Dataset Exploration

Introduction:

The “Bank Churners” dataset includes the data with respect to credit card usage information and the history of the customer with the bank. This dataset helps us to understand credit card usage behavior and to proactively analyze which user is about to leave the bank services and to offer the respective users more bank services and offers. I find this dataset very interesting which could help me understand the different attributes and some important analysis of churned users utilizing the credit card usage dataset. I got this dataset from Kaggle but originally it is from the following website: https://leaps.analyttica.com/home. This dataset contains more than 10,000 records and 21 different variables. The dataset can be divided into Customer information and its relationship with the bank, and the Credit card usage behavior of a customer. Customer information contains the following attributes:

- Customer Age
- Gender
- Education
- Marital Status
- Income Category
- Number of dependants
- Attrition Flag

Credit card usage behavior contains the following attributes:

- Type of credit card
- Period of relationship with the bank.
- Credit limit
- Revolving balance on credit card
- Total transaction count
- Average card utilization ratio
- Transaction Amount

The attrition flag variable is one of the deciding variables that will be used in analyzing existing customers who are close to leaving the bank. For this, I will be using the data and analytics from the “Attrited Customers” value of the Attrition Flag variable.

Data Dictionary:

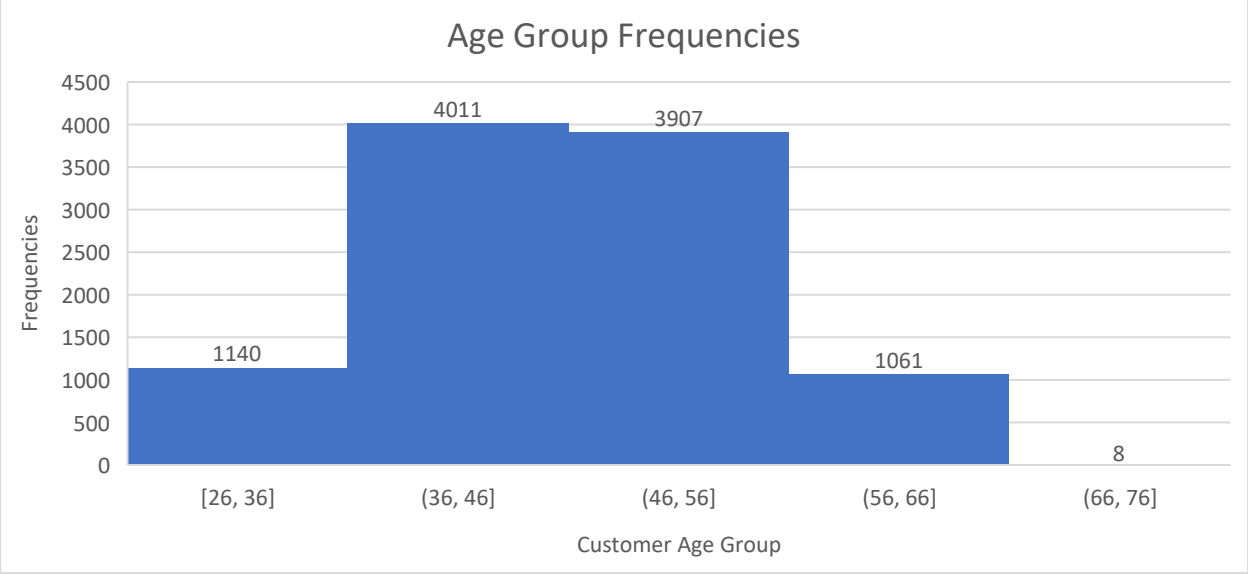
Variable	Variable Type	Description	Range
Client_Number	Continuous	Unique 9-digit Identifier for the customer holding the bank account	N/A
Attrition Flag	Nominal	This is an internal event that describes if the account is still active or not.	Existing Customers, Attrited Customers
Customer_Age	Nominal	Customer's Age in years	26 – 73
Gender	Nominal	Gender of a customer	M = Male, F = Female
Dependent_count		Number of dependants for a customer	0 – 5
Education_Level	Ordinal	Educational qualification of the account holder	High School, College, Graduate, Doctorate, Post-Graduate, Uneducated, Unknown
Marital_Status	Nominal	Marital status of a customer	Married, Single, Divorced, Unknown
Income_Category	Ordinal	Annual Income Category of the account holder	$40K, $40K - 60K, $60K - $80K, $80K-$120K, $120K >
Card_Category	Ordinal	Type of the card	Blue, Silver, Gold, Platinum
Months_on_book	Discrete	Period of relationship with the bank	N/A
Total_Relationship_Count	Discrete	Total no. of products held by the customer	1 – 6
Months_Inactive_12_mon	Discrete	No. of months inactive in the last 12 months	0 - 6

Contacts_Count_12_mon	Discrete	No. of Contacts in the last 12 months by bank	0 – 6
Credit_Limit	Continuous	Credit Limit on the Credit Card	N/A
Total_Revolving_Bal	Continuous	Total Revolving Balance on the Credit Card	N/A
Avg_Open_To_Buy	Continuous	Open to Buy Credit Line (Average of last 12 months)	N/A
Total_Amt_Chng_Q4_Q1	Continuous	Change in Transaction Amount (Q4 over Q1)	N/A
Total_Trans_Amt	Continuous	Total Transaction Amount (Last 12 months)	N/A
Total_Trans_Ct	Continuous	Total Transaction Count (Last 12 months)	N/A
Total_Ct_Chng_Q4_Q1	Continuous	Change in Transaction Count (Q4 over Q1)	N/A
Avg_Utilization_Ratio	Continuous	Average Card Utilization Ratio	0 - 1

Univariate Analysis:

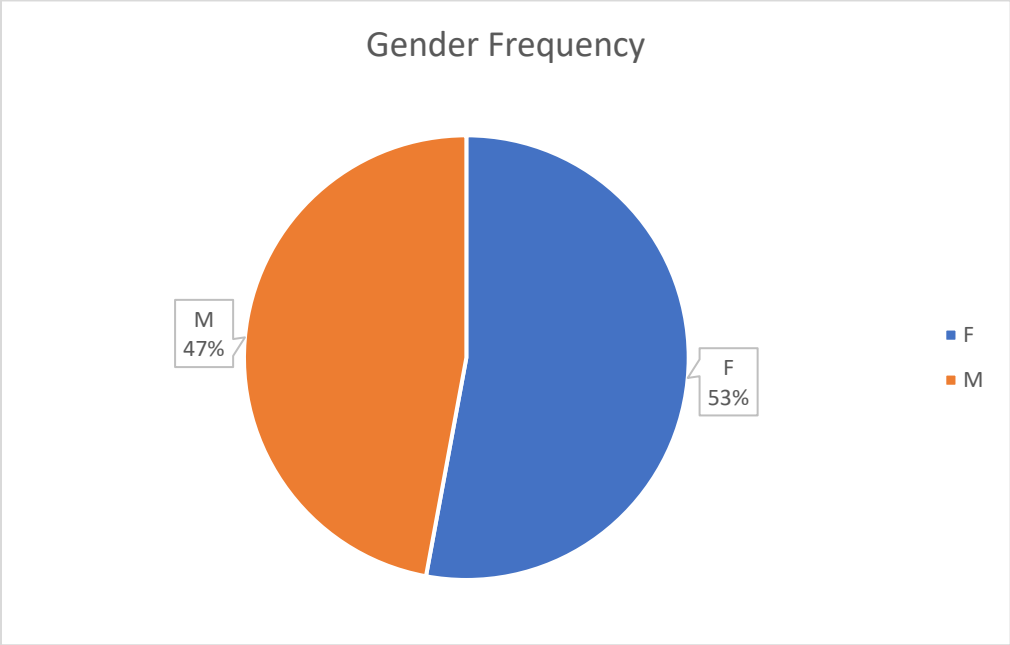
1. **Age Group univariate analysis:**

   Histogram has been used for this analysis which groups the ages within a bin of 10. This helps us understand our customer base. In the below chart, it is clearly visible that the age group [36-46] and [46-56] have a very tight gap in credit card usage.



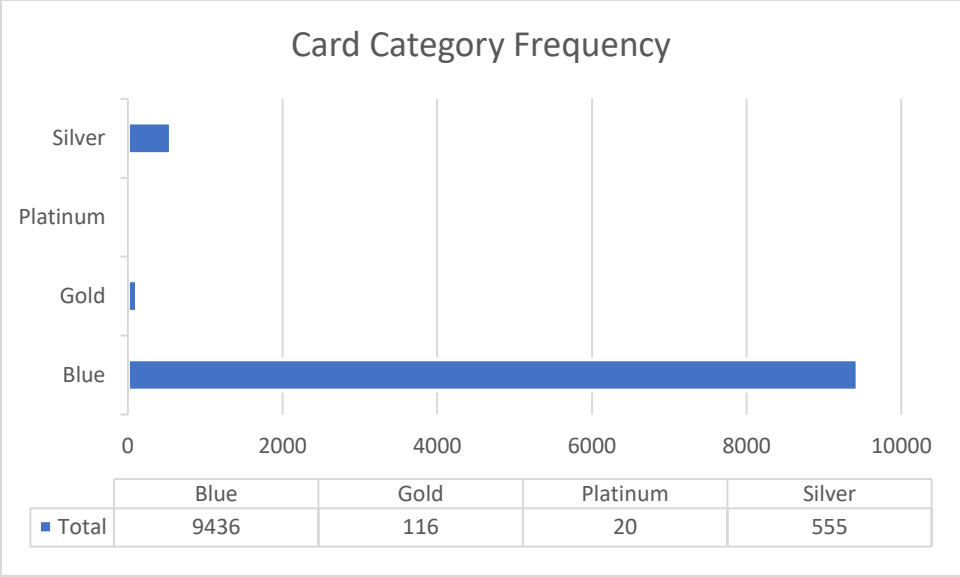
2. **Gender:**

For Gender, I have used a pie chart that shows the distribution of Male and Female cardholders in the dataset.



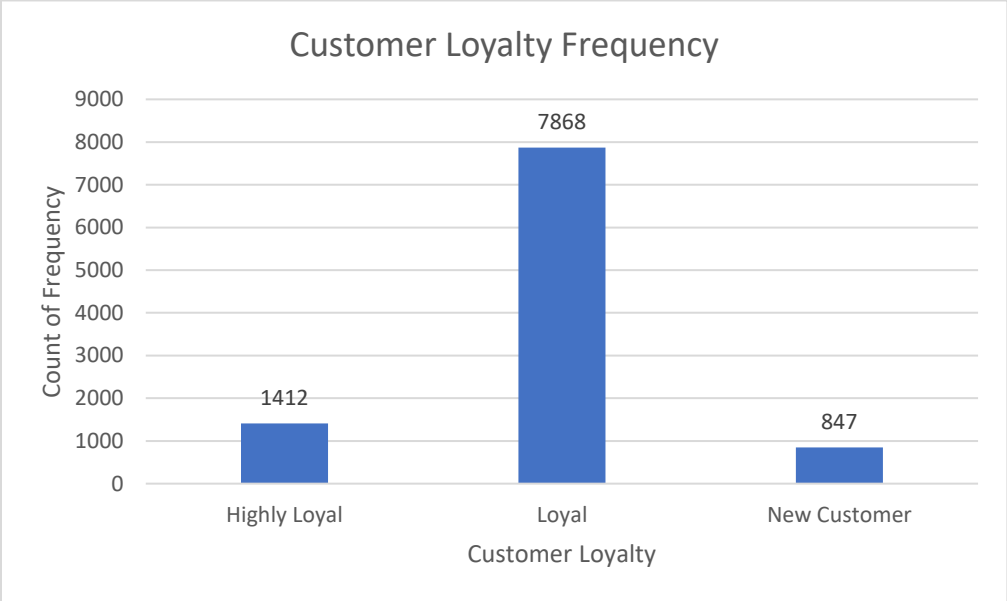
3. **Card Category Frequency:**

It shows the frequency of the card category in a horizontal bar chart. A Bar chart provides the flexibility to represent data in the form of a table with values, making it easier for the reader to understand the chart.



4. **Customer Type Frequency:**

This variable helps us understand the type of customer. The category is formed using a variable 'Months_on_book'. I have used a bar chart to better understand the variable.



5. **Credit Limit:**

The credit limit is a limit on each card. With the below analysis, we can understand the minimum, maximum, and average, range of all the credit cards.

Credit Limit	
Minimum	$1,438.30
Maximum	$34,516.00
Average	$8,631.95
Sum	$87,415,795.10
Count	10127
Range	33077.7
Median	4549
Kurtosis	1.81
Skewness	1.67

6. **Total Transaction Amount:**

The total transaction amount determines the overall sum of transactions completed in 12 months. The below analysis gives deeper look at the variable.

Total Transaction Amount	
Minimum	$510
Maximum	$18,484
Average	$4,404.09
Sum	$44,600,182

Count	10127
Range	17974
Median	3899
Kurtosis	3.89
Skewness	2.04

7. **Total Transaction:**

The below analysis focuses on the transaction count variable for minimum, maximum, average, sum, and other mathematical calculations which helps to understand the variables.

Total Transactions	
Minimum	10
Maximum	139
Average	64.85
Sum	656824
Standard Deviation	23.47
Variance	550.96
Count	10127
Range	129
Median	67
Kurtosis	-0.37
Skewness	0.15

8. **Utilization Ratio:**

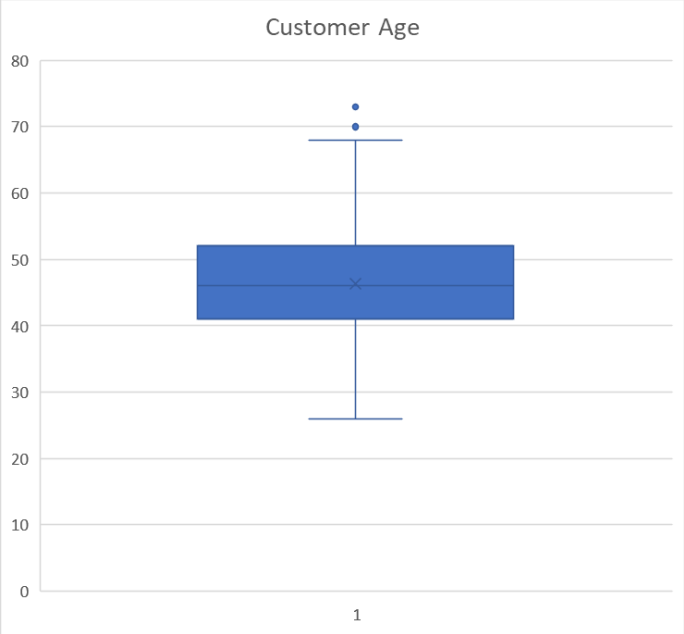
Utilization ratio analysis is done in table form, providing the maximum, minimum, and other mathematical operations.

Utilization Ratio	
Minimum	0
Maximum	0.999
Average	0.274
Standard Deviation	0.275
Variance	0.076
Median	0.176
Sum	2783.84
Kurtosis	-0.79
Skewness	0.72

Clean your Data set / Outliers:

I am visualizing Box Plot on customer age to find out if there are any outliers in the dataset. As the below chart clearly says that there are two points outside of a given range that can be considered outliers.

But for further analysis, I may need those points because of which I am not deleting any outliers which may affect the analysis in future.



Although there are some "Unknown" values that I would keep for further analysis.

Education_Lev	Marital_Stat	Income_Catego
High School	Married	Less than $40K
Unknown	Single	$40K - $60K
Doctorate	Divorced	$80K - $120K
Uneducated	Single	Less than $40K
Uneducated	Married	Unknown
Unknown	Unknown	$120K +
Graduate	Married	Less than $40K
High School	Single	Less than $40K
College	Single	Less than $40K
High School	Single	$60K - $80K

The following 'Unknown' values are important for analysis as they have other important information such as utilization ratio, card category, gender, and age which can be used in the analysis, and deleting the rows with 'Unknown' values will greatly affect the analysis.

Categorize or Code variable:

It was very important to create a categorized variable from the 'Months_on_book' variable. I have created a variable called 'Customer Loyalty'. This represents the relationship with the bank/credit card company. If the relationship is short, then it categorizes as a "New Customer", if it is more than 2 years then it represents "Loyal" and if more than 3.5 years then it represents a "Highly Loyal" customer. This would open a lot of analysis possibilities with the combination of another categorical variable.

K
Customer Loyal
Loyal
Highly Loyal
Loyal
New Customer
Loyal
Loyal
New Customer
Loyal
Loyal
Highly Loyal

Hypothesis Testing:

1. Married females have more average credit card utilization ratio.

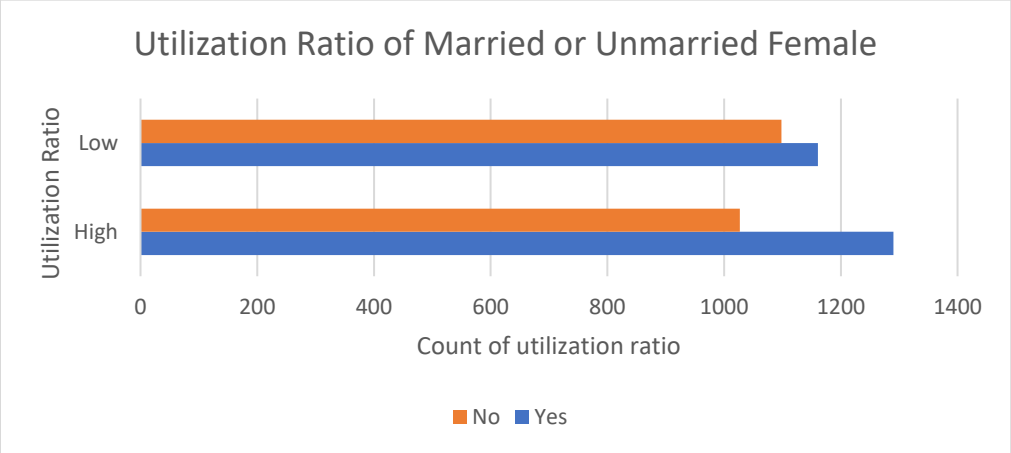
Let us understand the data and apply an Odds Ratio testing for a significant outcome.

For this testing, I have created two dichotomous variables to compare groups. Married Females and Average Credit Card utilization ratio are high or low. Both variables are coded for this hypothesis.

Utilization Ratio variable - If the credit card utilization is more than 30% then **High** else it is **Low**.

Married Females variable – If the female is married then **Yes** else it is **No.**

	Utilization Ratio	
Married Female?	**High**	**Low**
Yes	**A = 1290**	**B = 1161**
No	**C = 1027**	**D = 1098**



Odds Ratio	**1.19**	NOTE: Formula for Odds Ratio (OR): OR = (A*D) / (B*C)
Standard Error	**0.06**	NOTE: Formula for Standard Error (SE): SE = SQRT((1/A) + (1/B) + (1/C) + (1/D))
z-score for 95% confidence	**1.96**	z-score for a 95% confidence interval is 1.96
Lower Limit:	**0.96**	Lower Limit Calculation: LL = EXP(LOG(OR) – (SE*Z))
Upper Limit:	**1.21**	Upper Limit Calculation: UL = EXP(LOG(OR) + (SE*Z))

With the Odds Ratio of 1.19, statistically, we can say that there is a significant association between the exposure and the outcome.

Females who are married are 1.19 times more likely to spend more of their average credit card utilization ratio than single females, with a 95% CI of 0.96 – 1.21.

However, the fact that the lower limit is below 1 suggests that the effect may be small and further studies may be needed to confirm the findings. [Maybe, considering "Divorced" females as Single Females could result in more change in the dependent variable]

2. A report read that highly loyal existing customers engage with more products.

To understand the hypothesis, I am applying an Odds Ratio on this hypothesis. Two dichotomous variables that were created are: Highly Loyal Customers or Engaged with multiple products.

If the relationship count is more than 3, then it is categorized as Highly Engaged else Less Engaged.

	Engaging with multiple products?	
Highly Loyal Customers?	**Highly Engaged**	**Less Engaged**
Yes	**738**	**443**
No	**4283**	**3036**

Odds Ratio	**1.18**	NOTE: Formula for Odds Ratio (OR): OR = (A*D) / (B*C)
Standard Error	**0.06**	NOTE: Formula for Standard Error (SE): SE = SQRT((1/A) + (1/B) + (1/C) + (1/D))
z-score for 95% confidence	**1.96**	z-score for a 95% confidence interval is 1.96
Lower Limit:	**0.95**	Lower Limit Calculation: LL = EXP(LOG(OR) – (SE*Z))
Upper Limit:	**1.22**	Upper Limit Calculation: UL = EXP(LOG(OR) + (SE*Z))

The Odds Ratio is 1.18, statistically, we can say that there is a positive association between the exposed group and the outcome.

Customers who are Highly Loyal are 1.18 times more likely to engage with multiple products that credit card company has to offer than other customers, with a 95% CI of 0.95 – 1.22.

As the lower limit is below 1, we can say that the effect may be small and further studies may help to derive the outcome.

3. The income category has an effect on overall transactions. [FINER Question]

The above hypothesis can be better evaluated with a Chi-Squared test as there are multiple categories in the income variables. This hypothesis is tested on a sample size of 5000.

Considering the null hypothesis – The "less than $40K" income category would have more transactions count than other incomes (40:60 ratio).

Hypothesized proportions and the information is displayed in the below table:

Income Category	Hypothesized Proportions	Observed	Expected	Chi-Squared
Less than $40K	**0.4**	**1806**	**1770.8**	**0.70**
$40K - $60K	**0.15**	**878**	**664.05**	**68.93**
$60K - $80K	**0.15**	**672**	**664.05**	**0.10**
$80K - $120K	**0.15**	**725**	**664.05**	**5.59**
$120K+	**0.15**	**346**	**664.05**	**152.33**
Total		**4427**		**227.65**

Is the chi-squared value 227.65 meaningful? We can calculate the p-value from the chi-squared using the formula in Excel. The calculated p-value is shown below:

p-value	**4.22E-48**

For a 95% CL, the alpha is 0.05. *The p-value is very much less than the alpha so we can strongly reject the null hypothesis.*

4. Bank survey says that the types of customers have an equal proportion of males and females.

Null Hypothesis: Both males and females fall under an equal proportion of customers.

To test this hypothesis, I am using a chi-squared test and the data for the same is displayed below:

Actual Value				
Gender	**Highly Loyal**	**Loyal**	**New Customer**	**Total**
Female	417	3066	275	3758
Male	328	2638	276	3242
Total	745	5704	551	7000

The expected values are calculated using the formula: (Row Total * Column Total)/Grand Total. The table for the expected values is:

Expected Value			
Gender	**Highly Loyal**	**Loyal**	**New Customer**
Female	400	3062	296
Male	345	2642	255

The chi-Squared value is calculated by: (Actual – Expected)/Expected:

Chi-Squared Value				
Gender	**Highly Loyal**	**Loyal**	**New Customer**	**Total**
Female	0.73	0.00	1.46	2.19
Male	0.84	0.01	1.70	2.54
Total	1.57	0.01	3.16	4.74

p-value	**0.09**

The p-value for the above Chi-Squared (4.74) is 0.09. Alpha is considered 0.05 for 95% CL.

If the p-value is > alpha, then we must accept the null hypothesis. *Here, 0.09 is greater than 0.05, hence we accept the null hypothesis.*

5. Analysis of Variance – ANOVA:

Null Hypothesis: There is no difference in means of age between the credit card holders.

I am using a Single Factor ANOVA for testing the hypothesis. In an excel file, there is observation data for the ANOVA, below you can find the ANOVA summary:

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Blue	4000	184874	46.2185	66.5999077
Gold	116	5271	45.4397	43.2571964
Platinum	20	950	47.5	22.8947368
Silver	555	25352	45.6793	55.0918984

While configuring the single factor ANOVA in excel, I mentioned the alpha to be 0.05 which is considered a default.

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	236.3394335	3	78.7798	1.2215863	0.3001409	2.60680386
Within Groups	302263.5203	4687	64.4898			
Total	302499.8597	4690				

The p-value is 0.30 which is greater than 0.05, so we can accept the null hypothesis.

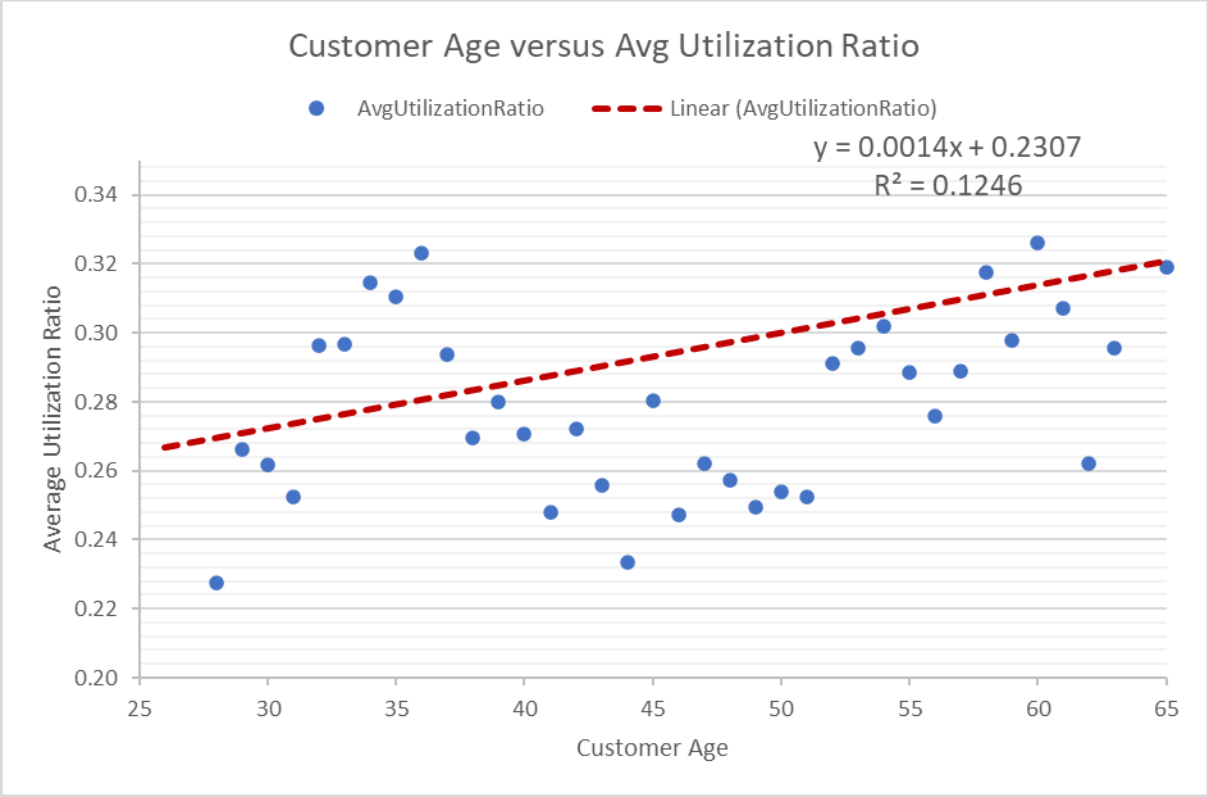
Inferential Techniques:

Inferential techniques are statistical methods used to draw conclusions or make predictions about a population based on a sample of data. The accuracy of inferential techniques depends on the sample size and the representativeness of the sample. Common inferential techniques include hypothesis testing, confidence intervals, and regression analysis.

Let's understand the inferential techniques by answering the research questions:

1. How does the average utilization ratio behave with the increase in customer age?

We can understand this by applying a linear regression on the data where customer age can be an independent variable whereas average utilization ratio is a dependent variable.



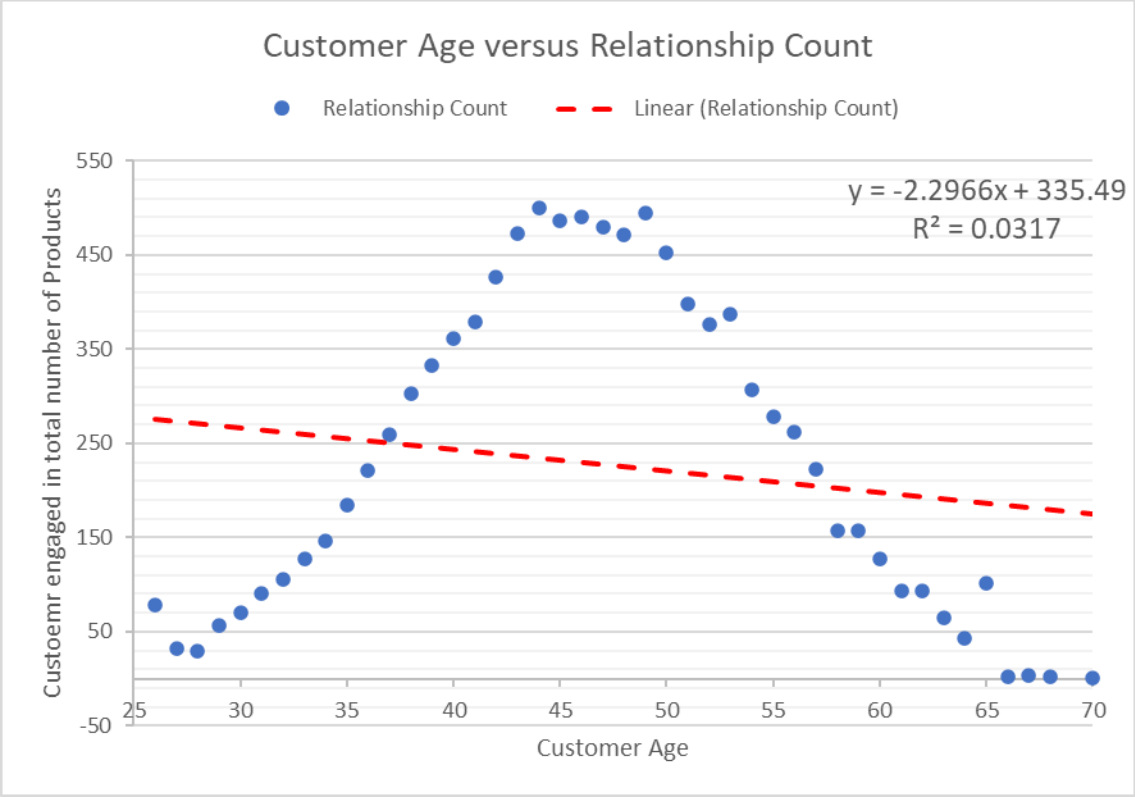
To clean the outliers, I narrowed down the analysis for the customer age between 25-65 and narrowed down the margin of average utilization ratio between 0.20 – 0.35 as 99% of the data lies within this range. The obtained chart (above) is the final result of the analysis performed.

The correlation coefficient for the above analysis is 0.35 which means there is a moderate positive relationship between the two variables. A positive correlation coefficient means that when one variable increases, the other variable also tends to increase, and vice versa. It is important to note that correlation does not imply causation and further analysis may be required to understand the nature and strength of the relationship between the two variables.

The Slope for the above analysis is 0.0018 which means that with the increase in customer age, the average utilization ratio increases by 0.18% times.

2. Are the customer more likely to engage in multiple products with an increase in their age?

To understand this analysis, I applied linear regression on the data where customer age acts as an independent variable and relationship count acts as a dependent variable.



To clean the outliers, I narrowed down the age margin between 25-70 and capped the relationship count to 550.

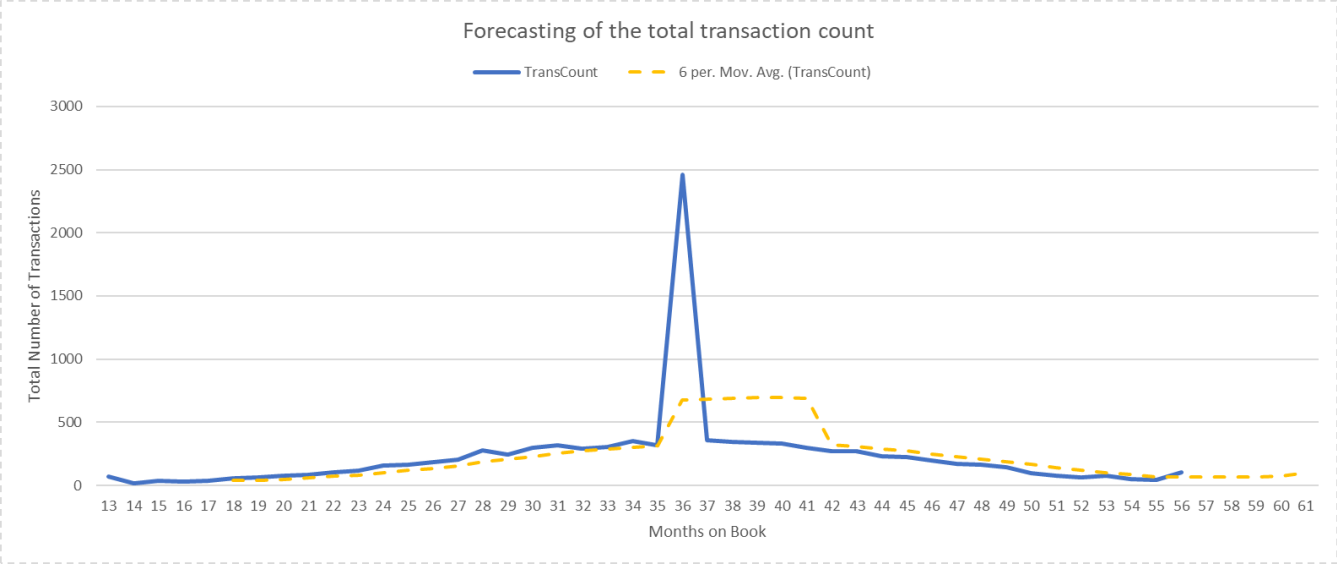
For the above analysis, a negative correlation of -0.17 is obtained which means if one variable tends to increase then the other variable tends to decrease. With the increase in customer age, they are less likely to engage in multiple products.

The slope of the above analysis is -2.29 which means that with the increase in customer age, they are 2.29 times less likely to engage with other products of the company.

This information helps the company to plan such targeted products with respect to the customer's age.

3. Will the customer have more transaction count in the next few months if they stay with the bank longer?

We can achieve this analysis by applying a forecasting method. Here, I am applying 6 months of moving average forecasting to understand the analysis.



Due to the absence of time series data in the dataset, I am analyzing if the customer tends to have more transaction counts if they stay longer with the credit card company.

The "months on book" variable implies the number of months the customer is active and utilizing the credit card.

The 6 months of moving average shows that there is a steady growth in the total number of transaction counts.

Determine any assumptions:

This dataset is largely used for predicting the customers who are close to leaving the banking service or analyzing churned users. Attrited Customer information is used and performed analysis on that to understand the behavior of the existing customers. This is a sample dataset. The dataset consists of both quantitative and qualitative data types. Numerical variables are discrete as well as with currency measurement.

FINER Research Question:

1. Does the income category have any effect on overall transactions?
2. Is there any similarity between Attrited customers or Existing customers with respect to products used, months inactive, income category, or transaction amount?
3. Are the customers who are contacted by banks proven to be highly loyal?
4. Which marital status group is likely to utilize more of their credit limit? Understand the analysis between Attrited and Existing customers.
5. Which age group is likely to use more products and has a better utilization ratio?

More Data required for Analysis?

The current dataset only allows viewing with respect to credit utilization and demographic variables. To accurately derive a reason why the customers are leaving the bank services, additional credit card support data or else compliant/inquiry data would be helpful to understand any issues faced by customers.

Track the analysis:

To be able to further analyze the data, I would require familiarizing myself with some financial terminology to analyze the data further.

I found this dataset on Kaggle for credit card customers that have both qualitative and quantitative data with 10,128 records and 21 columns. The dataset consists of age, gender, educational qualification, marital status, and income category data for the customer and contains credit card usage information.

I am trying to ask more such questions such as: What is the spending ratio based on educational qualification? Which type of credit card is more utilized?

Univariate analysis of the variables proved to be having no possible outliers and no further data cleaning is needed at this point.

Detailed research is required on how to handle the ‘Unknown’ values in the dataset for analysis as there are few variables with ‘Unknown’ values which may affect the analysis if deleted.

Analysis of “Divorced Single” customers could also be helpful to determine certain use cases.

The analysis around the “Education Level” variable with the help of the Chi-Squared test could open up gates to deep-dive analysis.

The regression analysis technique could help more with respect to credit card company data and how the forecasting technique can help retain some of the customers before hand if the analysis done accurately.

References:

Credit Card Customers (Kaggle): https://www.kaggle.com/datasets/sakshigoyal7/credit-card-customers