GROUP 14

DEMOGRAPHIC BEHAVIORAL DATA

DATA SCIENCE FOR THE HEALTH SCIENCES

INTRODUCTION

Describing participants' demographics (e.g., ethnicity/race, socioeconomic status, gender/sex, age) is important for understanding their relationship with behavior-analytic procedures.(Jones et al., 2020).

This study explores the dataset **“2_Demographic_Behavioral_data_Group_014.csv"**, which contains data on patient demographics (region, Socioeconomic, Education, Smoking status, Drinking status), body composition (weight, height, and BMI). Physical Activity, Patient Satisfaction Score and Health literacy score.

Analyzing the dataset, we aim to use exploratory data analysis on a dataset that includes demographic and behavioral information, aiming to gain insight into patient characteristics, lifestyle choices, and their possible relationships. The analysis emphasizes data cleaning, summary statistics, and the visualization of distributions and connections among important demographic, behavioral, and health-related factors, with the ultimate goal of uncovering patterns that could help inform public health and patient care strategies.

METHODS

The dataset was first imported into the R environment and cleaned using the tidyverse and skimr packages.

Key steps included:

Data Cleaning: Using colSums is.na() was used for checking missing values. Specific columns containing unwanted information were removed using select (), rename_with() for renaming these column names for easier manipulation.

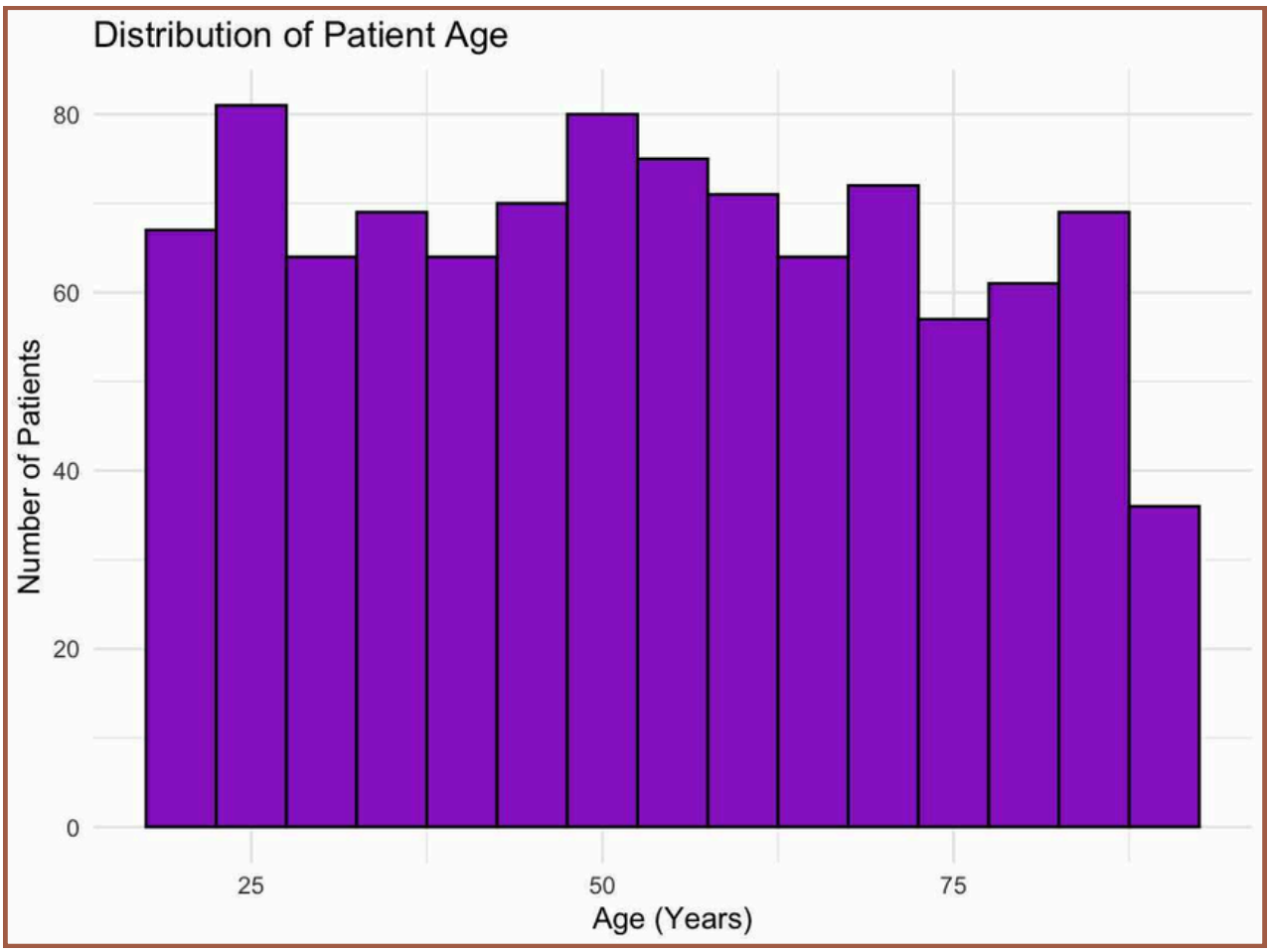
Type Conversion: The categorial variables (sex, region, socioeconomic, education, smoking status, drinking status,patient satisfaction score and health literacy score) are converted to factor variables. While other variables (age, weight,height,BMI, and Physical activity hours/week) are converted to numeric types for quantitative analysis.

Descriptive Statistics: Summary metrics including mean, median, quartiles and standard deviation were computed for all numeric variables using summary () and skim().

Visual Analysis: A variety of plots were created using ggplot2, histograms, boxplots, scatter plots, bar charts to explore trends in Age, Region, Patient Satisfaction Score, BMI, Smoking Status, Physical Activity Hours/Week and Health literacy Score.

Statistical test: A conditional Chi-square test was conducted to examine the relationship between smoking status and drinking status.

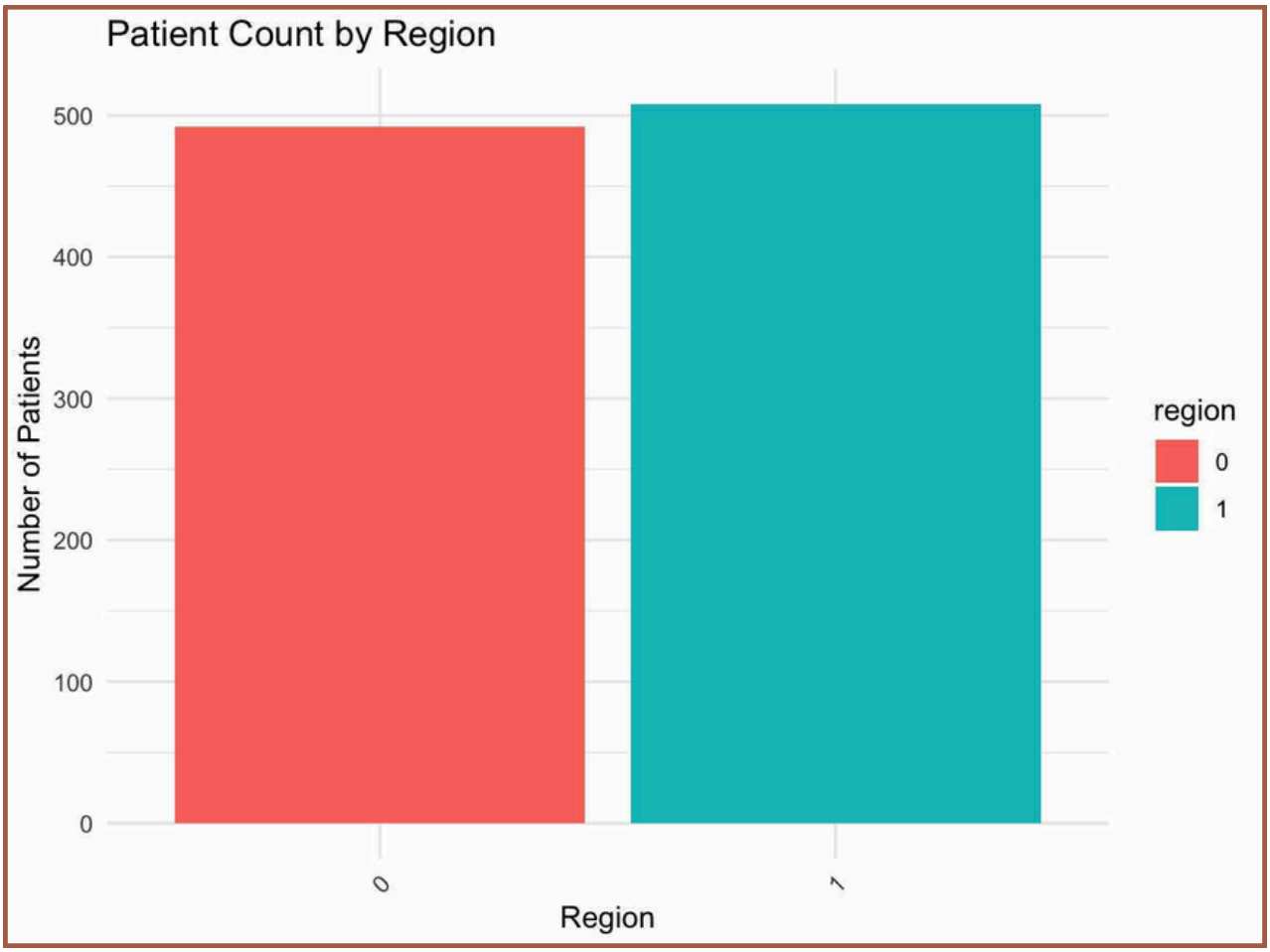
RESULTS AND FIGURES



Distribution of Patient Age

The histogram shows that 25 years old dominates the age group with the highest number of patients. The bar corresponding to the 25 years on the x-axis reaches a height of just over 80 patients.

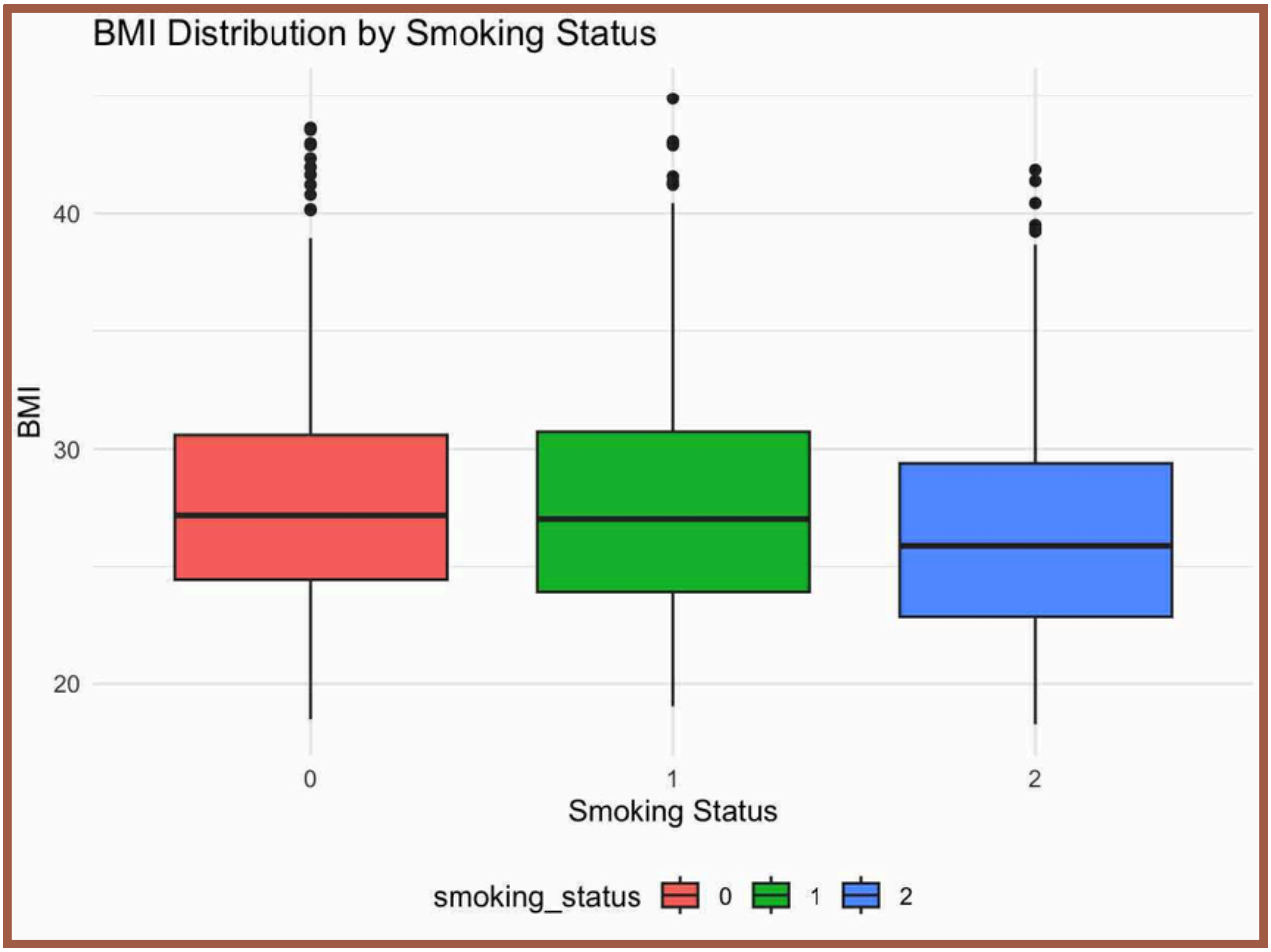
RESULTS AND FIGURES



Patient Count by Region

The bar plot indicates that Region 0 has approximately 490 patients less than Region 1 having approximately 500 patients having a slightly higher number of patients than Region 0.

RESULTS AND FIGURES



BMI Distribution by Smoking Status

The boxplot shows the comparison of the Body Mass Index (BMI) across three different smoking statuses. The smoking status is labeled as Red Box = “0”, Green Box = “1”, Blue Box = “2”.

KEY FINDINGS

BMI Distribution by Smoking Status

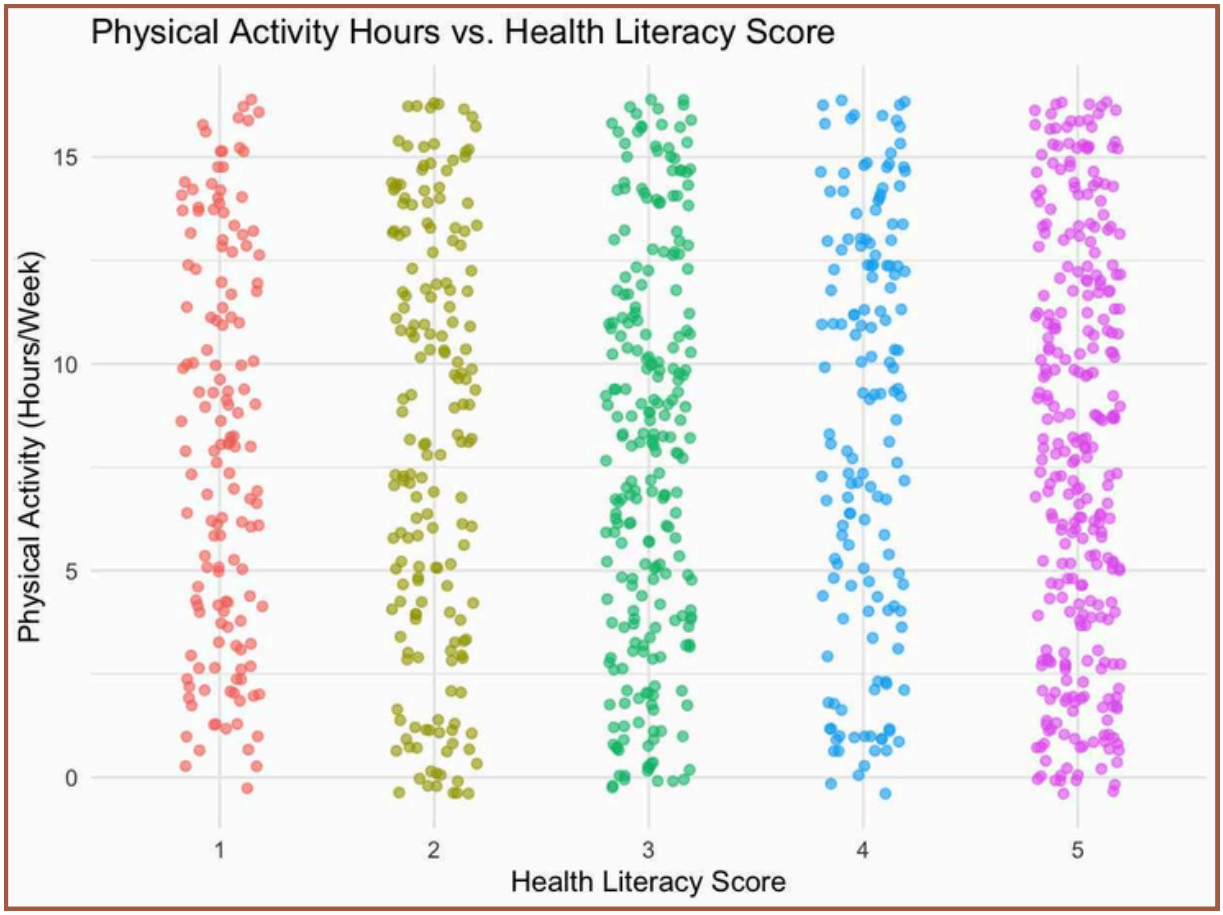
Smoking Status 0 has several outliers based on the black dots above the upper whisker. With the median BMI is 26.5. Its interquartile range (IQR) shows 50% of individuals in this group have a BMI between 24 and 31

Smoking Status 1 also has several outliers similar to Status 0. The median BMI is slightly lower around 26. The IQR is approximately 23 and 31.

Smoking Status 2 has the lowest median BMI approximately 24.5. The IQR is between 22 and 29.5.There are also outliers with higher BMI up to around 42.

The three groups show outliers with high BMIs, indicating that some individuals have much higher body mass regardless of smoking status. This information helps in understanding BMI differences related to smoking, which can be important for health assessments and targeted interventions.

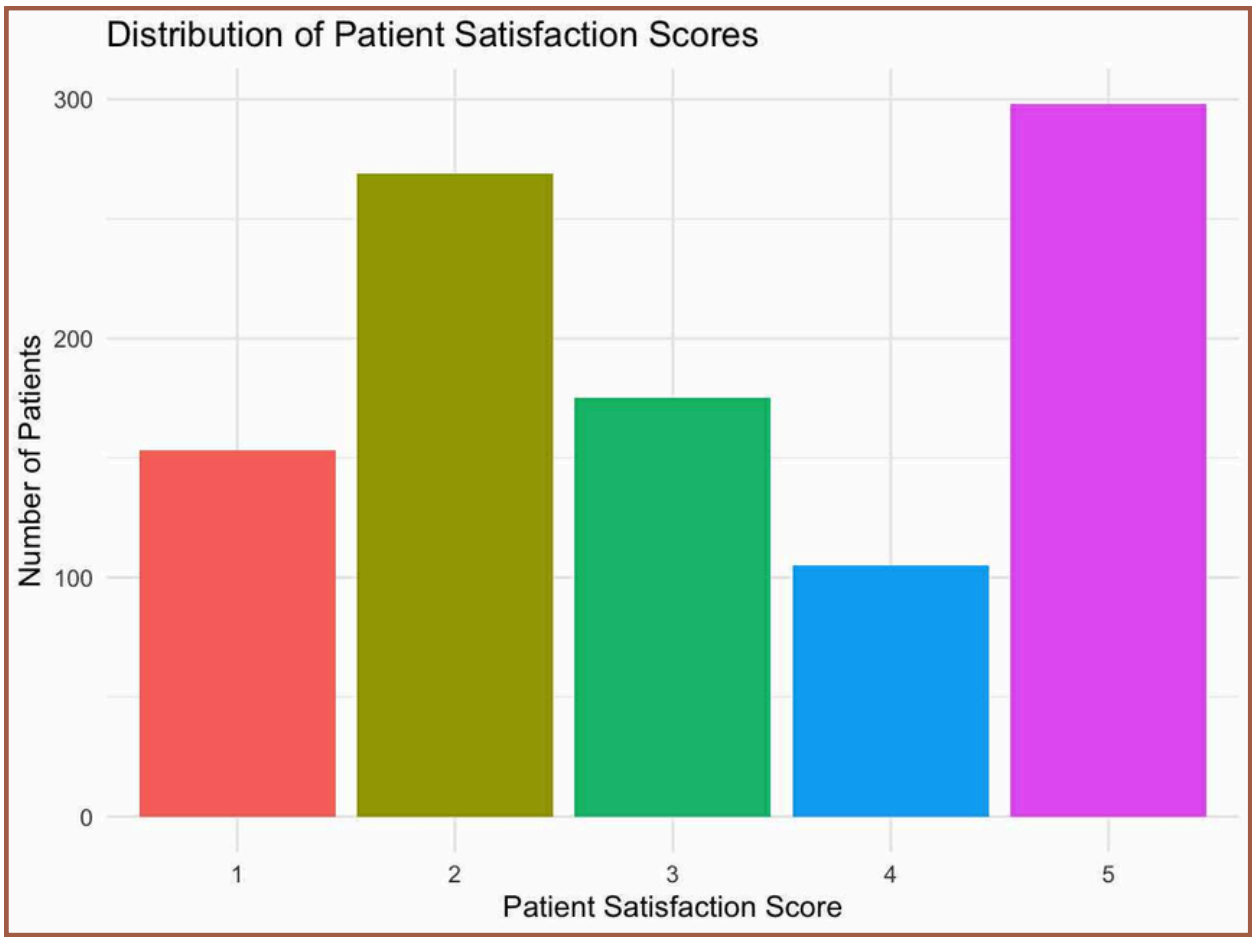
RESULTS AND FIGURES



Distribution of Physical Activity Hours vs. Health Literacy Score

The scatter plot shows there is no appearance positive or negative linear correlation between Health Literacy Score and Physical Activity because of the cluster points for each score are relatively spread out vertically. Based on the jittered points, it shows there's a denser concentration of patients in the middle range of physical activity.

RESULTS AND FIGURES



Distribution of Patient Satisfaction Scores

This barplot shows the number of patients for each satisfaction score that is ranging from 1 to 5.The distribution is skewed towards higher satisfaction, with the highest number of patients giving a perfect score of 5 and a significant portion giving a score of 2. This indicates an overall positive trend in patient satisfaction, although there remain distinct groups with lower ratings.

CONCLUSION

This analysis provides a comprehensive overview of patient demographics, health indicators, and satisfaction. The age distribution highlights that 25-year-olds form the largest patient group. Geographically, Region 1 has a slightly higher patient count than Region 0.

Examining health metrics, BMI distribution by smoking status reveals interesting patterns: while all three smoking status groups have outliers with high BMIs, Smoking Status 2 shows the lowest median BMI. The analysis of physical activity hours versus health literacy scores indicates no clear linear correlation, with patient concentration denser in the middle range of physical activity. Finally, the distribution of patient satisfaction scores leans positively, with a large number of patients reporting a perfect score of 5, alongside a notable group giving a score of 2

REFERENCE

Reference: Jones, S. H., St Peter, C. C., & Ruckle, M. M. (2020). Reporting of demographic variables in theJournal of Applied Behavior Analysis. Journal of Applied Behavior Analysis, 53(3), 1304–1315. https://doi.org/10.1002/jaba.722