**Course 4 Task 1 Project Instructions**

**Task 1: Define a Data Science Process**

[INTRODUCTION](https://ut.daacertificate.com/mc/poa?productID=2652&taskID=3334#introduction)

**Plan of Attack**

Kathy has asked you to produce an initial report to IOT Analytics' clients. Your report should be in the form of a PowerPoint presentation that explains the Data Science process that you will follow during the analysis and also presents some initial insights *of business relevance* based on an initial look at the data.

**[1. Perform an initial exploration of the data](https://ut.daacertificate.com/mc/poa?productID=2652&taskID=3334" \l "collapsepoa3099)**

**Familiarize Yourself with the Domain and the Data**

1. Begin by coming up to speed on the domain by reading the article on [sub-metered housing](http://syracusecoe.org/gpe/images/allmedia/LivableNewYork/Sub-MeteringforElectricity.pdf) in the required resources section.
2. Read the provided [overview of the dataset](https://archive.ics.uci.edu/ml/datasets/Individual+household+electric+power+consumption).

Listen to Jaime Carbonell talk about the importance of the right data:

Expert Advice

Lots of the wrong data

**Load and Analyze the Energy Consumption Data**

There are several packages for R that fall into the category of data “wrangling” or “munging”, but there are two packages, DPLYR (Dee-Plyer) and TIDYR (Tidy-er) that really lead the way as far as capability goes. In order to fully work with your data you will need to understand how to load and use both packages and many of the features contained within each package.

1. In order for you become skilled with both of these packages you will need to work through the DPLYR and TIDYR tutorials listed in the required resources
2. Load the dataset into R
3. Use R and R Studio and the skills you learned in Chapter 14 of the *Predictive Analytic for Dummies*text and the DPLYR and TIDYR skills you have acquired to perform an initial exploration of the data in the [Energy Consumption Dataset](http://archive.ics.uci.edu/ml/datasets/Individual+household+electric+power+consumption)(in the required resources).

TIP:

When loading a dataset into R, you can load it directly using the *read.csv* function (the one you used in the tutorial), or you download the dataset to your computer and use the *import dataset* button in the top-right window of R Studio. Be sure to check the type of each attribute after you load the data to ensure that it is what you expect it to be before you begin processing the data. R provides a number of type conversion functions that you can apply as necessary (see, e.g., *as.numeric*). See the *Using R: common errors in table import* in the optional resources for answers to common import errors

Since the Date and Time columns are separate they will need to be combined within the dataset in order to convert them to the correct format to complete the appropriate analysis. Here is an example of how this can be done in R:

1. Creating one date+time column called DateTime (replace 'yourdataframe' with the name of your data frame):

yourdataframe <-cbind(yourdataframe,paste(yourdataframe$Date,yourdataframe$Time), stringsAsFactors=FALSE)  
colnames(yourdataframe)[10] <-"DateTime"  
yourdataframe <- yourdataframe[,c(ncol(yourdataframe), 1:(ncol(yourdataframe)-1))]  
head(yourdataframe)

1. Converting date and Time format (replace 'yourdataframe' with the name of your data frame):

yourdataframe$DateTime <- strptime(yourdataframe$DateTime, "%d/%m/%Y %H:%M:%S")  
yourdataframe$Date <- as.Date(HHPC$Date, "%d/%m/%Y")  
str(yourdataframe)

TIP:

In the above example note the ‘Y’ in the “%d%m%Y” is case sensitive!

TIP:

You might also use some of the functionality from DPLYR and TIDYR to perform the same task. Select (), Unite() and Mutate() are a few (of many) functions that might be helpful to you in performing this step.

Now that you have converted the date and time appropriately you can begin to obtain some information from the data set.

1. Produce descriptive statistics for the data
   1. Using the summary() command calculate the mean, mode, standard deviation, quartiles & characterization of the distribution and maybe more.
   2. Research basic statistics available in R and calculate others that may yield additional insights.

TIP:

While you can use the basic summary function from R to perform the following steps, you may also find the Summarise() and groupby() functions of DPLYR helpful to you as well; it would be helpful for you to try both so you can see the difference between the two methods.

Finally, you will use your analysis and insight gained from the data to provide a report of your findings and suggestions to the client.

1. Propose three management-level decisions you can suggest based on your initial exploration of the power consumption data.
   * Your mentor will assist you with this objective during your team meetings.

**[2. Define a Data Science process](https://ut.daacertificate.com/mc/poa?productID=2652&taskID=3334" \l "collapsepoa3100)**

**Define a Data Science Process**

Now that you have had a chance to examine the data you’ll be using for this work you’ll need to define a Data Science process that outlines exactly how you’ll be using the data.

Given a new project, your first step towards a successful analysis should be to select, tailor, and instantiate a process framework appropriate to your project.

Two of the most important factors in determining the success of an analysis are likely to be a clear definition of the goals of the analysis and exercising the discipline to follow the principled approach you have defined.

Considering the steps outlined below in either process (there are two alternatives so read both first) and in your readings, define the process that you will follow to thoroughly analyze the data found in the Energy Consumption Dataset. You will work with your mentor as needed to answer some of the questions. (References are provided in the Optional Resources section of this task.)

*Note that both of these process frameworks are iterative. A poor or unexpected outcome at any step might necessitate returning to previous steps. And if the problem is business critical, the process might be re-executed regularly.*

***Framework One - Zumel and Mount, Practical Data Science with R, chapter 1:***

Define the goal The first step in a data science process is to define a measurable and quantifiable goal.

* Why do the stakeholders want to do the project?
* What do they need from it?
* Why is their current solution inadequate?
* What resources do you need?
* How will the result of your project be deployed?

Collect and manage data This step includes identifying the data you need, then exploring and conditioning it. This is often the most time consuming step.

* What data is available?
* Will it help to solve the problem? Is it enough?
* Is the data quality good enough?

Build the model Here is where you try to extract useful insights from the data in order to achieve your goals.

* Which techniques might I apply to build the model?
* How many techniques should I apply?

Evaluate and critique the model Once you have derived a model, you need to determine whether it meets your goals. If not, it’s time to loop back to the modeling step.

* Is the model accurate enough to meet the stakeholders’ needs?
* Does it perform better than “the obvious guess” and any techniques being used currently?
* Do the results of the model make sense in the context of the real-world problem domain?

Present results and document Once you have a model that meets your criteria, you will present your results to your project sponsor and   other stakeholders.

* How should stakeholders interpret the model?
* How confident should they be in its predictions?
* When should they potentially overrule the model’s predictions?

Deploy and maintain the model Finally the model is put into But you still need to ensure that the model will run smoothly. In many cases this requires enhancement of the requirements based on customer feedback or in some cases fixing bugs.

* How is the model to be handed off to “production”?
* How often, and under which circumstances, should the model be revised?

***Framework Two -* BADIR** *(***Jain and Sharma, *Behind Every Good Decision,*chapter 4):**

**B**usiness question

* What is the stated business question?
* What is the intent underlying the question (e.g., what is the context, what is the impacted segment, and what are stakeholders’ current thoughts about the underlying reasons?
* What business considerations (e.g., stakeholders, timeline, and cost) are likely to impact the analysis?

**A**nalysis plan

* What is the analysis goal?
* What hypotheses are to be tested?
* What data is required/available to test the hypotheses?
* What methodology(-ies) will you employ?
* What is the project plan (timeline and milestones, risks, phasing, prioritization, …)?

**D**ata collection

* From where can the data be obtained?
* How must the data be cleansed and validated?

**I**nsights

* What patterns do you see in the data?
* Are each of the hypotheses proven or disproven?
* How much confidence should stakeholders place in the results?
* How do you rank your findings in terms of quantified impact on the business?

**R**ecommendation

* How can you most effectively present the results of your analysis to your stakeholders (in terms they can understand and in alignment with information they’ll value)?
* Note: A generic template for a recommendation presentation or report might include:
* Objective
  + Background (optional)
  + Scope (optional)
  + Approach (optional)
  + Recommendations
  + Key insights with impact
  + Next steps

**[3. Produce a report to management](https://ut.daacertificate.com/mc/poa?productID=2652&taskID=3334" \l "collapsepoa3102)**

Create a report to your client's management in the form of a PowerPoint presentation. Your report should include:

1. A written statement of the goal(s)
2. A well-defined data science process framework and the reasons you are proposing it
3. Descriptions and location of related data sources
4. An explanation of how you will manage the data for the project
5. Descriptive statistics you have applied as an initial step in data analysis
6. Any known issues with the data and how you plan to address them
7. A flowchart visualizing the detailed process you will follow, annotated with any potential pitfalls you’ve identified and your proposed solutions to such pitfalls.
8. Any initial insights you can glean from your quick look at the data. (Recall that you were previously ask to recommend three management-level decisions on the basis of your initial analysis.)

TIP:

You can use LucidChart, Vizio, Dia, PowerPoint, et cetera to produce your flowchart.

TIP:

Remember that your report is intended for business rather than technical people. Consider the things likely to be important to a business person and the things likely to confuse them. If you believe that a technical detail is likely to be important, think about how best to explain it to a business audience.