Homework 2

1. Five analytics queries or questions:
    - Are there any outliers for the employment to population ratio within the very high human development indicator?
    - Are countries with low human development indicators grouped together or close to each other?
    - Is there a correlation between infants being exclusively breastfed and HIV prevalence in adults within the very high human development indicator?
    - Comparing and contrasting female to male literacy rates, which gender accounts for the highest literacy rates and for which countries?
    - What is the average government expenditure for each different level/indicator of human development?
2. Five insights:
    - The correlation between infants that were exclusively breastfed and HIV prevalence in adults within the low human development indicator is 0.177 which indicates a weak association.
    - The country (Kuwait) with the lowest percentage of the population that has reached the level of secondary education in the very high human development indicator is still 51.5% higher than the country (Grenada) with the lowest percentage in the high human development indicator.
    - The average percentage of government expenditure on education for very high human development is about 1.2 times greater than the average for the low human development indicator.
    - The country with the greatest gap in the literacy rate between the female and male population is Guinea-Bissau which has about a 21.6 percent gap.
    - The total population in millions for 2019 for the very high human development indicator sums to 1562.5 million compared to 921.4 million for the very low human development indicator which is a reasonable difference.
3. Four Steps/Tasks Performed:
    - In order to find correlation, I used the built in function that excel has called CORELL which takes in two arrays or columns. I was interested in whether being breastfed as an infant led to less health illnesses, specifically HIV prevalence in adults.
    - I sorted the column for “population with at least some secondary education” for the very high human development and high human development indicators in descending order and then I picked the

minimum of each indicator and subtracted the two to find how much higher the population percentage was for one over the other.

- I was curious what the average was for government expenditure on education for the very high human development indicator, so I just highlighted the column with the corresponding numbers and the average appears at the very bottom of the screen.
- I subtracted one column from another by doing **=(first column data – second column data)** in order to find the country that had the greatest gap in the literacy rate between the male and female population.
- I wanted to see the population trend for the total of the very high human development and very low human development indicators to see if there was a major difference between them. In order to carry this out, I just highlighted the cells corresponding to the "total" column for 2019 for each indicator and the sum gets automatically calculated as you highlight a new cell in the column.

4. Rating Helpfulness of Visualization:
    a. Low because the average cannot be seen intuitively, but you could easily calculate these values using the built-in function found in excel for finding averages and you could also sort by gender.
    b. High because there is no way to sort the data according to countries located near the equator, in order to get all these values for the unemployment rate you would need to sort them which in this case cannot be done easily. A visualization would really aid in conveying the unemployment rate especially if there are a lot of countries you are gathering data for.
    c. Medium because the correlation is not something that is seen right away, you would need to use the built in function found in excel called "CORREL." A visualization could be helpful here just to see the relationship if there is one or is not one.
    d. High because I would assume not many people can picture a distribution from just seeing a data set. Additionally, it would be difficult to come up with a non-visual distribution with different attributes (columns) that might have different types of data.
    e. Low because you could easily sort by gender and country as well as sort in ascending or descending order to see which countries have the most female internet users.
5. Data Classification
    a. ***Labour force participation rate*** is easily defined with equal intervals because the rates are percentages, you could easily do bins of equal length that add up to 100%. Whereas, with the ***Refugees by country of origin***, the data is by thousands with numbers going up to as high as

20,000, so this would not be easily defined with equal intervals since the data has more range.

b. Quantiles make more sense for ***Female youth literacy rate*** because the data is linear and evenly distributed and there is not that much deviation or gaps in the data values.