Week 3 Tasks — Data Preparation, EDA, and Intro Modeling (R)

Your Name

Instructions

- Use this template to complete Week 3 tasks. Replace placeholders with your work.
- Ensure the document can knit top-to-bottom without errors.
- Add short captions/annotations below each plot and metric output.

Task 1 — Load Data and Inspect

```
# TODO: Set path and read your dataset
# Example:
# data <- read_csv("path/to/your.csv")
# head(data, 10)
# dim(data)

# Load NCR ride bookings dataset
data <- read_csv("ncr_ride_bookings.csv")

# Preview and dimensions
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>, ...
```

```r
dim(data)
```

```
## [1] 150000     21
```

Briefly describe the dataset, its purpose, and key variables.

Task 2 — Data Types, Summary Stats, and Missingness

```r
# Glimpse types
# glimpse(data)

# Summary statistics (numeric and categorical)
# summary(data)

# Missingness per column
# tibble(col = names(data), n_missing = colSums(is.na(data))) %>%
#   mutate(p_missing = n_missing / nrow(data)) %>%
#   arrange(desc(p_missing))

# Glimpse types
glimpse(data)
```

```
## Rows: 150,000
## Columns: 21
## $ Date                                <date> 2024-03-23, 2024-11-29, 2024-08-2~
## $ Time                                <time> 12:29:38, 18:01:39, 08:56:10, 17:~
## $ `Booking ID`                        <chr> "\"CNR5884300\"", "\"CNR1326809\""~
## $ `Booking Status`                    <chr> "No Driver Found", "Incomplete", "~
## $ `Customer ID`                       <chr> "\"CID1982111\"", "\"CID4604802\""~
## $ `Vehicle Type`                      <chr> "eBike", "Go Sedan", "Auto", "Prem~
## $ `Pickup Location`                   <chr> "Palam Vihar", "Shastri Nagar", "K~
## $ `Drop Location`                     <chr> "Jhilmil", "Gurgaon Sector 56", "M~
## $ `Avg VTAT`                          <chr> "null", "4.9", "13.4", "13.1", "5.~
## $ `Avg CTAT`                          <chr> "null", "14.0", "25.8", "28.5", "1~
## $ `Cancelled Rides by Customer`       <chr> "null", "null", "null", "null", "n~
## $ `Reason for cancelling by Customer` <chr> "null", "null", "null", "null", "n~
## $ `Cancelled Rides by Driver`         <chr> "null", "null", "null", "null", "n~
## $ `Driver Cancellation Reason`        <chr> "null", "null", "null", "null", "n~
## $ `Incomplete Rides`                  <chr> "null", "1", "null", "null", "null~
## $ `Incomplete Rides Reason`           <chr> "null", "Vehicle Breakdown", "null~
## $ `Booking Value`                     <chr> "null", "237", "627", "416", "737"~
## $ `Ride Distance`                     <chr> "null", "5.73", "13.58", "34.02", ~
## $ `Driver Ratings`                    <chr> "null", "null", "4.9", "4.6", "4.1~
## $ `Customer Rating`                   <chr> "null", "null", "4.9", "5.0", "4.3~
## $ `Payment Method`                    <chr> "null", "UPI", "Debit Card", "UPI"~
```

```r
# Summary statistics
summary(data)
```

```
##       Date                 Time                   Booking ID
##  Min.   :2024-01-01   Min.   :00:00:00.000000   Length:150000
##  1st Qu.:2024-03-31   1st Qu.:10:20:25.000000   Class :character
##  Median :2024-07-01   Median :15:23:33.000000   Mode  :character
##  Mean   :2024-06-30   Mean   :14:32:00.931033
##  3rd Qu.:2024-09-30   3rd Qu.:18:57:31.000000
##  Max.   :2024-12-30   Max.   :23:59:59.000000
##  Booking Status     Customer ID        Vehicle Type       Pickup Location
##  Length:150000      Length:150000      Length:150000      Length:150000
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  Drop Location        Avg VTAT           Avg CTAT
##  Length:150000      Length:150000      Length:150000
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##  Cancelled Rides by Customer Reason for cancelling by Customer
##  Length:150000               Length:150000
##  Class :character            Class :character
##  Mode  :character            Mode  :character
##
##
##
##  Cancelled Rides by Driver Driver Cancellation Reason Incomplete Rides
##  Length:150000             Length:150000              Length:150000
##  Class :character          Class :character           Class :character
##  Mode  :character          Mode  :character           Mode  :character
##
##
##
##  Incomplete Rides Reason Booking Value      Ride Distance
##  Length:150000           Length:150000      Length:150000
##  Class :character        Class :character   Class :character
##  Mode  :character        Mode  :character   Mode  :character
##
##
##
##  Driver Ratings     Customer Rating    Payment Method
##  Length:150000      Length:150000      Length:150000
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
```

```r
# Missingness per column
missing_tbl <- tibble(
  col = names(data),
  n_missing = colSums(is.na(data))
```

```
) %>%
  mutate(p_missing = n_missing / nrow(data)) %>%
  arrange(desc(p_missing))
missing_tbl
```

```
## # A tibble: 21 x 3
##    col              n_missing p_missing
##    <chr>                <dbl>     <dbl>
##  1 Date                     0         0
##  2 Time                     0         0
##  3 Booking ID               0         0
##  4 Booking Status           0         0
##  5 Customer ID              0         0
##  6 Vehicle Type             0         0
##  7 Pickup Location          0         0
##  8 Drop Location            0         0
##  9 Avg VTAT                 0         0
## 10 Avg CTAT                 0         0
## # i 11 more rows
```

Task 3 — Data Cleaning

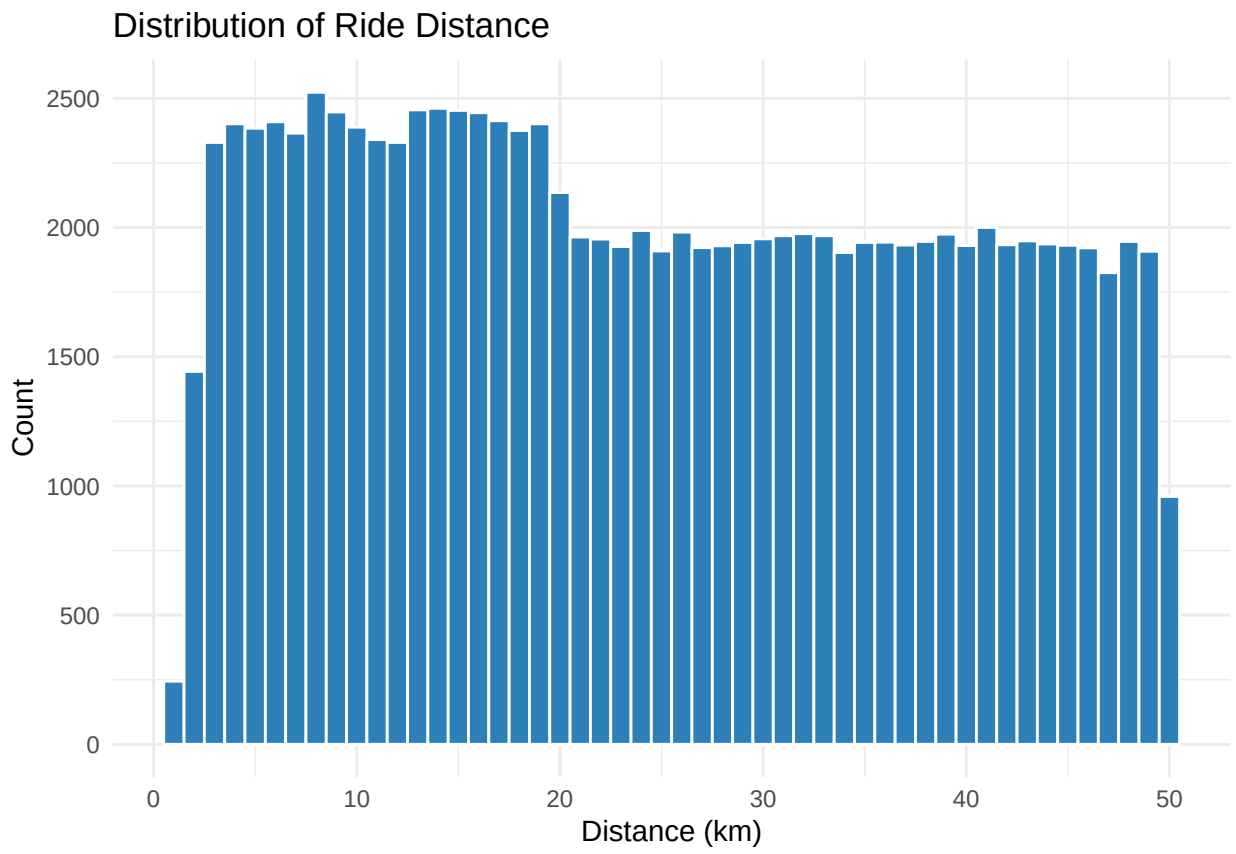
```
# Examples:
# data <- distinct(data)
# names(data) <- names(data) %>%
#   tolower() %>% str_replace_all(" ", "_") %>% str_trim()
# data <- data %>% mutate(num_col = if_else(is.na(num_col), median(num_col, na.rm = TRUE), num_col))

# Remove duplicate rows
data <- distinct(data)

# Standardize column names
names(data) <- names(data) %>%
  str_to_lower() %>%
  str_replace_all(" ", "_") %>%
  str_replace_all("[^a-z0-9_]", "")

# Trim whitespace
data <- data %>%
  mutate(across(where(is.character), str_squish))

# Simple imputation
data <- data %>%
  mutate(across(where(is.numeric),
                ~ ifelse(is.na(.x), median(.x, na.rm = TRUE), .x)))

# Check missingness again
colSums(is.na(data))
```

```
##                             date                             time
```

```
##                                0                                0
##                       booking_id                   booking_status
##                                0                                0
##                      customer_id                     vehicle_type
##                                0                                0
##                  pickup_location                    drop_location
##                                0                                0
##                         avg_vtat                         avg_ctat
##                                0                                0
##       cancelled_rides_by_customer reason_for_cancelling_by_customer
##                                0                                0
##         cancelled_rides_by_driver        driver_cancellation_reason
##                                0                                0
##                 incomplete_rides           incomplete_rides_reason
##                                0                                0
##                    booking_value                    ride_distance
##                                0                                0
##                    driver_ratings                   customer_rating
##                                0                                0
##                   payment_method
##                                0
```

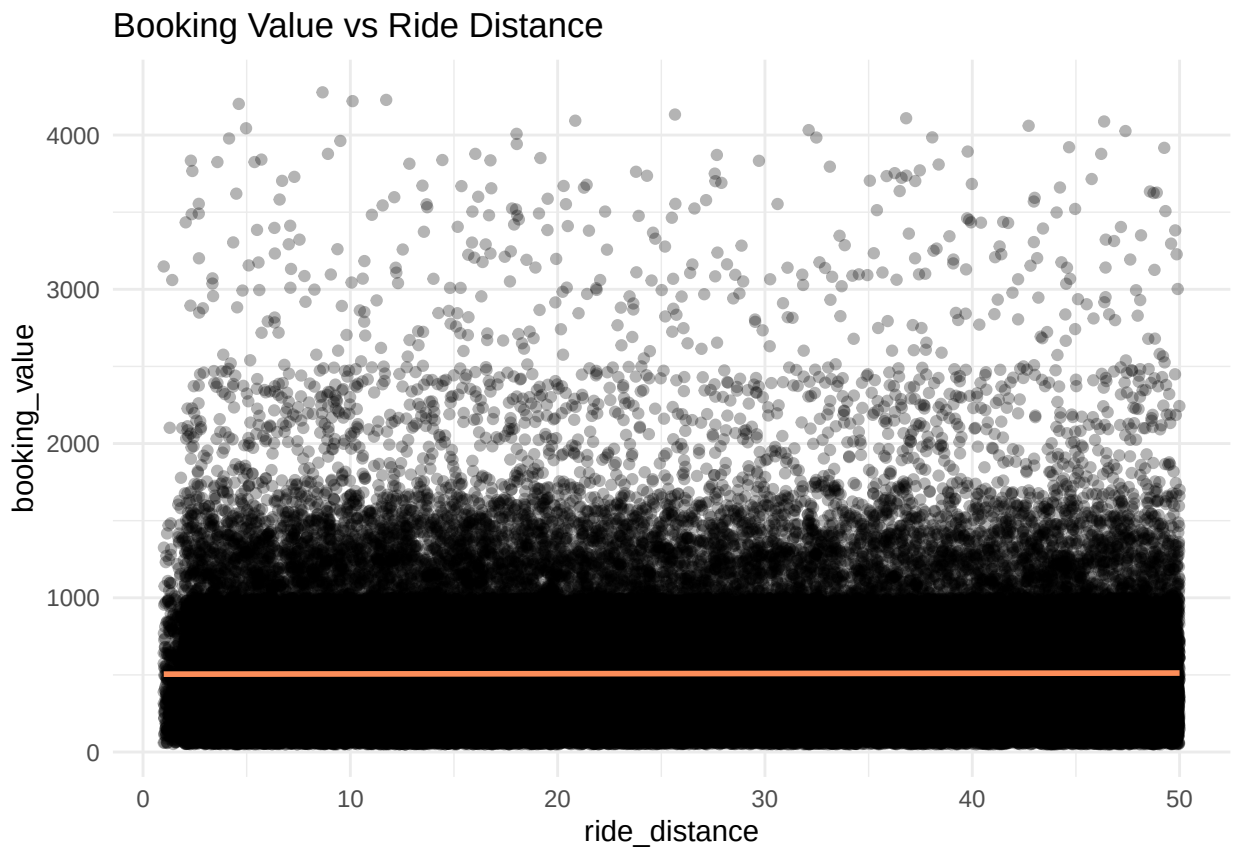
Task 4 — Exploratory Data Analysis (EDA)

```r
# Example univariate
# data %>% ggplot(aes(numeric_col)) + geom_histogram(binwidth = 5, fill = "#2c7fb8") +
#   labs(title = "Distribution of numeric_col", x = "numeric_col", y = "Count")

# Example bivariate
# 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)

data <- data %>%
  mutate(
    ride_distance = parse_number(as.character(`ride_distance`)),
    booking_value = parse_number(as.character(`booking_value`))
  )

# Univariate: Ride Distance
ggplot(data, aes(x = ride_distance)) +
  geom_histogram(binwidth = 1, fill = "#2c7fb8", color = "white") +
  labs(title = "Distribution of Ride Distance", x = "Distance (km)", y = "Count")
```

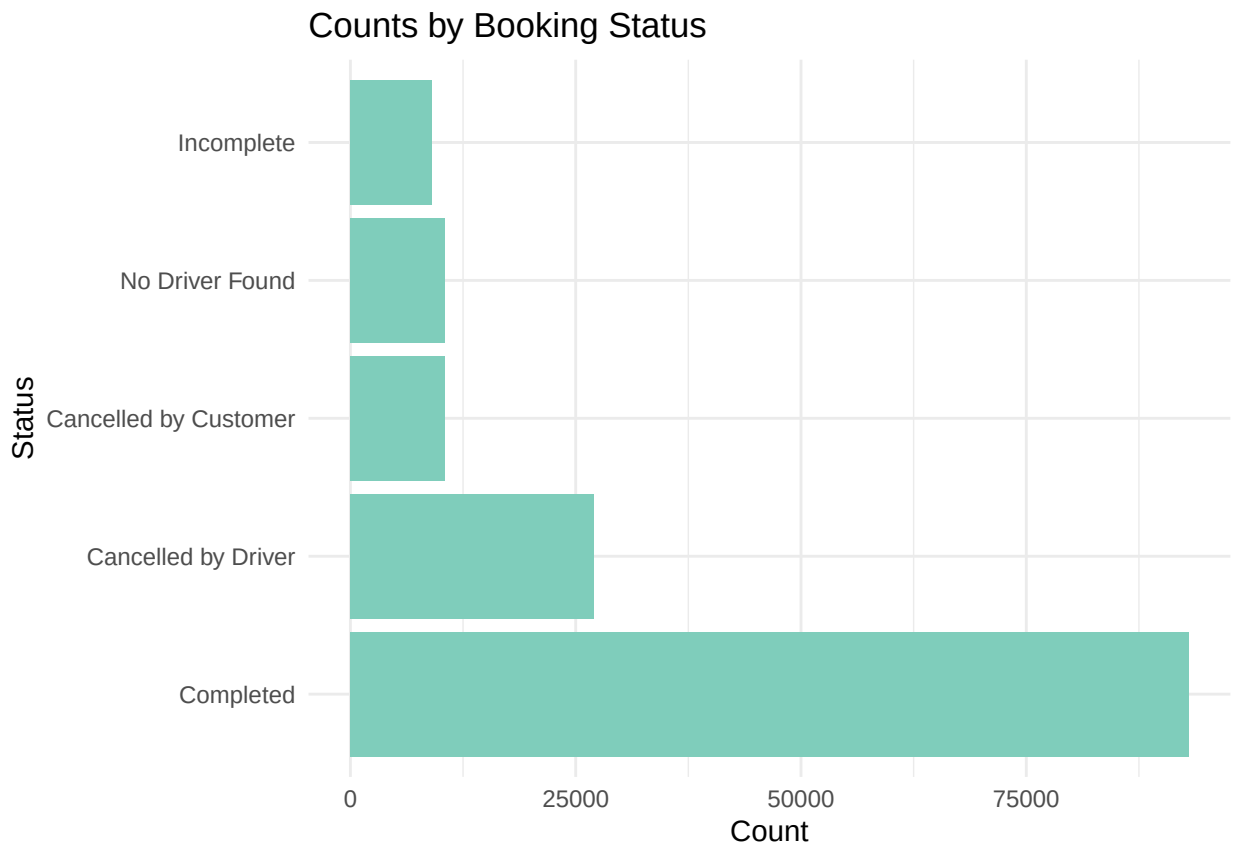



```
# Univariate: Booking Value
ggplot(data, aes(x = booking_value)) +
  geom_histogram(binwidth = 10, fill = "#41b6c4", color = "white") +
  labs(title = "Distribution of Booking Value", x = "Value", y = "Count")
```


```
# Bivariate: Booking Value vs Ride Distance
ggplot(data, aes(x = ride_distance, y = booking_value)) +
  geom_point(alpha = 0.3) +
  geom_smooth(method = "lm", se = FALSE, color = "#fc8d59") +
  labs(title = "Booking Value vs Ride Distance")
```




```
# Categorical: Booking Status counts
data %>%
  count(booking_status) %>%
  ggplot(aes(x = reorder(booking_status, -n), y = n)) +
  geom_col(fill = "#7fcdbb") +
  coord_flip() +
  labs(title = "Counts by Booking Status", x = "Status", y = "Count")
```

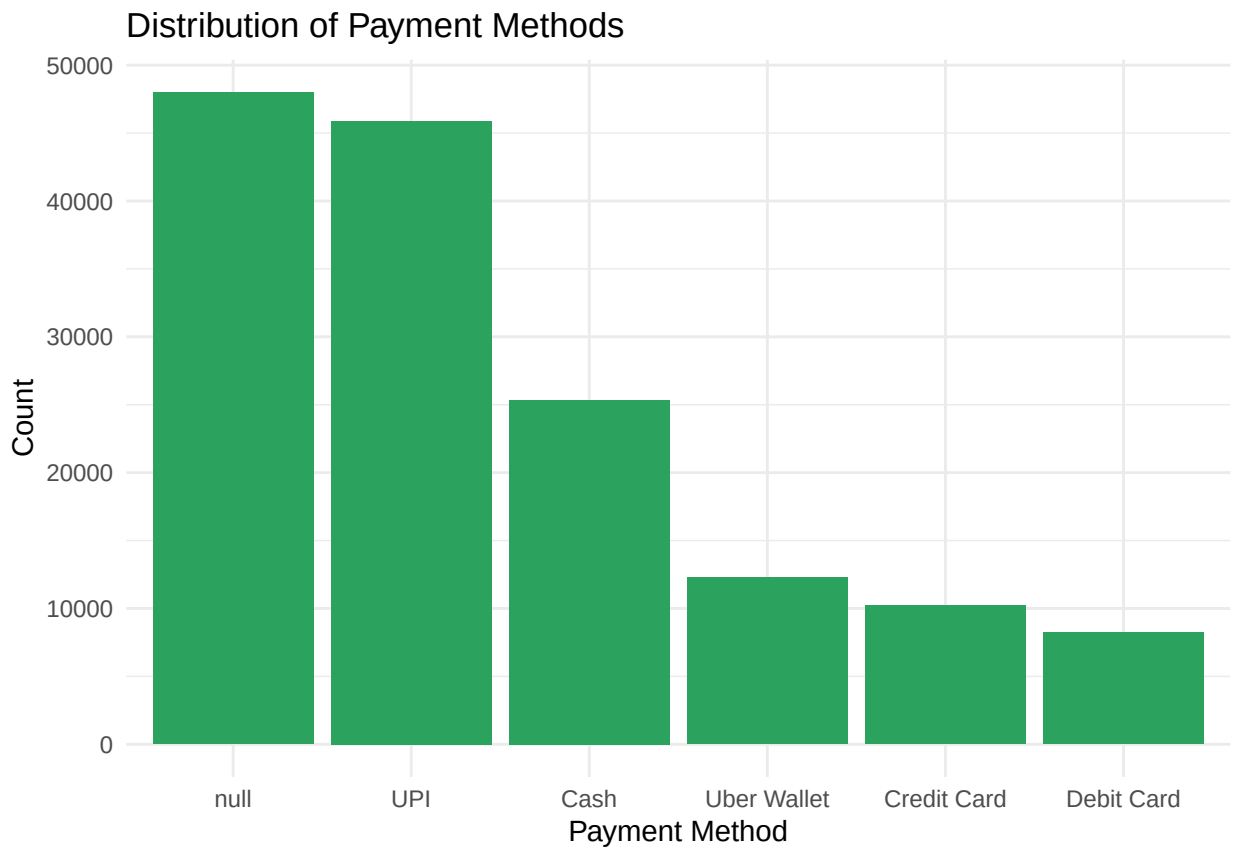


Write 3–5 short observations from EDA here.

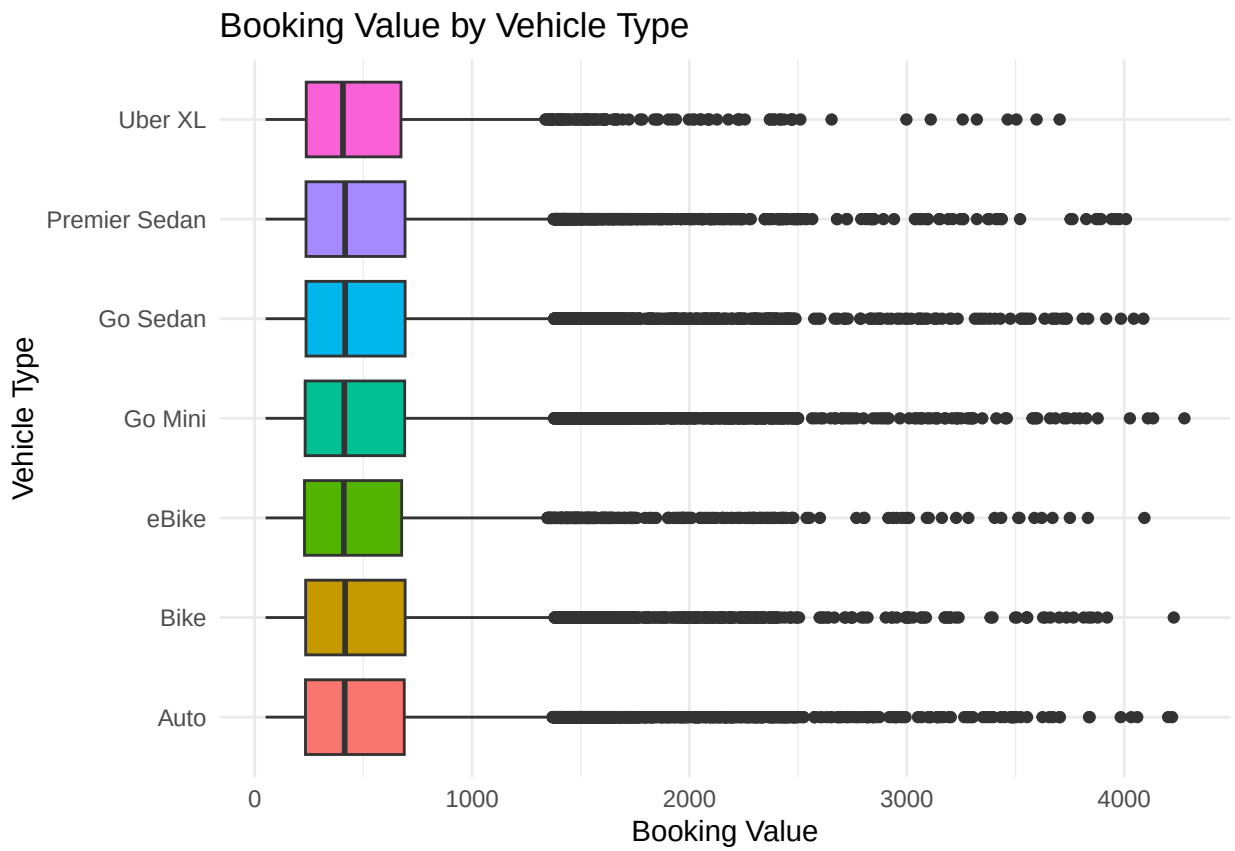
Ride distances are mostly short, with a few long rides. Booking value is skewed, with most fares below 300 units. There is a positive correlation between ride distance and booking value. Majority of bookings are Completed, while Cancelled rides form a smaller share.

Task 5 — Data Visualization

```
# Create at least 2 clear plots with captions below
# 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")

# Plot 1: Payment Method distribution
data %>%
  count(payment_method) %>%
  ggplot(aes(x = reorder(payment_method, -n), y = n)) +
  geom_col(fill = "#2ca25f") +
  labs(title = "Distribution of Payment Methods", x = "Payment Method", y = "Count")
```




```r
# Plot 2: Boxplot of Booking Value by Vehicle Type
ggplot(data, aes(x = vehicle_type, y = booking_value, fill = vehicle_type)) +
  geom_boxplot() +
  labs(title = "Booking Value by Vehicle Type", x = "Vehicle Type", y = "Booking Value") +
  guides(fill = "none") +
  coord_flip()
```



Key takeaway: Cash is the dominant payment method, followed by digital payments. SUVs and Sedans have higher median booking values compared to Auto Rickshaws. Vehicle type clearly influences ride fares.