## Project Context & Purpose

This team project is designed to give you an opportunity to apply what you have learned in this course in a holistic and capstone-like manner. It is designed to reflect challenges and the processes of real-world data mining projects while reinforcing the learning. The project structure can also be used in the future to guide you through the CRISP-DM process for real world-projects. It outlines important general questions and concerns that you need to address in each step of a data mining project.

The project uses world development inidicators dataset provided by the world bank that includes various ecomic and development variables reported by member countries to the world bank on a yearly basis. Your job is to make sense of the data, find trends and predict some outcomes as layed out in the deliverable 2 instructions.

Due to the size of the project, and the many challenges that may arise, I strongly advise you and your team to start working on it from week 1 of the course. If you leave it for the last weeks, the chances are that you are not going to finish it or do a good enough job to grant you an excellent grade.

During the course of the project, you will face some questions that you may not be able to answer directly using the methodologies and techniques you have learned in the course. However, you have also learned to utilize different resources to answer your questions, such as using the Internet and other available resources through this course.

Please remember that some of the questions can be answered with various data modeling techniques and methodologies. It is up to you to choose what you think is the best approach or methodology for answering a question. And remember making mistakes is a way of learning. So don’t be afraid; use this platform and practice to get ready for real-life projects.

## Project Structure & Policies

* You need to form a team of 2 to 4 individuals and pick one group member as the group leader.
* It is completely your (students) responsibility to form/find a team. The instructor will not provide any support or direction in the process of team formation.
* Students who fail to form a group by the deadline will receive a zero for their entire group project grade. The deadline for group formation is stated on the D2L. A nominal part of your project grade is associated with team formation.
* If some group members drop the class after group formation, the rest are required to carry on with the project, even if it becomes a team of one student.
* The group leader is in charge of all communications between the group and the instructor and/or Celonis representative. Each communication or submission regarding the group project outside of the leader channel will result in 5 points deduction from the team project grade.
* Text in *italic* provides clarification and further instructions on how to conduct the analysis or what is expected to be submitted or included in the project report.

## Deliverable 1 – Group Formation (2% of total team project grade)

As a result of this deliverable, you must form your group, choose a team leader and report your team members' information to the instructor. The team leader will be one communicating with the instructor and submitting to D2L from this point forward.

### What to submit for Deliverable 1

* The group leader is supposed to submit one document containing the group members information, including group members’ first and last names, along with their KSU ID.

## Deliverable 2 – Applying the CRISP-DM (98% of total team project grade)

This is the second and final deliverable of the team project corresponding to CRISP-DM steps. You will need to structure your report to follow the steps on DL 2, meaning you will need to at least have 4 sections on your report corresponding to steps on this document***.*** You also need to reinstate all questions per steps on your report. Failure to comply with this requirement will result in 10 points deduction.

### Step 1 – Business and Data Understanding (20 points)

*Dataset*: world\_development\_indicators.csv

1. What is the goal of this analysis?
   1. *Hint: read the intro and the rest of the questions per deliverable to answer this question*
2. Who are the possible experts that can tell you more about the data and its context?
3. What public resources can you use to gain more insights about data and its context?
4. Why is this data gathered?
5. How is this data gathered?
6. Purely based on your business understanding, do you expect to see any relationship between variables in this dataset? If so, list variable pairs that you think are related in any way.
7. Were you able to find reliable answers to all the above questions? If not, what assumptions have you made? Express your assumptions in detail.

### Step 2 – Data Understanding (20 points)

*Dataset*: world\_development\_indicators.csv

1. How many variables are there? How many observations?
2. What does each variable refer to? How can you find more about the nature of each variable and how they are collected or calculated?
   1. *Create a list or table of all avriables*
3. Does the software assigned variable types comply with their concept? (e.g., are the numeric/real variables are presented as numeric/rea and vice versa). If not, how can you resolve this issue?
   1. *Create a list or table of all inconsistences*
4. What is an acceptable/reasonable range for each variable?
   1. *Create a list or table of all variables and their acceptable/reasonable rages*
5. Do we have ratios in the data? If so, what are they, and how are they calculated?
   1. *Put your answer in a list or table format*
6. Are the missing values represented by a special value/symbol within the original dataset, or they just presented as null?
7. Do we have variables with missing values? List them.
8. Were you able to detect any other inconsistencies or errors?
9. Were you able to eyeball any outliers for each variable? If so, explain your reasoning and assumptions.
10. Were you able to find reliable answers to all the above questions? If not, what assumptions have you made? Express your assumptions in detail

### Step 3 – Data Manipulation, Data Cleaning, and Data Preparation (20 points)

*Dataset*: world\_development\_indicators.csv

1. Assume that you are going to use a modeling technique in the future that does not tolerate missing values. Thus you need to deal with missing values. You may (should!) choose a different strategy to deal with missing values for each variable according to your data and business understanding. Use the criteria you have learned to decide what to do with missing values. Clearly and in detail, explain what you will be going to do to deal with existing missing values in your dataset.
2. Implement the plan/strategy you developed above.
3. Check for strong correlations among your variables. Did you detect any super strong correlations (above absolute value of 0.9) in the previous step? What are you going to do about them? Which variable of each pair are you going to keep and why?
4. Implement the plan/strategy you develpped for dealing with highly correlated variables you identified in the previous question and save results as a new csv file with the following name: “**world\_dev\_ind\_clean.csv**”
   1. *You will need to submit this dataset along with your final project report*
5. Using the provided *country\_metadata.csv* in combination with the datset you created as result of the question 4 of step 3 (*world\_dev\_ind\_clean.csv*), create a new dataset that contains the real country names (short names), their income group, and their region variables instead of the country code variable.
   1. *Hint: joining/merging, select attributes, export data, and store data. You only need the country names (short), income group, and region variables from the country\_metadata.csv*
6. Reorder the new dataset, so the country name, income group, and region are the first four columns in your new dataset. Save your dataset with this new variable order. Save the new resulting dataset as “**WorldIndicatorsComplete.csv**”
7. *You will need to submit this dataset along with your final project report*

### Step 4 – Data Modeling and Evaluation (20 points)

*Dataset*: WorldIndicatorsComplete.csv (you have created this dataset as result of step 3 completion)

#### Scenario: Imagine that we want to understand and later predict the GDP for each nation for any given year using available attributes.

1. What modeling technique(s) can help better understand what attribute(s) impact the GDP the most and why?
2. Create models using methods that you identified in your answer to the previous question.
   1. *In your report include screenshots of your RapidMiner process window and results tab*
   2. *Use as many screen shots as you deem necessary*
3. Using the results of those models, explain what the most important attributes in predicting GDP are?
4. Do your findings make business sense to you? Explain them.
5. Do your findings comply with the available knowledge/literature on this subject?
   1. *Hint: there is an abundance of literature on what affects GDP*
6. Are there localities in what affect GDP? If so, how are you going to account for these localities?
7. If we wish to move a model to deployment step and predict GDP, as opposed to merely understanding attributes impacting the GDP, which modeling technique gives us the best predictive performance? Explain your answer.
   1. *Compare performance measures for various* ***viable*** *modeling techniques to best answer this question. Your answer need to be technical relying on obtained performance measures and their comparison. General or non-technical answers would not be considered worthy of consideration and credit.*

### What to submit for Deliverable 2

* Data file world\_dev\_ind\_clean.csv (10 points)
* Data file WorldIndicatorsComplete.csv (10 points)
* Completed project document in either PDF or Microsoft word format. (80 points)