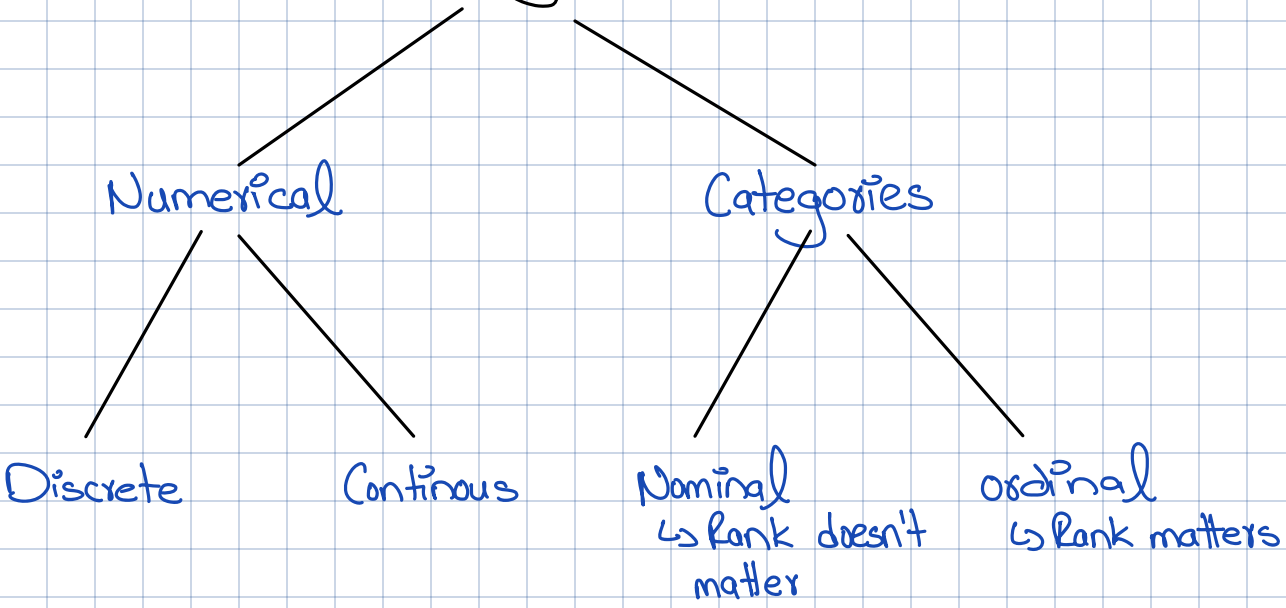
Data Types:



Discriptive Analysis:

↳ Organizes & summarizes data

Eg:
- ↳ Sorting
- ↳ Grouping
- ↳ Creating tables

- ↳ Mean
- ↳ Median
- ↳ Mode

} central tendency

- ↳ Variance
- ↳ Range
- ↳ Standard deviation

} spread of data

Inferential Analysis:

↳ Make inferences & predictions

Eg:
- ↳ Predicting future sales

Graphs:

1. Bar chart:
   - ↳ Represent categorical data
   - ↳ Shows frequency
2. Histogram:
   - ↳ Represents numerical data
   - ↳ Shows distribution of data over bins.

Summarizing 1D continous data:

1. Mean:

$$\text{mean}(\{x_i\}) = \frac{1}{N}\sum_{i}^{N} x_i$$

↳ It's the centre of data.
↳ Describes central tendency of data.
↳ It's sensitive to outliers.

Properties of Mean:

1. Scale:

$$\text{mean}(\{k \cdot x_i\}) = k \cdot \text{mean}(\{x_i\})$$

↳ Changes position of points.
↳ Relative distance b/w points remain same.
↳ Spread of data changes.

2. Translate:

$$\text{mean}(\{x_i + c\}) = \text{mean}(\{x_i\}) + c$$

↳ Changes position of points.
↳ Spread of data remains same.

3. Signed distances sum to 0:

$$\sum_{i}^{N}(x_i - \text{mean}\{x_i\}) = 0$$

4. Mean minimizes sum of squared distance:

$$\underset{\mu}{\text{argmin}} \sum_{i}^{N}(x_i - \mu)^2 = \text{mean}(\{x_i\})$$

Q. Why is there a need to square differences?

↳ Emphasize/Penalize larger values.
↳ Allows positive contribution from all deviations.

Standard deviation:

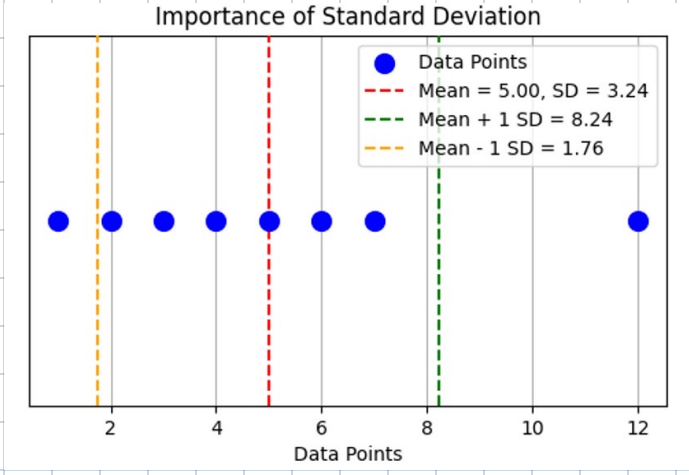
↳ Tells spread of w.r.t mean

$$\text{std}(\{x_i\}) = \sqrt{\frac{1}{N}\sum_{i}^{N}(x_i - \text{mean}(\{x_i\})}$$

$$= \sqrt{\text{mean}\left(\{(x_i - \text{mean}(\{x_i\}))\}\right)}$$

Q. How do identify outliers in data?

↳ Calculate mean
↳ Calculate std
↳ Range: (mean − std, mean + std)
↳ Points outside are outliers



Properties of the standard deviation:

1. Scaling:

$$\text{std}(\{k \cdot x_i\}) = |k| \cdot \text{std}(\{x_i\})$$

↳ Changes position of points.
↳ Relative distance b/w points remain same.
↳ Spread of data changes. i.e std changes

2. Translate:

$$\text{std}(\{x_i + c\}) = \text{std}(\{x_i\})$$

↳ Changes position of points.
↳ Spread of data remains same. i.e std remains same.

Variance:

$$\text{Var}(\{x_i\}) = \frac{1}{N} \sum_{i}^{N} (x_i - \text{mean}(\{x_i\}))^2$$

Properties of Variance:

1. Scaling:

$$\text{Var}(\{k \cdot x_i\}) = k^2 \cdot \text{var}(\{x_i\})$$

2. Translating:

$$Var(\{x_i + c\}) = Var(\{x_i\})$$

Standard Coordinates/Normalize:

↳ Allows us to compare two datasets at different ranges.

$$\hat{x}_i = \frac{x_i - mean(\{x_i\})}{std(\{x_i\})}$$

↳ Standard coordinates always have:
- ↳ Mean: 0
- ↳ Stdev: 1
- ↳ Var: 1

↳ It is unitless.

Median:

↳ Sort data

↳ If odd:

Middle value

$$Median = \left(\frac{n+1}{2}\right)^{th}$$

↳ If even:

Mean of two middle values

$$Median = \frac{\left(\frac{n}{2}\right)^{th} + \left(\frac{n}{2}+1\right)^{th}}{2}$$

Properties of Median:

1. Scale:

$$median(\{k \cdot x_i\}) = k \cdot median(\{x_i\})$$

2. Translate:

$$median(\{x_i + c\}) = median(\{x_i\}) + c$$

Percentile:

↳ k^{th} percentile means k% of data is smaller than OR equal to it.

Total data points

$$P = \frac{n \cdot k}{100} \rightarrow \text{Percentile}$$

If P is in points:

$$P = X_A + (X_B - X_A) \times \text{point}$$

- X_A: p^{th} value
- X_B: $(p+1)^{th}$ value
- point: value after decimal point

Interquartile Range (IQR):

- iqr = (75th percentile − 25th percentile).
- Tells spread of data between the 75th & 25th percentile.
- Less sensitive to outliers.
- Can identify outliers.

Q. What indicates a large OR small iqr in a dataset?

Relative Size

Context:

- Compare IQR to range of data.
- IQR of 17 with range 60 − 95 indicates moderate variation.
- Range: $95 - 60 = 35$
- Proportion: $\frac{17}{35} = 0.49$
  - How big is 50% values compared to entire data.
  - 50% of scores take up half of total spread, which indicates moderate variability

Benchmarking:

- Compare IQR to similar datasets.
- In education, 10−15 IQR is normal
- So, 17 seems too large.

Scale of measurement:

1. Unit matters:
   - IQR of 17 is significant spread in middle 50% of the students
   - IQR if 5 may be small in a dataset of heights in cm.
2. Domain knowledge:
   - Use your experience to assess the IQR.
   - IQR of 10 is normal for tests

Comparison with standard deviation:

- Small IQR, data is clustered
- stdev: 11 , IQR: 17
- Significant spread of students in middle 50% of the students

Visualization:

- Use box plots

Properties of iqr:

1. Scale:

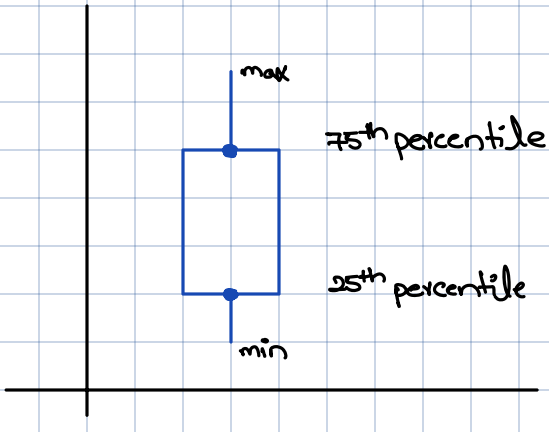
$$iqr(\{k \cdot x_i\}) = k \cdot iqr(\{x_i\})$$

2. Translate:

$$iqr(\{x_i + c\}) = iqr(\{x_i\})$$

Box plot:

- Max: 75th percentile + (iqr × 1.5)
- Min: 25th percentile − (iqr × 1.5)



- Above max & below min are outliers.
- If max of data is NOT in range of max calculated, then we use max as the max of data. Similarly, for min.

Example,

Data: [15, 8, 18, 12, 11, 20, 2, 3, 5, 7, 1, 18, 15]

1	2	3	5	7	8	11	12	15	15	18	18	20	
1	2	3	4	5	6	7	8	9	10	11	12	13	

$P_{25} = \frac{13 \times 25}{100}$

$= 3.25$

$P_{75} = \frac{13 \times 75}{100}$

$= 9.75$

$P'_{25} = 3 + (5-3) \times 0.25$

$= 3.5$

$P'_{75} = 15 + (15-15) \times 0.75$

$= 15$

Range: $15 - 3.5$

11.5

Max $= 15 + (11.5 \times 1.5)$

$= 32.25$

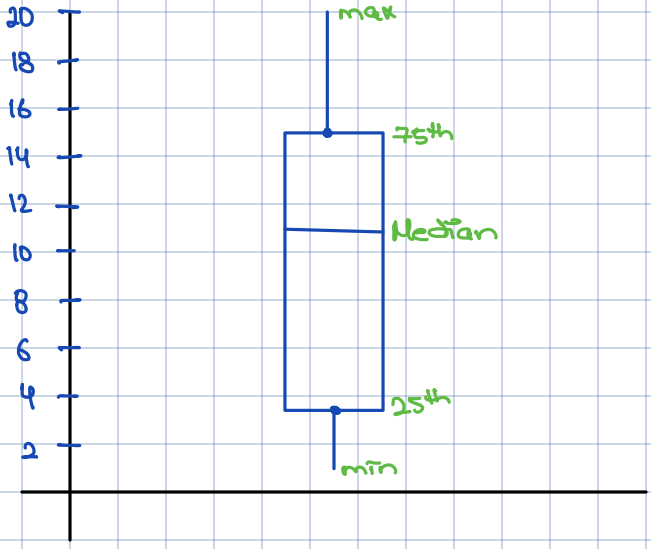
$= 20$

Min $= 3.5 - (11.5 \times 1.5)$

$= -13.75$

↳ As this is far too less than 1 so we say

$= 1$



Median $= \frac{13+1}{2} = 7^{th}$

$= 11$

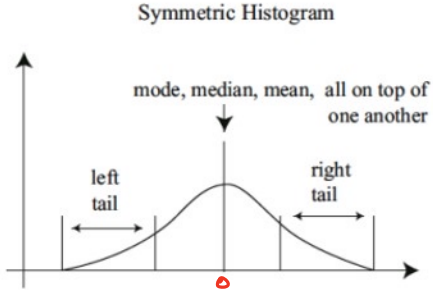
Modes:

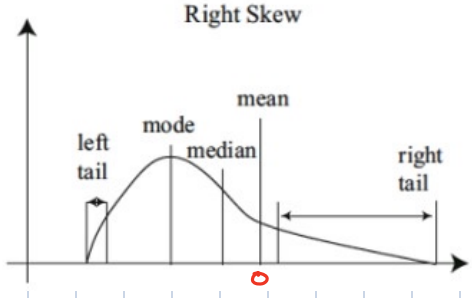
↳ The most frequent value.

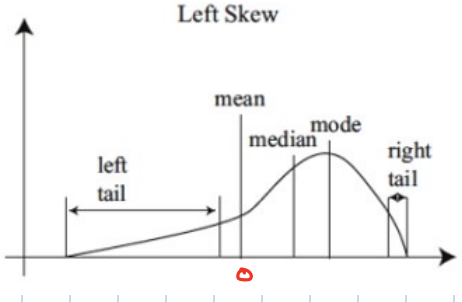
↳ Peaks in histogram.

↳ If there are multiple peaks then the data is bimodal & there are sub groups in the population.

Tails and Skews:



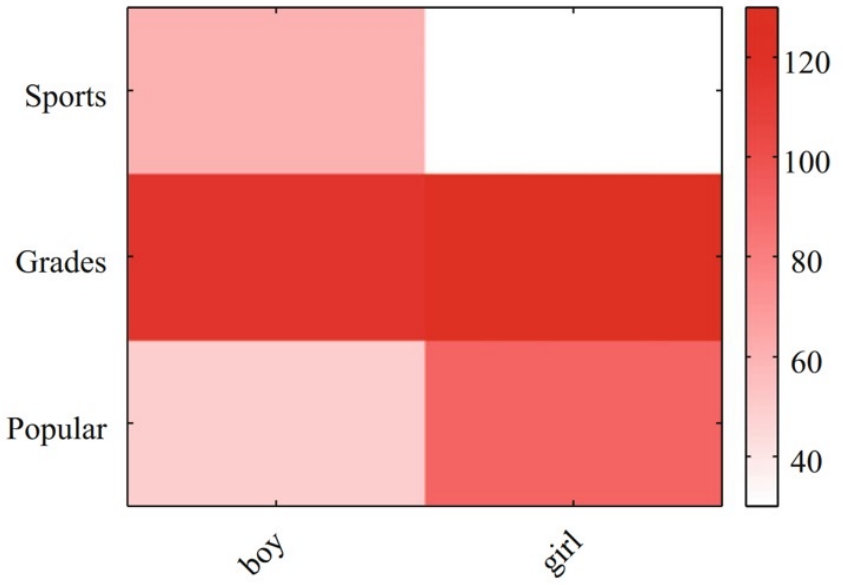




↳ Mean is always in the middle.
↳ Mode is the peak.
↳ Median's in b/w.

Visualing relationships with a heatmap.

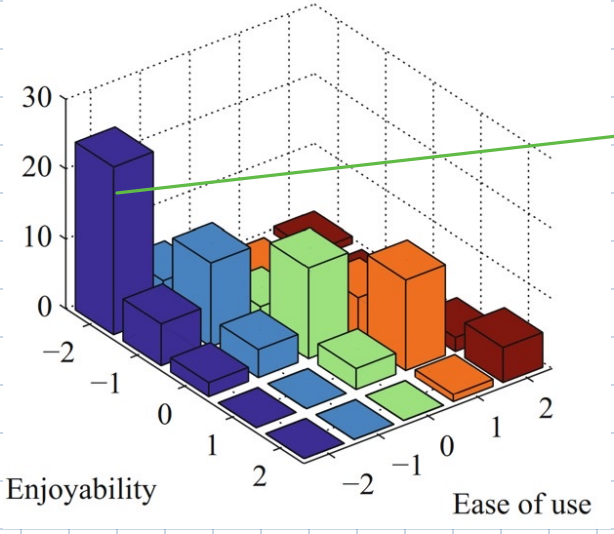
↳ Individual values are represented by color.



↳ Each cell corresponds to count of numbers of elements of that type.
↳ This map tells:
  ↳ Both boys & girls like grades
  ↳ Boys like sports more
  ↳ Girls like popularity more.

Visualising with a 3D bar chart.

↳ Consider a dataset of user rating on enjoyability & ease of use of a software.



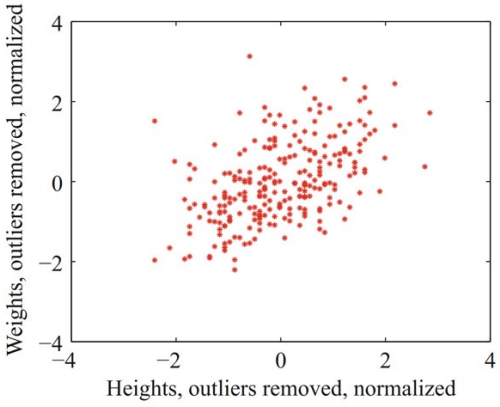
↳ Part of the plot is hidden

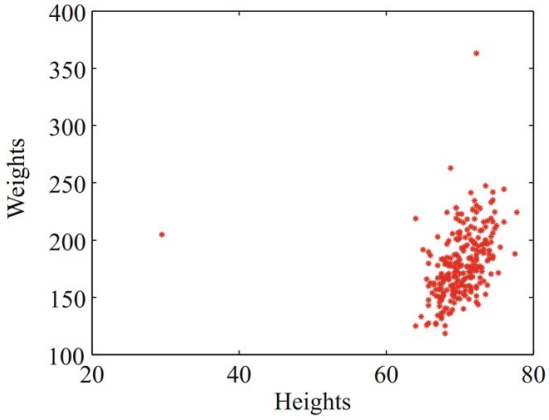
→ Those who found it hard to use also didn't enjoy it.

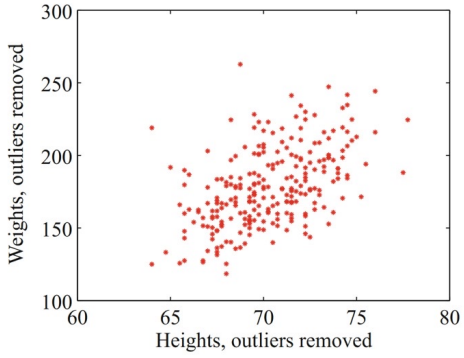
Disadvantages for heatmap & 3D bar plot:

1. For discrete categorical data.
2. Assumes direct relation

Scatter plot:







Correlation

1. Positive correlation:

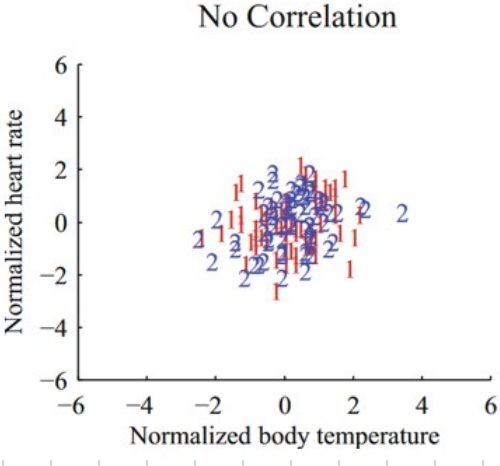
   ↳ When one variable increases, the other also increases.

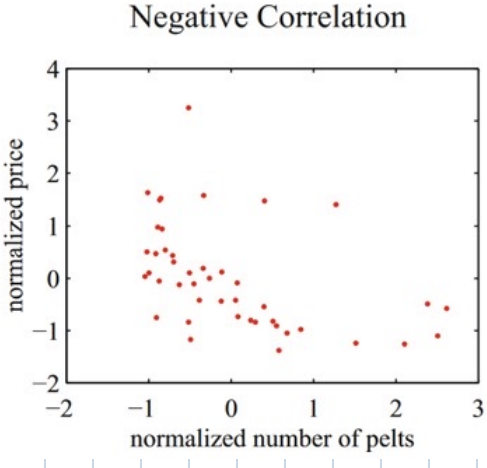
2. Negative correlation:

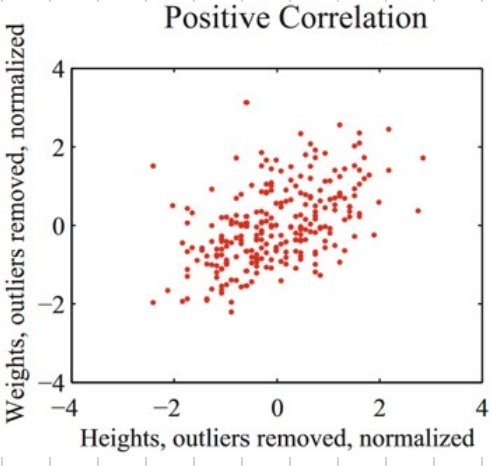
   ↳ When one variable increases, the other decreases.

3. Zero correlation:

   ↳ No relation between variables.







↳ Correlation doesnot mean causation.

Assignment 01:

Task 01:

$$\hat{x} = \frac{x_i - \text{mean}\{x_i\}}{\text{std}\{x_i\}}$$

$$\text{mean}\{\hat{x}_i\} = \sum_i^n \frac{1}{n} \hat{x}_i$$

$$= \frac{1}{n} \sum_i^n \frac{x_i - \text{mean}\{x_i\}}{\text{std}\{x_i\}}$$

$$= \frac{1}{n} \cdot \frac{1}{\text{std}\{x_i\}} \sum_i^n (x_i - \text{mean}\{x_i\})$$

$$\sum_i^n (x_i - \text{mean}\{x_i\}) = \sum_i^n x_i - \sum_i^n \text{mean}\{x_i\}$$

$$= \sum_i^n x_i - n\,\text{mean}\{x_i\}$$

$$= \sum_i^n x_i - \cancel{n} \cdot \frac{1}{\cancel{n}} \sum_i^n x_i$$

$$= \sum_i^n x_i - \sum_i^n x_i$$

$$= 0$$

$$= \frac{1}{n} \cdot \frac{1}{\text{std}\{x_i\}} \times 0$$

$$= 0$$

Task 02:

$$d = \{572, 572, 573, \overset{\times}{568}, \overset{\times}{569}, 575, \overset{\times}{565}, \overset{\times}{570}\}$$

1. $\text{mean} = \frac{572 + 572 + 573 + 568 + 569 + 575 + 565 + 570}{}$

8

$= 570.5$

1 2 3 4 5 6 7 8

565, 568, 569, 570, 572, 572, 573, 575

$$\text{median} = \frac{\frac{8^{th}}{2} + \left(\frac{8}{2}+1\right)^{th}}{2}$$

$$= \frac{570 + 572}{2} = 571$$

2. Var = 8.75

std = $\sqrt{8.75}$ = 2.95

Range: 575 − 565

= 10

4. The tires seem to be of good quality. Most of the tires lie around 570 with small std of 2.95

3. The data relatively balanced but median > mean shows presense of smaller values ∴ slightly skewed.

Task 3:

919 1196 785 1126 936 918 1156 920 948 1067 1092 1162 1170 929 930 905 972 1035 1045 855 1195 1195 1340
1122 938 970 1237 956 1192 1157 978 832 1009 1187 1151 1009 765 958 902 1022 1333 811 1217 1085 896 958
1311 1037 702 923

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16

702, 765, 785, 811, 832, 855, 896, 902, 905, 918, 919, 920, 923, 929, 936, 938,

17 18 19 20 21 22 23 24 25 26 27 28 29

948, 950, 956, 958, 958, 970, 972, 978, 1009, 1009, 1022, 1035, 1037,

30 31 32 33 34 35 36 37 38 39 40 41

1045, 1067, 1085, 1092, 1102, 1122, 1126, 1151, 1156, 1157, 1162, 1170,

42 43 45 46 47 48 49 50

1195, 1195, 1196, 1217, 1237, 1311, 1333, 1340

$P_{25} = \frac{25 \times 50}{100}$

$= 12.5$

$P'_{25} = 920 + (923 - 920) \times 0.5$

$= 921.5$

$P_{75} = \frac{75 \times 50}{100}$

37.5

$P'_{75} = 1151 + (1156 - 1151) \times 0.5$

$= 1153.5$

Range: $1153.5 - 921.5$

$= 232$

min: $921.5 - 232 \times 1.5$

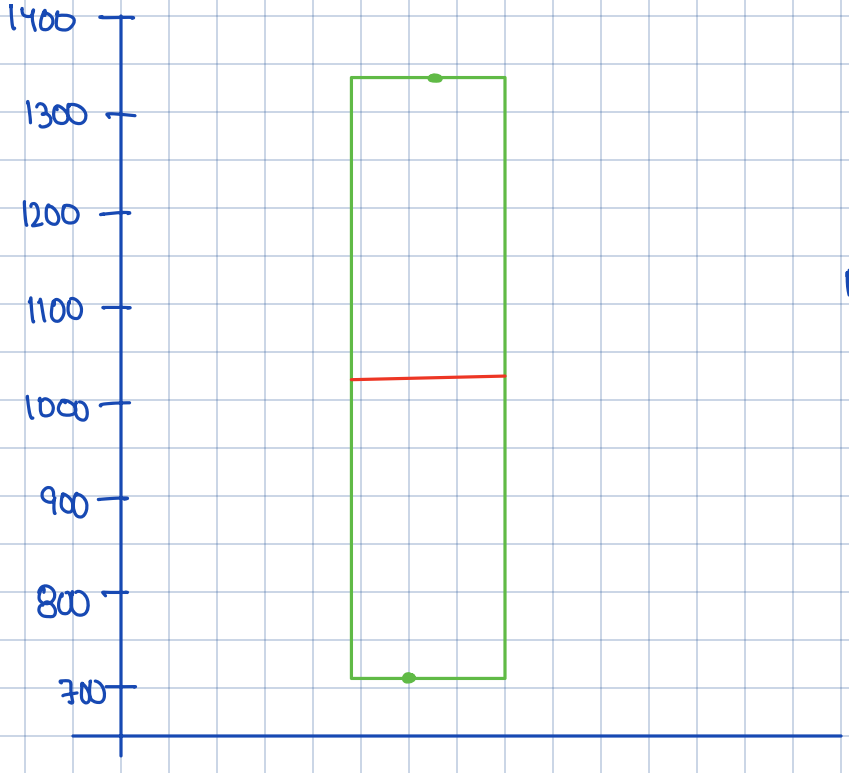
$= 573$

$= 702$

max $= 1153.5 + (232 \times 1.5)$

$= 1501.5$

$= 1340$



$\text{Median} = \frac{\frac{50^{th}}{2} + \left(\frac{50}{2} + 1\right)^{th}}{2}$

$= \frac{1009 + 1009}{2}$

The boxplot tells that there is quite alot of variation in the data.

Task 4:

6.X2 6.X7 6.82 6.X0 6.X8 6.X0 6.X2 6.X5 6.X6 6.X6 6.X4 6.X6 6.X3 6.X0 6.X2 6.X6 6.X6 6.X8 6.X6 6.X2 6.X2 6.X6 6.X0 6.X8
6.X6 6.X7 6.X0 6.X2 6.X4 6.X1 6.X9 6.X8 6.X6 6.X6 6.X6 6.X2

1 2 4 1 1 4
6.64, 6.62, 6.62, 6.66, 6.66, 6.66, 6.66, 6.67, 6.68, 6.70, 6.70, 6.70, 6.70,
5 1 1 1 1 7
6.72, 6.72, 6.72, 6.72, 6.72, 6.73, 6.74, 6.75, 6.76, 6.76, 6.76, 6.76,
1 3 7 1 1 7
6.76, 6.76, 6.76, 6.77, 6.78, 6.78, 6.78, 6.79, 6.80, 6.81, 6.82

$n = 36$

bins = 6

mean : 6.726

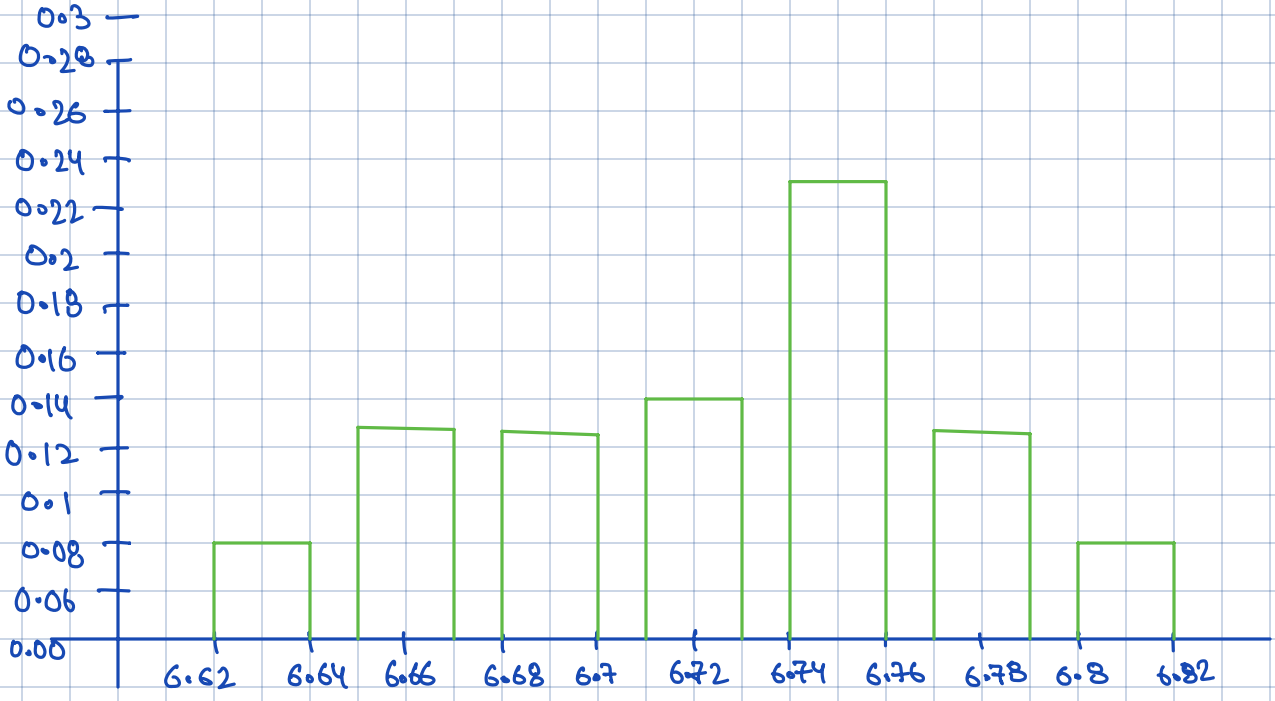
std : 0.0536

$$\text{width} = \frac{6.82 - 6.64}{6}$$

$$= 0.03$$

3 5 5 6 9
6.62 – 6.64, 6.65 – 6.67, 6.68 – 6.70, 6.71 – 6.73, 6.74 – 6.76,

5 3
6.77 – 6.79, 6.80 – 6.82



3. Almost symmetric but with some skewness.

Properties of correlation:

- Translating the data doesnot effect the correlation
- Scaling the data may change the sign of correlation

$$corr(\{(ax_i+b, cy_i+d)\}) = sign(a,c)\, corr(\{(x_i,y_i)\})$$

- Correlation is bounded between $[1,-1]$
  - $Corr(\{(x_i,y_i)\}) = 1$ if $\hat{x} = \hat{y}$
  - $Corr(\{(x_i,y_i)\}) = -1$ if $\hat{x} = -\hat{y}$

Predictions using correlation:

1. $mean(\{u\}) = 0$

$$mean(\{u\}) = mean(\{\hat{y} - \hat{y}_p\})$$
$$= mean(\{\hat{y}\}) - mean(\{\hat{y}_p\})$$
$$= mean(\{\hat{y}\}) - mean(\{a\hat{x}+b\})$$
$$= mean(\{\hat{y}\}) - a\, mean(\{\hat{x}\}) + b$$
$$0 - a(0) + b$$

$\therefore b = 0$

2. $Var(\{u\}) = Var(\{\hat{y} - \hat{y}_p\})$

$$= Var(\{\hat{y} - a\hat{x} + b\})$$
$$= Var(\{\hat{y} - a\hat{x}\})$$
$$= mean(\{\hat{y} - a\hat{x}\}^2)$$
$$= mean(\{\hat{y}^2 - 2a\hat{x}\hat{y} + a^2\hat{x}^2\})$$
$$= mean(\{\hat{y}^2\}) - 2a\, mean(\{\hat{x}\hat{y}\}) + a^2\, mean\{(\hat{x})^2\}$$
$$= 1 - 2as + a^2$$

$$\frac{d\,var(\{u\})}{da} = -2\gamma + 2a$$

$$-2\gamma + 2a = 0$$

$$a = \gamma$$

$$\hat{y}_p = a\hat{x} + b$$
$$= \gamma\hat{x}$$

Procedure:

1. Find $\hat{x}$ & $\hat{y}$
2. Find $corr(\{\hat{x}_i, \hat{y}_i\}) = \frac{1}{N}\sum_i \hat{x}_i \hat{y}_i$
3. $y_i^p = std(y) \cdot \gamma \cdot \hat{x} + mean(\{y\})$