Technical Writing Sample by Patricio Kobek: Analysis of Credit Card Debt, Tableau Visualization, and SQL Code

Introduction:

When it comes to expanding your business in the lending industry, it is crucial to understand the challenges and opportunities associated with providing credit cards to consumers. Credit cards are a crucial part of modern-day finance, and they offer consumers a convenient way to access credit. As credit is extended, there are several observations to be made from analyzing the data of consumer usage. This report will demonstrate how to analyze lending data and provide key takeaway observations and business recommendations to expand lending[i].

Dataset:

The original dataset can be found here, featuring over 65,000 entries of credit card users in Taiwan in 2005. Although the data is not current, observations presented below remain applicable to today's market. The data provided is the limit of each account, the age, gender, marital status, and education of the consumer, and six months of account balance and payment history.

Observations and Business Recommendations:

Drilling down into the refined data and examining the proportion of available credit available to a consumers based on their age (divided into three groups 18-34, 35-54, and 55+), level of education, and marital status provides a clear picture of two groups of individuals: low-risk consumers that are ideal to offer higher credit limits, and consumers near the maximum of their available credit limit, indicating a high likelihood of having debt at competing institutions that could qualify for consolidation, improving their financial health and more importantly, providing an opportunity to collect all their business from a competitor.

The first group of individuals, represented below in the chart "Percentage of Credit Limit Available" with values (green in color) representing accounts with little balance owing each month.

For example, women who have completed graduate school, are married, and are between the age ranges of 18-34 and 35-54 have approximately 76% and 89% of their credit limit available over the six-month period of this dataset. This indicates that they and other consumers with similarly low balances would be low-risk candidates for credit limit increases and may also have income available for investments (for organizations like banks that could use cross promotion).

The second group of individuals in the same chart can be spotted by the bright-red values. These represent credit limits near capacity. A pessimist might say that these accounts are at risk of default, however an optimist should recognize an opportunity here, because it is often the case that consumers

have credit cards from a variety of organizations, not limited only to banks, but also from credit unions, for vehicle loans, from departments stores to name but a few. A proactive account manager should reach out to these individuals and inquire about the consumers overall debt structure. If there are other accounts at or near their limits elsewhere, the strategic move would be to offer a consolidation loan of all credit, improving the financial health of the consumer and earning their loyalty, and additionally bringing all their business to us.

As for risk in such a proposal, in granting these consumers their initial credit cards, a list of assets should have been considered for risk assessment prior to approval, and those with assets like homes may have equity that would make such a consolidation loan relatively risk free to our organization. These individuals should be the highest priority to grow sales in the future.

Percentage of Credit Limit Available

Sex	Marriage	Education	Age Range 18-34	35-55	55+
Man	Married	Graduate School	36.39	56.49	
		High School	50.87	53.03	52.72
		Other	66.37	71.72	63.77
		University	42.81	50.21	54.98
	Other	Graduate School	97.69		
		High School	35.36	38.59	65.88
		Other	81.92	39.85	21.93
		University	25.09	36.88	19.90
	Single	Graduate School	51.61	99.44	
		High School	44.31	39.56	21.93
		Other	64.56	58.66	38.38
		University	44.09	40.83	46.64
Woman	Married	Graduate School	76.21	88.99	2.53
		High School	51.67	59.64	54.70
		Other	74.73	78.81	81.00
		University	48.64	53.43	51.48
	Other	Graduate School	99.64	100.17	
		High School	48.40	50.04	34.51
		Other	52.85	56.72	
		University	28.18	42.45	64.28
	Single	Graduate School	62.84	82.57	
		High School	51.92	55.98	46.67
		Other	67.95	77.06	36.14
		University	53.45	50.98	41.20

Chart I: *Percentage of Credit Limit Available*[ii]

SQL Code:

```
/*Dataset for Credit Card use, Taiwan, second half of 2005

Notes On Column and values:
Table name is PortfolioProject.dbo.Credit_Data
This research employed a binary variable, default payment (Yes = 1, No = 0), as the response variable. This study reviewed the
literature and used the following 23 variables as explanatory variables:
Limit_bal: Amount of the given credit (NT dollar)- it includes both the individual consumer credit and his/her family
(supplementary) credit.
Sex (1 = male; 2 = female).
Education (1 = graduate school; 2 = university; 3 = high school; 4 = others). (0, 5, and 6, ommitted, see below)
Marriage (1 = married; 2 = single; 3 = others). "Others" must include divorced, seperated, widowed.
Age (year).
Pay_0 to Pay_6: History of past payment. Monthly payment records (from April to September, 2005) as follows:
X6 = the repayment status in September, 2005; X7 = the repayment status in August, 2005; . . .;
X11 = the repayment status in April, 2005. The measurement scale for the repayment status is: -1 = pay duly;
1 = payment delay for one month; 2 = payment delay for two months; . . .; 8 = payment delay for eight months;
9 = payment delay for nine months and above.
Bill_amt1 to Bill_amt6: Amount of bill statement (NT dollar) (2005)
Pay_amt1 to Pay_amt6: Amount of previous payment (NT dollar).
Default Payment next month: Target Column - binary variable, default payment (Yes = 1, No = 0)

Note: Posted a question to the publication forum asking what values 0, 5 and 6 mean in Education.
The creator of this data set does not know what 5 and 6 refer to. Next, check proportion of entries
affected by this and make a note in final report reflecting this potential for data innacuracy.
Resolution: Values 0, 5 and 6 cannot be defined, so instead query to show they make up a total of 1.112% of the total data
and is not significant enough to cause problems in our analysis, but will still be mentioned in the write up.
There are also values for 0 in Education and Marriage.

For the final report, SQL queries #6, 7, and 8 below were the most important, and everything before that provided useful context
in establishing the best course of observation.
*/

select * from PortfolioProject.dbo.Credit_Data

/*
1. Mean amount of credit for each combination of gender, education level, and marital status.
*/
SELECT
    Sex,
    Education,
    Marriage,
    AVG([Bill_AMT1]) AS MeanCredit
FROM
    PortfolioProject.dbo.Credit_Data
GROUP BY
    Sex,
    Education,
    Marriage
ORDER BY
    MeanCredit DESC;

    --Not sure if useful, checking proportions, might be necessary if Education 5 and 6 are large.
SELECT Sex, Education, Marriage,
    AVG(Sex) as Mean_Credit_Amount,
    COUNT(*) as Entry_Count,
    COUNT(*) / (SELECT COUNT(*) FROM PortfolioProject.dbo.Credit_Data) as Proportion
FROM PortfolioProject.dbo.Credit_Data
GROUP BY Sex, Education, Marriage
ORDER BY Sex, Education, Marriage;
```

```sql
--Proportion of entries by Education
SELECT
  Sex,
  Education,
  Marriage,
  COUNT(*) * 100.0 / (SELECT COUNT(*) FROM PortfolioProject.dbo.Credit_Data) AS entry_proportion,
  SUM(Limit_bal) * 100.0 / (SELECT SUM(Limit_bal) FROM PortfolioProject.dbo.Credit_Data) AS credit_proportion
FROM PortfolioProject.dbo.Credit_Data
GROUP BY Sex, Education, Marriage
ORDER BY Education DESC;

--Sum proportion of 5 and 6 for education to see if relevant.

SELECT
    ROUND(CAST(COUNT(*) AS FLOAT) / (SELECT COUNT(*) FROM PortfolioProject.dbo.Credit_Data), 4) AS PropEntries,
    ROUND(SUM(Limit_bal) / (SELECT SUM(Limit_bal) FROM PortfolioProject.dbo.Credit_Data) * 100, 2) AS PropDebt
FROM PortfolioProject.dbo.Credit_Data
WHERE Education IN (5, 6)

--Total number of entries with education as 0, 5 or 6 is 730 out of 65,634, or 1.112%
SELECT
    COUNT(*) AS total_entries,
    SUM(CASE WHEN Education IN (0,5,6) THEN 1 ELSE 0 END) AS education_5_6_count
FROM
    PortfolioProject.dbo.Credit_Data



--2. Are there any other inconsistencies that require attention?
--Answer: No, there are no NULL values to consider.

SELECT *
FROM PortfolioProject.dbo.Credit_Data
WHERE Limit_bal IS NULL
OR Sex IS NULL
OR Education IS NULL
OR Marriage IS NULL
OR Age IS NULL
OR Pay_0 IS NULL
OR Pay_2 IS NULL
OR Pay_3 IS NULL
OR Pay_4 IS NULL
OR Pay_5 IS NULL
OR Pay_6 IS NULL
OR Bill_amt1 IS NULL
OR Bill_amt2 IS NULL
OR Bill_amt3 IS NULL
OR Bill_amt4 IS NULL
OR Bill_amt5 IS NULL
OR Bill_amt6 IS NULL
OR Pay_amt1 IS NULL
OR Pay_amt2 IS NULL
OR Pay_amt3 IS NULL
OR Pay_amt4 IS NULL
OR Pay_amt5 IS NULL
OR Pay_amt6 IS NULL
OR [Default Payment Next Month] IS NULL;
```

```sql
/*
3. Average Credit Card debt by age, education, and marriage.
Does one group stand out? What might that mean?

Maybe better to consider average debt (beginning to end) by age group,
and then grab the highest ones, and deep dive like avaerage debt of that age group
by education, marriage, and gender. Are there any big outliers?
*/

SELECT
    CASE
        WHEN Age >= 18 AND Age <= 28 THEN '18-28'
        WHEN Age >= 29 AND Age <= 39 THEN '29-39'
        WHEN Age >= 40 AND Age <= 50 THEN '40-50'
        WHEN Age >= 51 AND Age <= 61 THEN '51-61'
        WHEN Age >= 62 AND Age <= 73 THEN '62-73'
        WHEN Age >= 74 AND Age <= 85 THEN '74-85'
        ELSE '86+'
    END AS Age_Range,
    Education,
    Marriage,
    AVG(Bill_amt6) AS Avg_Bill_amt6
FROM
    PortfolioProject.dbo.Credit_Data
GROUP BY
    CASE
        WHEN Age >= 18 AND Age <= 28 THEN '18-28'
        WHEN Age >= 29 AND Age <= 39 THEN '29-39'
        WHEN Age >= 40 AND Age <= 50 THEN '40-50'
        WHEN Age >= 51 AND Age <= 61 THEN '51-61'
        WHEN Age >= 62 AND Age <= 73 THEN '62-73'
        WHEN Age >= 74 AND Age <= 85 THEN '74-85'
        ELSE '86+'
    END,
    Education,
    Marriage
ORDER BY Avg_Bill_amt6 DESC;


/*
4. Mean credit (Use Bill_AMT1) based on Education and age ranges (18-34, 35-55, 55+), seeking correlation between
age, education, and outstanding credit. Remember NOT IN 0, 5, 6 for education, see notes above.
*/
SELECT
  CASE
    WHEN Age >= 18 AND Age <= 34 THEN '18-34'
    WHEN Age >= 35 AND Age <= 55 THEN '35-55'
    ELSE '55+'
  END AS AgeRange,
  Education,
  AVG(Bill_amt1) as MeanCredit
FROM PortfolioProject.dbo.Credit_Data
WHERE Education NOT IN (0,5,6)
GROUP BY
  CASE
    WHEN Age >= 18 AND Age <= 34 THEN '18-34'
    WHEN Age >= 35 AND Age <= 55 THEN '35-55'
    ELSE '55+'
  END,
  Education
ORDER BY AgeRange, Education;
```

```sql
/*
5. Mean credit (Use Bill_AMT1) based on Education and age ranges (18-34, 35-55, 55+), and now by gender.
*/
SELECT
  CASE
    WHEN Age >= 18 AND Age <= 34 THEN '18-34'
    WHEN Age >= 35 AND Age <= 55 THEN '35-55'
    ELSE '55+'
  END AS AgeRange,
  Education,
  Sex,
  AVG(Bill_amt1) as MeanCredit
FROM PortfolioProject.dbo.Credit_Data
WHERE Education NOT IN (0,5,6)
GROUP BY
  CASE
    WHEN Age >= 18 AND Age <= 34 THEN '18-34'
    WHEN Age >= 35 AND Age <= 55 THEN '35-55'
    ELSE '55+'
  END,
  Education,
  Sex
ORDER BY AgeRange, Education;


/*
6. Mean credit (Use Bill_AMT1) based on Education and age ranges (18-34, 35-55, 55+), gender, and now
marital status. Which individual group owes the most? From there, which individual group appears most likely
to default on payment?

This is my main observation used to answer the question "which target could borrow more, which are consistent, and which
appear highest risk?" Last part about risk can be drilled down by seeing payment history and outstanding balance relative to limit.
*/
SELECT
  CASE
    WHEN Age >= 18 AND Age <= 34 THEN '18-34'
    WHEN Age >= 35 AND Age <= 55 THEN '35-55'
    ELSE '55+'
  END AS AgeRange,
  Education,
  Marriage,
  Sex,
  AVG(Bill_amt1) as MeanCredit
FROM PortfolioProject.dbo.Credit_Data
WHERE Education NOT IN (0,5,6) or
Marriage NOT IN (0)
GROUP BY
  CASE
    WHEN Age >= 18 AND Age <= 34 THEN '18-34'
    WHEN Age >= 35 AND Age <= 55 THEN '35-55'
    ELSE '55+'
  END,
  Education,
  Marriage,
  Sex
ORDER BY AgeRange, Education;
```

```sql
/*
7. Checks limit versus outstanding balance by age group overall, and shows proportion unused.
After seeing which clients are medium and lowest risk, we can compare their Limit_Bal, and offer an increase proportional
to their risk (from low to medium risk) to increase targets for a year).

At the same time, we may also be able to speculate that those closest to the limit may be more prone to default,
but there is a secondary opportunity here, in that if they are near the limit
here, they may also be near the limit elsewhere, and we could potentially
offer consolidation of outside debt to get all their business, though this
is best suited for clients with collateral (Home equity for consolidation)
*/

SELECT
  CASE
    WHEN Age >= 18 AND Age <= 34 THEN '18-34'
    WHEN Age >= 35 AND Age <= 55 THEN '35-55'
    ELSE '55+'
  END AS AgeRange,
  Education,
  Sex,
  Marriage,
  AVG(Bill_amt1/Limit_bal)*100 AS Available_Credit_Percentage
FROM PortfolioProject.dbo.Credit_Data
WHERE Education NOT IN (5,6)
GROUP BY
  CASE
    WHEN Age >= 18 AND Age <= 34 THEN '18-34'
    WHEN Age >= 35 AND Age <= 55 THEN '35-55'
    ELSE '55+'
  END,
  Education,
  Sex,
  Marriage


  /*
8. Query #7 reveals some excellent information about who to lend to, and who might be offered consolidation.

Now, query to see if there is a large difference between the groups and the number of people in each category.
Initial assumption is it should be fairly equal accross the board.
*/

  SELECT
    Sex,
    Marriage,
    Education,
    COUNT(*) AS TotalEntries,
    ROUND(CAST(COUNT(*) AS FLOAT) / SUM(COUNT(*)) OVER(), 2) AS Proportion
FROM
    PortfolioProject.dbo.Credit_Data
WHERE
    Education NOT IN (0, 5, 6)
GROUP BY
    Sex, Marriage, Education;
```

Endnotes

[i] A few notes regarding data validation and cleaning. The source for this data project identified the values used in each column. For example, highest level of Education is represented by values 1 for “Other”, 2 for “High School”, 3 for “University”, and 4 for "Graduate School”. However, values 0, 5, and 6 appear in the data when performing initial SQL queries in search of outlying or NULL data points. This issue appears as well in the Marriage column, where 0 appears but is undefined.

I inquired to the dataset provider about these unlabelled values, but they do not know what the values represent. As a result, these values are excluded from the analysis, totalling 730 out of 65,634 data points, or 1.1% of the total sample. The sample remains large enough to proceed.

Second, the Marriage category only features three options, 1 for “Married”, 2 for “Single”, and 3 for “Other” (and the aforementioned value of 0). In most cases, the “Other” category makes up a considerable percentage of entries across all age and education levels. This value most likely represents individuals who are divorced, separated, widowed, and common law, but this too could not be confirmed.

[ii] White spaces in this chart represent no available data for this age group.