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.

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

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

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:

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

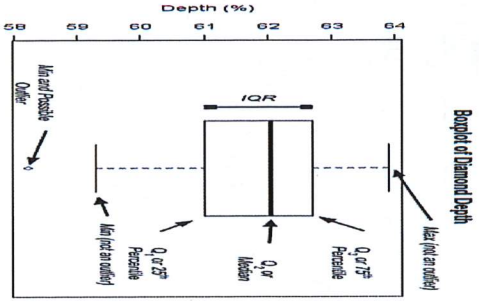
$$\text{Upper outlier} > Q_3 + 1.5 * IQR$$

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.

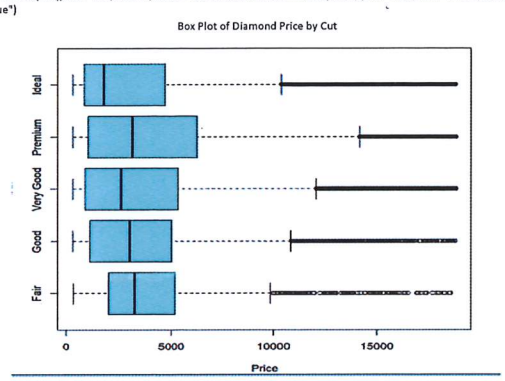
Box Plots

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



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")
```



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

```
#      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
```


Scatter Plots

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