Categorical data

classifies individuals or items into different group.

id	marital status
62229	1
62230	3
62231	1
62232	1
62233	4
...	...

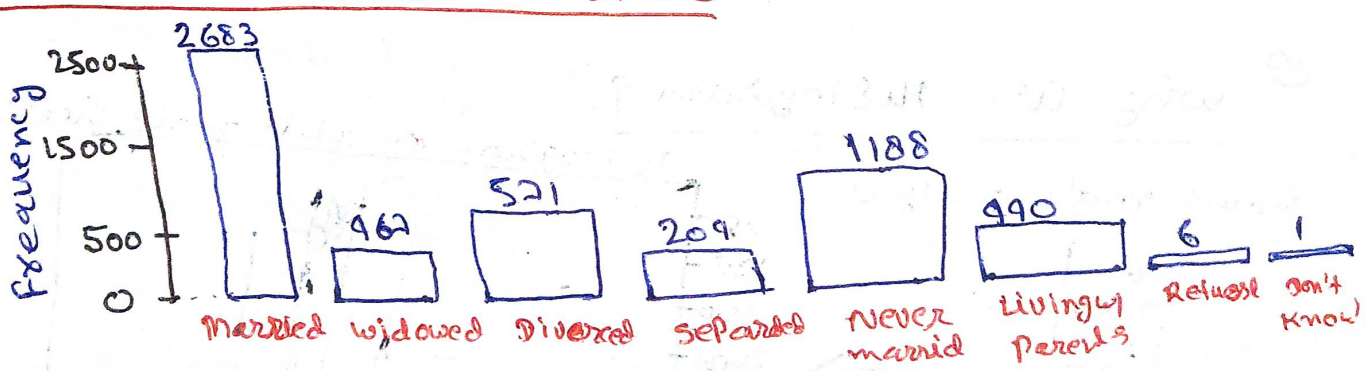
1. married
2. widowed
3. Divorced
4. separated
5. Never married
6. Living w/ Partner
7. Refused
8. Don't know

Frequency Table

- Counts
- Percentage

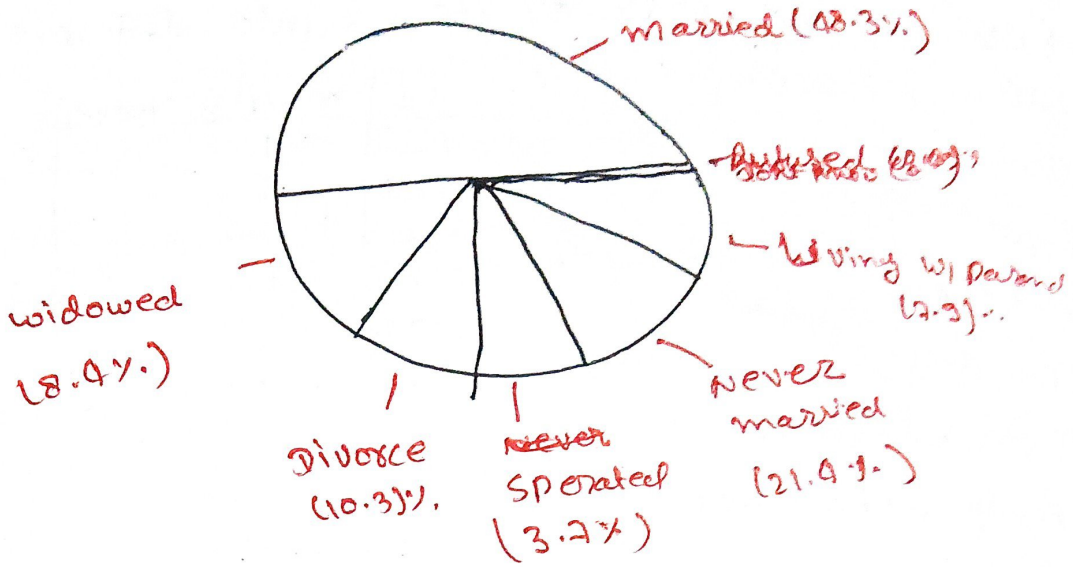
Marital Status	Count	Percent
Married	2683	48.3%
Widowed	467	8.4%
Divorced	571	10.3%
Separated	209	3.7%
Never married	1188	21.4%
Living w/ Parents	440	7.9%
Refused	6	0.1
Don't know	1	0.0%
Total	5560	100%

Bar chart of marital status



we can also ~~Percent~~ use Percentage instead of Count.

Pie chart of marital Status



→ Frequency Table
  ↳ creat for numeral summaried

→ Bar chart
  ↳ creat for visualization

Histogram

what are Quantitative variables ?

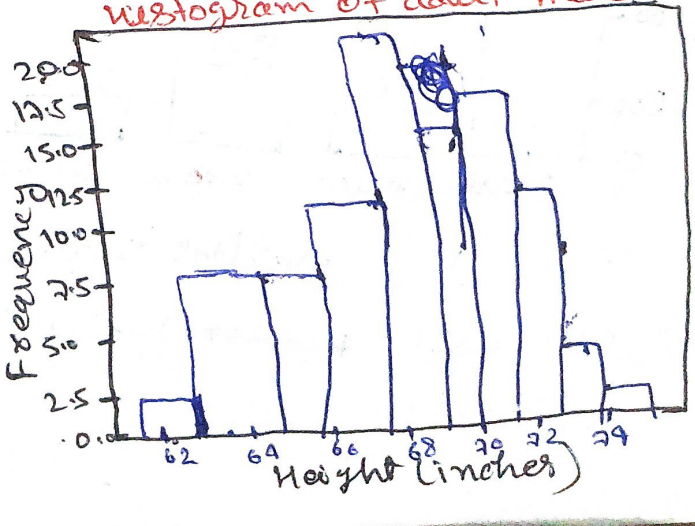
Variables that have a numerical value that we can perform mathematical operations on

Ex-> Height, weight, income, test scores.

why use Histogram ?

Adult male height

66.3
2.5.1
62.9
6.7.6
70.0
69.9
64.8
--



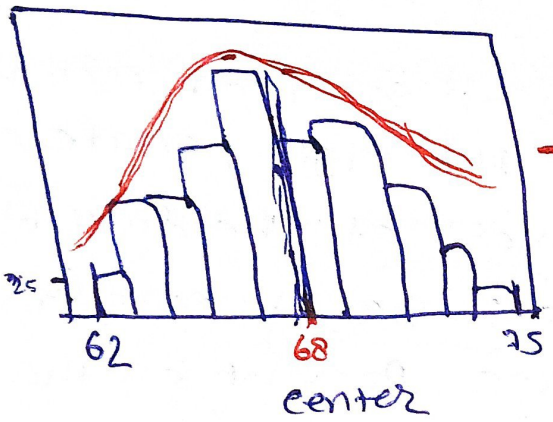
4 main aspect

Shape → overall apperance of Histogram. can be symmetric, bell-shaped, left skewed, right skewed etc

Center → mean or median

Spread → How far our data spreads .. Range, interQuartile range (IQR). Standard deviation, variance.

Outliers → Data Points that fall far from the bulk of the data.



→ Bell-shaped / uni-modal

center ≒ 68 = median

mean = 68

Spread ⇒ Range = max − min

= 75 − 62

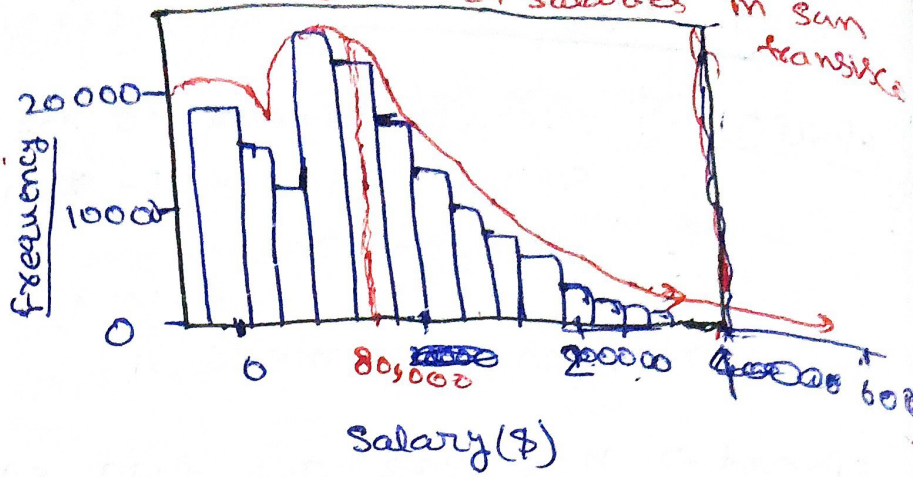
= 13

Outlier → No Apperent outliers

→ The distribution of adult male heights is roughly bell shaped with a center of about 68 inches a range of 13 inches (62 to 75) and no apparent outliers.

⇒ Bar chart use for categorical data and Histogram use for Quantitative data

Histogram of Salaries in San Francisco



Shape →

Right Skewed

Binodal

center →

median → 80,000

mean → 85000

Spread →

Range → 60,0000 − 0

= 600000

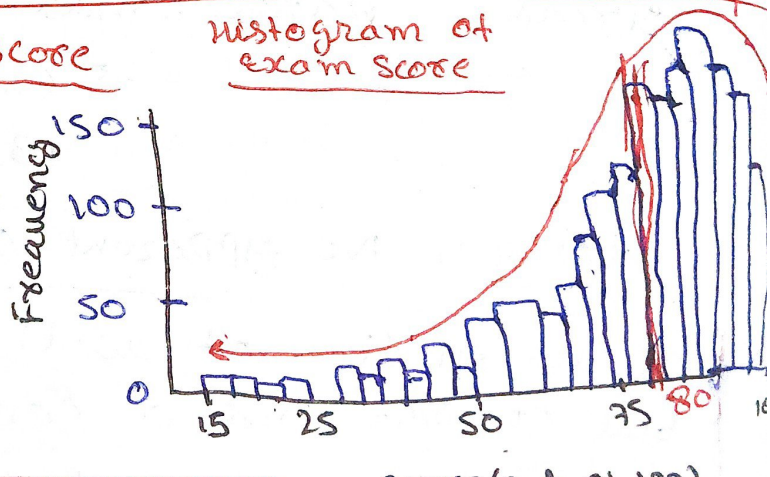
Outliers →

High end

The distribution of salaries in San Francisco is bimodal and skewed to the right; centered at about $ 80,000 with most of the data btw $40,000 and $120,000, a range of roughly $ 60,0,000 and outliers are present on the higher end.

Exam score

Histogram of exam score



median = 80

Range = 100 − 15

= 85

Outlier = lower end

The distribution of exam score is skewed left centered at abaut 80 Points with most score being btw 65 and 90 Points; a range of roughly 85 and some outlier are present below 50 Points.

Numerical summaries

Adult male height

5 number summary

- min
- 1st Quartile (25%)
- median (50%)
- 3rd Quartile (75%)
- max

Height

min. : 61.7
1st Qu. : 66.5
median : 68.3
mean : 68.3
3rd Qua : 70.1
max. : 75.1

$$IQR = Q_3 - Q_1$$
$$= 70.1 - 66.5$$
$$= 3.6$$

Salaries in San Francisco (2011-2017)

~~5 number summary~~

min	25%	50%	75%	max	mean	SD	n
-618.1	36169	71427	105839	567595	74768	50517	148654

- min: Input error because salaries can't be in negative
- 25% = Q_1, 50% = median, 75% = Q_3; Q_1 to Q_3 = IQR
- SD: Standard deviation
- n: Sample size

mean ≠ median
↳ outliers exist

Exam score

min	14.0	(14%)
25%	68.0	(68%)
med	78.0	(78%)
75%	82.0	(82%)
max	100.0	(100%)
mean	76.3	(76.3%)
SD	12.4	

n = 1802.0

SD = 12.4

IQR = 82 - 68 = 14

IQR tells us where is most of data falls

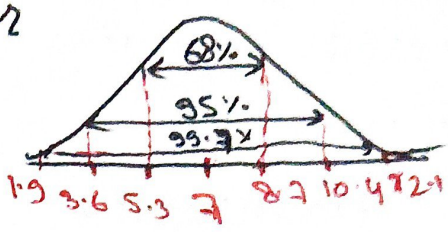
mean < median
↳ Left Skewed

→ mean is sensitive to extreme observations. and the few unusually low scores will tend to bring the mean a bit lower than the median

Amount of sleep (2-2-2)

→ μ = 7 hour

→ standard deviation (σ) = 1.2 hour



SD ⇒ roughly the avg distance of the values from their mean.

Empirical rule (68-95-99.7)

For normal distributions:

→ 1SD ($\mu \pm \sigma$) = 68% of values

→ Range : 5.3 to 8.7 hours

→ 2 SD ($\mu \pm 2\sigma$) = 95% of values

→ Range : 3.6 to 10.4 hours

→ 3SD ($\mu \pm 3\sigma$) = 99.7% of values

Range : 1.9 to 12.1 hours

> very useful to judge how unusual/extreme a value is

Z-score (standard score)

→ It tells us how unusual they are compared to the mean.

→ Formula → $z = \frac{x - \mu}{\sigma}$

x = observation
μ = mean
σ = SD

→ interpretation

→ Positive z → above mean

→ negative z → below mean

→ magnitude → how far (in SD units) from mean.

Ex→

(1) Reed (10 hrs sleep)

$$Z = \frac{10-7}{1.7} = 1.76$$

→ above average, fairly unusual (close to upper 95% limit)

(2) Roomate (6 hrs sleep)

$$Z = \frac{6-7}{1.7} = -0.59$$

→ slightly below average, not unusual

(3) Mark ($Z = -2.7$)

Observation

$$x = z\sigma + \mu$$

$$= -2.7 \times 1.7 + 7$$

$$= 2.41 \text{ hrs}$$

→ very low sleep, rare situation

Boxplot

What is a Boxplot?

→ graphical representation of the five number summary

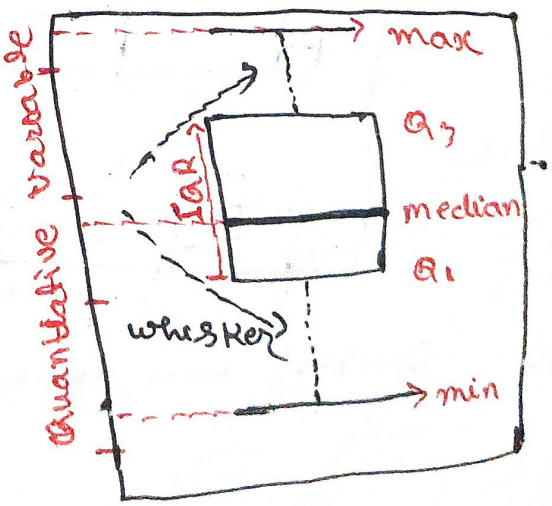
Five number summary

min	Q_1	median	Q_3	max
		center		

IQR

Range

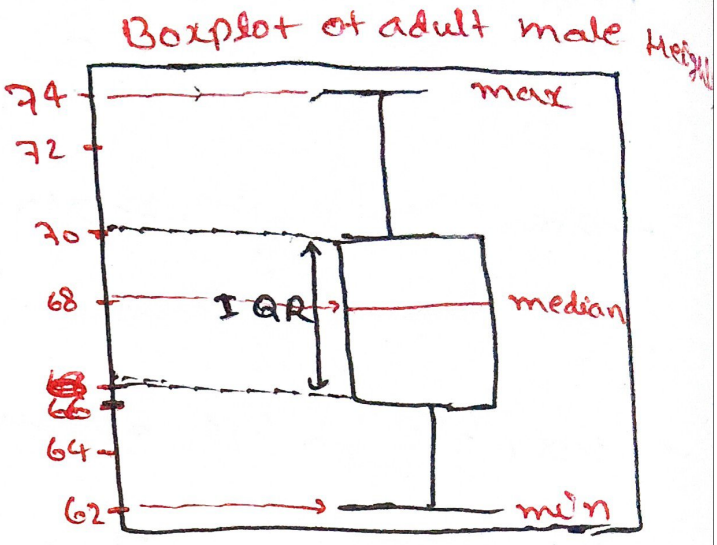
Whiskers :→ Line extending from box to min and max



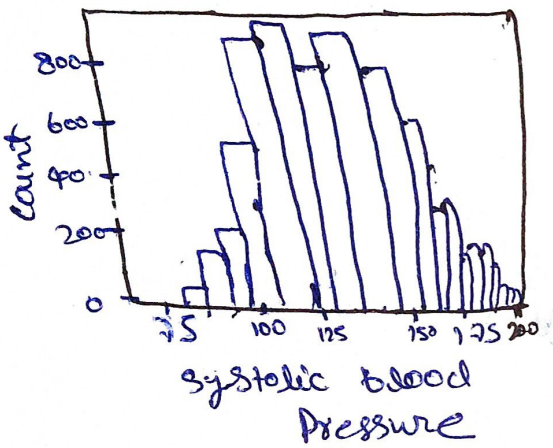
Ex→ heights of adult males

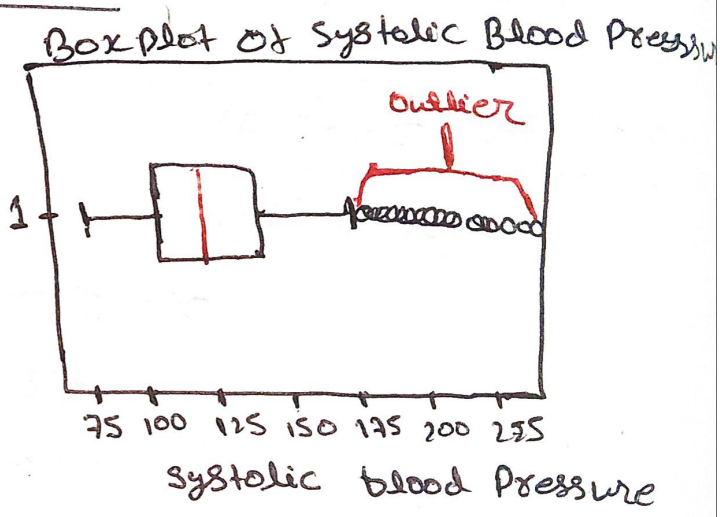
height

min. : 61.7
1st Qu : 66.5
median : 68.3
mean : 68.3
3rd Qu : 70.1
max : 75.1



Ex→ Systolic Blood Pressure

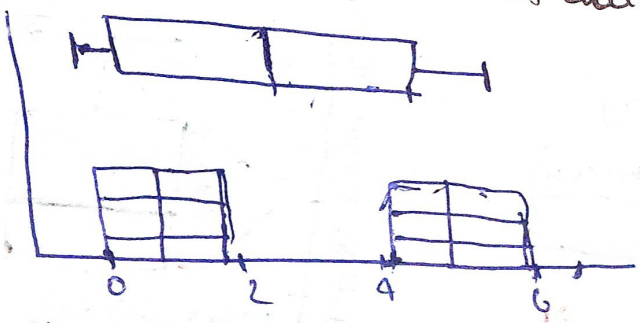




→ Boxplots can help identify outliers

Ex→ Quiz score

Boxplot and Histogram of Quiz score



Boxplots can hide gaps and clusters

advantange of Boxplot

→ summarized data quickly with five number summary.

→ show :

- center → median
- spread → IQR and overall range
- outliers → plotted seperately

- Excellent for comparison across groups.

limitations

→ Dont show the exact shape

→ Histograms are better for shape

→ Boxplots focus on summary + comparisons