

**Acknowledgment**

I would like to express my special thanks and gratitude to my teacher Deepak Karna, who continuously motivated me and gave me an idea about the study of data management using SAS studio. It was his consistent effort and unique way of teaching, and his experiences that made the subject so interesting to study and complete this report.

**Executive Summary**

One of the most conspicuous trends of today’s world is a colossal upsurge in the management of data in an organization, thanks to the increasing value of data, which have not only made the company more money, but is also effective for future decision-making processes. Needless to say, the management of data is vital and must be improved using the machine learning and feature engineering techniques. SAS is one strong tool that helps in machine learning, and we shall use it more to study, analyze, clean, and make the best use of the data. The ‘PROC CONTENT’ and ‘PROC FREQ’ are some important codes in the SAS which perform the variable analysis and cross-tab analysis efficiently. With the help of SAS studio, understanding each variable inside the dataset in the initial data exploration stage is possible and the data then have to be cleaned, handled, and fix the noisy and inconsistent data in the pre-processing processes. The exploratory data analysis is another technique that works with EDA graphs such as scatter charts, histogram, descriptive analysis, univariate analysis, and correlation and find the outliers within the data. For Apache Hadoop, Amazon web services tools including the EMR, AWS Glue, Athena, S3 bucket and Amazon Quick Sight have been used. Furthermore, the hypothesis, involving the Chi-Square Test and Anova tests has been proved to be effective in predicting that high loan approval was for those with high income, less debt, more likely a female population, and married population. These future engineering techniques in machine learning, and with the use of SASs has been a great prediction tool for the “XYZ Loan” company.

**TABLE OF CONTENTS**

[CHAPTER 1: Introduction to Feature Engineering 2](#_Toc106827621)

[CHAPTER 2: Literature Review / Related work 3](#_Toc106827622)

[CHAPTER 3: Principal Component Analysis 4](#_Toc106827623)

[CHAPTER 4: Feature extraction 5](#_Toc106827624)

[CHAPTER 6: Label encoding 5](#_Toc106827625)

[CHAPTER 7: One-hot encoding 5](#_Toc106827626)

[CHAPTER 8: Metadata - Exploration of the dataset given 6](#_Toc106827627)

[CHAPTER 9: Conclusion 15](#_Toc106827628)

[CHAPTER 10: References 16](#_Toc106827629)

List of Figures

[Figure 1: Code for importing and checking the content 7](#_Toc106827630)

[Figure 2: The Contents procedure 8](#_Toc106827631)

[Figure 3: Engine/Host dependent information 8](#_Toc106827632)

[Figure 4: Alphabetic List of Variables and Attributes 9](#_Toc106827633)

[Figure 5: SAS Code for using PROC FREQ with single variable 10](#_Toc106827634)

[Figure 6: Number of missing Frequency 10](#_Toc106827635)

[Figure 7: PROC FREQ for LOAN\_LOCATION 11](#_Toc106827636)

[Figure 8: Output of PROC FREQ on LOAN\_LOCATION 11](#_Toc106827637)

[Figure 9: Output of LOAN\_LOCATION in terms of City, Town, and Village 12](#_Toc106827638)

[Figure 10: Code for the LOAN\_LOCATION in terms of City, Town, and Village 12](#_Toc106827639)

[Figure 11: Cross-Tab analysis 13](#_Toc106827640)

[Figure 12: Code for generating the frequency. 13](#_Toc106827641)

[Figure 13: Output of percentage and cumulative frequency 14](#_Toc106827642)

[Figure 14: Code for generating values 14](#_Toc106827643)

List of tables

[Table 1: Metadata for the dataset 7](#_Toc106827644)

# CHAPTER 1: Introduction to Feature Engineering

The process of selection, manipulation, and transformation of raw data into features that can be applicable in supervised learning is called feature engineering. This uses statistical approaches and machine learning tools. Feature Engineering improves performance by making the models more accurate and the insights more useful.

There are various features in feature engineering.

1. **Feature Creation**

The creation of new variables by addition or removal of some features is included in this feature creation.

1. **Feature Transformation**

The features are transformed from one form to the other which includes plotting and visualization of data, which are involved in feature transformation.

1. **Feature Extraction**

This involves the extraction of various features from a dataset to depict helpful information.

1. **Exploratory Data Analysis**

To identify new patterns, present in the data, exploratory data analysis is used.

1. **Benchmark**

This involves a comparison of performance among various machine learning models.

# CHAPTER 2: Literature Review / Related work

The research done by (Rawat & Khemchandani, 2019) uses various methodologies, tools, and technologies to improve the accuracy of unseen data. Additionally, with the motive to achieve high performance in terms of accuracy, this research also presents various applications of feature engineering in link prediction, fraud detection, and clinical text classification.

According to the research done by (Heaton, n.d.), various types of features engineered to the types of machine learning models are shown. In this research, several datasets are generated to exhibit the strength of the machine learning model to synthesize features on their own. Various features like Deep Neural network, Support Vector Machine, Random Forest, and Gradient Boosted Machine have been discussed in this research which benefit from a different set of synthesized features and further help us to make recommendations.

The research done by (Khurana et al., 2017) proposed a framework for automating feature engineering. By comparing various exploration policies in terms of the efficiencies, based on the performance-driven exploration of any transformation graph which would enumerate the space of given options systematically and compactly.

It also predicts the transformation which affects the performance of classification positively using a set of datasets to train neural networks for the prediction. A novel framework is known as Learning Feature Engineering (LFE) is proposed in this research paper by learning patterns between class distributions, feature characteristics, and historical data presenting useful transformations. The variable-sized features are transformed into a fixed array protecting the important characteristics. For various kinds of classification problems, the efficiency and efficacy of the framework are demonstrated at very low computational costs.

# CHAPTER 3: Principal Component Analysis

This analysis is used in machine learning for the reduction of dimension, analysis of the exploratory data as well as prediction of the models. One of the most popular applications of this analysis includes the processing of the images, building a system for a recommendation of movies, and optimization of information.

The algorithms in the principal component analysis include the covariance matrix, correlation, and Orthogonal, and the most popular steps include standardizing, calculating covariance, sorting, and removing less important features soon after the dataset has been collected.

# CHAPTER 4: Feature extraction

This technique transforms the dataset into more smaller and meaningful features by reducing the dimensions of the data. The most important application of feature extraction includes the recognition of patterns, processing of images, and unsupervised learning of machine learning. Furthermore, it improves speed and accuracy and improves the visualization of data. (Malik, 2021)

# CHAPTER 6: Label encoding

Label encoding involves the conversion of labels into a readable form of a number, which the machine can understand.

One limitation of label encoding is that while converting, a unique number is also created, which may create issues in the training session of the datasets. Label encoding is usually carried out when the categorical variable od ordinal in nature.

# CHAPTER 7: One-hot encoding

Categorical variables can be better solved with the one-hot encoding, which creates dummy-like variables before the actual encoding is performed for prediction purposes. The one-hot encoding process is usually carried out when the categorical variable is not ordinal.

(Jackson & Agrawal, n.d.)

# CHAPTER 8: Metadata - Exploration of the dataset given

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| No | Name of variable | Description | Data Type | Length | Sample Data (Instances) |
| 1 | SME\_LOAN\_ID\_NO | Unique id for Loan | Char | 8 | LP001024 |
| 2 | GENDER | Gender of the customer | Char | 6 | Female |
| 3 | MARITAL\_STATUS | Marital Status of a customer | Char | 11 | Married |
| 4 | FAMILY\_MEMBERS | Count of family members | Char | 2 | 1 |
| 5 | QUALIFICATION | Qualification of the customer | Char | 14 | Graduate |
| 6 | EMPLOYMENT | Employment status of the customer | Char | 3 | Yes |
| 7 | CANDIDATE\_INCOME | Income of the candidate | Number | 8 | 5849 |
| 8 | GUARANTEE\_INCOME | Amount guaranteed | Number | 8 | 2358 |
| 9 | LOAN\_AMOUNT | Amount took as loan | Number | 8 | 128 |
| 10 | LOAN\_DURATION | Duration of loan taken | Number | 8 | 360 |
| 11 | LOAN\_HISTORY | Number of loans taken in the past | Number | 8 | 1 |
| 12 | LOAN\_LOCATION | The location where the loan was taken | Number | 7 | City |
| 13 | LOAN\_APPROVAL\_STATUS | Status whether the loan is approved or not | Char | 1 | Y |

Table : Metadata for the dataset

The following code was used to check the contents in the given sample file in SAS.

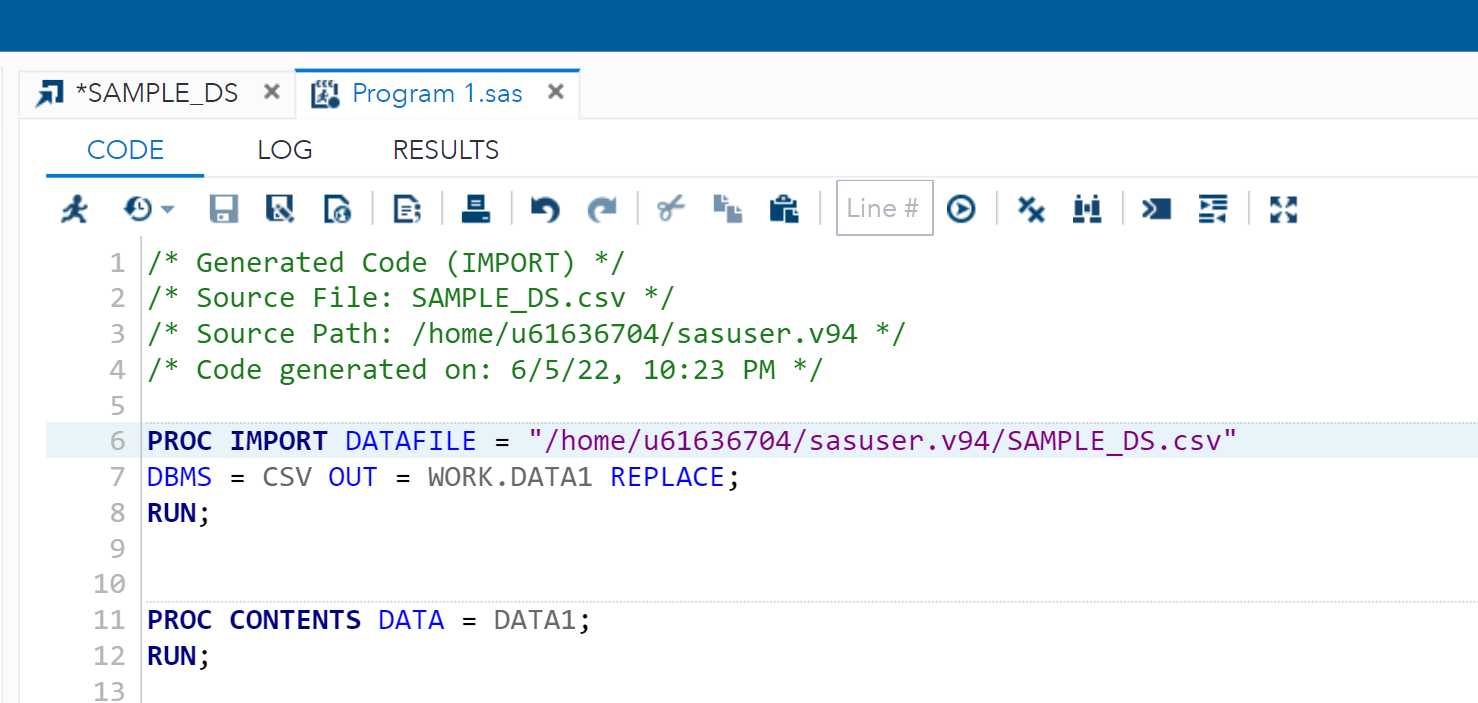


Figure : Code for importing and checking the content

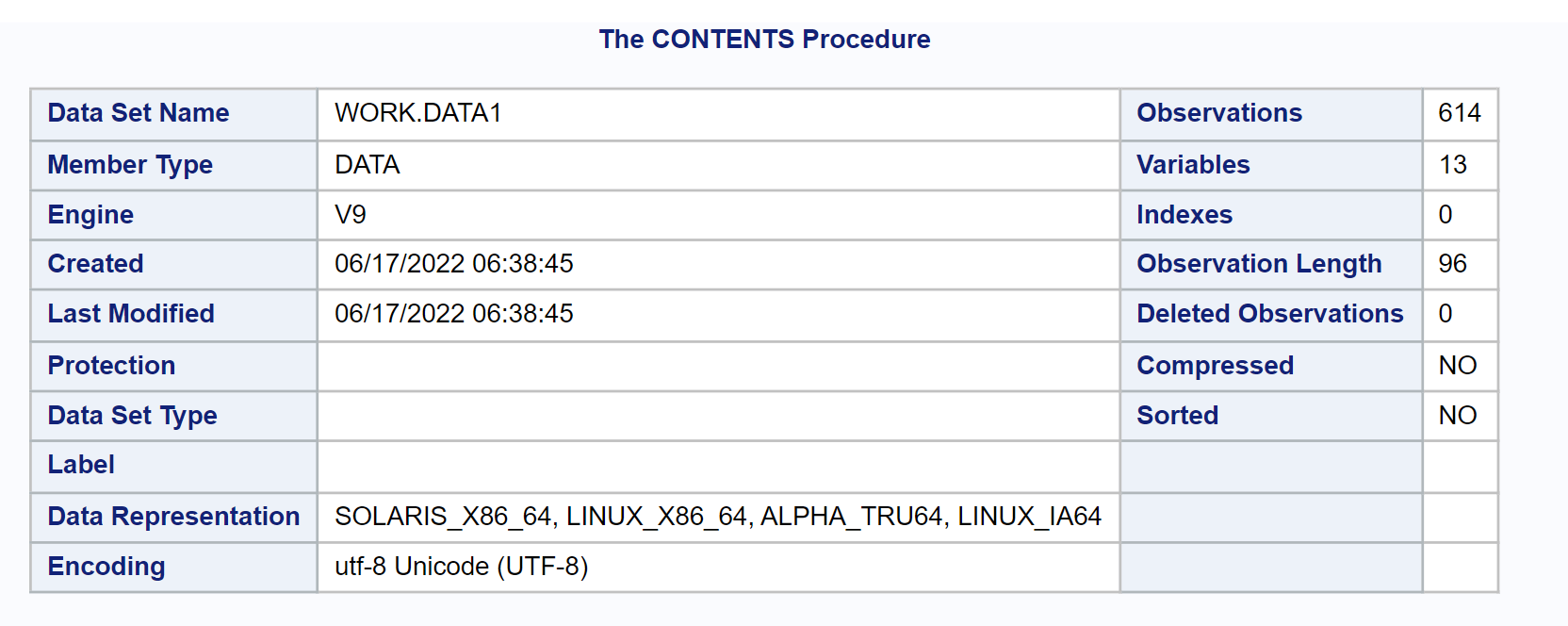


Figure : The Contents procedure

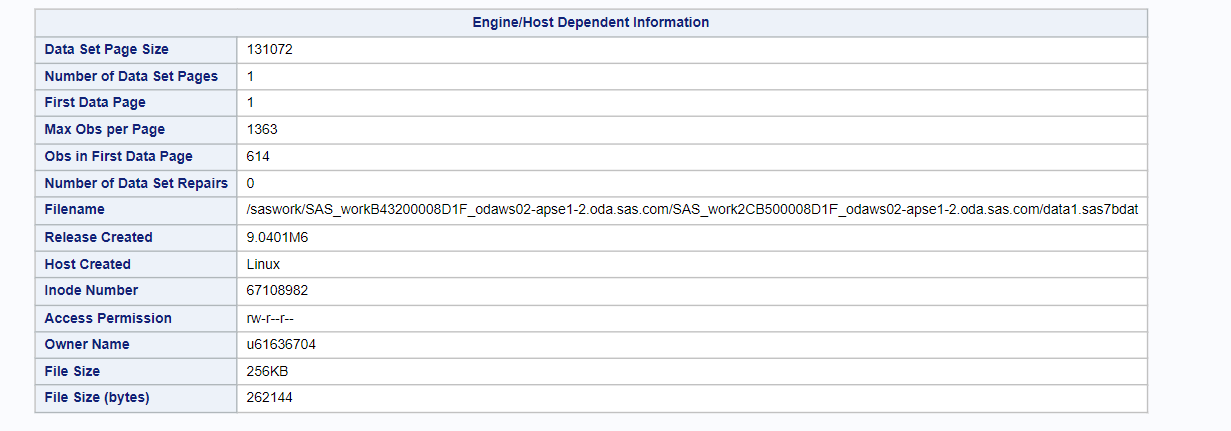


Figure : Engine/Host dependent information

PROC CONTENT shows very important aspects of the data by checking the contents in the file.

First of all, the number of observations is calculated. Here, in the given dataset, the number of observations is 614.

Also, it shows the number of variables in the dataset. In the given dataset, there are altogether 13 variables.

The data is also presented in the sorting order.

The type of permission in the file is shown. Here, we-r-r.

Finally, the list of variables in alphabetical order is listed down along with the data type of those variables, the length of the variable in the given dataset, and also checks the level if there are any.

The figure below shows the alphabetic list of variables and attributes present in the data using the above code.

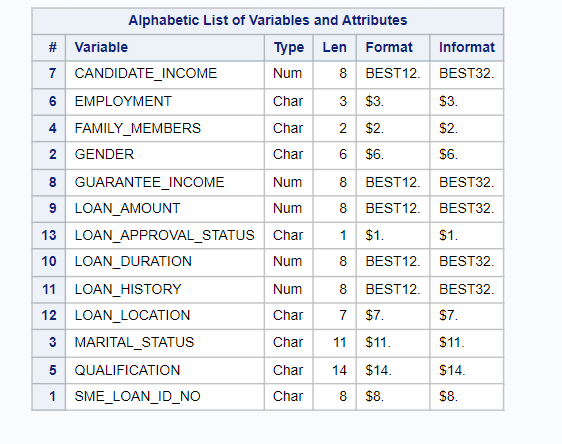


Figure : Alphabetic List of Variables and Attributes

**Single Variable Analysis:**

An important way to identify the frequency in the data is PROC FREQ which is used to analyze single or multiple variables. The snapshot below shows the code for identifying frequency in the Gender column of the sample data.

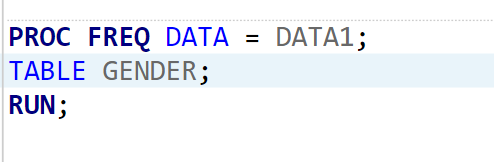


Figure : SAS Code for using PROC FREQ with a single variable

Any other columns can also be used to determine the frequency of occurrence in the data. The figure below shows the output derived from the above code.

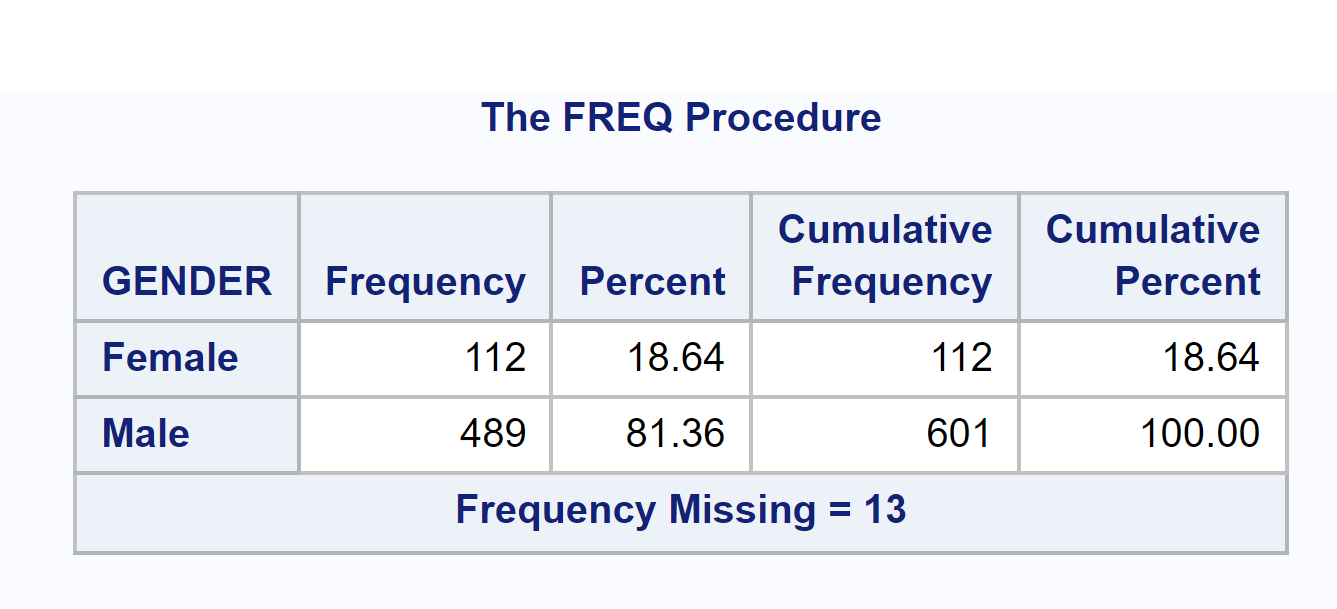


Figure : Number of missing Frequency

The above table shows that there is altogether 112 Female population in the data set which contributes to only 18.64% of the overall population while the male population is 489 which contributes to 81.36% of the total population.

Similarly, the function was applied to another variable in the data – Loan Location. The code used to derive this is shown in the figure below.

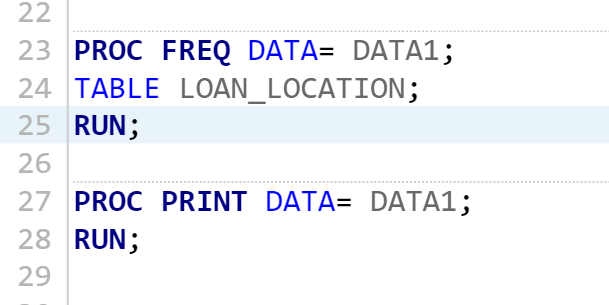


Figure : PROC FREQ for LOAN\_LOCATION

The output derived from the above code is shown in the figure below.

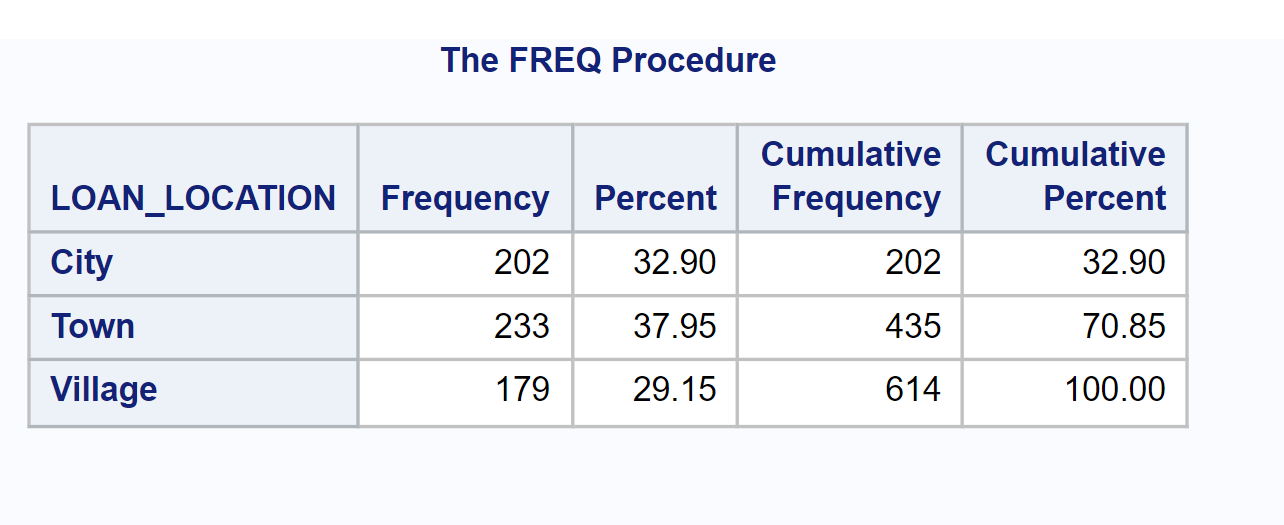


Figure : Output of PROC FREQ on LOAN\_LOCATION

From the above figure, we can determine the number of people living in the city, town, and village individually and their coverage in percentage and cumulative frequency. The same data can be shown without using the cumulative frequency as shown in the figure below.

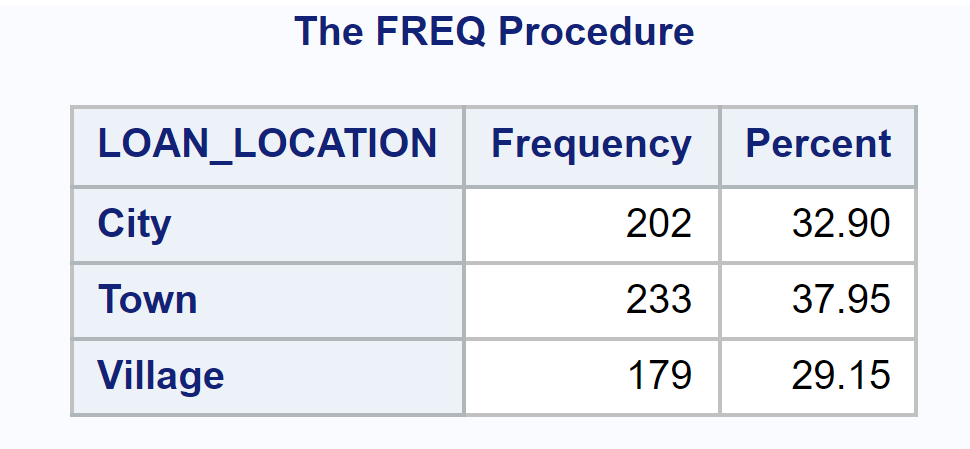


Figure : Output of LOAN\_LOCATION in terms of City, Town, and Village

The code used to generate the above data is shown in the figure below.

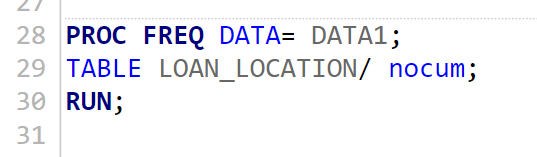


Figure : Code for the LOAN\_LOCATION in terms of City, Town, and Village

**Cross Tab Analysis:**

We can also do cross tabs i.e., analysis across different variables as shown in the figure below.

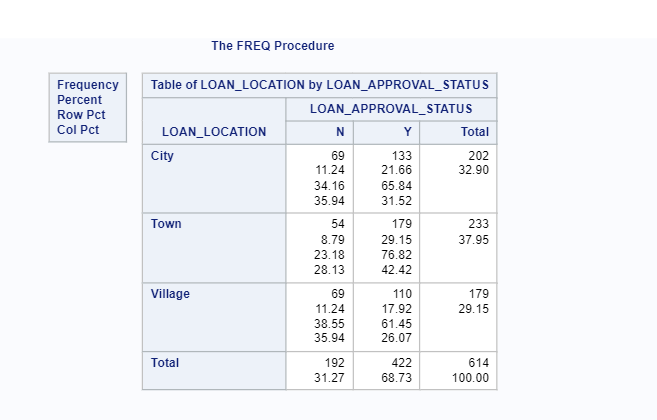


Figure : Cross-Tab analysis

In the above figure, two variables “Loan\_Location” and “Loan\_Approval\_Status” have been taken. The figure shows the number of people with Loan Approval status as either ‘Y’ or ‘N’ among people living in the City, Town, and Village.

The figure below shows the code for generating the above frequency.

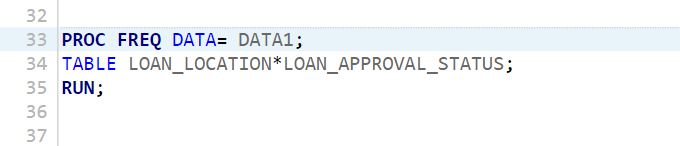


Figure : Code for generating the frequency.

Similarly, the same result without percentage and cumulative frequency can be shown below.

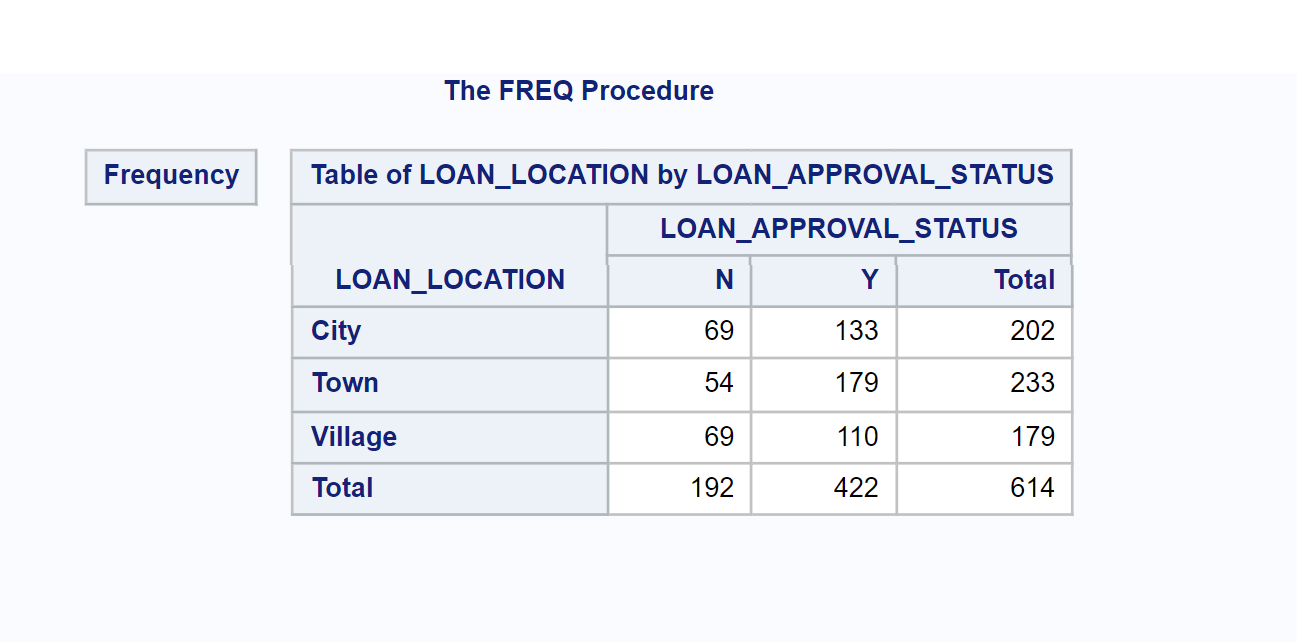


Figure : Output of percentage and cumulative frequency

From the above table, we can see that the number of people whose loan was approved in the city was 133 while 69 people in the city had their loan rejected. Similarly, the number of people in the village whose loans were approved was 110 while 69 people had their loans rejected.

The above figure shows cross tabulation which represents how the numbers and values are spreading around.

This is very helpful in the case of the classification of data either in the form of numbers or percentages.

The code used to generate the above values is shown below.

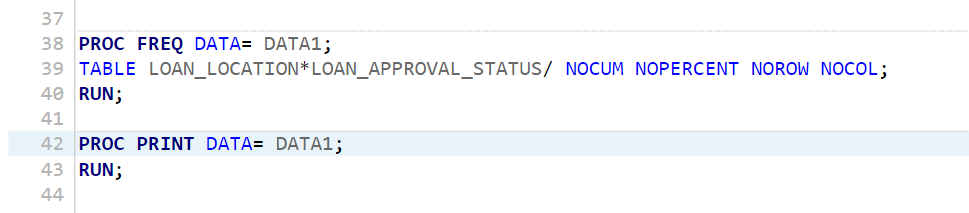


Figure : Code for generating values

The cross-tab analysis of output variables can also be done using more than two variables.

# CHAPTER 9: Conclusion

To conclude, future engineering is an indispensable part of machine learning, and not only helps in organizing and describing the dataset but also helps in predictions and future decision-making processes. Data today is undoubtedly the most valuable asset for any organization, and the data must be properly organized, and used with a full opportunity cost. However, all the data we get or deal with may not be of importance for an organization or decision-making. Therefore, to select, clean and process the best data for the future, the engineering has features such as creation, feature transformation, feature extraction, Exploratory data analysis, and the benchmark. Furthermore, level encoding and hot-end coding features make the machine more readable and effective for future predictions. SAS is one strong tool that helps in machine learning, and we shall use it more to study, analyze, clean, and make the best use of the data. The ‘PROC CONTENT’ and ‘PROC FREQ’ are some important codes in the SAS which perform the variable analysis and cross-tab analysis efficiently.

# CHAPTER 10: References

Heaton, J. (n.d.). *An Empirical Analysis of Feature Engineering for Predictive Modeling*. https://arxiv.org/pdf/1701.07852.pdf

Jackson, E., & Agrawal, R. (n.d.). Performance Evaluation of Different Feature Encoding Schemes on Cybersecurity Logs. *IEEE*. https://doi.org/10.1109/SoutheastCon42311.2019.9020560

Khurana, U., Turaga, D., & Samulowitz, H. (2017). Feature Engineering for Predictive Modeling using Reinforcement Learning. *AAAI*.

Malik, H. (2021). *Advances in Machine Learning and Data Analytics*.

Rawat, T., & Khemchandani, V. (2019). *Feature Engineering (FE) Tools and Techniques for Better Classification Performance*. http://dx.doi.org/10.21172/ijiet.82.024