A.

1. The research questions I wanted to answer were what are the characteristic of the salary of those in the data science field? Are there salary benefits to being a contractor as opposed to a freelancer? Do executives consistently make more money than senior level employees? And how much more do mid level employees make compared to entry level?

2. The scope of this work will include an initial .csv file from Kaggle that will be downloaded into an Excel workbook. It will be broken down into individual spreadsheets within that workbook for the comparisons listed above. The scope will only be for recent salaries, from 2020-2023, and will not provide a deeper analysis in the sense of historical salaries.

3. Overview of my solution - This analysis will utilize a data set from Kaggle.com listing salaries across job categories and employee type categories. I will be using Microsoft Excel for the data analysis. Specifically the data analysis toolpack t-test models. I will compare the mean salaries and standard deviations of contractors vs freelancers. I will do the same comparison for entry level employees vs mid level employees and mid level vs senior level employees. Lastly I will do the comparison for executives vs senior level empolyees. What will be delivered will be scatter plots showing salary ranges for the employee types listed. Mean salaries and their standard deviations for each employee type listed, and some T-test results comparing two of the employee types listed.

B. The project plan happened according to the plan outlined in Task 2 part B. I was able to download the .csv from Kaggle and clean it. I was then able to download the data analysis tool pack for Excel and use it for my statistical analysis.

C. The data selection and collection process happened according to plan. I went to Kaggle.com and searched for available datasets that were in a .csv format. The data science salary dataset was interesting to me, and I was able to download it for free. Because Kaggle is a site with free available data, there were no data governance issues to deal with. An advantage that my dataset had that I was able to use is all salaries were converted in USD for easier comparison. One disadvantage is that for a few job categories like freelancers and contractors, the amount of data points was limited.

D. There was no missing data points for any row, so I didn't have to clean it. For processing the data, I decided to create a new tab in my Excel workbook for each job category that I had namely, "Entry level", "Mid level", "Senior level", "Executive level", "contractor", and "freelancer". Then I could pull the relevant information from those individual tabs to new tabs that I would use for comparison between categories. Because I was comparing salaries, I only needed to use the USD equivalent and the job category columns for my comparisons.

E.

1. The methods I used to analyze the data were to find the average salary of each group, the standard deviation for each group. I did that with the "AVERAGE" and "STDEV.S" functions built into Excel. I also used the built in scatter plots for each job category. Later I used the Excel "Data Analysis" add on tool pack.
2. The advantage of using Microsoft Excel is that the functions used are built in and I could just push in data for the functions to work. No programming knowledge was needed to do the analysis. A disadvantage of using Excel is that with a specific developed program for this

analysis, later data could simply be pushed into it and analysis could be done quicker. With Excel, any new data will have to be processed manually.

3. In order to do the comparisons I wanted to do, I used a T-test. However, to begin I first had to decide on which T-test to use. There are two in the tool pack. One was a T-test assuming equal variance, and one was assuming unequal variance. In order to know which T-test to use, I first used an F-test. An F-test result will show if you have equal variance or not. Once that was understood, the correct T-test could be employed. The subsets of the data were also of varying lengths, so each salary range was sorted smallest the greatest and the middle section of each data subset was used in the comparison. This ensured an equal number of comparisons, and also got rid of outliers in the data.

F.

1. The F-test results in each comparison showed an unequal variance in my data samples, so a T-test assuming unequal variance was also used for better accuracy. From there I was able to view the P values to see if they were above or below the industry standard of .05. If there were below, then we could fail to reject the null hypothesis and the mean salary values could be trusted as accurate. If they were above, then we could reject the null hypothesis and the mean values could vary from what was calculated.

2. The practical significance is really for those interested in entering the data science job field. This data is relevant salary ranges as it dates back only to 2020. It would also be a good metric for those who are in the data science field and want to get an idea of how their salary might change and they advance over time. This data set spans between small, medium, and large businesses, so it is applicable industry-wide. It would also be a good data set to view if someone was interested in knowing what to expect as a contractor or freelance worker rather than becoming an employee of a company.

3. From task 2 I stated that the measure of success was to uncover insights on salary expectations with a clean set of data. My data is not missing any data points. The second measure of success was to clearly show statistics of mean, standard deviation, T-test, and scatter plots of the job categories listed. I was able to do both of these things and consider the project a success. I also find the results to be an effective display of the statistical analysis done.

G.

1. Obviously moving up the corporate ladder is going to increase salary, but my analysis shows the biggest jump to be from a mid level employee to a senior level employee. The smallest jump is entry level to mid level. The analysis also shows that being a contractor or freelancer has a lot less stability in salary than being part of a company.

2. A T-test is used to compare the means of two groups. So using it to compare the mean salaries of two job categories is an effective method. Using the F-test was an effective method to choose which T-test to use. The scatter plots are an effective visualization because scatter plots are good at showing all the data points, including outliers, as well as effectively showing trends.

3. The first course of action based on my findings is to have this data available to those who are finishing their education and starting to see what jobs might be available in the data science field, or for those who are currently in another job field and are curious about data science. This analysis can provide them with a helpful tool to decide which course of career they should choose.

The second course of action I would suggest is to provide this data to companies currently operating in the data science job field. They could use these data points and analysis to try and set their base salaries for employees and maybe create more attractive job offers to recruit the best talent to their company.

H. https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=cacf1f8d-5612-473b-8846-b03c0107e562

I . My source data comes from https://www.kaggle.com/datasets/arnabchaki/data-science-salaries-2023. It is also the first tab in my Excel workbook.

J. Sources:

1. *Data Science Salaries 2023* 💸. (2023, April 13). Kaggle.

https://www.kaggle.com/datasets/arnabchaki/data-science-salaries-2023