

# ANALYSIS AND DESIGN OF A MODULAR PREDICTIVE SYSTEM FOR THE "SWEET REGRESSION COMPETITION"

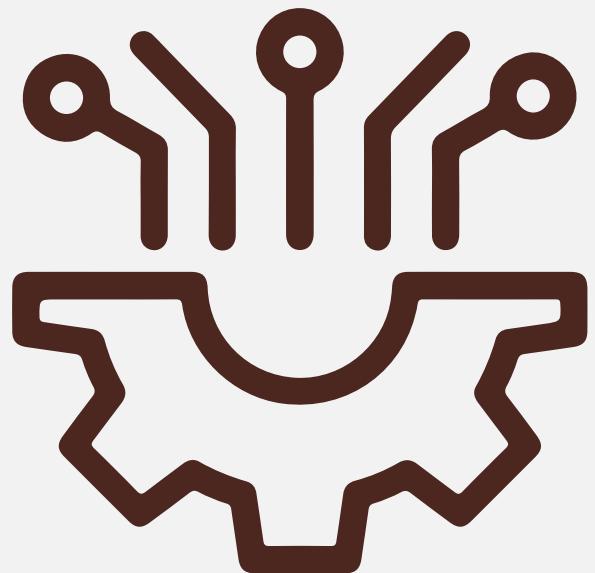
Samuel Aljure Bernal - 20202020111

Carlos Alberto Barriga Gámez - 20222020179

David Santiago Aldana Gonzalez - 20222020158

Juan Diego Álvarez Cristancho - 20221020076

# THE PROBLEM: THE "SWEET REGRESSION COMPETITION"



## OBJECTIVE:

Develop a system to predict chocolate sales for the "Chocolates 4U" company as it expands into new regions.

## INPUT DATA:

The model must process marketing (GRP), regional, and consumer characteristic variables.



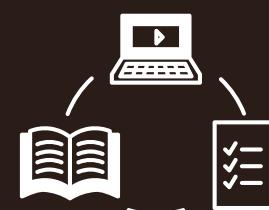


# THE CHALLENGE: CRITICAL SYSTEM CONSTRAINTS

Our systemic analysis (Workshop #1) identified 3 key constraints:



**Technical:** The final solution must be implemented in R.



**Operational:** A strict limit of 10 submission attempts.



**Evaluation:** Grading is based on a relative reference model, where the best submission sets the top score



# THE RISK: SENSITIVITY AND CHAOTIC BEHAVIOR

- These constraints introduce high sensitivity.
- Small variations in data or model configuration can drastically alter the results, a behavior consistent with "chaos".
- Algorithmic randomness and data noise impact stability.
- The Need: A robust design that controls uncertainty and ensures reproducibility.

# THE PROPOSED SOLUTION: A HYBRID MODULAR SYSTEM

We propose an architecture that strategically separates analysis from implementation:

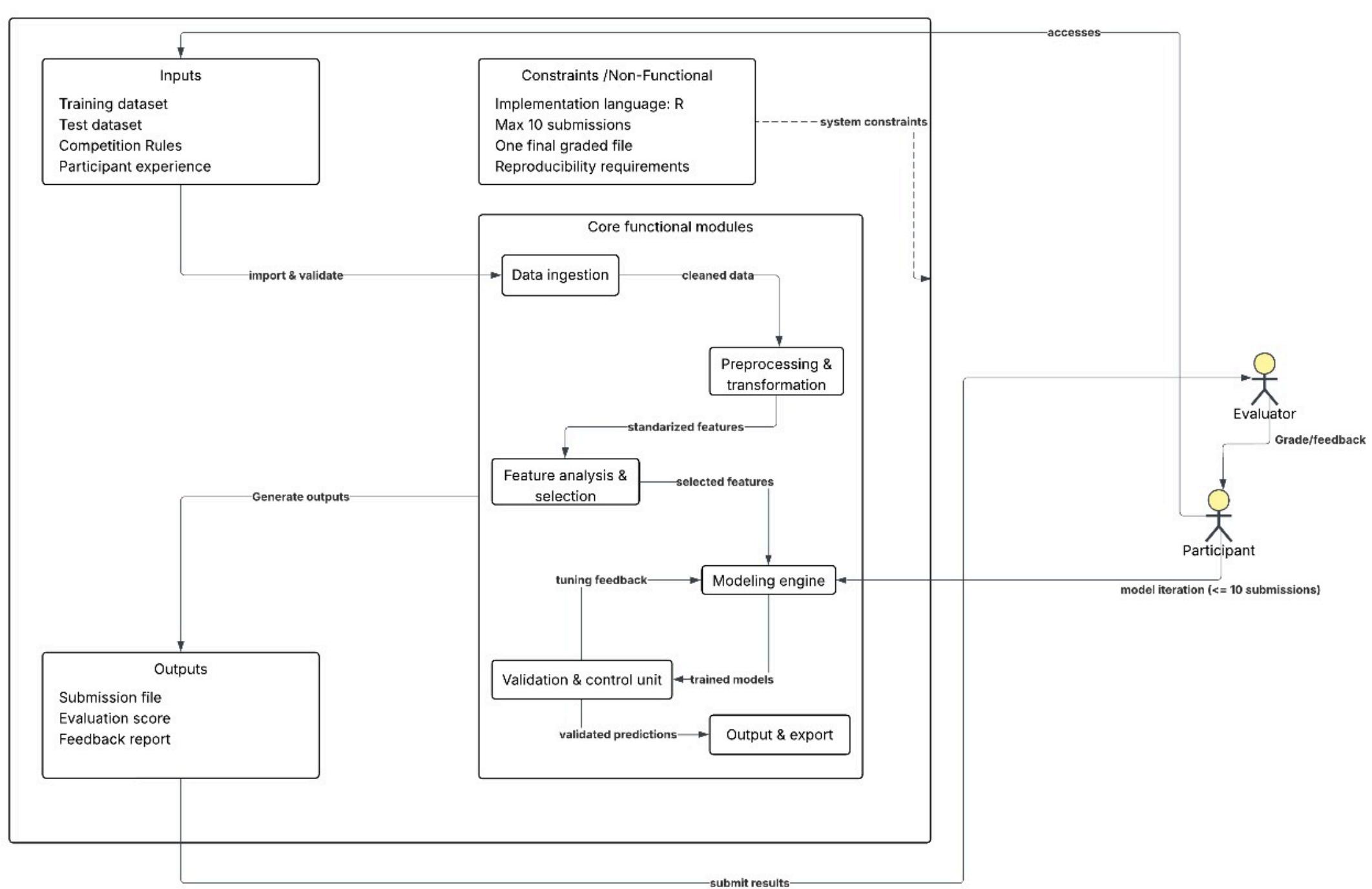
# Phase 1 (Completed): Exploratory & Statistical Analysis

We used Python (Pandas, Matplotlib, Scikit-learn) to explore the data, find correlations, and run ANOVA tests.

# Phase 2 (Future Work): Modeling & Delivery

We will use R (with caret, xgboost) to build, validate, and export the final model, fulfilling the competition's constraint.

# SYSTEM ARCHITECTURE (THE DESIGN)



The design uses 6 independent modules to ensure control and traceability:

- Data Ingestion
- Preprocessing & Transformation
- Feature Analysis & Selection
- Modeling Engine
- Validation & Control Unit
- Output and export unit

# THE "MODELING ENGINE" (THE PLAN FOR R)

This slide explains Phase 2 (the future work in R).  
The engine's design (not yet implemented) will allow us to:

## N° 1

Compare multiple techniques (Linear, Lasso, Random Forest, XGBoost) to find the lowest Mean Absolute Error (MAE).

## N° 2

Control overfitting using cross-validation and early-stopping mechanisms

## N° 3

Ensure reproducibility (key for "chaos") by fixing random seeds.

# RESULTS OF STATISTICAL ANALYSIS (PHASE 1 IN PYTHON)

The work in this phase is conceptual, not an experimental model; it is the analysis that justifies our design.

**Finding 1: A Dominant Predictor.**

$R = 0.98$

The Time\_in\_Region variable has a near-perfect positive linear correlation with sales.

**Finding 2: Weak Linear Correlation.**

$R = 0.09$

Facebook\_GRP

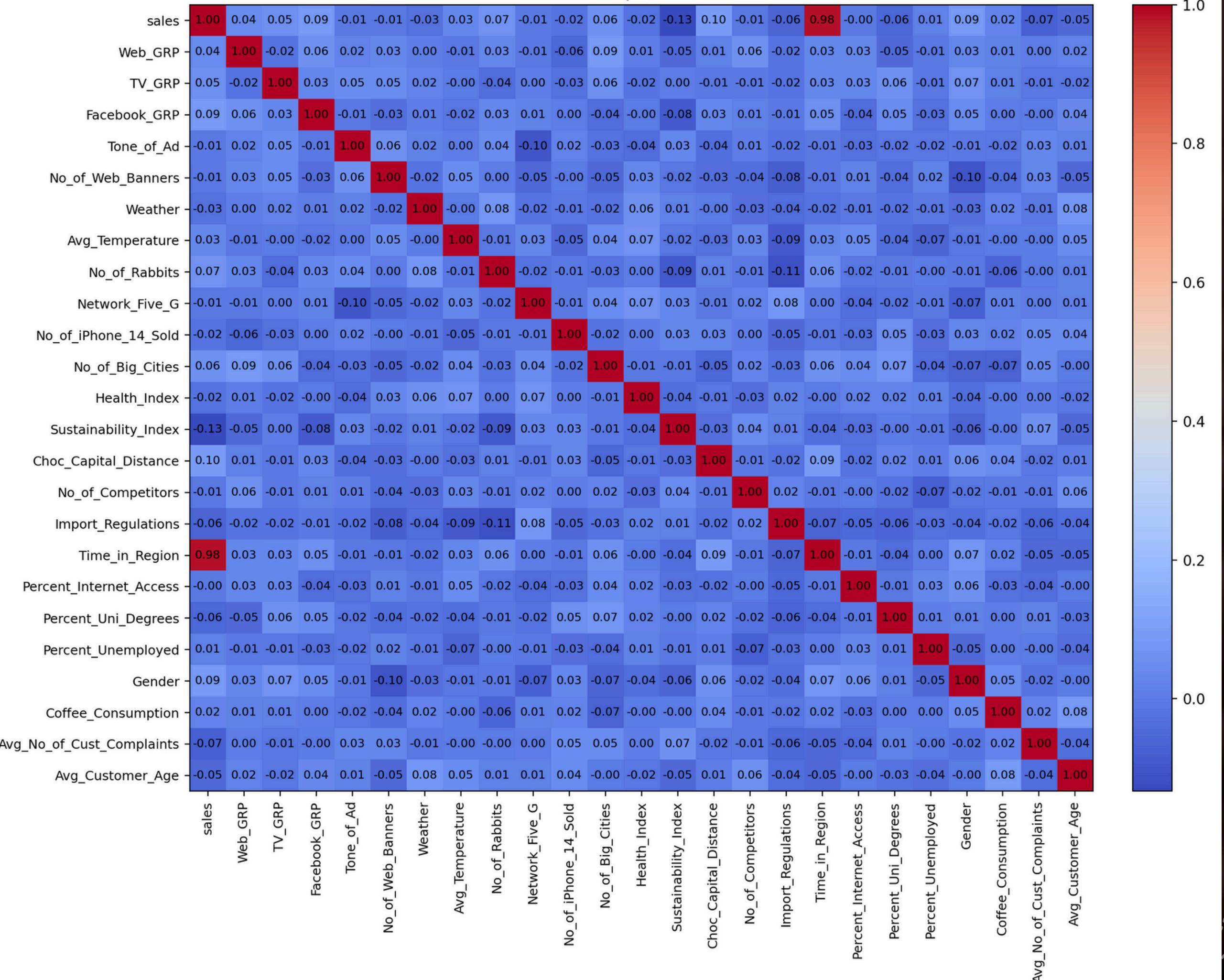
$R = 0.10$

Choc\_Capital\_Distance

$R = 0.09$

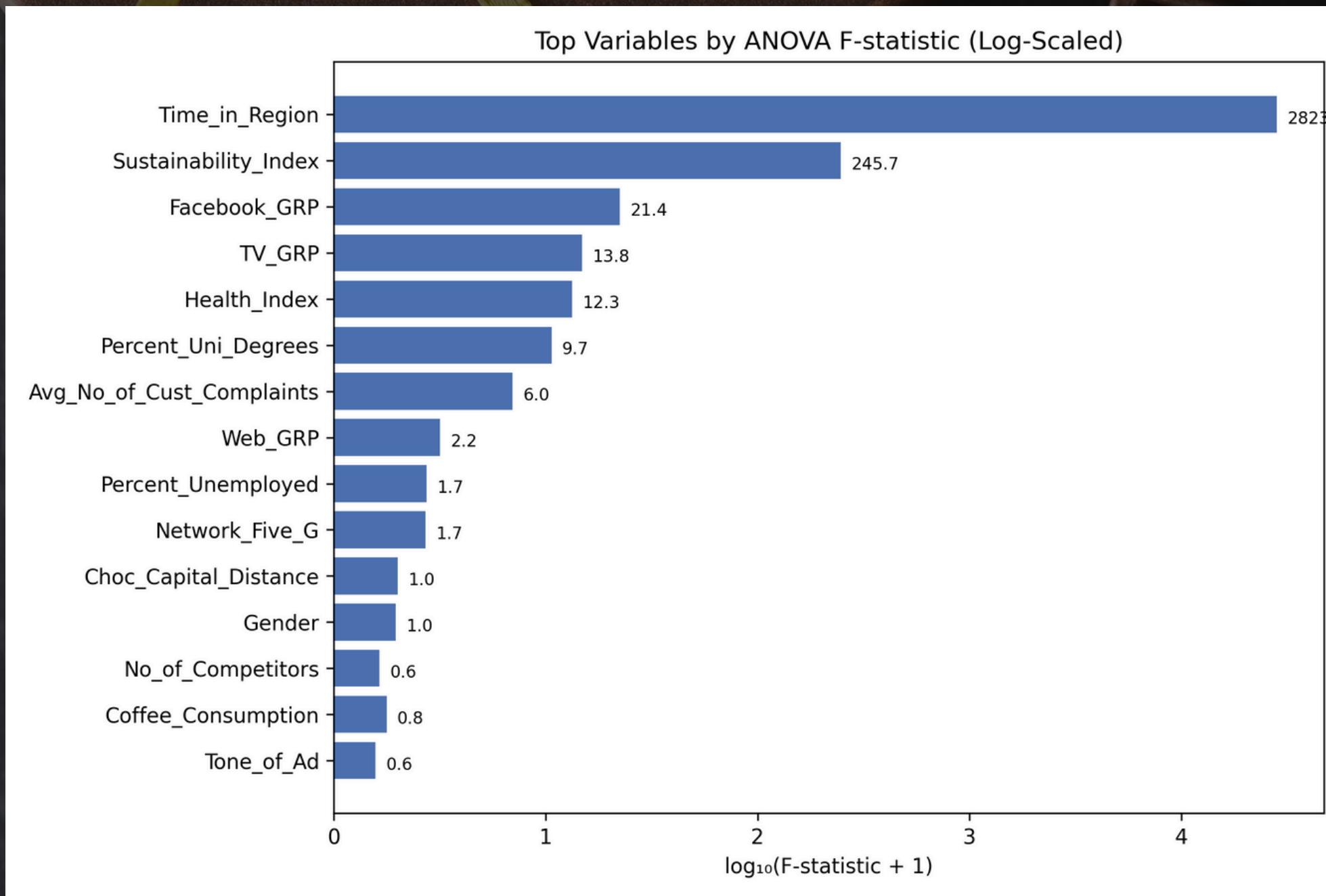
Gender

Correlation Heatmap (Numerical Features)



# RESULTS (VARIABLE IMPORTANCE - ANOVA)

The ANOVA (ANalysis Of VAriance) analysis quantifies the statistical influence of each variable:



## Extreme Influence:

Time\_in\_Region (log-scaled value: 2823.0).

## High Influence:

Sustainability\_Index (245.7), Facebook\_GRP (21.4), and TV\_GRP (13.8).

## Low Impact:

Tone\_of\_Ad, Coffee\_Consumption (Candidates for model exclusion).



## CONCLUSIONS & FUTURE WORK

The statistical analysis (Phase 1) confirms Time\_in\_Region is the dominant predictor. However, it also shows that key marketing variables (like Facebook\_GRP) have a weak linear influence , even though the ANOVA analysis proves they are statistically relevant.

**Completed:** The systemic analysis, architecture design, and exploratory statistical analysis (Python).

**The Design:** This is a conceptual "blueprint" for a robust, modular, and reproducible system, ready for implementation.

**Future Work (Phase 2):**  
Implement the 6 modules and execute the Modeling Engine in R to validate the design and compete.



# THANK YOU

