

# Data Exploration Part 1

## **Dataset: 2000s Olympics Gold Medalists**

Mathematics for Data Analytics

Work By:

Rimi Mondal



## Introduction

The Olympic Games are considered one of the largest sporting events globally, influenced by the ancient Greek Olympic Games. They serve as a platform for countries to display their athletic abilities through intense competition. This dataset focuses on the gold medal-winning athletes from various countries in the 2000s Olympic Games, which are held in both summer and winter seasons. The dataset has 2511 rows and 15 columns.

## Data Dictionary

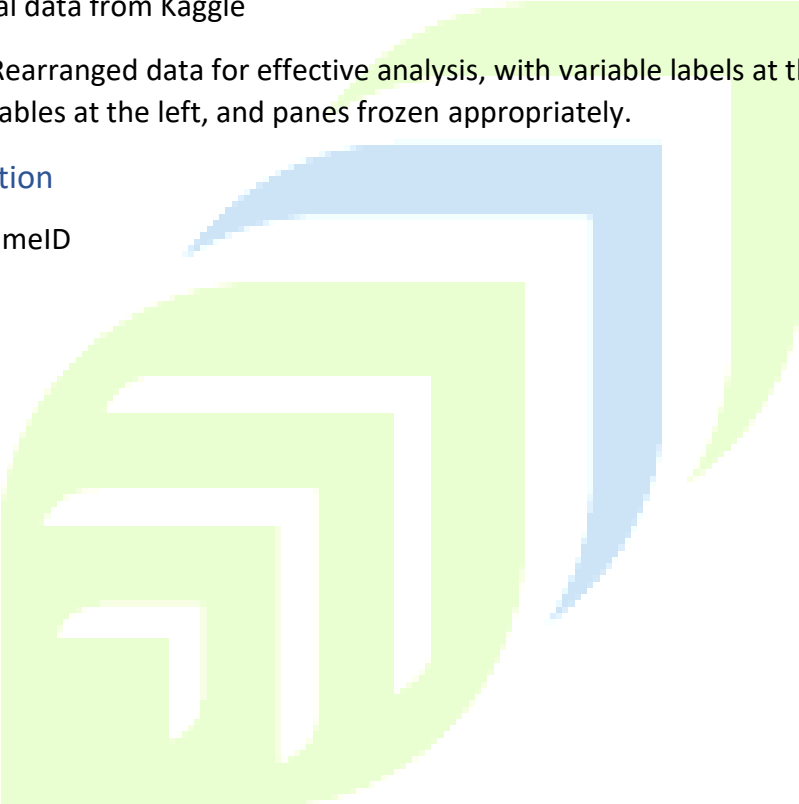
### File descriptions

raw\_data: Original data from Kaggle

olympics\_data: Rearranged data for effective analysis, with variable labels at the top, index and independent variables at the left, and panes frozen appropriately.

### Column Description

1. UniqueGameID
2. Name
3. Sex
4. Age
5. Height
6. Weight
7. Team
8. NOC
9. Games
10. Year
11. Season
12. City
13. Sport
14. Event
15. Medal



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	UniqueGameID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal
2	20010001	Wang Xin (Ruoxue-)	F	15	137	28	China	CHN	2008 Summer	2008	Summer	Beijing	Diving	Diving Women's Synchronized Platform	Gold
3	20010002	Loredana Boboc	F	16	139	32	Romania	ROU	2000 Summer	2000	Summer	Sydney	Gymnastics	Gymnastics Women's Team All-Around	Gold
4	20010003	Jiang Yuyuan	F	16	140	32	China	CHN	2008 Summer	2008	Summer	Beijing	Gymnastics	Gymnastics Women's Team All-Around	Gold
5	20010004	He Kexin	F	16	142	33	China	CHN	2008 Summer	2008	Summer	Beijing	Gymnastics	Gymnastics Women's Team All-Around	Gold
6	20010005	He Kexin	F	16	142	33	China	CHN	2008 Summer	2008	Summer	Beijing	Gymnastics	Gymnastics Women's Uneven Bars	Gold

*Image: Snippet of Dataset 'olympics\_data' showing all the variables*

## Variables Description

1. UniqueGameID: The Unique Game ID column is used to distinctly identify athletes, because some athletes share an identical name.
2. Name: The name of an athlete participated in Olympic games
3. Sex: The gender identification of an athlete.
4. Age: Age of the participating athlete
5. Height: Height of the athlete
6. Weight: Weight of the athlete
7. Team: The name of an athlete's team
8. NOC: National Olympic Committee, name of the country an athlete is representing.
9. Games: The year and season in which an athlete has participated
10. Year: The year of the Olympic game
11. Season: The season name of the Olympics
12. City: The city in which the Olympic game was hosted
13. Sport: The sport in which an athlete has participated
14. Event: The event in which an athlete has participated
15. Medal: The type of the medal; every entry is "Gold" in the dataset.

## Data Description

### Sex:

F Female  
M Male

### Age:

Range is from 15 years old to 50 years old.

### Height:

Range is from 137cm to 211 cm.

### Weight:

Range is from 28 kg to 155kg.

### Team:

Algeria	Ethiopia	Norway
Argentina	Finland	Panama
Australia	France	Poland
Australia-1	France-1	Portugal
Austria	Georgia	Romania
Austria-1	Germany	Russia
Azerbaijan	Germany-1	Russia-1
Bahamas	Germany-2	Satchmo
Belarus	Great Britain	Serbia and Montenegro

Belgium	Greece	Slovakia
Bonaparte	Hungary	Slovenia
Brazil	India	South Africa
Brazil-1	Indonesia	South Korea
Bulgaria	Indonesia-1	South Korea-1
Cameroon	Iran	South Korea-2
Canada	Israel	Spain
Canada-1	Italy	Sweden
Chile	Jamaica	Switzerland
China	Japan	Thailand
China-1	Kazakhstan	Trinidad and Tobago
China-2	Kenya	Tunisia
Chinese Taipei	Latvia	Turkey
Colombia	Lithuania	Ukraine
Croatia	Mexico	United Arab Emirates
Cuba	Mongolia	United States
Czech Republic	Morocco	United States-1
Denmark	Mozambique	United States-2
Dominican Republic	Netherlands	Uzbekistan
Egypt	New Zealand	Zimbabwe
Elvis Va	Nigeria	
Estonia	North Korea	

#### NOC:

ALG	Algeria	FRA	France	NZL	New Zealand
ARG	Argentina	GBR	Georgia	PAN	Panama
AUS	Australia	GEO	Germany	POL	Poland
AUT	Austria	GER	Great Britain	POR	Portugal
AZE	Azerbaijan	GRE	Greece	PRK	North Korea
BAH	Bahamas	HUN	Hungary	ROU	Romania
BEL	Belgium	INA	India	RSA	South Africa
BLR	Bonaparte	IND	Indonesia	RUS	Russia
BRA	Brazil	IRI	Iran	SCG	Satchmo
BUL	Bulgaria	ISR	Israel	SLO	Slovenia
CAN	Canada	ITA	Italy	SUI	Switzerland
CHI	Chile	JAM	Jamaica	SVK	Slovakia
CHN	China	JPN	Japan	SWE	Sweden
CMR	Cameroon	KAZ	Kazakhstan	THA	Thailand
COL	Colombia	KEN	Kenya	TPE	Chinese Taipei

CRO	Croatia	KOR	South Korea	TTO	Trinidad and Tobago
CUB	Cuba	LAT	Latvia	TUN	Tunisia
CZE	Czech Republic	LTU	Lithuania	TUR	Turkey
DEN	Denmark	MAR	Morocco	UAE	United Arab Emirates
DOM	Dominican Republic	MEX	Mexico	UKR	Ukraine
EGY	Egypt	MGL	Mongolia	USA	United States
ESP	Elvis Va	MOZ	Morocco	UZB	Uzbekistan
EST	Estonia	NED	Netherlands	ZIM	Zimbabwe
ETH	Ethiopia	NGR	Nigeria		
FIN	Finland	NOR	Norway		

#### Games:

2000 Summer  
2002 Winter  
2004 Summer  
2006 Winter  
2008 Summer  
2010 Winter

#### Year:

2000  
2002  
2004  
2006  
2008  
2010

#### Season:

Winter  
Summer

#### City:

Athens  
Beijing  
Salt Lake City  
Sydney  
Torino  
Vancouver



### Sport:

Alpine Skiing	Figure Skating	Ski Jumping
Archery	Football	Snowboarding
Athletics	Freestyle Skiing	Softball
Badminton	Gymnastics	Speed Skating
Baseball	Handball	Swimming
Basketball	Hockey	Synchronized Swimming
Beach Volleyball	Ice Hockey	Table Tennis
Biathlon	Judo	Taekwondo
Bobsleigh	Luge	Tennis
Boxing	Modern Pentathlon	Trampoline
Canoeing	Nordic Combined	Triathlon
Cross Country Skiing	Rhythmic Gymnastics	Volleyball
Curling	Rowing	Water Polo
Cycling	Sailing	Weightlifting
Diving	Shooting	Wrestling
Equestrianism	Short Track Speed Skating	
Fencing	Skeleton	

**No variables or data are missing. All the variables are available within the dataset.**

### Assumptions

1. Data is taken from Kaggle and it is assumed that it was acquired for data science education.
2. It is a complete data from the year 2000 to 2010 for all the Olympics games.
3. The units of measurements for all the variables are known.

### Research Questions

1. Is there a relationship between an athlete's weight and their likelihood of winning a gold medal in different sports categories?
2. Does an athlete's age impact their chances of winning a gold medal overall?
3. Does the location have any influence on the probability of winning a gold medal and if so, how?
4. Are there countries where women win more gold medals than men? If so, does gender affect an athlete's chances of winning a gold medal?
5. Does the time of year (summer or winter) have any effect on the chances of winning a gold medal, and if yes, in what way?

### Consideration

1. The data for each variable is adequate to address the research questions.

2. However, different types of data may be necessary to fully address all the research questions.

### Track of Analysis

1. The data has been obtained from Kaggle on 01/26/2023 by me.
2. The data has been rearranged from the original data for the ease of analysis.
3. A data dictionary has been created in this report for ease of access and better understanding of the dataset.



## Data Exploration Part 2

### **Dataset: 2000s Olympics Gold Medalists**

Perform univariate data analysis, including examination, visualization, and preliminary cleaning and coding



## Descriptive Statistics

The following section presents the descriptive statistics for the four variables mentioned below. Part 1 provides a detailed explanation of the variables.

### 1. Age

### 2. Weight

### 3. Height

### 4. Year

<b>Age</b>		<b>Height</b>		<b>Weight</b>		<b>Year</b>	
Mean	26.20478	Mean	177.7916	Mean	74.1394	Mean	2004.44
Standard Error	0.103258	Standard Error	0.224551	Standard Error	0.31489	Standard Error	0.06724
Median	26	Median	178	Median	73	Median	2004
Mode	24	Mode	180	Mode	75	Mode	2008
Standard Deviation	5.173234	Standard Deviation	11.24998	Standard Deviation	15.776	Standard Deviation	3.36881
Sample Variance	26.76235	Sample Variance	126.562	Sample Variance	248.883	Sample Variance	11.3489
Kurtosis	1.065476	Kurtosis	-0.11014	Kurtosis	0.71007	Kurtosis	-1.33312
Skewness	0.732692	Skewness	-0.01004	Skewness	0.55855	Skewness	0.0044
Range	35	Range	74	Range	127	Range	10
Minimum	15	Minimum	137	Minimum	28	Minimum	2000
Maximum	50	Maximum	211	Maximum	155	Maximum	2010
Sum	65774	Sum	446257	Sum	186090	Sum	5031144
Count	2510	Count	2510	Count	2510	Count	2510

## Univariate Analysis

### 1. Mean age of athletes:

Variable- Age

Row Labels	Average of Age
F	25.66756272
M	26.6348637
<b>Grand Total</b>	<b>26.20478088</b>

### 2. Median of Height:

Variable-Height

=MEDIAN(Sheet1!D:D)		
C	D	E
		Median Height
		178

### 3.MODE of weight:

Variable-weight

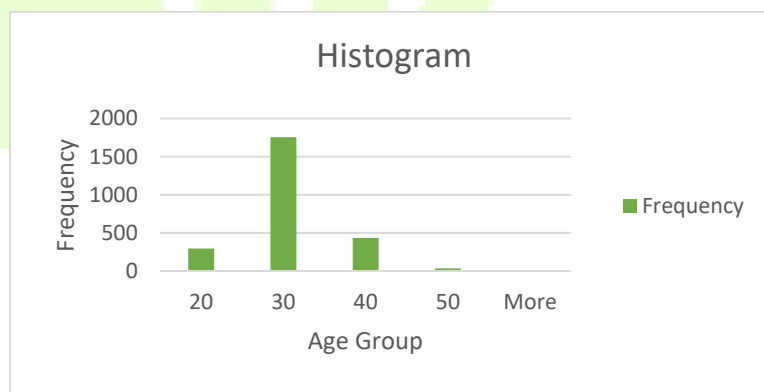
=MODE.SNGL(Sheet1!E:E)		
	C	D
	Mode of weight	
	75	

### 4.Frequency Distribution (Histogram)

The histogram below demonstrates that age group is a critical factor in winning a gold medal, with athletes over 50 having a lower probability of success.

This finding addresses the following research question:

Research Q2: Does an athlete's age impact their chances of winning a gold medal overall?



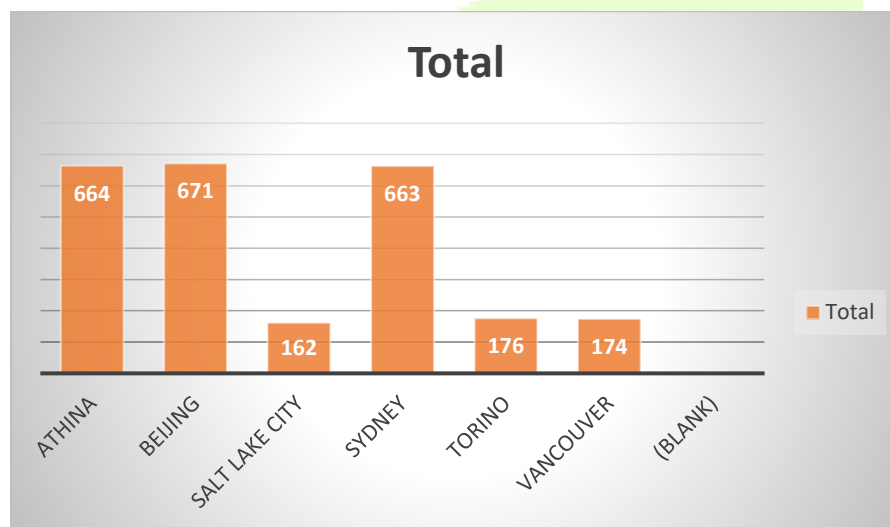
## 5.Count of sports by city Chart

The chart presented below provides an answer to a research question posed in Part 1. The chart reveals that venues with lower medal counts are more challenging, indicating that they are the toughest venues.

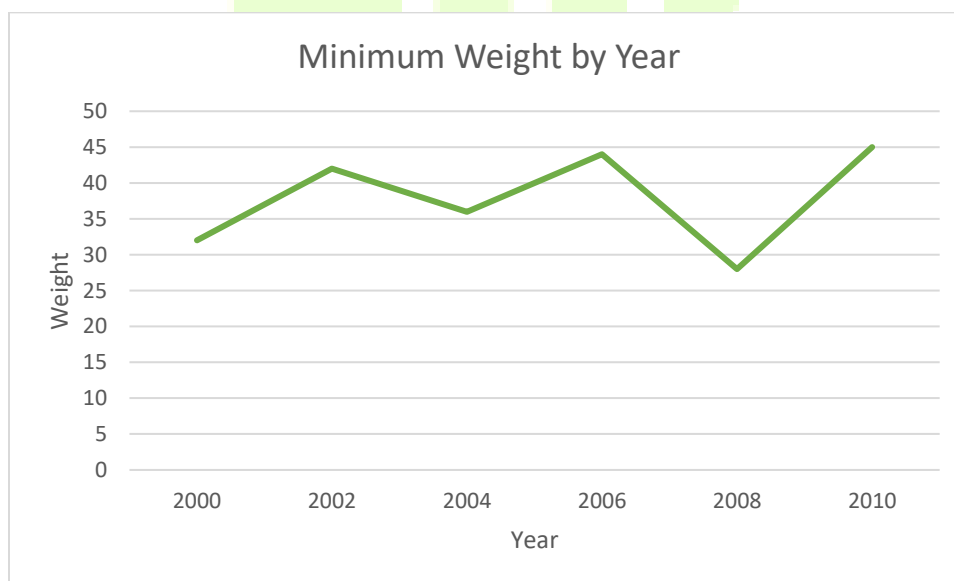
Two out of three cities with fewer medal counts are located in North America, leading to the conclusion that North America is the most difficult continent to win Olympic gold medals in.

This finding addresses the following research question:

Research Q3: Does the location have any influence on the probability of winning a gold medal and if so, how?



## 6.Line graph of min weight by year

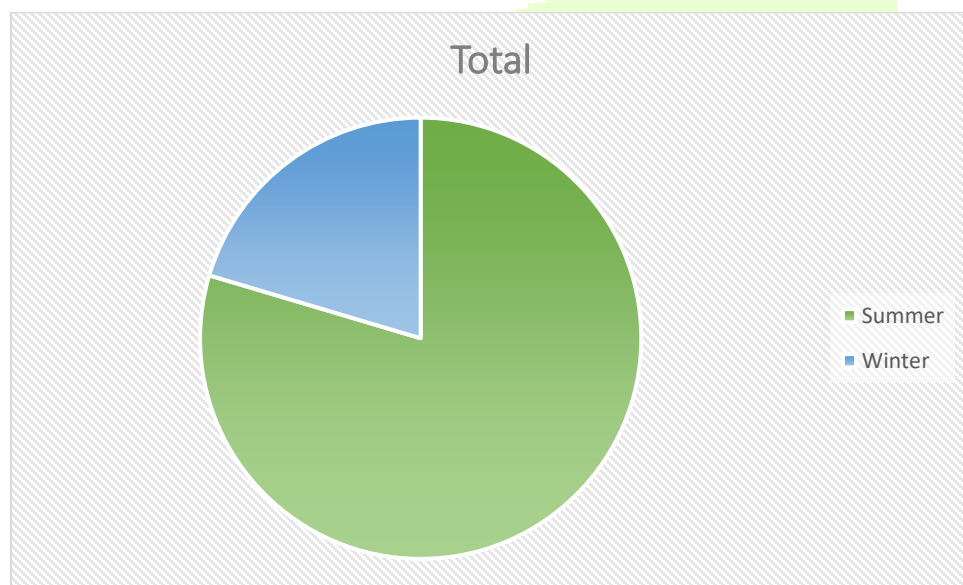


## 7. Pie Chart by season

Based on the following pie chart, it is easy to determine which season has yielded the most wins for athletes in various events. The chart shows that the summer season has a significantly larger share than the winter season in terms of wins across all events. From this, we can conclude that weather conditions play a crucial role in determining success in athletic events.

This finding addresses the following research question:

Research Q5: Does the time of year (summer or winter) have any effect on the chances of winning a gold medal, and if yes, in what way?



## Coding

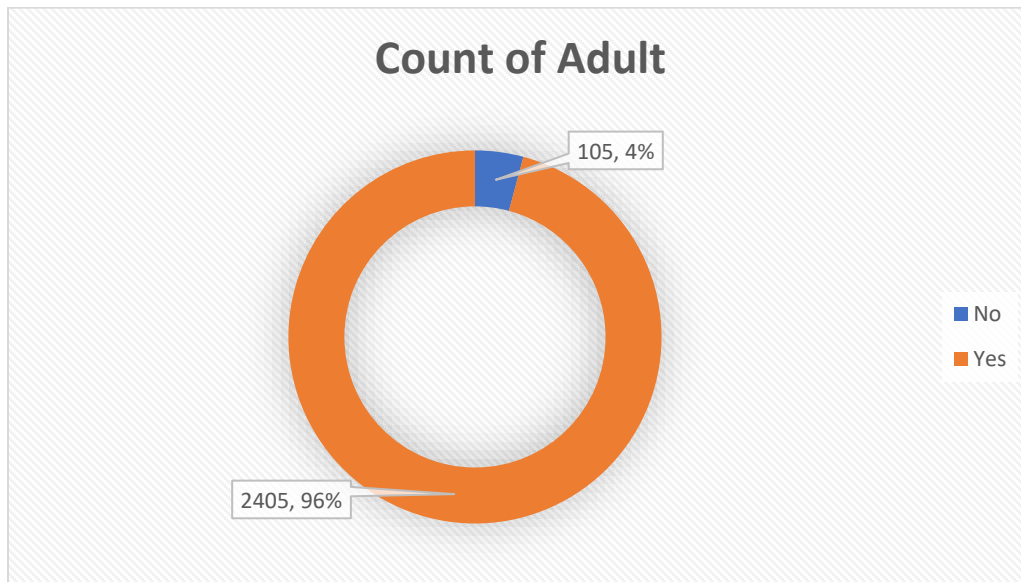
The original dataset did not contain any categorical variables.

However, I have introduced two categorical variables for the better analysis to answer the research questions, namely:

Adult (age > 19) (Yes or No)

Summer (Yes or No)

## 8. Doughnut chart by count of adults



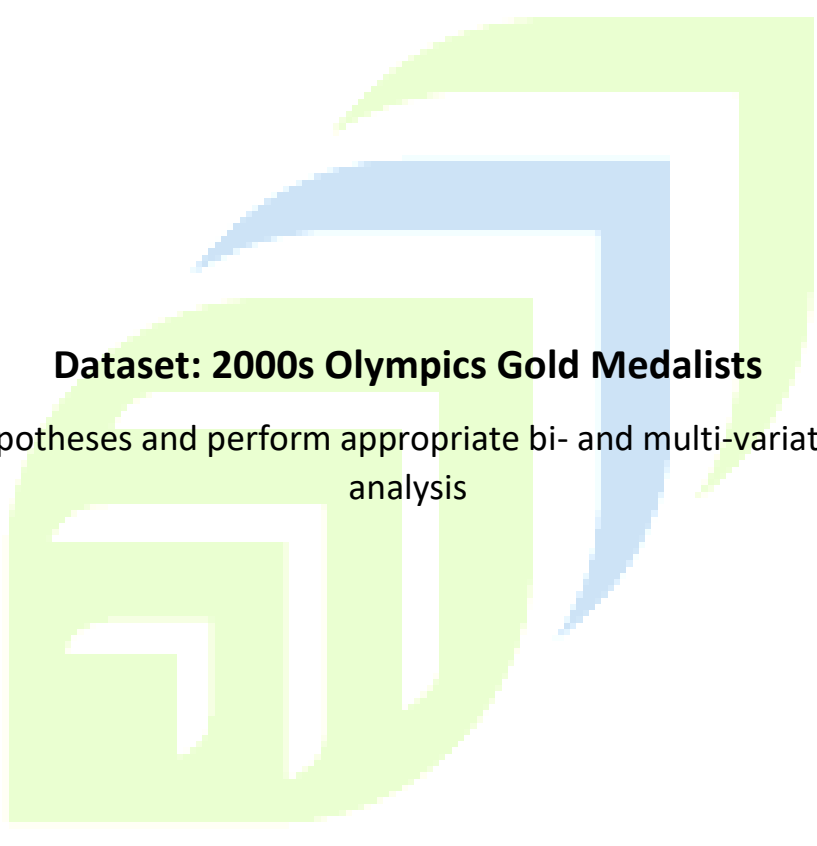
### Assumptions

The research questions and assumptions remain unchanged, and there is no requirement for additional data to conduct a comprehensive analysis for the said research questions.

### Track of Analysis

1. The dataset now includes two newly added categorical variables by coding them - one to track the count of adults and the other to determine if the event took place during the summer season or not.
2. Descriptive statistics has been performed on four numerical variables and univariate analysis has been performed on at eight variables.

## Data Exploration Part 3



### **Dataset: 2000s Olympics Gold Medalists**

Develop hypotheses and perform appropriate bi- and multi-variate statistical analysis

## Hypothesis

In this section dedicated to hypothesis testing, five tests have been conducted, which are listed below for reference:

1. Odds Ratio
2. Risk Ratio
3. Chi-Square Test
4. T-test
  - a. Two-sample assuming equal variance
  - b. Two-sample assuming unequal variance
5. ANOVA Test

### 1. Odds Ratio

### 2. Risk Ratio

In order to compute the odds ratio and risk ratio, it is necessary to obtain the actual values. Therefore, I have utilized the formulas taught in the class to derive the required values.

Actual Values			
Count of UniqueGameID Column Labels			
Row Labels	No	Yes	Grand Total
No	16	496	512
Yes	89	1909	1998
Grand Total	105	2405	2510

With these values now available, I can proceed with the calculation of odds ratio and risk.

Actual Values			
Count of UniqueGameID Column Labels			
Row Labels	No	Yes	Grand Total
No	16	496	512
Yes	89	1909	1998
Grand Total	105	2405	2510
Odds Ratio	$(a*d)/(b*c)$		0.69191736 1.445259 Inverted OR
	Adult	Not Adult	
summer	a	b	
Not summer	c	d	
Risk Ratio	$(a/(a+b))/(c/(c+d))$		0.70154494

### 3. Chi Square Test

To perform the chi-square test, it is necessary to compute the expected values beforehand. The expected values have been calculated using the formulae that were taught in class, and can be referred to in the attached excel sheet.

Expected values			
Row Labels	Yes	No	Grand Total
Yes	21.41832669	491	512
No	83.58167331	1914	1998
Grand Total	105	2405	2510
Chi-square			
Row Labels	Yes	No	Grand Total
Yes	1.370707646	0.06	
No	0.35125241	0.02	
Grand Total			1.79713918
p-value for Chi-square	0.180058739		

#### Result

The results indicate that we cannot reject the null hypothesis, as the P value is greater than 0.05. Therefore, the result is not significant.

### 4. T-Test

#### a. Equal variance:

When performing the t-test assuming equal variances, the p-value is observed to be less than 0.05. Therefore, we reject the null hypothesis and accept the alternate hypothesis.

#### Result

This indicates a statistically significant outcome.

t-Test: Two-Sample Assuming Equal Variances		
	No	Yes
Mean	27.02539	25.99449
Variance	25.19113	26.96091
Observations	512	1998
Pooled Variance	26.60032	
Hypothesized Mean Difference	0	
df	2508	
t Stat	4.035221	
P(T<=t) one-tail	2.81E-05	
t Critical one-tail	1.645461	
P(T<=t) two-tail	5.62E-05	
t Critical two-tail	1.96091	



### b. Unequal Variance:

In the unequal variance t-test, the obtained p-value is found to be less than 0.05, indicating that the null hypothesis can be rejected.

### Result

The result is considered to be statistically significant.

t-Test: Two-Sample Assuming Unequal Variances		
		15
Mean	27.02539	26
Variance	25.19113	26.91383
Observations	512	1997
Hypothesized Mean Difference	0	
df	814	
t Stat	4.095716	
P(T<=t) one-tail	2.31E-05	
t Critical one-tail	1.646728	
P(T<=t) two-tail	4.63E-05	*
t Critical two-tail	1.962883	

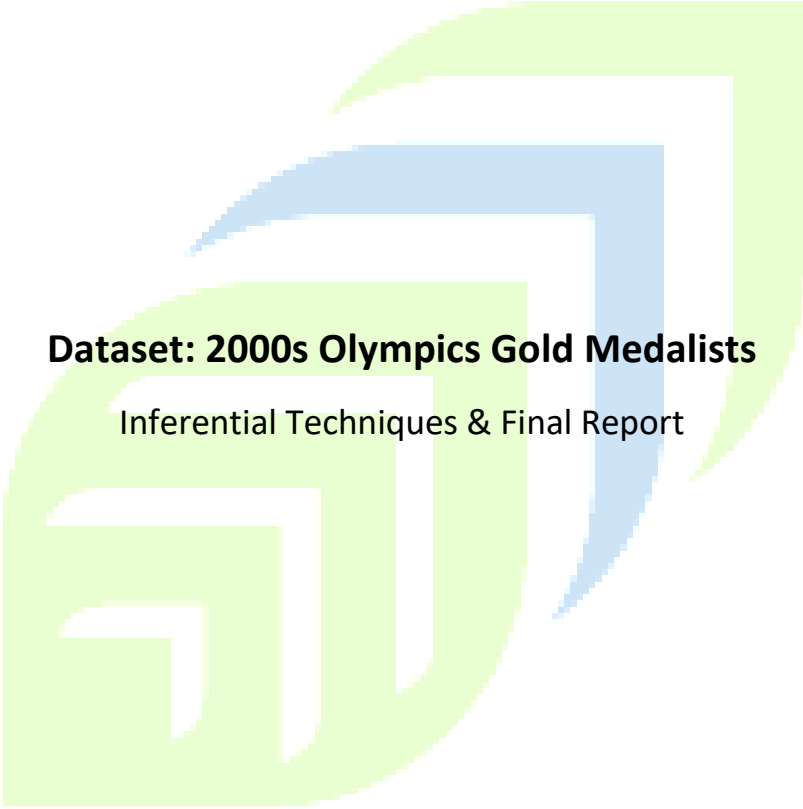
## 5. ANOVA Test

Based on the ANOVA test, it can be observed that the F statistic exceeds the F-critical value. Therefore, the null hypothesis can be rejected.

Anova: Single Factor						
SUMMARY						
Groups	Count	Sum	Average	Variance		
No	512	13837	27.02539	25.19113		
Yes	1998	51937	25.99449	26.96091		
ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	433.1333	1	433.1333	16.28301	5.62E-05	3.845169
Within Groups	66713.61	2508	26.60032			
Total	67146.74	2509				

### Track of Analysis

1. The research questions have been enhanced to provide more depth and insight.
2. Hypotheses have been developed to aid in answering one of the research questions.
3. Several statistical tests have been performed, including Odds Ratio, Risk Ratio, Chi-Square Test, Two-Sample T-test with equal variance assumption, Two-Sample T-test with unequal variance assumption, and ANOVA Test.



### **Dataset: 2000s Olympics Gold Medalists**

Inferential Techniques & Final Report

## Introduction

The Olympic Games are considered one of the largest sporting events globally, influenced by the ancient Greek Olympic Games. They serve as a platform for countries to display their athletic abilities through intense competition. This dataset focuses on the gold medal-winning athletes from various countries in the 2000s Olympic Games, which are held in both summer and winter seasons. The dataset has 2511 rows and 15 columns.

## Data Dictionary

### File descriptions

raw\_data: Original data from Kaggle

olympics\_data: Rearranged data for effective analysis, with variable labels at the top, index and independent variables at the left, and panes frozen appropriately.

### Column Description

1. UniqueGameID
2. Name
3. Sex
4. Age
5. Height
6. Weight
7. Team
8. NOC
9. Games
10. Year
11. Season
12. City
13. Sport
14. Event
15. Medal

### Variables Description

1. UniqueGameID: The Unique Game ID column is used to distinctly identify athletes, because some athletes share an identical name.
2. Name: The name of an athlete participated in Olympic games
3. Sex: The gender identification of an athlete.
4. Age: Age of the participating athlete
5. Height: Height of the athlete
6. Weight: Weight of the athlete
7. Team: The name of an athlete's team
8. NOC: National Olympic Committee, name of the country an athlete is representing.

9. Games: The year and season in which an athlete has participated
10. Year: The year of the Olympic game
11. Season: The season name of the Olympics
12. City: The city in which the Olympic game was hosted
13. Sport: The sport in which an athlete has participated
14. Event: The event in which an athlete has participated
15. Medal: The type of the medal; every entry is “Gold” in the dataset.

## Data Analysis

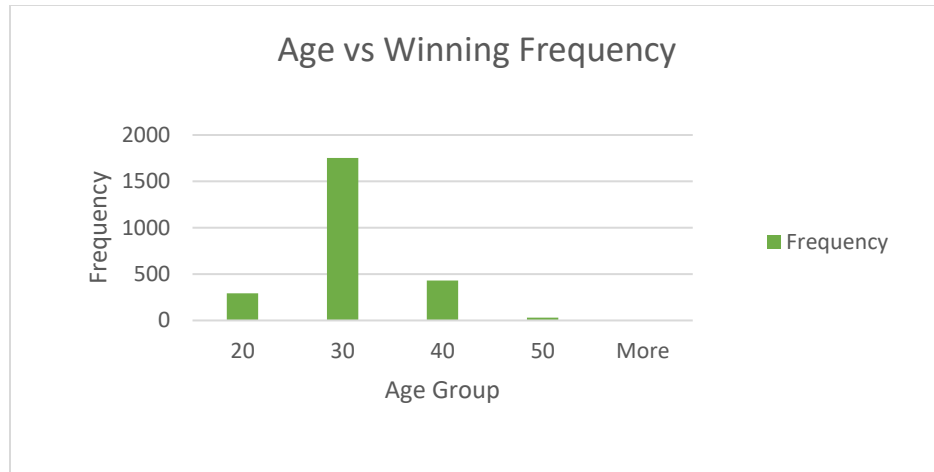
1. The data has been rearranged from the original data for the ease of analysis.
2. A data dictionary has been created in this report for ease of access and better understanding of the dataset.
3. The dataset now includes two newly added categorical variables by coding them - one to track the count of adults and the other to determine if the event took place during the summer season or not.
4. Descriptive statistics has been performed on four numerical variables and univariate analysis has been performed on at eight variables.
5. The research questions have been enhanced to provide more depth and insight.
6. Hypotheses have been developed to aid in answering one of the research questions.
7. Several statistical tests have been performed, including Odds Ratio, Risk Ratio, Chi-Square Test, Two-Sample T-test with equal variance assumption, Two-Sample T-test with unequal variance assumption, and ANOVA Test.

## Study Results

1. Research Q2: Does an athlete's age impact their chances of winning a gold medal overall?

Based on the histogram, it is evident that the probability of winning sharply decreases after the age of 30 and remains relatively constant after the age of 50.

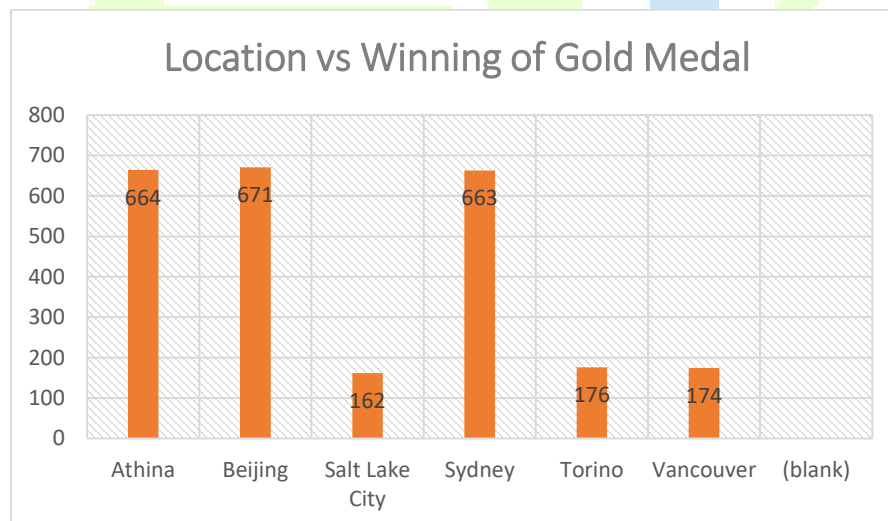
This observation suggests that age is a significant determinant of winning a medal, which is supported by the data presented in the histogram.



## 2. Research Q3: Does the location have any influence on the probability of winning a gold medal and if so, how?

Upon analyzing the provided pivot table and bar graph, it can be observed that venues such as Salt Lake City and Vancouver have lower counts. Notably, these two locations are located in North America.

The graph below serves as evidence to support the research question.



## 3. Research Q5: Does the time of year (summer or winter) have any effect on the chances of winning a gold medal, and if yes, in what way?

Upon examination of the accompanying pie chart, it becomes evident that determining which season sees more wins in athletic events can be easily answered. The chart illustrates the share

of wins for each season in relation to all events. A careful analysis of the chart reveals that the share of wins for summer events is significantly larger than that of winter events.

Based on this finding, one could reasonably infer that weather conditions play a crucial role in determining the likelihood of winning an event. For instance, it is possible that athletes who participate in summer events are better adapted to the warmer temperatures and longer daylight hours. Similarly, those who compete in winter events may have an advantage in colder temperatures and snowier conditions.



## Conclusion

In conclusion, the analysis of the dataset has provided significant insights into the factors that impact an athlete's chances of winning a gold medal. Age has been identified as a significant determinant of winning a medal, with the probability of winning sharply decreasing after the age of 30. Location has also been found to have an influence on the probability of winning, with North American venues exhibiting lower counts of medals won by athletes in Olympic events. Finally, the time of year has been shown to have an effect on the chances of winning a gold medal, with the share of wins for summer events significantly larger than that of winter events. These findings can be useful for athletes and coaches looking to optimize their training and performance in Olympic events.

## Track of Analysis

1. Three specific research questions have been selected carefully to provide valuable insights into the data at hand and address important aspects of the research topic.
2. The charts and graphs used in the analysis have been improved, in order to ensure that the results were accurate and easy to interpret.
3. To facilitate easy interpretation and understanding of the charts and graphs, clear and concise titles have been assigned to each one.