Rubric

Spec Category	Spec Details
Goal	Your goal is to explore whether presidential inaugural speeches from Democratic and Republican presidents differ in how much unifying or polarizing language they use. You will scrape data from a historical archive of inaugural addresses, calculate relevant word usage metrics, visualize patterns, and build simple models to test relationships. The goal is to think critically about how language, politics, and data intersect — and to practice scraping, text-based feature engineering, and exploratory analysis.
Task	Your final deliverable will be a reproducible, well-documented Jupyter Notebook that includes: - Scraping inaugural speech texts from the web - Cleaning and structuring the data into a CSV with word counts and calculated ratios - Performing exploratory data analysis (EDA) and visualizations of unifying vs. polarizing language by political party - Building and evaluating a basic logistic regression model using word features to predict party affiliation - Reflecting on model performance, bias, and uncertainty - Including clear markdown documentation for future students to follow
Data	You will be scraping data from the UCSB American Presidency Project. A starter script and list of unifying/polarizing words are provided. You must: - Produce a cleaned, structured dataset of speeches with the following fields: president name, party, word counts, word ratios, overall ratio, and classification (unifying or polarizing) - Save the dataset as speech_data.csv and store it in the /data folder of your GitHub repo
Modeling	Your Notebook should include: - A point-biserial correlation between party and overall ratio - A logistic regression model predicting party using at least 3 engineered features - Evaluation using accuracy, classification report, and confusion matrix

	- (Optional) Calibration curve and Brier score for confidence and bias analysis
Bias & Reflection	You must reflect on your model's strengths and weaknesses. Address at least two of the following: - Bias in the dataset (e.g., historical context, speech length) - Party imbalance or missing data - Uncertainty in classification confidence - Low accuracy or overfitting from small sample size Write a short markdown cell discussing what went wrong, why, and how you'd approach it differently next time.

Submission Requirements	GitHub Repo must include:
	- Notebook that:
	  - Scrapes, cleans, and processes the speech data
	  - Engineers the necessary features
	  - Performs EDA
	  - Builds and evaluates a logistic regression model
	  - Includes markdown documentation
	  - Reflects on bias and uncertainty
	  - A README file briefly explaining your work and how to run your code (optional but encouraged).

Evaluation Criteria	**Data Collection:** Successful scraping and correct formatting of speech data **Feature Engineering:** Accurate calculation of text-based features (counts, ratios, classifications) **EDA and Visualizations**: Clear, labeled plots comparing political parties **Modeling**: Functional logistic regression with interpretation of results **Bias Reflection:** Honest discussion of data/modeling bias and uncertainty **Communication**: Well-commented code and markdown explanations