Bitathon 2024

Data Creed

Urjith Reddy
Aditya Ganti
Rahul Tekkali

Problem Definition / Analytics Objective

Objective:

To harness the potential of the integrated datasets—Call Center, Customer Transactions, and Geography Lookup—to provide a comprehensive analysis of customer behavior, transaction patterns, and geographical trends within the telecom sector. By merging these datasets, we aim to deliver both retrospective insights and future forecasts that will drive strategic decision-making.

Problem Statement:

The telecom industry faces multifaceted challenges, including optimizing customer service, enhancing transaction efficiency, and leveraging geographical data for targeted marketing and network expansion. Our analysis seeks to address these challenges by drawing on the rich information contained in the three datasets.

Scope of Analysis:

Call Center Interactions: Analyze call center data to understand customer queries, complaints, and feedback patterns. Identify areas for improving customer service and reducing churn.

Customer Transactions: Dive into transaction data to explore purchasing behaviors, payment trends, and service usage. This will help in tailoring product offerings and improving customer satisfaction.

Geographical Insights: Utilize geographical data to examine customer distribution, regional service demands, and potential areas for network improvement or expansion.

Data Utilization: By integrating the Call Center, Customer Transactions, and Geography Lookup databases, we have created a unified database that provides a deep view of our customers. This integration allows for an analysis that covers customer service interactions, transaction behaviors, and geographical trends.

Approach:

Data Cleaning and Preprocessing:

Our initial focus is on meticulously analyzing the datasets to uncover potential relationships, which necessitates a thorough cleaning process. This involves meticulously preparing each dataset for integration, ensuring they are devoid of inaccuracies and inconsistencies. Identifying the primary and foreign keys plays a crucial role, as these keys are instrumental in accurately joining the datasets. This foundational step sets the stage for an in-depth data analysis, followed by comprehensive data visualization efforts.

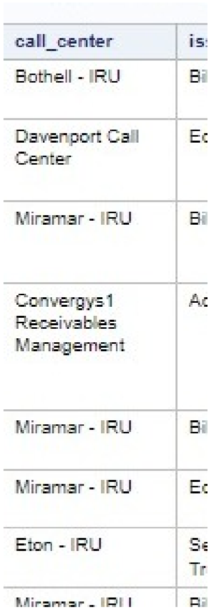
By adhering to this structured approach, we aim to seamlessly integrate the datasets, providing a unified and enriched foundation for deriving insightful analyses and visual representations that highlight key trends, patterns, and anomalies within the data. This meticulous preparation not only enhances the quality and reliability of our findings but also ensures that our analysis is grounded in a robust and coherent dataset, ready for the subsequent stages of exploration and interpretation.

Data Exploration:

Processing for cities:

From the three datasets provided, we began by meticulously analyzing each column to understand the relationships among them. Following this, we embarked on cleaning the data to rectify any errors, spelling mistakes, and null values. During our examination of the call center data, we identified a primary key, Customer_ID, and its relationship with a foreign key, call_center, which links to the Geography Lookup.

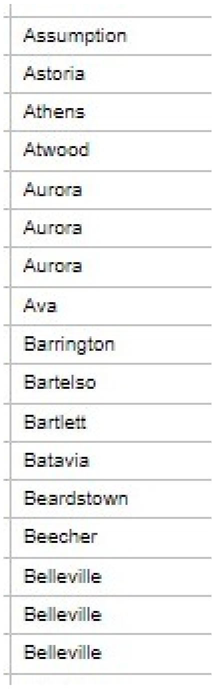
We noticed that the call_center foreign key in the Call Center data and the city column in the Geography Lookup were not processed correctly. Although the Geography Lookup contained all cities, some did not have associated call centers. Additionally, in the Call Center data, there were entries that were not cities, such as "Convergys1 Receivables Management". Furthermore, the city names in the call_center column were inconsistently formatted, with entries like "Bothell - IRU", where "- IRU" was extraneous, and "Atwater Call Center", where "Call Center" was unnecessary.



call_center	is
Bothell - IRU	Bi
Davenport Call Center	Ec
Miramar - IRU	Bi
Convergys1 Receivables Management	Ac
Miramar - IRU	Bi
Miramar - IRU	Ec
Eton - IRU	Se Tr
Miramar - IRU	Bi

To address these issues, we undertook a three-step data processing approach:

Removing Irrelevant Cities: We eliminated cities from the Geography Lookup that lacked corresponding call centers, focusing solely on locations with call center operations. This step ensured our analysis would only consider relevant geographical areas.



Assumption
Astoria
Athens
Atwood
Aurora
Aurora
Aurora
Ava
Barrington
Bartelso
Bartlett
Batavia
Beardstown
Beecher
Belleville
Belleville
Belleville

Consolidating City Entries: We noticed multiple entries for the same city with different zip codes in the Geography Lookup. To streamline our data, we retained only one entry per city, choosing a single zip code for each, thereby eliminating redundancy and simplifying the geographical data for analysis.

Data Cleaning and Integration: Once we had a definitive list of cities in the Geography Lookup, we proceeded to clean the Call Center data. This involved standardizing city names by removing unnecessary suffixes and labels, thus ensuring consistency across our datasets. Following this, we joined the datasets, creating a unified database that facilitated comprehensive analysis.

Code Explanation:

(1) Our first Python script focused on filtering out cities from the Geography Lookup that did not have matching call centers. By comparing city names with call center entries, we identified and retained only those cities with an existing call center presence.
(2) The second script was designed to remove duplicate city entries from the Geography Lookup, each associated with different zip codes. By keeping only one zipcode per city, we ensured that each city was uniquely represented, simplifying our geographical analysis.
(3) In the final step, we refined the Call Center data by correcting city names, removing irrelevant parts of the entries, and aligning them with the standardized city names from the Geography Lookup. This cleaning process was crucial for accurately merging the call center data with geographical information, enabling a seamless integration of the datasets for our analysis.

Processing for verbatims in call centre data:

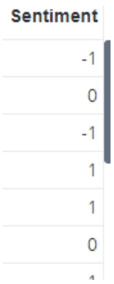
To analyze customer feedback from the call center data, we employed a Python script that utilizes the VADER Sentiment Analysis tool. This approach enables us to gauge the sentiment of customer verbatims, effectively categorizing them into positive, neutral, or negative sentiments based on their content or numeric ratings. Here's a breakdown of the process:

(4)Python Script Breakdown:
Import Libraries:

We start by importing pandas for data manipulation and SentimentIntensityAnalyzer from vaderSentiment for sentiment analysis.
Load Data:

The script loads the call center data from an Excel file (CC.xlsx), which contains customer feedback.
Sentiment Analysis Setup:

We initialize the VADER sentiment analyzer, a tool adept at understanding the nuances of sentiment expressed in textual feedback.
Classify Integer Sentiment:



Sentiment
-1
0
-1
1
1
0

A function, classify_integer_sentiment, is defined to classify feedback given as integer values. This classification is based on a predefined scale where ratings from 0 to 4 indicate negative sentiment, 5 to 6 signify neutral sentiment, and 7 to 10 represent positive sentiment.

Analyze Textual Sentiment:

For textual feedback, the analyze_sentiment function employs the VADER analyzer to determine sentiment scores. Feedback with a compound score above 0.05 is considered positive, below -0.05 is negative, and scores within this range are deemed neutral.
Applying Sentiment Analysis:

We apply the analyze_sentiment function to each piece of feedback in the dataset, stored in the 'translated verbatim' column. This process assigns a sentiment category to each entry, effectively summarizing the emotional tone of the feedback.
Exporting Results:

The updated dataset, now including sentiment classifications, is saved to a new Excel file (Updated_CC_Data_with_Sentiment.xlsx). This file facilitates further analysis and reporting on customer sentiment trends.
Additional Points:
Versatility: This script demonstrates versatility by handling both numeric and textual feedback, ensuring comprehensive sentiment analysis across different types of customer verbatims.
Automation: Automating sentiment analysis streamlines the process of understanding customer sentiment, enabling quick and effective decision-making based on customer feedback trends.
Insight Generation: By classifying customer feedback into sentiment categories, we can generate actionable insights into customer satisfaction, areas of concern, and overall service quality.

Processing for resolution column in call center data:

In processing the resolution column within the call center data, we embarked on a structured approach to categorize the resolutions of customer inquiries and issues, further enhancing our analysis with sentiment assessment. This multifaceted process involved several key steps:(5)

Step 1: Importing the Data

Initially, we imported the call center data from an Excel file, ensuring we correctly specified the path and sheet name to access the relevant data accurately.

Step 2: Categorizing Resolutions

Our next step focused on categorizing the resolutions found within the call center data. By examining the resolution descriptions, we identified keywords indicative of positive outcomes (such as 'RESOLVED' and 'SATISFIED'), pending issues (like 'UNABLE' and 'PENDING'), and other uncategorized outcomes. This categorization allowed us to better understand the effectiveness of the call center's problem-solving efforts.

Step 3: Counting Categories

Utilizing SAS procedures, we then counted the occurrences of each resolution category. This quantitative analysis provided us with a clear picture of how many customer issues were resolved satisfactorily, remained pending, or fell into other categories.

Step 4: Visualizing the Data

To effectively communicate our findings, we created a pie chart visualizing the distribution of resolution categories. This visual representation highlighted the proportions of resolved, pending, and other types of resolutions, offering valuable insights into the call center's operational efficiency.

Analysing the data:

This report presents a comprehensive analysis of customer churn and feedback across various call centers, employing a blend of Python and SAS for data manipulation and analysis, coupled with PowerBI for data visualization. Our methodology focused on predictive modeling for churn analysis, sentiment analysis of call center feedback, transaction pattern analysis, and geographical trend analysis. The insights garnered from this analysis provided a multifaceted view of customer satisfaction, purchasing behaviors, and service demand distribution, which are critical for strategic decision-making.

Analysis Methodology:

Predictive Modeling for Churn Analysis: Utilizing a dataset compiled from multiple sources, we applied logistic regression and other machine learning techniques to identify key predictors of customer churn. This enabled us to forecast potential churn rates and understand the underlying factors contributing to customer attrition.

Sentiment Analysis of Call Center Feedback: Advanced sentiment analysis techniques were deployed to classify customer feedback into positive, neutral, and negative sentiments. This analysis was instrumental in gauging customer satisfaction levels and pinpointing areas requiring service enhancements.

Transaction Pattern Analysis: Through statistical and machine learning models, we examined customer transaction data to uncover trends in purchasing behaviors, preferred payment methods, and service usage patterns. This analysis aimed to reveal customer preferences and behaviors, providing insights into market demands and customer needs.

Geographical Trend Analysis: We leveraged geographical information systems (GIS) and spatial analysis to analyze customer distribution and regional service demands. This approach helped identify high-demand regions, underserved areas, and potential opportunities for network expansion or service improvement.

Data Analysis Tools:

Python: Served as the primary tool for data preprocessing, sentiment analysis, and visualization of call center feedback categorization. Python's versatile libraries enabled efficient analysis of feedback types across different call centers, facilitating the identification of centers with predominantly positive or negative feedback. Libraries used included Seaborn, Matplotlib, Rake, and Wordcloud.

SAS: Played a crucial role in analyzing resolution quality provided by call centers, including the timeliness and effectiveness of issue resolutions. SAS's analytical capabilities were also harnessed for generating pie charts and other visualizations to represent the analysis of resolution attributes within the dataset.

PowerBI: Was predominantly used for the visualization of analysis results, offering interactive and dynamic reporting capabilities. Although PowerBI's role was primarily in visualization, it also contributed to some aspects of data analysis, enhancing the comprehensiveness of our insights.

Visualizations and Reporting:

The culmination of our analysis was brought to life through detailed visualizations in PowerBI, supplemented by Python and SAS-generated charts. Visual representations included pie charts for resolution analysis, sentiment distribution graphs, trend analyses, and geographical heat maps. These visualizations provided intuitive access to complex data insights, facilitating easier interpretation and decision-making.

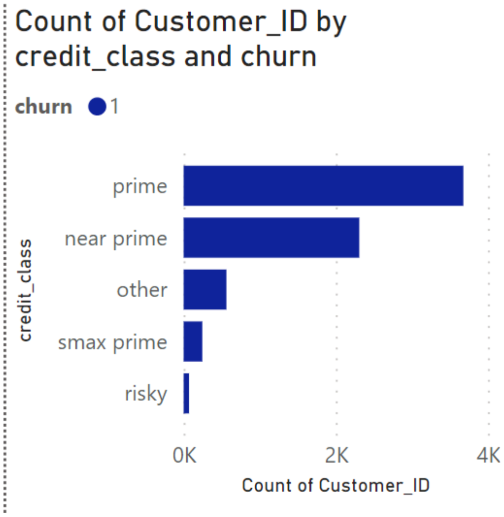
PowerBI was used primarily for data visualisation and a variety of features were utilised in the visualisation including stacked bar and column charts, pie charts, donut charts, word clouds, and many more. These represent a wide range of data representation. Built in PowerBI features like data transformation, outlier detection, were also utilised. This ensured that the data was properly and concisely represented and there is no problem for the viewer to understand the visuals. We ensured that the visualizations remained as uncluttered and clean as possible to make sure that they would be easily understood by the user and did not make them complex

Conclusion:

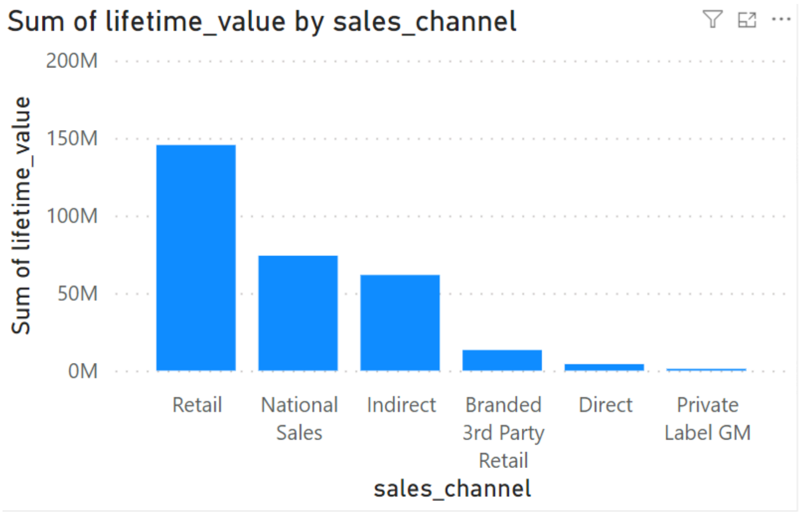
The integrated use of Python, SAS, and PowerBI enabled a deep dive into customer churn, feedback analysis, and purchasing behaviors, illuminating the path towards improved customer satisfaction and strategic growth opportunities. The methodologies employed not only provided a snapshot of current customer engagement levels but also offered predictive insights that can guide future business strategies.

Results and Conclusions:

We have found many interesting things after analyzing the data. Few key points are

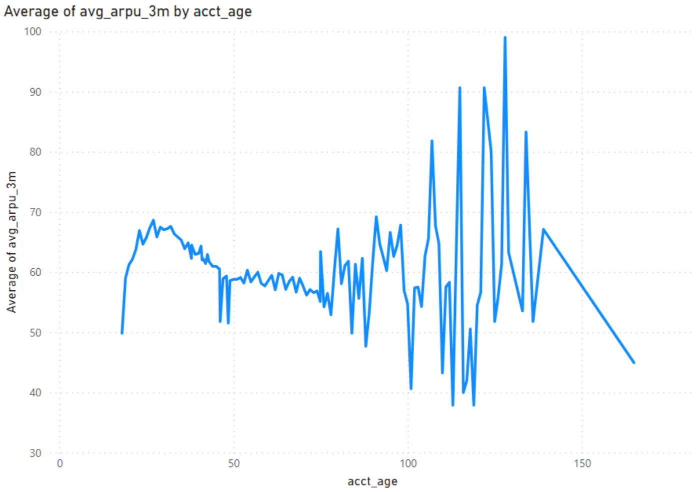


The data visualization indicates that among customers who have churned, the majority belong to the 'prime' credit class category, followed by a smaller, yet significant, proportion in the 'near prime' category. The 'other', 'smax prime', and 'risky' categories contribute to a lesser extent, with very few churns observed in the 'risky' category. This suggests that while customers with good credit are more likely to churn, the risk of churn is not confined to any single credit category, and even those with lower credit ratings are churning. This insight could be pivotal for developing targeted customer retention strategies across different credit segments.

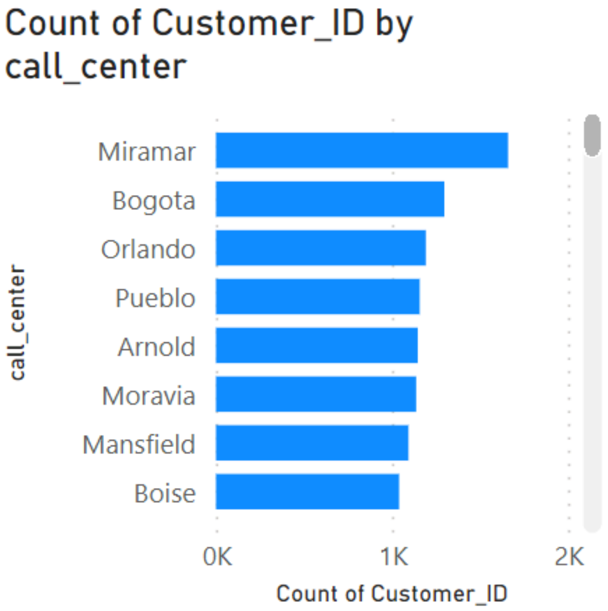


The bar chart showcases the sum of lifetime value categorized by different sales channels. Retail clearly leads in generating lifetime value, significantly outperforming other channels. National Sales follows as the second most valuable channel, with Indirect sales trailing closely behind.

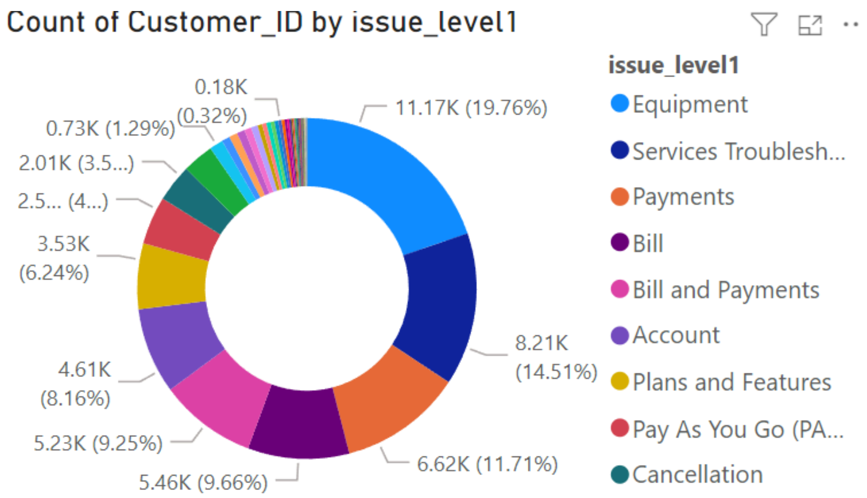
Branded 3rd Party Retail and Direct channels contribute relatively less to the lifetime value, while Private Label GM appears to have the least impact. This suggests that while Retail and National Sales are the strongest contributors to customer lifetime value.



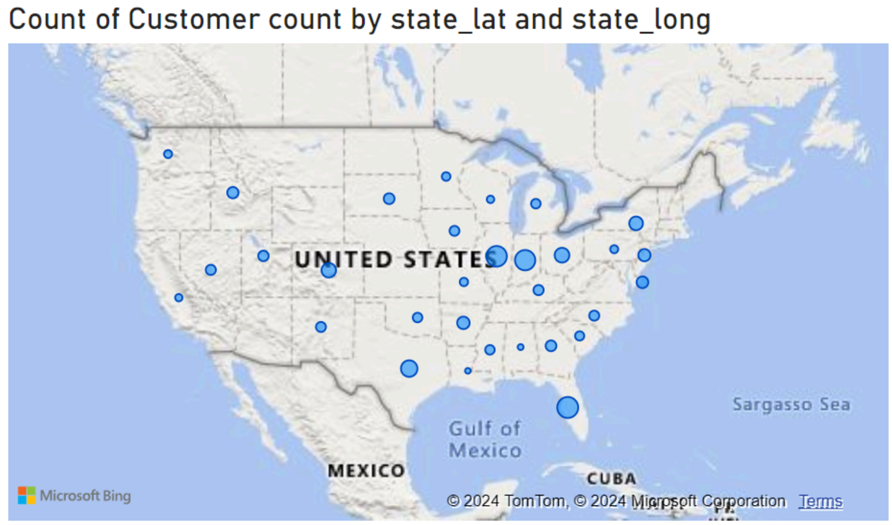
The graph shows the average of the average revenue per user (ARPU) over 3 months (avg_arpu_3m) against the account age. There appears to be a high level of variability in ARPU across different account ages. Notably, there is a significant peak in ARPU for accounts in the middle age range, which then decreases for older accounts. This could indicate that customers may initially use services more or opt for higher-value plans, but this engagement may taper off as the account ages. It suggests the need for strategies to maintain or boost revenue from long-standing customers.



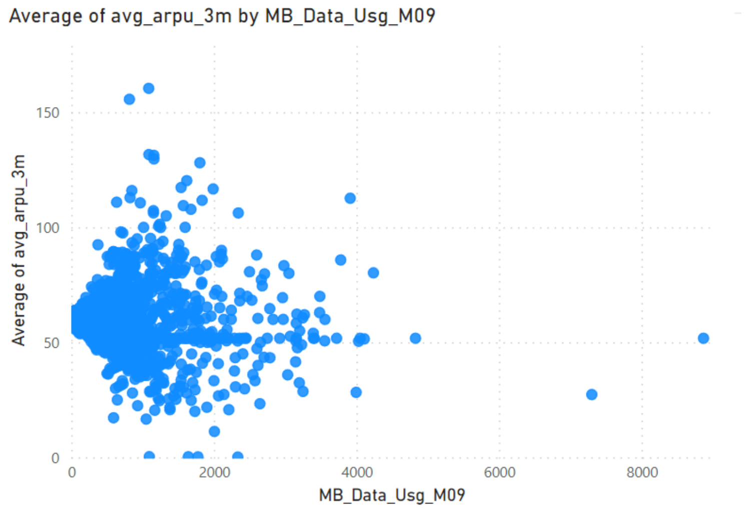
The bar chart illustrates the distribution of Customer_IDs across various call centers. Miramar has the highest count, indicating it serves the most customers, followed by Bogota. Orlando, Pueblo, and Arnold have a moderate number of customers, while Moravia, Mansfield, and Boise have the lowest counts. This distribution suggests that resource allocation, staffing, and customer service strategies might need to be tailored according to the customer base size at each location. It could also indicate market penetration or the success of each call center in customer acquisition.



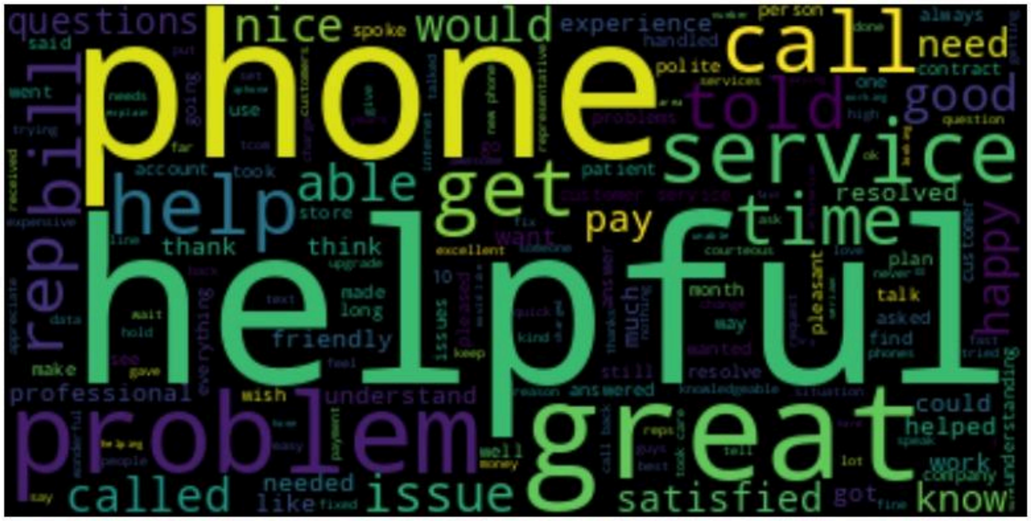
The donut chart presents the count of Customer_IDs by various issue categories. The most significant proportion of issues pertains to 'Equipment', followed by 'Services Troubleshooting' and 'Account' related issues. 'Payments', 'Bill', and 'Bill and Payments' also represent notable segments, while 'Plans and Features', 'Pay As You Go (PAYG) Adjustments', and 'Cancellation' account for smaller portions. 'Equipment' issues being the most common could indicate potential areas for product improvement or customer support focus. The chart provides a visual distribution of customer issues, which is valuable for prioritizing customer service and support initiatives.


Count of Customer count by state_lat and state_long
UNITED STATES
Sargasso Sea
Gulf of Mexico
MEXICO
CUBA
Microsoft Bing
© 2024 TomTom, © 2024 Microsoft Corporation Terms

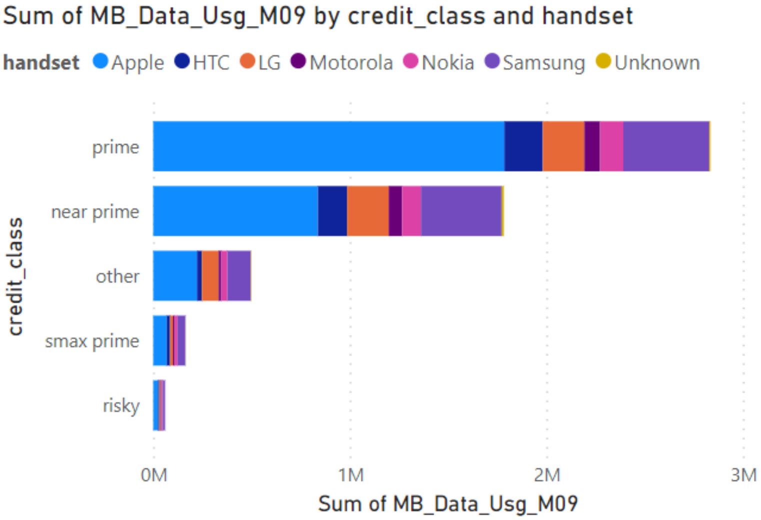
The map visualization displays the geographical distribution of customers across the United States, with circles of varying sizes indicating the count of customers in each location. Larger circles represent higher concentrations of customers, predominantly seen in states along the East and West Coasts as well as in some central regions. This spread suggests a diverse customer base and may indicate regions where the business has a stronger presence, potentially guiding strategic decisions for market expansion, targeted marketing, and resource allocation.



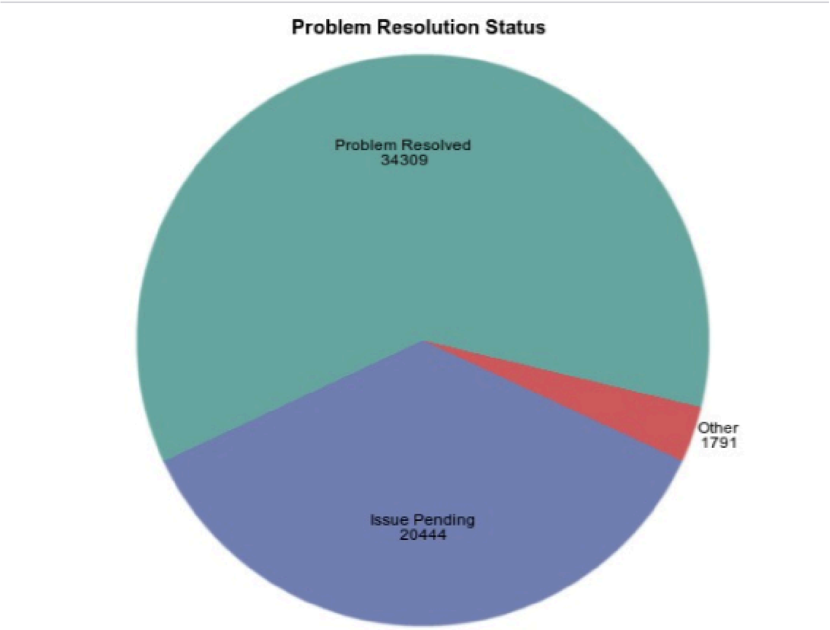
The scatter plot depicts the relationship between the average revenue per user (avg_arpu_3m) and mobile data usage (MB_Data_Usg_M09). It shows a wide dispersion of data points, suggesting a varied ARPU across different levels of data usage. While there's a concentration of users with lower data usage and ARPU, there are also outliers with high data usage but not proportionally high ARPU. This could indicate that higher data usage does not necessarily translate to a higher average revenue, which might prompt a review of data pricing strategies or the need to investigate the data plans associated with high data usage customers.



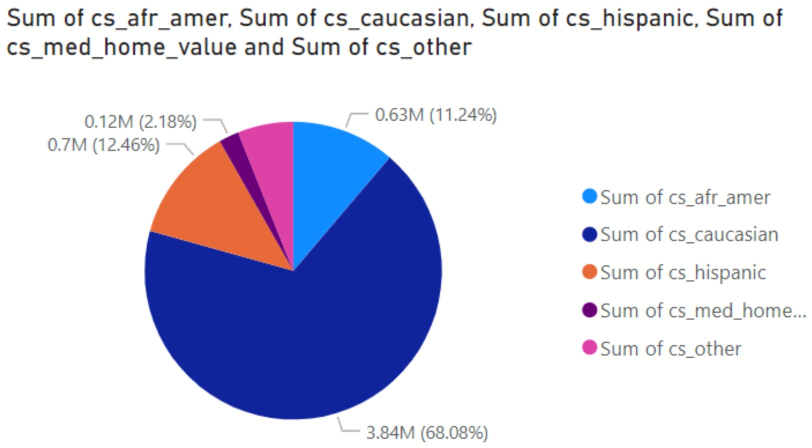
The word cloud is a visual representation of the most frequent words found in customer feedback related to phone service. The prominence of words such as "helpful," "service," "great," and "phone" suggests positive customer sentiment. "Helpful" and "service" being the most dominant words may indicate that customers highly value assistance and overall service quality. The presence of words like "problem" and "issue," albeit smaller, does indicate the occurrence of customer concerns, but the overall larger and central placement of positive terms like "great" and "satisfied" suggests a generally positive reception of the service provided.



The stacked bar chart illustrates the sum of mobile data usage (MB_Data_Usg_M09) across different credit classes and handset brands. Users in the 'prime' credit class category exhibit the highest data usage, with Apple and Samsung handsets being the most used among them. The 'near prime' and 'other' categories show more varied handset usage but lower overall data consumption. Users in the 'smax prime' and 'risky' credit categories have the least data usage, with a relatively small representation across handset brands. This could inform targeted marketing strategies and product offerings based on customer credit class and preferred handsets.

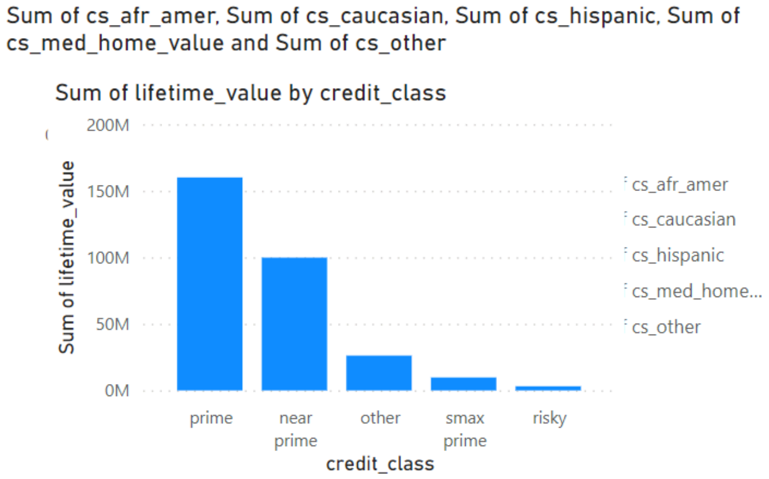


The pie chart titled "Problem Resolution Status" displays the outcomes of customer service issues. The largest portion shows that the majority of problems were resolved (34,309 cases), indicating effective resolution processes. A significant number, however, remains pending (20,444 cases), suggesting areas for improvement in responsiveness or resolution efficiency. A smaller segment is categorized as 'Other' (1,791 cases), which may include unresolved or differently categorized issues. This data is crucial for evaluating and enhancing customer service strategies.



Sum of cs_afr_amer, Sum of cs_caucasian, Sum of cs_hispanic, Sum of cs_med_home_value and Sum of cs_other

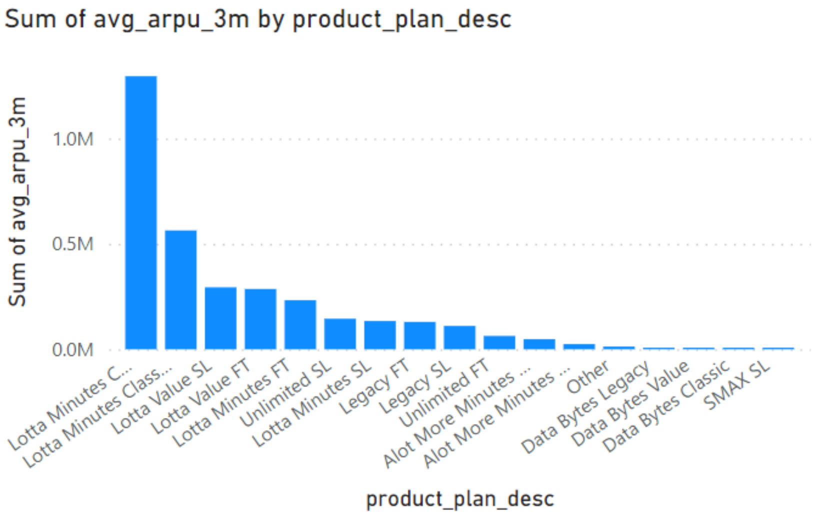
The pie chart represents a demographic segmentation based on customer data, with the largest segment being the 'Sum of cs_caucasian' at 68.08%. It is followed by 'Sum of cs_med_home_value', indicating a medium home value demographic, which makes up 12.46%. The 'Sum of cs_afr_amer' and 'Sum of cs_hispanic' represent smaller portions at 12.46% and 11.24% respectively, and the smallest segment is 'Sum of cs_other' at 2.18%. This demographic breakdown can be crucial for understanding customer diversity and tailoring services to meet the needs of different segments.



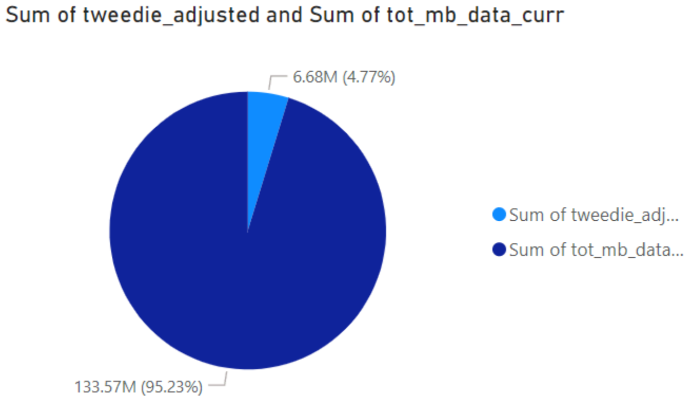
Sum of cs_afr_amer, Sum of cs_caucasian, Sum of cs_hispanic, Sum of cs_med_home_value and Sum of cs_other

Sum of lifetime_value by credit_class

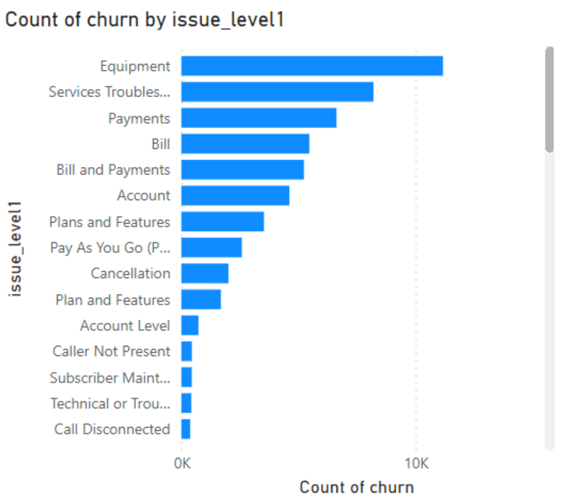
The bar chart displays the sum of lifetime value segmented by credit class. The 'prime' credit class contributes the highest to lifetime value, significantly more than other categories. The 'near prime' follows, but with less than half the contribution of the 'prime' segment. The 'other', 'smax prime', and 'risky' categories show relatively minor contributions in comparison. This distribution highlights the value of maintaining good credit standing customers and may inform credit management and targeted marketing strategies.



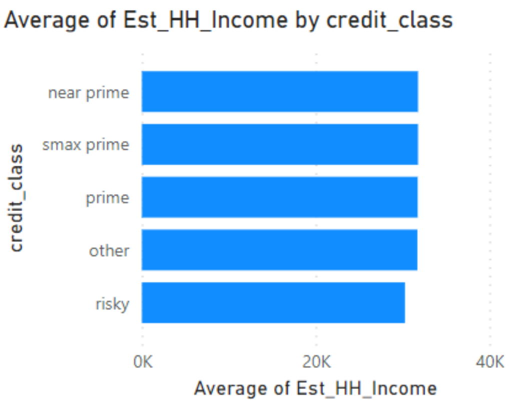
The bar chart illustrates the sum of average revenue per user (ARPU) over three months, broken down by different product plans. The 'Lotta Minutes' plan stands out with the highest ARPU, indicating it is the most revenue-generating option for the company. The subsequent plans, including 'Lotta Value' and various unlimited options, contribute progressively less to the total ARPU. The chart effectively highlights which plans are currently the most profitable and may guide the company in focusing their sales and marketing strategies to enhance revenue.



The pie chart displays two data components: the sum of 'tweedie_adjusted' and the sum of 'tot_mb_data_curr'. The vast majority, 95.23%, is attributed to the sum of 'tot_mb_data_curr', which could represent current mobile data usage or a similar metric. In contrast, the 'tweedie_adjusted' sum is a much smaller slice at 4.77%. This suggests that the 'tweedie_adjusted' value is a minor adjustment or component in comparison to the total mobile data metric, possibly indicating an adjustment factor or a specific subset of the data.

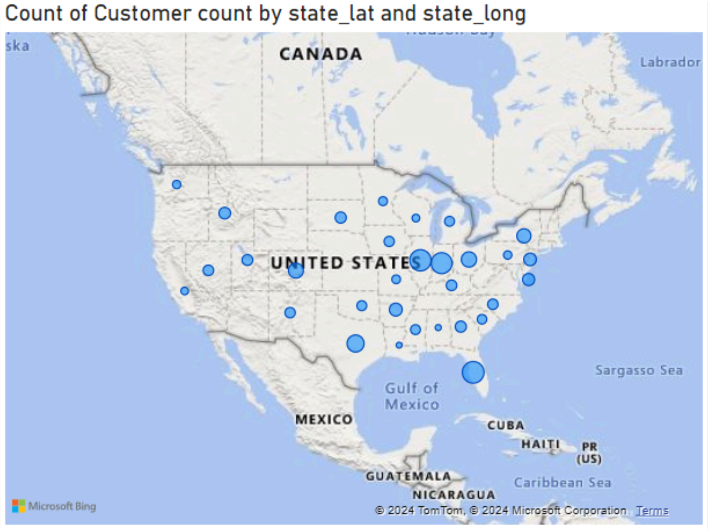


The bar chart represents the count of customer churn categorized by various issues. 'Equipment' issues lead to the highest churn, indicating significant customer dissatisfaction in this area. 'Services Troubleshooting' and 'Payments' also have high churn counts, suggesting these are critical factors influencing customer retention. 'Bill' and 'Bill and Payments' issues follow closely, while 'Account' and 'Plans and Features' represent a moderate churn risk. Lower on the list, issues like 'Account Level' and 'Technical or Troubleshooting' indicate fewer instances of churn. This data can inform targeted improvements in customer service and product offerings to reduce churn.

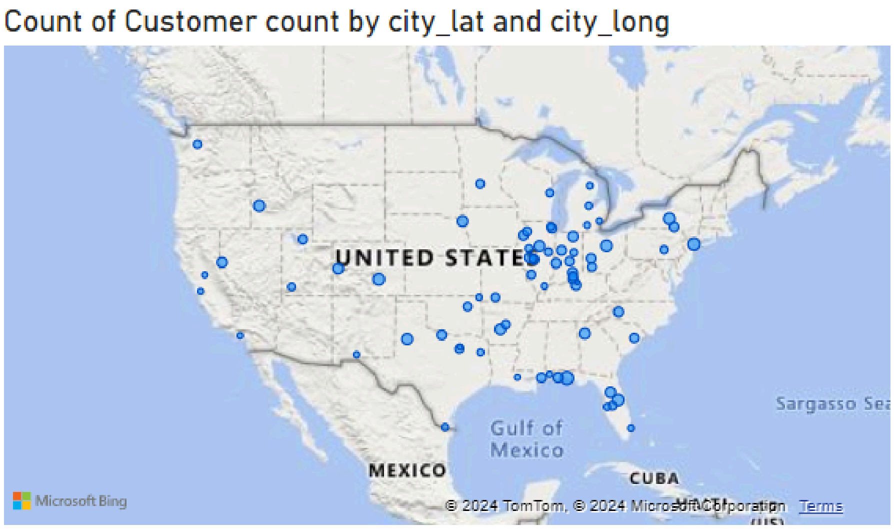


The bar chart depicts the average estimated household income segmented by credit class. It shows that customers in the 'near prime' category have the highest average household income, followed closely by those in the 'smax prime' and 'prime' categories. The 'other' category has a slightly lower average income, and the 'risky' category has the lowest of the groups presented.

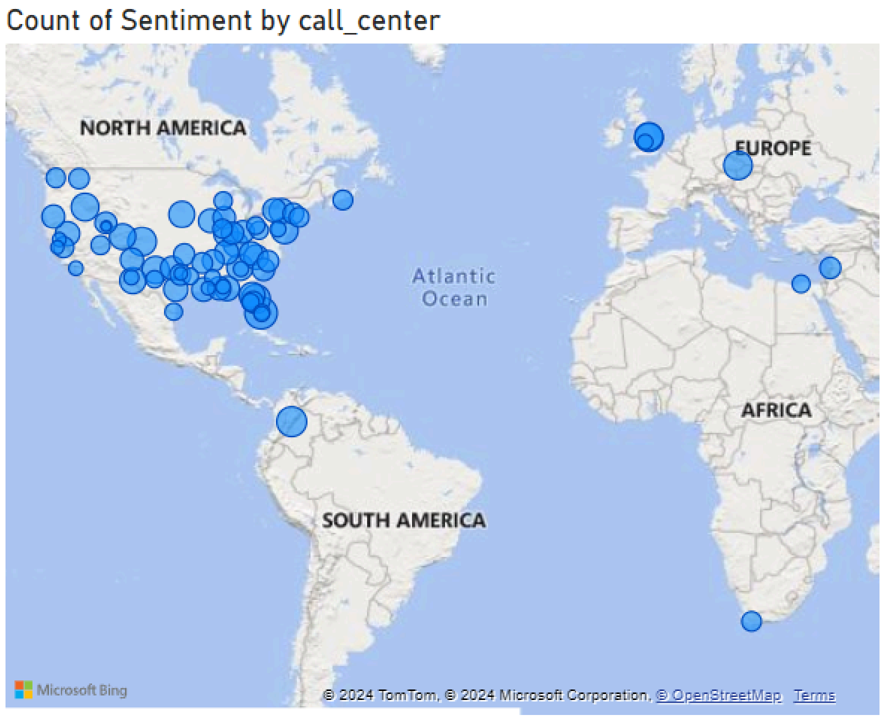
This information could be useful for tailoring financial products and marketing strategies to different customer segments based on their estimated income levels.


Count of Customer count by state_lat and state_long

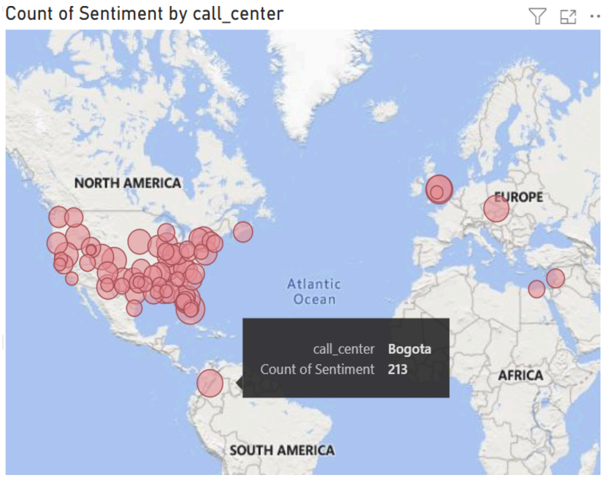
The geographic data visualization shows the distribution of customers across various states in the United States. The size of the circles represents the number of customers, with larger circles indicating a greater customer base. Concentrations of customers are visible in several states, with notable clusters around central and eastern regions. This distribution can help inform regional market strategies and resource allocation to better serve the customer base across the country.


Count of Customer count by city_lat and city_long

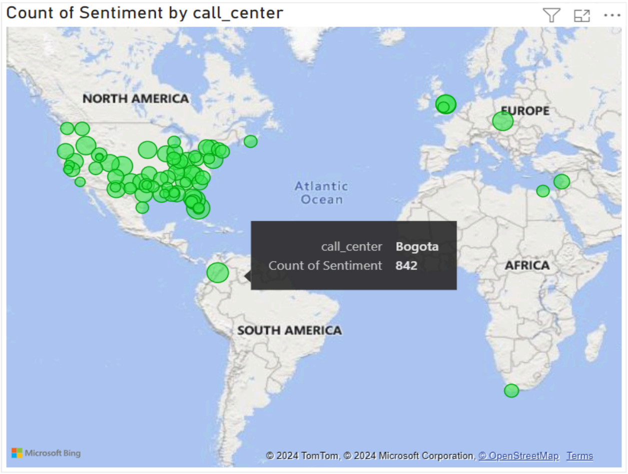
The map shows the distribution of customers by city across the United States, with each dot representing the customer count in different cities. The varied sizes of the dots suggest a diverse dispersion of customers, with some cities having higher concentrations than others. This geographic representation can provide insights into market penetration in different urban areas and may guide decisions related to marketing, sales, and distribution logistics to address customer density effectively.


Count of Sentiment by call_center
NORTH AMERICA
EUROPE
Atlantic Ocean
AFRICA
SOUTH AMERICA
Microsoft Bing
© 2024 TomTom, © 2024 Microsoft Corporation, © OpenStreetMap Terms

The map indicates the count of sentiments related to call centers across various global locations. There is a significant concentration of data points in North America, implying a higher volume of sentiment-related data, potentially reflecting customer feedback or service experiences. Other notable clusters appear in Europe and South America. This distribution suggests where the company's call center operations are most active and could highlight areas with high customer interaction volume, possibly guiding global customer service strategy and resource allocation.


Count of Sentiment by call_center

call_center Bogota
Count of Sentiment 213


Count of Sentiment by call_center

call_center Bogota
Count of Sentiment 842
NORTH AMERICA
EUROPE
Atlantic Ocean
AFRICA
SOUTH AMERICA
Microsoft Bing
© 2024 TomTom, © 2024 Microsoft Corporation, © OpenStreetMap Terms

The map showing green and red dots indicates the distribution of customers sentiment towards call centers across various locations. The green dots represent positive reviews, which are predominantly concentrated in North America, suggesting high customer satisfaction in these regions. Europe and South America also display green dots but with less frequency. The red dots, which would indicate negative reviews, are not visible on this map, suggesting that the displayed call centers have positive sentiment associated with them. This information is valuable for understanding customer satisfaction levels at different call center locations.

Implications:

The analysis conducted has multiple strategic implications:

Retention Strategies: A focus on tailored engagement for prime credit customers to mitigate churn, possibly through personalized services or offers.

Sales Channel Optimization: Increased investment in high-performing retail channels could capitalize on their strong contribution to lifetime customer value.

Lifecycle Engagement: Development of loyalty programs targeting mid-lifecycle customers could sustain or increase ARPU.

Product and Service Improvement: Addressing equipment-related issues could significantly enhance customer satisfaction and reduce churn.

Targeted Marketing: Geographical and demographic insights could lead to more personalized marketing strategies and better allocation of expansion resources.

Sentiment Monitoring: Positive sentiment in North American call centers could serve as a model for customer service practices globally.

These findings can inform future strategies to enhance customer service, optimize transactions, and leverage geographic insights for market growth.

Conclusion:

The comprehensive analysis of customer data across call centers, transactions, and geographic locations has led to several actionable insights for the telecom industry:

Churn Analysis: Prime credit customers exhibit the highest churn rates, emphasizing the need for focused retention plans. A deep dive into the reasons behind this trend could help in formulating targeted strategies to improve retention.

Revenue and Sales Channels: Retail channels emerged as the primary contributors to customer lifetime value, suggesting potential areas for investment and growth. The data points to the importance of optimizing retail operations and enhancing the customer experience in these channels.

Customer Lifecycle Value: The analysis indicates a fluctuation in ARPU related to the customer lifecycle. The trend of higher ARPU among mid-age accounts suggests that there may be opportunities to develop loyalty programs or long-term benefits to retain these valuable customers.

Service Issues and Churn: The high churn associated with equipment issues indicates a critical area for quality improvement and customer service training. Addressing these technical challenges could lead to improved customer satisfaction and reduced churn.

Geographic and Demographic Insights: The data indicates significant geographic and demographic variations in service usage and satisfaction, providing guidance for localized marketing strategies and network expansion decisions.

Sentiment Analysis: Positive feedback concentrated in certain regions highlights areas of strength in customer service, while the absence of negative sentiment indicators suggests successful service strategies in these areas.

Overall, the insights from this analysis underscore the importance of a multi-faceted approach to customer service and experience, guiding strategic investments in areas that could yield the highest returns in customer satisfaction and revenue growth.

Appendix:

Codes: bitathon_codes

PPT: Click here