Font : Time Roman, 12 , double space APA format

Please choice any method below to include in the project.

[**Week 1: Introduction to tm Package**](https://northeastern.blackboard.com/webapps/blackboard/content/listContent.jsp?course_id=_2608939_1&content_id=_22578981_1)

### [Week 2: Finding Patterns in Data and EDA,](https://northeastern.blackboard.com/webapps/blackboard/content/listContent.jsp?course_id=_2608939_1&content_id=_22578992_1)

### [Week 3: Data Munging and Data Wrangling & Logistic Regression, Decision Trees, and Random Forests Lectures,](https://northeastern.blackboard.com/webapps/blackboard/content/listContent.jsp?course_id=_2608939_1&content_id=_22579002_1)

### [Week 4: Cluster, Association, and Linear Discriminant Analysis Lecture](https://northeastern.blackboard.com/webapps/blackboard/content/listContent.jsp?course_id=_2608939_1&content_id=_22579016_1)

### [Week 5: Support Vector Machines Lectures,](https://northeastern.blackboard.com/webapps/blackboard/content/listContent.jsp?course_id=_2608939_1&content_id=_22579026_1)

**I will also need a Proposal.** Prepare a short paper. 1-2 Page

**Project proposal**

You should HAND IN A PROJECT PROPOSAL by the proposal due date. The proposal is a page or so describing what you plan to do. Be as specific as possible, describing what data source you’re using, what questions/objectives you want to investigate with this data, what R package might help with the visualization or analysis, and generally how you plan to answer these questions. I will return the proposals to you with comments.  The more detailed your proposal, the better feedback you get!  Your proposal should address the following questions:

* What is the topic of your project?
* What is the data source?
* What are the main issues or problems you plan to address?
* What R package(s) might you try using?
* What are your plans for obtaining background information (if needed) about your project?
* Describe the data set that you plan on using, including the variables measured and how many observations.
* What questions and/or concerns do you have about your project?

Things you need to consider.

What is the data set?

Where did it come from?

What is the structure?

What do you think you will find?

Which modeling methods will you consider and why?

The project requires you to synthesize all the programming and data mining skills from the course. Hence, it's one of the best ways to solidify your understanding of R.  Plus, you can get answers to issues that pique your intellectual curiosity.

Your project will be submitted in several ways.  You must submit your R script file with all commands used in the analysis.  All commands must work (no errors) so that I can replicate your analysis.  **You must also submit a 7-10 page paper summarizes your project.**  All three aspects of the project (R script file, video presentation, and written report) should have the following:

* Explains your data and the objectives of the analysis
* Shows graphs & other descriptive statistics I get a ‘feel’ for the data
* Describes analysis techniques used and give answers to the objectives/goals
* Comments about any statistical model adequacy, any shortfalls you ran into, suggestions for further analysis or data collection…
* Shows off your newly developed R skills, & is creative!!

**In this project, I am asking you to summarize your project using BOTH a written paper and an oral presentation.  Why both??  Because in the "real world" it is important to have outstanding written and oral communication skills.  Both skill sets are necessary to communicate Analytical results in your careers, especially to senior management.  I will be judging you on your communication effectiveness as well as the technical R commands you develop in the analysis.**

Project grading guidelines

As I am grading, I will be looking for the following characteristics:

1. **Consistency:**  Did you answer your question of interest?  Did you synthesize several R programming techniques learned in this course to describe, visualize, and analyze the data? Did you use at least one R package to do so?
2. **Clarity:**          Is it easy for your reader to understand what you did how it was implemented in R?
3. **Relevancy:**     Did you use statistical techniques and R programming skills wisely to address your question?
4. **Interest:**         Did you tackle a challenging, interesting question (good), or did you just perform basic descriptive statistics (bad)?

Some suggestions for scoring high on these criteria, and suggestions you should keep in mind whenever you write anything, are the following:

Guidelines for making an effective paper or presentation

An effective written analysis paper or oral presentation communicates your project in a clear and concise fashion.  The paper and presentation should address the following six points:

1. **Statement of the topic/problem:** Describe the questions you address and any key issues surrounding the questions.
2. **Data Description:** Explain the data set to the reader/listener. Describe the types of variables and data structure.  Provide graphs, plots, and descriptive statistics so that the reader gets a good ‘feel’ for the data set.
3. **Analyses:** Describe the analyses you did.  Explain why you believe these methods are justified and how they answer the objectives of the analysis.
4. **Results:** Present the results of relevant descriptive statistics, statistical analysis, and data mining techniques. Include tables or graphs that support your analyses (be judicious here--too many tables and graphs hurts the clarity of your message).
5. **Conclusions:** Answer your question(s) of interest.
6. **Discussion:** What implications do your results have for the population in which the data came from?  What could be done to improve the study if it was done again? What types of biases might exist?