EDA ASSIGNMENT 6

PART 1

5420 Anomaly Detection, Fall 2020

- Harsh Dhanuka, hd2457

Considerations

1. Revised 2 features

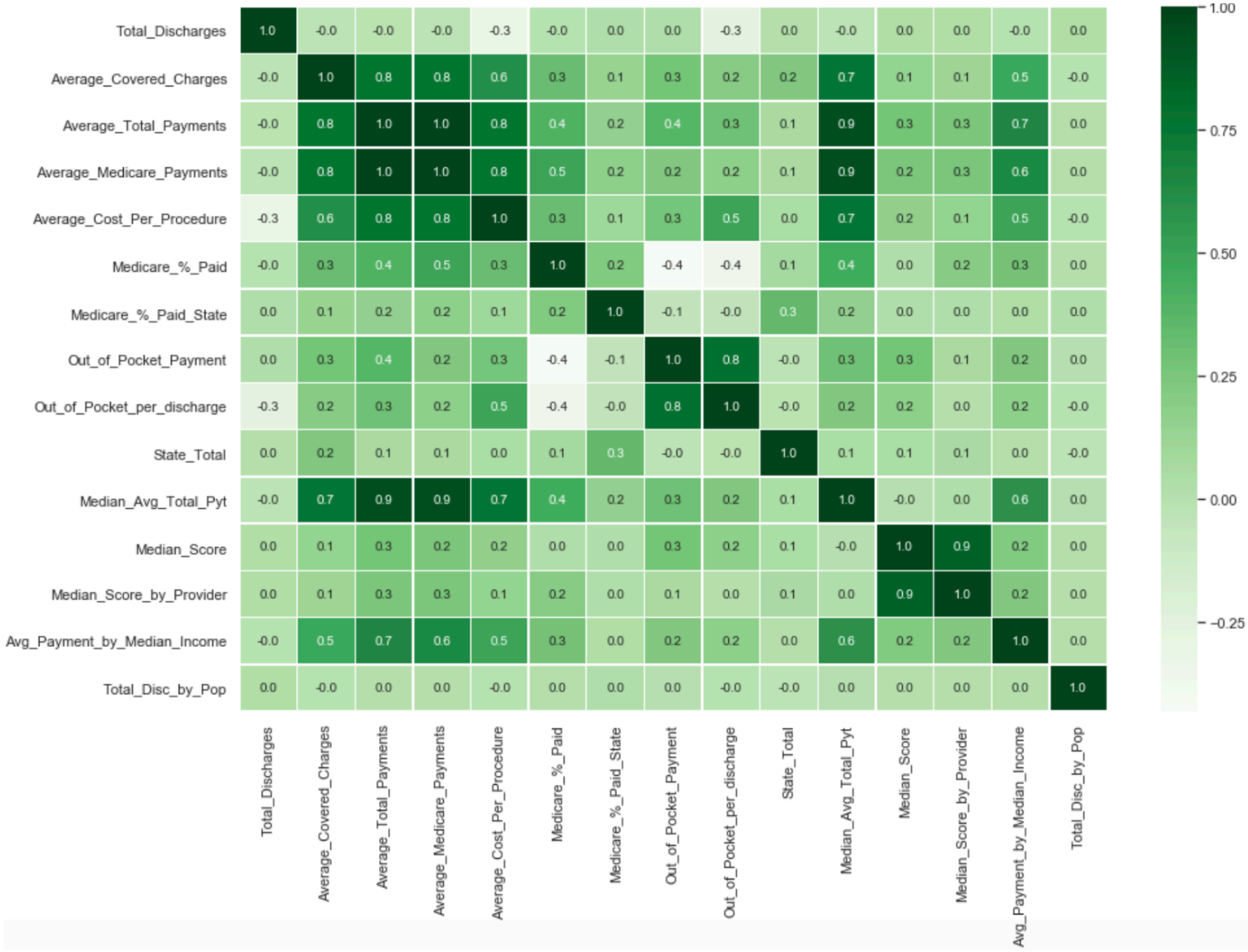
- Grouped the 'Sum total of Total Discharged' by Provider Name, and State
- 'Median Score by Provider' grouped by the Provider Name, and State

2. Drop 7 variables with high multi-collinearity

3. Split to train_test:

- 75% split, train has 75% of the data.
- Now, for the test data, I will be using the **entire 100% data**, as even the train data has anomalies.

Heatmap for multi-collinearity



kNN Clustering

1 **Model**

Build initial model, and check stability by using the 'Average' aggregate method

2 **Scores**

The average scores go from -0.5 to 120. So, I make a subset with scores less than 4.0 to visualize better.

3 **Reasonable Boundaries**

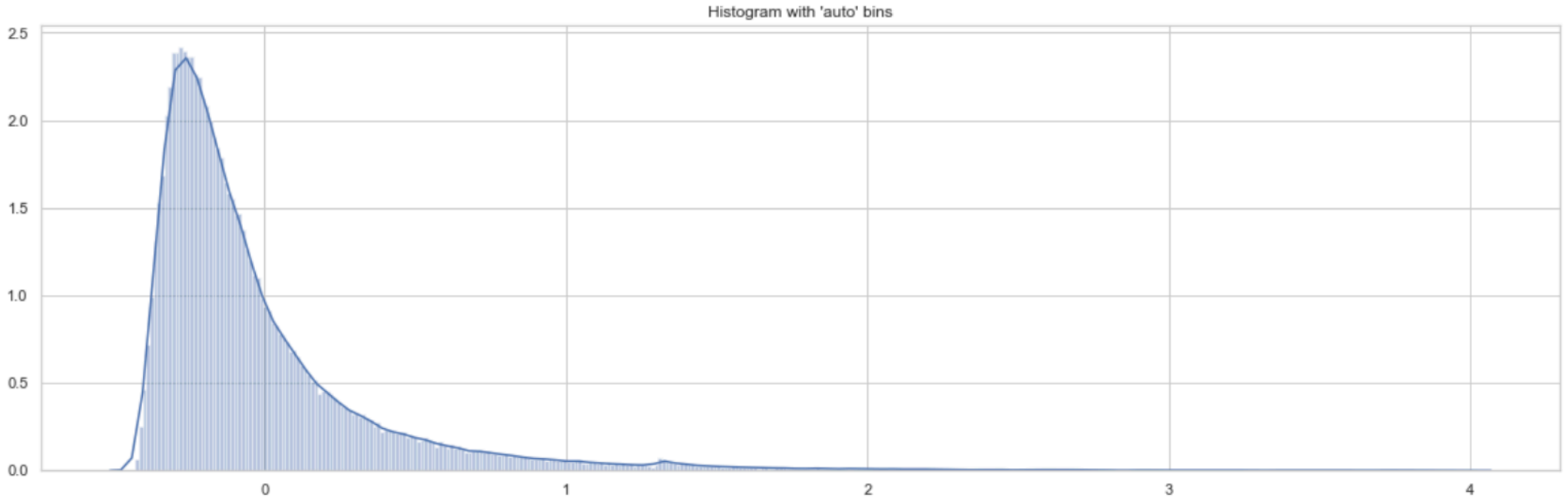
I will chose 3 different cut points, which are:

0.0

1.0

5.0

This will result in a 4 cluster analysis.



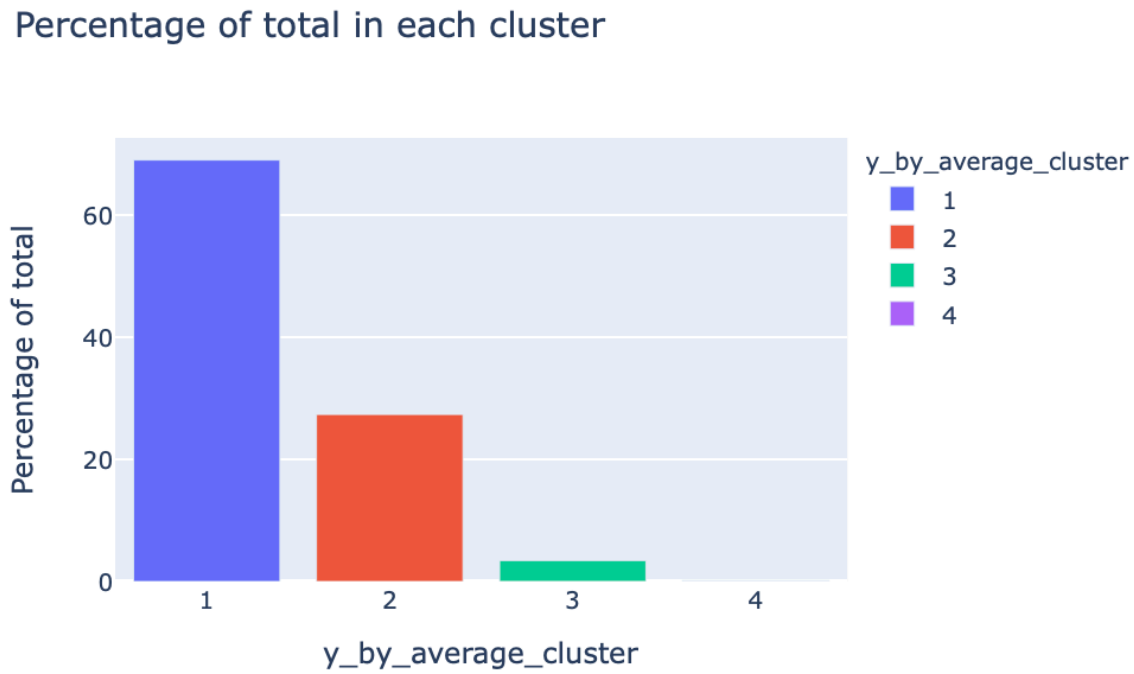
kNN Clustering

4 **Clusters**

Check the statistics of the 4 clusters.

Here, I am showing the percentage of data points in each cluster

Percentage of total	y_by_average_cluster
69.033821	1
27.369454	2
3.439733	3
0.156993	4



Cluster Evaluation

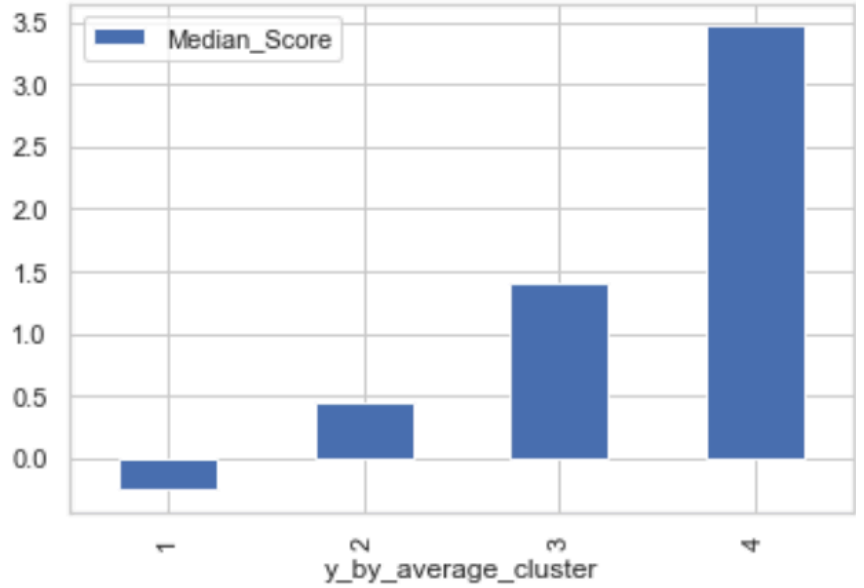
Cluster wise variable Averages

5 **Cluster 3 and 4 have less than 5% of the total data points**

Out of these two clusters, cluster 4 has extremes or high standard deviation from mean for some variables, and hence, I will consider this cluster as suspicious.

Feature-wise cluster EDA will be imperative to justify this claim.

On the right, I evaluate a feature 'Median Score'



cluster	Total_Discharges	Average_Total_Payments	Medicare_%_Paid	Medicare_%_Paid_State	Out_of_Pocket_Payment	State_Total	Median_Score	Total_Disc_by_Pop
1	-0.221173	-0.283433	0.061537	-0.022363	-0.256388	0.033268	-0.252521	-0.030195
2	0.384266	0.492422	-0.066014	0.078355	0.325327	-0.077728	0.439989	-0.020812
3	1.236714	1.643783	-0.652100	-0.200777	2.240825	-0.039097	1.408806	0.203568
4	3.167817	2.770374	-1.263044	0.572458	6.927906	-0.221593	3.466872	12.445800

Cluster Evaluation

6 **Scores**

y-by-average-score gives us insights about the clusters which are anomalies, as the anomalies might have a very high score compared to others.

I see that cluster 4 has a score almost 13-14 times higher than all other clusters. So, I can safely conclude that Cluster 4 is highly suspicious.

y_by_average_cluster	y_by_average_score
1	-0.211456
2	0.282751
3	1.697698
4	13.632964

So, I would pass on the 256 specific entries of the Cluster 4 to the relevant authorities, and call for further investigation on each of the entries, to understand of they are true anomalies. I will provide all the reasoning as I have highlighted above, as to the differences in the means, and walk through the process I have done.

PCA Clustering

1 **Model**

Build initial model, and check stability by using the 'Average' aggregate method

2 **Scores**

The scores go from -0.1 to 70. So I make a subset of scores less than 4.0 to visualize better.

3 **Reasonable Boundaries**

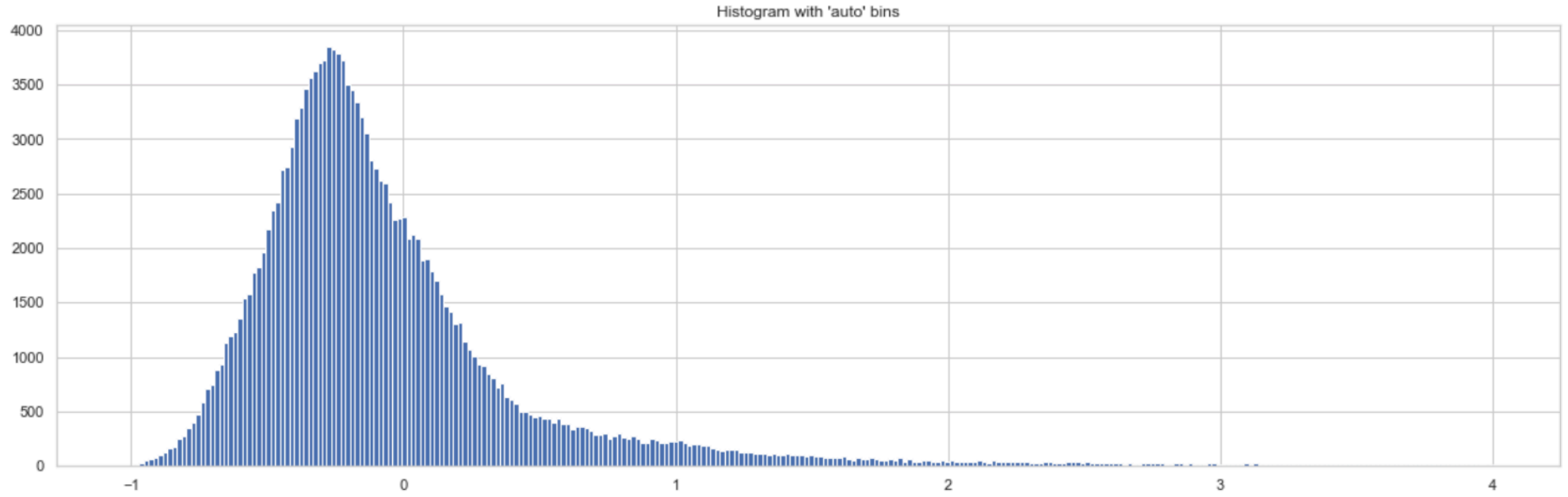
I will chose 3 different cut points, which are:

0.0

1.0

5.0

This will result in a 4 cluster model.



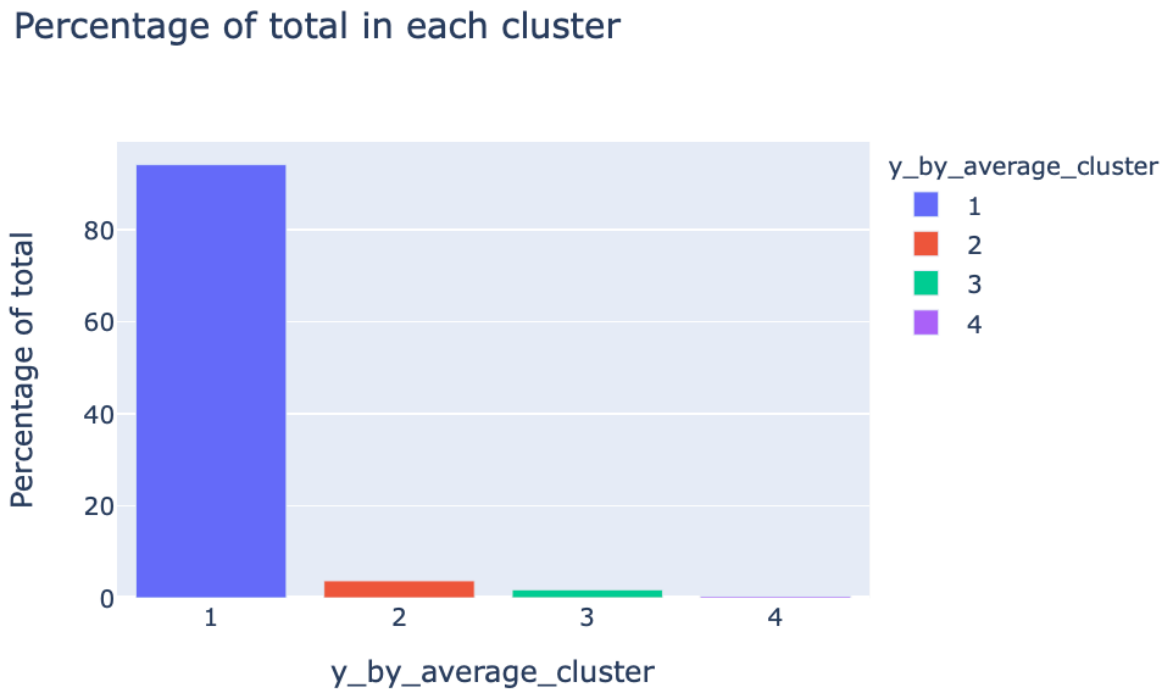
kNN Clustering

4 **Clusters**

Check the statistics of the 4 clusters.

Here, I am showing the percentage of data points in each cluster

Percentage of total	y_by_average_cluster
94.115230	1
3.715696	2
1.781498	3
0.387576	4



Cluster Evaluation

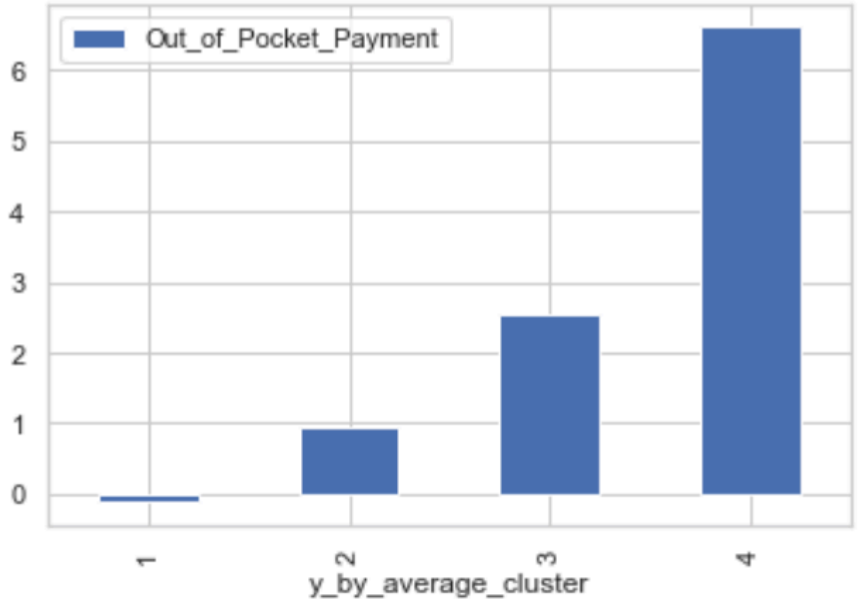
Cluster wise variable Averages

5 **Cluster 2, 3 and 4 have less than 5% of the total data points**

Out of these three clusters, cluster 4 has extremes or high standard deviation from mean for some variables, and hence, I will consider this cluster as suspicious.

Feature-wise cluster EDA will be imperative to justify this claim.

On the right, I evaluate a feature 'Out of Pocket Payment'



cluster	Total_Discharges	Average_Total_Payments	Medicare_%_Paid	Medicare_%_Paid_State	Out_of_Pocket_Payment	State_Total	Median_Score	Total_Disc_by_Pop
1	-0.061422	-0.120624	0.026056	-0.004790	-0.112412	-0.005606	-0.080405	-0.024842
2	0.590131	1.522103	-0.208337	0.007671	0.944993	0.062190	1.004088	0.008102
3	1.311465	2.572240	-0.606691	0.187180	2.530000	0.169234	1.571540	0.159379
4	3.229441	2.875316	-1.541189	0.229214	6.608139	-0.012786	2.674932	5.222078

Cluster Evaluation

6 **Scores**

y-by-average-score gives us insights about the clusters which are anomalies, as the anomalies might have a very high score compared to others.

I see that cluster 4 has a score almost 13-14 times higher than all other clusters. So, I can safely conclude that Cluster 4 is highly suspicious.

y_by_average_cluster	y_by_average_score
1	-0.150710
2	1.368969
3	2.935768
4	9.372393

So, I would pass on the 638 specific entries of the Cluster 4 to the relevant authorities, and call for further investigation on each of the entries, to understand of they are true anomalies. I will provide all the reasoning as I have highlighted above, as to the differences in the means, and walk through the process I have done.