DAS 732: Data Visualisation Assignment 3 Report

Aryan Singhal
IMT2022036
Aryan.Singhal@iiitb.ac.in

Pranav Kulkarni
IMT2022053
Pranav.Kulkarni@iiitb.ac.in

Rishi Patel
IMT2022041
Rishi.Patel@iiitb.ac.in

***Abstract*—Assignment 3 report prepared by the group "Mazdoor" for the course DAS-732 Data Visualization focuses on utilizing Visual Analytics to identify factors influencing the number of enrollments in an airlines's loyalty program. The analysis begins by examining temporal variations in the number of enrollments followed by other factors. Key areas of investigation include temporal trends ,demographic trends and geographical trends.**

***Index Terms*—Visual Analytics, Data Visualization, Machine Learning, Clustering, Statistical Analysis**

I. Dataset

A. Customer Loyalty History

The Customer Loyalty History dataset provides a comprehensive view of customer behaviors, loyalty program details, and demographic attributes.

Column Descriptions

- **Loyalty Number**: A unique identifier assigned to each customer to track their activity within the loyalty program.
- **Country**: Indicates the country where the customer resides. In this dataset, it consistently shows "Canada."
- **Province**: The province within Canada where the customer resides, aiding in regional segmentation and analysis.
- **City**: The city of the customer's residence, providing more granular geographical data.
- **Postal Code**: The postal code of the customer's location, which can be used for highly localized segmentation and targeted campaigns.
- **Gender**: Specifies the customer's gender (e.g., Male, Female) for demographic analysis.
- **Education**: Represents the highest level of education attained by the customer (e.g., Bachelor, Master).
- **Salary**: Indicates the annual income of the customer.
- **Marital Status**: Captures whether the customer is single, married or divorced,
- **Loyalty Card**: Represents the type of loyalty card held by the customer (Eg: Star,Nova,Aurora), which reflects their engagement level in the program.
- **CLV (Customer Lifetime Value)**: The estimated monetary value a customer will contribute to the business over their entire relationship.
- **Enrollment Type**: Indicates the type of enrollment in the loyalty program (e.g., Standard, 2018 Promotion), showing the customer's initial preferences.
- **Enrollment Date**: The date on which the customer joined the loyalty program.Null values in this column signify that the person has not enrolled into the loyalty program.
- **Cancellation Date**: The date when the customer left the loyalty program. Null values in this column represent that the customer is still a part of the program.

B. Customer Flight History

This dataset contains detailed information about customer flight activity within a loyalty program. It provides insights into customers' travel patterns, their engagement with the loyalty rewards system, and related temporal data.

Column Descriptions

- **Loyalty Number**: A unique identifier assigned to each customer, used to track their flight-related activity within the loyalty program.
- **Total Flights**: Represents the total number of flights taken by the customer on a particular date, useful for assessing engagement and activity level.
- **Distance**: The total distance traveled by the customer on a particular date, which could relate to their travel habits or preferences.
- **Points Accumulated**: The total loyalty points accumulated by the customer on a particular date through travel activity.
- **Points Redeemed**: The total loyalty points redeemed by the customer on a particular date reflecting their engagement in the rewards program.
- **Dollar Cost Points Redeemed**: The monetary equivalent of the redeemed loyalty points on a particular date, providing insight into the value returned to customers.
- **Flight Date**: The date associated with the above columns, useful for temporal analysis and tracking trends.

This dataset has only been used in Task 3 of the analyses. All other tasks have been done using only the first dataset.

II. Goal

The main goal of our analysis is to uncover patterns,trends and factors that affect the number of enrollments into an airline's loyalty program based on various factors. The factors based on which we have analyzed our data can be broadly classified into three categories:

- **Temporal Factors**: The main objective is to determine if the number of enrollments is influenced by the month or year.

- **Demographic Factors**: Clustering is used to identify the combination of factors that have the greatest impact on enrollments.
- **Geographical Factors**: This analysis focuses on the province and city of individuals in the dataset.

III. Introduction to the Workflow

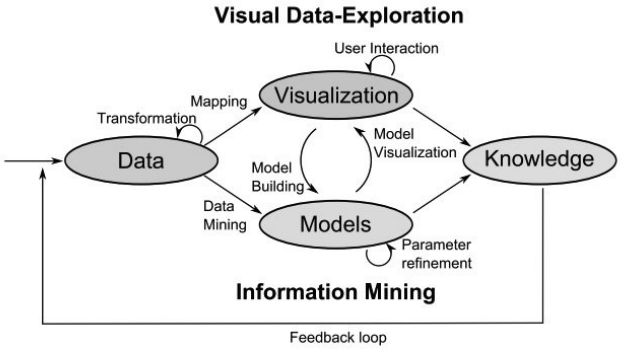


Fig. 1: Keim et al. Diamond Workflow Used for Visual Analytics.Image Courtesy: [1]

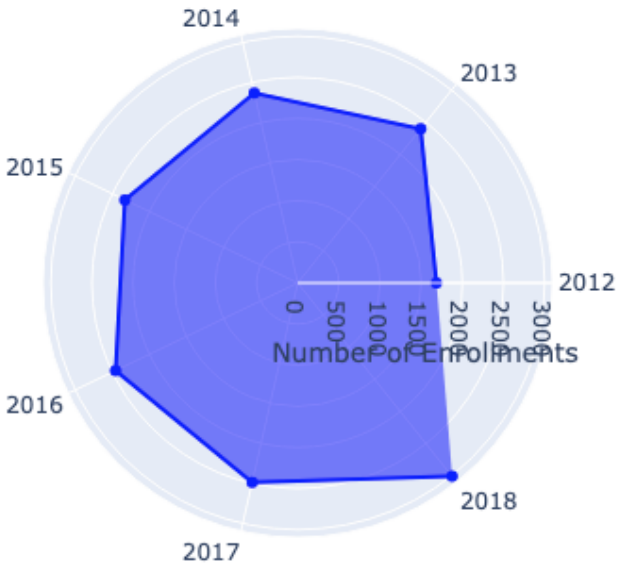


Fig. 2: Radial Area Chart Across Years for Enrollments. The chart shows the number of enrollments year wise.

The analyses has been structured along the Keim et al. diamond workflow [1], with a focus on the iterative interplay between visualization and analytics. There were different workflows designed for the analysis of temporal, demographic, and geographical factors, each suited to the particular needs of the respective dimension.

Final results from these workflows were then combined to present an overall view of enrollment patterns. Figure 1 illustrates the Keim et al. workflow and gives context to the iterative process that was adopted. Detailed implementation for each analysis are described in the respective sections.

IV. Task-1 Analyses based on Temporal Factors

In this task, we shall be using the Diamond workflow by Keim et al. to explore the temporal factors that affect the number of enrollments into the airline loyalty program.

A. First Run

This iteration tries to investigate the relationship between the year of enrollments and the number of enrollments into the loyalty program. It tries to see if there is a periodic or cyclical trend in the data.

Data

- Enrollment data for each year, month was first filtered from the airline loyalty data.
- Then from the filtered data, year wise enrollments into the program were calculated leading to data that contained the number of enrollments per year.

Visualization

- A radial area chart Figure. 2 was plotted to show the number of enrollments. A radial area shows values as radial distances from the center with the categories(year of enrollment in this case) arranged around the axis. Greater is the radial distance, greater is the value.
- The LOESS curve Figure. 3 is one kind of the non-parametric regression techniques in which trends can be caught by fitting localized linear regressions. It is effective in visualizing and smoothing non-linear relationships without assuming a specific functional form.
- All these plots were plotted using plotly giving a certain level of interactivity to the plots.

Knowledge

- From the plots, we see that the enrollments are the lowest in 2012 possibly because the program started in 2012.
- The enrollments are nearly steady from the years 2013 - 2015. It then slightly increases in the years 2016 and 2017.
- But there is a steep increase in the number of enrollments in 2018. This could possibly be because of the ”2018 Promotion” run by the airline to enroll more people into the loyalty program.

Feedback Loop

- In this iteration we see that there is no strong pattern emerging in the data, except in 2018 having a steep increase in the number of enrollments with the rest of the years having similar levels of enrollments.
- Thus there is a need to investigate at further zoomed in level to uncover any other hidden patterns.
- The next iteration does this by plotting the number of enrollments month wise.

B. Second Run

In this iteration we try to see if there any more trends visible at a zoomed in level of the data i.e. enrollment data, plotted

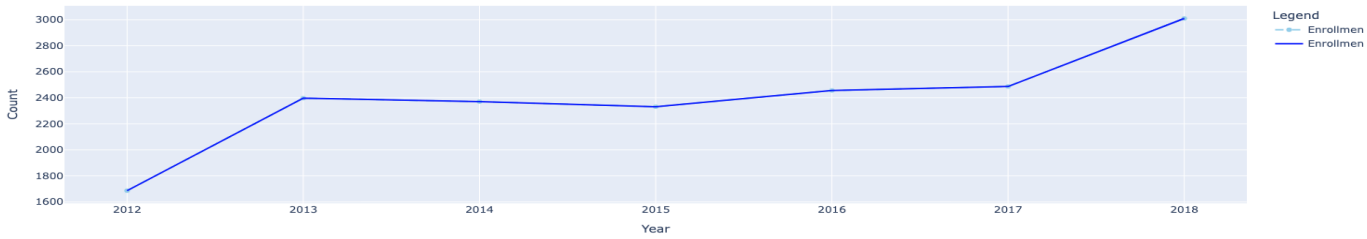


Fig. 3: LOESS Trend Curve of the Year Wise Enrollments.

for every year month - wise.

Data

- Enrollment data for each year, month was first filtered from the airline loyalty data.
- Then from the filtered data, for every year, monthly enrollments into the program were calculated leading to data that contained the number of enrollments per year per month.

Visualization

- A line chart Figure. 4 showing the number of enrollments for year month wise. The different years are plotted using different colored lines and are overlay-ed onto the same layout.
- An area chart Figure. 5 of the absolute number of enrollments by month for each year; the contribution of each year is clearly visible there.
- A normalized stacked area chart Figure. 6 for the enrollment proportion by month and year. All years' data are normalized to give relative trends rather than absolute values.

Knowledge

- The line plot Figure. 4 does confirm our earlier prediction, with the enrollments being nearly the same for all the years, except the year 2018. But the number of enrollments in the year 2018 seem to be higher than the rest in the first half of the year i.e. from January to June, but seem to come back down to the same level as other years in the later half i.e. from July to December.
- No other pattern is discernible from the line plot. Hence an area chart with the raw counts of enrollments Figure. 5 was plotted. From the plot there seems to be a weak trend, which is not clearly visible. Hence a normalized area plot was plotted as it amplifies any patterns that are present in the layout.
- From the normalized area plot of the number of enrollments month wise for each year, we see that the number of enrollments are the highest in May and July possible due to the holiday season. They are the lowest in January.

Feedback Loop

- In this analysis, we observed a very weak trend in the enrollment data.
- Thus, further work is needed to determine if a stronger trend emerges with different temporal groupings such as seasons or quarters.

C. Third Run

In this iteration we are investigating the data through temporal grouping. We group the months according to seasons and try to investigate if any trends may emerge.

Data

- From the enrollment data that was obtained for every year, month - wise, seasonal enrollment data was created from it.
- The seasons were considered according to Canadian weather and are the following:
  - Winter : December to February
  - Spring : March to May
  - Summer : June to August
  - Fall : September to November

Visualization

- Enrollment data for each year, month was first filtered from the airline loyalty data.
- Then from the filtered data, for every year, monthly enrollments into the program were calculated leading to data that contained the number of enrollments per year per month.

A normalized area chart Figure. 8 was plotted using plotly. The plot shows normalized seasonal enrollments across different years distributed across different seasons. For each year from 2012-2018 and for each season (Winter, Spring, Summer, Fall) the enrollment values were normalized and plotted to see if there are any trends. The reason a area chart was chosen is because it amplifies patterns if any in the data. In a similar manner a area chart Figure. 7 with raw counts was also plotted.

Knowledge

- From the raw area chart in Figure 7, it is clear that enrollment is higher in fall and summer compared to spring and winter seasons, though the trend is not very strong.
- However, when looking at the normalized area chart in Figure 8, the trend is stronger. Summer and fall always have a peak in enrollment, which must be because of the vacation travel rush during summer. Fall enrollment

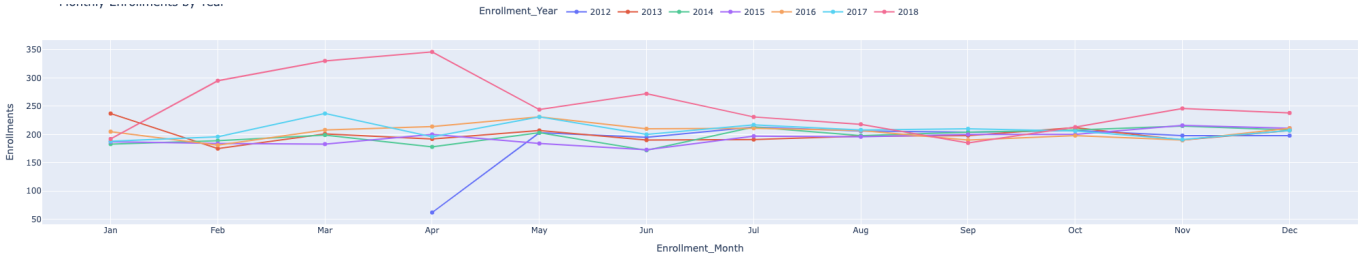


Fig. 4: The Line Plot Showing the Monthly Enrollments of Every Year Overlay-ed on a Single Layout. Different Colored Lines Represent Different Years.

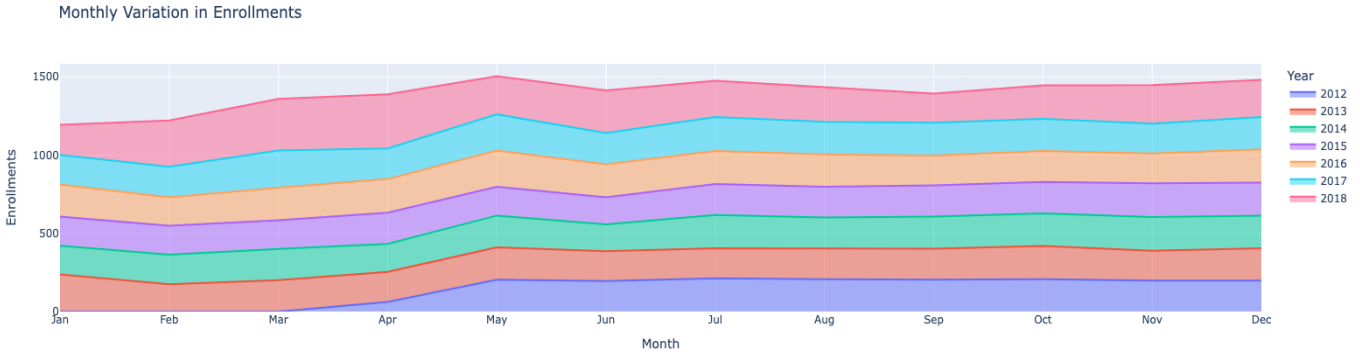


Fig. 5: The Line Plot Showing the Monthly Enrollments of Every Year Overlay-ed on a Single Layout. Different Colored Lines Represent Different Years.

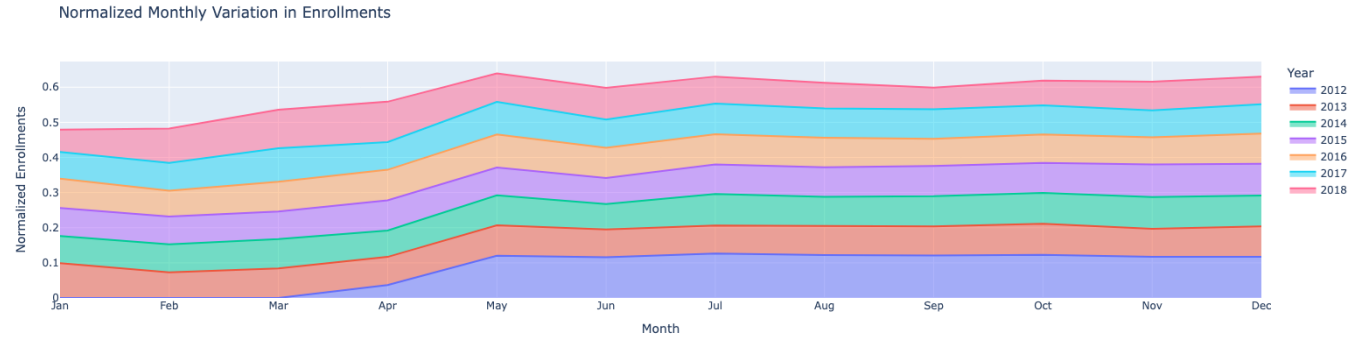


Fig. 6: The Line Plot Showing the Monthly Enrollments of Every Year Overlay-ed on a Single Layout. Different Colored Lines Represent Different Years.

uptick may be partly because of the pickup in business activities following the slower summer months.

Feedback Loop In this iteration we found that there is a weak trend in the enrollments data based on seasonal grouping of months. There could be a possibility that based on some other temporal grouping a stronger trend may emerge.

D. Fourth Run

In this iteration we try to see the data by grouping the month according to the quarter i.e. we analyse the data accordking the quarter it is in.

Data

- Enrollment data for each year, month was first filtered from the airline loyalty data.
- Then from the filtered data, for every year, quarterly enrollments into the program were calculated leading to data that contained the number of enrollments per year per quarter.

Visualization

- Circular bar plots in Figure. 9 were used to plot normalized quarterly enrollments over the years using Plotly. Every circular plot is one year, and the bars are divided

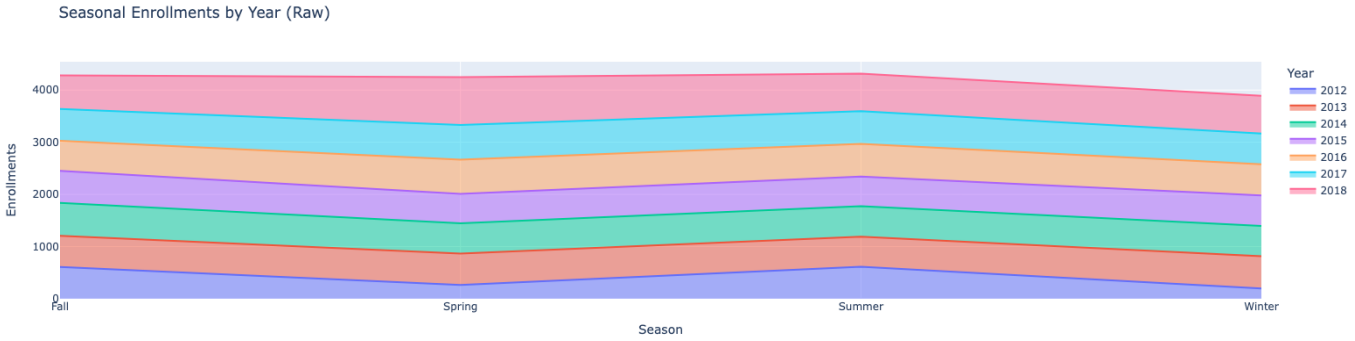


Fig. 7: This area chart shows the total number of enrollments across each season for every year.

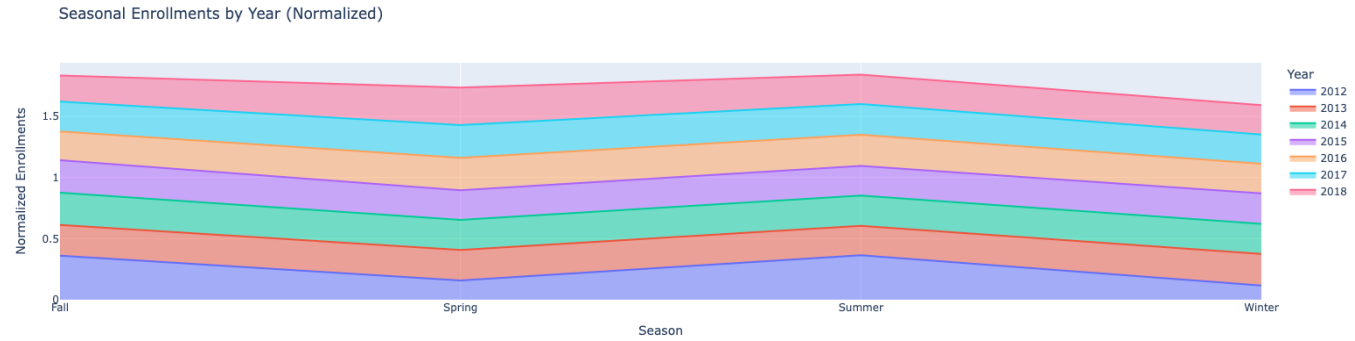


Fig. 8: This area chart shows the normalized total number of enrollments across each season for every year.

into four quarters: Q1, Q2, Q3, and Q4. The values were normalized to look for proportional trends within the quarters for every year.

- Figure. 10 shows the raw enrollment counts in circular bar plots. The plot gives absolute values for quarterly enrollments over the same years (2012–2018). Because of their compact display and ability to convincingly graph cyclical data, circular bar plots were used in the interest of evaluating any patterns over time and between quarters.

Knowledge

- From the raw circular bar chart 9 we observe that 2018 has the highest number of enrollments among all years.
- But from the normalized circular bar chart 10, there is no trend visible.
- Thus there is no trend in the enrollment data when the temporal grouping is quarterly.

E. Summary of Findings

- From the *First Run*, we found out that the enrollments were lowest in 2012 possibly because the program started in 2012. The enrollments remained steady throughout the years, spiking in 2018 possibly due to the ”2018 Promotion” done by the airline. This would suggest that such promotions are good for the loyalty program as it attracts more number of people to enroll into the program.
- In the *Second Run*, we observed that the number of enrollments are highest in May and July possibly due to the holiday season and they are the lowest in January.
- The *Third Run*, showed that the the number of enrollments in the summer and fall season are the high and low in winter and spring and winter. A possible reason for enrollments to be high in summer could be because people traveling more because of the holiday season. The fall season uptick may be partly because of pickup in business activities following the slow summer months.
- In the *Fourth Run*, we analyzed the data according to quarters and found no trend in quarterly enrollments.
-

V. Task-2 Analyses based on Demographic Factors

A. First Run

The fist logical factor that should affect number of enrollments the most is Salary. So this iteration aims to explore the relationship between salary bands and the number of enrollments in the program. The salaries have been categorized into predefined bands to analyze how enrollments are distributed across these groups.

Data

- Salary values were categorized into predefined bins:
  - Low (<30K)

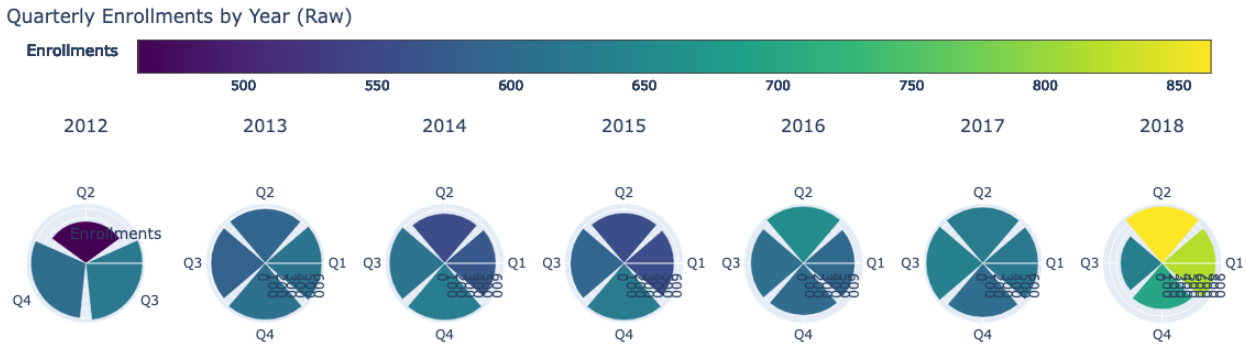


Fig. 9: Circular bar plots showing quarterly enrollments for each year from 2012 to 2018. The bars are proportional according to the four quarters, Q1, Q2, Q3, Q4 for each of the years

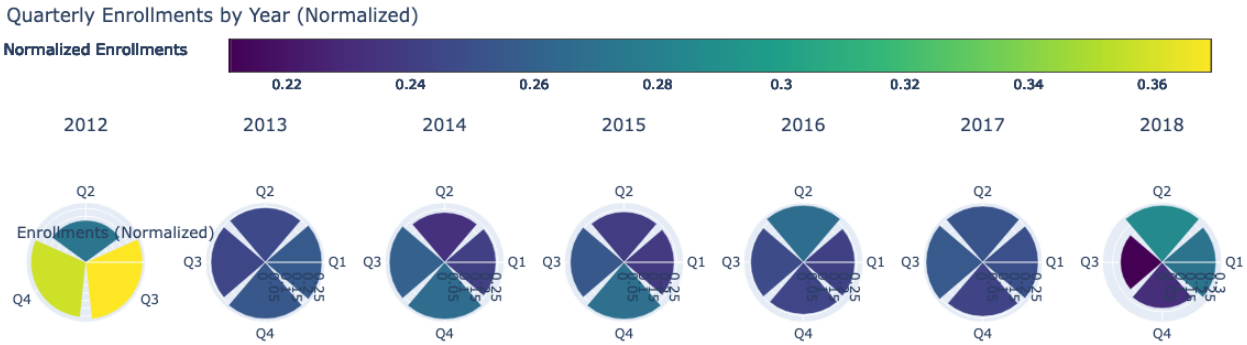


Fig. 10: Circular bar plots showing normalized quarterly enrollments for each year from 2012 to 2018. The bars are proportional according to the four quarters, Q1, Q2, Q3, Q4 for each of the years

  - Lower-Mid (30K–50K)
  - Upper-Mid (50K–75K)
  - High (75K–100K)
  - Very High (>100K)
- The data was grouped by salary band to compute the number of enrollments in each category.

Visualization

- A radial bar chart (Figure 11) was used to show the distribution of enrollments across salary bands.
- Key observations:
  - The Upper-Mid (50K–75K) band had the highest enrollments.
  - This was followed by the High (75K–100K) band.
  - Enrollment was lowest in the Lower-Mid Band (30K - 50K) band.

Knowledge

- Higher salary bands tended to have higher enrollments.
- This trend suggests a relationship between financial capacity and enrollment likelihood.
- But at the same time we surprisingly see that people from the lowest band also contribute greatly to the overall number of enrollments.
- One thing to note here is that context of the total population within each salary band was missing which limits the insights that we can draw.
- Also we don't know anything about the skewness that is present in the salary data which might affect the definition of the bands.

Feedback Loop

- In this iteration the need to compare enrollments with the total population in each salary band has been identified.
- The need to understand the distribution of the salary data

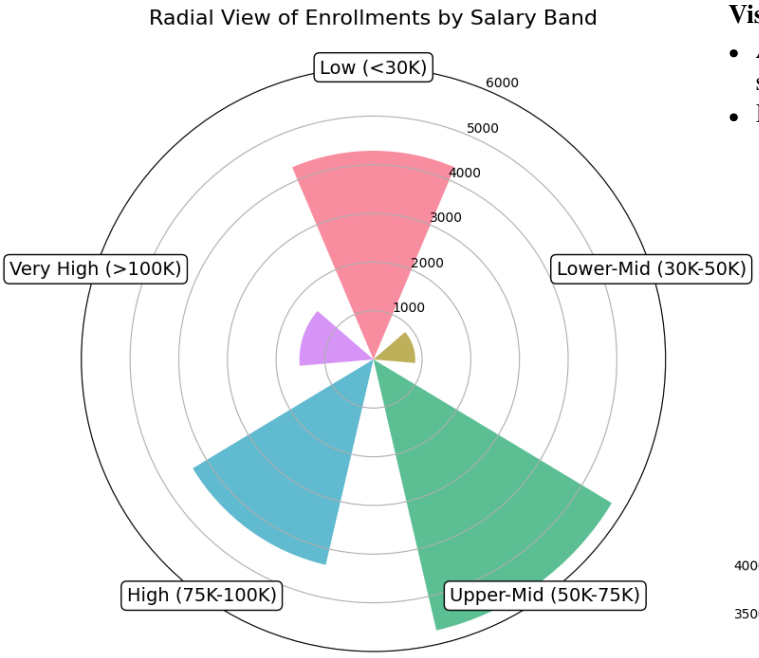
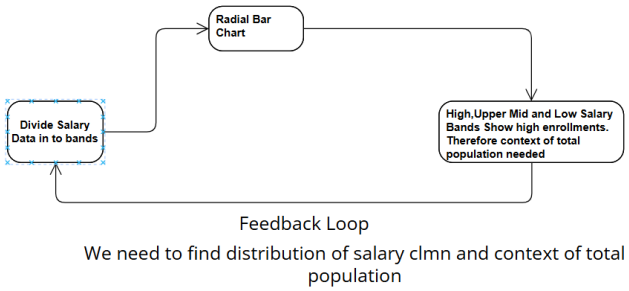


Fig. 11: Radial View of Enrollments by Salary Band. The chart shows the number of enrollments across different income levels.

in the dataset was also identified.

- The next iteration was planned to address these gaps.



This is the Visual Analytics Workflow used in the first iteration of Task 2.

B. Second Run

This iteration focused on comparing the total population in each salary band with the total number of enrollments in that salary band to assess and identify any potential biases.The focus was also finding out the distribution of values in the Salary Column.

Data

- Total population counts for each salary band were calculated.
- Enrolled population counts were extracted by filtering records with non-null enrollment dates.

Visualization

- A box plot for showing the distribution of values in the salary column(Figure 12)
- Key observations:
  - The box plot in Figure 12 clearly shows that salary values are heavily concentrated in the lower range of the graph (below 100K), with a few extreme outliers in the higher range.
  - This indicates that even minor changes to the definition of salary bands can significantly change the distribution of enrolled individuals across those bands.
  - As a result, defining salary bands arbitrarily and drawing conclusions based on them is not ideal, especially in the absence of a standard definition.

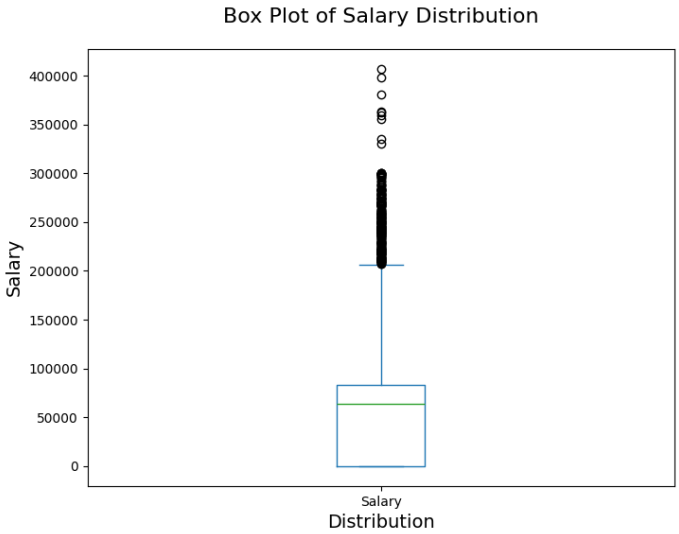


Fig. 12: A box plot showing the distribution of salary across the dataset.

- A horizontal stacked bar chart (Figure 13) was created to compare the total population with the enrolled population.
- Key observations:
  - The total and enrolled populations were totally identical across all salary bands.
  - This suggests the dataset might only represent enrolled individuals.

Knowledge

- The box plot shows that most salaries are concentrated in the lower range (below 100K CAD), with just a few very high outliers.
- This means salaries are unevenly distributed, so even small adjustments to salary bands could lead to big shifts in how people are grouped into these bands.This reduces the reliability of salary as a metric to find patterns in the number of enrollments especially when there is no standard definition available for the salary bands.
- The counts for total and enrolled in the population are almost exactly the same across all salary bands. This is highly improbable.

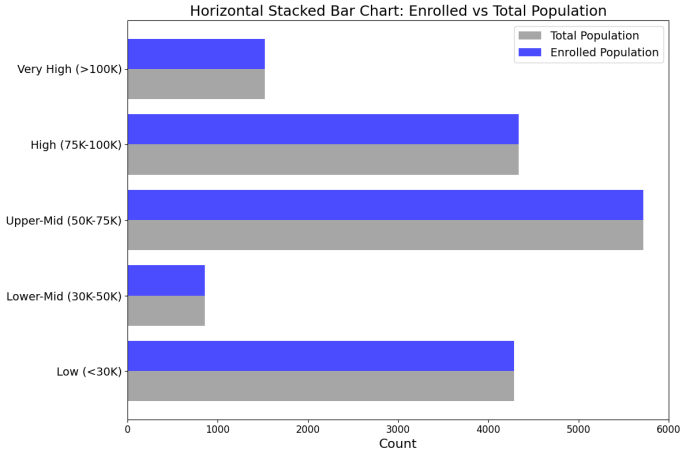


Fig. 13: Horizontal Stacked Bar Char comparing the total population in each salary band with the enrolled population.

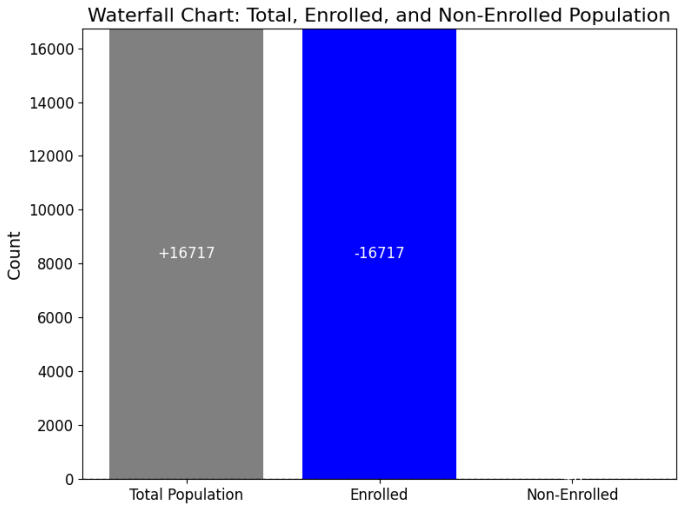


Fig. 14: Waterfall Chart depicting proof that there are no non-enrolled individuals within the dataset.

- There may be little to no data for the non-enrolled in the dataset.

Feedback Loop

- The hypothesis that non-enrolled data is missing requires validation.
- So the next iteration focuses on verifying this hypothesis.
- Also it has been established that salary is not a good metric to find demographic patterns so in the Fourth iteration other factors will be looked at.

C. Third Run

This iteration aims to validate whether the dataset lacked information about non-enrolled individuals.

Data

- Total population counts for the dataset were computed.
- Enrolled and non-enrolled population counts were calculated.

Visualization

- A waterfall chart (Figure 14) shows:
  - Total population
  - Enrolled population
  - Non-enrolled population
- Key observations:
  - The non-enrolled population count is zero.
  - This confirmed the absence of non-enrolled individuals in the dataset.

Knowledge

- The dataset only contains data about enrolled individuals.
- This limitation restricts the analysis to patterns among enrolled individuals only.

Feedback Loop

- The next steps would focus on analyzing demographic factors among enrolled individuals only.
- Future iterations would eliminate non-enrollment trends. due to a lack of data

D. Fourth Run

This iteration has been designed to analyze the demographic factors: gender, marital status, and education, in the hope of discovering enrollment patterns. Hierarchical clustering is done to identify the segments within the demographics that are similar with respect to their enrollment numbers.

Data

- The data is grouped based on three demographic factors: Gender, Marital Status, and Education.
- For every unique combination of these three factors, the total amount of enrollments was calculated, reflecting the share of every demographic group in the total enrollments.
- This grouped data acts as an input to the cluster algorithm, and clustering is performed based on the total enrollments belonging to each group.

Model

- Hierarchical clustering was chosen over other clustering methods, such as k-means, because it:
  - Allows us to understand relationship between factors at various granularities by allowing exploration of the clusters at various levels.
  - Enables investigation into relationships between multiple factors, such as gender, marital status, and education, without analyzing them separately.
  - Provides a visual representation of the clustering process through a dendrogram, which facilitates easier interpretation of hierarchical groupings.
  - Does not require predefining the number of clusters, offering flexibility for exploratory analysis.
- The Ward method was used for clustering, which minimizes the variance within clusters and ensures that the groups formed are compact and distinct.

Visualization

- A dendrogram was used to visualize the process of hierarchical clustering.(Figure 15).
- To delve further into the two clusters, the Parallel Coordinates Plot (PCP) was generated, as shown in (Figure 16).
- Key observations from the dendrogram:
  - The number of clusters was determined to be two as an optimal number.
  - These two clusters are two different demographic segments with different enrollment behaviorsiors.

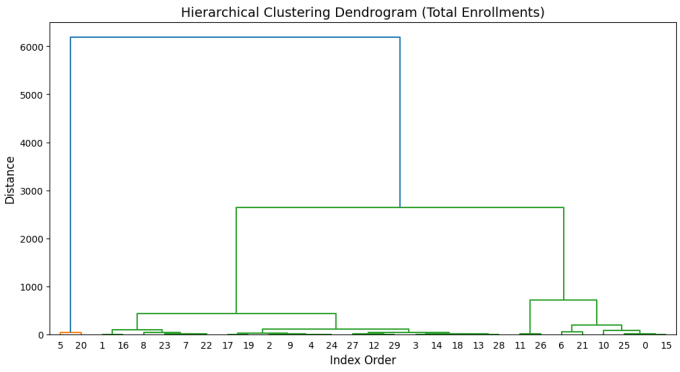


Fig. 15: Hierarchical Clustering Dendrogram (Total Enrollments).

Knowledge

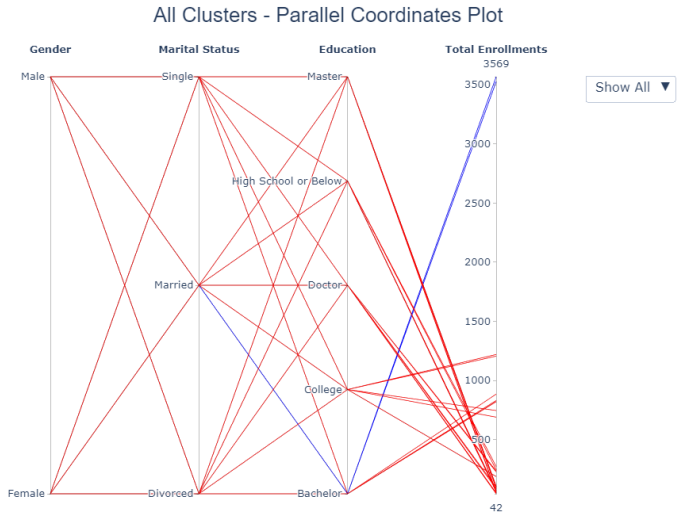


Fig. 16: Parallel Coordinates Plot (PCP). This plot shows the two clusters identified, with blue lines representing one cluster and red lines representing the other.

- In the PCP plot:
  - Blue lines represent one cluster, and red lines represent the other.
  - When focusing on Cluster 1 (Figure 17), this cluster was observed to represent higher enrollments.
  - Also we observe that only two demographic groups belong to this cluster.

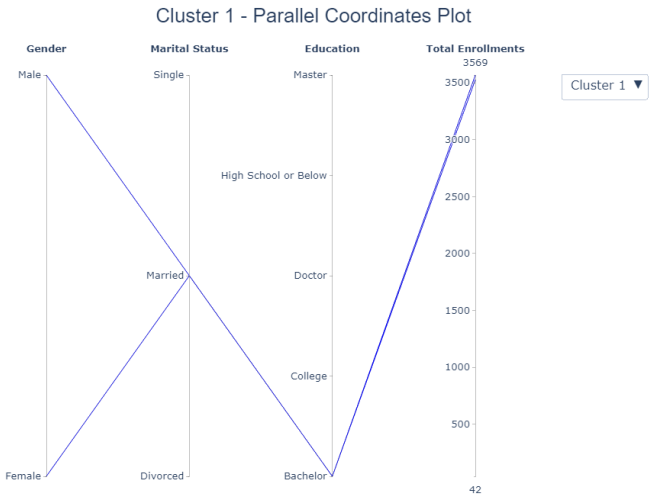


Fig. 17: Parallel Coordinates Plot for Cluster 1. This cluster represents higher enrollments

- Insights from Cluster 1:
  - Gender had no significant effect on enrollment, as both males and females contributed similarly to the cluster.
  - Married individuals with a bachelor's degree formed the majority of this cluster, leading to higher enrollments.
  - A possible reason is that married individuals may travel more frequently for family commitments, such as visiting relatives or going on vacations, while bachelor's degree holders might travel more often for work-related purposes, making them more likely to enroll in an airline loyalty program to benefit from their frequent travel.

Feedback Loop

- While Cluster 1 represents higher enrollments,Cluster 2 remains entirely unexplored.
- In the next iteration we will be looking at the second cluster(with lesser enrollments) in more detail to analyse any patterns or anomalies in it. This is done by dividing entire data into three clusters instead of two.
- This demonstrates one of the advantages of using hierarchial clustering, which is the ability to analyze specific clusters in more detail for a deeper analysis.
- From the dendrogram in Figure 15 we can clearly see the clusters 2 and 3 from the next iteration will combine to form the cluster with lesser enrollments in the present iteration. Analyzing these two separately could provide some more insights.

E. Fifth Run

This iteration extends the analysis by segmenting the dataset into three clusters using hierarchical clustering for further insight into enrollment patterns of Cluster 2 above. **Data**

- The dataset was grouped by Gender, Marital Status, and Education to calculate the total number of enrollments for each combination.
- Hierarchical clustering was applied to segment the data into three clusters instead of two.
- Mean enrollments for each cluster were computed to determine which cluster has the highest average total enrollments.
- Clusters 2 and 3 from this iteration form the earlier Cluster 2 from the previous iteration. This split now allows a more in-depth look into the trends within that cluster.

Model

- Hierarchical clustering was done using the Ward method, with a division into three clusters.
- The three-cluster division gave a finer granularity of analysis than the two-cluster model in the previous iteration, thus helping in much greater detail in analyzing the second cluster from the previous iteration.
- This shows one of the strengths of hierarchical clustering: to drill down into the pattern within a given cluster as the analysis progresses.

Visualization

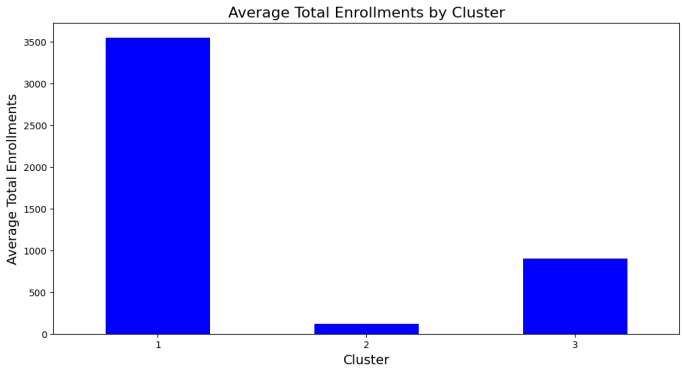


Fig. 18: Average Total Enrollments by Cluster. Cluster 3 shows the highest mean enrollments.

- A bar chart was created to compare the average total enrollments across the three clusters (Figure 18).
  - Cluster 3 demonstrated the highest mean total enrollments between Clusters 2 and 3, making it the focus of further analysis.
- A sunburst chart was generated to analyze the demographic distribution within Cluster 3 (Figure 19).
  - The chart highlights contributions to enrollments across Gender, Marital Status, and Education levels within Cluster 3.

Knowledge

- Cluster 3 has the highest average total enrollments among the three clusters in this iteration.
- Insights from the sunburst chart for Cluster 3:
  - Single individuals that are college going formed the majority within Cluster 3, contributing significantly to enrollments.
  - Gender had minimal impact, as both males and females contributed equally to the cluster.
  - The reason for this could be that college going students travel more as they travel back and forth between college and home.Also most college goinf students wouldn't be married yet much less divorced.

F. Summary of Findings

- **Iteration 1**: Salary bands showed a potential relationship between income level and enrollment, but the analysis was limited in its correctness because of skewness in salary data had to be checked.
- **Iteration 2**: The comparison of total and enrolled populations suggested that the dataset exclusively represents enrolled individuals. Also the skewness in the salary data meant that our arbitrary definition of the bands couldn't be used to draw any insights or conclusions.
- **Iteration 3**: The waterfall chart confirmed that the dataset lacks information on individuals not enrolled in the loyalty program, emphasizing the need to focus analysis solely on the enrolled population.
- **Iteration 4**: This stage shifted attention to other demographic attributes. Hierarchical clustering identified two main demographic clusters. Cluster 1, which had higher enrollments, predominantly consisted of married individuals with a bachelor's degree, while gender showed minimal influence.
- **Iteration 5**: A deeper analysis of Cluster 2 from the previous iteration revealed that the majority of enrollers in this cluster were single, college-going individuals, with gender again having no significant role.
- So in conclusion we can conclude that most of the people enrolled in the loyalty program are Married Bachelors with the second largest group being Single College going individuals.
- Also the number of Married Bachelors is disproportionately high.

VI. Task-3 Analyses based on Geographical Factors

A. First Run

In this iteration, the goal is to identify urban provinces based on their air traffic trends. This analysis lays the foundation for exploring geographical factors affecting enrollments with assumption that more urbanized provinces should have a higher number of people enrolling for the loyalty program.

Data:

- Air traffic data was analyzed province wise, grouped by year.
- For this analyses the Customer Flight Activity dataset was used.
- The 'Flight Year' was extracted from the flight dates to analyze yearly trends.

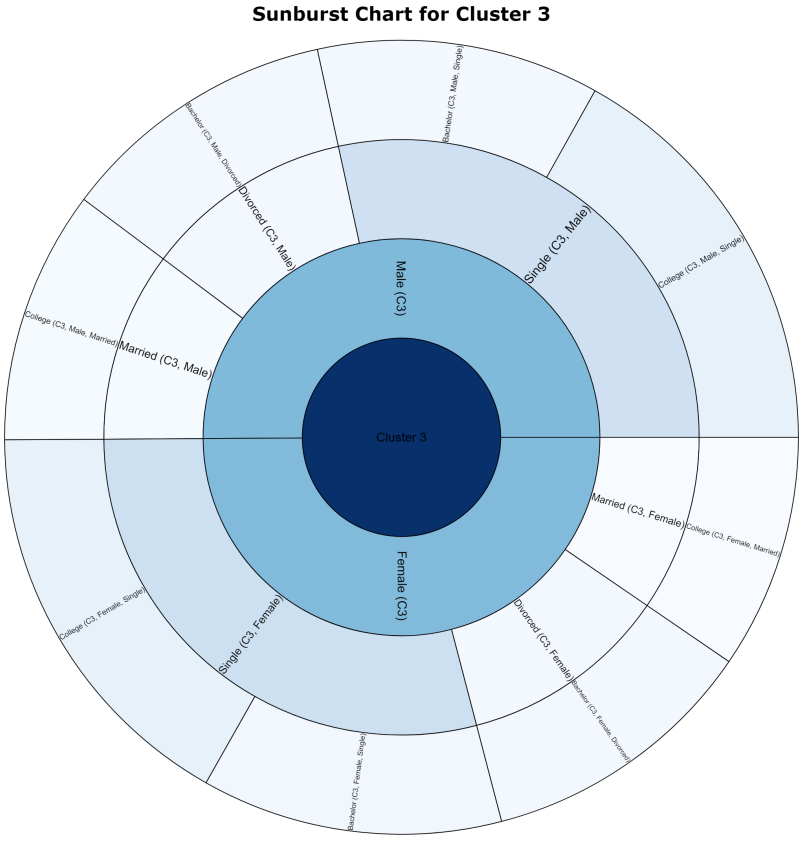


Fig. 19: Sunburst Chart for Cluster 3 showing Demographic Distribution. Cluster 3 primarily includes single individuals with college education.

- Total flights per province were calculated from the above to rank provinces by air traffic.
- A consistent color scheme was used iin both years to show the provinces.

Visualization:

- The radial bar charts of Figure 20 represent Provincal air traffic for 2017 and 2018, since those were the years the only 2 years for which data was available.
- The charts highlight provinces with the highest air traffic.This means these regions could be considered urban.

Knowledge:

- Ontario, British Columbia, and Quebec can been classified as urban provinces based on their air traffic volumes.
- These provinces align with economic hubs and population centers of Canada according to the recent Canadian census of the year 2021 [2]

Feedback Loop:

- Based on this analysis we can now shift our focus to comparing enrollments on urban versus rural provinces.

B. Second Run

This iteration builds on the previous feedback and compares Enrollments in urban and rural provinces over the years to understand and visualize the implications of regional classification. on enrollments.

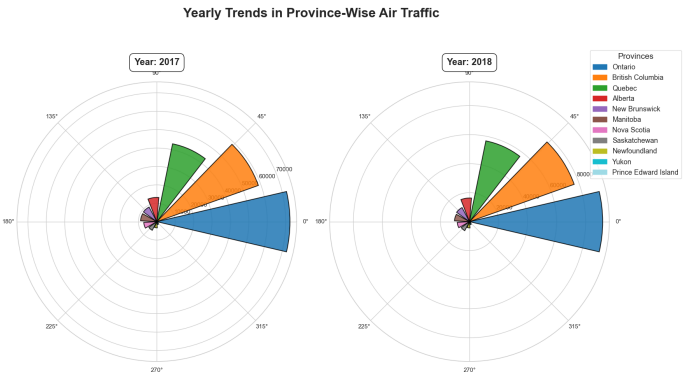


Fig. 20: Yearly Trends in Province-Wise Air Traffic. Urbanized Provinces should show higher air traffic.

Data:

- Provinces were classified as urban or rural depending upon the air traffic analysis of previous iteration..
- Enrollment data was grouped by region type (‘Urban‘ or ‘Rural‘) and enrollment year.

Visualization:

- A stacked area chart (Figure 21) has been used to compare enrollments over time in urban and rural provinces.
- The chart points out the clear dominance of urban provinces in enrollments.

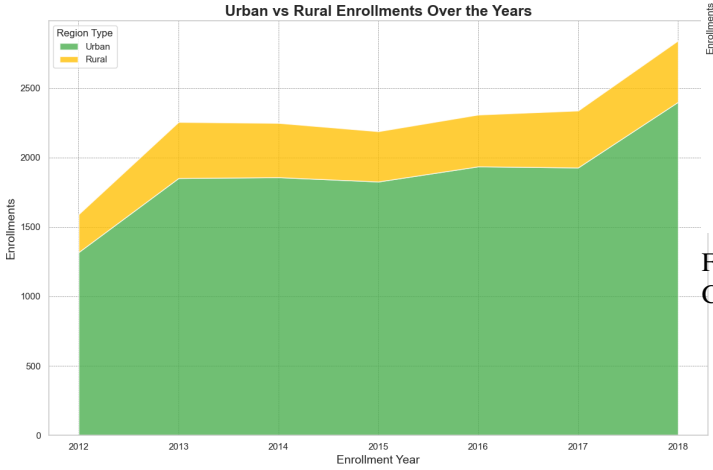


Fig. 21: Urban vs Rural Enrollments Over the Years. Urban provinces consistently show higher enrollments.

Knowledge:

- Urban provinces significantly outpace rural provinces in enrollments which is expected.
- This trend is consistent with better access to airports and economic opportunities in urban areas.
- Even though the urban provinces clearly have more enrollments that rural provinces, it might be that there are some provinces that have been classified as rural based on air traffic but they still have a huge number of enrollments.This is highly unlikely but never the less should be checked.

Feedback Loop:

- The next iteration will analyze province-specific trends to check if there are any provinces that have been classified as rural but still contribute greatly to the number of enrollments.

C. Third Run

This iteration focuses on province-wise enrollments to identify top contributors and their trends over time.

Data:

- Enrollment data was grouped by province and year in order to make the visualization.

Visualization:

- A line chart (Figure 22) was plotted to show enrollment trends for each province over the years.
- This chart also highlights the disproportionate dominance of Ontario, British Columbia, and Quebec,which have been classified as urban, in the number of enrollments into the loyalty program.
- All other provinces seem to have a similar number of contribution to total enrollments which is low.

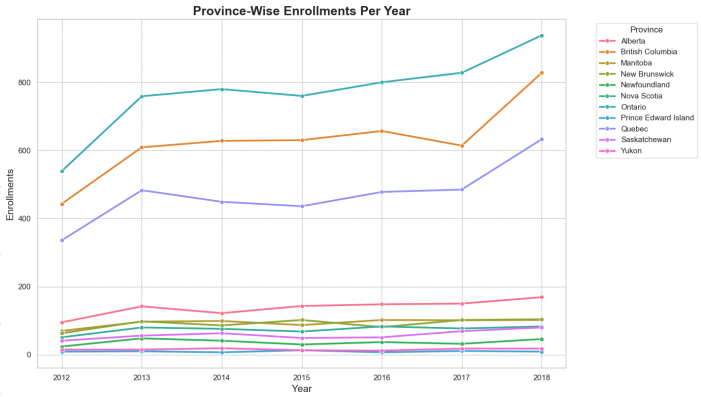


Fig. 22: Province-Wise Enrollments Per Year. Ontario, British Columbia, and Quebec dominate enrollments.

Knowledge:

- Ontario, British Columbia, and Quebec show consistently high enrollments.
- It has been proved that there doesn't exist any province that has been classified as rural that contributes greatly to the total number of enrollments.
- However, contributions from these provinces may be concentrated in specific cities rather than being evenly distributed throughout the province.
- If that is the case it would be wrong to conclude that urban provinces contribute greatly to the number of enrollments as a whole.

Feedback Loop:

- Next, city-level contributions will be analyzed to identify key contributors and to check if all cities in urban provinces contribute proportionately to the number of enrollments.

D. Fourth Run

This iteration narrows the focus to cities so that any anomalies in the total number of contributions can be detected and to check where these cities with anomalous contributions lie.

Data:

- Enrollment data was first grouped by city and province to detect cities with unusually high enrollment contributions.
- A heatmap was created to visualize yearly enrollment trends across cities, helping to pinpoint patterns and outliers.
- The top contributing cities were isolated, and their contributions were analyzed in relation to their provinces by calculating the percentage contribution of each city to its respective province.
- Coordinates for the top cities and approximate central points of their provinces were determined for mapping purposes.

Visualization:

- The heatmap (Figure 23) revealed that Toronto, Vancouver, and Montreal consistently contributed disproportionately high enrollments across all years.

- A map of Canada (Figure 24) was created to show:
  - The three provinces with the highest enrollments (Ontario, British Columbia, Quebec) marked with blue dots.
  - The top cities in these provinces (Toronto, Vancouver, Montreal) marked with red dots.
  - Province labels displaying total enrollments and the percentage contribution from their top cities.
  - City labels indicating their total enrollments.

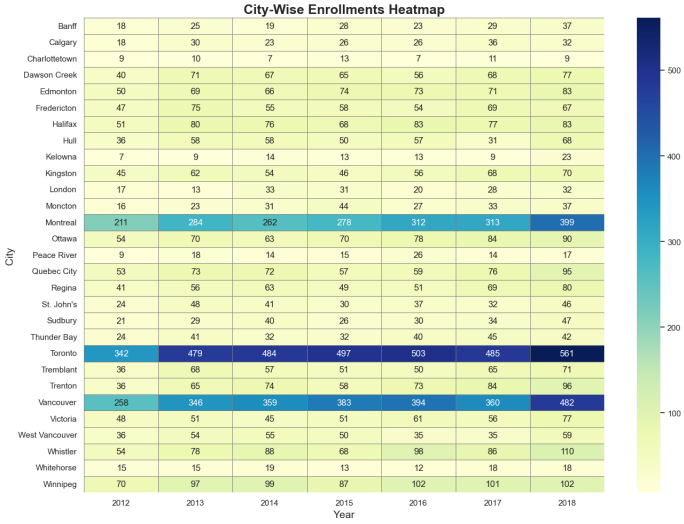


Fig. 23: City-Wise Enrollment Heatmap. Cities with the highest enrollments are clearly visible.

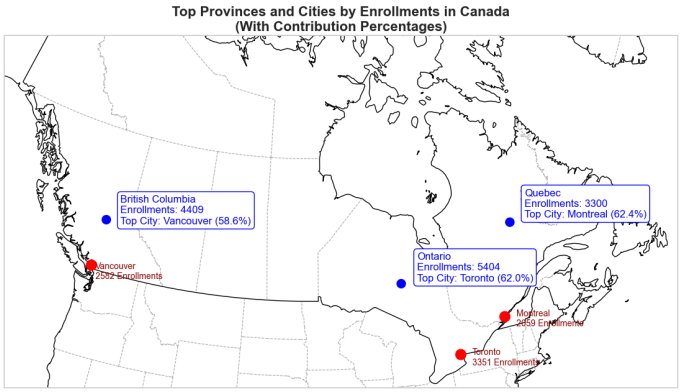


Fig. 24: Top Provinces and Cities by Enrollments in Canada.

Knowledge:

- Heatmap Analysis:
  - The top three cities by enrollments across all years were identified as Toronto, Vancouver, and Montreal.
  - These cities dominate the enrollment counts, with much smaller contributions from other cities.
- Mapping Insights:
  - The top three provinces by enrollments are:
    - Ontario: 5,404 enrollments, with Toronto contributing 62.0% (3,351 enrollments).
    - British Columbia: 4,409 enrollments, with Vancouver contributing 58.6% (2,852 enrollments).
    - Quebec: 3,300 enrollments, with Montreal contributing 62.4% (2,059 enrollments).
  - These findings confirm that the majority of enrollments in these provinces are concentrated in their largest cities.
  - The contributions from these cities significantly overshadow the enrollments from other cities, demonstrating a highly concentrated pattern of enrollment activity.

Feedback Loop:

- The findings emphasize the need to investigate and classify these cities in order to get a pattern that may validate and explain their role as key contributors in the number of enrollments.

E. Fifth Run

The findings from the previous iteration emphasized the need to classify the top contributing cities to better understand their role in the number of enrollments. This classification aims to determine whether these cities share common characteristics that explain their disproportionate contribution, validating their significance as key contributors.

Ideally we would like to classify the cities as metro and non metro and generally say that metro cities are likely to contribute to the loyalty program in higher numbers.

Data:

- Air traffic data from earlier iterations was reused o and this time grouped by city in order to help classify the cities.
- The data for the top 10 cities across Canada was compiled, and cities with disproportionately higher air traffic counts were classified as metro regions.

Visualization:

- The air traffic trends for the top 10 cities were visualized using radial bar charts for two different years (Figure 25).
- These charts highlight the disproportionate air traffic handled by Toronto, Vancouver, and Montreal compared to other cities.

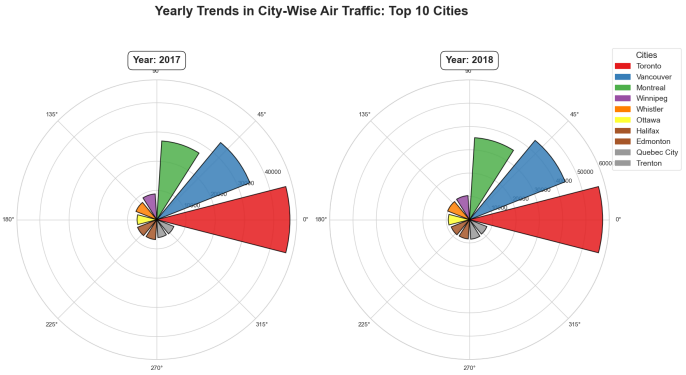


Fig. 25: Yearly Trends in City-Wise Air Traffic: Top 10 Cities.

Knowledge:

- Metro Classification:
  - Toronto, Vancouver, and Montreal, identified as the top contributors in Iteration 4, were classified as metro cities based on their high air traffic volumes.
  - This classification aligns with the 2021 Census of Canada [2], where these cities are officially categorized as metropolitan areas.
- Contribution to Enrollments:
  - These metro cities, which exhibit high air traffic, also contribute disproportionately to the total number of enrollments as can be clearly inferred from Figure 23.
  - The alignment of air traffic and enrollment contributions and the successful categorization of cities as metro cities based on air traffic data reinforces the hypothesis that metro cities are significant drivers of enrollments.
- Validation:
  - The findings validate that metro cities, characterized by higher economic activity, infrastructure, and connectivity, play a dominant role in enrollment patterns.
  - This provides a cohesive explanation for their repeated identification as key contributors in previous iterations.

F. Summary of Findings

- Urban provinces (Ontario, British Columbia, Quebec) dominate enrollments, driven by their high air traffic.
- Within these Urban Centers enrollment contributions are concentrated in major cities (Toronto, Vancouver, Montreal).
- Metro cities with high air traffic are the primary contributors to enrollments in the loyalty program.

VII. Conclusion

The analysis successfully identified patterns and trends influencing enrollments in the airline's loyalty program. Key findings across the tasks include:

- **Temporal Factors**: Enrollment spikes were identified in 2018 due to promotions and seasonally during summer and fall, correlating with holiday travel and resumed business activities.
- **Demographic Factors**: Clustering revealed two major groups that contribute to enrollments: married individuals with bachelor's degrees (highest enrollment) and single college students. Salary proved unreliable for insights due to dataset limitations.
- **Geographical Factors**: Enrollment dominance in urban provinces (Ontario, British Columbia, Quebec) was driven by metro cities (Toronto, Vancouver, Montreal). These cities, with high air traffic and economic activity, are pivotal to the loyalty program's success and contribute the highest to the total enrollments.

VIII. Contributions

- **Task-1 : Pranav Kulkarni (IMT2022053)**:
- **Task-2 : Aryan Singhal (IMT2022036)**:
- **Task-3 : Rishi Patel (IMT2022041)**:

References

[1] ResearchGate. *Visual Analytics Process defined by Keim et al. (2008)*. Retrieved from https://www.researchgate.net/figure/sual-analytics-process-defined-by-Keim-et-al-Keim-et-al-2008_fig1_301721746

[2] Statistics Canada. *Census Program*. Retrieved from https://www12.statcan.gc.ca/census-recensement/index-eng.cfm

Appendix

The A-1 report has been added as appendix to this A-3 report. The images of the plots used in A-1 can be referenced from this appendix.

DAS732: Data Visualisation Assignment 1 Report

Aryan Singhal
IMT2022036
Aryan.Singhal@iiitb.ac.in

Rishi Patel
IMT2022041
Rishi.Patel@iiitb.ac.in

Pranav Kulkarni
IMT2022053
Pranav.Kulkarni@iiitb.ac.in

DATASET

Our dataset records the flight activity of customers, capturing details like the number of flights taken, distance traveled, points accumulated, and points redeemed for various months over different years of a fictitious Canadian Airline.

Customer Flight Activity Dataset

1) Loyalty Number: Unique identifier for each customer in the loyalty program.
2) Year: The year when the flight activity was recorded.
3) Month: The month corresponding to the recorded flight activity.
4) Total Flights: The number of flights taken by the customer in that particular month.
5) Distance: The total distance flown in kilometers by the customer during the month.
6) Points Accumulated: The number of loyalty points earned based on the flights taken and distance traveled.
7) Points Redeemed: The number of loyalty points the customer redeemed during that month.
8) Dollar Cost Points Redeemed: The equivalent dollar value for the points redeemed.

Customer Loyalty History Dataset

This dataset provides demographic information and details about customer loyalty program memberships, including customer value, enrollment, and cancellation history.

1) Loyalty Number: Unique identifier for each customer in the loyalty program.
2) Country: The country where the customer resides.
3) Province: The province or state where the customer resides.
4) City: The city where the customer resides.
5) Postal Code: The postal code of the customer's residence.
6) Gender: The gender of the customer (e.g., Male, Female).
7) Education: The highest level of education attained by the customer.
8) Salary: The annual salary of the customer (if available).
9) Marital Status: The marital status of the customer (e.g., Single, Married, Divorced).
10) Loyalty Card: The type or tier of the loyalty card held by the customer.
11) CLV: The customer lifetime value, which is an estimation of the total worth of the customer to the company.
12) Enrollment Type: The type of loyalty program the customer is enrolled in (e.g., Standard).
13) Enrollment Year: The year when the customer enrolled in the loyalty program.
14) Enrollment Month: The month when the customer enrolled in the loyalty program.
15) Cancellation Year: The year when the customer canceled their loyalty program membership (if applicable).
16) Cancellation Month: The month when the customer canceled their loyalty program membership (if applicable).

TASK

To gain valuable insights through visual exploratory analysis, we aim to accomplish the following five key tasks. These tasks will help us thoroughly understand the data and extract meaningful patterns:

1) Task-1: Identify pattern in enrollments.
2) Task-2: Analyze Factors Affecting Customer Lifetime Value (CLV)
3) Task-3: Analyze Flight Frequency Patterns Across Seasons and Years
4) Task-4: Relationship Between Distance Traveled and Total Flights
5) Task-5: Analyze Points Accumulation by Marital Status

DATASET PREPARATION AND ASSUMPTIONS MADE

In the customer loyalty history sheet, we have merged the Enrollment Month and Year columns to Enrollment Date .In the same way Cancellation Date has been made. Then in the customer flight activity also we have implemented the same thing with Year and Month columns. We also changed the datatype of appropriate columns to categorical.

The following assumptions were made during the data cleaning and preparation process:

- For the *Cancellation Year* and *Cancellation Month* columns, all missing (NaN) values were assumed to represent customers who have not yet canceled their loyalty card membership.
- In the *Salary* column, missing (NaN) values were found to belong to college-going students, and thus, their salary was set to zero for analysis purposes.

- While analyzing the *Salary* data, outliers and negative salary values were identified and removed to maintain the integrity of visualizations and insights.

Data Stories

Task-1: Identify Pattern in Enrollments

This task focuses on understanding how different factors (such as location and income) affect customer enrollments in a loyalty program. The key aim is to discover enrollment patterns based on geography and income levels.

Hypothesis 1:

Customers from more urbanized provinces (like Ontario/British Columbia/Quebec) have higher enrollments compared to customers from less urbanized provinces.

- **Explanation:** This hypothesis is based on the assumption that more urbanized provinces generally have more frequent travelers due to better access to transportation hubs like airports, and higher levels of business and tourism activities.
- **Rationale:**
  - Urbanized provinces like Ontario, British Columbia, and Quebec contain cities with international airports, high business activity, and large populations.
  - People in urban areas tend to travel more frequently for work and leisure, making them more likely to enroll in loyalty programs.
- **Implications:** Identifying this pattern would help airlines and loyalty programs focus their marketing and promotional efforts in these urban centers to capitalize on a larger customer base and increase enrollments.

Image 1: Province-Wise Enrollments in Canada (Choropleth Map)

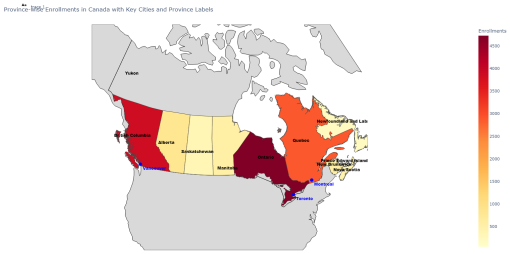


Fig. 1. Province-wise Enrollments in Canada

Type of Plot: Choropleth Map
Description: A choropleth map is a thematic map where areas (in this case, Canadian provinces) are shaded in different colors based on a statistical variable—in this case, the number of enrollments.

Color Scale: The color scale on the right indicates the enrollment range. Light yellow represents fewer enrollments (starting at around 500), and dark red signifies higher enrollments (up to around 4500). This gradient helps in visually distinguishing the enrollment density across provinces.

Details:

- **Urbanized Provinces:**
  - **Ontario:** Highlighted in a deep red, indicating the highest number of enrollments, especially around key cities such as Toronto.
  - **British Columbia:** Also appears in a darker shade, suggesting high enrollments centered around urban areas like Vancouver.
  - **Quebec:** Displays a medium to dark orange color, indicating a moderate to high level of enrollments, particularly near Montreal.
- **Less Urbanized Provinces:**
  - Yukon, Saskatchewan, Manitoba, and Newfoundland and Labrador are shaded in lighter colors, indicating relatively lower enrollments.

Analysis: The map shows a clear pattern where more urbanized provinces have darker shades, indicating higher enrollments. Urban centers (e.g., Toronto in Ontario, Vancouver in British Columbia, and Montreal in Quebec) are key drivers of this enrollment distribution. Less urbanized provinces, characterized by smaller populations and fewer large cities, show significantly lower enrollments.

Image 2: Customer Distribution by Province and City (Treemap)

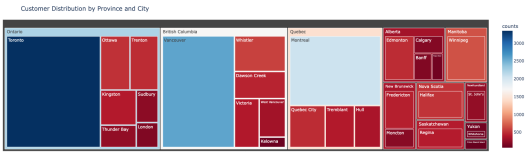


Fig. 2. Customer Distribution by Province and City

Type of Plot: Treemap
Description: A treemap is a data visualization that displays hierarchical data as nested rectangles. The size of each rectangle corresponds to the value it represents (here, the number of enrollments). The rectangles are grouped by categories (provinces) and subcategories (cities within those provinces).

Color Scale: The color scale on the right ranges from light to dark blue for Ontario and various shades of red for other provinces. The color intensity is proportional to the count of enrollments. Darker shades represent higher counts.

Details:

- **Ontario:** The largest and darkest rectangle belongs to Toronto, indicating the city has the highest enrollment count in the province. Other cities like Ottawa, Trenton, Kingston, and Thunder Bay show smaller rectangles, suggesting they have fewer enrollments.

- **British Columbia:** Vancouver is prominently represented, showing a high number of enrollments. Other cities like Whistler and Victoria appear with smaller rectangles, indicating lower customer distribution.
- **Quebec:** Montreal stands out as the main city for enrollments, while cities like Quebec City and Tremblant are smaller in size, reflecting fewer enrollments.
- **Other Provinces:** Cities in less urbanized provinces, such as Fredericton in New Brunswick, Regina in Saskatchewan, and St. John's in Newfoundland, have relatively smaller rectangles, indicating fewer enrollments.

Analysis: The treemap further supports the hypothesis. Larger rectangles for cities in urbanized provinces indicate a higher concentration of enrollments. For instance, Toronto and Vancouver, being major urban centers, dominate in their respective provinces, while smaller cities and those in less populated provinces have significantly smaller rectangles, representing lower enrollment counts.

Connecting the Hypothesis to the Plots

Both plots provide complementary evidence to support the hypothesis that more urbanized provinces have higher enrollments:

- **Choropleth Map:** Visually confirms that provinces with major urban centers (e.g., Ontario, British Columbia, Quebec) show higher enrollment levels. The darker shades over these provinces indicate a larger number of enrollments.
- **Treemap:** Offers a detailed breakdown of enrollment distribution at the city level within each province. It shows that within more urbanized provinces, key cities like Toronto, Vancouver, and Montreal dominate the enrollment numbers. In contrast, cities in less urbanized provinces are smaller in size and have lighter shades, indicating lower customer counts.

Summary of Insights

- **Urbanization Impact:** Both visualizations highlight that urban centers in provinces like Ontario, British Columbia, and Quebec drive higher enrollments. This suggests a correlation between urbanization and customer concentration.
- **Provincial Patterns:** Less urbanized provinces have fewer enrollments, as indicated by both the lighter colors in the choropleth map and the smaller rectangles in the treemap. This implies that factors such as population density, access to educational institutions, and urban infrastructure might significantly influence enrollment numbers.

By using both the choropleth map and the treemap, we gain a comprehensive view of how enrollments are distributed across Canada's provinces and cities, reinforcing the hypothesis about the relationship between urbanization and enrollment levels.

Hypothesis 2:

Income level affects the type of Loyalty Card enrolled in. Higher-income people are more likely to enroll in Aurora than lower-income households.

- **Explanation:** This hypothesis explores the relationship between income levels and loyalty card enrollment.
- **Rationale:**
  - Higher-income customers may travel more frequently and prefer premium services that offer added benefits.
  - Lower-income households might travel less frequently and be more price-sensitive, choosing standard-tier loyalty cards.
- **Implications:** Understanding this trend would enable loyalty programs to target marketing based on income segments.

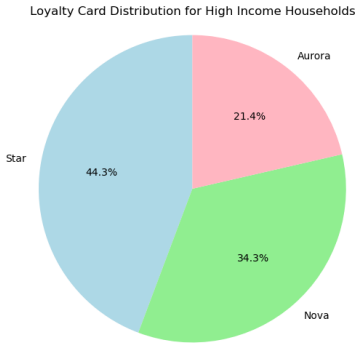


Fig. 3. Loyalty Card Distribution for High Income Households

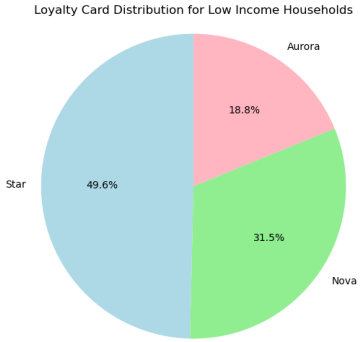


Fig. 4. Loyalty Card Distribution for Low Income Households

Loyalty Card Distribution by Income Level

The pie charts represent the distribution of loyalty cards (**Star**, **Nova**, and **Aurora**) across different income levels (**Low**, **Middle**, and **High**).

Low-Income Households:

- The majority (**49.6%**) of low-income households hold the *"Star"* loyalty card.

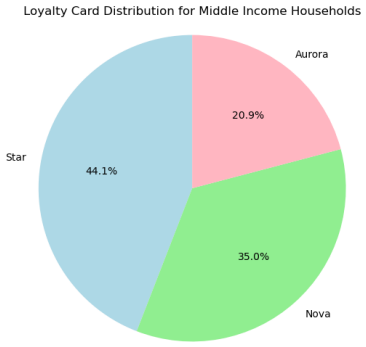


Fig. 5. Loyalty Card Distribution for Middle Income Households

- *"Nova"* follows with **31.5%**.
- The smallest share (**18.8%**) is for the *"Aurora"* card, indicating that this segment is less likely to opt for Aurora.

Middle-Income Households:

- A similar pattern is seen here, with *"Star"* still being the most popular (**44.1%**), but slightly less than in low-income households.
- *"Nova"* shows an increase (**35.0%**) compared to low-income households.
- *"Aurora"* sees a modest rise in participation, with **20.9%** of middle-income households enrolled.

High-Income Households:

- The *"Star"* card remains popular, but its share drops to **44.3%**.
- *"Nova"* has the highest proportion here (**34.3%**), with many high-income households opting for this program.
- *"Aurora"* has the largest share in this income bracket (**21.4%**), though still lower compared to the other cards.

Inference and Conclusion:

The graphs reveal some interesting trends:

- **"Aurora" Enrollment:** The hypothesis partially holds true. While there is a slight increase in *"Aurora"* enrollment as income increases (from **18.8%** in low-income to **21.4%** in high-income), the shift is not drastic. The difference between low- and high-income groups in Aurora enrollment is only about **2.6%**.
- **"Star" Loyalty Card:** Surprisingly, *"Star"* remains the most popular card across all income levels, even among higher-income groups. This may indicate that even wealthier individuals prefer practical, everyday benefits offered by this card.
- **"Nova" Loyalty Card:** *"Nova"* seems to be a preferred choice for middle- and high-income households, possibly offering a balanced mix of benefits that appeal to both segments.

Conclusion: While income level does have some effect on the type of loyalty card enrolled in, as the hypothesis suggests, the difference is not significant enough to conclude that high-income households overwhelmingly prefer *"Aurora"*. Instead, they appear to distribute their preferences across all cards, with *"Star"* and *"Nova"* remaining popular.

Task-2: Analyze Factors Affecting Customer Lifetime Value (CLV)

This task aims to determine which factors contribute most to a customer's lifetime value (CLV).

Hypothesis 1:

Customers with premium loyalty cards (e.g., Aurora) have higher CLV compared to customers with standard loyalty cards.

- **Explanation:** This hypothesis is based on the assumption that premium loyalty cardholders are more valuable to the business.
- **Rationale:**
  - Premium loyalty programs typically attract frequent travelers who spend more on airline tickets, upgrades, and other services.
- **Implications:** Identifying the impact of loyalty card tiers on CLV would help airlines prioritize high-value customers with tailored offerings and services.

Image 1: Radar Chart – Total CLV

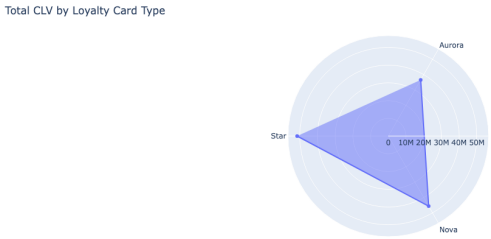


Fig. 6. Total CLV

- **Description:**
  This radar chart presents the total CLV (Customer Lifetime Value) for each loyalty card type: *Aurora*, *Nova*, and *Star*.
- **Analysis:**
  - *Aurora* has the lowest total CLV despite being the highest tier card. This could indicate that it generates the least value for the airline.
  - *Nova* follows Aurora but significantly ahead in total CLV.
  - *Star*, generates the highest CLV for the airline.
  - The chart highlights how *Aurora* customers provide a disproportionate share of value, suggesting they are the most frequent or high-spending customers.
- **Conclusion:**
  The total CLV is highest for *Star cardholders* , confirming that this group generates substantial revenue.

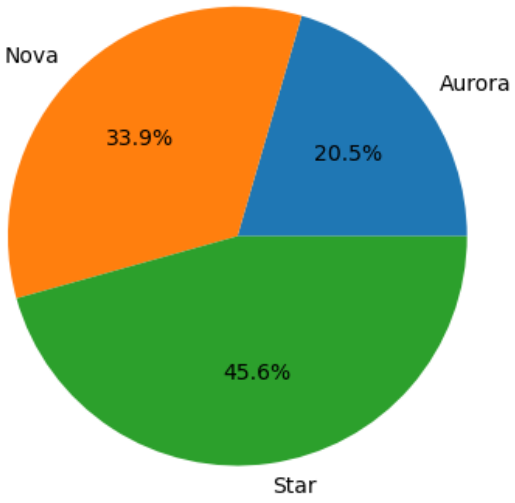


Fig. 7. Enrollments Categorized

Image 2: Pie Chart – Enrollments Categorized :

- **Description:**
  This pie chart displays the percentage distribution of customer enrollments across three loyalty card types: *Aurora*, *Nova*, and *Star*.
- **Analysis:**
  - *Star* loyalty card enrollments represent the largest portion at **45.6%**, showing that the majority of customers opt for the basic or entry-level loyalty program.
  - *Nova* accounts for **33.9%** of the enrollments, indicating a significant middle-tier customer base.
  - *Aurora* enrollments are the smallest, comprising **20.5%** of the customer base.
  - The relatively low number of *Aurora* enrollments along with the lowest total CLV from this group (as seen in the radar chart), suggests that *Aurora* may not generate high value to the airline despite being the highest - tier card.
- **Conclusion:**
  The *enrollment distribution* is skewed toward the *Star* card. This along with the highest total CLV might indicate that *Star* might be generating the highest per capita CLV for the airline.

Image 3: Line Chart – Average CLV :

- **Description:**
  This line chart shows the average CLV for customers in each loyalty card tier: *Aurora*, *Nova*, and *Star*.
- **Analysis:**
  - The *Aurora* tier has the highest average CLV, exceeding **10,500** per customer. This reinforces the notion that *premium customers contribute more value per person*.
  - *Nova* cardholders have a lower average CLV, around **9,000** per customer, which represents a middle-tier group that provides moderate value.
  - *Star* cardholders have the lowest average CLV, slightly above **7,000**, indicating that this group consists of less frequent travelers or those with lower spending.
  - The downward trend in the chart indicates a *clear relationship between loyalty card type and average CLV*, with premium customers (Aurora) contributing significantly more on a per-customer basis.
- **Conclusion:**
  The *average CLV decreases* as you move from *Aurora* to *Star*, showing that *premium loyalty programs attract the most valuable customers*. *Star*, while having the largest customer base, generates the least value per individual.

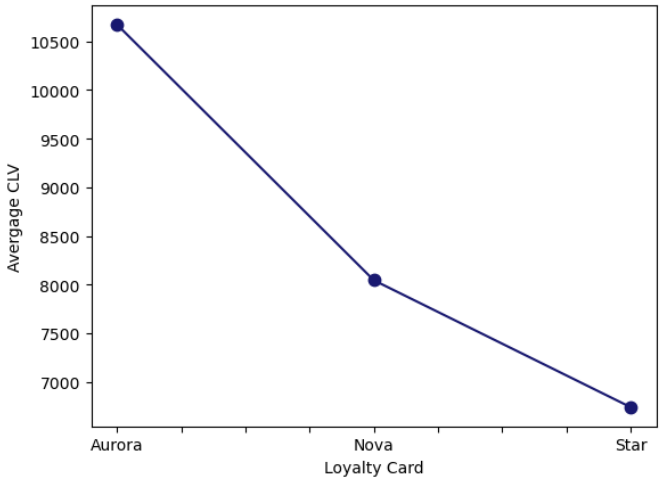


Fig. 8. Average CLV

Summary Across All Graphs

- *Aurora loyalty cardholders* are the most valuable group in terms of average CLV.
- *Star loyalty cardholders*, while being the largest group, contribute the least both in terms of average CLV while contributing the highest total CLV.
- The *Nova loyalty card* sits in the middle, providing moderate value on both fronts.

Each graph adds a layer of understanding, showing how *loyalty program tier* is a major determinant of CLV, with premium programs (Aurora) yielding the highest returns for the airline.

Hypothesis 2:

There is a positive correlation between salary and CLV.

- **Explanation:** This hypothesis explores the connection between a customer's income level and their CLV.
- **Rationale:**

  - Higher-income customers are more likely to have a higher CLV, as they can afford more frequent travel and premium services.
- **Implications:** This analysis would help airlines identify high-income customers as key drivers of revenue.

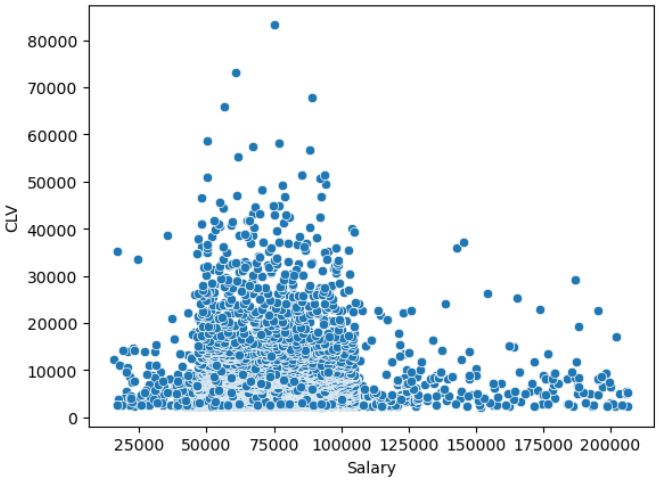


Fig. 9. Scatter Plot of CLV

Image 4: Scatter Plot of CLV vs. Salary:

- **Observation:**
  The scatter plot visualizes the relationship between Salary (x-axis) and CLV (y-axis). There is no apparent upward or downward trend, and the data points are widely dispersed. The majority of points are clustered in the salary range between $25,000 and $100,000, with CLV values largely spread across the range.
- **Insights:**
  - **No Strong Correlation:** There's no visible pattern indicating that higher salaries correspond to higher CLV. In fact, individuals with lower salaries show a wide range of CLV values, including some of the highest CLVs in the dataset.
  - **Outliers:** A few outliers exist where customers with extremely high CLV values ($50,000+) appear, but these are relatively rare and not tied to high salary levels.
- **Conclusion:**
  There is **no clear relationship** between salary and CLV, as customers with lower salaries can have high CLV, and higher salaries don't guarantee higher CLV. The hypothesis of a positive correlation is **not supported by the scatter plot**.

Image 5: Correlation Heatmap between CLV and Salary:

- **Observation:**
  The heatmap displays a numerical correlation between the two variables: Salary and CLV. The correlation coefficient shown is **-0.03**, indicating a very weak negative correlation between Salary and CLV.
- **Insights:**

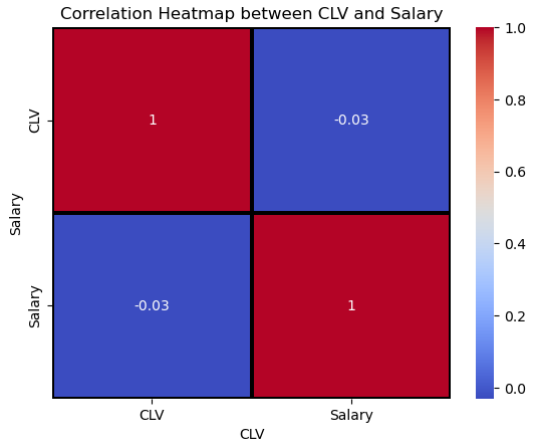


Fig. 10. Correlation Heatmap between CLV and Salary

  - **Negative Correlation:** A correlation of **-0.03** suggests that any relationship between Salary and CLV is both minimal and negative. This implies that as salary increases, CLV slightly decreases, though this relationship is not strong enough to be statistically significant.
  - **Lack of Dependence:** The weak negative correlation supports the idea that salary has little to no effect on determining a customer's lifetime value to the airline.
- **Conclusion:**
  The **correlation heatmap further invalidates the hypothesis**. The negative correlation confirms that salary is not a major factor in predicting CLV. The airline cannot rely on salary as an indicator for customer value, as there are likely other variables influencing CLV more significantly.

Task-3: Analyze Flight Frequency Patterns Across Seasons and Years

This task focuses on understanding how travel patterns vary based on seasonality and how customers redeem points during these times.

Hypothesis 1:

Customers are more likely to fly in certain seasons (e.g., summer or winter) compared to others.

- **Explanation:** Certain seasons tend to have higher travel demand due to holidays, vacations, or specific weather conditions that encourage travel.
- **Rationale:**
  - Summer and winter seasons are popular for vacations and holiday travel.
- **Implications:** Understanding seasonal flight frequency patterns would allow airlines to optimize flight scheduling, pricing strategies, and marketing campaigns to capitalize on these peak periods.

Image 1: Total Flights by Season (Bar Chart for 2017 vs. 2018):

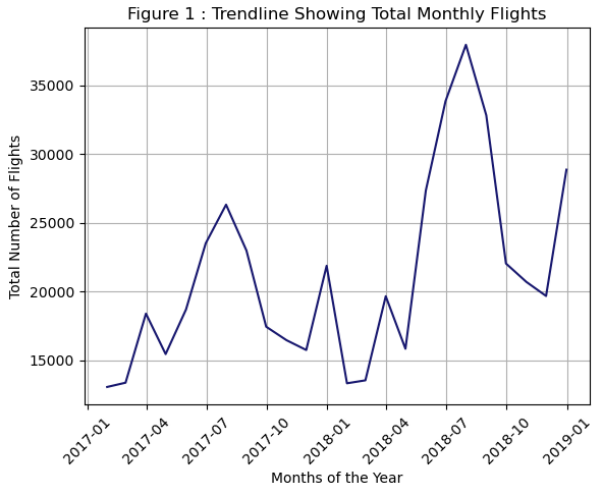


Fig. 11. Total Monthly Flights

Observations from the Graph:

- **Summer** is clearly the peak season, with the highest total flights for both 2017 and 2018. The number of flights in summer 2018 is notably higher than in 2017.
- **Winter** shows an increase in flights from 2017 to 2018, though not as dramatic as summer.
- **Fall** has moderate flight activity, while **Spring** appears to have relatively lower flight activity.
- There's a consistent seasonal pattern with an increase from Spring to Summer, then a decrease from Summer to Winter.

Hypothesis Validation:

- The data supports the hypothesis that customers are more likely to fly in **summer**. It also shows increased travel during **winter**, likely influenced by holiday travel.
- **Spring** and **Fall** seem to be less busy periods, which could correspond to off-peak times for vacations or fewer national holidays.

Validation: The hypothesis holds, particularly for **summer** and **winter**, which show more flights than other seasons.

For further analysis, the seasons of Canada and their months are the following :

- ***Summer*** : June - August
- ***Fall*** : September - November
- ***Winter*** : December - February
- ***Spring*** : March - May

Image 2: Total Flights by Season :

Observations from the Graph::

- The **summer months (June-August)** show the highest spikes in flights, reinforcing the popularity of summer travel.
- There's a noticeable drop after summer, followed by a spike again in **December**, likely due to holiday travel, before tapering off at the start of the next year.

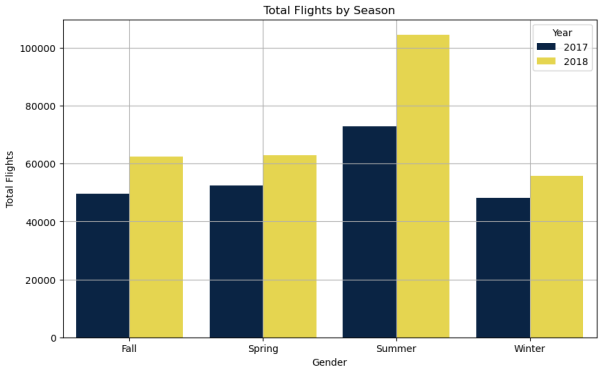


Fig. 12. Total Flights per Season

- The pattern is cyclical, with clear peaks in summer and winter, confirming the seasonal nature of travel.

a) Implications and Hypothesis Validation::

- **Summer** continues to dominate, but the **winter peak in December** is also clear.
- **Off-peak periods**, like early spring (April) and late fall (October-November), see dips in flights, confirming seasonal trends.

Validation: The hypothesis holds across both years, as flight activity rises sharply during **summer** and **winter months**. This aligns with vacation schedules and holiday travel.

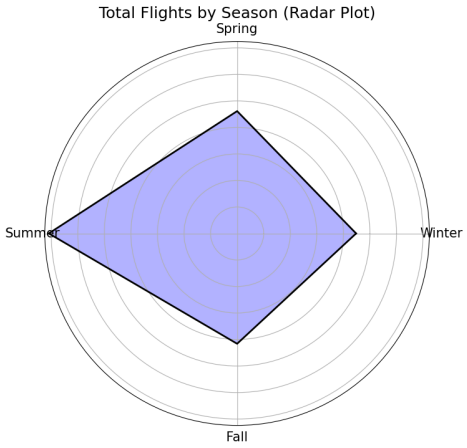


Fig. 13. Total Monthly Flights

Image 3: Polar Plot - Total Flights by Season:

Observations from the Graph:

- The diamond shape emphasizes that **Summer** dominates in terms of flight frequency, with **Winter** following. **Fall** and **Spring** are more comparable, though Spring appears slightly less active.

Hypothesis Validation:

- The polar plot strengthens the visual understanding of flight distribution by season, confirming that **summer** and **winter** are the peak seasons.

- It visually reinforces the idea that there's a distinct increase in travel during these periods, supporting the hypothesis.

Conclusion : The hypothesis is well-supported by the data:

- **Summer** is clearly the most popular season for flights, followed by **winter**, which aligns with holiday travel.
- **Spring** and **fall** show moderate to low activity, possibly due to fewer vacation periods and holidays.

This seasonality analysis could guide the airline's strategic decisions on pricing, marketing, and flight scheduling to optimize for these peak travel times.

Hypothesis 2:

Customers tend to redeem more points in the off-season (e.g., fall) compared to peak seasons.

- **Explanation:** This hypothesis assumes that customers prefer to redeem loyalty points during off-peak times when flights are less in demand and more availability is offered for point redemptions.
- **Rationale:**
  - Airlines may offer more availability for point-based bookings during off-peak seasons, when demand for paid tickets is lower.
- **Implications:** By understanding the seasonal point redemption patterns, airlines can adjust their loyalty rewards strategies to balance demand and increase seat occupancy.

Graph Analysis:

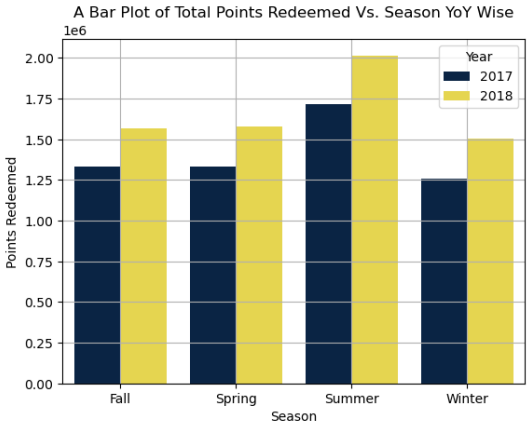


Fig. 14. Bar Plot of Total Points Redeemed

The bar plot titled *"Total Points Redeemed Vs. Season YoY Wise"* compares the total loyalty points redeemed across four seasons—Fall, Spring, Summer, and Winter—over the years 2017 and 2018.

Observations from the Graph:

1) **Seasonal Variations:**
   - **Summer** shows the highest number of points redeemed in both 2017 and 2018.
   - **Spring** follows closely behind Summer in terms of redemption rates.
   - **Winter** has the lowest redemption rate in both years, with **Fall** also showing lower redemption rates compared to Summer and Spring.
   - **Summer 2018** saw the largest amount of points redeemed, peaking at around 2 million.
2) **Yearly Comparison:**
   - Across all seasons, points redeemed in 2018 are higher than in 2017. This is evident from the yellow bars being taller than the blue bars across every season.
   - The increase is particularly pronounced in **Summer**, where the gap between 2017 and 2018 is the largest.

Inferences and Conclusions:

Hypothesis Evaluation:

- The graph **does not support Hypothesis 2**, which stated that customers would redeem more points in off-peak seasons like Fall. Instead, **Summer** and **Spring**, typically peak travel seasons, have the highest point redemptions.
- **Winter** has the lowest point redemptions, indicating that fewer customers are redeeming points during the colder months, likely due to fewer travel plans or a lower perceived value in traveling during that season.

Customer Behavior Insight:

- Customers appear to redeem more points during high-demand periods such as **Spring and Summer**, possibly for vacations or trips when airline fares are more expensive, maximizing the value of their loyalty points.
- This behavior indicates that loyalty points are most valued when travel costs are higher, and customers are keen to save money on flights during these busy travel months.

Conclusion:

While the initial hypothesis suggested that off-peak seasons like Fall would see higher point redemptions, the data reveals that customers are actually redeeming more points during peak travel seasons like **Spring** and **Summer**. This may reflect a desire to maximize the value of points by using them when travel demand is high and flight prices are more expensive. Airlines could address this pattern by incentivizing off-peak redemptions, encouraging more balanced year-round loyalty program engagement.

TASK-4: RELATIONSHIP BETWEEN DISTANCE TRAVELED AND TOTAL FLIGHTS

This task aims to understand the relationship between the number of flights a customer takes and the total distance they travel.

Hypothesis 1:

There is a positive correlation between the number of flights taken and the total distance traveled.

- **Explanation:** It is intuitive to assume that customers who take more flights also accumulate more total distance traveled.
- **Rationale:**
  - Frequent travelers, especially those who fly internationally or across long distances, would cover substantial distances over time.
- **Implications:** Airlines could use this data to identify frequent flyers and provide targeted rewards or upgrades.

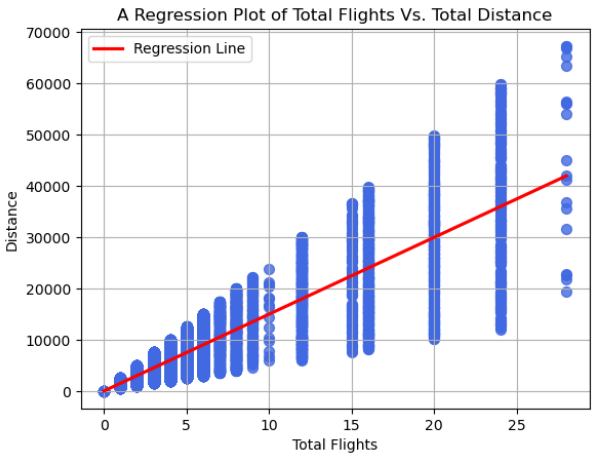


Fig. 15. Regression Plot

Graph Description:

- **X-Axis (Horizontal)**: Total Flights — This represents the number of flights taken by the customer.
- **Y-Axis (Vertical)**: Distance — This represents the total distance traveled (likely in miles or kilometers).
- **Data Points (Blue Dots)**: Each dot represents a customer or an observation in the dataset, plotting the total number of flights they have taken against the total distance they have traveled.
- **Regression Line (Red)**: This line shows the trend of the relationship between total flights and distance traveled. It is the best-fit line, indicating the overall direction of the data.

Key Insights::

Strong Positive Correlation:

- The graph clearly shows a positive correlation between the number of flights and the distance traveled. This can be seen as the data points generally move upwards as the number of flights increases.
- The red regression line rises consistently from left to right, demonstrating that with more flights, the distance traveled tends to increase.

Correlation Coefficient (0.908):

- The correlation coefficient value of 0.908 suggests a strong positive relationship between the two variables. In simpler terms, this high value means that there is a significant and direct connection between the number of flights taken and the total distance traveled.
- A correlation value close to +1 indicates that as the number of flights increases, the distance increases as well. This fits the assumption of frequent flyers accumulating miles.

Clusters of Data:

- In the graph, there are vertical clusters of blue points at specific flight counts (e.g., 5, 10, 15 flights). These clusters suggest that many passengers might be taking similar numbers of flights, but the distance they travel can vary, likely depending on whether they fly short-haul or long-haul routes.
- For example, someone with 10 flights might travel 20,000 miles, while another person with the same number of flights might travel only 10,000 miles, reflecting differences in the nature of the flights.

Outliers:

- There are some observations where the distance traveled is significantly higher or lower for the same number of flights, particularly around the 25-flight mark. These outliers could represent specific travelers who either take long international flights or short domestic trips, impacting the total distance they accumulate despite taking the same number of flights.

Linear Trend:

- The regression line demonstrates a near-linear relationship, meaning that on average, for every additional flight taken, the total distance traveled increases at a relatively consistent rate. However, there is some variability around this line, especially in higher flight counts.

Conclusion: The analysis supports the hypothesis that there is a strong positive correlation between the total number of flights taken and the distance traveled. The correlation coefficient of 0.908 confirms that as customers take more flights, they generally tend to travel longer distances overall. This insight reflects consistent travel patterns where frequent flyers, over time, accumulate more mileage. The linear trend suggests that airlines can expect this relationship to hold across different customer segments, though there is variability depending on the nature of the flights (short-haul vs. long-haul).

In terms of customer behavior, frequent travelers are often either flying frequently on short routes or taking fewer but longer international trips. Understanding this relationship can help airlines optimize their loyalty programs by offering different benefits based on the flight patterns (e.g., frequent short-haul travelers vs. long-distance flyers).

Hypothesis 2:

Frequent flyers tend to fly shorter distances, while infrequent flyers opt for longer trips.

- **Explanation:** This hypothesis is based on the assumption that frequent flyers, such as business travelers, often take shorter domestic or regional flights, while infrequent

flyers might save their flights for longer international trips.

  - **Frequent Flyers:**
    * Typically consist of business travelers or those who fly regularly within a region. Since they may take many short-haul flights for work or personal commitments, they accumulate a high number of total flights but their average travel distance per flight tends to be shorter.
    * Examples include people who fly between cities in the same country regularly.
  - **Infrequent Flyers:**
    * These individuals fly less often but are more likely to take long trips, such as vacations or international travel. Each flight is potentially a long-haul trip covering more miles, resulting in fewer total flights but with higher travel distances per flight.
    * For instance, people who fly only a few times a year but travel internationally for holidays.
- **Rationale:**
  - Business travelers might regularly fly short distances, while infrequent travelers might take fewer but longer trips.
- **Implications:** This insight could help airlines better understand the needs of frequent vs. infrequent travelers and customize their loyalty offerings accordingly.

Classification of Flyers: The classification function breaks flyers into three categories:-
Frequent, Moderate, and Infrequent. The decision is based on the number of flights a traveler takes, with threshold values set by year:

- **Frequent Flyers:** Above the 75th percentile of total flights.
- **Infrequent Flyers:** Below the 25th percentile of total flights.
- **Moderate Flyers:** Between the 25th and 75th percentile.

The function uses thresholds that differ by year to ensure fairness as flight habits and data distributions may change over time.

Graph Analysis:

1. Stripplots (Travel Distance by Flyer Frequency in 2017 and 2018):

- The first two graphs for 2017 and 2018 show stripplots comparing travel distances across frequent, moderate, and infrequent flyers.
- **2017:**
  - Frequent and moderate flyers display a similar distribution of travel distances, reaching up to 17,500 miles.
  - Infrequent flyers show a slightly lower range of distances, with fewer flyers hitting the upper ranges.
- **2018:**

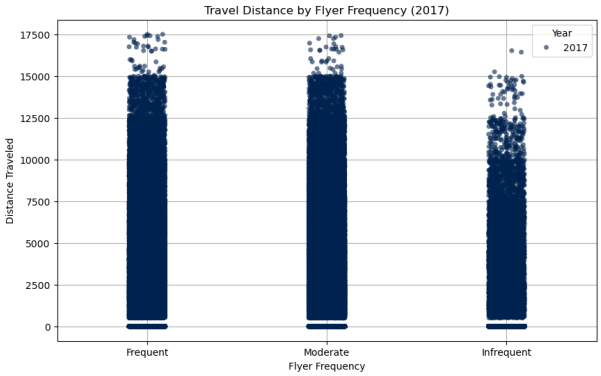


Fig. 16. Travel Distance by Flyer(2017)

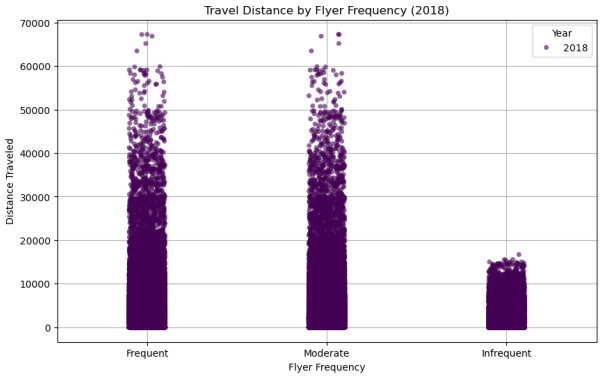


Fig. 17. Travel Distance by Flyer(2018)

  - A noticeable increase in travel distance for all categories, especially for frequent flyers, some of whom reached up to 70,000 miles.
  - Infrequent flyers show a clear distinction, with shorter travel distances on average compared to 2017.

Insight: The assumption that infrequent flyers fly longer distances is not entirely validated in the 2018 data, as frequent flyers show a larger spread in terms of travel distance. In both years, frequent and moderate flyers tend to travel similar distances, with infrequent flyers covering less ground, particularly in 2018.

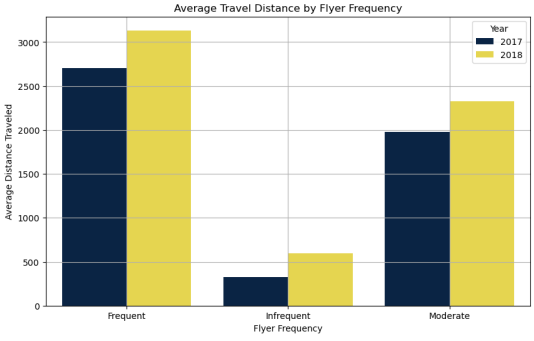


Fig. 18. Average Travel Distance

2. Bar Chart (Average Travel Distance by Flyer Frequency):

- The bar chart shows the **average** distance traveled by each flyer category in both 2017 and 2018.
- In 2017, frequent flyers averaged around 2,500 miles, while moderate and infrequent flyers averaged less.
- In 2018, the distance traveled by frequent and moderate flyers increased significantly, surpassing 3,000 miles, while infrequent flyers saw a more modest increase.

b) Insight:: Frequent and moderate flyers saw a growth in travel distance, likely indicating that even frequent flyers may be taking longer trips than previously thought. The increase in average distance across all categories in 2018 suggests an overall trend toward longer trips, but frequent flyers still dominate in terms of the number of flights and travel distance.

Conclusion:

The hypothesis stating that **frequent flyers tend to fly shorter distances** is partially contradicted by the data. In both 2017 and 2018, frequent flyers have a wider range of travel distances and average higher travel distances than infrequent flyers, especially in 2018. While it's true that frequent flyers likely include individuals who take regular short-haul flights, the data suggests they also take long-haul flights.

TASK-5: ANALYZE POINTS ACCUMULATION BY MARITAL STATUS

This task explores how marital status affects points accumulation and redemption patterns in the loyalty program.

Hypothesis 1:

Single customers accumulate more points compared to married or divorced customers, as they travel more frequently.

- **Explanation:** The assumption here is that single individuals, particularly those with fewer familial obligations, might have more flexibility to travel frequently, allowing them to accumulate more loyalty points over time.
- **Rationale:**
  - Single individuals may have fewer responsibilities (e.g., children, family obligations) that limit their ability to travel, allowing them to accumulate more loyalty points.
  - Singles might prioritize travel as a lifestyle choice, making frequent trips for leisure or business, which further increases their loyalty points accumulation.
  - Married or divorced customers, particularly those with families, may face time and financial constraints that limit their ability to travel frequently, resulting in fewer opportunities to accumulate points.
- **Implications:** Understanding these differences in travel patterns can help airlines tailor loyalty programs and marketing campaigns. For example, single customers may be more responsive to promotions that encourage frequent travel, while married customers might be interested in family-oriented packages or offers.

Graph 1: Points Accumulation by Marital Status (2017 vs. 2018)

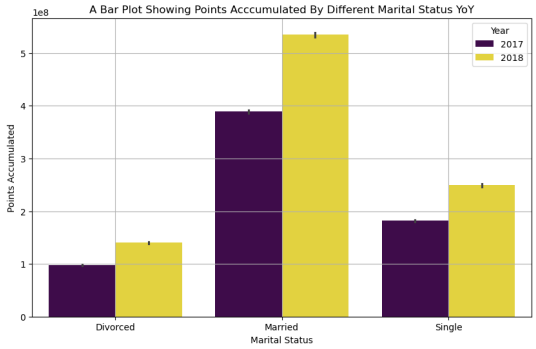


Fig. 19. Points Accumulated by different Marital Status YoY

Description of the Graph:

- **X-Axis (Horizontal):** Marital Status (Divorced, Married, Single) — This shows three different customer segments based on their marital status.
- **Y-Axis (Vertical):** Points Accumulated — The total points accumulated by each group, presumably representing loyalty points or miles over two years (2017 and 2018).

Bar Colors:

- **Purple (2017)**
- **Yellow (2018)**

Key Insights:

- **Married Customers Accumulate the Most Points:**
  - Married customers accumulated the highest points in both years, with a significant increase from 2017 to 2018.
- **Single Customers Show Steady Growth:**
  - Single customers accumulated fewer points than married customers but more than divorced customers. There is a notable increase in points from 2017 to 2018.
  - This suggests that single customers loyalty and travel patterns are increasing year over year.
- **Divorced Customers Accumulate the Fewest Points:**
  - Divorced customers accumulated the least number of points in both years, with only a slight increase from 2017 to 2018.
- **Year-on-Year Growth:**
  - Across all marital statuses, there is growth in points accumulated from 2017 to 2018. This could be due to the ***2018 Enrollment Promotion***, indicating that the promotion could be a success.

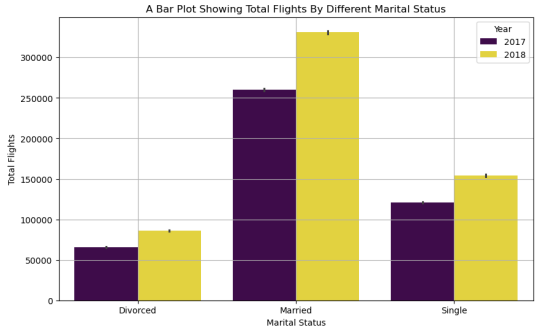


Fig. 20. Total Flights by different Marital Status

Graph 2: Total Flights by Marital Status (2017 vs. 2018)

Description of the Graph::

- **X-Axis (Horizontal):** Marital Status (Divorced, Married, Single).
- **Y-Axis (Vertical):** Total Flights — Total flights taken by each group over the years 2017 and 2018.

Bar Colors:

- **Purple (2017)**
- **Yellow (2018)**

Key Insights::

- **Married Customers Take the Most Flights:**
  - Similar to points accumulation, married customers take the most flights in both years.
  - In 2017, they took approximately 250,000 flights, which increased to over 300,000 in 2018. This confirms that married individuals not only accumulate more points but also travel more frequently.
- **Single Customers Increase Flights in 2018:**
  - Single customers show a marked increase in the total number of flights from 2017 to 2018. In 2017, they took around 100,000 flights, increasing to around 150,000 in 2018.
  - This growing trend could reflect a rise in travel frequency for this group.
- **Divorced Customers Have the Least Number of Flights:**
  - Divorced customers take the fewest flights, similar to their points accumulation pattern.
  - The number of flights remains stable, hovering around 75,000 across both years.

Conclusion:

Hypothesis Validation: The hypothesis that single customers accumulate more points compared to married or divorced customers due to traveling more frequently is not supported by the data. In both years, married customers accumulated the most points and took the most flights. Single customers showed a steady increase in both points and total flights but still lagged behind married customers.

Key Finding: Married individuals tend to travel more frequently and accumulate more points compared to single or divorced individuals, which could be due to a combination of personal and family travel. Single customers are catching up with married customers in terms of points and flights, but they have not yet surpassed them.

Implications for Airline Loyalty Programs: Airlines might benefit from tailoring their loyalty programs to recognize the different travel patterns based on marital status. For instance, offering family travel rewards for married individuals or exclusive benefits for single customers could encourage higher engagement.

Hypothesis 2:

Married customers are more likely to redeem points compared to single customers, indicating a higher propensity for family trips.

- **Explanation:** This hypothesis suggests that while single customers may accumulate more points, married customers (especially those with families) are more likely to redeem their points for vacations or family trips.
- **Rationale:**
  - Married individuals or those with families may use their points to offset the cost of family vacations, which tend to be more expensive than solo trips, leading to higher redemption rates among married customers.
  - Families are more likely to plan trips in advance and take advantage of loyalty points to reduce costs for multiple travelers.
  - Single customers might accumulate points for future redemptions or be more selective about when and how they redeem points.
- **Implications:** Airlines could offer specific family travel packages, discounted points redemptions for groups, or loyalty bonuses for family trips. Understanding the redemption behavior of married vs. single customers allows airlines to create more targeted and personalized loyalty offers.

Graph Analysis:

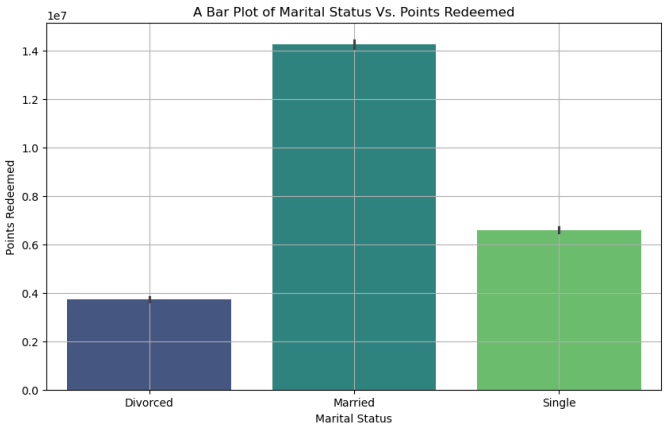


Fig. 21. Marital Status Vs. Points Redeemed

1. Bar Plot: Marital Status vs. Points Redeemed:

- **X-axis:** Marital status (Divorced, Married, Single).
- **Y-axis:** Points redeemed (measured in the range of millions).

Observation:

- Married individuals redeemed more points (close to 14 million) compared to both single (close to 6 million) and divorced individuals (about 4 million).

Inferences::

- Married customers redeemed significantly more points than their single and divorced counterparts.
- This suggests that married individuals may travel more frequently or for larger trips (potentially family vacations).

Conclusion::

- This graph strongly supports the hypothesis that married individuals are more likely to redeem points.
- The large gap between married individuals and the other groups indicates that marriage may correlate with higher loyalty program engagement, possibly driven by family-related travel.

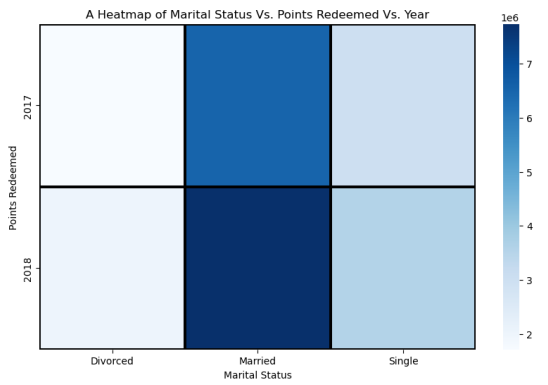


Fig. 22. Heatmap of Marital Status Vs. Points Redeemed Vs. Year

2. Heatmap: Marital Status vs. Points Redeemed vs. Year:

Structure:

- **X-axis:** Marital status (Divorced, Married, Single).
- **Y-axis:** Year (2017 and 2018).
- **Color Gradient:** The darker the color, the more points redeemed. The color intensity shows differences across marital status and year.

Inferences::

- **2017:** Married individuals had the highest points redemption, with the heatmap showing a relatively strong blue color. Single individuals redeemed fewer points, and divorced individuals had minimal point redemption.
- **2018:** Married individuals again led with the highest points redeemed, even higher than in 2017. Single and divorced individuals show lighter colors, meaning fewer points were redeemed compared to married individuals.

Conclusion::

- **Consistency in point redemption:** Over both years, married individuals consistently redeemed more points, indicating sustained loyalty and a stronger inclination towards airline travel.
- **Yearly increase for married individuals:** The points redeemed by married individuals increased from 2017 to 2018, which may further reinforce the idea that families or married people engage more with loyalty programs over time.

Overall Conclusion

Both the bar plot and the heatmap support the hypothesis. Married customers are significantly more likely to redeem points than single or divorced customers. The higher number of points redeemed by married individuals could suggest that they are frequent travelers, likely driven by family-related trips, group bookings, or vacations that leverage loyalty programs.

VISUALISATIONS

The following are the visualisation used to gain insights from the dataset :

1) Scatter Plots
2) Pie Charts
3) Heat Maps
4) Tree Maps
5) Line Plots
6) Bar Plots
7) Strip Plots
8) Reg Plot
9) Polar Plots
10) Choropleth Map

AUTHOR CONTRIBUTIONS

The tasks were initially discussed during a meeting, and we collaboratively determined them. For the visualizations, dashboards, and narratives, the following distribution of responsibilities was agreed upon:

- The project consists of five visualization tasks, report writing, and the creation of a dashboard using Streamlit.
- ***Aryan Singhal*** was responsible for completing Tasks 1 and 2.
- ***Pranav Kulkarni*** handled Tasks 4 and 5.
- ***Rishi Patel*** completed Task 3 , along with data preprocessing.
- Report writing and the development of the dashboard were carried out collaboratively by all team members.