F1 Data Analysis

JOE BARRETT
CONNOR NICKOL
BRITTANY NGUYEN
(GROUP 3)

DS 4002
SPRING 2022

UVA

Agenda

UVA

OBJECTIVE

- **CLASSIFICATION ML:**

->1096 RACES TOTAL

->PREDICT WHETHER THE 1ST, 2ND, OR 3RD PLACE QUALIFIER WILL WIN A GIVEN RACE (1) OR NOT (0) (& CORRESPONDING PROBABILITIES OF EACH EVENT)
->LIKELIHOOD OF 1ST PLACE QUALIFIER WINNING AT EACH OF 5 MOST FREQUENTLY RACED TRACKS

- **MODELS USED:**

->EMPLOYED DECISION TREE, RANDOM FOREST LOGISTIC REGRESSION, BAGGING/BOOSTING

HYPOTHESES

- **HIGHER LIKELIHOOD OF WINNING A RACE WITH:**

1) LESS PIT STOPS
2) MORE SAFETY CARS
3) WEATHER = RAINY

*"SAFETY CARS" & "WEATHER" MAKE CERTAIN TRACKS MORE UNPREDICTABLE
->INCREASED LIKELIHOOD OF OVERTAKES
->DECREASED CHANCE OF 1ST PLACE QUALIFIER WINNING

VARIABLES

1) WEATHER
2) QUALIFYING TIME GAP ->WITH RESPECT TO 4TH PLACE QUALIFIER
3) QUALIFYING POSITION
4) CIRCUIT ID/REFERENCE
5) DRIVER NATIONALITY
6) LATITUDE & LONGITUDE
7) COUNT OF SAFETY CARS
8) COUNT OF PIT STOPS

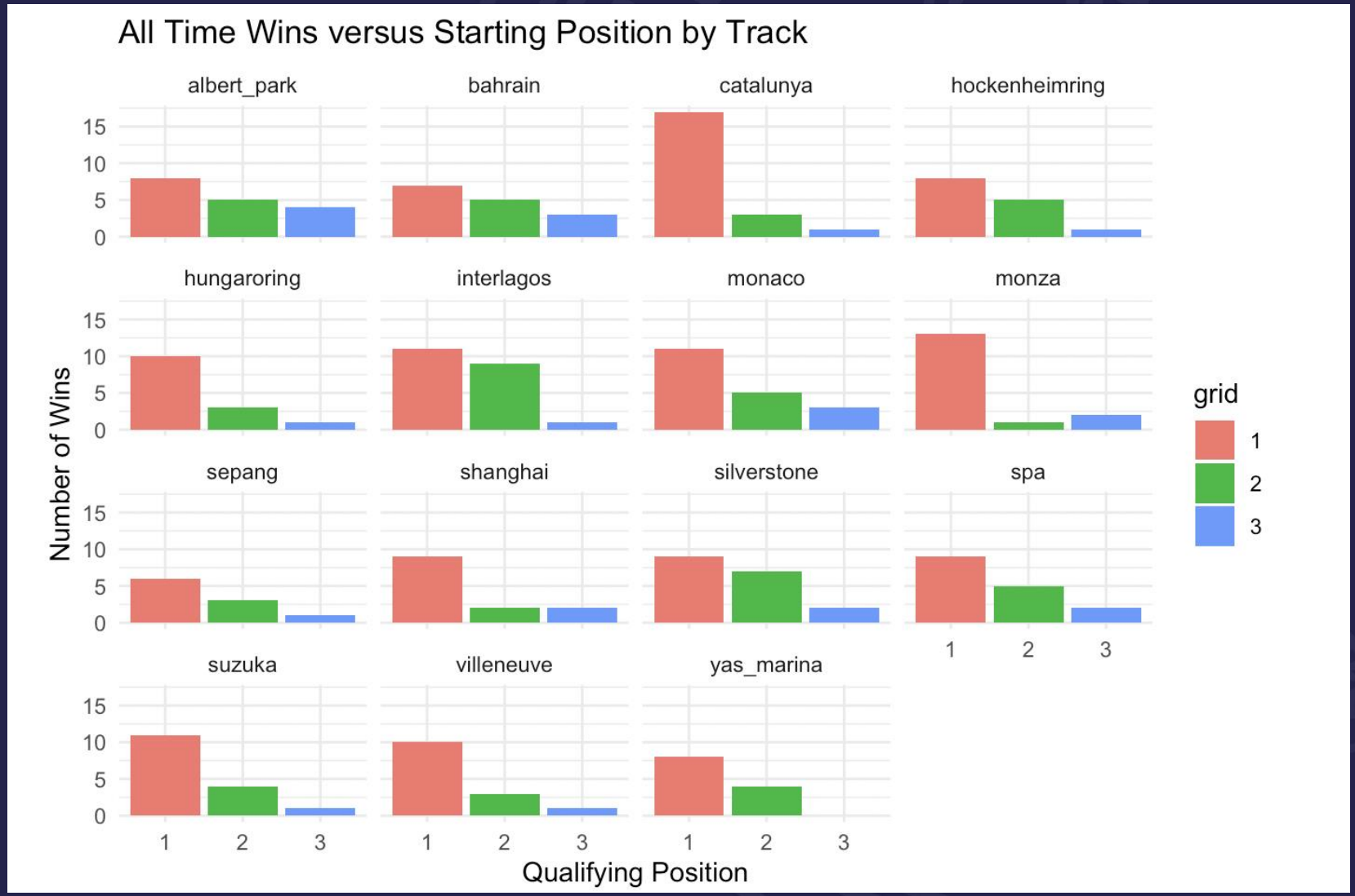
NAME	WEATHER_WET	CIRCUITREF	circuitId	circuit	country	lat	lng	alt	code	forename	surname	nationality	qual_gap	won_race	r
Australian Grand Pri	N	albert_park	0	Albert Park Grand Pri	Australia	-37.8497	144.968	10	VET	Sebastian	Vettel	German	-0.084	0	
Australian Grand Pri	N	albert_park	0	Albert Park Grand Pri	Australia	-37.8497	144.968	10	BAR	Rubens	Barrichello	Brazilian	-0.409	0	
Australian Grand Pri	N	albert_park	0	Albert Park Grand Pri	Australia	-37.8497	144.968	10	BUT	Jenson	Button	British	-0.712	1	
Malaysian Grand Pri	Y	sepang	0	Sepang Internationa	Malaysia	2.76083	101.738	18	TRU	Jamo	Trulli	Italian	-0.378	0	
Malaysian Grand Pri	Y	sepang	0	Sepang Internationa	Malaysia	2.76083	101.738	18	BUT	Jenson	Button	British	-0.47	1	
Chinese Grand Prix	Y	shanghai	0	Shanghai Internatior	China	31.3389	121.22	5	WEB	Mark	Webber	Australian	-0.027	0	
Chinese Grand Prix	Y	shanghai	0	Shanghai Internatior	China	31.3389	121.22	5	ALO	Fernando	Alonso	Spanish	-0.112	0	
Chinese Grand Prix	Y	shanghai	0	Shanghai Internatior	China	31.3389	121.22	5	VET	Sebastian	Vettel	German	-0.309	1	
Bahrain Grand Prix	N	bahrain	3	Bahrain Internationa	Bahrain	26.0325	50.5106	7	VET	Sebastian	Vettel	German	-0.029	0	
Bahrain Grand Prix	N	bahrain	3	Bahrain Internationa	Bahrain	26.0325	50.5106	7	GLO	Timo	Glock	German	-0.332	0	

DATA OVERVIEW AND BACKGROUND

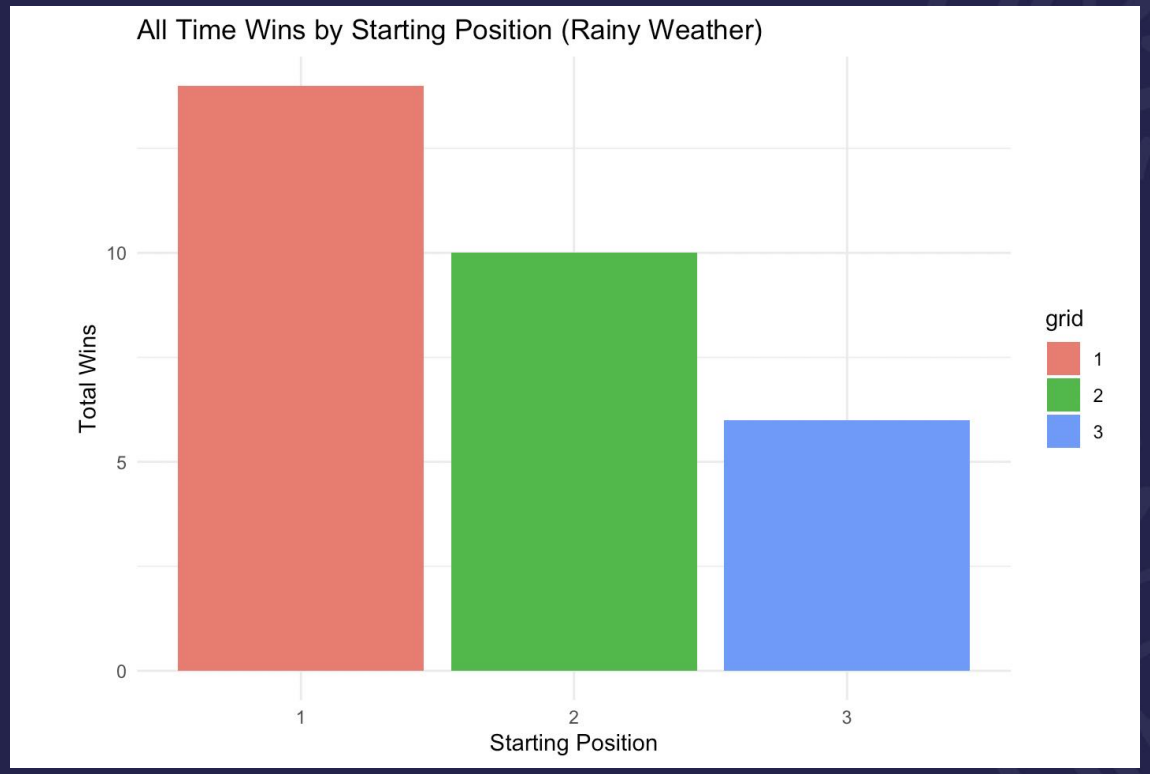
Source	Dataset	Selected Variables	Source	Dataset	Selected Variables
Kaggle	circuits	• circuitId • circuitRef • Name • Lat/long • Altitude • Country	Oracle	pitstops	• raceID • driverID • num_stops • date_time
	results	• resultId • raceID • driverID • positionOrder • position • grid • statusID		races	• raceID • Weather_wet • Name • year
	drivers	• driverId • code • forename • surname • nationality		safety cars	• year • Race • Count • Laps
	qualify	• driverId • raceID • position • q1 • q2 • q3			
	results	• constructorId • number • positionText • Points • fastestLapTime • fastestLapSpeed • Laps • Time • Milliseconds • fastestLap • rank			

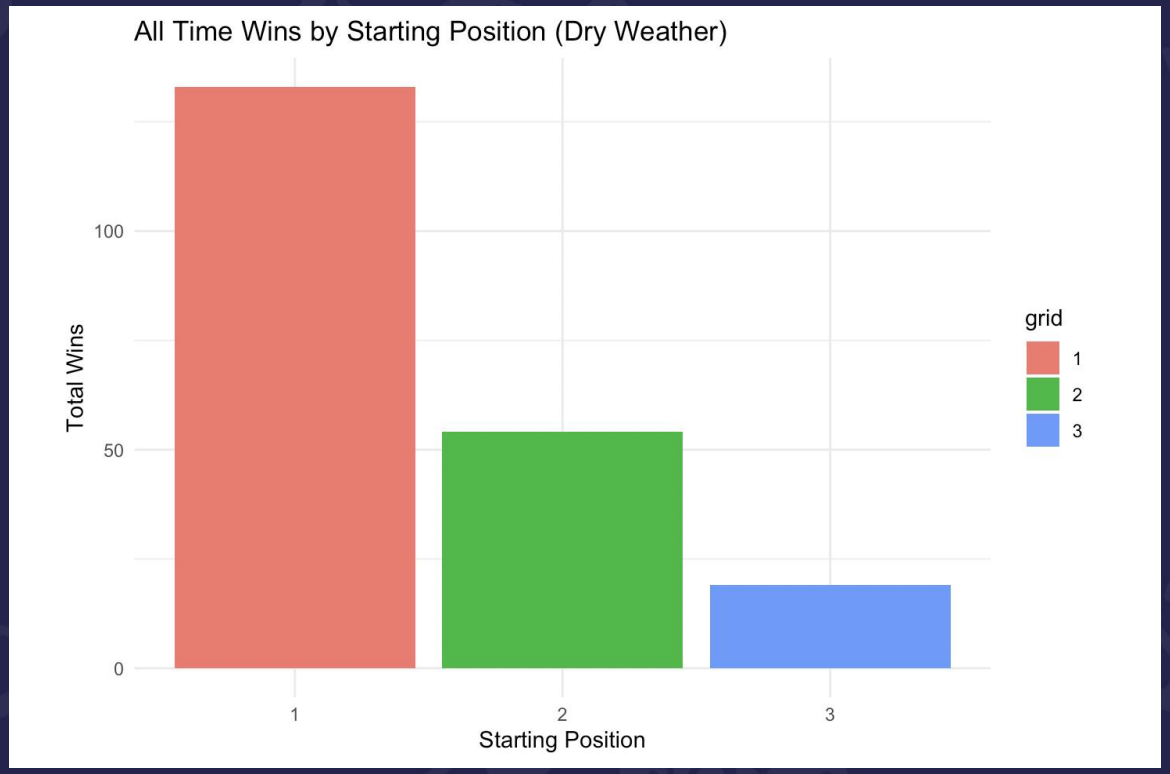
UVA

EXPLORATORY DATA ANALYSIS (PT. 1)

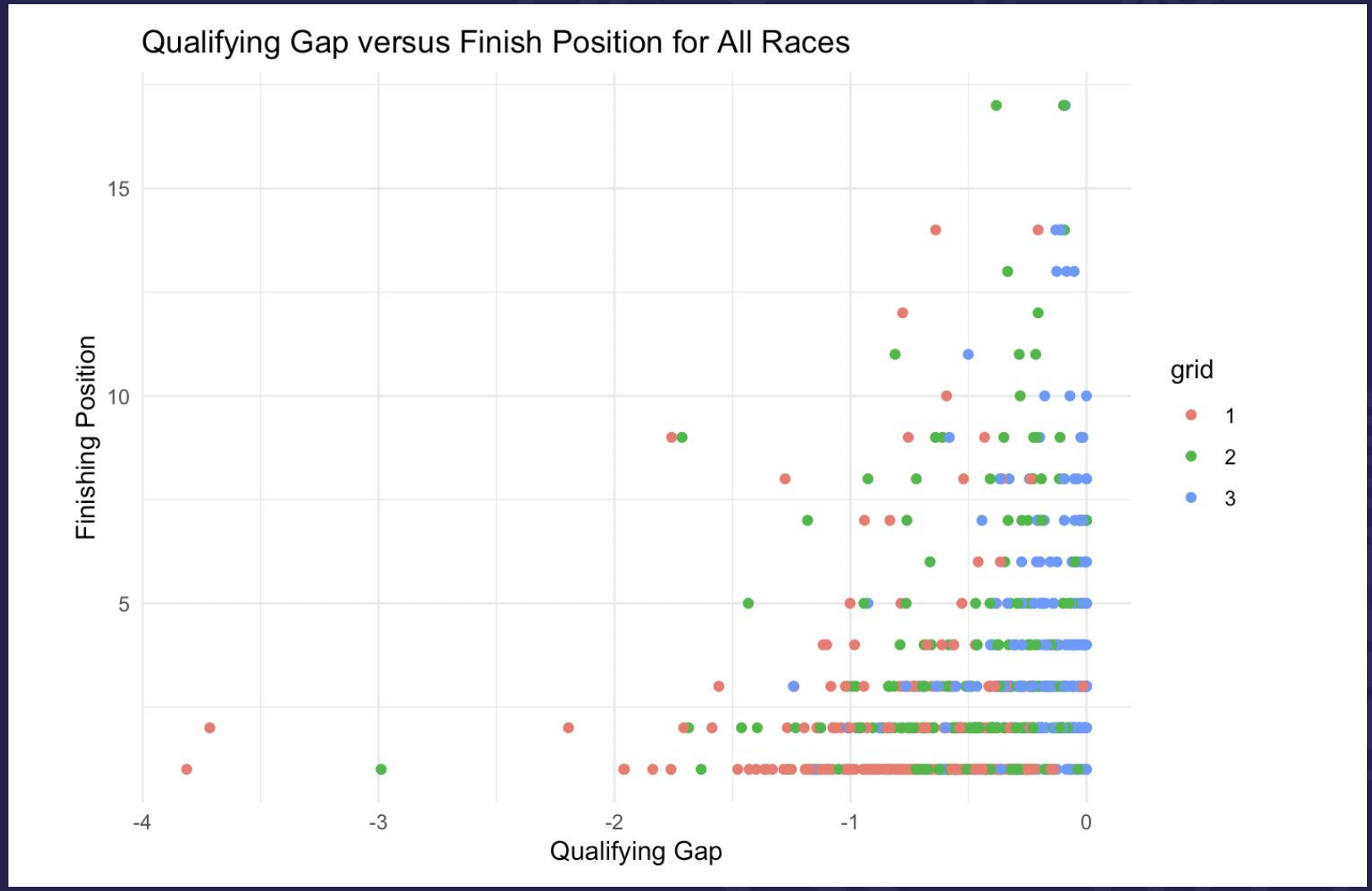


EXPLORATORY DATA ANALYSIS (PT. 2)





EXPLORATORY DATA ANALYSIS (PT. 3)



UVA

FEATURE ENGINEERING

1) COLUMN "QUAL_GAP" IS THE DIFFERENCE BETWEEN QUAL. TIMES OF EACH QUALIFIER AND THE 4TH PLACE QUALIFIER (WAY OF STANDARDIZING DIFFERENCES)

qual_gap
-0.084
-0.409
-0.712
-0.378
-0.47
-0.027
-0.112
-0.309
-0.029
-0.332
-0.613
-0.172
-0.274
-0.407
-0.194
-0.344
-0.369
-0.192

2) CONVERSION OF THE QUALIFYING TIMES TO SECONDS

UVA

FULL MODEL & EVALUATION

TWO SETS OF TRAINING DATA

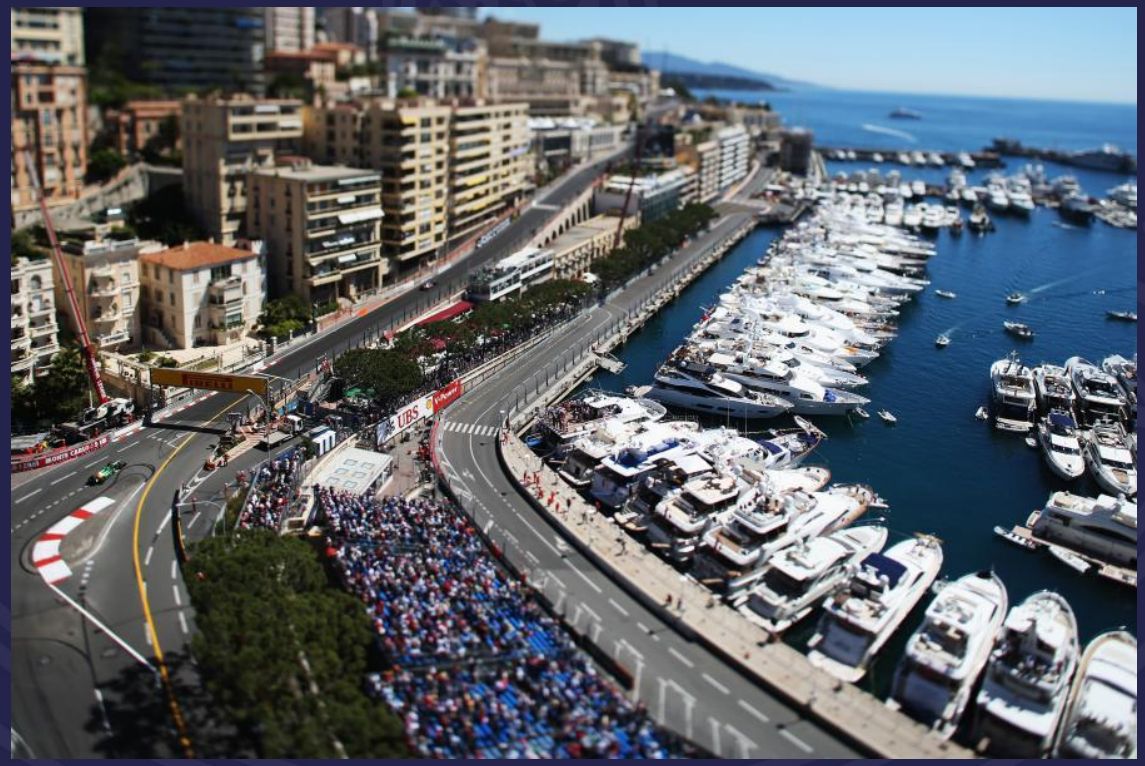
- ONE ALL YEARS OF DATA(1950-2020)
- SECOND JUST 2011-2020 BUT WITH PIT STOP AND SAFETY CAR COUNTS AS PREDICTOR VARIABLES
- BOTH USED JUST 2021 AS THE TESTING DATA

Model	Full Data Accuracy	2011+ Data Accuracy
Logistic Regression	0.7543	0.7593
Recursive Binary Splitting	0.7543	0.7407
Random Forest	0.7018	0.7222
Bagging	0.6842	0.7037
Boosting	0.8070	0.7034

UVA

LOOKING AT WIN PROBABILITIES BY TRACK

	circuitId	avg_prob	name	country	location
1	6	0.6426821	Circuit de Monaco	Monaco	Monte-Carlo
2	14	0.5906555	Autodromo Nazionale di Monza	Italy	Monza
3	13	0.5625987	Circuit de Spa-Francorchamps	Belgium	Spa
4	9	0.5208961	Silverstone Circuit	UK	Silverstone
5	11	0.5095333	Hungaroring	Hungary	Budapest



CONCLUSIONS AND REFLECTIONS

1) CONCLUSIONS:

- INCLUDING PIT STOPS AND SAFETY CAR DATA PRODUCED HIGHER ACCURACIES FOR 3 OUT OF 5 MODELS

- **BOOSTING** = CHAMPION MODEL (W/OUT SAFETY CARS AND PIT STOPS)
LOG. REG = CHAMPION MODEL WITH SAFETY CARS AND PIT STOPS

2) FUTURE DIRECTIONS:

- PROBABILITY OF 2ND OR 3RD PLACE QUALIFIERS WINNING AT EACH OF SAME 5 MOST RACED-AT TRACKS
- TOTAL WINS BY NATIONALITY/COUNTRY

3) GENERAL REFLECTIONS FROM CYCLES 1+2:

- DECIDING WHEN IT IS "WORTH IT" TO EXCLUDE PORTIONS OF DATA (PROJECT 1: GDP DATA NOT AVAILABLE FOR PORTION OF TIME PROJECT 2: SAFETY CAR/PIT STOP DATA MISSING IN TRAINING DATA)

- VALUE OF EDA IN FORECASTING (PROJECT 1: WHAT COUNTRIES HAD HIGHEST MEDAL COUNT PROJECT 2: EFFECTS OF TRACK OR WEATHER ON WINS)

UVA

WORKS CITED

Oracle-Devrel. (2022). *Oracle-devrel/redbull-analytics-hol: Learn machine learning the fun way, with Oracle and Redbull Racing*. GitHub. Retrieved April 25, 2022, from https://github.com/oracle-devrel/redbull-analytics-hol

Vopani. (2022, April 11). *Formula 1 World Championship (1950 - 2022)*. Kaggle. Retrieved April 25, 2022, from https://www.kaggle.com/datasets/rohanrao/formula-1-world-championship-1950-2020?select=races.csv

UVA

Thank You!

ANY QUESTIONS? :)

UVA