TOKYO OLYMIPCS 2021 MEDAL TALLY ANALYSIS

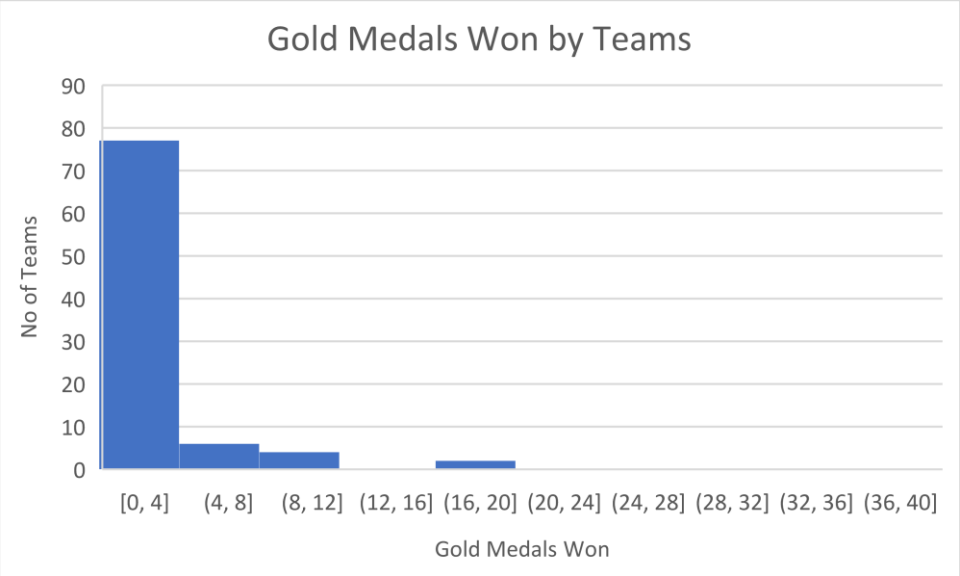
This dataset was sourced from Kaggle. The attributes in this dataset are:

1. Name of Country/NOC
2. Gold Medals Won
3. Silver Medal Won
4. Bronze Medals Won
5. Total Medals Won
6. Rank according to Total medal Won

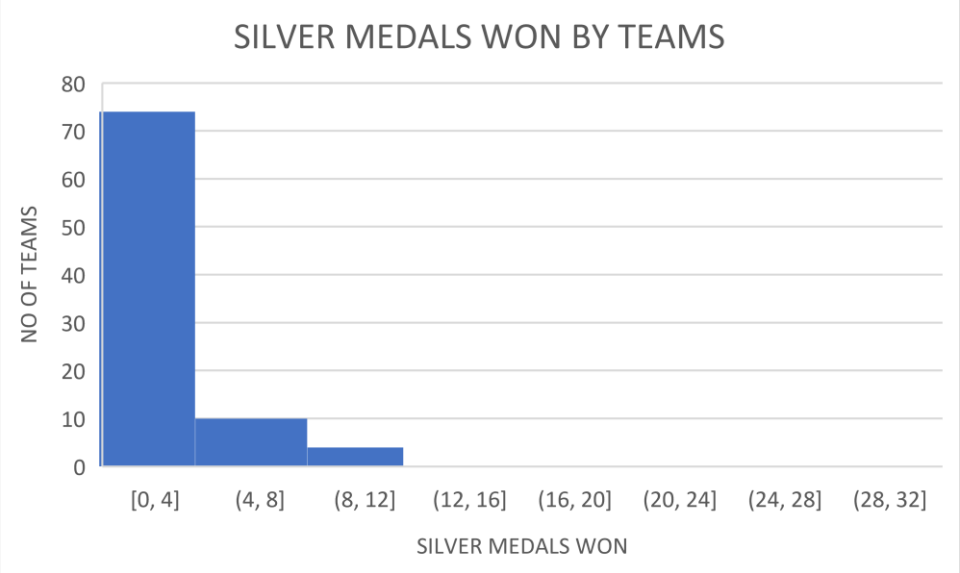
This dataset has a total of 94 rows and 6 columns before data cleaning and feature engineering process.

The histograms before data cleaning and feature engineering ratio are as follows -:

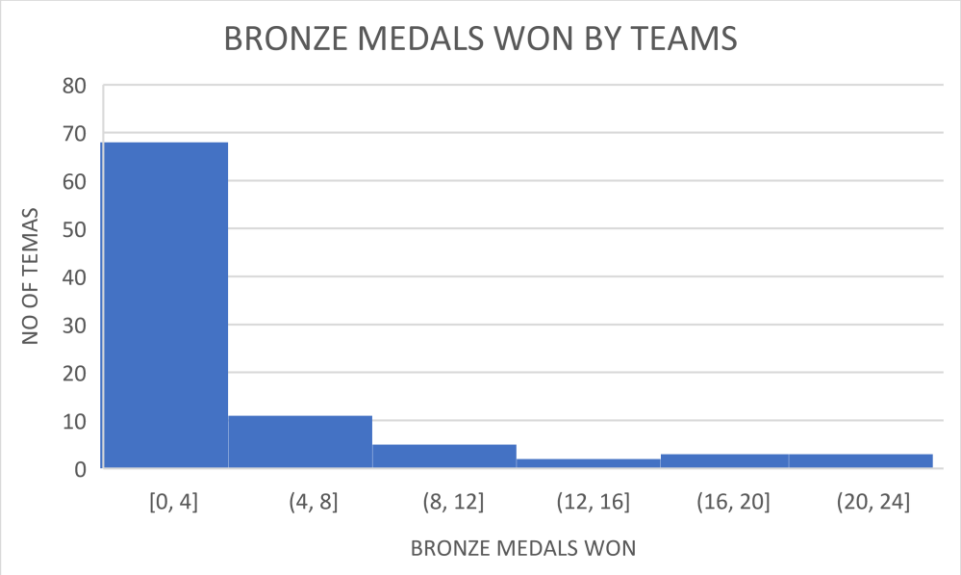
GOLD MEDALS HISTOGRAM



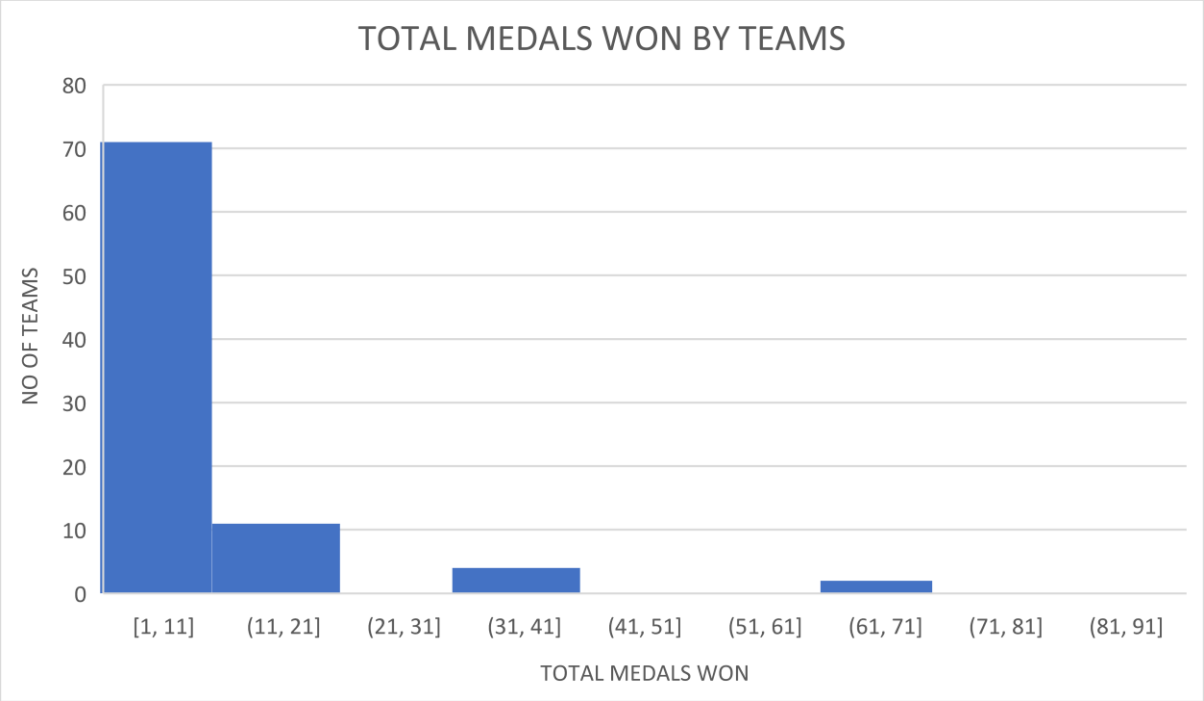
SILVER MEDALS HISTOGRAM



BRONZE MEDALS HISTOGRAM



TOTAL MEDALS HISTOGRAM

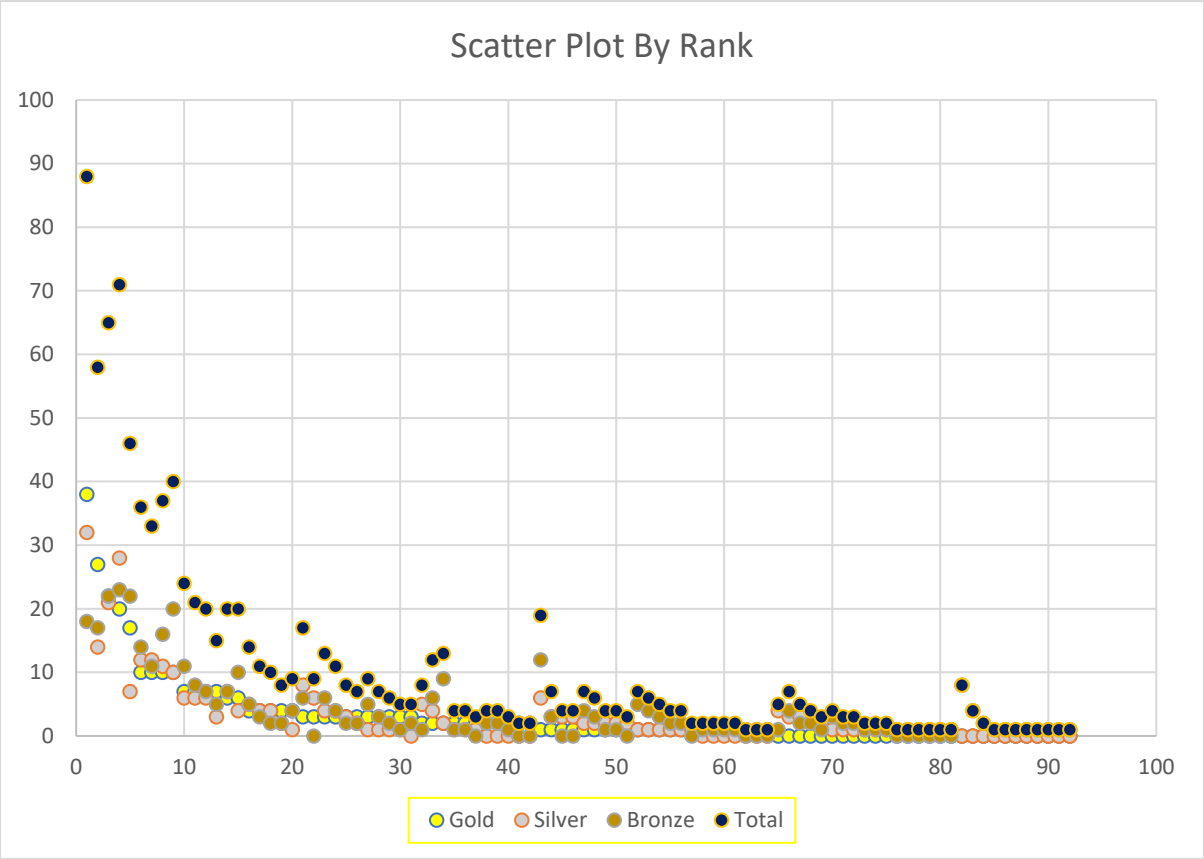


Seeing above histograms we can conclude that most of the teams have the medal count between 1 to 10.

In the gold medal histogram we see that 74 teams have less than 4 gold medals. This data shows that the Olympics were dominated by only a few teams.

We will have to filter out this data by removing the teams which performed exceptionally well because they cannot interpret what we need to find.

The scatter plot between medals won and rank is -:



We observe that all of the elements in the scatter plot are correlated to one another expect at a few points like -:

/. Ukraine has more total medals won but less gold medals won

/. Kazakhstan won 8 bronze medals but no gold or silver medals

After seeing this data we conclude that USA is a possible outlier because they performed exceptionally well. So, we will remove USA from our analysis.

Here are the stats before we remove USA from the data :-

	Gold	Silver	Bronze	Total
count	93.000000	93.000000	93.000000	93.000000
mean	3.655914	3.634409	4.322581	11.612903
std	7.022471	6.626339	6.210372	19.091332
min	0.000000	0.000000	0.000000	1.000000
25%	0.000000	0.000000	1.000000	2.000000

50%	1.000000	1.000000	2.000000	4.000000
75%	3.000000	4.000000	5.000000	11.000000
max	39.000000	41.000000	33.000000	113.000000

We have now removed USA from our analysis

The stats after removal are now as follows -:

	Gold	Silver	Bronze	Total
count	92.000000	92.000000	92.000000	92.000000
mean	3.271739	3.228261	4.010870	10.510870
std	5.998357	5.374141	5.464158	15.946679
min	0.000000	0.000000	0.000000	1.000000
25%	0.000000	0.000000	1.000000	2.000000
50%	1.000000	1.000000	2.000000	4.000000
75%	3.000000	4.000000	5.000000	10.250000
max	38.000000	32.000000	23.000000	88.000000

Now I have created an extra feature called the gold vs total ratio which says that how much of our medals won were gold

I formulated three hypothesis about the data

1. India did not won more gold medals than average
2. Kazakhstan won more total medals than average
3. Serbia won more total medals than average

I will conduct a test for hypothesis 1-:

The null hypothesis will be India did not won more gold medals than average

And the alternate hypothesis will be India won more gold medals than average

Calculating a 95% cut-off beforehand. For a team to perform better than average in gold medals they need to win 3 or more gold medals. But India have won only one gold medal . So they did not perform better than most of the teams.