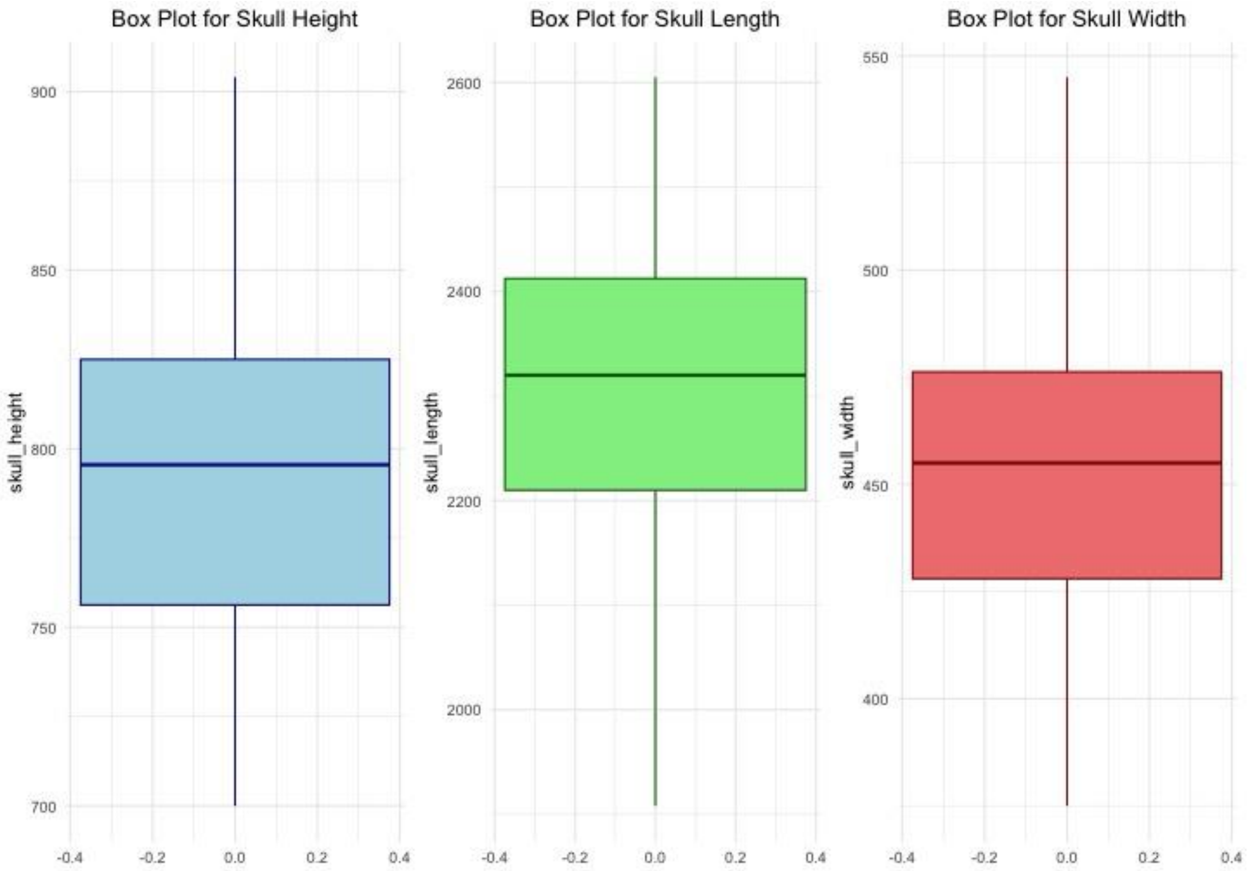
The EDA below is after replacing outliers as mentioned above.

Descriptive summary

variables	n	mean	min	max
index	199			
skull_length	159	2308.45	1908	2605
skull_height	162	795.08	700	904
skull_width	168	453.18	375	545

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

Boxplot



Comment on Box plot: The box plot for the unknown dataset does not reveal any apparent outliers. The distributions of all three measurements—skull length, width, and height—are slightly left-skewed, indicating that observations are more concentrated toward the lower end of the scale.

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

MODELING

Seven models were created based on the known dataset using logistic regression.

Models	**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

Table1: 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 ***)

Discussion on the models: The estimates in Table 1 are the coefficient values for intercept, length, height, and width for each model. They represent the change in the log-odds of the response variable for a unit change in a predictor variable. The intercept represents the estimated log-odds of the response variable (1 for subterraneus or 0 for multiplex) when all other predictors are set to zero. The corresponding p-values are included within the parentheses.

For instance, the estimated coefficients for intercept, length, height, and width for model_LWH were 70.27, -0.0012, -0.0450, and -0.0728, respectively, and the corresponding p-values were 2.52e-05 ***, 0.88476, 0.00741 **, and 0.02542 *. The p-value for skull length under the model_LWH was less than 0.05, indicating non-significance of the skull length variable for this model. The negative sign (-0.0012 for skull length for model_LWH) indicated an inverse relationship, while the value indicated the magnitude of the relationship, i.e., as skull length increases, the log-odds of the response variable or being in 'multiplex or subterraneus' decreases by 0.0012.

MODEL ASSESSMENT

Table 2: 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

Table 2 summarizes deviances, Akaike Information Criterion (AIC), Mean Squared Error (MSE), and accuracy of the models.

The null deviance represents the model's fit with intercepts only, i.e., without any predictors, while residual deviance represents the model's fit after adding the predictors. If the deviance significantly reduces from null deviance to residual deviance, it implies adding predictors helps improve the model. A lower residual deviance indicates a better model fit. For all seven models, null deviance is around 112.18, while models LWH and HW have the lowest residual deviance of 49.868 and 49.89, respectively.

Between models LWH and HW, accuracy is identical at 0.8780, while HW has a lower AIC value of 55.89 compared to that for LWH, 57.868.

AIC helps identify models that best balance goodness of fitness and model complexity. Since lower AIC values represent better models, Model HW provides the best performance (also evident by the least MSE value of 0.10199) and hence is the preferred model.

Table:3 Confusion matrix: Prediction of known species

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

The diagonals (shaded) represent correct predictions, and off-diagonals (white) represent incorrect predictions. Models with more than one variable are better at predicting multiplex type. Model_LWH and Model_HW show relatively higher numbers of correct predictions.

PREDICTION

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.

The table below summarizes the prediction model used for the respective subset of unknown data.

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

Table: Prediction counts by each model on unknown dataset

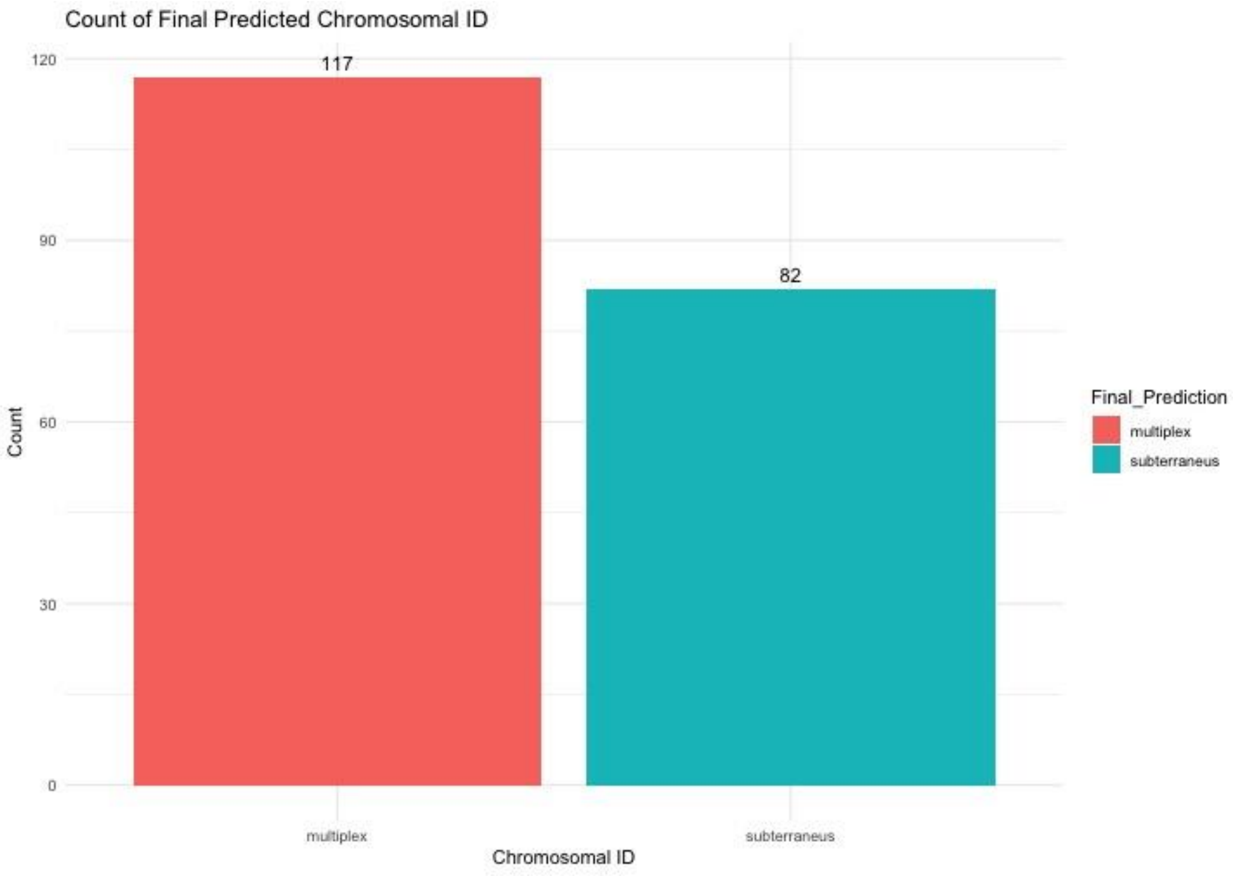
	Model_LWH	Model_LH	Model_HW	Model_LW	Model_L	Model_H	Model_W
subterraneus	45	59	55	62	74	69	72
multiplex	72	76	80	75	85	93	96
Total predicted	117	135	135	137	159	162	168

Final Prediction

None of the seven models were able to predict the whole unknown dataset due to unavailable measurements. 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.

EDA for Predicted dataset

Prediction count



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

CONCLUSION

Among all seven models, the model based on height and width (model_HW) is the best (least AIC and least MSE).
We recommend using model_HW to predict the unknown dataset. This model gives prediction for those observations that have skull measurements of height and width. Because unknown dataset has several observations with missing values (NA), this model doesn't give predictions for those observations. This issue could be resolved in two ways:

1. Using high frequency count to get the predictions for the observations with NAs: We could utilize other remaining models, get the count of the predicted species and replace NA with the species that has highest frequency.
2. Using high frequency count to get the predictions for all the observations: We chose the second method to get our final prediction. We utilized all seven models to get the final predictions. We counted the number of species predicted by each model and whichever species had the highest count, we assign that particular species to that observation. The rationale behind using all seven models is the combined accuracy of all seven models.

Table: Predictions using each model and Final prediction based on highest frequency

index	Model_LWH	Model_LH	Model_HW	Model_LW	Model_L	Model_H	Model_W	Final_Prediction
1	multiplex	multiplex	multiplex	subterraneus	subterraneus	multiplex	subterraneus	multiplex
2	NA	NA	subterraneus	NA	NA	subterraneus	subterraneus	subterraneus
3	NA	NA	NA	NA	subterraneus	NA	NA	subterraneus
4	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex
5	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex
6	NA	NA	NA	multiplex	multiplex	NA	multiplex	multiplex
7	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex
8	NA	multiplex	NA	NA	multiplex	subterraneus	NA	multiplex
9	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex
10	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus
11	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex
12	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex
13	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex
14	NA	NA	multiplex	NA	NA	subterraneus	multiplex	multiplex
15	multiplex	multiplex	multiplex	multiplex	multiplex	subterraneus	multiplex	multiplex
16	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex
17	NA	subterraneus	NA	NA	subterraneus	subterraneus	NA	subterraneus
18	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus
19	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	multiplex	subterraneus	subterraneus
20	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus

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

21	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	multiplex	subterraneus	subterraneus
22	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus
23	NA	multiplex	NA	NA	multiplex	multiplex	NA	multiplex
24	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus
25	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex
26	NA	NA	multiplex	NA	NA	multiplex	multiplex	multiplex
27	NA	NA	NA	multiplex	multiplex	NA	multiplex	multiplex
28	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex
29	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex
30	NA	NA	NA	multiplex	multiplex	NA	multiplex	multiplex
31	NA	NA	NA	multiplex	multiplex	NA	multiplex	multiplex
32	NA	NA	NA	NA	NA	NA	subterraneus	subterraneus
33	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex
34	NA	multiplex	NA	NA	multiplex	subterraneus	NA	multiplex
35	NA	NA	subterraneus	NA	NA	subterraneus	subterraneus	subterraneus
36	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex
37	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	multiplex	subterraneus	subterraneus
38	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus
39	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus
40	NA	NA	multiplex	NA	NA	multiplex	multiplex	multiplex
41	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus
42	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex
43	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex
44	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex
45	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	subterraneus	multiplex
46	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex
47	NA	multiplex	NA	NA	multiplex	multiplex	NA	multiplex
48	multiplex	multiplex	subterraneus	multiplex	multiplex	multiplex	subterraneus	multiplex
49	NA	NA	subterraneus	NA	NA	subterraneus	multiplex	subterraneus
50	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus
51	NA	NA	NA	subterraneus	subterraneus	NA	subterraneus	subterraneus
52	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus
53	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex
54	multiplex	subterraneus	multiplex	subterraneus	subterraneus	multiplex	subterraneus	subterraneus
55	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex
56	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex
57	NA	NA	NA	NA	NA	NA	multiplex	multiplex
58	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	multiplex	subterraneus	subterraneus
59	NA	NA	NA	NA	NA	multiplex	NA	multiplex

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

60	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex
61	multiplex	subterraneus	multiplex	subterraneus	subterraneus	multiplex	multiplex	multiplex
62	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex
63	NA	NA	multiplex	NA	NA	multiplex	multiplex	multiplex
64	NA	multiplex	NA	NA	multiplex	multiplex	NA	multiplex
65	NA	NA	multiplex	NA	NA	multiplex	multiplex	multiplex
66	NA	NA	NA	multiplex	subterraneus	NA	multiplex	multiplex
67	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex
68	NA	multiplex	NA	NA	multiplex	multiplex	NA	multiplex
69	NA	NA	NA	subterraneus	subterraneus	NA	subterraneus	subterraneus
70	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex
71	NA	multiplex	NA	NA	multiplex	multiplex	NA	multiplex
72	NA	NA	multiplex	NA	NA	multiplex	multiplex	multiplex
73	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex
74	NA	NA	subterraneus	NA	NA	subterraneus	subterraneus	subterraneus
75	NA	subterraneus	NA	NA	subterraneus	subterraneus	NA	subterraneus
76	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus
77	NA	NA	multiplex	NA	NA	multiplex	multiplex	multiplex
78	NA	NA	NA	subterraneus	subterraneus	NA	subterraneus	subterraneus
79	NA	subterraneus	NA	NA	subterraneus	subterraneus	NA	subterraneus
80	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex
81	NA	NA	NA	NA	NA	multiplex	NA	multiplex
82	multiplex	subterraneus	multiplex	multiplex	multiplex	subterraneus	multiplex	multiplex
83	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex
84	multiplex	multiplex	multiplex	multiplex	multiplex	subterraneus	multiplex	multiplex
85	NA	NA	NA	subterraneus	subterraneus	NA	subterraneus	subterraneus
86	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex
87	NA	NA	NA	NA	subterraneus	NA	NA	subterraneus
88	subterraneus	subterraneus	subterraneus	subterraneus	multiplex	subterraneus	subterraneus	subterraneus
89	NA	subterraneus	NA	NA	subterraneus	subterraneus	NA	subterraneus
90	NA	NA	subterraneus	NA	NA	subterraneus	subterraneus	subterraneus
91	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex
92	NA	NA	subterraneus	NA	NA	subterraneus	subterraneus	subterraneus
93	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus
94	NA	NA	NA	NA	NA	NA	subterraneus	subterraneus
95	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus
96	NA	multiplex	NA	NA	multiplex	multiplex	NA	multiplex
97	NA	NA	subterraneus	NA	NA	subterraneus	subterraneus	subterraneus
98	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex

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

99	NA	NA	NA	subterraneus	subterraneus	NA	subterraneus	subterraneus
100	subterraneus	subterraneus	subterraneus	subterraneus	multiplex	subterraneus	subterraneus	subterraneus
101	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex
102	NA	NA	NA	NA	NA	multiplex	NA	multiplex
103	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex
104	multiplex	multiplex	multiplex	subterraneus	subterraneus	multiplex	subterraneus	multiplex
105	NA	NA	NA	NA	NA	NA	subterraneus	subterraneus
106	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus
107	NA	subterraneus	NA	NA	subterraneus	multiplex	NA	subterraneus
108	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex
109	NA	NA	NA	NA	NA	NA	multiplex	multiplex
110	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus
111	NA	NA	NA	NA	NA	NA	multiplex	multiplex
112	multiplex	multiplex	multiplex	subterraneus	multiplex	multiplex	subterraneus	multiplex
113	NA	NA	NA	NA	multiplex	NA	NA	multiplex
114	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex
115	NA	multiplex	NA	NA	multiplex	subterraneus	NA	multiplex
116	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex
117	NA	NA	multiplex	NA	NA	multiplex	multiplex	multiplex
118	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex
119	NA	NA	NA	subterraneus	subterraneus	NA	subterraneus	subterraneus
120	subterraneus	subterraneus	subterraneus	multiplex	subterraneus	subterraneus	multiplex	subterraneus
121	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex
122	NA	NA	NA	NA	NA	multiplex	NA	multiplex
123	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex
124	NA	NA	multiplex	NA	NA	multiplex	multiplex	multiplex
125	NA	NA	NA	NA	multiplex	NA	NA	multiplex
126	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus
127	NA	NA	NA	subterraneus	subterraneus	NA	subterraneus	subterraneus
128	NA	multiplex	NA	NA	multiplex	multiplex	NA	multiplex
129	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus
130	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex
131	NA	NA	NA	NA	NA	NA	multiplex	multiplex
132	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus
133	NA	NA	NA	subterraneus	subterraneus	NA	subterraneus	subterraneus
134	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex
135	subterraneus	subterraneus	subterraneus	multiplex	multiplex	subterraneus	multiplex	subterraneus
136	NA	NA	NA	NA	NA	subterraneus	NA	subterraneus
137	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex

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

138	multiplex	multiplex	multiplex	subterraneus	subterraneus	multiplex	subterraneus	multiplex
139	NA	NA	NA	multiplex	multiplex	NA	multiplex	multiplex
140	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex
141	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex
142	NA	NA	NA	NA	NA	multiplex	NA	multiplex
143	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus
144	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex
145	NA	NA	NA	NA	NA	NA	multiplex	multiplex
146	multiplex	multiplex	multiplex	multiplex	subterraneus	multiplex	multiplex	multiplex
147	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex
148	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus
149	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex
150	NA	NA	NA	NA	NA	subterraneus	NA	subterraneus
151	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex
152	NA	subterraneus	NA	NA	subterraneus	subterraneus	NA	subterraneus
153	multiplex	multiplex	multiplex	subterraneus	subterraneus	multiplex	subterraneus	multiplex
154	NA	NA	NA	multiplex	subterraneus	NA	multiplex	multiplex
155	NA	NA	NA	multiplex	multiplex	NA	multiplex	multiplex
156	NA	NA	NA	NA	NA	NA	subterraneus	subterraneus
157	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus
158	NA	subterraneus	NA	NA	subterraneus	subterraneus	NA	subterraneus
159	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus
160	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus
161	NA	NA	NA	NA	NA	NA	multiplex	multiplex
162	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus
163	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus
164	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus
165	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex
166	multiplex	multiplex	subterraneus	multiplex	multiplex	subterraneus	multiplex	multiplex
167	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus
168	NA	NA	NA	subterraneus	subterraneus	NA	subterraneus	subterraneus
169	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus
170	multiplex	multiplex	multiplex	subterraneus	subterraneus	multiplex	subterraneus	multiplex
171	NA	NA	NA	NA	NA	NA	multiplex	multiplex
172	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex
173	NA	NA	NA	subterraneus	subterraneus	NA	subterraneus	subterraneus
174	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus
175	NA	NA	NA	NA	NA	multiplex	NA	multiplex
176	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex

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

177	multiplex	subterraneus	multiplex	multiplex	multiplex	subterraneus	multiplex	multiplex
178	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex
179	NA	subterraneus	NA	NA	subterraneus	multiplex	NA	subterraneus
180	NA	NA	multiplex	NA	NA	multiplex	multiplex	multiplex
181	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus
182	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex
183	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus
184	multiplex	multiplex	multiplex	subterraneus	multiplex	multiplex	subterraneus	multiplex
185	NA	NA	NA	NA	NA	NA	multiplex	multiplex
186	multiplex	subterraneus	multiplex	subterraneus	subterraneus	subterraneus	multiplex	subterraneus
187	multiplex	subterraneus	multiplex	multiplex	subterraneus	subterraneus	multiplex	multiplex
188	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus
189	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus
190	NA	NA	subterraneus	NA	NA	subterraneus	multiplex	subterraneus
191	subterraneus	subterraneus	subterraneus	subterraneus	multiplex	subterraneus	subterraneus	subterraneus
192	subterraneus	subterraneus	subterraneus	multiplex	subterraneus	subterraneus	multiplex	subterraneus
193	NA	NA	NA	NA	NA	subterraneus	NA	subterraneus
194	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus
195	NA	NA	NA	multiplex	subterraneus	NA	multiplex	multiplex
196	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex	multiplex
197	NA	NA	NA	multiplex	multiplex	NA	multiplex	multiplex
198	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus	subterraneus
199	NA	NA	NA	NA	NA	NA	multiplex	multiplex

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

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)

What is Exploratory Data Analysis (EDA)? (2024b). IBM – United States https://www.ibm.com/topics/exploratory-data-analysis

Zumel, N., & Mount, J. (2019). Practical Data Science with R, Second Edition. Manning.