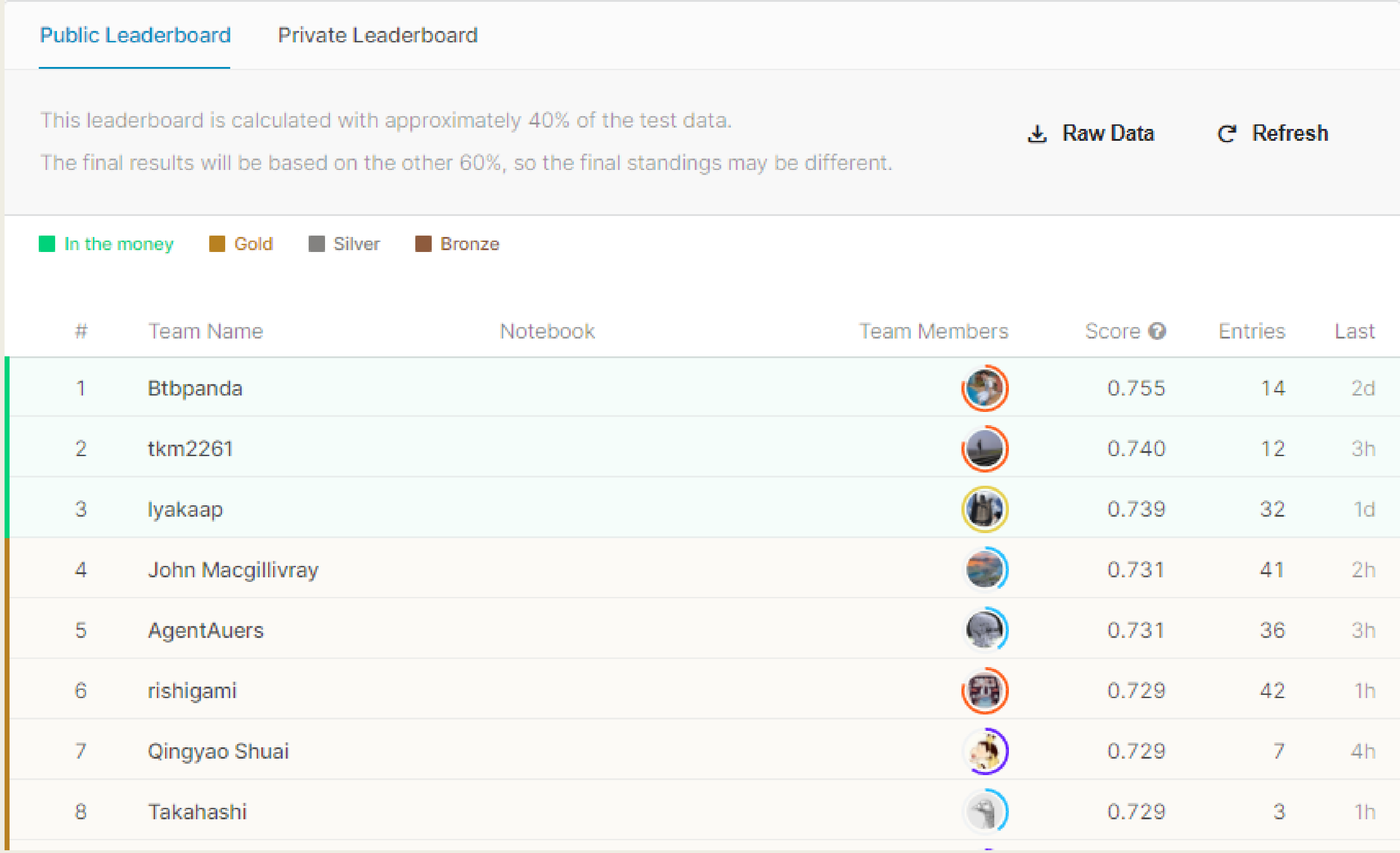
KAGGLE PROJECT:
PREDICTIVE ANALYSIS COMPETITION

Applied Analytics: Frameworks and Methods 1

Competition

- Compete to generate the best predictions.
- Goal is to generate the best predictions at the end of the ~4-week long competition.
- Every submission is scored and results posted to leaderboard in real time.
- Can submit up to three predictions each day.
- Complete transparency. Positions of all participants are visible throughout the competition.

Sample Leaderboard



Public Leaderboard   Private Leaderboard

This leaderboard is calculated with approximately 40% of the test data.

The final results will be based on the other 60%, so the final standings may be different.

Raw Data   Refresh

🟩 In the money 🟨 Gold ◼ Silver 🟫 Bronze

#	Team Name	Notebook	Team Members	Score	Entries	Last
1	Btbpanda			0.755	14	2d
2	tkm2261			0.740	12	3h
3	lyakaap			0.739	32	1d
4	John Macgillivray			0.731	41	2h
5	AgentAuers			0.731	36	3h
6	rishigami			0.729	42	1h
7	Qingyao Shuai			0.729	7	4h
8	Takahashi			0.729	3	1h

kaggle

- Hosted on Kaggle, an online platform that runs data science competitions
- 1M registered users and 60K active users compete on Kaggle for
  - *Sport and Bragging rights*
  - *A Job with competition sponsor*
  - *A chance to showcase skills to recruiters*
  - *Prize Money*

kaggle

- Through this competition, you will
  - *earn bragging rights*
  - *gain valuable hands-on experience with building models*
  - *gave a chance to showcase skills to recruiters, and*
  - *earn points*

ABOUT THE COMPETITION

About the Competition

- Description
  - People interested in renting an apartment or home, share information about themselves and their property on Airbnb. Those who end up renting the property share their experiences through reviews. The dataset describes property, host, and reviews for over 40,000 Airbnb rentals in New York along 90 variables.*
- Goal
  - *Construct a model using the dataset supplied and use it to predict the price of a set of Airbnb rentals included in scoringData.csv.*
- Metric
  - *Submissions will be evaluated based on RMSE (root mean squared error). Lower the RMSE, better the model.*

** Disclaimer: The data is not supplied by Airbnb. It was scraped from Airbnb's website. We do not either implicitly or explicitly guarantee that the data is exactly what is found on Airbnb's website. This data is to be used solely for the purpose of the Kaggle Project for this course. It is not recommended for any use outside of this competition.*

Deliverables

- Predictions submitted on competition site hosted on Kaggle
- Presentation
- Report and supporting R code for
    - best model
    - data wrangling and experimentation in arriving at the best model

Grading Criteria

- Commitment to the Project (25 points)
  - *Worked consistently on the Project.*
  - *First submission before specified date and a total of at least five submissions*
- Quality of Modeling (50 points)
  - *Demonstrated adequate knowledge of data exploration, suitably prepared data for analysis, used a variety of predictive analysis techniques, and communicated results effectively.*
  - *Assessed by (a) a brief report summarizing the data analysis process supplemented by neatly commented R code for the best submission, and (b) a 1-2 min presentation on experiences and lessons learned*
- Prediction Accuracy (75 points)
  - *Accuracy of predictions at the end of the Project.*
  - *Assessed by Rank on Leaderboard*

GETTING STARTED

Registration

- Registration opens on October 18th
- To register for the Kaggle Competition, click here and follow directions

First Submission

- Download data from Kaggle
- Read Data
- Construct Model
- Read scoring Data and apply model to generate predictions
- Construct submission from predictions
- Upload to Kaggle

First Submission Code

```
# For the following code to work, ensure analysisData.csv and scoringData.csv are in your working directory.
# Read data and construct a simple model
        data = read.csv('analysisData.csv')
        model = lm(price~minimum_nights+review_scores_accuracy,data)

# Read scoring data and apply model to generate predictions
        scoringData = read.csv('scoringData.csv')
        pred = predict(model,newdata=scoringData)

# Construct submission from predictions
        submissionFile = data.frame(id = scoringData$id, price = pred)
        write.csv(submissionFile, 'sample_submission.csv',row.names = F)
```

Kaggle Timeline

- October 18th: Registration Opens
- October 31st: Deadline for entering first submission
- November 19th: Competition Closes

Good Luck