DESCRIPTIVE ANALYSIS ON WHEAT DATASET BASED ON ARRIVAL DATE

Description of data:

The dataset contains the data about the commodity of wheat sold in various states of India in the month of January 2023 from 1 to 27. The dataset contains the attributes are followed

- State
- District
- Market
- Commodity (Wheat)
- Variety - Variety of wheat
- Arrival date
- Min_price (minimum guaranteed price)
- Max_price (maximum price that the commodity purchased)
- Modal_price(Most frequently purchased price)
- Update_date

Objective of the analysis:

- The farmers want to know which state contributes the maximum frequent sales in the country and this is the sample correlation for the whole dataset.
- To know which part of the data contributes the most profit to the farmer based on the arrival date of the commodity.

Assumption:

- My assumption is that the most frequent state is the sample for the whole dataset based on the correlation value of min_price, max_price and modal_price.

```
#Importing the libraries for the analysis
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

library(lattice)
library(markdown)
#Importing the dataset
data=read.csv("Wheat_2023 DATA.csv")
#Internal structure of the data
str(data)

## 'data.frame':    9038 obs. of  10 variables:
##  $ state       : chr  "Bihar" "Bihar" "Bihar" "Bihar" ...
##  $ district    : chr  "Muzaffarpur" "Muzaffarpur" "Muzaffarpur" "Muzaffarpur" ...
##  $ market      : chr  "Muzaffarpur" "Muzaffarpur" "Muzaffarpur" "Muzaffarpur" ...
##  $ commodity   : chr  "Wheat" "Wheat" "Wheat" "Wheat" ...
```

```
##  $ variety     : chr  "147 Average" "147 Average" "147 Average" "147 Average" ...
##  $ arrival_date: chr  "03/01/2023" "04/01/2023" "05/01/2023" "09/01/2023" ...
##  $ min_price   : int  1950 2000 1950 2000 2000 2100 2100 2100 2700 2700 ...
##  $ max_price   : int  2800 2800 2800 2900 2900 2900 2900 2900 2900 2900 ...
##  $ modal_price : num  2500 2500 2500 2600 2550 2550 2550 2550 2800 2800 ...
##  $ update_date : chr  "2023-01-27" "2023-01-27" "2023-01-27" "2023-01-27" ...

#Statistical summary of the data
summary(data)

##     state              district           market           commodity
##  Length:9038        Length:9038        Length:9038        Length:9038
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    variety          arrival_date         min_price       max_price
##  Length:9038        Length:9038        Min.   :  245   Min.   :  680
##  Class :character   Class :character   1st Qu.: 2450   1st Qu.: 2620
##  Mode  :character   Mode  :character   Median : 2568   Median : 2750
##                                        Mean   : 2547   Mean   : 2772
##                                        3rd Qu.: 2680   3rd Qu.: 2870
##                                        Max.   :34000   Max.   :38000
##                                        NA's   :17      NA's   :23
##   modal_price    update_date
##  Min.   :  660   Length:9038
##  1st Qu.: 2560   Class :character
##  Median : 2655   Mode  :character
##  Mean   : 2663
##  3rd Qu.: 2760
##  Max.   :38000
##  NA's   :1

#Finding the NA values in the dataset
which(is.na(data$min_price))

##  [1] 1193 1200 1206 1213 1216 1830 1831 1836 1837 1843 1844 2312 2313 3792 3794
## [16] 5053 8544

which(is.na(data$max_price))

##  [1] 1193 1200 1206 1213 1216 1830 1831 1836 1837 1843 1844 1944 1945 1946 1947
## [16] 1948 1949 1950 1952 2312 2313 7190 8544

which(is.na(data$modal_price))

## [1] 5053

#Removing the NA values
data=na.omit(data)
summary(data)

##     state              district           market           commodity
##  Length:9012        Length:9012        Length:9012        Length:9012
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##    variety          arrival_date         min_price       max_price
##  Length:9012        Length:9012        Min.   :  245   Min.   :  680
```

```
##  Class :character   Class :character   1st Qu.: 2450   1st Qu.: 2620
##  Mode  :character   Mode  :character   Median : 2566   Median : 2750
##                                        Mean   : 2547   Mean   : 2772
##                                        3rd Qu.: 2678   3rd Qu.: 2870
##                                        Max.   :34000   Max.   :38000
##   modal_price    update_date
##  Min.   :  660   Length:9012
##  1st Qu.: 2560   Class :character
##  Median : 2655   Mode  :character
##  Mean   : 2664
##  3rd Qu.: 2760
##  Max.   :38000
```

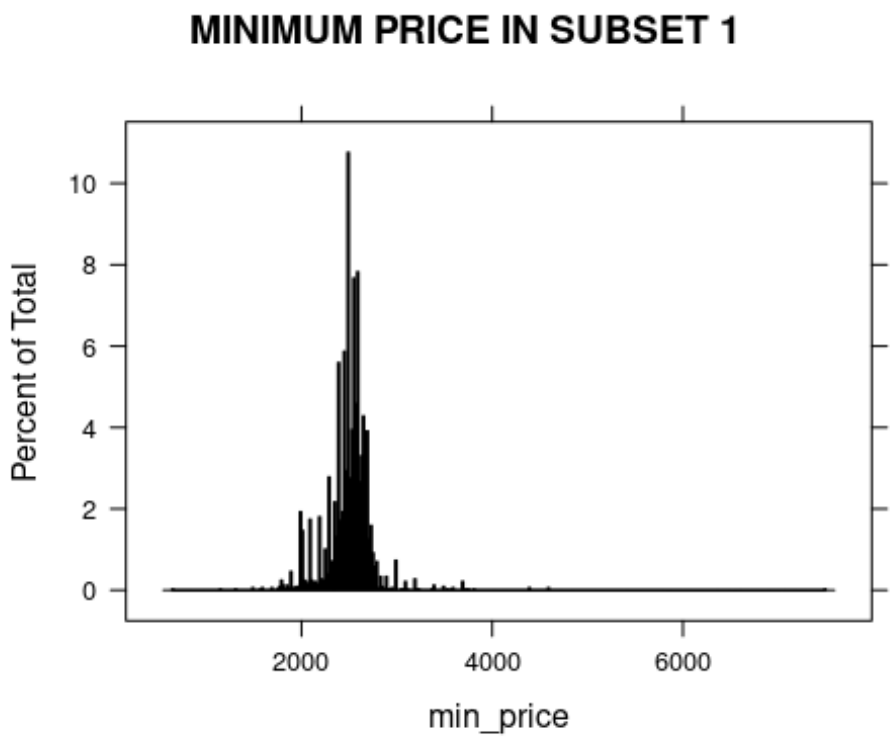
```
#date type conversion
data["arrival_date"]=as.Date(data$arrival_date, format="%d/%m/%Y")
str(data)
```

```
## 'data.frame':    9012 obs. of  10 variables:
##  $ state       : chr  "Bihar" "Bihar" "Bihar" "Bihar" ...
##  $ district    : chr  "Muzaffarpur" "Muzaffarpur" "Muzaffarpur" "Muzaffarpur" ...
##  $ market      : chr  "Muzaffarpur" "Muzaffarpur" "Muzaffarpur" "Muzaffarpur" ...
##  $ commodity   : chr  "Wheat" "Wheat" "Wheat" "Wheat" ...
##  $ variety     : chr  "147 Average" "147 Average" "147 Average" "147 Average" ...
##  $ arrival_date: Date, format: "2023-01-03" "2023-01-04" ...
##  $ min_price   : int  1950 2000 1950 2000 2000 2100 2100 2100 2700 2700 ...
##  $ max_price   : int  2800 2800 2800 2900 2900 2900 2900 2900 2900 2900 ...
##  $ modal_price : num  2500 2500 2500 2600 2550 2550 2550 2550 2800 2800 ...
##  $ update_date : chr  "2023-01-27" "2023-01-27" "2023-01-27" "2023-01-27" ...
##  - attr(*, "na.action")= 'omit' Named int [1:26] 1193 1200 1206 1213 1216 1830 1831
1836 1837 1843 ...
##   ..- attr(*, "names")= chr [1:26] "1193" "1200" "1206" "1213" ...
```

```
#getting the unique data of the categorical attributes
unique(data["state"])
```

```
##               state
## 1             Bihar
## 16      Chattisgarh
## 81          Gujarat
## 989         Haryana
## 990       Karnataka
## 1139         Kerala
## 1159 Madhya Pradesh
## 4227    Maharashtra
## 5071   NCT of Delhi
## 5111         Odisha
## 5116      Rajasthan
## 5994  Uttar Pradesh
## 8755    West Bengal
```

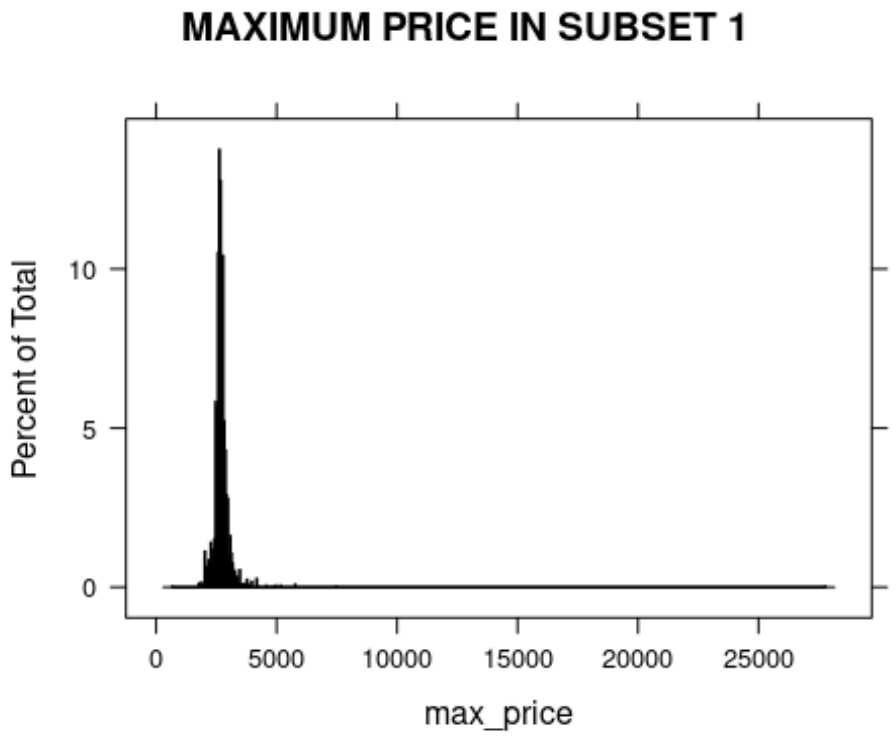
```
#Filtering first 9 days of data 1to 9
subset1=subset(data,arrival_date < "2023-01-10")
#second 9 days 10 to 18
subset2=subset(data, arrival_date > "2023-01-09" & arrival_date < "2023-01-19")
#Third 9 days 19 to 27
subset3=subset(data,arrival_date>"2023-01-18")
#Histogram for minimum price in subset 1
histogram(~min_price,main="MINIMUM PRICE IN SUBSET 1",data=subset1,breaks=415)
```



Inference:

The above histogram is left skewed, indicating that mean is less than median.

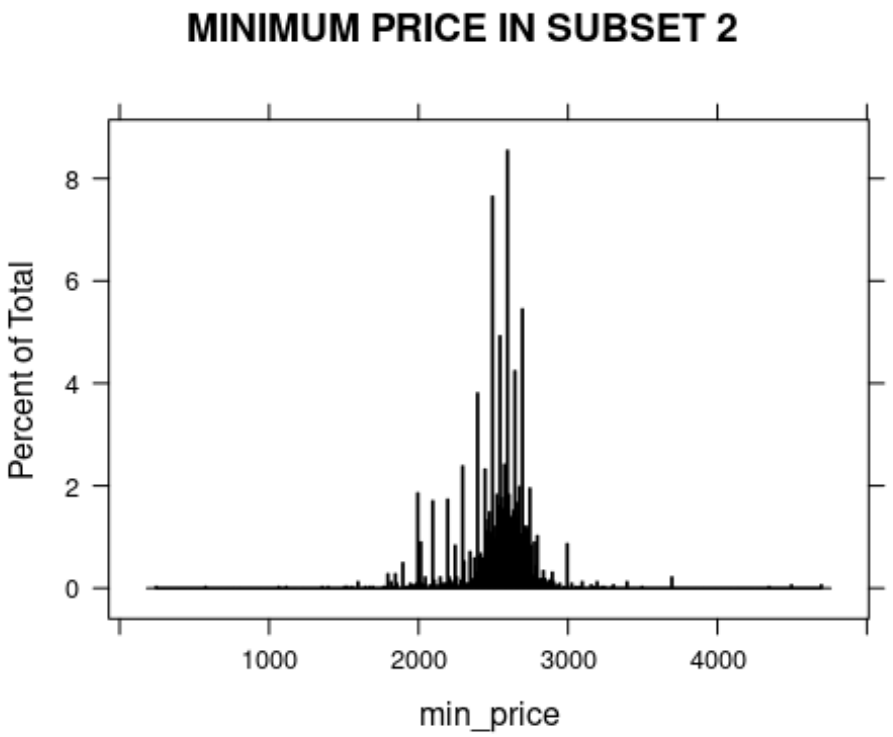
```
#Histogram for maximum price in subset 1
histogram(~max_price,main="MAXIMUM PRICE IN SUBSET 1",data=subset1,breaks=542)
```



Inference:

The above histogram is right skewed.

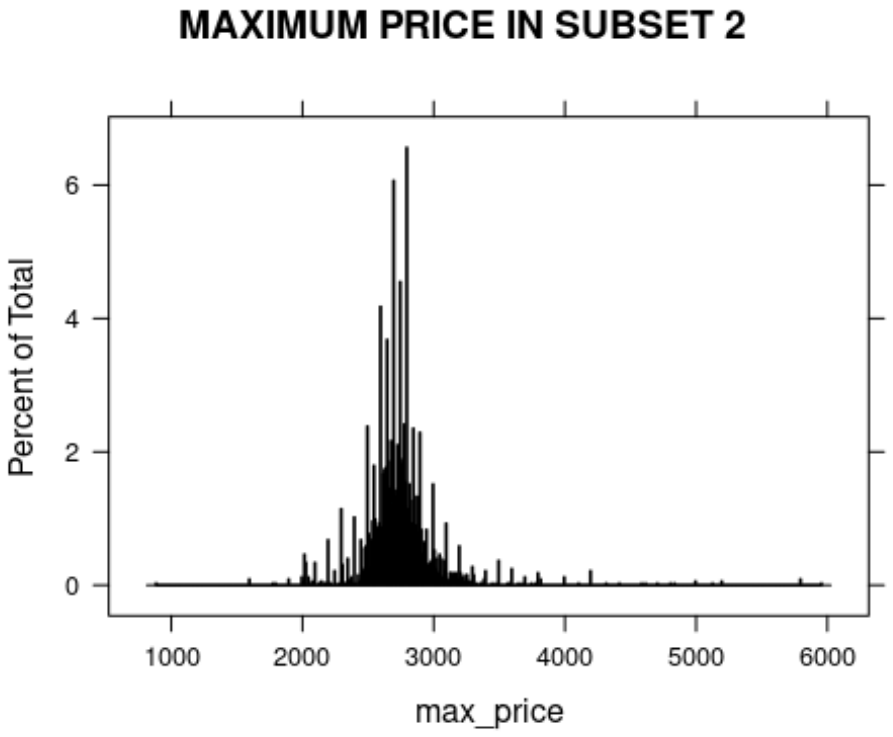
```
#Histogram for minimum price in subset 2
histogram(~min_price,main="MINIMUM PRICE IN SUBSET 2",data=subset2,breaks=418)
```



Inference:

In the above histogram, the mean is less than the median. So, it is left skewed.

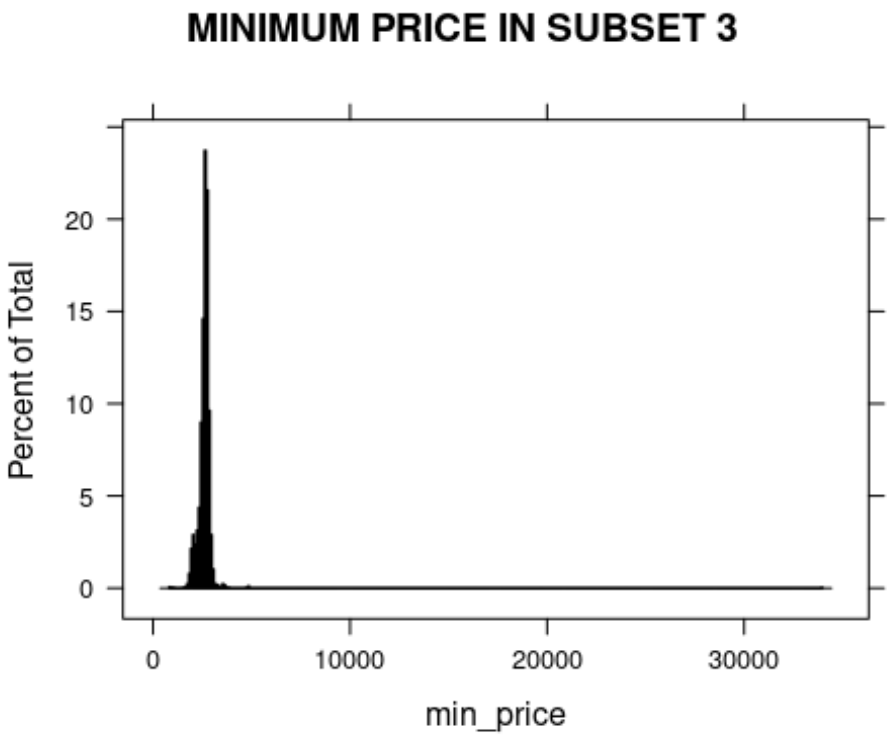
```
#Histogram for maximum price in subset 2
histogram(~max_price,main="MAXIMUM PRICE IN SUBSET 2",data=subset2,breaks=540)
```



Inference:

In the above histogram, the mean is greater than the median, the histogram is right skewed.

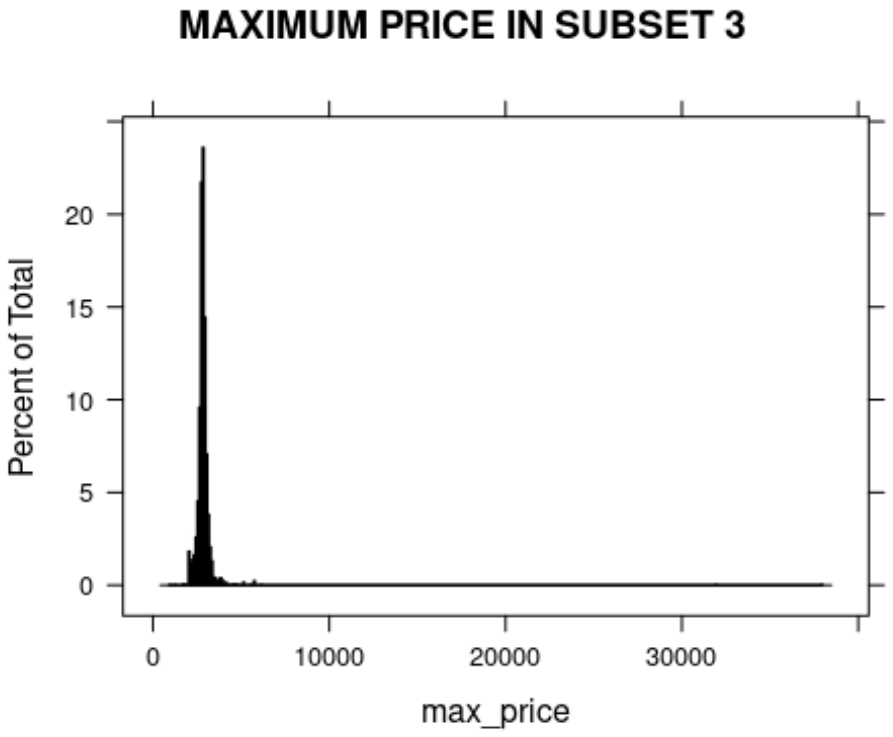
```
#Histogram for minimum price in subset 3
histogram(~min_price,main="MINIMUM PRICE IN SUBSET 3",data=subset3,breaks=381)
```



Inference:

Mean is less than the median of the above histogram. The histogram is left skewed.

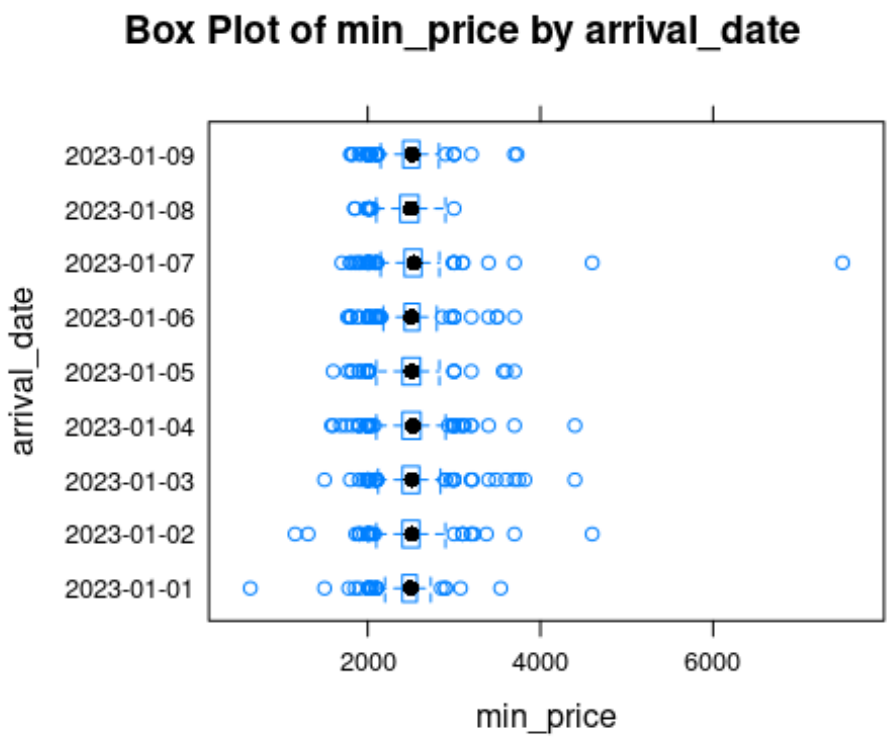
```
#Histogram for maximum price in subset 3
histogram(~max_price,main="MAXIMUM PRICE IN SUBSET 3",data=subset3,breaks=460)
```



Inference:

The mean is greater than the median, so the histogram is right skewed.

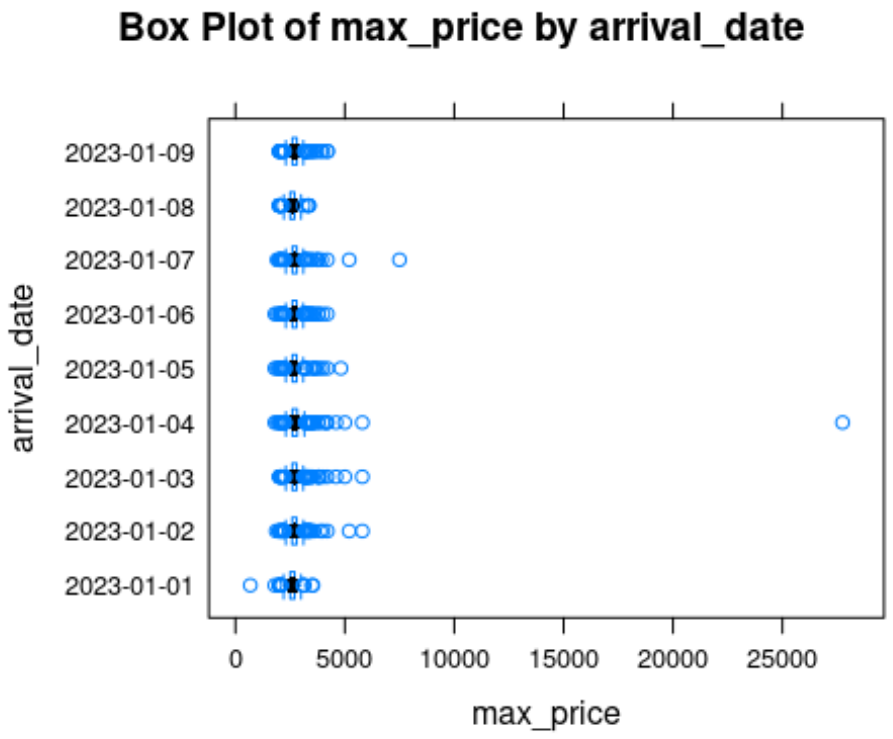
```
#BoxPlot for min_price on the subset 1
bwplot(arrival_date~min_price, data = subset1, main = "Box Plot of min_price by
arrival_date",
xlab = "min_price", ylab = "arrival_date")
```



Inference:

The maximum min_price is recorded on “2023-01-07” and minimum min_price is recorded on “2023-01-01”

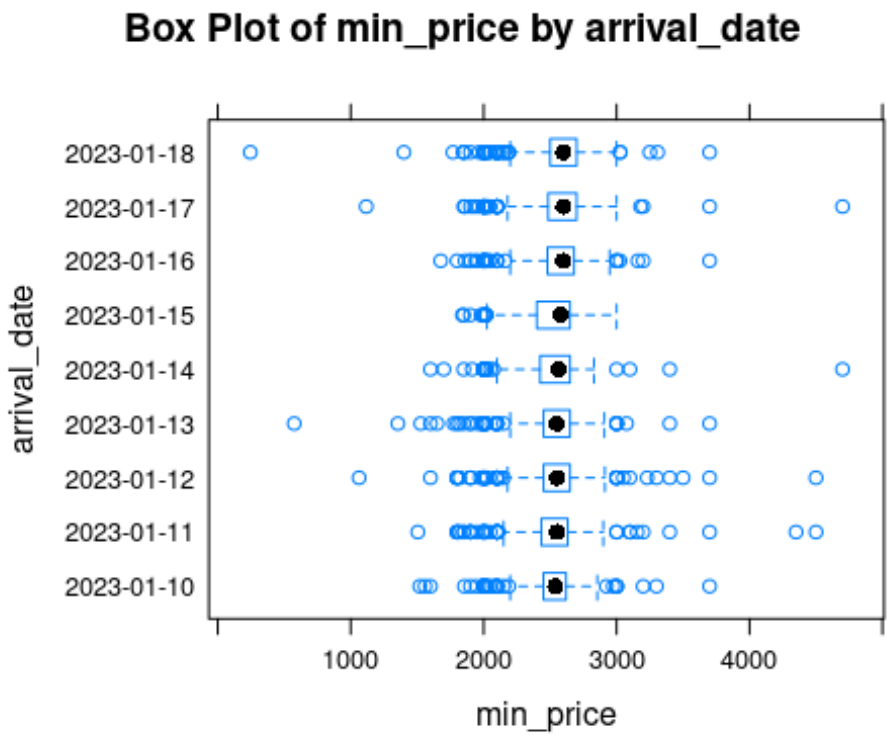
```
#BoxPlot for max_price on the subset 1
bwplot(arrival_date~max_price, data = subset1, main = "Box Plot of max_price by
arrival_date",
xlab = "max_price", ylab = "arrival_date")
```



Inference:

The maximum max_price is recorded on “2023-01-04” and minimum max_price is recorded on “2023-01-01”

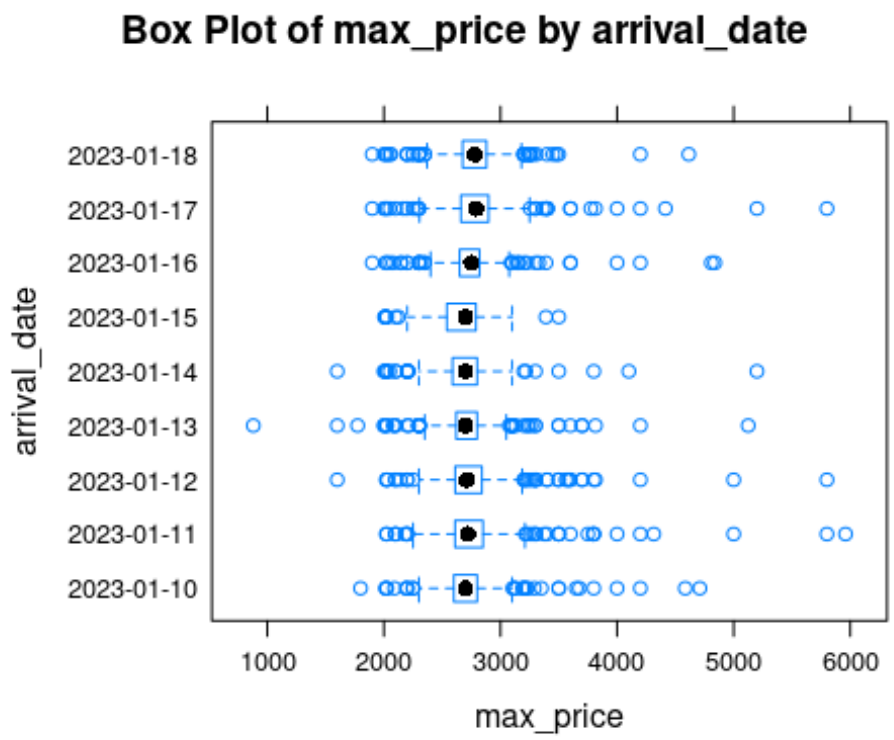
```
#BoxPlot for min_price on the subset 2
bwplot(arrival_date~min_price, data = subset2, main = "Box Plot of min_price by
arrival_date",
xlab = "min_price", ylab = "arrival_date")
```



Inference:

The maximum min_price is recorded on “2023-01-17”, “2023-01-14” and minimum min_price is recorded on “2023-01-18”

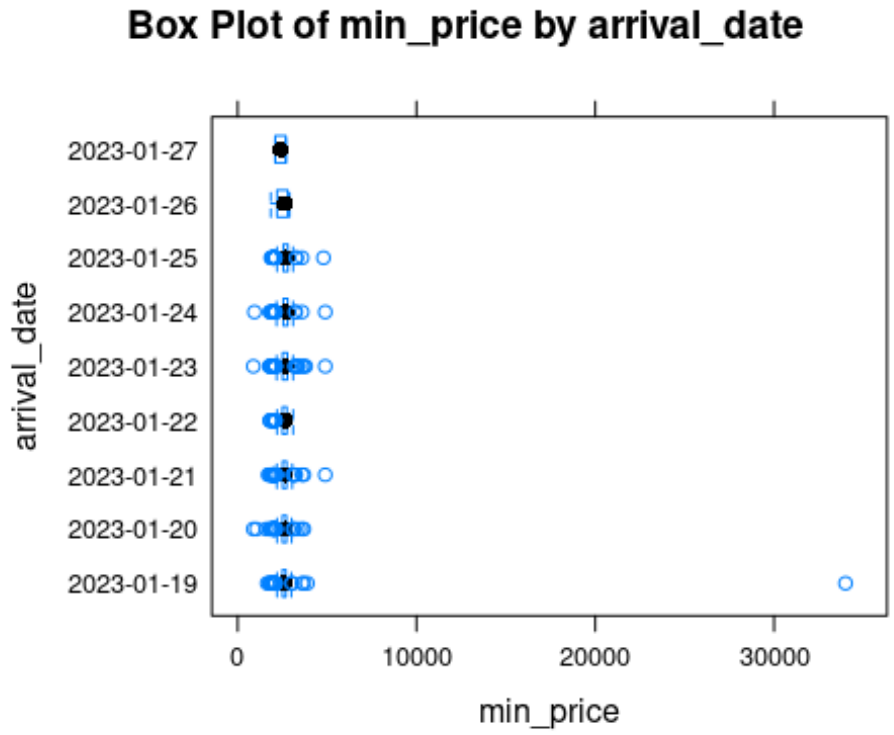
```
#BoxPlot for max_price on the subset 2
bwplot(arrival_date~max_price, data = subset2, main = "Box Plot of max_price by
arrival_date",
xlab = "max_price", ylab = "arrival_date")
```



Inference:

The maximum max_price is recorded on “2023-01-11” and minimum max_price is recorded on “2023-01-13”

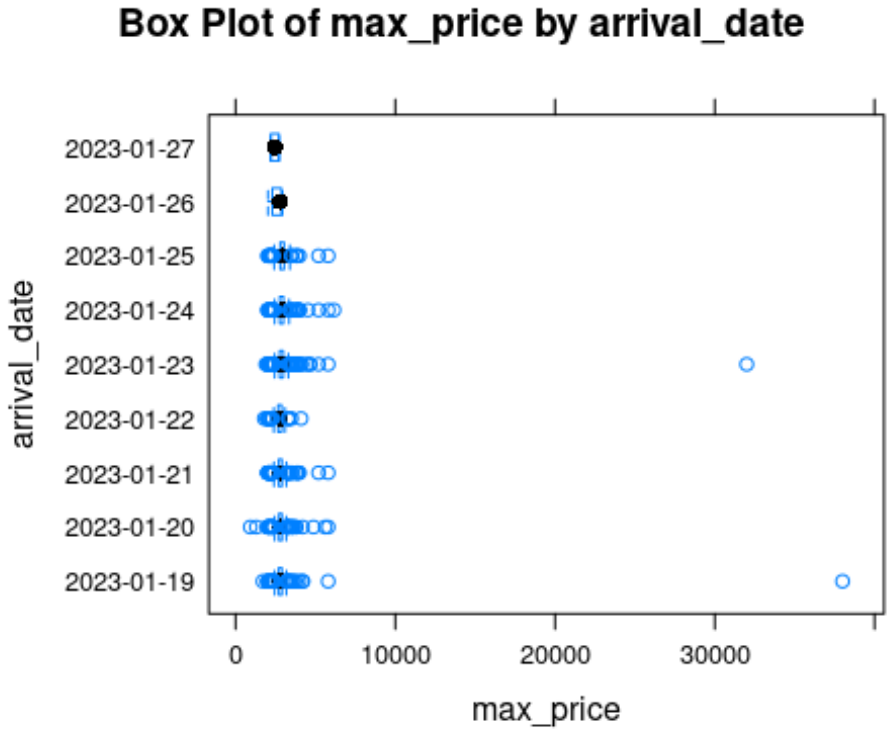
```
#BoxPlot for min_price on the subset 3
bwplot(arrival_date~min_price, data = subset3, main = "Box Plot of min_price by
arrival_date",
xlab = "min_price", ylab = "arrival_date")
```



Inference:

The maximum min_price is recorded on “2023-01-19” and minimum min_price is recorded on “2023-01-20”

```
#BoxPlot for max_price on the subset 3
bwplot(arrival_date~max_price, data = subset3, main = "Box Plot of max_price by
arrival_date",
xlab = "max_price", ylab = "arrival_date")
```



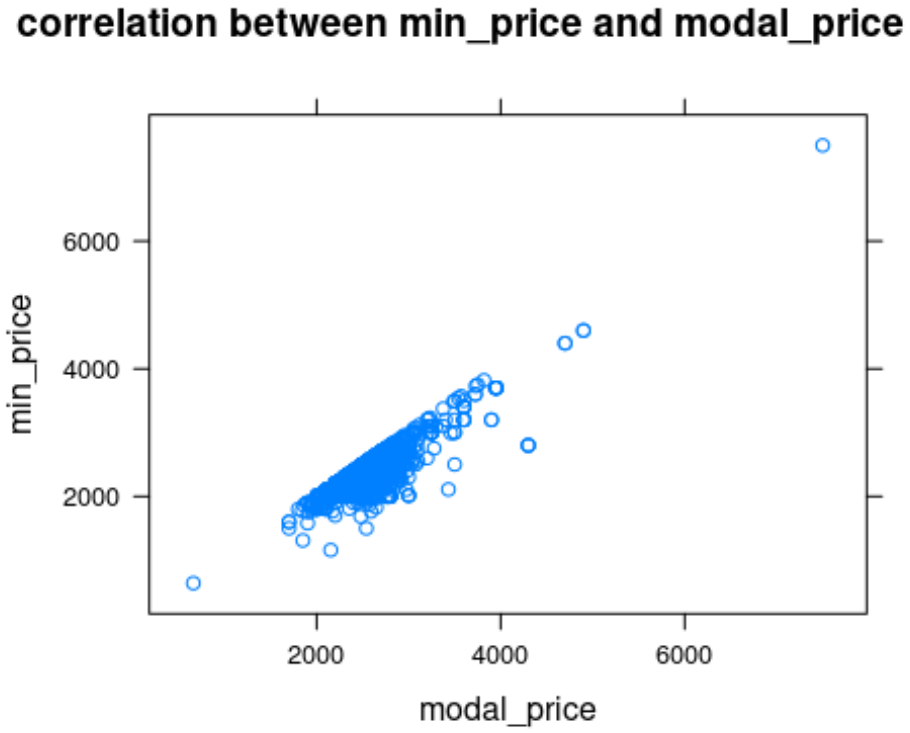
Inference:

The maximum max_price is recorded on “2023-01-19” and minimum max_price is recorded on “2023-01-20”

```
#scatter plot to find the correlation  between min_price and max_price with modal_price in
the subset1
cor(subset1$min_price,subset1$modal_price)
```

```
## [1] 0.8417589
```

```
xyplot(min_price ~ modal_price, data = subset1,main="correlation between min_price and
modal_price")
```



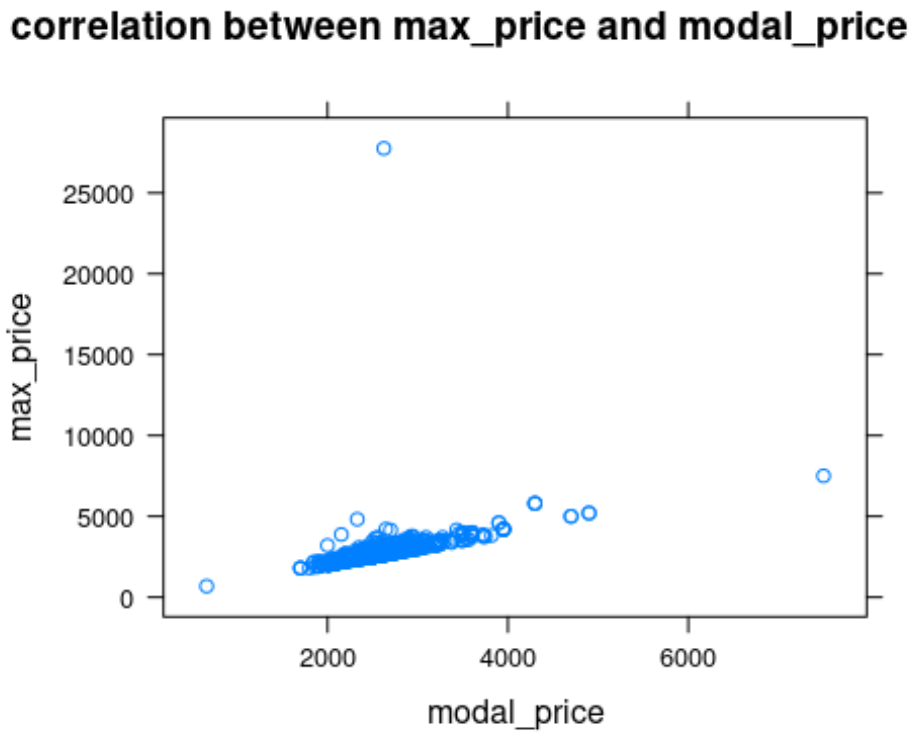
Inference:

There is a positive correlation between min_price and modal_price.

```
cor(subset1$max_price,subset1$modal_price)
```

```
## [1] 0.5171995
```

```
xyplot(max_price ~ modal_price, data = subset1,main="correlation between max_price and
modal_price")
```



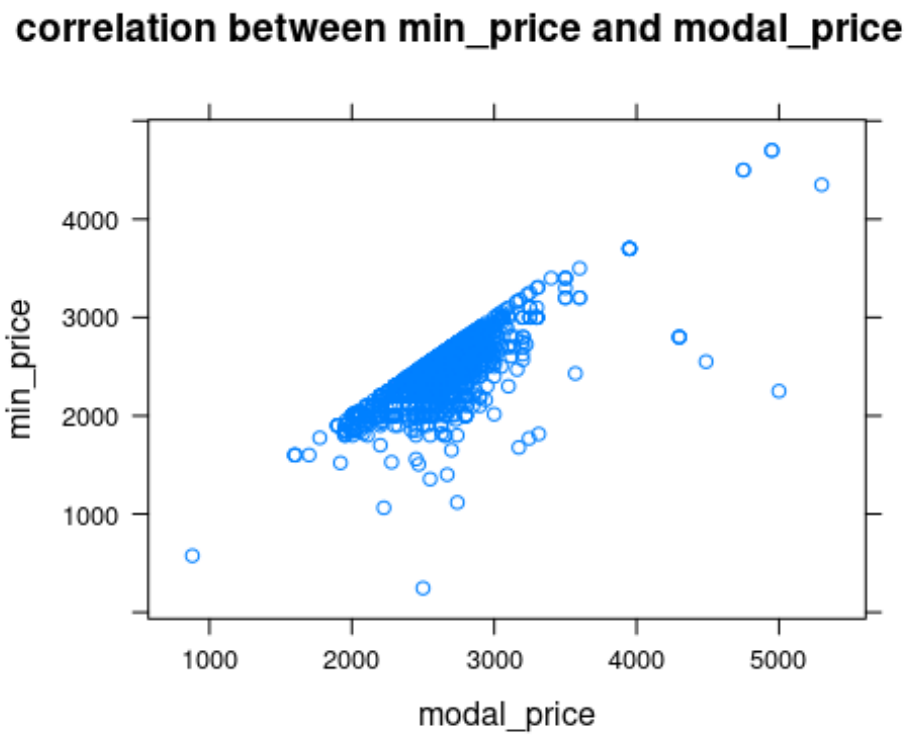
Inference:

There is a positive correlation between max_price and modal_price.

```
#scatter plot to find the correlation  between min_price and max_price with modal_price in the subset2
cor(subset2$min_price,subset2$modal_price)
```

```
## [1] 0.7641578
```

```
xyplot(min_price ~ modal_price, data = subset2,main="correlation between min_price and modal_price")
```



Inference:

There is a positive correlation between min_price and modal_price.

```
cor(subset2$max_price,subset2$modal_price)
```

```
## [1] 0.8659675
```

```
xyplot(max_price ~ modal_price, data = subset2,main="correlation between max_price and modal_price")
```

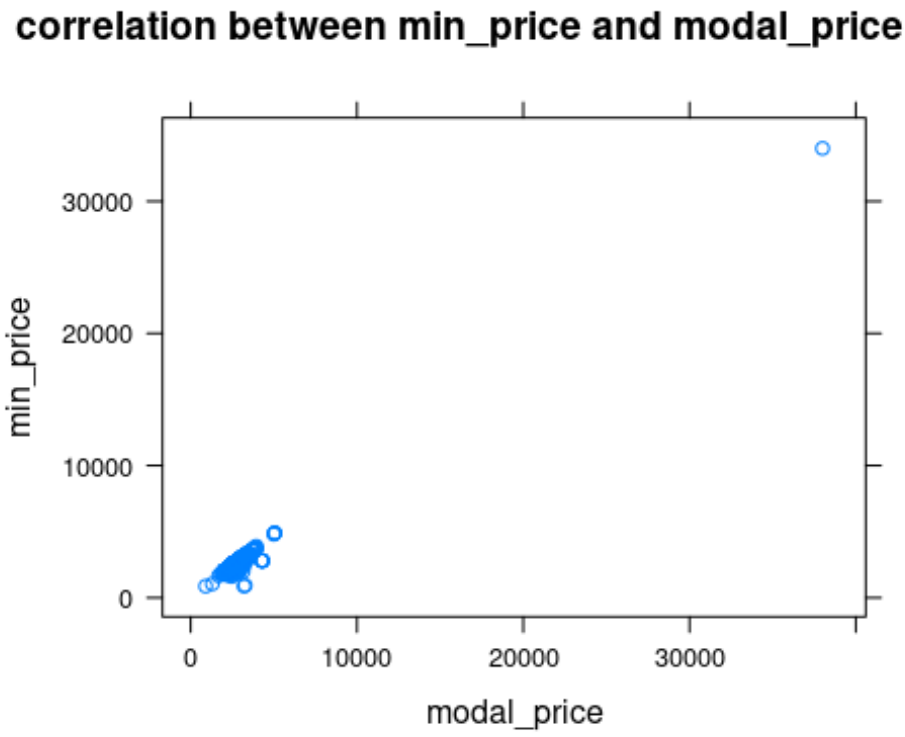
Inference:

There is a positive correlation between max_price and modal_price.

```
#scatter plot to find the correlation  between min_price and max_price with modal_price in the subset3
cor(subset3$min_price,subset3$modal_price)
```

```
## [1] 0.9710703
```

```
xyplot(min_price ~ modal_price, data = subset3,main="correlation between min_price and modal_price")
```



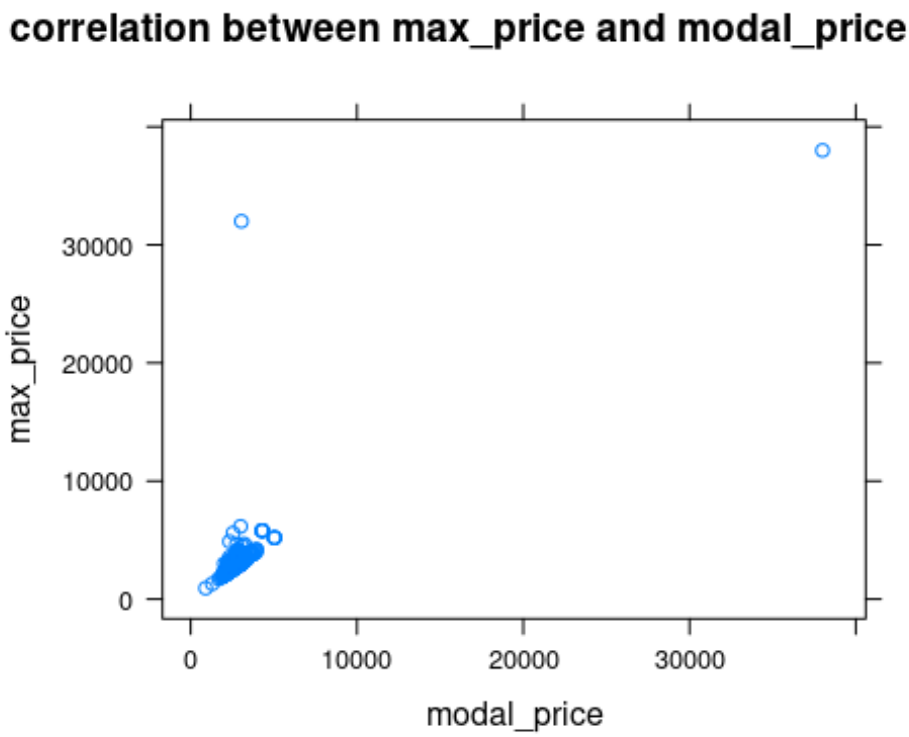
Inference:

There is a positive correlation between min_price and modal_price.

```
cor(subset3$max_price,subset3$modal_price)
```

```
## [1] 0.7836698
```

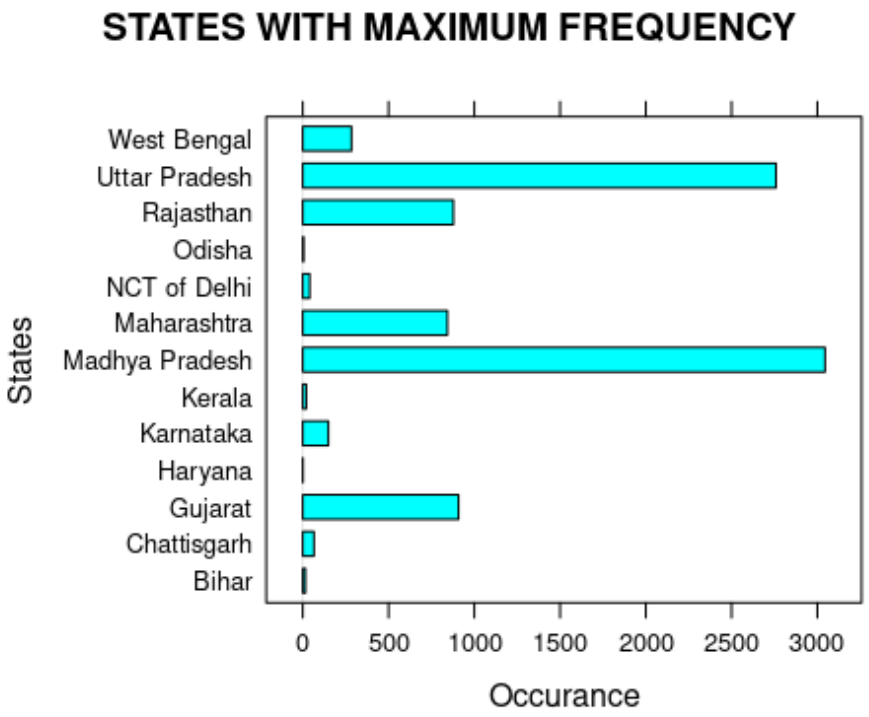
```
xyplot(max_price ~ modal_price, data = subset3,main="correlation between max_price and
modal_price")
```



Inference:

There is a positive correlation between max_price and modal_price.

```
# Barchart to find the highest frequent state in the dataset
barchart(data["state"],main = "STATES WITH MAXIMUM FREQUENCY",
xlab = "Occurance",
ylab = "States")
```



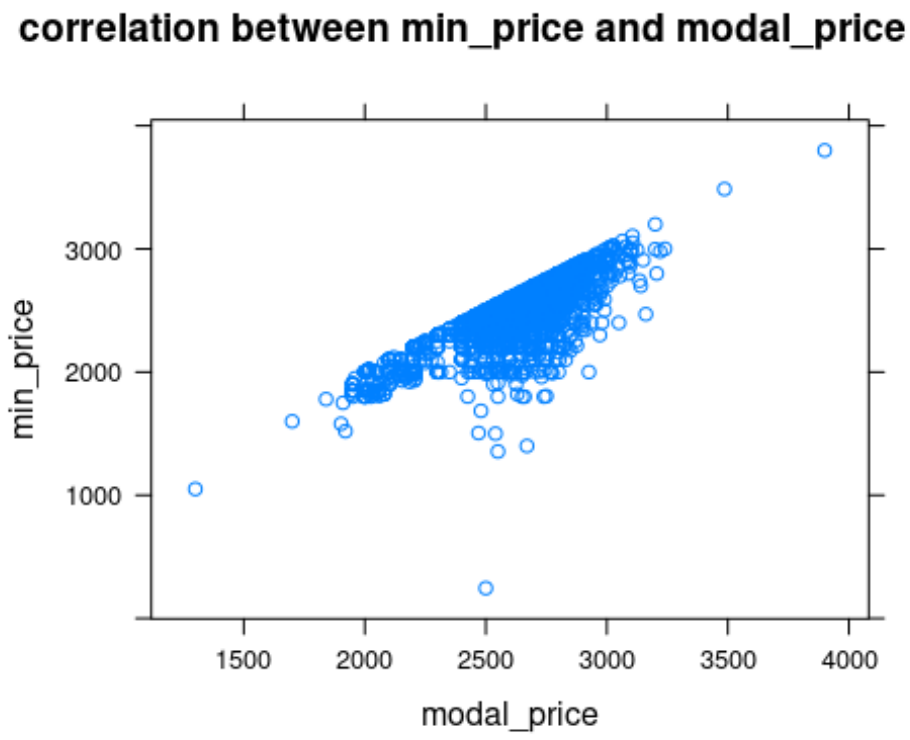
Inference:

Madhya Pradesh is the state having the highest frequency in the whole dataset.

```
#Most frequent state
mp=filter(data,state=="Madhya Pradesh")
#scatter plot for correlation between min_price and max_price with modal_price on the
MadhyaPradesh
cor(mp$min_price,mp$modal_price)
```

```
## [1] 0.7787383
```

```
xyplot(min_price ~ modal_price, data = mp,main="correlation between min_price and
modal_price")
```



Inference:

There is a positive correlation between min_price and modal_price.

```
cor(mp$max_price,mp$modal_price)
```

```
## [1] 0.8284751
```

```
xyplot(max_price ~ modal_price, data = mp,main="correlation between max_price and
modal_price")
```

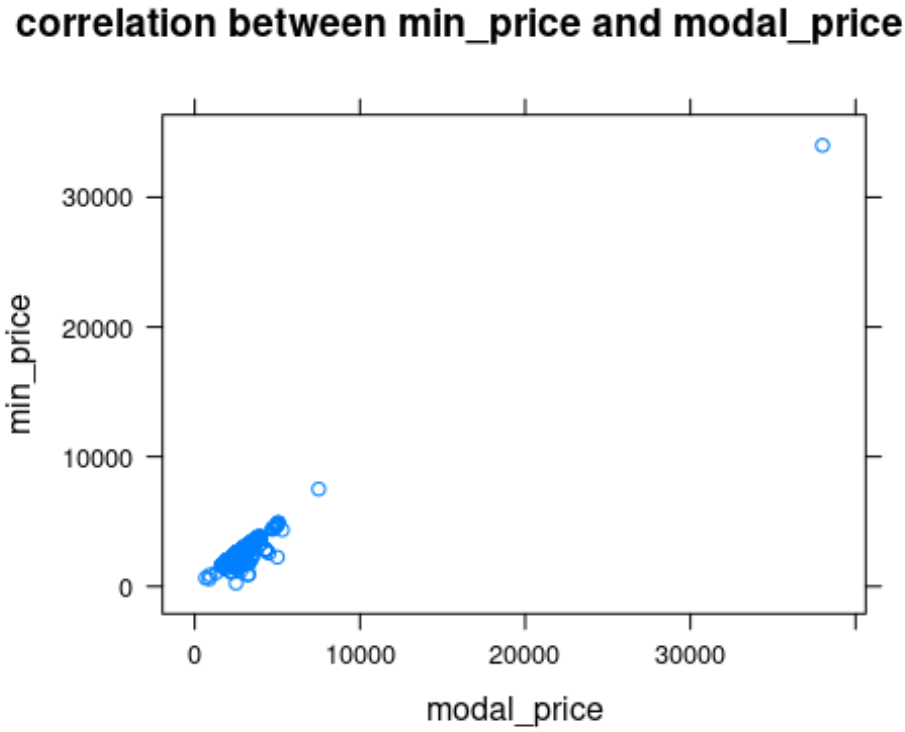
Inference:

There is a positive correlation between max_price and modal_price.

```
#scatter plot for whole country correlation analysis of min_price and max_price with
modal_price
cor(data$min_price,data$modal_price)
```

```
## [1] 0.9312049
```

```
xyplot(min_price ~ modal_price, data = data,main="correlation between min_price and
modal_price")
```



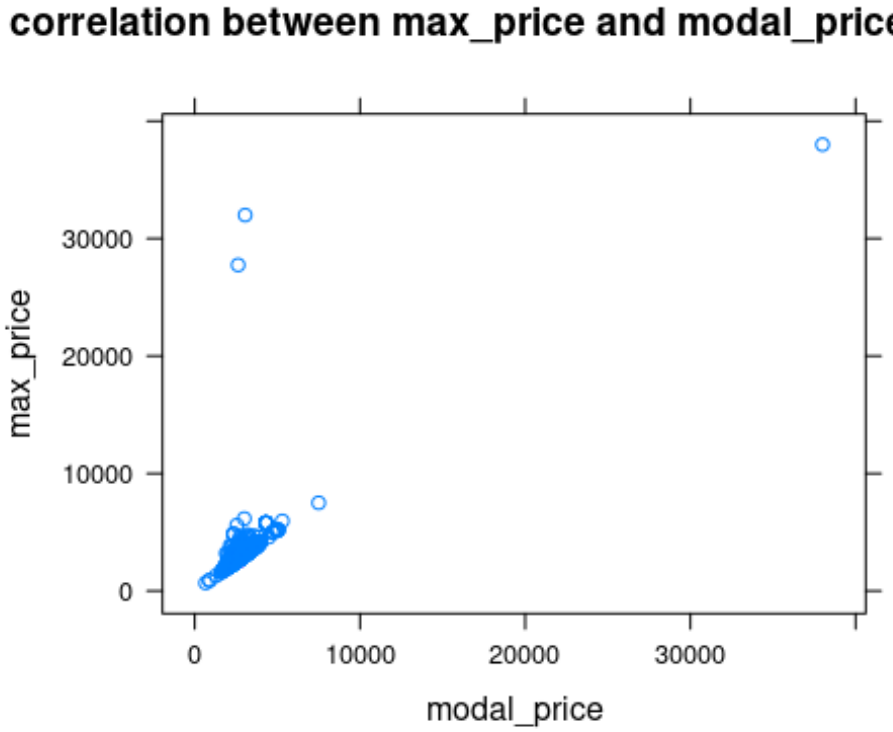
Inference:

There is a positive correlation between min_price and modal_price.

```
cor(data$max_price,data$modal_price)
```

```
## [1] 0.7342681
```

```
xyplot(max_price ~ modal_price, data = data,main="correlation between max_price and
modal_price")
```



Inference:

There is a positive correlation between min_price and modal_price.

Insights:

Frequent state based analysis:

- As a result of analysis, it is found that Madhya Pradesh is the frequently occured state in the whole dataset.
- The highest contributor of Madhya Pradesh is not the sample of the whole dataset, their correlation value of min_price and max_price with modal_price is compared with the whole data.

Arrival_date based analysis:

The dataset is divided into 3 subsets based on the arrival date of the commodity, then the correlation is found between modal_price with max_price and min_price.

- The first subset of the date “2023-01-01” to “2023-01-09” of modal_price is highly correlated with the min_price. It indicates that most commodities are sold near to the min_price.
- The second subset of the date “2023-01-10” to “2023-01-18” of modal_price is highly correlated with the max_price. It indicates that most commodities are sold near to the max_price.
- The third subset of the date “2023-01-19” to “2023-01-27” of modal_price is highly correlated with the min_price. It indicates that most commodities are sold near to the min_price.

  As a conclusion,

  - The third subset of date has sold the commodity mostly related to the min_price, which is a loss for farmers.
  - The farmers who sold their commodity in the second subset receives the maximum profit.

Whole data based analysis:

As the whole dataset’s min_price and max_price is correlated with modal_price

- The min_price receives the highest value of correlation.
- It indicates that there is **NO DEMAND** for the commodity wheat in the whole country.

—-------