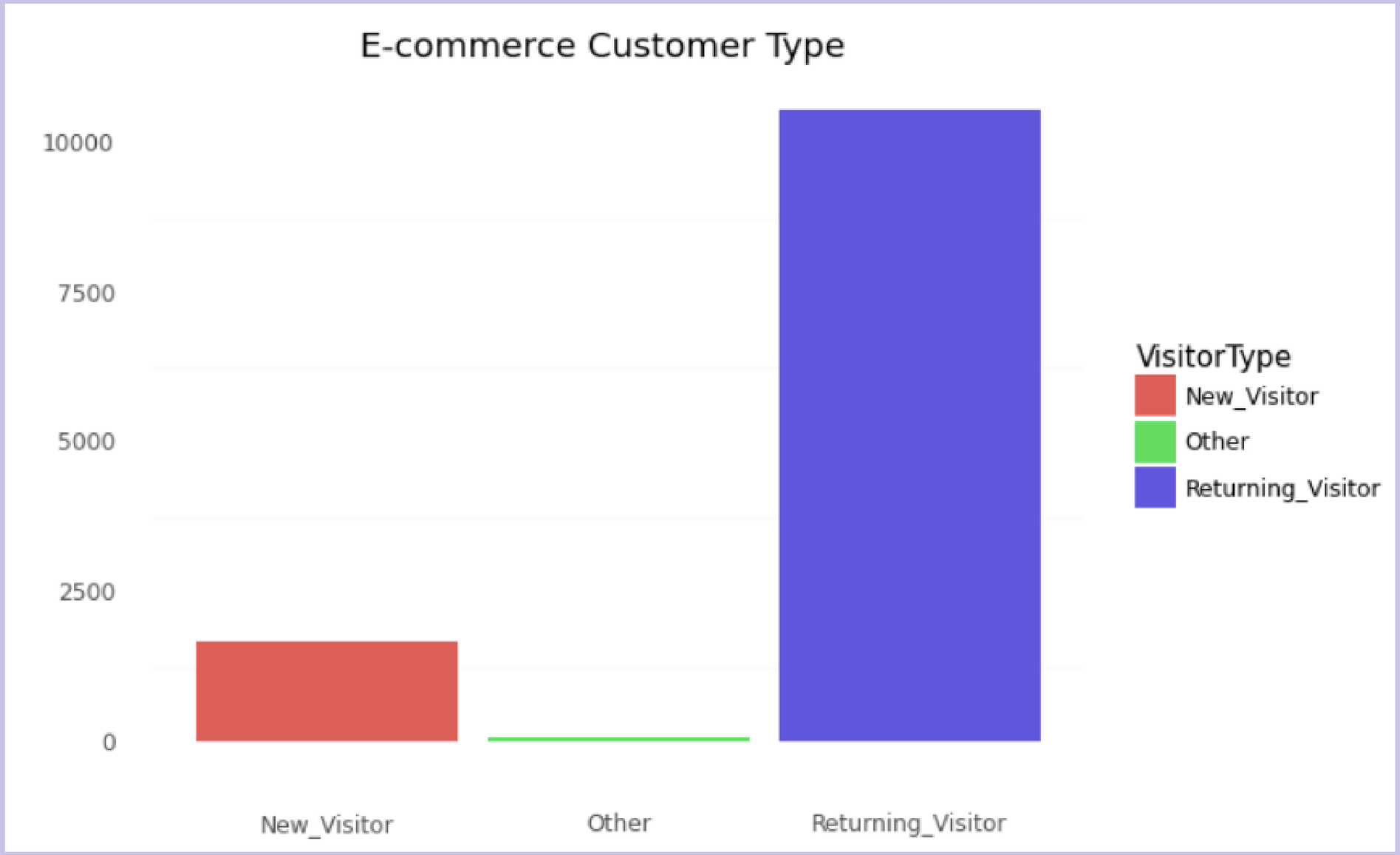
392 FINAL E-COMMERCE DATA SET

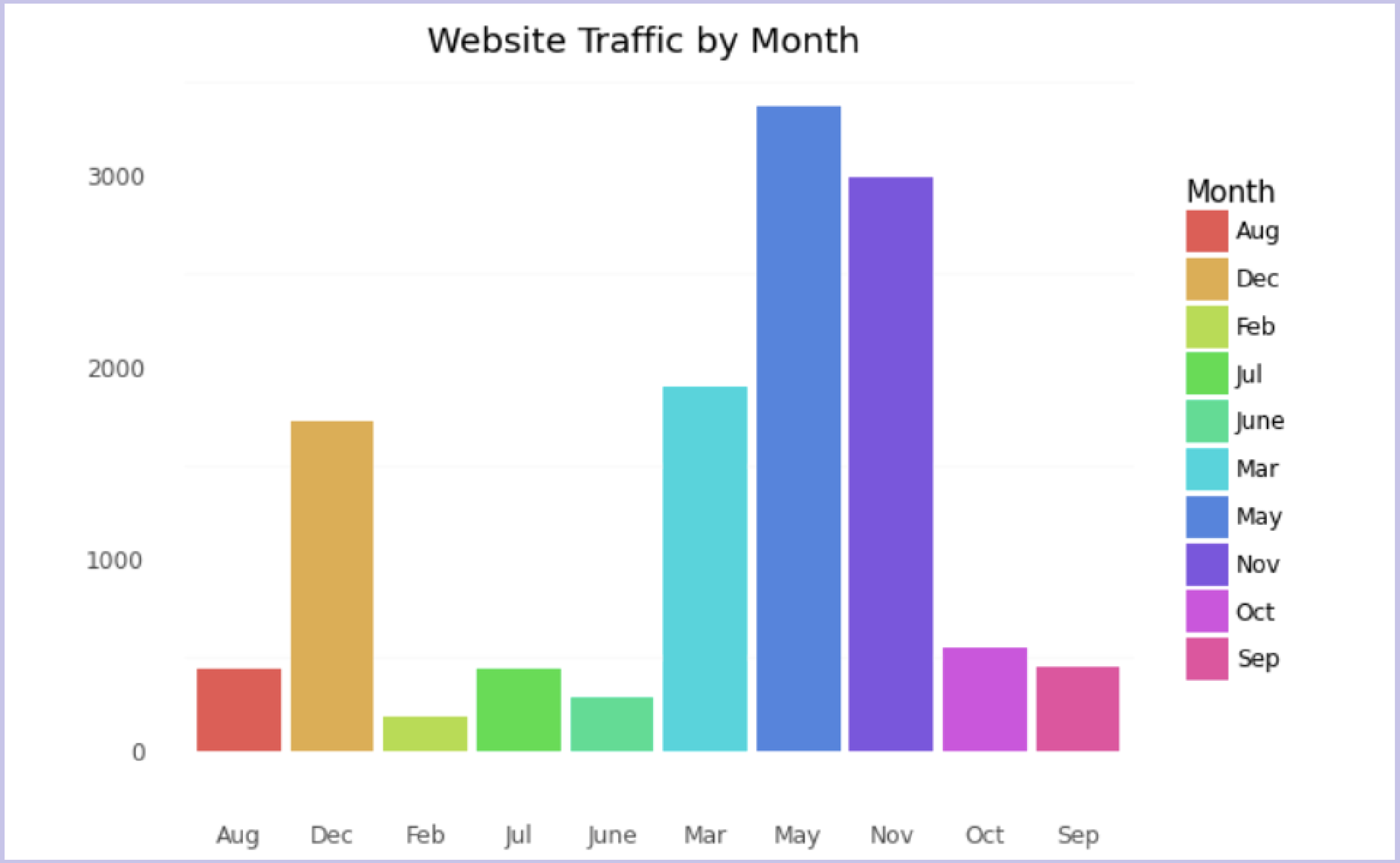
liz lyon

VISUALIZE DATA



VISUALIZE DATA

392 FINAL PROJECT — 03



VISUALIZE DATA

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:

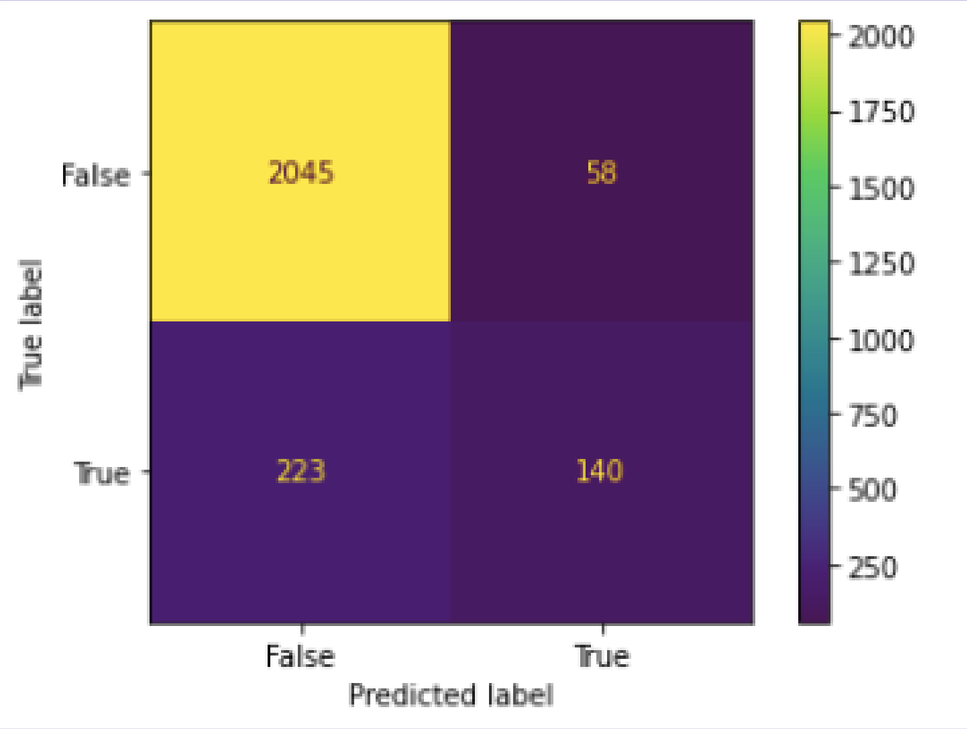
- Separated X & Y
- TTS
- Z-scored
- Logistic Regression Model
- Predicted Vals
- Accuracy Score
- Confusion Matrix
- Coefs in Odds

Q1

Q1) Findings

Accuracy score: 0.886050283860502 9

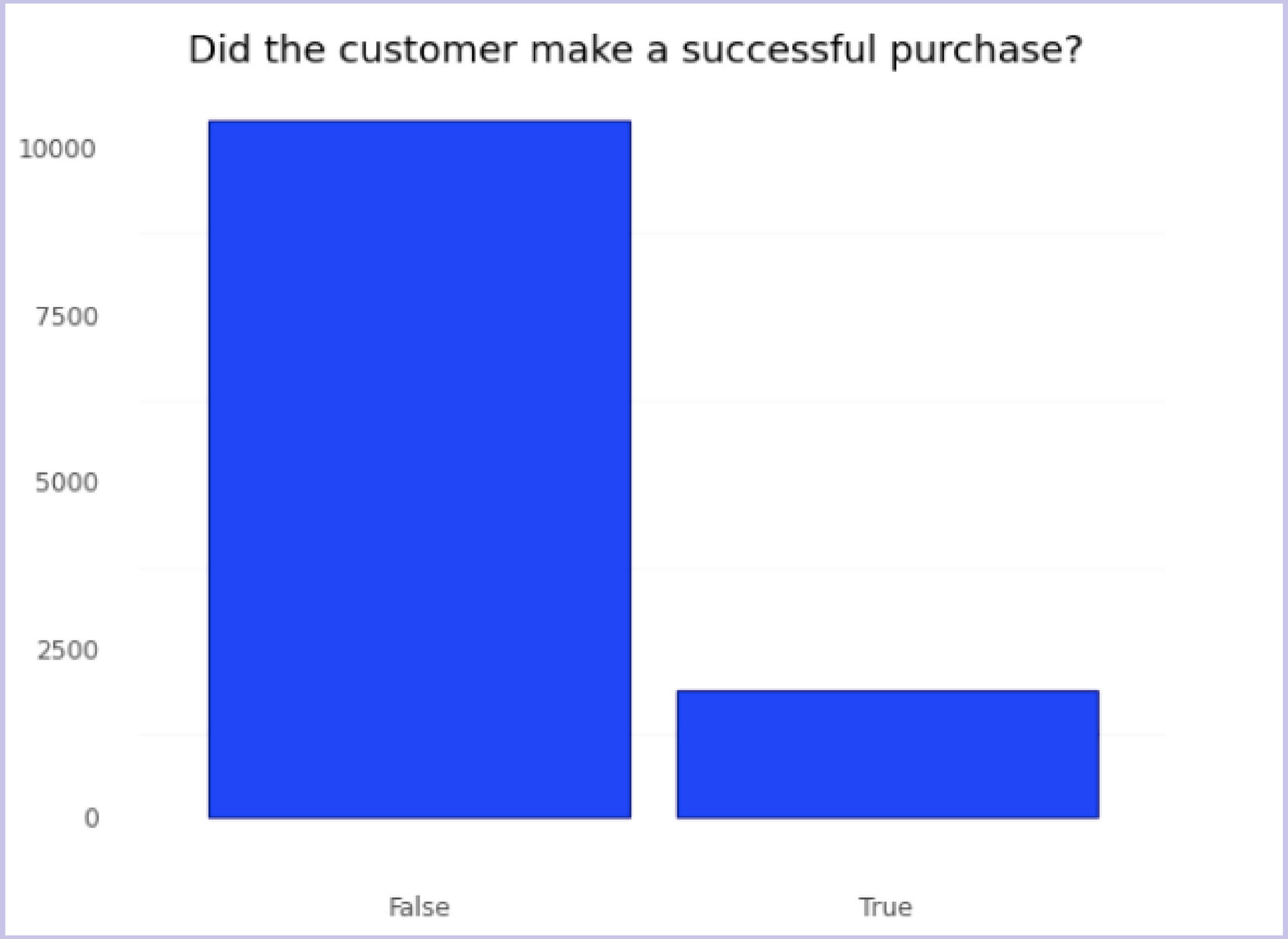
Confusion matrix:



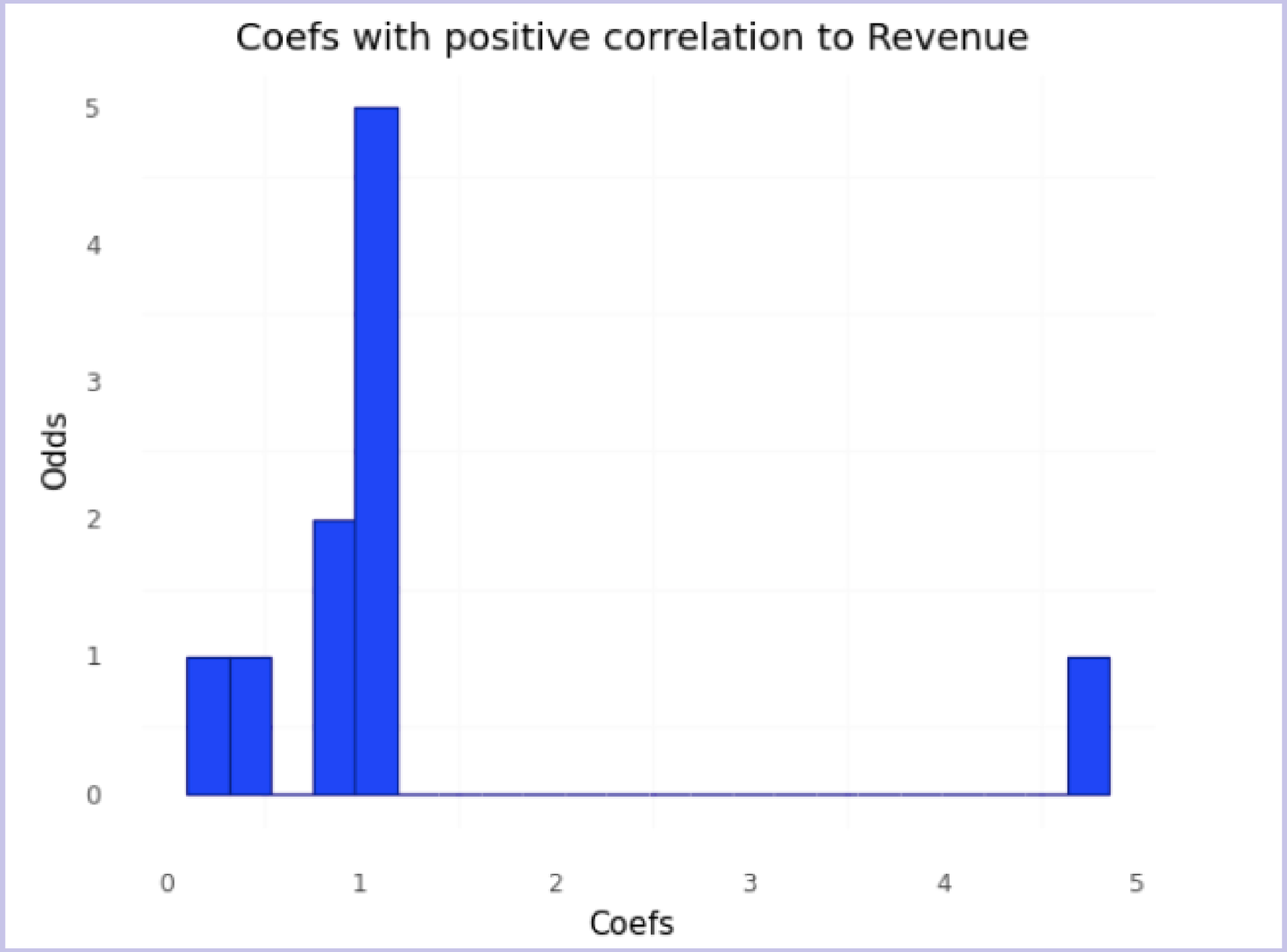
Coefs:

	Coefs	Names	Odds Coefs
0	0.019275	Administrative	1.019461
1	-0.050981	Administrative_Duration	0.950296
2	0.038008	Informational	1.038740
3	0.005686	Informational_Duration	1.005702
4	0.155905	ProductRelated	1.168715
5	0.114954	ProductRelated_Duration	1.121822
6	-0.166766	BounceRates	0.846397
7	-0.798695	ExitRates	0.449916
8	1.535403	PageValues	4.643196
9	-2.184701	intercept	0.112511

Q1) GGPlot


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

Q1) GGPlot


Coefs with positive correlation to Revenue
Odds
0
1
2
3
4
5
Coefs
0
1
2
3
4
5

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

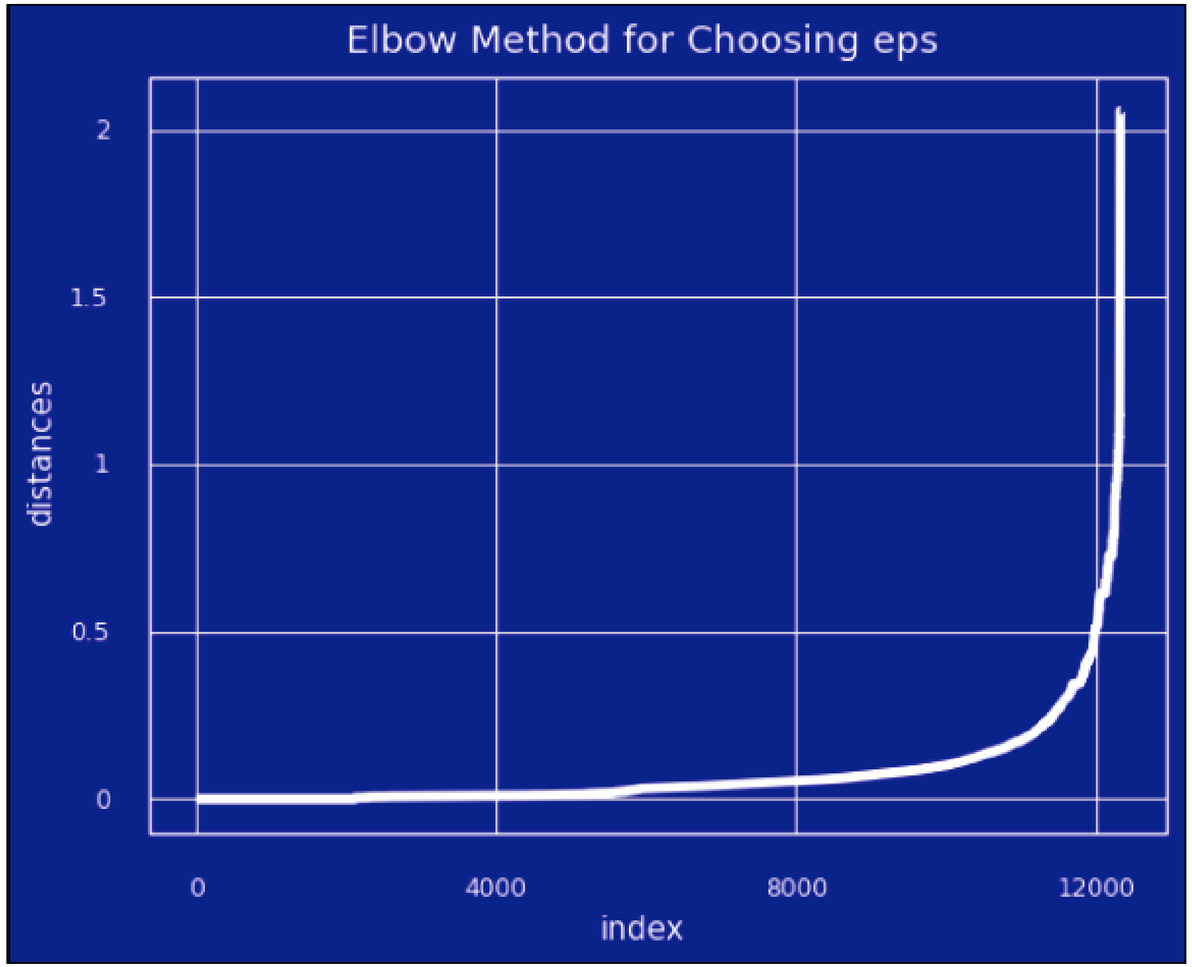
Methods:

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

Q2

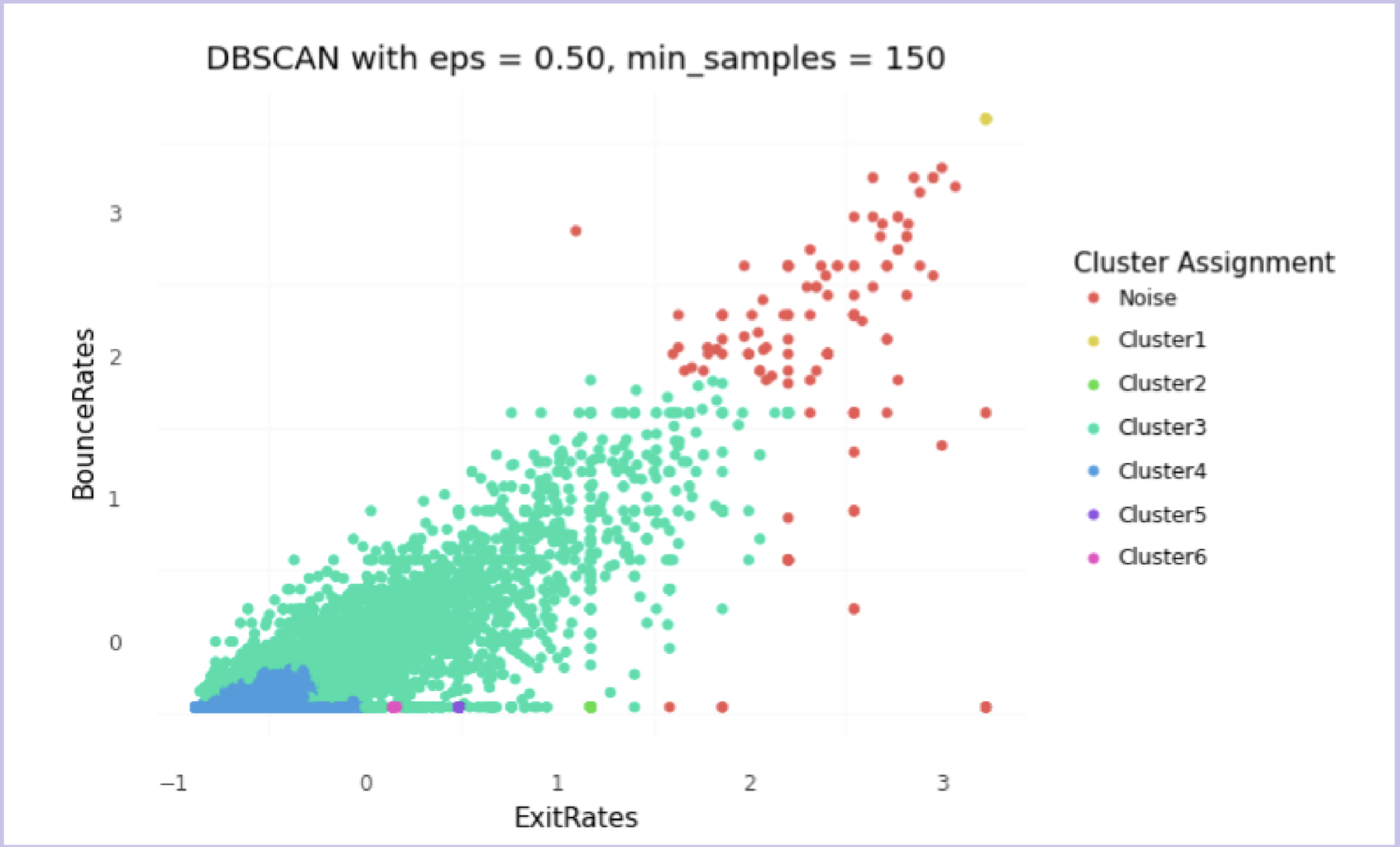
Q2) GGPlot


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



Q2) GGPlot & Findings

```
Cluster: 0.3603768094079344 3
Overall: 0.3542671334119401
```

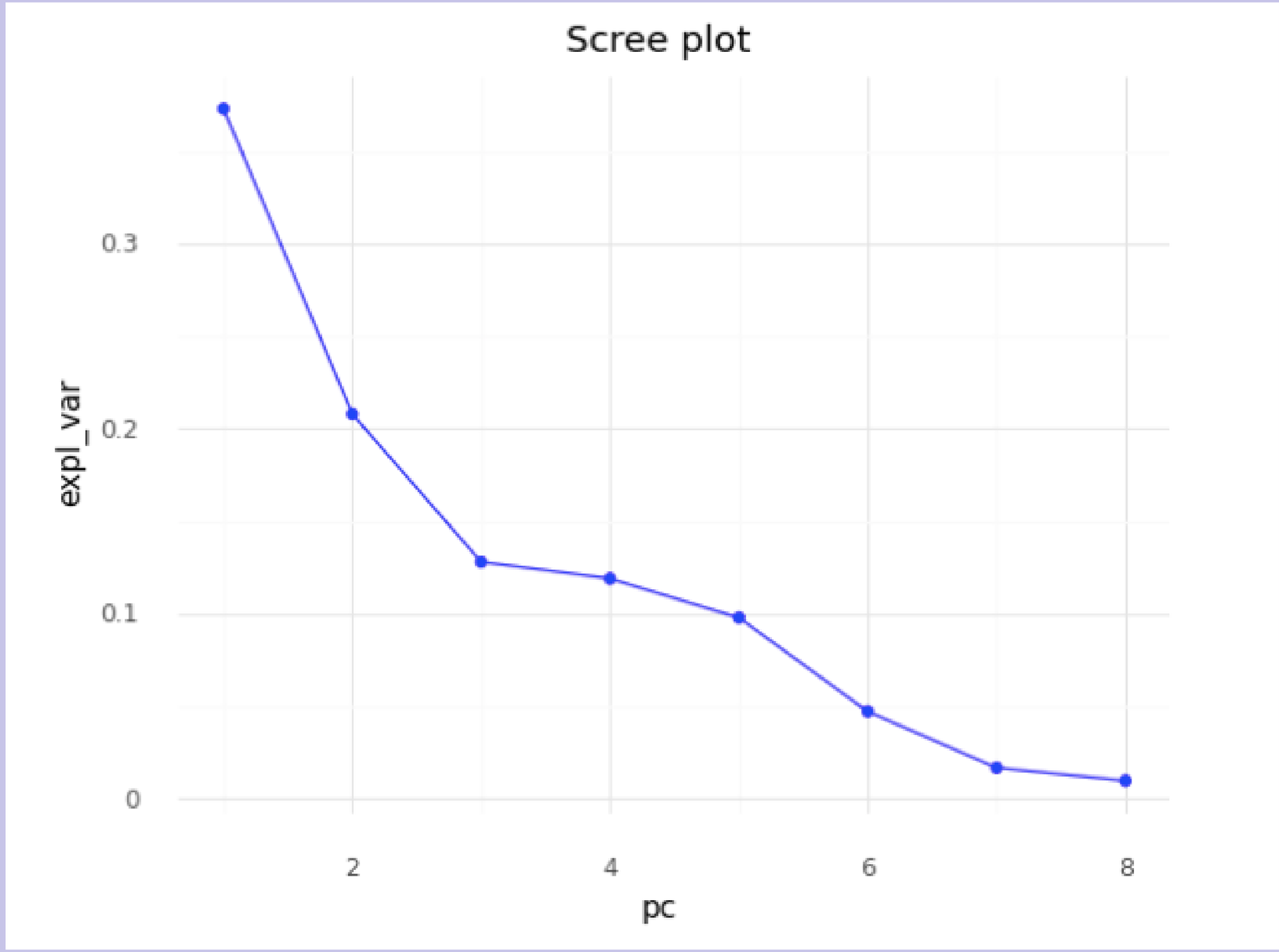
Are there any methods to reduce the dimensionality of the continuous variables in your data set? If so, which method, how can you tell, and how many variables do you need to retain 80% of the original variance?

Methods:

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

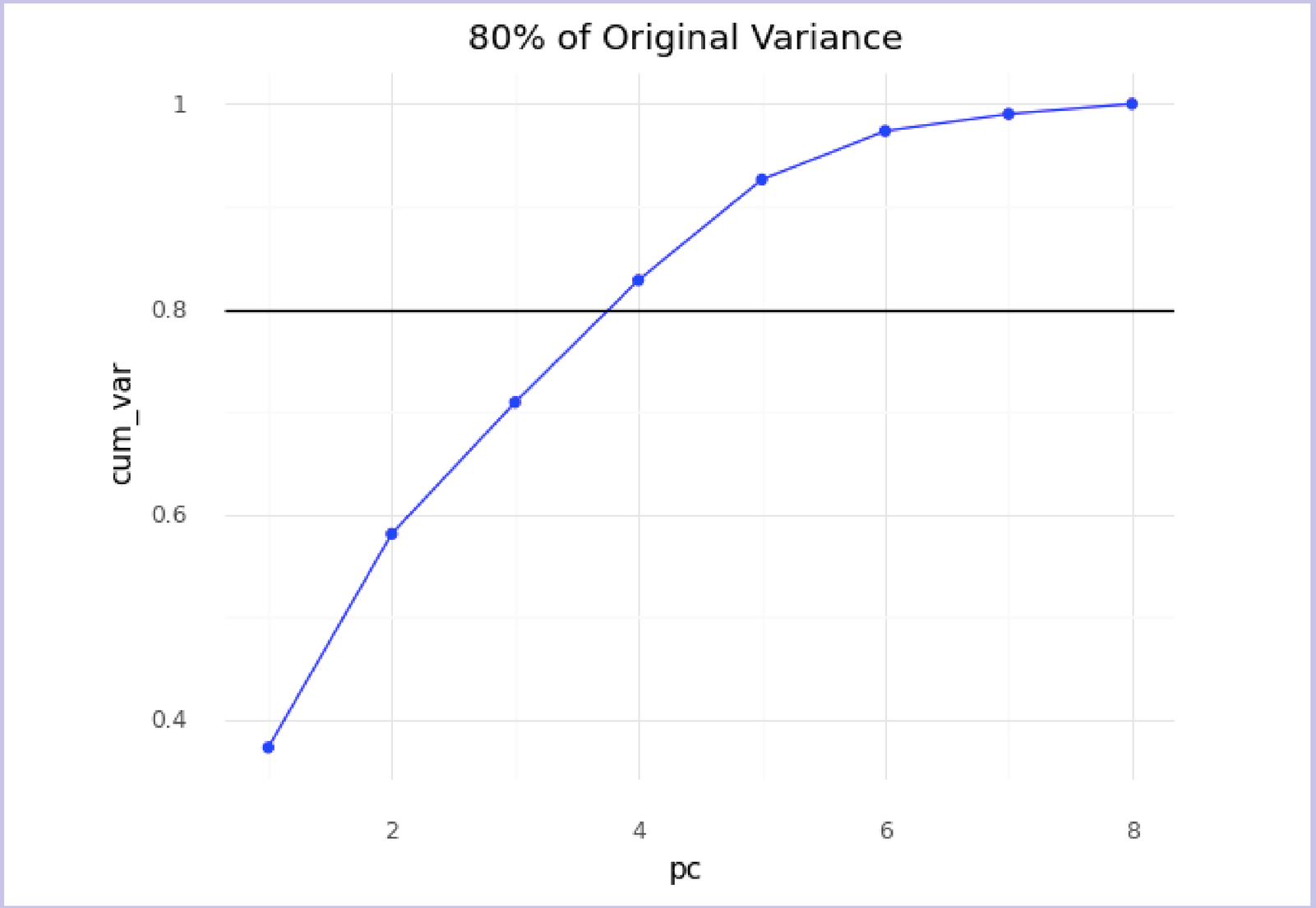
Q3

Q3) GGPlot



Q3) GGPlot & Findings

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



THE END
THANK YOU