SQL for Data Science Capstone Project

Assignment 3: WEEK 3 / MILESTONE 3: Beyond Descriptive Stats

Name: Daniel Siva

Q1: Specify 1-2 correlations you discovered. List the fields that you found to be correlated and describe what you learned from these correlations.

Age Group by Medals (discrete variables)

I calculated a cross-tabulation using different age groups (group_10_15, group_15_20, group_20_25, group_25_30, group_30_35, group_35_40, group_40_50, group_50_60, group_60_70, group_70_80) and medals (gold, silver, bronze, NA).

I used the Pearson's chi-squared test to determine whether there is a statistically significant difference between the expected and observed frequencies. I reject the null hypothesis (p-value <0.01), so I concluded that age (specially the age group between 20 and 25) has an impact in the number of medals win by the athlete. I performed the same test separating male and female athletes and I got the same results.

Statistics for Table of Age_Group by Medal

Statistic	DF	Value	Prob
Chi-Square	33	729.5813	<.0001
Likelihood Ratio Chi-Square	33	773.4427	<.0001
Mantel-Haenszel Chi-Square	1	39.9437	<.0001
Phi Coefficient		0.0529	
Contingency Coefficient		0.0529	
Cramer's V		0.0306	

Conclusion: if an athlete is between 20 and 25 years old the probability of winning a medal is higher than in other age (relation statistically significant) and works for both genders (male and female)

Number of participants (continuous variable) x Number of medals (continuous variable)

I calculated the number of participants and the number of medals in each country and used the Pearson Correlation. The Pearson Correlation evaluates whether there is statistical evidence for a linear relationship among the same pairs of variables in the population. The r=0.929214 with a p-value <0.0001. This mean that the number of participants and the number of medals are correlated positively.

The CORR Procedure

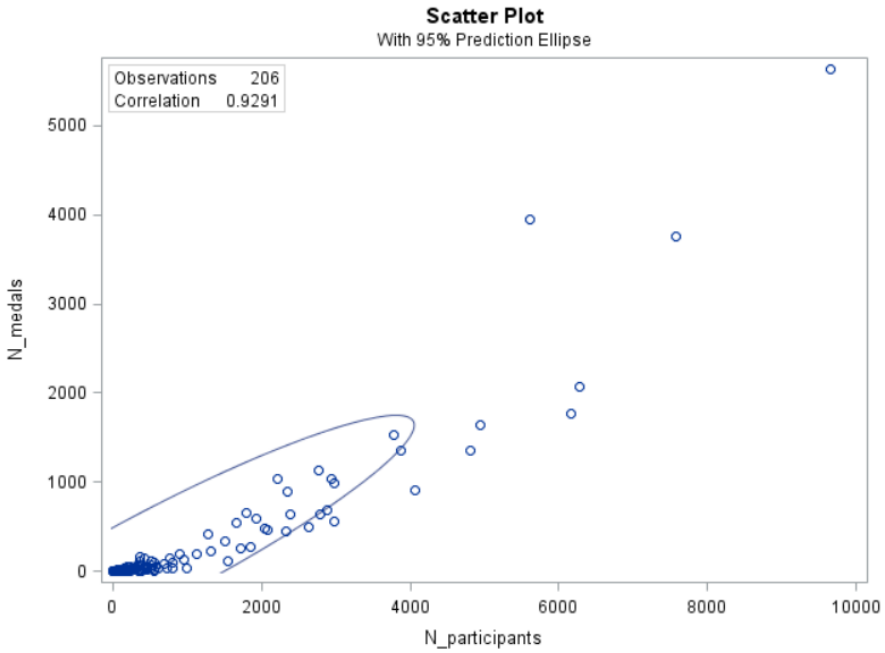
2 Variables: N_participants N_medals

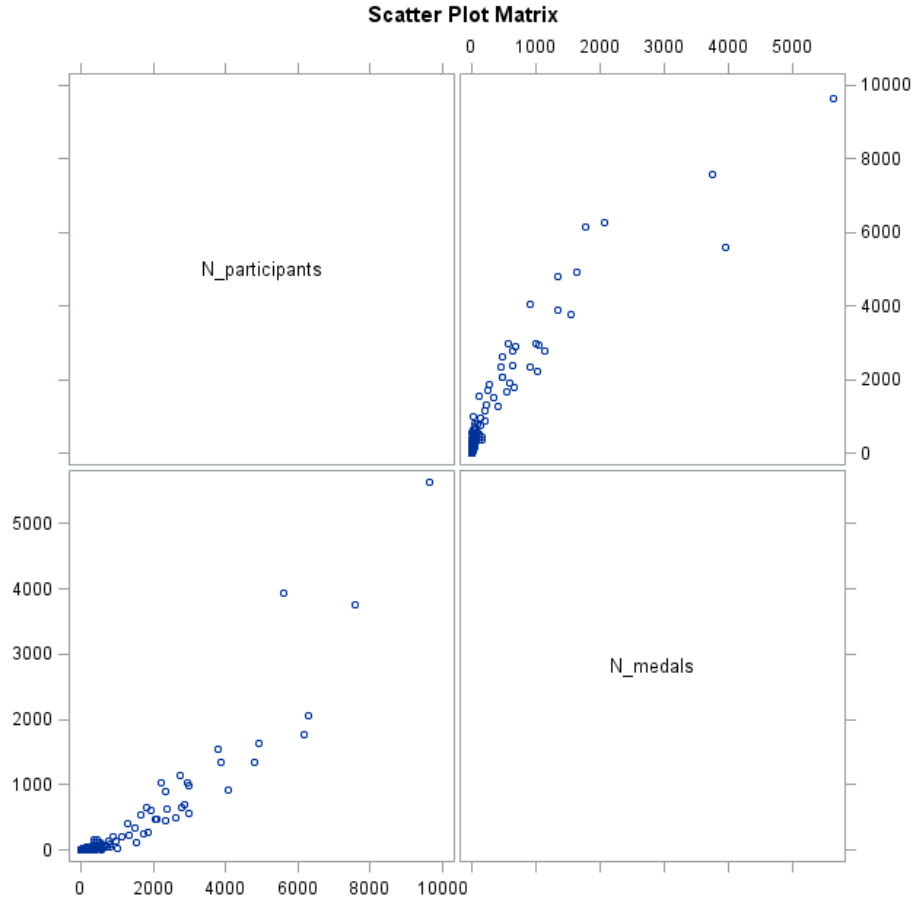
Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
N_participants	206	660.85922	1376	136137	3.00000	9653
N_medals	206	193.02427	625.72966	39763	0	5637

Pearson Correlation Coefficients, N = 206
Prob > |r| under H0: Rho=0

	N_participants	N_medals
N_participants	1.00000	0.92914
		<.0001
N_medals	0.92914	1.00000
	<.0001	





Conclusion: if a country increase the number of participants the impact in the number of medals won is positive and highly correlated.

Q2: Create 1 or 2 new metrics to track relationships of data you discovered. Explain why you created them.

In my analyses I use essential there metrics: number of participants and number of medals (total or by type: gold, silver and bronze) and age. The number of participants and the number of medals are essential to evaluate the success of the athletes and the countries in the Olympic Games.

The age for me is a core variable to analyze the demographics of Olympic Games. I discovered that the age group between 20 and 25 is the most represented. But if we do the breakdown by sport we discover that 'Rhythmic Gymnastics' has the lower mean (18 years, with a minimum age of 13). The 'Roque' and the 'Art Competitions' were the sports with the highest mean: 53, and 45 years old, respectively. Nowadays, both sports are not considered as an Olympic sport.