Credit Card Risk Analysis

SQL Data Exploration

Marcus Fernandes

Email LinkedIn GitHub

Introduction

Previously, we validated and profiled the anonymized data related to credit card holders and their transactions. Now, we will use **SQL** to explore it more and try to answer the key business questions:

1. Who are our riskiest customers?
2. Are there early signs of credit distress?
3. How do risk patterns differ by region or demographic?
4. Where can we reduce exposure without hurting loyal customers?

1. Who are our riskiest customers?

The objective is to find out what profiles are associated with repeated late payments or high utilization.

	customer_id integer	card_id integer	num_late_payments bigint	num_snapshots bigint	late_rate numeric
1	245	285	18	78	0.23
2	613	719	16	72	0.22
3	15	16	15	64	0.23
4	344	401	15	77	0.19
5	428	505	15	80	0.19
6	511	604	15	77	0.19
7	259	301	14	70	0.20
8	553	652	14	75	0.19
9	178	209	14	82	0.17
10	182	214	12	60	0.20

Late payment rate by client and card tier, ordered by the number of late payments.

	customer_id integer	card_id integer	num_late_payments bigint	num_snapshots bigint	late_rate numeric
1	392	460	3	4	0.75
2	89	101	3	6	0.50
3	623	729	2	4	0.50
4	41	42	3	7	0.43
5	139	162	3	7	0.43
6	287	337	2	5	0.40
7	204	239	2	5	0.40
8	209	245	2	5	0.40
9	368	433	2	5	0.40
10	244	284	9	24	0.38

Late payment rate by client and card tier, ordered by the percentage of late payments.

Both tables above identify the top 10 riskiest customers in terms of repeated late payments. The first one, ordered by the number of late payments, tells us who frequently pays late, while the second, ordered by the percentage of late payments, identifies the cards (issued for at least 4 months) with a continuous pattern of risk.

	customer_id integer	credit_limit_bin text	credit_limit integer	card_id integer	avg_utilization numeric
1	597	Upper-Mid (6K-10K)	7708	700	0.94
2	568	Very High (20K+)	48380	667	0.92
3	200	Upper-Mid (6K-10K)	9978	235	0.92
4	398	Very High (20K+)	40857	469	0.91
5	33	High (10K-20K)	18154	34	0.91
6	495	High (10K-20K)	17285	585	0.91
7	535	Upper-Mid (6K-10K)	8919	630	0.91
8	195	Upper-Mid (6K-10K)	7296	228	0.91
9	270	Medium (3K-6K)	4701	315	0.91
10	283	Medium (3K-6K)	4675	331	0.91

The table above provides the top 10 average utilization rates by customer. We can note that many customers consistently use around 90% of their credit limit monthly, which can be either a sign of financial distress that can lead to late payments, or an inappropriate credit limit. We can combine both metrics to improve the identification of the riskiest customers.

	customer_id integer	card_id integer	avg_utilization numeric	num_late_payments bigint	num_snapshots bigint	late_rate numeric
1	392	460	0.37	3	4	0.75
2	89	101	0.43	3	6	0.50
3	623	729	0.42	2	4	0.50
4	139	162	0.41	3	7	0.43
5	41	42	0.39	3	7	0.43
6	209	245	0.89	2	5	0.40
7	204	239	0.41	2	5	0.40
8	368	433	0.39	2	5	0.40
9	287	337	0.39	2	5	0.40
10	244	284	0.40	9	24	0.38
11	14	14	0.37	3	8	0.38
12	473	559	0.86	2	6	0.33
13	96	109	0.41	10	30	0.33
14	283	331	0.91	5	16	0.31
15	72	79	0.90	9	32	0.28

The table above shows that customers with a high late payment rate did not necessarily have a high utilization rate, and vice versa.

We can conclude that there's no strong relationship between high credit utilization and frequent late payments, as expected, since there is no consistent correlation between these two indicators.

2. Are there early signs of credit distress?

The goal with this question is to identify patterns before a customer starts missing payments. However, this requires trend analysis, which is hard to do in SQL. Therefore, we need another tool to visualize the data, which will be Power BI.

3. How do risk patterns differ by region or demographic?

We need to evaluate the risk metrics for dimensions, such as provinces, age groups, income levels, etc.

	gender text	num_customers bigint	avg_utilization numeric	late_rate numeric
1	Other	8	0.42	0.12
2	Non-Binary	18	0.42	0.09
3	Female	306	0.44	0.08
4	Male	291	0.46	0.07

'Males' tend to have a higher utilization rate, while people who identified themselves as 'Other' have a higher late payment rate. However, the difference is small.

	province character varying (2)	num_customers bigint	avg_utilization numeric	late_rate numeric
1	BC	83	0.47	0.08
2	NS	11	0.40	0.08
3	QC	134	0.44	0.08
4	AB	64	0.45	0.07
5	ON	316	0.45	0.07
6	MB	15	0.46	0.06

The provinces have similar late payment rates, with Nova Scotia (NS) presenting a lower utilization rate than the others.

	income_bracket text	num_customers bigint	avg_utilization numeric	late_rate numeric
1	<30K	130	0.50	0.15
2	>100K	66	0.44	0.06
3	30K-60K	228	0.43	0.06
4	60K-100K	199	0.44	0.06

Clients with an income lower than 30 K present higher utilization and late payment rates.

	employment_status text	num_customers bigint	avg_utilization numeric	late_rate numeric
1	Retired	63	0.47	0.10
2	Self-Employed	63	0.48	0.09
3	Employed	380	0.45	0.07
4	Student	55	0.43	0.07
5	Unemployed	62	0.44	0.07

Retired and Self-Employed present higher utilization and late payment rates.

	marital_status text	num_customers bigint	avg_utilization numeric	late_rate numeric
1	Married	252	0.44	0.08
2	Single	312	0.46	0.08
3	Divorced	35	0.43	0.07
4	Widowed	24	0.40	0.06

People from different marital statuses have similar late payment rates, with Widows presenting a lower utilization rate than the others.

	age_band text	num_customers bigint	avg_utilization numeric	late_rate numeric
1	<20	25	0.44	0.10
2	20-29	108	0.44	0.08
3	30-39	117	0.44	0.07
4	40-49	117	0.44	0.07
5	50-59	98	0.45	0.08
6	60+	158	0.46	0.08

Younger people up to 20 years of age have a slightly higher late payment rate. The groups present a similar utilization rate.

4. Where can we reduce exposure without hurting loyal customers?

The objective is to identify risky clients that potentially need a lower credit limit, but not so risky that they'll churn or default if the company acts aggressively.

	customer_id integer	card_id integer	account_age numeric	avg_utilization numeric	late_payments bigint
1	209	245	1	0.89	2
2	473	559	0	0.86	2

The table above identifies the customers with high risk and a short relationship with the company. They have an account less than 2 years old, a 0.85 utilization rate, and at least 2 late payments. There are only two recent clients that fit these criteria, indicating that they might need a higher credit limit.

	customer_id integer	avg_spend numeric	late_rate numeric
1	289	38.81	1.00
2	89	40.31	0.50
3	41	37.21	0.43
4	204	40.35	0.40
5	244	42.92	0.38
6	14	39.17	0.34
7	283	38.76	0.31

The table identifies customers with low spend and repeated lateness. They have at least a 1-year-old account, spend less than $100.00 per month, and have a late payment rate equal to or higher than 0.30. These are the customers to keep an eye on, as they present financial risk but provide limited value.

Final views

The final step was to create two new views to be used in Power BI. This will facilitate performing the data transformations and visualizations without needing to rejoin.

The created views are:

- Risk Analysis View, containing data from customers, credit_cards, and payment_behavior. It is useful for customer segmentation, late payment and utilization analysis, etc.
- Spending Analysis View, containing foreign keys and data from the transactions table. It is useful for category spend analysis.

Conclusion

This SQL analysis was essential for identifying credit risk, segmenting customers, and preparing the data for dashboarding. The next stage in Power BI will focus on uncovering trends and presenting actionable insights to support risk monitoring and credit limit adjustments.