# ==========================================

# STEP 4: DATA ANALYSIS AND VISUALIZATIONS

# ==========================================

## TRIP LENGTH

## ===========

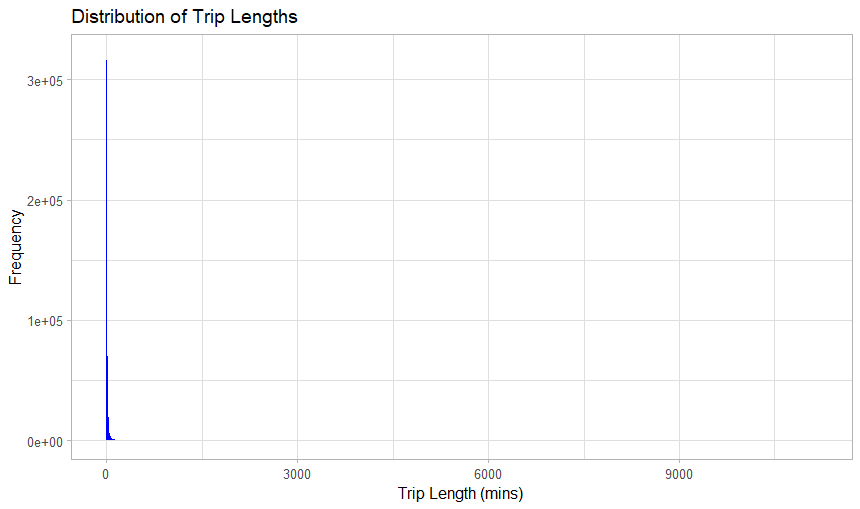
# Histogram of trip lengths

ggplot(data = real\_trips\_v2, aes(x = trip\_length)) +

geom\_histogram(binwidth = 1, fill = "blue", alpha = 1) +

labs(title = "Distribution of Trip Lengths", x = "Trip Length (mins)", y = "Frequency") +

theme\_light()



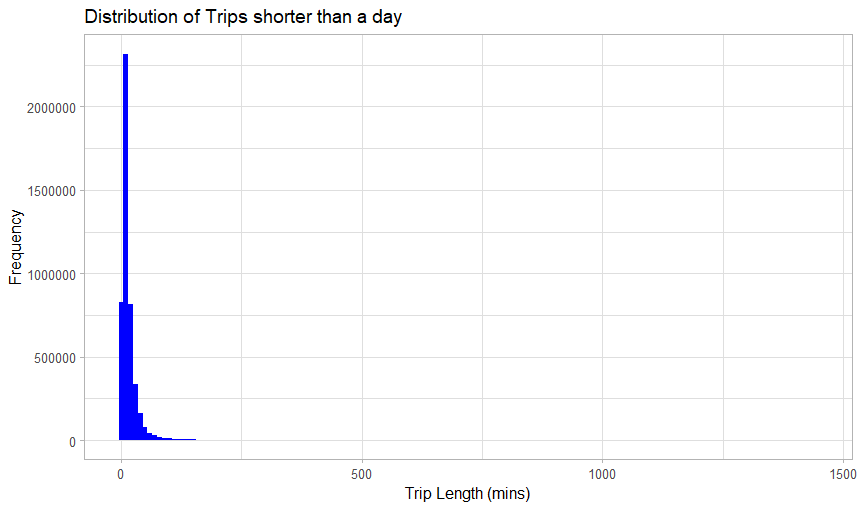
# Histogram of trip lengths shorter than a day

ggplot(data = real\_trips\_v2 %>% filter(trip\_length < 1440), aes(x = trip\_length)) +

geom\_histogram(binwidth = 10, fill = "blue", alpha = 1) +

labs(title = "Distribution of Trips shorter than 24 hours", x = "Trip Length (mins)", y = "Frequency") +

theme\_light()



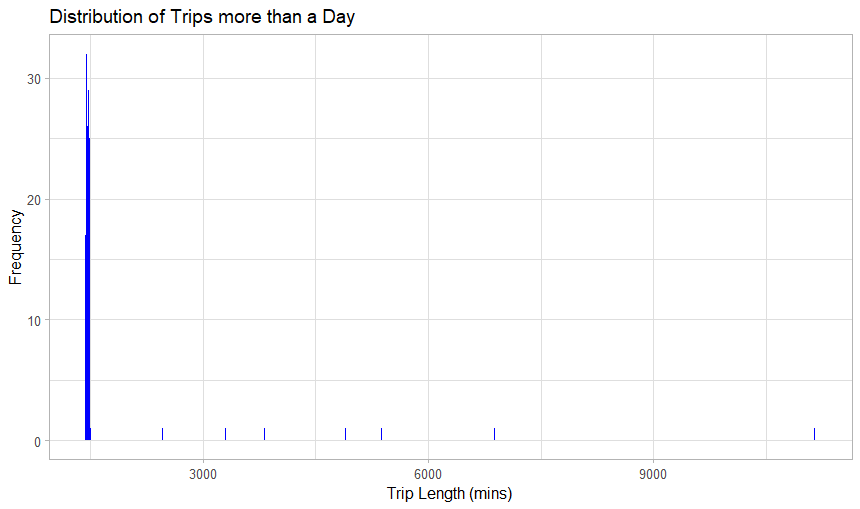
# Histogram of trip lengths longer than a day

ggplot(data = real\_trips\_v2 %>% filter(trip\_length > 1440), aes(x = trip\_length)) +

geom\_histogram(binwidth = 10, fill = "blue", alpha = 1) +

labs(title = "Distribution of Trips more than a Day", x = "Trip Length (mins)", y = "Frequency") +

theme\_light()



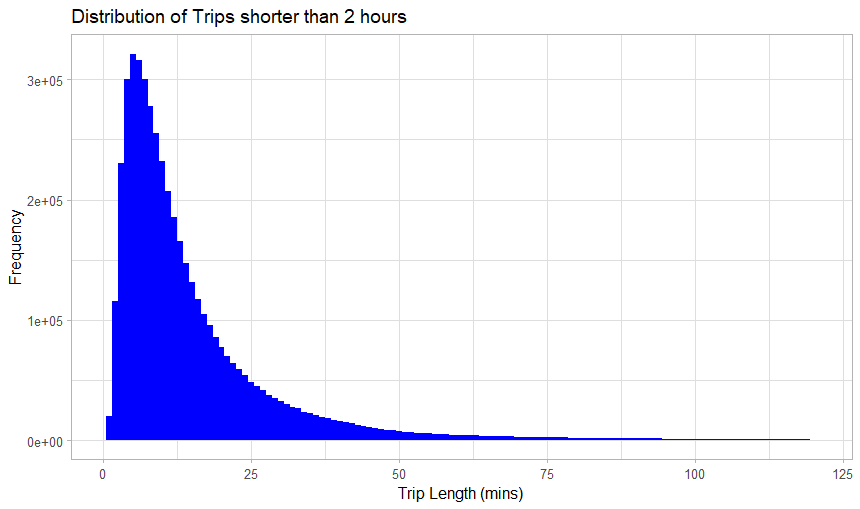
# Histogram of trip lengths shorter than 2 hrs

ggplot(data = real\_trips\_v2 %>% filter(trip\_length < 120), aes(x = trip\_length)) +

geom\_histogram(binwidth = 1, fill = "blue", alpha = 1) +

labs(title = "Distribution of Trips shorter than 2 hours", x = "Trip Length (mins)", y = "Frequency") +

theme\_light()



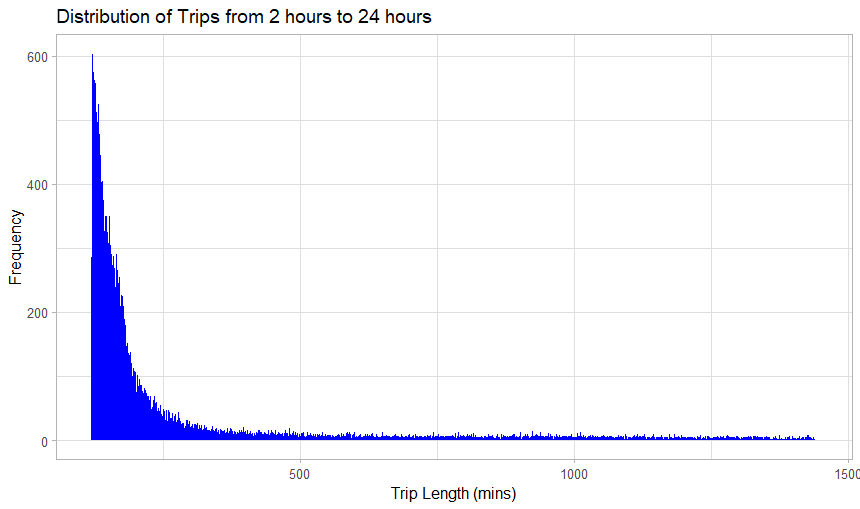
# Histogram of trip lengths from 2 hours to 24 hours

ggplot(data = real\_trips\_v2 %>% filter(trip\_length > 120 & trip\_length < 1440), aes(x = trip\_length)) +

geom\_histogram(binwidth = 1, fill = "blue", alpha = 1) +

labs(title = "Distribution of Trips from 2 hours to 24 hours", x = "Trip Length (mins)", y = "Frequency") +

theme\_light()



# frequency of trips started by day of week

real\_trips\_v2 %>%

count(day\_of\_week\_started) %>%

ggplot(aes(x = day\_of\_week\_started, y = n)) +

geom\_bar(stat = "identity", fill = "blue", alpha = 0.7) +

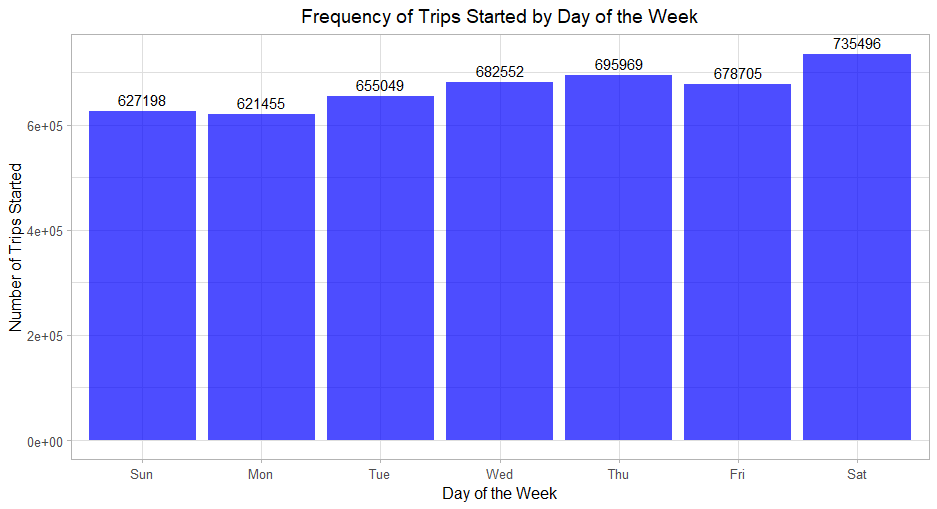
geom\_text(aes(label = n), vjust = -0.5, size = 3.5) + # Add labels above bars

labs(title = "Frequency of Trips Started by Day of the Week",

x = "Day of the Week", y = "Number of Trips Started") +

theme\_light() +

theme(plot.title = element\_text(hjust = 0.5)) # Center the title



# Plot for trips less than 1 hour

real\_trips\_v2 %>%

filter(trip\_duration\_category == "Less than 1 hour") %>%

count(day\_of\_week\_started) %>%

ggplot(aes(x = day\_of\_week\_started, y = n, group = 1)) +

geom\_line(color = "blue", size = 1) +

geom\_point(color = "blue", size = 2) +

geom\_text(aes(label = n), vjust = -0.5, size = 3) +

labs(

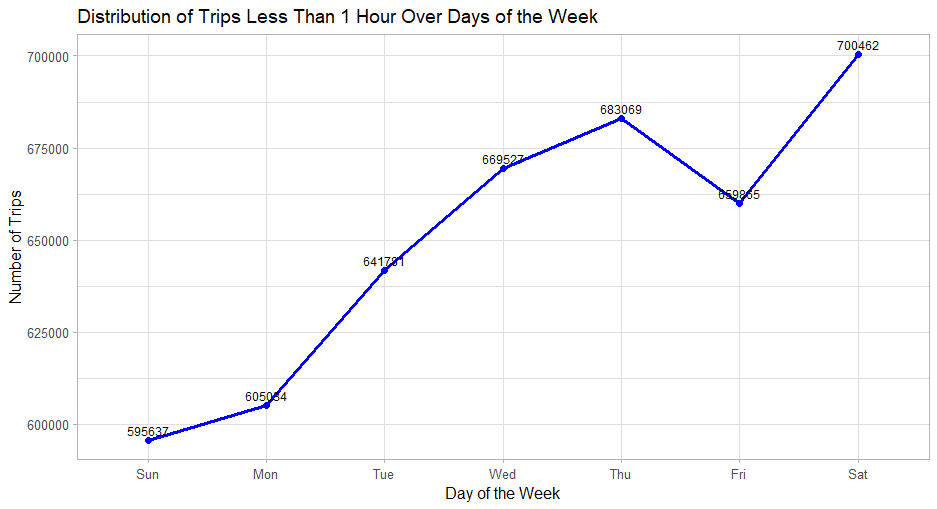
title = "Distribution of Trips Less Than 1 Hour Over Days of the Week",

x = "Day of the Week",

y = "Number of Trips"

) +

theme\_light()



# Plot for remaining categories

real\_trips\_v2 %>%

filter(trip\_duration\_category != "Less than 1 hour") %>%

count(day\_of\_week\_started, trip\_duration\_category) %>%

ggplot(aes(x = day\_of\_week\_started, y = n, color = trip\_duration\_category, group = trip\_duration\_category)) +

geom\_line(size = 0.5) +

geom\_point(size = 2) +

geom\_text(aes(label = n), vjust = 1, hjust = -0.5, size = 3) +

labs(

title = "Distribution of Trips (Remaining Categories) Over Days of the Week",

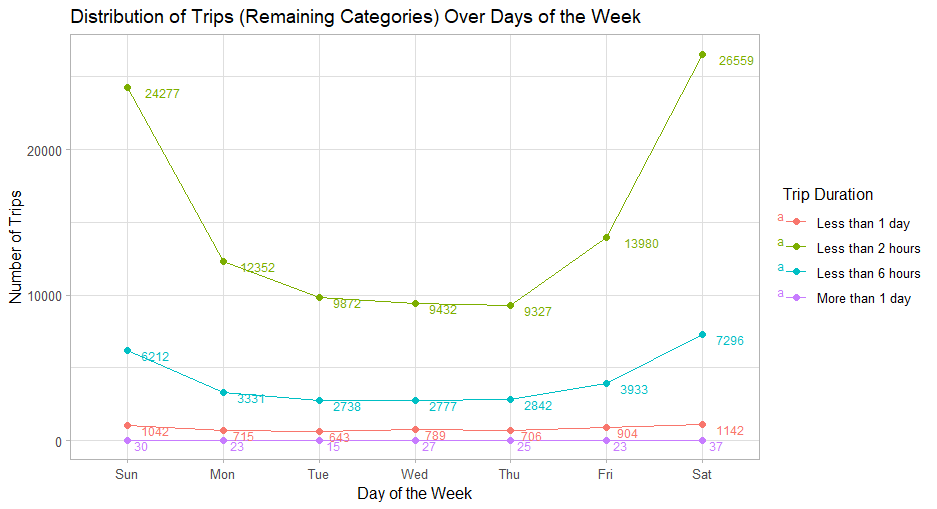
x = "Day of the Week",

y = "Number of Trips",

color = "Trip Duration"

) +

theme\_light()



# plotting start/end times (extract hour and minute of the day from started\_at and ended\_at)

real\_trips\_v2 %>%

mutate(

start\_time\_minute = hour(started\_at) \* 60 + minute(started\_at) + second(started\_at) / 60,

end\_time\_minute = hour(ended\_at) \* 60 + minute(ended\_at) + second(ended\_at) / 60

) %>%

ggplot() +

geom\_density(aes(x = start\_time\_minute, color = "Start Time"), fill = "blue", alpha = 0.3) +

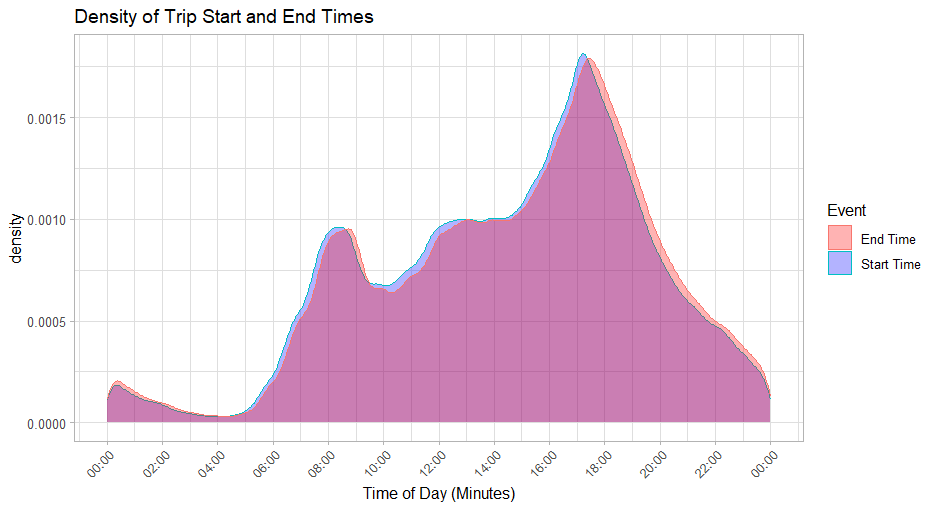
geom\_density(aes(x = end\_time\_minute, color = "End Time"), fill = "red", alpha = 0.3) +

scale\_x\_continuous(name = "Time of Day (Minutes)", breaks = seq(0, 1440, by = 120), labels = function(x) format(as.POSIXct(x \* 60, origin = "1970-01-01", tz = "UTC"), "%H:%M")) +

labs(title = "Density of Trip Start and End Times", color = "Event") +

theme\_light() +

theme(axis.text.x = element\_text(angle = 45, hjust = 1))



The graph shows the distribution of trip start and end times throughout the day. The density curves represent the likelihood of trips starting and ending at each minute in the day. Peaks in the curves indicate the times of day with the highest likelihood of trip starts and ends.

# member trips

real\_trips\_v2 %>%

filter(member\_casual == "member") %>%

mutate(

start\_time\_minute = hour(started\_at) \* 60 + minute(started\_at) + second(started\_at) / 60,

end\_time\_minute = hour(ended\_at) \* 60 + minute(ended\_at) + second(ended\_at) / 60

) %>%

ggplot() +

geom\_density(aes(x = start\_time\_minute, color = "Start Time"), fill = "blue", alpha = 0.3) +

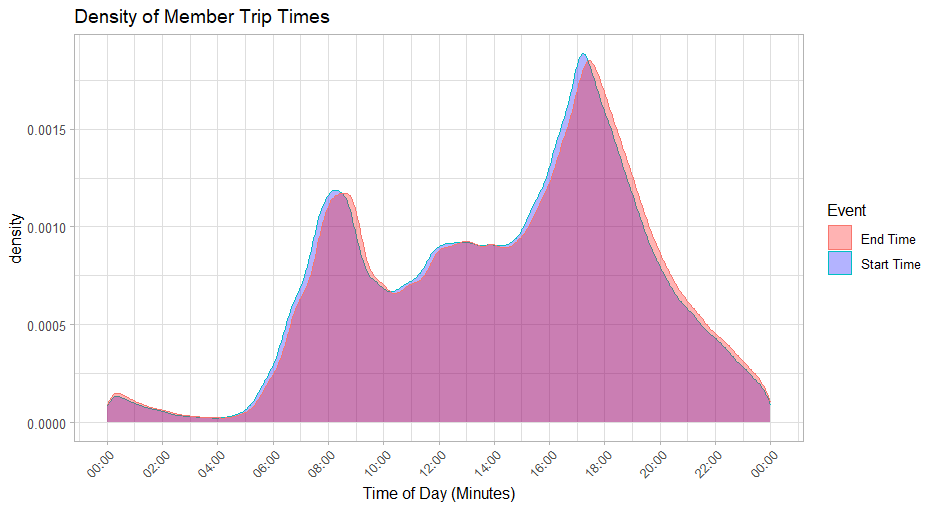
geom\_density(aes(x = end\_time\_minute, color = "End Time"), fill = "red", alpha = 0.3) +

scale\_x\_continuous(name = "Time of Day (Minutes)", breaks = seq(0, 1440, by = 120), labels = function(x) format(as.POSIXct(x \* 60, origin = "1970-01-01", tz = "UTC"), "%H:%M")) +

labs(title = "Density of Member Trip Times", color = "Event") +

theme\_light() +

theme(axis.text.x = element\_text(angle = 45, hjust = 1))



# casual trips

real\_trips\_v2 %>%

filter(member\_casual == "casual") %>%

mutate(

start\_time\_minute = hour(started\_at) \* 60 + minute(started\_at) + second(started\_at) / 60,

end\_time\_minute = hour(ended\_at) \* 60 + minute(ended\_at) + second(ended\_at) / 60

) %>%

ggplot() +

geom\_density(aes(x = start\_time\_minute, color = "Start Time"), fill = "blue", alpha = 0.3) +

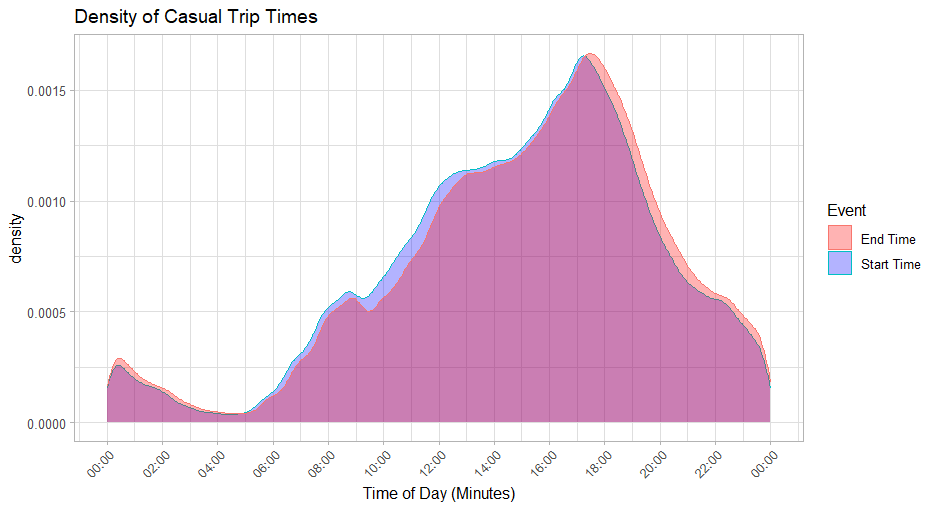
geom\_density(aes(x = end\_time\_minute, color = "End Time"), fill = "red", alpha = 0.3) +

scale\_x\_continuous(name = "Time of Day (Minutes)", breaks = seq(0, 1440, by = 120), labels = function(x) format(as.POSIXct(x \* 60, origin = "1970-01-01", tz = "UTC"), "%H:%M")) +

labs(title = "Density of Casual Trip Times", color = "Event") +

theme\_light() +

theme(axis.text.x = element\_text(angle = 45, hjust = 1))



# plotting start/end times again with frequency plot

real\_trips\_v2 %>%

mutate(

start\_time\_minute = hour(started\_at) \* 60 + minute(started\_at) + second(started\_at) / 60,

end\_time\_minute = hour(ended\_at) \* 60 + minute(ended\_at) + second(ended\_at) / 60

) %>%

ggplot() +

geom\_freqpoly(aes(x = start\_time\_minute, color = "Start Time"), binwidth = 1, alpha = 0.6, size= 1) +

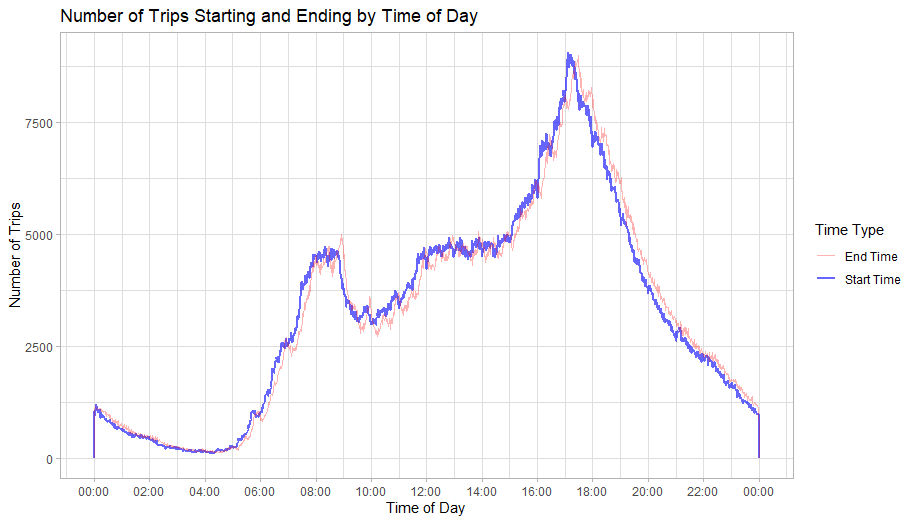
geom\_freqpoly(aes(x = end\_time\_minute, color = "End Time"), binwidth = 1, alpha = 0.3, size = 0.5) +

scale\_x\_continuous(name = "Time of Day", breaks = seq(0, 1440, by = 120), labels = function(x) format(as.POSIXct(x \* 60, origin = "1970-01-01", tz = "UTC"), "%H:%M")) +

scale\_color\_manual(values = c("Start Time" = "blue", "End Time" = "red")) + # Specify colors for the legend

labs(title = "Number of Trips Starting and Ending by Time of Day", color = "Time Type", y = "Number of Trips") + # Add title to the legend

theme\_light()



# same as above for members

real\_trips\_v2 %>%

filter(member\_casual == "member") %>%

mutate(

start\_time\_minute = hour(started\_at) \* 60 + minute(started\_at) + second(started\_at) / 60,

end\_time\_minute = hour(ended\_at) \* 60 + minute(ended\_at) + second(ended\_at) / 60

) %>%

ggplot() +

geom\_freqpoly(aes(x = start\_time\_minute, color = "Start Time"), binwidth = 1, alpha = 0.6, size= 1) +

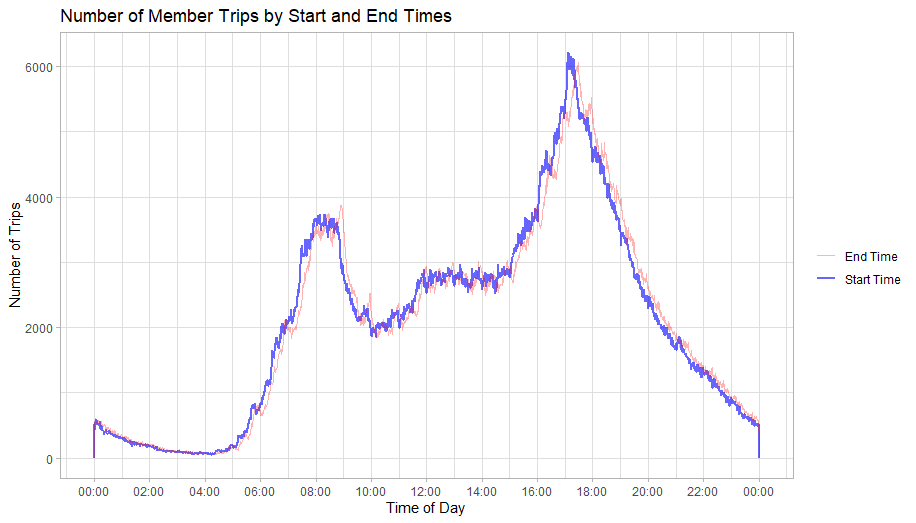
geom\_freqpoly(aes(x = end\_time\_minute, color = "End Time"), binwidth = 1, alpha = 0.3, size = 0.5) +

scale\_x\_continuous(name = "Time of Day", breaks = seq(0, 1440, by = 120), labels = function(x) format(as.POSIXct(x \* 60, origin = "1970-01-01", tz = "UTC"), "%H:%M")) +

scale\_color\_manual(values = c("Start Time" = "blue", "End Time" = "red")) +

labs(title = "Number of Member Trips by Start and End Times", color = " ", y = "Number of Trips") +

theme\_light()



# same as above for casuals

real\_trips\_v2 %>%

filter(member\_casual == "casual") %>%

mutate(

start\_time\_minute = hour(started\_at) \* 60 + minute(started\_at) + second(started\_at) / 60,

end\_time\_minute = hour(ended\_at) \* 60 + minute(ended\_at) + second(ended\_at) / 60

) %>%

ggplot() +

geom\_freqpoly(aes(x = start\_time\_minute, color = "Start Time"), binwidth = 1, alpha = 0.6, size= 1) +

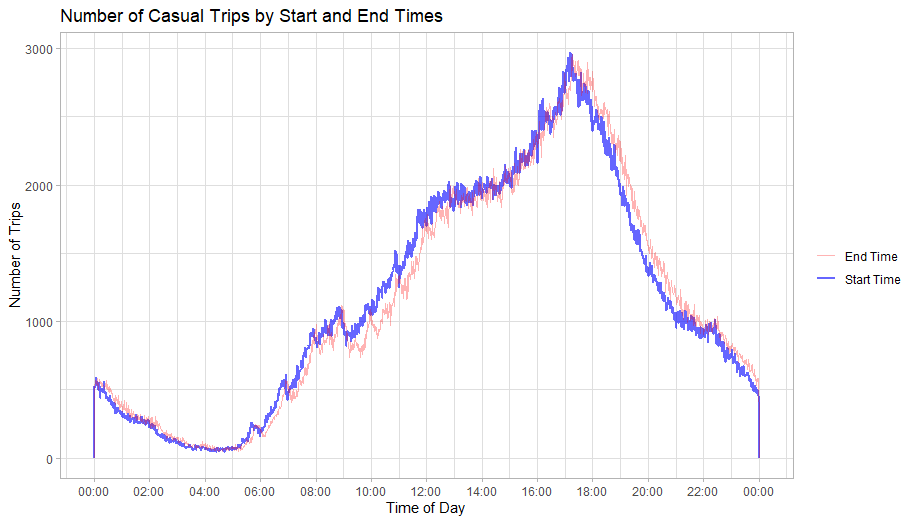
geom\_freqpoly(aes(x = end\_time\_minute, color = "End Time"), binwidth = 1, alpha = 0.3, size = 0.5) +

scale\_x\_continuous(name = "Time of Day", breaks = seq(0, 1440, by = 120), labels = function(x) format(as.POSIXct(x \* 60, origin = "1970-01-01", tz = "UTC"), "%H:%M")) +

scale\_color\_manual(values = c("Start Time" = "blue", "End Time" = "red")) +

labs(title = "Number of Casual Trips by Start and End Times", color = " ", y = "Number of Trips") +

theme\_light()



# bar chart showing distribution of trips along rideable type and member status

ggplot(real\_trips\_v2, aes(x = factor(rideable\_type, levels = c("classic\_bike", "electric\_bike", "docked\_bike")), fill = member\_casual)) +

geom\_bar(position = "stack") +

geom\_text(stat = "count", aes(label = after\_stat(count)), position = position\_stack(vjust = 0.5)) +

scale\_y\_continuous(labels = scales::comma) +

labs(title = "Breakdown of Rideable Types by Member Status",

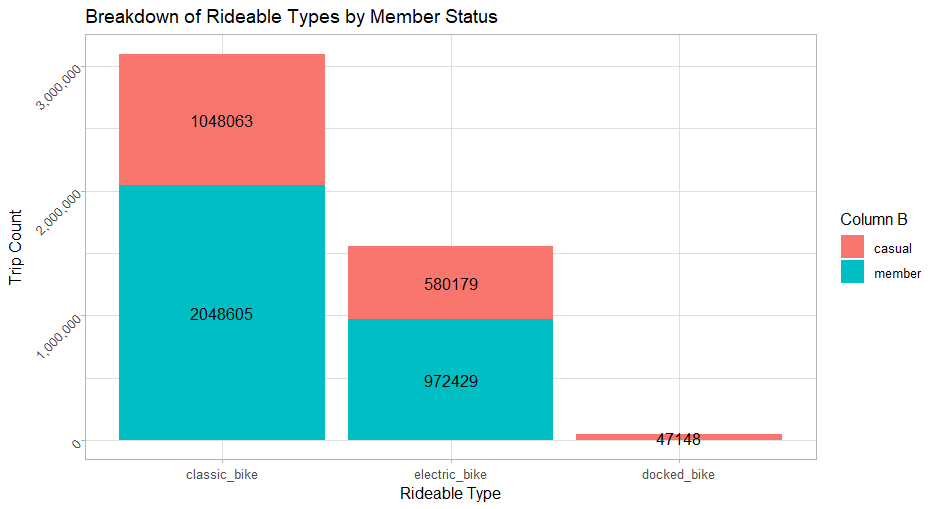
x = "Rideable Type",

y = "Trip Count",

fill = "Column B") +

theme\_light() +

theme(axis.text.y = element\_text(angle = 45, hjust = 1))



# histogram of trips by user type (use filter to change trip lengths, use position to be dodge or stack)

ggplot(real\_trips\_v2 %>% filter(trip\_length < 100), aes(x = trip\_length, fill = member\_casual)) +

geom\_histogram(binwidth = 1, position = "dodge", color = "black") +

scale\_y\_continuous(labels = scales::comma) +

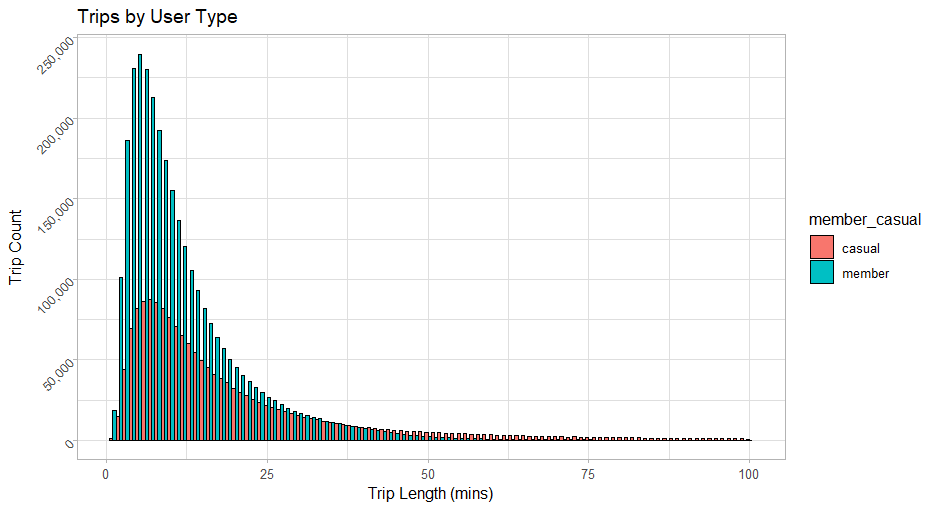
labs(title = "Trips by User Type",

x = "Trip Length (mins)",

y = "Trip Count") +

theme\_light() +

theme(axis.text.y = element\_text(angle = 45, hjust = 1))



# histogram of trip count by day of week and user type

real\_trips\_v2 %>%

group\_by(member\_casual, day\_of\_week\_started) %>%

summarise(trip\_count = n()) %>%

ggplot(aes(x = day\_of\_week\_started, y = trip\_count, fill = member\_casual)) +

geom\_col(position = "dodge") +

scale\_y\_continuous(labels = scales::comma) +

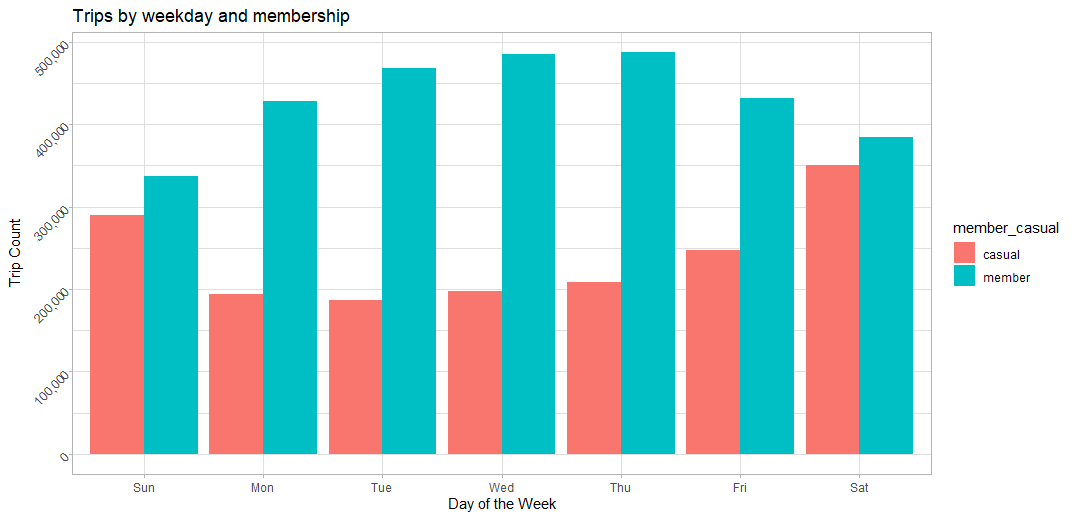
labs(title = "Trips by weekday and membership",

x = "Day of the Week",

y = "Trip Count") +

theme\_light() +

theme(axis.text.y = element\_text(angle = 45, hjust = 1))



# histogram of average trip length by day of week and user type

real\_trips\_v2 %>%

group\_by(member\_casual, day\_of\_week\_started) %>%

summarise(avg\_trip\_length = mean(trip\_length)) %>%

ggplot(aes(x = day\_of\_week\_started, y = avg\_trip\_length, fill = member\_casual)) +

geom\_col(position = "dodge") +

scale\_y\_continuous(labels = scales::comma) +

labs(title = "Trip Length by weekday and membership",

x = "Day of the Week",

y = "Trip Length") +

theme\_light() +

theme(axis.text.y = element\_text(angle = 45, hjust = 1))

