ABOUT THE PROJECT AND FINAL PRESENTATION:

The class project offers you the opportunity to practice and pull together every method that we use in this class. Even if the Project assignment in class suggests that you do not need to use all the methods, I am requiring that you do the following:

Required Methods:

1) Full cleaning, prep, EDA and visual EDA
2) Association Rule Mining (unsupervised) Offer the top 10 rules for the highest sup, the top 10 for conf, and the top 10 for lift. All rules must have at least one element on the left and one on the right. Also choose to set the left as a given value and show the top 10 (based on your dataset and determinations).
3) Clustering (k means and one other – your choice) (unsupervised) Include visualization. Run k means for three choices for k and choose the best.
4) Decision Tree (supervised). Include at least three different trees and their visualizations.
5) Naïve Bayes
6) SVM – use at least three different kernels, and for each kernel, look at and discussion 3 different costs.
7) Text Mining – it is your choice how to apply this.

Notes:

1) You must have at least two different vis types that support each method.
2) Each method above will have its own subsection(s) in the Analysis and in the Results.
3) Place the visualizations where they belong in the your project – not at the end.

NOTE: Your Live Presentation (Week 10) will also be part of your grade. If working in team, team members will get the same grade and must BOTH present.

Live Presentations:

(1) Introduce - what is your area (briefly!) and what are your goals.
(2) Discuss 5 key analyses, their results and their visualizations. You will have more than 5 analyses in your report, but will choose the most interesting 5 to present.
(3) Offer clear and anyone-can-understand conclusions.
- Include nothing about the data cleaning, etc.
(4) Be precise and succinct.
1) Your Project should be written professionally. You may use RMarkdown, ppt, pdf, Word, or a more advanced option.
2) Whether you use RMarkdown or not, you must have all the written areas (paper submission).

Required Project Headings:

Note: These are a bit more involved than the Headings shared for the Assignments

Introduction - (3-4 paragraphs all about the area of interest. Do NOT discuss the data in the intro). The intro is non-technical and offers background, objectives, history, value, goals, etc. It must be readable by a person who is 10 or 11.

GRADE NOTES:

a) The Intro discusses the data or dataset. (-10 points)
b) The Intro discusses data science in general or models and methods. (-10 points)
c) The Intro is shorter than 3 paragraphs (a paragraph is 7 sentences). (-10 points)
d) The Intro is not clearly or well written and/or does not offer a good basis for what the project is about. (up to -10)
e) The Intro is written below the graduate level, such as by saying things like "In my project, I …", or "This Project is about….". Consider this project to be an academic paper. Its never about you. Avoid all use of "I", "we", "you", …

Analysis

Subsection 1: The Data

This is where you talk all about the data, the variables, the cleaning, measures of cleanliness, etc.
Even if your data is clean, you must write code to clean it (pretend you cannot see that it is clean). Minimally: NAs, changing things to factors as needed, outliers, discretization, incorrect values, such as .23 for age, etc.

Have a sub-subsection for each variable and show that you looked at all the criteria noted above. Show AND MEASURE the before and after. For example, if you find that a variable has 10 missing values and you update these with the mean, then the before is the mean before, and the after is the mean after. In this case, you should also include the variance before and after. The measure is *very* dependent on what you clean and how you clean it. So this will be for you to think about.

Subsection 2: EDA: statistical and visual. Create a vis and/or table for each variable in the dataset.

Subsections 3 - n:

Here you run all methods learned in this class. One subsection for each. Include tuning and/or different options as applicable - such as different kernels and C for SVM, different k for kmeans, etc.

Explain as you go. Pretend you are writing a tutorial paper.

GRADE NOTES:

a) If there is not a clear subsection (and sub-subsections) that are all about the data, dataset, data cleaning, variables, EDA, visual EDA, prep, etc. (up to -15)
b) If there is not a subsection dedicated to each required method noted above. (-10 for each missing method).
c) If there is not at least one vis for each required method in the Analysis area and within the subsection for that method. (-5 per missing vis)
d) If the model, model parameters (such as kernel for SVM), etc. are not well explained points can be lost.
e) If a model or method is not correctly applied or gives results that should not occur, points can be lost.

Results

Subsections 1 – n

You will have and will discuss results, issues, and limitations for all the analysis. You will note which ones worked well, and why, which ones did not, and why, etc.

GRADE NOTES:

a) If any method is missing. (-10 per missing method)
b) At least one results-supporting vis for each method. (Missing a vis – 5). These are not the same as the visualizations required in Analysis.
c) The Results section will have a subsection for each method. Each method subsection will be different in the sense that each method is different, has different parameters, etc. For example, when you do kmeans, you will talk about the different values of k, how each performed, what each revealed, and which was selected and why. Results talk about what happened in each method, but technically.

Conclusions:

3-4 paragraphs - NON-TECHNICAL. What was the outcome - what did you find, discover, predict, classify? WHy does it matter to humans?

GRADE NOTES

a) Less than 3 paragraphs (-10)
b) Any mention of technical results (which belong in the Results above) -10
c) A lack of flow or clarity about what actual findings were, why they matter, who they matter to, how they can improve things, etc…

A note about communication

1) Effective communication - via presentation and writing - is critical.
2) In life, we are often judged and measured through these avenues.
3) When writing technical, academic, and/or professional papers (assignments and projects):
- Avoid speaking in the first person. Avoid “I”, “you”, “me”, “we”, etc.
- Include important R code and visualizations throughout the paper. This does not mean that you should show every line of your R code (please do not), but rather you can include critical lines that show models and methods.
- Include R results in your Assignments and Project that support your models, methods, analysis, and results. Do this smartly.
4) Proof-read all submission out loud first. Be sure your assignments/project have a clear flow and are easy to follow and understand.