Contents

[**EDA - Supermarket Sales** 2](#_Toc99103656)

[1. Introduction ( 15 Minutes ) 3](#_Toc99103657)

[Objective ( ~ 2 Min ) 3](#_Toc99103658)

[Data ( ~ 2 Min ) 3](#_Toc99103659)

[Data Dictionary ( ~ 10 Min including loading of libraries ) 3](#_Toc99103660)

[2. Data Wrangling ( 35 Minutes ) 4](#_Toc99103661)

[Data Imputation 6](#_Toc99103662)

[Adding fields for time series analysis 10](#_Toc99103663)

[2. Univariate Analysis ( Quantitative ) *( 25 Minutes)* 11](#_Toc99103664)

[3. Univariate Analysis ( Categorical ) *( 25 Minutes )* 16](#_Toc99103665)

[Stores Location on map 20](#_Toc99103666)

[4. Bivariate Analysis *( 30 Minutes )* 21](#_Toc99103667)

[Heatmap 21](#_Toc99103668)

[5. Timeseries Analysis ( 25 Minutes ) 27](#_Toc99103669)

[6. Customers across Cities ( 5 Minutes ) 31](#_Toc99103670)

[7. Member/Non-Member Deep Dive ( 5 minutes ) 31](#_Toc99103671)

[7. Final Insights ( 5 minutes ) 33](#_Toc99103672)



**EDA - Supermarket Sales**

Notes for the educator –

* This EDA will give students a good idea on how to tackle a business problem using EDA. The objective is to equip them a peek into how a Retail analysis is done. Please approach the session with that objective in mind. Discuss a lot about the problems, the ideas. Engage students in discussions for their views and ideas too if possible.
* There are new functions and charts in this EDA that have not been covered in this week earlier – for example nunique(), drop\_duplicates(), style, Geo map etc. Put additional focus on such new additions.
* The Jupyter notebook is exhaustive with comments, description, topics and introductions. Please use the same to brief students about each section accordingly.
* Explain each code line by line. There are comments for each code line to provide a basic idea about how the code is being executed.
* Data Dictionary – Spend considerable amount of time (5-10 mins) in going through the data dictionary either in Excel sheet or Jupyter notebook. Cover all the fields used in Notebook and give students a good understanding of those variables.
* Follow timelines mentioned on each topic. There can be 10-20% deviation as per the skillset level of class but try to adhere to mentioned timelines.
* For all the charts in Univariate/Bivariate analysis, there should be 3 types of discussions in the class –
  + Code/ Function explanation line by line
  + Insights discussion
  + Alternate visualizations chart ideas

## 1. Introduction ( 15 Minutes )

This case study presents a real world alike scenario where important decisions are influenced by the data driven insights. We'll explore the transaction level data sales data of a supermarket chain, and derive useful business insights.

### Objective ( ~ 2 Min )

The supermarket chain wants to run some promotions and they have a fixed budget for the same. Business wants insights from the data on how to most efficiently allocate these funds.

The objectives are as following -

* Provide insights to Marketing head for allocating funds. The funds should be allocated in such a way so that the promotional return can be maximized.

(Promotional Return – When a promotion is run for a specific product/customer, the difference in sales before the promotions and after the promotion is called promotional return )

* Any new Strategy or strategy changes suggestions to improve revenue.

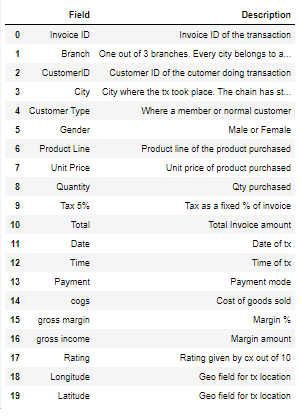
### Data ( ~ 2 Min )

* The data belongs to a supermarket chain in Myanmar, with stores in 3 cities.
* This is a sampled data of 1000 transactions and 20 fields for each transaction. A transaction here records the billing details of a particular product, while capturing important details like Customer ID, time, product name, quantity, price, units etc about the purchase. Please note that if a customer has purchased 3 different types of items, the data will have 3 separate rows ( transactions) for the same.
* All transactions belong to year 2019

### Data Dictionary ( ~ 10 Min including loading of libraries )

Reading data dictionary –





Note : Tx is usually short for transaction, cx for customer.

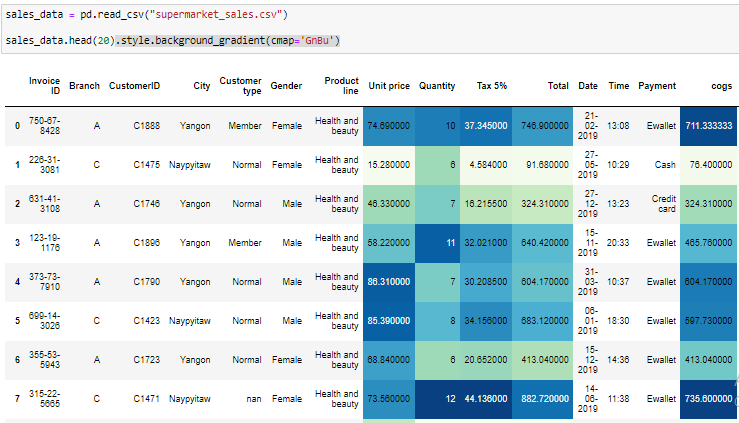
## 2. Data Wrangling ( 35 Minutes )

Style.background is being shown for the first time. Extra attention would be required there.

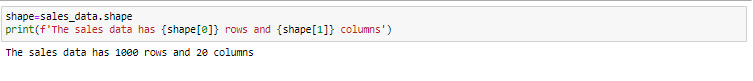
Reading data file using pandas –

(Please note the style function used to color code the quantitative fields. A really useful method to visualise the data with the data ingestion)





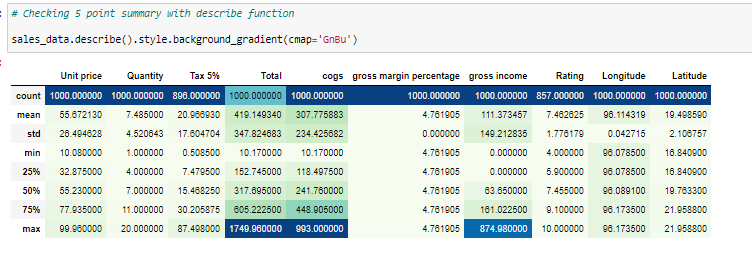
##### *Checking shape of data*



##### *5 Number summary*

A 5 number summary consists of five values: the most extreme values in the data set (the maximum and minimum values), the lower and upper quartiles, and the median. These values are presented together and ordered from lowest to highest: minimum value, lower quartile (Q1), median value (Q2), upper quartile (Q3), maximum value

The color coding is used to compare each cell with respect to the column’s range and gives an idea about how low or high it is in that regard from a benchmark point of view.

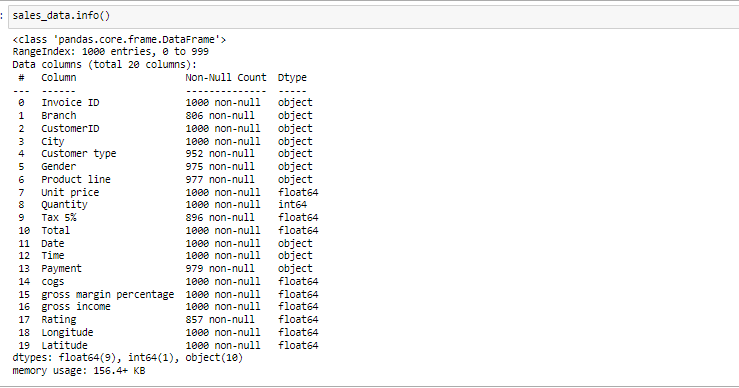


Following are the observations from 5 number summary -

* Unit Price has range of 10-99, with mean price of 55
* Avg qty ordered per transaction is 6.5, the range being 1-14
* Invoice total mean is 323, and the range is 10-1042
* Gross Margin is constant as 4.76%
* Mean rating is 7.46/10

##### *Data types and Null counts*

.info() function is used to check data types of all variables and null counts.



*Rule for discarding columns with more NULLs –*

Although there is no standard rule or threshold value to decide where should you draw the line on whether it is okay to impute NULLs or discard the column, still in most cases generally 35-40% NULLs are considered a lot and makes a field not much valuable. Although, it depends on problems statement and business use case, sometimes a column is so important that even 20% availability of data ( 80% NULLs) can be considered of great value. Students are advised to make a rational and hand picked rational judgement about this.

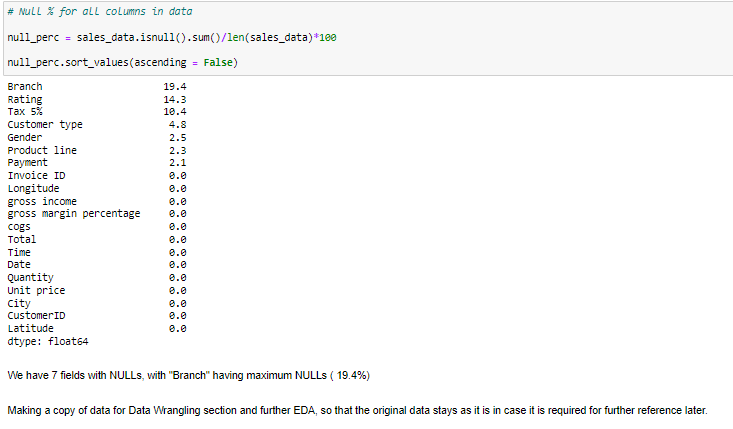
For example, in a case where a company is looking to rank their loyal customers, even a 20% available data about customers who are purchasing more with a competitor can be considered deal breaking.

### Data Imputation

As discussed in Credit Risk case study, data imputation is a necessary step so as to avoid misleading insights. It also makes the data frame compatible with most of the analytical and visual functions.

##### *Checking Null %s*

Let's first check the columns which have NULL values and then we can impute in those fields one by one –



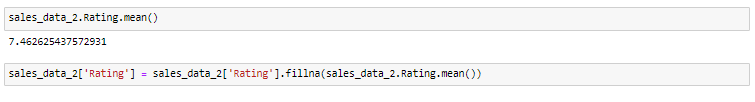
We have 7 fields with NULLs, with "Branch" having maximum NULLs ( 19.4%)

Making a copy of data for Data Wrangling section and further EDA, so that the original data stays as it is in case it is required for further reference later.



##### *Imputing NULLs for Ratings*

Ratings can be imputed with Mean rating values



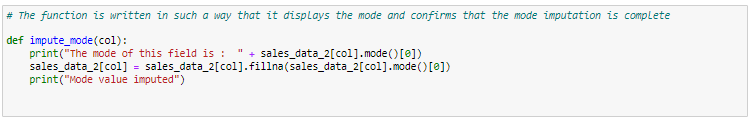
##### *Imputing NULLs for Tax*

From the data we can see that Tax is 5% of the Invoice value always. We can impute nulls using the same formula as below –





Mode is a common way to fill in Categorical variables unless there is a linkage among variables ( As in case of Branch). Thus, it is better to define a function which can be used multiple times for such variables –



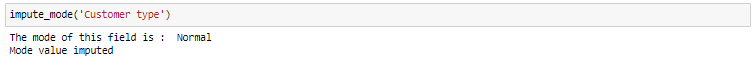
##### *Imputing NULLs for Payment*

Payments can be imputed with Mode as there is no logical way to find out what method was used.



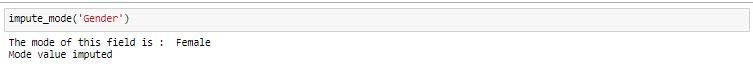
##### *Imputing NULLs for Customer Type*

Customer Type has just 5% of Nulls and it is better to complete the data with mode imputation than leaving it as it is



##### *Imputing NULLs for Gender*

Similar to Customer type, Gender also has lower % of NULLs and can be imputed with Mode



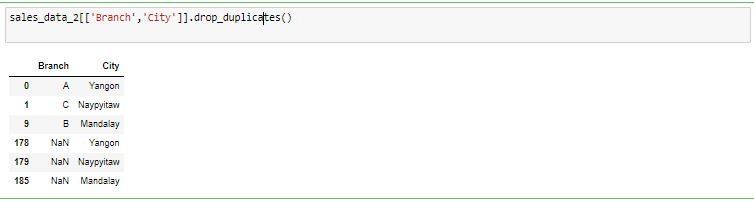
##### *Imputing NULLs for Product Line*



##### *Imputing NULLs for Branch*

One can logically deduce that City and Branch must be linked. A quick look at the data using head() function would confirm the same as below –

NOTE : drop\_duplicates() can be used to find the unique values based on a set of fields.



Based on the above values, we can write the following logic to impute Branch as per the City values –



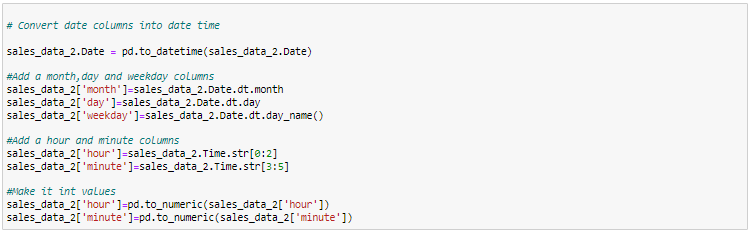
##### *Verifying NULL% after imputations*

Finally we can verify that all columns have 0% NULLs suing the same logic as earlier for NULL % calculation –

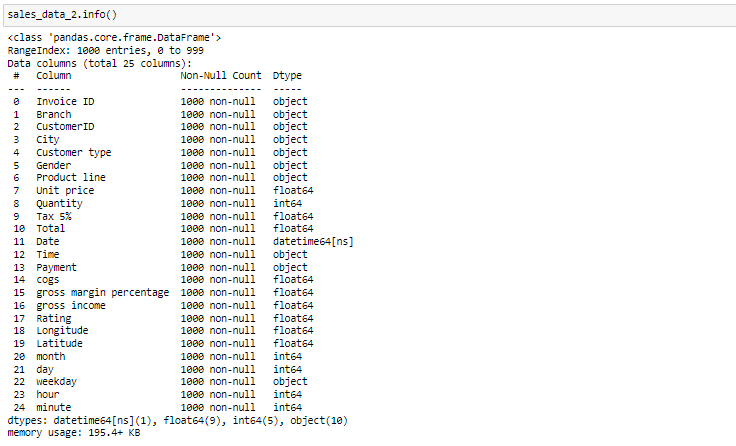


### Adding fields for time series analysis

Time based analysis becomes really crucial for a sales data. It can inform about patterns, peak & low seasons, events that trigger sales to be high or low etc. It is useful to extract data and have separate time fields for such analysis as done below -



##### *Checking final data types and NULL counts*



Outcome :

All the variables have 1000 non-null rows as desired, and there are no Null values.

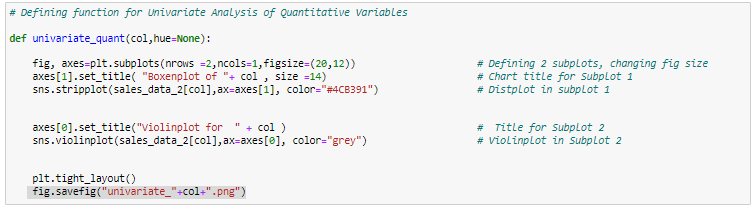
## 2. Univariate Analysis ( Quantitative ) *( 25 Minutes)*

As already discussed in Day 4 session, we know that Univariate Analysis is simplest form of analysing data. It restricts the analysis to only 1 variable as the name states.

It doesn't take into account the mutual relationships and associations among variables. Rather it focuses on finding patterns through a particular field.

##### *Function for Univariate Charts of quantitative fields –*

For our Univariate analysis in this case, we are going to plot Violinplot and Stripplot using Seaborn library. Violinplot is a hybrid plot of Box lot & KDE plot which gives us insights about quartiles as well as distribution of a variable. Stripplot is considered to be good complement to Violinplot because one can see the individual data points along with the distribution.



*Instructions on how to read and interpret this and further charts for Insights –*

* The boundary of Violin plot represents KDE plot. It tells us about the distribution of a variable, and its various characteristics like – Skew, Kurtosis, Normality etc.
* The lines in the middle of violinplot represent the quartiles. They’re basically analogous to a box plot. One can get Median, 25%ile, 75%ile etc. values from here.
* The Stripplot is a great complement to Violinplot as it shows individual data points and shows how much the data points are concentrated in a given range.

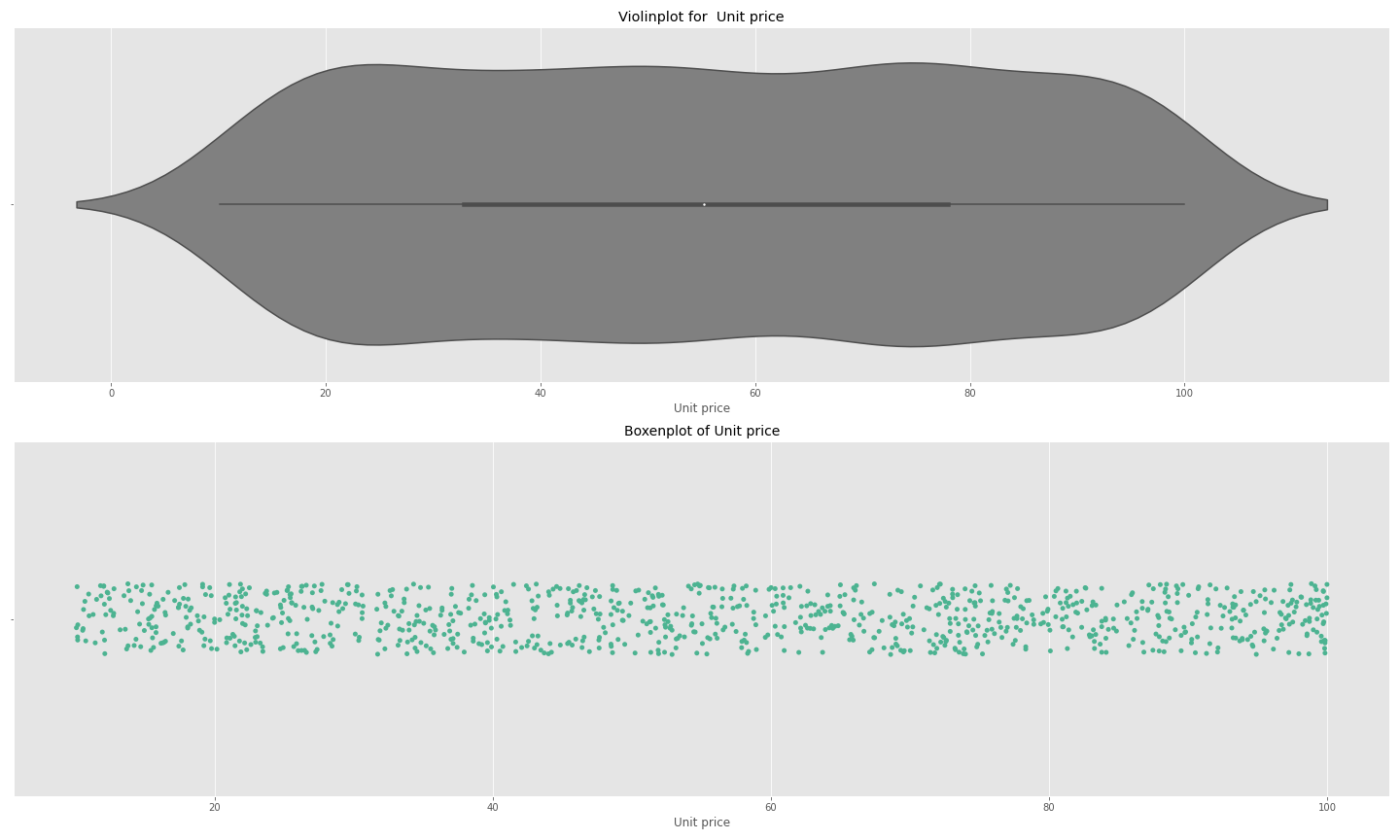
##### *Unit Price*

Univariate analysis of Unit Price will help us understand in general how the products are priced in supermarket catalogue.

Insights-

* Unit price doesn't follow a normal distribution; it is more or less similar frequency for range 20-80. Statistically speaking it has a –ve Kurtosis.





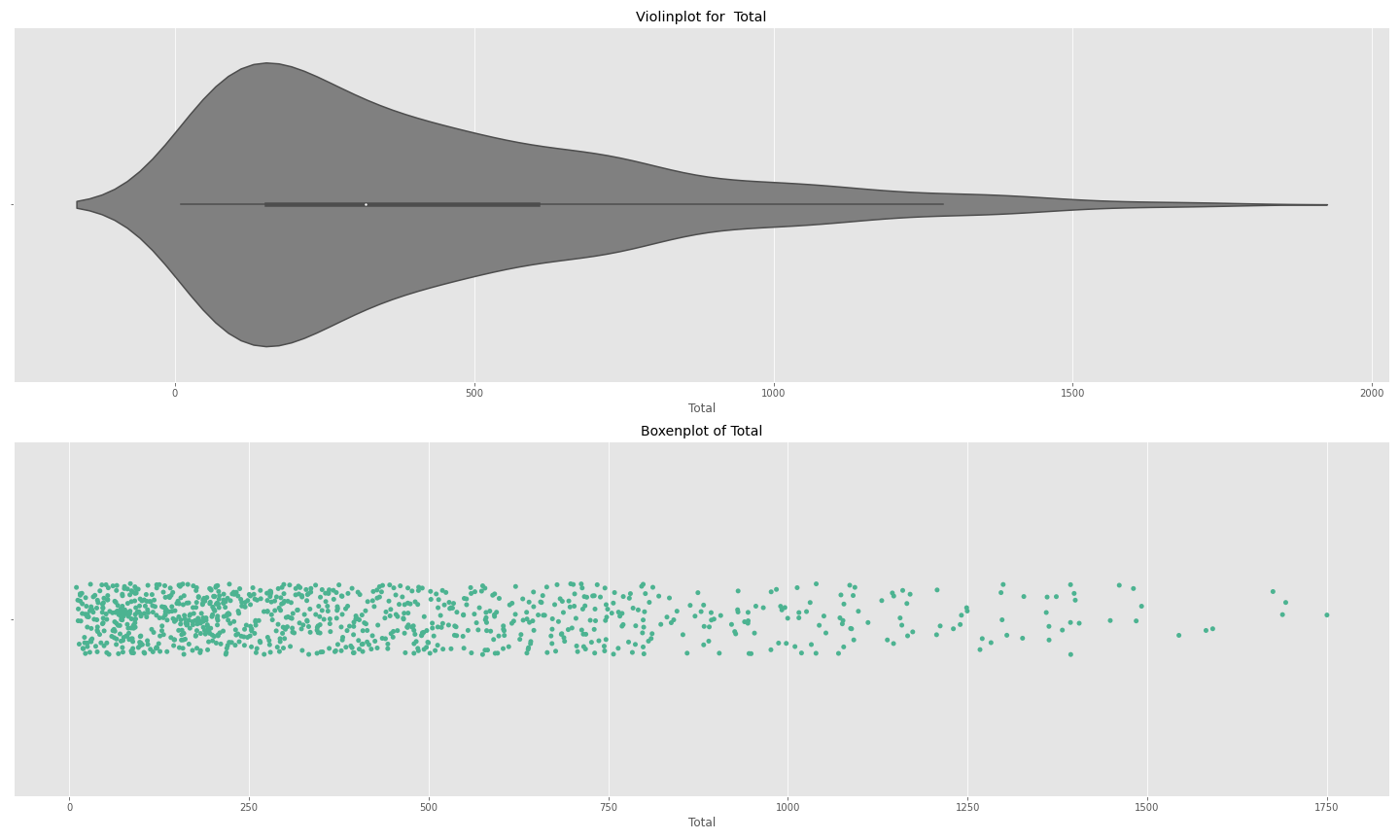
##### *Invoice Amount*

Univariate analysis of Invoice Amount will help us understand in general how much is the bill amount per transaction

Insights-

* It has a right skewed distribution, with majority of values lying in under 500 range.



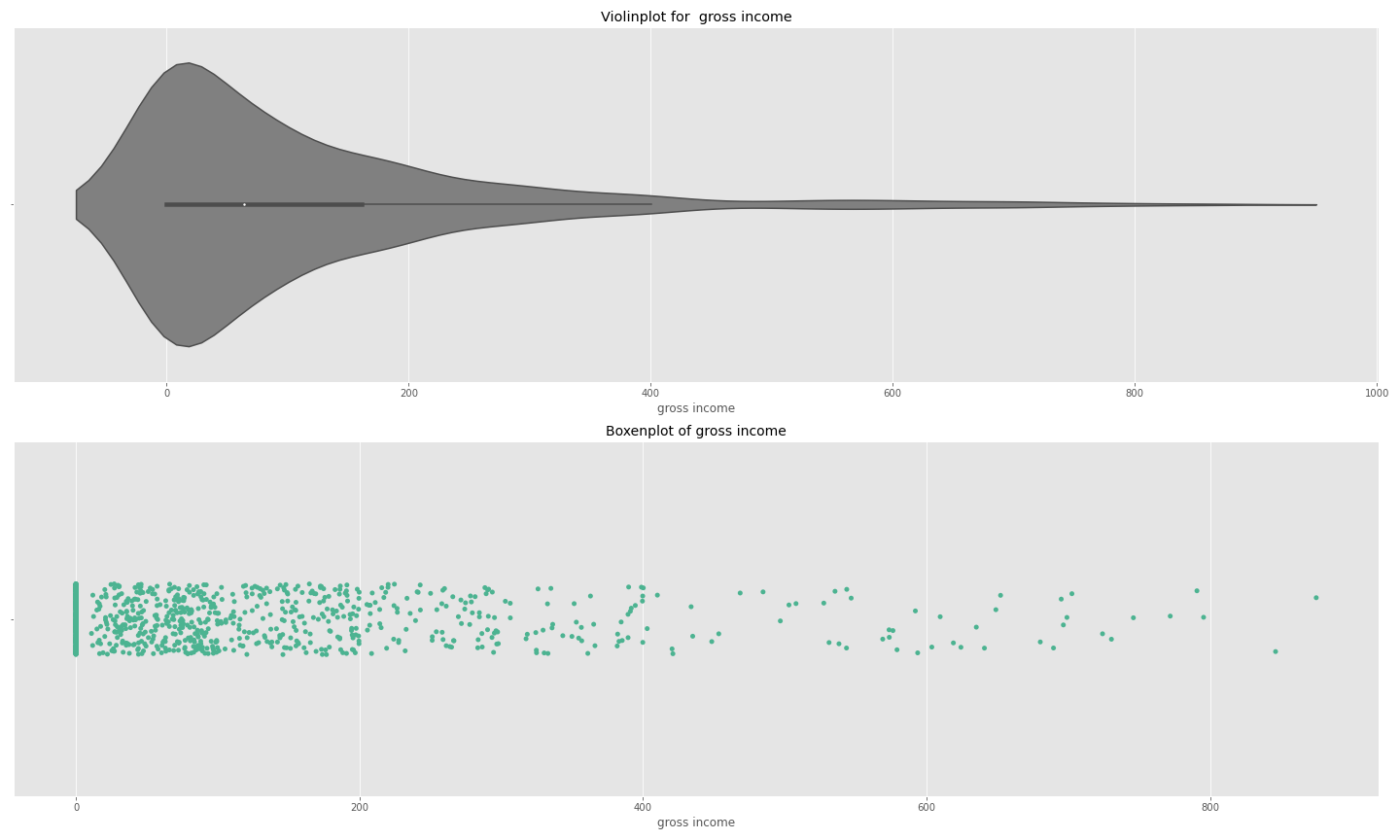


##### *Gross Income*

Insights-

* Since gross income is derived out of Invoice amount and COGS, the distribution is similar.



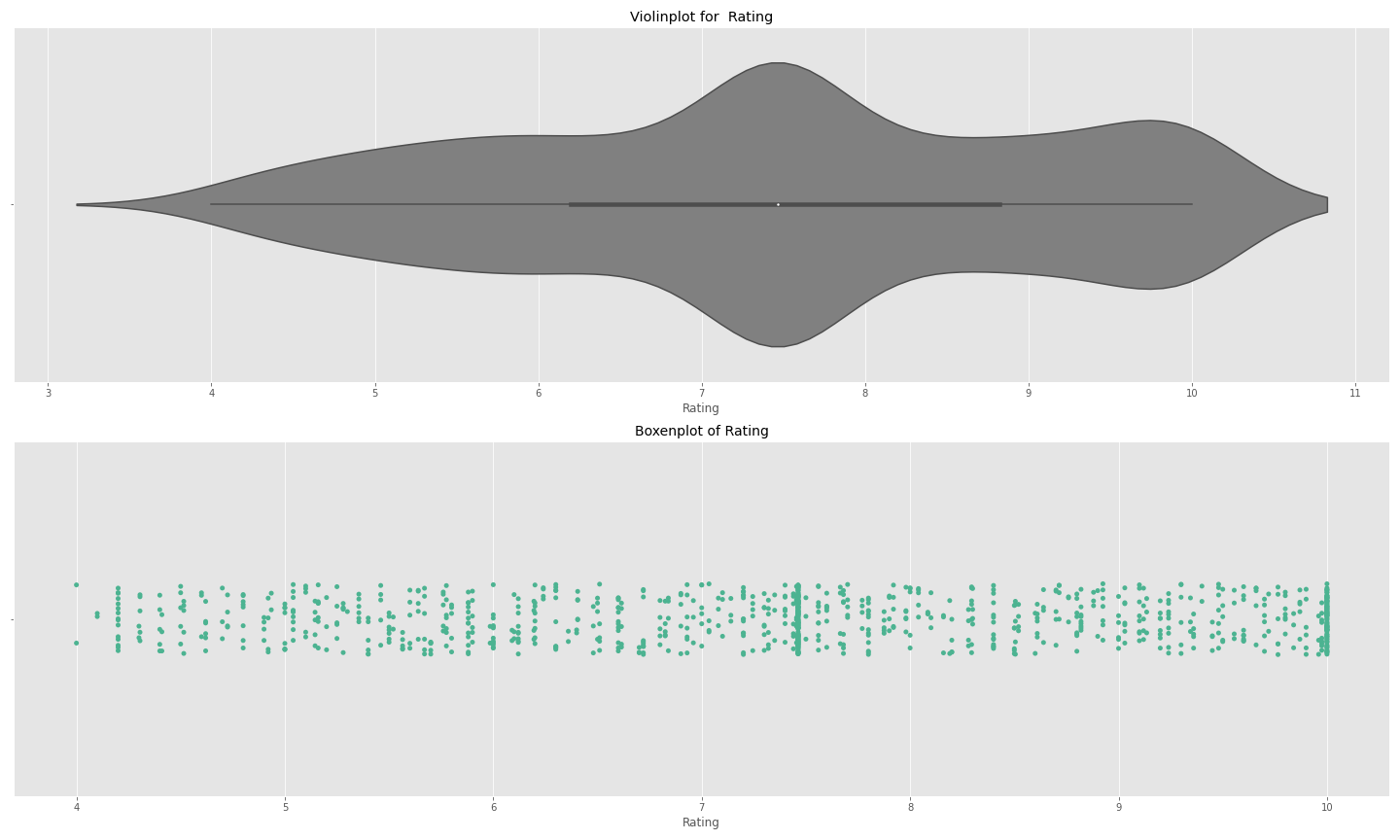


##### *Ratings*

Insights-

* Ratings do have a similar to normal distribution with mean at around 7.5 rating..
* There are much lesser ratings in 3-5 range which is a good sign for the company.



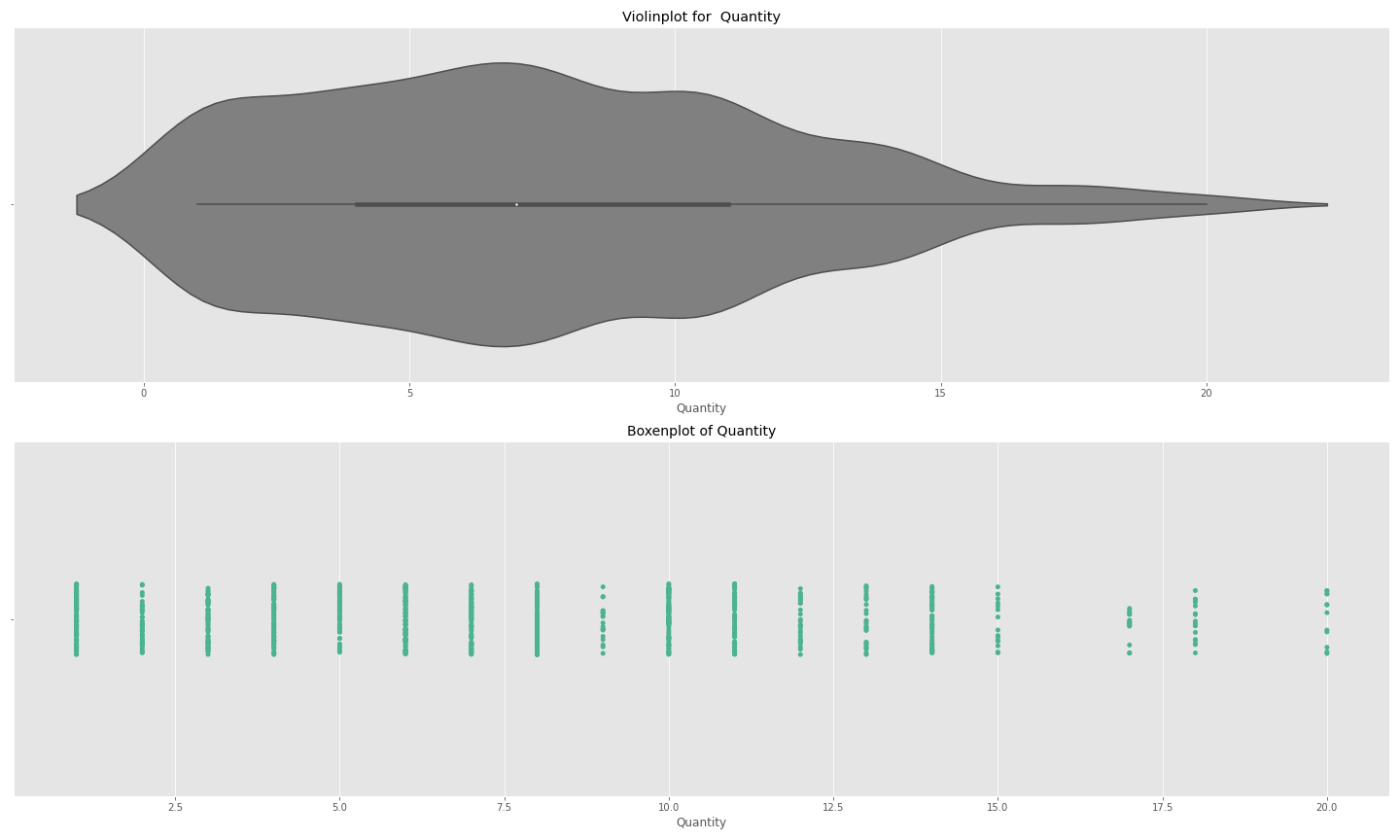


##### *Quantity purchased*

Insights-

* The distribution is relatively a flatter one as the Kurtosis seems to be negative. The data points in range 1-13 are most dense.
* Mean and peak coincide at 7, meaning 7 is the most frequent order size.





##### *Alternate Charts -*

For Quantitative Univariate analysis there is always a range of options to choose from - Box Plot, Histograms, Swarmplots, KDE, Distplot etc.

## 3. Univariate Analysis ( Categorical ) *( 25 Minutes )*

The Univariate analysis of qualitative variables is fundamentally different from quantitative variables.

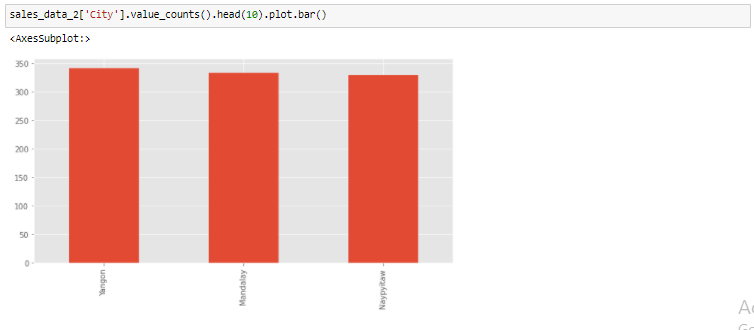
While for quantitative charts, they are more statistical in nature as we have seen in last section, for categorical analysis we simply plot the frequencies via bar/Pie chart.

The purpose is to understand the bias in the data with respect to various categories-

##### *Across City*

Insights -

All cities have almost equal number of transactions. In other words, the footprint of the supermarket is similar across cities.

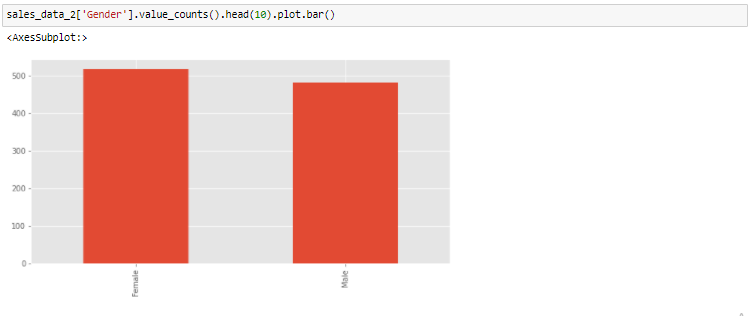


##### *Across Gender*

Insights -

Females transact more as compared to males. There can be multiple reasons for this which can be a part of further analysis -

* Is supermarket known for female oriented products?
* Does region in general have more females than males ?
* It might be a surprise insight for the supermarket and a particular factor might be causing it - Store aesthetics, offers on a select catalogue etc.



##### *Across Product Line*

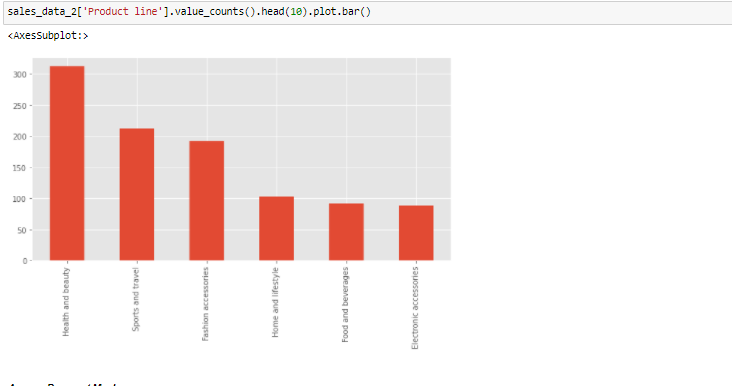
Insights -

There are 2 types of product lines because there is a huge gap in # transaction among top 3 and bottom 3 product lines.

1. Health & Beauty , Sports , Fashion constitute majority of the transactions and are high selling product lines for the supermarket.
2. The other 3 product lines are not selling as much.
3. Health and Beauty is top selling product line with a healthy margin.

The possible reasons for this might be following -

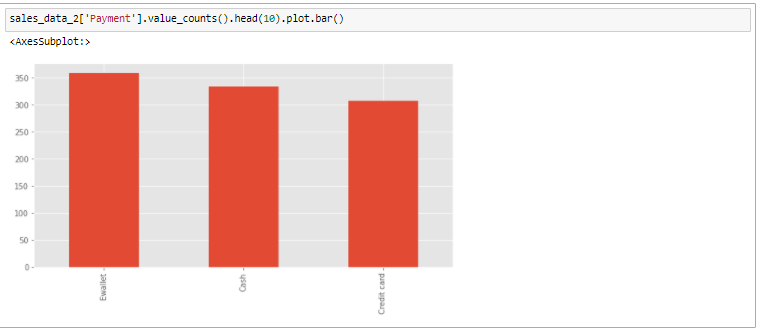
* There is differentiation on focus by supermarket - there are more products on sale from top 3 as compared to bottom 3
* There are special offers on top 3 which are driving the sales.
* The prices and quality for top 3 are more competitive than the rest.
* The management is more focused and concerned with the growth of top 3 rather than others ( Possibly due to low margins )



##### *Across Payment Modes*

Insights -

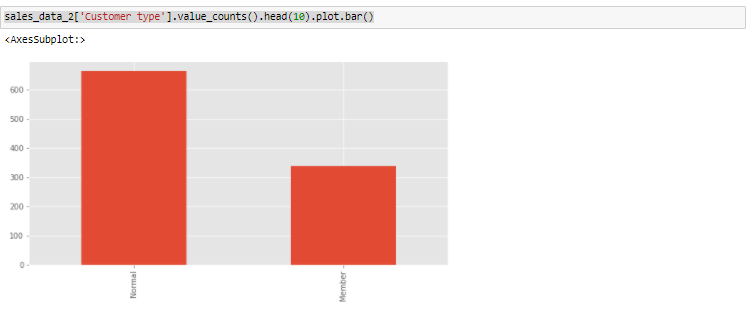
* E Wallet is most preferred payment mode while Credit card is least.



##### *Across Customer Type*

Insights -

2/3rd of the transactions are done by Normal Members and only 1/3rd by Members



##### *Unique Customers who are Member/Non-Members*

While we have already looked at the number of transactions, it is also worthwhile to look at the number of Member/Normal customers that the Supermarket has.

There is a subtle difference between number of transactions and number of customers, which is to be noted.

The count of distinct Member/Non Member customers can be obtained by grouping the data by Customer Type and then using nunique() function –



Insights -

There are about 250 Members and 375 Normal customers with the Supermarket.

It can be seen that the ratio of Members is more than the Transactions done by Members.

It is essentially telling us that Members are not contributing to more transactions as would be expected from those.

##### *Alternate Charts -*

Some other ways to look at categorical distributions are - Pie Chart, Countplot, Donut Chart etc.

### Stores Location on map

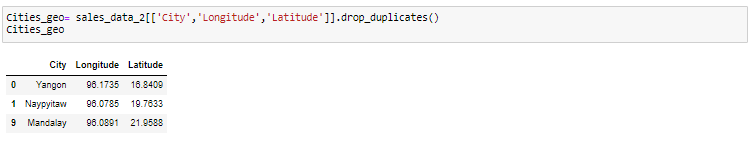
Plotting on map is being shown for the first time. Extra attention would be required there.

Folium is one of the libraries which provides us with the ability of plotting Geological charts. It has already been installed using *import folium* in the initial library set up steps earlier in the notebook.

We will use the same here to understand on the map where exactly the 3 stores are located and how is the proximity like.

First, we'll need to create a dataframe with all the cities, along with their coordinates.

Drop\_duplicates can be used here to get the unique rows -



In the next step, the map can be plotted using folium.map(). The parameters are passed accordingly.

Some basic idea about the parameters are as following ( Use Shift + tab for more details)–

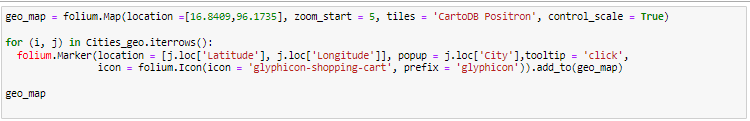
*Location* – latitude and longitude of the marker/map

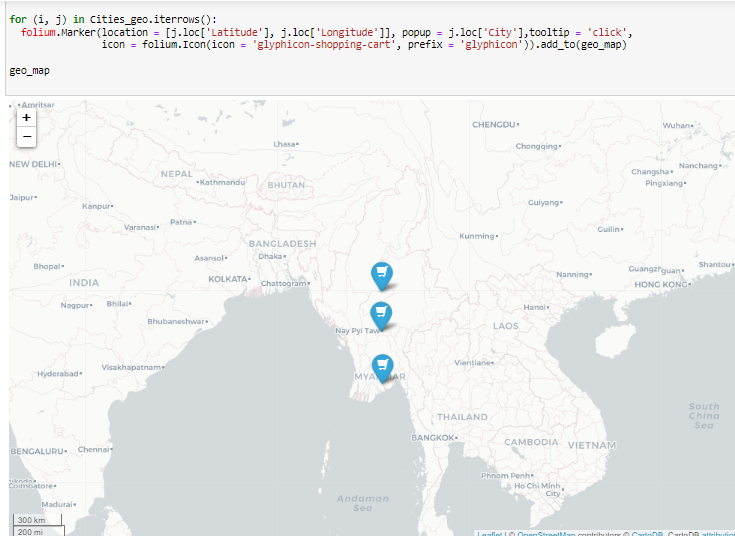
*Popup* – Text to be displayed for the marker

*Icon* – Icon style for marker

*Prefix* – source of the icon

folium.Map() just plots the chart. We also need to put city markers on the map. This done by putting markers one by one using folium.marker() function –





Insights -

The cities are distributed across lengths of Myanmar and are not close to each other.

Yangon is the only coastal city where the chain is present, other ones are located in the middle of Myanmar.

Company can expand in North and East of Myanmar as there is no presence in those areas.

Also, a big coastal city Sittwe is also an option for later expansion.

## 4. Bivariate Analysis *( 30 Minutes )*

It is one of the simplest form of statistical analysis where 2 variables are involved. It looks for relationship among the 2 variables. The applications involve hypothesis validation of association among variables, finding trends, regression etc.

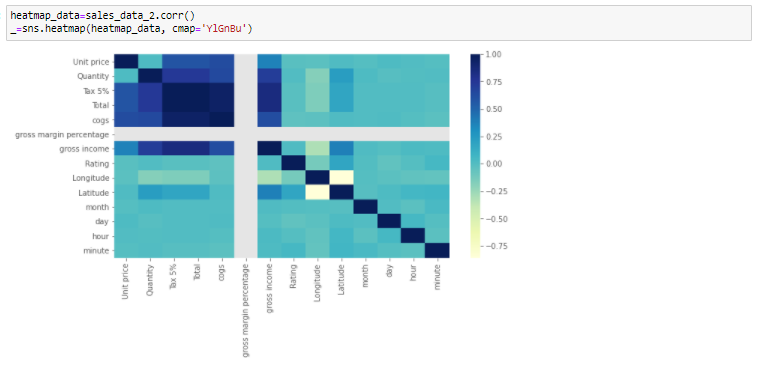
### Heatmap

The purpose of Heatmap here is to look for any significant correlations among continuous variables in our data.

Since majority of the quantitative variables in this sales data are mathematically related to each other, the use of Heatmap in this case is not of much help.

For example following related variables -

* Invoice amount = Unit Price \* Quantity
* COGS = Invoice Amount \* 0.0476
* Gross Income = Invoice Amount - COGS



As expected, we see high correlation in almost all sales related fields because they are derived from each other.

##### *Function for bivariate analysis -*

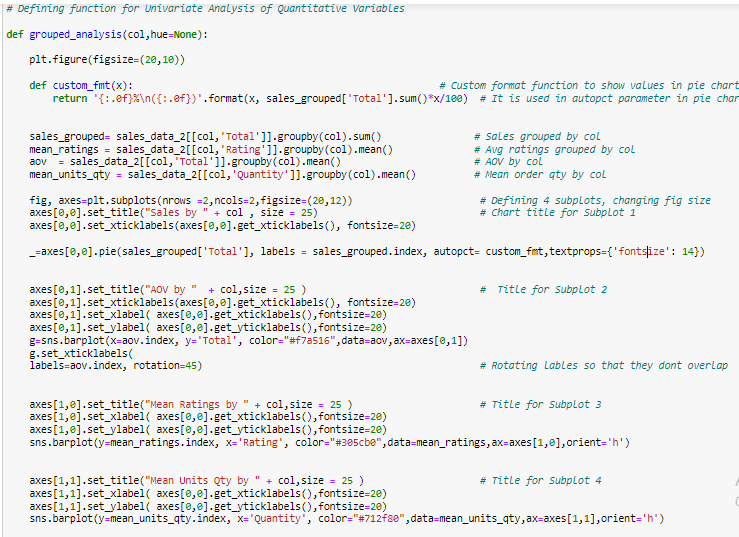
There are 4 important metrics that can provide maximum insights in this sales data -

1. Sales
2. AOV - Average Order Value
3. Mean Order Quantity
4. Ratings

It would be worthwhile to plot these against multiple categorical variables to see if those categories have an impact on these 4 metrics.

It can be a great way to find out what's working for the chain and what is not.

( Functionality of Groupby – It arranges the data logically into identical groups based on the selected columns, so that the aggregate functions like min, max, sum , avg can be performed on top of those. )



Functionality of autopct & Custom\_fmt function :

autopct parameter in Pie Chart enables us to display section %s and values using string formatting functions. For example, a value of ‘2.2f%’ means that for each pie wedge the % value will be formatted as per ‘2.2f%’.

A callable function can also be used as per above case to pass on modified or multiple lines of calculated strings. Try changing the format to {1.1f%} to get a better idea about how the expression works.

Choice of charts –

We have 4 charts in this subplot matrix of 2X2. Let’s talk about all the chart selections 1 by 1 –

**Top Left ( Subplot 1 ) :** Since the purpose is to show how much portion a category is contributing to the entire sales, a pie chart showing %s of each would make good sense. It will be easier to infer which categories are dominant and which ones are lagging,. The % values and absolute values are displayed using autopct function to provide more information.

**Top Right ( Subplot 2 ) :** A comparison of AOVs for different categories can be done using a bar chart. Averaged metrics are generally compared via a bar chart unless being represented on a timeline.

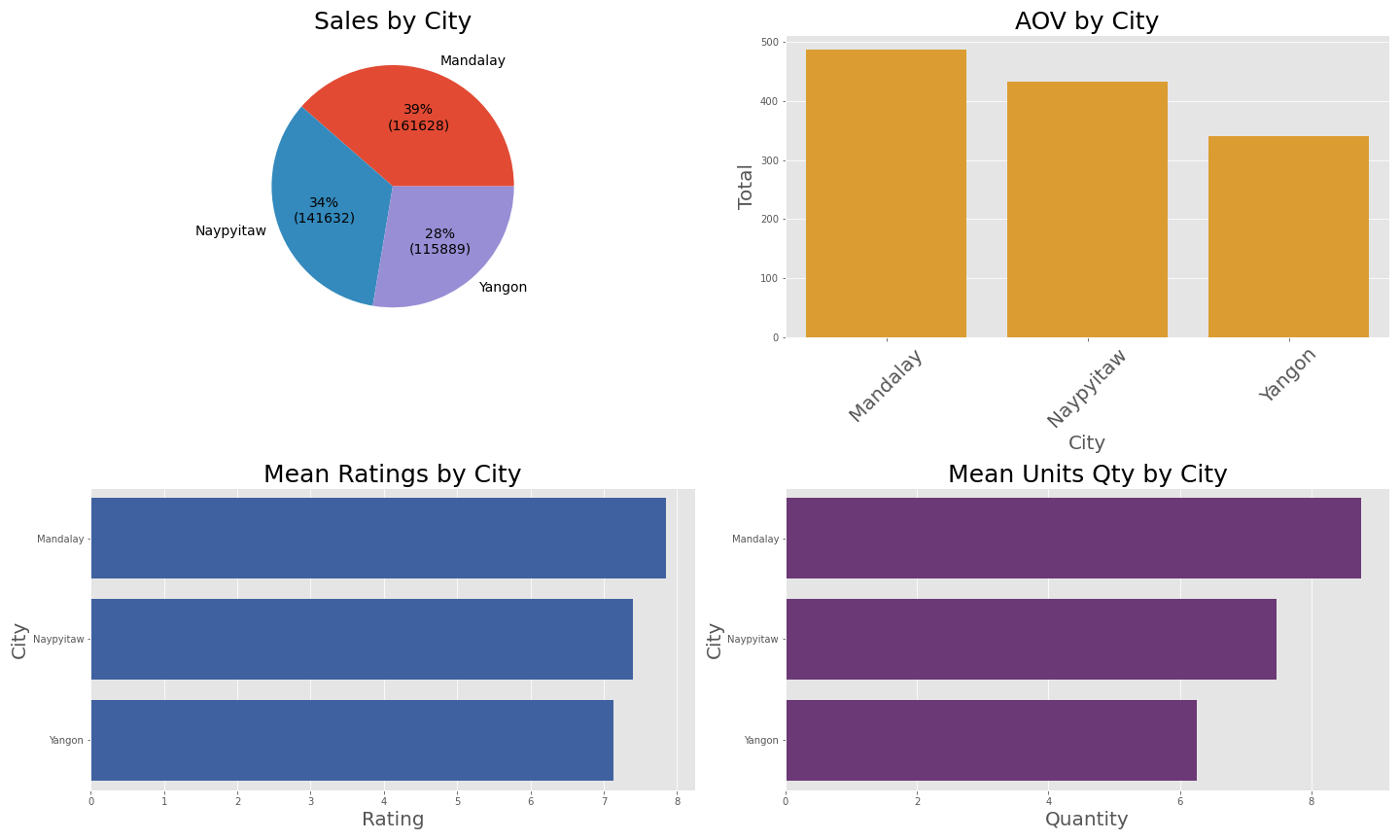
**Subplot 3 & 4 :** Similar to AOV, mean ratings are mean quantities are also averaged metrics, and hence represented via bar chart. Unique colors have been assigned to these metrics so that these are spotted easily among different plots for different categories.

#### Across City

Insights -

* Although the number of transactions were similar as we saw earlier in Univariate analysis, Mandalay city have significantly more sales at 39%
* Yangon sales are least at 28%
* The difference in sales is essentially driven by the differences in AOV, which is maximum for Mandalay and really low for Yangon
* Further, we can say that the higher AOV is driven by higher mean quantity as seen in the Right bottom chart
* Ratings also reflect a similar story, with highest ratings for Mandalay and least for Yangon.
* All of above points point to a big scope of improvement in Yangon for the supermarket chain.

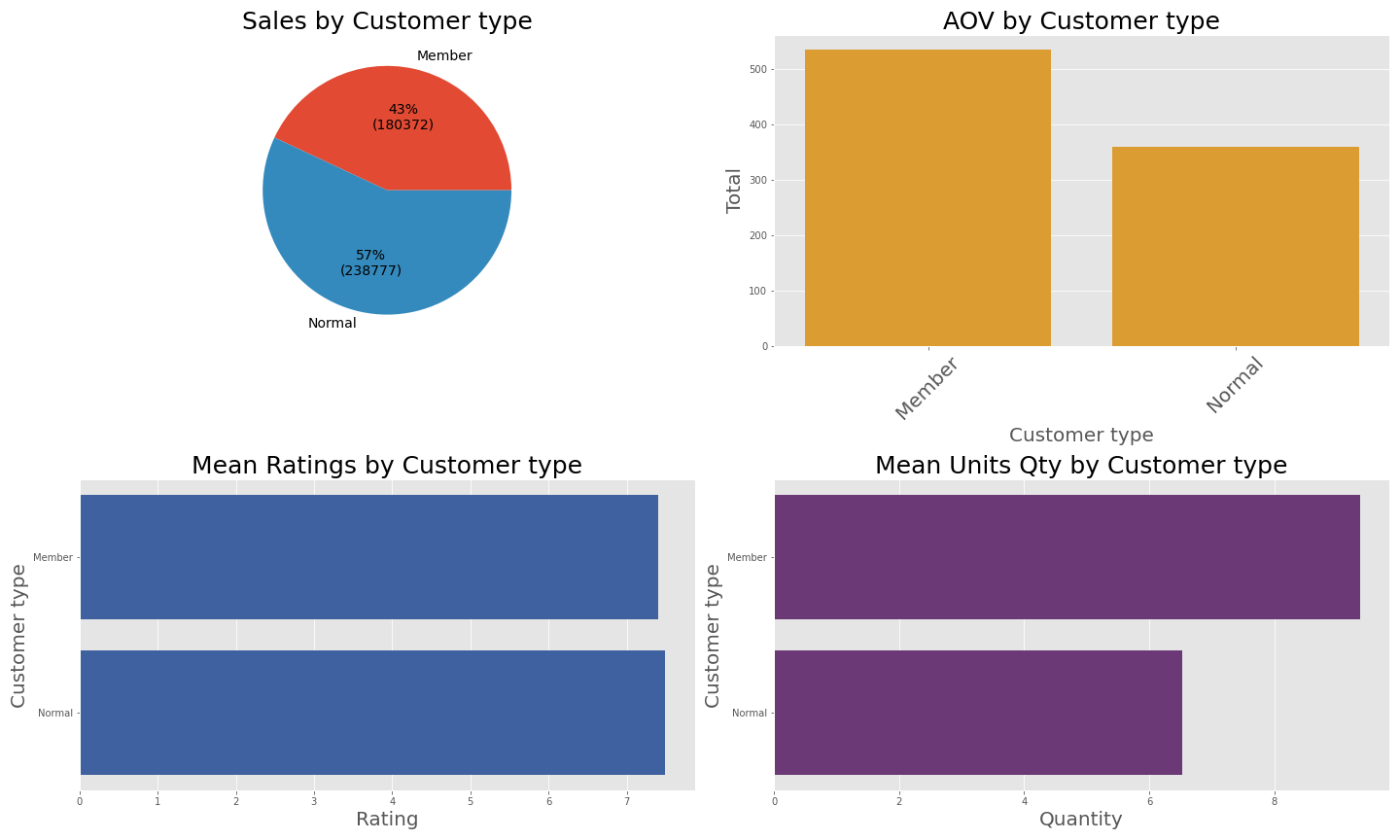




#### Across Customer Type

Insights -

* Members are contributing to 43% of sales, driven by really high AOV of 500 a compared to 300 of a normal customer
* The difference in AOV is driven by the difference in Average order quantity
* Ratings are similar by Member and Normal customers.

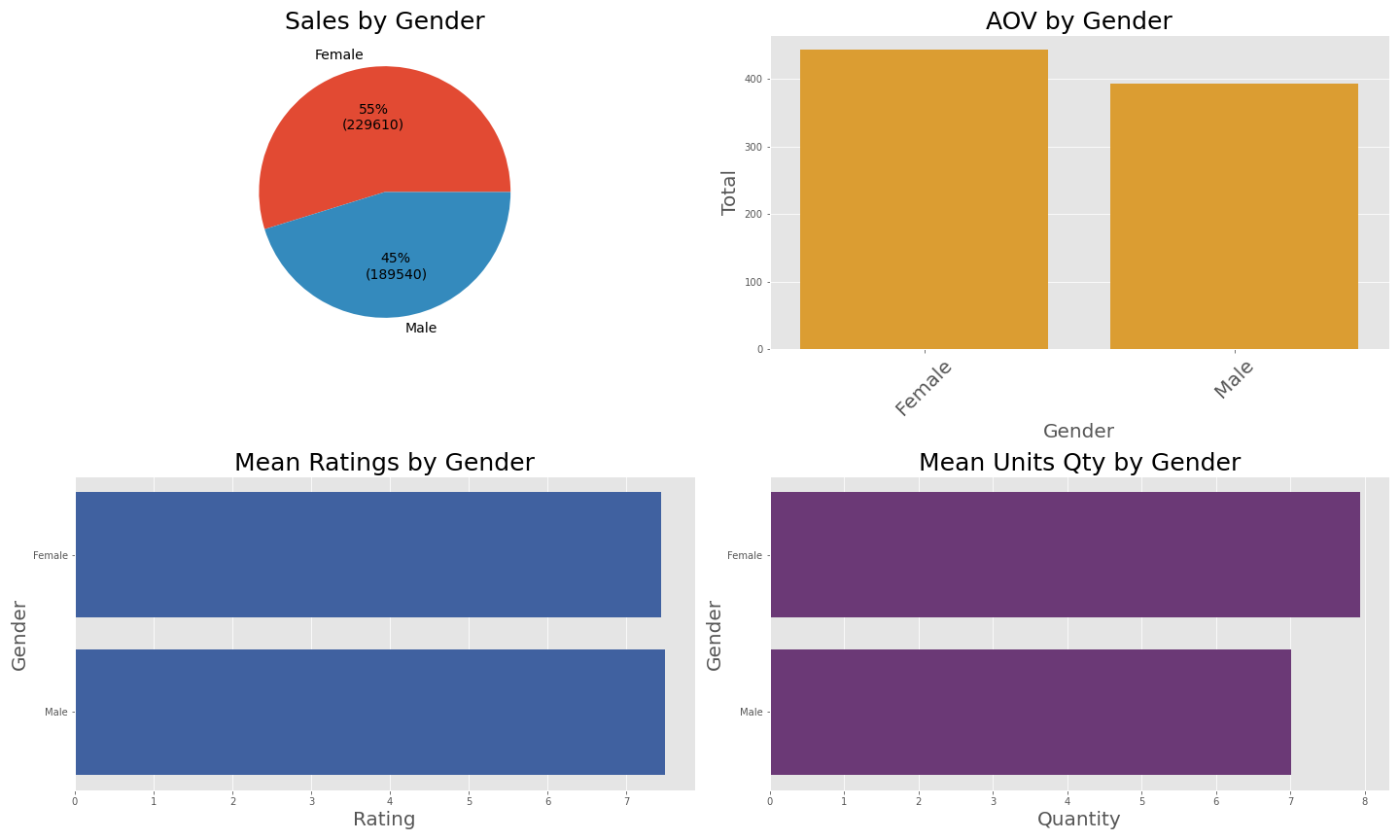
 

#### Across Gender

Insights -

* As expected after Univariate analysis, females are driving more sales at 55%
* The AOV and average order quantity are also significantly higher for females
* The ratings are not impacted by Gender





#### Across Product Line

Insights -

* Health and Beauty is the most prominent segment with 30% of sales.
* 3 bottom segments - Food & Beverages, Electronics, and Home & lifestyle together make 30% of the sales.
* Food & Beverages have the maximum AOV among all categories, Home and lifestyle also has a relatively higher AOV.
* Customers love these 2 segments as the ratings are also highest for these 2 segments.
* If the company could drive more transactions in above 2 segments, it will boost the sales significantly due to their higher AOV



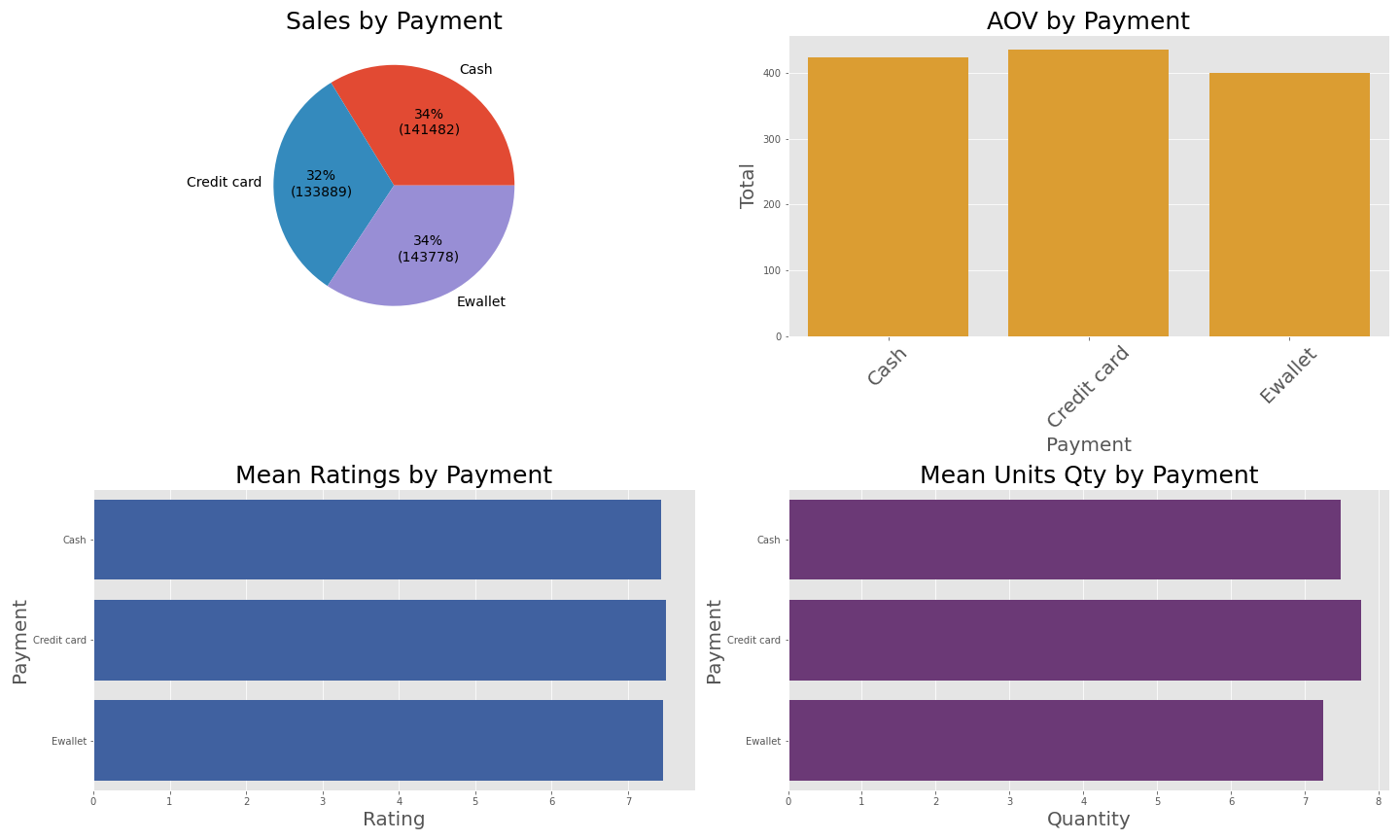


#### Across Payment Modes

Insights -

* Payment mode doesn't seem to have an impact in any metric as modes have equal distribution in sales, ratings etc.
* AOV for Credit cards is relatively higher but the difference is not significant for it to prove a hypothesis.





## 5. Timeseries Analysis ( 25 Minutes )

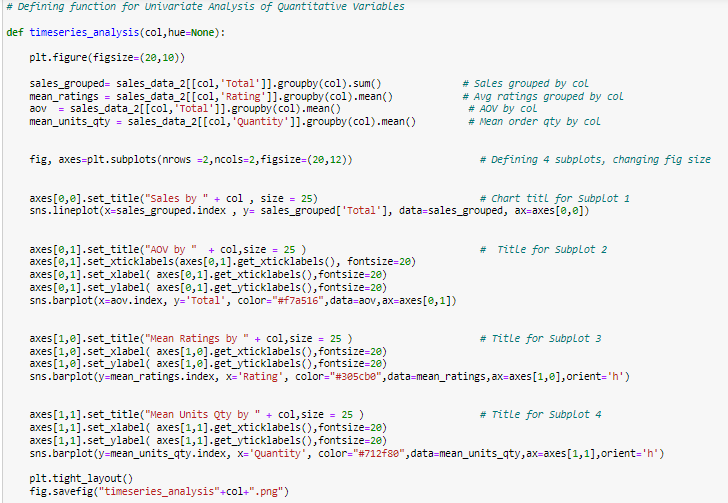
It’s a new concept being shown for the first time. Extra attention would be required there.

Timeseries analysis is when a particular metric (sales/aov etc) is analysed with respect to a time variable ( day/month/year etc)

Although timeseries analysis is always beneficial wherever applicable, but in case of Sales related data it becomes even more valuable. This can be attributed to following reasons -

* Cyclicity in sales can be visualised
* Seasonal sales patterns can be identified
* Impact of various events ( Festivals, Market collapse, Natural disasters etc. ) can be identified
* New strategies can be devised based on the insights to lift up sales during specific times.
* Ideal promotional times can be identified.

Here, we will be plotting the same 4 metrics as in last sections against a time variable for our analysis -

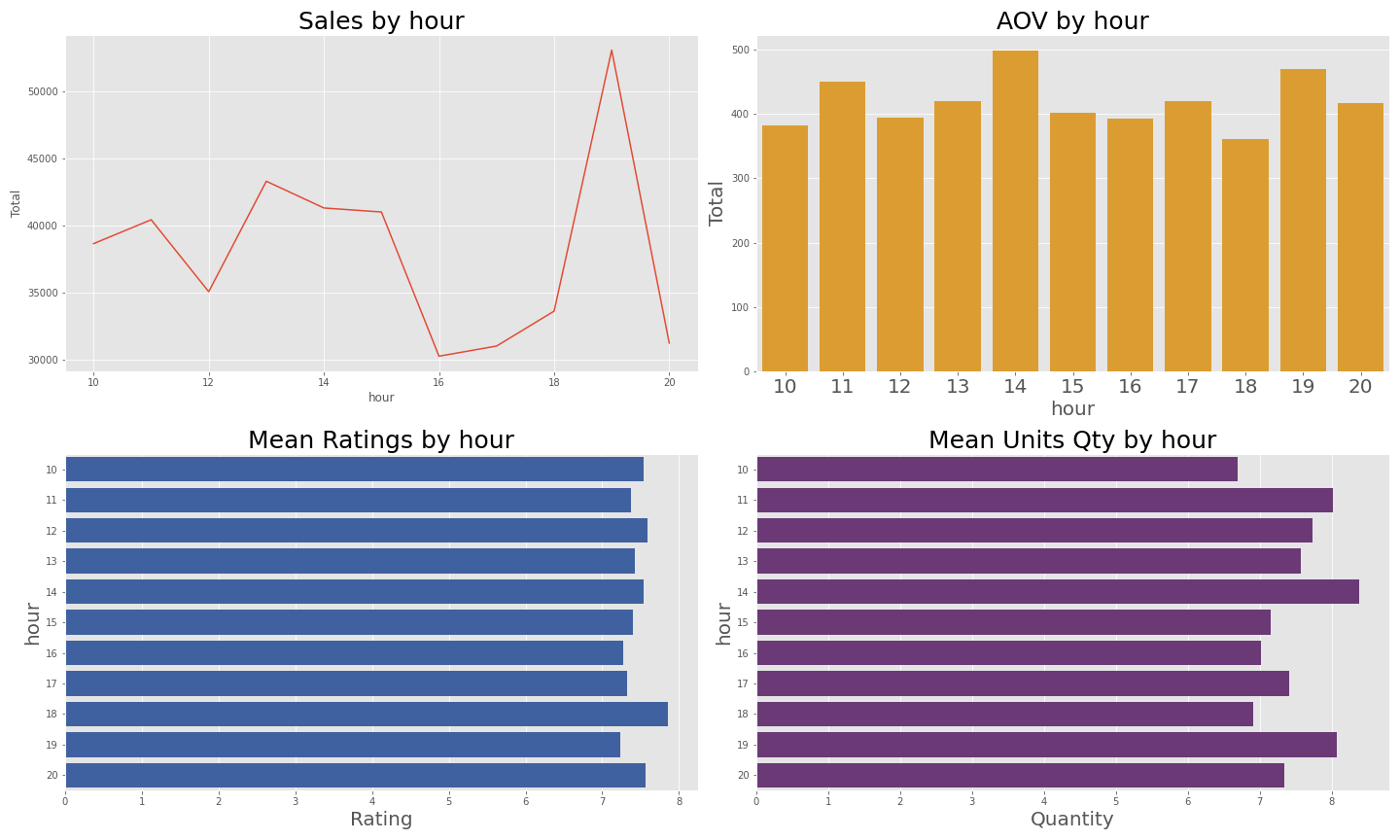


##### *Timeseries Analysis - by Hour*

Insights -

* Sales peak sharply at 7PM in the evening
* 4PM-6PM is a dead period with really low sales
* Stores see moderate sales during 10 AM to 3 PM
* Ratings, Order quantity logically should not be related to time of the day, hence any variation should be attributed to a random phenomenon.



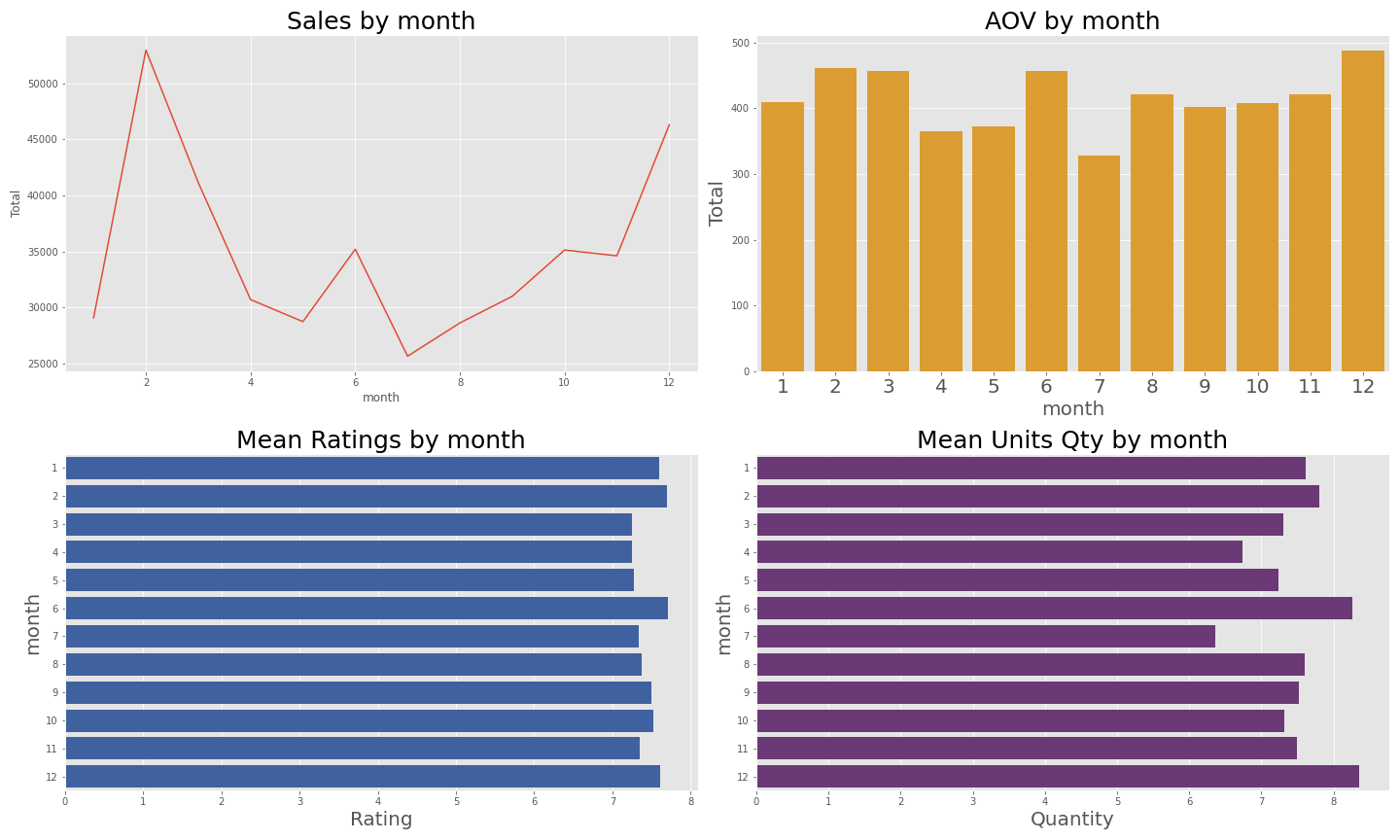


##### *Timeseries Analysis - by Month*

Insights -

* February is peak sales month where sales go beyond 50k+. Jut for comparison, most of the other month accumulate less than 35k revenue.
* There must be a reason for this spike in February. It can either be a festival in the region, or a season that's driving sales of particular segment or a promotional campaign that has worked well.
* December also sees a significant spike, although not as steep as February but still significant.
* April-Nov sales are on the lower side, with sales dipping close to 25k in July.

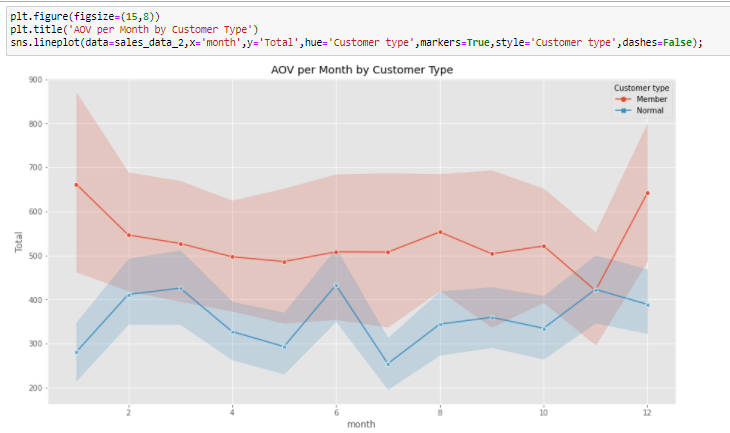




##### *Month on Month sales - Members and Non Members*

Insights -

* Members always have significantly more AOV than normal customers.
* Members AOV has lesser fluctuations than Non members.



## 6. Customers across Cities ( 5 Minutes )

Insights -

* Mandalay has least unique customers (90), while Yangon hs most (200+)
* Sales per customers metric has opposite trend where each customer spends maximum in Mandalay ( 1600).

NOTE : nunique function is used to calculate distinct customers

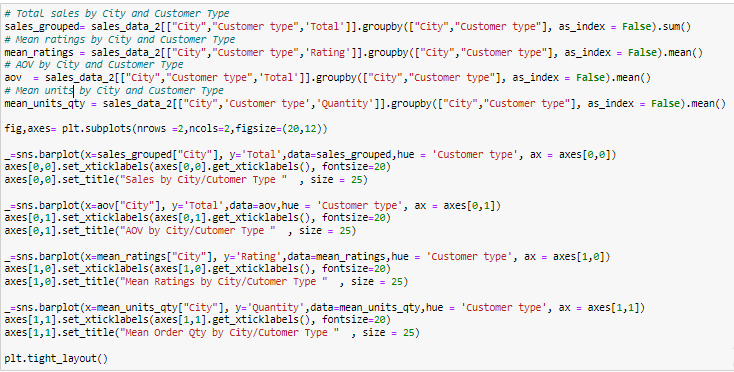


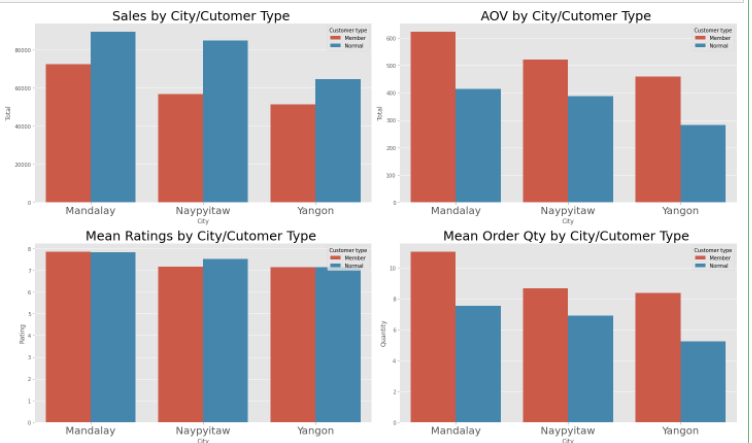


## 7. Member/Non-Member Deep Dive ( 5 minutes )

Finally, we'll see how our Member and Non Members perform across 4 metrics that we've used for bivariate analysis.

We'll make use of Grouped bar chart to represent the comparison in a better way.





Insights -

* Normal customers have more sales in all cities. But that may be due to the fact that Non Members are higher in count as we've seen earlier.
* Members consistently have higher AOV than non members.
* The AOV is driven by higher order quantity.

## 7. Final Insights ( 5 minutes )

As per the insights discussed so far, we can have following final recommendations -

* The Members are doing really well in terms of sales numbers for the company. Thus a campaign can be formulated to promote the member sign ups.
* The spike in Feb sales are a hint that the stores have the potential to have higher sales throughout the year. Specially April to November period sales can be significantly improved by some focused marketing.
* Food & Beverages, Home and lifestyle have the maximum AOV among all categories, but constitute just 20% of the total sales together. The ratings are also better than other categories. It is a sign that a strategy around promoting these categories can be formulated.
* It is a good sign for the supermarket chain that Females are driving majority of the sales. Even females' dominated categories such as Health & Beauty, Fashion are top product categories. The company can scale this further to improve overall revenue. Female tend to be higher spenders than Men, and company can take more advantage of this data insight.
* Yangon has least sales and AOV of all cities and the difference is significant too. Hence a focused marketing campaign can be launched to meet the potential sale in this city.