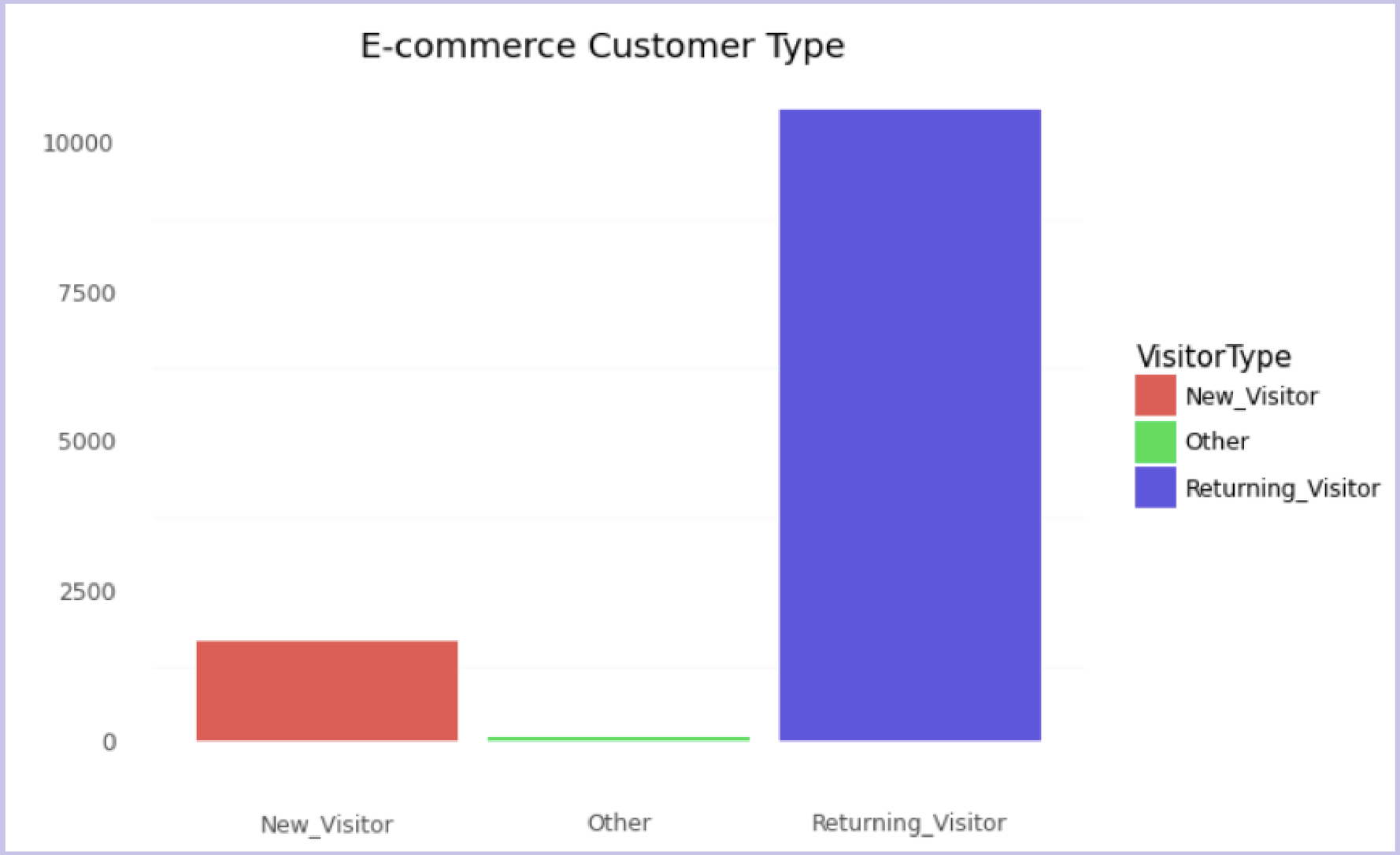
Liz Lyon
Chapman University

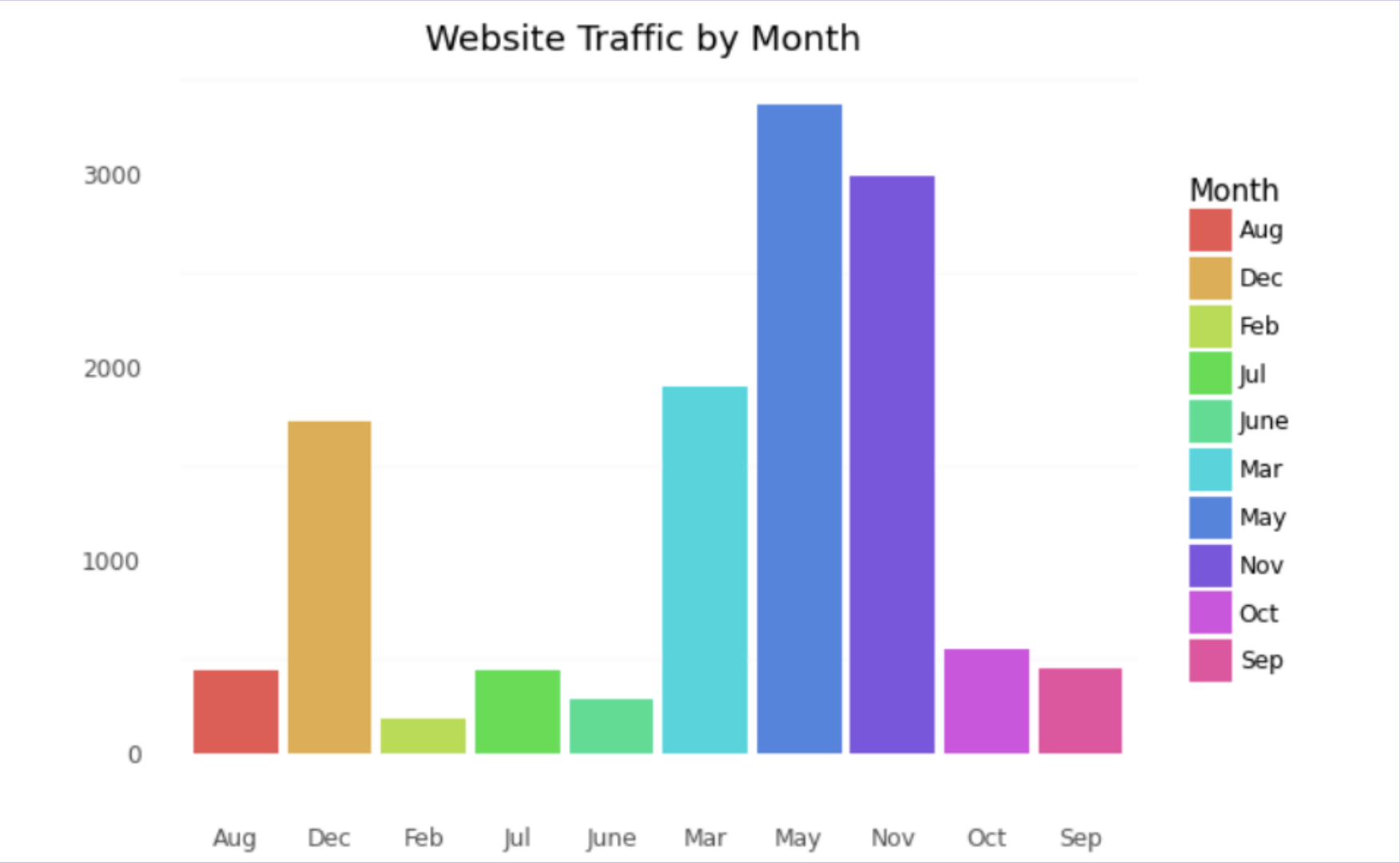
E-COMMERCE ANALYSIS

CUSTOMER SEGMENTS



https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset

MONTHLY TRAFFIC



https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset

CONSUMPTION BEHAVIOR

https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset

LET'S DISCUSS — AND ANALYZE

Question 1

Which variables have the strongest impact on a successful customer transaction (Revenue column)?

Question 2

Are there any definitive clusters between the ExitRate and BounceRate columns? If so, are there any outliers?

Question 3

Are there any methods to reduce the dimensionality of the continuous variables in your data set? If so, which method, how can you tell, and how many variables do you need to retain 80% of the original variance?

Which variables have the strongest impact on a successful customer transaction (Revenue column)?

Methods:

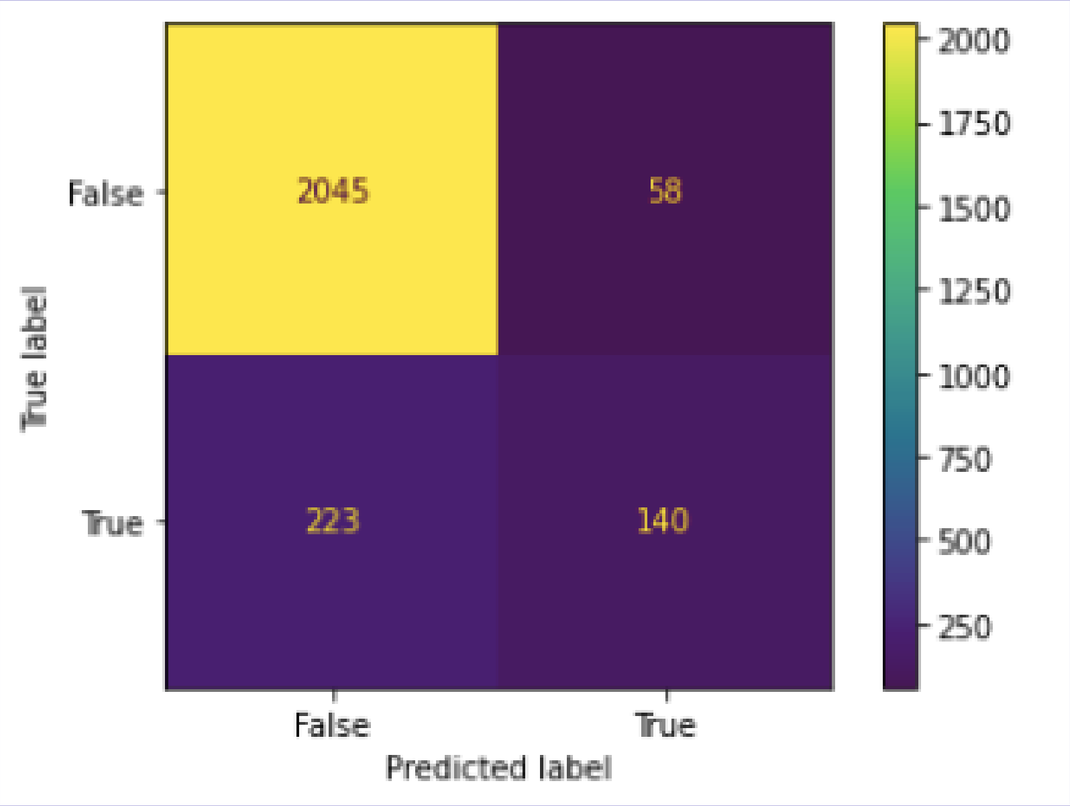
- Separated X & Y
- 20/80 TTS
- Z-scored
- Logistic Regression Model
- Predicted Vals
- Accuracy Score
- Confusion Matrix
- Coefs in Odds

Q1

Q1) Findings

Accuracy score: 0.8860502838605029

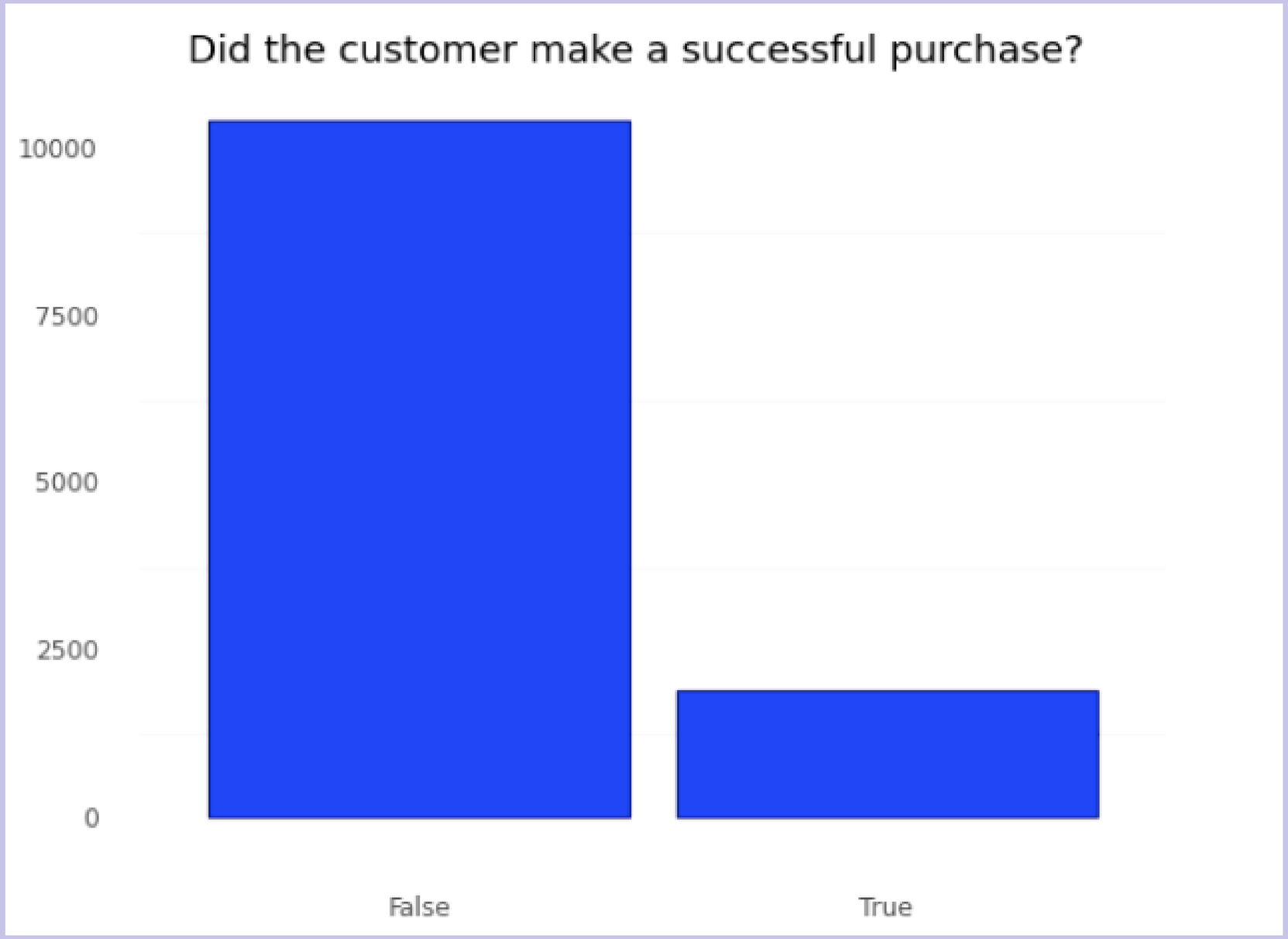
Confusion matrix:



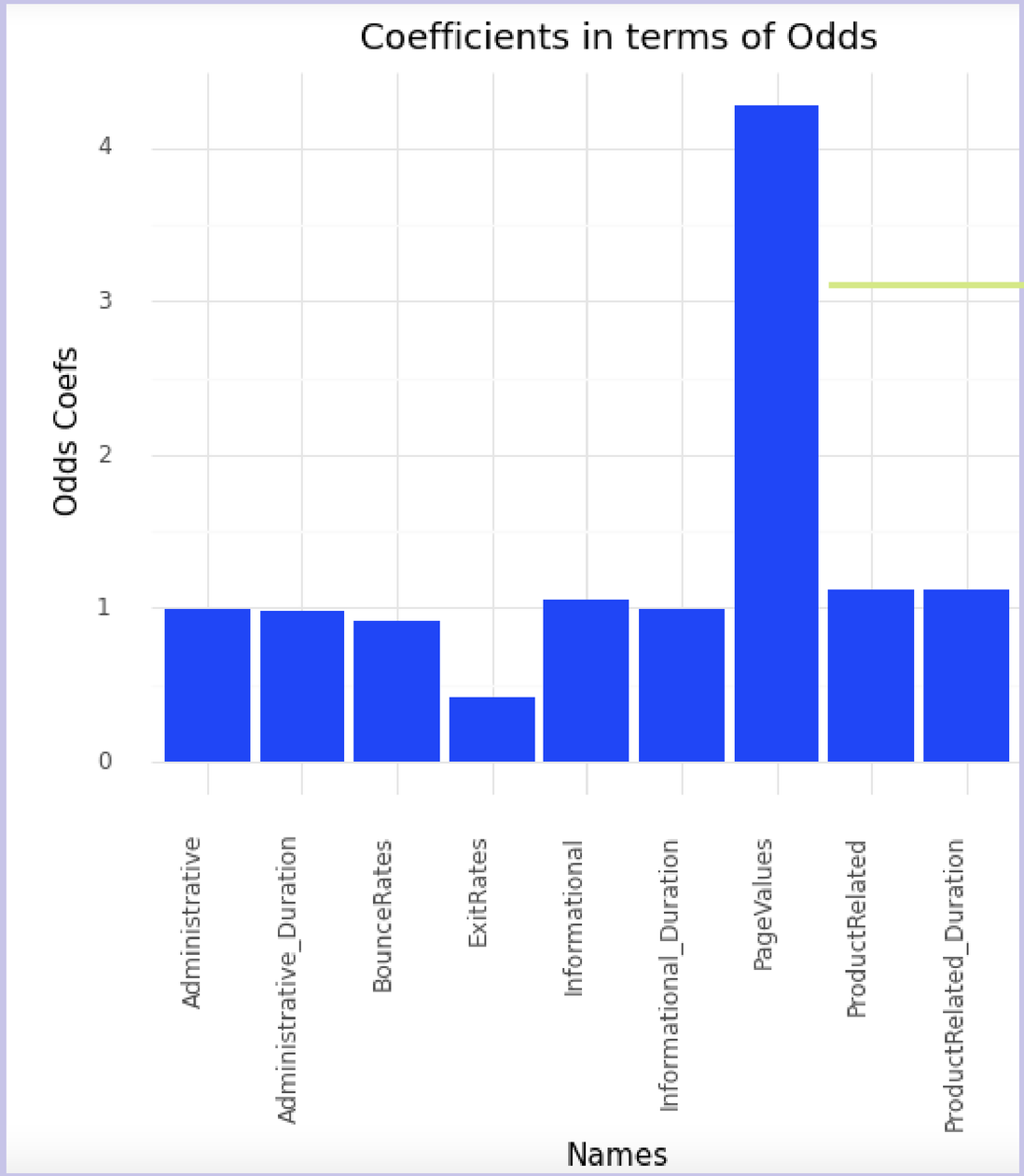
Coefs:

	Coefs	Names	Odds Coefs
0	-0.003486	Administrative	0.996520
1	-0.023688	Administrative_Duration	0.976591
2	0.053742	Informational	1.055213
3	-0.002796	Informational_Duration	0.997208
4	0.110005	ProductRelated	1.116284
5	0.118537	ProductRelated_Duration	1.125849
6	-0.089224	BounceRates	0.914641
7	-0.882084	ExitRates	0.413919
8	1.453265	PageValues	4.277055
9	-2.250173	intercept	0.105381

Q1) GGPlot


Did the customer make a successful purchase?
10000
7500
5000
2500
0
False
True

Q1) GGPlot



greatest impact

Are there any definitive clusters between the ExitRate and BounceRate columns? If so, are there any outliers?

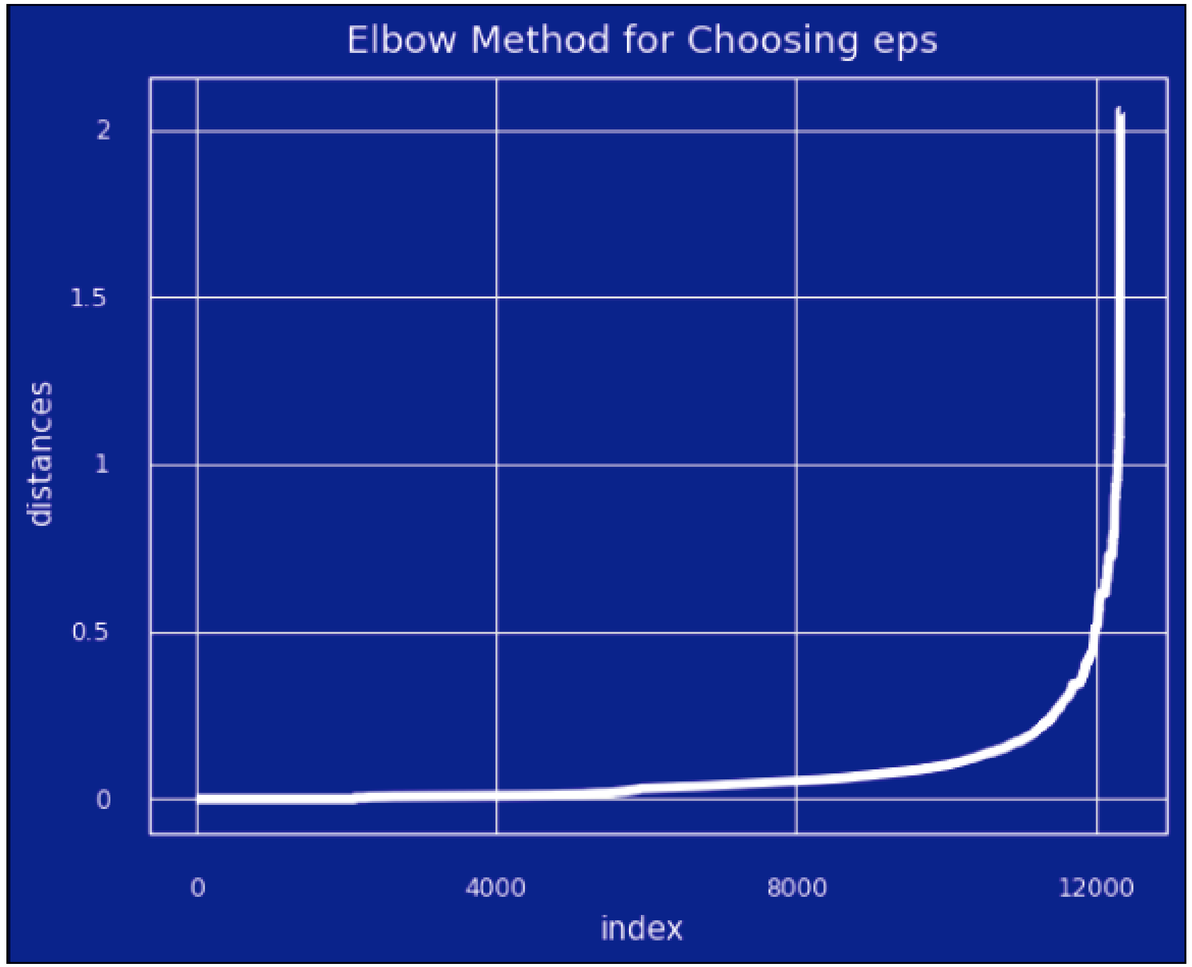
Methods:

- Create DF
- Z-scored
- Mins & Eps using Elbow Method
- DBSCAN
- Silhouette Score

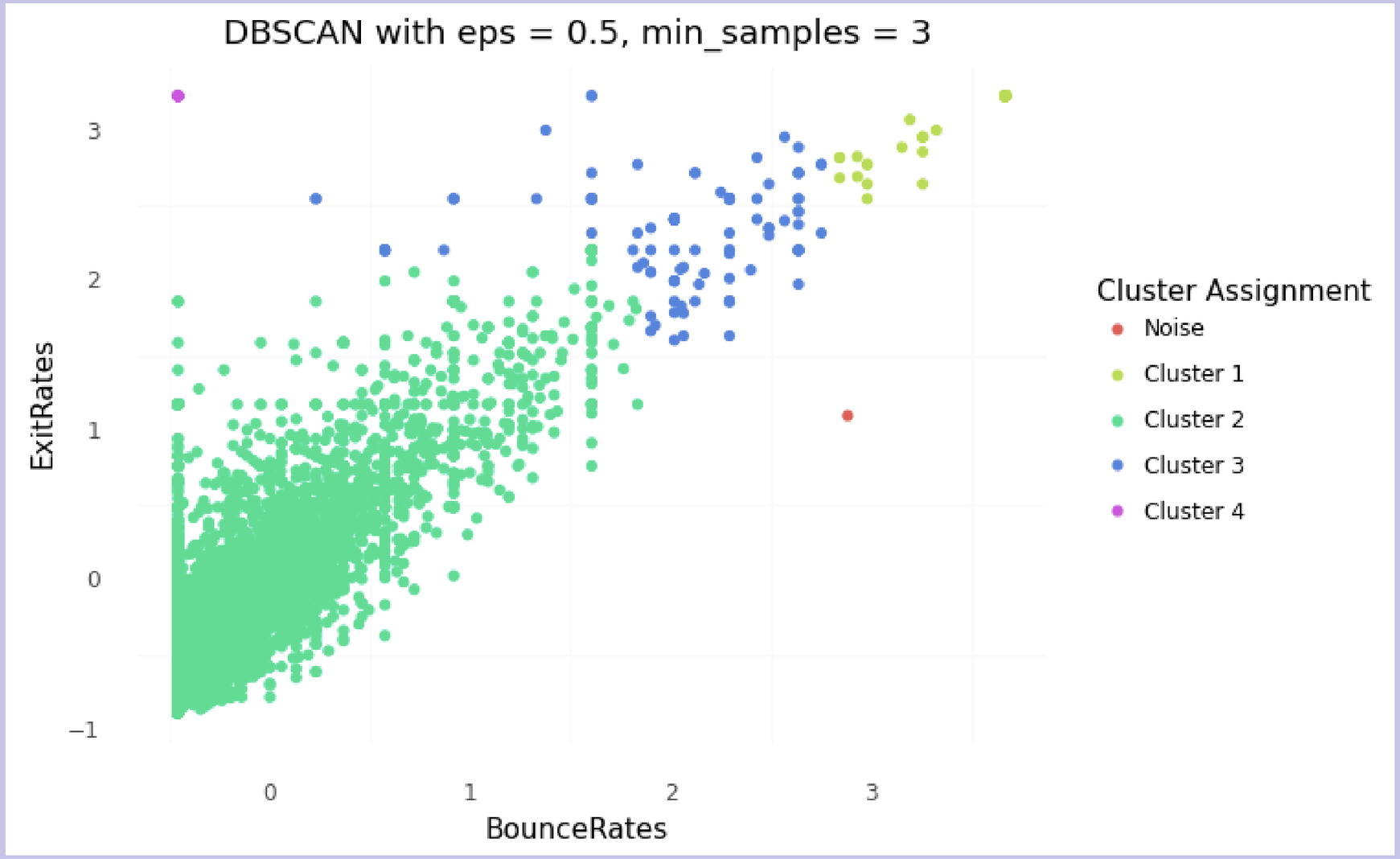
Q2

Q2) GGPlot


```
eps = 0.50, min_samples = 150
```



Q2) GGPlot & Findings




```
clustered silhouette score: 0.7704227278237737
overall silhouette score: 0.766578779029016 2
```

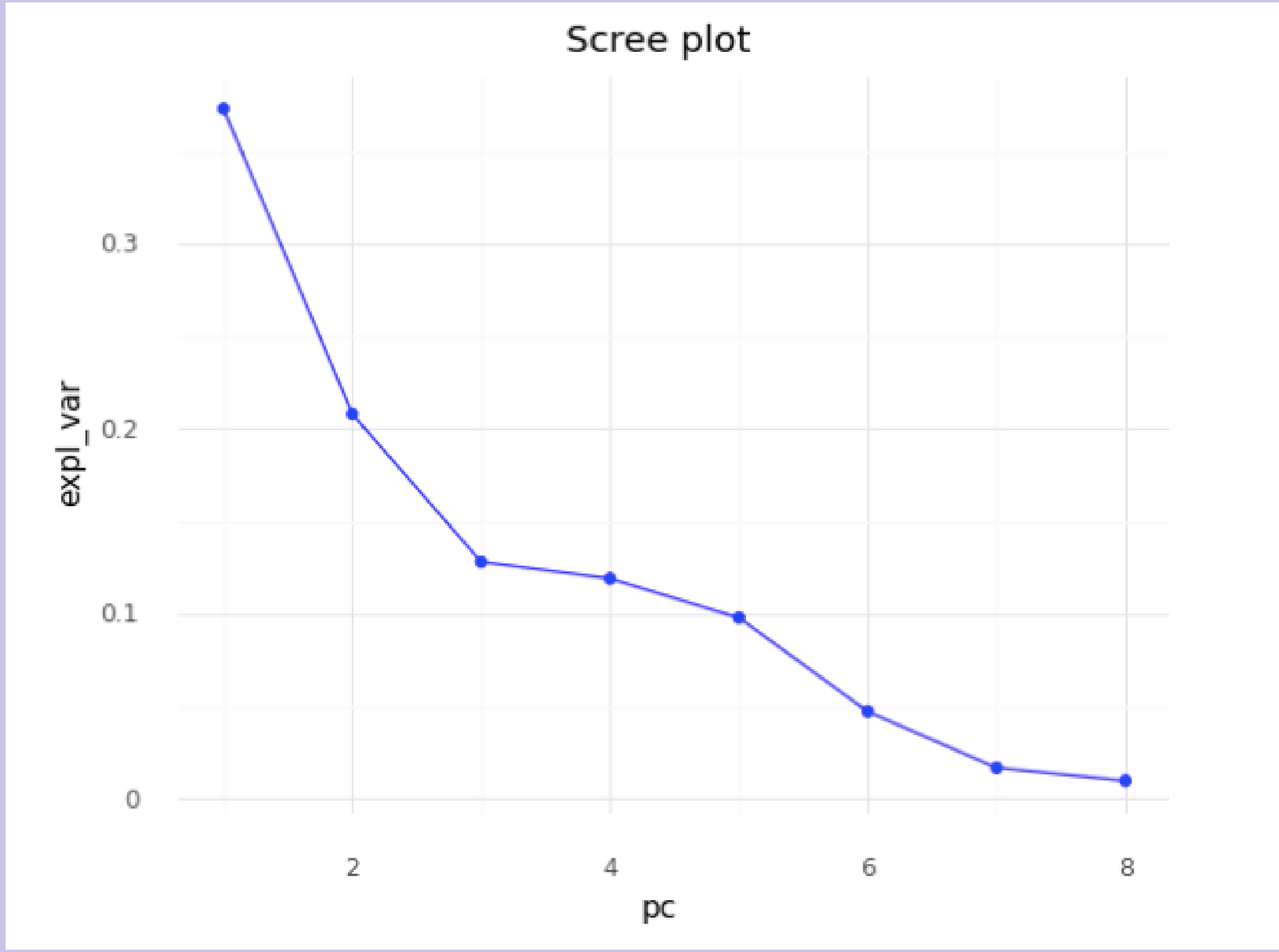
Are there any methods to reduce the dimensionality of the continuous variables in your data set? If so, which method, how can you tell, and how many variables do you need to retain 80% of the original variance?

Methods:

- Selected predictor columns
- Z-scored
- PCA
- DF of PC's
- Scree & 80% Variance Plot
- Logistic Regression
- PC score

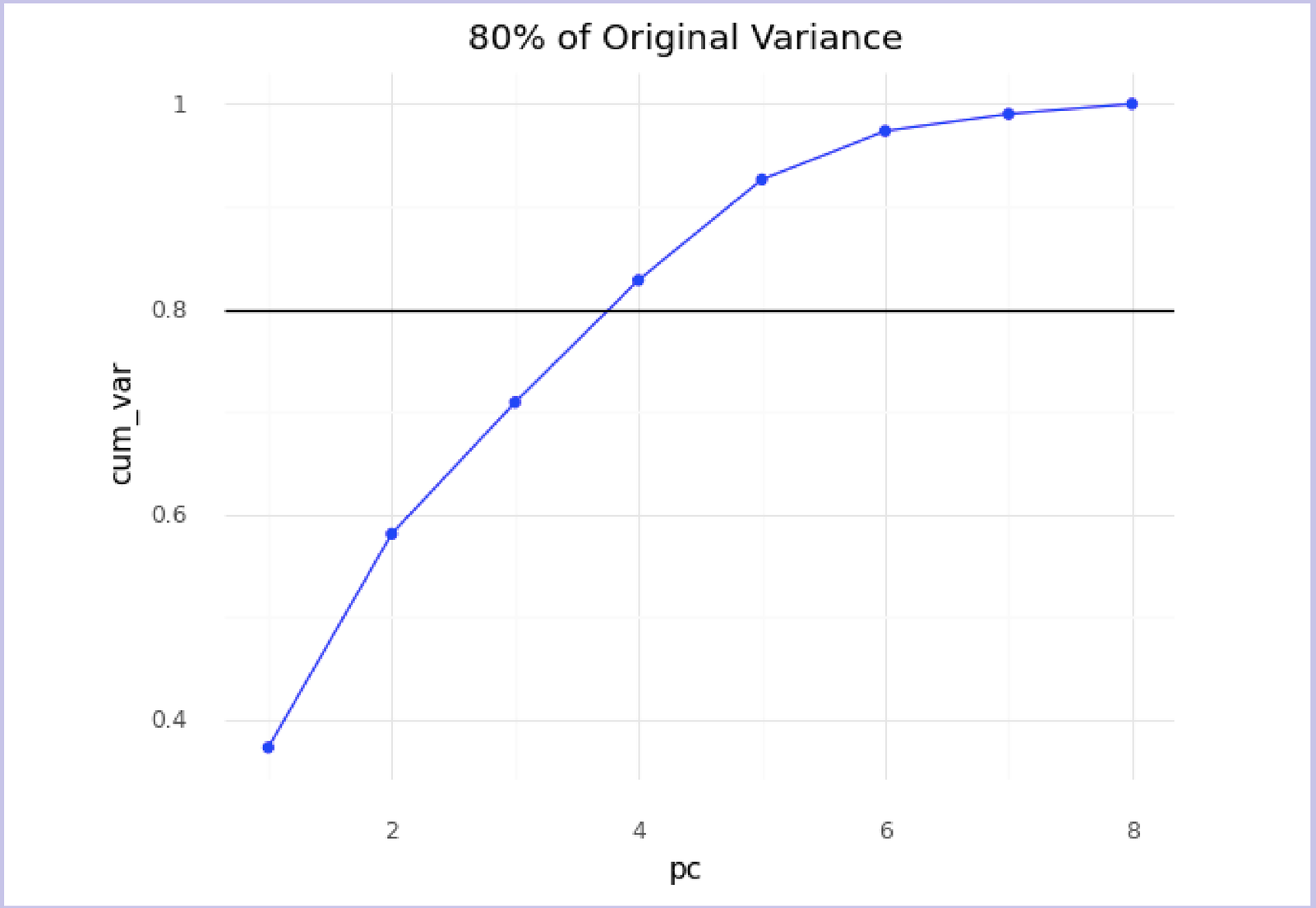
Q3

Q3) GGPlot



Q3) GGPlot & Findings

```
All data:  0.8840227088402 27
6 PCs:     0.8816707218167072
4 PCs:     0.8814274128142742
```



EXEC SUMMARY

Finding 1

The variables that are positively impactful on Revenue include Informational, ProductRelated, ProductRelated_Duration, PageValues. In terms of Odds, PageValues is dominating.

Finding 2

DBSCAN picked up on 4 cluster groups and one data point classified as 'Noise'.

Finding 3

PCA was used to reduce variable dimensionality.

The 80% variance plot indicated that 4 PC's are needed to retain all information available within the dataset.

THE END
THANK YOU