First Homework report

Andrey Donetskov

October 13, 2025

In this report, I will discuss two files: **statistics.ipynb** and **visualization.ipynb**. Both files relate to the **First Mandatory Homework** and cover **wrecked data** statistical analysis and corresponding visualizations.

1 Visual analysis

Here I trided my best to visualize and clean [1]. Below you can see what it contains:

	Animal type	Country	Weight kg	Body Length cm	Gender
0	–	–	–	–	–
1	–	–	–	–	–
2	European bison	Poland	930.000	335.0	male
3	European bison	Poland	909.000	311.0	not determined
4	European bison™	Poland	581.000	277.0	female
⋮	⋮	⋮	⋮	⋮	⋮
1006	red squirrel	Poland	0.346	20.0	female
1007	hedgehog	Germany	1.000	23.0	female
1008	hedgehog	Germany	0.500	17.0	female
1009	red squirrel	Poland	0.346	20.0	female
1010	hedgehog	Hungary	0.900	16.0	female

Table 1: Animal observation **wrecked data** part 1

	Animal code	Latitude	Longitude	Animal name	Observation date	Data compiled by
0	–	–	–	–	03.01.2024	James Johnson
1	–	–	–	–	03.02.2024	James Johnson
2	–	52.828845	23.820144	Szefu	01.03.2024	Anne Anthony
3	–	52.830509	23.826849	–	01.03.2024	Anne Anthony
4	–	52.834109	23.807093	–	01.03.2024	Anne Anthony
⋮	⋮	⋮	⋮	⋮	⋮	⋮
1006	–	52.212001	21.033187	Lola	7 May 2024	Anne Anthony
1007	–	49.561356	11.105334	–	7 May 2024	Bob Bobson
1008	–	49.561569	11.087046	–	7 May 2024	Bob Bobson
1009	–	52.212001	21.033187	Lola	7 May 2024	Anne Anthony
1010	–	47.509860	18.939943	–	8 May 2024	Anne Anthony

Table 2: Animal observation **wrecked data** part 2

Yep. It pretty cursed. I cleaned it:

1. Dropped rudimet columns.
2. Dropped lines with NaNs.

3. Stardartized objects manually renaming whem.

4. Multipy negative values by -1.

After this I visualized everithing.

1.1 Count plots

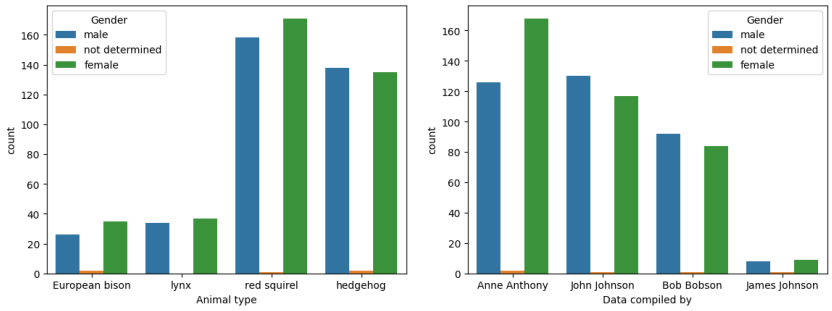

Figure 1: Count plots of animal type and of data contibutor by gender.

On 1 gender distribution is normal but anumal type is not as well as data contibuting. Worth to take a closer look at James data.

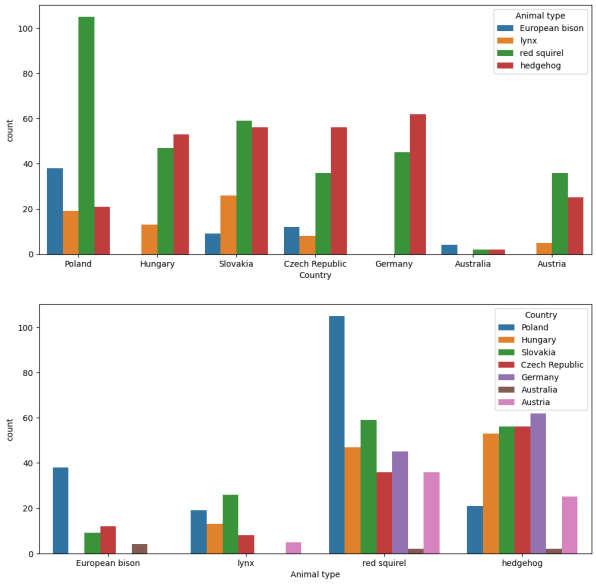

Figure 2: Count plot of animal type by country and vise versa.

On 2 plots I note nothing except that European bison lives in Australia.

1.2 Line plots

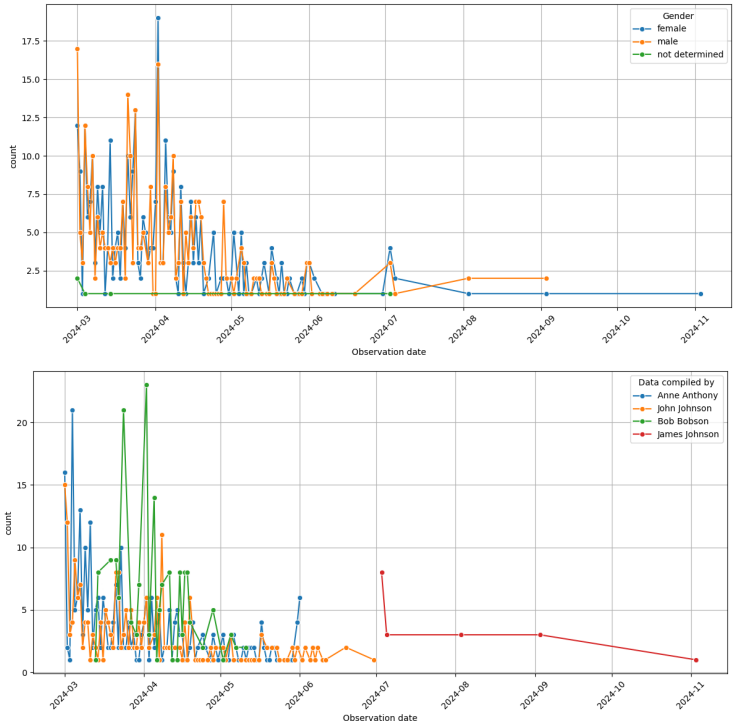

Figure 3: Line plots of observation date by gender and by data contibutor

Rivising 3:

1. All data impliments in 2024.
2. James somewhat stanging out even more.

1.3 Hist and box plots

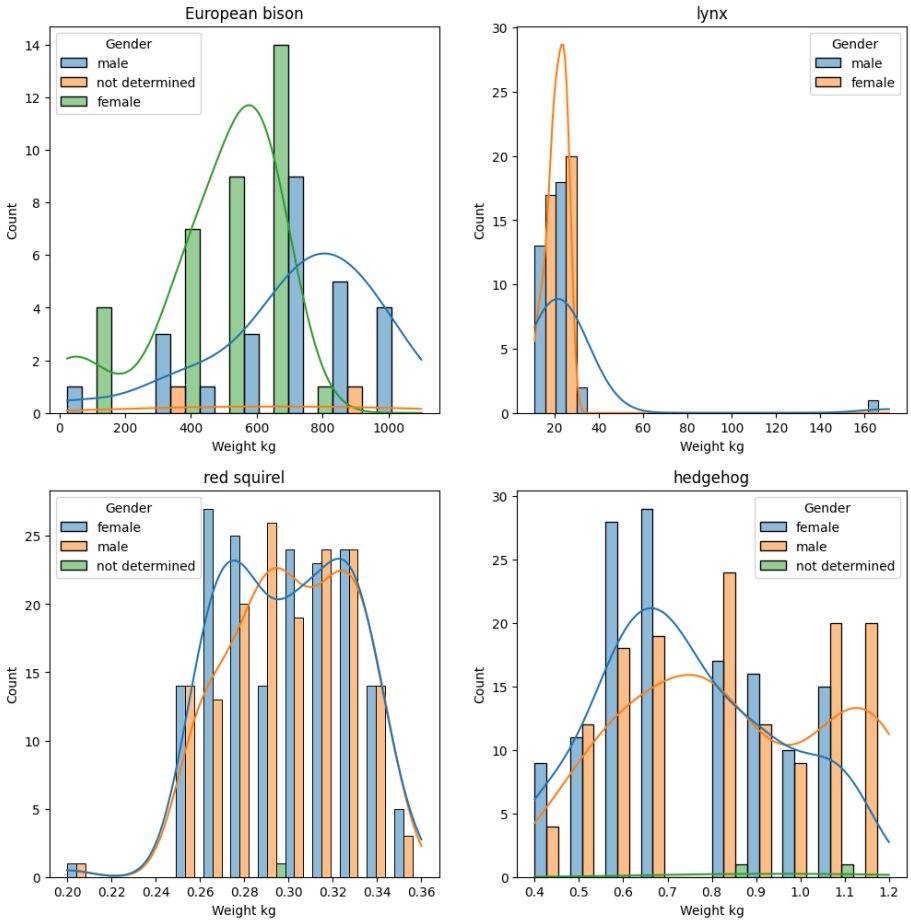

Figure 4: Distribution of weight by animal type

On 4 nothing special except a very heavy lynx. No. A VERY VERY **heavy** one.

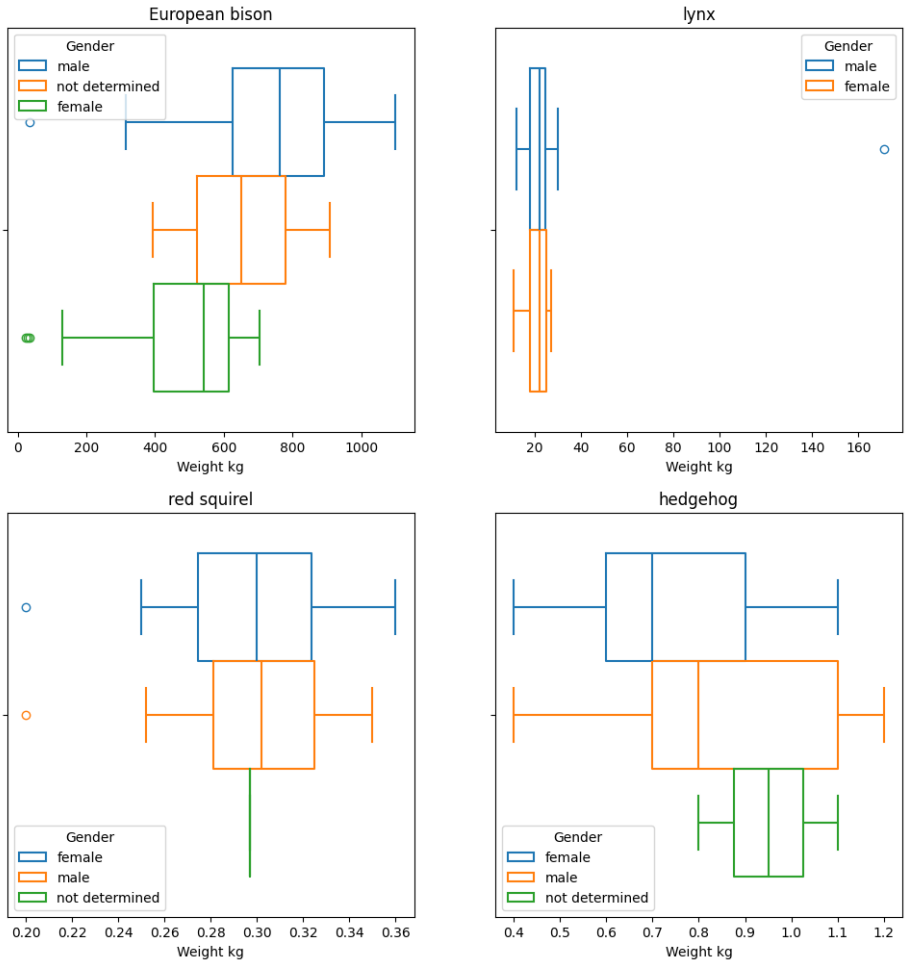

Figure 5: Bot plots of weight by animal type

Here(5) the same lynx shine bright.

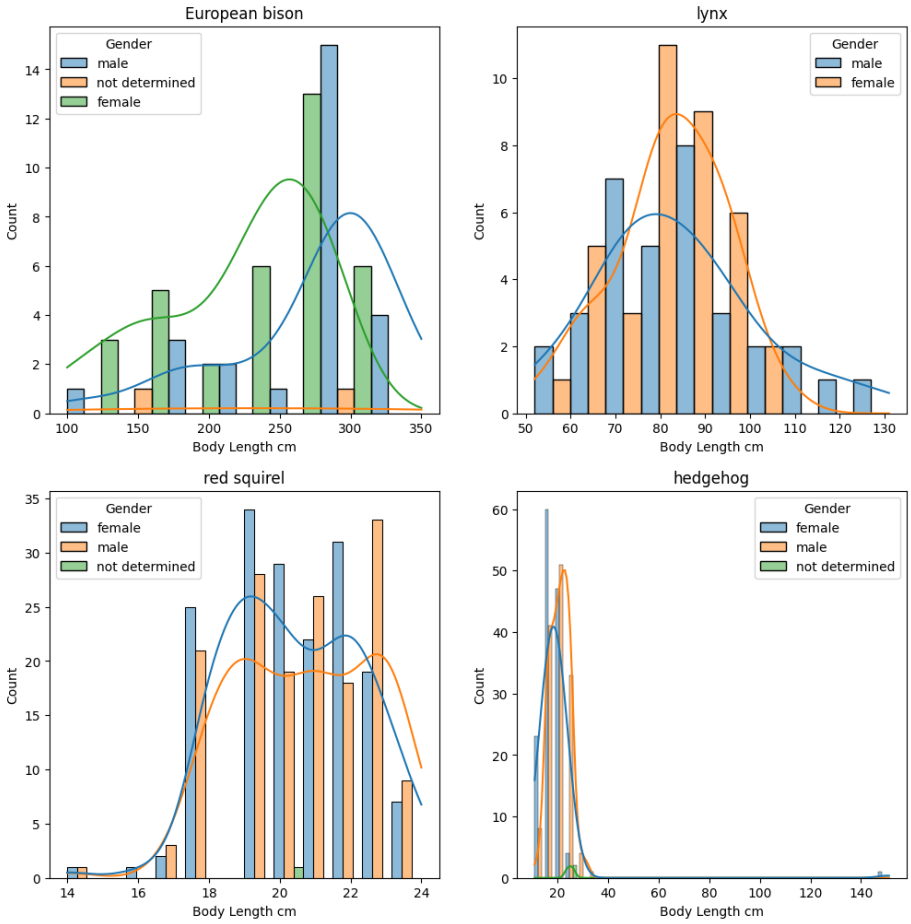

Figure 6: Distribution of lenght by animal type

As 4 on 6 I notice mutative animal. Not lynx, but a new gigantic hedgehogs 150 cm long. What a monsters... I guess it a new spicies bred by creazy Russian scientists.

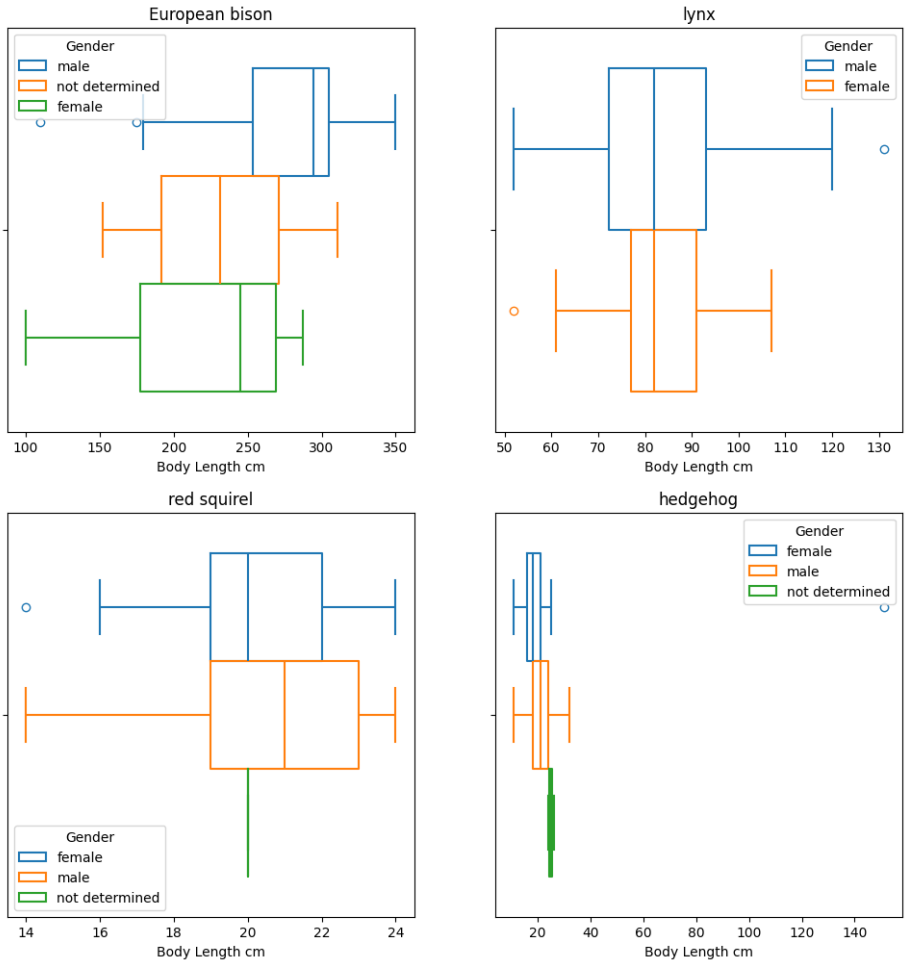

Figure 7: Bot plots of lenght by animal type

At this moment I have no idea why I must use box plots. They're the same as hist plots or even less intuitive.

1.4 Pair plot

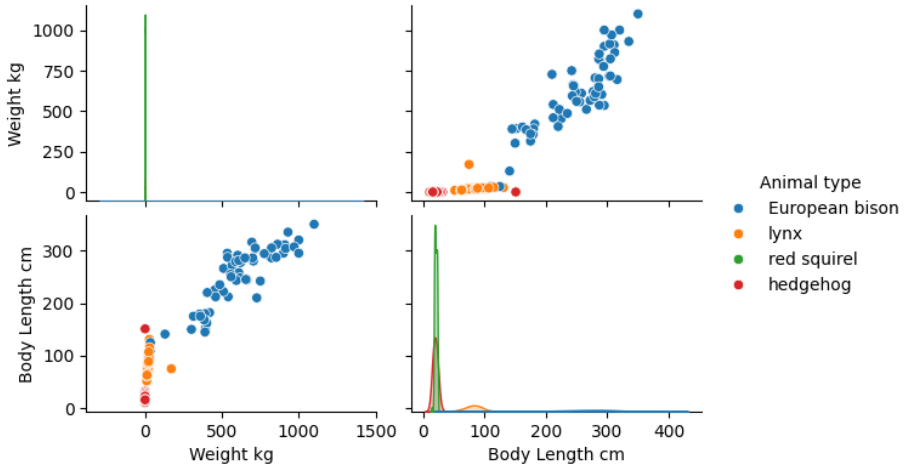

Figure 8: Pair plot of lenght and weight

Dataset contains 4 decimal parametrs:

1. Weight (kg)
2. Lenght (kg)
3. Latitude
4. Longitude

I create a pairplot of only two weight and lenghtas most interesting. It rudiment to weight to lenght *"simple"* plot. Let's see it too:

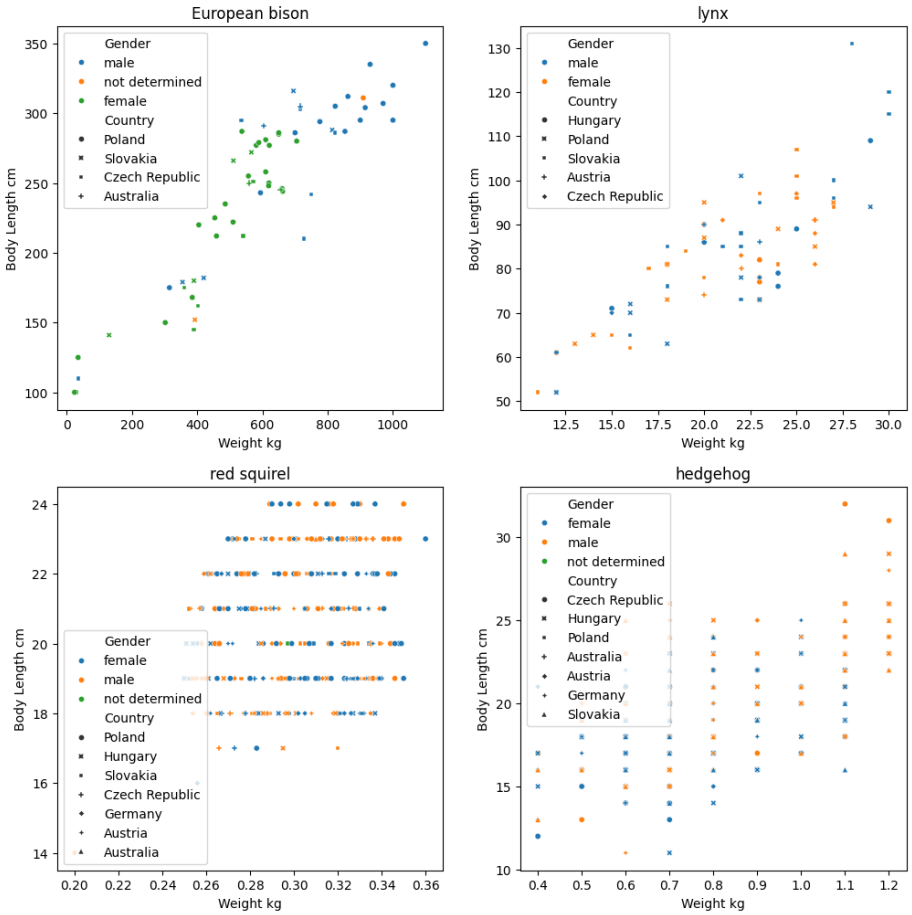

Figure 9: Scatter plot of lenght and weight

We can see some correlation, but I cannot say anything before I'd see correlation coefficients.

1.5 Geo-visualization

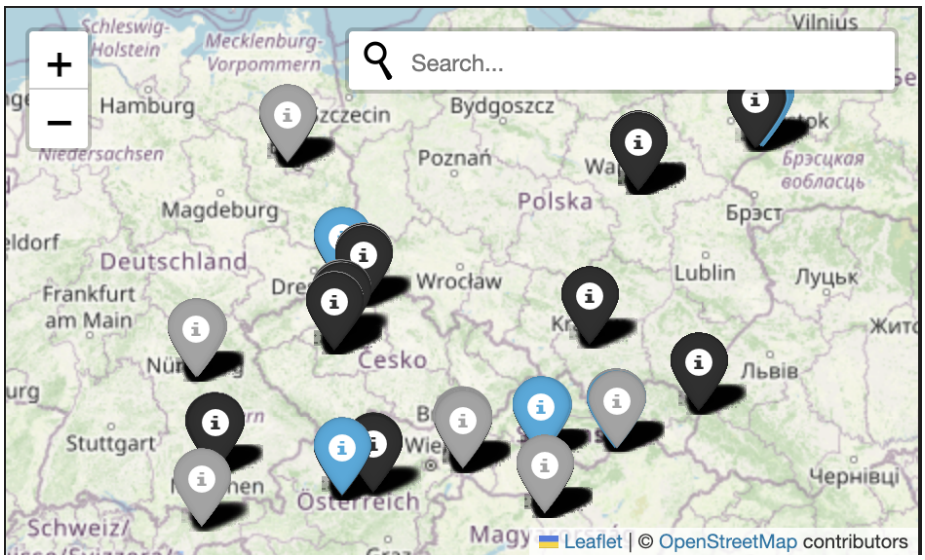

Figure 10: Data scattering

In this section, I want somehow visualize 2 more columns from dataset. Here some clever way of geo-visualization.

2 Statistical analysis

First of all, I used *pandas.DataFrame.describe()* option to view min, mean, max and quartiles of **wrecked data**.

Statistic	Weight (kg)	Body Length (cm)	Animal code	Latitude	Longitude
count	984.000000	984.000000	0.0	913.000000	913.000000
mean	39.745503	39.107724	NaN	49.393369	18.203280
std	156.290076	58.628601	NaN	7.168900	3.899601
min	-0.252000	-19.000000	NaN	-78.582973	11.074008
25%	0.293000	19.000000	NaN	48.186913	14.384559
50%	0.331500	21.000000	NaN	49.560723	18.944015
75%	0.800000	23.000000	NaN	52.212433	21.033243
max	1100.000000	350.000000	NaN	52.853843	34.896734

Table 3: Descriptive statistics for **wrecked data**

Right now I greatly sorrow, that I'd thought of chosing [1]. It was a great mistake.

Hence of this, in this section if I didn't tell another, I used **cleaned data** in order of simplicity. It look like this:

Statistic	Weight (kg)	Body length (cm)	Latitude	Longitude
count	833.000000	833.000000	750.000000	750.000000
mean	46.720602	42.501043	49.743041	17.595264
std	168.843218	62.879887	1.851982	3.881169
min	0.200000	11.000000	47.316383	11.074008
25%	0.297367	19.000000	48.186430	14.348471
50%	0.349000	21.000000	49.545885	18.847488
75%	1.000000	23.000000	52.211980	21.031909
max	1100.000000	350.000000	52.853843	23.919668

Table 4: Descriptive statistics for **cleaned data**

2.1 Correlation

Below 11 I pin correlation maps. On this I can notice very strong normal correlation between weight and lenght. Also Kendall correlation didn't show anithing. It's strange. Let's try view correlation on diffetent animal types.

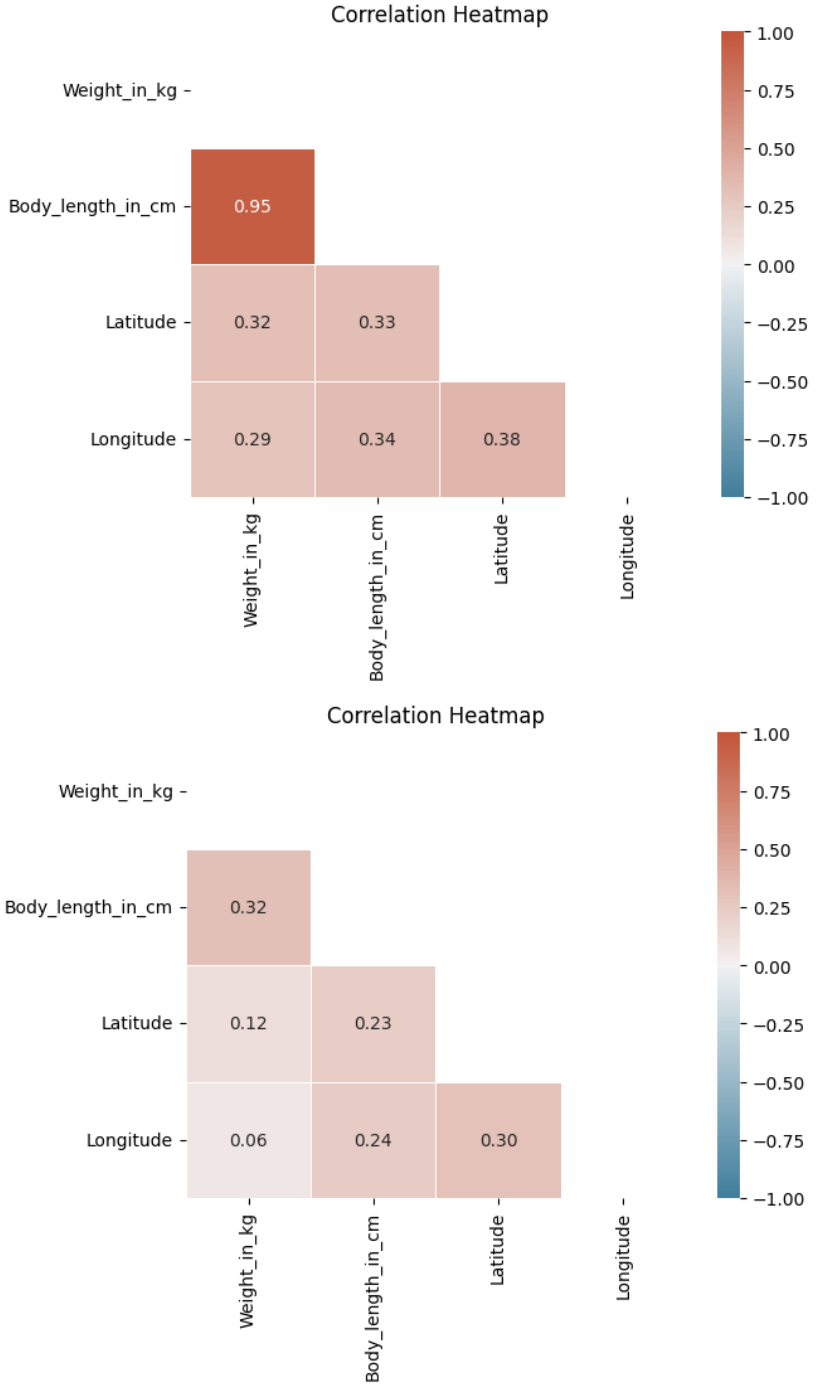

Figure 11: Normal and Kendall correlation

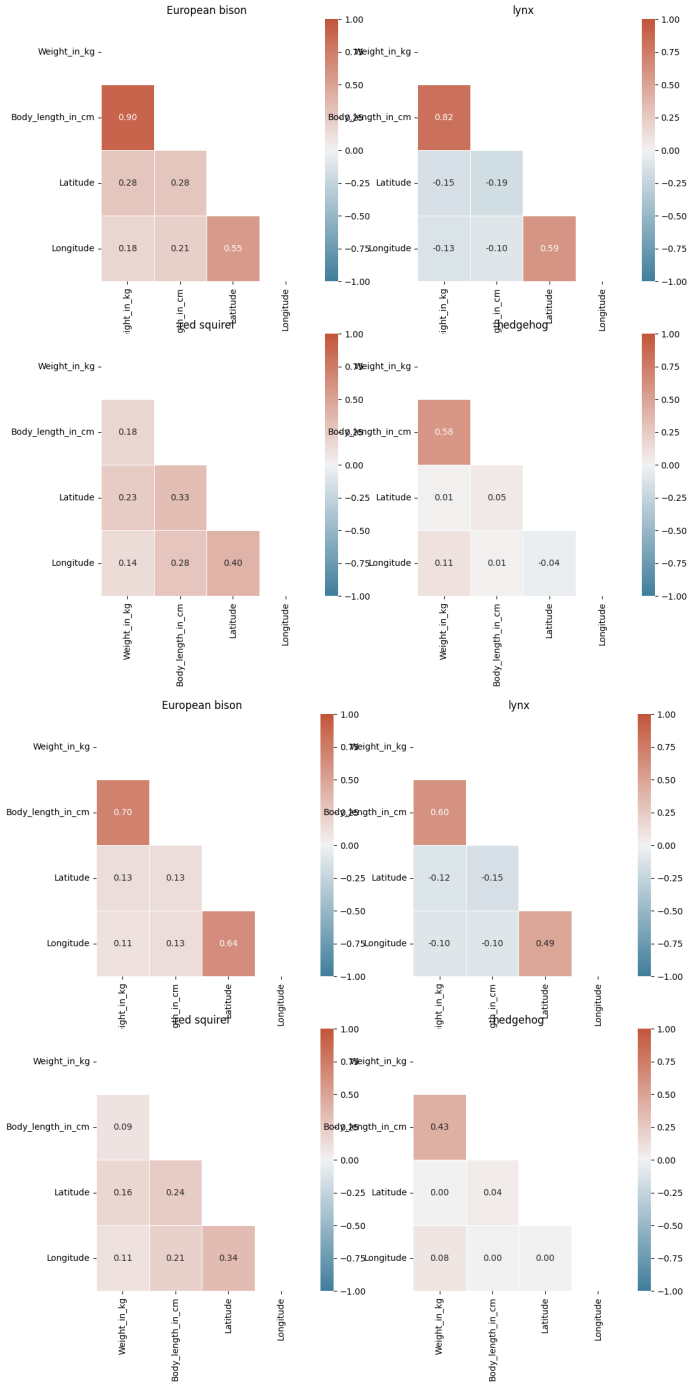

Figure 12: Normal and Kendall correlation

On 12 4 normal correlation heat maps and 4 Kendall correlation maps. As I can notice This data is more optimistic. However

$$E[Bison(kg)] = 0.9 * E[Bison(cm)]$$

feels as well strange. Although, as expected, latitude and longitude didn't correlate with anithing much. Let's move on with this wonderful quote:

"Correlation does not imply causation"

2.2 Is lynx? lynx?

When I wrote *dirty_data["Animal_type"].unique()*, I discovered class lynx?:

```
array([nan, 'European bison', 'European bison™', 'European bisson',
       'European buster', 'lynx', 'lynx?', 'red squirel', 'red squirrel',
       'red squirrell', 'hedgehog', 'wedgehod', 'ledgehod'], dtype=object)
```

I found question *"Is lynx? lynx?"* interesting to me. Let's find it out!

First of all I tryed to use t-values:

1. For **body length in cm**:

   $$t - statistic : -0.5793803445653539, p - value : 0.5642194350169902$$

2. For **weight in kg**:

   $$t - statistic : 0.28872058707288334, p - value : 0.7736608274782608$$

For both cases $p - value >> 0.05$, this mean significant differences in classes. I'd already lost trast in statistic...

References

[1] Dirty data to clean What's wrong with this dataset.