Final Report (Lab-Style Redo)

Course Title: Machine Learning

Section: _____

Team Members:

- Mohamed Mostafa : 23101594
- Marwan Khaled : 23101599
- Mohamed Adel : 23101899

Dataset: Top_10000_Movies.csv

Figures exported: 24

Section 1: Problem Domain

We analyze a dataset of movies to study popularity and engagement signals (votes),
ratings behavior, and basic relationships between movie attributes.
We then build ML models to predict high engagement and a numeric outcome (when available).

Section 2: Project Summary

Workflow:

- Load + audit + clean the dataset (types, missing values, duplicates)
- EDA with distributions, correlations, and groupby analysis
- Scoring model (weighted rating) to reduce small-sample bias
- ML: 4 lab classifiers (LogReg, DecisionTree, GaussianNB, KNN) + Linear Regression
- Export all plots to PDF and ZIP

Section 3: Source Code

Source code is provided in the notebook:

- Ml_Project_LabStyle_Redo.ipynb

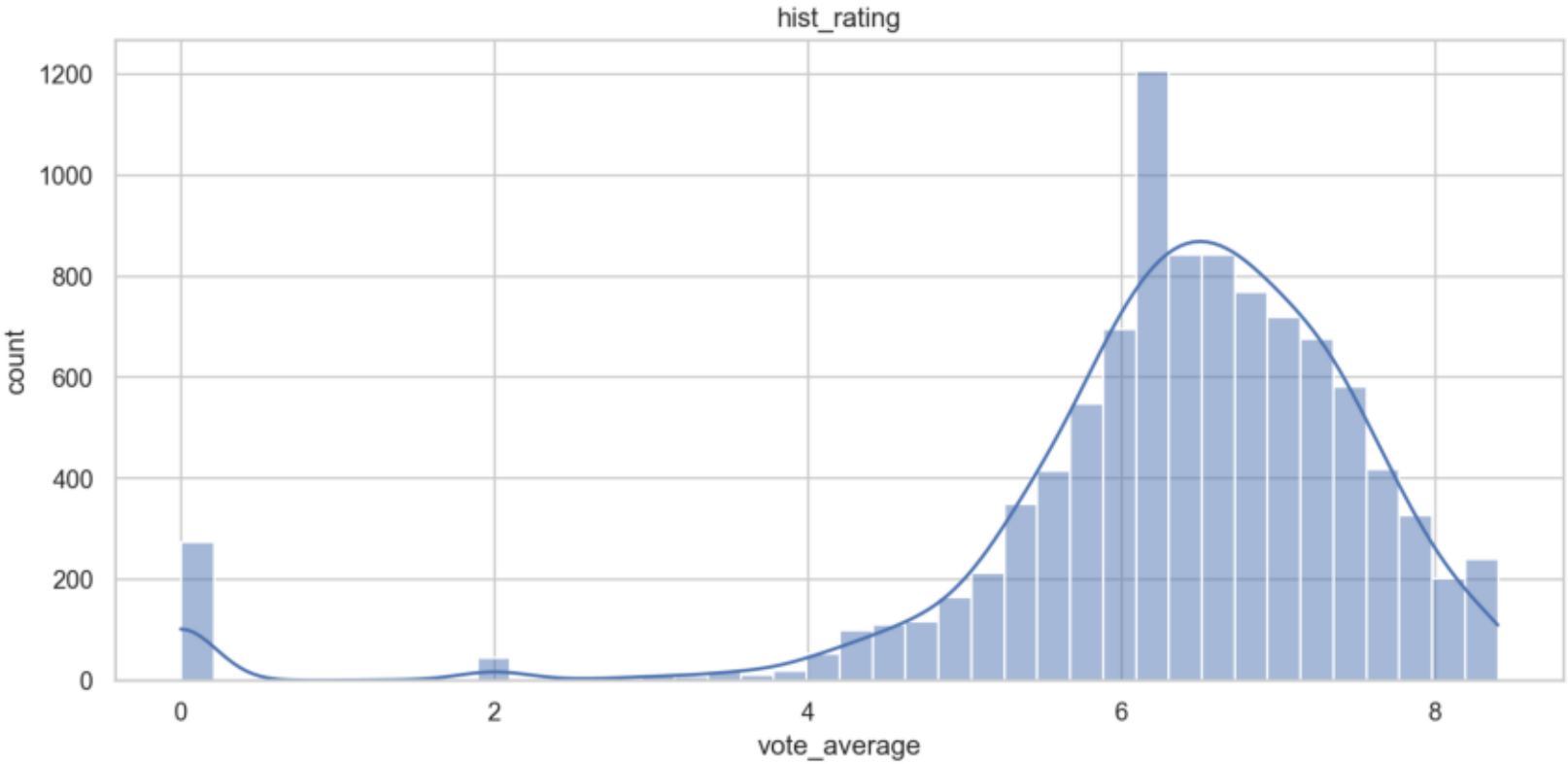
Key implementation points:

- Cleaning & feature engineering with defensive column detection
- Scikit-learn pipelines + ColumnTransformer preprocessing
- Standard metrics (Accuracy/Precision/Recall/F1, confusion matrix, ROC when available)
- Linear Regression metrics (MAE/RMSE/R2)

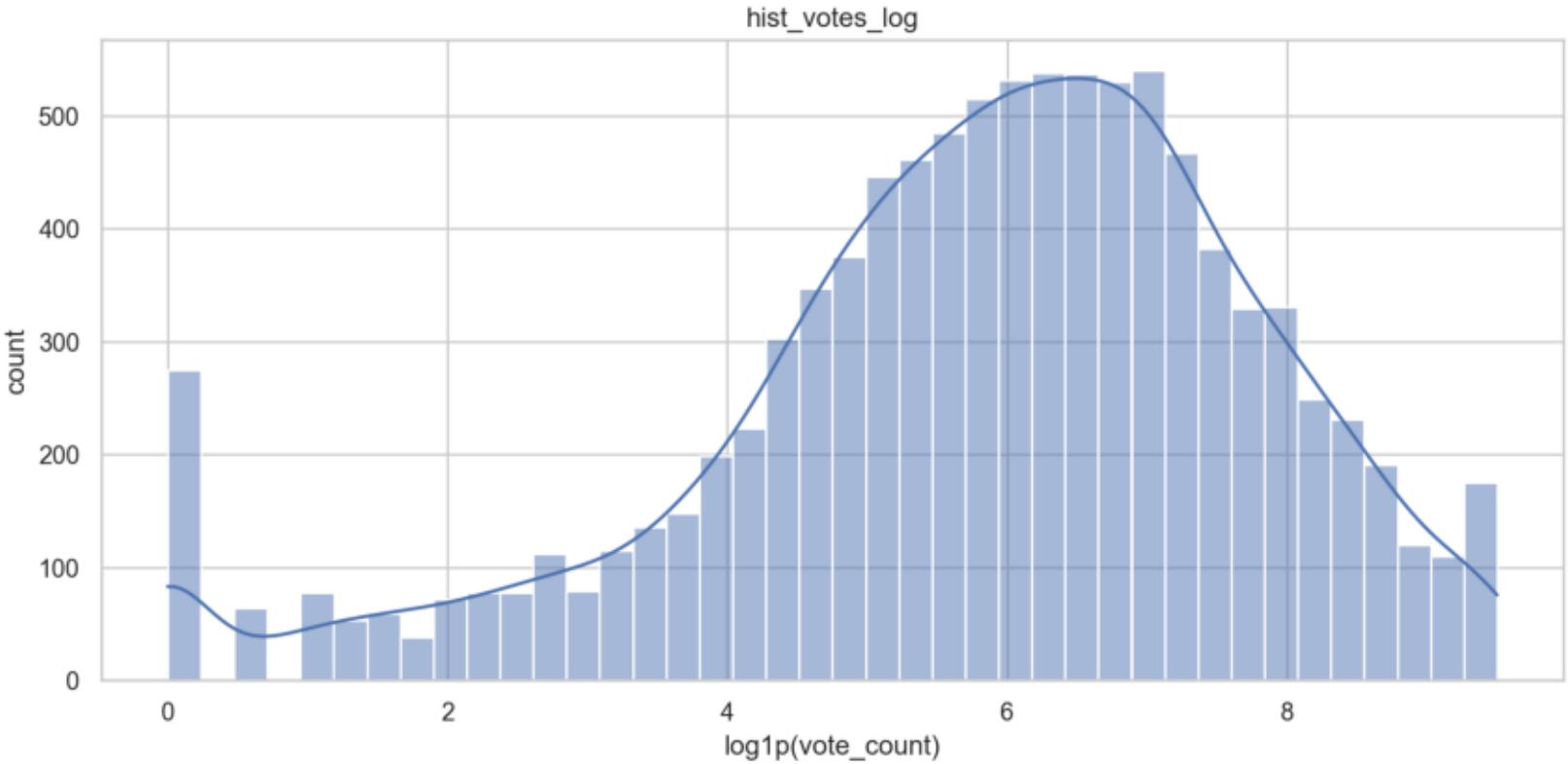
Section 4: Visualization Snapshots

The following pages include every figure saved during EDA + ML evaluation.

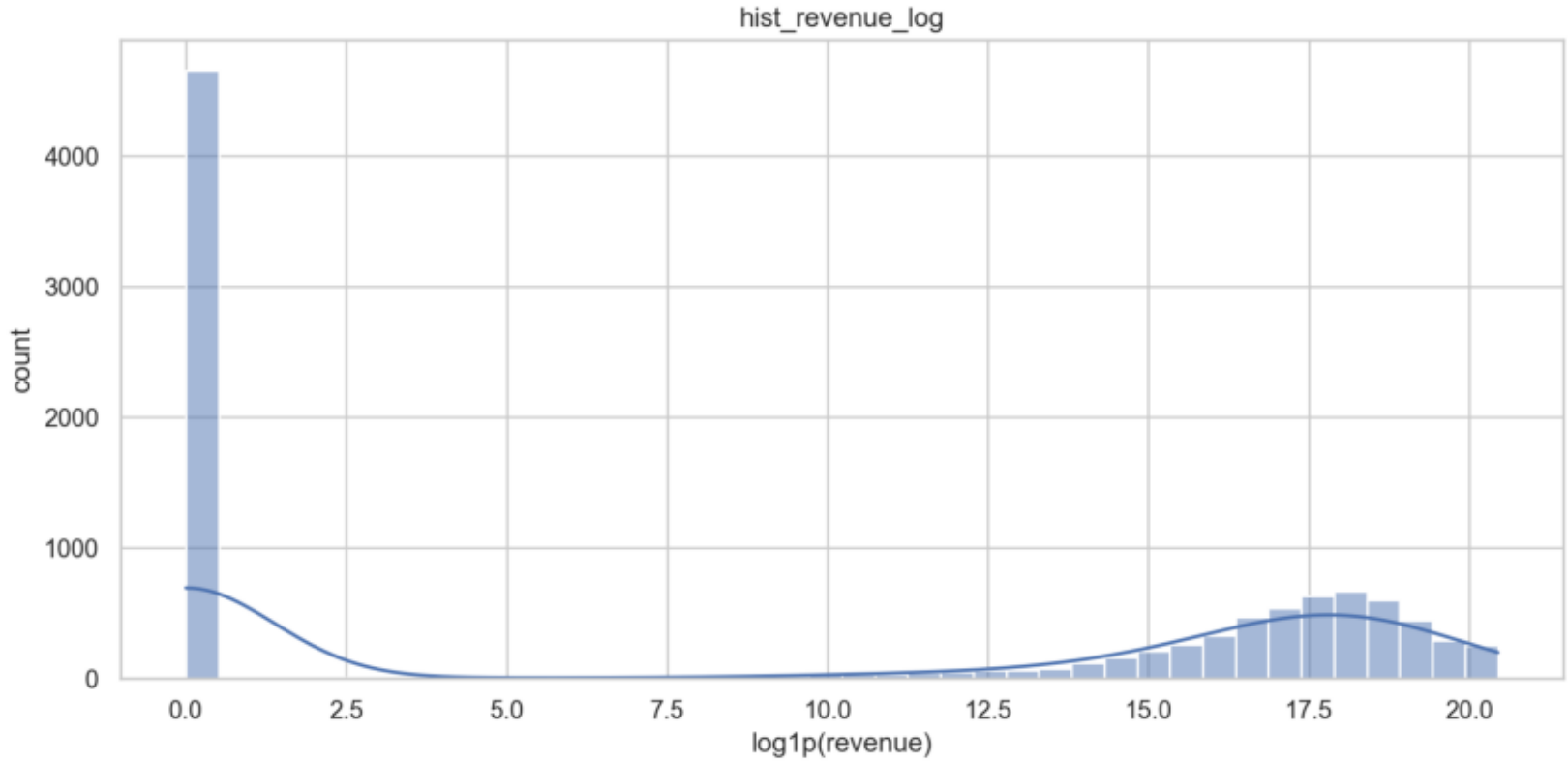
hist_rating


hist_rating
count
1200
1000
800
600
400
200
0
0
2
4
6
8
vote_average

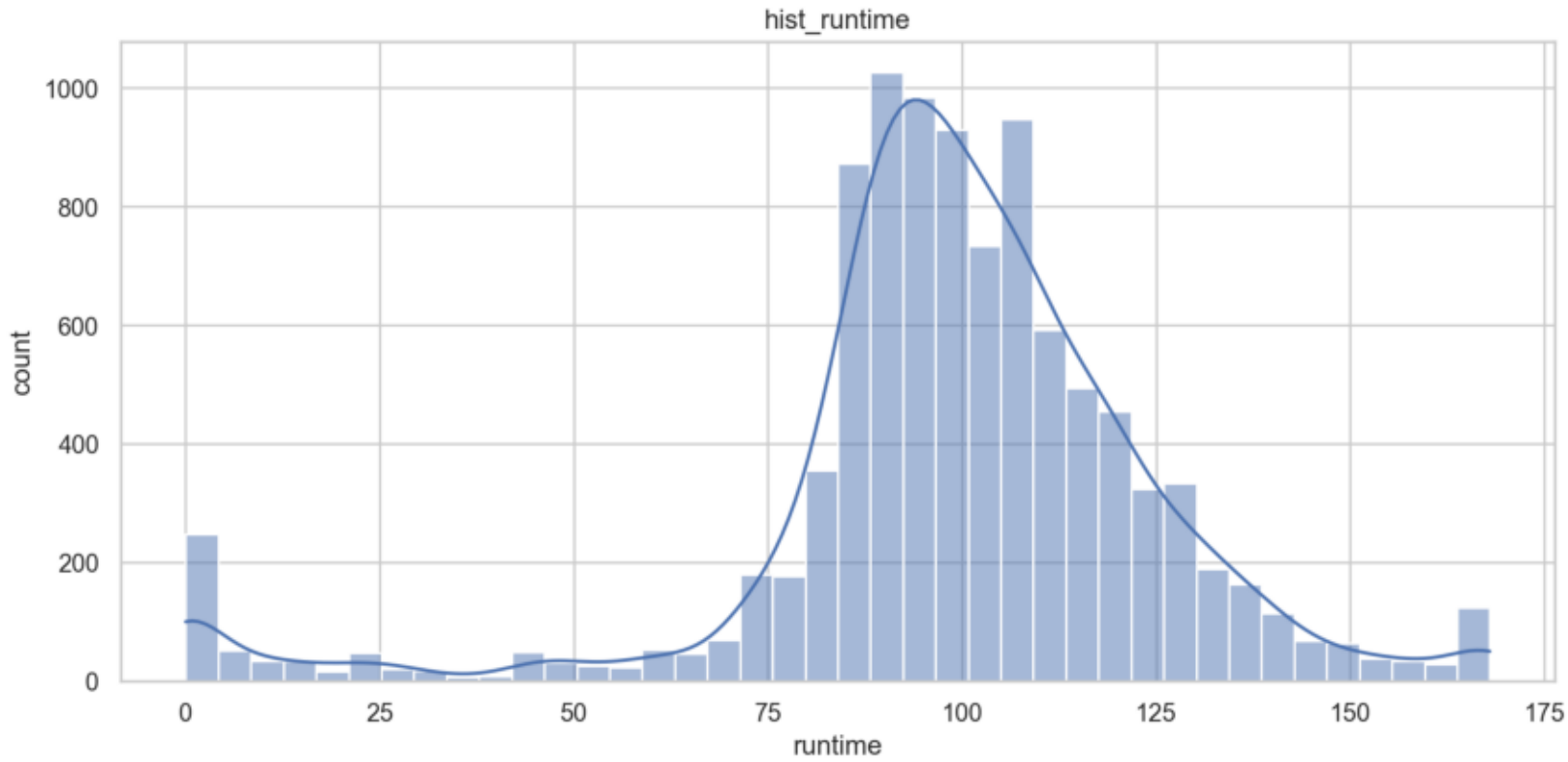
hist_votes_log

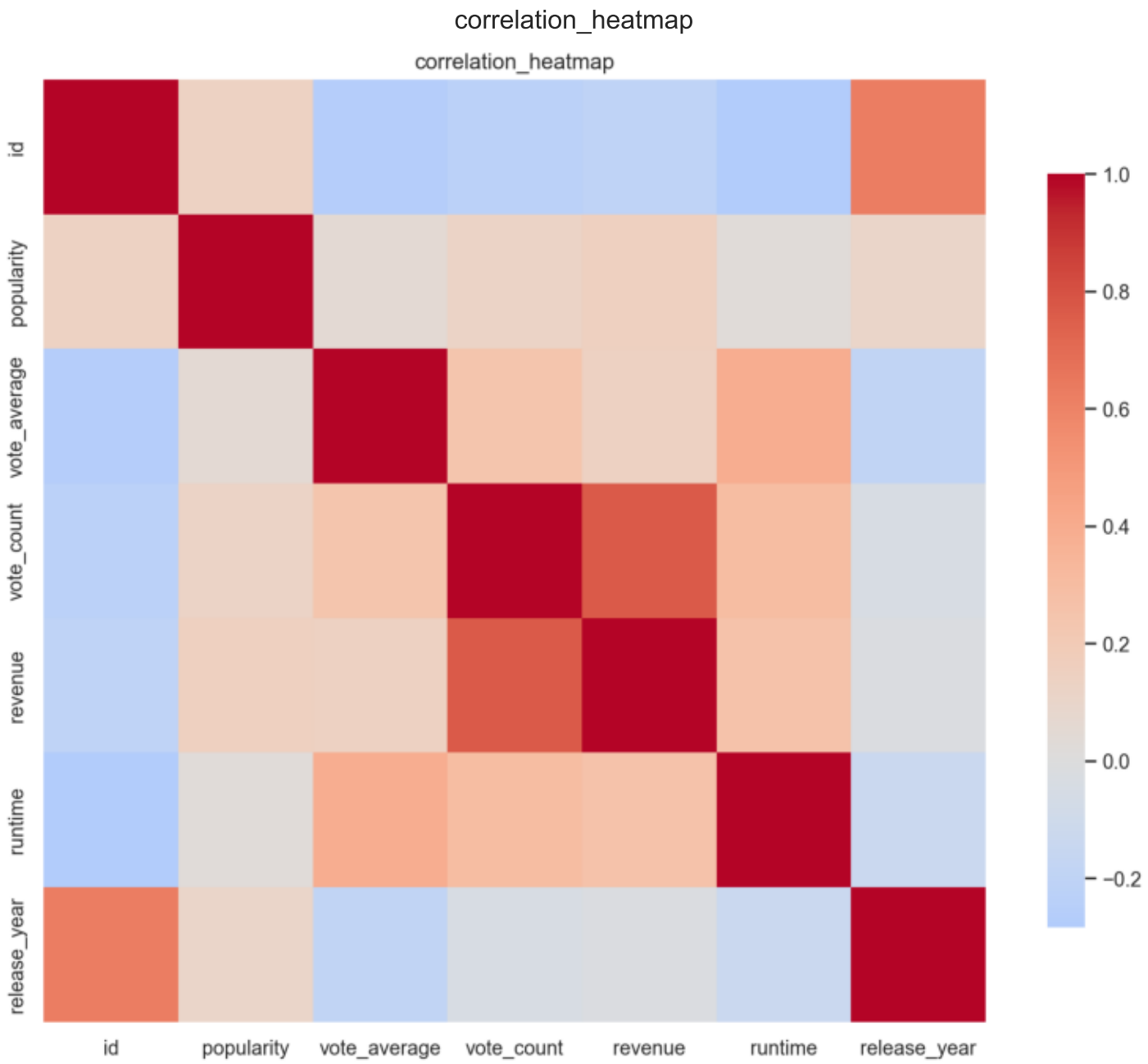


hist_revenue_log

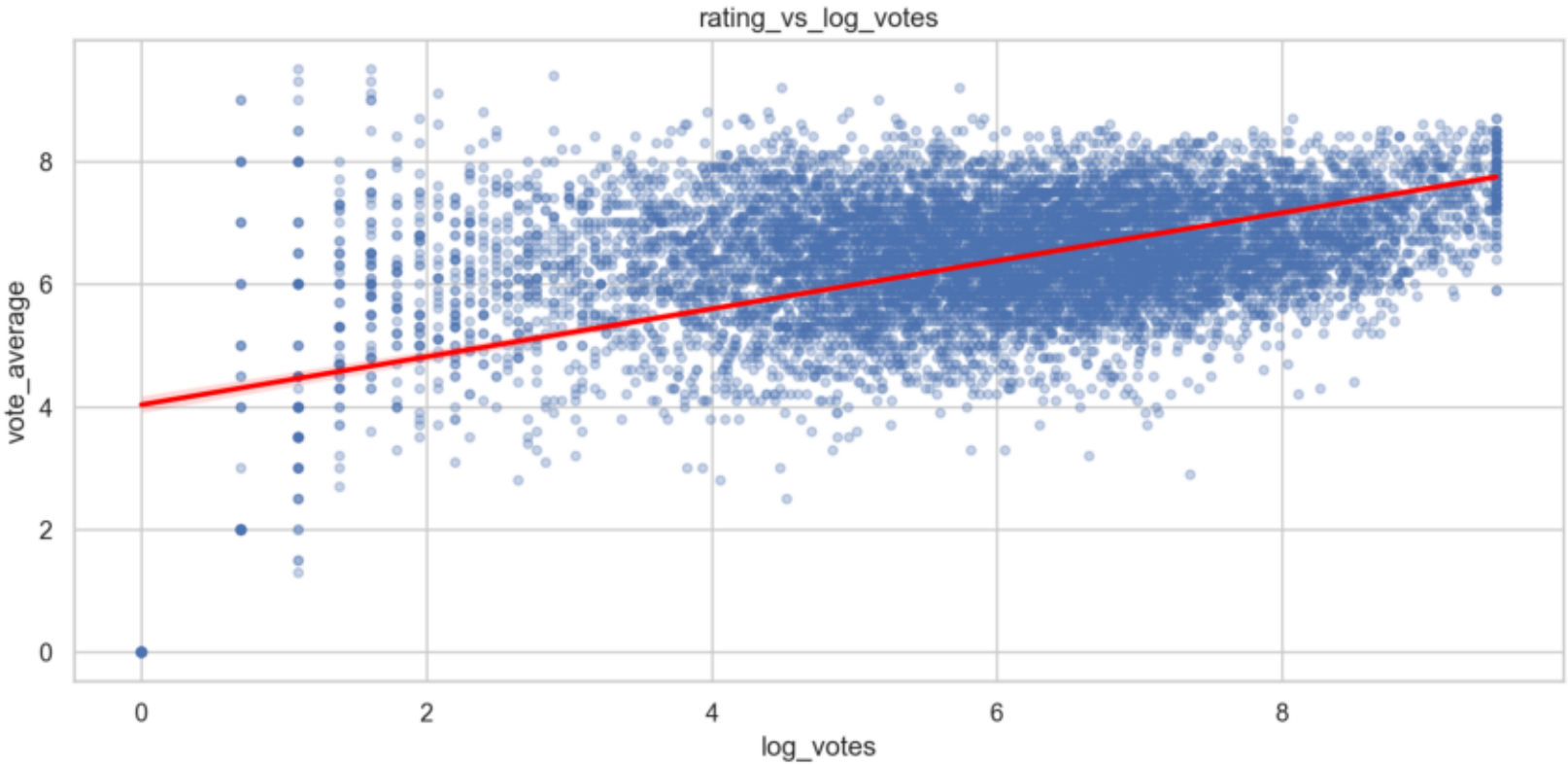


hist_runtime

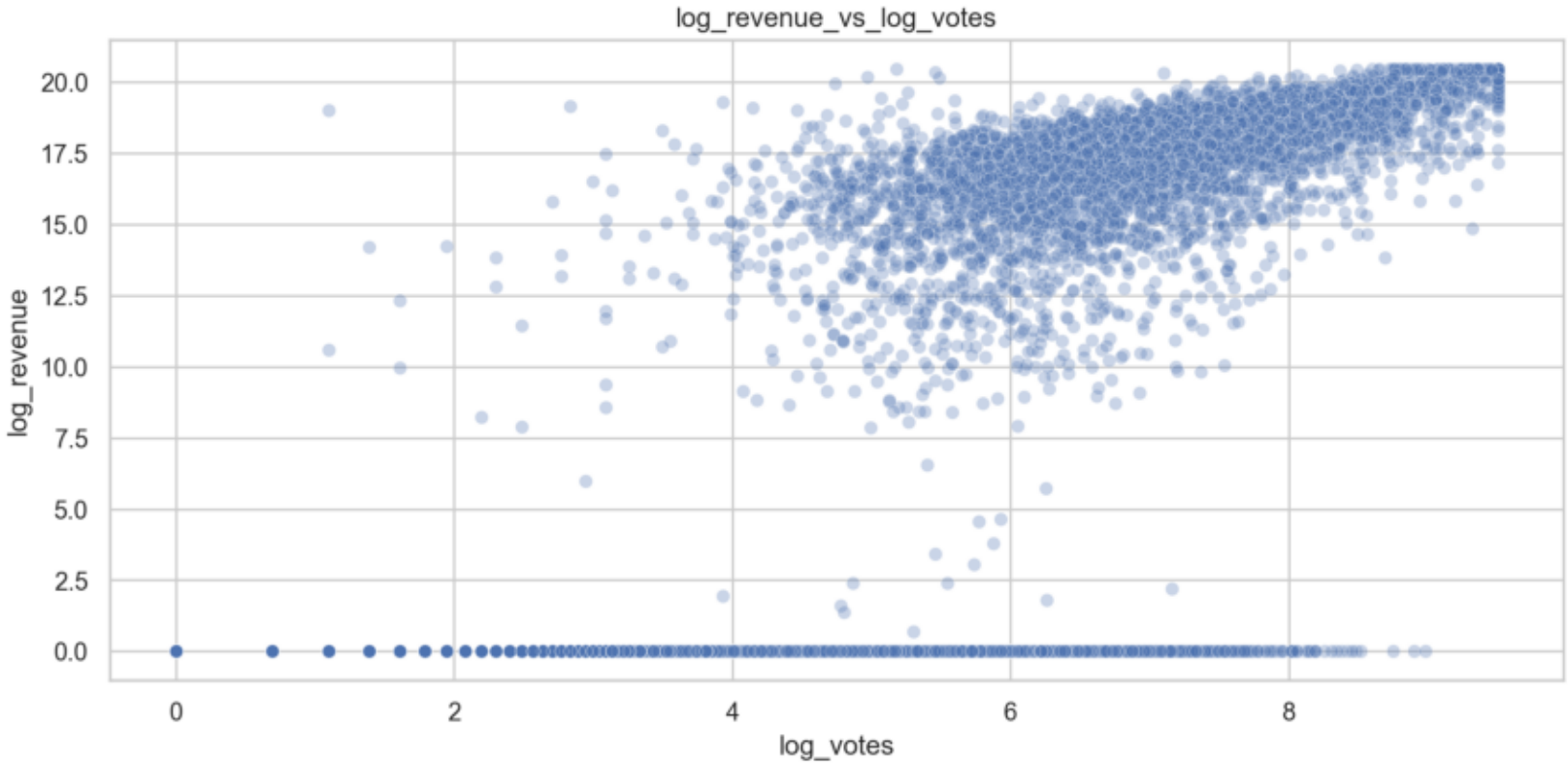



correlation_heatmap
correlation_heatmap
id
popularity
vote_average
vote_count
revenue
runtime
release_year
id
popularity
vote_average
vote_count
revenue
runtime
release_year
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

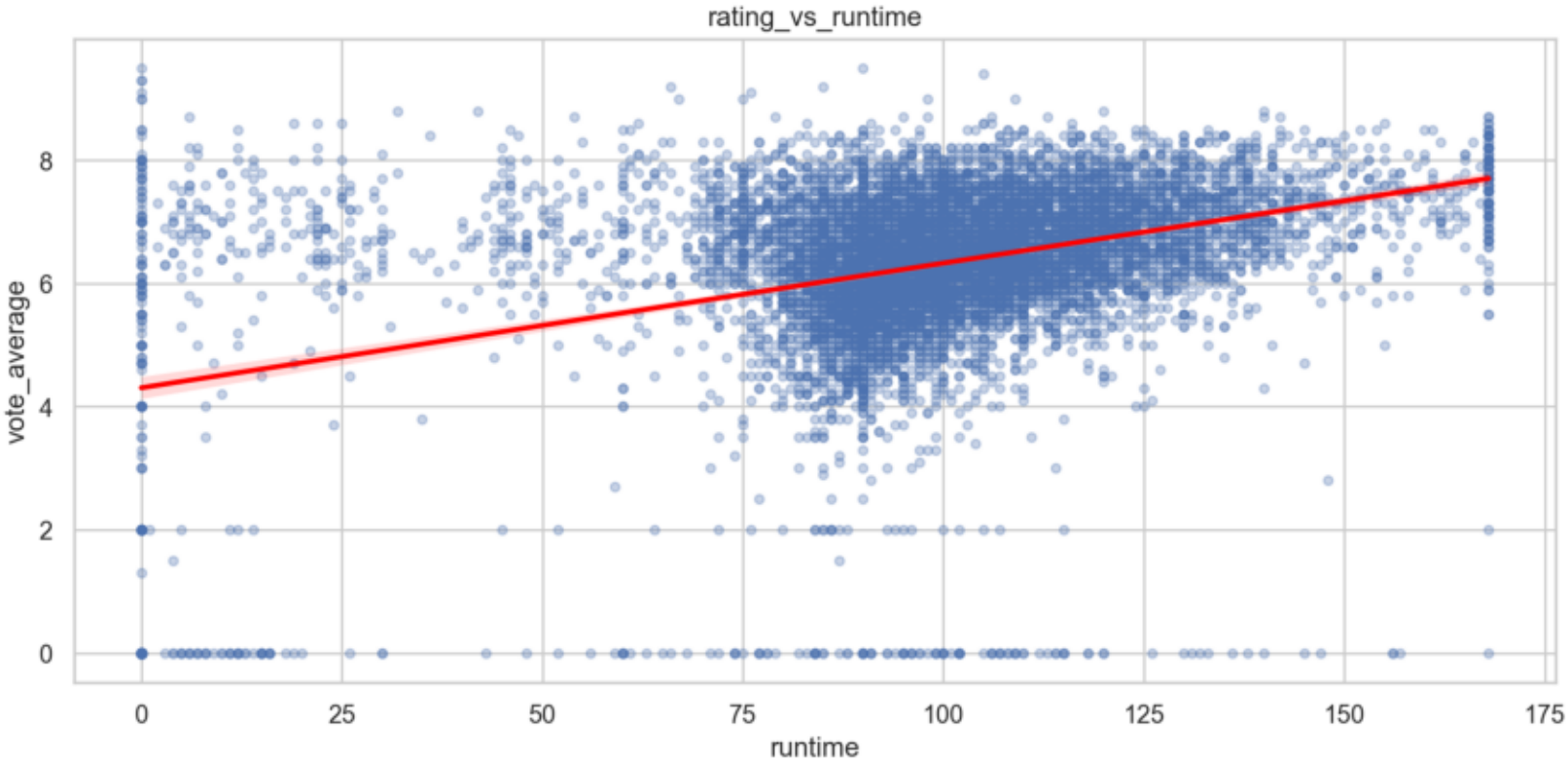
scatter_rating_vs_log_votes



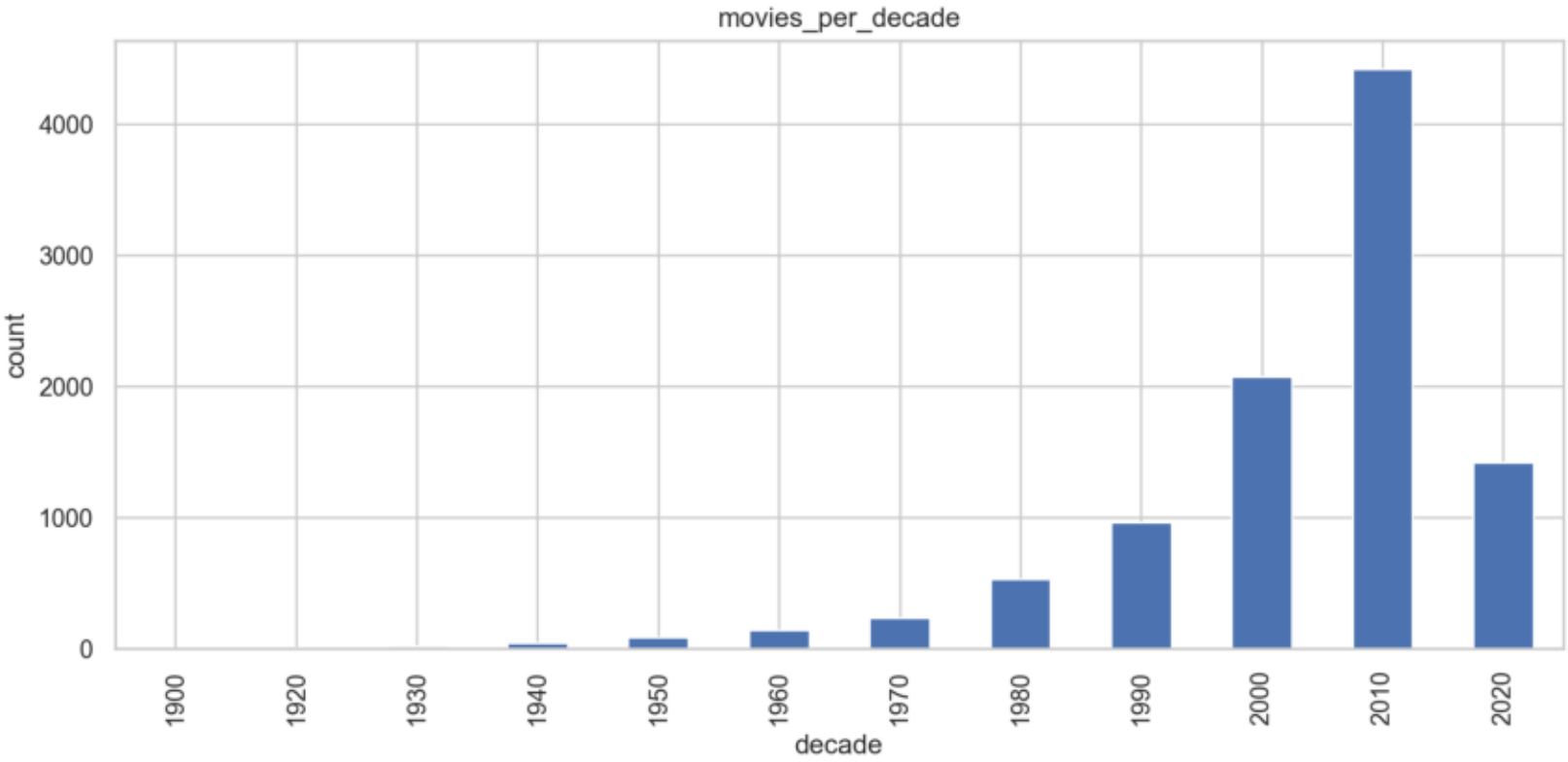
scatter_log_revenue_vs_log_votes



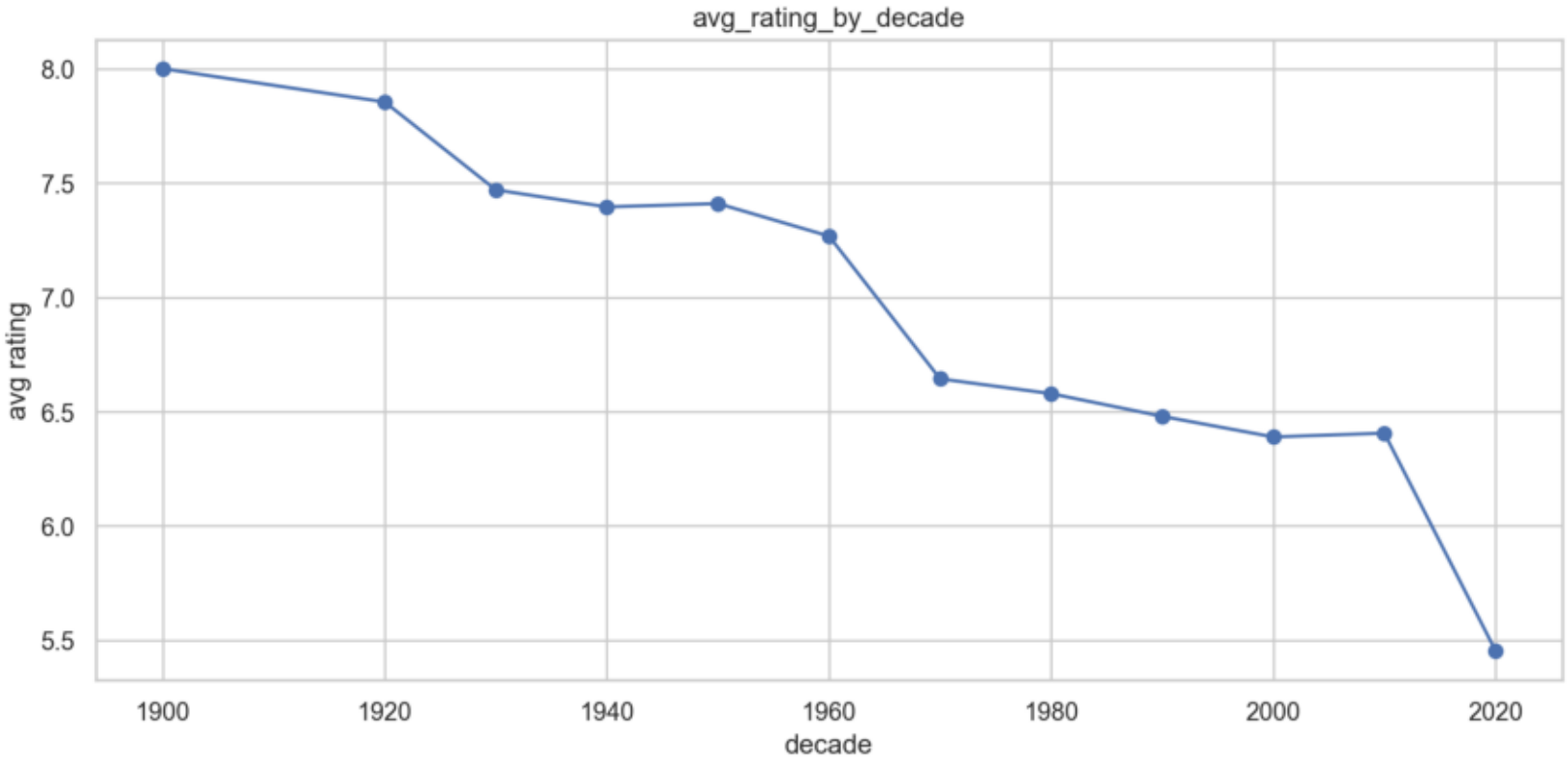
scatter_rating_vs_runtime

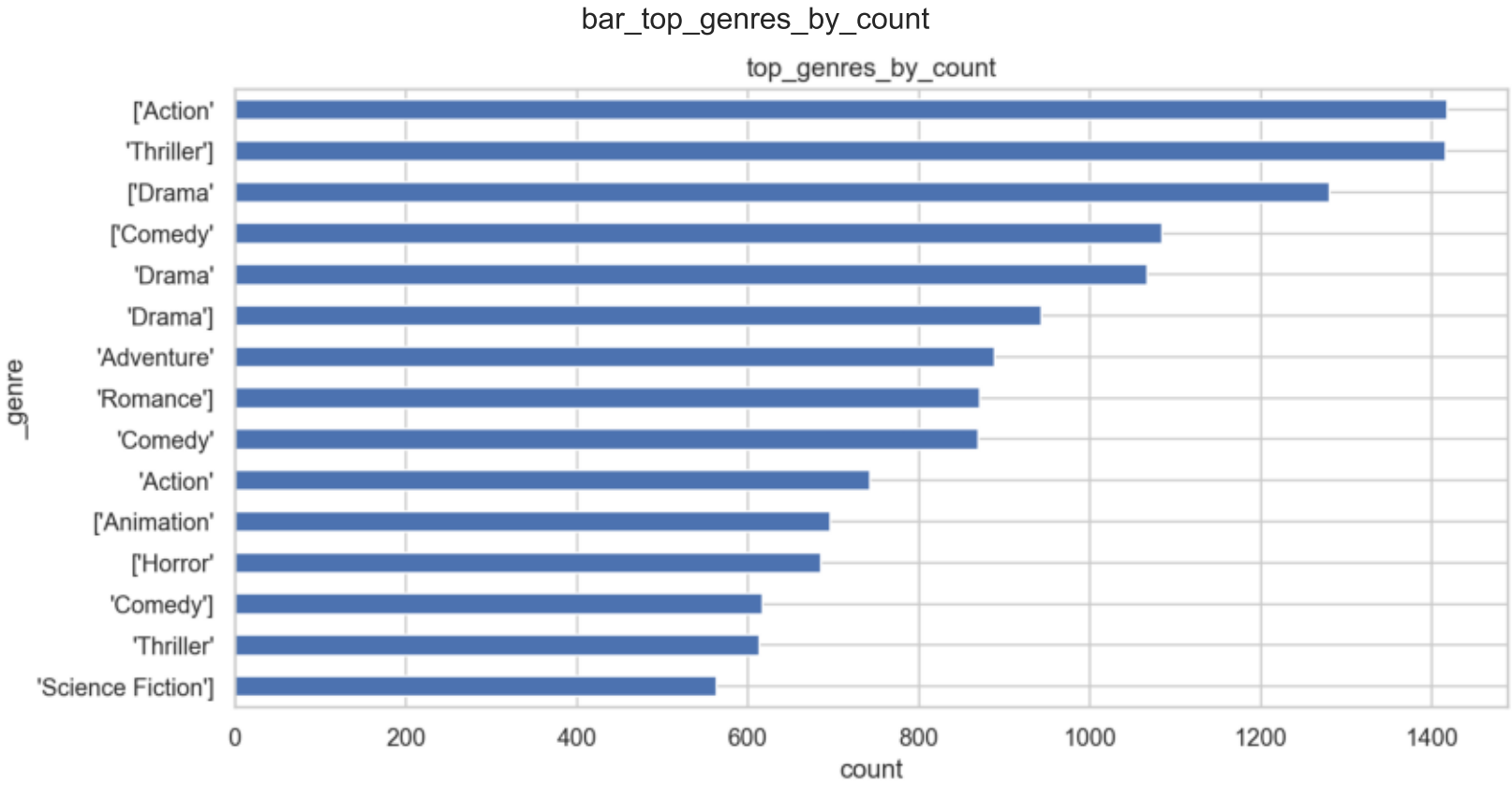


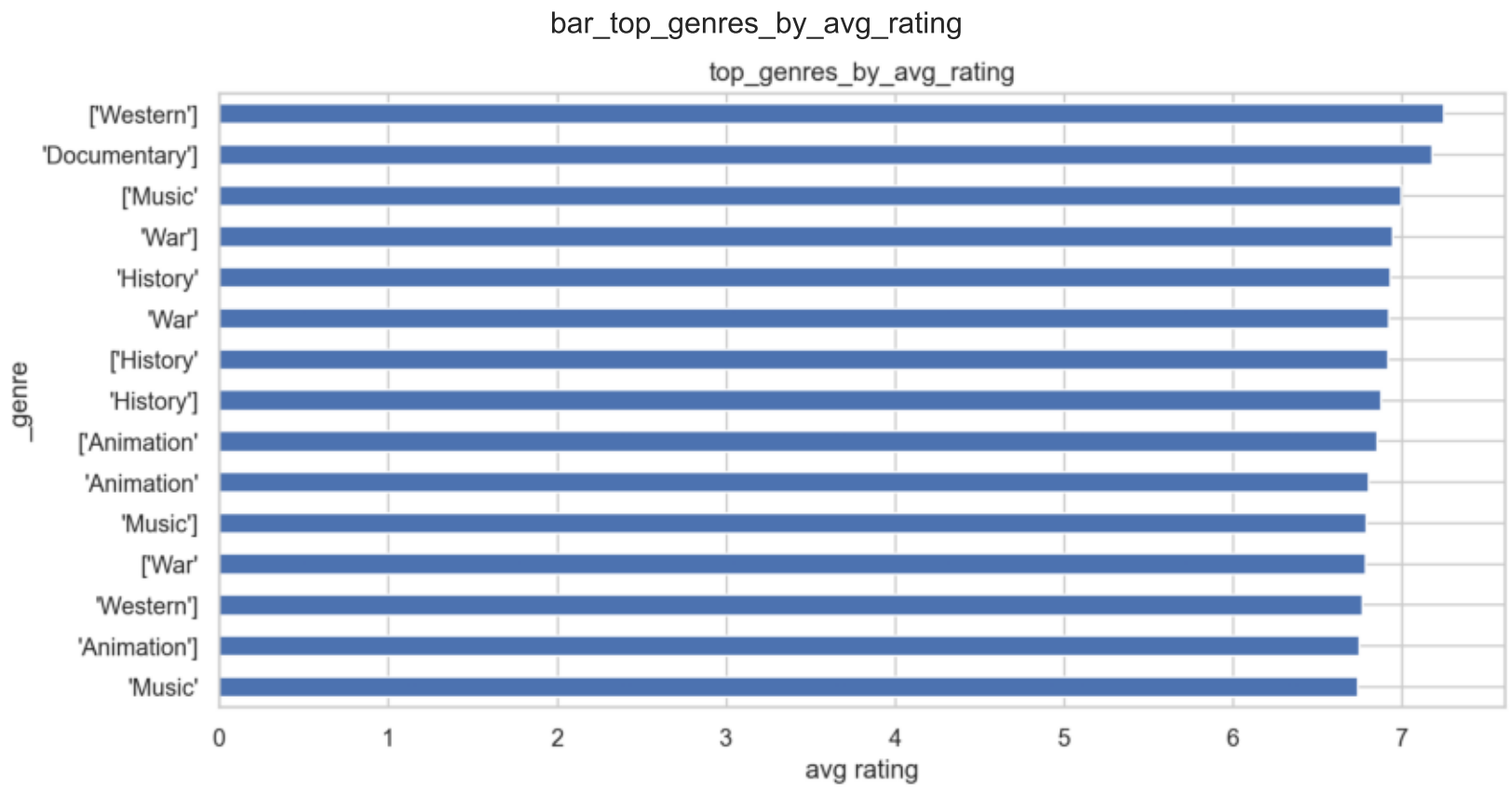
bar_movies_per_decade

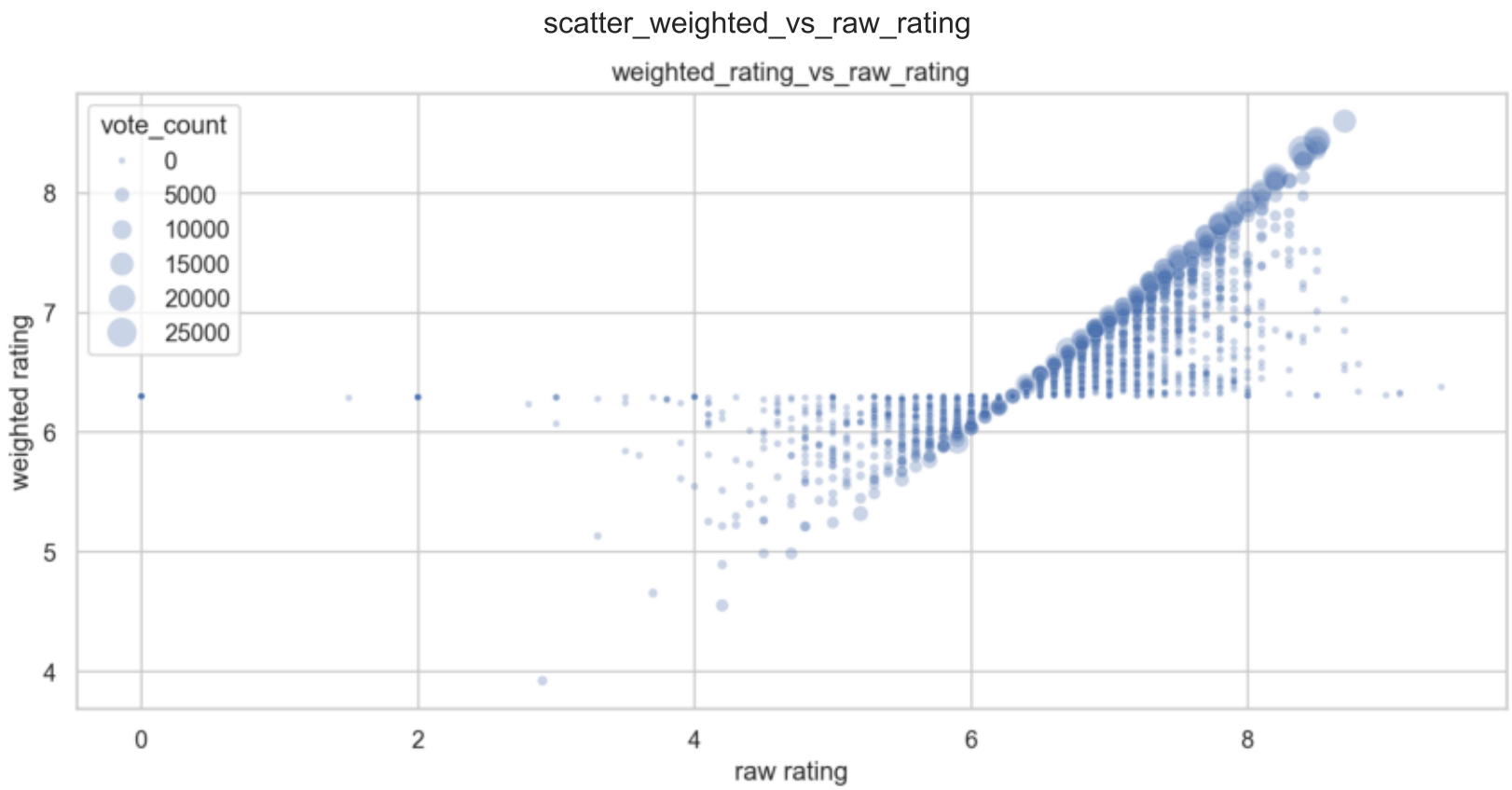


line_avg_rating_by_decade

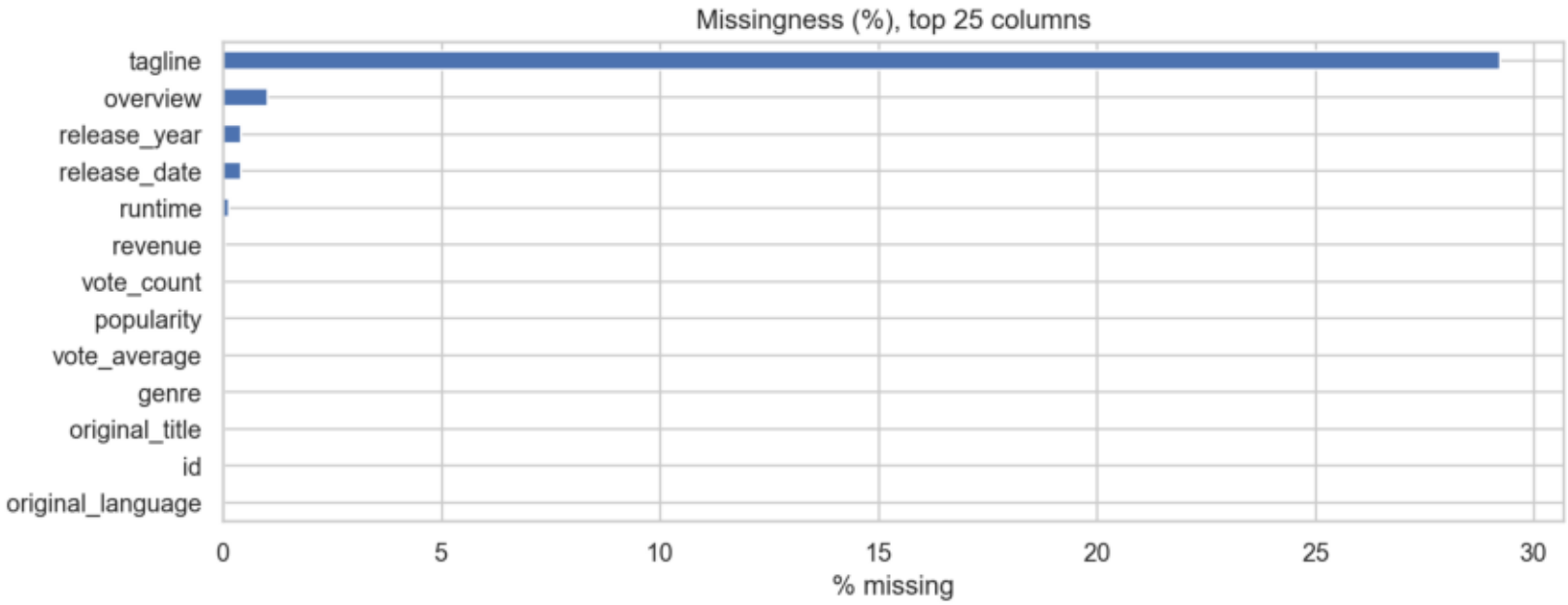



bar_top_genres_by_count
top_genres_by_count
['Action'
'Thriller']
['Drama'
['Comedy'
'Drama'
'Drama']
'Adventure'
'Romance']
'Comedy'
'Action'
['Animation'
['Horror'
'Comedy']
'Thriller'
'Science Fiction']
_genre
0
200
400
600
800
1000
1200
1400
count

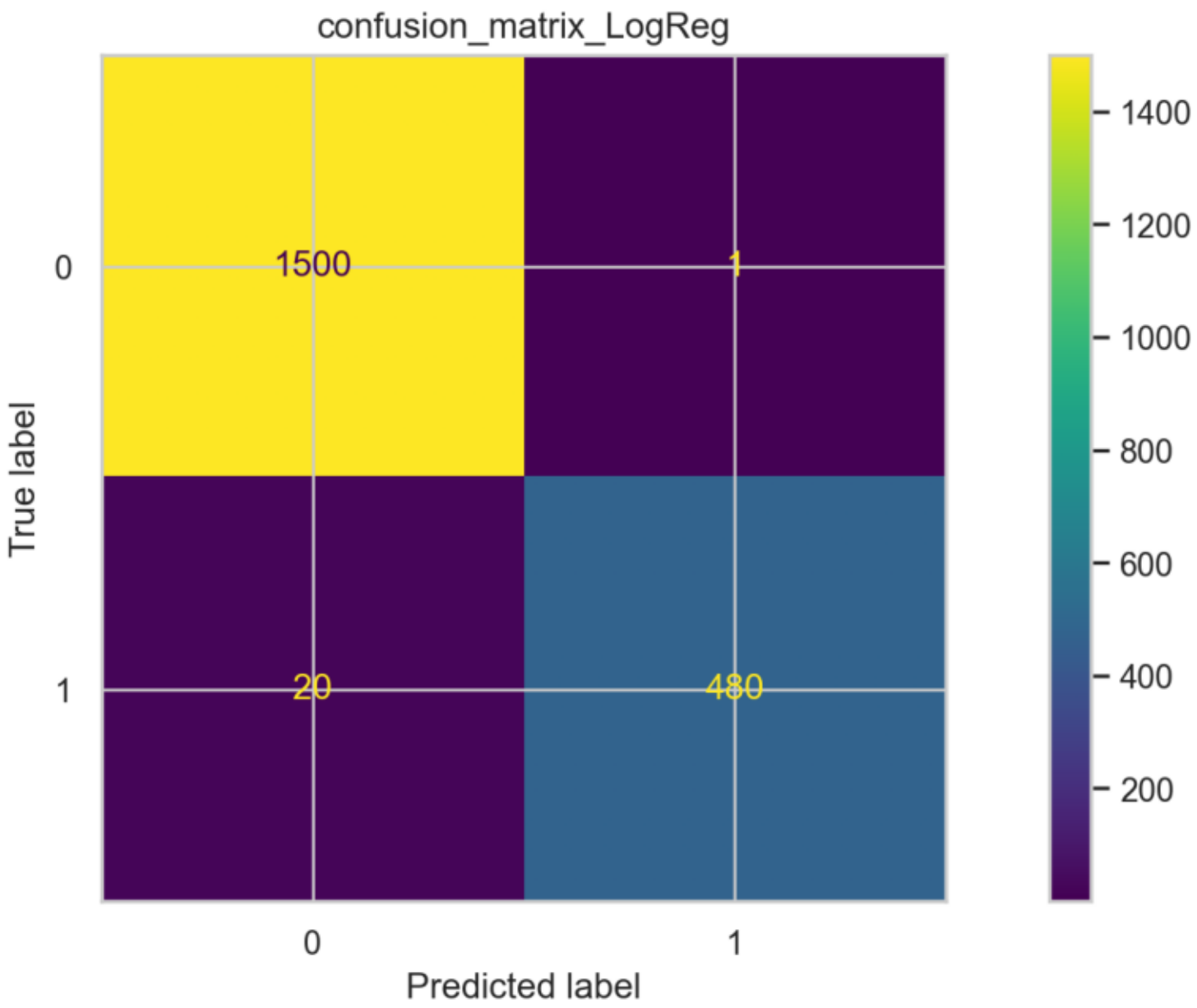

bar_top_genres_by_avg_rating
top_genres_by_avg_rating
['Western']
'Documentary']
['Music'
'War']
'History'
'War'
['History'
'History']
['Animation'
'Animation'
'Music']
['War'
'Western']
'Animation']
'Music'
_genre
0
1
2
3
4
5
6
7
avg rating


scatter_weighted_vs_raw_rating
weighted_rating_vs_raw_rating
vote_count
0
5000
10000
15000
20000
25000
weighted rating
raw rating
4
5
6
7
8
0
2
4
6
8

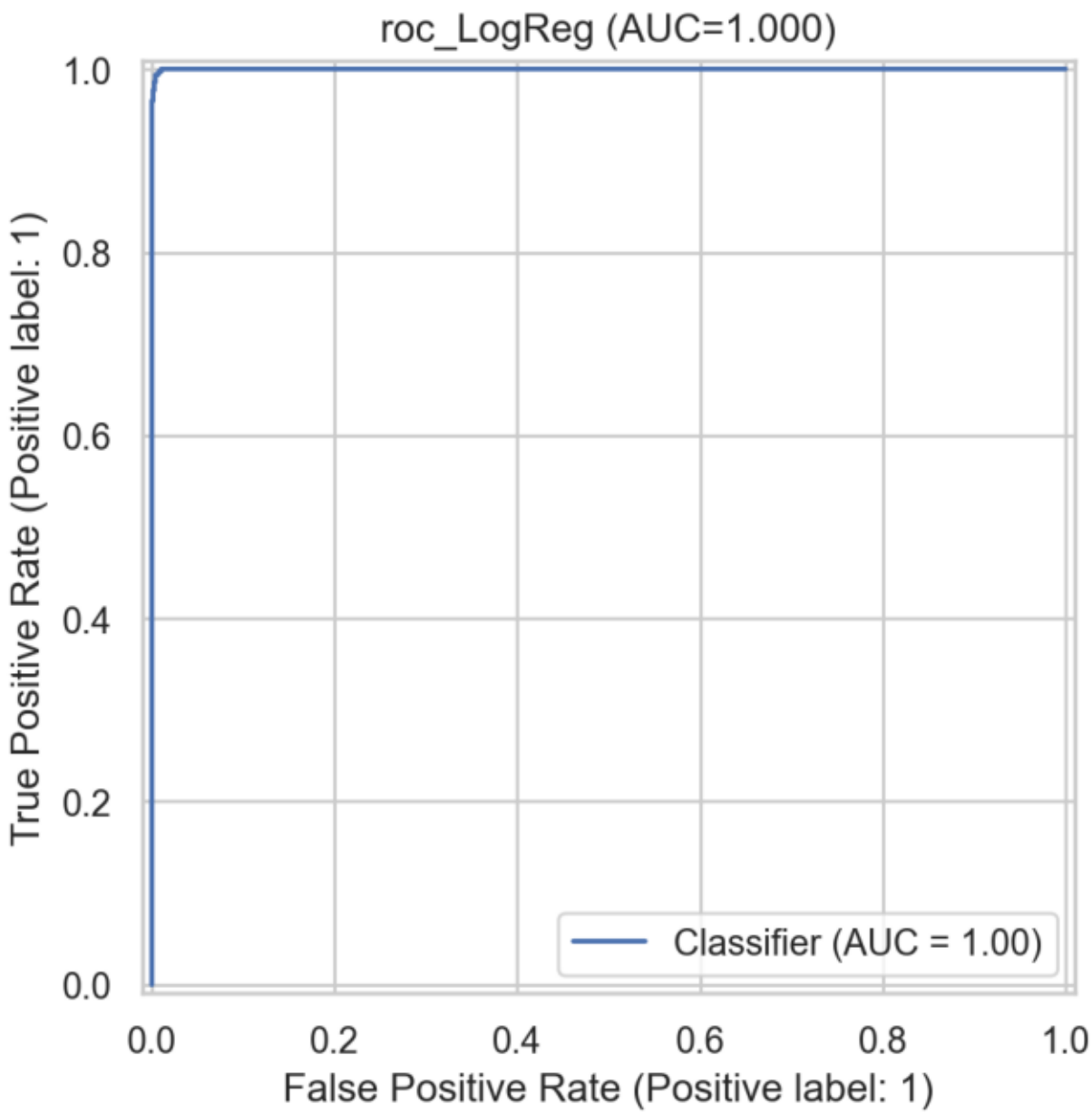
bar_missingness_top25



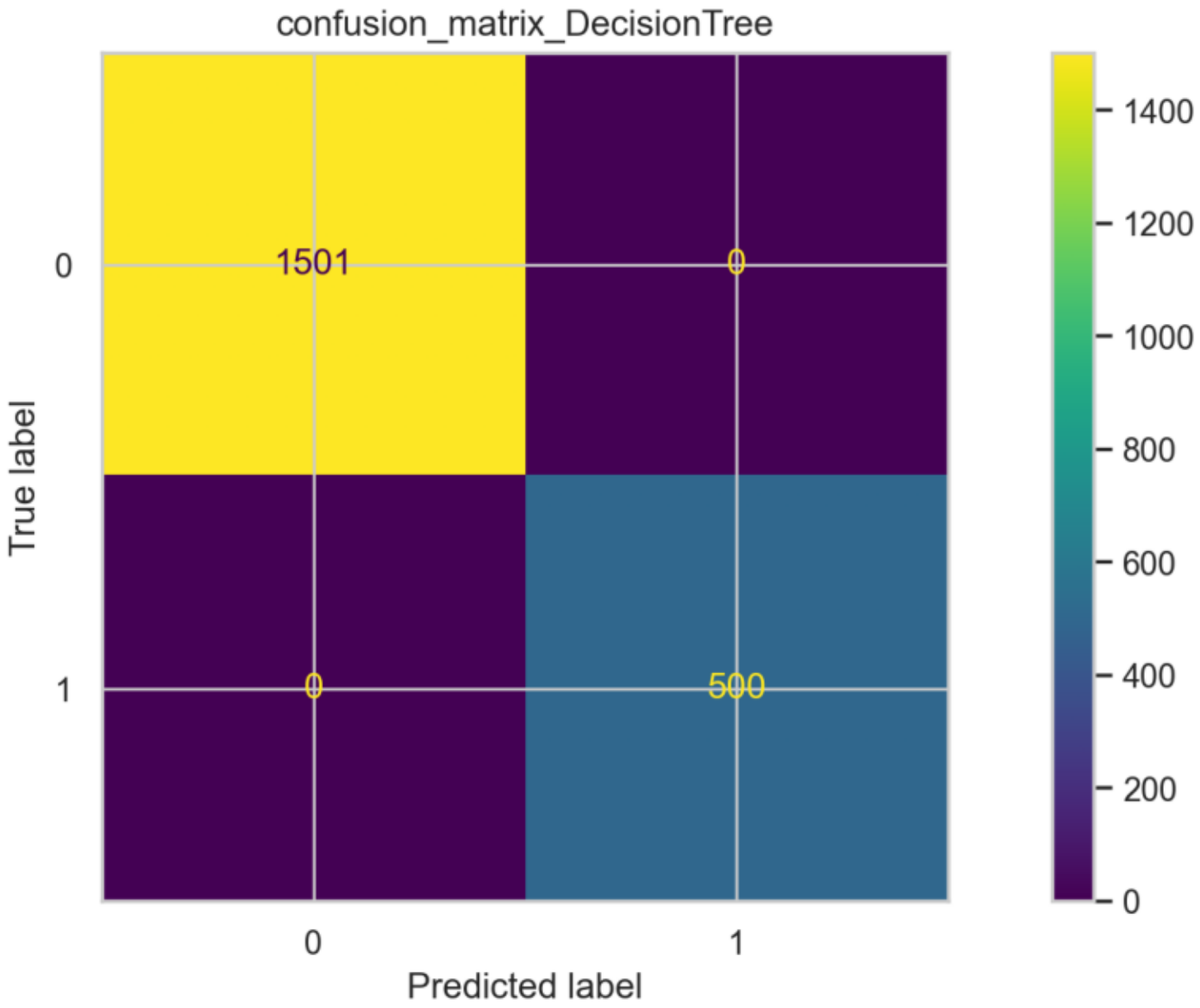
confusion_matrix_LogReg



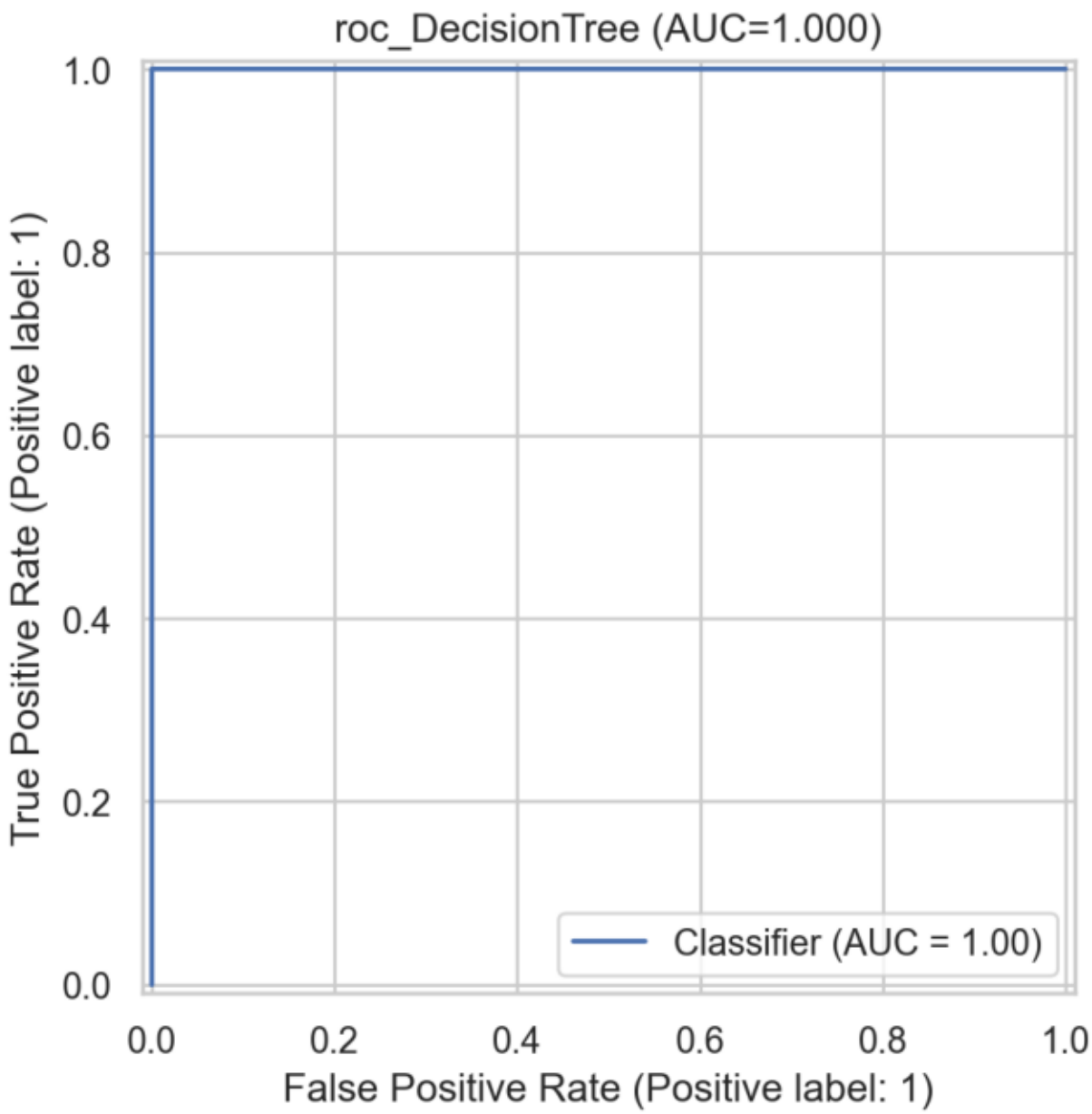
roc_LogReg


roc_LogReg (AUC=1.000)
True Positive Rate (Positive label: 1)
False Positive Rate (Positive label: 1)
0.0
0.2
0.4
0.6
0.8
1.0
Classifier (AUC = 1.00)

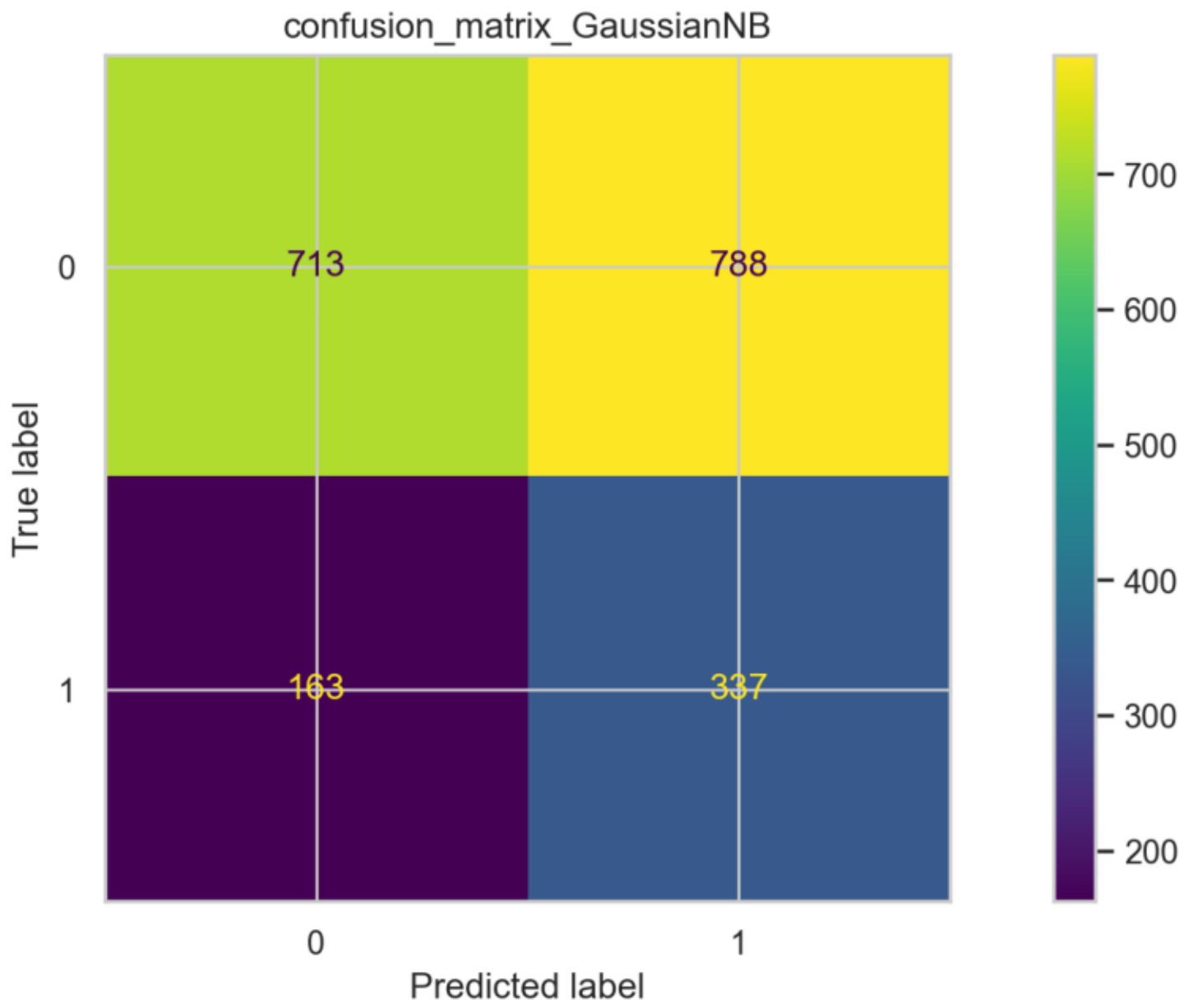
confusion_matrix_DecisionTree



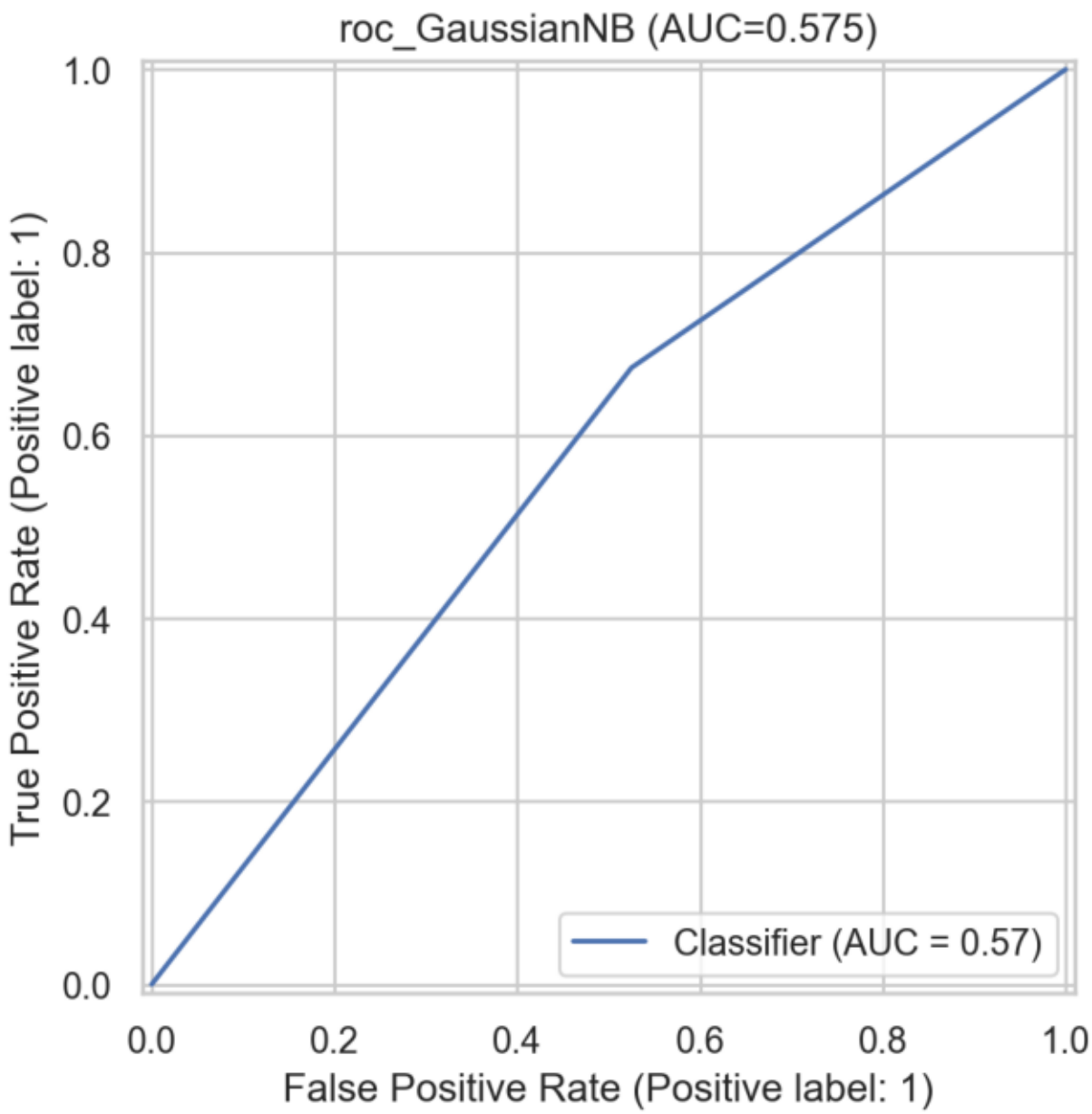
roc_DecisionTree



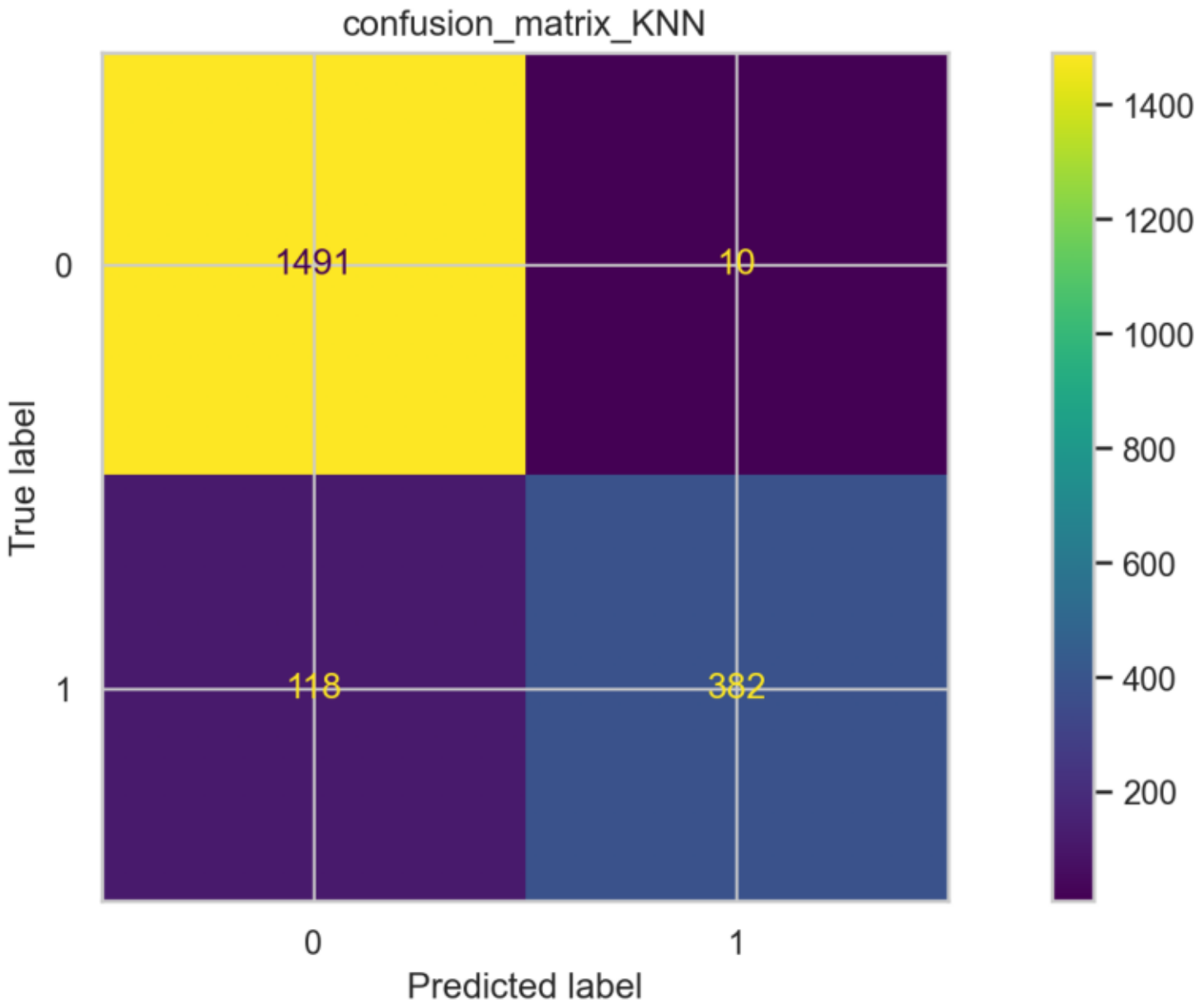
confusion_matrix_GaussianNB

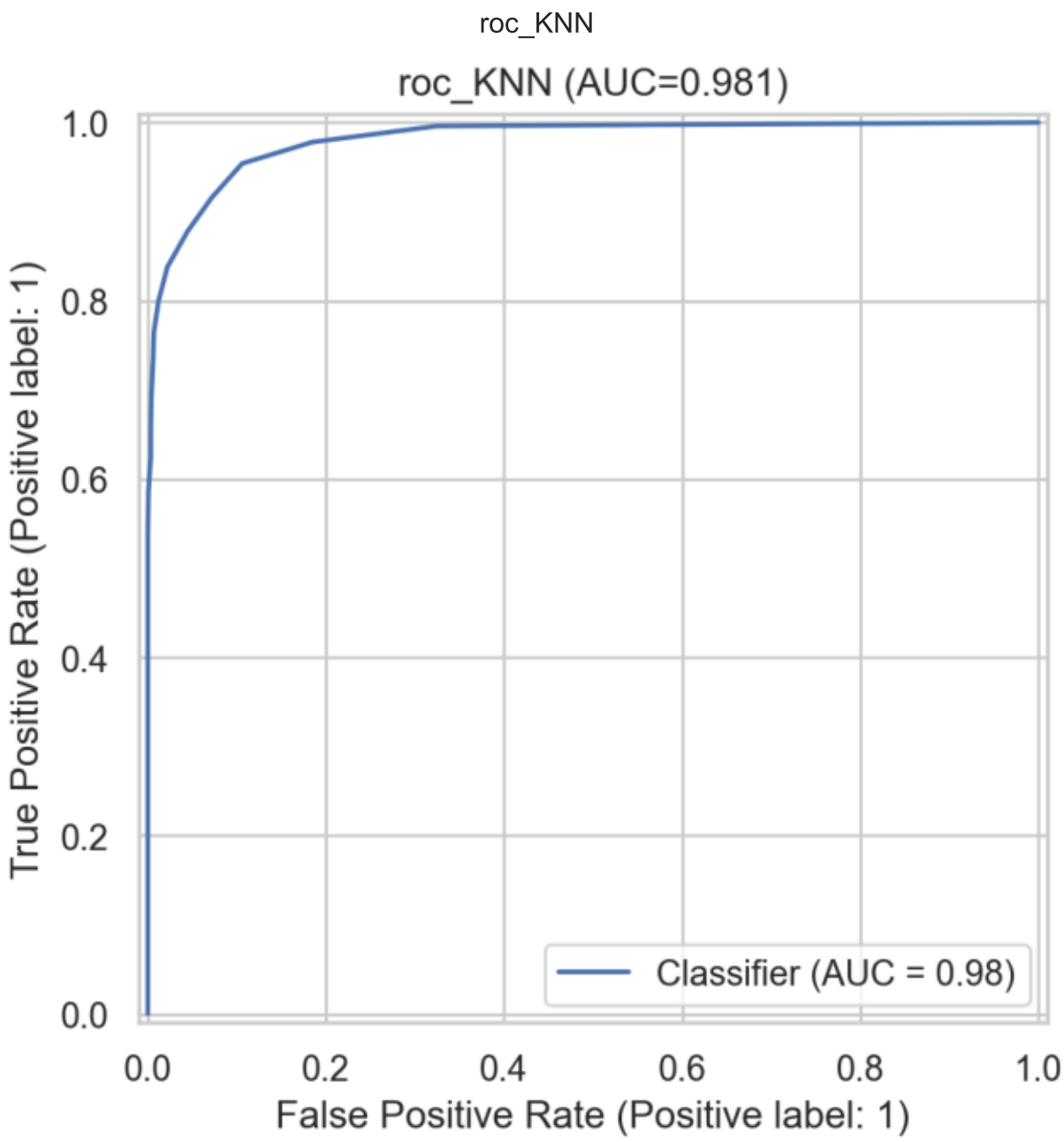


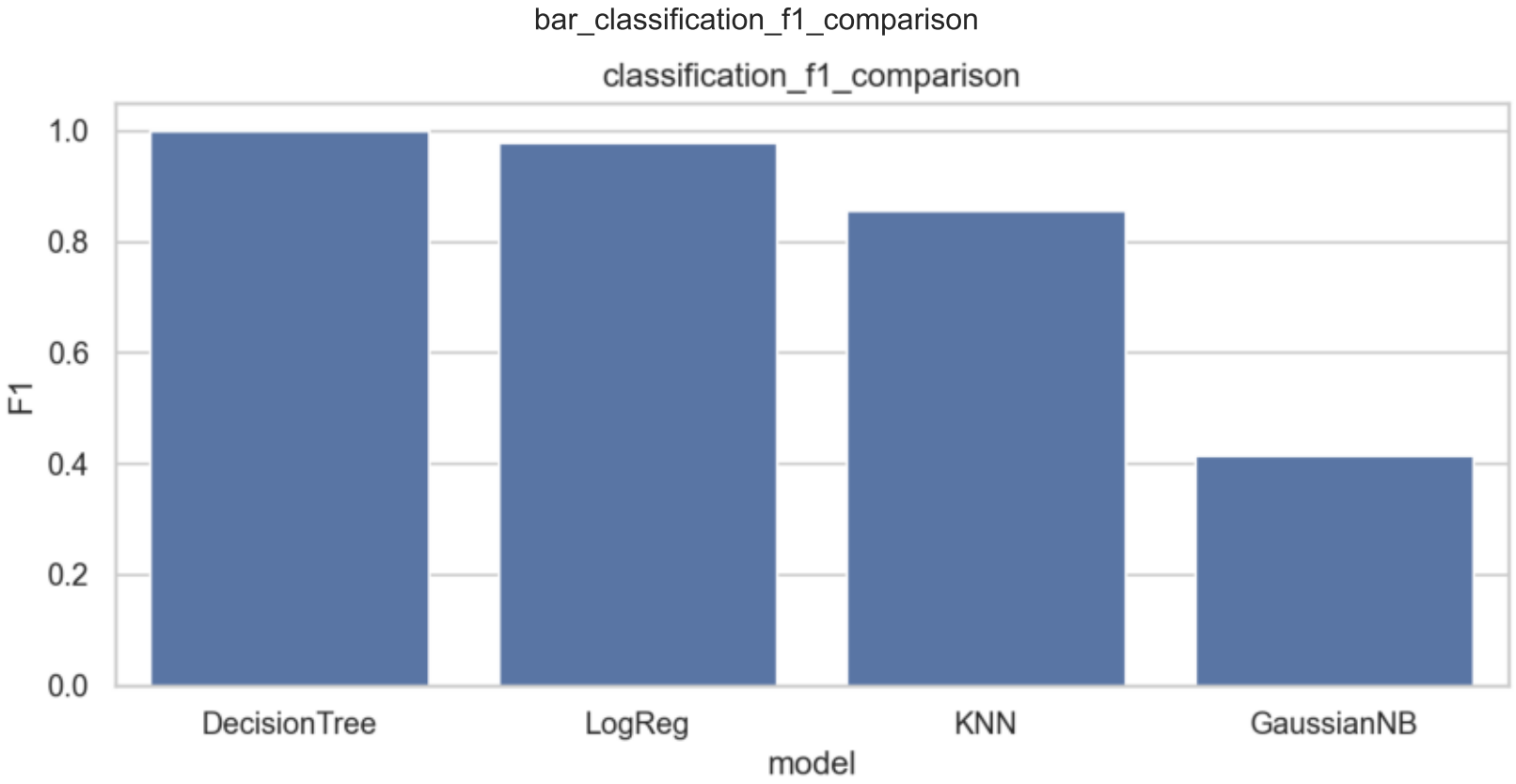
roc_GaussianNB



confusion_matrix_KNN




roc_KNN
roc_KNN (AUC=0.981)
True Positive Rate (Positive label: 1)
False Positive Rate (Positive label: 1)
0.0
0.2
0.4
0.6
0.8
1.0
Classifier (AUC = 0.98)


bar_classification_f1_comparison
classification_f1_comparison
1.0
0.8
0.6
0.4
0.2
0.0
F1
DecisionTree
LogReg
KNN
GaussianNB
model

scatter_predicted_vs_actual_linear_regression