Mestrado em Data Science and Advanced Analytics, com especialização em Data Science

ABCDEats Inc.

Project of Data Mining

Jan-Louis Schneider, 20240506
Marta Boavida, 20240519
Matilde Miguel, 20240549
Sofia Gomes, 20240848

Fall/Spring Semester 2024-2025

Index:

1 Introduction

In an increasingly competitive market, food delivery companies face the challenge of delivering unique and targeted experiences to diverse customer bases.

ABCDEats Inc., a fictional food delivery service, has tasked us with analysing its customer data to uncover insights that can inform a data-driven marketing strategy. By focusing on data collected over three months from three different cities, this project aims to segment customers based on economic and behavioural perspectives.

This integrated segmentation strategy is expected to empower ABCDE to develop personalized marketing strategies that align with distinct customer segments and maximize engagement and profitability.

This project was divided into 4 phases, **exploration of the dataset**, **new features** and **relationship among features**. Throughout the report there are graphs for easier and more intuitive viewing.

2 Exploration of the dataset

After importing the dataset "DM2425_ABCDEats_DATASET.csv" and the necessary libraries, we used several functions already defined in the libraries in order to a better data understanding.

To see all the columns:

```
df.columns.values
```

To see the first ten lines:

```
df.head(10)
```

To get information about the index dtype and columns, non-null values and memory usage:

```
df.info()
```

2.1 Incorrect data

After an analysis, we noticed that some variables are incorrect (the variable type is float and should be int), namely, **customer_age**, **first_order** and **HR_0**.

So we converted the types using the function:

```
df["customer_age"]= df["customer_age"].astype('Int64')
df["first_order"] = df["first_order"].astype("Int64")
df["HR_0"] = df["HR_0"].astype("Int64")
```

Another detail about incorrect data is between product_count and vendor_count. There is information about products that were not purchased but there is information about sellers who sold them.

2.2 Duplicates

We looked for duplicate values and found 0.041%, so we removed them using the function:

```
df.drop_duplicates(inplace=True)
```

Jan-Louis Schneider, 20240506
Marta Boavida, 20240519
Matilde Miguel, 20240549
Sofia Gomes, 20240848

2.3 Missing values

We check for missing values and found some variables:

- Percentage of missing values in **customer_region** is: 1.387%
- Percentage of missing values in **customer_age** is: 2.281%
- Percentage of missing values in **first_order** is: 0.333%
- Percentage of missing values in **last_promo** is: 52.53%
- Percentage of missing values in **HR_0** is: 3.652%

2.4 Numerical variables

Numerical variables represent measurable quantities and can be analyzed mathematically. The numerical variables that exist in this dataset are:

- **customer_age**: 75% of customers are young, but there are a few older individuals
- **vendor_count**: Most customers ordered from few unique vendors, but there are customers with much higher count
- **product_count**: Most customers don't buy many products
- **is_chain**: relative small amount of orders made in chain restaurants
- **first_order**: On average customers place their first order 28 days after joining the app
- **last_order**: customers are buying more recently than they did earlier
- **CUI** (Overall analysis): 75% of customers spend nothing or a very small amount but the maximum values of these cuisines are significantly high (specially for American (280) and Asian (896)) cuisines which indicate the presence of potential outliers who frequently order or spend heavily
- **DOW_0 to DOW_6** (Overall analysis): most customers only ordered 1 time in each day of the week
- **HR_0 to HR_23** (Overall analysis): No activity at midnight (HR_0). 75% of customers placed no order in HR_1 to HR_23

2.5 Categorical variables

Categorical variables represent characteristics or qualities that group data into distinct categories or labels. In this dataset, the categorical variables are:

- **customer_region**: Customers are from 8 different regions. Most customers are located in region 8670
- **last_promo**: There are promotions in 3 different categories. Most customers use promotions in the delivery category
- **payment_method**: There are 3 different payment methods used by customers. Most customers use Card as their preferred payment method

After analysing the categorical variables, we realise that there isn't strange data.

3 New features

Creating new features can significantly enhance our analysis by providing additional insights and improving the performance of models.

So, we create some:

1. **Customer Lifetime:** Interval of customer activity, so we have an idea of how many days the customer ordered.

```
df['lifetime_days'] = df['last_order'] - df['first_order']
```

2. **Most frequent order day of the week:** Indicates the days of the week on which the customer placed the most orders.

3. **Most frequent part of the day:** 6h-12h correspond to Morning (Breakfast), 12h-18h correspond to Afternoon (Lunch), 18h-00h correspond to Evening (Dinner) and 00h-6h correspond to Night.

4. **Total monetary units spend:** Sum all total expenses.

```
cuisine = df.filter(like='CUI_').columns.tolist()
df['total_expenses'] = df[cuisine].sum(axis=1)
```

5. **Average monetary units per product:** Show the average monetary of all products.

```
df['avg_per_product'] = pd.to_numeric(df['total_expenses'] / df['
    product_count'].replace(0, pd.NA), errors='coerce')
```

6. **Average monetary units per order:** Show the average monetary per order.

```
df['avg_per_order'] = pd.to_numeric(df['total_expenses'] / df[dows].sum(
    axis=1).replace(0, pd.NA), errors='coerce')
```

7. **Average order size:** Help identifying users who make larger orders.

```
df['avg_order_size'] = pd.to_numeric(df['product_count'] / df[dows].sum(
    axis=1).replace(0, pd.NA), errors='coerce')
```

8. **Culinary profile:** A proportion of ordered cuisines. A higher number indicates more diversity of types of cuisine you ordered.

```
total_cuisine = len(cuisine)
df['culinary_variety'] = round((df[cuisine].gt(0).sum(axis=1) /
    total_cuisine), 5)
```

9. **Loyalty to chain restaurants:** Proportion of orders from restaurant chains. A high value indicates that you prefer to try different restaurant chains. A lower value is only more faithful to certain chains.

```
df['chain_preference'] = pd.to_numeric(df['is_chain'] / df[dows].sum(axis
    =1).replace(0, pd.NA), errors='coerce')
```

10. **Loyalty to venders:** Proportion of orders from specific restaurants. A high value indicates that you prefer to try different restaurants. A lower tend to be more loyal to specific restaurants.

```
df['loyalty_to_venders'] = pd.to_numeric(df['vendor_count'] / df[dows].
    sum(axis=1).replace(0, pd.NA), errors='coerce')
```

4 Visualizations and relationship among features

In order to explore the relationships among the features, we looked for trends and patterns. The most relevant outputs are in Annexes.

1. **Correlation of all numerical features:** We filter all correlations that are greater than 0.7 and smaller than 1.0 to avoid correlations with itself, and we verify that the maximum correlation is between vendor_count and culinary_variety and the minimum correlation is between vendor_count and is_chain (see Fig 1).

2. **Visualization of total Orders placed per Hour:** The most orders placed is between 10h to 12h and 16h to 18h (see Fig 2).

3. **Visualization of total Orders placed per week day:** The most orders placed on Thursday and Saturday, and least on Sunday (see Fig 3).

4. **Percentage of each payment_method for each age_group:** DIGI is balanced, CASH is mostly used by older people above 50, while these use less CARD (see Fig 4).

5. **Proportions of each last_promo for each payment_method:** DELIVERY always has the highest proportion, but for CARD the last_promo is more balanced. For CASH, the proportion of FREEBIE is lower than usual (see Fig 5).

6. **Means of each payment_method in vendor_count:** CARD highest (3.399494) and CASH lowest (2.455192), maybe because people who are more open to experimenting new restaurants also are more open to use modern or alternative payment methods.

7. **Means of each payment_method in lifetime_days:** CARD highest (40.708141) and CASH lowest (23.645542) maybe because who plans to use more frequently registers his card on website/app and first time users only use CASH.

8. **Means of each payment_method in total_expenses:** Total expenses highest with CARD (42.655100), than DIGI (32.906232) and CASH (28.518848). Explanation could be that with CASH you have better feeling for how much you spent.

9. **Means of each payment_method in avg_per_product:** The difference between the average per product of CARD (7.366293), CASH (7.534105) and DIGI (8.296765) is not very big, being DIGI the highest average.

10. **Means of each payment_method in avg_per_order:** The average monetary per order is greater if paying by DIGI (11.620458).

11. **Means of each payment_method in avg_order_size:** Who made majors order payed with DIGI (1.332555), but the difference is minimal (CARD: 1.296034 and CASH: 1.272643).

12. **Means of each payment_method in culinary_variety:** CARD highest (0.162484), CASH lowest (0.127230). Similar to vendor_count, people with CASH are more conservative people, trying less new methods (”new” payment methods).

13. **Means of each payment_method in chain_preference:** In chain preference, the customers prefer paying in CASH (0.654581), but the difference is minimal (CARD: 0.619821 and DIGI: 0.624892).

14. **Total expenses per age group:** Lower at younger age, after 23-28 more regular, probably because young people have less money (see Fig 6).

15. **Culinary variety per age group:** No big differences but peak at 23-28, maybe because people start to live on their own and try more different things (see Fig 7).

16. **Means of each last_promo in lifetime_days:** FREEBIE highest(36.043528), DELIVERY lowest(22.651561).

17. **Means of each last_promo in total_expenses:** FREEBIE highest (36.043528), DELIVERY lowest(22.651561) but more equal to DISCOUNT (30.976978). FREEBIE leads to more expenses in total.

18. **Means of each last_promo in avg_per_product:** In the last_promo, the average per product is greater on DELIVERY (8.186845), but the difference is minimal (DISCOUNT: 7.232929 and FREEBIE: 7.771700).

19. **Means of each last_promo in avg_per_order:** In the last_promo, the average per order is greater on DELIVERY (11.361651), but the difference is minimal (DISCOUNT: 9.532429 and FREEBIE: 10.583562).

20. **Means of each last_promo in chain_preferences:** DISCOUNT highest (0.650907), people with DISCOUNT promo tend to go more to chains.

21. **Means of each last_promo in loyalty_to_venders:** DELIVERY (0.892207) leads to more loyalty.

22. **Relations between the costumer age and some types of cuisine:** We do with CUI_Asian, CUI_Desserts and CUI_Healthy per age, because they are the most relevant. Asian increase by age, Dessert decrease with age and Healthy peak on 23-28 (see Fig 8).

23. **Means of each total_expenses in chain_preferences:** Big spenders (total expenses > 45) tend to go to less different chains. (see Fig 9).

24. **Proportions of each last_promo value for the two groups of people:** Big spenders use Freebie a lot more, while low spenders use delivery the most (see Fig 10).

25. **Cuisines and total_expenses:** The biggest spenders have their highest increase of spending compared to the other costumers in StreetFood/Snacks, Cafe, Asian, Healthy and Desserts, while the lowest increase is in Chicken Dishes, Noodles Dishes and Indian (see Fig 11).

26. **Cuisines and loyalty_to_venders:** The costumers with the highest loyalty prefer to go to Italian Cuisine, followed by Chinese and Cafe. The lowest are Desserts and Snacks where Costumers tend to choose any place without too much thought or favourite places (see Fig **??**).

27. **Customer_regions:** The region 8550 spends more than the others (see Fig 13).

A Annexes

Figure 1: Highest Correlations of all numerical features

```
vendor_count      product_count      0.827602
                  is_chain           0.762893
                  culinary_variety   0.869244
product_count     is_chain           0.827070
                  total_expenses     0.824801
avg_per_product   avg_per_order      0.813709
avg_per_order     avg_order_size     0.725985
dtype: float64
```

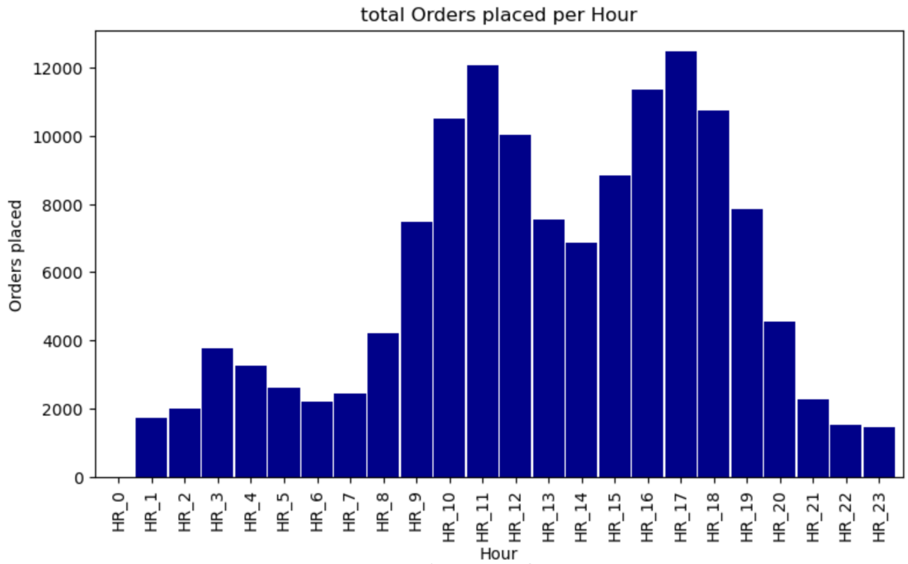

Figure 2: Visualization of total Orders placed per Hour

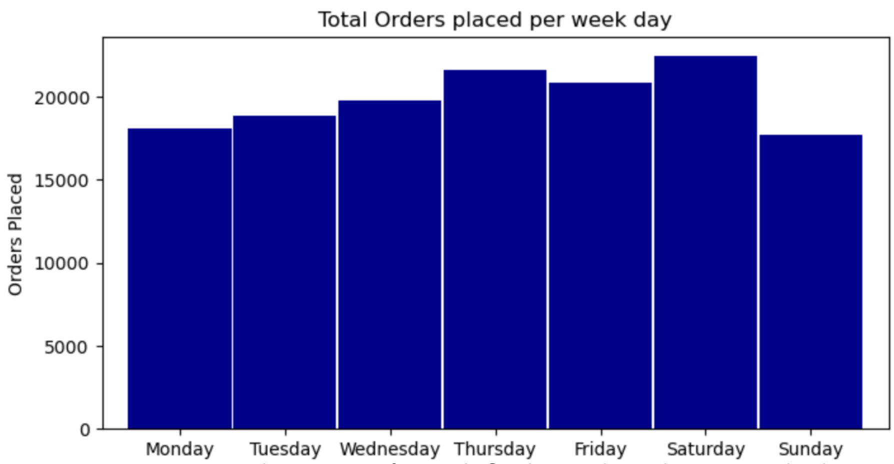

Figure 3: Visualization of total Orders placed per week day

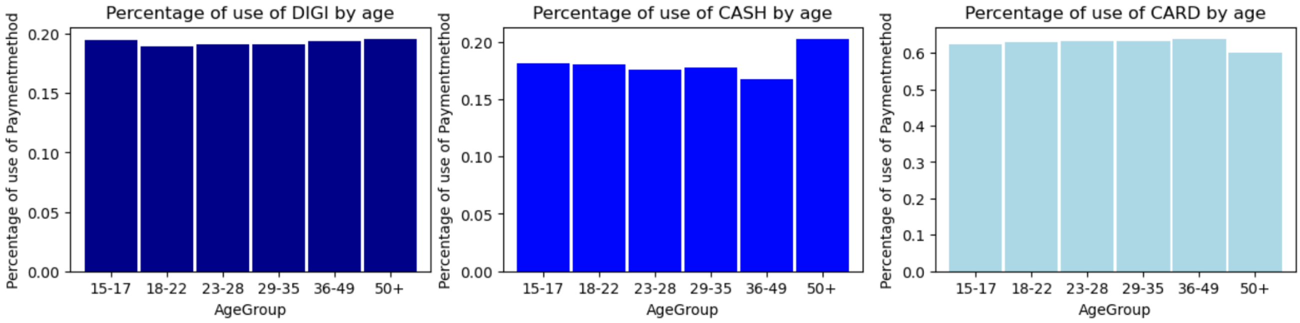

Figure 4: Percentage of each payment_method for each age_group

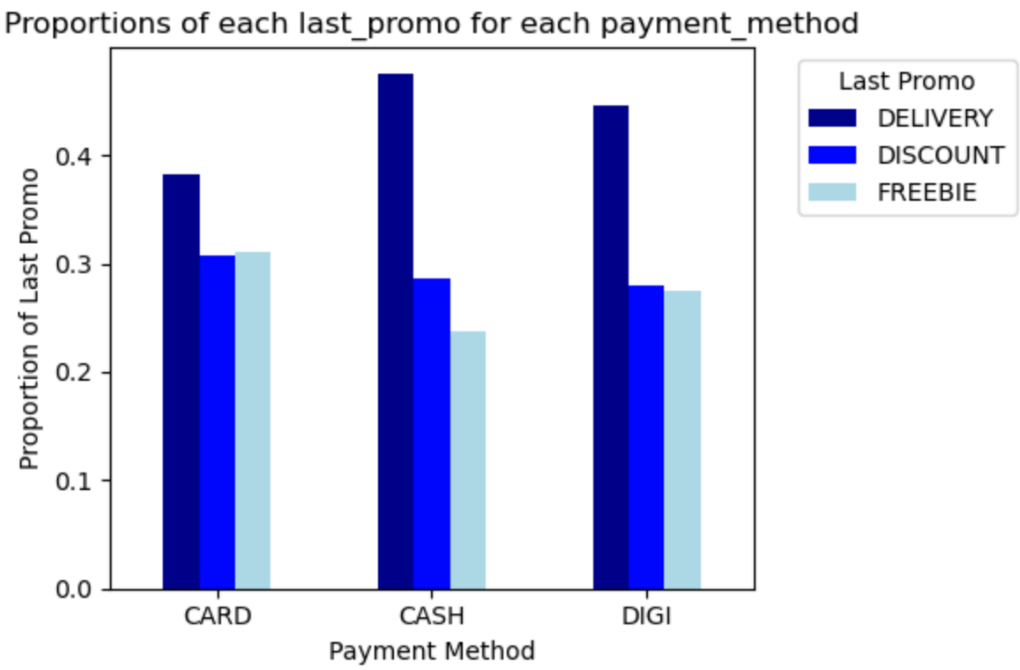

Figure 5: Proportions of each last_promo for each payment_method

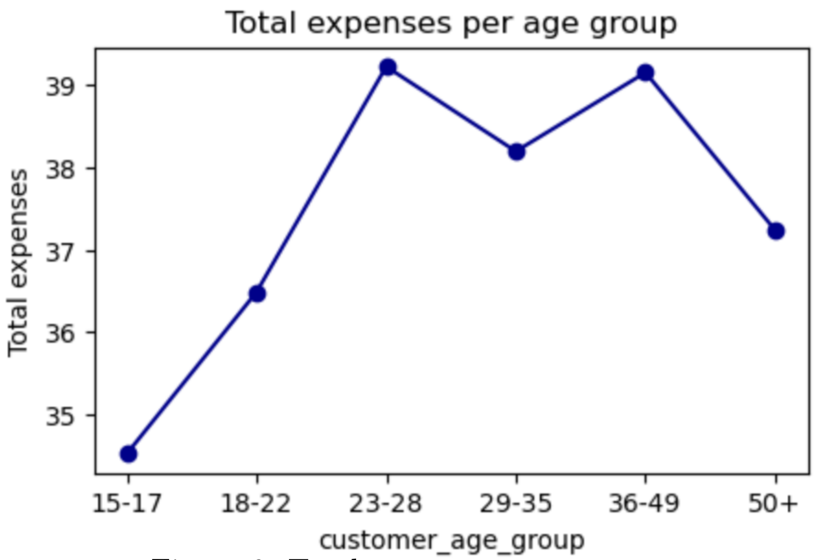

Figure 6: Total expenses per age group

Jan-Louis Schneider, 20240506
Marta Boavida, 20240519
Matilde Miguel, 20240549
Sofia Gomes, 20240848

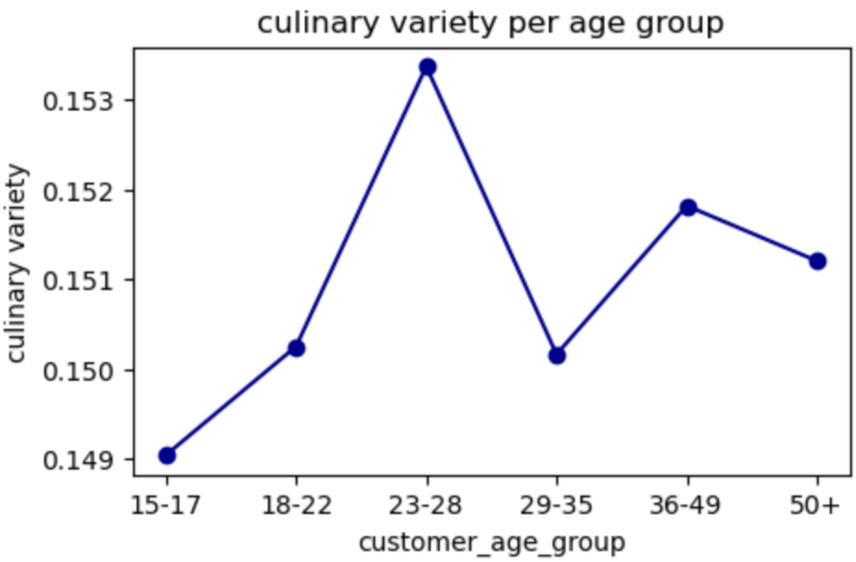

Figure 7: Culinary variety per age group

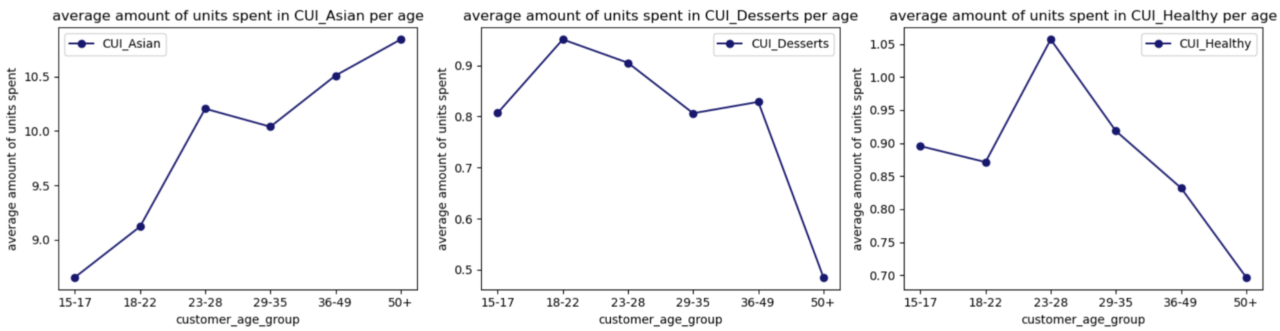

Figure 8: Relations between the costumer age and some types of cuisine

Figure 9: Means of total_expenses > 45 as True and < 45 as False in chain_preferences

```
total_expenses
False    0.651449
True     0.553577
Name: chain_preference, dtype: object
```

Figure 10: Proportions of each last_promo value for the two groups of people (True for customer with total_expenses > 45)

```
total_expenses  last_promo
False           DELIVERY      0.436434
                DISCOUNT      0.299038
                FREEBIE       0.264528
True            FREEBIE       0.391588
                DELIVERY      0.319797
                DISCOUNT      0.288615
Name: proportion, dtype: float64
```

Figure 11: Percent Difference expenses for each cuisine between highest spenders and other customers

	Column	Percent Difference
13	CUI_Street Food / Snacks	796.320340
3	CUI_Cafe	752.400594
1	CUI_Asian	539.315645
7	CUI_Healthy	390.942451
6	CUI_Desserts	358.410243
10	CUI_Japanese	354.910572
5	CUI_Chinese	339.652114
0	CUI_American	297.970345
12	CUI_OTHER	292.006300
2	CUI_Beverages	275.757716
9	CUI_Italian	268.204610
14	CUI_Thai	256.833114
8	CUI_Indian	234.749068
11	CUI_Noodle Dishes	157.181095
4	CUI_Chicken Dishes	84.202205

Figure 12: Percent Difference expenses for each cuisine between most loyal and other customers

	Column	Percent Difference
9	CUI_Italian	1185.583826
5	CUI_Chinese	425.792680
3	CUI_Cafe	367.177231
4	CUI_Chicken Dishes	280.490457
7	CUI_Healthy	225.828735
0	CUI_American	201.494039
8	CUI_Indian	187.469433
11	CUI_Noodle Dishes	169.202893
2	CUI_Beverages	112.540042
12	CUI_OTHER	98.020422
14	CUI_Thai	77.716246
10	CUI_Japanese	56.526945
1	CUI_Asian	47.356558
13	CUI_Street Food / Snacks	-60.773845
6	CUI_Desserts	-95.652146

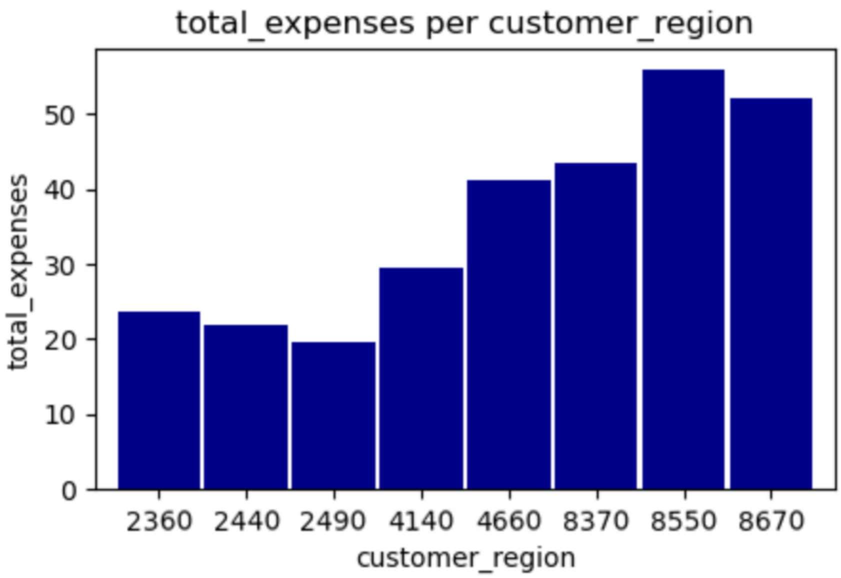

Figure 13: Customer_regions