**CHAPTER 1**

**INTRODUCTION**

* 1. **General Background**

In the agricultural value chain, farmers frequently encounter obstacles in the agricultural value chain that keep them from earning a just portion of the produce's final market price. It is essential to comprehend the scope of these problems and pinpoint the main causes in order to develop just and long-lasting solutions. In order to reduce the risks that farmers encounter such as crop failures, erratic weather patterns, and price volatility ,thus insurance is essential which helps stabilise farmers' incomes and lessens their reliance on unfair business practices.

* 1. **Problem Statement**

Farmers play a critical role in the agricultural value chain. Despite their essential contributions, they often face significant challenges, including unpredictable risks such as crop failures, natural disasters, and market price volatility. These uncertainties not only affect their earnings but also make it difficult for them to secure financial stability. As a result, many farmers remain uninsured or underinsured, leaving them vulnerable to financial distress. To address this, it is essential to analyze the factors affecting insurance pricing and develop predictive models to forecast premiums accurately.

* 1. **Objective**

The purpose of this report is to analyze the dataset and predict the share farmers receive from their agricultural outputs.

* 1. **Scope**

The project involves a comprehensive approach to the agricultural and insurance field, thus combining the merit of both. This study utilises the PMFBY dataset for an intensive study which are required to provide predictive insights. By analyzing a dataset on agricultural production and pricing, this study seeks to provide actionable insights that can guide policy changes and support efforts to create a fairer value distribution for farmers.

**CHAPTER 2**

**EXPLORATORY DATA ANALYSIS**

**2.1 Data Overview**

The dataset provides comprehensive information on the agricultural insurance scheme, covering details about policies, premiums, shares paid by various stakeholders, and other key attributes. The dataset provides insights into farmers' enrollment and insurance coverage under two key agricultural insurance schemes in India: the Pradhan Mantri Fasal Bima Yojana (PMFBY) and the Weather-Based Crop Insurance Scheme (WBCIS). The data spans from the inception of PMFBY in 2018 up to the Rabi 2022-23 season and includes district-wise statistics on enrollment, coverage, and financial details.

#### **1. Scheme and Seasonal Details**

sssyName.Year : Represents the year in which the farmer joined the scheme, combining State, Season, Scheme, and Year identifiers.

sssyName.seasonName : Denotes the season (e.g., Kharif, Rabi) during which the scheme was active.

sssyName.schemeName : Specifies the name of the insurance scheme under which the farmer is enrolled.

sssyName.stateName : Indicates the state in which the farmer is enrolled in the scheme.

#### **2. Regional and Crop Information**

#### level3Name : Refers to the district or local administrative region involved in the insurance scheme.

#### cropName : Indicates the type of crop insured under the scheme.

#### **3. Financial Details**

#### sumInsured : The total value of the crop insured against potential losses. This serves as the basis for calculating premium amounts.

#### premiumRate : The rate at which the insurance premium is calculated as a percentage of the sum insured.

#### stateShare : The percentage of the sum insured that the state government contributes towards the premium.

#### goiShare : The percentage of the sum insured that the Government of India contributes towards the premium.

#### farmerShare : The percentage of the sum insured that the farmer needs to pay as their share of the premium.

#### farmerShareValue : The monetary value of the premium amount paid by the farmer, calculated as sumInsured × farmerShare.

#### goiShareValue : The monetary value of the premium amount contributed by the Government of India.

#### stateShareValue : The monetary value of the premium amount contributed by the state government.

#### **4. Policy and Administrative Details**

#### sssyID : A unique identifier for the State-Season-Scheme-Year combination.

#### seasonID and seasonCode : Unique identifiers and codes representing specific seasons.

#### schemeID and schemeCode : Unique identifiers and codes for the insurance scheme.

#### stateID and stateCode : Unique identifiers and codes for states.

#### level3ID and level3Code : Unique identifiers and codes for districts.

#### cropID and cropCode : Unique identifiers and codes for crops.

#### **5. Temporal Information**

#### Year : The specific year when the scheme or policy was active.

#### startDate and endDate : The start and end dates of the scheme or policy period.

#### policyStartDate and policyEndDate : The start and end dates of the insurance policy coverage period.

#### yieldEndDate : The date on which the yield estimation for the insured crop concludes.

#### **Indemnity Details**

IndemnityLevel : Indemnity Level refers to the percentage of the sum insured that farmers are eligible to receive as compensation in case of crop loss due to insured perils. Specifies the percentage of the sum insured that farmers are eligible to receive as compensation in case of crop loss.

Indemnity levels typically range between 70%, 80%, and 90%, depending on the crop type and associated risks.

#### thresholdYield : The benchmark yield level linked to the indemnity level, below which compensation is provided.

Farmers are eligible for compensation when the actual yield falls below the Threshold Yield. The amount of compensation is determined based on the extent of the yield shortfall and the applicable indemnity level.

#### **7. Insurance Company Information**

#### insuranceCompanyName : The name of the insurance company providing coverage under the scheme.

#### insuranceCompany.insuranceCompanyID and insuranceCompany.insuranceCompanyCode : Unique identifiers and codes for the insurance company.

#### headQuaterAddress, headQuaterEmail, and tollFreeNumber : Contact details of the insurance company's headquarters.

#### websiteLink : Official website link of the insurance company for additional information.

#### **8. Policy Details**

#### isOpen : Indicates whether the scheme is still active or closed.

#### cnStarted : Denotes the year in which the scheme commenced.

#### isPreviousSeasonYearInSubsidy : A Yes/No field indicating whether subsidies for the previous season/year are included.

#### Policy : Details of the specific insurance policy issued.

#### Notification : Notifications associated with the policy or scheme.

#### isOfflineChallan : Indicates whether the payment or subsidy process is done offline.

#### firstGoiSubsidy : Represents the first subsidy amount released by the Government of India.

#### goiOfflineChallan and stateOfflineChallan : Specific details about offline payments made by the Government of India and state governments.

#### **9. Miscellaneous Details**

#### categoryName : Represents different categories (14 in total) to which crops or policies belong.

#### cropType : The type of crop insured (e.g., Kharif or Rabi crops).

#### Unit : Represents the insurance coverage unit. Most rows have a value of 1, with only 2 rows having a value of 2.

#### pickingType : Represents specific collection or grouping mechanisms within the dataset.

#### type: Classification or categorization associated with the data record.

**2.2 Data Characteristics**

Size: The dataset includes district-level enrollment and coverage data, which covers a large number of records across various districts, crops, and states. The data spans several years, beginning from 2018, and continues up to the Rabi 2022-23 season. There are abut 29999 rows and 62 columns. Thus the dataset implies data records of 29999 incidents w.r.t 62 features which are taken.

Time Span:The data covers the period from the introduction of the PMFBY scheme in 2018 up to February 2023, ensuring a wide temporal range for analysis. This includes multiple agricultural seasons (Kharif, Rabi), allowing for seasonal comparisons.

The dataset contains a variety of data types, including:

* Integer (int64): Used for categorical variables like IDs (e.g., sssyID, seasonCode, stateCode) and numerical values (e.g., year, schemeCode).
* Float (float64): Used for numerical variables such as sumInsured, premiumRate, stateShare, and goiShare.
* Object (object): Used for textual data, including categorical variables like seasonName, stateName, cropName, and insurance company details.
* Boolean (bool): Used for flags, such as whether a record has specific characteristics (e.g., isOfflineChallan).

Non-null Counts: All columns in the dataset have 29999 non-null entries, indicating that the dataset does not have missing values for the variables in question. This suggests that the data has been adequately prepared and cleaned for analysis.

Handling Missing Values: There are no missing values, as all 29,999 rows are complete.

Data Synchronization: Efforts have been made to synchronize the state and district names between the enrollment statistics and coverage datasets, though some discrepancies might remain.

**2.3 Data Transformation**

Column Removal:

As part of the data preprocessing, some columns were renamed for clarity and consistency, while irrelevant or redundant features were removed to streamline the dataset for analysis.

To streamline the dataset and focus on relevant variables for the analysis, several columns were removed, including those related to administrative IDs, season names, and insurance company details. This transformation resulted in a reduced set of features that better align with the goals of the project.

The following features columns were dropped: sssyName.seasonName, sssyName.schemeName, seasonID, schemeID, schemeCode, level3Name, stateID, stateCode, level3ID, level3, level3Code, cropName, cropID, cropCode, pickingType, sssyID, year, policyStartDate, policyEndDate, isOfflineChallan, goiOfflineChallan, stateOfflineChallan, yieldEndDate, currentTime, default, insuranceCompanyName, cutOfDate, tollFreeNumber, headQuaterAddress,headQuaterEmail,websiteLink,insuranceCompany.insuranceCompanyCode,insuranceCompany.insuranceCompanyID, isOpen, cnStarted, unit, ayTy, Scheme, Start.

Colomn Rename:

To enhance clarity and improve readability, several columns were renamed to more intuitive and consistent names. The following changes were made:

sssyName.year was renamed to Year

sssyName.stateName was renamed to State

sssyName.schemeNameShort was renamed to Scheme

insuranceCompany.insuranceCompanyName was renamed to insuranceCompanyName

Row Removal:

Also WBCIS (Weather based Crop Insurance Scheme) data rows is being removed from the dataset as it not needed in the part of predictive analysis.

Rearranged Data:

To ensure a more focused analysis, the columns of the dataset were rearranged to highlight the most relevant features for the study. The following columns were selected and reordered:

Key Features: Year, State, sumInsured, premiumRate, farmerShare, farmerShareValue, stateShare, stateShareValue, goiShare, goiShareValue.

Additional Information: indemnityLevel, seasonCode, isPreviousSeasonYearInSubsidy, firstGoiSubsidy, categoryName, cropTypeCode

Insurance Details: startDate, endDate, insuranceCompanyName

**2.4 Null values Identification**

There are no missing values. We can identify the missing values in the dataset by using the syntax “dataset.isnull().sum()”.

**2.5 Rearrangement as Categorical and Numerical columns**

Categorical columns contain non-numeric data or data that can be categorized, such as states, schemes, or boolean indicators.

They are typically used for:

1. Grouping and aggregation.
2. Encoding for predictive modeling.
3. Analyzing categorical trends.

Included categorical columns are

‘State','Scheme','schemeCode','isPreviousSeasonYearInSubsidy','firstGoiSubsidy','seasonCode','categoryName','cropTypeCode','startDate','endDate','insuranceCompanyName','indemnityLevel'.

Numerical columns contain numeric data used for calculations, analysis, and predictive modeling.

Included numerical columns are

‘sumInsured','premiumRate','farmerShare','farmerShareValue','stateShare','stateShareValue','goiShare',goiShareValue’.

Also in this Startdte and EndDate columns are vdroppeed and are replced with Policy Term.

To analyze the duration of insurance coverage, a new column, Policy Term (Days), was created. This column represents the number of days between the start and end dates of each policy.

The difference (endDate - startDate) is computed in days using .dt.days and stored in a new column, Policy Term (Days).

This represents the policy duration, providing valuable insights into how long each insurance coverage lasted.

**2.6 Removal of Duplicate rows**

This method removes duplicate rows based on all columns by default.

If two or more rows have the same values across all columns, only the first occurrence is kept, and subsequent duplicates are dropped.

It returns a new DataFrame with duplicates removed, leaving the original DataFrame unmodified.

We can identify the duplicate values in the dataset by using the syntax “dataset.drop\_duplicates()”.

**2.7 Display count of each category in categorical columns**

We can identify the duplicate values in the dataset by using the syntax “dataset.['Category'].value\_counts”.

For example here we have found of the count for each state as a represented way.

So we got anumber for highest in Tamil Nadu as 2979 and lowest in Andaman and Nicobar as 20.

**2.8 Graphical representation for the features of dataset**

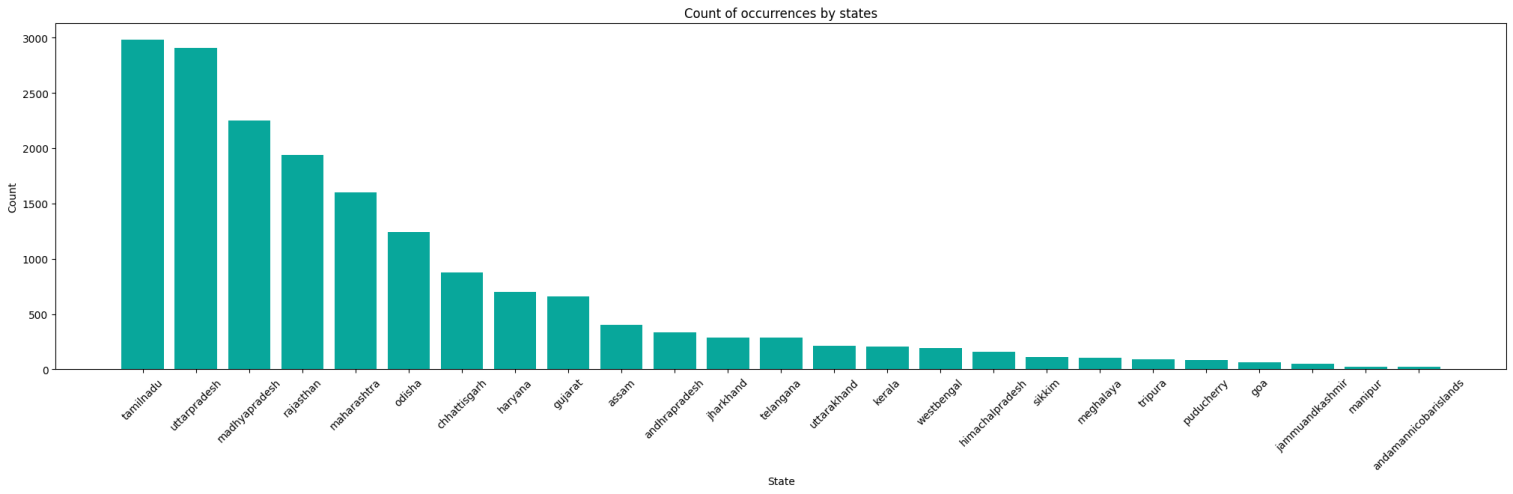


Fig.1. Count of Occurances by state

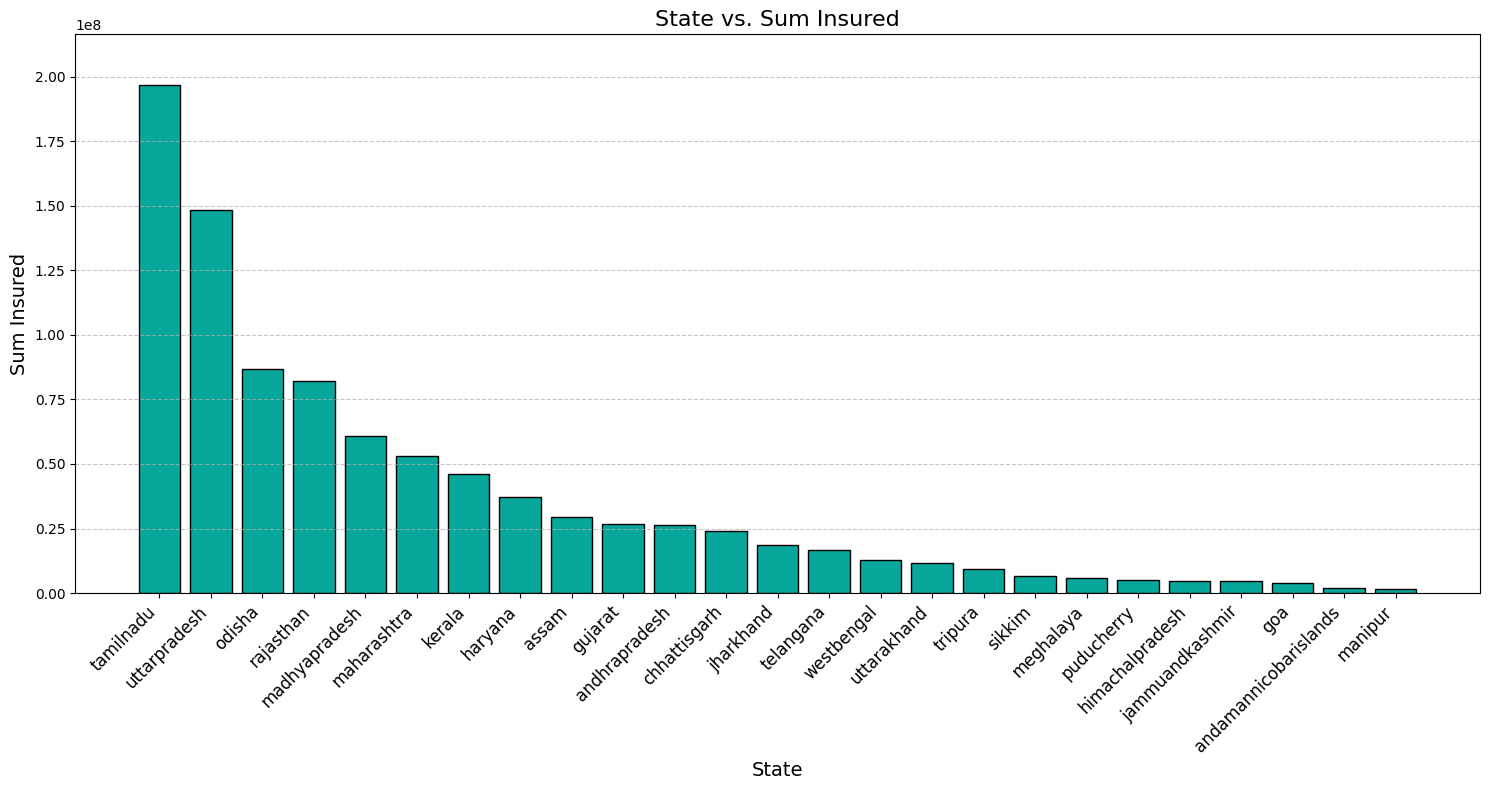


Fig.1. State vs SumInsured

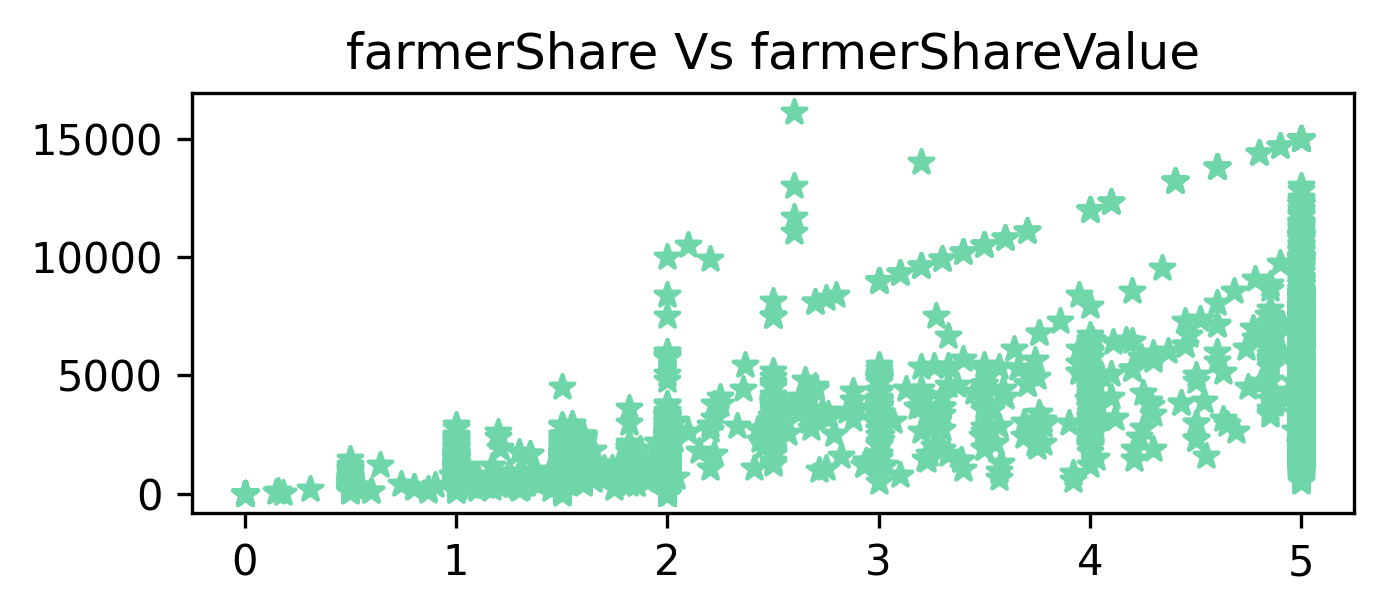


Fig.1. farmersshare vs farmers ShareValue

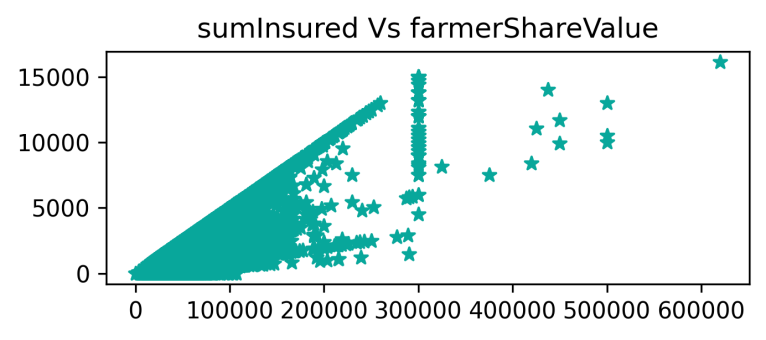


Fig.1. sumInsured vs sumInsuredShareValue

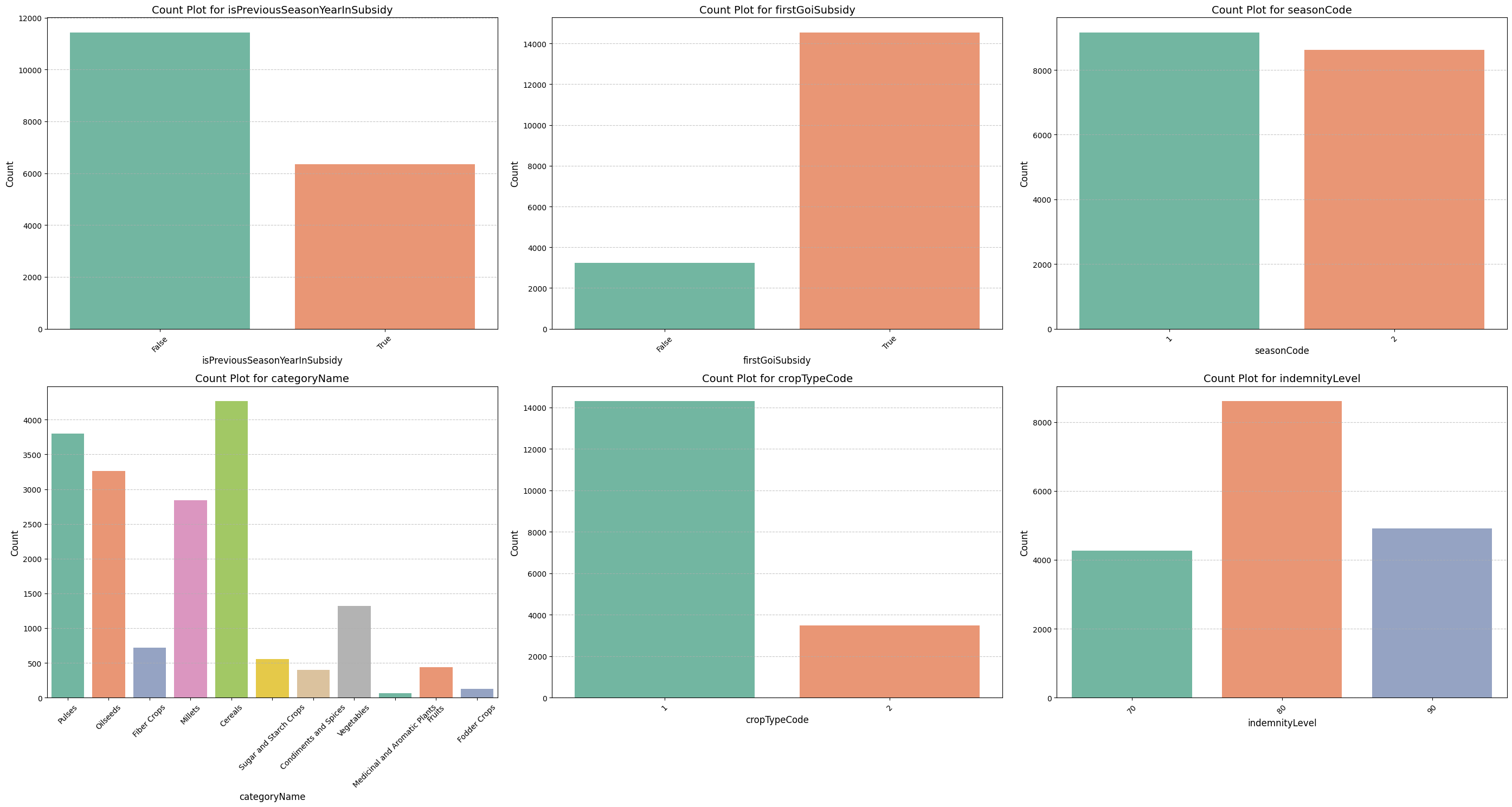


Fig.1.Count plot for CategoriyName

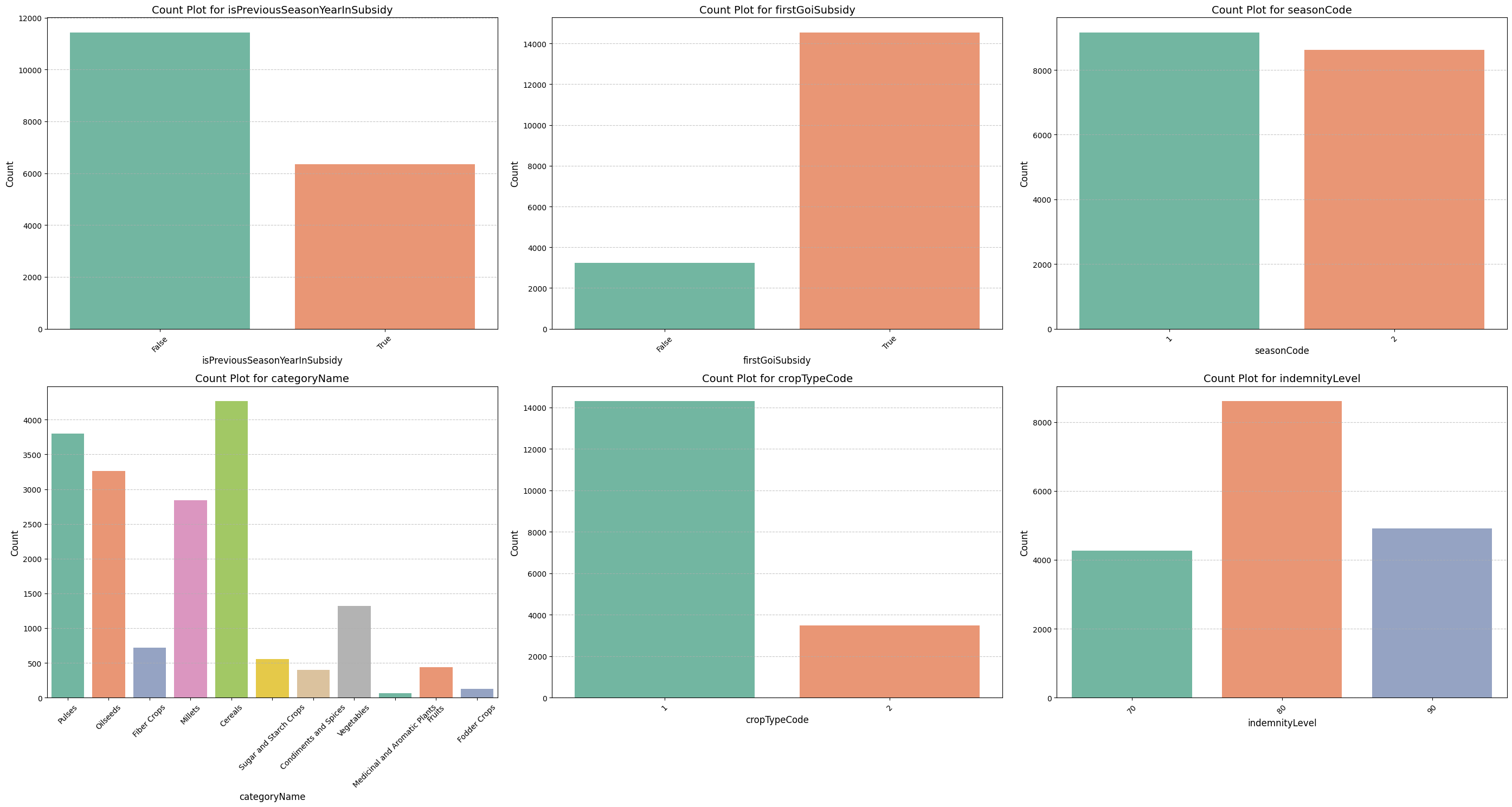


Fig.1. Count Plot for Previous YearSubsidy

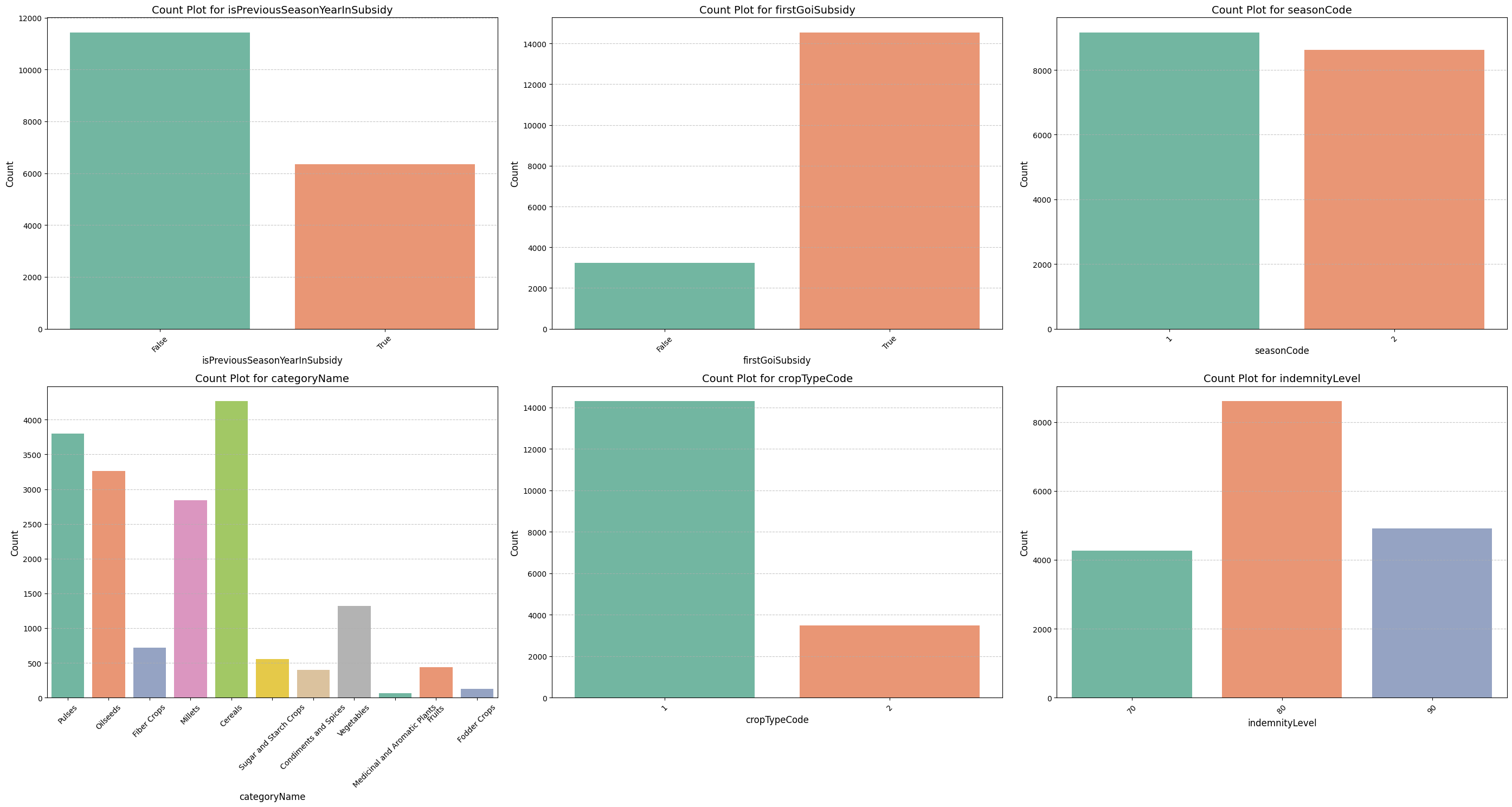


Fig.1. Count Plot for First Govr of India Subsidy

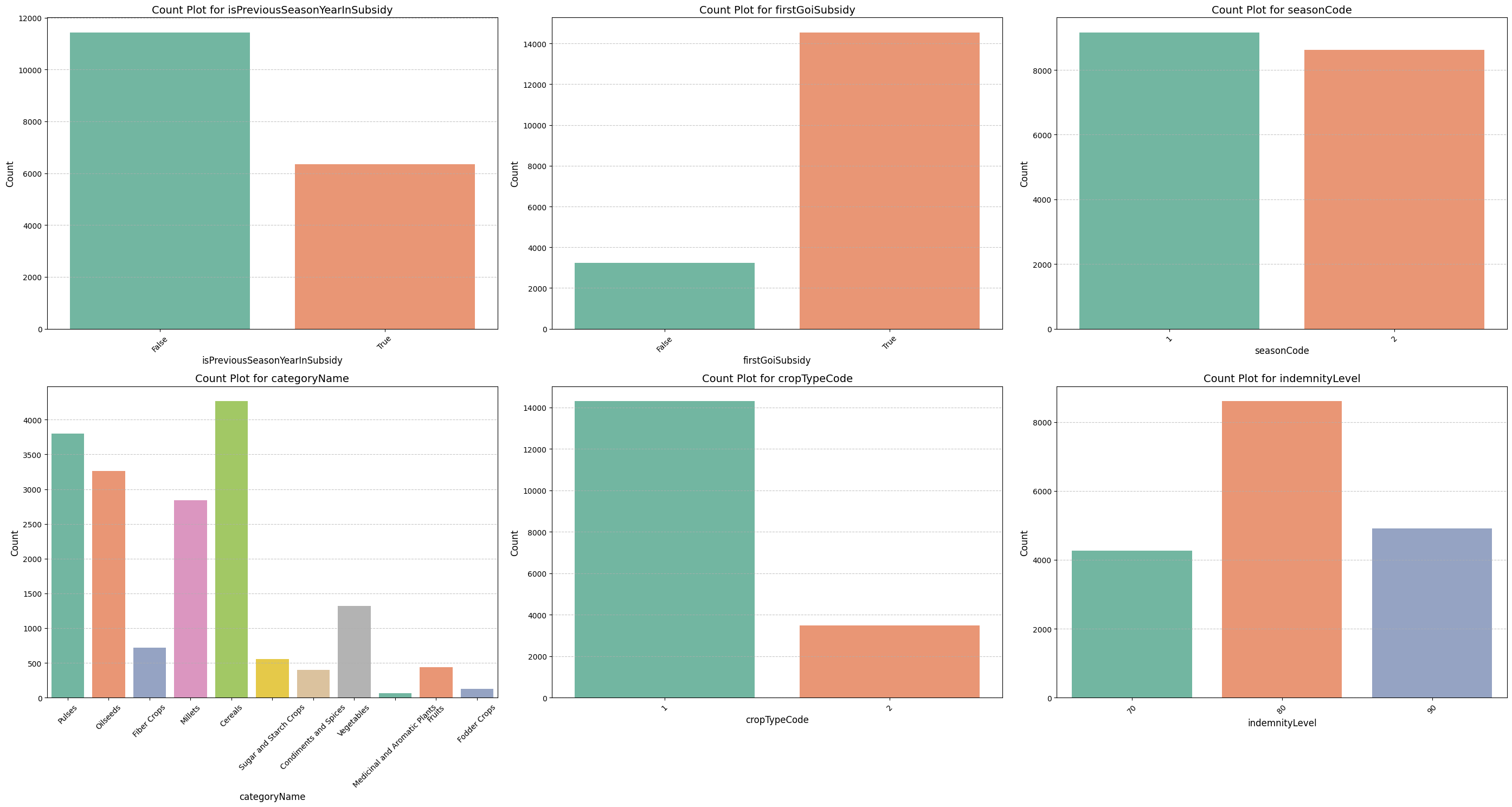


Fig.1. Count Plot for Crop Type

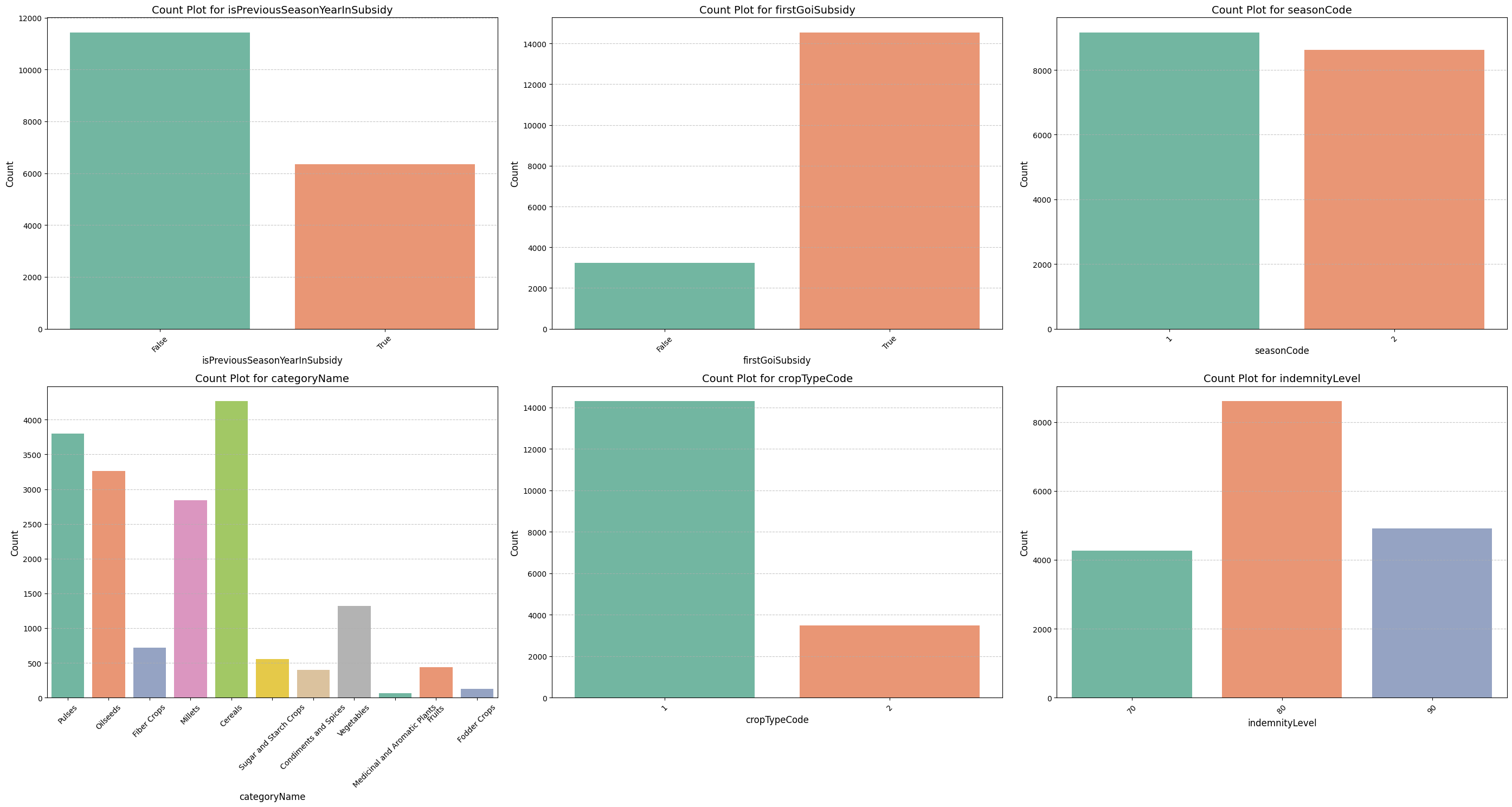


Fig.1. Count Plot for indemnity level

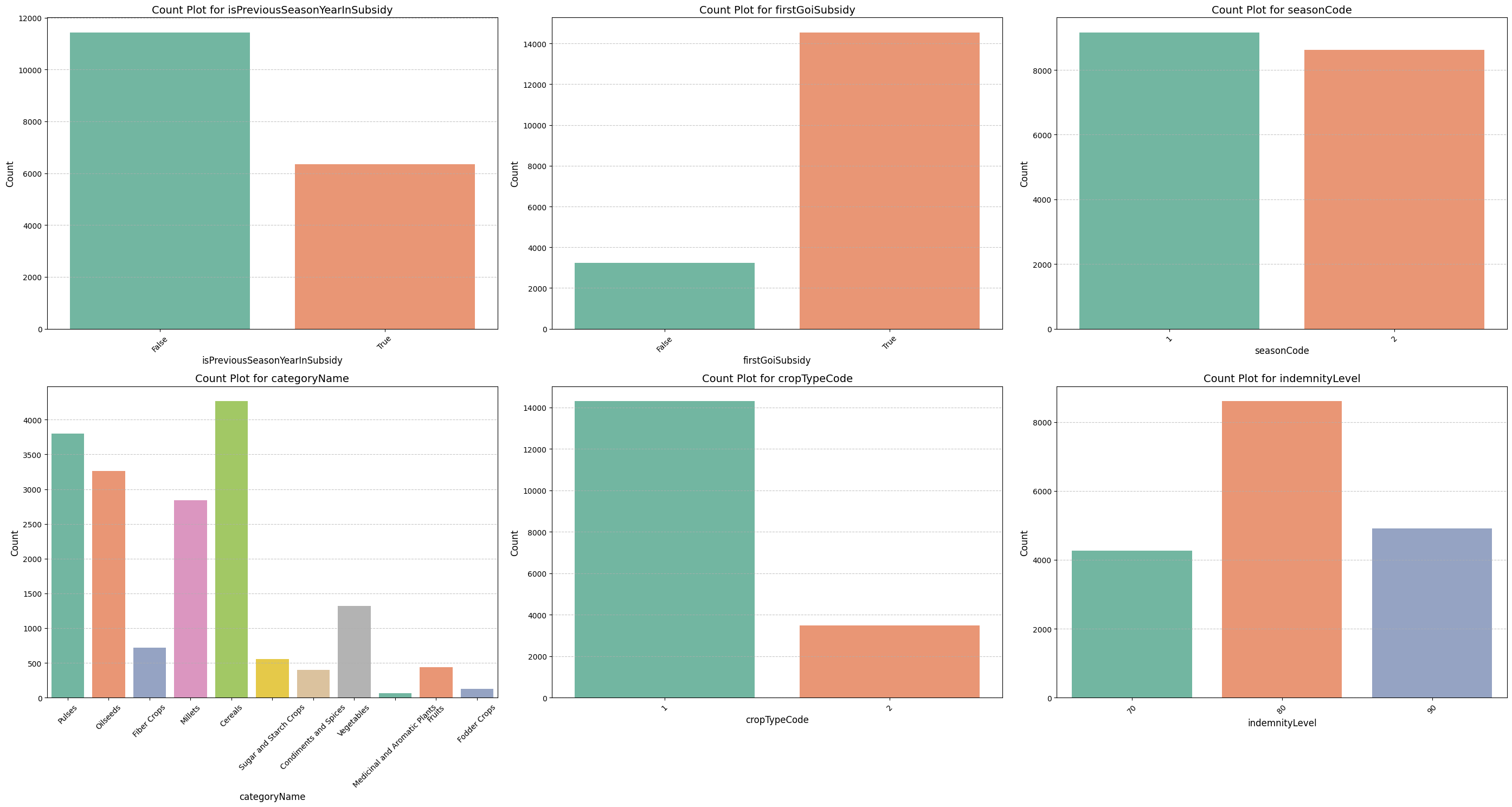


Fig.1. Count Plot for seasonCode

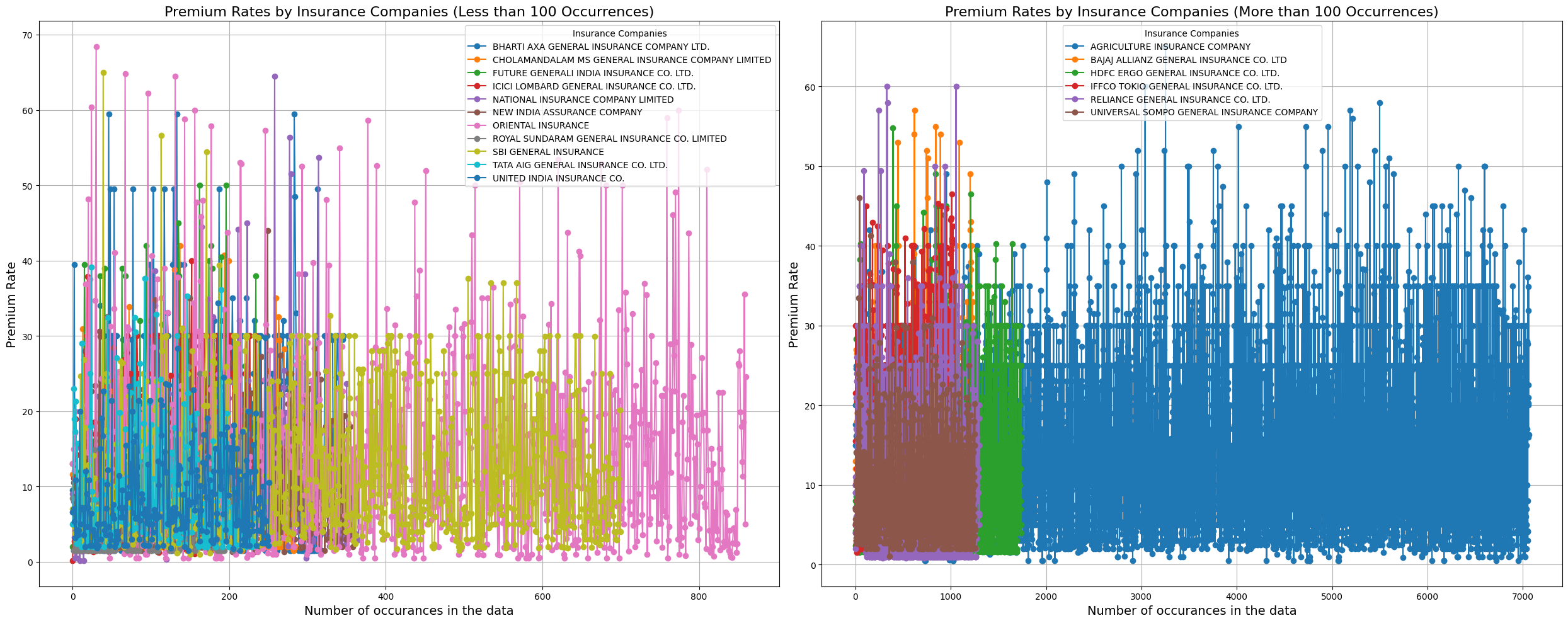


Fig.1. Premium Rates by Insurance Companies (Less than 100 Occurrences)

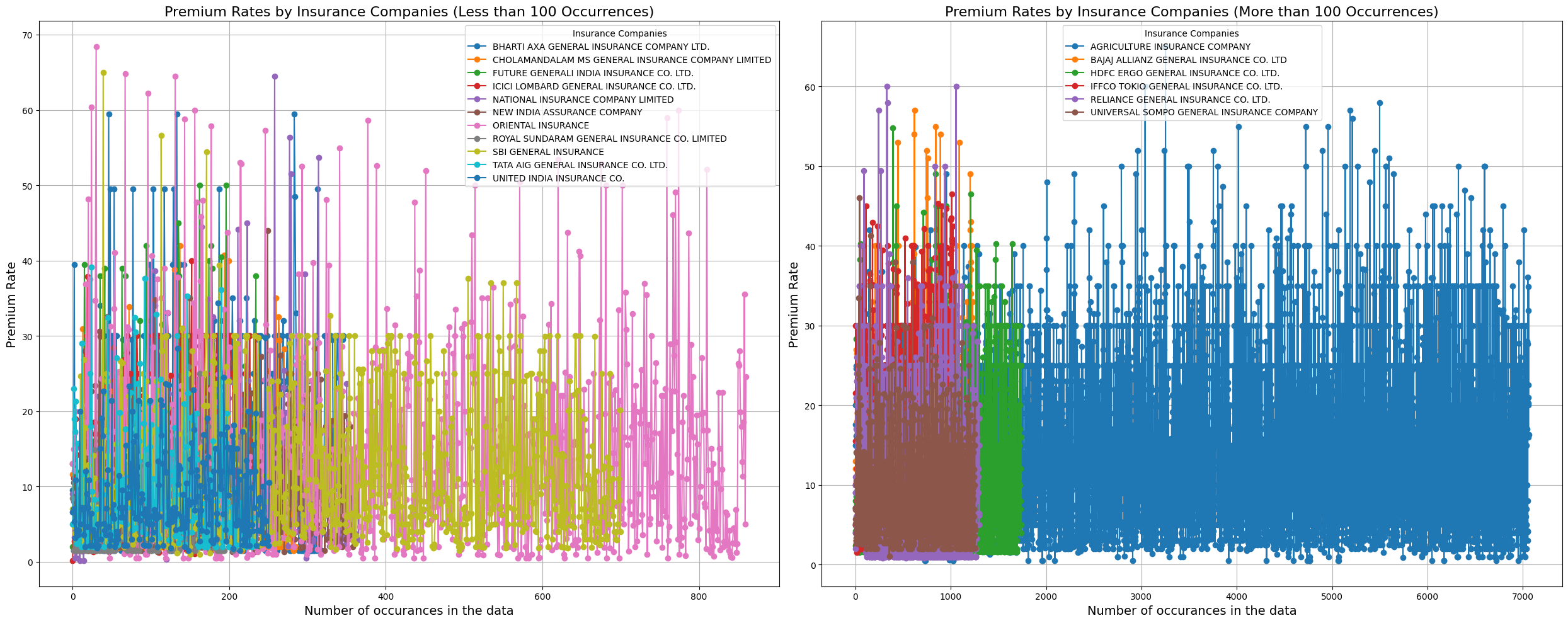


Fig.1. Premium Rates by Insurance Companies (More than 100 Occurrences)

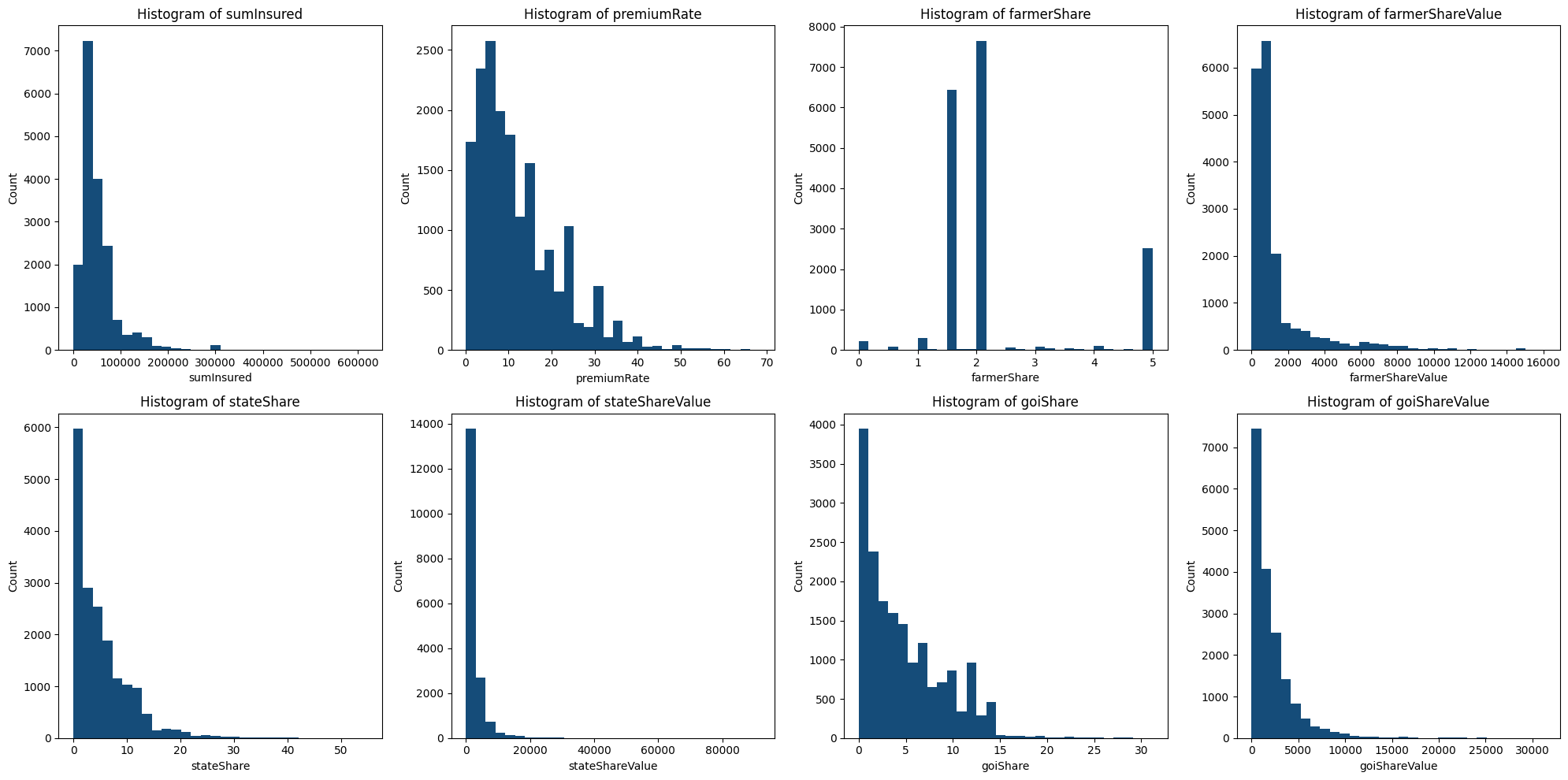


Fig.1. Histogram

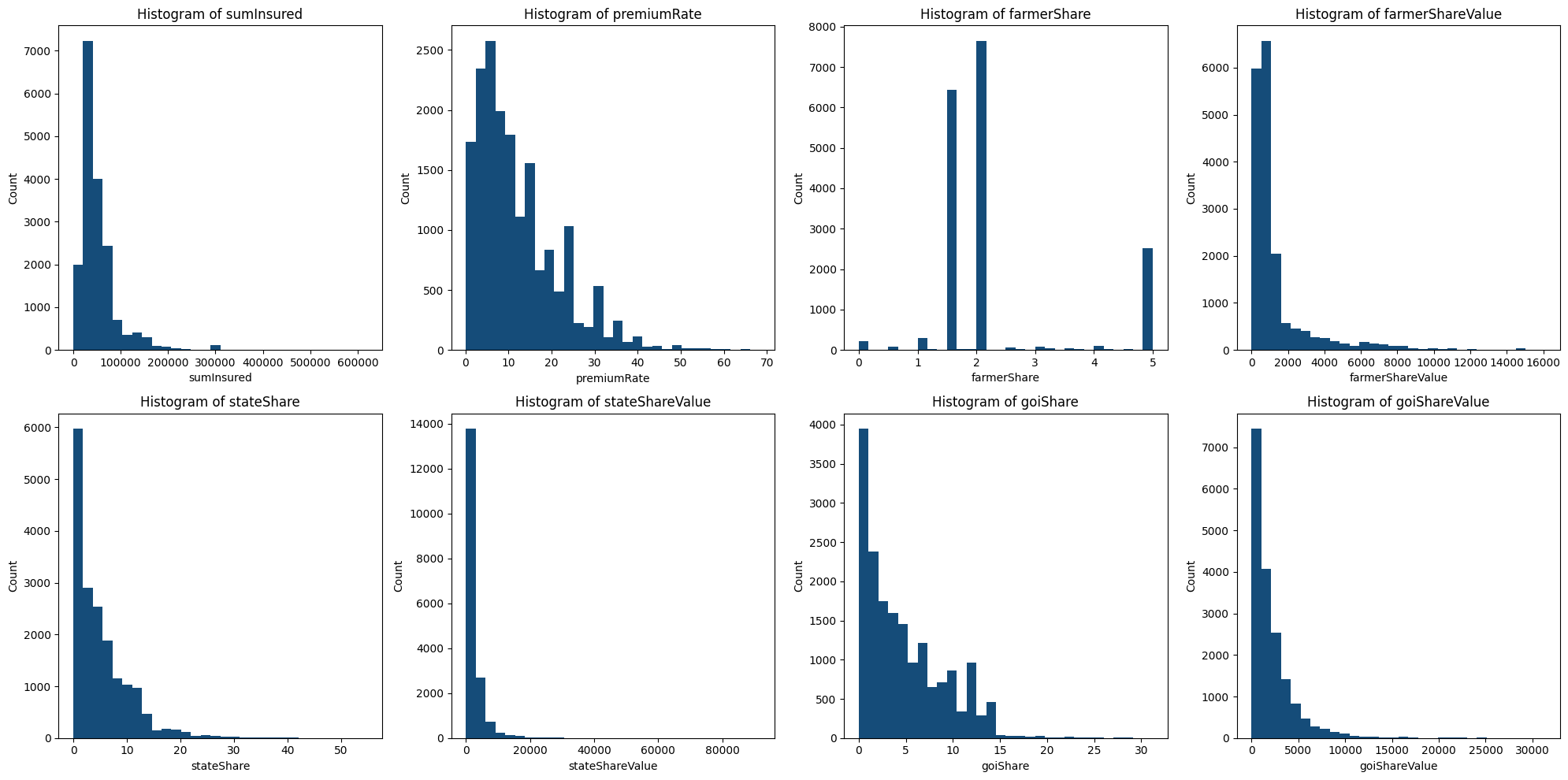


Fig.1. Histogram

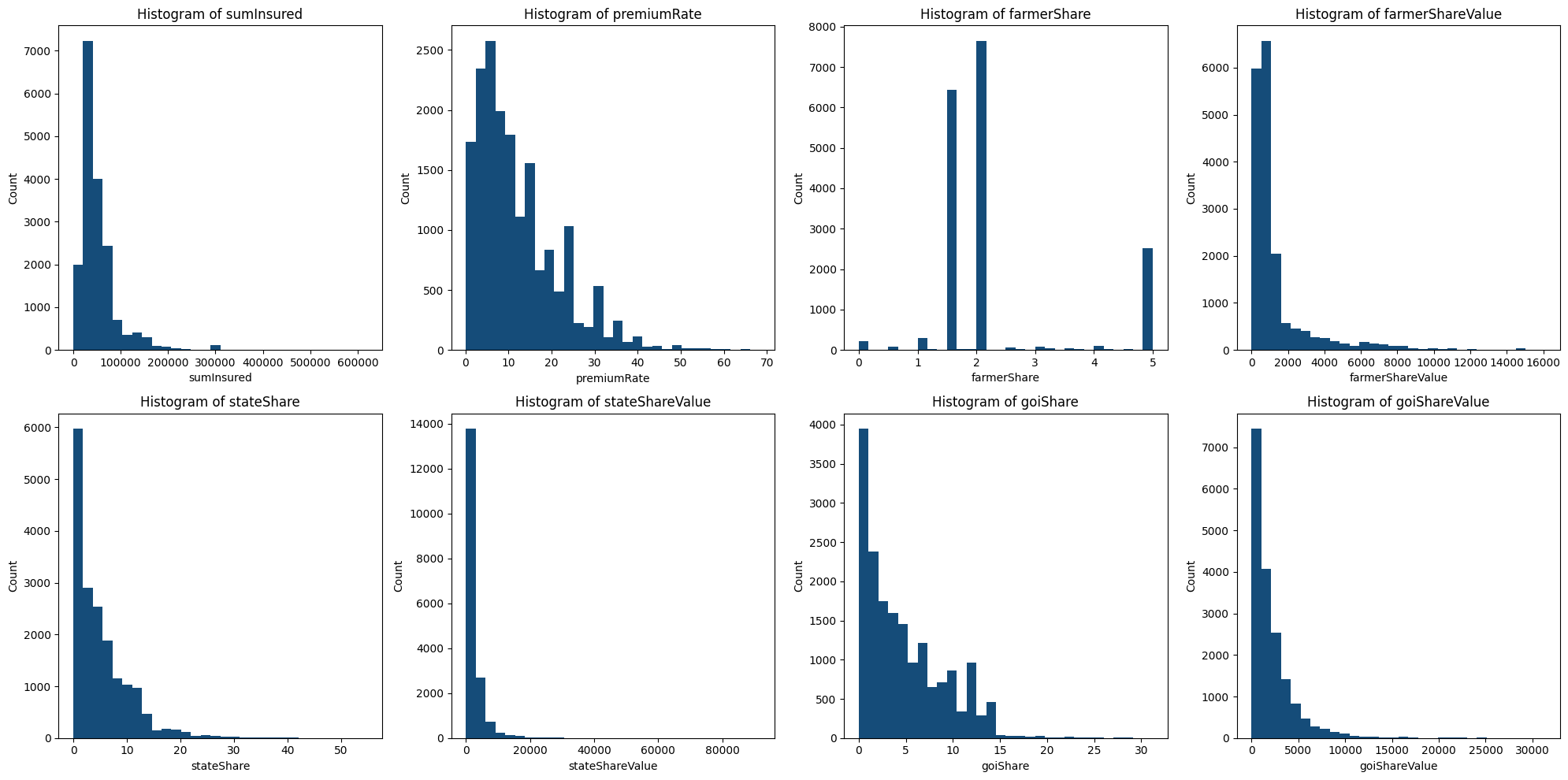


Fig.1. Histogram

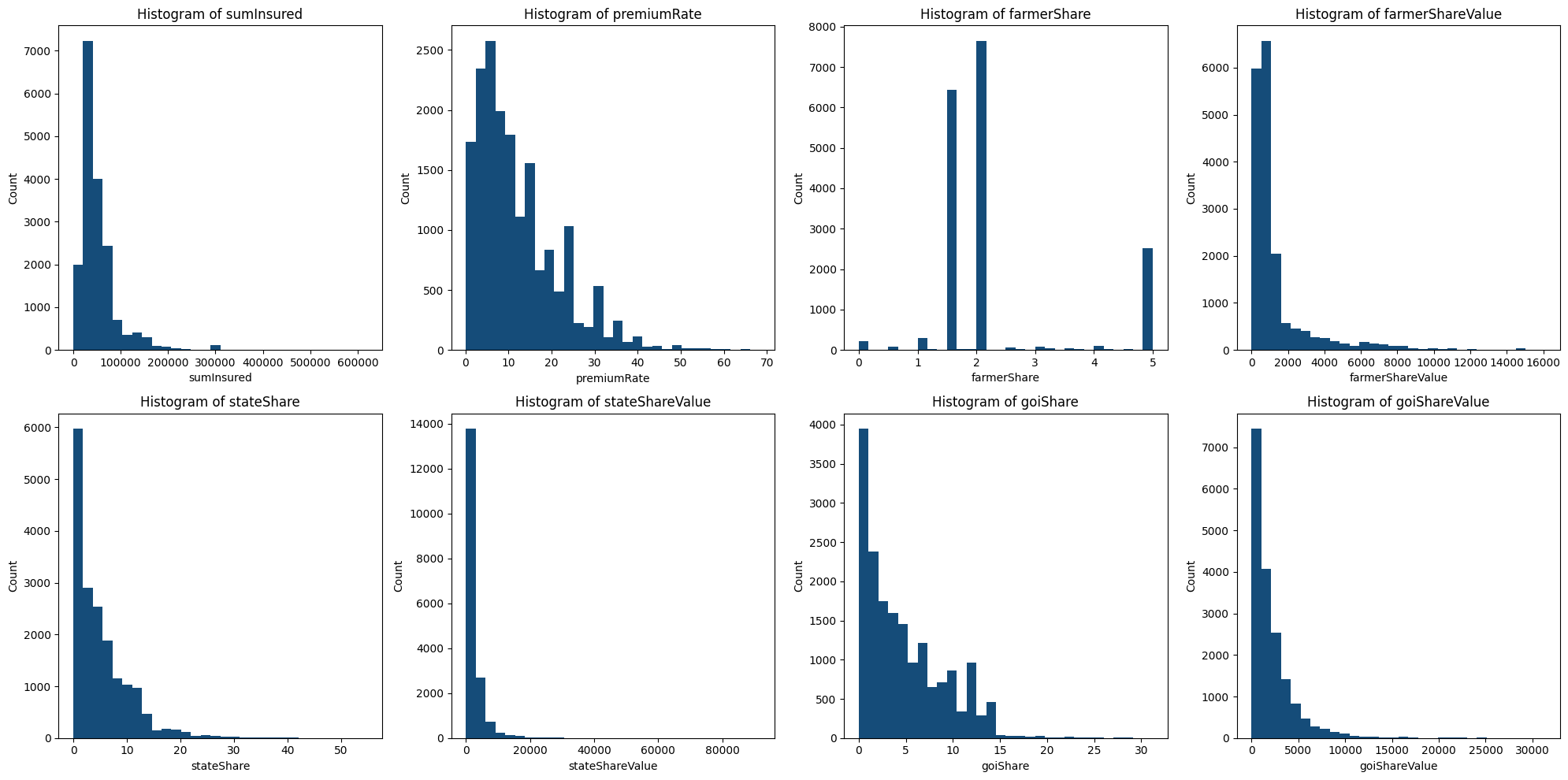


Fig.1. Histogram

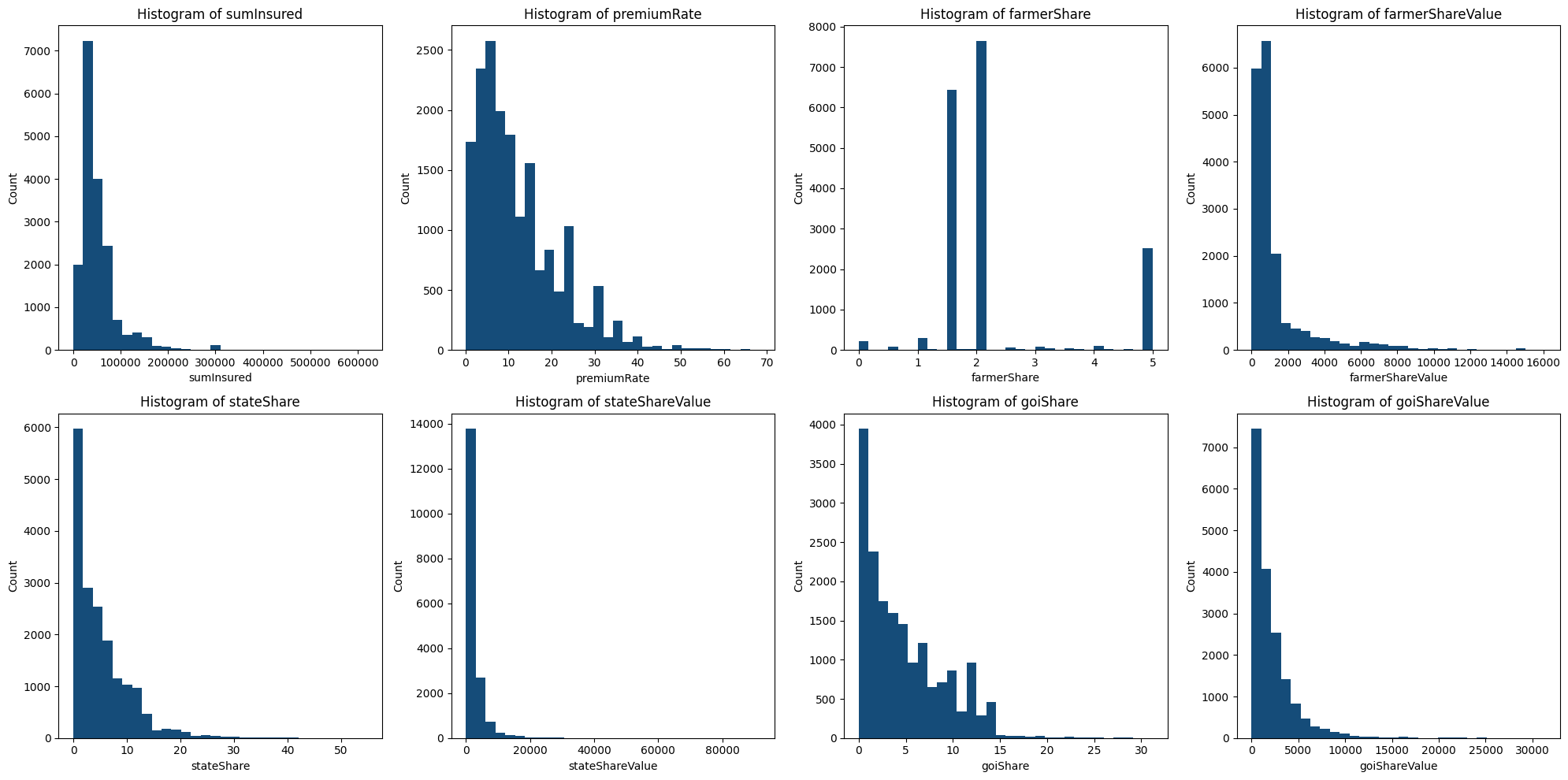


Fig.1. Histogram

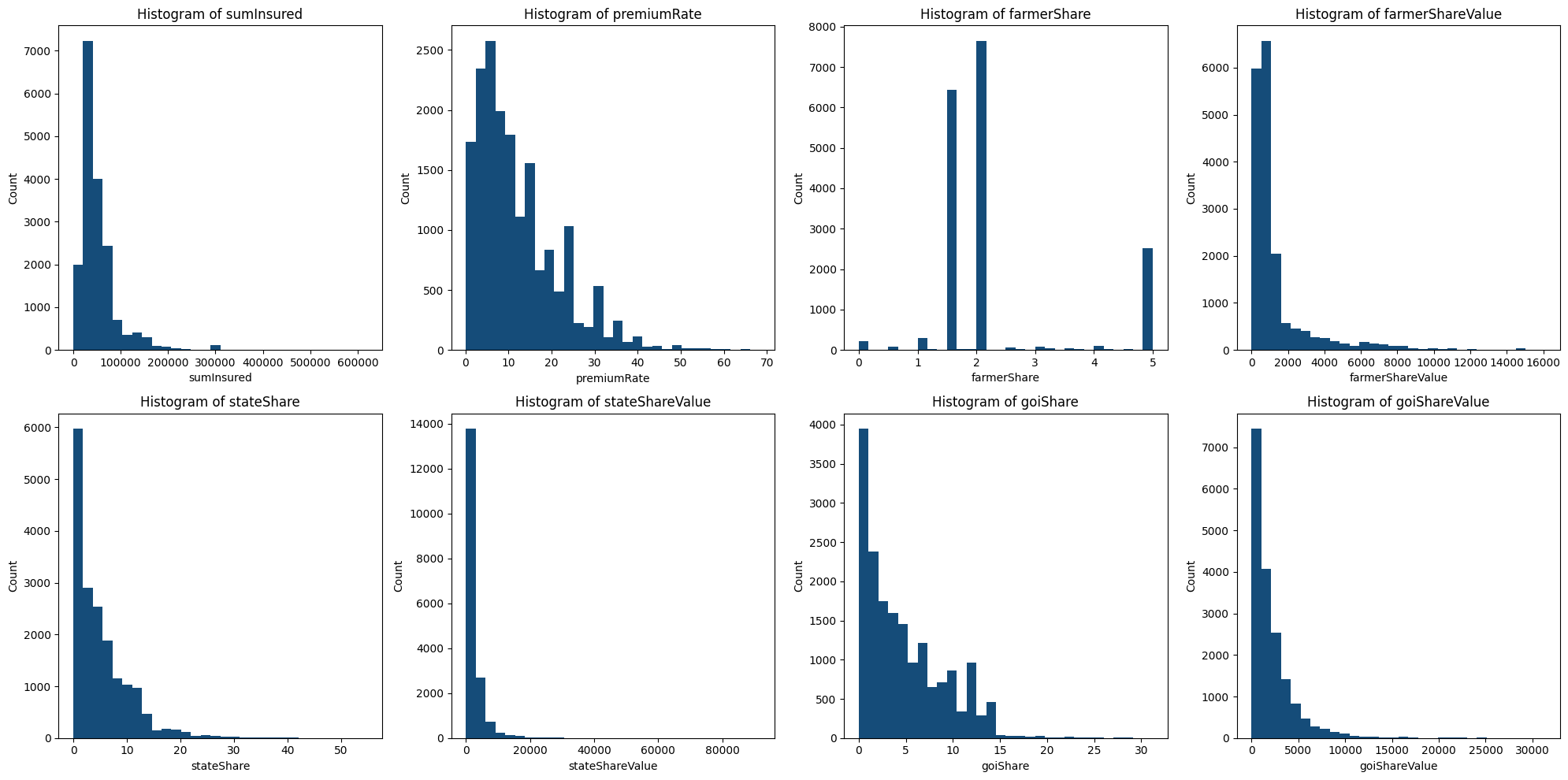


Fig.1. Histogram

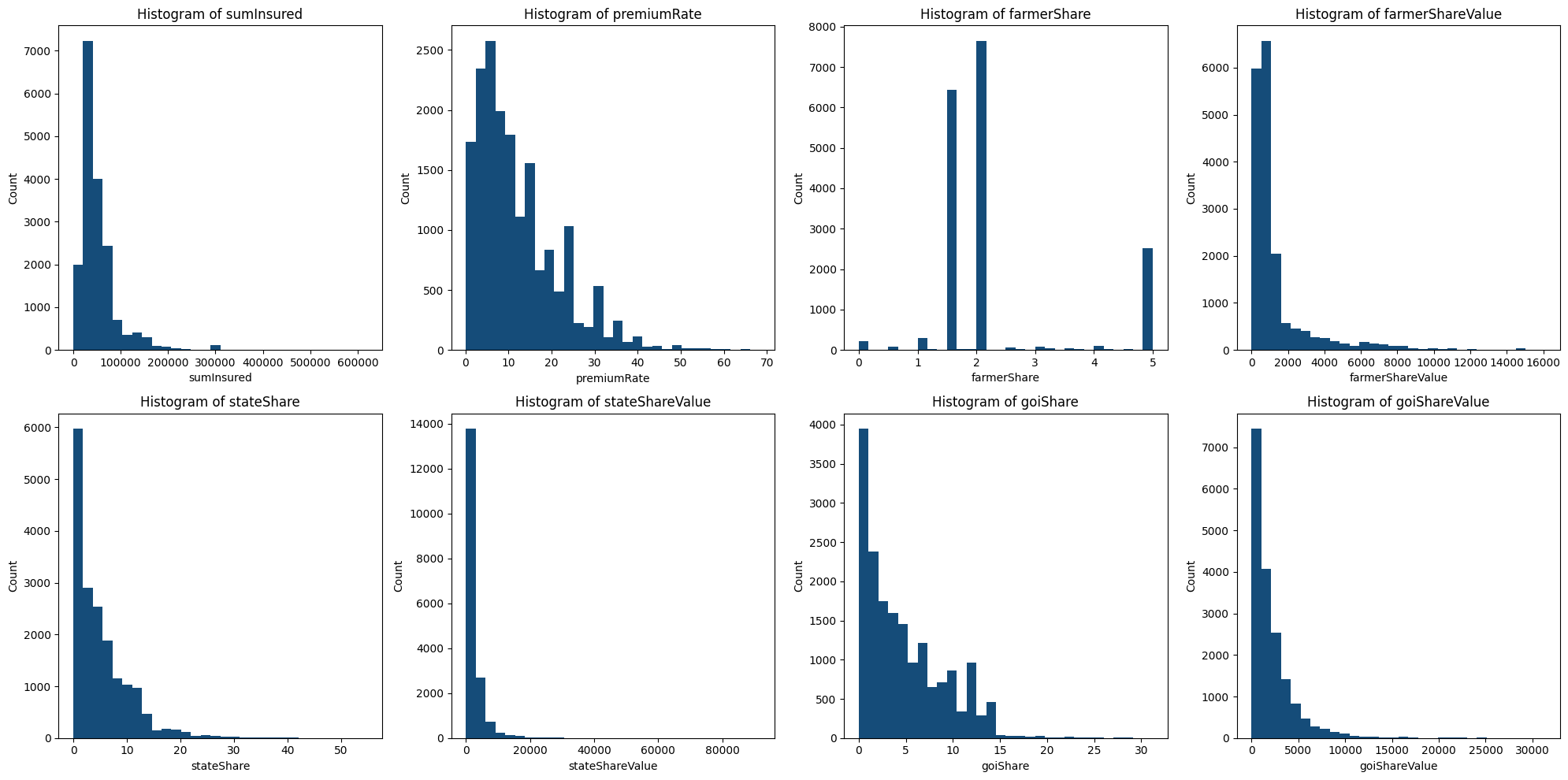


Fig.1. Histogram

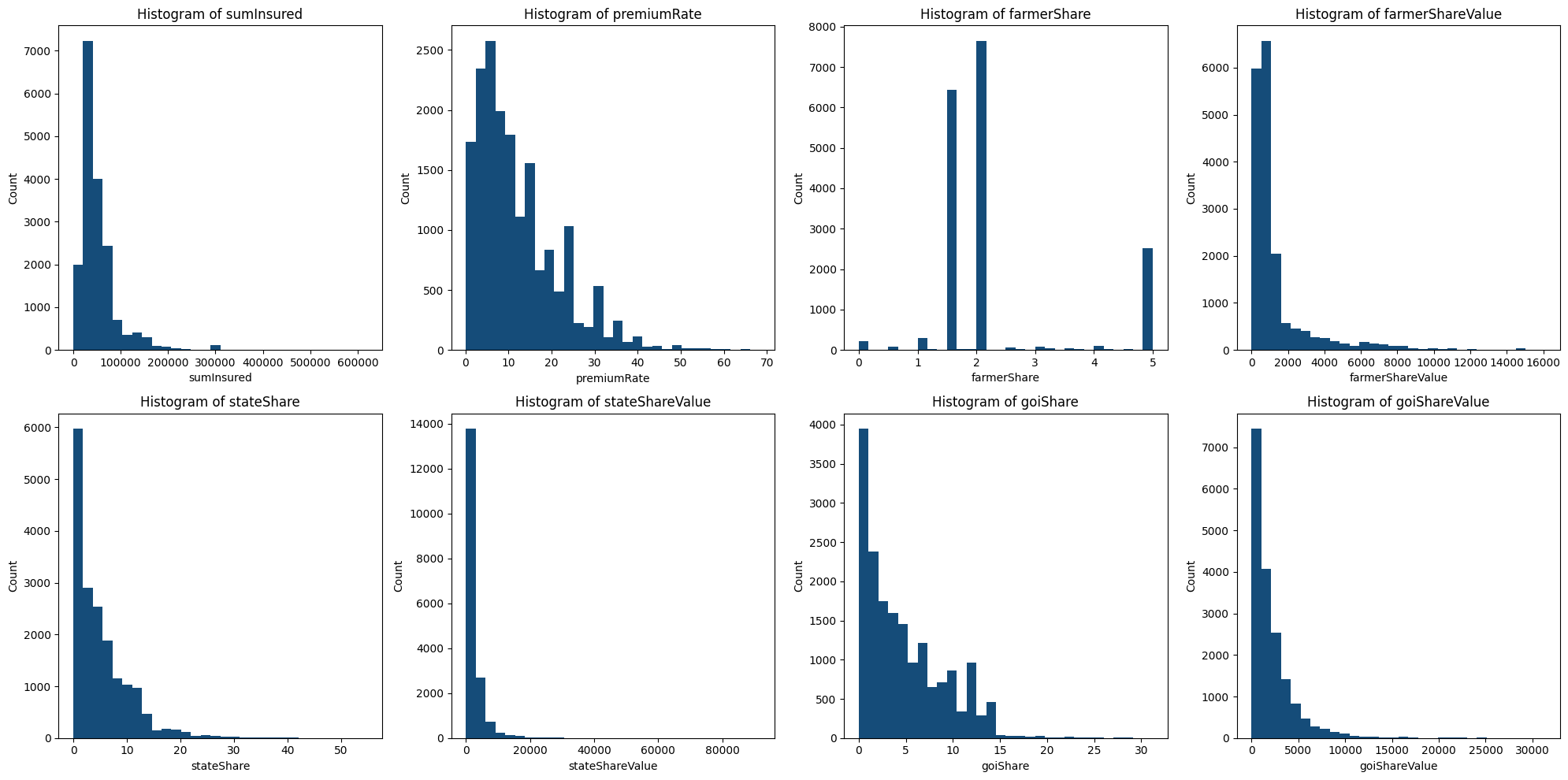


Fig.1. Histogram

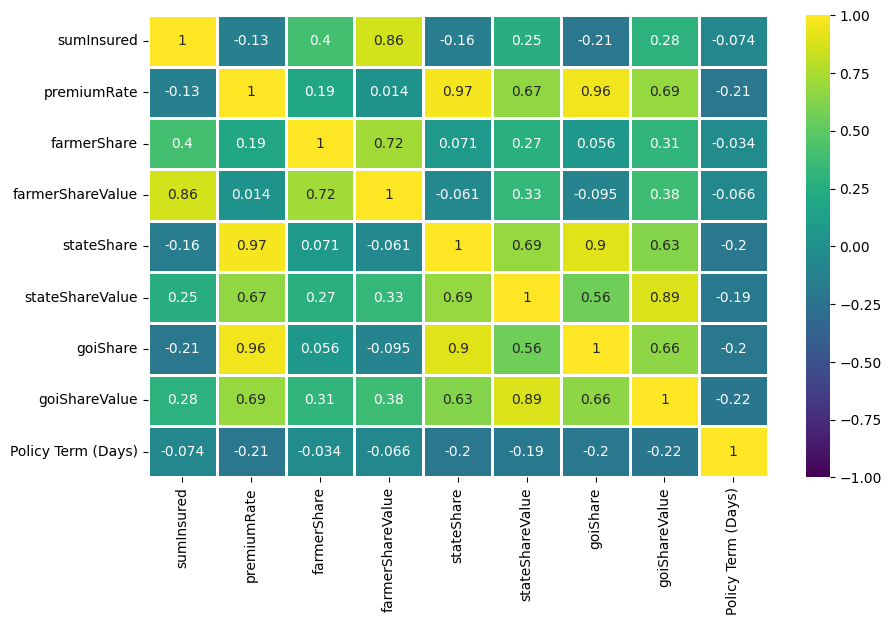


Fig.1. Correlation between numerical columns

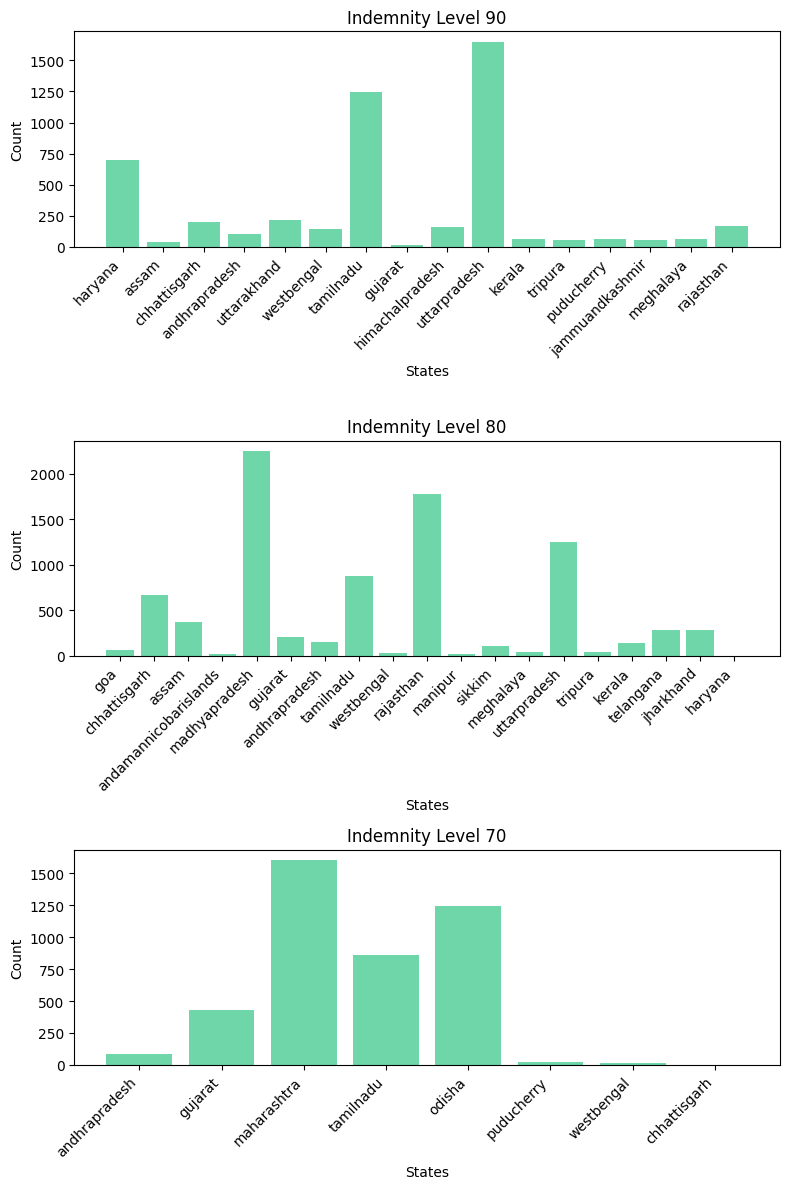


Fig.1. Indemnity levels

**CHAPTER 3**

**METHODOLOGY**

**CHAPTER 4**

**RESULT AND OBSERVATION**

**CHAPTER 5**

**CONCLUSION AND FUTURE SCOPE**

**REFERENCES**