**A step-by-step approach to your project**

Your AE 102 term project is not a complex project, but it is complex enough that you may find it overwhelming if you don’t plan it out. I present below a step-by-step approach that you may follow to complete your project.

**Step-1: Choose your project topic.** The first step is to choose a topic for your project. There are no restrictions on the choice of topic apart from the basic requirement that you will be able to collect a good amount of numerical data—this is after all a data analysis project! Choose something that you are passionate about. The reason is two-fold: (i) you will have fun doing it, and (ii) in case you are stuck—this is very likely to happen—your passion for the topic will give you the necessary motivation to persist and push through.

**Step-2: Identify two or more quantities of interest.** There are a lot of things you can do once you have chosen a particular topic. To make the analysis concrete, you first need to identify two or more measurable quantities—random variables, in short—that are of interest to you.

**Step-3: Understand the scope of your project.** The scope of AE 102 term projects is quite limited: you will be performing a statistical analysis of the chosen quantities using various techniques that you will learn in class–nothing more, nothing less! This would entail (i) inferring the statistical distribution of each random variable, (ii) understanding the relationship between the random variables, and (iii) formulating and testing hypotheses involving the random variables that provide some insight into your project topic. That’s it!

*Note:* You will *not* be doing any predictive modeling in this project! The reason is simply that we won’t be studying the relevant ideas in class. You can always do a follow-up course on machine learning and do your predictions as part of that. AE 102 is fundamentally about statistical characterization of given data—your project will accordingly be about just that.

**Step-4: Collect data.** Now that you have a topic in mind, and some measurable quantities that you are interested in studying, the next step is collecting the data. This is often the hardest part of the entire project! If you are lucky, you may find a dataset directly. In most real situations, your data will be scattered across multiple sources. Collecting the relevant data is not a straightforward task. There are no easy solutions here—you may find some web-scraping tools that will make the task easier, but that requires work too. Since you will be doing this project in teams of 3-4, my advice would be to split up the arduous task of collecting data among yourselves so that none of you find it overwhelming.

While you are collecting your data, make a note of the sources from which you obtain the data. It is important to choose sources that are unbiased and trustworthy—this may not always be straightforward! You are also expected to cite your sources in your report.

**Step-5: Organize data.** Once you have your data, you need to store it in a form that is appropriate for data analysis. The recommended data format for AE 102 projects is a simple table. Here is a schematic that illustrates tabular data:

|  |  |  |  |
| --- | --- | --- | --- |
| **Sample Label** | **Random Variable 1 (units)** | **Random Variable 2 (units)** | **Random Variable 3 (units)** |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

Think of your data as random samples drawn from the (often inaccessible!) population. Your goal in this project is to infer something about the population—in terms of your chosen random variables—using the data that you have collected.

*Note:* It is a good idea to choose your preferred units first, and then recast your collected data in these units. The best choice of units is one where the values of the random variables are neither too large nor too small. Please mention the units of each random variable in the table; if a particular random variable has no units, leave the units part blank.

To start with, you may use simple spreadsheet software like Microsoft Excel to store your data. Alternatively, a plain text file that stores the data in a Comma Separated Values (CSV) format is sufficient. Python has a very handy library to handle tabular data called Pandas, which you will eventually use to store your data. The last step in organizing your data is to store it as a Pandas D*ataFrame*.

**Step-6: Do a quick Exploratory Data Analysis.** Once you have cleaned up your data and organized it neatly in a Pandas DataFrame, it is a good idea to do a quick EDA to get an intuitive feel for your data. EDA typically involves plotting your random variables, getting a first picture of their distribution using a histogram, etc. It is also a good idea to plot pairs of random variables to understand correlations that might exist in your data. At this point, you don’t have to quantify anything. EDA is meant to provide you a quick qualitative feel for your data so that you can frame some tentative questions that you think are interesting. Later, when you do a systematic analysis of the data that you have collected, you will find that these questions will assume a more concrete shape and direct the specific choice of analysis.

**Step-7: Characterize each of your random variables separately.** Ok, it’s time to get to the details. To start with, you need to characterize each of your random variables individually. This will help you get a detailed understanding of the distribution of a random variable, its expected value, its variability, etc. What would be great is if you could *explain* their variability, but note that this may not always be possible. Keep this issue in mind when you study the relationship between different random variables in the next step.

Once you finish the basic characterization, you can do the more interesting task of estimating true parameters in the actual system. Recall that your data is just a sample drawn from the population, *not* the entire population! The goal of statistical estimation is to figure out what the ground truth of your population is based on quantities that you can compute using the finite sample that you have. You can now perform point and interval estimates of various parameters related to the distribution of the random variables.

**Step-8: Understand the relationship between each pair of random variables.** Since you have at least two random variables of interest, your next step is to study these pair-wise. The reason for doing a pair-wise study is simply that this is easier to perform and visualize—you can do more sophisticated analysis, but that’s outside the scope of AE102! This pair-wise study will give you some concrete insights that may help you understand hidden patterns and correlations in your data. If you’re lucky, you may also be able to explain the observed variability in some of your random variables.

As in the case of the single random variable analysis, your next task is to infer the population parameters using the point and interval estimates that you learnt in class. This will specifically involve understanding the correlations that exist between your random variables, and a regression analysis to uncover any linear trends that may exist in your data.

**Step-9: Formulate and test hypotheses on your data.** The foregoing characterization of the random variables individually and in pairs will give you a more precise understanding of your data, unlike the qualitative understanding that you got initially using EDA. Now that you have a good quantitative understanding of your data, you can formulate and investigate simple hypotheses about your system involving the random variables that you characterized earlier. Hypothesis testing will help you acquire some concrete data-based insights into your problem.

**Step-10: Write a report and make a presentation.** This is the part where you summarize your findings and systematically record that in the form of a report. I want you to keep in mind that communicating your work effectively is as important as doing good work in the first place. Unfortunately, there are no systematic courses at IITB that train you in this valuable skill-set. Most students see this task as boring. Fortunately, there is a structured approach that you can adopt that will make writing bearable, and perhaps even fun. Towards the end of the semester, I will be providing you with detailed style guides on how to write your report and create your presentation.