Elo Merchant Category Recommendation

This project is intended to help understand customer loyalty and build a recommendation engine with discount from credit card provider

Overview

This project focuses on

Data Wrangling – Methods used to transform data into statistical usable format

EDA – Visual insights into data and correlation

Introduction

ELO, one of the largest payment brands in Brazil, has built partnerships with merchants in order to offer promotions or discounts to cardholders.

Data is at https://www.kaggle.com/c/elo-merchant-category-recommendation/data

This project intends to clean data and perform EDA.

This project is divided into three parts **Data Wrangling** and **EDA.**

Data Dictionary

There are 6 Data sets

1. **train.csv** - contain card_ids and information about the card itself - the first month the card was active, etc. train.csv also contains the target

2. **test.csv** - contain card_ids and information about the card itself - the first month the card was active, etc.

3. **historical_transactions.csv** - designed to be joined with train.csv, test.csv, and merchants.csv. They contain information up to 3 months' worth of historical transactions for each card_id

4. **new_merchant_transactions.csv -** designed to be joined with train.csv, test.csv, and merchants.csv. They contain information about two months' worth of data for each card_id containing ALL purchases that card_id made at merchant_ids that were not visited in the historical data

5. **merchants.csv** - additional information about all merchants / merchant_ids in the dataset. Merchants can be joined with the transaction sets to provide additional merchant-level information.

6. **sample_submission.csv** - a sample submission file in the correct format - contains all card_ids you are expected to predict for.

Data Wrangling

Following data cleaning methods are used **merchant.csv**

- **Missing Data**
  - Columns having inf are replaced first with NaN and then are imputed based on datatype of column as described below.
  - Columns with object datatype having NaN values are imputed with "other"
  - Columns with int and float datatype having NaN values are imputed with median
- **Outliers -** Outlier identification is applied for following columns. Other columns are either categorical or ID's. **3-Sigma** Rule is applied to impute outliers.
  - numerical_1
  - numerical_2
  - avg_sales_lag3

Data Wrangling

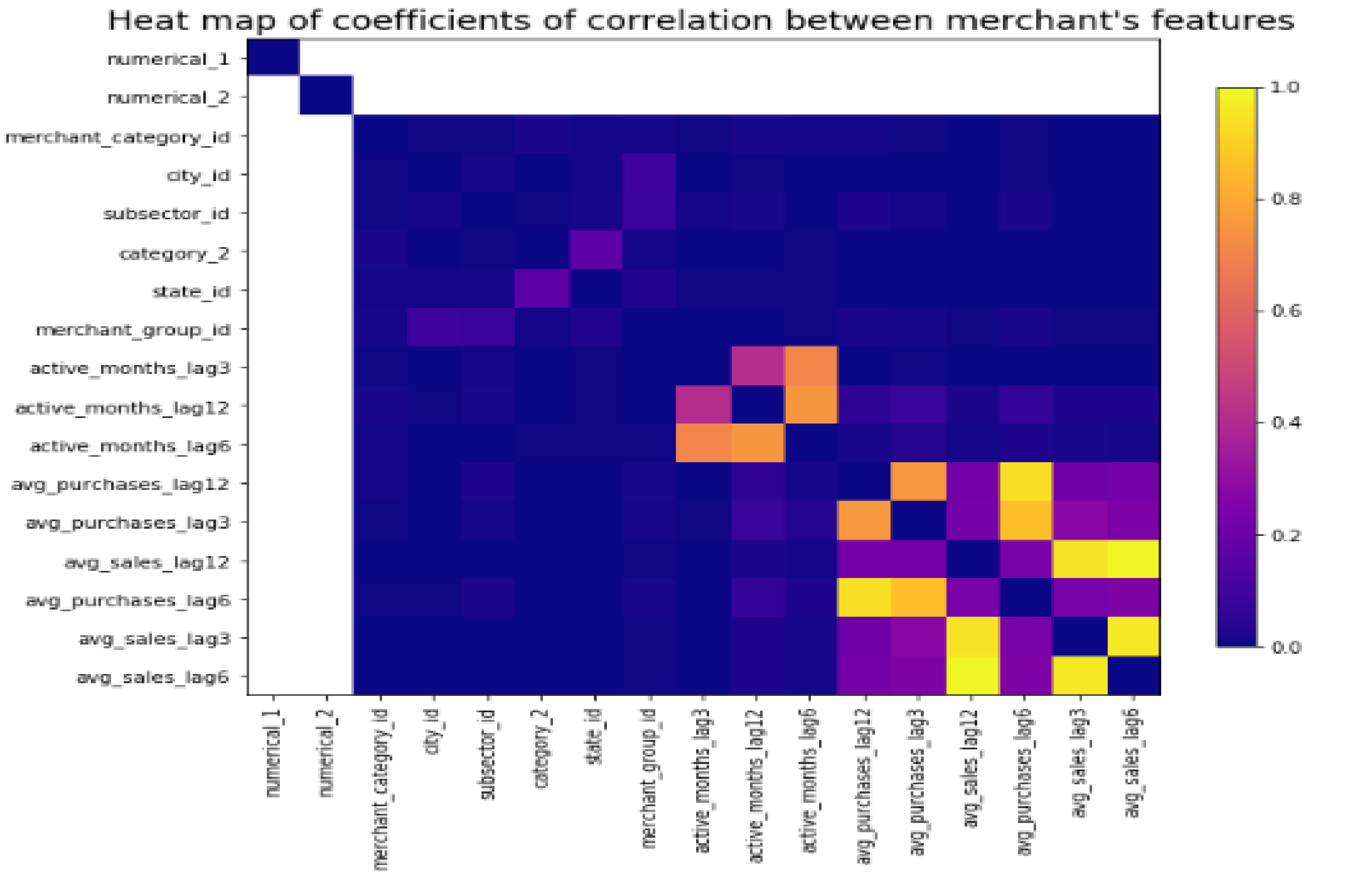
- **Outliers -** contniued
  - avg_purchases_lag3
  - avg_sales_lag6
  - avg_purchases_lag6
  - avg_sales_lag12
  - avg_purchases_lag12

Data Wrangling

- For datasets **historical_transactions.csv** and **new_merchant_transactions.csv** –
  - **Missing values** (**NaN**)are imputed with "**other**" for columns with object datatype, **median** for columns with int and float datatype, and **new** category is added for columns with categorical datatype.
  - **Outliers** are imputed with **3-Sigma** rule for columns "**purchase_amount**" and "**installments**"
- Datetime features are created for "**purchase_date**"
  - Purchase year
  - Purchase month
  - Purchase day of the week
  - Purchase week of the year
  - Purchase weekend
  - Purchase hour
  - month difference - difference in numbers of months from current date to purchase date

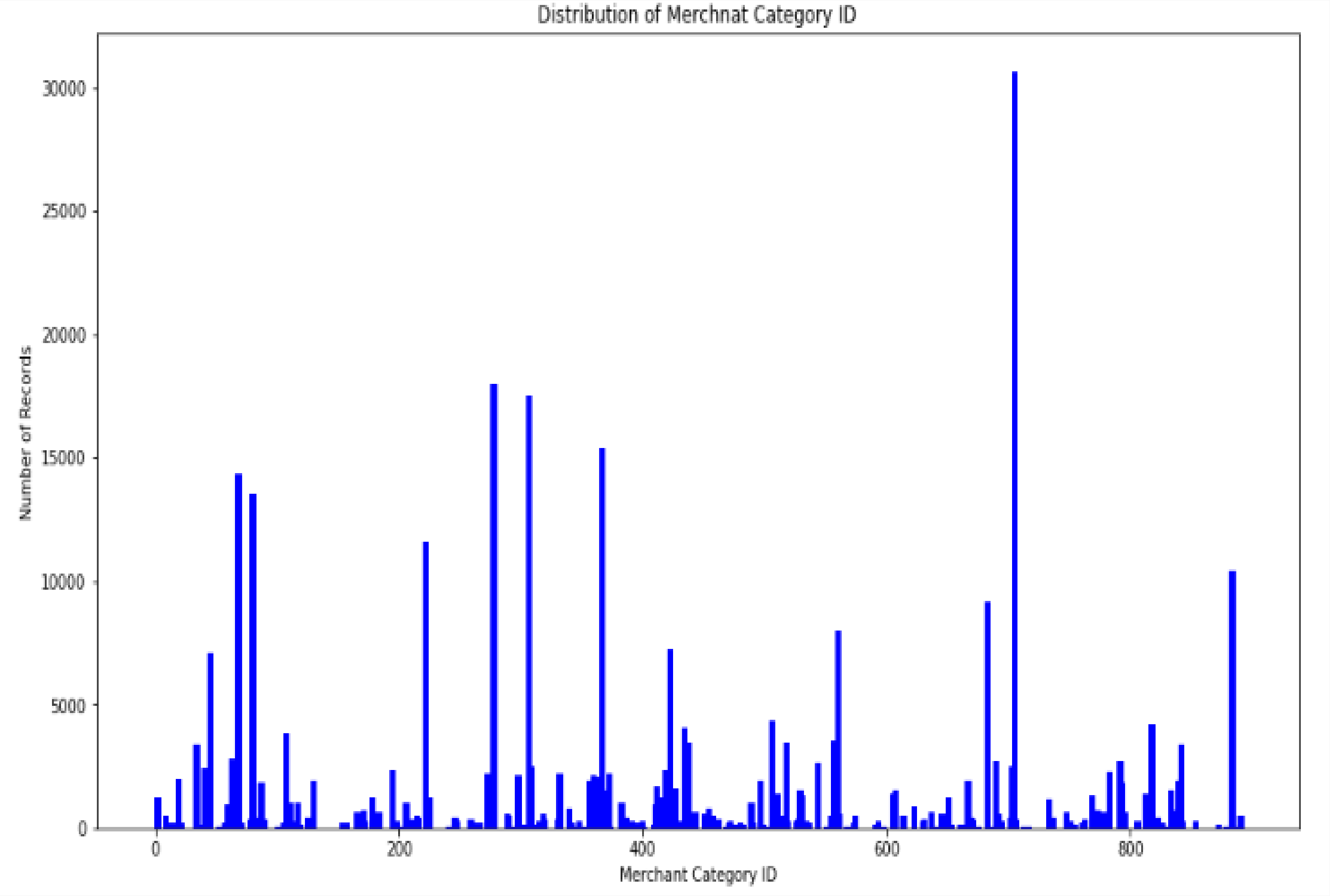
EDA –merchant.csv data

- There is no corelation numerical_1 and numerical_2 feature.
- There is correlation between avg_sales and avg_purchases of 3, 6 an 12 month.



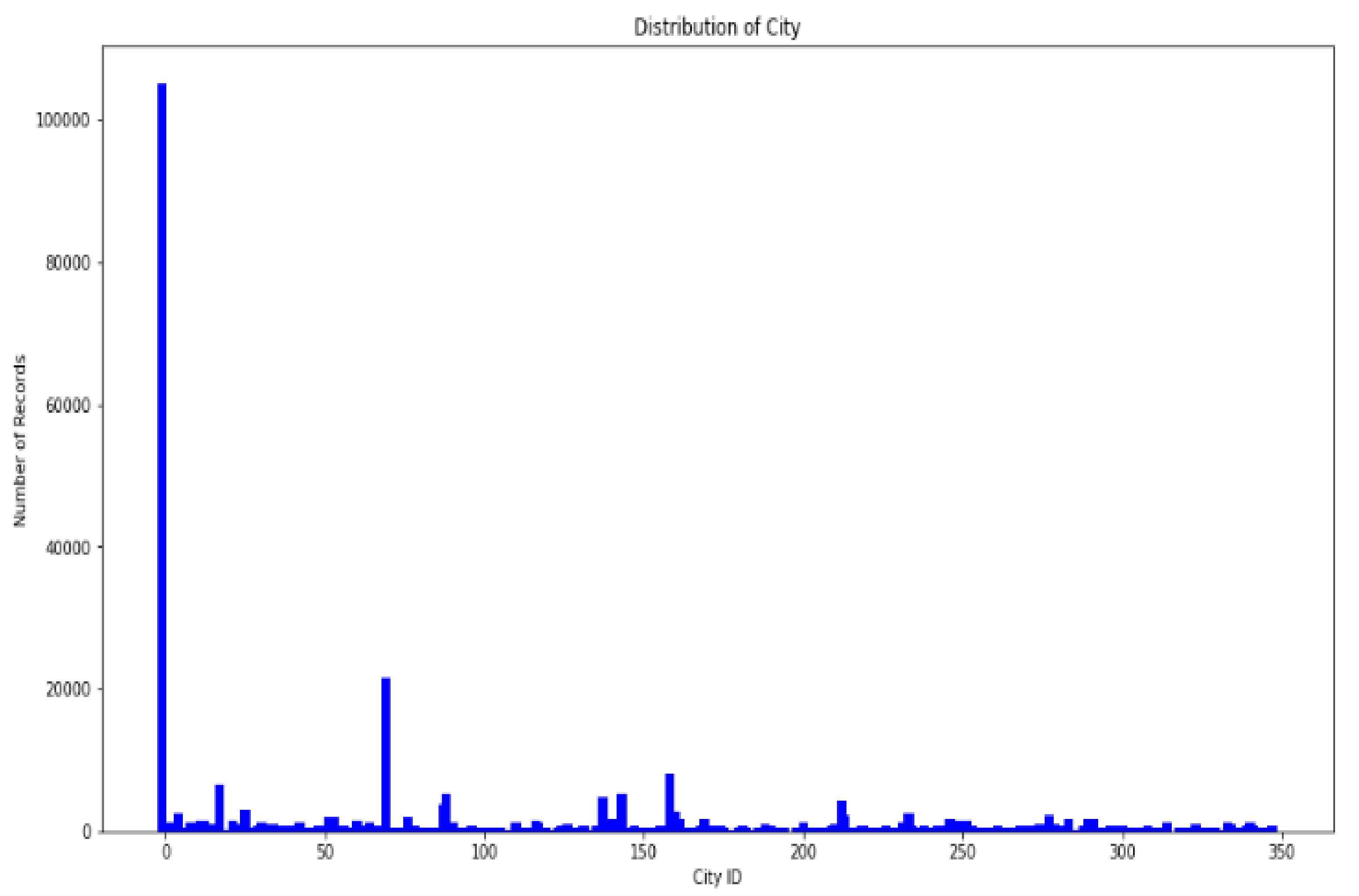
EDA –merchant.csv data

Merchant category ID 705 is the most famous merchant category with 9% sales



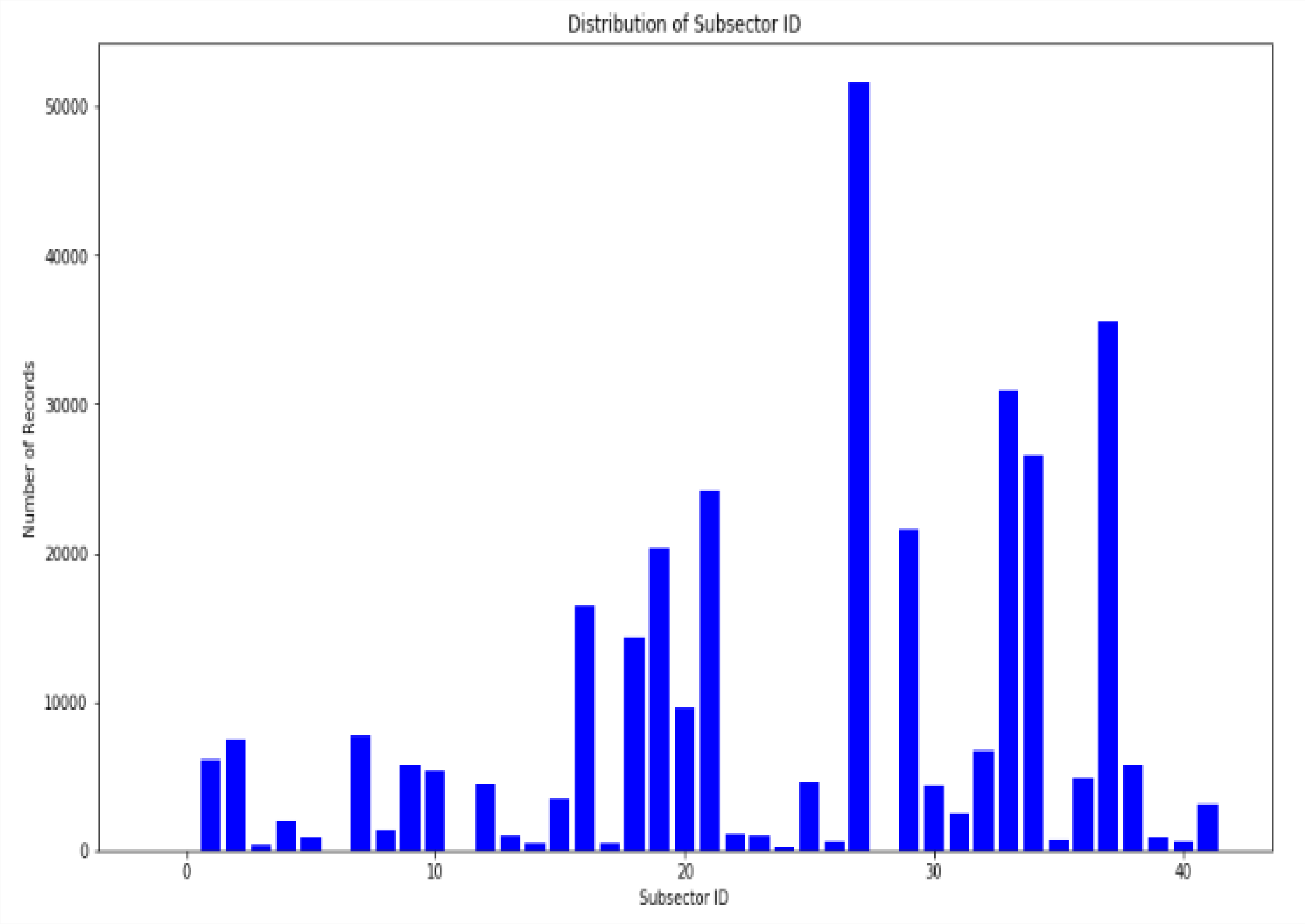
EDA –merchant.csv data

City ID -1 has over 100000 transactions and amounts to 31% of transactions



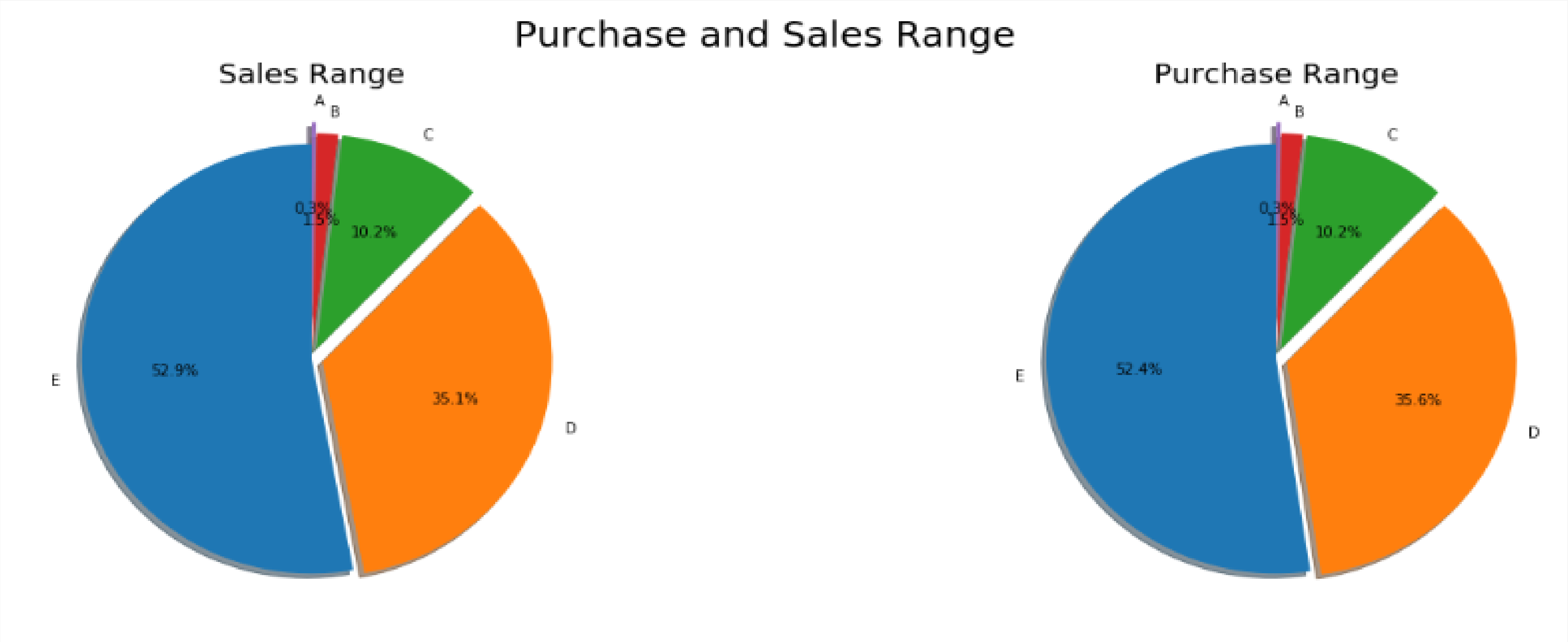
EDA –merchant.csv data

Subsector ID 27 has over 50000 transactions and amounts to 15% of transactions



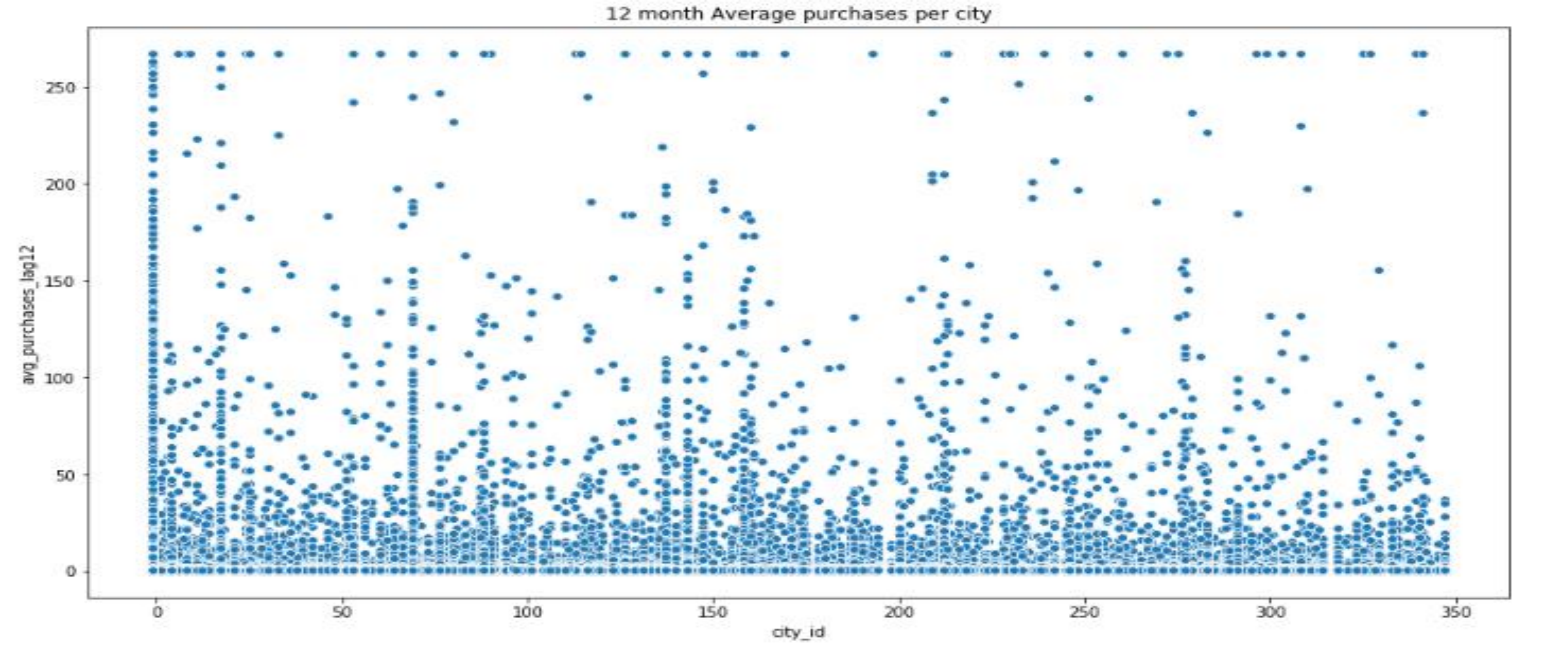
EDA –merchant.csv data

EDA –merchant.csv data



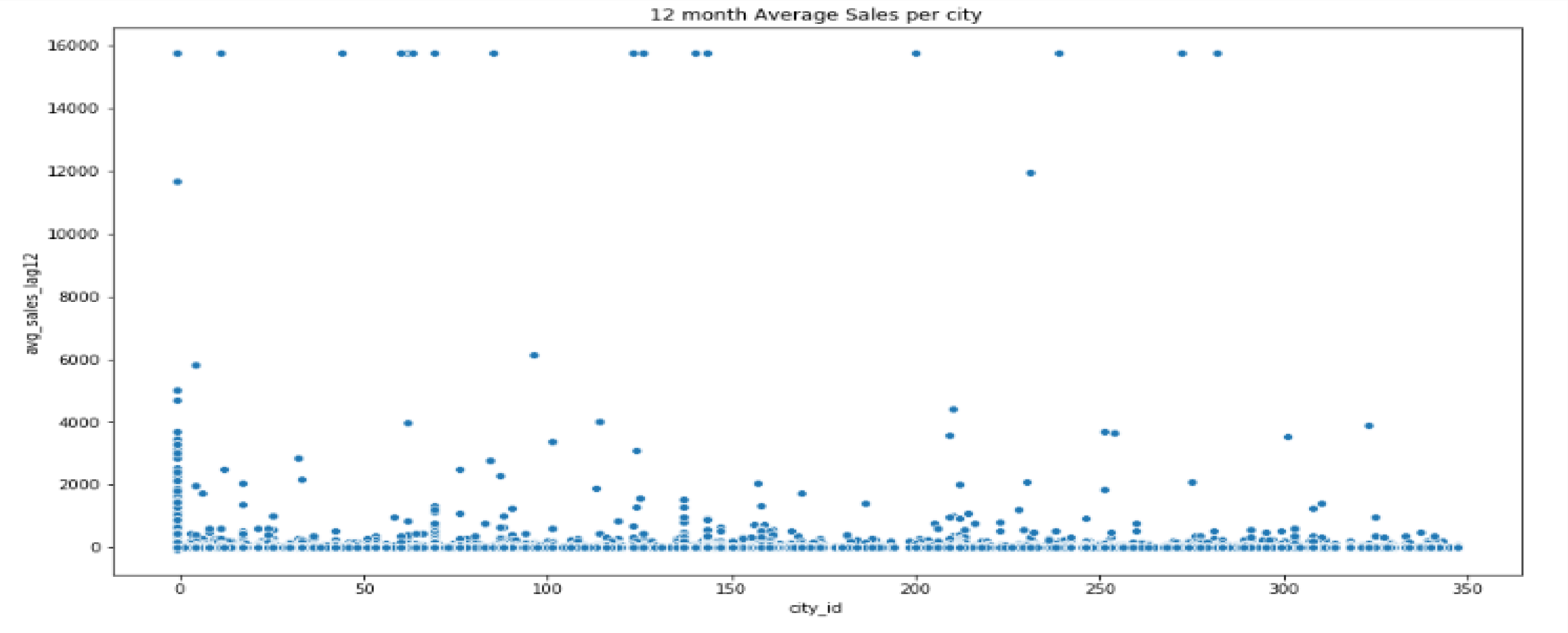
EDA –merchant.csv data

12 Month average purchases distribution per city



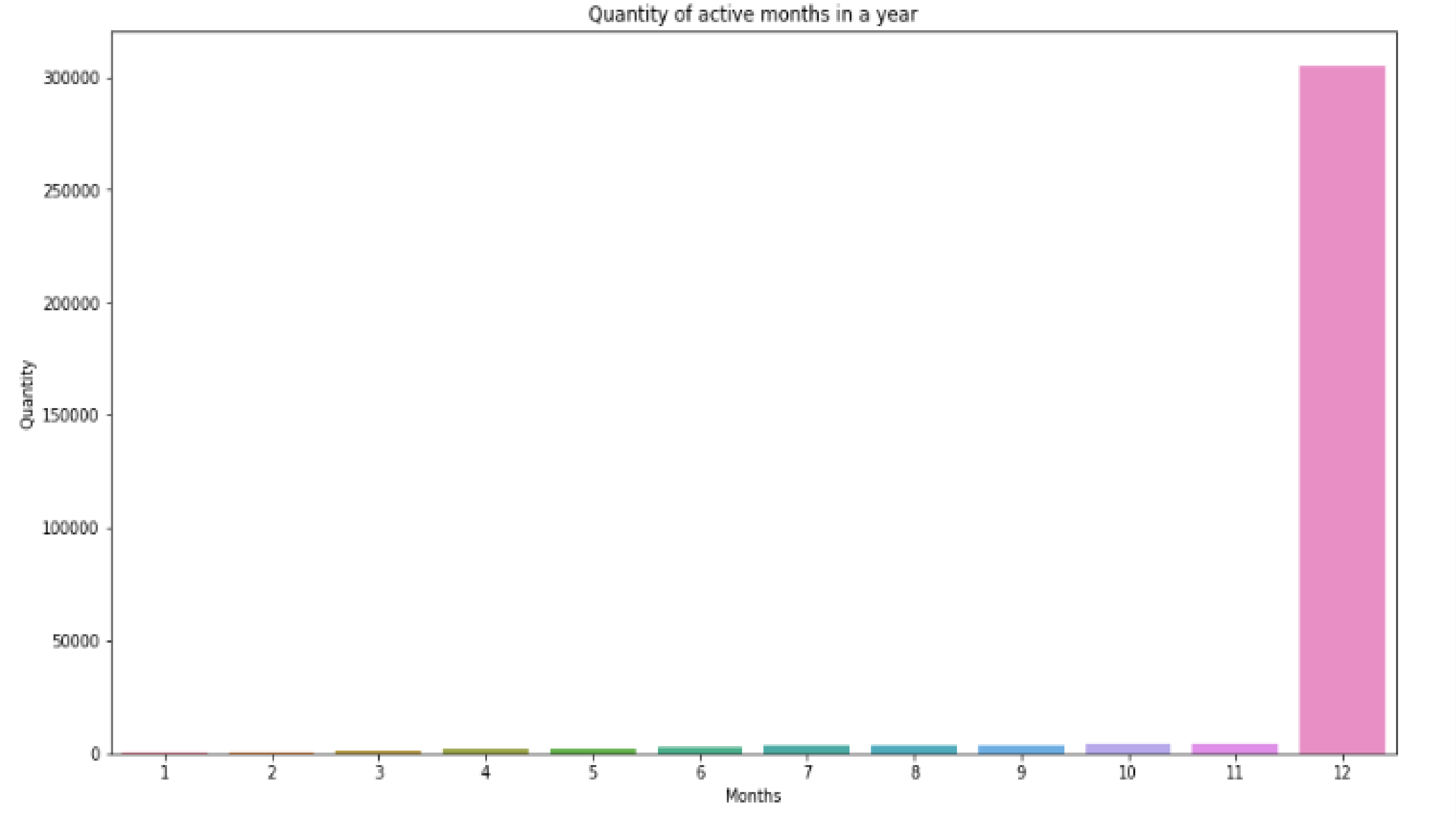
EDA –merchant.csv data

12 Month average sales distribution per city



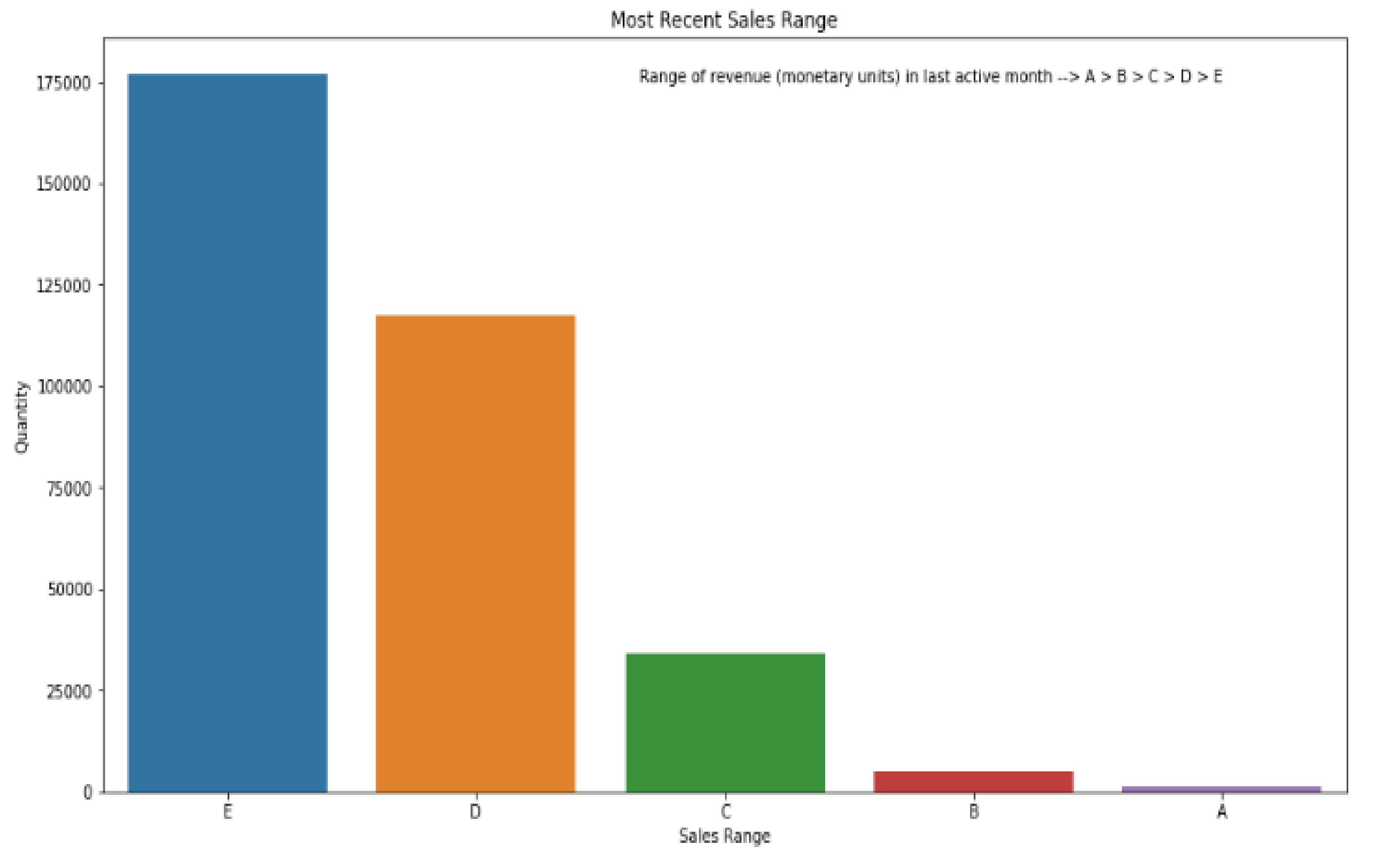
EDA –merchant.csv data

Most Sales are in the month of December



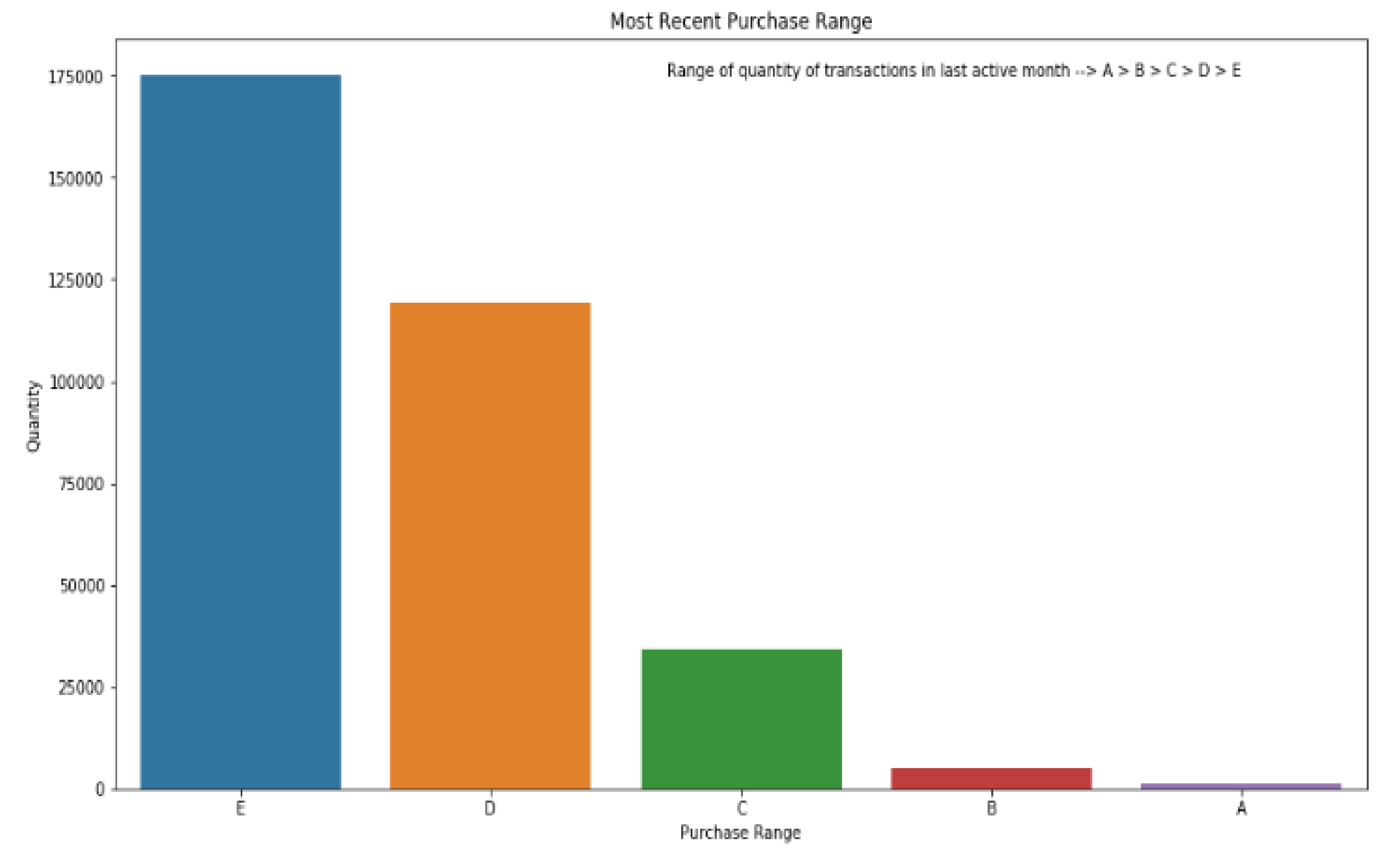
EDA –merchant.csv data

Most number of sales are in E category Range.



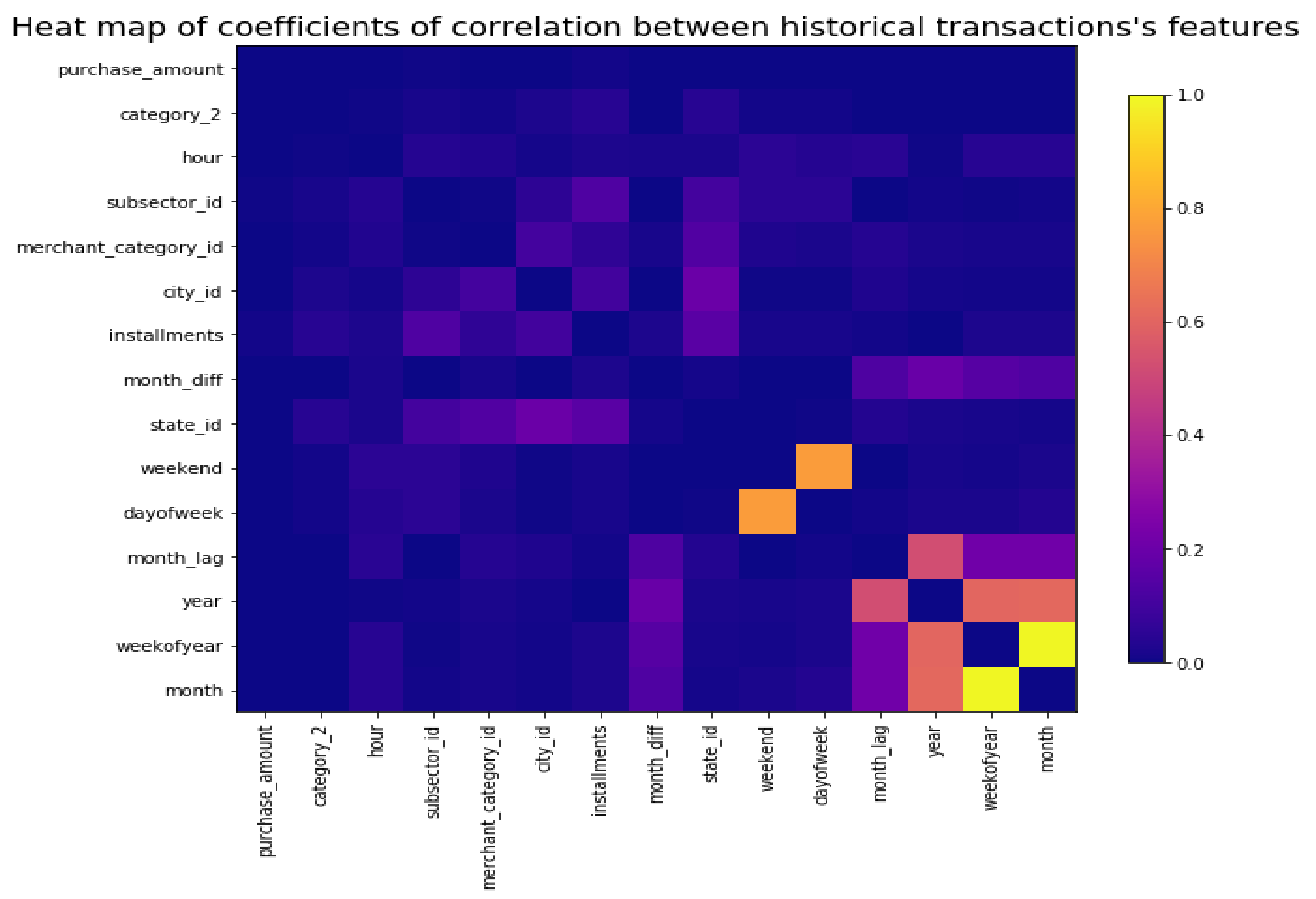
EDA –merchant.csv data

Most number of purchases are in E category Range.



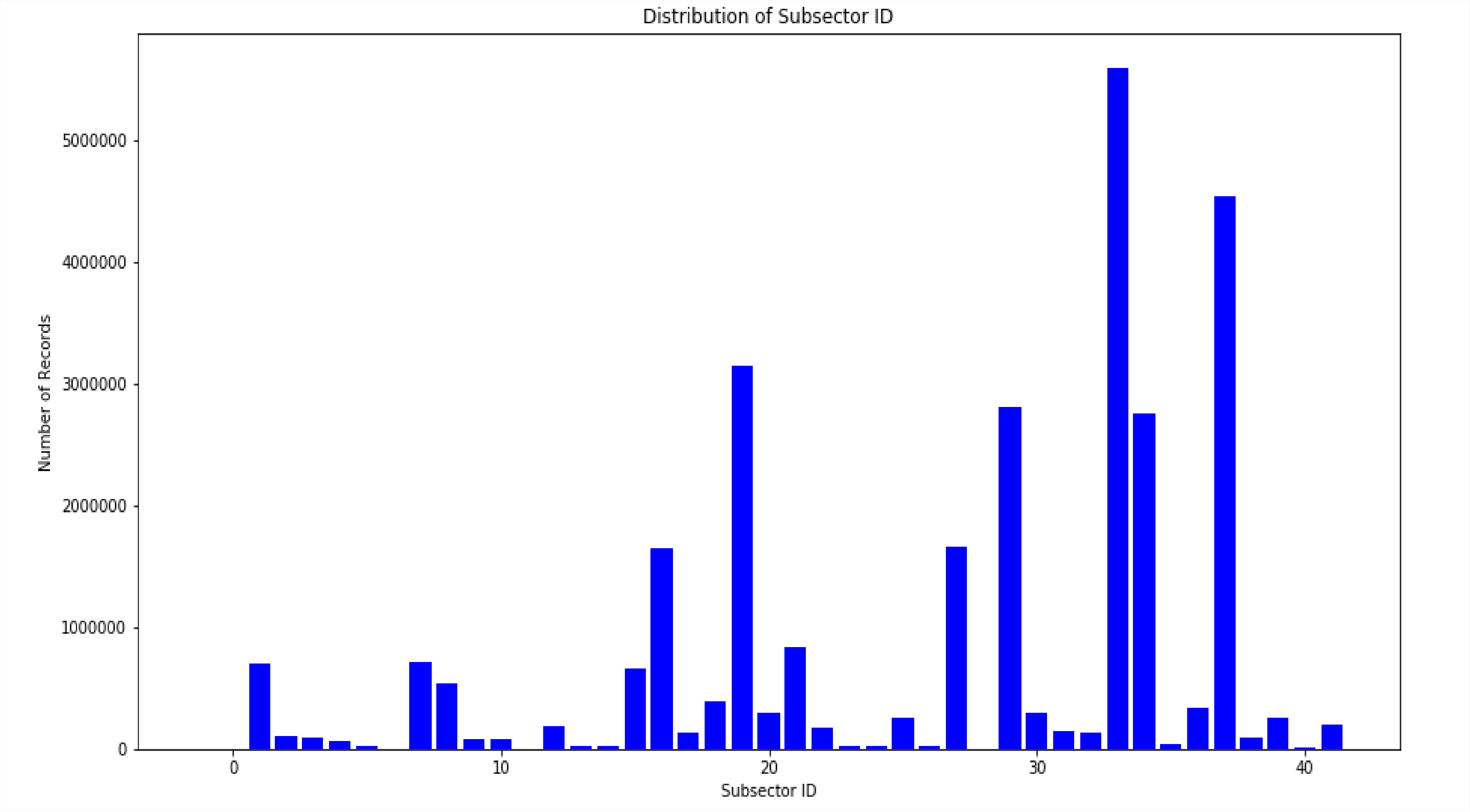
EDA –historical_transactions.csv data

There seems to be no correlation between features.



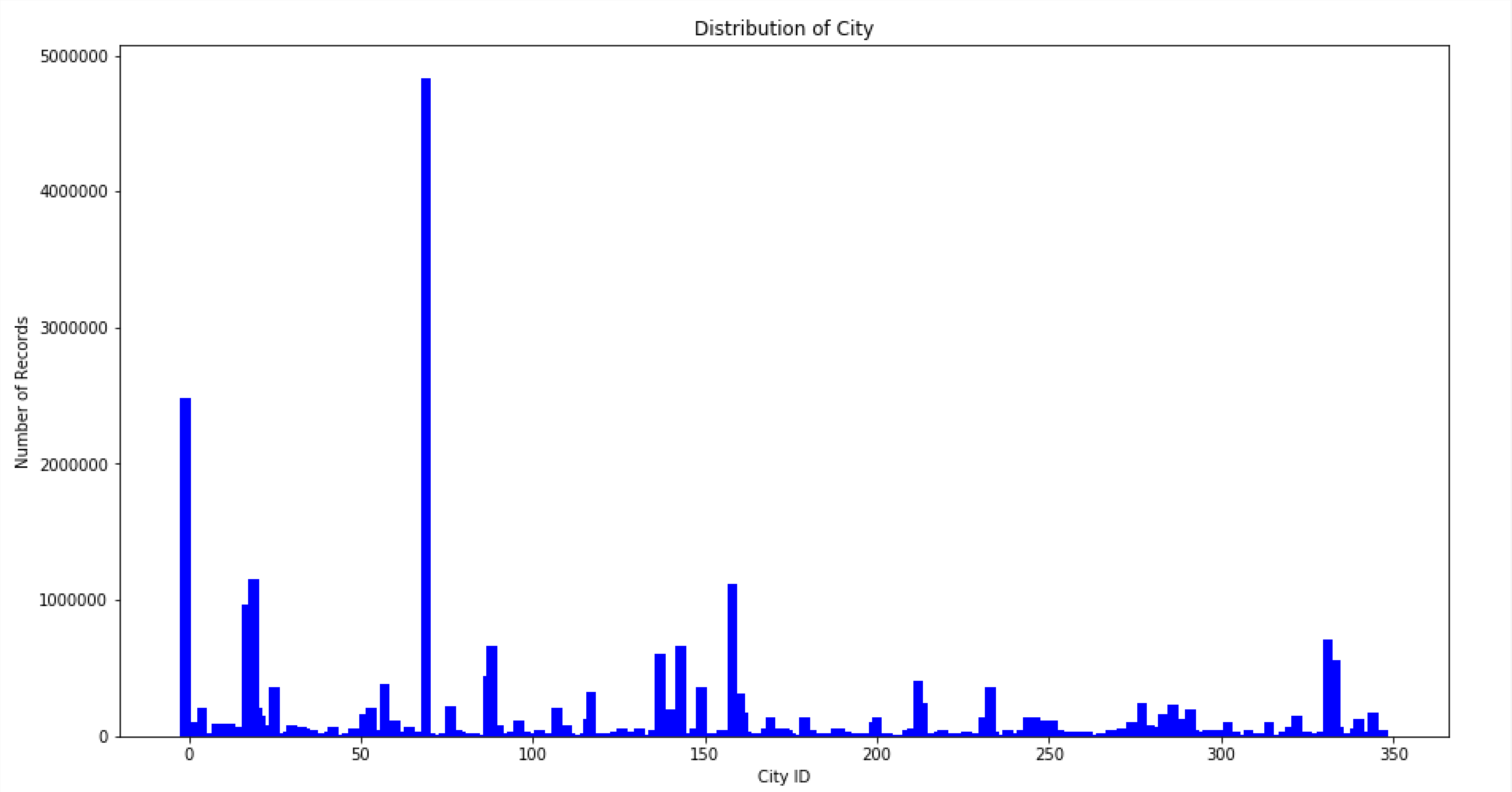
EDA –historical_transactions.csv data

Subsector ID 33 has over 5000000 transactions and amounts to 19% of transactions

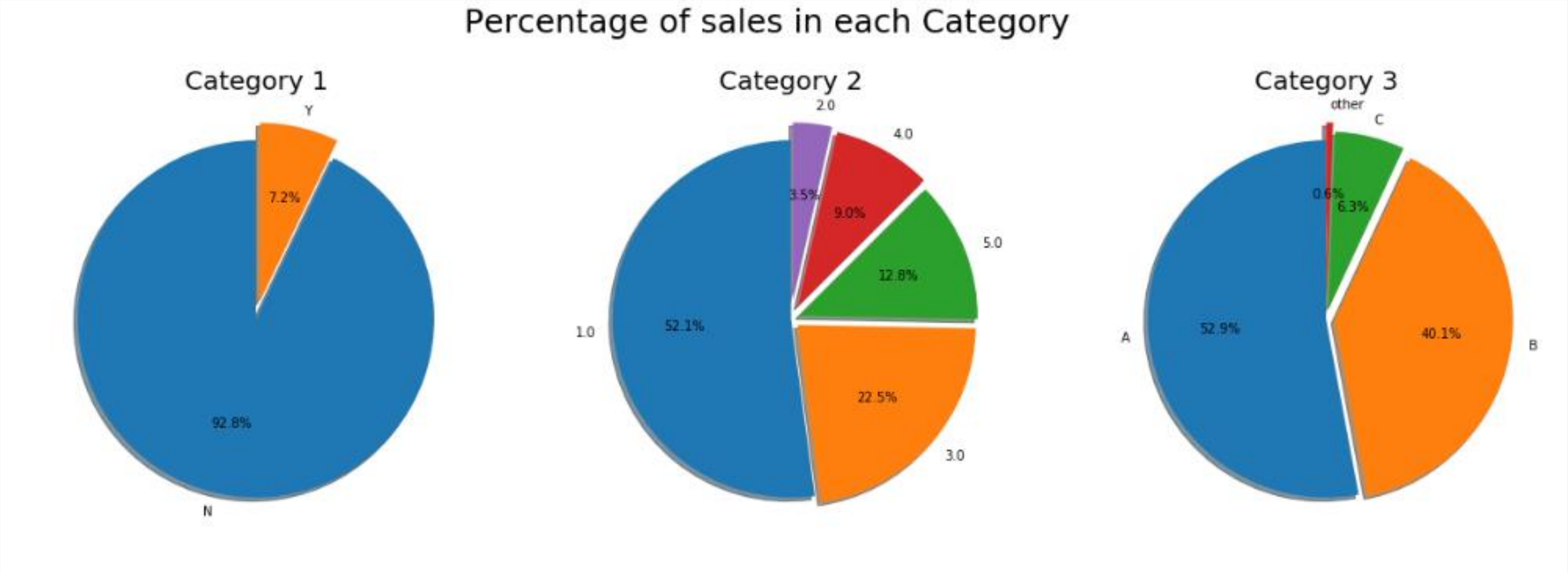


EDA –historical_transactions.csv data

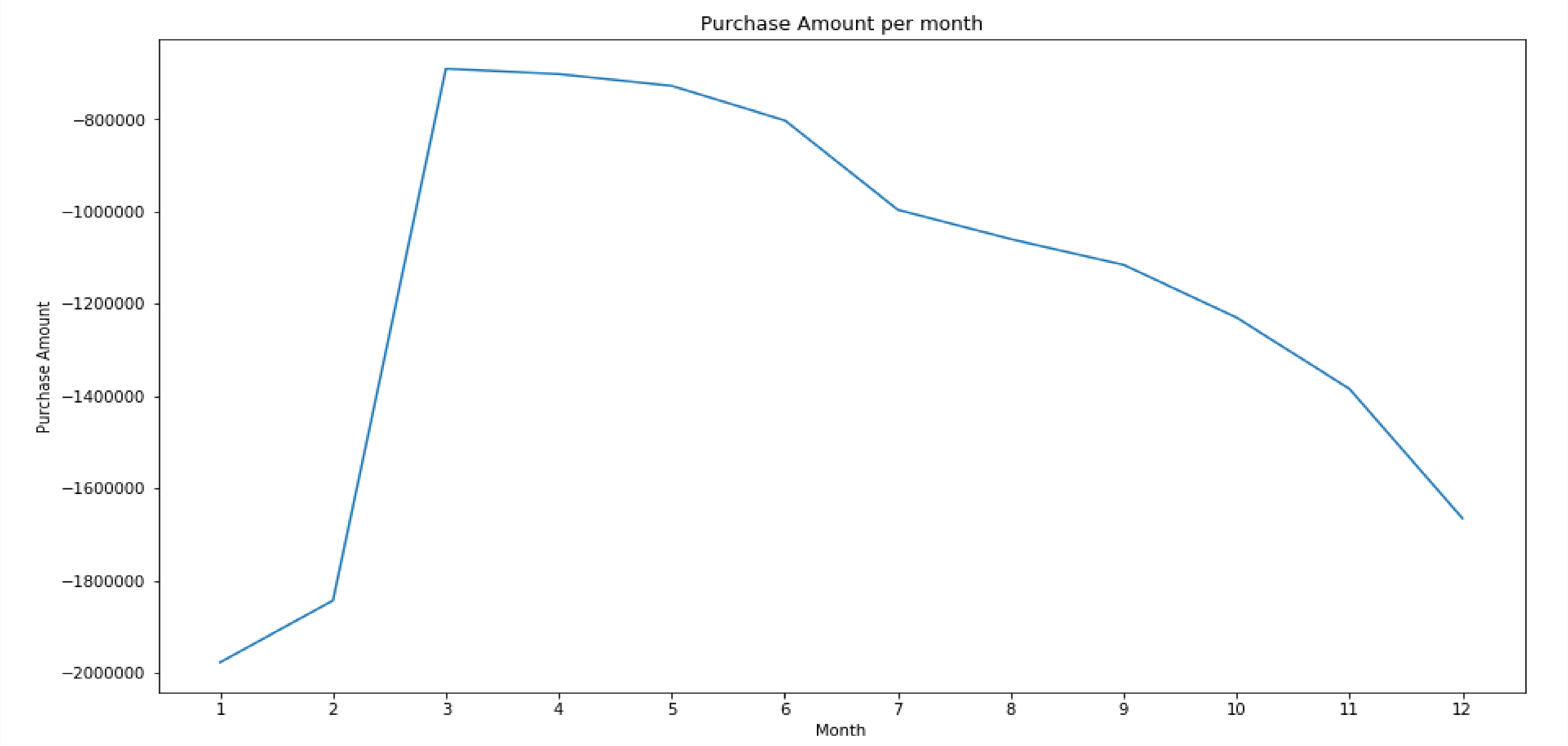
City ID 33 has over 4000000 transactions and amounts to 16% of transactions



EDA –historical_transactions.csv data

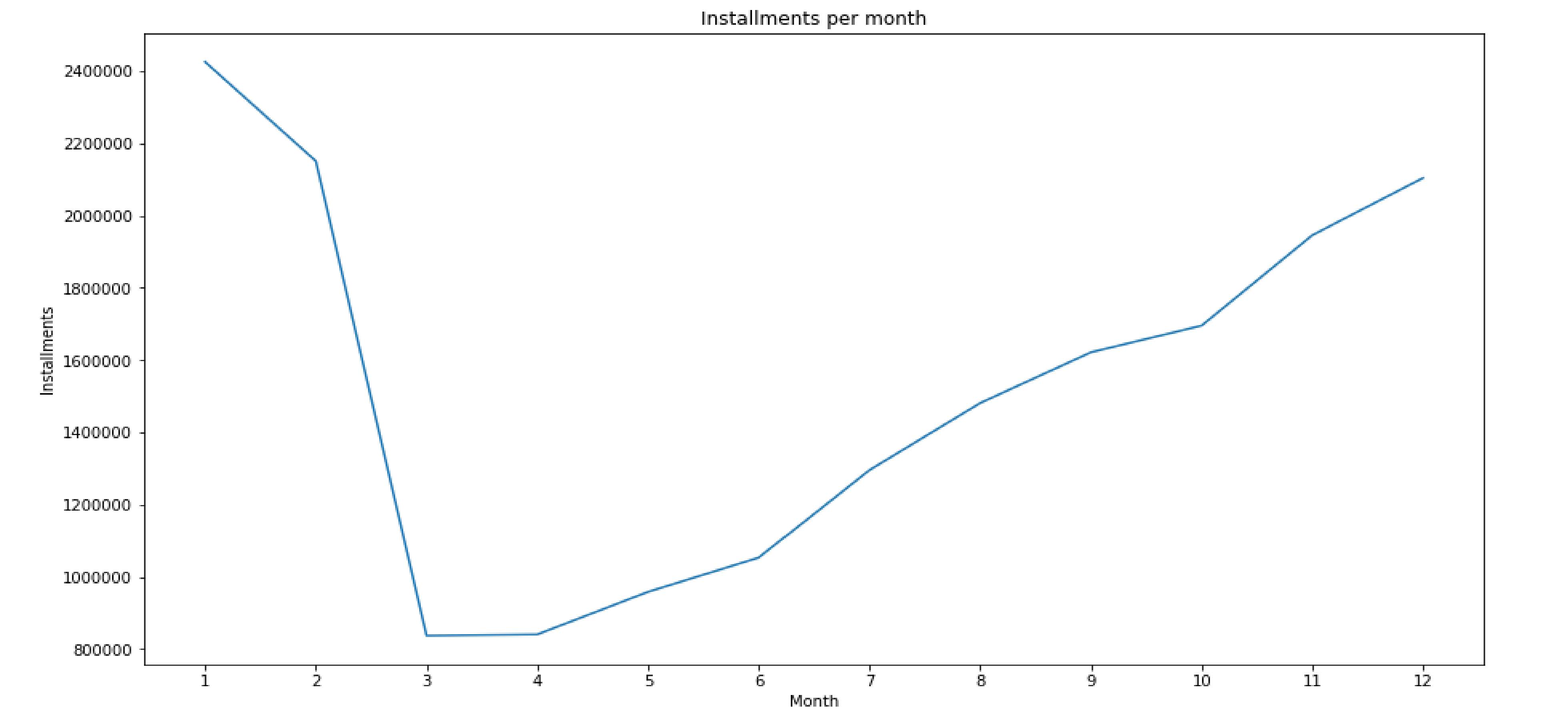


EDA –historical_transactions.csv data



March has most purchases per month.

EDA –historical_transactions.csv data



January has most installments per month.

EDA –historical_transactions.csv data

Most number of purchases are not part of category 1.

Highest number of purchase in category 2 are in **1.0**.

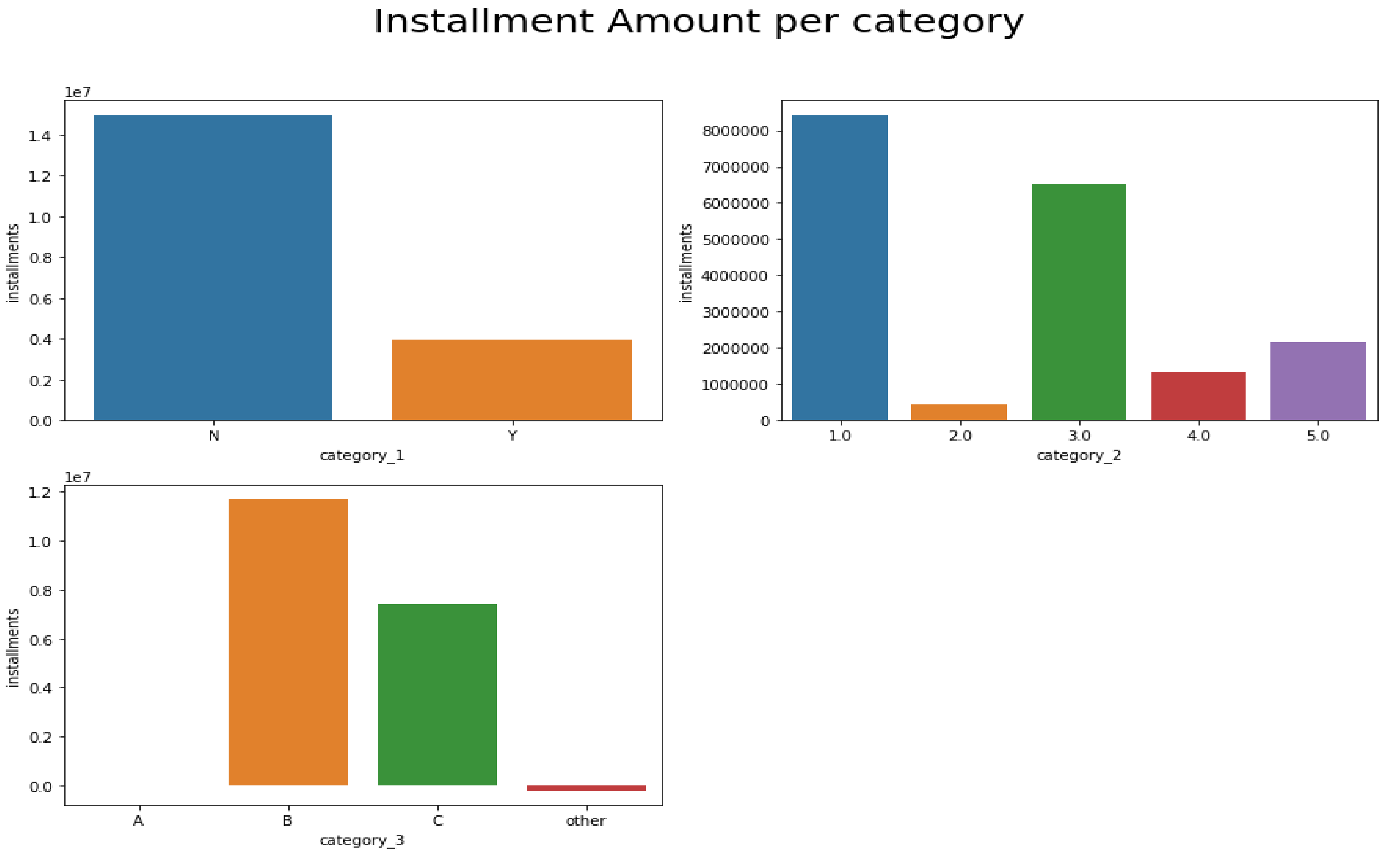
Highest number of purchase in category 3 are in **A**.

EDA –historical_transactions.csv data

Most number of installments are not part of category 1.

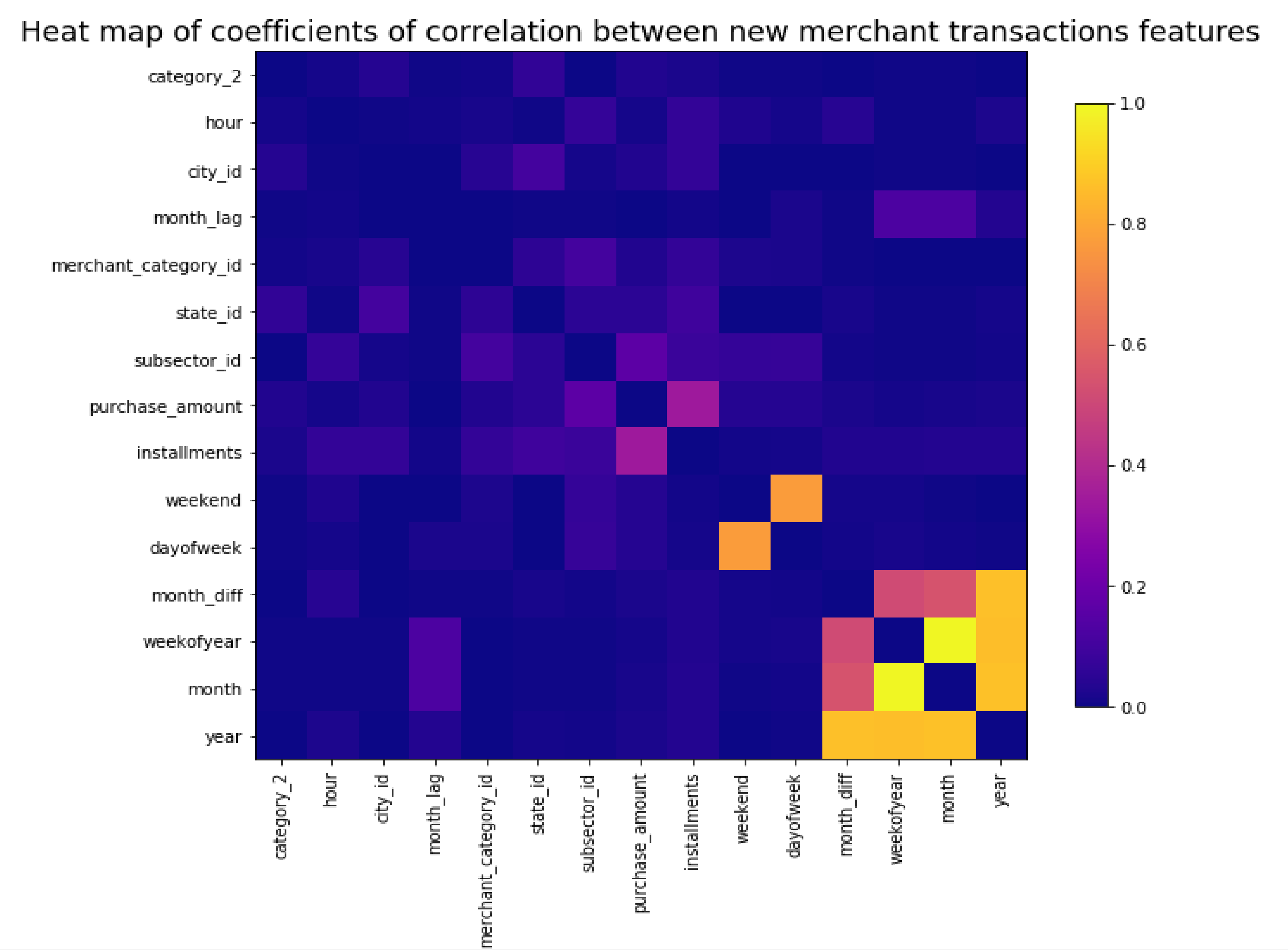
Highest number of installments in category 2 are in **1.0**.

Highest number of installments in category 3 are in **B**.



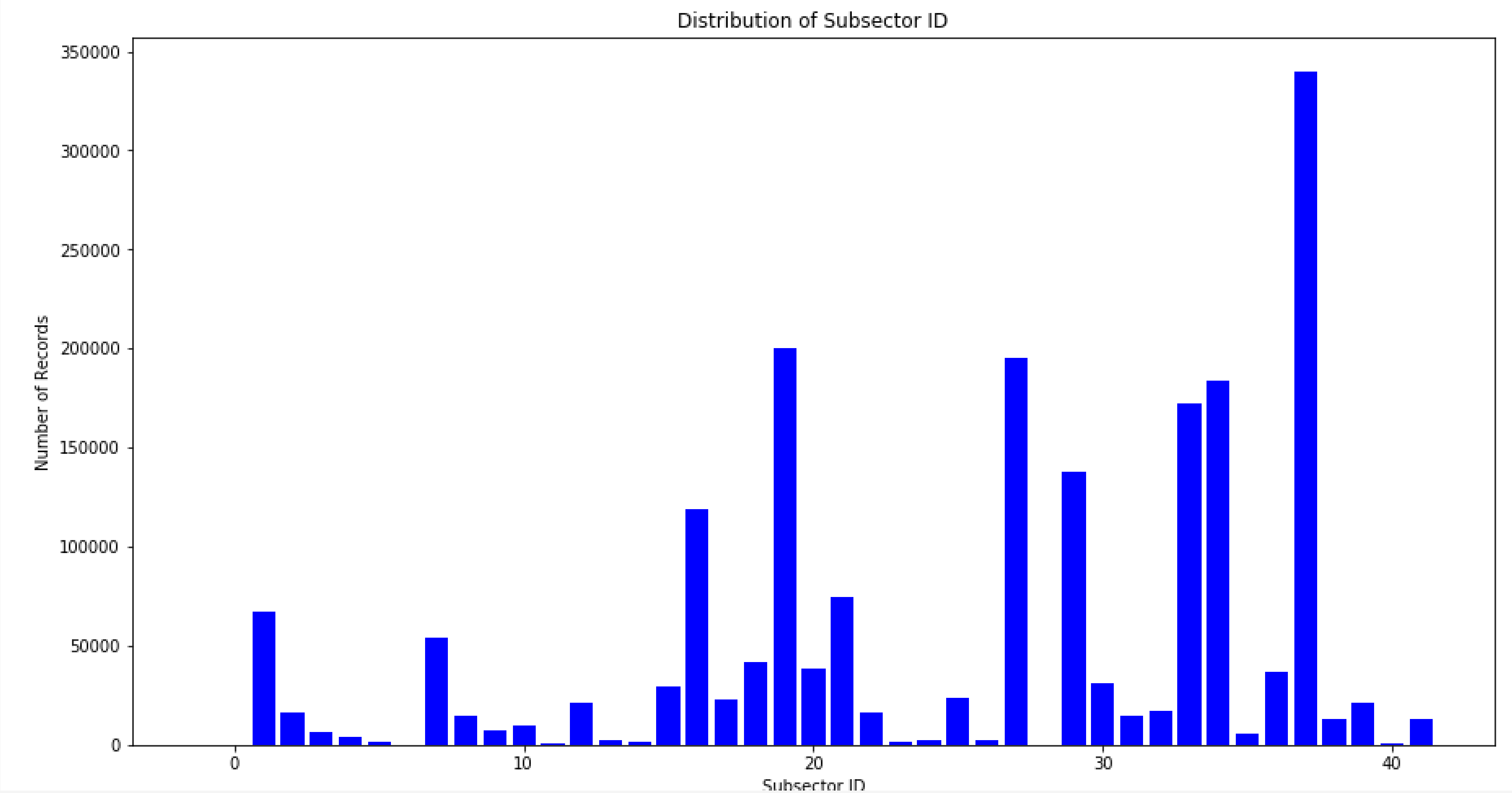
EDA – newMerchant_transactions.csv data

There seems to be a correlation purchase amount and number of installments.



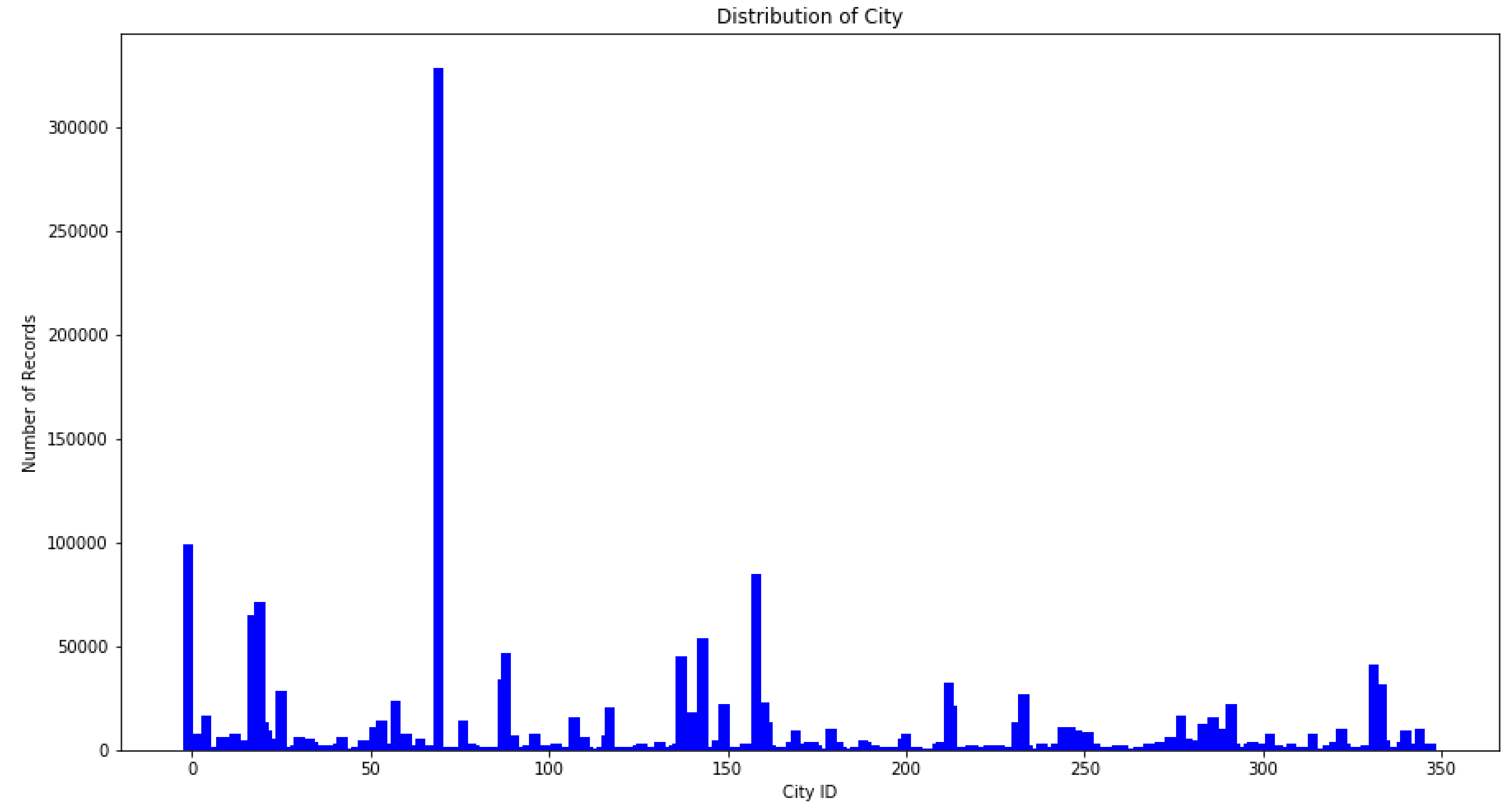
EDA – newMerchant_transactions.csv data

Subsector ID 37 has over 340053 transactions and amounts to 17% of transactions

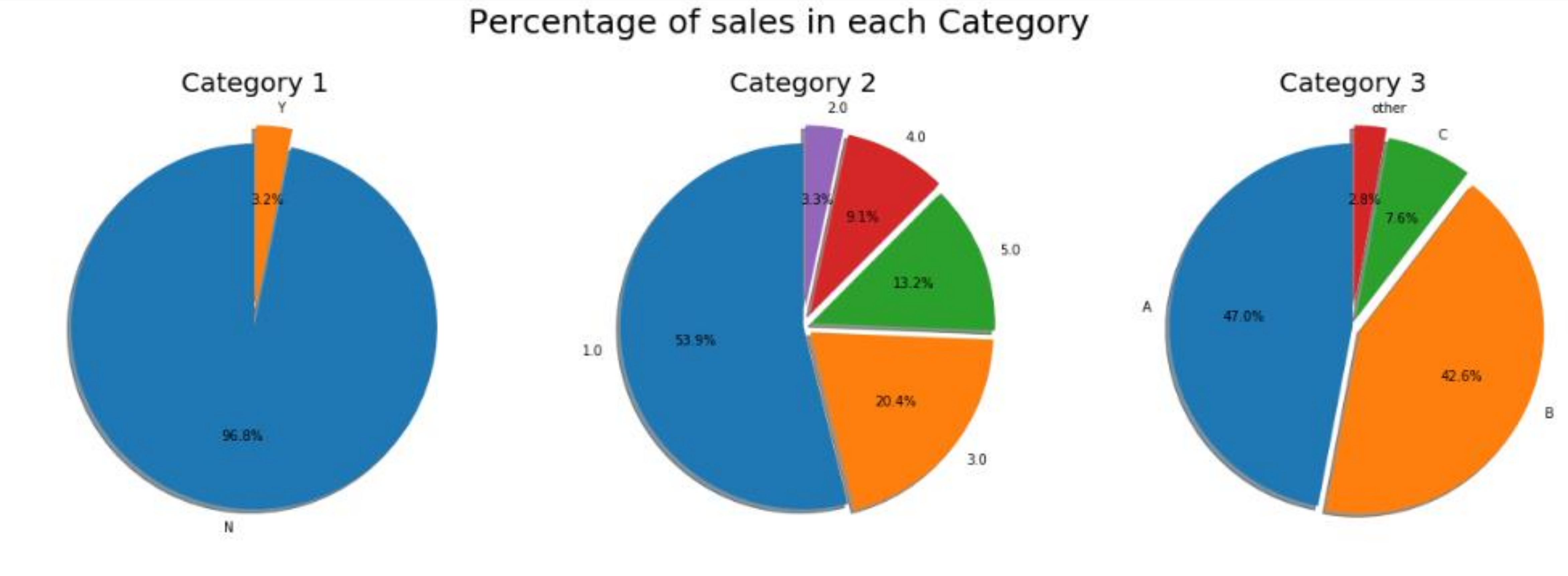


EDA – newMerchant_transactions.csv data

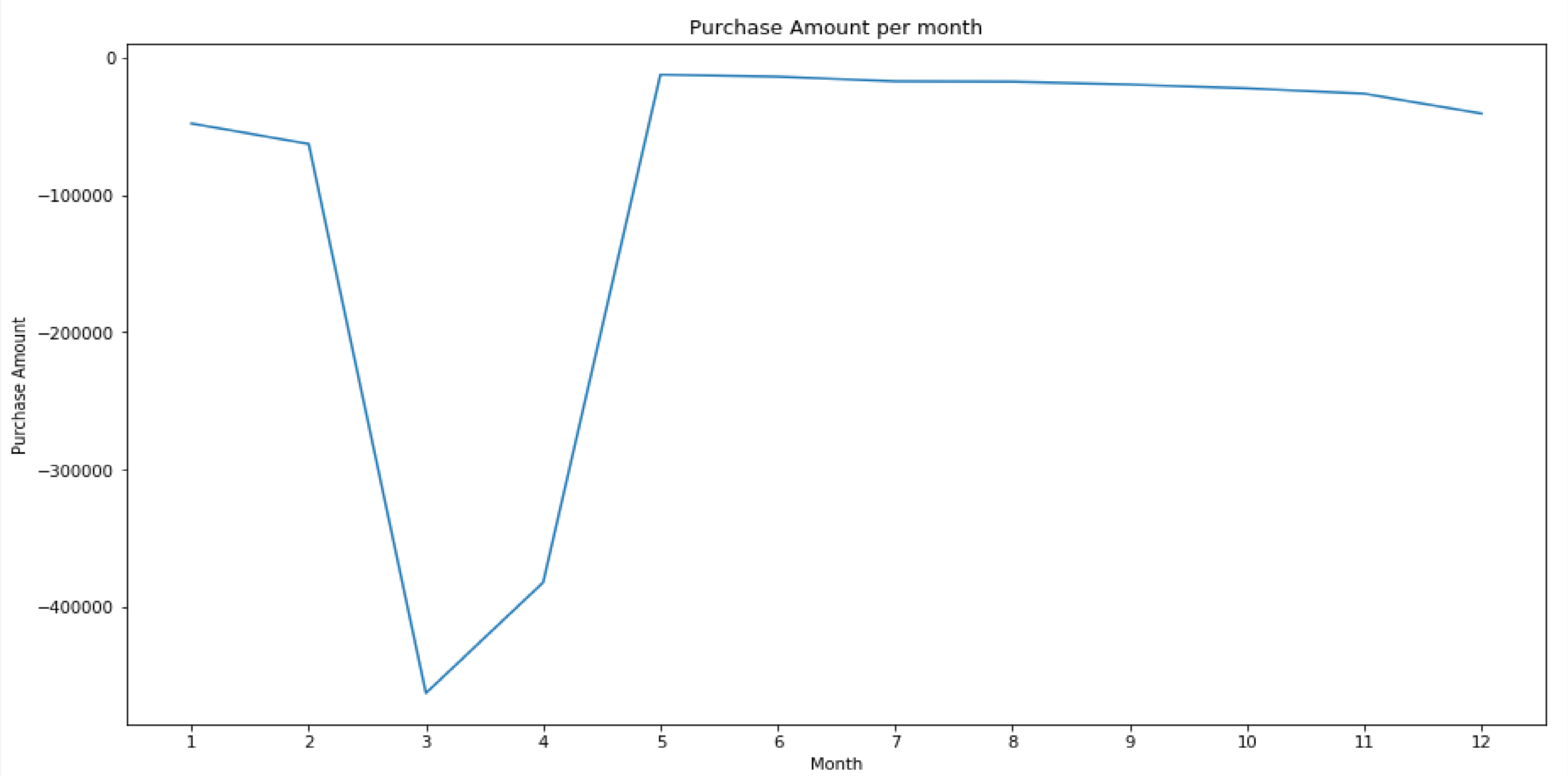
City ID 69 has 328916 transactions and amounts to 17% of transactions



EDA –EDA – newMerchant_transactions.csv data

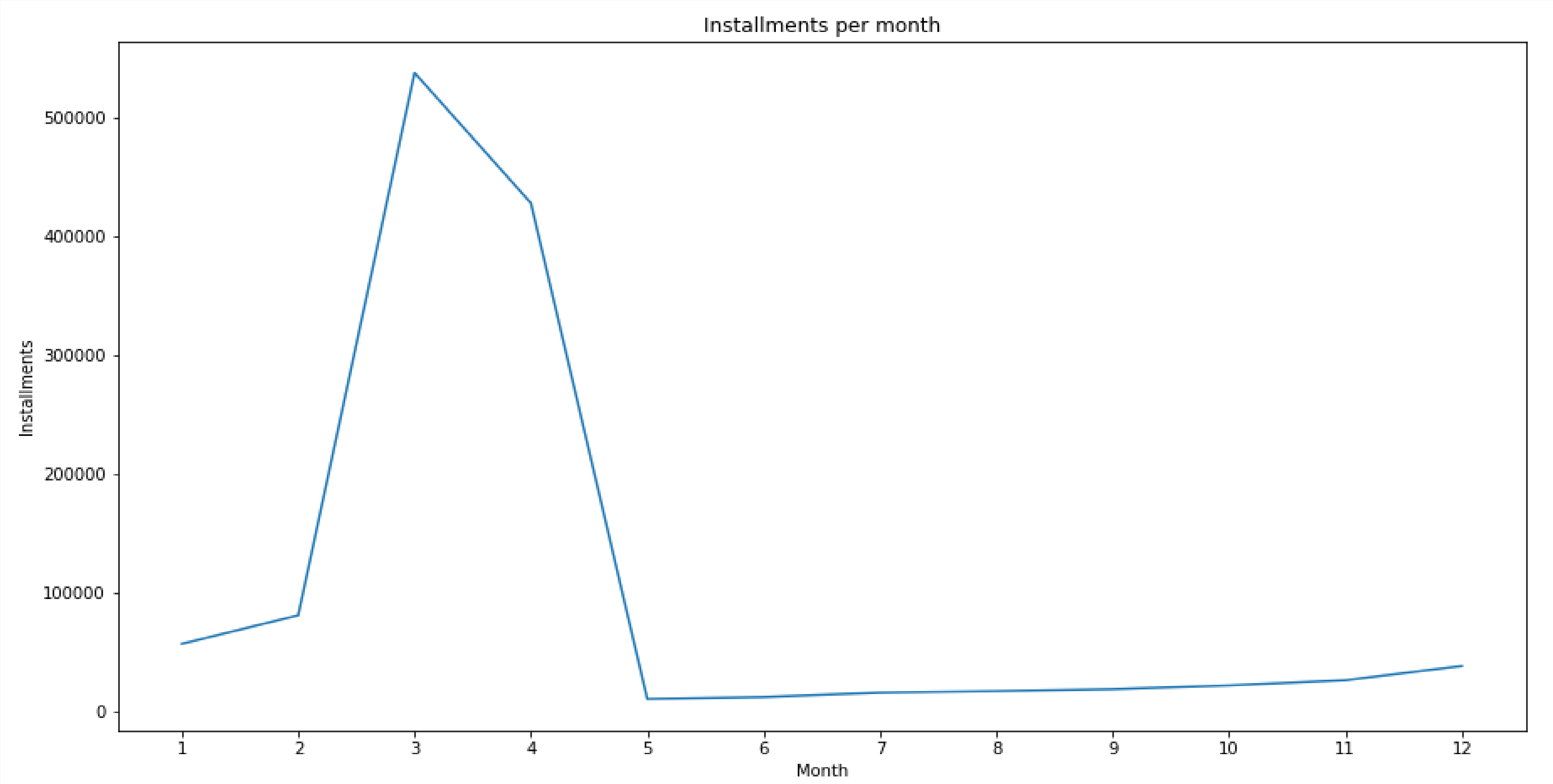


EDA – newMerchant_transactions.csv data



March has least purchase per month and there is constant purchases from May to December

EDA – newMerchant_transactions.csv data



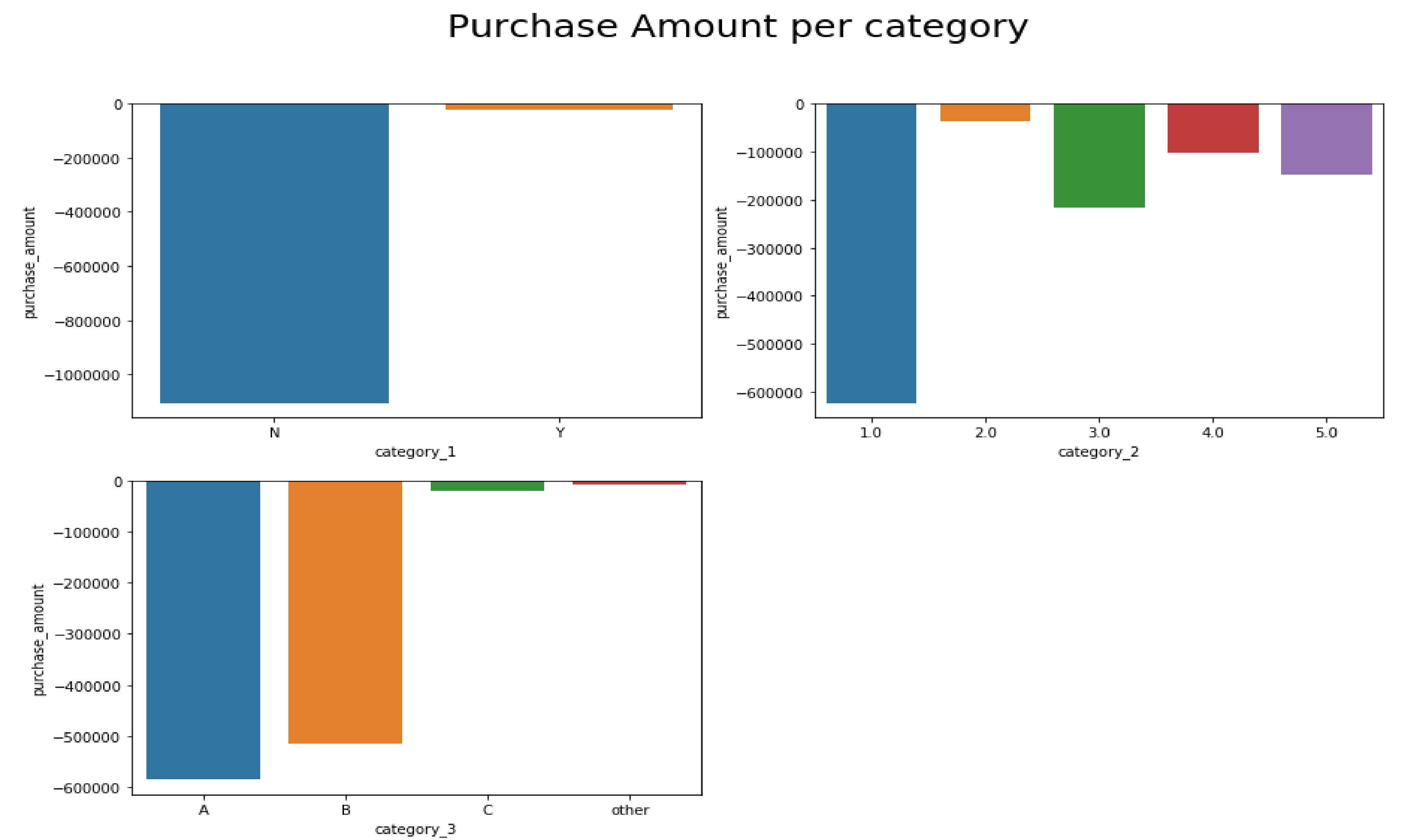
March has most installments per month.

EDA – newMerchant_transactions.csv data

Most number of purchases are not part of category 1.

Highest number of purchase in category 2 are in **1.0**.

Highest number of purchase in category 3 are in **A**.

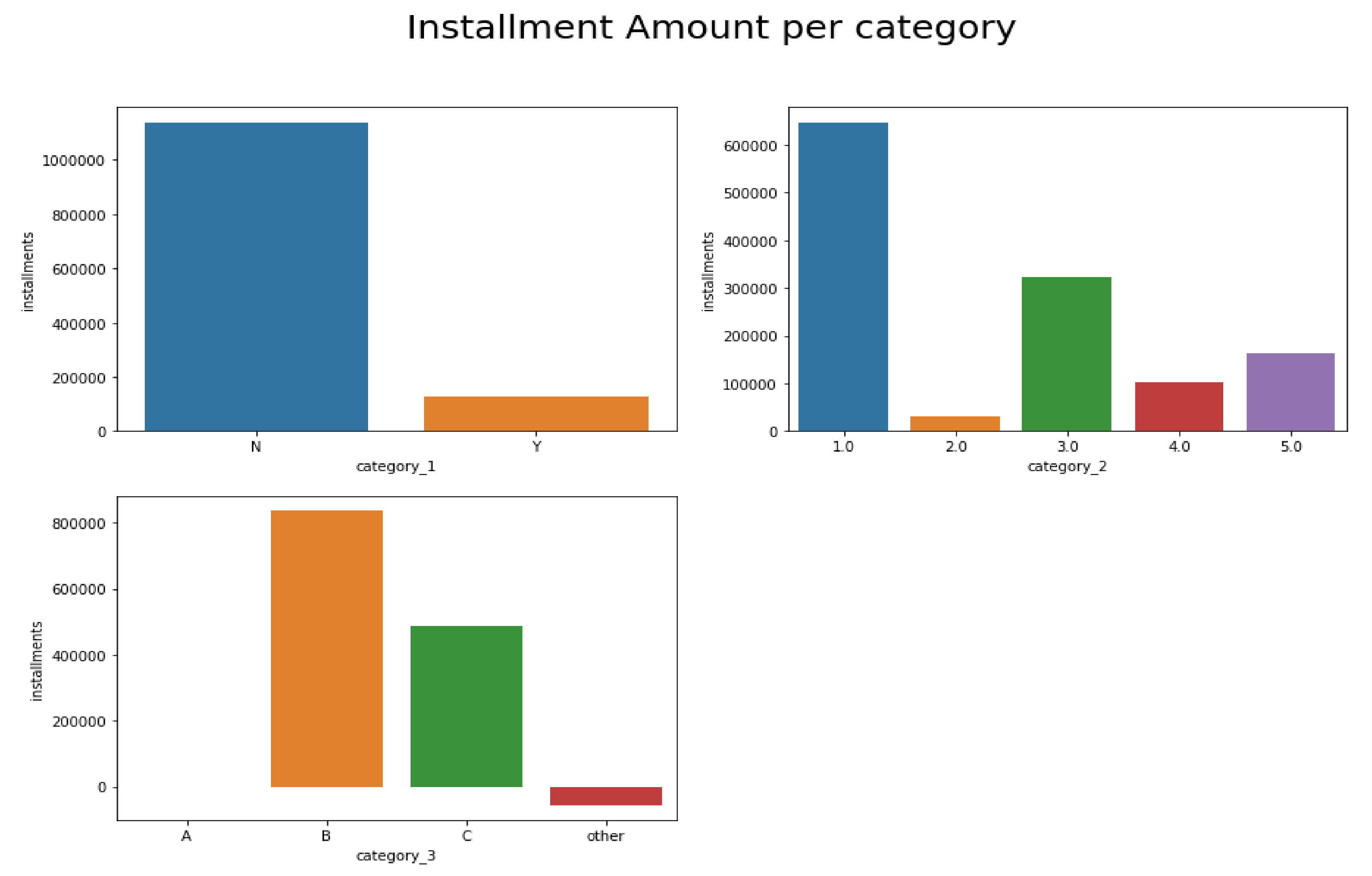


EDA – newMerchant_transactions.csv data

Most number of installments are not part of category 1.

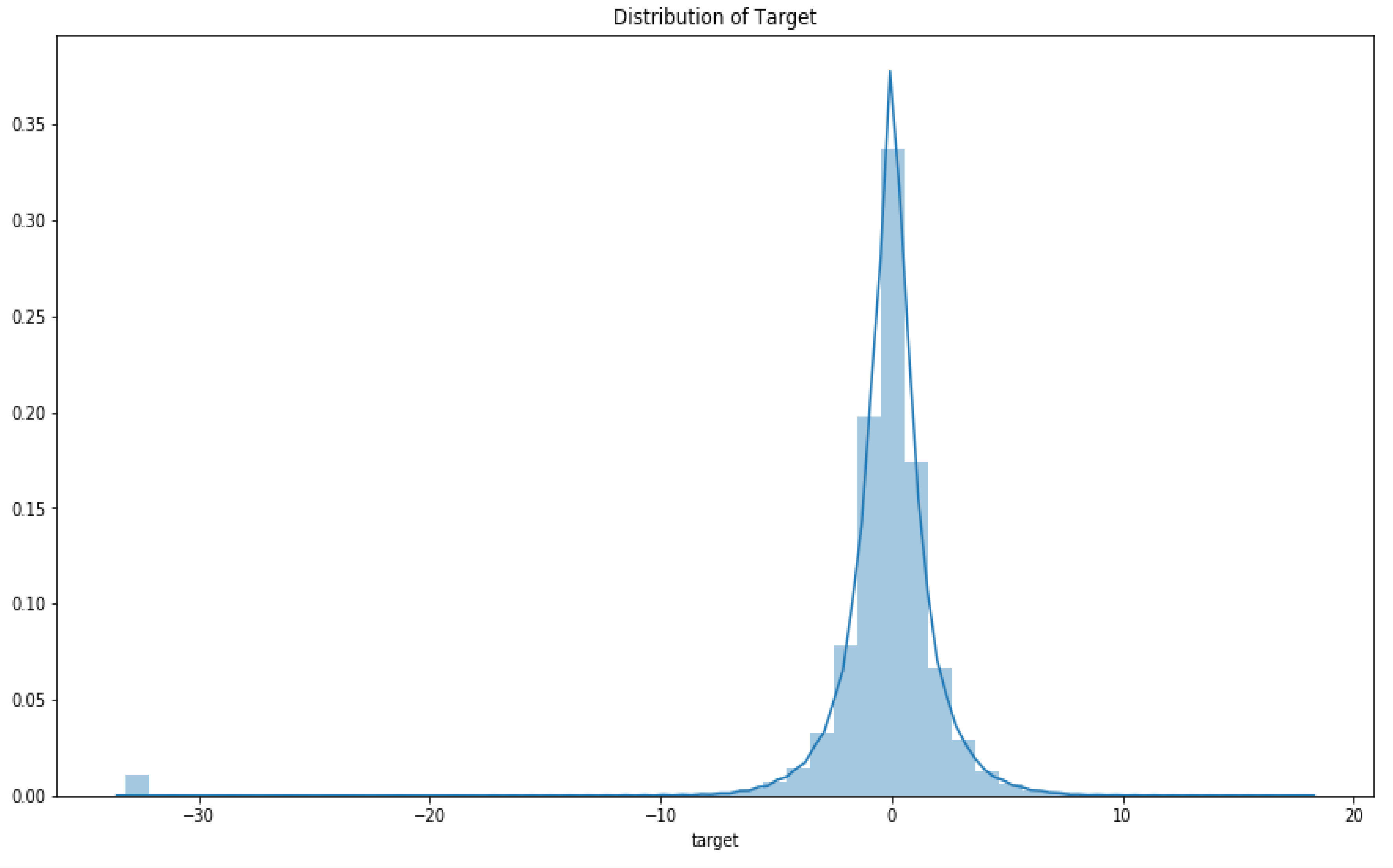
Highest number of installments in category 2 are in **1.0**.

Highest number of installments in category 3 are in **B**.



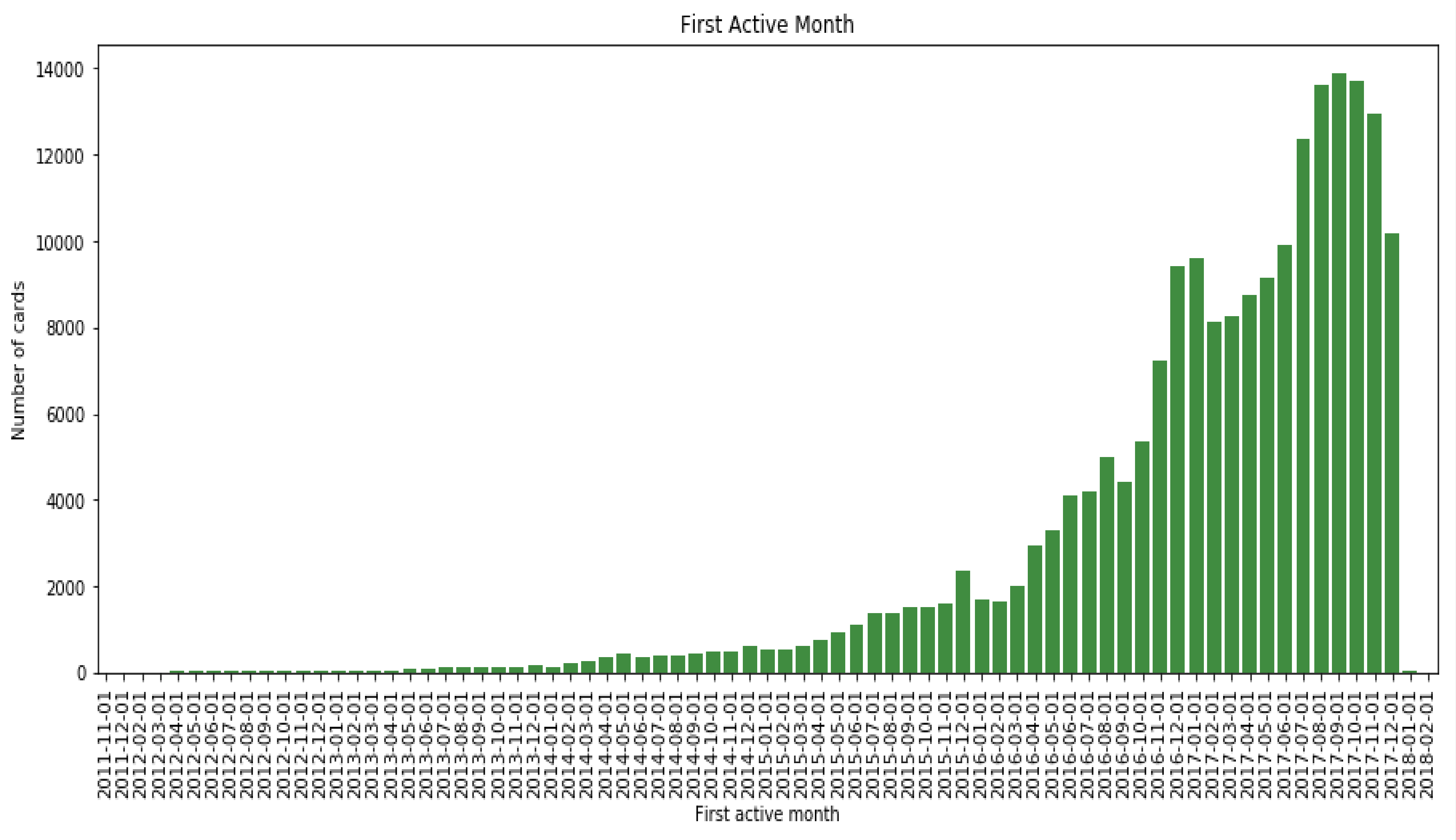
EDA – train.csv data

Target is mostly normally distributed except there is an outlier over -30 score.

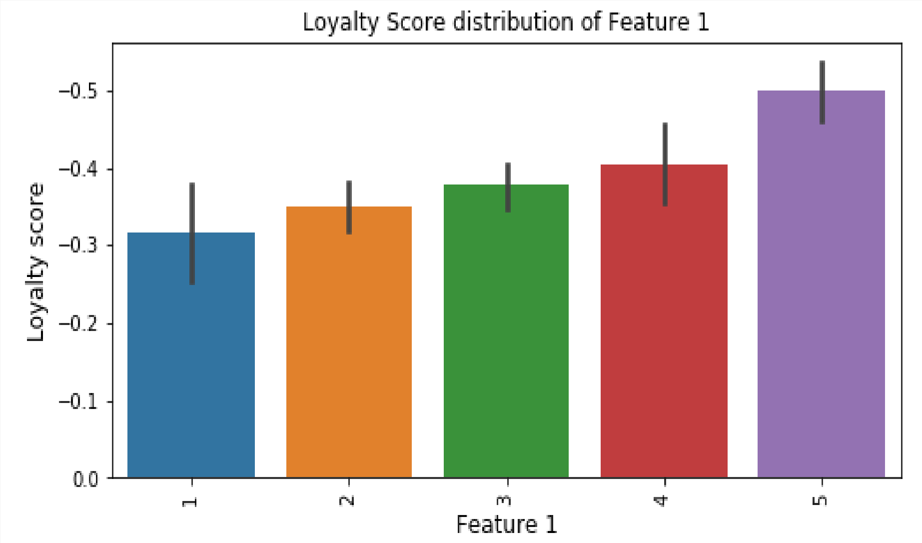


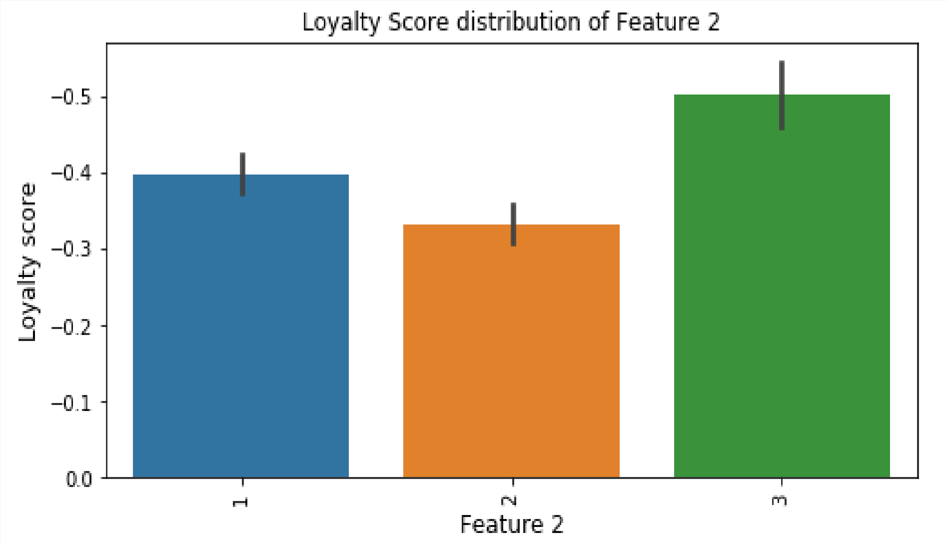
EDA – train.csv data

There is a steady increase in number of first time used cards since 2015-Jul-01.

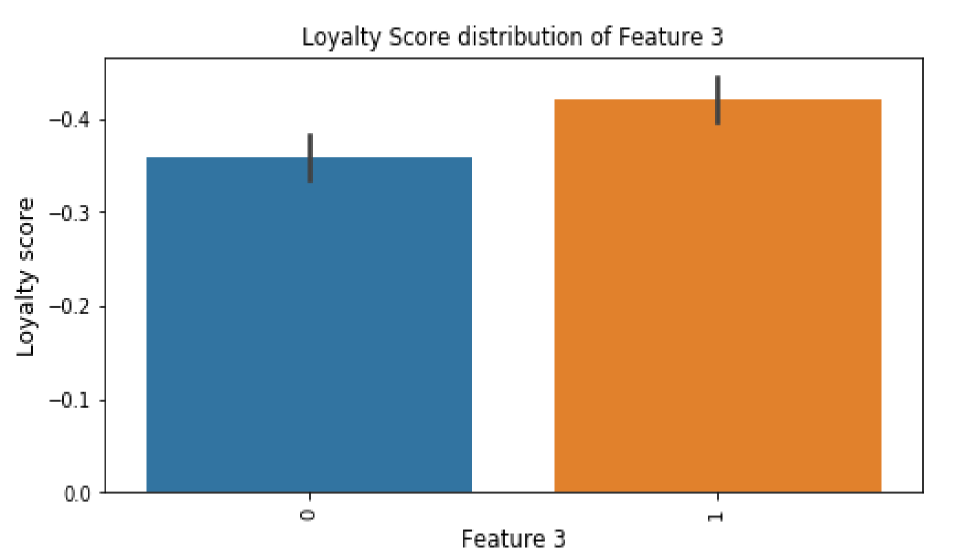


EDA – train.csv data





Loyalty score is balanced distributed across feature_1, feature_2 and feature_3.



Consulsion

Merchant transactions Data

- There is strong corelation numerical_1 and numerical_2 feature.
- There is a correlation between avg_sales and avg_purchases of 3, 6 an 12 month.
- Merchant category ID 705 has most sales with 9% sales
- City ID -1 has over 100000 transactions and amounts to 31% of transactions
- Subsector ID 27 has over 50000 transactions and amounts to 15% of transactions

- Percentage of sales in each Category
  - 98% of the transactions does not belong to category 1
  - 48 % of category 2 transactions are in 1.0
  - 71 of the transactions does not belong to category 4
- Purchase and Sales Range
  - 53% of sales and transactions are in E range
- Quantity of active months in a year
- December is most active sales month of the year

Consulsion

Historical transactions Data

- There seems to be no correlation between data
- Subsector ID 33 has over 5000000 transactions and amounts to 19% of transactions
- City ID 33 has over 4000000 transactions and amounts to 16% of transactions
- March has most purchases per month.
- January has most installments per month

- Percentage of sales in each Category
  - 92% of the transactions does not belong to category 1
  - 52 % of category 2 transactions are in 1.0
  - 53 of category 3 transactions are in A

Consulsion

New Merchant transactions Data

- There is a correlation between installments and purchase_amount.
- Subsector ID 37 has over 340053 transactions and amounts to 17% of transactions
- City ID 69 has 328916 transactions and amounts to 17% of transactions

- Percentage of sales in each Category
  - 97% of the transactions does not belong to category 1
  - 54 % of category 2 transactions are in 1.0
  - 47 of category 3 transactions are in A
- March has most installments per month.
- March has least purchase per month and there is constant purchases from May to December.