REPORT

Data:
This data is about the transaction details of a cafe.
Containing i)transaction id
ii)Item the customer purchased
iii)Quantity which refers to no of items he/she purchased
iv)total spent
v)Payment method
vi)location
vii)Transaction data
viii) price per unit

Data Cleaning:
1)There are missing values and noisy values in the data.
The noisy values are ‘UNKNOWN’ and ‘ERROR’.i replaced these noisy values with nan
2) Syntax of inputs:
I checked is there any entry in transaction id is not in the format TXN_*******
There is no entry which doesn't following the given pattern
I checked is there any entry in transaction id is not in the format year-month-date
There is no entry which doesn't following the given pattern
3)filling nans:

- ➔ I filled an Item by its price per unit if there is a unique item present corresponding to that price.
- ➔ I filled price if i know item
- ➔ I filled Quantity if i know both price and total spent
- ➔ I filled total spent if i know both quantity and price
- ➔ I observed the proportions of the payment type of each item in it. all the shows same proportions 33:33:33 approximately So I filled in the nans with proportions.
- ➔ I observed the proportions of the location type of each item in it. all the shows same proportions 52:48 approximately So I filled in the nans with proportions.

- ➔ I filled nans in the Transaction date by forward fill assuming that there is more probability to enter at the same date of its neighbouring elements.

4)checked outliers in each numerical column
I found there is an outliers in total spent i removed that row
5)i removed rows still containing nans
6)i remove unnecessary columns like Transaction id
7)I added day and month columns and removed the transaction date.

I moved the cleaned data into refined data.csv

Data analysis:
a)univariate analysis:

- ➔ Quantity is not skewed to much to right or left
- ➔ Price is not skewed to much to right or left
- ➔ Total spent is skewed to right
- ➔ The total spent are independent on month all are approximately 8.5
- ➔ The total spent are independent on day all are approximately 14
- ➔ Type of location have same probability
- ➔ Type of payment has the same probability to choose.

b)bivariate analysis:

- ➔ Quantity and total spent are positively correlated
- ➔ Quantity and price per unit are not strong correlated approximately -0.05
- ➔ Total spent and price per unit negatively correlated
- ➔ Coffee and juice are highly sold and salad is the least sold over the year (but the difference is approximately 1/5th of max-min
- ➔ We get most total spent on sandwich and least on cookies
- ➔ Total spent is independent on day and month

c) multivariate analysis:

- ➔ In july there is maximum of total spent by customer and quantity
- ➔ Total number of customer visits is more in friday in july
- ➔ Total number of visits less in wednesday april
- ➔ For salad if the customer takes more quantity then he is more likely to pay by digital wallet . remaining items have equal priority in payment type