**MSDS 6306: Doing Data Science**

**Case Study 02**

**Due: One hour before Week 15 Live Session**

**Description**: DDSAnalytics is an analytics company that specializes in talent management solutions for Fortune 1000 companies. Talent management is defined as the iterative process of developing and retaining employees. It may include workforce planning, employee training programs, identifying high-potential employees and reducing/preventing voluntary employee turnover (attrition). To gain a competitive edge over its competition, DDSAnalytics is planning to leverage data science for talent management. The executive leadership has identified predicting employee turnover as its first application of data science for talent management. Before the business green lights the project, they have tasked your data science team to conduct an analysis of existing employee data.

**Your team has been given a dataset (CaseStudy2-Data.xlsx) to conduct exploratory data analysis (EDA) to determine factors that lead to attrition. You** should identify (at least) the top three factors that contribute to turnover(attrition). There may or may not be a need to create derived attributes/variables. The business is also interested in learning about any job role specific trends that may exist in the data set (e.g., “Data Scientists have the highest job satisfaction”). You can also provide any other interesting trends and observations from your analysis. **The analysis should be backed up by robust experimentation and (where applicable) appropriate visualization**. Experiments, analysis, and visualization must be conducted in R. If needed, consult with the business owner to get clarifications on the dataset or business goals.

**robust experimentation: is gender affecting attrition. Is there going to be any**

**job role vs Gender**

**1) null hypothesis : gender distribution across attrition categories is same.**

**2) null hypothesis : attrition across genders is same**

**Is there fair distribution of income across same experience levels**

**Categorical variable**

**How to visualize a relation between a categorical variable and a continuous variable .**

**In a good light <> good spirit**

1. **Department is one of the attributes we found some differences**
2. *## M F*
3. *##hr 23.5% 16.2% 42.9% <----- HR WTF?*
4. *##rnd = 16.1% 18.3% 12.8%*
5. *##sales = 26% 26.6% 25.2%*
6. **Monthly income**
7. **Num of companies worked . : we might not find a statistical significance but this variable seems to have some practical significance .**
8. **Total number of years**
9. Explore the columns and their relations
10. Linear regression or other techniques
11. Any other advanced analysis
12. Looking at box plot following 3 variables show some diff:
    1. Years in present role
    2. Years in company
    3. Years with current manager
13. Are linear relations really linear or do they need some transformation or is this even relevant
14. Is there any variable that should be analyzed for classification sort of thing
15. Is there any other visual analysis (graphs etc to be included)

**Deliverables:**

This is a group project so it’s the responsibility of the group members to collaborate accordingly. Team work is important.

The due date for submission is 1 hour before live session 15. During live session 15, the group will have up to 20 minutes to present their case study report. The goal is to communicate the findings of the project in a clear, concise and scientific manner. Tell me the story.

Create a GitHub repository with an RMarkdown file containing an executive summary, introduction to the project, all supporting code and analysis, and the group presentation.

Submit a link to the GitHub repository via the space provided for the Case Study 02 page in 2DS.

Your Presentation should be in this outline

* Title
  + Authors all listed
* Presentation Outline
  + Business Objectives
  + Data Sourced
  + Methodology
  + Evaluation/Results
  + Summary
* Business Objectives
* Data Source
  + Where you got it
  + Basic Statistics (EDA)
* Methodology
  + Steps
  + Workflow
* Evaluation/Results
  + Tell me the percentages and why
  + Show me graphs with explanations
  + The top three factors that contribute to turnover
  + Tell me about any job role specific trends that may exist in the data set
  + Provide any other interesting trends and observations from your analysis
  + Other things to consider?
* Summary
  + Insights
  + Recommendations
  + Improvements
* Questions

Have your presentation ready to present on during live session 15.

NO CODE should be present in your presentation. (ie variable names is Mon\_Age this should in your presentation as age). You code should be in your GitHub and RMarkdown File.

Remember, no late assignments will be accepted, as this is the final. It has been a pleasure teaching you this semester! Congratulations on making it through your first semester in the Masters of Data Science program. Good luck!

# Remember:

A GitHub repository and presentation

1. Introduction to the project. The introduction should not reference a project, persay. No part of this should be informal. It should be appealing-looking as well.
2. The introduction needs to be written as if you are presenting the work to someone who has given you the data to analyze and wants to understand the result. Pretend it’s a presentation for a client. This may take some imagination of whom your client might be. **If it sounds like a student presentation, that is not acceptable**.
3. R code for answers concerning the tasks below. Make it in RMarkdown file format and always include echo=TRUE and include=TRUE for charts. **Keep the .md file so I can readily see everything on GitHub! Your .md file should mirror the .RMD file.**
4. Give clear, explicit answers to the questions. Just the code to answer the questions is not enough, even if the code is correct and gives the correct answer. You must state the answer in a complete sentence outside the code chunk.
5. Once you’re finished, be sure to frame a conclusion to the project. The presentation does not stop when you’re done with the questions. Find a way to wrap it up: summarize your findings from this exercise. Again: **the file must be readable in GitHub. In other words, don’t forget to keep the md file!!**
6. You should expand your repository with at least this RMarkdown file, the input file(s), and a README.md that describes the purpose of the project and codebook. The repo can be structured how you like, but it should make sense and be easily navigated. If things are not clear from the root directory, you will lose points.

You’re working on this as a group. I expect that all will do equal work. All will need to push, add, commit, and pull the GitHub. If I do not see equal effort (and remember, GitHub tracks commits!), I reserve the right to penalize students. This is a collaborative project, so take it seriously and plan ahead with your teammate

# Tasks

*Tip:* I recommend in the code block commenting which number the code answers. It will help me find your answers more readily. Something as simple as # 3B helps me.

1. Formulate your Repo
   1. The client wants this to be reproducible and know exactly what you did. There needs to be an informative Readme, complete with several sections, as referenced in Live Session. Give contact information, session Info, and the objective of the repo at least.
   2. You have a large data set, and it needs its own Codebook, formatted in an approachable way. Make sure you describe peculiarities of the data by variable and what needs transforming. However, do not make it too long either.
   3. Create a file structure that is accessible and transparent. Document it in the root directory, ideally in the Readme.
2. Clean your Raw Data
   1. Read the csv into R and take a look at the data set. Output how many rows and columns the data.frame is.
   2. The column names are either too much or not enough. Change the column names so that they do not have spaces, underscores, slashes, and the like. All column names should be under 12 characters. Make sure you’re updating your codebook with information on the tidied data set as well.
   3. Some columns are, due to Qualtrics, malfunctioning.
   4. Make sure your columns are the proper data types (i.e., numeric, character, etc.). If they are incorrect, convert them.
3. Preliminary Analysis
   1. Remove all observations where the participant is under age 18. No further analysis of underage individuals is permitted by your client. Remove any other age outliers as you see fit, but be sure to tell what you’re doing and why.
   2. Please provide (in table format or similar), descriptive statistics on at least 7 variables (age, Income, etc.). Create a simple histogram for *two* of them. Comment on the shape of the distribution in your markdown.
   3. Give the frequencies (in table format or similar) for Gender, Education, and Occupation. They can be separate tables, if that’s your choice.
   4. Give the counts (again, table) of management positions.
4. Deeper Analysis and Visualization
   1. *Note:* You should make all of these appealing looking. Remember to include things like a clean, informative title, axis labels that are in plain English, and readable axis values that do not overlap.
   2. Create bar charts in ggplot or similar. The bars should be in descending order, Use any color palette of your choice other than the default.
   3. Is there a relationship between Age and Income? Create a scatterplot and make an assessment of whether there is a relationship. Color each point based on the Gender of the participant. You’re welcome to use lm() or similar functions to back up your claims.
   4. What about Life Satisfaction? Create another scatterplot. Is there a discernible relationship there to what?