Classification

November 28, 2025

1 Classification

(a)

[12]:
```
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
import pandas as pd

Autodata = pd.read_csv('Auto.csv')
df = pd.DataFrame(Autodata)

mpg_median = df['mpg'].median()

print(mpg_median, ": is the median of Miles Per Gallon")

df['mpg01'] = (df['mpg'] >= mpg_median).astype(int)
print(df)


df['horsepower'] = pd.to_numeric(df['horsepower'], errors='coerce')
df = df.dropna(subset=['horsepower'])

features = ['displacement', 'horsepower', 'weight', 'acceleration']
fig, axes = plt.subplots(2, 2, figsize=(15, 12))
axes = axes.flatten()

for i, feature in enumerate(features):
    sns.boxplot(ax=axes[i], data=df, x='mpg01', y=feature)
    axes[i].set_title(f'Boxplot of {feature.capitalize()} by mpg01')
    axes[i].set_xlabel('mpg01 (0=Low, 1=High)')

plt.tight_layout(rect=[0, 0.03, 1, 0.95])
plt.show()

features2 = ['origin', 'cylinders']
fig, axes = plt.subplots(1, 2, figsize=(15, 6))
axes2 = axes.flatten()
```

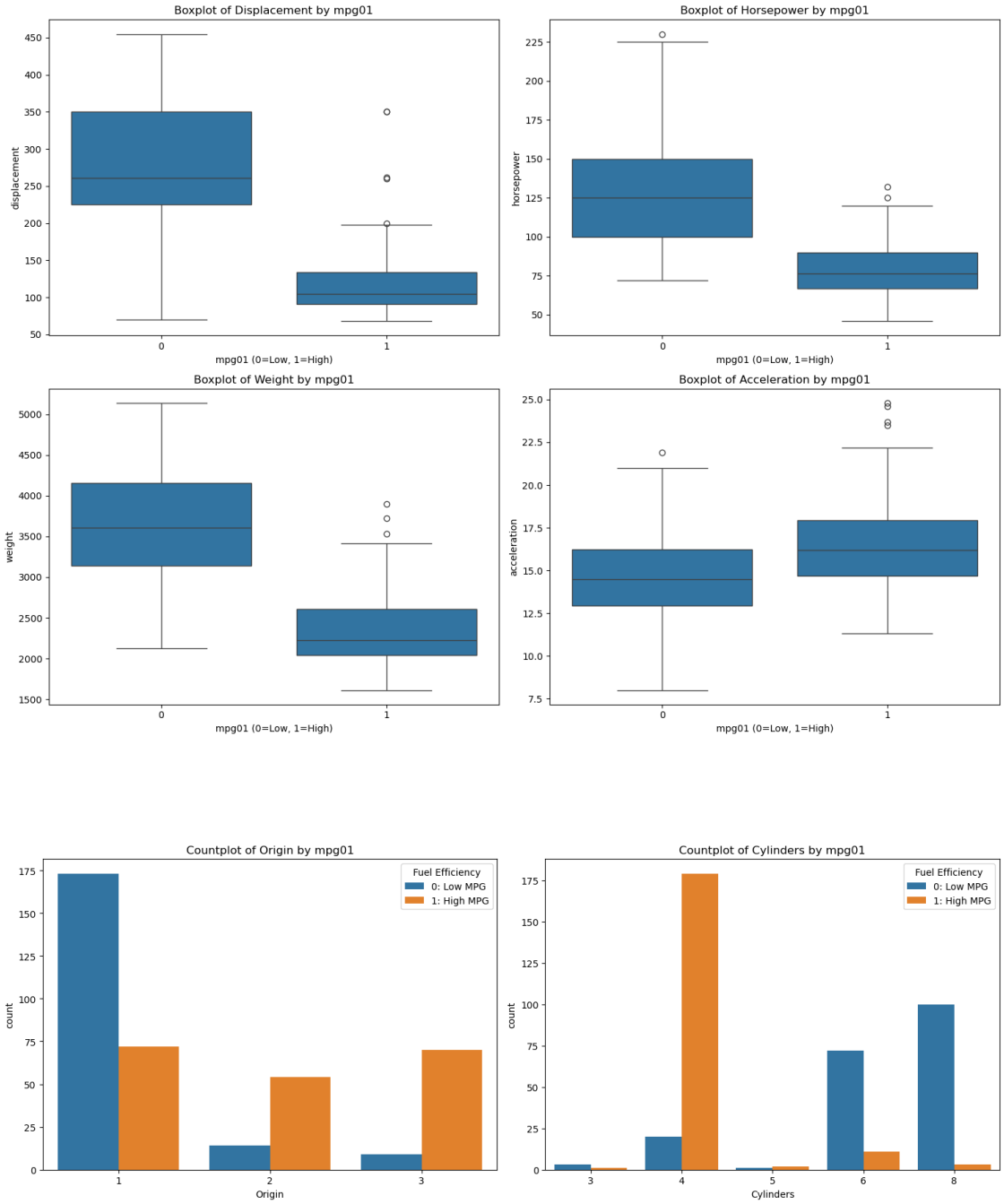
```
for i, features2 in enumerate(features2):
    sns.countplot(ax=axes[i], data=df, x=features2, hue='mpg01')
    axes[i].set_title(f'Countplot of {features2.capitalize()} by mpg01')
    axes[i].set_xlabel(features2.capitalize())
    handles, labels = axes[i].get_legend_handles_labels()
    axes[i].legend(handles, ['0: Low MPG', '1: High MPG'], title='Fuel␣
 ↪Efficiency')

plt.tight_layout(rect=[0, 0.03, 1, 0.95])
plt.show()
```

```
23.0 : is the median of Miles Per Gallon
      mpg  cylinders  displacement horsepower  weight  acceleration  year  \
0    18.0          8         307.0        130    3504          12.0    70
1    15.0          8         350.0        165    3693          11.5    70
2    18.0          8         318.0        150    3436          11.0    70
3    16.0          8         304.0        150    3433          12.0    70
4    17.0          8         302.0        140    3449          10.5    70
..    …          …             …          …     …             …   …
392  27.0          4         140.0         86    2790          15.6    82
393  44.0          4          97.0         52    2130          24.6    82
394  32.0          4         135.0         84    2295          11.6    82
395  28.0          4         120.0         79    2625          18.6    82
396  31.0          4         119.0         82    2720          19.4    82

     origin                       name  mpg01
0         1  chevrolet chevelle malibu      0
1         1          buick skylark 320      0
2         1         plymouth satellite      0
3         1              amc rebel sst      0
4         1                ford torino      0
..      …                          …    …
392       1            ford mustang gl      1
393       2                  vw pickup      1
394       1              dodge rampage      1
395       1                ford ranger      1
396       1                 chevy s-10      1

[397 rows x 10 columns]
```



1.1 Interpretation

Based on these 6 graphs that we have, 6 categories: displacement, horsepower, weight, acceleration, year, cylinders. When interpretting these graphs, the key function to look at are the overlaps and specifically the location of them. If the 2 boxes are aligned with eachother, or at least in similar locations, it means that there isn't much difference in the 2 categories. This means that the average, or the line in the middle of the boxes is around the same even if they are above or below the median

for MPG. Based on this, we can determine that the categories that are the most comparitive with statistics like this are the weight, displacement, horsepower, cylinders, year, and acceleration, from most to least useful. We can see that the weight, displacement, and horsepower are the most useful in determining this, due to them being directly related to how much the car has to move, and this determined how much fuel is used up because of it.You want the ones below the median to be much more near the bottom of the graph, which shows us that it uses less gas per mile, which is what we want.