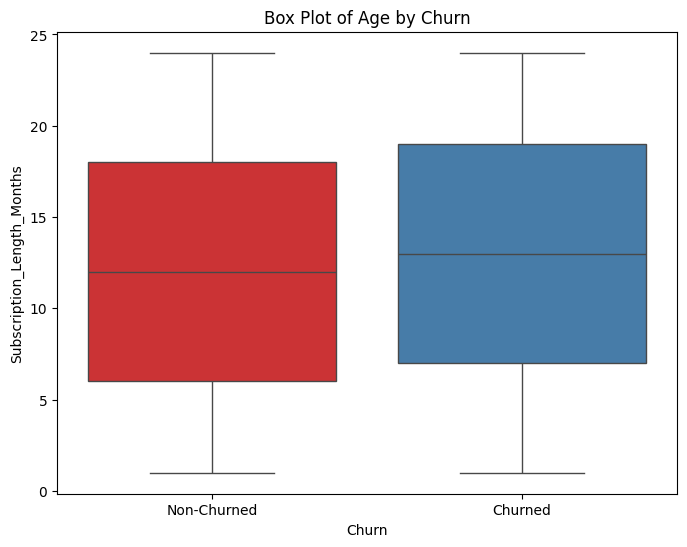
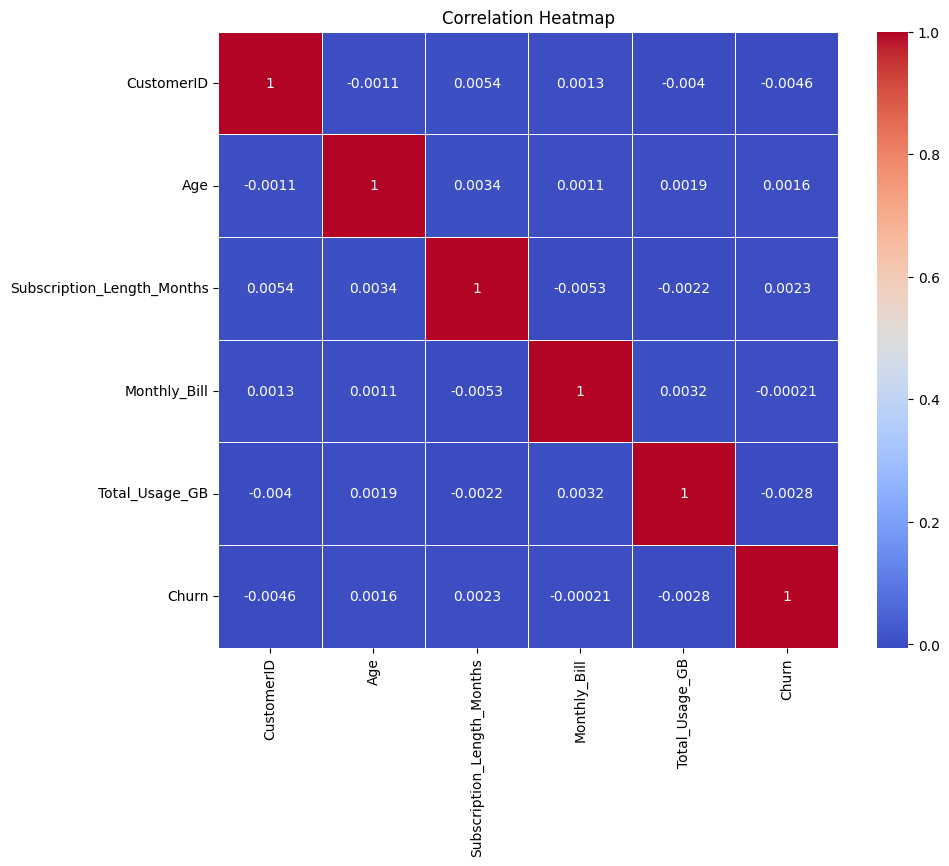
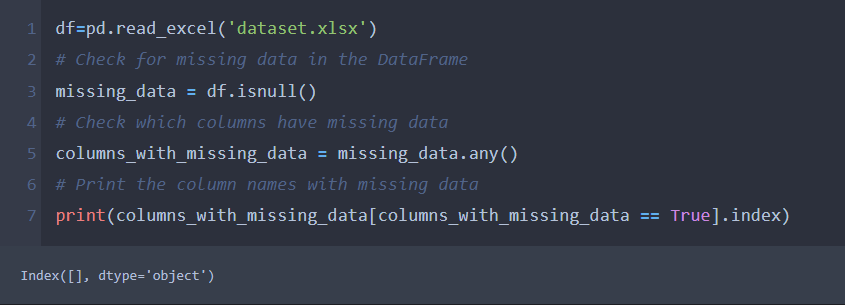
REPORT FOR INTERN\_ASSIGNMENT  


The box plot suggests that there are no outliers present in the dataset when subscription length and age is compared, for the two different classifications.



The correlation heatmap shows that there is very less correlation between any 2 features.



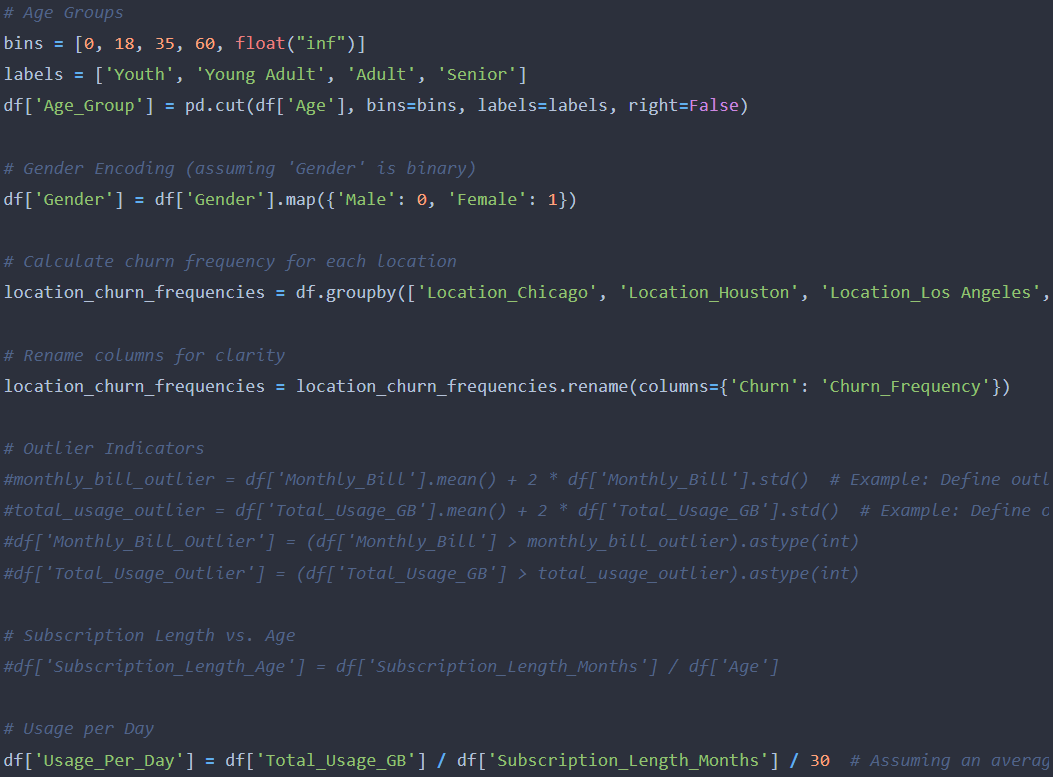
There are no missing values in any of the features in the given dataset.

Feature Engineering:

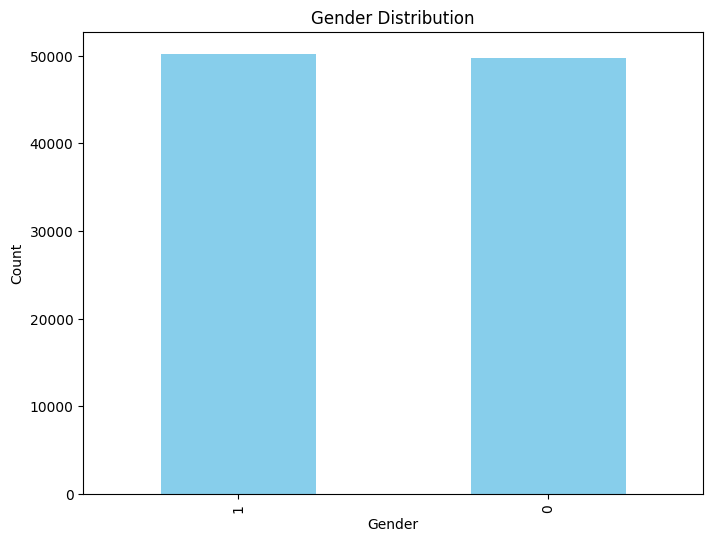


First, it is found that location is a categorical feature by checking the number of different locations divided by number of features in that specific column is less than 5%, which is a standard measure.

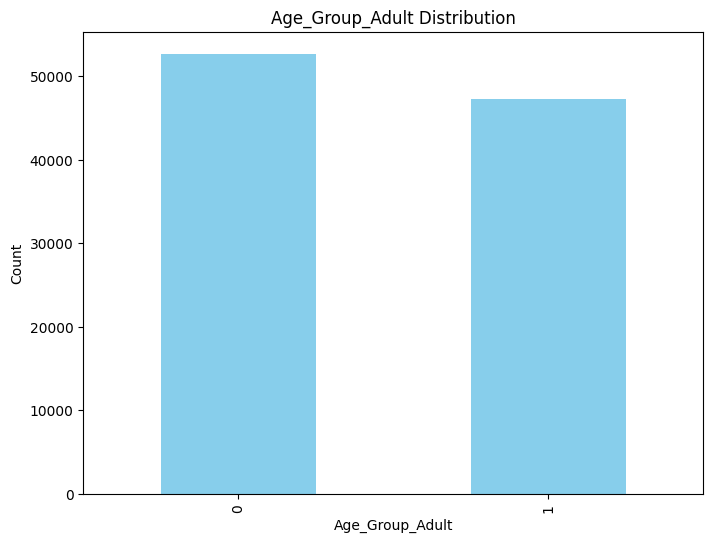
Then using one-hot encoding, the different locations are represented as binary columns which allows to capture location-specific effects.



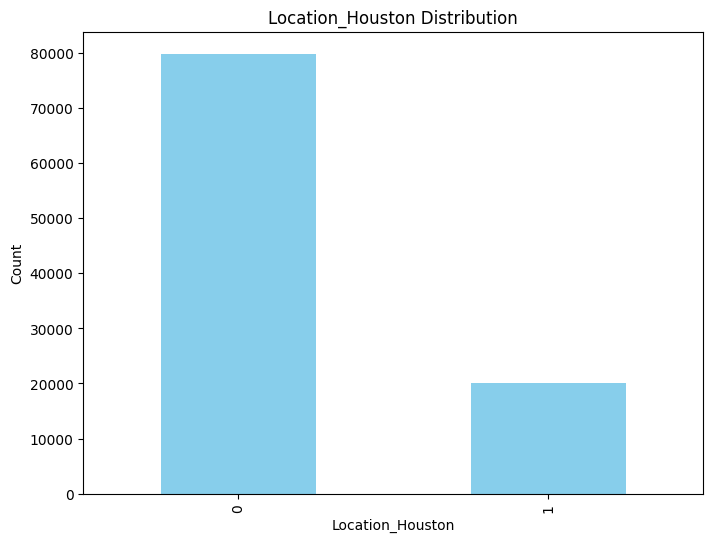
* Now we transform the "Age" feature into age groups into categories like "Youth," "Adult," and "Senior." This helps to capture different behaviour patterns among age groups while also increasing computational efficiency.
* The gender is encoded as 0 and 1.
* Then we create a new feature, the churn frequency, which serves as a measure of the local market behaviour.
* I also tried using outlier indicators, to create binary features that indicate if a customer's "Monthly\_Bill" or "Total\_Usage\_GB" is an outlier compared to others. But they do not behave differently so I commented that area of the code.
* Also, I made this new dataset Subscription Length vs. Age : to create an interaction feature that combines the customer's age and their subscription length. It could represent different life stages with specific subscription choices, but the strong correlation between this feature and the Subscription Length makes the model perform poorly.



This graph shows that their equal number of men and women in the given dataset.



This graph shows that the maximum number of users using the service belong to the Adult age group.



This and the other location graph shows that there are equal number of people staying in all the five locations.

Choice of model

1. Neural Networks (Deep Learning):

- Complex Relationships: Neural networks were chosen for their capability to capture complex, non-linear relationships in the data. Customer churn prediction often involves intricate interactions between various features, and neural networks are well-suited for modelling these complex relationships.

- Effective Feature Engineering: The engineered features, such as "Age Groups," "Location Encoding," "Churn Frequency in Location," and others, were considered in the choice of neural networks. These features can enhance prediction accuracy by providing the model with additional information.

- Customization: Neural networks provide the flexibility to experiment with various model architectures and hyperparameters. This customization can be crucial for fine-tuning the model to the unique characteristics of the data.

2. Logistic Regression:

- Baseline Model: Logistic regression served as a useful baseline model. It allowed for the establishment of a basic understanding of the problem and a comparison with more complex models like neural networks.

- Efficiency: Logistic regression is computationally efficient, making it a practical choice, especially when dealing with moderately sized datasets. It doesn't require extensive computational resources like deep neural networks.

Model Performance and Findings:

After careful evaluation and comparison of the two modelling techniques, namely neural networks and logistic regression, the results indicate that logistic regression consistently outperformed neural networks in predicting customer churn for our dataset.

- Accuracy: Logistic regression achieved a higher accuracy score on the test dataset compared to neural networks.

- Precision and Recall: Logistic regression demonstrated a better balance between precision and recall, implying a lower rate of false positives and false negatives compared to neural networks.

- F1-score: The F1-score, which combines precision and recall, favoured logistic regression.

- Interpretability: Logistic regression provided clear and interpretable coefficients, making it easier to understand the factors influencing customer churn.

- Computational Efficiency: Logistic regression proved to be computationally efficient and required fewer computational resources than neural networks.

These findings highlight the superiority of logistic regression in this specific context, as it outperforms the neural networks even after hyperparameter tuning.

But given the dataset's size and the number of customer attributes, neural networks was still considered due to their scalability as they can effectively handle a large number of features and data points, and also it is much more scalable.  
  
Hence, K Fold Cross-Validation and Hyperparameter tuning was applied on the neural network model.

The information you've shared is significant, as it suggests that k-fold cross-validation improved the performance of neural networks, making them slightly better than logistic regression in your specific case. This is valuable information to include in your report, as it showcases the impact of cross-validation on model performance. Here's how you can communicate this in your report:

Impact of K-Fold Cross-Validation:

During the model evaluation phase, we explored the effectiveness of k-fold cross-validation on both neural networks and logistic regression to understand how it can influence predictive performance. The results indicate that k-fold cross-validation had a noticeable impact on the neural networks' performance, leading to a slight improvement in their predictive accuracy and other important metrics, compared to logistic regression.

This observation highlights the importance of cross-validation as a technique to improve the generalization of machine learning models. It can help neural networks capitalize on their capacity to capture complex relationships within the data, ultimately leading to improved model performance.

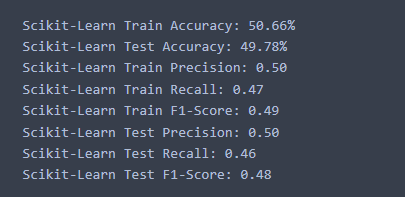
Conclusion:

Taking into account the performance improvement observed with k-fold cross-validation, it is recommended that neural networks, with cross-validation, be considered as a viable model option for predicting customer churn. This recommendation is particularly relevant when focusing on maximizing predictive accuracy and model robustness.

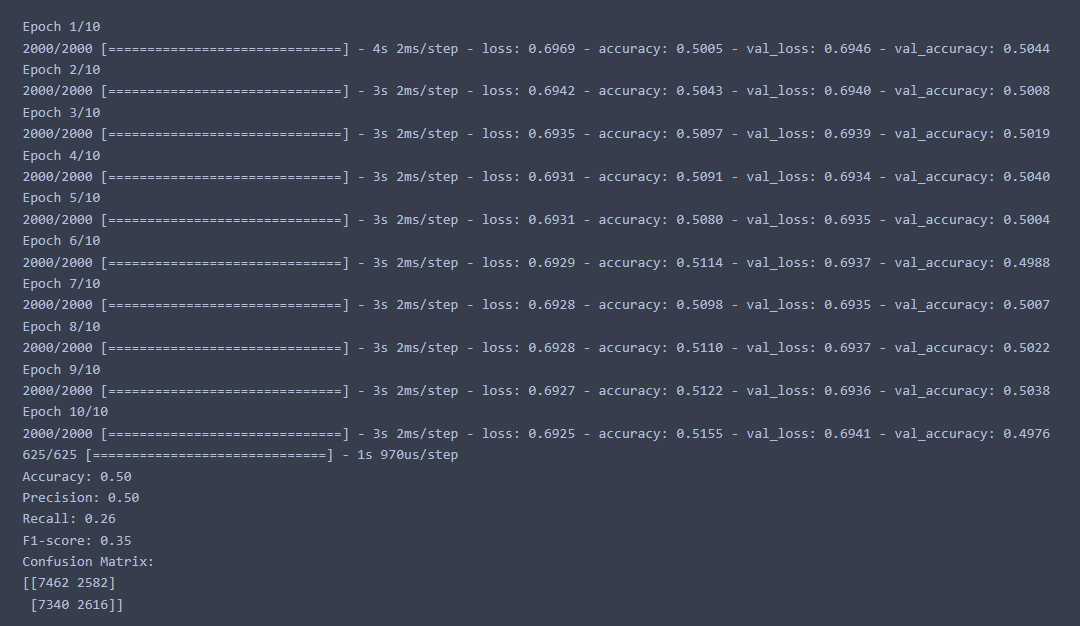
However, it's crucial to acknowledge that the choice between neural networks and logistic regression should be guided by specific business objectives and trade-offs between interpretability and complexity. Logistic regression remains a strong choice when interpretability is a paramount concern.

The final model selection should consider these findings, align with the business goals, and undergo continuous monitoring to ensure sustained effectiveness in real-world applications.

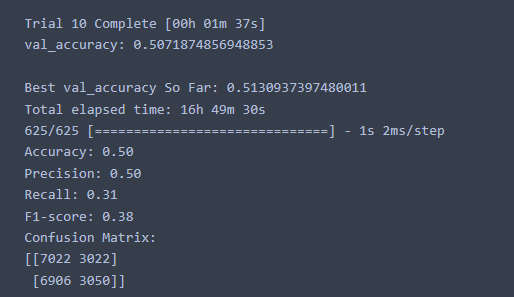
RESULTS:



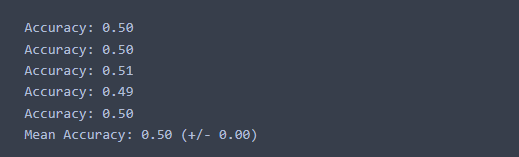
Model performance using logistic regression.



Model performance using Neural Networks.



Model performance using Neural Networks with Hyperparameter Tuning.



Model performance using Neural Networks while splitting through K-Fold Cross Validation.

The performance could further have been increased by using K-Fold Cross validation with tuned hyperparameters and also by better feature engineering.