Data Science Assignment

Written By: Sujeeth Kumaravel

Contents:

Introduction:
This report presents my solution to the data science problem of analyzing the effectiveness of the promotional activity for a customer, analyzing the sales data of the competitors and recommending a marketing channel for each customer.

Data Preprocessing:

1. The column names in the original data set seem too long. They were converted to more compact forms using the following mapping:

Original Column Name	New Column Name
Customer ID	ID
Title	Title
Specialty Code	SpecCode
Specialty Description	SpecDesc
State	State
Call Attempts	CallAttempts
Calls Successfully Completed	CallsSuccessful
Emails Sent	EmailsSent
Emails Opened	EmailsOpened
Faxes Sent	Faxes
Brand 1 Sales (Company's Brand)	Brand_1
Brand 2 Sales (Competitor Brand)	Brand_2
Total Branded Market Sales	TotalBranded
Total Market (Branded + Unbranded) Sales	Total Market

2. 'ID' column does not have any effect in prediction and modelling. So it was dropped.

3. 'Specialty Description' is simply a description of 'Specialty Code' column and so it is redundant. So it was also dropped.

4. The columns ‘Title’, ‘Specialty Code’, ‘State’, ‘Brand 2 Sales (Competitor Brand)’, ‘Total Branded Market Sales’ have missing data:

Column	No. of Rows With Missing Data
Title	6401
Specialty Code	757
State	757
Brand 2 Sales (Competitor Brand)	48091
Total Branded Market Sales	42385

5. Populate the missing value:

- Missing values in the ‘Title’ column are populated with value ‘UNK’ (which stands for ‘unknown’)
- Missing values in the ‘Specialty Code’ column are populated with value ‘UNK’ (which stands for ‘unknown’)
- Missing values in the ‘State’ column are populated with value ‘UN’ (which stands for ‘unknown’)
- Missing values in the ‘Brand 2 Sales (Competitor Brand)’ are populated with a value of 0. When the competitor’s sales data is not known, assuming 0 is a good choice.
- Missing values in the ‘Total Branded Market Sales’ are populated with the sum of the corresponding values of ‘Brand 1 Sales (Company’s Brand)’ and ‘Brand 2 Sales (Competitor Brand)’ columns. When the other brands’ sales data is not known, the total branded sales will not be known. So assuming the sum of brand 1 and brand 2 sales as the total branded sales is a good choice.

6. In the data set, there is no row such that ‘Calls Successfully Completed’ value is higher than the ‘Calls Attempted’ value.

7. In the data set, there are 199 rows such that ‘Emails Opened’ value is greater than ‘Emails Sent’ value. This is illogical. In such rows, ‘Emails Sent’ value was replaced with ‘Emails Opened’ value.

8. In the data set, there is 1 row where the ‘Total Branded Market Sales’ value is less than the sum of brand 1 and brand 2 sales. This is illogical. In that row, the ‘Total Branded Market Sales’ was replaced with the sum of corresponding values of brand 1 and brand 2.

9. In the data set, there are 618 rows where the ‘Total Market (Branded + Unbranded) Sales’ value is less than the ‘Total Branded Market Sales’ value. This is illogical. In such rows, the

‘Total Market (Branded + Unbranded) Sales’ value was replaced with the ‘Total Branded Market Sales’ value.

10. An ‘OtherBrands’ column was created by taking the difference between the ‘Total Branded Market Sales’ column and sum of Brand 1 and Brand 2 columns. This column indicates the sales of other brands in the market.

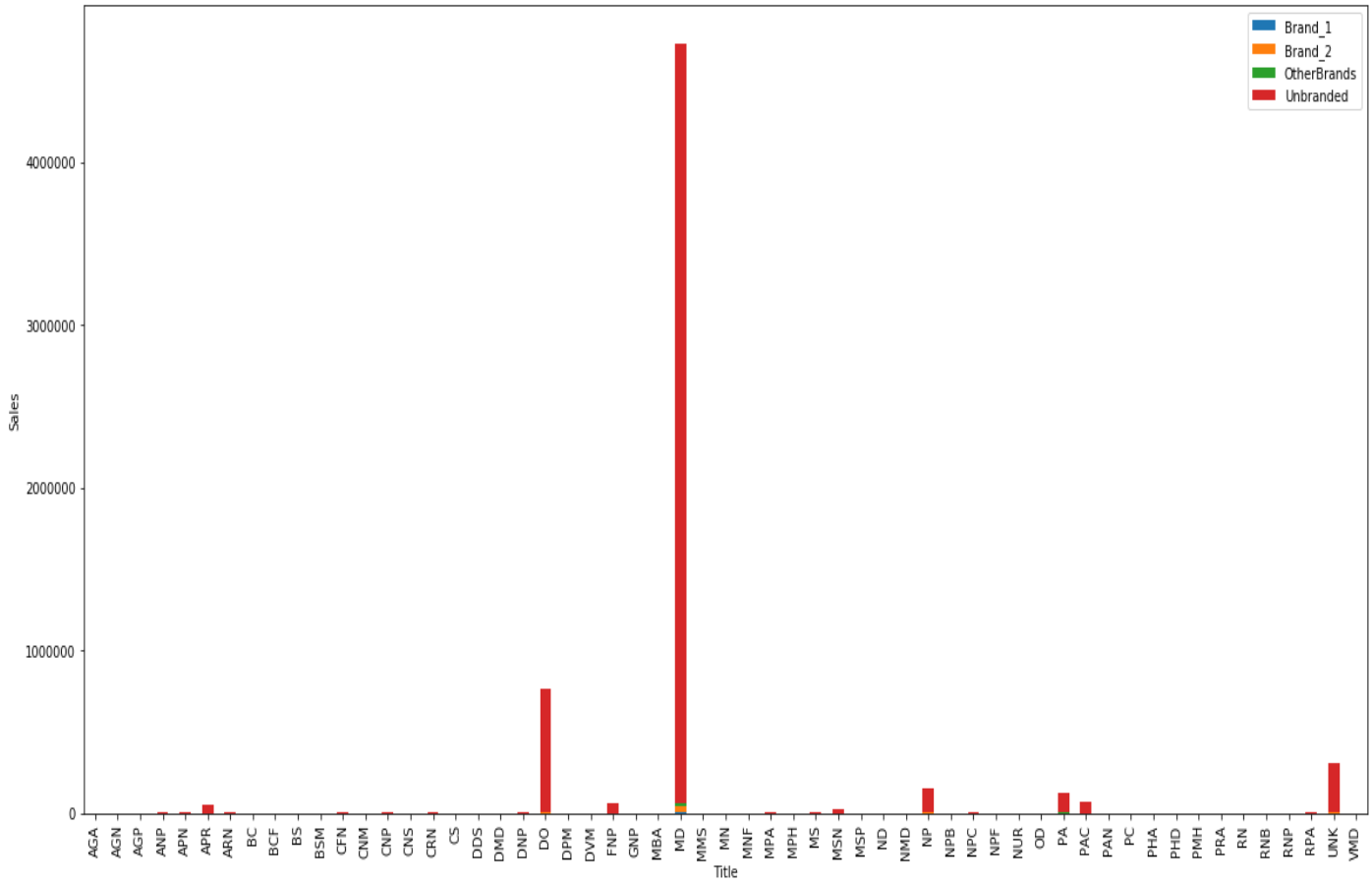
11. An ‘Unbranded’ column was created by taking the difference between the ‘Total Market (Branded + Unbranded) Sales’ and ‘Total Branded Market Sales’ values. This column indicated the sales of unbranded items in the market.

(continue to next page)

Exploratory Data Analysis:

Note:
In the following bar graphs, the colour code is:
Blue - brand 1
Orange - brand 2
Green - other brands
Red - unbranded

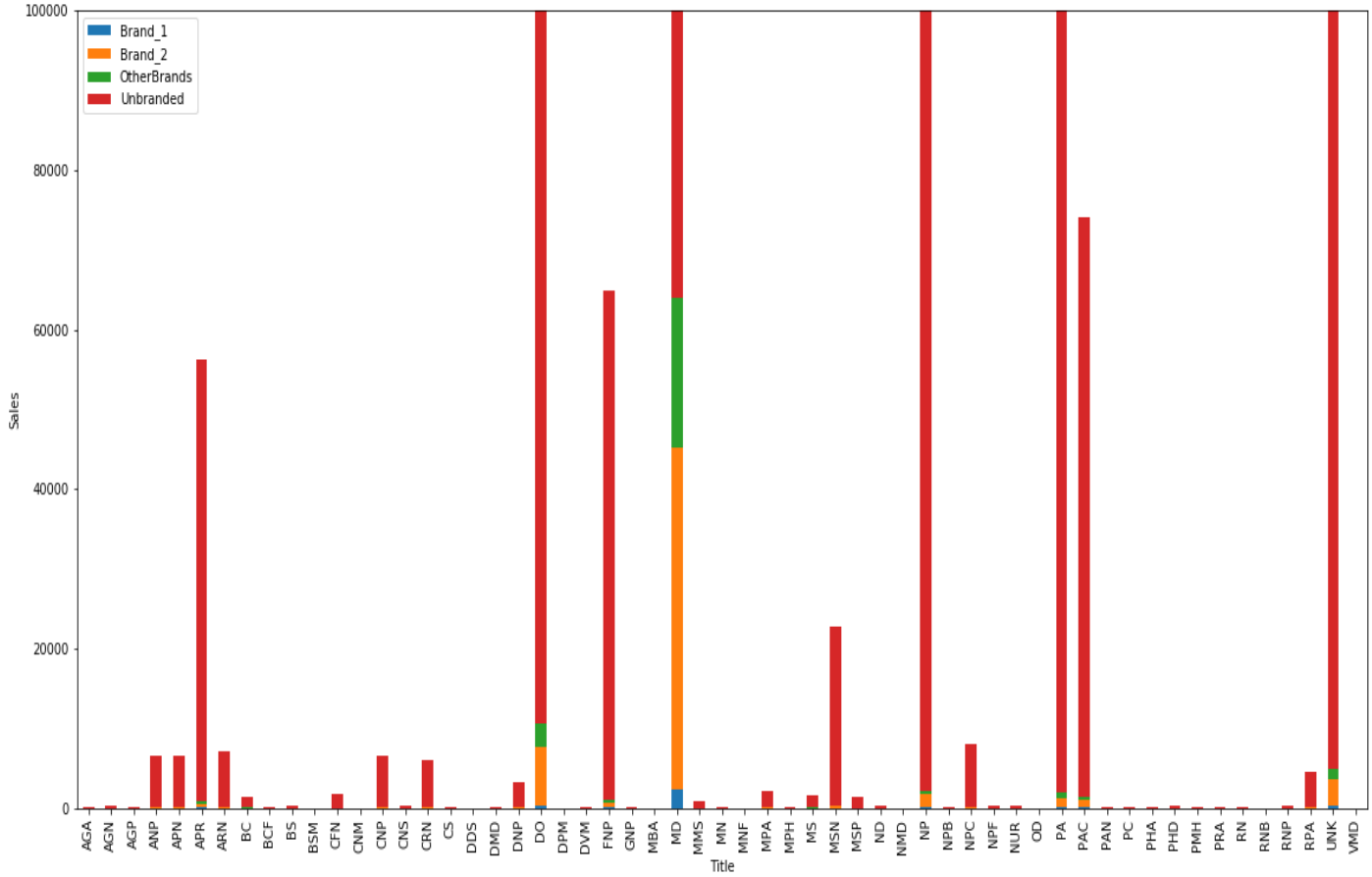
Bar Graph showing the market share of brand 1, brand 2, other brands and unbranded items with respect to the customers' 'Title' (a clearer graph with axis limits is given next):



From the graph above, it is clear unbranded items have been heavily favoured by customers across various Titles.

(continue to next page)

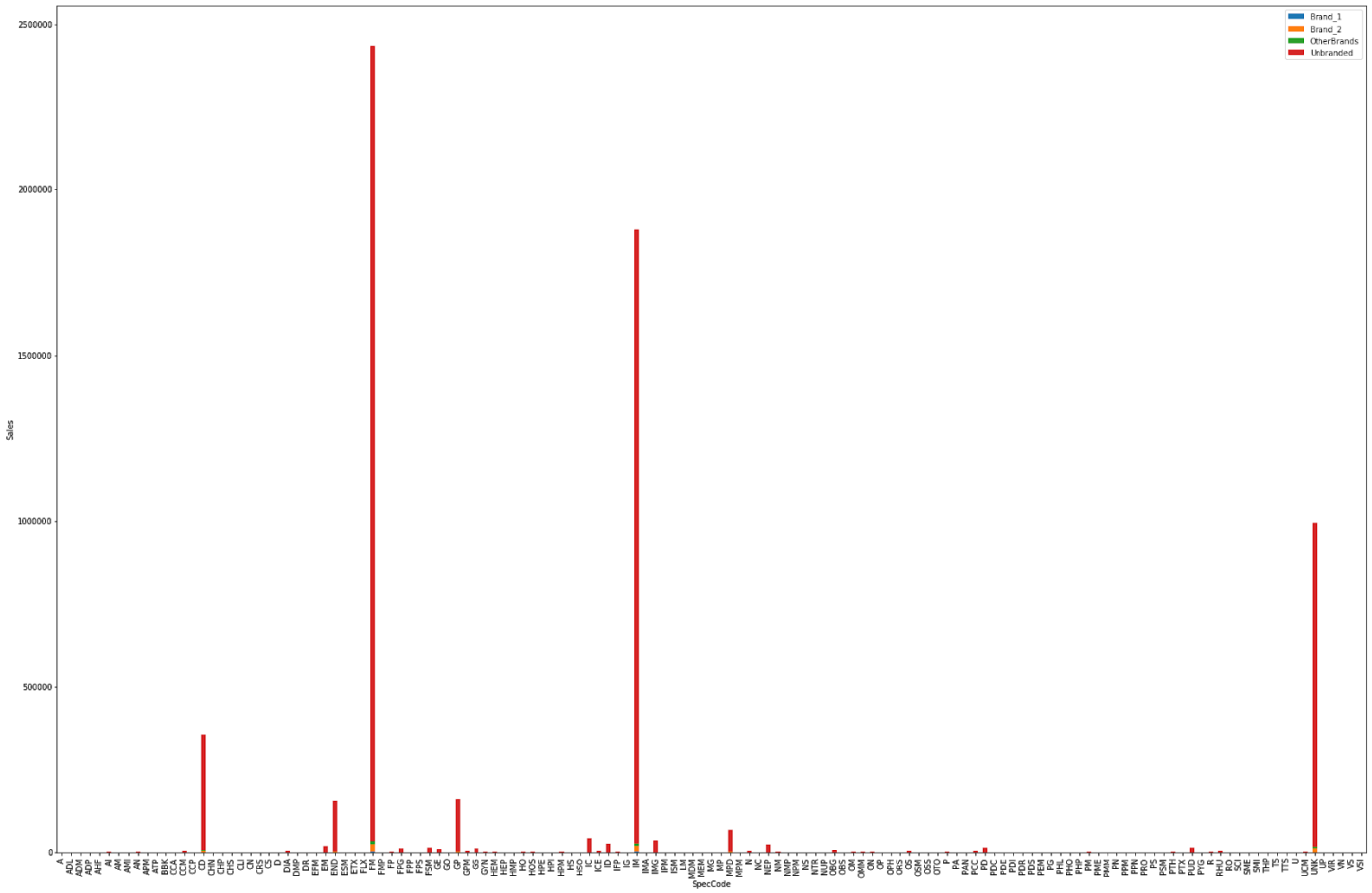
In the plot above, the market share of brand 1, brand 2 and other brands is not clearly visible. The following bar graph limits the y axis to (0, 100000) for a better view:



From the graph above, it seems brand 1 has a low market share across customers. It has some sales among customers with titles such as ‘MD’ and ‘DO’. Brand 2 enjoys a comparatively higher market share across titles.

(continue to next page)

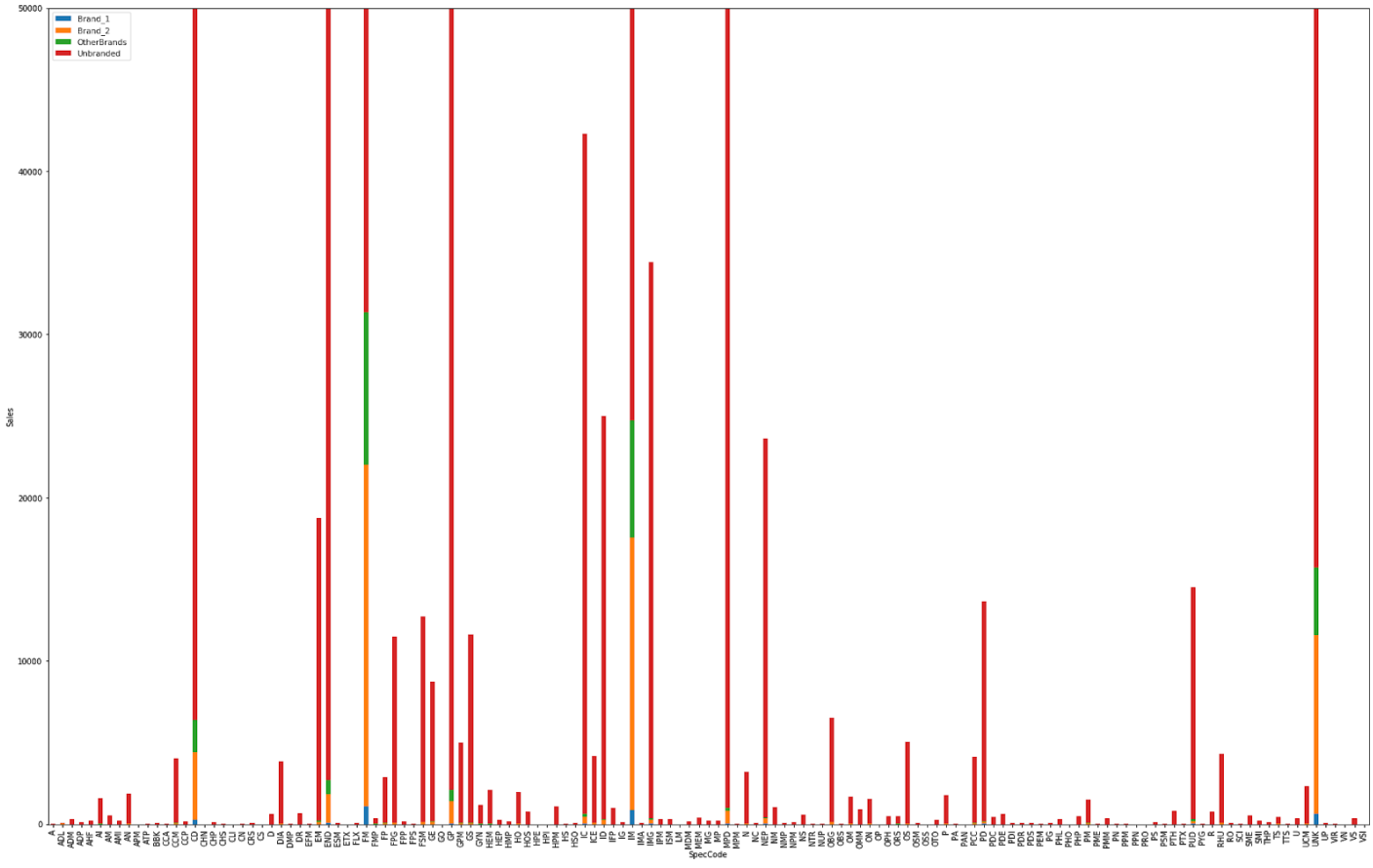
Bar Graph showing the market share of brand 1, brand 2, other brands and unbranded items with respect to the ‘Specialty Code’ (a clearer graph with axis limits is given next):



Due to too many categories in the x-axis and big range in the y-axis, the plot above is not clear. But it can be seen that unbranded items have a huge market share among customers across specialties.

(continue to next page)

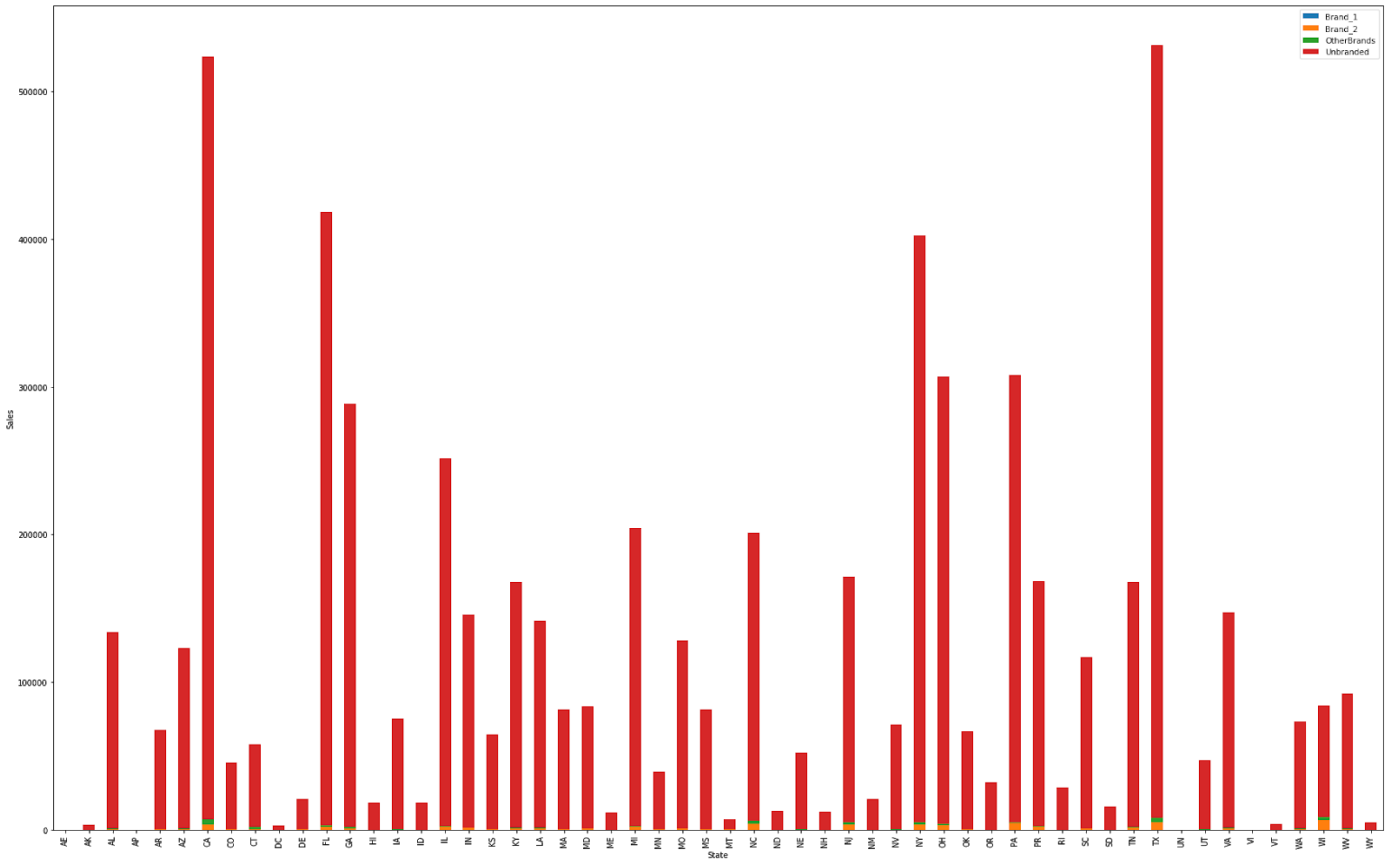
In the plot above, the market share of brand 1, brand 2 and other brands is not clearly visible. The following bar graph limits the y axis to (0, 50000) for a better view:



From the graph above, it seems brand 1 has a low market share with customers across multiple specialties. It has some sales among customers with specialties such as ‘Family Medicine’ and ‘Internal Medicine’. Brand 2 enjoys a comparatively higher market share across specialties.

(continue to next page)

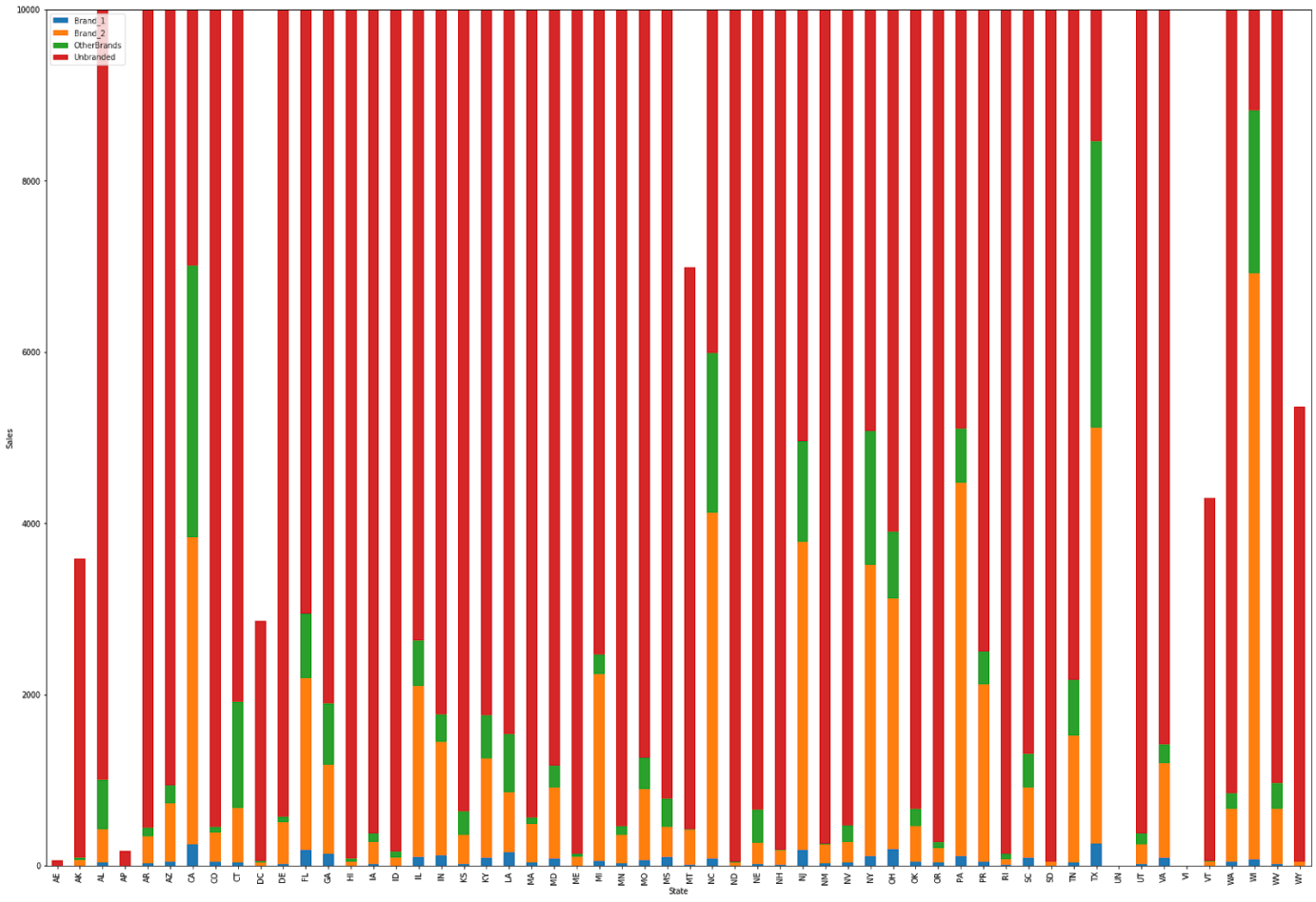
Bar Graph showing the market share of brand 1, brand 2, other brands and unbranded items with respect to the ‘State’ (a clearer graph with axis limits is given next):



Due to too many states in the x-axis and big range in the y-axis, the plot above is not clear. But it can be seen that unbranded items have a huge market share among customers across states.

(continue to next page)

In the plot above, the market share of brand 1, brand 2 and other brands is not clearly visible. The following bar graph limits the y axis to (0, 10000) for a better view:



From the graph above, it seems brand 1 has a low market share with customers across multiple states when compared to brand 2, other brands and unbranded items. But it enjoys customers across almost all states. Brand 2 enjoys a comparatively higher market share across states.

Market Shares:

Market share of brand 1 in the market for brand 1 and brand 2 alone = 5.17%
Market share of brand 2 in the market for brand 1 and brand 2 alone = 94.83%

Market share of brand 1 in the whole branded market = 3.68%
Market share of brand 2 in the whole branded market = 67.51%

Market share of brand 1 in the whole market (branded + unbranded) = 0.05%
Market share of brand 2 in the whole market (branded + unbranded) = 0.93%

Market share of branded products in the whole market (branded + unbranded) = 1.39%
Market share of unbranded products in the whole market (branded + unbranded) = 98.61%

Insights:
Considering the insights from bar graphs and market shares the following points are clear:

- Unbranded products have heavily outperformed the branded products. So the overall market seems to have favoured unbranded products.
- In the branded market itself, brand 1's sales are considerably low when compared to brand 2.
- Brand 1 enjoys customers across all states even though the market share is low.
- Brand 1 has a narrow customer base when it comes to customer titles. It enjoys a certain number of customers only with certain titles like 'MD' and 'DO'.
- Brand 1 has a narrow customer base when it comes to specialties. It enjoys a certain number of customers only with certain specialties like 'Family Medicine' and 'Internal Medicine'.

Feature Engineering:

- Since 'OtherBrands' and 'Unbranded' columns have been added, 'Total Branded Market Sales' and 'Total Market (Branded + Unbranded) Sales' are now redundant and so are dropped.
- 'Calls Attempted' and 'Calls Successfully Completed' columns give an idea about how willing the customer is to attend marketing calls.
- 'Emails Sent' and 'Emails Opened' columns give a idea about how willing the customer is to open marketing emails.
- Since it cannot be found whether a marketing fax sent was actually read or not, it should be assumed that the faxes sent are read.
- Hence the columns 'Calls Attempted', 'Calls Successfully Completed', 'Emails Sent', 'Emails Opened' and 'Faxes Sent' are used as features while developing a model
- 'Brand 1 Sales', 'Brand 2 Sales', 'OtherBrands', and 'Unbranded' columns indicate the propensity of customers to buy brand 1, brand 2, other brands and unbranded items respectively. This information will contribute to finding how effective was each channel in influencing the decision of the customer to buy the company's brand. Hence these columns also are used as features while developing a model.
- 'Title', 'Specialty Code' and 'State' are categorical features. They have to be converted into numerical features. Each of them is label encoded into integer values.

We have 12 columns for each customer with numerical data. In other words, each customer is represented using a 14-dimensional feature vector with the following feature

1. Title
2. SpecCode
3. State

4. CallAttempts
5. CallsSuccessful
6. EmailsSent
7. EmailsOpened
8. Faxes
9. Brand_1
10. Brand_2
11. OtherBrands
12. Unbranded

But the numerical values across features are of varying scales. To bring them into a common scale, MinMaxScaling was performed on them.

Modelling:
The dataset is not labelled with labels indicating which of the three marketing channels is to be recommended. Since we have unlabelled data set, this is an unsupervised learning problem.

Clustering was performed on the data set to cluster them into three clusters ie. three marketing channels. K-means algorithm was chosen to perform the clustering.

Performance Evaluation of the Model:
Silhouette score is a metric that measures how well k-means algorithm has performed the clustering process.

The Silhouette score is calculated using the mean intra-cluster distance (a) and the mean nearest-cluster distance (b) for each sample. The Silhouette score for a sample is (b-a)/max(a,b). Here b is the distance between a sample and the nearest cluster that the sample is not part of.

The mean of the Silhouette scores of all samples has been used as a performance metric.

For the model developed, the score is approximately ***0.40*** which indicates that the clustering processing is good and that the clusters are well separated.

Channel Affinity Score:
We get the distance of each data point to the cluster centroids after the clustering process is over. A data point will have a low distance to the cluster centroid of the cluster it belongs to and will have higher values for the other two clusters. From this a channel affinity score for each channel can be developed for each data point (customer).

Inversing and Softmaxing has been followed to achieve this. For a data point, each of the three distances is inversed. Hence the distance to the cluster centroid on the cluster in which the data

point belongs to will be converted into the highest value. The other two distances will be converted into smaller values.

The resulting three values are given to the softmax function to get probabilities as outputs (softmax function takes an array of numbers and converts them into probabilities that sum to 1). These values are the channel affinity score of the customer. The channel to which the data point belongs will get the highest channel affinity score.

Labeling the clusters:
A few data points were looked at and manually labeled. Then the cluster to which they belong was found and its label was chosen accordingly.

For example, the customer with the 'Customer ID' of 10469 has the following values in the 'Calls Attempted', 'Calls Successful', 'Emails Sent', 'Emails Opened', 'Faxes Sent' and 'Brand 1 Sales columns':

Calls Attempted	Calls Successful	Emails Sent	Emails Opened	Faxes Sent	Brand 1 Sales
20	1	17	0	0	11

It seems that the customer has a propensity to buy brand 1 through calls.

After looking for the cluster to which the data point above belongs, it was found that it belongs to cluster 0. Hence cluster 0's label is chosen as 'call'. It contains the customers for whom 'call' marketing channel should be recommended.

Similarly, for the customer with Customer ID 32844, the following is found:

Calls Attempted	Calls Successful	Emails Sent	Emails Opened	Faxes Sent	Brand 1 Sales
9	0	17	0	1	2

Since the customer did not attend the calls or open the emails, it seems brand 1 sales was due to fax. This data point belongs to the cluster 2. Hence cluster 2's label is chosen as fax.

The other remaining cluster (cluster 1) is labelled as 'email'.

Finding the cluster and the channel affinity scores of a customer:
Let us take the customer who has the following values for the various columns:

ID	Title	Code	Spec. Description	State
32832	MD	FM	FAMILY MEDICINE	NE

Calls Made	Calls Successful	Emails Sent	Emails Opened	Faxes	Brand	Brand 2	Total Branded	Total Market
0	0	17	0	0	0		7	29

Computing what cluster it belongs to yields the following:

Marketing channel recommended: email
Channel Affinity Scores: [0.14432833 0.82969686 0.02597481]

That is, the channel affinity scores for call is 0.1443, email is 0.8297, fax is 0.0260 (after rounding to 4 decimal places).

Conclusion:
In this work, a solution for a data science problem related to market analytics has been developed. Data preprocessing, exploratory data analysis, feature engineering and model building were performed. Exploratory data analysis yielded useful insights into the situation of the brand and the market. A clustering model was built that will aid in recommending what marketing channel to use for a customer and gives the channel affinity scores.