Analysis of JGA Bank Credit Card Customer Data and Churn Characteristics

Student's Name

Western Governors University

Table of Contents

Project Overview 3
A. Project Highlights 3
Project Plan 3
B. Project Execution 3
Methodology 5
C. Data Collection Process 5
C1. Advantages and Limitations of Data Set 5
D. Data Extraction and Preparation Processes 6
E. Data Analysis Process 6
E1. Data Analysis Methods 6
E2. Advantages and Limitations of Tools/Techniques 6
E3. Application of Analytical Methods 7
Results 9
F. Project Success 9
F1. Statistical Significance 9
F2. Practical Significance 12
F3. Overall Success 12
G. Key Takeaways 12
G1. Summary of Conclusions 12
G2. Effective Storytelling 17
G3. Findings-based Recommendations 17
H. Panopto Presentation 18
Appendices 18
I. Evidence of Completion 18
Sources 18

Project Overview

A. Project Highlights

A1. Research Question:

The question this project sought to answer was whether segmenting consumers with similar traits, such as age, education and income would help identify if certain customer segments are more likely to churn than others. Specifically, I predicted that customers of an older age group, with higher incomes and educational backgrounds would be more likely to churn than other segments.

A2. Project Scope:

The scope of this project was to create a reporting solution via Python that would group credit card consumers into customer segments and analyze churn trends. The solution ingests customer data and provides an output detailing the number of customers present in each customer segment and the percentage of churn for each segment. The output from the reporting solution is intended as guide for a broader analysis an organization may wish to undertake when studying churn trends within its customer base. The results from the application are not meant to provide an explanation as to the cause of churn, or a definitive answer to which customers are ultimately more likely to churn.

A3. Solution Overview:

A3 - I. Tools:

The reporting solution was created using the Python scripting language within a Jupyter Notebook environment. The solution is setup to ingest data from a predefined .csv file input that contains historical customer information such as customer status (active / inactive), age, education, and income, as well as other customer and product specific details. The application ingests data from the input .csv file and is set with parameters to group customers into segments based on age, education, and income. Once the solution has segmented customers, it then produces outputs that detail the number of customers present in each segment as well as the percentage of churn each segment represented during the historical period reviewed.

A3 - II. Methodologies:

To successfully complete this project, multiple methodologies were employed. The Waterfall method of project management was used first and guided this project from start to finish. Secondly, data collection methods were employed to explore and clean the dataset. Descriptive analytics were used to gain insights from the data and to guide the project design. Lastly, statistical analytics were deployed for testing and validation. Each of methods mentioned were used throughout the project, often in tandem with each other and will each be addressed in this report.

Project Plan

B. Project Execution

B1. Project Plan

The project plan outlined in Task 2 and included below, was executed successfully without deviation.

To accomplish this goal, there are three objectives:

1. Create customer segmentations to analyze churn behavior among customers.
    a. The deliverable for this objective is to create customer segmentations for the entire customer base by grouping customers upon factors including age, education, and income.

2. Create a customer segmentation summary for customers who have churned.
    a. The deliverable for this objective will be a summary detailing how many customers are in each customer segment, along with the percentage each customer segment represents of overall churn rate.

3. Create an active customer segmentation summary.
    a. The deliverable for this objective will be a summary detailing how many active customers are in each segment, highlighting those customers who fit the same profile as those identified in the previous summary with higher churn rates, as these customers may be at higher risk of churning, and should be further reviewed.

B2. Project Planning Methodology

To complete this project, a Waterfall methodology was employed. This methodology uses a linear progression, following five phases that require each phase to be completed before proceeding to the next. Each phase is outlined below, along with an explanation of how each step was executed during the completion of this project.

• **Requirements**: During this phase of the project, I began documenting requirements, assumptions and dependencies based on an initial review of my selected dataset.

• **Design**: After gathering all relevant details and completing the Requirements phase, I then began the design phase. During the design phase, I thoroughly reviewed the available data, and then determined how to segment customers and best complete a churn analysis for each segment.

• **Implementation**: After determining an appropriate design that met all my requirements, I then developed my Python script to execute upon my outlined design.

• **Verification & Testing**: During the verification phase I began testing my completed code to ensure functionality. Once I verified my code functioned as desired, I was then able to create a finalized version of the report and test my hypothesis.

• **Deployment & Maintenance**: The last phase of the Waterfall methodology is deployment and maintenance. As this project was academic in nature, this phase was not applicable.

B3. Project Timeline and Milestones

Overall, the anticipated duration for each milestone remained accurate. However, the development phase did take longer than anticipated and came in closer to four days / 32 hours. All other milestones and durations were consistent to the original project timeline outlined below.

Milestone	Projected Start Date	Projected End Date	Duration (days/hours)
Gather project requirements	02/01/2023	02/01/2023	1 day / 8 hours
Review requirements and develop workflow	02/02/2023	02/02/2023	1 day / 8 hours
Gather data, create test dataset	02/03/2023	02/03/2023	1 day / 8 hours
Python script development	02/06/2023	02/08/2023	3 days / 24 hours
Test Python application	02/09/2023	02/09/2023	1 day / 8 hours
Develop User Training and Maintenance Materials	02/10/2023	02/10/2023	1 day / 8 hours

Methodology

C. Data Collection Process

- Actual data selection vs. planned collection process

  Data collection for this project was conducted via a download from Kaggle.com. This approach was outlined in Task 2 and was carried out as planned, with no deviation. The data used for the completion of this project is representative of data that an actual financial institution would likely have available for analysis.

- Obstacles to data collection

  There were no obstacles present during the data collection process. Retrieving data from Kaggle.com was beneficial as the data is made publicly available and is intended for analytics projects. Kaggle has a good reputation in the analytics world and is recognized as a reliable source of training datasets. Generally, datasets receive a usability score from other users, and in this case the dataset retrieved for this project received a high score of 10, meaning the data has a high probability of being useful.

- Unplanned data governance handling

  There were no unplanned data governance issues while working with the dataset for this project. The dataset was retrieved from Kaggle.com and is publicly available and does not contain sensitive information. However, this analysis was geared toward the financial industry, specifically banking. If this analysis were to be performed for a real institution, data governance would be of paramount concern, as an actual dataset would likely contain sensitive customer and company information and would require the utmost security and concern when conducting an analysis. With these considerations in mind, a project of this nature would require considerable security and planning of precautions to safeguard data.

C1. Advantages and Limitations of Data Set

There were numerous advantages working with the dataset collected for this project. As mentioned above, the dataset is publicly available and contains no sensitive information that required usage approval, and the data was easily retrieved. The dataset also provided a good

example of what an actual banking institutions data would likely look like for a credit card product and contained many customer and product attributes that aided in the creation of customer segments and churn measurement. However, as with most datasets, there were some attributes that did not contain values and were marked as 'unreported'. This limitation of the dataset led to fewer records being available for review. While the number of records available in the dataset were adequate for this project, a larger dataset with a greater number of records would be very beneficial when conducting a real-world analysis of churn prediction and would likely yield greater insights.

D. Data Extraction and Preparation Processes

The tool used for data extraction and preparation for this project was Python via the Jupyter Notebook user interface. The dataset for this project was retrieved from Kaggle.com in a .csv file format. Python was an optimal tool for this project as it is compatible with many file types and provides numerous libraries for data ingestion, preparation, and analysis. The goal of this project was to identify churn rates amongst customer segments. To accomplish this, each customer record had to be reviewed based on age, income, and education to identify which customer segment they were to be placed within. In addition to reviewing each customer record for these attributes, another vital step was to identify the fitness of each entry, and account for any duplicate records if present and to ensure attributes that contained null values were accounted for. Python provided the necessary means to accomplish these steps through multiple functions available in the Pandas, NumPy and SciPy function libraries. By running the .info and. duplicated functions, I was able to determine there were no duplicated records or null values present within the dataset. The .info function was useful as it allowed me to identify the data formats for each attribute, as well as the total number of records present. These steps were vital to the preparation process as they ensured that records available for review, contained all necessary attribute information and each entry was a valid customer record that could be placed within a customer segment.

E. Data Analysis Process

E1. Data Analysis Methods

During the completion of this project the descriptive analytical method was employed. The descriptive analytical method was appropriate for this project as the goal was to gather further insights into the characteristics of customers who have churned and those who may be at risk of churning based on similar traits. Aggregation was also used as the data analysis technique, as the primary metric for evaluation in this project was churn rate. Churn rates requires a count of customers who are both active and inactive, as well as an aggregated total of all customers present within each segment and the overall dataset. Once these aggregations have been performed, churn rate can then be calculated by dividing the number of inactive customers by total customers. A key element of this project was to provide a count of customers who have churned, as well as the number of customers presented in each customer segment created, along with the churn rate each segment represented.

E2. Advantages and Limitations of Tools/Techniques

The primary tool used to complete this project was Python. Python is an optimal tool when reviewing large datasets as it is compatible with many file types and provides numerous libraries containing functions that quickly and accurately aggregate data for descriptive and statistical

analytics. Functions from the Pandas, NumPy, MatPlotLib and SciPy package libraries were utilized to complete this project. In addition to the many functions available through Python, it is open sourced, and compatible across platforms. However, while the open-source nature of Python makes it widely available, this can also be a limitation as some organizations may restrict use of Python due to security vulnerabilities associated with being an open-source language. Due to this, Python built applications may not be best suited to all organizations seeking to perform the type of analysis performed in this project.

The data analysis technique used in this project was aggregation and best suited the requirements for success. This technique allowed for the aggregation of customer counts by status and segment and aided in the calculation of churn rate. A clear limitation of this technique is that the measurements derived from its use, are only as strong as the dataset being used to produce the results. As noted in the previous section discussing data extraction and preparation, as with any dataset, there is often incomplete or duplicate entries present, these records can reduce the total number of entries available for review. This can lead to not enough data being available for aggregation or overly inflated metrics from a small dataset that do not accurately represent a trend. Due to this limitation, it is important to ensure there is an appropriate number of records to sample when completing an analysis of this nature.

E3. Application of Analytical Methods

There were three assumptions detailed in Task 2 that the project must address via its output to be deemed successful. These assumptions were:

1) Were all customers correctly identified in a segment?
    a. Considered successful if there are no outlier customers that have not been assigned to a segment.
2) Was churn percentage correctly captured for each customer segment?
    a. Considered successful if churn percentage from each customer segment correctly adds up to the observed total churn percentage.
3) Were all customers with an active status successfully assigned to a customer segment, and all customer segments are consistent based upon customer characteristics?
    a. Considered successful if all customers are correctly assigned to a customer segment, and segments are consistent based upon customer characteristics regardless of active status.

To meet these assumptions, the project used the methods and techniques outlined in sections E1 and E2 to complete the steps below:

- Determined range of age groups present in data to determine age factor in customer segmentation. Ages within the dataset ranged from 26 – 73 years old. The following age groups were determined:
    - Young Adult – 34 years old or younger
    - Middle Aged – 35 to 55 years old
    - Aged – 56 years old and older
- Determined unique levels of education present in data to create education tier used in determining customer segmentation. Education levels present in the data ranged from no formal education to doctorate level degrees. The following education tiers were determined:

        - No Formal Ed – No formal education
        - Primary Ed – K – 12 primary education
        - College Ed – Undergraduate or above
- Determined distinct range of incomes present in data to create income tier used in determining customer segmentation. Yearly incomes ranged from under $40k to over $120k. The following income tiers were determined:
    - Low - $40k and under
    - Mid - $40 - $80k
    - High - $80k and above
    - Other – Unreported income
- Created four distinct Customer Segments, along with 28 customer sub segments based on age, education, and yearly income. An outline of each customer segment and sub segment is provided below:
    - Segment A – Individuals aged 56 or older:
        - Aged, No Formal Ed, Low
        - Aged, No Formal Ed, Mid
        - Aged, No Formal Ed, High
        - Aged, Primary Ed, Low
        - Aged, Primary Ed, Mid
        - Aged, Primary Ed, High
        - Aged, College Ed, Low
        - Aged, College Ed, Mid
        - Aged, College Ed, High
    - Segment B – Individuals aged 35 to 55:
        - Middle-aged, No Formal Ed, Low
        - Middle-aged, No Formal Ed, Mid
        - Middle-aged, No Formal Ed, High
        - Middle-aged, Primary Ed, Low
        - Middle-aged, Primary Ed, Mid
        - Middle-aged, Primary Ed, High
        - Middle-aged, College Ed, High
        - Middle-aged, College Ed, Low
        - Middle-aged, College Ed, Mid
    - Segment C – Individuals aged 34 or younger:
        - Young Adult, No Formal Ed, Low
        - Young Adult, No Formal Ed, Mid
        - Young Adult, No Formal Ed, High
        - Young Adult, Primary Ed, Low
        - Young Adult, Primary Ed, Mid
        - Young Adult, Primary Ed, High
        - Young Adult, College Ed, Low
        - Young Adult, College Ed, Mid
        - Young Adult, College Ed, High
    - Segment Other – Included all customer records that were missing entries for either education or income attributes.
- Aggregated customer counts for active, churned, and total customers for each customer segment, and overall dataset population.
- Calculated churn rate for each customer segment and sub segment, and sub segment. (Total Churned Customers / Total Customers).

Results

F. Project Success

F1. Statistical Significance

This project sought to answer whether segmenting customers with similar traits, such as age, education and income would help identify if certain customer segments were more likely to churn than others. Specifically, I predicted that customers of an older age group, with higher educations and incomes would be more likely to churn than other segments. To evaluate if the project was successful, and to test if there was a statistically significant difference in churn rate among customer segments, I chose to perform a one-way ANOVA test.

The one-way ANOVA is a vigorous assessment, that uses an F test to measure the means of each observation group, as well as the mean of the overall population. If one group has a mean that significantly varies from the other observation groups or the overall population mean, the test allows for the rejection of a null hypothesis and indicates that there is no statistically significant difference between observation groups. The one-way ANOVA test produces two values, a F value, and a P value. These values can be used to determine if there is a statistically significant difference between groups. The F value is a measure of the variance between the groups and within the groups themselves. A high F value indicates a higher variance between the observation groups, and a lower value signifies less of a variance is present. The P value indicates the probability, or likelihood that there is a significant difference between groups. Unlike the F value, a lower P value, below 0.05 for this project, indicates a stronger variance between the observation groups, and allows the null hypothesis to be rejected. If the P value is above the 0.05 threshold, the null hypothesis cannot be rejected, and it can be stated that there is no statistically significant difference between the observation groups.

To perform a One-Way ANOVA, there are four assumptions that the data must meet:

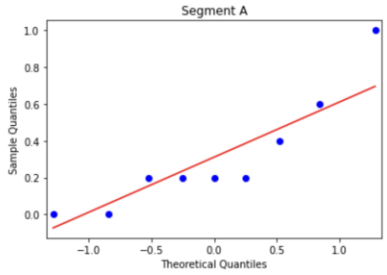
- The independent variable in each group must be unique to the group it is assigned to, and the observation groups cannot influence each other.
- The data selected for testing must be a random sample collected for each observation group.
- The values of the dependent variable must be normally distributed.
- There must be a homogeneity of variance among the groups.

These assumptions, along with the nature of the one-way ANOVA, made the test an ideal fit to measure the effectiveness of this project. The customer segmentation strategy this project implemented satisfied the first and second assumptions of the ANOVA test, as the segments created grouped all customers into a distinct segment based on the customers age, which prevented customers from being in two segments at once. Additionally, this approach meant that individual segments could not influence the churn rate of other segments, allowing each segment to be independent of one another and to be sampled without risking unintended interference.

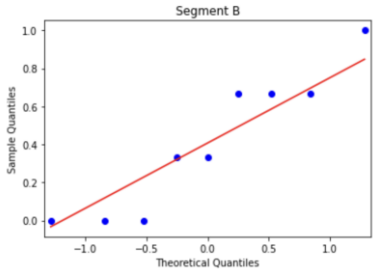
To satisfy the second assumption of the one-way ANOVA, that all datapoints for the dependent variable, churn rate, be normally distributed, I utilized the MinMaxScaler function from the SKLearn library in the Python. This function took the churn rate values from each of the nine sub segments within each customer segment and redistributed the values on a scale of zero to one, making for a more normalized distribution of churn rate, and satisfying the second assumption of

the one-way ANOVA. After normalizing the churn rate percentage for each segment, I then ran a Shapiro-Wilk test for normality for each segment to normal distribution. Like an ANOVA test, the Shapiro-Wilk test produces two values, a Stats value, and a P value. The Stats value is rank from zero to one, and the closer the stat value is to one, the more equally distributed the data is. The P value in the test is a measurement used to accept or reject the null hypothesis, that data is not evenly distributed. If the P value is above 0.05, then you can state the data is equally distributed. I utilized a QQ plot to visualize the results of the Shapiro-Wilk test for each segment, further validating normally distributed data, further satisfying the second assumption of the one-way ANOVA test requirements.

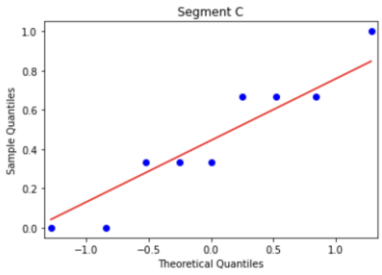
```
Segment A - Stats=0.837, p=0.054
Segment A data is normally distributed
```




```
Segment B - Stats=0.884, p=0.172
Segment B data is normally distributed
```




```
Segment C - Stats=0.917, p=0.364
Segment C data is normally distributed
```



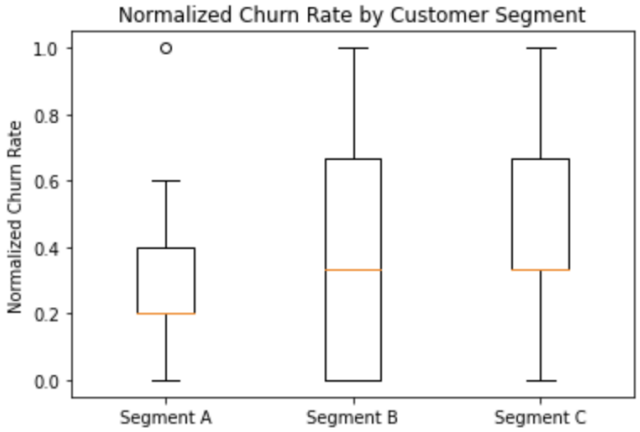
The last assumption of the one-way ANOVA is that there is a homogeneity of variance among the sample groups, or that the spread of values examined are equally distributed among the independent variable. To verify the segments satisfied this requirement, I performed a Levene test to evaluate the spread of churn among the customer segments. Like the Shapiro-Wilk test, the Levene test provides two values, a Stat value, and a P value. In this scenario, a high Stat score indicates that values are not evenly distributed. The P value further indicates whether the null hypothesis is to be accepted or rejected, with a value larger than 0.05 indicating that the null hypothesis can be rejected, and values are homogeneous.

```
Statistics=0.399, p=0.675
Variances are equal among segments
```

After completing all steps to test and verify the data met all assumptions required, I then performed a one-way ANOVA test. The test included three customer segments; each segment contained randomly sampled churn rates from nine sub segments of customers who fit into each customer segment. Segment A encompassed aged individuals, Segment B was middle aged customers, and Segment C contained young adults. The test returned low F stat value of 0.37 and a high P value of 0.69 (>=0.05). These two results indicate that there is little variation among customer segments when it comes to churn rates, and there is no statistically significant variation among the three customer segments.

```
Results from One-Way ANOVA Test:

No statistically significant difference between customer segments
One-Way ANOVA F Stat: 0.37079455977093756
One-Way ANOVA P Value: 0.6940708680960437
```



To further examine the results from the ANOVA, I implemented a Tukey HSD test. The Tukey HSD test can be used to verify results from a one-way ANOVA test and to further evaluate if there is a statistically significant variance among observations groups. The Tukey HSD test produces similar values as the ANOVA test, with a F Stat and a P value, as well as a more detailed view of variances among the groups observed. The Tukey HSD test results further validated that no statistically significant variances were present, returning the same values as the ANOVA for each test value.

```
Results from Tukeys HSD Test:

Tukey HSD F Stat: 0.37079455977093756
Tukey HSD P Value: 0.6940708680960437


   Multiple Comparison of Means - Tukey HSD, FWER=0.05
=====================================================
  group1    group2  meandiff p-adj   lower  upper  reject
-----------------------------------------------------
Segment_A Segment_B   0.0963 0.8074 -0.3028 0.4954  False
Segment_A Segment_C   0.1333 0.6752 -0.2658 0.5324  False
Segment_B Segment_C    0.037    0.9 -0.3621 0.4361  False
-----------------------------------------------------
```

With the combined results from the one-way ANOVA test and Tukey HSD test, my hypothesis that there would be a significant variance among customers segments, more specifically that individuals of an older age and with higher education and incomes would churn more, was rejected.

F2. Practical Significance

While the results of the statistical test did not prove my hypothesis, I believe the results of this project still offer significant insights to an organization when reviewing customer churn. While the test did not show a significant variance between customer segments, the project did uncover churn trends among sub segments of customers who have similar traits. For organizations that are seeking to better understand their customers, and how to prevent the loss of their business, it is vital for an institution to understand who its customers are, and what is important to each. To do this, businesses must utilize the data available to them to harness insights, and this tool will help accomplish that. The customer segmentation strategy utilized in this project provides a useful approach to doing this, as one customer record itself does not imply a trend on its own, but when coupled with other records, a segmentation strategy more easily identifies trends among individuals who share common characteristics. With these types of insights an organization can begin to understand where they need to shift their focus to improve their business strategy.

F3. Overall Success

Overall, I view this project as a complete success. While I would have preferred for my hypothesis to be proven correct, the original intent of this project was successfully implemented. The reporting solution runs quickly and accurately and produces insightful results. Each of the goals I stated at the beginning of the project were successfully implemented, with all customers being correctly assigned to a customer segment, along with the percentage of churn associated with each segment. In addition to successfully creating customer segments, customers were also assigned to sub segments to reveal an extra layer of insight when reviewing churn rates. As an analyst with experience across multiple industries, I have a keen understanding of how important, and challenging it can be to maintain a healthy customer base, and the amount of time this type of analysis demands. I believe a tool like this will help save an organization time and money through the insights it provides.

G. Key Takeaways

G1. Summary of Conclusions

This project sought to create a reporting solution that reveals insights into credit card customer churn rates, by implementing a customer segmentation strategy. The output of the solution was to create tables and graphs that can be used to understand the number of customers that are both active and churned, how many customers share similar traits and what the churn trends are among segments and sub segments of customers.

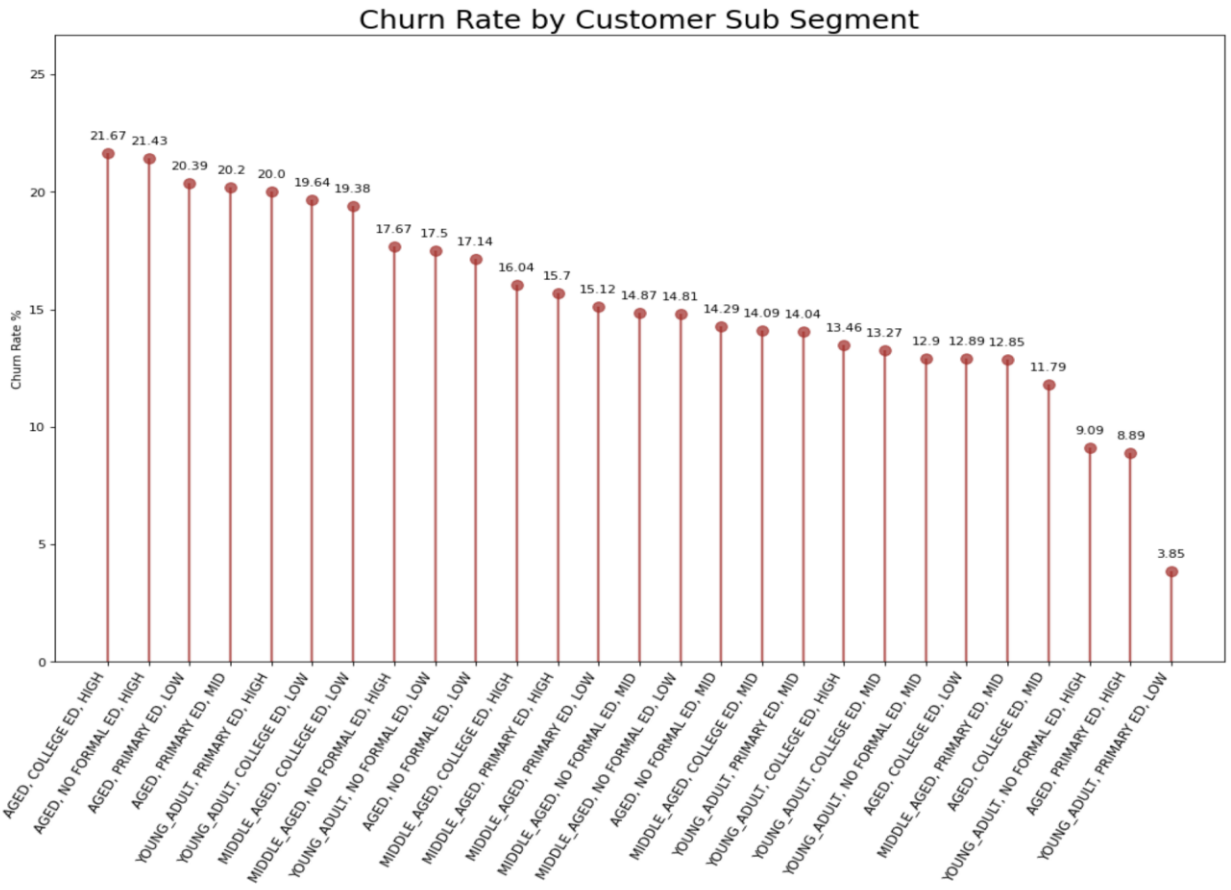
In the table produced below, we can see a summary of the total customers in the customer base, an active customer count, a churned customer count, as well as breakout by customer segment and the churn rates associated with each segment.

Customer_Segment	Segment Customer Count	Segment Customer Count - Active	Segment Customer Count - Churned	Segment Churn Rate	Customerbase Cusotmer Count - Total	Segment Pct of Customerbase	Customerbase Customer Count - Churned	Segment Pct of Total Churn Rate
Segment_A	1000	842	158	15.8%	10127	9.9%	1627	9.7%
Segment_B	6102	5129	973	15.9%	10127	60.3%	1627	59.8%
Segment_C	539	459	80	14.8%	10127	5.3%	1627	4.9%
Segment_Other	2486	2070	416	16.7%	10127	24.5%	1627	25.6%

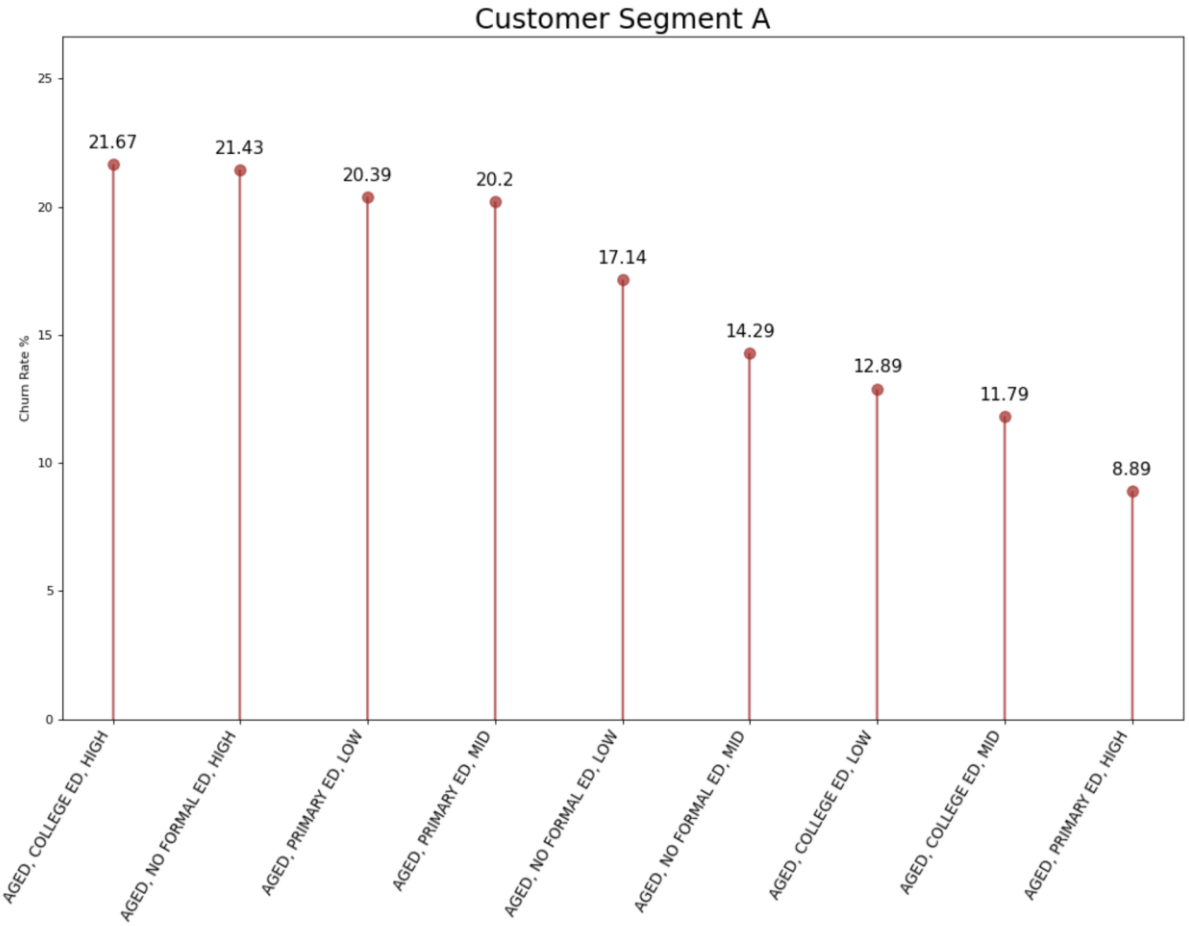
The table below contains a summarized view of the sub segments within each customer segment, as well as the age, education and income of the customers that are included in each sub segment.

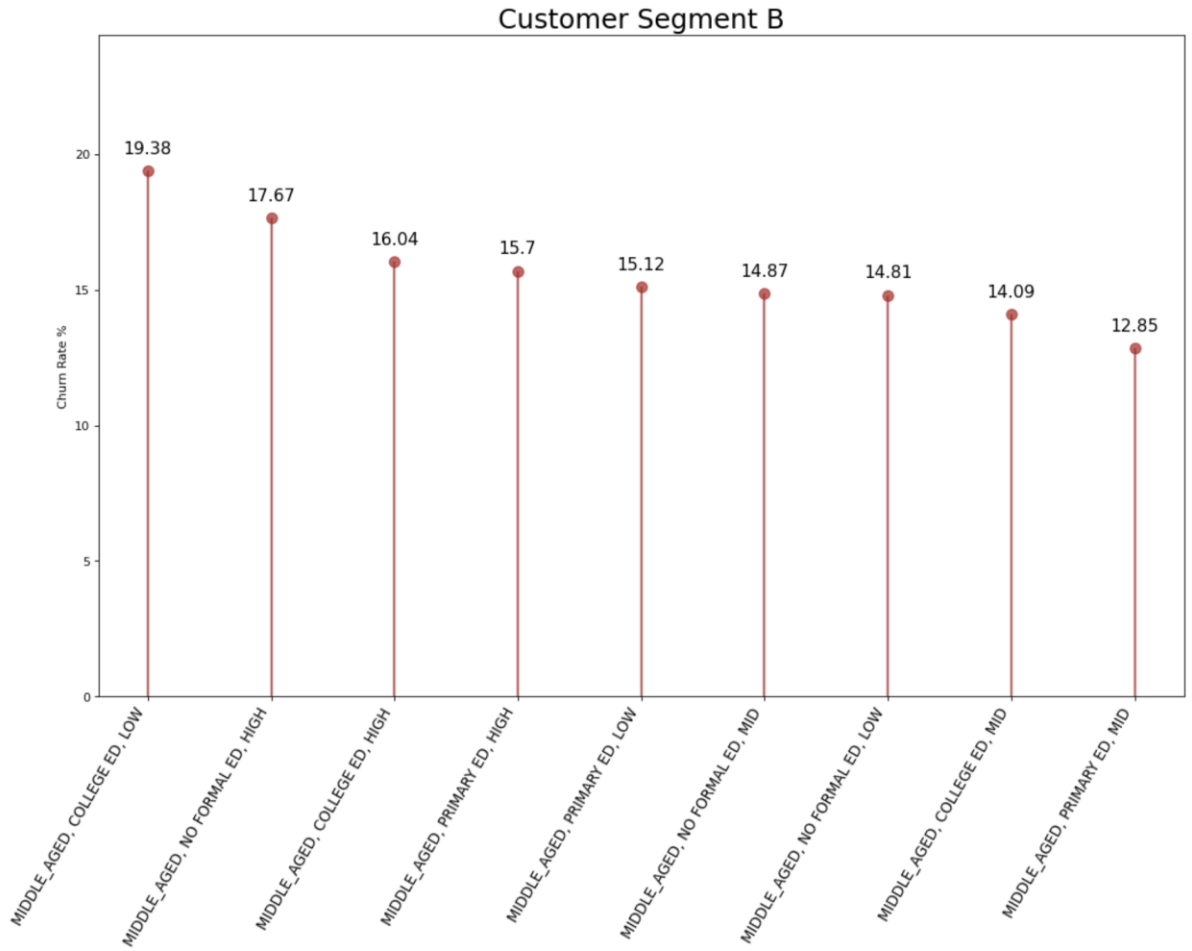
	Customer_Segment	Sub_Segment	Active_Customers	Churned_Customers	Total_Customers	Churn_Rate
0	Segment_A	Aged, College Ed, High	94	26	120	21.666667
1	Segment_A	Aged, College Ed, Low	223	33	256	12.890625
2	Segment_A	Aged, College Ed, Mid	172	23	195	11.794872
3	Segment_A	Aged, No Formal Ed, High	33	9	42	21.428571
4	Segment_A	Aged, No Formal Ed, Low	58	12	70	17.142857
5	Segment_A	Aged, No Formal Ed, Mid	60	10	70	14.285714
6	Segment_A	Aged, Primary Ed, High	41	4	45	8.888889
7	Segment_A	Aged, Primary Ed, Low	82	21	103	20.388350
8	Segment_A	Aged, Primary Ed, Mid	79	20	99	20.202020
9	Segment_B	Middle_Aged, College Ed, High	806	154	960	16.041667
10	Segment_B	Middle_Aged, College Ed, Low	1119	269	1388	19.380403
11	Segment_B	Middle_Aged, College Ed, Mid	1116	183	1299	14.087760
12	Segment_B	Middle_Aged, No Formal Ed, High	233	50	283	17.667845
13	Segment_B	Middle_Aged, No Formal Ed, Low	351	61	412	14.805825
14	Segment_B	Middle_Aged, No Formal Ed, Mid	292	51	343	14.868805
15	Segment_B	Middle_Aged, Primary Ed, High	333	62	395	15.696203
16	Segment_B	Middle_Aged, Primary Ed, Low	438	78	516	15.116279
17	Segment_B	Middle_Aged, Primary Ed, Mid	441	65	506	12.845850
18	Segment_C	Young_Adult, College Ed, High	45	7	52	13.461538
19	Segment_C	Young_Adult, College Ed, Low	135	33	168	19.642857
20	Segment_C	Young_Adult, College Ed, Mid	98	15	113	13.274336
21	Segment_C	Young_Adult, No Formal Ed, High	10	1	11	9.090909
22	Segment_C	Young_Adult, No Formal Ed, Low	33	7	40	17.500000
23	Segment_C	Young_Adult, No Formal Ed, Mid	27	4	31	12.903226
24	Segment_C	Young_Adult, Primary Ed, High	12	3	15	20.000000
25	Segment_C	Young_Adult, Primary Ed, Low	50	2	52	3.846154
26	Segment_C	Young_Adult, Primary Ed, Mid	49	8	57	14.035088
27	Segment_Other	Other	2070	416	2486	16.733709

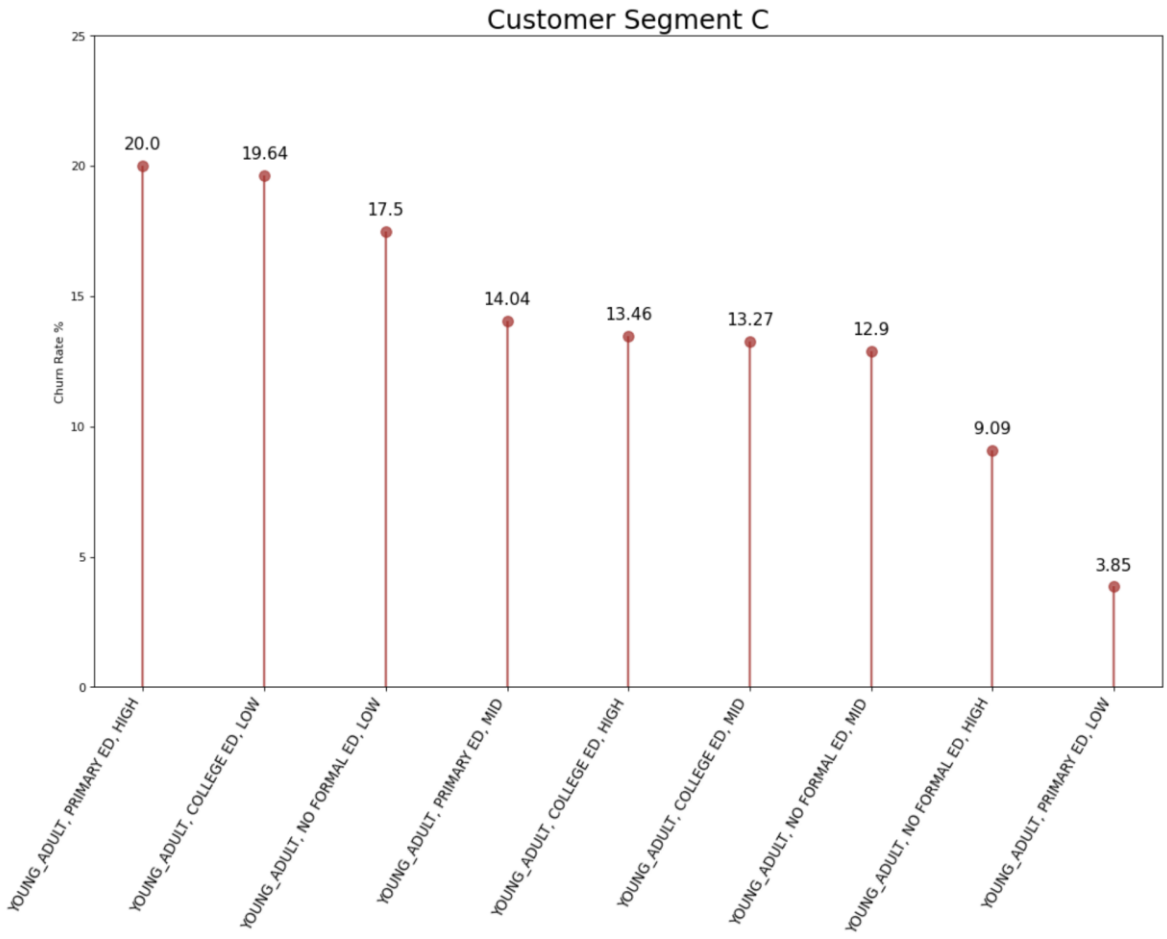
The next graph shows churn rate by sub segments of customers in descending order from the highest churn rate to the lowest churn rate. Of note, while my hypothesis was not accepted during the one-way ANOVA test, customers in the aged group, with college educations and high incomes, did represent the highest churn of all sub segments, although not by a statistically significant difference, as the following sub segments were close in churn rate as well.



The next three graphs present a similar view as the one above but focus on each customer segment individually for closer analysis.


Customer Segment A
Churn Rate %
0
5
10
15
20
25
21.67
21.43
20.39
20.2
17.14
14.29
12.89
11.79
8.89
AGED, COLLEGE ED, HIGH
AGED, NO FORMAL ED, HIGH
AGED, PRIMARY ED, LOW
AGED, PRIMARY ED, MID
AGED, NO FORMAL ED, LOW
AGED, NO FORMAL ED, MID
AGED, COLLEGE ED, LOW
AGED, COLLEGE ED, MID
AGED, PRIMARY ED, HIGH


Customer Segment B
Churn Rate %
0
5
10
15
20
19.38
17.67
16.04
15.7
15.12
14.87
14.81
14.09
12.85
MIDDLE_AGED, COLLEGE ED, LOW
MIDDLE_AGED, NO FORMAL ED, HIGH
MIDDLE_AGED, COLLEGE ED, HIGH
MIDDLE_AGED, PRIMARY ED, HIGH
MIDDLE_AGED, PRIMARY ED, LOW
MIDDLE_AGED, NO FORMAL ED, MID
MIDDLE_AGED, NO FORMAL ED, LOW
MIDDLE_AGED, COLLEGE ED, MID
MIDDLE_AGED, PRIMARY ED, MID



These tables and graphs show that all customers were successfully placed in a segment, and churn rate was captured for each segment and sub segment of customers, demonstrating that all goals stated in section B1 of this report were met, and the project was successfully completed.

G2. Effective Storytelling

The tools and visualizations chose for this project represent the best methods for displaying the insights and trends uncovered during the completion of this project. The tables included in the output of the reporting solution display all metrics that are relevant when reporting churn rates and provide insights into the customer base. The bar graphs included help convey churn rates among different segments and sub segments by allowing a user to see how each of the segments compare to one another. This type of visual makes for a quick and easy way for users to identify which customer groups have the highest churn, and visually communicates the trends present within the data in a way a simple table cannot.

G3. Findings-based Recommendations

The goal of this project was to create a reporting solution that JGA Bank (fictional), or any banking institution offering credit card services could use to gather insights into their customer base and churn trends. In the case of JGA Bank, whose dataset was used to create this project, there are two action items the bank should address based on the findings of this project:

1) To reduce churn, focus on high-risk customer segments.

    a. Customers who are aged 56 years old or older, with college educations and incomes of $80k annually, represent the highest levels of churn among all customers who are in the same age group.
    b. Middle aged consumers, who are between the ages of 35 to 55 years old, with primary educations and incomes of $40k annually or less, represent the highest percentage of churn among the same age group.
    c. Young adults, who are 34 years old or younger, with primary educations and high incomes of $80k annually or more, represent the highest percentage of churn among the same age group.

2) Ensure customer records are populated with all relevant information.

    a. There were 2,486 customer records or 25% of the customer population that were missing entries for the education and income attributes that were placed in the 'Other' segment and could not be properly assessed.
    b. The bank should review these entries and create procedures to ensure all relevant customer attributes are captured at the time of on-boarding for new customers.
    c. The bank should review and the update data where possible and then complete another pass of this analysis to examine if churn trends are consistent or change, to better implement a churn strategy.

H. Panopto Presentation

Please visit the link below to view a presentation of my project in Panopto.

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=e309d3b6-d3d9-4db5-acda-b00a011edb8b

Appendices

I. Evidence of Completion

Please see the following articles for evidence of project completion:

- JGA Bank Customer Segmentation and Churn Analytics Tool – Jupyter Notebook / Python Code - pdf format for upload.
- JGA Bank Customer Segmentation and Churn Analysis - Hypothesis Test - Jupyter Notebook / Python Code - pdf format for upload.
- BankChurners.csv – source dataset, saved in .xlsx format for upload.

Sources

OpenStaxCollege. (2013, July 19). *F Distribution and One-Way ANOVA*. Introductory Statistics. https://pressbooks-dev.oer.hawaii.edu/introductorystatistics/chapter/one-way-anova/