CM4125 - Coursework Pt 1

Design of a Data Visualisation

Sergio Castillo - 1513228

Table Of Contents

Introduction

Paleontology is an ever-changing field, with a thriving community of passionate individuals, focused on understanding the evolution of life. Scientists have learnt more over the last 25 years than in the previous 250 - from the color of their skin and feathers to how they lived and evolved [5], and the rate of the discoveries only seems to accelerate, with new species being named every 2 weeks, on average [0].

Dinosaurs are a topic that many people interact with during childhood, which seems to be forgotten about once infancy is left behind. From books to films, toys and even videogames, it becomes apparent that humans are instinctively attracted to the creatures that roamed our planet before humans ever existed. But the notion of giant, bloodthirsty creatures, as portrayed by Hollywood films such as Jurassic Park has twisted the general population's perception of dinosaurs, providing an image that does not match the current understanding of these animals.

The intention behind this study is to shed some light on the field of paleontology, and attempt to visualize the current understanding on the time span, diversity and location of dinosaurs. This will be done through an exploration of paleontologic data, and the use of an infographic with a diverse set of data visualizations, to offer a more accurate and modern understanding of our knowledge on prehistoric animals.

Data Sources

Mesozoic Periods Web Scraping

The first source of data used on this study is the 3 periods of the Mesozoic era, with figures for the approximate start and end for such periods [3]. The information was scraped from Wikipedia using the Beautiful Soup package for Python [4], and the list that displayed the necessary information was parsed into lists of strings that were then assembled into a dataframe.

The motivation for this source of data was to have a proper definition of the time periods in which dinosaurs existed. Time can be divided using geological terms, which themselves can be subdivided to produce more meaningful and identifiable time frames, that can serve as a reference for paleobiology. The information produced by the scraped data frame was then used on other parts of the data exploration stage.

The dataset contains 3 columns and 3 rows, and it includes data for the period name, the start and the end of the period in millions of years.

Period Name	Start	End
Triassic	251.902	201.3
Jurassic	201.300	145.0
Cretaceous	145.000	66.0

Dinosaur List Dataset

The second source of data used in this research was the Dinosaur List, a dataset containing every dinosaur species found so far (2020). It includes the species name, diet, time period and the country where their remains were found [1]. It was published by the user KumaKo on the Data Science platform, Kaggle, so no extra techniques for data gathering were used but downloading a CSV file and then loading it on a Python Notebook using Pandas.

The dataset originally contained 4 columns and 1154 rows, and as well as providing information for individual species, it serves as the foundation for another 3 datasets which display the number of dinosaurs found based on the period of time in which they lived, the region where they were found and the diet they were believed to consume.

Name	Period	Diet	Country
Dilong	Cretaceous	carnivore	China
Oxalaia	Cretaceous	carnivore	South America
Staurikosaurus	Triassic	carnivore	South America
Zuniceratops	Cretaceous	herbivore	North America
Palaeoscincus	Cretaceous	herbivore	North America

The Paleobiology Database

The final source of data was the Paleobiology Database, a database that contains 500 million years of evolution, with information for every specimen found in North America. The dataset was also published in Kaggle by Jorge Orellano [2]. Once the zip file was downloaded, it included 4 files, with information about the occurrences of extinct species, information on Geological strata and descriptions of every time interval of biological diversity. From this source only two files were used in the ipython notebook:

- The first one was “Diversity Over Time”, which provides information on every interval of time of the Earth's existence where life has been documented, as well as the number of specimens found that lived during such intervals, containing 10 columns and 100 rows. The table contains information for the interval name and identification, the time frame and the number of occurrences from such interval.

	interval_no	interval_name	max_ma	min_ma	X_Ft	X_bL	X_FL	X_bt	sampled_in_bin	n_occs
46	136	Norian	228.000	208.500	154	161	366	245	803	7297
4	741	Gelasian	2.588	1.806	38	13	49	1488	199	655
10	101	Langhian	15.970	13.820	158	120	76	1524	736	4647
51	653	Induan	252.170	251.200	36	14	37	222	112	1625
25	115	Coniacian	89.800	86.300	48	16	29	775	206	772

- The second file was “Occurrences”, which provides information about every specimen of every organism found in North America, with a size of 17 columns and 493,763 rows. The columns provide information on the identified and accepted name, identification numbers for its occurrence, whether is a species, a genus or a family, the period of time where it was believed to have lived and the names of the intervals belonging to such time frames.

	occurrence_no	record_type	reid_no	flags	collection_no	identified_name	identified_rank	identified_no	difference	accepted_name	accepted_rank	a
38542	60934	occ	NaN	NaN	4310	Liothyris sp.	genus	17821	NaN	Liothyris	genus	
78063	128316	occ	NaN	NaN	10555	Esmeraldina ? cometes	species	19127	species not entered	Esmeraldina	genus	
335306	805553	occ	NaN	NaN	87926	Cheloniceras inconstans	species	14683	species not entered	Cheloniceras	genus	
156051	278425	occ	NaN	NaN	26617	Cranaena sp.	genus	30116	NaN	Cranaena	genus	
424330	1065680	occ	NaN	NaN	131677	Trigoniocardia (Trigoniocardia) haitensis	species	102122	NaN	Trigoniocardia (Trigoniocardia) haitensis	species	

Data Pre-Processing

Mesozoic Periods

As previously mentioned, the first dataset was scraped from the web by using beautiful soup. The Wikipedia article was parsed, and the list with the necessary information was retrieved. After that, the text of the list elements were turned into strings and then split into words, so the period names and the numbers could easily be added to previously created arrays. A new dataframe was then formed, the period name was set as the index and the Start and End columns were changed to numeric values.

```
# request the url
r = requests.get("https://en.wikipedia.org/wiki/Mesozoic")

# soup it
soup = BeautifulSoup( r.content, "html.parser")

# Getting the list of lists and the list of periods
ul_list = soup.findAll("ul")
period_list = ul_list[4].findAll("li")

for period in period_list:
    # Get the list of items in the text
    period_string_list = period.get_text().split()

    # Adding the period name, start and end
    periods.append(period_string_list[0])
    start.append(period_string_list[1][1:])
    end.append(period_string_list[3])

# Creating the dataframe from the scraped lists
period_ranges = pd.DataFrame({'Period Name':periods,'Start':start,'End':end})

# Set the Period Name as Index
period_ranges = period_ranges.set_index('Period Name')

# Change the types of Start and End to float numbers
period_ranges = period_ranges.astype({'Start':'float64','End':'float64'})
```

The result of this preparation produced the following table:

Period Name	Start	End
Triassic	251.9	201.3
Jurassic	201.3	145.0
Cretaceous	145.0	66.0

Dinosaur List

The Dinosaur list contained a number of issues that had to be addressed:

- Inconsistencies in the spelling of certain values, such as the use of uppercase letter, extra characters and even spelling mistakes.
- Some entries displayed more specific information on time periods or diets.
- There were entries with missing values, such as time periods, diets or regions.
- In the case of the countries, the inconsistencies of the values was aggravated:
    - Some entries displayed continents rather than countries.
    - Some entries displayed multiple countries.
    - Some entries provided information about their names or taxonomy.

The dataset was cleaned and dealt with in the following way:

- Unique values for the different columns were displayed and studied. Once terms that were referring to the same concept were identified, they were then grouped in a list and replaced with a single, unifying term.
- In cases where the values were specific, a more general term was chosen to replace it with, e.g Insectivores being Carnivores, Early Cretaceous being Cretaceous, etc.
- The missing values were dealt with differently depending on the column:
    - On the time period column there was only 1 missing value, so the entry was deleted.
    - On the diet column, a new category was created for unknown diets.
- The “Country” column required some extra cleaning:
    - Firstly, the countries have been replaced with the overall region they belong to. The following regions have been chosen: North America, South America, Europe, Africa, Asia, Oceania and Antarctica. The column was then renamed to “Regions”.
    - A separate category for dinosaurs extending throughout different regions has been provided.
    - Upon further examination, all the species with descriptions rather than locations seemed to come from China or Mongolia, so the region was assigned to Asia.

```
# Period Column (Diet column followed a similar structure

# Getting all unique periods
dinosaurs.Period.unique()

# Replacing Unstable Periods with general names
dinosaurs['Period'] = dinosaurs['Period'].replace([...], 'Triassic')
dinosaurs['Period'] = dinosaurs['Period'].replace([...], 'Jurassic')
dinosaurs['Period'] = dinosaurs['Period'].replace([...], 'Cretaceous')

# Getting rid of the "unknown" rows
dinosaurs[dinosaurs['Period'] == '(unknown)']
dinosaurs.drop(dinosaurs.loc[dinosaurs['Period']=='(unknown)'].index)

# Display all the unique Period names
dinosaurs.Period.unique()
```

```
# Country Column

# Getting all unique Locations
dinosaurs.Country.unique()

# Get the list of Countries belonging to a region
...
asia          = ['Pakistan','China & South Korea' ...]
descriptions  = ['Formerly referred to Phaedrolosaurus.', ... ]
multiple      = ['Portugal & Uzbekistan','Portugal & USA', ...]
...

# Replacing the values
...
dinosaurs['Country'] = dinosaurs['Country'].replace(africa, 'Africa')
dinosaurs['Country'] = dinosaurs['Country'].replace(asia, 'Asia')
dinosaurs['Country'] = dinosaurs['Country'].replace(descriptions, 'Asia')
...

# Change column name to "Regions"
dinosaurs.rename(columns = {'Country':'Region'}, inplace = True)

# Getting all unique Locations
dinosaurs.Region.unique()
```

The resulting dataset looked like this:

Name	Period	Diet	Region
Riojasaurus	Triassic	Herbivore	South America
Stegosaurus	Jurassic	Herbivore	North America
Velociraptor	Cretaceous	Carnivore	Asia

Datasets counting Species per Column

Some more specific datasets were produced from the Dinosaur List dataset. This required making use of grouping by and aggregation functionalities, with which all the species that belonged to a particular group could be counted.

```
# All of the sub-dataset followed a similar pattern

# Grouping Dinosaurs by Diet
dinosaurs_diet = dinosaurs.groupby('Diet')

# Aggregating the Count of Species
dinosaurs_diet_agg = dinosaurs_diet.agg({'count'})

# Change column name to "Number of Species"
dinosaurs_diet_agg = dinosaurs_diet_agg['Period']
dinosaurs_diet_agg.rename(columns = {'count':'Number of Species'}, inplace =
True)

# Displaying the periods
dinosaurs_diet_agg
```

The resulting tables looked like this:

Diet	Number of Species
Carnivore	352
Herbivore	720
Omnivore	70
Unknown	11

Paleobiology Database

Since both files came from the same source, the initial preparation was the same for both. This included skipping the first 15-20 rows, and in the case of the Occurrences dataset, set the low_memory flag to false.

```
# Loading the data
occurrences  = pd.read_csv('occurences.csv', skiprows=16, low_memory=False)
```

Diversity over Time

There were a number of preparations that had to be done to the diversity over time dataset. It is worth noting that there were no missing values, so they did not need to be dealt with. The preparations were the following:

- The only necessary columns were the name of the interval (*interval_name*), the start and ending of the interval (*max_ma* and *min_ma*) and the Number of Occurrences (*n_occs*).
- The columns were then renamed to more appropriate titles, being *"Interval Name"*, *"Start"*, *"End"* and *"Species Found"*.
- The Interval Name was set as the index.
- As the interest of this study lays on the Mesozoic era, it is then relevant to filter out all the periods outside this specific time frame. In order to do that, the previously scraped period data was used to keep only the intervals that occured between the beginning of the Triassic and the end of the Cretaceous.
- The intervals do not provide any information about the period they belong to, so a new column was created and then compared to the mesozoic periods dataset, and the overall period was added accordingly.
- Finally, the columns were ordered in a more sensible manner, to provide all the time-related information first, and then the number of species found from each interval.

```
# Keeping only necessary columns
diversity = diversity[['interval_name','max_ma','min_ma','n_occs']]

# Rename the Columns to more appropriate names
diversity = diversity.rename(columns = {...})

# Set the Interval name as the index
diversity = diversity.set_index('Interval Name')

# Only keep entries from the Mesozoic (Dinosaur times)
time_condition = (diversity['Start'] <= period_ranges.loc['Triassic']['Start'])
& (diversity['End'] >= period_ranges.loc['Cretaceous']['End'])
diversity = diversity[time_condition]

# Adding a column to specify the overall period it belongs to
diversity.loc[diversity['Start']>period_ranges.loc['Cretaceous']['End'],'Period
'] = "Cretaceous"
diversity.loc[diversity['Start']>period_ranges.loc['Jurassic']['End'],'Period']
= "Jurassic"
diversity.loc[diversity['Start']>period_ranges.loc['Triassic']['End'],'Period']
= "Triassic"

# Change the order of the columns
diversity = diversity[['Start','End','Period','Species Found']]
```

The resulting table looked like this:

Interval Name	Start	End	Period	Species Found
Valanginian	139.8	132.9	Cretaceous	872
Ladinian	242.0	237.0	Triassic	729
Callovian	166.1	163.5	Jurassic	759

Occurrences

Part of the preparation for the occurrences table was similar to the Diversity of the time. However, it required some specific actions:

- Only the necessary columns were kept, in this case "*accepted_name*", "*early_interval*", "*max_ma*", "*min_ma*" and "*accepted_rank*".
- There were a number of missing values on unnecessary columns, so no further action was required once they were removed.
- One of the reasons the dataset was so large was because it did not only show species, but also genera, families and even kingdoms. Because of this, it was necessary to filter it only to species (e.g dogs, not canids). Once the species and subspecies had been filtered, the accepted rank column was deleted.
- The columns were renamed to more appropriate terms, such as the *max_ma* and *min_ma* to the earliest and latest time in which the species lived, or *accepted_name* to the Species name.
- For further ease of use, only the first name of the species was kept.
- The Species Name column was then set as the index, to identify each occurence.
- Just like it was done with the Diversity over time dataset, only the species from the Mesozoic era were kept by using the limits provided by the web scraped dataset of Mesozoic periods.
- And again just like in the Diversity over time dataset, another column was added to show the overall period the specimen lived in.

Additionally, a secondary dataset was produced to count the number of specimens of a single species. This was done in the following manner:

- Firstly, the table was grouped by the species name, and a new column was created, showing the count of specimens.
- Secondly, the columns from the parent dataset were added to the new one, by adding the information of the first entry of each specimen.
- Finally, the columns were arranged to show the occurrences at the end.

It is worth mentioning that once the dataset was cleaned, the size of it went from almost 400,00 rows to 3,358.

```
# Keeping only necessary columns
occurrences = occurrences[['accepted_name','early_interval',
'max_ma','min_ma','accepted_rank']]

# Only keep species, then delete the rank column
occurrences = occurrences[(occurrences['accepted_rank'] == 'species') |
(occurrences['accepted_rank'] == 'subspecies')]
occurrences = occurrences.drop(columns=['accepted_rank'])

# Rename the Columns to more appropriate names
occurrences = occurrences.rename(columns = { ... })

# Get rid of empty species names and keep first name
occurrences = occurrences.dropna()
occurrences['Species Name'] = occurrences['Species Name'].str.split(' ').str[0]

# Set the species name as the index
occurrences = occurences.set_index('Species Name')

# Only keep entries from the Mesozoic (Dinosaur times)
time_condition =
(occurrences['Earliest']<period_ranges.loc['Triassic']['Start']) &
(occurrences['Latest']>period_ranges.loc['Cretaceous']['End'])
occurrences = occurrences[time_condition]

# Adding a column to specify the overall period it belongs to
occurrences.loc[occurrences['Earliest']>period_ranges.loc['Cretaceous']['End'],
'Period'] = "Cretaceous"
occurrences.loc[occurrences['Earliest']>period_ranges.loc['Jurassic']['End'],'P
eriod'] = "Jurassic"
occurrences.loc[occurrences['Earliest']>period_ranges.loc['Triassic']['End'],'P
eriod'] = "Triassic"

# Grouping and aggregating Dinosaurs by period
occurences_grouped = occurrences.group by('Species Name')
occurences_grouped_agg = occurences_grouped.agg({'count'})
occurences_grouped_agg = occurences_grouped_agg['Interval']

# Adding the other columns
occurences_grouped_agg['Earliest'] =
occurrences[~occurrences.index.duplicated(keep='first')]['Earliest']
...

# Change the order of the columns
occurences_grouped_agg = occurences_grouped_agg[[ ... ,'Occurrences']]
```

The resulting table looked like this:

Species name	Start	End	Interval	Period	Occurrences
Coelophysis	228	208.5	Norian	Triassic	9
Tyrannosaurus	83.5	66	Campanian	Cretaceous	2
Stegosaurus	157.3	145	Kimmeridgian	Jurassic	12

Visualization Plan

As it has been previously explained, the idea behind this infographic is to provide a new perspective on dinosaurs and to show what science and research has allowed us to know about them. The infographic is based on 3 topics:

- The first and main topic is the “When”. It aims to show when exactly dinosaurs lived, for how long and how their diversity changed over time.
- The second topic is where dinosaurs have been found, how many of them and what areas contain the most dinosaurs.
- The third topic is focused on what they ate, since pop culture has shown an image of dinosaurs as monsters, rather than animals just like the ones that exist currently.

There have been design considerations, such as the limit of palette to shades of one tonality to avoid confusion for colourblind individuals on colour-based visualizations. The details have been kept to a minimum not to make the infographic too difficult to observe or read, since Paleontology is a field that contains a great number of technical names in Latin or Greek. The choice of visual-oriented visualizations is an example of this, as simply showing numbers and tables would reduce the appeal of the infographic. Finally, up-to-date paleoart has been used to depict some of the mentioned dinosaurs, to keep the infographic engaging and interesting to look at, as well as offering an updated view on species whose resemblance has changed over time.

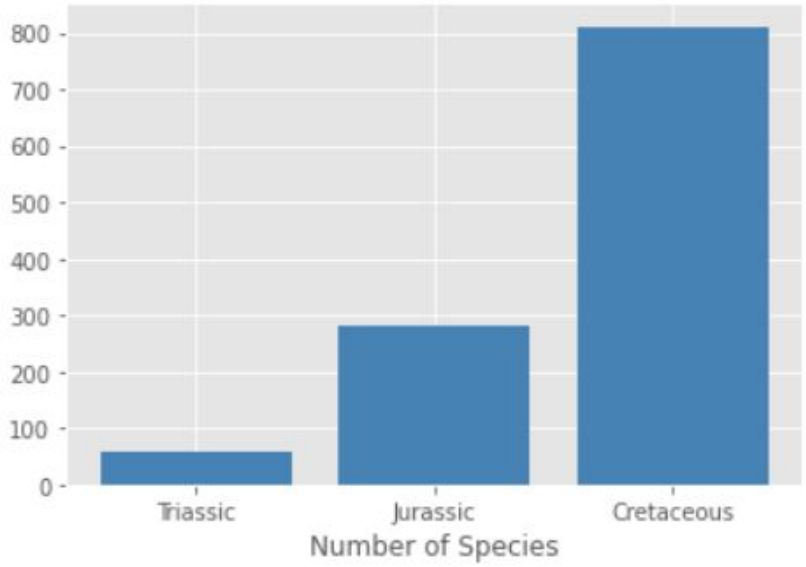
Time-based Visualizations

Species per Period

The aim of this visualization is to show the change of dinosaur diversity through time. Even though the Mesozoic was the “Age of Dinosaurs”, during most of the Triassic they represented a minority in relation to other types of animals living at the time, but slowly thrived until the Cretaceous, where the biggest number of species are found.

For this visualization it is required to use the Dinosaur List dataset, and specifically the species per period that was produced by using the group by functionality. From this dataset, both the period name and species count number are used.

This change can be represented using both a bar graph, to represent the number of species for each period, or using a more visual representation of a timeline, showing the number of species per period.

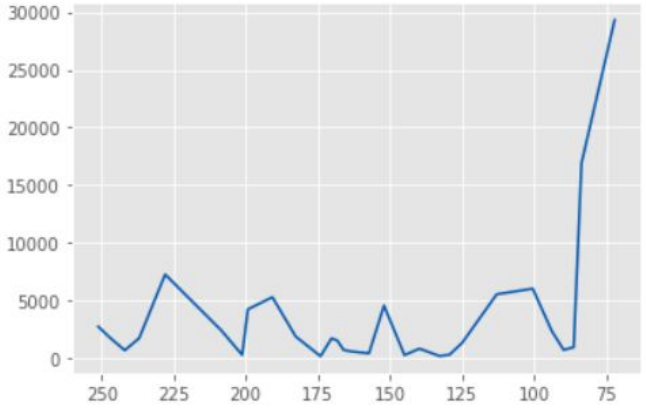


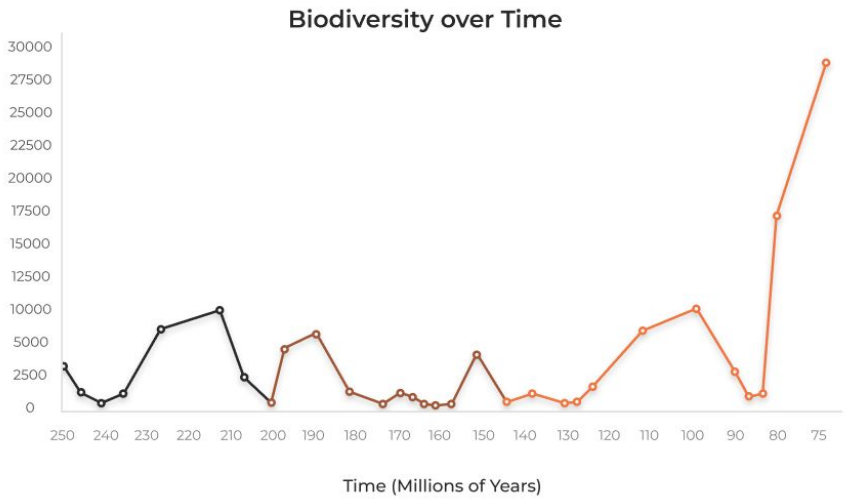
Diversity over Time

This visualization is meant to represent biodiversity over time. As mentioned in the previous point, dinosaurs were not the only group on earth, and larger-scale diversity can be observed to increase and decrease, due to shifts in the earth's plate tectonics that altered ecosystems, or extinctions that vanished large percentages of earth's life.

In order to show this change of diversity, the "diversity over time" dataset will be used. In it, the start of the interval and number of species columns will be used. It is not necessary to use the end of the interval since it coincides with the start of the next one, and the interval name doesn't provide any meaningful information (And it may actually confuse the reader).

A line graph is the best approach for this visualization, since it can easily display the number of species in a certain interval in the Y axis, and a precise location of the interval's start on the X location.





Timeline of Species

The last time-based visualization will be a timeline of the mesozoic periods, with specific, well-known dinosaurs popping out of the timeline, in order to show how many specimens have been found, and the distance in time between some species and humans, to show that 300,00 years of human evolution is nothing compared to 200,000,000 years of dinosaur evolution.

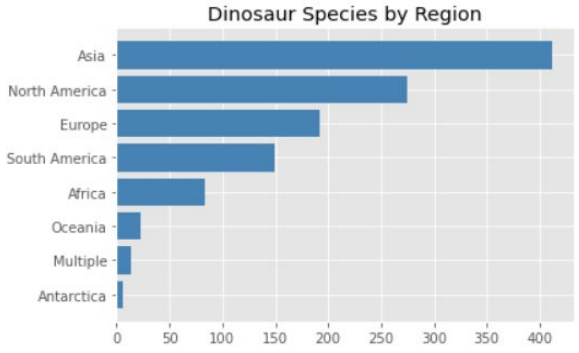
To display this information two datasets will be necessary, the web-scraped mesozoic table to render the time space (Using the period name, start and end), and the occurrences dataset to retrieve information about specific dinosaur species, using the species name, start interval and number of occurrences. As this is a purely visual representation, no code-based visualization has been provided.

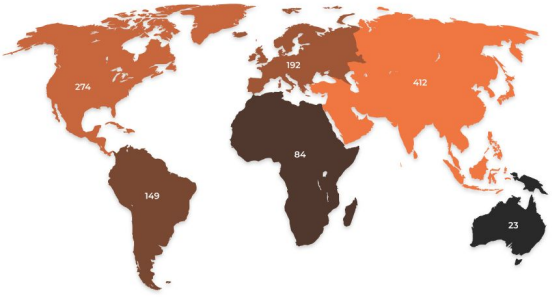
Location-based Visualizations

The intention behind this visualization is to show the distribution of dinosaur species over the globe, to give an idea about how much paleontologists have discovered, and find that some areas in the world still need to be explored.

For this, the dinosaur list dataset will be used, and specifically the subdataset produced by grouping species by the location. From this dataset, both the location name and number of species columns will be used.

A bar chart could be used to represent every region and the species number, but a map can be used to more accurately provide a sense of location, with colours shades going from darker to brighter as more species have been found in a particular region, and numbers indicating how many species have been found there.



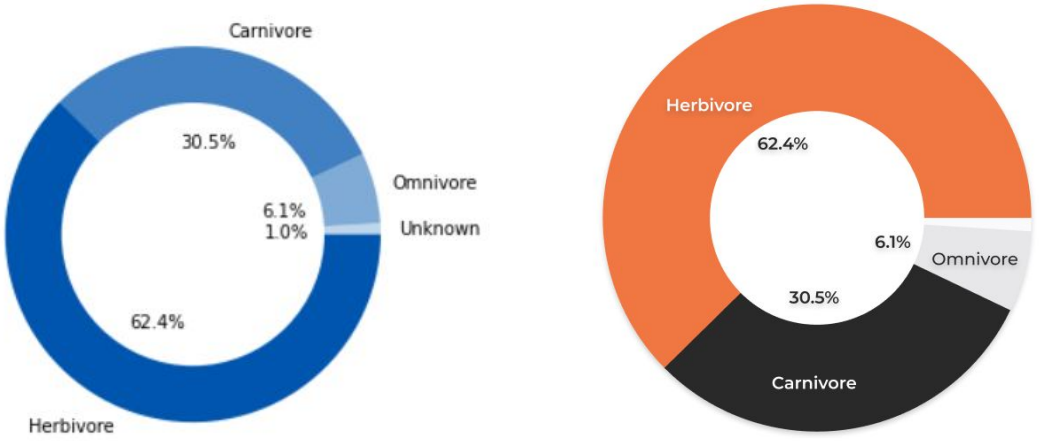


Diet demystification

The aim of this visualization is to show the diets of discovered dinosaurs, in order to show that the hypercarnivore hunters were not the norm, but a relative minority (Just like just relatively few carnivores comprise a natural ecosystem).

For this, the dinosaur list dataset will be used, and specifically the subdataset produced by grouping species by their diet. From this dataset, both the type of diet and number of species columns will be used.

As the visualization is focused on showing percentages of species, relative to the total, a pie chart will be used, where the type of diet and its percentage will be shown.



Infographic

The infographic can be viewed from the following URL on Figma:
https://www.figma.com/file/nvluSWi86jdBSdlZkvuCnP/Data-Viz?node-id=58%3A182

The file has been made sure to be visible by any user that has access to the link, even without an account on the platform. Once it is opened the “hand” tool should be selected. The main controls to observe the canvas are:

- By holding click and moving the mouse the canvas should move around.
- By holding the control key and using the scroll wheel, the canvas will zoom in and out.

Data Exploration Report

Along with this report, a Python Notebook exported to Html has been included, where all the data exploration can be observed, including the code written to achieve the visualizations and its output, in-depth explanation on the data exploration and processing and how the datasets were cleaned and modified for its proper use.

References

[0] - PBS Eons, 2020. *An Illustrated History Of Dinosaurs*. [video] Available at: <https://www.youtube.com/watch?v=JDnQmBFxIfE> [Accessed 3 December 2020].

[1] - Kumazaki, 2020. *Dinosaur List*. [online] Available at: <https://www.kaggle.com/kumazaki98/dinosaur-list> [Accessed 3 December 2020].

[2] - Orellano, J., 2020. *The Paleobiology Database*. [online] Kaggle.com. Available at: <https://www.kaggle.com/apartmentguru/dino-crisis> [Accessed 3 December 2020].

[3] - En.wikipedia.org. 2020. *Mesozoic*. [online] Available at: <https://en.wikipedia.org/wiki/Mesozoic> [Accessed 3 December 2020].

[4] - Crummy.com. 2020. *Beautiful Soup Documentation — Beautiful Soup 4.9.0 Documentation*. [online] Available at: <https://www.crummy.com/software/BeautifulSoup/bs4/doc/> [Accessed 3 December 2020].

[5] - Magazine. 2020. *See How We'Re Reimagining Dinosaurs In Today'S 'Golden Age' Of Paleontology*. [online] Available at: <https://www.nationalgeographic.com/magazine/2020/10/reimagining-dinosaurs-prehistoric-icons-get-a-modern-reboot-interactive-feature/> [Accessed 3 December 2020].