Customer Reviews Analysis:

Unlocking Insights into Consumer Satisfaction and Preferences

BY
HAMZA YOUSAF

BUSINESS QUESTIONS & MOTIVATION BEHIND THEM

1. **The first question aims to determine if the customer is likely to recommend the product or not based on their review?** **(**Positive recommendations from satisfied customers can lead to new customers trying out the product, while negative recommendations can have the opposite effect.)

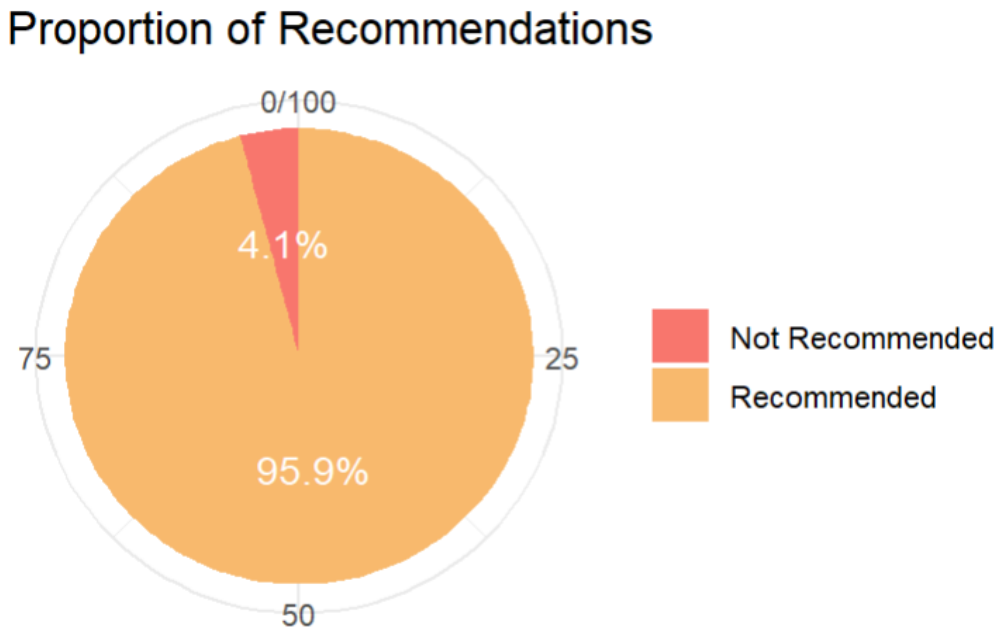
2. **The second question seeks to identify if the customer will give the product a rating of 5 or not?**
   (Marketers need to focus on providing excellent customer experiences and addressing any issues that arise promptly to increase the chances of customers giving higher ratings.)

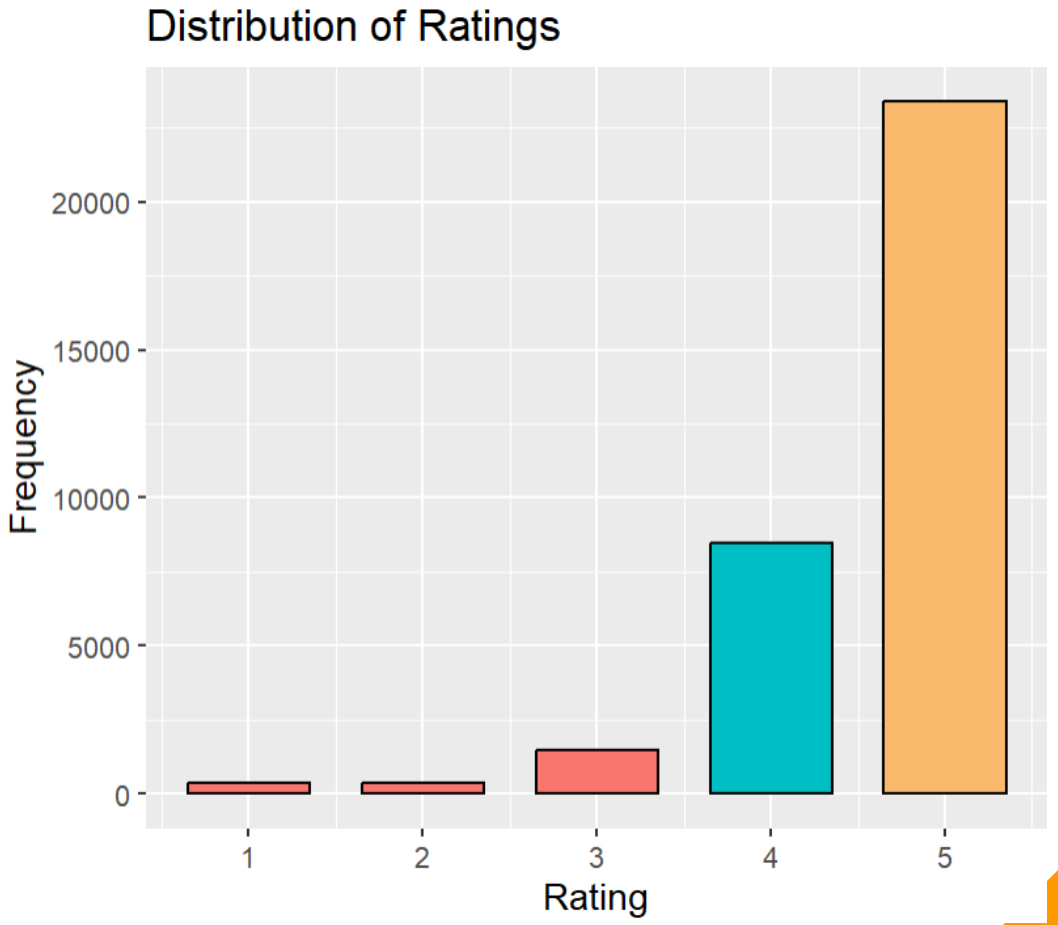
3. **The third question examines the customer's rating behavior, specifically if they will give a rating of 5, 4, or below 4?**
   (Understanding the specific rating a customer is likely to give a product can help marketers identify areas for improvement.)

DESCRIPTIVE ANALYSIS:

21 variables and **34,660** observations in this dataset.
4 variables were selected for the study, and an additional **5** product group variables were created from one of the original variables, 'categories'.

Imbalanced dataset





The dataset covers a time frame of 3.5 years, ranging from Oct 2014 to Apr2018.

DESCRIPTIVE ANALYSIS:

VARIABLE GROUPS:

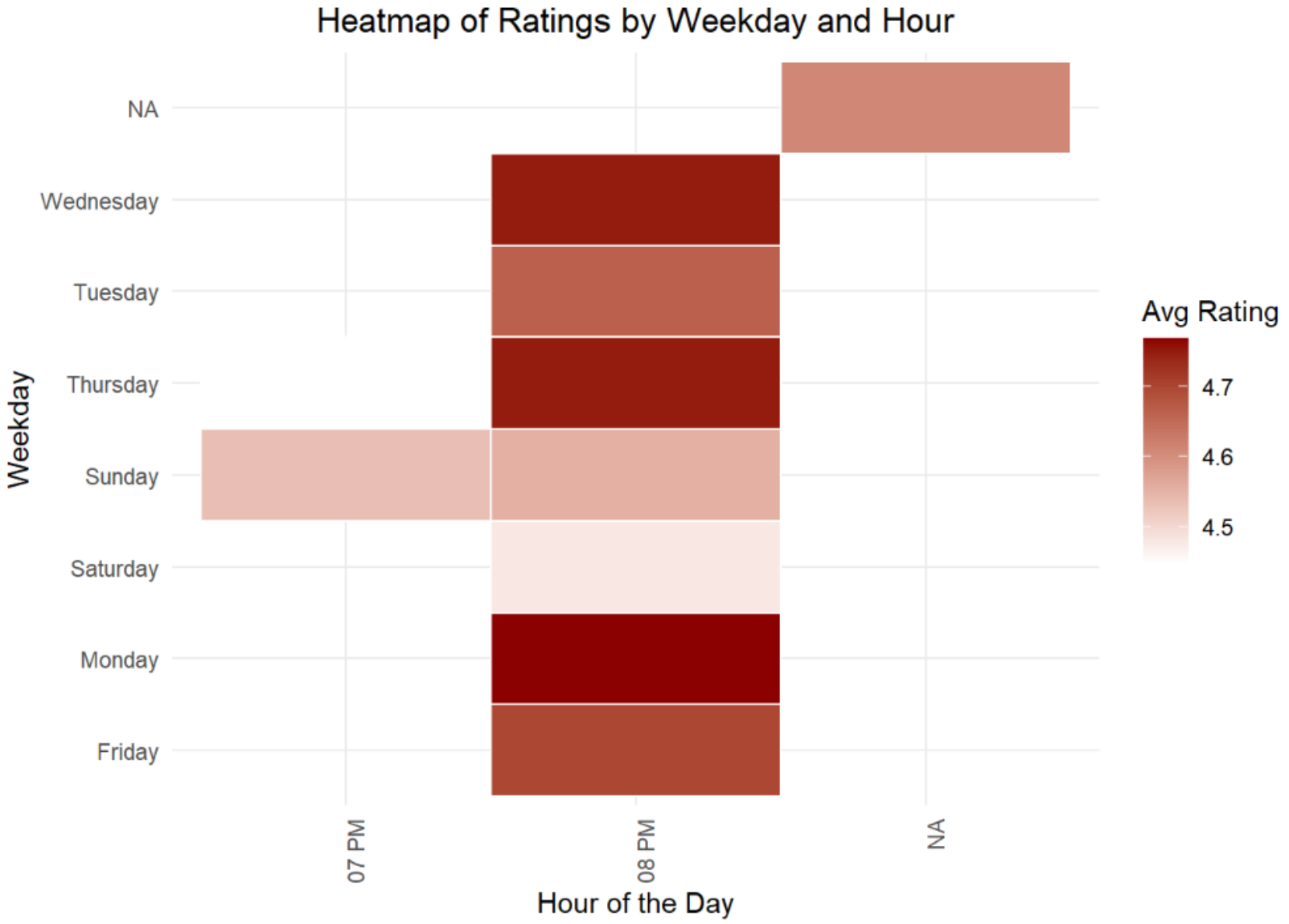
1. Tablets & E-Readers
2. Amazon Devices
3. Accessories
4. Electronics
5. office & Business

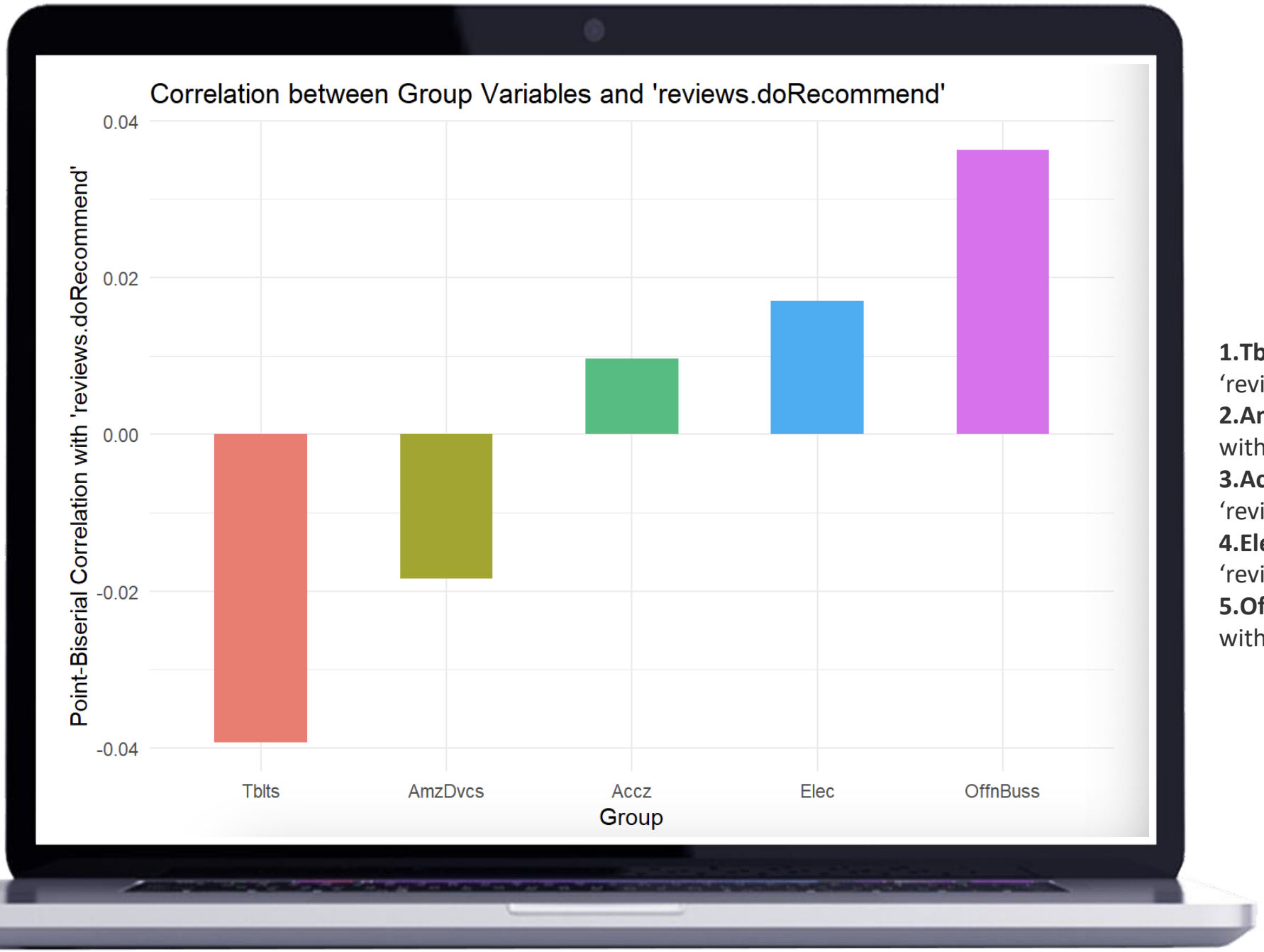
Group	Percent_Recommended	Not_Recommended_Percentage	Average_Rating	Total_Reviews
OffnBuss	98.1738	1.826197	4.772355	3176
Elec	96.68956	3.310442	4.682338	5679
Accz	96.32243	3.677574	4.670519	6662
AmzDvcs	93.52708	**6.472919**	4.560106	757
Tblts	95.16973	4.830272	4.490526	**17204**

Top 10 Products with Highest Not Recommended Percentage

DATA VISUALIZATION:

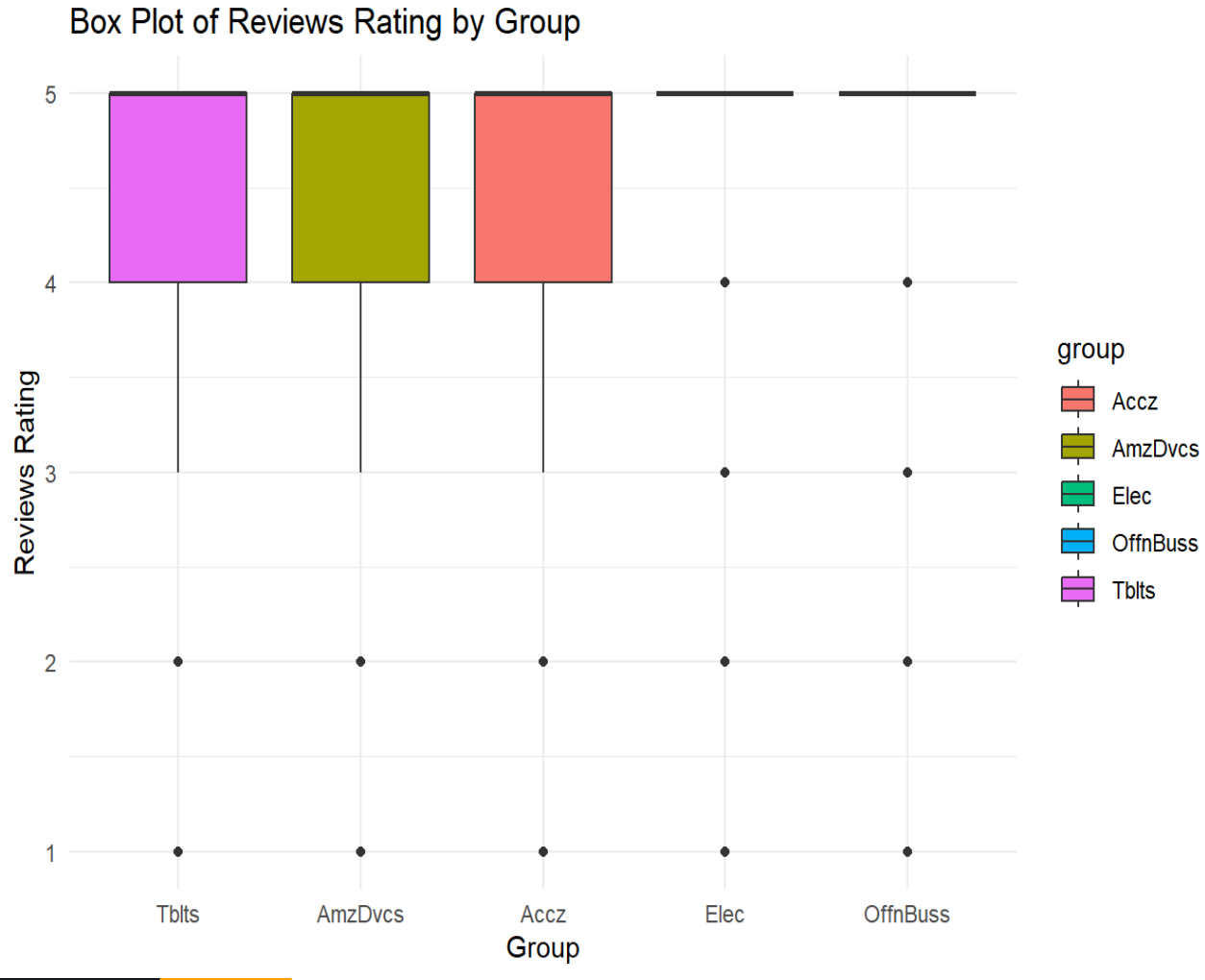
- heatmap - average customer ratings by weekdays and hours - most ratings - 8 PM
- Sundays - 7 PM - ratings range from 4.48 to 4.7
- Mondays and Wednesdays - highest ratings
- Sundays - slightly lower average of 4.55

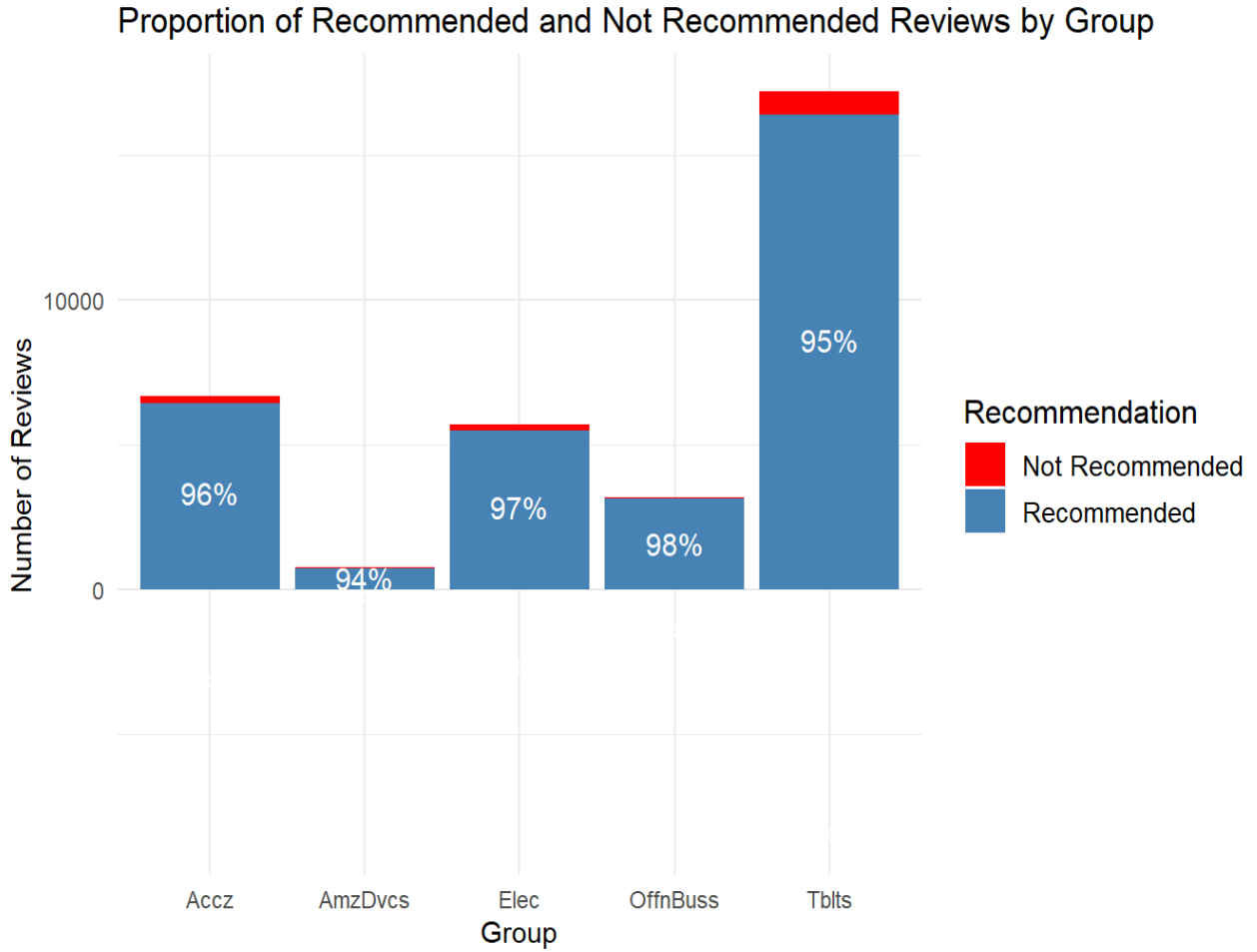




1.Tblts: weak negative correlation (-0.039) with ‘reviews.doRecommend’
2.AmzDvcs: weak negative correlation (-0.018) with ‘reviews.doRecommend’
3.Accz: weak positive correlation (0.010) with ‘reviews.doRecommend’
4.Elec: weak positive correlation (0.017) with ‘reviews.doRecommend’
5.OffnBuss: higher positive correlation (0.036) with ‘reviews.doRecommend’

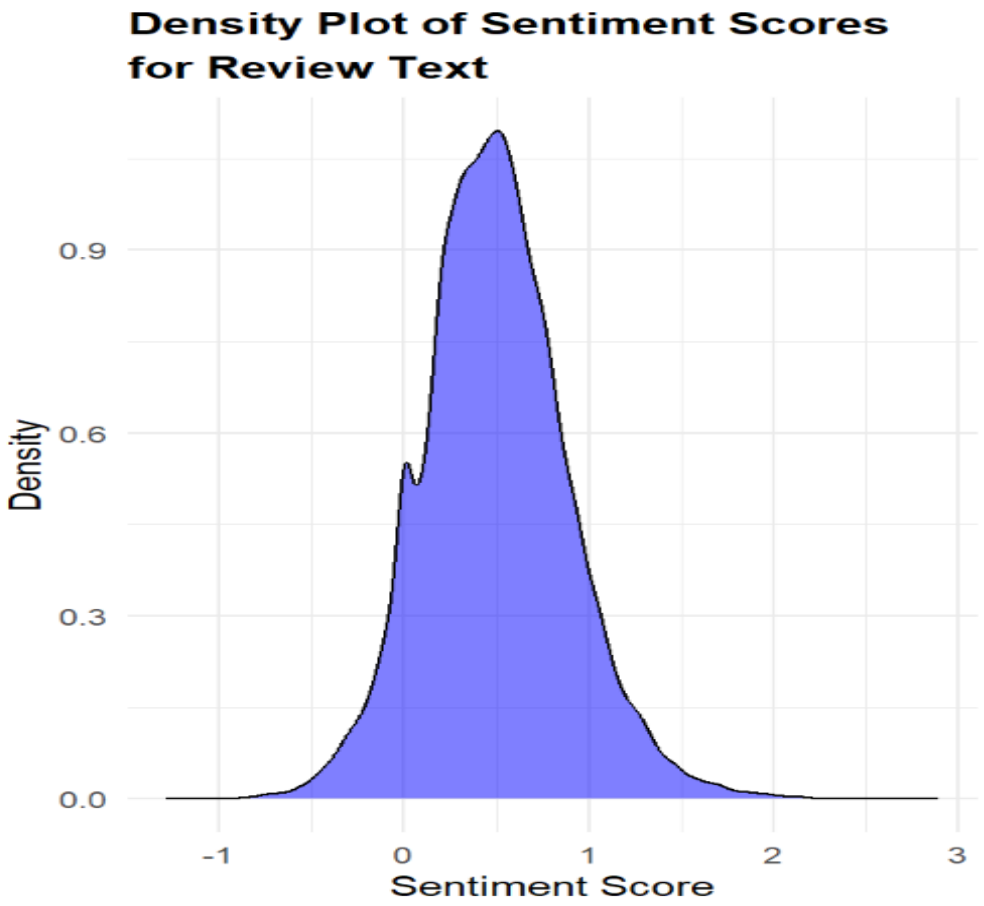
EVALUATING PRODUCTS REPUTATION

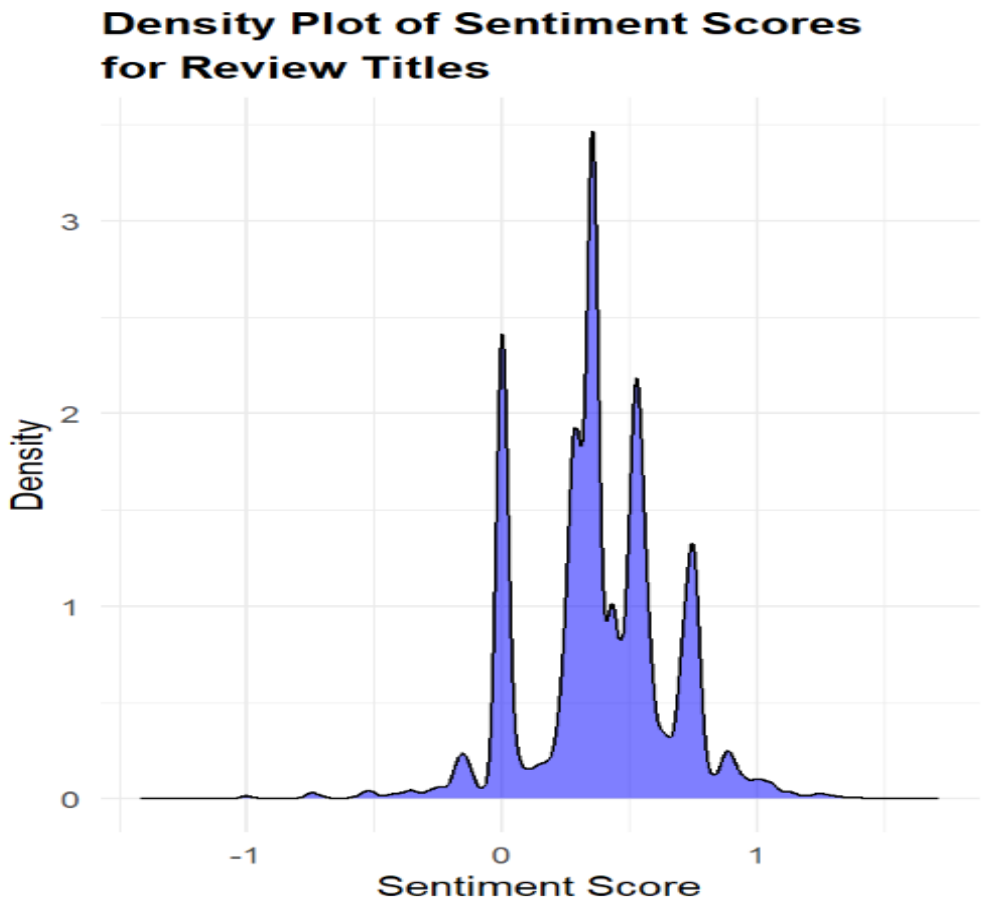




TEXT DATA PREPARATION AND SENTIMENT SCORE EXTRACTION

Lowercasing → Removing special characters → Stopwords removal → Stemming/ Lemmatization → Sentiment Score conversion



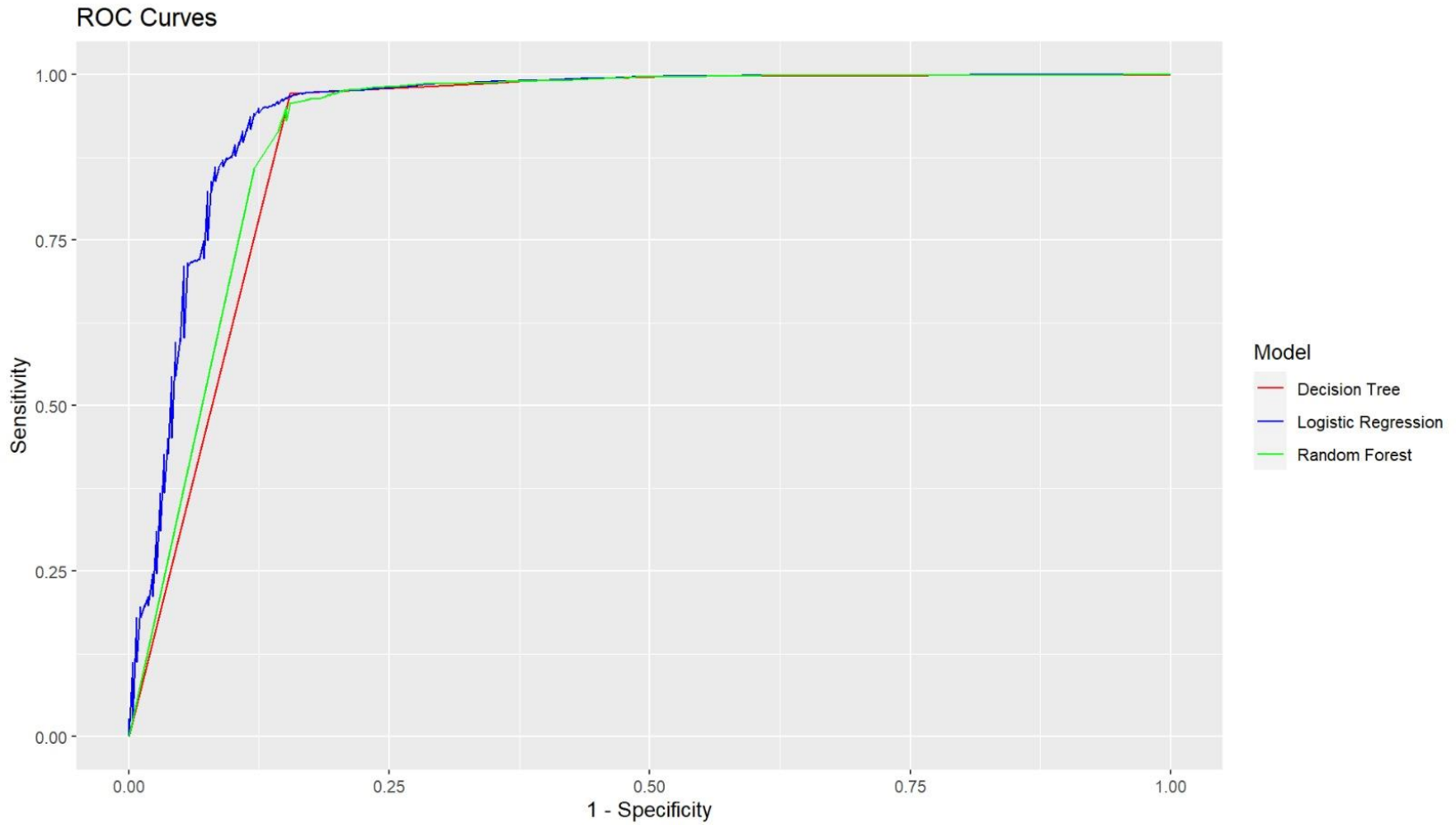


VISUALIZING CUSTOMER SENTIMENTS:

Word Cloud (Reviews-based)

1. **‘great’** and **‘use’** are most frequently - customers find the product to be of high quality
2. ‘tablet’ and ‘love’ are the second most common - customers are expressing their appreciation for the tablets
3.‘amazon’, ‘bought’, ‘good’, ‘read’, ‘work’, ‘kindle’, and ‘one’. - significant in frequency

MODEL EVALUATION – BUSINESS QUESTION 1



Predict whether customer is going to recommend the product or not?

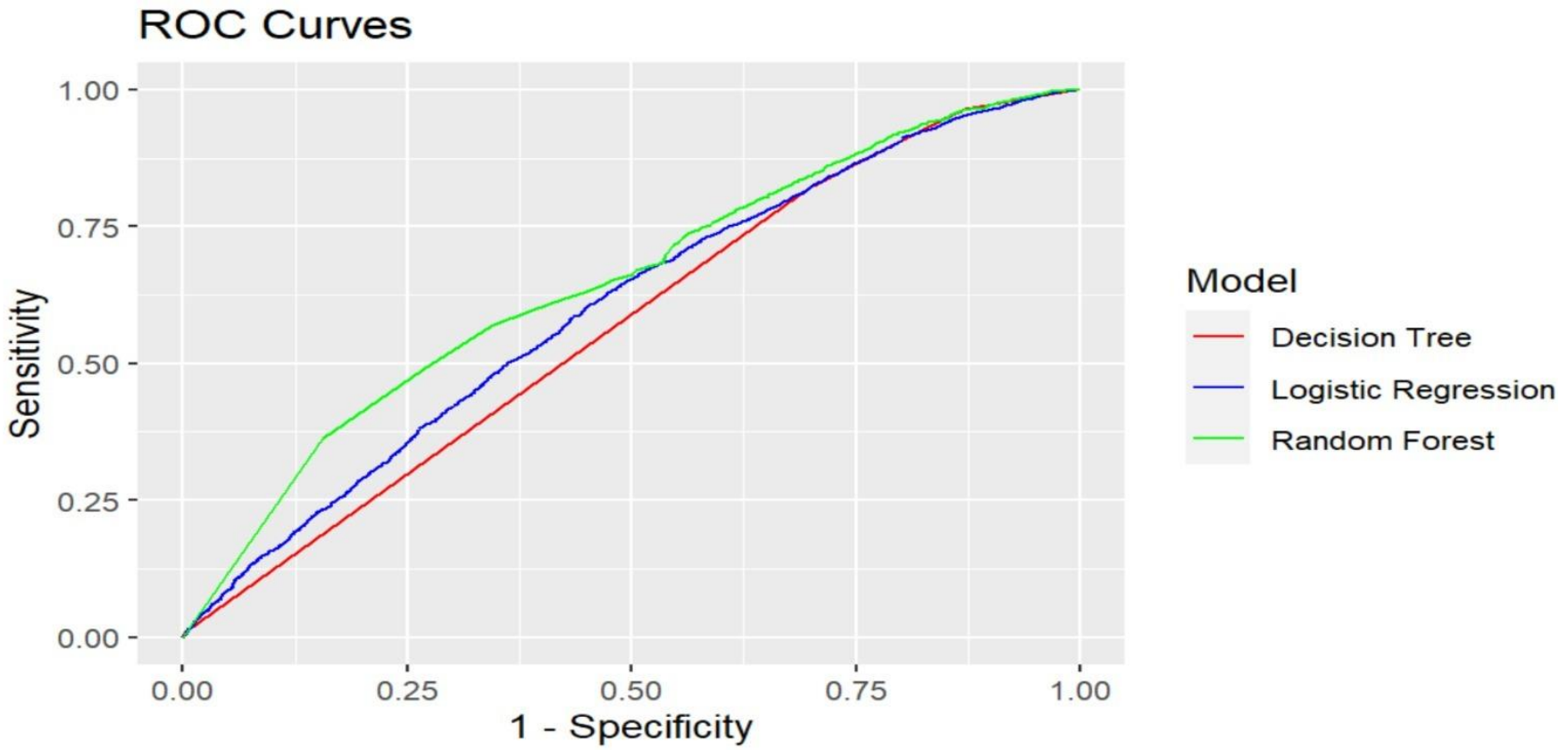
	Logistic	Classification Tree	Random Forest
Sensitivity	49.4%	61.5%	52.8%
specificity	99.6%	99.1%	99.5%
Accuracy	97.7%	97.6%	97.7%
AUROC	94.6%	91.5%	92.2%
Balanced Accuracy	74.5%	80.3%	76.1%

As our focus is on identifying the customer who will not recommend the product, so we would focus on sensitivity metric, and it is considerably higher for Classification Tree and sensitivity is also same across all the models.

Classification Tree model demonstrates the best overall performance in terms of balanced accuracy, making it the preferred choice for this business question.

MODEL EVALUATION – BUSINESS QUESTION 2

	Logisitic	Classification Tree	Random Forest
Sensitivity	55.2%	29.79%	10.20%
specificity	61.6%	82.45%	97.00%
Accuracy	59.1%	62.00%	63.40%
AUROC	62.0%	63.11%	63.92%
Balanced Accuracy	58.3%	56.12%	53.60%



Predict whether the customer is going to give rating of 5 or not?

As our focus is on identifying customers who will not give a rating of 5, we would pay more attention to the specificity metric, as it measures the model's ability to correctly identify customers who will not give a rating of 5. Random Forest model is the best choice, as it has the highest specificity of 97%

MODEL EVALUATION – BUSINESS QUESTION 3

	Multinomial logistic	Classification Tree	Random Forest
Senstivity	57.1%	56.7%	57.0%
specificity	74.2%	74.1%	74.1%
Accuracy	54.1%	57.6%	54.4%
Balanced Accuracy	65.5%	65.4%	65.5%

Predict which rating customer is going to give 5, 4 or below 4 (3-1)?

Classification Tree model has the highest overall accuracy at 57.6%, which indicates it performs better overall in classifying the ratings correctly compared to the other models.
The Classification Tree model appears to be the best choice for this business question.

RECOMMENDATIONS

- collect more data on empty variables, such as reviews.didPurchase and reviews.id. - knowing if the customer giving the review purchased the product can help to differentiate between genuine and fake reviews. - reviews.id can provide a unique identifier for each review, which can be used to avoid duplicates and improve accuracy of analysis

- collect more data on empty variables, such as reviews.userCity and reviews.userProvince - gain insights into geographic distribution of customers and tailor marketing strategies

- identified that Monday is most favorable day to request reviews from customers as they are more likely to leave positive feedback - information valuable for businesses - strategically plan their review requests and increase likelihood of receiving positive reviews - businesses can leverage this insight to enhance their customer satisfaction and overall success

- recommendations, ratings, and use of tablets have a significant impact on the overall good reviews - inform marketing and product development strategies

- collect more data that focuses on negative customer feedback, especially on variables such as reviews.doRecommend, reviews.ratings, snt_scr_text, snt_scr_title, Tblts, AmzDvcs, Accz, Elec, and OffnBuss, to enhance accuracy

Thank you!

APPENDIX!

DATA DICTIONARY:

No.	Variable Name	Data Type	Data type used for analysis	Variables choose for this study	Description
1	id	Alphanumeric	Chr		Unique identifier for the product.
2	name	Character	Chr		Name of the product.
3	asins	Character	Chr		Amazon Standard Identification Numbers for the product.
4	brand	Character	Factor		Brand of the product.
5	categories	Character	Chr	ü (made 5 product groups from this feature)	Categories the product belongs to, separated by commas.
6	keys	Character	Chr		Various identifiers and keys related to the product, separated by commas.
7	manufacturer	Character	Factor		Manufacturer of the product.
8	reviews.date	Date	Date		Date when the review was written.
9	reviews.dateAdded	Date	Date		Date when the review was added to the dataset.
10	reviews.dateSeen	Character	Chr		Dates when the review was seen online, separated by commas.
11	reviews.didPurchase	Logical	logi		Indicates if the reviewer purchased the product (TRUE) or not (FALSE).
12	reviews.doRecommend	Logical	logi	ü	Indicates if the reviewer recommends the product (TRUE) or not (FALSE).
13	reviews.id	Numeric	Int		Unique identifier for the review.
14	reviews.numHelpful	Numeric	num		Number of people who found the review helpful.
15	reviews.rating	Numeric	num	ü	Rating given by the reviewer on a scale of 1 to 5, with 1 being the lowest and 5 the highest.
16	reviews.sourceURLs	Character	Chr		URLs where the review was found, separated by commas.
17	reviews.text	Character	Chr	ü	Text of the review.
18	reviews.title	Character	Chr	ü	Title of the review.
19	reviews.userCity	Character	logi		City of the reviewer.
20	reviews.userProvince	Character	logi		Province of the reviewer.
21	reviews.username	Character	Chr		Username of the reviewer.

MULTINOMIAL LOGISTIC REGRESSION :

Metric	Class: 4	Class: 5	Class: below_4
Accuracy	54.14%	54.14%	54.14%
Sensitivity	63.90%	50.25%	57.10%
Specificity	51.80%	71.40%	99.50%
Precision	30.50%	78.60%	89.10%
Balanced Accuracy	57.60%	60.70%	78.30%

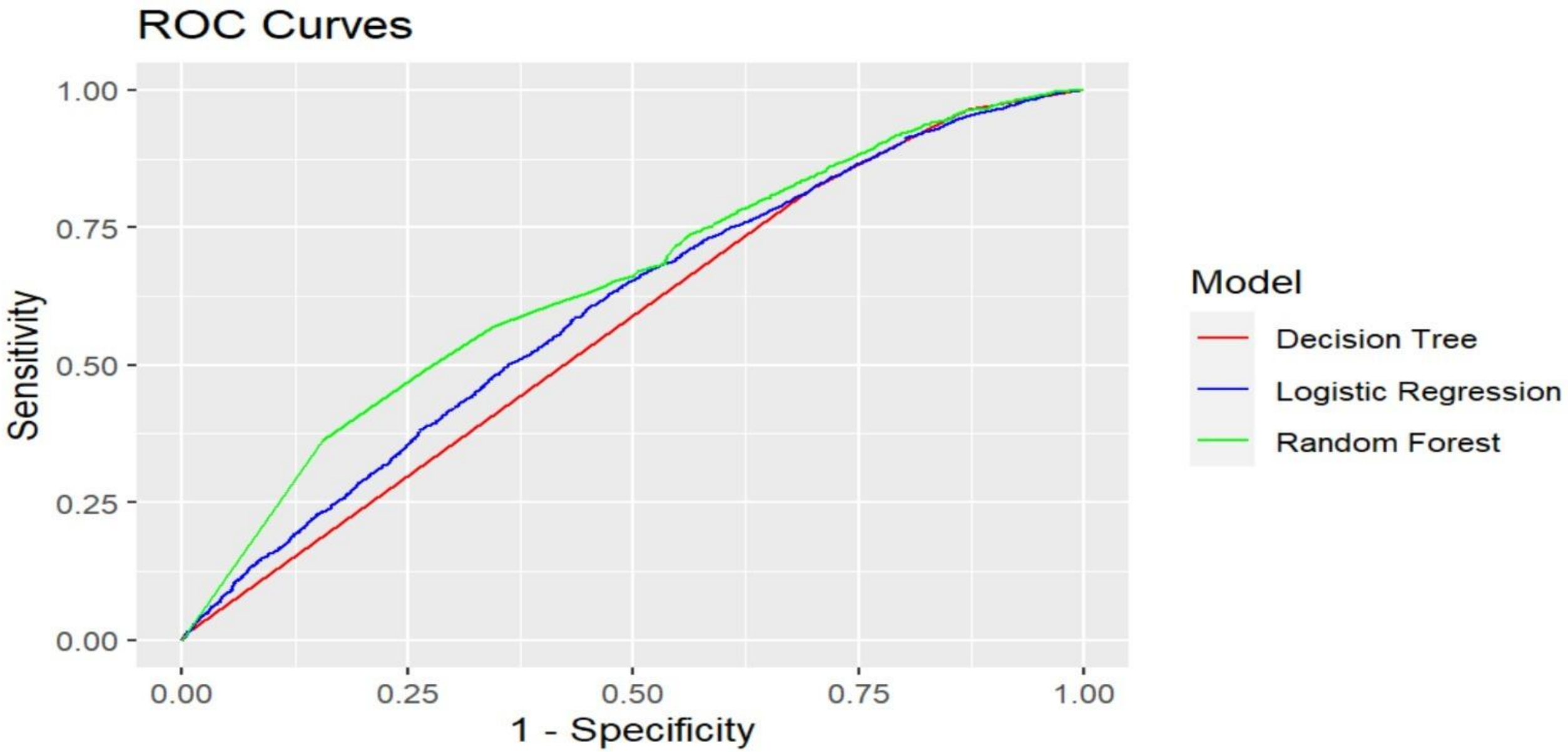
DECISION TREE:

Metric	Class: 4	Class: 5	Class: below_4
Accuracy	57.6%	57.6%	57.6%
Sensitivity	54.2%	58.8%	57.1%
Specificity	60.0%	62.9%	99.5%
Precision	31.0%	77.1%	89.4%
Balanced Accuracy	60.8%	57.1%	78.2%

RANDOM FOREST:

Metric	Class: 4	Class: 5	Class: below_4
Accuracy	54.4%	54.4%	54.4%
Sensitivity	63%	51.0%	57.1%
Specificity	52.6%	70.3%	99.4%
Precision	30.6%	78.4%	89.3%
Balanced Accuracy	57.8%	60.6%	78.2%

ROC CURVES FOR Q2 FOR MODELS BUILT ON IMBALANCED DATA



```
Confusion Matrix and Statistics

          Actual
Predicted FALSE TRUE
    FALSE   381  230
    TRUE   2261 3941

               Accuracy : 0.6344
                 95% CI : (0.6228, 0.6458)
    No Information Rate : 0.6122
    P-Value [Acc > NIR] : 8.669e-05

                  Kappa : 0.1037

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.14421
            Specificity : 0.94486
         Pos Pred Value : 0.62357
         Neg Pred Value : 0.63544
             Prevalence : 0.38779
         Detection Rate : 0.05592
   Detection Prevalence : 0.08968
      Balanced Accuracy : 0.54453

       'Positive' Class : FALSE
```

Q2 : LOGISTIC MODEL PREDICATIVE PERFORMANCE USING IMBALANCED DATA

```
Confusion Matrix and Statistics

          Actual
Predicted FALSE TRUE
    FALSE   381  230
    TRUE   2261 3941

               Accuracy : 0.6344
                 95% CI : (0.6228, 0.6458)
    No Information Rate : 0.6122
    P-Value [Acc > NIR] : 8.669e-05

                  Kappa : 0.1037

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.14421
            Specificity : 0.94486
         Pos Pred Value : 0.62357
         Neg Pred Value : 0.63544
             Prevalence : 0.38779
         Detection Rate : 0.05592
   Detection Prevalence : 0.08968
      Balanced Accuracy : 0.54453

       'Positive' Class : FALSE
```

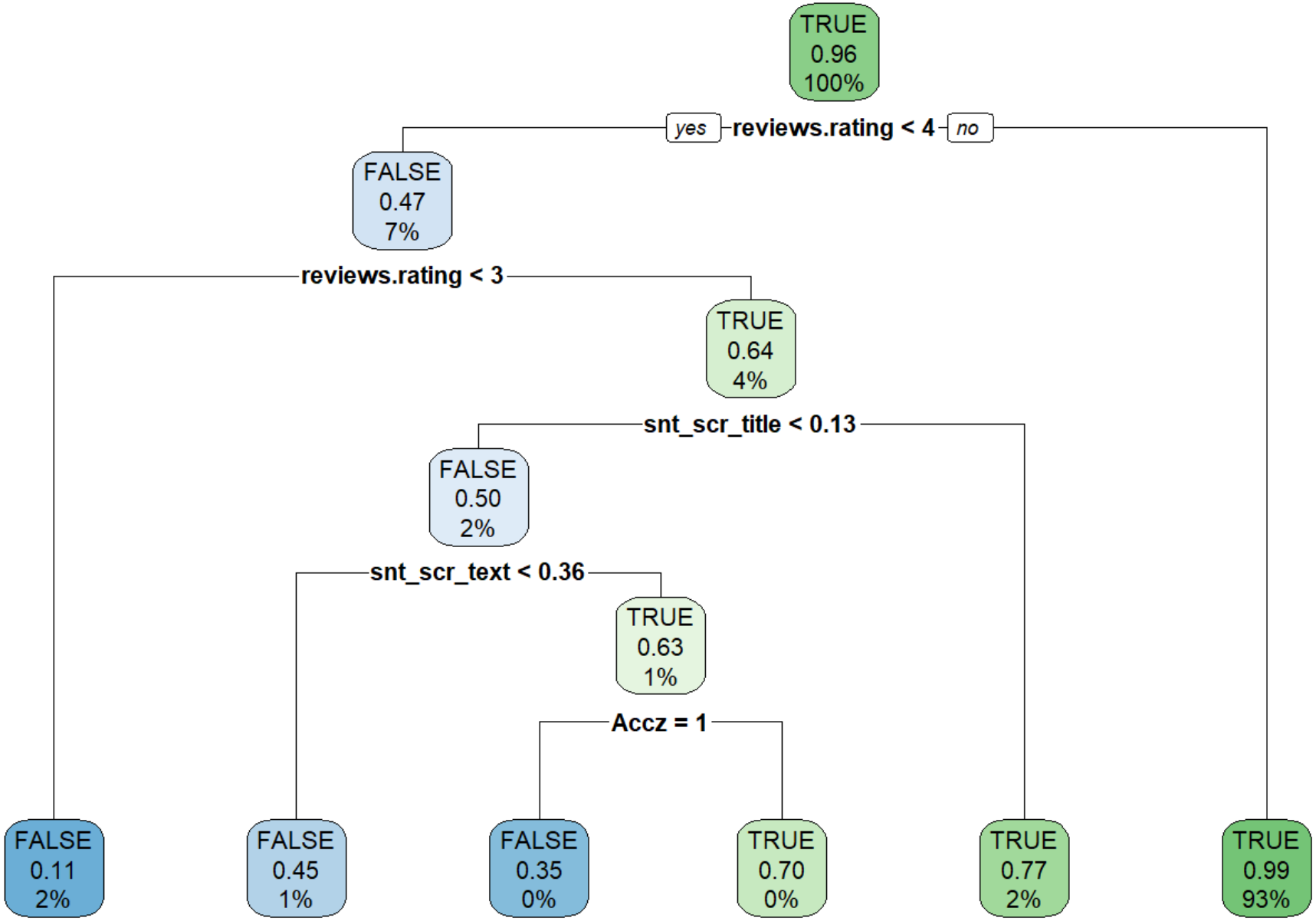
Q2 : LOGISTIC MODEL PREDICATIVE PERFORMANCE USING IMBALANCED DATA

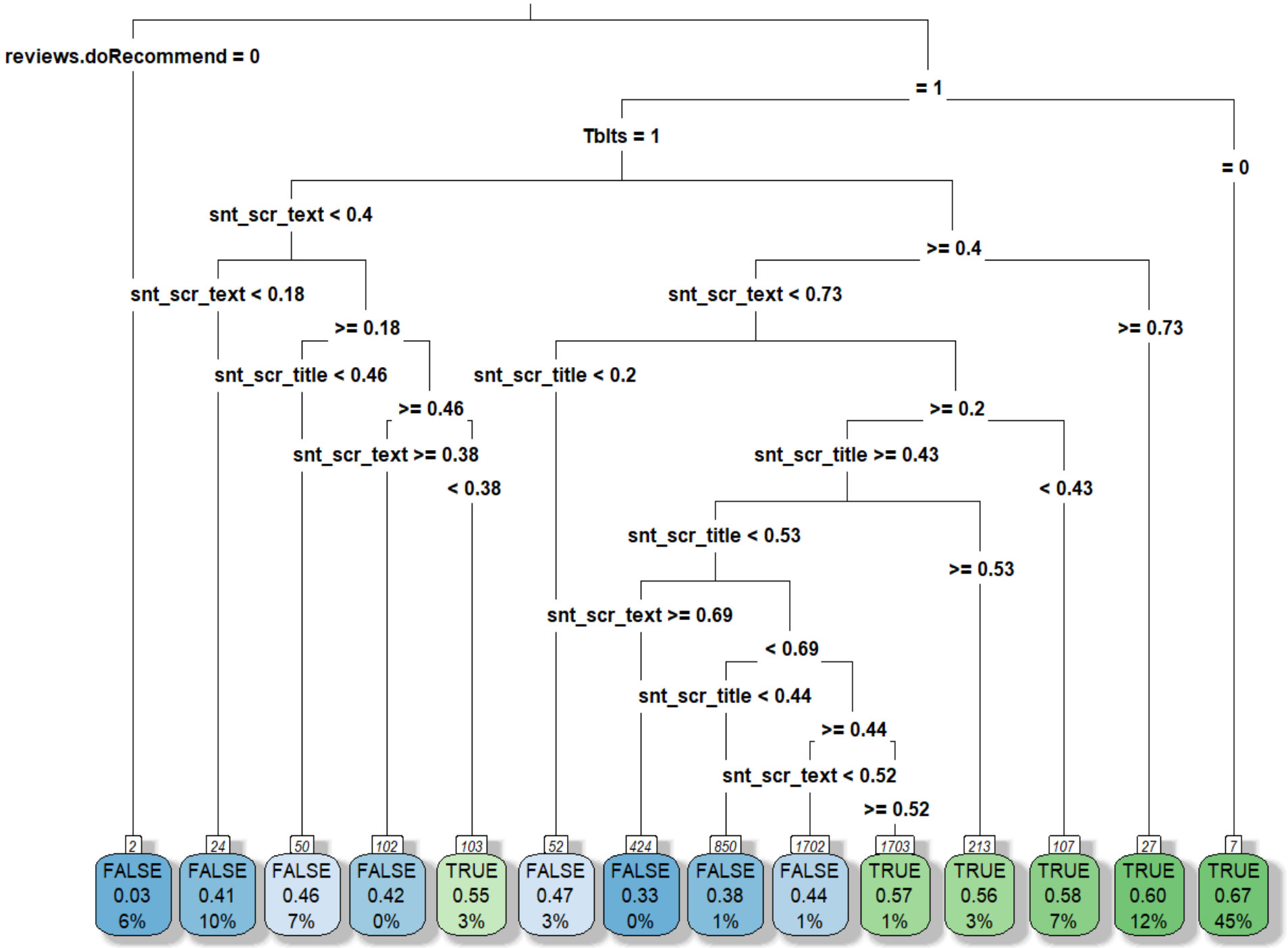
```
          Reference
Predicted FALSE TRUE
    FALSE  1457 1601
    TRUE   1185 2570
Confusion Matrix and Statistics

          Reference
Predicted FALSE TRUE
    FALSE  1457 1601
    TRUE   1185 2570

               Accuracy : 0.5911
                 95% CI : (0.5793, 0.6028)
    No Information Rate : 0.6122
    P-Value [Acc > NIR] : 0.9998

                  Kappa : 0.1629

 Mcnemar's Test P-Value : 3.767e-15

            Sensitivity : 0.5515
            Specificity : 0.6162
         Pos Pred Value : 0.4765
         Neg Pred Value : 0.6844
             Prevalence : 0.3878
         Detection Rate : 0.2139
   Detection Prevalence : 0.4488
      Balanced Accuracy : 0.5838

       'Positive' Class : FALSE
```

— low

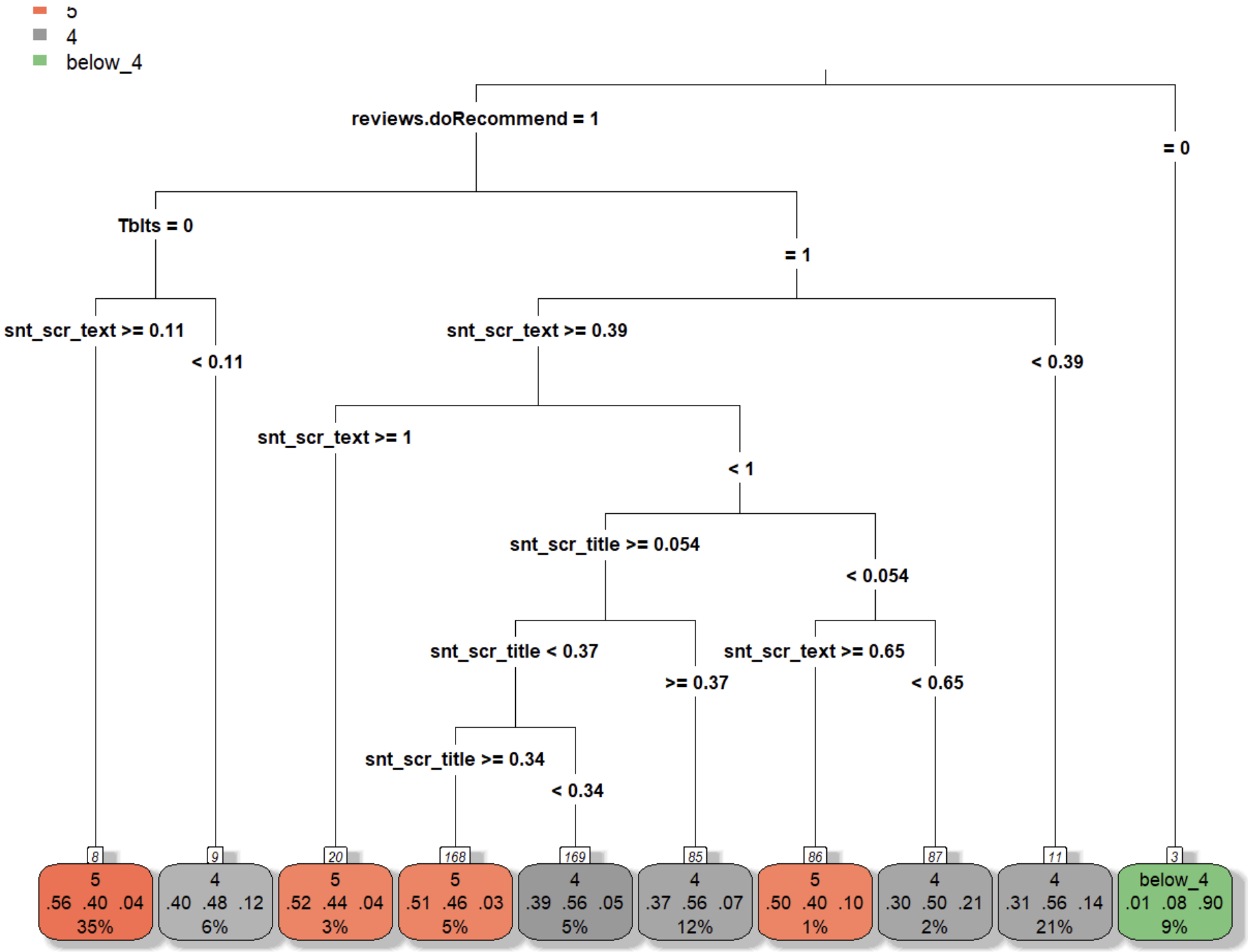
Q2 : LOGISTIC MODEL PREDICATIVE PERFORMANCE USING BALANCED DATA

RUNNING DECISION TREES ON UNBALANCE TRAINING DATA ON BUSINESS QUESTION 1:-





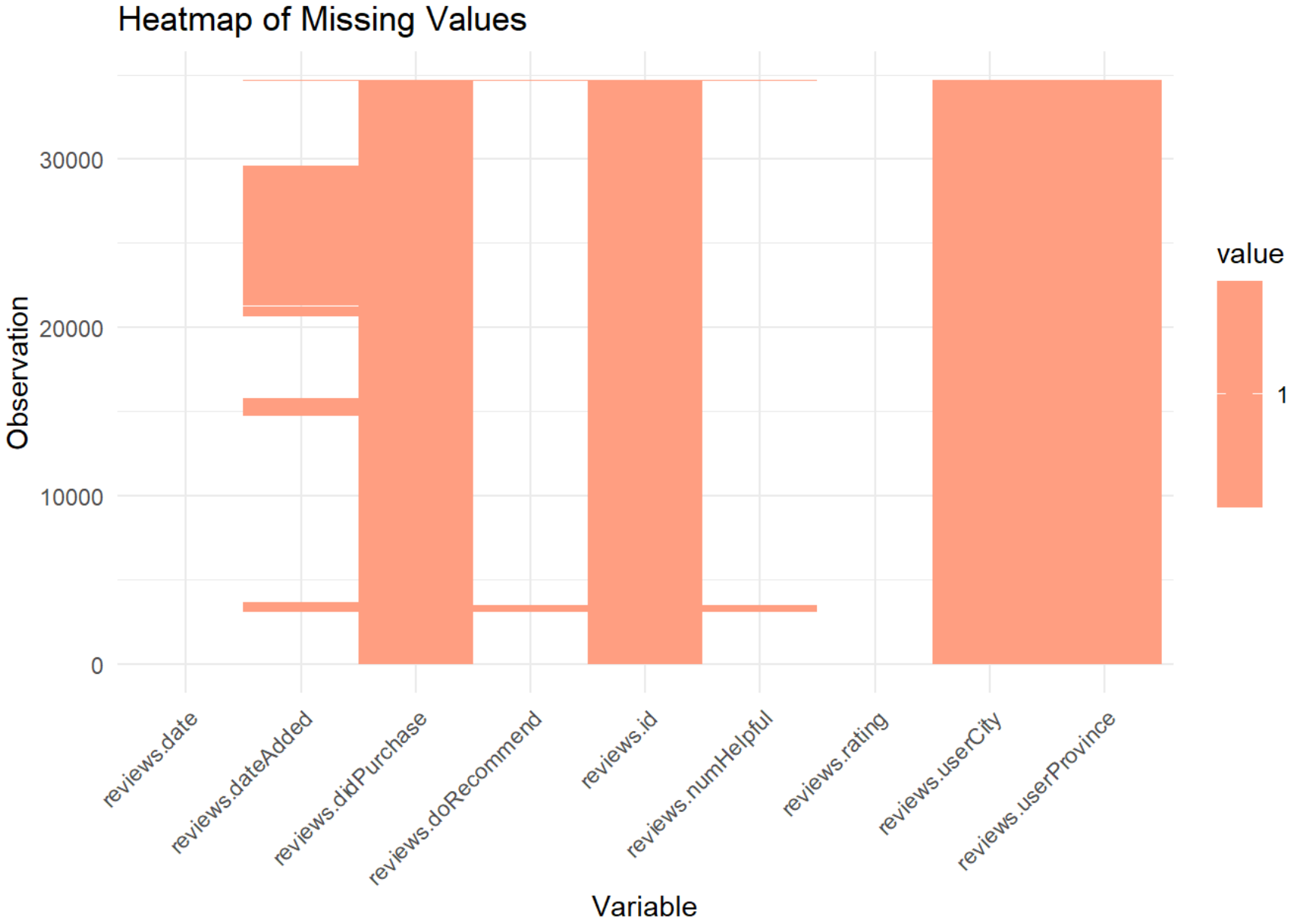
RUNNING DECISION TREE ON BUSINESS QUESTION :2

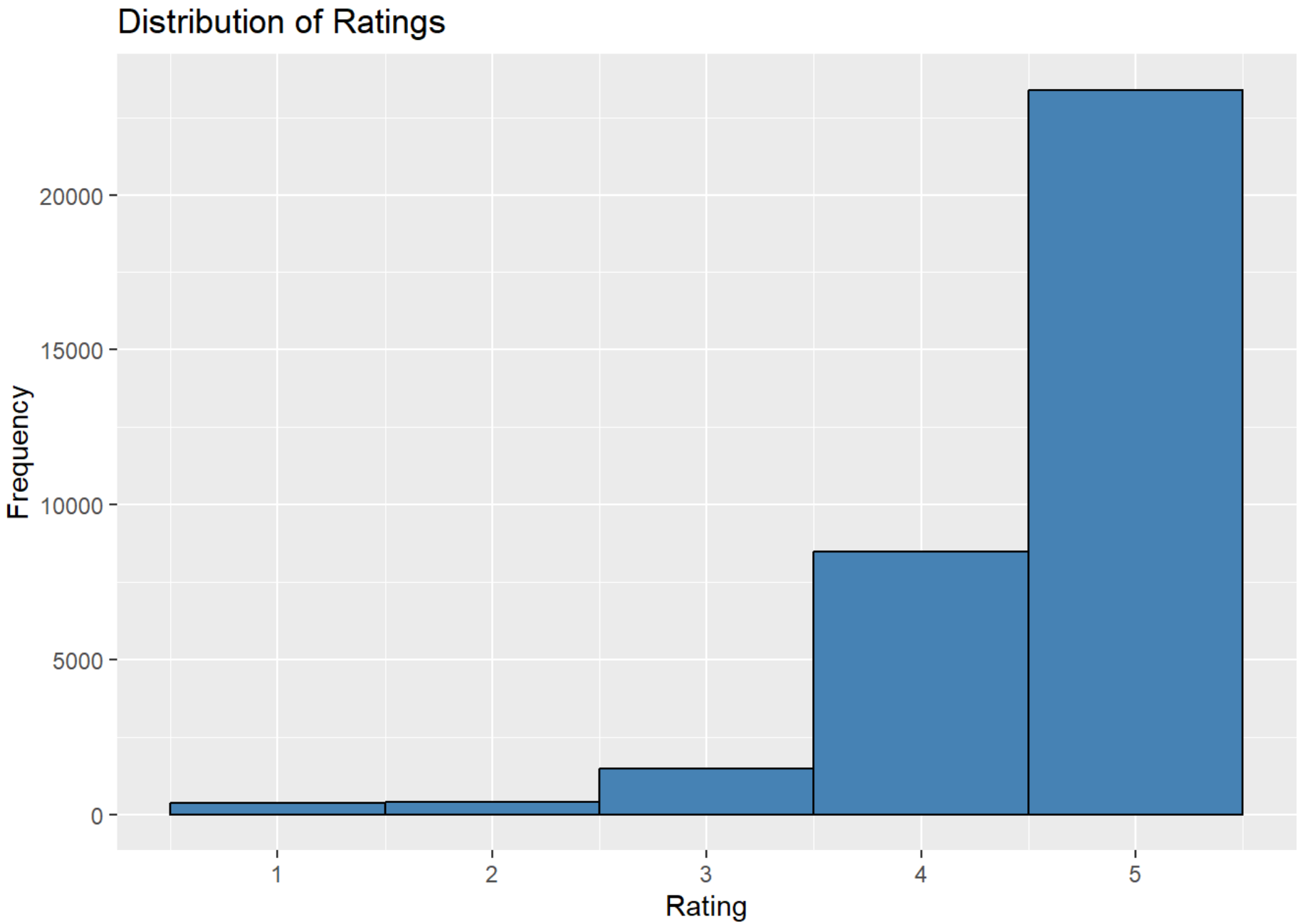


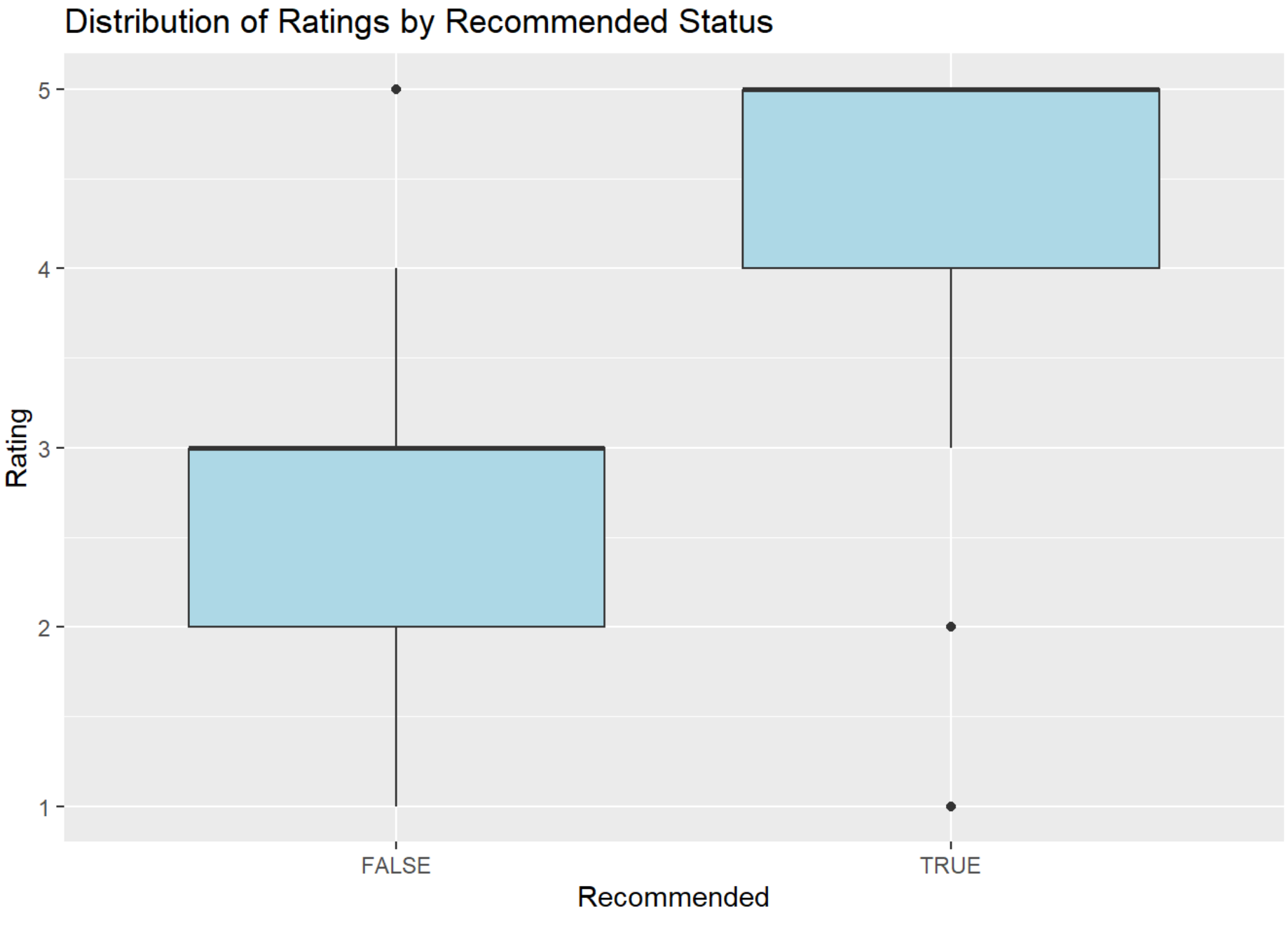
RUNNING DECISION TREE ON BUSINESS QUESTION :3

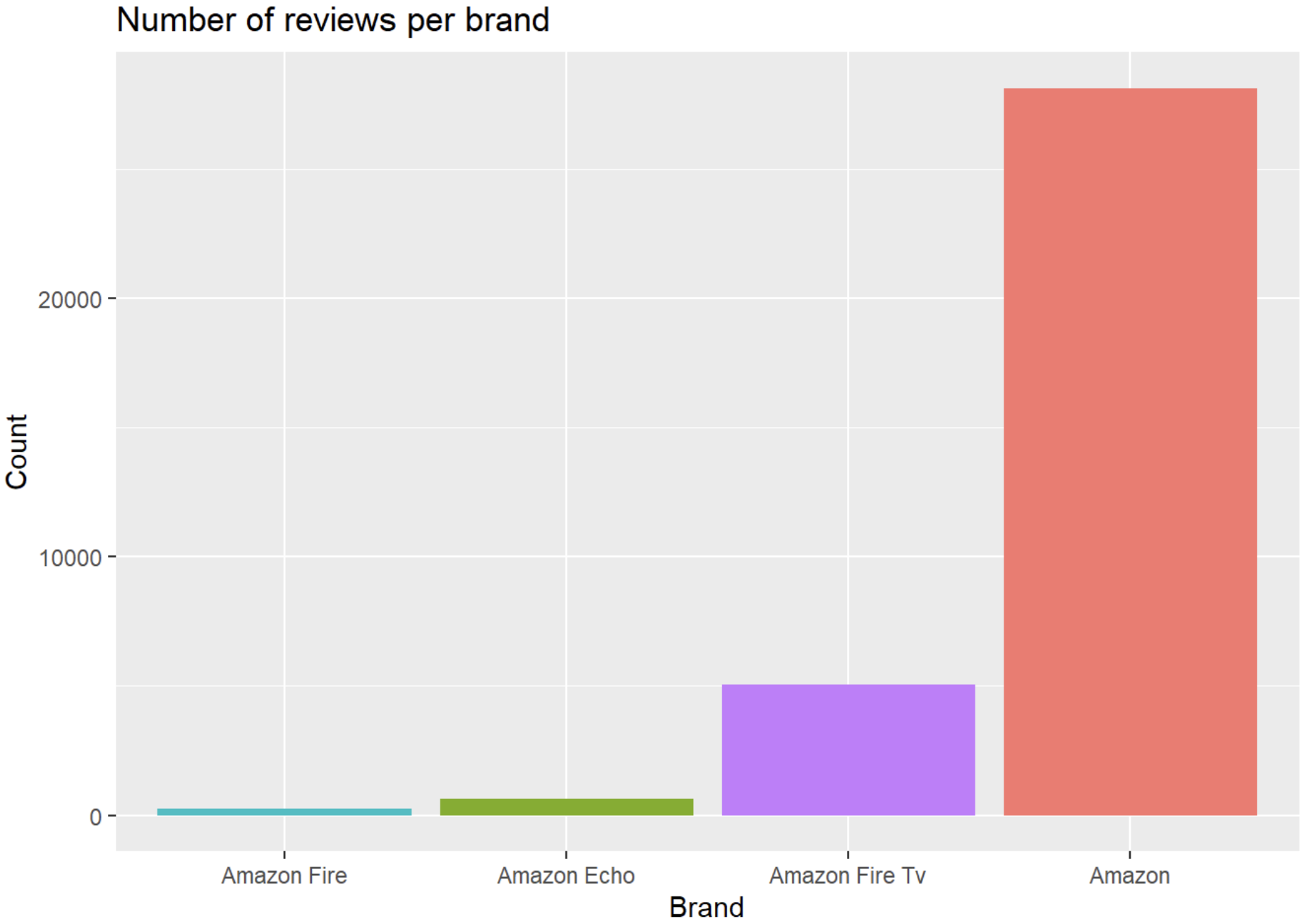
MISSING VALUES

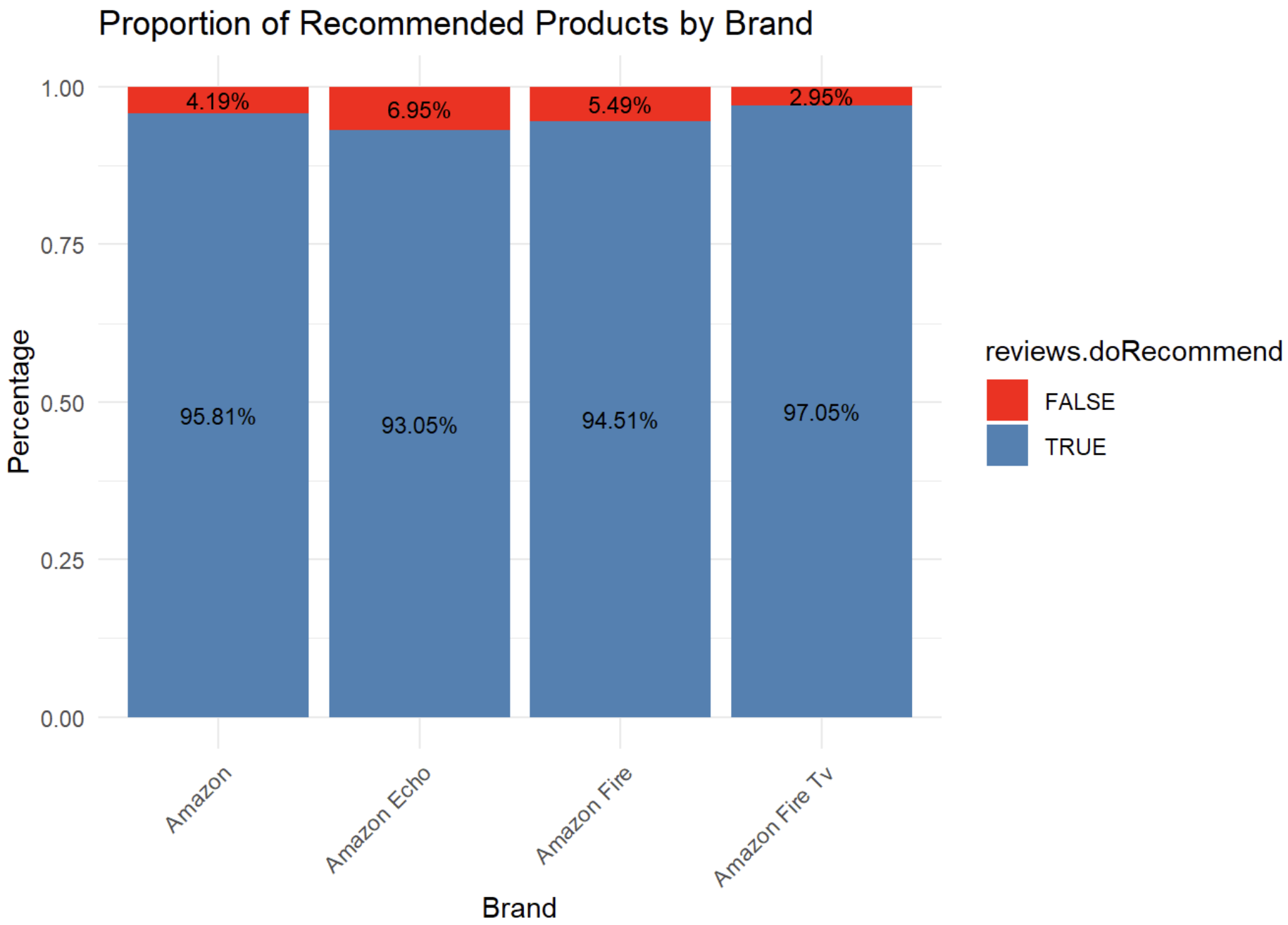
No.	Variable Name	Description	New Data type	Missing Values in each column	%age of data missing from total obv of entire var	which method to use to fill in the missing values.
8	reviews.date	Date when the review was written.	Date	39	0.1%	deleting observations
9	reviews.dateAdded	Date when the review was added to the dataset.	Date	10,621	31%	Feature not important/ not included in model building
12	reviews.doRecommend	Indicates if the reviewer recommends the product (TRUE) or not (FALSE).	logical (T/F)	594	1.7%	deleting observations
14	reviews.numHelpful	Number of people who found the review helpful.	numeric	529	1.5%	deleting observations
15	reviews.rating	Rating given by the reviewer on a scale of 1 to 5, with 1 being the lowest and 5 the highest.	numeric	33	0.10%	deleting observations
				11,816		

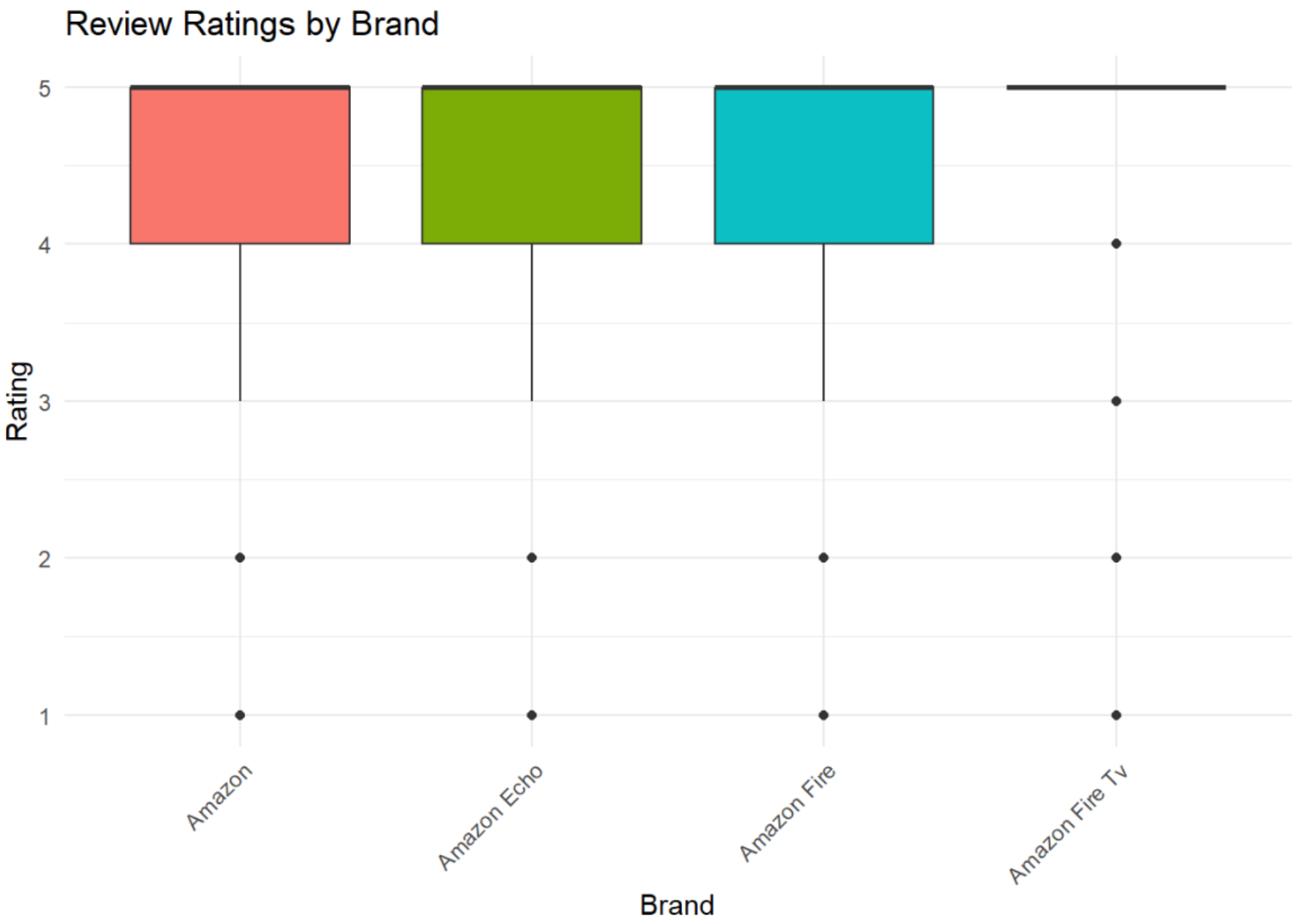

Heatmap of Missing Values
Observation
30000
20000
10000
0
reviews.date
reviews.dateAdded
reviews.didPurchase
reviews.doRecommend
reviews.id
reviews.numHelpful
reviews.rating
reviews.userCity
reviews.userProvince
Variable
value
1

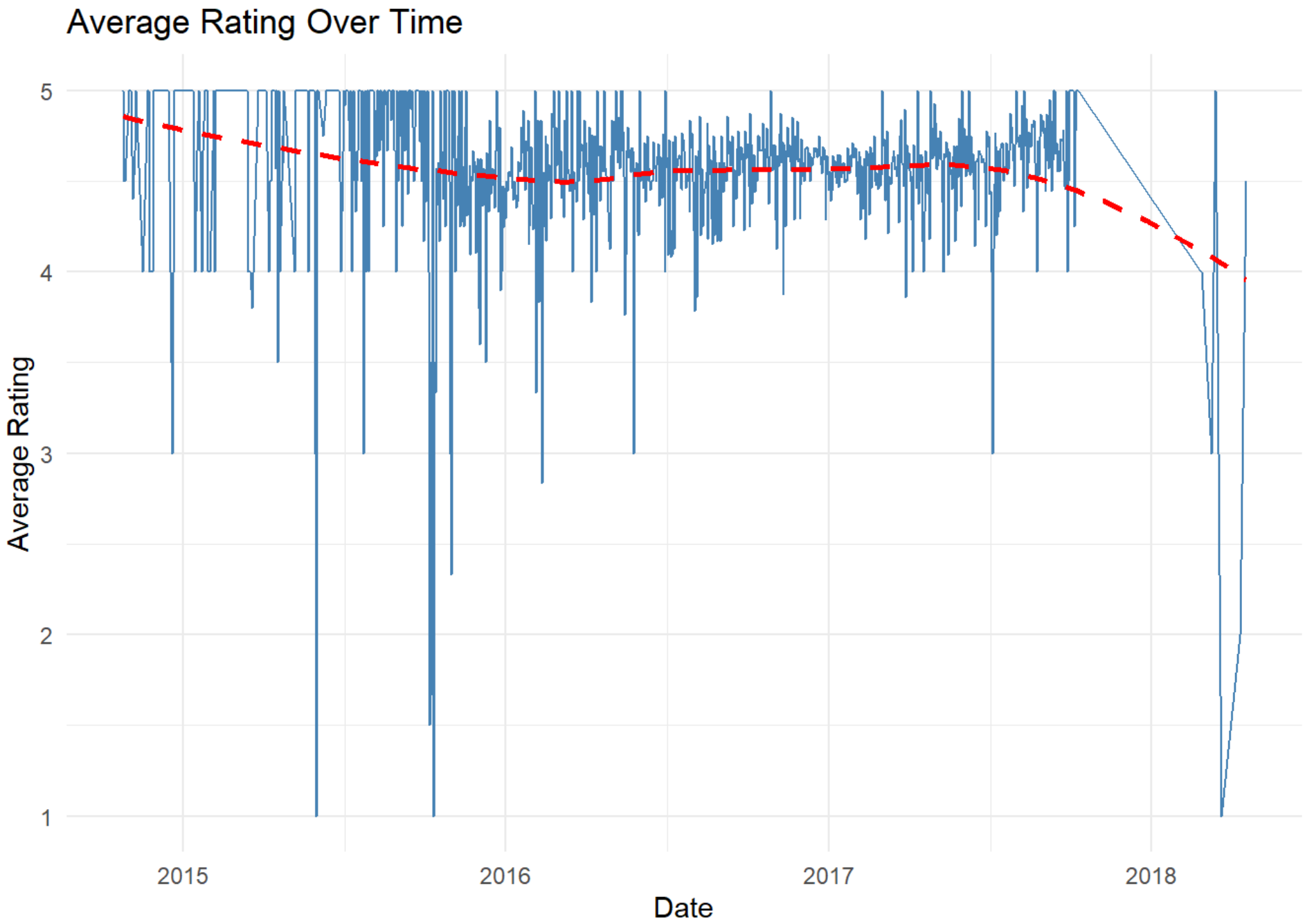

Distribution of Ratings
Frequency
20000
15000
10000
5000
0
1
2
3
4
5
Rating

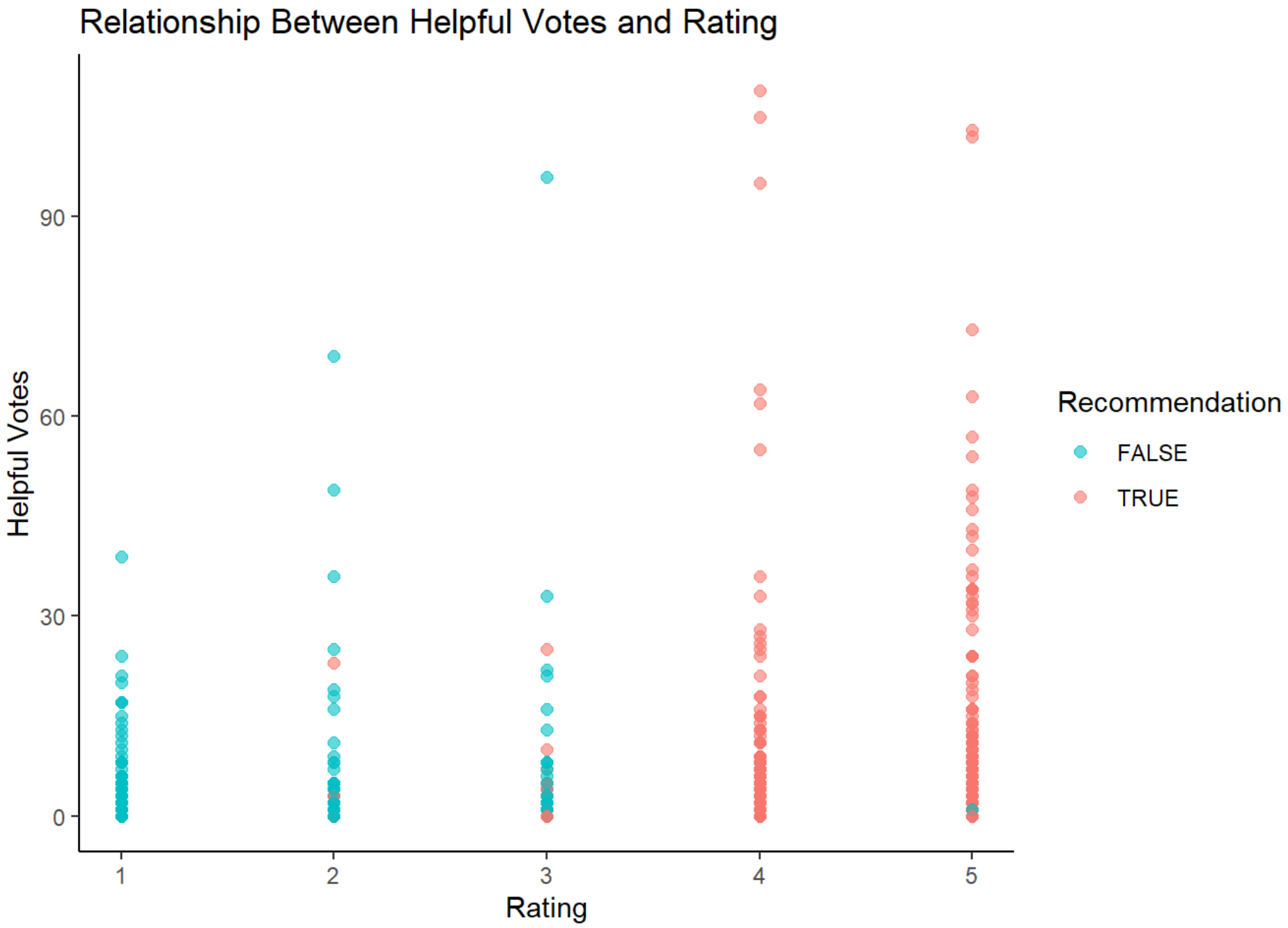

Distribution of Ratings by Recommended Status
Rating
5
4
3
2
1
FALSE
TRUE
Recommended

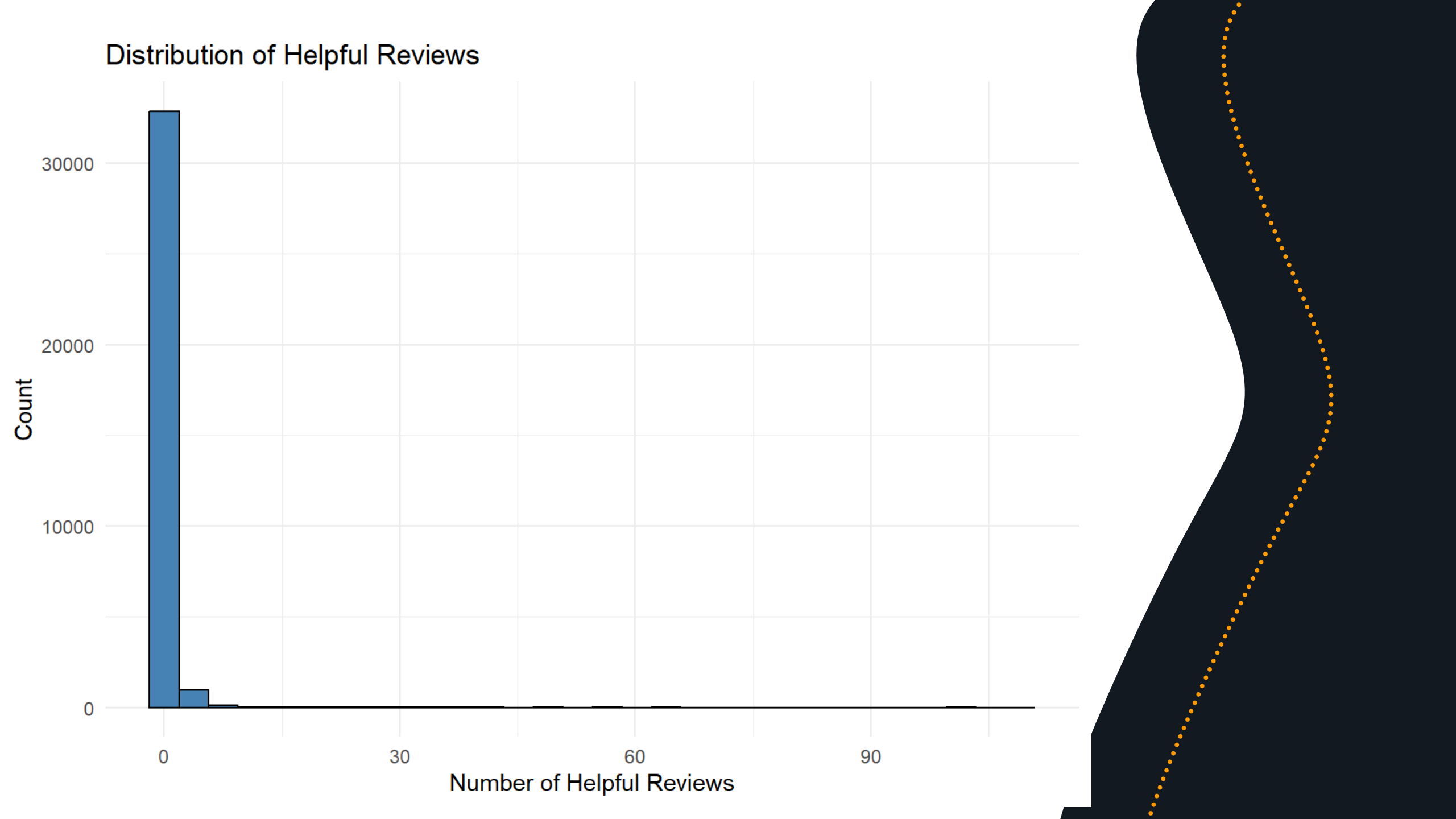

Number of reviews per brand
Count
20000
10000
0
Amazon Fire
Amazon Echo
Amazon Fire Tv
Amazon
Brand


Proportion of Recommended Products by Brand
1.00
0.75
0.50
0.25
0.00
Percentage
4.19%
6.95%
5.49%
2.95%
95.81%
93.05%
94.51%
97.05%
Amazon
Amazon Echo
Amazon Fire
Amazon Fire Tv
Brand
reviews.doRecommend
FALSE
TRUE


Review Ratings by Brand
Rating
5
4
3
2
1
Amazon
Amazon Echo
Amazon Fire
Amazon Fire Tv
Brand

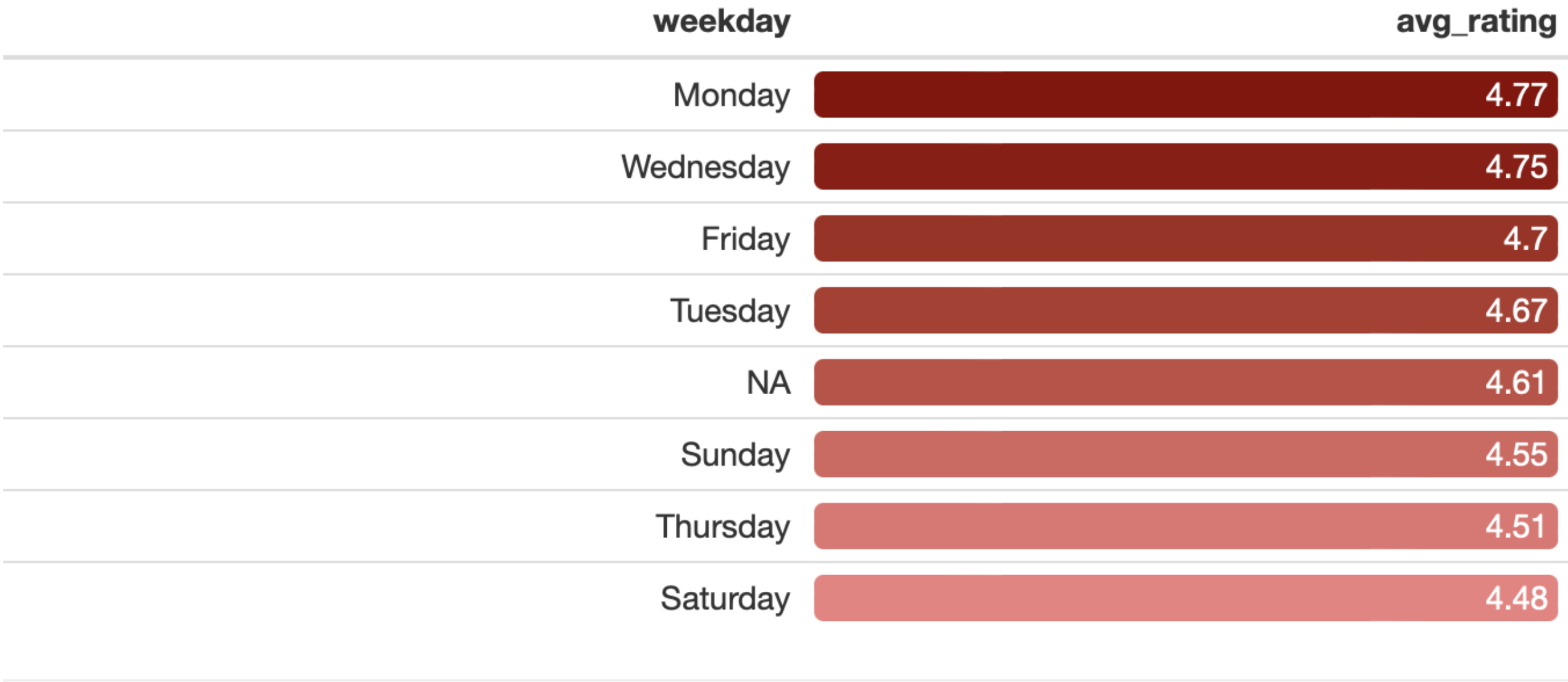

Average Rating Over Time
5
4
3
2
1
Average Rating
2015
2016
2017
2018
Date


Relationship Between Helpful Votes and Rating
Helpful Votes
90
60
30
0
1
2
3
4
5
Rating
Recommendation
FALSE
TRUE

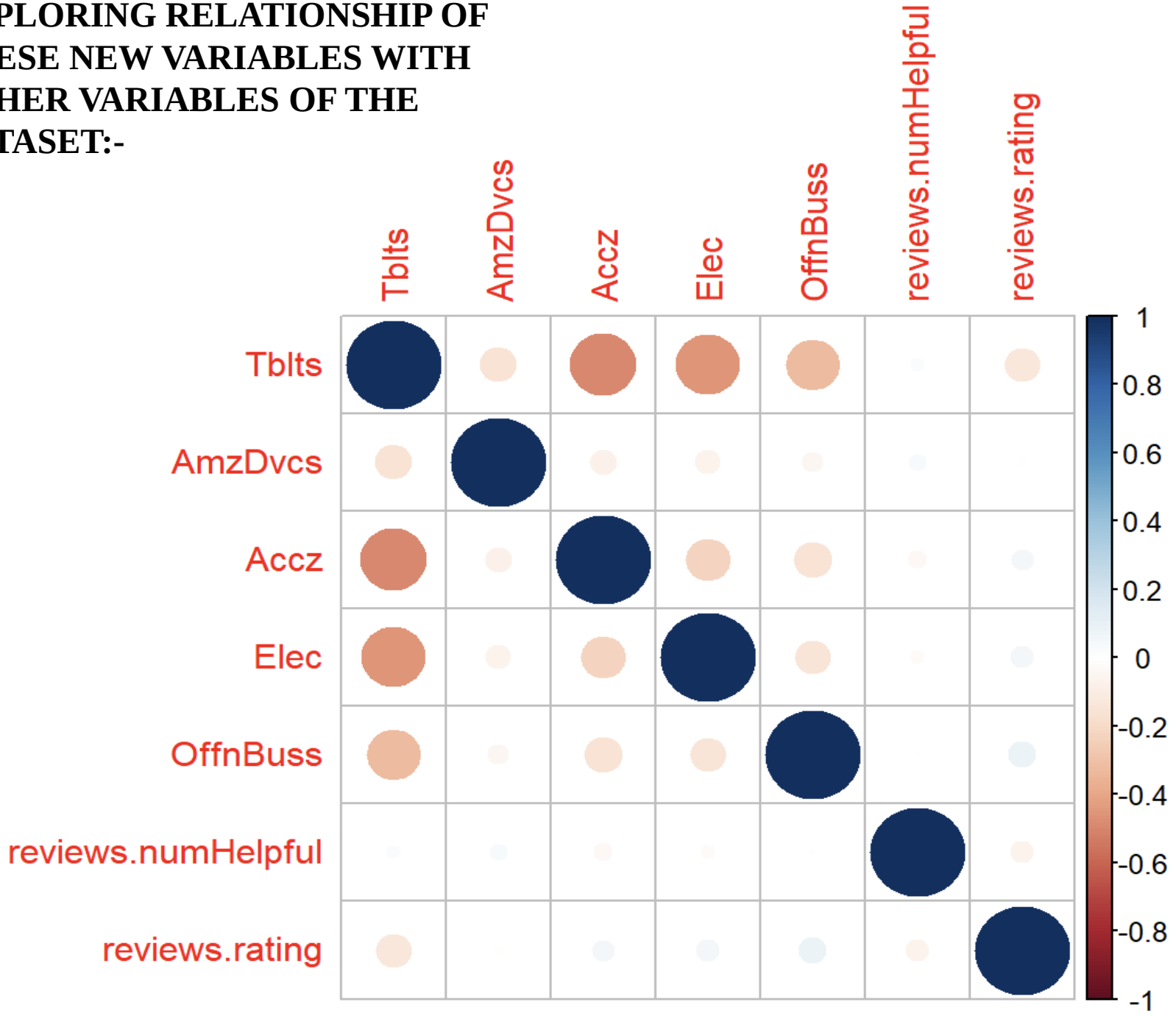

Distribution of Helpful Reviews
Count
30000
20000
10000
0
0
30
60
90
Number of Helpful Reviews

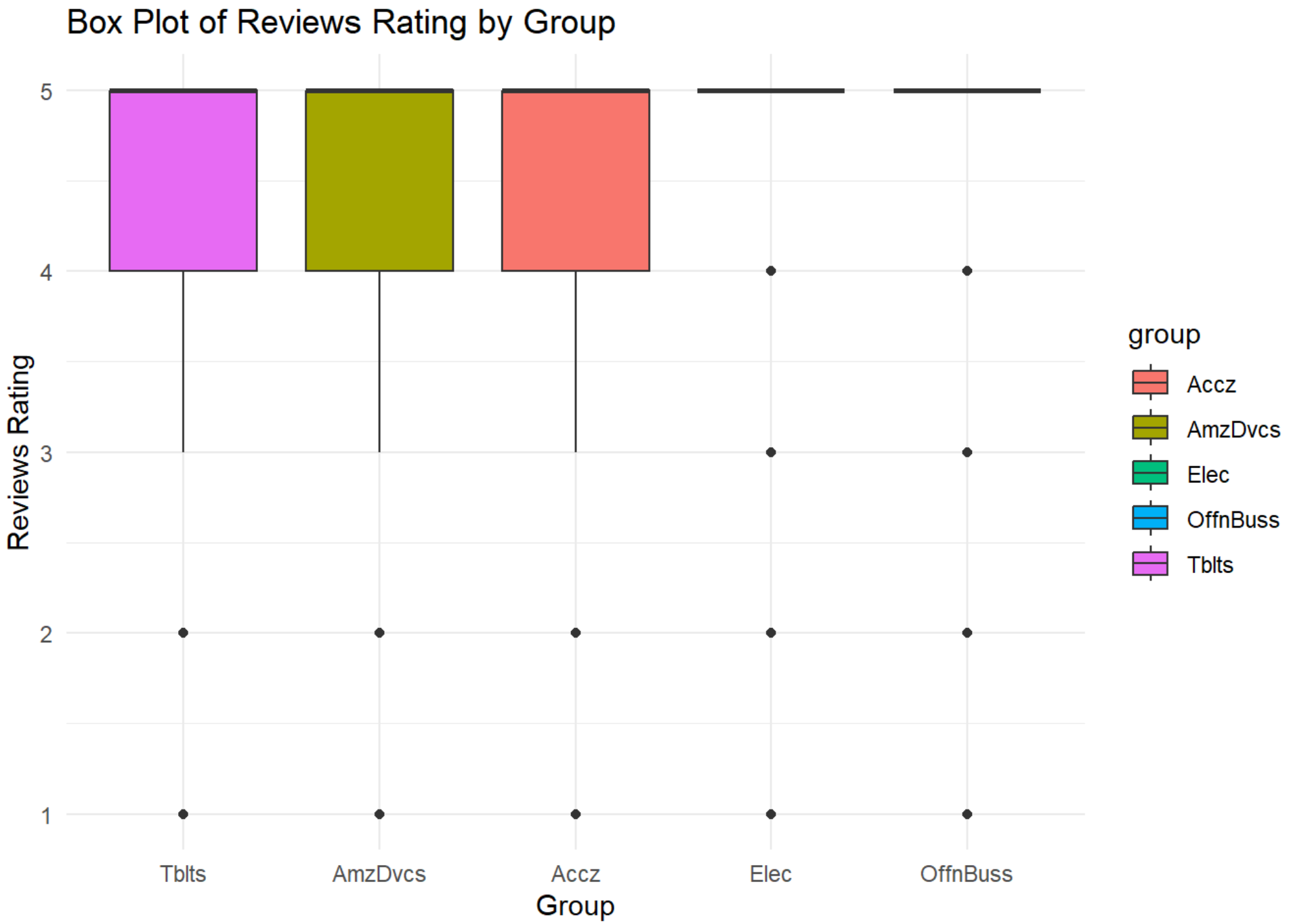
MOST FREQUENTLY MENTIONED WORDS IN THE REVIEWS

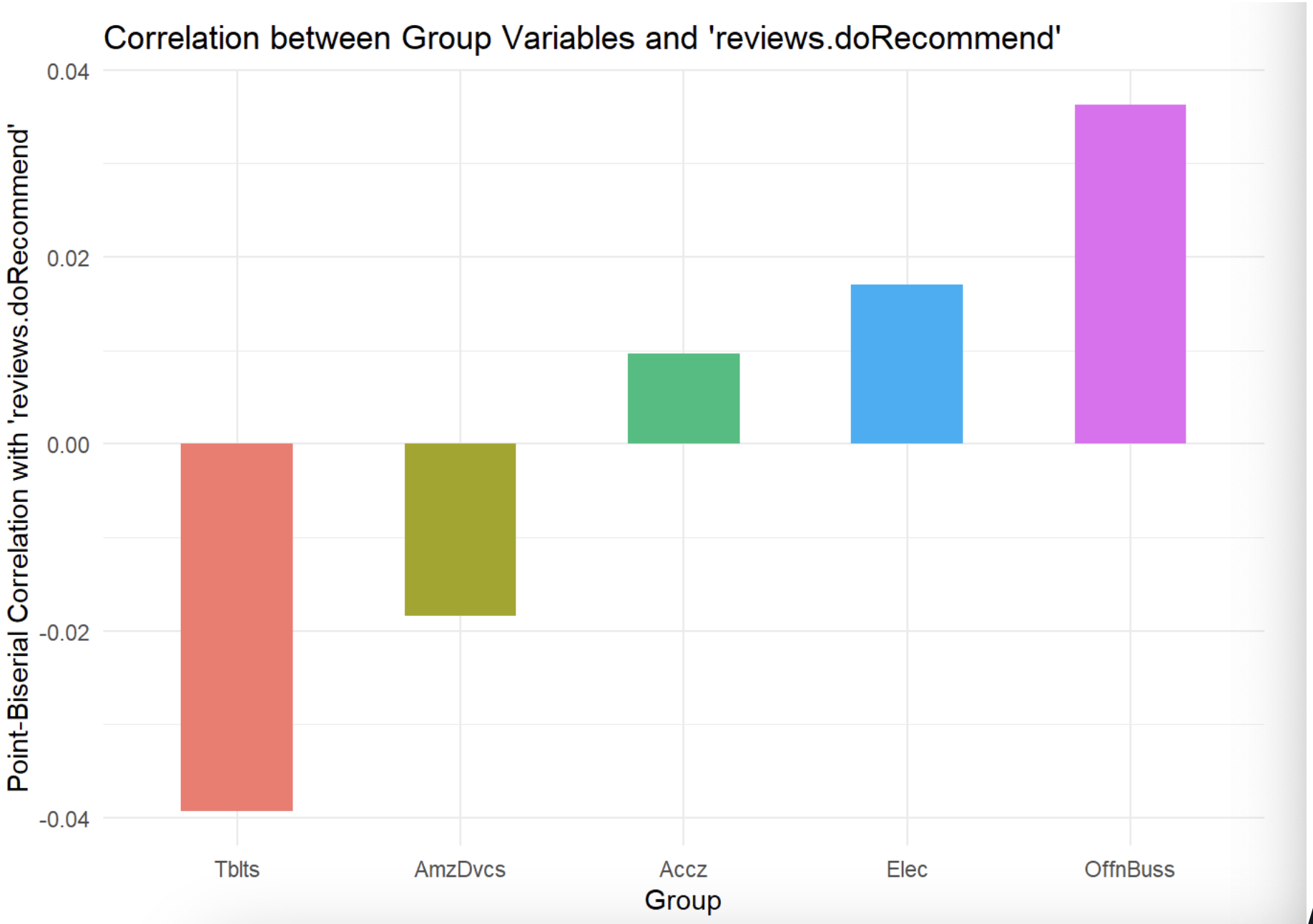
AVERAGE RATING BY DAY:

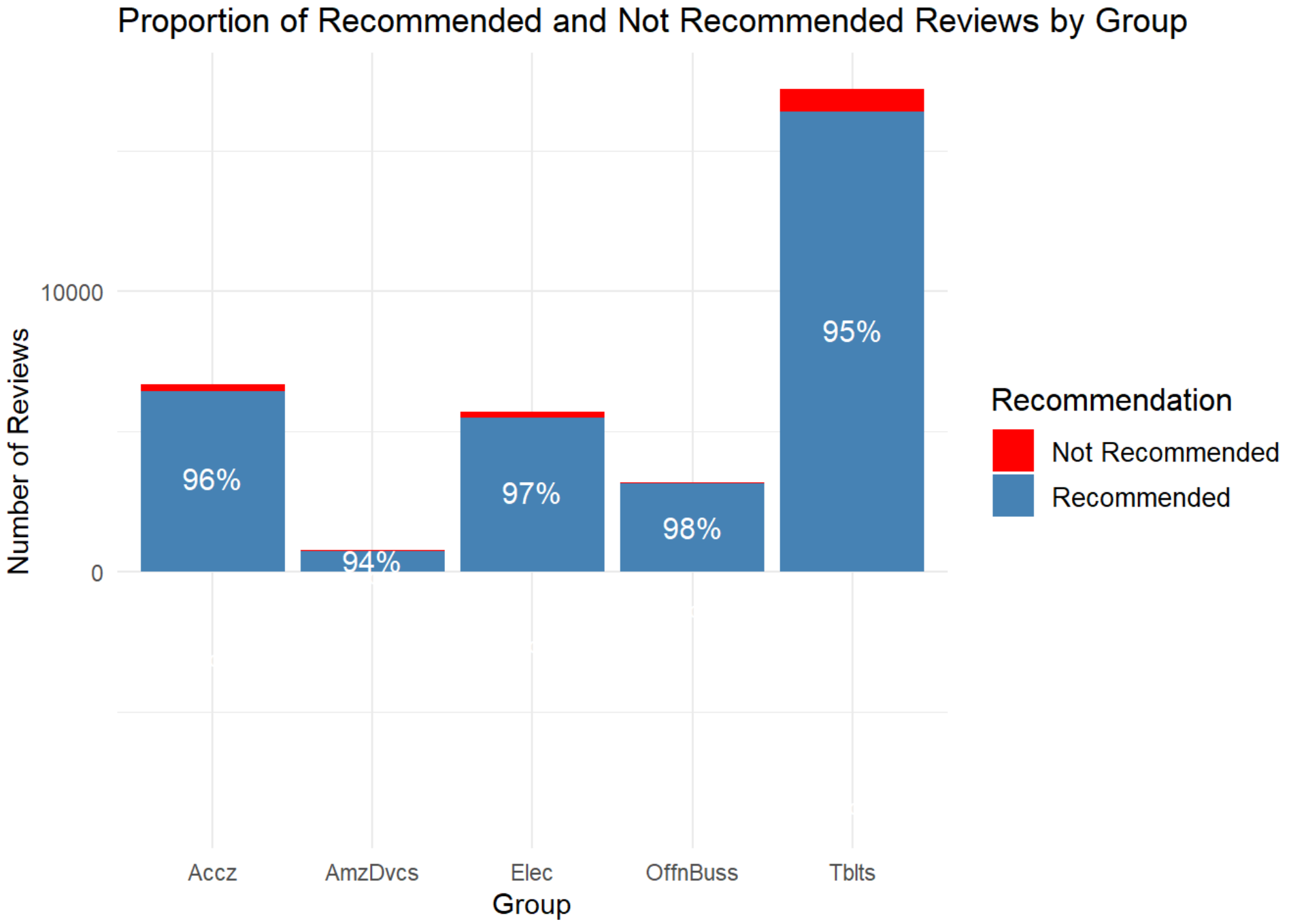


EXPLORING RELATIONSHIP OF THESE NEW VARIABLES WITH OTHER VARIABLES OF THE DATASET:-




Box Plot of Reviews Rating by Group
Reviews Rating
5
4
3
2
1
Tblts
AmzDvcs
Accz
Elec
OffnBuss
Group
group
Accz
AmzDvcs
Elec
OffnBuss
Tblts


Correlation between Group Variables and 'reviews.doRecommend'
Point-Biserial Correlation with 'reviews.doRecommend'
0.04
0.02
0.00
-0.02
-0.04
Tblts
AmzDvcs
Accz
Elec
OffnBuss
Group


Proportion of Recommended and Not Recommended Reviews by Group
Number of Reviews
10000
0
96%
94%
97%
98%
95%
Accz
AmzDvcs
Elec
OffnBuss
Tblts
Group
Recommendation
Not Recommended
Recommended

Top 10 Products with Highest Not Recommended Percentage

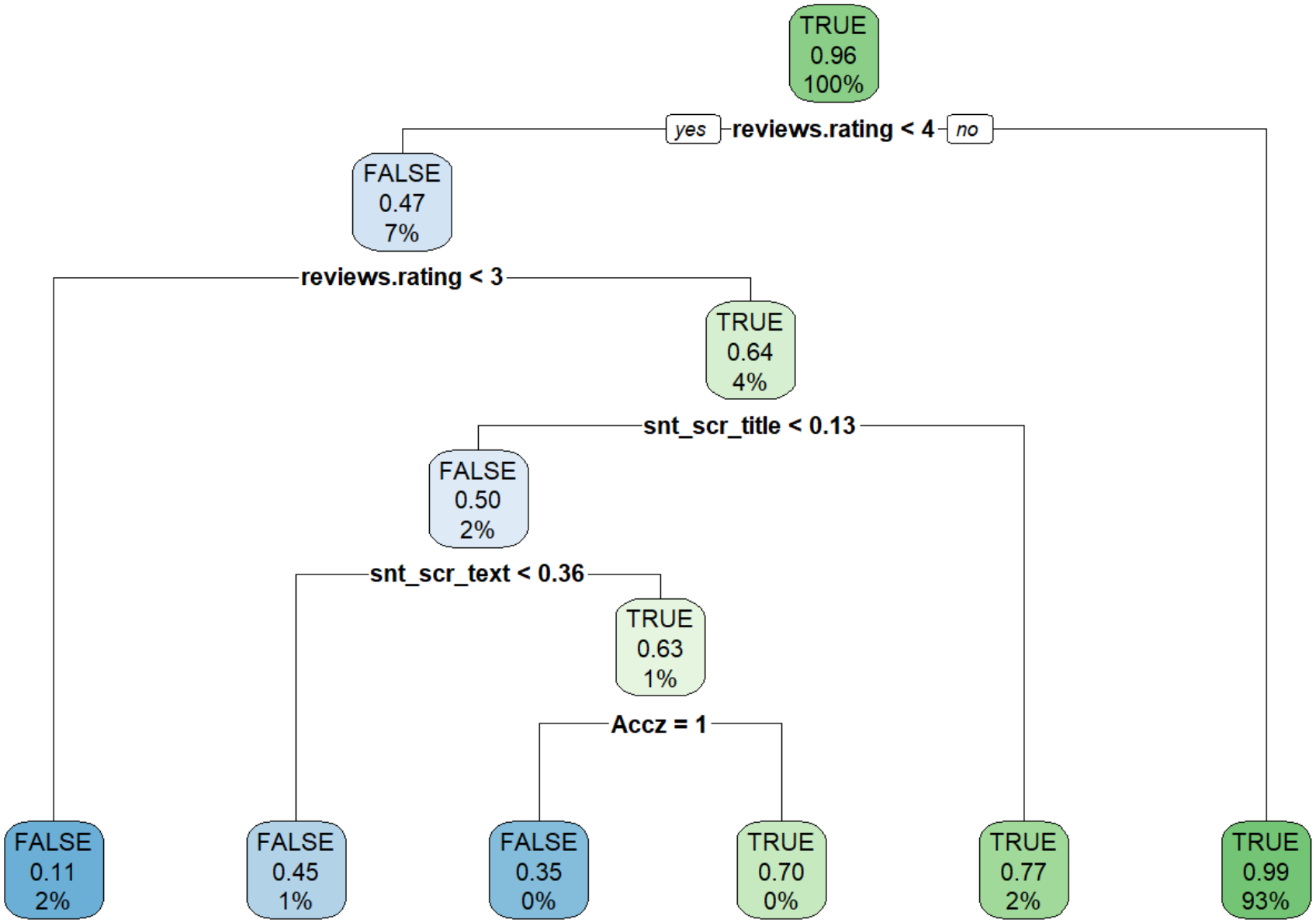
PRODUCT GROUP SUMMARY: REVIEWS, RECOMMENDATIONS, AND AVERAGE RATINGS (PIVOT TABLE):-

Tblts	AmzDvcs	Accz	Elec	Group	Percent_Recommended	Not_Recommended_Percentage	Average_Rating	Total_Reviews
0	0	0	0	NA	97.78157	2.218430	4.730375	586
0	0	0	0	OffnBuss	98.17380	1.826197	4.772355	3176
0	0	0	1	Elec	96.68956	3.310442	4.682338	5679
0	0	1	0	Accz	96.32243	3.677574	4.670519	6662
0	1	0	0	AmzDvcs	93.52708	6.472919	4.560106	757
1	0	0	0	Tblts	95.16973	4.830272	4.490526	17204


The picture can't be displayed.

CHECKING CORRELATION B/W CONTINUOUS INDEPENDENT VARIABLES “SNT_SCR_TEXT”, “SNT_SCR_TITLE”, AND “REVIEWS.RATING”:-

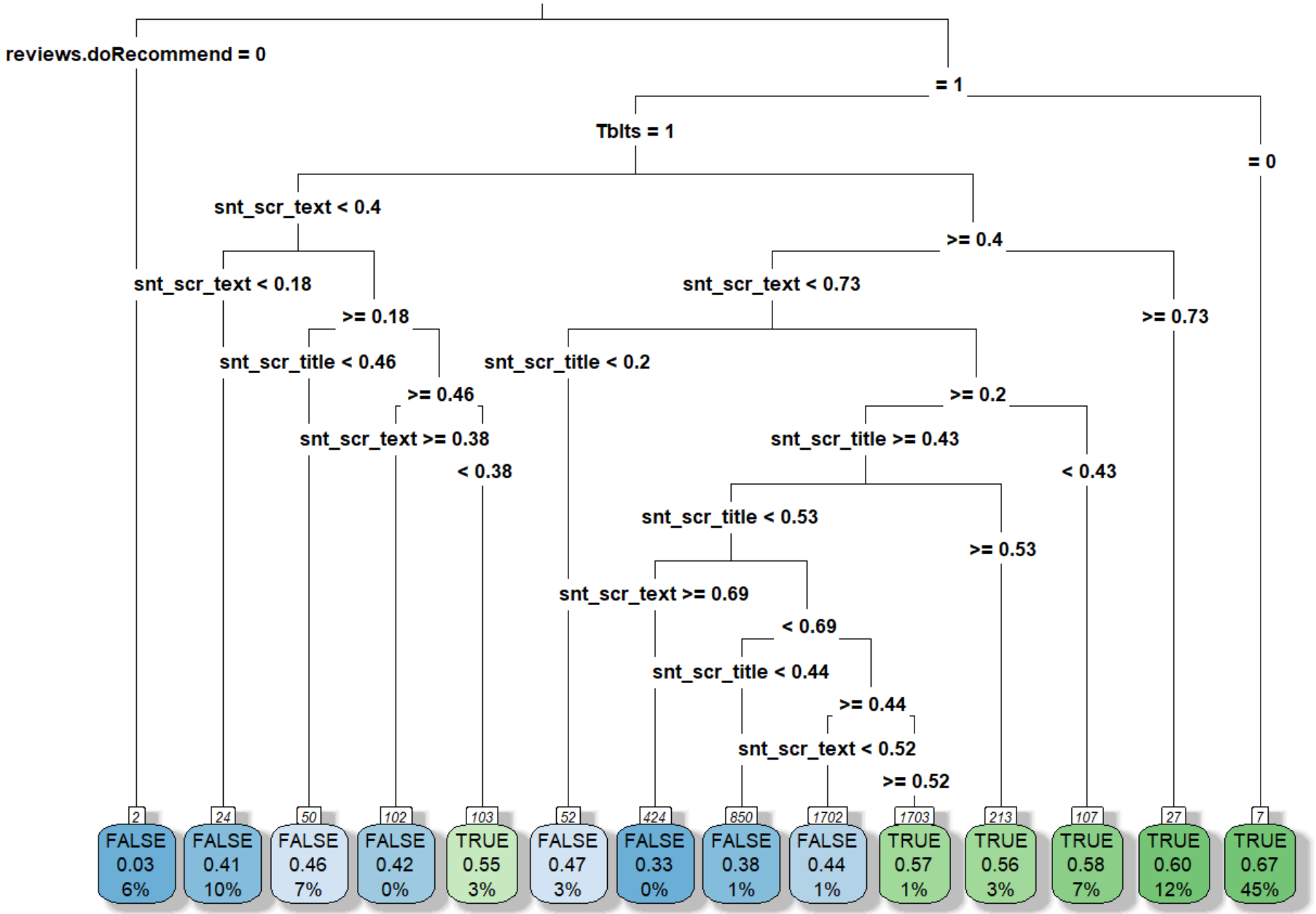
RUNNING DECISION TREES ON UNBALANCE TRAINING DATA ON BUSINESS QUESTION 1:-



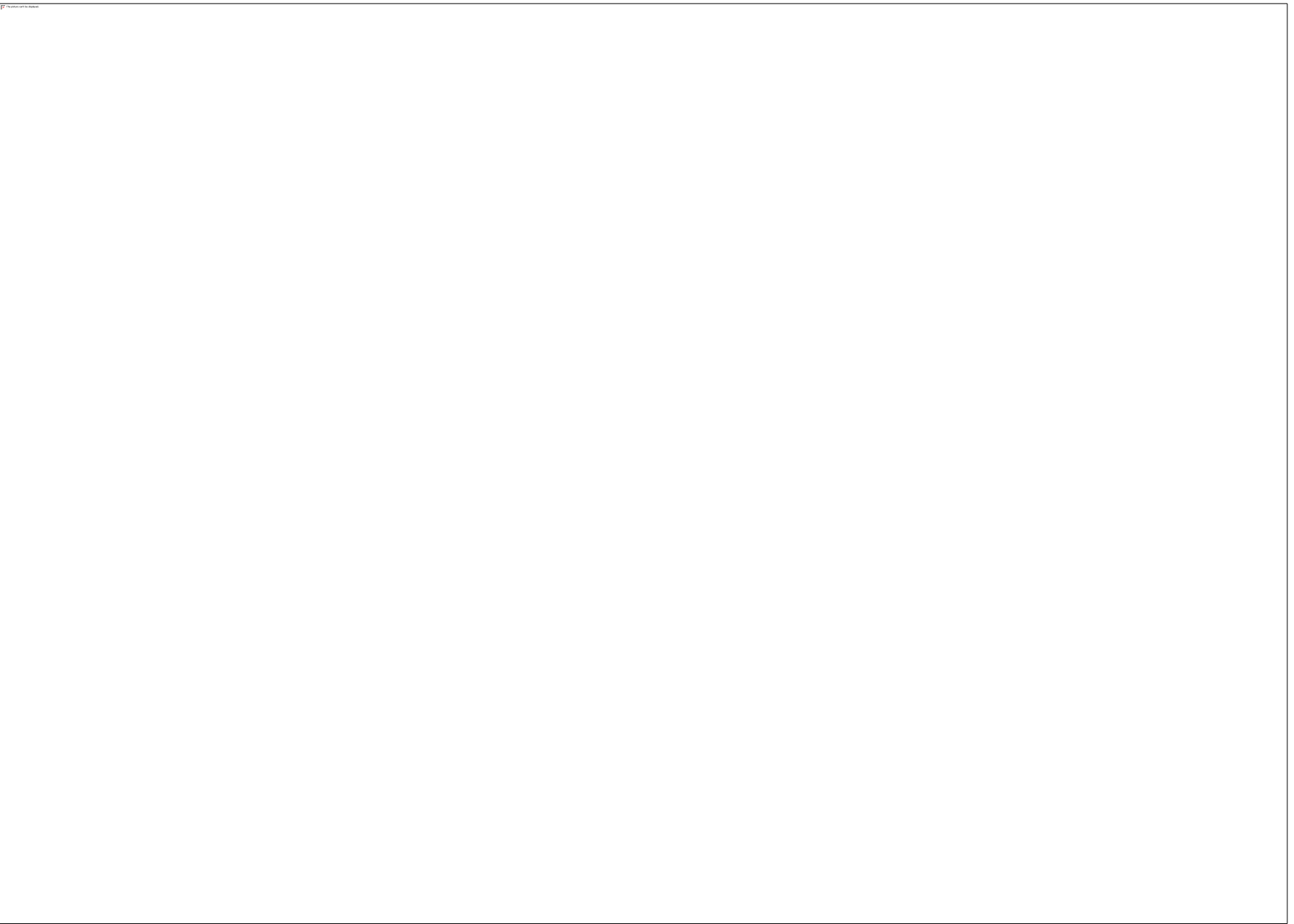
BUSINESS QUESTION 2: PREDICT WHETHER CUSTOMER WILL GIVE A RATING OF 5 OR NOT?

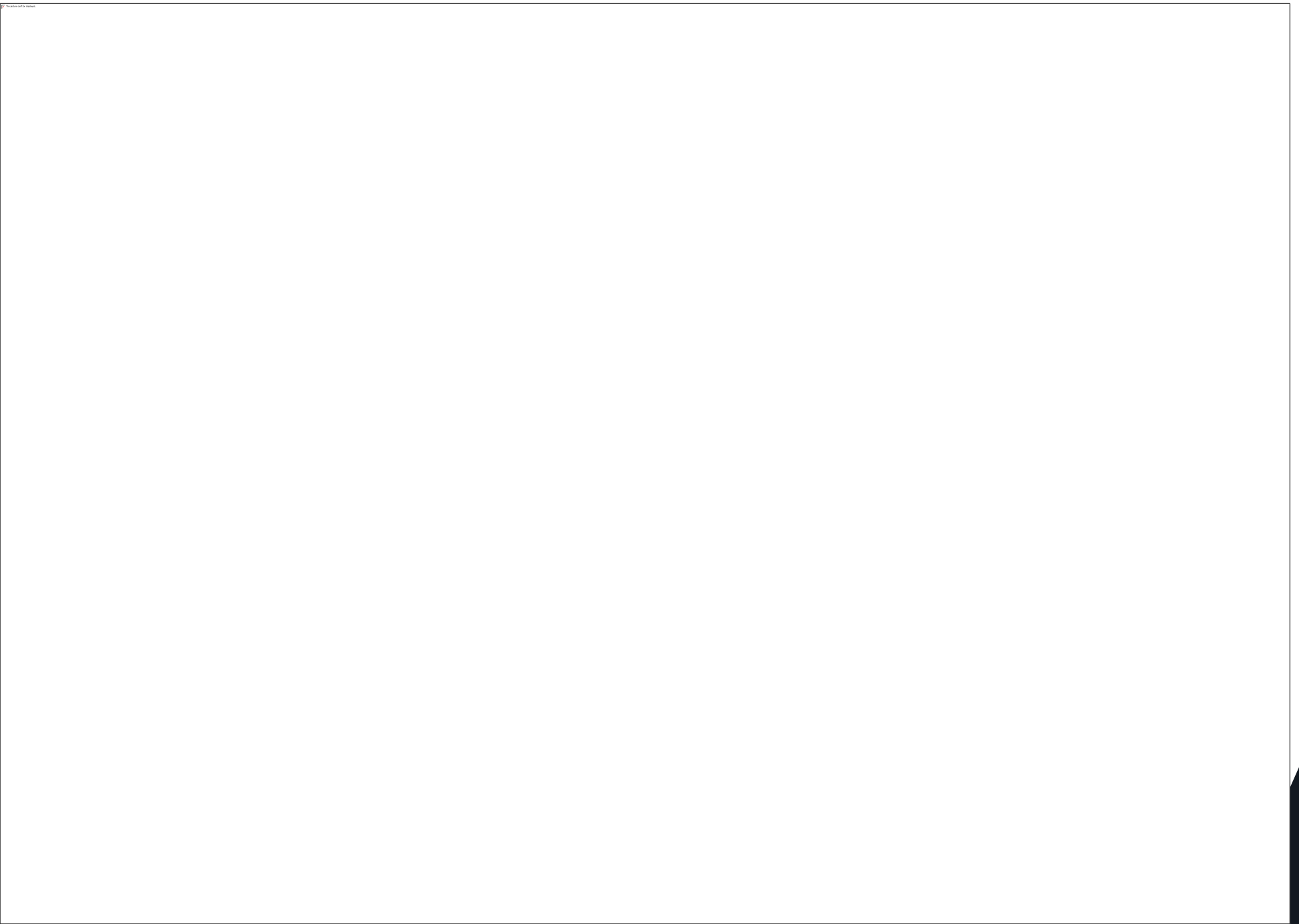
ON BALANCED DATASET:-

RUNNING DECISION TREE ON BUSINESS QUESTION :2



BUSINESS QUESTION 3:PREDICT WHICH RATING CUSTOMER WILL GIVE 5, 4 OR BELOW 4 (3-1)?




The picture can't be displayed.

RUNNING DECISION-TRESS ON BUSINESS QUESTION 3: