**MATH 243 - Project One**

[00:00:01.13] SPEAKER: Hi, class. In this video, I will discuss the first project in Math 243. We will go over the project template, as well as the Python script that you need to run to get your statistics. Let's first see how we can access the template. So once you're in Brightspace, in module 3, if you scroll down all the way, under project one submission, there's a link for requirements and rubric for the project.

[00:00:32.49] So if I click on this link and I scroll down, I will see the link to the template here. Summary report or document. If I click this, this will open up the template. So this is your project template that you're going to be working with. The template here includes questions that you will need to answer. So the first thing to do here is to just go through the template and see what type of questions are included here.

[00:01:03.46] So for example, here, you have to provide an introduction to the project. So after you've read about the data set and your Python script, then you have to answer some of these questions here. For example, what is a problem that you're going to solve? What data set are you using? And what statistical methods you will be using to do the analysis?

[00:01:29.16] Then you have the second section, where you have to pick your team and the assigned team. So in this project, you're going to be working with a basketball data set. And so you have to pick a team of your choice from a list of the NBA teams. And then you will also be given an assigned team. So the assigned team is already assigned to you.

[00:01:58.06] And the idea is that the team that you pick, we're going to do some analysis, where we will compare the performance of your team to the team that is assigned. And so you have to answer some of these questions. Be sure to follow these comments that are highlighted. So you have to answer the questions in a paragraph response. Remove all questions and this note before you submit the project. Do not include Python code in the report.

[00:02:32.62] So in this template, when you're writing out your summary report, please be sure to remove all of these questions, as well as these comments that are highlighted. And do not include any Python code in your report because you will be submitting the code as a second submission. So you will submit your summary report, as well as your Python code in HTML format. More on that later.

[00:03:02.10] OK. So once you answer the questions here and you go to the third step, which is data visualization, point scored by your team. So here, the question is specifically about performing analysis. And you can see, there's a comment here that says see step three in Python script to address the following items. And there's a similar comment here. See steps one and two address the following items.

[00:03:30.25] So what this means here is that once you're at this point, you should open your Python script, and start reading through it, and follow this the steps that are provided here. So for example, to answer these questions here in the introduction, your team and assigned team, you should follow steps 1 and 2 in your Python script. So be sure to run those steps. Read the instructions in those steps. And then answer these questions.

[00:04:06.61] At this point, we can open up the Python script now. So let's go through the script, at least four steps one, two, and three, and see how we can enter these questions. So in the Python script here, once you open it, you're going to see this audio screen. On the left side here, you have the files for the project. And the file that you're going to open is project one Jupyter script.

[00:04:37.28] So if I click on this, this will open up the Python script in a Jupyter notebook. So be sure to read through these instructions here, and there's information about the data set as well. So you can see here, we're provided with the variables that are included in the data set and also an explanation of what the variable represents.

[00:05:12.01] So for example, pts variable point is point scored by the team in a game. Elo underscore n is a measure of relative skill of the team in the league. And then you have three other variables. Now, here, you can clearly see step one is labeled here. In step one, you're going to do data preparation and the assigned team.

[00:05:38.63] Now in the first step, you don't have to make any edits you just have to run this step. And how do you run this step? So I will click on this cell here. And you can see it gets highlighted. And then I will click on run button here. And this will run these set of Python commands.

[00:06:03.74] And essentially, what we're doing here is we're just reading the data set from this NBA data file, dot CSV data file. You'll see that this file is here as well. So this is the CSV file that contains the data set. So we're reading it using a module called pandas that is capable of reading CSV files.

[00:06:26.09] And then we're kind of filtering the data little bit. And we can see that the assigned team here is already provided to you. So you don't have to make any edits here. The assigned team is a team that played in 1996 through 1998. So you can see that instruction is here. And the franchise ID for the team is Bulls. So it's the Chicago Bulls of 1996 through 1998. That's the team that you're going to compare your team against.

[00:07:09.00] So when I ran this step, I got this printout. So what we're doing here is we're just printing the first five rows of the data. So there are obviously many rows in the data set. You can see number of rows in the data set of 246. We're just printing the first five to get a visual of what the data looks like. And this is what it looks like. So you have a game ID, year ID. Franchise ID is bold because we're printing the assigned team data set.

[00:07:39.05] We can see how many points they scored, points by the opponents. This is their skill level. This is the skill level of the opponent. Game location was home. H is for home, and A is for away. And the result for the game was that the Bulls won. So this is what the data looks like.

[00:08:02.51] Now let's go to the second step, where you have to pick your team. And because you have to pick your team, you have to edit the script. And so be sure to read the instructions first. So in this step, you will pick your team. Make the following edits to the code. So replace this with your choice of team from one of the following teams. Remember to enter the team name within single quotes. For example, if you pick the Suns, then this should be replaced with single quote, Suns, single quote.

[00:08:43.03] So suppose we pick the Suns. So let's try to find this here, in this piece of code. So we can see that this is right here. So this is what we're going to replace. So I'm going to delete this first, then I will have a single quote. And Python already automatically puts the ending single quote. So I'm good there. And I'm going to type Suns. And I'm going to type it just like it's shown here. So whichever team you pick, be sure to type it exactly like the way it's shown here.

[00:09:20.15] So single quote, Suns, single quote. I have made my edit. This is the team I picked. Now I'm going to run this step. And again, I will highlight the cell here and run. And here's the data set for the Suns. And you'll notice that here, we're comparing the years when your team played, are 2013 and 2015. So we're comparing the Suns from 2013 through 2015 with the Bulls from 1996 to 1998. So it's kind of like we're comparing it to the team in which Michael Jordan played. So we're trying to see how the performance of your team compares to the performance of Chicago Bulls from those years.

[00:10:14.11] So now that you ran the first two steps in your Python script, you can answer the questions in this section of the template. What team did you pick, and what years were picked to do the analyzes? What team and range of years were you assigned for the comparative study? And please be sure to discuss it in a paragraph, in a short paragraph, and also fill out a table like this, clearly showing the team that you've picked and the team that was assigned.

[00:10:46.65] So now, in the third step here in the Python script, you have to answer these questions. And you can clearly see that the step involves data visualization or creating a plot. So let's look at step three in the Python script. So step three is data visualization, point scored by your team. And this matches data visualization point scored by your team. This matches the section here in the template.

[00:11:16.24] So here, the coach has requested that you provide a visual that shows distribution of points scored. So through these three years, however points your team scored in each game, we want to get a distribution of those points and see how the team was performing in those years in terms of points that they were scoring.

[00:11:46.74] So let's create this. So what's going to happen here, when I run this step, you're going to get a histogram, and you're also going to get a scatterplot. And the question that is asked is to pick between one of these two plots. So whichever plot you think is most appropriate to answer the question about the distribution of points scored. So let's print this first.

[00:12:17.59] So here are the two plots. Now sometimes, when you run this step, the plots may not come up the first time you run it. So if that happens, just rerun the step again. And the second time, the plots should print. So histogram of point score. And here's a scatterplot of points scored. So which one do you think makes more sense to get an understanding of the distribution of points scored?

[00:12:50.28] So usually, when the distribution is discussed, the histogram is always a better choice over any other plot. The scatterplot will generally show trends. So when you have two variables and you want to see a trend between those two variables, you plot the two variables on y and x-axis, and you are able to see the trend using a scatterplot. Whereas for a distribution for any given variable, a histogram does the job.

[00:13:28.02] So now, when we go back to the template, the question is, in general, how is data visualization used to study data distributions and trends? So be sure to answer this question. There's information on this in zyBook. In this activity you were asked to pick one of the two plots. So which one did you pick? So be sure to say why you picked the plot that you picked. And also include the plot in your report.

[00:13:55.18] So do not just answer the question. You should also include this plot in your report. And the easiest way that I found to include it is right click, and you can copy image, and then you can paste the image in your report. You can also save this image and then load it into your report. So you can do these two things. But usually, copy image and pasting it works for me. OK. And then why did you pick this plot and explain. So this is the third step.

[00:14:33.94] Now the fourth step is another data visualization, point scored by the assigned team. So now, we want to look at points scored by the assigned team. So if I run this step, again, I get this histogram and the scatterplot. And again, the question is about the distribution of points scored by the assigned team. And therefore, histogram again is preferred over a scatterplot. So you should include the histogram here and also discuss why you think this is it a better choice than a scatterplot.

[00:15:18.68] Now we go to step five. In step five, now the coach wants you to prepare one plot that provides a visual of the differences in the distribution of points scored by the assigned team and your team. So we want the distribution of points scored for both your team and the assigned team to be on one plot. And that way, you can compare the distribution. And we're going to use a side by side box plot, as well as a histogram for it. So let's run this step.

[00:15:55.89] And so here it is. You can clearly see here, the blue is the assigned team, and your team is orange here. You can see that the two histograms are overlaid. And so you can kind of compare the distribution of points scored. One distribution may be shifted to the left of the other. And you should explain why that is.

[00:16:21.50] Similarly, you can also see that the side by side box plots can be used to do this comparison, right? This also shows how the distribution of points scored is for your team, as well as the assigned team. So you can do the same comparison here. So in this scenario, you have to pick one of these two and then just justify why you picked the plot that you picked and how it explains the difference in the distribution of points scored for the two teams.

[00:16:57.72] Let's now go to the next section in the template. Section 6 now is about descriptive statistics. Points scored by your team in home games. So the games that your team played at home, we're going to do some descriptive statistics on it. So in the Python script, if you scroll down, the management of your team wants you to run descriptive statistics on the points scored. And some of these statistics are mean, median, variance, standard deviation for points scored by your team played at home.

[00:17:38.41] So these are the games that your team played at home. We want to look at the average median, variance, and standard deviation of the points that your team scored while playing those games at home. And make the following edit. So here are the edits that you have to make. So all of these instructions in red will tell you specifically what you need to replace to get the code to run. If you run the code as is, it's going to give an error because you have not made the appropriate edits.

[00:18:12.02] So replace this with the name of Python function that calculates the mean. So here it is. And so we're going to replace this, and the functions are not going to be in single quotes. So we're going to type them as is. And the function that calculates the mean is just the mean. The function that calculates the median is simply called the median. For variances, var. And for standard deviation, that's std.

[00:18:48.25] So I made the replacement here for all of these four instructions. And so the code should now run. So if I run it, as you can see, it's calculating the mean, median, variance, standard deviation and then printing it here. So here are the stats. Points scored by your team in home games. So the mean is this. The median is this. The variance is this. And the standard deviation is this.

[00:19:22.56] Now here, first, be sure to read through these questions and answer each one of them. So you're going to provide these statistics in a table like that. Then answer, in general, how our measures of central tendency and variability used to analyze a data distribution. So explain how descriptive statistics can be used to analyze data. You should interpret each statistic in detail and explain what explain what it represents in the scenario.

[00:19:54.94] So each of these statistics, for example, mean and the median, are for the central tendency of the data, right? So you should explain what that means and what it represents. So here, you can see mean and the median are almost the same. And that points to a relatively bell-shaped distribution for the points scored by your team in the home game.

[00:20:22.56] And we kind of saw that right when we printed this histogram for points scored by your team. Although these were not just at home. This also included away games. But you can kind of see that the distribution seems to be bell-shaped. And if you only print the distribution for the home games, it should look bell-shaped, based on the mean and the median. And that's because when you have outliers in your data, the mean will be shifted in the direction of those outliers, whereas median is fairly resistant to that shift. And so you will see a difference between the mean and the median.

[00:21:07.63] Standard deviation and the variance, they are a measure of variation in your data. So how much the data values vary in the data set? That's what's measured by the standard deviation. And then you can also, if you have a bell-shaped distribution, then standard deviation can be used to get an approximate range of the data. And so you should discuss this And this is in your zyBook. So be sure to discuss and interpret each of those statistics this way.

[00:21:49.52] Now use the mean and the median to describe the distribution of points scored by your team in home games. So that is what I just discussed. You should describe whether there is a skew, left, right, or bell-shaped, and explain which measure of central tendency is best to use. So if you have a skewed distribution where the mean is significantly different than the median, then which of the mean or the median should you use to represent the center for the data?

[00:22:20.52] The next section is, again, descriptive statistics, but this time, it's point scored by your team in away games. So we're going to look at the games that were played at opponents menu. We're going to see how the descriptive stats for points scored by our team look for those. So again, this is fairly straightforward step seven. However, in this step, you are to write the code block yourself. So you have to write the code yourself here.

[00:22:55.77] Now the code is going to look very similar to the previous step where we calculated the descriptive stats for home games. And there are instructions provided here. So since you are calculating statistics for games played at opponent's venue, game location variables should be set to A, right? Because here, game location variable was set to H for home games. Well, now it should be A, so that we only get away games. So that's critical.

[00:23:31.29] Then functions for all statistics are the same as those in step six. So the functions mean, median, variance, standard deviation that I edited are all going to remain the same. Your statistics should be rounded to two decimal places. So now, to write this code block, the best thing to do here is to simply just copy this. Let's copy everything here. And then paste it.

[00:23:58.76] So I'm highlighted, right click copy. And then paste. And I'm going to say print points scored by your team in away games here. This is the header that gets printed. Now here, game location is not home. It's away. So I'm going to type A here. That's what the first instruction said here. And everything else remains the same because I'm calculating the mean, median, variance, standard deviation. Nothing else needs to change. You can see it was that simple to get the statistics for away games.

[00:24:41.60] So it's simply just copying over the previous step, changing game location to away, and then getting the stats. So let's run this step now. I'm going to highlight it, run it. And here are the descriptive stats for points scored by your team in away games. So again, you have the four statistics. And now, make sure, again, you interpret each of these stats and answer the questions that are included in the template.

[00:25:17.84] And by the way, there's an additional question in this step, which is is your team performing better in games played at home than those that are away? So make sure this is kind of comparing steps six and seven. So you need to look at your descriptive statistics for home games and away games and then conclude whether you think the team is performing better at away or home games and why.

[00:25:43.24] Now when it comes to discussing whether the team is doing better at home or away, not only in terms of the average or the median, but also in terms of the standard deviation and the variance. So are you seeing that the team is consistent playing at home? If it is, then perhaps the standard deviation for home games is lower than away games. So those are the things that you need to discuss both in terms of the variation, as well as in terms of the central tendency or the center of the data set.

[00:26:21.22] Now step 8 is about confidence intervals for the average relative scale of all team. The first thing to remember is that confidence intervals are always for an average. So we won't say that confidence intervals are for relative scale. It's always the average relative scale. So there's an average, and then there's a confidence interval around that average. There's lower limit and an upper limit for the interval.

[00:26:54.04] So first, you have to describe how confidence intervals are generally used in estimating the measures of central tendency. Then you will run the step. So this is step eight. So in this step, you have to do some edits before you can run the code. So for example, replace SD variable with the variable name representing standard deviation of relative scale of all teams.

[00:27:23.67] So let's go through this code here. Here, we're calculating the average for relative scale. Elo underscore n is relative scale. And this is for all the teams. Confidence interval for average relative skill in the year 2013 to 2015 for all teams. Then we're calculating the standard deviation of the relative scale. So that's what this calculates. So we have the mean and the standard deviation.

[00:28:02.19] And then we're calculating the sample size, which is n. So we have the means standard deviation and n. And here's the first edit that we need to make. Here we have to replace this with the variable name representing the standard deviation. So we just calculated that here. So it would be stdev. So I will replace this with stdev.

[00:28:32.91] And this is the formula for standard error. Standard error is just a standard deviation divided by square root of the sample size. And so that's what we're doing here. Now replace the CL with the confidence level of the confidence interval. So the confidence level that is asked is 95%. So for CL, I will include 0.95. That gives us the 95% confidence interval. Replace the mean variable with the variable name representing the mean, which is we calculated here. Mean is this. So I will replace this with just the mean.

[00:29:20.09] So this variable name will come here. And now this is the variable name representing the standard error. Standard error we calculated right here using the standard deviation. So I will copy this. And I will paste it here. And this concludes all of the edits that I need to make here. Once I have this and I run this step, I will get the confidence interval. The lower and the upper limit of the interval. I will get both the unrounded and the rounded. In your report, be sure to only use the rounded confidence interval. So let's run this.

[00:30:10.07] So here we go. We get the confidence interval. The rounded is, the lower limit is 1502, and the upper limit is 1507. This is the confidence interval. Now provide a detailed interpretation of the interval in terms of average relative scale of teams in the years that you picked. Now when you're interpreting this, be sure to provide the correct interpretation.

[00:30:40.53] So we're going to say that we're 95% confident that the average relative skill for the years 2013 and 2015 is between 1502.02 and 1507.18. And then extend on this interpretation and explain what it actually means. How would your interval be different if you had used a different confidence level? So maybe explain what happens if we use a 99% confidence level or a 90% confidence level. And then what is the probability that a given team in the league has a relative skill less than that of the team you picked? And that was printed right here in this piece of code. This is right here.

[00:31:27.32] OK. Now section nine is, again, a confidence interval. This time, it's for all the teams in the assigned team year. So 1996 to 1998. And this code block you have to write yourself. So the easiest thing to do here is, again, to copy everything here in this step. So let's copy everything here and paste it here.

[00:32:00.99] And the first thing to do here is confidence interval for average relative skill in the years 1996 to 1998 because that's what we're printing here. The data frame for these years is called assigned years league df. So we were using your years leagues df. That was the data frame for the teams that played in 2013 to 2015. So we're going to replace all of these with assigned years leagues df. So I'm going to do all of these replacements here. Wherever we used that name, we're going to update it to this name.

[00:32:53.13] OK. Now the variable elo underscore n represents the relative scale. So that we're going to keep the same. Start by calculating the mean and the standard deviation. So we already did that. Calculate n. That's already here. Calculate the standard error. So we already are calculating that. And then the interval is being calculated here. In the print statement, be sure to update the years. So the printout is correct.

[00:33:34.22] And probability. A team has average of skill less than average or the skill of your team in 1996. Actually, you don't really need to print this. Let's see. Assuming that the population standard deviation known. Use Python methods for normal distribution to calculate the interval. Your statistics should be rounded to two decimal places.

[00:34:06.33] So use step eight to help you write this code block. Here is some additional information. So calculate the mean relative scale of the Bulls. So yes, you do need to update this. So the mean relative skill for the Bulls is not going to be in your team df because your team is the one you picked. It's going to be the assigned team df. So we're going to take this. And update this year. I'm going to paste it here.

[00:34:48.15] OK. So now, assigned team df elo and dot mean we'll give you the mean elo for the assigned team . And then here, we will update mean elo for the assigned team and mean elo for the assigned team. OK. Probably, a team has every relative skill less than the average relative skill of Bulls in the years 1996. And this should complete all of the edits.

[00:35:29.79] And so if I hit run, this will give you the printout. The 95% confidence interval is here, and the probabilities are here. So again, the questions that you're going to answer are similar to step eight. You need to first specify the interval here and then provide an interpretation of that interval. And then how does this confidence interval compare with the previous one?

[00:36:01.73] So remember, the confidence interval that we calculated here is for the relative skill of all teams in years 1996 to 1998. In the previous step, we had a confidence interval for average relative scale of all teams in 2013 and 2015. So we can use the intervals to compare how the NBA teams overall are performing in 2013 to 2015 compared to 1996 to 1998. And that is what is being asked here.

[00:36:40.15] So you need to do a little bit of discussion around how you can use these intervals to decide whether the teams are performing better in 2013 and 2015 compared to back when the Bulls used to play. So half the teams actually started performing much better in terms of their skill level. So that's what you're going to discuss here.

[00:37:13.36] Then after that, this will conclude your Python script. And after this, you will provide a conclusion. So in the conclusion, what is the practical importance of the analysis that you performed? And describe what these results mean for the scenario. So discuss the results from an overall perspective. Discuss what analysis or what analyses was important? Which one did you think was easiest to follow? And how these statistics and the visuals can help describe a story about the data set and explain the patterns and the trends. And then be sure to include any citations if you are including any text from a source.

[00:38:10.12] So this concludes the project. So after you've done your summary report, you'll submit the summary report, as well as your Python script here in HTML format. And once you have completed the script, you'll go to file, download as, and then HTML. This will download an HTML file that you will then submit along with your summary report in Brightspace.

[00:38:38.66] Now before we conclude, I also wanted to discuss a couple of instances where you may have issues running the Python script. Maybe there's an error in the code. Maybe you mistyped something that you were editing, and then sometimes, it becomes hard to figure out what's going on. So I wanted to discuss some debugging solutions that may be useful to you.

[00:39:10.38] The first thing is when you run your Python script. So say that you start running the script from step one. And let's say you run the first four steps, and then you stop, and you want to continue at a later time. So say you want to continue the script to the next day or the following day. And when you do that, when you get out of this script and you come back to it at a later time, you can not simply just continue from step five. You have to run all of the previous steps.

[00:39:44.43] So at a minimum, even if you have edited your script, you have to run the previous steps. So by running, what I mean is you have to highlight the step and hit run. And you have to do that from step one. And that's because Codio does not save the data in its memory. So the data needs to be read, and the variable assignments in Python code here, these lines need to be run again, so that you can continue the script.

[00:40:19.75] So in other words, just be sure to start from step one, start running the script from step one every time you continue at a later time. The second scenario is suppose that I'm working say in step two. And in step two, I have to type the team name here. So suppose that I make this edit, but while doing so, I accidentally delete the equal sign, the additional equal sign that was there. And I type Suns.

[00:41:01.09] Now if I run this, I'm going to get an error. It's going to say it is an invalid syntax. And it's going to point to where the error is. And let's say that I am not sure how to fix it, right? So what I can do here is you have also been provided a backup Project One Jupyter script. So this is a backup.

[00:41:26.87] So if I click on this, it will open this tab here. It's a second tab for the backup project. And if I scroll down to step two, I'll see that this is where I was getting the error, right? And I can see that there should be an extra equal sign. So there are two ways I can fix it. This is my script here. So the first thing is I can just simply type what I'm missing here. So I can just type the equal sign and run it. And this will run. Well, I need to run the previous step first. And then this will run.

[00:42:11.92] Now suppose I don't do that, and I still have one equal sign. So it still gives me that invalid syntax error. Another way I can fix this is I don't have to worry about what error and where it is. I can just highlight all of this from the backup script. So every Python line in step two, I can highlight, copy. And then go to the project. And then highlight all of it. Right click and paste.

[00:42:47.39] So this brings my script, almost like resets it. It brings it to the original state. And now I can make this edit again. So I can type Suns. And now I get the correct output. So there are two ways to fix it. If you know where the errors are, you can simply look at the backup project and fix those errors in your script. Or you can just sort of reset it by copying the entire step.

[00:43:26.12] And the last issue that I've seen over the years is on these plots. So sometimes, when you run the plot, although it's printing it now, sometimes, you may have instances where running this step will not output any plots. If that happens, simply just rerun it. And it should print. And that should resolve this issue.