DATA SCIENTIST
BUSINESS CASE.

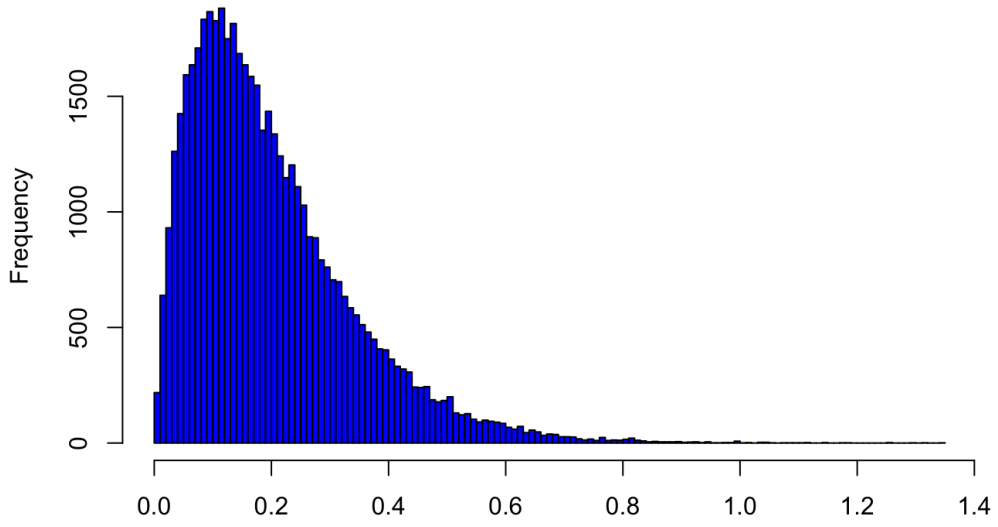
Welcome to the Atrato challenge for **Data Scientists.** In this challenge, we expect you to apply your knowledge of programming to analyze data, develop basic models, and draw meaningful insights. We value your ability to think critically, solve problems, and communicate your findings effectively.

- It is important to have a basic understanding of Machine Learning algorithms and their applications. Please answer the following questions:

    a. Describe briefly each of the following algorithms and its typical use cases:
        i. Logistic Regression.
        ii. Multiple Linear Regression.
        iii. Random Forest
        iv. XGBoost
        v. Genetic Algorithms.
        vi. KMeans
    b. What is the difference between supervised and unsupervised learning algorithms?
        i. Describe the types of supervised and unsupervised learning.
    c. Are you familiar with any techniques for algorithm evaluation and performance metrics?

- We would like to assess your understanding of statistical concepts.

1. Observe the distribution plot of the following variable:



    a. What can you infer about the data's asymmetry? Would you say the distribution is symmetric, skewed to the right (positively skewed), or skewed to the left (negatively

skewed)? Please provide a brief explanation of your answer based on the shape of the plot.

b. In the distribution plot, what can you infer about the data's concentration? Does the distribution exhibit a pronounced peak or is it more flattened?

Extra points:

a. Are you familiar with any technique to reduce bias in the distribution of a variable? If so, please mention the technique

b. How would you apply the technique to a DataFrame in Python?

- Exam Approval Prediction

Introduction:

In this challenge, you will tackle the task of predicting the probability that a student will pass a grade. As a data scientist, you will have to choose and apply the best algorithm to build a predictive model.

Context:

Imagine you are part of a data science team working for an educational institution. The team is tasked with developing a predictive model that can assist in identifying students who are likely to pass or fail the grade. Such a model can provide valuable insights into student performance and help in designing targeted interventions to support struggling students.

Use G3 > 12: 1 else 0.

Tasks:

- Load and explore the dataset
- Visualize the relationships:
  - Bivariate analysis.
  - Correlation matrix.
  - Others
- Normalize or standardize features if necessary.
- Build a predictive model.
- Train the model.
- Assess the model's performance using metrics such as accuracy, confusion matrix, and classification report.
- Interpret the results of the model.
- Communicate conclusions regarding the founding relationships.
- Provide actionable recommendations based on the analysis.

Dataset: https://archive.ics.uci.edu/dataset/320/student+performance

- Share a diagram that shows an end to end pipeline data science project from the experimentation to a productive environment.

You can use https://www.drawio.com/
Extra points:

- How to integrate DVC in the pipeline.
- How to integrate MLFLOW in the pipeline.

- Can you explain and give examples about this following concepts:
    - Encapsulation
    - Abstraction
    - Inheritance
    - Polymorphism

Remember, this challenge is an opportunity for growth and learning. Embrace the challenge, be curious, and demonstrate your enthusiasm for becoming a skilled Data Scientist.

Good luck with the challenge!