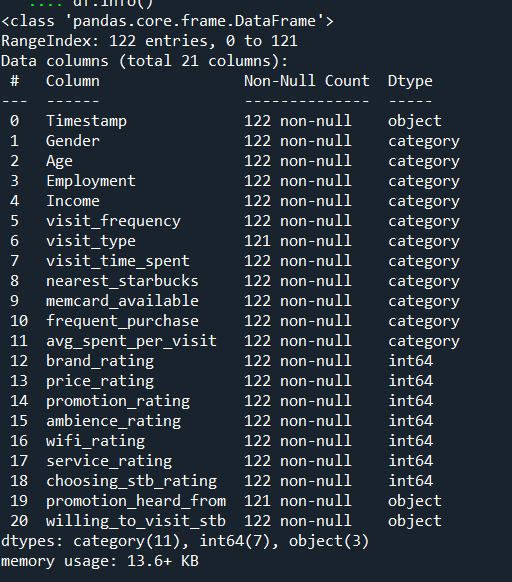
# Starbucks Customer Survey

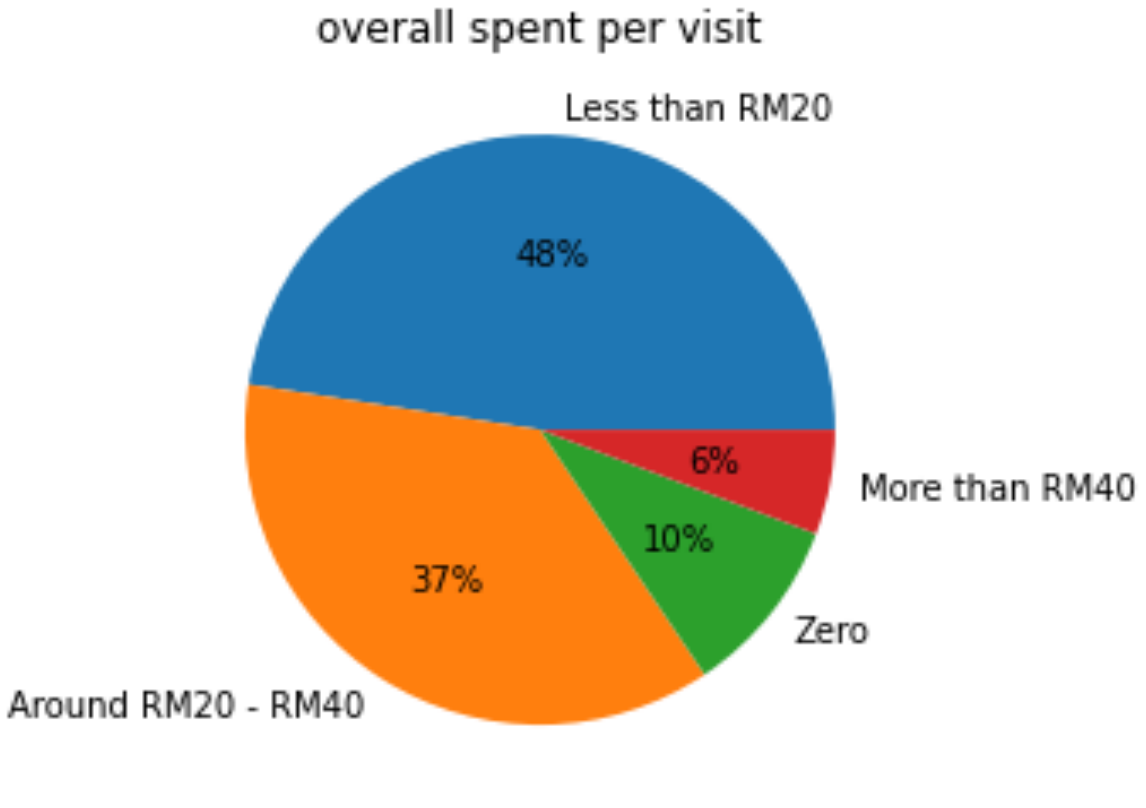
# This dataset is composed of a survey question of over 100 respondents for their buying behaviour at Starbucks.

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.

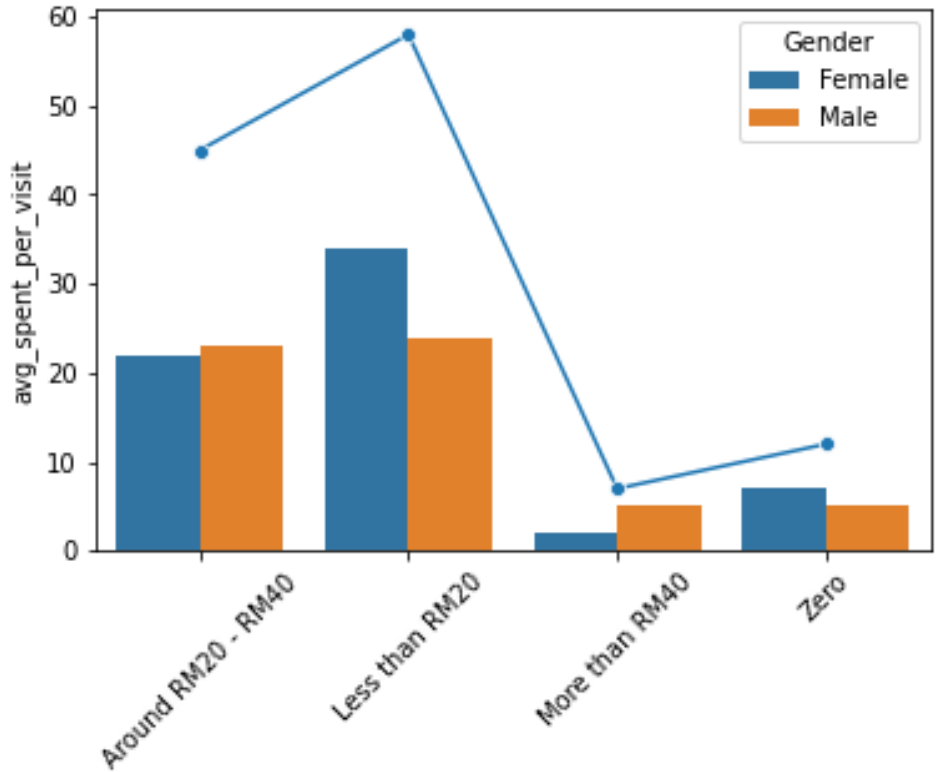


Any business to be sustainable or remain afloat needs to generate subsequent sales. So, in this dataset the column which describes the sales detail is **avg\_spent\_per\_visit**. We will be ignoring the column **frequent\_purchase** as there doesn’t seem to be much variability in the product names.

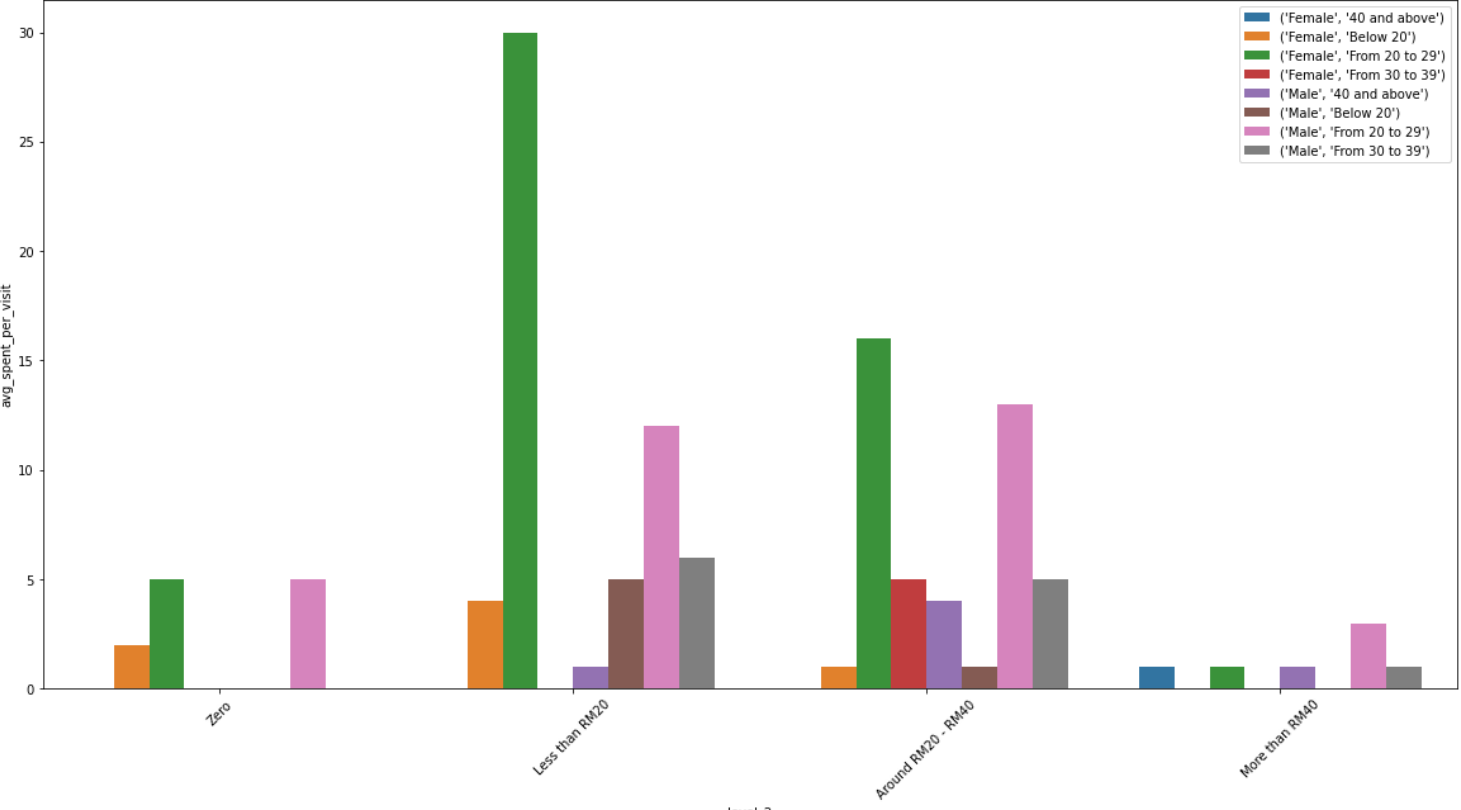
When we start analyzing the **avg\_spent\_per\_visit**, we see that there seems to be a large sales volume in the categories **Less than RM20(5.81 CAD)** and **Around RM20 – RM40 (11.63CAD).**



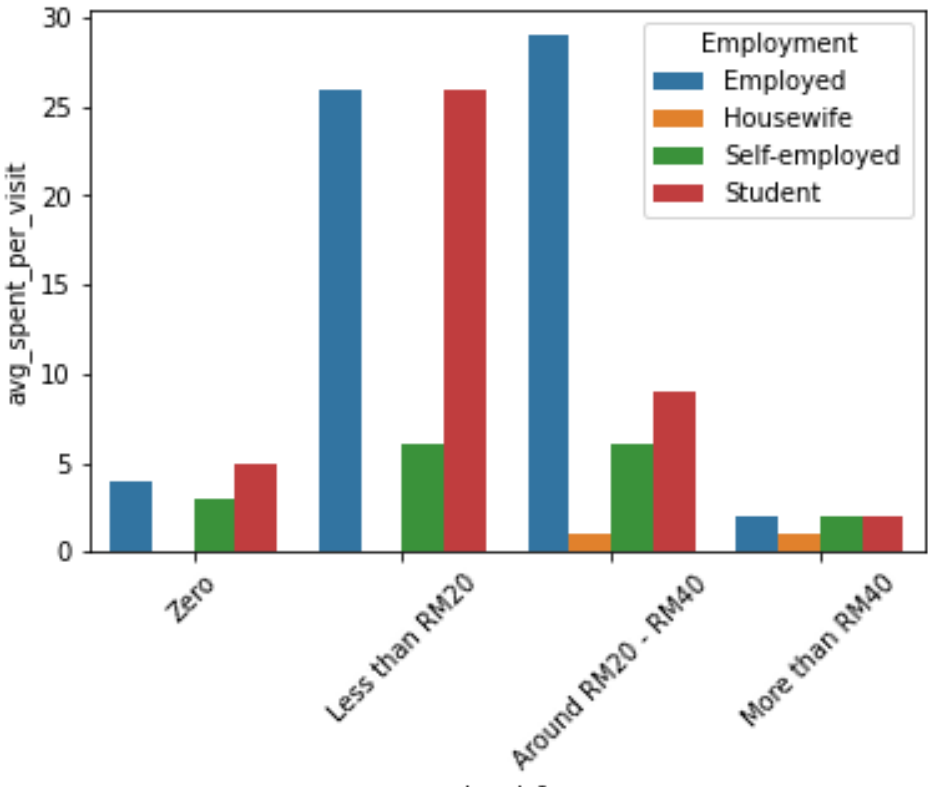
When drilling down deeper among Males and Females, we can observe that females spent more in the category Less than RM20.

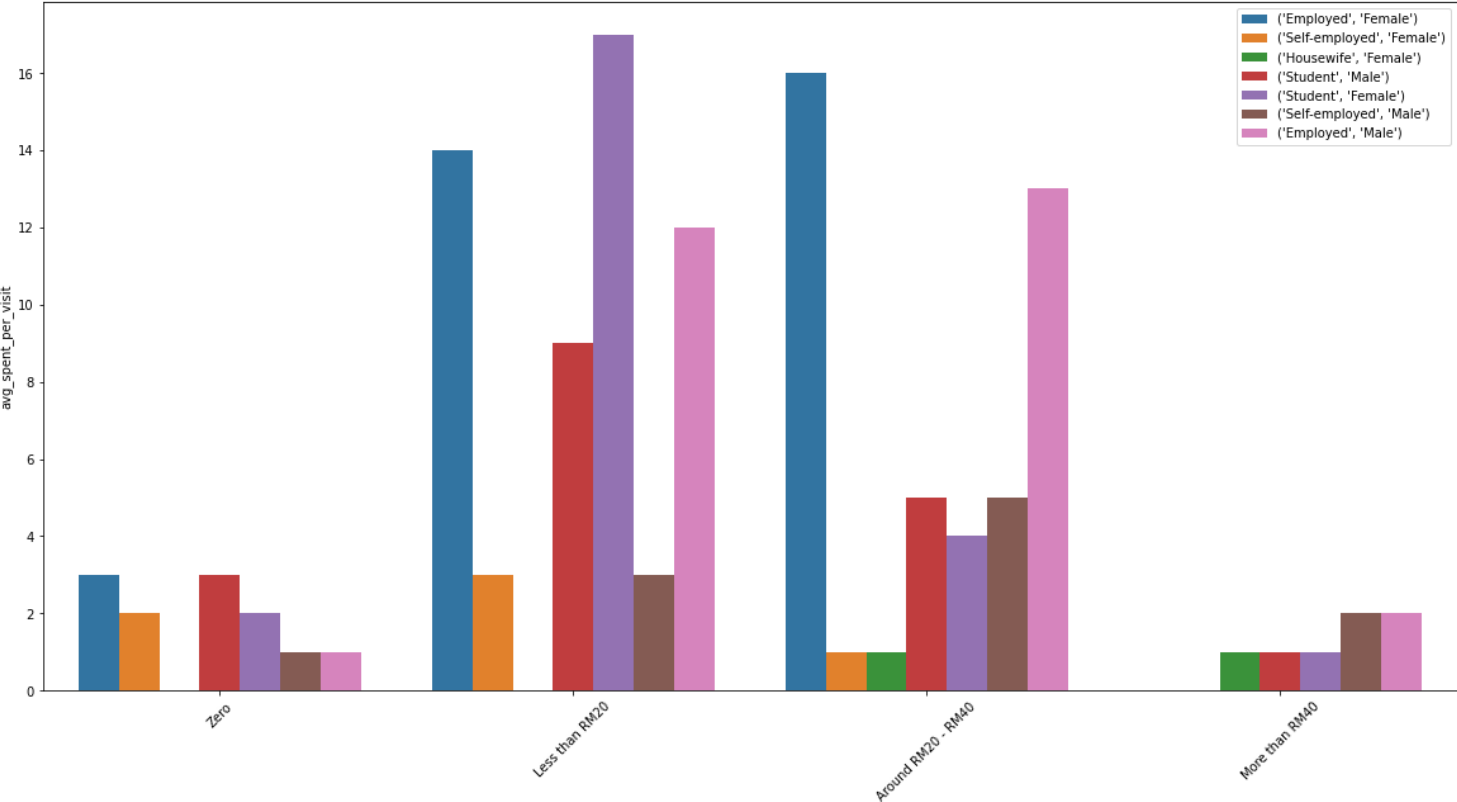


The same follows for different age categories among Males and Females.

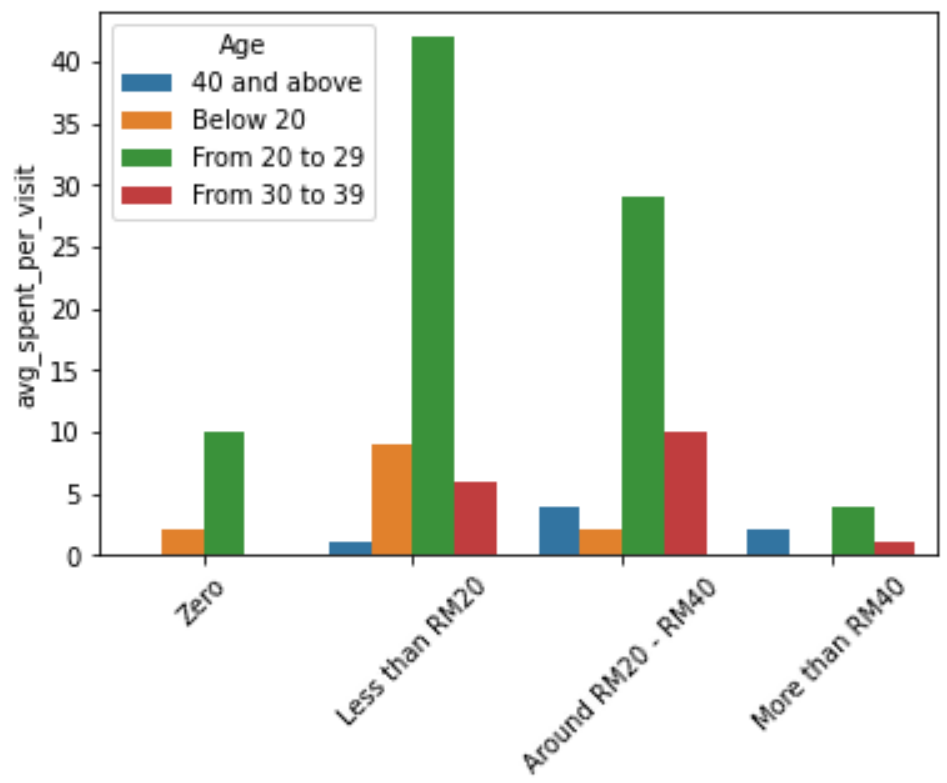


While checking if Employment type had been a confounder for less sales in the category of **More than RM40**, we can outright ignore this assumption since many employed customers too are not willing to spend in the category **More than RM40.**

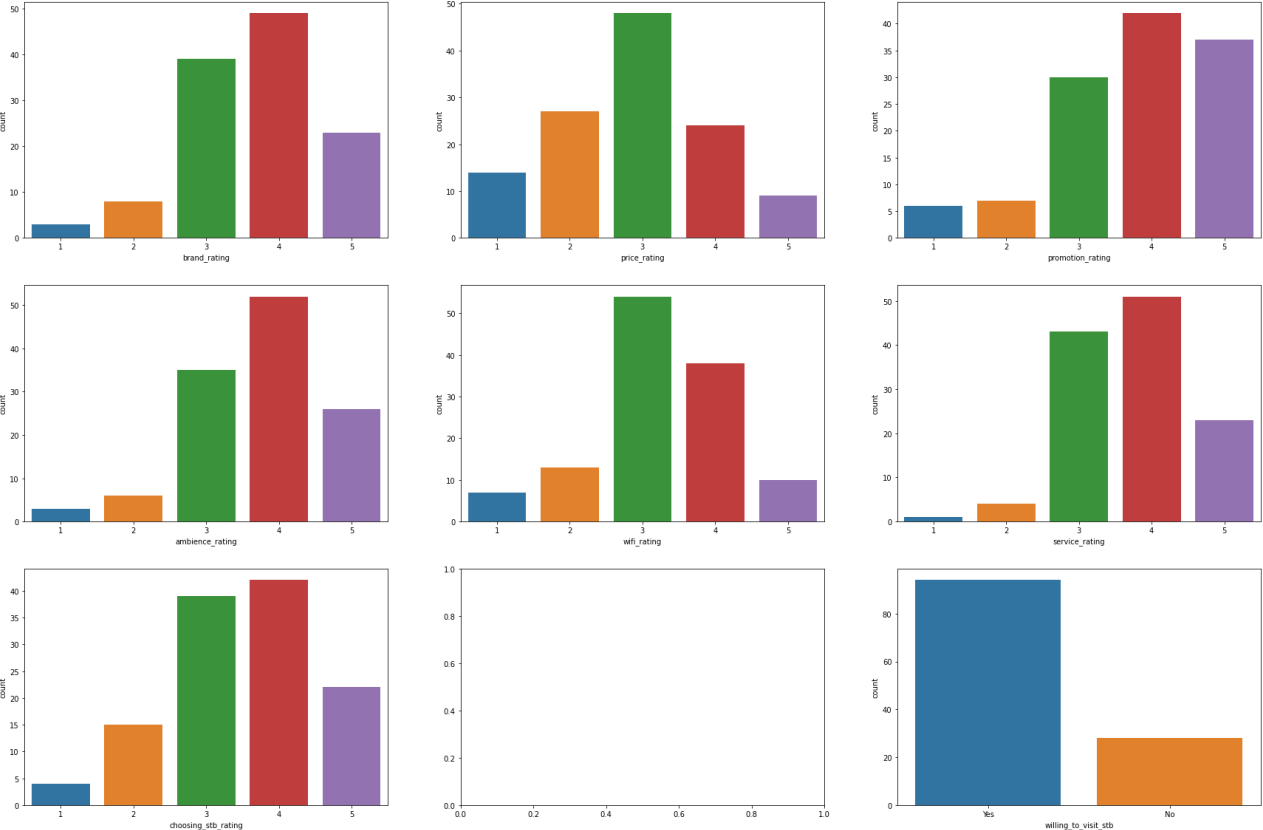




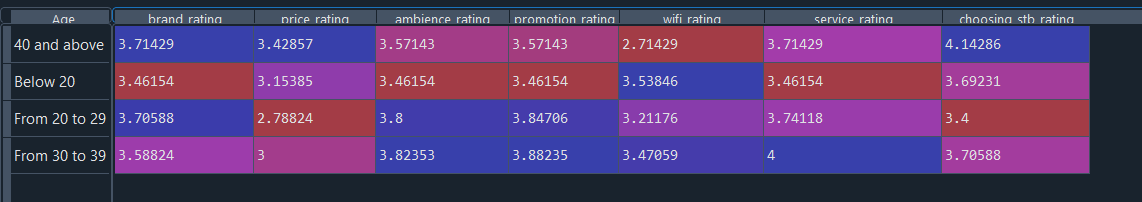
While considering Age into the picture, millennials seem to spend very less in the More than RM40 category products. And interestingly, people in the range 30 to 40+ age range when visited the store always seem to spend on something.



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.**

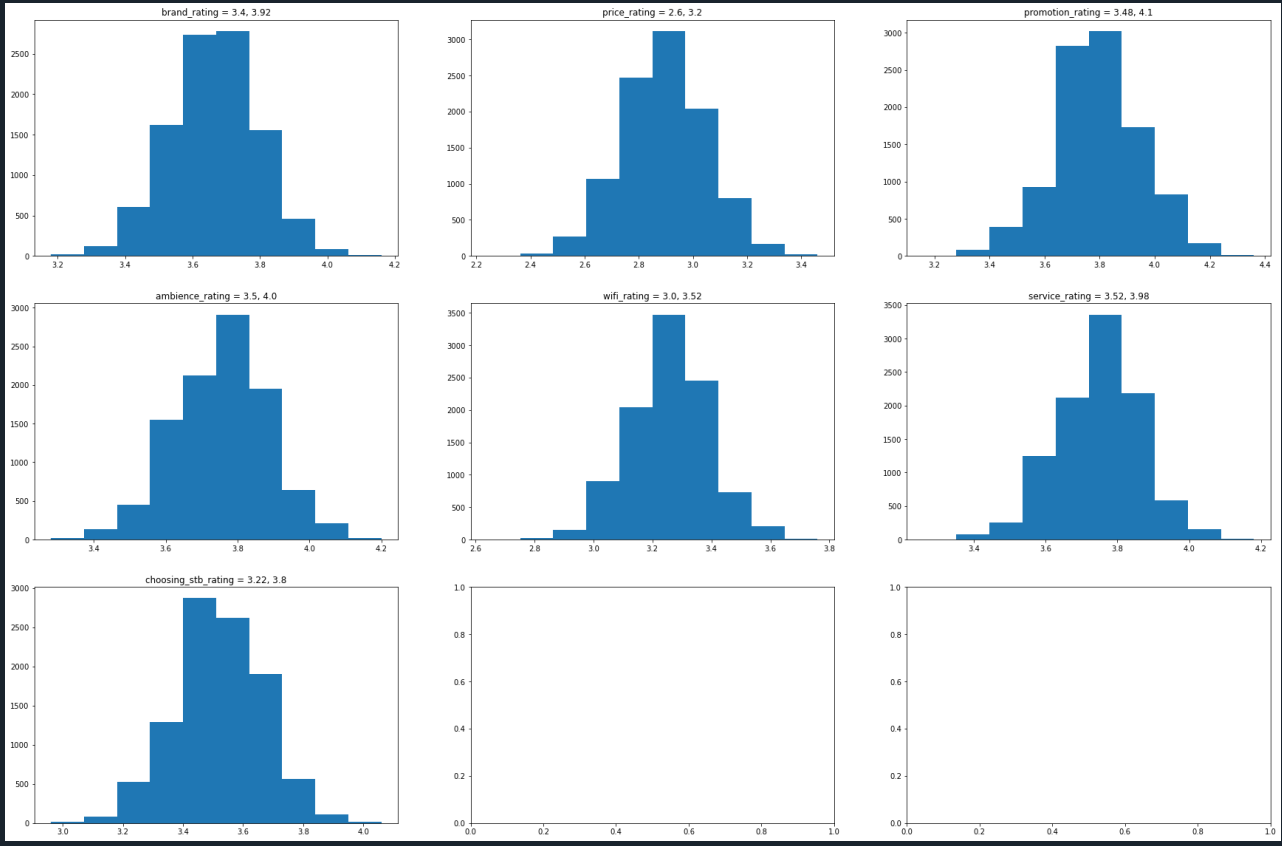


40 and above people also have given a low rating (mean) for wifi while 20 to 29 group influenced the price rating a lot when compared with others.



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

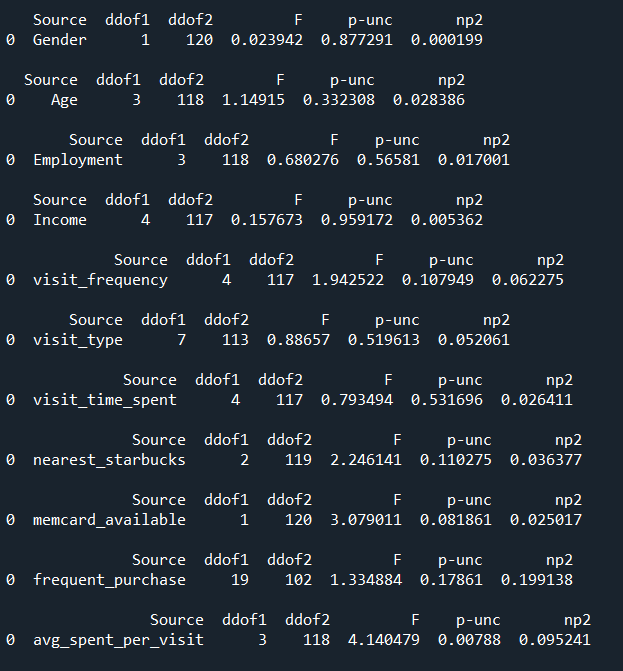
Through simulation we can find the same results where price\_rating and wifi\_rating has a lower interval of 2.6 and 3.0 respectively.



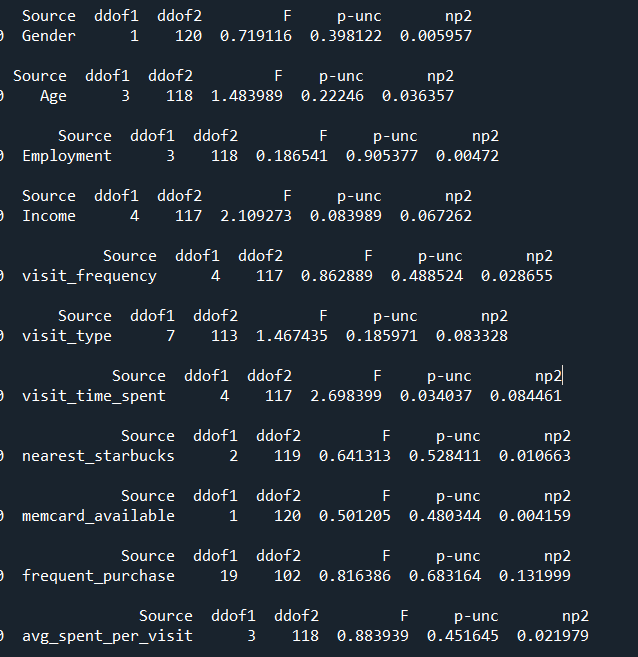
Hypothetically there’s a chance for a substantial sales increase from Dine in customers by improving the Wi-Fi network. Since most people who Dine in tend to use laptops, mobile phones.

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.

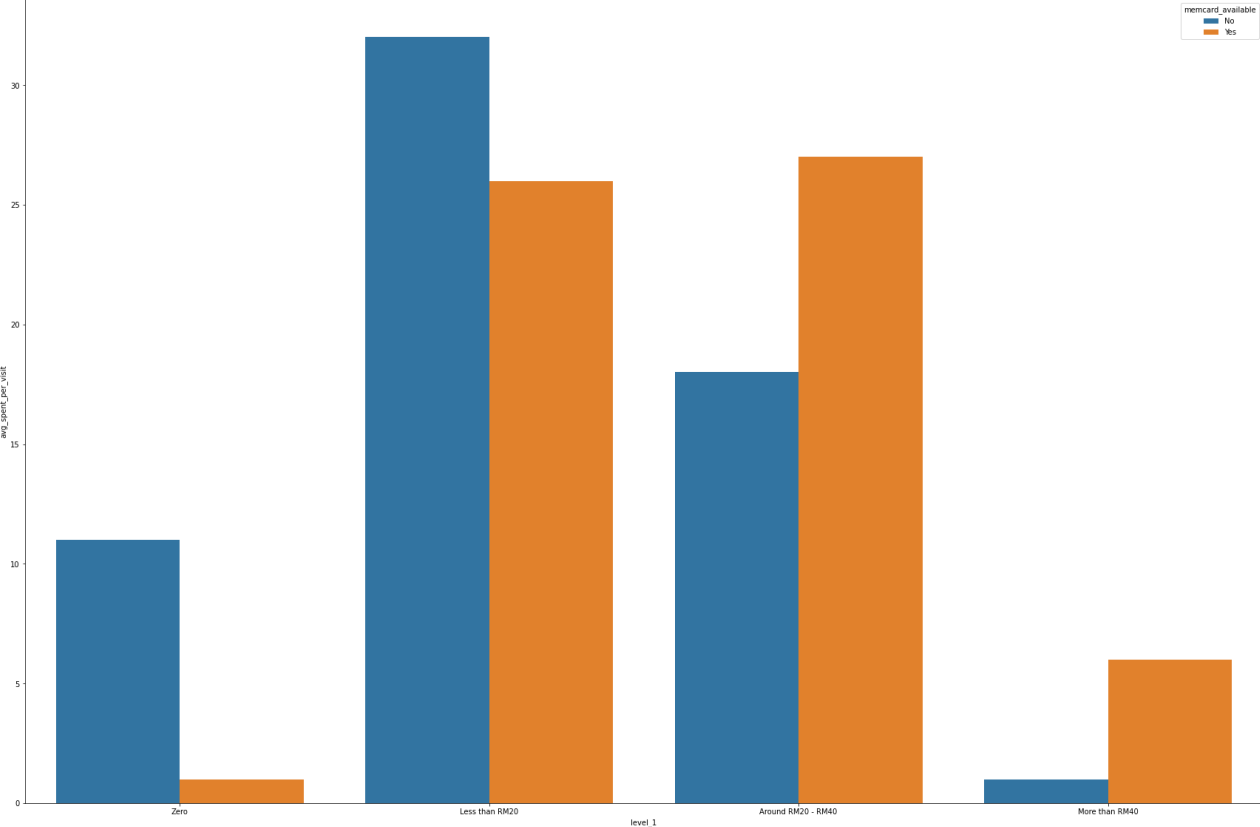
When performed anova test on different groups for price\_rating, we can observe a p-value of 0.0078 for avg\_spent\_per\_visit.

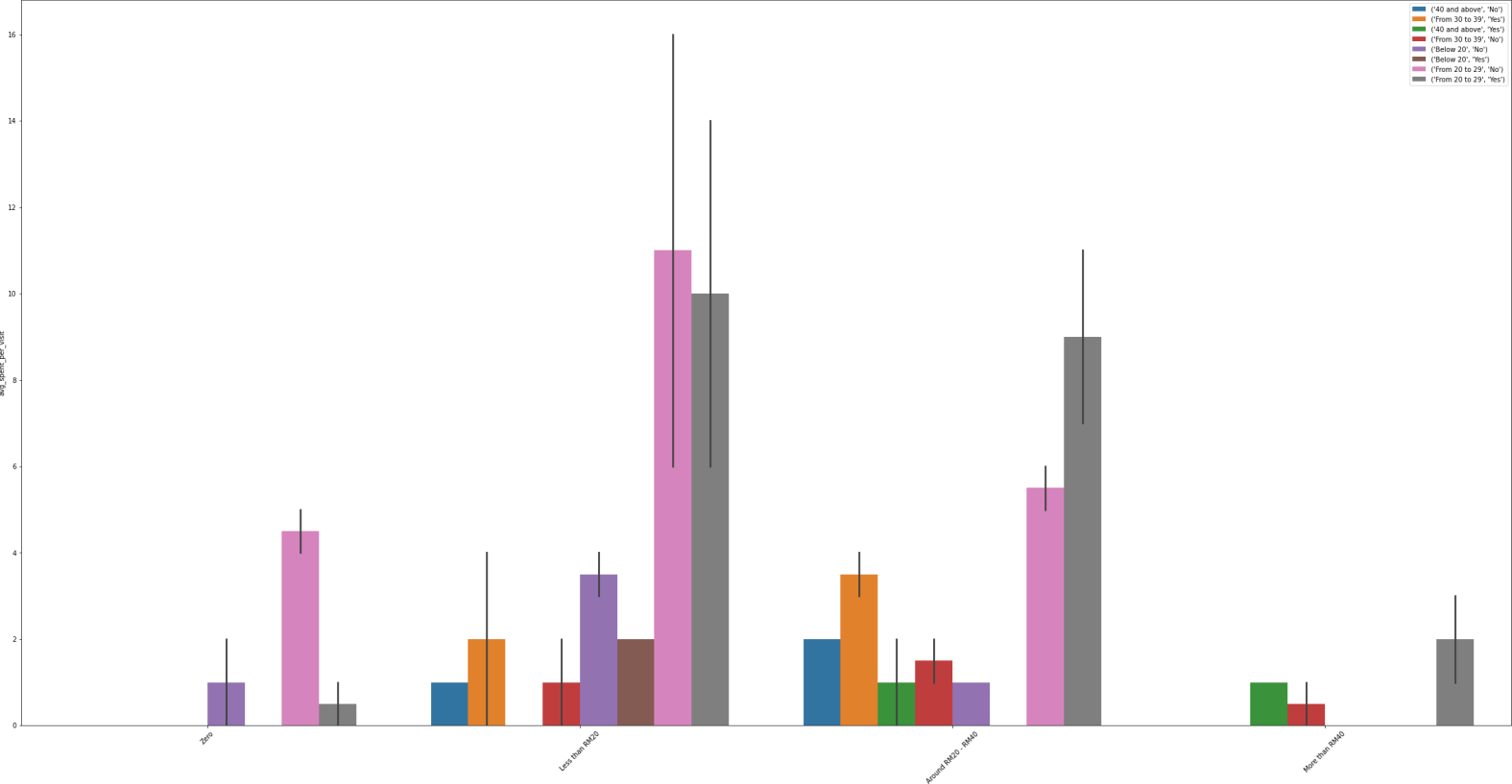


For wifi it seems different visit\_time\_groups had a p\_value of 0.034, which corroborates our suggestion to improve the Wi-fi network.



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