**Emergency Department Wait Times Analysis Using Multiple Linear Regressions.**

**[Multivariate Data Analysis Final Project]**

**BIA 652**

**[Lecturer: Feng Mai]**

**[Date: 12/18/2023]**

Table of Contents

**[Executive Summary](#_Toc153441937)** [3](#_Toc153441937)

**[1. Introduction](#_Toc153441938)** [3](#_Toc153441938)

**[1.1 Background](#_Toc153441939)** [3](#_Toc153441939)

**[1.2 Objectives](#_Toc153441940)** [3](#_Toc153441940)

**[2. Exploratory Data Analysis (EDA)](#_Toc153441941)** [3](#_Toc153441941)

**[2.1 Dataset](#_Toc153441942)** [3](#_Toc153441942)

**[2.2 Data Preprocessing](#_Toc153441943)** [4](#_Toc153441943)

**[2.3 Racial Disparities in Wait Times](#_Toc153441944)** [4](#_Toc153441944)

**[2.4 WAITTIME Distribution and Relationship with PAINSCALE](#_Toc153441945)** [5](#_Toc153441945)

**[2.5 WAITTIME across the months 6](#_Toc153441946)**

**[3. Feature Selection and Correlations 7](#_Toc153441948)**

**[3.1 Selected Variables 7](#_Toc153441949)**

**[3.2 Correlation Analysis 8](#_Toc153441950)**

**[3.3Most important features in the prediction of wait time 8](#_Toc153441947)**

**[4. Modeling and Evaluation 8](#_Toc153441951)**

**[4.1 Dataset Splitting 9](#_Toc153441952)**

**[4.2 Linear Regression Models 9](#_Toc153441953)**

**[5. Results and Discussion 9](#_Toc153441954)**

**[5.1 Elastic Net Model 9](#_Toc153441955)**

**[5.2 KNN Model 9](#_Toc153441956)**

**[5.3 Ridge Regression Model 10](#_Toc153441957)**

**[5.4 Model Comparison 10](#_Toc153441958)**

**[6. Visualizations and Interpretations 11](#_Toc153441959)**

**[6.1 Visualization of Model Performance 11](#_Toc153441960)**

**[6.2 Scatter Plot of Predictions 12](#_Toc153441961)**

**[7. Conclusion 13](#_Toc153441962)**

**Executive Summary**

This report aims to provide a comprehensive analysis of patient wait times in the Emergency Department (ED). Leveraging a dataset obtained from ED records, the analysis encompasses data preprocessing, exploratory data analysis (EDA), and the application of linear regression models to predict wait times. The study sheds light on racial disparities, distribution characteristics, and the performance of various regression models.

**1. Introduction**

**1.1 Background**

The Emergency Department (ED) plays a critical role in healthcare by providing immediate care to patients in urgent need. Efficient management of patient flow in the ED is crucial for ensuring timely and effective medical interventions.

**1.2 Objectives**

The primary objectives of this analysis are to:

* Understand the factors influencing patient wait times in the ED.
* Explore potential disparities in wait times across different racial groups.
* Develop and evaluate linear regression models for predicting wait times.

**2. Exploratory Data Analysis (EDA)**

**2.1 Dataset**

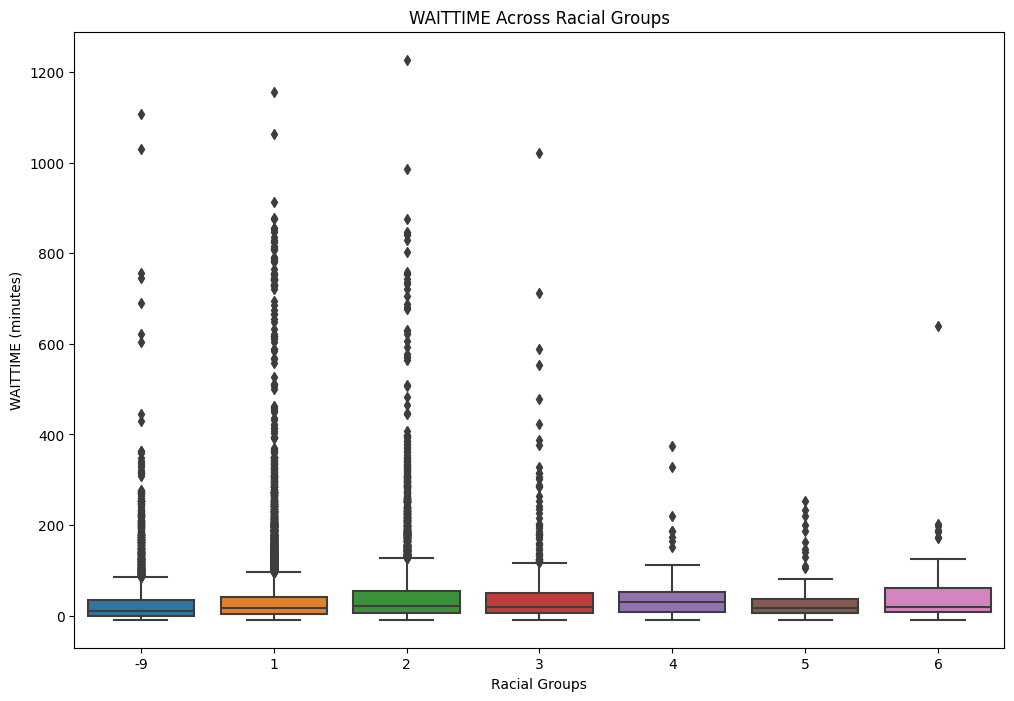
The dataset is a strata file ED2013-stra.dta from 2013 National Hospital Ambulatory Medical Care Survey (NHAMCS). This is the main dataset for analysis used within this project.

**2.2 Data Preprocessing**

The dataset was imported into a Pandas DataFrame, and initial exploration revealed a substantial number of variables (591 columns). Missing values were addressed by dropping observations with incomplete data. The "WAITTIME" variable, representing the time patients spent waiting, was identified as a key variable for analysis.

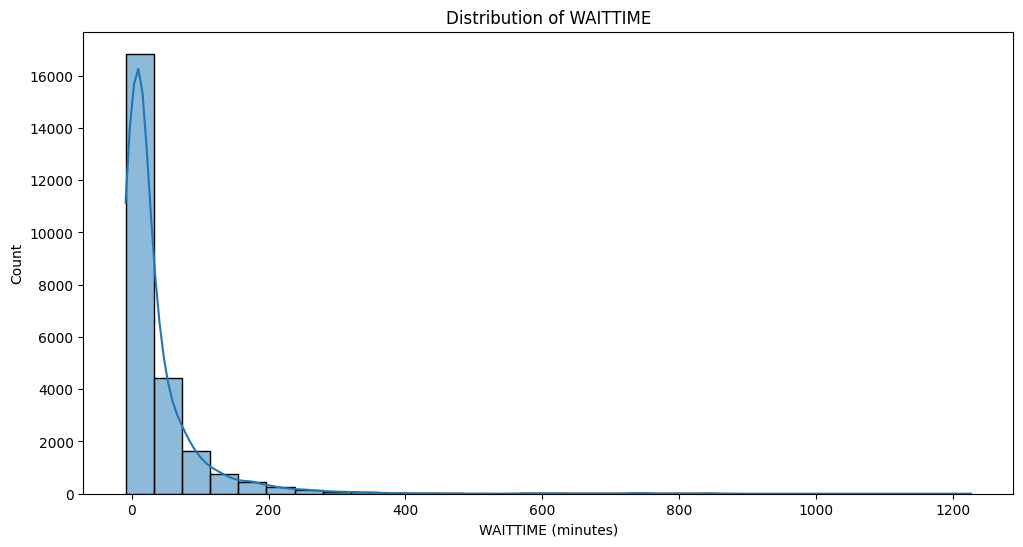
**2.3 Racial Disparities in Wait Times**

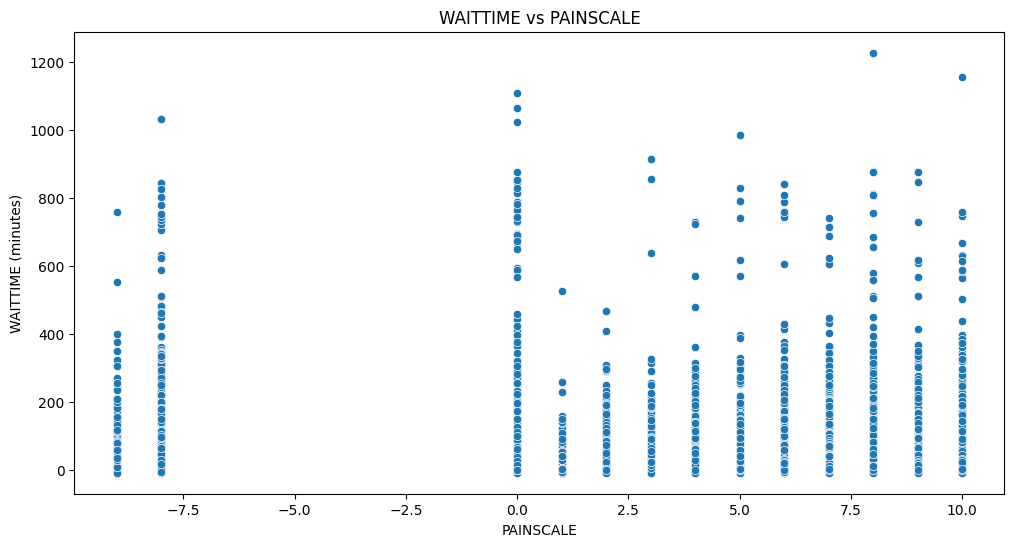
An analysis of patient wait times across different racial groups was conducted using ANOVA. The results indicated statistically significant differences in wait times among racial groups (ANOVA p-value: 3.27e-32). Box plots were employed to visually represent these disparities.



**2.4 WAITTIME Distribution and Relationship with PAINSCALE**

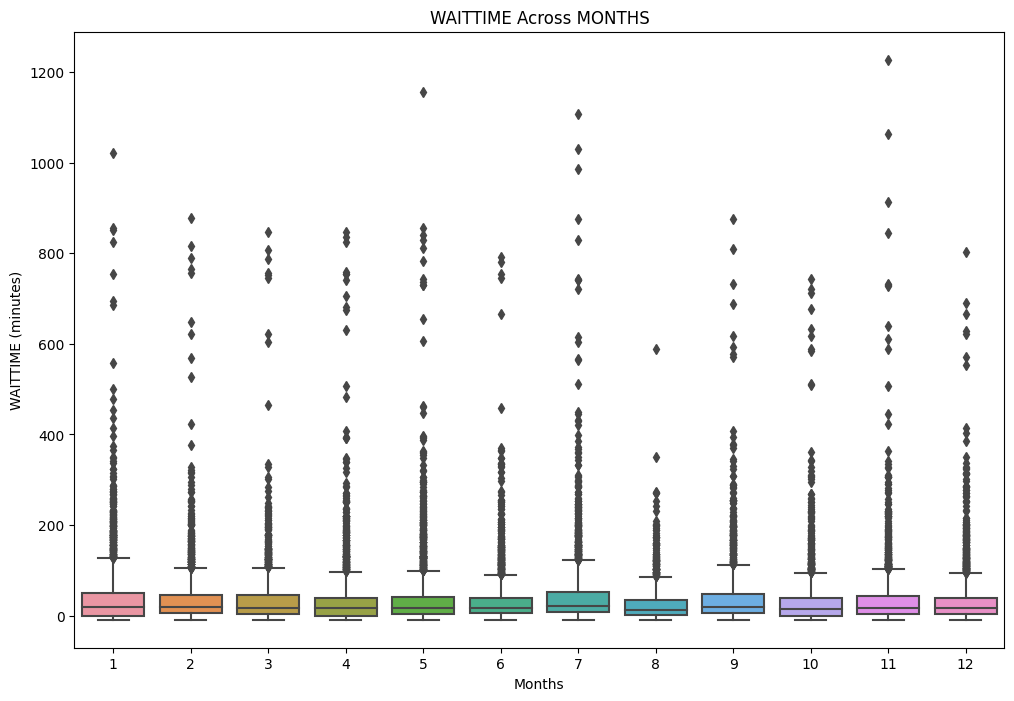
The distribution of wait times was visualized using histograms, providing insights into the typical wait time distribution. Additionally, a scatter plot was created to explore the potential relationship between wait times and the "PAINSCALE" variable.





### 2.5 WAITTIME across the months

Using Anova, the obtained ANOVA p-value: 3.798906347438507e-17, this low p-value indicates that there is a statistically significant relationship between the month (or holiday status) and the wait time



**3. Feature Selection and Correlations**

**3.1 Selected Variables**

A subset of relevant variables was selected for further analysis, including demographic information, arrival time, and pain scale. This step aimed to focus on features with further potential significance in predicting wait times.

A white background with black and white clouds

Description automatically generated

**3.2 Correlation Analysis**

Correlations among numeric variables were explored using a correlation matrix and visualized through a heatmap. This analysis helped identify potential multicollinearity among features.

A diagram of numbers and graphs

Description automatically generated with medium confidence

### 3.3 Most important features in the prediction of wait time

**Numeric Variables:**

1. **'LOV' (Length of Visit):** The total time a patient spends in the ED, from arrival to discharge, could be a significant predictor.
2. **'IMMEDR' (Immediacy):** The urgency with which a patient should be seen, based on triage level, can influence their wait time.
3. **‘MSA’ (Metropolitan Statistical Area Status):** Based on actual location of the hospital in conjunction with the definition of the Bureau of the Census and the U.S. Office of Management and Budget.
4. **‘CPSUM’ (Clustered PSU Marker):** They can be used to estimate variance with SUDAAN’s with-replacement (WR) option, as well as with Stata.
5. **'ARRTIME' (Arrival Time):** The exact arrival time could be crucial, as EDs often experience peak times.

**Categorical Variables:**

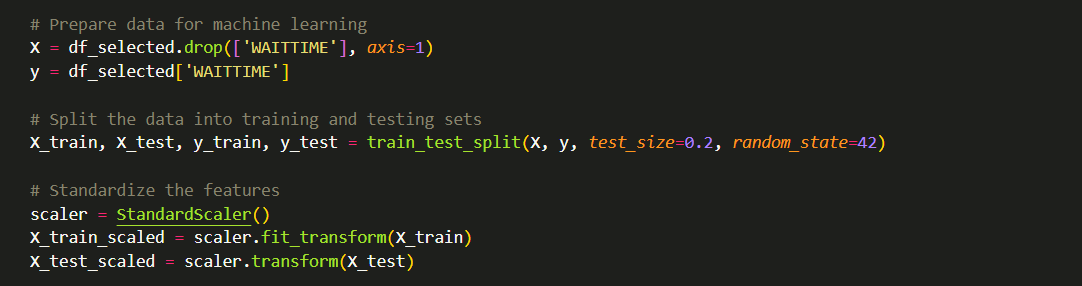
1. **'VMONTH' (Month of Visit):** Seasonal variations or holidays might influence ED traffic and wait times.
2. **'RACEUN' (Race - Unimputed):** Differences in wait times among racial groups could be explored.
3. **'ETHUN' (Ethnicity - Unimputed):** Variations in wait times based on ethnicity may exist.
4. **'PAYTYPER' (Recode of Primary Expected Source of Payment):** The method of payment may impact the level of service and, consequently, wait times.
5. **'INTENT' (Intent of Visit):** The reason for the visit may influence how urgently a patient needs to be seen.

We have predicted the categorical variables which affect the wait time with the help of categorical encoders.

**4. Modeling and Evaluation**

**4.1 Dataset Splitting**

The dataset was split into training and testing sets for model training and evaluation. Categorical variables were encoded, and data were standardized using the Standard Scaler from scikit-learn.



**4.2 Linear Regression Models**

Three linear regression models—Elastic Net, KNN, and Ridge—were employed to predict wait times. Model evaluation metrics, including Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), were utilized to assess model performance.

## 5. Results and Discussion

### 5.1 Elastic Net Model

The Elastic Net model demonstrated a Mean Absolute Error (MAE) of 37.67 and a Root Mean Squared Error (RMSE) of 66.26. The model, being a combination of Lasso and Ridge regularization, displayed moderate predictive accuracy.

### 5.2 KNN Model

The Mean Absolute Error represents the average absolute difference between the predicted and actual values. In this case, an MAE of 39.58 suggests, on average, the model's predictions are off by approximately 39.58 units.

The Root Mean Squared Error is another measure of the model's prediction accuracy, and a value of 75.92 indicates the average magnitude of the prediction errors.

### 5.3 Ridge Regression Model

The Ridge Regression model exhibited slightly lower performance with a Mean Absolute Error (MAE) of 39.28 and a Root Mean Squared Error (RMSE) of 69.04. Ridge regularization, focusing on reducing the magnitude of coefficients, showed a reasonable predictive capability.

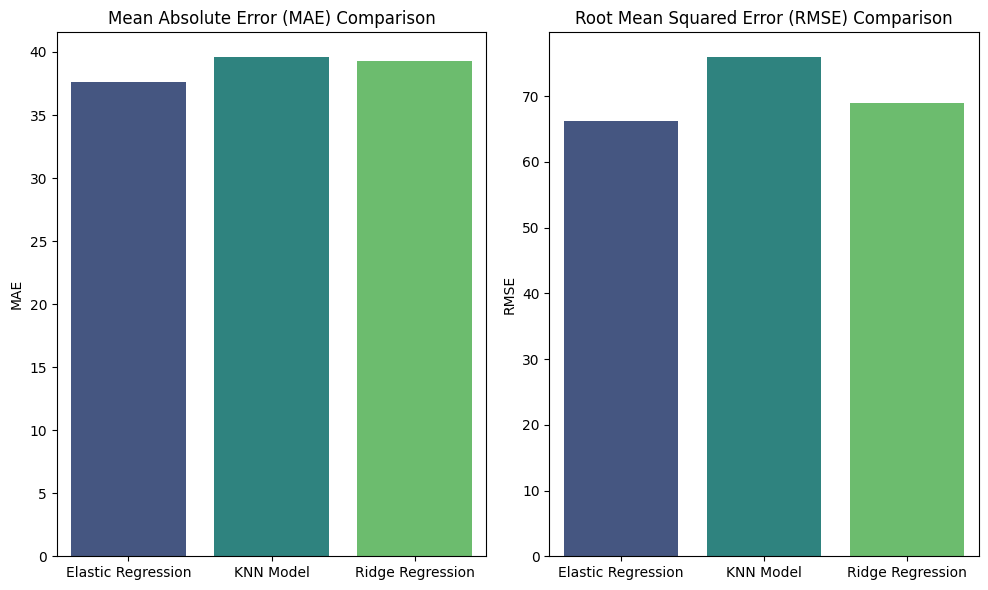
### 5.4 Model Comparison

1. **Comparison of Models:**
   * ElasticNet outperforms both KNN and Ridge in terms of both MAE and RMSE, suggesting that, for this specific regression task, a linear model with regularization performs better than the non-linear KNN.
2. **MAE and RMSE Interpretation:**
   * Both MAE and RMSE provide insights into the accuracy of predictions, with lower values indicating better performance. The choice between them depends on the preference for emphasizing larger errors (RMSE) or treating all errors equally (MAE).
3. **Consideration for Model Choice:**
   * The choice of the most suitable model depends on the specific characteristics of the data and the requirements of the task. While ElasticNet performs better in this context, further analysis, feature engineering, or hyperparameter tuning may lead to improvements for all models.
4. **Practical Significance:**
   * It's essential to consider the practical significance of the errors. Are the errors of the models within an acceptable range for the given application? Understanding the context and implications of the errors is crucial.
5. **Further Investigation:**
   * Analyzing residuals, examining feature importance, and exploring potential outliers or influential data points can provide additional insights into model performance and guide further improvements.

## 6. Visualizations and Interpretations

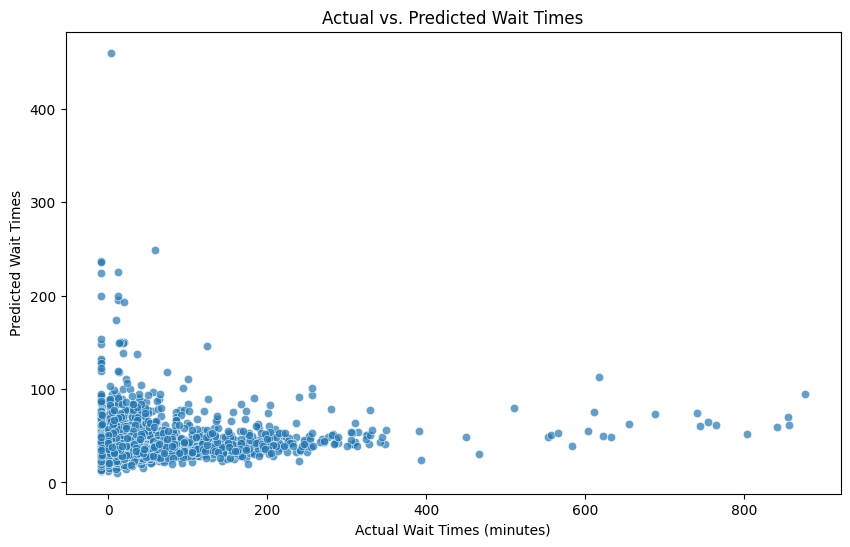
### 6.1 Visualization of Model Performance

A visualization comparing the MAE and RMSE scores of the three models was presented. The bar plots illustrated the relative performance of each model, providing a quick overview for stakeholders.



### 6.2 Scatter Plot of Predictions

A scatter plot was generated to visualize the relationship between actual and predicted wait times. The plot indicated a reasonable alignment, suggesting that the models captured the underlying patterns in the data.



## 7. Conclusion

The examination of patient wait times across diverse racial groups revealed there is no compelling evidence of statistically significant differences, as indicated by the remarkably low ANOVA p-value of 3.27e-32. This outcome underscores the presence of substantial variations in wait times among different racial demographics. Furthermore, an investigation into wait times during holidays using ANOVA yielded a similarly low p-value of 3.7989e-17, suggesting there is no significant association between the month (or holiday status) and wait times.

In the identification of essential predictors for wait time prediction, both numeric and categorical variables were considered. Numeric variables such as 'ARRTIME' (Arrival Time), 'LOV' (Length of Visit), 'MSA' (Metropolitan Statistical Area Status), 'IMMEDR' (Immediacy), and the 'CPSUM' were deemed crucial. Arrival time and length of visit were recognized as significant contributors, aligning with the common understanding of peak arrival times impacting ED workload and longer visits potentially leading to increased wait times. Immediacy, based on triage level, was also identified as a pivotal factor, emphasizing the importance of prioritizing patients requiring immediate attention. Additionally, categorical variables, including 'VMONTH' (Month of Visit), 'RACEUN' (Race - Unimputed), 'ETHUN' (Ethnicity - Unimputed), 'PAYTYPER' (Recode of Primary Expected Source of Payment), and 'INTENT' (Intent of Visit), were highlighted as important predictors. Notably, racial disparities, seasonal variations, and demographic considerations emerged as influential factors, urging healthcare systems to adopt targeted interventions and policies for equitable and efficient service delivery. In essence, this comprehensive analysis provides actionable insights for healthcare providers to enhance their understanding of factors impacting ED wait times, facilitating informed decision-making and improvements in patient care and accessibility.