Case Competition: Data Preprocessing Roles and Decisions

# TINA - EDA Specialist & Data Quality Lead

Primary Role: Exploratory Data Analysis and foundational data understanding

## Responsibilities:

* Overall dataset inspection and structure analysis
* Target variable distribution analysis (critical for imbalanced data)
* Missing value pattern identification across all variables
* Data quality assessment (duplicates, outliers, anomalies)
* Baseline statistics and distributions
* Visualization of key patterns
* Foundation insights that guide team preprocessing decisions

# ANDREW - Categorical & Temporal Processing Specialist

Primary Role: Handle categorical encoding and temporal variable transformations

## Variables Assigned:

* Temporal Variables: DAYS\_BIRTH, DAYS\_EMPLOYED, DAYS\_REGISTRATION, DAYS\_ID\_PUBLISH, DAYS\_LAST\_PHONE\_CHANGE
* Categorical Variables: All categorical variables (CODE\_GENDER, NAME\_\* variables, FLAG\_\* variables)
* High-Cardinality Variables: OCCUPATION\_TYPE, ORGANIZATION\_TYPE (special encoding needed)
* Missing Value Strategy: Systematic imputation approach for all variable types
* Document Flags: FLAG\_DOCUMENT\_\* variables (21 document verification flags)

# RELLIKSON - Feature Engineering

## Variables Assigned:

* ML-Derived Features: Customer risk segments (K-means clustering), anomaly scores (Isolation Forest)
* Advanced Transformations: Percentile rankings, standardization for ML algorithms
* Feature Selection: Statistical and tree-based feature importance ranking

# Data Preprocessing Decisions

## Will you create dummy variables, one hot encoding, or let tree-based methods split categories?

We use different approaches for different situations:  
- Simple Yes/No flags (like car ownership) → Just convert Y=1, N=0  
- Categories with few options (like education) → One-hot encoding  
- Categories with many options (like occupation) → Target encoding  
Why? Because creating 50+ dummy variables would make the model too complicated

## Will you bin categories together or not?

Yes, but only when necessary:  
- We combine rare categories (less than 2% of data) into 'Other'  
- Only for high-cardinality variables like ORGANIZATION\_TYPE  
Why? Rare categories don't have enough data to be reliable

## Are there any duplicate rows?

No duplicate rows found (checked above)  
- We ran df.duplicated().sum() and got 0  
- Each loan application is unique, which makes sense

## How will you handle date variables?

Convert negative days to positive years:  
- DAYS\_BIRTH → AGE\_YEARS (more understandable)  
- DAYS\_EMPLOYED → EMPLOYMENT\_YEARS  
Why? Nobody thinks in 'days before application' - years make sense

## How will you handle NULL/missing values?

Different strategies based on how much is missing:  
- Lots missing (like occupation 31%) → Fill with 'Unknown'  
- Some missing (like credit scores) → Fill with median  
- Little missing → Fill with most common value  
Why? Missing patterns tell us something - don't just delete them

# WHY WE MADE OUR DECISIONS

Why different encoding strategies?  
- Simple flags (car/house ownership): Just 1/0 because it's yes/no  
- Education levels: One-hot because there's a natural order  
- Occupation (18+ categories): Target encoding because 18 dummy variables is too many  
- We didn't want to create 50+ columns that would confuse the model

Why bin rare categories?  
- If only 5 people have job 'Astronaut', the model can't learn from it  
- Better to group rare jobs into 'Other' so the model has enough data  
- We only did this for variables with 20+ categories

Why convert days to years?  
- DAYS\_BIRTH = -14086 means nothing to humans  
- AGE\_YEARS = 38.6 makes perfect sense  
- Same with employment - 'worked 3.2 years' vs 'worked -1168 days'

Why different missing value strategies?  
- If 31% of people don't have occupation data, 'Unknown' is meaningful  
- If only 0.3% missing social circle data, just use the average  
- Missing credit bureau data means 'no credit history' = fill with 0  
- Each missing pattern tells a different story

Why customer segmentation?  
- Not all customers are the same - some are naturally higher risk  
- Grouping similar customers helps identify patterns  
- Anomaly detection finds the weird cases that might default  
- Business can create different strategies for different segments