Coffee Sales

Abishek Chaudhari

```
#Add essential libraries
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(dplyr)
library(lubridate)
data = read.csv("D:/My_DS_Internship/Coffee_sales/index.csv")
names(data)
```

```
## [1] "date"        "datetime"    "cash_type"   "card"        "money"
## [6] "coffee_name"
```

Data Preprocessing

```
data = read.csv("D:/My_DS_Internship/Coffee_sales/index.csv") #read the csv file and store in data

data$coffee_name <- factor(data$coffee_name) #convert the column coffee_name into factors

data <- data %>% select(-card) #remove the card column since it is of no use

data$date <- ymd_hms(data$datetime) #convert the datetime column into standard date time representation

data <- data %>% select(-datetime) #remove column datetime

#Add two columns for Month and Hour for each day
data <- data%>% mutate(Month = format(data$date, "%Y-%m"),
                       Hour = as.numeric(format(data$date, "%H")))
#Check the overall summary of dataset
summary(data)
```

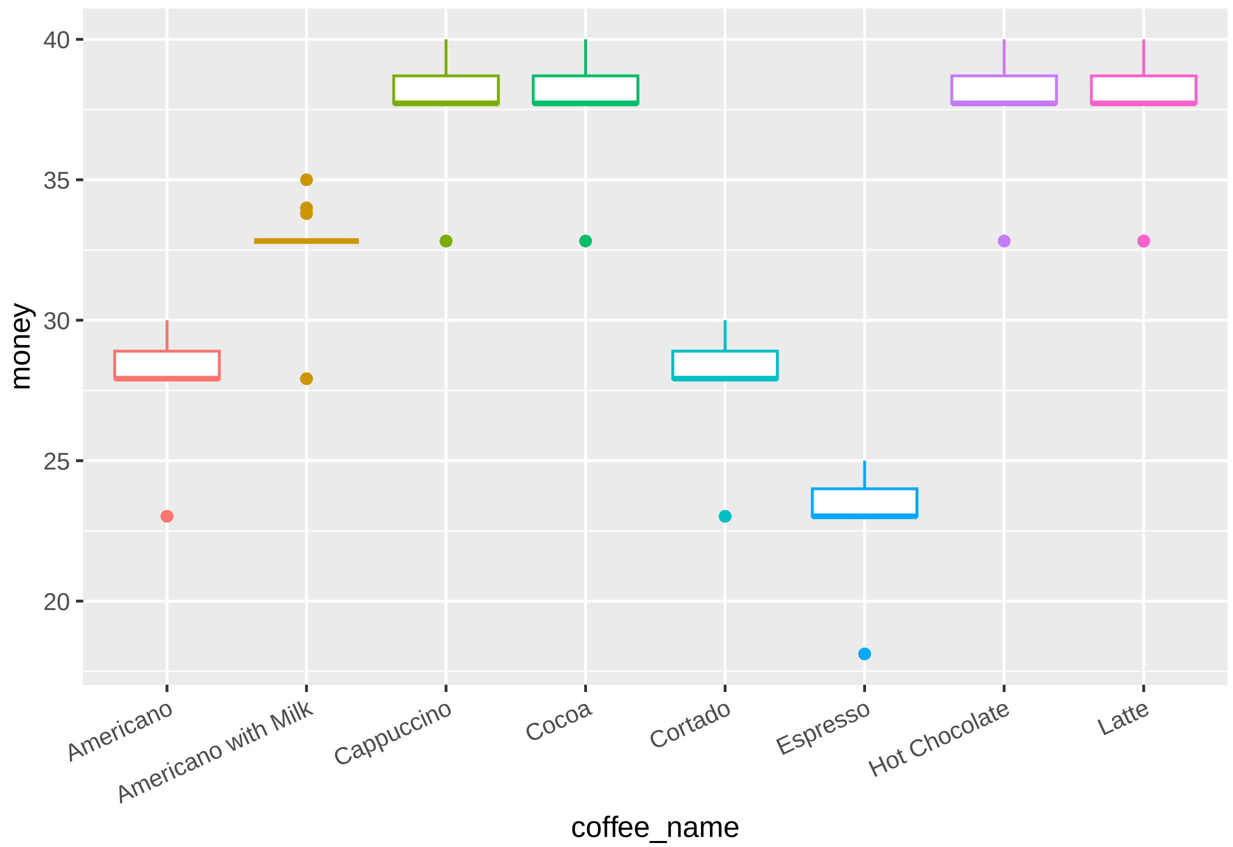
```
##       date                         cash_type             money
##  Min.   :2024-03-01 10:15:50.51   Length:1133        Min.   :18.12
##  1st Qu.:2024-04-14 10:55:27.41   Class :character   1st Qu.:28.90
##  Median :2024-05-23 12:22:06.59   Mode  :character   Median :32.82
##  Mean   :2024-05-20 02:38:39.04                      Mean   :33.11
##  3rd Qu.:2024-06-22 08:39:50.26                      3rd Qu.:37.72
##  Max.   :2024-07-31 21:55:16.56                      Max.   :40.00
##
##               coffee_name     Month                Hour
##  Americano with Milk:268   Length:1133        Min.   : 7.00
##  Latte              :243   Class :character   1st Qu.:11.00
##  Cappuccino         :196   Mode  :character   Median :14.00
##  Americano          :169                      Mean   :14.55
##  Cortado            : 99                      3rd Qu.:18.00
##  Hot Chocolate      : 74                      Max.   :22.00
##  (Other)            : 84
```

```
#Check if any column has null data
colSums(is.na(data))
```

```
##        date   cash_type       money coffee_name       Month        Hour
##           0           0           0           0           0           0
```

Exploratory Data Analysis 1. Outlier Detection

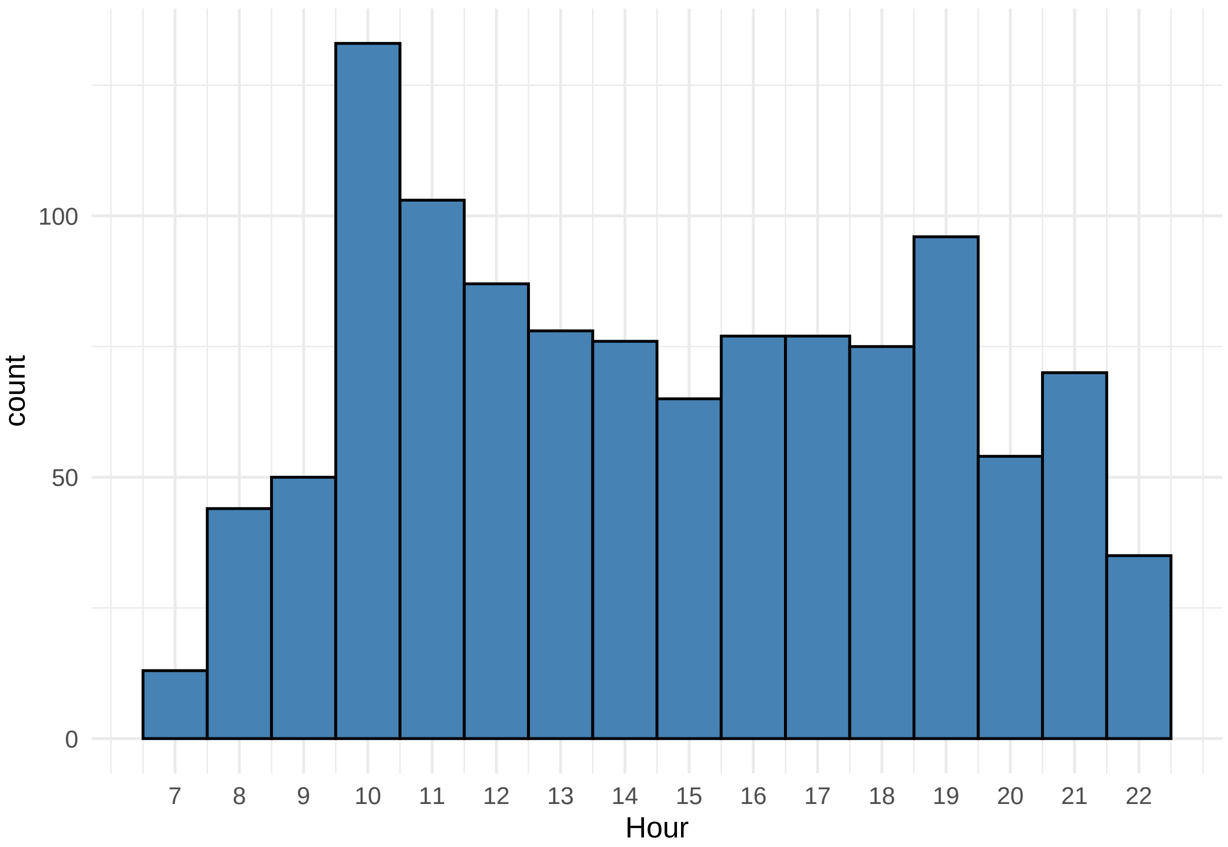
```
ggplot(data,aes(coffee_name,money, color = coffee_name))+
  geom_boxplot()+
  theme(axis.text.x = element_text(angle = 25, hjust = 1),
          legend.position = 'none')
```



In the above plot,none of the outliers are wildly out of sync with general coffee pricing. We're not seeing a coffee priced at 2 or 100. They are generally within a reasonable deviation.

2. Distribution of Sales over different hour of day

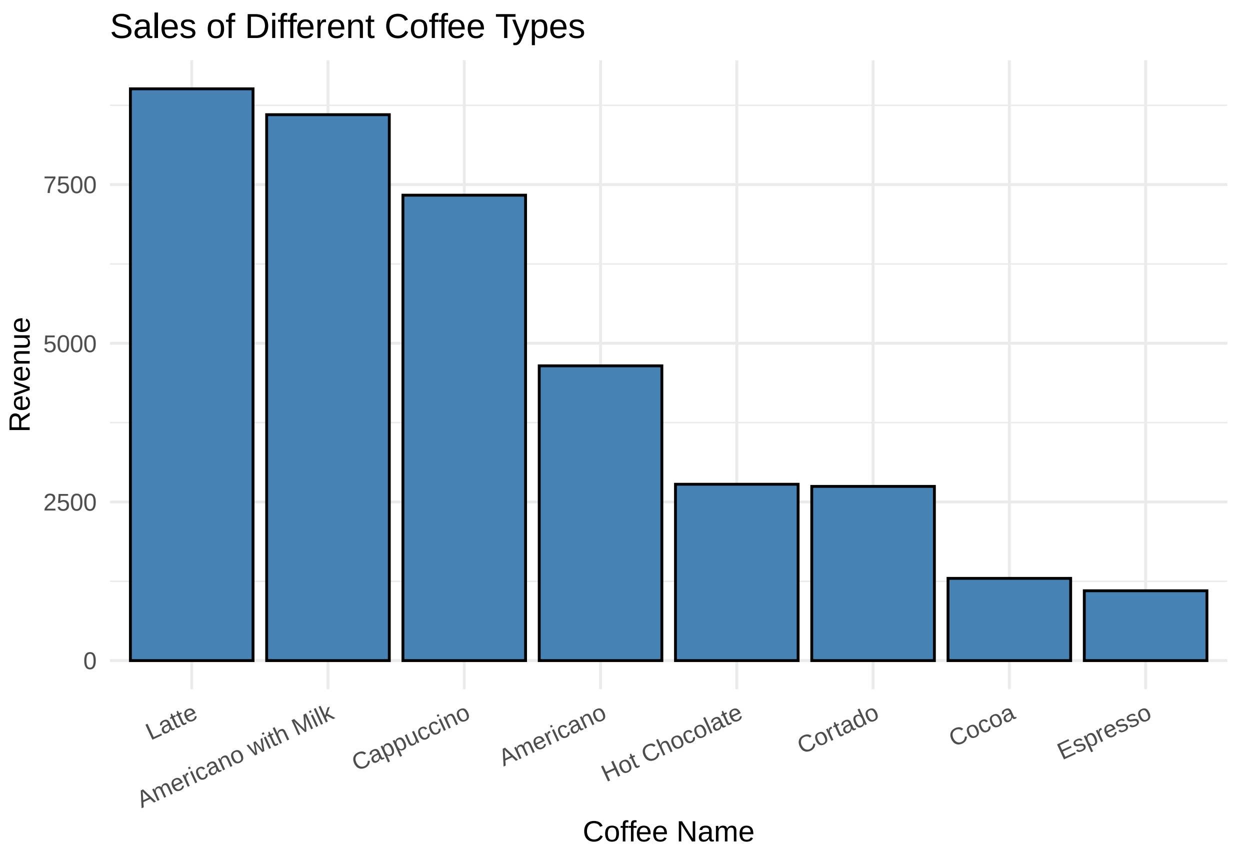
```
ggplot(data, aes(Hour)) +
  geom_histogram(binwidth = 1,color = "black" ,fill = 'steelblue') +
  theme_minimal()+
  scale_x_continuous(breaks = seq(min(data$Hour, na.rm=TRUE),
                                  max(data$Hour, na.rm=TRUE), by = 1))
```



We see that the sales of coffee is higher at two times of the day around 10 AM and 19 PM . Inventory of coffee can be managed as per this distribution.

3.Total Revenue Generated by Different Coffee Types

```
#Sales by Coffee_Name
#Plotting each coffee type and their total generated revenue over the full period
data %>% group_by(coffee_name)%>%
  summarize(total_sales = sum(money),count = n(),.groups = 'drop') %>%
  ggplot(aes(x = reorder(coffee_name,-total_sales),y = total_sales)) +
  geom_bar(stat = 'identity',color = 'black',fill = "steelblue") +
  labs(title = "Sales of Different Coffee Types",
       x = "Coffee Name",
       y = "Revenue")+
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 25, hjust = 1),
          legend.position = 'none')
```

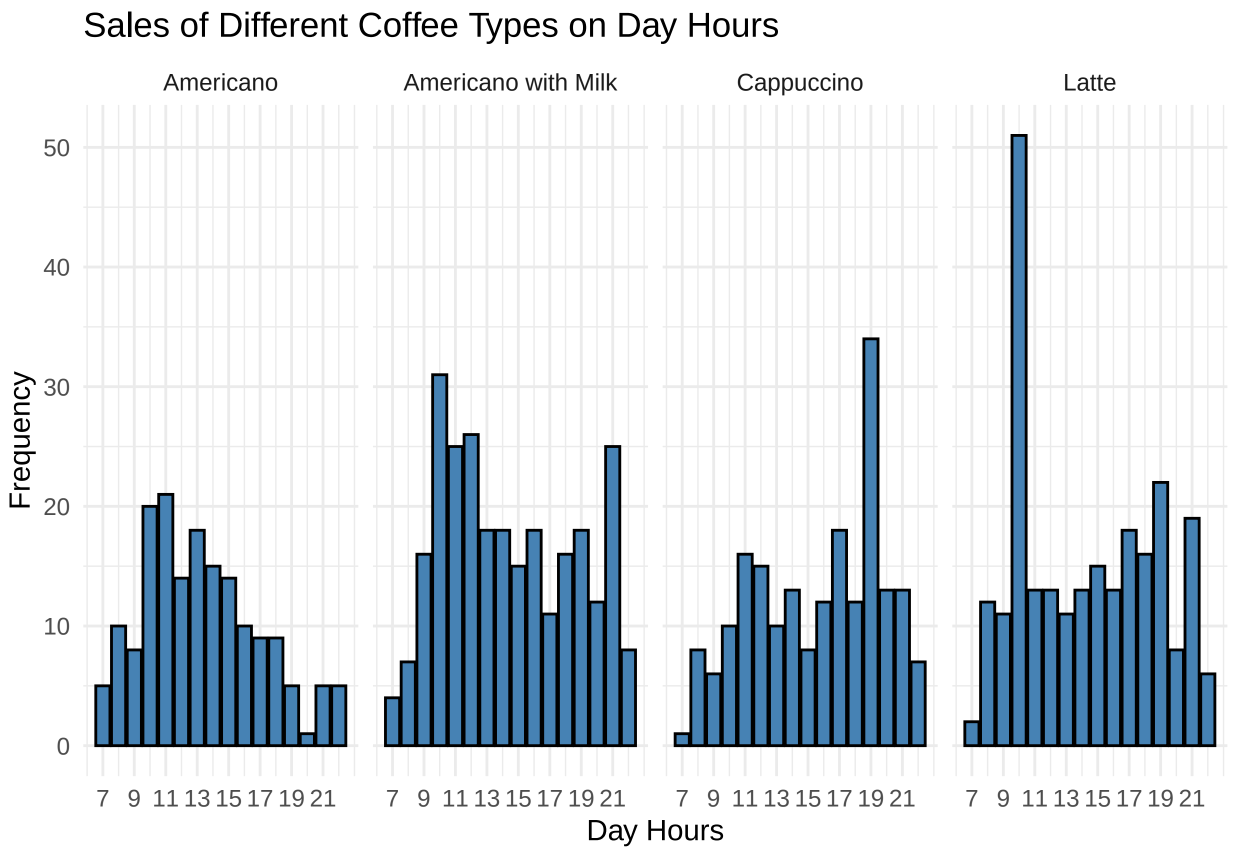


From the chart, it is clear that Latte, Americano with Milk and Cappuccino are the three most popular and highly sold coffee while Espresso is the least.

4.Coffee Sales on different Day Hours

To visualize the demand of different coffee on different time of the day, plots are generated and kept on the same row for comparison.Four coffee types have mean sales in a day greater than 10,these four are plotted below:

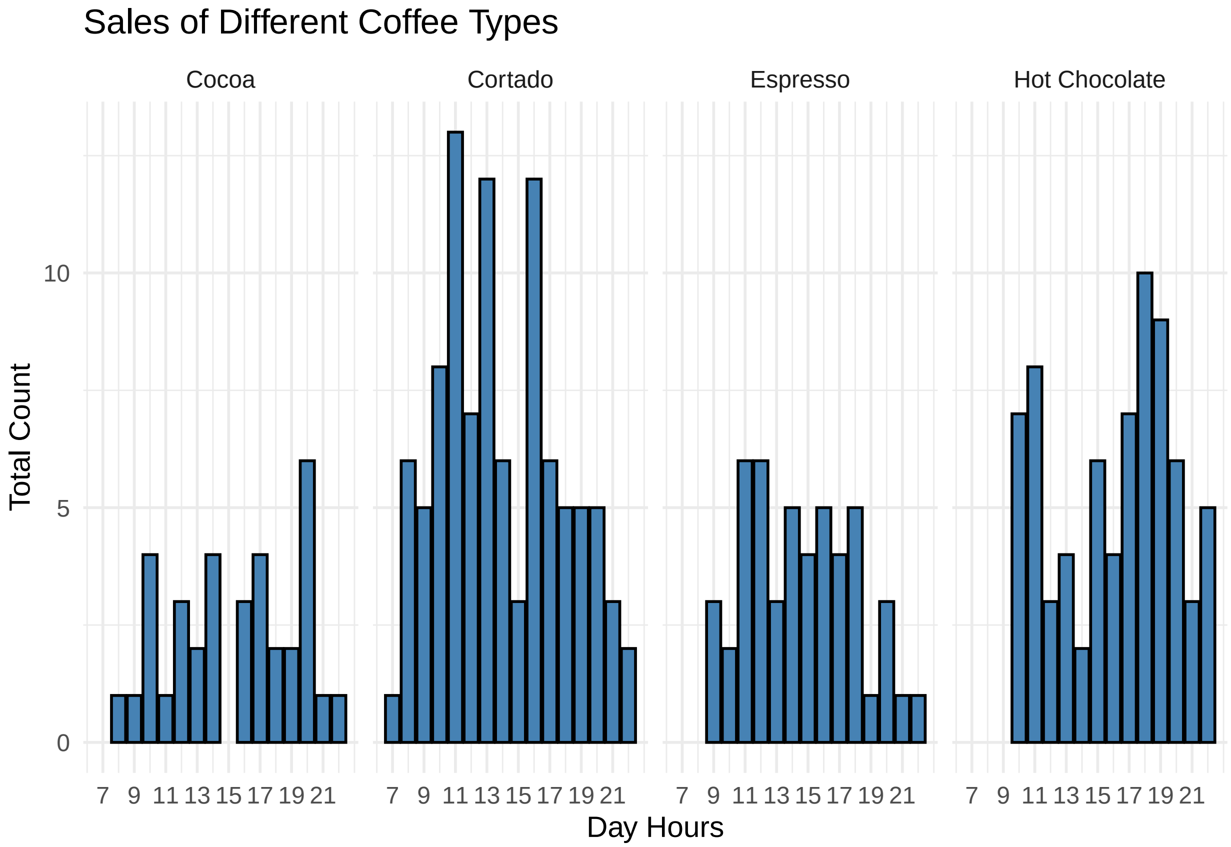
```
#Coffee sales trend
intermediate_data <- data %>%
  group_by(Hour,coffee_name)%>%
  summarize(total_count = n(),.groups = 'drop')%>%
  arrange(coffee_name,Hour)
#Sales Trend in hours of a day in top four coffee type
mean <- intermediate_data %>% group_by(coffee_name)%>%
  summarize(mean = mean(total_count))%>% arrange(desc(mean))
intermediate_data %>%
  arrange(coffee_name)%>%
  filter(coffee_name %in% mean$coffee_name[1:4])%>%
  ggplot(aes(x = Hour,y = total_count)) +
  geom_bar(stat = 'identity',color = 'black' ,fill = 'steelblue')+
  facet_wrap(.~ coffee_name, ncol = 4)+
  labs(title = "Sales of Different Coffee Types on Day Hours",
       x = "Day Hours",
       y = "Frequency")+
  theme_minimal() +
  scale_x_continuous(breaks = seq(min(data$Hour, na.rm=TRUE),
                                  max(data$Hour, na.rm=TRUE), by = 2))
```



The bar plots depict that the demand is higher during two times of the day, first at around 10 AM in the morning and another at around 19 PM in the evening.

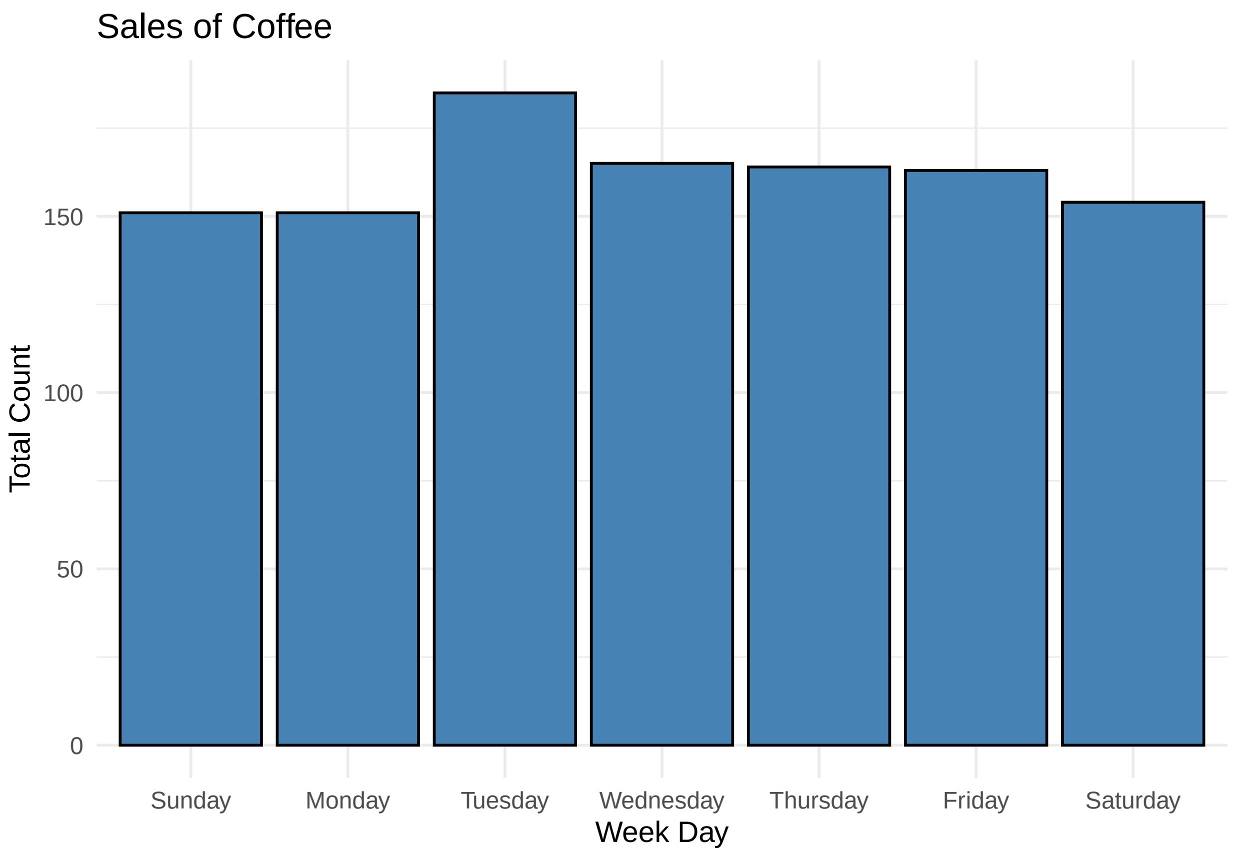
Below are the plots of coffee type with mean sales below 10 in a day:

```
#Sales Trend in hours of a day in bottom four coffee type

intermediate_data %>%
  arrange(coffee_name)%>%
  filter(coffee_name %in% mean$coffee_name[5:8])%>%
  ggplot(aes(x = Hour,y = total_count)) +
  geom_bar(stat = 'identity', color = 'black',fill = 'steelblue')+
  facet_wrap(.~ coffee_name, ncol = 4)+
  labs(title = "Sales of Different Coffee Types",
       x = "Day Hours",
       y = "Total Count")+
  theme_minimal() +
  scale_x_continuous(breaks = seq(min(data$Hour, na.rm=TRUE),
                                  max(data$Hour, na.rm=TRUE), by = 2))
```



5. Coffee Sales Over Days In a Week

```
#Coffee Sales report per week days in 4 months
day_data <- data %>%
  mutate(day = format(data$date,"%A"))%>%
  group_by(day)%>%
  summarize(total_count = n(),.groups = 'drop')
# Assuming your day column is called `day`
day_data$day <- factor(day_data$day, levels = c("Sunday", "Monday", "Tuesday",
                                    "Wednesday", "Thursday",
                                    "Friday", "Saturday"))

day_data %>%
  ggplot(aes(x = day,y = total_count)) +
  geom_bar(stat = 'identity', color = 'black',fill = 'steelblue')+
  #geom_text(aes(label = total_count), vjust = -0.5)+
  labs(title = "Sales of Coffee",
       x = "Week Day",
       y = "Total Count")+
  theme_minimal()
```

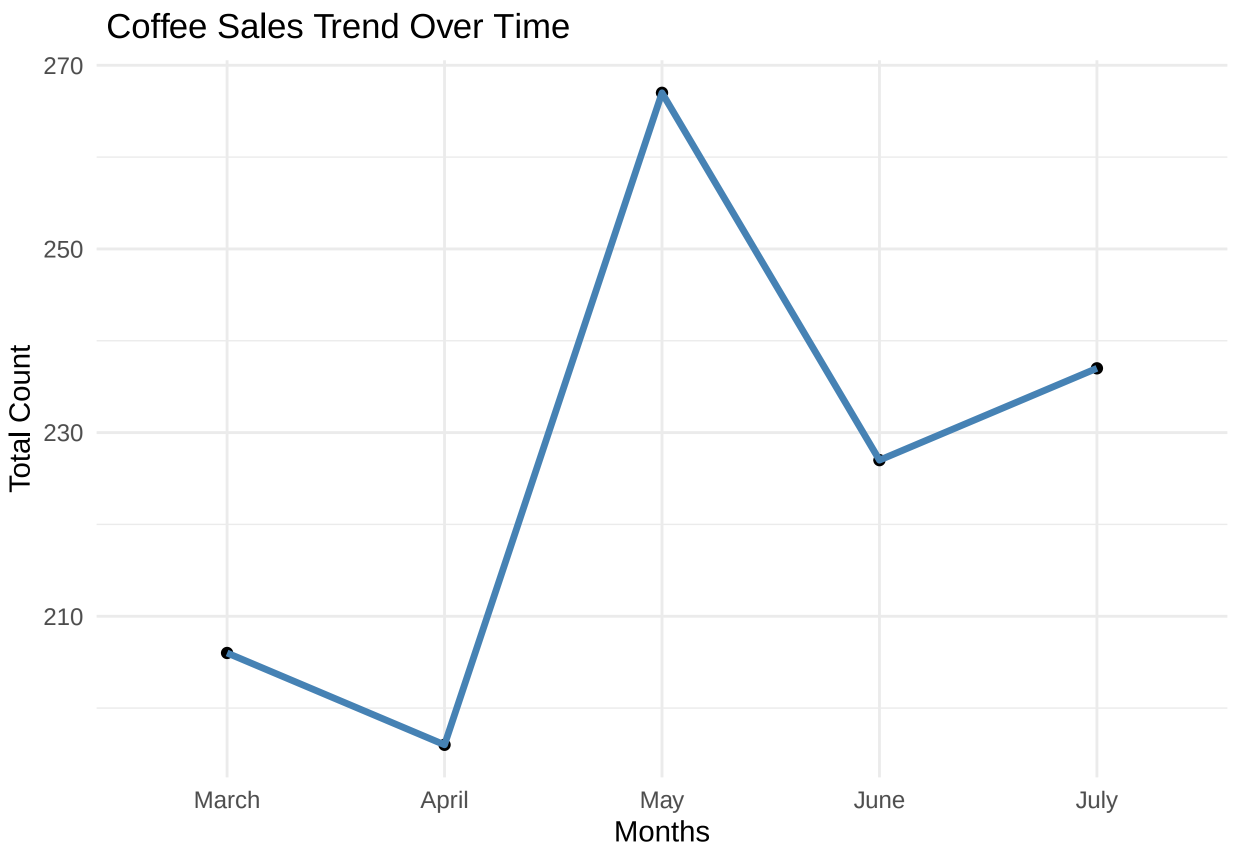


We see the highest demand for coffee on Tuesday among the seven days in a week. This information can be very beneficial for inventory planning.

6.Coffee Sales Trend Over Time

To see trend of sales of coffee each month, we plot the total orders of coffee every month in a line chart as shown below:

```r
#Coffee Sales Trend Over Months
data_month <- data %>% mutate(Month = format(data$date,"%B"))%>%
  group_by(Month)%>%
  summarize(total_count = n(),.groups = 'drop')
data_month$Month <- factor(data_month$Month, levels = c("March","April","May","June","July"))
data_month %>%
  ggplot(aes(x = Month,y = total_count,group = 1)) +
  geom_point()+
  geom_line(color = "steelblue", linewidth = 1.2)+
  labs(title = " Coffee Sales Trend Over Time",
       x = "Months",
       y = "Total Count")+
  theme_minimal()
```



The trend shows that the demand/orders have been increasing .

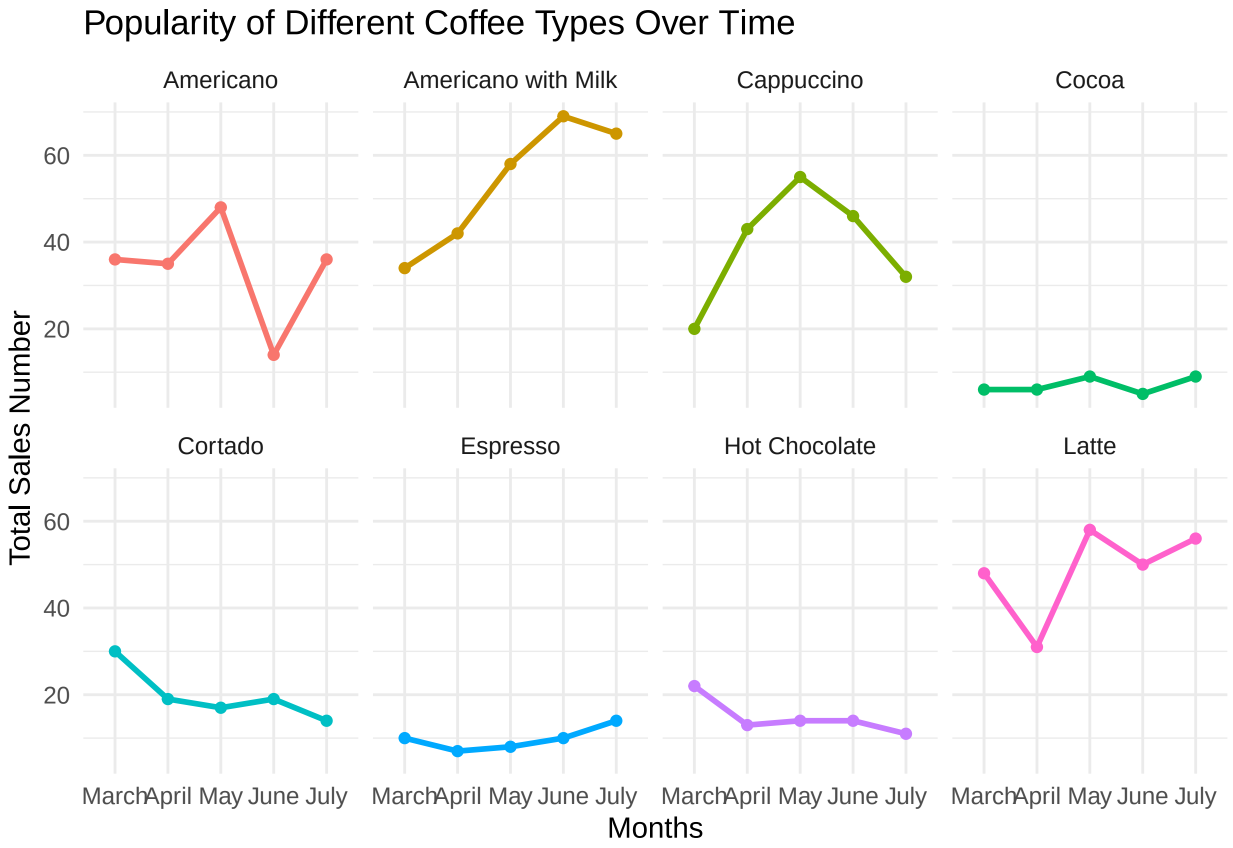
7.Popularity of Different Coffee Types Over Time

The increasing demand may be due to some particular coffee types and not by some other types. To find this, we plot a grid of line plots for each coffee type over all 5 months

Lets see, how is the trend for each coffee type

```r
#Coffee Sales Trend Over Months
data_month <- data %>% mutate(Month = format(data$date,"%B"))%>%
  group_by(Month,coffee_name)%>%
  summarize(total_count = n(),.groups = 'drop')%>%
  arrange(coffee_name)
data_month$Month <- factor(data_month$Month, levels = c("March","April","May","June","July"))

data_month %>%
  ggplot(aes(x = Month,y = total_count,color = coffee_name, group = coffee_name)) +
  geom_point()+
  geom_line(size = 1)+
  facet_wrap(.~coffee_name, ncol = 4)+
  labs(title = "Popularity of Different Coffee Types Over Time",
       x = "Months",
       y = "Total Sales Number")+
  theme_minimal()+
  theme(legend.position = 'none')
```

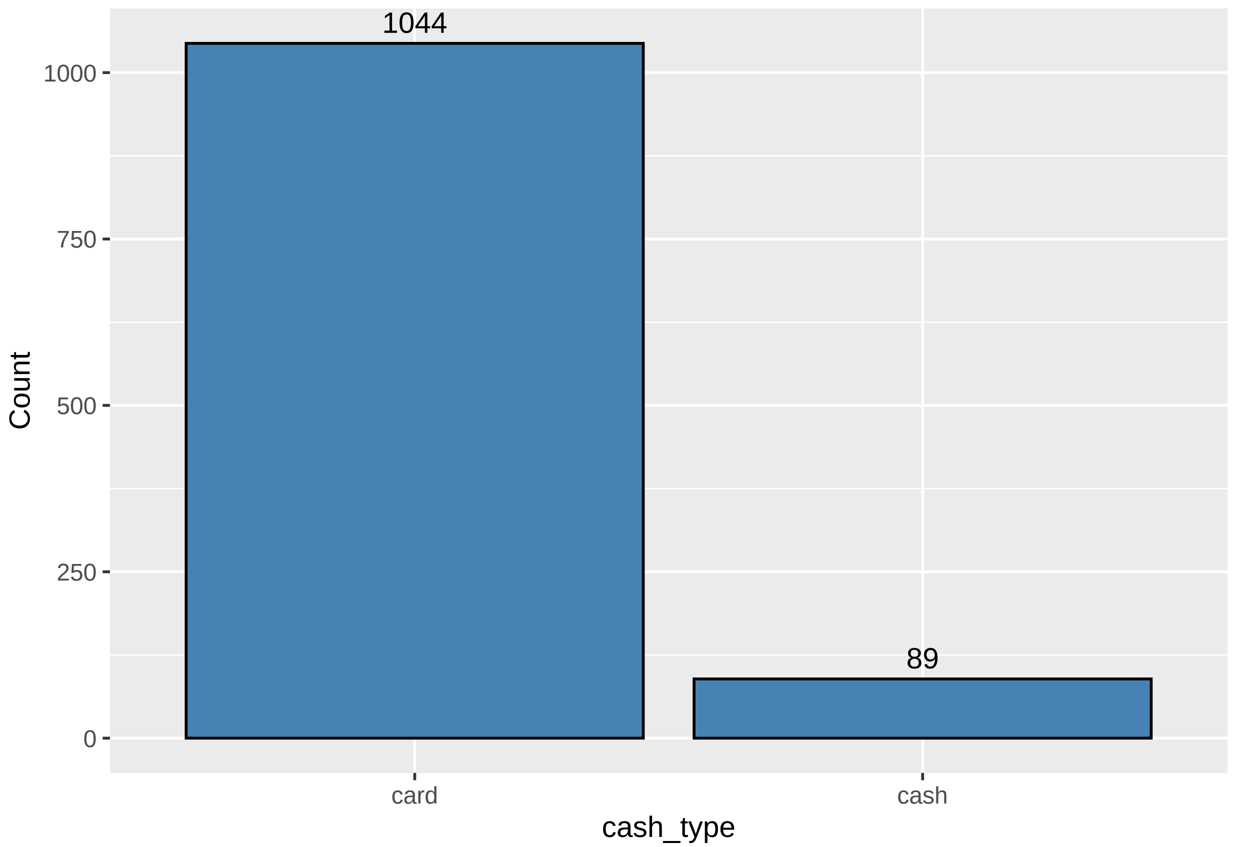
```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```



Popularity of ‘Americano with Milk’ and ‘Latte’ is increasing while ‘Cortado’ and ‘Hot Chocolate’ is decreasing. It means the overall increase in demand is particularly due to the ‘Americano with Milk’ and ‘Latte’ coffee. Others are neutral in behaviour.

8. Cash and Card Transaction

```
data%>% group_by(cash_type)%>%
  summarize(Count = n(),.groups = "drop")%>%
  ggplot(aes(cash_type,Count))+
  geom_bar(stat="identity", fill = "steelblue", color = "black")+
  geom_text(aes(label = Count), vjust = -0.5)
```



89 out of 1133 transactions are cash type. It is about 8 % of the total transactions.

SUMMARY

1.Coffee sales peak twice a day: around 10 AM and 7 PM.

2.Highest total revenue generated by ‘Latte’, ‘Americano with Milk’, and ‘Cappuccino’.

3.’Espresso’ generated the least revenue.

4.Top four coffee types show clear peak demand during 10 AM and 7 PM.

5.’Hot Chocolate’ is seen to have higher demand in evening.

6.Tuesday records the highest coffee sales among all weekdays.

7.Monthly sales show a rising trend from March to July.

8.’Americano with Milk’ and ‘Latte’ show increasing demand over the months.’Cortado’ and ‘Hot Chocolate’ show a decreasing trend in demand.

9.92% of transactions are card-based; only 8% are cash.

OVERALL INSIGHTS AND RECOMMENDATION

1.Focus on popular items (‘Latte’, ‘Americano with Milk’) for promotions and stocking.

2.Expect daily demand peaks at 10 AM and 7 PM.

3.Tuesdays see the most sales — leverage this for weekday offers.

4.Certain drinks are falling in popularity (e.g., Hot Chocolate) — consider re-evaluating their presence.

5.The overwhelming majority of card payments suggests less need for cash handling resources.