Presentation Overview

Identifying Different Kinds of Businesses

- To understand the variety of merchants using our company's payment processing services, we analyzed transaction data spanning two years. Our approach aimed to categorize merchants based on their transaction patterns, sizes, frequencies, and any discernible trends that could suggest their business type.

Predicting Merchant Churn

- We focused on defining what constitutes merchant churn and identifying patterns that precede a merchant's decision to stop using our services. A predictive model was then developed to forecast which active merchants might churn in the near future, enabling targeted retention strategies.

Implemented three clustering algorithms to segment merchants into distinct groups based on their transaction patterns.

1. DB-SCAN
2. Label Propagation
3. K-Means

For Churn Prediction

1. Logistic Regression was employed as the predictive model

The period considered for predicting churn was set at 90 days, corresponding to a quarterly timeline.

Data Overview

The provided dataset contains three distinct columns:

- 'merchant' for Merchant ID
- 'time' for the timestamp of the transaction
- 'amount_usd_in_cents' for the transaction value in US cents

```
[ ] df.head()
```

	Unnamed: 0	merchant	time	amount_usd_in_cents
0	1	faa029c6b0	2034-06-17 23:34:14	6349
1	2	ed7a7d91aa	2034-12-27 00:40:38	3854
2	3	5608f200cf	2034-04-30 01:29:42	789
3	4	15b1a0d61e	2034-09-16 01:06:23	4452
4	5	4770051790	2034-07-22 16:21:42	20203

Creating Merchant Profiles with Derived Features

Purpose of Feature Creation:

- To gain a nuanced understanding of merchant behavior, we derived several features from the transaction data. These features will inform our analysis of business types and churn risk.

Key Features Generated:

- **First and Last Payment Dates**: Determine the active period of a merchant within the dataset.
- **Total Amount Paid**: Measure the total revenue generated by each merchant for our company.
- **Number of Payments**: Quantify the transaction volume for each merchant.
- **Lifespan**: Calculate the duration between the first and last payments to assess merchant longevity.
- **Acquisition and Recency Metrics**: Analyze how recently a merchant was acquired and their last activity date, which are critical for churn analysis.
- **Acquisition Quarter and Year**: Understand seasonal trends and yearly growth in merchant acquisitions.
- **Number of Payments Per Day**: Evaluate the regularity and frequency of transactions.
- **Average Order Value**: Gauge the typical transaction size, which may indicate the type of goods or services sold.

New Features

```
features_df
```

merchant	first_payment_date	last_payment_date	total_amount_paid	number_of_payments	lifespan	acquisition	recency	acquisition_quarter	acquisition_year	nu
0002b63b92	2033-05-16 20:07:57	2033-05-16 20:07:57	3379	1	0 days 00:00:00	593 days 03:52:03	593 days 03:52:03	2	2033	
0002d07bba	2034-10-11 17:02:26	2034-12-15 09:56:19	89278	4	64 days 16:53:53	80 days 06:57:34	15 days 14:03:41	4	2034	
00057d4302	2033-05-30 01:30:52	2033-08-04 04:26:40	29521	28	66 days 02:55:48	579 days 22:29:08	513 days 19:33:20	2	2033	
000bcff341	2033-08-09 20:18:36	2033-08-09 20:18:36	7826	1	0 days 00:00:00	508 days 03:41:24	508 days 03:41:24	3	2033	
000ddbf0ca	2033-06-02 13:25:12	2033-06-02 13:25:12	10299	1	0 days 00:00:00	576 days 10:34:48	576 days 10:34:48	2	2033	
...	...	...	...	...	...	...	...	...	...	
ffd3e45675	2033-01-04 04:35:29	2033-01-27 00:32:30	72626	5	22 days 19:57:01	725 days 19:24:31	702 days 23:27:30	1	2033	

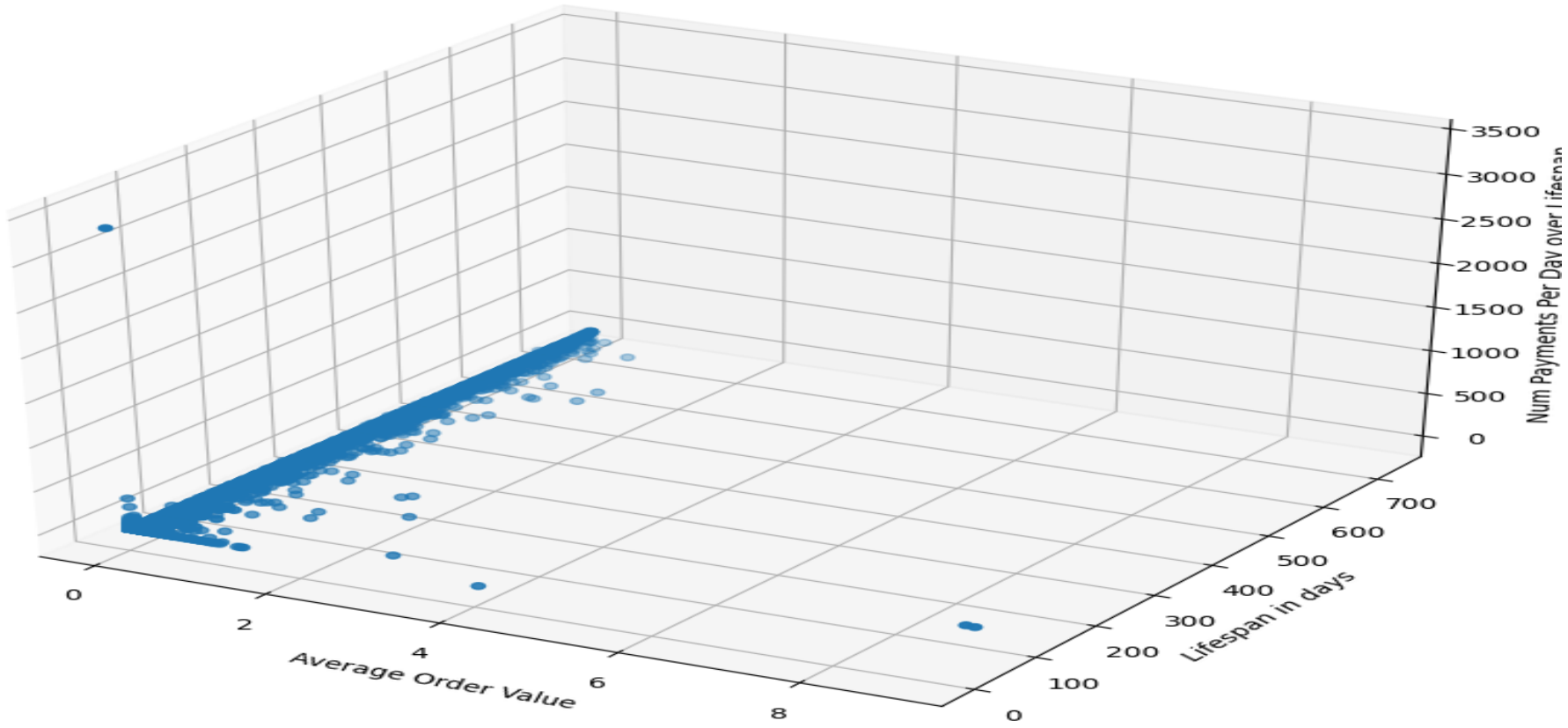
3D Scatter Plot of Merchant Activity Features

Visualizing Relationships Between Features:

- Created a 3D scatter plot to visualize the potential relationships between three key metrics:
  - **Average Order Value** on the x-axis.
  - **Lifespan in Days** on the y-axis.
  - **Number of Payments Per Day over Lifespan** on the z-axis.

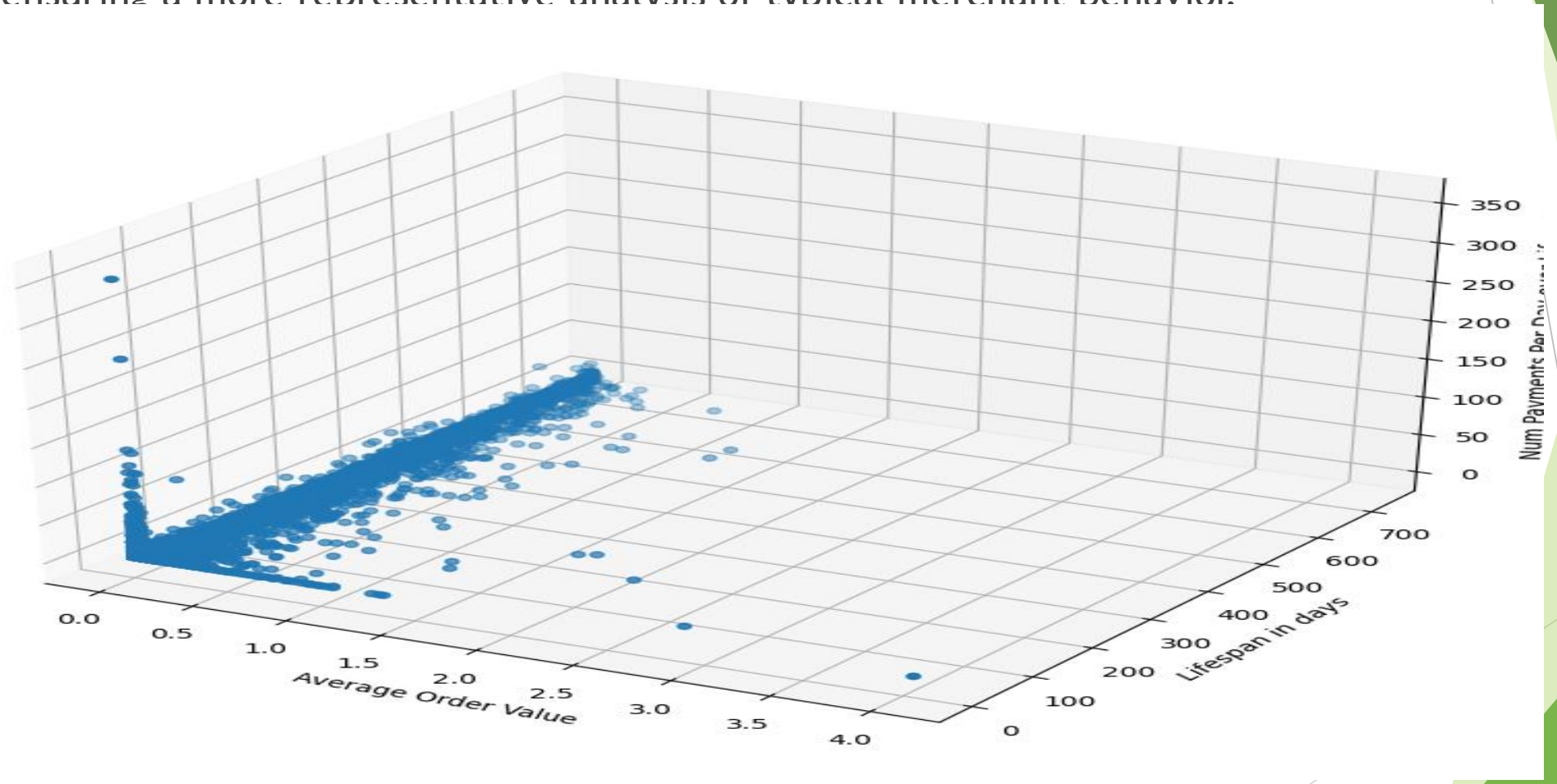
Interpreting the Data:

- This visualization helps us explore how frequently merchants transact (payment frequency), the value of their transactions (average order value), and their tenure with our company (lifespan).
- A dense cluster or distinct patterns may suggest commonalities among merchant types or indicate segments with a higher risk of churn.



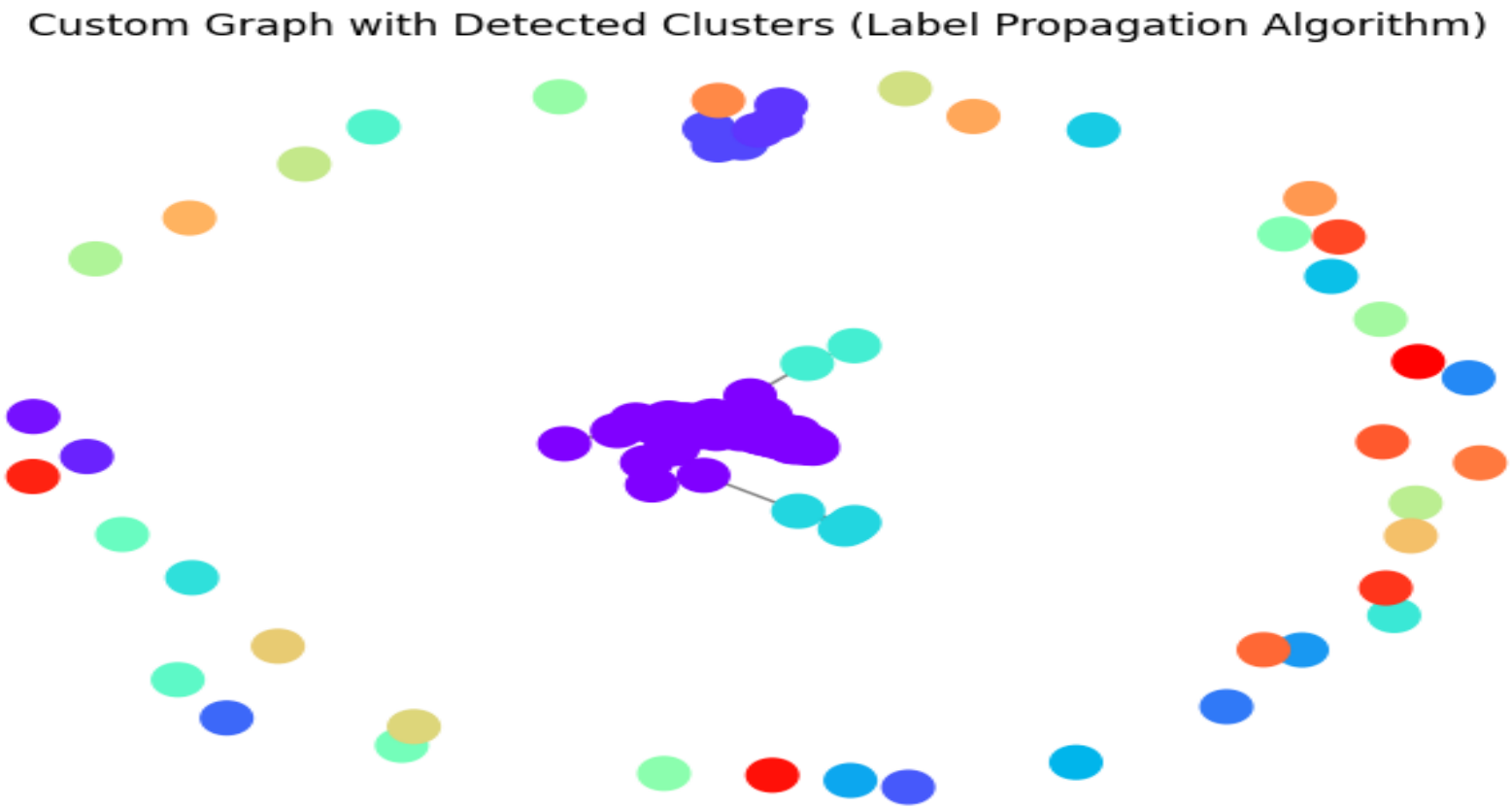
Based on the 3D visualization, we can clearly identify the presence of outlier data point.

Established thresholds for 'number_of_payments_per_day' (< 3000) and 'average_order_value' (<= 5,887,465 cents) to exclude extreme values from the dataset, ensuring a more representative analysis of typical merchant behavior.



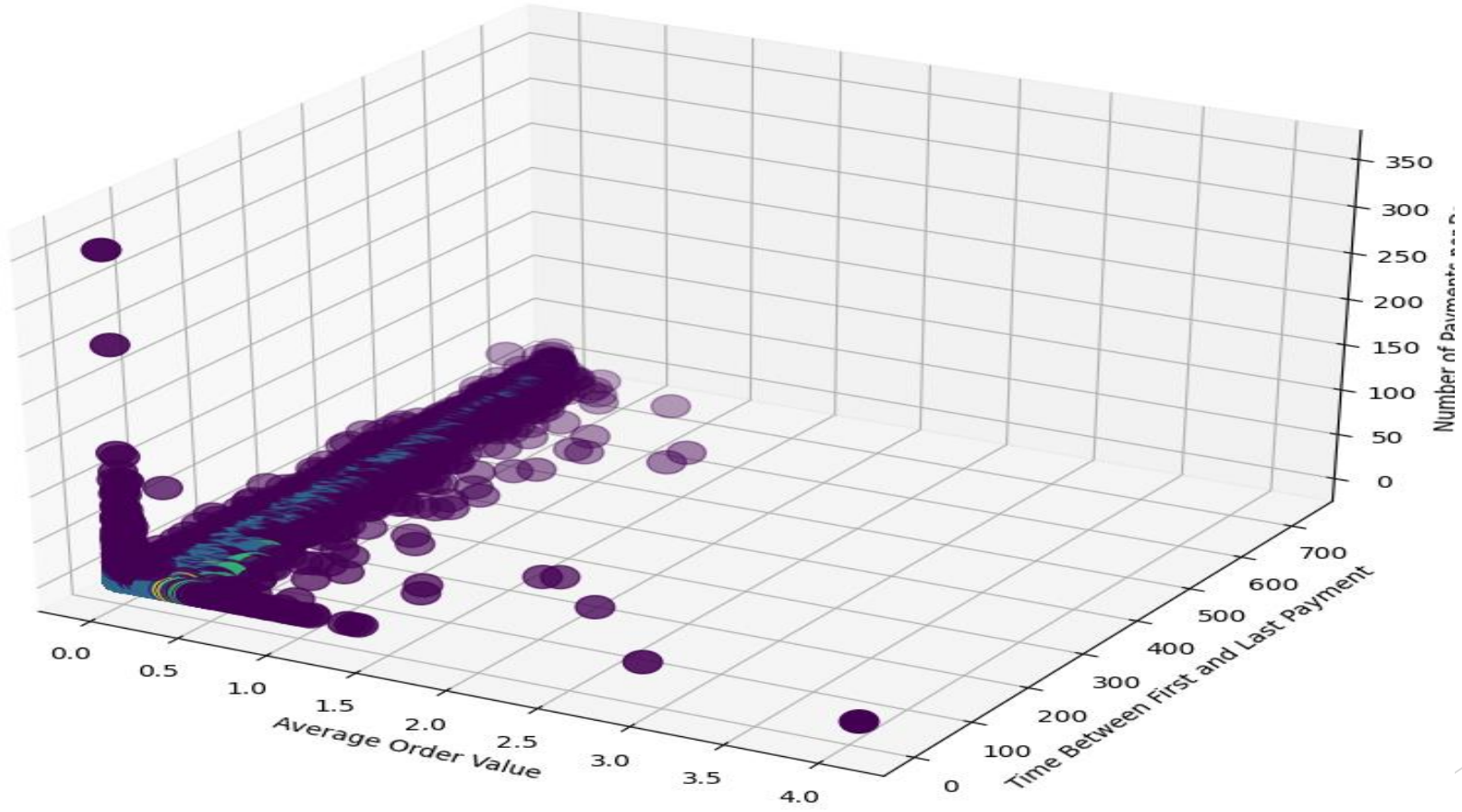
Clustering

1. Lable Propagation



After implementing the Label Propagation algorithm, we identified 45 distinct clusters, indicating a potentially unstable clustering outcome

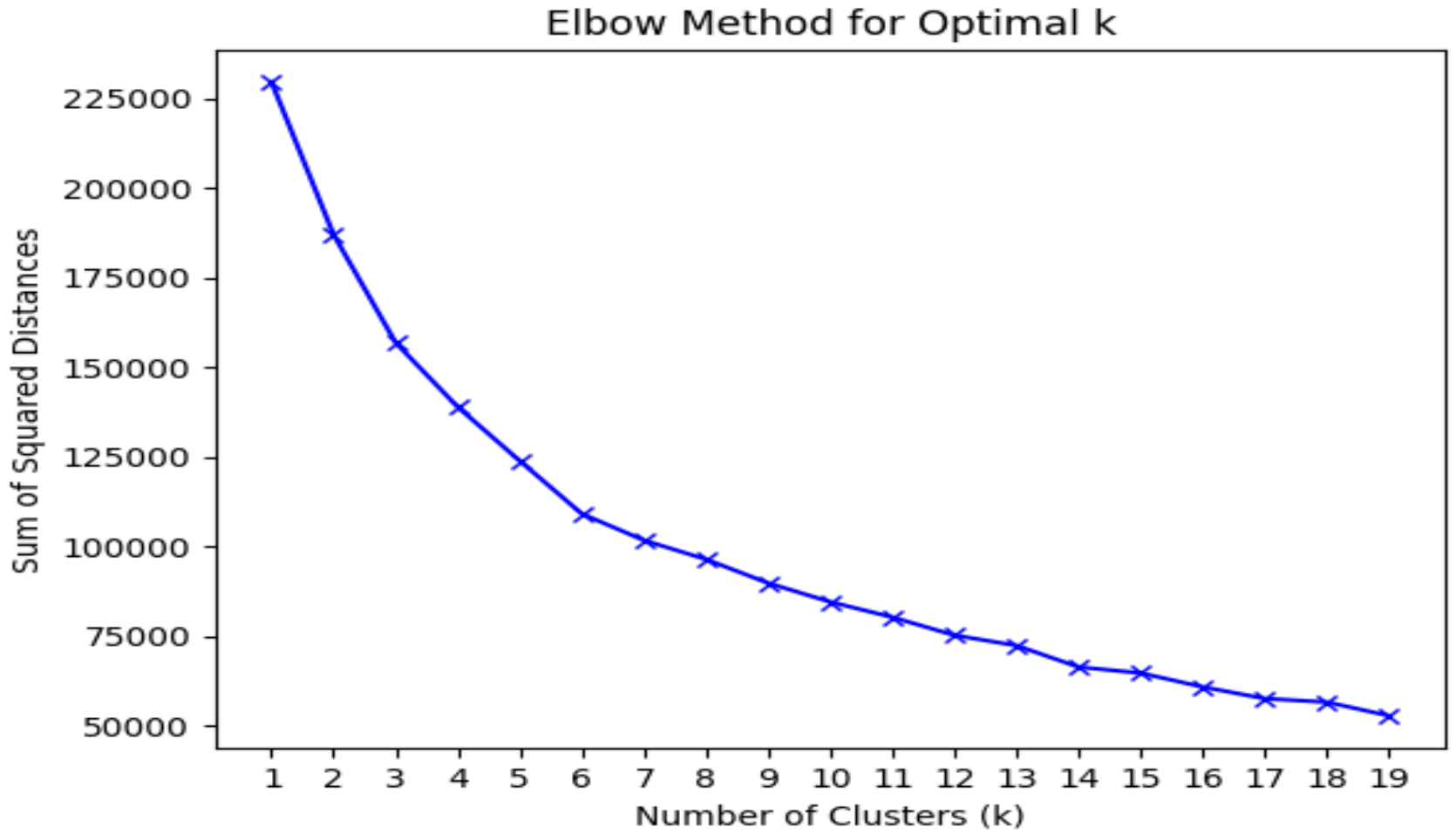
2.DB-Scan



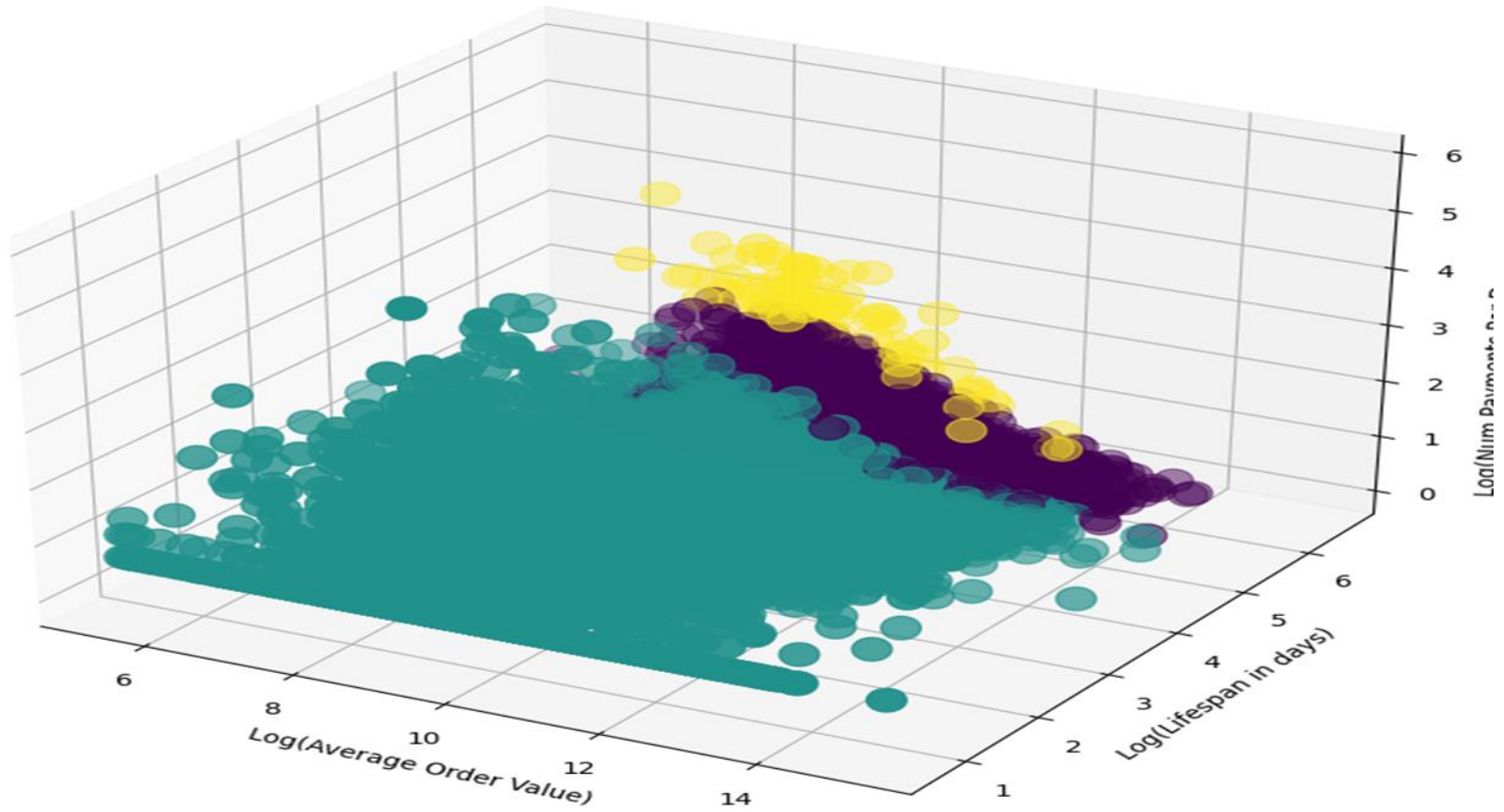
The DBSCAN algorithm resulted in 3 clusters; however, the 3D plot reveals that the clustering distribution appears irregular

3. K-Means

I generated an elbow curve to pinpoint the optimal number of clusters, but the resulting curve didn't showcase a clear elbow point as anticipated.



Experimented with different values of k for clustering and discovered that the distribution when k is set to 3 appears to be more justifiable



Cluster-Based Aggregation and Summary Statistics Reporting

Performed data aggregation on a DataFrame, grouped by a 'cluster' category, to calculate and present summary statistics such as count, sum, mean, and derived metrics like average order value (AOV) and lifetime value (LTV) per person, with results rounded and selected for reporting.

```
Cluster Summary:
         cluster_count  num_payments_per_day_in_life  \
cluster
0                 4656                          0.44
1                 9612                          1.47
2                   80                         11.64

         time_between_first_and_last_payment      aov  ltv_per_person
cluster
0                                     453.95  17247.0    3.204553e+06
1                                      75.02  16439.0    4.040581e+05
2                                     527.34  11031.0    5.639636e+07
```

- Based on the cluster summary statistics , we can infer certain characteristics about the different kinds of businesses within each cluster by examining the metrics like the number of payments, average order value (AOV), lifetime value (LTV) per person, and the timing between payments. Here are three distinct points for each cluster that could suggest the type of businesses they might be:

Cluster 0:

- **Low Frequency, Moderate Value Transactions**: This cluster has a low number of payments per day on average and a moderate AOV, which suggests these could be businesses that transact less frequently but with a moderate transaction value, such as specialty goods providers or businesses with a subscription model with less frequent renewal periods.
- **Longer Customer Relationship**: The average time between the first and last payment is quite long, indicating a longer customer-business relationship. This might be characteristic of businesses that provide long-term services or goods with a long lifespan.
- **Moderate LTV**: The LTV per person is substantial but not the highest among the clusters, which might imply these are established businesses with a solid customer base that maintains steady, albeit not frequent, purchases over time.

Cluster 1:

- **Higher Frequency, Moderate Value Transactions**: With a higher average number of payments per day and a slightly lower AOV compared to Cluster 0, businesses in Cluster 1 could be those that see more frequent purchases but with lower transaction values, such as daily needs providers or convenience stores.
- **Shorter Duration Between Payments**: The time between first and last payment is significantly shorter, indicating a quick turnover of transactions, typical of businesses that deal in fast-moving consumer goods or services.
- **Lower LTV**: The LTV per person is lower than in Cluster 0, suggesting these might be businesses that rely on volume over margin, such as discount stores or those operating in highly competitive markets where customers may not be as loyal.
- **Cluster 2:**
- **High Frequency, High-Value Transactions**: The small number of entities in this cluster have a very high number of payments per day and the lowest AOV among the clusters, suggesting these could be high-traffic businesses with lower transaction values, like fast-food restaurants or cafes.

2.Extended Duration Between Payments: Despite the high frequency of payments, there's a long time between the first and last payment, which could indicate either seasonal businesses or those that have infrequent but recurring high-value transactions.

3.Highest LTV: The exceptionally high LTV per person could signify luxury goods or service providers, high-stake investment or financial services, or perhaps businesses that have a few clients but conduct very large transactions.

Based on the above summary and the inferences drawn from the metrics, here are suitable labels for each cluster that encapsulate their transaction behaviors and potential business types:

Cluster 0: Steady Engagers

Businesses in this cluster may have a steady but infrequent engagement with customers, likely offering specialized or durable goods that do not require frequent purchasing.

Cluster 1: Fast Movers

This cluster appears to consist of businesses that have more frequent, routine interactions with customers, potentially reflecting entities like grocery stores or businesses offering everyday consumables.

Cluster 2: High-Value Niche

The combination of high lifetime value and high payment frequency, despite the small size, suggests that these businesses might deal in luxury goods, offer premium services, or have a small but high-spending customer base.

These labels serve to quickly convey the distinct characteristics of each cluster's transactional and customer engagement patterns.

- Churn Prediction using Logistic Regression

```
merchant_level['churn'].value_counts()
```

```
Churned    7526
Active     6822
Name: churn, dtype: int64
```

The dataset reflects customer attrition figures based on a time frame of 90 days.

```
from sklearn.linear_model import LogisticRegression
from sklearn import metrics


model = LogisticRegression(random_state = 42)
model.fit(X_train_scaled, y_train)


prediction_test = model.predict(X_test_scaled)
metrics.confusion_matrix(y_test, prediction_test)
```

```
array([[1359,   15],
       [   0, 1496]])
```

Trained logistic regression model on scaled training data, made predictions on scaled test data, and calculated the confusion matrix to evaluate the model's performance.

```
accuracy = metrics.accuracy_score(y_test, prediction_test)
print(accuracy)
```

```
0.9947735191637631
```

- Calculated the accuracy of the logistic regression model's predictions
- The accuracy is determined by comparing the predicted values (**`prediction_test`**) against the actual values (**`y_test`**). It represents the proportion of the total number of predictions that were correct and is given by the formula:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

- The **accuracy_score** function from **sklearn.metrics** is used to compute this metric, and the result is a number between 0 and 1, where 1 means perfect accuracy (all predictions are correct) and 0 means no predictions are correct.

Code

- https://colab.research.google.com/drive/1oj7xtjQk69LpddwUD-I2Q31UHIw-5vt_?usp=sharing

- https://colab.research.google.com/drive/1MSeA3TfU49zSWBmO2JjmXC_o5pTI8n8X?usp=sharing (Label Propagation)