Milestone 1 Project — Written Portion

# Assignment Overview:

* The Milestone 1 Project is split into a **Written Portion** and an **Excel Workbook** portion. You can download the Excel Workbook portion through D2L.
* Write your answers to the written portion directly into this document, and submit it, along with your completed Excel Workbook, to D2L.
* Make sure that your write **all** your answers in this Word document in **red**, and that you **do not** remove any text or information from this Word document.
* In addition, to completing this document, the other half of Milestone 1 includes completing a separate set of exercises in an Excel workbook. The content of those exercises is not directly related to the ones in this document.

# Part 1: Reviewing Terminology

**Answer the Following Questions about quantitative data in Complete Sentences:**

1. Why do we bother calculating **mean**, **median**, or **mode**?

The mean, median and mode gives us an Idea of a "normal" value In the dataset. Each of these values give a different representation of what this value Is.

1. If mean, median, and mode all try to answer the same overall question, why do we need three separate numbers/formulas/strategies to answer that question?

These values have different meanings and purpose. For example, the mean can give us an Idea of the average but It can be heavily Influenced by outliers which can create Inaccurate real average value. The median gives us the middle value of the dataset If they are sorted In ascending / descending order. It Is also not Influenced by extreme values. The mode gives us the most common value In the dataset which may be useful In certain data such as color.

1. **Standard Deviation**, **Range**, and **Interquartile Range** can all be used to measure how “spread out” your data is. What are the differences between each of these measures of spread?

**The range only gives us the maximum value and the minimum value of the dataset. It does not consider any outliers that could Influence the data so It may be Inaccurate. The Interquartile range gives us a view of the middle of the dataset while Ignoring the outliers. It gives us a more accurate spread of the data. The standard deviation on the other hand shows us how far the data is from the mean of the whole range of data. It considers the outliers as It Is dependent on the mean for Its calculation.**

1. What is the purpose of calculating **z-scores** or **percentiles** for individual observations?

The z-score which Is dependent on the standard deviation gives us an Idea of how many standard deviations away the data is from the mean of the data. It can give us a quick view of where a specific data Is In comparison to the mean. The percentile on the other hand shows us how much of the data falls below the percentile value. This gives us an Idea of the position of a specific datapoint relative to the entire data.

1. If a histogram or boxplot appears to be **right-skewed**, what does that tell you about how the data is distributed? Are you more likely to find smaller numbers or larger numbers in your data? Justify your answers.

If a histogram is right-skewed, it would mean a large part of the data is on the left side of the histogram. The values represented by the histogram start from the left to right starting from the smaller values. Therefore, a histogram with a right skew would mean most of the values would be towards the smaller values as they likely occur more frequently.

# Part 2: Thinking About Careers

**Respond to the Part 2 prompts WITHOUT looking up the answers online:**

1. Name one career/job that might interest you.

Network Administrator

1. If you had to guess, how much money would you expect to earn each year with a career like that in the United States? **Do not look up anything online to answer this question**; just use your gut feeling.

75,000$/year

1. Without looking it up, what would you expect the **minimum** salary to be for that career?

45,000$/year

1. Without looking it up, what would you expect the **maximum** salary to be for that career?

200,000$/year

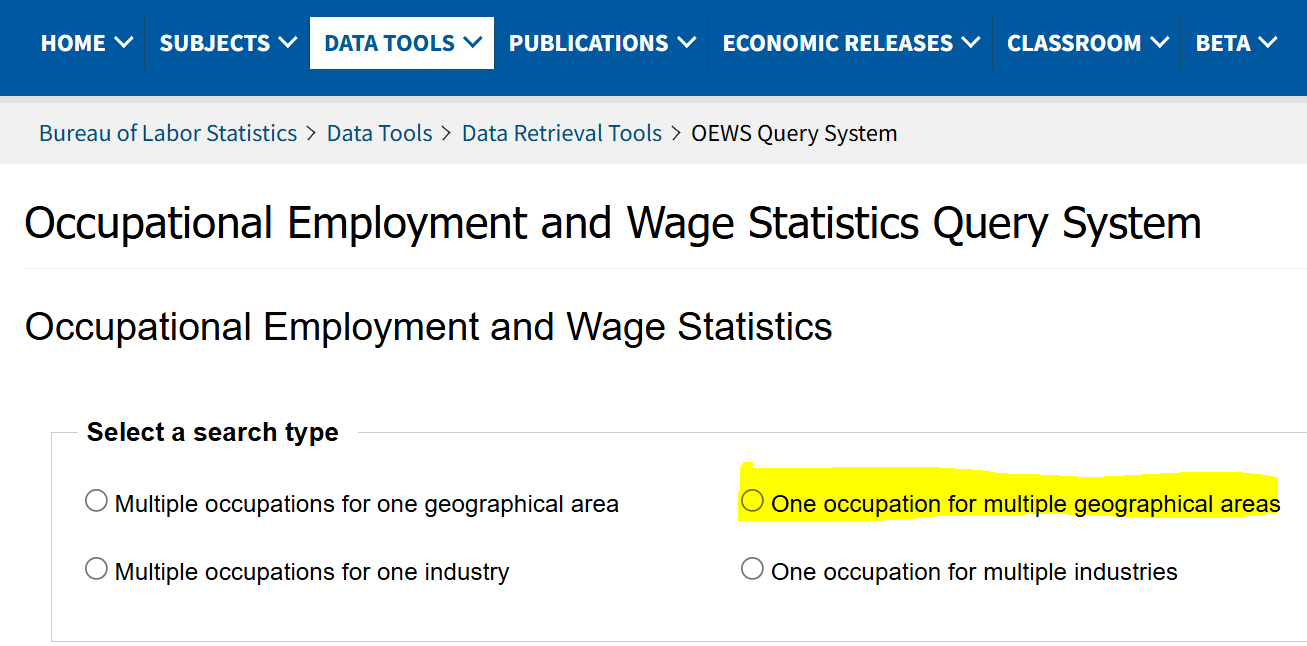
1. How would you expect the median salary for your chosen career to compare to the median salary for *all careers* in the U.S.?

It would probably be higher than the average median salary.

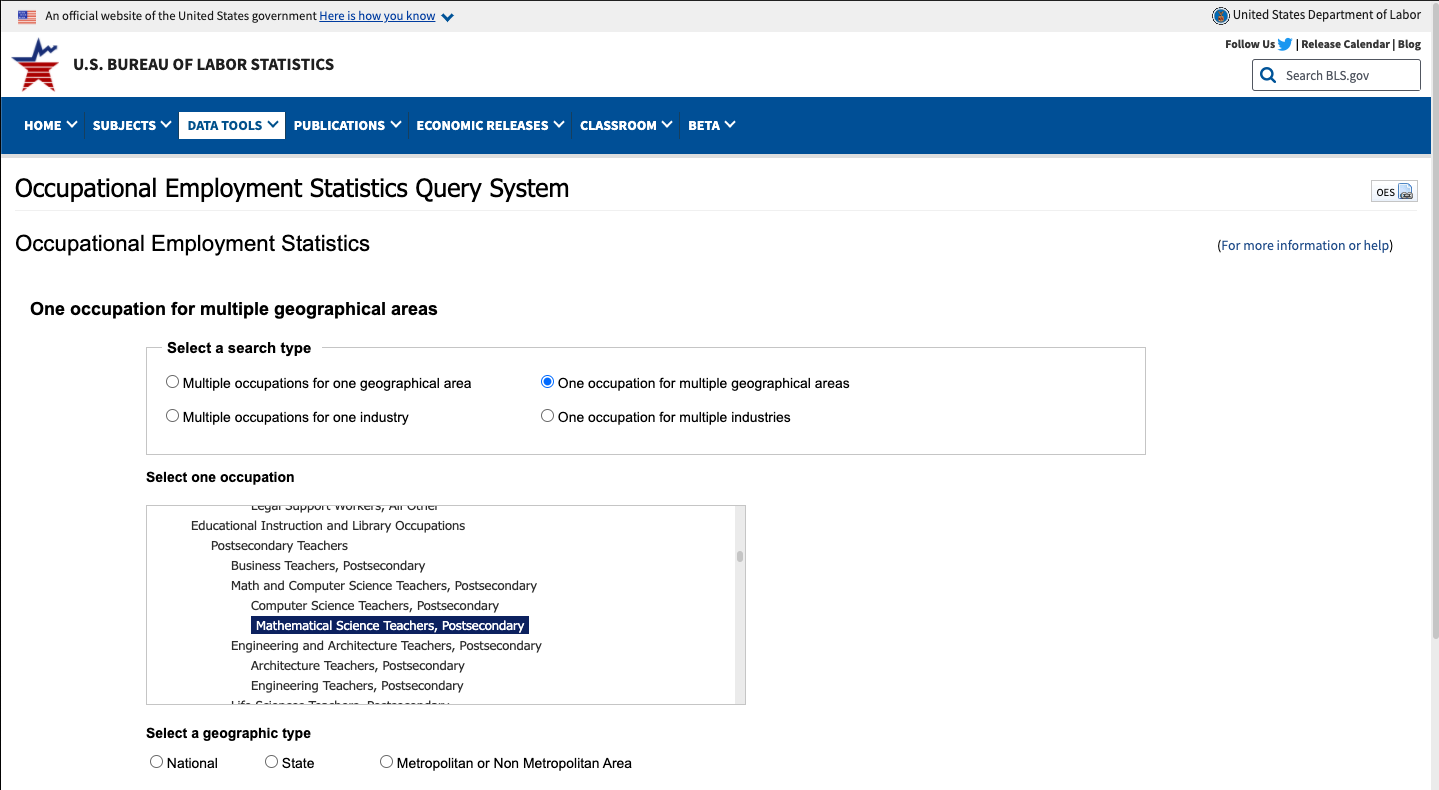
# Part 3: Collecting Salary Information from the Bureau of Labor Statistics

For this part of the assignment, you will go to the United States Bureau of Labor Statistics to further investigate your career. **Follow the listed steps for Part 3:**

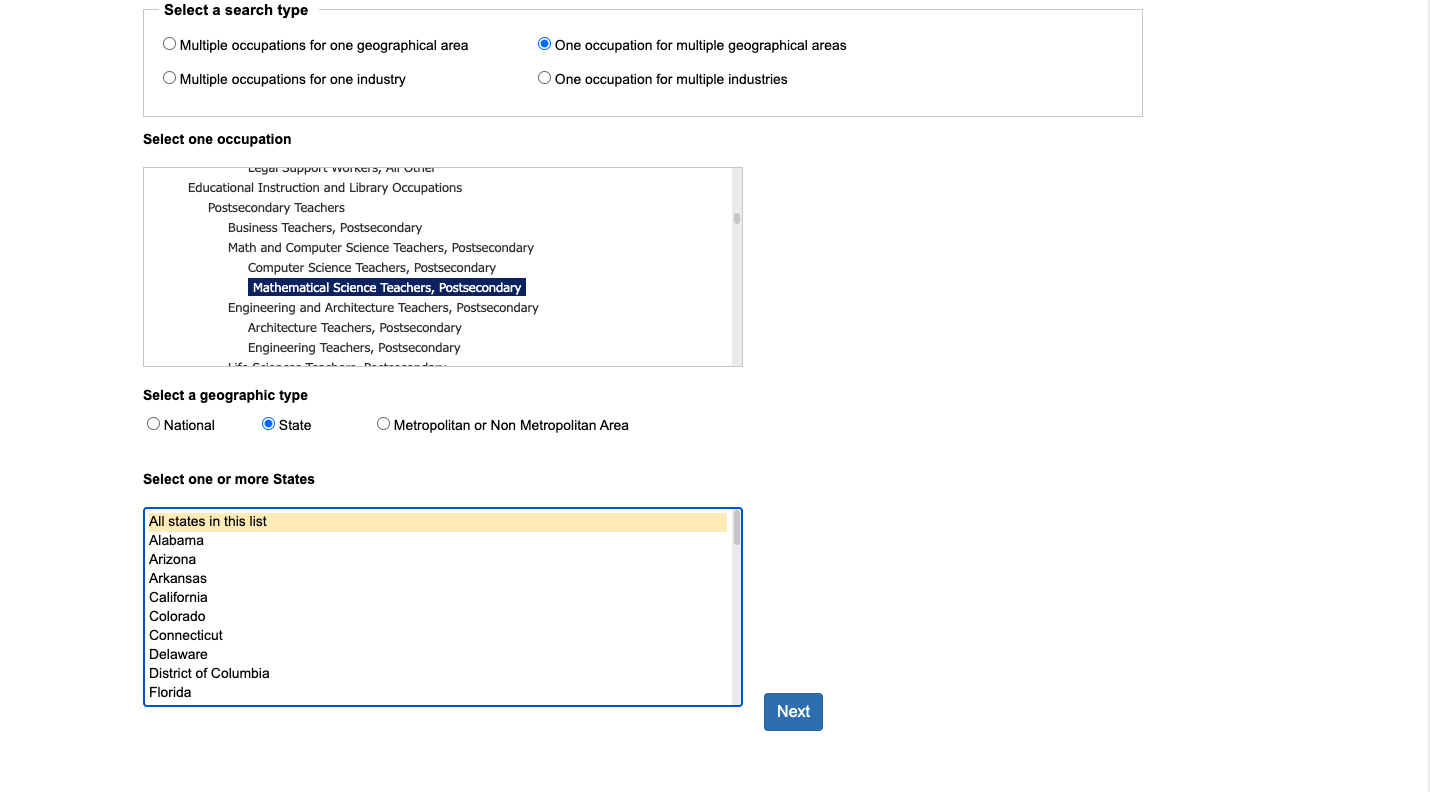
1. Go to <https://data.bls.gov/oes/#/home> .  
   This website allows you to search summary statistics about certain careers in the U.S.
2. For “search type”, Select the option “One occupation for multiple geographical areas”:



1. From the list of occupations that pops up, scroll through and choose one that best fits with your desired career. For example, Lucas chose “Mathematical Science Teachers, Postsecondary”. \*use Ctrl+F or Cmd+F to Search.

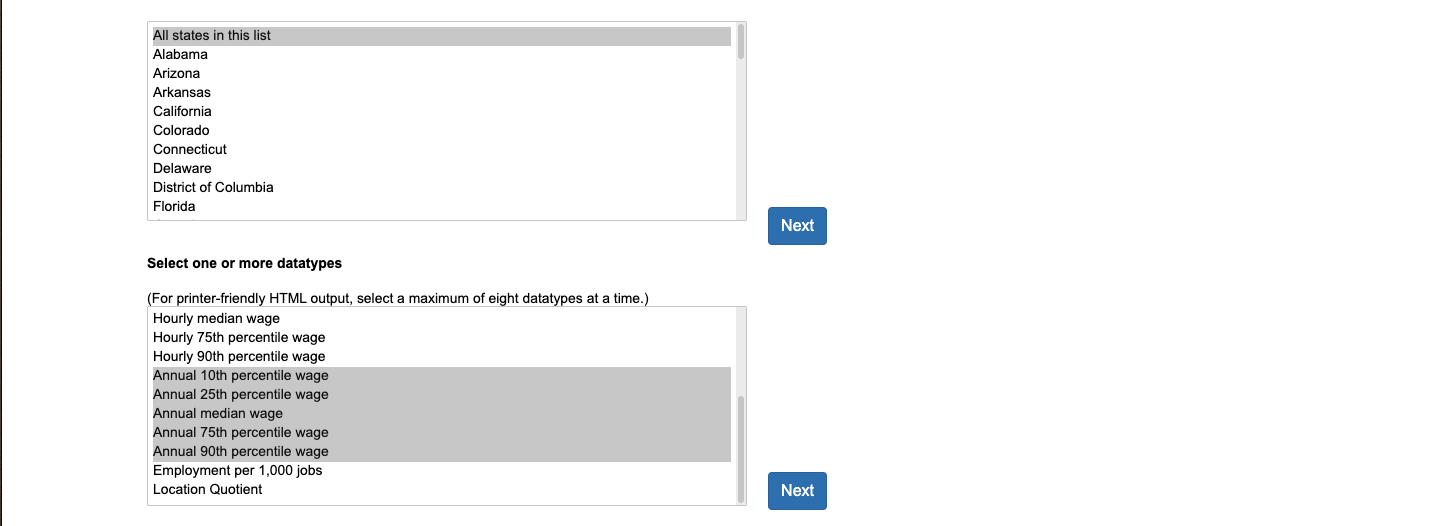


1. For geographical type, Select the option **“State”**, then choose “**All states in this list**” and click “**Next”**.

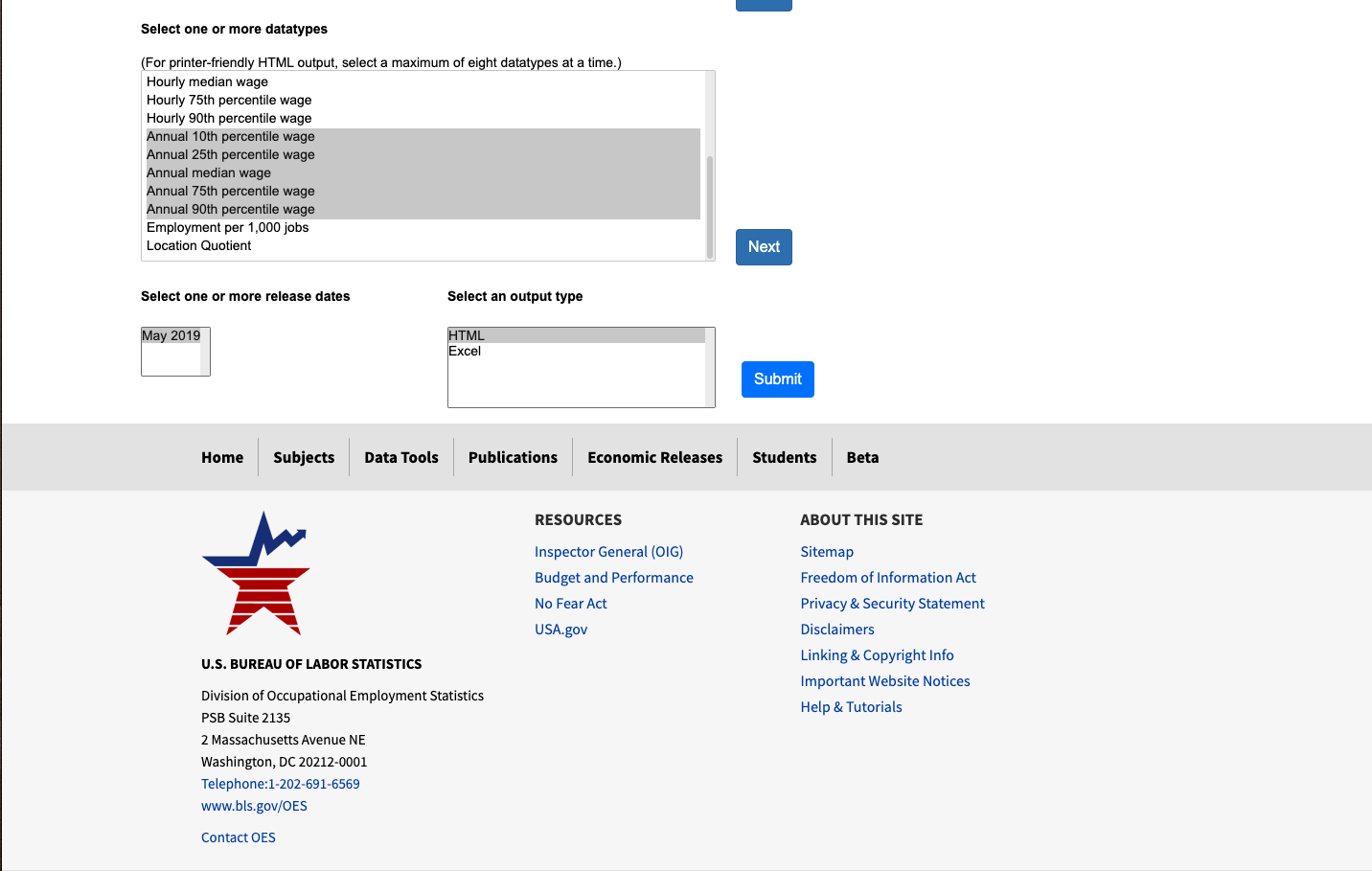


1. From the list of datatypes that pops up, highlight all 5 of the following, then click the “Next” button **\*Make sure you choose ANNUAL wages and not HOURLY wages\*.** To highlight, you may click first on Annual 10th percentile wage, then hold shift and click on Annual 90th percentile wage. All 5 selections should be highlighted:

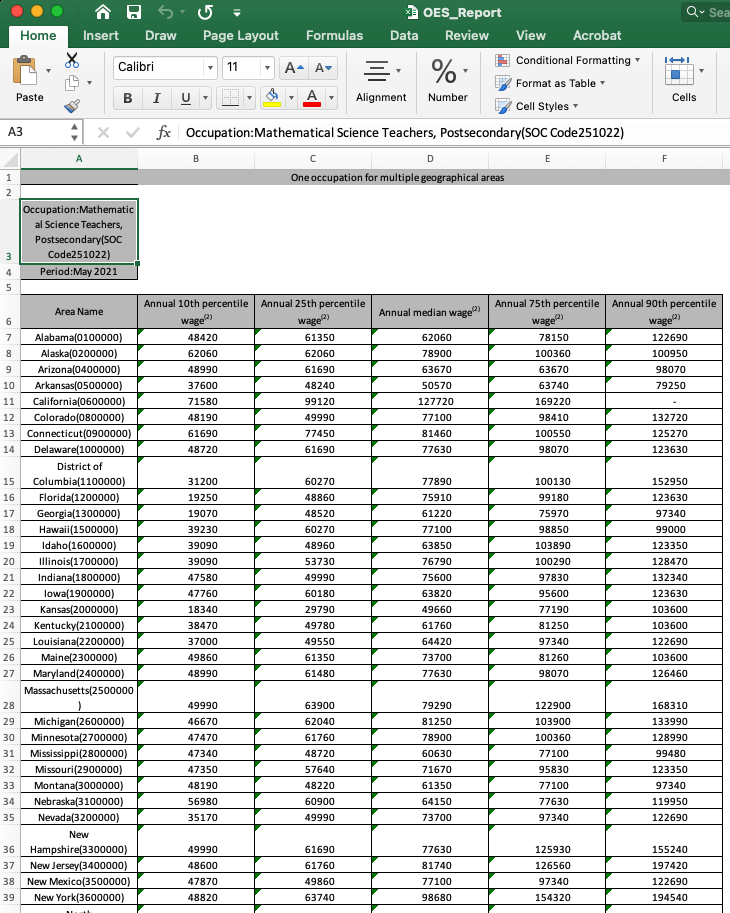
* Annual 10th percentile wage
* Annual 25th percentile wage
* Annual median wage
* Annual 75th percentile wage
* Annual 90th percentile wage



1. If more than one release date is listed, make sure the most recent data is highlighted, choose “**Excel**” for your Output Type, then Click “**Submit**”.



1. You will receive a file called “OES\_Report.xlsx”. It should look something like this:



# Part 4: Interpreting Salary Information from the Bureau of Labor Statistics

Remember that these are *percentiles*, and not raw data. Consider the following data for Industrial Engineers in Alabama (this is just an example — do not use these numbers when you complete Part 5!):

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Area Name | Annual 10th percentile wage | Annual 25th percentile wage | Annual median wage | Annual 75th percentile wage | Annual 90th percentile wage |
| Alabama | 64840 | 79180 | 97360 | 109210 | 133940 |

These numbers serve the same purpose as a traditional 5-number summary. They are *slightly* different, however.

**Answer the reflection questions below to investigate the differences:**

1. What makes the Bureau of Labor Statistics’ version of the “5-Number Summary” different from the version we learned in class?

The BLS version starts at 10th percentile Instead of 0 for first quartile and ends at 90th percentile Instead of the maximum for fourth quartile.

1. Notice that for this data, the Bureau of Labor Statistics reports median salary, rather that mean salary. Why might median be a better choice than mean for reporting “typical numbers” for salary data?

It might be more accurate as It represents that 50% of the people earn less than this value and the other 50% earn more than this value. The mean might not be accurate as there would be outliers Influencing the value where a person may earn a lot less and there could be people earning a lot more than typical.

1. Why might the Bureau of Labor Statistics version be *even better* than a traditional 5-Number Summary when working with salary data?

By starting at the 10th percentile and ending at the 90th percentile, It Ignores the bottom 10% and the upper 10% of the salary data which may be significantly different and skew the data. By representing what 90% of what people typically earn, this data can more accurately represent the salary distribution.

# Part 5: Comparing Your Chosen Profession to National Averages:

According to the U.S. Social Security Agency[[1]](#footnote-1), **the average annual salary in the U.S. is $63,932**.

I couldn’t find a reputable citation for the standard deviation for all salaries in the U.S. For the purposes of this assignment, we’re going to make up a number. For this assignment, we pretend that the **standard deviation for salary in the U.S. is $13,246.**

**Respond to the Part 5 prompts listed below:**

1. Based on the mean and standard deviation figures provided above, as well as the information you gathered in Part 3, complete the following table.

For your z-score calculations, **show your work**.

|  |  |
| --- | --- |
| Based on the file you downloaded in Part 3, list the salary values (for your chosen profession) **for *Illinois* only**: | For each of the Illinois salarypercentiles in the corresponding row, calculate their ***z-scores***, relative to the mean and standard deviation figures provided above**:** |
| Annual 10% = $59,010 | **z =** -0.37 |
| Annual 25% = $75,050 | **z =** 0.84 |
| Annual 50% = $96,640 | **z =** 2.47 |
| Annual 75% = $124,750 | **z =** 4.59 |
| Annual 90% = $145,300 | **z =** 6.14 |

1. How did the salaries from the career you chose compare to national averages? Justify your answer using the five-number summary and z-scores above.  
     
   For the bottom 10% of the employees, the salary Is less than the average US salary by 0.37 standard deviations. For 25% of the employees, It Is higher than the average by 0.84 standard deviations. With a median salary of 96,640, It Is much higher than the average salary with a standard deviation of 2.47. This means 50% of people have a salary of more or less than 96k$. For the 75 percentiles, It Is 4.59 standard deviations more than the average. Which suggests 25% of salary are more than 124,750. And the upper 10% people have a salary of 145300 or more, with 6 standard deviations away from the national average.
2. Do the estimates you made way back in Part 2 appear to reflect actual salaries for your chosen profession? Justify your answer based on the data you’ve obtained.

My estimates came close to the real values especially the on the upper end of the salary. However, I was wrong about the lower percentile of values as where I said 45k$ per year but the first quartile of the data suggests It Is 59,010$. My gut feeling of 75k$ per year Is also undervalued because the median salary Is 96k$ per year according to the BLS.

1. <https://www.ssa.gov/oact/cola/central.html> [↑](#footnote-ref-1)