Module 2 summary: Descriptive Statistics through Visualisation

Describing data with numbers

- > favstats(price, data=Diamonds)

min	Q1	median	Q3	max	mean	sd	n	missing
326	950	2401	5324.25	18823	3932.8	3989.44	53940	0

Describing data

- **Descriptive statistics** summarise characteristics of data using numbers such as mean, range, mode or percentage.
- **Statistical visualisations** are visual displays of descriptive statistics or data, most commonly graphs or plots, that summarise important features or trends

Mean and Variance

- Mean and Variance Measuring the centre and variability of the sample data, are influenced by each individual data in the sample.
- Variance is unit-less but Standard Deviation(its square root) convert it back to its original scale.

$$\bar{x} = \frac{\sum_{n}^{i=1} x_i}{n}$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Quartiles and the Median

- Quartiles are values that break a distribution into four parts.
- Q1= lower quartile or 25th percentile value, 25% of data lie to its left,
- Q2= Median, 50% percentile, 50% of data lie to its left,
- Q3= Upper quartile or 75th percentile, 75% of data lie to its left,

Calculation of median

- Median when n = even
- Scenario 2 2, 4, 9, 8, 6, 5
- Ordered 2, 4, 5, 6, 8, 9
- Location of Median Average of the (n/2) and (n/2) + 1 observations, so the average of the 3rd and 4th.
- The median is the average of the 3rd and 4th ordered observation so, Median = (5+6)/2 = 5.5.

Calculation of median

- The calculation of the median depends on whether there are an even or odd number of data points:
- Median when n = odd
- Scenario 1 2, 4, 9, 8, 6, 5, 3
- Ordered 2, 3, 4, 5, 6, 8, 9
- Location of Median (n + 1)/2 = (7+1)/2 = 4th
- Therefore, the median is the 4th ordered value, Median = 5.

Calculation of quartiles

- Q1 and Q3 are calculated in a similar fashion after the dataset is split at the median, top and bottom 50%.
- Q1 is the median of the bottom 50% (i.e. 25th percentile) and Q3 is median of the top 50% (i.e. 75th percentile).

Calculation of quartiles

- **Q1 and Q3 when n = odd (take median value)**
- Q1 = Median of bottom 50%: For example, Median of 2, 3, 4, 5 = average of 2nd and 3rd value = (3+4)/2 = 3.5
- Q3 = Median of top 50%: For example, Median of 5, 6, 8, 9 = average of 2nd and 3rd value = (6+8)/2 = 7
- Note how the median is included in both halves.

Outliers

- The **interquartile range (IQR)** is the range of the middle 50% of data and is depicted as the "box" in the box plot. The IQR is also a measure of variation.

$$IQR = Q_3 - Q_1$$

The outlier fences are defined as the following:

$$\text{Lower outlier} < Q_1 - 1.5 * IQR$$

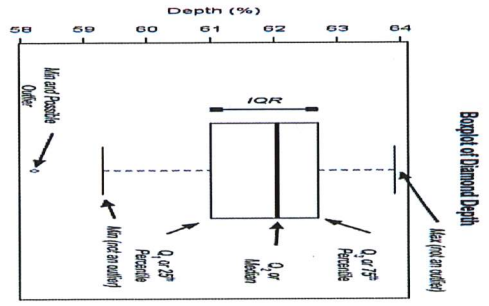
$$\text{Upper outlier} > Q_3 + 1.5 * IQR$$

Calculation of quartiles

- **Q1 and Q3 when n = even**
- Q1 = Median of bottom 50%: For example, Median of 2, 4, 5 = 2nd value = 4
- Q3 = Median of top 50%: For example, Median of 6, 8, 9 = 2nd value = 8.
- Note how the median is not included because the median is not an actual data point.

Box Plots

- Diamonds_sample$depth %>% summary()
- ## Min. 1st Qu. Median Mean 3rd Qu. Max.
- ## 58.20 61.05 62.05 61.78 62.68 63.90



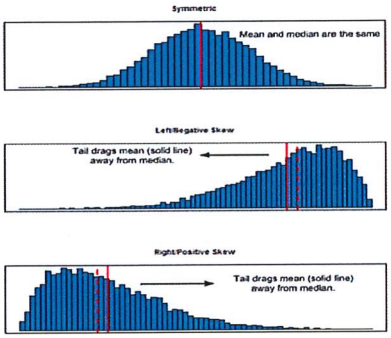
Outliers

- Outliers are values that fall beyond the **outlier fences**. Box plots also include suspected **outliers**, depicted using an "o" or a similar symbol. Always check the reason for outliers. The outlier fences are

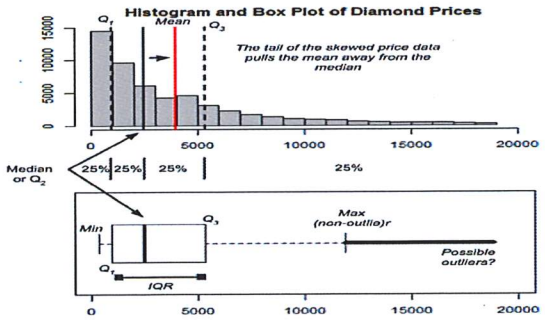
$$\text{Lower outlier} < Q_1 - 1.5 * IQR$$

$$\text{Upper outlier} > Q_3 + 1.5 * IQR$$

Symmetry and skewness



Box plot and Histogram together

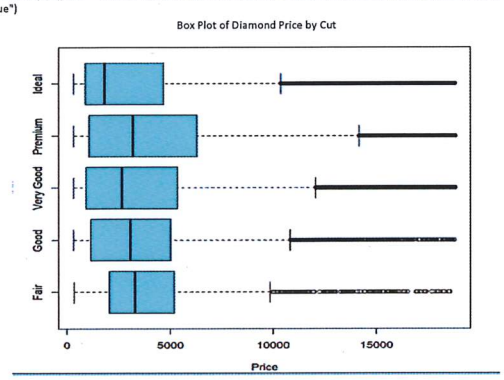


Comparing Groups using descriptive Statistics

```
Diamonds %>% group_by(cut) %>% summarise(Min = min(price,na.rm = TRUE)
## # A tibble: 5 x 10
##       cut   Min    Q1  Median    Q3   Max  Mean    SD     n
##    <ord> <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl> <int>
## 1   Fair   337  2050.25 3282.0 5205.50 18574 4358.758 3560.387 1610
## 2   Good   327  1145.00 3050.5 5028.00 18788 3928.864 3681.590 4906
## 3 Very Good  336  912.00 2648.0 5372.75 18818 3981.760 3935.862 12082
## 4   Premium  326  1046.00 3185.0 6296.00 18823 4584.258 4349.205 13791
## 5   Ideal   326  878.00 1810.0 4678.50 18806 3457.542 3808.401 21551
## # ... with 1 more variables: Missing <int>
```

Comparing Groups using visualisation

```
Diamonds %>% boxplot(price ~ cut,data = ., main="Box Plot of Diamond Price by Cut", ylab="Cut", xlab="Price",horizontal=TRUE,
col = "skyblue")
```



Scatter Plots

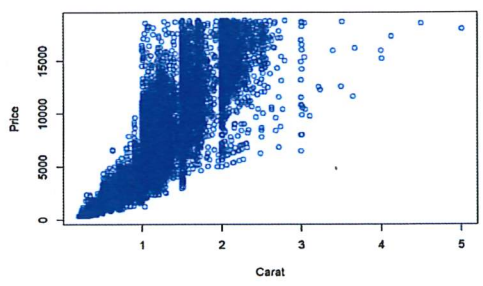
```
#      ID Carat  Price
## 1   1  0.23   326
## 2   2  0.21   326
## 3   3  0.23   327
## 4   4  0.29   334
## 5   5  0.31   335
## 6   6  0.24   336
## 7   7  0.24   336
## 8   8  0.26   337
## 9   9  0.22   337
## 10 10 0.23   338
```

Comparing Groups using visualisation

- Using this plot, confirm the following features:
- Ideal has the smallest median price
- Premium has the highest IQR
- All price distributions are positively skewed
- All price distributions have many suspected outliers
- Fair has the highest Q1
- Premium has the highest Q3
- **Scatter Plots**

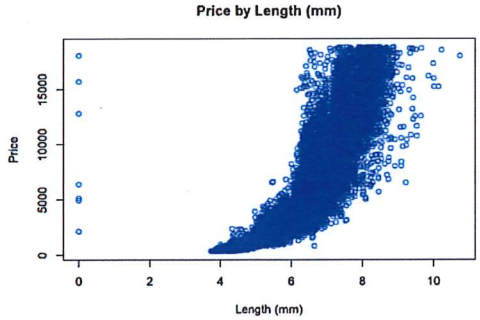
Scatter Plots

```
Diamonds %>% plot(price ~ carat, data = .,ylab="Price", xlab="Carat", col="blue",main="Price by Carat")
```



Scatter Plots

Diamonds %>% plot(price ~ y, data = ., ylab="Price", xlab="Width (mm)", col="blue",main="Price by Width (mm)")



Scatter Plots

Diamonds %>% plot(price ~ y, data = ., ylab="Price", xlab="Width (mm)", col="blue",main="Price by Width (mm)