**K-means Clustering**

We used k means clustering as second technique for our data set. Is\_canceled, stays\_in\_weekend\_nights, stays\_in\_week\_nights, adults, children, babies, reserved\_room\_type, assigned\_room\_type, customer\_type are our attributes. We changed data types of reserved\_room\_type, assigned\_room\_type, customer\_type for nominal to numeric.

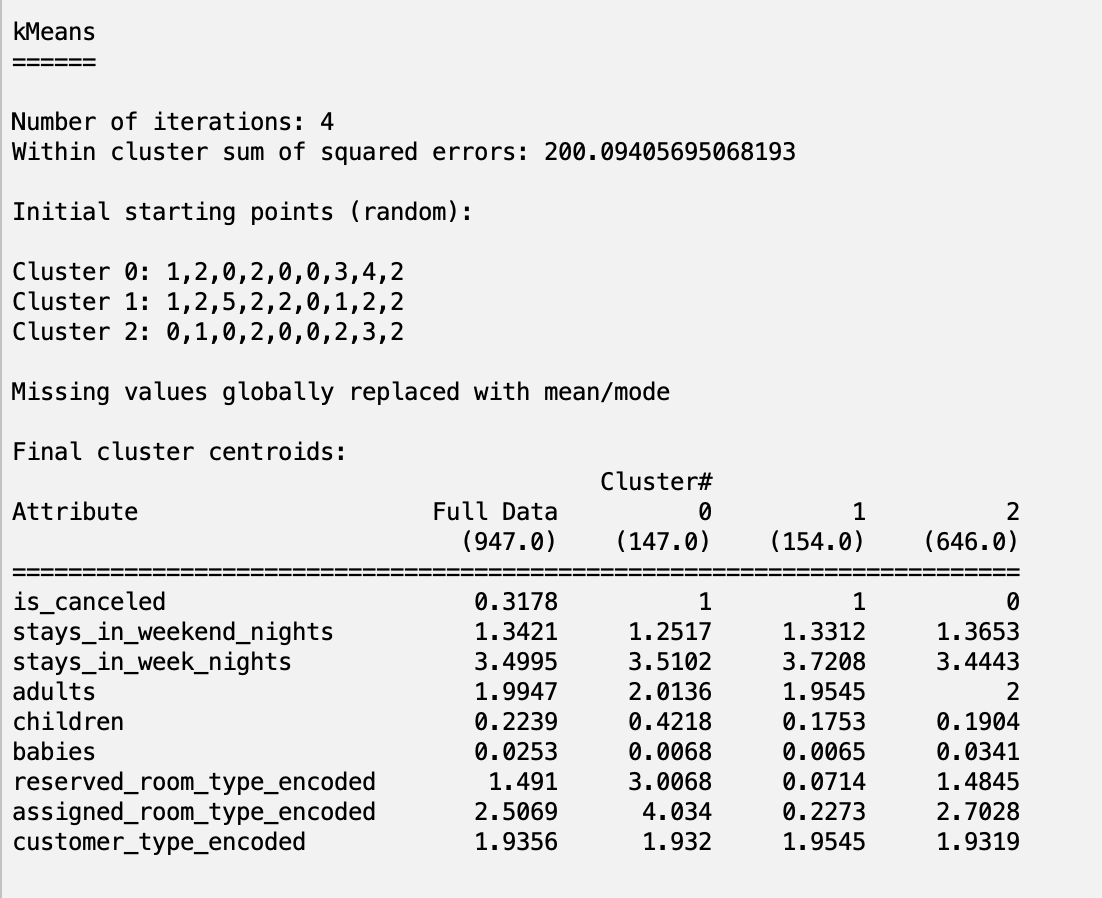
**1st test:**

First tested the dataset with 3 clusters. Then we got output with iterations = 4 and sum of squared errors = 200.094056.

Cluster 0: 1,2,0,2,0,0,3,4,2

Cluster 1: 1,2,5,2,2,0,1,2,2

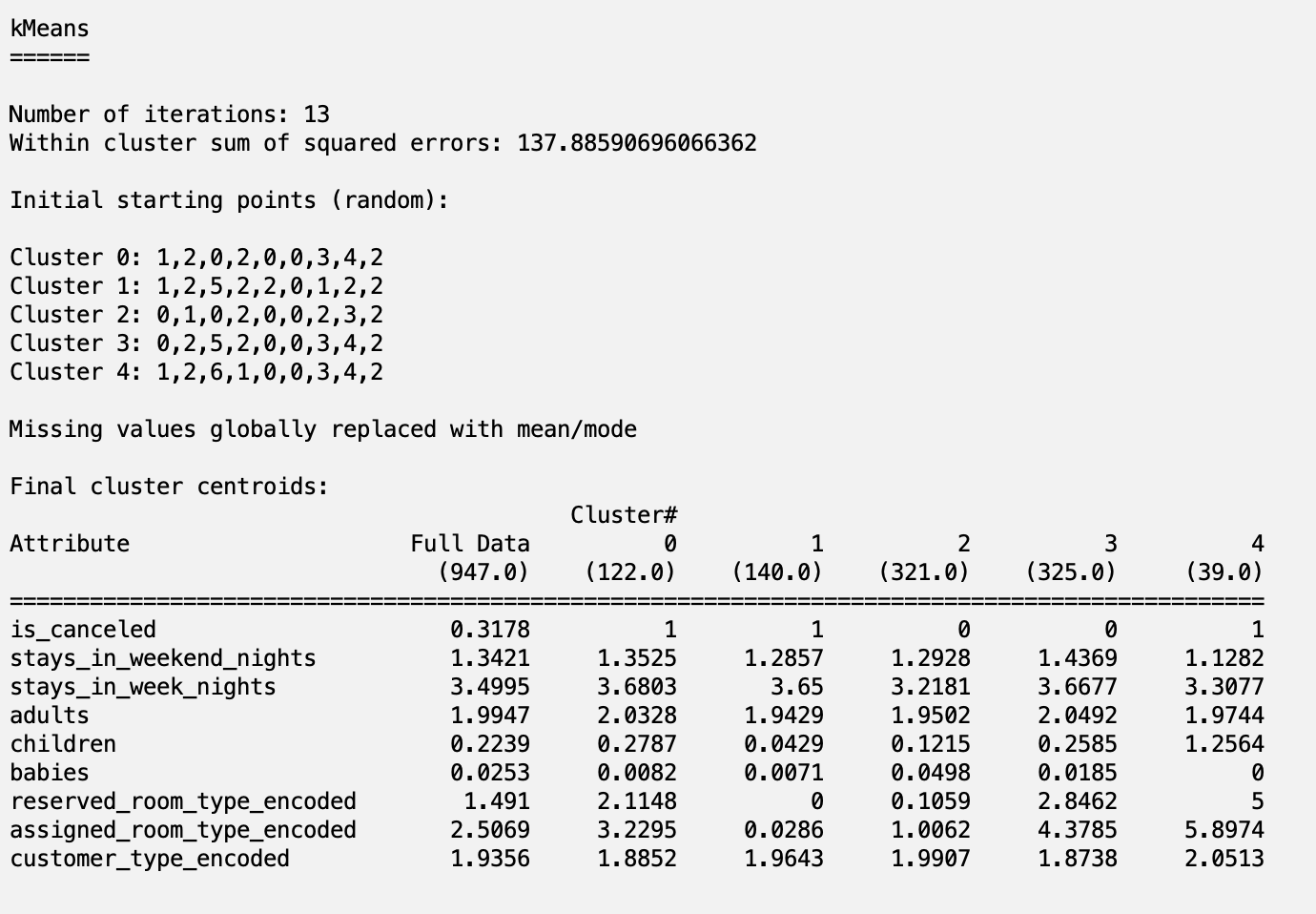
Cluster 2: 0,1,0,2,0,0,2,3,2



**2nd test:**

Then we tested our dataset with 5 total clusters. We got iterations = 13 and sum of squared error = 137.8859069.

Since the sum of squared error is lower in 2nd test with 5 clusters, we selected it for analysis.



**Clusters for the test:**

Cluster 0: 1,2,0,2,0,0,3,4,2

Cluster 1: 1,2,5,2,2,0,1,2,2

Cluster 2: 0,1,0,2,0,0,2,3,2

Cluster 3: 0,2,5,2,0,0,3,4,2

Cluster 4: 1,2,6,1,0,0,3,4,2

**Clustered Instances**

0 122 ( 13%)

1 140 ( 15%)

2 321 ( 34%)

3 325 ( 34%)

4 39 ( 4%)

**Attributes used**

* is\_canceled
* stays\_in\_weekend\_nights
* stays\_in\_week\_nights
* adults
* children
* babies
* reserved\_room\_type\_encoded
* assigned\_room\_type\_encoded
* customer\_type\_encoded

**Cluster 0 (122 observations):**

* is\_canceled: 1
* stays\_in\_weekend\_nights: 1.3525
* stays\_in\_week\_nights: 3.6803
* adults: 2.0328
* children: 0.2787
* babies: 0.0082
* reserved\_room\_type\_encoded: 2.1148
* assigned\_room\_type\_encoded: 3.2295
* customer\_type\_encoded: 1.8852

**Cluster 1 (140 observations):**

* is\_canceled: 1
* stays\_in\_weekend\_nights: 1.2857
* stays\_in\_week\_nights: 3.65
* adults: 1.9429
* children: 0.0429
* babies: 0.0071
* reserved\_room\_type\_encoded: 0
* assigned\_room\_type\_encoded: 0.0286
* customer\_type\_encoded: 1.9643

**Cluster 2 (321 observations):**

* is\_canceled: 0
* stays\_in\_weekend\_nights: 1.2928
* stays\_in\_week\_nights: 3.2181
* adults: 1.9502
* children: 0.1215
* babies: 0.0498
* reserved\_room\_type\_encoded: 0.1059
* assigned\_room\_type\_encoded: 1.0062
* customer\_type\_encoded: 1.9907

**Cluster 3 (325 observations):**

* is\_canceled: 0
* stays\_in\_weekend\_nights: 1.4369
* stays\_in\_week\_nights: 3.6677
* adults: 2.0492
* children: 0.2585
* babies: 0.0185
* reserved\_room\_type\_encoded: 2.8462
* assigned\_room\_type\_encoded: 4.3785
* customer\_type\_encoded: 1.8738

**Cluster 4 (39 observations):**

* is\_canceled: 1
* stays\_in\_weekend\_nights: 1.1282
* stays\_in\_week\_nights: 3.3077
* adults: 1.9744
* children: 1.2564
* babies: 0
* reserved\_room\_type\_encoded: 5
* assigned\_room\_type\_encoded: 5.8974
* customer\_type\_encoded: 2.0513

**Conclusions for clustering**

Cluster 0 and Cluster 4 represent groups with high cancellation rates. Cluster 0 has slightly more children and higher room types. Cluster 4 has significantly more children and very high room types.

Cluster 1 also shows high cancellation rates but includes very few children and no babies, and lower room types.

Cluster 2 and Cluster 3 represent groups with no cancellations. Cluster 2 has a shorter stay duration and fewer children, while Cluster 3 has a longer stay duration and more children. This group should be target for marketing and loyalty programs.