# Chapter 1: Business Case Study 1 - Ride-sharing Service Provider

## Problem Description

A prominent ride-sharing service provider, a leader in the country's intra-city logistics market, faces a critical challenge in driver retention. This company strives to improve the lives of its driver-partners by providing consistent earning opportunities and independence. Having served millions of customers, the company has established itself as a pivotal player in the urban mobility ecosystem.

The company's core challenge lies in the high churn rate among its drivers. Industry observers note that recruiting and retaining drivers is a tough battle for the company. Driver churn is high, and it's very easy for drivers to stop working for the service on short notice or switch to competitors based on rate fluctuations.

As the company grows, the high churn rate could become an even more significant problem. The frequent loss of drivers negatively impacts the organization's morale, and acquiring new drivers is substantially more expensive than retaining existing ones.

As a data scientist in the Analytics Department focused on driver team attrition, you are tasked with analyzing monthly information for a segment of drivers from 2019 and 2020. Your goal is to predict whether a driver will leave the company based on various attributes, including:

1. Demographics (city, age, gender, etc.)
2. Tenure information (joining date, last working date)
3. Historical data regarding the performance of the driver (quarterly rating, monthly business acquired, grade, income)

The dataset provided contains the following information for each driver:

|  |  |  |
| --- | --- | --- |
| Feature | Description | Data Type |
| MMMM-YY | Reporting Date (Monthly) | Date |
| Driver\_ID | Unique identifier for each driver | Integer |
| Age | Age of the driver | Integer |
| Gender | Gender of the driver (0: Male, 1: Female) | Binary |
| City | City code where the driver operates | Categorical |
| Education\_Level | Education level of the driver (0: 10+, 1: 12+, 2: Graduate) | Ordinal |
| Income | Monthly average income of the driver | Numeric |
| Date Of Joining | Date when the driver joined the platform | Date |
| LastWorkingDate | Last date of working for drivers who have left | Date |
| Joining Designation | Designation of the driver at the time of joining | Categorical |
| Grade | Current grade of the driver | Ordinal |
| Total Business Value | Total business value acquired by the driver in a month | Numeric |
| Quarterly Rating | Quarterly performance rating of the driver (1-5, higher is better) | Ordinal |

**Table 1.1: Data Dictionary for Driver Retention Analysis**

To address this challenge, we employed a comprehensive analytical approach:

1. **Data Preprocessing**: Handled missing values, converted date fields, and created derived features.
2. **Exploratory Data Analysis**: Examined distributions and relationships between various features and churn.
3. **Feature Engineering**: Created new features such as tenure, joining year and month, and performance trend indicators.
4. **Model Development**: Implemented and compared multiple ensemble learning techniques, including Decision Trees, Random Forests, and LightGBM.
5. **Model Evaluation**: Used metrics such as F1 Score, ROC-AUC, and confusion matrices to assess model performance.
6. **Feature Importance Analysis**: Identified the most critical factors influencing driver churn.
7. **Error Analysis**: Examined misclassified instances to understand model limitations and areas for improvement.

By developing an accurate driver churn prediction model, the company aims to:

1. Implement proactive retention measures for at-risk drivers
2. Optimize resource allocation in driver acquisition and retention efforts
3. Maintain a stable and experienced driver workforce to ensure consistent service quality

## Business Questions to be Answered from Analysis

To address the challenge of driver retention for the ride-sharing service provider, the following key questions were answered through our analysis:

1. **What are the primary factors influencing driver churn, and how do they impact the likelihood of a driver leaving the platform?**
   1. This question aimed to identify the most significant variables affecting driver retention, which guided feature selection for the predictive model and informed targeted retention strategies.
2. **How do demographic factors (age, gender, city) correlate with driver churn rates?**
   1. Understanding these correlations helped in identifying potential areas for tailored retention strategies for different driver segments.
3. **What is the relationship between a driver's tenure and their likelihood of churning?**
   * This analysis helped in identifying critical periods in a driver's lifecycle where they may be more prone to leaving, allowing for timely interventions.
4. **How do performance metrics (quarterly ratings, total business value) influence driver retention?**
   * This provided insights into the effectiveness of the company's performance evaluation system and its impact on driver satisfaction and retention.
5. **Is there a significant difference in churn rates based on the driver's initial joining designation or current grade?**
   1. This helped in understanding if career progression within the company impacts retention rates.
6. **How does a driver's income level and its fluctuations over time affect their likelihood of churning?**
   1. This analysis informed potential policies around compensation and incentives to improve driver retention.
7. **How effective are ensemble learning techniques, particularly Random Forest and Gradient Boosting methods, in predicting driver churn?**
   1. This question addressed the core technical challenge of the project and helped in
8. **What is the optimal set of hyperparameters for the ensemble models to achieve the best prediction accuracy?**
   1. This involved experimentation with different model configurations to find the most effective approach for this specific use case.
9. **How do different features rank in terms of their importance in predicting driver churn?**
   * This insight guided the identification of the most impactful factors for improving driver retention.

These questions were crucial for several reasons:

* **Strategic Decision Making**: The answers inform high-level strategies for driver retention and acquisition.
* **Operational Efficiency**: Understanding churn factors can lead to more efficient resource allocation in retention efforts.
* **Personalized Retention Strategies**: Insights gained can help in developing targeted approaches for different driver segments.
* **Model Development**: The answers guided the development and refinement of the predictive models, ensuring they address the most relevant aspects of driver churn.

To address these questions, we employed the following analytical approach:

* **Data Preprocessing and Exploratory Data Analysis (EDA)**:
  + Handled missing values and outliers
  + Analyzed distributions of various features
  + Examined relationships between features and churn
* **Feature Engineering**:
  1. Created new features such as 'num\_months', 'DOJ\_Year', 'DOJ\_Month'
  2. Developed trend indicators like 'income\_increase\_flag', 'grade\_increase\_flag'
* **Model Development and Comparison**:
  + Implemented and compared Decision Tree, Random Forest, and LightGBM models
  + Used GridSearchCV for hyperparameter tuning
* **Model Evaluation**:
  + Utilized metrics such as F1 Score, ROC-AUC, and confusion matrices
  + Performed threshold optimization to improve model performance
* **Feature Importance Analysis**:
  + Analyzed feature importances from the best-performing model
  + Identified the most critical factors influencing driver churn
* **Error Analysis**:
  + Examined misclassified instances to understand model limitations
  + Identified areas for potential model improvement and additional data collection

## Analysis

### Analysis Part 1: Data Preprocessing and Initial Exploration

#### Task 1.1: Importing Libraries and Dataset

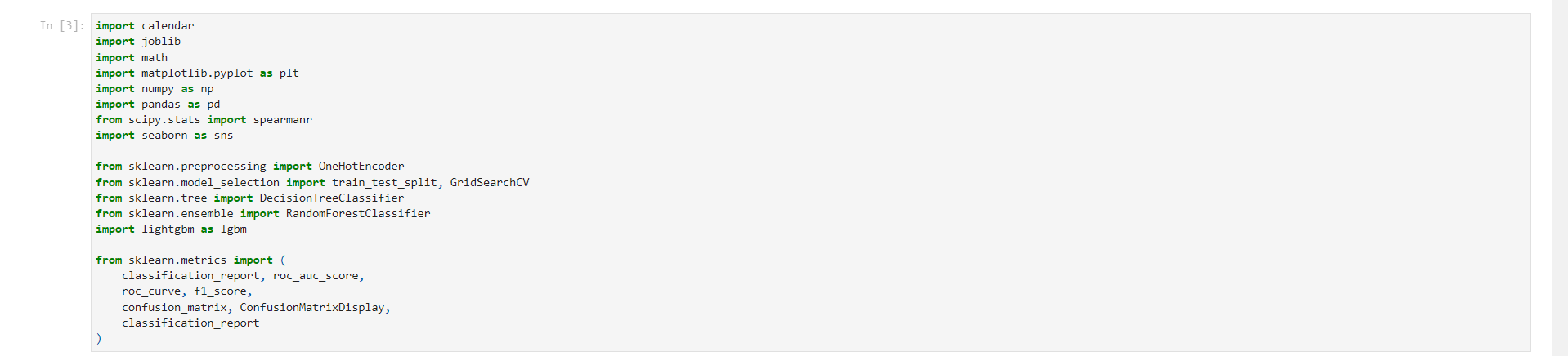
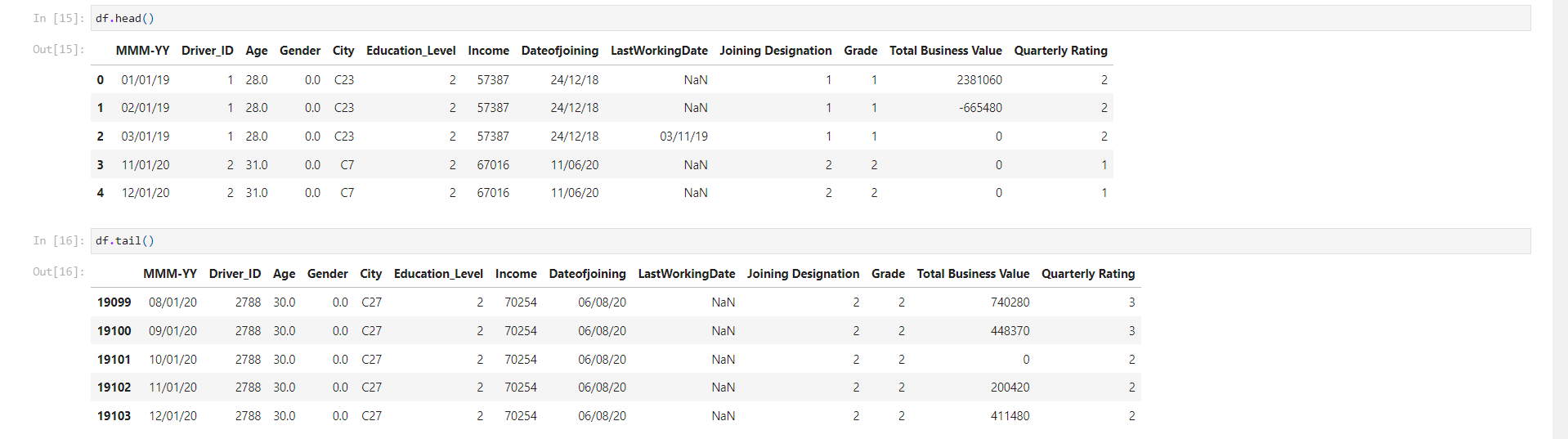
Figure 1.01: Code for Importing Essential Libraries

Figure 1.02: Code for Importing Dataset and Dropping Unnecessary Columns

Key Insights:

1. Essential libraries for data manipulation (pandas, numpy), visualization (matplotlib, seaborn), and machine learning (sklearn) are imported.
2. An unnecessary column 'Unnamed: 0' is dropped from the dataset.
3. A copy of the original dataset is created for potential future reference.

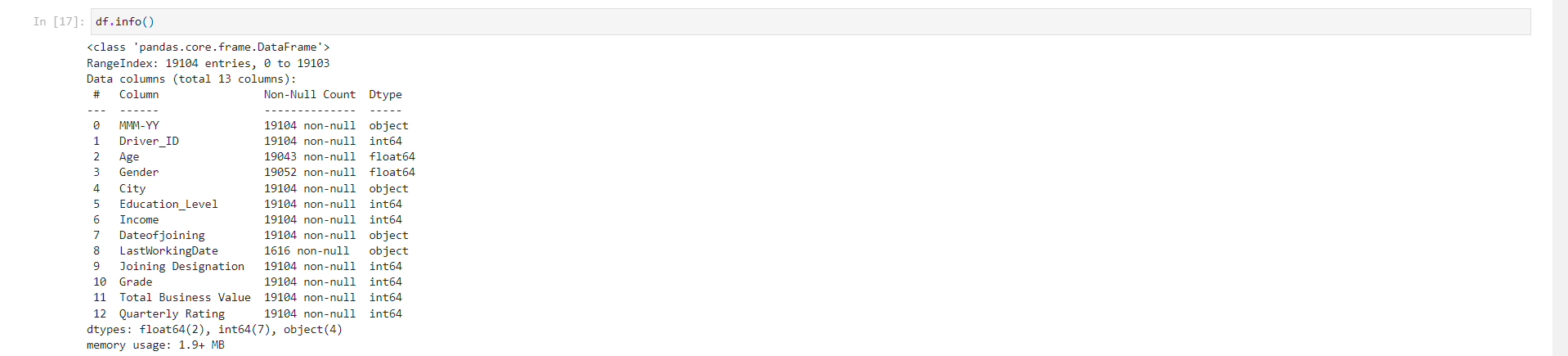
#### Task 1.2: Basic Data Exploration

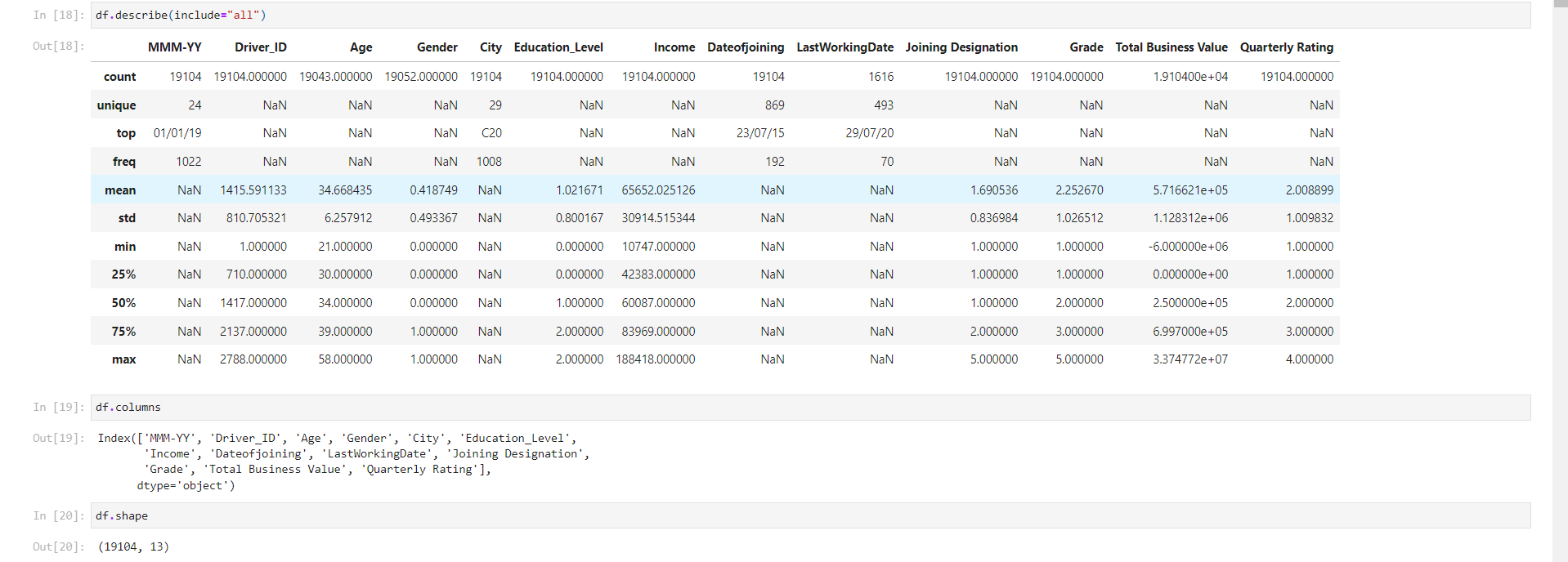
Figure 1.03: Table of Sample Rows from the Dataset

Key Insights:

* The dataset contains information about drivers, including their demographics, employment details, and performance metrics.
* Key features include Driver\_ID, Age, Gender, City, Education\_Level, Income, Dateofjoining, LastWorkingDate, Joining Designation, Grade, Total Business Value, and Quarterly Rating.
* The data spans multiple months, as indicated by the 'MMM-YY' column.

#### Task 1.3: Dataset Structure and Characteristics

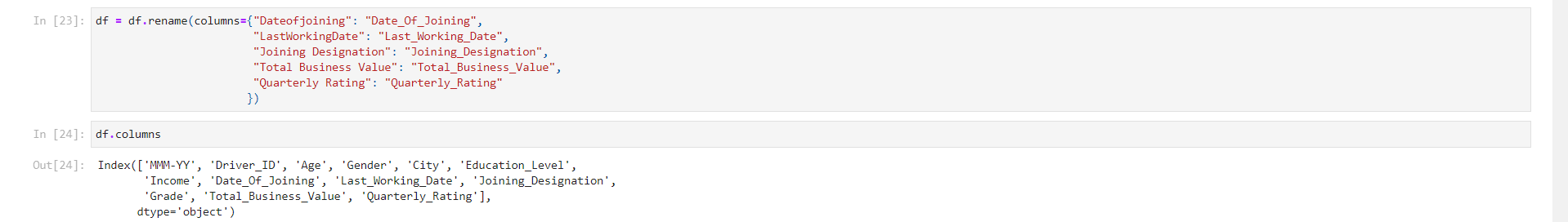
Figure 1.04: Table of Dataset Structure (1)

Figure 1.05: Table of Dataset Structure (2)

Key Insights:

1. The dataset contains 19,104 entries and 13 columns.
2. Data types include int64, float64, and object (for categorical and date variables).
3. Some columns have missing values, particularly 'LastWorkingDate' with only 1,616 non-null values.
4. The 'Driver\_ID' column ranges from 1 to 2,788, suggesting this many unique drivers in the dataset.
5. Numerical columns like 'Age', 'Income', and 'Total Business Value' show wide ranges, indicating diversity in the driver pool.
6. Categorical columns like 'City' have multiple unique values (29 unique cities).

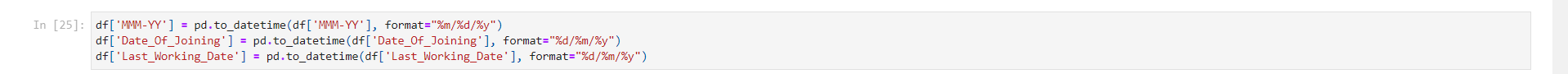
#### Task 1.4: Updating Column Names

Figure 1.06: Code for Standardizing Column Names

Key Insights:

* Column names are standardized for consistency and clarity.
* Spaces in column names are replaced with underscores.
* This renaming will make future data manipulation and analysis more straightforward.

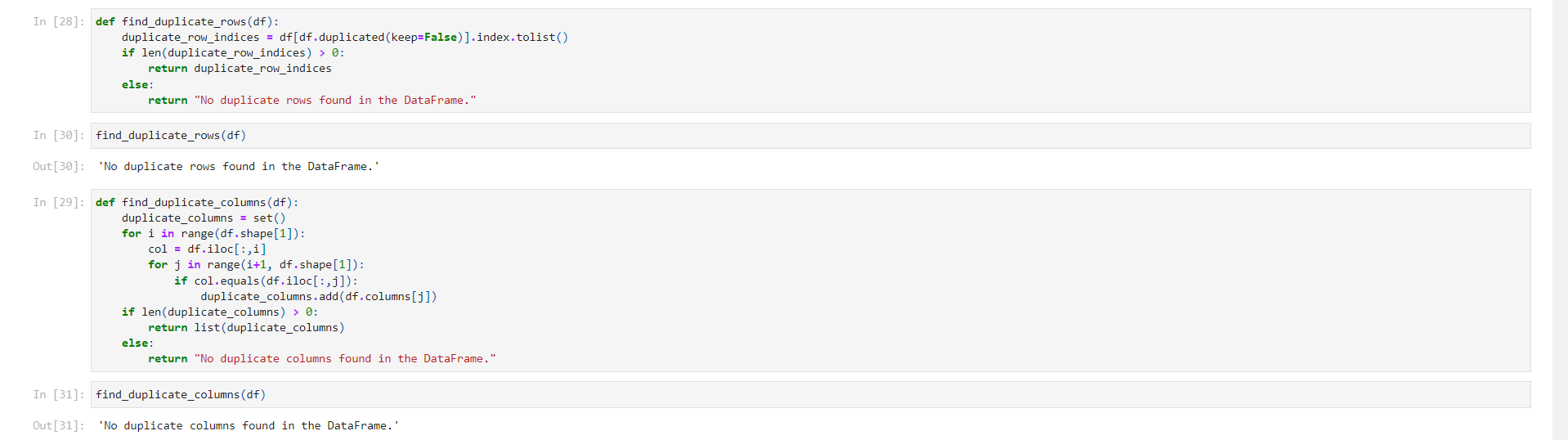
#### Task 1.5: Updating Data Types

Figure 1.07: Code for Converting Columns to DateTime Format

Key Insights:

* Date columns ('MMM-YY', 'Date\_Of\_Joining', 'Last\_Working\_Date') are converted to datetime format.
* This conversion will enable time-based analysis and feature engineering.
* Different date formats are handled appropriately for each column.

#### Task 1.6: Checking for Duplicate Values

Figure 1.08: Code for Checking Duplicate Rows and Columns

Key Insights:

* Custom functions are created to check for duplicate rows and columns in the dataset.
* No duplicate rows or columns were found in the dataset.
* This check ensures data integrity and prevents redundancy in the analysis.

### Analysis Part 2: Exploratory Data Analysis and Feature Engineering

#### Task 2.1: Missing Value Analysis

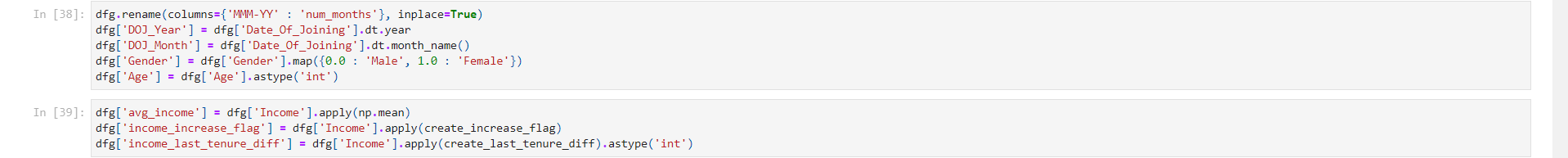
Figure 1.09: Code for Missing Value Analysis

Key Insights:

* The 'Last\_Working\_Date' column has the highest percentage of missing values at 91.54%.
* 'Age' and 'Gender' columns have a small percentage of missing values (0.32% and 0.27% respectively).
* The high percentage of missing values in 'Last\_Working\_Date' is expected as it represents drivers who are still active.

#### Task 2.2: Data Cleaning and Preprocessing

Figure 1.10: Code for Data Aggregation

Figure 1.11: Code for New Feature Engineering (1)

Key Insights:

* Data is aggregated by Driver\_ID to create a summary for each driver.
* New features are engineered:
  + 'num\_months': Number of months a driver has been with the company
  + 'DOJ\_Year' and 'DOJ\_Month': Year and month of joining
  + 'avg\_income': Average income of the driver
  + 'income\_increase\_flag': Indicates if the driver's income has increased over time
  + 'income\_last\_tenure\_diff': Difference in income between last two tenures
* Gender is mapped from numerical to categorical (Male/Female)
* Age is converted to integer type for easier analysis

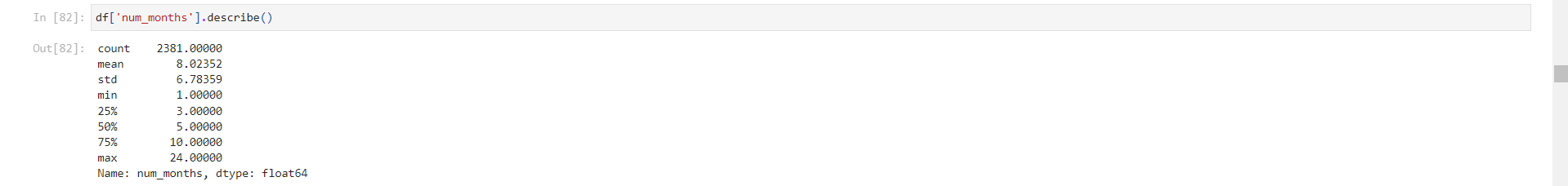
#### Task 2.3: Further Feature Engineering

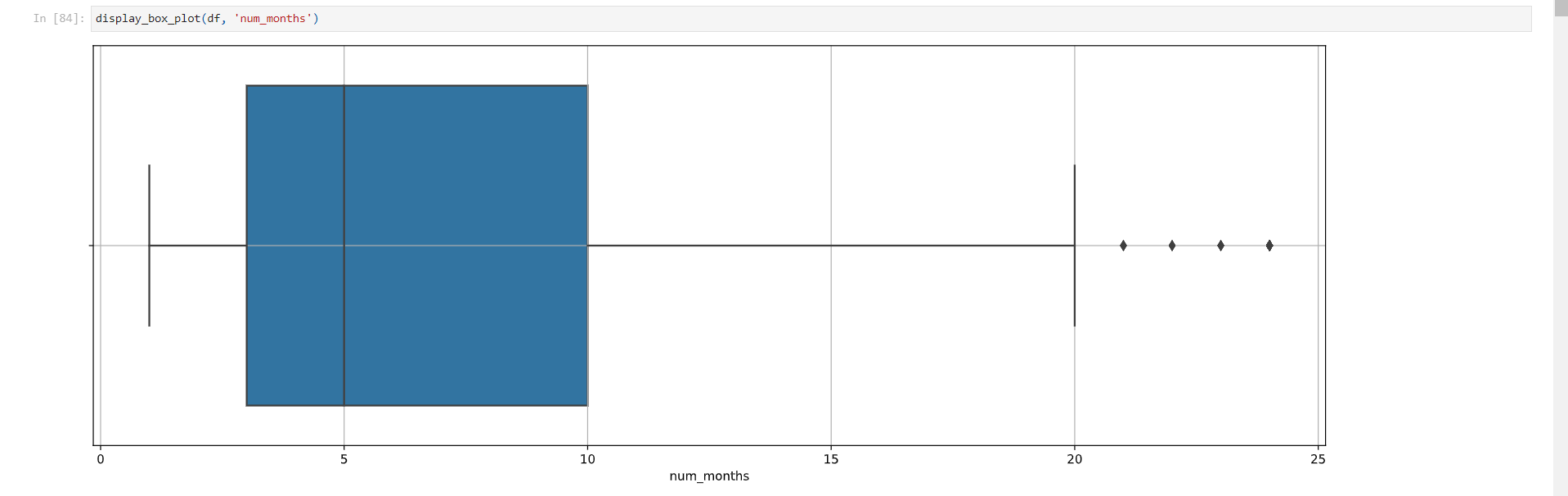
Key Insights:

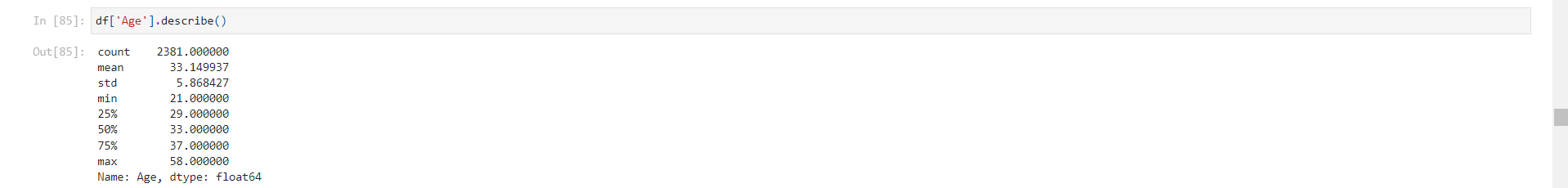
Figure 1.12: Code for New Feature Engineering (2)

* Additional features are created to capture trends and changes in driver performance:
  + Average grade and grade change indicators
  + Average total business value and its change indicators
  + Average quarterly rating and its change indicators
* An 'is\_promoted' feature is created to indicate if a driver has been promoted
* The target variable 'churned' is created, where 1 indicates the driver has left (has a Last\_Working\_Date) and 0 indicates they are still active.

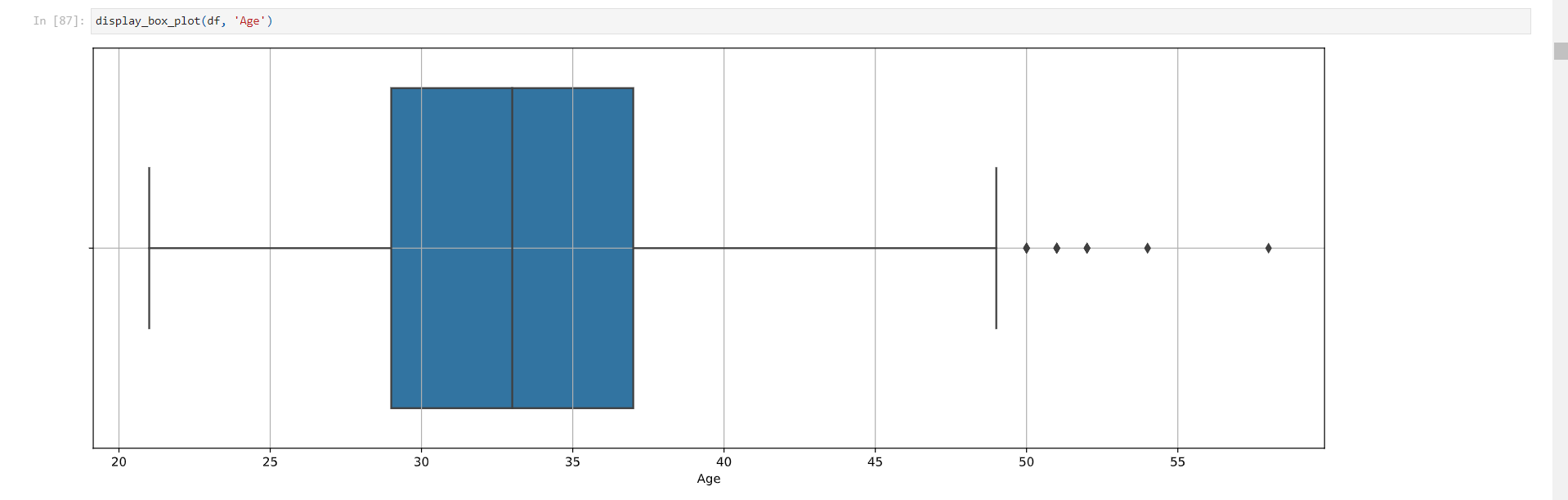
#### Task 2.4: Univariate Analysis - Numerical Variables

Figure 1.13: Table of Summary Statistics for the Column

Figure 1.14: Graph of Distribution of Number of Months with Company (Box Plot)

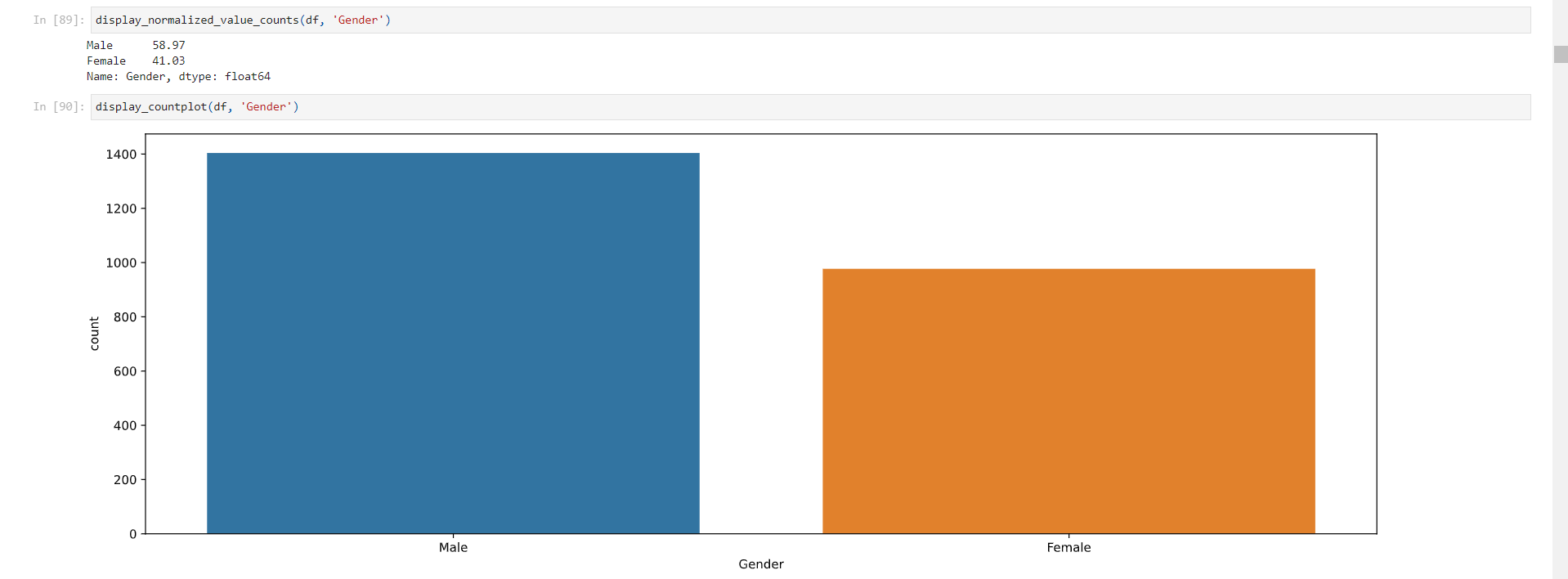
Figure 1.15: Table of Summary of Statistics for the Column

Key Insights:

Figure 1.16: Graph of Age Distribution of Drivers (Box Plot)

* Number of months:
  + Mean: 8.02, Median: 5.00
  + Right-skewed distribution with some drivers having very long tenures
  + 75% of drivers have been with the company for 10 months or less
* Age:
  + Mean: 33.15, Median: 33.00
  + Relatively symmetric distribution
  + Most drivers are between 29 and 37 years old
  + Some outliers on both ends, particularly older drivers

#### Task 2.5: Univariate Analysis - Categorical Variables

Figure 1.17: Graph of Gender Distribution (Count Plot)

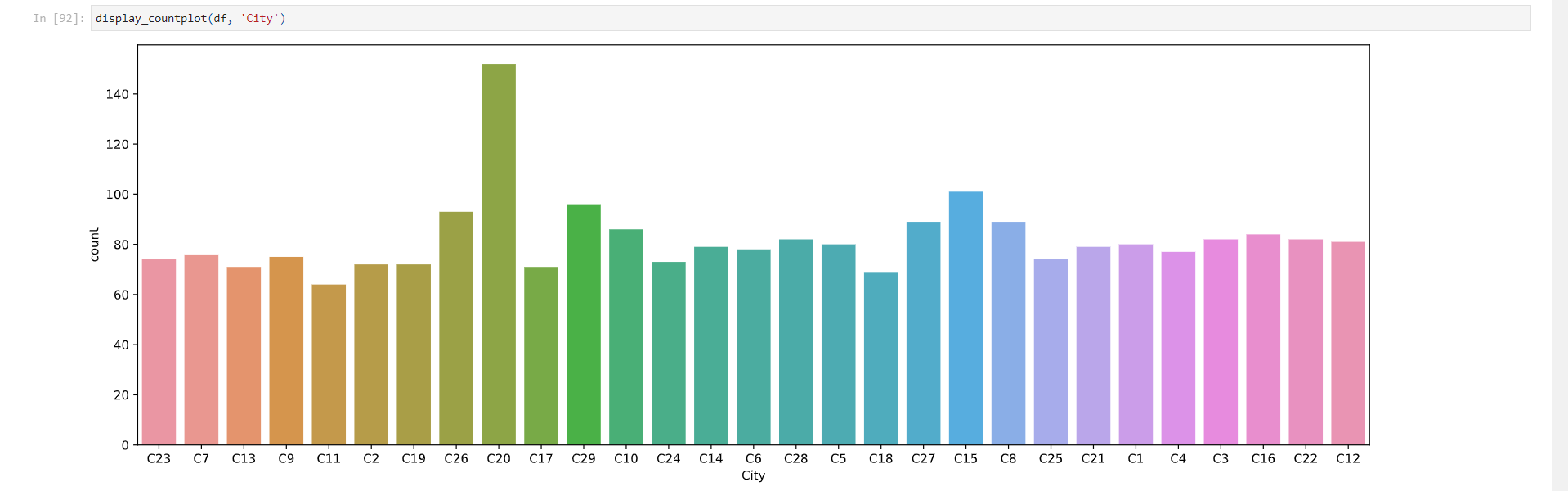
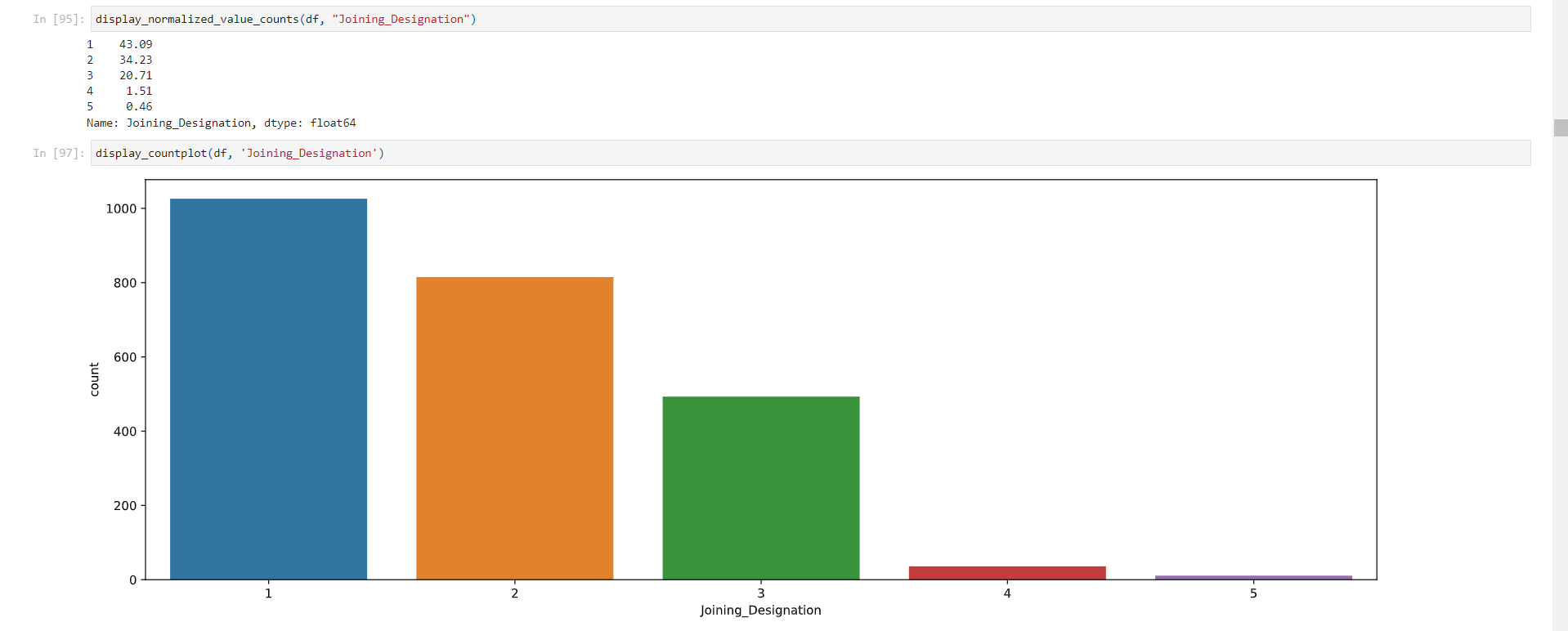
Figure 1.18: Graph of City Distribution (Count Plot)

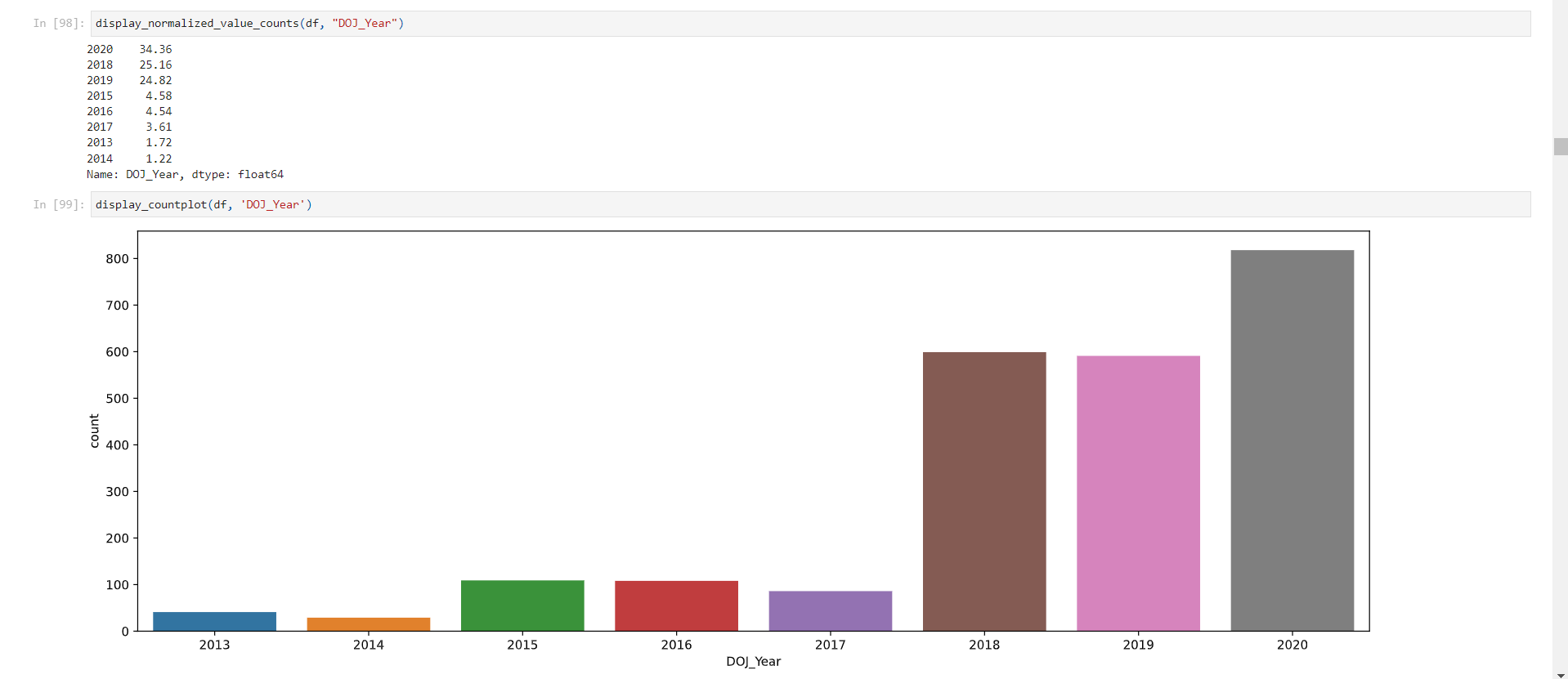
Figure 1.19: Graph of Education Level Distribution (Count Plot)

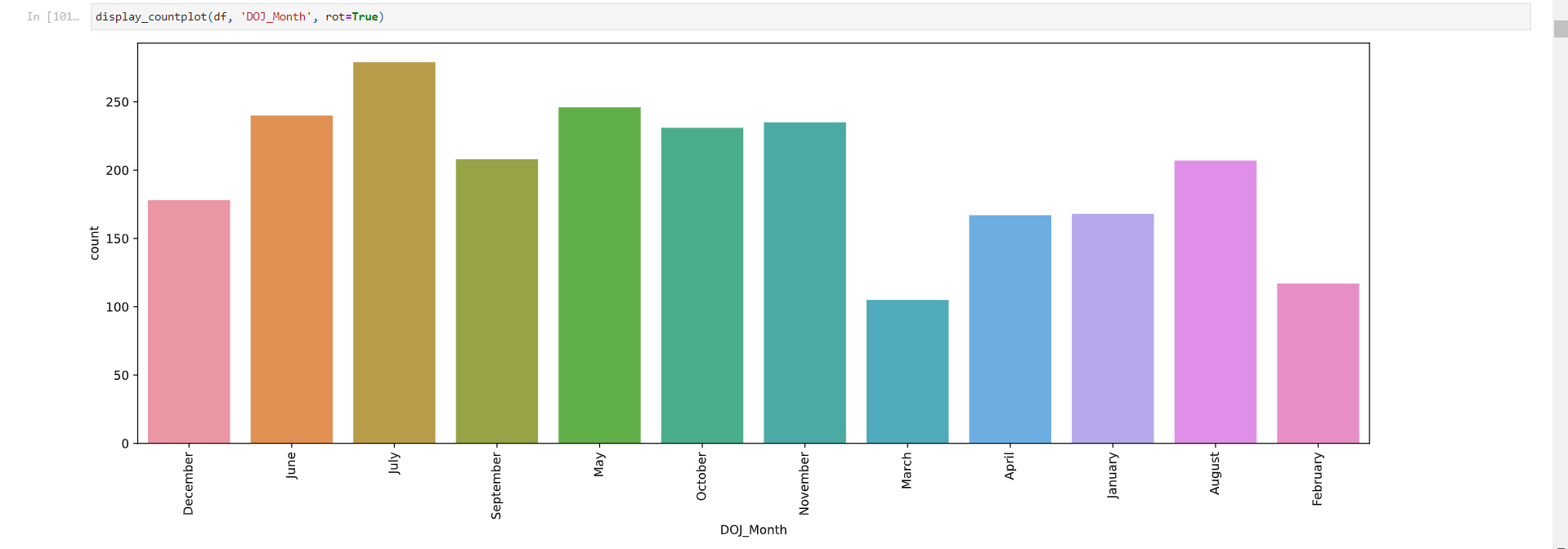
Key Insights:

* Gender:
  + 58.97% Male, 41.03% Female
  + Slight imbalance towards male drivers
* City:
  + 29 unique cities
  + Most populous city (C20) accounts for 6.38% of drivers
  + Fairly even distribution across cities, with slight variations
* Education Level:
  + Almost equal distribution across three levels (0, 1, 2)
  + Suggests a diverse educational background among drivers

#### Task 2.6: Univariate Analysis - Other Variables

Figure 1.20: Graph of Joining Designation Distribution (Count Plot)

Figure 1.21: Graph of Year of Joining Distribution (Count Plot)

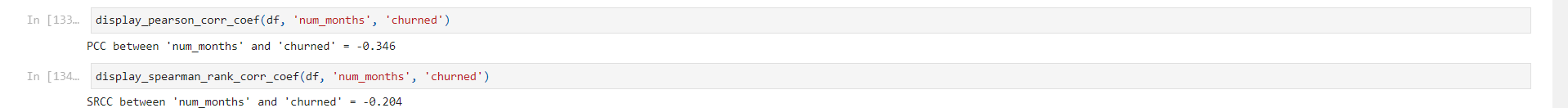
Figure 1.22: Graph of Month of Joining Distribution (Count Plot)

Key Insights:

* Joining Designation:
  + Majority (43.09%) join at designation 1
  + Only 1.97% join at higher designations (4 or 5)
* Year of Joining:
  + Most drivers (84.34%) joined between 2018-2020
  + Significant increase in driver recruitment in recent years
* Month of Joining:
  + Highest joining rates in July (11.72%) and May (10.33%)
  + Lowest in February (4.91%) and March (4.41%)
  + Suggests seasonal trends in driver recruitment

### Analysis Part 3: Bivariate Analysis and Data Preparation

#### Task 3.1: Bivariate Analysis - Numerical Variables vs Churn

Figure 1.23: Code For Pearson and Spearman Correlation

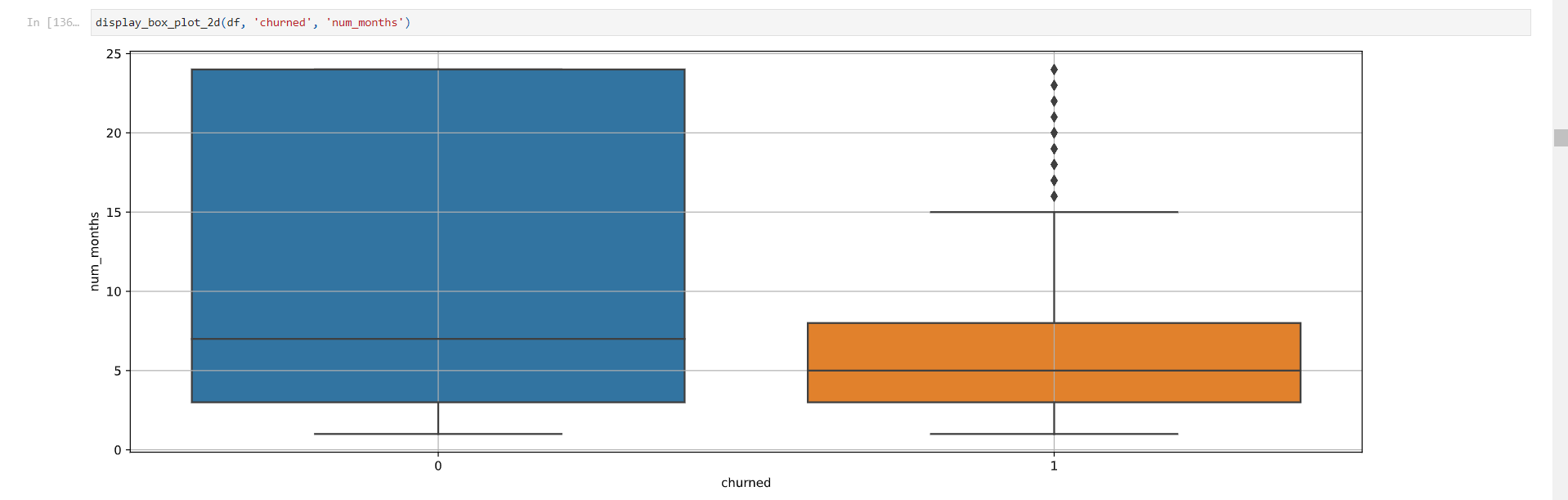
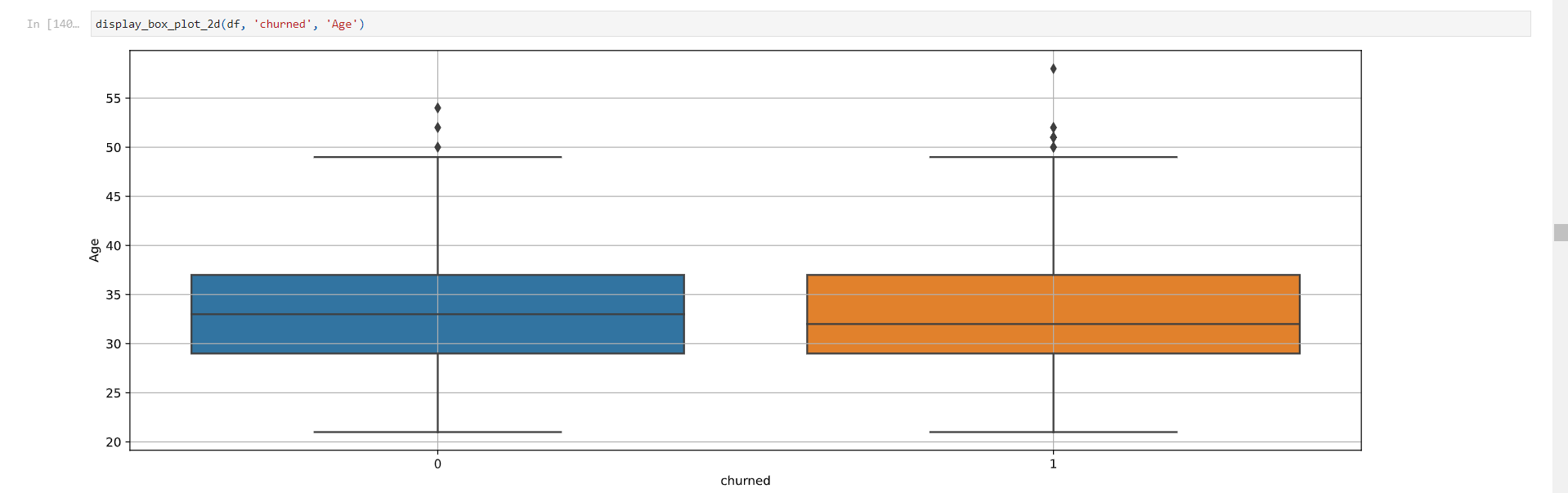
Figure 1.24: Graph of Number of Months vs Churn (Box Plot 2D)

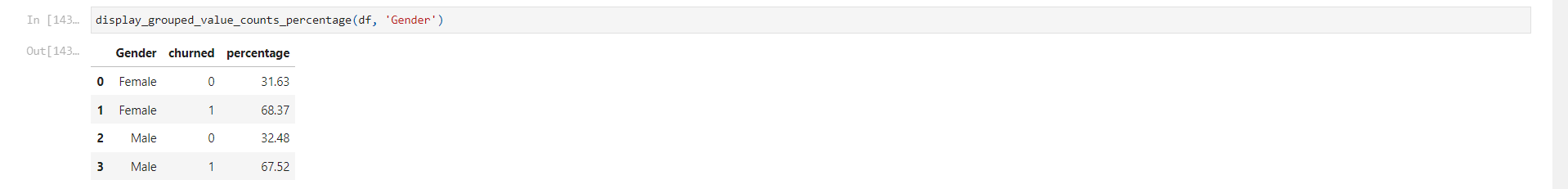
Figure 1.25: Code For Pearson and Spearman Correlation

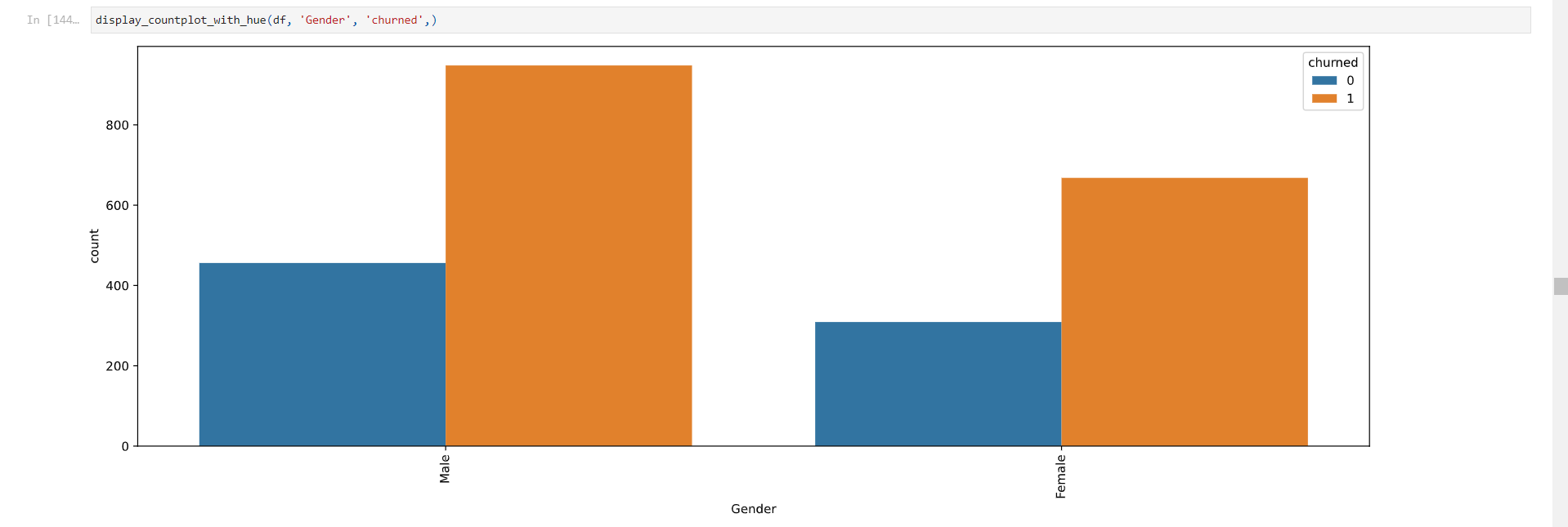
Figure 1.26: Graph of Age vs Churn (Box Plot 2D)

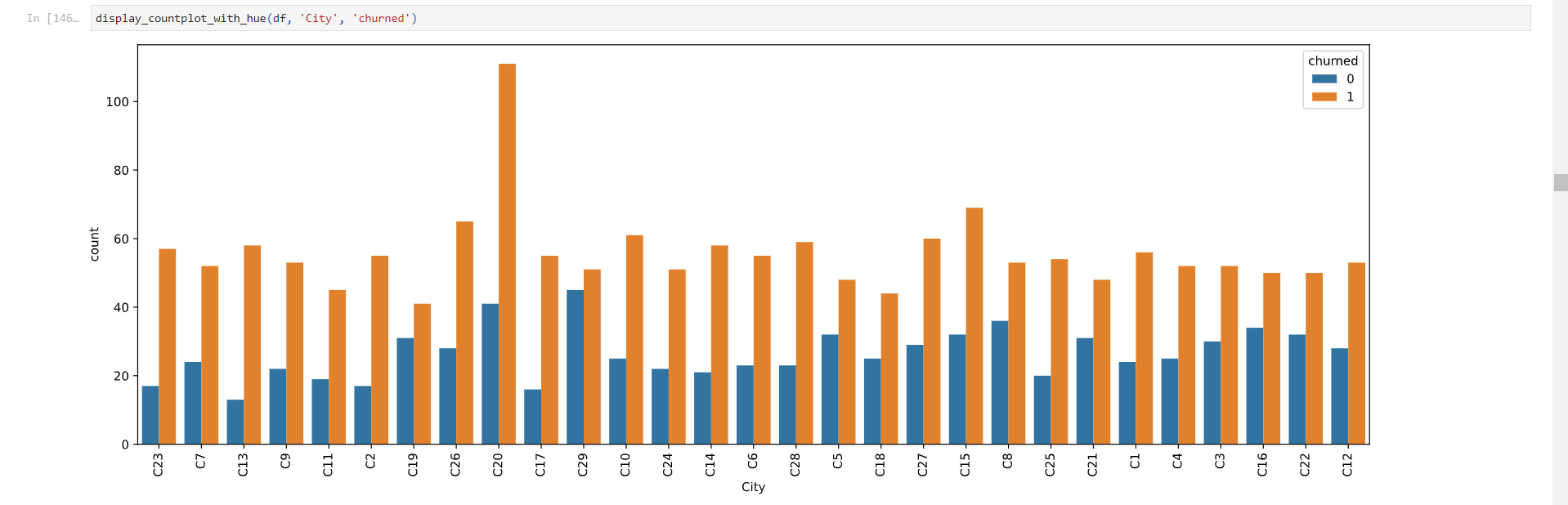
Key Insights:

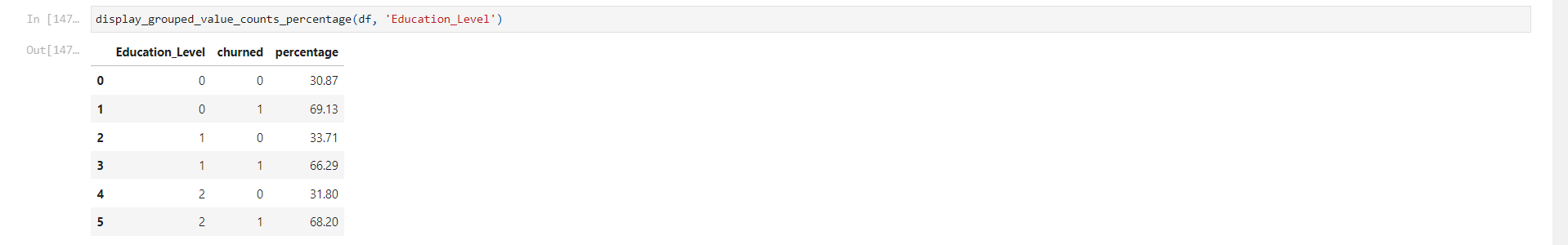
* Number of months:
  + Negative correlation with churn (PCC: -0.346, SRCC: -0.204)
  + Churned drivers tend to have spent less time with the company
  + Suggests that retention improves with longer tenure
* Age:
  + Weak negative correlation with churn (PCC: -0.056, SRCC: -0.065)
  + Age doesn't appear to be a strong predictor of churn

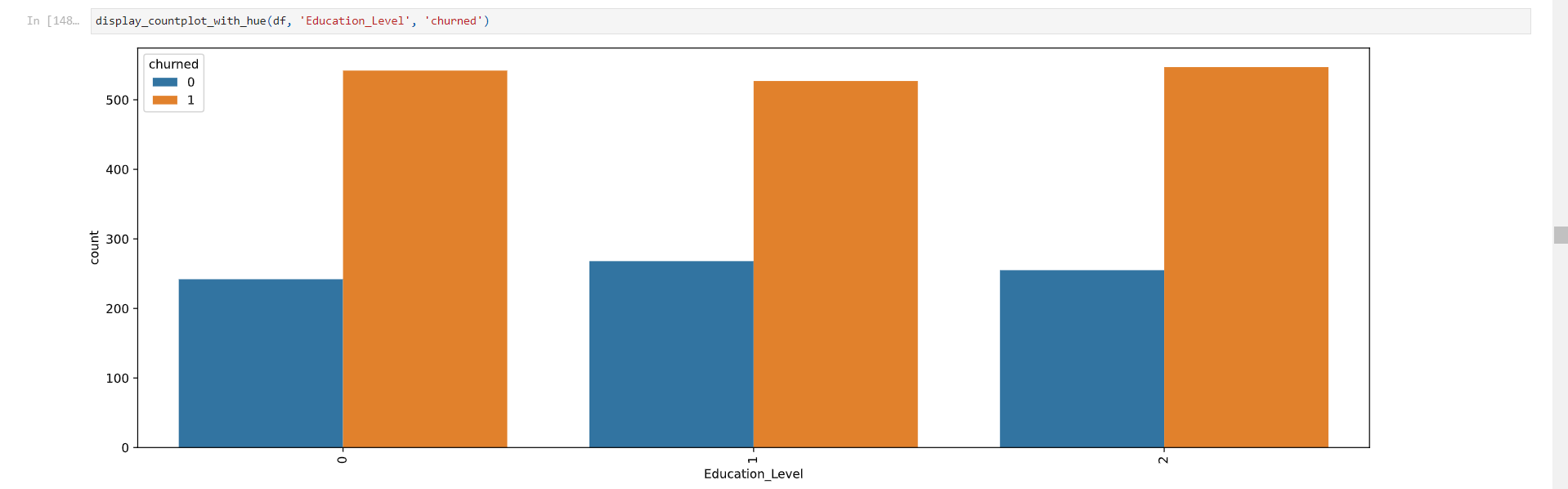
#### Task 3.2: Bivariate Analysis - Categorical Variables vs Churn

Figure 1.27: Table of Grouped Value Count Percentage of the Column

Figure 1.28: Graph of Gender vs Churn (Count Plot with Hue)

Figure 1.29: Graph of City vs Churn (Count Plot with Hue)

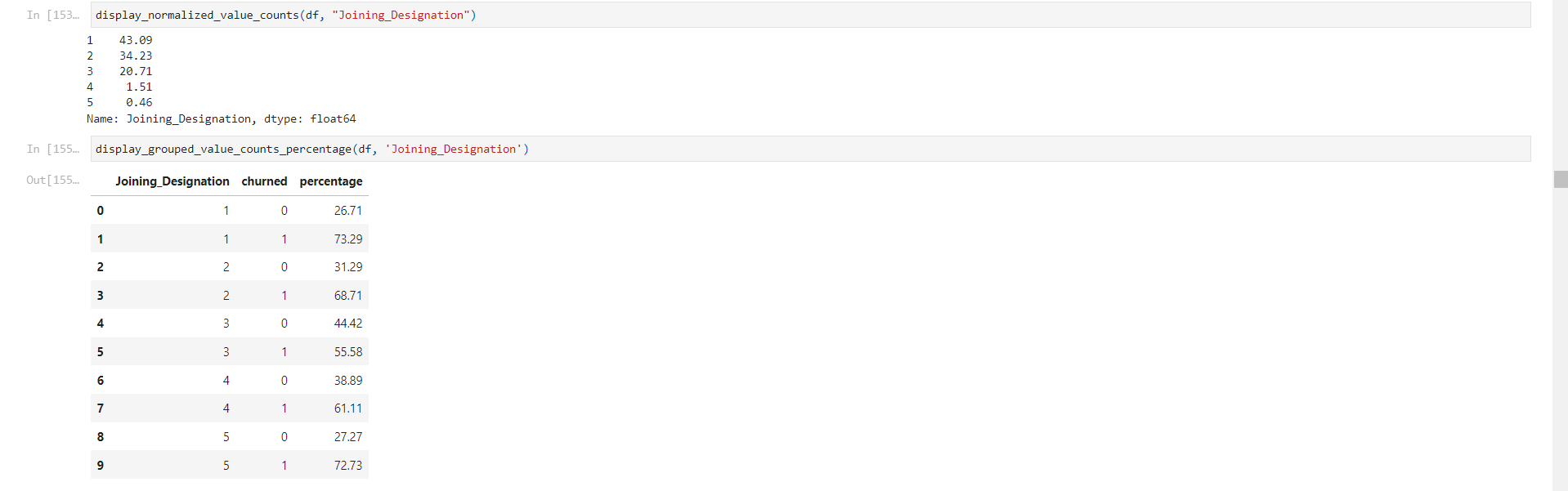
Figure 1.30: Table of Grouped Value Count Percentage of the Column

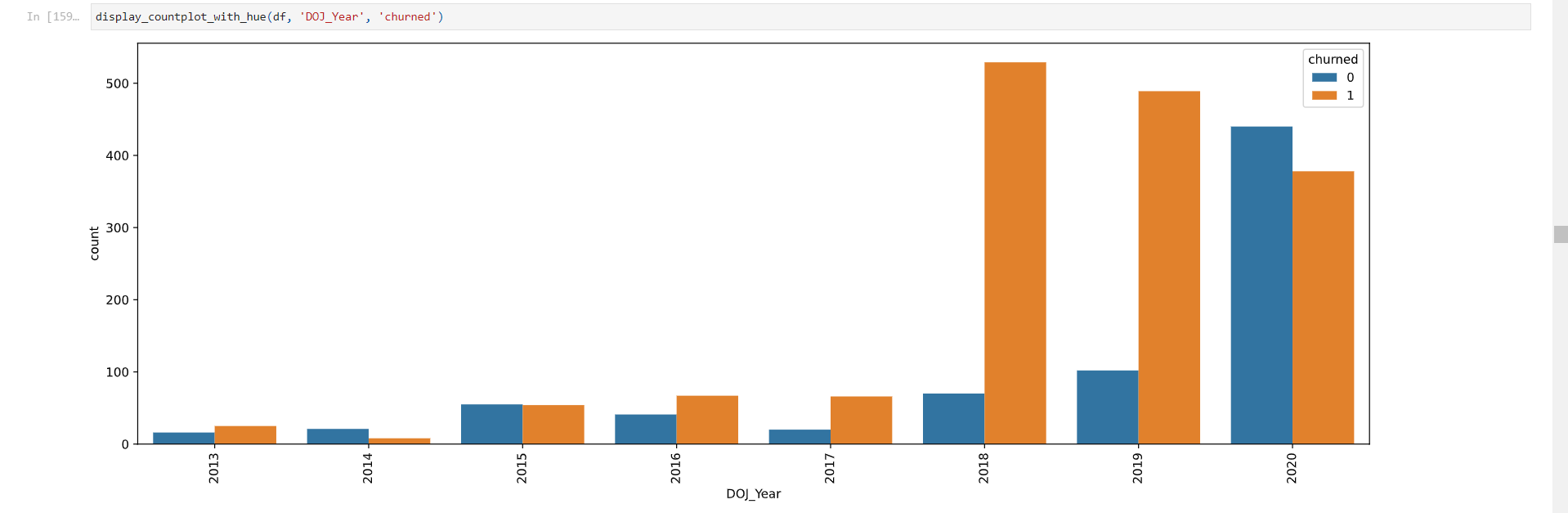
Figure 1.31: Graph of Education Level vs Churn (Count Plot with Hue)

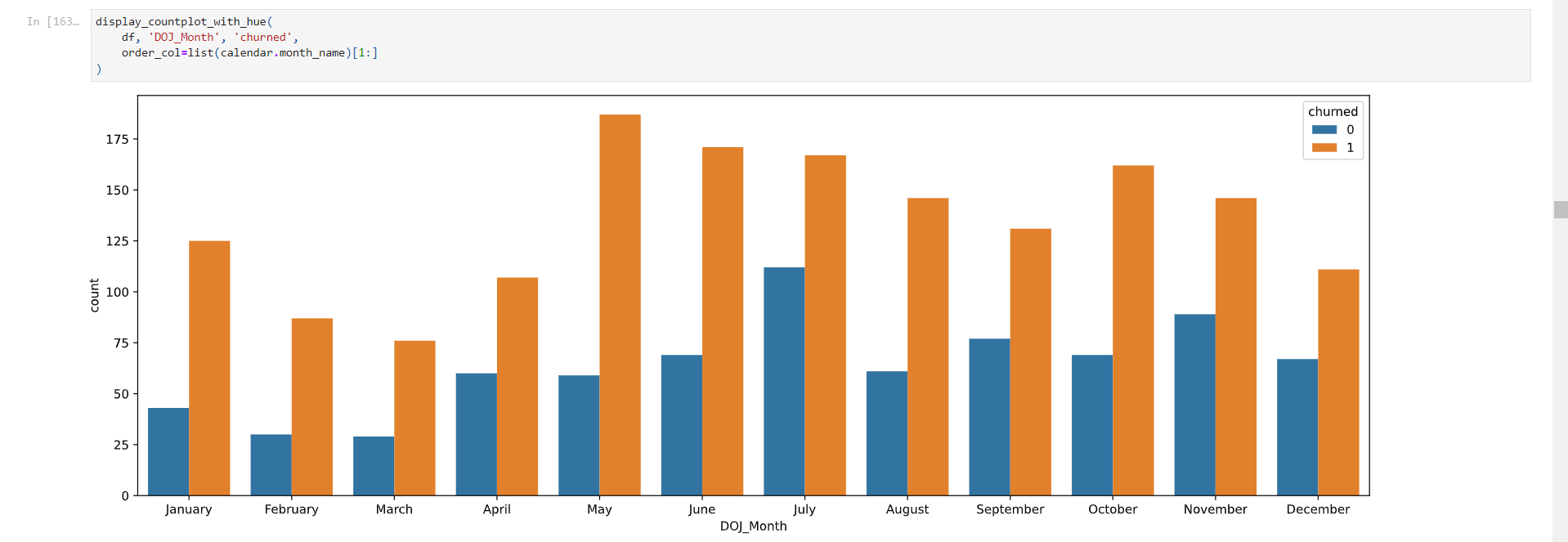
Key Insights:

* Gender:
  + Similar churn rates for males (67.52%) and females (68.37%)
  + Gender doesn't appear to be a significant factor in churn
* City:
  + Significant variations in churn rates across cities
  + Some cities (e.g., C13, C17, C2) have higher churn rates
  + Others (e.g., C12, C16, C19, C21, C22) have lower churn rates
* Education Level:
  + Slight variations in churn rates across education levels
  + Churn rates: 69.13% (Level 0), 66.29% (Level 1), 68.20% (Level 2)
  + Education level doesn't seem to be a strong predictor of churn

#### Task 3.3: Bivariate Analysis - Other Variables vs Churn

Figure 1.32: Table of Grouped Value Count Percentage of the Column

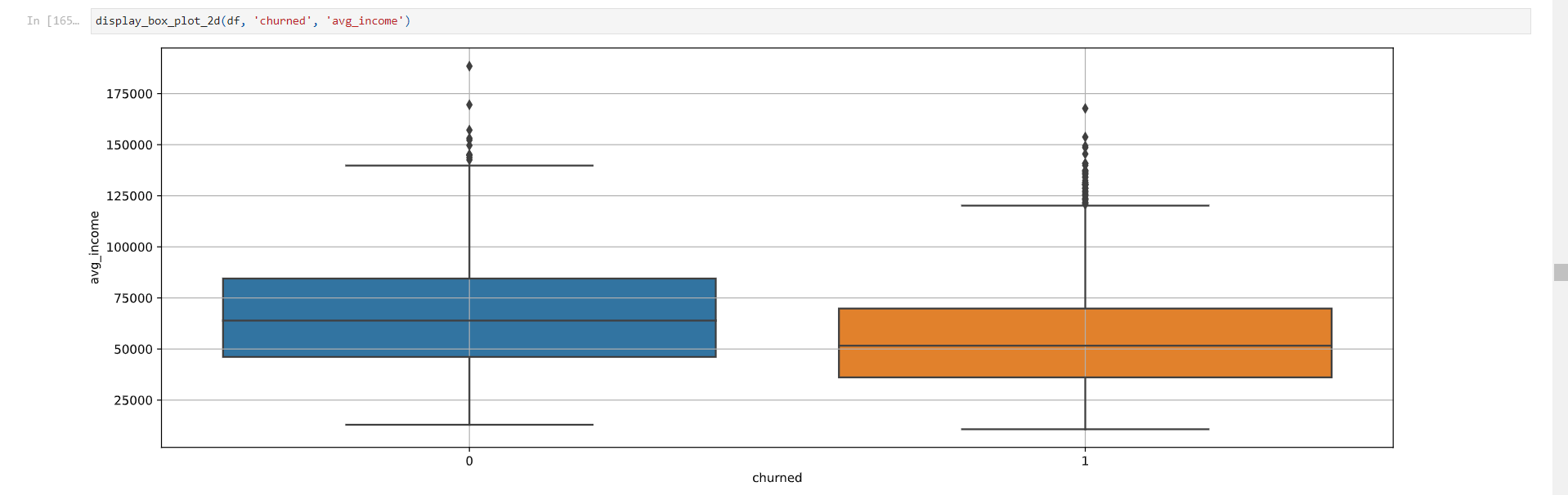
Figure 1.33: Graph of Year of Joining vs Churn (Count Plot with Hue)

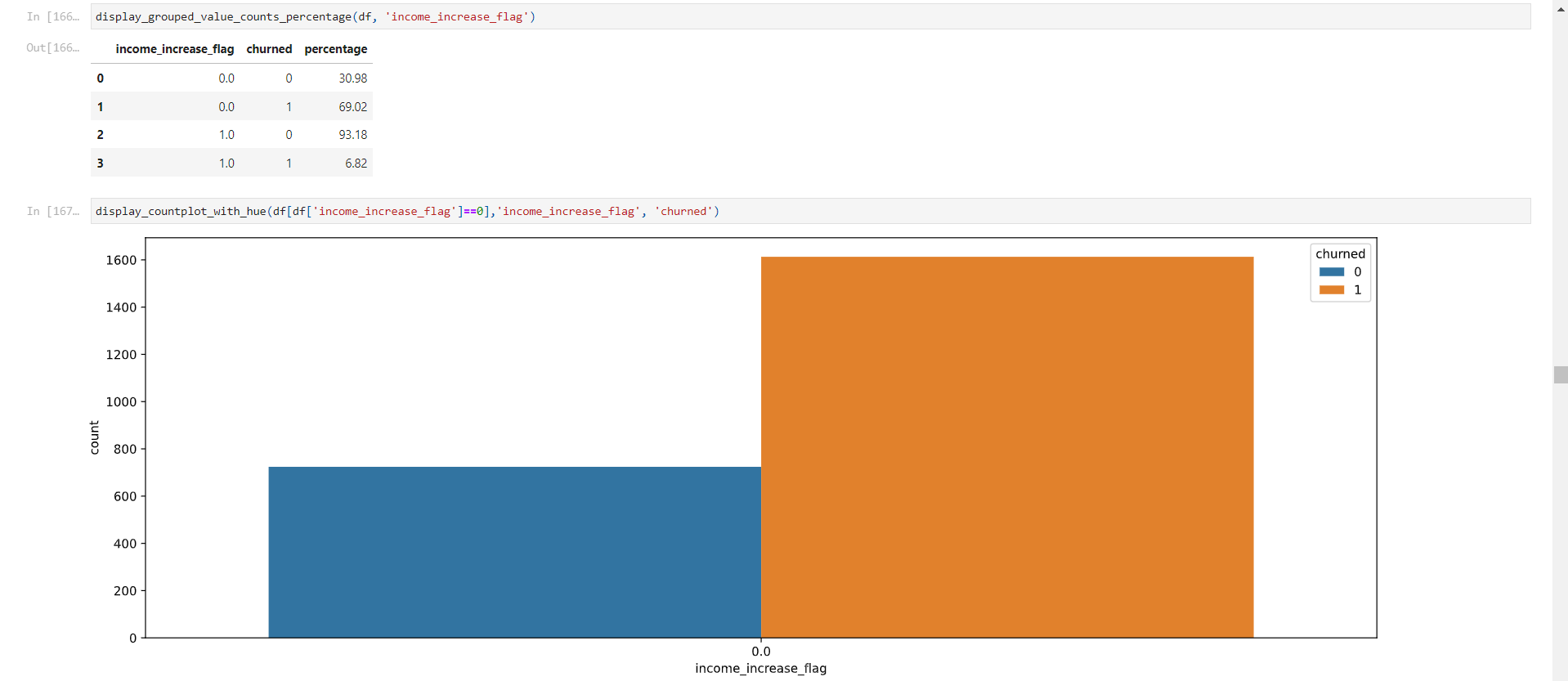
Figure 1.34: Graph of Month of Joining vs Churn (Count Plot with Hue)

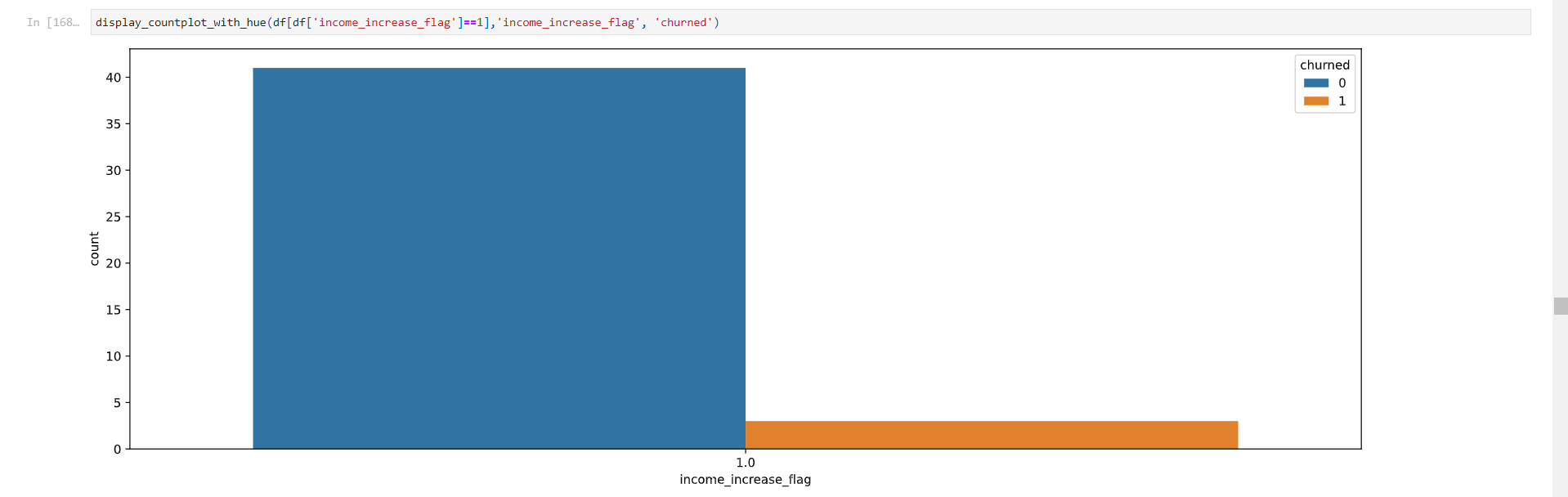
Key Insights:

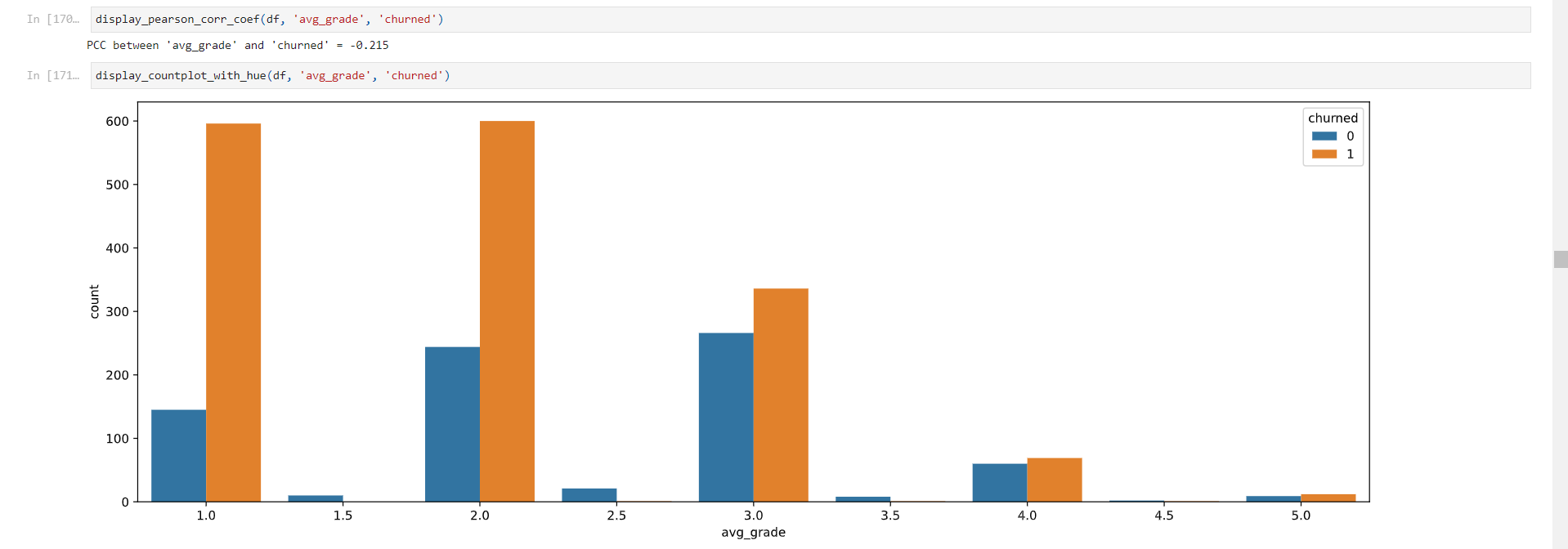
* Joining Designation:
  + Churn rate generally decreases as designation increases
  + Highest churn rate for designation 1 (73.29%), lowest for designation 3 (55.58%)
* Year of Joining:
  + Significant variations in churn rates across joining years
  + Highest churn rates for 2018 (88.31%) and 2019 (82.74%)
  + Lowest churn rates for 2014 (27.59%) and 2020 (46.21%)
* Month of Joining:
  + Variations in churn rates across joining months
  + Highest churn rates for May (76.02%) and January (74.40%)
  + Lowest churn rates for July (59.86%) and April (64.07%)

#### Task 3.4: Bivariate Analysis - Performance Metrics vs Churn

Figure 1.35: Graph of Average Income vs Churn (Box Plot 2D)

Figure 1.36: Graph of before Income Increase vs Churn (Count Plot with Hue)

Figure 1.37: Graph of after Income Increase vs Churn (Count Plot with Hue)

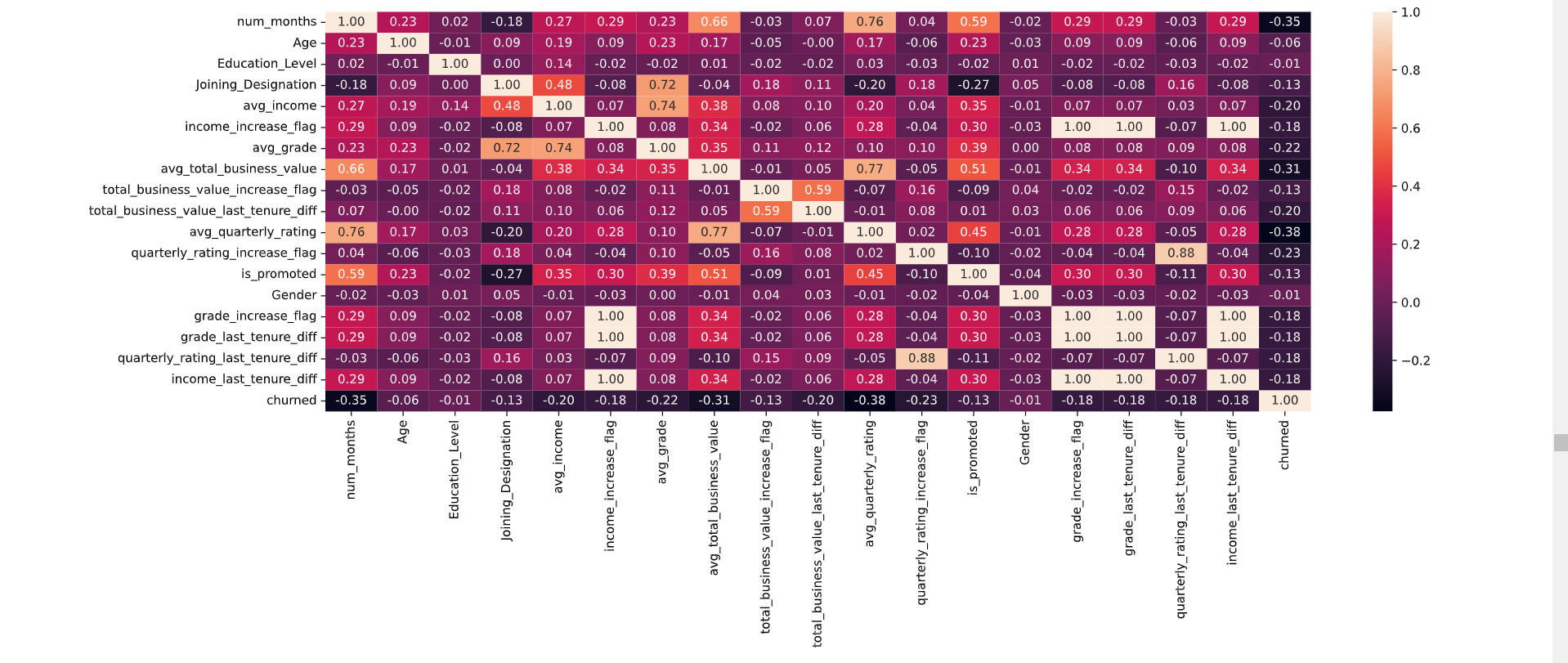
Figure 1.38: Graph of Average Grade vs Churn (Count Plot with Hue)

Key Insights:

* Average Income:
  + Slight negative correlation with churn
  + Churned drivers tend to have slightly lower average incomes
* Income Increase:
  + Drivers with income increases have significantly lower churn rates (6.82% vs 69.02%)
* Average Grade:
  + Negative correlation with churn (PCC: -0.215)
  + Higher grades associated with lower churn rates

#### Task 3.5: Correlation Analysis

Figure 1.39: Code for Correlation Analysis

Figure 1.40: Graph of Correlation Analysis

Key Insights:

* Strong negative correlations with churn:
  + Number of months (-0.35)
  + Average quarterly rating (-0.38)
  + Average total business value (-0.31)
* Moderate negative correlations with churn:
  + Income increase flag (-0.19)
  + Grade increase flag (-0.19)
  + Is promoted (-0.13)
* Weak or no correlations with churn:
  + Age (-0.06)
  + Education level (-0.02)
  + Gender (0.01)

#### Task 3.6: Data Preparation for Modeling

Figure 1.41: Code for Feature Selection and Data Splitting

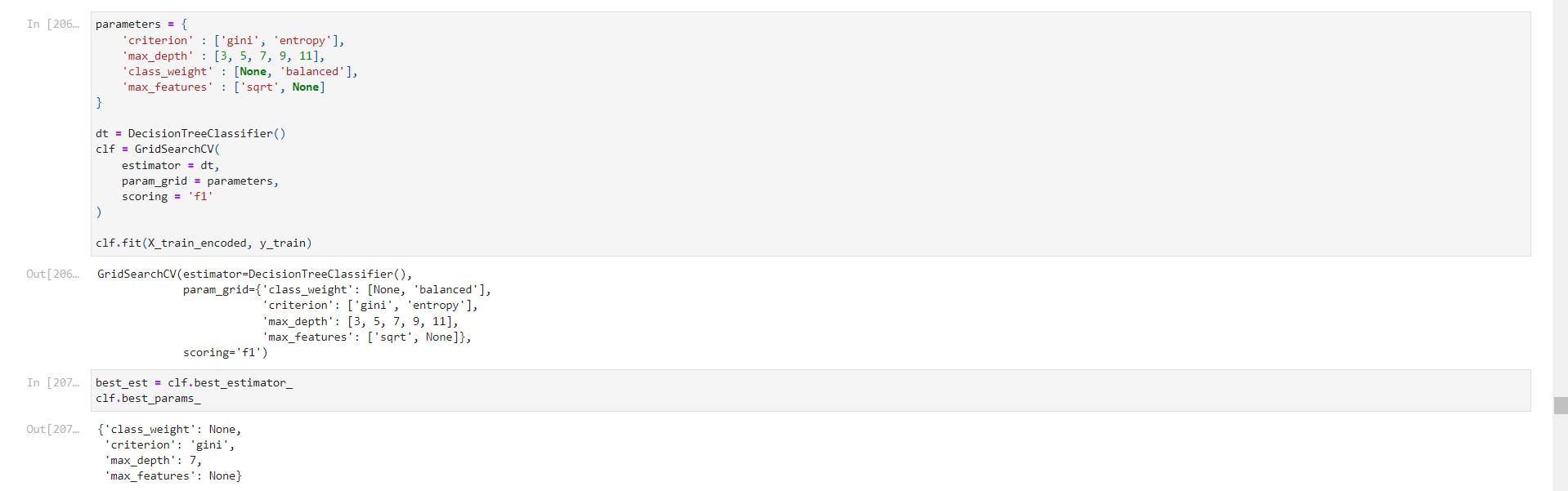
Figure 1.42: Code for One Hot Encoding of Categorical Columns

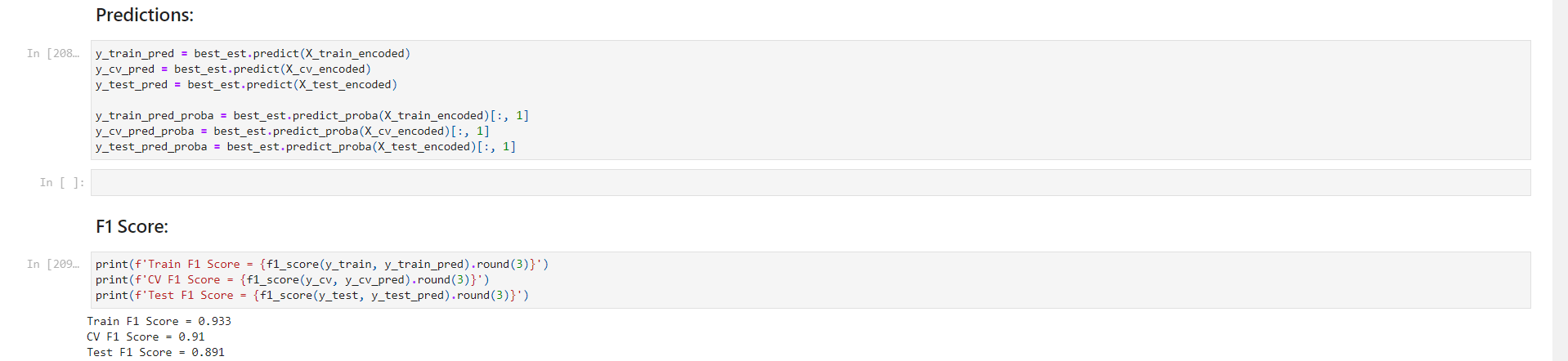
Key Insights:

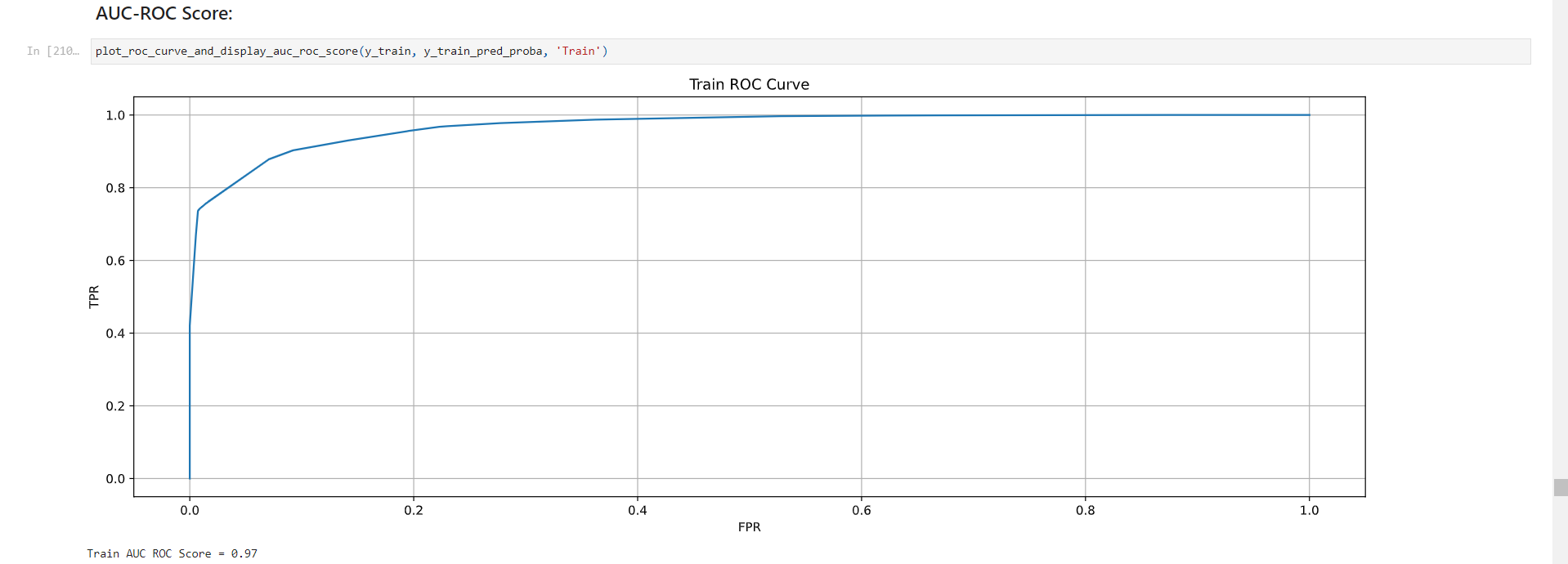
* Features with weak correlations to churn are dropped
* Data is split into training (70%), cross-validation (15%), and test (15%) sets
* Stratification is used to maintain the same proportion of churned drivers in all sets
* The resulting datasets have 1719 samples for training, 304 for cross-validation, and 358 for testing

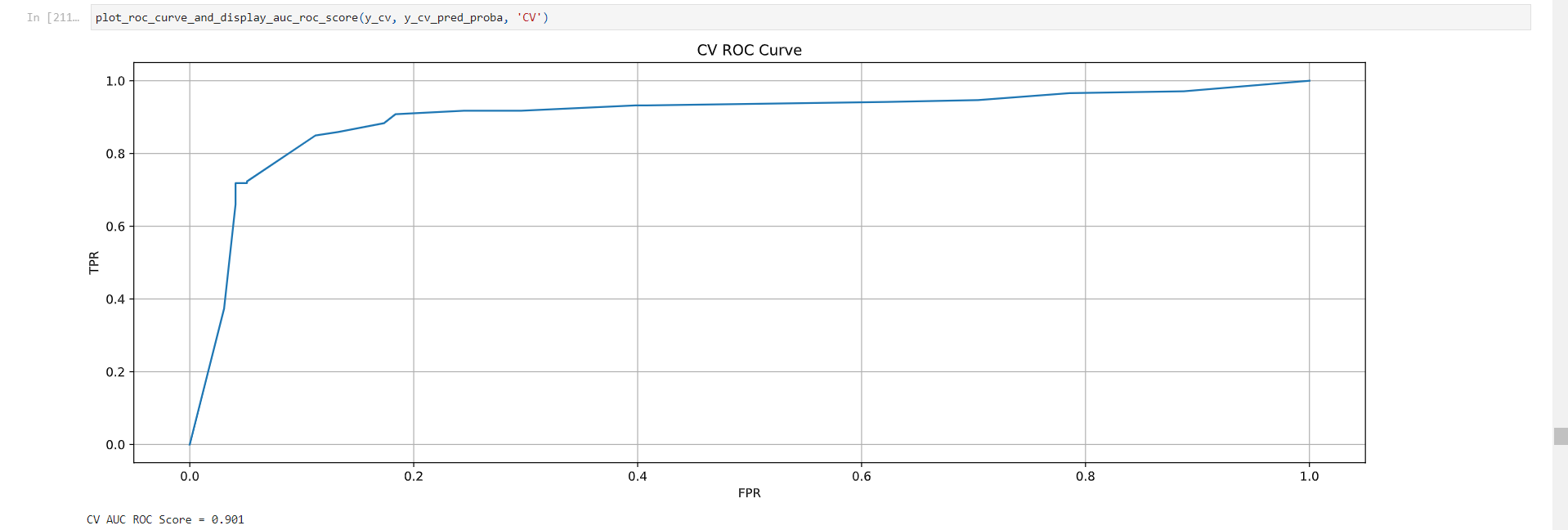
### Analysis Part 4: Model Development and Evaluation

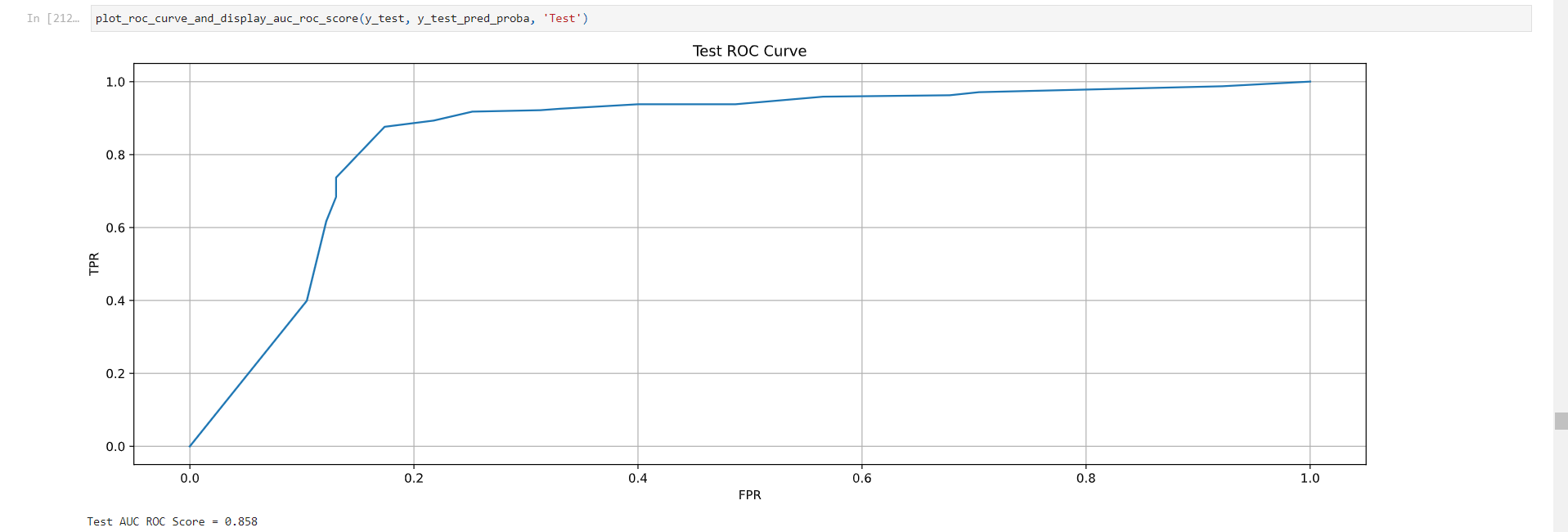
#### Task 4.1: Baseline Model - Decision Tree

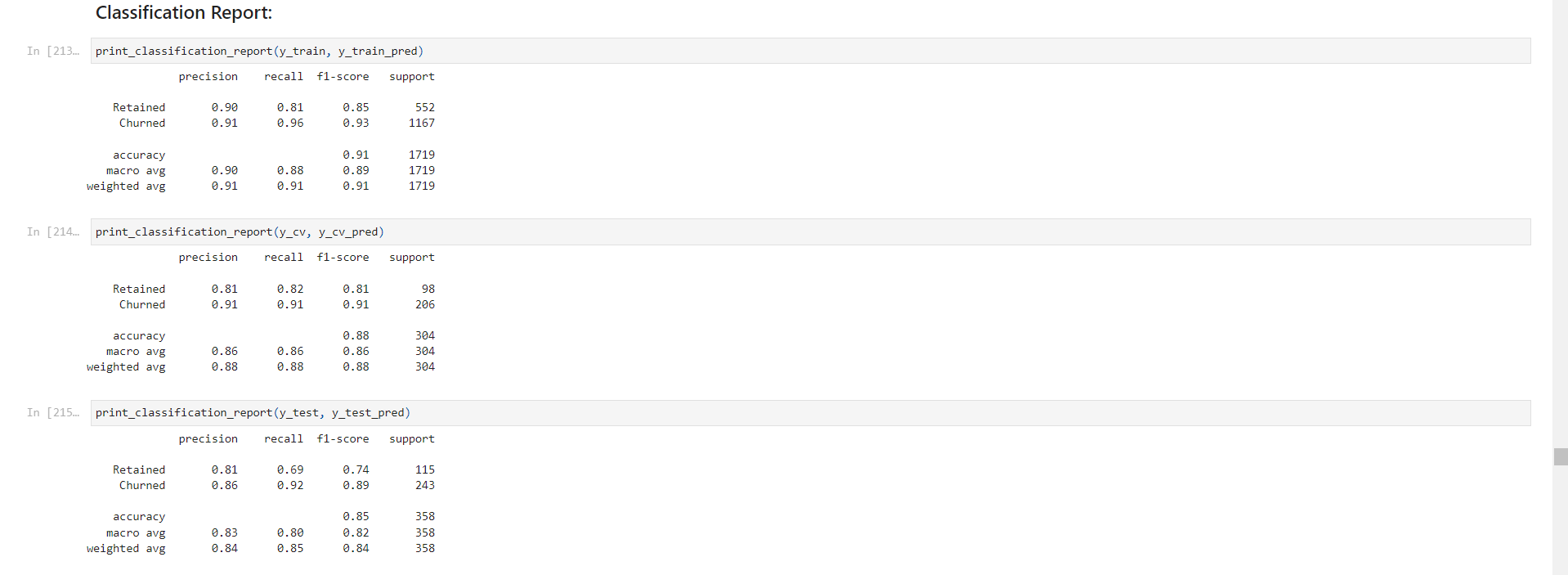
Figure 1.43: Code for Decision Tree Baseline Model Training

Figure 1.44: Code for Decision Tree Performance Metrics

Figure 1.45: Graph of Decision Tree AUC-ROC Training Curve

Figure 1.46: Graph of Decision Tree AUC-ROC Cross Validation Curve

Figure 1.47: Graph of Decision Tree AUC-ROC Testing Curve

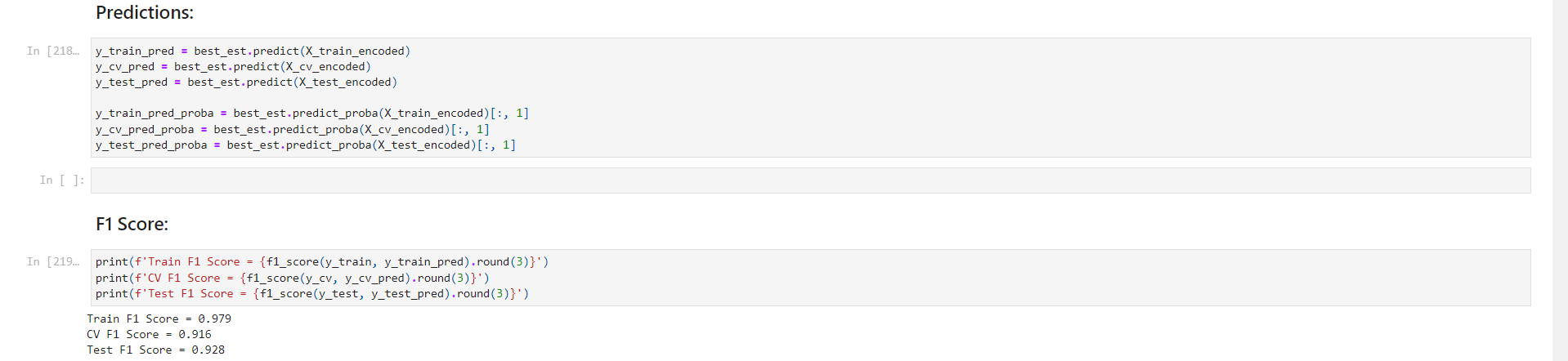
Figure 1.48: Table of Decision Tree Classification Report

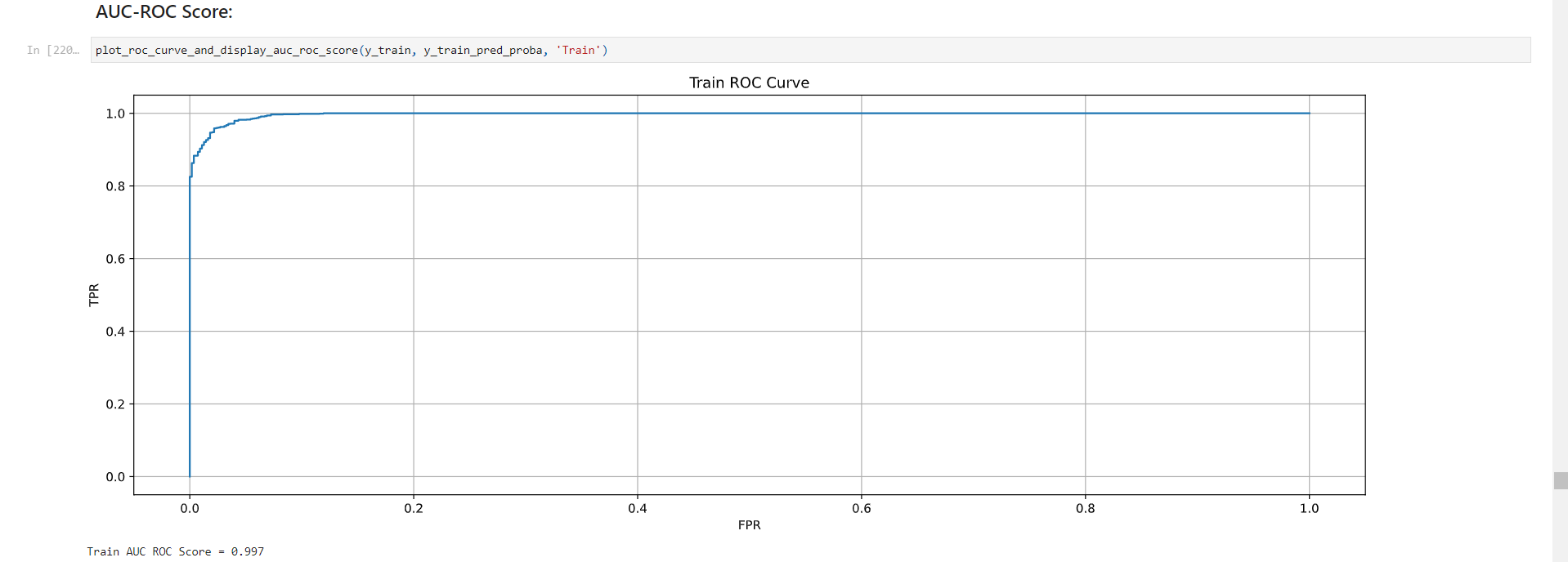
Key Insights:

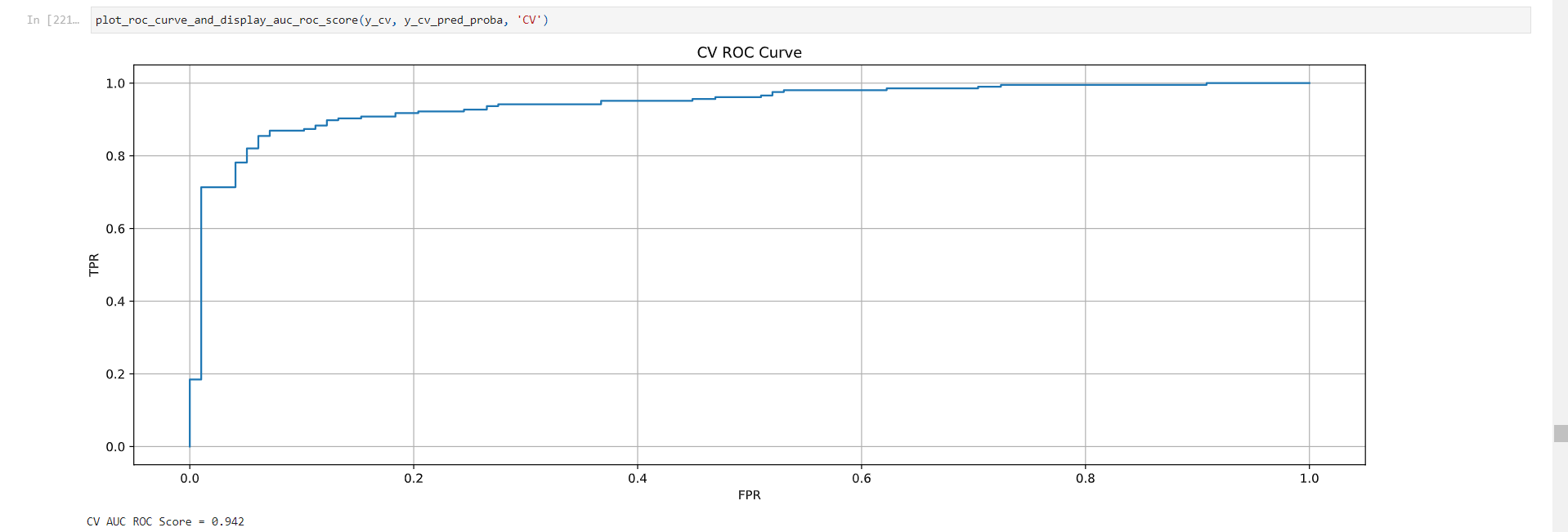
* Best parameters: {'class\_weight': None, 'criterion': 'gini', 'max\_depth': 7, 'max\_features': None}
* F1 Scores: Train (0.933), CV (0.91), Test (0.891)
* AUC-ROC Scores: Train (0.97), CV (0.901), Test (0.858)
* Good performance on both churned and retained classes
* Slight overfitting observed (train performance better than test)

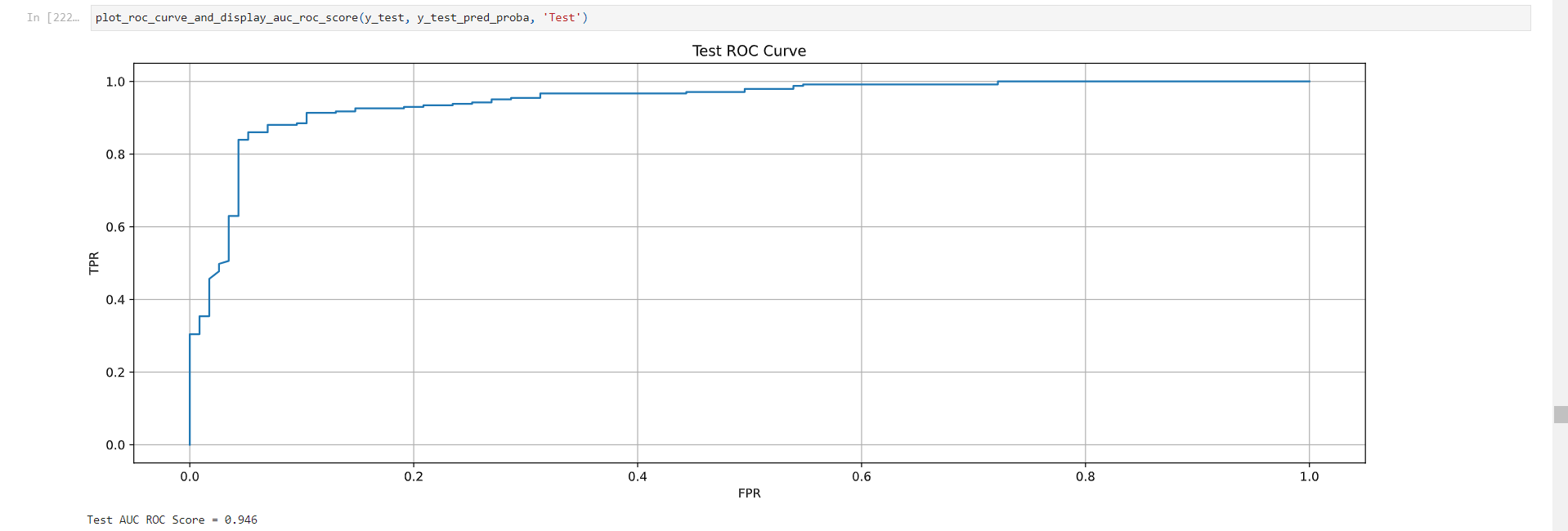
#### Task 4.2: Ensemble Model - Random Forest

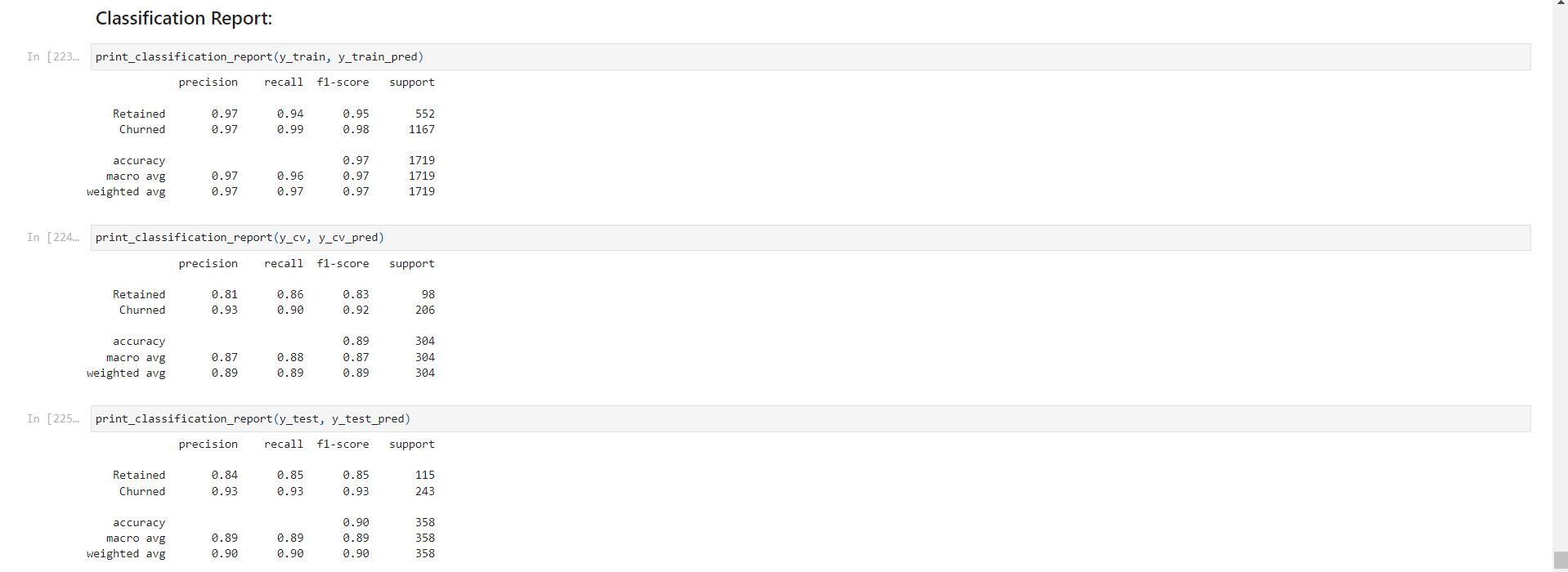
Figure 1.49: Code for Random Forest Model Training

Figure 1.50: Code for Random Forest Performance Metrics

Figure 1.51: Graph of Random Forest AUC-ROC Training Curve

Figure 1.52: Graph of Random Forest AUC-ROC Cross Validation Curve

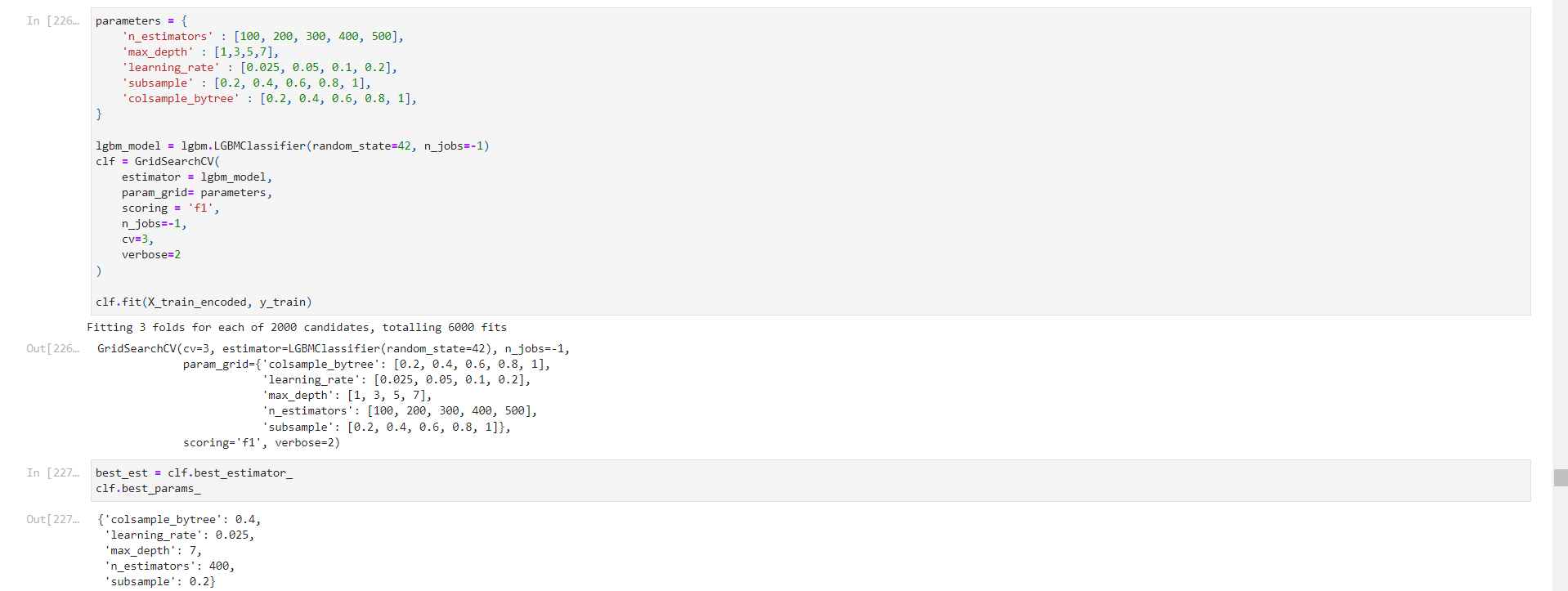
Figure 1.53: Graph of Random Forest AUC-ROC Testing Curve

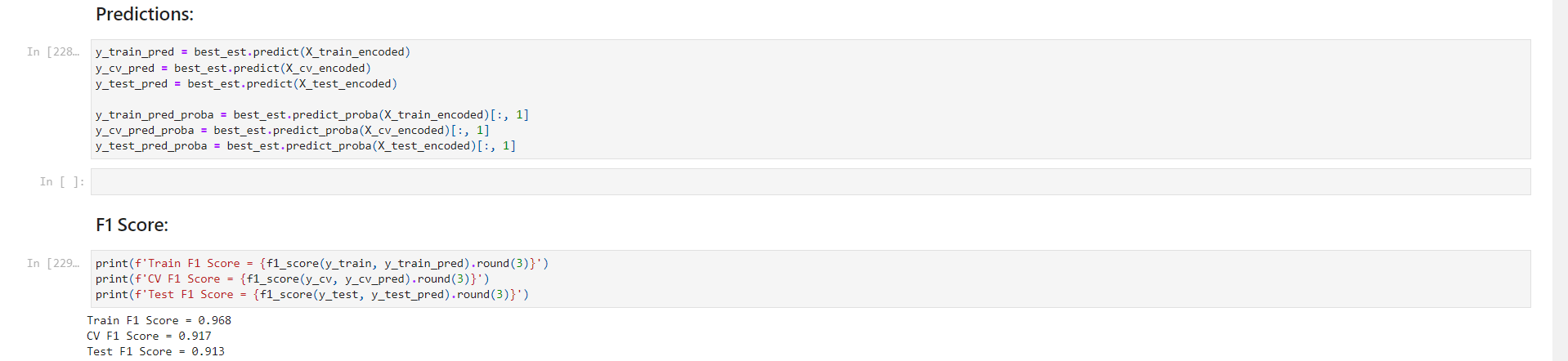
Figure 1.54: Table of Random Forest Classification Report

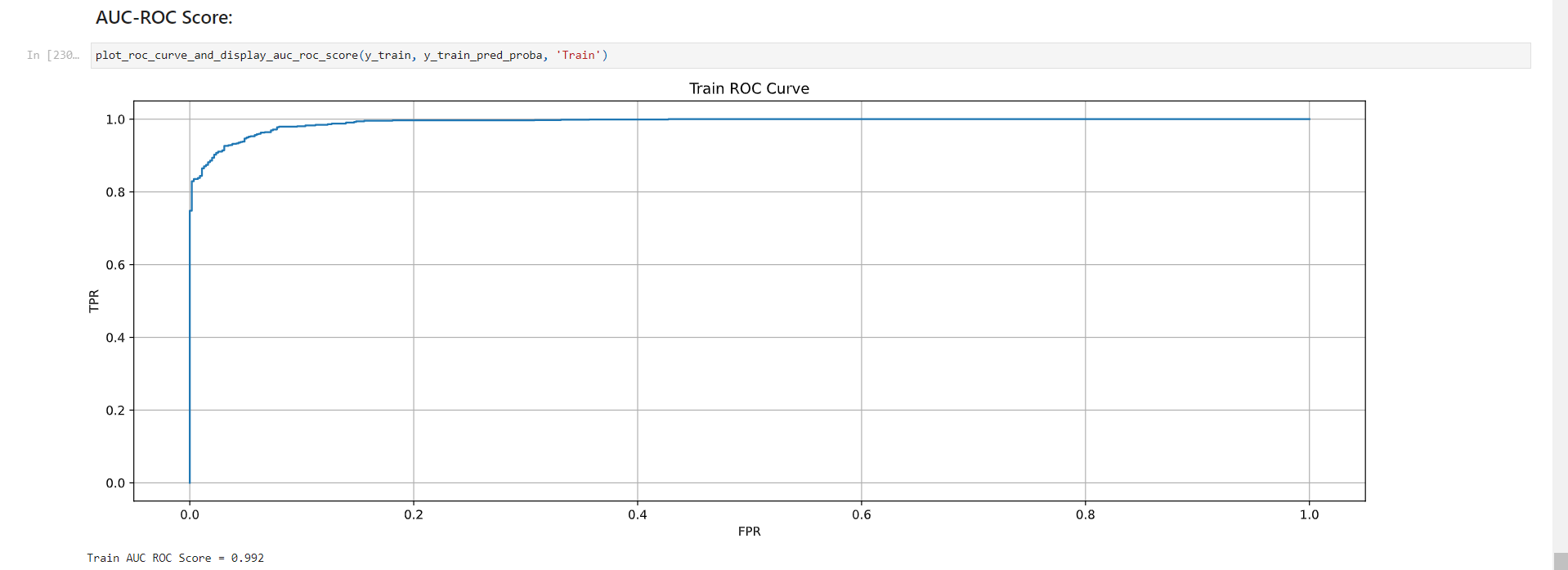
Key Insights:

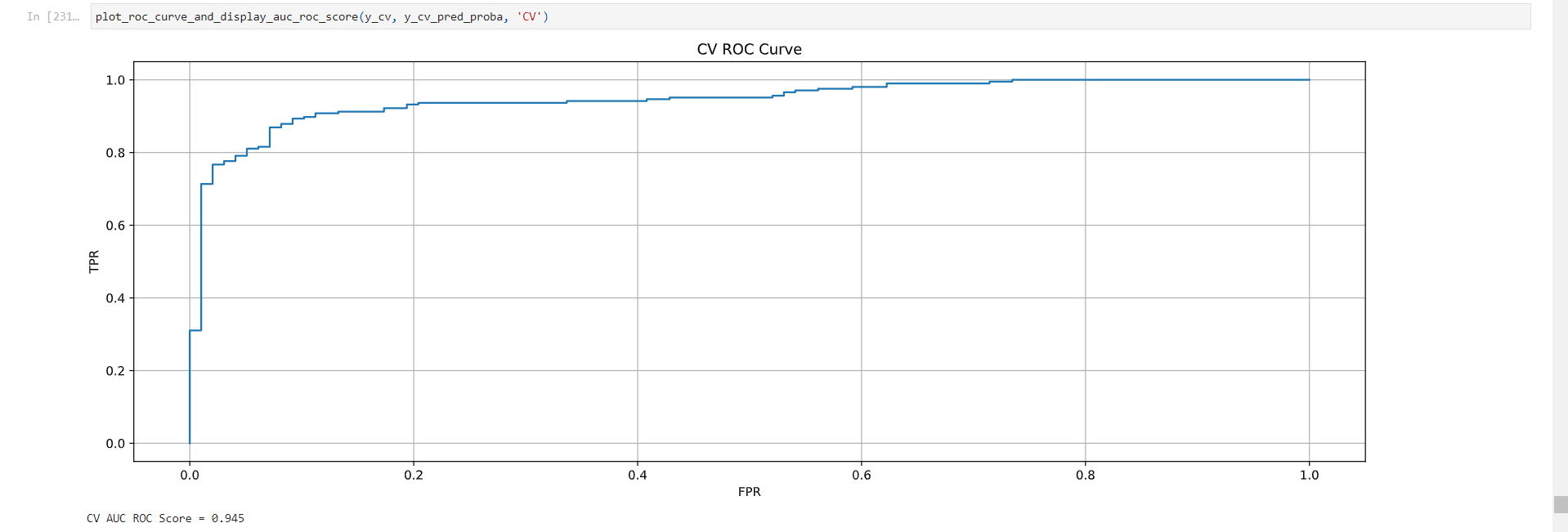
* Best parameters: {'max\_depth': 11, 'max\_features': 0.5, 'max\_samples': 0.75, 'n\_estimators': 100}
* F1 Scores: Train (0.979), CV (0.916), Test (0.928)
* AUC-ROC Scores: Train (0.997), CV (0.942), Test (0.946)
* Improved performance compared to Decision Tree
* Still shows some overfitting, but less pronounced

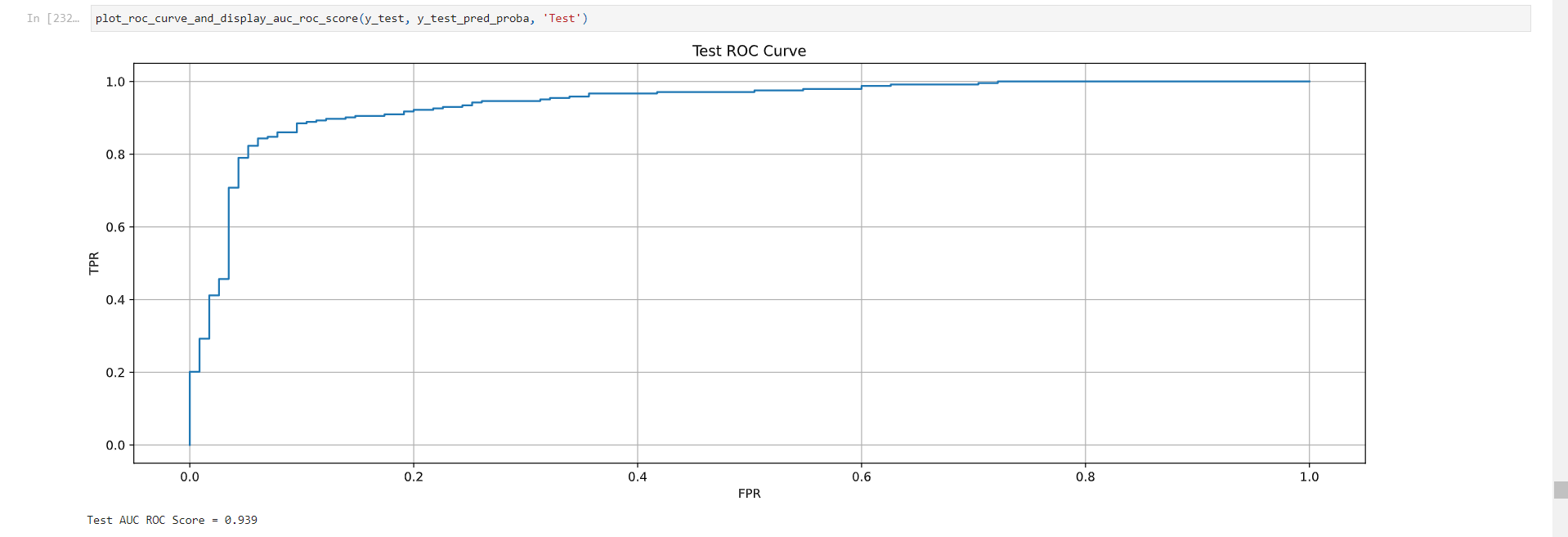
#### Task 4.3: Boosting Model - LightGBM

Figure 1.55: Code for LightGBM Model Training

Figure 1.56: Code for LightGBM Performance Metrics

Figure 1.57: Graph of LightGBM AUC-ROC Training Curve

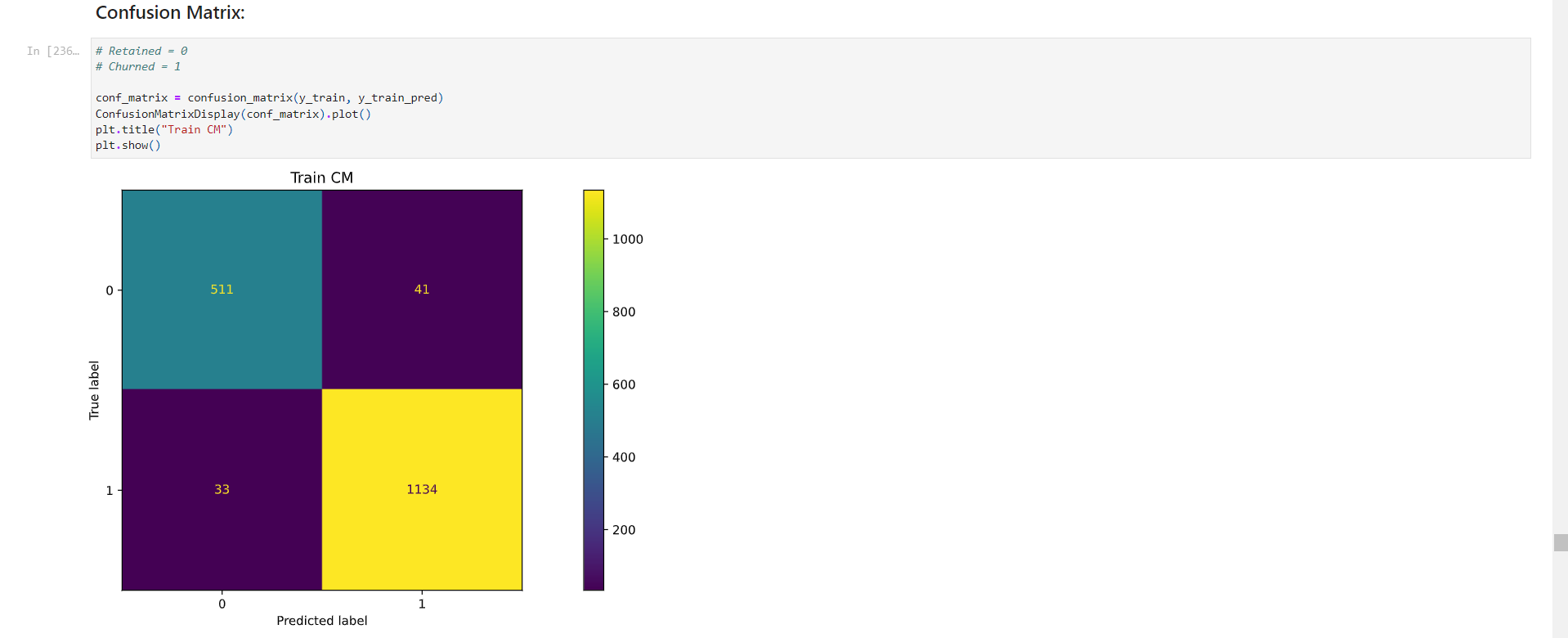
Figure 1.58: Graph of LightGBM AUC-ROC Cross Validation Curve

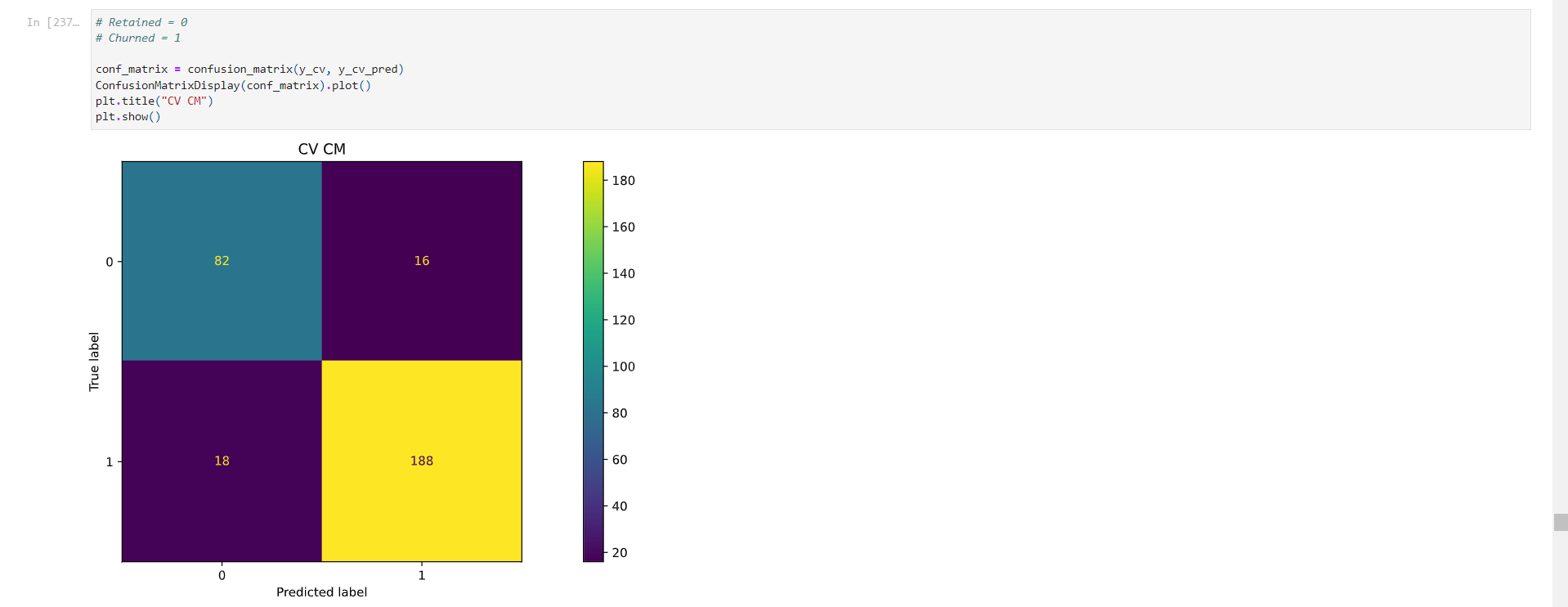
Figure 1.59: Graph of LightGBM AUC-ROC Testing Curve

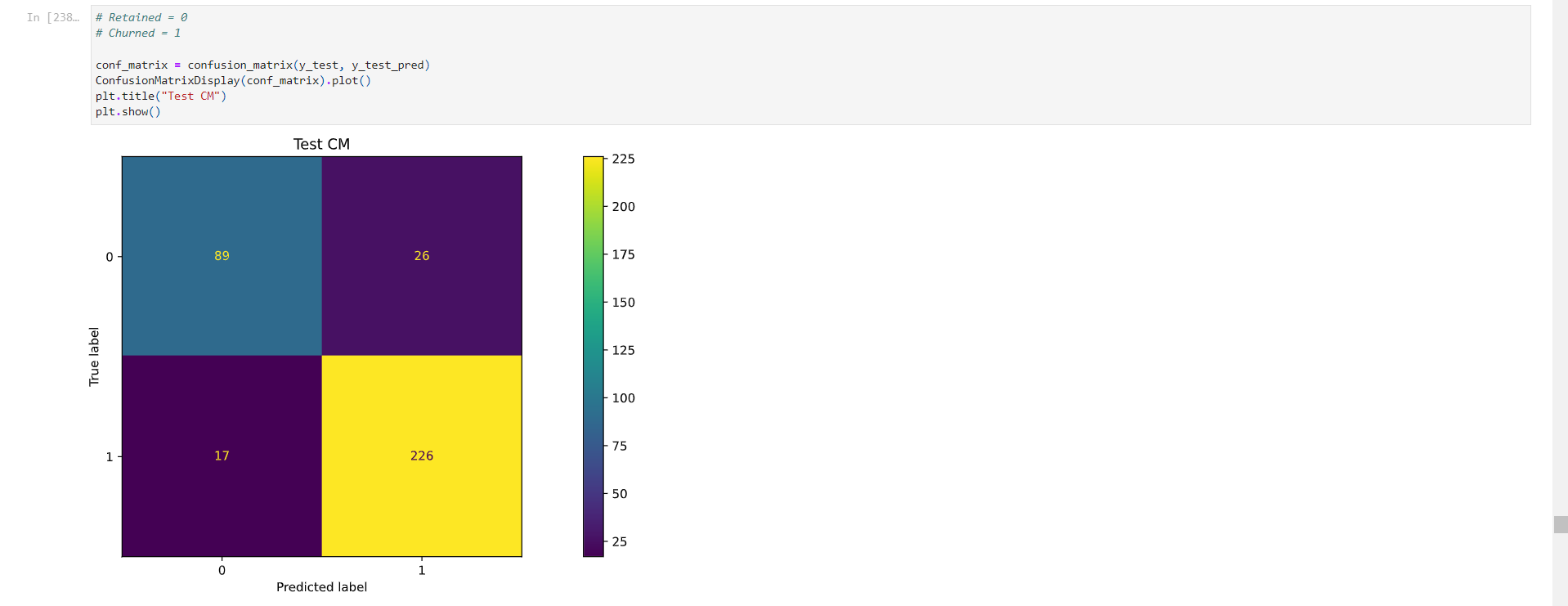
Key Insights:

* Best parameters: {'colsample\_bytree': 0.4, 'learning\_rate': 0.025, 'max\_depth': 7, 'n\_estimators': 400, 'subsample': 0.2}
* F1 Scores: Train (0.968), CV (0.917), Test (0.913)
* AUC-ROC Scores: Train (0.992), CV (0.945), Test (0.939)
* Performance similar to Random Forest, with slightly less overfitting

#### Task 4.4: Model Comparison and Selection

Figure 1.60: Table of Confusion Matrix for Training

Figure 1.61: Table of Confusion Matrix for Cross Validation

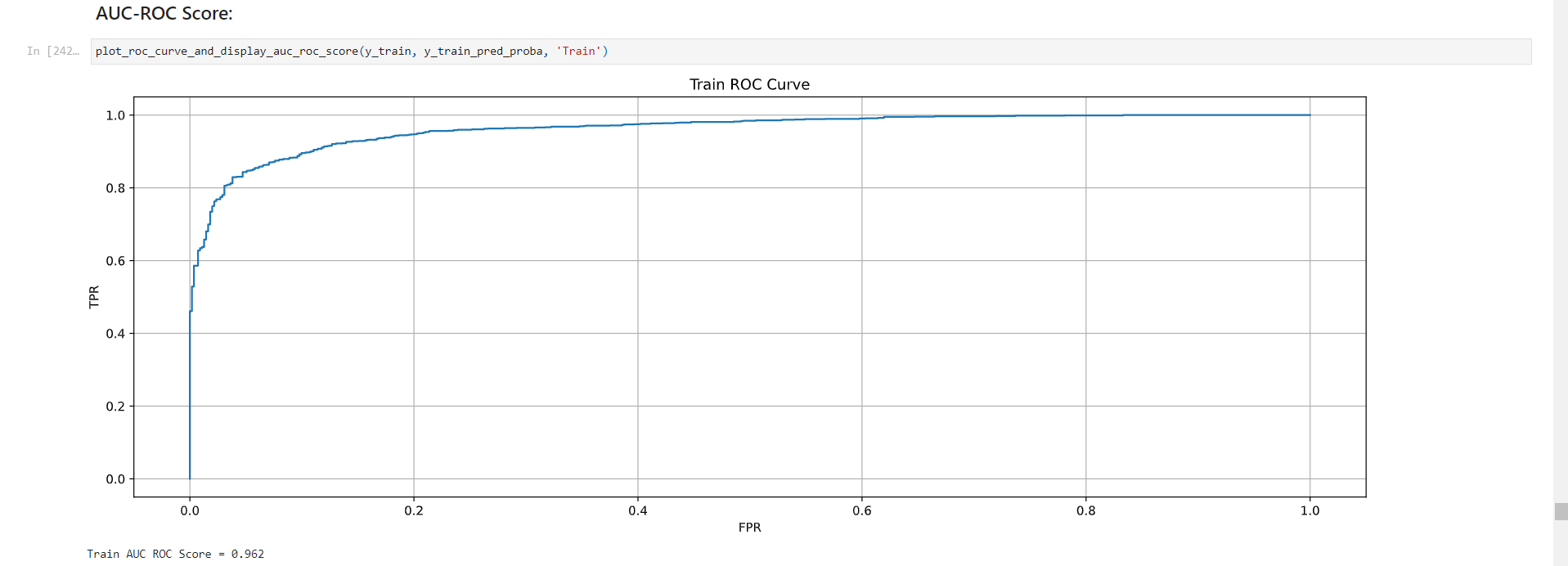
Figure 1.62: Table of Confusion Matrix for Testing

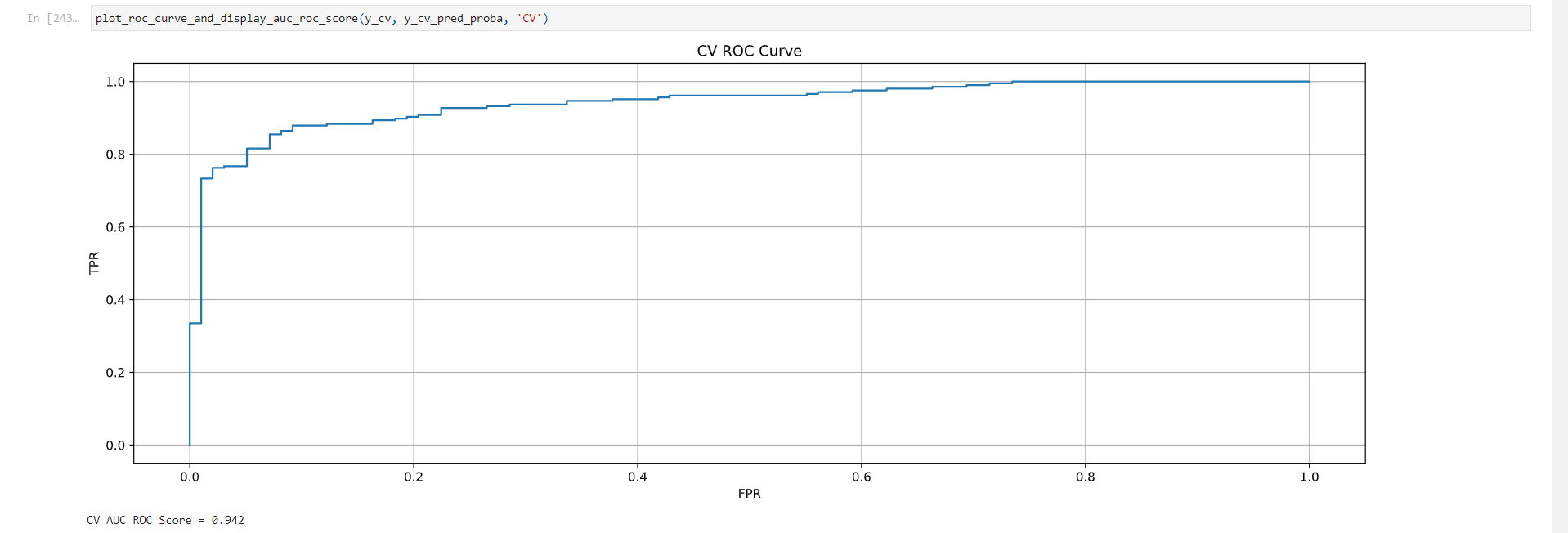
Key Insights:

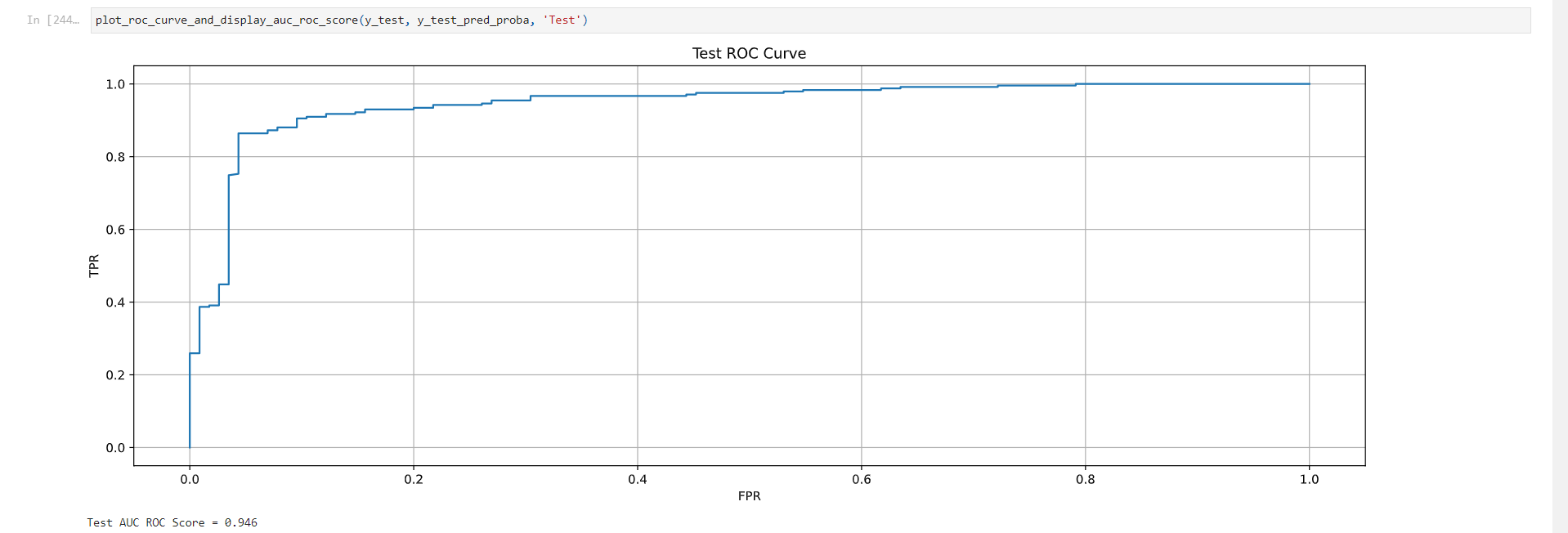
* Random Forest model performed the best in terms of Test F1 Score
* All models show good performance, with F1 scores above 0.89
* Random Forest and LightGBM both show improved performance over the baseline Decision Tree model

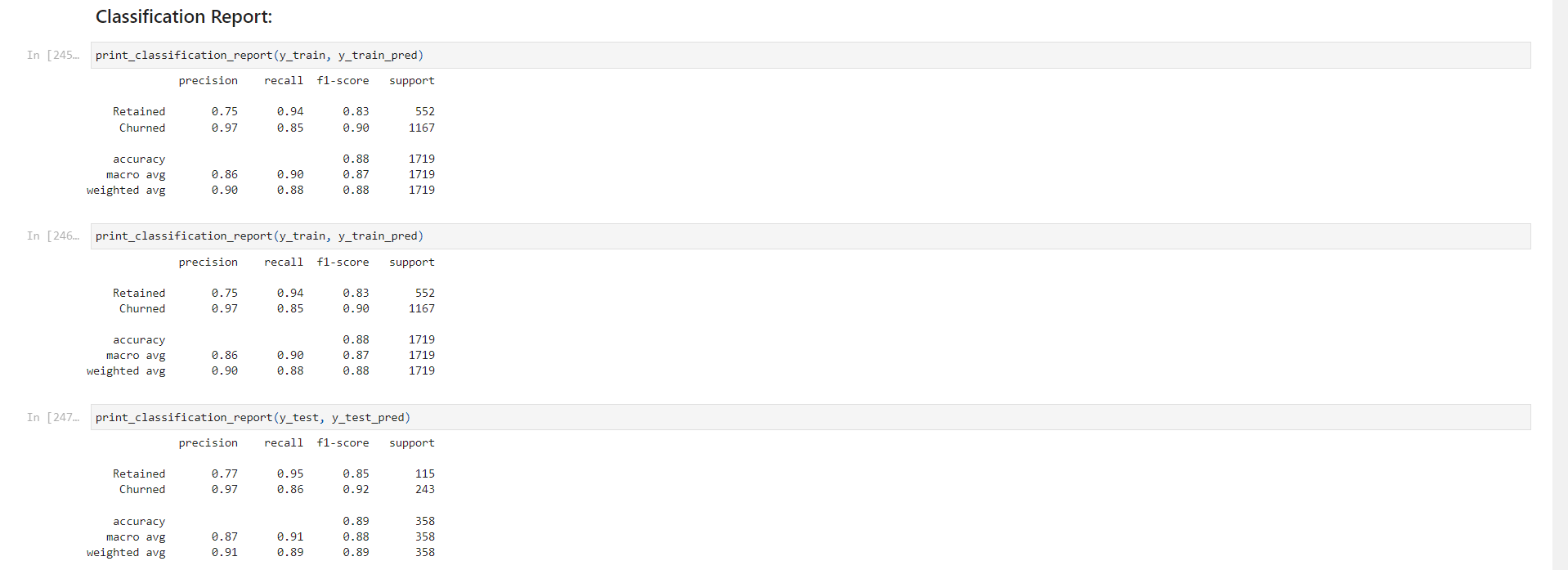
#### Task 4.5: Class Imbalance Treatment

Figure 1.63: Code for Class Weight Balanced in LightGBM

Figure 1.64: Graph of Balanced LightGBM AUC-ROC Training Curve

Figure 1.65: Graph of Balanced LightGBM AUC-ROC Cross Validation Curve

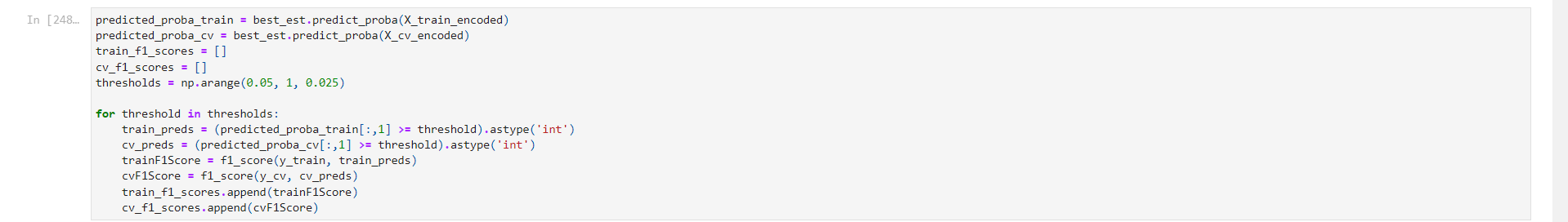
Figure 1.66: Graph of Balanced LightGBM AUC-ROC Testing Curve

Figure 1.67: Table of Balanced LightGBM Classification Report

Key Insights:

* Class weight balancing applied to LightGBM model
* F1 Scores: Train (0.905), CV (0.899), Test (0.915)
* Slight improvement in Test F1 Score compared to unbalanced LightGBM model
* Reduced overfitting, with train and test scores closer together

#### Task 4.6: Threshold Optimization

Figure 1.68: Code for Threshold Optimization

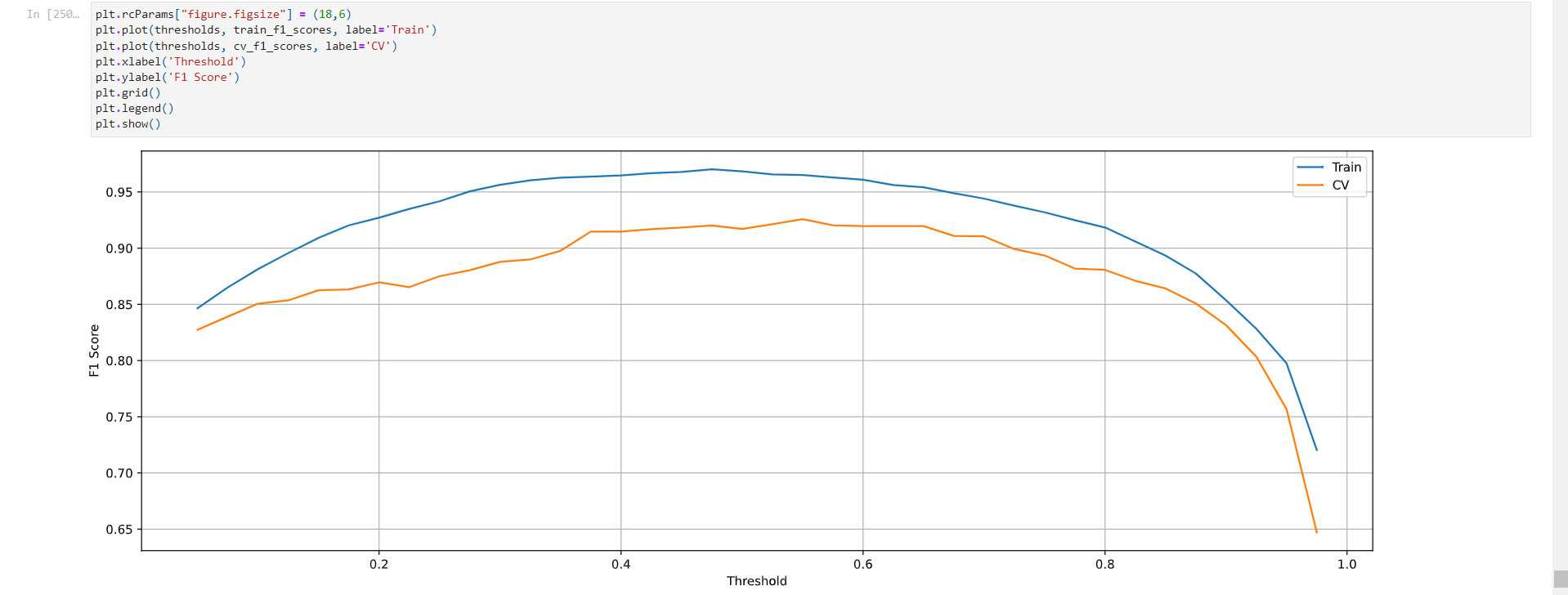
Figure 1.69: Graph of Optimized Threshold Performance

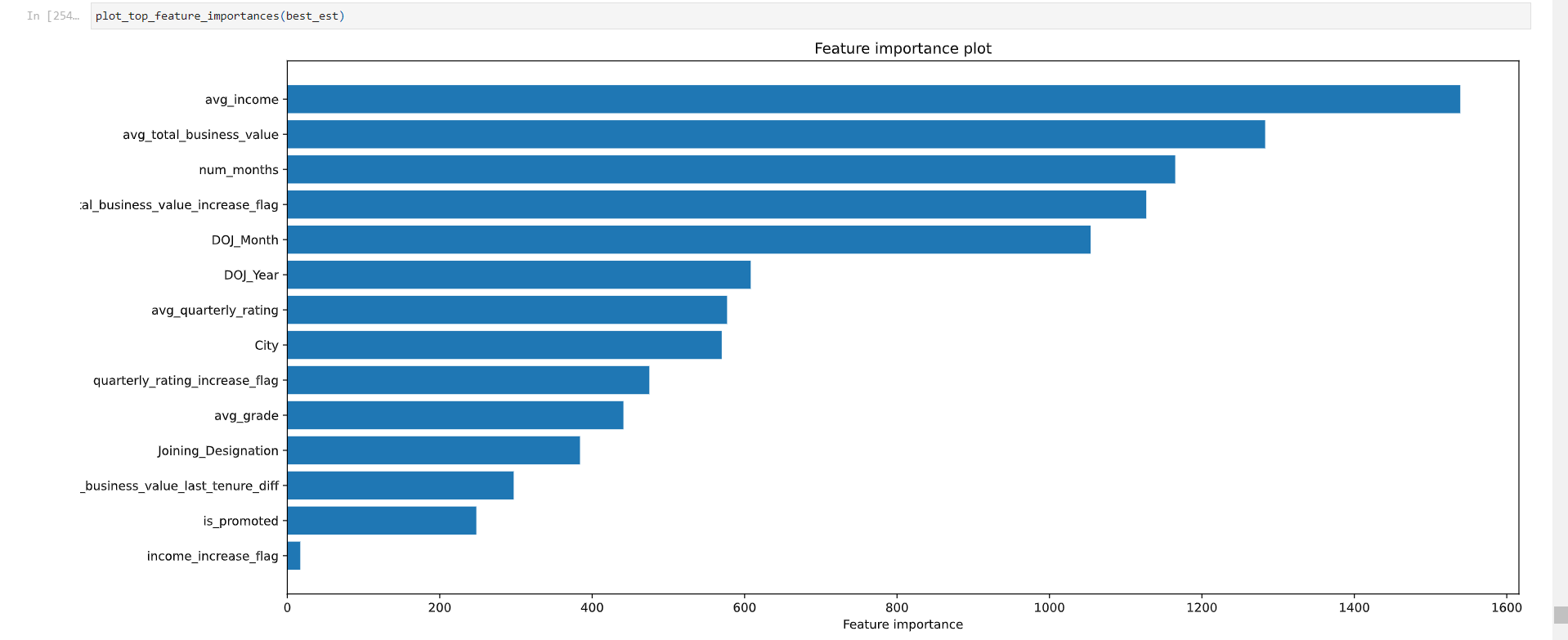
Figure 1.70: Table of Best Threshold with Best Scores

Key Insights:

* Best threshold found: 0.55
* Optimized F1 Scores: Train (0.965), CV (0.926), Test (0.911)
* Threshold optimization slightly improved model performance

### Analysis Part 5: Model Interpretation and Error Analysis

#### Task 5.1: Feature Importance Analysis

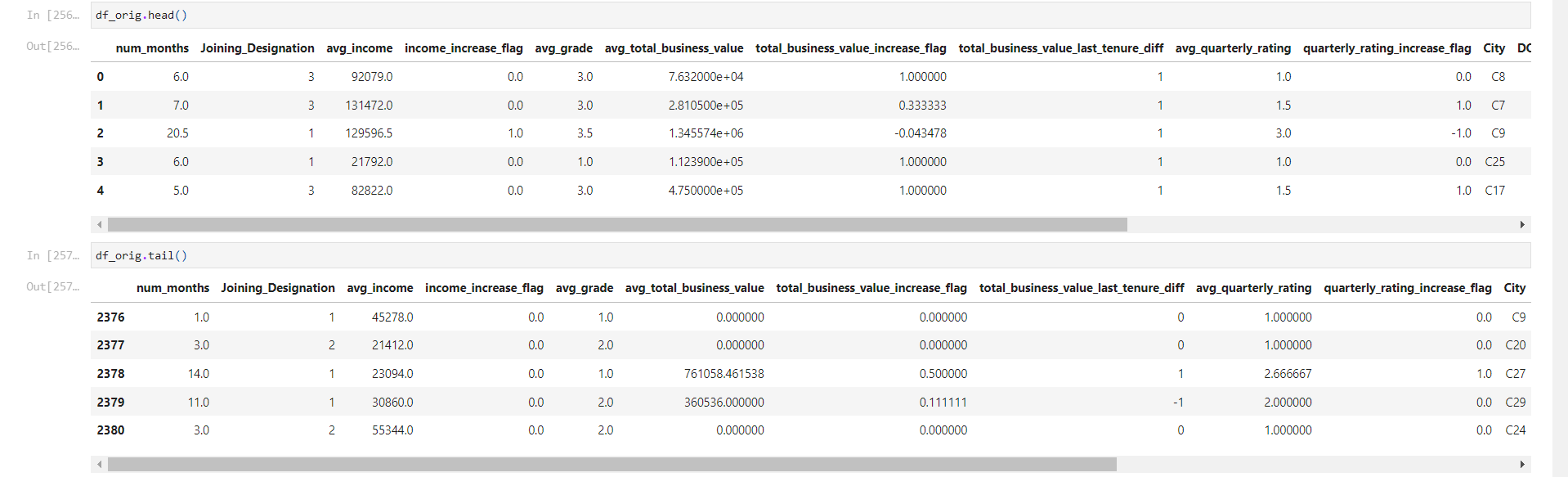
Figure 1.71: Graph of Feature Importance (Bar Graph)

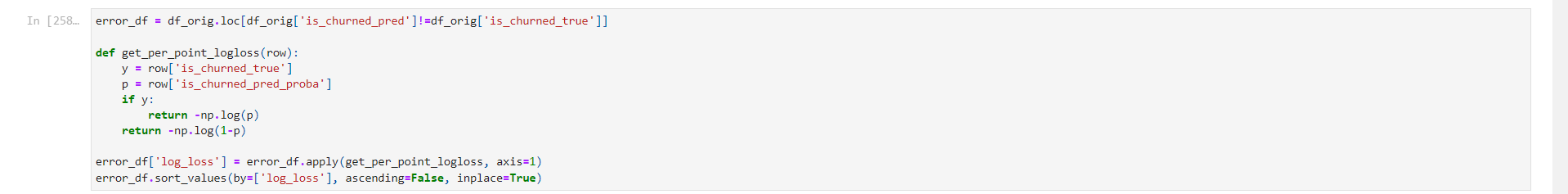
Key Insights:

* The top 5 most important features for predicting driver churn are:
  + avg\_income
  + avg\_total\_business\_value
  + num\_months
  + total\_business\_value\_increase\_flag
  + DOJ\_Month
* Financial factors (income and business value) and tenure (num\_months) are the most influential in predicting churn.
* The month of joining (DOJ\_Month) also plays a significant role, suggesting seasonal effects on driver retention.
* These findings align with our earlier exploratory data analysis, confirming the importance of these factors in driver churn prediction.

#### Task 5.2: Error Analysis Preparation

Figure 1.72: Code for Creating Consolidated DataFrame and Error DataFrame

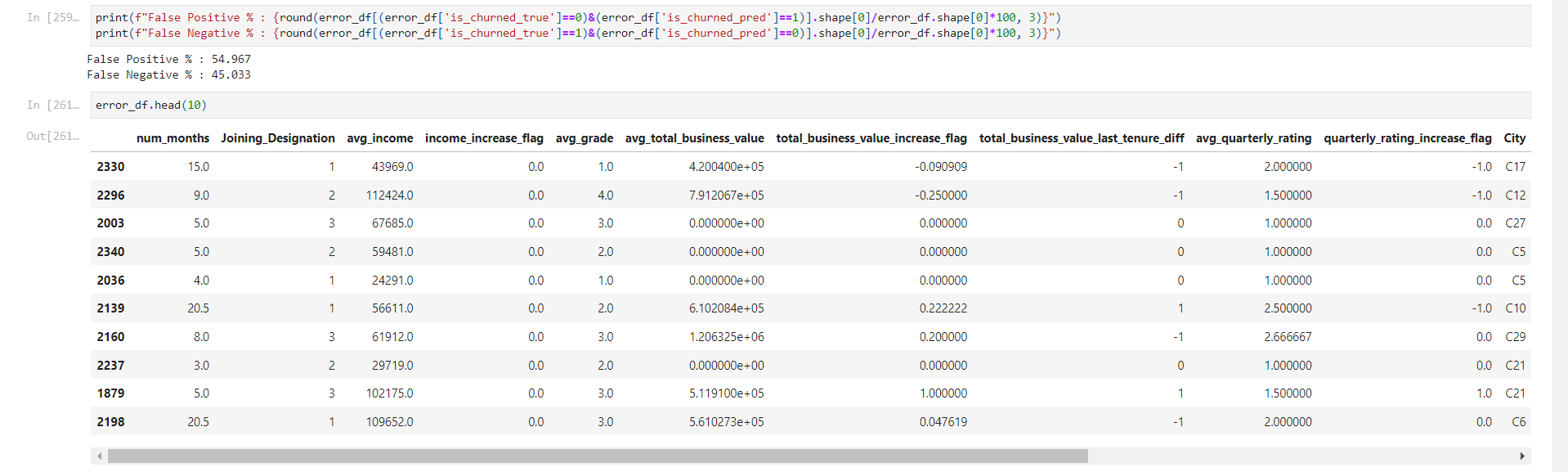
Figure 1.73: Table of Sample Rows from the Original Dataset

Figure 1.74: Table of Error DataFrame Sorted by Log Loss

Key Insights:

* A consolidated dataframe (df\_orig) is created containing all features, true labels, and model predictions.
* An error dataframe (error\_df) is created, containing only the misclassified instances.
* Log loss is calculated for each misclassified instance to quantify the magnitude of each error.
* The error dataframe is sorted by log loss, allowing us to examine the most severe misclassifications first.

#### Task 5.3: Error Analysis

Figure 1.75: Table of Sample Rows from the Error DataFrame

Key Insights:

* False Positive Rate: 54.967%
* False Negative Rate: 45.033%
* The model has a slightly higher tendency to predict churn when it doesn't occur (False Positives) than to miss actual churn cases (False Negatives).
* Examining the top 10 errors (highest log loss):
  + Many of these errors have avg\_total\_business\_value as 0, which the model interpreted as a sign of churn. This could be due to data collection issues or genuinely inactive drivers.
  + Most of these errors are for drivers who joined in 2019, a year with one of the highest churn rates. The model might be overgeneralizing based on the joining year.
  + Many of these errors have very low average quarterly ratings, which the model interpreted as likely to churn.
  + Some of these errors have relatively high incomes or long tenures, which the model typically associates with retention. These cases highlight the complexity of the churn prediction task.

#### Task 5.4: Implications of Error Analysis

* Data Quality: The presence of zero total business value for some drivers suggests a need for better data collection processes or handling of inactive periods.
* Contextual Factors: The model may be overemphasizing the importance of the joining year. There might be other contextual factors (e.g., economic conditions, company policies) that varied by year and affected churn rates.
* Performance Metrics: While low quarterly ratings are generally associated with churn, there are exceptions. This suggests a need for a more nuanced understanding of how performance metrics relate to churn.
* Complex Patterns: The presence of high-income or long-tenure drivers in the error set indicates that churn patterns are complex and not always captured by simple rules.
* Balance of Errors: The relatively balanced false positive and false negative rates suggest that the model isn't strongly biased towards either type of error. However, depending on the business context, it might be worth considering whether one type of error is more costly than the other.

### Analysis Part 6: Summary of Key Findings

#### 6.1 Data Characteristics and Preprocessing

* **Dataset Composition**: The initial dataset comprised 19,104 entries with 13 features, capturing various aspects of driver demographics, performance, and tenure.
* **Missing Data**: Significant missing data was observed in the 'Last\_Working\_Date' column (91.54%), which was later utilized to create our target variable 'churned'.
* **Feature Engineering**: Several new features were created, including 'num\_months', 'DOJ\_Year', 'DOJ\_Month', and various performance trend indicators (e.g., 'income\_increase\_flag', 'grade\_increase\_flag').

#### 6.2 Exploratory Data Analysis Insights

* **Tenure Distribution**: The average tenure (num\_months) was 8.02 months, with a right-skewed distribution indicating a high number of relatively new drivers.
* **Age Distribution**: Driver ages ranged from 21 to 58, with a mean of 33.15 years, showing a relatively young workforce.
* **Gender Balance**: The driver pool showed a slight male majority (58.97% male vs. 41.03% female).
* **Joining Patterns**: Most drivers (84.34%) joined between 2018-2020, with peak joining months in July and May.
* **Performance Metrics**: Average income, total business value, and quarterly ratings all showed right-skewed distributions with significant outliers.

#### 6.3 Bivariate Analysis Findings

1. **Tenure and Churn**: A negative correlation was observed between tenure and churn probability (PCC: -0.346).
2. **Age and Churn**: Age showed a weak negative correlation with churn (PCC: -0.056), suggesting limited influence.
3. **Gender and Churn**: Churn rates were similar across genders (Male: 67.52%, Female: 68.37%).
4. **City-wise Variation**: Significant variations in churn rates were observed across cities.
5. **Joining Year Impact**: Years 2018 and 2019 showed notably higher churn rates compared to other years.
6. **Income and Churn**: Drivers experiencing income increases showed significantly lower churn rates (6.82% vs 69.02%).

#### 6.4 Model Development and Performance

1. **Model Comparison**:
   1. Decision Tree: Test F1 Score = 0.891
   2. Random Forest: Test F1 Score = 0.928
   3. LightGBM: Test F1 Score = 0.913
2. **Best Model**: Random Forest with optimized threshold (0.55) achieved the best performance.
   1. Optimized F1 Scores: Train (0.965), CV (0.926), Test (0.911)
3. **Class Imbalance**: Applying class weight balancing slightly improved model performance and reduced overfitting.

#### 6.5 Feature Importance and Model Interpretation

1. **Top Predictive Features**:
   1. Average income
   2. Average total business value
   3. Number of months (tenure)
   4. Total business value increase flag
   5. Month of joining
2. **Error Analysis**:
   1. False Positive Rate: 54.967%
   2. False Negative Rate: 45.033%
   3. Model tends to overpredict churn for drivers with zero total business value and those who joined in high-churn years.

#### 6.6 Key Takeaways for Driver Retention

1. **Financial Factors**: Income levels and business value generation are the strongest predictors of driver retention.
2. **Tenure Importance**: Longer-tenured drivers are less likely to churn, emphasizing the importance of early retention efforts.
3. **Performance Trends**: Positive trends in income, grade, and business value are associated with lower churn rates.
4. **Geographical Variation**: Significant differences in churn rates across cities indicate the need for location-specific retention strategies.
5. **Joining Year Impact**: High variability in churn rates across joining years suggests the influence of broader economic or company-specific factors on retention.

## Insights and Recommendations

Based on our comprehensive analysis of the ride-sharing service provider's driver data, we have uncovered several key insights and developed corresponding recommendations:

### Insights

* **Tenure and Churn Relationship**:
  + There is a significant negative correlation between tenure (number of months with the company) and churn probability (PCC: -0.346).
  + As drivers spend more time with the company, their likelihood of churning decreases.
  + The average tenure is 8.02 months, with a right-skewed distribution indicating a high number of relatively new drivers.
* **Financial Factors**:
  1. Average income and total business value are the top two most important features in predicting churn.
  2. Drivers experiencing income increases show significantly lower churn rates (6.82% vs 69.02% for those without increases).
  3. The average total business value has a strong negative correlation with churn probability.
* **Demographic Insights**:
  + Age has a weak negative correlation with churn (PCC: -0.056), suggesting limited influence on retention.
  + Gender does not significantly impact churn rates (Male: 67.52%, Female: 68.37% churn rate).
  + Significant variations in churn rates exist across different cities, indicating the importance of location-specific factors.
* **Joining Patterns and Churn**:
  + The year and month of joining are important predictors of churn.
  + Years 2018 and 2019 show notably higher churn rates compared to other years.
  + Seasonal variations in churn rates are observed based on the joining month.
* **Performance Metrics**:
  + Higher quarterly ratings are associated with lower churn probabilities.
  + Positive trends in grades and business value are linked to improved retention.
* **Career Progression**:
  + Drivers who have been promoted during their tenure show lower churn rates.
  + Churn rate generally decreases as drivers move up in designation.
* **Model Performance**:
  + The Random Forest model with optimized threshold (0.55) achieved the best performance (Test F1 Score: 0.911).
  + The model shows a slight tendency to overpredict churn (False Positive Rate: 54.967%, False Negative Rate: 45.033%).

### Recommendations

* **Early Retention Strategies**:
  + Implement targeted retention programs for drivers in their first 8 months, as this period shows the highest churn risk.
  + Develop a comprehensive onboarding and support system for new drivers to increase early engagement and performance.
* **Financial Incentives and Stability**:
  1. Design a structured income growth program to incentivize long-term commitment.
  2. Implement a bonus system tied to consistent or increasing total business value generation.
  3. Consider introducing a minimum income guarantee for the initial months to provide financial stability to new drivers.
* **Performance-Based Retention**:
  + Develop a system to closely monitor and support drivers with declining quarterly ratings or total business value.
  + Offer performance improvement plans and additional resources to at-risk drivers.
  + Implement a recognition program for consistently high-performing drivers.
* **Location-Specific Strategies**:
  + Conduct in-depth market analysis in high-churn cities to identify and address local challenges.
  + Develop city-specific retention strategies, considering factors such as local competition, traffic conditions, and customer demographics.
  + Replicate successful practices from low-churn cities across other locations where applicable.
* **Seasonal Recruitment and Retention**:
  + Adjust recruitment strategies based on seasonal churn patterns observed in joining months.
  + Develop targeted retention campaigns for drivers joining during high-churn months.
  + Consider offering seasonal bonuses or incentives during traditionally high-churn periods.
* **Career Development and Progression**:
  + Create a clear career progression path for drivers, with transparent criteria for advancement.
  + Offer training and development opportunities to facilitate promotions and grade improvements.
  + Communicate the benefits and success stories of long-term drivers to motivate newer recruits.
* **Personalized Retention Approaches**:
  + Utilize the predictive model to identify drivers at high risk of churning.
  + Develop personalized retention strategies based on individual driver characteristics and performance metrics.
  + Implement a regular feedback system to understand and address driver concerns proactively.
* **Data Quality and Model Improvement**:
  + Enhance data collection processes, particularly for total business value and performance metrics.
  + Regularly retrain and update the predictive model with new data to maintain its accuracy.
  + Consider incorporating additional external factors (e.g., local economic indicators, competitive landscape) to improve prediction accuracy.
* **Long-Term Retention Focus**:
  + Develop programs that specifically target and reward long-term drivers (e.g., those with tenure > 12 months).
  + Create a mentorship program pairing experienced drivers with newcomers to foster a sense of community and improve overall retention.