A Trial Run of the Olympics

DS 4002
JOE BARRETT
CONNOR NICKOL
BRITTANY NGUYEN

UVA

Agenda

UVA

OBJECTIVE AND HYPOTHESIS DEVELOPMENT

COUNTRY RANKINGS

Rank	NOC	Gold	Silver	Bronze	Total
1	Russia*‡	11	10	9	30
4	United States‡	9	9	10	28
2	Norway	11	5	10	26
3	Canada	10	10	5	25
5	Netherlands	8	7	9	24
6	Germany	8	6	5	19
9	Austria	4	8	5	17
10	France	4	4	7	15
14	Sweden	2	7	6	15
7	Switzerland‡	7	2	2	11
7	Switzerland‡	7	2	2	11

OBJECTIVE/HYPOTHESIS

- RELEASED DATA SHOWS: RUSSIA, UNITED STATES, NORWAY, CANADA, NETHERLANDS, GERMANY, AUSTRIA, FRANCE, SWEDEN, AND SWITZERLAND HAD THE HIGHEST TOTAL MEDAL COUNT IN THE 2014 OLYMPICS
- BASED ON DATA FROM 1960-2010, WE AIM TO PREDICT THE 2014 MEDAL COUNT FOR THESE 10 HIGH-PERFORMING COUNTRIES (SHOULD MATCH AS CLOSELY AS POSSIBLE)

PREDICTOR TRENDS

- COUNTRIES WITH A:
- 1)HIGHER NUMBER OF ATHLETE PARTICIPANTS
- 2)HIGHER NUMBER OF EVENTS
- 3)HIGHER AVERAGE GDP
- 4)NORTHERN LATITUDES
- 5)SMALLER EUCLIDEAN DISTANCE FROM HOST LOCATION
- …ARE PREDICTED TO WIN MORE TOTAL MEDALS

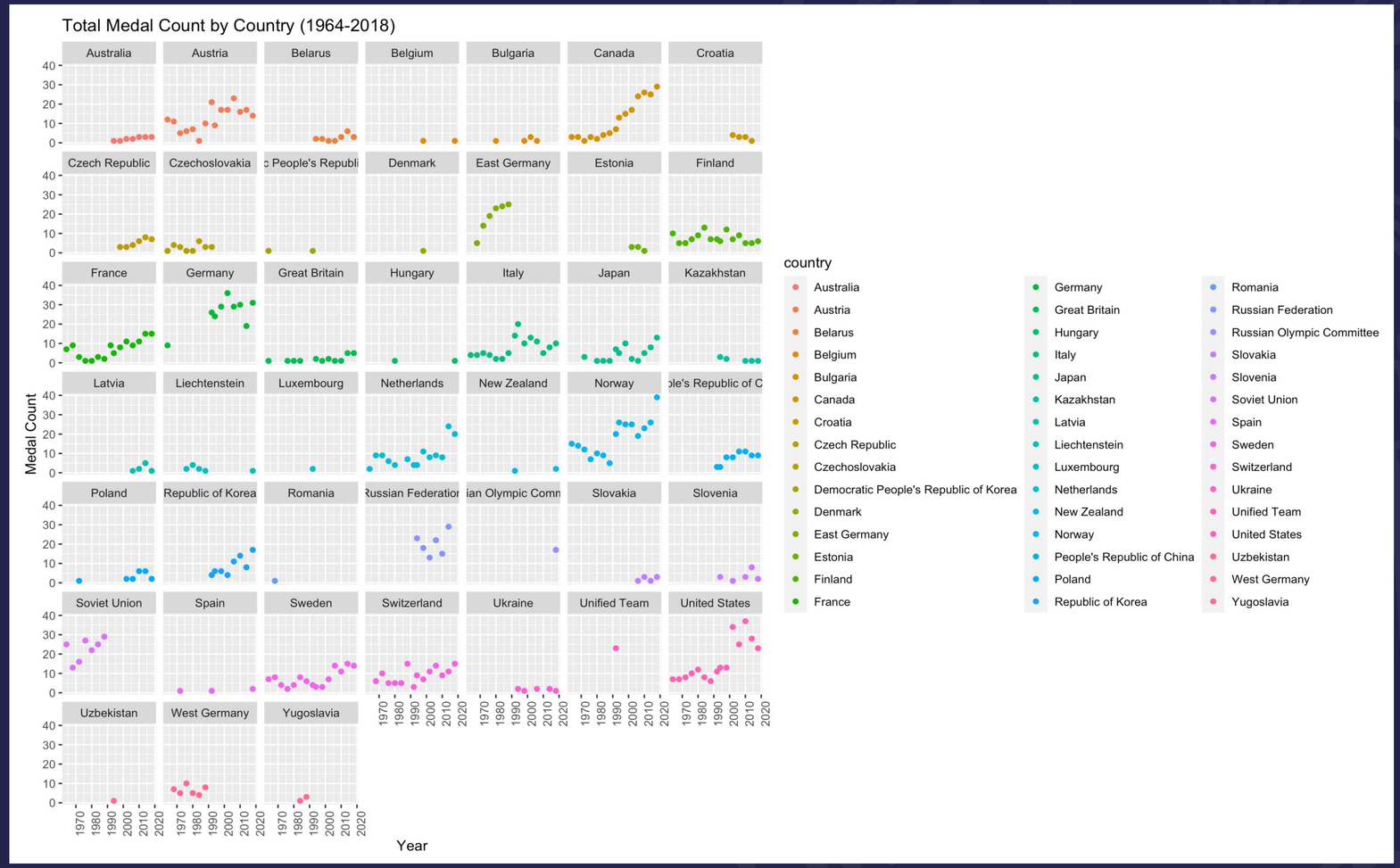
UVA

DATA OVERVIEW AND BACKGROUND

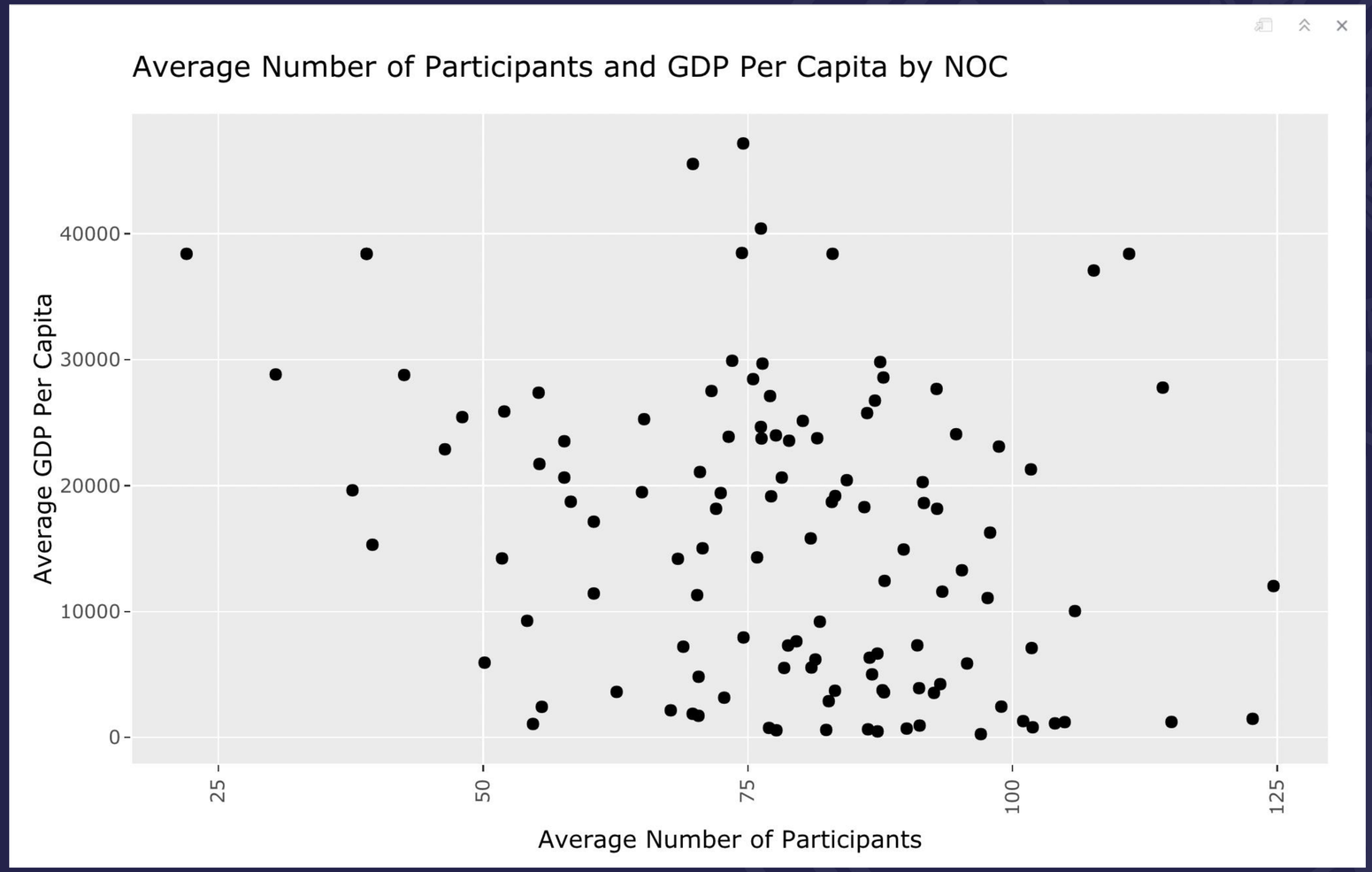
DATASET (LINK)	VARIABLES	SIZE	SCOPE
OLYMPIC HISTORY: ATHLETES AND RESULTS	• ID • NAME • SEX • AGE • HEIGHT /WEIGHT • TEAM • NOC - 3-LETTER CODE • GAMES - YEAR/SEASON • YEAR - INTEGER • SEASON • CITY • SPORT • EVENT - EVENT • MEDAL TYPE	• 42MB • 134732 UNIQUE ATHLETES • 230 UNIQUE COUNTRY NAMES • 1896 - 2016	• ALL DATA • LINKED TO FOUR OTHER SOURCES VIA COUNTRY CODE AND NAME
GDP PER CAPITA, PPP	• GDP PER CAPITA • COUNTRY	• 209KB • 266 UNIQUE VALUES	• ALL
LATITUDE & LONGITUDE BY COUNTRY AND STATE	• LATITUDE • LONGITUDE • STATE NAME • COUNTRY NAME	• 11.2KB • 245 UNIQUE VALUES	• ALL
COUNTRIES ISO CODES	• COUNTRY NAME • COUNTRY NAME	• 9.45KB • 246 UNIQUE VALUES	• ALL
WINTER OLYMPICS (1924 – 2018)	• EVENT NUMBER • YEAR • COUNTRY • [MEDAL TYPES]	• 12.25KB • 410 UNIQUE VALUES	• ALL

UVA

EDA AND MODELING PLAN



EDA AND MODELING PLAN



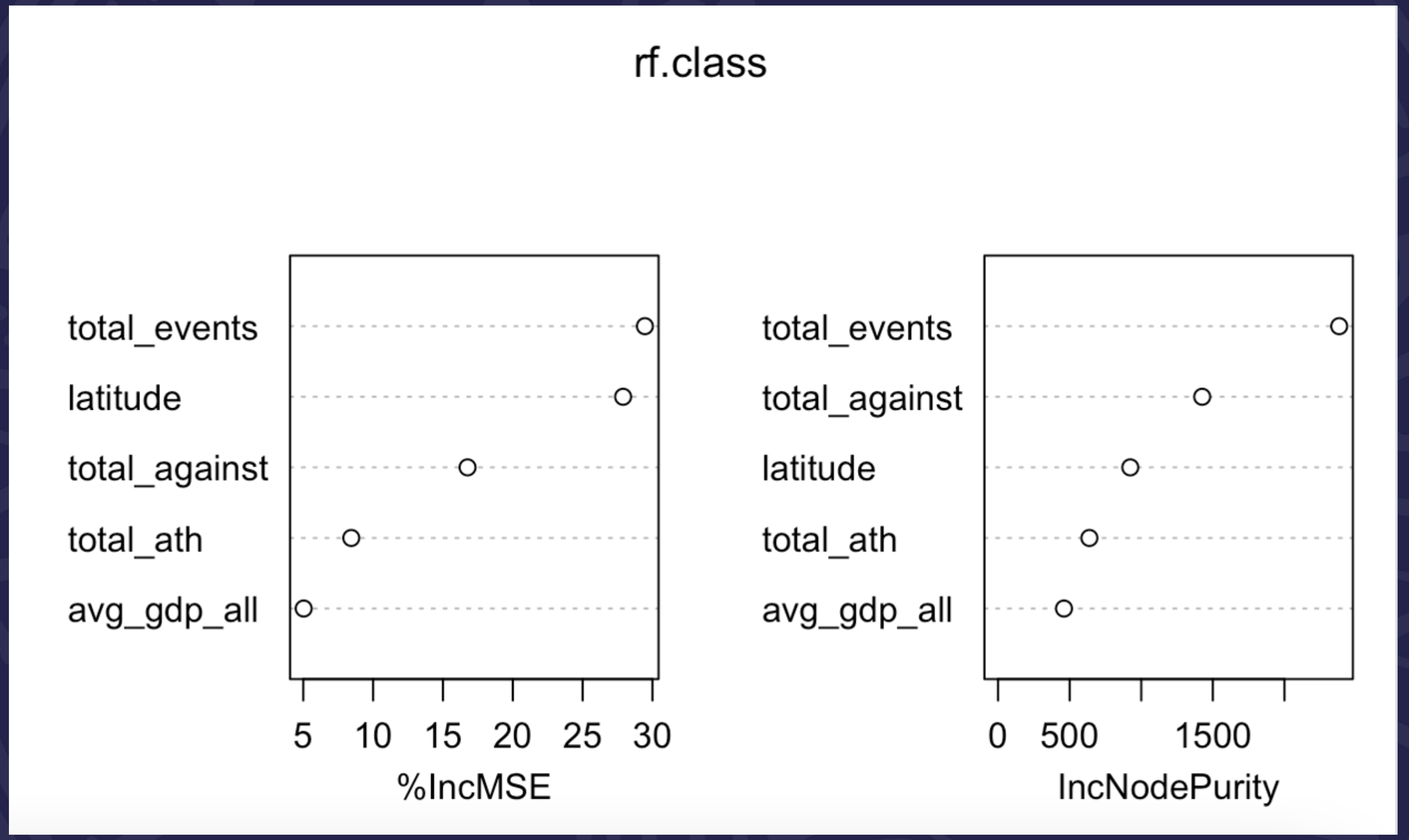
INITIAL MODELING AND FEATURE ENGINEERING

VARIABLES INCLUDED IN THE MODEL:

- TOTAL EVENTS ENTERED
- TOTAL ATHLETES COMPETED AGAINST
- TOTAL ATHLETES ENTERED
- AVG GDP OF COUNTRIES FACED
- EUCLIDEAN DISTANCE TO THE OLYMPIC HOST CITY

ADDED IN:

- LATITUDE OF COMPETING COUNTRY



MODEL COMPARISON AND RESULTS OF THE 2014 PREDICTIONS

	Method	Mean Squared Error
1	Linear Regression	57.89
2	Full Linear Regression	55.78
3	Random Forest	62.92
4	Recursive Binary Splitting	107.75
5	Bagging	62.23
6	Boosting	61.11

	country	Year	total_medals	predict
1	RUS	2014	29	36
2	USA	2014	27	29
3	GER	2014	19	26
4	NOR	2014	26	24
5	SUI	2014	10	20
6	AUT	2014	17	19
7	FRA	2014	15	17
8	CAN	2014	24	12
9	NED	2014	24	12

UVA

CONCLUSIONS AND FUTURE DIRECTIONS

- MSE OF 55 INDICATES DECENT MODEL! (ABOUT 7 MEDALS OFF ON AVERAGE PER COUNTRY)

- FUTURE DIRECTIONS:
- USING THIS DATA PIPELINE TO PREDICT ‘22 GAMES DATA ONCE TEST SET IS AVAILABLE

- OTHER VARIABLES CONSIDERED:
- 1)AIR QUALITY OF EACH COUNTRY
- 2)EFFECTS OF HISTORICAL/POLITICAL TURMOIL PER COUNTRY ON MEDAL COUNT

- REVISITING QUALITY OF PREDICTORS