Q1 paper

ETL

ETL Life Cycle real life

1. Cycle initiation
2. Build reference data
3. Extract (from sources)
4. Validate
5. Transform (Clean, Apply business rules, check for integrity, create aggregates or disaggregates)
6. Stage (Load into staging tables, if used)
7. Audit reports (for example, on compliance with bussiness rules, also in case of failure, helps to diagnose/repair)
8. publish (to target tables)
9. Archive
10. clean up

✓

① Cycle Initiation:

i) Start the ETL process, often triggered by a schedule, an event, or specific business need.

ii) ex: Initiation the ETL process every night to update a data warehouse with the latest sales data.

② Build Reference Data:

i) Create or update reference data need for transformation and validation process.

ii) ex: Building a reference table or product categories that can be used to categories and enrich sales data

③ Extract (from Sources)

i) Gather data from various sources, such as databases, files, or APIs

ii) ex

Extracting Customer information from an online store's database and order data from a legacy system

④ Validate

i) Check the extracted data for accuracy completness and adherence to predefine rules

ii) ex

Verifying that all order records have valid Customer ID and that the total order amount is within an acceptable range.

P.T.O

(5) Transform:

i) clean, apply business rules, and manipulate the data to fit the target formate.

ii) ex
converting date formats, aggregating sales data by region and calculating average order value.

(6) Stage (Load into Staging)

i) Load the transformed data into staging tables for further processing.

ii) ex
Storing cleaned and transformed customer and order data in temporary tables before loading them into the main database.

(7) Audit Reports:

i) Generate reports to track compliance with rules & diagnose issues in case of failures.

ii) ex
Creating a report that highlights any data anomalies, such as missing values or discrepancies b/w source and target data.

(8) Publish (To target tables):

i) move the validated and Transform data to the final destination or target tables.

ii) ex Loading the cleaned and validated sales date into the main data warehouse for business analytics.

(9) Archive.

i) Store the historical or older versions of data for future reference or regulatory compliance.

ii) ex Archiving quarterly financial data to maintain a historical record for auditing purposes.

(10) Clean Up:

i) Remove temporary files, tables or any other artifacts created during the ETL process

ii) ex

Deleting staging tables and Temporary files used during the ETL cycle to free up storage space

* Q2 paper

Data Source Selection Creteria:

1. Revelence to project Objectives
   * meaning ful insights
2. Data quality & Accuracy
   * accurate info
3. Timeliness of Data
   * upto date information
4. Data Volume and Scope
   * Suppient data
5. Data Compatibility and formate
   * integration
6. legal and Ethical Compliance
   * copyrights
7. Cost and Resource constraint
   * financially

OR

Q 2

Data Source selection Criteria

1. credible (reputable & trustworthy)
2. complete (all necessary information)
3. verifiable (accuracy & authenticity)
4. Accurate (free from errors)
5. current (up to date)
6. compliance (legal & regulatory)
7. accessible (easily accessible)
8. cost (reasonable cost)
9. legal (~~copy write~~ copyright, intellectually properties)
10. security (protection)
11. storage (efficient stored)
12. provenance
    - ↳ Understanding the origin & history of the data, including how it was collected & processed.
    - ↳ Documenting the source and processing steps for research data.

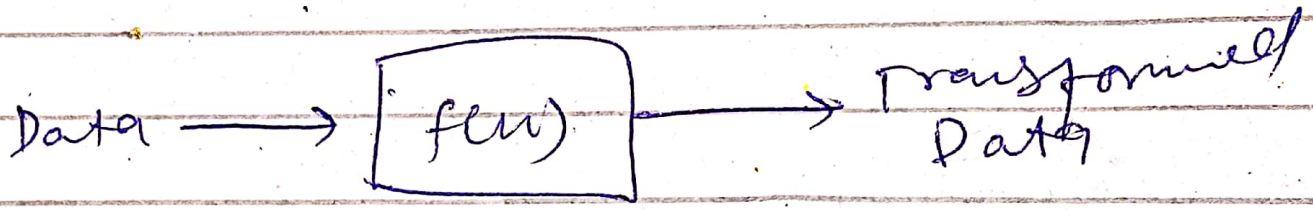
Q3 paper

Data Transformations

① ↳ Apply a function $f(x)$ to adjust scales of data

② Done usually when data is skewed, so that it becomes easier to perform modelling

③ Done to convert non linear relationship into ~~linear~~ linear relationship.



(Natural) Log Transformation

① To Transform data that is positively Skewed.

② Usually done when data is concentrated to zero
(most values are close to zero but few are much larger)

Log Transformation

⇒ Makes the relationship b/w two variables more Linear.

↳ many simple modeling methods assume a Linear relationship b/w variable

⇒ Linear regression

Other Transformation

① Square ROOT
② Square
③ Inverse

Q4

↳ Accuracy is not good evaluation measure when dealing with imbalanced datasets

↳ Imbalanced datasets occur when one class significantly outnumbers the other.

Example

Total transaction = 10000

non fraudulent = 99000

Fraudulent = 1000 (1% of total.

* Model prediction all transaction as non fraudulent for maximize accuracy

⇒ TP => 0 (model incorrectly predicts all fraudulent transaction as non fraudulent

⇒ True N: 99000 (model correctly predicts no fraudulent tran

⇒ FP: 0 (model incorrectly predict non f as f tran

⇒ False N = 1000 (model fails to predict f tran. tran)

~~Accuracy => TP/(TP+FP) = 0/0+0 = ∞~~

Accuracy = (TP+TN) / Total

= 0 + 99000 / 100000 = 99%

* Q5 paper

Explain the role of data Scientist

DS

① Explorer: Explore and analyse large sets of data to uncove meaningful patterns, trends and insights

② problem Solver: They use data to solve Complex probems make informed decision, helping business and orga...

③ predictor: Build models to predict future outcomes

④ Story teller: Turning data into actionable insights.

⑤ continous learner: Adapting new technologies, method data Scientist stay updated.

* Paper

Q6:

a) df["Color"] = df["Department"].map({"IT": "red", "HR": "green", "Finance": "blue", "Marketing": "orange"})

b) plt.hist(df["Department"], bins = len(df["Department"].unique()))

c) sns.lineplot(x="Age", y="Experience", hue="Department")

d) plt.hist(df[df["Department"]=="IT"]["Age"], df["Department"]=="HR"]["Age"]], color=["red", "green"], label=["IT", "HR"])

e) df.df.drop("Salary", axis=1)

Q7

(a) data["Title"].value_counts()

output

Title	
SE	1
PM	1
HR M	1
Q A	1
A	1
M A	1
HR Inter	1

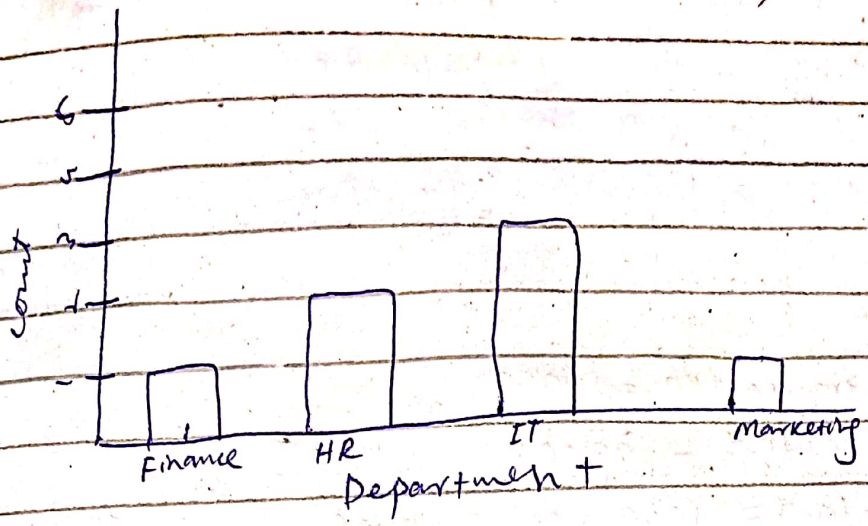
Name: count, dtype: int64

(b) data["Department"].describe()

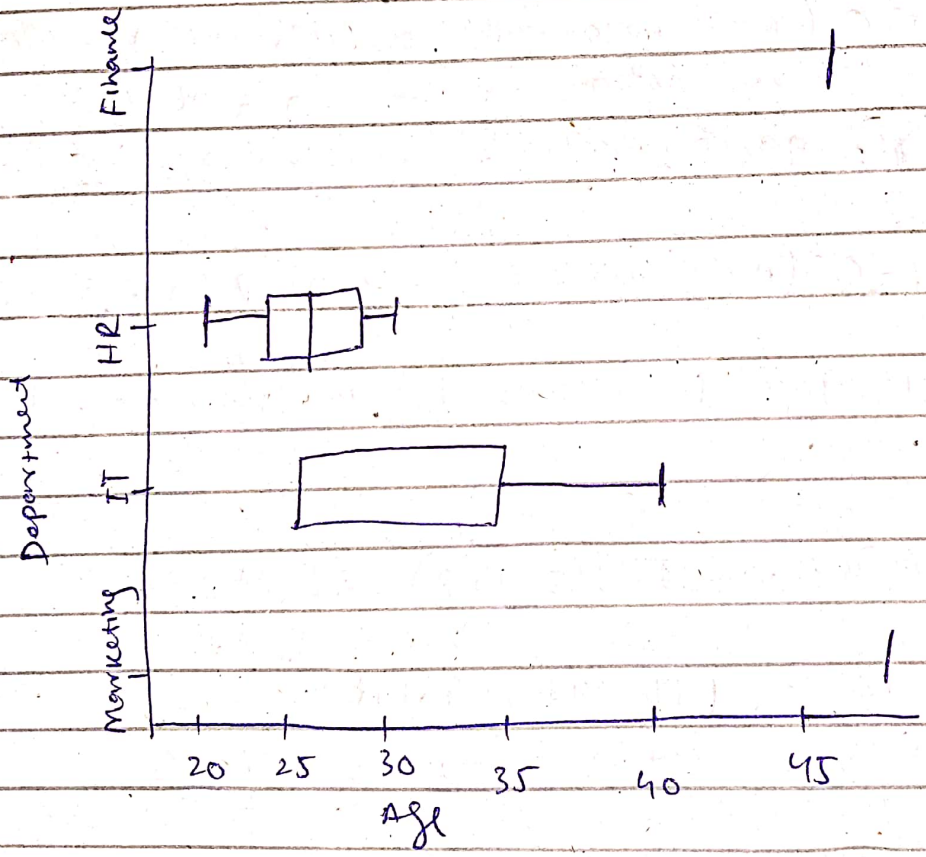
output

count	7
unique	4
top	IT
freq	3

© sns.countplot(data["Department"])



(d) sns.boxplot(data["Age"], data["Department"])



Apriori algorithm:

Q8 paper

Generate Rules using Apriori Algorithm

Support = 50% confidence = 75%

Transaction ID	Items purchased
1	Bread, Cheese, Egg, juice
2	Bread, Cheese, juice
3	Bread, Milk, yogurt
4	Bread, juice, Milk
5	Cheese, juice, Milk

① Find Frequent Item set and their support

Item	Freq	Support in (%)
Bread	4	4/5 = 80%
Cheese	3	3/5 = 60%
Egg	1	1/5 = 20% ✗
juice	4	4/5 = 80%
Milk	3	3/5 = 60%
yogurt	1	1/5 = 20% ✗

Support(item) = Freq of item / no of transactions

(2) Remove all the items whose support is below given min support

New Table

Item	Freq	Support
Bread	4	4/5 = 80%
Cheese	3	3/5 = 60%
Juice	4	4/5 = 80%
Milk	3	3/5 = 60%

(3) Now form two items candidate set & write their freq

Items pair	Freq	Support (in %)
Bread, cheese	2	2/5 = 40%
Bread, juice	3	3/5 = 60%
Bread, milk	2	2/5 = 40%
cheese, juice	3	3/5 = 60%
cheese, milk	1	1/5 = 20%
juice, milk	2	2/5 = 40%

④ Remove all items whose support is below the given min support

item pairs	freq	support %
Bread, juice	3	3/5 = 60%
Cheese, juice	3	3/5 = 60%

⑤ Generate rules

For Rules Consider

ⓐ (Bread, juice)

Bread → juice & juice → Bread

ⓑ (Cheese, juice)

Cheese → juice and juice → cheese

Confidence $(A \to B)$ = support $(A \cup B)$ / support (A)

① confidence (Bread → juice) = $\frac{\text{support(Bread } \cup \text{ juice)}}{\text{support (Bread)}}$

$\Rightarrow 3/5 * 5/4 = 3/4 \Rightarrow 75\%$

② confidence (juice ∪ Bread) / support (juice)

$\Rightarrow 3/5 * 5/4 = 3/4 \Rightarrow 75\%$

③ Confidence (cheese → juice) = $\frac{\text{support(cheese ∪ juice)}}{\text{support(cheese)}}$

⇒ 3/5 * 5/3 = 1 = 100%

④ confidence (juice → cheese) = $\frac{\text{support(juice ∪ cheese)}}{\text{support(juice)}}$

⇒ 3/5 * 5/4 = 3/4 = 75%

Q9

(a)

Cons

(1) computational complexity: if no of itemset grows exponentially.

(2) Memory Usage: Requires multiple procers through data. leads to high memory usage.

(3) Fixed threshold
- Setting min support threshold.
- choice can impact the results.

Pros

(1) simple concept: easy

(2) scalability: Efficient

(3) widely used:

Defination: classic algo in data mining. for association rule learning
- used to discovered frequent items sets in a datasets
- generate association rules based on these itemsets.

(d) Q_0 = min value = 20

Q_4 = max value = 47

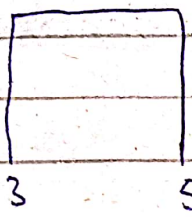
(c)

Q 10 paper — Required HAC (complete), Dendogram

Given

	P1	P2	P3	P4	P5
P1	0				
P2	9	0			
P3	3	7	0		
P4	6	5	9	0	
P5	11	10	2	8	0

①



② The distance matrix will update

	P1	P2	P3,P5	P4
P1	0			
P2	9	0		
P3,P5	11	10	0	
P4	6	5	9	0

① = Max [dis (P3,P5), P1]

=) Max [dis (P3,P1), (P5,P1)]

=) max [3, 11] =) 11

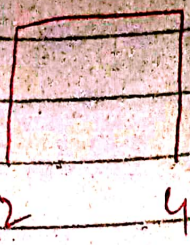
② max [dis (P3,P5), P2] => Max [dis (P3,P2), (P5,P2)]

=) Max (7, 10) => 10

③ =max [dis (P3,P5), P4] =) Max [dis (P3,P4), (P5,P4)]

=) max [9, 8) = 9

P T O



(3) the distance matrix will update

	P1	P2, P4	P3, P5
P1	0		
P2, P4	9	0	
P3, P5	11	10	0

① ⇒ Max [dis [(P2, P4), P1]]

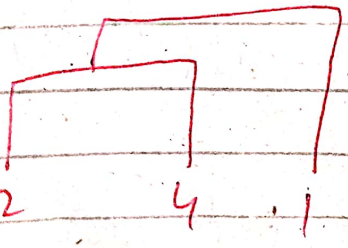
⇒ Max [dis (P2, P1), (P4, P1)]

⇒ Max [9, 6] = 9

② ⇒ Max [dis [(P2, P4) (P3, P5)]]

⇒ Max [dis [(P2, (P3, P5)), (P4, (P3, P5))]]

⇒ Max [10, 9] = 10



(4) The distance matrix will update :

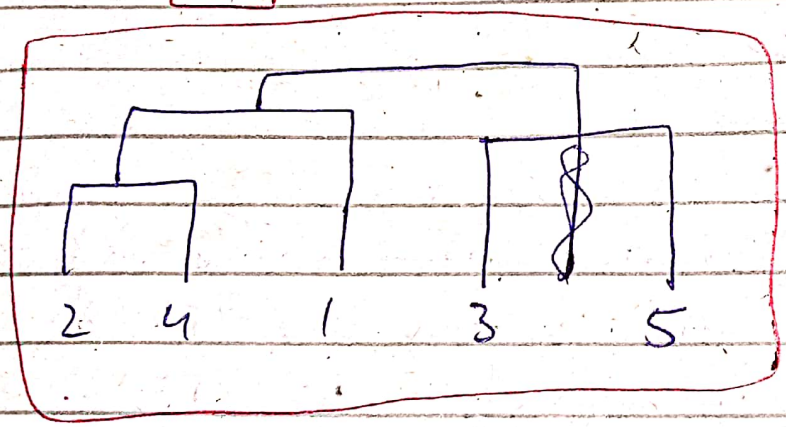
	P1 > (P2 > P4)	P3 > P5
P1 > (P2 > P4)	0	
P3 > P5	11	0

1) Max [dis (P3 > P5), [P1, (P2 > P4)]]

2) max [dis (P1, (P3, P5), (P2 > P4) (P3, P5)]]

3) max [11, 10]

2) 11



Q9 paper

Cons (b)

(1) Computational Complexity: Can be computationally expensive, with a large number of data points.

(2) Sensitivity to noise: Susceptible to noise & outliers, which can influence the merging process.

(3) Memory Usage: Requires storing a potentially large distance matrix, leading to high memory usage.

(4) Fixed Hierarchy: Once clusters are merged, the hierarchy is fixed, and changes require recomputing from scratch.