ETE3-2.R

Pranab Rai-3447137

2025-01-02

```r
library(dplyr)
```

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

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

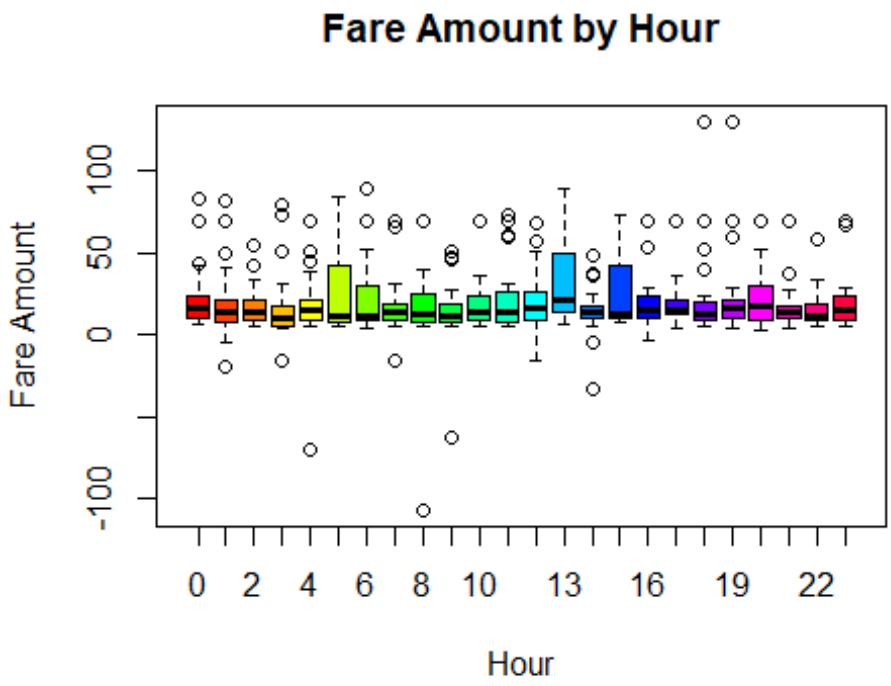
```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(ggplot2)
### --- ANOVA Analysis ---



df <-
read.csv("C:\\Users\\prana\\OneDrive\\Desktop\\2trimester\\R\\ETE3\\test2.csv
")

View(df)

#considering Pr(>F) we deduce the result

### 1. One-Way ANOVA: Total Amount by Pickup Hour
##(" 1. One-Way ANOVA: Does the average total amount vary significantly
across different hours of the day?")

##("Hypotheses:")
##( "H0: The average total amount is the same for all hours of the day.")
##( "Ha: At least one hour of the day has a different average total amount.")
anova_one_way_result <- aov(total_amount ~ factor(pickup_hour), data = df)
summary_table <- summary(anova_one_way_result)
print(summary_table)
```

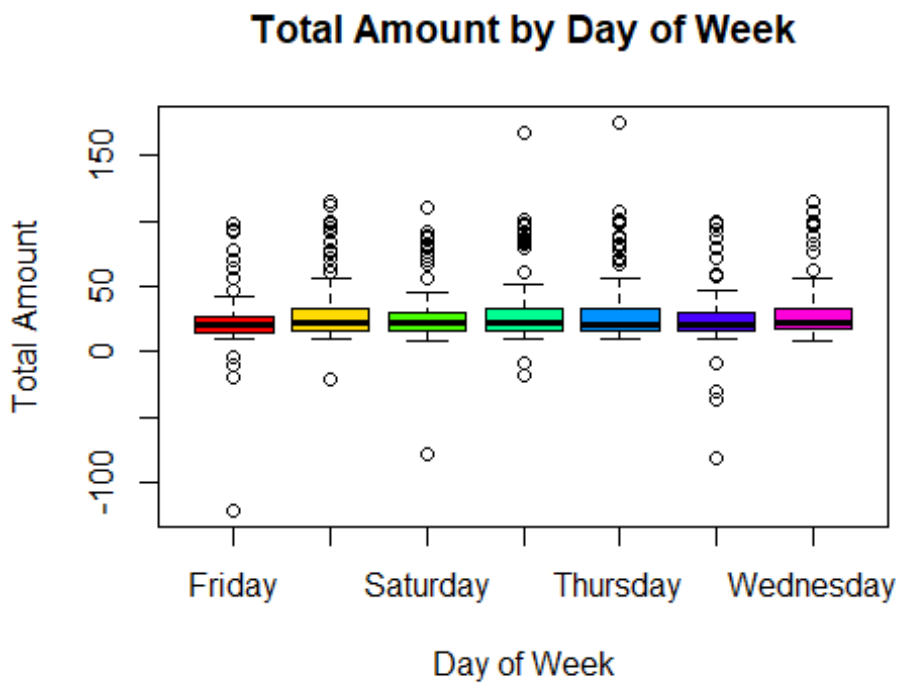
```
##                      Df Sum Sq Mean Sq F value Pr(>F)
## factor(pickup_hour)  23  19136   832.0   1.302  0.157
## Residuals           696 444859   639.2
```

```r
##("Result: The average total amount does not vary significantly across hours
of the day.")

boxplot(fare_amount ~ factor(pickup_hour), data = df,
```

```
        main = "Fare Amount by Hour", xlab = "Hour",
        ylab = "Fare Amount", na.rm = TRUE, col = rainbow(24))
```

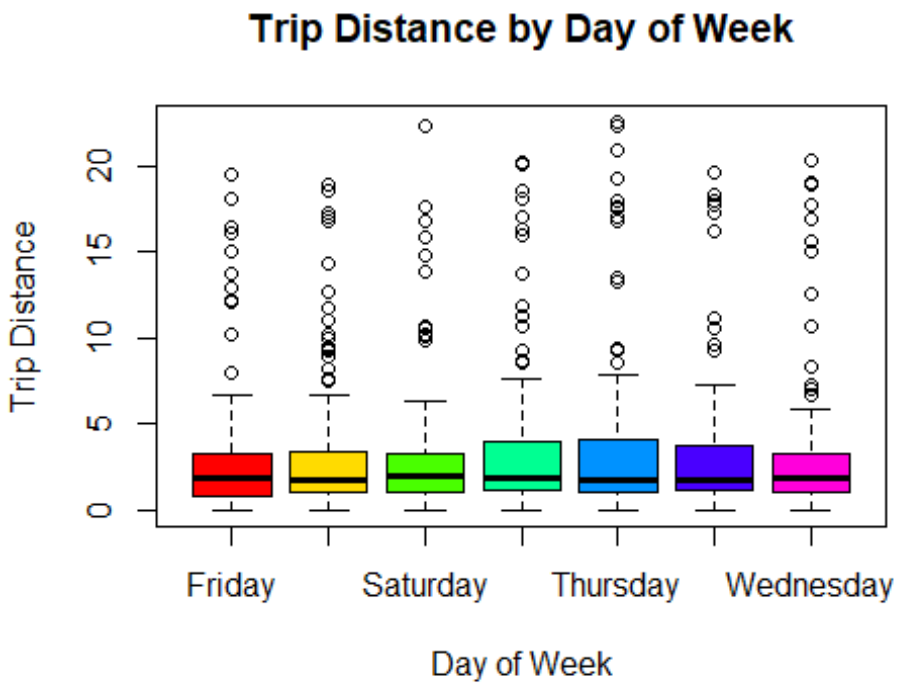



```
#-----------------------------------------------------------

### 2. One-Way ANOVA: Total Amount by Day of Week

##("2. One-Way ANOVA: Does the average total amount vary significantly across
different days of the week?")
##("Hypotheses:")
##("H0: The average total amount is the same for all days of the week.")
##("Ha: At least one day of the week has a different average total amount.")


anova_one_way_day <- aov(total_amount ~ day_of_week, data = df)
summary_table_day <- summary(anova_one_way_day)
print(summary_table_day)

##              Df Sum Sq Mean Sq F value Pr(>F)
## day_of_week   6   5123   853.8   1.327  0.243
## Residuals   713 458872   643.6

boxplot(total_amount ~ day_of_week, data = df,
        main = "Total Amount by Day of Week",
        xlab = "Day of Week", ylab = "Total Amount",
        col= rainbow(7))
```




```
##("Result: The average total amount does not vary significantly across days
of the week.")



#-------------------------------------------------------------

### 3. One-Way ANOVA: Trip Distance by Day of Week

##("3. One-Way ANOVA: Does the average trip distance vary significantly
across different days of the week?")
##("Hypotheses:")
##("H0: The average trip distance is the same for all days of the week.")
##("Ha: At least one day of the week has a different average trip distance.")

anova_one_way_distance <- aov(trip_distance ~ day_of_week, data = df)
summary_table_distance <- summary(anova_one_way_distance)
print(summary_table_distance)

##              Df Sum Sq Mean Sq F value Pr(>F)
## day_of_week   6     51   8.502   0.408  0.874
## Residuals   713  14856  20.836

boxplot(trip_distance ~ day_of_week, data = df,
        main = "Trip Distance by Day of Week",
        xlab = "Day of Week", ylab = "Trip Distance",
```

```
        col= rainbow(7)
)
```

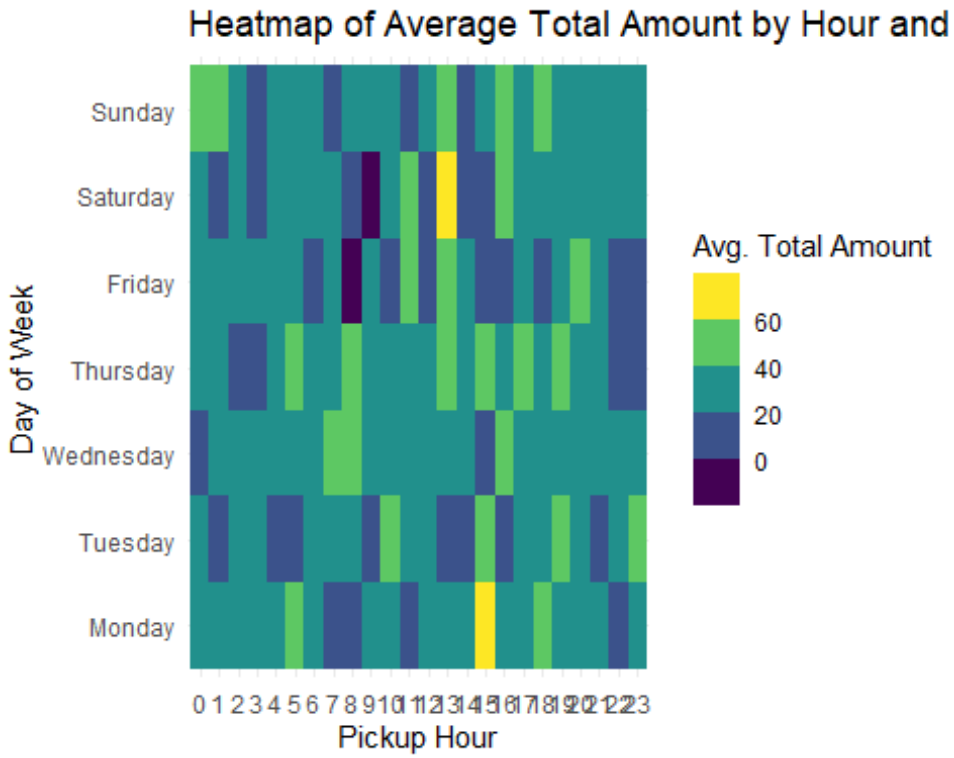



```
##("Result: The average trip distance does not vary significantly across days
of the week.")

#------------------------------------------------------------

### 4. Two-Way ANOVA: Total Amount by Hour of Day and Day of Week
##("4. Two-Way ANOVA: Does the average total amount vary significantly by
both hour of day and day of week?")
##("Hypotheses:")
##("H0a (Main effect of Hour): The average total amount is the same for all
hours of the day.")
##("H1a: At least one hour of the day has a different average total amount.")
##(" ")
##("H0b (Main effect of Day of Week): The average total amount is the same
for all days of the week.")
##("H1b: At least one day of the week has a different average total amount.")
##(" ")
##("H0c (Interaction effect): There is no interaction between hour of day and
day of week on the average total amount.")
##("H1c: There is an interaction between hour of day and day of week on the
average total amount.")
```

```r
anova_two_way_result <- aov(total_amount ~ factor(pickup_hour) * day_of_week,
data = df)
print(summary(anova_two_way_result))
```

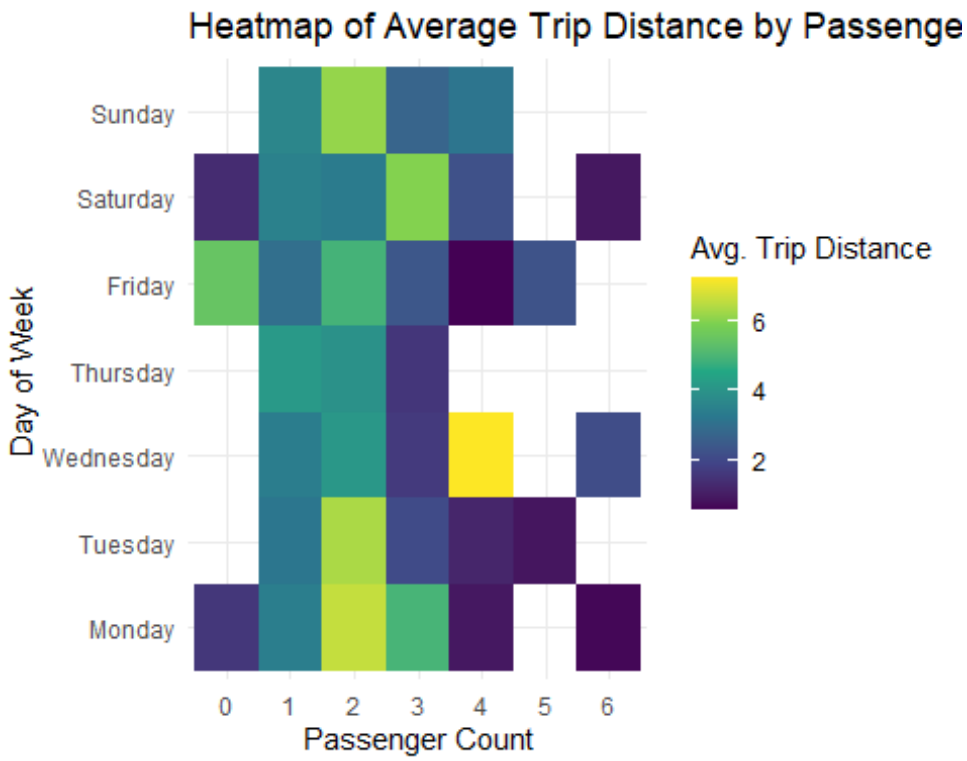
```
##                                   Df Sum Sq Mean Sq F value Pr(>F)
## factor(pickup_hour)               23  19136   832.0   1.312  0.151
## day_of_week                        6   5276   879.3   1.387  0.218
## factor(pickup_hour):day_of_week  138  89596   649.2   1.024  0.419
## Residuals                        552 349988   634.0
```

```r
# Data type conversions
df$pickup_hour <- factor(df$pickup_hour)
df$day_of_week <- factor(df$day_of_week, levels = c("Monday", "Tuesday",
"Wednesday", "Thursday", "Friday", "Saturday", "Sunday"))
df$total_amount <- as.numeric(df$total_amount)

# Calculate average total amount using dplyr
hourly_daily_avg <- df %>%
  group_by(pickup_hour, day_of_week) %>%
  summarize(mean_total_amount = mean(total_amount, na.rm = TRUE), .groups =
"drop")

# Heatmap
ggplot(hourly_daily_avg, aes(x = pickup_hour, y = day_of_week, fill =
mean_total_amount)) +
  geom_tile() +
  labs(title = "Heatmap of Average Total Amount by Hour and Day",
       x = "Pickup Hour", y = "Day of Week", fill = "Avg. Total Amount") +
  theme_minimal() +
  scale_fill_viridis_b(na.value = "grey50")
```




```
##("Result: The average total amount does not vary significantly across hours
of the day.")
##("Result: The average total amount does not vary significantly across days
of the week.")



#------------------------------------------------------------
### 5. Two-Way ANOVA: Trip Distance by Passenger Count and Day of Week
##("5. Two-Way ANOVA: Does the average trip distance vary significantly by
both passenger count and day of week?")
##("Hypotheses:")
##("H0a (Main effect of Passenger Count): The average trip distance is the
same for all passenger counts.")
##("H1a: At least one passenger count has a different average trip
distance.")
##("H0b (Main effect of Day of Week): The average trip distance is the same
for all days of the week.")
##("H1b: At least one day of the week has a different average trip
distance.")
##("H0c (Interaction effect): There is no interaction between passenger count
and day of week on the average trip distance.")
##("H1c: There is an interaction between passenger count and day of week on
the average trip distance.")

anova_two_way_distance <- aov(trip_distance ~ factor(passenger_count) *
```

```
day_of_week, data = df)
summary_table_two_way_distance <- summary(anova_two_way_distance)
print(summary_table_two_way_distance)

##                                       Df Sum Sq Mean Sq F value Pr(>F)
## factor(passenger_count)                6    224   37.36   1.781  0.100
## day_of_week                            6     51    8.43   0.402  0.878
## factor(passenger_count):day_of_week   22    261   11.88   0.566  0.946
## Residuals                            685  14371   20.98

# Data type conversions
df$passenger_count <- factor(df$passenger_count)
df$day_of_week <- factor(df$day_of_week, levels = c("Monday", "Tuesday",
"Wednesday", "Thursday", "Friday", "Saturday", "Sunday")) # Order days
df$trip_distance <- as.numeric(df$trip_distance)

# Calculate average trip distance using dplyr
passenger_daily_avg_distance <- df %>%
  group_by(passenger_count, day_of_week) %>%
  summarize(mean_trip_distance = mean(trip_distance, na.rm = TRUE), .groups =
"drop")

# Heatmap
ggplot(passenger_daily_avg_distance, aes(x = passenger_count, y =
day_of_week, fill = mean_trip_distance)) +
  geom_tile() +
  labs(title = "Heatmap of Average Trip Distance by Passenger Count and Day",
       x = "Passenger Count", y = "Day of Week", fill = "Avg. Trip Distance")
+
  theme_minimal() +
  scale_fill_viridis_c(na.value = "grey50")
```




```
##("Result: The average trip distance does not vary significantly across passenger counts.")
##("Result: The average trip distance does not vary significantly across days of the week.")



#-------------------------------------------------------------


##Interpretation after ANOVA:


#The time of day does not appear to influence total amounts significantly. This suggests that fare rates and demand are relatively stable throughout the day, possibly due to consistent pricing policies or a lack of strong hourly demand patterns.


#The day of the week does not affect the total amount significantly. This could indicate that usage patterns and fare structures remain steady regardless of the weekday or weekend.


#Trip distances are consistent across the week. This might reflect uniform travel needs or a lack of day-specific events impacting trip distances.
```

```
#Neither the time of day, the day of the week, nor their combination
significantly impacts the total amount. This supports the notion of
consistent fare structures and user behavior across these dimensions.


#Trip distances are not significantly influenced by the number of passengers
or the day of the week. This could imply that most trips are short and
consistent, regardless of these factors.
```