Assignment: LendSmart Credit Risk Analysis (Full Business Case)

1. Project Objective

You are a data science consultant for **LendSmart**, a FinTech company specializing in personal and small business loans. The company's current loan portfolio has a 28% default rate, which management considers too high.

Your objective is to analyze the provided credit_risk_data.csv to build and evaluate a statistical model that predicts the likelihood of a new applicant defaulting on their loan (loan_status).

Your analysis will compare two key classification techniques, **Linear Discriminant Analysis (LDA)** and **Quadratic Discriminant Analysis (QDA)**, to determine the most effective and interpretable model.

Your final submission will be a complete consulting engagement, consisting of your technical analysis, a non-technical business report, and a stakeholder presentation.

2. The Business Case

LendSmart's core business problem involves a critical trade-off:

- **Approving a "bad" loan (a "False Negative" for risk):** This results in a direct financial loss when the applicant defaults. This is the **worst-case scenario**.
- **Rejecting a "good" loan (a "False Positive" for risk):** This results in lost business, forgone interest revenue, and a dissatisfied potential customer.

The goal is to develop a model that minimizes the first case (approving "bad" loans) while keeping the second case (rejecting "good" loans) at an acceptable level. Your analysis will provide the data-driven recommendation on which model achieves the best balance.

3. Deliverables

You will submit three distinct deliverables:

1. **Deliverable 1: The Code (LendSmart_Analysis.ipynb)**
    - A single Jupyter Notebook containing your complete technical analysis, all code, visualizations, and *technical* interpretations. This is your "analytical appendix."
2. **Deliverable 2: The Business Report (Executive_Summary.pdf)**
    - A 1-2 page, code-free document written for a non-technical executive. It must summarize the business problem, your key findings, and your final recommendation.
3. **Deliverable 3: The Pitch (Video & Slides)**
    - **Slides (Presentation.pdf or .pptx):** A 5-7 slide deck that visually supports your presentation.
    - **Video (Presentation.mp4 or link):** A 5-7 minute recorded video presentation "pitching" your findings to the LendSmart management team.

4. Step-by-Step Instructions

Part A: The Jupyter Notebook (Your Analytical Work)

Your notebook should be well-organized and tell a clear story, similar to the marketing_discriminant_analysis.ipynb example. Use Markdown cells extensively to explain your steps and findings.

Section 1: Project Setup & Data Loading

- Import all necessary libraries (pandas, numpy, sklearn, matplotlib, seaborn).
- Load the credit_risk_data.csv file.
- Perform an initial inspection: .head(), .info(), .describe().
- Write a brief summary of your initial findings (e.g., "No missing values," "Data spans 3 years," etc.).

Section 2: Exploratory Data Analysis (EDA)

- **Target Variable:** Plot the distribution of loan_status and calculate the exact default rate.
- **Continuous Variables:** For key predictors (credit_score, annual_income, debt_to_income_ratio, etc.), create plots (e.g., box plots or histograms) that **compare the distribution for defaulters (1) vs. non-defaulters (0)**.
- **Categorical Variables:** Create bar plots showing the *mean default rate* for each category in education_level and marital_status.
- **Correlations:** Generate a correlation matrix heatmap for all numeric predictors. Note any variables with high multicollinearity.

Section 3: Data Preprocessing

- **Handle Categorical Data:** Convert education_level and marital_status into numerical dummy variables (pd.get_dummies()).
- **Define Predictors (X) and Target (y):** Create your X (features) and y (target) variables. Be sure to drop non-predictive columns like application_id.
- **Train-Test Split:** Split your data into X_train, X_test, y_train, y_test. Use test_size=0.2 and a random_state=42 for reproducibility.
- **Standardization:** Initialize a StandardScaler. **Fit** it *only* on X_train, then **transform** *both* X_train and X_test. This is critical for interpreting LDA coefficients.

Section 4: Statistical Assumption Testing (Written Discussion)

- In a Markdown cell, discuss the key statistical assumptions that differentiate LDA and QDA.
- **Multivariate Normality:** Discuss the assumption. Do your EDA plots suggest this is reasonably met?
- **Homogeneity of Covariance Matrices:** This is the *key difference*. Explain what it means (LDA assumes all classes share one covariance matrix; QDA does not). State your hypothesis: "If the covariance matrices are unequal, we expect QDA to outperform LDA."

Section 5: Model 1 - Linear Discriminant Analysis (LDA)

- Initialize LinearDiscriminantAnalysis and fit it on your **standardized** X_train and y_train.
- **Interpret Coefficients:** Extract the lda.coef_. Place them in a DataFrame with their feature names. Sort by absolute value.
- **Write a clear interpretation:** Which 3-5 variables are the *most important* drivers of default risk? What does the sign (+/-) tell you?

Section 6: Model 2 - Quadratic Discriminant Analysis (QDA)

- Initialize QuadraticDiscriminantAnalysis and fit it on your **standardized** X_train and y_train.
- *(Note: QDA does not produce simple linear coefficients, so no interpretation is needed here.)*

Section 7: Model Evaluation & Comparison

- Generate predictions (.predict()) for both models on the **X_test** data.
- **Confusion Matrices:** For *both* models, generate and plot a ConfusionMatrixDisplay.
- **Classification Reports:** For *both* models, print the classification_report.
- **ROC Curves:** Generate the RocCurveDisplay for *both* models and **plot them on the same axis** for a direct visual comparison. Report the **AUC (Area Under the Curve)** score for both.

Section 8: Technical Conclusion & Model Selection

- In a final Markdown cell, state which model (LDA or QDA) you select as the "best" technical model.
- **Justify your choice** using clear evidence from Section 7 (e.g., "QDA is the superior model as its AUC score was 0.89 compared to LDA's 0.83, and its Recall for the default class was 12 points higher...").

Part B: The Executive Summary (Your Business Report)

Write a 1-2 page professional report. **This document must not contain any Python code.** It should be structured as follows:

1. **Business Problem:**
    - Briefly state the challenge LendSmart faces (e.g., "LendSmart needs to reduce its 28% loan default rate...").
    - State your project objective (e.g., "Our team was tasked with building a statistical model to identify high-risk applicants...").
2. **Key Findings & Insights:**
    - What did you learn? **Translate your technical findings into simple business language.**
    - *Bad Example:* "The standardized LDA coefficient for credit_score was -2.5."
    - *Good Example:* "Our analysis revealed that **credit score is the single most important predictor of default**. Other key factors include an applicant's debt-to-income ratio and their total asset value."
    - Describe the "profile" of a high-risk applicant in simple terms.
3. **Model Performance & Selection:**
    - State which model you chose (e.g., "We compared two models and selected the **QDA model** as it was significantly more effective.").
    - Translate your best metric into a business-friendly statement:
        - e.g., "The new model successfully **identifies 85% of all actual defaulters** (Recall), a major improvement over random chance."
        - e.g., "Of the applicants the model flags as 'high-risk,' **90% are correct** (Precision), meaning we are not rejecting good customers unnecessarily."
4. **Final Recommendation:**
    - Give a clear "Go / No-Go" recommendation (e.g., "We recommend LendSmart deploy this QDA model...").
    - State the **business trade-off**. (e.g., "This model will save an estimated $X in default losses. However, it will also incorrectly flag an estimated 10% of 'good' applicants (False Positive Rate),

which we believe is a manageable cost for the added security.").
    - Suggest one clear next step (e.g., "Future work should focus on...").

Part C: The Video Presentation (Your Stakeholder Pitch)

Record a **5-7 minute video** presenting your findings to the LendSmart management team. Your accompanying slide deck should be 5-7 slides, clean, professional, and visual.

- **Slide 1: Title Page** (Project Title, Your Name)
- **Slide 2: The Business Problem** (The challenge: 28% default rate. The goal: A predictive model. Use simple icons and visuals.)
- **Slide 3: Key Insights: What Drives Default?** (Show a simple bar chart of your top 3-5 LDA coefficients. Label them clearly, e.g., "Credit Score," "Debt-to-Income Ratio." Explain what they mean.)
- **Slide 4: Model Performance: LDA vs. QDA** (Show your **combined ROC curve plot**. This is the one key technical chart. Explain it simply: "The blue line (QDA) is better because it's closer to the top-left corner and has a larger Area (AUC).")
- **Slide 5: The Business Trade-Off** (Show your best model's confusion matrix, but re-label it in plain English. e.g., **"1,000 Correctly Approved," "50 Wrongly Rejected," "150 Correctly Flagged," "25 Missed Defaulters."**)
- **Slide 6: Final Recommendation** (Summarize your recommendation, the business impact (e.g., "Catch 85% of defaulters"), and your proposed next step.)

5. Grading Rubric (100 Points Total)

Deliverable	Section	Criteria	Points
Deliverable 1: Jupyter Notebook (40 total)	EDA & Setup	Data loaded. Comprehensive EDA with plots *comparing* the two groups.	10
Deliverable 1: Jupyter Notebook	Preprocessing	Correct dummy encoding, train/test split, and **StandardScaler used correctly** (fit on train, transform both).	10
Deliverable 1: Jupyter Notebook	Assumption Testing	Clear **written discussion** of *why* the assumptions (Normality, Covariance) matter for LDA vs. QDA.	5
Deliverable 1: Jupyter Notebook	Model Fitting	LDA/QDA models are fit correctly. LDA **standardized coefficients are correctly interpreted**.	5
Deliverable 1: Jupyter Notebook	Model Evaluation	All metrics (Conf. Matrix, Class. Report, ROC/AUC) are present *for the test set*. **ROC curves are plotted on one graph** for comparison.	10

Deliverable	Section	Criteria	Points
Deliverable 2: Executive Summary (30 total)	Clarity & Professionalism	1-2 pages, **no code**, well-formatted, and free of jargon and typos.	5
Deliverable 2: Executive Summary	Problem & Objective	The business problem and project goal are clearly and concisely stated.	5
Deliverable 2: Executive Summary	Key Findings	**Translates technical findings** (e.g., coefficients) into plain English insights (e.g., "profile of a defaulter").	10
Deliverable 2: Executive Summary	Recommendation	Clear "Go / No-Go" choice. **Performance and trade-offs are explained in business terms** (e.g., "We will catch 85% of defaulters...").	10
Deliverable 3: Video & Slides (30 total)	Slide Deck	Slides are professional, **visual, and support the narrative** (not just walls of text).	10
Deliverable 3: Video & Slides	Presentation	Clear, confident, and persuasive. Stays within the 5-7 minute time limit.	10
Deliverable 3: Video & Slides	Content & Flow	Presentation tells a logical story (Problem -> Insight -> Model -> Recommendation) and **correctly translates technical findings** for a non-technical audience.	10
Total			**100**