Data Description

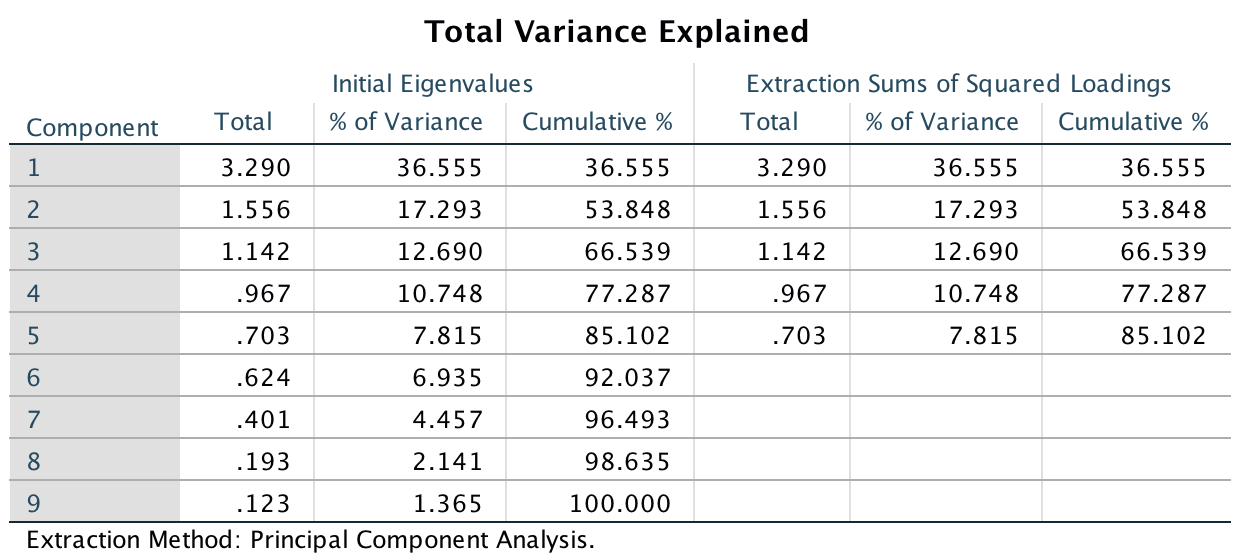
The data set for this project we have selected we have collected it from Kaggle and the data is about King country from Seattle It includes homes sold between May 2014 and May 2015. houses information below we will have some example of the data explanation. The data includes information about houses, prices and the inside information of the houses and selling and purchasing and the size and location information also. 19 house features plus the price and the id columns, along with 21613 observations. The columns comprises of id , date, price, bedrooms, sqft\_living. Below is the example of some the data.



Reference sources of the data set (<https://www.kaggle.com/harlfoxem/housesalesprediction>)

**Factor Analysis:**

To find the variance and observation between our selected factors we used Factor Analysis, so we can Analyze the Total variance, % of Variance below in the table is given detail the method performed by ***SPSS*** functionality of Factor Analysis the result is as below.



**Component Matrix**

**Component Matrix**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 | 5 |
| price | .665 | .411 | .322 | -.068 | .214 |
| bedrooms | .720 | .266 | -.349 | .204 | -.336 |
| floors | .563 | -.444 | .283 | .161 | .451 |
| waterfront | .126 | .290 | .695 | -.501 | -.258 |
| condition | -.178 | .666 | -.332 | -.210 | .499 |
| zipcode | -.303 | .251 | .469 | .739 | .025 |
| new\_bathrooms | .910 | .128 | -.153 | .148 | -.124 |
| yr\_built (Binned) | .515 | -.686 | -.006 | -.178 | .082 |
| sqft\_living (Binned) | .883 | .204 | .017 | .017 | .051 |

Extraction Method: Principal Component Analysis.5 components extracted.

We must explain why we choose only 3 factors instead of 4.

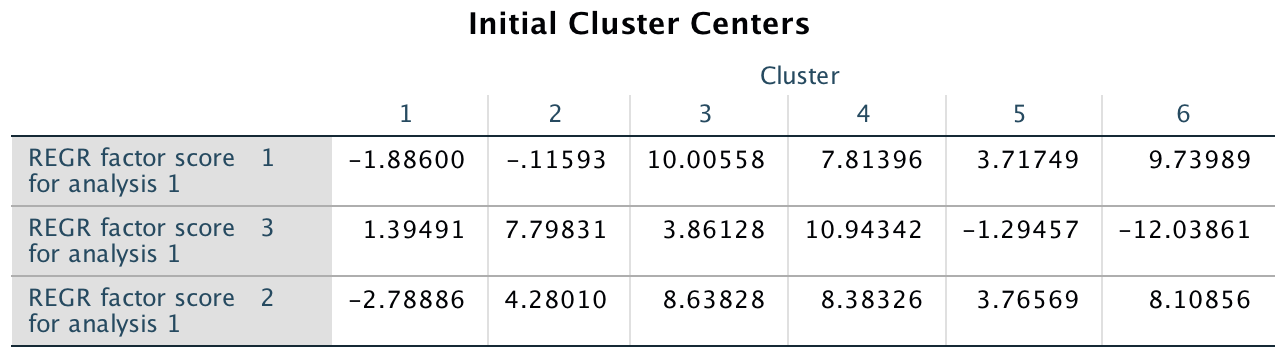
We highlight our new factors the most important components for each one. For factors number 5 we don’t have any very important component then the others.

**2Step Cluster Analysis:**

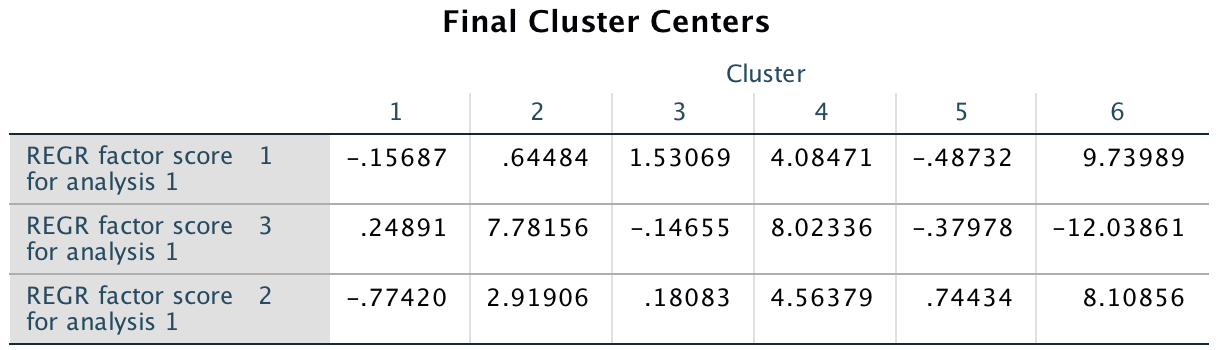
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Cluster** | **percent(values)** | **percent(values)** | **V4** | **V5** |
| 1 | 0.819 | 177 | 1 | 0.8189515569333272 |
| 2 | 13.3253 | 2,880 | 2 | 13.325313468745662 |
| 3 | 18.4657 | 3,991 | 3 | 18.465738213112477 |
| 4 | 27.2197 | 5,883 | 4 | 27.21972886688567 |
| 5 | 22.3292 | 4,826 | 5 | 22.329153750057834 |
| 6 | 17.8411 | 3,856 | 6 | 17.841114144265024 |

With the 2Step cluster analysis we try find the best number of clusters. The algorithm gives us as the best choice K = 6. Analyzing the Silhouette graph, we can see that the quality of the clusters is not very good, but this is the best result we could get from this dataset.

  
K-Means

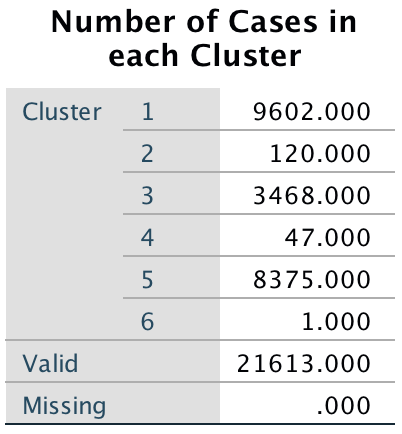


We did the analysis considering the first 3 factors we found with the factor analysis



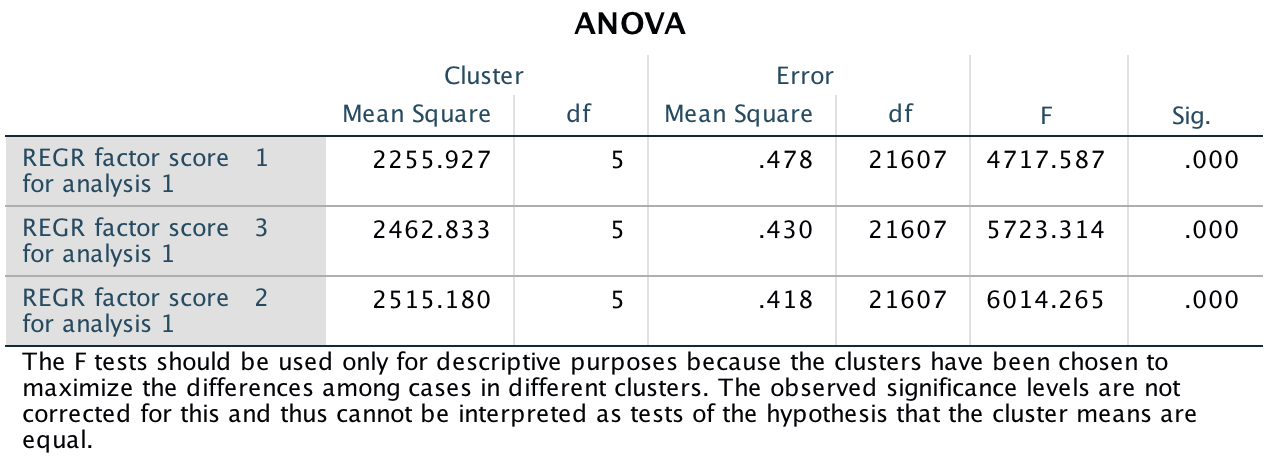
We will explain and describe the cluster. We think that for example Cluster 4 and 6 has the huge amount of bathroom bedrooms and maybe they are big houses, because in the factor analysis we found that factors 1 is more correlated to bedrooms, bathrooms and sqft\_living.

**All the 3 factors are significant.**



In particular, for CLUSTER 6 the observation is the n° 15871. We present now the value for this one

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Price | Bedroom | Sqft\_liv | Floors | Waterfront | Condition | Year | Zipcode | Bathrooms |
| 6.4E+005 | 33 | 1620 (group n°7) | 1.0 | 0 | 5 | 1947  (group n° 4) | 98103 | 58.00 |



|  |  |  |
| --- | --- | --- |
| Factor 1 | Factor 2 | Factor 3 |
| 9.73989 | 8.10856 | -12.03861 |

We think this observation is the only one in the cluster number 6 because for Factor n° 1 (found with the factor analysis) we have that the data are explained most with bedrooms ( 72%) bathrooms (91%) and this house has a very high number of this two things.

**“Forecasting” The clusters for new houses**

K-means is not a forecasting alghoritm, but we tried to understand in which cluster the house we used for the forecasting process in k-neighbour could be part of.

For doing this we calculate for every cluster the average value of the only “important” factors we found with the factor analysis ( in particular we are speaking about price, square feet leaving of the house, the zip code, waterfront and number of floors,

|  |  |  |
| --- | --- | --- |
| Criterion function | 350.883164 | from anova |