Healthcare A/B Analysis and Data Pipeline Project Report

Project Overview

This project demonstrates the implementation of an end-to-end data pipeline using DBT, Snowflake, and Airflow, focused on healthcare claims data. The core objective was to conduct A/B testing on claim payment data and showcase proficiency in data engineering and analytics workflows.

Technologies Used

- DBT (Data Build Tool) – For data transformation and modeling.
- Snowflake – For data storage and processing.
- Apache Airflow – For orchestrating the pipeline.
- Python & Jupyter Notebooks – For A/B testing and statistical analysis.
- GitHub – For version control and project publication.

Key Steps and Execution

1. Data Pipeline Setup

- Configured DBT models for staging, intermediate, and marts layers.
- Defined transformations and built structured data marts.
- Implemented Airflow DAGs to automate the pipeline execution.

2. A/B Testing on Healthcare Claims

• **Objective:** Analyze whether there is a significant difference in **total paid amounts** between two test groups.

• **Steps Taken:**

a. Loaded FACT_AB_TEST_DATA from Snowflake.
b. Performed data exploration (missing values, distribution analysis).
c. Defined hypotheses:
d. H_0 (Null Hypothesis): No significant difference between Group A and Group B.
e. H_1 (Alternative Hypothesis): A significant difference exists.
f. Conducted a normality test (Shapiro-Wilk test) – determined data was not normally distributed.
g. Used the Mann-Whitney U test for comparison – concluded no significant difference between groups.
h. Visualized data distribution using box plots and histograms.

3. Fraud Detection Analysis (Abandoned)

• Initially planned fraud detection using FACT_FRAUD_RISK, but later dropped from scope due to time constraints.

• The focus was redirected to publishing the project and documenting it properly.

4. Project Publication

• Organized GitHub repository: Healthcare A/B Analysis Pipeline

• Created a structured README.md with:

• Project overview

• Technologies used

• Execution steps

• Code and analysis breakdown

Conclusion

This project successfully showcases the ability to work with DBT, Snowflake, and Airflow, process healthcare claims data, and conduct A/B testing. While the fraud detection portion was dropped, the main goal of demonstrating ETL, data transformation, and statistical analysis was met.

Future improvements could include automated reporting, model deployment, and further insights into healthcare claims optimization.