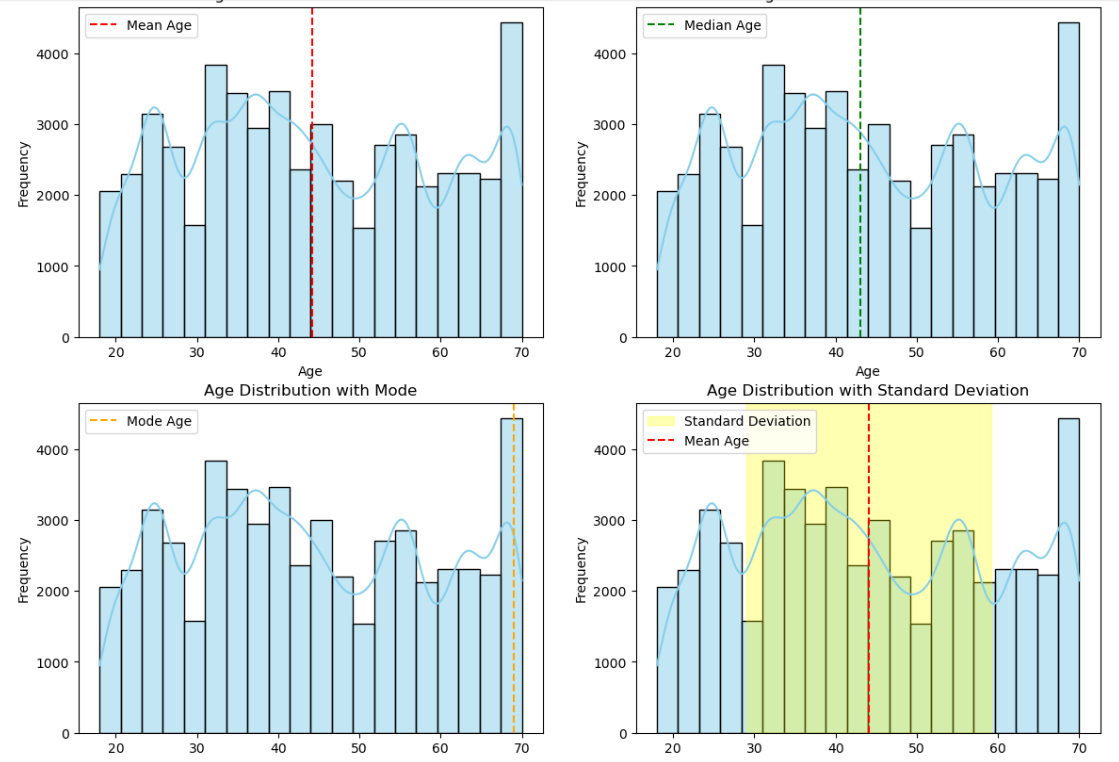
Descriptive analysis.

1. **Age Distribution**

* **Mean age**: The average age in the dataset is approximately 44.14 years.
* **Median age**: The median age, which represents the middle value when all ages are ordered, is 43 years.
* **Mode age**: The mode age, which is the age that appears most frequently in the dataset, is 69 years.
* **Standard deviation of age**: The standard deviation indicates the dispersion or spread of ages around the mean. In this case, it's approximately 15.08 years.



**Here are some conclusions we can draw from these findings:**

Central Tendency:

The mean, median, and mode are relatively close to each other, suggesting that the distribution of ages is likely symmetric or nearly symmetric.

Mode:

The mode being at 69 years suggests that there might be a significant number of individuals in the dataset around this age. This could imply a peak in the distribution of ages around this value.

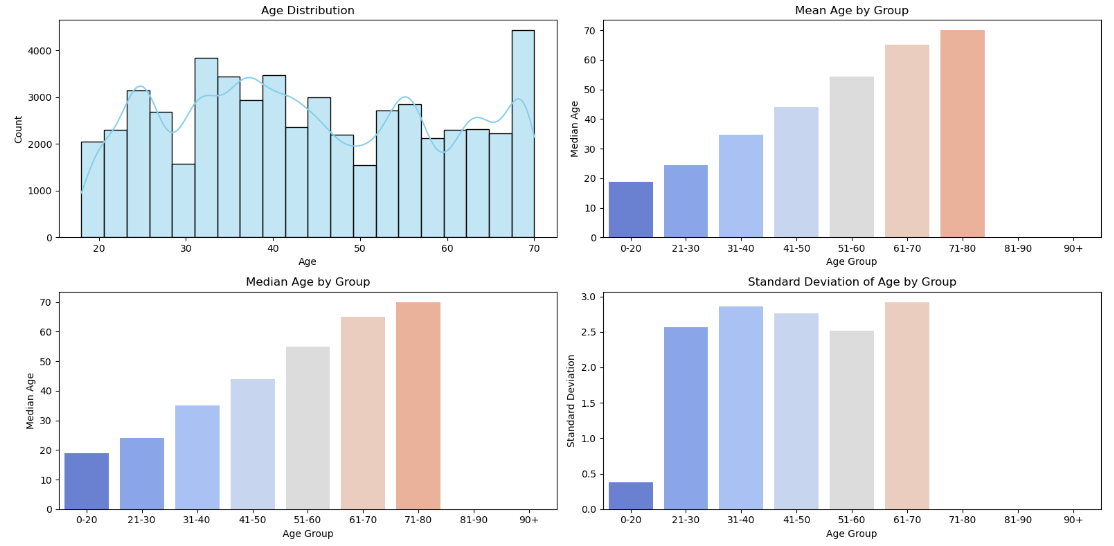
Spread:

The standard deviation of approximately 15.08 years indicates that ages in the dataset vary around the mean by about 15 years on average. This suggests a moderate spread of ages in the dataset.

Skewness:

The fact that the mean and median are close together suggests that the distribution of ages is likely not heavily skewed. If the mean were substantially different from the median, it would indicate skewness in one direction or the other.

**Analysis Across Age Groups.**

****

Age groups seem to follow an increasing trend with respect to mean and median age values, which is expected as individuals tend to grow older in higher age groups.

Variability Across Groups:

Standard deviation provides information about the variability or spread of ages within each group.

Groups with higher standard deviations indicate greater variability in ages within those groups. For instance, age groups '31-40' and '61-70' have relatively higher standard deviations compared to others, suggesting a wider range of ages within these groups.

Uniformity of Some Groups:

Some age groups, such as '71-80', have a standard deviation of 0, indicating that all individuals within these groups have the same age. This could be due to the nature of the data or the way age groups are defined.

Missing Values:

Age group '81-90' and '90+' have Nan values for all statistics, indicating that there might be no individuals or insufficient data in these groups.

1. **Gender Distribution**

**Based on Findings in the dataset we can make the following conclusions about the following topics:**

****

Gender Distribution:

The dataset contains a relatively balanced distribution of genders, with 51.4% male and 48.6% female.

The count of males is slightly higher than females, with approximately 14,995 more males than females.

Gender Equality:

The proportions suggest that there is no significant gender imbalance in the dataset, as both genders are represented fairly equally.

This indicates that the dataset is diverse in terms of gender representation, which can be beneficial for analysing gender-related trends or making gender-inclusive decisions.

Implications for Analysis:

The balanced distribution of genders allows for more reliable analysis and conclusions about gender-specific trends, behaviours, or preferences within the dataset.

We can explore various aspects such as purchasing behaviour, interaction patterns, or preferences across genders with confidence that the findings are representative of both male and female populations.

Further Investigation:

While the proportions suggest gender balance overall, it's essential to consider potential variations in gender representation within subgroups or specific segments of the dataset.

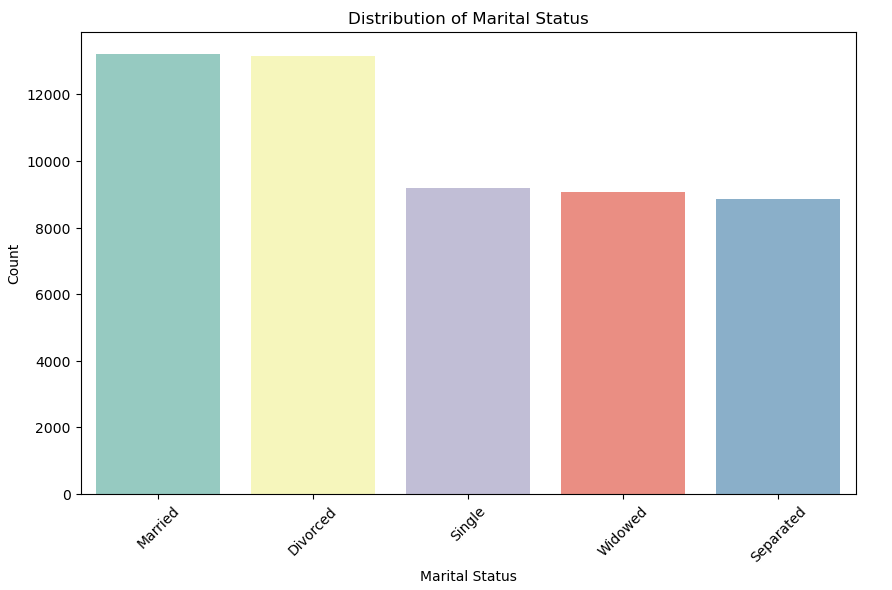
Further investigation could involve examining gender distribution across different demographics, geographic regions, or customer segments to identify any disparities or patterns that may exist. This will be done in later chapters of Bivariate Analysis.

Potential Biases:

We should be cautious about potential biases that may arise if the dataset disproportionately represents one gender over the other, which could skew the results and lead to erroneous conclusions.

1. **Marital Status Distribution**

**Based on the dataset we can make the following conclusions about the marital status distribution analysis.**

****

Marital Status Distribution:

The dataset consists of individuals with various marital statuses, including Married, Divorced, Single, Widowed, and Separated.

Married and Divorced individuals are the most prevalent marital statuses, with counts of 13,219 and 13,151, respectively.

Single, Widowed, and Separated individuals have slightly lower counts, ranging from 8,861 to 9,195.

Proportional Representation:

Married and Divorced individuals each account for approximately 24.7% and 24.6% of the dataset, respectively. This indicates that they are the two most common marital statuses.

Single individuals represent approximately 17.2% of the dataset, while Widowed and Separated individuals each represent around 17.0% and 16.6%, respectively.

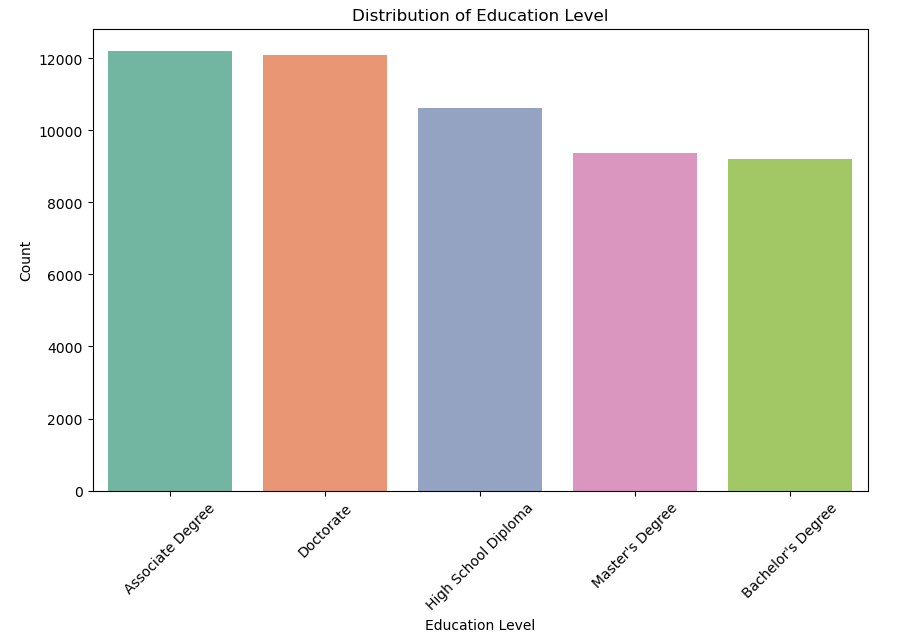
Implications for Analysis:

Understanding the distribution of marital statuses is essential for conducting analyses that consider demographic factors and relationship dynamics.

Analyses may need to account for differences in behaviours, preferences, or outcomes based on marital status, as individuals in different relationship statuses may have distinct needs and experiences.

1. **Education Level**

**Based on the dataset we can make the following conclusions:**

****

Education Level Distribution:

The dataset includes individuals with various levels of education, including Associate Degree, Doctorate, High School Diploma, Master's Degree, and Bachelor's Degree.

The counts reveal the number of individuals associated with each education level, ranging from 9,214 to 12,213.

Proportional Representation:

Associate Degree and Doctorate are the two most prevalent education levels, each accounting for approximately 22.8% to 22.6% of the dataset.

High School Diploma, Master's Degree, and Bachelor's Degree represent slightly lower proportions, ranging from approximately 17.2% to 19.8%.

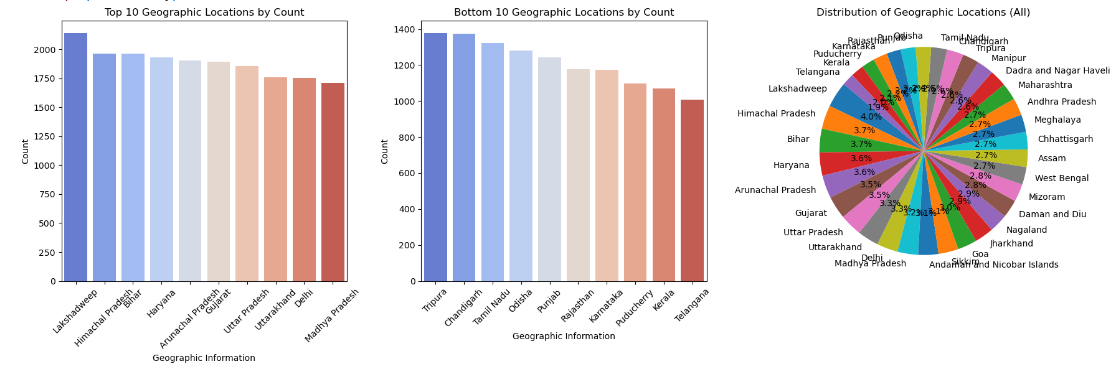
Implications for Analysis:

Understanding the distribution of education levels is crucial for conducting analyses that consider educational attainment as a demographic factor.

Analyses may need to account for differences in behaviours, preferences, or outcomes based on education level, as individuals with different educational backgrounds may have distinct needs and experiences.

1. **Geographic Analysis**

**From the provided results, we can draw several conclusions regarding the geographic distribution of locations in the dataset:**



Variability in Geographic Representation:

The dataset contains a diverse range of geographic locations, including states, union territories, and islands, indicating a broad representation across different regions of the country.

Concentration of Counts:

Lakshadweep has the highest count among all locations, followed by Himachal Pradesh, Bihar, and Haryana. This suggests that these locations are relatively more prevalent or have a higher frequency of occurrence in the dataset compared to others.

Proportional Distribution:

The proportion of each geographic location provides insight into the relative prevalence of different regions within the dataset. Lakshadweep, Himachal Pradesh, and Bihar have the highest proportions, indicating a significant presence in the dataset compared to others.

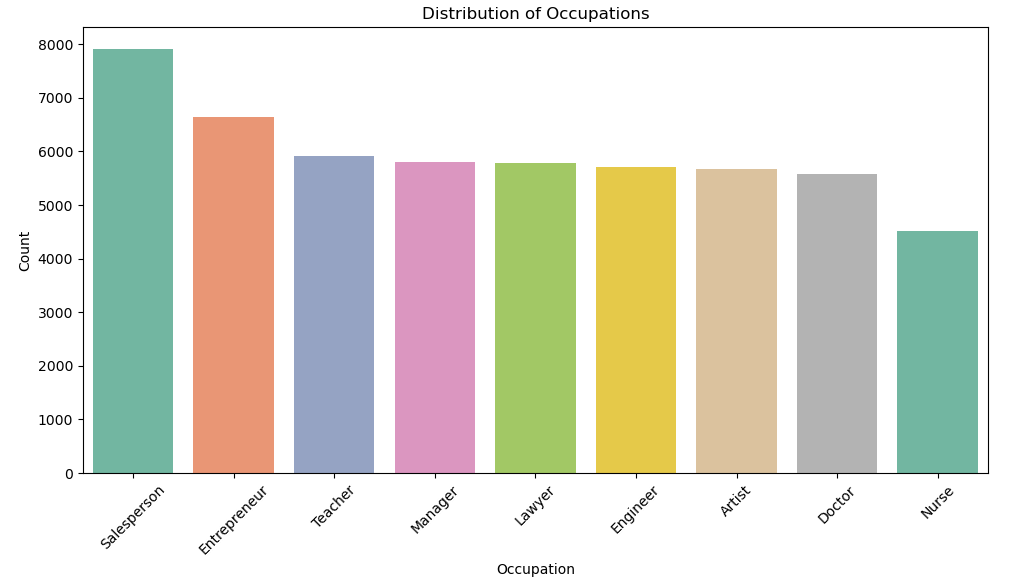
However, it's essential to note that even locations with lower counts, such as Telangana and Kerala, still contribute to a substantial proportion of the dataset.

Regional Representation:

The distribution of geographic locations reflects a representation of various states, union territories, and regions across India, encompassing both mainland and island territories. This indicates the diversity of locations represented in the dataset.

1. **Occupational Analysis**

**From the provided data, we can draw several conclusions regarding the distribution of occupations in the dataset:**



Variety of Occupations:

The dataset encompasses a diverse range of occupations, including salespersons, entrepreneurs, teachers, managers, lawyers, engineers, artists, doctors, and nurses. This indicates a broad representation of professions within the dataset.

Prevalence of Salespersons and Entrepreneurs:

Salespersons and entrepreneurs are the most common occupations, with 7919 and 6636 occurrences, respectively. This suggests that a significant portion of individuals in the dataset are engaged in sales-related roles or entrepreneurial activities.

Presence of Education and Healthcare Professionals:

Teachers, doctors, and nurses are also prevalent occupations in the dataset, with 5906, 5573, and 4521 occurrences, respectively. This indicates the presence of education and healthcare professionals among the individuals represented in the dataset.

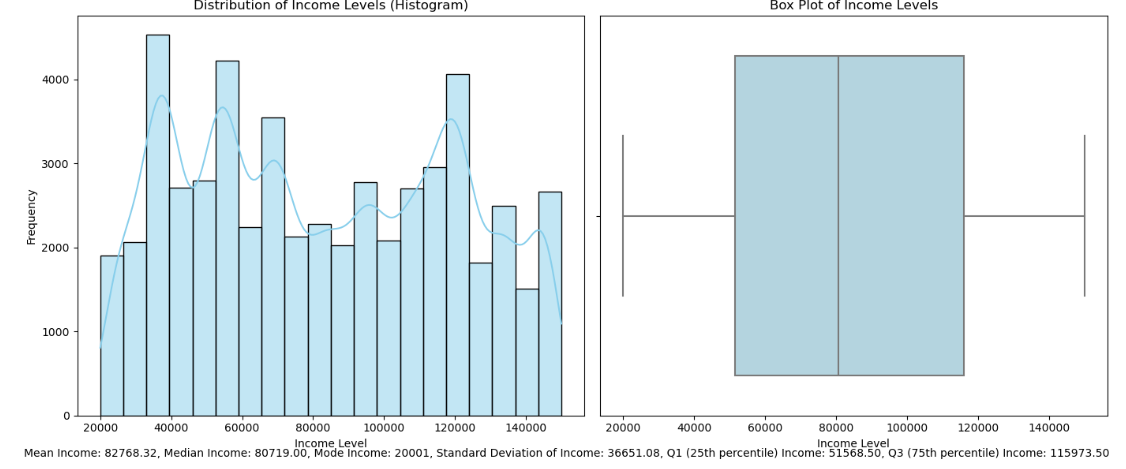
Managerial and Professional Roles:

Managerial roles, such as managers, and professional roles, such as lawyers and engineers, are also well-represented, with 5803, 5775, and 5704 occurrences, respectively. This suggests the presence of individuals holding leadership or specialized technical positions.

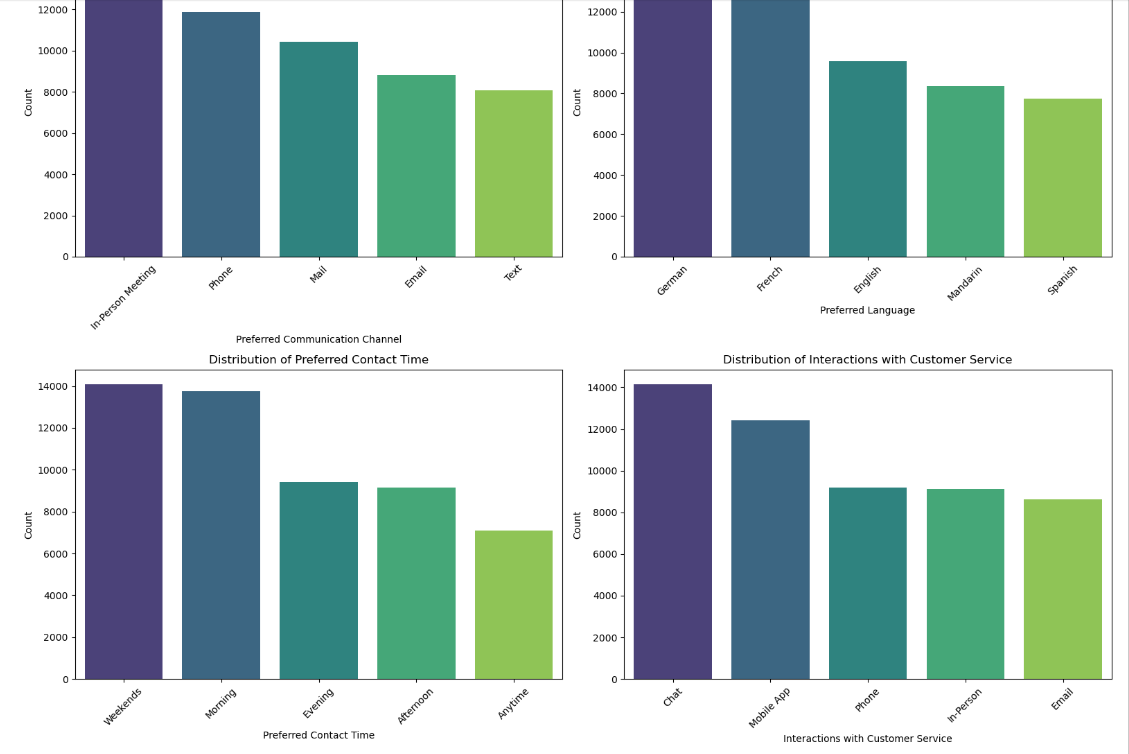
Diversity of Skill Sets:

The diversity of occupations in the dataset reflects a wide range of skill sets, expertise, and professional backgrounds among the individuals. This diversity may have implications for analysing demographic trends, economic activities, or research focus areas within the dataset.

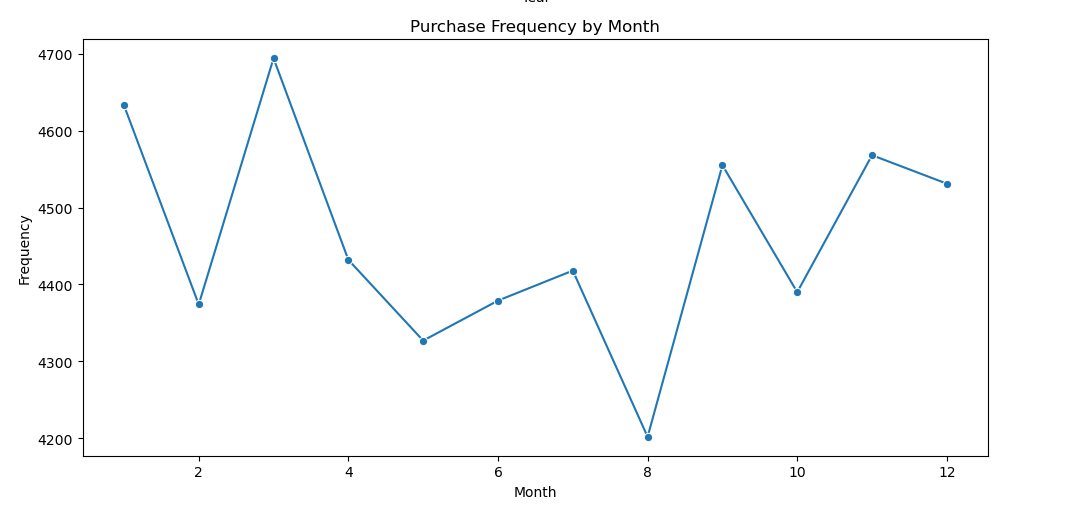
1. **Distribution of Income Levels.**

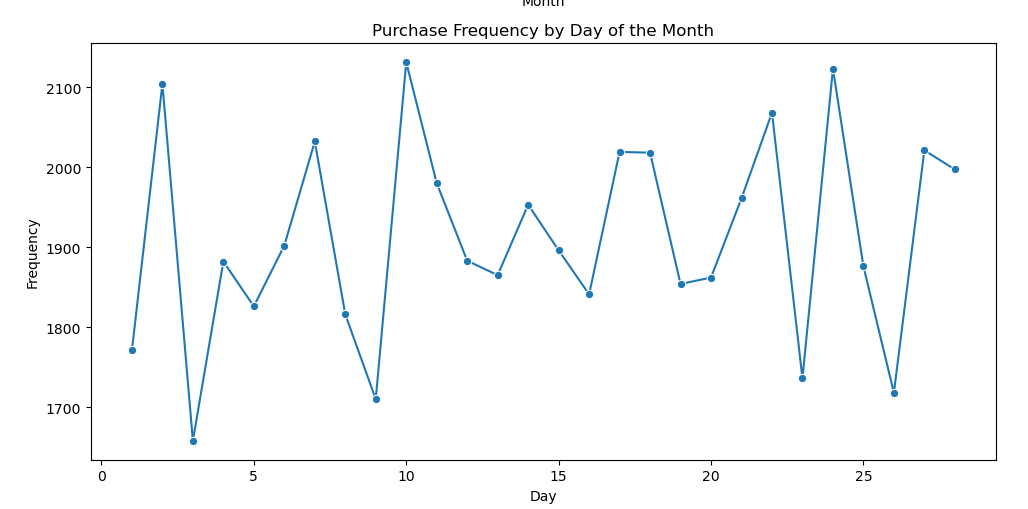


1. **Behavioral Data**



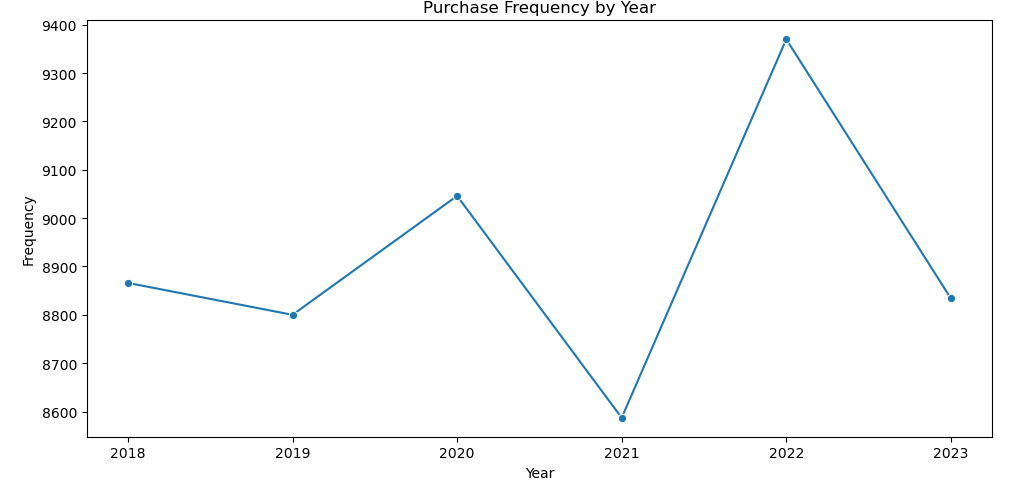
1. **Purchase Patterns**





* **Count**: There are 53,503 purchases in total.
* **Mean**: The average purchase date is around January 1, 2021, at approximately 14:46:59.
* **Minimum**: The earliest purchase date is on January 1, 2018, at 00:00:00.
* **25th Percentile (Q1):** 25% of the purchases were made before July 10, 2019.
* **Median (50th Percentile, Q2):** The median purchase date is on January 1, 2021, at 00:00:00, which means half of the purchases were made before this date and half after.
* **75th Percentile (Q3):** 75% of the purchases were made before June 28, 2022.
* **Maximum:** The latest purchase date is on December 28, 2023, at 00:00:00.

**From these statistics, we can make several conclusions:**

****

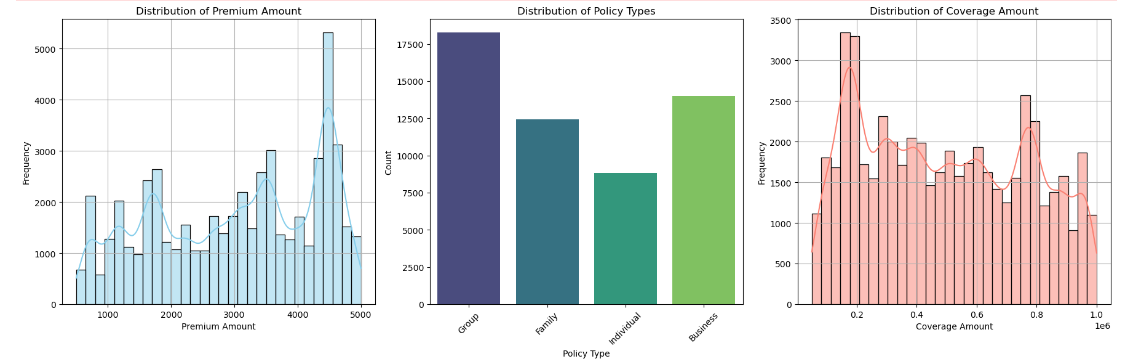
Trend over Time.

The data spans from January 1, 2018, to December 28, 2023. There seems to be a steady increase in purchases over time, as indicated by the mean shifting towards later dates.

Distribution

The spread of purchase dates is relatively balanced between the first and third quartiles, suggesting a relatively consistent distribution of purchases over time.

1. **Insurance products, Coverage and Premium Analysis**

****

Premium Amount Distribution:

The dataset consists of 53,503 premium amounts.

The mean premium amount is approximately $3,023.70, with a standard deviation of $1,285.83, indicating variability in premium amounts.

Premium amounts range from a minimum of $500 to a maximum of $5,000.

The median premium amount (50th percentile) is $3,194.00, which is close to the mean, suggesting a roughly symmetrical distribution.

The interquartile range (IQR), which spans from the 25th percentile ($1,817.00) to the 75th percentile ($4,311.50), indicates the middle 50% of the data.

Coverage Amount Distribution:

The dataset also consists of 53,503 coverage amounts.

The mean coverage amount is approximately $492,580.79, with a standard deviation of $268,405.51, indicating variability in coverage amounts.

Coverage amounts range from a minimum of $50,001 to a maximum of $1,000,000.

The median coverage amount (50th percentile) is $477,261.00, which is close to the mean, suggesting a roughly symmetrical distribution.

The interquartile range (IQR), which spans from the 25th percentile ($249,613.50) to the 75th percentile ($739,124.00), indicates the middle 50% of the data.

Policy Type Distribution:

The dataset contains four main types of policies: Group, Business, Family, and Individual.

The most common policy type is Group, with 18,255 occurrences, followed by Business (13,986), Family (12,424), and Individual (8,838).

Group policies have the highest representation in the dataset, followed by Business and Family policies.