We have limited data on these merchants and their transactions, but we are still interested in understanding their payments activity to try to infer the types of merchants using Stripe. Using only the given data, how would you identify different kinds of businesses in the sample? Please generate assignments for each merchant.

A: Initially, the data was prepared to be processed for the different algorithms. After trying to make sense of the data through exploratory analytics, I decided to use the k-means clustering algorithm to observe various clusters formed. The features used were merchant_ids (encoding for the merchant names), date(day, month, year), and amount. To find the optimal value of k, I used the elbow method to observe the error vs the values of k and decided to opt for 3 clusters. The observed findings from these clusters are as follows.

1. Cluster 1:

- The number of transactions is 8297 with the maximum number of transactions in December 2034.
- The number of churned users is **6**, however, 5 of them are also present in cluster 3, so only 1 user churned.
- 1648 unique users within transactions

2. Cluster 2:

- The number of transactions is 52 with the maximum number of transactions in September 2034.
- **No users churned**
- 37 unique users within transactions

3. Cluster 3:

- The number of transactions is 1505370 with the maximum number of transactions in December 2034.
- The number of churned users is **228**, having common churned users with cluster 1.
- 14182 unique users within transactions.

Note: All the unique active users across all the clusters have some intersection as well. Also, k-means clustering could come up with different cluster assignments every time we execute the process, so we need additional business acumen to generalize the results over multiple iterations.

As we can see from the above clusters, Cluster 3 has the highest volume of transactions but also the highest number of churned customers with average spending of $127.26.

Furthermore, Cluster 2 has the smallest number of transactions with 71% of unique users within the transactions and no churned users. Also, the average spending is $55300

Finally, Cluster 1 falls in the middle with 6 churned users and 1 unique churned user with average spending of $4813.44.

Therefore we can say that Cluster 2 could consist of users that are more loyal to the service and could be the sustained users based on past data. Furthermore, Cluster 3 could be users on a limited deal, trying out our service who churned by the end of their tenure. And users from cluster 1 could be seasonal users based on their activity during the holiday season.

Adding definite merchant types to each merchant did not seem the most appropriate just based on the dataset, I would also require some additional domain as well as business knowledge to see the kinds of users using our services, like enterprise customers, retail sellers, etc.

Sometimes a merchant may stop processing with Stripe, which we call churn. We are interested in identifying and predicting churn. Please a) come up with a concrete definition for churn b) identify merchants that have already churned in the dataset, and c) build a model to predict which active merchants are most likely to churn shortly.

A: **Churn**: In simple words, churn is the number of customers that stop using a service during a certain time frame, in our case a transaction. For eg: if we do not observe any transactions from a customer for 5 months, we could say the customer has churned based on the last transaction timestamp.

Furthermore, for identifying customers that have churned, instead of making hard assumptions on time, I truly wanted to understand the users who get churned right from the first quarter of 2033 to the last quarter of 2034. So I decided to calculate the churn quarter over quarter. Initially, I calculated the total number of active customers in Q1 and Q2 of 2033 and calculated the churn of Q1, similarly, I continued the exercise for 4 quarters, to identify all the churned users for the year 2033. Furthermore, on continuing this exercise for 2 years in the end I was able to find the intersection of all the churned users for every quarter over **2** years, which turn out to be **229** users with a churn rate of **1.59%.**

Furthermore, based on this list of churned users, I created a new feature “churned” to the value “yes” for transactions of these users and “no” otherwise.

Finally, to predict the active merchants that are most likely to churn, I identified it as a binary classification problem and decided to apply 4 different classification algorithms namely,

1. Logistic Regression,

2. K-Nearest Neighbours,

3. Naive Bayes

4. Decision Trees

Even before, moving ahead with the algorithm, I identified a huge class imbalance in the dataset, so I decided to SMOTE (Synthetic Minority Oversampling Technique) the dataset, which proportioned the dataset equally by oversampling the instances of the minority class. Additionally, the train-test split of the data was 80%-20% respectively.

The results are as follows.

```
                Model  Accuracy  Precision    Recall  F1 Score  F2 Score
0  Logistic Regression  0.845642   0.024202  0.998277  0.047259  0.110314
                Model  Accuracy  Precision    Recall  F1 Score  F2 Score
0  K-Nearest Neighbours  0.997457   0.770401  0.479759  0.591295  0.518912
         Model  Accuracy  Precision    Recall  F1 Score  F2 Score
0  Naive Bayes  0.746086   0.014605  0.981051  0.028781   0.06892
           Model  Accuracy  Precision    Recall  F1 Score  F2 Score
0  Decision Trees  0.999069   0.833207  0.946598   0.88629  0.921516
```

From the results, we can see we have high values of accuracy for Decision Trees and K-Nearest Neighbours, we would still need more data to truly test the limits of our model.