Reading Material for This Week

DO NOT WATCH THE VIDEOS THIS WEEK. Instead, I want everyone to do some independent research and try to understand the major idea behind the following terms: Covariance and Correlation Matrix of multiple variables and the Multivariate Normal Distribution. Just see what google/Wikipedia has and then read around the web a bit.

PRE-LIVE ASSIGNMENT

I have included 3 linear regression projects from previous semesters. I want you to read them and provide some feedback both negative and positive. When thinking about the critique, it is important to examine not just whether they technically did it correct, but if what they are saying actually reflects and supports what they did technically. Below are some questions I put together in case you find yourself lost on what to consider or look for. I’ve bolded a couple of important ones.

1. Did the report clearly state how they handled missing data, how they treated variable types, and/or additional data cleaning steps, splitting of data etc.
2. Was it clear that they were using the same data set throughout the entire study? That is was the same data used in EDA also the data used to train the model? If not, is still okay?
3. Was a summary statistics table provided?
4. **While reporting EDA, do they provide any insight into what is going on with the variables other than simply stating which ones are highly correlated? Is every graph in the EDA (both in the main text and in the appendix) referenced in the document?**
5. **How much discussion was there on connecting what the EDA is saying and what their final interpretable model suggests?**
6. Did they technically interpret the regression coefficients correctly? Did they create their own report table or did they just submit the table directly out of R?
7. **When building complex models, did they even attempt to fit a more complex model? Did they appropriately assess the bias/variance trade off?**
8. When comparing models, were appropriate metrics used and was it clear what metrics they used? For example is it clear that the ASE reported is from the train, test, or found via CV?
9. **If you didn’t have the graphs present and only read the words, would you know what is going on?**
10. **Do you feel like you could reproduce what they did (without looking at their code)?**
11. Is there excessive amount of codes produced in the document itself? Is every reader of the report going to know R?

We will spend a good portion of time discussing this and I will prepare some other big level strategies including some technical things.