Starbucks Customer Survey

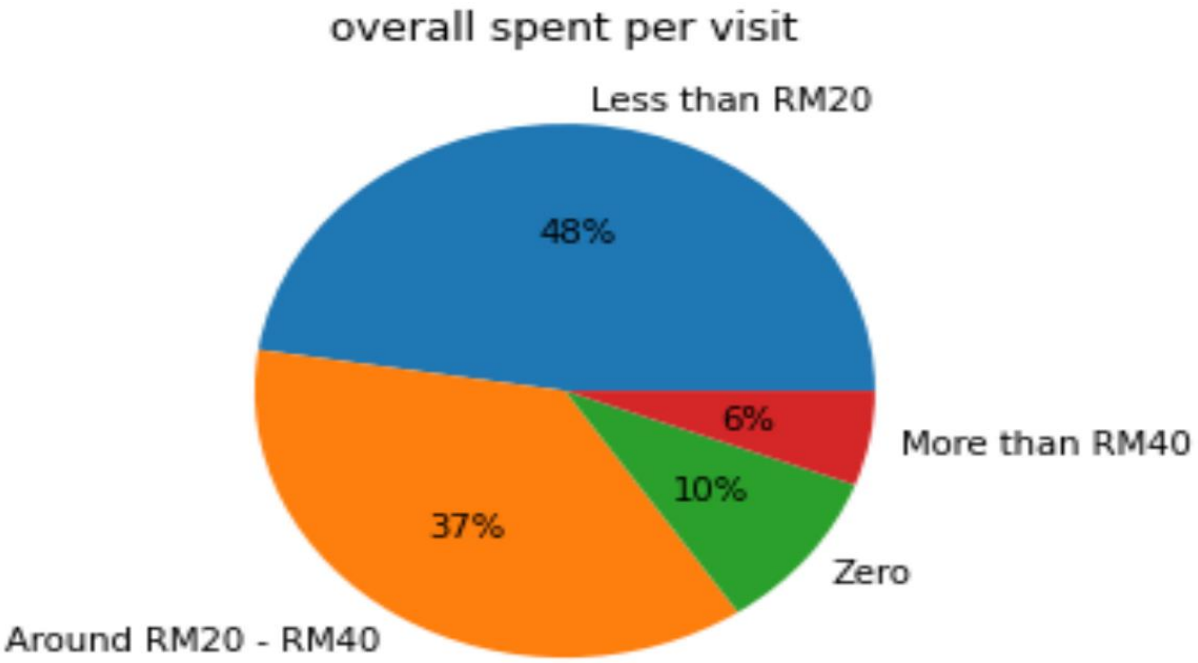
This dataset is composed of a survey question of over 100 respondents for their buying behaviour at Starbucks.

The survey dataset initially contained category columns as object types and all the columns had long names which makes it difficult for a data analyst while making calculations. I have cleaned the dataset and refactored the column names accordingly.

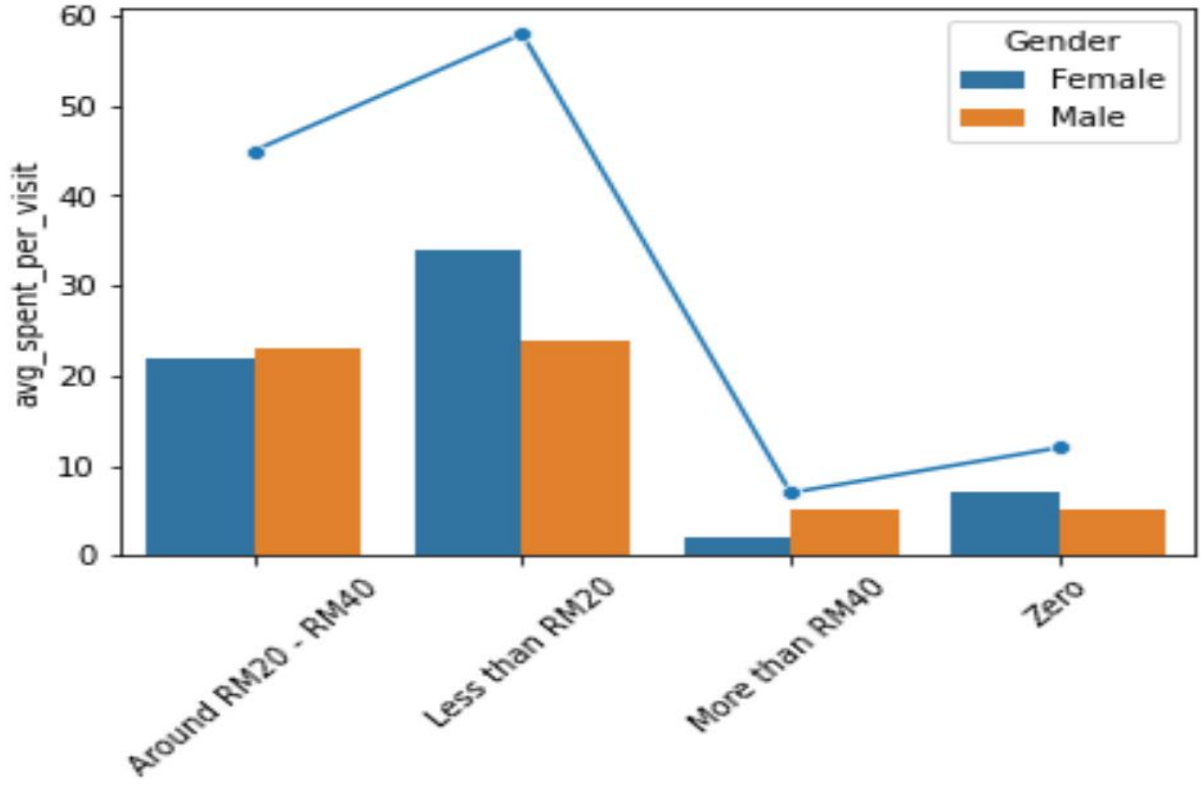
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 122 entries, 0 to 121
Data columns (total 21 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Timestamp             122 non-null    object
 1   Gender                122 non-null    category
 2   Age                   122 non-null    category
 3   Employment            122 non-null    category
 4   Income                122 non-null    category
 5   visit_frequency       122 non-null    category
 6   visit_type            121 non-null    category
 7   visit_time_spent      122 non-null    category
 8   nearest_starbucks     122 non-null    category
 9   memcard_available     122 non-null    category
 10  frequent_purchase     122 non-null    category
 11  avg_spent_per_visit   122 non-null    category
 12  brand_rating          122 non-null    int64
 13  price_rating          122 non-null    int64
 14  promotion_rating      122 non-null    int64
 15  ambience_rating       122 non-null    int64
 16  wifi_rating           122 non-null    int64
 17  service_rating        122 non-null    int64
 18  choosing_stb_rating   122 non-null    int64
 19  promotion_heard_from  121 non-null    object
 20  willing_to_visit_stb  122 non-null    object
dtypes: category(11), int64(7), object(3)
memory usage: 13.6+ KB
```

Any business to be sustainable or remain afloat needs to generate subsequent sales. So, in this dataset the column which describes the sales detail is **avg_spent_per_visit**. We will be ignoring the column **frequent_purchase** as there doesn't seem to be much variability in the product names.

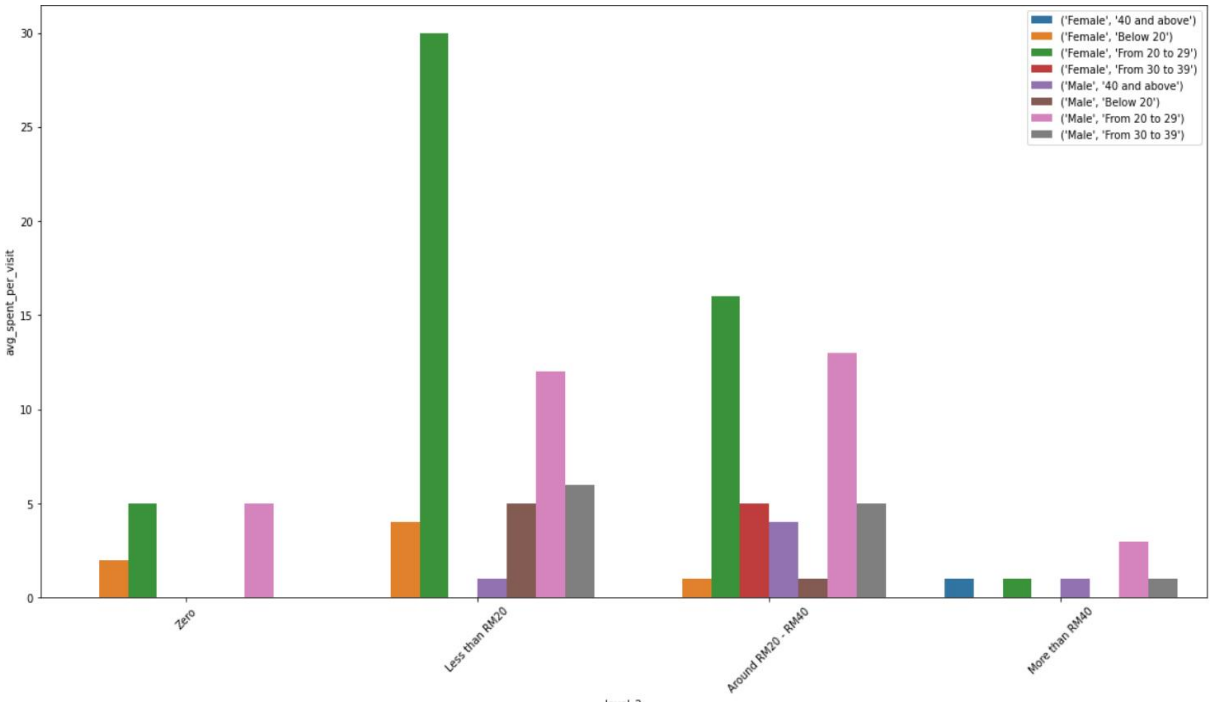
When we start analyzing the **avg_spent_per_visit**, we see that there seems to be a large sales volume in the categories **Less than RM20(5.81 CAD)** and **Around RM20 – RM40 (11.63CAD).**



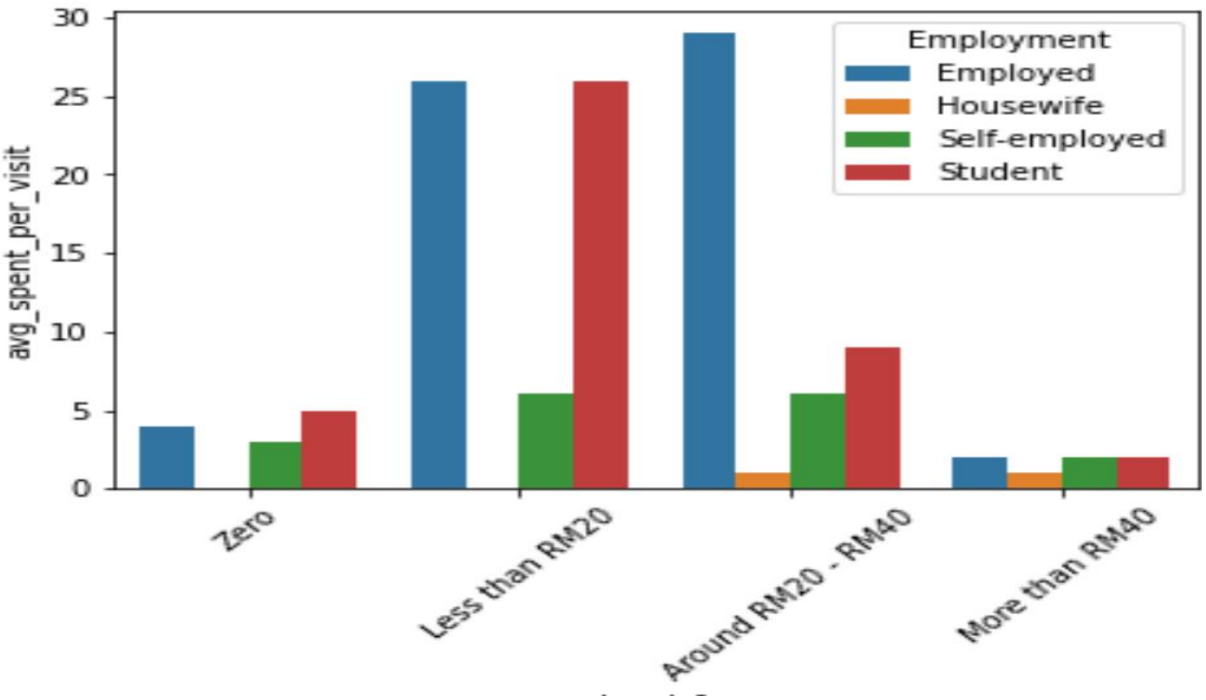
When drilling down deeper among Males and Females, we can observe that females spent more in the category Less than RM20.

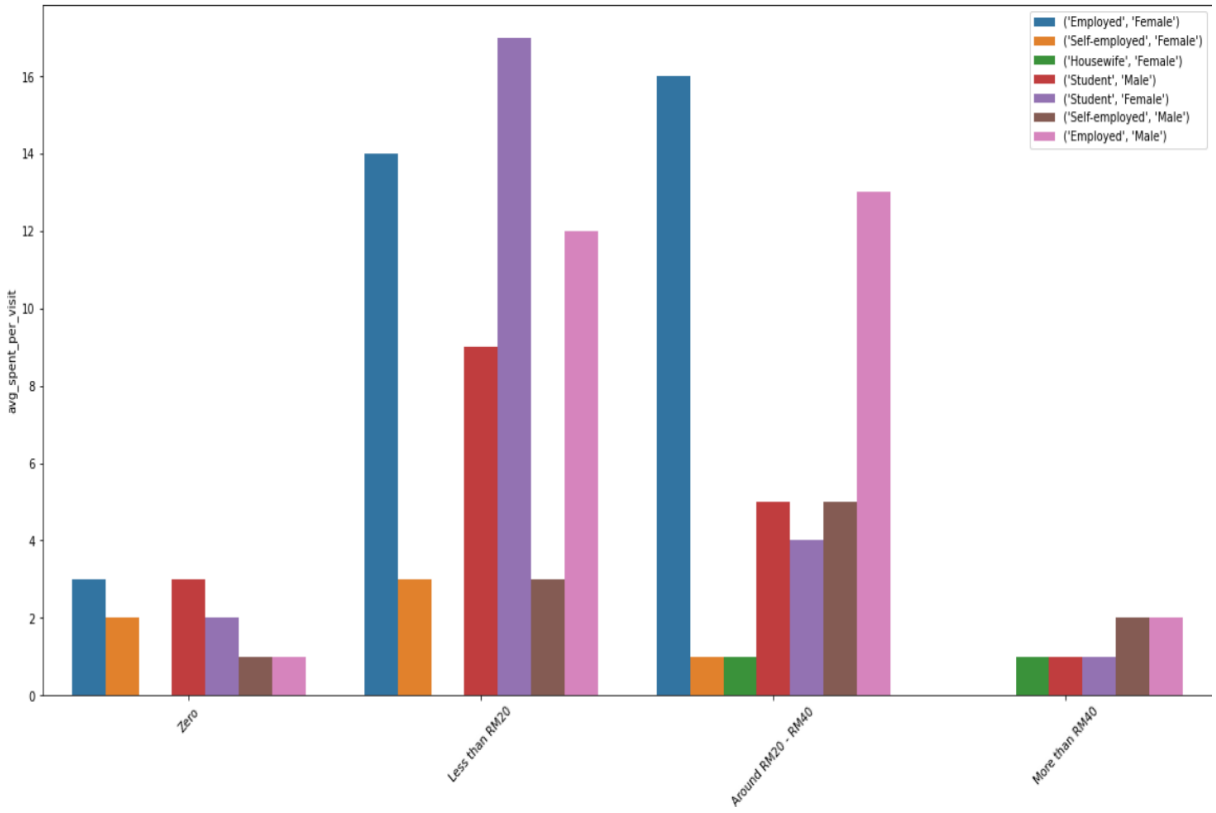


The same follows for different age categories among Males and Females.

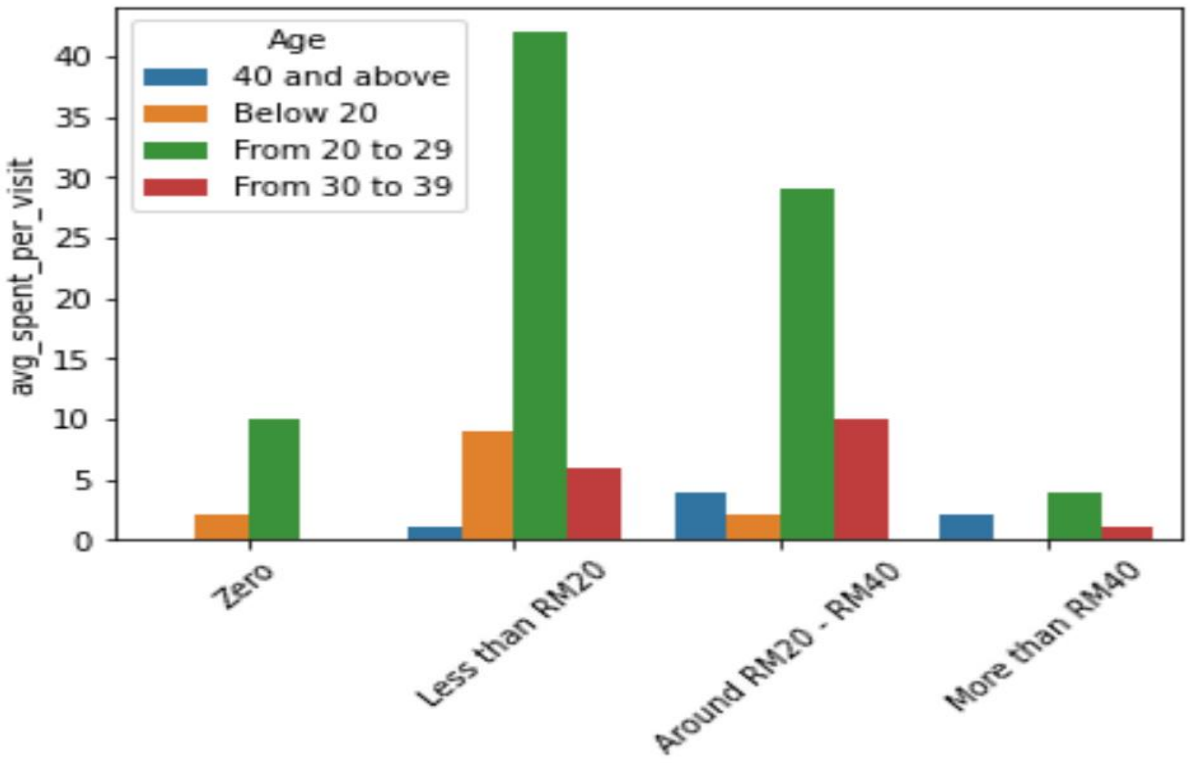


While checking if Employment type had been a confounder for less sales in the category of **More than RM40**, we can outright ignore this assumption since many employed customers too are not willing to spend in the category **More than RM40.**

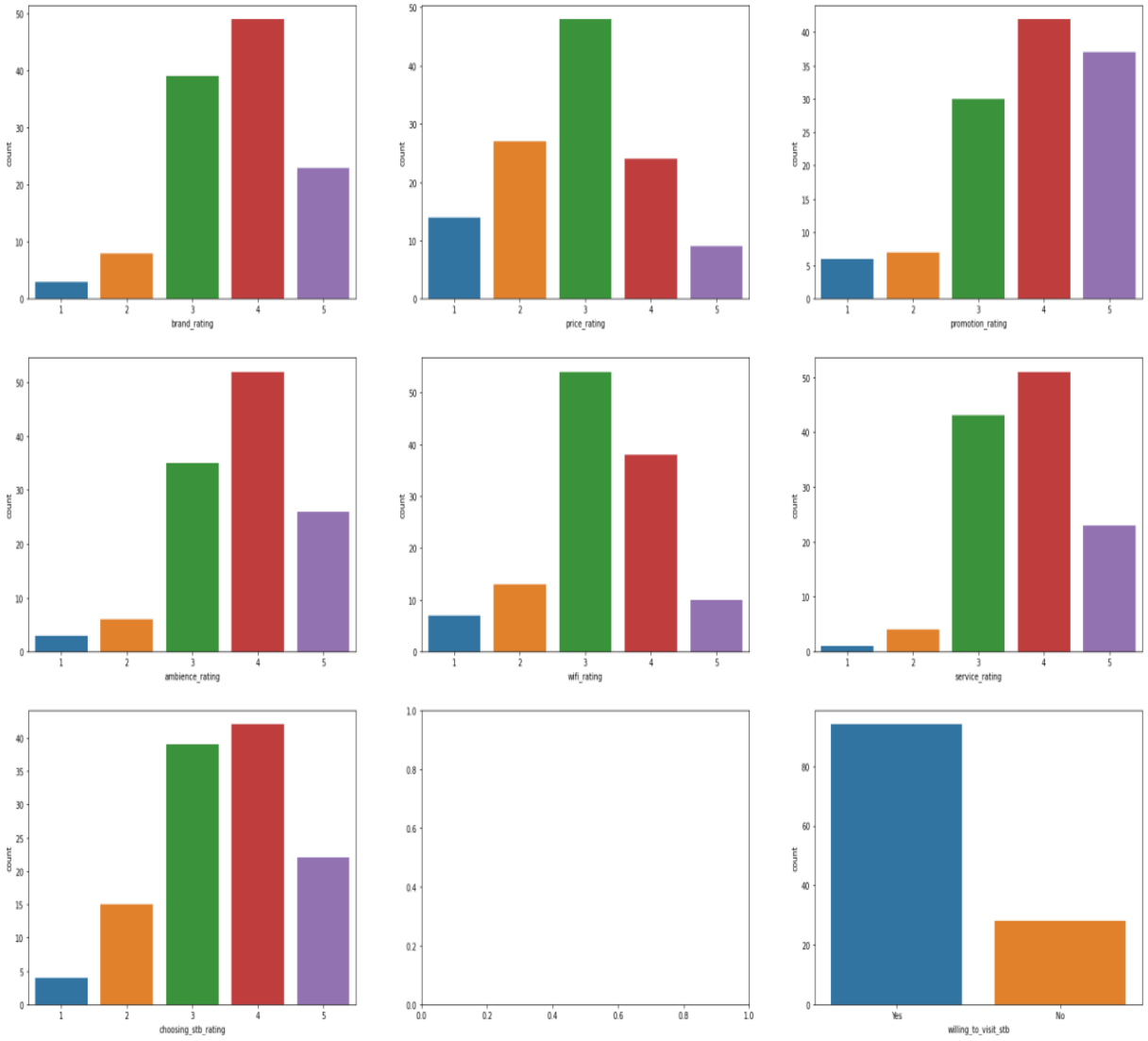




While considering Age into the picture, millennials seem to spend very less in the More than RM40 category products. And interestingly, people in the range 30 to 40+ age range when visited the store always seem to spend on something.



Plotting the count plot of all the ratings column, we see that most of them follow a right skewed distribution towards 5 star rating, **except price and wifi which seem to follow a perfect normal distribution. So there's a chance for improvement.**

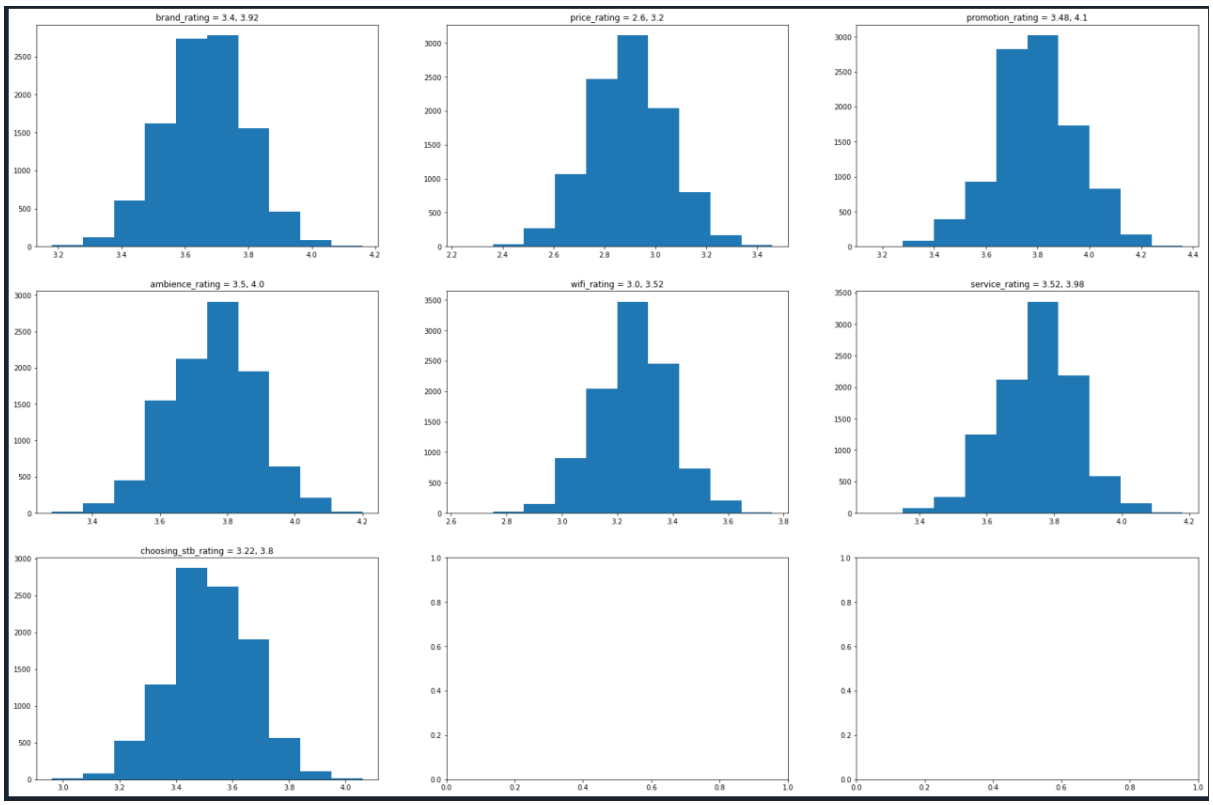


40 and above people also have given a low rating (mean) for wifi while 20 to 29 group influenced the price rating a lot when compared with others.

Age	brand rating	price rating	ambience rating	promotion rating	wifi rating	service rating	choosing stb rating
40 and above	3.71429	3.42857	3.57143	3.57143	2.71429	3.71429	4.14286
Below 20	3.46154	3.15385	3.46154	3.46154	3.53846	3.46154	3.69231
From 20 to 29	3.70588	2.78824	3.8	3.84706	3.21176	3.74118	3.4
From 30 to 39	3.58824	3	3.82353	3.88235	3.47059	4	3.70588

The below graphs calculates the 95% Confidence Interval Mean based on a subsample of 50 and simulated over 10000 times for each category ratings.

Through simulation we can find the same results where price_rating and wifi_rating has a lower interval of 2.6 and 3.0 respectively.



Hypothetically there's a chance for a substantial sales increase from Dine in customers by improving the Wi-Fi network. Since most people who Dine in tend to use laptops, mobile phones.

For the price_rating, we could initially add few more products to the lower category price range or try decreasing the prices of products to a little extent and perform a Z-test to see if lowering the price and improving the Wi-Fi network had any effect on sales.

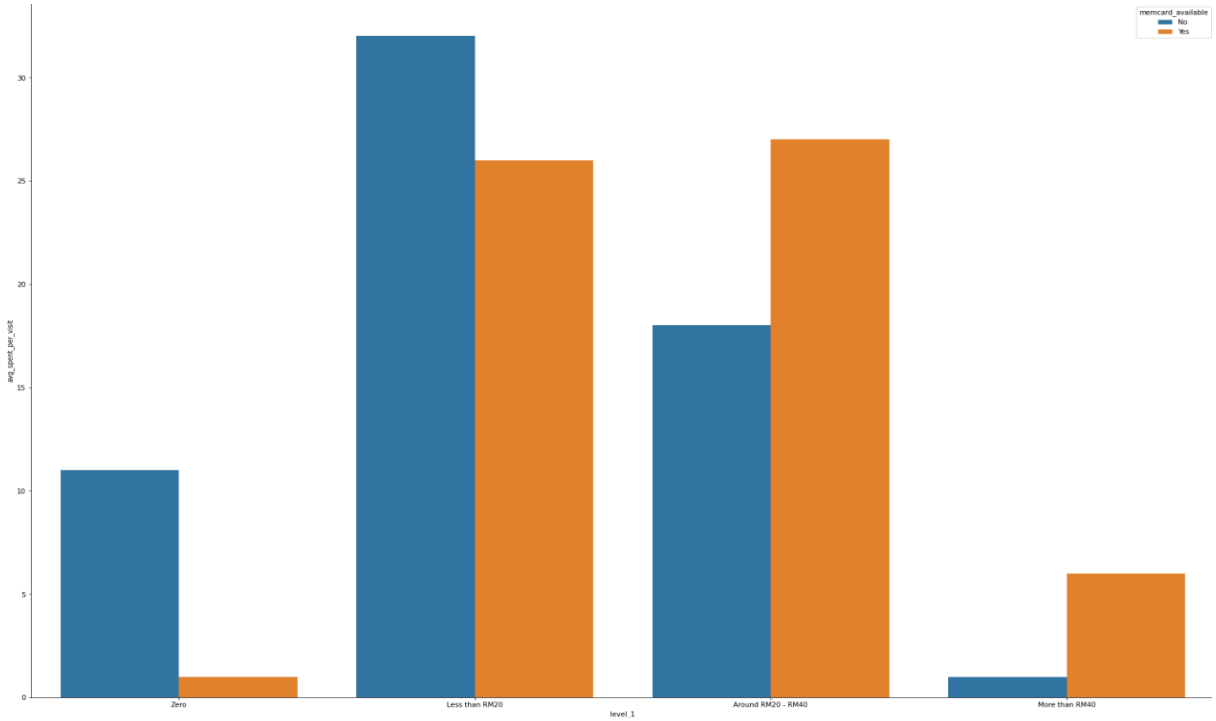
When performed anova test on different groups for price_rating, we can observe a p-value of 0.0078 for avg_spent_per_visit.

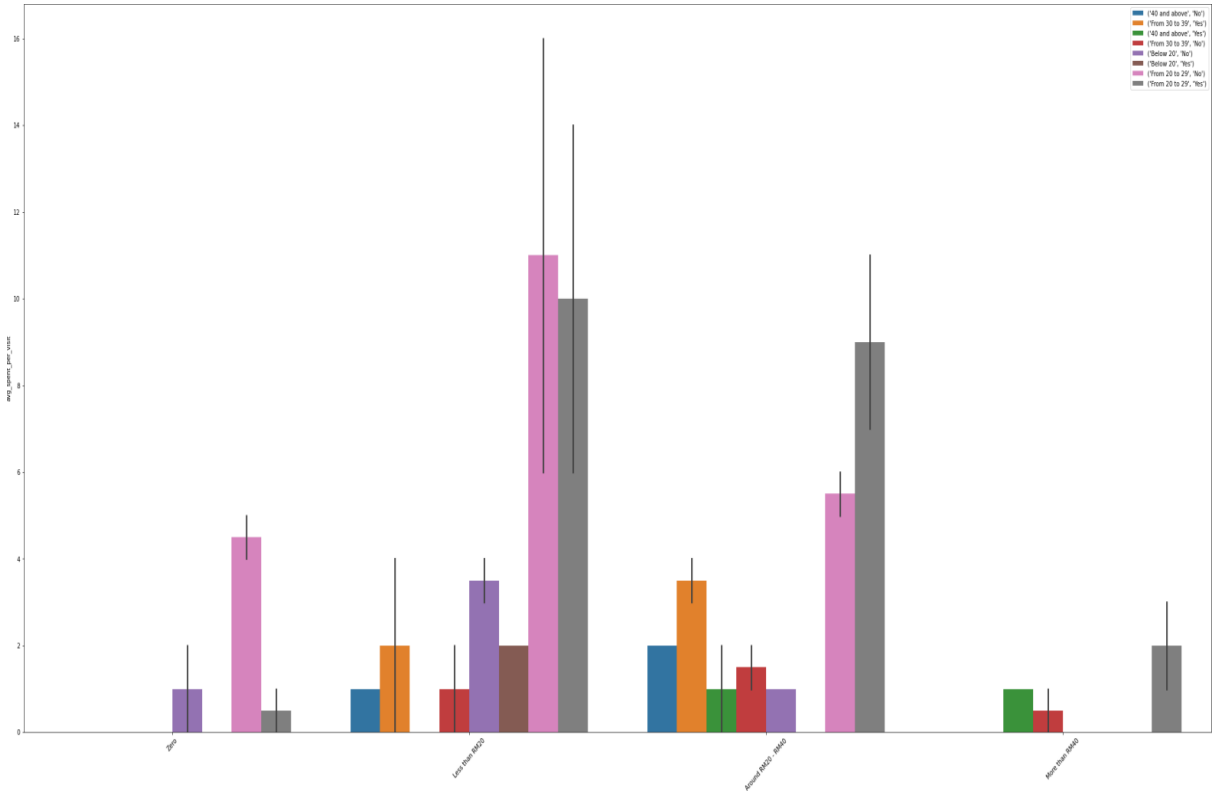
```
   Source  ddof1  ddof2         F     p-unc       np2
0  Gender      1    120  0.023942  0.877291  0.000199

  Source  ddof1  ddof2        F     p-unc       np2
0    Age      3    118  1.14915  0.332308  0.028386

       Source  ddof1  ddof2         F    p-unc       np2
0  Employment      3    118  0.680276  0.56581  0.017001

   Source  ddof1  ddof2         F     p-unc       np2
0  Income      4    117  0.157673  0.959172  0.005362

           Source  ddof1  ddof2         F     p-unc       np2
0  visit_frequency      4    117  1.942522  0.107949  0.062275

       Source  ddof1  ddof2        F     p-unc       np2
0  visit_type      7    113  0.88657  0.519613  0.052061

             Source  ddof1  ddof2         F     p-unc       np2
0  visit_time_spent      4    117  0.793494  0.531696  0.026411

              Source  ddof1  ddof2         F     p-unc       np2
0  nearest_starbucks      2    119  2.246141  0.110275  0.036377

              Source  ddof1  ddof2         F     p-unc       np2
0  memcard_available      1    120  3.079011  0.081861  0.025017

              Source  ddof1  ddof2         F    p-unc       np2
0  frequent_purchase     19    102  1.334884  0.17861  0.199138

                Source  ddof1  ddof2         F    p-unc       np2
0  avg_spent_per_visit      3    118  4.140479  0.00788  0.095241
```

For wifi it seems different visit_time_groups had a p_value of 0.034, which corroborates our suggestion to improve the Wi-fi network.

```
  Source  ddof1  ddof2         F     p-unc       np2
0 Gender      1    120  0.719116  0.398122  0.005957

  Source  ddof1  ddof2         F    p-unc       np2
0    Age      3    118  1.483989  0.22246  0.036357

      Source  ddof1  ddof2         F     p-unc      np2
0 Employment      3    118  0.186541  0.905377  0.00472

  Source  ddof1  ddof2         F     p-unc       np2
0 Income      4    117  2.109273  0.083989  0.067262

          Source  ddof1  ddof2         F     p-unc       np2
0 visit_frequency      4    117  0.862889  0.488524  0.028655

      Source  ddof1  ddof2         F     p-unc       np2
0 visit_type      7    113  1.467435  0.185971  0.083328

           Source  ddof1  ddof2         F     p-unc       np2
0 visit_time_spent      4    117  2.698399  0.034037  0.084461

            Source  ddof1  ddof2         F     p-unc       np2
0 nearest_starbucks      2    119  0.641313  0.528411  0.010663

            Source  ddof1  ddof2         F     p-unc       np2
0 memcard_available      1    120  0.501205  0.480344  0.004159

            Source  ddof1  ddof2         F     p-unc       np2
0 frequent_purchase     19    102  0.816386  0.683164  0.131999

              Source  ddof1  ddof2         F     p-unc       np2
0 avg_spent_per_visit      3    118  0.883939  0.451645  0.021979
```

Customers who have **membership** card seems to have a sense of belonging, as most customers with membership card always seem to spend and sometimes even more in higher category products.





To conclude the analysis, by testing these below suggestions there a chance for subsequent increase in sales

- Adding few more products or discounting the price
- Improving the WIFI network
- Signing up more customers who are above 20 into members

Group G:

Dheshoju Kalyan Kumar

Gurdaan Walia

Manuel Paredes

Keerat Singh Sandhu