Vyrasaathamanu, 2019AIML618, PEAM ZG511 Feature Engg EC-3R Q

Q.01 Dataset is drawn tabularly as shown.

Cust ID	Loan Type	Gender	Marital Status	Acc. Type	Yrs. in addr	Yrs in Job	Salary	Balance	Target
1 IDs 1 2 3 ... unique	Home / Home Improvement. Binary / Categorical with 2 levels	M/F Binary 2-level category	Family / Single / Others. Categorical 3-levels	Family / Others / Company / Own / Rent. 5-level Categorical	0, 1 ... n, numeric	0, 1 ... n numeric	∈ R any numbers numeric.	∈ R Real no. numeric	??

Summarizing: 1) Customer ID → primary key / unique = ID (ignored for modelling)

2) Loan Type = Loan (0/1) → 0 for Home, 1 for Improvement

3) Gender = 0/1, 1 = Male, 0 = female

4) Marital Status = Family, Single, Others = Status

5) Accommodation Type = Acc = Family / Own / ... etc.

6) Yrs. in Address = numeric = YIA

7) Yrs. in Job = YIJ = numeric

8) Salary numeric & 9) Balance numeric

a) Charts for Cat. / bin / num. attr.

Categorical → Salary vs. YIA or YIJ or YIA distributi...

Binary → Gender vs. Accommodation Type, or Gender histogram

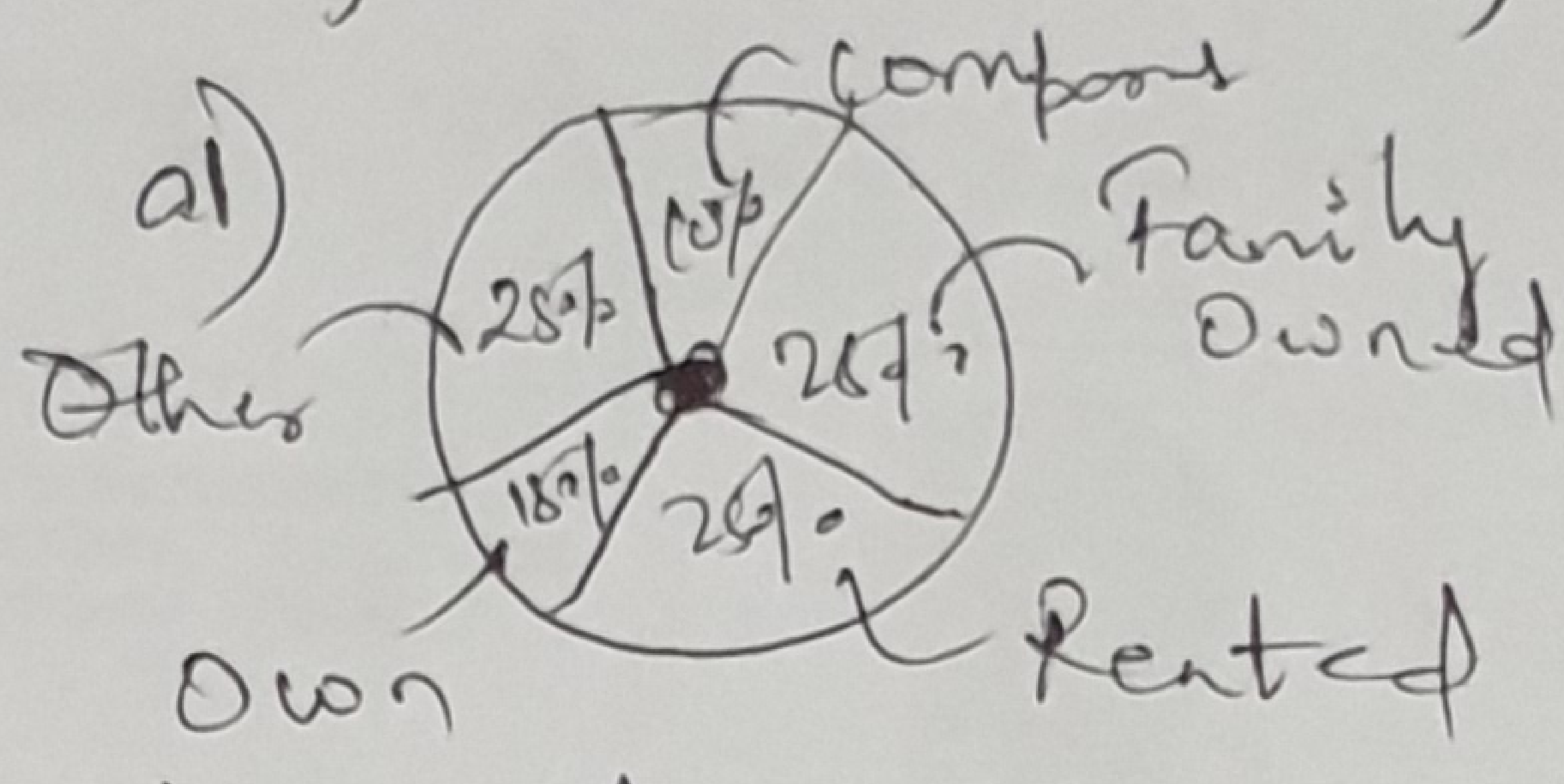
Numeric → Salary ranges or salary vs. balance or gender

swathanam, 2019AC1ML618, PCAM ZG11 Featuring. EC-3R C2

Q1 Contd.

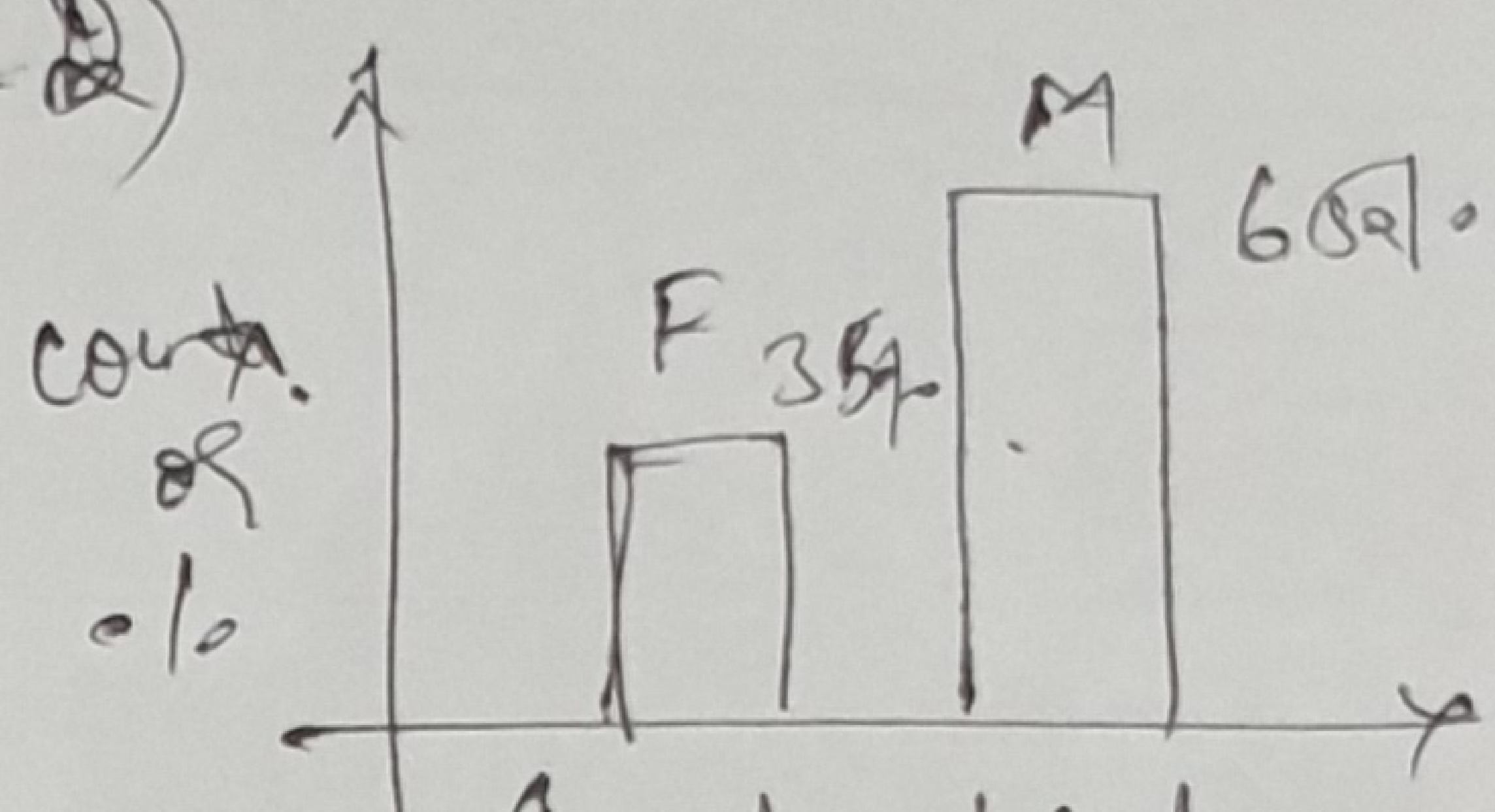
PIE

a1)



Accomodation type

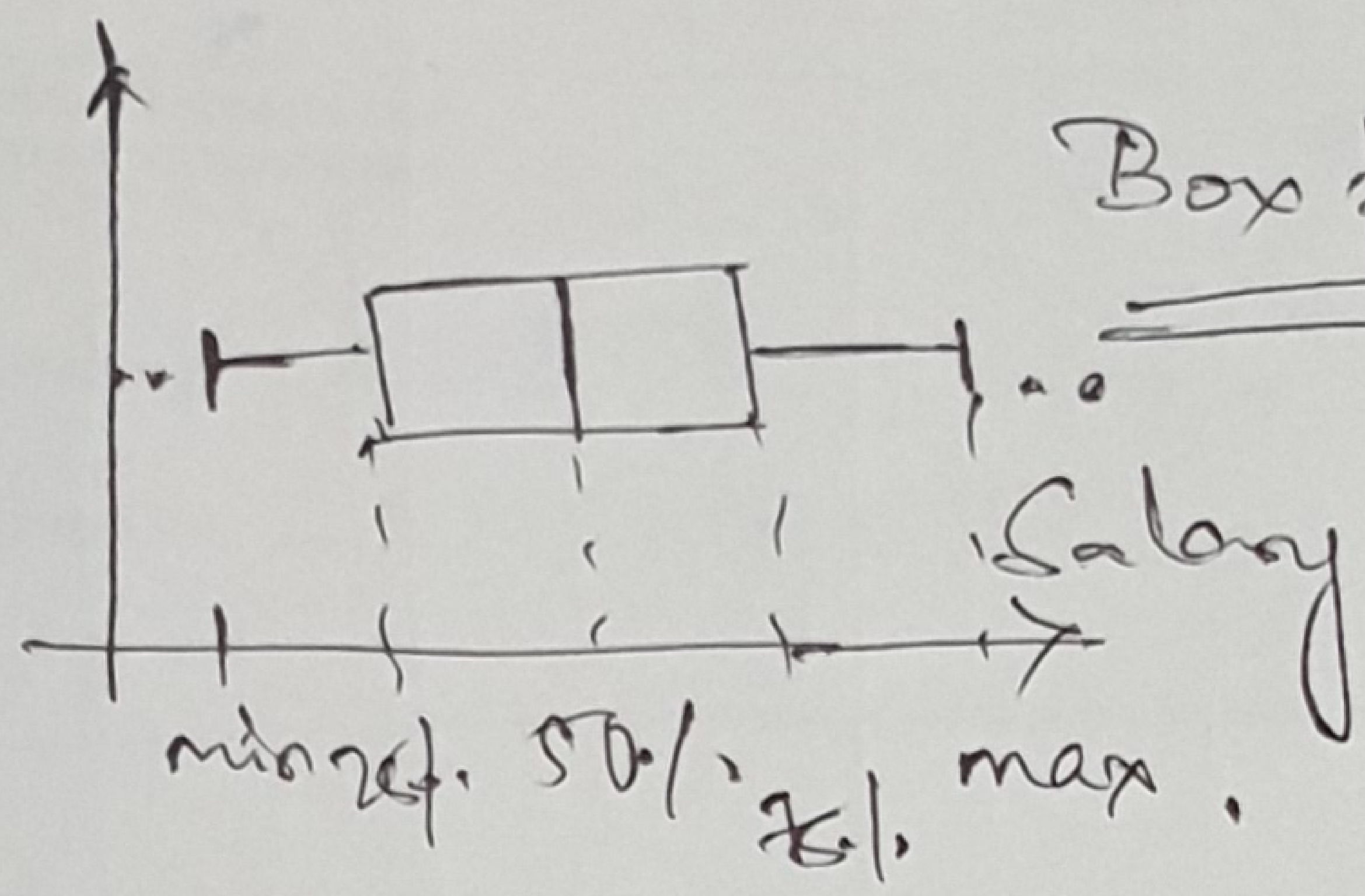
a2)



Gender histogram

a3)

Box plot



a1) Check % distribution of diff. accomodation types / any category

a2) Check freq. distribution of gender / any binary var.

a3) Distribution plot showing median, 25%, range & outliers (box plot) for salary / or any numeric attribute.

NOTE: There are many other ways to slice & dice data as added in the previous page.

B) Multiple Regression : [Loan Sanction Amount] (numeric)

Logistic Regression : [Loan Default] or [Loan Approval] (categorical / binary).

Decision Tree: This can predict real values as well as classify categories, hence all of the above target types listed can be used, viz.,

[Loan Sanction Amount or Loan Defaults or Loan Approval]

c) Multiple regression features : ~~ID → dropped~~.

Loan Type	Gender	YA	YIJ	Salary	Balance
0/1 Binary	0/1 Binary	numeric	numeric	numeric	numeric

Acc. & Status have to be encoded or binarized

Ex. Status = F/S/other.

Family	0	0	1
Single	0	1	0
Other	1	0	0

Satharam, 2019AIML618, PGAM-ACIII Feature Engg. EC-3R C2

Contd. c) So the total column will be.

After one hot encoded plus 3 column for status & 5 for acc.

list (columns) = LoanType, Gender, YIJ, YIA, Salary, Balance, Status_F, Status_O, Status_S, acc_f, acc_oth, acc_co, acc_o, acc_r

Target = Loan Amount Sanctioned

Status_f = 001 = family	acc_f = 00001 = family owned
Status_s = 010 = single	acc_oth = 00010 = other
Status_o = 100 = other	acc_co = 00100 = company provided
	acc_o = 01000 = own, acc_r = 10000 = rented

d) Numeric attr are YIA, YIJ, Salary & Balance.

Issues than can arise.

→ YIA, YIJ & Salary cannot be < 0 (invalid/incorrect data) → need correction

→ Salary can grow to a large number (need normalization)

→ Salary can be skewed (detect outliers)

→ Normalization can be with z-score $\left[\frac{x_i - \bar{x}}{\sigma_i}\right]$ or IQR

x_i = value, $\bar{x}$ = mean, σ_i = std. dev.

→ From box-plot, IQR can be values < 1.5 × 25%
or > 1.5 × 75%

I, Sudharsan, 2019AIML618, PGAM4CIII featuring. EC-3RC2

2. F-Score : I CANNOT SEE THE QUESTION.

PROCTOR HAS GIVEN THE BELOW TABLE WITH 5 ENTRIES

F1	F2	CLASS
1	5	C_1
2	6	C_1
3	7	C_2
4	1	C_2
5	2	C_2

$$F = \sum_{i=1}^{k} \phi_i (M_i - \mu)^2 \Big/ \sum_{i=1}^{k} p_i \sigma_i$$

Mean M_1 for class $C_1 = \frac{1+5+2+6}{4} = \frac{14}{4} = \boxed{7/2}$

Mean M_2 for class $C_2 = \frac{3+7+4+1+5+2}{6} = \frac{22}{6} = \boxed{11/3}$

Global mean $\mu = \frac{1+2+3+4+5+5+6+7+1+2}{10} = \frac{36}{10} = \boxed{18/5}$

C_1	1, 5, 2, 6
C_2	3, 7, 4, 1, 5, 2

, $\phi_1 = 4/10$, $\phi_2 = 6/10$

$\Rightarrow \phi_1 = 0.4$, $\phi_2 = 0.6$

Variance $\sigma_1 = \frac{1}{n} \Sigma (f_1 - \bar{f})^2 = \frac{(1-7/2)^2 + (5-7/2)^2 + (2-7/2)^2 + (6-7/2)^2}{4}$

$\Rightarrow \sigma_1 = \frac{1}{4} \cdot \left(25/4 + 9/4 + 9/4 + 25/4\right) = 68/16 = \boxed{17/4}$

$\sigma_2 = \frac{1}{n} \Sigma (f_2 - \bar{f})^2 = \frac{1}{6}\left[(3-11/3)^2 + (7-11/3)^2 + (4-11/3)^2 + (1-11/3)^2 + (5-11/3)^2 + (2-11/3)^2\right]$

$\Rightarrow \sigma_2 = \frac{1}{6}\left(4/9 + 100/9 + 1/9 + 64/9 + 16/9 + 25/9\right) = \frac{210}{9 \times 6} = \boxed{\frac{35}{9}}$

$$\boxed{F = \frac{\phi_1 (\mu_1 - \mu)^2 + \phi_2 (\mu_2 - \mu)^2}{\phi_1 \sigma_1^2 + \phi_2 \sigma_2^2}}$$

$$= \frac{\frac{4}{10} \times \left(\frac{7}{2} - \frac{18}{5}\right)^2 + \frac{6}{10}\left(\frac{11}{3} - \frac{18}{5}\right)^2}{\frac{4}{10} \times \frac{17}{4} + \frac{6}{10} \times \frac{35}{9}}$$

$$= \frac{\frac{4}{10} \times \frac{1}{100} + \frac{6}{10} \times \frac{1}{225}}{\frac{4}{10} \times \frac{17}{4} + \frac{6}{10} \times \frac{35}{9}} = \frac{\frac{1500}{100 \times 225}}{\frac{68 \times 3 + 70 \times 4}{3 \times 4}} = \frac{\frac{1}{15}}{\frac{484}{12}} = \frac{1}{15} \times \frac{12}{484} = \frac{1}{605} = \boxed{0.001653}$$

Vardhamana 2019AIML618 PGAMZC111 Feat. Eng. EC-R3C2

Q.03. Total deficit = 1500 (All amt. in crores)

Prop. tax deficit = 600 (targ. = 1800)

Other sources { Sky sign deficit = 66 (target = 90)
building permission — [Dev. charges / perm. fees] deficit = 400 (target = 800)

Refund from govt's state & union = 70% of total revenue.

i) Prominent reasons for revenue deficit — PMC is dependent on grants by state & union govts. which
i) have given only 70% of revenue & not ~~yielding~~ results on follow-up yielding

ii) LBT (Local body tax has been abolished)

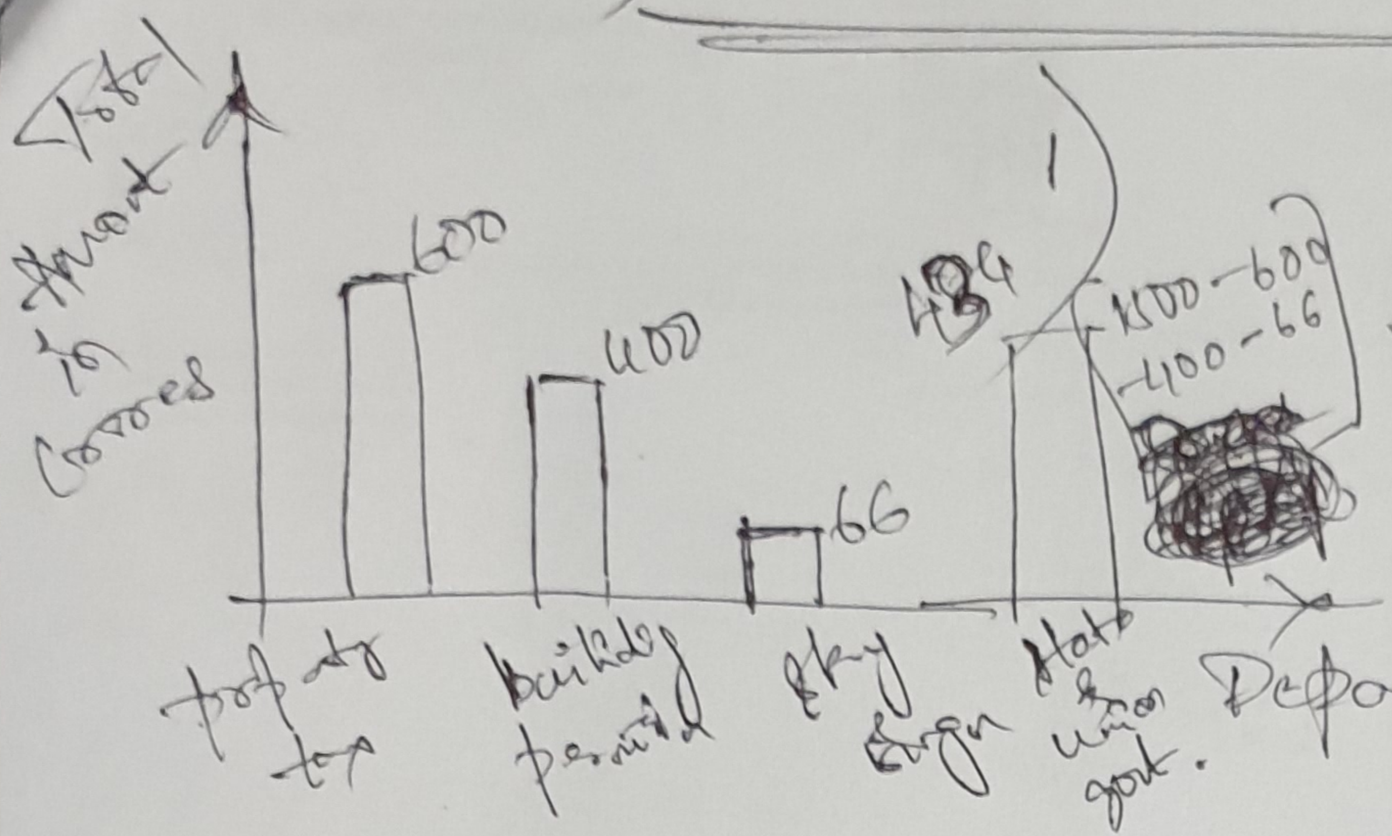
iii) Individual departments have deficits and are not generating any profits (like property tax, other sources including sky sign, etc.)

---

ii) Story points are already described above. Plots in next page for data visualization.

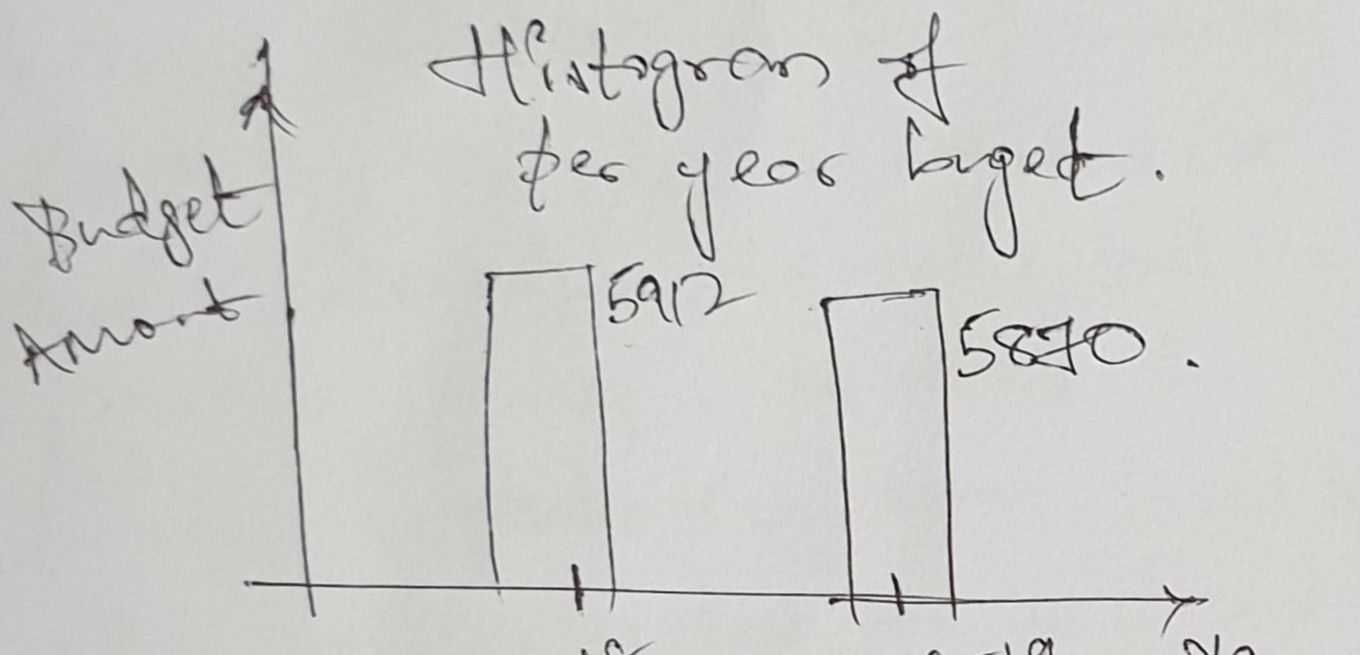
Swathanai 2019AIML618 PCAM7C111 Feature Eng EC-RB C2

203. Contd. ii) Plots for Visualization



Deficit per department

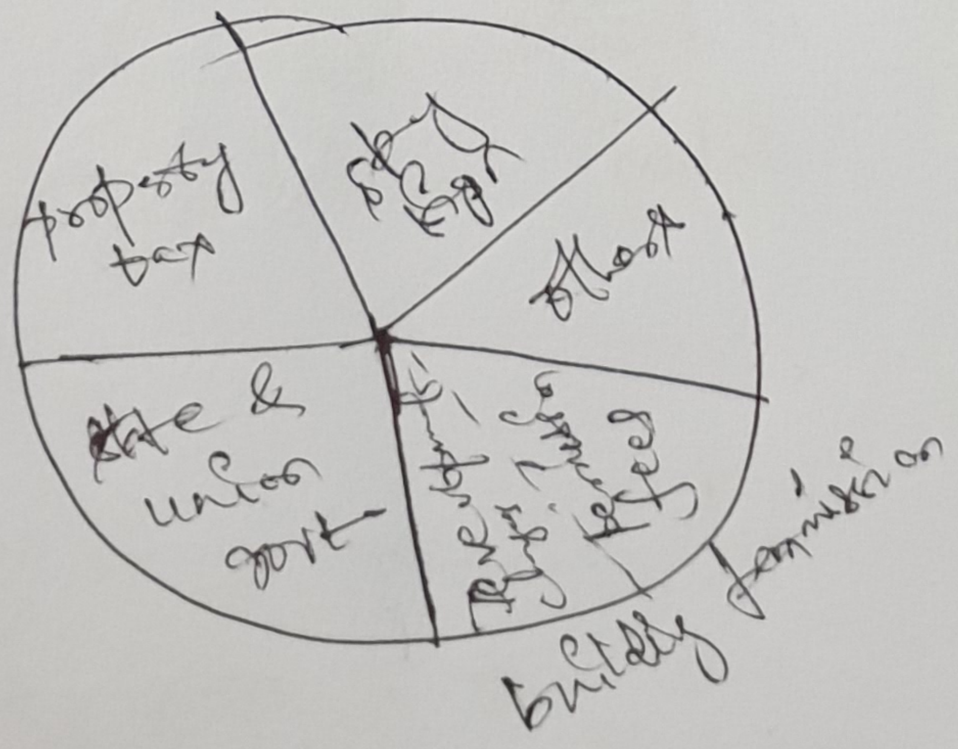
Bar chart for distribution of per department deficit.



2) Histogram to show yearly budget allocation / distribution

3) PIE for department distribution o/o.

[Revenue generating departments]



4. A)

Hypothesis Testing	v/s. Exploratory Data Analysis (EDA)
i). Hypothesis testing is to test for differences in samples via some statistic (mean, variance, etc.) or between models & model parameters.	i) EDA is analyzing & visualizing data to be done to identify patterns / usages for the data to build a suitable model
ii) Hypothesis testing is usually done with a null Hyp. Ho which states "there is no / 0 significance" (in observed values). This is typically tested through a p-value which shows the probability of finding the value as extreme (at least as extreme) as observed just by random chance (if Ho is true). If the p-value is ~~greater~~ than some threshold then Ho is rejected and a alternate Hyp. Ha which typically is "there is some significance" (in observed values) is selected	ii) EDA is done through plots / charts & graphs to get more insights into the data & the process that generates the data (underlying)
iii) Hyp. is post-hoc. which is after finding relevant models / stats / parameters	iii) EDA is prior to finding stats / models / params & involves finding statistics & estimating parameters
iv.) Ex. In linear regression, (univariate) after finding slope of the line, $H_0 \Rightarrow slope = 0$, $H_a \Rightarrow slope \neq 0$; we test it using t-stat / distribution $t = \frac{obs.value - 0}{std.error(obs.value)}$; obs. value = observed value	iv) Ex. Finding data summary (mean, variance, distribution), plotting them, finding null / missing values

Hyp.

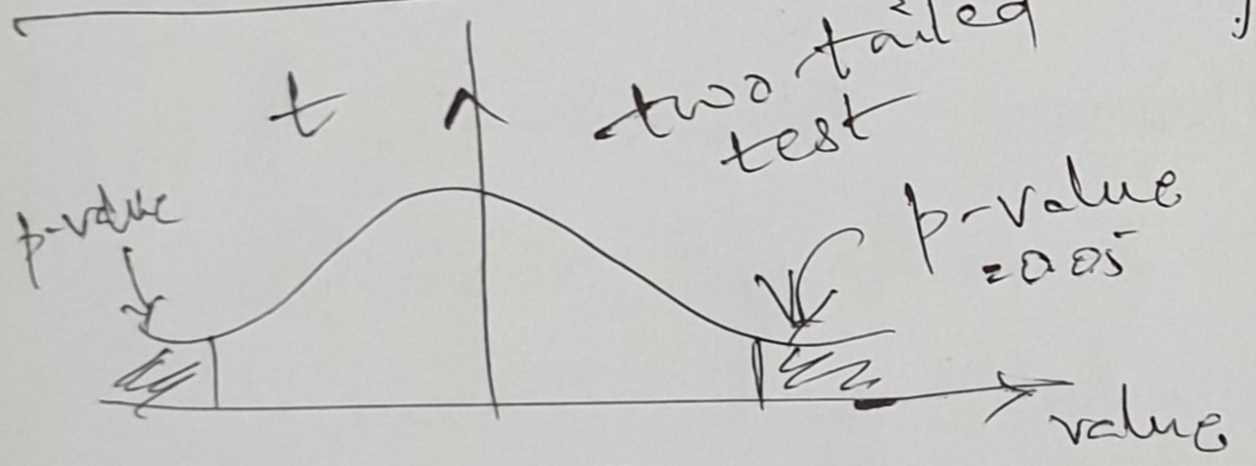
iv) t-stat says, how far the slope is from '0' & the probability of finding such a value by chance if Ho is true

Note: For multiple reg. this will be F-stat.

Other tests are between samples for $\mu = 0$ & $\sigma^2 = 0$

where Ho = samples are from same population.

Ha = samples are not from same population



Ex: of a two-tailed test for t-distribution

p-values are typically 0.05 or 0.01

EDA

iv) EDA gives params in tables, also in graphs and charts which can be univariate/multi-variate

Eg. box, histogram (uni) scatter, contour (multi) bi

Q.04. Contd. B) i) Variable with 90% missing values.

→ this variable will be dropped as there is only 10% information. We cannot assume the distribution from which the data was generated or the probability of it or the likelihood of data.

→ as population is not known, we cannot estimate any statistic (mean, dispersion, etc.) of the population parameters & hence will discard it.

ii) Variables strongly correlated carry the same information & hence will be dropped as well

→ This will pose difficulty in optimization.

→ For probabilistic generative methods, this will over-estimate the prior & probabilities affecting the model negatively.

C) NMAR data can be imputed by these 2 ways -

i) Add values at the end of distribution → (mode in case of categorical column).

ii) Any random value (which is not the mean/median)

iii) Also add a missingness indicator column & check model performance.

~~Male~~

In given data set →

i) ⇒ Male = 57
Female = 18

18, 57 tail/end of distribution.

ii) ⇒ Avg. of highest & least 2 values, etc.

iii) ⇒ Additional column.

~~Gend~~	~~Age~~	~~Age_missing~~