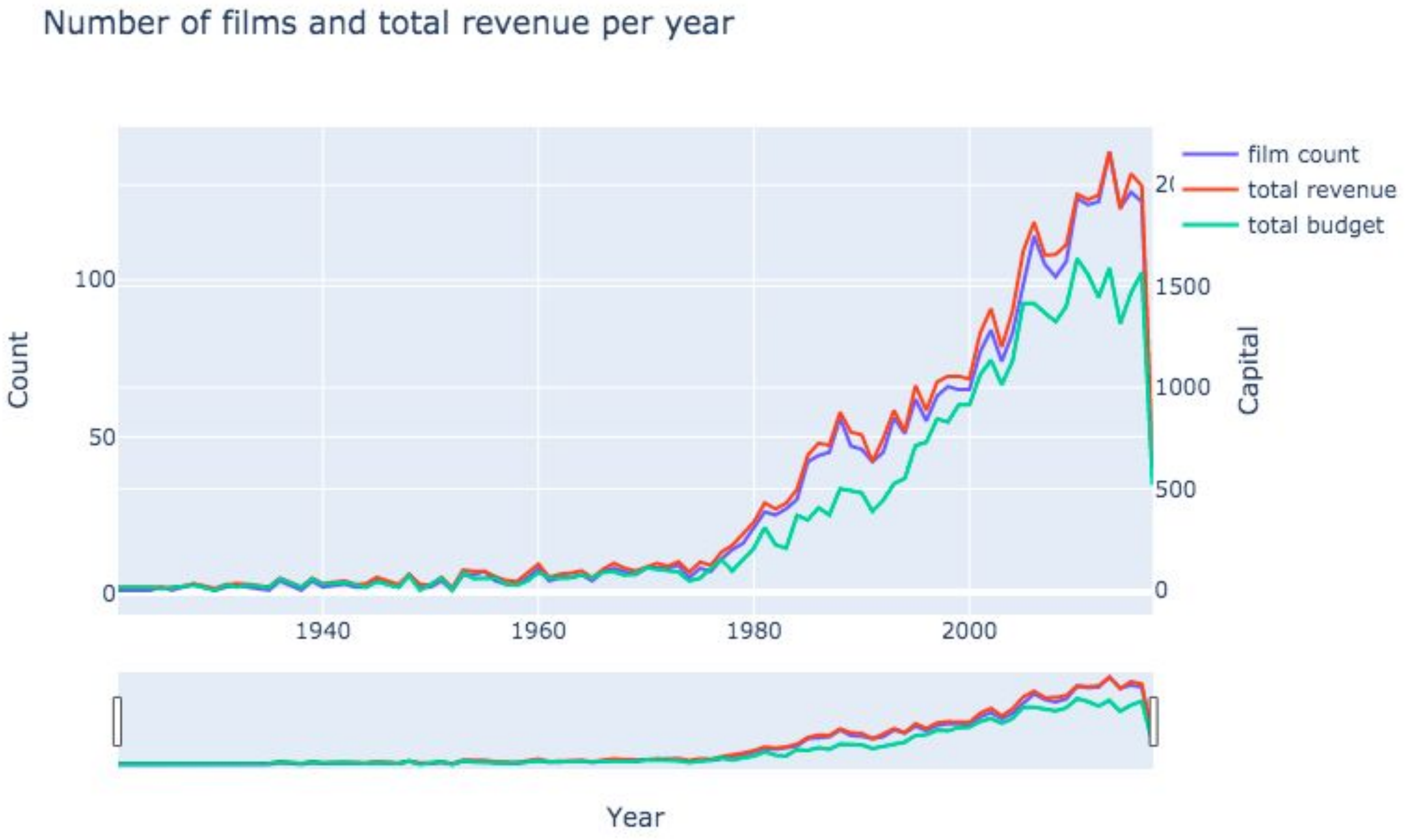
TMDB Box Office Collection: Prediction

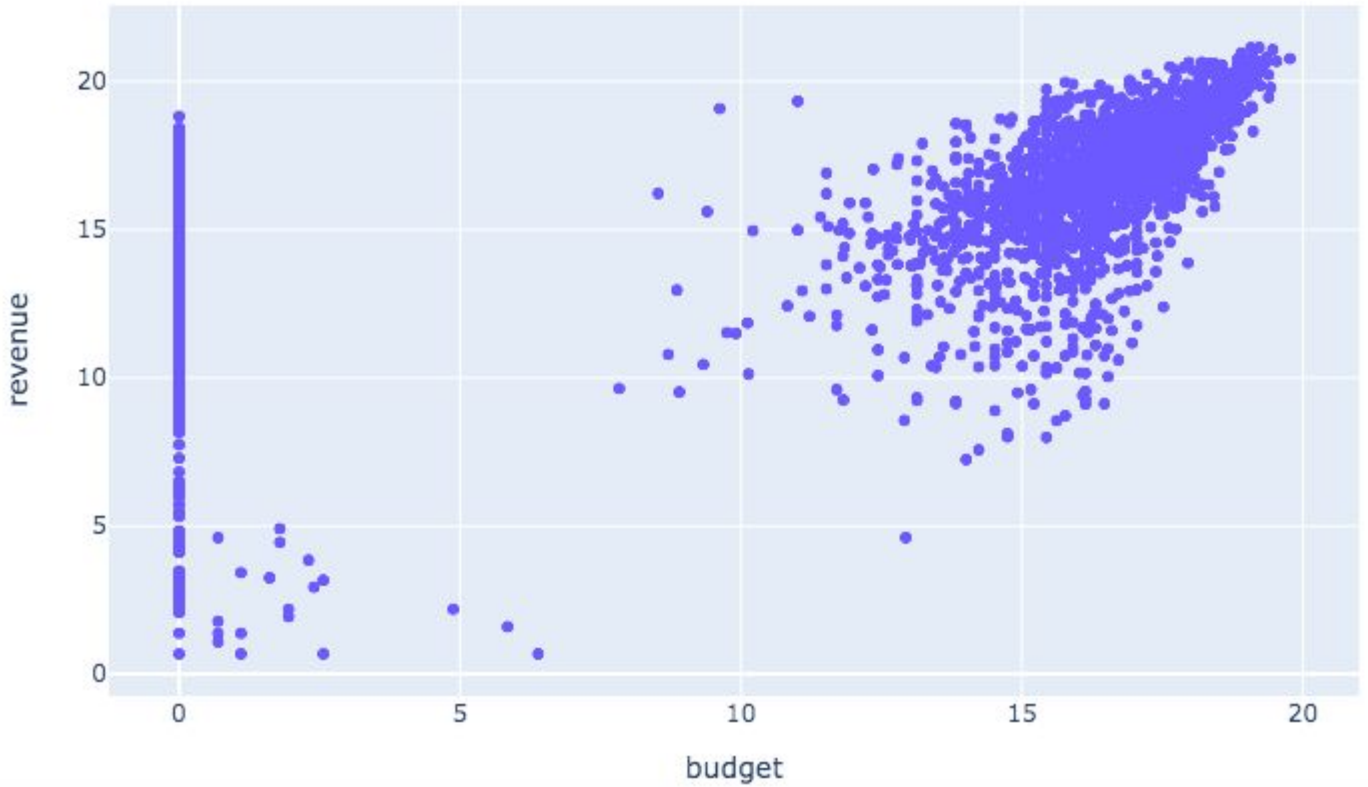
Predicting the box office revenue of over 4000 movies based on the data and features we have for 3000 movies.

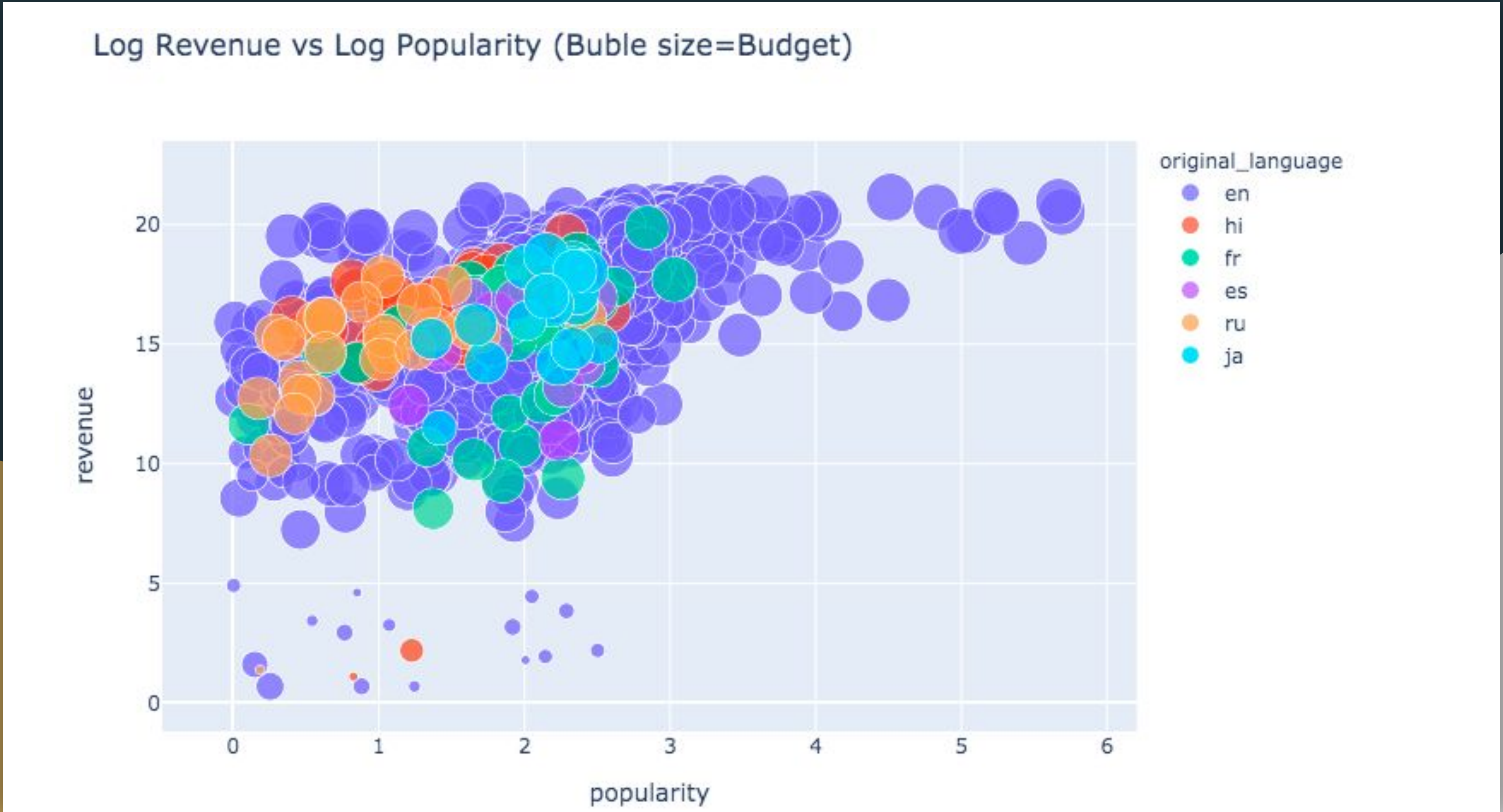
Adamya Nayyar

Data Exploration

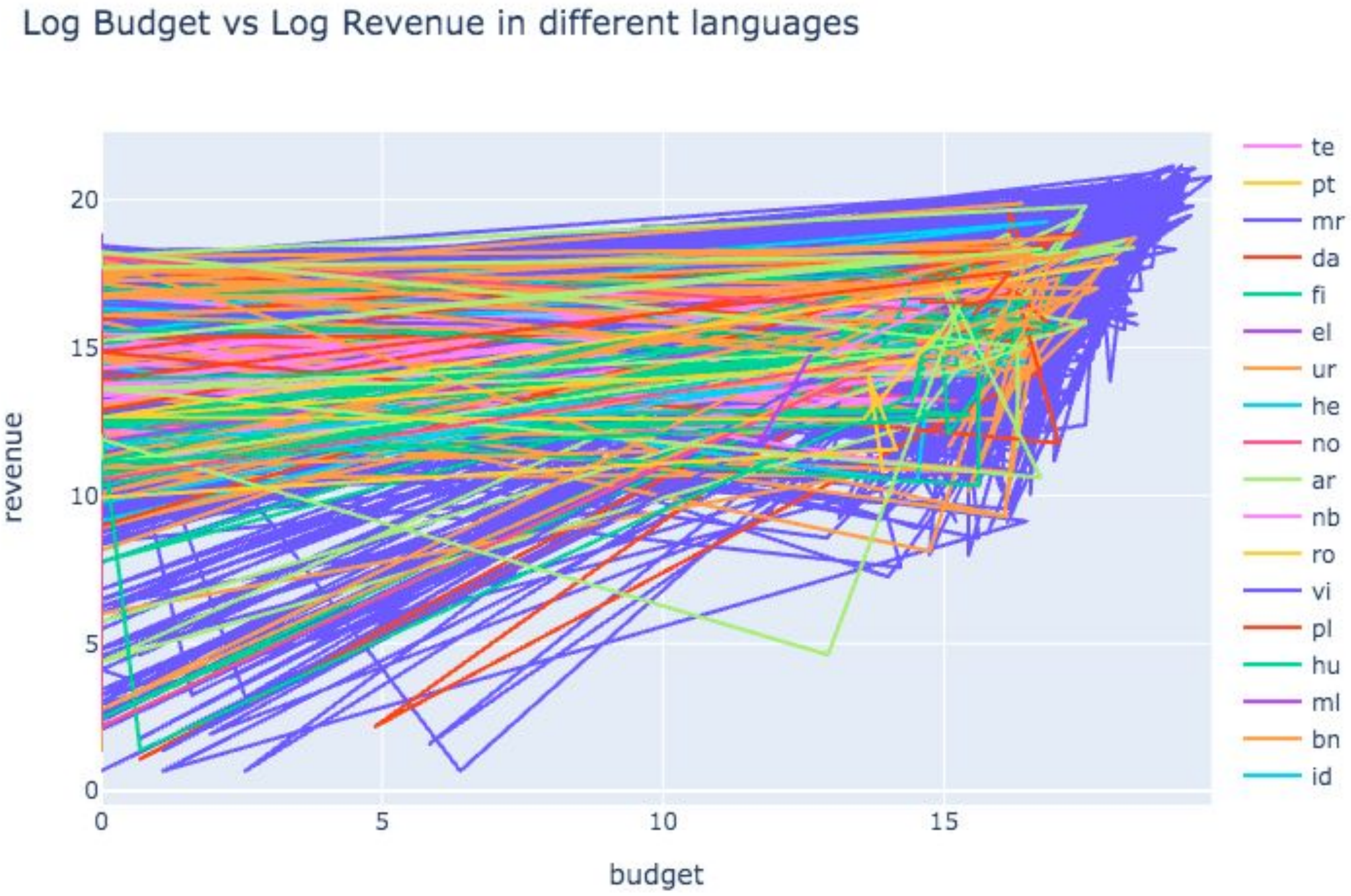


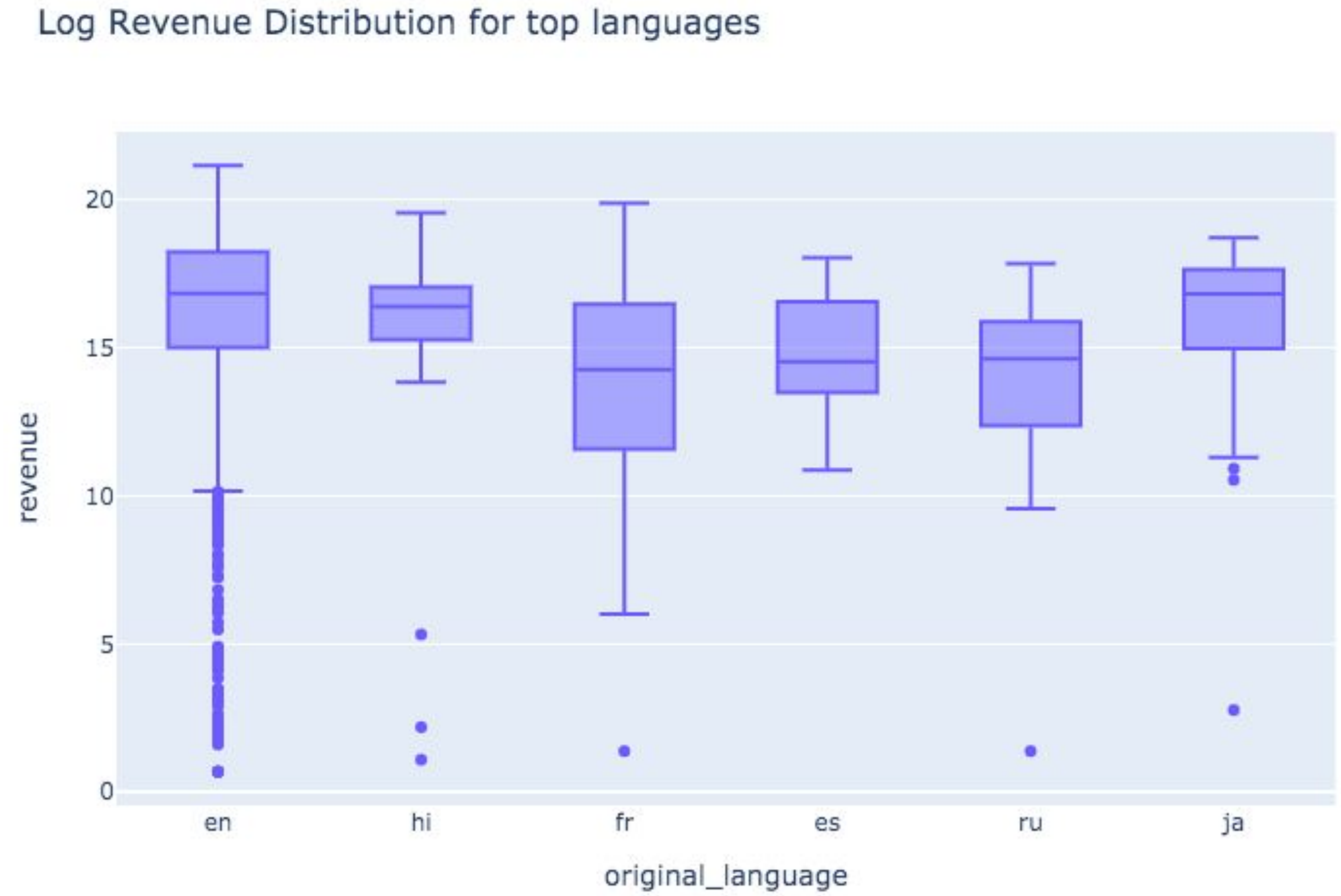
Log Budget vs Log Revenue



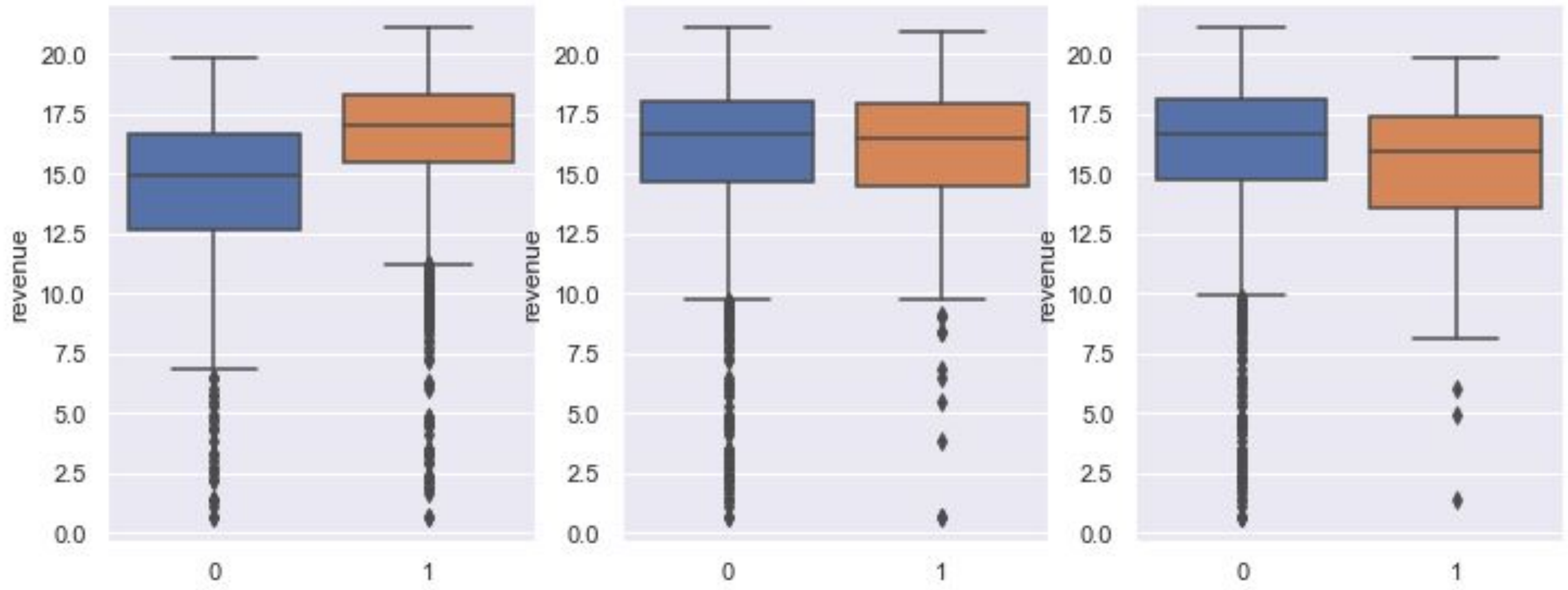


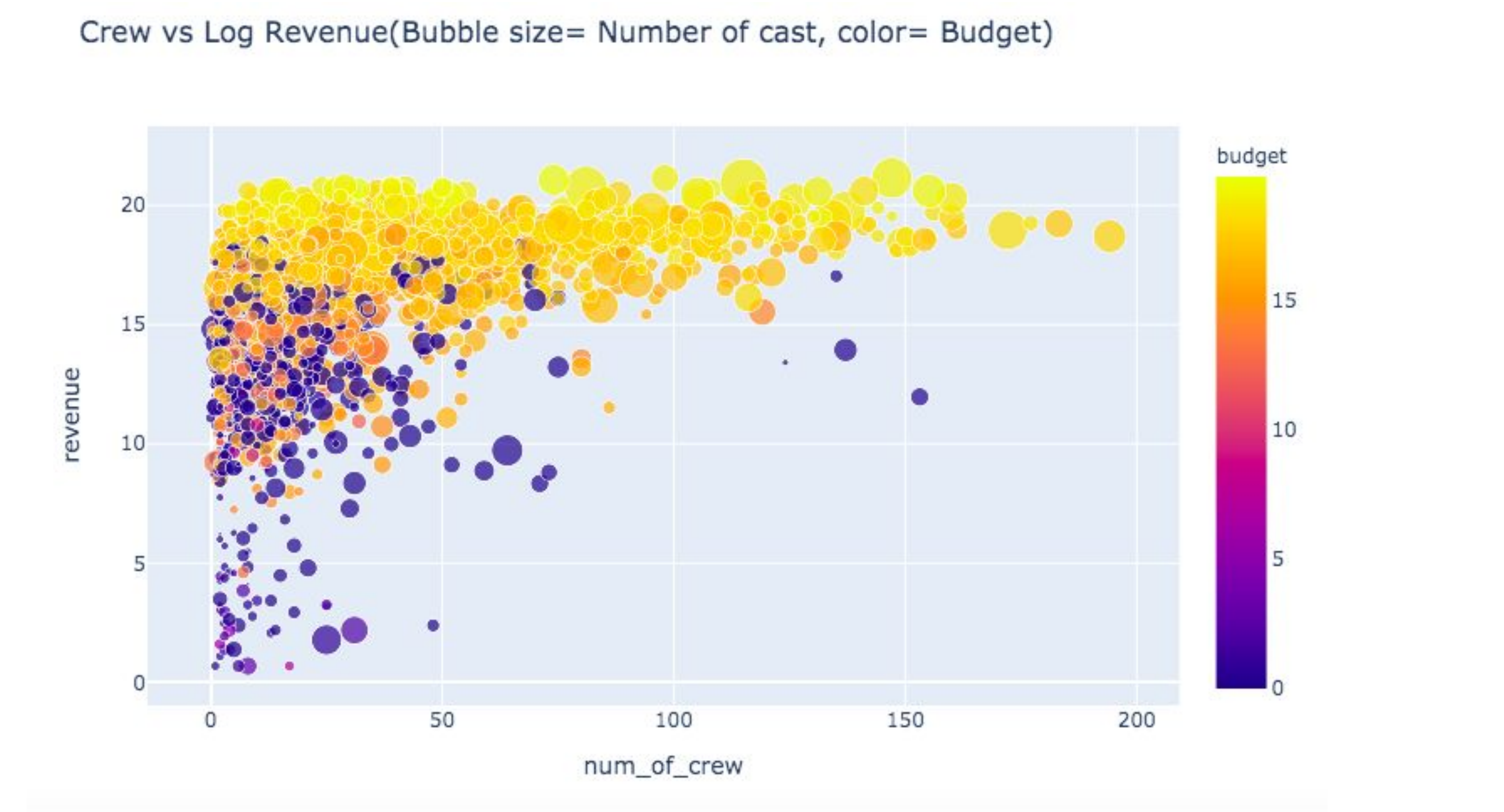
Log Revenue vs Log Popularity (Buble size=Budget)


Log Budget vs Log Revenue in different languages
revenue
budget
0
5
10
15
20
0
5
10
15
te
pt
mr
da
fi
el
ur
he
no
ar
nb
ro
vi
pl
hu
ml
bn
id


Log Revenue Distribution for top languages
revenue
20
15
10
5
0
en
hi
fr
es
ru
ja
original_language

Log revenue vs Top Production Countries



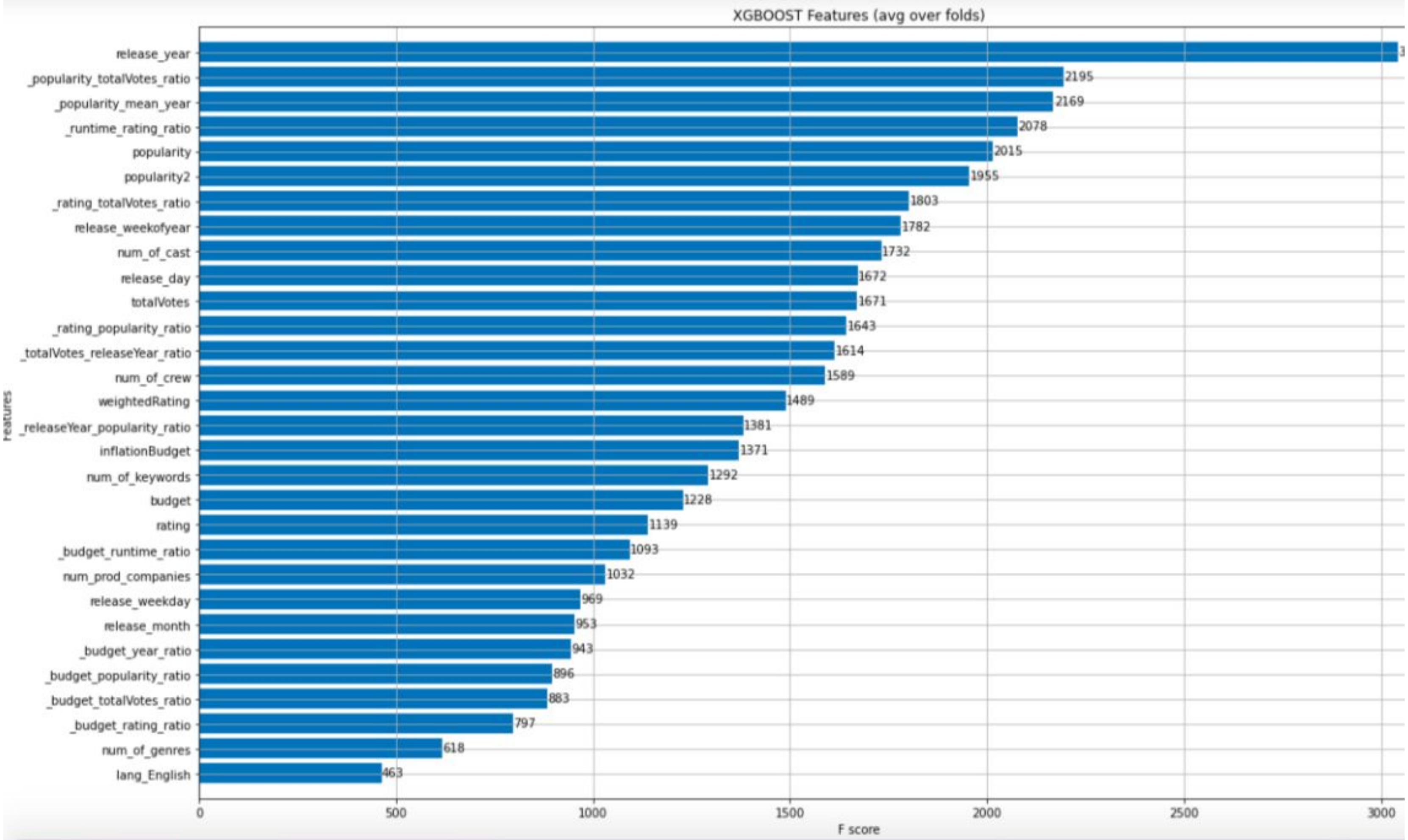

Crew vs Log Revenue(Bubble size= Number of cast, color= Budget)
revenue
20
15
10
5
0
num_of_crew
0
50
100
150
200
budget
15
10
5
0

Model Selection

- After fitting the data on various regression models, random forest gave us good result whereas Xgboost and Catboost model gave us the least errors.
- Random Forest model had rmse of 1.91 on validation set. Xgboost had even less error at rmse 1.83 whereas Catboost with the least rmse 1.81 .
- Since the root mean squared error of Xgboost and Catboost are very close, we can select either one as our final model. But for this project I calculated the final prediction as:

Final prediction = 0.3*Xgboost + 0.7*Catboost

- The above ensemble model gave the best prediction score.
- Next page have important features by Xgboost model.


XGBOOST Features (avg over folds)
release_year
3
_popularity_totalVotes_ratio
2195
_popularity_mean_year
2169
_runtime_rating_ratio
2078
popularity
2015
popularity2
1955
_rating_totalVotes_ratio
1803
release_weekofyear
1782
num_of_cast
1732
release_day
1672
totalVotes
1671
_rating_popularity_ratio
1643
_totalVotes_releaseYear_ratio
1614
num_of_crew
1589
weightedRating
1489
_releaseYear_popularity_ratio
1381
inflationBudget
1371
num_of_keywords
1292
budget
1228
rating
1139
_budget_runtime_ratio
1093
num_prod_companies
1032
release_weekday
969
release_month
953
_budget_year_ratio
943
_budget_popularity_ratio
896
_budget_totalVotes_ratio
883
_budget_rating_ratio
797
num_of_genres
618
lang_English
463
Features
0
500
1000
1500
2000
2500
3000
F score

THANK YOU! ANY QUERIES?