Problem 1.1 (All Problem 1 Code in A1_Q1.py)

Solve for Mean, 5 number summary, standard deviation, and variance.

Antibiotics (1) Mean

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\overline{x} = \frac{1}{7}(14 + 30 + 8 + 8 + 7 + 3 + 11)$$

$$\boxed{\overline{x} = 11.57\ days}$$

No Antibiotics (0) Mean

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\overline{x} = \frac{1}{18}(5 + 10 + 6 + 11 + 5 + 11 + 17 + 3 + 9 + 3 + 5 + 5 + 4 + 7 + 9 + 11 + 9 + 4)$$

$$\boxed{\overline{x} = 7.44\ days}$$

5 Number Summary

Antibiotics (1)

- 3, 7, 8, 8, 11, 14, 30
- Minimum = 3
- Q1 = 7
- Q3 = 14
- Maximum = 30
- Median = 8

$$IQR = 14 - 7 = 7$$
$$1.5 * IQR = 10.5$$

No Antibiotics (0)

- 3, 3, 4, 4, 5, 5, 5, 6, 7, 9, 9, 9, 10, 11, 11, 11, 17
- Minimum = 3
- Q1 = 4.5
- Median = (6+7) * 0.5 = 6.5
- Q3 = 10.5
- Maximum = 17

$$IQR = 10 - 5 = 5$$
$$1.5 * IQR = 7.5$$

Standard Deviation Antibiotics

$$SD = \sqrt{\frac{\sum(x-u)^2}{n-1}} = \sqrt{\frac{\sum(x-11.57)^2}{n-1}} = 8.16$$

Standard Deviation No Antibiotics

$$SD = \sqrt{\frac{\sum(x-u)^2}{n-1}} = \sqrt{\frac{\sum(x-7.44)^2}{n-1}} = 3.59$$

Variance Antibiotics

$$Variance = \frac{\sum(x-u)^2}{n-1} = 66.5$$

Variance No Antibiotics

$$Variance = \frac{\sum(x-u)^2}{n-1} = 12.91$$

As seen above the average (or the means) number of days spent in the hospital if antibiotics are taken is 11.57 days whereas it is 7.44 for individuals who have not taken antibiotics. The mean is calculated for all patients who stayed a minimum of 3 days up to a maximum of 30 days In addition to this, the 5 number summary was also found. This includes the minimum, first quartile, median, third quartile and maximum values. For this, the antibiotics group had a higher median (8 > 6.5), higher maximum (30 > 17), a higher Q3 (14 > 10.5), and equal minimum of 3 and a higher Q1 (7 > 4.5). The 5 number summary is a collection of numbers that divides the datapoints based on equal quartile ranges. After this, the standard deviation was calculated. The standard deviation is a measure of the dispersion of the dataset relative to the mean and it is calculated by taking the square root of the variance. As seen in the calculations above, there is a greater standard deviation for the antibiotics group than the non-antibiotics group (8.16 > 3.59) showcase that there is greater dispersion away from the mean. The final thing calculated for each group was the variance. Variance is the measure of the spread between numbers in the dataset. From the dataset given, the antibiotic group had greater variance compared to the non-antibiotic group (66.5 > 12.91).

Problem 1.2 | Box Whisker Plot

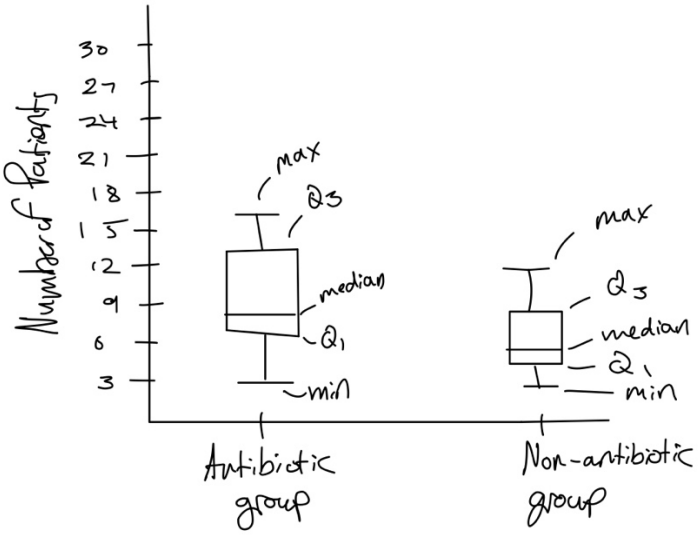

Figure 1: Box whisker plot for antibiotic taking and non-antibiotic taking groups

With respect to the central tendency, the box plot showcases that is it within the 6-9 range for both antibiotic and non-antibiotic taking groups. This can be seen when looking at the median which is 8 for the antibiotic taking group and 6.5 for the non-antibiotic taking group. Furthermore, another conclusion about the variance can also be gathered from the figure above. The antibiotic group displays a greater range of days when comparing both the Q3's and the max values showcasing that it has greater variance compared to the non-antibiotic taking group.

Problem 1.3 | Drawn Histogram

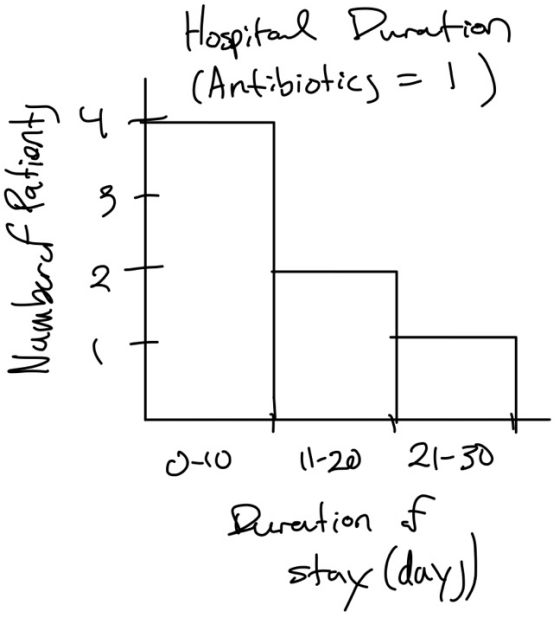

Figure 2: Histogram of Hospital Stay Duration for Group Taking Antibiotics

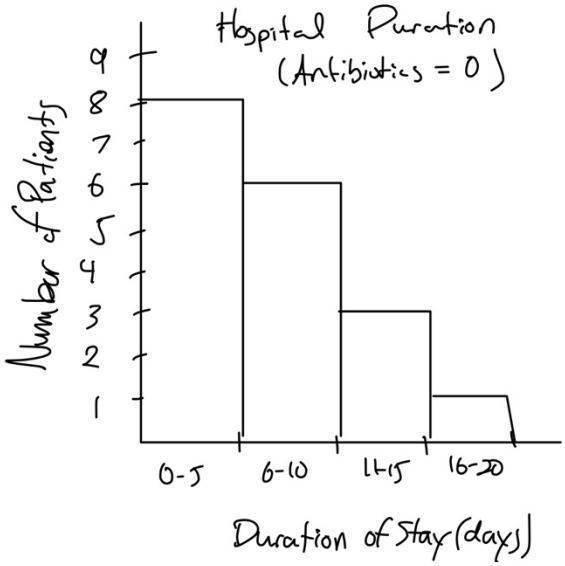

Figure 3: Histogram of Hospital Stay Duration for Group Not Taking Antibiotics

Problem 1.4

```
antibiotic = positive, stdv =  8.16

antibiotic = positive, mean =  11.57

antibiotic = positive, variance =  66.53

antibiotic = positive, min,Q1,median,Q3,max =  3 [ 7.5  8.  12.5] 30
antibiotic = negative, means =  7.44

antibiotic = negative, stdv =  3.59

antibiotic = negative, variance =  12.91

antibiotic = negative, min,Q1,median,Q3,max =  3 [5.  6.5 9.5] 17
```

Figure 4: Standard Deviation, Variance, Mean, and 5 Number Summary Obtained from Python Script for Antibody +/- Patients

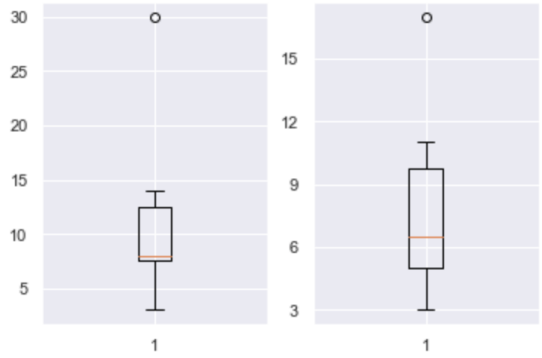

Figure 5: Boxplots for Antibody +/- Patients

Problem 1.5

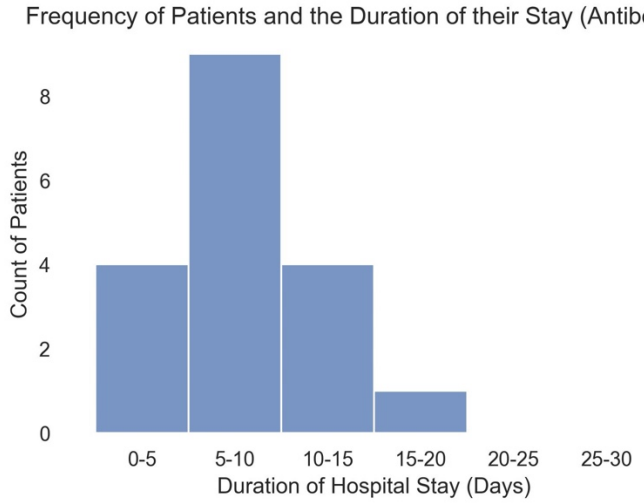

Figure 6: Histogram of the Number of Patients and the Duration of Their Stay (Antibody -)

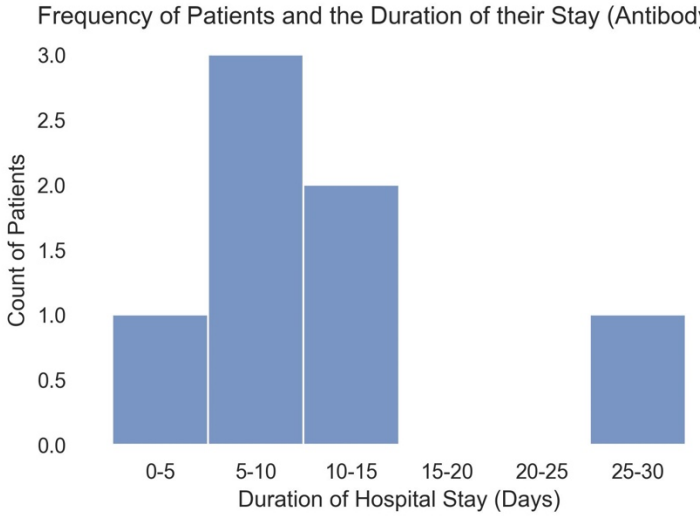

Figure 7: Histogram of the Number of Patients and the Duration of Their Stay (Antibody +)

Problem 1.6 | Are there any outliers present?

Yes, there are outliers present in both samples. In the antibiotic taking group, the outlier is the maximum duration spent, which was 30. This compared to the next closest value of 14 showcases the large gap between the rest of the dataset and this outlier. With respect the non-antibiotic taking group, the maximum days spent for a patient was 17 days. When compared to the next closest value of 11, there is a clear large gap. However, this isn't as large of an outlier compared to the one found in the antibiotic taking group, which is proved by that fact that the standard deviation for the non-antibiotics group is lower than that of the antibiotics taking group. If the outliers were removed, this would result in lower variance, and standard deviation for both groups. This is because outliers can have an impact on the dispersion of data away from the mean. It would also result in the Q3 value being closer to the median and a decreased IQR and whisker length.

Problem 2.1 (All Problem 2 Code in A1_Q2.py)
After a brief inspection using python, Categories of Sex, Chest pain, Trestbps, Chol, Fbs, Restecg, Thalach, Exang, Oldpeak, Slope, Ca, and Thal all have rows that contain missing data values. These rows can be removed from the dataset using the python function .dropna(). This will drop rows with NA data values and will make it such that all the columns have an equal number of rows with values ≥ 0. The specific number of missing values for each column/category can be seen down below.

- Sex – missing 2
- Chest pain – missing 2
- Trestbps – missing 3
- Chol – missing 3
- Fbs - missing 3
- Restecg – missing 3
- Thalach – missing 4
- Exang – missing 3
- Oldpeak – missing 3
- Slope – missing 3
- Ca – missing 1
- Thal - missing 1

Problem 2.2
The datatypes for Age, Slope & Heart_Disease are object, object, and int64 as determined via the python code found in Appendix B. Age and Slope don't match the required datatype. To be able to perform any numeric operations on the data contained within these columns, they would need to be a float or integer value. The reason for this data mismatch can be determined when observing the data table within the csv file a little more closely. Within the 'Age' column, the third entry from the top isn't an integer value, but instead it is a string with the name "Terry". The same can be observed in the column for 'Slope'. As a couple of entries, namely rows 194, 293, and 300 contain non integer values such as: "?", "yes", and "A" respectively. The reason these columns have been assigned an object datatype is because object datatypes in pandas and python are a mix of strings, numeric and non-numeric values. To fix this problem, the pandas function pd.to_numeric() can be used to convert the object data types of the aforementioned columns into float values. This can be seen in the code found in python file.

Problem 2.3
The limit of plausible data for Heart_disease would be, $0 \leq Heart\ disease \leq 1$. The limit for age was determined by determining the max value in the Age column (see Appendix B for code). When this was done, a value of 165 was returned. According to the published statistics, the oldest person to have lived was 122 years old. So, with that, it can be inferred that the value of 165 is an incorrect entry. Therefore, the limit of plausible data was then set to, $0 \leq Age \leq 100$. Finally, the limit of plausible data for slope was set to $1 \leq Slope \leq 3$, as all slopes of the ST segments of the ECG should be kept. See Appendix B for code.

Problem 2.4
See A1_Q2.py.

Problem 2.5

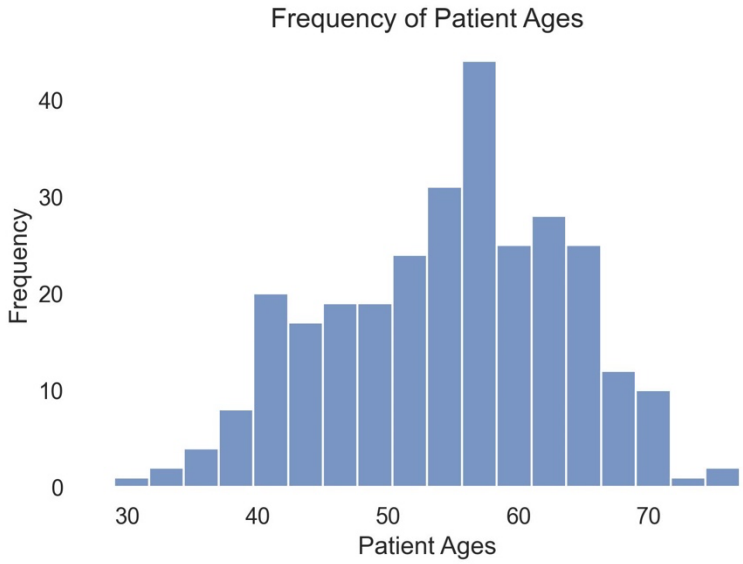

Figure 8: Histogram Plot Showcasing Frequency of Patient's Age (bins = 17)

17 bins were used to create this histogram. This is because the sample size of 293 was square rooted to provide a value of 17.11, which was rounded down to 17. The data was cleaned and all values of NaN were removed. All values above the age of 100 were removed as well since there was one outlier of 165, which I assumed was an input error.

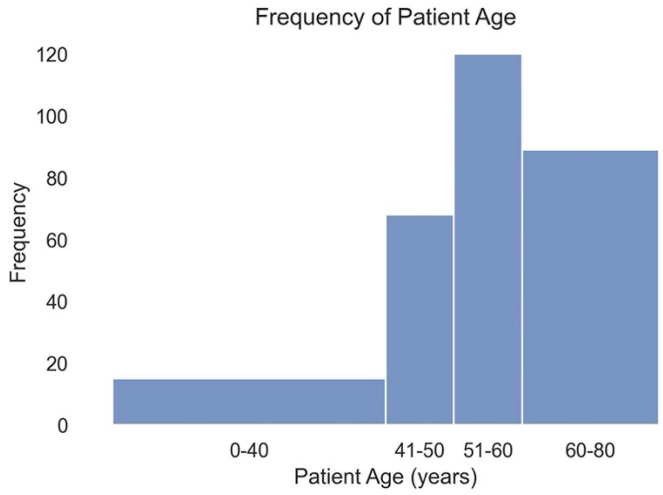

Figure 9: Histogram Plot Showcasing Frequency of Patient's Age Categorized by Age Group

Custom bin sizes were used in the construction of this Histogram. Data was cleaned and organized in the same step as the data from the histogram plot in figure 8.

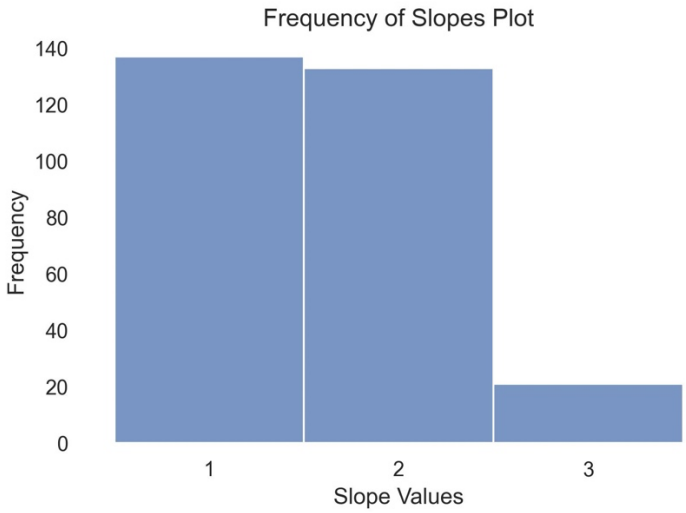

Figure 10: Histogram Plot Showcasing Frequency of Slope Values

Custom bin sizes were used to depict the correct sample size. Cleaning and organization of data was done in the same process and step as the data from both figures 8 and 9.

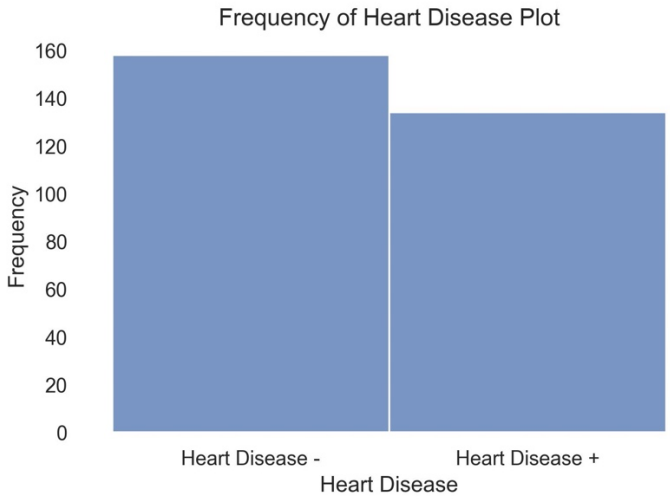

Figure 11: Histogram Plot Showcasing Frequency of Heart Disease Presence

Custom bin sizes were used to depict the correct sample size. Cleaning and organization of data was done in the same process and step as the data from both figures 8 and 9.

Discussion

The data was cleaned and organized into a pandas' data frame in the previous steps by eliminating all rows with NaN values. All the datatypes were converted into float values to integer values such that numeric manipulations could occur. For the purposes of this question specifically, histograms were used to showcase the Age, Age_Category, Slope, and Heart_Disease columns of the pandas data frames. This was because these are categorical topics, with fixed constant sample sizes, hence it made sense to use a histogram. Additionally, the data frames measured the frequency of certain events occurring within a specific category which can easily be depicted by a histogram.

Problem 3 (All Problem 3 Code in A1_Q3.py)

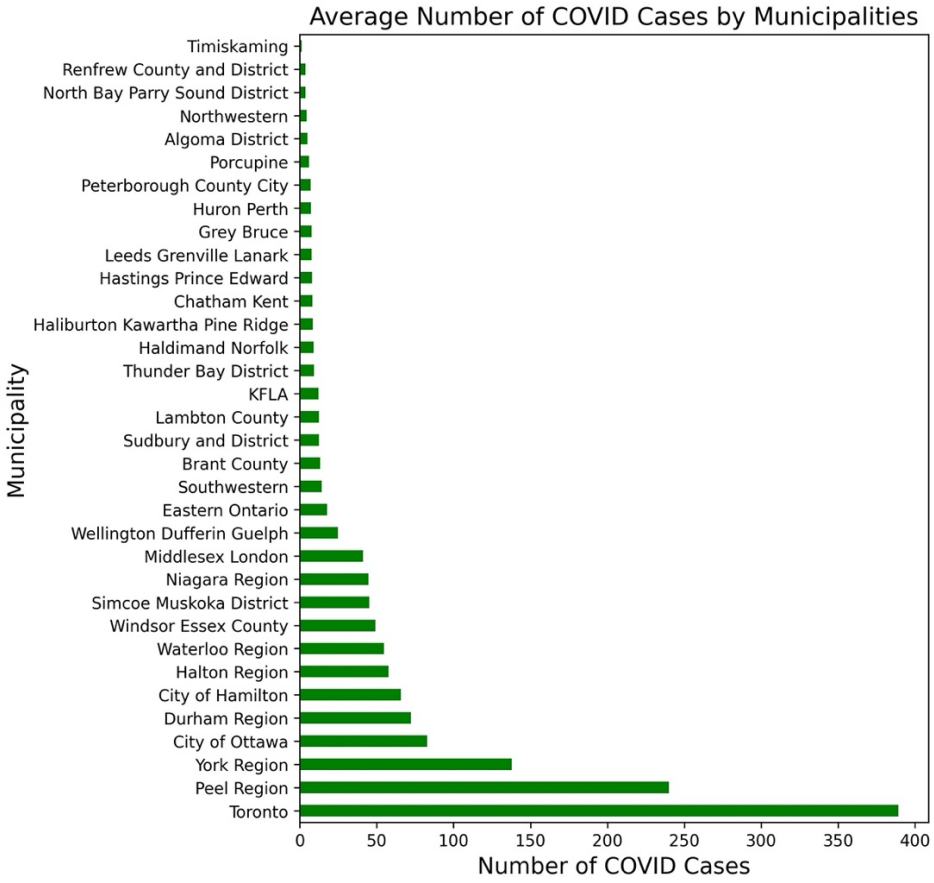

Figure 12: Bar Group Showcasing Average Number of COVID-19 Cases by Municipality in Ascending Order

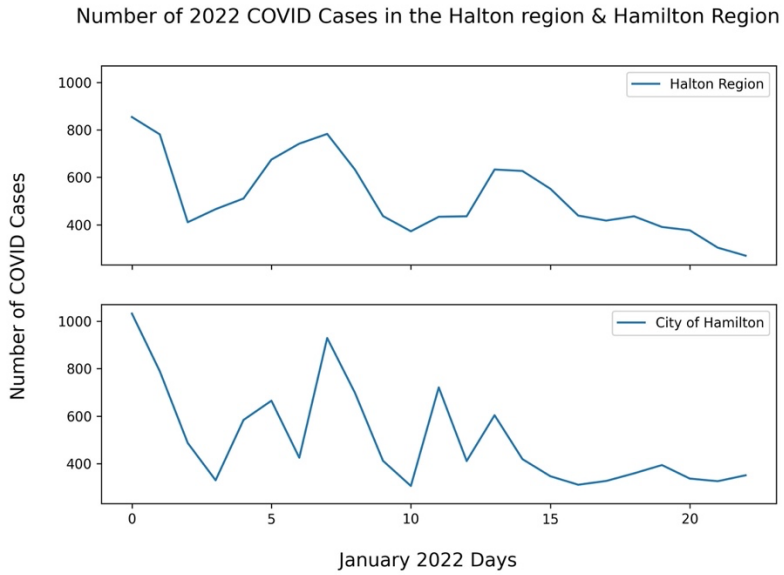

Figure 13: Time Series Graph Showcasing Number of 2022 COVID-19 Cases in Halton & Hamilton Region

Discussion

The data in the box plot above in figure 12 was cleaned and organized by removing all NaN values and converting all values to the correct data type (float). All the "_" characters found in the columns were removed and replaced with spaces to make the data look cleaner. The graph was plotted with the municipalities on the vertical axis as when on the horizontal axis, the municipalities are difficult to read making the delivery of the data less than ideal. Furthermore, the data was organized in increasing order to showcase a trend of least covid cases to highest covid cases based on location. This makes the information less jumbled to the viewer and more easily digestible. Finally, the use of the white background allows for the data and information to be the subject in focus and to be the thing that stands out the most.

The data in the timeseries plot in figure 13 showcases the trend of COVID cases in 2022 for the regions of Halton and Hamilton. These two were selected to showcase a comparison between two neighboring municipalities and to see how COVID-19 has spread. Both the x-axis and the y-axis are shared to ensure that the time sample used is the same and the axis is the same to ensure a fair comparison. The linewidth is slightly increased so that the trend is more easily seen. And a legend is applied so that it is obvious which graph pertains to Halton and which one pertains to Hamilton.

Problem 4.1.1 – Fox News Critique

The figure depicted in the Fox news report has a very poor ink to data ratio. There is hardly any data presented, however most of the space is covered by large, bolded text and graphic elements such as colouring in blank spaces. Additionally, the timeframe of the statistic makes it difficult to analyze for any trends present. It is impossible to see how President Obama compares to previous presidents or even how he compared to himself through his entire term. Additionally, the colour scheme of red on a dark blue makes it difficult to read the graphic and unappealing to the eye. Many improvements could be made, starting with using Tufte's principles, specifically the one pertaining to ink reduction. If the ink were reduced and the information was increased, this could help improve the graphic vastly.

Problem 4.1.2 – MSNBC Critique

The figure depicted in the MSNBC report, like the Fox news one also has a very poor ink to data ratio. The grid space in the background is covered by a gradient which does not help aide the graphic in any way and does not improve the delivery of information. Furthermore, this is just a time-based trend, which could be better depicted through a simple time series graph. By using a bar plot, the graph creates holes and inconsistencies with the data and simply does not display the nuances of the trend line as efficiently as a normal trend line graph would. Additionally, the scale of the x-axis distorts the perception of the data due to its nonlinear scale. It is difficult to understand whether there was an exponential increase in the coronavirus cases simply from looking at this graph as it is implying. Finally, improvements can be made regarding the colour scheme. If the gradient background is removed and replaced with a simpler white background, it can make it easier to see the data points across many different kinds of screens.

Problem 4.2

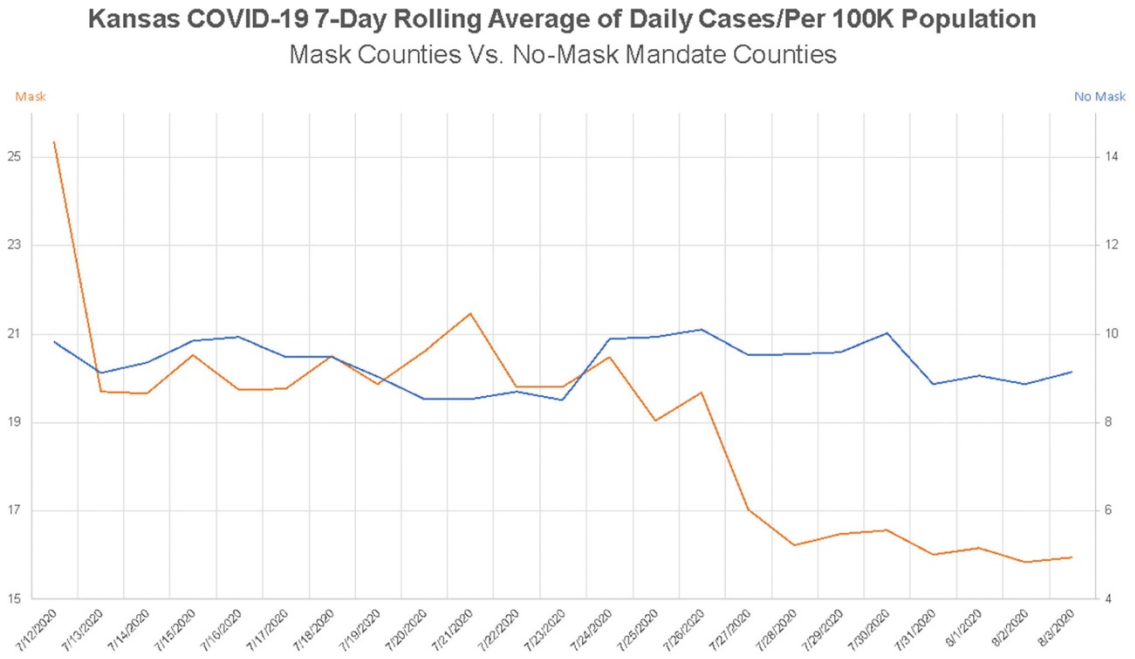

Figure 14: Graph Depicting COVID-19 7-Day Rolling Average of Cases/Per 100K Population [1]

While at first glance this may not seem like a problematic graph, it begins to have some more problems when considering the context of who this was intended for. This graph was intended to inform the public by the Kansas Department of Health & Environment. The graph contains two vertical axes', of different scales, which can make it very difficult to understand the relationship between the two things being compared, in this case being the effectiveness of masks vs no-masks at reducing COVID-19 transmission. Tufte has mentioned that vertical axis' can be used to show a potential association between things, however, in this case, we do not see an association, instead we are seeing a comparison, hence making the use of two vertical axes as poor choice. To remediate this, either two timeseries graphs should be presented with equal scales and a shared x and y axis to showcase a truthful comparison between the two. Or the two could be plotted on the same y-axis with equal scale. Additionally, the ink to data ratio is quite low in this figure. There is a lot of white space, and the dates are difficult to read. The figure is not friendly for all devices making it very difficult to read on handheld devices such as smartphones. To remediate this, the text size could either be increased, the linewidth could be increased, and sharper colours could be used to better communicate the data.

Problem 5.1

1. Probability of an event must fall between 0 and 1.
2. Summation of all probabilities equals 1.
3. The complement rule states that the complement of an events probability is mutually exclusive. This is calculated by 1 minus the probability of the complement. (Insert equation here)
4. This rule refers to the probability that both independent events A and B occur representing the Logical AND. P (A OR B) = P(A)P(B). When both events are dependent of one another, the rule slightly changes. P (B|A) = P (A AND B)/P(A)
5. Mutually exclusive events don't possess the ability to occur simultaneously. Thus, the equation for this addition rule becomes. P (A OR B) = P (A) + P (B) (1 – P (A)).

6. For two non-mutually exclusive events, the addition rule becomes. P (A OR B) = P (A) + P (B) – P (A AND B)
7. The Bayes' rule is used to calculate the probability of an event given the following equation. P(B|A) = P(A|B) * P(B)/P(A).

Problem 5.2

First, we need to solve for the probability that the switch C is closed using the complement rule.

P (Switch C is Closed) = 1 – P (Switch C is Opened)
P (Switch C is Closed) = 1 – 0.15
P (Switch C is Closed) = 0.85

Then we need to solve for the probability that the first stage is closed.

P (Stage 1) = P (A closed) + P (C closed) – P (A closed) * P (C closed)
P (Stage 1) = 0.8 + 0.85 – (0.8*0.85)
P (Stage 1) = 0.97

Furthermore, since the same set of switches are used for the second stage, the probability of the second stage being closed (and working) is the same as the first, that being 0.97. Therefore, the overall probability of the whole circuit working is then determined down below.

P (works) = P (section 1) * P (section 2) * P (Switch B is Closed)
P (works) = 0.97 * 0.97 * 0.9
P (works) ~= 0.85

Therefore, the probability that the circuit works is about 0.85 or about 85%.

Problem 5.3

The probability of defectiveness can be determined via the following equation.

P(defective) = sum(P(i)P(i_defective))
P(defective) = 0.25(0.05) + 0.25(0.4) + 0.25(0.1) + (0.25)(0.25) = 0.2

Problem 5.4

To solve for the probability of having picked from the second box if in the event that the light bulb is defective can be solved by Bayes' rule. The calculation for this is down below.

P (A|B) = P(B|A) P(A) / P(B)
P (A|B) = (0.45*0.25) / 0.2
P (A|B) = 0.5

References

[1]. C. Engledowl and T. Weiland, "Data (mis)representation and covid-19: Leveraging misleading data visualizations for developing statistical literacy across grades 6–16," *Taylor & Francis*, 19-May-2021. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/26939169.2021.1915215. [Accessed: 21-Jan-2022].