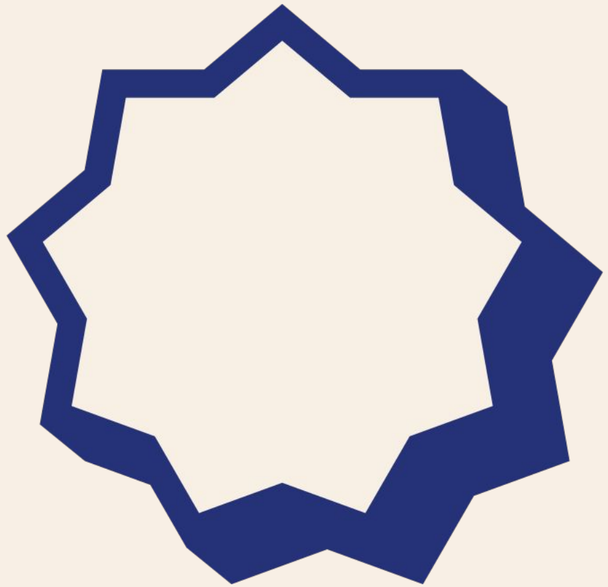
2024



Dongkyu Kim | Marukha Hussaini | Natalie Lira | Sachin Pandya

Project 4

Objectives

Develop an Unsupervised Model

We began by developing an unsupervised model to analyze restaurant data and to separate the values into two different clusters

Standardize the data to allow for comparison

All data was scaled to the same power for ease of comparison and conclusion

Future Goals

Our Future Goals are to then tweak this model so that it can integrate into a restaurant review service like Yelp and produce groups of restaurants as “Go” and “Don’t Go”

Original Data

- The original dataset consists of 17 columns and 8368 rows.
- Includes both categorical and numerical values
- Requires standardization and cleaning of values for comparison

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8368 entries, 0 to 8367
Data columns (total 17 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Name                    8368 non-null   object
 1   Location                8368 non-null   object
 2   Cuisine                 8368 non-null   object
 3   Rating                  8368 non-null   float64
 4   Seating Capacity        8368 non-null   int64
 5   Average Meal Price      8368 non-null   float64
 6   Marketing Budget        8368 non-null   int64
 7   Social Media Followers  8368 non-null   int64
 8   Chef Experience Years   8368 non-null   int64
 9   Number of Reviews       8368 non-null   int64
 10  Avg Review Length       8368 non-null   float64
 11  Ambience Score          8368 non-null   float64
 12  Service Quality Score   8368 non-null   float64
 13  Parking Availability    8368 non-null   object
 14  Weekend Reservations    8368 non-null   int64
 15  Weekday Reservations    8368 non-null   int64
 16  Revenue                 8368 non-null   float64
dtypes: float64(6), int64(7), object(4)
memory usage: 1.1+ MB
```

Standardize the Data and PCA

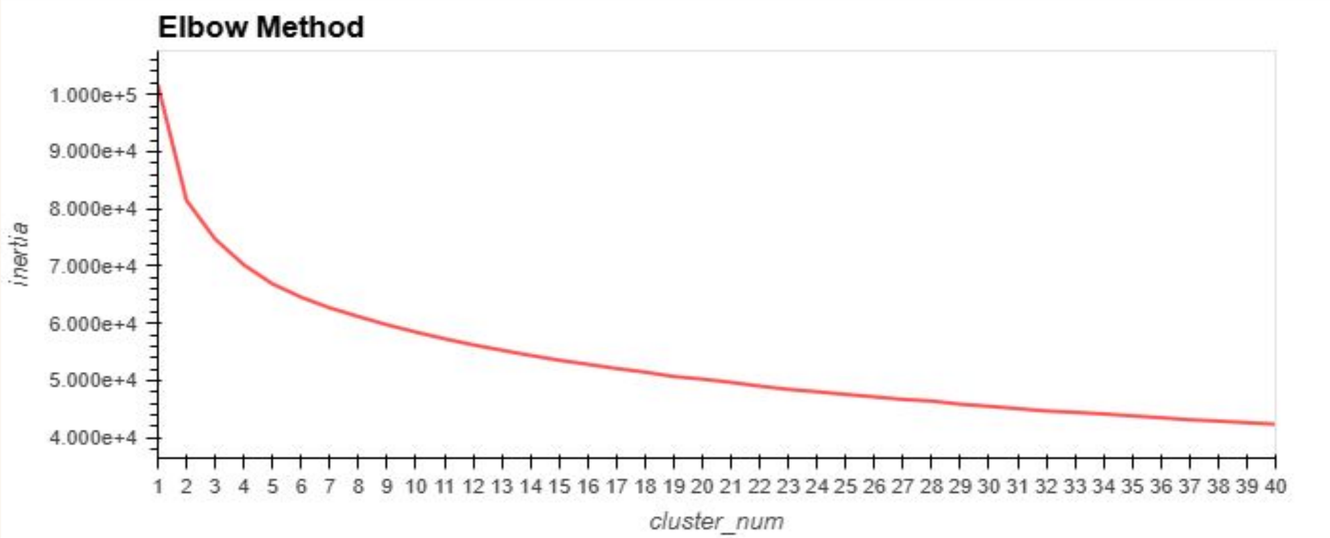
	Rating	Seating Capacity	Average Meal Price	Marketing Budget	Social Media Followers	Chef Experience Years	Number of Reviews	Avg Review Length	Ambience Score	Service Quality Score	Weekend Reservations	Weekday Reservations	Revenue
0	-0.014202	-1.276714	1.819441	-0.544861	-0.686274	0.534421	-1.219380	-0.178419	-1.639150	0.576566	-0.823590	-1.261571	-0.064043
1	-1.390099	0.907389	-1.380216	0.656375	0.351622	-0.371987	0.036038	-0.361285	-1.134352	-0.815332	0.924293	-1.161586	-0.620285
2	1.189708	-0.701950	0.027437	-0.231400	0.058746	1.440830	1.190445	-1.637931	-0.085926	0.460574	-0.124437	-0.761648	-0.428956
3	0.673746	-1.506619	0.254838	-1.124106	-1.126017	0.534421	-1.590955	0.425916	-0.357740	-1.047314	-1.023349	-0.611671	-0.940598
4	1.533682	1.597106	1.958951	0.230572	0.213665	-0.190705	-1.605385	0.929408	1.195484	-1.317961	0.374958	-0.161740	3.122598
...	...	...	...	...	...	...	...	...	...	...	...	...	...
8363	-1.046125	-0.357091	-0.910068	-1.159727	-1.336227	0.171858	-0.515913	1.099394	1.544960	-0.196710	0.374958	-1.461540	-0.828044
8364	-0.530163	-0.644473	-0.768466	-0.674191	-0.845917	-0.190705	0.685392	0.011393	-1.095521	-1.124642	0.374958	-0.411702	-0.901626
8365	1.189708	1.597106	-0.071614	1.496473	1.489845	-0.734551	-0.313892	0.669276	-0.280079	-1.472616	2.672176	-0.411702	1.025907
8366	-1.562086	-1.679048	-0.234841	-1.376191	-1.557816	-1.640959	0.743112	0.051572	0.224719	-1.317961	-1.173167	-0.411702	-1.288631
8367	-0.014202	-1.564095	1.616454	-0.665971	-0.640002	-0.371987	-1.176090	-0.318526	0.147058	0.769885	-1.223107	-0.861632	-0.455978

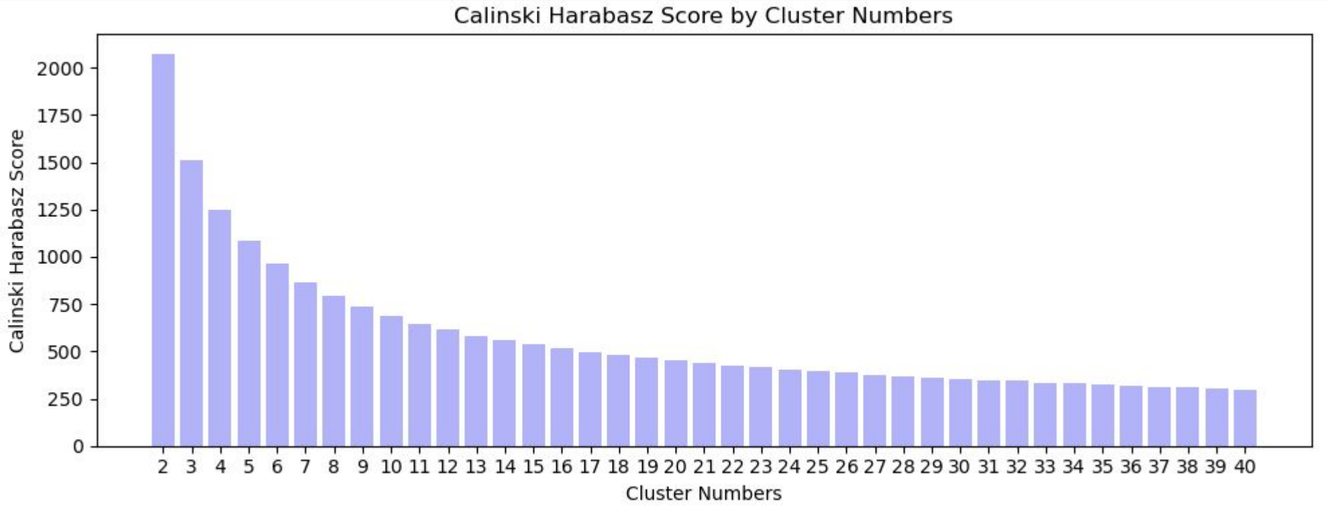
Location_Rural	Location_Suburban	Cuisine_American	Cuisine_French	Cuisine_Indian	Cuisine_Italian	Cuisine_Japanese	Cuisine_Mexican	Parking Availability_No	Parking Availability_Yes
True	False	False	False	False	False	True	False	False	True
False	False	False	False	False	False	False	True	False	True
True	False	False	False	False	True	False	False	True	False
True	False	False	False	False	True	False	False	False	True
False	False	False	False	False	False	True	False	True	False
...	...	...	...	...	...	...	...	...	...
False	True	False	False	True	False	False	False	False	True
True	False	False	False	True	False	False	False	True	False
False	False	False	False	False	True	False	False	False	True
True	False	True	False	False	False	False	False	True	False
True	False	False	False	False	False	True	False	False	True

	PCA_1	PCA_2	PCA_3	PCA_4	PCA_5	PCA_6	PCA_7	PCA_8	PCA_9
0	-1.602329	-1.717388	1.636562	-1.393867	-1.705817	-0.210651	0.097664	0.326914	0.408310
1	0.441447	1.358972	-1.409816	-0.710771	-0.172635	0.511681	0.809583	-0.487749	1.704060
2	-0.779930	0.468890	1.343955	-0.353624	0.538145	0.535154	0.998055	2.139401	0.279931
3	-2.535516	-0.174292	1.015148	-1.686116	-0.077267	-0.640989	-0.547675	-0.862953	-0.334777
4	2.860161	-2.526592	1.211125	-0.760200	0.943812	-0.518266	-1.510619	-1.638053	0.214561
...	...	...	...	...	...	...	...	...	...
8363	-2.168573	0.060189	-1.092439	0.477991	1.108680	-1.067013	-1.265438	-0.399140	1.338251
8364	-1.677503	0.512871	-0.794318	-0.501850	0.174091	0.291397	1.426922	-0.668104	0.547127
8365	3.380806	0.677995	0.008023	-0.402776	0.372431	0.014191	0.324511	-1.757077	2.050849
8366	-3.407766	-0.334320	-0.686713	0.519816	0.756618	1.255804	0.432478	-1.461543	-0.400274
8367	-1.962954	-1.389461	1.659146	-0.247970	-0.900507	0.429666	-1.181607	0.155208	-0.164609

8368 rows × 9 columns

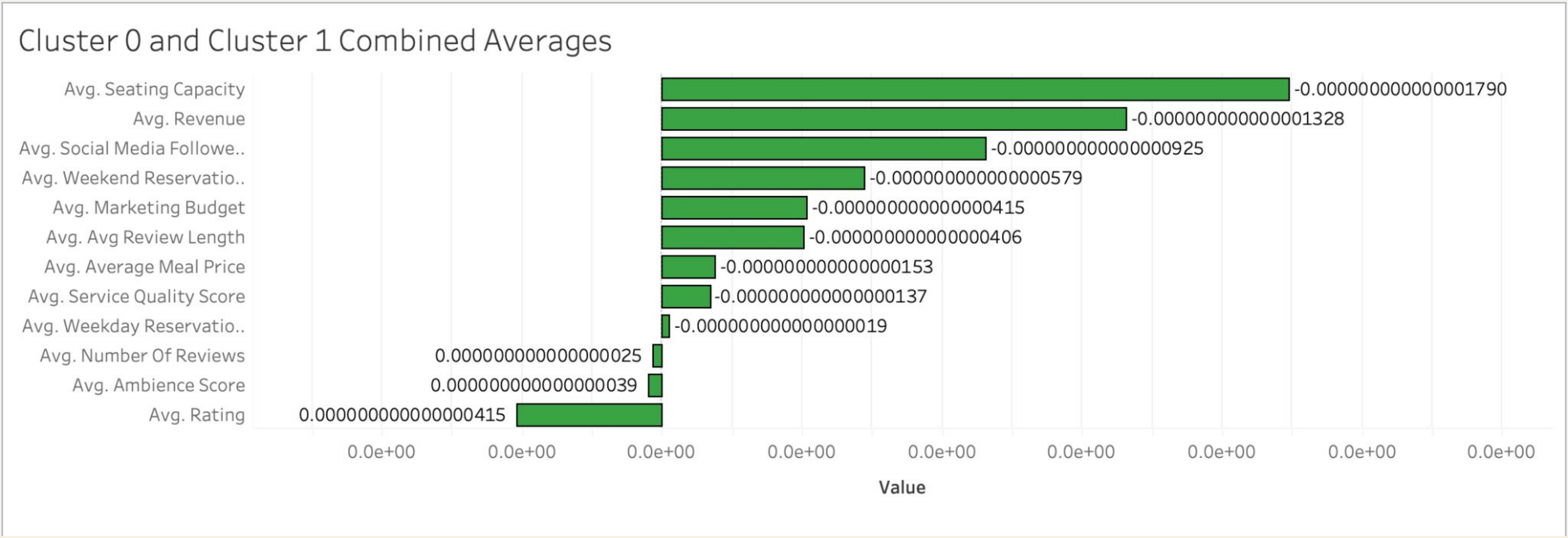
Determining the Number of Clusters & K-means Clustering



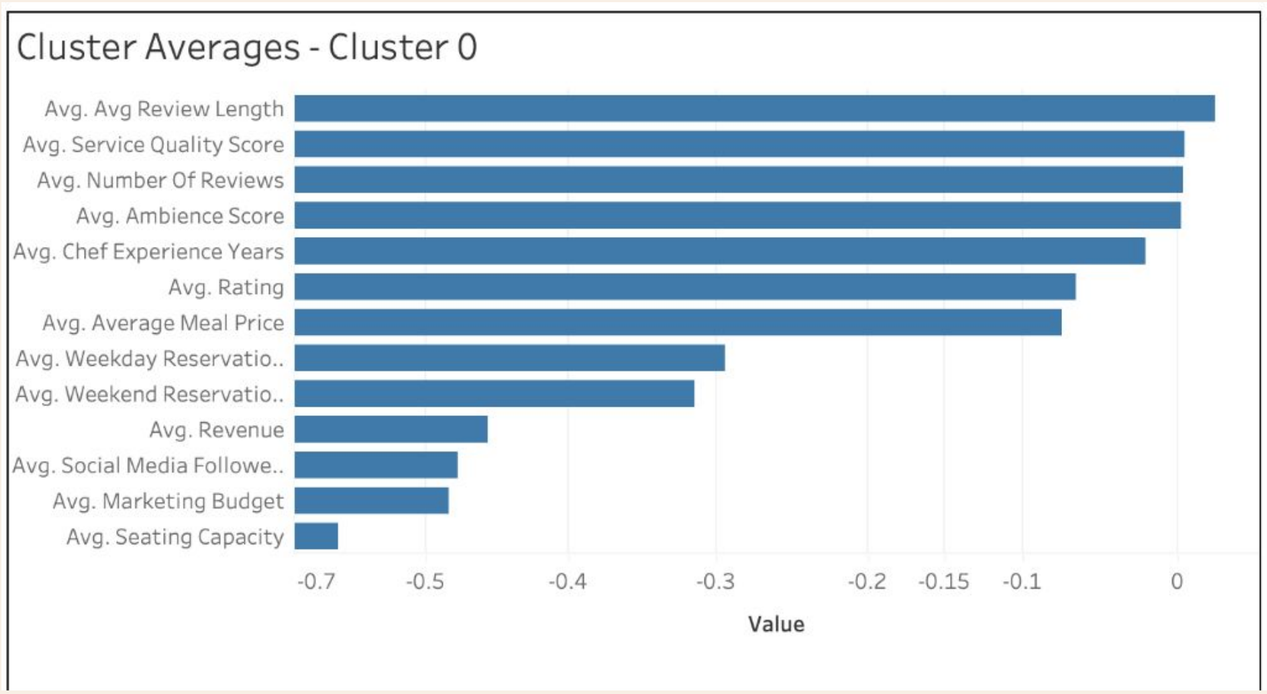


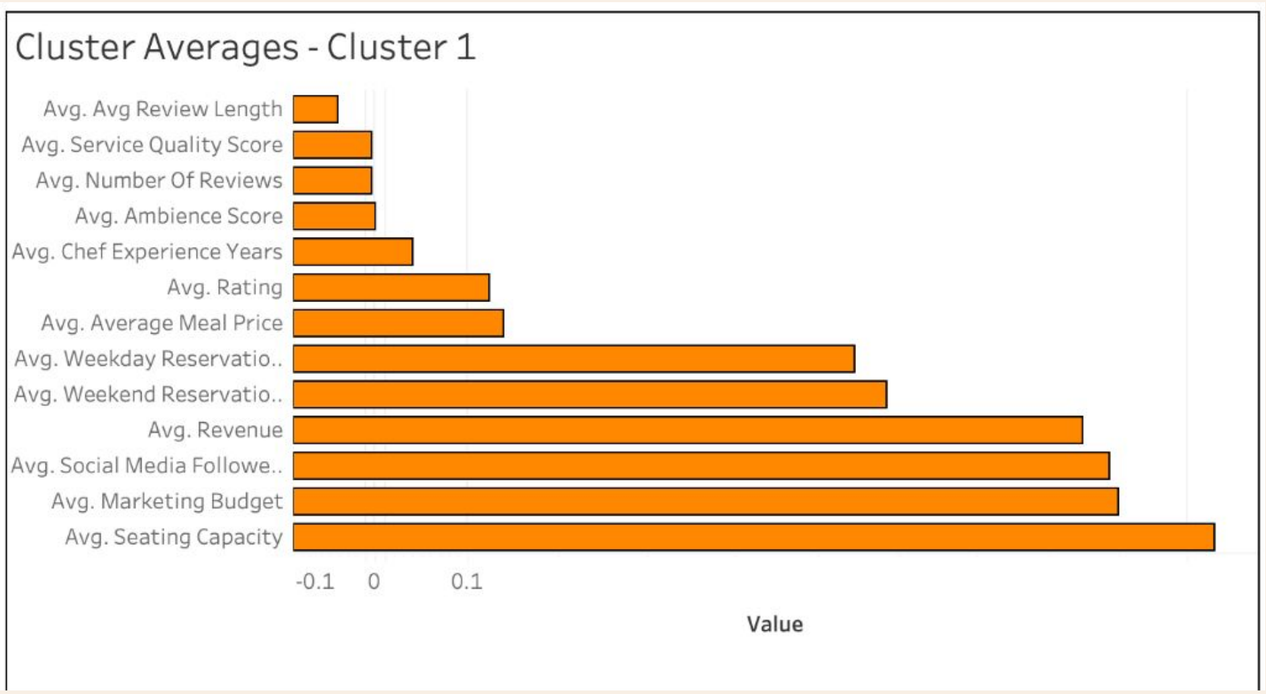
	PCA_1	PCA_2	PCA_3	PCA_4	PCA_5	PCA_6	PCA_7	PCA_8	PCA_9	Cluster
0	-1.602329	-1.717388	1.636562	-1.393867	-1.705817	-0.210651	0.097664	0.326914	0.408310	0
1	0.441447	1.358972	-1.409816	-0.710771	-0.172635	0.511681	0.809583	-0.487749	1.704060	1
2	-0.779930	0.468890	1.343955	-0.353624	0.538145	0.535154	0.998055	2.139401	0.279931	0
3	-2.535516	-0.174292	1.015148	-1.686116	-0.077267	-0.640989	-0.547675	-0.862953	-0.334777	0
4	2.860161	-2.526592	1.211125	-0.760200	0.943812	-0.518266	-1.510619	-1.638053	0.214561	1
...	...	...	...	...	...	...	...	...	...	...
8363	-2.168573	0.060189	-1.092439	0.477991	1.108680	-1.067013	-1.265438	-0.399140	1.338251	0
8364	-1.677503	0.512871	-0.794318	-0.501850	0.174091	0.291397	1.426922	-0.668104	0.547127	0
8365	3.380806	0.677995	0.008023	-0.402776	0.372431	0.014191	0.324511	-1.757077	2.050849	1
8366	-3.407766	-0.334320	-0.686713	0.519816	0.756618	1.255804	0.432478	-1.461543	-0.400274	0
8367	-1.962954	-1.389461	1.659146	-0.247970	-0.900507	0.429666	-1.181607	0.155208	-0.164609	0

Cluster Averages - Combined



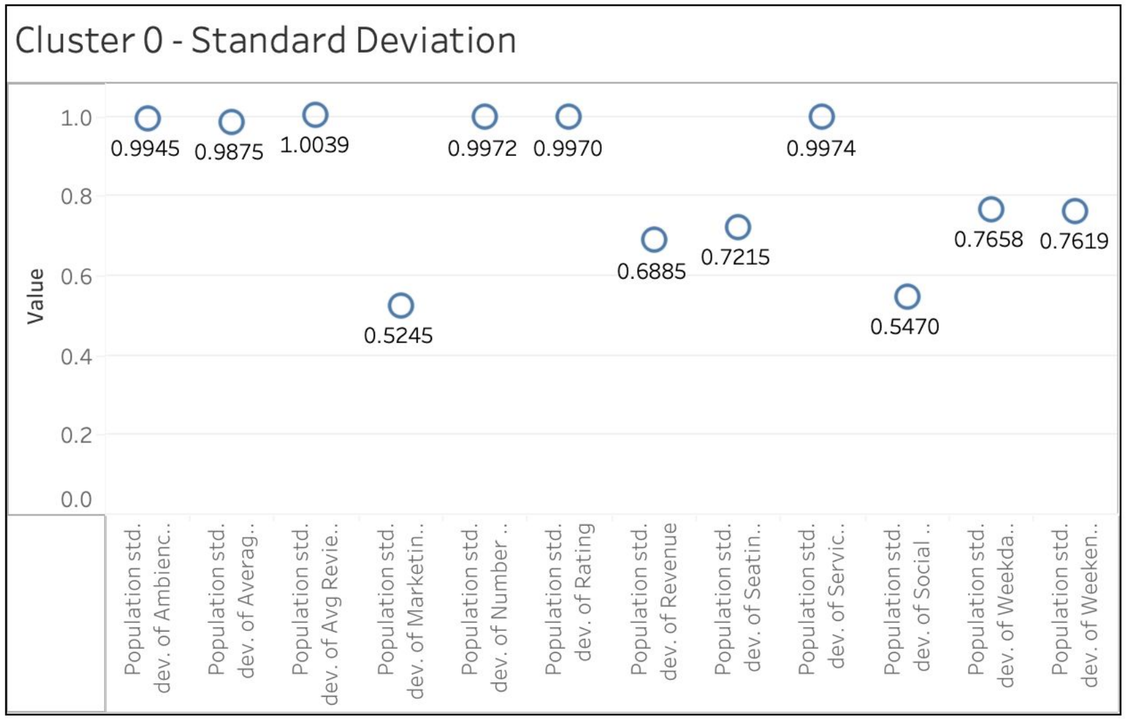
Cluster Averages - Separated

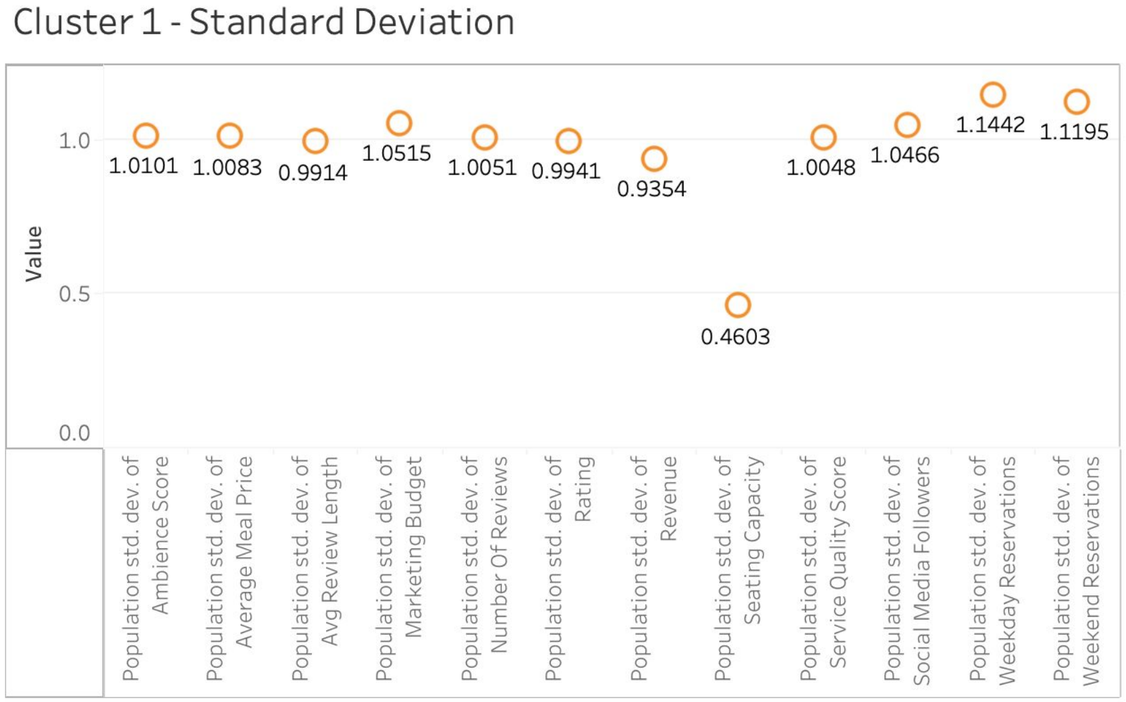




- With the standardized data, the average values for each of the reported items are complementary to each other and help create a visual for how the model selected between the two clusters

Standard Deviations

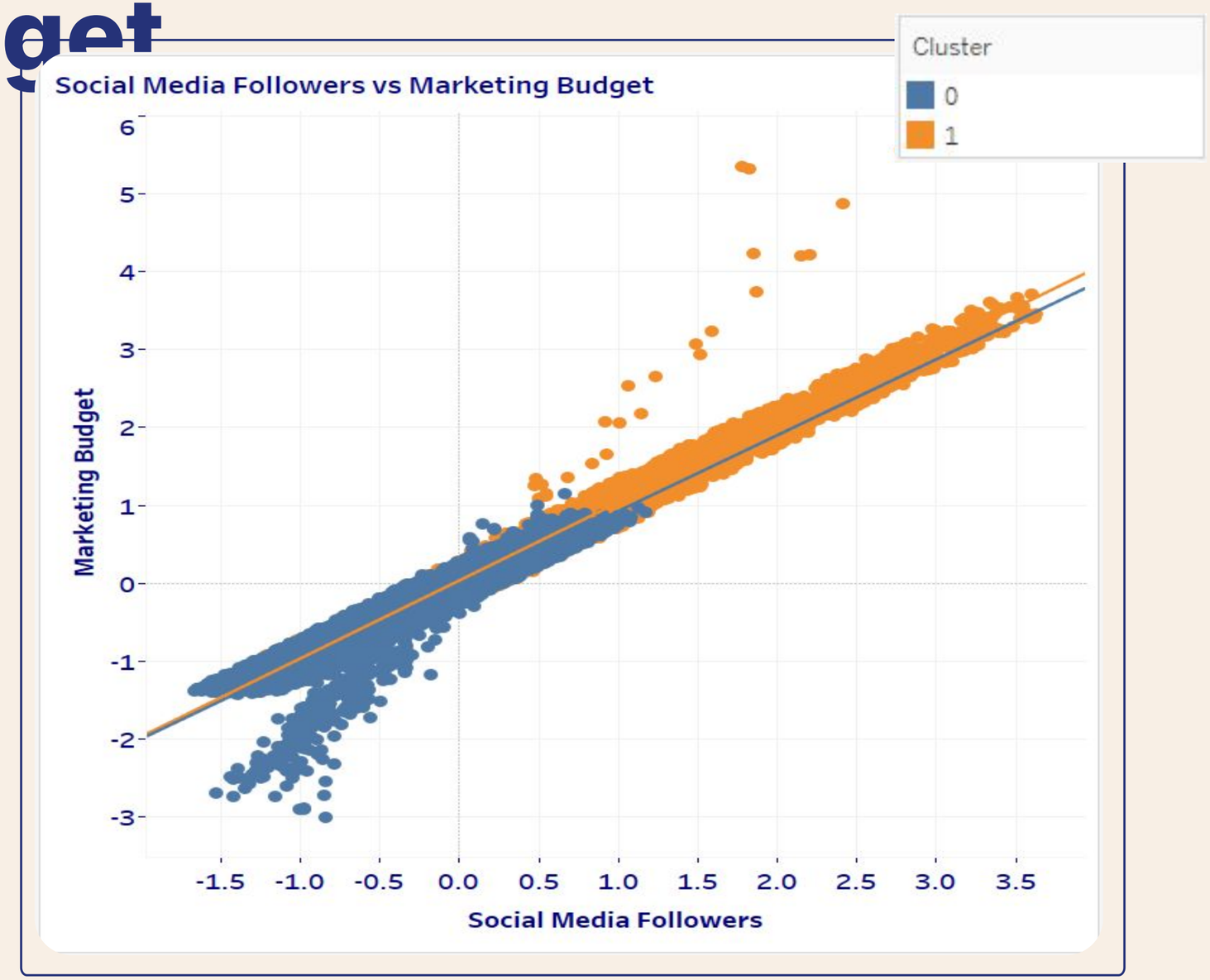




- Between the two clusters we see more variability in the SD of the measures for Cluster 0 versus Cluster 1
- Cluster 1 overall has a more consistent, but larger, standard deviation for each measure
- In large, it appears that Cluster 1 has more “normal” distribution of the data and may help to indicate that these are the restaurants that should be included in the “Go” group.

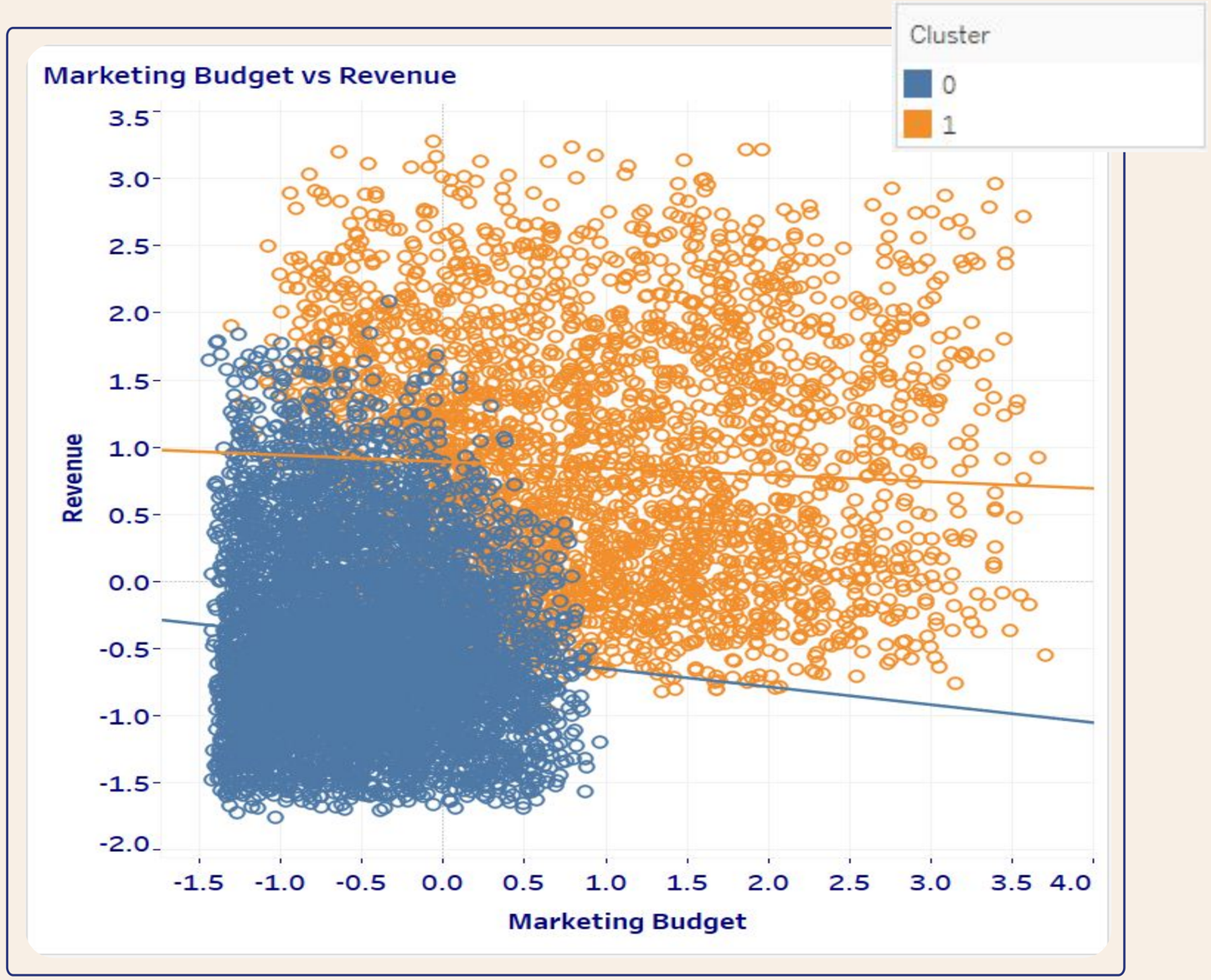
Social Media Followers vs Marketing Budget

- Both clusters show a **positive relationship** between Social Media Followers and Marketing Budget.
- There are indicators that businesses investing **more in marketing** efforts, tend to gain **more followers** on social media.
- While both clusters show similar results, ***Cluster 1*** may reflect **more substantial investments in marketing,** contributing to an even greater increase in social media followers compared to ***Cluster 0***.



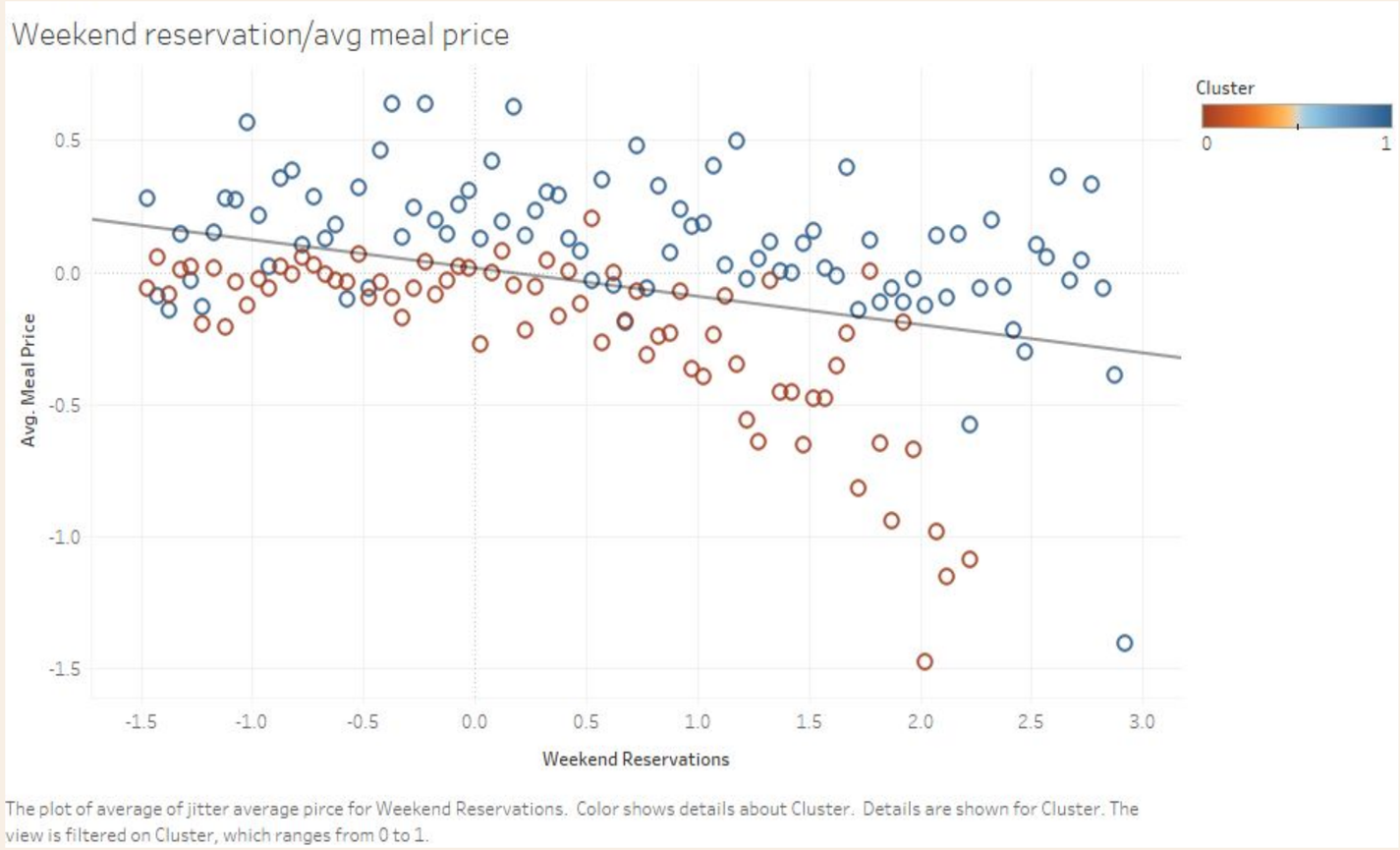
Marketing Budget vs Revenue

- Both clusters show a **negative relationship** between Marketing Budget and Revenue.
- As Marketing Budget increases, Revenue tends to decrease for both clusters, but the **rate of decrease varies between them**.
- Businesses in ***Cluster 0*** may need to reassess their marketing strategies, as their higher marketing budget is **not yielding revenue gains to match**.
- ***Cluster 1*** shows a more **moderate decline** in Revenue with increasing Marketing Budget, suggesting that **marketing efforts might be more effective** or have diminishing returns at a **slower rate**.



Weekend Reservations vs Avg.Meal Price

- Clusters shows **negative** relationships between **Weekend Reservations** and **avg.Meal Price**
- Avg Meal Price **decrease** for both clusters, Weekend RSVP i**ncrease**
- **Cluster 0** experiences a sharper decline in **Average Meal Price** as **Weekend Reservations** increase, indicating that businesses in this cluster may need to reassess their pricing strategies, as more reservations are correlating with lower prices rather than higher ones.
- **Cluster 1** gradual decrease in **Average Meal Price** as **Weekend Reservations** rise, suggesting these businesses have a more stable pricing strategy that maintains meal prices despite more reservations.



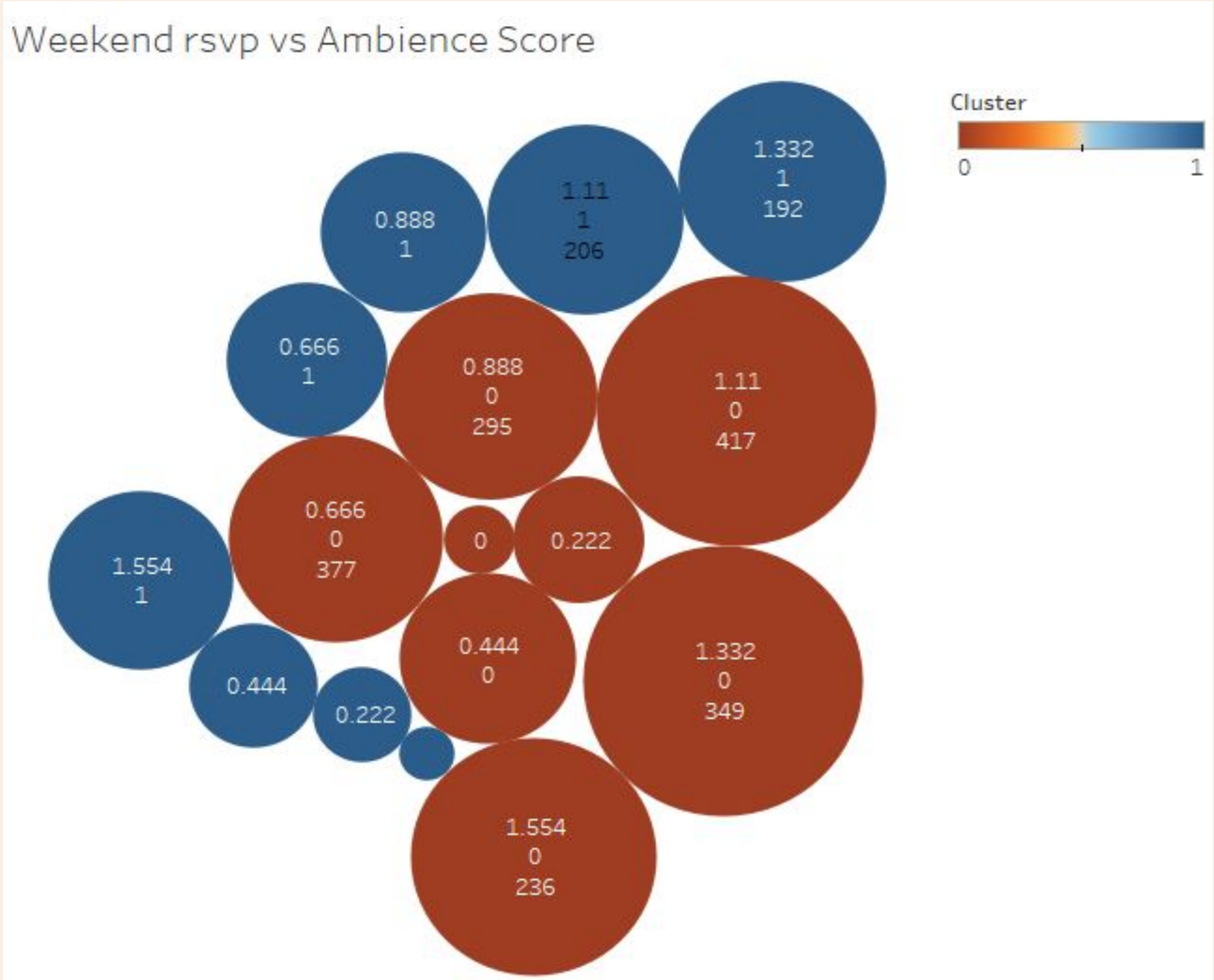
Weekend Reservations vs Ambience Score

- The chart shows the relationship between **Weekend Reservations** and **Ambience Score** for two clusters
- Larger bubbles indicated more Weekend RSVP

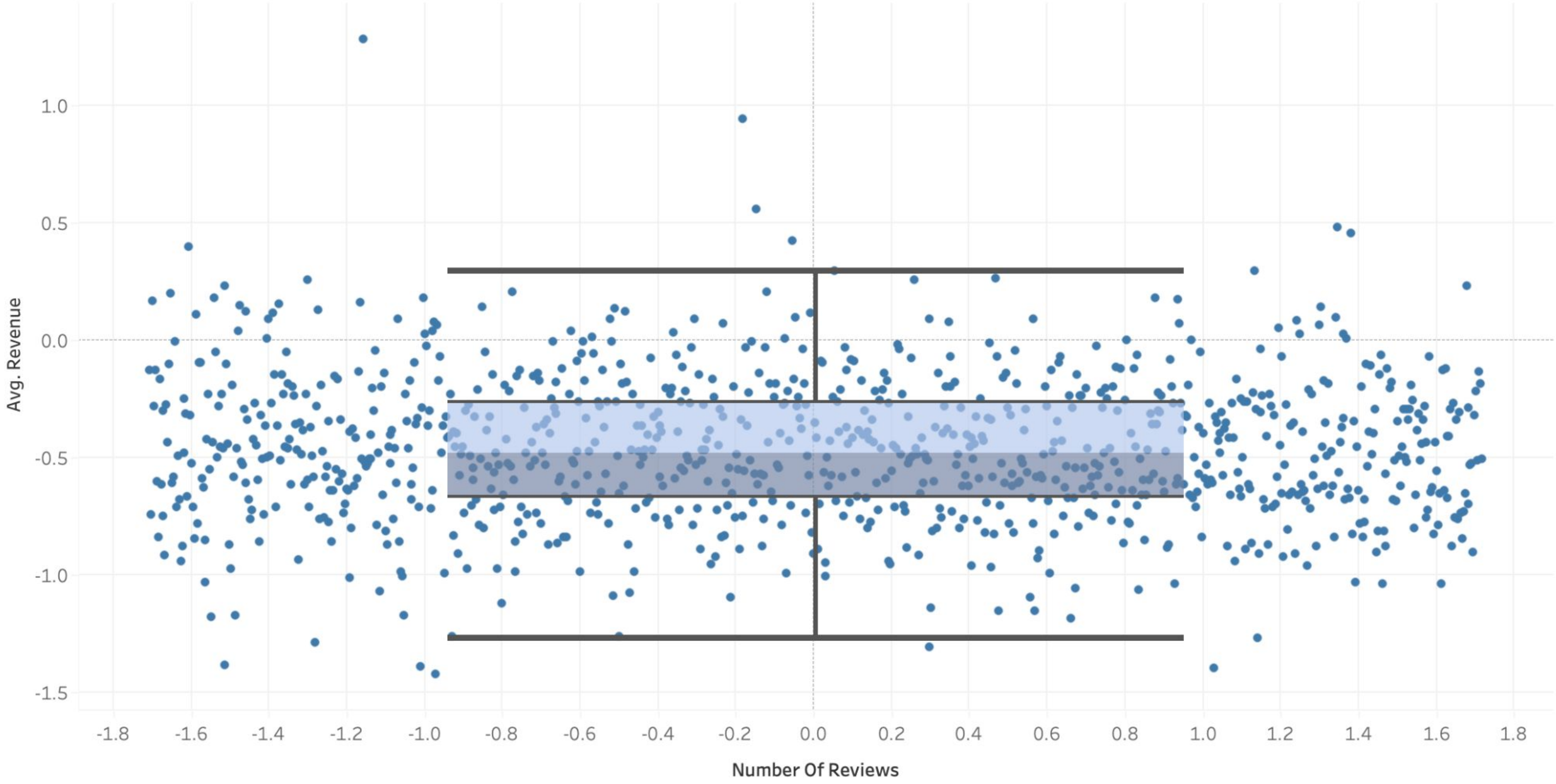
- Cluster 0 (Red) tends to have **higher Ambience Scores** and **more Weekend Reservations**

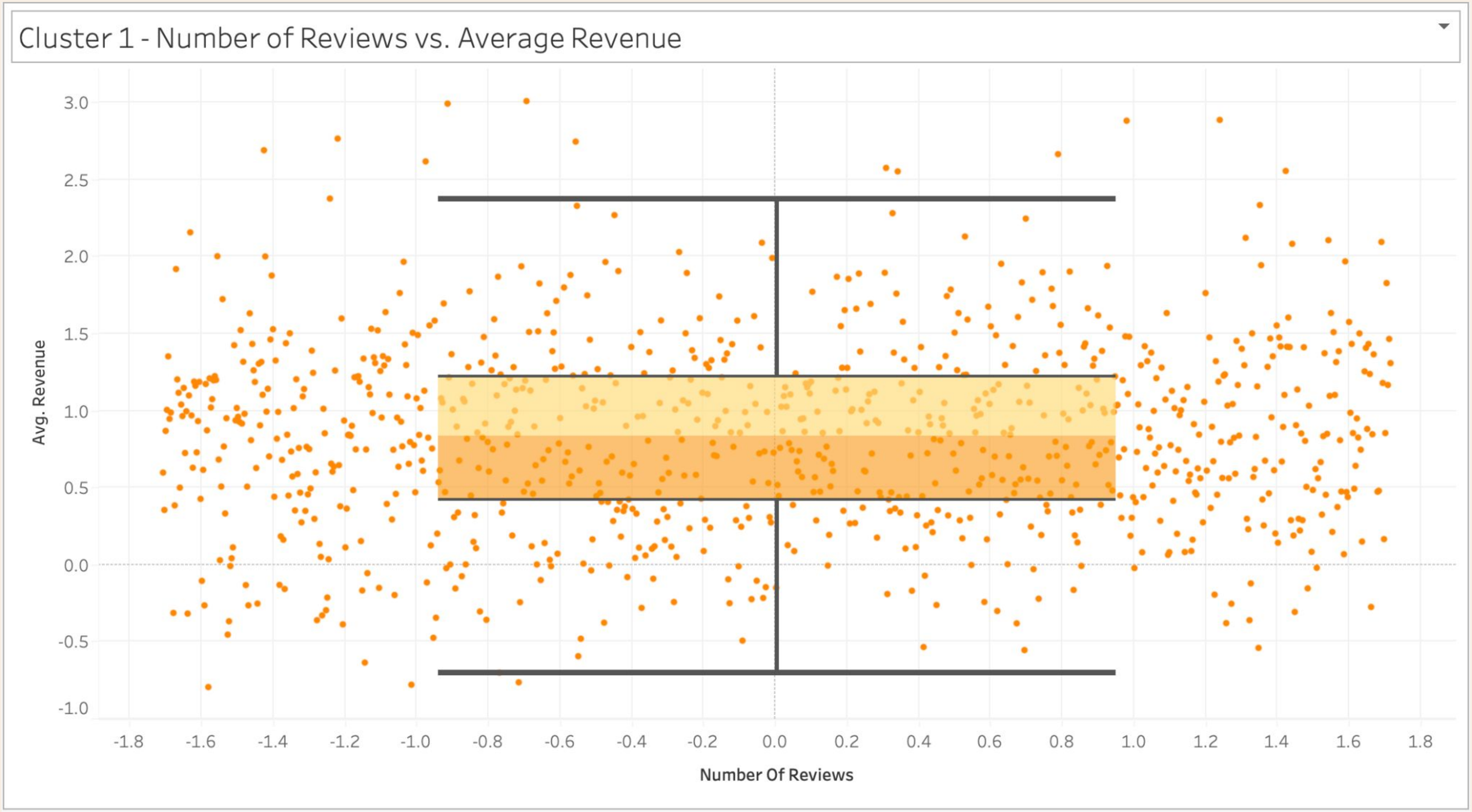
- Cluster 1 (blue) has **lower Ambience scores** and **fewer Weekend Reservations**.

- The bubble chart suggests that business with **higher Ambience scores** tend to have **more Weekend reservations.** However, the relationship is **not necessarily linear,** and there are variations within each cluster.



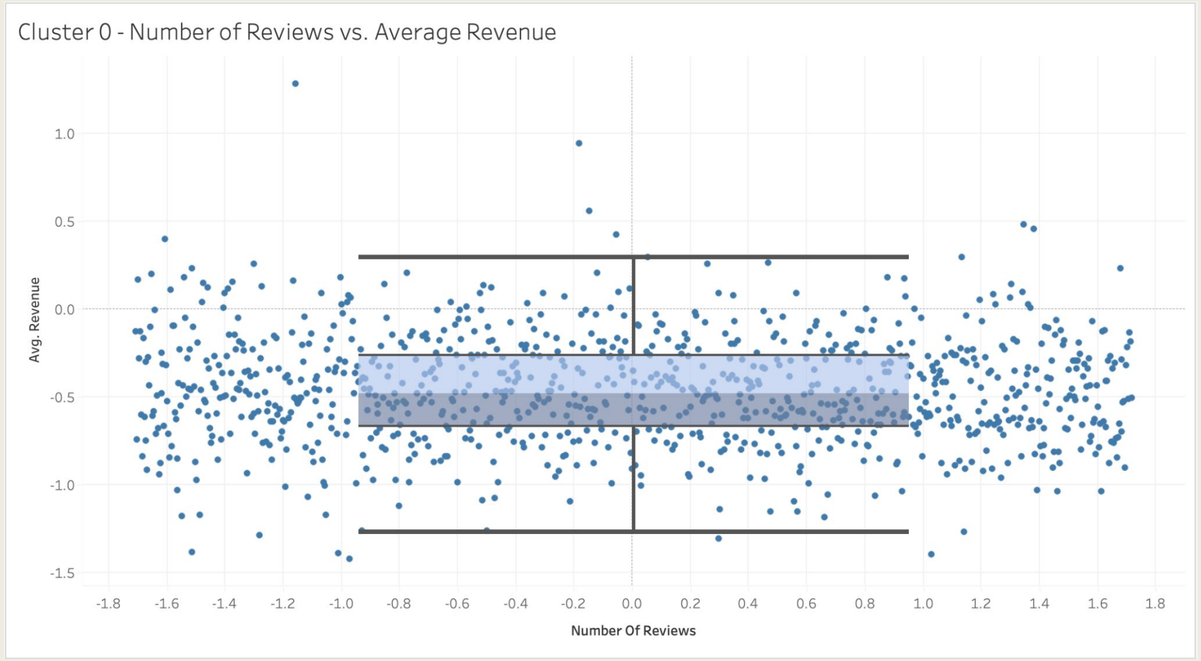
Cluster 0 - Number of Reviews vs. Average Revenue

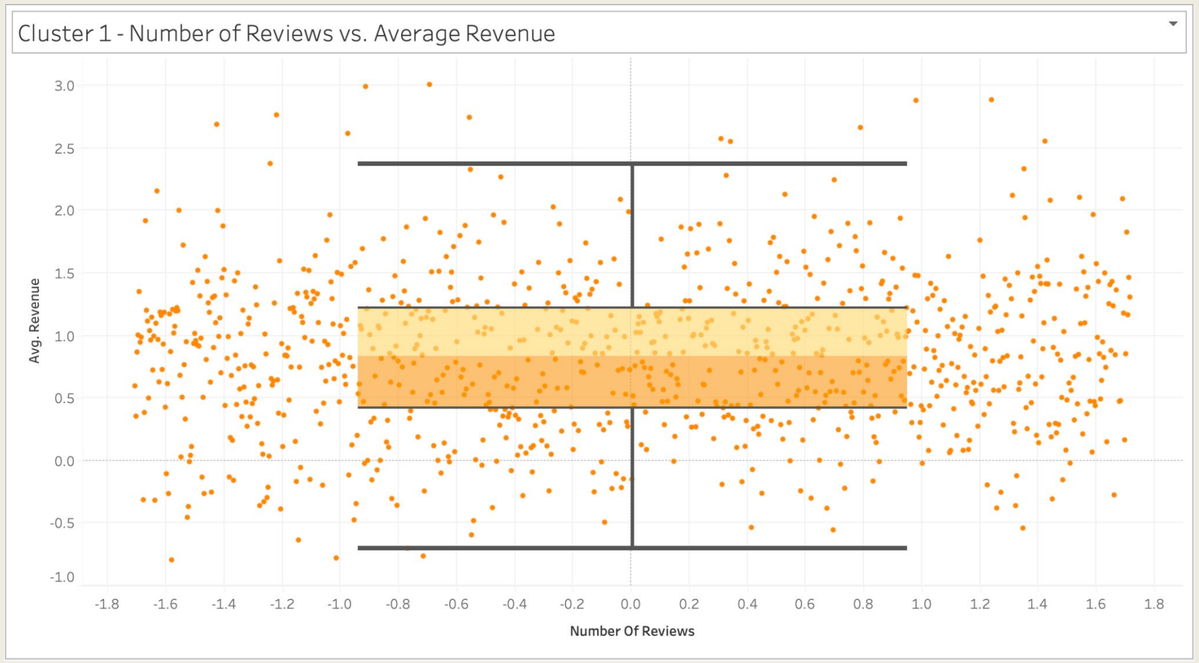



Cluster 1 - Number of Reviews vs. Average Revenue
Avg. Revenue
3.0
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.8
-1.6
-1.4
-1.2
-1.0
-0.8
-0.6
-0.4
-0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
Number Of Reviews

Number of Reviews vs. Average Revenue

- Both clusters have relatively even distribution between number of reviews and the correlated average revenue.
- The major difference between the clusters is the total revenue as a whole with Cluster 0 having a lower median value.
- Cluster 1 shows a larger standard deviation and more outliers than Cluster 0 at a higher median value





Questions?

DON'T FORGET TO DELETE THIS PAGE BEFORE PRESENTING.

Resources

Dataset:
https://www.kaggle.com/datasets/anthonytherrien/restaurant-revenue-prediction-dataset/data

DON'T FORGET TO DELETE THIS PAGE BEFORE PRESENTING.