A public-sector agency wants to release a dataset for researchers studying algorithmic bias in access to financial services. The dataset includes **10,000 individuals** and the following attributes:

- **Age** (integer)
- **Gender** (Male, Female, Other)
- **Ethnicity** (7 categories)
- **ZIP Code** (5-digit)
- **Annual Income** (continuous, in euros)
- **Loan Approval Status** (Approved / Denied)
- **Credit Score Group** (Low / Medium / High)
- **Education Level** (5 categories)

Researchers want the dataset to be **useful for studying fairness**, but the agency must ensure that the dataset satisfies **k-anonymity with k = 10** before release.

The quasi-identifiers (QIs) are:

- Age
- Gender
- Ethnicity
- ZIP Code
- Education Level

Loan Approval Status is considered **sensitive**, and Income and Credit Score Group are **non-identifying features** used for fairness analysis but may still influence re-identification risk indirectly.

Part A — Fairness Assessment Before Privacy Transformation

You are given the following high-level statistics from the *raw (non-anonymized)* dataset:

1. **Approval rates**:
    - Male: 68%
    - Female: 54%
    - Other: 61%
2. **Approval by Ethnicity**:
    - Group A: 70%
    - Group B: 48%
    - Group C: 52%
    - Group D: 67%
    - Group E: 60%
    - Group F: 50%
    - Group G: 49%

3. A logistic regression classifier trained on the raw dataset shows:
    - **Gender coefficient = –0.33**
    - **Ethnicity_Group_B coefficient = –0.58**

Questions (Part A)

1. **Identify which groups appear disadvantaged** using at least *two* fairness concepts
    - Provide conceptual reasoning without computation.
2. The agency argues that "the model is fair enough because all variables were included purely for predictive accuracy."
    - Explain why this argument is flawed.
3. Discuss how **intersectional fairness issues** could arise from combinations such as:
    - Female + Group B
    - Other + Group F
    - Female + Low Education Level

    What types of biases might be hidden when looking only at single-attribute fairness?
4. Describe at least **two potential fairness interventions** that could be performed on the *raw* dataset before anonymization, and discuss:
    - Their conceptual effect
    - Possible risks when later applying k-anonymity

    (e.g., fairness repair may make privacy harder to satisfy)

Part B — Achieving k-Anonymity

The agency applies a **k = 10** anonymization process using generalization and suppression. After anonymization:

- ZIP Code is generalized from 5-digit to 3-digit prefixes
- Age is bucketed into 5-year age ranges
- Ethnicity and Education Level undergo category grouping
- 4% of records are suppressed entirely because they would violate k-anonymity otherwise

Questions (Part B)

5. Explain **why Gender is left unmodified** in many k-anonymity transformations, even though it is a quasi-identifier.
6. Discuss how the generalization of ZIP Code and Age impacts:
    - **Re-identification risk**
    - **Data utility** for fairness analysis

    Provide at least one positive and one negative effect for each.
7. Describe how category grouping for Ethnicity may:
    - **Reduce privacy risk**, but
    - **Introduce or hide fairness issues**, especially for minority subgroups.
8. The suppressed records disproportionately include individuals in the groups:
    - Ethnicity Group F
    - ZIP Codes in sparsely populated rural areas
    - "Other" gender category

    Explain how suppression can **create fairness distortions**, even if unintentionally.

Part C — Synthesis

9. The agency wants a dataset that satisfies **both**:
    - k-anonymity
    - The ability to perform fine-grained fairness auditing on protected groups

    Explain whether this is possible in principle.

    - If yes, under what strict conditions?
    - If not, why do fairness and privacy requirements fundamentally conflict?
      Your answer should reason about **information loss**, **representativeness**, and **risk of re-identification**.
10. Propose a **conceptual pipeline** (no code) that balances fairness and privacy in data preparation.
    Your pipeline must include:
    - Data collection considerations
    - Pre-processing
    - Fairness repair
    - Privacy transformation
    - Post-release validation
      Explain the rationale behind each step and highlight potential failure points.

Part A — Fairness Assessment Before Privacy Transformation

1. Disadvantaged groups & fairness notions

From raw statistics:

- **Gender**
  - Male: 68% approved
  - Female: 54%
  - Other: 61%
- Females have substantially **lower approval** than males (–14 percentage points).
- The "Other" category is also lower than males (–7 pp), though less so.

  So **Females** are clearly disadvantaged; "Other" may also be disadvantaged but to a lesser extent.

- **Ethnicity**
  - Group A: 70%
  - Group D: 67%
  - Group E: 60%

    vs.
  - Group B: 48%
  - Group C: 52%
  - Group F: 50%
  - Group G: 49%

  Groups **B, F, G (and to a lesser extent C)** appear disadvantaged relative to A/D.

If we think in terms of **disparate impact** (ratio of approval rates, e.g. group B / group A):

- Group B: 48% / 70% ≈ 0.69 → less than the common "80% rule" threshold → likely problematic.
- Similarly for G and F vs A.

The **negative regression coefficients** (–0.33 for Gender, –0.58 for Ethnicity Group B) suggest that:

- Being Female (if coded that way) and being in Group B **reduces the log-odds of approval**, even after controlling for other variables.
- This aligns with the descriptive fairness indicators.

Conclusion: likely disadvantaged:

- **Females**, possibly "Other" gender relative to Males.
- **Ethnic groups B, F, G** (and perhaps C).

2. The agency's argument: *"The model is fair enough because all variables were included purely for predictive accuracy."*
Problems:

- **Accuracy ≠ fairness.** A model can be highly accurate and still treat protected groups systematically worse (e.g., deny loans more often to equally creditworthy women or Group B individuals).
- **Historical/structural bias.** The training data may already encode discrimination (e.g., past biased approvals). Optimizing accuracy will *reproduce* and *legitimize* existing inequities.
- **Correlated features.** Even if protected attributes were not used, other features (ZIP code, income, education) may act as proxies and perpetuate discrimination.
- **Fairness requires explicit constraints/analysis.** Fairness definitions (statistical parity, equal opportunity, etc.) require intentional checking and possibly modifying the model/data, not just "letting accuracy decide."

3. Groups like:

- Female + Group B
- Other + Group F
- Female + low education

may experience **compounded disadvantage**:

- A woman in Group B might face bias associated with both gender *and* ethnicity.
- "Other" gender in a small ethnic group (e.g., F) might be extremely rare, leading to:
  - Higher model uncertainty,
  - Potential over-penalization or conservative decisions.
- Female + low education: the model might have learned a particularly low approval rate for such combinations.

When we only inspect **single attributes** (gender OR ethnicity), we can miss:

- **Hidden subgroups** that are much worse off (e.g., females overall at 54% approval, but "Female + Group B" at, say, 35%).
- **Simpson's paradox** effects, where aggregate statistics hide subgroup disparities.
- Bias that only appears when multiple attributes intersect, i.e., **intersectional unfairness**.

4. Two possible **pre-anonymization fairness interventions**:

1. **Reweighing or resampling**
   - Increase weight of disadvantaged groups (e.g., Female, Group B/F/G) so the model "pays more attention" to fairness.
   - Conceptual effect: push the learned decision boundary to treat protected groups more similarly to advantaged groups.
   - Risk for k-anonymity:

        - If you upsample small groups, you may create *more* rare combinations of quasi-identifiers.
        - This can require **more aggressive generalization or suppression** to maintain k = 10, possibly **destroying** some of the fairness gains or making those groups vanish post-anonymization.
2. **Pre-processing / label modification ("massaging")**
    - Slightly adjust labels (e.g., flipping some "Denied" to "Approved" for qualified individuals in disadvantaged groups) to remove unjustified disparities.
    - Conceptual effect: enforce closer parity or equal opportunity in the data.
    - Risk for k-anonymity:
        - These carefully corrected patterns may be obscured by later generalization and suppression.
        - Suppressed records might disproportionately include the carefully corrected minority examples, so **fairness repair is undone in the released data**.

In general, fairness interventions can **change distributions** and **highlight rare subgroups**, which later makes meeting k-anonymity harder and increases data loss.

Part B — Achieving k-Anonymity

5. Why Gender is often left unmodified

Reasons:

- **Low cardinality.** Gender typically has 2–3 categories; generalizing that further (e.g., all to "*") doesn't reduce uniqueness much because uniqueness often comes from high-cardinality QIs (ZIP, Age, etc.).
- **Utility for fairness analysis.** Gender is a key protected attribute; generalizing it away or suppressing it would prevent gender-based fairness studies.
- **Impact on k-anonymity.** k-anonymity mainly reduces re-identification risk by modifying high-dimensional or high-cardinality QIs. Leaving Gender as is often still allows constructing equivalence classes of size ≥ k using Age, ZIP, Ethnicity, Education.

So we often **keep Gender explicit** while generalizing other QIs more heavily.

6. Impact of ZIP and Age generalization

ZIP: from 5-digit to 3-digit prefix

- **Re-identification risk:**
    - Positive: Many exact ZIPs become a single region; individuals are less unique geographically.

    - Negative: If some 3-digit prefixes still correspond to low-population regions, uniqueness may remain high there.
- **Data utility for fairness:**
    - Positive: Still allows some regional fairness analysis (urban vs rural zones, regional trends).
    - Negative: Fine-grained discrimination (e.g., specific poor neighborhoods vs rich ones) can no longer be detected.

Age: bucketed into 5-year ranges

- **Re-identification risk:**
    - Positive: Makes rare exact ages less unique; a 47-year-old becomes part of, say, [45–49].
    - Negative: Very old or very young age buckets might still be sparse and require additional suppression.
- **Data utility for fairness:**
    - Positive: Still possible to analyze age-related fairness at a coarse level (young vs middle-aged vs elderly).
    - Negative: Subtle age thresholds (e.g., late 20s vs early 30s) might matter for model behavior but are now obscured.

7. Ethnicity grouping: privacy vs fairness

Grouping Ethnicity categories:

- **Privacy benefit:**
    - Larger, more homogeneous equivalence classes → fewer unique combinations → lower re-identification risk.
    - Especially helpful for very small ethnic minorities.
- **Fairness risks:**
    - **Masking minority discrimination:** If Group F and G are severely disadvantaged but merged with A and D into a "Combined ethnic group," the overall group might look "average" or even advantaged.
    - **Loss of granularity:** Different ethnic groups have different histories and structural disadvantages; aggregation hides these distinctions.
    - **Token representation:** Some minorities may become a tiny fraction of a larger category; their specific unfair treatment becomes invisible in aggregated fairness metrics.

So ethnicity grouping can **protect privacy but hide or even invert apparent fairness patterns**.

8. Fairness distortions from suppression

Suppression targets records that cannot be made k-anonymous even after generalization, often those with **rare combinations** of QIs:

- Groups mentioned:
  - Ethnicity Group F
  - Rural ZIP codes
  - "Other" gender

These are exactly the **already-vulnerable or minority groups**. Consequences:

- Under-representation or **complete removal** of these individuals from the released dataset.
- Resulting dataset looks:
  - More homogeneous.
  - Less discriminatory, simply because the worst-treated groups are missing.
- Fairness metrics become **optimistic** ("no disparity") not because the system is fair, but because affected people have been **silenced via suppression**.

Part D — Synthesis

9. Can we have both k-anonymity and fine-grained fairness auditing?

In principle:

- It's **sometimes possible**, but only under **strong conditions**:
  - Very **large dataset**, so even fine-grained groups (e.g., Female + Group B + specific age bucket + region) have at least k individuals.
  - QIs can be carefully chosen so that protected/analytic attributes (Gender, Ethnicity) are not over-fragmented.
  - Population is **dense** across combinations; no extremely rare subgroups.

If those conditions hold, you can:

- Enforce k-anonymity mainly on non-protected QIs.
- Keep detailed protected attributes intact.
- Maintain enough counts per subgroup to compute fairness metrics.

In practice, there's a fundamental tension:

- **Fairness auditing needs granularity**:
  - Small, vulnerable subgroups.
  - Intersectional combinations.
  - Outliers and rare patterns.
- **Privacy (k-anonymity) hates granularity**:
  - Rare combinations are exactly what must be generalized or suppressed.
  - The more detailed your subgroup definitions, the more likely they fall below k.

This leads to:

- **Information loss**: more generalization means fewer meaningful fairness comparisons.

- **Representativeness issues**: suppression tends to discard rare/minority records → fairness metrics no longer reflect reality.
- **Re-identification risk** vs **granular fairness**: To protect individuals, you must blur exactly the distinctions fairness wants to examine.

So, except in very large, balanced datasets, **strong privacy and fine-grained fairness are in tension**; there's a trade-off to navigate, not a free lunch.

10. Conceptual pipeline balancing fairness & privacy

Here's a high-level pipeline:

1. **Data collection & governance**
    - Clearly define:
        - **Protected attributes** (Gender, Ethnicity, etc.) that *must* be collected for fairness.
        - **Quasi-identifiers** (Age, ZIP, Education) that pose privacy risks.
    - Minimize unnecessary personal data; don't collect extra high-risk identifiers.
    - Failure point: under-collecting protected attributes → fairness cannot be audited later.
2. **Pre-processing (before any fairness or privacy operations)**
    - Clean missing values, standardize categories, handle outliers.
    - Check distributions of key groups: are minority groups large enough for statistical analysis?
    - Failure point: silent data quality issues can later be wrongly attributed to "bias."
3. **Fairness diagnostics on raw data (internal only)**
    - Compute fairness metrics:
        - Statistical parity, disparate impact, equal opportunity, etc.
        - Use intersectional groups (Gender × Ethnicity × rough Age).
    - Identify:
        - Which groups are disadvantaged.
        - Where sample sizes are too small for reliable conclusions.
    - Failure point: misinterpretation of noisy subgroup metrics if groups are extremely small.
4. **Fairness interventions (still on raw data, internal)**
    - Consider:
        - **Pre-processing**: reweighing, balanced subsampling, label massaging.
        - **In-processing**: fairness-constrained model training (if model is part of pipeline).
    - Evaluate the impact on fairness metrics and utility.
    - Failure point: over-correction leading to reverse discrimination or severe utility drop.
5. **Privacy transformation design**
    - Decide on:
        - k-anonymity parameters (k, generalization hierarchies).
        - Possibly complement with l-diversity or consider DP for query-based access.

- Strategy:
  - Keep **protected attributes explicit** where possible (Gender, Ethnicity).
  - Focus generalization on high-cardinality QIs (Age, ZIP, Education) and possibly Income.
  - Explicitly monitor impact of generalization/suppression on **disadvantaged groups** (don't let them vanish entirely).
- Failure point: suppressing too many records from sensitive minorities → fairness invisibility.

6. **Post-transformation validation (on the anonymized data)**
   - Re-compute fairness metrics on the released/anonymized dataset.
   - Compare to the **pre-anonymization fairness**:
     - Where are metrics stable?
     - Where have they become unreliable or misleading?
   - Document:
     - Data utility limits.
     - Groups for which fairness metrics are no longer interpretable.
   - Failure point: releasing anonymized data with fairness metrics that appear fine but are artifacts of suppression/aggregation.
7. **Documentation & access control**
   - Provide **data sheets / model cards** that:
     - Describe how fairness and privacy were balanced.
     - Explicitly note limitations ("fine-grained ethnicity fairness cannot be evaluated from the public release").
   - For sensitive fairness audits, allow **restricted access** to less-aggregated data under strong governance (ethics review, secure environment).
   - Failure point: external users misunderstanding the data and drawing incorrect fairness conclusions.