Hospital Emergency Room Efficiency Data Analysis

Author: Sana Siddiqui, MD

Introduction

Hospital emergency rooms (ERs) are fast paced, high turnover environments that demand maximum efficiency to ensure timely patient care. In order to consistently meet this goal, patient surveys are an excellent tool to gain insight into patient satisfaction with the care provided in ERs. In order to better understand how efficient our current hospital ER system is, we present an in-depth analysis utilizing patient satisfaction surveys. Our goal is to understand what factors surrounding patient care logistics can be improved to elevate patient satisfaction and overall experience when visiting the ER.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(knitr)
library(skimr)
```

Data Overview

The dataset consists of ER visit records with key variables:

- **patient_waittime**: Time (in minutes) patients waited before receiving care
- **patient_sat_score**: Patient satisfaction scores on a scale from 0 (lowest) to 10 (highest)
- **date**: Date of patient visit
- **patient_race**: Patient race category
- **patient_age**: Patient age in years

Methods and Analyses

Data Loading and Cleaning

The dataset was cleaned in a separate R script: `Scripts/Data_Cleaning.R`.

Cleaning steps included:

- Removing duplicate rows
- Standardizing column names
- Converting timestamps to proper date-time format
- Imputing missing patient satisfaction scores with the median

```
# Load cleaned data
df = read.csv('Data/Cleaned/Hospital_ER_Cleaned.csv')

# Convert date column to date object
df$date = as_datetime(df$date)
```

```
# Quick data visualization after loading cleaned data
kable(head(df))
```

date	patient_id	patient_gender	patient_age	patient_sat_score	patient_id	patient_first_initial	patient_last_name	patient_race	patient_admit_flag	patient_wait_time	department_referral
2020-03-20 08:47:01	145-39-5406	M	69	10	H	Glasspool	White	false	39	None
2020-06-15 11:29:36	316-34-3057	M	4	5	X	Methuen	Native American/Alaska Native	true	27	None
2020-06-20 09:13:13	897-46-3852	F	56	9	P	Schubuser	African American	true	55	General Practice
2020-02-04 22:34:29	358-31-9711	F	24	8	U	Titcombe	Native American/Alaska Native	true	31	General Practice
2020-09-04 17:48:27	289-26-0537	M	5	5	Y	Gionettitti	African American	false	10	Orthopedics

date	patient_id	patient_gender	patient_age	patient_satisfaction	patient_first_initial	patient_last_name	patient_race	patient_admitted	patient_wait_time	department_referral
2019-04-20 00:13:05	255-51-2877	M	58	5	H	Buff	Asian	false	59	None

Exploratory Data Analysis

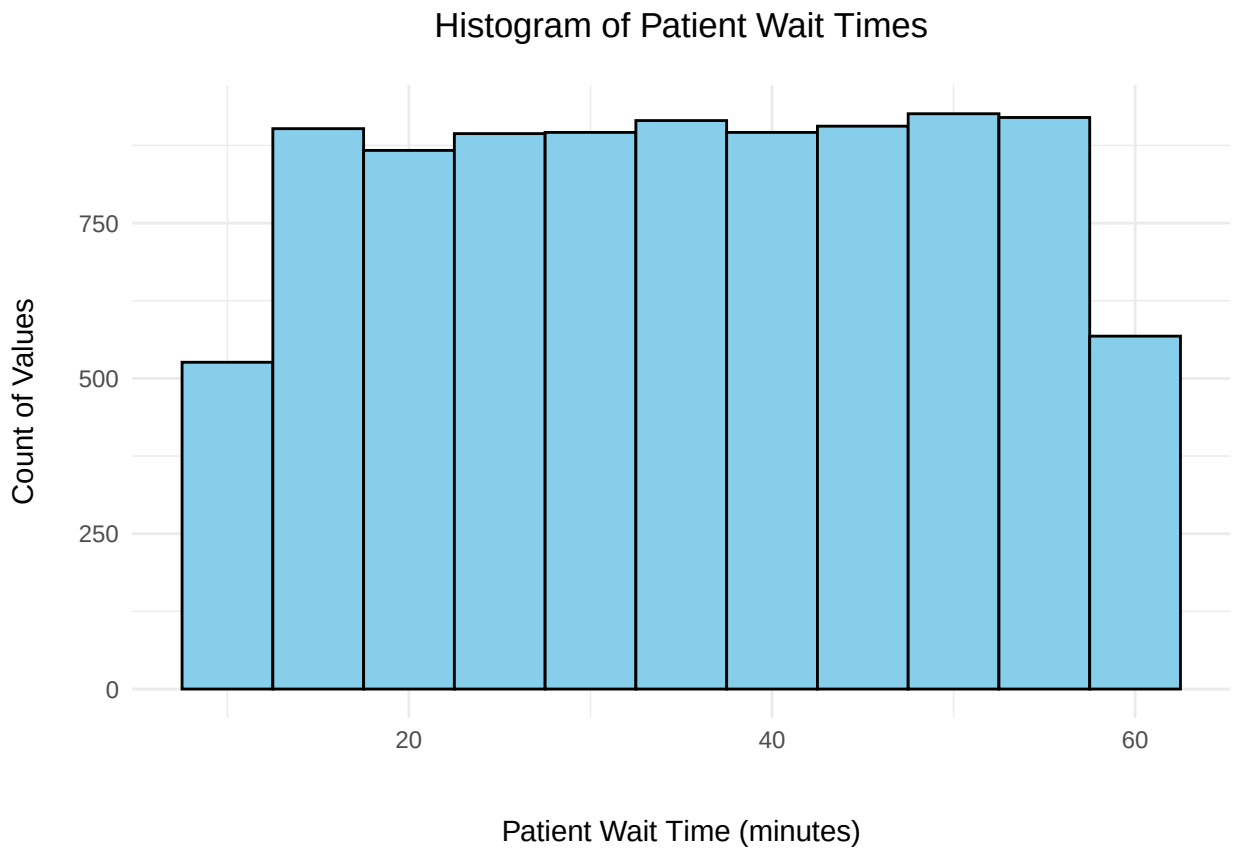
Descriptive Statistics

Patient Wait Time

The following is a histogram of patient wait times

Each bin represents a 5 minute interval

```
ggplot(df, aes(x = patient_waittime)) +
  geom_histogram(binwidth = 5, fill = "skyblue", color = "black") +
  theme_minimal() +
  labs(
    title = "Histogram of Patient Wait Times",
    x = "Patient Wait Time (minutes)",
    y = "Count of Values"
  ) +
  theme(
    plot.title = element_text(hjust = 0.5, margin = margin(b = 15)),
    axis.title.x = element_text(margin = margin(t = 25)) ,
    axis.title.y = element_text(margin = margin(r = 15))
  )
```




```
mean_wait = round(mean(df$patient_waittime, na.rm = TRUE), 2)
median_wait = round(median(df$patient_waittime, na.rm = TRUE), 2)
```

The average patient wait time is 35.26 mins.

The median patient wait time is 35 mins.

Key Takeaway:

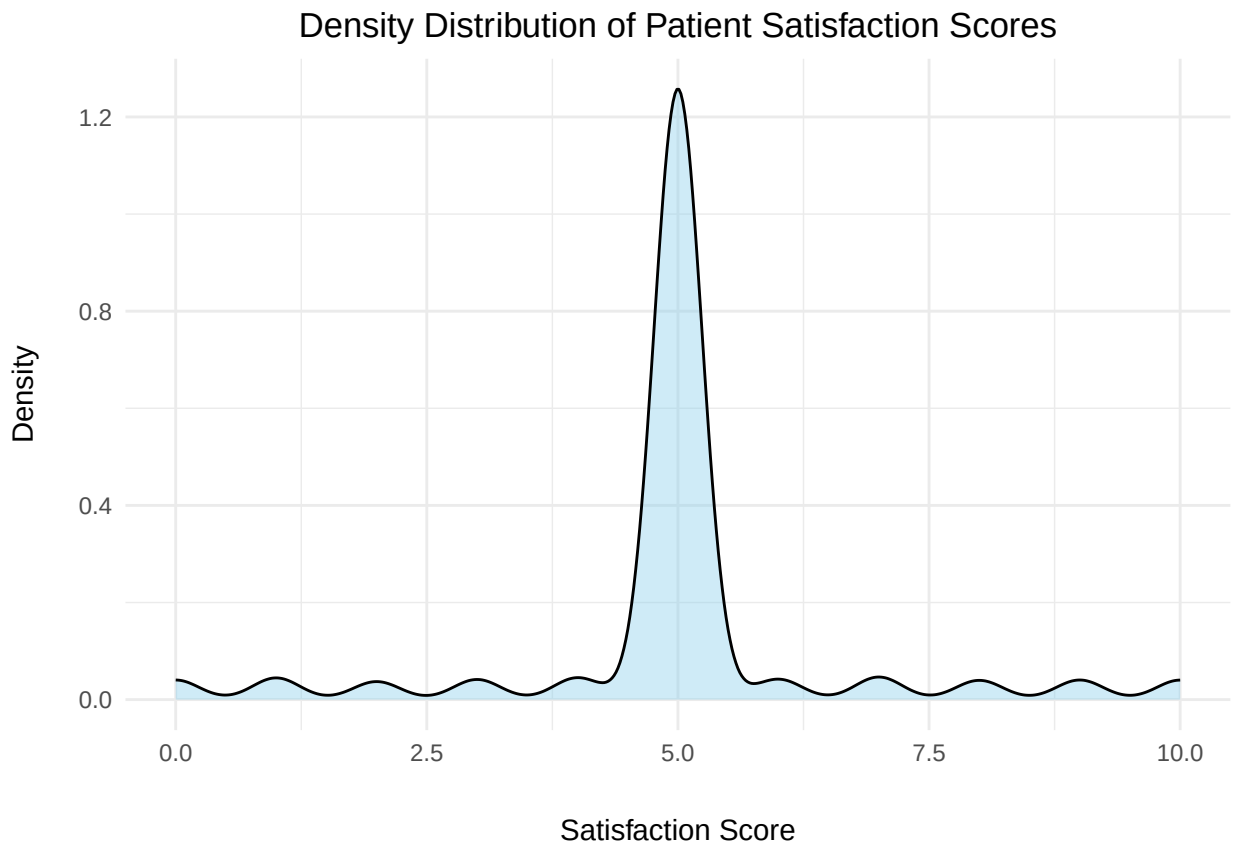
The distribution of patient wait times appears approximately normal. The mean wait time is 35.25 minutes and the median is 35 minutes, indicating that the distribution is largely symmetric. Although the peak of the histogram occurs just left of the mean, the surrounding bars are of similar height, suggesting that the skew is not pronounced

Patient Satisfaction Score

The following is a kernel density estimate plot that depicts the spread of data across the satisfaction scale

```
ggplot(df, aes(x = patient_sat_score)) +
  geom_density(fill = "skyblue", alpha = 0.4) +
  theme_minimal() +
  labs(title = "Density Distribution of Patient Satisfaction Scores",
       x = "Satisfaction Score",
       y = "Density") +
  theme(
    plot.title = element_text(hjust = 0.5),
    axis.title.x = element_text(margin = margin(t = 20)),
```

```
    axis.title.y = element_text(margin = margin(r = 15))
  )
```



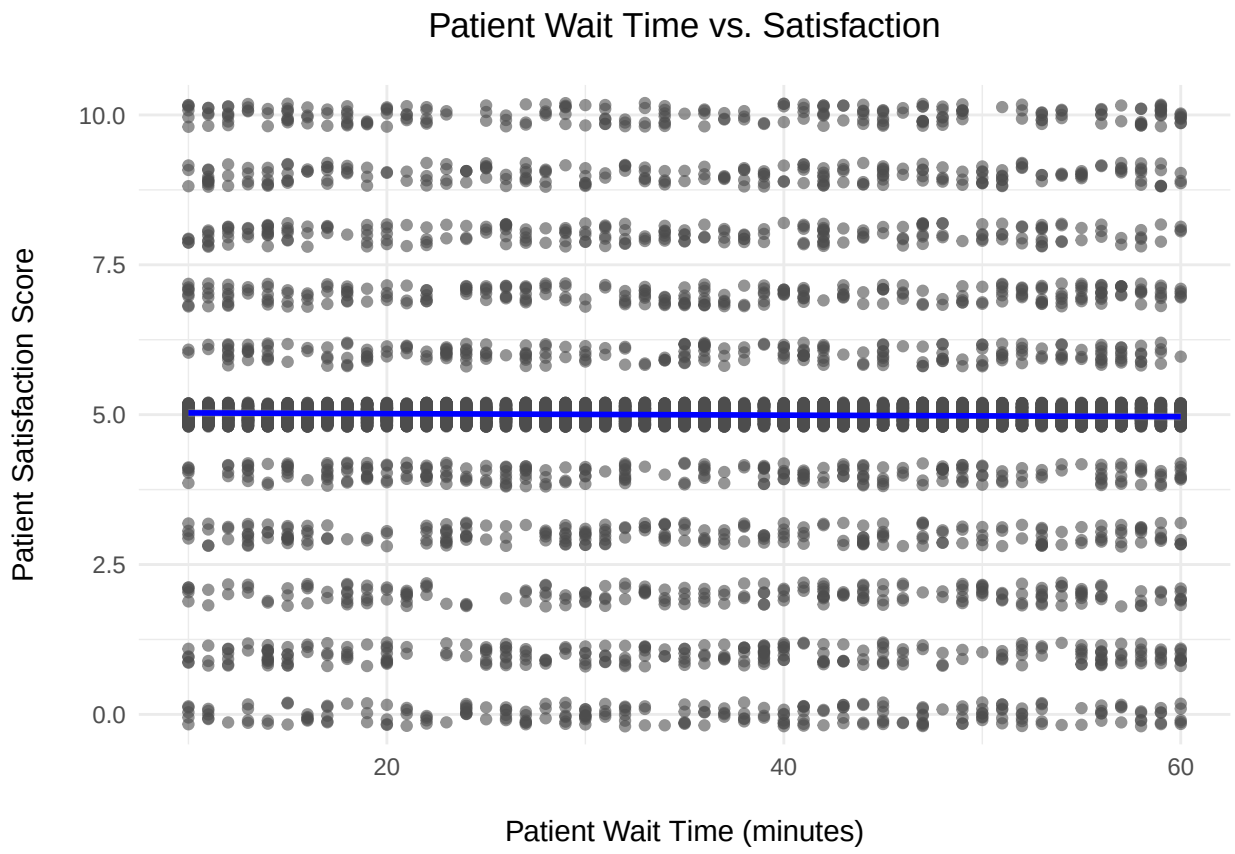
The density plot shows a sharp peak near satisfaction score of 5, reaching a density of ~1.2. While this might seem high, the height of the density curve reflects how concentrated the data is at that point, and not the raw probability. The total area under the curve remains 1.

However, an important note to keep in mind is that we have imputed the median value for all missing values in the satisfaction score column so we expect majority values to cluster around the median. In order to do a more insightful analysis, we would need all raw scores to avoid needing to impute missing values.

Wait Time vs. Satisfaction Score Relationship

```
ggplot(df, aes(x = patient_waittime, y = patient_sat_score)) +
  geom_jitter(width = 0, height = 0.2, alpha = 0.6, color = "gray30") +
  geom_smooth(method = "lm", se = TRUE, color = "blue") +
  coord_cartesian(ylim = c(0,10)) +
  theme_minimal() +
  labs(title = "Patient Wait Time vs. Satisfaction",
       x = "Patient Wait Time (minutes)",
       y = "Patient Satisfaction Score") +
  theme(plot.title = element_text(hjust = 0.5, margin = margin(b = 15)),
        axis.title.x = element_text(margin = margin(t = 15)),
        axis.title.y = element_text(margin = margin(r = 15)))
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
corr = round(cor(df$patient_waittime, df$patient_sat_score), 2)
```

Patient wait time and satisfaction score have a correlation of -0.01. This negative correlation aligns with intuition — as wait times increase, patient satisfaction tends to decrease. However, the magnitude of the correlation is extremely small (-0.01), indicating a very weak inverse relationship. This suggests that while wait time may play a role in shaping satisfaction, other factors are likely to have a much stronger influence on patient satisfaction scores.

Group Analysis

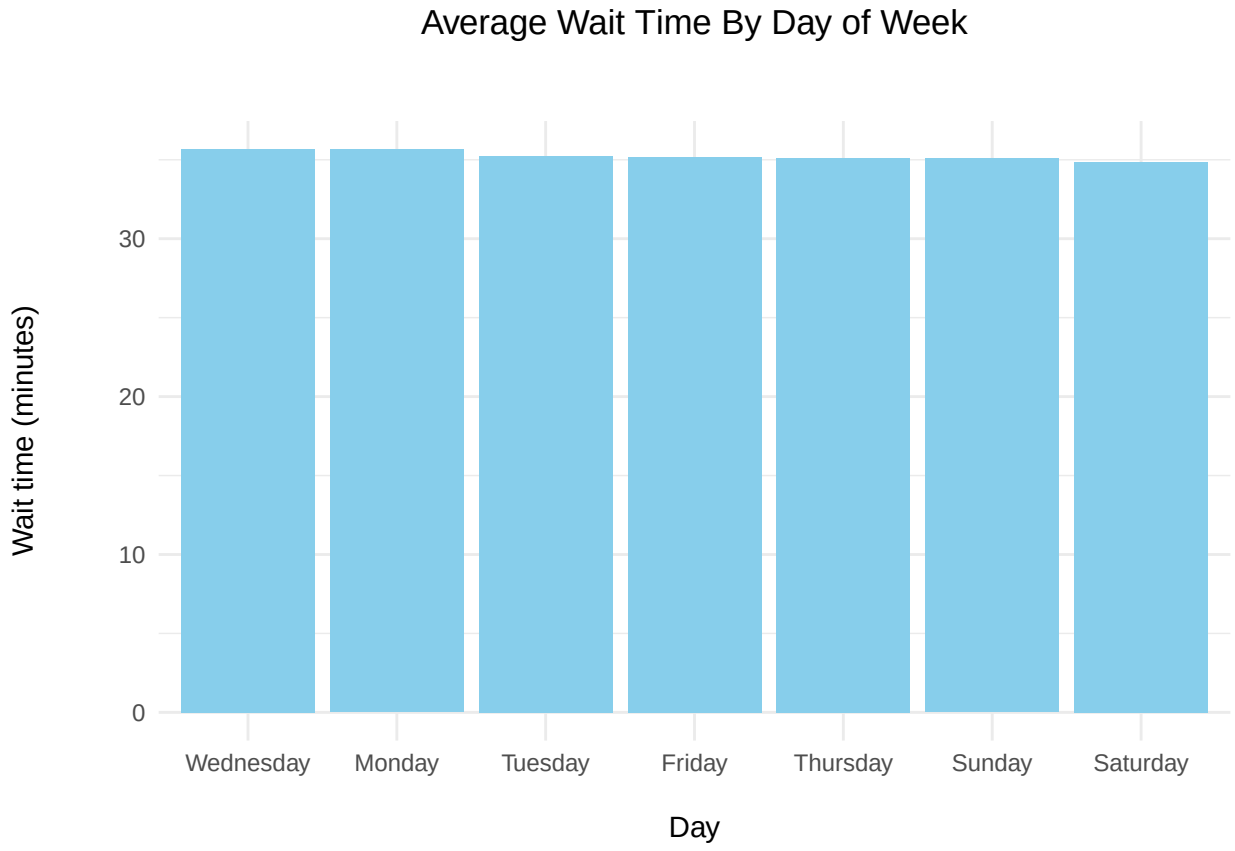
Average Patient Wait Time By Day of the Week

```
weekday_summary = df %>%
  mutate(date = as_datetime(date),
         weekday = weekdays(date)) %>%
  group_by(weekday) %>%
  summarise(avg_wait = round(mean(patient_waittime, na.rm = TRUE), 2)) %>%
  arrange(avg_wait)
kable(weekday_summary)
```

weekday	avg_wait
Saturday	34.88
Sunday	35.08

weekday	avg_wait
Thursday	35.09
Friday	35.19
Tuesday	35.25
Monday	35.65
Wednesday	35.67

```
ggplot(weekday_summary, aes(x = reorder(weekday, -avg_wait), y = avg_wait)) +
  geom_col(fill = "skyblue") +
  theme_minimal() +
  labs(title = "Average Wait Time By Day of Week", x = "Day", y = "Wait time (minutes)") +
  theme(
    plot.title = element_text(hjust = 0.5, margin = margin(b = 30)),
    axis.title.x = element_text(margin = margin(t = 15)),
    axis.title.y = element_text(margin = margin(r = 30))
  )
```

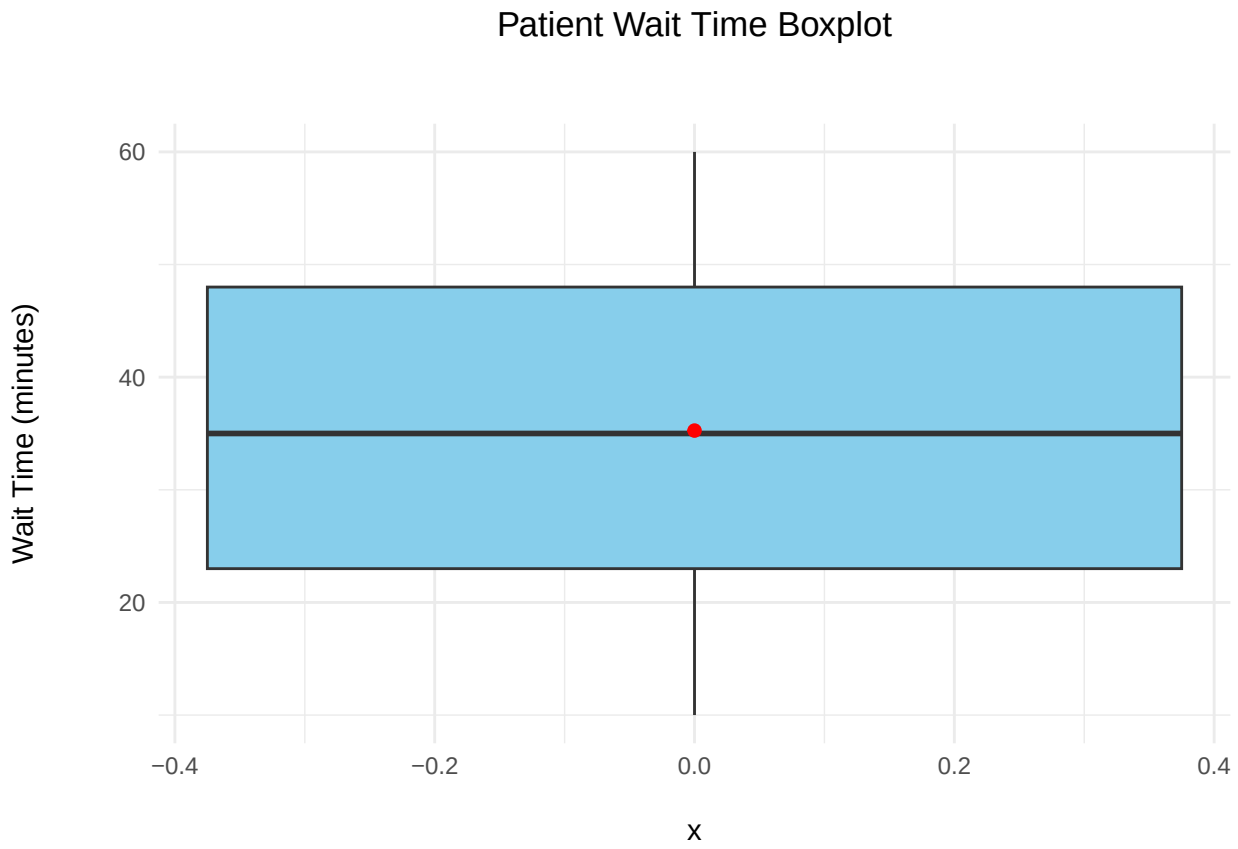


Outlier Detection

Outliers are a critical part of data analysis. If an outlier exists in the data it warrants further investigation because outliers significantly impact summary statistics of the data. For example, an outlier value that is far to the right from all other data points will significantly inflate the mean, standard deviation, and variance of the data. Therefore, it is important to further investigate if it is a true outlier or the result of human error via data entry.

Below we use a box plot to analyze patient wait time to detect any potential outliers. We are particularly interested in detecting any unusually high wait times.

```
ggplot(df, aes(x = 0, y = patient_waittime )) +
  geom_boxplot(fill = "skyblue") +
  stat_summary(fun = mean, geom = "point", shape = 20, size = 3, color = "red") +
  theme_minimal() +
  labs(title = "Patient Wait Time Boxplot",
       y = "Wait Time (minutes)") +
  theme(
    plot.title = element_text(hjust = 0.5, margin = margin(b = 30)),
    axis.title.x = element_text(margin = margin(t = 15)),
    axis.title.y = element_text(margin = margin(r = 30))
  )
```



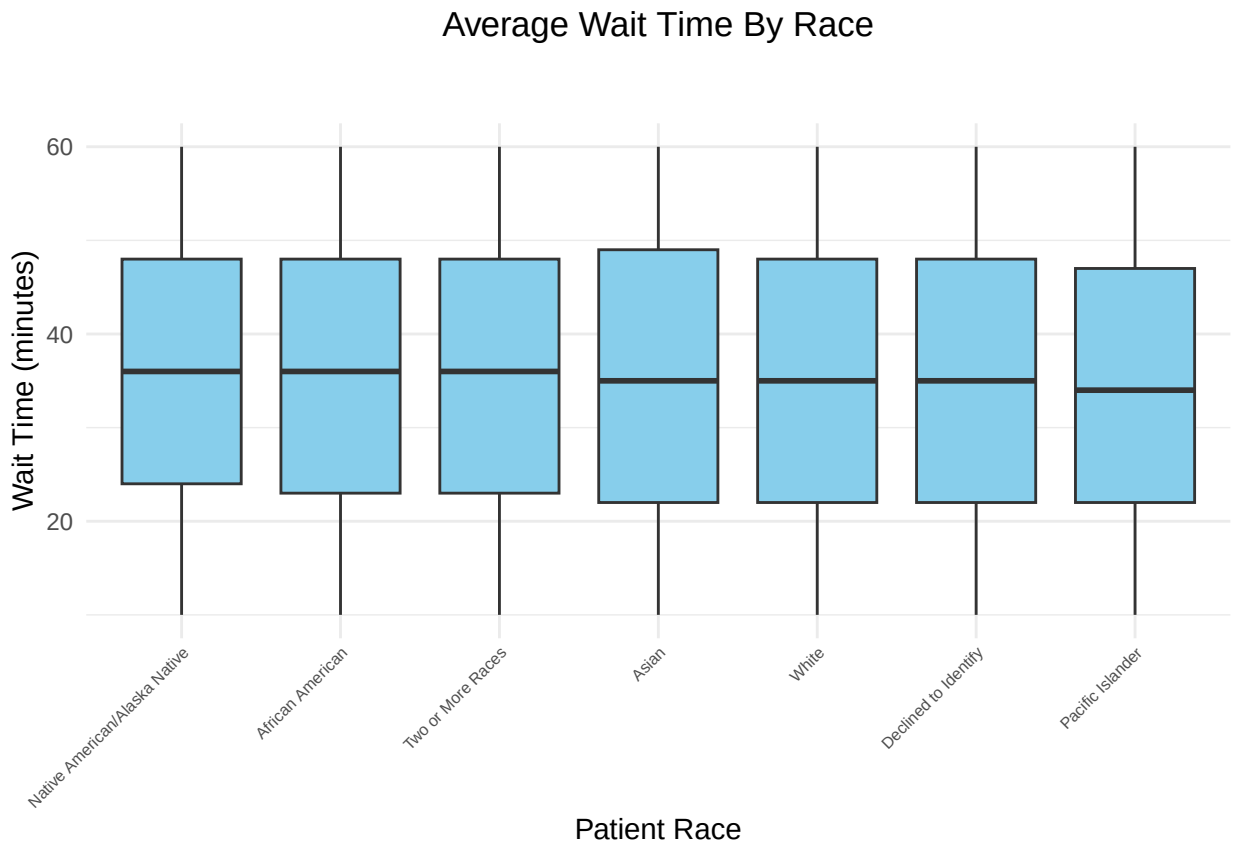
The boxplot above shows there are no outliers in our dataset.

Patient Wait Time By Race

```
df %>%
  group_by(patient_race) %>%
  summarise(avg_wait = round(mean(patient_waittime, na.rm = TRUE), 2)) %>%
  arrange(desc(avg_wait)) %>%
  kable()
```

patient_race	avg_wait
Native American/Alaska Native	35.69
African American	35.61
Two or More Races	35.36
Asian	35.27
White	35.11
Declined to Identify	34.94
Pacific Islander	34.65

```
ggplot(df, aes(x = reorder(patient_race, -patient_waittime, FUN = mean),
               y = patient_waittime)) +
  geom_boxplot(fill = "skyblue") +
  theme_minimal() +
  labs(title = "Average Wait Time By Race",
       x = "Patient Race",
       y = "Wait Time (minutes)") +
  theme(
    plot.title = element_text(hjust = 0.5, margin = margin(b = 30)),
    axis.text.x = element_text(angle = 45, hjust = 1, size = 6, margin = margin(t = 0)))
```

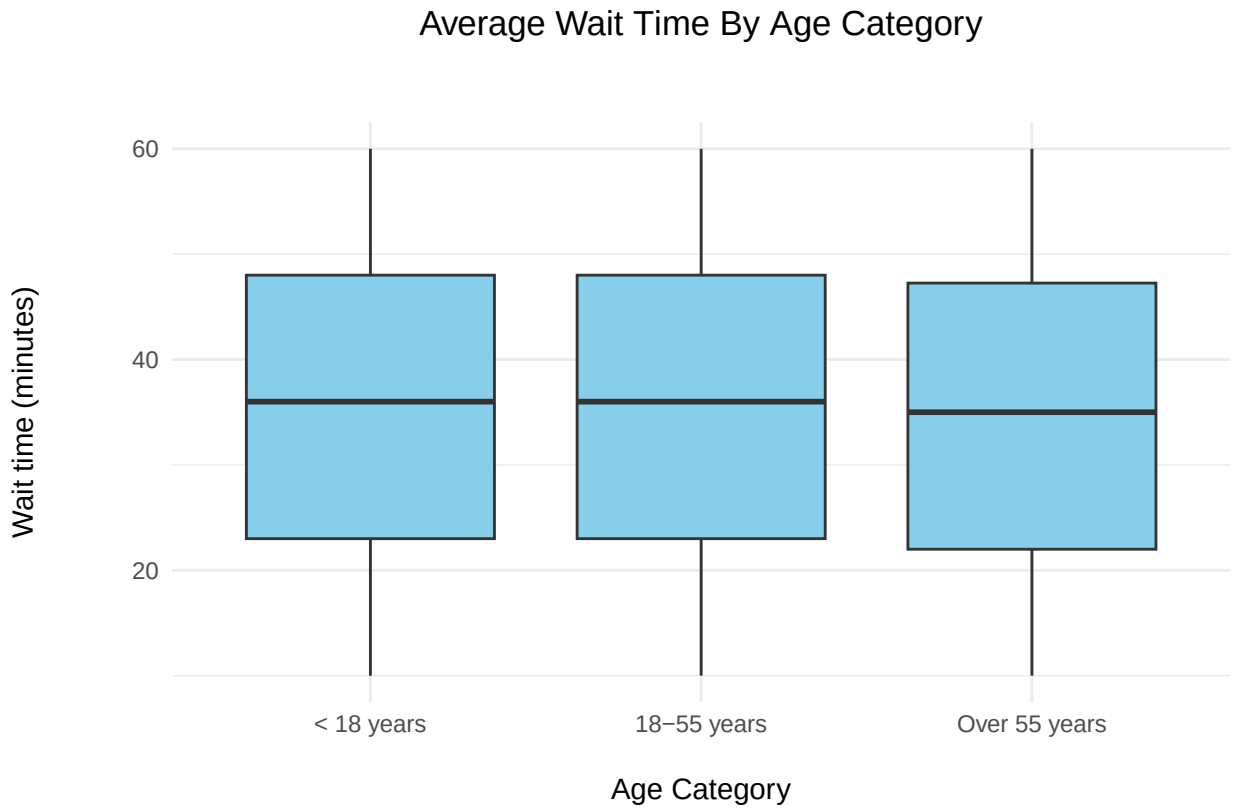


Patient Wait Time by Age

```
df = df %>%
  mutate(age_category = case_when(
```

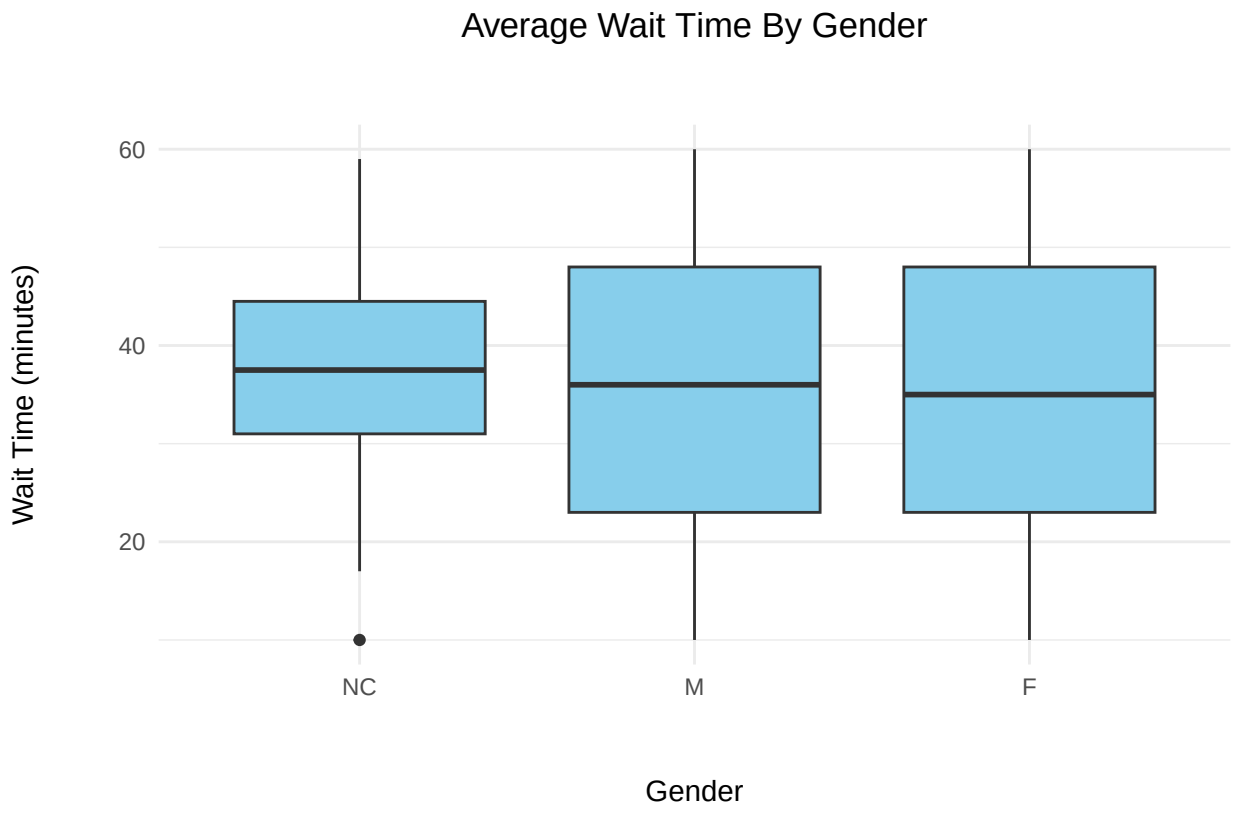
```
    patient_age < 18 ~ "< 18 years",
    patient_age >= 18 & patient_age <= 55 ~ "18-55 years",
    TRUE ~ "Over 55 years"
  ))

ggplot(df, aes(x = age_category, y = patient_waittime)) +
  geom_boxplot(fill = "skyblue") +
  theme_minimal() +
  labs(title = "Average Wait Time By Age Category", x = "Age Category", y = "Wait time (minutes)") +
  theme(
    plot.title = element_text(hjust = 0.5, margin = margin(b = 30, t = 15)),
    axis.text.x = element_text(hjust = 0.5),
    axis.title.x = element_text(margin = margin(t = 15)),
    axis.title.y = element_text(margin = margin(r = 35)))
```



Patient Wait Time By Gender

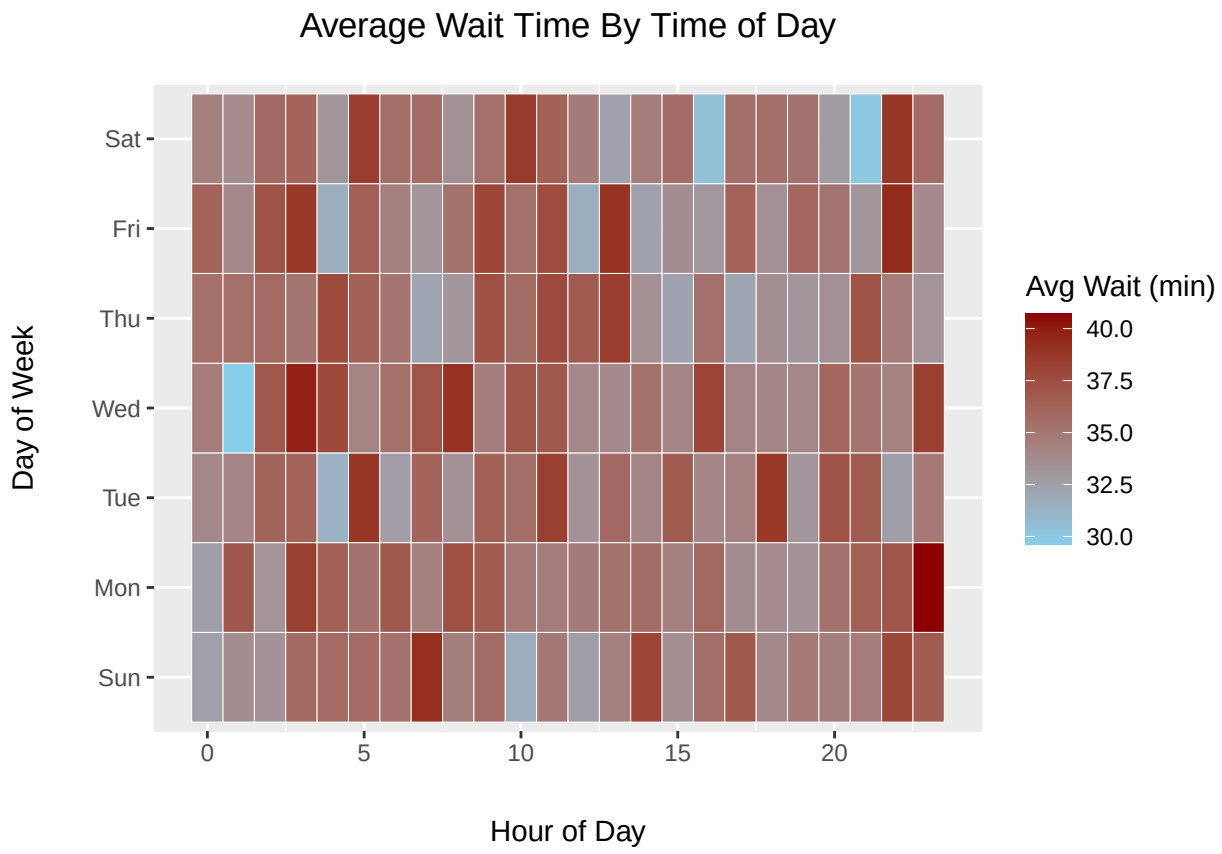
```
ggplot(df, aes(x = reorder(patient_gender, -patient_waittime, FUN = mean), y = patient_waittime)) +
  geom_boxplot(fill = "skyblue") +
  theme_minimal() +
  labs(title = "Average Wait Time By Gender", x = "Gender", y = "Wait Time (minutes)") +
  theme(plot.title = element_text(hjust = 0.5, margin = margin(b = 30, t =15)),
        axis.title.x = element_text(margin = margin(t = 30)),
        axis.title.y = element_text(margin = margin(r = 30)))
```



Patient Wait Time by Time of Day

```
# isolate hour and day of week for each record
df_heatmap = df %>%
  mutate(
    hour = as.integer( hour(df$date)),
    weekday = wday(df$date, label = TRUE, abbr = TRUE)
  ) %>%
  group_by(weekday, hour) %>%
  summarise(avg_wait = mean(patient_waittime, na.rm = TRUE), .groups = "drop")

ggplot(df_heatmap, aes(x = hour, y = weekday, fill = avg_wait)) +
  geom_tile(color = "white") +
  scale_fill_gradient(low = "skyblue", high = "darkred", name = "Avg Wait (min)") +
  labs(title = "Average Wait Time By Time of Day",
       x = "Hour of Day",
       y = "Day of Week") +
  theme(plot.title = element_text(hjust = 0.5, margin = margin(b = 15)),
        axis.title.x = element_text(margin = margin(t = 20)),
        axis.title.y = element_text(margin = margin(r = 20)))
```



The heatmap above visualizes average patient wait times across different hours of the day and days of the week. The x-axis represents the hour of the day (0–23), while the y-axis shows the day of the week. Each tile is color-coded to reflect the average wait time for that specific hour–weekday combination, with darker shades indicating longer wait times.

Notable patterns emerge from the visualization:

Peak wait times tend to occur during evening through early morning, 6p-7a on most days of the week

Slightly shorter wait times are generally seen during evening hours of 4-8p on Thursday, Friday, and Saturday.

These patterns may reflect operational factors such as staffing, patient inflow patterns, or seasonal trends.

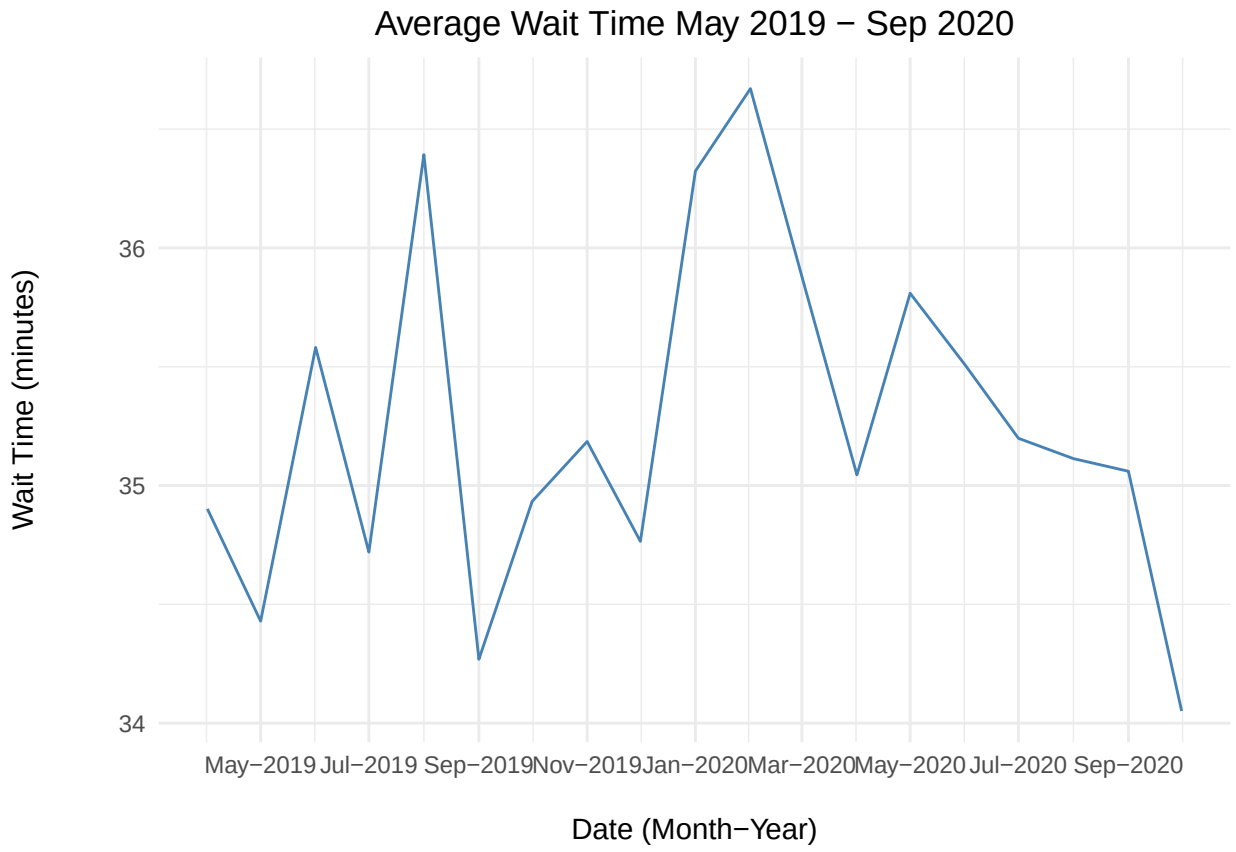
This visualization helps identify temporal bottlenecks and can guide staffing or resource planning decisions in the emergency department.

Average Wait Time Fluctuations Over Time

```
time_summary = df %>%
  mutate(month = as.Date(paste0(format(date, "%Y-%m"),"-01"))) %>%
  group_by(month) %>%
  summarise(avg_wait = mean(patient_waittime, na.rm = TRUE))

ggplot(time_summary, aes(x = month, y = avg_wait)) +
  geom_line(color = "steelblue") +
  scale_x_date(date_breaks = "2 month", date_labels = "%b-%Y") +
  theme_minimal() +
  labs(title = "Average Wait Time May 2019 - Sep 2020", x = "Date (Month-Year)", y = "Wait Time (minutes
  theme(
```

```
    plot.title = element_text(hjust = 0.5),
    axis.title.y = element_text(margin = margin(r = 30)),
    axis.title.x = element_text(margin = margin(t = 15))
)
```



Group Analysis Key Findings

- Wait times show an approximately normal, unimodal pattern: Most patients waited around 35 minutes, but a small subset experienced significantly longer waits.
- Patient satisfaction is centered around 5, with limited variation, possibly influenced by median imputation.
- Small but notable differences in wait times by race with Native Americans waiting the longest indicate areas for potential equity improvement.
- No substantial variation in wait times by weekday, by age category, or by gender was observed.

Time Series Analysis Key Findings

- Higher average wait times were observed during the summer months and peak winter months (January to March)
- This seasonal trend aligns with expected patterns:
  - Summer months often see a rise in pediatric ER visits due to recreational injuries

  - Winter months typically bring an increase in respiratory illnesses, such as common colds, asthma exacerbations, and other viral infections

- It's important to note that this dataset reflects data collected during the first year of the COVID-19 pandemic. The global health crisis and operational uncertainty likely contributed to elevated wait times across all months

Limitations of Data

- Median imputation for missing satisfaction scores likely compressed variability and may bias insights
- The dataset lacks some contextual variables such as severity of illness or staffing levels
- Temporal granularity limited to monthly aggregates may mask shorter-term fluctuations

Recommendations and Next Steps

- Investigate underlying causes of racial disparities in wait times with more detailed data
- Explore alternative imputation methods or collect more complete satisfaction data
- Incorporate additional operational data (e.g., staffing, case severity) to refine analyses
- Consider patient-level regression modeling to control for confounding factors

Conclusion

This exploratory analysis provides foundational insights into ER patient wait times and satisfaction, highlighting patterns, potential disparities, and areas for deeper analysis. Addressing data quality and incorporating additional variables will be key for more actionable conclusions.