Mohammed Uddin
CISC 3225
Final Report

1. Introduction: Describe your dataset. What is its purpose and what kind of data does it contain? What do you hope to discover in your analysis?

The dataset I chose contains thousands of rows of evolutionary data on different hominid species. With 12000 rows and 28 columns, this dataset covers a wide range of characteristics of different hominins. The purpose of the dataset, as stated by the publisher of the dataset, is to predict either genders and species, or whether a row is bipedal or not.

The columns contained in the dataset consist of three numeric columns and 25 categorical columns. For the sake of keeping the final report clean, I will not be explaining the columns here, as I did so at the beginning of the notebook for the project.

From the dataset and the analysis of the dataset, I hope to see if there are certain features or characteristics of different human species that influence the presence of bipedalism.

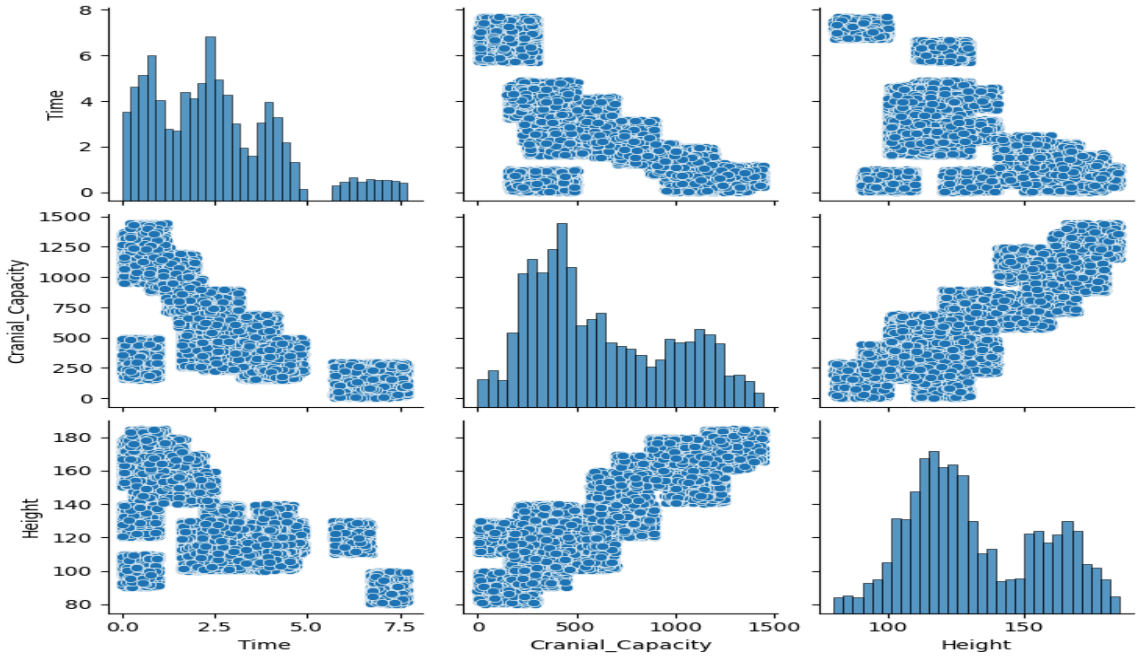
2. Exploratory analysis. Describe the characteristics of the data you observe, with visualization to support your observations. Use domain knowledge to explain interesting observations, citing external sources if necessary.

Time is really the only column that deserves an explanation here, just because it may be confusing to interpret. The value of time is millions of years ago, therefore the higher the value the more in the past it is.

From the EDA, I noticed that there were a lot of columns that I felt weren't relevant to answering the analysis question. The columns that I ended up removing are Location, Zone, and current country. The reason why I don't want these columns in the data for analysis is because it will make the table more complex and because habitat gives sufficient information of what kind of environment the species was living in. The habitat column alone can make up for the information lost by removing location, zone, and current country. I've also decided to drop the columns: Sexual_Dimorphism, Vertical_Front, Tooth_Enamal, and Migrated. Sexual_dimorphism is just if the appearance between sexes in the species is very distinct. Vertical_Front is just more information regarding the bone structure in the face. Tooth_Enamal is not that important of a characteristic, also there are two other columns relating to teeth. Migrated is not logically a reason for bipedalism, it is an effect of it.

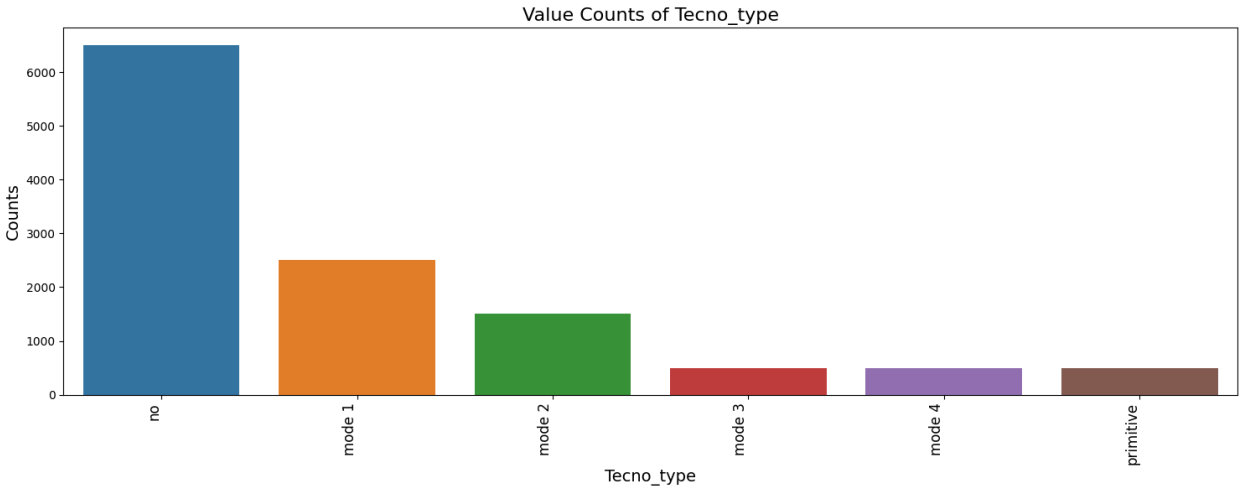
In the EDA, I also presented a pairplot that shows the correlations between the numeric columns (time, cranial capacity, height) in the data. The pair plot revealed many things and led me to read articles that explain the correlation. For example, Time - Cranial Capacity has a correlation of -.66, which is negatively and moderately strong. As time increases, cranial capacity decreases, which is something anthropologists have observed. The article, Why human

brains were bigger 3,000 years ago, explains why cranial capacity tends to get smaller and it is due to us not needing to store so much information because we can rely on each other, on writing, and now the internet. In the correlation pair plot, I also observed that there were natural groups that existed, as can be seen in the figure below.

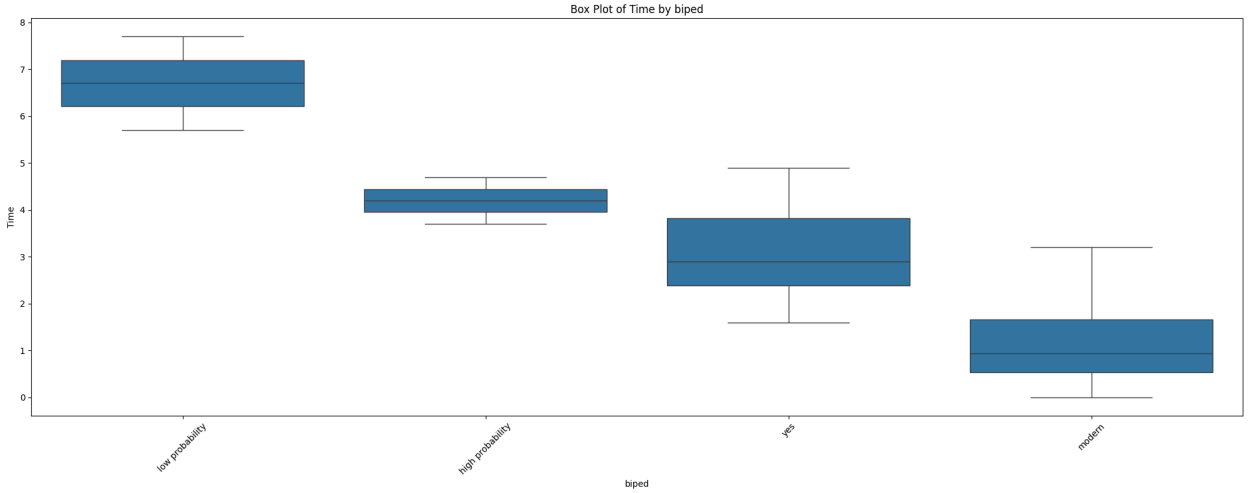


Upon observing this, I decided a linear regression and a cluster analysis would provide some more information about the dataset.

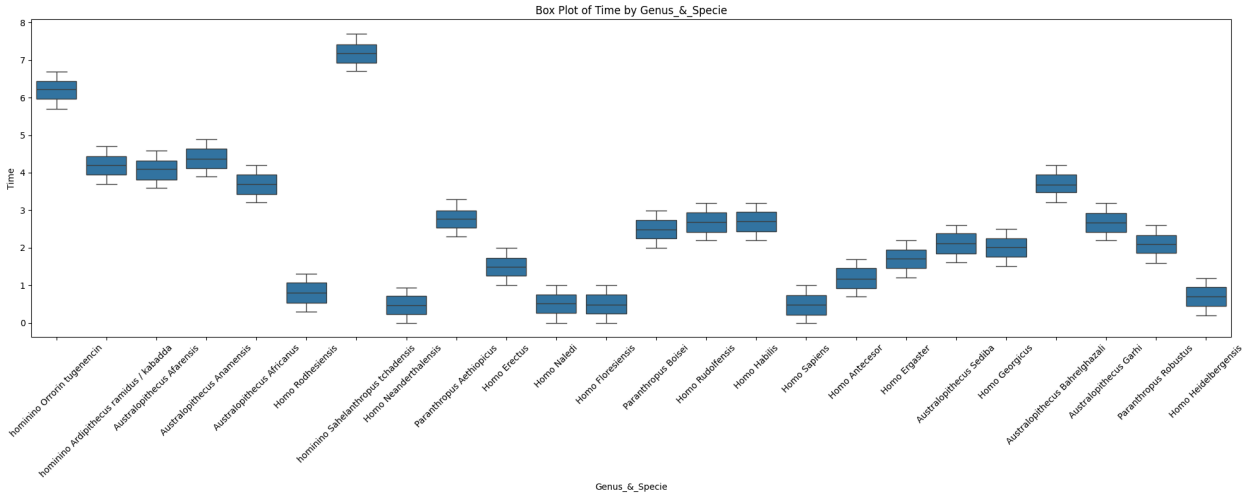
In the EDA, I also provided some visualizations for the categorical variables and observed a lot of interesting things. From the value counts visualizations, one stood out in particular. The counts for tecno type show that most hominid species never made it to the point of crafting stone tools. This is unrelated to the analysis question, as were most of the observations I made in the EDA since it wasn't possible to see in that step of the project what categorical variables influence bipedalism.



Many interesting observations also came from the box plots of the categorical variables plotted against the numeric column. Since there are so many box plots, I will only focus on box plots that tell a clear story and are related to the analysis question.



When the biped column is plotted against the time column, it is observed that as time progressed, or in other words, as we approach present time, bipedalism has a more common occurrence than it does in the past. The small whiskers and medians of the box plot show this.



For the figure above, this may be a wrongful interpretation of the box plots of species against time, but it seems as though all human species lived only for a certain and very similar amount of time. This may be caused by a sampling bias in the data, where fossils from a certain species were just all found together in an area and all dated to be around the same time. But assuming that the interpretation is correct because we know that many human species lived for a short time due to environmental reasons or evolving into something else, these box plots indicate that we have an expiration date as a species.

The 16th box plot is us, and the species or plots to the right of it are species that we know existed during the same time as us. But they died (and sort of live through us because of intermingling) due to reasons unclear to us as of now. We can only really speculate what

happened to the other species. That is, that they evolved and/or became another hominid species or just died off. As for us, well, it is clearly said by scientists today that we are destroying our world and that it is becoming more uninhabitable. Though on the brightside, we are becoming extremely different in an overall good way to homo sapiens that lived a few hundred years ago. That is, we are getting longer life spans, and soon, we will be the editers of our genes, as explained by Yuval Noah Harari in his famous book, Homo Deus.

3. High-level analysis. Introduce each of your analyses and present them, with relevant visualizations, in their own sections.

1. In my first analysis, I used Pandas to see if there's any information to see influences of characteristics on bipedalism. The results were summary statistics of numeric columns as they were on a certain category for bipedalism.

Summary stats for biped type: low probability

	Time	Cranial_Capacity	Height
count	1000.000000	1000.000000	1000.000000
mean	6.693097	150.628199	105.127598
std	0.563387	87.617154	16.190755
min	5.700095	0.074910	80.009030
25%	6.215290	76.365827	89.929220
50%	6.701511	146.929925	104.983640
75%	7.183456	231.025420	120.434865
max	7.699417	299.503600	129.965510

Summary stats for biped type: high probability

	Time	Cranial_Capacity	Height
count	500.000000	500.000000	500.000000
mean	4.191768	300.393117	109.993517
std	0.287294	86.873593	5.634087
min	3.700208	150.641350	100.011320
25%	3.951217	227.530483	105.107612
50%	4.202263	299.149065	109.441105
75%	4.443350	377.954415	115.055610
max	4.699689	449.835200	119.858570

Summary stats for biped type: yes

	Time	Cranial_Capacity	Height
count	5000.000000	5000.000000	5000.000000
mean	3.067028	444.870722	117.956010
std	0.834224	140.152875	9.396972
min	1.600071	150.843650	100.009800
25%	2.392718	344.877180	111.299995
50%	2.897004	436.839650	117.519390

75%	3.815822	540.218650	124.871167
max	4.899290	839.391780	139.973370

Summary stats for biped type: modern

	Time	Cranial_Capacity	Height
count	5500.000000	5500.000000	5500.000000
mean	1.139610	886.675139	150.458540
std	0.770637	329.680507	22.099036
min	0.000529	151.415400	90.027430
25%	0.535215	686.618847	137.712148
50%	0.944606	963.498270	155.740060
75%	1.663369	1143.355290	166.744223
max	3.198981	1448.397470	184.981450

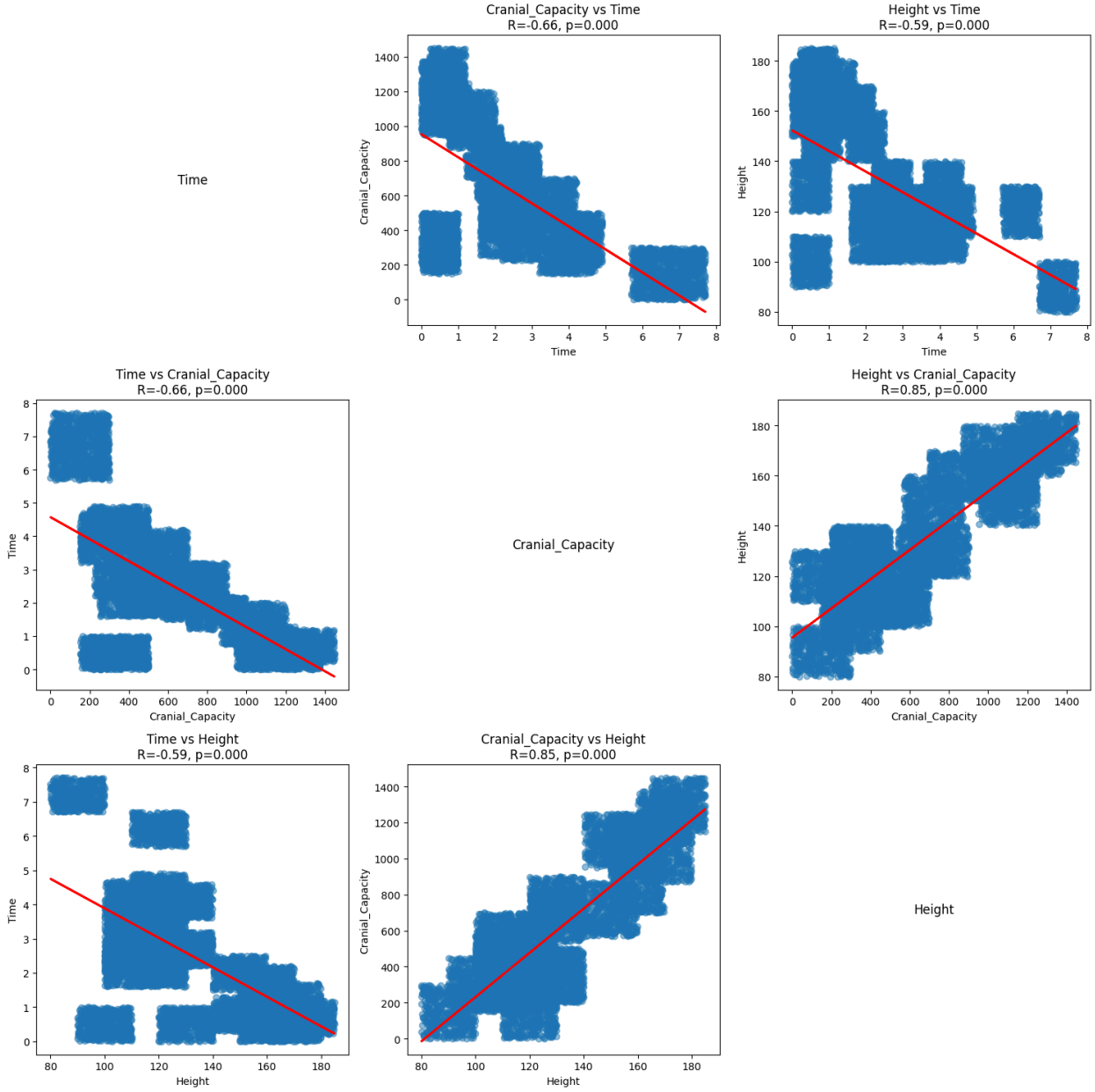
The statistic summaries above reveal a few things about time, cranial_capacity, and height against the unique values of bipedalism. As time approaches the modern day, it is a more common occurrence for a species to likely be bipedal, as the count for being bipedal is higher. Also, as height increased, so to did the count of bipedalism.

2. In my second analysis, I was curious to see if the correlation calculated for the cranial capacity and height is statistically significant. So I used the Pearson correlation coefficient to get a coefficient value, along with a p-value to see if there was a significant correlation between the two numeric columns. Part of the reason I decided to do this was to see if an article's findings of bigger brains correlating with height was valid. In the experiment, my Null Hypothesis (H0): There is no significant correlation between cranial capacity and height in the population. My Alternative Hypothesis (H1): There is a significant correlation between cranial capacity and height in the population. The results came out to this:

Correlation Coefficient: 0.8459252276033566
P-value: 0.0

Since the correlation coefficient is near 1 and the p-value is 0, I reject the null hypothesis. The findings of the article seem to have been right. Since the article used it very own data to come to the conclusion, and I'm using a dataset unrelated from the study in the article, and both datasets show the same trends, strong evidence is provided to the idea that having a bigger brain means you are tall and vice versa.

3. For my third analysis, I decided to do linear regressions for my numeric columns.



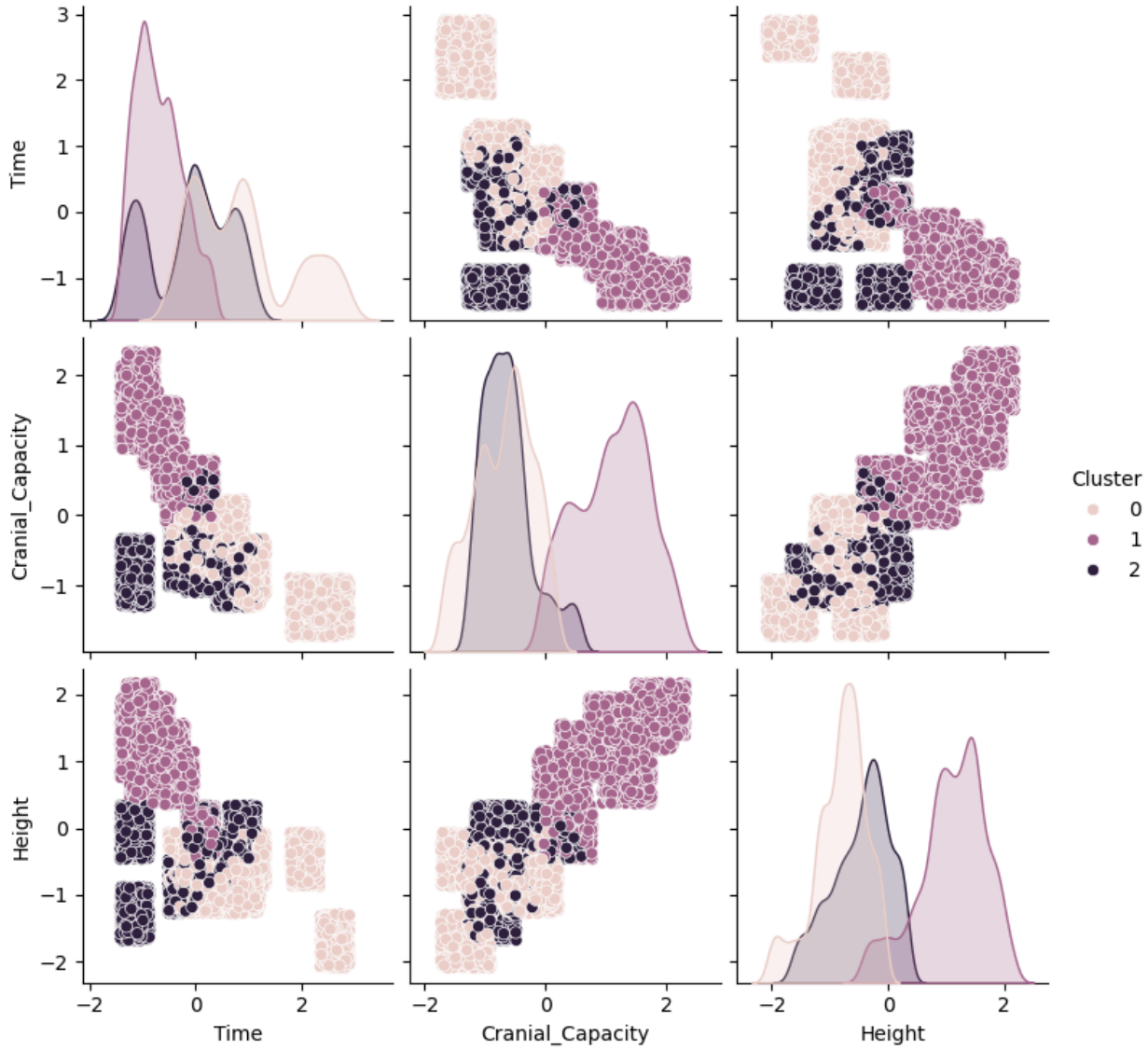
The closer time is to 0, the more closer it is to present time. There seems to be more species togther the more closer we are to the present time. It makes sense that the species are cluttered close, as the data comes closer to present time because as we know, there weren't huge differences between the hominid species that lived alongside homo sapiens. The values of the r-statistic computed show that the numeric columns do have a significant relationship. As time went on, hominid brain capacity got bigger, as well as heights. There's also a positive relationship for height and brain capacity, which was sad to see.

4. I did chi-squared tests to see if there is any assocoiation between bipedalism and other categorical variables. The null hypothesis is that there is no association between bipedalism and another categorical column and the alternative hypothesis is that there is an association between bipedalism and another categorical column.

Variable	**Chi-Squared**	**P-Value**
Genus_&_Specie	36000.000000	0
Habitat	12650.000000	0
Incisor_Size	6992.727273	0
Jaw_Shape	13227.272727	0
Torus_Supraorbital	9069.421488	0
Prognathism	13464.545455	0
Foramen_Mágnum_Position	22536.363636	0
Canine Size	4475.015893	0
Canines_Shape	7680.000000	0
Tecno	10439.160839	0
Tecno_type	10439.160839	0
biped	36000.000000	0
Arms	8509.090909	0
Foots	6880.000000	0
Diet	17763.636364	0
Hip	8945.454545	0
Anatomy	9600.000000	0
Skeleton	3281.652893	0

Based on the output of the tests, there is significant association between bipadalism and all of the categorical columns. All of the tests between bipedalism and another cat column result in a high computed chi statistic and a low p-value, which gives strong evidence against the null hypothesis.

5. I did a cluster analysis of the data. I used K-means and found that k=3 was the best number of clusters to group the data.



In the charts above, the clusters can be seen in a pair plot of the numeric columns in the data. Before clustering them, it was clear that there were groups in the data because of the way the points were together. The cluster assignments make it clear which species are more similar to each other than others. I can't inclue the counts of what the clusters found in the report because the chart is very wide. However, to say some stuff about what the clustering found, it found that Homo Sapiens (found in cluster 1), is closely related to 8 other species (Homo Antecesor, Homo Erectus, Homo Ergaster, Homo Georgicus, Homo Habilis, Homo Heidelbergensis, Homo Neanderthalensis, and Homo Rodhesiensis), whom some have been proved to be extremely similar to us, such as Homo Antecesor, Homo Erectus, Homo Habilis, and Homo

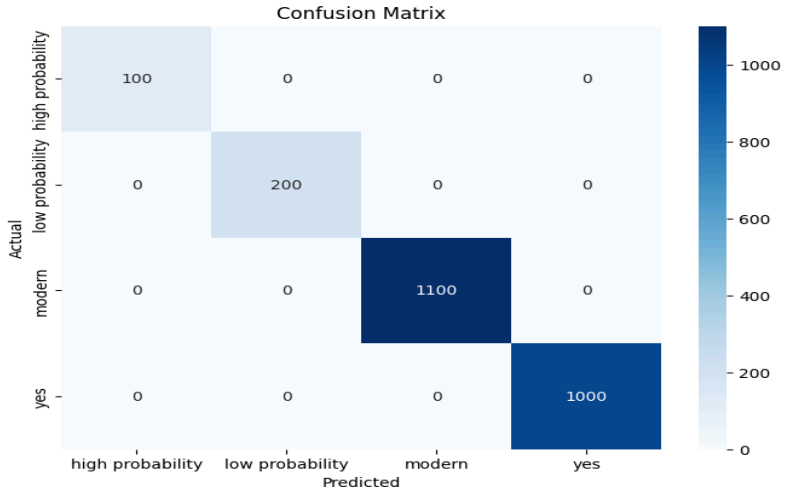
Neanderthalensis. Many articles list some of the species found in cluster 1 to be closely related to Sapiens, however, some are not mentioned. Maybe this is the basis for more research.

6. Prediction of bipedalism. I was curious to see if a logistic regression model could predict bipedalism from the rest of the columns in the data. The model seemed to have found enough patterns to be able to accurately predict bipedalism, as the metrics of the model showed that it was performing very well on the test set.

Accuracy: 1.00

Classification Report:

	precision	recall	f1-score	support
high probability	1.00	1.00	1.00	100
low probability	1.00	1.00	1.00	200
modern	1.00	1.00	1.00	1100
yes	1.00	1.00	1.00	1000
accuracy			1.00	2400
macro avg	1.00	1.00	1.00	2400
weighted avg	1.00	1.00	1.00	2400



The model seems to have been successfully able to use the information from both the categorical and numeric columns to accurately predict if an instance of a hominin found in the dataset may have been bipedal or not.

4. Conclusions. What did you learn from this project? End with a thoughtful discussion of the data and insights you obtained from your analysis, and draw conclusions.

In this project, I found several significant insights about hominin evolution, particularly regarding bipedalism and other characteristics. Here are the main conclusions drawn from the data analysis:

Time Trends in Bipedalism:

- As we move closer to the present time, bipedalism becomes more common among hominins. This is evident from the statistical summaries and visualizations showing higher counts of bipedal species in more recent periods. This trend aligns with the understanding that bipedalism is a key trait in the evolution of modern humans.
- The analysis of the time column, which represents millions of years ago, shows a clear pattern where bipedalism is more frequently observed in hominins closer to present times. This suggests that bipedalism was a later adaptation in the evolutionary history of hominins.

Correlation Between Physical Characteristics:

- The Pearson correlation analysis revealed a strong positive correlation between cranial capacity and height, indicating that species with larger brains tended to be taller. This finding supports existing anthropological theories suggesting that larger brain sizes are associated with greater body heights in hominins.
- The linear regression analysis confirmed significant relationships among the numeric columns, highlighting how physical attributes such as brain size and height evolved together over time.

Cluster Analysis of Hominin Species:

- The K-means clustering analysis identified three distinct clusters, grouping hominins based on their physical and categorical characteristics. Notably, Homo sapiens were clustered with species like Homo neanderthalensis, Homo erectus, and Homo habilis, indicating significant similarities among these species.
- This clustering provides a clearer picture of evolutionary relationships and similarities, suggesting that some species are more closely related to modern humans than others.

Predictive of Bipedalism:

- A logistic regression model was successfully trained to predict bipedalism based on the dataset's features. The model achieved perfect accuracy on the test set, demonstrating that

the dataset contains sufficient information to reliably predict bipedalism. I realize this doesn't mean much, considering not all known hominin species are in the dataset, though it is worth mentioning as it shows that with enough data, we can predict whether a species is bipedal or not.

- This predictive capability underscores the robustness of the dataset and the strength of the relationships between the various characteristics and bipedalism.

The dataset and the subsequent analyses have provided a wealth of information about the evolutionary traits of hominins. The trends observed in bipedalism, cranial capacity, and height reflect known evolutionary patterns and provide additional quantitative support for these theories, as brought up in the articles linked in the report. Also, the clustering analysis further enhances our understanding by visually and statistically grouping species based on their similarities, offering potential areas for further research into the evolutionary pathways of these hominins.

The predictive success of the logistic regression model highlights the potential for using machine learning techniques to gain deeper insights into human and other animal evolutionary biology. By accurately predicting bipedalism, the model validates the dataset's quality and the meaningful relationships found between the variables.

In conclusion, this project has demonstrated the power of data analysis and machine learning in studying evolutionary biology. The insights gained from the temporal trends, physical correlations, and species clustering provided a deep understanding of hominin evolution. The predictive model's success further emphasizes the value of machine learning methods in uncovering patterns and making predictions based on historical data. This project not only answers specific questions about bipedalism but also opens up new avenues for research and exploration in the field of anthropology and evolutionary studies.