PERSONAL BANKING STRATEGIC ANALYTICS

Candidate Package

Overview

You will have received 3 datasets containing account, customer, and transaction level data. We are interested in understanding the process you take in answering the following questions. Please provide the answer and an explanation where applicable. Questions are broken into two categories:

- Data Manipulation Queries
  - These are questions designed to test your understanding of data manipulation and preparation. Please provide the answer to the question on the Data Manipulation answer slide (slide 5) and provide your queries in the corresponding Data Manipulation queries slides. You can use any programming language you would like to complete the task.

- Business Case
  - These questions are designed to test your critical thinking, logic, and problem solving skills. Please document your approach and present your findings as if you were presenting to an executive (non-technical) business partner. You can use any programming language you would like to complete the task.

NOTE: Please complete your work and include your findings in this deck and send back the completed deck saved as CandidatePackage_TD_LASTNAME_FIRSTNAME. Also submit the code used to complete your tasks (e.g. Jupyter notebook, R code, etc.).

Your Task

Data Manipulation

1. What branch has the most number of customers?
2. How old is the oldest customer as of 2019-07-01
3. How many accounts does the oldest customer have?
4. How many transactions went to Starbucks in April?
5. How much was spent on Starbucks in April?
6. Hypothesis Testing: Is the average spend at Starbucks (statistically) significantly different in April compared to June?
7. Which date exhibited the highest average spend above trend at Starbucks (based on a 10-period moving average, ignoring missing dates)?

Internal

❶ Data Manipulation Questions

Data Manipulation Query Answer Summary Page

1. What branch has the most number of customers?
   - **ANSWER: (XXX)**
2. How old is the oldest customer as of 2019-07-01?
   - **ANSWER: (XXX)**
3. How many accounts does the oldest customer have?
   - **ANSWER: (XXX)**
4. How many transactions went to Starbucks in April?
   - **ANSWER: (XXX)**
5. How much was spent on Starbucks in April?
   - **ANSWER: (XXX)**
6. Hypothesis Testing: Is the average spend at Starbucks (statistically) significantly different in April compared to June?
   - **ANSWER: (XXX)**
7. Which date exhibited the highest average spend above trend at Starbucks (based on a 10-period moving average, ignoring missing dates)?
   - **ANSWER: (XXX)**

Data Manipulation Query 1: Branch with most customers

Answer :

```
select count(distinct(id)) as number_of_customer ,
       branchNumber
from account
group by branchNumber
order by number_of_customer desc
limit 1;
```

Data Manipulation Query 2: Oldest customer as of 2019-07-01

Answer :

```
select
    id,
    timestampdiff(year, birthDate, '2019-07-01') as Age
from customer
group by id
-- having  Age <= 100
order by Age desc
limit 1;
```

Data Manipulation Query 3: Oldest customer number of accounts

Answer :

```
select
    customer.id,
    timestampdiff(year, customer.birthDate, '2019-07-01') as
Age,
    count(account.id) as number_of_accounts
from customer
left join account on customer.id = account.customer_id
group by id, Age
order by Age desc
limit 1;
```

Data Manipulation Query 4: Transactions to Starbucks

Answer:

```
select
    extract( month from originationDateTime) as
Required_month,
    count(*) as number_of_transactions
from transaction
where description like '%STARBUCKS%'
    AND Month(originationDateTime)= 4
group by Required_month
;
```

Data Manipulation Query 5: Dollars spent at Starbucks

Answer:

```sql
select
    description,
    extract(month from originationDateTime) as month_required,
    round(sum(currencyAmount), 2)  as total_spent
from transaction
    where MONTH(originationDateTime)=4
    and description like '%STARBUCKS%'
group by month_required;
```

Data Manipulation Query 6: Hypothesis Test - Starbucks April and June statistical significance spend comparison

Answer:

```
SELECT
    description,
    EXTRACT(MONTH FROM originationDateTime) as required_month
,
    round(avg(currencyAmount),2)  as average_spending
from transaction
where month(originationDateTime) in (4, 6)
and description like '%STARBUCKS%'
GROUP BY required_month;
```

Data Manipulation Query 7: Highest average spend above trend at Starbucks (based on a 10-period moving average, ignoring missing dates)

```
WITH starbucks_tx AS (
  SELECT
    originationDateTime,
    currencyAmount,
    ROW_NUMBER() OVER (ORDER BY originationDateTime) AS rn
  FROM td_interview.transaction
  WHERE description LIKE '%STARBUCKS%'
),

moving_avg_calc AS (
  SELECT
    tx1.originationDateTime,
    tx1.currencyAmount,
    ROUND(AVG(tx2.currencyAmount), 2) AS moving_avg,
    ROUND(tx1.currencyAmount - AVG(tx2.currencyAmount), 2) AS above_trend
  FROM starbucks_tx tx1
  JOIN starbucks_tx tx2
    ON tx2.rn BETWEEN tx1.rn - 10 AND tx1.rn - 1
  WHERE tx1.rn > 10  -- 👈 Filter out rows that can't have 10 previous values
  GROUP BY tx1.originationDateTime, tx1.currencyAmount, tx1.rn
)

SELECT *
FROM moving_avg_calc
ORDER BY above_trend DESC
LIMIT 1;
```

❷ Business Case

Your Task

Business Case

Use the data provided to answer the following questions:

1. We are planning to launch a new product focused on a specific merchant category (e.g. travel credit card). Which specific merchant category would you like to focus on for this new product? Please explain your rationale for this category incorporating both the insights derived from the data and other concepts where you see fit.

2. Identify and describe various segments of customers within the data. Consider applying segmenting/clustering techniques to aid in the development of your answer.

3. Of the segments that you created in question 2, which specific segment would you like to target for this new product? Why would you target them? What are the potential challenges/risks to consider when targeting this segment vs. others?

Internal

Approach

Approach 1: **Pareto Analysis** (80/20)

Purpose : find the 80% indicators that have the biggest influences to the task. Which in this case, we will identify which category contributes the most spending

What I did:

I summed all spending for each categories, and calculated cumulative percentage of spending of each category, ordered from the most to the least.

In this way, it helps me to focus on the top spending categories.

Approach 2, **Customer segmentation**

Purpose: Understanding which customer group is targeted marketing group. Making plans accordingly.

What I did:

I used "income" and "saving" as two dimensions, creating segments :

< 40k income
40k–80k income
80k+ income;

< 5k savings
5k–25k savings
25k+ savings

This allowed me to compare **spending behavior by segment** and identify groups with the highest potential for the credit product.

❸ Executive Presentation (Business Case Answers)

Executive Summary

I have analyzed transaction data across the categories to identify optimal targeting opportunities of the potential credit product, which in this case, will be a VISA credit card.

Pareto analysis revealed that 'travel' takes over about 45% of the total spending of customers. Following with shopping, food and dining, which makes up over 90% of the total purchase from customer.

In customer segmentation analysis, it shows that customer with higher income and savings spend more money than customers whose income is lower than 40,000 per year. The difference between spending on traveling and shopping does not significantly show liner relevance. Which means, the potential VISA credit card's targeting group should for customers with income above 40,000 per year.

Overall, targeting the categories of top spendings can lead to more usage of credit card, while maintain customer retention, and improve transactions related profits and strenghening relationships with partnered companies.

Executive Summary

Key findings:

- 1, over 90% of spending is in 'travel', 'shopping', and 'food and dining'.
- 2, traveling and shopping purchases are mainly contributed by people with income level over 40,000 a year.
- 3, there is no significant difference between the group with 40,000 to 80,000 income and the group of having 80,000 and above income.
- 4, Mid income customers also show strong travel and shopping behaviour.

Presentation of Findings (Question 1)

We are planning to launch a new product focused on a specific merchant category (e.g. travel credit card). Which specific merchant category would you like to focus on for this new product? Please explain your rationale for this category incorporating both the insights derived from the data and other concepts where you see fit.

Merchant category: travel and shopping.

According to transcation data:

- Travel alone accounts over 45% of total customer spending. Makes it the highest impact category.
- When combined with shopping, these two categories accounts over 70% of the total discre5tionary spending.

In addition of the high volume, these two categories often goes hand to hand. Many customers shop while traveling – such as buying souvenirs, gifts, and clothes. By providing a VISA credit card that earns rewards for both traveling and shopping would encourage customers to uise the card during trips.

It also increase the value for customers during their multiple stages of usage -- from hotel booking, flight tickets purchasing, to spending during traveling. This would increase usage, satisfaction, and gain potential revenues through transaction fees, and partner rewards.

Presentation of Findings (Question 2)

Identify and describe various segments of customers within the data. Consider applying segmenting/clustering techniques to aid in the development of your answer.

Customer Segments :

- Income : "<40k", "40k+", "80k+"
- Savings: "<5k", "5k+", "25k+"

Key insights:

-Income "80k+" and "<40k" tend to spend the most on "shopping" -- total spending is twice as much as people with income "40K+".

-Customers in the 40k–80k income range are the top spenders in Travel, contributing the highest total in that category.

-Average spending for each income level groups do not have a significant difference. Total spending at different volumes base on different group sizes.

Presentation of Findings (Question 3)

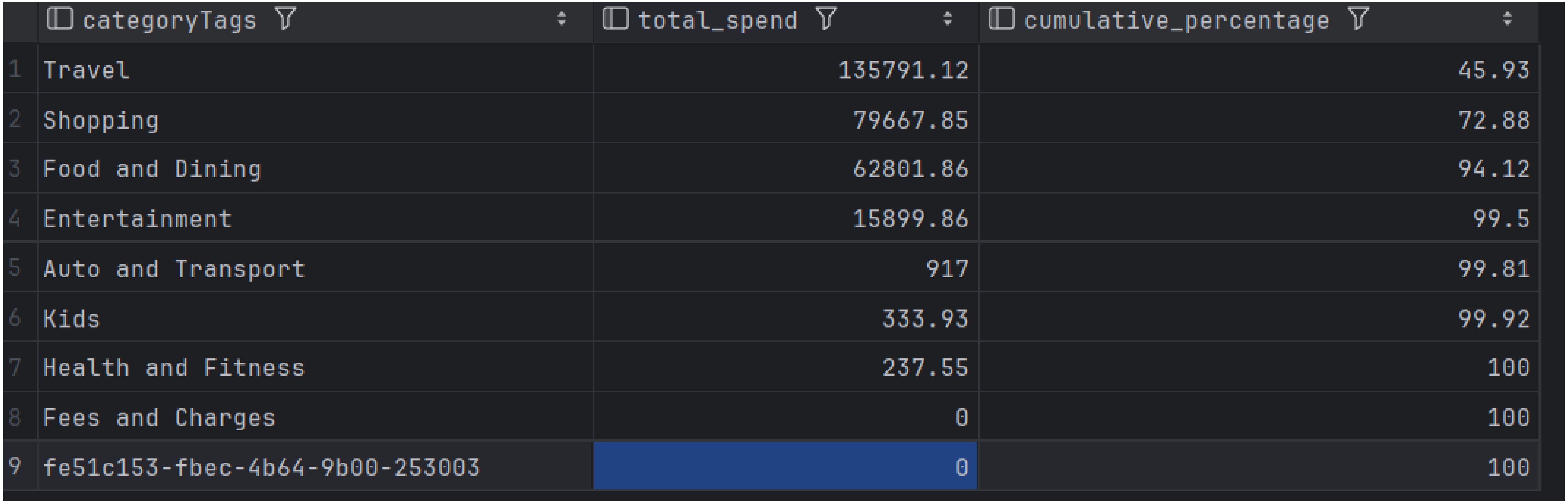
Of the segments that you created in question 2, which specific segment would you like to target for this new product? Why would you target them? What are the potential challenges/risks to consider when targeting this segment vs. others?

- I would target customers earning 40k–80k with a travel-focused credit card, as they are the highest spenders in this category.
- For shopping-related rewards, I'd target customers earning <40k, as they have strong purchase activity despite limited income — making them more likely to value discounts and points.

I think these two groups will be more interested in points rewards and purchases discounts since they are . Every dollar consicous. They are the groups that would be most attracted to points collecting and discounts offering.

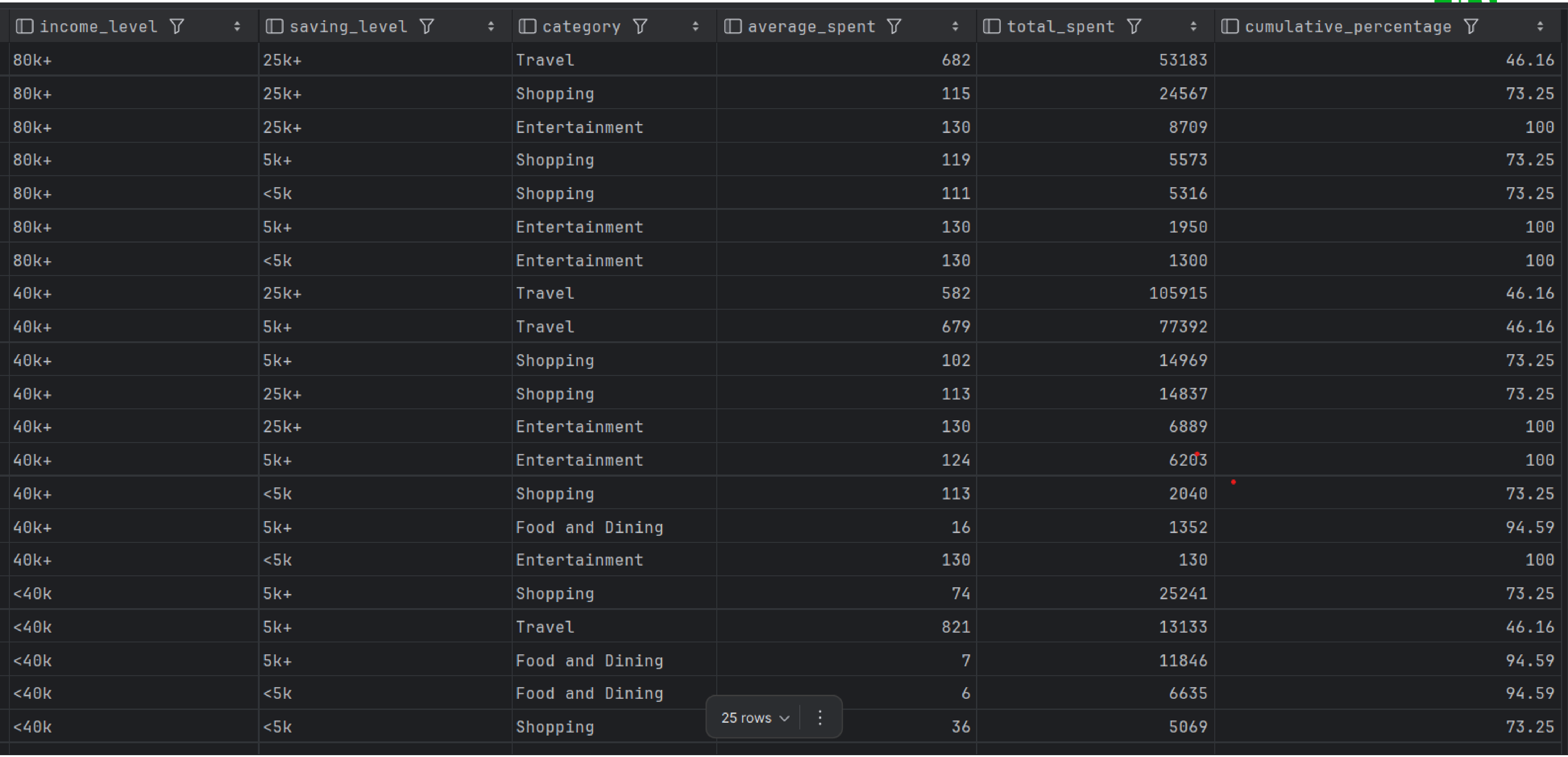
The potential risks would include : economy downturns and inflation – while medium level income customers have more stresses paying daily bills, taxes, and utilities. During tough financial periods, spending on travel and non-essentials may drop, affecting card usage.

Reference



	categoryTags	total_spend	cumulative_percentage
1	Travel	135791.12	45.93
2	Shopping	79667.85	72.88
3	Food and Dining	62801.86	94.12
4	Entertainment	15899.86	99.5
5	Auto and Transport	917	99.81
6	Kids	333.93	99.92
7	Health and Fitness	237.55	100
8	Fees and Charges	0	100
9	fe51c153-fbec-4b64-9b00-253003	0	100

Reference



income_level	saving_level	category	average_spent	total_spent	cumulative_percentage
80k+	25k+	Travel	682	53183	46.16
80k+	25k+	Shopping	115	24567	73.25
80k+	25k+	Entertainment	130	8709	100
80k+	5k+	Shopping	119	5573	73.25
80k+	<5k	Shopping	111	5316	73.25
80k+	5k+	Entertainment	130	1950	100
80k+	<5k	Entertainment	130	1300	100
40k+	25k+	Travel	582	105915	46.16
40k+	5k+	Travel	679	77392	46.16
40k+	5k+	Shopping	102	14969	73.25
40k+	25k+	Shopping	113	14837	73.25
40k+	25k+	Entertainment	130	6889	100
40k+	5k+	Entertainment	124	6203	100
40k+	<5k	Shopping	113	2040	73.25
40k+	5k+	Food and Dining	16	1352	94.59
40k+	<5k	Entertainment	130	130	100
<40k	5k+	Shopping	74	25241	73.25
<40k	5k+	Travel	821	13133	46.16
<40k	5k+	Food and Dining	7	11846	94.59
<40k	<5k	Food and Dining	6	6635	94.59
<40k	<5k	Shopping	36	5069	73.25

Reference

Pareto analysis code :

```
select
    categoryTags,
    total_spend,
    ROUND(sum(total_spend) OVER (ORDER BY total_spend DESC) /
SUM(total_spend) OVER() * 100, 2)  as cumulative_percentage
FROM (
select
    sum(currencyAmount) as total_spend ,
    categoryTags
from transaction
WHERE categoryTags NOT IN ( 'Income', 'Transfer', 'Taxes', 'Mortgage and
Rent', 'Bills and Utilities', 'Home', '')
and categoryTags is not null
group by categoryTags ) AS temp_1
order by total_spend DESC ;
```

Pareto analysis + customer segmentation code :

```sql
select
    MAX(total_Income) as Max_Income ,
    MIN(total_Income ) as Min_Income
from customer ;

with table_1 as (
    select
        trim(transaction.categoryTags) as Category,
        account.balance as Savings,
        customer.total_Income as Income,
        transaction.currencyAmount as Spent
    from transaction
    left join account on account.customer_id = transaction.customerId
    left join customer on account.customer_id = customer.id
    where trim(transaction.categoryTags) in ('Food and Dining', 'Entertainment', 'Shopping', 'Travel')
),
    table_2 as (
        select
            Savings,
            Spent,
            Category,
            case
                WHEN Income IS NULL OR Income = 0 THEN 'Unknown'
                WHEN Income >= 80000 THEN '80k+'
                WHEN Income >= 40000 THEN '40k+'
                ELSE '<40k'
            end as Income_Level
        from table_1
    )
select
    Income_Level,
    Category,
    ROUND(AVG(Spent), 0) AS Average_Spent,
    round( SUM(Spent), 0) as Total_Spent
from table_2
group by Income_level, Category
order by Income_Level desc, Total_Spent DESC;
```