**Instructions: Data Science in Global Health**

## **Part I: Assessing Data Quality**

**Data Quality Dimensions**

**Accuracy** – The extent to which the result of a measurement, calculation, or specification conforms to the correct value or a standard.

**Completeness** – The extent to which there are any gaps in the data from what was expected to be collected, and what was actually collected.

**Recency** – The extent to which the data contains observations from recent time periods.

**Representativeness** – The extent to which the data provides insight into what is happening with the broader population of interest.

**Answer the following questions about data accuracy based on the presentation and the data set provided.**

1. **Accuracy:** What are some potential sources of bias in the way the data is collected? Are there any data points different from the others? Which ones?
2. **Completeness:** Is there any data missing that does not seem to be random? If so, which data points appear to be missing?
3. **Recency:** How recent were the latest observations in our dataset?
4. **Representativeness:** Can we make conclusions about air quality in all of Washington with observations we have in our data?

**Part II: Generating Descriptive Statistics Using a Spreadsheet**

**Student Spreadsheet:** <https://tinyurl.com/data-driven-lesson>

**Step 1: Creating a new database in Google Sheets:**

* Go to <https://tinyurl.com/data-driven-lesson>.
* Press Control + A to select all cells in the data set.
* Press Control + C to copy the data set.
* Press File > New > Spreadsheet.
* Press Control + V to paste data into a new sheet.

**Step 2: Finding your city:**

* Create a filter on the cell titled “city”.
* Select your city and filter out all other cities.

**Step 3: Identifying and removing outliers:**

* Based on observations from Part I, identify outliers for your city.
  + Examples: negative values, missing values (“NA”), or very large values.
  + Mark them “1” in the “outlier” column.
* Create a filter on the cell titled “outlier”.
* Filter out any values which are outliers.
  + 1 = outlier
  + 0 = not an outlier
* We now have a clean dataset!

**TIP: How to Filter Your Data in Google Docs**

1. Select a range of cells by clicking on them.
2. Click Data and then Create a filter.
3. To see filter options, go to the top of the range and click Filter Filter.
   * Filter by condition: Choose from a list of conditions or write your own. For example, if the cell is empty, if data is less than a certain number, or if the text contains a certain letter or phrase.
   * Filter by values: Uncheck any data points that you want to hide and click OK. If you want to choose all data points, click Select all. You can also uncheck all data points, by clicking Clear.
   * Search: Search for data points by typing in the search box. For example, typing "P" will shorten your list to just the names that start with P.
4. To turn the filter off, click Data and then Turn off filter.

Credit: Google Sheets instructions from support.google.com.

**Step 4: Creating a clean datasheet:**

* Press Control + A to select all cells in the clean data set.
* Press Control + C to copy the clean data set.
* Press the + button at the bottom left of the spreadsheet to “Add Sheet”.
* Press Control + V to paste data into the new “Sheet2”.

**Step 5: Finding the minimum:**

* In “Sheet2”, sort the PM2.5 column from largest to smallest (or smallest to largest) and find the minimum value for their city.
  + Hint: you might have to create a new filter.
* Or, use the minimum function.
  + In a new blank cell type the following function
    - = min (highlight all the cells for their city)
    - Press enter

**Step 6: Finding the maximum:**

* Sort the PM2.5 column from largest to smallest (or smallest to largest) and find the maximum value for their city.
* Or, use the maximum function.
  + In a new blank cell type the following function
    - = max (highlight all the cells for their city)
    - Press enter

**Step 7: Finding the median:**

* Sort all values by largest to smallest and find the middle value.
* Or, use the median function.
  + In a new blank cell type the following function
  + = median (highlight all the cells for their city)
  + Press enter

## **TIP: How to Sort a Range of Data in Google Docs**

1. Highlight the group of cells you would like to sort.
2. Click **Data** and then **Sort range**.
3. If your columns have titles, click **Data has header row**.
4. Select the column you would like to be sorted first and whether you would like that column sorted in ascending or descending order. This also sorts numbers.
   * If needed, click **+Add another sort column** to add another sorting rule. Sorting will be done according to the order of your rules.
   * To delete a rule, click Close Close.
5. Click **Sort**. Your range will be sorted.

Credit: Google Sheets instructions from support.google.com.

**Step 8: Calculating the mean:**

* Add together all the values by hand.
  + Take a calculator and add up all the values.
  + Divide by the number of values (or rows).
* Or, use the average function.
  + In a new blank cell type the following function:
    - = average (highlight all the cells for their city)
    - Press enter

**Step 9: Calculating variance:**

* In a new blank cell type the following function:
  + = var (highlight all the cells for their city)
  + Press enter

## **Part III: Comparing Metrics across Two Cities**

**Definitions:**

* **Minimum** – The smallest observation.
* **Maximum** – The largest observation.
* **Median** – The middle observation in a ranked list of observations.
* **Mean** – The sum of the observations over the number of observations.
* **Variance** – The sum of the squared differences between each observation and the mean of all observations.

**Fill out the table and answer the questions that follow:**

|  |  |  |
| --- | --- | --- |
| Statistic | City 1  NAME: | City 2  NaME: |
| Minimum |  |  |
| Maximum |  |  |
| Median |  |  |
| Mean |  |  |
| Variance |  |  |

1. What interesting differences between cities can you identify?
2. What do these differences in the data mean? What might they imply about air quality in these two cities?
3. What are some possible explanations for these differences?

## **Part IV: Comparing Washington State and Global Air Quality**

Now that you have examined air quality data for specific cities/areas of Washington State and compared data across different regions, let’s take a global look at air quality. In this section, you will be examining air quality data for different countries of your choosing using a website called State of Global Air.

**Exploring air quality in different countries**

1. First, examine the air quality of different countries around the world. You will want to find a country with lower, higher, and equivalent values as compared to the maximum value in Washington State. You will first need to go back to the original Washington State dataset and find the maximum value of PM2.5.
2. Write it down here: **Washington State maximum PM2.5:** \_\_\_\_\_\_\_\_
3. Next, check out the “map view” option of the State of Global Air website: <http://www.stateofglobalair.org/air#PM>
   * Move cursor over countries to view levels of PM2.5
   * Write down the three countries you chose:
     + **Lower:**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ **PM2.5:**\_\_\_\_\_\_\_\_\_\_\_\_\_
     + **Higher:**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ **PM2.5:**\_\_\_\_\_\_\_\_\_\_\_\_\_
     + **Equivalent:**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ **PM2.5:**\_\_\_\_\_\_\_\_\_\_\_\_\_
4. Now, explore one of the countries you identified in more detail. To do this, check out the “plot view” option of the State of Global Air website: <http://www.stateofglobalair.org/data/#/air/plot>
   * Go to the “Choose a country” tab
     + In the drop-down menu, pick one of the countries you identified from the map view.
     + Click on that country’s name.
   * The plot should generate a yellow line representing PM2.5 levels between 1990 and 2015.
   * To compare the three country PM2.5 in a time series, you can add a country by selecting them in the “+Add countries” drop down menu. Compare the data across all three countries.
5. Record some observations of the data for the three countries of your choice. What do you notice? What do you wonder about?
6. What we can do with your Washington State data and this global data? How might it help inform policy, action, or education in order to improve air quality/pollution across Washington State and globally? Record your thoughts and ideas below.