Idea One

https://www.drivendata.org/competitions/56/predict-cleaning-time-series/

What is the business problem?

The cleanliness of production equipment is vitally important in the Food & Drug industry. Therefore, it is important to understand the amount of cleaning resources needed to meet industry cleanliness requirements.

Problem Statement: How can Schneider Electric reduce turbidity in the last rinse cycle to maintain an acceptable level of cleanliness, while still minimizing the use of water, energy, and time?

Who are the intended stakeholders, and why is this problem relevant to them?

- CEO
  - As head of the company, the CEO needs to be included in order to be aware of process improvements that affect the business.
- Chemical Laboratory Manager
  - As head of the lab, this stakeholder oversees the laboratory process. This person needs to be aware of any changes proposed after project completion.
- Head of QA
  - Since changes in turbidity are directly related to product quality, the head of QA needs to stay informed.

Where are the datasets available from?

- The dataset was available from the Schneider Electric website.

What data science approaches do you anticipate you will use to model the business problem as a data science problem?

- The goal of the project is to predict turbidity. This is a floating point, numerical value. Turbidity is an outcome variable and the rest of the dataset consists of features related to the outcome. Therefore, I would explore multiple regression models. I would anticipate using metrics such as R-squared and MAE to evaluate the models performance.

How do you anticipate that the intended clients will use the results of your CP2 to address the original business problem?

- The model will provide a tool where the client can experiment with the values of water, energy, and time used during the cleaning stage to predict the amount of turbidity. They could then set their turbidity standards depending on the companies requirements.
- Using the developed models we will provide the following to the client:
  (a) Analysis of the performance of the models, with respect to several performance metrics, such as R-squared, MAE, MAPE, and study of distribution of residuals to estimate upper/lower bounds of errors.
  (b) Analysis of impact of each of the independent variables (a.k.a. features) in the determination of target (turbidity)
  (c) Analysis of the importance of each of the features in the determination of the model

Idea Two

https://www.kaggle.com/c/santander-customer-transaction-prediction/overview

What is the business problem?

Santander provides many products to their customers to help them with their financial health. The company would like to understand their customer's buying trends better so that they can improve their financial health products.

Problem Statement: How can Santander help customers' financial health by understanding what types of customers make what types of transactions?

Who are the intended stakeholders, and why is this problem relevant to them?

- CEO
  - As head of the company, the CEO needs to be aware of any product changes.
- Head of Customer Service
  - Since this position works more directly with customers, it is important that this person is in the loop regarding any product changes provided to the customer.

Where are the datasets available from?

- The dataset was taken from Kaggle and was provided to Kaggle by Santander.

What data science approaches do you anticipate you will use to model the business problem as a data science problem?

- The goal of this project is to predict: A) Customer made a transaction, or B) Customer did not make a transaction. Hence, this is a binary classification problem, consisting of features and the binary outcome variable, transaction made. I will build and evaluate classification models to estimate the probability that a client will make a transaction or not. The evaluation metrics to be used will include: accuracy, precision, recal, F-1, and AUC..

How do you anticipate that the intended clients will use the results of your CP2 to address the original business problem?

- The modeling results will lead to an understanding of purchasing trends amongst customers. With this knowledge, Santander can provide the right products and services to specific groups of customers.
- The following deliverables will be presented to the client upon completion of the project:
  - A thorough analysis of classification models capable of predicting customer transactions decisions. The analysis will include AUC metrics to assess model performance, and impact/importance of the features in determining the likelihood of a customer making a transaction or not .
  - An analysis of customer profiles by grouping customers by type according to their transaction data.

Idea Three

https://www.drivendata.org/competitions/7/pump-it-up-data-mining-the-water-table/

What is the business problem?

Tanzania is a developing country and access to water is very important for the health of the population. For this reason, it is vital that all water pumps are properly working.

Problem Statement: How can the government of Tanzania improve water pump maintenance by knowing the pump functional status in advance?

Who are the intended stakeholders, and why is this problem relevant to them?

- Minister of Water, Hon. Jumaa H. Aweso
    - This is the most senior position in the government related to water issues. Therefore, he should be aware of improvements to water pump maintenance.
- Deputy Minister for Water, Hon. Maryprisca Mahundi (Mp)
    - This person works directly under the Minister of Water and should also be made aware of water pump maintenance improvements.
- Permanent Secretary, Eng. Anthony Sanga
    - The Permanent Secretary is in charge of the water management team. Since this person works more directly with individuals managing water issues, they should also be involved in any decisions related to water pump maintenance.

Where are the datasets available from?

- The datasets are available through DrivenData. They were provided to DrivenData by Taarifa and the Tanzanian Minister of Water.

What data science approaches do you anticipate you will use to model the business problem as a data science problem?

- The goal of this project is to classify water pump functional status into three groups using a set of features. Therefore, this is a supervised learning problem and I would build classification models to address this business problem.

How do you anticipate that the intended clients will use the results of your CP2 to address the original business problem?

- The Ministry of Water could be provided a web application to analyze the functional status of water pumps in the country. This information could be provided to the maintenance department so that they could fix any faulty pumps in a timely manner.
- At the end of this project, the client will be presented with the following deliverables:
    - An analysis of water pump status as it relates to water pump features.
    - A description of which features are important is predicting the outcome variable (in this case, water pump status).
    - A measure of feature impact as it is related to predicting water pump status.