

clusters

February 23, 2024

1

```
[ ]: import pandas as pd
from scipy.cluster.hierarchy import dendrogram, linkage, fcluster
import matplotlib.pyplot as plt

data = pd.read_csv('extended_data.csv', sep = ';', index_col=0)

metric_vars = ['year', 'price', 'owner_count', 'engine_volume_lites',
               ↪ 'power_in_hp', 'mileage']

data = data.dropna(subset=metric_vars)
for i in metric_vars:
    data['z_' + i] = (data[i] - data[i].mean()) / data[i].std()

metric_vars_norm = list(map(lambda x: 'z_' + x, metric_vars))
plt.figure(figsize=(15, 10))
dn = dendrogram(linkage(data[metric_vars_norm], 'ward'))
```



```
[ ]: data.groupby('culster_labels')[metric_vars].mean()
```

```
[ ]:
      year      price  owner_count  engine_volume_lites \
culster_labels
1      2017.023838  7.682089e+06      2.584242           3.509616
2      1997.753226  6.396754e+05      6.891935           2.365323
3      2015.409892  2.921244e+06      2.947097           2.097663

      power_in_hp      mileage
culster_labels
1      350.581414   93081.826667
2      163.000000   319837.154032
3      197.075842   133176.449319
```

```
[ ]: grouped = data.groupby('culster_labels').size().reset_index(name='all_count')
grouped_audi = data[data['manufacturer'] == 'Audi'].groupby('culster_labels').
    ↪size().reset_index(name='audi_count')
grouped_mercedes = data[data['manufacturer'] == 'Mercedes-Benz'].
    ↪groupby('culster_labels').size().reset_index(name='mercedes_count')
grouped_bmw = data[data['manufacturer'] == 'BMW'].groupby('culster_labels').
    ↪size().reset_index(name='bmw_count')

result = pd.merge(grouped, grouped_audi, on='culster_labels')
result = pd.merge(result, grouped_bmw, on='culster_labels')
result = pd.merge(result, grouped_mercedes, on='culster_labels')

result['percentage_audi'] = (result['audi_count'] / result['all_count']) * 100
result['percentage_mercedes'] = (result['mercedes_count'] /
    ↪result['all_count']) * 100
result['percentage_bmw'] = (result['bmw_count'] / result['all_count']) * 100
result
```

```
[ ]:
      culster_labels  all_count  audi_count  bmw_count  mercedes_count \
0                1      2475        312        1128          1035
1                2      1240        414        429           397
2                3      6975       1574       2698         2703

      percentage_audi  percentage_mercedes  percentage_bmw
0      12.606061      41.818182      45.575758
1      33.387097      32.016129      34.596774
2      22.566308      38.752688      38.681004
```

```
[ ]: color_groups = ['black', 'white', 'grey_silver', 'blue_brown', 'red', 'other']
def color_group(x):
    if pd.isna(x):
        return np.nan
    for i in color_groups:
```

```

        if x in i:
            return i
        return 'other'

```

```
data['color_group'] = data['color'].apply(color_group)
```

```
pd.crosstab(data['culster_labels'], data['color_group']).apply(lambda r: r*100/
↳r.sum(), axis=1)
```

```
[ ]: color_group      black  blue_brown  grey_silver      other      red \
culster_labels
1          46.967847   13.146113   16.931217   3.540904  1.709402
2          29.562345   17.671346   28.075970  12.881916  4.541701
3          37.339674   15.996541   15.621848   3.631647  4.496325

color_group      white
culster_labels
1          17.704518
2           7.266722
3          22.913965

```

```
[ ]: pd.crosstab(data['culster_labels'], data['is_sport_line']).apply(lambda r:
↳r*100/r.sum(), axis=1)
```

```
[ ]: is_sport_line      False      True
culster_labels
1          91.562656   8.437344
2          99.908257   0.091743
3          99.338374   0.661626

```

```
[ ]: pd.crosstab(data['culster_labels'], data['is_crossover']).apply(lambda r: r*100/
↳r.sum(), axis=1)
```

```
[ ]: is_crossover      False      True
culster_labels
1          35.878788   64.121212
2          88.145161   11.854839
3          64.817204   35.182796

```

```
[ ]: pd.crosstab(data['culster_labels'], data['transmission']).apply(lambda r: r*100/
↳r.sum(), axis=1)
```

```
[ ]: transmission      automatic      manual      robot  variator
culster_labels
1          97.048120   0.080873   2.871007  0.000000
2          47.657512  48.626817   0.646204  3.069467
3          74.562661   1.132779  20.461715  3.842845

```

```
[ ]: data['model_full'] = data['manufacturer'] + " " + data['model']
data.groupby('culster_labels')['model_full'].value_counts().groupby(level=0).
    ↪nlargest(20)
```

```
[ ]: culster_labels  culster_labels  model_full
1                1
    BMW X5                                342
    Mercedes-Benz S-Class                 287
    BMW X6                                252
    Mercedes-Benz GLS-Class              166
    Mercedes-Benz GLE                    143
    BMW X7                                120
    BMW 7-Series                         111
    BMW 5-Series                         105
    Mercedes-Benz G-Class               103
    Audi Q7                             100
    Mercedes-Benz GLE Coupe              88
    Audi Q8                              59
    Mercedes-Benz GL-Class              54
    Audi A8                             52
    BMW X4                              51
    Mercedes-Benz E-Class               43
    BMW X3                              42
    Mercedes-Benz CLS-Class             34
    BMW 6-Series Gran Turismo           33
    Mercedes-Benz M-Class              26
2                2
    BMW 5-Series                        163
    BMW 3-Series                        140
    Mercedes-Benz E-Class              138
    Audi 80                            108
    Audi A6                             94
    Audi 100                           87
    Audi A4                             78
    BMW X5                              75
    Mercedes-Benz C-Class              73
    Mercedes-Benz S-Class              35
    Mercedes-Benz M-Class              32
    BMW 7-Series                       26
    Mercedes-Benz Mercedes             25
    Mercedes-Benz Vito                 20
    Mercedes-Benz 190                  18
    Audi A3                            14
    Mercedes-Benz CLK-Class            13
    BMW X3                             12
    Mercedes-Benz G-Class              12
    Mercedes-Benz A-Class              7
3                3
    Mercedes-Benz E-Class             731
```

BMW 5-Series	684
BMW 3-Series	557
Mercedes-Benz C-Class	410
BMW X3	316
Audi A6	292
Audi A4	278
Mercedes-Benz GLC	249
BMW X1	247
Audi Q5	236
Audi A3	197
BMW X5	189
BMW 1-Series	174
Mercedes-Benz GLA-Class	168
Audi A5	155
Mercedes-Benz GLE	134
Mercedes-Benz CLA-Class	133
Mercedes-Benz A-Class	128
Mercedes-Benz M-Class	121
BMW X4	118

Name: count, dtype: int64