Name-Surname: Nouredeen Ahmed Mahmoud Ali Hammad
Student Number: 2121221362

Probability and Statistics Project Report

About Dataset:

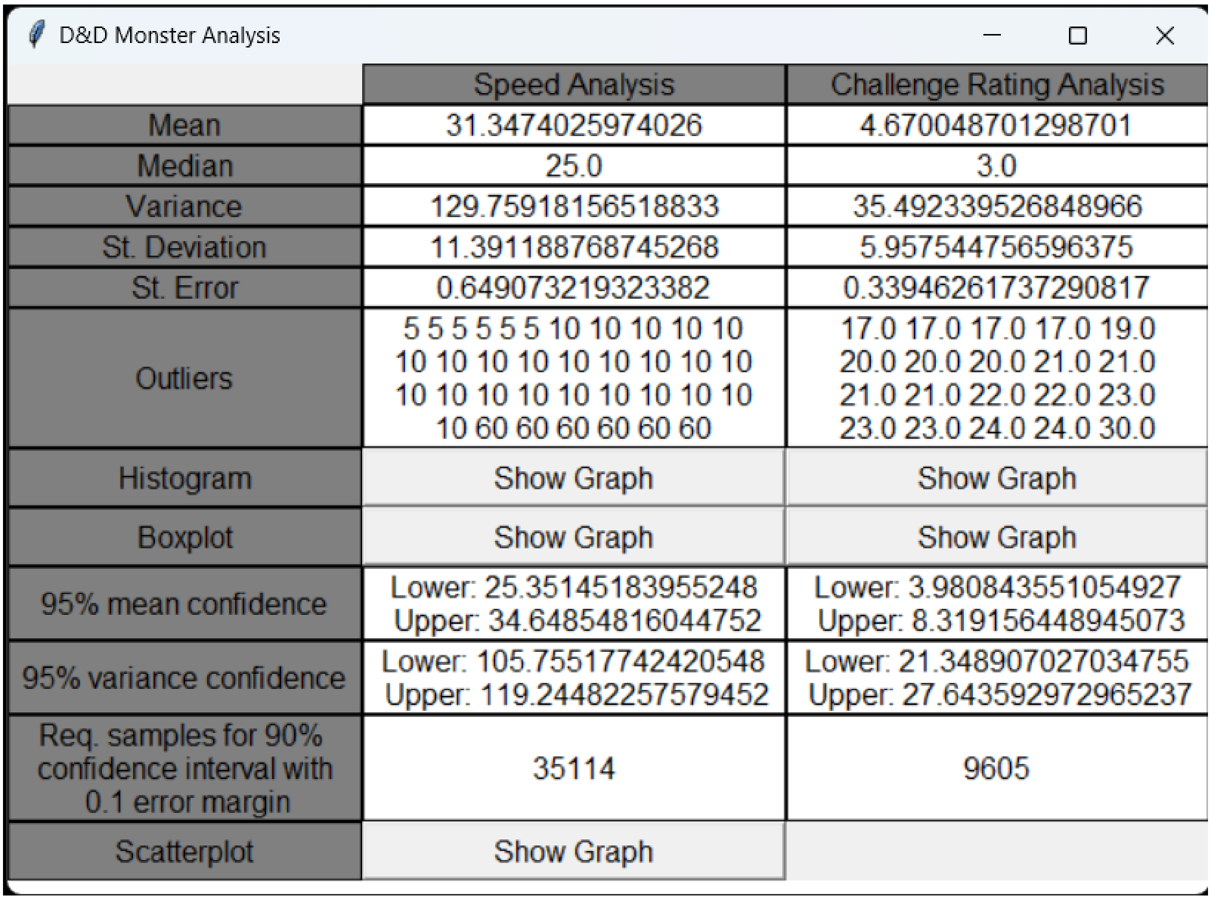
The dataset is about monsters used in a table-top role playing game called Dungeons and Dragons (D&D). It contains data about the monsters' characteristics. It contains 323 records.
In this project I will look at the "Speed" and "Challenge Rating (XP)" columns.
I chose them because I was curious about the relationship between speed and challenge rating (challenge rating is a score that indicates how difficult the monster is to kill).

Note: Challenge Rating is abbreviated as CR.

The target columns are not completely numerical. I filter out invalid data when reading the CSV file in the code.

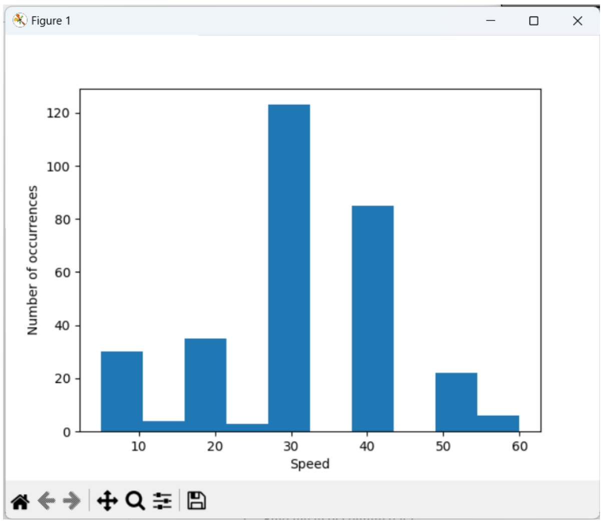
Screenshots and Analysis:

GUI

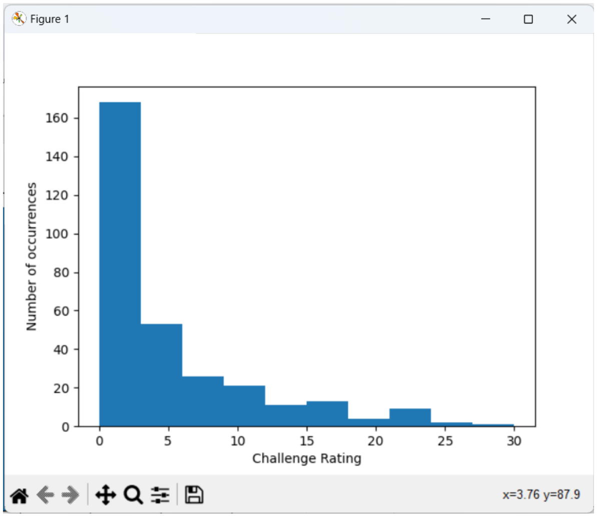


	Speed Analysis	Challenge Rating Analysis
Mean	31.3474025974026	4.670048701298701
Median	25.0	3.0
Variance	129.75918156518833	35.492339526848966
St. Deviation	11.391188768745268	5.957544756596375
St. Error	0.649073219323382	0.33946261737290817
Outliers	5 5 5 5 5 5 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10 60 60 60 60 60 60	17.0 17.0 17.0 17.0 19.0 20.0 20.0 20.0 21.0 21.0 21.0 21.0 22.0 22.0 23.0 23.0 23.0 24.0 24.0 30.0
Histogram	Show Graph	Show Graph
Boxplot	Show Graph	Show Graph
95% mean confidence	Lower: 25.35145183955248 Upper: 34.64854816044752	Lower: 3.980843551054927 Upper: 8.319156448945073
95% variance confidence	Lower: 105.75517742420548 Upper: 119.24482257579452	Lower: 21.348907027034755 Upper: 27.643592972965237
Req. samples for 90% confidence interval with 0.1 error margin	35114	9605
Scatterplot	Show Graph	

Speed



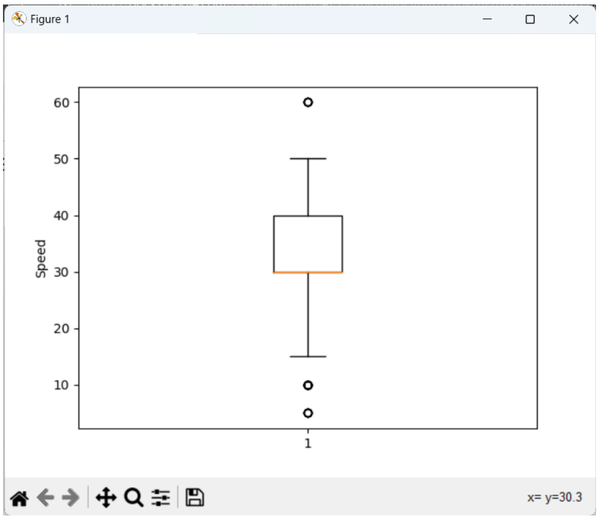
Challenge Rating



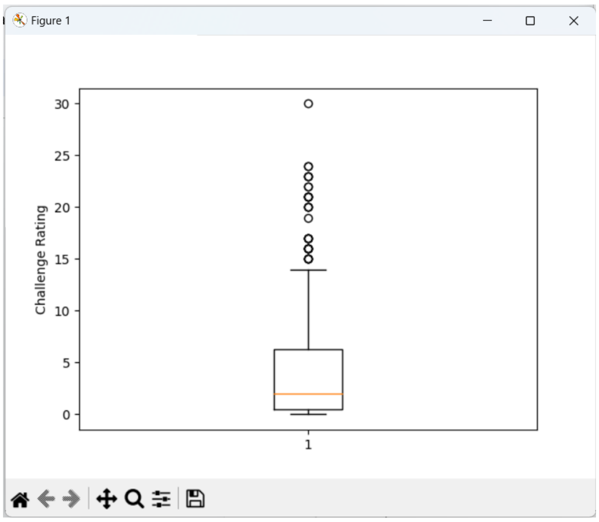
From the speed histogram graph we can notice that it is approximately a bell shaped distribution, technically it is more like a random distribution. It is not uniform because the values are not completely consistent, for example: there are a lot of monsters with speed 30 and speed 40, but no monsters with speed 35.
We can also notice that the most common monster speed is 30, which is very close to the mean value of the data which equals 31.347 .

From the Challenge Rating histogram we can notice that it is very sharply right skewed.
We can also notice that the majority of monsters have a challenge rating between 0 and 4. The mean value of the data is about 4.67 so it makes sense.

Speed



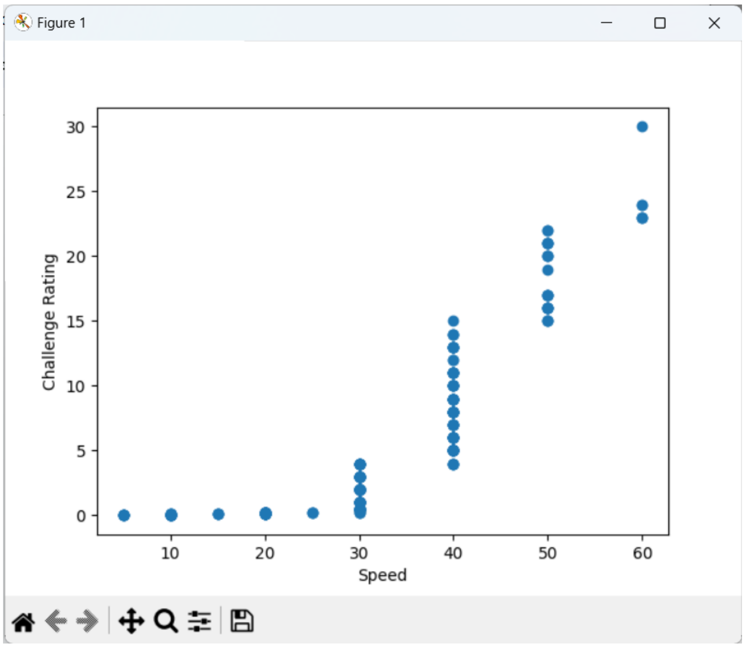
Challenge Rating



From the speed box plot we can see that the median is around 30, the lower quartile (Q1) is around 30, the upper quartile (Q3) is around 40. The min and max values (excluding outliers) are around 15 and 50 respectively and there are outliers around 60 and 10 and lower.

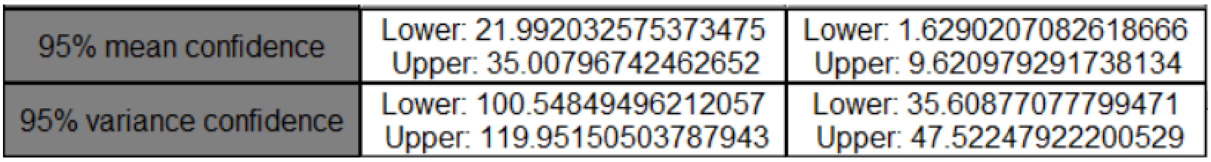
From the challenge rating box plot we can see that the median is around 3, the lower quartile (Q1) is close to 0, the upper quartile (Q3) is around 6. The min and max values (excluding outliers) are around 0 and 15 respectively and all values above 15 are considered outliers.

Bonus: Scatterplot



I made an extra scatter plot because I was curious about whether there is a correlation between speed and challenge rating. By looking at this scatter plot, we can conclude that generally challenge rating increases as speed increases.

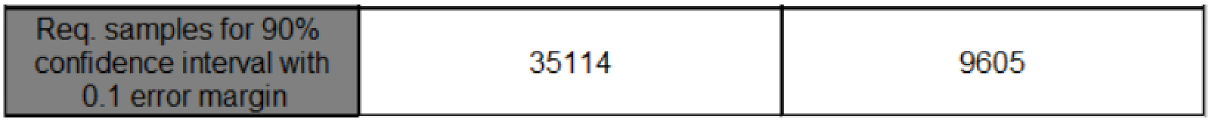
95% Confidence Intervals:



95% mean confidence	Lower: 21.992032575373475 Upper: 35.00796742462652	Lower: 1.6290207082618666 Upper: 9.620979291738134
95% variance confidence	Lower: 100.54849496212057 Upper: 119.95150503787943	Lower: 35.60877077799471 Upper: 47.52247922200529

My code takes in 20 different random samples every time it is run.
Values differ a little every time it is run, but they stay mostly similar each time.
The critical value is 1.96.

No. of Samples for 90% Confidence Intervals:



Req. samples for 90% confidence interval with 0.1 error margin	35114	9605

I was confused as to why these sample numbers are so high. But after doing research I concluded that these values are actually correct.

First of all, the dataset has a small number of records (323) so making predictions according to it will not be very accurate.
Secondly, the standard deviation generally is quite high, because the data in the dataset is very discrete, such as there are a lot of 30 and 40 speed monsters but no monsters at all with a speed between 30 and 40. (This pattern is very apparent in the speed histogram)
I think this is the reason the sample size needs to be very large in order to make correct predictions with 90% confidence rate and a 0.1 error margin.

I used a critical value of 1.645 and the margin of error is 0.1.

Conclusion:

This was a very interesting project for me, I learned a lot about statistics while doing it and it was also fun because of my interest in the dataset itself. Comparing the monster's speed and challenge rating was something I was really wondering about, so I'm glad I could answer that question!