```
import numpy as np
import pandas as pd
import random as rnd
import plotly.express as px
import seaborn as sns
sns.set_palette('Set2')
import matplotlib.pyplot as plt
%matplotlib inline
from plotly.subplots import make_subplots
import plotly.graph_objects as go

from sklearn.pipeline import make_pipeline
from sklearn.linear_model import Ridge, Lasso, ElasticNet, LinearRegression
from sklearn.ensemble import GradientBoostingRegressor
from sklearn.preprocessing import LabelEncoder
# from sklearn.preprocessing import LabelEncoder

from sklearn.model_selection import GridSearchCV
from sklearn.exceptions import NotFittedError
from sklearn.metrics import r2_score,mean_absolute_error

from sklearn.linear_model import LinearRegression
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.ensemble import RandomForestRegressor
from sklearn.ensemble import GradientBoostingClassifier
from sklearn.svm import SVR
from sklearn.metrics import accuracy_score
from sklearn.metrics import mean_absolute_error, mean_absolute_percentage_error
from sklearn.metrics import mean_squared_error
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn import preprocessing
from sklearn.naive_bayes import GaussianNB
from sklearn.metrics import accuracy_score
from sklearn.metrics import confusion_matrix
from sklearn.metrics import precision_score
from sklearn.metrics import recall_score
from sklearn.metrics import f1_score
from sklearn.model_selection import GridSearchCV
from sklearn import svm

import warnings
warnings.filterwarnings("ignore", category=DeprecationWarning)


df = pd.read_csv(r'/content/abalone.csv')


df.head()
```

	Sex	Length	Diameter	Height	Whole	Shucked	Viscera	Shell	Rings

✓ 3m 21s completed at 12:58 AM

4	I	0.330	0.255	0.080	0.2050	0.0895	0.0395	0.055	7

```
df.columns
```

```
Index(['Sex', 'Length', 'Diameter', 'Height', 'Whole weight', 'Shucked weig
       'Viscera weight', 'Shell weight', 'Rings'],
      dtype='object')
```

```
# Changing the column names for better readability
newCols = list(map(lambda x : x.lower().replace(' ', '_'), df.columns))

df.columns = newCols
```

```
df.info() # Checking the columns meta deta
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4177 entries, 0 to 4176
Data columns (total 9 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   sex             4177 non-null   object 
 1   length          4177 non-null   float64
 2   diameter        4177 non-null   float64
 3   height          4177 non-null   float64
 4   whole_weight    4177 non-null   float64
 5   shucked_weight  4177 non-null   float64
 6   viscera_weight  4177 non-null   float64
 7   shell_weight    4177 non-null   float64
 8   rings           4177 non-null   int64  
dtypes: float64(7), int64(1), object(1)
memory usage: 293.8+ KB
```

```
# CHecking for null values
df.isnull().sum()
```

```
sex               0
length            0
diameter          0
height            0
whole_weight      0
shucked_weight    0
viscera_weight    0
shell_weight      0
rings             0
```

	length	diameter	height	whole_weight	shucked_weight	visc
count	4177.000000	4177.000000	4177.000000	4177.000000	4177.000000	
mean	0.523992	0.407881	0.139516	0.828742	0.359367	
std	0.120093	0.099240	0.041827	0.490389	0.221963	
min	0.075000	0.055000	0.000000	0.002000	0.001000	
25%	0.450000	0.350000	0.115000	0.441500	0.186000	
50%	0.545000	0.425000	0.140000	0.799500	0.336000	
75%	0.615000	0.480000	0.165000	1.153000	0.502000	
max	0.815000	0.650000	1.130000	2.825500	1.488000	

It is observed that the minimum value for height is zero, which is most probably due to an error while data collection activity. We will investigate this in the next steps and take appropriate measures

```
df[df['height']==0]
```

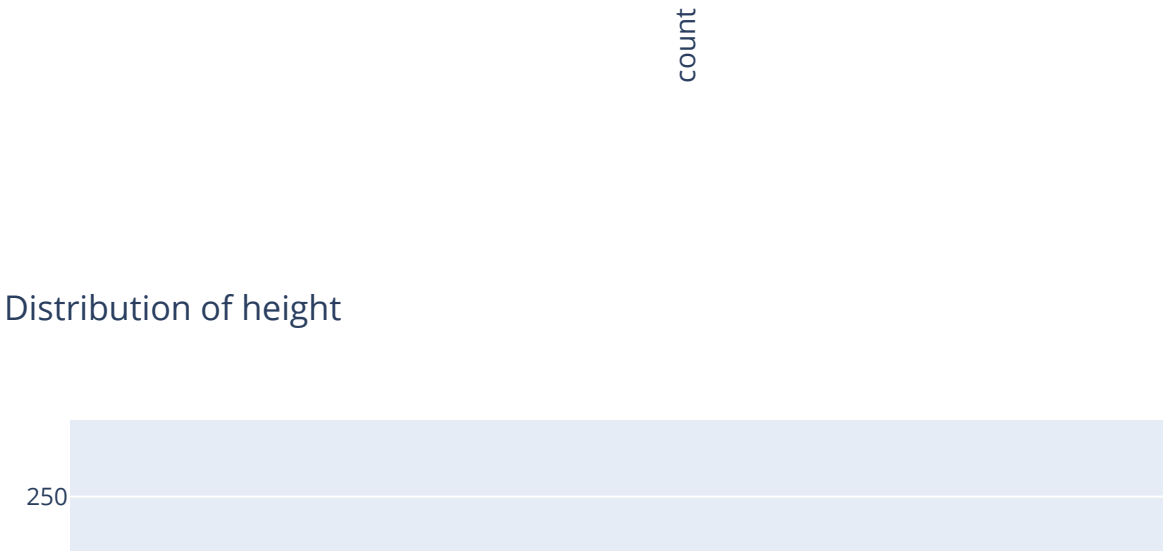
	sex	length	diameter	height	whole_weight	shucked_weight	viscera_w
1257	I	0.430	0.34	0.0	0.428	0.2065	
3996	I	0.315	0.23	0.0	0.134	0.0575	

There are only two records with height as zero this is definitely a error as diameter, wieghts are non zero. They can be dropped or the height can be replaced with median or median after checking the distribution of the height column

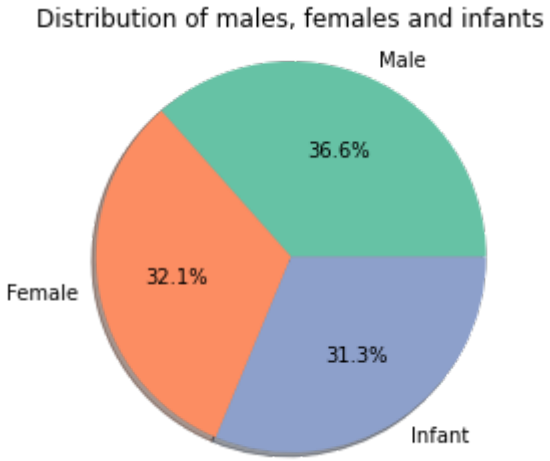
```
df
```

4177 rows × 9 columns

```
# see the distibution of hieghts information
fig = px.histogram(df, x="height", title="Distribution of height" )
fig.show()
```



```
ax1.axis('equal')
plt.title("Distribution of males, females and infants")
plt.show()
```



It is observed that the count of males, females and infants are almost equal, there is no clear domination of a specific gender types in the data

```
fig = make_subplots(rows=1, cols=2, subplot_titles=('Distribution for Height', "
fig.add_trace(go.Histogram(x=df['height']), row=1, col=1)
fig.add_trace(
    go.Violin(y=df['height'], box_visible=True, line_color='black',
                                meanline_visible=True, fillcolor='lightseagreen',
```

After looking at the distribution of the data and the box plot we can observe that the data is nearly normally distributed, has little outliers

```
df.columns
```

```
Index(['sex', 'length', 'diameter', 'height', 'whole_weight', 'shucked_weig
       'viscera_weight', 'shell_weight', 'Age'],
      dtype='object')
```

```
fig = make_subplots(rows=1, cols=2, subplot_titles=('Distribution for lenght', "
fig.add_trace(go.Histogram(x=df['length']), row=1, col=1)
fig.add_trace(
    go.Violin(y=df['length'], box_visible=True, line_color='black',
                               meanline_visible=True, fillcolor='lightseagreen',
                               x0='length')
```

```
df.columns
```

```
Index(['sex', 'length', 'diameter', 'height', 'whole_weight', 'shucked_weig
       'viscera_weight', 'shell_weight', 'Age'],
      dtype='object')
```

```
fig = make_subplots(rows=1, cols=2, subplot_titles=('Distribution for viscera_we
fig.add_trace(go.Histogram(x=df['viscera_weight']), row=1, col=1)
fig.add_trace(
    go.Violin(y=df['viscera_weight'], box_visible=True, line_color='black',
                                meanline_visible=True, fillcolor='lightseagreen',
                                x0='viscera_weight measuremnet')

    , row=1, col=2)
```