Sage Johnson

Dr. Cohen

CS 1030 – 003

5 May 2023

<div align="center">Mercury in Seafood – Final Project</div>

Module 1:

I felt confident in my skills after completing the first module in Khan Academy, due to my practice in analyzing data sets throughout this semester in my MTH 3210 class, probability and statistics. I was able to further hone my skills in recognizing trends in graphs, namely scatterplots. I did well on the two quizzes, scoring a 3 out of 4 on both. I did have to take both twice, however.

Module 2:

Well, I knew there was a lot of data out there, but I didn't know the specifics as to how fast it was being collected and stored! From scientific data to medical records and social media, we are quickly increasing the amount of data that we have available to us. I like that the underlying theme of this module was responsibility regarding how we process and store data about others. Privacy is a big topic in the data science field, and I am glad that I got to learn more about it. I scored a 75% on the module mini-test. I did have to take it twice. There were some tricky questions in there, but I feel like I learned from the mistakes I made on the first attempt.

Module 3:

Machine learning algorithms still seem a bit challenging for me to understand, even though I read the first article a couple times. I did appreciate hearing about the bias in machine learning in greater detail, as I found it fascinating (and alarming) when reading about it in

Hannah Fry's book, Hello World. I believe this module allowed me to further solidify my belief that machine learning should be taken with a grain of salt. While it can be quite powerful, it is important to remember that machine learning is not without its flaws, just like most things. I scored a 75% on the module test.

Unit Test:

I scored 7 out of 9 on the unit test on my first attempt, and I feel as though this was sufficient for me, considering I had little previous knowledge on any of these subjects, except for the first module. I found the unit test to be quite tricky; I think the multiple-choice answers caused more confusion than understanding. If I didn't have answers to read and choose from, I think I would have come to the same, correct conclusion. For example, if I only had a scatterplot and a regression line, I could easily identify the correlation…but with 4 potential answers to read through, I got confused.

The data set I retrieved was from the Canadian Food Inspection Agency. The link to this data set can be found here: https://open.canada.ca/data/dataset/7c2ec6ef-48de-4594-8c6a-b2fc2a5803ff The data set released is a compilation of raw data from 2000 – 2021, measuring the various metal content in seafood. Export, import, and domestic seafood were sampled, with over 50,000 different data points.

Research Questions:

- What is the average mercury content of all species of seafood sampled by the CFIA?

- Is there one species that generally has the highest concentration of mercury?

The process of answering these research questions was carried out as follows:

1) I first filtered the data in an Excel spreadsheet, so that only mercury would be shown (other types of metals were tested for in the report). Then, I applied various filters to account for the fact that mercury concentrations were measured using two different units: parts per million (ppm) and micrograms per gram (ug/g).

2) After filtering so that only (ug/g) would be shown, I calculated the average concentration of mercury using this unit of measurement.

3) Following the average for micrograms per gram, I then calculated the average concentration of mercury using parts per million (ppm).

4) Parts per million and micrograms per gram were then discovered to be the same measurement! So, the average of both was calculated to determine the total sample's average mercury concentration.

Average ug/g = **0.262212**

Average ppm = **0.163407**

Total average = (0.262212 + 0.163407) / 2 = 0.2128095

5) Finally, to answer the second research question, the data was sorted by decreasing levels of mercury concentration. After some analysis, it was determined that a reasonable conclusion could not be made due to outliers, as well as the fact that several species had high levels of mercury concentration, relative to the others. The top 5 highest concentrations, with one specification per species (as there were a few marlins with high concentrations) were:

i) Marlin: 7.3ppm

ii) Whelk: 4.184ppm

iii) Grouper: 3.744ppm

iv) Catfish: 3.66ppm

v) Shark: 3.4ppm

6) Since were did not exist an appropriate second variable (the date that the sample was gathered was not in proper form), it was not possible to create a scatterplot to find any linear relationship(s) between variables…at least without having to re-engineer several thousand data points.