Stats Research Report

The start of this project was difficult because the data was very messed up. Some numbers did not have any commas, and some did which messed things up when I was reading the csv. To deal with this program, I opened the csv file in excel and removed the decimals. There were also a couple where a category with mostly numbers had words instead, so I had to change them to -1. For example, the engine size category had a couple “electrics”, some “empty” data spaces had “N/A” or “- “, ‘+’, so I had either converted them to -1 or ignore it (like 1000+ or <9) which probably messed up with my results. Once that was done, it was time to start the real research paper.

Supercars are expensive, so to start off, I had to find the mean and the median of the prices. Surprisingly, the mean was a lot lower than expected despite there being multiple one million plus cars. The mean came out to be $382,035.94 which is a lot lower than expected, I thought it was going to be close to $500,000. Meanwhile, the median was $92,950.00. This made me understand why the median was a lot lower than expected. I also decided to find out the standard deviation which ended up as $738,322.72.

Chapter 1:

Knowing the mean, I asked, “What percentage of supercars have a price higher than the mean?” and the answer to that was 18%. Once again, I found out why the mean was surprisingly low.

Chapter 2:

Each list had 1,008 entries, so for the sake of this project, I decided to create some subsets with the size of 20 each. List A had the first 20 entries and List B had the next 20 entries. We find out that within the first 40 entries, there are only 22 supercar makes or car brands. Once again, using List A, given that the probability for each entry is equal, what is the probability that an “event” has stored a Porsche and an Audi? The answer to that is 0.15. I decided to look at the first 20 entries for prices, and this time, the probability that the car is from 2022 is 40% which is expected because most of the supercars from this datasest are either from 2021 or 2022. Once again, using List A, I decided to find out the number of ways of selecting two car makes from the list? In order to get this information, I had to call for my combinations function which gave me 190. Unfortunately, this was a far as I could get in chapter 2, because after this, a lot of conditional probabilities are needed which is difficult to get using my dataset.

Chapter 3:

The first few sections I was unable to do, but I was able to the probability distribution to find out some interesting information. First, I find out that there is a 38% chance that the supercar is from 2022. Knowing that there is a 38% of the supercar being from 2022, out a sample of 10, the probability that exactly 3 are from 2022 is 23.24% chance. With the same information, we were able to find the geometric distribution. The chance of the fifth entry being from 2022 is 14.61% chance. For negative binomial probability distribution, I decided to find what is the probability of the second 2022 car appearing at the fifth entry is 13.76% chance. Looking at the years of the first 20 entries again, when selecting 5 supercars, the chance of selecting 4 cars that are from 2022 is 5.42% which is also our hypergeometric probability distribution.