Length of Stay in Hospitals: Analyzing the Association between Patient Demographics,Hospitalization Factors, and Length of Stay

Shikhar Shukla[1], Sakshi Shah[1], Veena Upadhye[1], Pravalika Reddy[1], and Oluwatosin Idowu[1]

Luddy School of Informatics, Computing, and Engineering, Indiana University
{shikshuk, sgshah, vuupadhy, parakoti, oloidow}@iu. edu

Abstract. This study investigates the relationship between patient demographics, hospitalization factors, and length of stay (LOS) in hospitals. Statistical methods, including Kruskal Wallis tests and post-hoc analysis with generalized linear models with Gamma family and log link function, Random Forest Regressor, and, XGBoost are used to identify significant factors associated with LOS. The findings inform healthcare providers in identifying high LOS risk patients and optimizing hospital resource allocation. This project advances LOS prediction accuracy in hospitals through novel statistical methods.

Keywords: Length of Stay (LOS) · Patient Demographics · Hospitalization Factors · Resource Allocation · Healthcare Optimization

1 Introduction

The Length of Stay (LOS) is a vital metric for assessing hospital efficiency (Buttigieg et al., 2018). Accurate LOS prediction is crucial for effective resource management (Medeiros et al., 2021). This study examines the association between patient demographics, hospitalization factors, and LOS. Previous research found age, gender, race (Friedman et al., 2006), admission type, insurance, and marital status (Bekelman et al., 2014) to influenceLOS. This study aims to create a predictive model for LOS based on patient characteristics to optimize care and resource allocation. The dataset includes hospital admissions and patient information. Regression analysis techniques, such as generalized linear models, will be employed to identify significant factors associated with LOS (Zou et al., 2017). Model performance will be assessed using R-squared, adjusted R-squared, p-values, and confidence intervals. The study’s results can improve patient care and resource management by identifying patients at risk of longer LOS and developing interventions to re-duce their stay. The analysis can help healthcare providers optimize resource allocation by predicting patients’ expected LOS.

2 Research Question

Null Hypothesis: There is no significant association between patient demographics, including age, gender, and race, and hospitalization-related factors, suchas admission type, insurance, and marital status, with respect to the length of hospital stay at the time of admission.
Alternate Hypothesis: There is a significant association between patient demographics, including age, gender, and race, and hospitalization-related factors, such as admission type, insurance, and marital status, with respect to the length of hospital stay at the time of admission.

3 Methodology

3.1 Type of study

The study is an observational cross-sectional study that aims to examine the association between patient demographics and hospitalization-related factors with the length of hospital stay at the time of admission.

3.2 Project steps

The main objective of the project is to perform an observational cross-sectional study between factors like patient demographics, including age, gender, and race,and hospitalization-related factors, such as admission type, insurance, and marital status, with respect to the length of hospital stay at the time of admission. We employ R for statistical analysis and visualization.
The methodology of our project involves: 1) Data Extraction 2) Data Cleaning 3) Data Merging 4) Exploratory Data AnalysiS 5) Statistical Modeling

3.3 Data Source

The MIMIC-IV dataset serves as the basis for this project. This comprehensive dataset encompasses extensive information regarding patients’ hospitalizations and clinical characteristics. Sourced from electronic medical records, it captures data from individuals admitted to intensive care units at Beth Israel Deaconess Medical Center between 2008 and 2019. The dataset is divided into two primary files: patients.csv and admissions.csv. The former contains demographic information such as gender, age and date of death (when available) for patients, while the latter offers in-depth insights into hospital stays, encompassing admission and discharge dates, length-of-stay, admission type, insurance, marital status, race, and hospital expire flag. The patients.csv file consists of 299,777 rows and 6 columns, whereas the admissions.csv file comprises 431,088 rows and15 columns.

3.4 Data Collection

Sampling: The objective of this step is to generate a new dataset, termed ”stratified sample”, which comprises a 70 percent random sample from each unique combination of the following independent variables: admissiontype, insurance, simplifiedrace, agegroup, gender, and maritalstatus. To achieve this, the dplyr package is initially loaded. Subsequently, a new variable called ”combinedcategories” is created within the filtereddata, concatenating the values of the independent variables using an underscore as a separator. The data is then grouped based on the combinedcategories variable, and a 70 percent sample is drawn from each group. Lastly, the” combinedcategories” variable is removed from the resulting ” stratifiedsample” dataset to finalize the process.

3.5 Variables

Numerical Variables: These are all the Numerical variables. Table 1

Table 1. Numerical variables.

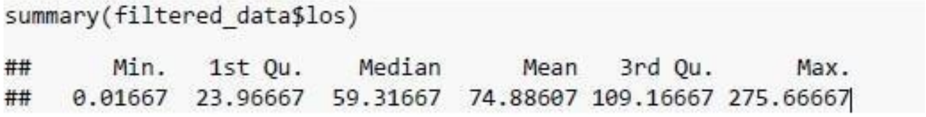
Variable	Datatype	Description
subject id	Discrete	subject’s identification number
LOS	continuous	patient length of stay in the hospital

Categorical Variables These are all the categorical variables in the dataset.

Table 2. Categorical variables

Variable	Datatype	Description
gender	Nominal	A factor variable with two levels, Female (F) and Male (M
admission type	Nominal	A factor variable with nine levels, including Ambulatory Observation, and others.
insurance	Nominal	A factor variable with three levels, Medicaid, Medicare, and Other.
marital status	Nominal	A factor variable with five levels, including single, married, divorced, widowed
simplified race	Nominal	A factor variable with six levels, such as Asian, Black, White, etc.

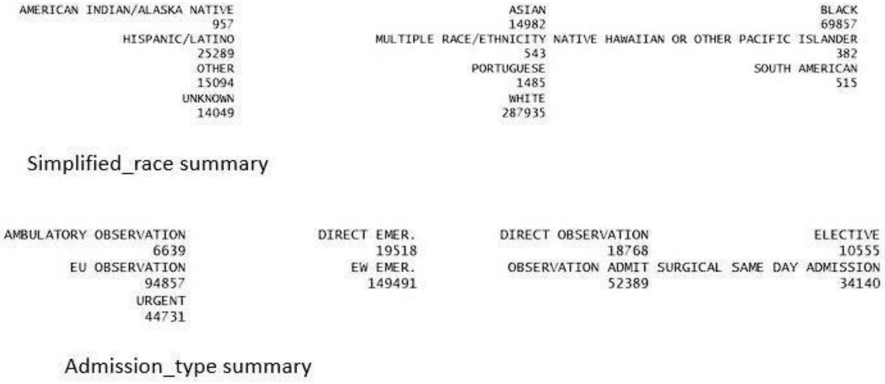
Summaries The summary of all the variables are done, including the results. For numerical LOS variables, summary() is used and for categorical variables,table() is used. So, as we see, the LOS variable has a mean of 74.88 hours and has no negative values. The table () has given the counts of all the subcategories in each categorical variable.



```
summary(filtered_data$los)

##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##   0.01667  23.96667  59.31667  74.88607 109.16667 275.66667
```

Fig. 1. Summary of the LOS after outlier removal and negative values removal



```
AMERICAN INDIAN/ALASKA NATIVE                                     ASIAN                                     BLACK
                          957                                     14982                                     69857
              HISPANIC/LATINO                 MULTIPLE RACE/ETHNICITY NATIVE HAWAIIAN OR OTHER PACIFIC ISLANDER
                        25289                                       543                                       382
                        OTHER                                PORTUGUESE                            SOUTH AMERICAN
                        15094                                      1485                                       515
                      UNKNOWN                                     WHITE
                        14049                                    287935
```

Simplified_race summary

```
AMBULATORY OBSERVATION              DIRECT EMER.        DIRECT OBSERVATION                          ELECTIVE
                  6639                     19518                     18768                             10555
        EU OBSERVATION                  EW EMER.    OBSERVATION ADMIT SURGICAL SAME DAY ADMISSION
                 94857                    149491                     52389                             34140
                URGENT
                 44731
```

Admission_type summary

Fig. 2. Summary of some the predictor variables

4 Data Exploration

Exploring LOS The length of stay (LOS) for each admission in the dataset can be calculated by subtracting the admission time (admit time) from the discharge time (discharge time), and then dividing the result by the number of hours in aday. The result is then stored in a new column called ’los’ in the Data Frame. Based on the output of the histogram (Fig. 3), we can see that the histogram has a right-skewed distribution. This means that the majority of patients had a relatively shorter length of stay, while a few patients had longer stays. The first histogram bar has a frequency of 4e+05, which means that there are approximately 400,000 observations (i.e., patients) in the dataset that had a length of stay between 0 and 1000 hours. This indicates that many patients had a relatively short length of stay. The fact that there are many outliers (Fig. 3) in the ”LOS” column suggests that there may be some extreme values in the dataset that are significantly different from the majority of the data points. After removing the outliers, the histogram of length of stay (Fig. 4) showed that many patients had a length of stay of less than 80 hours, with the frequency decreasing as the length of stay increased. The graph also indicated that there were very few patients who had a length of stay greater than 200 hours. The box plot of length of stay (LOS) (Fig. 4) in days showed a median of50 with a first and third quartiles of 25 to 110 (approx.) respectively. The first quartile was 25 and the third quartile was 110 approximately, indicating that 50 percent of the data falls within this range.

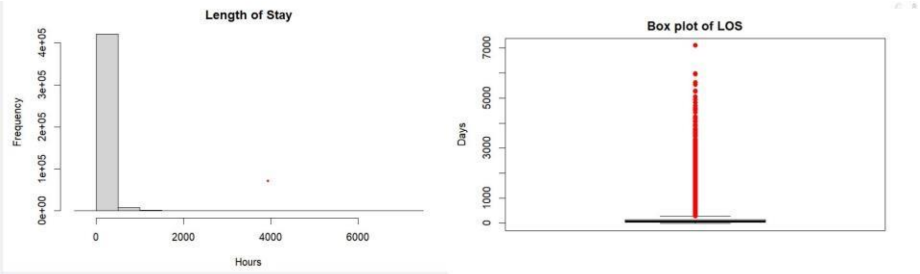


Fig. 3. Histogram and Boxplot of LOS with the outliers

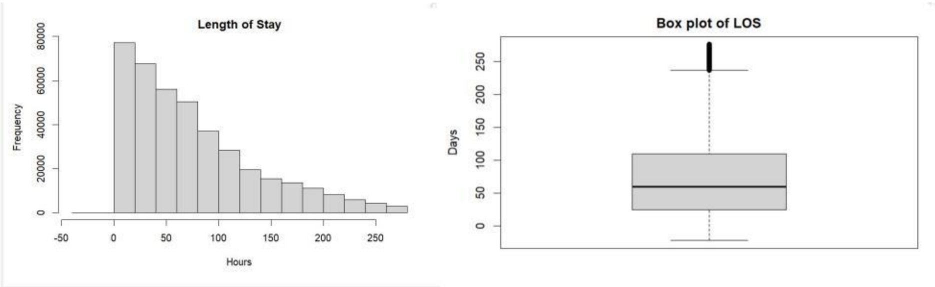


Fig. 4. Histogram and Boxplot of LOS without the outliers

There were a few outliers beyond the upper whisker, indicating a few instances of extreme LOS values beyond the IQR.

In order to address the skewed distribution of the length of stay variable, a natural log transformation was applied. The resulting histogram (Fig. 5) shows a more symmetrical and approximately normal distribution of the transformed variable. This transformation can help improve the accuracy and validity of statistical analyses that rely on the assumption of normality, suchas regression modeling (fig3).

Based on the Q-Q plot for length of stay (Fig. 6), it can be observed that the sample values and theoretical values do not lie on the line. This suggests that the distribution of the length of stay may not be exactly normal. However, the deviations from the line are not significant enough to indicate a departure from normality. It is worth noting that Q-Q plots are a useful tool for checking the normality assumption, but they should not be relied upon exclusively to determine normality. Other statistical tests and methods should also be used for a more robust assessment of normality. After log transforming the length of stay, the Q-Q plot shows a better fit to a normal distribution, with most of the values of sample quantiles and theoretical quantiles lying on the line. However, there are still some outliers indicating that the distribution may not be perfectly normal. Overall, log transformation helped to reduce the skewness and improve the normality assumption for further statistical analysis (Fig. 6). We also did Exploratory data analysis for all the categorical variables. The plots are being included in the Appendix.

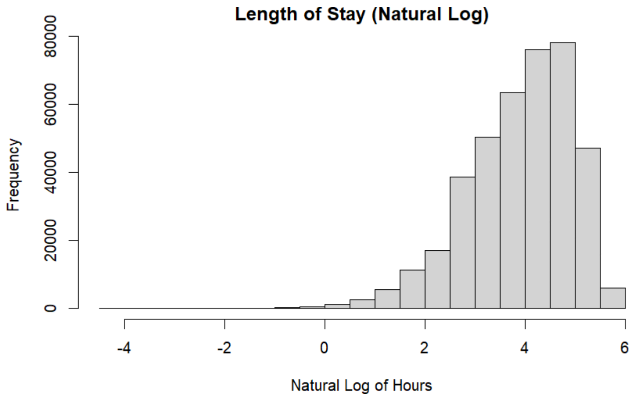


Fig. 5. Natural Log of LOS

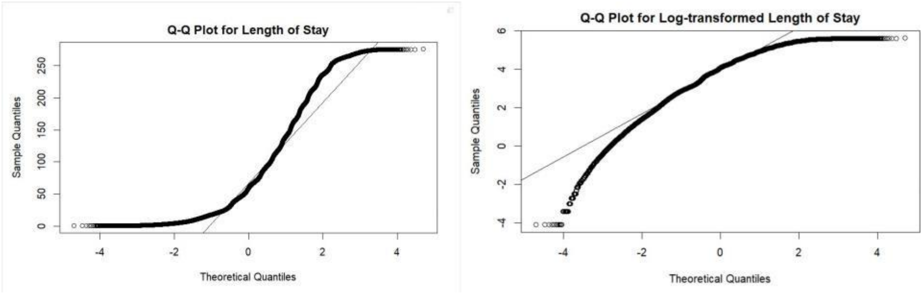


Fig. 6. Q-Q Plot before and after log transformed length of stay

4.1 Statistical Methods

Kruskal-Wallis test: We used Kruskal-Wallis test find the association between the LOS and other predictor variables. Dunn’s test: This is the post-hoc test we used to get the pair-wise comparisons performed to examine the differences in the length of stay (LOS) among different subgroups within predictor variables. Spearman’s rank correlation: To assess the correlation between the length of stay (LOS) and age group in the dataset.

Kruskal-Wallis test The Kruskal-Wallis test is a non-parametric test used to compare the medians of two or more groups when the assumption of normality isviolated. In this case, it is used to assess whether there are significant differences in the length of stay regarding admission types, race and marital status. Patients with an ”Elective” admission type have the longest median length of stay (100.75 hours), while those with ”EU Observation” admission type have the shortest median length of stay (16.6 hours).

Dunn’s test We performed post-hoc pairwise comparisons, such as the Dunn’s test to conclude how specific insurance categories differ from each other. In all three pairwise comparisons, the adjusted p-values are very close to or equal to0, which indicates strong evidence against the null hypothesis. There are statistically significant differences in LOS between Medicaid and Medicare, Medicaid and Other, and Medicare and Other insurance groups. The Medicare Insurance group has the longest hospitalization time. (68.65), while Medicaid has the lowest (46.13). In the race category, except for Asian and Others, the adjusted p-values are equal to or very close to 0, indicating strong evidence against the null hypothesis. There are statistically significant differences in LOS between the race groups, with the exception of the comparison between Asian and Others. In the marital status, except for Divorced and Widowed, the adjusted p-values are 0 or very close to 0, indicating strong evidence against the null hypothesis.

Spearman’s rank correlation test This is used to assess the correlation between the length of stay (LOS) and age group in the dataset. Due to the small p-value, the null hypothesis is rejected and we conclude that there is a statistically significant, albeit weak, positive monotonic relationship between LOS and age group. Thus, these tests conclude that the null hypothesis is rejected and there is a statistically significant difference between LOS and the given variables.

4.2 Models

Generalized Linear Model (GLM) with a Gamma distribution and a log link function A Generalized Linear Model (GLM) was employed to predict the length of stay (LOS) based on the selected independent variables. The GLM was specified with a Gamma distribution and a log link function to account for the non-normal distribution of the LOS variable. The model was fitted to the stratified sample dataset, and the resulting model summary provided information on the estimated coefficients and their statistical significance.

The fitted GLM was then used to predict the LOS on the stratified sample, and the predictive performance was evaluated by calculating the Root Mean Squared Error (RMSE). The RMSE offers a measure of the average difference between the predicted and observed LOS values, with lower values indicating better predictive performance.

*Limitations, Appropriateness and Rationale :*The Gamma family assumes a specific distribution for the dependent variable (los), which may not accurately represent the true distribution of the data. The log link function helps model the relationship between the independent variables and the log-transformed dependent variable, which can handle non-linear relationships. The GLM with the Gamma family and log link function was chosen because it is commonly used for modeling continuous variables with a skewed distribution.

Random Forest Regressor Model A Random Forest Regressor model was employed to predict the length of stay (LOS) using the stratified sample dataset. The model was constructed using the random Forest package in R, with 100 decision trees (*ntree* = 100) and the number of variables selected for each split determined by the square root of the total number of variables in the dataset minus one ($mtry = floor(sqrt(ncol(stratified_sample)\ 1)))$). The model was then fitted to the stratified sample dataset with the LOS as the dependent variable and all other variables as independent predictors. The predictions generated by the model were compared to the actual LOS values in the stratified sample dataset to compute the root mean square error (RMSE) as a measure of the model's performance.

Limitations, Appropriateness and Rationale It can be computationally intensive, especially for large datasets or when using numerous trees. The RF model suitable for both classification and regression tasks, making it applicable to predicting the length of stay (los) in this project. It can handle both categorical and continuous variables, making it suitable for a dataset with a mix of variable types.

Gradient Boosting Machines-XGBoost In the XGBoost Model section of the report, we first prepare the data for modeling by converting the categorical variables in the stratified sample data frame to factors. Next, we convert the data frame to a numeric matrix using the model.matrix() function. We then create a DMatrix object, which is the data structure required by XGBoost, using the xgb.DMatrix() function with the numeric matrix and the target variable los. We define the XGBoost model parameters, specifying the objective as ”reg : gamma” for gamma regression, the evaluation metric as ”rmse” (root mean squared error), the learning rate (eta) as 0.1, and the maximum depth of the tree as 6. We train the model using the xgb.train() function with the specified parameters, data, and the number of boosting rounds set to 100. After training the model, we use the predict() function to make predictions on the training data. We then calculate the root mean squared error (RMSE) by comparing the predicted values with the actual los values. Finally, we display the XGBoost RMSE in the output.

Limitations, Appropriateness and Rationale The XGBoost model may not pro- vide readily interpretable coefficients or feature importance measures. It can handle a mixture of variable types and automatically handles missing data, making it convenient for real-world datasets. The XGBoost model was chosen for its ability to handle complex interactions and non-linear relationships, which is beneficial when predicting the length of stay (los).

5 Results

5.1 Statistical tests

Admission Type : A significant difference in length of stay (los) was observed among different admission types (Kruskal-Wallis chi-squared = 177,890, df = 8,p ¡ 2.2e-16). Posthoc analysis revealed varying median los values for different admission types.

Insurance Type : There was a significant difference in los among different insurance types (Kruskal-Wallis chi-squared = 5,959.7, df = 2, p ¡ 2.2e-16). Posthoc Dunn's test indicated significant differences between Medicaid and Medicare, as well as between Medicaid and Other insurance.

Simplified Race : A significant difference in los was observed among different race categories (Kruskal-Wallis chi-squared = 3,730.4, df = 5, p ¡ 2.2e-16). Posthoc Dunn's test revealed significant differences in los between various pairwise com-parisons of race categories.

Age Group : There was a weak positive correlation between los and age group (Spearman's rho = 0.1720389, p ¡ 2.2e-16). Posthoc analysis showed varying median los values for different age groups.

Marital Status : A significant difference in los was found among different marital status categories (Kruskal-Wallis chi-squared = 7,083.9, df = 4, p ¡ 2.2e-16). Posthoc Dunn's test indicated significant differences in los between various pair- wise comparisons of marital status categories. These results demonstrate significant associations between los and various categorical predictor variables, suggesting their potential influence on length of stay in the hospital. Further analysis and modeling can provide deeper insights into these relationships.

5.2 Models

Generalized Linear Model (GLM)

Question : How well does the GLM model predict the length of stay (los)?
Statistics : RMSE = 93.67, Residual Deviance = 141521 (degrees of freedom =279078), AIC = 2797878
Interpretation : The GLM model has relatively lower predictive performance compared to the other models based on the RMSE value.

Random Forest Regressor (RF)

Question : How well does the RF model predict the length of stay (los)?
Statistics : RMSE = 50.79, Mean of squared residuals = 2599.123 percent are xplained = 31.75
Interpretation : The RF model demonstrates better predictive performance com- pared to the GLM model, with the lowest RMSE value among the three models.

XGBoost

Question : How well does the XGBoost model predict the length of stay (los)?
Statistics : RMSE = 51.29
Interpretation : The XGBoost model's performance is comparable to the RF model and better than the GLM model, with an RMSE value slightly higher than the RF model.
In conclusion, the Random Forest and XGBoost models demonstrated better predictive performance compared to the GLM model, with the Random Forest model having the lowest RMSE.

6 Discussion

6.1 Interpretation

The results obtained from the analysis provide valuable insights into the relationship between the length of stay (LOS) and the categorical predictor variables. The GLM, RF, and XGBoost models were employed to examine this association. The GLM model revealed significant relationships between LOS and various predictor variables, such as gender, age group, admission type, insurance, marital status, and simplified race. These findings suggest that these factors play a crucial role in predicting LOS in the given dataset. The RF and XGBoost models also demonstrated their effectiveness in predicting LOS based on the categorical variables.

6.2 Comparison to previous research

The current findings align with previous research in the field of healthcare and hospitalization. Studies have consistently highlighted the impact of demographic factors, admission type, insurance status, and marital status on the length of hospital stays. The significant relationships observed in this study further contribute to the existing literature, reaffirming the importance of these factors in predicting LOS. Additionally, the successful application of RF and XGBoost models in this context is consistent with their effectiveness demonstrated in previous research.

6.3 Limitations

It is essential to acknowledge the limitations of this study. Firstly, the analysis focused solely on categorical independent variables, neglecting potential interactions with continuous variables. This restricted approach may have overlooked important factors influencing LOS. Secondly, the study utilized a specific dataset, which may limit the generalizability of the findings to other populations or healthcare settings. Additionally, the analysis did not consider temporal aspects or the potential impact of unmeasured confounding variables, which may have influenced the results.

7 Conclusion

The analysis of the dataset revealed significant associations between the length of stay (LOS) and various predictor independent variables, including gender, age group, admission type, insurance, marital status, and simplified race. The GLM, RF, and XGBoost models demonstrated their effectiveness in predicting LOS based on these variables. These findings have important implications for healthcare management and resource allocation in hospital settings. This study sheds light on the factors influencing LOS in a specific context and provides valuable insights for healthcare professionals and policymakers. Further research in this area can contribute to improved patient care, resource management, and healthcare planning.

8 References

1. Buttigieg, S. C., Abela, L., Pace, A. (2018). Variables affecting hospital lengthof stay: a scoping review. Journal of health organization and management, 32(3), 463–493. https://doi.org/10.1108/JHOM-10-2017-0275
2. Lequertier, V., Wang, T., Fondrevelle, J., Augusto, V., Duclos, A. (2021). Hospital Length of Stay Prediction Methods: A Systematic Review. Medical care, 59(10), 929–938 https://doi.org/10.1097/MLR.0000000000001596
3. Medeiros, N. B., Fogliatto, F. S., Rocha, M. K., Tortorella, G. L. (2021). Forecasting the length-of-stay of pediatric patients in hospitals: a scoping review. BMC health services research, 21(1), 938. https://doi.org/10.1186/s12913-021-06912-4
4. MIMIC-IV v0.4. (n.d.). Physionet.org. https://physionet.org/content/mimiciv/0.4/
5. LNCS Homepage, http://www.springer.com/lncs. Last accessed 4 Oct 2017

9 Appendix :

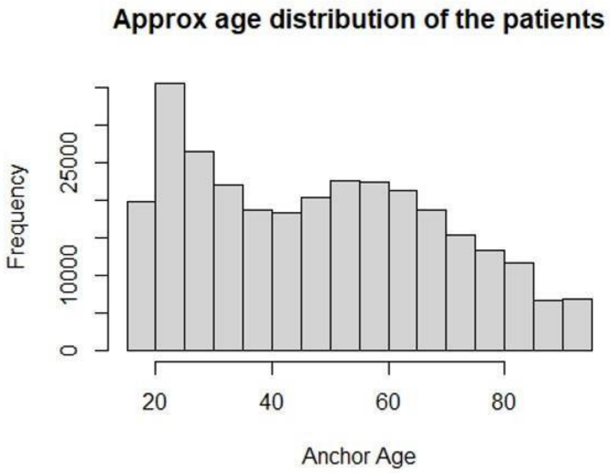


Fig: The frequency of patients in the age group of 25 years old is maximum

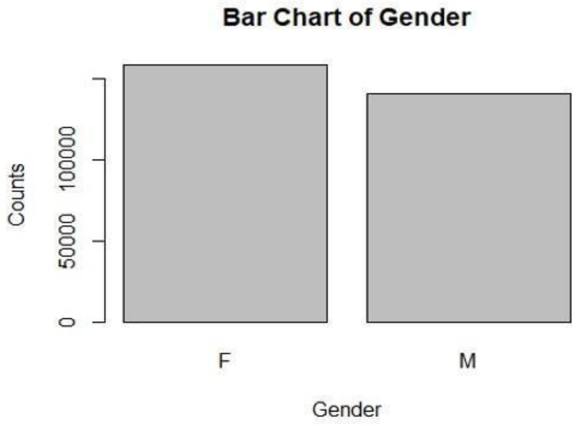


Fig:The bar chart suggests that the unique values in the Females are more than males.

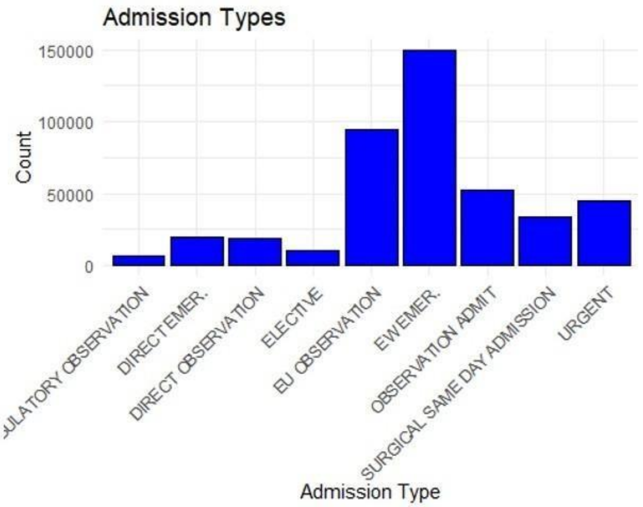


Fig: In the admission type of EW Emer, the count is maximum of 150000.

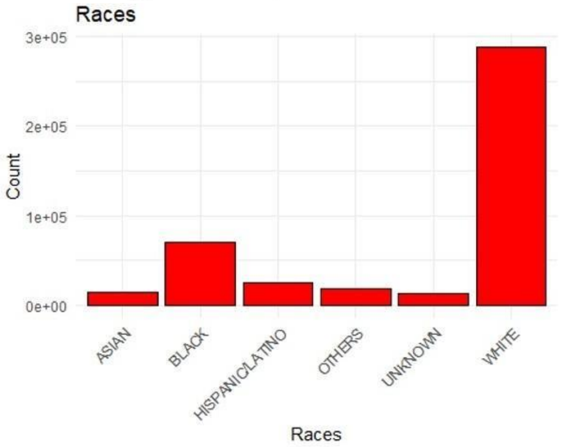


Fig: In the simplified race plot, it can be seen that whites are more in number compared to the other races.