

REGRESSION

Get the data and performing data visualization

```
In [2]: import matplotlib
import pylab as plt
import numpy as np
import pandas as pd
```

```
In [3]: from numpy import random
np.random.seed(42) # Calculate standard deviation of column
df = pd.read_csv('OGG_Emission.csv',na_values=['NA'],'?',' ', 'NaN')) #na values may be like NA, ?, ' '
df = df.replace(np.random.permutation(df.index)) #shuffle data
#function of pandas
#Index of pandas, randomly select indices, reindex our index
df.reset_index(inplace=True, drop=True) # Reset index
df[0:15] # Display top five rows
```

	X Coordinate (km)	Y Coordinate (km)	Y Measured Depth (m)	Deviation (deg)	Abandoned (True/False)	Surface- Casing Weight (kg/m)	Production- Casing Size (mm)	Cumulative GAS Prod. (e3m3)	Month Well Spudded	Classification	Emission Rate (m3/day)
0	588.1	476.3	NaN	NaN	True	NaN	1778	695.7	0.0	Non Serious	44.4326
1	62.4	666.9	491.7	NaN	False	35.7	139.7	1.0	0.0	Serious	29.9987
2	534.4	391.8	NaN	15.690192	True	NaN	17.8	2204.9	0.0	Serious	55.4241
3	298.7	583.0	NaN	NaN	True	35.7	139.7	32683.5	0.0	Serious	53.0769
4	513.8	434.9	2598.2	9.273310	False	NaN	114.3	32683.5	843.0	Serious	50.5069

```
In [4]: REMOVE OUTLIERS

def outlier_remove(df, n_name):
    """Pandas rows for a specified column where values are out of +/- n*sd standard deviations
    n : n in the equation n*sd
    name: Column name
    """
    mean=df[name].mean() # Calculate mean of column
    std=df[name].std() # Calculate standard deviation of column
    drop = df.index[(mean - n * sd < df[name]) | (mean + n * sd < df[name])] # vertical line is 0
    df.drop(drop, inplace=True, drop=True) # dropping rows that dont satisfy the code
    df.reset_index(inplace=True, drop=True) # Reset index

# Drop outliers in last column 'Oil Prod. (e3m3/month)'
outlier_remove(df, n=2.5, name='Emission Rate (m3/day)') #based on oil production
df.describe()
```

	X Coordinate (km)	Y Coordinate (km)	Y Measured Depth (m)	Deviation (deg)	Abandoned (True/False)	Surface- Casing Weight (kg/m)	Production- Casing Size (mm)	Cumulative GAS Prod. (e3m3)	Month Well Spudded	Emission Rate (m3/day)
count	1478.000000	1478.000000	1056.000000	552.000000	1122.000000	1371.000000	1778.000000	5.896969e+03	178.505007	50.291586
mean	444.461367	489.886232	1424.863636	31.367534	41.972282	144.076565	139.7	1.144858e+05	0.0	29.998576
std	174.814773	219.042216	674.915439	19.010916	12.636631	26.083016	21.144858e+05	249.017701	9.362271	55.424137
min	10.900000	10.000000	158.600000	0.231718	13.700000	73.000000	1.000000e+00	0.000000	25.516553	
25%	336.475000	369.575000	675.800000	14.534334	35.700000	114.300000	5.752000e+02	0.000000	50.281985	
50%	509.850000	494.000000	1104.850000	31.641937	35.700000	139.700000	5.077600e+03	0.000000	50.280065	
75%	592.500000	595.275000	1922.625000	44.343328	48.100000	177.800000	3.268350e+04	322.750000	56.625944	
max	649.600000	1186.100000	5418.900000	76.894237	107.200000	244.500000	2.273275e+06	1264.000000	74.168289	

The outliers have been removed since the number of the rows have reduced.

```
In [5]: REGRESSION MODELLING

df_reg=df.copy()
df_reg.drop(['Classification'], axis=1, inplace=True) # we will use it for regression
df_reg[0:5]
```

	X Coordinate (km)	Y Coordinate (km)	Y Measured Depth (m)	Deviation (deg)	Abandoned (True/False)	Surface- Casing Weight (kg/m)	Production- Casing Size (mm)	Cumulative GAS Prod. (e3m3)	Month Well Spudded	Emission Rate (m3/day)
0	588.1	476.3	NaN	NaN	True	NaN	1778	695.7	0.0	44.432680
1	62.4	666.9	491.7	NaN	False	35.7	139.7	1.0	0.0	29.998576
2	534.4	391.8	NaN	15.690192	True	NaN	17.8	2204.9	0.0	55.424137
3	298.7	583.0	NaN	NaN	True	35.7	139.7	32683.5	0.0	53.076944
4	513.8	434.9	2598.2	9.273310	False	NaN	114.3	32683.5	843.0	50.506939

```
In [6]: STRATIFIED SAMPLING FOR EVEN DISTRIBUTION

df_reg['emission_rate_cat']=np.ceil(df_reg['Emission Rate (m3/day)']/17) # 10 leads to most data in category
# 20 % is test and put some random # otherwise data will change
df_reg['emission_rate_cat']=df_reg['emission_rate_cat'] < 4, 4, inplace=True #anything more than 4 as
#any number will work, 4 is generally chosen, bigger the # more the computation time
df_reg[0:4]
```

	X Coordinate (km)	Y Coordinate (km)	Y Measured Depth (m)	Deviation (deg)	Abandoned (True/False)	Surface- Casing Weight (kg/m)	Production- Casing Size (mm)	Cumulative GAS Prod. (e3m3)	Month Well Spudded	Emission Rate (m3/day)
0	588.1	476.3	NaN	NaN	True	NaN	1778	695.7	0.0	44.432680
1	62.4	666.9	491.7	NaN	False	35.7	139.7	1.0	0.0	29.998576
2	534.4	391.8	NaN	15.690192	True	NaN	17.8	2204.9	0.0	55.424137
3	298.7	583.0	NaN	NaN	True	35.7	139.7	32683.5	0.0	53.076944

```
In [8]: def histplt (val,bins,title,xlab,ylim,axt=None): #python function
""" Function for histogram plotting"""
from matplotlib.offsetbox import AnchoredText

ax1 = axt or plt.axes() # left or right plot
val=np.array(val) # values in the plot
plt.hist(val, bins=bins,ec='black')
n=len(val) # number of values in the plot
Mean=np.mean(val)
SD=np.sqrt(np.var(val))
Mx=np.amax(val)
Min=np.amin(val)

txt="n=%1.0f \n Sd=%1.0f \n S1sigma=%1.0f \n S2sigma=%1.0f \n S3sigma=%1.0f \n S4sigma=%1.0f \n S5sigma=%1.0f \n S6sigma=%1.0f \n S7sigma=%1.0f \n S8sigma=%1.0f \n S9sigma=%1.0f \n S10sigma=%1.0f"
ax1.add_artist(AnchoredText(txt %(n,Mean,SD,Mx,Min,loc=0)))
ax1.add_artist(AnchoredText(txt %(n,Mean,SD,Mx,Min,loc=0)))
plt.title(title,fontsize=font['size']*1.25)
plt.xlabel(xlab,fontsize=font['size'])
plt.ylabel('Frequency',fontsize=font['size'])
plt.xlim(ylim)
ax1.grid(linewidth=0.55)
```


```
In [9]: # Plot histogram of categories of entire
font = {'size' : 14}
matplotlib.rc('font', **font)

fig = plt.subplots(figsize=(6, 3), dpi=100, facecolor='w', edgecolor='k')
```

```
df_reg['emission_rate_cat'].hist(bins=15)
plt.title("Number of Instances per Class")
plt.xlabel("Regression Labels")
plt.ylabel("Frequency")
plt.xlim(0.68,4.3)
```

```
#histplt(df_reg['emission_rate_cat'],bins=15,title='Number of Instances per Class',xlab='Regression Labels',
#       xlim=(0.68,4.3)) # since x or number of categories is from 0 to 4
```

Out[9]: Text(0, 0.5, 'Frequency')



Entire data has been split into 4 category where the split is equally distributed in two categories.

```
In [10]: df_reg['emission_rate_cat'].value_counts()
```

Out[10]: 3.0 726
4.0 691
2.0 61
Name: emission_rate_cat, dtype: int64

COMMENT

The distribution of data has the following regression labels:

Regression Label 1 - 0% distribution

Regression Label 2 - 4.1% distribution

Regression Label 3 - 49.1% distribution

Regression Label 4 - 46.8% distribution

```
In [11]: CATEGORIES HAVE CREATED TO SPLIT DATA INTO TEST AND TRAIN

SPLITTING TRAINING AND TESTING DATA BASED ON EMISSIONS RATE CATEGORY CREATED
```

```
In [11]: from sklearn.model_selection import StratifiedShuffleSplit #from sklearn used to divide data into data and
spt = StratifiedShuffleSplit(n_splits=1, test_size=0.2, random_state=42)
train_idx, test_idx = spt.split(df_reg, df_reg['emission_rate_cat']) #split ratio of this data and in
train_set_strat = df_reg.loc[train_idx]
test_set_strat = df_reg.loc[test_idx]
```

```
In [12]: train_set_strat
```

	X Coordinate (km)	Y Coordinate (km)	Y Measured Depth (m)	Deviation (deg)	Abandoned (True/False)	Surface- Casing Weight (kg/m)	Production- Casing Size (mm)	Cumulative GAS Prod. (e3m3)	Month Well Spudded	Emission Rate (m3/day)
678	294.3	548.7	2504.7	47.479024	False	NaN	1778	84776.5	0.0	51.237935
469	611.4	448.0	870.7	19.626487	False	48.7	1778	10270.5	0.0	45.630606
1327	407.2	176.2	3269.1	NaN	True	35.7	139.7	23327.1	533.0	64.327602
96	480.2	478.5	1102.3	NaN	True	NaN	114.3	138.0	0.0	47.891883
562	379.7	638.8	1953.6	48.966279	True	35.7	114.3	59959.4	616.0	54.767542
...
992	385.7	615.9	603.3	NaN	True	NaN	114.3	54822.2	323.0	61.370796
1156	604.5	483.0	610.0	NaN	True	53.6	1778	32683.5	0.0	52.178626
855	601.1	517.9	NaN	NaN	True	48.1	1778	851.4	0.0	61.319530
958	348.7	697.9	600.2	NaN	True	81.1	1778	15003.3	438.0	64.793491
487	502.6	1.4	NaN	NaN	True	35.7	114.3	76.3	0.0	43.892625

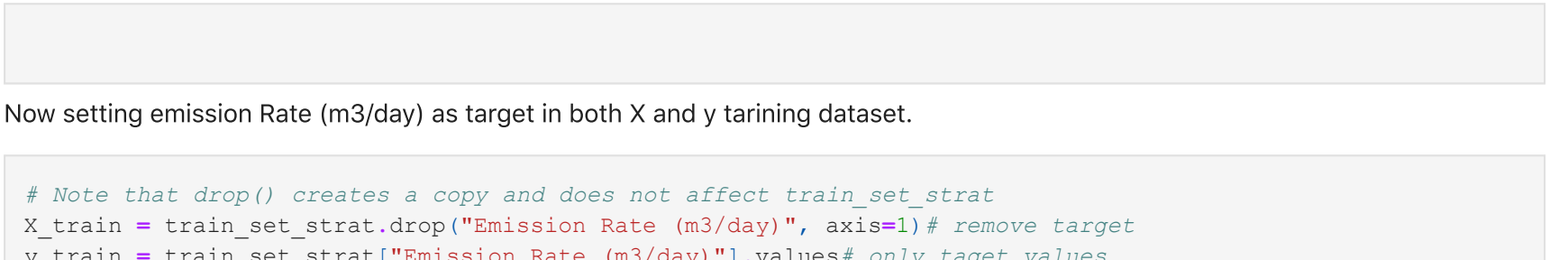
1182 rows x 11 columns

```
In [13]: font = {'size' : 10}
matplotlib.rc('font', **font)

fig = plt.subplots(figsize=(10, 3), dpi=100, facecolor='w', edgecolor='k')
```

```
ax1=plt.subplot(1,2,1)
train_set_strat['emission_rate_cat'].hist(bins=15)
plt.title("Training Set")
plt.xlabel("Regression Labels")
plt.ylabel("Frequency")

ax2=plt.subplot(1,2,2)
test_set_strat['emission_rate_cat'].hist(bins=15)
plt.title("Test Set")
plt.xlabel("Regression Labels")
plt.ylabel("Frequency")
plt.show()
```



TRAINING DATA STRATIFIED SAMPLING

```
In [14]: train_set_strat['emission_rate_cat'].value_counts()
```

Out[14]: 3.0 590
4.0 553
2.0 49
Name: emission_rate_cat, dtype: int64

COMMENT

The distribution of training data has the following regression labels:

Regression Label 1 - 0% distribution

Regression Label 2 - 4.1% distribution

Regression Label 3 - 49.1% distribution

Regression Label 4 - 46.8% distribution

```
In [15]: TEST DATA STRATIFIED SAMPLING

test_set_strat['emission_rate_cat'].value_counts()
```

Out[15]: 3.0 146
4.0 138
2.0 49
Name: emission_rate_cat, dtype: int64

The distribution of testing data has the following regression labels:

Regression Label 1 - 0% distribution

Regression Label 2 - 4.1% distribution

Regression Label 3 - 49.3% distribution

Regression Label 4 - 46.6% distribution

The distribution of training and testing data is almost the same as the original data.

```
In [16]: Removing the emission_rate_cat

The column - emission_rate_cat that was created for stratified sampling has now been deleted as it is no longer needed
```

```
In [16]: for dataset in (train_set_strat, test_set_strat):
dataset.drop('emission_rate_cat', axis=1, inplace=True)
dataset.reset_index(inplace=True, drop=True) # Reset index
```

```
In [17]: train_set_strat
```

	X Coordinate (km)	Y Coordinate (km)	Y Measured Depth (m)	Deviation (deg)	Abandoned (True/False)	Surface- Casing Weight (kg/m)	Production- Casing Size (mm)	Cumulative GAS Prod. (e3m3)	Month Well Spudded	Emission Rate (m3/day)
0	294.3	548.7	2504.7	47.479024	False	NaN	1778	84776.5	0.0	51.237935
1	611.4	448.0	870.7	19.626487	False	48.7	1778	10270.5	0.0	45.630606
2	407.2	176.2	3269.1	NaN	True	35.7	139.7	23327.1	533.0	64.327602
3	480.2	478.5	1102.3	NaN	True	NaN	114.3	138.0	0.0	47.891883
4	379.7	638.8	1953.6	48.966279	True	35.7	114.3	59959.4	616.0	54.767542
...
1177	385.7	615.9	603.3	NaN	True	NaN	114.3	54822.2	323.0	61.370796
1178	604.5	483.0	610.0	NaN	True	53.6	1778	32683.5	0.0	52.178626
1179	601.1	517.9	NaN	NaN	True	48.1	1778	851.4	0.0	61.319530
1180	348.7	697.9	600.2	NaN	True	81.1	1778	15003.3	438.0	64.793491
1181	502.6	1.4	NaN	NaN	True	35.7	114.3	76.3	0.0	43.892625

1182 rows x 10 columns

```
In [17]: Now setting emission rate (m3/day) as target in both X and y training dataset.
```

```
In [18]: # Note that drop() creates a copy and does not affect train set strat
X_train = train_set_strat.drop("Emission Rate (m3/day)", axis=1, inplace=True) # remove target
y_train = train_set_strat["Emission Rate (m3/day)"].values # only target values
```

```
In [19]: X_train
```

	X Coordinate (km)	Y Coordinate (km)	Y Measured Depth (m)	Deviation (deg)	Abandoned (True/False)	Surface-Casing Weight (kg/m)	Production- Casing Size (mm)	Cumulative GAS Prod. (e3m3)	Month Well Spudded
0	294.3	548.7	2504.7	47.479024	False	NaN	1778	84776.5	0.0
1	611.4	448.0	870.7	19.626487	False	48.7	1778	10270.5	0.0
2	407.2	176.2	3269.1	NaN	True	35.7	139.7	23327.1	533.0
3	480.2	478.5	1102.3	NaN	True	NaN	114.3	138.0	0.0
4	379.7	638.8	1953.6	48.966279	True	35.7	114.3	59959.4	616.0
...
1177	385.7	615.9	603.3	NaN	True	NaN	114.3	54822.2	323.0
1178	604.5	483.0	610.0	NaN	True	53.6	1778	32683.5	0.0
1179	601.1	517.9	NaN	NaN	True	48.1	1778	851.4	0.0
1180	348.7	697.9	600.2	NaN	True	81.1	1778	15003.3	438.0
1181	502.6	1.4	NaN	NaN	True	35.7	114.3	76.3	0.0

1182 rows x 9 columns

```
In [20]: y_train
```

Out[20]: array([51.23793486, 45.63060645, 64.32760219, ..., 61.31952954, 64.79349141, 43.89262681])

The target column -Emission Rate (m3/day) has been removed from the X_train dataset The target column -Emission Rate (m3/day) has been set as the y_train dataset

DATA VISUALIZATION

Converting all text to number in the data.

```
In [72]: X_train['Abandoned (True/False)']=X_train['Abandoned (True/False)'].replace(False, 0) # Replace False with 0
X_train['Abandoned (True/False)']=X_train['Abandoned (True/False)'].replace(True, 1)
```

```
In [72]: X_train
```

	X Coordinate (km)	Y Coordinate (km)	Y Measured Depth (m)	Deviation (deg)	Abandoned (True/False)	Surface-Casing Weight (kg/m)	Production- Casing Size (mm)	Cumulative GAS Prod. (e3m3)	Month Well Spudded
0	294.3	548.7	2504.7	47.479024	0.0	NaN	1778	84776.5	0.0
1	611.4	448.0	870.7	19.626487	0.0	48.7	1778	10270.5	0.0
2	407.2	176.2	3269.1	NaN	1.0	35.7	139.7	23327.1	533.0
3	480.2	478.5	1102.3	NaN	1.0	NaN	114.3	138.0	0.0
4	379.7	638.8	1953.6	48.966279	1.0	35.7	114.3	59959.4	616.0
...
1177	385.7	615.9	603.3	NaN	1.0	NaN	114.3	54822.2	323.0
1178	604.5	483.0	610.0	NaN	1.0	NaN	114.3	138.0	0.0

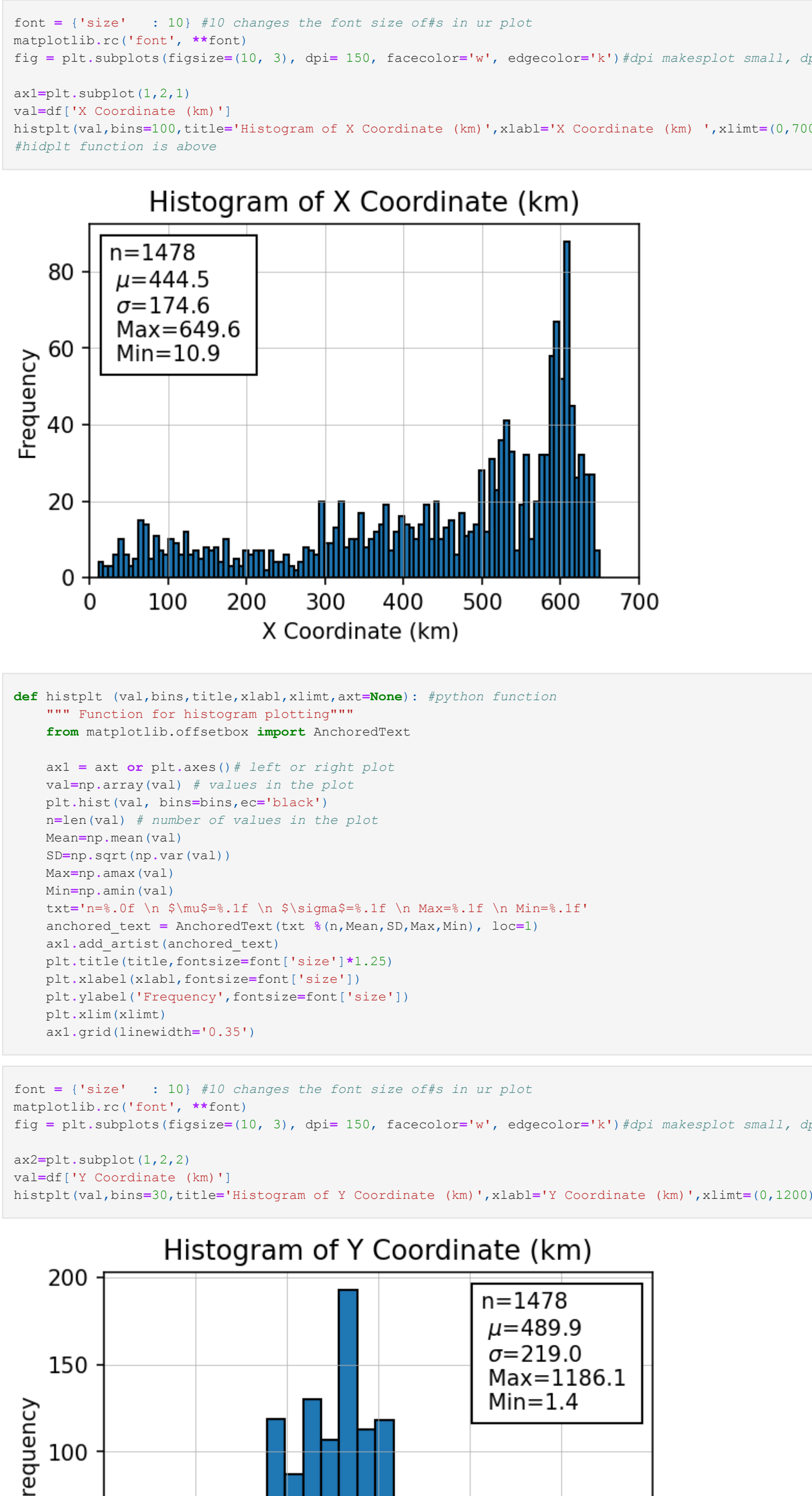
anchored_text = AnchoredText(txt = '%f', loc=2)

ax1.grid(True)

plt.xlabel(xlabel,fontsize=font['size'],

plt.ylabel(ylabel,fontsize=font['size'])

ax1.grid(linewidth=0.35)



Even though both the graphs have the same number of data points, X Coordinate has a lower mean and lower standard deviation as compared to Y-axis. The values for the Y coordinate histogram are higher than X coordinate. X coordinate histogram is skewed towards right. i.e. depicts as negatively skewed distribution. On the contrary, Y coordinate histogram depicts a uniform distribution.

In []:

IMPUTATION

Applying imputation to missing values in our dataset. Note that imputation is not applied to text data, or to target data. The target data has already been removed so we need to remove the text variable=Abandoned (True/False) from the dataset.

Column=Abandoned (True/False) has been removed from the dataset.

In [36]:

```
d2=d1.dropna()
```

Out [36]:

	X Coordinate (km)	Y Coordinate (km)	Measured Depth (m)	Deviation (deg)	Surface-Casing Weight (kg/m)	Production-Casing Size (mm)	Cumulative GAS Prod. (cm3)	Month Well Spudded
0	294.3	548.7	250.7	47.479024	NaN	1778	84776.5	0.0
1	611.4	448.0	870.7	19.626487	4.87	1778	10270.5	0.0
2	407.2	176.2	326.1	NaN	35.7	139.7	23327.1	5330
3	4802.1	767.6	595.3	32.981093	35.7	114.3	138.0	0.0
4	379.7	638.8	1953.6	48.966279	35.7	114.3	59959.4	6160
...
1177	385.7	615.9	603.3	NaN	NaN	114.3	54822.2	3230
1178	604.5	483.0	610.0	NaN	53.6	1778	32683.5	0.0
1179	601.1	517.9	NaN	NaN	48.1	1778	851.4	0.0
1180	348.7	697.9	600.2	NaN	81.1	1778	15003.3	4380
1181	502.8	1.4	NaN	NaN	35.7	114.3	76.3	0.0

All variables except X Coordinate (km), Y Coordinate (km) and Month Well Spudded need to be imputed.

Imputation has been applied on the entire dataset (not just on the columns that are missing values as it does not matter. Also median has been chosen as the imputation measurement. Applying median on all the numerical values. We will apply this same imputation on the target dataset.

In [38]:

```
from sklearn.impute import SimpleImputer
```

In [39]:

```
input_mdn=SimpleImputer(missing_values='NaN', strategy='median')
input_mdn.fit(d2)
```

Out [39]:

```
array([[ 500.8, ..., 493.4, ..., 33.17004428, ..., 110.49, ..., 0.],
       [ 35.7, ..., 139.7, ..., 5005.54, ..., 0.],
       ...,
       [ 35.7, ..., 139.7, ..., 5005.54, ..., 0.]])
```

In [40]:

```
im=SimpleImputer(missing_values='NaN', strategy='median')
```

Out [40]:

```
im
```

Out [41]:

```
d2=im.fit_transform(d2)
d2
```

In [42]:

```
X_train=im.get_data(d2)
X_train
```

Out [42]:

	X Coordinate (km)	Y Coordinate (km)	Measured Depth (m)	Deviation (deg)	Surface-Casing Weight (kg/m)	Production-Casing Size (mm)	Cumulative GAS Prod. (cm3)	Month Well Spudded
0	294.3	548.7	250.7	47.479024	NaN	1778	84776.5	0.0
1	611.4	448.0	870.7	19.626487	4.87	1778	10270.5	0.0
2	407.2	176.2	326.1	33.710044	35.7	139.7	23327.1	5330
3	4802.1	767.6	595.3	33.710044	35.7	114.3	138.0	0.0
4	379.7	638.8	1953.6	48.966279	35.7	114.3	59959.4	6160
...
1177	385.7	615.9	603.3	33.710044	35.7	114.3	54822.2	3230
1178	604.5	483.0	610.0	33.710044	53.6	1778	32683.5	0.0
1179	601.1	517.9	NaN	33.710044	48.1	1778	851.4	0.0
1180	348.7	697.9	600.2	33.710044	81.1	1778	15003.3	4380
1181	502.8	1.4	NaN	33.710044	35.7	114.3	76.3	0.0

1182 rows x 8 columns

Column=Abandoned (True/False) has been removed from the d2 dataset.

In [37]:

```
<class 'pandas.core.frame.DataFrame'>
d2.info()
d2.columns
d2.dtypes
```

Out [37]:

```
<class 'pandas.core.frame.DataFrame'>
d2.info()
d2.columns
d2.dtypes
```

CONCLUSION: All the values have been replaced with median

In []:

ONE HOT ENCODING THE NUMERICAL COLUMN - Abandoned (True/False)

One Hot Encoding has been selected for the Abandoned (True/False) column as True and False.

In [44]:

```
Abandoned_cat=X_train['Abandoned (True/False)']
Abandoned_cat
```

Out [44]:

	Abandoned (True/False)
0	0.0
1	0.0
2	1.0
3	1.0
4	1.0
...	...
1177	1.0
1178	0.0
1179	0.0
1180	1.0
1181	0.0

In [45]:

```
from sklearn.preprocessing import OneHotEncoder
```

Out [45]:

```
OneHotEncoder
```

In []:

STANDARDIZATION OF IMPUTED VARIABLES

In [46]:

```
from sklearn.preprocessing import StandardScaler
```

Out [46]:

```
StandardScaler
```

In []:

Out [47]:

```
X_train=im.get_data(d2)
X_train
```

Out [48]:

```
X_train
```

The mean is 0 and the standard deviation is 1 for the new X_train dataset. So, the X_train dataset is now standardized.

In [49]:

```
X_train=X_train/np.linalg.norm(X_train,axis=0)
X_train
```

Out [49]:

	X Coordinate (km)	Y Coordinate (km)	Measured Depth (m)	Deviation (deg)	Surface-Casing Weight (kg/m)	Production-Casing Size (mm)	Cumulative GAS Prod. (cm3)	Month Well Spudded
0	-0.834877	0.270139	1.452185	1.283786	-0.410919	1.395937	0.137200	-0.732085
1	0.972239	-0.194731	-0.455259	-0.151244	-0.410919	1.395937	-0.271716	-0.732085
2	-0.191473	-1.445381	2.401098	0.32812	-0.410919	-0.193713	-0.155025	1.470060
3	0.224545	-0.053162	-0.286945	0.32812	-0.410919	-1.163480	-0.265301	-0.732085
4	-0.343193	0.680807	0.769143	1.413811	-0.410919	-1.163480	0.019181	1.743640
...
1177	-0.331999	0.579623	-0.059584	0.32812	-0.410919	-1.163480	-0.005249	0.560603
1178	0.932916	-0.032457	-0.897672	0.32812	1.181660	1.395937	-0.110530	-0.732085
1179	0.931540	0.129268	-0.032457	0.32812	0.692251	1.395937	-0.261908	-0.732085
1180	-0.524588	0.957469	-0.908630	0.32812	3.628106	1.395937	-0.194609	1.028251
1181	0.352200	-2.250409	-0.283720	0.32812	-0.410919	-1.163480	-0.265594	-0.732085

1182 rows x 8 columns

In []:

CONCATENATING OF STANDARDIZED DATA AND ONE-HOT ENCODING column

In [50]:

```
concatenated=X_train[['X_train', 'Y_train', 'Measured Depth', 'Deviation', 'Surface-Casing Weight', 'Production-Casing Size', 'Cumulative GAS Prod.', 'Month Well Spudded']]
concatenated
```

Out [50]:

	X Coordinate (km)	Y Coordinate (km)	Measured Depth (m)	Deviation (deg)	Surface-Casing Weight (kg/m)	Production-Casing Size (mm)	Cumulative GAS Prod. (cm3)	Month Well Spudded
0	-0.834877	0.270139	1.452185	1.283786	-0.410919	1.395937	0.137200	-0.732085
1	0.972239	-0.194731	-0.455259	-0.151244	-0.410919	1.395937	-0.271716	-0.732085
2	-0.191473	-1.445381	2.401098	0.32812	-0.410919	-0.193713	-0.155025	1.470060
3	0.224545	-0.053162	-0.286945	0.32812	-0.410919	-1.163480	-0.265301	-0.732085
4	-0.343193	0.680807	0.769143	1.413811	-0.410919	-1.163480	0.019181	1.743640
...
1177	-0.331999	0.579623	-0.059584	0.32812	-0.410919	-1.163480	-0.005249	0.560603
1178	0.932916	-0.032457	-0.897672	0.32812	1.181660	1.395937	-0.110530	-0.732085
1179	0.931540	0.129268	-0.032457	0.32812	0.692251	1.395937	-0.261908	-0.732085
1180	-0.524588	0.957469	-0.908630	0.32812	3.628106	1.395937	-0.194609	1.028251
1181	0.352200	-2.250409	-0.283720	0.32812	-0.410919	-1.163480	-0.265594	-0.732085

1182 rows x 8 columns

In []:

Model Training and Evaluation

LINEAR REGRESSION

In [63]:

```
from sklearn.linear_model import LinearRegression
```

Out [63]:

```
LinearRegression
```

In [64]:

```
neg_mean_squared_error = np.sqrt(np.mean(neg_mean_squared_error))
```

Out [64]:

```
0.226802076502988
```

Root mean squared error is 0.22 for the Linear Regression

In []:

SUPPORT VECTOR MACHINE-POLYNOMIAL KERNEL Fine-tune 'degree', 'coef0', 'C'

In [66]:

```
from sklearn.model_selection import RandomizedSearchCV
```

Out [66]:

```
RandomizedSearchCV
```

In []:

Out [67]:

```
cv_results_
```

Out [68]:

	X Coordinate (km)	Y Coordinate (km)	Measured Depth (m)	Deviation (deg)	Surface-Casing Weight (kg/m)	Production-Casing Size (mm)	Cumulative GAS Prod. (cm3)	Month Well Spudded
0	-0.834877	0.270139	1.452185	1.283786	-0.410919	1.395937	0.137200	-0.732085
1	0.972239	-0.194731	-0.455259	-0.151244	-0.410919	1.395937	-0.271716	-0.732085
2	-0.191473	-1.445381	2.401098	0.32812	-0.410919	-0.193713	-0.155025	1.470060
3	0.224545	-0.053162	-0.286945	0.32812	-0.410919	-1.163480	-0.265301	-0.732085
4	-0.343193	0.680807	0.769143	1.413811	-0.410919	-1.163480	0.019181	1.743640
...
1177	-0.331999	0.579623	-0.059584	0.32812	-0.410919	-1.163480	-0.005249	0.560603
1178	0.932916	-0.032457	-0.897672	0.32812	1.181660	1.395937	-0.110530	-0.732085
1179	0.931540	0.129268	-0.032457	0.32812	0.692251	1.395937	-0.261908	-0.732085
1180	-0.524588	0.957469	-0.908630	0.32812	3.628106	1.395937	-0.194609	1.028251
1181	0.352200	-2.250409	-0.283720	0.32812	-0.410919	-1.163480	-0.265594	-0.732085

1182 rows x 8 columns

In []:

Decision Tree - Fine-tune 'max_depth','min_samples_split','min_samples_leaf'

In [71]:

```
from sklearn.tree import DecisionTreeRegressor
```

Out [71]:

```
DecisionTreeRegressor
```

In []:

Out [72]:

	X Coordinate (km)	Y Coordinate (km)	Measured Depth (m)	Deviation (deg)	Surface-Casing Weight (kg/m)	Production-Casing Size (mm)	Cumulative GAS Prod. (cm3)	Month Well Spudded
0	-0.834877	0.270139	1.452185	1.283786	-0.410919	1.395937	0.137200	-0.732085
1	0.972239	-0.194731	-0.455259	-0.151244	-0.410919	1.395937	-0.271716	-0.732085
2	-0.191473	-1.445381	2.401098	0.32812	-0.410919	-0.193713	-0.155025	1.470060
3	0.224545	-0.053162	-0.286945	0.32812	-0.410919	-1.163480	-0.265301	-0.732085
4	-0.343193	0.680807	0.769143	1.413811	-0.410919	-1.163480	0.019181	1.743640
...
1177	-0.331999	0.579623	-0.059584	0.32812	-0.410919	-1.163480	-0.005249	0.560603
1178	0.932916	-0.032457	-0.897672	0.32812	1.181660	1.395937	-0.110530	-0.732085
1179	0.931540	0.129268	-0.032457	0.32812	0.692251	1.395937	-0.261908	-0.732085
1180	-0.524588	0.957469	-0.908630	0.32812	3.628106	1.395937	-0.194609	1.028251
1181	0.352200	-2.250409	-0.283720	0.32812	-0.410919	-1.163480	-0.265594	-0.732085

1182 rows x 8 columns

In []:

Model Training and Evaluation

LINEAR REGRESSION

In [63]:

```
from sklearn.linear_model import LinearRegression
```

Out [63]:

```
LinearRegression
```

In [64]:

```
neg_mean_squared_error = np.sqrt(np.mean(neg_mean_squared_error))
```

Out [64]:

```
0.226802076502988
```

Root mean squared error is 0.22 for the Linear Regression

In []:

SUPPORT VECTOR MACHINE-POLYNOMIAL KERNEL Fine-tune 'degree',

[illegible][illegible]

```

7.438442646963147" bootstrap": True, "max_depth": 48, "min_samples_split": 2, "n_estimators": 4093
7.435555552707754 "bootstrap": True, "max_depth": 33, "min_samples_split": 2, "n_estimators": 4681
7.434142789553511 "bootstrap": True, "max_depth": 19, "min_samples_split": 3, "n_estimators": 863)
7.451186670507728 "bootstrap": True, "max_depth": 48, "min_samples_split": 3, "n_estimators": 503)
7.410313962082824 "bootstrap": True, "max_depth": 22, "min_samples_split": 5, "n_estimators": 229)
7.401945819271332 "bootstrap": True, "max_depth": 75, "min_samples_split": 7, "n_estimators": 122)
7.37852040366222 "bootstrap": True, "max_depth": 38, "min_samples_split": 7, "n_estimators": 836)
7.38446748849038 "bootstrap": True, "max_depth": 16, "min_samples_split": 0, "n_estimators": 200)
7.388446748849038 "bootstrap": True, "max_depth": 37, "min_samples_split": 7, "n_estimators": 348)
7.387025049997448 "bootstrap": True, "max_depth": 50, "min_samples_split": 8, "n_estimators": 928)
nan "bootstrap": False, "max_depth": 39, "min_samples_split": 1, "n_estimators": 871)
7.411869624866248 "bootstrap": True, "max_depth": 44, "min_samples_split": 4, "n_estimators": 517)
7.347994862731823 "bootstrap": True, "max_depth": 16, "min_samples_split": 9, "n_estimators": 623)
7.313693470612244 "bootstrap": True, "max_depth": 11, "min_samples_split": 8, "n_estimators": 472)
nan "bootstrap": True, "max_depth": 60, "min_samples_split": 0, "n_estimators": 1)
nan "bootstrap": False, "max_depth": 39, "min_samples_split": 0, "n_estimators": 967)
nan "bootstrap": True, "max_depth": 39, "min_samples_split": 1, "n_estimators": 890)
nan "bootstrap": False, "max_depth": 82, "min_samples_split": 1, "n_estimators": 821)
7.403662377931626 "bootstrap": True, "max_depth": 51, "min_samples_split": 5, "n_estimators": 775)
7.3668958747192466 "bootstrap": True, "max_depth": 80, "min_samples_split": 8, "n_estimators": 995)
nan "bootstrap": True, "max_depth": 39, "min_samples_split": 0, "n_estimators": 967)
nan "bootstrap": True, "max_depth": 27, "min_samples_split": 1, "n_estimators": 276)
nan "bootstrap": False, "max_depth": 99, "min_samples_split": 0, "n_estimators": 452)
nan "bootstrap": True, "max_depth": 68, "min_samples_split": 0, "n_estimators": 511)

```

```

In [95]:
clf = RandomForestRegressor(bootstrap=True, max_depth=11, min_samples_split=8, n_estimators=472,
                           random_state=X_train_new_y_train)

neg_rmse=cross_val_score(X_train_new_y_train, cv=5, scoring="neg_mean_squared_error")
rf_rmse = np.sqrt(-neg_rmse)
rf_rmse = np.mean(rf_rmse_scores)
rf_rmse

Out[95]:
7.30416219410449

```

The RMSE values of RandomForest classifier of 7.3 is lower than decision tree

```

In [96]:

```

ENSEMBLE TECHNIQUES

Gradient Boosting. Fine-tune 'n_estimators', 'learning_rate', 'algorithm'

XGBoost stands for "Extreme Gradient Boosting" and it is an implementation of gradient boosting trees algorithm. The XGBoost is a popular supervised machine learning model with characteristics like computation speed, parallelization, and performance.

```
In [100]: import xgboost as xgb
          from scipy.stats import randint
          np.random.seed(42)
          model = xgb.XGBRegressor(random_state=42)
          import warnings
          warnings.filterwarnings('ignore')

          param_grid = [
              'learning_rate': randint(low=0.1, high=20), 'n_estimators': randint(low=0.1, high=20), 'algorithm': ['SAMME',
              ],
              gs = RandomizedSearchCV(model, param_grid, cv=5, n_iter = 150, scoring='neg_mean_squared_error', random_state=42)
              gs.fit(X_train, y_train)
              gs.best_params_

[04:05:43] WARNING: /Users/runner/miniforge3/conda-bld/xgboost_1593723618214/work/src/learner.cc:480:
Parameters: { 'algorithm' } might not be used.

This may not be accurate due to some parameters are only used in language bindings but
passed down to XGBoost core. Or some parameters are not used but slip through this
verification. Please open an issue if you find above cases.

[04:05:43] WARNING: /Users/runner/miniforge3/conda-bld/xgboost_1593723618214/work/src/learner.cc:480:
Parameters: { 'algorithm' } might not be used.

This may not be accurate due to some parameters are only used in language bindings but
```

```

passed down to XGBoost core. Or some parameters are not used but slip through this
verification. Please open an issue if you find above cases.

[04:05:42] WARNING: /Users/runner/miniforge3/conda-bld/xgboost_1593723618214/work/src/learner.cc:480:
Parameters: { algorithm } might not be used.

This may not be accurate due to some parameters are only used in language bindings but
passed down to XGBoost core. Or some parameters are not used but slip through this

```


[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

```

7904355980.09847 ['algorithm': 'SAXMSE.R', 'learning_rate': 5, 'n_estimators': 15]
32025679602.63308 ['algorithm': 'SAXMSE.R', 'learning_rate': 14, 'n_estimators': 8]
1791790723.13797 ['algorithm': 'SAXMSE.R', 'learning_rate': 11, 'n_estimators': 12]
13399428785.38684 ['algorithm': 'SAXMSE.R', 'learning_rate': 8, 'n_estimators': 10]
10131012.12213 ['algorithm': 'SAXMSE.R', 'learning_rate': 14, 'n_estimators': 11]
14113700709.89098 ['algorithm': 'SAXMSE2', 'learning_rate': 6, 'n_estimators': 12]
1262638169.03575 ['algorithm': 'SAXMSE', 'learning_rate': 16, 'n_estimators': 7]
8264503363.099715 ['algorithm': 'SAXMSE', 'learning_rate': 14, 'n_estimators': 11]
732046143.6262157 ['algorithm': 'SAXMSE.R', 'learning_rate': 13, 'n_estimators': 7]
16769692.0973466 ['algorithm': 'SAXMSE', 'learning_rate': 14, 'n_estimators': 7]
78228889.14568 ['algorithm': 'SAXMSE.R', 'learning_rate': 5, 'n_estimators': 12]
80340491.4875382 ['algorithm': 'SAXMSE', 'learning_rate': 11, 'n_estimators': 7]
1674094.74029 ['algorithm': 'SAXMSE', 'learning_rate': 14, 'n_estimators': 11]
234662957.758344 ['algorithm': 'SAXMSE.R', 'learning_rate': 14, 'n_estimators': 6]
19806862.750152 ['algorithm': 'SAXMSE', 'learning_rate': 4, 'n_estimators': 14]
780783194.7153076 ['algorithm': 'SAXMSE', 'learning_rate': 14, 'n_estimators': 11]
1453975.96576947 ['algorithm': 'SAXMSE', 'learning_rate': 13, 'n_estimators': 6]
8602846.1830213 ['algorithm': 'SAXMSE', 'learning_rate': 14, 'n_estimators': 6]
10487848.2597413 ['algorithm': 'SAXMSE.R', 'learning_rate': 7, 'n_estimators': 8]
18691426.42588903 ['algorithm': 'SAXMSE', 'learning_rate': 6, 'n_estimators': 8]
18691426.42588903 ['algorithm': 'SAXMSE', 'learning_rate': 6, 'n_estimators': 8]
18691426.42588903 ['algorithm': 'SAXMSE.R', 'learning_rate': 6, 'n_estimators': 8]
12161493.0368097 ['algorithm': 'SAXMSE.R', 'learning_rate': 13, 'n_estimators': 5]
12161493.0368097 ['algorithm': 'SAXMSE', 'learning_rate': 13, 'n_estimators': 5]
11514199.12809738 ['algorithm': 'SAXMSE.R', 'learning_rate': 3, 'n_estimators': 18]
107156.165738306 ['algorithm': 'SAXMSE', 'learning_rate': 9, 'n_estimators': 6]
4805304.44850203 ['algorithm': 'SAXMSE.R', 'learning_rate': 11, 'n_estimators': 5]
4805304.748758173 ['algorithm': 'SAXMSE', 'learning_rate': 15, 'n_estimators': 4]
2489400.138252766 ['algorithm': 'SAXMSE', 'learning_rate': 11, 'n_estimators': 1]
2261837.35260289 ['algorithm': 'SAXMSE', 'learning_rate': 7, 'n_estimators': 6]
2261837.35260289 ['algorithm': 'SAXMSE', 'learning_rate': 7, 'n_estimators': 6]
188015.677873686 ['algorithm': 'SAXMSE', 'learning_rate': 15, 'n_estimators': 6]
188015.677873686 ['algorithm': 'SAXMSE', 'learning_rate': 15, 'n_estimators': 6]
797937.153573734 ['algorithm': 'SAXMSE.R', 'learning_rate': 9, 'n_estimators': 5]
797937.153573734 ['algorithm': 'SAXMSE', 'learning_rate': 9, 'n_estimators': 5]
189759.0356879804 ['algorithm': 'SAXMSE.R', 'learning_rate': 5, 'n_estimators': 7]
73206.0145224607 ['algorithm': 'SAXMSE', 'learning_rate': 3, 'n_estimators': 13]
32240.640147916 ['algorithm': 'SAXMSE', 'learning_rate': 14, 'n_estimators': 1]
103026.6752124007 ['algorithm': 'SAXMSE', 'learning_rate': 4, 'n_estimators': 8]
17173.83267463 ['algorithm': 'SAXMSE', 'learning_rate': 14, 'n_estimators': 8]
117846.4874904078 ['algorithm': 'SAXMSE', 'learning_rate': 8, 'n_estimators': 4]
36007.37881560485 ['algorithm': 'SAXMSE', 'learning_rate': 10, 'n_estimators': 3]
15170.0481876084 ['algorithm': 'SAXMSE', 'learning_rate': 10, 'n_estimators': 3]
30640.19063834643 ['algorithm': 'SAXMSE2', 'learning_rate': 6, 'n_estimators': 4]
30640.19063834643 ['algorithm': 'SAXMSE', 'learning_rate': 6, 'n_estimators': 4]
22083.7387336042 ['algorithm': 'SAXMSE.R', 'learning_rate': 9, 'n_estimators': 3]
22083.7387336042 ['algorithm': 'SAXMSE', 'learning_rate': 9, 'n_estimators': 3]
16933.48773095613 ['algorithm': 'SAXMSE', 'learning_rate': 8, 'n_estimators': 3]
16933.48773095613 ['algorithm': 'SAXMSE', 'learning_rate': 8, 'n_estimators': 3]
7070.7254044018 ['algorithm': 'SAXMSE', 'learning_rate': 14, 'n_estimators': 2]
7070.7254044018 ['algorithm': 'SAXMSE', 'learning_rate': 14, 'n_estimators': 2]
9738.20232664001 ['algorithm': 'SAXMSE.R', 'learning_rate': 15, 'n_estimators': 2]
9738.20232664001 ['algorithm': 'SAXMSE', 'learning_rate': 15, 'n_estimators': 2]
1798.613710410749 ['algorithm': 'SAXMSE', 'learning_rate': 8, 'n_estimators': 11]
1798.613710410749 ['algorithm': 'SAXMSE', 'learning_rate': 8, 'n_estimators': 11]
2154.43948224628 ['algorithm': 'SAXMSE', 'learning_rate': 5, 'n_estimators': 3]
2154.43948224628 ['algorithm': 'SAXMSE', 'learning_rate': 5, 'n_estimators': 3]
7199.61374262126 ['algorithm': 'SAXMSE.R', 'learning_rate': 3, 'n_estimators': 4]
7199.61374262126 ['algorithm': 'SAXMSE', 'learning_rate': 3, 'n_estimators': 4]
649.665502078192 ['algorithm': 'SAXMSE', 'learning_rate': 14, 'n_estimators': 1]
649.665502078192 ['algorithm': 'SAXMSE', 'learning_rate': 14, 'n_estimators': 1]
4199.7104105705714 ['algorithm': 'SAXMSE', 'learning_rate': 11, 'n_estimators': 7]
4199.7104105705714 ['algorithm': 'SAXMSE', 'learning_rate': 11, 'n_estimators': 7]
299.4910340481178 ['algorithm': 'SAXMSE.R', 'learning_rate': 7, 'n_estimators': 1]
299.4910340481178 ['algorithm': 'SAXMSE', 'learning_rate': 7, 'n_estimators': 1]
50.73396636781616 ['algorithm': 'SAXMSE', 'learning_rate': 0, 'n_estimators': 1]
50.73396636781616 ['algorithm': 'SAXMSE', 'learning_rate': 0, 'n_estimators': 1]
50.73396636781616 ['algorithm': 'SAXMSE', 'learning_rate': 0, 'n_estimators': 14]
50.73396636781616 ['algorithm': 'SAXMSE', 'learning_rate': 0, 'n_estimators': 14]
50.73396636781616 ['algorithm': 'SAXMSE2', 'learning_rate': 0, 'n_estimators': 4]
50.73396636781616 ['algorithm': 'SAXMSE', 'learning_rate': 0, 'n_estimators': 4]
50.73396636781616 ['algorithm': 'SAXMSE', 'learning_rate': 0, 'n_estimators': 7]
50.73396636781616 ['algorithm': 'SAXMSE', 'learning_rate': 0, 'n_estimators': 0]
50.73396636781616 ['algorithm': 'SAXMSE', 'learning_rate': 2, 'n_estimators': 0]
50.73396636781616 ['algorithm': 'SAXMSE', 'learning_rate': 0, 'n_estimators': 2]
50.73396636781616 ['algorithm': 'SAXMSE', 'learning_rate': 0, 'n_estimators': 0]
50.73396636781616 ['algorithm': 'SAXMSE', 'learning_rate': 0, 'n_estimators': 0]
50.73396636781616 ['algorithm': 'SAXMSE', 'learning_rate': 0, 'n_estimators': 10]
50.73396636781616 ['algorithm': 'SAXMSE', 'learning_rate': 0, 'n_estimators': 10]
50.73396636781616 ['algorithm': 'SAXMSE.R', 'learning_rate': 0, 'n_estimators': 11]
50.73396636781616 ['algorithm': 'SAXMSE.R', 'learning_rate': 0, 'n_estimators': 11]
50.73396636781616 ['algorithm': 'SAXMSE', 'learning_rate': 0, 'n_estimators': 19]
50.73396636781616 ['algorithm': 'SAXMSE', 'learning_rate': 0, 'n_estimators': 19]
50.73396636781616 ['algorithm': 'SAXMSE', 'learning_rate': 0, 'n_estimators': 4]
50.73396636781616 ['algorithm': 'SAXMSE', 'learning_rate': 0, 'n_estimators': 4]
50.73396636781616 ['algorithm': 'SAXMSE', 'learning_rate': 0, 'n_estimators': 8]
50.73396636781616 ['algorithm': 'SAXMSE', 'learning_rate': 0, 'n_estimators': 8]
50.73396636781616 ['algorithm': 'SAXMSE', 'learning_rate': 0, 'n_estimators': 0]
50.73396636781616 ['algorithm': 'SAXMSE', 'learning_rate': 4, 'n_estimators': 0]
50.73396636781616 ['algorithm': 'SAXMSE', 'learning_rate': 0, 'n_estimators': 0]
50.73396636781616 ['algorithm': 'SAXMSE', 'learning_rate': 0, 'n_estimators': 0]
50.73396636781616 ['algorithm': 'SAXMSE', 'learning_rate': 2, 'n_estimators': 6]
50.73396636781616 ['algorithm': 'SAXMSE', 'learning_rate': 2, 'n_estimators': 6]
44.61109479897922 ['algorithm': 'SAXMSE.R', 'learning_rate': 2, 'n_estimators': 11]
44.61109479897922 ['algorithm': 'SAXMSE', 'learning_rate': 2, 'n_estimators': 11]
44.61109479897922 ['algorithm': 'SAXMSE', 'learning_rate': 2, 'n_estimators': 15]
44.61109479897922 ['algorithm': 'SAXMSE', 'learning_rate': 2, 'n_estimators': 15]
43.6692537689885 ['algorithm': 'SAXMSE', 'learning_rate': 2, 'n_estimators': 16]
43.6692537689885 ['algorithm': 'SAXMSE', 'learning_rate': 2, 'n_estimators': 16]
43
```

Parameters: { algorithm } might not be used.

This may not be accurate due to some parameters are only used in language bindings but passed down to XGBoost core. Or some parameters are not used but slip through this verification. Please open an issue if you find above cases.

```
[04:08:18] /Users/runner/miniforge3/conda-bld/xgboost_1593723618214/work/src/learner.cc:480:
Parameters: { algorithm } might not be used.
```

This may not be accurate due to some parameters are only used in language bindings but passed down to XGBoost core. Or some parameters are not used but slip through this verification. Please open an issue if you find above cases.

```
[04:08:18] WARNING: /Users/runner/miniforge3/conda-bld/xgboost_1593723618214/work/src/learner.cc:480:
Parameters: { algorithm } might not be used.
```

This may not be accurate due to some parameters are only used in language bindings but passed down to XGBoost core. Or some parameters are not used but slip through this verification. Please open an issue if you find above cases.

```
Out [302]: 0.249335407812195
```

XBG is giving some warnings.

The gradient Boosting has the rmse of 8.25.

```
In [ ]: 
```

SHALLOW NEURAL NETWORK

EARLY STOPPING APPLIED

DIVING DATA INTO VALIDATION AND TRAINING FOR EARLY STOPPING

```
In [305]: from sklearn.model_selection import train_test_split
# Smaller Training
# You need to divide your data to smaller training set and validation set for early stopping.
Training_csv=np.concatenate((X_train_new,np.array(y_train).reshape(-1,1)),axis=1)
Smaller_Training_Validation = train_test_split(Training_csv, test_size=0.15, random_state=42)
# to avoid overfitting or epoch

#
Smaller_Training_Target=Smaller_Training[:,:-1]
Smaller_Training=Smaller_Training[:,:-1]
#
Validation_Target=Validation[:,:-1]
Validation=Validation[:,:-1]
```

Lets now train a ANN for regression with 3 hidden layers; each layer has 50 neurons. Th following code implement early stopping to avoid overfitting.

```
In [306]: def ANN (input_dim,neurons=50,loss='mse',activation='relu',Nout=1,
            metrics=None,activation_out=None,dropout_rate=True):
    """ Function to run Neural Network for different hyperparameters"""
    # activation_out='sigmoid' not needed for regression
    np.random.seed(42)
    tf.random.set_seed(42)

    # create model
    model = keras.models.Sequential()

    # Input and Hidden Layer 1
    model.add(keras.layers.Dense(neurons,input_dim=input_dim, activation=activation))

    # Hidden Layer 2
    model.add(keras.layers.Dense(neurons,activation=activation))

    # Hidden Layer 3
    model.add(keras.layers.Dense(neurons,activation=activation))

    # Output Layer
    model.add(keras.layers.Dense(Nout,activation=activation_out))

    # Compile model
    model.compile(optimizer='adam',loss='mse',metrics=metrics)
    return model
```

```
In [307]: from keras.wrappers.scikit_learn import GridSearchRegressor
from sklearn.model_selection import GridSearchCV
from tensorflow import keras
import tensorflow as tf

# define the grid search parameters
param_grid = {
    'neurons': [50,100,150,200],
    'activation': ['softmax','softplus', 'softsign', 'relu', 'tanh', 'sigmoid', 'hard_sigmoid', 'linear'],
    'dropout_rate': [(0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9)]
}

# Run
model = KerasRegressor(build_fn=ANN,input_dim=Smaller_Training.shape[1],loss='mse',
                        metrics=None,activation_out=None)
#model._estimator_type = "Regressor"

# loss use mse
# use KerasRegressor

# Apply Scikit Learn GridSearchCV
grid = GridSearchCV(estimator=model,param_grid=param_grid,cv=5)

# Early stopping to avoid overfitting
model=Keras.callbacks.EarlyStopping(monitor='val_loss',patience=5, verbose=0)
grid_result = grid.fit(Smaller_Training,Smaller_Training_Target,batch_size=32,validation_data=(
    Validation,Validation_Target),callbacks=[monitor],
    verbose=0,epochs=1000)
```

```
7/7 |=====| - Os 659us/step - loss: 102.2029
7/7 |=====| - Os 715us/step - loss: 94.3496
7/7 |=====| - Os 651us/step - loss: 82.3714
7/7 |=====| - Os 698us/step - loss: 107.3257
7/7 |=====| - Os 707us/step - loss: 85.1095
7/7 |=====| - Os 698us/step - loss: 101.3450
7/7 |=====| - Os 714us/step - loss: 93.8098
7/7 |=====| - Os 742us/step - loss: 81.8962
7/7 |=====| - Os 872us/step - loss: 107.0561
7/7 |=====| - Os 743us/step - loss: 84.7270
7/7 |=====| - Os 776us/step - loss: 100.9460
7/7 |=====| - Os 748us/step - loss: 93.5308
7/7 |=====| - loss: 81.6819
7/7 |=====| - Os 731us/step - loss: 106.9377
7/7 |=====| - Os 738us/step - loss: 84.5807
7/7 |=====| - Os 876us/step - loss: 100.4711
7/7 |=====| - Os 806us/step - loss: 92.3567
7/7 |=====| - Os 780us/step - loss: 81.1761
```

7/7	=====	-08 701us/step	-loss: 106.6998
7/7	=====	-08 788us/step	-loss: 85.1095
7/7	=====	-08 693us/step	-loss: 102.2029
7/7	=====	-08 693us/step	-loss: 85.1095
7/7	=====	-08 699us/step	-loss: 82.3714
7/7	=====	-08 695us/step	-loss: 107.3227
7/7	=====	-08 701us/step	-loss: 85.1095
7/7	=====	-08 663us/step	-loss: 101.3450
7/7	=====	-08 713us/step	-loss: 93.8098
7/7	=====	-08 988us/step	-loss: 107.0561
7/7	=====	-08 812us/step	-loss: 107.0561
7/7	=====	-08 722us/step	-loss: 84.3964
7/7	=====	-08 742us/step	-loss: 100.9460
7/7	=====	-08 739us/step	-loss: 93.3308
7/7	=====	-08 773us/step	-loss: 81.6839
7/7	=====	-08 735us/step	-loss: 106.9377
7/7	=====	-08 739us/step	-loss: 84.5807
7/7	=====	-08 753us/step	-loss: 100.4711
7/7	=====	-08 752us/step	-loss: 92.3567
7/7	=====	-08 766us/step	-loss: 81.7611
7/7	=====	-08 790us/step	-loss: 106.6998
7/7	=====	-08 707us/step	-loss: 84.3964
7/7	=====	-08 839us/step	-loss: 100.2029
7/7	=====	-08 718us/step	-loss: 94.3496
7/7	=====	-08 699us/step	-loss: 82.3714
7/7	=====	-08 713us/step	-loss: 107.3227
7/7	=====	-08 607us/step	-loss: 85.1095
7/7	=====	-08 714us/step	-loss: 101.3450
7/7	=====	-08 702us/step	-loss: 93.8098
7/7	=====	-08 923us/step	-loss: 81.6942
7/7	=====	-08 717us/step	-loss: 107.0561
7/7	=====	-08 693us/step	-loss: 84.7270
7/7	=====	-08 735us/step	-loss: 100.9460
7/7	=====	-08 718us/step	-loss: 92.3567
7/7	=====	-08 781us/step	-loss: 81.6839
7/7	=====	-08 764us/step	-loss: 84.3964
7/7	=====	-08 775us/step	-loss: 84.5807
7/7	=====	-08 821us/step	-loss: 100.4711
7/7	=====	-08 806us/step	-loss: 92.3567
7/7	=====	-08 961us/step	-loss: 81.7611
7/7	=====	-08 778us/step	-loss: 106.6998
7/7	=====	-08 814us/step	-loss: 84.3964
7/7	=====	-08 1ms/step	-loss: 102.2029
7/7	=====	-08 1ms/step	-loss: 94.3496
7/7	=====	-08 696us/step	-loss: 82.3714
7/7	=====	-08 693us/step	-loss: 107.3227
7/7	=====	-08 654us/step	-loss: 85.1095
7/7	=====	-08 683us/step	-loss: 101.3450
7/7	=====	-08 703us/step	-loss: 93.8098
7/7	=====	-08 853us/step	-loss: 107.0561
7/7	=====	-08 771us/step	-loss: 107.0561
7/7	=====	-08 695us/step	-loss: 84.7270
7/7	=====	-08 676us/step	-loss: 100.9460
7/7	=====	-08 980us/step	-loss: 93.3308
7/7	=====	-08 1ms/step	-loss: 81.6839
7/7	=====	-08 819us/step	-loss: 106.9377
7/7	=====	-08 690us/step	-loss: 84.5807
7/7	=====	-08 730us/step	-loss: 100.4711
7/7	=====	-08 882us/step	-loss: 92.3567
7/7	=====	-08 813us/step	-loss: 81.7611
7/7	=====	-08 1ms/step	-loss: 106.6998
7/7	=====	-08 1ms/step	-loss: 102.2029
7/7	=====	-08 678us/step	-loss: 94.3496
7/7	=====	-08 773us/step	-loss: 94.3496
7/7	=====	-08 780us/step	-loss: 82.3714
7/7	=====	-08 674us/step	-loss: 107.3227
7/7	=====	-08 2ms/step	-loss: 85.1095
7/7	=====	-08 760us/step	-loss: 101.3450
7/7	=====	-08 678us/step	-loss: 93.8098
7/7	=====	-08 753us/step	-loss: 81.6942
7/7	=====	-08 679us/step	-loss: 107.0561
7/7	=====	-08 668us/step	-loss: 84.7270
7/7	=====	-08 721us/step	-loss: 100.9460
7/7	=====	-08 903us/step	-loss: 93.3308
7/7	=====	-08 786us/step	-loss: 81.6839
7/7	=====	-08 853us/step	-loss: 106.9377
7/7	=====	-08 765us/step	-loss: 84.5807
7/7	=====	-08 1ms/step	-loss: 100.4711
7/7	=====	-08 1ms/step	-loss: 92.3567
7/7	=====	-08 757us/step	-loss: 81.7611
7/7	=====	-08 739us/step	-loss: 106.6998
7/7	=====	-08 781us/step	-loss: 84.3964
7/7	=====	-08 709us/step	-loss: 102.2029
7/7	=====	-08 731us/step	-loss: 94.3496
7/7	=====	-08 731us/step	-loss: 82.3714
7/7	=====	-08 796us/step	-loss: 107.3227
7/7	=====	-08 770us/step	-

```

7/7 [=====] 00 7740s/step - loss: 84.5807
7/7 [=====] 00 7735s/step - loss: 100.4711
7/7 [=====] 00 9211s/step - loss: 92.1565
7/7 [=====] 00 10s/step - loss: 81.1761
7/7 [=====] 00 8230s/step - loss: 106.6998
7/7 [=====] 00 3170s/step - loss: 94.3964
7/7 [=====] 00 6950s/step - loss: 102.2029
7/7 [=====] 00 7030s/step - loss: 94.8596
7/7 [=====] 00 6666s/step - loss: 82.3714
7/7 [=====] 00 8818s/step - loss: 107.3257
7/7 [=====] 00 7020s/step - loss: 65.1030
7/7 [=====] 00 7070s/step - loss: 101.3450
7/7 [=====] 00 8304s/step - loss: 93.8098
7/7 [=====] 00 6730s/step - loss: 94.3964
7/7 [=====] 00 7160s/step - loss: 107.0561
7/7 [=====] 00 6330s/step - loss: 84.3664
7/7 [=====] 00 7330s/step - loss: 100.9460
7/7 [=====] 00 7030s/step - loss: 93.3038
7/7 [=====] 00 7620s/step - loss: 81.6839
7/7 [=====] 00 7330s/step - loss: 106.9377
7/7 [=====] 00 7430s/step - loss: 84.5807
7/7 [=====] 00 8000s/step - loss: 100.4711
7/7 [=====] 00 8000s/step - loss: 92.3567
7/7 [=====] 00 7130s/step - loss: 81.1761
7/7 [=====] 00 7490s/step - loss: 106.6998
7/7 [=====] 00 7670s/step - loss: 84.3664
7/7 [=====] 00 6840s/step - loss: 102.2029
7/7 [=====] 00 6750s/step - loss: 94.3436
7/7 [=====] 00 7300s/step - loss: 82.7174
7/7 [=====] 00 4870s/step - loss: 94.3436
7/7 [=====] 00 7030s/step - loss: 85.1095
7/7 [=====] 00 10s/step - loss: 100.4711
7/7 [=====] 00 7960s/step - loss: 93.8098
7/7 [=====] 00 7230s/step - loss: 81.6942
7/7 [=====] 00 7430s/step - loss: 94.3964
7/7 [=====] 00 7300s/step - loss: 84.7270
7/7 [=====] 00 7270s/step - loss: 100.9460
7/7 [=====] 00 7840s/step - loss: 91.3581
7/7 [=====] 00 7480s/step - loss: 81.6839
7/7 [=====] 00 7230s/step - loss: 106.9377
7/7 [=====] 00 7310s/step - loss: 92.3567
7/7 [=====] 00 7710s/step - loss: 100.4711
7/7 [=====] 00 7860s/step - loss: 92.3567
7/7 [=====] 00 7620s/step - loss: 81.1761
7/7 [=====] 00 7520s/step - loss: 106.6998
7/7 [=====] 00 7600s/step - loss: 100.4711
7/7 [=====] 00 6810s/step - loss: 102.2029
7/7 [=====] 00 6750s/step - loss: 94.3436
7/7 [=====] 00 7070s/step - loss: 82.7174
7/7 [=====] 00 6590s/step - loss: 107.3257
7/7 [=====] 00 6860s/step - loss: 65.1030
7/7 [=====] 00 8304s/step - loss: 101.3450
7/7 [=====] 00 6940s/step - loss: 93.8098
7/7 [=====] 00 6740s/step - loss: 94.3964
7/7 [=====] 00 7140s/step - loss: 107.0561
7/7 [=====] 00 7480s/step - loss: 100.9460
7/7 [=====] 00 6730s/step - loss: 93.3038
7/7 [=====] 00 7370s/step - loss: 81.6839
7/7 [=====] 00 7390s/step - loss: 106.9377
7/7 [=====] 00 8480s/step - loss: 84.5807
7/7 [=====] 00 7780s/step - loss: 100.4711
7/7 [=====] 00 7450s/step - loss: 92.3567
7/7 [=====] 00 7370s/step - loss: 94.3964
7/7 [=====] 00 7780s/step - loss: 106.6998
7/7 [=====] 00 7760s/step - loss: 84.3664
7/7 [=====] 00 6870s/step - loss: 94.3964
7/7 [=====] 00 8070s/step - loss: 65.4708
7/7 [=====] 00 7030s/step - loss: 59.1223
7/7 [=====] 00 6770s/step - loss: 70.1615
7/7 [=====] 00 7120s/step - loss: 73.5841
7/7 [=====] 00 7410s/step - loss: 75.1622
7/7 [=====] 00 7620s/step - loss: 70.1615
7/7 [=====] 00 7340s/step - loss: 62.2404
7/7 [=====] 00 8030s/step - loss: 91.7961
7/7 [=====] 00 7510s/step - loss: 71.4354
7/7 [=====] 00 7760s/step - loss: 71.9300
7/7 [=====] 00 8010s/step - loss: 65.6514
7/7 [=====] 00 7880s/step - loss: 59.0906
7/7 [=====] 00 7890s/step - loss: 70.3857
7/7 [=====] 00 8300s/step - loss: 70.4118
7/7 [=====] 00 9030s/step - loss: 68.9624
7/7 [=====] 00 8550s/step - loss: 62.9398
7/7 [=====] 00 83990s/step - loss: 91.7222
7/7 [=====] 00 8500s/step - loss: 70.0819
7/7 [=====] 00 7060s/step - loss: 75.2708
7/7 [=====] 00 6920s/step - loss: 65.6514
7/7 [=====] 00 7140s/step - loss: 59.1223
7/7 [=====] 00 7100s/step - loss: 87.7615
7/7 [=====] 00 7020s/step - loss: 73.5841
7/7 [=====] 00 7480s/step - loss: 75.1622
7/7 [=====] 00 7410s/step - loss: 64.9926
7/7 [=====] 00 7150s/step - loss: 70.4118
7/7 [=====] 00 7620s/step - loss: 91.7961
7/7 [=====] 00 7770s/step - loss: 71.9300
7/7 [=====] 00 7990s/step - loss: 65.6514
7/7 [=====] 00 8070s/step - loss: 59.0906
7/7 [=====] 00 7960s/step - loss: 93.4482
7/7 [=====] 00 7900s/step - loss: 70.3857
7/7 [=====] 00 8830s/step - loss: 70.4118
7/7 [=====] 00 8870s/step - loss: 68.9624
7/7 [=====] 00 8850s/step - loss: 62.9398
7/7 [=====] 00 8650s/step - loss: 91.7222
7/7 [=====] 00 8460s/step - loss: 70.0819
7/7 [=====] 00 6910s/step - loss: 70.1615
7/7 [=====] 00 7230s/step - loss: 65.4708
7/7 [=====] 00 6890s/step - loss: 59.1223
7/7 [=====] 00 6950s/step - loss: 70.1615
7/7 [=====] 00 7070s/step - loss: 73.5841
7/7 [=====] 00 7240s/step - loss: 75.1622
7/7 [=====] 00 7310s/step - loss: 65.6514
7/7 [=====] 00 7510s/step - loss: 62.2404
7/7 [=====] 00 7360s/step - loss: 91.7961
7/7 [=====] 00 7420s/step - loss: 71.4354
7/7 [=====] 00 8330s/step - loss: 71.9300
7/7 [=====] 00 7600s/step - loss: 70.4118
7/7 [=====] 00 7770s/step - loss: 59.0906
7/7 [=====] 00 7370s/step - loss: 93.4482
7/7 [=====] 00 7310s/step - loss: 70.3857
7/7 [=====] 00 8550s/step - loss: 70.4118
7/7 [=====] 00 8510s/step - loss: 68.9624
7/7 [=====] 00 8610s/step - loss: 62.9398
7/7 [=====] 00 8600s/step - loss: 91.7222
7/7 [=====] 00 8510s/step - loss: 70.0819
7/7 [=====] 00 6900s/step - loss: 75.2708
7/7 [=====] 00 6890s/step - loss: 65.6514
7/7 [=====] 00 6710s/step - loss: 65.4708
7/7 [=====] 00 6980s/step - loss: 87.7615
7/7 [=====] 00 7110s/step - loss: 73.5841
7/7 [=====] 00 7350s/step - loss: 75.1622
7/7 [=====] 00 7640s/step - loss: 64.9926
7/7 [=====] 00 7740s/step - loss: 70.4118
7/7 [=====] 00 8650s/step - loss: 91.7961
7/7 [=====] 00 7230s/step - loss: 71.4354
7/7 [=====] 00 7890s/step - loss: 70.4118
7/7 [=====] 00 7570s/step - loss: 65.6514
7/7 [=====] 00 7010s/step - loss: 59.0906
7/7 [=====] 00 8450s/step - loss: 93.4482
7/7 [=====] 00 7340s/step - loss: 70.3857
7/7 [=====] 00 8330s/step - loss: 70.4118
7/7 [=====] 00 9820
```