OMIS 482 - Semester Project

Deadlines

- **10/03**: Project Part 1(10%)
- **11/28**: Project Part 2 (20%)
- **10/ 10 & 12/04:** Peer evaluations (5% one after part 1 and one after part 2)
- **11/30 & 12/02**: Project Presentations (5%)

Project Part 1

Part 1 should include the following items:

- Provide a team name. Pick a name that is catchy and related to your project.
- Select your data set for the semester project among the three available:
  - Housing data.
  - Airport data.
  - Movie data.

  Each data set is available on Blackboard and you can review them and pick the one that seems the most interesting for your group. Each dataset presents different challenges for you. From large amount of observations to high dimensionality in terms of columns. Moreover, remember that different modeling techniques requires different type of variables (e.g., don’t attempt to perform a logistic regression if you don’t have a binary variable in your data set). So, it is important to have different data type variables. Keep these components in mind when choosing your dataset. The same dataset will be used for both RStudio and SAS EM portions of the project.
- Create a Project in RStudio and a RMarkdown file. Write in the RMarkdown the necessary code and the reasoning why you wrote/used that code in your project. The reasoning should make it clear for a reader or your boss (me in this case but it can be assumed as your supervisor in your first job after graduation) on why you decided to perform this kind of analysis/ manipulation to your original dataset. I recommend brainstorming with your group what is your research question for this semester project. Start by brainstorming, what is the most important thing you would like to know from the dataset you selected? Then the code you will write is a consequence to your research question and it will enable you to find answers.
- You should write code to:
  - import your dataset in RStudio,

    - tidy your dataset (if required),
    - explore the variables inside your data (know your data is a critical step in any analytical project) and
    - manipulating your data (creating new variables of interest or adding variables of interest from other datasets).
- Remember, Part 1 needs to be conducted in RStudio, even if Excel is easier or you are more familiar with it, you need to practice and master RStudio to be successful in this class and in your professional career.
  I do not specify a quantity for each function explained in class (filter, mutate, arrange, etc.) because the number of functions really depends on what do you want to know from your data, and on which question you are trying to answer. However, try to tell a story based on your data and your objective.
  See the Part 1 Mock Project if you want to get an idea on my expectations on this section of the project. Remember, that the project is where you showcase that you are mastering the skills covered in the videos, so the more the better ;-)
  Also, I expect all the people in your group to contribute so there should be a multitude of manipulations applied to your data.
- Visualize your data in RStudio, charts are extremely useful in explore and get to know your data. Some people are good at reading and interpreting descriptive statistics, but others need visual elements to make sense of them. Use charts to get to know your data even more. Start thinking about variables distributions, possible required transformation and so on and so forth! A big component of your future PPT presentations at your job will include creating charts that show interesting findings that arise from your analysis. Make sure to embed a set of attention-grabbing charts in your story telling narrative.
- Make sure that your report has a title page (group name, authors name, date), an introduction section, a body section (the core of the analysis) and a conclusion section. All these sections are required.
- Knit your RMarkdown in a Word file when you are happy with your code and the project narrative. Submit one RMarkdown and one Word document per group by the deadline (October 3rd at 11:59 pm). Length limit for the Word document is 20 pages, so make sure to include only relevant information (code, code output and charts) in the Word document.

NB: 10 points from your Part 1 grade will be deducted for each day past the deadline.

Project Part 2

In part 2 we switch from RStudio to SAS EM. Pretend like you are just starting the analysis and you didn't complete part 1 in RStudio

- Import the original dataset into SAS EM. However, make sure to use the same dataset selected in Part 1.
- Get to know your data in SAS EM. Explore your data using the SAS EM nodes covered in the lecture videos. Is it easier to explore in SAS EM or RStudio?
- Make sure to also provide relevant charts for the variables of interest of your project using the SAS EM nodes covered in the lecture videos.
- Perform data manipulations to your original variables or observations. Use the nodes covered in the lecture videos. Also, in this case there is not a minimum or maximum number of nodes that you should use because their use heavily depends on your project objective.
- Modeling using the Decision Trees node in SAS EM.

- Modeling using the Regression node in SAS EM.
- Providing the model assessment of the modeling nodes used in your project. This step is critical to compare different models and identify which one is superior in predicting your data.
- Providing the model implementation of the model that best predict your data. Ideally, you will conclude your project by implementing the most performing model on unseen data. While this step is extremely important if you are trying to pursue a career in analytics using SAS EM, it is not required for this semester project.
- You will take screenshot of relevant outputs (E.g., your process flow, nodes settings, charts and results) and you will provide a narrative to explain why you use those nodes and why in that specific order. Also, you should explain if and the results of some nodes affected your analysis. The narrative and screenshot will be included into a Word document in logical order. Remember that the objective is to write a story about your data and what you learned about them. Keep in mind that is critical that you include the Z-ID of any of your group member in your SAS EM screenshots. The Z-ID is reported in the bottom right corner of the SAS EM window. While there might be situations in which it is difficult to include your Z-ID (e.g., if you are zooming on some property panel settings or some specific output) to prevent cheating I expect that Z-ID is available in the majority of the screenshot. I hope you understand that this is a measure required to enforce fairness in the grading.

- Make sure that your report has a title page (group name, authors name, date), an introduction section, a body section (the core of the analysis) and a conclusion section. All these sections are required. Please make sure to include in the conclusions some considerations and final thoughts on your semester project (include relevant findings, lesson learned and issues you encountered in completing it). Feel free to look at Part 2 Mock Project to see what my minimum expectations on this part are.

- Submit one report (Word document) per group by the deadline (November 28th at 11:59 pm). Make sure to include in the Word document only the most relevant materials that address the above points (limit is 30 pages).

NB: 10 points from your Part 2 grade will be deducted for each day past the deadline.

Project Presentation

Your presentation (15-20mins including questions) should cover:

1. What is your research question/the problem you are trying to solve? Why it is important?

2. Description of your data. For instance, what are its specifics and why you select it to answer your research question? Use some charts to make it more visually appealing and a table to summarize each column.

3. Presenting your main results and limitations of the different modeling techniques. What did you find? Any technical aspect that you are particularly proud of? Any issues with the data (e.g., missing values, suspicious patterns)

NB: 10 points from your Presentation grade will be deducted for each day past the deadline for submitting slides.

All the team members should present but you will decide the order and the slides each of you will cover during the presentation. The presentation can be on both parts of the project or only on part 2.

Project Grading

I grade the project part 1 and part 2 holistically, which means that I consider the project submissions as a whole, and I do not follow a rubric with dozens of items each worth a few points. Each project is different, and this gives you the flexibility to devote more or less to various aspects of the project depending on what's appropriate, within reason. However, keep in mind that I am more impressed by quality than quantity (so please do not add 10 screenshots showing SAS EM default properties settings of different nodes or make 10 basic charts in RStudio of 10 different variables). The more the better but I am looking for quality work, as I know that you guys are capable of. In determining grades, I take the following into account:

- **Originality:** Are your questions thought-provoking? Do they encourage the reader to think about the topic in a new way?
- **Complexity:** Is your code well written (use of %>%)? Are relevant manipulations applied? Are the appropriate charts used (right geoms and aesthetics)? Is your process flow structure taking into account relevant nodes in your analysis? Are any nodes ignored without motivations?
- **Content ownership:** Do your graphs and textual descriptions reflect a solid understanding of what your data mean? Is it clear why are you asking the questions that you are asking? Did you create the appropriate chart/ used the right tools? Are your interpretations reasonable?
- **Story telling:** Does your report have a logical flow? Does it read as an interesting story about your data, your analysis and your results? Do you keep the reader engaged?

Additional Notes

I will take attendance during your group meetings with me and might ask questions to each of you. So, I will be able to assess each individual contribution. However, I will assign the same grade to each group member unless I recognize that the contribution is really unbalanced.

Moreover, please note that 5% of your semester project grade will be determined by a peer evaluation within your group. Peer evaluation should be based on contribution to the project and not on personal preferences (e.g., Biagio is my best friend so I will give him 100 even if he sleeps during our meetings). I will provide indication on how submit your peer evaluation later in the semester.

Finally, I truly believe that all of you should be mature and professional enough to work in any group. Also keep in mind that in your future job you will not be able to choose your coworkers! However, if you have any issue in working in your group, or there is an unbalanced contribution please notify me immediately. I will not accept complaints the day before the presentation or after you made the submissions.