Introduction

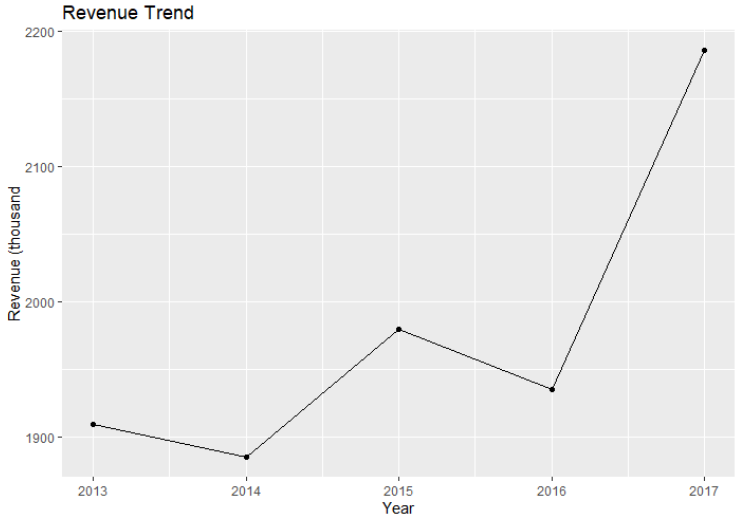
This report provides an analysis of sales data collected from the COMP2031-8031 database using MongoDB. The report covers various aspects of data collection, data wrangling, and data analysis to gain insights into the sales performance of a company. The content includes revenue analysis, store's revenue comparison, revenue by purchase method, sales quantity, coupon usage, customer demographics, and customer satisfaction.

Data transformation and analysis

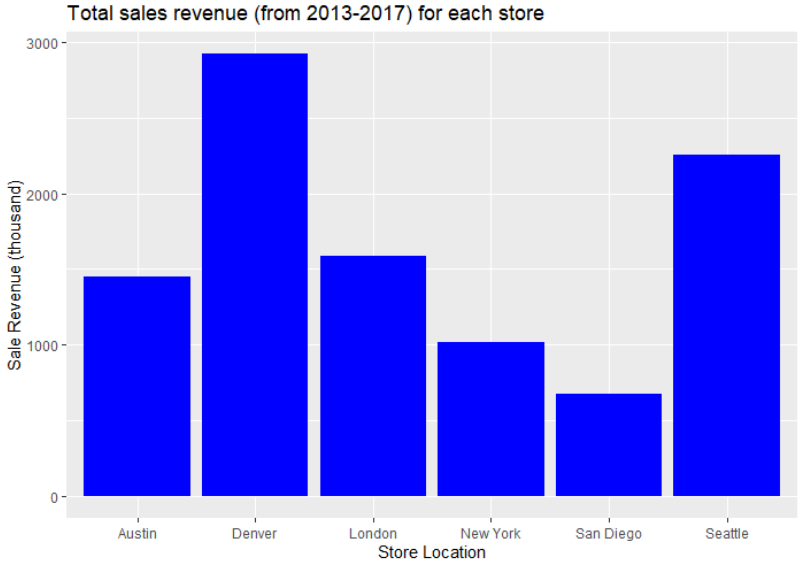
Revenue analysis

a. Total revenue of sale per year (from 2013 – 2017) for the whole supply company

year	totalRevenue
2013	1908.918
2014	1885.110
2015	1979.871
2016	1934.820
2017	2185.853



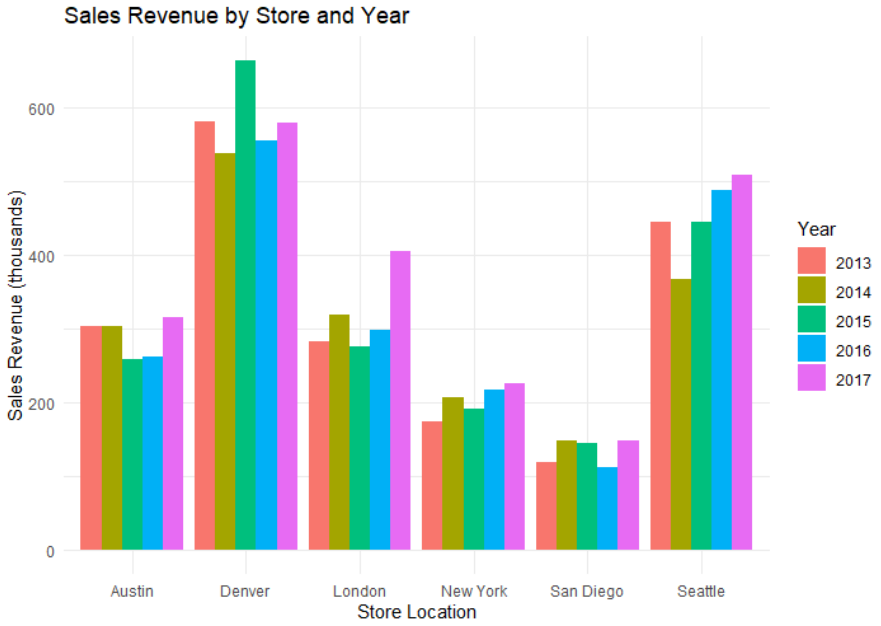
b. Total revenue of sale per year (from 2013 – 2017) for each store



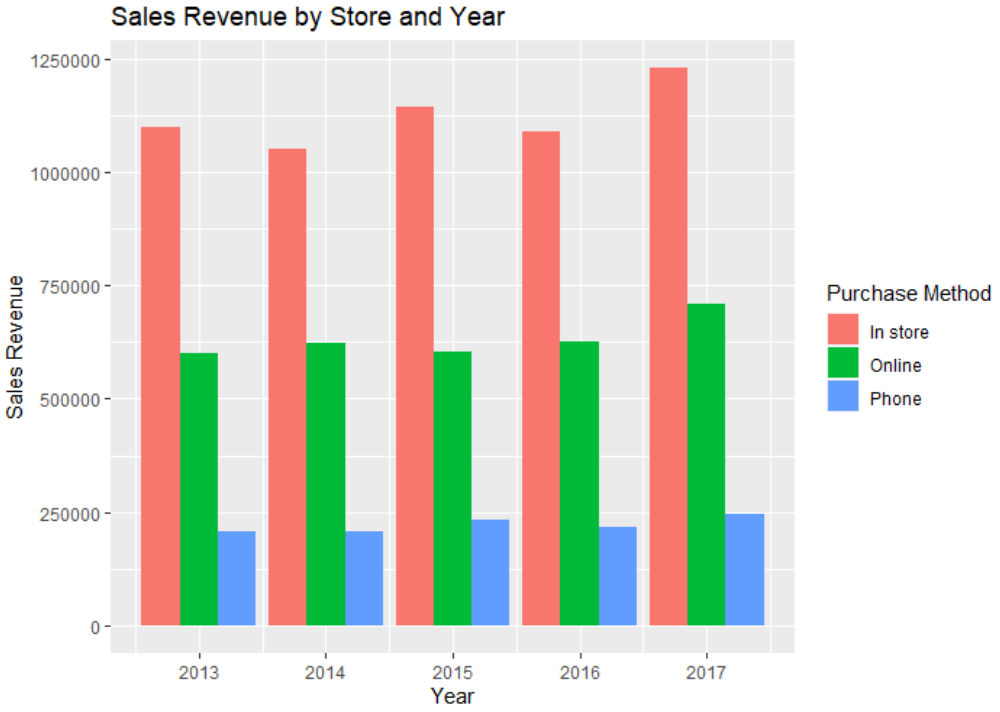
storeLocation	totalRevenue
Denver	2921009.9
Seattle	2255947.7
London	1583066.8
Austin	1445603.1
New York	1016059.6
San Diego	672885.2

c. Sales revenue for each store in each year from 2013 – 2017

storeLocation	2013	2014	2015	2016	2017
Austin	304115.0	304409.2	258664.3	262800.3	315614.3
Denver	582295.8	537944.2	664211.4	556312.8	580245.6
London	283522.0	319280.4	275396.8	298542.3	406325.4
New York	174068.6	207745.7	191614.8	216853.8	225776.7
San Diego	118973.6	148072.0	145262.9	111719.0	148857.7
Seattle	445943.0	367658.7	444721.1	488591.6	509033.2

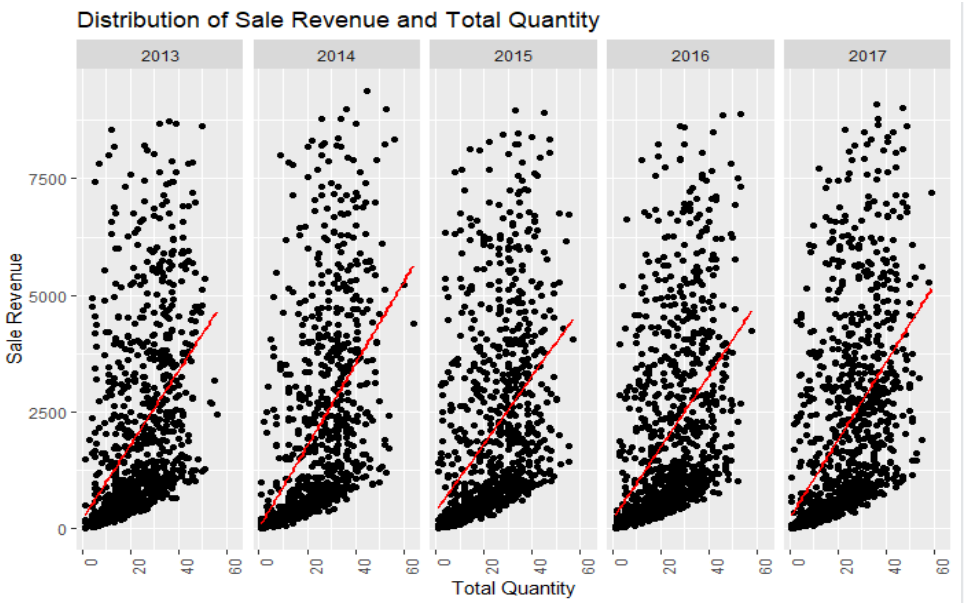


d. Revenue and purchase method

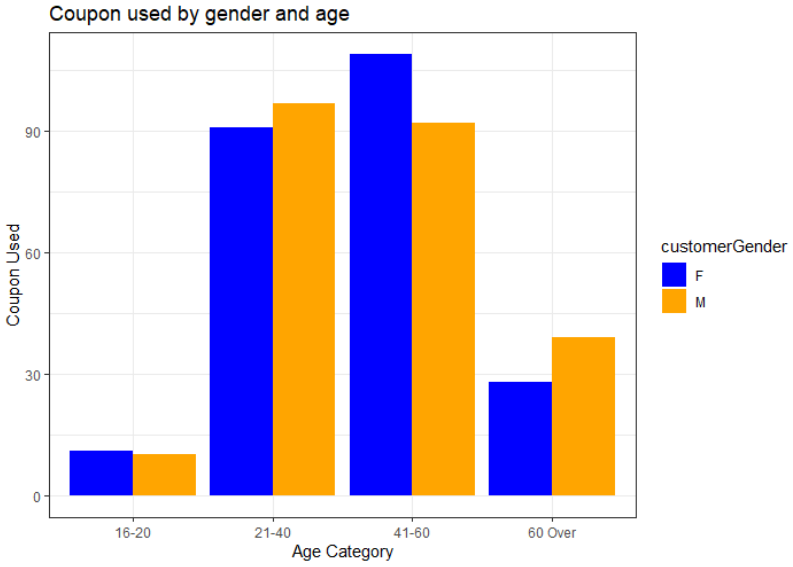


year	In store	Online	Phone
2013	1100231	600027.8	208659.3
2014	1052164	624622.7	208323.2
2015	1145562	602504.0	231805.4
2016	1091207	627707.3	215905.9
2017	1229975	708952.9	246925.1

e. Identify the distribution related to sales and inventory management

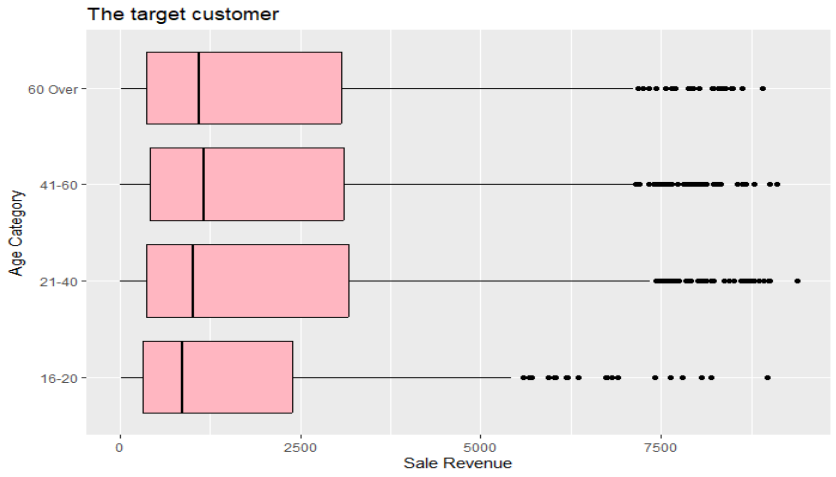


Coupon used, compared by gender and age



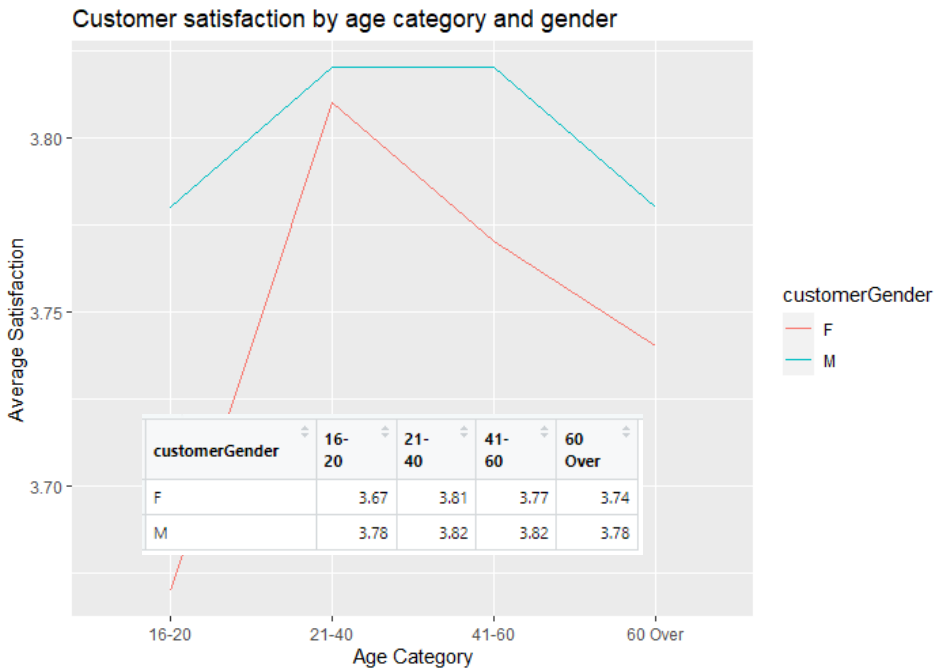
customerGender	Age_16-20	Age_21-40	Age_41-60	Age_60 Over
F	11	91	109	28
M	10	97	92	39

Identify the target customer with the Age category



Customer satisfaction

a. Customer satisfaction by gender and age category



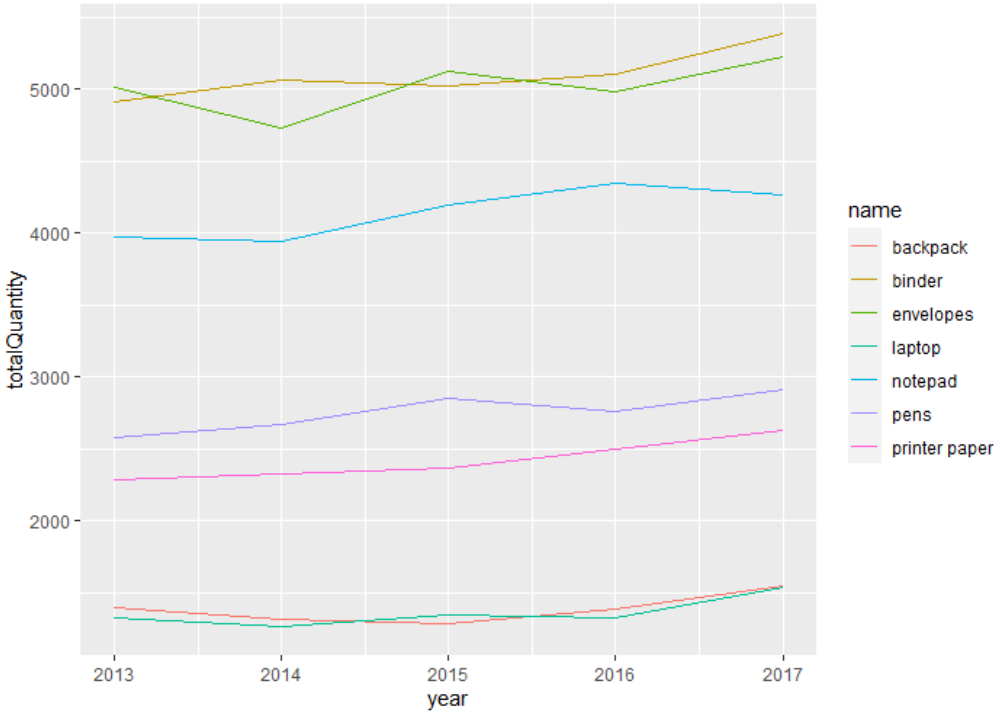
customerGender	16-20	21-40	41-60	60 Over
F	3.67	3.81	3.77	3.74
M	3.78	3.82	3.82	3.78

b. Mean customer satisfaction per year (whole company)

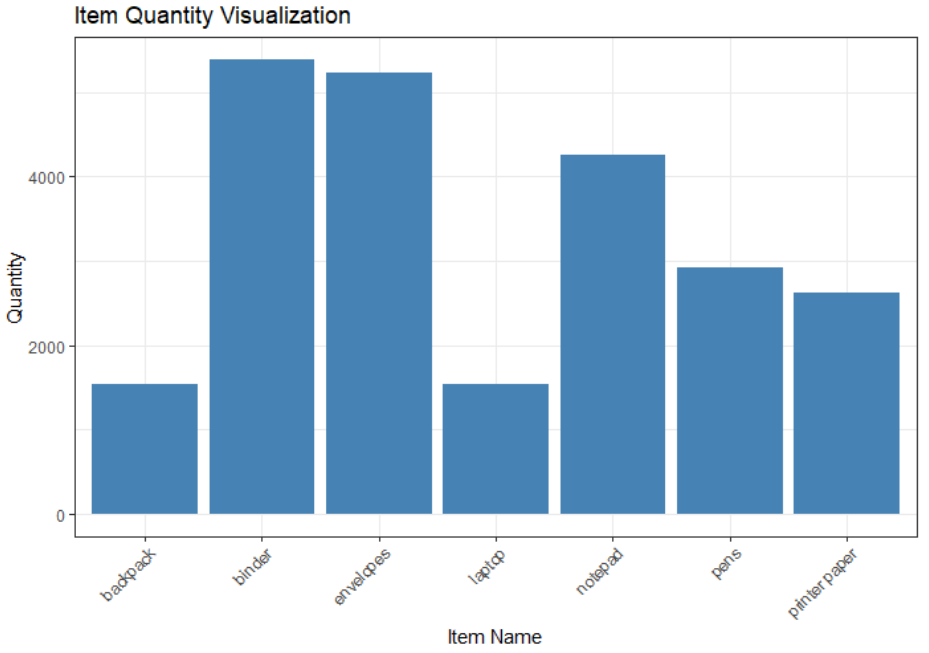
year	meanSatisfaction
2013	3.776993
2014	3.840595
2015	3.782306
2016	3.826733
2017	3.744297

Quantity sale of items

a. Quantity of Items sold each year



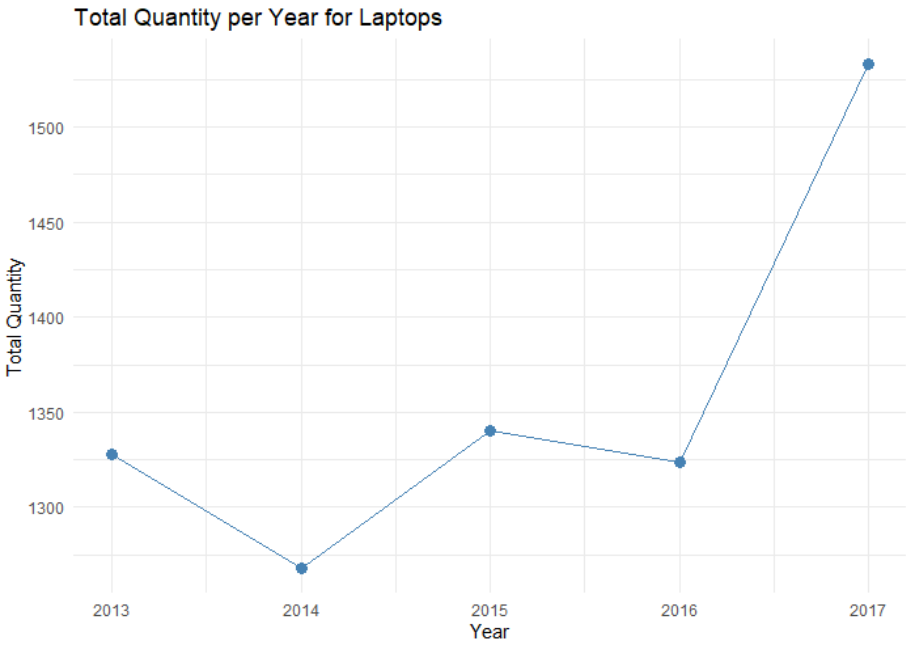
b. Total quantity of items sold in 2017



name	totalQuantity
backpack	1545
binder	5386
envelopes	5231
laptop	1533
notepad	4262
pens	2916
printer paper	2630

c. Laptop sales quantity per year

year	totalQuantity
2013	1328
2014	1268
2015	1340
2016	1324
2017	1533

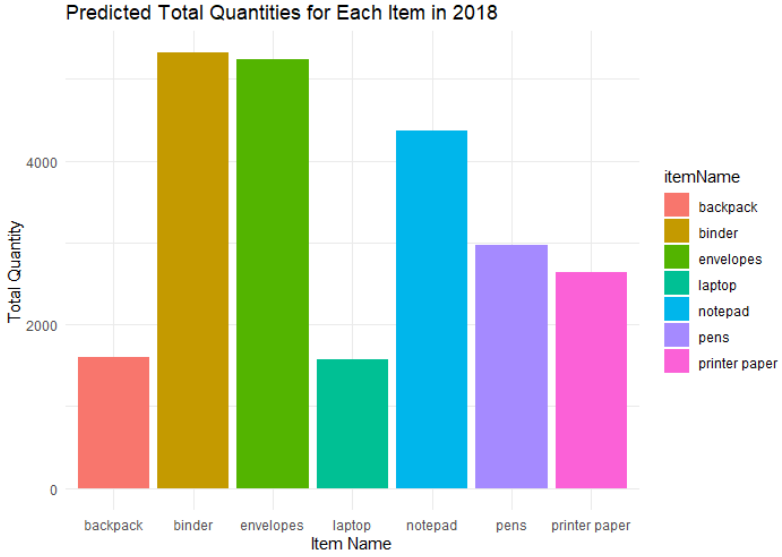


Data modelling

Linear regression

a. Linear regression modelling predicts total quantity sold for each item in the next year

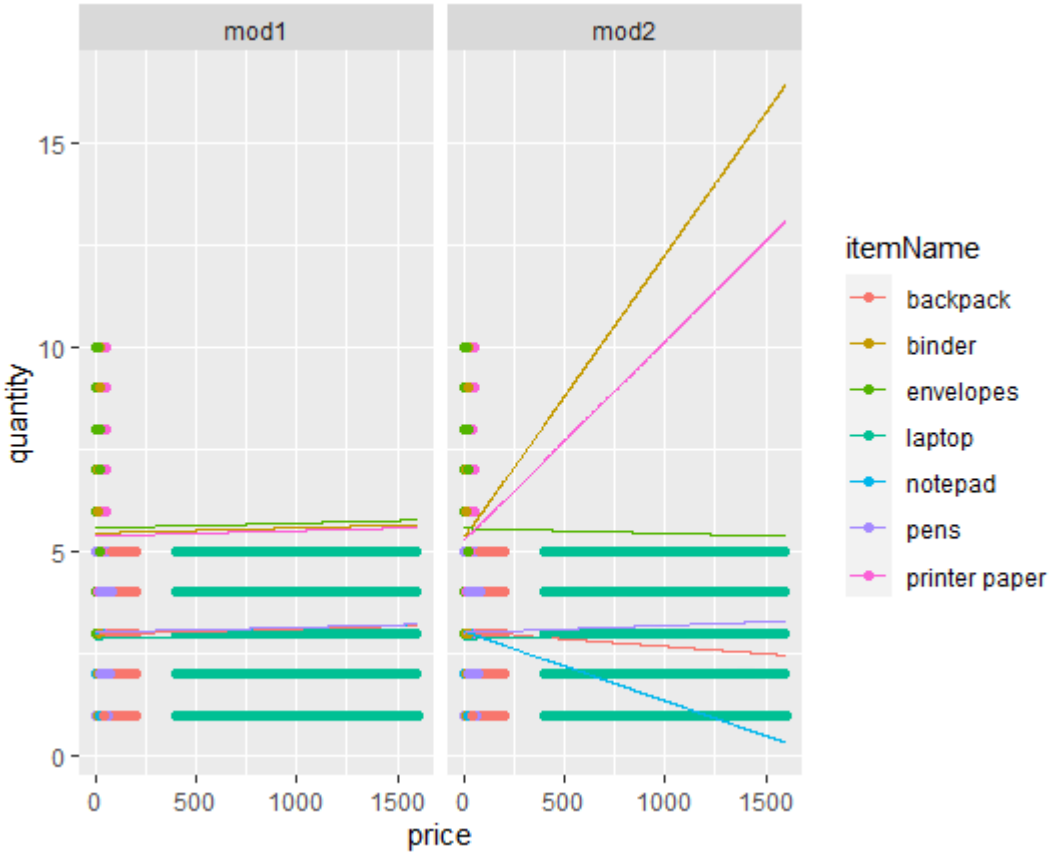
itemName	year	total_quantity
backpack	2018	1604.657
binder	2018	5319.657
envelopes	2018	5236.657
laptop	2018	1579.657
notepad	2018	4366.457
pens	2018	2974.257
printer paper	2018	2639.457



b. Linear regression model predicts quantity of items sale

The equation for the model as below,

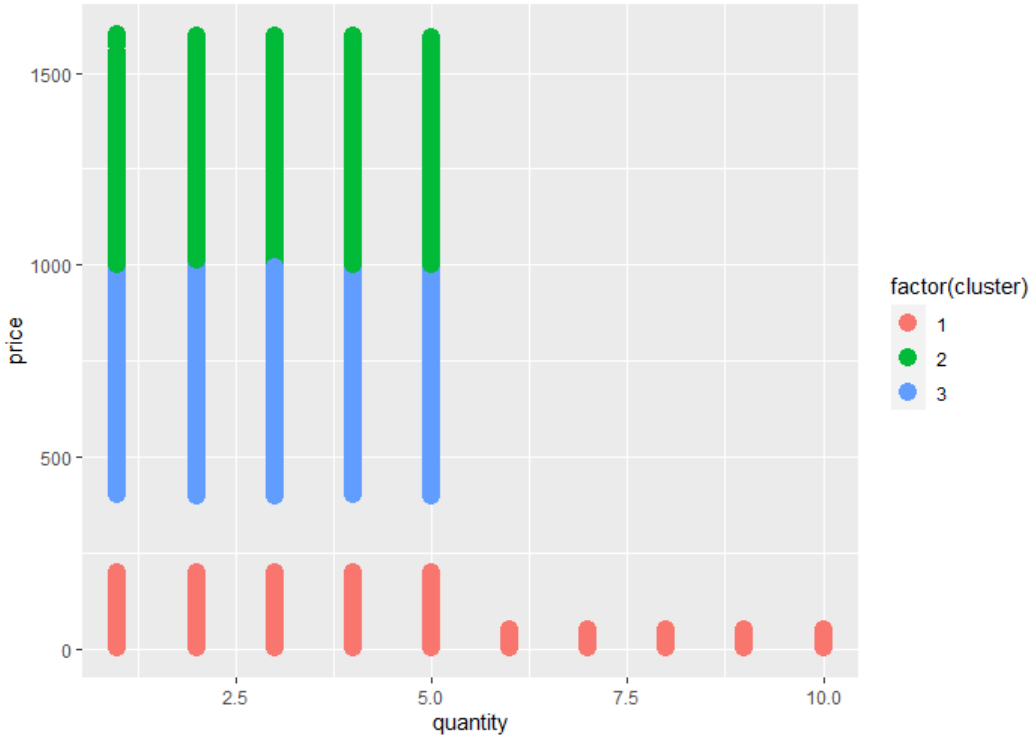
```
Coefficients:
          (Intercept)                 price       itemNamebinder    itemNameenvelopes
            2.9809171             0.0001282             2.4683965             2.5838603
       itemNamelaptop      itemNamenotepad        itemNamepens  itemNameprinter paper
           -0.1169610             0.0323595             0.0331878             2.3989944
```



Model performance, compute the RMSE:

- rmse_mod1: 2.154454
- rmse_mod2: 2.154305

K-means clustering



K-means clustering for mydf with saleRevennue and totalQuantity.