# BSTAT 5325 – Fall 2022 - Course Project
# Predicting Secondary School Student Math Performance
# Due Date: Dec 6 at 11:59 pm

## Introduction

a) This is a group project that requires all group members to contribute equally. You have a responsibility and authority to let me know as early as possible during project development if any member(s) of the team are not contributing. It is my role to support you and address any group problems that you might be facing.
b) Start working on the project as early as possible. You don't have to wait until we cover all statistical techniques before you begin. It will be too late otherwise.
c) I will review and grade each project report.
a) All reports will undergo a plagiarism check. Evidence of plagiarism will negatively impact your grade.
b) Develop a project plan to organize your approach, track progress, allocate tasks fairly, and create accountability.
c) Iterate and don't try to develop the final solution in one shot. Use Agile and Lean principles to organize yourselves and learn incrementally.
d) Ask me question if anything is not clear or if you need any support.

## Business Case Context

You are a team of Data Analysts/Scientists working for a consulting firm that specializes in data and analytics. You have been hired by a real estate company called "Ames Real Estate Associates" (AREA) to help them improve their ability to predict house prices and identify the factors that are likely to increase the price of homes that their customers want to sell.

The real estate company, AREA,  is recognized in the city of Ames, Iowa for its market expertise and special approach to helping its clients sell their homes. The way the process works is that AREA advices its clients who are looking to sell their homes as to what price to expect and what home features or characteristics could be improved to maximize the selling price. In some cases, their clients want to sell their homes for a higher price than what the market will offer. In those cases, AREA can provide them with a report describing why their home is worth less than comparable homes.

At a minimum, key business questions that you will need to answer include:

1. What is the expected selling price of my home?
2. What factors influence the price of my home?
3. Which factors are more important than others?
4. How much should I invest in improving the condition of my home in order to increase the expected price by more than the cost of improvements?
5. Which homes should I compare my house to?
6. When is the best time of the year to sell my home?

You are encouraged to answer additional questions and bring additional insights or just questions that the AREA team has not thought about.

**<u>It’s critical that you explain throughout your report why you performed the analysis tasks that you did. Primarily focus on explaining the reason behind you descriptive and inferential statistics work.</u>**

## Tools and Software

You are encouraged to use any tool or combinations of tools that will allow you to complete the project. Different team members might choose different tools based on expertise. What matters is that the project report is cohesive and that you explain the rationale behind your work and recommendations.

## Requirements and Grading

| **Requirements** | **Maximum Points** |
|---|---|
| **Project Report:**<br>You are requested to develop **<u>a maximum 10-page Word report (no minimum)</u>** covering the four major sections of the requirements described below. You will also need to provide worksheets, code, or calculations you used to develop the report. This is separate from the 10-page Word report. Keep in mind that your value as a business analyst is not only in the technical analysis but how convincing and concise you are in communicating the findings and telling a story. The audience of the report consist of the Chief Financial officer, key members of the management team, and the head of analysis. You need to develop the report with the audience in mind. | |
| **Problem Definition:**<br>Develop a clear business problem with operational definitions of key terms. The structure of the problem definition section must include:<br>- Business problem<br>- Motivation<br>- CALC (Constraints, Assumptions, Limitations, and Condition) see CALC section<br>- Operational definitions<br>- SMART objectives (Specific, Measurable, Achievable, Realistic, Time-Bound)<br>- A list of questions the business questions that you will attempt to answer | 20 |
| **Data Preparation, exploration, and Understanding:**<br>In this phase, you will explore the data and generate descriptive statistics. This section should include graphs, tables, and a description of the data, including assumptions that you are making about data quality and distribution of data elements. This could include measures of center, variation, and shape. You will also need to determine if all variables are required by understanding the correlations that exit among them. Remember that parsimony in modeling leads to models that balance accuracy and explainability.<br><br>Keep in mind that data exploration will help you and customers understand the data. | 30 |

| | |
|---|---|
| **Analysis:**<br>a) In this phase, you will need to apply three or more techniques that you have learned in this course to determine how to best predict the expected selling price of a home and to answer the required business questions. Use your knowledge about ensemble modeling to select the output of one or more models. Make sure to describe how you combine the scores from multiple models. Techniques that you can apply include statistical tests, correlations, Classification Trees, Regression Trees, Clustering, Factor Analysis, Multiple Regression, Logistic Regression, Neural Nets, or some combination.<br>b) In addition, you will need to use or discuss at least 3 techniques or topics that we covered in class but are not covered/or not covered in depth in the book. Examples include Simpson’s paradox, data life cycle, Kurtosis, imbalanced data, imputation, and others.<br>c) You will also need to describe how to assess and monitor the performance of your models and discuss how you can improve them. | 30 |
| **Conclusions and Recommendations:**<br>In this section, you will provide the insights developed during the study. You will also adjust and confirm any CALC (see below). Most importantly, you will need to explain WHY you decided to analyze and model the data the way you did and how this relates to the business problem and questions. You will also need to explain WHY you select to present the findings the way you did (graphs, tables, verbiage). | 20 |

## Submit the following in Canvas

a. Up to 10-page report (Word) file containing at least the following sections:
   i. Business Problem
   ii. Data Preparation, exploration, and Understanding
   iii. Analysis
   iv. Conclusions and Recommendations

b. Collection of documents supporting your work. This could include files, code, software output, and any other material that provides evidence of the work completed behind the scenes.

## What is CALC?

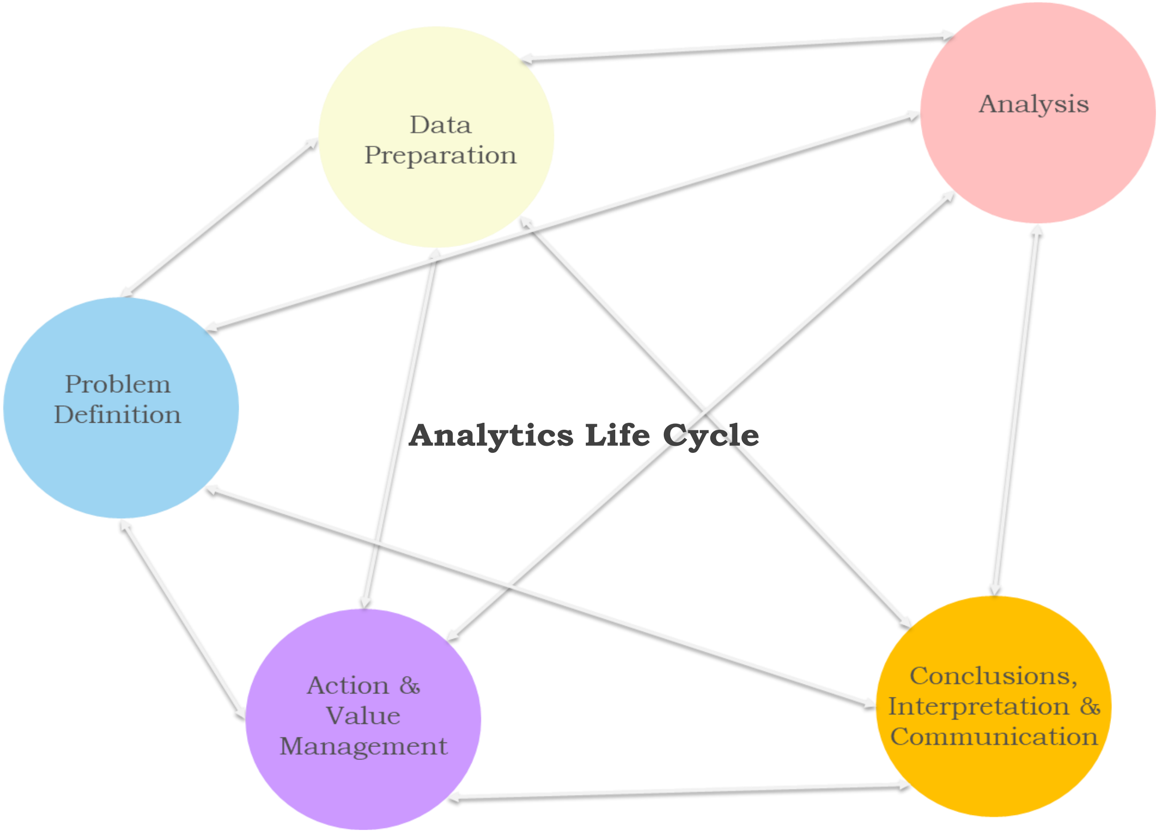
Every study has constraints, assumptions, limitations, and conditions. Here’s a definition of each as applied to this project:

- **Constraints:** restrictions imposed by the business that limit available options for resources and solutions. For example, any solution provided should not cost more than $100,000.

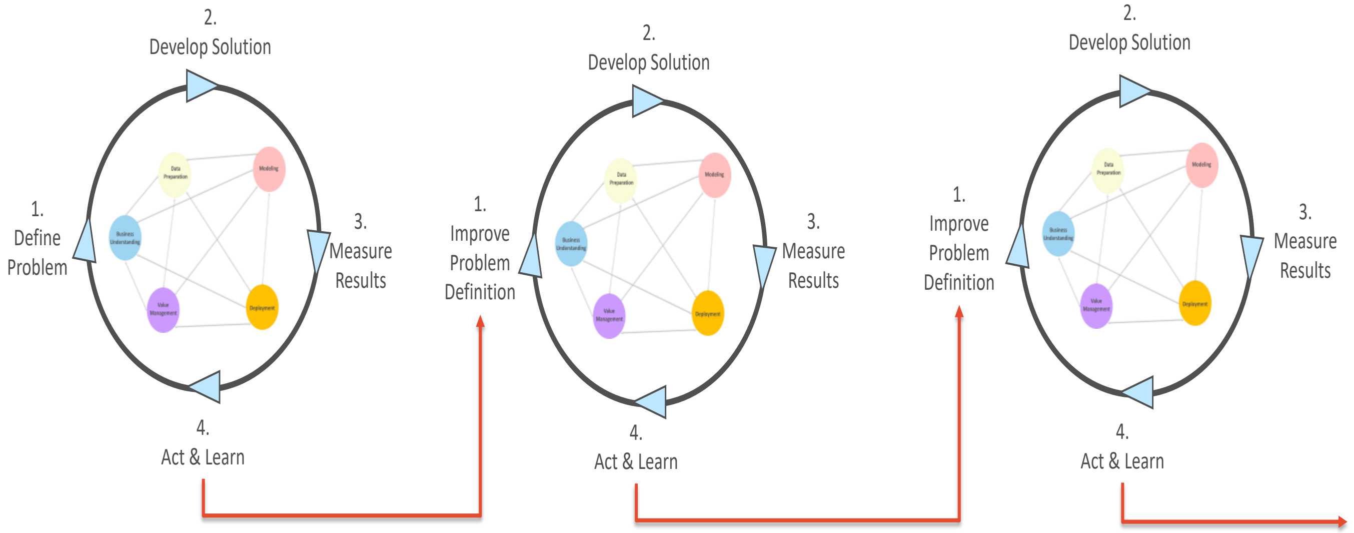
- **Assumptions:** statements related to the project that are considered to be true in the absence of verifiable data or facts, generally to accommodate a limitation.

- **Limitations:** inability or a lack of capability that reduces how and when potential solutions are developed. For example, some required data is not available, so we might generate synthetic data or assume a value/range, among other possible approaches.

- **Conditions:** special circumstances that are neither limitations nor constraints but should be taken into account during the project. This could relate to internal and external factors that require special attention. For example, forecasting sale of vehicles during a Pandemic. This is important because the results of a study might not generalizable beyond the specific context of the business problem.


Data Preparation
Analysis
Problem Definition
Analytics Life Cycle
Action & Value Management
Conclusions, Interpretation & Communication

# Agile Data Science



1. Define Problem

2. Develop Solution

3. Measure Results

4. Act & Learn

1. Improve Problem Definition

2. Develop Solution

3. Measure Results

4. Act & Learn

1. Improve Problem Definition

2. Develop Solution

3. Measure Results

4. Act & Learn

**Keep cycling to:**

- Adapt based on learning
- Improve business problem understanding
- Deliver higher business value