Week 5 Tasks - Data Visualization in R

Dat Thanh

Instructions

- Use this template to complete Week 5 tasks. Replace placeholders with your work.
- Ensure the document can knit top-to-bottom without errors.
- Add short captions/annotations below each plot.

Task 1 - Setup and Data Loading

```
# TODO: Set path and read your dataset
# Example:
# data <- read_csv("path/to/your.csv")
# head(data, 10)
library(RKaggle)
data <- RKaggle::get_dataset("yashdevladdha/uber-ride-analytics-dashboard")

head(data, 10)
```

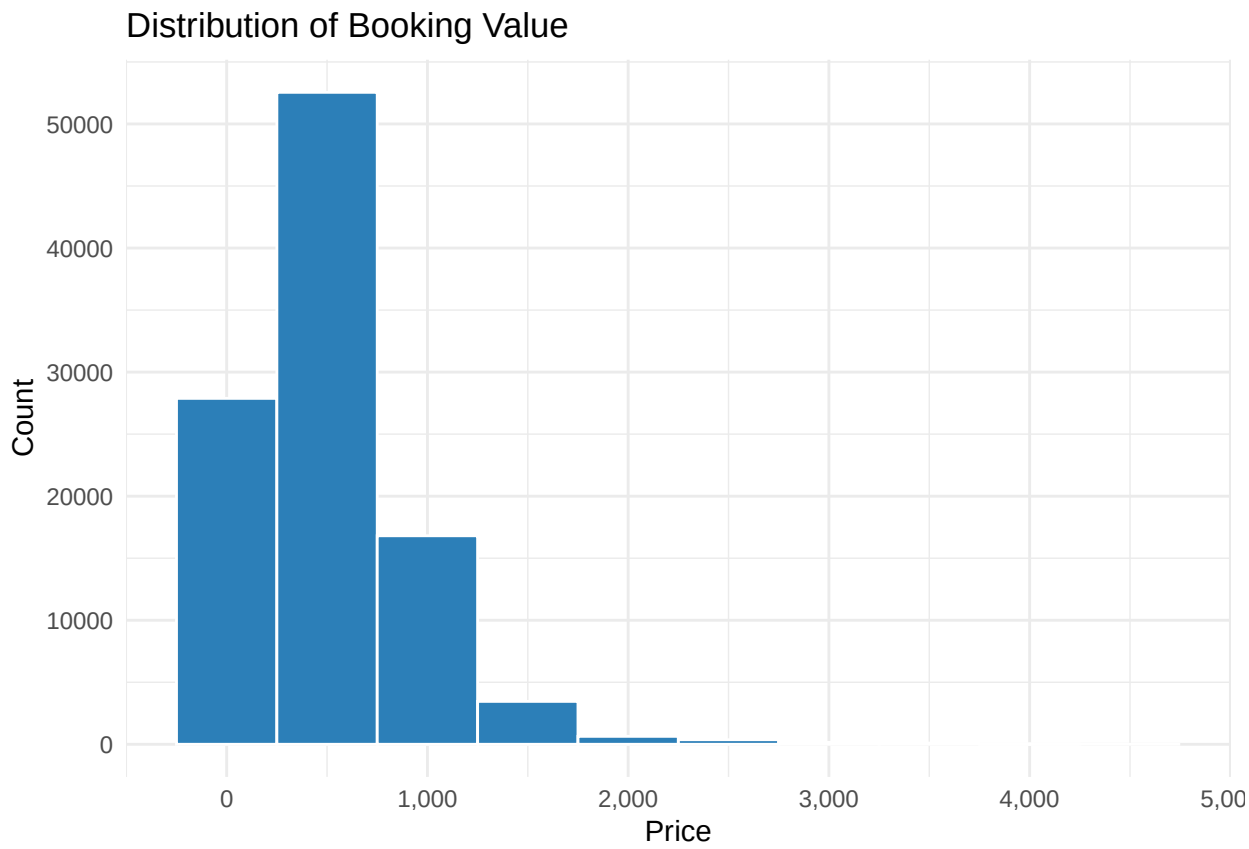
```
## # A tibble: 10 x 21
##    Date       Time     `Booking ID`     `Booking Status` `Customer ID`
##    <date>     <time>   <chr>            <chr>            <chr>
##  1 2024-03-23 12:29:38 "\"CNR5884300\"" No Driver Found  "\"CID1982111\""
##  2 2024-11-29 18:01:39 "\"CNR1326809\"" Incomplete       "\"CID4604802\""
##  3 2024-08-23 08:56:10 "\"CNR8494506\"" Completed        "\"CID9202816\""
##  4 2024-10-21 17:17:25 "\"CNR8906825\"" Completed        "\"CID2610914\""
##  5 2024-09-16 22:08:00 "\"CNR1950162\"" Completed        "\"CID9933542\""
##  6 2024-02-06 09:44:56 "\"CNR4096693\"" Completed        "\"CID4670564\""
##  7 2024-06-17 15:45:58 "\"CNR2002539\"" Completed        "\"CID6800553\""
##  8 2024-03-19 17:37:37 "\"CNR6568000\"" Completed        "\"CID8610436\""
##  9 2024-09-14 12:49:09 "\"CNR4510807\"" No Driver Found  "\"CID7873618\""
## 10 2024-12-16 19:06:48 "\"CNR7721892\"" Incomplete       "\"CID5214275\""
## # i 16 more variables: `Vehicle Type` <chr>, `Pickup Location` <chr>,
## #   `Drop Location` <chr>, `Avg VTAT` <chr>, `Avg CTAT` <chr>,
## #   `Cancelled Rides by Customer` <chr>,
## #   `Reason for cancelling by Customer` <chr>,
## #   `Cancelled Rides by Driver` <chr>, `Driver Cancellation Reason` <chr>,
## #   `Incomplete Rides` <chr>, `Incomplete Rides Reason` <chr>,
## #   `Booking Value` <chr>, `Ride Distance` <chr>, `Driver Ratings` <chr>, ...
```

Briefly describe the dataset and variables here.

ncr_ride_bookings.csv is a table that keeps track of ride-hailing bookings, containing 150,000 rows and 21 columns. It includes bookings from January 1, 2024, to December 30, 2024 (UTC). The table combines different types of data, like booking status and vehicle type, with some numerical metrics such as average VTAT and CTAT, along with date and time fields. Common identifiers and context columns are Booking ID, Booking Status, Vehicle Type, Pickup Location, Drop Location, and timing/SLAs.

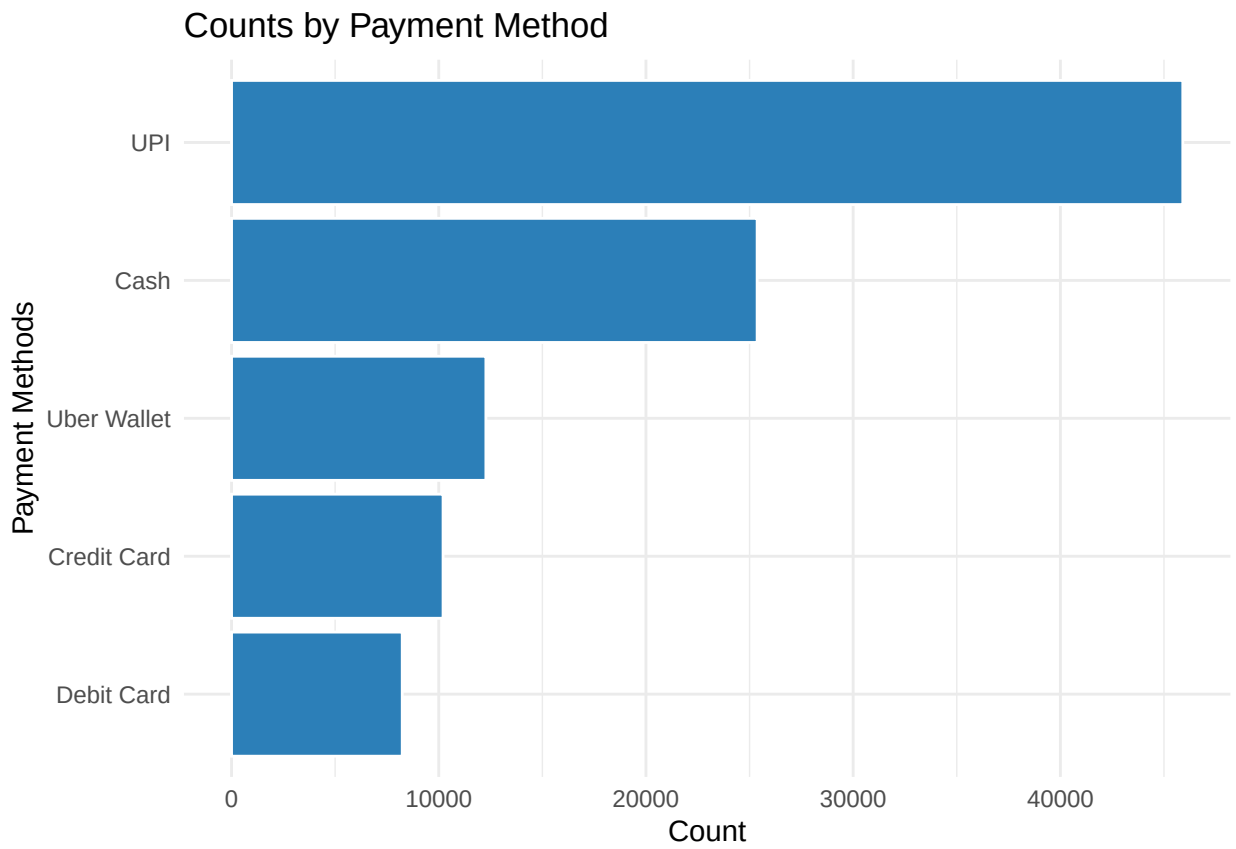
Task 2 - Univariate Visualizations

```
# Example histogram/density for a numeric variable
# ggplot(data, aes(numeric_col)) +
#   geom_histogram(binwidth = 5, fill = "#2c7fb8") +
#   labs(title = "Distribution of numeric_col", x = "numeric_col", y = "Count")

library(dplyr)
library(readr)
library(ggplot2)
library(scales)

data_clean <- data %>%
  mutate(BookingValue = parse_number(`Booking Value`))

ggplot(data_clean, aes(x = BookingValue)) +
  geom_histogram(binwidth = 500, fill = "#2c7fb8", color = "white",na.rm = TRUE) +
  labs(title = "Distribution of Booking Value", x = "Price", y = "Count") +
  scale_x_continuous(labels = comma) +
  theme_minimal()
```



Key takeaway: Pronounced right skew: The majority of reservations are cheap to mid-priced, with a long tail extending to high values (infrequent costly visits). The primary mass seems to be about inside the 200-1,200 range, based on visual estimation of the bins, exhibiting a single dominating mode. Significant outliers: Several reservations fall within the range of 2,000 to 5,000; categorize them as outliers or a separate group. Potential negatives/zeros: The x-axis seems to decline below zero, perhaps indicating refunds, credits, or data discrepancies. Deserving of validation.

```
# Example bar plot for a categorical variable
# data %>% count(category) %>%
#   ggplot(aes(x = fct_reorder(category, n), y = n)) +
#   geom_col(fill = "#7fcdbb") + coord_flip() +
#   labs(title = "Counts by category", x = "Category", y = "Count")
library(ggplot2)
library(dplyr)

data_clean <- data %>%
  mutate(`Payment Method` = na_if(`Payment Method`, "null")) %>%
  filter(!is.na(`Payment Method`))

data_clean %>% count(`Payment Method`) %>%
  ggplot(aes(x = fct_reorder(`Payment Method`, n), y = n)) +
  geom_col(fill = "#2c7fb8", color = "white") + coord_flip() +
  labs(title = "Counts by Payment Method", x = "Payment Methods", y = "Count")
```

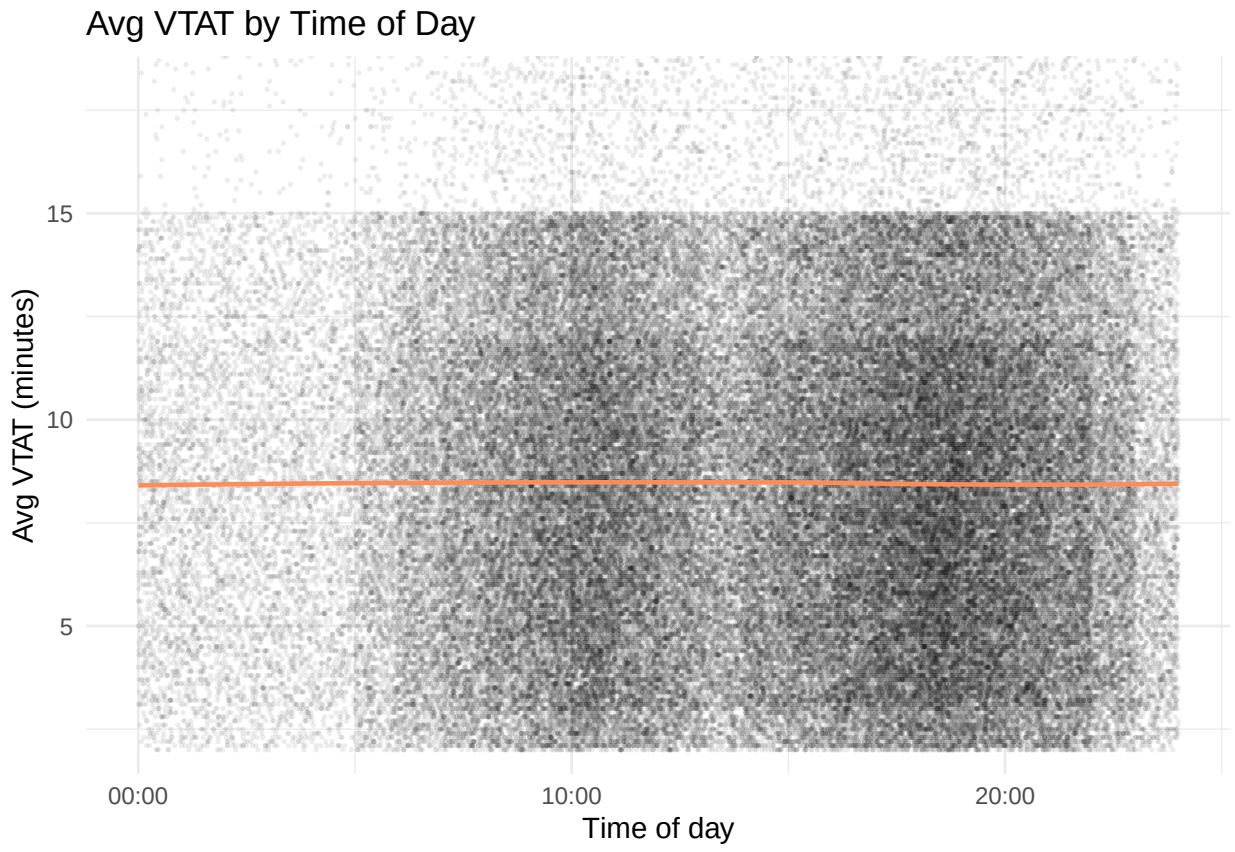


Key takeaway: UPI prevails significantly as the predominant method of payment for clients. Cash remains the unequivocal second choice, nevertheless comprising a substantial portion, indicating potential for transition to digital alternatives. Uber Wallet has a moderate position in terms of acceptance, significantly trailing behind UPI; it presents a possible opportunity for promotions and loyalty initiatives. Card transactions (credit/debit) lag behind both UPI and cash; card incentives are not increasing market share. Operational implication: guarantee the robustness of UPI uptime and restrictions; enhance cash management to minimize friction and leakage.

Task 3 - Bivariate Visualizations

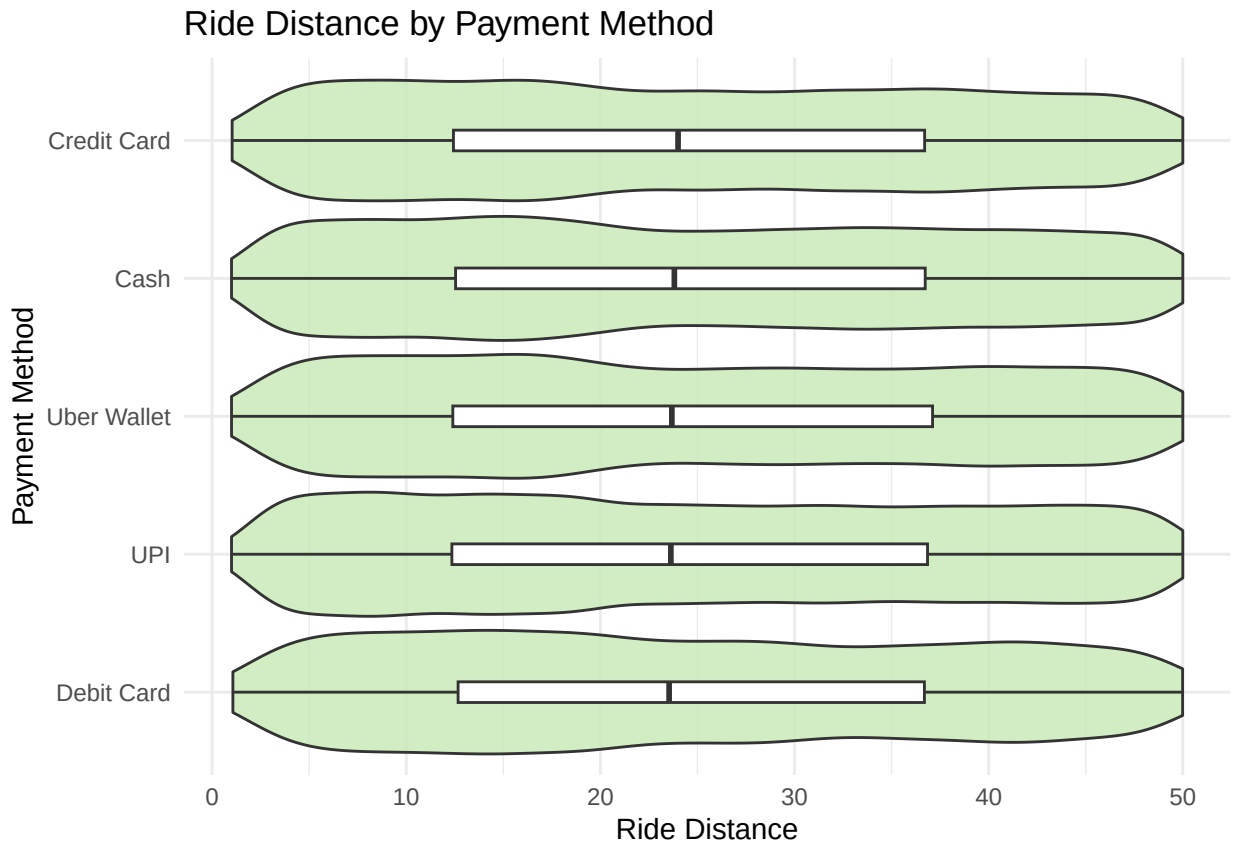
```
# Example scatter with smooth
# ggplot(data, aes(x = xvar, y = yvar)) +
#   geom_point(alpha = 0.6) +
#   geom_smooth(method = "lm", se = FALSE, color = "#fc8d59") +
#   labs(title = "xvar vs yvar")

library(dplyr)
library(ggplot2)
library(readr)   # parse_number()
library(hms)     # as_hms()
library(scales)  # axis label helpers

label_time_safe <- if ("label_time" %in% getNamespaceExports("scales")) {
  scales::label_time("%H:%M")
```

```r
} else {
  scales::time_format("%H:%M")
}

df_vtat <- data %>%
  mutate(
    # Time-of-day as 'hms'. Works for "HH:MM(:SS)" strings or POSIXt.
    time_of_day = if (inherits(Time, "hms")) Time else as_hms(Time),
    AvgVTAT = parse_number(`Avg VTAT`)  # e.g., "12.5 min" -> 12.5
  ) %>%
  filter(!is.na(time_of_day), !is.na(AvgVTAT))

ylims <- quantile(df_vtat$AvgVTAT, c(.01, .99), na.rm = TRUE)

ggplot(df_vtat, aes(x = time_of_day, y = AvgVTAT)) +
  geom_point(alpha = 0.08, size = 0.6, shape = 16) +
  geom_smooth(method = "loess", se = FALSE, color = "#fc8d59", linewidth = 0.8) +
  coord_cartesian(ylim = ylims) +
  scale_x_time(labels = label_time_safe) +
  labs(
    title = "Avg VTAT by Time of Day",
    x = "Time of day",
    y = "Avg VTAT (minutes)"
  ) +
  theme_minimal()
```



Interpretation: The scatter plot indicates that the average VTAT remains pretty consistent at around 8–9 minutes all day long, and the smooth line is basically flat over the 24-hour period. Every hour, there's a pretty big spread (like 3–15+ minutes), showing a lot of variability that doesn't really make sense with just the clock time. Any little dip in the middle of the day or a rise in the evening doesn't really matter much compared to all the other stuff going on. It seems like pickup latency is influenced more by local factors like where you are, what supplies are available, and the type of vehicle, rather than the time of day.

```r
# Example box/violin by category
# ggplot(data, aes(x = category, y = numeric_col)) +
#   geom_violin(fill = "#c7e9b4", alpha = 0.8) +
#   geom_boxplot(width = 0.15, outlier.alpha = 0.2) +
#   labs(title = "numeric_col by category")

library(dplyr)
library(ggplot2)
library(forcats)
library(readr)   # parse_number

data_clean <- data %>%
  mutate(
    `Payment Method` = na_if(`Payment Method`, "null"),
    `Ride Distance`  = parse_number(`Ride Distance`)
  ) %>%
  filter(!is.na(`Payment Method`), !is.na(`Ride Distance`))

ggplot(data_clean,
       aes(x = fct_reorder(`Payment Method`, `Ride Distance`, .fun = median),
           y = `Ride Distance`)) +
  geom_violin(fill = "#c7e9b4", alpha = 0.8, trim = TRUE) +
  geom_boxplot(width = 0.15, outlier.alpha = 0.2) +
  coord_flip() +
  labs(title = "Ride Distance by Payment Method",
       x = "Payment Method", y = "Ride Distance") +
  theme_minimal()
```

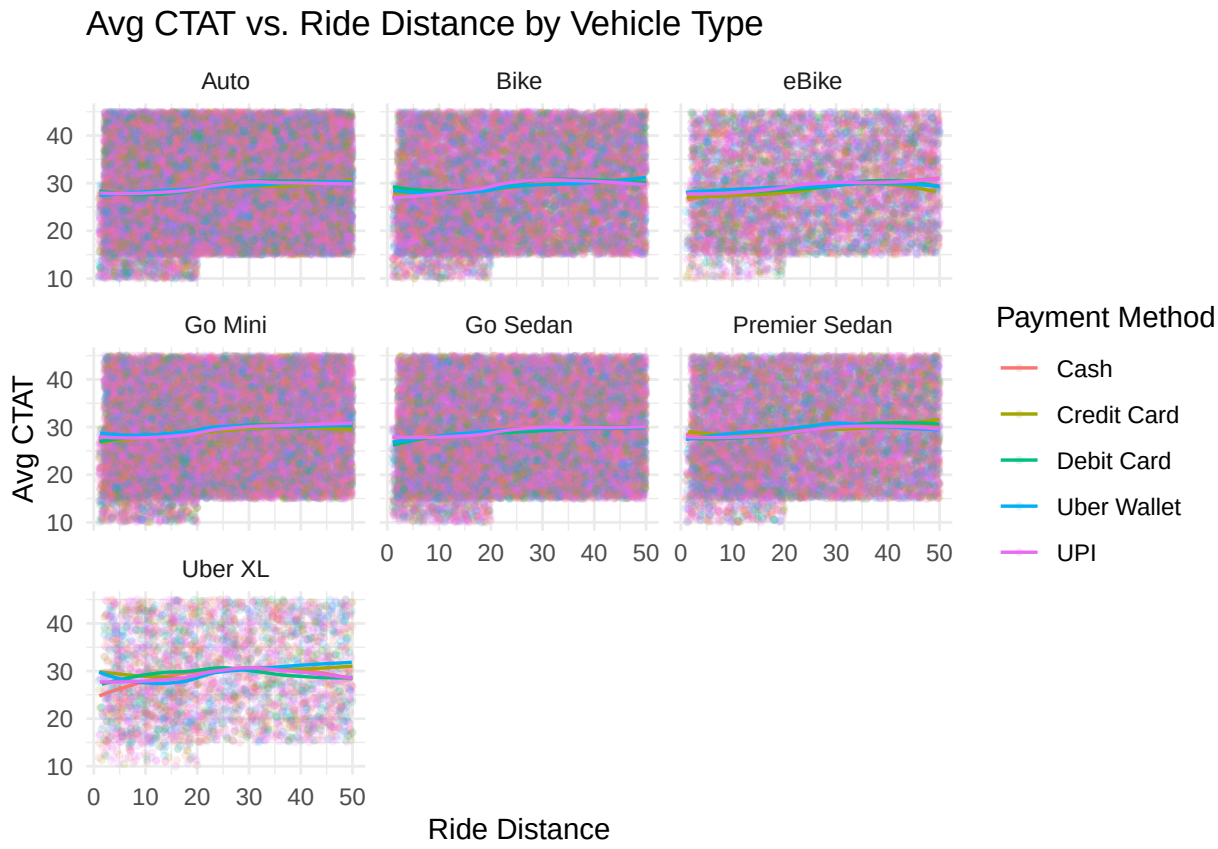


Interpretation: The distributions across approaches are quite comparable: medians about 24-27, and the interquartile ranges are approximately 15-35. The differences are minimal, with Cash/Uber Wallet seeming somewhat elevated compared to Debit; nonetheless, the overlap is substantial. The extensive variation within each method indicates that the payment method is a poor predictor of distance. A distinct bulk around brief excursions and an upper limit approaching 50, indicating a restriction or truncation in documented distance.

Task 4 - Multivariate Visualizations

```
# Example facets
# ggplot(data, aes(x = xvar, y = yvar, color = group)) +
#   geom_point(alpha = 0.6) +
#   facet_wrap(~ facet_var) +
#   labs(title = "Relationship by facet_var")

library(dplyr)
library(ggplot2)
library(readr)   # parse_number()

data_facet <- data %>%
  mutate(
    RideDistance = parse_number(`Ride Distance`),
    AvgCTAT      = parse_number(`Avg CTAT`)
  ) %>%
  filter(!is.na(RideDistance), !is.na(AvgCTAT),
```

```
        !is.na(`Payment Method`), !is.na(`Vehicle Type`))

ggplot(data_facet,
       aes(x = RideDistance, y = AvgCTAT, color = `Payment Method`)) +
  geom_point(alpha = 0.15, size = 0.8) +
  geom_smooth(method = "loess", se = FALSE, linewidth = 0.6) +
  facet_wrap(~ `Vehicle Type`) +
  labs(title = "Avg CTAT vs. Ride Distance by Vehicle Type",
       x = "Ride Distance", y = "Avg CTAT", color = "Payment Method") +
  theme_minimal()
```

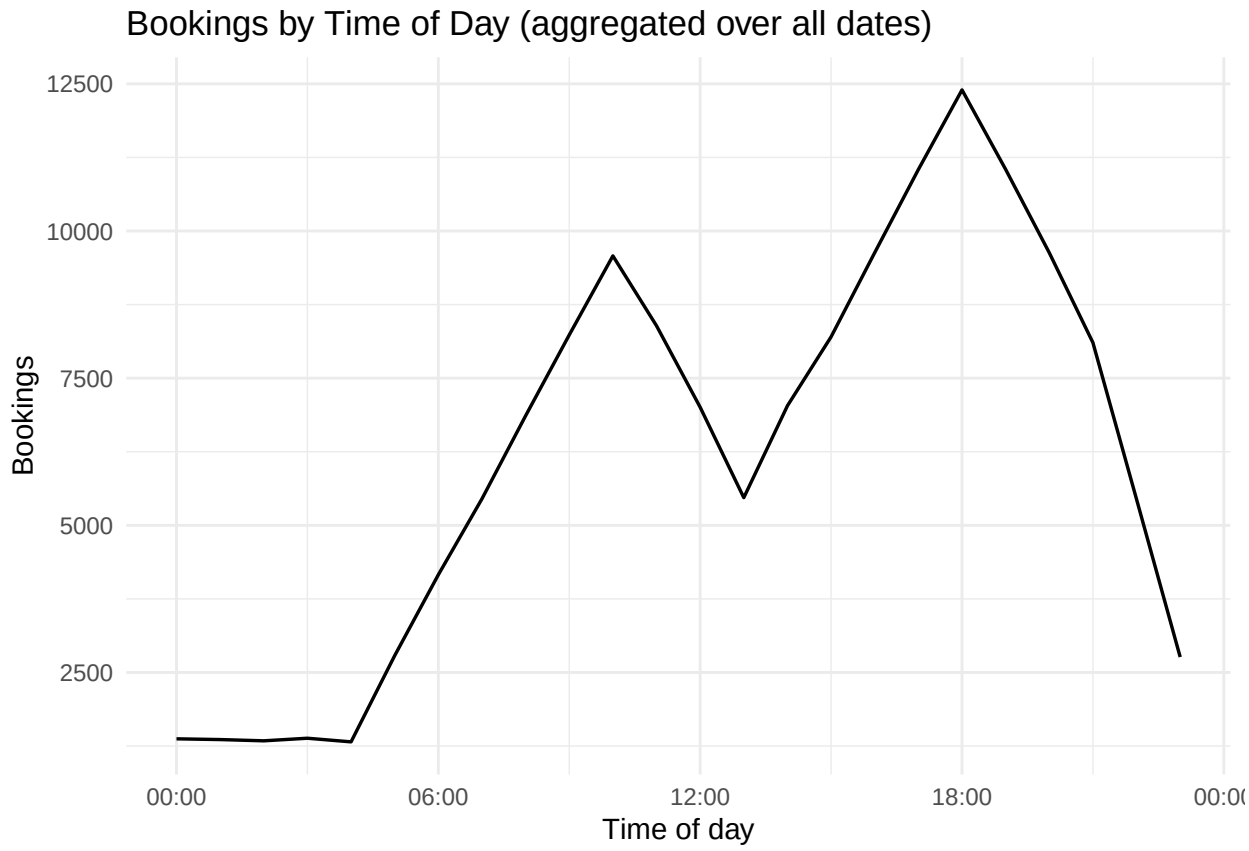


Short note: The average CTAT goes up a bit with Ride Distance for different vehicle types, but the effect is pretty small. The baseline CTAT varies a little depending on the vehicle, with sedans being slightly higher compared to bikes and autos. The payment method doesn't really stand out, which means it probably doesn't have a big impact on CTAT at a certain distance.

Task 5 - Temporal or Composition Analysis

```
# Example time series (if applicable)
# ggplot(data, aes(x = date_col, y = value)) +
#   geom_line(color = "#2b8cbe") +
#   labs(title = "Value over time", x = "Date", y = "Value")
```

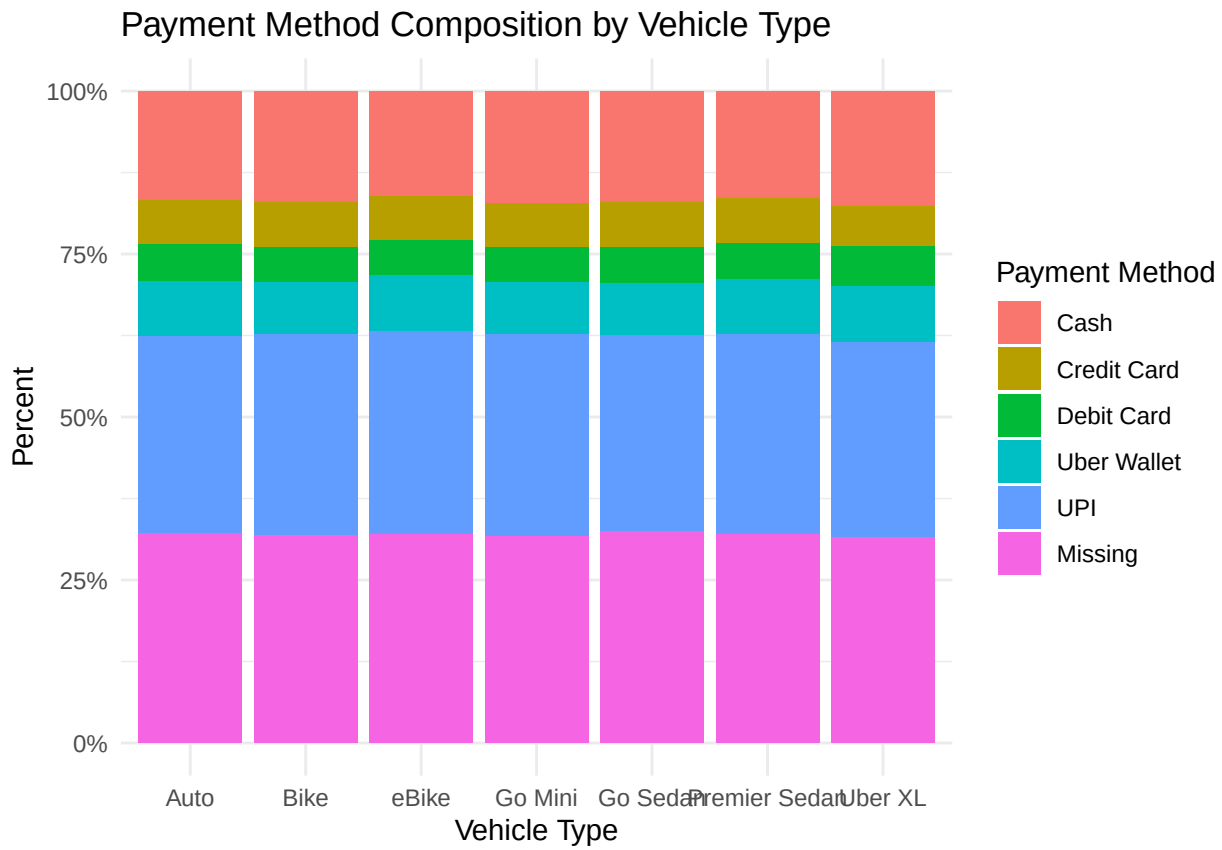
```r
library(dplyr)
library(ggplot2)
library(lubridate)
library(hms)

daily_cnt <- data %>%
  mutate(time_of_day = as_hms(Time),
         tod_hour = floor_date(as.POSIXct(time_of_day, origin = "1970-01-01", tz = "UTC"), "hour")) %>%
  count(tod_hour, name = "bookings")

ggplot(daily_cnt, aes(tod_hour, bookings)) +
  geom_line(linewidth = 0.6) +
  labs(title = "Bookings by Time of Day (aggregated over all dates)",
       x = "Time of day", y = "Bookings") +
  scale_x_datetime(date_labels = "%H:%M") +
  theme_minimal()
```



Alternative composition example if no time variable is present:

```r
# Example stacked percentage bar
# data %>% count(group, category) %>%
#   group_by(group) %>% mutate(p = n/sum(n)) %>% ungroup() %>%
#   ggplot(aes(x = group, y = p, fill = category)) +
#   geom_col() + scale_y_continuous(labels = scales::percent) +
#   labs(title = "Composition by group", y = "Percent", x = "Group")
```

```r
library(dplyr)
library(ggplot2)
library(forcats)
library(scales)

comp <- data %>%
  mutate(
    `Payment Method` = na_if(`Payment Method`, "null"),   # treat "null" as missing
    `Payment Method` = fct_explicit_na(`Payment Method`, "Missing")
  ) %>%
  count(`Vehicle Type`, `Payment Method`, name = "n") %>%
  group_by(`Vehicle Type`) %>%
  mutate(p = n / sum(n)) %>%
  ungroup()

ggplot(comp, aes(x = `Vehicle Type`, y = p, fill = `Payment Method`)) +
  geom_col() +
  scale_y_continuous(labels = percent_format()) +
  labs(title = "Payment Method Composition by Vehicle Type",
       x = "Vehicle Type", y = "Percent", fill = "Payment Method") +
  theme_minimal()
```



Task 6 - Geospatial (Bonus)

```
# Optional: Use sf/leaflet/ggplot2 for maps if you have location data
# library(sf)
# library(leaflet)
```

Task 7 - Aesthetics and Clarity

List the style choices you made (titles, labels, themes, color choices).

Global/recurring -theme_minimal() for a clean look. -Ordered categories with fct_reorder(...). -Transparent points (alpha ~ 0.1-0.6) to reduce overplotting. -Smooth trend lines via geom_smooth(method = "loess", se = FALSE, linewidth ~ 0.6). -Rotated bars with coord_flip() where helpful. Histogram - "Distribution of Booking Value" -Title: Distribution of Booking Value -Axes: x = "Price", y = "Count"; scale_x_continuous(labels = scales::comma) -Colors: Bars filled #2c7fb8, white border; binwidth = 100. Bar chart - "Counts by Payment Method" -Title: Counts by Payment Method -Axes: x = "Payment Methods", y = "Count" -Colors: geom_col(fill = "orange", color = "white") -Layout: coord_flip() to read labels easily. Scatter - "Avg VTAT by Time of Day" -Axes labels: x = "Time of day", y = "Avg VTAT (minutes)" -Theme: theme_minimal() (light grid, clean background) -Points: geom_point(shape = 16, size = 0.6, alpha = 0.08) (dense scatter, high transparency) -Trend line: geom_smooth(method = "loess", se = FALSE, linewidth = 0.8, color = "#fc8d59") -X scale: scale_x_time(labels = "%H:%M") -Y view window: coord_cartesian(ylim = quantile(AvgVTAT, c(.01, .99))) (focus without dropping data) -Colors: points in grayscale (default), trend line in #fc8d59 (soft orange) Violin + box - "Ride Distance by Payment Method -Title: Ride Distance by Payment Method -Axes: x ="Ride Distance", y = "Payment Method" -Colors: Violins fill = "#c7e9b4", color = "gray30"; boxplots white fill. Facets - "Avg CTAT vs. Ride Distance by Vehicle Type" -Title: Avg CTAT vs. Ride Distance by Vehicle Type -Mapping: color = Payment Method (ggplot default palette) -Faceting: facet_wrap(~ Vehicle Type). Temporal line - "Bookings by Time of Day (aggregated over all dates)" -Axes: x = "Time of day", y = "Bookings" -Style: geom_line(linewidth = 0.6). Composition (stacked 100%) - "Payment Method Composition by Vehicle Type" -Axes: x = "Vehicle Type", y = "Percent" -Legend: fill = "Payment Method" -Scale: scale_y_continuous(labels = scales::percent_format()) -Colors: Categorical fills per payment method (ggplot default palette) plus a "Missing" category.

Task 8 - Narrative and Reproducibility

Write 5-10 sentences connecting the plots into a cohesive story and add final insights.

Bookings have this pretty clear daily pattern: they're super low overnight, start to pick up after the morning, and then hit their highest point during the evening commute. Even though the volume changes a lot, the ride distance stays pretty consistent throughout the day, and the VTAT scatter shows a flat trend around 8–9 minutes—pickup latency isn't really affected by the time. Booking values are skewed to the right, meaning that most trips are priced low to mid-range, while there are some really expensive ones that stretch out the data. So, using medians and IQRs gives a clearer picture than just looking at the means. Payment habits are pretty similar no matter what kind of vehicle we're talking about. UPI and Wallets are the big players, cash still holds its ground, and cards are less popular. The distance distributions are pretty much the same for all payment methods, which means that how you pay doesn't really affect how long the trip is. CTAT goes up a bit with distance in each vehicle class, but the baselines vary a little by class (bikes and autos are lower, while sedans and XLs are higher), which suggests that there are some operational factors at the vehicle level.