```
import pandas as pd
```

loading dataset

```
from google.colab import drive
drive.mount('/content/drive')
```

```
df = pd.read_csv("/content/drive/MyDrive/Intro ML 2022 Summer/dataset/basketball_stat.
```

```
df
```

```
# how many players per position? (you can skip this code)
df.Pos.value_counts()
```

data visualization: decide which feature is useful for good classification, which one is not

```
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline

# draw a line plot
sns.lmplot('STL', '2P', data=df, fit_reg=True,
           markers=["o", "x"],
          hue="Pos")
plt.title('STL and 2P for each player')
# x-axis, y-axis, data, and a trend line
# use different color
```

```
sns.lmplot('AST', '2P', data=df, fit_reg=True,
           markers=["o", "x"],
           hue="Pos")
```

```
sns.lmplot('BLK', '3P', data=df, fit_reg=True,
           markers=["o", "x"],
           hue="Pos")
```

```
sns.lmplot('TRB', '3P', data=df, fit_reg=True,
```

```
            markers=["o", "x"],
            hue="Pos")
```

data trimming

```
df.drop(['2P', 'AST', 'STL'], axis=1)
```

data split (train vs test) now and keep it this way (pickling)

```
from sklearn.model_selection import train_test_split

train, test = train_test_split(df, test_size=0.2)
```

```
# double check how many train data points?
train.shape[0]
```

```
# double check how many test data points?
test.shape[0]
```

store trimmed data in a file

```
import pickle
# wb(write bytes): store data in a file
with open('/content/drive/MyDrive/Intro ML 2022 Summer/dataset/basketball_train.pkl',
    pickle.dump(train, train_data)

with open('/content/drive/MyDrive/Intro ML 2022 Summer/dataset/basketball_test.pkl', '
    pickle.dump(test, test_data)
```

```
train
```

```
test
```

✓ 0s completed at 1:10 PM