Descriptive Statistics

19 December 2022 08:06

Statistical methods

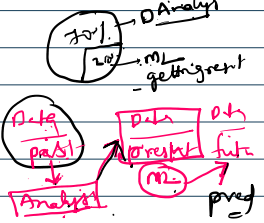
Descriptive | Inferential Statistics

Stats:- Representation of your data in terms of graphs or Tables and applying some statistical methods to get interpretations / results about your data

DA: ① - read ② - analysis

ML: ① - read data ② - clean data ③ - preprocess data ④ - model building

Statistical methods



Descriptive Stat: Is Analysation of past data and get the pattern for past data which are useful for present / future data.

Analysation → patterns

→ prediction (ML) | Hypothesis → Inferential

Descriptive :-

- Measures of Central Tendency
- Measures of Dispersion / Deviation
- Measures of Shape
- measures of relation

n: 10, 11, 12, 13 — $\frac{10+11+12+13}{4} = \square$

Central Tendency:

① mean :- Sum of all obsv° / no of the obsv°s ($\bar{n}$) / (μ) $\bar{n} = \frac{\sum n}{n}$ (summation)

② median :- middle value / which is devided into two equal parts of data

③ mode

① 7, 15, 13, 6, 4, 2, 10, ~~100~~

mean = $\frac{7+15+13+6+4+2+10}{7} = \frac{57}{7} = 8.14$

median :- 2, 4, 6, 7, 10, 13, 15 ⇒ m = 7

median: $\frac{n^{th} + (\frac{n}{2}+1)^{th}}{2}$, $n = 8$

2, 4, 6, 7, 10, 13, 15, 100

median = $\frac{(\frac{8}{2})^{th} + (\frac{8}{2}+1)^{th}}{2} = \frac{4^{th} + 5^{th}}{2} = \frac{7+10}{2} = \frac{17}{2} = 8.5$

② 7, 15, 13, 6, 4, 2, 10, 100

mean = $\frac{157}{8} = 19.6$

mean = 19.6

mean = 8.14

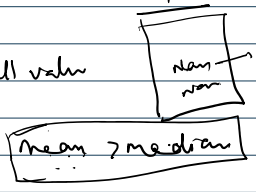
median = 8.5

median = 7

19.6 > 8.5

Note:- mean is effected by the outliers but not median

CT:- ① Imputation of the ~~data~~ null value ② Identifying the outliers



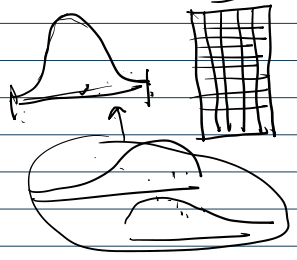
mode :- High concentration / highly repeated value

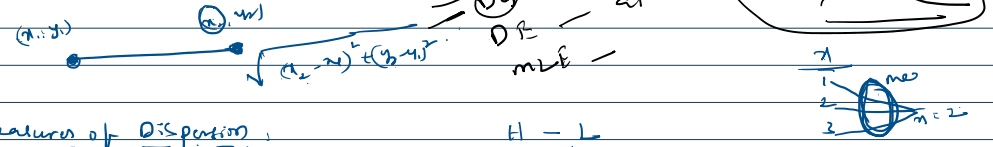
2, 4, 5, 1, 2, 4, 1, 2, 2 → mode: 2 - 4 times

2, 4, 5, 1, 2, 4, 1, 2, 2 → mode: 2 – 4 times, 1 – 2 times

Numerical	Categorical
① Imputation of null value never mean / median	① Imputation of null value we use mode

Salary





Measures of Dispersion: H – L

① Range :- Max – min

② SD – Standard Deviation : Distance the Complexity

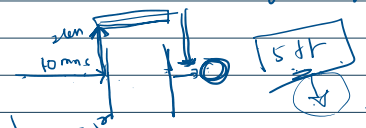
③ variance

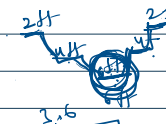
SD :- Distance the original point / observation to mean

$$SD = \sqrt{\frac{\sum (x - \bar{x})^2}{n}} = \sigma \text{ (sigma)} \quad \text{(Root mean Squared Error)} \quad RMSE$$

$$\text{variance} \quad \sigma^2 = \frac{\sum (x - \bar{x})^2}{n} \quad \text{(mean squared error)}$$

*** SD is also identifying of the distribution





$\frac{2+4+6+4+2}{5} = \frac{18}{5} = 3.6$ mean

x	$(x - \bar{x})$	$(x - \bar{x})^2$
2	$2 - 3.6 = -1.6$	2.56
4	$4 - 3.6 = 0.4$	0.16
6	$6 - 3.6 = 2.4$	4.86
4	$4 - 3.6 = 0.4$	0.16
2	$2 - 3.6 = -1.6$	2.56
		10.30

$\sqrt{\frac{8.1}{5}} =$

$\sigma = \sqrt{\frac{(x - \bar{x})^2}{n}}$

2 ft	3.6
4 ft	1.4
6 ft	0.1
4 ft	0.2
2 ft	

$SD = 1.4$

$\sigma = \sqrt{\frac{10.3}{5}} = \sqrt{2.06}$

$\frac{16}{5} = 3.2$

$SD = 1.4$ → 2 mean = 3.6 → 2.2

$SD = 0.4$

x	$(x - \bar{x})$	$(x - \bar{x})^2$	$(x - \bar{x})^2$
2	$2 - 3.2 = -1.2$		1.44
4	$4 - 3.2 = 0.8$		0.64
4	$4 - 3.2 = 0.8$		0.64
4	$4 - 3.2 = 0.8$		0.64
2	$2 - 3.2 = -1.2$		1.44
			4.80

$\sigma = \sqrt{\frac{4.80}{5}} = \sqrt{0.96}$

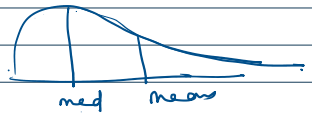
$SD = 0.97$

3.2

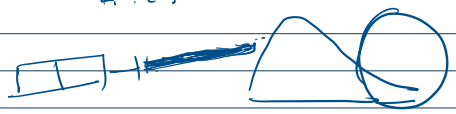
```
diamond_df['price'].describe()

count    53940.000000
mean      3932.799722
std       3989.439738
min        326.000000
25%        950.000000
50%       2401.000000
75%       5324.250000
max      18823.000000
Name: price, dtype: float64
```

Q1
Q2/M
Q3

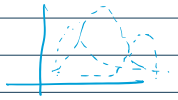
mean > med positive skewness



SD > mean



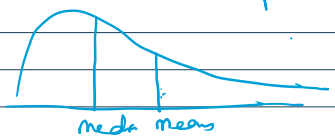
Measures of shape: Skewness → scatterdness



① Mean > median → positive skewness

$Sk > 0$

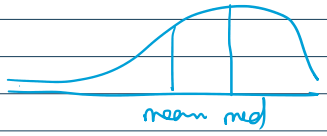
$Sk > 0.5$



② mean < median → negative skewness

$Sk < 0$

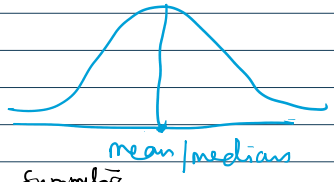
$Sk < -0.5$

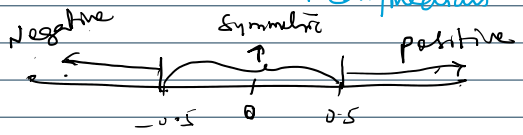


③ Mean ≅ median → Symmetric skewness / Normal curve

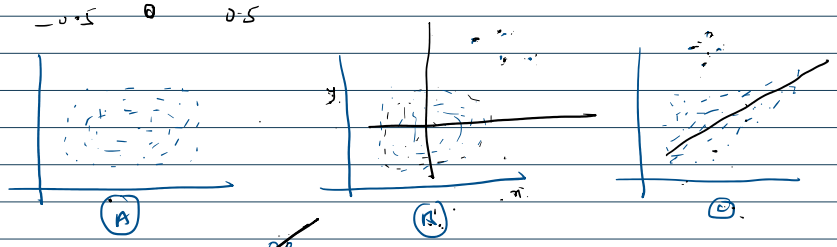
$Sk \cong 0$

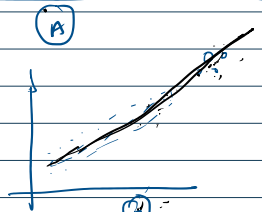
$-0.5 \leq Sk \leq 0.5$





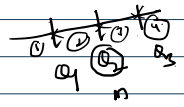
Outliers :-

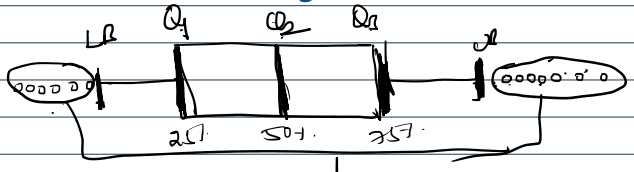


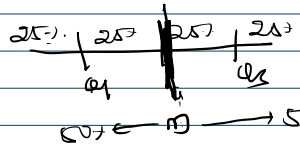


Correlation :

Quartiles :







Outliers Interquartile range:

iqr :-

$UB = Q_3 + (1.5 * iqr)$

$LB = Q_1 - (1.5 * iqr)$

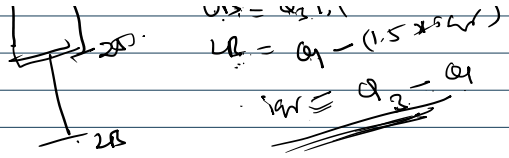
$= Q_3 - Q_1$

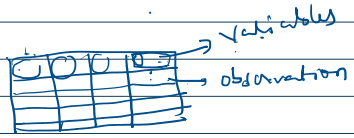
Q1 – Lower Quartile – 25%.

Q3 – Upper Quartile – 75%.

Q2 – Median – 50%.

Variables

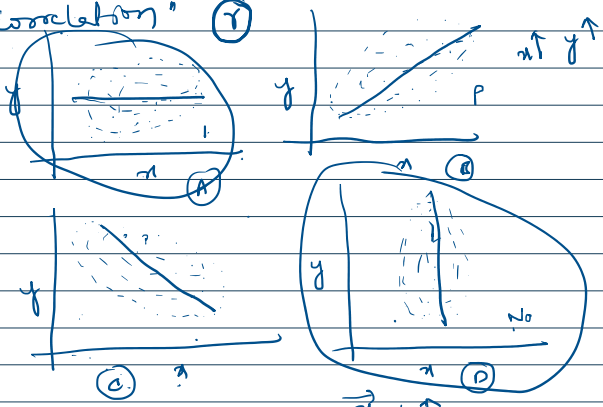




Measures of relation: relation b/w two variables/columns is called "correlation"

① positive correlation
② Negative correlation

change in one variable the leads to change in other variable is called having relation



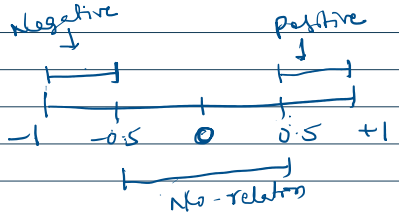
① If we have opposite direction in the change of variable is called Negative

$x \uparrow y \downarrow$ | $x \downarrow y \uparrow$ $\quad r < 0$

② If we have same direction in the change of variable, is called positive correlation

$x \uparrow y \uparrow$ | $x \downarrow y \downarrow$ $\quad r > 0$

$1 > r \geq 0.5$ $\quad r \geq 0$ no-correlation



age | salary

① Because if the age decreases my salary is also decreasing
Increasing my salary is also increasing

Corr = 0.7 (70%)

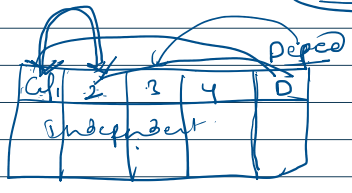
age ↓ salary ↓ } → positive $\quad +0.5 < Cor < 1$
age ↑ salary ↑

-0.6 (60%)

age ↑ salary ↓ } → Negative $\quad -0.5 < Cor < -1$
age ↓ salary ↑

Coffee → | Milk, water, coffee powder, Sugar ①

→ Milk with water, coffee powder, Sugar, ② coffee powder, Sugar, water, milk



In independent variable if we have relation b/w the variable is called "multi-collinearity"

To identify the multicollinearity we have

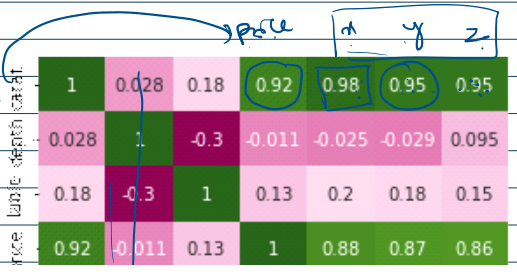
VIF

↓

Variance Inflation factor

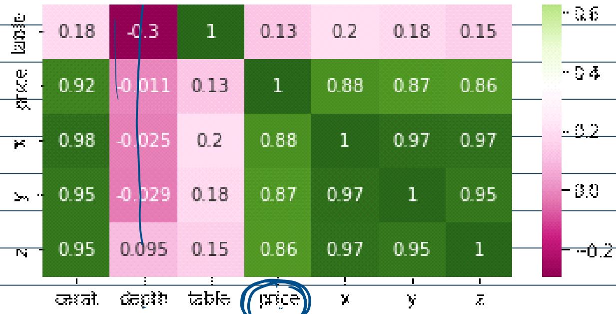
$VIF < 10$.

Corr



pre-def remaining call as independent

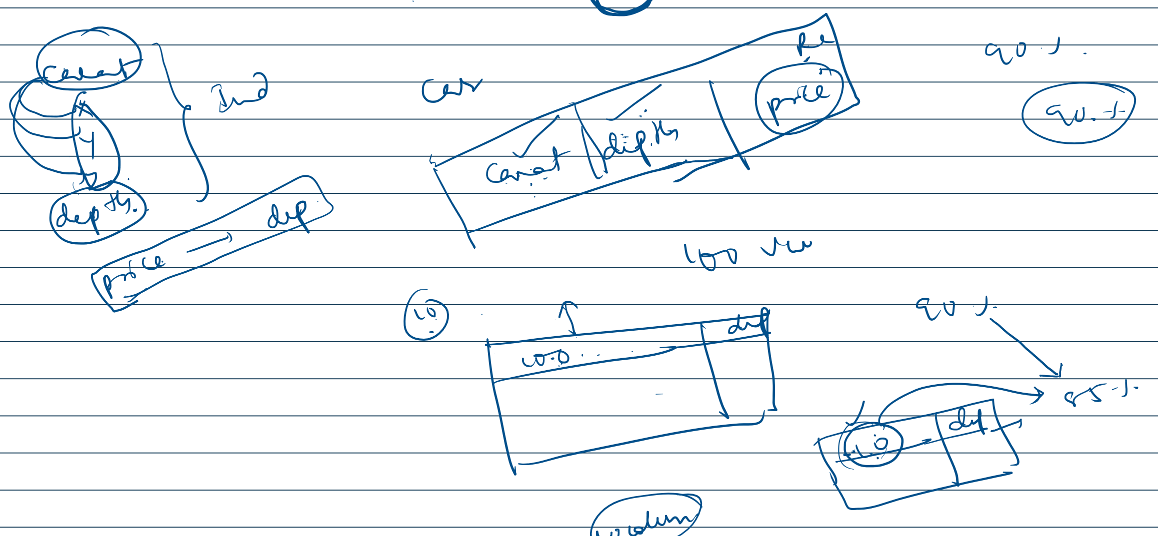
renaming ... are independent



Variance Inflation factor

vif < 10.

vif ≥ 10 → that column indicate multi collinear variable



10 column

{ patient demography details — (10) — ID, name, age, gender, employement stat
patient census details — popn, density/area, area, city, Language, salary, rural, urban, F/M
~~patient disease details~~ —

90 columns

10 column

100 columns	Coverage Y/N	Insurance / not
	Y	
	Y	
	N	
	N	
	Y	

50 > Insur

male → Insur

1. age ✓
2. gender
3. employement
4. population
5. area
6. city

→ Correlation
→ multi...

50 row > 10