Assignment 1

Weight: *20% of Final Grade*
Format: *Python Notebook (.ipynb) and PDF/HTML Export*
Submission: *Individual*

Part 1: Feature Selection & Dimensionality Reduction

Dataset: Home Credit Default Risk
Business Context:
You are a Senior Data Scientist at **"Global Access Finance,"** a firm dedicated to providing credit to the "unbanked" population—people with little to no formal credit history. While the company has access to a massive dataset (including credit bureau data and internal application details), the model has become bloated. A model with hundreds of features is expensive to maintain, slow to run, and difficult to explain to financial regulators.
Your goal for Part 1 is to perform **Dimension Reduction and Feature Selection** to identify the most "signal-heavy" variables that predict whether a client will default (the TARGET variable), ensuring the model is lean, efficient, and explainable.

Tasks (50 Points):

1. **Initial Data Audit (10 pts):**
    - Identify and visualize columns with more than 40% missing values.
    - Provide a business justification: Should these be dropped or imputed? Perform the necessary action.
2. **Multicollinearity Analysis (15 pts):**
    - The dataset contains many "bureau" and "credit_bureau_balance" features that likely track similar information.
    - Construct a correlation matrix. Identify pairs of features with a correlation coefficient > 0.9 and remove one from each pair to reduce redundancy.
3. **Filter & Wrapper Methods (15 pts):**
    - Apply a **Feature Importance** technique (e.g., using a Random Forest, etc...) to rank the top 20 most impactful features.
    - Explain why the business might prefer a model with 20 features over 100, even if there is a slight drop in accuracy.

Part 2: Feature Engineering & Temporal Analysis

Dataset: Municipal Crime Data (One Year Prior)
Business Context:
You have been hired by **"SafeCity Solutions,"** a tech startup that provides predictive analytics to urban planners and private security firms. The raw crime data (provided in the image) contains basic administrative info. However, raw timestamps and street addresses are not directly useful for a machine learning model.

Your goal for Part 2 is to **engineer new features** that reveal patterns in human behavior, geography, and time to help predict whether an incident will result in an ARREST.

Tasks (50 Points):

1. **Temporal Engineering (15 pts):**
    - Using the DATE OF OCCURRENCE column, extract:
        - Hour_of_Day (0–23)
        - Day_of_Week (Monday–Sunday)
        - Is_Weekend (Binary flag)
    - **Visualization:** Create a plot showing which hour of the day has the highest volume of crimes.
2. **Categorical Consolidation (10 pts):**
    - The PRIMARY DESCRIPTION column contains many specific crimes. Simplify this for the model by grouping them into three broader categories: *Violent Crime*, *Property Crime*, and *Public Order/Other*.
    - Explain the business rationale for "Bucketing" categorical variables.
3. **Spatial Bucketing (10 pts):**
    - Using the WARD or BEAT columns, create a "Risk Level" feature by calculating the historical frequency of crimes in that area (High, Medium, Low).
    - Drop the raw X/Y COORDINATE columns once your feature is created to prevent model over-fitting.
4. **Encoding & Scaling (15 pts):**
    - Convert the LOCATION (e.g., APARTMENT, STREET) into numerical format using **One-Hot Encoding**.
    - Rescale the LATITUDE and LONGITUDE features using **Standardization (Z-score)** so the model treats geographic distance appropriately.
    - Handle the ARREST and DOMESTIC columns as binary (0/1) indicators.

Submission Guidelines

Documentation & Rationale:

- **Markdown Cells:** For every technical step, include a Markdown cell explaining **WHY** this helps a business decision-maker. (e.g., "We used One-Hot Encoding for Location because ML algorithms cannot process the word 'Apartment' as a mathematical input.")
- **Code Comments:** Comment your code to explain complex logic.
- **Cleanliness:** Ensure all plots have titles, axis labels, and legends.

Grading Rubric (Total 100%):

- **Data Wrangling Accuracy (30%):** Correct handling of missing values, types, and duplicates.
- **Methodological Rigor (40%):** Correct application of PCA, Feature Selection, and Encoding techniques.

- **Business Insight (20%):** Quality of the explanations regarding how these steps improve business ROI or model interpretability.
- **Presentation (10%):** Organization of the notebook and clarity of visualizations.

Note to Students: *Please use the CSV files provided in the course portal. Ensure your final notebook can run "from top to bottom" without errors before submitting.*