ASIGNMENT 1

ISSS602 – Data Analytics Lab

2020 TOKYO OLYMPICS
Insights from Swimming Reaction Time

Le Vu Anh Phuong (Joshua)
vaple.2021@mitb.smu.edu.sg

Contents

1. **OVERVIEW** 2

2. **OBJECTIVES** 2

3. **DATA** 3

3.1 Data Used 3

3.2 Data Quality Issues and Wrangling Solutions 3

4. **DATA ANALYSIS** 7

4.1 Reaction Time with Gender and Start Position 7

4.1.1 Subdividing the Populations and Normality Tests 7

4.1.2 Starting Position and Reaction Time 10

4.1.3 Gender and Reaction Time 10

4.2 Swimming Events and Reaction Time 12

4.2.1 Reaction Time Between Sprinter and Long-Distance Swimmers 12

4.2.2 Reaction Time Correlation with Split Performance and Final Placements 16

4.3 Other Aggregate Observations 17

4.3.1 The Outliers 17

4.3.2 The R.T of Past Good Performers within a Race 18

5. **RECOMMENDATIONS** 19

Bibliography 19

1. OVERVIEW

With technological progress, data has become easier to collect, shared and analyzed. As a result, data analytics has been used in many sports to carry out in-depth studies on athletes' performance metrics, thereby providing their team with insights to improve future results (Morgulev, 2018).

2. OBJECTIVES

The main objective of this study is to understand how swimmers' reaction time vary with different factors, and whether it provides any advantage for better final performance.

The reaction time (R.T) is defined as the time taken for the feet to leave the starting position after the start signal.

JMP Pro 16.0 is used throughout the processes of data import, preparation, and analysis, using various wrangling, interactive exploration of trends and relationships, as well as hypothesis testing to validate these observations.

3. DATA

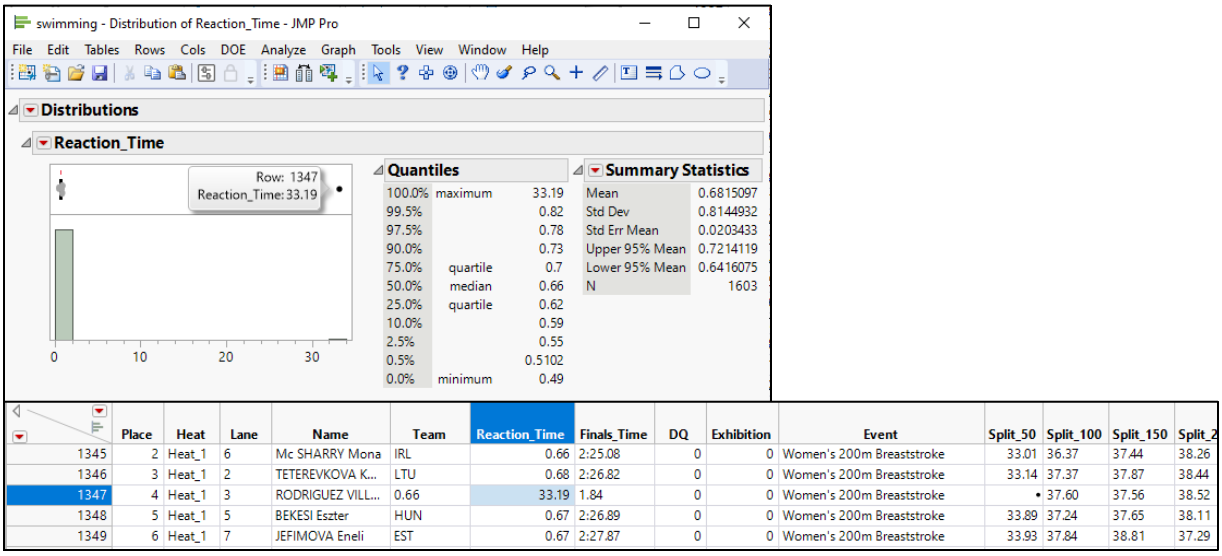
3.1 Data Used

Omega, the official time partner of the Olympics, provided the original data source in PDF form, separated by event (stroke styles), round, and gender. Data was then extracted into a single .csv file named “swimming”, which will be imported into JMP Pro and serve as the main source file for the subsequent data processing and analysis.

3.2 Data Quality Issues and Wrangling Solutions

Issue 1:

From the R.T column, using column view (*Figure 1*), one outlier identified for row 1347 with reaction time of 33.19s – not realistic being generally sub-second. Tracing back to the data table and the original pdf file, the data was not imported correctly for this row, as the R.T of 0.66s was read as “Team”, and the split 50 timing was read as R.T.



Quantiles

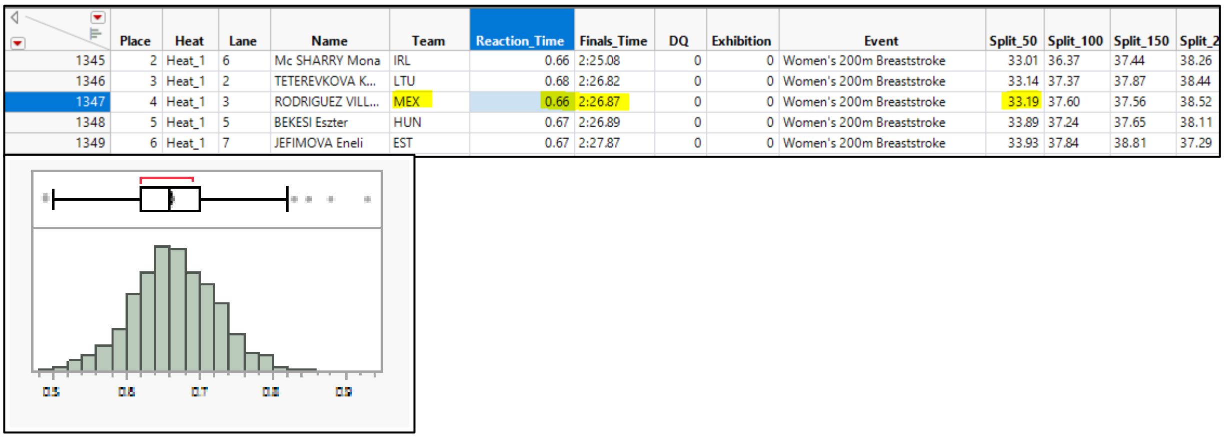
100.0%	maximum	33.19
99.5%		0.82
97.5%		0.78
90.0%		0.73
75.0%	quartile	0.7
50.0%	median	0.66
25.0%	quartile	0.62
10.0%		0.59
2.5%		0.55
0.5%		0.5102
0.0%	minimum	0.49

Summary Statistics

Mean	0.6815097
Std Dev	0.8144932
Std Err Mean	0.0203433
Upper 95% Mean	0.7214119
Lower 95% Mean	0.6416075
N	1603

	Place	Heat	Lane	Name	Team	Reaction_Time	Finals_Time	DQ	Exhibition	Event	Split_50	Split_100	Split_150	Split_2
1345	2	Heat_1	6	Mc SHARRY Mona	IRL	0.66	2:25.08	0	0	Women's 200m Breaststroke	33.01	36.37	37.44	38.26
1346	3	Heat_1	2	TETEREVKOVA K...	LTU	0.68	2:26.82	0	0	Women's 200m Breaststroke	33.14	37.37	37.87	38.44
1347	4	Heat_1	3	RODRIGUEZ VILL...	0.66	33.19	1.84	0	0	Women's 200m Breaststroke	•	37.60	37.56	38.52
1348	5	Heat_1	5	BEKESI Eszter	HUN	0.67	2:26.89	0	0	Women's 200m Breaststroke	33.89	37.24	37.65	38.11
1349	6	Heat_1	7	JEFIMOVA Eneli	EST	0.67	2:27.87	0	0	Women's 200m Breaststroke	33.93	37.84	38.81	37.29

Figure 1: Abnormal R.T

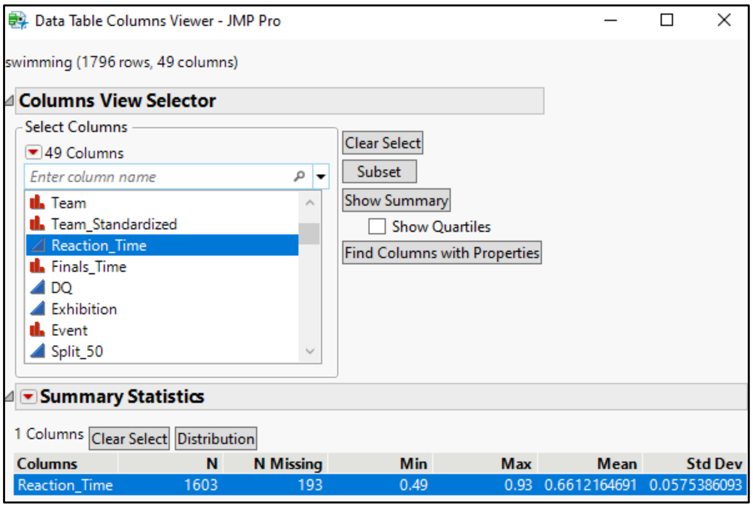


	Place	Heat	Lane	Name	Team	Reaction_Time	Finals_Time	DQ	Exhibition	Event	Split_50	Split_100	Split_150	Split_2
1345	2	Heat_1	6	Mc SHARRY Mona	IRL	0.66	2:25.08	0	0	Women's 200m Breaststroke	33.01	36.37	37.44	38.26
1346	3	Heat_1	2	TETEREVKOVA K...	LTU	0.68	2:26.82	0	0	Women's 200m Breaststroke	33.14	37.37	37.87	38.44
1347	4	Heat_1	3	RODRIGUEZ VILL...	MEX	0.66	2:26.87	0	0	Women's 200m Breaststroke	33.19	37.60	37.56	38.52
1348	5	Heat_1	5	BEKESI Eszter	HUN	0.67	2:26.89	0	0	Women's 200m Breaststroke	33.89	37.24	37.65	38.11
1349	6	Heat_1	7	JEFIMOVA Eneli	EST	0.67	2:27.87	0	0	Women's 200m Breaststroke	33.93	37.84	38.81	37.29

Figure 2: Corrected R.T based on Source PDF and Resulted Overall Distribution

Resolution: As only 1 row is erroneous, editing on the worksheet is acceptable and achieve a reasonable-looking distribution (*Figure 2*)

Issue 2: There are 193 missing values for Reaction time (10.7%) (*Figure 3*).



swimming (1796 rows, 49 columns)

□ Show Quartiles

Columns	N	N Missing	Min	Max	Mean	Std Dev
Reaction_Time	1603	193	0.49	0.93	0.6612164691	0.0575386093

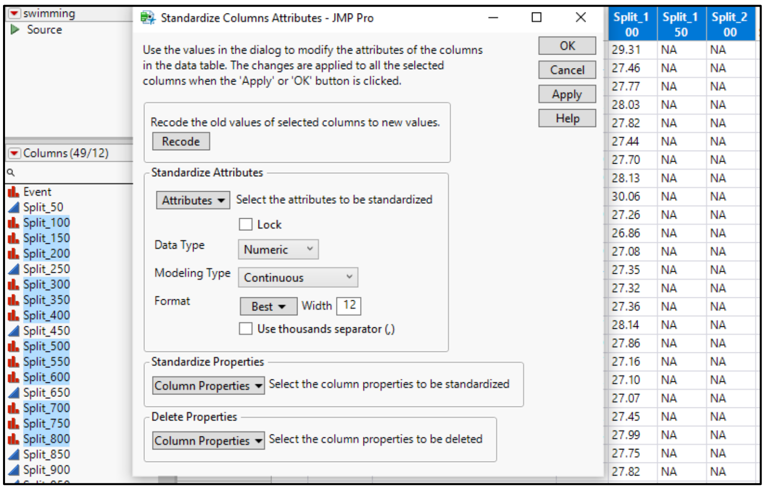
Figure 3: Missing R.T Values

Resolution: By selecting a sample R.T cell with missing value, using Rows > Row Selection > Select Matching cells…, all missing entries can be identified on Row pane. Using Data View on the selected rows, it can be noted that all 7 relay events face this issue (167 out of 193 rows). All these 193 rows will be excluded and hidden from the subsequent analysis.

This is acceptable because this study is focusing on the R.T from the starting signal (the horn sound). In relays, only the first swimmers start with that signal, everyone afterwards starts when the preceding swimmer ends his/her turn. Hence, most of R.T in relays are not relevant to our study. The other 26 data points count for a very small portion, the remaining 1603 data points constitute significant size for statistical conclusions.

Issue 3: Some split timing columns are in categorical data type.

Resolution: formatting them into continuous data type via Standardized Column Attributes (*Figure 4*).



Use the values in the dialog to modify the attributes of the columns in the data table. The changes are applied to all the selected columns when the 'Apply' or 'OK' button is clicked.

Recode the old values of selected columns to new values.

Select the attributes to be standardized

□ Lock

Data Type: Numeric

Modeling Type: Continuous

Format: Best Width 12

□ Use thousands separator (,)

Select the column properties to be standardized

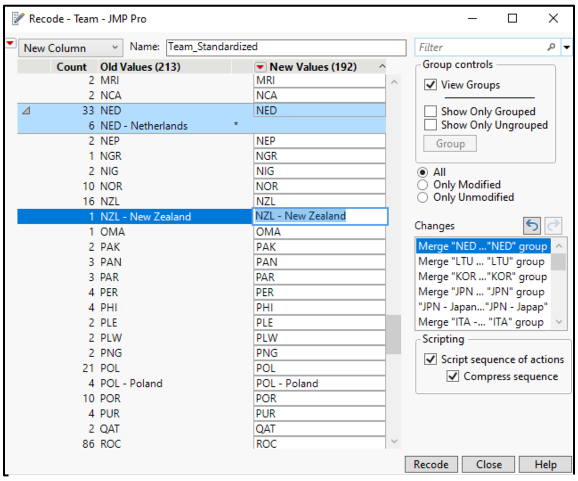
Select the column properties to be deleted

Split_100	Split_150	Split_200
29.31	NA	NA
27.46	NA	NA
27.77	NA	NA
28.03	NA	NA
27.82	NA	NA
27.44	NA	NA
27.70	NA	NA
28.13	NA	NA
30.06	NA	NA
27.26	NA	NA
26.86	NA	NA
27.08	NA	NA
27.35	NA	NA
27.32	NA	NA
27.36	NA	NA
28.14	NA	NA
27.86	NA	NA
27.16	NA	NA
27.10	NA	NA
27.07	NA	NA
27.45	NA	NA
27.99	NA	NA
27.75	NA	NA
27.82	NA	NA

Figure 4: Correction on Split Timing Data Type

Issue 4: Team names are not consistent. E.g., New Zealand, having both “NZL” and “NZL – New Zealand”.

Resolution: Recoding this column to “Team_Standardized” column to correct all repeated and inconsistent values (*Figure 5*).


Recode - Team - JMP Pro

New Column | Name: Team_Standardized

	Count	Old Values (213)		New Values (192)
	2	MRI		MRI
	2	NCA		NCA
⊿	33	NED		NED
	6	NED - Netherlands	*	
	2	NEP		NEP
	1	NGR		NGR
	2	NIG		NIG
	10	NOR		NOR
	16	NZL		NZL
	1	NZL - New Zealand		NZL - New Zealand
	1	OMA		OMA
	2	PAK		PAK
	3	PAN		PAN
	3	PAR		PAR
	4	PER		PER
	4	PHI		PHI
	2	PLE		PLE
	2	PLW		PLW
	2	PNG		PNG
	21	POL		POL
	4	POL - Poland		POL - Poland
	10	POR		POR
	4	PUR		PUR
	2	QAT		QAT
	86	ROC		ROC

Group controls
☑ View Groups
☐ Show Only Grouped
☐ Show Only Ungrouped

● All
○ Only Modified
○ Only Unmodified

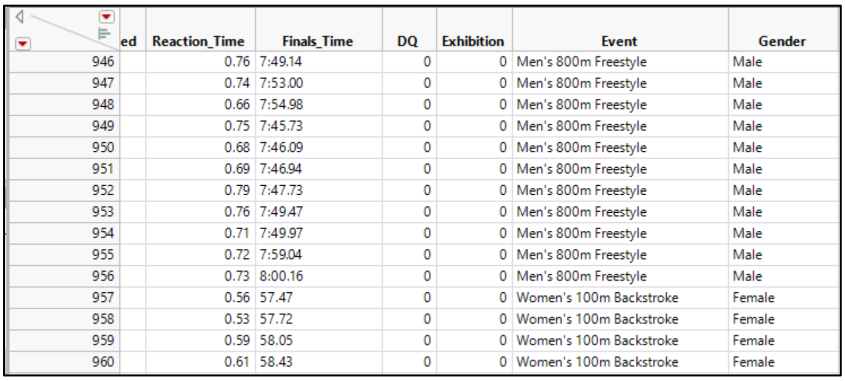
Changes
Merge "NED ..."NED" group
Merge "LTU ... "LTU" group
Merge "KOR ..."KOR" group
Merge "JPN ... "JPN" group
"JPN - Japan..."JPN - Japap"
Merge "ITA -... "ITA" group

Scripting
☑ Script sequence of actions
☑ Compress sequence

Figure 5: Correcting Inconsistent Team Data via Recoding

Issue 5: there is no gender column for separating the population in later analysis.

Resolution: a new “Gender” column is created by recoding the “Event” column, grouping all Men's events into “Male”, all Women's into “Female” (*Figure 6*).

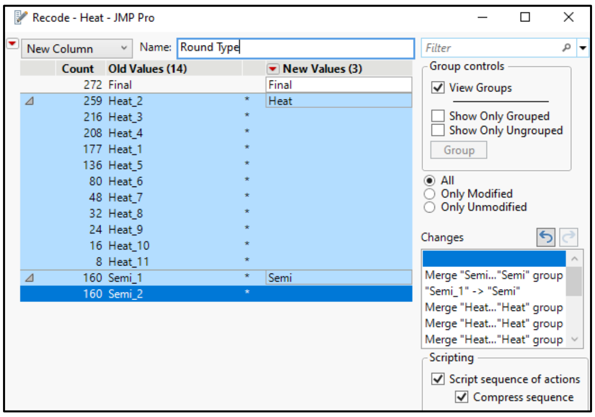


ed	Reaction_Time	Finals_Time	DQ	Exhibition	Event	Gender
946	0.76	7:49.14	0	0	Men's 800m Freestyle	Male
947	0.74	7:53.00	0	0	Men's 800m Freestyle	Male
948	0.66	7:54.98	0	0	Men's 800m Freestyle	Male
949	0.75	7:45.73	0	0	Men's 800m Freestyle	Male
950	0.68	7:46.09	0	0	Men's 800m Freestyle	Male
951	0.69	7:46.94	0	0	Men's 800m Freestyle	Male
952	0.79	7:47.73	0	0	Men's 800m Freestyle	Male
953	0.76	7:49.47	0	0	Men's 800m Freestyle	Male
954	0.71	7:49.97	0	0	Men's 800m Freestyle	Male
955	0.72	7:59.04	0	0	Men's 800m Freestyle	Male
956	0.73	8:00.16	0	0	Men's 800m Freestyle	Male
957	0.56	57.47	0	0	Women's 100m Backstroke	Female
958	0.53	57.72	0	0	Women's 100m Backstroke	Female
959	0.59	58.05	0	0	Women's 100m Backstroke	Female
960	0.61	58.43	0	0	Women's 100m Backstroke	Female

Figure 6: New Gender Column from Event Column

Issue 6: As Heats are preliminary rounds to select the top athletes to semis and finals, they contain swimmers from countries traditionally less developed in the sport and hence the R.T may differ. All the round information is not separated, and combined under the column “Heat”

Resolution: a new column Round Type is recoded from the original column Heat (*Figure 7*).


Recode - Heat - JMP Pro

New Column | Name: Round Type

	Count	Old Values (14)		New Values (3)
	272	Final		Final
⊿	259	Heat_2	*	Heat
	216	Heat_3	*	
	208	Heat_4	*	
	177	Heat_1	*	
	136	Heat_5	*	
	80	Heat_6	*	
	48	Heat_7	*	
	32	Heat_8	*	
	24	Heat_9	*	
	16	Heat_10	*	
	8	Heat_11	*	
⊿	160	Semi_1	*	Semi
	160	Semi_2	*	

Group controls
☑ View Groups
☐ Show Only Grouped
☐ Show Only Ungrouped

● All
○ Only Modified
○ Only Unmodified

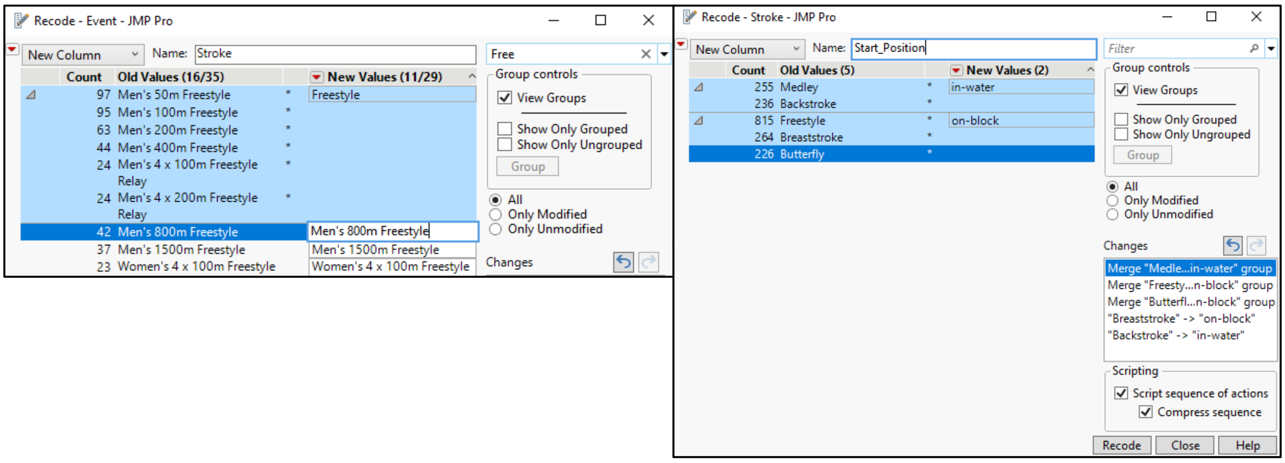
Changes
Merge "Semi..."Semi" group
"Semi_1" -> "Semi"
Merge "Heat..."Heat" group
Merge "Heat..."Heat" group
Merge "Heat..."Heat" group

Scripting
☑ Script sequence of actions
☑ Compress sequence

Figure 7: New Round Type Column from Heat Column via Recoding

Issue 7: There is no separate column for swimming styles of event (e.g., backstroke, freestyle) as a potential factor affecting R.T. Also, backstroke style has a different start position (inside the pool) from the rest of the styles (on the blocks). Hence it is worthwhile to analyze separately by separating the styles into two groups.

Resolution: Using recoding, two new columns "Stroke" and "Start_Position" are created (*Figure 8*).



New Column Name: Stroke

Free

Count	Old Values (16/35)		New Values (11/29)
97	Men's 50m Freestyle	*	Freestyle
95	Men's 100m Freestyle	*	
63	Men's 200m Freestyle	*	
44	Men's 400m Freestyle	*	
24	Men's 4 x 100m Freestyle Relay	*	
24	Men's 4 x 200m Freestyle Relay	*	
42	Men's 800m Freestyle		Men's 800m Freestyle
37	Men's 1500m Freestyle		Men's 1500m Freestyle
23	Women's 4 x 100m Freestyle		Women's 4 x 100m Freestyle

Group controls

☑ View Groups

□ Show Only Grouped

□ Show Only Ungrouped

● All

○ Only Modified

○ Only Unmodified

New Column Name: Start_Position

Filter

Count	Old Values (5)		New Values (2)
255	Medley	*	in-water
236	Backstroke	*	
815	Freestyle	*	on-block
264	Breaststroke	*	
226	Butterfly	*	

Group controls

☑ View Groups

□ Show Only Grouped

□ Show Only Ungrouped

● All

○ Only Modified

○ Only Unmodified

Changes

Merge "Medle...in-water" group
Merge "Freesty...n-block" group
Merge "Butterfl...n-block" group
"Breaststroke" -> "on-block"
"Backstroke" -> "in-water"

Scripting

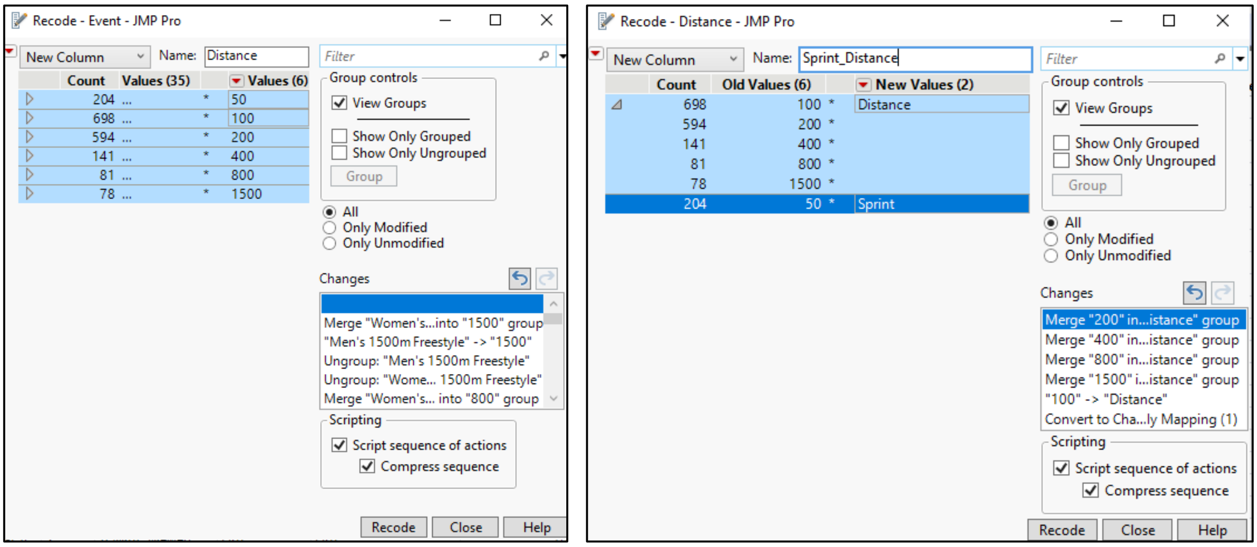
☑ Script sequence of actions

☑ Compress sequence

Figure 8: New Column Stroke and Start_Position Created via Recoding

Issue 8: Besides swimming style, swing distance may affect how the athletes train - sprinter vs distance swimmers - and hence could affect their R.T.

Resolution: A new column "Distance" is created from Event column using the race length information (e.g., 50m, 100m…). Afterwards, another column "Sprint_Distance" which groups all 50m events into "Sprint" and the rest into "Distance" is created.



New Column Name: Distance

Filter

Count	Values (35)		Values (6)
204	...	*	50
698	...	*	100
594	...	*	200
141	...	*	400
81	...	*	800
78	...	*	1500

Group controls

☑ View Groups

□ Show Only Grouped

□ Show Only Ungrouped

● All

○ Only Modified

○ Only Unmodified

Changes

Merge "Women's...into "1500" group
"Men's 1500m Freestyle" -> "1500"
Ungroup: "Men's 1500m Freestyle"
Ungroup: "Wome... 1500m Freestyle"
Merge "Women's... into "800" group

Scripting

☑ Script sequence of actions

☑ Compress sequence

New Column Name: Sprint_Distance

Filter

Count	Old Values (6)		New Values (2)
698	100	*	Distance
594	200	*	
141	400	*	
81	800	*	
78	1500	*	
204	50	*	Sprint

Group controls

☑ View Groups

□ Show Only Grouped

□ Show Only Ungrouped

● All

○ Only Modified

○ Only Unmodified

Changes

Merge "200" in...istance" group
Merge "400" in...istance" group
Merge "800" in...istance" group
Merge "1500" i...istance" group
"100" -> "Distance"
Convert to Cha...ly Mapping (1)

Scripting

☑ Script sequence of actions

☑ Compress sequence

Figure 9: New Column Distance and Sprint_Distance Created via Recoding

The final JMP table is named "Swim_Final".

4. DATA ANALYSIS

4.1 Reaction Time with Gender and Start Position

4.1.1 Subdividing the Populations and Normality Tests

Besides gender being the potential determining factor for differences in R.T (*Figure 10a)* with male swimmers having lower mean R.T than females, the starting positions seem to be another factor. Backstroke starting position is different from inside the pool with their feet against the wall, instead of from the blocks above the pool in the rest of the strokes. Hence backstroke R.T form a separate population from the rest.

Separating all the R.T by the above 2 factors, we can make the following observations based on *Figure 10b*.

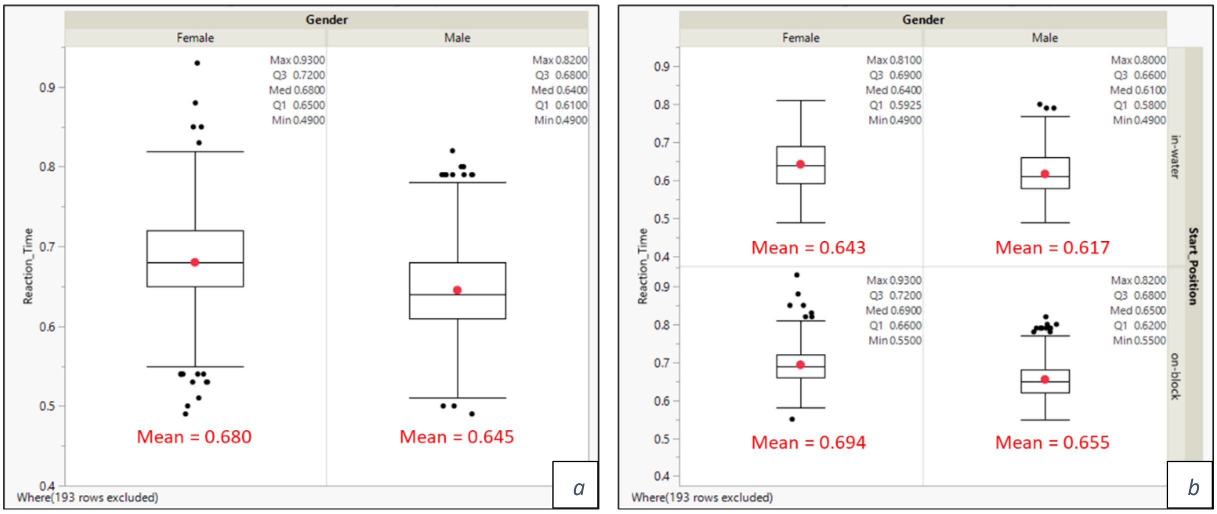


Figure 10: Male vs Female R.T, Dissected by Start_Position

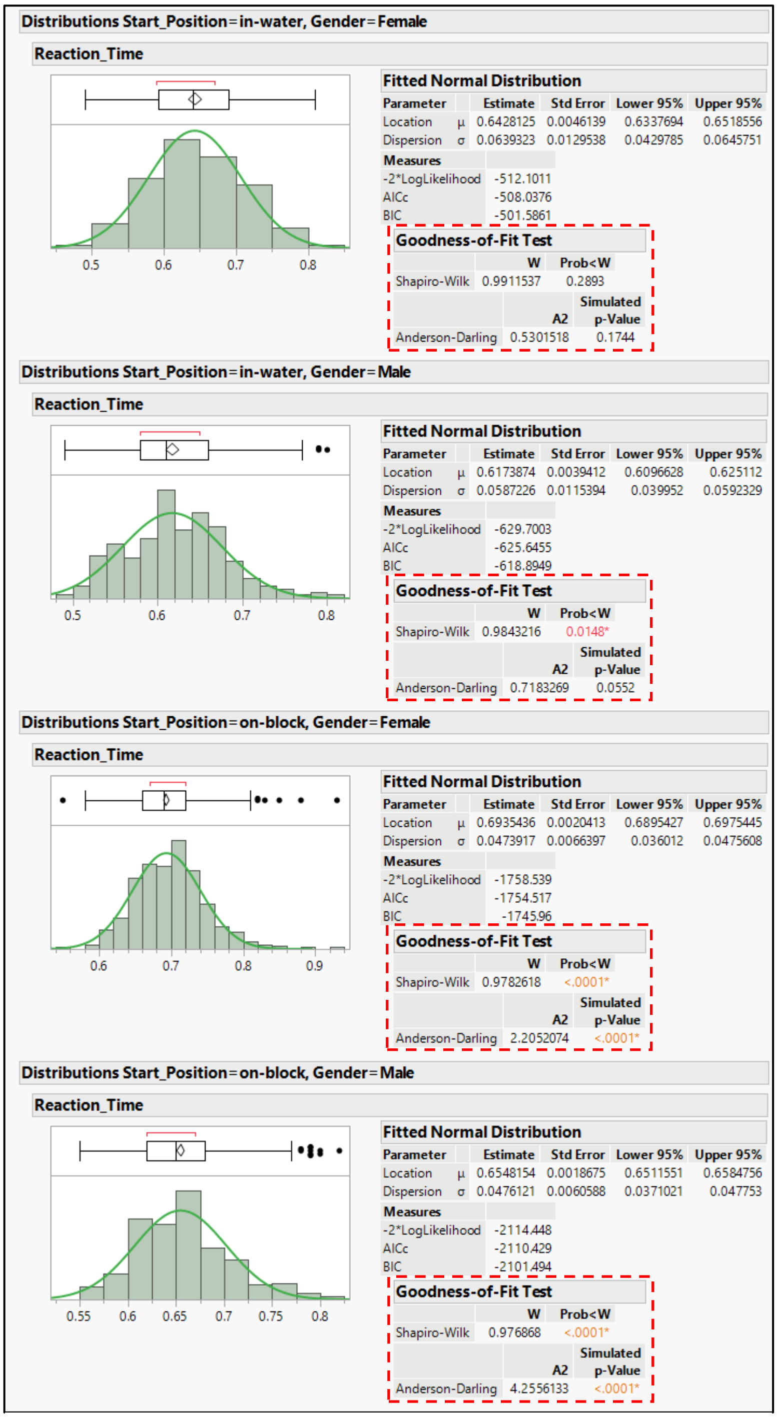
Observation 1: Female in-water starting position has lower mean RT than on-block starting position
Observation 2: Male in-water starting position has lower mean RT than on-block starting position
Observation 3: Male have lower mean R.T than Female for in-water starting position
Observation 4: Male have lower mean R.T than Female for on-block starting position

Group	Gender	Start Position
1	Female	In-water
2	Male	In-water
3	Female	On-block
4	Male	On-block

Table 1: Groupings for R.T vs Gender/Start Position Study

Before selecting appropriate confirmatory analysis tests, it is necessary to perform the normality check on the above groups, with 99% confidence level selected. By the distribution and goodness-of-fit test, *Figure 11* is obtained and summarized in *Table 2* to test the following null and alternative hypotheses.

Null hypothesis H_0: The group R.T is normally distributed.
Alternative hypothesis H_a: The group R.T is not normally distributed.



Distributions Start_Position=in-water, Gender=Female

Reaction_Time

Fitted Normal Distribution

Parameter		Estimate	Std Error	Lower 95%	Upper 95%
Location	μ	0.6428125	0.0046139	0.6337694	0.6518556
Dispersion	σ	0.0639323	0.0129538	0.0429785	0.0645751

Measures	
-2*LogLikelihood	-512.1011
AICc	-508.0376
BIC	-501.5861

Goodness-of-Fit Test

	W	Prob<W
Shapiro-Wilk	0.9911537	0.2893

	A2	Simulated p-Value
Anderson-Darling	0.5301518	0.1744

Distributions Start_Position=in-water, Gender=Male

Reaction_Time

Fitted Normal Distribution

Parameter		Estimate	Std Error	Lower 95%	Upper 95%
Location	μ	0.6173874	0.0039412	0.6096628	0.625112
Dispersion	σ	0.0587226	0.0115394	0.039952	0.0592329

Measures	
-2*LogLikelihood	-629.7003
AICc	-625.6455
BIC	-618.8949

Goodness-of-Fit Test

	W	Prob<W
Shapiro-Wilk	0.9843216	0.0148*

	A2	Simulated p-Value
Anderson-Darling	0.7183269	0.0552

Distributions Start_Position=on-block, Gender=Female

Reaction_Time

Fitted Normal Distribution

Parameter		Estimate	Std Error	Lower 95%	Upper 95%
Location	μ	0.6935436	0.0020413	0.6895427	0.6975445
Dispersion	σ	0.0473917	0.0066397	0.036012	0.0475608

Measures	
-2*LogLikelihood	-1758.539
AICc	-1754.517
BIC	-1745.96

Goodness-of-Fit Test

	W	Prob<W
Shapiro-Wilk	0.9782618	<.0001*

	A2	Simulated p-Value
Anderson-Darling	2.2052074	<.0001*

Distributions Start_Position=on-block, Gender=Male

Reaction_Time

Fitted Normal Distribution

Parameter		Estimate	Std Error	Lower 95%	Upper 95%
Location	μ	0.6548154	0.0018675	0.6511551	0.6584756
Dispersion	σ	0.0476121	0.0060588	0.0371021	0.047753

Measures	
-2*LogLikelihood	-2114.448
AICc	-2110.429
BIC	-2101.494

Goodness-of-Fit Test

	W	Prob<W
Shapiro-Wilk	0.976868	<.0001*

	A2	Simulated p-Value
Anderson-Darling	4.2556133	<.0001*

Figure 11: Distribution of Group 1-4 RT

Group	Gender	Start Position	Prob<W (Shapiro-Wilk)	p-value (Anderson-Darling)	Conclusion on H_0
1	Female	In-water	0.2893	0.1744	Not rejected
2	Male	In-water	0.0148	0.0552	Not rejected
3	Female	On-block	<0.0001	<0.0001	Rejected
4	Male	On-block	<0.0001	<0.0001	Rejected

Table 2: Normality Test Results for Group 1-4

From *Table 2*, based on both Shapiro-Wilk and Anderson-Darling goodness-of-fit tests, we can infer that for in-water starting position, the R.T follow the normal distribution, while for other strokes the R.T distribution is not normal. Hence, the type of hypothesis tests with the null hypotheses are:

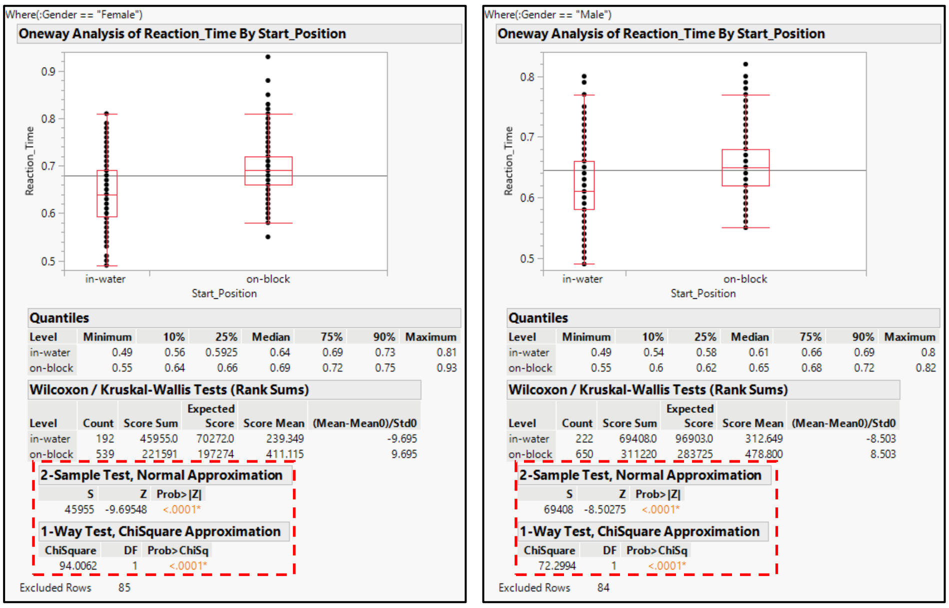
Observation No.	Comparison Pairs between Group:			Hypothesis No.	Type of Hypothesis Test	Null Hypothesis H_0
1	1	vs	3	A	Nonparametric	The mean R.T of female is the same for in-water and on-block starting positions.
2	2	vs	4	B	Nonparametric	The mean R.T of male is the same for in-water and on-block starting positions.
3	1	vs	2	C	Parametric	The mean R.T of female and male is the same for in-water starting positions.
4	3	vs	4	D	Nonparametric	The mean R.T of female and male is the same for on-block starting positions.

Table 3: Hypothesis Tests Applied to Group 1-4 Comparison

4.1.2 Starting Position and Reaction Time

Firstly, the difference in mean R.T between the two starting positions is studied for each gender.

From *Figure 12*, for the nonparametric tests on hypothesis A and B, with the confidence level of 99%, the p-values are lower than the alpha value of 0.01. Therefore, we reject the null hypotheses A and B, and infer that for both male and female swimmers, the mean R.T in for in-water starting position is lower than on-block starting position.



Where(:Gender == "Female")

Oneway Analysis of Reaction_Time By Start_Position

Quantiles

Level	Minimum	10%	25%	Median	75%	90%	Maximum
in-water	0.49	0.56	0.5925	0.64	0.69	0.73	0.81
on-block	0.55	0.64	0.66	0.69	0.72	0.75	0.93

Wilcoxon / Kruskal-Wallis Tests (Rank Sums)

Level	Count	Score Sum	Expected Score	Score Mean	(Mean-Mean0)/Std0
in-water	192	45955.0	70272.0	239.349	-9.695
on-block	539	221591	197274	411.115	9.695

2-Sample Test, Normal Approximation

S	Z	Prob>|Z|
45955	-9.69548	<.0001*

1-Way Test, ChiSquare Approximation

ChiSquare	DF	Prob>ChiSq
94.0062	1	<.0001*

Excluded Rows 85

Where(:Gender == "Male")

Oneway Analysis of Reaction_Time By Start_Position

Quantiles

Level	Minimum	10%	25%	Median	75%	90%	Maximum
in-water	0.49	0.54	0.58	0.61	0.66	0.69	0.8
on-block	0.55	0.6	0.62	0.65	0.68	0.72	0.82

Wilcoxon / Kruskal-Wallis Tests (Rank Sums)

Level	Count	Score Sum	Expected Score	Score Mean	(Mean-Mean0)/Std0
in-water	222	69408.0	96903.0	312.649	-8.503
on-block	650	311220	283725	478.800	8.503

2-Sample Test, Normal Approximation

S	Z	Prob>|Z|
69408	-8.50275	<.0001*

1-Way Test, ChiSquare Approximation

ChiSquare	DF	Prob>ChiSq
72.2994	1	<.0001*

Excluded Rows 84

Figure 12: Hypothesis Test A and B Results for Mean R.T vs Start_Position

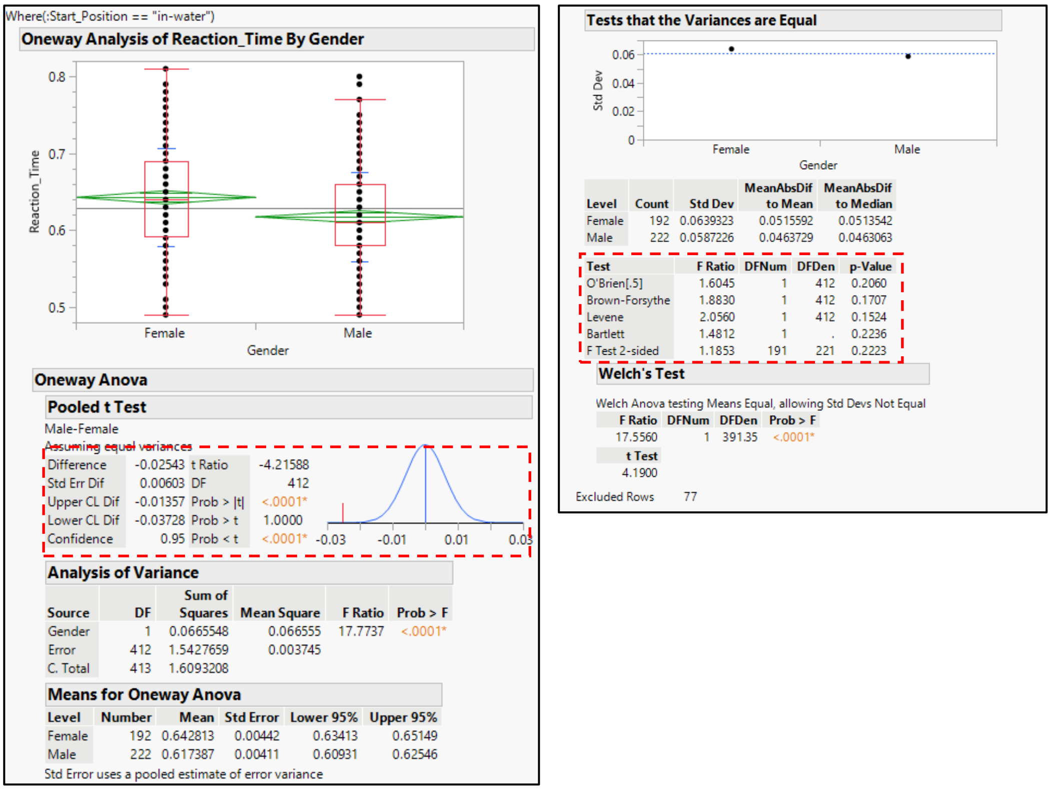
This is a reasonable inference as swimmers may need to exert more force when starting on the blocks to push propel their bodyweight both vertically and horizontally into air before entering the pool. In the case of backstroke, the athletes exert force mostly in the horizontal direction. Although the drag force under water is higher than air, for the short distance of R.T, this appears to be a less significant factor.

4.1.3 Gender and Reaction Time

For hypothesis C, we are checking if the mean R.T of male and female athletes are the same for in-water starting position. By conducting both the parametric Pooled t Test and the Welch's Test, the result in *Figure 13* is obtained.

Under the Pooled t Test (equal variances assumption), from the two-tail test result (Prob > |t|: <0.0001), the observed difference in mean R.T is statistically significant at 99% confidence level. Hence, we reject the null hypothesis that the 2 means are the same. From the left-tail test result (Prob < t: <0.0001), the observation of male R.T being "less than" female R.T is statistically significant, and we can infer that male athletes' mean R.T is lower than that of female athletes.

Under the Equal Variances test, as the p-values across the tests are all higher than the alpha value of 0.01, we cannot reject the null hypothesis that the variances of these 2 populations are equal. As these variances are inferred to be the same, we can maintain our conclusion from the previous Pooled t Test.



Where(:Start_Position == "in-water")

Oneway Analysis of Reaction_Time By Gender

Oneway Anova

Pooled t Test

Male-Female

~~Assuming equal variances~~

Difference	-0.02543	t Ratio	-4.21588
Std Err Dif	0.00603	DF	412
Upper CL Dif	-0.01357	Prob > |t|	<.0001*
Lower CL Dif	-0.03728	Prob > t	1.0000
Confidence	0.95	Prob < t	<.0001*

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Gender	1	0.0665548	0.066555	17.7737	<.0001*
Error	412	1.5427659	0.003745		
C. Total	413	1.6093208			

Means for Oneway Anova

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
Female	192	0.642813	0.00442	0.63413	0.65149
Male	222	0.617387	0.00411	0.60931	0.62546

Std Error uses a pooled estimate of error variance

Tests that the Variances are Equal

Level	Count	Std Dev	MeanAbsDif to Mean	MeanAbsDif to Median
Female	192	0.0639323	0.0515592	0.0513542
Male	222	0.0587226	0.0463729	0.0463063

Test	F Ratio	DFNum	DFDen	p-Value
O'Brien[.5]	1.6045	1	412	0.2060
Brown-Forsythe	1.8830	1	412	0.1707
Levene	2.0560	1	412	0.1524
Bartlett	1.4812	1	.	0.2236
F Test 2-sided	1.1853	191	221	0.2223

Welch's Test

Welch Anova testing Means Equal, allowing Std Devs Not Equal

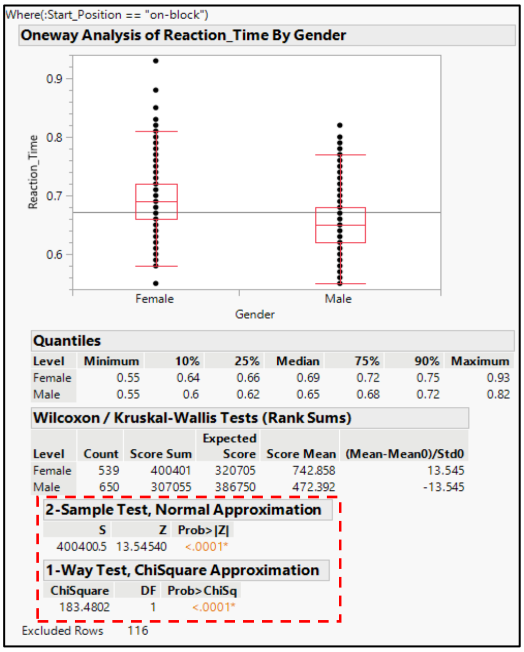
F Ratio	DFNum	DFDen	Prob > F
17.5560	1	391.35	<.0001*

t Test
4.1900

Excluded Rows 77

Figure 13: Hypothesis Test C Results for Mean R.T vs Gender (in-water)

For hypothesis D, we are repeating the hypothesis C for the on-block staring position. Applying the nonparametric test for these 2 populations of female and male mean R.T, we obtain the result in *Figure 14*. From both tests, the p-value is lower than the specified alpha level of 0.01. Thefore the null hypothesis that these two mean R.T values are the same is rejected, and it can be inferred that the observed lower mean R.T from male swimmers is statistically significant. This is not a surprising result as male athletes with generally stronger physique, hence higher ability to generate force and to have better R.T performance.



Where(:Start_Position == "on-block")

Oneway Analysis of Reaction_Time By Gender

Quantiles

Level	Minimum	10%	25%	Median	75%	90%	Maximum
Female	0.55	0.64	0.66	0.69	0.72	0.75	0.93
Male	0.55	0.6	0.62	0.65	0.68	0.72	0.82

Wilcoxon / Kruskal-Wallis Tests (Rank Sums)

Level	Count	Score Sum	Expected Score	Score Mean	(Mean-Mean0)/Std0
Female	539	400401	320705	742.858	13.545
Male	650	307055	386750	472.392	-13.545

2-Sample Test, Normal Approximation

S	Z	Prob>|Z|
400400.5	13.54540	<.0001*

1-Way Test, ChiSquare Approximation

ChiSquare	DF	Prob>ChiSq
183.4802	1	<.0001*

Excluded Rows 116

Figure 14: Hypothesis Test D Results for Mean R.T vs Gender (on-block)

4.2 Swimming Events and Reaction Time

4.2.1 Reaction Time Between Sprinter and Long-Distance Swimmers

In this section, the difference in mean R.T between sprint swimmers and distance swimmers is explored. In this data set, 50m distance is defined as “sprint” and further distances are defined as “distance”. As 50m event is only applicable to freestyle stroke in this Olympics, our analysis will be limited to freestyle events for a fair comparison of swimmer population. A separate JMP table with only this stroke is created (“Swimming_FreestyleOnly”) from the table “Swimming_Final” using the Row > Select Matchig Cells function.

Figure 15 shows that:

Observation 5: R.T of female sprinter is better than female distance swimmer
Observation 6: R.T of male sprinter is better than female distance swimmer

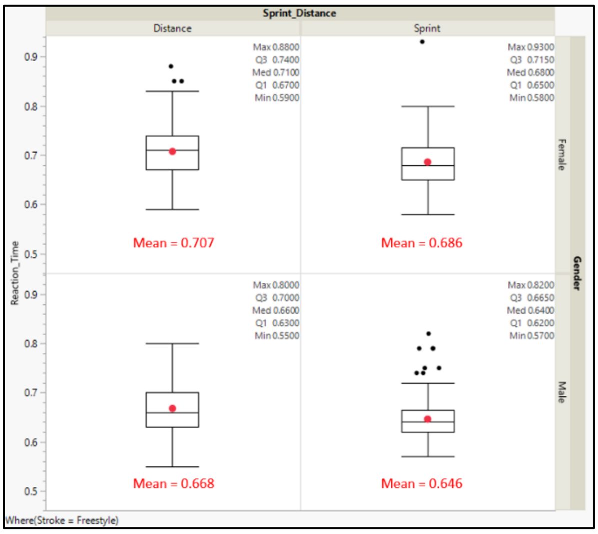
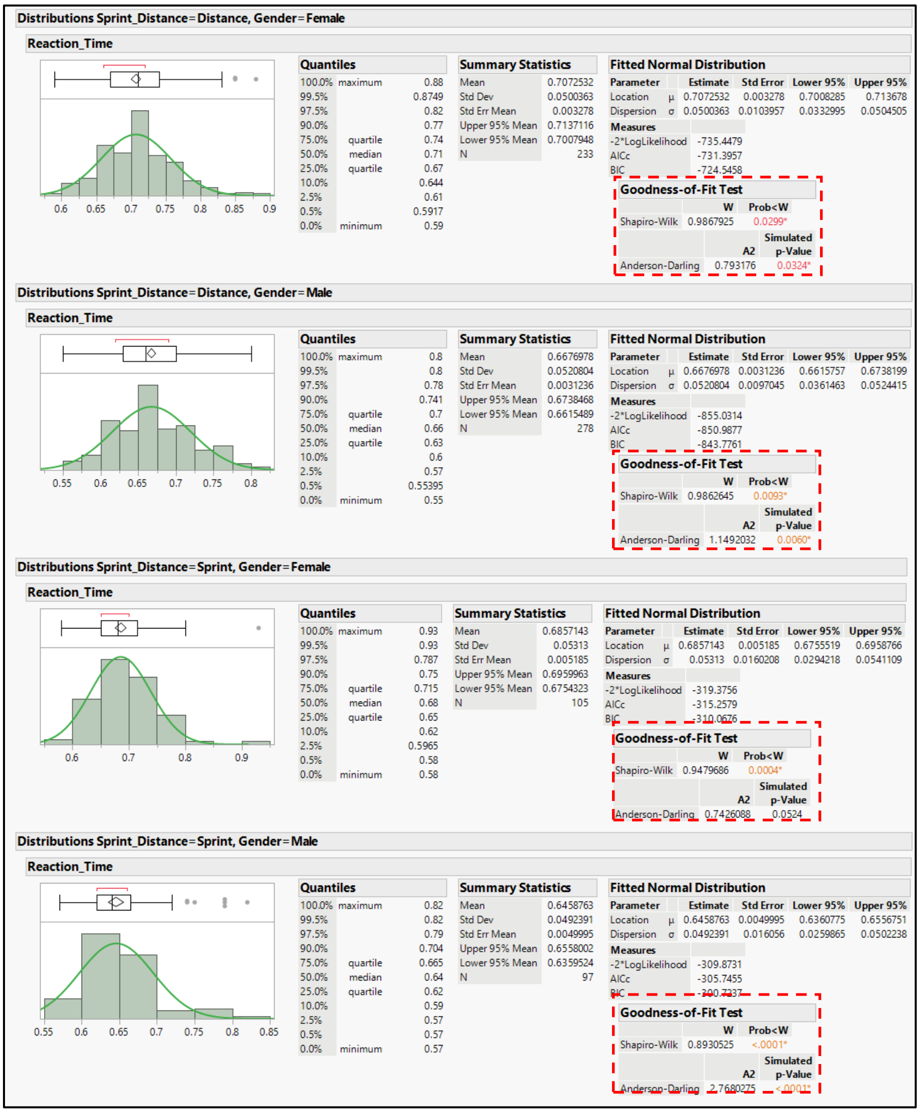


Figure 15: R.T vs Sprint and Distance Events

Similar to previous analyses, normal distribution check is done followed by confirmatory hypothesis testing to check if the above observation is statistically significant. Similar to Hypothesis 0, the following normality test aims at determining the distrbution types of sprinter and distance swimmers’ RT.

Hypothesis E: sprinter and distance swimmers’ RT is normally distributed.

By the distribution and goodness-of-fit test, the following results in *Figure 16* are obtained, and summarized in *Table 4*, based on which the appropriate tests are listed in *Table 5*.



Distributions Sprint_Distance=Distance, Gender=Female

Reaction_Time

Quantiles		
100.0%	maximum	0.88
99.5%		0.8749
97.5%		0.82
90.0%		0.77
75.0%	quartile	0.74
50.0%	median	0.71
25.0%	quartile	0.67
10.0%		0.644
2.5%		0.61
0.5%		0.5917
0.0%	minimum	0.59

Summary Statistics	
Mean	0.7072532
Std Dev	0.0500363
Std Err Mean	0.003278
Upper 95% Mean	0.7137116
Lower 95% Mean	0.7007948
N	233

Fitted Normal Distribution

Parameter		Estimate	Std Error	Lower 95%	Upper 95%
Location	μ	0.7072532	0.003278	0.7008285	0.713678
Dispersion	σ	0.0500363	0.0103957	0.0332995	0.0504505

Measures	
-2*LogLikelihood	-735.4479
AICc	-731.3957
BIC	-724.5458

Goodness-of-Fit Test

	W	Prob<W
Shapiro-Wilk	0.9867925	0.0299*

	A2	Simulated p-Value
Anderson-Darling	0.793176	0.0324*

Distributions Sprint_Distance=Distance, Gender=Male

Reaction_Time

Quantiles		
100.0%	maximum	0.8
99.5%		0.8
97.5%		0.78
90.0%		0.741
75.0%	quartile	0.7
50.0%	median	0.66
25.0%	quartile	0.63
10.0%		0.6
2.5%		0.57
0.5%		0.55395
0.0%	minimum	0.55

Summary Statistics	
Mean	0.6676978
Std Dev	0.0520804
Std Err Mean	0.0031236
Upper 95% Mean	0.6738468
Lower 95% Mean	0.6615489
N	278

Fitted Normal Distribution

Parameter		Estimate	Std Error	Lower 95%	Upper 95%
Location	μ	0.6676978	0.0031236	0.6615757	0.6738199
Dispersion	σ	0.0520804	0.0097045	0.0361463	0.0524415

Measures	
-2*LogLikelihood	-855.0314
AICc	-850.9877
BIC	-843.7761

Goodness-of-Fit Test

	W	Prob<W
Shapiro-Wilk	0.9862645	0.0093*

	A2	Simulated p-Value
Anderson-Darling	1.1492032	0.0060*

Distributions Sprint_Distance=Sprint, Gender=Female

Reaction_Time

Quantiles		
100.0%	maximum	0.93
99.5%		0.93
97.5%		0.787
90.0%		0.75
75.0%	quartile	0.715
50.0%	median	0.68
25.0%	quartile	0.65
10.0%		0.62
2.5%		0.5965
0.5%		0.58
0.0%	minimum	0.58

Summary Statistics	
Mean	0.6857143
Std Dev	0.05313
Std Err Mean	0.005185
Upper 95% Mean	0.6959963
Lower 95% Mean	0.6754323
N	105

Fitted Normal Distribution

Parameter		Estimate	Std Error	Lower 95%	Upper 95%
Location	μ	0.6857143	0.005185	0.6755519	0.6958766
Dispersion	σ	0.05313	0.0160208	0.0294218	0.0541109

Measures	
-2*LogLikelihood	-319.3756
AICc	-315.2579
BIC	-310.0676

Goodness-of-Fit Test

	W	Prob<W
Shapiro-Wilk	0.9479686	0.0004*

	A2	Simulated p-Value
Anderson-Darling	0.7426088	0.0524

Distributions Sprint_Distance=Sprint, Gender=Male

Reaction_Time

Quantiles		
100.0%	maximum	0.82
99.5%		0.82
97.5%		0.79
90.0%		0.704
75.0%	quartile	0.665
50.0%	median	0.64
25.0%	quartile	0.62
10.0%		0.59
2.5%		0.57
0.5%		0.57
0.0%	minimum	0.57

Summary Statistics	
Mean	0.6458763
Std Dev	0.0492391
Std Err Mean	0.0049995
Upper 95% Mean	0.6558002
Lower 95% Mean	0.6359524
N	97

Fitted Normal Distribution

Parameter		Estimate	Std Error	Lower 95%	Upper 95%
Location	μ	0.6458763	0.0049995	0.6360775	0.6556751
Dispersion	σ	0.0492391	0.016056	0.0259865	0.0502238

Measures	
-2*LogLikelihood	-309.8731
AICc	-305.7455
BIC	-300.7237

Goodness-of-Fit Test

	W	Prob<W
Shapiro-Wilk	0.8930525	<.0001*

	A2	Simulated p-Value
Anderson-Darling	2.7680275	<.0001*

Figure 16: Distribution of R.T in Groups 5-8

Group	Gender	Sprint or Distance	Prob<W (Shapiro-Wilk)	p-value (Anderson-Darling)	Conclusion on H_0
5	Female	Distance	0.0299	0.0324	Accepted
6	Female	Sprint	0.0004*	0.0524	Accepted
7	Male	Distance	0.0093	0.0060	Rejected
8	Male	Sprint	<0.0001	<0.0001	Rejected

Table 4: Normality Test Results for Group 5-8

*Note that the Shapiro-Wilk test does not support the Null hypothesis of normal distribution for this group. Hence both parametric and non parametric tests will be performed for conservatism under hypothesis test F below (*Table 5*).

Observation No.	Comparison Pairs			Hypothesis Test No.	Type of Hypothesis Test	Null Hypothesis
5	5	vs	6	F	Parametric and Nonparametric	The mean R.T of sprinter and long-distance female swimmers is the same.
6	7	vs	8	G	Nonparametric	The mean R.T of sprinter and long-distance male swimmers is the same.

Table 5: Hypothesis Tests Applied to Group 5-8 Comparison

For hypothesis test F, we are interested to find out if the mean R.T of sprinter and long-distance is the same for female swimmers.

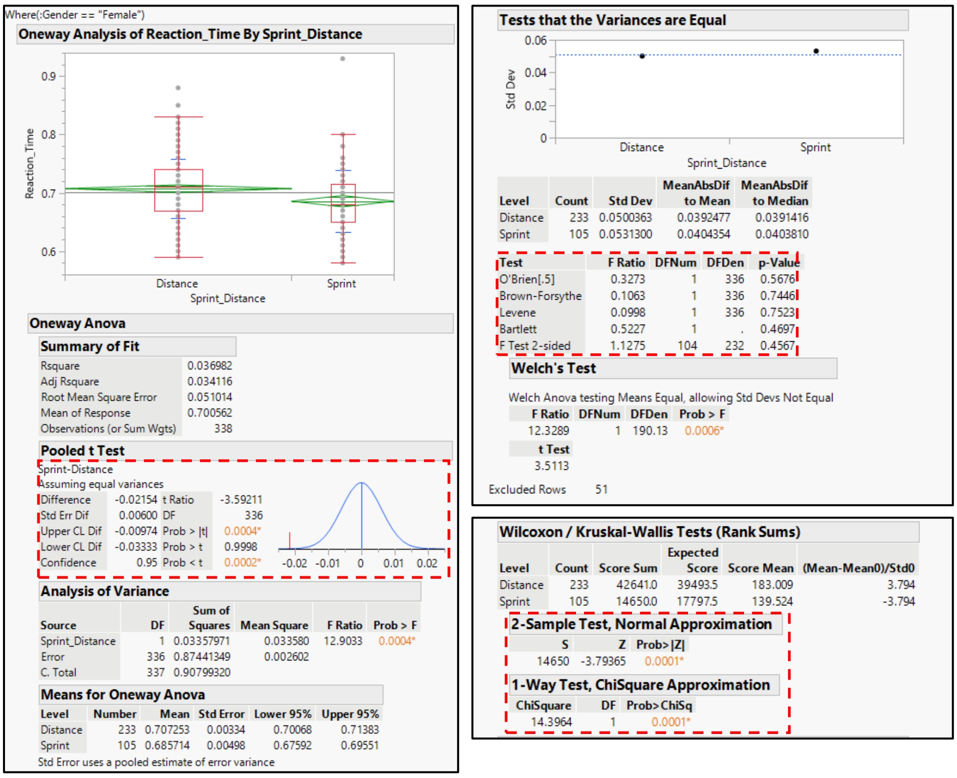
In *Figure 17* Pooled t Test (equal variances assumption), the two-tail test result (Prob > |t|: <0.0004) shows the difference in mean R.T is statistically significant at 99% confidence level. Hence, we reject the null hypothesis that the 2 means are the same. From the left-tail test result (Prob < t: <0.0002), the observation of sprinters' R.T being "less than" long-distance swimmers' R.T is statistically significant, and we can infer that sprinters' mean R.T is lower than that of long-distance swimmers.

Under the Equal Variances test, the p-values across the tests are all greater than the alpha value of 0.01, we cannot reject the null hypothesis of equal variances. As these variances are inferred to be the same, we can maintain our view from the previous pooled T-test.

Under the nonparametric test (Wilcoxon), the p-values are lower than the alpha value, hence we reject the null hypothesis and infer that they have different mean R.T, and that the observed lower R.T from the sprint swimmers is statistically supported.

The same analysis can be done on the male population using nonparametric test for hypothesis G. From the result in F*igure 18*, the p-value is again lower than the specified alpha level of 0.01, hence the null hypothesis can be rejected. Similar inference to that of female population can be drawn that sprinters have lower mean R.T than long-distance swimmers.

Note that swimmers who compete in the 50m events can also compete in the longer distance events. Hence their R.T may reduce the mean R.T among the longer distance population. Despite this, the long-distance population still have statistically longer R.T for both male and female, further reinforcing our inferences above.



Where(:Gender == "Female")

Oneway Analysis of Reaction_Time By Sprint_Distance

Oneway Anova

Summary of Fit

Rsquare	0.036982
Adj Rsquare	0.034116
Root Mean Square Error	0.051014
Mean of Response	0.700562
Observations (or Sum Wgts)	338

Pooled t Test

Sprint-Distance

Assuming equal variances

Difference	-0.02154	t Ratio	-3.59211
Std Err Dif	0.00600	DF	336
Upper CL Dif	-0.00974	Prob > |t|	0.0004*
Lower CL Dif	-0.03333	Prob > t	0.9998
Confidence	0.95	Prob < t	0.0002*

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Sprint_Distance	1	0.03357971	0.033580	12.9033	0.0004*
Error	336	0.87441349	0.002602		
C. Total	337	0.90799320			

Means for Oneway Anova

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
Distance	233	0.707253	0.00334	0.70068	0.71383
Sprint	105	0.685714	0.00498	0.67592	0.69551

Std Error uses a pooled estimate of error variance

Tests that the Variances are Equal

Level	Count	Std Dev	MeanAbsDif to Mean	MeanAbsDif to Median
Distance	233	0.0500363	0.0392477	0.0391416
Sprint	105	0.0531300	0.0404354	0.0403810

Test	F Ratio	DFNum	DFDen	p-Value
O'Brien[.5]	0.3273	1	336	0.5676
Brown-Forsythe	0.1063	1	336	0.7446
Levene	0.0998	1	336	0.7523
Bartlett	0.5227	1	.	0.4697
F Test 2-sided	1.1275	104	232	0.4567

Welch's Test

Welch Anova testing Means Equal, allowing Std Devs Not Equal

F Ratio	DFNum	DFDen	Prob > F
12.3289	1	190.13	0.0006*

t Test
3.5113

Excluded Rows 51

Wilcoxon / Kruskal-Wallis Tests (Rank Sums)

Level	Count	Score Sum	Expected Score	Score Mean	(Mean-Mean0)/Std0
Distance	233	42641.0	39493.5	183.009	3.794
Sprint	105	14650.0	17797.5	139.524	-3.794

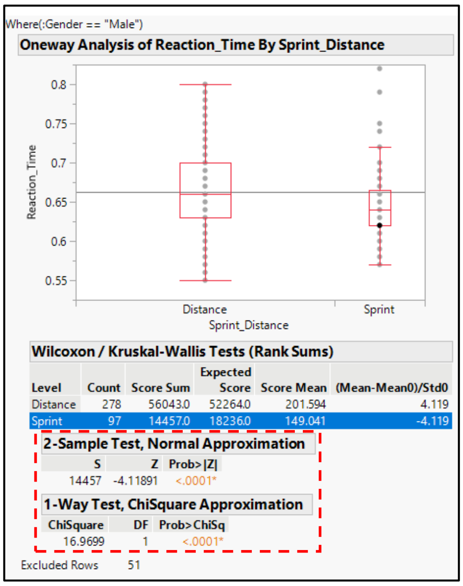
2-Sample Test, Normal Approximation

S	Z	Prob>|Z|
14650	-3.79365	0.0001*

1-Way Test, ChiSquare Approximation

ChiSquare	DF	Prob>ChiSq
14.3964	1	0.0001*

Figure 17: Hypothesis Test F Results for Mean R.T vs Sprint_Distance (Female)



Where(:Gender == "Male")

Oneway Analysis of Reaction_Time By Sprint_Distance

Wilcoxon / Kruskal-Wallis Tests (Rank Sums)

Level	Count	Score Sum	Expected Score	Score Mean	(Mean-Mean0)/Std0
Distance	278	56043.0	52264.0	201.594	4.119
Sprint	97	14457.0	18236.0	149.041	-4.119

2-Sample Test, Normal Approximation

S	Z	Prob>|Z|
14457	-4.11891	<.0001*

1-Way Test, ChiSquare Approximation

ChiSquare	DF	Prob>ChiSq
16.9699	1	<.0001*

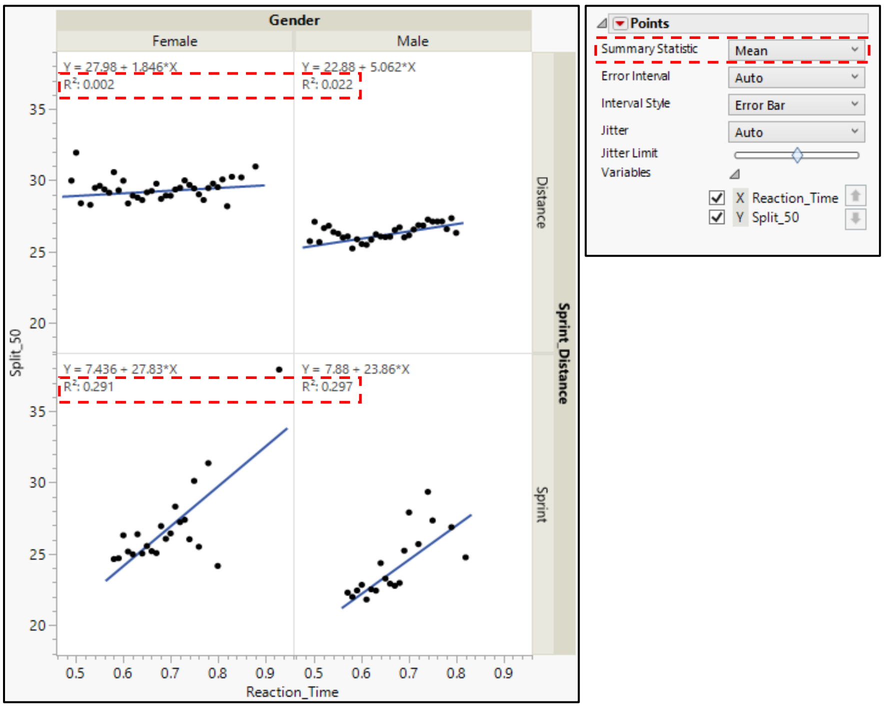
Excluded Rows 51

Figure 18: Hypothesis Test G Results for Mean R.T vs Sprint_Distance (Male)

4.2.2 Reaction Time Correlation with Split Performance and Final Placements

In this section, the impact of having short R.T on performance is explored. Specifically, it is reasonable to split the analysis by not only gender, but also the sprint-distance category above. This is because having a short R.T is more critical to sprinters given the lack of race distance to make up for any poor start. Additionally, the metrics for performance is chosen to be the 50m split timing as this is applicable to both groups. By plotting the mean split 50 timing against the R.T, we can form the foll owing observation. Note that the mean (at each R.T increment on the X-axis) is chosen for better illustration, and this does not change the correlation observed.

Observation 7: R.T has a positive correlation to 50m split performance for sprint swimmers



Summary Statistic: Mean
Error Interval: Auto
Interval Style: Error Bar
Jitter: Auto
Jitter Limit
Variables
☑ X Reaction_Time
☑ Y Split_50

Figure 19: Correlation Between Split 50m Timing and R.T

From *Figure 19*, while there is a positive linear correlation between the R.T and the split_50 timing performance among male and female sprinters (R^2~0.3), there is almost no linear correlation in longer distance events. This is reasonable as distance swimmers may have other focus areas to improve their final placement, such as form, stamina, swimming tactics… rather than relying on good R.T.

Extending this check to the split timing beyond 50m:

Observation 8: R.T has no correlation to further split performance.

In *Figure 20*, there is effectively no linear correlation between R.T and split performance in further distances with R^2 being extremely small. We can infer that the early advantage in R.T is not a significant factor to swimmers’ timing performance further down the race, as other factors explained above will likely to have a greater relationships to timing performance.

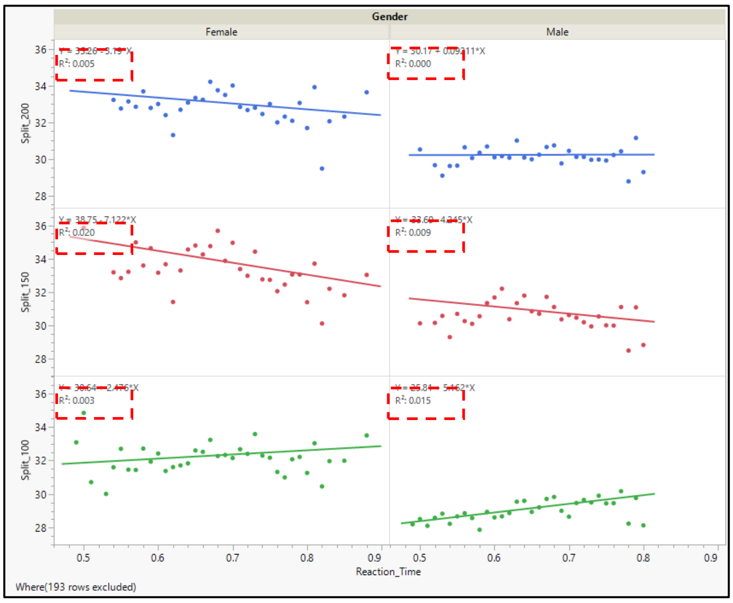


Figure 20: The Lack of Linear Correlation between R.T and Further Split Timings

4.3 Other Aggregate Observations

4.3.1 The Outliers

When the aggregate distribution of male and female R.T is plotted in *Figure 21:*

Observation 9: there are more outliers in the heats, followed by the semifinals, and least in the finals.

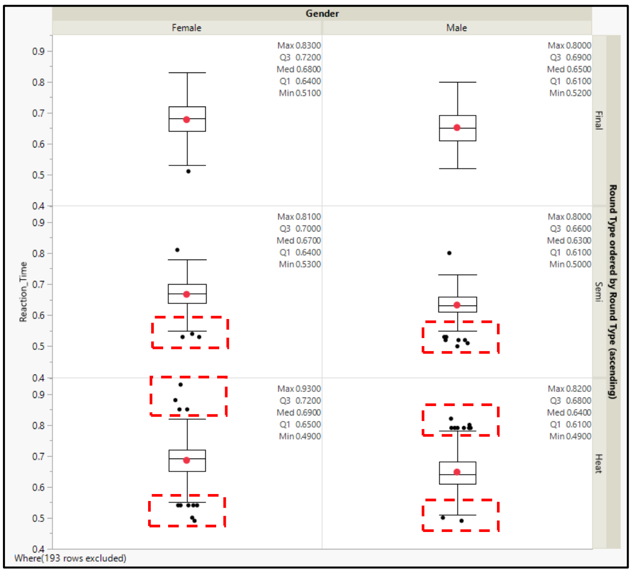


Figure 21: Frequency of Outliers Across Rounds

During heats, swimmers' ability spectrum is the widest as strong medal contenders race with others, many of whom are from countries less developed in the sport. Hence it is likely to find outliers with respect to the quartiles of the population of heat swimmers. As outliers are found in both upper and lower ends of the quartiles, their R.T performance is likely to cancel off each other on average, leading to similar mean and median.

4.3.2 The R.T of Past Good Performers within a Race

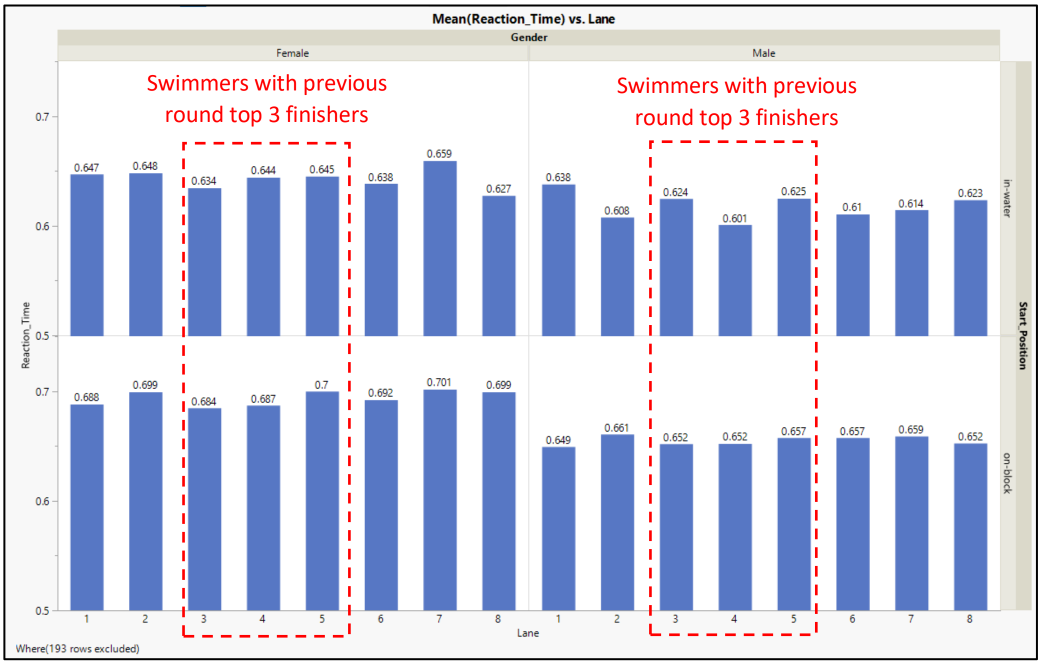
In swimming events, athletes with better timing in previous rounds are placed into the middle lanes:

Lane #	1	2	3	4	5	6	7	8
Qualifying Timing Position	7	5	3	1	2	4	6	8

Hence it is reasonable to associate these lanes to high past performers, and their R.T can be placed in comparison with the others. When plotting R.T by the lane positions (*Figure 22*):

Observation 10: past good performing swimmers based on their placement into the middle lanes do not have outstanding R.T.

Despite the previous distinction in R.T between sprinters and long-distance swimmers, when comparing with competitors on the same event, strong contenders are not observed to possess significant advantage in their R.T. This may suggest their ability for winning races do not heavily rely on this metrics.



Swimmers with previous round top 3 finishers

Swimmers with previous round top 3 finishers

Figure 22: Past Top Performers' R.T vs The Rest

5. RECOMMENDATIONS

From the preceding analysis, based on the perspective of data, the following improvements are recommended:

1. As Tokyo 2020 did not include sprint events other than freestyle into, more sprint data on these events can be gathered from other relevant competitions to explore how R.T correlate with the split 50 timing performance across all strokes for the purpose of future Olympics preparation.
2. The age of the athletes may play an important role to the ability of achieving good R.T and overall performance. Hence, age information can be included for future studies.
3. R.T of individuals on each relay team can affect the overall team performance. Hence, data importing can be improved to cover the missing relays data.
4. As R.T is observed to have positive correlation to sprinter's timing, more in-depth studies on biomechanical factors that could affect the R.T can be explored (posture, body ratios…). This may have implications for tailored training.
5. R.T is one of several constituents of the start of a race, the others being the flight and the slip phase (Thanopoulos, 2012). Hence more data on these phases can be analyzed to discover further how the start performance can affect final timing.

Bibliography

Morgulev, E. A. (2018). Sports Analytics and the Big-Data Era. *International Journal of Data Science and Analytics volume* , 213-222.

Thanopoulos, V. R. (2012). Differences in the efficiency between the grab and track starts for both genders in greek young swimmers . *Journal of human kinetics*, 43-51.