Explorando clusters de autos

Estadística Multivariada

Miguel Zamorano

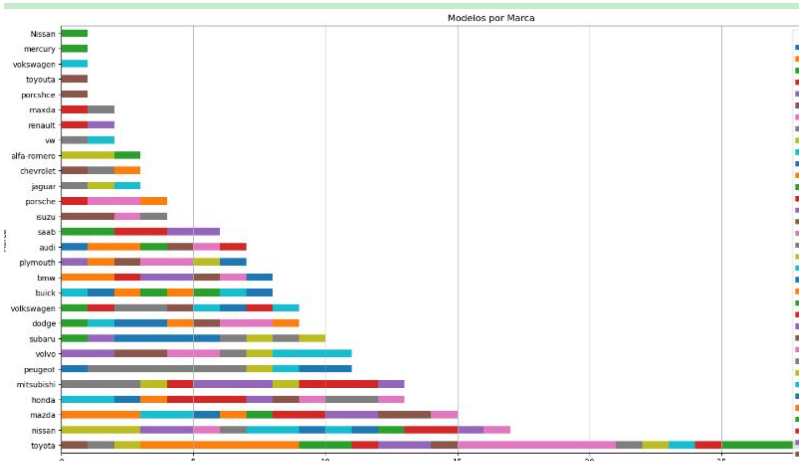
Índice

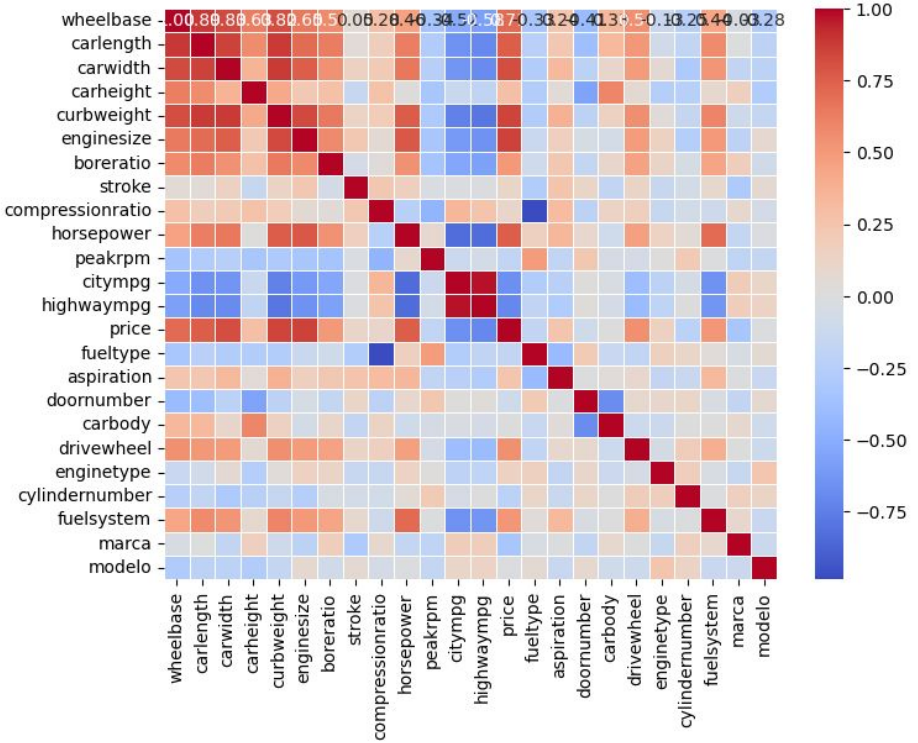
Estructura del Dataset

205 registros, 25 variables de las cuales :

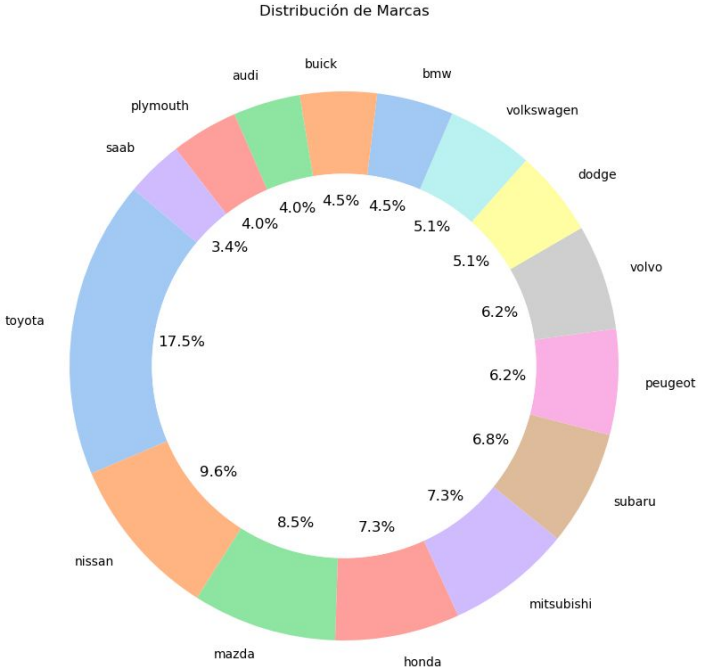
inicialmente : 10 categóricas y 16 numéricas

post limpieza: 25 variables numéricas y 177 registros

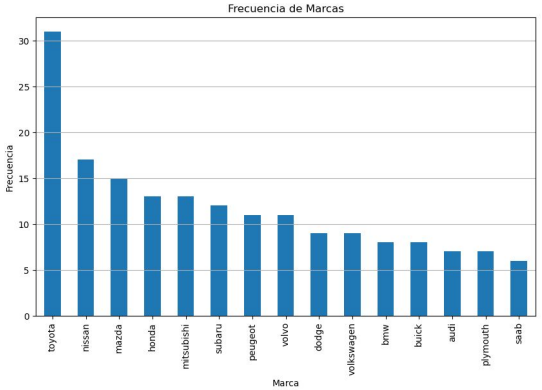




Post Limpieza



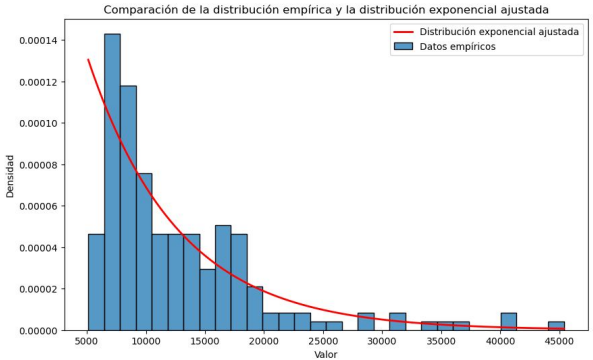
toyota	31
nissan	17
mazda	15
honda	13
mitsubishi	13
subaru	12
peugeot	11
volvo	11
dodge	9
volkswagen	9
bmw	8
buick	8
audi	7
plymouth	7
saab	6



15 marcas de 28

177 registros de 205

Prueba de Bondad:Kolmogorov-Smirnov



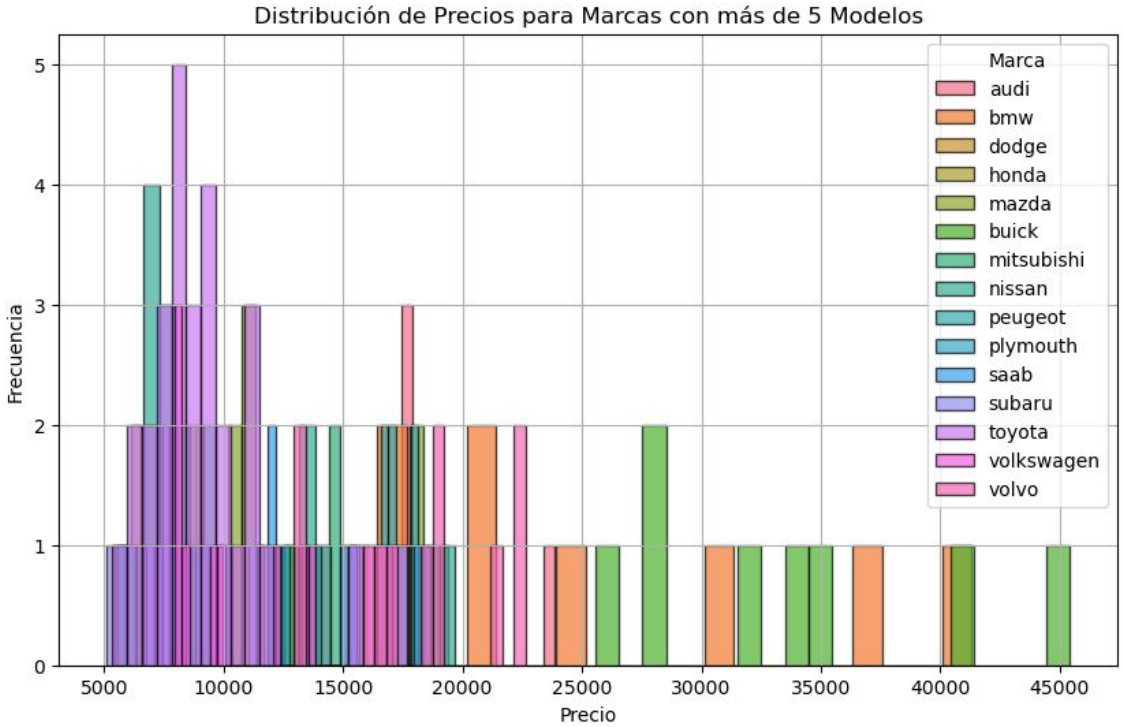

```
# Realizamos  prueba de Kolmogorov-Smirnov
ks_statistic, p_value = kstest(p_marca_mayor, 'powerlaw', args=params)

print(f'Estadístico KS: {ks_statistic}')
print(f'Valor p: {p_value}')
```

```
Estadístico KS: 0.2669693040214083
Valor p: 1.2543809951898195e-11
```

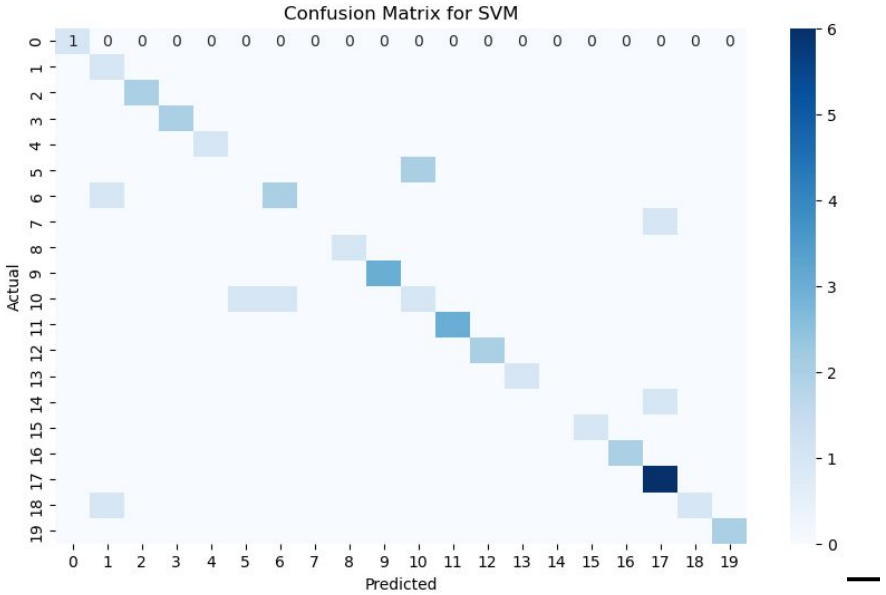
Rechazamos la hipótesis nula: los datos de la variable precio no siguen un distribución exponencial
porque valor P es menor que 0.05

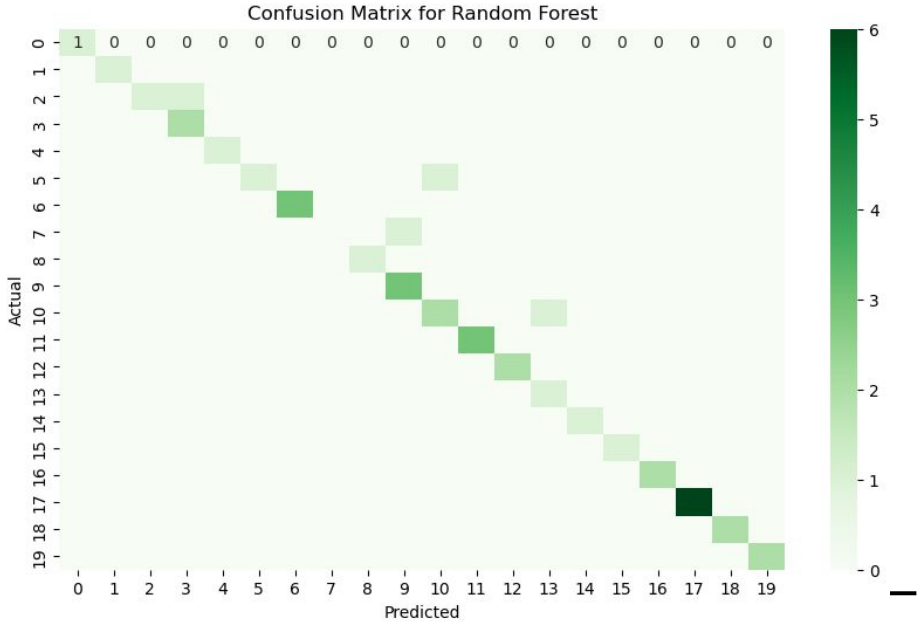
Distribución Precio con Marcas



Clasificación: marcas de auto

	Model	Accuracy	Precision	Recall	F1-Score
0	SVM	0.8	0.770833	0.8	0.774405
1	Random Forest	0.9	0.902083	0.9	0.887619





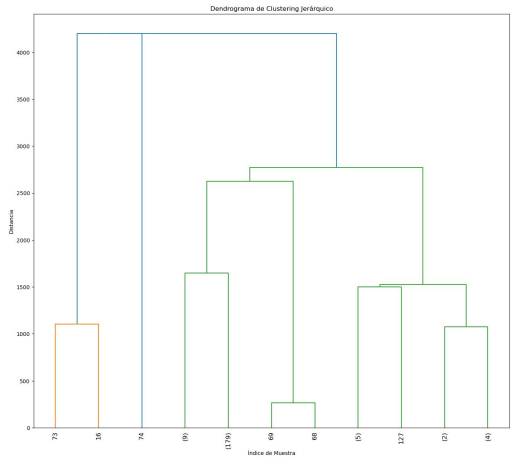
Clustering

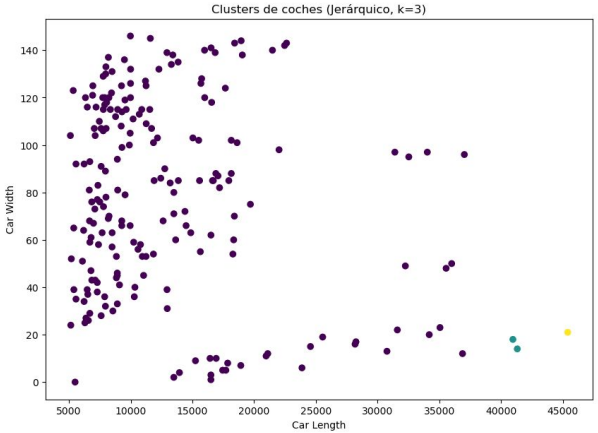
1. Clustering
    a. Aglo.Jerárquico simple
    b. Aglo.Jerárquico completo
    c. Aglo.Jerárquico centroide
    d. kmeans
2. Clustering con Reducción de dimensionalidad
    a. PCA y kmeans
3. Análisis de Calidad
    a. Tabla Siluetas
4. Interpretación de Resultados
    a. **Características correspondientes a cada cluster**

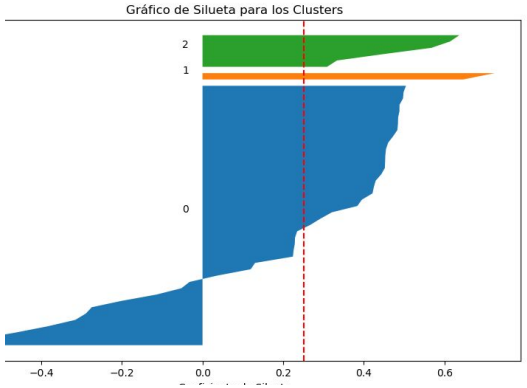
Resultados

Método	Silhouette Coefficient
kmeans	0.643946205
Aglo.Jerárquico simple	0.25
Aglo.Jerárquico completo	0.42
Aglo Jerárquico centroide	0.42
PCA y kmeans	0.7243985953663576

Aglomerativo Jerárquicos simple



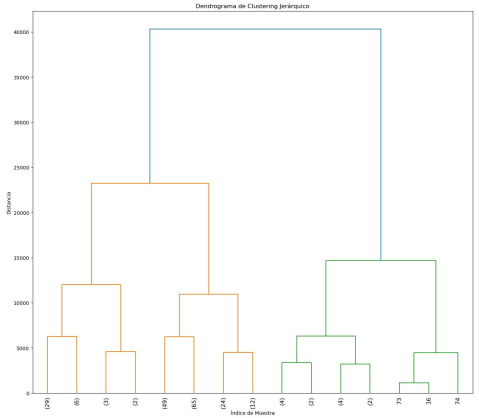


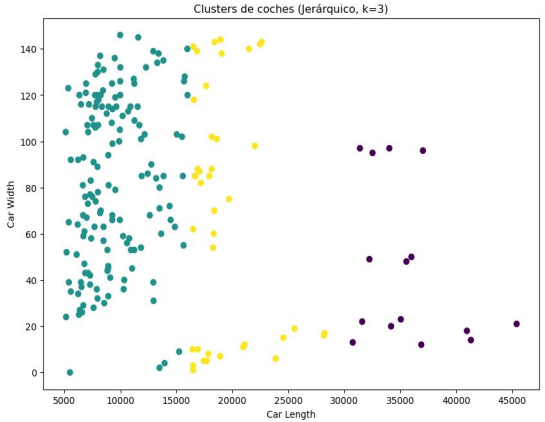


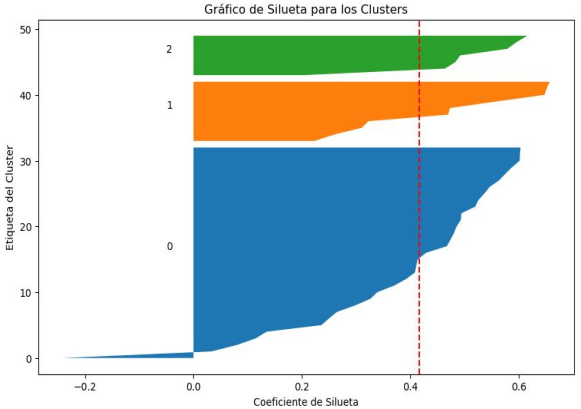
la puntuación de silueta promedio es: 0.25

Aglomerativo Jerárquicos completo

Para n_clusters = 3, la puntuación de silueta promedio es: 0.42

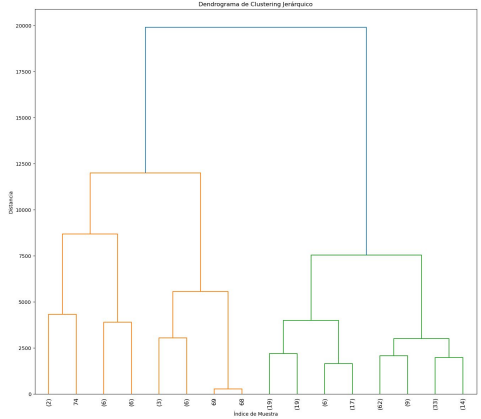


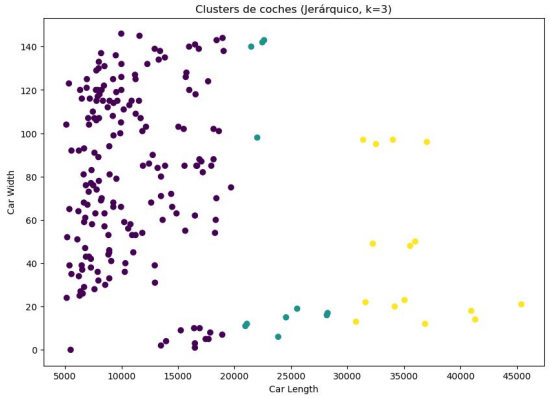


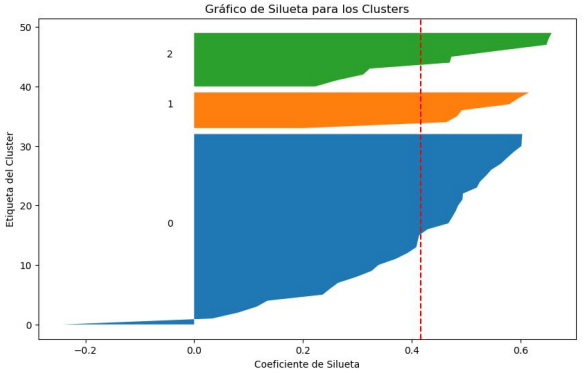


Aglomerativo Jerárquicos centroide

Para n_clusters = 3, la puntuación de silueta promedio es: 0.42

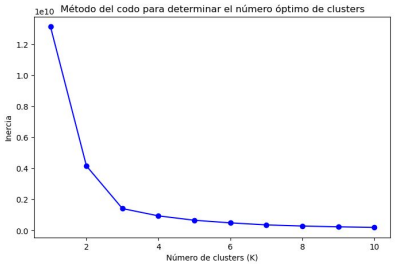




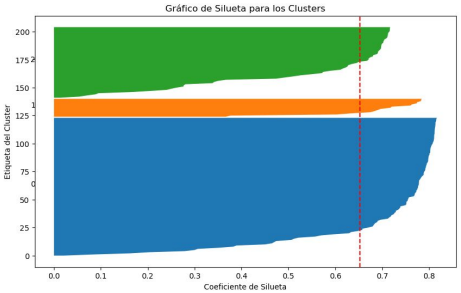


kmeans k=3

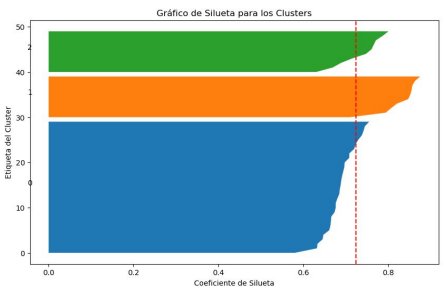
Resultados M.del Codo



Sin PCA 0.643946205

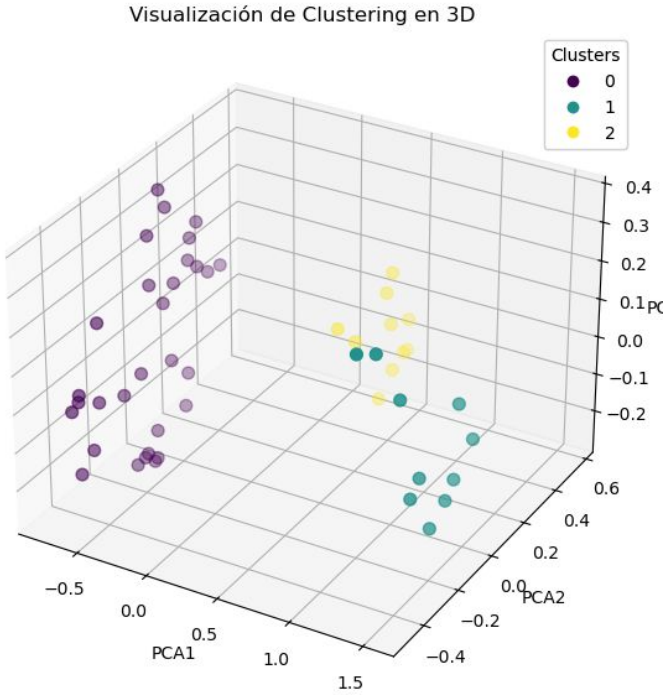


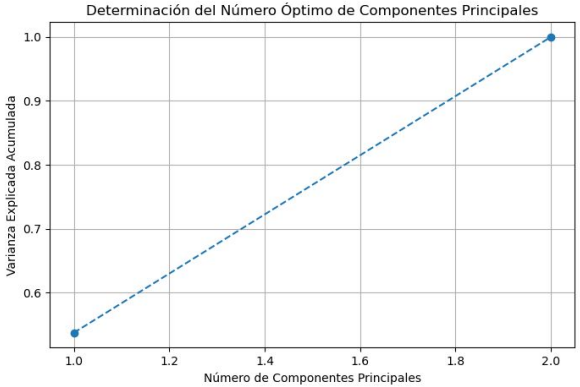
Con PCA.7243985953663576

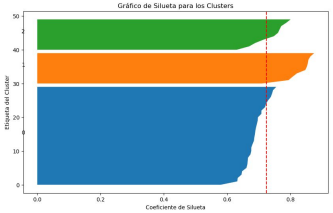


kmeans con PCA

PCA nos encontró 2 componentes Principales

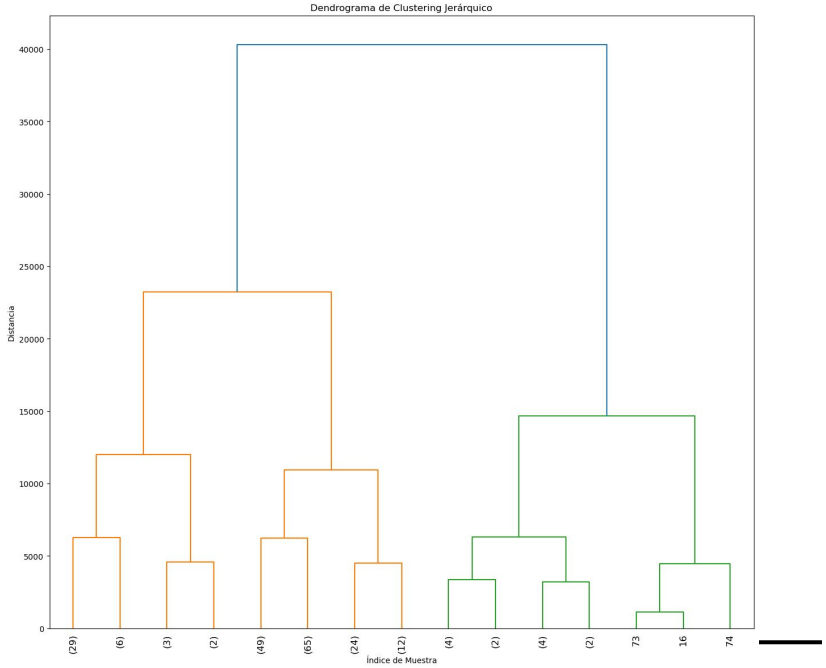




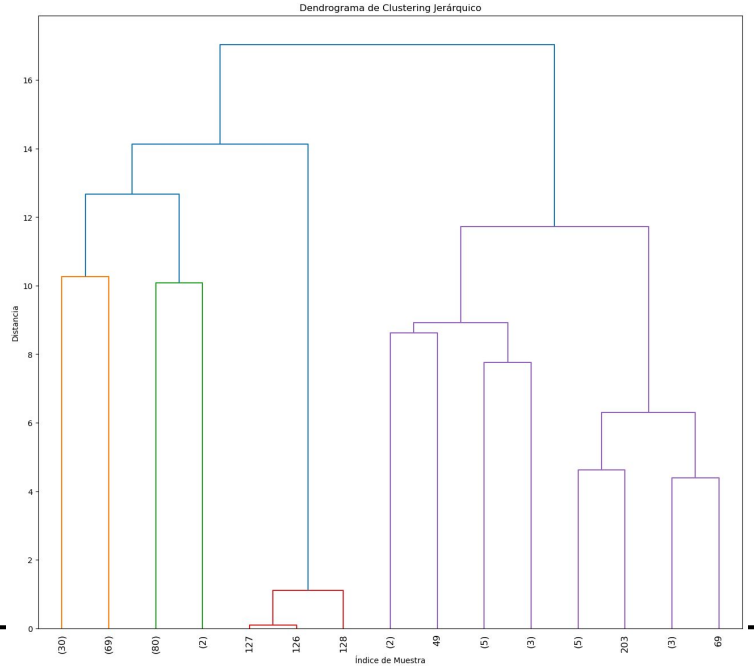


comparativo Aglo completo

Sin PCA



Con PCA

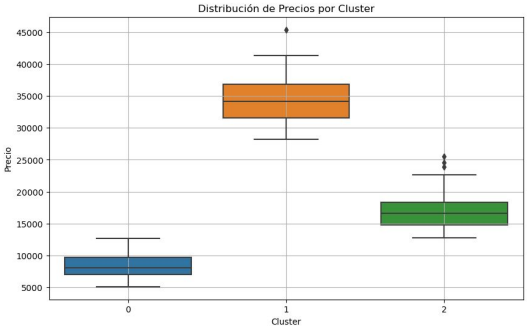


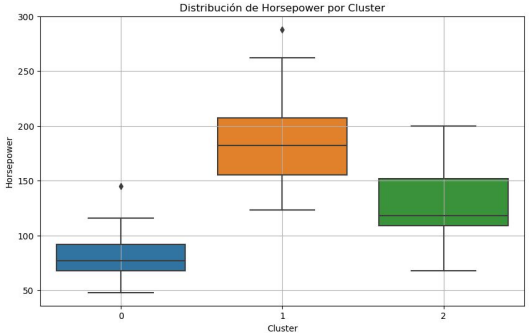
Resultados

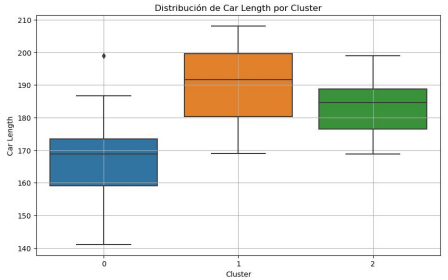
Método	Silhouette Coefficient
kmeans	0.643946205
Aglo.Jerárquico simple	0.25
Aglo.Jerárquico completo	0.42
Aglo Jerárquico centroide	0.42
PCA y kmeans	0.7243985953663576

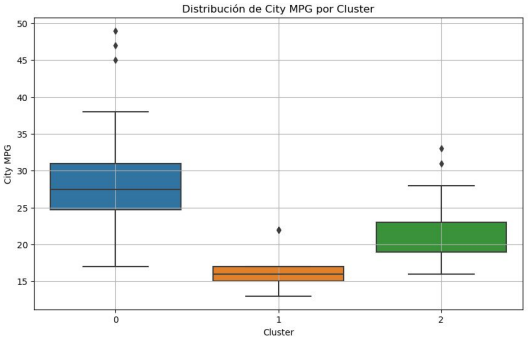
Características representativas correspondientes a cada cluster

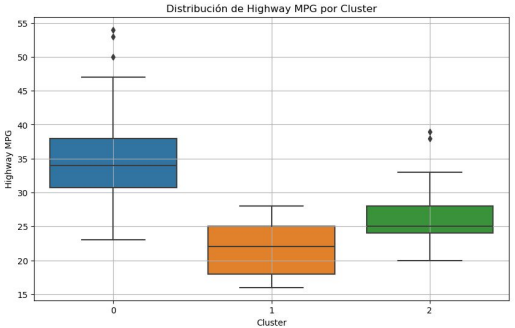
cluster vs columna n

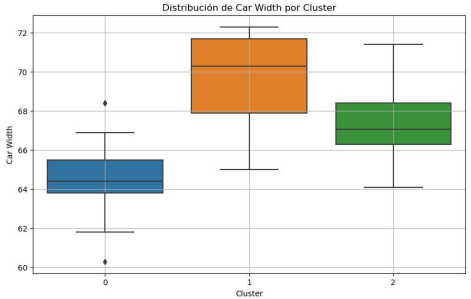




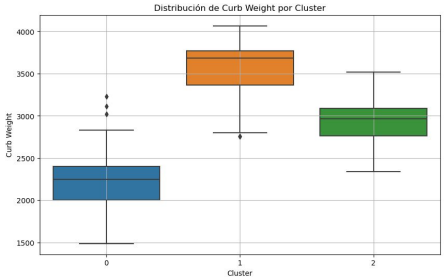


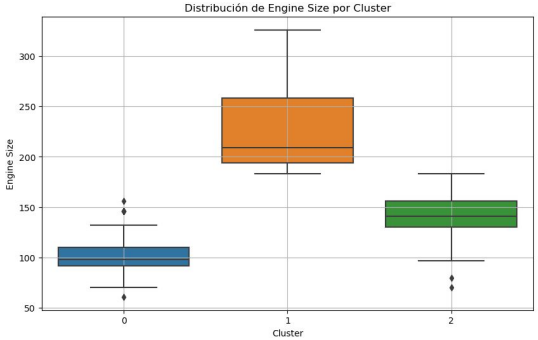




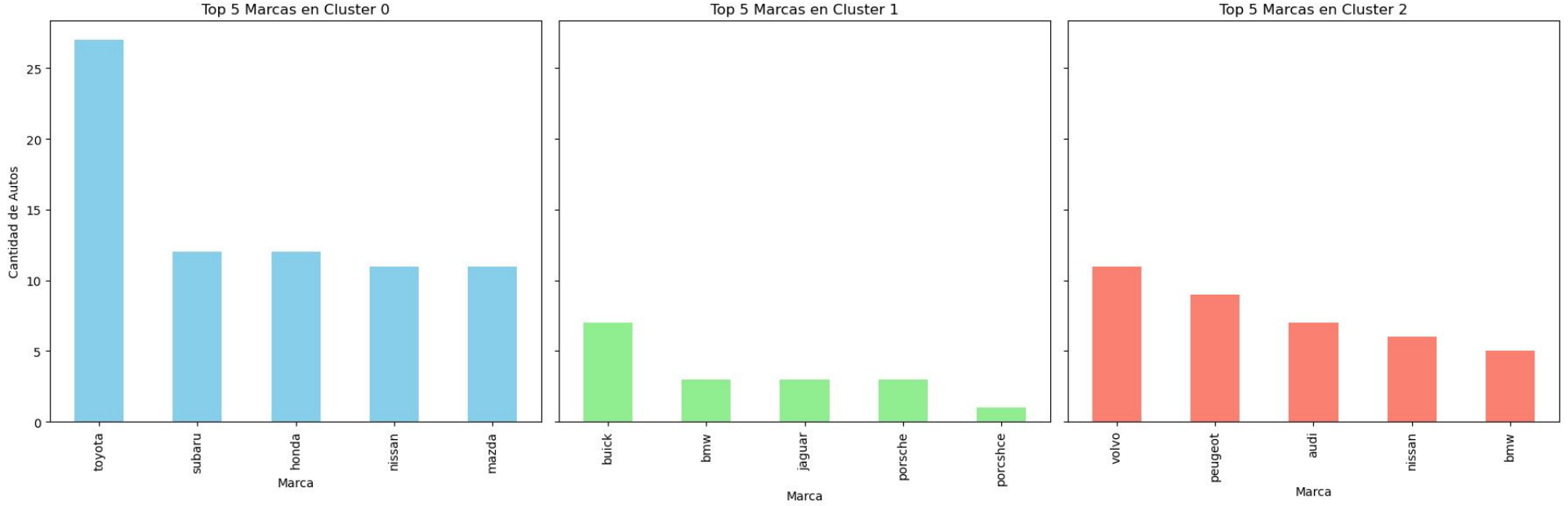


cluster vs columna n





Top 5 Marcas por Cluster



Referencias

El dataset elegido fue obtenido de kaggle
https://www.kaggle.com/code/davidcanorosillo/car-clustering

titulado : car datasets