Date: 22nd October 2019

Predict Future Sales:
Final project for "How to win a data science competition"

Team DataScavengers: Anurag Bejju – abejju@sfu.ca, Savitaa Venkateswaran - savitaav@sfu.ca

---

Problem Statement:

As part of this assignment, we have explored clustering algorithms to group shops and item categories based on the available transactions.

Density Based Clustering (Shops):

We intend to cluster shops based on median value of items per shop sold for month, median value of items returned per shop for month, revenue monthly based using density based clustering algorithm.

<table>
<tr><th colspan="2">Density Based Clustering (Shops):</th></tr>
<tr><td>Features for Shops vs Shops clustering

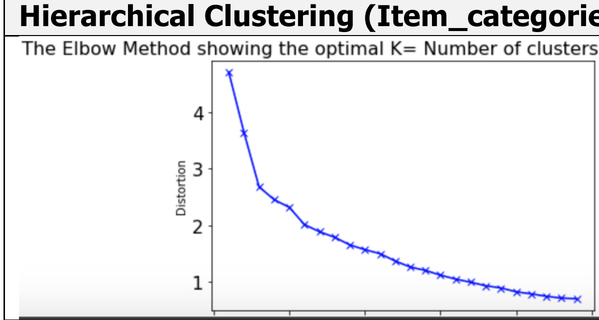
Feature 1 shop_id
Feature 2 median_total_items_sold_permonth
Feature 3 highest_total_items_sold_permonth
Feature 4 lowest_total_items_sold_permonth
Feature 5 median_no_of_transactions_permonth
Feature 6 highest_no_of_transactions_permonth
Feature 7 lowest_no_of_transactions_permonth
Feature 8 median_revenue_per_month
Feature 9 highest_revenue_per_month
Feature 10 lowest_revenue_per_month
Feature 11 median_total_returned_items_permonth
Feature 12 highest_total_returned_items_permonth
Feature 13 lowest_total_returned_items_permonth</td><td>As part of our feature generation to do this clustering, we have computed:

Net of items sold (bought-returned) per shop per month
Number of items returned per shop per month,
Net Revenue per shop per month
Number of items sold per shop per month</td></tr>
</table>

We have also removed shops having less than *5 months of sales data* (it could mean either they are closed shops or newly opened shops) as well as removed returned items from net items sold as they might have been bought at some point. Finally, we have computed high, median, low values for all the attributes found and used them to cluster shops. We have accounted for shops that didn't have any items being returned per month, (that is no negative values) by assigning 0 to it. In order to tackle varying ranges across different features, we have scaled them before clustering to improve results as well as reduce dimensionality issues. We haven't removed outliers as DBSCAN can account for it.

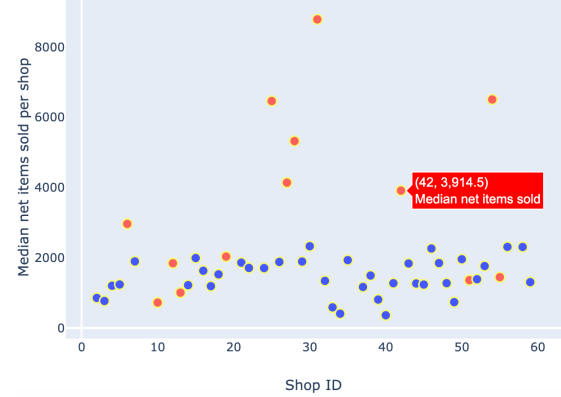
Hierarchical Clustering (Item_categories):

We have used hierarchical clustering to group items based on *number of items in an item category, number of items being sold per item category per month, median price of the item sales per category per month and number of times that was returned per category per month*. Just like the above case, we used median, lowest and highest values in each of the above defined attributes in order to ensure accuracy is not impacted due to skewed- datasets.

<table>
<tr><th colspan="2">Hierarchical Clustering (Item_categories)</th></tr>
<tr><td>
</td><td>K-means was used to determine good n_clusters (number of clusters to you think the dataset has) for agglomerative hierarchical clustering. Using elbow method with and without scaling datasets, a good k value was determined. Based on this, we take "5" (where we can see the elbow take a sharp cut) with scaled dataset and perform bottom up clustering.</td></tr>
</table>

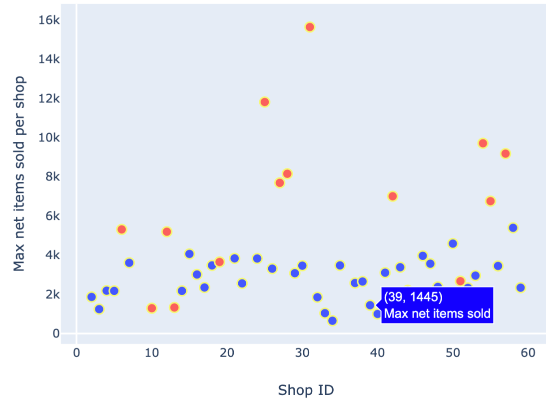
Density Based Clustering (Shops) Results

Average (Median) Net Items sold per Shop



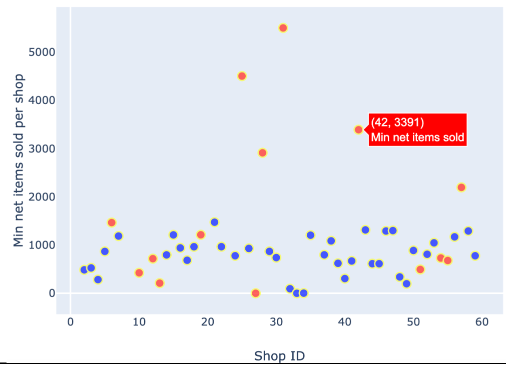
Observation: The result of clustering using DBSCAN for shops with respect to a single feature from the full feature set- median of values over the entire dataset of net items sold/bought per shop on a monthly basis.

Highest Net Items sold per Shop



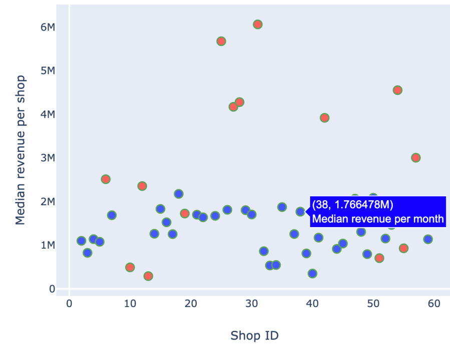
Observation: The result of clustering using DBSCAN for shops with respect to a single feature from the full feature set- highest/maximum value of net item sold over the entire dataset per shop on a monthly basis.

Lowest Net Items sold per Shop



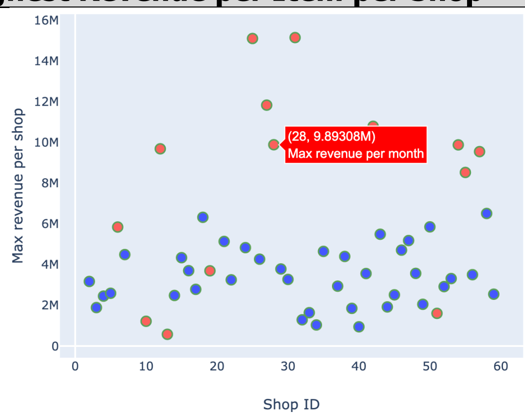
Observation: The result of clustering using DBSCAN for shops with respect to a single feature from the full feature set- lowest/minimum value of net item sold over the entire dataset per shop on a monthly basis.

Median Revenue per Item per Shop



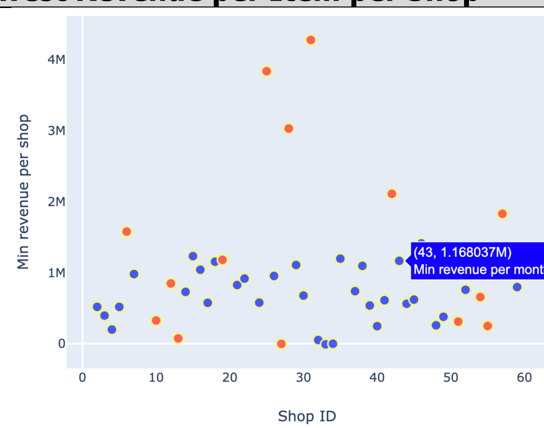
Observation: The result of clustering using DBSCAN for shops with respect to a single feature from the full feature set- median of values over the entire dataset for revenue made per shop on a monthly basis.

Highest Revenue per Item per Shop

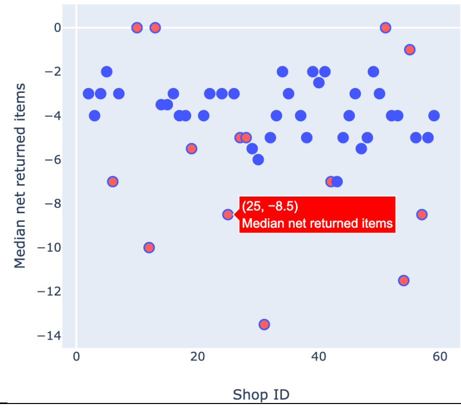
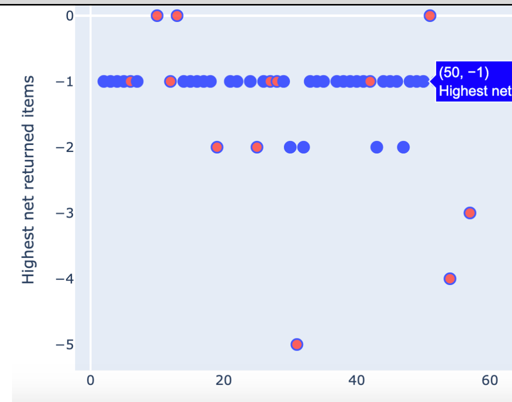


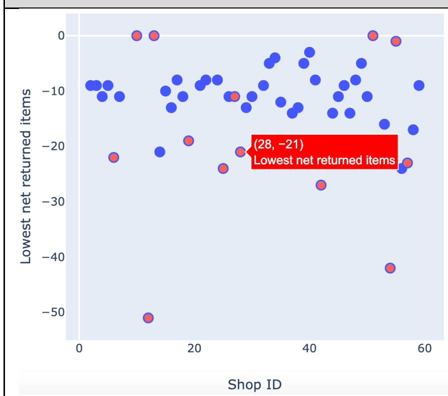
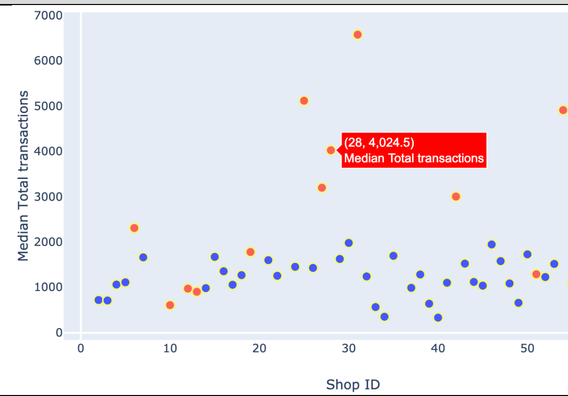
Observation: The result of clustering using DBSCAN for shops with respect to a single feature from the full feature set- highest/maximum value over the entire dataset for revenue made per shop on a monthly basis.

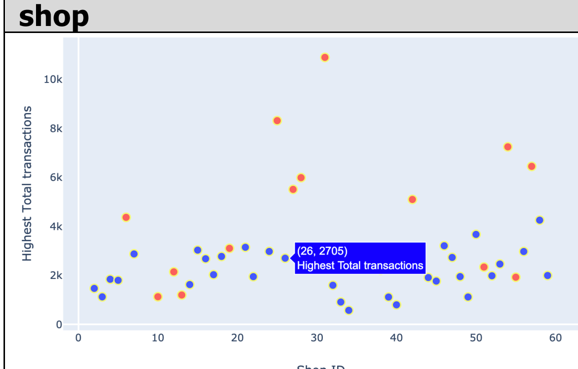
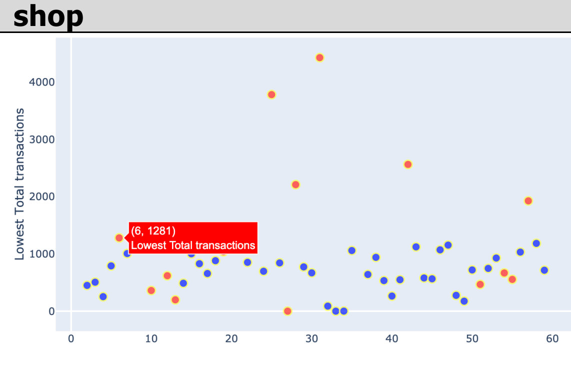
Lowest Revenue per Item per Shop



Observation: The result of clustering using DBSCAN for shops with respect to a single feature from the full feature set- lowest/minimum value over the entire dataset for revenue made per shop on a monthly basis.

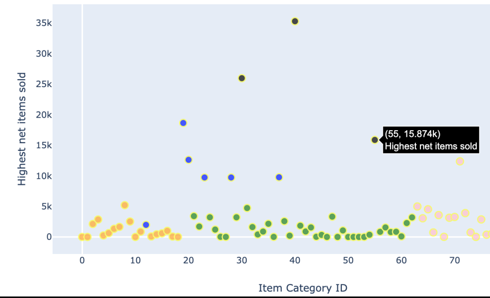
Average (Median) Total Items returned per Shop	**Highest Total Items returned per Shop**
	

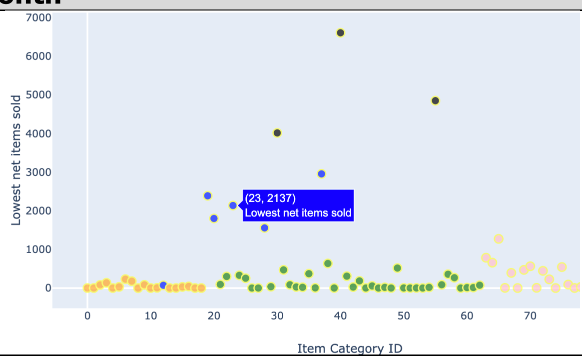
<u>Observation:</u> the result of clustering using DBSCAN for shops with respect to a single feature from the full feature set- median of values over the entire dataset for number of items returned per shop on a monthly basis.	***<u>Observation:</u>*** Using DBSCAN for shops with respect to a single feature from the full feature set- highest/maximum value over the entire dataset for number of items returned per shop on a monthly basis.
Lowest Total Items returned per Shop	**Median Total transactions per month per shop**
	

<u>Observation:</u> Using DBSCAN for shops with respect to a single feature from the full feature set- lowest/minimum value over the entire dataset for number of items returned per shop on a monthly basis.	***<u>Observation:</u>*** Using DBSCAN for shops with respect to a single feature from the full feature set- median of values over the entire dataset for number of items transacted for/total transactions made per shop on a monthly basis.
Highest Total transactions per month per shop	**Lowest Total transactions per month per shop**
	

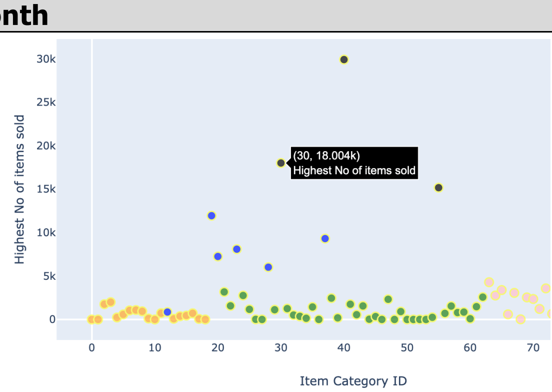
<u>Observation:</u> using DBSCAN for shops with respect to a single feature from the full feature set- highest/maximum value over the entire dataset for number of items transacted for/total transactions made per shop on a monthly basis.	***<u>Observation:</u>*** using DBSCAN for shops with respect to a single feature from the full feature set- lowest/minimum value over the entire dataset for number of items transacted for/total transactions made per shop on a monthly basis.

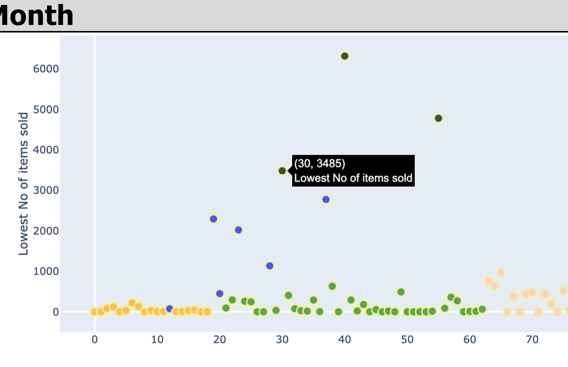
<table>
<tr><th>Average (Median) Net Items sold per Category per Month</th><th>Highest Net Items sold per Category per Month</th></tr>
<tr><td>


</td><td>


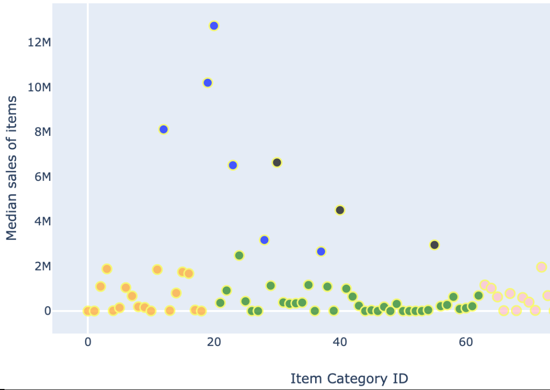
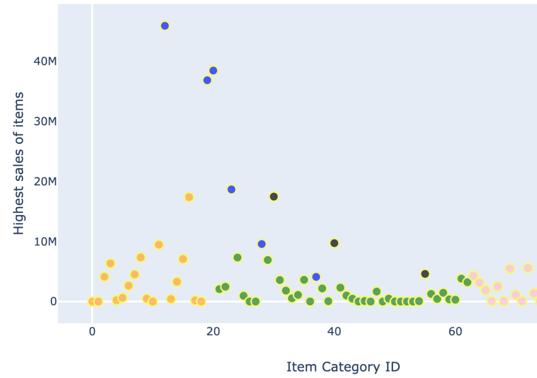
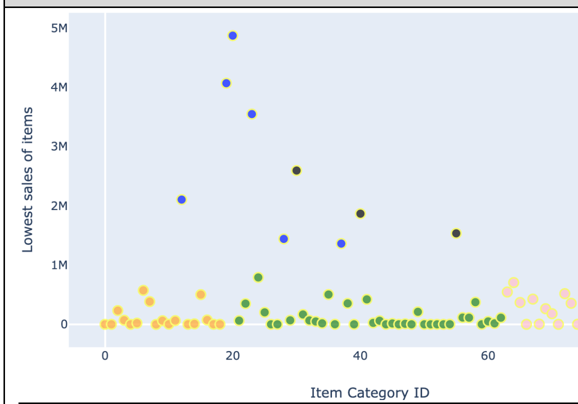
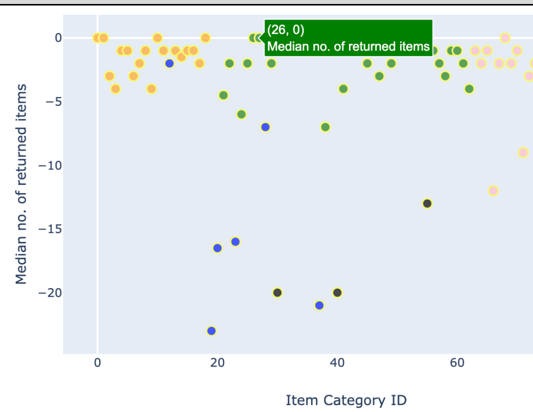
</td></tr>
<tr><td><u>Observation:</u> using Agglomerative Hierarchal clustering for item categories with respect to a single feature from the full feature set- median of values over the entire dataset for net items sold per category on a monthly basis.</td><td><u>Observation:</u> using Agglomerative Hierarchal clustering for item categories with respect to a single feature from the full feature set- highest. maximum value over the entire dataset for net items sold per category on a monthly basis.</td></tr>
<tr><th>Lowest Net Items sold per Category per Month</th><th>Median Number of Items per Category per Month</th></tr>
<tr><td>


</td><td>


</td></tr>
<tr><td><u>Observation:</u> Using Agglomerative Hierarchal clustering for item categories with respect to a single feature from the full feature set- lowest/minimum value over the entire dataset for net items sold per category on a monthly basis.</td><td><u>Observation:</u> Using Agglomerative Hierarchal clustering for item categories with respect to a single feature from the full feature set- median of values over the entire dataset for number of items per category on a monthly basis.</td></tr>
<tr><th>Highest Number of Items per Category per Month</th><th>Lowest Number of Items per Category per Month</th></tr>
<tr><td>


</td><td>


</td></tr>
<tr><td><u>Observation:</u> using Agglomerative Hierarchal clustering for item categories with respect to a single feature from the full feature set- highest/maximum value over the entire dataset for number of items per category on a monthly basis.</td><td><u>Observation:</u> Agglomerative Hierarchal clustering for item categories with respect to a single feature from the full feature set- lowest/minimum value over the entire dataset</td></tr>
</table>

Average (Median) Sales Value of item Category per Month	Highest Sales Value of item Category per Month
	
<u>Observation:</u> using Agglomerative Hierarchal clustering for item categories with respect to a single feature from the full feature set- median of values over the entire dataset for sales made per item category on a monthly basis.	***<u>Observation:</u>*** using Agglomerative Hierarchal clustering for item categories with respect to a single feature from the full feature set- highest/maximum value over the entire dataset for sales made per item category on a monthly basis.
Lowest Sales Value of item Category per Month	**Median Number of items returned per item cat gory transactions per month**
	
<u>Observation:</u> using Agglomerative Hierarchal clustering for item categories with respect to a single feature from the full feature set- lowest/minimum value over the entire dataset for sales made per item category on a monthly basis.	***<u>Observation:</u>*** using Agglomerative Hierarchal clustering for item categories with respect to a single feature from the full feature set- median of values over the entire dataset for items returned back per item category on a monthly basis.

Conclusion:

Our inferences from this intense clustering exercise is that, there is a clear hierarchy between item categories, which is kind of well explained intuitively. The same can be seen through our work (images attached) as well. This implicit hierarchy was also a strong motivation for us to choose agglomerative hierarchal clustering, adding-on we used well-defined metrics to evaluate our n_clusters for hierarchal clustering.

Secondly, we chose DBSCAN with a motive to understand a common trend/pattern in relation to buy/sell of different items between shops. We wanted to see if there is a monthly trend, hence we made sure our features for clustering resonated that well. We were able to see that even though there isn't a very close association or strong association in reference to commonality between the shops, there is indeed a range within which the sales seems to happen.

```
print(clustering)
```

```
DBSCAN(algorithm='auto', eps=3, leaf_size=30, metric='euclidean',
    metric_params=None, min_samples=3, n_jobs=5, p=None)
```

```
shop_df['cluster_labels'].unique()
```

```
array([ 0, -1])
```

```
print(clustering.core_sample_indices_)
```

```
[ 0  1  2  3  5  9 10 11 12 13 15 16 17 19 22 26 27 28 29 30 31 32 33 36
 37 38 39 40 41 42 44 45 48 51]
```

```
print(len(shop_df[shop_df['cluster_labels']==0]))
print(len(shop_df[shop_df['cluster_labels']==-1]))
```

```
0
0
0
38
14
```

The dataset shows a divide of these sales based metrics derived from the it (all transactions) clearly through our DBSCAN approach. All these findings are immensely useful in understanding the dataset better and will help us curate models that fit this dataset well. For instance strong, closely knit clusters probably teases us with the information that these groups can be dealt with separate models for predicting sales to gain improved accuracy in the challenge.