Vincent Angelo Flores

Jade Espinar

Assignment #2

Dataset Theme: **Factors Influencing Dining Choice**

Description: Our dataset aims to explore the factors that influence individuals' choices when selecting dining options, whether eating out, ordering takeout, or eating at home.

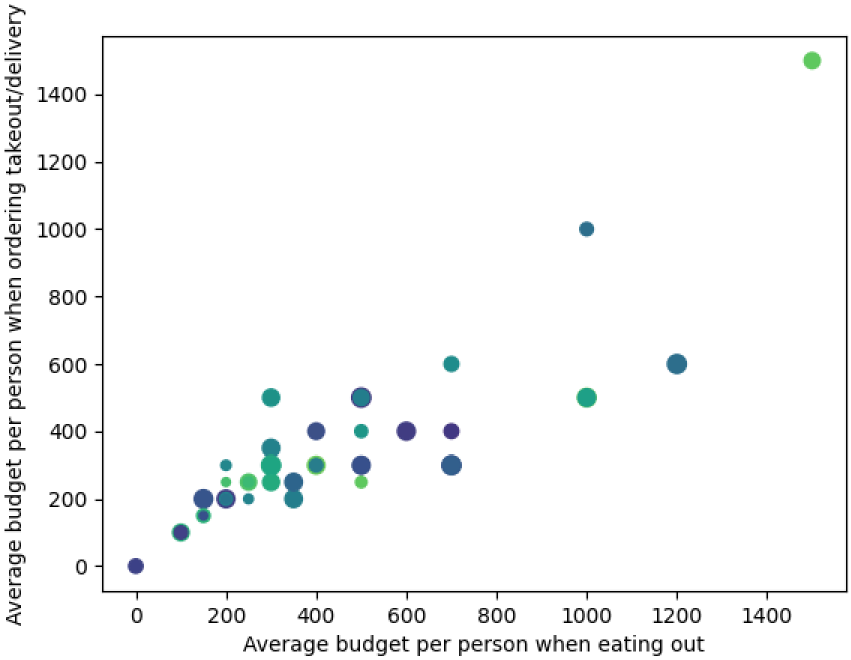
**Data Preprocessing & Preliminary Data Mining**

The table below shows all the correlation coefficients for all possible pairs of numeric attributes. The pair with the highest correlation is *Average budget per person when eating out* and *Average budget per person when ordering takeout/delivery,* while the pair with the weakest correlation or no correlation is *Average budget per person when ordering takeout/delivery* and *Approximate cost of groceries per month for home cooking*. Their corresponding graphs are shown below the table.

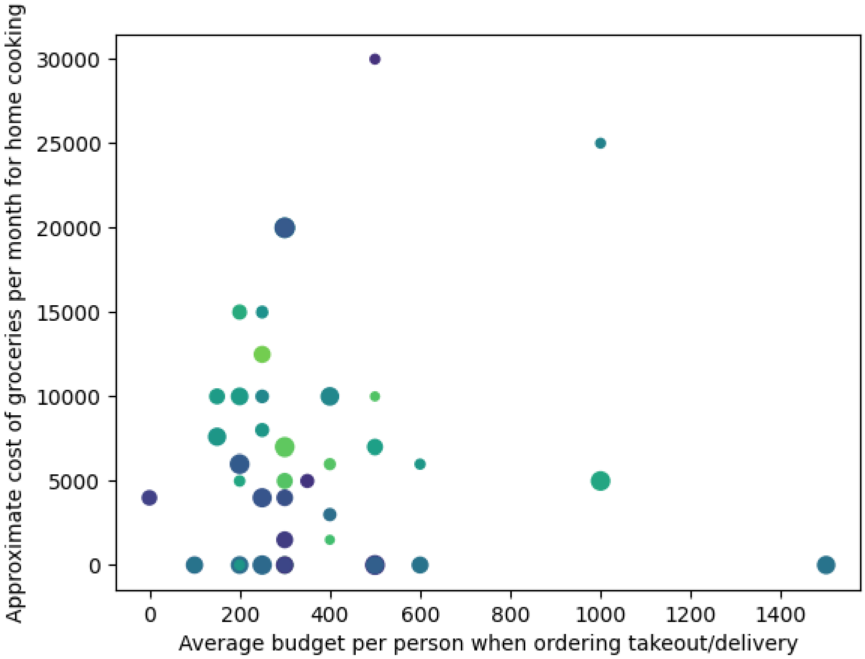
| **Attribute 1** | **Attribute 2** | **Correlation Coefficient** | **Rank** |
|---|---|---|---|
| Average budget per person when eating out | Average budget per person when ordering takeout/delivery | 0.84 | 1 |
| Household Size | Approximate cost of groceries per month for home cooking | 0.31 | 2 |
| Age | Approximate cost of groceries per month for home cooking | 0.23 | 3 |
| Age | Average budget per person when eating out | -0.19 | 4 |

| | | | |
|---|---|---|---|
| Age | Average budget per person when ordering takeout/delivery | -0.17 | 5 |
| Household Size | Average budget per person when ordering takeout/delivery | -0.15 | 6 |
| Age | Household Size | 0.13 | 7 |
| Household Size | Average budget per person when eating out | -0.13 | 8 |
| Average budget per person when eating out | Approximate cost of groceries per month for home cooking | 0.12 | 9 |
| Average budget per person when ordering takeout/delivery | Approximate cost of groceries per month for home cooking | 0.06 | 10 |

### Strongest correlation



**Weakest correlation/most independent attributes**



The figures above show the pair of attributes with the strongest and weakest correlation, respectively. The strongest correlation, *Average budget per person when eating out* and *Average budget per person when ordering takeout/delivery,* has a coefficient of 0.84, while the weakest correlation, *Average budget per person when ordering takeout/delivery* and *Approximate cost of groceries per month for home cooking,* has a coefficient of 0.06. This can be interpreted as people who tend to spend more when eating out generally spend the same amount when ordering for takeout/delivery, since these two are almost the same except for when the food is eaten.

For the categorical attribute correlations, the Chi-Square test was applied. The pair with the highest correlation and that has rejected the null hypothesis is the *Relationship Status* and *Housing Type* attributes. No other pair of categorical attributes has rejected the null hypothesis. Their Chi-Square information is detailed below:

| | |
|---|---|
| Attributes | Relationship Status, Housing Type |
| Relationship Status values | "Married",<br>Married but not living together”,<br>"Single" |
| Housing Type values | "Dormitory",<br>"Living with Relatives (not with parents/siblings)",<br>"Living with parents/siblings",<br>"Own House",<br>"Renting" |
| Degrees of Freedom | 8 |
| Critical Value for DF 8 | 15.507 |
| Chi-Square Value for the pair | 23.267 |

| | | **Housing Type** | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Relationship Status** | | **Dormitory** | | **Living with Relatives (not with parents/siblings)** | | **Living with parents/siblings** | | **Own House** | | **Renting** | | **Total** |
| | **Married** | 0 | 0.78 | 0 | 1.04 | 1 | 4.68 | 9 | 3.38 | 3 | 3.12 | 13 |
| | **Married but not living together** | 0 | 0.06 | 0 | 0.08 | 0 | 0.36 | 1 | 0.26 | 0 | 0.24 | 1 |
| | **Single** | 3 | 2.16 | 4 | 2.88 | 17 | 12.96 | 3 | 9.36 | 9 | 8.64 | 36 |
| | **Total** | 3 | | 4 | | 18 | | 13 | | 12 | | 50 |

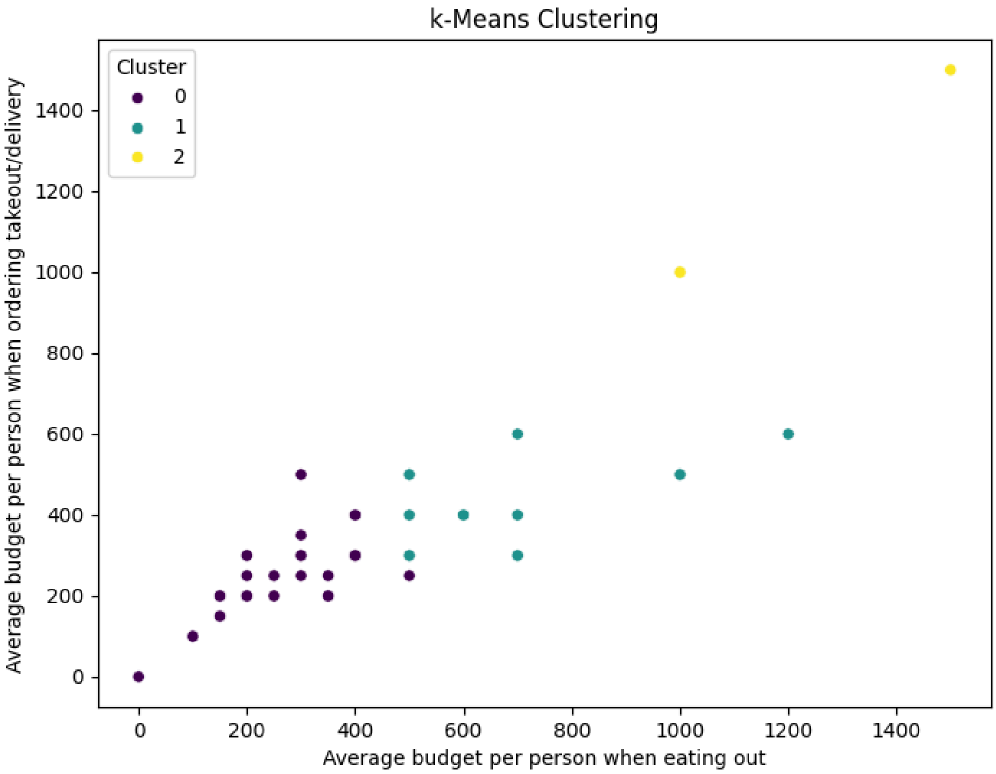
The two tables above show the details of the Chi-Square test and the observed-expected table, respectively, for the attributes, *Relationship Status* and *Housing Type*. Values in red are the expected values. The Chi-Square test is done by comparing the Chi-Square score of a given correlation with a corresponding critical value that is dependent on the correlation's degrees of freedom. The Chi-Square Score is heavily dependent on the difference between the observed and expected values. Closer inspection of the 2nd table shows that almost all of the values are near their expected values, except for 4 pairs of values, which are the Single-Own House and Married-Own House, Single-Living with parents/siblings, and Married-Living with parents/siblings value pairs. These 4 value pairs have the most difference with their corresponding expected values from 3.62 as the lowest difference to 6.36 as the highest difference. This discrepancy can come from the distribution of Single vs Married respondents, where 36/50 or around 72% are Single respondents, while 13/50 or 26% are married respondents. This distribution skewed the expected values of the Own House value, wherein this is the only instance where the Married count (9) is greater than the Single count (3). However, this is to be expected since it is common for Married couples to have their Own House as opposed to those who are Single. This is also apparent by inspecting the pair, Single-Living with parents/siblings, which has the highest observed value. This is common where Single people still live with parents/siblings. If additional respondents are obtained, we can expect that most Single respondents would answer "Living with parents/siblings as their" Housing type and most Married respondents would answer "Own House."

## Clustering

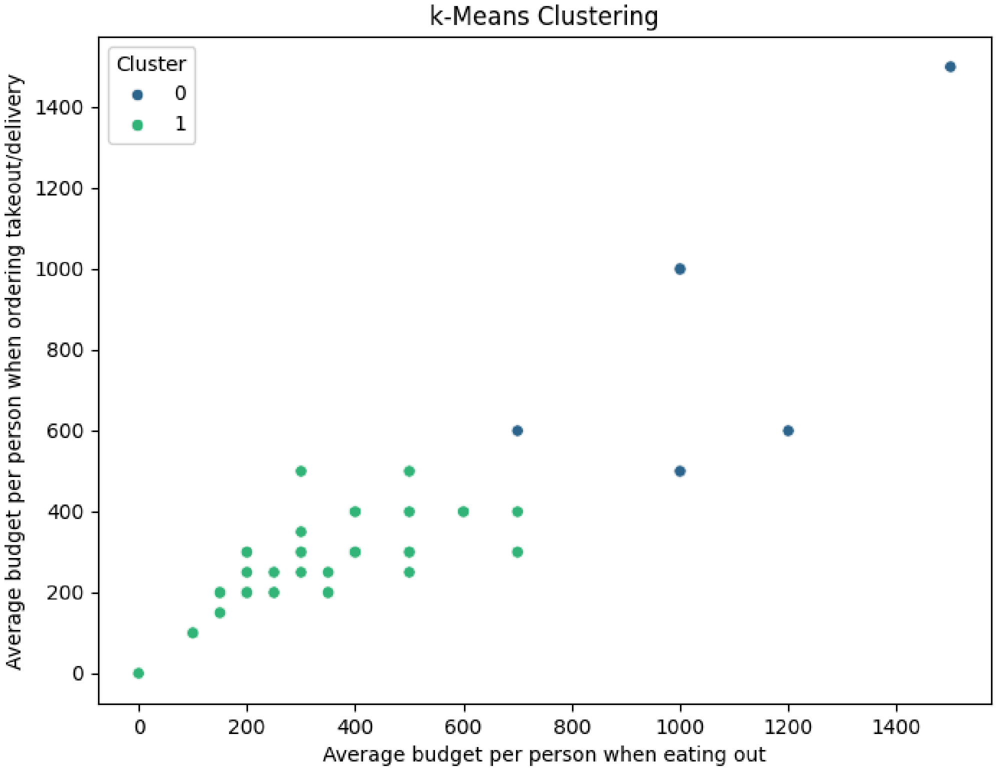
The two numeric attributes with the strongest correlation are *Average Budget Per Person When Ordering Takeout/Delivery* and *Average Budget Per Person When Eating Out*.

Using **k=3** and the **display_clusters** function in **Flores_Espinar_A1_Code.py**, the cluster visualization below was generated. The clusters have high inertia and a low silhouette score. Here are our impression on the characteristics of each cluster:

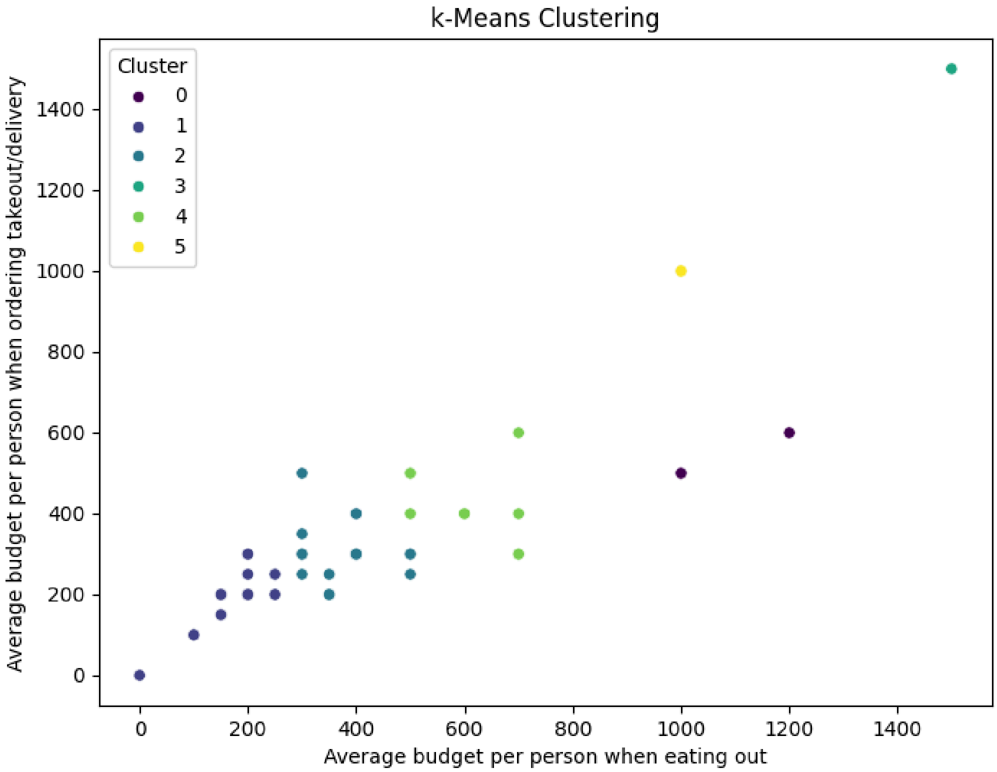
- **Cluster 0**: People who tend to spend little on both eating out and takeout/delivery
- **Cluster 1**: People who tend to spend more on eating out than on takeout/delivery
- **Cluster 2**: People who tend to spend a lot on both eating out and takeout/delivery



Using **k=2** and the **display_clusters** function in **Flores_Espinar_A1_Code.py**, the cluster visualization below was generated. The clusters have high inertia and a low silhouette score.



Using **k=6** and the **display_clusters** function in **Flores_Espinar_A1_Code.py**, the cluster visualization below was generated. The clusters have high inertia and a low silhouette score.



Even with different values of **k**, the inertia remains high, and the silhouette score remains low. The visualization appears almost as a straight line with a positive constant slope, suggesting a direct variation relationship between the two variables.

This means that as the budget for eating out increases, the budget for takeout and delivery also increases. **The results suggest that people seem to treat both experiences similarly, with the only difference being the setting.**

As a result, no matter how many clusters are formed, the data points continue to overlap and are not tightly grouped.