MidProject Guidelines:

1. Work should be organised using a Trello Board where you have the different parts of your project into tickets trackable of your progress.
2. Domain knowledge: you must come up with a business/domain question where data analytics will support and test your hypothesis.
3. You follow the data process lifecycle and structure your work according to that.
4. You will be evaluated from 1-5 and the grade you get will depend on:
    - Presentation of your project on Saturday 13th. The presentation should show an approach from general to technical. You will show your ability to present your work in a story that reveals your domain knowledge and your time management.
    - Dataset → understand the features → Ask questions that can be answered visually.
    - Data visualisation during your presentation: you must use the principle 1 plot per hypothesis to test/conclusion to reach.
    - Choose a KPI that you want optimise (performance metric that you can build yourselves or a metric that is relevant for your predictive model)
    - Pre-processing habits and arguments on the model selection.
    - Statistical studies conducted during the project (plots, descriptive analysis, outliers, confidence intervals)
    - Interpretation of ML results and performance metrics.
    - We will evaluate your code. Use the DRY (Do not repeat yourself) principle and program more functionally as you can when it makes sense. A couple of examples where you can use them: when you want to examine several models with different scalers, impact of outliers in your models or tune hyperparameters of the model.
    - You should have the code well documented and uploaded on your git. The readme file will be part of the evaluation as the full documentation of the dataset and your code. Here you have the opportunity to be as technical, methodical and detailed as you want presenting parts of the analysis that you did not have time to show us during your presentation. This deliverable should be available Monday 15th??.
    - Extra points if you can learn something new and show us this tool you decided to use.
    - **You need a minimum grade of 3 to pass. To get more than 4.5 you need to use a new model that we never used in class (but if you are able also to explain to us what is happening), new plot library in Python, Tableau.**

Minimal requirements:

- Understand your features.
- Build the use case.
- Pre-process the data well.
- Build a confidence interval for some quantity that is derived from the dataset (performance metric).
- Identify which ML metric is relevant to your business.
- Use Plotly
- Create a trello board that is readable and understandable.

Requirements to get from 3-4.5

- Perform a statistical study that makes sense (outlier analysis, conf intervals arguments, build smart deductions from plots)
- Creative feature generation that contains new info on the use case.
- Successfully perform a pre-processing pipeline ending on a ML model that predicts something useful with a decent result on the metric you choose.
- Your ability to be technical but not forgetting to focus on the product / use case.
- Code is readable and not too spaghetti code.
- You perform some optimisation on the parameters of model (for loop, GridSearch)

Requirement to get more than 4.5

- Being innovative but making sense (you can use new tools or models but you are able to explain them)
- Presentation needs to be balanced, interesting and general-technical in 12 min.

Data Analytics project (minimal steps):

- Dataset
- Identify the use case. Identify a storyline and a target variable (numerical → regression, categorical → classification). Identify the metric you want to optimise (RMSE, MSE, R2 score, Precision, recall, f1 score)....
- Understand and catalogue your features (if they are not understandable). Follow here the eg Wail and Estela on the wine dataset.
- Ask yourself questions on the dataset. One question → 1 plot
- Descriptive analysis (mean, std, mode, IQR, boxplots)
- Outliers? What are they in terms of the use case. Is it relevant to take them away or not ? Think in terms of the product but also test afterwards if they affect the performance of your model.
- **Preprocessing pipeline:**
    1. Cleaning the nans (erase the nans, Simple Imputer, knn imputer, other imputer)
    2. Impact on outliers
    3. Different scalers (MinMax, StandardScaler, RobustScaler, Polynomial features – feature transformation of your dataset)
    4. Feature selection (highly correlated features, features that you think make no sense…)
    5. Train, test splits and how you handle imbalance data (SMOTE, other technique, put more weight on the minority class or the class you want to predict)
    6. What is a good score? No answer for that. I will give you direct input on this because it will depend on how complex your dataset is.
    7. Train different models. You test different scenarios (scalers, outliers in or out, impute methods). YOU MUST FUNCTIONS TO AUTOMATE YOUR PROCESSES.
    8. Tune hyper parameters
    9. Explainability: some models have feature importance. You can run a couple of models that have this structure and you choose the best features that show up on more than 1 model.
    10. Build a nice presentation

Tuesday 02 April

1. Dataset, use case defined.
2. Trello Board to show where you have the different steps (high level) of what you are going to do.