Repeat Programming Assignment

Wing shape variation associated with mimicry in butterflies

```
# import modules
from IPython.display import display
import pandas as pd
import os
import re
import seaborn as sns
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
pd.set_option("display.max_rows", 999)
pd.set_option('max_colwidth',100)
```

```
filepath = '/content/drive/MyDrive/WING.xlsx'
# Load file with `sheet_name=None` - returns a dictionary
df_dict = pd.read_excel(filepath, sheet_name=None)
```

Saved successfully!

Experimental brood cleaning

```
# Get data from  worksheet
df_brood = df_dict.get('Experimental brood')
df_brood = df_brood.drop(df_brood.index[82:90])
# Preview
df_brood.head()
```

	name	sex	genotype	phenotype
0	725	female	a-e	elegans
1	726	female	a-s	aurora

```
df_brood.tail()
```

	name	sex	genotype	phenotype
77	1029	male	s-s	silvana
78	1030	male	e-s	elegans
79	1036	male	a-s	aurora
80	1037	male	a-e	elegans
81	1052	male	a-e	elegans

```
df_brood['sex'].replace(['male','female'],['M','F'],inplace=True)
df_brood.head()
```

	name	sex	genotype	phenotype
0	725	F	a-e	elegans
1	726	F	a-s	aurora
2	735	F	s-s	silvana
3	736	F	a-e	elegans
				gans

Saved successfully!

Chnaging the name to id column

Information about the data types

```
df_brood.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 82 entries, 0 to 81
Data columns (total 4 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   name       82 non-null     object
 1   sex        82 non-null     object
 2   genotype   82 non-null     object
```

```
 3   phenotype  82 non-null     object
dtypes: object(4)
memory usage: 3.2+ KB
```

converting datatypes

```
df_brood['sex'] = df_brood['sex'].astype('category')
df_brood['genotype'] = df_brood['genotype'].astype('category')
df_brood['phenotype'] = df_brood['phenotype'].astype('category')
df_brood.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 82 entries, 0 to 81
Data columns (total 4 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   name       82 non-null     object
 1   sex        82 non-null     category
 2   genotype   82 non-null     category
 3   phenotype  82 non-null     category
dtypes: category(3), object(1)
memory usage: 2.0+ KB
```

Converting the required data types

```
df_brood.isnull().sum().sum()
```

```
0
```

Replacing the name by id

Saved successfully!

```
df_brood.rename( {'name ':'id' },axis=1,inplace=True)
```

```
list(df_brood.columns)
```

```
['id', 'sex', 'genotype', 'phenotype']
```

```
df_brood.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 82 entries, 0 to 81
Data columns (total 4 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   id         82 non-null     object
```

```
 1   sex        82 non-null     category
 2   genotype   82 non-null     category
 3   phenotype  82 non-null     category
dtypes: category(3), object(1)
memory usage: 2.0+ KB
```

```
df_brood['id'] = df_brood['id'].astype('int')
```

```
df_brood.head()
```

	id	sex	genotype	phenotype
0	725	F	a-e	elegans
1	726	F	a-s	aurora
2	735	F	s-s	silvana
3	736	F	a-e	elegans
4	738	M	a-e	elegans

Axis 1 measures Cleaning

```
# Get data from worksheet
df_Measures1 = df_dict.get('Axis 1 measures')
df_Measures1 = df_Measures1.drop(df_Measures1.index[164:175])
```

```
# Preview the first 5 Rows
```

Saved successfully!

	name	sex	genotype	area	ellipse axe 1 length	max axe 1 length	deviation max axe 1 from	axe 1 Moment 2	Mom
0	0725-	F	a-e	954465.0	1622.1291	1674.0	69.7104	9.677134e+11	1.166094

```
df_Measures1.rename( {'name':'id' }, axis=1,inplace = True)
```

```
df_Measures1['id'] = df_Measures1.id.str.extract('(\d+)')
df_Measures1.head()
```

	id	sex	genotype	area	ellipse axe 1 length	max axe 1 length	deviation max axe 1 from center	axe 1 Moment 2	axe Moment
0	0725	F	a-e	954465.0	1622.1291	1674.0	69.7104	9.677134e+11	1.166094e+
1	0725	F	a-e	638157.0	1120.3231	1100.0	-59.4572	2.490248e+11	1.951945e+
2	0726	F	a-s	984051.0	1673.3414	1741.0	89.1505	1.069848e+12	1.336034e+
3	0726	F	a-s	634452.0	1145.1990	1138.0	50.8133	2.677018e+11	2.165425e+
4	0735	f	s-s	889416.0	1602.0627	1649.0	-123.5292	6.447254e+11	7.112162e+

Removing of the Characters from ID column and making the Sex column to upper case

```
df_Measures1['sex'] = df_Measures1['sex'].str.upper()
df_Measures1.head()
```

Saved successfully!

	id	sex	genotype	area	ellipse axe 1 length	max axe 1 length	deviation max axe 1 from center	axe 1 Moment 2	axe Moment
0	0725	F	a-e	954465.0	1622.1291	1674.0	69.7104	9.677134e+11	1.166094e+
1	0725	F	a-e	638157.0	1120.3231	1100.0	-59.4572	2.490248e+11	1.951945e+
2	0726	F	a-s	984051.0	1673.3414	1741.0	89.1505	1.069848e+12	1.336034e+
3	0726	F	a-s	634452.0	1145.1990	1138.0	50.8133	2.677018e+11	2.165425e+
4	0735	F	s-s	889416.0	1602.0627	1649.0	-123.5292	6.447254e+11	7.112162e+

```
df_Measures1.tail()
```

	id	sex	genotype	area	ellipse axe 1 length	max axe 1 length	deviation max axe 1 from center	axe 1 Moment 2	Mom
159	1036	M	a-s	725224.0	1132.4175	1116.0	-98.6783	2.926688e+11	2.32635
160	1037	M	a-e	1133782.0	1735.2357	1775.0	-105.4738	9.725906e+11	1.16342
161	1037	M	a-e	790356.0	1190.8893	1183.0	-85.1766	3.614237e+11	3.04756
162	1052	M	a-e	1151146.0	1776.4828	1821.0	-128.2016	1.049082e+12	1.28977
163	1052	M	a-e	800738.0	1192.3945	1185.0	-31.9997	3.603016e+11	3.02450

```
df_Measures1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 164 entries, 0 to 163
Data columns (total 13 columns):
 #   Column                           Non-Null Count  Dtype
---  ------                           --------------  -----
 0   id                               164 non-null    object
 1   sex                              164 non-null    object
 2   genotype                         164 non-null    object
 3   area                             164 non-null    float64
 4   ellipse axe 1 length             164 non-null    float64
 5   max axe 1 length                 164 non-null    float64
 6   deviation max axe 1 from center  164 non-null    float64
 7   axe 1 Moment 2                   164 non-null    float64
 8   axe 1 Moment 3                   164 non-null    float64
 9   axe 1 Moment 4                   164 non-null    float64
 10  std axe 1 Moment 2               164 non-null    float64
 [illegible]                          164 non-null    float64
 [illegible]                          164 non-null    float64
memory usage: 17.9+ KB
```

Saved successfully!

```
df_Measures1['id'] = df_Measures1['id'].astype('int')
df_Measures1['sex'] = df_Measures1['sex'].astype('category')
df_Measures1['genotype'] = df_Measures1['genotype'].astype('category')
df_Measures1['max axe 1 length'] = df_Measures1['max axe 1 length'].astype('int')
df_Measures1['area'] = df_Measures1['area'].astype('int')
df_Measures1['axe 1 Moment 2'] = df_Measures1['axe 1 Moment 2'].astype('int')
df_Measures1['axe 1 Moment 3'] = df_Measures1['axe 1 Moment 3'].astype('int')
df_Measures1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 164 entries, 0 to 163
Data columns (total 13 columns):
 #   Column                           Non-Null Count  Dtype
---  ------                           --------------  -----
```

```
 0   id                              164 non-null    int64
 1   sex                             164 non-null    category
 2   genotype                        164 non-null    category
 3   area                            164 non-null    int64
 4   ellipse axe 1 length            164 non-null    float64
 5   max axe 1 length                164 non-null    int64
 6   deviation max axe 1 from center 164 non-null    float64
 7   axe 1 Moment 2                  164 non-null    int64
 8   axe 1 Moment 3                  164 non-null    int64
 9   axe 1 Moment 4                  164 non-null    float64
 10  std axe 1 Moment 2              164 non-null    float64
 11  std axe 1 Moment 3              164 non-null    float64
 12  std axe 1 Moment 4              164 non-null    float64
dtypes: category(2), float64(6), int64(5)
memory usage: 16.0 KB
```

Changing the variable datatype

```
df_Measures1.head()
```

	id	sex	genotype	area	ellipse axe 1 length	max axe 1 length	deviation max axe 1 from center	axe 1 Moment 2	axe 1 Mo
0	725	F	a-e	954465	1622.1291	1674	69.7104	967713404464	1166093876
1	725	F	a-e	638157	1120.3231	1100	-59.4572	249024797899	195194465
2	726	F	a-s	984051	1673.3414	1741	89.1505	1069847709180	1336034050
3	726	F	a-s	634452	1145.1990	1138	50.8133	267701750118	216542452
4	735	F	s-s	889416	1602.0627	1649	-123.5292	644725446306	711216237

Saved successfully!

```
df_Measures1.isnull().sum().sum()
```

```
0
```

Checking for Null values

Axis 2 measures Cleaning

```
df_Measures2 = df_dict.get('Axis 2 measures')
df_Measures2 = df_Measures2.drop(df_Measures2.index[164:170])
# Preview
df_Measures2.head()
```

	name	name.1	area	ellipse axe 2 length	max axe 2 length	deviation max axe 2 from center	axe 2 Moment 2	axe Moment
0	MJ02.0725-d_Ant_d	0725-d_Ant_d	954465.0	777.8875	844.0	-25.2980	1.791255e+11	9.815703e+1
1	MJ02.0725-d_Post_g	0725-d_Post_g	638157.0	729.2658	731.0	-95.3215	9.913302e+10	4.978770e+1
2	MJ02.0726-d_Ant_d	0726-d_Ant_d	984051.0	774.3652	844.0	-39.1989	1.782607e+11	9.668417e+1
3	MJ02.0726-d_Post_d	0726-d_Post_d	634452.0	710.0563	718.0	-108.5135	9.449492e+10	4.642871e+1
4	MJ02.0735-d_Ant_g	0735-d_Ant_g	889416.0	732.2889	801.0	-44.8532	1.446070e+11	7.412817e+1

Saved successfully!

Get data from worksheet and Removing unwanted rows

```
df_Measures2.tail()
```

	name	name.1	area	ellipse axe 2 length	max axe 2 length	deviation max axe 2 from center	axe 2 Moment 2	a Mome
159	MJ02.1036-d_Post_g	1036-d_Post_g	725224.0	821.4157	822.0	-6.8278	1.436304e+11	8.148462
160	MJ02.1037-	1037-	1133782.0	854.1530	924.0	-49.0985	2.587367e+11	1.559255

```
df_Measures2.isnull().sum().sum()
```

```
0
```

Null Value Checking in the Axis 2 Measures

```
df_Measures2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 164 entries, 0 to 163
Data columns (total 13 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   name                            164 non-null    object
 1   name.1                          164 non-null    object
 2   area                            164 non-null    float64
 3   ellipse axe 2 length            164 non-null    float64
 4   max axe 2 length                164 non-null    float64
 5   deviation max axe 2 from center 164 non-null    float64
 6   axe 2 Moment 2                  164 non-null    float64
 7   axe 2 Moment 3                  164 non-null    float64
 8   axe 2 Moment 4                  164 non-null    float64
 9   std axe 2 Moment 2              164 non-null    float64
 10  std axe 2 Moment 3              164 non-null    float64
                                     164 non-null    float64
                                     164 non-null    float64
dtypes: float64(11), object(2)
memory usage: 17.9+ KB
```

Saved successfully!

Information about the data types in Axis 2 Measures

```
df_Measures2.drop(['name'], axis=1,inplace = True)
df_Measures2.rename( {'name.1':'id' } , axis=1 , inplace = True)
df_Measures2.head()
```

	id	area	ellipse axe 2 length	max axe 2 length	deviation max axe 2 from center	axe 2 Moment 2	axe 2 Moment 3	axe Momen
0	0725-d_Ant_d	954465.0	777.8875	844.0	-25.2980	1.791255e+11	9.815703e+13	5.823411e-
1	0725-d_Post_g	638157.0	729.2658	731.0	-95.3215	9.913302e+10	4.978770e+13	2.690291e-
2	0726-d_Ant_d	984051.0	774.3652	844.0	-39.1989	1.782607e+11	9.668417e+13	5.690378e-
3	0726-	634452.0	710.0563	718.0	-108.5135	9.449492e+10	4.642871e+13	2.456467e-
4	d_Ant_g	889416.0	732.2889	801.0	-44.8532	1.446070e+11	7.412817e+13	4.119448e-

dropping the first column and rename the name.1 -id

```
df_Measures2['id'] = df_Measures2['id'].str.extract('(\d+)')
```

```
df_Measures2['id'] = df_Measures2['id'].astype('int')
df_Measures2['area'] = df_Measures2['area'].astype('int')
df_Measures2['max axe 2 length'] = df_Measures2['max axe 2 length'].astype('int')
df_Measures2['axe 2 Moment 2'] = df_Measures2['axe 2 Moment 2'].astype('int')
df_Measures2['axe 2 Moment 3'] = df_Measures2['axe 2 Moment 3'].astype('int')
df_Measures2['axe 2 Moment 4'] = df_Measures2['axe 2 Moment 4'].astype('int')
df_Measures2['max axe 1 length'] = df_Measures2['max axe 1 length'].astype('int')

df_Measures2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 164 entries, 0 to 163
Data columns (total 12 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   id                          164 non-null    int64
 1   area                        164 non-null    int64
 2   ellipse axe 2 length        164 non-null    float64
 3   max axe 2 length            164 non-null    int64
 4   deviation max axe 2 from center 164 non-null  float64
 5   axe 2 Moment 2              164 non-null    int64
 6   axe 2 Moment 3              164 non-null    int64
 7   axe 2 Moment 4              164 non-null    int64
 8   std axe 2 Moment 2          164 non-null    float64
 9   std axe 2 Moment 3          164 non-null    float64
 10  std axe 2 Moment 4          164 non-null    float64
 11  max axe 1 length            164 non-null    int64
dtypes: float64(5), int64(7)
memory usage: 16.7 KB
```

Saved successfully!

Changing the varaiabe type of the Axis 2 Measures

```
df_Measures2.head(10)
```

	id	area	ellipse axe 2 length	max axe 2 length	deviation max axe 2 from center	axe 2 Moment 2	axe 2 Moment 3	axe 2
0	725	954465	777.8875	844	-25.2980	179125471444	98157027576774	582341051
1	725	638157	729.2658	731	-95.3215	99133019103	49787697268249	269029093
2	726	984051	774.3652	844	-39.1989	178260702151	96684173147953	569037777
3	726	634452	710.0563	718	-108.5135	94494918365	46428712462551	245646716
4	735	889416	732.2889	801	-44.8532	144607023282	74128171449556	411944827
5	735	552109	675.1800	670	-50.1721	70551107988	32464121232856	161090771
6	736	998666	791.1103	863	-24.5902	194687731402	108513568126722	654043164
7	736	630493	732.1526	727	-85.5836	98590417572	49734103171044	270068426
8	738	1105201	829.3862	900	-55.5016	231202150457	134401847957309	846653020
9	738	745133	810.1177	821	-155.7159	150723211926	85637734990026	523697946

Saved successfully!

```
# Get data from  worksheet and removing the unwanted rows
df_perimeter = df_dict.get('Wing perimeter')
df_perimeter = df_perimeter.drop(df_perimeter.index[82:87])
# Preview
df_perimeter.head()
```

	name	name.1	Perimeter
0	0725-d_Ant_d	MJ02.0725-d_Ant_d	4565.9549

```
df_perimeter = df_perimeter.drop(['name.1'], axis=1)
df_perimeter['name'] = df_perimeter.name.str.extract('(\d+)')

df_perimeter.rename( {'name':'id' } , axis=1 , inplace = True)
df_perimeter.head()
```

	id	Perimeter
0	0725	4565.9549
1	0726	4595.8817
2	0735	4447.3519
3	0736	4727.6391
4	0738	4885.6046

Rename of the name to id and removing extra character from the column id

```
df_perimeter.isnull().sum().sum()
```

```
0
```

Checking Null values in the Wing Perimeter sheet

```
df_perimeter.tail()
```

Saved successfully!

77	1029	4914.9011
78	1030	4785.2543
79	1036	4654.8767
80	1037	4735.4208
81	1052	4949.8300

```
df_perimeter['id'] = df_perimeter['id'].astype('int')

df_perimeter.info()
```

```
<class 'pandas.core.frame.DataFrame'>
```

```
Int64Index: 82 entries, 0 to 81
Data columns (total 2 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   id         82 non-null     int64
 1   Perimeter  82 non-null     float64
dtypes: float64(1), int64(1)
memory usage: 1.9 KB
```

```
df_perimeter.head(5)
```

	id	Perimeter
0	725	4565.9549
1	726	4595.8817
2	735	4447.3519
3	736	4727.6391
4	738	4885.6046

Wild Wings,Axis 2 Cleaning

Saved successfully!

```
# Get data from  worksheet
df_Wildwings_axis1 = df_dict.get('Wild wings, axis 1')
df_Wildwings_axis1 = df_Wildwings_axis1.drop(df_Wildwings_axis1.index[2888:28921])
# Preview
df_Wildwings_axis1.head()
```

	Name	Unnamed: 1	Tribe	Genus	Species	sub-species	name	area
0	JM00.0001-d_Ant_d.tif	f	Heliconiinae	Heliconius	numata	elegans	JM00.0001-d_Ant_d	892614.0
1	JM00.0001-d_Ant_g.tif	f	Heliconiinae	Heliconius	numata	elegans	JM00.0001-d_Ant_g	889094.0
2	JM00.0001-d_Post_d.tif	f	Heliconiinae	Heliconius	numata	elegans	JM00.0001-d_Post_d	562096.0
3	JM00.0001-	f	Heliconiinae	Heliconius	numata	elegans	JM00.0001-	582341.0

```
df_Wildwings_axis1.tail()
```

	Name	Unnamed: 1	Tribe	Genus	Species	sub-species	name	area
2883	MJ99.1241-d_Post_g.tif	f	Heliconiinae	Heliconius	numata	silvana	MJ99.1241-d_Post_g	614848.
2884	MJ99.1241-v_Ant_d.tif	f	Heliconiinae	Heliconius	numata	silvana	MJ99.1241-v_Ant_d	937794.
2885	MJ99.1241-v_Ant_g.tif	f	Heliconiinae	Heliconius	numata	silvana	MJ99.1241-v_Ant_g	935533.
2886	MJ99.1241-v_Post_d.tif	f	Heliconiinae	Heliconius	numata	silvana	MJ99.1241-v_Post_d	618037.
2887	MJ99.1241-	f	Heliconiinae	Heliconius	numata	silvana	MJ99.1241-v_Post_g	608690.

Saved successfully!

```
df_Wildwings_axis1.isnull().sum().sum()
```

```
56
```

```
df_Wildwings_axis1.isnull().sum()
```

```
Name                              0
Unnamed: 1                        0
Tribe                             8
Genus                             0
Species                           0
sub-species                      48
```

```
name                                0
area                                0
ellipse axe 1 length                0
max axe 1 length                    0
deviation max axe 1 from center     0
axe 1 Moment 2                      0
axe 1 Moment 3                      0
axe 1 Moment 4                      0
std axe 1 Moment 2                  0
std axe 1 Moment 3                  0
std axe 1 Moment 4                  0
Perimeter                           0
dtype: int64
```

```
df_Wildwings_axis1 = df_Wildwings_axis1.fillna('')
```

```
df_Wildwings_axis1.isnull().sum().sum()
```

```
0
```

```
df_Wildwings_axis1['Name']=df_Wildwings_axis1['name'].str[5:9]
df_Wildwings_axis1.head()
```

	Name	Unnamed: 1	Tribe	Genus	Species	sub-species	name	area	ellip axe leng
0	0001	f	Heliconiinae	Heliconius	numata	elegans	JM00.0001-d_Ant_d	892614.0	1581.41
1	0001	f	Heliconiinae	Heliconius	numata	elegans	JM00.0001-d_Ant_g	889094.0	1567.05
2	0001	f	Heliconiinae	Heliconius	numata	elegans	JM00.0001-d_Post_d	562096.0	1047.86
3	0001	f	Heliconiinae	Heliconius	numata	elegans	JM00.0001-d_Post_g	582341.0	1054.83
4	0001	f	Heliconiinae	Heliconius	numata	elegans	JM00.0001-v_Ant_d	893173.0	1573.86

Saved successfully!

```
df_Wildwings_axis1.rename({'Name':'id'},axis=1,inplace = True)
df_Wildwings_axis1.drop(['name'], axis=1,inplace=True)
df_Wildwings_axis1.head()
```

	id	Unnamed: 1	Tribe	Genus	Species	sub-species	area	ellipse axe 1 length	max axe 1 length
0	0001	f	Heliconiinae	Heliconius	numata	elegans	892614.0	1581.4126	1632.0
1	0001	f	Heliconiinae	Heliconius	numata	elegans	889094.0	1567.0531	1625.0
2	0001	f	Heliconiinae	Heliconius	numata	elegans	562096.0	1047.8688	1057.0
3	0001	f	Heliconiinae	Heliconius	numata	elegans	582341.0	1054.8397	1049.0
4	0001	f	Heliconiinae	Heliconius	numata	elegans	893173.0	1573.8670	1630.0

```
df_Wildwings_axis1.rename( {'Unnamed: 1':'sex'} ,axis=1 , inplace = True)

df_Wildwings_axis1.head()
```

	id	sex	Tribe	Genus	Species	sub-species	area	ellipse axe 1 length	max axe 1 length	devi max c
0	0001	f	Heliconiinae	Heliconius	numata	elegans	892614.0	1581.4126	1632.0	124
1	0001	f	Heliconiinae	Heliconius	numata	elegans	889094.0	1567.0531	1625.0	128
2	0001	f	Heliconiinae	Heliconius	numata	elegans	562096.0	1047.8688	1057.0	-2
3	0001	f	Heliconiinae	Heliconius	numata	elegans	582341.0	1054.8397	1049.0	10
				conius	numata	elegans	893173.0	1573.8670	1630.0	125

Saved successfully!

```
df_Wildwings_axis1['sex'] = df_Wildwings_axis1['sex'].str.upper()
df_Wildwings_axis1.head()
```

	id	sex	Tribe	Genus	Species	sub-species	area	ellipse axe 1 length	max axe 1 length	devi max c
0	0001	F	Heliconiinae	Heliconius	numata	elegans	892614.0	1581.4126	1632.0	124
1	0001	F	Heliconiinae	Heliconius	numata	elegans	889094.0	1567.0531	1625.0	128

```
df_Wildwings_axis1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2888 entries, 0 to 2887
Data columns (total 17 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   id                              2888 non-null   object
 1   sex                             2888 non-null   object
 2   Tribe                           2888 non-null   object
 3   Genus                           2888 non-null   object
 4   Species                         2888 non-null   object
 5   sub-species                     2888 non-null   object
 6   area                            2888 non-null   float64
 7   ellipse axe 1 length            2888 non-null   float64
 8   max axe 1 length                2888 non-null   float64
 9   deviation max axe 1 from center 2888 non-null   float64
 10  axe 1 Moment 2                  2888 non-null   float64
 11  axe 1 Moment 3                  2888 non-null   float64
 12  axe 1 Moment 4                  2888 non-null   float64
 13  std axe 1 Moment 2              2888 non-null   float64
 14  std axe 1 Moment 3              2888 non-null   float64
 15  std axe 1 Moment 4              2888 non-null   float64
 16  Perimeter                       2888 non-null   float64
dtypes: float64(11), object(6)
memory usage: 406.1+ KB
```

Saved successfully!

```
ings_axis1['id'].astype('int')

df_Wildwings_axis1['axe 1 Moment 2'] = df_Wildwings_axis1['axe 1 Moment 2'].astype('int')
df_Wildwings_axis1['axe 1 Moment 3'] = df_Wildwings_axis1['axe 1 Moment 3'].astype('int')
df_Wildwings_axis1['axe 1 Moment 4'] = df_Wildwings_axis1['axe 1 Moment 4'].astype('int')
df_Wildwings_axis1['max axe 1 length'] = df_Wildwings_axis1['max axe 1 length'].astype('int')

df_Wildwings_axis1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2888 entries, 0 to 2887
Data columns (total 17 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   id                              2888 non-null   int64
 1   sex                             2888 non-null   object
 2   Tribe                           2888 non-null   object
```

```
 3   Genus                           2888 non-null   object
 4   Species                         2888 non-null   object
 5   sub-species                     2888 non-null   object
 6   area                            2888 non-null   float64
 7   ellipse axe 1 length            2888 non-null   float64
 8   max axe 1 length                2888 non-null   int64
 9   deviation max axe 1 from center 2888 non-null   float64
 10  axe 1 Moment 2                  2888 non-null   int64
 11  axe 1 Moment 3                  2888 non-null   int64
 12  axe 1 Moment 4                  2888 non-null   int64
 13  std axe 1 Moment 2              2888 non-null   float64
 14  std axe 1 Moment 3              2888 non-null   float64
 15  std axe 1 Moment 4              2888 non-null   float64
 16  Perimeter                       2888 non-null   float64
dtypes: float64(7), int64(5), object(5)
memory usage: 406.1+ KB
```

```
for col in ['sex', 'Genus','Tribe', 'Species', 'sub-species']:
  df_Wildwings_axis1[col] = df_Wildwings_axis1[col].astype('category')
```

```
df_Wildwings_axis1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2888 entries, 0 to 2887
Data columns (total 17 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   id                              2888 non-null   int64
 1   sex                             2888 non-null   category
 2   Tribe                           2888 non-null   category
 3   Genus                           2888 non-null   category
 4   Species                         2888 non-null   category
 5   sub-species                     2888 non-null   category
                                     2888 non-null   float64
                                     2888 non-null   float64
                                     2888 non-null   int64
 9   deviation max axe 1 from center 2888 non-null   float64
 10  axe 1 Moment 2                  2888 non-null   int64
 11  axe 1 Moment 3                  2888 non-null   int64
 12  axe 1 Moment 4                  2888 non-null   int64
 13  std axe 1 Moment 2              2888 non-null   float64
 14  std axe 1 Moment 3              2888 non-null   float64
 15  std axe 1 Moment 4              2888 non-null   float64
 16  Perimeter                       2888 non-null   float64
dtypes: category(5), float64(7), int64(5)
memory usage: 308.9 KB
```

Saved successfully!

Wild wings, axis 2 Cleaning

```
# Get data from  worksheet
df_Wildwings_axis2 = df_dict.get('Wild wings, axis 2')
df_Wildwings_axis2 = df_Wildwings_axis2.drop(df_Wildwings_axis2.index[432:])
# Preview
df_Wildwings_axis2.head()
```

	Name	doresal/ventral	Anterior/Posterior	left/right	Sex	sub/Family	Gen
0	JM00.0001	d	Ant	d	f	Heliconiinae	Heliconi
1	JM00.0018	d	Ant	g	f	Heliconiinae	Heliconi
2	JM02.0107	d	Ant	d	f	Ithomiinae	Melina
3	JM02.0107	d	Ant	g	f	Ithomiinae	Melina
			Ant	d	f	Ithomiinae	Melina

Saved successfully!

```
df_Wildwings_axis2.tail()
```

	Name	doresal/ventral	Anterior/Posterior	left/right	Sex	sub/Family	Ge
427	MJ99.0196	v	Ant	g	f	Ithomiinae	Melin
428	MJ99.0217	d	Ant	d	m	Ithomiinae	Melin
429	MJ99.0217	d	Ant	g	m	Ithomiinae	Melin

```
df_Wildwings_axis2.isnull().sum().sum()
```

```
4
```

```
df_Wildwings_axis2.isnull().sum()
```

```
Name                                0
doresal/ventral                     0
Anterior/Posterior                  0
left/right                          0
Sex                                 0
sub/Family                          4
Genus                               0
Species                             0
Subspecies                          0
name                                0
area                                0
ellipse axe 2 length                0
max axe 2 length                    0
deviation max axe 2 from center     0
axe 2 Moment 2                      0
                                    0
                                    0
std axe 2 Moment 2                  0
std axe 2 Moment 3                  0
std axe 2 Moment 4                  0
dtype: int64
```

Saved successfully!

```
df_Wildwings_axis2 = df_Wildwings_axis2.fillna('')
```

```
df_Wildwings_axis2.isnull().sum().sum()
```

```
0
```

Replacing all the null values with NaN

```
df_Wildwings_axis2['Name'] = df_Wildwings_axis2['name'].str[5:9]
df_Wildwings_axis2.head()
```

	Name	doresal/ventral	Anterior/Posterior	left/right	Sex	sub/Family	Genus	
0	0001	d	Ant	d	f	Heliconiinae	Heliconius	
1	0018	d	Ant	g	f	Heliconiinae	Heliconius	
2	0107	d	Ant	d	f	Ithomiinae	Melinaea	m
3	0107	d	Ant	g	f	Ithomiinae	Melinaea	m
4	0107	v	Ant	d	f	Ithomiinae	Melinaea	m

Cleaning the name Column that having the unecessary characters

```
df_Wildwings_axis2.rename( {'Name':'id' } ,axis=1 ,inplace = True)
df_Wildwings_axis2.drop(['name'], axis=1,inplace=True)
```

Renaming the name column to id column and droppping the unnecessary column

Saved successfully!

```
df_Wildwings_axis2.head()
```

```
df_Wildwings_axis2.rename( {'Subspecies':'sub-species' } ,axis=1 , inplace = True)
df_Wildwings_axis2.rename( {'Sex':'sex' } , axis=1 , inplace = True)
```

Renaming the Subspecies and Sex column

```
df_Wildwings_axis2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 432 entries, 0 to 431
Data columns (total 19 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   id                              432 non-null    object
 1   doresal/ventral                 432 non-null    object
 2   Anterior/Posterior              432 non-null    object
 3   left/right                      432 non-null    object
 4   sex                             432 non-null    object
 5   sub/Family                      432 non-null    object
 6   Genus                           432 non-null    object
 7   Species                         432 non-null    object
 8   sub-species                     432 non-null    object
 9   area                            432 non-null    float64
 10  ellipse axe 2 length            432 non-null    float64
 11  max axe 2 length                432 non-null    float64
 12  deviation max axe 2 from center 432 non-null    float64
 13  axe 2 Moment 2                  432 non-null    float64
 14  axe 2 Moment 3                  432 non-null    float64
 15  axe 2 Moment 4                  432 non-null    float64
 16  std axe 2 Moment 2              432 non-null    float64
 17  std axe 2 Moment 3              432 non-null    float64
```

Saved successfully!

```
df_Wildwings_axis2['id'] = df_Wildwings_axis2['id'].astype('int')
df_Wildwings_axis2['axe 2 Moment 2'] = df_Wildwings_axis2['axe 2 Moment 2'].astype('int')
df_Wildwings_axis2['axe 2 Moment 3'] = df_Wildwings_axis2['axe 2 Moment 3'].astype('int')
df_Wildwings_axis2['axe 2 Moment 4'] = df_Wildwings_axis2['axe 2 Moment 4'].astype('int')
df_Wildwings_axis2['sex'] = df_Wildwings_axis2['sex'].astype('category')
df_Wildwings_axis2['doresal/ventral'] = df_Wildwings_axis2['doresal/ventral'].astype('categor
df_Wildwings_axis2['Anterior/Posterior'] = df_Wildwings_axis2['Anterior/Posterior'].astype('c
df_Wildwings_axis2['left/right'] = df_Wildwings_axis2['left/right'].astype('category')
df_Wildwings_axis2['area'] = df_Wildwings_axis2['area'].astype('int')
df_Wildwings_axis2['sub/Family'] = df_Wildwings_axis2['sub/Family'].astype('category')
df_Wildwings_axis2['Genus'] = df_Wildwings_axis2['Genus'].astype('category')
df_Wildwings_axis2['Species'] = df_Wildwings_axis2['Species'].astype('category')
df_Wildwings_axis2['sub-species'] = df_Wildwings_axis2['sub-species'].astype('category')
df_Wildwings_axis2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 432 entries, 0 to 431
Data columns (total 19 columns):
 #   Column                            Non-Null Count  Dtype
---  ------                            --------------  -----
 0   id                                432 non-null    int64
 1   doresal/ventral                   432 non-null    category
 2   Anterior/Posterior                432 non-null    category
 3   left/right                        432 non-null    category
 4   sex                               432 non-null    category
 5   sub/Family                        432 non-null    category
 6   Genus                             432 non-null    category
 7   Species                           432 non-null    category
 8   sub-species                       432 non-null    category
 9   area                              432 non-null    int64
 10  ellipse axe 2 length              432 non-null    float64
 11  max axe 2 length                  432 non-null    float64
 12  deviation max axe 2 from center   432 non-null    float64
 13  axe 2 Moment 2                    432 non-null    int64
 14  axe 2 Moment 3                    432 non-null    int64
 15  axe 2 Moment 4                    432 non-null    int64
 16  std axe 2 Moment 2                432 non-null    float64
 17  std axe 2 Moment 3                432 non-null    float64
 18  std axe 2 Moment 4                432 non-null    float64
dtypes: category(8), float64(6), int64(5)
memory usage: 45.1 KB
```

```
df_Wildwings_axis2.tail(10)
```

Saved successfully!

423	195	v	Ant	g	f	Ithomiinae	Melinaea	m

Landmark details Cleaning

427	196	v	Ant	g	f	Ithomiinae	Melinaea	m

```
# Get data from  worksheet
df_LandMark = df_dict.get('Landmark details')
df_LandMark = df_LandMark.drop(df_LandMark.index[81:85])
# Preview
df_LandMark.head()
```

	ID	sex	name	genotype	phenotype	sex-phenotype
0	ID=0	f	725.0	a-e	elegans	f-elegans
1	ID=1	f	726.0	a-s	aurora	f-aurora
2	ID=2	f	736.0	a-e	elegans	f-elegans
3	ID=3	m	738.0	a-e	elegans	m-elegans
4	ID=4	m	739.0	a-s	aurora	m-aurora

```
df_LandMark.tail()
```

Saved successfully!

76	ID=76	m	1028.0	a-s	aurora	m-aurora
77	ID=77	m	1030.0	e-s	elegans	m-elegans
78	ID=78	m	1036.0	a-s	aurora	m-aurora
79	ID=79	m	1037.0	a-e	elegans	m-elegans
80	ID=80	m	1052.0	a-e	elegans	m-elegans

```
df_LandMark.isnull().sum().sum()
```

```
0
```

```
df_LandMark.rename( {'name':'id' } ,axis=1,inplace = True)
```

```
df_LandMark.drop(['ID'], axis=1,inplace=True)
```

Renaming the name column to id

```
df_LandMark.head()
```

	sex	id	genotype	phenotype	sex-phenotype
0	f	725.0	a-e	elegans	f-elegans
1	f	726.0	a-s	aurora	f-aurora
2	f	736.0	a-e	elegans	f-elegans
3	m	738.0	a-e	elegans	m-elegans
4	m	739.0	a-s	aurora	m-aurora

```
df_LandMark.tail()
```

	sex	id	genotype	phenotype	sex-phenotype
76	m	1028.0	a-s	aurora	m-aurora
77	m	1030.0	e-s	elegans	m-elegans
78	m	1036.0	a-s	aurora	m-aurora
79	m	1037.0	a-e	elegans	m-elegans
80	m	1052.0	a-e	elegans	m-elegans

```
df_LandMark.info()
```

Saved successfully!

```
Frame'>
Int64Index: 81 entries, 0 to 80
Data columns (total 5 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   sex            81 non-null     object
 1   id             81 non-null     float64
 2   genotype       81 non-null     object
 3   phenotype      81 non-null     object
 4   sex-phenotype  81 non-null     object
dtypes: float64(1), object(4)
memory usage: 3.8+ KB
```

```
neworder = ['id','sex','genotype','phenotype','sex-phenotype']
df_LandMark=df_LandMark.reindex(columns=neworder)
```

```
df_LandMark.head()
```

	id	sex	genotype	phenotype	sex-phenotype
0	725.0	f	a-e	elegans	f-elegans
1	726.0	f	a-s	aurora	f-aurora
2	736.0	f	a-e	elegans	f-elegans
3	738.0	m	a-e	elegans	m-elegans
4	739.0	m	a-s	aurora	m-aurora

```
df_LandMark['id'] = df_LandMark['id'].astype('int')
df_LandMark['sex'] = df_LandMark['sex'].astype('category')
df_LandMark['genotype'] = df_LandMark['genotype'].astype('category')
df_LandMark['phenotype'] = df_LandMark['phenotype'].astype('category')
df_LandMark['sex-phenotype'] = df_LandMark['sex-phenotype'].astype('category')
```

```
df_LandMark.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 81 entries, 0 to 80
Data columns (total 5 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   id             81 non-null     int64
 1   sex            81 non-null     category
 2   genotype       81 non-null     category
 3   phenotype      81 non-null     category
 4   sex-phenotype  81 non-null     category
dtypes: category(4), int64(1)
memory usage: 2.2 KB
```

Saved successfully!

```
ex'].str.upper()
df_LandMark.head()
```

	id	sex	genotype	phenotype	sex-phenotype
0	725	F	a-e	elegans	f-elegans
1	726	F	a-s	aurora	f-aurora
2	736	F	a-e	elegans	f-elegans
3	738	M	a-e	elegans	m-elegans
4	739	M	a-s	aurora	m-aurora

*Outline analysis -brood Cleaning *

```
# Get data from  worksheet
df_Out_Analy = df_dict.get('Outline analysis -brood')
df_Out_Analy = df_Out_Analy.drop(df_Out_Analy.index[82:94])
# Preview
df_Out_Analy.head()
```

	File Name	Wing	Genotype 1	Genotype 2	sex	ind Name	side	left/right	PCA scores	Un
0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	axis Variance	0.51846	0
1	MJ02.0725-v_Ant_g.tif	Ant	aurora	elegans	f	MJ02.0725	v	g	31.67010	7
2	MJ02.0726-v_Ant_g.tif	Ant	aurora	silvana	f	MJ02.0726	v	g	-5.16480	-15
3	MJ02.0735-v_Ant_d.tif	Ant	silvana	silvana	f	MJ02.0735	v	d	19.51230	-9
				elegans	f	MJ02.0736	v	d	26.80950	16

Saved successfully!

5 rows × 88 columns

```
df_Out_Analy.tail()
```

	File Name	Wing	Genotype 1	Genotype 2	sex	ind Name	side	left/right	PCA scores	Un
77	MJ02.1029-v_Ant_d.tif	Ant	silvana	silvana	m	MJ02.1029	v	d	-21.0567	-1
78	MJ02.1030-v_Ant_g.tif	Ant	elegans	silvana	m	MJ02.1030	v	g	-3.6090	·
79	MJ02.1036-v_Ant_g.tif	Ant	aurora	silvana	m	MJ02.1036	v	g	15.0247	-1
80	MJ02.1037-v_Ant_g.tif	Ant	aurora	elegans	m	MJ02.1037	v	g	9.3776	-1

```
df_Out_Analy.isnull().sum().sum()
```

```
7
```

```
df_Out_Analy = df_Out_Analy.fillna('')
```

```
df_Out_Analy.isnull().sum().sum()
```

```
0
```

```
df_Out_Analy.info()
```

```
 32  Unnamed: 32  82 non-null     float64
 33  Unnamed: 33  82 non-null     float64
 34  Unnamed: 34  82 non-null     float64
 35  Unnamed: 35  82 non-null     float64
 36  Unnamed: 36  82 non-null     float64
 37  Unnamed: 37  82 non-null     float64
 38  Unnamed: 38  82 non-null     float64
                                  float64
                                  float64
                                  float64
 42  Unnamed: 42  82 non-null     float64
 43  Unnamed: 43  82 non-null     float64
 44  Unnamed: 44  82 non-null     float64
 45  Unnamed: 45  82 non-null     float64
 46  Unnamed: 46  82 non-null     float64
 47  Unnamed: 47  82 non-null     float64
 48  Unnamed: 48  82 non-null     float64
 49  Unnamed: 49  82 non-null     float64
 50  Unnamed: 50  82 non-null     float64
 51  Unnamed: 51  82 non-null     float64
 52  Unnamed: 52  82 non-null     float64
 53  Unnamed: 53  82 non-null     float64
 54  Unnamed: 54  82 non-null     float64
 55  Unnamed: 55  82 non-null     float64
 56  Unnamed: 56  82 non-null     float64
 57  Unnamed: 57  82 non-null     float64
 58  Unnamed: 58  82 non-null     float64
```

Saved successfully!

```
 59  Unnamed: 59  82 non-null     float64
 60  Unnamed: 60  82 non-null     float64
 61  Unnamed: 61  82 non-null     float64
 62  Unnamed: 62  82 non-null     float64
 63  Unnamed: 63  82 non-null     float64
 64  Unnamed: 64  82 non-null     float64
 65  Unnamed: 65  82 non-null     float64
 66  Unnamed: 66  82 non-null     float64
 67  Unnamed: 67  82 non-null     float64
 68  Unnamed: 68  82 non-null     float64
 69  Unnamed: 69  82 non-null     float64
 70  Unnamed: 70  82 non-null     float64
 71  Unnamed: 71  82 non-null     float64
 72  Unnamed: 72  82 non-null     float64
 73  Unnamed: 73  82 non-null     float64
 74  Unnamed: 74  82 non-null     float64
 75  Unnamed: 75  82 non-null     float64
 76  Unnamed: 76  82 non-null     float64
 77  Unnamed: 77  82 non-null     float64
 78  Unnamed: 78  82 non-null     float64
 79  Unnamed: 79  82 non-null     float64
 80  Unnamed: 80  82 non-null     float64
 81  Unnamed: 81  82 non-null     float64
 82  Unnamed: 82  82 non-null     float64
 83  Unnamed: 83  82 non-null     float64
 84  Unnamed: 84  82 non-null     float64
 85  Unnamed: 85  82 non-null     float64
 86  Unnamed: 86  82 non-null     float64
 87  Unnamed: 87  82 non-null     float64
dtypes: float64(80), object(8)
memory usage: 57.0+ KB
```

```
df_Out_Analy['Sum_pca'] = df_Out_Analy.iloc[:,8:87].mean(axis=1)
```

Saved successfully!

	File Name	Wing	Genotype 1	Genotype 2	sex	ind Name	side	left/right	PCA scores	Unr
0								axis Variance	0.51846	0

```
df_Out_Analy.drop(df_Out_Analy.iloc[:,8:87], inplace = True, axis = 1)
df_Out_Analy.drop('Unnamed: 87',inplace=True, axis = 1)
df_Out_Analy.head()
```

	File Name	Wing	Genotype 1	Genotype 2	sex	ind Name	side	left/right	Sum_pca
0								axis Variance	0.012658
1	MJ02.0725-v_Ant_g.tif	Ant	aurora	elegans	f	MJ02.0725	v	g	0.357203
2	MJ02.0726-v_Ant_g.tif	Ant	aurora	silvana	f	MJ02.0726	v	g	-0.188276

```
df_Out_Analy.tail()
```

	File Name	Wing	Genotype 1	Genotype 2	sex	ind Name	side	left/right	Sum_pca
77	MJ02.1029-v_Ant_d.tif	Ant	silvana	silvana	m	MJ02.1029	v	d	-0.468096
78	MJ02.1030-v_Ant_g.tif	Ant	elegans	silvana	m	MJ02.1030	v	g	-0.254634
79	MJ02.1036-v_Ant_g.tif	Ant	aurora	silvana	m	MJ02.1036	v	g	-0.195669

Saved successfully!

```
df_Out_Analy['File Name'] = df_Out_Analy['File Name'].str[5:9]
df_Out_Analy['File Name'].head()
```

```
0
1    0725
2    0726
3    0735
4    0736
Name: File Name, dtype: object
```

Renamig the name column

```
df_Out_Analy.rename( {'File Name':'id' } ,axis=1 , inplace = True)
df_Out_Analy.drop( {'Wing'} ,axis=1 , inplace = True)
df_Out_Analy.drop('ind Name',inplace=True, axis=1)
df_Out_Analy.head()
```

	id	Genotype 1	Genotype 2	sex	side	left/right	Sum_pca
0						axis Variance	0.012658
1	0725	aurora	elegans	f	v	g	0.357203
2	0726	aurora	silvana	f	v	g	-0.188276
3	0735	silvana	silvana	f	v	d	0.691022
4	0736	aurora	elegans	f	v	d	0.607218

Dropping and renaming the require column as the wing column could not be the same for all identity so dropping it

```
df_Out_Analy = df_Out_Analy.iloc[1:,:]
df_Out_Analy.head()
```

	id	Genotype 1	Genotype 2	sex	side	left/right	Sum_pca
1	0725	aurora	elegans	f	v	g	0.357203
2	0726	aurora	silvana	f	v	g	-0.188276
3	0735	silvana	silvana	f	v	d	0.691022
4	0736	aurora	elegans	f	v	d	0.607218
				m	v	d	0.062585

Saved successfully!

```
df_Out_Analy.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 81 entries, 1 to 81
Data columns (total 7 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   id          81 non-null     object
 1   Genotype 1  81 non-null     object
 2   Genotype 2  81 non-null     object
 3   sex         81 non-null     object
 4   side        81 non-null     object
 5   left/right  81 non-null     object
 6   Sum_pca     81 non-null     float64
```

```
dtypes: float64(1), object(6)
```

```
df_Out_Analy['id'] = df_Out_Analy['id'].astype('int')
for col in [ 'Genotype 1','Genotype 2','sex','side','left/right']:
  df_Out_Analy[col] = df_Out_Analy[col].astype('category')
```

```
df_Out_Analy.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 81 entries, 1 to 81
Data columns (total 7 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   id          81 non-null     int64
 1   Genotype 1  81 non-null     category
 2   Genotype 2  81 non-null     category
 3   sex         81 non-null     category
 4   side        81 non-null     category
 5   left/right  81 non-null     category
 6   Sum_pca     81 non-null     float64
dtypes: category(5), float64(1), int64(1)
memory usage: 2.9 KB
```

```
df_Out_Analy['sex'] = df_Out_Analy['sex'].str.upper()
df_Out_Analy.head()
```

	id	Genotype 1	Genotype 2	sex	side	left/right	Sum_pca
1	725	aurora	elegans	F	v	g	0.357203
2	726	aurora	silvana	F	v	g	-0.188276
3	735	silvana	silvana	F	v	d	0.691022
4	726	aurora	elegans	F	v	d	0.607218
				M	v	d	0.062585

Saved successfully!

Outline analysis -wild Cleaning

```
# Get data from  worksheet
df_Out_Analy_wild = df_dict.get('Outline analysis -wild')
```

```
N = 8
df_Out_Analy_wild = df_Out_Analy_wild.iloc[:-N , :]
# Preview
df_Out_Analy_wild.head()
```

	File Name	Wing	Genotype 1	Genotype 2	sex	ind Name	side	left/right	PCA scores	Un
0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	axis Variance	0.50301	0
1	JM00.0001-v_Ant_g.tif	Ant	numata	elegans	f	JM00.0001	v	g	2.79410	1
2	JM00.0018-v_Ant_d.tif	Ant	numata	elegans	f	JM00.0018	v	d	2.10520	-2
3	JM02.0107-v_Ant_d.tif	Ant	marsaeus	phasiana	f	JM02.0107	v	d	17.00490	1
4	JM02.0108-v_Ant_g.tif	Ant	marsaeus	phasiana	f	JM02.0108	v	g	17.61300	-1

5 rows × 88 columns

Removing the last 8 rows from the sheet Outline Analysis Wild

```
df_Out_Analy_wild.tail(8)
```

Saved successfully!

	File Name	Wing	Genotype 1	Genotype 2	sex	ind Name	side	left/right	PCA scores
361	MJ99.0210-v_Ant_g.tif	Ant	menophilus	sspn	f	MJ99.0210	v	g	6.41330
362	MJ99.0211-v_Ant_g.tif	Ant	menophilus	sspn	f	MJ99.0211	v	g	8.19060
363	MJ99.0212-	Ant	marsaeus	rilevi	f	MJ99.0212	v	g	0.36977

```
df_Out_Analy_wild.isnull().sum().sum()
```

```
7
```

	MJ99.0215-

```
df_Out_Analy_wild·=·df_Out_Analy_wild.fillna('')
```

	MJ99.0216-

```
df_Out_Analy_wild.isnull().sum().sum()
```

```
0
```

Checking the Null Values and Removing it from the requied dataframe

8 rows x 88 columns

```
df_Out_Analy_wild.info()
```

```
 32  Unnamed: 32  369 non-null    float64
 33  Unnamed: 33  369 non-null    float64
 34  Unnamed: 34  369 non-null    float64
 35  Unnamed: 35  369 non-null    float64
 36  Unnamed: 36  369 non-null    float64
 37  Unnamed: 37  369 non-null    float64
 38  Unnamed: 38  369 non-null    float64
 39  Unnamed: 39  369 non-null    float64
                                  float64
                                  float64
 42  Unnamed: 42  369 non-null    float64
 43  Unnamed: 43  369 non-null    float64
 44  Unnamed: 44  369 non-null    float64
 45  Unnamed: 45  369 non-null    float64
 46  Unnamed: 46  369 non-null    float64
 47  Unnamed: 47  369 non-null    float64
 48  Unnamed: 48  369 non-null    float64
 49  Unnamed: 49  369 non-null    float64
 50  Unnamed: 50  369 non-null    float64
 51  Unnamed: 51  369 non-null    float64
 52  Unnamed: 52  369 non-null    float64
 53  Unnamed: 53  369 non-null    float64
 54  Unnamed: 54  369 non-null    float64
 55  Unnamed: 55  369 non-null    float64
 56  Unnamed: 56  369 non-null    float64
 57  Unnamed: 57  369 non-null    float64
 58  Unnamed: 58  369 non-null    float64
 59  Unnamed: 59  369 non-null    float64
```

Saved successfully!

```
 60  Unnamed: 60  369 non-null    float64
 61  Unnamed: 61  369 non-null    float64
 62  Unnamed: 62  369 non-null    float64
 63  Unnamed: 63  369 non-null    float64
 64  Unnamed: 64  369 non-null    float64
 65  Unnamed: 65  369 non-null    float64
 66  Unnamed: 66  369 non-null    float64
 67  Unnamed: 67  369 non-null    float64
 68  Unnamed: 68  369 non-null    float64
 69  Unnamed: 69  369 non-null    float64
 70  Unnamed: 70  369 non-null    float64
 71  Unnamed: 71  369 non-null    float64
 72  Unnamed: 72  369 non-null    float64
 73  Unnamed: 73  369 non-null    float64
 74  Unnamed: 74  369 non-null    float64
 75  Unnamed: 75  369 non-null    float64
 76  Unnamed: 76  369 non-null    float64
 77  Unnamed: 77  369 non-null    float64
 78  Unnamed: 78  369 non-null    float64
 79  Unnamed: 79  369 non-null    float64
 80  Unnamed: 80  369 non-null    float64
 81  Unnamed: 81  369 non-null    float64
 82  Unnamed: 82  369 non-null    float64
 83  Unnamed: 83  369 non-null    float64
 84  Unnamed: 84  369 non-null    float64
 85  Unnamed: 85  369 non-null    float64
 86  Unnamed: 86  369 non-null    float64
 87  Unnamed: 87  369 non-null    float64
dtypes: float64(80), object(8)
memory usage: 253.8+ KB
```

```
df_Out_Analy_wild['Sum_pca'] = df_Out_Analy_wild.iloc[:,8:87].mean(axis=1)

df_Out_Analy_wild.head()
```

Saved successfully!

File Name	Wing	Genotype 1	Genotype 2	sex	ind Name	side	left/right	PCA scores	Unr

```
df_Out_Analy_wild['File Name'] = df_Out_Analy_wild['ind Name'].str[5:9]
df_Out_Analy_wild.drop(df_Out_Analy_wild.iloc[:, 8:87], inplace = True, axis = 1)
```

Dropping the unwanted columns from 8 to 87 and removing the unecessary character from File Name that is id

3	v_Ant_d.tif	Ant	marsaeus	phasiana	f	JM02.0107	v	d	17.00490	f

```
df_Out_Analy_wild.drop('ind Name',inplace=True, axis=1)
df_Out_Analy_wild.rename( {'File Name':'id' } , axis=1 , inplace = True)
df_Out_Analy_wild.drop( {'Wing'} ,axis=1 ,inplace = True)
df_Out_Analy_wild.head()
```

	id	Genotype 1	Genotype 2	sex	side	left/right	Unnamed: 87	Sum_pca
0						axis Variance	0.000009	0.012658
1	0001	numata	elegans	f	v	g	0.051368	0.081010
2	0018	numata	elegans	f	v	d	0.012828	-0.049301
3	0107	marsaeus	phasiana	f	v	d	0.003466	0.287179
4	0108	marsaeus	phasiana	f	v	g	-0.038387	0.104642

Renaming the name column as id and dropping the ind Name column

```
df_Out_Analy_wild = df_Out_Analy_wild.iloc[1:,:]
df_Out_Analy_wild.drop('Unnamed: 87',inplace=True, axis = 1)
```

Saved successfully!

```
/usr/local/lib/python3.7/dist-packages/pandas/core/frame.py:4913: SettingWithCopyWarnin
```

```
df_Out_Analy_wild['sex'] = df_Out_Analy_wild['sex'].str.upper()
df_Out_Analy_wild.head()
```

	id	Genotype 1	Genotype 2	sex	side	left/right	Sum_pca
1	0001	numata	elegans	F	v	g	0.081010
2	0018	numata	elegans	F	v	d	-0.049301
3	0107	marsaeus	phasiana	F	v	d	0.287179
4	0108	marsaeus	phasiana	F	v	g	0.104642
5	0153	marsaeus	phasiana	F	v	g	0.259739

```
df_Out_Analy_wild.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 368 entries, 1 to 368
Data columns (total 7 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   id          368 non-null    object
 1   Genotype 1  368 non-null    object
 2   Genotype 2  368 non-null    object
 3   sex         368 non-null    object
 4   side        368 non-null    object
 5   left/right  368 non-null    object
 6   Sum_pca     368 non-null    float64
dtypes: float64(1), object(6)
memory usage: 20.3+ KB
```

```
df_Out_Analy_wild['id'] = df_Out_Analy_wild['id'].astype('int')
                                                 2','sex','side','left/right']:
                                                 naly_wild[col].astype('category')
```

Saved successfully!

```
df_Out_Analy_wild.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 368 entries, 1 to 368
Data columns (total 7 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   id          368 non-null    int64
 1   Genotype 1  368 non-null    category
 2   Genotype 2  368 non-null    category
 3   sex         368 non-null    category
 4   side        368 non-null    category
 5   left/right  368 non-null    category
 6   Sum_pca     368 non-null    float64
```

```
dtypes: category(5), float64(1), int64(1)
memory usage: 9.1 KB
```

Combining all Datasets

```
df_Measures1.head()
```

	id	sex	genotype	area	ellipse axe 1 length	max axe 1 length	deviation max axe 1 from center	axe 1 Moment 2	axe 1 Mo
0	725	F	a-e	954465	1622.1291	1674	69.7104	967713404464	1166093876
1	725	F	a-e	638157	1120.3231	1100	-59.4572	249024797899	195194465
2	726	F	a-s	984051	1673.3414	1741	89.1505	1069847709180	1336034050
3	726	F	a-s	634452	1145.1990	1138	50.8133	267701750118	216542452
4	735	F	s-s	889416	1602.0627	1649	-123.5292	644725446306	711216237

```
df_brood.tail()
```

	id	sex	genotype	phenotype
77	1029	M	s-s	silvana
				egans
				aurora
80	1037	M	a-e	elegans
81	1052	M	a-e	elegans

Saved successfully!

```
merge_df = pd.merge(df_brood,df_Measures1, on=['id','sex','genotype'], how='left')
merge_df1 = pd.merge(df_brood,df_Measures2, on=['id'], how='left')
merge_df2=pd.merge(df_brood,df_perimeter, on=['id'], how='left')
merge_df3 = pd.merge(df_Measures1,df_perimeter, on=['id'], how='left')
```

```
merge_df4=pd.merge(merge_df,merge_df1, on=['id','genotype','phenotype','area','sex','max axe
merge_df5=pd.merge(merge_df2,merge_df3, on=['id','sex','Perimeter','genotype'], how='left')
```

```
merge_new_all_1=pd.merge(merge_df4,merge_df5,on=['id','sex','phenotype','area','ellipse axe 1
```

```
merge_new_all_1.head()
```

	id	sex	genotype	phenotype	area	ellipse axe 1 length	max axe 1 length	deviation max axe 1 from center	axe 1 Moment 2
0	725	F	a-e	elegans	954465	1622.1291	1674	69.7104	967713404464
1	725	F	a-e	elegans	638157	1120.3231	1100	-59.4572	249024797899
2	726	F	a-s	aurora	984051	1673.3414	1741	89.1505	1069847709180
3	726	F	a-s	aurora	634452	1145.1990	1138	50.8133	267701750118
4	735	F	s-s	silvana	889416	1602.0627	1649	-123.5292	644725446306

5 rows × 24 columns

```
merge_new_all_1.columns
```

```
Index(['id', 'sex', 'genotype', 'phenotype', 'area', 'ellipse axe 1 length',
       'max axe 1 length', 'deviation max axe 1 from center', 'axe 1 Moment 2',
       'axe 1 Moment 3', 'axe 1 Moment 4', 'std axe 1 Moment 2',
       'std axe 1 Moment 3', 'std axe 1 Moment 4', 'ellipse axe 2 length',
       'max axe 2 length', 'deviation max axe 2 from center', 'axe 2 Moment 2',
       'axe 2 Moment 3', 'axe 2 Moment 4', 'std axe 2 Moment 2',
       'std axe 1 Moment 4', 'std axe 2 Moment 4', 'Perimeter'],
      dtype='object')
```

Saved successfully!

```
merge_new_all_1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 164 entries, 0 to 163
Data columns (total 24 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   id                              164 non-null    int64
 1   sex                             164 non-null    category
 2   genotype                        164 non-null    category
 3   phenotype                       164 non-null    category
 4   area                            164 non-null    int64
 5   ellipse axe 1 length            164 non-null    float64
 6   max axe 1 length                164 non-null    int64
```

```
 7   deviation max axe 1 from center  164 non-null    float64
 8   axe 1 Moment 2                   164 non-null    int64
 9   axe 1 Moment 3                   164 non-null    int64
 10  axe 1 Moment 4                   164 non-null    float64
 11  std axe 1 Moment 2               164 non-null    float64
 12  std axe 1 Moment 3               164 non-null    float64
 13  std axe 1 Moment 4               164 non-null    float64
 14  ellipse axe 2 length             164 non-null    float64
 15  max axe 2 length                 164 non-null    int64
 16  deviation max axe 2 from center  164 non-null    float64
 17  axe 2 Moment 2                   164 non-null    int64
 18  axe 2 Moment 3                   164 non-null    int64
 19  axe 2 Moment 4                   164 non-null    int64
 20  std axe 2 Moment 2               164 non-null    float64
 21  std axe 2 Moment 3               164 non-null    float64
 22  std axe 2 Moment 4               164 non-null    float64
 23  Perimeter                        164 non-null    float64
dtypes: category(3), float64(12), int64(9)
memory usage: 29.1 KB
```

```
merge_new_all_1.shape
```

```
(164, 24)
```

```
merge_2_new = pd.merge(df_Wildwings_axis1,df_Wildwings_axis2, on=['id','sex','Genus','Species
```

```
merge_2_new.head()
```

	id	sex	Tribe	Genus	Species	sub-species	area	ellipse axe 1 length	max axe 1 length	deviat max ax f cen
0	1	F	Heliconiinae	Heliconius	numata	elegans	892614.0	1581.4126	1632	124.9
				nius	numata	elegans	889094.0	1567.0531	1625	128.4
2	1	F	Heliconiinae	Heliconius	numata	elegans	562096.0	1047.8688	1057	-2.5
3	1	F	Heliconiinae	Heliconius	numata	elegans	582341.0	1054.8397	1049	10.6
4	1	F	Heliconiinae	Heliconius	numata	elegans	893173.0	1573.8670	1630	125.1

5 rows × 30 columns

Saved successfully!

```
merge_2_new1 = pd.merge(df_Out_Analy,df_Out_Analy_wild,on=['id','sex','Genotype 2','Genotype
```

```
merge_2_new2 = pd.merge(merge_2_new1,df_LandMark,on=['id','sex'],how='left')
```

```
merge_2_new3 = pd.merge(merge_2_new,df_LandMark,on=['id','sex'],how='left')
```

```
merge_new_all_2 = pd.merge(merge_2_new,merge_2_new3,on=['id','sex','Genus','Species','sub-spe
                    ,'deviation max axe 2 from center','axe 1 Moment 2','axe 1 Moment 3','axe
                    'doresal/ventral','Anterior/Posterior','left/right','sub/Family','ellipse
                    'std axe 2 Moment 2','std axe 2 Moment 3','std axe 2 Moment 4','axe 2 Mom
```

```
merge_new_all_2.head()
```

	id	sex	Tribe	Genus	Species	sub-species	area	ellipse axe 1 length	max axe 1 length	deviat max ax f cen
0	1	F	Heliconiinae	Heliconius	numata	elegans	892614.0	1581.4126	1632	124.9
1	1	F	Heliconiinae	Heliconius	numata	elegans	889094.0	1567.0531	1625	128.4
2	1	F	Heliconiinae	Heliconius	numata	elegans	562096.0	1047.8688	1057	-2.5
3	1	F	Heliconiinae	Heliconius	numata	elegans	582341.0	1054.8397	1049	10.6
4	1	F	Heliconiinae	Heliconius	numata	elegans	893173.0	1573.8670	1630	125.1

5 rows × 33 columns

```
merge_new_all_2.info()
```

Saved successfully!

```
Frame'>
Int64Index: 3026 entries, 0 to 3025
Data columns (total 33 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   id                              3026 non-null   int64
 1   sex                             3026 non-null   object
 2   Tribe                           3026 non-null   category
 3   Genus                           3026 non-null   object
 4   Species                         3026 non-null   object
 5   sub-species                     3026 non-null   object
 6   area                            3026 non-null   float64
 7   ellipse axe 1 length            3026 non-null   float64
 8   max axe 1 length                3026 non-null   int64
 9   deviation max axe 1 from center 3026 non-null   float64
 10  axe 1 Moment 2                  3026 non-null   int64
 11  axe 1 Moment 3                  3026 non-null   int64
 12  axe 1 Moment 4                  3026 non-null   int64
```

```
 13  std axe 1 Moment 2              3026 non-null   float64
 14  std axe 1 Moment 3              3026 non-null   float64
 15  std axe 1 Moment 4              3026 non-null   float64
 16  Perimeter                       3026 non-null   float64
 17  doresal/ventral                 4 non-null      category
 18  Anterior/Posterior              4 non-null      category
 19  left/right                      4 non-null      category
 20  sub/Family                      4 non-null      category
 21  ellipse axe 2 length            4 non-null      float64
 22  max axe 2 length                4 non-null      float64
 23  deviation max axe 2 from center 4 non-null      float64
 24  axe 2 Moment 2                  4 non-null      float64
 25  axe 2 Moment 3                  4 non-null      float64
 26  axe 2 Moment 4                  4 non-null      float64
 27  std axe 2 Moment 2              4 non-null      float64
 28  std axe 2 Moment 3              4 non-null      float64
 29  std axe 2 Moment 4              4 non-null      float64
 30  genotype                        8 non-null      category
 31  phenotype                       8 non-null      category
 32  sex-phenotype                   8 non-null      category
dtypes: category(8), float64(16), int64(5), object(4)
memory usage: 639.5+ KB
```

Concating all the data frame

```
Final_All = pd.concat([merge_new_all_1,merge_new_all_2])
```

```
Final_All.head()
```

Saved successfully!

				ype	area	ellipse axe 1 length	max axe 1 length	deviation max axe 1 from center	axe 1 Moment 2
0	725	F	a-e	elegans	954465.0	1622.1291	1674	69.7104	967713404464
1	725	F	a-e	elegans	638157.0	1120.3231	1100	-59.4572	249024797899
2	726	F	a-s	aurora	984051.0	1673.3414	1741	89.1505	1069847709180
3	726	F	a-s	aurora	634452.0	1145.1990	1138	50.8133	267701750118
4	735	F	s-s	silvana	889416.0	1602.0627	1649	-123.5292	644725446306

5 rows × 33 columns

```
print(Final_All.isnull().sum())
```

```
id                                   0
sex                                  0
genotype                          3018
phenotype                         3018
area                                 0
ellipse axe 1 length                 0
max axe 1 length                     0
deviation max axe 1 from center      0
axe 1 Moment 2                       0
axe 1 Moment 3                       0
axe 1 Moment 4                       0
std axe 1 Moment 2                   0
std axe 1 Moment 3                   0
std axe 1 Moment 4                   0
ellipse axe 2 length              3022
max axe 2 length                  3022
deviation max axe 2 from center   3022
axe 2 Moment 2                    3022
axe 2 Moment 3                    3022
axe 2 Moment 4                    3022
std axe 2 Moment 2                3022
std axe 2 Moment 3                3022
std axe 2 Moment 4                3022
Perimeter                            0
Tribe                              164
Genus                              164
Species                            164
sub-species                        164
doresal/ventral                   3186
Anterior/Posterior                3186
left/right                        3186
sub/Family                        3186
sex-phenotype                     3182
dtype: int64
```

Saved successfully!

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 3190 entries, 0 to 3025
Data columns (total 33 columns):
 #   Column                           Non-Null Count  Dtype
---  ------                           --------------  -----
 0   id                               3190 non-null   int64
 1   sex                              3190 non-null   object
 2   genotype                         172 non-null    category
 3   phenotype                        172 non-null    category
 4   area                             3190 non-null   float64
 5   ellipse axe 1 length             3190 non-null   float64
 6   max axe 1 length                 3190 non-null   int64
 7   deviation max axe 1 from center  3190 non-null   float64
 8   axe 1 Moment 2                   3190 non-null   int64
 9   axe 1 Moment 3                   3190 non-null   int64
 10  axe 1 Moment 4                   3190 non-null   float64
```

```
 11  std axe 1 Moment 2              3190 non-null   float64
 12  std axe 1 Moment 3              3190 non-null   float64
 13  std axe 1 Moment 4              3190 non-null   float64
 14  ellipse axe 2 length            168 non-null    float64
 15  max axe 2 length                168 non-null    float64
 16  deviation max axe 2 from center 168 non-null    float64
 17  axe 2 Moment 2                  168 non-null    float64
 18  axe 2 Moment 3                  168 non-null    float64
 19  axe 2 Moment 4                  168 non-null    float64
 20  std axe 2 Moment 2              168 non-null    float64
 21  std axe 2 Moment 3              168 non-null    float64
 22  std axe 2 Moment 4              168 non-null    float64
 23  Perimeter                       3190 non-null   float64
 24  Tribe                           3026 non-null   category
 25  Genus                           3026 non-null   object
 26  Species                         3026 non-null   object
 27  sub-species                     3026 non-null   object
 28  doresal/ventral                 4 non-null      category
 29  Anterior/Posterior              4 non-null      category
 30  left/right                      4 non-null      category
 31  sub/Family                      4 non-null      category
 32  sex-phenotype                   8 non-null      category
dtypes: category(8), float64(17), int64(4), object(4)
memory usage: 674.0+ KB
```

```
Final_All.shape
```

```
(3190, 33)
```

```
data = Final_All.ffill().bfill()
print(data.isnull().sum())
```

```
id                                  0
                                    0
                                    0
                                    0
area                                0
ellipse axe 1 length                0
max axe 1 length                    0
deviation max axe 1 from center     0
axe 1 Moment 2                      0
axe 1 Moment 3                      0
axe 1 Moment 4                      0
std axe 1 Moment 2                  0
std axe 1 Moment 3                  0
std axe 1 Moment 4                  0
ellipse axe 2 length                0
max axe 2 length                    0
deviation max axe 2 from center     0
axe 2 Moment 2                      0
axe 2 Moment 3                      0
axe 2 Moment 4                      0
std axe 2 Moment 2                  0
```

Saved successfully!

```
std axe 2 Moment 3                 0
std axe 2 Moment 4                 0
Perimeter                          0
Tribe                              0
Genus                              0
Species                            0
sub-species                        0
doresal/ventral                    0
Anterior/Posterior                 0
left/right                         0
sub/Family                         0
sex-phenotype                      0
dtype: int64
```

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 3190 entries, 0 to 3025
Data columns (total 33 columns):
 #   Column                            Non-Null Count  Dtype
---  ------                            --------------  -----
 0   id                                3190 non-null   int64
 1   sex                               3190 non-null   object
 2   genotype                          3190 non-null   category
 3   phenotype                         3190 non-null   category
 4   area                              3190 non-null   float64
 5   ellipse axe 1 length              3190 non-null   float64
 6   max axe 1 length                  3190 non-null   int64
 7   deviation max axe 1 from center   3190 non-null   float64
 8   axe 1 Moment 2                    3190 non-null   int64
 9   axe 1 Moment 3                    3190 non-null   int64
 10  axe 1 Moment 4                    3190 non-null   float64
 11  std axe 1 Moment 2                3190 non-null   float64
 12  std axe 1 Moment 3                3190 non-null   float64
 13  std axe 1 Moment 4                3190 non-null   float64
                                       3190 non-null   float64
                                       3190 non-null   float64
                                  center 3190 non-null   float64
 17  axe 2 Moment 2                    3190 non-null   float64
 18  axe 2 Moment 3                    3190 non-null   float64
 19  axe 2 Moment 4                    3190 non-null   float64
 20  std axe 2 Moment 2                3190 non-null   float64
 21  std axe 2 Moment 3                3190 non-null   float64
 22  std axe 2 Moment 4                3190 non-null   float64
 23  Perimeter                         3190 non-null   float64
 24  Tribe                             3190 non-null   category
 25  Genus                             3190 non-null   object
 26  Species                           3190 non-null   object
 27  sub-species                       3190 non-null   object
 28  doresal/ventral                   3190 non-null   category
 29  Anterior/Posterior                3190 non-null   category
 30  left/right                        3190 non-null   category
 31  sub/Family                        3190 non-null   category
 32  sex-phenotype                     3190 non-null   category
```

Saved successfully!

```
dtypes: category(8), float64(17), int64(4), object(4)
memory usage: 674.0+ KB
```

Convering Varibles to datatypes

```
data['sex'] = data['sex'].astype('category')
data['Genus'] = data['Genus'].astype('category')
data['Species'] = data['Species'].astype('category')
data['sub-species'] = data['sub-species'].astype('category')
data['left/right'] = data['left/right'].astype('category')
```

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 3190 entries, 0 to 3025
Data columns (total 33 columns):
 #   Column                            Non-Null Count  Dtype
---  ------                            --------------  -----
 0   id                                3190 non-null   int64
 1   sex                               3190 non-null   category
 2   genotype                          3190 non-null   category
 3   phenotype                         3190 non-null   category
 4   area                              3190 non-null   float64
 5   ellipse axe 1 length              3190 non-null   float64
 6   max axe 1 length                  3190 non-null   int64
 7   deviation max axe 1 from center   3190 non-null   float64
 8   axe 1 Moment 2                    3190 non-null   int64
 9   axe 1 Moment 3                    3190 non-null   int64
 10  axe 1 Moment 4                    3190 non-null   float64
 11  std axe 1 Moment 2                3190 non-null   float64
 12  std axe 1 Moment 3                3190 non-null   float64
 13  std axe 1 Moment 4                3190 non-null   float64
 14  ellipse axe 2 length              3190 non-null   float64
 15  max axe 2 length                  3190 non-null   float64
                                  center 3190 non-null   float64
                                         3190 non-null   float64
                                         3190 non-null   float64
 19  axe 2 Moment 4                    3190 non-null   float64
 20  std axe 2 Moment 2                3190 non-null   float64
 21  std axe 2 Moment 3                3190 non-null   float64
 22  std axe 2 Moment 4                3190 non-null   float64
 23  Perimeter                         3190 non-null   float64
 24  Tribe                             3190 non-null   category
 25  Genus                             3190 non-null   category
 26  Species                           3190 non-null   category
 27  sub-species                       3190 non-null   category
 28  doresal/ventral                   3190 non-null   category
 29  Anterior/Posterior                3190 non-null   category
 30  left/right                        3190 non-null   category
 31  sub/Family                        3190 non-null   category
 32  sex-phenotype                     3190 non-null   category
dtypes: category(12), float64(17), int64(4)
memory usage: 588.2 KB
```

Saved successfully!

```
data.head()
```

	id	sex	genotype	phenotype	area	ellipse axe 1 length	max axe 1 length	deviation max axe 1 from center	axe 1 Moment 2
0	725	F	a-e	elegans	954465.0	1622.1291	1674	69.7104	967713404464
1	725	F	a-e	elegans	638157.0	1120.3231	1100	-59.4572	249024797899
2	726	F	a-s	aurora	984051.0	1673.3414	1741	89.1505	1069847709180
3	726	F	a-s	aurora	634452.0	1145.1990	1138	50.8133	267701750118
4	735	F	s-s	silvana	889416.0	1602.0627	1649	-123.5292	644725446306

5 rows × 33 columns

```
data.to_csv('Repeat_assignment.csv')
```

```
df=pd.read_csv('/content/Repeat_assignment.csv')
df.head()
```

	Unnamed: 0	id	sex	genotype	phenotype	area	ellipse axe 1 length	max axe 1 length	deviation max axe 1 from center	ax
0	0	725	F	a-e	elegans	954465.0	1622.1291	1674	69.7104	967
				-e	elegans	638157.0	1120.3231	1100	-59.4572	249
2	2	726	F	a-s	aurora	984051.0	1673.3414	1741	89.1505	1069
3	3	726	F	a-s	aurora	634452.0	1145.1990	1138	50.8133	267
4	4	735	F	s-s	silvana	889416.0	1602.0627	1649	-123.5292	644

Saved successfully!

5 rows × 34 columns

The sheet 'Landmark coordinates' is not showing not much relevance so its been not used anlaysis

SO we have remove the null values by the forward fill mechanism and change the necessary datatype