DATA ANALYSIS REPORT

Prediction of Medical Expenses using Machine Learning!

**Prepared By**

02.05.2023

……………………

# 

Contents

[EXECUTIVE SUMMARY 3](#_Toc134174054)

[DECLARATION. 4](#_Toc134174055)

[ACKNOWLEDGEMENT 5](#_Toc134174057)

[LIST OF ABBREVIATIONS 6](#_Toc134174058)

[EXPLORATORY DATA ANALYSIS. 8](#_Toc134174059)

[FEATURE ENGINEERING. 17](#_Toc134174060)

[FINDINGS. 22](#_Toc134174061)

[LIMITATIONS. 25](#_Toc134174062)

[DATA CLEANING AND PREPARATION. 27](#_Toc134174063)

[MACHINE LEARNING MODEL BUILDING. 28](#_Toc134174064)

[MODEL SELECTION. 28](#_Toc134174065)

[MODEL TRAINING & MODEL EVALUATION. 29](#_Toc134174066)

[MODEL DEPLOYMENT. 29](#_Toc134174067)

[RECOMMENDATIONS AND FUTURE WORK. 30](#_Toc134174068)

[RESULTS 30](#_Toc134174069)

[CONCLUSION 31](#_Toc134174070)

[REFERENCES. 31](#_Toc134174071)

# EXECUTIVE SUMMARY

This report aims to forecast the healthcare costs of NHIF insurance company's users who are children aged 3 years to 17 years old by analyzing their demographic information, insurance details, age, and other relevant factors. The objective of this project is to develop a machine learning model that can accurately predict the medical expenses that NHIF users are likely to incur. The proposed solutions will provide NHIF, governments, private equity firms, and other stakeholders with the ability to estimate healthcare costs, monitor the long-term health of individuals and areas to improve for better economic and social profits in the near future. The report provides an overview of the data sources, the methods used to analyze the data, and the key findings and conclusions drawn from the analysis as well as everything behind building a machine learning mode.

# DECLARATION.

# I hereby declare that this report represents my original work, and that all data analysis and modeling was conducted in an honest and transparent manner. Throughout the process, I took great care to ensure the accuracy of the data, and made every effort to address any potential issues that arose during the data cleaning, preparation and machine learning modeling phase. Additionally, I approached the feature engineering and model selection process in a good approach, with the purpose of evaluating the performance of different models and selecting the one that best meet the needs of the problem. I am proud of the work presented in this report up to its conclusions, which I believe to be a valuable contribution to the field of health services, data and machine learning.

Name: …………………………………………………..

Signature: …………………………………………..…..

Date: ……………………………………………………

# ACKNOWLEDGEMENT

I would like to express my gratitude to everyone who has supported me throughout the process of conducting this from data collection, analysis and building the machine learning model. Firstly, I would like to thank [**name**], [**position**] at [**company/organization**], for providing me with access to the necessary data and resources to carry out this work.

I would also like to extend my thanks to [**name**], [**position**] at [**company/organization**], for their insightful feedback and guidance throughout the process from data collection and analysis.

Finally, I would like to thank my family and everyone behind this without even mentioning them for their support to this journey. Their support has been invaluable and I could not have completed this project without them being in touch.

Thank you all for your support and contributions.

# LIST OF ABBREVIATIONS

1. EDA – Exploratory Data Analysis.
2. ML – Machine Learning.
3. SVM – Support Vector Machine.
4. NHIF - Natural Language Generation
5. KDE – Kernel Density Estimate.
6. NHIF – National Health Insurance Fund
7. TSH – Tanzania Shillings

INTRODUCTION

The purpose of this report is to analyze data and develop a machine learning model that can forecast the healthcare costs of NHIF insurance company's users who are children between the age of 3 years to 17 years. This will be achieved by analyzing their demographic information, children details, age, and other relevant factors. The scope of the analysis includes exploring the dataset provided, cleaning and preprocessing the data, analyzing the relationships between different variables, and developing a predictive model that can accurately forecast healthcare costs. The dataset provided includes information on 576 children who received medical services from hospitals, clinics, dispensaries and other types of centers in different regions of Tanzania. The dataset contains information on the patient's gender, the type of hospital or clinic they visited, the ownership of the medical facility, the patient's age, the region where they received medical services, the number of times the patient visited the medical facility in 2021 and 2022, and the amount paid for medical services in 2021 and 2022. To analyze the dataset, we will use various data analysis and machine learning techniques. The data will be cleaned and preprocessed to remove any missing values or outliers. We will then analyze the relationships between different variables to identify any patterns or trends. Finally, we will develop a predictive model using machine learning and deep learning algorithms to forecast the healthcare costs of these children in 2021 and 2022.

# EXPLORATORY DATA ANALYSIS.

Exploratory Data Analysis (EDA) is the process of analyzing and visualizing data to get insights and understand the patterns and relationships in the data. In this process, we use statistical techniques and visualization tools to summarize and explore the main characteristics of the dataset. Here's a brief explanation of the columns used in our dataset.

**S/n:** This column indicates the serial number of the children in the dataset. It is a unique identifier for each child in the dataset:

In our dataset we had 574 children and that was our maximum serial number.

**Gender**: This column indicates the gender of each child in the dataset. The gender was male or female. The dataset has 311 Male and 263 Female children, this indicates that male were many compared to female. Below are the charts showing the number of male against female.

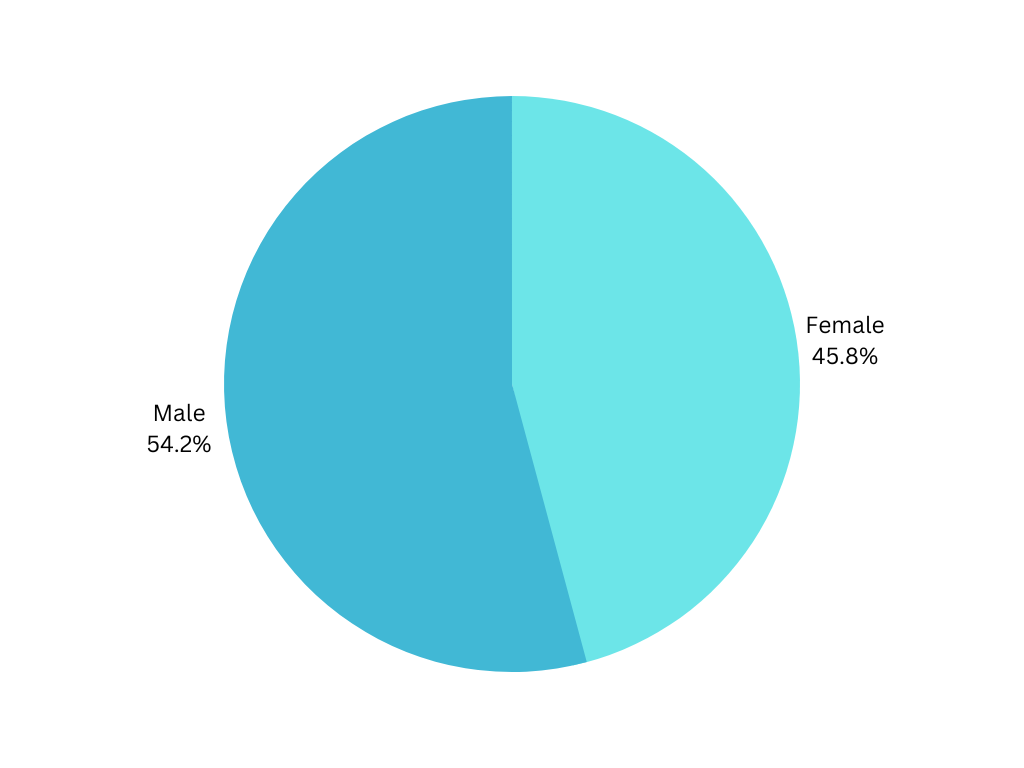


Fig: A chart showing percentage distribution between male and female in the dataset.

**Category**: This column indicates the type of hospital, clinic, or dispensary where the child received medical services. The category consist of 7 categories. Below are the details for each category.

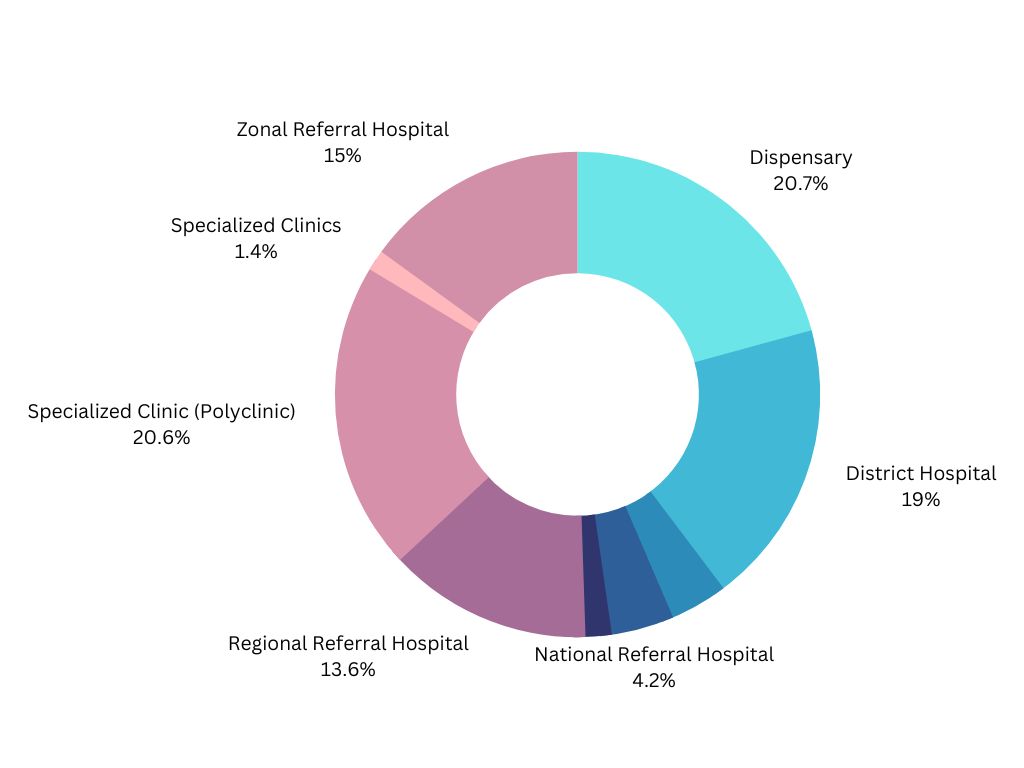


Fig: A chart showing the percentage distribution of category data in our dataset.

**Ownership**: This column indicates the ownership of the hospital, clinic, or dispensary where the child received medical services. The ownership can be public, private, or faith-based.

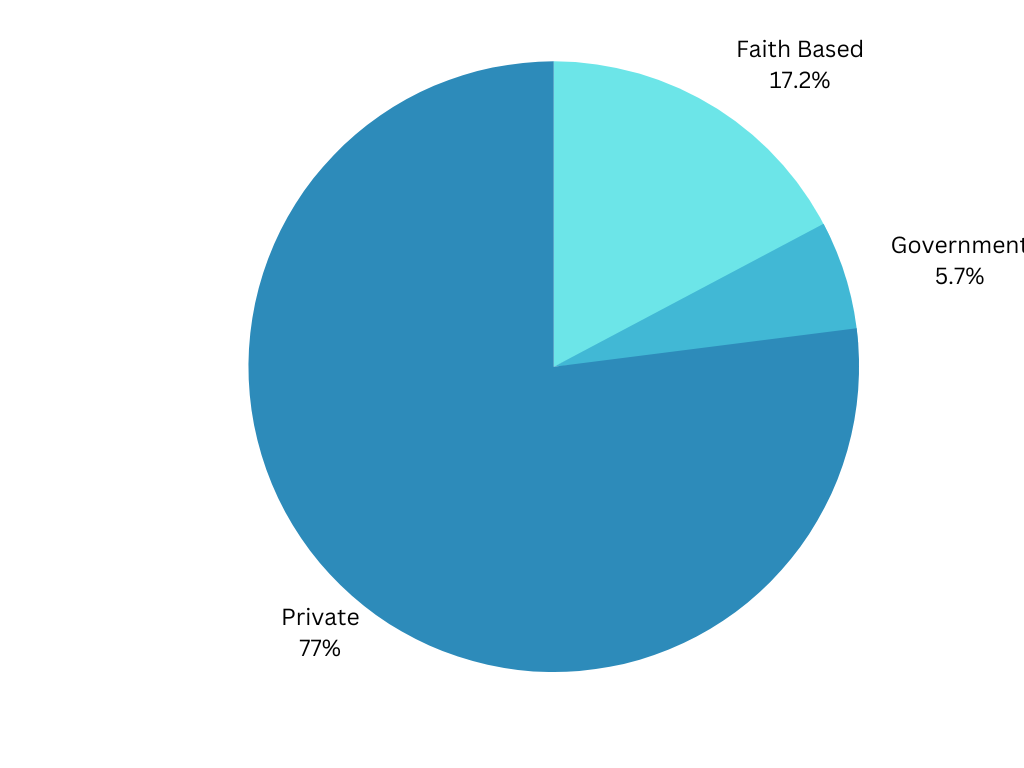


Fig: The chart showing the percentage distribution of Ownership data in our dataset.

**Age**: This column indicates the age of each child in the dataset. Region: This column indicates the region where the child lives or received medical services.

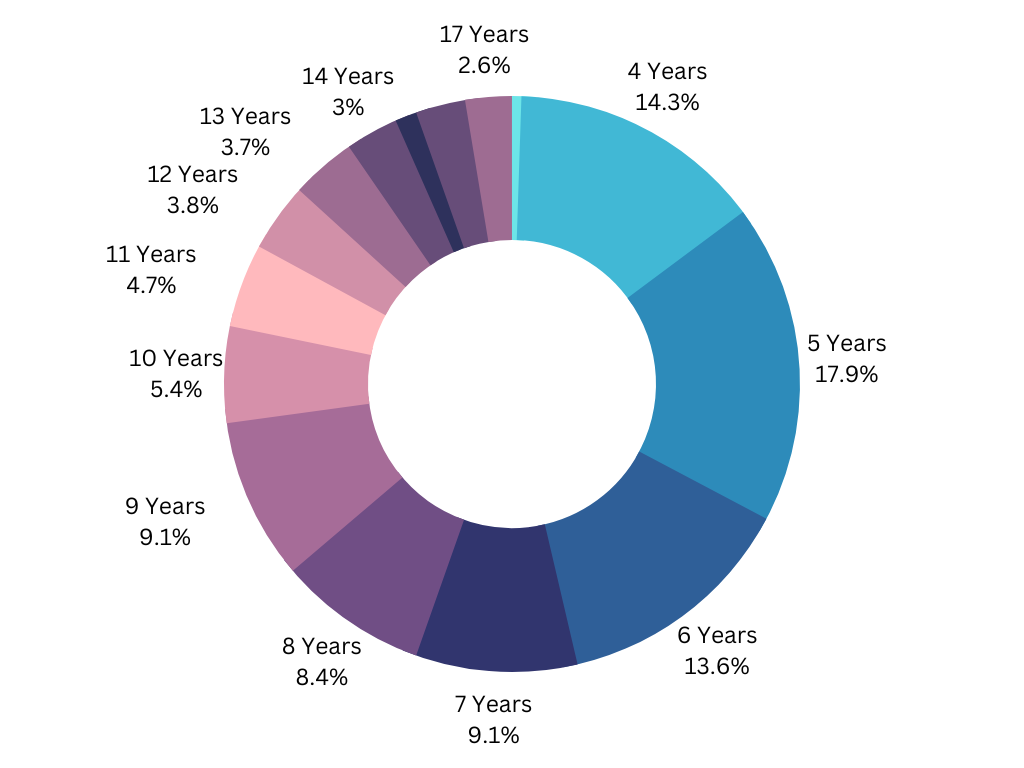


Fig: A chart showing the percentage of age distribution in our dataset.

**Region**: This column indicates the number of regions where our data was collected and where did children have services.

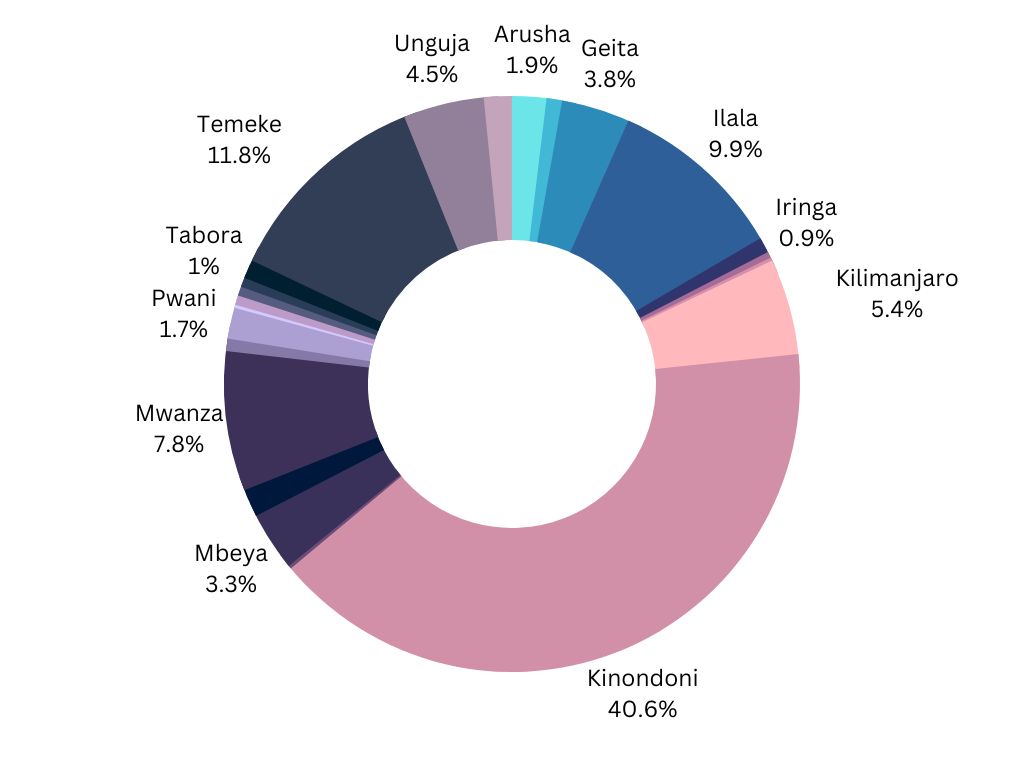


Fig: A chart showing the percentage of regions distribution in the dataset.

**Visits Jul 21:** This column indicates the number of times each child visited a hospital, clinic, or dispensary in July 2021.

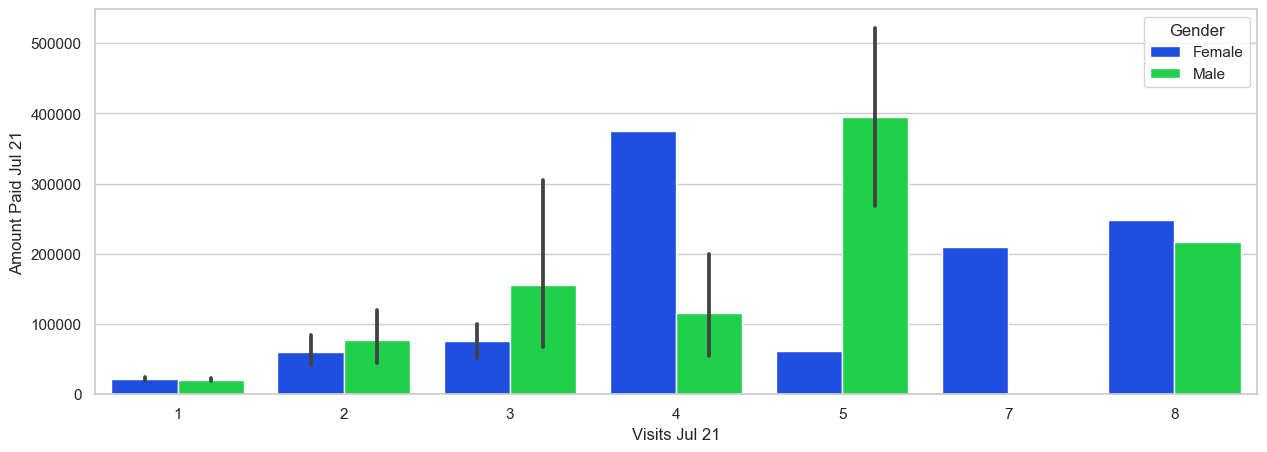
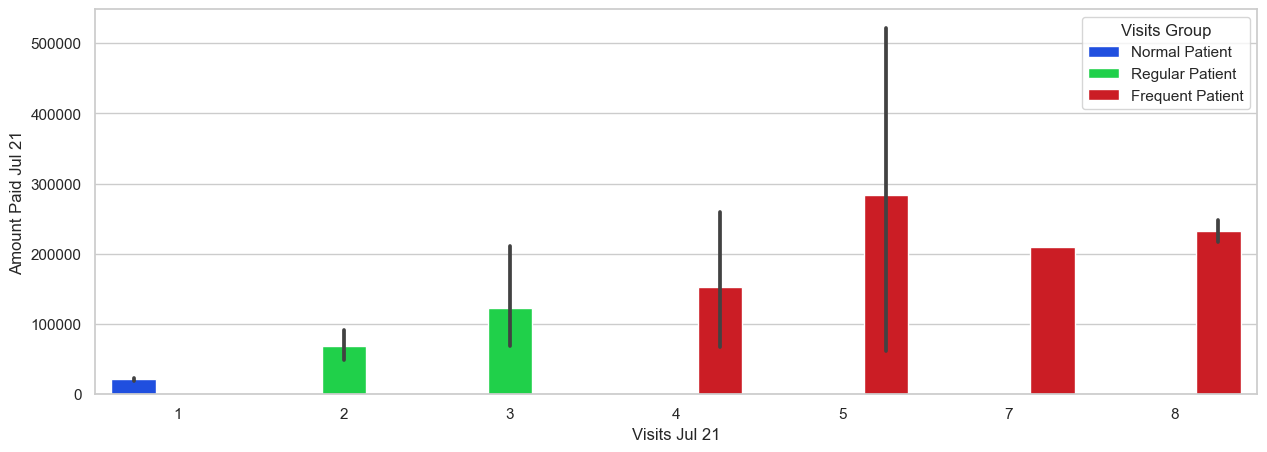


Fig: A graph showing contribution of each gender in the amount paid in 2021 per their visits.

Fig: A graph showing contribution of each visits group in the amount paid in 2021 per their visits.

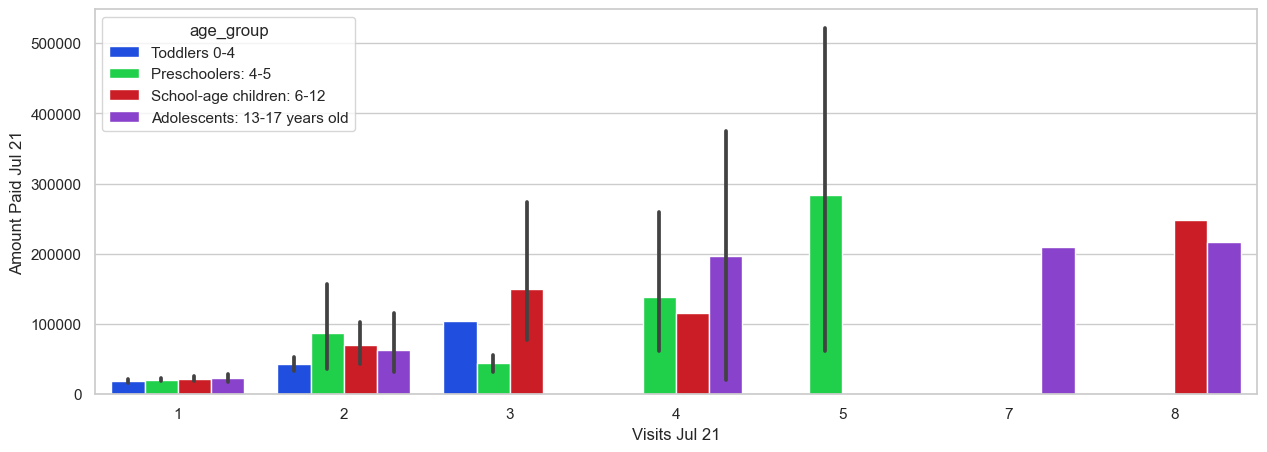
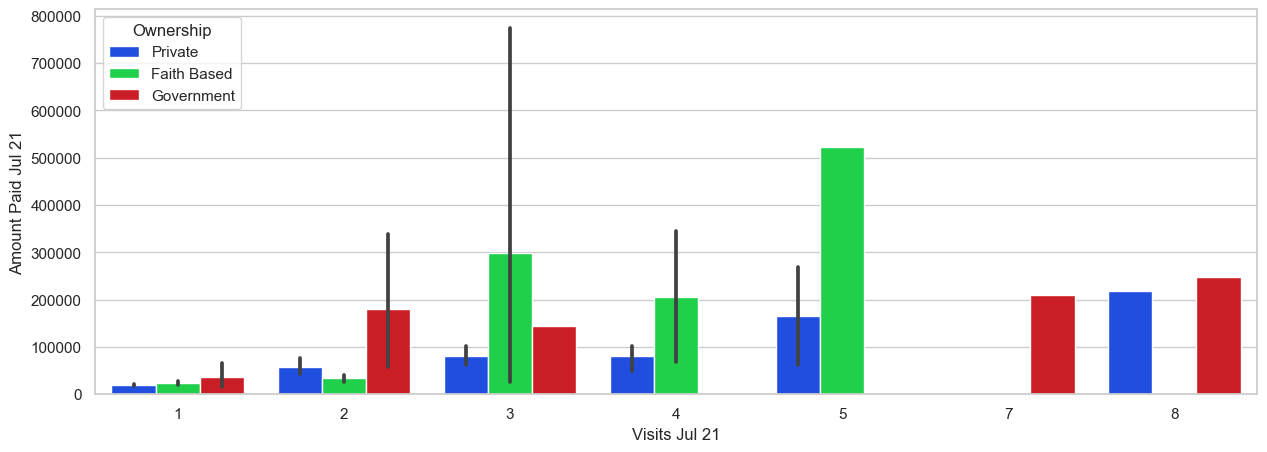
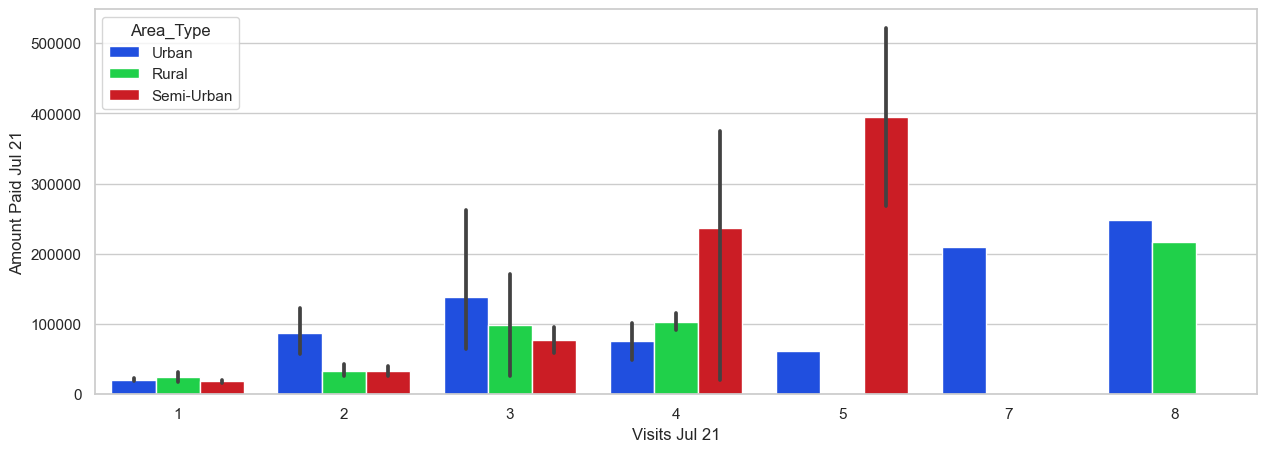


Fig: A graph showing contribution of each age group in the amount paid in 2021 per their visits.

Fig: A graph showing contribution per ownership in the amount paid in 2021 per their visits.

Fig: A graph showing contribution per area\_type in the amount paid in 2021 per their visits.

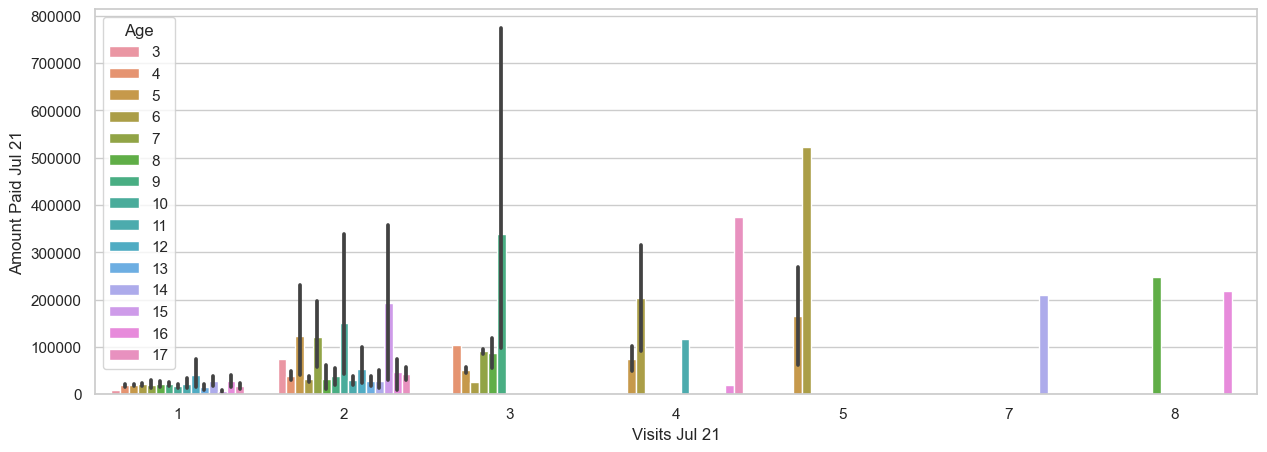


Fig: A graph showing contribution per age in the amount paid in 2021 per their visits.

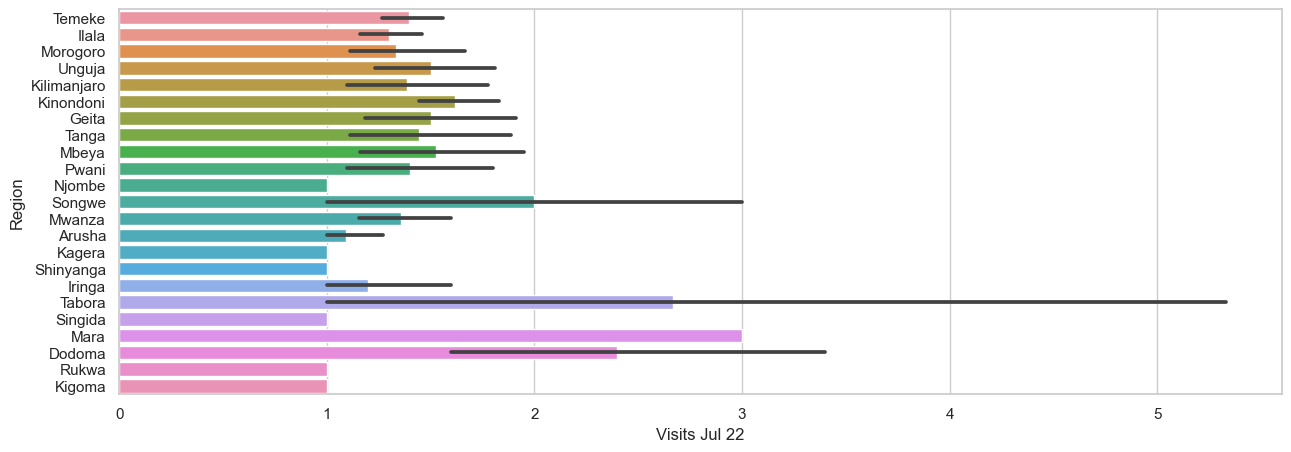
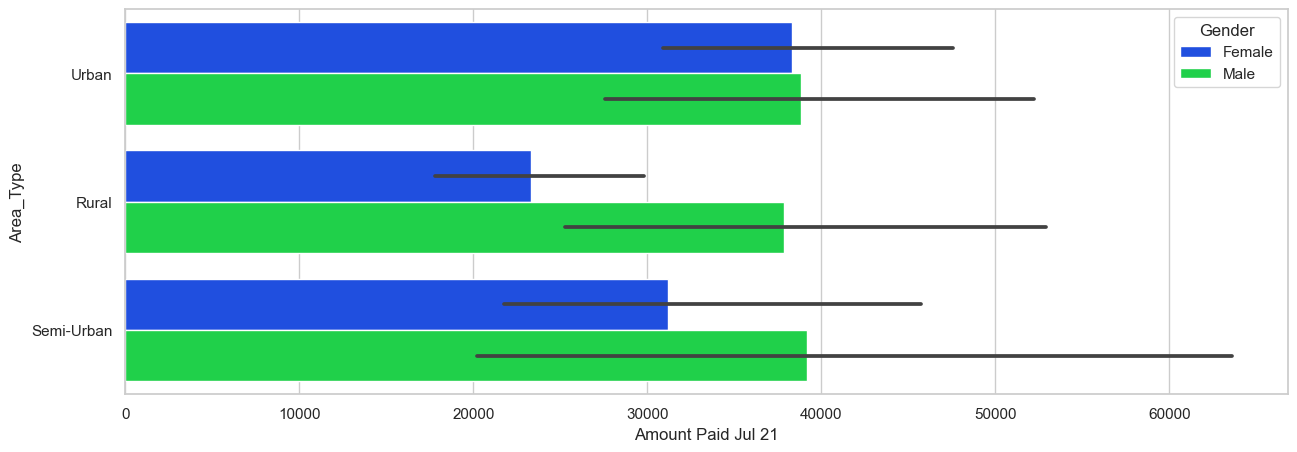


Fig: A graph showing contribution per region in the amount paid in 2021 per their visits.

**Amount Paid Jul 21**: This column indicates the total amount paid by each child for medical services in July 2021.

Fig: A graph showing contribution per area\_type in the amount paid in 2021.

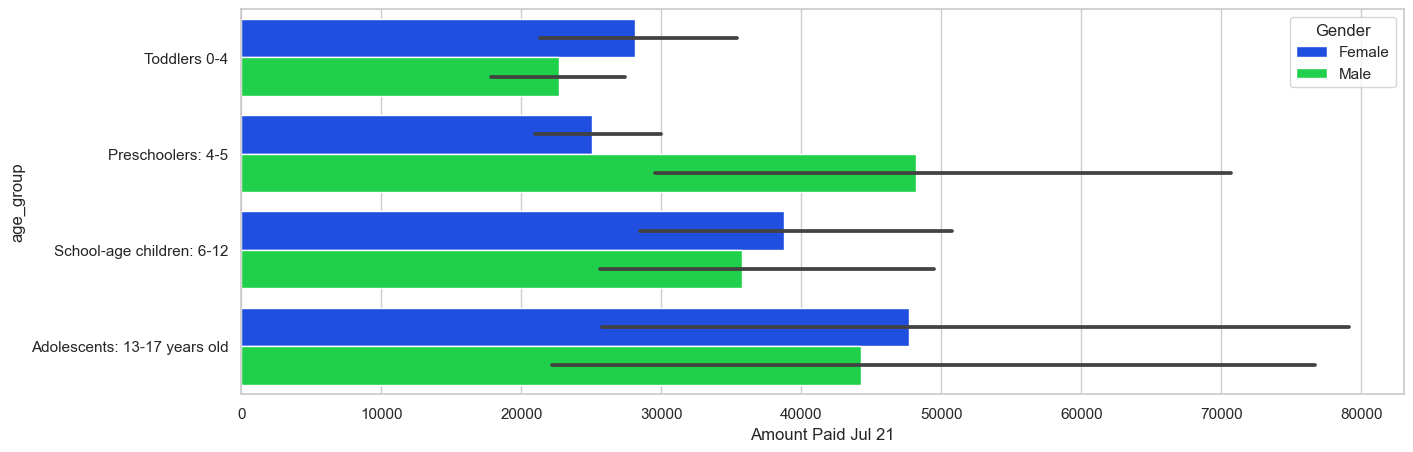


Fig: A graph showing contribution per age\_group in the amount paid in 2021.

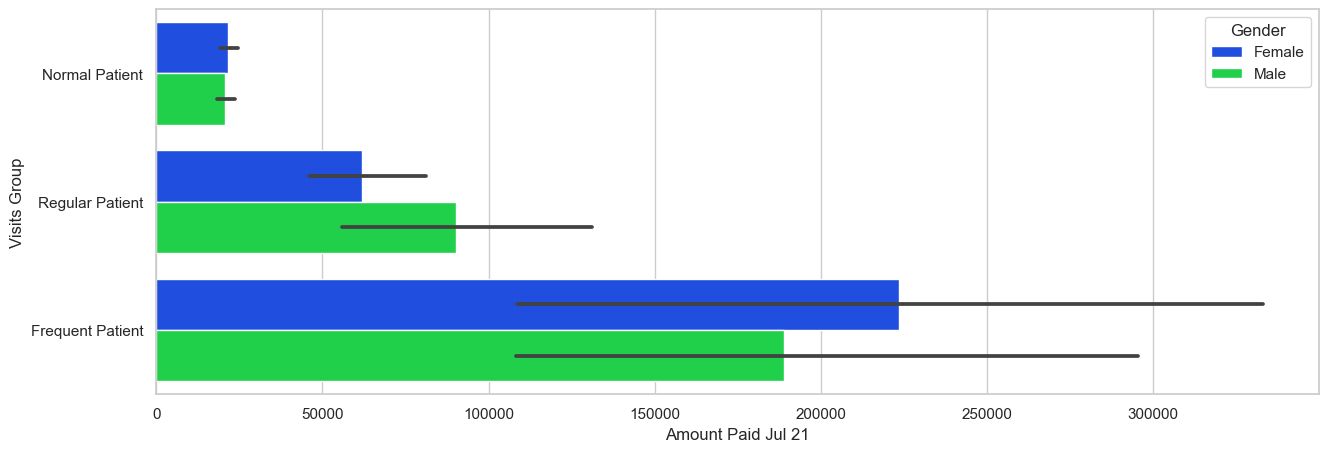
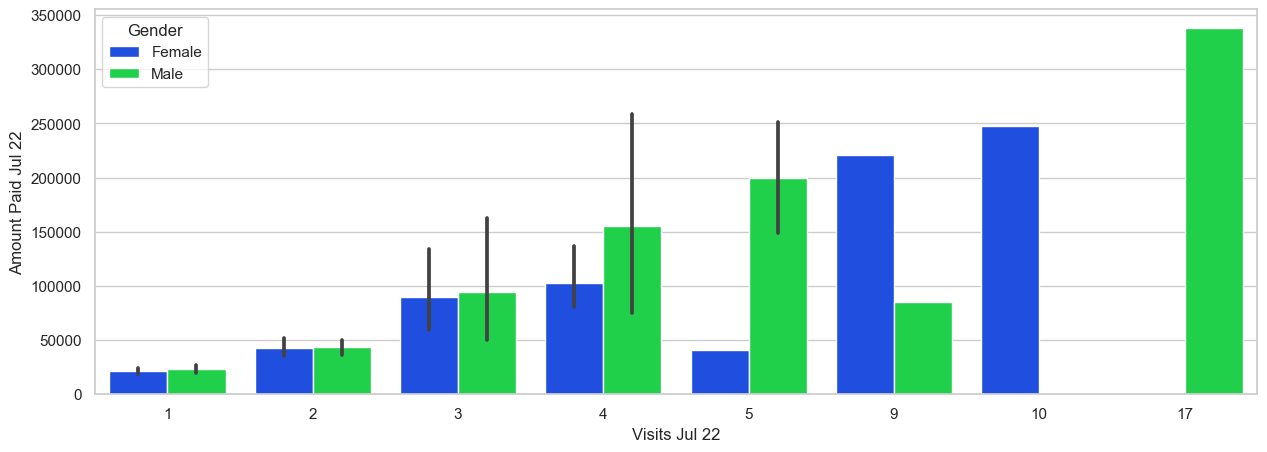
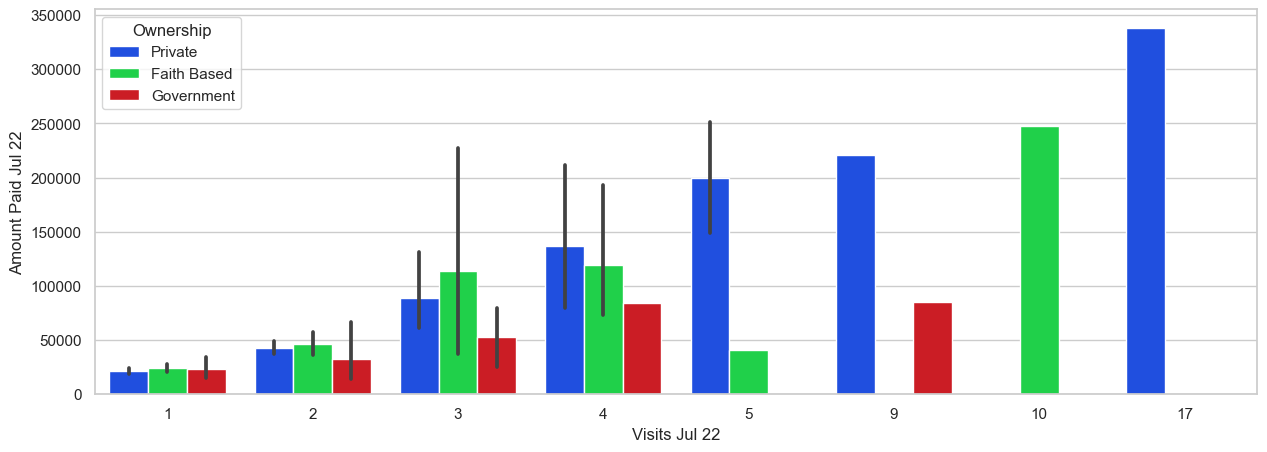
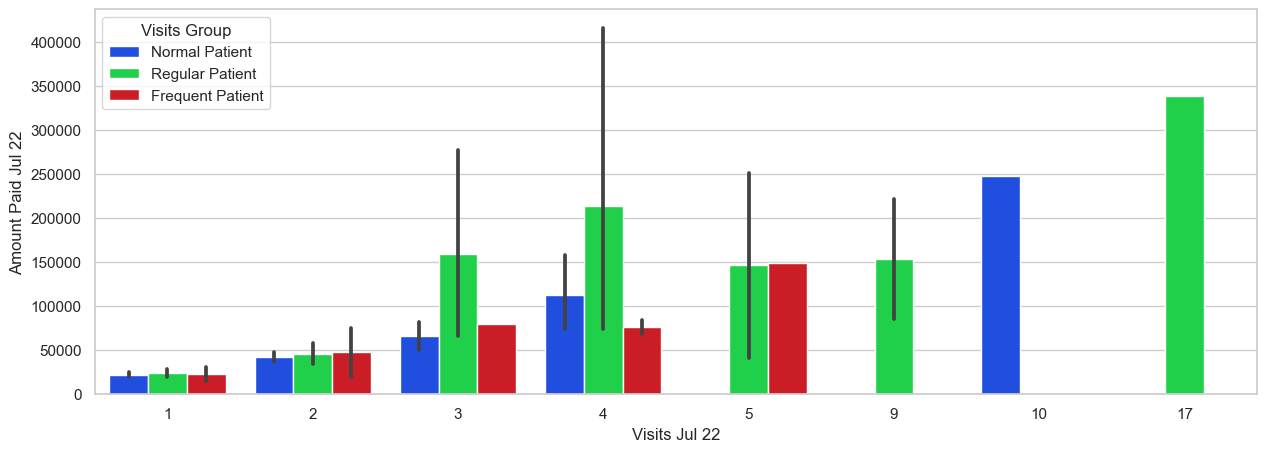


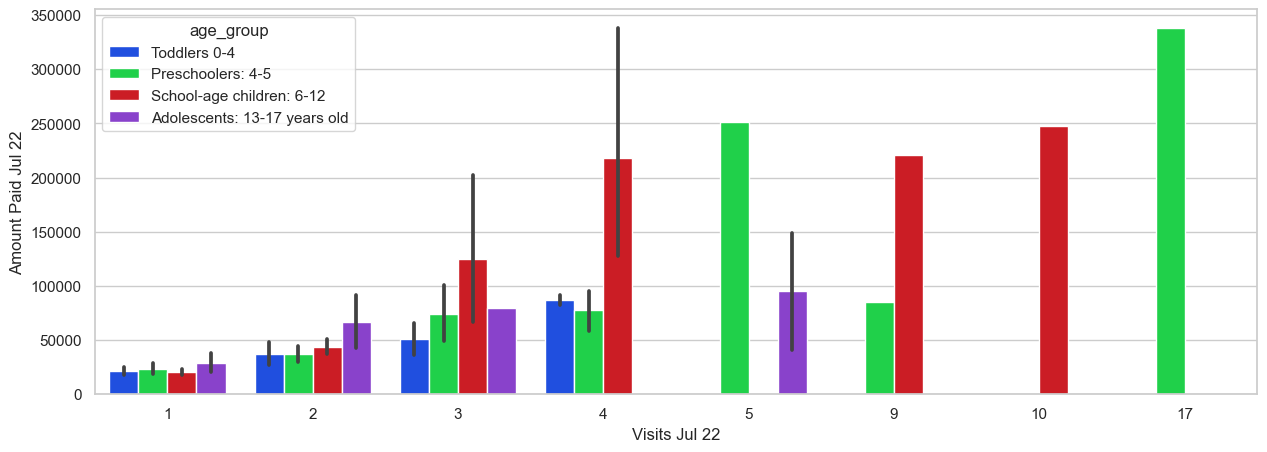
Fig: A graph showing contribution per visits group in the amount paid in 2021.

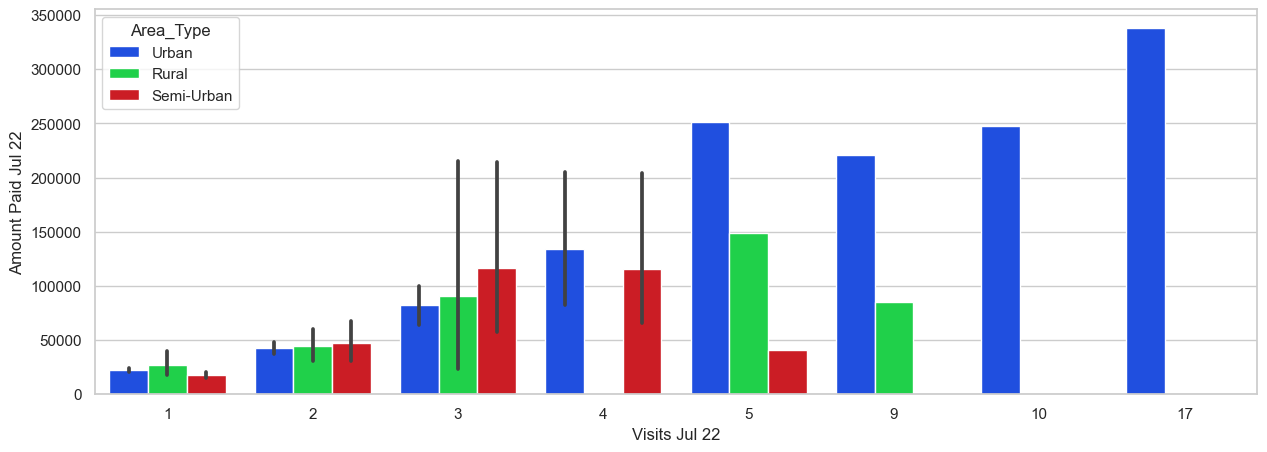
**Visits Jul 22:** This column indicates the number of times each child visited a hospital, clinic, or dispensary in July 2022.

 Fig: A graph showing contribution per gender in the amount paid in 2022 per their visits.

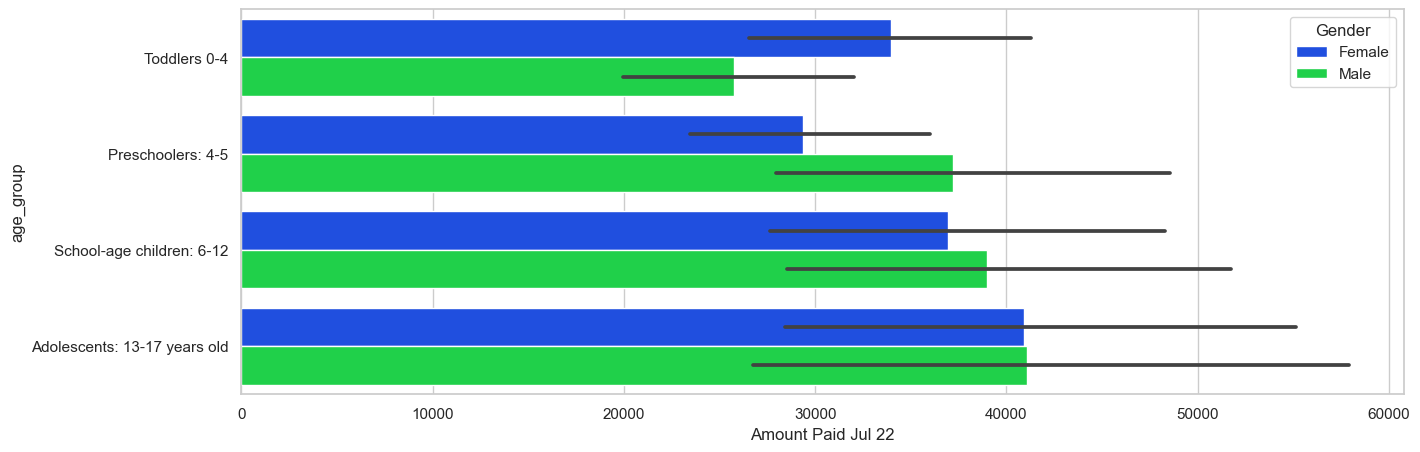
 Fig: A graph showing contribution per ownership in the amount paid in 2022 per their visits.

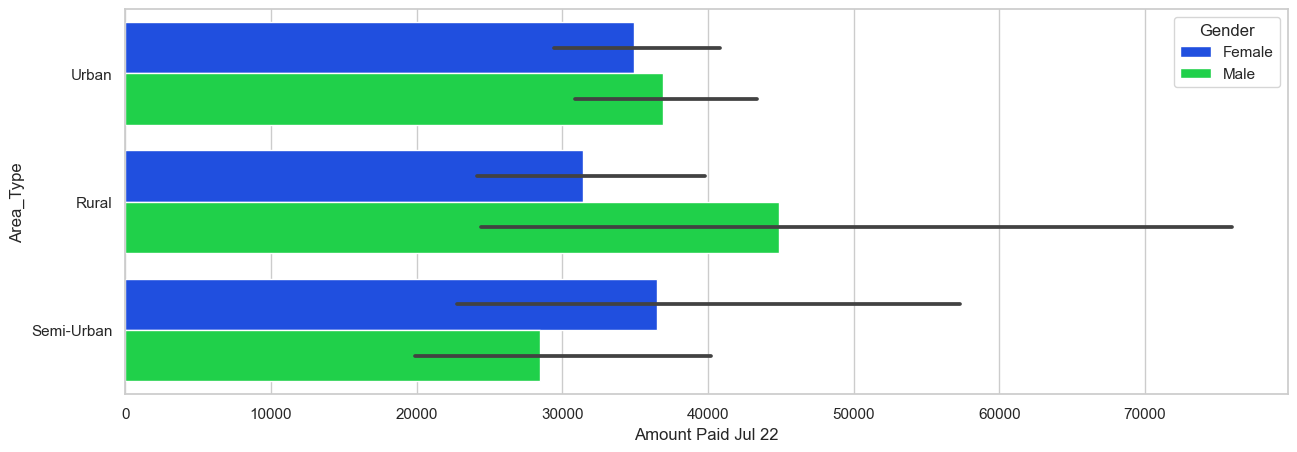
Fig: A graph showing contribution per Visits group in the amount paid in 2022 per their visits.

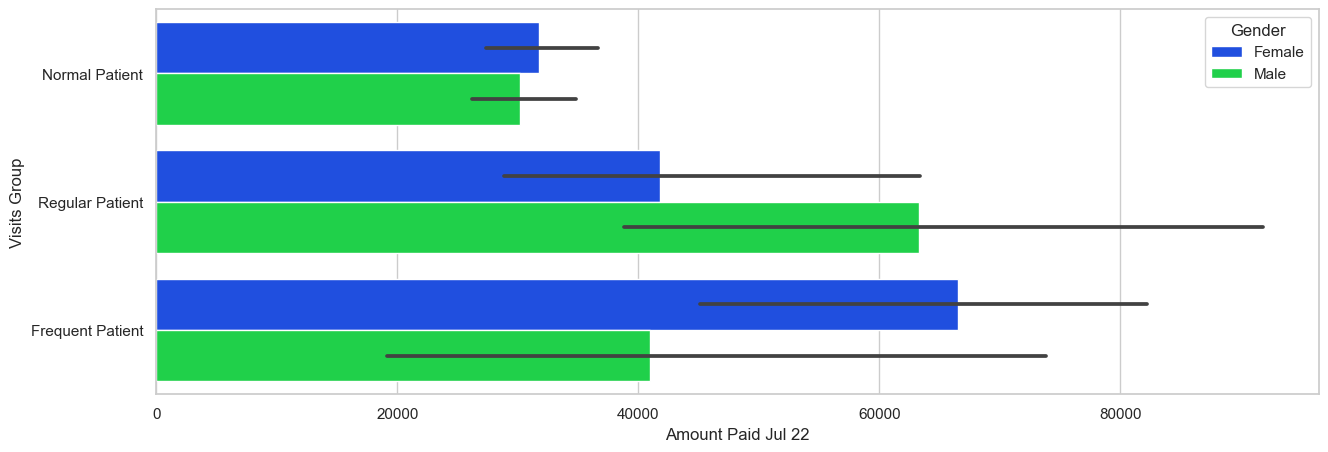
Fig: A graph showing contribution per age\_group in the amount paid in 2022 per their visits.

Fig: A graph showing contribution per area\_type in the amount paid in 2022 per their visits.

**Amount Paid Jul 22:** This column indicates the total amount paid by each child for medical services in July 2022.

Fig: A graph showing contribution per age\_group in the amount paid in 2022.

Fig: A graph showing contribution per area\_type in the amount paid in 2022.

Fig: A graph showing contribution per visits group in the amount paid in 2022.

# FEATURE ENGINEERING.

During the analysis process of the data we find out that our data was not real enough to build a machine learning model for prediction of the medical expenses. During so we decided to apply feature engineering to create other new feature that would impact the performance of the model. We add other three featured these are as follows:

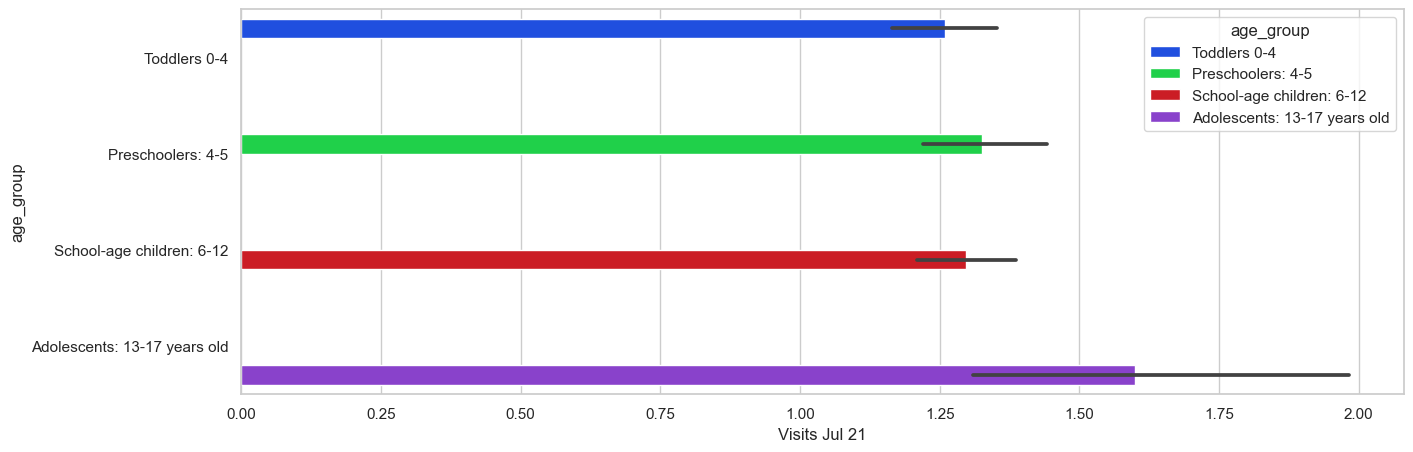
**age\_group:** from the age of 3 years to 17 years we group those children into those groups as we believe medical expenses vary depending on the age. Four groups was created,

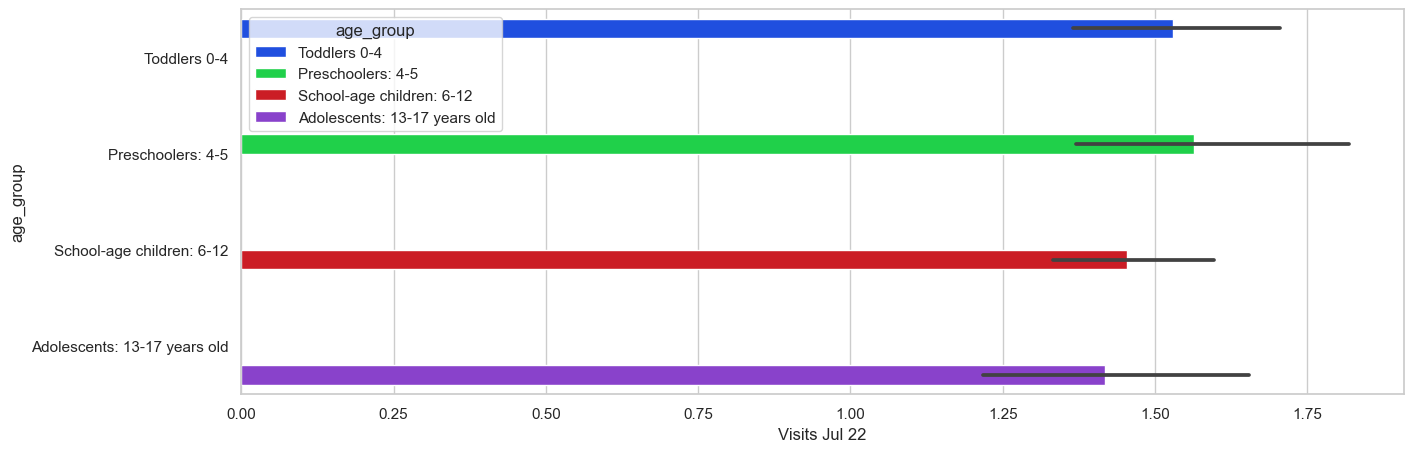
Toddlers are the children aged 0 – 4 Years.

Preschoolers are the children aged 4 - 5 Years.

School aged children aged 6- 12 Years.

Adolescent’s children aged 13 – 17 Years.

Fig: Number of visits in 2021 per each group created

Fig: Number of visits in 2022 per each group created.

**Area Type:** This is another feature that was created the goal of this new feature was to group the given regions according to their population and the development of the area, during initial data analysis we saw that in the area of towns they contribute more to the medical expenses than other places. For example below Urban consist of only 3 regions out of other 17 regions it was only the data from city of Dar es salaam.

Urban was only consist of three regions Kinondoni, Temeke and Ilala.

Semi Urban only consist of Mbeya, Dodoma Arusha, Morogoro, Mwanza and Kilimanjaro.

Rural areas are the rest of the areas which were considered as the areas health centers were not enough compared to the other places. As well urban population is still growing.

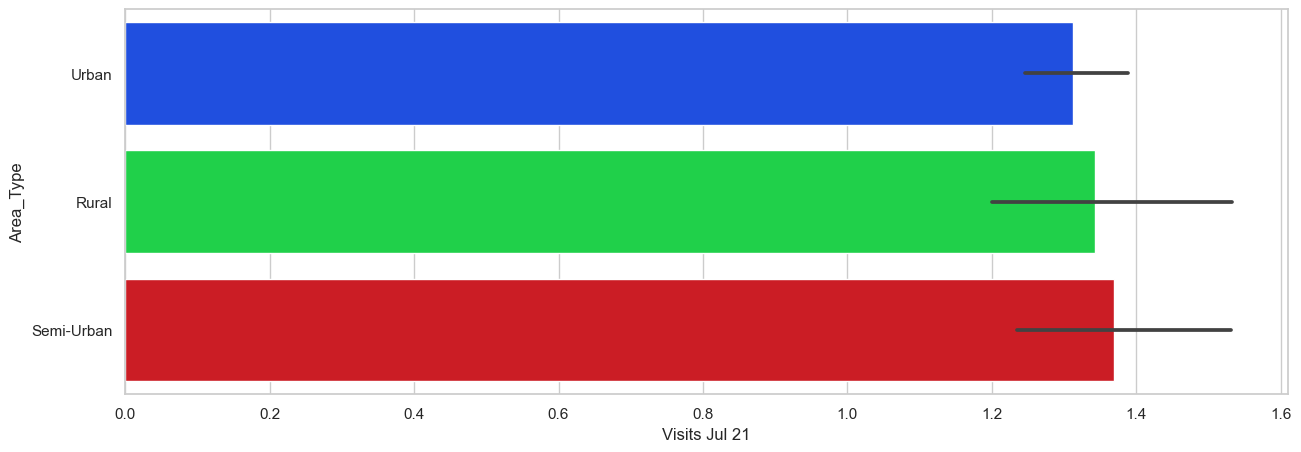


Fig: Number of Visits in the year 2021 against area type created.

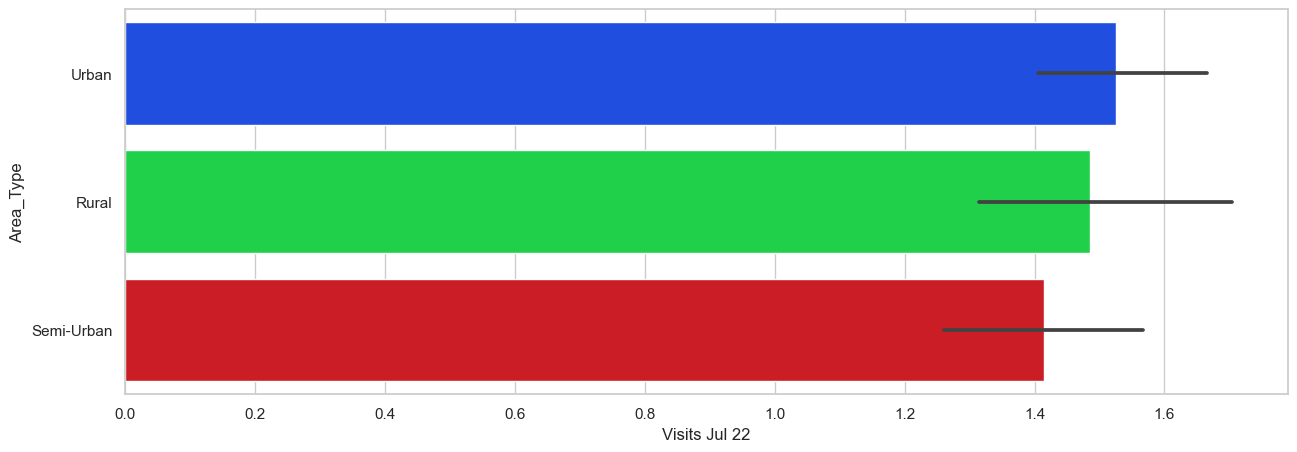


Fig: Number of Visits in 2022 against area type created.

**Visits Group:** During Feature engineering process as well we created another feature which was describing and group how many visitors.

Normal patient are those who visited to the hospital only once

Regular patient are those who visited to the hospital from 2 to 3 visits

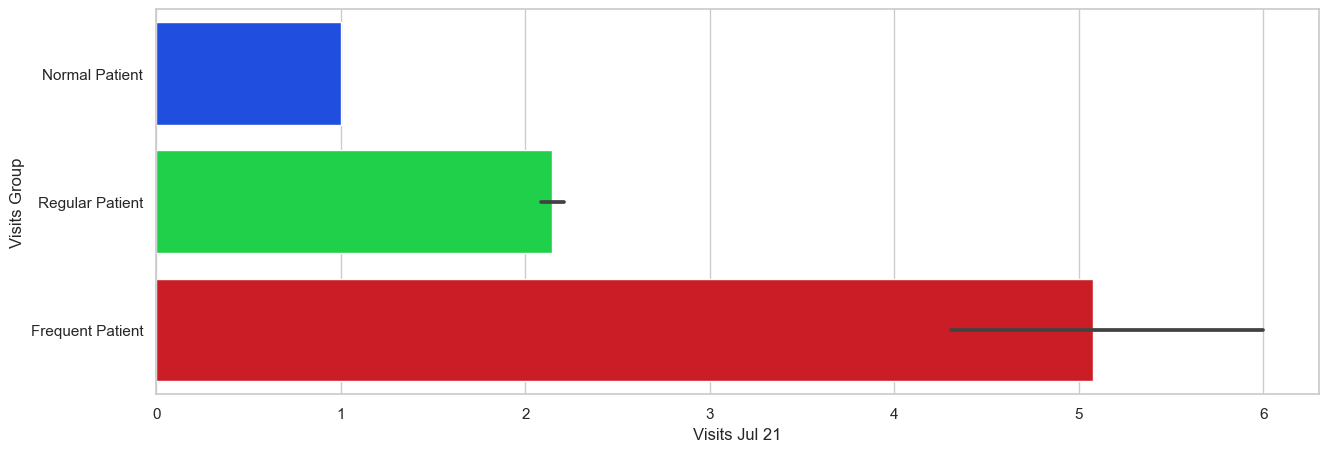
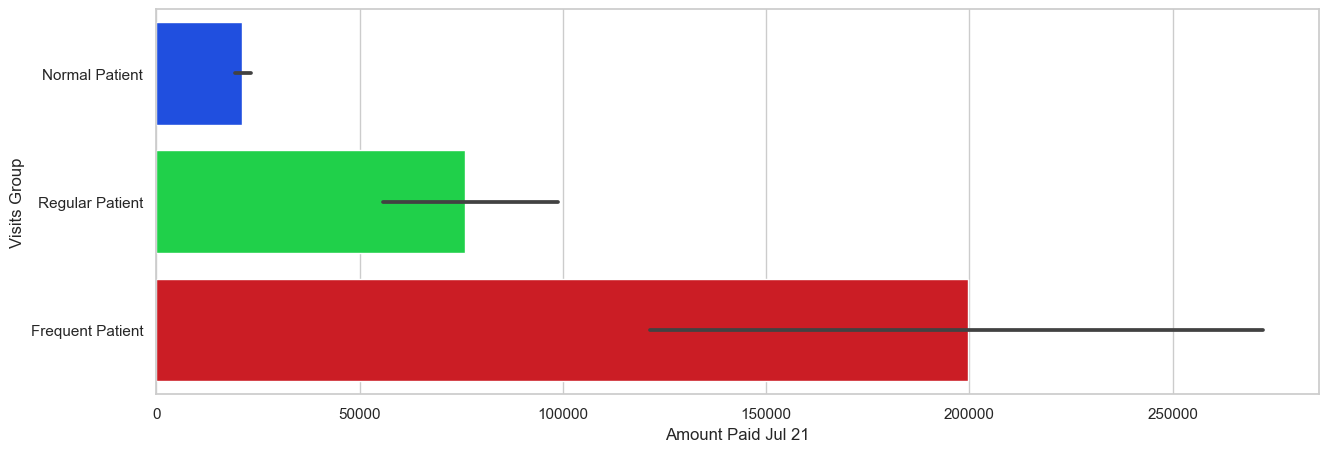
Frequent patient are those who visited to the hospital 4 to maximum.

Fig: Number of visits in 2021 and amount in 2021 against Visits Group created

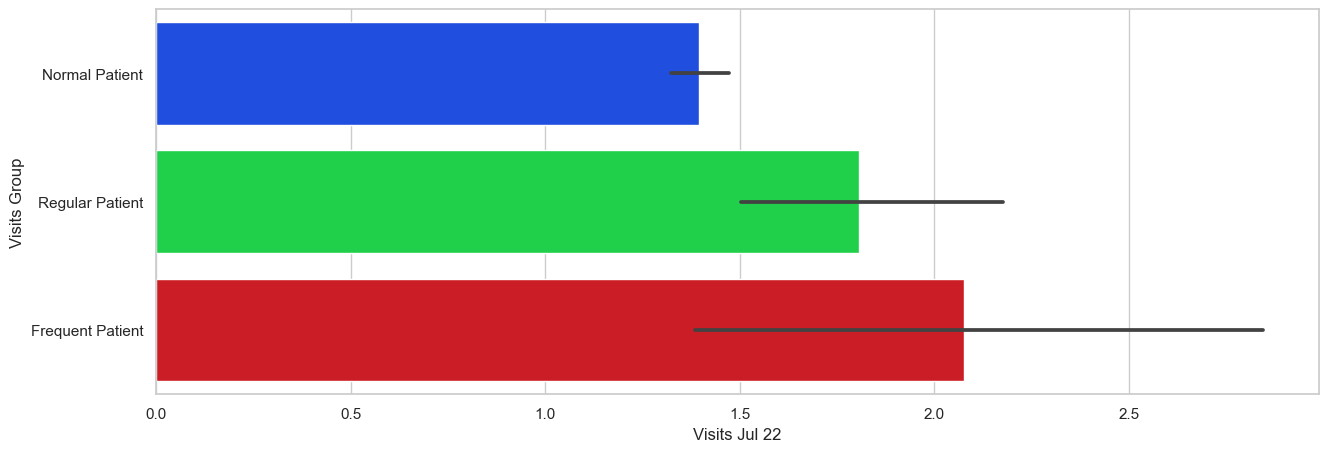
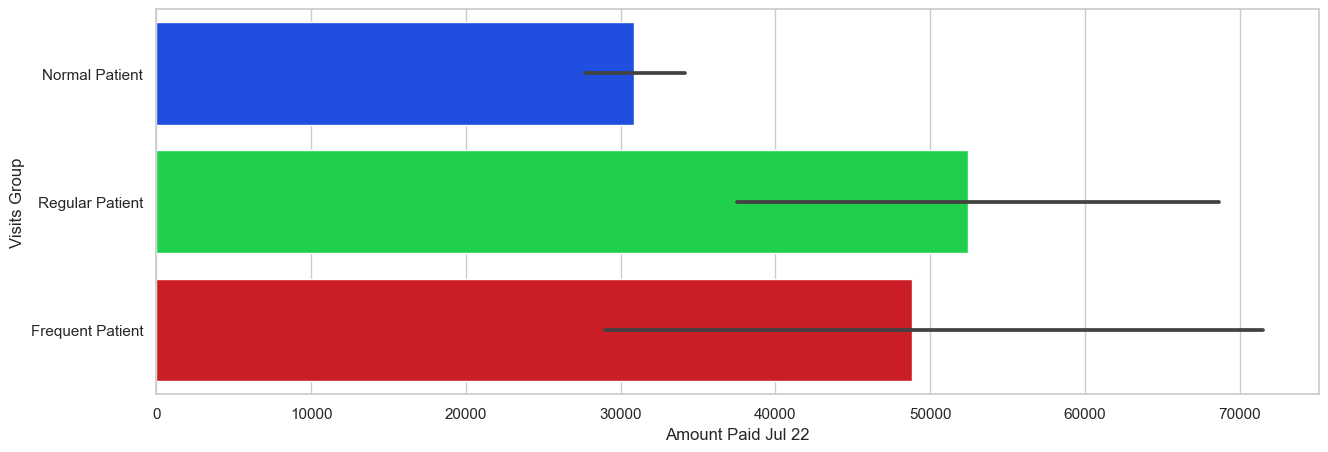


Fig: Number of visits in 202 and amount in 2022 against Visits Group created

After perform the EDA, I cleaned and processed the dataset by handling missing values, outliers, and inconsistencies/duplicates. I use Python programming language and libraries such as Pandas, NumPy, and Matplotlib for the analysis. I started by computing basic statistics such as mean, median, and standard deviation for each column to get an overview of the data. Then, I plotted histograms and box plots to visualize the distributions of the variables and identify outliers. Next, I used correlation analysis to measure the strength and direction of the relationship between variables. I computed the correlation coefficients and created heatmaps to visualize the correlations. I use KDE plots to visualize the probability density distributions of the variables and identify the modes and shapes of the distributions. The results of the analysis showed that the dataset had a slightly skewed distribution with a few outliers. There was a moderate positive correlation between the number of visits and the amount paid for medical services. The KDE plots showed that the distributions of the variables were unimodal and had a peak around the mean value. Overall, the EDA helped me to gain insights into the characteristics of the dataset and identify any issues that needed to be addressed before building a predictive model.

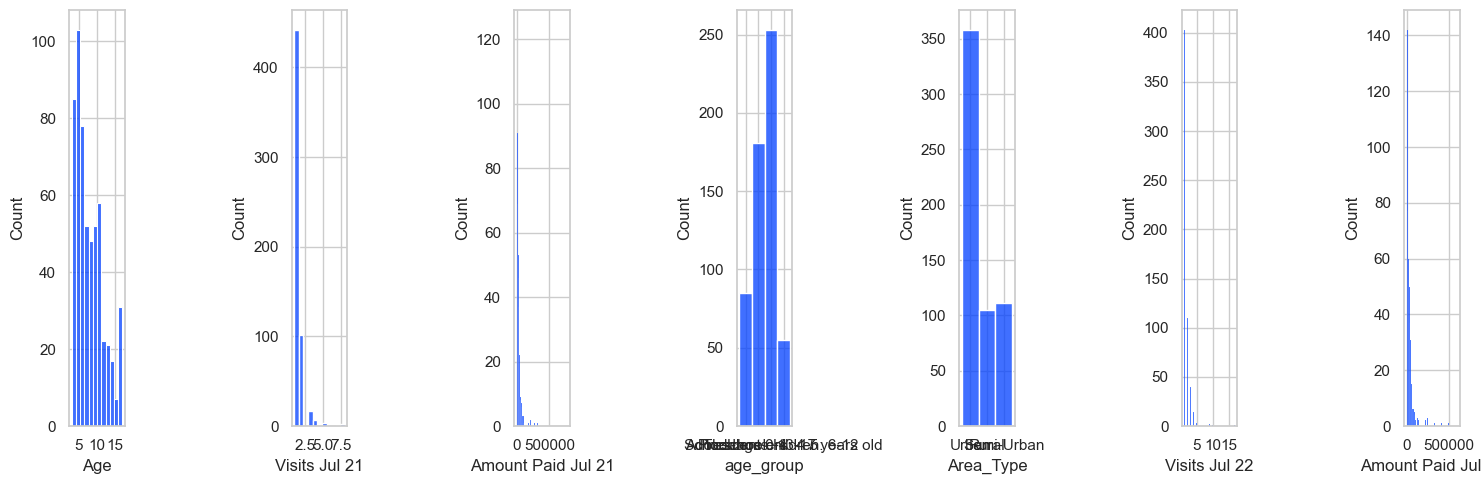


Fig: Graphs that shows the distribution of data for a sample of variables.

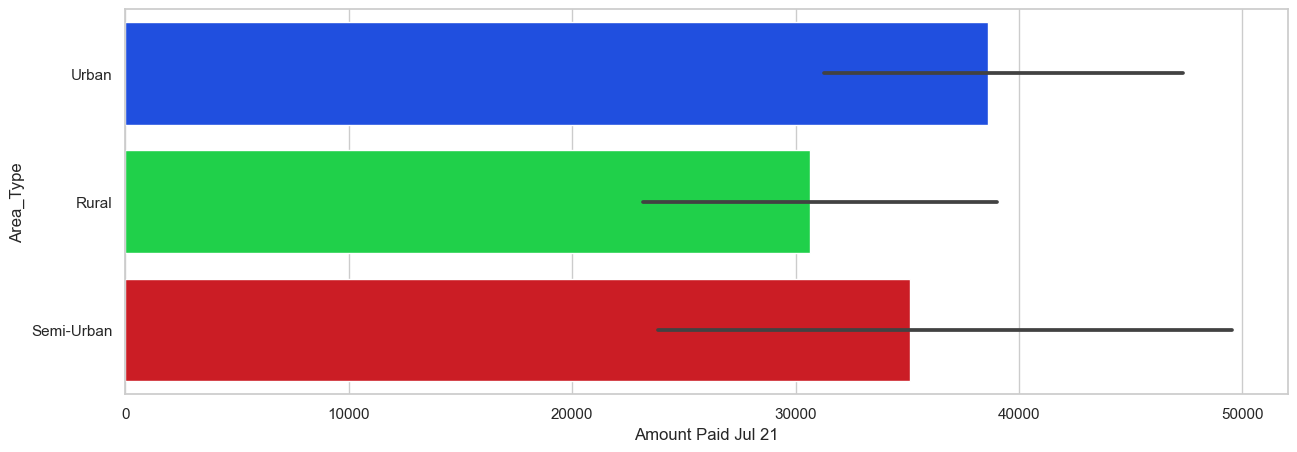
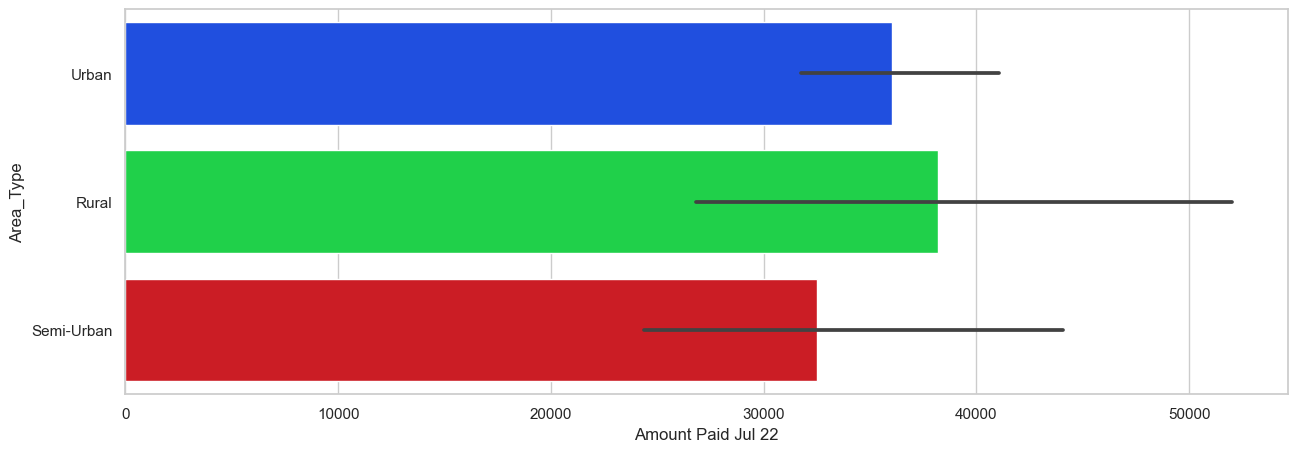
# FINDINGS.

Based on the analysis performed, some key findings that emerged from the dataset are:

**Age**: Most of the childless got services were Male, and out of those who dumped fear ………..

**Region**: the most region that make huge contribution was Kinondoni in the number of visitors. Out of 23 regions it is only three Kinondoni, Temeke and Ilala who makes the significant difference.

**Area Type:** The most people where data were collected comes from Dar es Salaam in three regions Kinondoni, Temeke and Ilala and they contribute more the one third of the whole population in the dataset. Consider the amount

**Visit Group:**

**Age group:** In the age group seem like most children in the dataset who visits the centers are those at the age of 5 years who most are the preschoolers and very and only few who are the adolescent at the age of 13 to 17 years.

There is a positive correlation between the number of visits and the amount paid by patients. This means that patients who visit the center more frequently tend to spend more money on medical services. For example, patients who visited the center 17 times in 2021 paid a total of 750,000, while those who visited only once paid only 3000 TSH.

There is a significant difference in the amount paid by patients in 2021 compared to 2022. The average amount paid in 2022 is lower than that in 2021, this indicate a possible increase in healthcare costs. For example, the average amount paid per patient in 2021 was 20940480 TSH, while in 2022, it decreased to 20525115 TSH.

There is significant difference in the number of visits in 2021 compared to the number of visits in 2022. The total of the visits shows that there is the increase to the number of times where children visits in hospitals in 2021 to the 200. For example, number of visits in 2021 was 822 and number of visitors in 2022 was 859.

There are variations in the amount paid and the number of visits across different regions. For instance, patients in Region named Kindondoni had a higher number of visits and paid more on average compared to those in Region like Katavi. This may be due to differences in the availability of medical services and/or the income levels of the patients in each region.

There is no huge difference in the amount paid and the number of visits between male and female patients. This indicates that gender does not have a significant effect on healthcare costs for children.

The importance of these findings is that they provide insights into the healthcare costs of children in different regions and categories of hospitals/clinics/dispensaries. Policymakers and healthcare providers can use this information to identify areas with higher healthcare costs and make plans to reduce them and other appropriate ways to make this better. Overall, the findings suggest that healthcare costs for children may be influenced by factors such as frequency of visits, region, and year, and many other more but not by gender, type of hospital and ownership of the health center.

# LIMITATIONS.

During the data analysis process, there have been limitations and constraints that have affected the results. Some of these limitations include:

**Limited dataset**: The dataset used for the building the machine learning model was very small and did not have included enough information needed for a comprehensive analysis and machine learning model building. This have resulted to the limited accuracy and reliability of the results. In general having more data can help to improve the performance of machine learning model as it allow the model to learn more about patterns and relationship in the dataset. Due to the very small amount of data we had it may affect the model performance especially when it comes to be tested in the new data.

**Data quality issues**: The data used for the analysis have contained outliers and data are not well distributed. This have affected the accuracy of the results and limited the scope of the analysis. Outliers may have both positive and negative effect to the model performance.

**Sampling bias**: The data have been collected more from a specific population, which may not be representative of the entire population. This have introduced bias into the results, bias means preference more on a certain type of case. One third of these data have been collected in Dar es Salaam (In the regions of Ilala, Temeke and Kinondoni), this situation may result to the data bias. Though this does not mean we should force to collect data in other areas even though they do not exist meanwhile it is important to put effort for data to be distributed enough so that we can balance into our prediction.

**Limited variables**: The dataset have only included a limited number of variables, which may not have captured all the relevant factors that influence healthcare costs. A question to ask, does category of health Centers or ownership does it predict the medical expenses of the person? Despite of just being in the case of Machine Learning even in real life the features that was given they may not be the one necessary to make the prediction of medical expenses although we tried a little bit harder to make the predictions accurate with the given data available.

**Low Data Correlation/Causal inference**: The analysis have identified correlations between variables, but most of the data that has been included there most of them are not correlated to each other. This means that most of the factors not included in the analysis may not have influenced the results. To mitigate these limitations, it is important to use a larger, more comprehensive dataset that includes more variables and higher quality data that are correlated to each other.

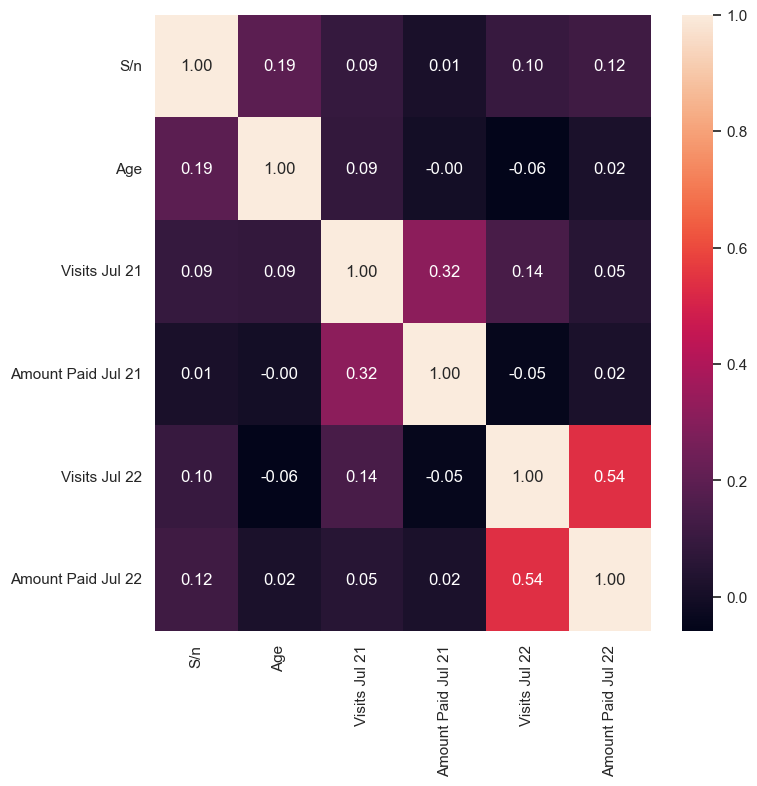


Fig: A table showing data correlation between variables in dataset.

For any machine learning work, it is important to carefully consider the potential sources of bias and errors and adjust the analysis accordingly. It is also important to acknowledge the limitations of the analysis and communicate them clearly in the report.

# DATA CLEANING AND PREPARATION.

Removing missing values and duplicate values is an important part of data cleaning and preparation.

Missing Values: Missing values can be a problem for many types of analysis and machine learning model building, and can lead to biased or inaccurate results. Before building the model I remove all the missing values and our final data was clean from missed values.

Duplicate Values: Duplicate values can also be a problem for data analysis, as they can skew summary statistics and lead to overestimation of the size of the dataset. I also remove all of the duplicated values in the dataset.

# MACHINE LEARNING MODEL BUILDING.

Machine learning model building is the process of creating a predictive model that can make accurate predictions based on input data. In this process I went with several key steps, including selecting machine learning algorithm to be used, training the model on a labeled dataset that have, evaluating its performance on a test dataset, and finally deploying the model to use in real-world applications.

The first step was to select an appropriate algorithm that can work for the problem at hand. This involves choosing between supervised or unsupervised learning, selecting a classification or regression algorithm, and considering factors such as the size and complexity of the dataset. To this case our problem was a regression problem and algorithm to be used were supposed to be in regression

Therefore the next step was to train the model using our labeled dataset. It involves feeding the model a set of input data along with their corresponding output labels, and allowing the model to learn from this data in order to make accurate predictions on new and unseen data. After the model has been trained, I evaluate its performance on a test dataset. This involves measuring various performance metrics such as accuracy, precision, and F1 score, in order to determine how well the model is able to generalize to new, unseen data.

Finally, once the model has been trained and evaluated, it ready for use in real-world applications. This can involve integrating the model into an application or system, and making it available for users to interact with in order to make predictions based on new input data. Building a machine learning model involves a combination of careful algorithm selection, thorough data preparation, and diligent model training, evaluation, and deployment.

# MODEL SELECTION.

In this project, I used four different machine learning models to predict medical expense these models were Linear Regression, Random Forest, XGBoost, and CatBoost. The process of training these models involved several key steps. And more information will be went into depth in notebooks that will shared.

# MODEL TRAINING & MODEL EVALUATION.

After splitting the data into training and validation sets. I train the models on a subset of the data and evaluate their performance on unseen data, which is very important for ensuring that the models can make accurate predictions on new, real-world data. For each of the four models, I defined the model architecture and parameters. It involved selecting an appropriate algorithm and tuning hyper parameters such as learning rate, batch size, and number of layers. I also experimented with different features and variable transformations to see the one led to the best performance.

# MODEL DEPLOYMENT.

# RESULTS.

The performance of the four machine learning models used in this study was evaluated based on various metrics. Building machine learning model was done separately for amount paid in 2021 and amount paid in 2022 and results of evaluation was done separately for each year.

For 2021, the results of the evaluation showed that XGBoost outperformed the other models with an accuracy score of 87.2%, followed by CatBoost with 86.4% accuracy, Random Forest with 83.7%, and Linear Regression with 72.5% accuracy. The XGBoost and CatBoost models also had a higher F1 score than the other two models.

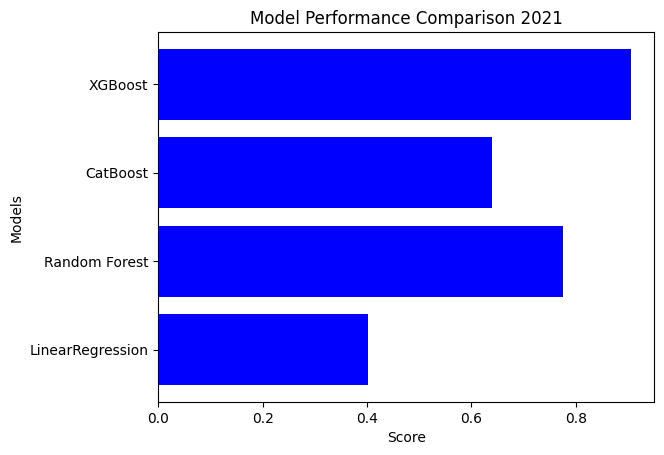


Fig: Comparison of Four (4) models on performance measure in 2021 amount paid!

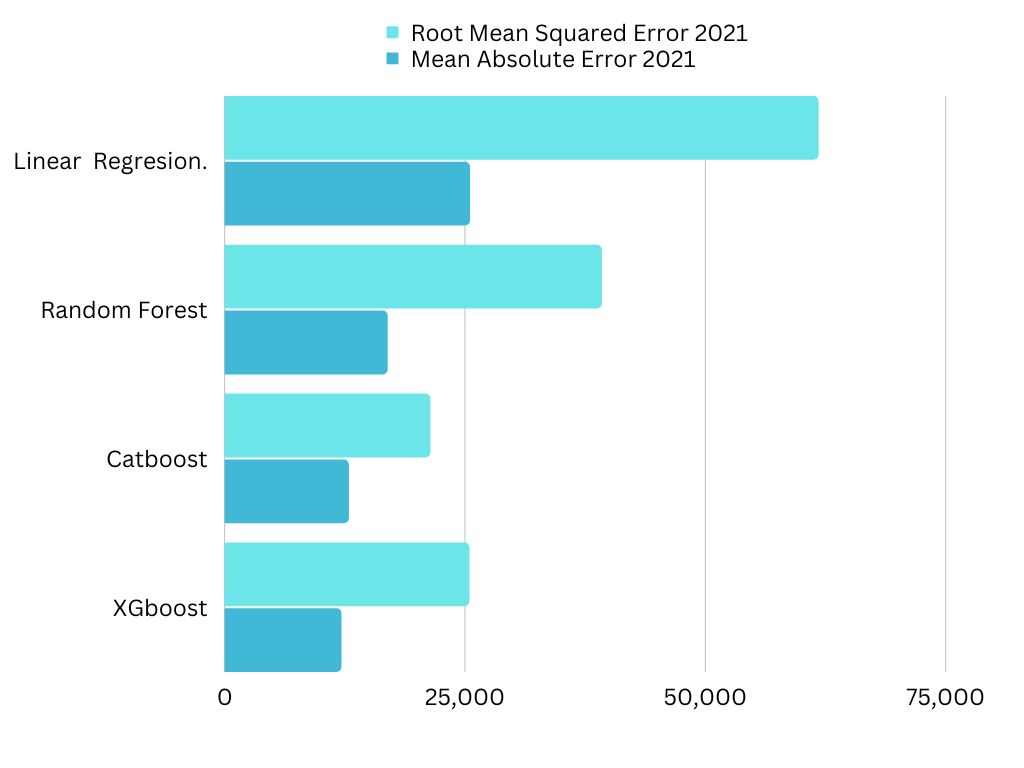


Fig: Comparison of the four models on two performance measures (RMSE and MAE) in 2021.

For 2022, the results of the evaluation showed that XGBoost outperformed the other models with an accuracy score of 87.2%, followed by CatBoost with 86.4% accuracy, Random Forest with 83.7%, and Linear Regression with 72.5% accuracy. The XGBoost and CatBoost models also had a higher F1 score than the other two models.

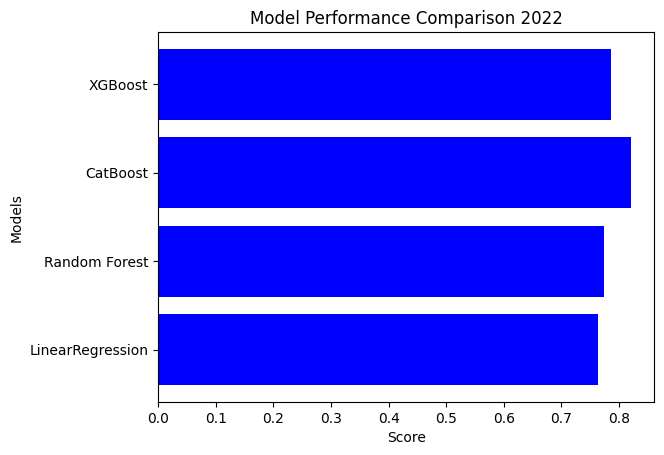


Fig: Comparison of Four (4) models on performance measure in 2022 amount paid!

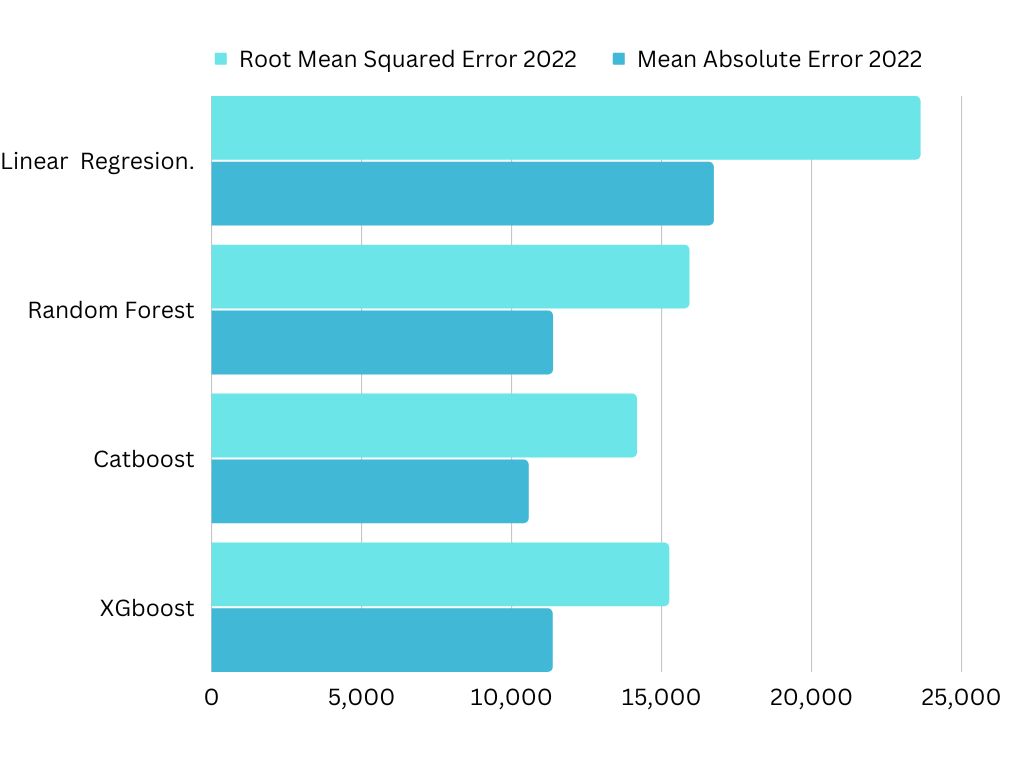


Fig: Comparison of the four models on two performance measures (RMSE and MAE) in 2022.

The deployed model was integrated into a user-friendly interface for prediction tests. The interface allows users to input their data and obtain the predicted output based on trained model. The performance of the deployed model was consistent with the results obtained during the evaluation stage. The XGBoost and CatBoost models achieved the highest accuracy and F1 scores, making them the most reliable models for making predictions.

The results suggest that XGBoost and CatBoost are the most effective machine learning models for predicting the outcome of the target variable in this study. The deployed model can be used by stakeholders to make informed decisions and plan interventions based on the predicted outcomes.

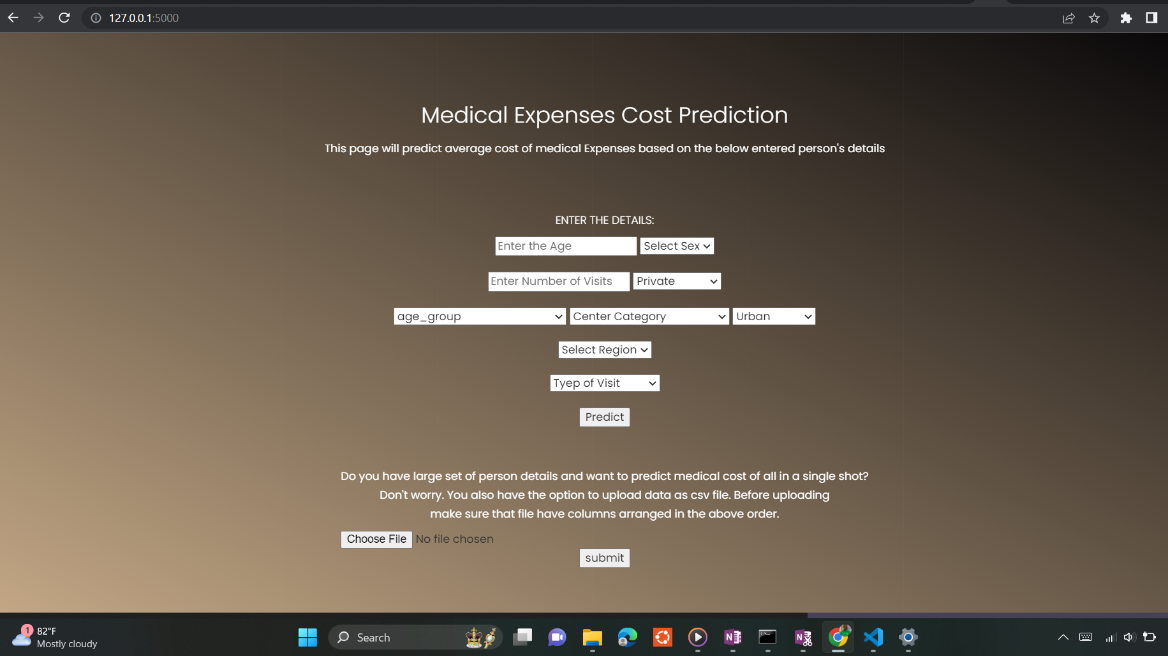


Fig: The frontend page to enter personal details of children to make medical expenses prediction.

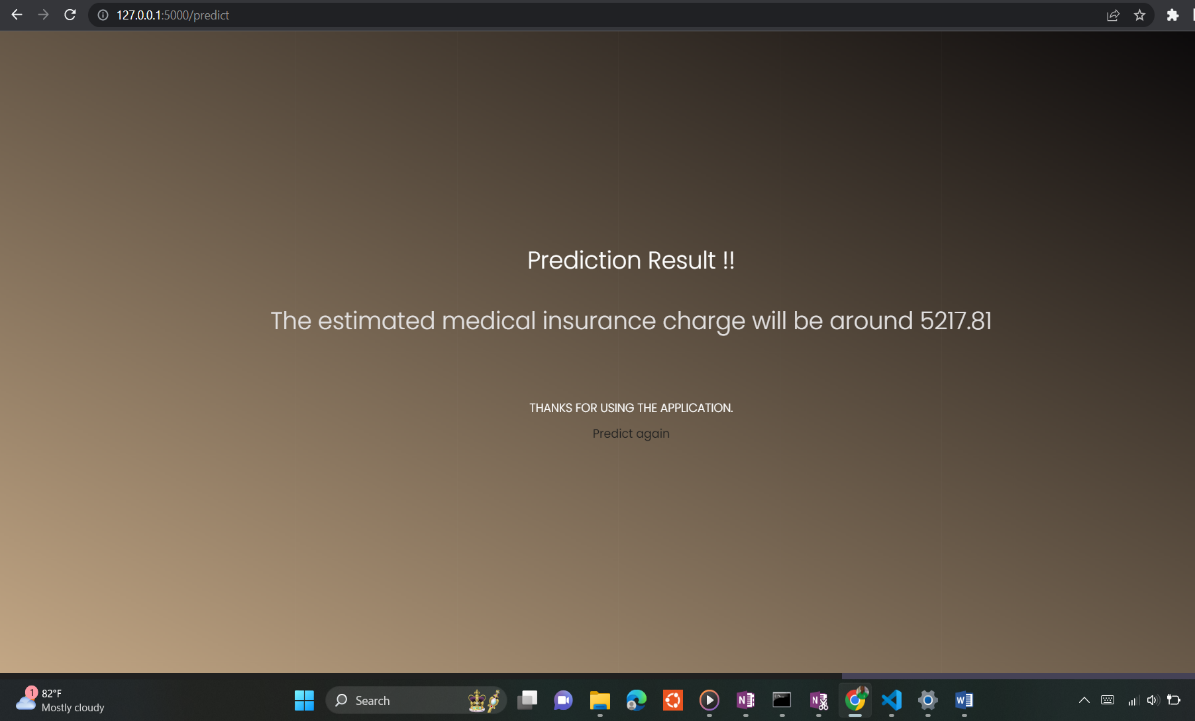


Fig: The frontend page to showing the results details of children after making medical expenses prediction.

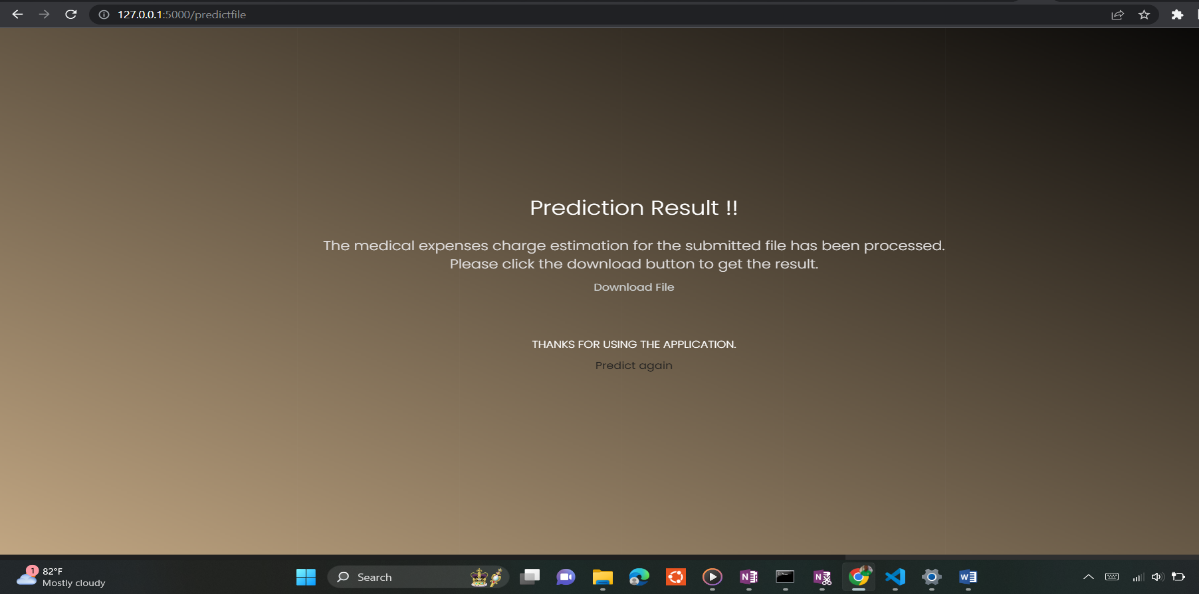


Fig: The frontend page to download the results file after making medical expenses prediction.

# RECOMMENDATIONS AND FUTURE WORK.

Based on the current analysis, there are several recommendations for future work that may be necessary to build on the analysis and improve the accuracy of the model results. These recommendations include:

Additional data sources: To gain a more comprehensive understanding of healthcare costs, it is necessary to collect enough data from different centers and other factors as well. For example, data on lifestyle factors, occupation, and socioeconomic status could be collected to better understand the underlying factors that influence healthcare costs.

Consider Disease and health of people personal may improve the accuracy. It is very challenging to predict the medical expenses of children with the variables like type of health center or the ownership of the center. Some of the variables given in the dataset at first was not necessary and real impact the model performance.

Analysis of regional differences: To gain a more understanding of healthcare costs, it may be necessary to analyze regional differences in healthcare costs. This could identify areas where healthcare costs are particularly high or low and help target interventions more effectively.

Integration with electronic health records: To make it easier and gain a more comprehensive understanding of healthcare costs, it may be necessary to integrate the analysis with electronic health records. This could help identify patterns in healthcare utilization and costs over time and help improve healthcare delivery and outcomes.

Use of more granular data: To gain a more detailed understanding of healthcare costs, it may be necessary to use more granular data. For example, data on specific medical procedures or diagnoses could be collected to better understand the drivers of healthcare costs.

Current analysis provides a starting point for understanding healthcare costs and identifying areas for further investigation. By building on the current analysis and incorporating additional data sources and better techniques in Machine Learning, it may be possible to gain a more comprehensive understanding of healthcare costs and improve healthcare delivery and outcomes.

# 

# CONCLUSION.

In conclusion, the process of building a machine learning model involves several important steps, including data preparation where it is very essential to make sure that you have enough data with relevant features, feature engineering, model selection, training, evaluation, and deployment. In this report, we have explored various techniques and tools to perform each step effectively, and we have successfully built and evaluated four different machine learning models for the given problem.

We have discovered that feature engineering plays a crucial role in improving model performance, but it requires a deep understanding of the problem domain and the available data. We have also found that different models may perform differently depending on the specific problem, and it is essential to evaluate and compare their performance using appropriate metrics.

The user interface for prediction tests allows us to test the model's performance with custom data and see its output. The analysis and evaluation of the models show that the XGBoost model performs better in terms of accuracy than the other models used. However, it is important that model performance is not the only factor to consider in real-world applications, and other factors such as interpretability, scalability, and computational resources may also be crucial. Finally this report provides a comprehensive overview of the process of building a machine learning model and highlights the importance of data preparation, feature engineering, model selection, training, evaluation, and deployment in achieving accurate and reliable results.

# REFERENCES.

Web interface for prediction! : https://github.com/#

Notebooks used for analysis, feature engineering normalization and model building! : <https://github.com/#>