🎓 CREDIT EDA ASSIGNMENT

Name: Krish Mendiratta

🔷 Slide 1: Problem Statement

We are exploring credit risk based on historical customer data.
Goal: Identify factors affecting default probability using exploratory data analysis (EDA).
Data Sources:

- `application_data.csv`
- `previous_application.csv`

🔷 Slide 2: Data Handling Approach

✅ Data Loaded from CSV
✅ Inspected shape, data types, and general completeness
✅ Removed columns with extreme missing values
✅ Imputed relevant missing values logically (e.g., fill with 0 or mode)
✅ Handled outliers and inconsistent data (e.g., extreme `DAYS_EMPLOYED`)

🔷 Slide 3: Missing Data Strategy

- Dropped columns with >40% missing values
- Imputed others:
  - `NFLAG_INSURED_ON_APPROVAL` → filled with 0
  - Categorical features → filled with 'Unknown'
- Carefully preserved important features to retain signal

🔷 Slide 4: Outlier Detection

- Flagged unrealistic values like:
  - `DAYS_EMPLOYED` > 100,000 → replaced with NaN
- Checked distributions using boxplots and summary stats
- Removed extreme z-score outliers for numerical features where relevant

🔷 Slide 5: Target Imbalance

- `TARGET` = 1 (Defaulters): ~8.1%
- `TARGET` = 0 (Non-Defaulters): ~91.9%

⚠️ Indicates class imbalance → Consider when modeling

📉 Slide 6: Outlier Treatment

- Flagged extreme values in:
  - `DAYS_EMPLOYED` > 365000 (anomaly → treated or excluded)
  - High-income anomalies handled using quantile capping
- Visual validation using boxplots

⚖️ Slide 7: Class Imbalance Check

- **TARGET (Default)** distribution:
  - 0 (No Default): ~91%
  - 1 (Default): ~9%
- Severe imbalance detected → Consider for future modeling

🔎 Slide 8: Univariate Analysis

- Key numerical variables:
  - `AMT_INCOME_TOTAL`, `DAYS_BIRTH`, `AMT_CREDIT`
- Categorical variables:
  - `NAME_EDUCATION_TYPE`, `NAME_FAMILY_STATUS`
- Identified distributions and skewness

🔗 Slide 9: Bivariate Analysis

- Relationship with TARGET:
  - Older applicants less likely to default
  - Lower income and single/divorced individuals → higher risk
  - Default more common among lower education levels

🧠 Slide 10: Correlation with TARGET

- Weak correlations overall
- Slight signals:
  - `EXT_SOURCE_3`, `EXT_SOURCE_2` negatively correlated with default
  - Social circle features and days-related fields relevant

📂 Slide 11: Analysis of previous_application.csv

- Cleaned high-missing columns (e.g., `RATE_DOWN_PAYMENT`)
- Imputed remaining relevant fields
- Distribution of `NAME_CONTRACT_STATUS`:
  - Majority: **Refused**, some **Approved**, **Canceled**

📈 Slide 12: Univariate & Bivariate (Prev App)

- AMT_APPLICATION vs AMT_CREDIT: strong linear relationship
- NAME_CONTRACT_TYPE, NAME_CLIENT_TYPE affect approval rates
- Segment analysis by NAME_CONTRACT_STATUS:
  - Approved: lower risk
  - Refused: show higher application amounts and longer terms

🔗 Slide 13: Merging Datasets

- Merged `application_data` and `previous_application` on `SK_ID_CURR`
- Enabled deeper feature insights and cross-reference between past and current behavior

🧮 Slide 14: Correlation (Segmented)

- TARGET = 1 (Defaulters): Strongest correlations:
  - `DAYS_EMPLOYED` ~ `FLAG_EMP_PHONE`
  - `AMT_APPLICATION` ~ `AMT_GOODS_PRICE_y`
- TARGET = 0 (Non-defaulters):
  - Similar high pairings, suggesting structural consistency

📌 Slide 15: Business Insights

- Features influencing defaults:
  - Longer unemployment, low EXT_SOURCE, high credit burden
- Suspicious patterns:
  - Many features are multicollinear (e.g., AVG vs MEDI)
- Risk flags:
  - Young age + low education + refused previous apps = high default

🧾 Slide 16: Final Recommendations

- Integrate EXT_SOURCE scores more deeply in decision logic
- Use previous application behavior (status, amount) in risk profiling
- Consider SMOTE or weighted models due to class imbalance

🎯 Slide 17: Summary

- Comprehensive EDA on both datasets
- Cleaned and merged for enriched understanding
- Identified key signals for credit risk
- Next: modeling or business decision pipeline