Lecture 8,9

INSTITUTE OF ACTUARIAL & QUANTITATIVE STUDIES

Class: SY BSc

Subject : Statistical Modelling in R - 1

Chapter: Unit I

Chapter Name: Data Management

Today's Agenda

1. Case study on car price prediction contd.
2. Statistical analysis
3. Uni-variate analysis
4. Bivariate analysis
5. Outlier detection and removal
6. Plots

Case Study

A Japanese automobile company Geely Auto aspires to enter the Indian market by setting up their manufacturing unit there and producing cars locally to give competition to their Indian counterparts.

They have contracted an automobile consulting company to understand the factors on which the pricing of cars depends. Specifically, they want to understand the factors affecting the pricing of cars in the Indian market, since those may be very different from the Japanese market.

The company wants to know:

- **Which variables are significant in predicting the price of a car**
-**How well those variables describe the price of a car**

Explain the approach taken to fulfill the company's requirement.

Fields in the dataset

aspiration	aspiration type of engine
boreratio	ratio between engine cyinder bore diameter and piston stroke length
car_ID	unique id provided to each car
carbody	car body types
carheight	height of a car in inches
carlength	car length in inches
CarName	Company's car with model name
carwidth	car widthin in inches
citympg	city travel miles per gallon
compressionratio	ratio of relative volumes of combustion chamber and cylinder
curbweight	total mass of a vehicle in kg
cylindernumber	number of cylinders in the car
doornumber	door number
drivewheel	drive train
enginelocation	location of the engine in the car

Fields in the dataset

enginesize	size of engine in cc
enginetype	types of car engines
fuelsystem	combination of parts responsible for delivering fuel to system
fueltype	type of fuel used in engine
highwaympg	indiactes long distance travel miles per gallon
horsepower	power of engine in horsepower
peakrpm	peak revolutions per minute
price	price of car
stroke	stroke length
symboling	risk fator symbol associated with its price
wheelbase	distance between centers of front and rear wheels

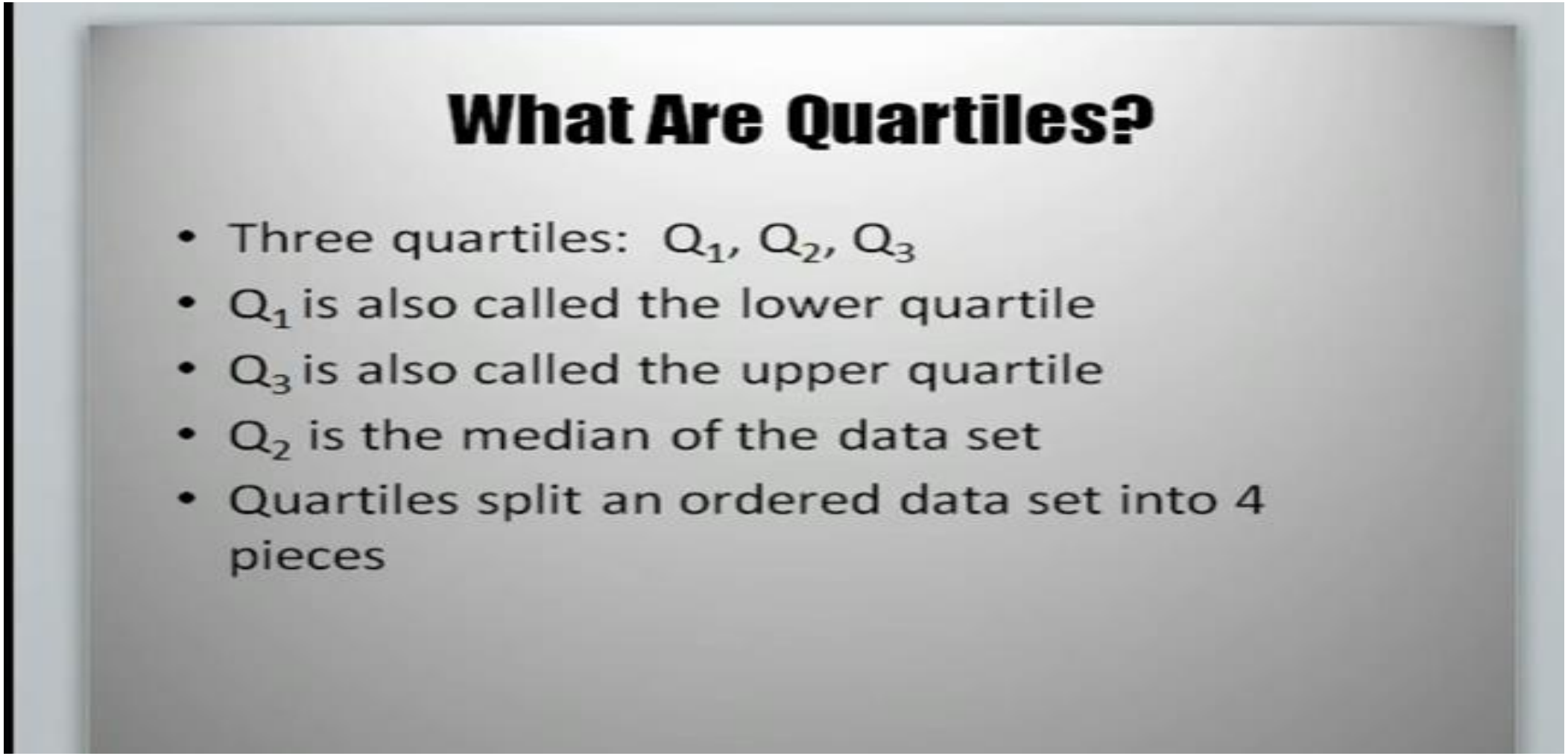
IAQS | INSTITUTE OF ACTUARIAL & QUANTITATIVE STUDIES

Data Visualisation

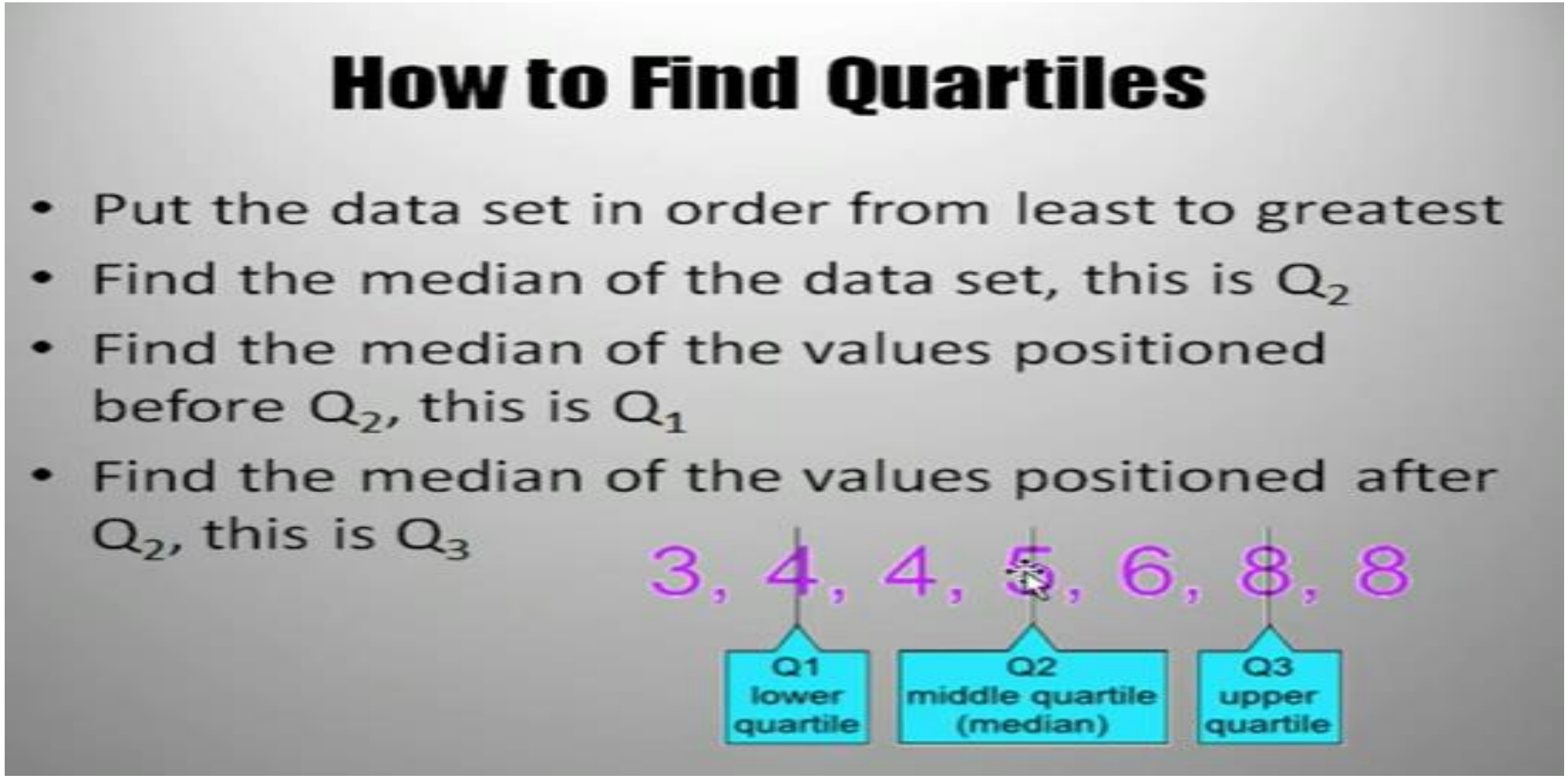
IAQS | INSTITUTE OF ACTUARIAL & QUANTITATIVE STUDIES

Boxplot

Boxplot


What Are Quartiles?
• Three quartiles: Q_1, Q_2, Q_3
• Q_1 is also called the lower quartile
• Q_3 is also called the upper quartile
• Q_2 is the median of the data set
• Quartiles split an ordered data set into 4 pieces

Boxplot


How to Find Quartiles
• Put the data set in order from least to greatest
• Find the median of the data set, this is Q_2
• Find the median of the values positioned before Q_2, this is Q_1
• Find the median of the values positioned after Q_2, this is Q_3
3, 4, 4, 5, 6, 8, 8
Q1 lower quartile
Q2 middle quartile (median)
Q3 upper quartile

Boxplot

5 Number Summary

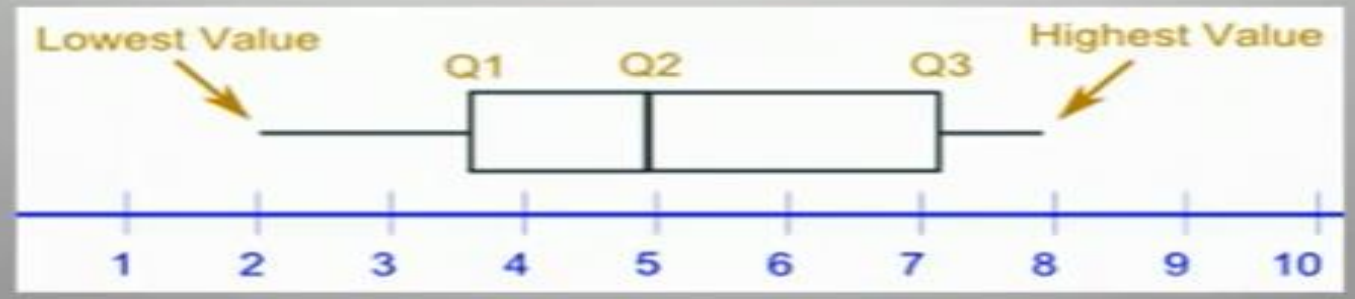
1. Minimum Value
2. Q_1
3. Q_2
4. Q_3
5. Maximum Value

These can be displayed graphically as a **boxplot** (also called a box and whisker plot).

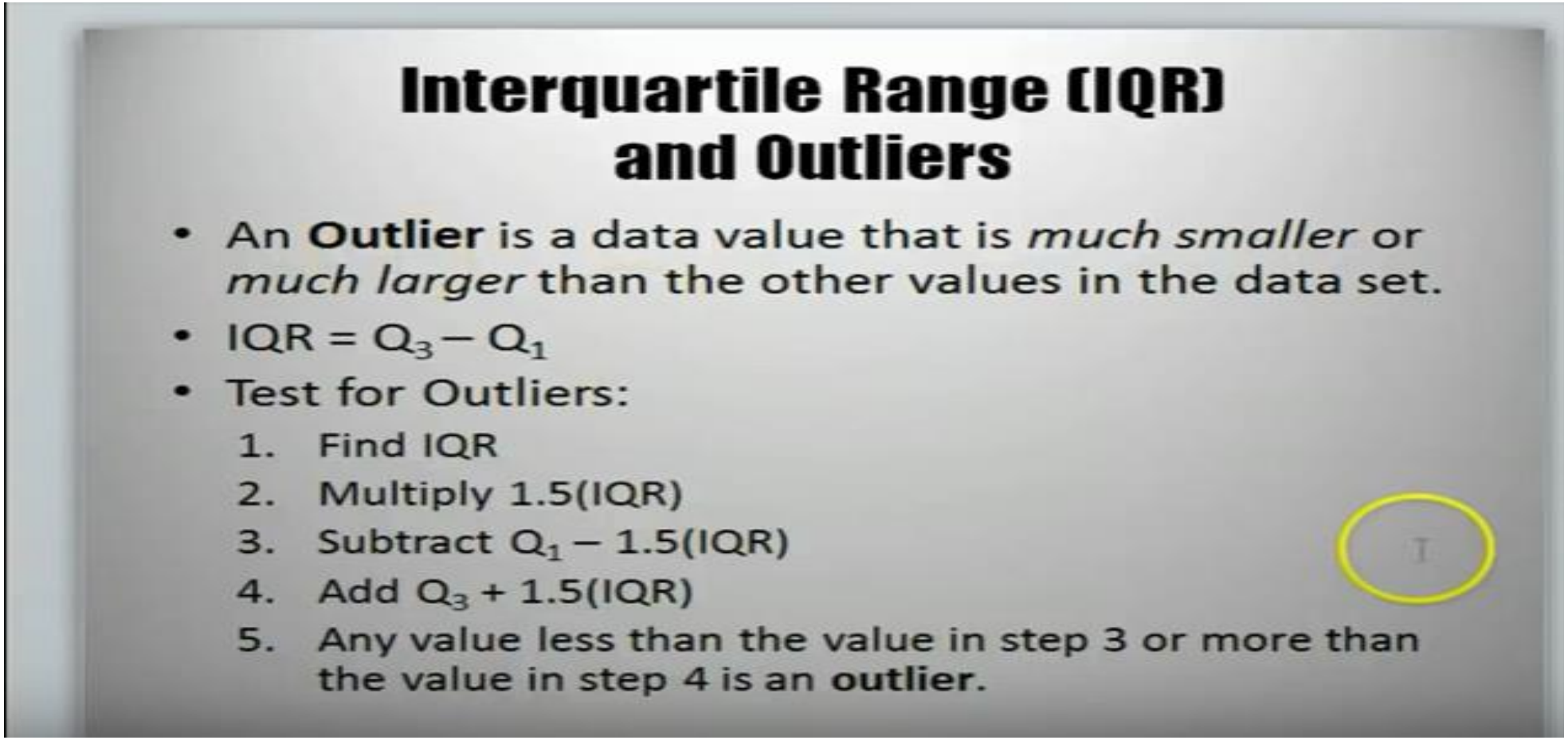
Boxplot

How to Construct a Boxplot

- Determine the 5 Number Summary
- Arrange the 5 Number Summary on a number line in the following fashion



Boxplot


Interquartile Range (IQR)
and Outliers
• An Outlier is a data value that is much smaller or much larger than the other values in the data set.
• IQR = $Q_3 - Q_1$
• Test for Outliers:
1. Find IQR
2. Multiply 1.5(IQR)
3. Subtract $Q_1 - 1.5(IQR)$
4. Add $Q_3 + 1.5(IQR)$
5. Any value less than the value in step 3 or more than the value in step 4 is an outlier.

Interquartile range

- The range where the middle 50% of the values lie is called interquartile range (IQR).
- IQR is especially useful to users who have more interests in values toward the middle and have less interests in extremes.
- In percentile terms, IQR is the distance between the 75^{th} percentile and 25^{th} percentile.
- The 75^{th} percentile is located at the 3^{rd} quartile, or Q3.
- The 25^{th} percentile is located at the 1^{st} quartile, or Q1.
- The IQR provides a clearer description of the overall data set by removing these extreme values in the data set.
- Thus, the IQR is more meaningful than the range.

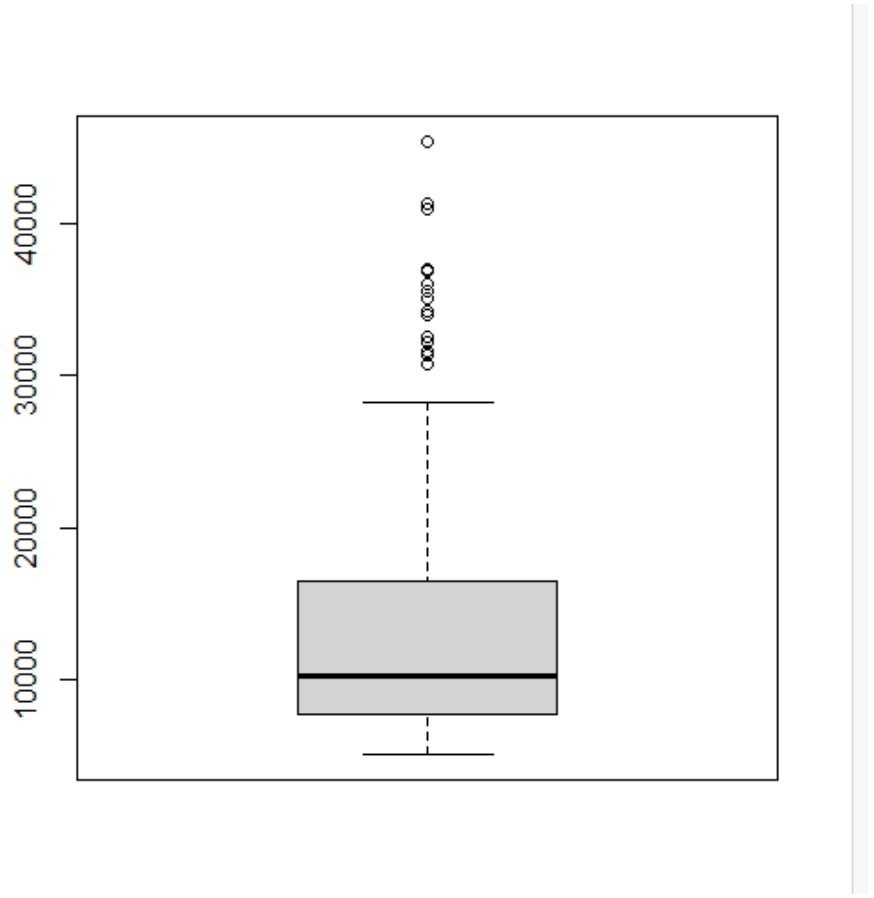
Detecting outliers

- > q_price=quantile(df_car$price)
- > q_price
- 0% 25% 50% 75% 100%
- 5118 7788 10295 16503 45400
- > max(df_car$price)
- [1] 45400
- > min(df_car$price)
- [1] 5118

- > iqr_price=q_price[4]-q_price[2]
- > iqr_price
- 75%
- 8715
- > higher_outlier=q_price[4]+(1.5*iqr_price)
- > higher_outlier
- 75%
- 29575.5
- > lower_outlier=q_price[2]-(1.5*iqr_price)
- > lower_outlier
- 25%
- -5284.5
- > boxplot(df_car$price)

Printing details of the car which are above normal range of price

```
> df_car[df_car$price>higher_outlier,c("CarName","price")]
# A tibble: 15 x 2
   CarName                        price
   <chr>                          <dbl>
 1 bmw x4                         30760
 2 bmw x5                         41315
 3 bmw x3                         36880
 4 jaguar xj                      32250
 5 jaguar xf                      35550
 6 jaguar xk                      36000
 7 buick skyhawk                  31600
 8 buick opel isuzu deluxe        34184
 9 buick skylark                  35056
10 buick century special          40960
11 buick regal sport coupe (turbo) 45400
12 porcshce panamera              32528
13 porsche cayenne                34028
14 porsche boxter                 37028
15 porsche cayenne                31400.
> |
```



Printing outliers in the price column with inbuilt R command

outliers <- boxplot(df_car$price, plot=FALSE)$out

```
> outliers <- boxplot(df_car$price, plot=FALSE)$out
> length(outliers)
[1] 15
> outliers
 [1] 30760.0 41315.0 36880.0 32250.0 35550.0 36000.0 31600.0 34184.0 35056.0 40960.0 45400.0 32528.0 34028.0 37028.0
[15] 31400.5
>
```

Outlier detection with histogram

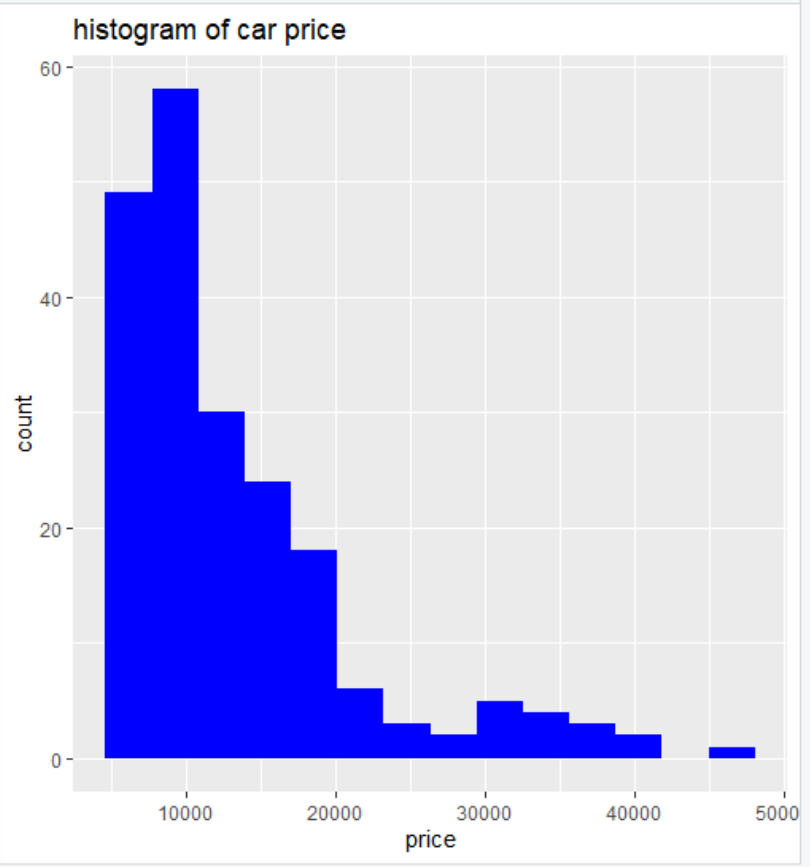
- A histogram is an approximate representation of the distribution of numerical data.
- It was first introduced by Karl Pearson.
- To construct a histogram, the first step is to "bin" (or "bucket") the range of values—
- that is, divide the entire range of values into a series of intervals
- and then count how many values fall into each interval.
- The bins are usually specified as consecutive, non-overlapping intervals of a variable.
- The bins (intervals) must be adjacent and are often (but not required to be) of equal size

ggplot

- ggplot() initializes a ggplot object.
- It can be used to declare the input data frame for a graphic
- and to specify the set of plot aesthetics intended to be common throughout all subsequent layers unless specifically overridden.

Analysing the price column with Histogram

- ggplot(df_car,aes(x=price))+geom_histogram(bins=sqrt(nrow(df_car)),fill =("blue"))+ggtitle("histogram of car price")



What is the significance of outliers

- Outliers are extreme values in a dataset.
- They are numerically distant from the remainder of the data and therefore seem out of place.
- For example , while detection of a disease, Outliers can occur because of the always present possibility of very high or low dietary intakes,
- but may also indicate errors in reporting, coding, or the underlying databases used to estimate intakes.
- Outliers are important because they can have a large influence on statistics derived from the dataset.

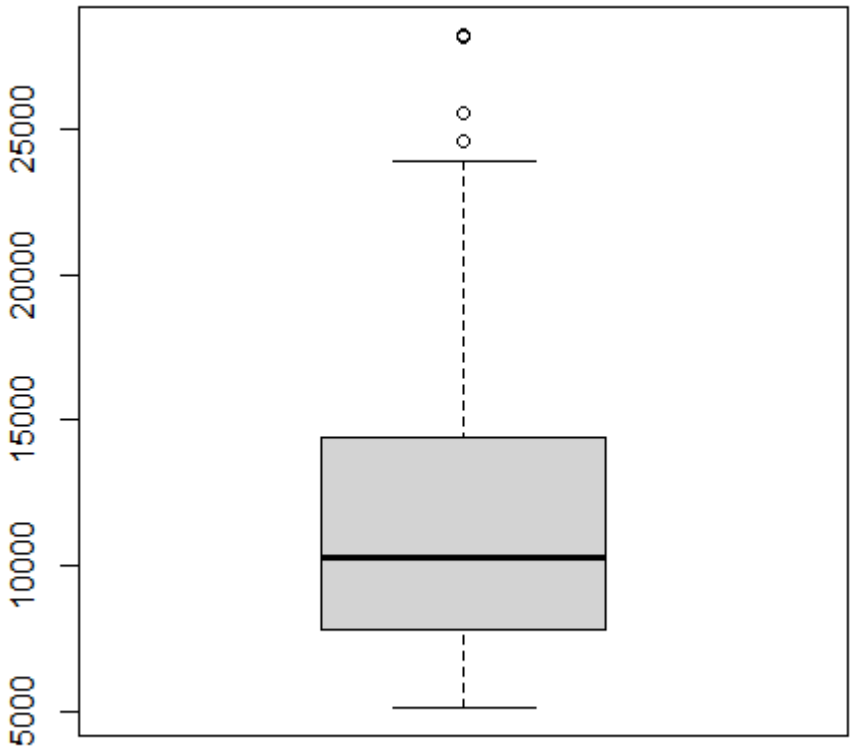
How to handle outliers

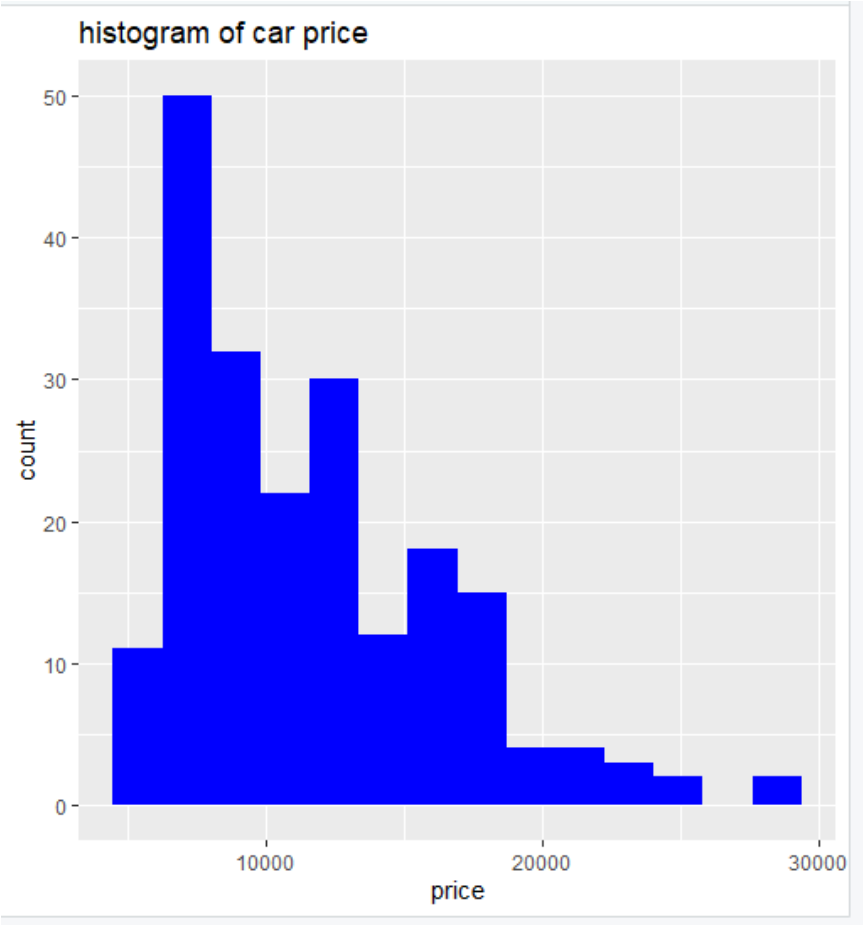
- **Trim the data set, but replace outliers with the nearest "good" data**, For example, if you thought all data points above the 95th percentile were outliers, you could set them to the 95th percentile value.
- **Replace outliers with the mean or median**

How to handle outliers

- > df_car_rm_out=df_car
- >df_car_rm_out[df_car_rm_out$price>higher_outlier,"price"]=mean(df_car_rm_out$price,na.rm=TRUE)
- > nrow(df_car_rm_out)
- [1] 205
- >ggplot(df_car_rm_out,aes(x=price))+geom_histogram(bins=sqrt(nrow(df_car)),fill=("blue"))+ggtitle("histogram of car price")
- > boxplot(df_car_rm_out$price)

printing the plots after handling outliers

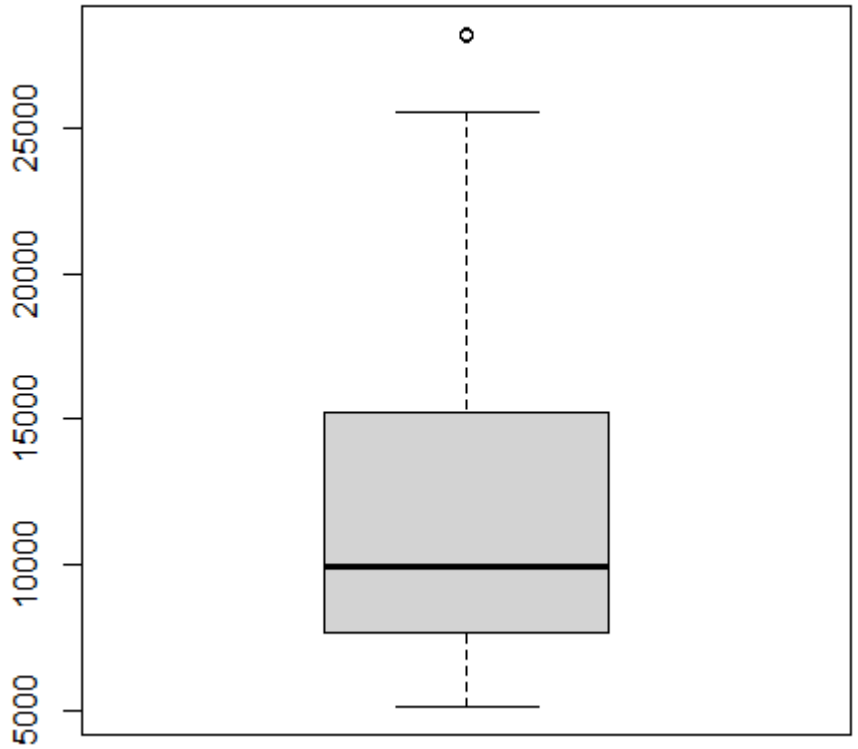


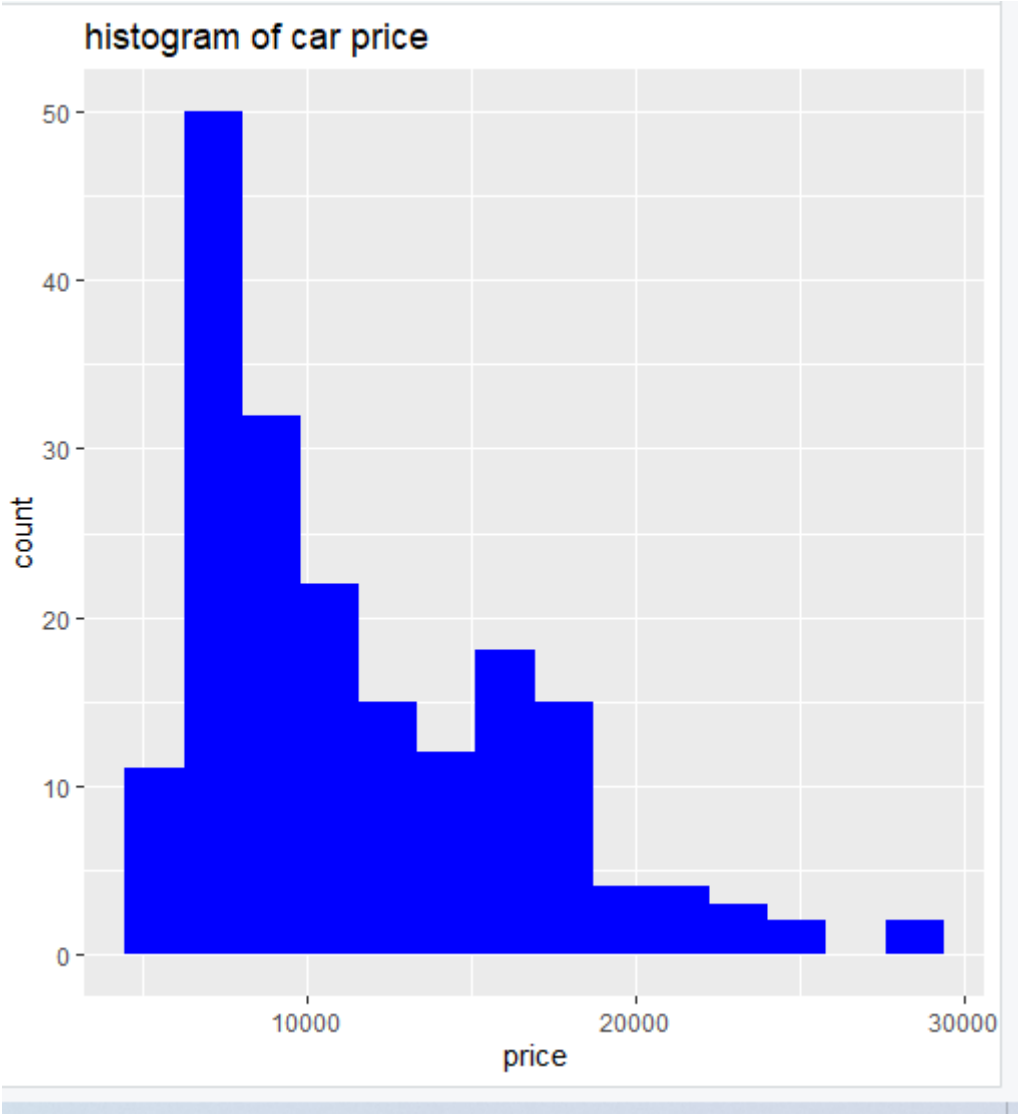


removing outliers

- > outliers <- boxplot(df_car$price, plot=FALSE)$out
- > df_car_new_rm= df_car[-which(df_car$price %in% outliers),]
- > nrow(df_car_new_rm)
- [1] 190

printing the plots after handling outliers





imputation of outliers

- “mean” : arithmetic mean
- “median” : median
- “mode” : mode
- “capping” : Imputate the upper outliers with 95 percentile,
- and Imputate the bottom outliers with 5 percentile.

imputation of outliers

```
cap =function(x){
  q <- quantile( x, c(.05, 0.25, 0.75, .95 ) )
  q1=quantile(x)
  highest_outlier=q1[4]+(1.5*IQR(x))
  lowest_outlier=q1[2]-(1.5*IQR(x))
  x[ x < lowest_outlier ] <- q[1]
  x[ x > highest_outlier ] <- q[4]
  x
}

df_car_cap_pr=cap(df_car$price)
> df_car_cap_pr
> df_car_impu_cap=cbind(df_car,"price_new"=df_car_cap_pr)
```

The new data frame with imputed values

```
> df_car_impu_cap[,c("price","price_new")]
      price price_new
1  13495.00  13495.00
2  16500.00  16500.00
3  16500.00  16500.00
4  13950.00  13950.00
5  17450.00  17450.00
6  15250.00  15250.00
7  17710.00  17710.00
8  18920.00  18920.00
9  23875.00  23875.00
10 17859.17  17859.17
11 16430.00  16430.00
12 16925.00  16925.00
```

Thank You