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.

.CSV

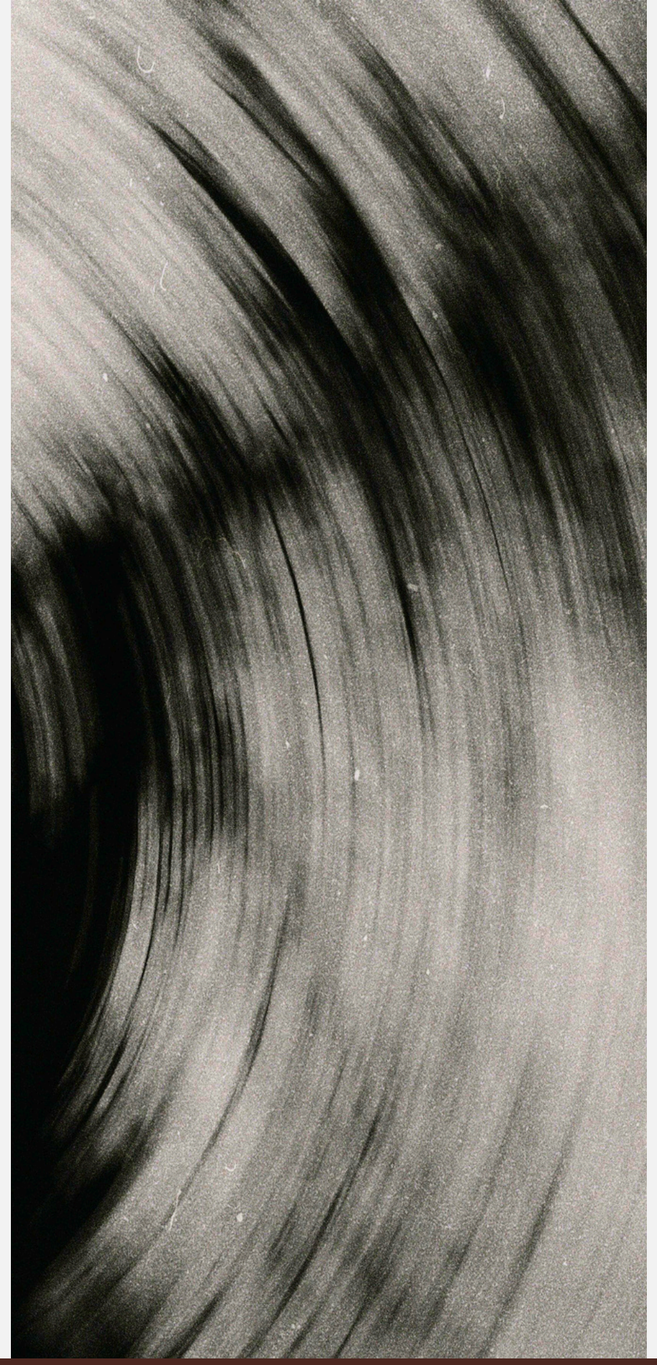
THE CHALLENGE: CRITICAL SYSTEM CONSTRAINTS

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

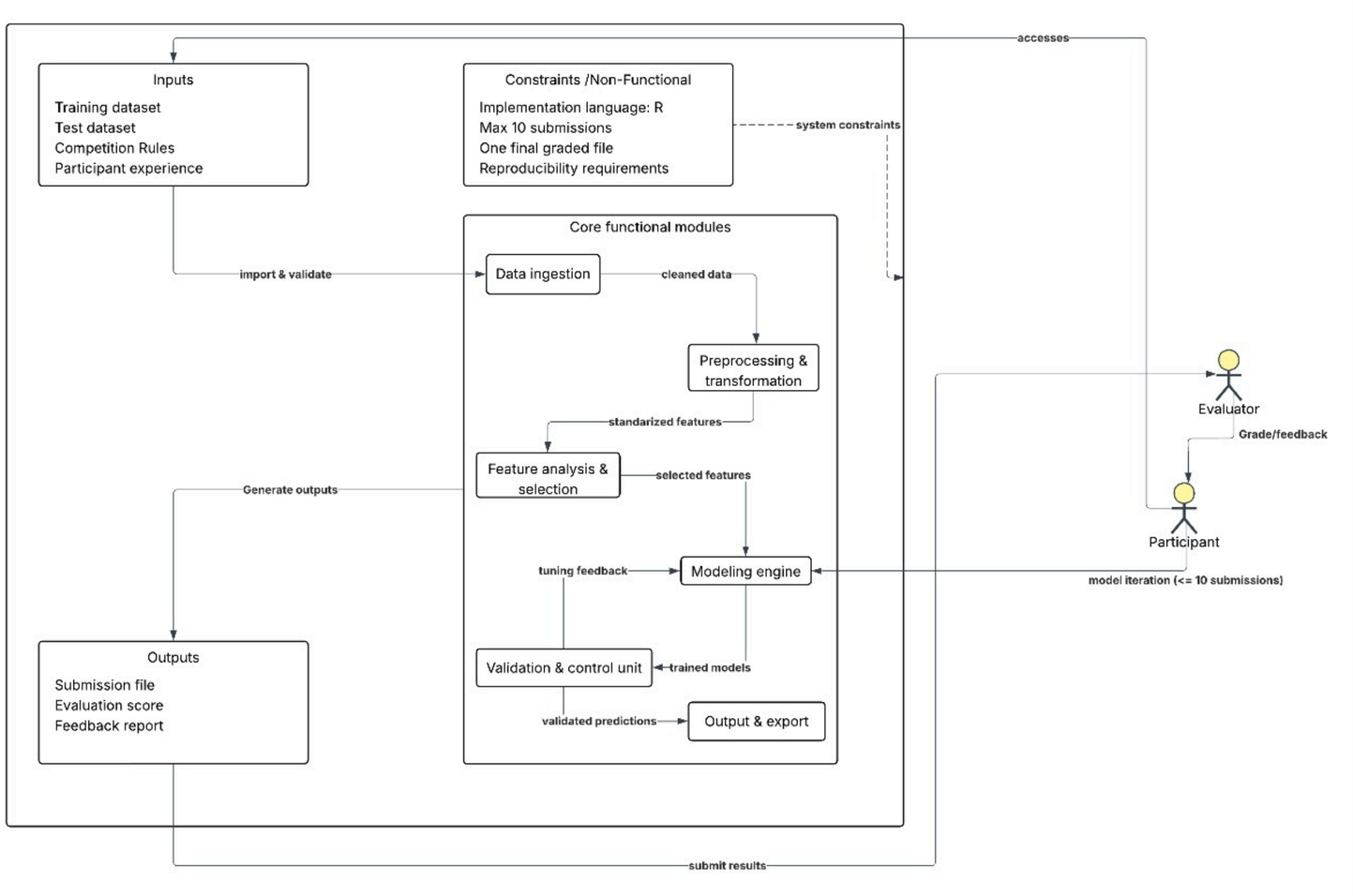
We propose an architecture that strategically separates analysis from implementation:

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

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

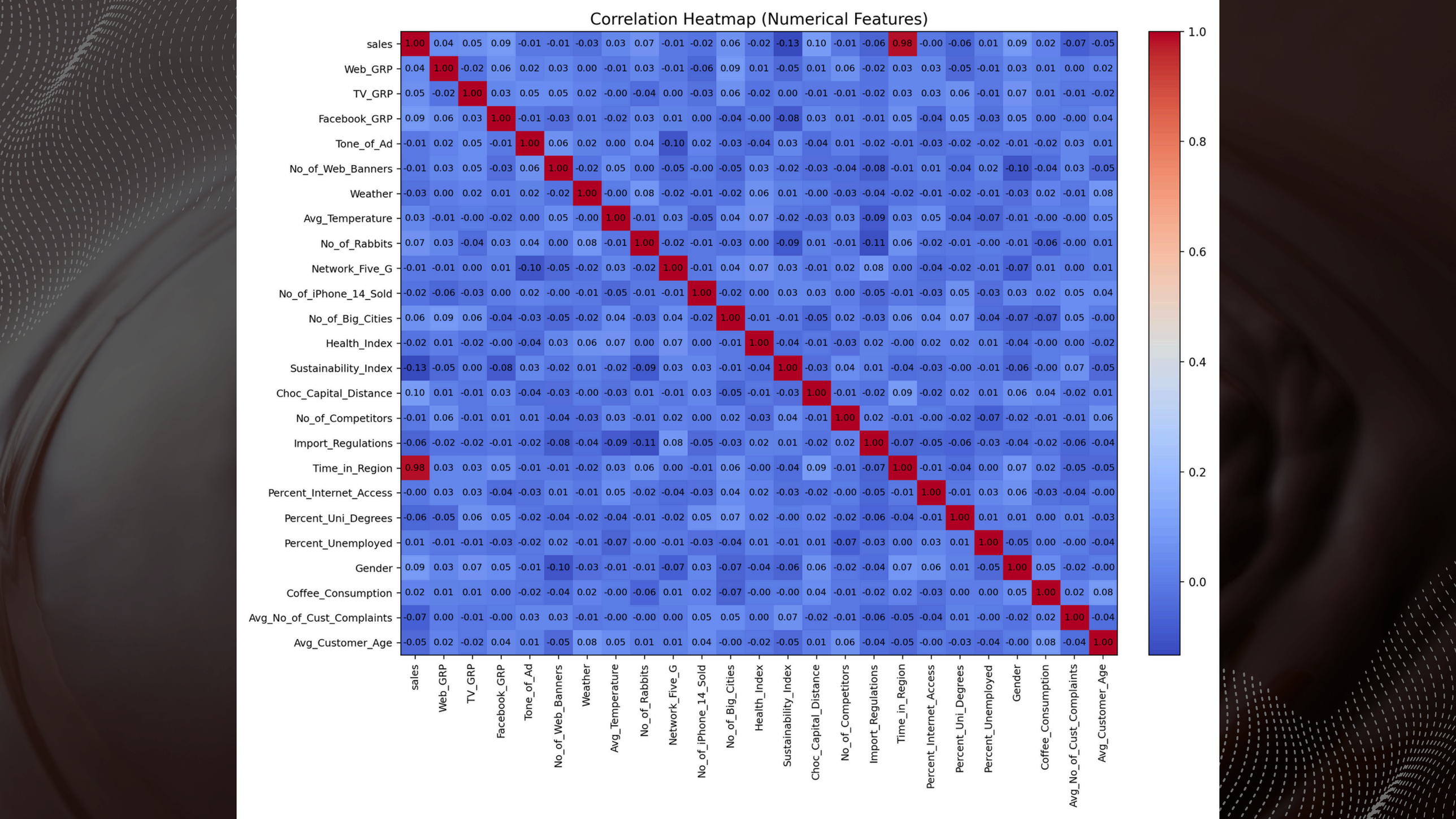
Facebook_GRP

R = 0.10

Choc_Capital_ Distance

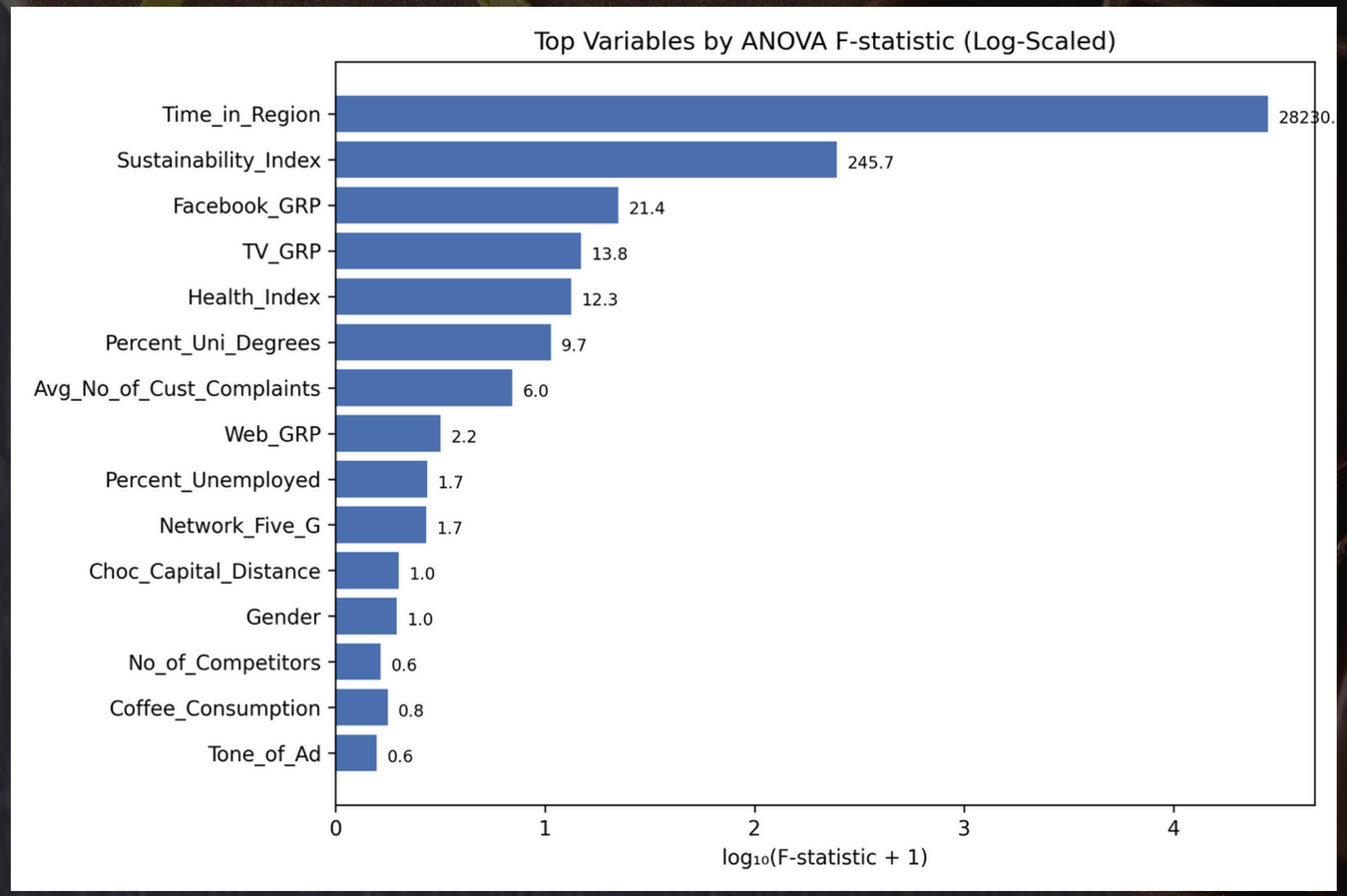
R = 0.09

Gender


Correlation Heatmap (Numerical Features)
sales 1.00 0.04 0.05 0.09 -0.01 -0.01 -0.03 0.03 0.07 -0.01 -0.02 0.06 -0.02 -0.13 0.10 -0.01 -0.06 0.98 -0.00 -0.06 0.01 0.09 0.02 -0.07 -0.05
Web_GRP 0.04 1.00 -0.02 0.06 0.02 0.03 0.00 -0.01 0.03 -0.01 -0.06 0.09 0.01 -0.05 0.01 0.06 -0.02 0.03 0.03 -0.05 -0.01 0.03 0.01 0.00 0.02
TV_GRP 0.05 -0.02 1.00 0.03 0.05 0.05 0.02 -0.00 -0.04 0.00 -0.03 0.06 -0.02 0.00 -0.01 -0.01 -0.02 0.03 0.03 0.06 -0.01 0.07 0.01 -0.01 -0.02
Facebook_GRP 0.09 0.06 0.03 1.00 -0.01 -0.03 0.01 -0.02 0.03 0.01 0.00 -0.04 -0.00 -0.08 0.03 0.01 -0.01 0.05 -0.04 0.05 -0.03 0.05 0.00 -0.00 0.04
Tone_of_Ad -0.01 0.02 0.05 -0.01 1.00 0.06 0.02 0.00 0.04 -0.10 0.02 -0.03 -0.04 0.03 -0.04 0.01 -0.02 -0.01 -0.03 -0.02 -0.02 -0.01 -0.02 0.03 0.01
No_of_Web_Banners -0.01 0.03 0.05 -0.03 0.06 1.00 -0.02 0.05 0.00 -0.05 -0.00 -0.05 0.03 -0.02 -0.03 -0.04 -0.08 -0.01 0.01 -0.04 0.02 -0.10 -0.04 0.03 -0.05
Weather -0.03 0.00 0.02 0.01 0.02 -0.02 1.00 -0.00 0.08 -0.02 -0.01 -0.02 0.06 0.01 -0.00 -0.03 -0.04 -0.02 -0.01 -0.02 -0.01 -0.03 0.02 -0.01 0.08
Avg_Temperature 0.03 -0.01 -0.00 -0.02 0.00 0.05 -0.00 1.00 -0.01 0.03 -0.05 0.04 0.07 -0.02 -0.03 0.03 -0.09 0.03 0.05 -0.04 -0.07 -0.01 -0.00 -0.00 0.05
No_of_Rabbits 0.07 0.03 -0.04 0.03 0.04 0.00 0.08 -0.01 1.00 -0.02 -0.01 -0.03 0.00 -0.09 0.01 -0.01 -0.11 0.06 -0.02 -0.01 -0.00 -0.01 -0.06 -0.00 0.01
Network_Five_G -0.01 -0.01 0.00 0.01 -0.10 -0.05 -0.02 0.03 -0.02 1.00 -0.01 0.04 0.07 0.03 -0.01 0.02 0.08 0.00 -0.04 -0.02 -0.01 -0.07 0.01 0.00 0.01
No_of_iPhone_14_Sold -0.02 -0.06 -0.03 0.00 0.02 -0.00 -0.01 -0.05 -0.01 -0.01 1.00 -0.02 0.00 0.03 0.03 0.00 -0.05 -0.01 -0.03 0.05 -0.03 0.03 0.02 0.05 0.04
No_of_Big_Cities 0.06 0.09 0.06 -0.04 -0.03 -0.05 -0.02 0.04 -0.03 0.04 -0.02 1.00 -0.01 -0.01 -0.05 0.02 -0.03 0.06 0.04 0.07 -0.04 -0.07 -0.07 0.05 -0.00
Health_Index -0.02 0.01 -0.02 -0.00 -0.04 0.03 0.06 0.07 0.00 0.07 0.00 -0.01 1.00 -0.04 -0.01 -0.03 0.02 -0.00 0.02 0.02 0.01 -0.04 -0.00 0.00 -0.02
Sustainability_Index -0.13 -0.05 0.00 -0.08 0.03 -0.02 0.01 -0.02 -0.09 0.03 0.03 -0.01 -0.04 1.00 -0.03 0.04 0.01 -0.04 -0.03 -0.00 -0.01 -0.06 -0.00 0.07 -0.05
Choc_Capital_Distance 0.10 0.01 -0.01 0.03 -0.04 -0.03 -0.00 -0.03 0.01 -0.01 0.03 -0.05 -0.01 -0.03 1.00 -0.01 -0.02 0.09 -0.02 0.02 0.01 0.06 0.04 -0.02 0.01
No_of_Competitors -0.01 0.06 -0.01 0.01 0.01 -0.04 -0.03 0.03 -0.01 0.02 0.00 0.02 -0.03 0.04 -0.01 1.00 0.02 -0.01 -0.00 -0.02 -0.07 -0.02 -0.01 -0.01 0.06
Import_Regulations -0.06 -0.02 -0.02 -0.01 -0.02 -0.08 -0.04 -0.09 -0.11 0.08 -0.05 -0.03 0.02 0.01 -0.02 0.02 1.00 -0.07 -0.05 -0.06 -0.03 -0.04 -0.02 -0.06 -0.04
Time_in_Region 0.98 0.03 0.03 0.05 -0.01 -0.01 -0.02 0.03 0.06 0.00 -0.01 0.06 -0.00 -0.04 0.09 -0.01 -0.07 1.00 -0.01 -0.04 0.00 0.07 0.02 -0.05 -0.05
Percent_Internet_Access -0.00 0.03 0.03 -0.04 -0.03 0.01 -0.01 0.05 -0.02 -0.04 -0.03 0.04 0.02 -0.03 -0.02 -0.00 -0.05 -0.01 1.00 -0.01 0.03 0.06 -0.03 -0.04 -0.00
Percent_Uni_Degrees -0.06 -0.05 0.06 0.05 -0.02 -0.04 -0.02 -0.04 -0.01 -0.02 0.05 0.07 0.02 -0.00 0.02 -0.02 -0.06 -0.04 -0.01 1.00 0.01 0.01 0.00 0.01 -0.03
Percent_Unemployed 0.01 -0.01 -0.01 -0.03 -0.02 0.02 -0.01 -0.07 -0.00 -0.01 -0.03 -0.04 0.01 -0.01 0.01 -0.07 -0.03 0.00 0.03 0.01 1.00 -0.05 0.00 -0.00 -0.04
Gender 0.09 0.03 0.07 0.05 -0.01 -0.10 -0.03 -0.01 -0.01 -0.07 0.03 -0.07 -0.04 -0.06 0.06 -0.02 -0.04 0.07 0.06 0.01 -0.05 1.00 0.05 -0.02 -0.00
Coffee_Consumption 0.02 0.01 0.01 0.00 -0.02 -0.04 0.02 -0.00 -0.06 0.01 0.02 -0.07 -0.00 -0.00 0.04 -0.01 -0.02 0.02 -0.03 0.00 0.00 0.05 1.00 0.02 0.08
Avg_No_of_Cust_Complaints -0.07 0.00 -0.01 -0.00 0.03 0.03 -0.01 -0.00 -0.00 0.00 0.05 0.05 0.00 0.07 -0.02 -0.01 -0.06 -0.05 -0.04 0.01 -0.00 -0.02 0.02 1.00 -0.04
Avg_Customer_Age -0.05 0.02 -0.02 0.04 0.01 -0.05 0.08 0.05 0.01 0.01 0.04 -0.00 -0.02 -0.05 0.01 0.06 -0.04 -0.05 -0.00 -0.03 -0.04 -0.00 0.08 -0.04 1.00
sales Web_GRP TV_GRP Facebook_GRP Tone_of_Ad No_of_Web_Banners Weather Avg_Temperature No_of_Rabbits Network_Five_G No_of_iPhone_14_Sold No_of_Big_Cities Health_Index Sustainability_Index Choc_Capital_Distance No_of_Competitors Import_Regulations Time_in_Region Percent_Internet_Access Percent_Uni_Degrees Percent_Unemployed Gender Coffee_Consumption Avg_No_of_Cust_Complaints Avg_Customer_Age
1.0
0.8
0.6
0.4
0.2
0.0

RESULTS (VARIABLE IMPORTANCE - ANOVA)

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



Extreme Influence:

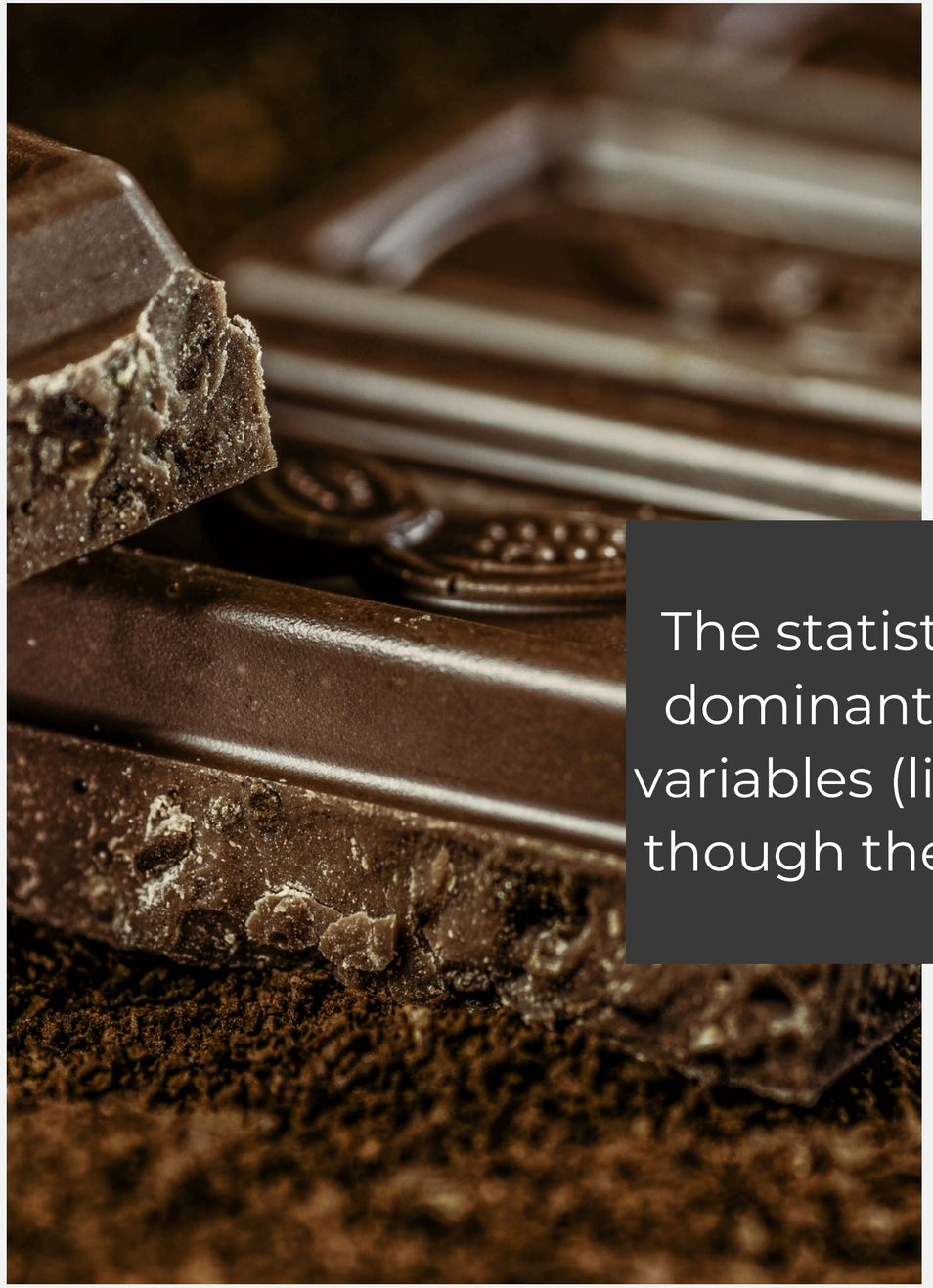
Time_in_Region (log-scaled value: 2823.0).

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