Revenue Classification and Profitability Analysis of Museums

ISyE 7406 Data Mining & Statistical Learning
Fall 2024
TEAM #27

Alejandra Sevilla

Georgia Tech

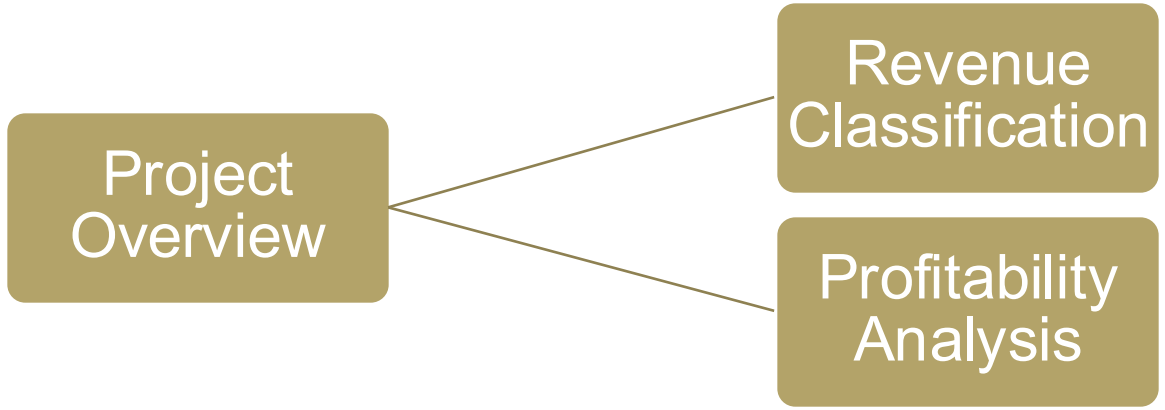
Project Overview

Problem Statement

- Classification of museums into **high and low revenue categories**.
- Determining the **profitability** of museums based on revenue and income data.

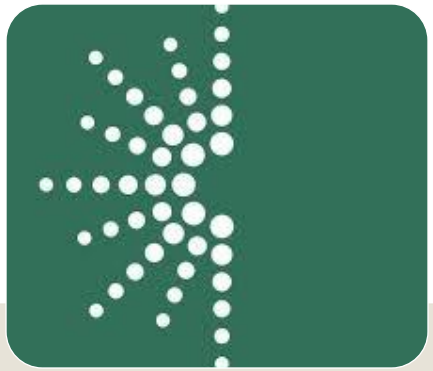
Research Questions

- What factors contribute to a museum being high revenue or low revenue?
- What are the characteristics of profitable versus non-profitable museums?



Georgia Tech

Institute of Museum and Library Services Dataset Overview



Data Source

- Institute of Museum and Library Services (IMLS)
- Link: Museum Data Files
- Data Format: CSV
- Useful For: Researchers, journalists, policymakers, public planners

Dataset Content

- Cover museums across all 50 states and the District of Columbia
- Three Main Files:
  - **File 1:** Museums by Discipline
  - **File 2:** General Museums
  - **File 3:** Historical Societies and Historic Preservation

Key Information Included

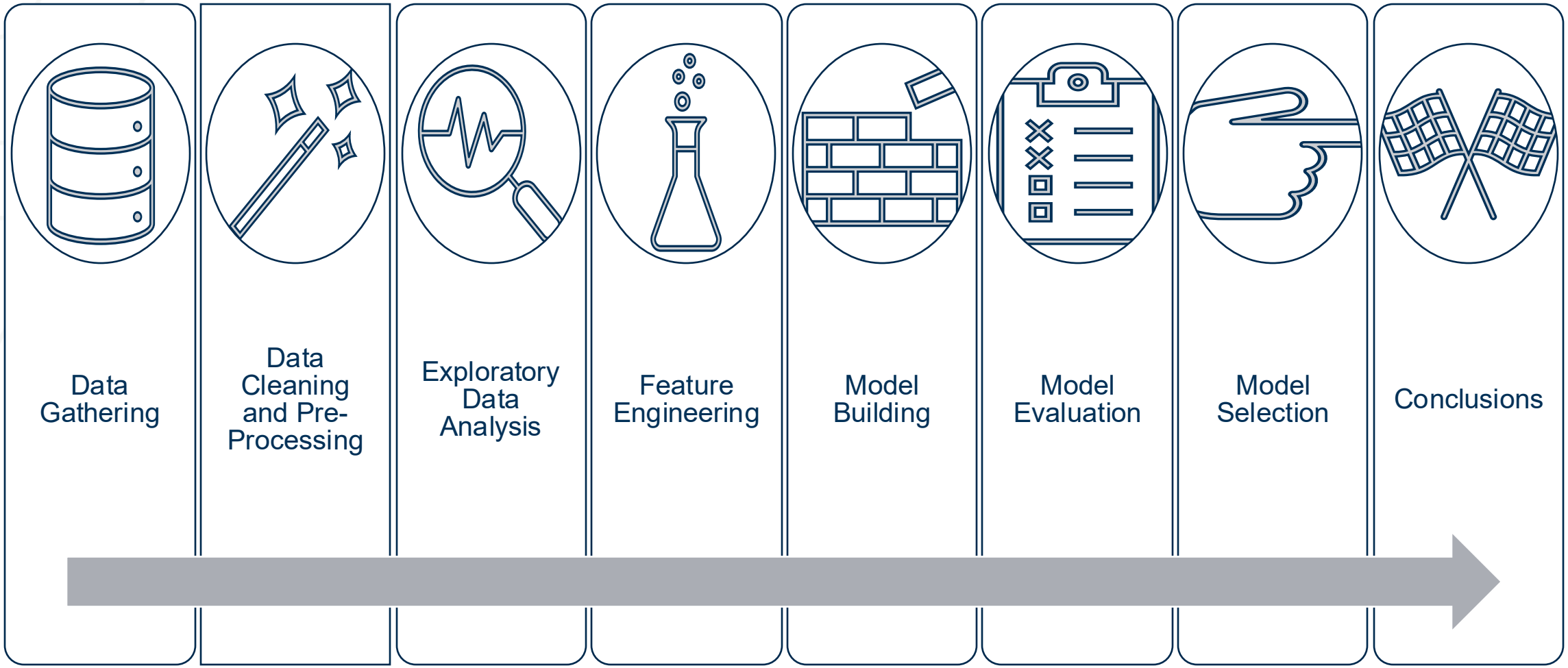
- **Institution Details:** Location, museum type, discipline, operational status.
- **Financial Information:** IRS 990 forms, revenue classifications.
- **Additional Details:** Geocode data, museum districts, co-located resources.

Data Collection

- Collected from public, federal, and nonprofit data sources, including IRS data, third-party contributions, private foundations, and user-contributed updates.
- Initially constructed in 2013 and updated in 2018.

Georgia Tech

Methodology and General Approach



Georgia Tech

Data Cleaning and Pre-Processing

Data Import and Initial Checks

- Imported three datasets with museum information
- Merged datasets after renaming columns

Handling Duplicates

- Identified duplicate rows
- No duplicate rows found

Removing Irrelevant Columns

- Dropped columns with unnecessary information (e.g., identifiers, addresses)
- Focused on columns that directly impact the analysis

Managing Missing Values

- Removed rows with missing values in Revenue and Income
- For Locale Type, filled missing values based on the most common value in each city
- For Region, imputed missing values based on state codes

Feature Engineering

- Replaced codes in variables as Region and Locale Type with descriptive names
- Created a new variable, Revenue to classify Revenue into High and Low
- Added Profitability, based on Income - Revenue
- Converted relevant columns to factors

Georgia Tech

Exploratory Data Analysis (EDA) Overview

What is EDA?

- EDA is the first step in understanding the dataset. It involves summarizing the main characteristics of the data through visualization and statistical techniques.

Why is EDA important?

- EDA helps identify data distribution, relationships between variables, and hidden patterns. It guides the next steps in modeling
- Detects outliers that could affect the analysis.

Main Focus of EDA

- For this project, the EDA focuses on identifying characteristics that make museums generate higher revenue than others.
- Explore patterns in profitability.

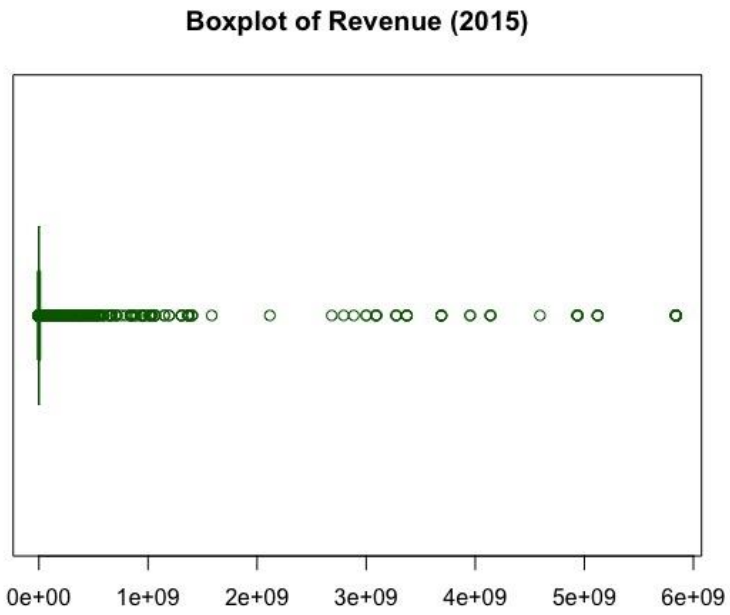
Key Questions Explored

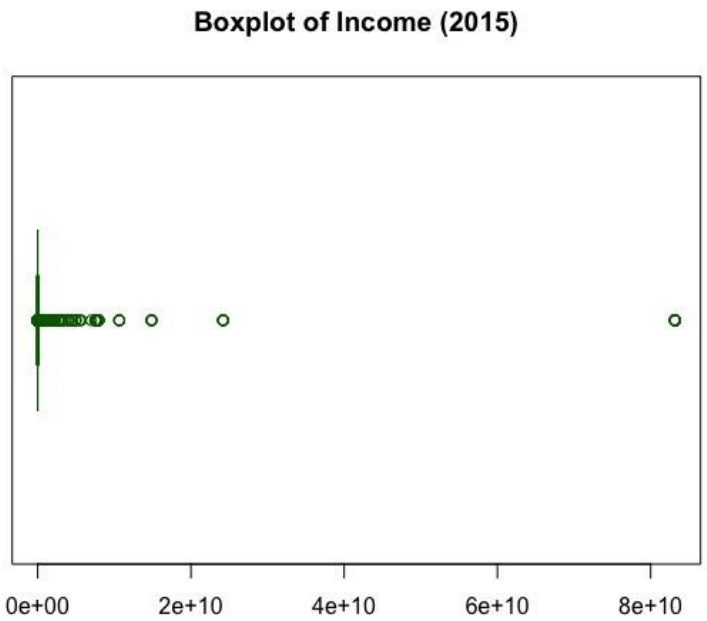
- What differentiates high-revenue museums from low-revenue ones?
- How does profitability vary by region, discipline, and other factors?
- Are there revenue patterns based on museum type, or location?

Georgia Tech

Revenue and Income Outliers and Descriptive Statistics

	Mean	Median	Min	Max
Income	113,800,000	1,455	0	83,180,000,000
Revenue	21,685,376	460	-2,127,393	5,840,349,457



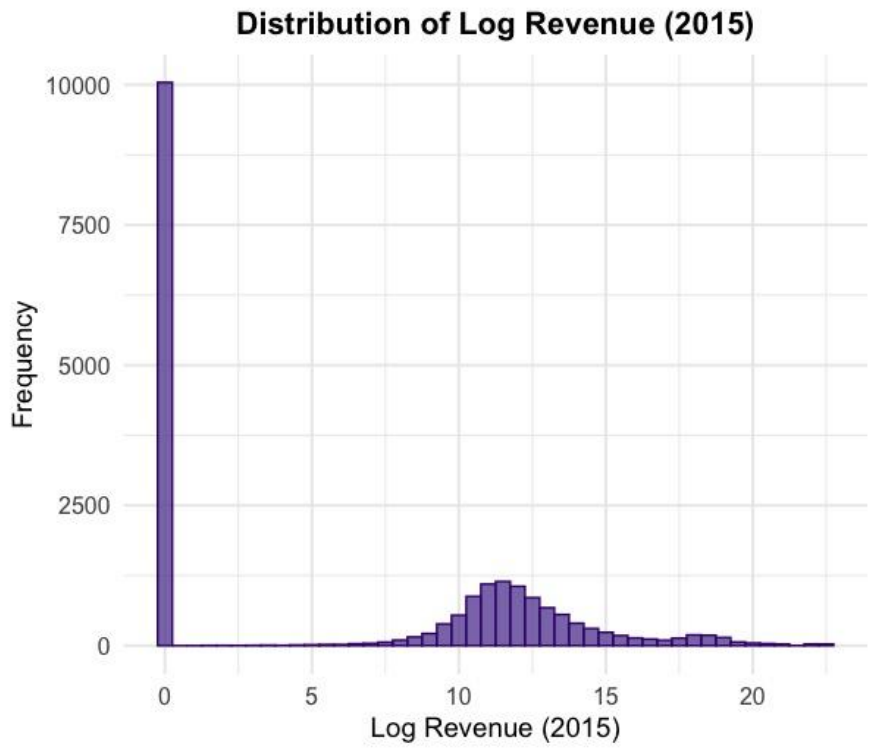


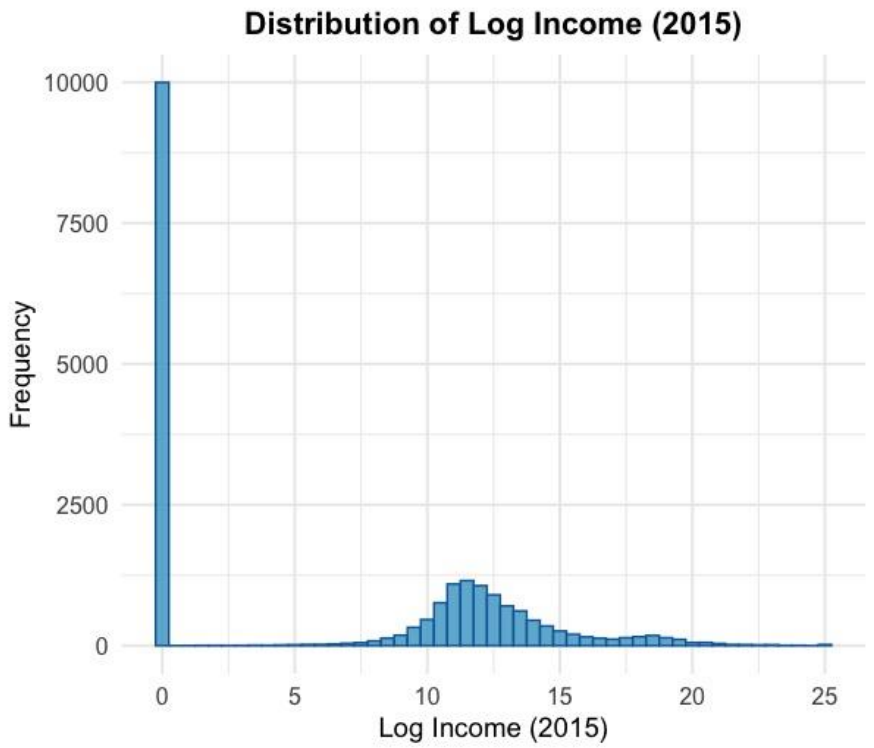
Main Observations:

- Income ranges from $0 to $83.18 billion, showing high variability across museums.
- Revenue ranges from -$2.13 million to $5.84 billion, indicating diverse financial situations.
- Extreme outliers in both Income and Revenue highlight high-earning museums.
- Outliers are retained as they represent real data needed for variability analysis.
- Mean Income ($113.8M) and Revenue ($21.69M) are much higher than the median, suggesting a right-skewed distribution due to a few high values.
- Most museums have low income and revenue, with high financial performance concentrated in a few institutions.

Georgia Tech

Revenue and Income Distributions



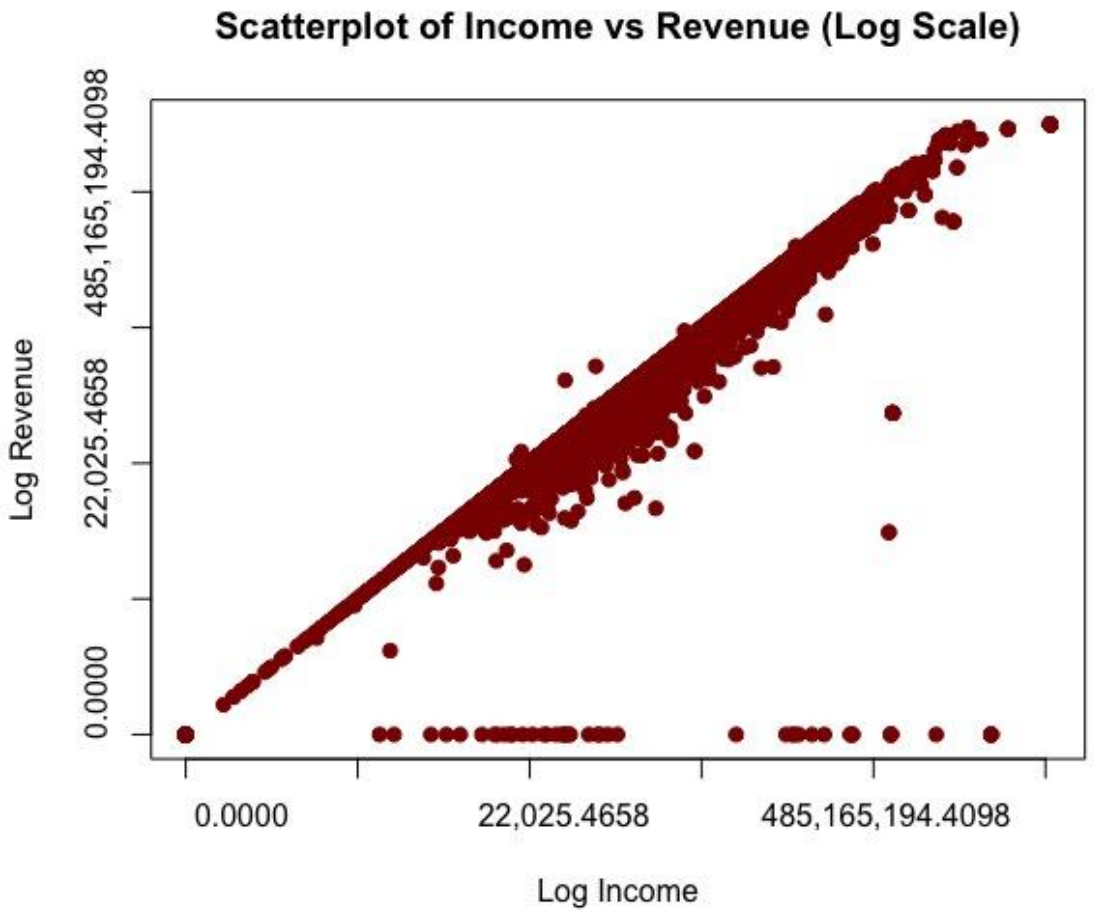


Main Observations:

- Both Income and Revenue distributions are heavily right-skewed, with most museums having low financial figures.
- Log transformation helps spread the data, making high-income outliers more distinguishable.
- Many values cluster near zero, suggesting most museums have very low income and revenue, while a few generate significantly more.

Georgia Tech

Income vs Revenue Relationship

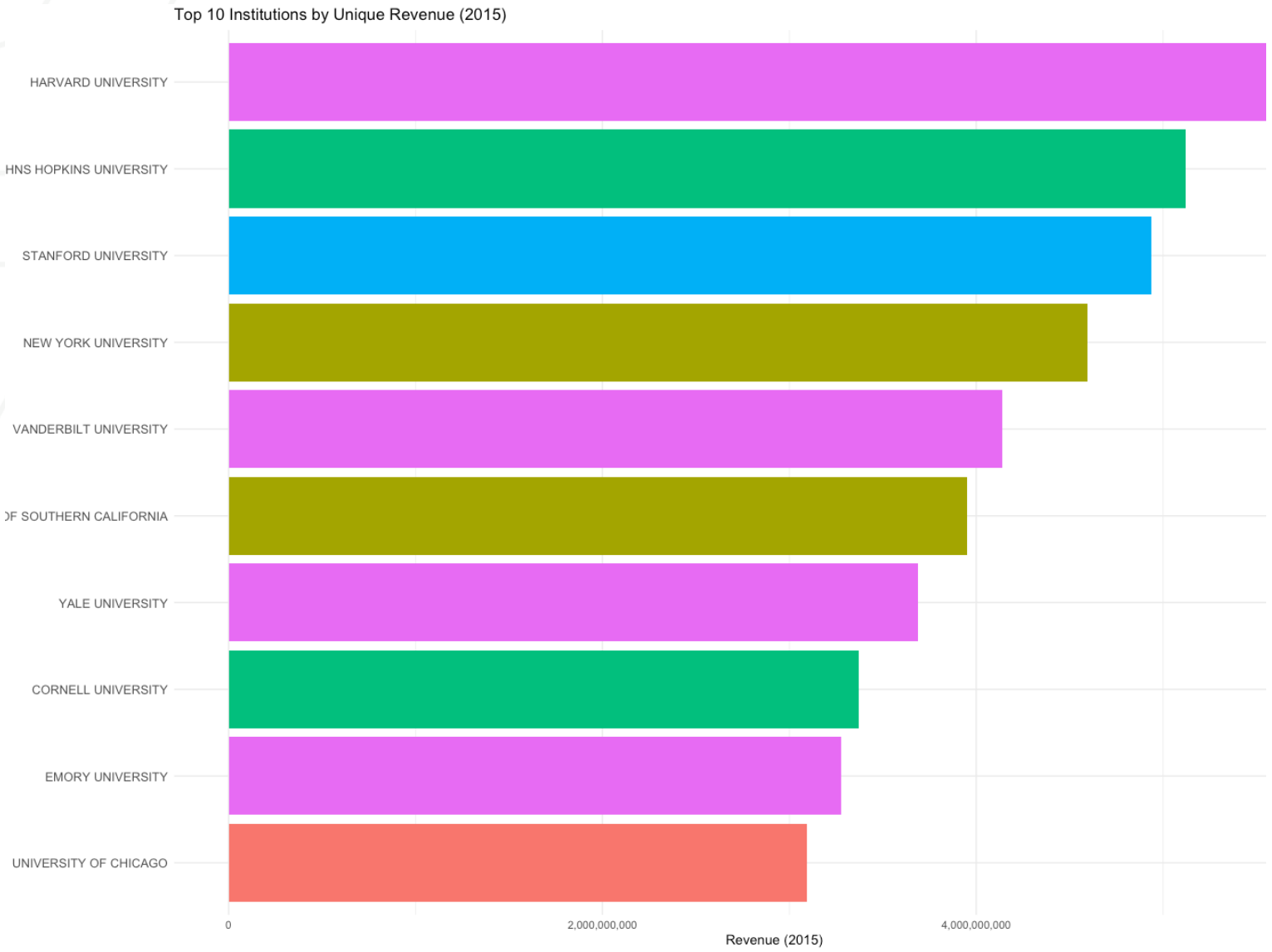


Main Observations:

- There is a strong positive linear relationship between Income and Revenue, suggesting that museums with higher income tend to have higher revenue.
- A few outliers diverge from the trend, showing either extremely high income or revenue that doesn't align with the general pattern.
- The clustering of points along the line reflects a stable relationship, with most museums fitting the overall trend, suggesting predictable financial behavior between income and revenue.

Georgia Tech

Top Institutions by Revenue

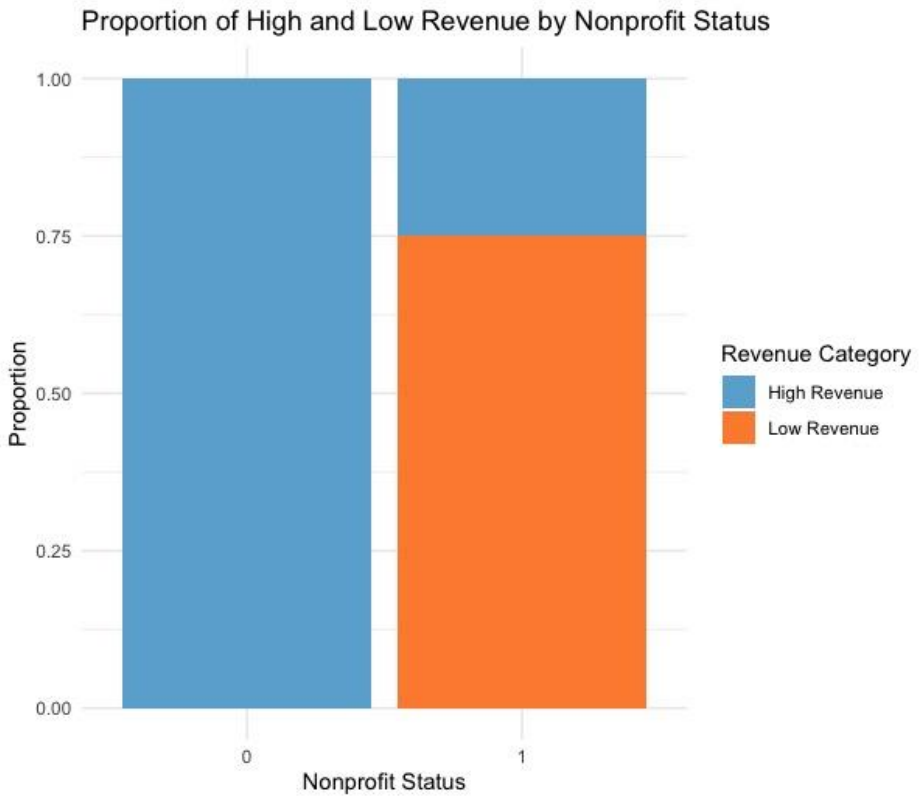


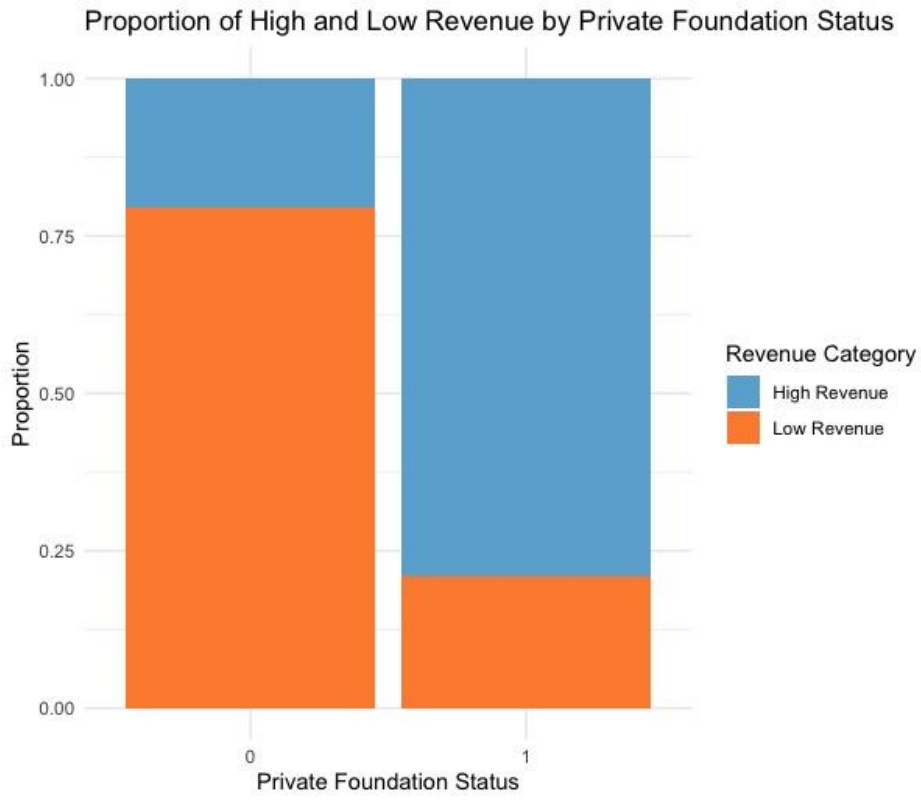
Main Observations:

- The top 10 institutions include both universities and standalone museums.
- ‘Institution' refers to museums and universities with associated collections or exhibits. This includes prominent universities that generate substantial revenue from their affiliated museum activities and dedicated museums like botanical gardens and art collections.
- Harvard, Johns Hopkins, and Stanford Universities lead in revenue, showing that the scale of the operation is tied to educational and cultural outreach.

Georgia Tech

Revenue by Nonprofit and Private Foundation Status



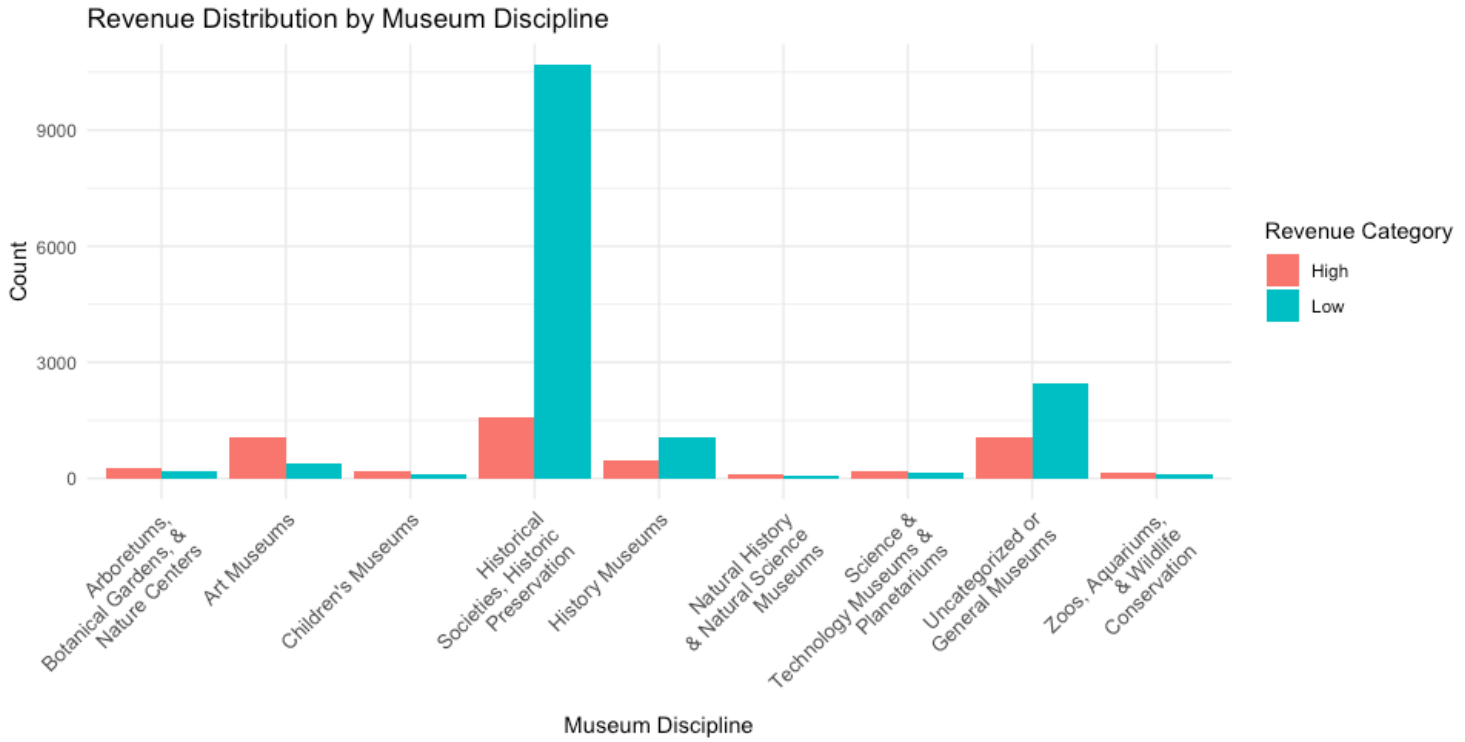


Main Observations:

- Nonprofit museums have a much higher proportion of low revenue compared to for-profit ones
- Museums that are private foundations generally tend to generate high revenue more consistently than those that are not.

Georgia Tech

Revenue by Museum Discipline

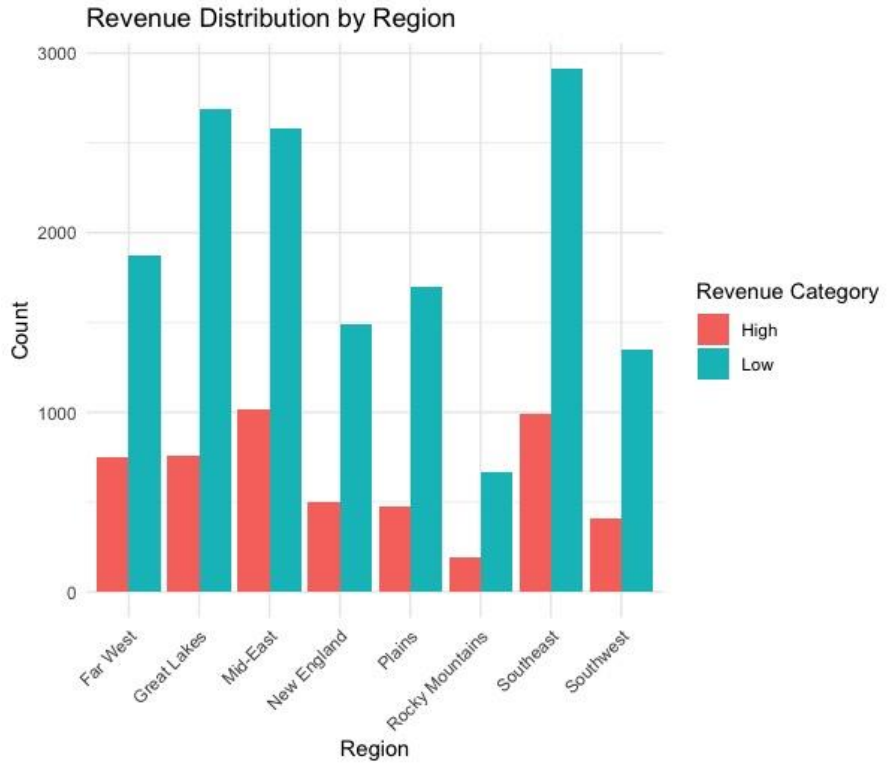


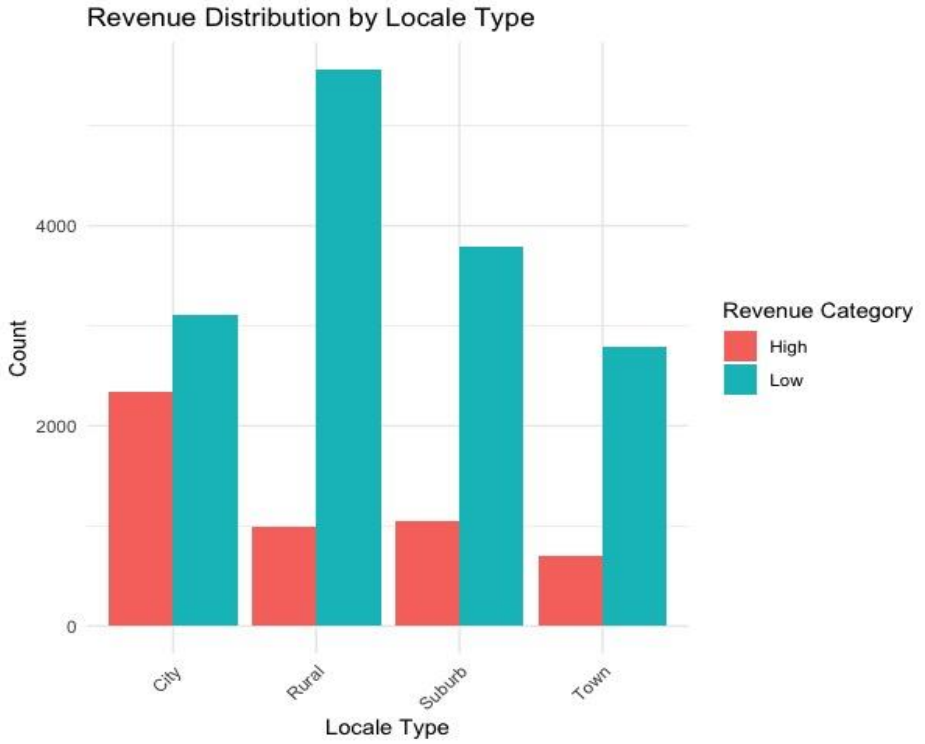
Main Observations:

- Art Museums, Historical Societies/Historic Preservation, History Museums, and Uncategorized or General Museums have the highest number of high-revenue museums.
- Historical Societies/Historic Preservation have the highest count in the low revenue category, which suggests that while they are numerous, most are not high earners.
- The data reflects concentration within specific disciplines, with high-revenue museums primarily found in fewer disciplines (e.g., Art, Historical Societies, and General Museums), indicating disparities in funding.
- Despite these exceptions, most museum disciplines predominantly fall into the low revenue category, highlighting that only a small portion of museums generate significant income.

Georgia Tech

Revenue by Region and Locale Type



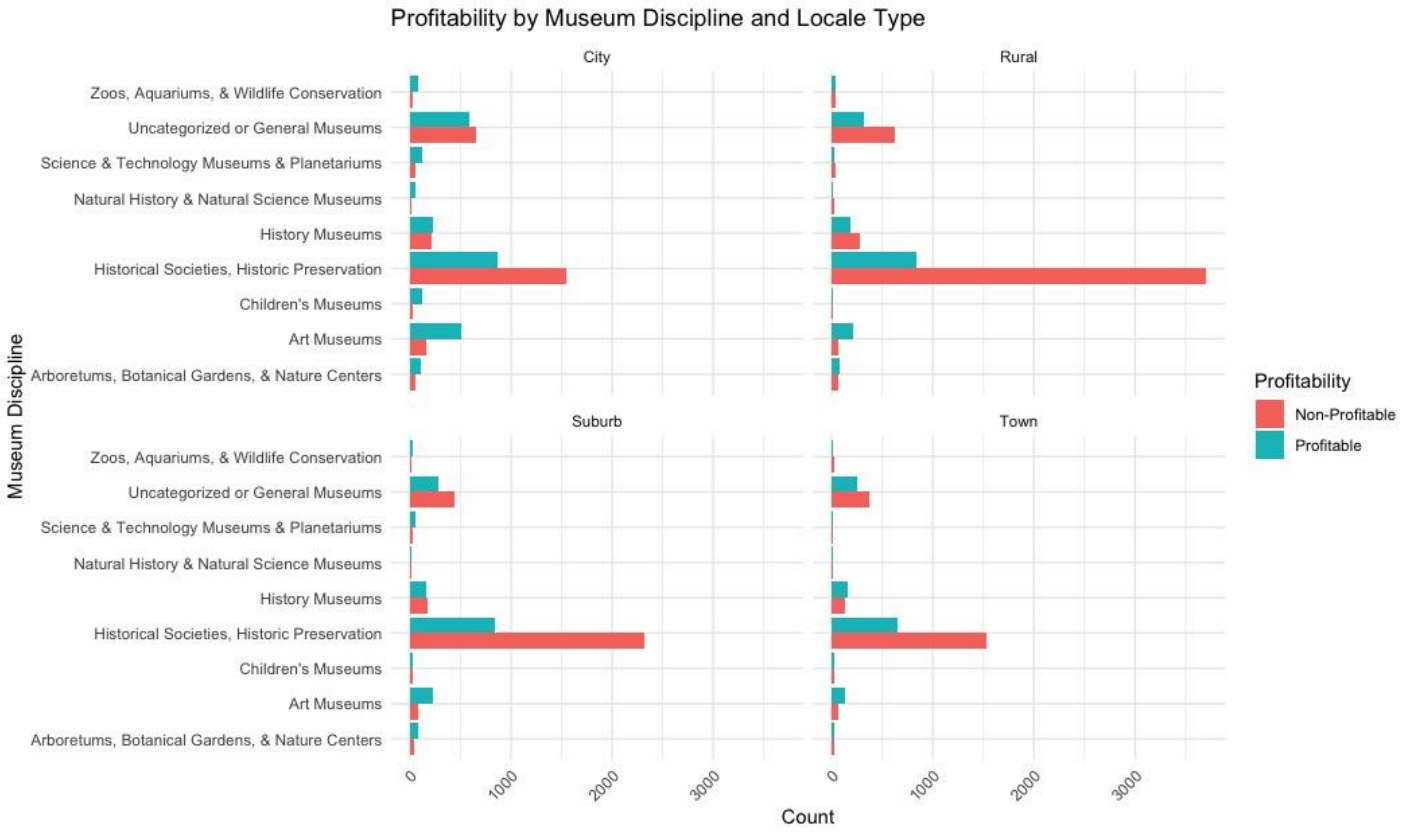


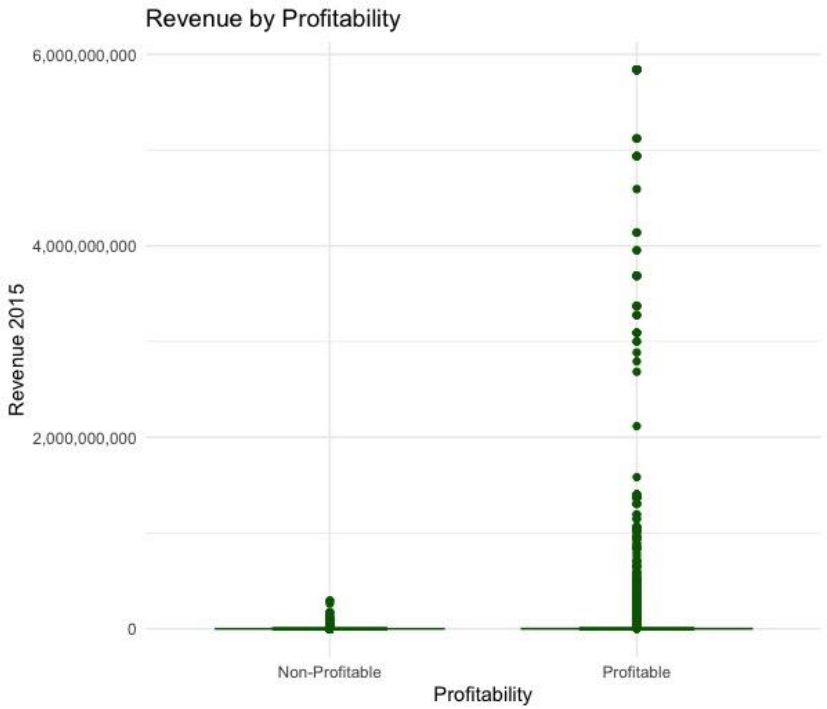
Main Observations:

- Museums in the Mid-East, Great Lakes, and Southeast regions show the highest counts, especially in the low-revenue category. The Mid-East also shows a considerable number of high-revenue museums.
- Rural areas have the highest count of museums, with a majority falling into the low-revenue category. In contrast, cities have a more balanced distribution between high and low revenue.
- Towns have fewer high-revenue museums compared to the other locale types.
- Across both regions and locales, most museums tend to fall into the low-revenue category, emphasizing a need for targeted support to increase museum profitability in underperforming areas.

Georgia Tech

Museum Profitability Overview



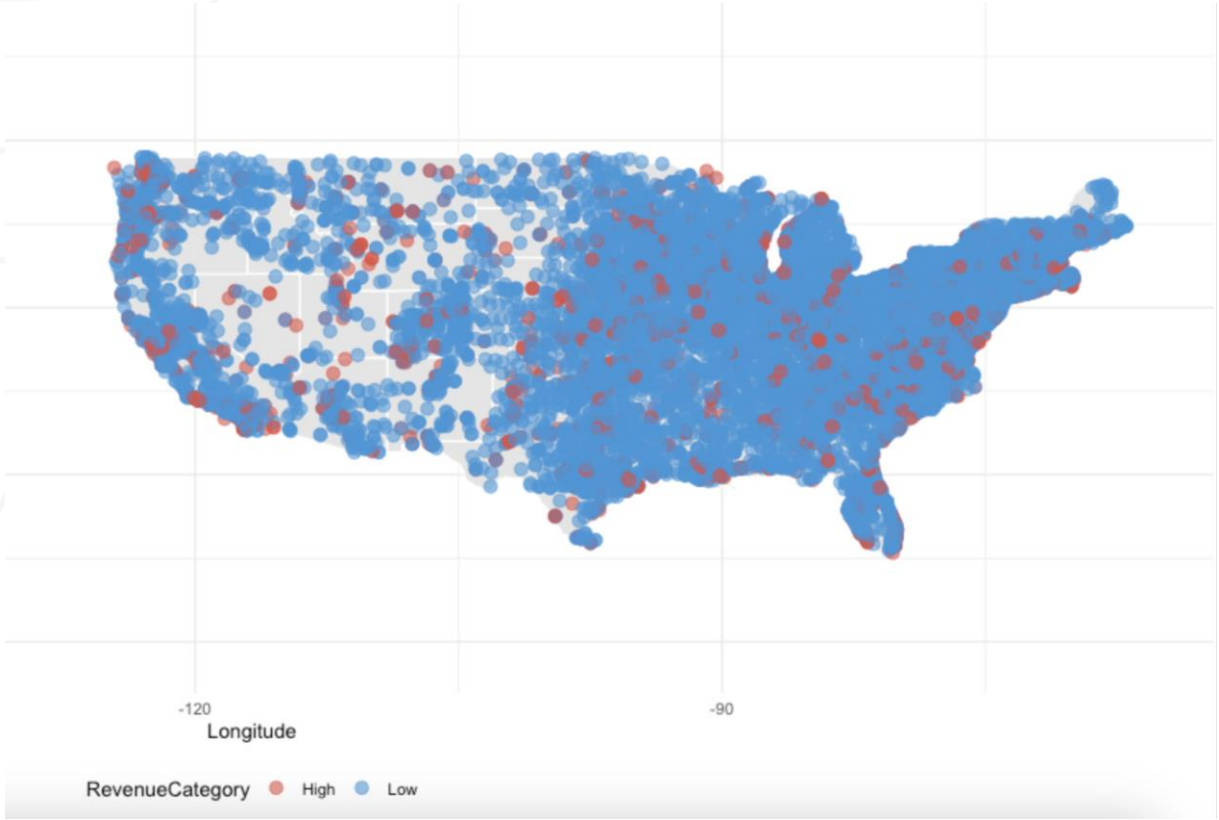


Main Observations:

- Most museums are non-profitable. This suggests that most museums face challenges in generating a profit.
- Rural museums, especially those under Historical Societies and Historic Preservation, have more non-profit institutions than other disciplines.
- The city has more profitable museums than non-profitable Museums in history, Art, and Children's museums.
- Profitable museums show a significant spread in revenue values, with some outliers showing extremely high revenue—suggesting that a select few museums might be driving the financial success of the category.

Georgia Tech

Geographic Distribution of Museums



Main Observations:

- Museums with high revenue appear more clustered in specific urban and densely populated areas, especially along the East Coast, Midwest, and West Coast regions. This could suggest higher financial sustainability in regions with larger populations and tourist attractions.
- Some regions, particularly in more rural and less populated areas of the western U.S., exhibit fewer museums overall, with those present tending to have lower revenue. This shows possible gaps in accessibility to cultural institutions based on geography.
- While museums with higher revenue are distributed across both urban and suburban areas, low-revenue museums also have a significant presence in these regions.

Georgia Tech

Key Takeaways from the EDA Section

Revenue Drivers

- Museums in densely populated areas (e.g., East Coast, Midwest, and West Coast) have a higher concentration of high-revenue museums.
- Art Museums, Historical Societies, and History Museums are linked with higher revenue, particularly in urban settings.
- Museums that are private foundations generate higher revenue.

Outlier Impact

- Outliers in both income and revenue are retained as they represent real data that is important for understanding variability.
- They show that most museums generate low revenue while a small number of institutions exhibit disproportionately high revenue, contributing to a heavily right-skewed distribution.

Profitability Trends

- Museums in urban settings (cities) show a more balanced profitability distribution, with art and children's museums being more profitable.
- Nonprofit museums have a higher proportion of low, while private foundations show better financial performance.

Geographic Distribution

- High-revenue museums tend to cluster in urban and high-tourism regions.

Data Preparation for Modeling Overview

Convert Categorical Variables to Factors

- Converted variables like Revenue Category, Museum Discipline, Region, etc., to factors.

Consolidate Low-Frequency Levels into 'Other'

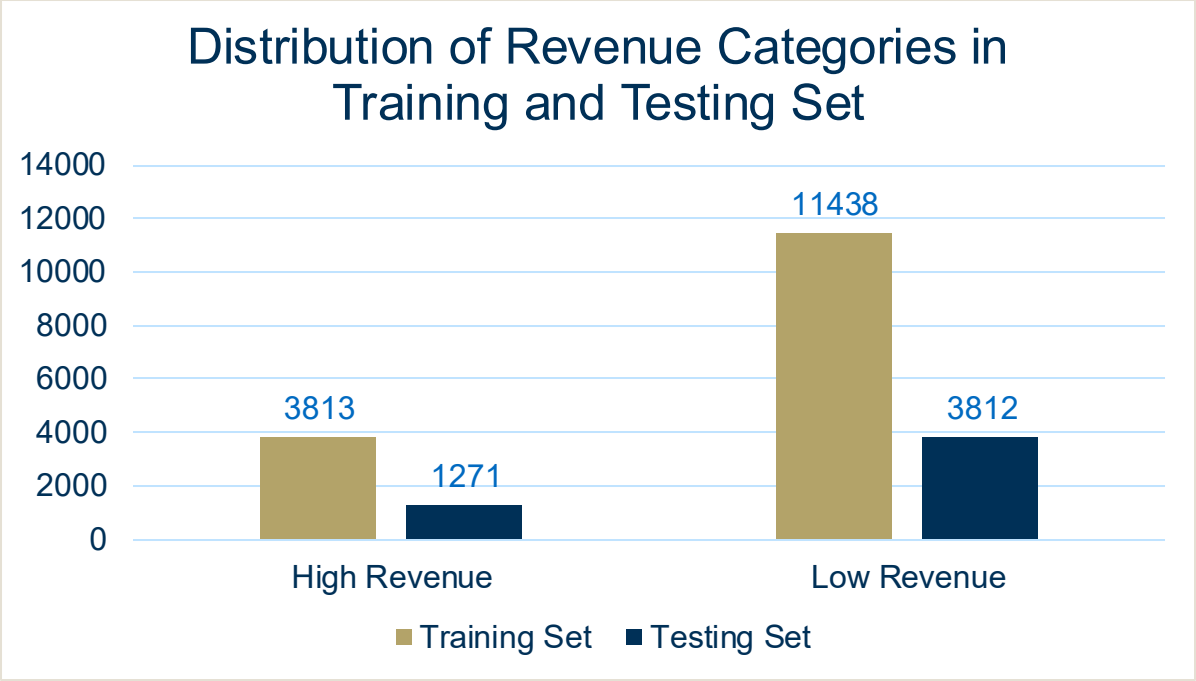
- Low-frequency levels across categorical variables like Museum Discipline and Locale Type were combined into an "Other" category to reduce sparsity.

Split Data into Training (75%) and Testing (25%) Sets

- Data split for model validation, with training used for fitting and testing for evaluation.

Align Factor Levels Between Training and Testing Sets

- This is to guarantee that factor levels in the testing set match those in the training set to prevent level mismatch errors.

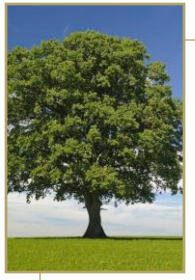


Georgia Tech

Modeling Approach Overview

Logistic Regression

- Estimates the probability of outcomes.
- Provides feature importance through odds ratios.
- Sets a benchmark for more complex models.



Random Forest

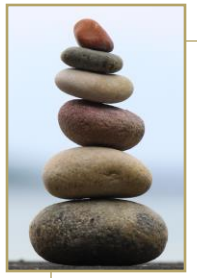
- Combines multiple decision trees.
- Identify prediction drivers.
- Handles complex patterns and interactions.

Ridge Regression

- Reduces overfitting by shrinking coefficients.
- Good for datasets with highly correlated features.
- Keeps all features while regularizing their impact.

Gradient Boosting

- Corrects errors from previous models to boost performance.
- Adjusts based on residuals to improve predictions.
- Good for nonlinear relationships.



Naive Bayes

- Works well with high-dimensional data.
- Quick to implement.
- Outputs probability distributions for interpretability.

Georgia Tech

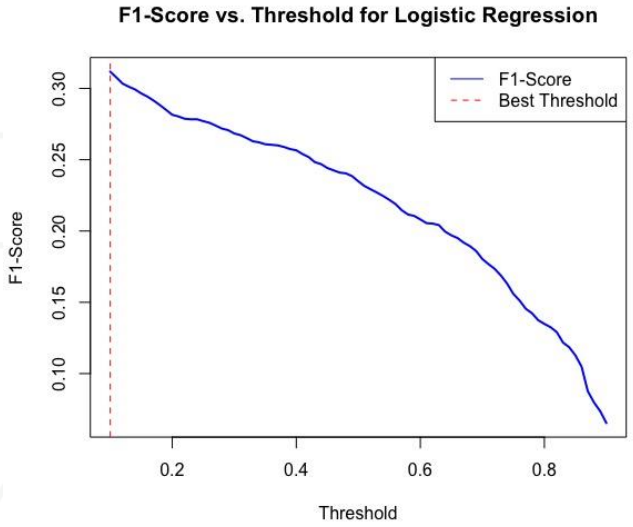
Museum Revenue Classification Model

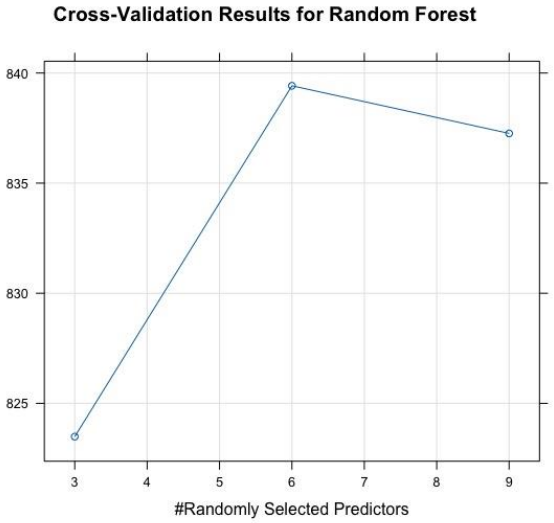
Objective - The goals is to predict whether a museum fall into High or Low revenue categories.

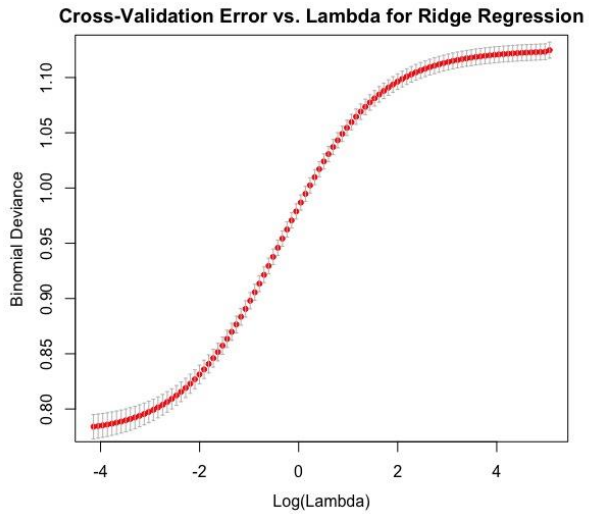
Model	Methodology	Hyperparameters Tuned
Logistic Regression	Generalized Linear Model	• Optimized classification threshold (0.1 to 0.9) to maximize F1-Score. • Evaluated F1-Score across thresholds to balance precision and recall.
Random Forest	Ensemble Trees	• Tuned `mtry` (number of features to consider at each split) by performing a grid search over values 3, 6, and 9. • `ntree` (Number of trees) set to 300. • Cross-validation (5-fold) to select the best `mtry` value.
Ridge Regression	Regularized Linear Model	• `lambda` (penalty term for regularization) was optimized using 10-fold cross-validation to minimize the cross-validation error. • A range of lambda values was automatically tested to identify the best level of regularization, with the optimal lambda selected for final model fitting.
Gradient Boosting Machine (GBM)	Boosted Decision Trees	• `n.trees` (number of trees) set to 500. • `interaction.depth` (depth of trees) set to 3, with evaluation using Cross-Validation (5-fold). • `shrinkage` (learning rate) set to 0.01 for smoother convergence. • `cv.folds` (number of folds for internal cross-validation) set to 5.
Naive Bayes	Probabilistic Classifier	• Trained Naive Bayes model using default parameters without additional tuning. • Predictions made based on learned class probabilities. • Evaluated the model with ROC and AUC metrics.

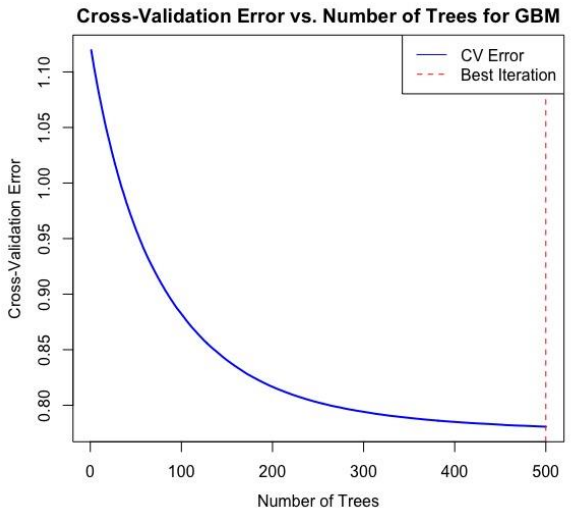
Georgia Tech

Museum Revenue Classification Model - Hyperparameter Tuning









- **Logistic Regression:** The F1-Score decreases as the threshold increases, indicating that stricter criteria for "High" result in more errors in balancing precision and recall. The best threshold maximizes the F1-Score.
- **Random Forest:** The accuracy increases as the number of predictors grows, peaking at 6 predictors. Beyond this point, the accuracy declines.
- **Ridge Regression:** The cross-validation error increases significantly as lambda grows, showing that too much regularization leads to underfitting.
- **GBM:** The cross-validation error decreases with more trees, stabilizing at 500 trees.

Georgia Tech

Museum Revenue Classification Model – Single Training Test Split

Models Performance and Evaluation Metrics

Metric	Definition	Ideal Scenario
Accuracy	Proportion of total predictions that are correct.	High Accuracy
AUC	Probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one.	Closer to 1
Precision	Proportion of positive predictions that are correct.	High Precision
Sensitivity	Proportion of actual positives that are correctly identified.	High Sensitivity
Specificity	Proportion of actual negatives that are correctly identified.	High Specificity
F1 Score	Harmonic mean of Precision and Recall, balancing both metrics.	High F1 Score

Georgia Tech

Museum Revenue Classification Model – Single Training Test Split

Models Performance Results

Metric	Logistic Regression	Random Forest	Ridge Regression	GBM	Naive Bayes
Accuracy	0.1898	0.7859	0.8426	0.8398	0.8356
AUC	0.1898	0.5554	0.8430	0.8435	0.8570
Precision	0.7384	0.7193	0.9709	0.9657	0.9373
Sensitivity	0.0070	0.8081	0.4577	0.4623	0.5305
Specificity	0.3130	0.6269	0.9025	0.9004	0.8953
F1 Score	0.8432	0.8265	0.8425	0.8469	0.8953

- **Ridge Regression** has the highest **accuracy**, **precision**, and **specificity**, making it ideal for minimizing false positives
- **Naive Bayes** has the best **AUC** and **F1 Score**, showing balanced performance.
- **GBM** performs well in both **accuracy** and **precision**..
- **Random Forest** has the highest **sensitivity**, making it reliable for identifying true positives.
- **Logistic Regression** performs poorly, indicating it's not suitable for this dataset.
- For a balanced approach, **Naive Bayes** and **GBM** are effective, while **Ridge Regression** is preferred for the highest **accuracy** and minimizing false positives.

Georgia Tech

Museum Revenue Classification Model – Performance Results

Single Training Test Split Performance Metrics

Metric	Logistic Regression	Random Forest	Ridge Regression	GBM	Naive Bayes
Accuracy	0.1898	0.7859	0.8426	0.8398	0.8356
AUC	0.1898	0.5554	0.8430	0.8435	0.8570
Precision	0.7384	0.7193	0.9709	0.9657	0.9373
Sensitivity	0.0070	0.8081	0.4577	0.4623	0.5305
Specificity	0.3130	0.6269	0.9025	0.9004	0.8953
F1 Score	0.8432	0.8265	0.8425	0.8469	0.8953

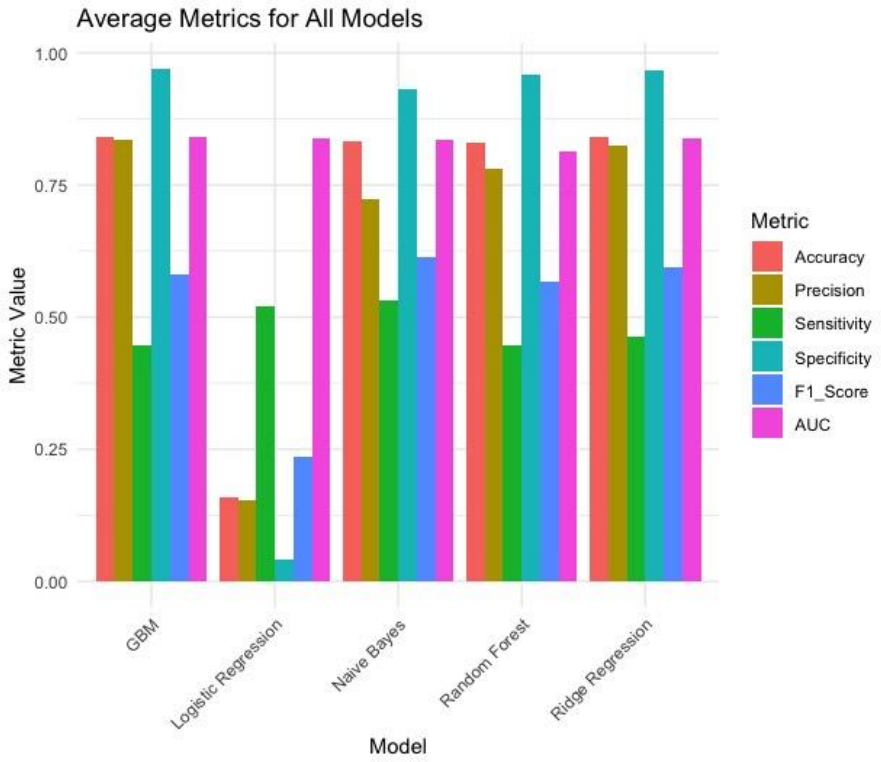
50 Iterations Cross-Validation Average Performance Metrics

Metric	Logistic Regression	Random Forest	Ridge Regression	GBM	Naive Bayes
Accuracy	0.1603	0.8298	0.8412	0.8396	0.8321
AUC	0.8384	0.8137	0.8375	0.8398	0.8354
Precision	0.1529	0.7797	0.8242	0.8354	0.7233
Sensitivity	0.5193	0.4453	0.4635	0.4466	0.5321
Specificity	0.0407	0.9580	0.9670	0.9707	0.9321
F1 Score	0.2362	0.5666	0.5932	0.5820	0.6131

To evaluate performance:

- The single training-test split approach involved training each model on 70% of the data and testing on the remaining 30%.
- Cross-validation was conducted with 50 iterations, using a 70%/30% random split each time.

Average Metrics for All Models – Cross Validation



Georgia Tech

Museum Revenue Classification Model – Model Selection

Ridge Regression

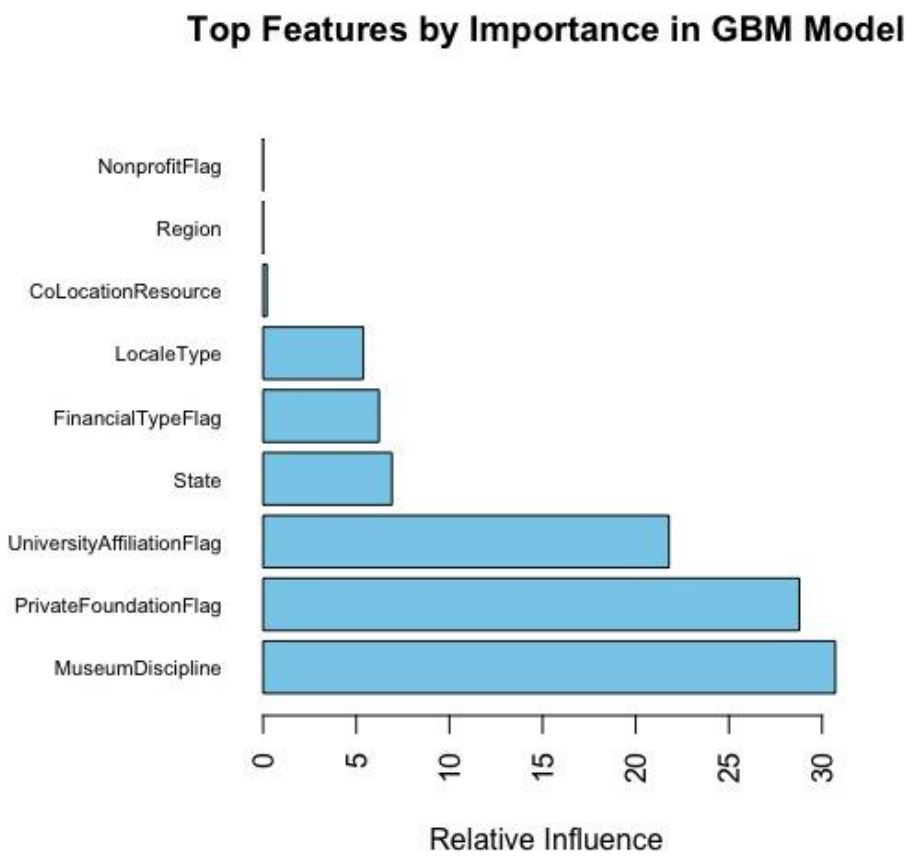
- Demonstrated consistently strong classification performance across both the single training-test split and cross-validation.
- Strong performance in terms of precision, specificity, and F1 Score, providing a reliable classification of revenue categories.
- Metrics were consistent between single test split and cross-validation, giving a high performance across different data splits.

Gradient Boosting Machine (GBM)

- Strong ability to differentiate between high and low revenue categories.
- It minimizes false positives, which is important for targeting true high-revenue outcomes. It offers a good balance between accuracy, precision, and F1 score, helping capture true high-revenue cases while minimizing false positives.

Georgia Tech

Museum Revenue Classification Model –Influential Variables in GBM



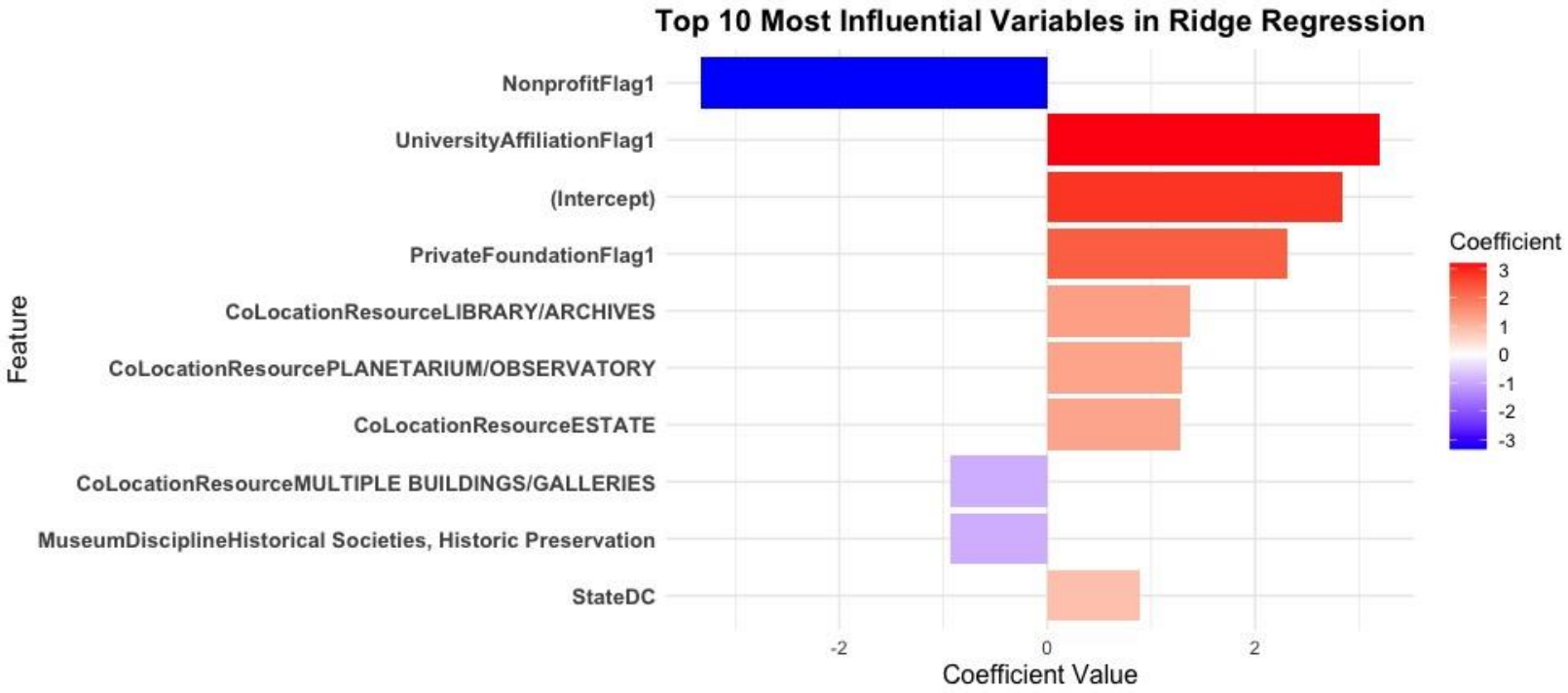
This chart represents the key features found using the GBM model, ranked by their relative importance.

Top Features Driving Predictions:

- **MuseumDiscipline:** The type of museum discipline is the most important feature, indicating that different types of museums (e.g., historical, art, or natural science) have a high impact on the revenue outcome.
- **PrivateFoundationFlag:** Museums affiliated with private foundations tend to perform better financially.
- **UniversityAffiliationFlag:** Museums affiliated with universities are associated with high revenue.
- **State:** Geographical location of the museum, as denoted by the state, plays a noticeable role.
- **FinancialTypeFlag:** Represents museums identified by Factual, a third-party data aggregator. This includes museums added through multiple data sources, highlighting different financial profiles that may impact revenue.
- **LocaleType: L**ocale type (urban, rural, or suburban) affects visitor demographics and, subsequently, revenue.

Georgia Tech

Museum Revenue Classification Model – Top 10 Influential Variables in Ridge Regression



- This chart represents the top 10 most influential variables in predicting museum revenue using Ridge Regression.
- **Positive Coefficients:** Associated with higher revenue.
- **Negative Coefficients:** Negatively associated with higher revenue.

Key Interpretations from Top Features:

- **NonprofitFlag1:** Museums that are nonprofit are strongly associated with **lower revenue**.
- **UniversityAffiliationFlag1:** Museums affiliated with universities are likely to have **higher revenue**.
- **PrivateFoundationFlag1**: Positive correlation with high revenue suggests that museums with private foundation ties perform better financially.
- **CoLocationResource Variables**: Museums co-located with other resources, such as **LIBRARY/ARCHIVES**, **PLANETARIUM/OBSERVATORY, ESTATE OR GALLERIES**, tend to have **higher revenue**.
- **StateDC** (Washington DC): Museums located in Washington DC tend to generate **higher revenue**, which may be attributed to high tourism, the presence of many cultural institutions, and a concentration of historical attractions.

Georgia Tech

Additional – Profitability Classification Model – Objectives and Performance Results

Objective:
To explore museum profitability by building a classification model that predicts whether a museum is profitable or non-profitable.

Background and Motivation:
- In addition to predicting museum revenue categories, I wanted to further understand the financial health of museums by examining profitability.
- To do this, I created a new column called **Profitability**, which defined whether income minus revenue results in a positive value (classified as profitable) or not (Non-profitable).

Single Training Test Split Performance Metrics

Metric	Logistic Regression	Random Forest	Ridge Regression	Naive Bayes
Accuracy	0.6591	0.6640	0.6586	0.6626
Precision	0.7306	0.7348	0.7293	0.7362
Sensitivity	0.8945	0.8874	0.9025	0.8694
Specificity	0.4238	0.4405	0.4146	0.4557
F1 Score	0.8043	0.8039	0.8067	0.7973

50 Iterations Cross-Validation Average Performance Metrics

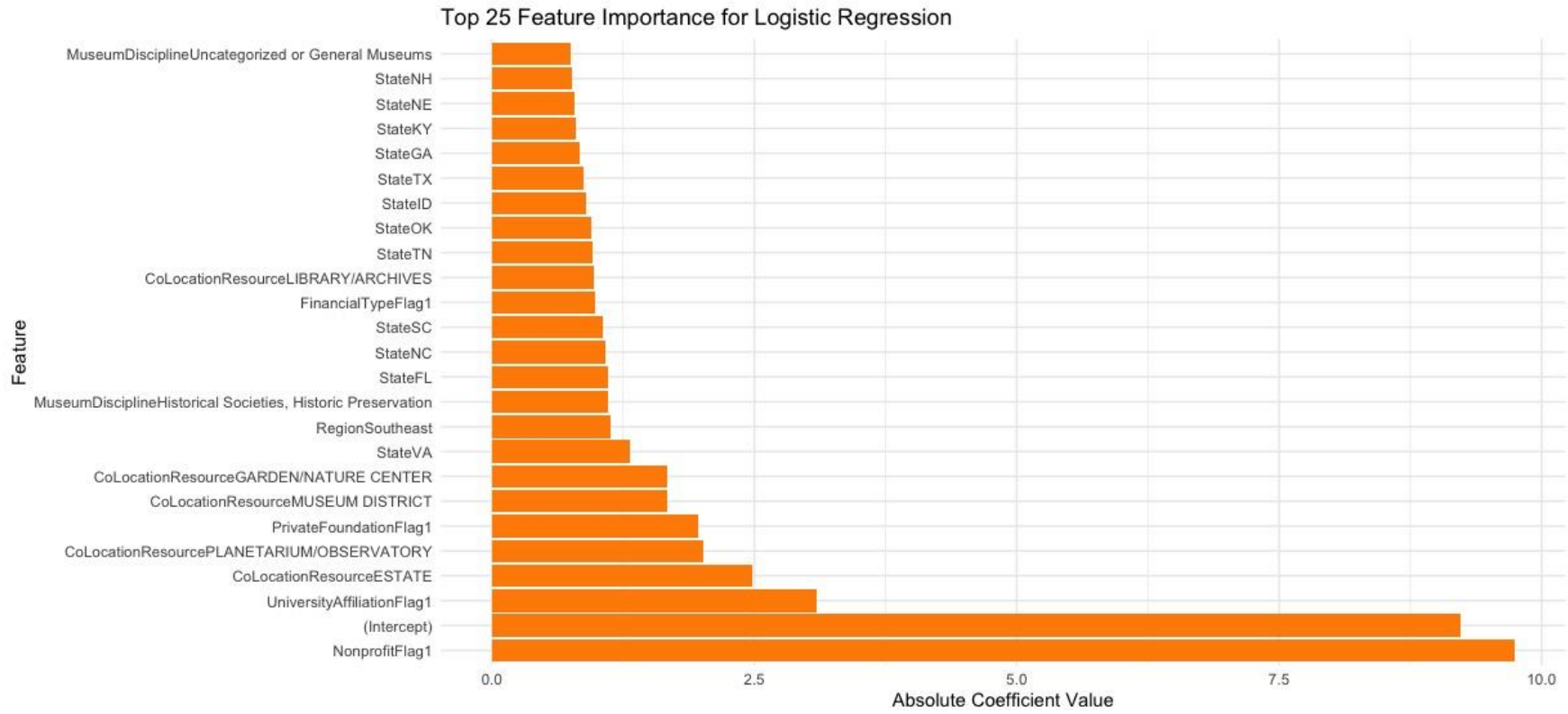
Metric	Logistic Regression	Random Forest	Ridge Regression	Naive Bayes
Accuracy	0.6763	0.6383	0.5807	0.5654
Precision	0.7411	0.7118	0.6771	0.6693
Sensitivity	0.9043	0.9443	0.9651	0.9571
Specificity	0.4484	0.3322	0.1962	0.1739
F1 Score	0.8147	0.8117	0.7959	0.7877

- Logistic Regression shows the highest Accuracy, Specificity, and F1 Score during cross-validation.
- Ridge Regression achieves the highest Sensitivity but poor Specificity, resulting in many false positives.

Georgia Tech

Additional – Profitability Classification Model - Selection

Logistic Regression is selected for its overall balanced performance. It achieved the highest Accuracy, Specificity, and F1 Score during cross-validation.



Interpretation of Logistic Regression Results:

- **Nonprofit Status and Affiliation**: Museums classified as nonprofits and those affiliated with universities show much higher odds of profitability. The odds ratio for nonprofits and university-affiliated museums indicates a strong positive impact on profitability, suggesting that institutional support and non-profit advantages contribute to financial performance.
- **Co-Location Resources Impact**: Museums co-located with specific resources such as **ESTATE** and **PLANETARIUM/OBSERVATORY** show high profitability odds, which means that shared resources or specialized attractions improve a museum's financial sustainability.
- **State-Level Influence on Profitability**: Museums located in states such as **Virginia, Florida, North Carolina, and South Carolina** exhibit higher profitability, as indicated by their odds ratios. This may reflect regional policies, funding support, or visitor interest.
- **Variation by Region**: Museums in the **New England** and **Southwest** regions also show higher profitability odds.

Georgia Tech