Zadanie 2

Spracovanie dát

Skontroloval som, či dáta neobsahovali žiadne duplikáty ani null hodnoty. Vymazal som stĺpce, kde vyskytovala prevažne iba jedna premenná. Takže mi nakoniec ostalo 110 stĺpcov. Normalizoval som dáta. Urobil som korelačnú maticu, z ktorej som zistil že so stĺpcom SalePrice najviac korelujú stĺpce GrLivArea, GarageArea, GarageCars, FullBath.

Rozhodovací strom

Použil som **RandomForestRegressor().** Na najlepšie nájdenie parametrov som použil GridSearchCV. Jedná sa o gridsearch ktorý umožňuje aj cross validáciu. Parametre, ktoré som menil v v gridsearch boli n_estimators – počet stromov v lese, min_samples_split – minimálny počet vzoriek potrebný na rozdelenie uzla, max_depth – maximálna hĺbka stromu. Výsledky trénovania som vyhodnocoval pomocu parametru best_score a najlepšie parametere som zobrazil pomocu best_params_.

Prvé trénovanie:

Cross validation = 4

n_estimators: [70, 80, 100, 120]

min_samples_split: [10, 15, 20]

max_depth: [4, 8, 10]

Najlepšie skóre: **0.831**

Najlepšie parametre:

max_depth: 8

min_samples_split: 10

n_estimators: 70

Druhé trénovanie:

Cross validation = 6

Predĺžil sa čas potrebný na trénovanie.

Najlepšie skóre: **0.827**

Najlepšie parametre:

max_depth: 10

min_samples_split: 15

n_estimators: 120

Tretie trénovanie

Cross validation =5

max_depth: [3,4,5]

Najlepšie skóre: **0.823**

Najlepšie parametre:

max_depth: 5

min_samples_split: 15

n_estimators: 100

Štvrté trénovanie

n_estimators: [100, 120, 150]

min_samples_split: [5,6,7]

Najlepšie skóre: **0.822**

Najlepšie parametre:

max_depth: 5

min_samples_split: 6

n_estimators: 120

Piate trénovanie:

n_estimators: [100, 120, 150]

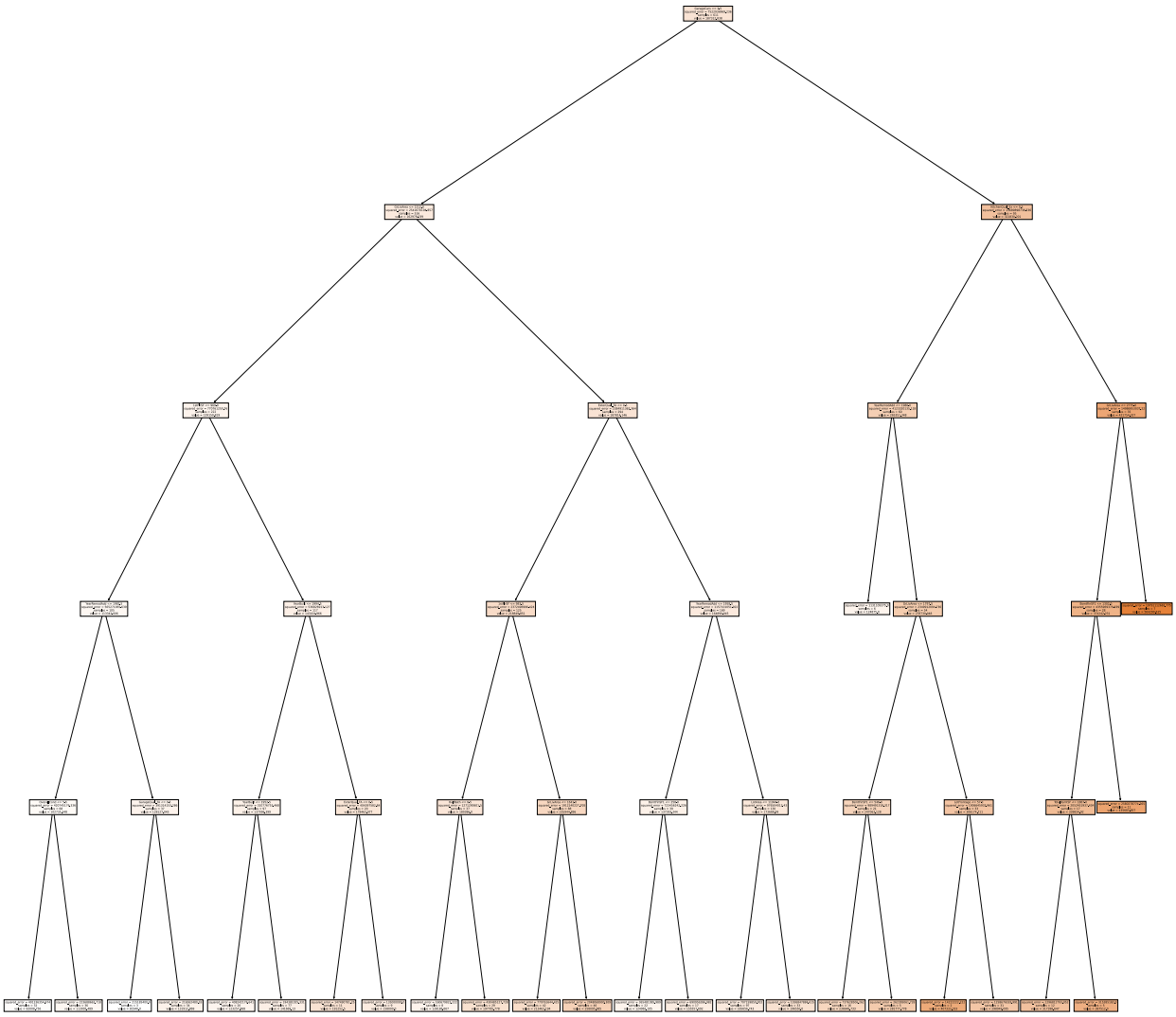
max_depth: [3,4,5]

min_impurity_decrease: [0.5, 1.5, 10]

Najlepšie skóre: **0.823**

Najlepšie parametre: max_depth: 5,'min_impurity_decrease: 1.5, n_estimators: 100

Vizualizácia stromu



SVR

Parametre pre Grid search kernel, C, gamma.

Prvé trénovanie

Cross validation - 5

kernel: [rbf, sigmoid]

C: [0.5, 1, 10]

gamma: [scale, auto]

Najlepšie skóre: **-0.058**

Najlepšie parametre:

C: 10, gamma: scale, kernel: rbf

Druhé trénovanie

kernel: [linear, poly]

C: [0.5, 1, 10]

gamma: [scale, auto]

Najlepšie skóre: **0.112**

Najlepšie parametre:

C: 10, gamma: scale, kernel: linear

Tretie trénovanie

kernel: [linear, rbf]

C: [0.1, 10]

gamma: [0.001, 10,100]

Najlepšie skóre: **0.11**

Najlepšie parametre:

C: 10, gamma:0.001, kernel: linear

Štvrté trénovanie

kernel: [linear, rbf]

C: [100, 300, 500]

gamma: [scale, auto]

Najlepšie skóre: **0.705**

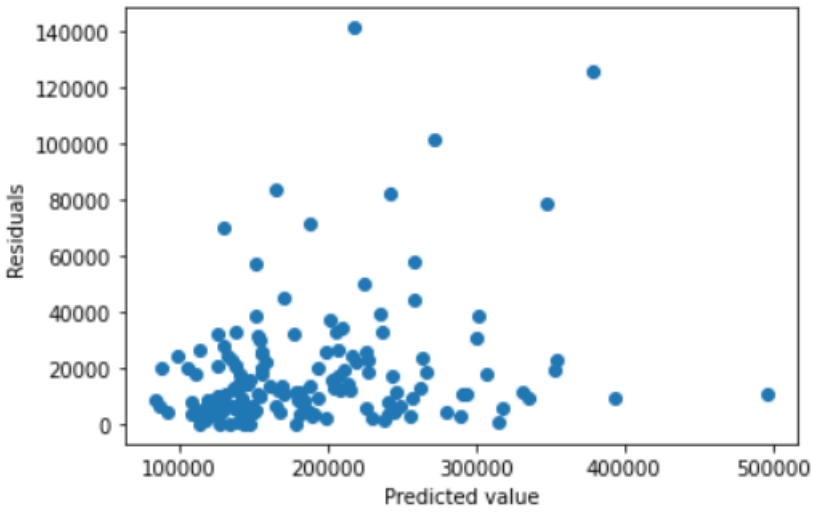
Najlepšie parametre:

C: 500, gamma: scale, kernel: linear

Testovacie dáta

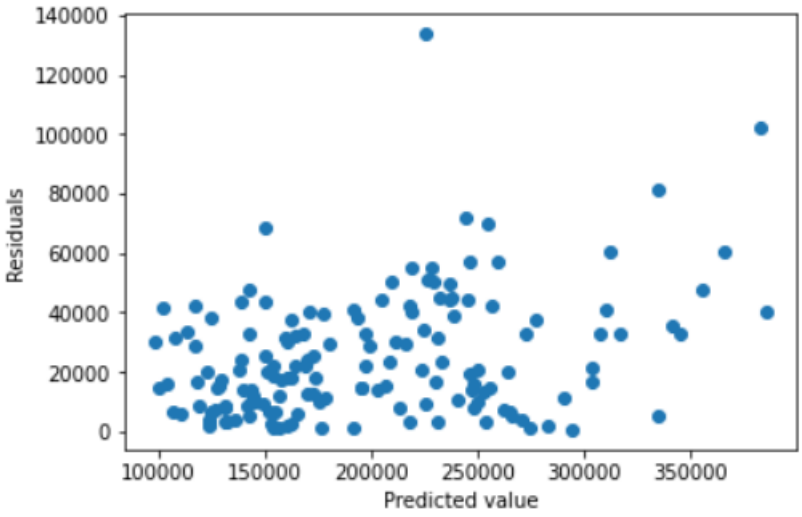
Random forest - max_depth: 10, min_samples_split: 15, n_estimators: 120, Cross validation = 6

Accuracy**: 88.67 %** **R-**squared**:0.862** MSE**: 819306180.5**



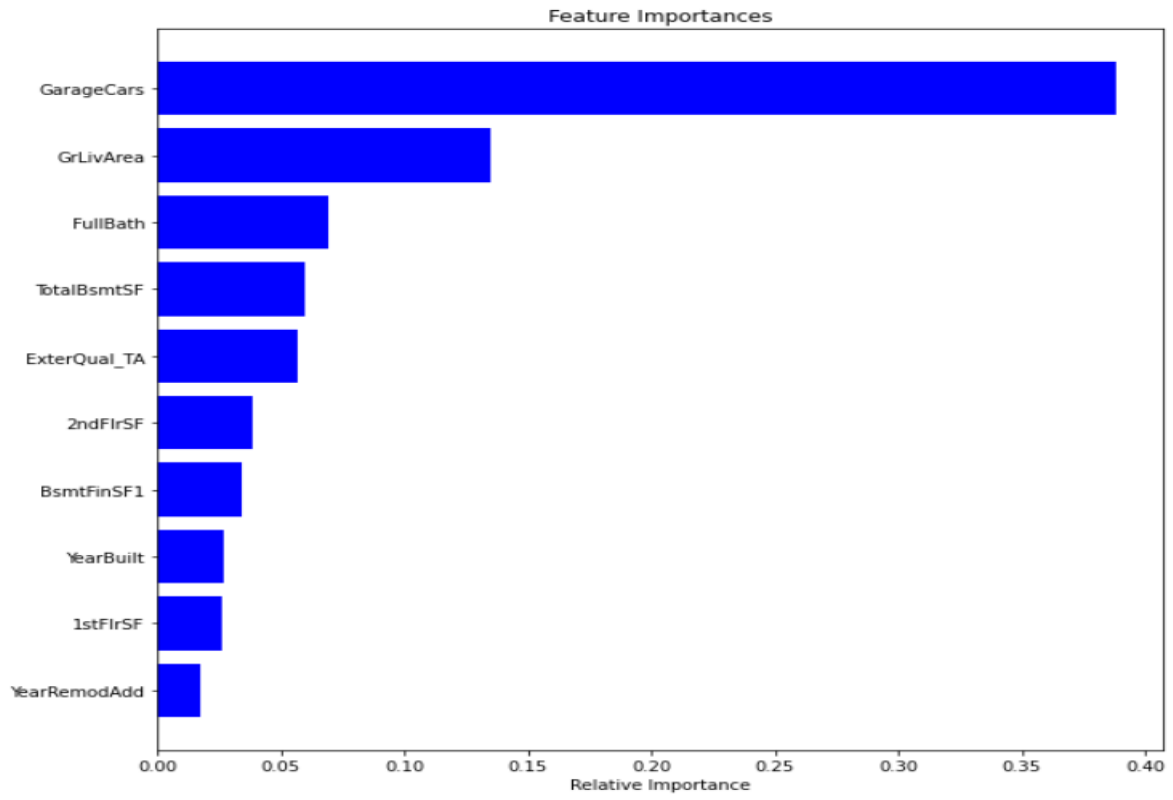
SVR - C: 500, gamma: scale, kernel: linear, Cross validation = 6

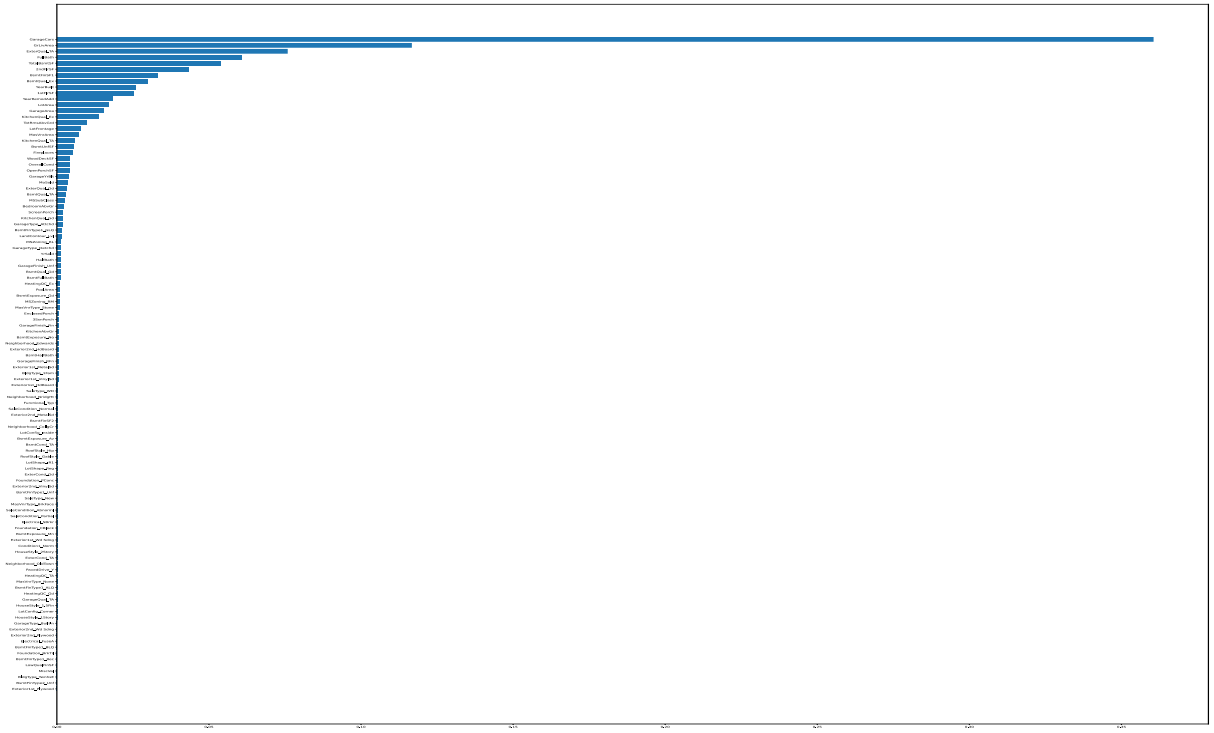
Accuracy**: 84.73 %** R-squared**:0.824** MSE**: 1047308376.16**



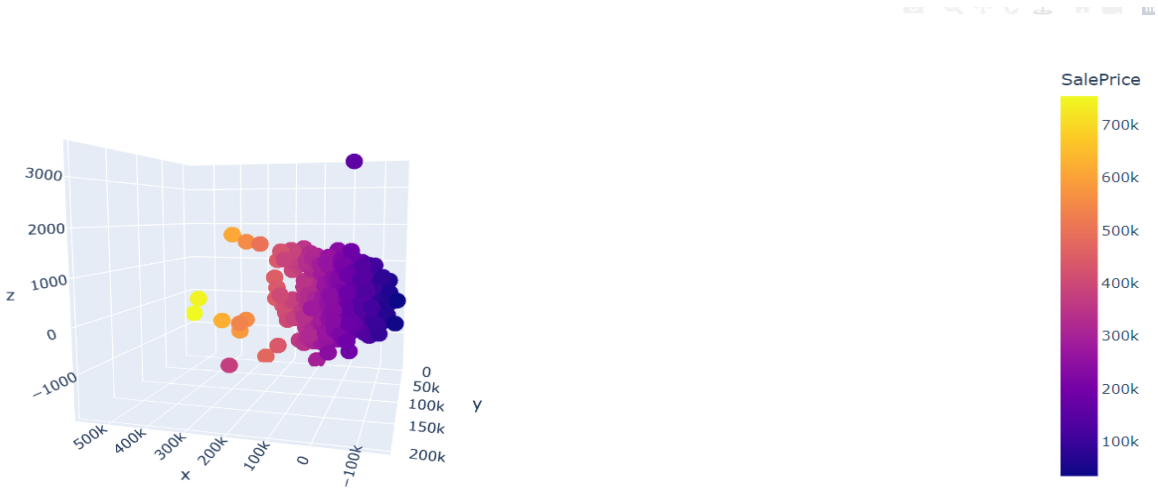
Rozhodovací strom dosiahol lepšie výsledky na testovacích dátach oproti SVR, aj reziduály sú menšie.

Dôležitosť vstupných parametrov pre vybraný strom



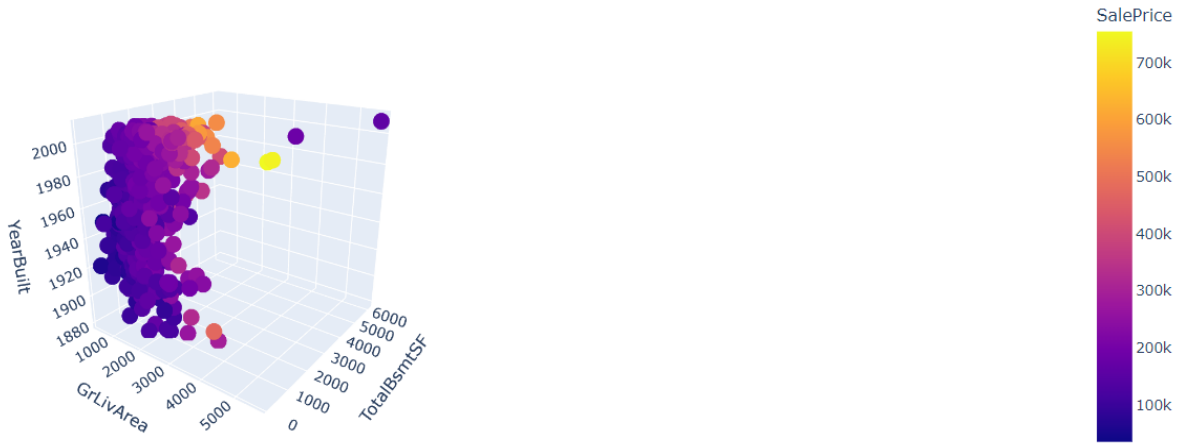


Graf z minimalizovanej množiny pomocou PCA



Ak sme zredukovali data len na 3 premenné x,y,z tak môžeme vidieť, že cena nehnuteľnosti sa najviac odvíja od parametru x. S narastajúcim parametrom x sa zvyšuje aj cena nehuteľnosti.

Graf výslednej ceny od dátum výstavby, priestoru, celkovej plochy suterénu



S narastajúcou veľkosťou priestoru nehnuteľnosti sa zvyšuje aj cena nehnuteľnosti. Cena sa taktiež mierne zvyšuje aj podľa roku výstavby, čím je stavba novšia taka je cena je vyššia. Veľkosť celkovej plochy suterénu nemá výrazný vplyv na cenu nehnuteľnosti.

Podmnožina X príznakov

Zredukovanie na **6 parametrov:**

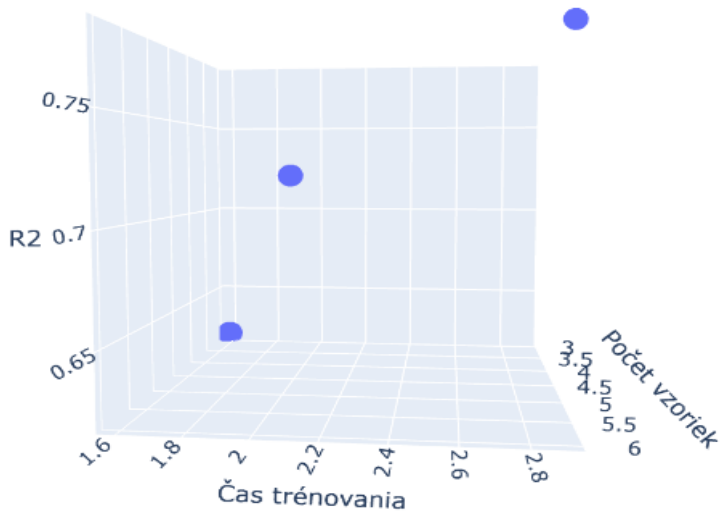
Čas trénovania 2.904s Best score:: 0.797 R-squared: 0.782

Zredukovanie na **5 parametrov:**

Čas trénovania 2.033s Best score::0.779 R-squared:0.722

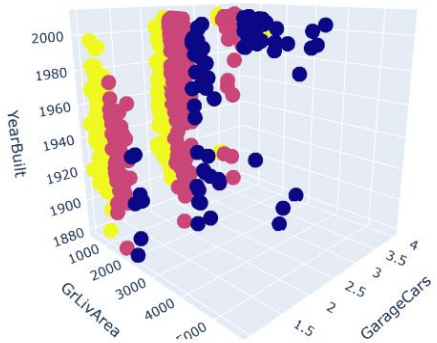
Zredukovanie na **3 parametre:**

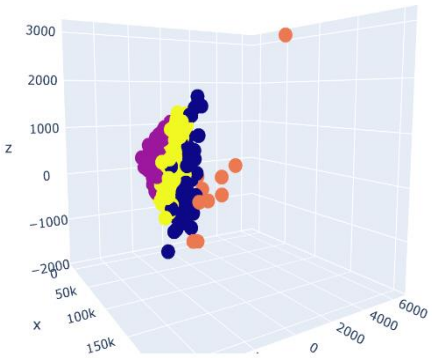
Čas trénovania 1.583s Best score::0.648 R-squared: 0.621



Zhlukovanie dát pomocou KMeans

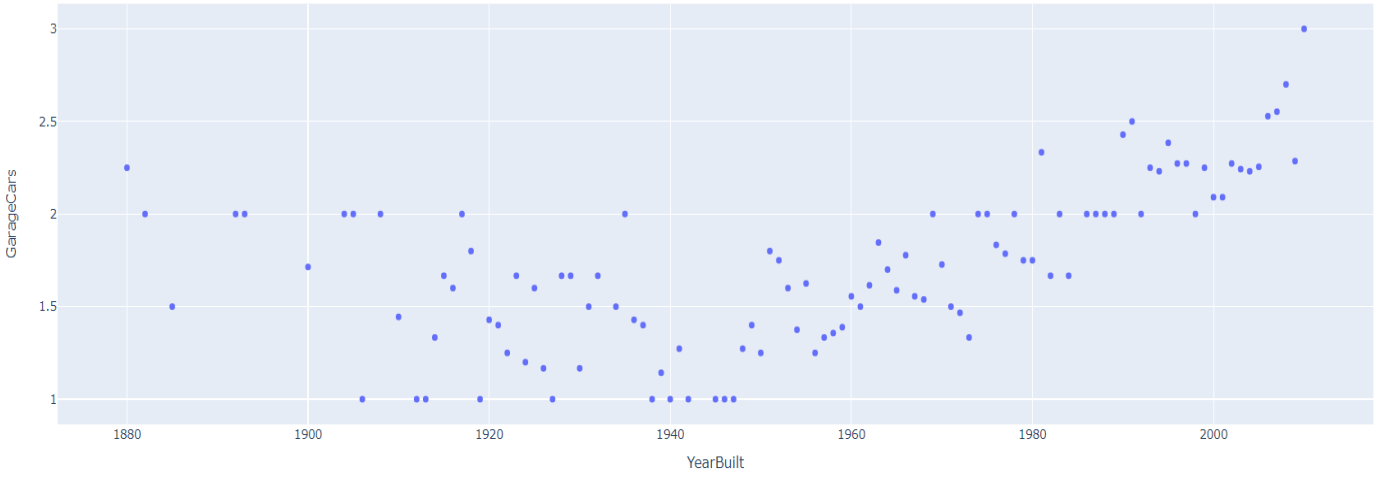
n_cluster = 3,4





Na grafe sú zobrazené nehnuteľnosti v závislosti od počtu áut, ktoré sa dajú zaparkovať v garáži, roku výstavby a priestoru. Dáta sme pomocou KMeans rozdelili na pravo do 3 skupín a na ľavom grafe, kde je redukcia pomocou PCA, do 4 skupín.

EDA



Na grafe vidíme vzťah medzi rokom kedy boli nehnuteľnosti postavené a priemerom počtu áut, ktoré sa dajú zaparkovať do garáže. Z grafu vyplýva, že ak je stavba novšia tak počet áut, ktoré sa dajú zaparkovať do garáže stúpa. Veľký nárast si môžeme všimnúť od roku 2000, kde nehnuteľnosti majú garáže pre 2 a viac áut.