Customer Segmentation

Using K-Means and DBSCAN Clustering: A Comparative Study

- Abhishek Puppala
- Neelapu Tirumalesh Reddy
- Deval Shaileshkumar Mali
- Jashwanth Kalyan Polavarapu

Clustering

- Clustering is the classification of objects into different groups, or more precisely, partitioning of a data set into subsets => clusters.
- Data points in each cluster share common trait/behavior according to some defined distance measure.
- Distance measure defines how the similarity of two elements is calculated and influences the shape of the cluster.
  - Euclidean Distance:
    - E = $\sum_{i=1}^{k} \sum_{x \in C_i} |x - x_i|^2$
- Clustering Algorithms used:
  - K – Means
  - DBSCAN

K – Means Clustering

- Clustering Algorithm to cluster ‘n’ objects based on attributes into k partitions, where k < n.
- An algorithms for partitioning (or clustering) N data points into K disjoint subsets S_i containing data points so as to minimize the sum-of-squares criterion.

$$J = \sum_{i=1}^{k} \sum_{x \in S_i} |x - x_i|^2$$

- Where x is a vector representing the data point and x_i is the geometric centroid of the data points in S_j.
- Clustering is done by minimizing the sum of squares of distances between data and the corresponding cluster centroid.

DBSCAN Clustering

- DBSCAN for Density-Based Spatial Clustering of Applications with Noise
- Density based clustering locates regions of high density that are separated from one another by regions of low density.
  - Density = number of points within a specified radius (eps)

- Algorithm:
  - Input => N objects to be clustered and global parameters Eps, MinPts.
  - Output => Cluster of objects.

Customer Segmentation

- Customer segmentation helps the organizations/companies understand their customer base and can target or design new products specific to a sub-group of the customers.
- Business can determine which niche market best fits the unique product they produce, which enables the business to maintain a sizable market share and maintain its competitive edge over other market participants.
- Segmentation can be achieved focusing on different attributes such as behavioral, geographic, demographic etc.,
- Segmentation focuses on behavioral data as it is the most efficient and practical one.

RFM Analysis

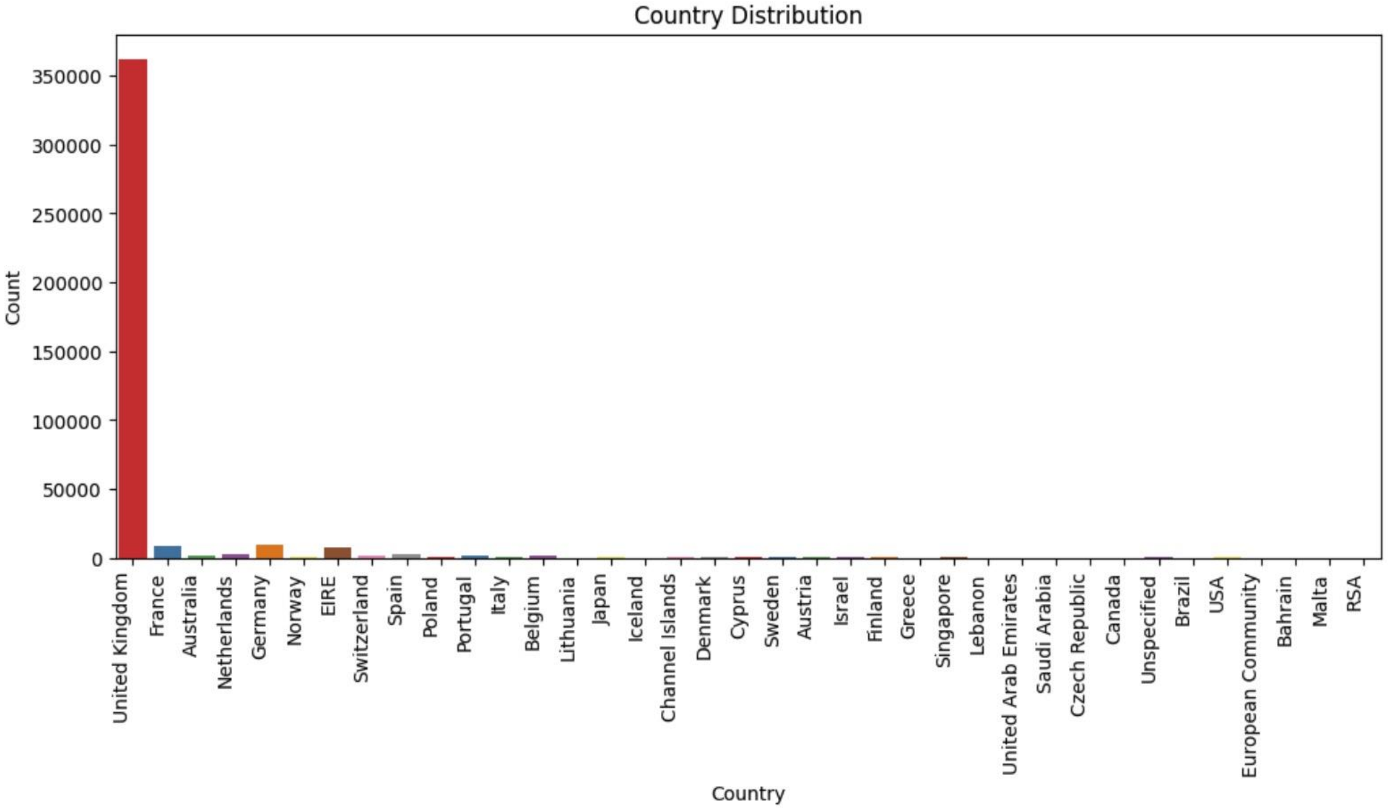
- RFM known for the values based on Recency, Frequency and Monetary.
- Recency:
  - Refers to how often times a customer awaits before making subsequent purchase.
- Frequency:
  - Measurement of how frequently a customer made a purchase over a given time frame.
  - Higher values suggests that clients are more devoted to the business, while the lower values suggests the opposite.
- Monetary:
  - Money spent by a consumer during the specified time period.

Customer Segmentation using Online Retail Dataset

- Dataset Link: https://archive.ics.uci.edu/ml/datasets/online+retail
- Attributes:
    - InvoiceNo – Six digit unique number for each transaction.
    - StockCode – Discrete value allocated to each individual product.
    - Description – Name or Description of the item.
    - Quantity – Volume of each purchase made in a single trade.
    - InvoiceDate – Timestamp when the invoice was generated.
    - UnitPrice – Production cost of the goods.
    - CustomerID – Each user is identified by a distinct autoincremented number.
    - Country = Country Name, title of the nation in which the client lives.

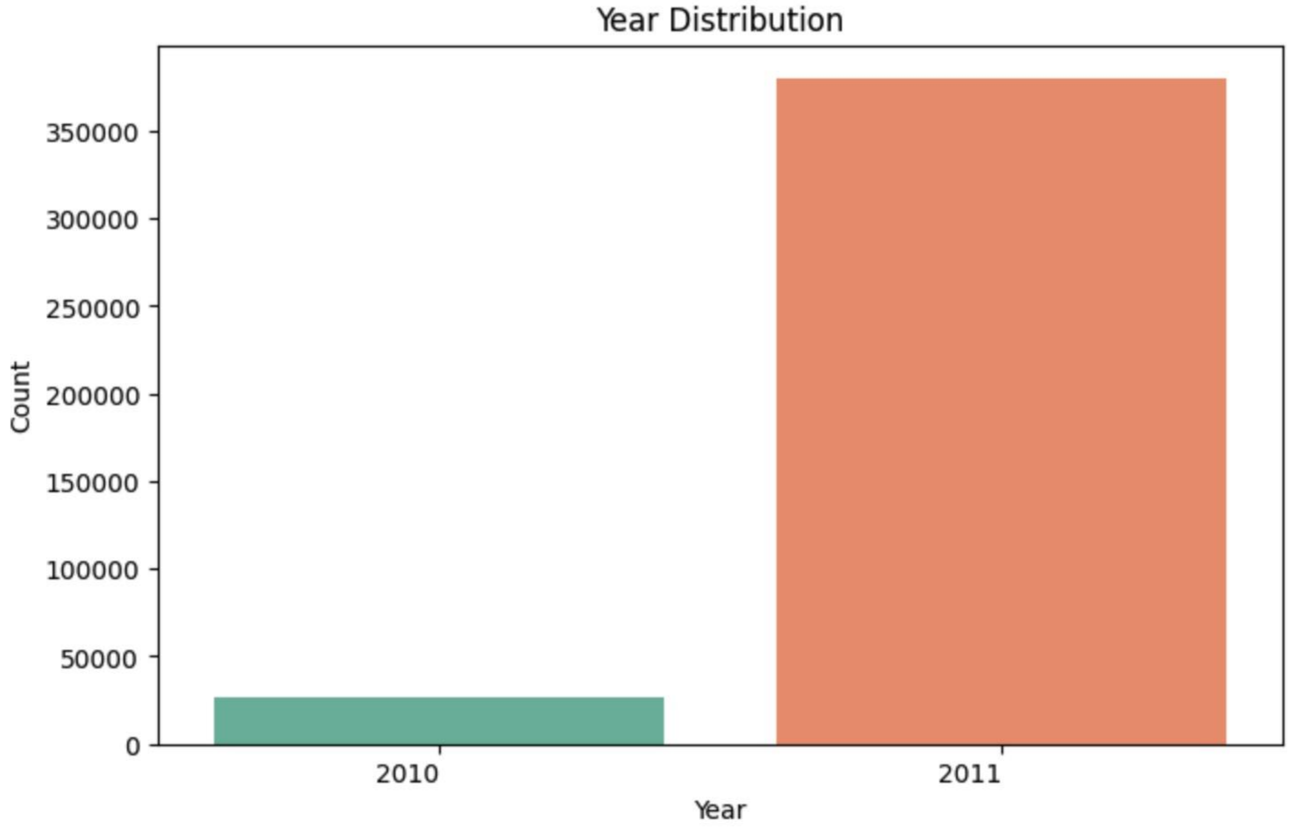
Country Distribution – Outliers

- United Kingdom comprises of 98% data and all the other countries data is considered as outliers and dropped from the dataset



Year Distribution – Outliers contd.

- Year 2010 data is dropped from the dataset as it is not having enough volume.



RFM Values Calculation

Recency

- InvoiceDate attribute provides the date information when the customer purchased.
- Max InvoiceDate of the entire dataset is calculated as most recent transaction date.
- Max transaction date and customer invoice date difference is calculated.
- Applying min function for each customer on the calculated date column gives the Recency of the customer.

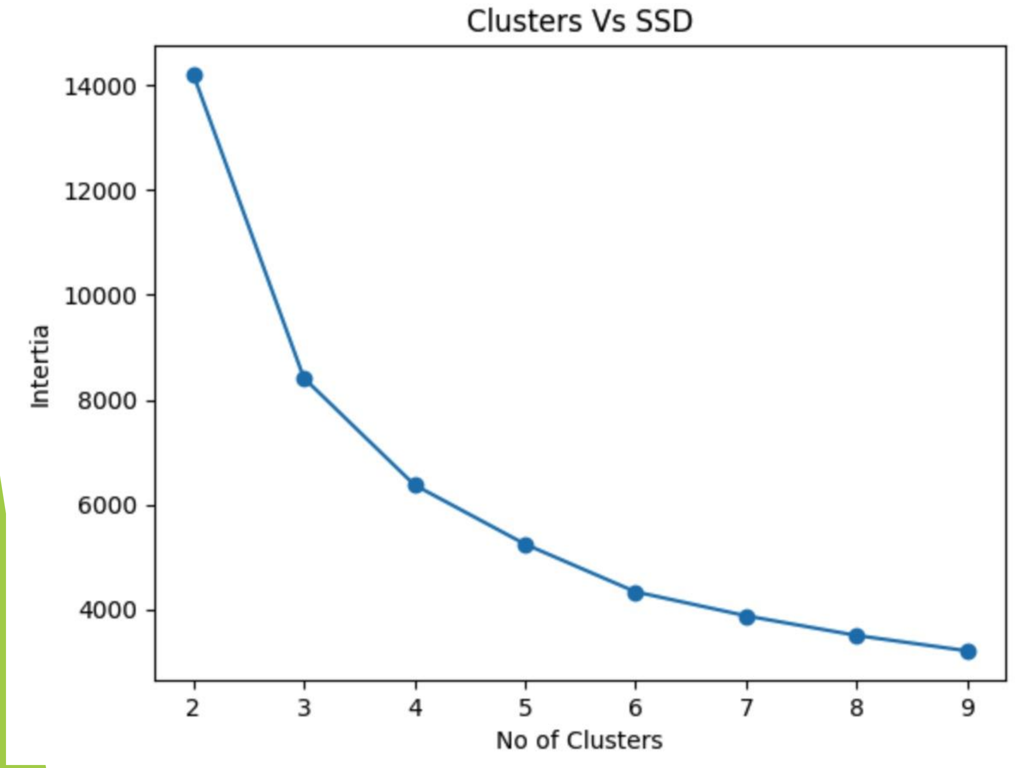
Frequency

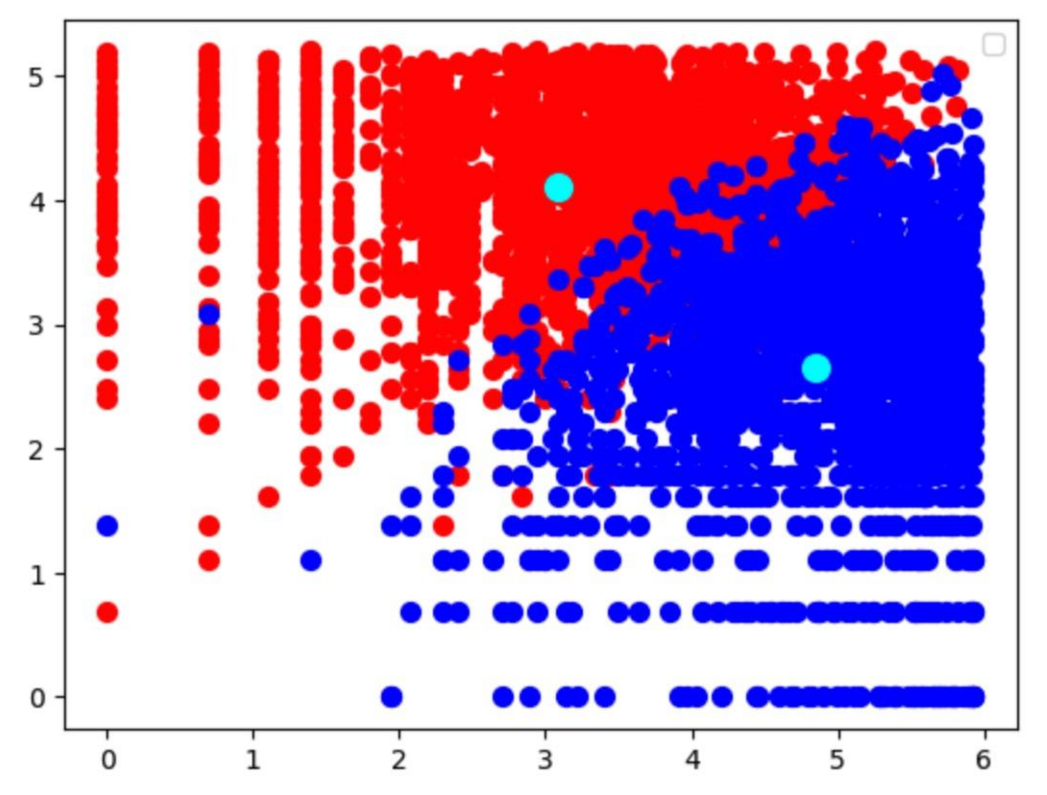
- For Each customer, no (count) of transactions made in the dataset is calculated.
- Calculated value is named as Frequency.

Monetary

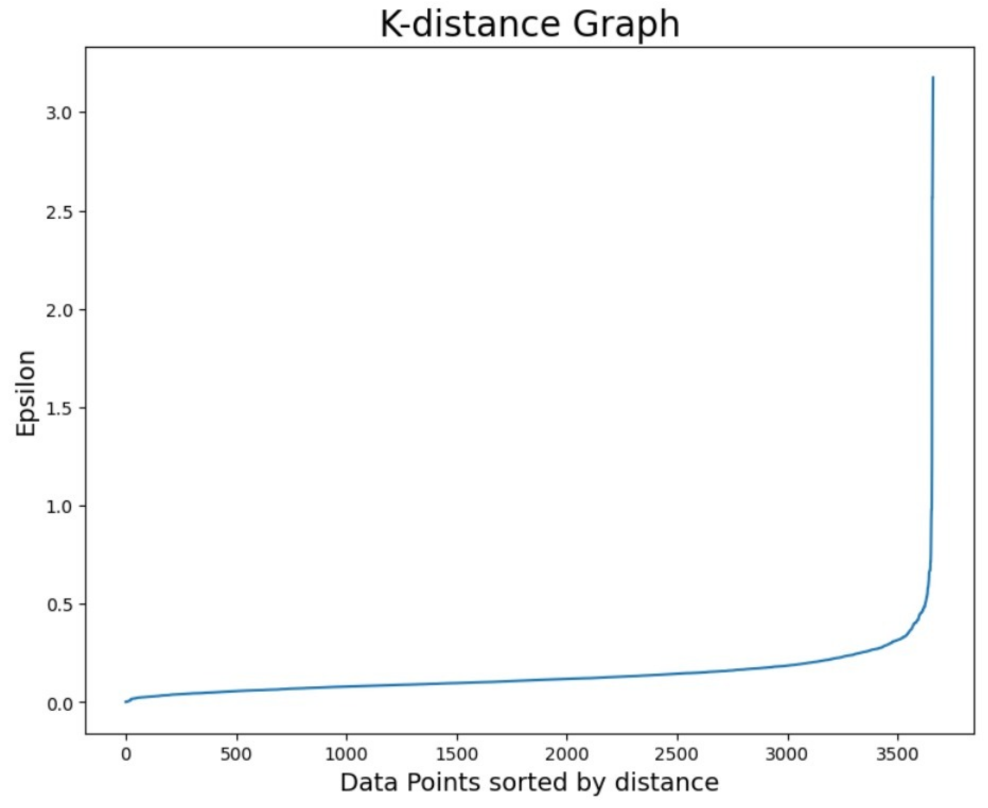
- UnitPrice and Quantity attributed are multiplied to calculate the amount spent by the customer for each transaction.
- For each customer total amount spent is calculated by using sum function.
- Aggregated amount spent is termed as Monetary.

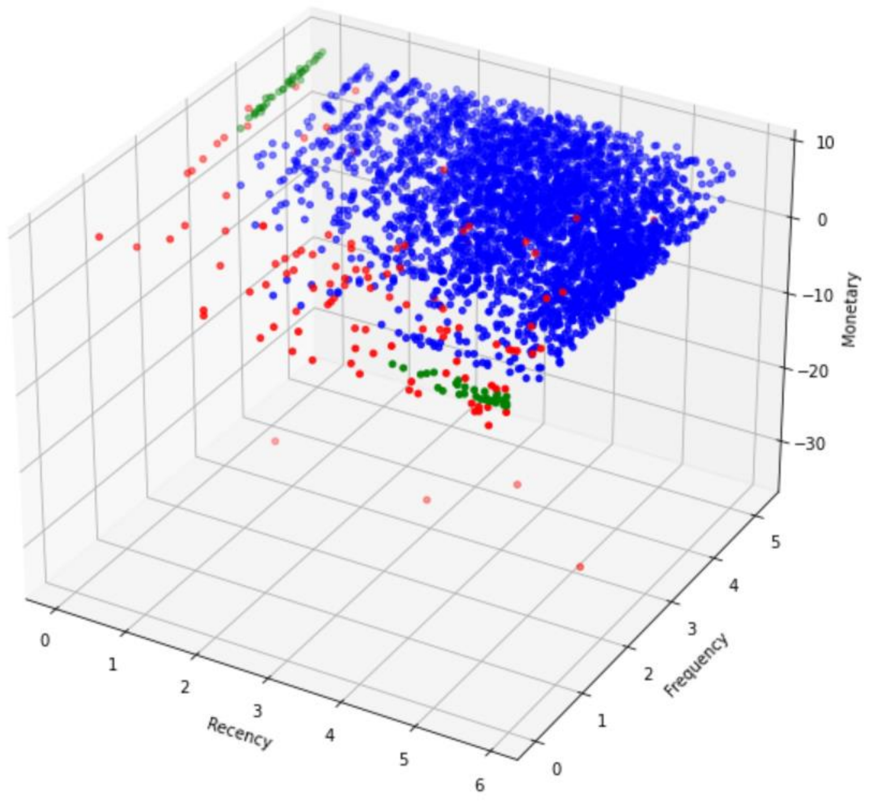
K – Means Clustering



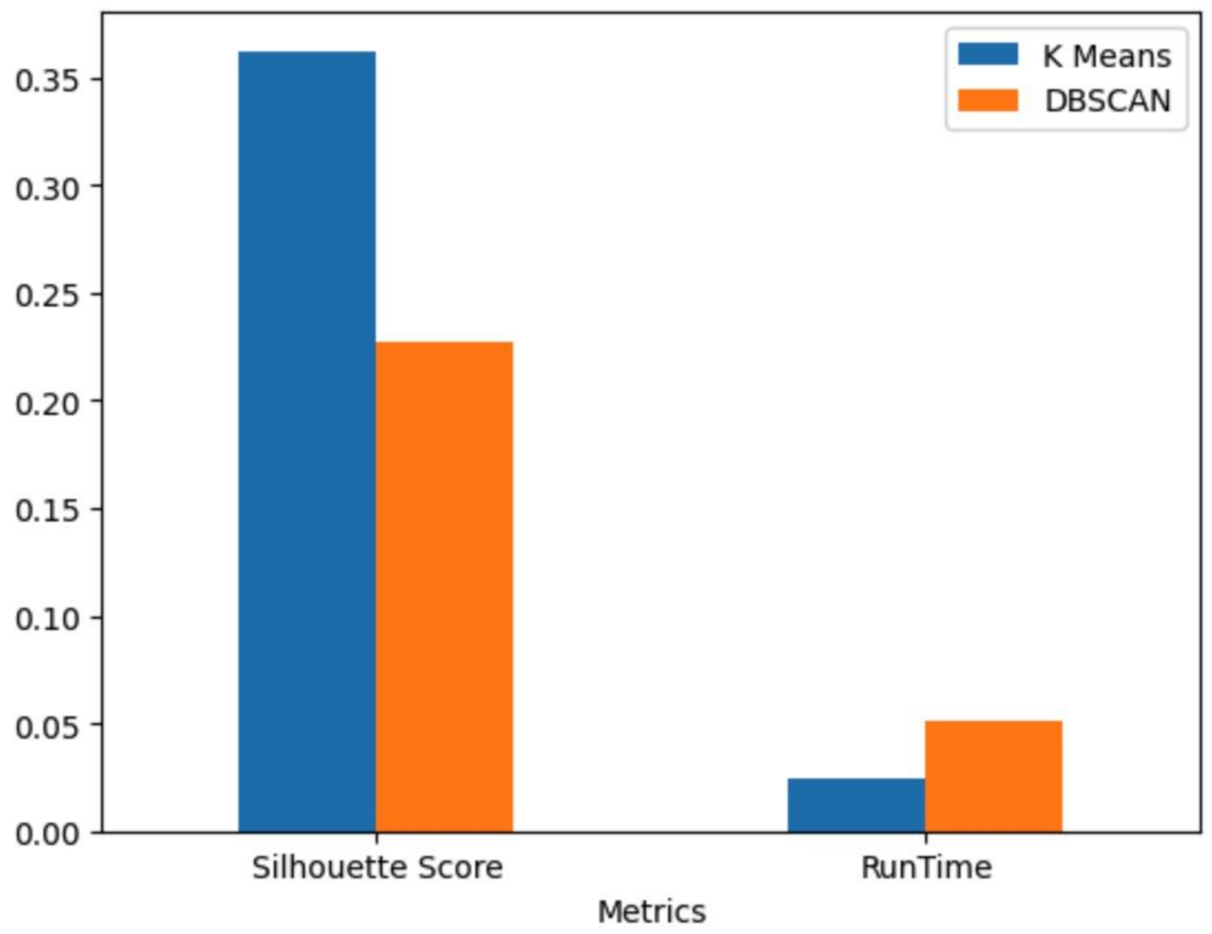


DBSCAN Clustering





Results


0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
K Means
DBSCAN
Silhouette Score
RunTime
Metrics

Conclusion

- For the selected Dataset with RFM values, K-Means algorithm performed better generating the clusters than DBSCAN.
  - Comparison done based on the silhouette scores of the formed clusters
- Time taken to model the data and form the clusters.
  - K – Means is faster than DBSCAN most times.