|  |
| --- |
| Close-up image showing the leaf-sides of two oversized books side-by-side on a bookshelf, with additional books in soft focus background |
| Financial transaction datasets: Analysis |
| |  |  |  | | --- | --- | --- | | Bidisha Gogoi- 2023JULB01242 | 12/11/24 | Strategic decision making with Power BI | |

**Business Problem**

In today’s data-driven financial environment, organizations need to maintain robust client segmentation, identify fraudulent activities, and optimize financial performance metrics. The specific goals of this analysis are:

1. **Client Insights**: Understand customer segmentation based on their financial behaviors, such as credit score, income levels, and card usage.
2. **Fraud Detection**: Identify patterns and trends related to flagged cards (e.g., cards on the dark web).
3. **Financial Metrics Analysis**: Evaluate financial health metrics such as total debt and yearly income distribution to improve financial offerings.
4. **Time-Based Trends**: Assess how specific behaviours (e.g., PIN changes, card issuance) change over time.

This analysis enables better decision-making regarding fraud prevention, targeted marketing, and operational efficiency.

**Data Requirement**

To achieve the above objectives, the following data is required:

1. **Demographic Data**:

* Gender
* Age
* Location (latitude, longitude, address)

Helps in customer segmentation based on demographics.

Identifies geographic regions with higher concentrations of customers or fraud incidences.

Enables targeted marketing campaigns and location-specific fraud prevention strategies.

1. **Financial Data**:

* Yearly income
* Total debt
* Credit score
* Card types and brands
* **Yearly Income**:
* Assesses the financial capability of users, influencing product offerings such as credit limits or loan approvals.
* Identifies high-income groups for premium card services.
* **Total Debt**:
* Helps measure financial risk and debt management behaviour.
* Guides debt-restructuring or repayment plans for struggling customers.
* **Credit Score**:
* A critical metric for assessing creditworthiness and identifying potential defaulters.
* Enables creation of credit-score-based financial products or services.
* **Card Types and Brands**:
* Insights into customer preferences for specific card types (e.g., debit vs. credit) or brands.
* Guides partnerships or promotions with card brands.

1. **Behavioral Data**:

* Number of cards issued
* Year PIN was last changed
* Account open date
* Tracks customer engagement and potential overuse or misuse of financial products.
* Identifies cross-selling opportunities for additional services.
* Identifies potential security risks due to infrequent PIN updates.
* Promotes security awareness among customers.
* Tracks customer longevity and loyalty trends.
* Helps identify long-term customers for exclusive benefits.

1. **Fraud-Related Data**:

* Flagged cards (e.g., card\_on\_dark\_web status)
* Directly identifies fraud or security threats within the system.
* Enables proactive measures to mitigate risks, such as blocking or reissuing compromised cards.
* Supports fraud pattern analysis to improve predictive modeling and fraud detection systems.

**Data Collection and Understanding**

**1. Internal Databases**

* **Source**: Customer profiles and transactional systems managed by the organization.
* **Reason for Collection**:
  + These databases contain essential customer demographic, financial, and behavioral data, such as gender, age, location, yearly income, total debt, credit score, card types, and transaction history.
  + Internal systems provide accurate, up-to-date, and structured information that is critical for customer analysis, financial health assessment, and fraud detection.
  + The data enables the organization to make informed business decisions, such as creditworthiness evaluation and tailored product offerings.

**2. Fraud Detection Tools**

* **Source**: Third-party fraud detection software or internal monitoring systems that flag suspicious activity (e.g., cards listed on the dark web or unusual transactions).
* **Reason for Collection:**
* These tools provide insights into fraudulent activities, helping the organization protect its customers and assets.
* Data from these tools enables proactive measures, such as blocking compromised cards or implementing security protocols, improving customer trust and minimizing financial losses.

**3. External Sources (Optional)**

* **Source**: Publicly available data, such as geographic location mappings (for latitude and longitude) or economic indicators for contextual analysis.
* Reason for Collection:
* Geographic data can enhance demographic analysis by enabling visualizations of customer density and fraud hotspots.
* Economic indicators provide a macroeconomic context to financial behaviors, aiding in predictive modeling.

**Data Understanding**:

* + Initial data inspection includes identifying the number of records, features, and missing values.
  + Relationships among variables, such as income and credit score, need to be examined for meaningful insights.

**Data Validation**

Data validation is a critical step in ensuring the accuracy, consistency, and reliability of the dataset before conducting any analysis. It involves checking the data for errors, inconsistencies, and biases to improve the quality of insights derived from the analysis.

**Accuracy**:

* Ensures the data correctly represents the real-world entities or events it is supposed to capture.
* Example: A yearly\_income value must fall within a reasonable range and not have negative or extreme outliers unless justified.

**Consistency**:

* Ensures data fields are uniformly formatted and represented.
* Example: Card brands should be consistently named (e.g., "Visa" vs. "VISA").

**Completeness**:

* Ensures no critical data is missing for key attributes.
* Example: Missing credit\_score values could affect financial risk assessments.

**Bias Prevention**:

* Prevents skewed insights by ensuring that all demographic, financial, and fraud-related data are adequately represented.
* Example: Over-representation of a particular region in the dataset could bias location-based fraud analysis.

**Reliability**:

* Verifies that data originates from trustworthy and accurate sources.
* Example: Fraud-related data should be verified through credible fraud detection tools.

**Process of doing data validation:**

**1. Schema Validation**

* **What to Check**: Verify that data fields match the expected schema.
  + Data types (e.g., credit\_score should be numeric).
  + Constraints (e.g., acct\_open\_date should follow a valid date format).
* **Method**: Use tools like Python’s pandas or Power BI’s data modeling features to inspect column data types and constraints.

**2. Range Checks**

* **What to Check**: Verify that numerical values fall within reasonable ranges.
  + yearly\_income: Check for positive values and realistic ranges.
  + total\_debt: Ensure values are not greater than yearly\_income.
* **Method**: Use Python for filtering and flagging outliers using statistical measures like interquartile range (IQR).

**3. Missing Data**

* **What to Check**: Ensure critical attributes are not missing.
  + Key fields like client\_id, yearly\_income, credit\_score, and card\_on\_dark\_web should not have null values.
* **Method**: Use Python (pandas.isnull()), Power BI (error bars or filters), or SQL queries (WHERE column IS NULL) to detect missing values.

**4. Duplication Check**

* **What to Check**: Verify that unique identifiers, such as client\_id or card\_number, are not duplicated unless justified.
* **Method**: Use Python (pandas.duplicated()), Power BI, or SQL (GROUP BY client\_id HAVING COUNT(\*) > 1) to identify duplicates.

**5. Categorical Data Validation**

* **What to Check**: Ensure categorical fields have valid, consistent values.
  + Example: card\_brand should not have unexpected categories like “V1sa.”
* **Method**: Use Python’s value\_counts() or Power BI filters to inspect and correct unique values.

**6. Consistency Between Fields**

* **What to Check**: Ensure relationships between fields are logical.
  + Example: total\_debt should not exceed yearly\_income.
* **Method**: Use conditional checks in Python (df['total\_debt'] > df['yearly\_income']) or calculated columns in Power BI.

**7. Outlier Detection**

* **What to Check**: Identify extreme or unusual values that may skew analysis.
  + Example: A credit\_score above 850 or below 300 is invalid.
* **Method**: Use Python’s statistical tools (e.g., Z-scores or box plots) or Power BI conditional formatting to flag outliers.

**8. Validation of Fraud Data**

* **What to Check**: Ensure flagged cards (e.g., card\_on\_dark\_web) are verified and consistent.
  + Example: All flagged cards should have valid card\_number and associated metadata.
* **Method**: Compare flagged card data with logs or fraud detection tool outputs for consistency.

**9. Cross-Validation**

* **What to Check**: Cross-check data across related fields.
  + Example: A client with no income (yearly\_income = 0) should not have total\_debt > 0.
* **Method**: Use Python’s conditional filtering or Power BI’s calculated measures for validation.

**Data Cleaning (EDA)**

Data cleaning is an essential part of Exploratory Data Analysis (EDA) that ensures the dataset is ready for analysis and visualization. This process involves identifying and resolving issues such as missing values, duplicates, inconsistencies, and outliers, as well as enhancing the dataset with new derived features. Exploratory Data Analysis (EDA) is performed using Python, Tableau, Power BI, or SQL. Key steps include:

**Steps for Data Cleaning**

**1. Handle Missing Values**

* **Why**: Missing data can skew analysis and lead to inaccurate conclusions.
* **Steps**:
  + **Identify Missing Data**:
    - Use Python: df.isnull().sum() to identify columns with missing values.
    - Use Power BI: Apply a filter or conditional formatting to flag nulls.
  + **Treatment**:
    - **Demographic Data** (e.g., age, gender):
      * Replace missing values with the column mean, median, or mode (e.g., median age or most frequent gender).
    - **Financial Data** (e.g., yearly\_income, credit\_score):
      * Impute missing values using statistical methods (e.g., median for income).
      * Alternatively, use machine learning techniques to predict missing values based on correlated features.
    - **Fraud Flags** (card\_on\_dark\_web):
      * If missing, either drop these rows or label them as “Unknown” to retain the data.

**2. Remove Duplicates**

* **Why**: Duplicate records inflate counts, skew distributions, and distort relationships.
* **Steps**:
  + **Identify Duplicates**:
    - Python: df.duplicated(subset=['client\_id', 'card\_number'])
    - SQL:

SELECT client\_id, COUNT(\*)

FROM table\_name

GROUP BY client\_id

HAVING COUNT(\*) > 1;

* + **Remove Duplicates**:
    - Python: df.drop\_duplicates(subset=['client\_id', 'card\_number'], keep='first')
    - Power BI: Use the "Remove Duplicates" feature in the Query Editor.

**3. Feature Engineering**

* **Why**: Derived features add value by enabling deeper insights into the data.
* **Steps**:
  + **Create Bins**:
    - For credit\_score:
      * Define ranges such as:
        + Poor: 300–500
        + Fair: 501–700
        + Good: 701–850
      * Python Example:

bins = [0, 500, 700, 850]

labels = ['Poor', 'Fair', 'Good']

df['credit\_score\_category'] = pd.cut(df['credit\_score'], bins=bins, labels=labels)

* + - For yearly\_income:
      * Create income brackets (e.g., <$50k, $50k–$100k, >$100k).
  + **Calculate Debt-to-Income Ratio**:
    - Formula: debt\_to\_income = total\_debt / yearly\_income
    - Python Example:

df['debt\_to\_income'] = df['total\_debt'] / df['yearly\_income']

* + - Power BI: Use a calculated column in DAX:

Debt\_to\_Income = DIVIDE([Total\_Debt], [Yearly\_Income])

**4. Data Normalization**

* **Why**: Standardizing numeric values ensures consistency and avoids bias from features with larger ranges.
* **Steps**:
  + **Standardize Numeric Values**:
    - Normalize features like total\_debt, yearly\_income, and credit\_score.
    - Python Example:

from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()

df[['total\_debt', 'yearly\_income', 'credit\_score']] = scaler.fit\_transform(df[['total\_debt', 'yearly\_income', 'credit\_score']])

* + - Power BI: Use the Query Editor to normalize values by calculating Z-scores.
  + **Scale Latitude and Longitude**:
    - Standardize geographic data to ensure precise mapping.

**Tool-Specific Implementation**

**Python**

* Use libraries like pandas, numpy, and sklearn for data cleaning, imputation, and feature engineering.
* Example for Missing Values:

df['yearly\_income'].fillna(df['yearly\_income'].median(), inplace=True)

**SQL**

* SQL queries for cleaning and imputation.
* Example:

DELETE FROM table\_name

WHERE client\_id IS NULL OR yearly\_income IS NULL;

**Power BI**

* Use Query Editor for handling missing values, removing duplicates, and adding calculated columns.
* Example: Create bins for income using "Group By" or "Add Column."

**Tableau**

* Use calculated fields and filters to clean and transform data.
* Example: Create bins in Tableau for credit score ranges.

**Expected Outcomes of Cleaning**

1. A complete, accurate, and consistent dataset for analysis.
2. Derived metrics like debt-to-income ratio and credit\_score\_category enable deeper insights.
3. Standardized and scaled numeric fields improve the reliability of visualizations and models.
4. **Handle Missing Values**:
   * Replace missing demographic values with averages or medians.
   * Drop rows or impute fraud flags if necessary.
5. **Remove Duplicates**:
   * Identify and remove duplicate client\_id or card\_number entries.
6. **Feature Engineering**:
   * Create bins for credit\_score and yearly\_income.
   * Calculate new metrics like debt-to-income ratio.
7. **Data Normalization**:
   * Standardize numeric values (e.g., total\_debt, yearly\_income) for consistent analysis.

**Tool Selection**

Power BI is chosen due to its:

**1. Ease of Use**

* **Drag-and-Drop Interface**:

Power BI allows users to create complex visualizations with minimal technical knowledge.

The user-friendly interface helps in quickly selecting fields, defining measures, and customizing visuals.

* **Pre-Built Templates**:

Power BI provides ready-to-use templates and formatting options, reducing setup time.

**2. Data Integration**

* **Support for Multiple Data Sources**:

Power BI integrates seamlessly with data from **Excel**, **SQL databases**, **CSV files**, **SharePoint**, **Azure**, and other cloud platforms.

Enables combining data from different sources into a single data model for unified analysis.

* **Direct Query and Import Mode**:

**Import Mode** allows data to be brought into Power BI for faster analysis.

**Direct Query** connects to live databases, ensuring real-time updates.

**3. Advanced Visualizations**

* **Rich Visual Library**:

Power BI offers bar charts, line charts, scatter plots, heat maps, tree maps, and more.

It supports custom visuals via the Power BI Marketplace for niche use cases.

* **Customizable Dashboards**:

Dashboards can include KPIs, slicers, and interactivity, allowing users to drill down into data.

**4. Interactivity**

* **Dynamic Dashboards**:

Power BI dashboards allow interactivity such as filtering by region, demographic, or time periods.

* **Cross-Filtering and Highlighting**:

Clicking on a data point in one chart can highlight related data in other visuals.

* **Drill-Through Functionality**:

Users can explore detailed data by drilling through visuals, such as viewing specific customer profiles from a summarized income chart.

**5. Analytical Capabilities**

* **DAX (Data Analysis Expressions)**:

Power BI supports advanced calculations using DAX, enabling custom measures and calculated columns.

* **Insights Feature**:

Power BI uses machine learning models to provide automated insights and detect anomalies in the data.

**6. Collaboration and Sharing**

* **Power BI Service**:

Dashboards and reports can be published to the cloud and shared with stakeholders.

* **Mobile Accessibility**:

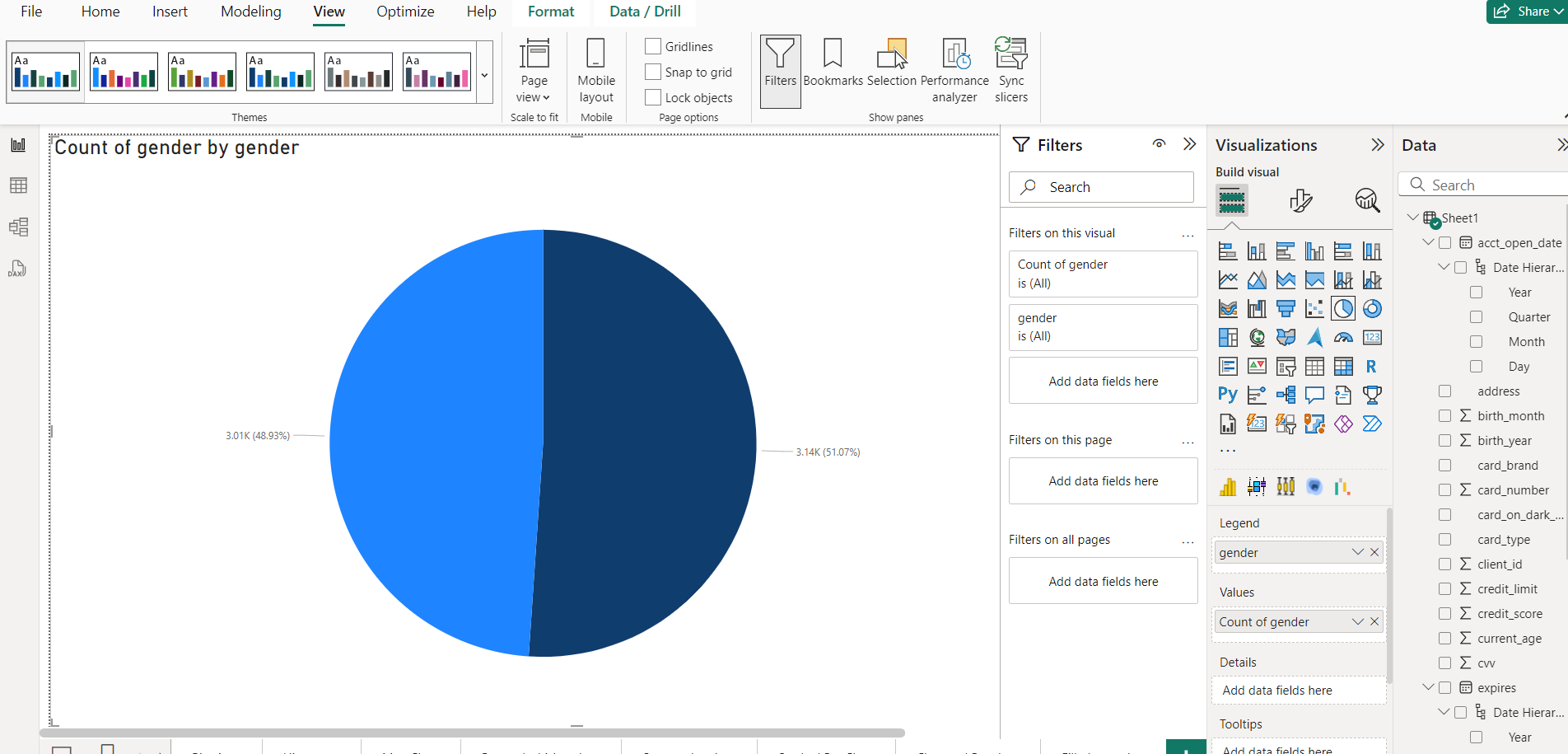
Users can view interactive dashboards on mobile devices.

* **Integration with Microsoft Teams**:

Enables team collaboration and sharing of reports within the Microsoft ecosystem.

**Graphs and Charts Used**

**Univariate Analysis:**

1. **Pie Chart:**

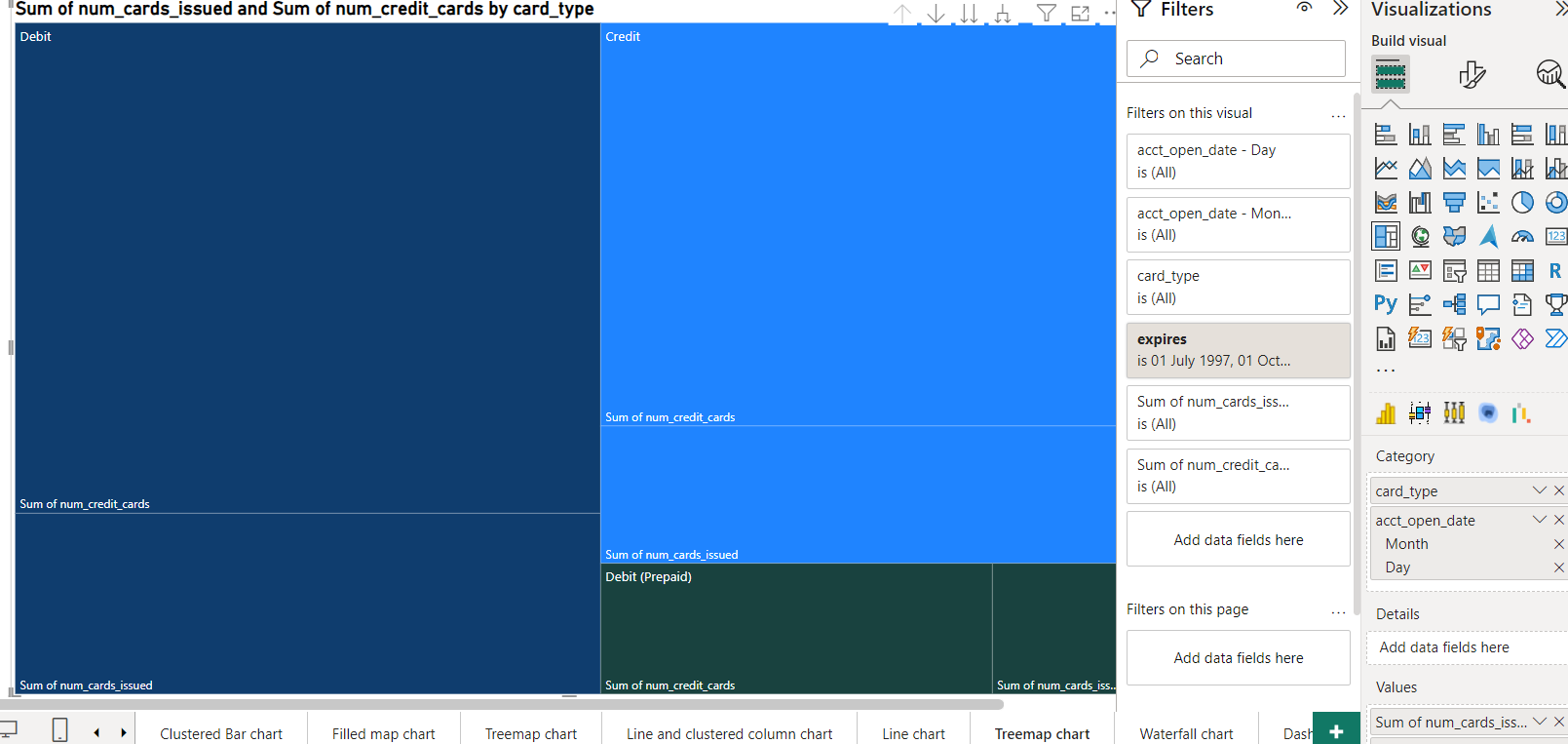
**Univariate analysis** refers to the analysis of a single variable. In this case, the graph is visualizing the distribution of one variable: **gender**. The pie chart simply shows how the data is divided between different gender categories (Male and Female), without considering any other factors or variables.

The analysis and insights based on the chart:

* **Total Count**: The total count of entries is approximately 6.15K (3.01K + 3.14K).
* **Gender Distribution**:
  + **Male**: 3.14K (51.07% of the total)
  + **Female**: 3.01K (48.93% of the total)

**Insights**:

1. The dataset is almost evenly split between male and female, with a slight majority of male entries.
2. The gender distribution indicates a balanced sample size, suggesting that the dataset does not exhibit a significant gender bias.
3. This type of analysis could be useful for understanding demographic trends or making decisions in areas like marketing, recruitment, or product design, ensuring an equitable approach for both genders.

2. Treemap chart:

**Analysis:**

This visualization is a **univariate analysis** because it focuses on a single variable: **Card Type**.

**Insights:**

* **Debit Cards:** The chart shows that the sum of issued cards and credit cards is significantly higher for debit cards compared to prepaid debit cards. This suggests that debit cards are more widely used than prepaid debit cards.
* **Prepaid Debit Cards:** While the number of prepaid debit cards issued and credit cards is lower compared to debit cards, it still represents a significant portion of the market.

**Limitations and Further Analysis:**

* **Without specific numerical values:** It's challenging to quantify the exact difference between the number of debit and prepaid debit cards.
* **Data Granularity:** The level of detail in the data (e.g., card brand, customer demographics) would affect the granularity of the analysis.
* **Contextual Information:** Additional context, such as economic conditions or industry trends, could provide further insights.

**Bivariate Analysis:**

1. A graph of blue and white lines

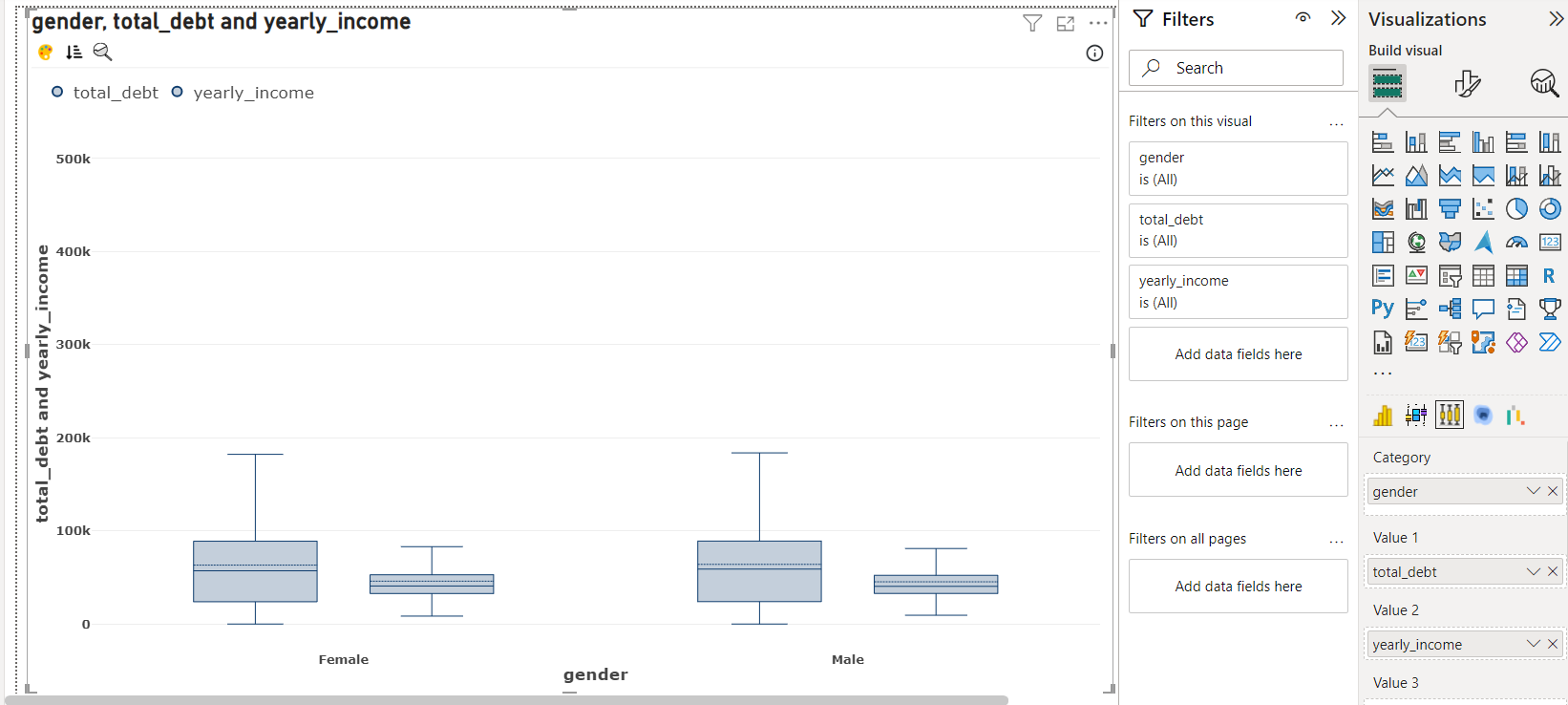
   Description automatically generated with medium confidence**Histogram**

This is a **bivariate** analysis because it visualizes the relationship between two variables: **current\_age** and **retirement\_age**. This type of analysis helps in comparing how two variables behave in relation to one another. If more variables were added (e.g., income, profession), it would shift to multivariate analysis.

**Analysis of the Graph:**

* **X-Axis**: Represents age, ranging from around 20 to 90 years.
* **Y-Axis**: Represents the count or frequency of people at each age group.

**Insights:**

1. **Peak around retirement age**: There is a noticeable peak in the **retirement\_age** distribution around the age of 60–65, which is typically associated with the legal or preferred retirement age in many countries.
2. **Current Age Distribution**: The **current\_age** distribution shows more variability across different age groups, with no clear peak around the retirement range. This suggests that individuals of various ages are present in the dataset, possibly representing a working-age population with varying distances from retirement.
3. **Comparison of Ages**:
   * Younger individuals (ages 20-40) are shown in both current and retirement age, suggesting that retirement planning might begin early in their careers.
   * The **retirement\_age** values are skewed toward the higher end (60–70 years), which makes sense given the typical retirement policies or planning age.
4. **Box and Whisker Chart:**

This visualization is a **bivariate analysis**. It examines the relationship between two variables:

1. **Gender:** Categorical variable (Female or Male)
2. **Financial Metrics:** Numerical variables (total\_debt and yearly\_income)

**Key Observations and Insights:**

**1. Total Debt:**

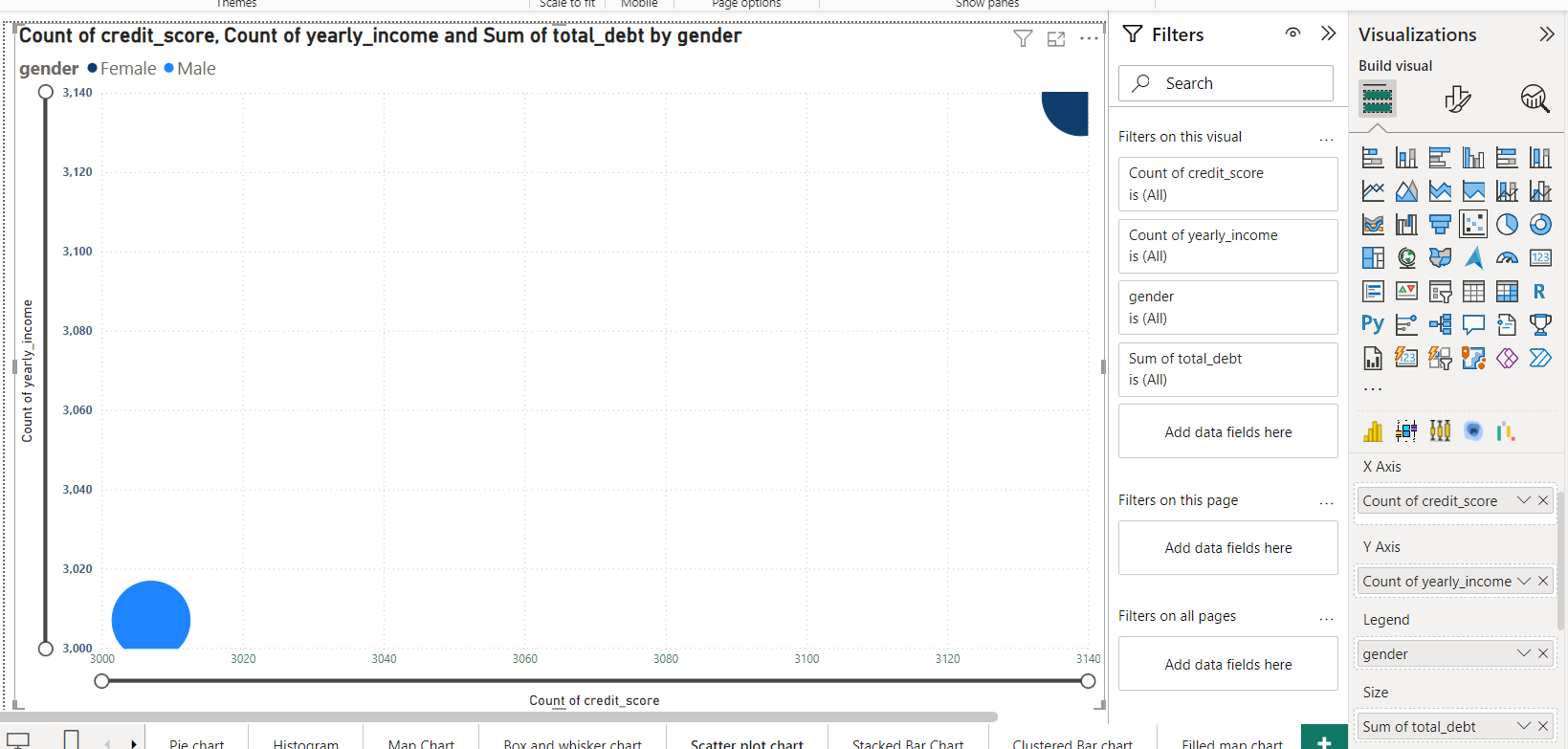
* **Median:** The median total debt for both genders appears to be around 100K.
* **Spread:** The spread (range between the 25th and 75th percentiles) is similar for both genders, indicating similar variability in total debt.
* **Outliers:** There are outliers present in both genders, suggesting individuals with significantly higher total debt.

**2. Yearly Income:**

* **Median:** The median yearly income for both genders is around 300K.
* **Spread:** The spread is again similar for both genders, indicating similar variability in yearly income.
* **Outliers:** There are outliers present in both genders, suggesting individuals with significantly higher yearly income.

**Overall Insights:**

* **Gender Similarity:** The box plots suggest that there is no significant difference in the distribution of total debt and yearly income between males and females. Both genders exhibit similar median values and spreads.
* **Outliers:** The presence of outliers in both categories indicates that there are individuals with extreme values for total debt and yearly income, which might warrant further investigation.
* **Potential for Deeper Analysis:** To gain deeper insights, additional variables like age, occupation, or location could be considered to explore potential correlations and differences.

1. **Scatter Plot chart:**

**Analysis of the Chart**

This chart is a **bivariate** analysis because it explores the relationship between two variables:

1. **Count of Credit Score:** This is a numerical variable representing the frequency of different credit scores.
2. **Count of Yearly Income:** This is another numerical variable representing the frequency of different yearly income ranges.

**Insights:**

* **Distribution of Credit Scores and Yearly Income:** The chart shows the distribution of credit scores and yearly income. The larger the circle, the higher the frequency of that credit score and yearly income combination.
* **Relationship between Credit Score and Yearly Income:** While the chart doesn't directly show a strong correlation between the two variables, it can be inferred that higher credit scores might be associated with higher yearly incomes, based on the distribution of the larger circles.
* **Data Concentration:** The majority of data points seem to be concentrated in the lower ranges of credit scores and yearly incomes. This suggests that a significant portion of the population might have lower credit scores and lower yearly incomes.

A graph on a computer screen

Description automatically generated4. Line chart:

**Analysis:**

This visualization is a bivariate analysis. It explores the relationship between two variables:

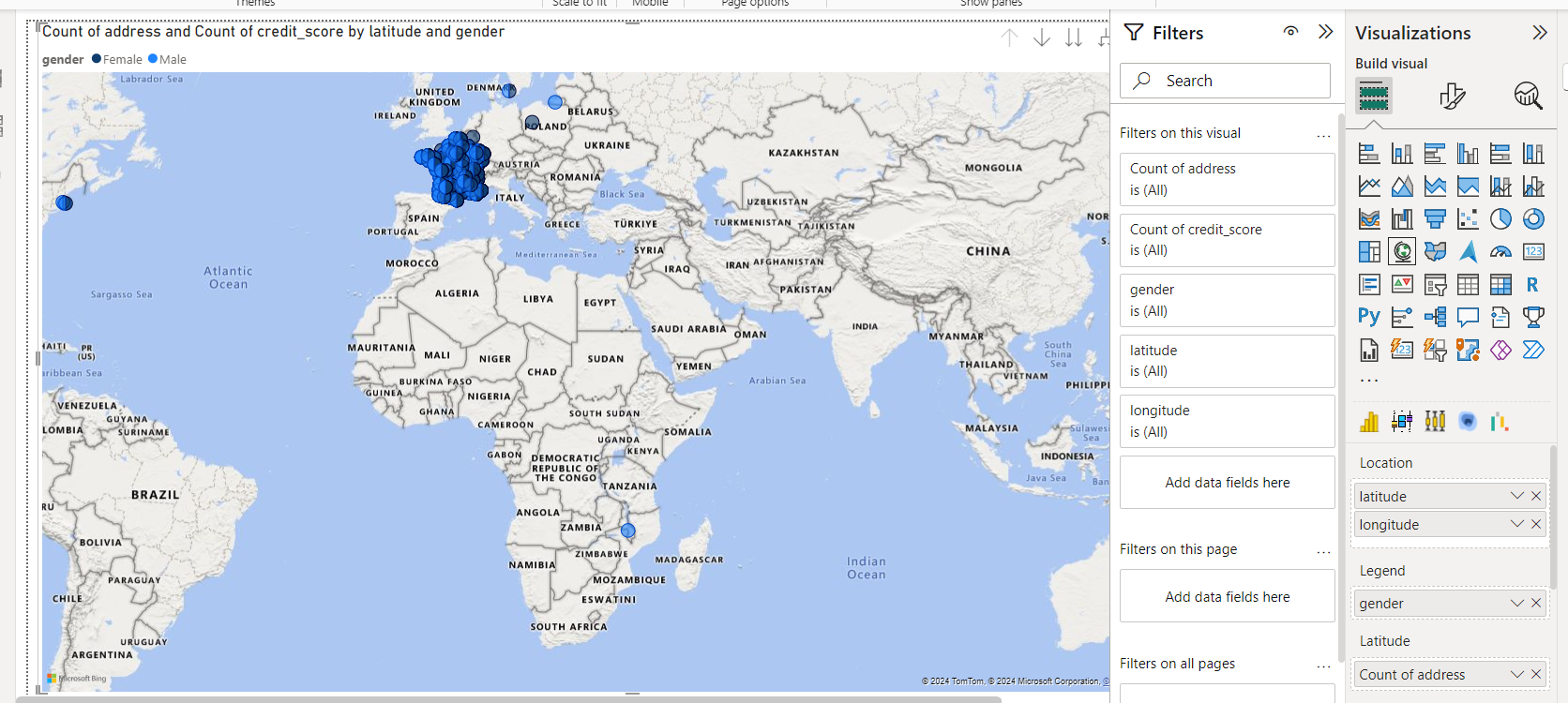
1. Year PIN Last Changed: Numerical variable representing the year when the PIN was last changed
2. Count of Client ID: Numerical variable representing the number of clients who changed their PIN each year

**Insights:**

* Trend Over Time: The chart shows a clear trend in the number of clients changing their PINs over the years. There was a significant increase in PIN changes around 2010, followed by a gradual decline.
* Peak in 2010: The peak in 2010 could be attributed to various factors, such as increased awareness of security threats, data breaches, or changes in security regulations.
* Decreasing Trend: The decreasing trend after 2010 might indicate that clients are becoming more comfortable with their PINs and less likely to change them frequently.
* Security Awareness: The data suggests that security awareness campaigns and incidents can significantly impact PIN change behavior.
* Customer Behavior: Analyzing the distribution of PIN changes across different customer segments (e.g., age, income, location) can provide insights into customer behavior and preferences.

**Multivariate Analysis:**

1. **Map Chart:**

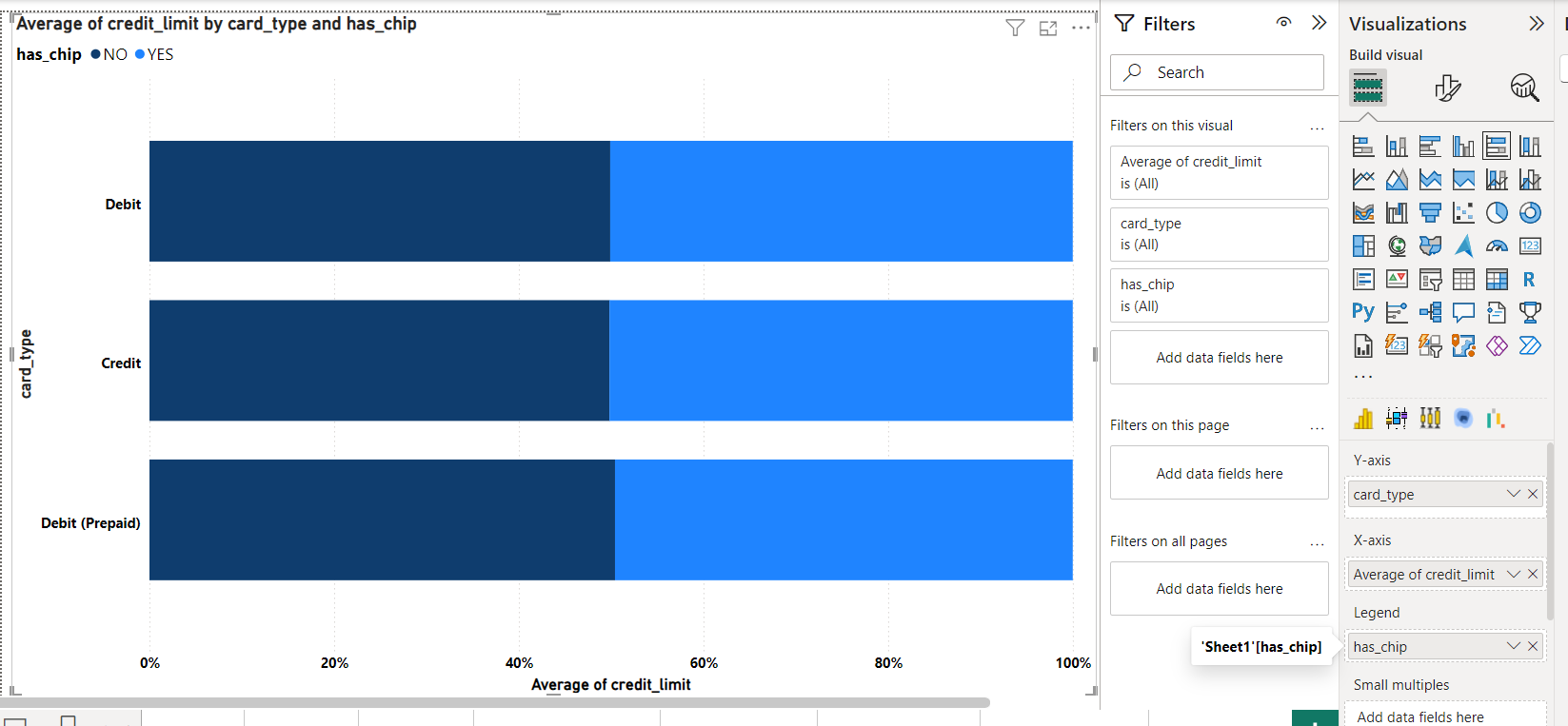
****

**Analysis of the graph :**

1. **Geographic Concentration:**
   * Most data points are clustered in Europe, suggesting a higher concentration of addresses and credit scores in this region.
   * There are fewer data points in other regions like North America, South America, Africa, and parts of Asia.
2. **Cluster Density:**
   * The size of the clusters might indicate the number of addresses or credit scores at that location. Larger clusters likely represent areas with a higher concentration of data points.

**Insights:**

* **Market Potential:** The high concentration of data points in Europe indicates a potentially large market for financial products and services.
* **Fraud Risk Assessment:** Analyzing the geographic distribution of fraud incidents (if available) can help identify regions with higher fraud rates and implement targeted security measures.
* **Customer Segmentation:** Clustering customers based on geographic location can help identify regional preferences and tailor marketing campaigns accordingly.
* **Operational Efficiency:** Understanding the geographic distribution of customers can help optimize branch and ATM locations.

**2. Stacked Bar chart:**

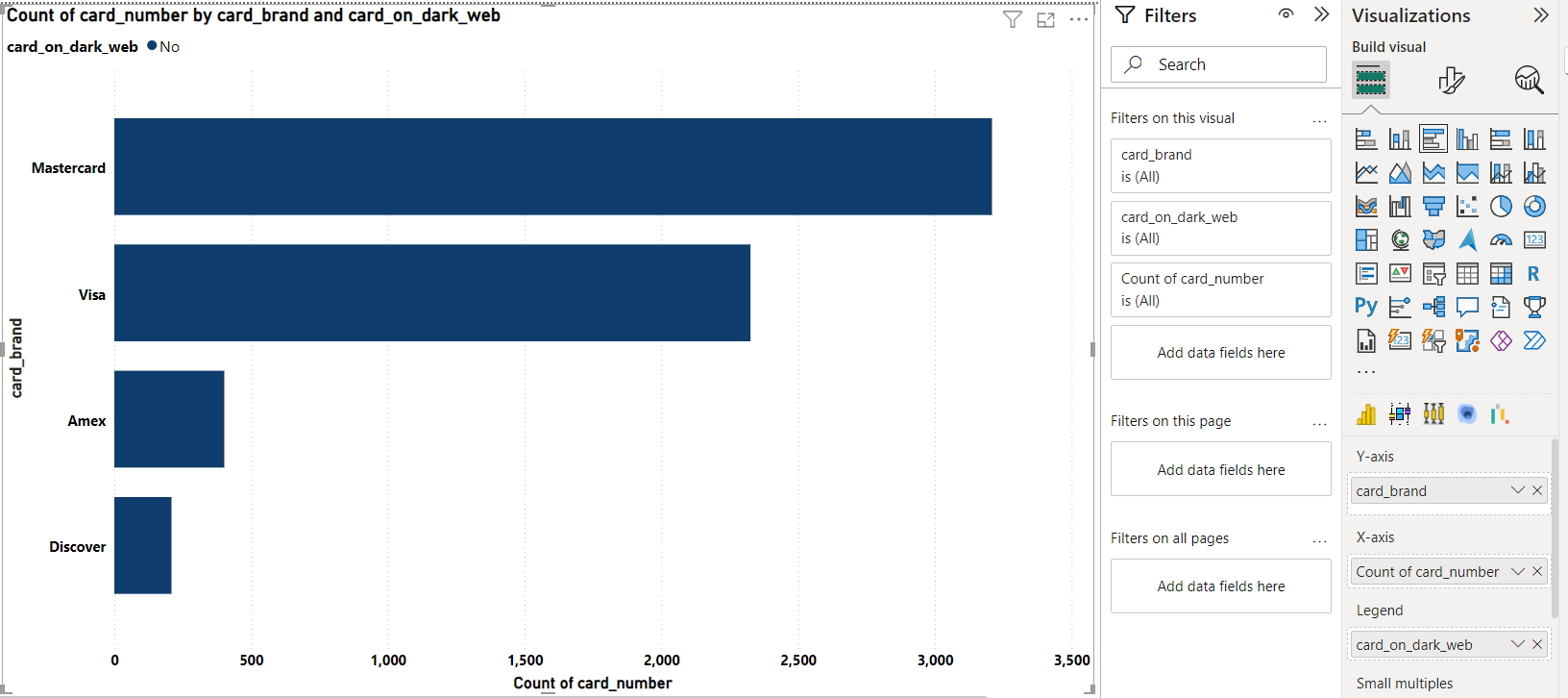
**Analysis:**

This visualization is a **multivariate analysis**. It explores the relationship between three variables:

1. **Card Type:** Categorical variable (Debit, Credit, Debit (Prepaid))
2. **Has Chip:** Categorical variable (Yes, No)
3. **Average Credit Limit:** Numerical variable

**Insights:**

* **Credit Card vs. Debit Card:**
  + Credit cards generally have higher average credit limits compared to debit cards and prepaid debit cards. This is likely due to the nature of credit cards, which allow users to borrow money up to a certain limit.
  + Debit and prepaid debit cards typically have lower credit limits, reflecting their function as payment instruments for available funds.

**3. Clustered Bar chart**:

**Analysis:**

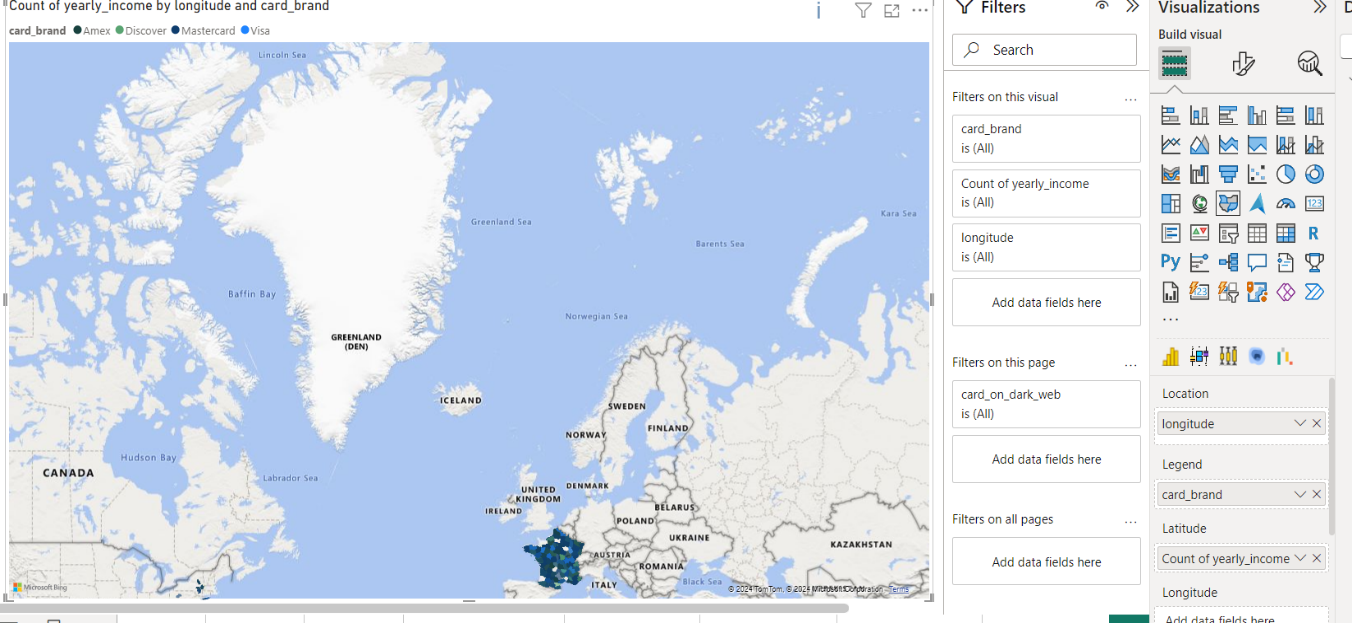
This visualization is a **multivariate analysis**. It explores the relationship between three variables:

1. **Card Brand:** Categorical variable (Mastercard, Visa, Amex, Discover)
2. **Card on Dark Web:** Categorical variable (Yes, No)
3. **Count of Card Number:** Numerical variable representing the frequency of cards

**Insights:**

* **Card Brand Popularity:** Mastercard appears to be the most popular card brand, followed by Visa. Amex and Discover have significantly fewer cards in circulation.
* **Card on Dark Web:** The chart shows the number of cards for each brand that have been found on the dark web. It's clear that Mastercard has the highest number of cards compromised, followed by Visa.
* **Relative Risk:** While Mastercard has the highest number of compromised cards, it's important to consider the relative risk. If Mastercard has a significantly larger market share, the percentage of compromised cards might be lower compared to other brands.
* **Security Measures:** The chart highlights the need for stronger security measures for Mastercard and Visa, especially considering their larger market share.
* **Consumer Awareness:** Educating consumers about the risks of card fraud and best practices for protecting their information can help mitigate the impact of data breaches.
* **Industry Collaboration:** Collaboration between card issuers, payment processors, and cybersecurity firms can help improve the overall security of the payment ecosystem.

**4. Filled Map Chart:**

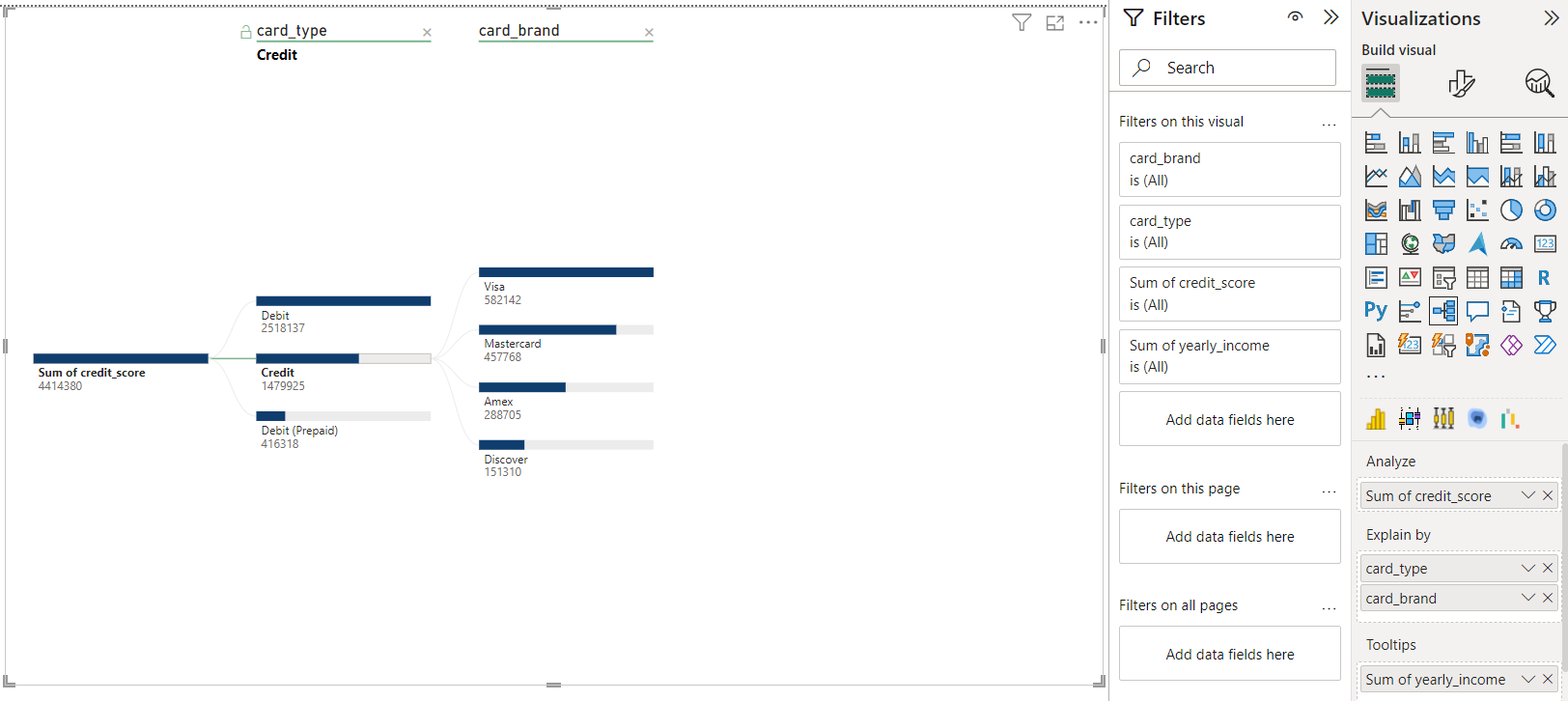


**Analysis:** This visualization is a **multivariate analysis**. It explores the relationship between multiple variables:

1. **Card Brand:** Categorical variable (Visa, Mastercard, Amex, Discover)
2. **Yearly Income:** Numerical variable
3. **Longitude and Latitude:** Geographical coordinates representing location

**Insights:**

* **Geographic Concentration:** The data points are primarily concentrated in Europe, suggesting a higher concentration of cardholders in this region.
* **Card Brand Distribution:** The distribution of card brands varies across different geographic regions. Some regions might have a higher prevalence of specific card brands.
* **Yearly Income Variation:** The size of the data points might indicate the average yearly income of cardholders in a particular location. Larger points could represent regions with higher average incomes.
* **Market Potential:** Identifying regions with a high concentration of cardholders can help financial institutions target specific markets.
* **Fraud Risk Assessment:** Analyzing the geographic distribution of fraud incidents (if available) can help identify regions with higher fraud rates and implement targeted security measures.
* **Customer Segmentation:** Clustering customers based on geographic location and other demographic factors can help identify specific segments with distinct needs and preferences.

**5**. **Tree Map chart:**

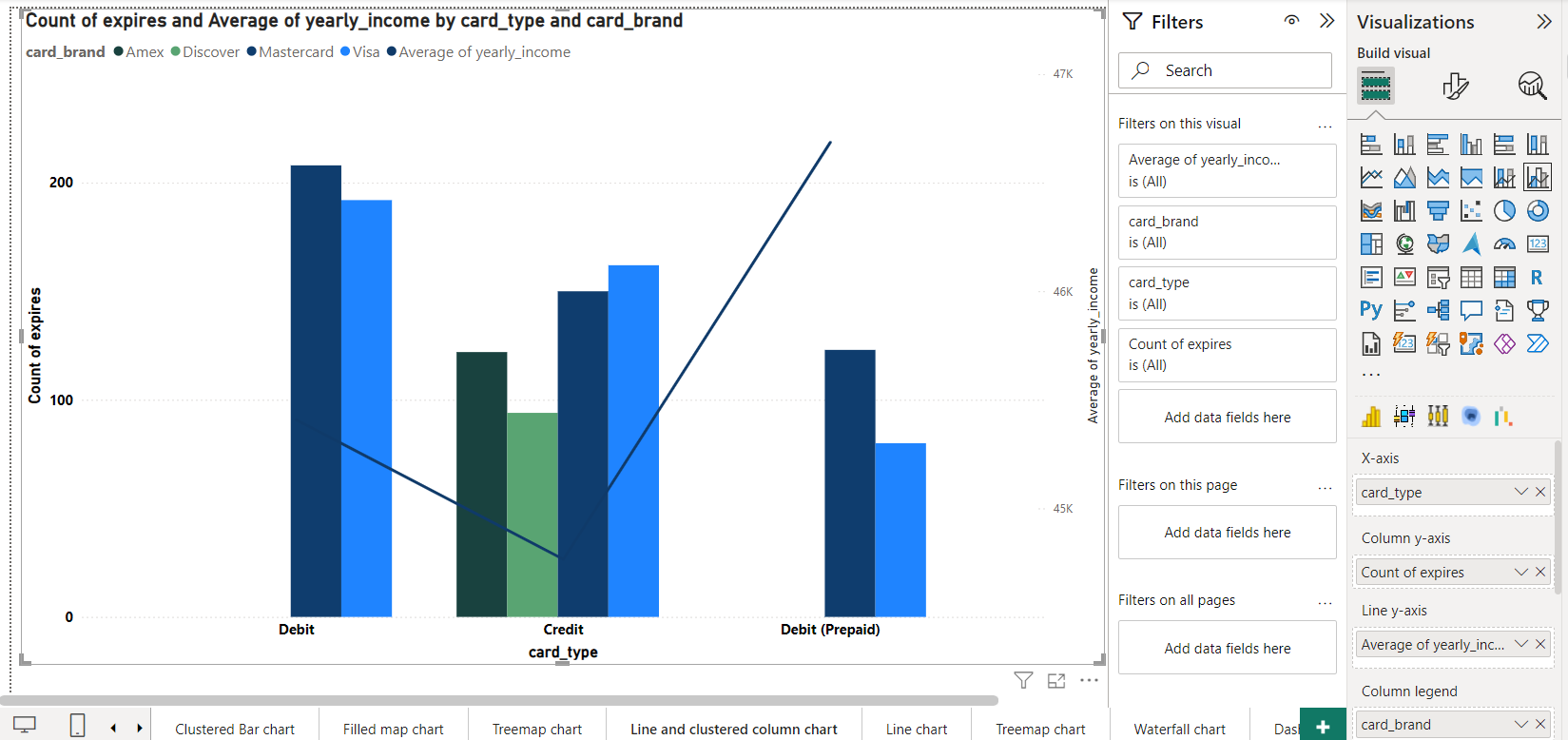
**Analysis:**

This visualization is a **multivariate analysis**. It explores the relationship between three variables:

1. **Card Type:** Categorical variable (Credit, Debit, Debit (Prepaid))
2. **Card Brand:** Categorical variable (Visa, Mastercard, Amex, Discover)
3. **Sum of Credit Score:** Numerical variable

**Insights:**

* **Credit Card Dominance:** Credit cards have the highest total credit score, indicating a higher number of credit cards or higher average credit scores for credit card holders.
* **Debit Card Usage:** Debit cards have the second-highest total credit score, suggesting widespread usage of debit cards.
* **Prepaid Card Usage:** Prepaid debit cards have the lowest total credit score, indicating lower usage or lower average credit scores for prepaid cardholders.
* **Card Brand Performance:** Within each card type, the distribution of total credit scores varies across different card brands. Visa and Mastercard appear to have higher total credit scores compared to Amex and Discover, suggesting a larger user base or higher average credit scores for these brands.
* **Customer Segmentation:** By segmenting the data further based on other variables like age, income, or location, financial institutions can identify specific customer segments with different card usage patterns.
* **Risk Assessment:** Analyzing the distribution of credit scores across different card types and brands can help financial institutions assess credit risk and implement appropriate risk management strategies.
* **Marketing Strategies:** Understanding the preferences of different customer segments can help financial institutions tailor their marketing campaigns and product offerings.

**6. Clustered Column chart:**

**Analysis:**

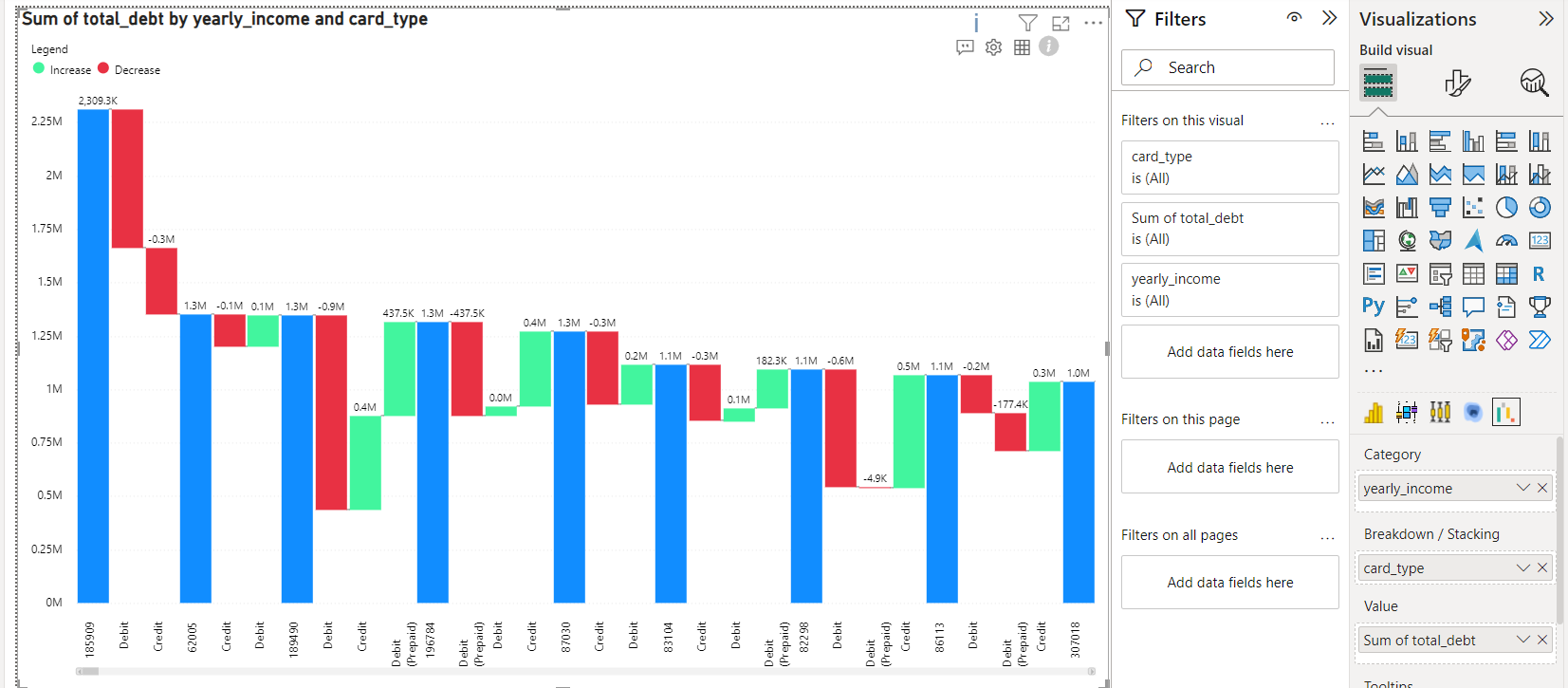
This visualization is a **multivariate analysis**. It explores the relationship between multiple variables:

1. **Card Type:** Categorical variable (Debit, Credit, Debit (Prepaid))
2. **Card Brand:** Categorical variable (Amex, Discover, Mastercard, Visa)
3. **Count of Expires:** Numerical variable representing the frequency of expired cards
4. **Average of Yearly Income:** Numerical variable representing the average yearly income of cardholders

**Insights:**

* **Card Type and Expires:**
  + Credit cards have the highest number of expired cards, followed by debit cards and prepaid debit cards. This might indicate a higher number of credit cards in circulation or longer card validity periods for credit cards.
* **Card Brand and Expires:**
  + Mastercard has the highest number of expired cards, followed by Visa. This could be due to a larger market share or different card validity policies.
* **Yearly Income and Card Type:**
  + The line chart shows a positive correlation between card type and average yearly income. Credit cards tend to have higher average yearly incomes compared to debit and prepaid debit cards.
* **Card Brand and Yearly Income:**
  + The line chart also shows variations in average yearly income across different card brands. Visa and Mastercard generally have higher average yearly incomes compared to Amex and Discover.
* **Customer Segmentation:** By segmenting the data further based on other variables like age, location, or transaction history, financial institutions can identify specific customer segments with different card usage patterns.
* **Risk Assessment:** Analyzing the distribution of expired cards and average yearly income can help financial institutions assess credit risk and implement appropriate risk management strategies.
* **Marketing Strategies:** Understanding the preferences of different customer segments can help financial institutions tailor their marketing campaigns and product offerings. 1

**7. Waterfall chart:**

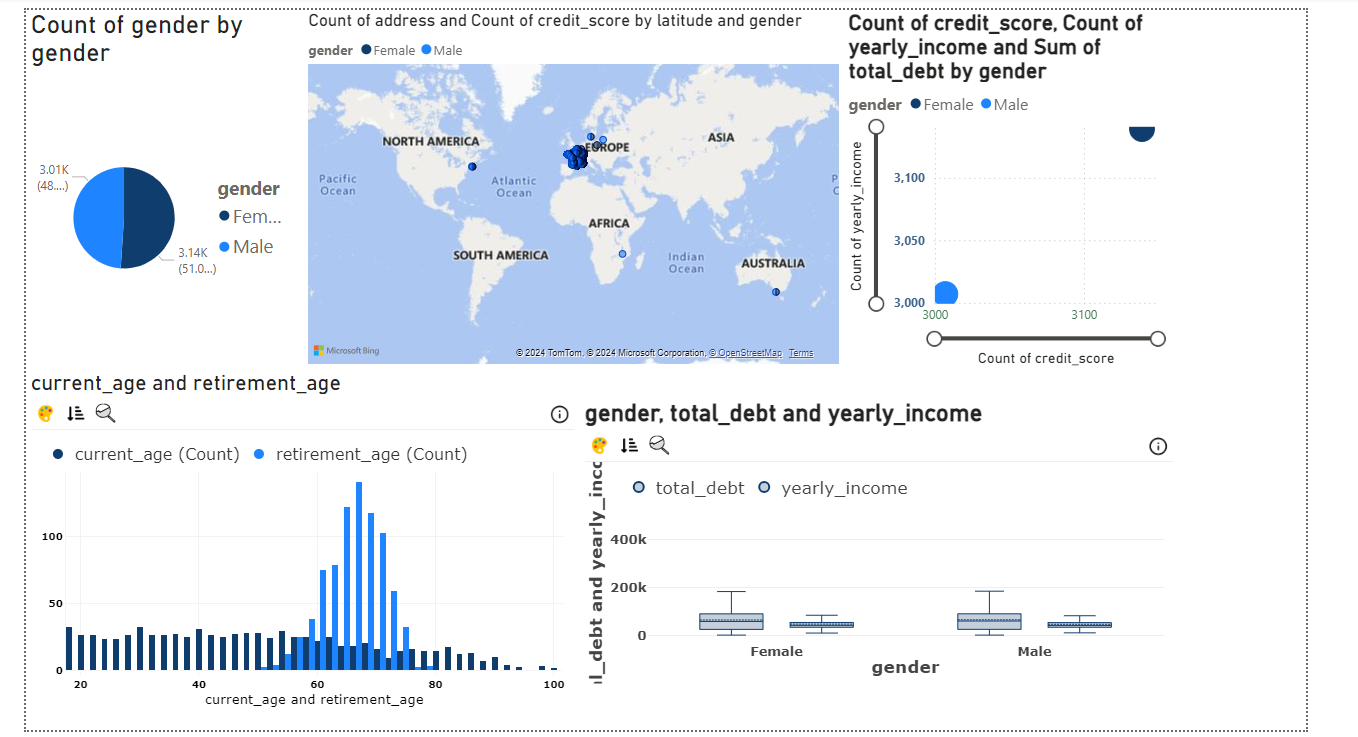
[](https://www.cognizant.com/us/en/glossary/customer-segmentation-banking" \t "_blank)

**Analysis:** This is a **multivariate analysis**. It explores the relationship between multiple variables:

1. **Year:** Categorical variable representing different years
2. **Card Type:** Categorical variable (Credit, Debit, Debit (Prepaid))
3. **Sum of Total Debt:** Numerical variable representing the total debt for each card type each year

**Insights:**

* **Overall Trend:** The chart shows a general increasing trend in total debt over the years for all card types.
* **Credit Card Dominance:** Credit cards consistently have the highest total debt compared to debit and prepaid debit cards. This is likely due to the higher credit limits and spending power associated with credit cards.
* **Debit Card Usage:** Debit cards have a significant share of total debt, indicating widespread usage and potentially higher transaction volumes.
* **Prepaid Card Usage:** Prepaid debit cards have the lowest total debt, which is expected given their lower spending limits and reliance on preloaded funds.
* **Year-to-Year Variation:** There are variations in total debt across different years, which could be influenced by economic factors, consumer behavior, or changes in credit policies.
* **Customer Segmentation:** By segmenting the data further based on other variables like age, income, or location, financial institutions can identify specific customer segments with different debt patterns.
* **Risk Assessment:** Analyzing the distribution of total debt across different card types and years can help financial institutions assess credit risk and implement appropriate risk management strategies.
* **Marketing Strategies:** Understanding the preferences of different customer segments can help financial institutions tailor their marketing campaigns and product offerings.

**Dashboard**

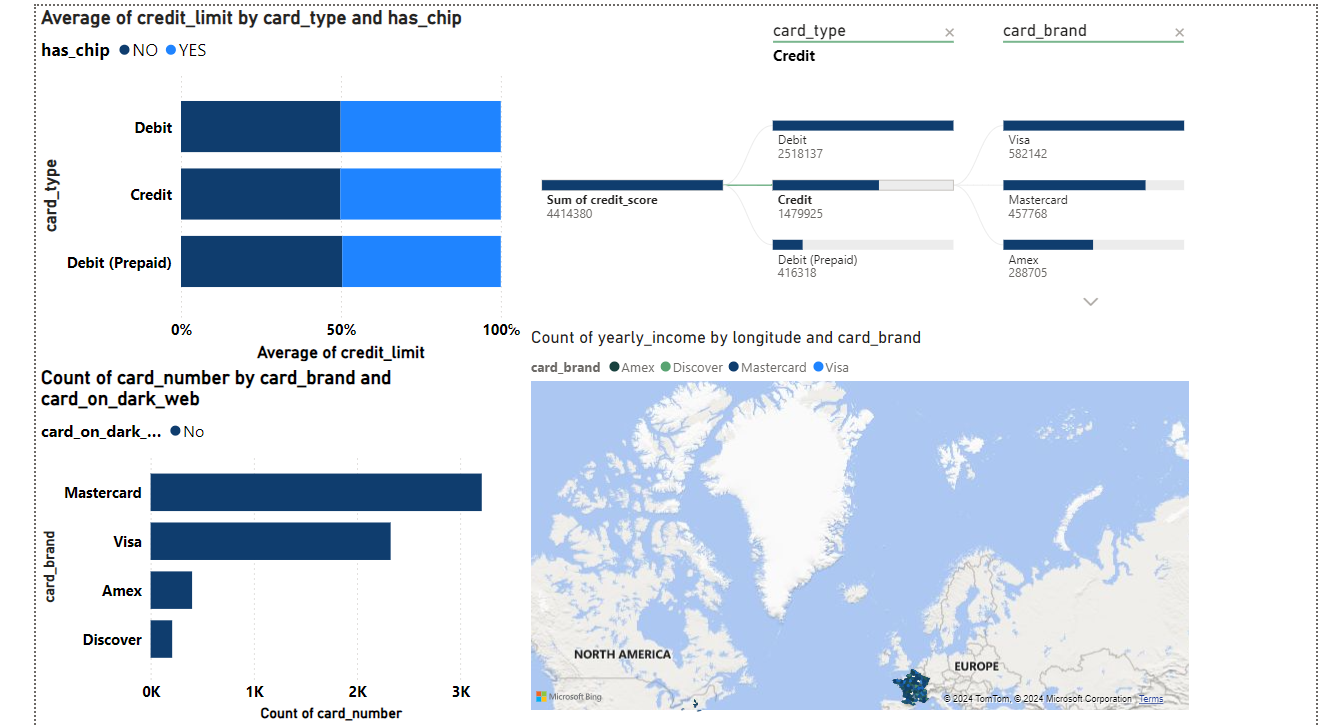
**Gender Distribution:** The dataset shows a relatively equal distribution between male and female genders.

**Geographic Concentration:** The majority of data points are concentrated in Europe, indicating a higher concentration of customers or transactions in this region.

**Financial Health:** The analysis of credit scores, yearly income, and debt levels can provide insights into the financial health of different customer segments.

**Age Distribution:** The distribution of current and retirement ages can be used to target specific age groups with relevant financial products and services.

**Card Usage Patterns:** The analysis of card types, brands, and usage patterns can help identify trends and opportunities for financial institutions.



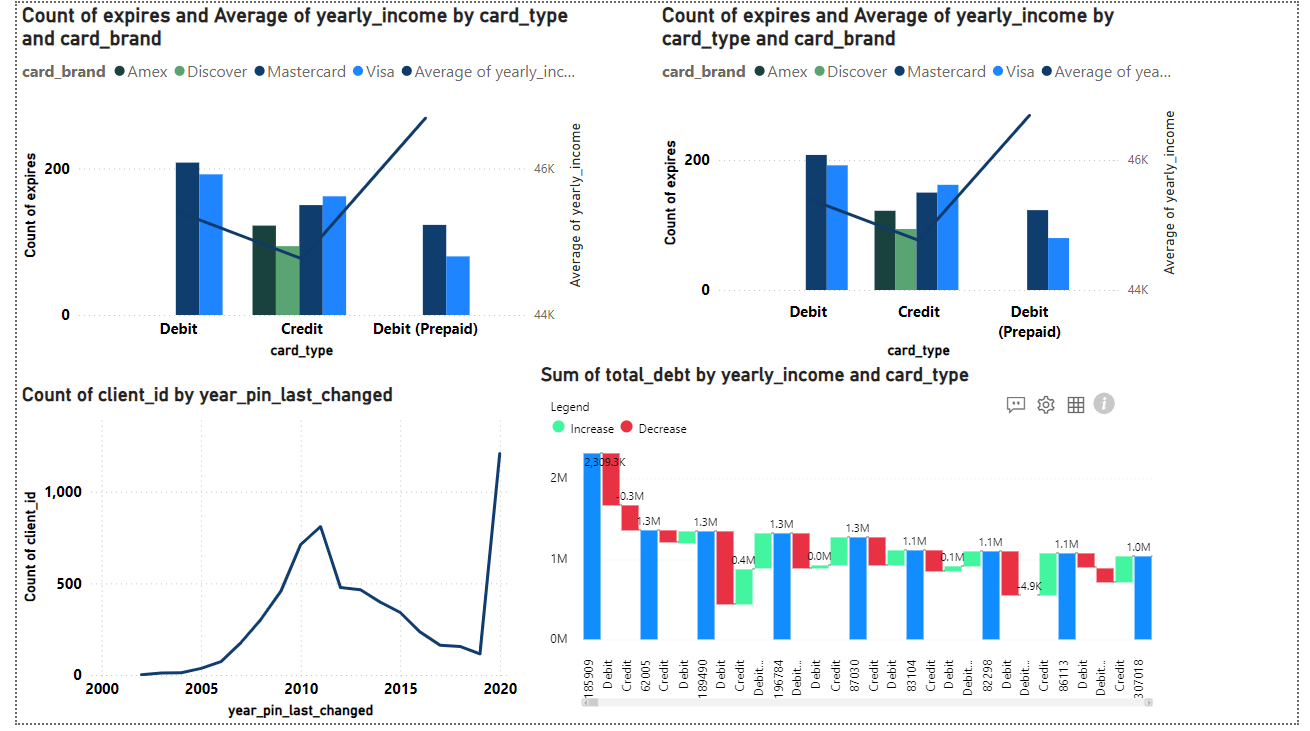
**Card Type and Credit Limit:** Credit cards have significantly higher average credit limits compared to debit and prepaid debit cards, especially when equipped with chip technology.

**Card Brand and Credit Limit:** Mastercard and Visa generally have higher average credit limits compared to Amex and Discover, suggesting a larger market share or higher creditworthiness among their cardholders.

**Card Fraud:** Mastercard has the highest number of cards compromised on the dark web, followed by Visa. This highlights the need for stronger security measures for these major card brands.

**Geographic Distribution:** Most cardholders are concentrated in Europe, indicating a higher market potential in this region.

**Yearly Income:** The map suggests variations in yearly income across different geographic regions. Identifying regions with higher average incomes can help target premium card products and services.



**Card Type and Yearly Income:** Credit cards tend to have a higher average yearly income compared to debit and prepaid debit cards. This suggests that credit cardholders might have higher purchasing power.

**Card Type and Expired Cards:** Credit cards have the highest number of expired cards, followed by debit cards and prepaid debit cards. This could be due to factors like longer card validity periods for credit cards or higher usage rates.

**Card Brand and Expired Cards:** Mastercard has the highest number of expired cards, followed by Visa. This might be due to their larger market share or different card validity policies.

**PIN Changes:** The chart shows a significant increase in PIN changes around 2010, followed by a gradual decline. This could be attributed to increased security awareness or changes in security regulations.

**Total Debt:** The chart shows the total debt for different card types and income levels. Credit cards consistently have the highest total debt, followed by debit cards. Prepaid debit cards have the lowest total debt.

**Conclusion and Storytelling**

**Key Storylines:**

1. **Customer Segmentation and Financial Health:**
   * **Income and Credit Card Usage:** The analysis reveals a strong correlation between yearly income and credit card usage. Higher-income individuals tend to use credit cards more frequently and have higher credit limits.
   * **Debt and Card Type:** Credit cards consistently have the highest total debt, indicating a higher reliance on credit. Debit and prepaid debit cards have lower debt levels, suggesting a more cautious approach to spending.
   * **Geographic Insights:** The geographic distribution of customers provides valuable information for targeted marketing and risk management. Regions with higher concentrations of customers can be prioritized for specific campaigns or security measures.
2. **Fraud and Security:**
   * **Card Compromise:** The data highlights the prevalence of card fraud, particularly for Mastercard and Visa. Implementing robust security measures, such as chip technology and fraud detection systems, is crucial to mitigate risks.
   * **PIN Changes:** The trend in PIN changes over time reflects changing security awareness and regulatory requirements. Financial institutions should continue to educate customers about best practices for password security.
3. **Customer Behavior and Preferences:**
   * **Card Usage Patterns:** The analysis of card types, brands, and usage patterns reveals insights into customer preferences and behaviors. This information can be used to tailor product offerings and marketing strategies.
   * **Age and Retirement:** Understanding the age distribution of customers helps in designing financial products and services that cater to specific life stages.

**Impact on Business:**

* **Enhanced Risk Management:** By identifying high-risk segments and regions, the institution can implement targeted security measures and fraud prevention strategies.
* **Improved Customer Experience:** Understanding customer preferences and behavior allows for personalized product offerings and services, leading to increased customer satisfaction and loyalty.
* **Optimized Marketing Strategies:** By segmenting customers based on demographics, financial behavior, and geographic location, the institution can tailor marketing campaigns to specific target audiences.
* **Data-Driven Decision Making:** Leveraging data-driven insights enables informed decision-making at all levels of the organization.

Overall, the dashboard empowers financial institutions to make data-driven decisions that drive growth, mitigate risk, and enhance customer satisfaction.