Data Quality Monitoring and Preparation – Dataiku Workshop (Part 1) - Week 5

Name: Joori Ahmed Shareef

Company: Jedco

Department: Business Applications – Data Center 2

Duration: 20 hours

1. Objectives

This week's objective was to build a solid understanding of data quality monitoring and preparation critical components in developing AI and analytics pipelines. As part of the fifth training item, titled "Introduction to Data Quality Monitoring and Common Issues in Aviation Datasets," I participated in a hands-on workshop using the Dataiku platform. The training focused on identifying common data quality issues, cleaning and consolidating datasets, and preparing them for analysis. These activities closely aligned with the core learning objectives set out in the original plan.

2. Activities Completed

We were enrolled in first part of the Google Cloud AI & Data Science Workshop via Dataiku, delivered over two sessions (Tuesday and Wednesday). This phase focused on preparing datasets in real-world scenarios:

- Dataset Import & Project Setup: I learned how to create new projects on the Dataiku platform and connect various data sources, such as CSV files and SQL tables.
- Data Stacking and Joining: We practiced consolidating different datasets using stack and join operations, a crucial step before beginning any modeling task.
- Data Cleaning: I explored various techniques to remove invalid rows, detect outliers, and parse date fields and geolocation data (latitude/longitude columns). This closely resembles what's needed in aviation datasets, which often contain missing or corrupted flight logs or passenger info.
- Geographical Feature Extraction: I learned how to extract location-based attributes such as merchant state and cardholder state, which can be paralleled to extracting airport or region-based indicators in aviation.
- Categorical Data Enrichment: We enriched data by creating meaningful features for example, flagging essential vs. non-essential purchases, and calculating a risk level score using conditional formulas (if-then-else logic).

3. Project Output

At the end of the week, I implemented a complete data preparation pipeline using a credit card fraud dataset. I documented each step and uploaded the project, including screenshots and analysis summaries, to my GitHub repository. The public link will be included in the final training report.

4. Key Learnings

- Learned how to handle real-world messy datasets and prepare them for model training.
- Practiced data quality inspection using automation tools provided by Dataiku.
- Understood the importance of structured, clean, and enriched data in any successful AI system.
- Recognized the strong link between what we practiced and the original item in the training plan, even if not covered directly by the company.