Shatakshi Shewale

Data Analysis Project Report

1. Introduction:

Injury outcomes and their determinants are pivotal for healthcare research. Understanding the interplay between demographic factors, injury characteristics, and recovery outcomes can provide valuable insights for personalized treatment strategies and resource allocation. This project explores the relationships between variables influencing injury duration, healthcare site preferences, and the role of demographic and clinical factors in predicting outcomes. Using a dataset containing 203 observations on patients with various injury types, this study aims to test hypotheses about key predictors of injury outcomes and examine how they vary across demographic subgroups. The analysis incorporates statistical tests, correlation analyses, and regression models to derive actionable insights.

2. Data Description and Study Objectives:

The dataset comprises 203 observations and 19 variables, capturing demographic details, injury characteristics, care preferences, and patient-reported outcomes. Patients' ages range from 8 to 18 years, with a mean age of 15.12 years. The samples include 61% males and 39% females. Injury types are categorized as falls (41%), sports injuries (20%), and assaults. Most patients (71%) sought emergency care, with 20% utilizing primary care facilities. Recovery times, measured as Injury Duration, range from 0 to 52 months, with a mean duration of 6.1 months. Patient-reported outcomes, represented by Rating2, range from 20 to 100, with an average score of 85.72. Missing values were imputed using medians for numerical variables and modes for categorical ones.

The variables of interest for the analysis include Injury Duration, which serves as the target outcome, and predictors such as Injury Type, Rating2, Sex, Age, and Care Site. Injury Type is explored for its impact on recovery patterns, particularly in interaction with demographic factors such as Age and Sex. Rating2, a patient-reported outcome measure, is analyzed for its predictive strength on recovery duration. Care Site is evaluated to understand patterns in healthcare utilization, particularly for emergency care, and how it correlates with Age. The analysis also investigates associations between Injury Type and demographic variables, such as Age and Sex, to reveal broader trends.

This study seeks to analyze factors influencing Injury Duration, care-seeking behavior, and demographic trends in injury recovery. The hypotheses are as follows:

- H1: Injury Type predicts Injury Duration and is associated with demographic factors such as Age and Sex.
- H2: Rating2 significantly predicts Injury Duration.
- H3: Age influences Care Site choice, with younger age groups more likely to visit emergency care.

The hypotheses are tested independently without sequential dependence. For numerical predictors like Rating2 and Injury Duration, linear models and Pearson correlations are applied. Logistic regression is used to assess patterns in Care Site utilization across age groups. Categorical relationships, such as between Injury Type and Sex, are analyzed using Chi-square or Fisher's exact tests. Interaction effects, such as the combined influence of Injury Type, Age, and Sex on Injury Duration, are evaluated using linear models with interaction terms.

3. Statistical Methods and Key Findings:

I]. To investigate whether Injury Type predicts Injury Duration, moderated by Age and Sex, a linear regression model with interaction terms was applied. The model included main effects for injury type, age, and sex, along with their interactions. ANOVA was conducted to assess the significance of main effects and interactions, ensuring assumptions of normality and equal variances were reasonably met. The analysis revealed that the main effects of injury type, age, and sex were not individually significant. However, the interaction between Injury Type: Sport and Age approached significance ($p = 0.0571$), indicating that the influence of sport-related injuries on recovery duration varies with age. Additionally, the injury type "Sport" had a significant positive effect on injury duration ($p = 0.0452$), suggesting that this category is associated with longer recovery times. The model's overall F-statistic was significant ($p = 0.0072$), explaining 17.7% of the variation in injury duration.

To further elucidate these findings, a visualization was created showing the interaction effects of injury type, age, and sex on injury duration using boxplots stratified by age group and colored by sex. The plot highlights trends such as longer recovery durations for sport injuries in younger individuals and differences in injury impact across genders (Figure 1). Additionally, a table summarizing the regression coefficients provides detailed numerical insights into the main effects and interaction terms (Table 1).

Predictor	Estimate	Std. Error	t-value	p-value
(Intercept)	-4.67	36.96	-0.126	0.8997
Injury type: Fall	33.53	38.32	0.876	0.3828
Injury type: Other	6.78	38.01	0.175	0.8591
Injury type: Sport	107.51	53.32	2.016	0.0452
Injury type: Vehicle	51.45	39.60	1.299	0.1955
Age	0.67	2.37	0.282	0.7786
Sex: Male	-8.63	41.34	-0.209	0.8348
Injury Type: Sport x Age	-6.35	3.32	-1.915	0.0571

Table 1: Regression Coefficients for Hypothesis 1

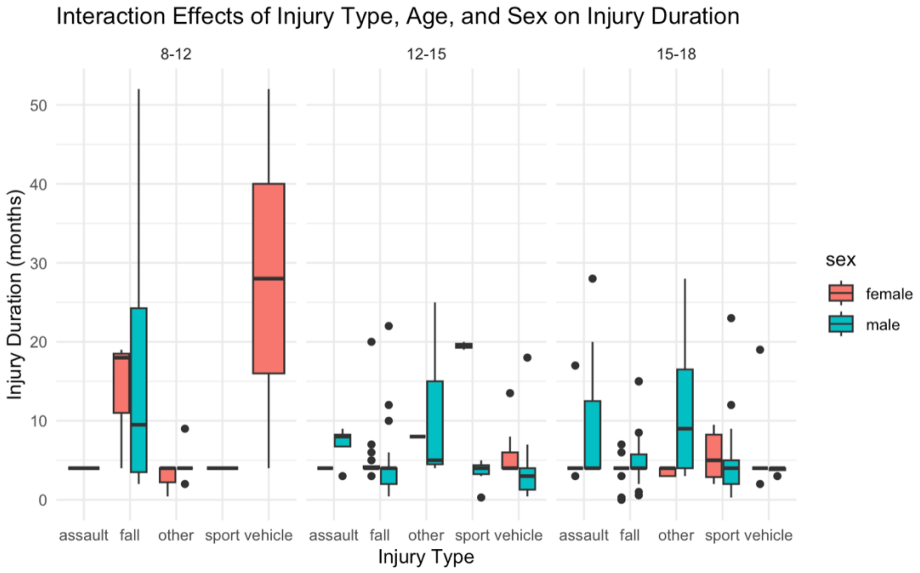

Figure 1: Interaction Effects of Injury Type, Age, and Sex on Injury Duration

II]. To evaluate the hypothesis that Rating2 significantly predicts Injury Duration, a linear regression analysis was conducted with Injury Duration as the dependent variable and Rating2 as the independent variable. The regression model revealed a statistically significant negative relationship between Rating2 and Injury Duration ($p<0.0001$). The coefficient for Rating2 (-0.1219) indicates that for every one-unit increase in Rating2, the recovery time decreases by approximately 0.12 months. This finding was further supported by the Pearson correlation analysis ($r=-0.27$, $p < 0.0001$), which confirmed a modest inverse relationship between these variables.

To validate the regression model, diagnostic plots were examined (Figure 3). These plots demonstrated no major violations of assumptions such as linearity or homoscedasticity, reinforcing the reliability of the results. A scatterplot (Figure 2) with a fitted regression line illustrates the negative association between Rating2 and Injury Duration, providing a clear visual representation of the trend. The findings highlight the utility of Rating2 as a meaningful predictor of recovery times, underscoring its potential to inform personalized treatment strategies and enhance patient care.

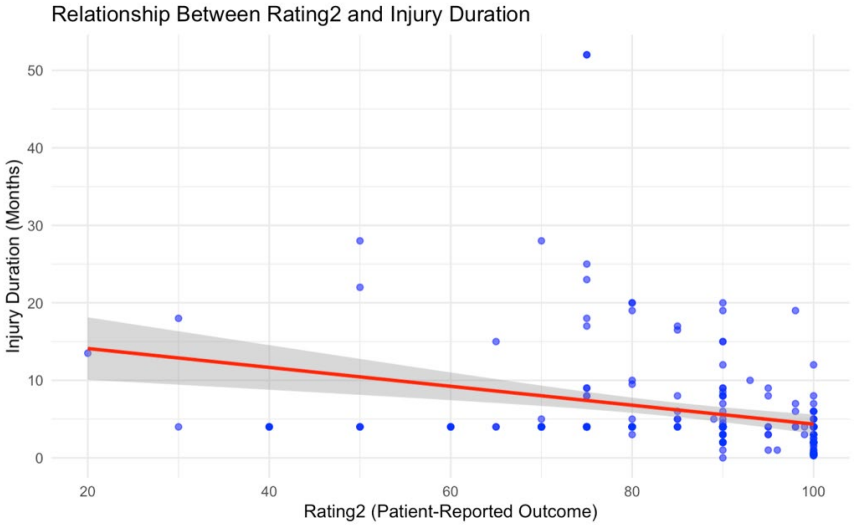

Figure 2: Relationship Between Rating2 and Injury Duration

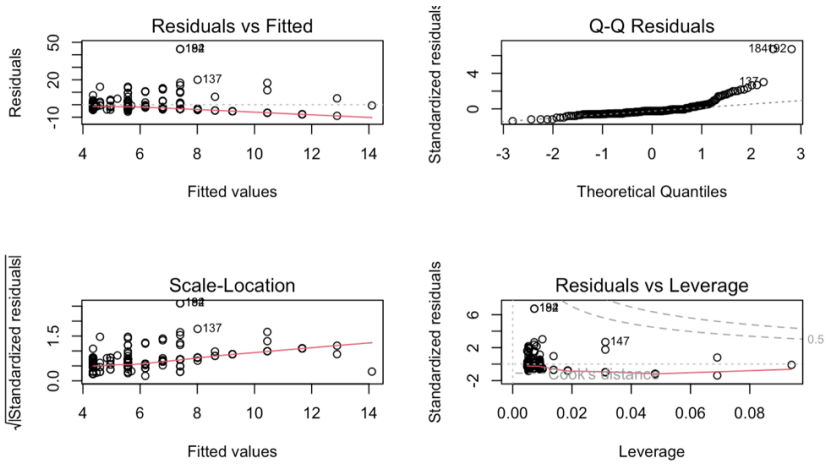

Figure 3: Diagnostic Plots for Model Validation

III]. To investigate the hypothesis that younger age groups are more likely to visit emergency care, a logistic regression model was applied. The Care Site variable was recoded into a binary outcome, with Emergency Care coded as 1 and all other care sites (Primary Care, Other) coded as 0. Age was categorized into three groups: 8-12, 12-15, and 15-18 years, with the youngest group serving as the reference category. The logistic regression model evaluated the likelihood of emergency care visits based on these age groups.

The analysis revealed a significant effect for the youngest age group (8-12) compared to the 15-18 group, where older adolescents were less likely to visit emergency care (Odds Ratio = 0.29, 95% CI: 0.0647–0.9369, p = 0.0607). The 12-15 group, however, showed no significant difference compared to the 8-12 group (Odds Ratio = 0.92, p = 0.9088). The Area Under the Curve (AUC) from the ROC analysis was 0.639, indicating modest predictive performance of the model. Figure 4 illustrates the proportional distribution of emergency care visits across the three age groups, highlighting a declining trend in emergency care utilization with increasing age. The ROC curve (Figure 5) further confirms the model's predictive capability. These findings suggest that younger individuals are more likely to visit emergency care, reflecting potential differences in injury severity, caregiver behavior, or healthcare-seeking attitudes.

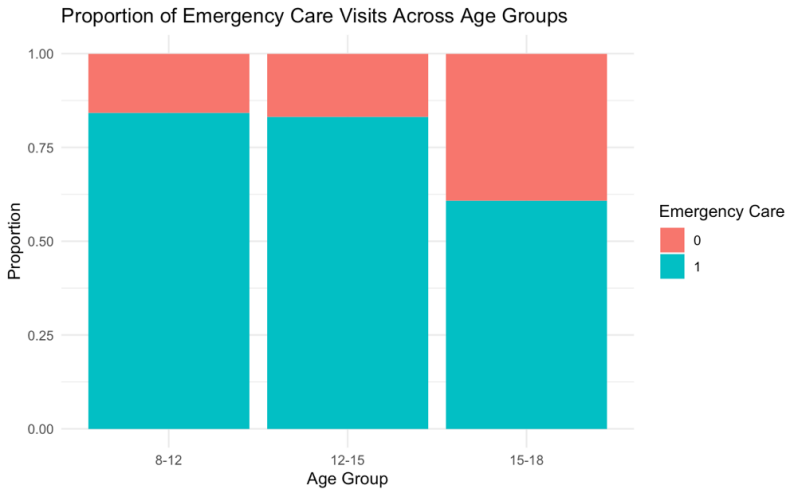

Figure 4: Proportion of Emergency Care Visits Across Age Groups

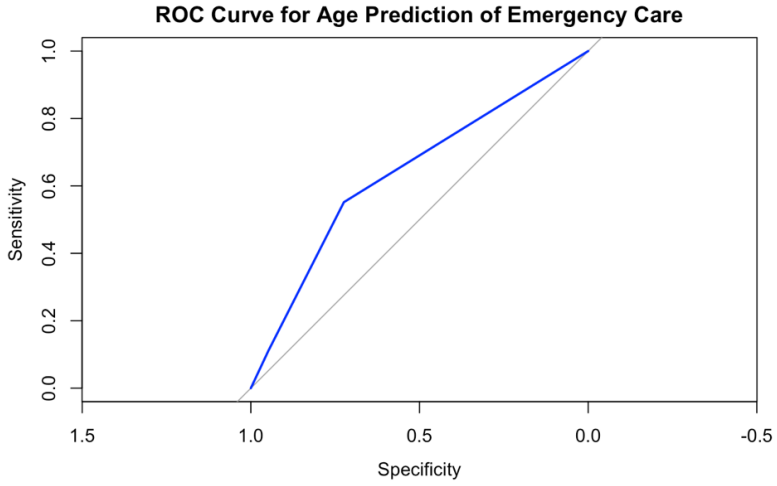

Figure 5: Model Performance (ROC Curve)

4. Discussion:

The findings from this analysis highlight significant relationships among injury type, demographic factors, patient-reported outcomes, and healthcare utilization. These results provide valuable insights into understanding recovery patterns and improving resource allocation in healthcare systems.

The interaction between injury type, age, and sex demonstrates the complexity of predicting injury duration. Specifically, sports injuries show a significant positive association with longer recovery times ($p = 0.0452$), while the interaction between sports injuries and age approached significance ($p = 0.0571$). This suggests that the impact of sports injuries on recovery is particularly pronounced in younger age groups. Additionally, the visualizations reveal trends such as longer recovery times for females in specific injury categories, underscoring the need for demographic-specific interventions. Tailored recovery plans that address the unique needs of age and gender subgroups could enhance treatment efficacy and optimize outcomes (Figure 1, Table 1).

The analysis of Rating2 as a predictor of injury duration further emphasizes the importance of patient-reported outcomes. The significant negative association ($p < 0.0001$) and correlation ($r = -0.27$, $p < 0.0001$) suggest that higher Rating2 scores are robustly associated with shorter recovery periods. These results advocate for the inclusion of subjective triage measures, like Rating2, in routine clinical assessments to guide personalized treatment strategies. The scatterplot illustrates this inverse relationship, while diagnostic plots confirm the reliability of the regression model, ensuring that the assumptions of linearity and homoscedasticity are met (Figures 2 and 3).

Finally, the logistic regression model investigating emergency care utilization highlights critical differences across age groups. Younger age groups (8-12 years) were significantly more likely to seek emergency care compared to older adolescents (15-18 years, $p = 0.0607$), with the 12-15 group showing no significant difference. The ROC analysis, with an AUC of 0.639, indicates modest predictive performance but provides actionable insights into age-specific healthcare utilization trends. Visualizations reveal a declining trend in emergency care usage with increasing age, reflecting potential differences in injury severity, caregiver behavior, or healthcare-seeking attitudes (Figures 4 and 5). These results highlight the need for targeted outreach and resource allocation strategies to address the specific needs of younger populations in emergency care settings.

In summary, these findings underscore the importance of integrating demographic, clinical, and patient-reported data to understand injury outcomes and healthcare utilization. They pave the way for tailored interventions and highlight areas for future research, such as exploring additional variables and developing more complex models to improve predictive accuracy.

5. Acknowledgments:

I would like to sincerely thank my professor for their invaluable guidance and support throughout the course of this project. Their expertise and thoughtful feedback greatly enhanced the depth and quality of my analysis. I am also grateful to the creators of the dataset, whose comprehensive data enabled meaningful exploration and insights into injury outcomes. Additionally, I wish to acknowledge the contributions of my peers, whose discussions inspired fresh perspectives and ideas. I also express my gratitude to the developers of R and its associated packages, which played a crucial role in facilitating the analysis and visualization for this study.

6. References:

[1]. R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Available at: https://www.R-project.org/
[2]. Wickham, H., & Bryan, J. (2019). *R Packages*. Available at: https://r-pkgs.org
[3]. Field, A. (2013). *Discovering Statistics Using R*. SAGE Publications.
[4]. Broom: Convert statistical analysis objects into tidy data frames. (2023). Available at: https://cran.r-project.org/web/packages/broom/index.html
[5]. ggplot2: Elegant Graphics for Data Analysis. (2023). Available at: https://ggplot2.tidyverse.org