# PA3

## Clustering

Jack, Felicia and Joost

# Talk About the Data Itself

## Numeric Features:

- Administrative
- Informational
- Product Related
- Administrative Duration
- Informational Duration
- Product Related Duration
- Bounce Rates
- Page Values
- Special Day

## Categorical Features:

- Traffic Type
- Visitor Type
- Operating Systems
- Browser
- Region
- Month
- Weekend
- Revenue

# Visualization Comparing Browser 13 and Others

# Visualization Comparing Browser 13 and Others

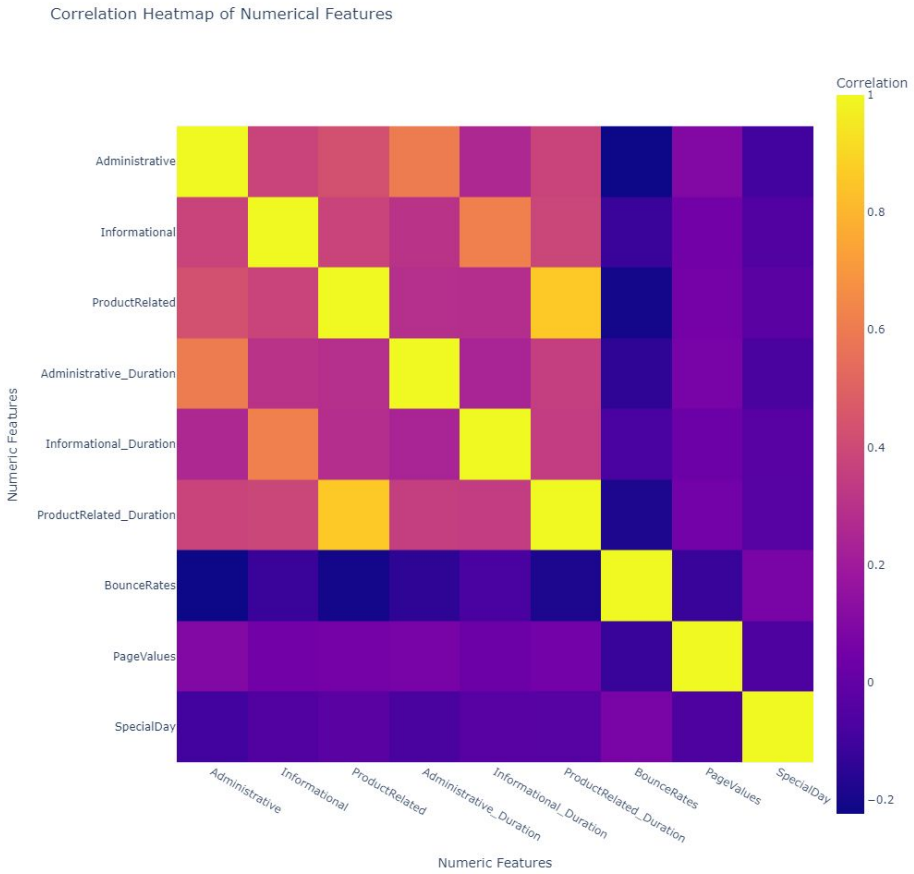
# Preprocessing Method (?

Normalized

# Suppose that if We're An E-commercial…

Different type of sites will have positive correlation with their duration time.

But types won’t affect any to bounce rate or page values

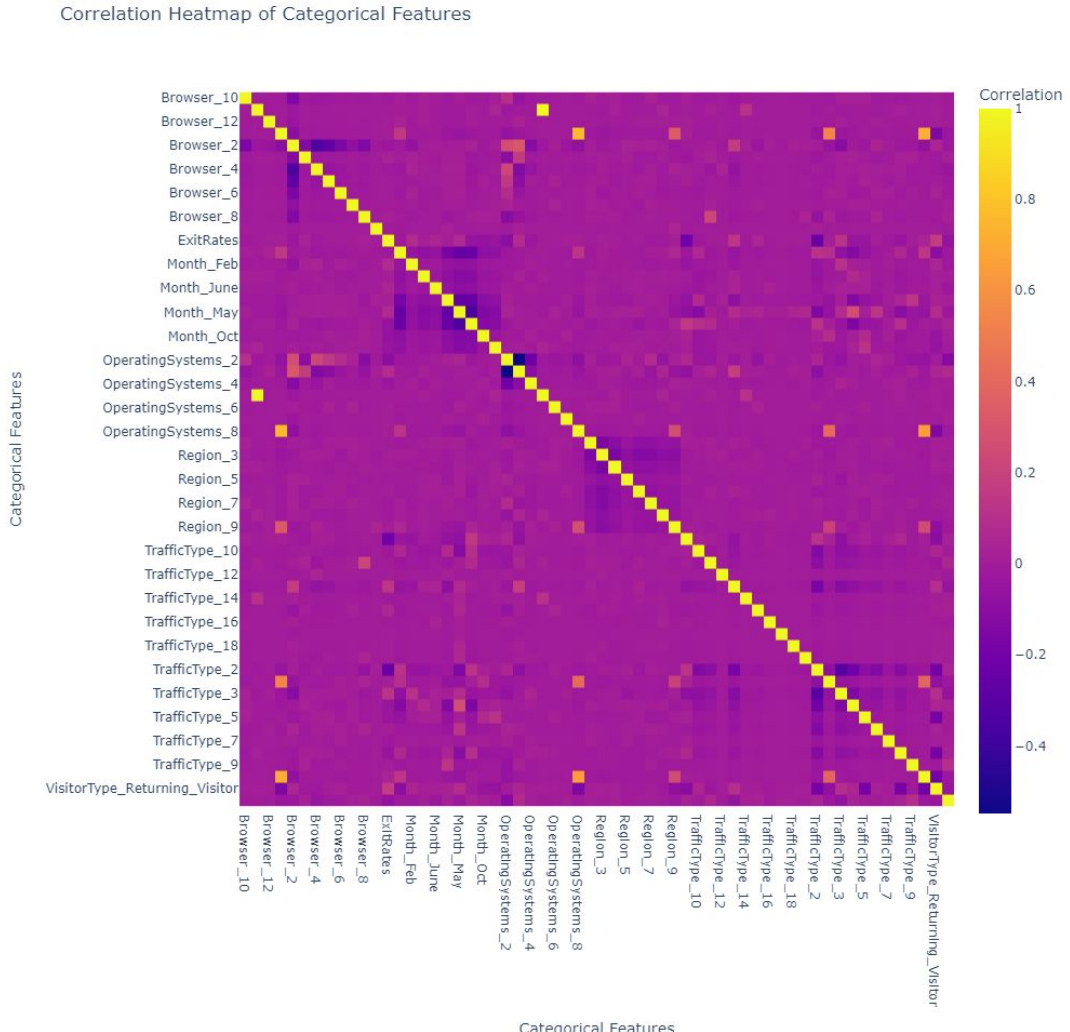


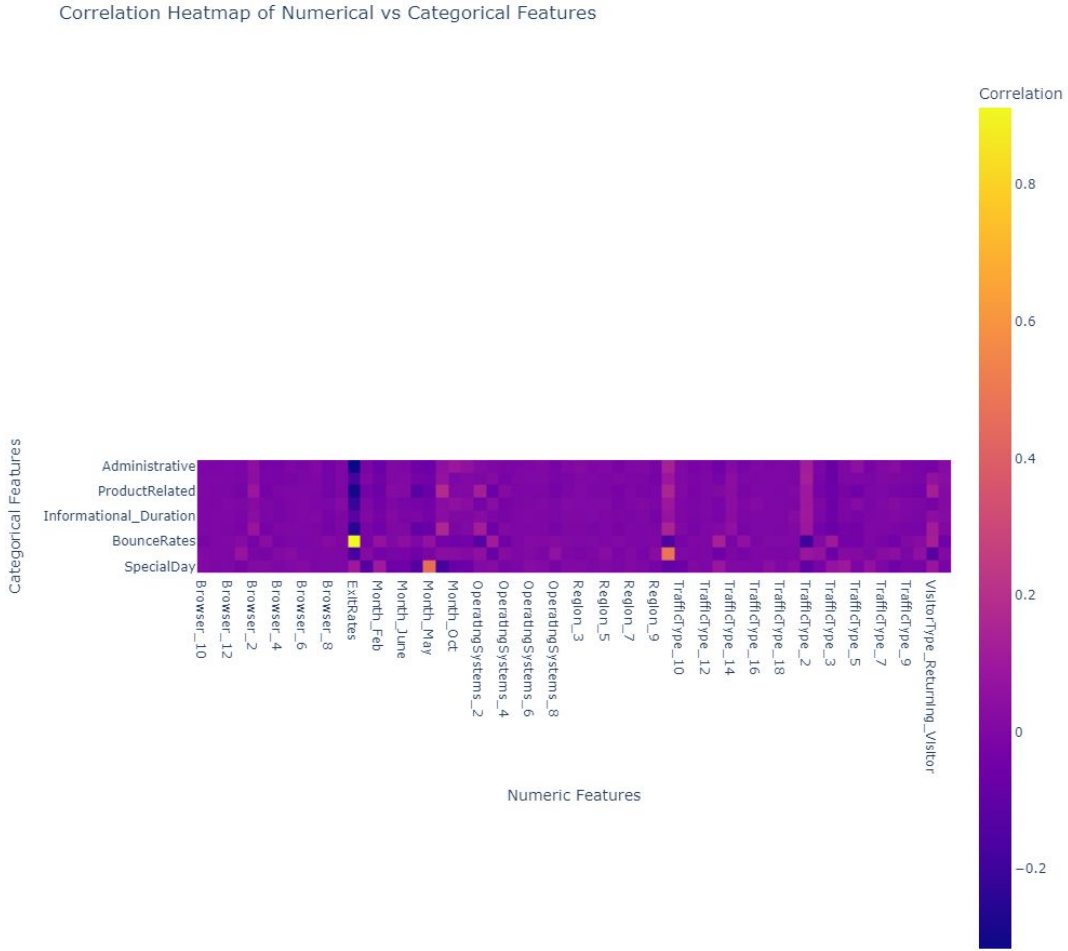
# About browser 13

- Operating system 8
- Region 9
- Traffic type 20
- Visitor type other

# Also

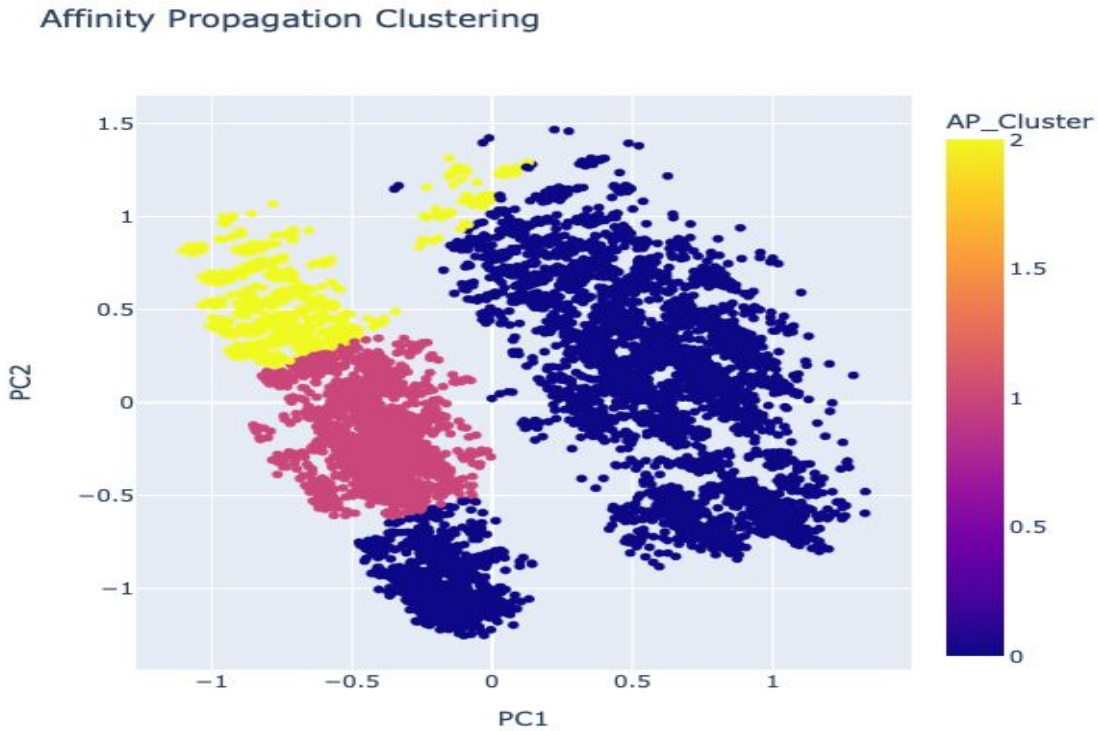
Operating system 8 has high positive correlation with traffic type 20, visitor type other




Correlation Heatmap of Numerical vs Categorical Features
Correlation
0.8
0.6
0.4
0.2
0
−0.2
Categorical Features
Administrative
ProductRelated
Informational_Duration
BounceRates
SpecialDay
Numeric Features
Browser_10
Browser_12
Browser_2
Browser_4
Browser_6
Browser_8
ExitRates
Month_Feb
Month_June
Month_May
Month_Oct
OperatingSystems_2
OperatingSystems_4
OperatingSystems_6
OperatingSystems_8
Region_3
Region_5
Region_7
Region_9
TrafficType_10
TrafficType_12
TrafficType_14
TrafficType_16
TrafficType_18
TrafficType_2
TrafficType_3
TrafficType_5
TrafficType_7
TrafficType_9
VisitorType_Returning_Visitor

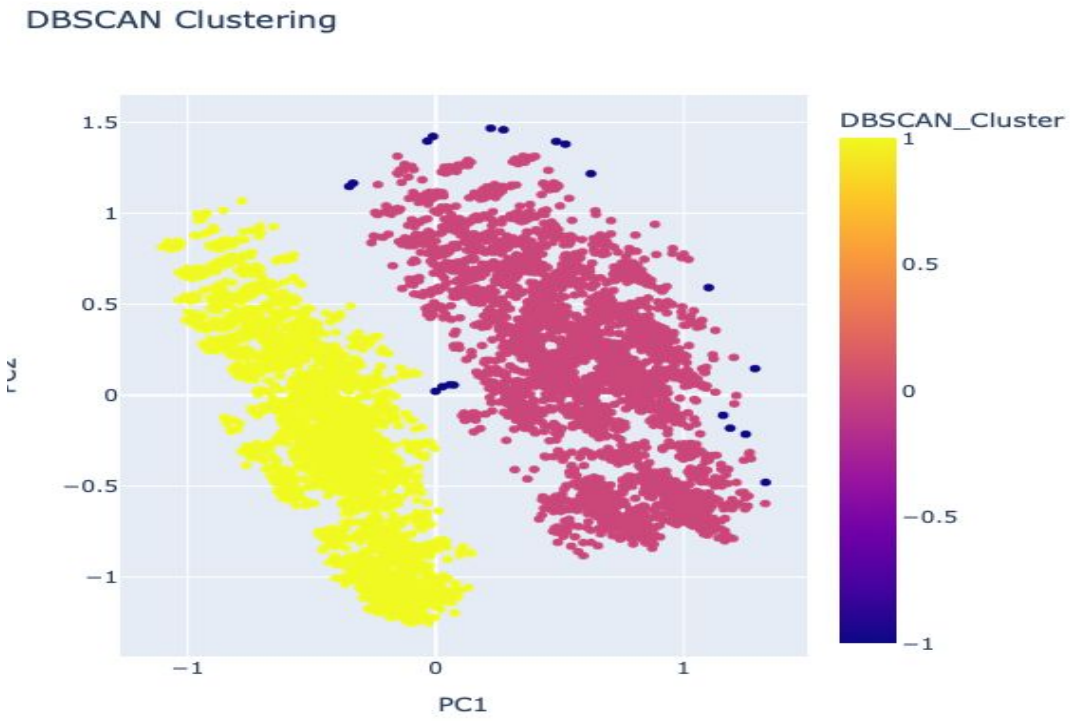
# Clustering Algorithms 1 - Affinity Propagation

| Model | Silhouette score | Davies-Bouldier Score | Calinsky-Harabasz Score |
|---|---|---|---|
| Affinity Propagation | 0.307416 | 0.963267 | 4668.185766 |



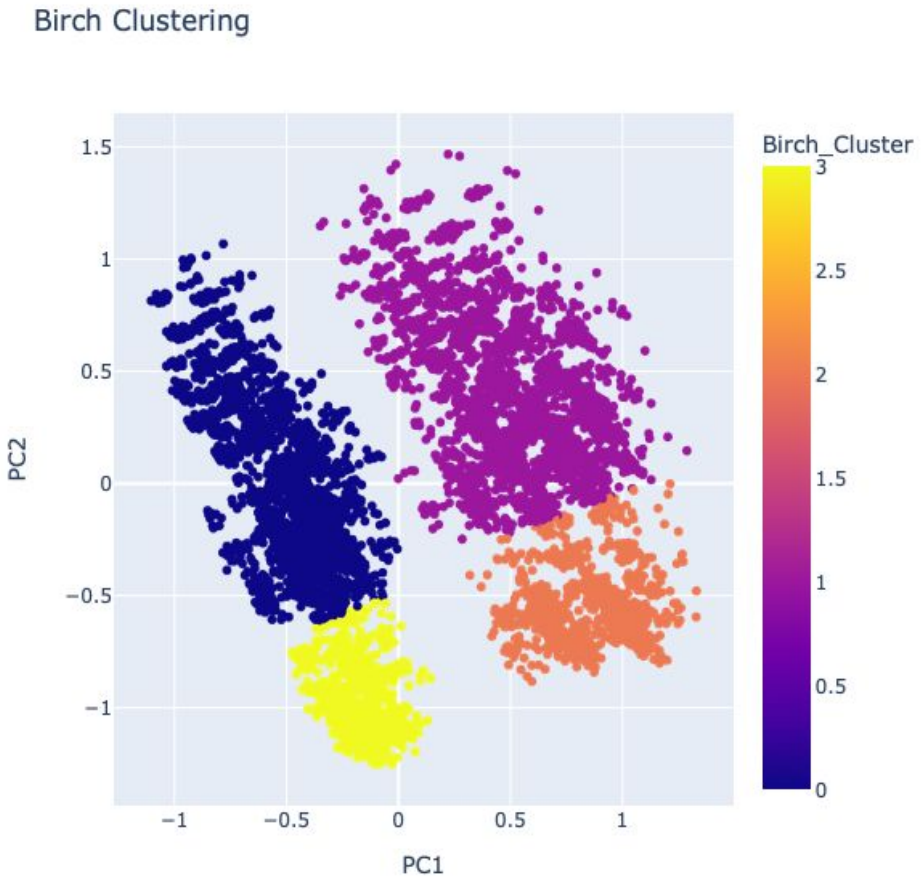
# Clustering Algorithms 2 - DBSCAN Clustering

| Model | Silhouette Score | Davies-Boudlin Score | Calinsky-Harabasz Score |
|---|---|---|---|
| DBSCAN | 0.401499 | 2.608033 | 5524.599755 |



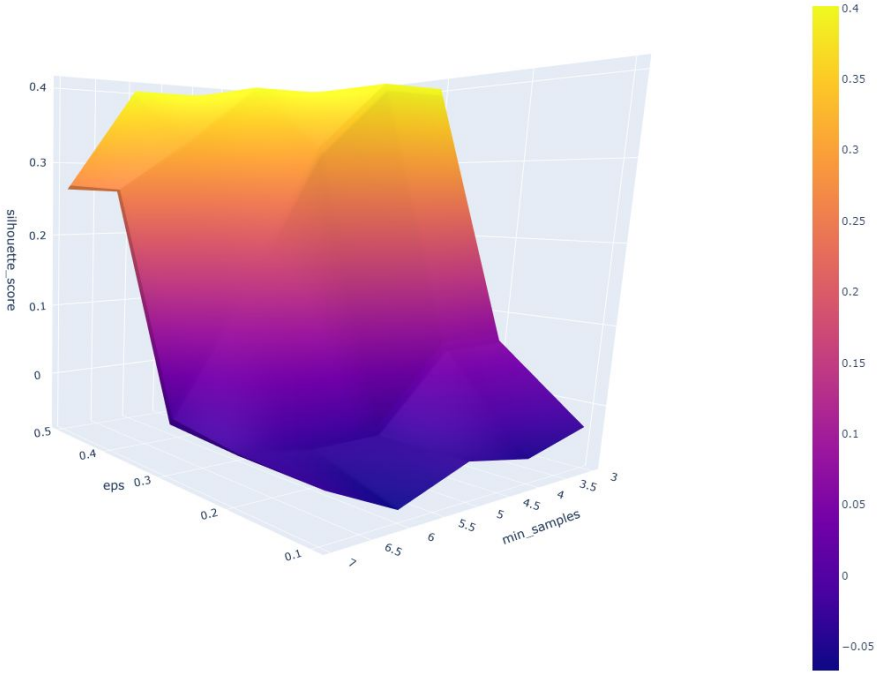
# Clustering Algorithms 3 - Birch Clustering

| Model | Score |
|---|---|
| Silhouette Score | 0.474736 |
| Davies-Bouldin Score | 0.641019 |
| Calinski-Harabasz Score | 12667.449132 |



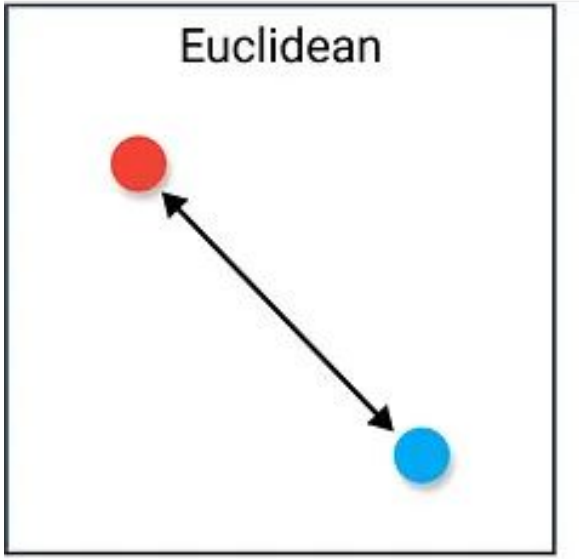
# DBSCAN - Parameters Explored

- Investigated the impact of eps and min_samples parameters
- eps measures the distance for outlier determination
- min_samples specifies the minimum neighboring points for an outlier/main point
- Utilized the silhouette score for performance assessment
- Lowering eps significantly enhances performance
- These parameters are highly dependent on the data, but provide a rough overview of the spacing between main points and outliers



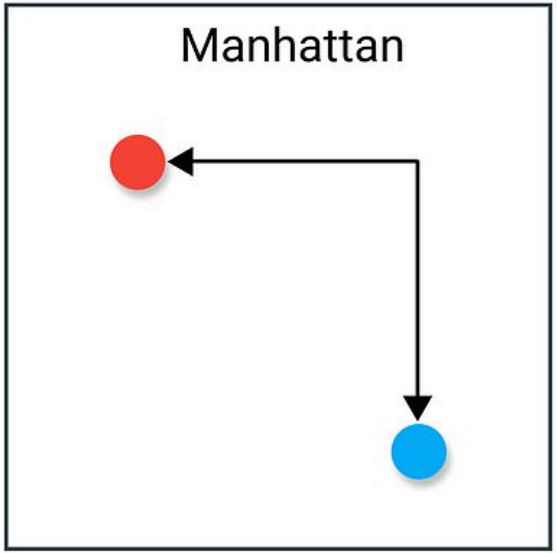
# DBSCAN - Euclidean distance function

- It measures the straight-line or "as-the-crow-flies" distance between two points.
- It is sensitive to outliers because it squares the differences.



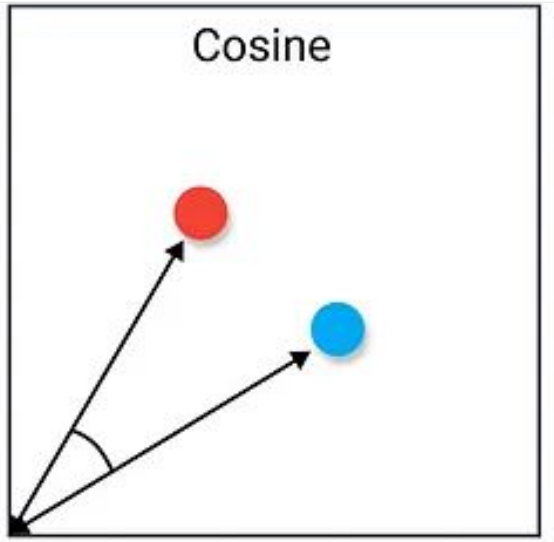
# DBSCAN - Manhattan distance function

- It measures the distance as if you are moving through a grid-like path (like city blocks) between two points.
- It is less sensitive to outliers compared to the Euclidean distance because it uses the absolute differences.

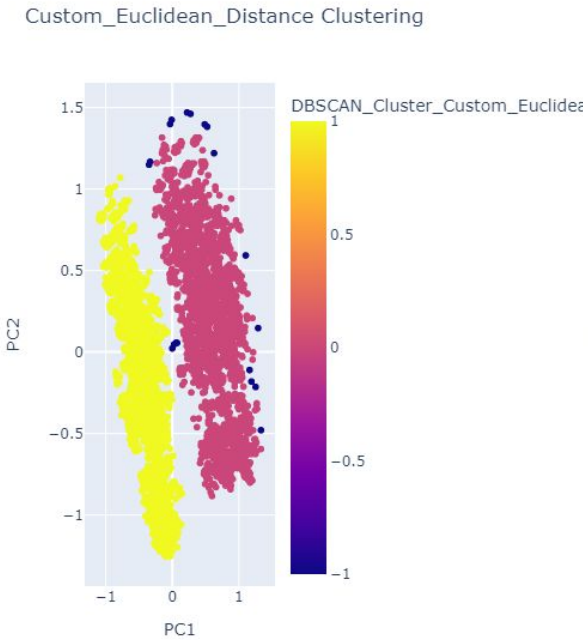


# DBSCAN - Cosine similarity distance function

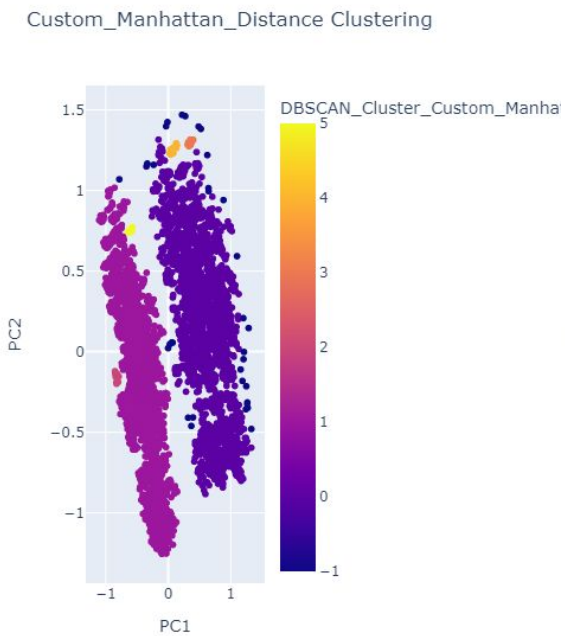
- It ranges between -1 and 1. A value of 1 indicates that the vectors are identical, 0 means they are orthogonal (unrelated), and -1 means they are diametrically opposed.
- It ignores the magnitude of the vectors and focuses on their orientation in the vector space.
- It is useful when you want to measure similarity in terms of the angle between vectors and not their absolute distance.



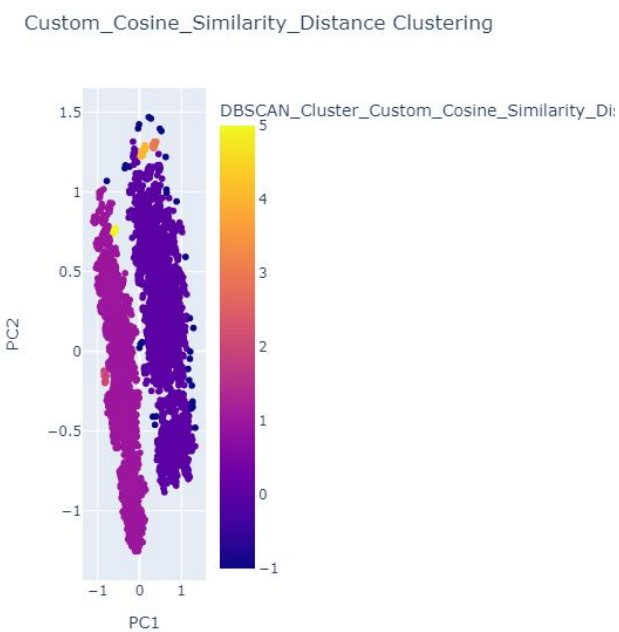
# Euclidean




# Manhattan




# Cosine



# We Preferred…

- Euclidean performed better using all scoring methods
- Notably, all distance functions received almost the exact same DB-Score