**General Process of a Data Project**

1. Meet with the ***subject matter experts*** (aka line of business experts, scientists, domain experts, etc). Determine what questions they want to address, take time to think about alternative questions or approaches.

*This is tough to do in a class project where you are working alone, try running the question through a LLM to try to generate additional ideas. In the Jupyter notebook, credit ideas or suggestions from the LLM.*

1. Look for ***data sources***, both internal to your organization and externally sourced. Consider data privacy, ethical data use and security at this point, not later! Consider issues of ethical data quality, explicit sources of bias in the data and implicit sources of bias in the data.
2. Once you have one or more possible data sources, have developed an understanding of the basic data structure and the basic ethical issues and security concerns, cycle back to your partners in step 1. Repeat steps 1-3 if necessary.
3. Form a ***training set*** and a ***test set*** from your data, if necessary. A test set is an “untouched” set of data you never look at during the analysis and use to test your results or models as the last step. If you will need a test set, use 10% to 50% of your data as a test set, and the rest as a training set. Do all your exploratory work, visualization and modeling on the test set only. Use the test set to check or verify your work. Not all analyses require a test set, and there are other ways t get test sets, but create one now if you need one!

*For the DSE5002 test project, you do not need to do this, it is critical when using advanced predictive models and you need to think about it before exploring the data. You want a test set you have no knowledge of to challenge your predictive model(s).*

1. ***Exploratory data analysis***- Understand what variables you have and what they mean. Look for errors and outliers, use several different data visualization method sto understand what is in the data. Use some visual methods to address the questions your team has. Cycle back to your team (domain experts) and talk about what you found in the exploratory analysis.

The exploratory analysis may be enough to answer the questions!

You may find that the questions change after the exploratory data analysis, or you may need more data, or someone may think of additional data to use. Depending on what happens in this discussion you may be done, or you may go back to step 1, or maybe proceed to:

1. You will probably do one or more of the following next:
2. ***Unsupervised learning***- looking for clusters, market segmentation, social networks, or other forms of latent structure
3. ***Supervised learning***- building predictive models for future events
4. ***Hypothesis testing*** of competing ideas
5. Building automated tools (*proscriptive* tools) to carry out or execute a task or solution.
6. Find ***performance measures*** (KPIs) for your solution, measure how well it works.
7. Use some sort of ***test set*** to evaluate your models or solution.
8. Report your final results. Discuss matters with your domain expert partners. Return to some prior point in the analysis and repeat as necessary. Remember though that you only can use a test set once as an independent test, so that portion cannot be repeated.

Application of these ideas to the R Project

Here is the project statement

Your CEO has decided that the company needs a full-time data scientist, and possibly a team of them in  
the future. She thinks she needs someone who can help drive data science within then entire  
organization and could potentially lead a team in the future. She understands that data scientist salaries  
vary widely across the world and is unsure what to pay them. To complicate matters, salaries are going  
up due to the great recession and the market is highly competitive. Your CEO has asked you to prepare  
an analysis on data science salaries and provide them with a range to be competitive and get top talent.  
The position can work offshore, but the CEO would like to know what the difference is for a person  
working in the United States. Your company is currently a small company but is expanding rapidly.

Prepare your analysis in an R file. Your final product should be a power point presentation giving your  
recommendation to the CEO. CEOs do not care about your code and don’t want to see it. They want to  
see visuals and a well thought out analysis. You will need to turn in the power point and the code as a  
flat R file.

**In an RMD notebook, do all of the following:**

A.) Restate the question in your own words

-create several alternative ways to ask this question

-think about what you could add to the question, or remove from it

B.) You are being given the data set, you don’t need to find another one, or consider other sources in this project. But you do need to do:

1.) Load the data into a dataframe

2.) Make sure all the variable types are set correctly. Think about whether any of the

variables should be factors (indicating membership in a group). Use the factor or as.factor

function to change these all to factors.

3.) Do a basis exploratory data analysis

-use summary and comment on what you see for each variable

-use one other table, or summary, that produces a text output for each variable

-do at least one plot of some sort per variable

C.) Prepare at least two possible tables and two possible graphs that would answer the question posed by your CEO.

**Using PowerPoint, do the following**

D.) Prepare a deck that answers the CEO’s question clearly. You are allowed a title page, and then three other pages. Your CEO likes short but complete presentations.

Your presentation *needs to tell a story* that the CEO can make sense of. Reports on data projects always need to tell a story, most people structure the world as stories, you gotta have a story.

Write down a short paragraph in your RMD file that summarizes your story.

Prepare high quality graphics in R, and/or tables as needed, paste them into PowerPoint. Let me know if you haven’t used PowerPoint before.

You should not have a presentation with no graphics- but it could all be graphics if you want. Just not all text