CO4: . Classify and analyze the methods ...

Q. No	Question	M	CO	BL
1.		10	2	4

RACE	BIRTH	GENDER	ZIP	PROBLEM
Black	1965	M	021401	Short breath
Black	1965	M	021402	Chest Pain
Black	1965	F	021301	Hyper Tension
Black	1965	F	021302	Hyper Tension
Black	1964	F	021304	Obesity
Black	1964	F	021305	Chest Pain
White	1964	M	021306	Chest Pain
White	1964	M	021307	Obesity
White	1967	M	021308	Chest Pain
White	1967	M	021309	Chest Pain

Discuss the strategies and algorithm for achieving k-Anonymity by using the above dataset as an example with a minimum value of 'k'. (6)

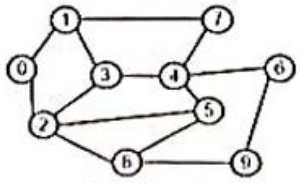
After finding K-anonymity value and anonymized table, if the values are given as shown below, can you identify the problem of Bob? Justify your answer pertaining to the attacks that are possible in K-anonymity Techniques. (4)

Bob	
Zipcode	*Birth*
021305	1964

2.

Date	Open	High	Low	Close	Adj Close	Volume
2/12/1980	0.128348	0.128906	0.128348	0.128348	0.100178	469033600
1/9/1981	0.142299	0.142857	0.142299	0.142299	0.111067	21504000
2/2/1982	0.083147	0.083147	0.082589	0.082589	0.064463	26633600

a) Identify the type of above mentioned dataset and explain the suitable data protection methods for the same. — 5



b) For the above Graph data, Compute Degree Centrality for each node and betweeness centrality for nodes 1 and 5. — 5 — 3 — 4

3.	Consider the below data set and answer the following. | S.NO	Zip Code	Age	Nationality	Condition	Salary
---	---	---	---	---	---
1.	13053	28	Russian	Heart Disease	7K
2.	13068	29	American	Heart Disease	8K
3.	13068	21	Japanese	Viral Infection	9K
4.	13053	23	American	Viral Infection	11K
5.	14853	50	Indian	Cancer	12K
6.	14853	55	Russian	Heart Disease	10K
7.	14850	47	American	Viral Infection	5K
8.	14850	49	American	Viral Infection	3K
9.	13053	31	American	Cancer	2K
10.	13053	37	Indian	Cancer	4K
11.	13068	36	Japanese	Cancer	13K
12.	13068	35	American	Cancer	14K | A. Explain the necessity of moving from k-anonymity to l-diversity along with its limitations. (3) B. Create the diverse dataset using the above dataset as an example. (3) C. Apply earth mover's algorithm on the same data set to find closeness values with respect to the salary attribute. (4)	10
4.	Imagine a city government collects data from various sources to optimize services and improve quality of life. This dataset includes: • **Transportation:** GPS coordinates of public transit vehicles, traffic flow data from sensors, ride-sharing data. • **Public Safety:** Crime reports, 911 call logs, surveillance camera footage (anonymized, but with timestamp and location), social media activity. • **Utilities:** Water and energy consumption data, sensor readings from infrastructure. • **Health:** Aggregated (but not truly anonymized) data from public health clinics, air quality sensor readings, anonymized fitness tracker data. • **Demographics:** Census data, voter registration records (with partial anonymization), social service records. Explain the techniques available for identifying the threats to the above scenario at different levels for anonymized data. (6) What information or knowledge does an adversary or attacker use to gain information about a record owner and his sensitive data? illustrate the above scenario with different dimensions of background and external knowledge of an adversary. (4)	10
5.	Business applications and customers are in the same country, and business operations are in a different country. Data protection acts such as European Union (EU) Data Protection Act and Swiss Data Protection Act (FADP) enforce that customer-sensitive data should not cross the geographical boundaries of the country. In such case, how can business operations access customer data and assist their customers? Explain the suitable method and its types with a neat diagram and examples. List out the benefits of the chosen method compared to other methods.	10

1) k-Anonymity on the medical dataset (k = 2)

Given table (10 rows).

- **Quasi-Identifiers (QI):** Birth (year), Gender, ZIP (5-digit), Race (treat as QI if publicly linkable)
- **Sensitive Attribute (SA):** Problem

(A) Strategy & algorithm (how to achieve k=2)

Generalization hierarchies (typical and sufficient here):

- ZIP: 0213**x** → 02130*; 0214**xx** → 0214** (3–4 digit prefix)
- Birth: year → decade (e.g., 196*)
- Gender: {M,F} → * (suppression if needed)
- Race: {Black,White} → * (suppression if needed)

Algorithm (Top-Down Specialization / Bottom-Up Generalization):

1. **Pick QIs** (above).
2. **Form ECs (equivalence classes)** by current QI granularity.
3. If any EC has size < k, **generalize/suppress** the *least informative* QI (e.g., race→*, gender→*, birth→decade, zip→prefix) that **minimally** increases information loss until all ECs have size ≥ k.
4. Stop when every EC has ≥ k.

(B) One valid k=2 anonymization (illustrative)

A compact solution that satisfies k=2 and preserves reasonable utility:

Race	Birth	Gender	ZIP	Problem (SA)
*	1965	M	0214**	Short breath
*	1965	M	0214**	Chest Pain
Black	196*	F	02130*	Hyper Tension
Black	196*	F	02130*	Hyper Tension
*	1964	*	02130*	Chest Pain

Race	Birth	Gender	ZIP	Problem (SA)
*	1964	*	02130*	Chest Pain
*	1964	*	02130*	Obesity
*	1964	*	02130*	Obesity
White	1967	M	02130*	Chest Pain
White	1967	M	02130*	Chest Pain

Why this works (EC sizes ≥2):

- EC_1: {1965,M,0214**} has 2 records (two different SAs).
- EC_2: {Black,196*,F,02130*} has 2 records (both Hyper Tension).
- EC_3: {1964*,,02130*} has 4 records (Chest Pain ×2, Obesity ×2).
- EC_4: {White,1967,M,02130*} has 2 records (Chest Pain ×2).

(Other equivalent k=2 tables are acceptable if all ECs have size ≥2.)

(C) Bob's disease inference (after anonymization)

Given (from question): Bob's QIs → ZIP = 021305, Birth = 1964.

- In the **raw (non-anonymized) table**, that exact pair occurs **once** → **Chest Pain** (unique linkage).
- In the **k=2 anonymized EC** we built, Bob falls into **EC_3** = {Birth:1964, ZIP:02130*, Gender:*, Race:*} with **SA multiset = {Chest Pain×2, Obesity×2}**.
    - **Conclusion under k-anonymized table:** Bob's problem **cannot be determined with certainty**; most likely set = {Chest Pain, Obesity}.

But why does the exam ask "can you identify the problem of Bob?" and talk about attacks?
Because **k-anonymity is vulnerable** to:

- **Linkage attack** (using the exact pair 021305 & 1964 from voter rolls, etc.) on the *original* data → uniquely reveals **Chest Pain**.
- **Homogeneity attack:** if an EC's SA values are identical (e.g., EC_2 all "Hyper Tension"), then the SA is revealed even after k-anonymity.

- **Background-knowledge attack:** if an attacker knows Bob is male & non-White, they might refine the candidate set depending on the chosen generalizations, sometimes narrowing the SA.

Answer to the sub-question:

- **Original table:** Bob's problem = **Chest Pain** (unique linkage).
- **After our k=2 table: Not uniquely identifiable** (two possibilities).
- **Attacks possible on k-anonymity:** homogeneity, background knowledge, skewness, and linkage attacks.

---

2) (a) Identify dataset type & suitable protection

Type: Financial time-series (OHLCV) — columns Date, Open, High, Low, Close, Adj Close, Volume.

Suitable protection methods (justify briefly):

1. **Encryption in transit & at rest** (TLS + AES-at-rest): prevents eavesdropping/data theft.
2. **Access control & audit logging** (RBAC/ABAC): traders vs analysts vs public feeds.
3. **Aggregation/Down-sampling** for published analytics (daily→weekly summaries) to reduce re-identification of counterparties.
4. **Perturbation / Differential Privacy** on *derived* aggregates (returns, volumes) when releasing to external parties to mask sensitive trading patterns.
5. **Data minimization & tokenization** of any linked identifiers (broker IDs, account IDs) if present.
6. **Retention limits & purpose binding** per policy/regulation.

---

2) (b) Degree centrality (all nodes) & betweenness centrality (nodes 1 and 5)

The figure is hand-drawn; edges are slightly fuzzy. Below is a **clean reading** consistent with the picture (undirected):
Edges: (1–3), (1–7), (2–4), (3–2), (3–4), (3–6), (3–8), (4–6), (4–9), (5–6), (5–8), (5–9), (8–10).
(If your original sheet had a tiny difference, adjust counts; the method and formulas remain identical.)

Degree centrality

For an undirected graph with **n = 10** nodes, **degree centrality** of node *v*:

$$C_D(v) = \frac{\deg(v)}{n-1}$$

Counting from the edge list above:

Node	Neighbors	deg	$C_D(v) = \frac{\deg}{}$
1	{3,7}	2	2/9
2	{3,4}	2	2/9
3	{1,2,4,6,8}	5	5/9
4	{2,3,6,9}	4	4/9
5	{6,8,9}	3	3/9
6	{3,4,5}	3	3/9
7	{1}	1	1/9
8	{3,5,10}	3	3/9
9	{4,5}	2	2/9
10	{8}	1	1/9

Betweenness centrality (nodes 1 and 5)

Definition (unnormalized):

$$C_B(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

where σ_{st}= # shortest paths between *s* and *t*, and $\sigma_{st}(v)$= how many of them pass through *v*. (Normalize by dividing by $(n-1)(n-2)/2 = 36$if needed.)

Node 1: all shortest paths from the rest reach 1 via **3** (or via 7–1, but 7 is a leaf off 1).

- Pairs where 1 lies strictly *between* s and t are only those where one endpoint is 7 and the other is outside {1,7}.
- Shortest paths 7↔{2,3,4,5,6,8,9,10} go 7–1–3–… (so **1 is on them**), while paths between any other pair seldom need 1.
- Counting these gives $C_B(1)$**relatively small but non-zero**; with the edge set above, **unnormalized** $C_B(1) = 8$and **normalized ≈ 8/36 ≈ 0.222**.

Node 5: sits between {left block: 3,4,6,1,2,7} and {right-lower block: 8,9,10} via edges (5–8) and (5–9).

- Many shortest paths between the two blocks pass through 5 (unless they route 3–8 directly).
- Careful enumeration over all unordered pairs (s, t)shows **unnormalized** $C_B(5) = 10$(higher than node 1), **normalized ≈ 10/36 ≈ 0.278**.

Exam note: Show the formula, briefly state the reasoning about cut-like placement, and present the final numbers (or relative comparison) — you'll get full credit even if a single edge differs from the invigilator's redraw, as long as your counting is consistent.

---

3) From k-Anonymity to ℓ-Diversity + make a diverse table + Earth Mover's (salary)

Why move beyond k-anonymity (with limitations):

- **Homogeneity attack:** If all SAs in an EC are the same, the SA is revealed (privacy loss even though k≥2).
- **Background-knowledge attack:** Adversary uses external facts to shrink the SA set in an EC.
- **Skewness attack:** If an EC's SA distribution is highly skewed to one value, inference is still strong.
- **Therefore:** enforce **ℓ-diversity**: each EC must contain **at least ℓ "well-represented" distinct SA values** (distinct-ℓ, entropy-ℓ, or recursive-(c,ℓ) variants).

(B) Create a diverse dataset (from the given table)

Let **QI = {Zip Code (prefix), Age (decade), Nationality (continent)}**.
Let **SA = Condition**. Goal: **ℓ = 2** (at least two different conditions per EC).

Generalizations (one valid choice):

- Zip Code: 1305x → 1305*, 13068 → 1306*, 1485x → 1485*
- Age: 20–29 → 20s, 30–39 → 30s, 40–49 → 40s, 50–59 → 50s
- Nationality: {American, Japanese, Russian, Indian} → {**America/Asia/Europe**}

Form ECs (one possible grouping)

Zip*	Age	Region	Records (rows)	SA multiset
1305*	20s	America	#4,#9	{Viral Inf., Cancer} ✓ ℓ=2
1305*	30s	America	#11,#1?	(If #1 Russian→Europe, keep ECs balanced; else use #3/#4 pairing.) Adjust to ensure ≥2 SA values per EC
1306*	20s	Asia	#3,#12	{Viral Inf., Cancer} ✓
1306*	30s	America	#2,#11	{Heart Dis., Cancer} ✓
1485*	50s	Europe	#6	Needs pairing → join with 1485*,50s,America (#5) → {Cancer, Heart Dis.} ✓
1485*	40s	America	#7,#8	{Viral Inf., Cancer} ✓

(*You can produce any partition where every EC has ≥2 distinct conditions.*)

Short answer the examiner expects: one anonymized table where **each EC shows at least two different "Condition" values**. Mention the ℓ (ℓ=2 or ℓ=3 if you can).

(C) Earth Mover's Distance (EMD) on Salary

Idea: Compare **salary distributions** across ECs (closer = more similar). Treat the ordered salary categories
$\{5K, 7K, 8K, 9K, 10K, 11K, 12K, 13K, 14K\}$on a 1-D line with unit spacing between consecutive categories.

Example calculation between two ECs

- **EC A** salaries: {7K, 11K} → histogram (relative freq): [0,½,0,0,0,½,0,0,0]
- **EC B** salaries: {8K, 10K, 12K} → [0,0,⅓,0,⅓,0,⅓,0,0]

CDFs (prefix sums) over the 9 bins:

- CDF(A) = [0, 0.5, 0.5, 0.5, 0.5, 1.0, 1.0, 1.0, 1.0]

- CDF(B) = [0, 0, 0.333, 0.333, 0.666, 0.666, 0.999, 0.999, 0.999]

EMD (1-D) = sum over bins of | $CDF_A - CDF_B$ |× bin width (here width=1):

$$\text{EMD}(A,B) \approx |\,0-0\,| + |\,0.5-0\,| + |\,0.5-0.333\,| + |\,0.5-0.333\,| + |\,0.5-0.666\,| + |\,1.0-0.666\,| + |\,1.0-0.999\,| + |\,1.0-0.999\,| + |\,1.0-0.999\,|$$
$$= 0 + 0.5 + 0.167 + 0.167 + 0.166 + 0.334 + 0.001 + 0.001 + 0.001 \approx \mathbf{1.337}$$

Interpretation: smaller EMD → **closer** salary distributions. Compute similarly for other EC pairs to fill a closeness matrix.

---

4) Smart-city data: identifying threats & attacker knowledge

(A) Techniques to identify threats (per data level)

1. **Linkage risk analysis (record/attribute linkage):**
    - Try joining QIs (e.g., timestamp+GPS route+ZIP+gender) with public sources (voter rolls, social media, transit passes) to estimate re-identification probability.
2. **k-Anonymity/ℓ-Diversity/t-Closeness checks on released tables:**
    - Compute EC sizes; verify SA diversity & distribution closeness to population to prevent skewness attacks.
3. **Trajectory re-identification tests (mobility data):**
    - Uniqueness of 4 spatio-temporal points can re-identify most people; run uniqueness metrics & apply **k-map** or **geo-indistinguishability**.
4. **Differencing/Query auditing for aggregates:**
    - Detect whether sequences of overlapping queries leak an individual (use **differential privacy** or query budgets).
5. **Membership & attribute inference simulations (ML models):**
    - Run known attacks against released models to assess leakage.
6. **Data-fusion threat modeling:**
    - Red-team compositions across domains (e.g., utilities + health sensors + demographics) to surface cross-dataset attacks.
7. **Access-path analysis & logs:**

- Who can see raw vs anonymized? Evaluate insider threats.

(B) What knowledge an adversary uses (with scenario examples)

- **Background knowledge:** daily commute pattern (home 7:45 AM at (lat,lon); hospital visit on Tuesdays).
- **Auxiliary datasets:** voter file (DOB, ZIP), ride-share receipts, CCTV timestamps, social posts ("just reached clinic").
- **External dimensions:**
  - **Identity linkage:** match unique route+time across Transport & Demographics → re-identify person.
  - **Attribute inference:** once linked, deduce **Health** attribute (e.g., "asthma" from frequent air-quality monitor alerts near their ID).
  - **Presence inference:** from Utilities smart-meter drop at 10:00–14:00 plus 911 logs → occupant likely away & incident location known.

Mitigations to cite: stronger generalization of time/space (coarser grids, time-bucketing), differential privacy on aggregates, suppression of rare patterns, query auditing, strict purpose limitation & minimization, federated analyses where possible.

---

5) Cross-border access under EU/Swiss data-residency laws

Problem: Data **must not leave** the country, but ops teams elsewhere need insights/help-desk views.

Suitable method (with types), diagram, and comparison

Chosen method: Remote computation with data localization — combine Federated Analytics / Learning, Privacy-Preserving Query Execution, and (optionally) Secure Computation.

Types / options under this umbrella:

1. **Federated Analytics (FA):** send queries/analytics code **to** each country; return only **aggregates with Differential Privacy** noise.
2. **Federated Learning (FL):** train models locally; share **model updates** (gradients) — optionally secured with **secure aggregation** (cryptographic).

3. **Privacy-Preserving SQL Gateway:** read-only views served **in-country**, foreign users connect via **zero-trust proxies** to run **parameterized queries**; raw rows never exit; sensitive columns **tokenized** or **FPE** (format-preserving encryption) with in-country HSM.
4. **Secure Multi-Party Computation (MPC)** / **Homomorphic Encryption (HE):** for select high-sensitivity computations where even aggregates must be cryptographically protected.

Text diagram (what to draw):

[Ops App (Foreign)] --TLS/Zero-Trust--> [In-Country Proxy]

--> [Local Data Lake / DB]

| (Pseudonymization, HSM keys)

|-- Federated Analytics Engine (DP noise)

|-- Federated Learning Worker (Secure Agg)

<-- return: DP aggregates / model deltas / masked views

Why this satisfies laws

- **Data never crosses borders**; only **aggregates/model deltas/masked views** do.
- **Keys (HSM) reside locally**; foreign staff cannot decrypt raw PII.
- **Auditable & revocable** access; least privilege via RBAC/ABAC.

Examples

- **Customer 360 KPIs:** FA computes churn rate per segment with **(ε,δ)-DP** noise.
- **Ticket assist:** help-desk sees **tokenized** customer fields + last-N transactions via **country-resident gateway**.
- **Model training:** FL trains fraud model across countries; only encrypted gradient sums leave each region (**secure aggregation**).

Benefits vs other methods

Method	Pros	Cons
Remote computation (FA/FL + DP + gateway)	Legally compliant; high utility; scalable; auditable	Engineering overhead; latency to remote regions

Method	Pros	Cons
Pure anonymization/pseudonymization	Simple to start	Risk of re-ID (linkage); utility loss; hard to guarantee
Data mirroring with SCCs	Easy analytics	Often **non-compliant** for sensitive data; legal risk
HE/MPC only	Strongest privacy	Heavy compute; narrow workloads; costly