

**Data Design Specialist Technical Assessment- 2024**

**About Us**

Godoki is a social enterprise that improves access to high-quality, affordable early childhood care & education in East Africa’s low-income communities. We partner with local women(“Mamapreneurs”) to start or grow their own childcare centers through a social franchising approach. We provide ‘Mamapreneurs’ with training, mentorship and support tools so that they can give children in their care the best start to life, while their mothers go to work for the day. Godoki currently operates in 90+ communities across Twelve(12) counties in Kenya and plans to scale even further.

**A: Task Description:**

Godoki operates on a social franchising model, where registered daycare centers (referred to as "Mamapreneurs") pay a monthly fee to remain active and receive benefits based on their center type: **Home, Center, or School**. Each daycare is assigned a unique identifier called **Center\_ID**, which is used to identify their records in the database and also make payments via the organization’s Mpesa Paybill system.

As a **Data Design Specialist** at Godoki, your role involves collaborating with the finance department to process Mpesa statements, clean the data, and cross-reference it with the center details in the database. This ensures that payments are correctly attributed to the corresponding daycare centers. Your deliverables include:

1. Providing a clean dataset in a usable format, as requested.
2. Creating a report in the form of an interactive dashboard (preferably in Power BI).

You have been provided with an Excel file titled **Data\_Designs\_Technical\_Assessment\_2024**, which contains the following three tabs:

1. **AllCenters**: This tab includes detailed records for all registered daycares enrolled in Godoki's franchising model. Each row corresponds to a specific daycare and its associated information.
2. **Collections\_Raw**: This tab contains raw Mpesa Paybill transaction data for the period May 2024 to October 2024. Each transaction is identified by a receipt number and includes payment details, a string of text that contains information such as ***“Acc. 1234”*** , which is the unique identifier called **Center\_ID** used to allocate payments to the respective centers.
   * **Notes:**
     + A single center can make multiple payments in a month (e.g., installment payments).
     + Some centers may have prepaid for future months (e.g., Center 1234 paying KES 3000 in June to cover fees for June, July, and August).
3. **README.md**: This tab serves as a data dictionary, providing detailed descriptions of the columns and structure in the **AllCenters** and **Collections\_Raw** tabs.

Your task is to use your data processing and reporting expertise to clean, analyze, and organize the data, ensuring accurate payment allocation and providing the team with actionable reports via a Power BI dashboard.

**B: Procedure:**

**Step 1:** Using Python programming language, Google Collab and Google Sheets **(10 points)**

Download the dataset and Upload it to your own google drive where you have the edit rights. You are required to Upload the excel file provided (Data\_Designs\_Technical\_Assessment\_2024) in a google sheet  **and then write two python functions, one that loads the data from Google sheet into Pandas dataframe,** and **another that will export your finalized output into a different tab on the same workbook**. Use default authentication for the google drive.

**Step 2:** Calculator **ExpectedMonthlyFee** for each center **( 20 points)**

The "**ExpectedMonthlyFee** " for member centers varies based on several parameters:

1. If the Pilot Program is **"Honoring Our Commitment"**:  
   * The fee is a flat rate determined by the center type:
     + **Center-Based**: KES 1000
     + **Home-Based**: KES 500
     + **School-Based**: KES 1500
2. If the Pilot Program is **"Tier\_Pilot MMPs"**:  
   * The fee depends on the **Pilot Model, Pilot Package**, , and **total\_children.**
   * First, identify the **Pilot Model** (Flat Rate or Banded Rate).

**2.1 Flat Rate Fees**:

* + **Bronze Package**: KES 500
  + **Silver Package**: KES 1000
  + **Gold Package**: KES 1500

**2.2 Banded Rate Fees**:

* + **Bronze Package**: KES 400 for the first 10 children, plus KES 50 for each additional child above 10.
  + **Silver Package**: KES 750 for the first 15 children, plus KES 50 for each additional child above 15.
  + **Gold Package**: KES 1000 for the first 15 children, plus KES 50 for each additional child above 15.

Using the df loaded from **AllCenters** tab, write a python function that would calculate the **ExpectedMonthlyFee** for each center based on the above context. Using the functions written in task 1, export the resulting df in the same google sheet **Data\_Designs\_Technical\_Assessment\_2024** on a new tab calledStep\_1&2. **(15 points)**

**Reflection:** In your collaboration notebook, document any data quality issues or errors encountered during **Step 2.** Include details about the issues identified and the steps you took to resolve them to ensure your code executed successfully. **(5 points)**

**Step 3:** Data Munging **(40 points)**

**3.1** Using the DataFrame loaded from the **Collections\_Raw** tab, write a Python function to extract account numbers (GodokiCode) from the column named "details". The extracted values should be stored in a new column named Center\_ID.

**(5 points)**

***Hint:*** *Use regular expressions (REGEX) to identify and extract specific patterns from the text string.*

**3.2** While cross-referencing tables, some transactions in the **Collections\_Raw** tab have account numbers (Center\_ID) that are not present in the **AllCenters** tab. These transactions are considered payments to incorrect account numbers. Filter out these records and export the resulting DataFrame to the same Google Sheet, **Data\_Designs\_Technical\_Assessment\_2024**, in a new tab named **"Wrong Account Payments"**. **(5 points)**

**3.3**  Using the resulting DataFrame from Step 2, which includes Center\_ID, center details, and Monthly Franchising Fees, along with the DataFrame from Step 3.1, which contains Center\_ID and payment details, prepare a Monthly Ledger Sheet. This ledger should reconcile all payments in the Collections\_Raw tab with their respective centers for each month. Using the functions written in task 1, export the resulting df in the same google sheet **Data\_Designs\_Technical\_Assessment\_2024** on a new tab calledStep\_3 **(30 points)**

To create the ledger sheet, consider the following guide. i.e add the following columns to your dataset:

* **SumPaid :** Equals to the total sum of the column paid\_in. Ensure you aggregate by month so that you cater for installment payments.
* **StartBalance :** This is the starting balance for a month. It's calculated by adding the previous month's ending balance (initially zero for the first month) to the current month's payment (SumPaid).
* **AllotmentRate:** This is a ratio calculated by dividing the starting balance for a month by the ExpectedMonthlyFee. It shows how close the franchisee is to fulfilling their payment obligation.
* **AllotedAmount :** This depends on the AllotmentRate. We can call it Sum Consumed for that particular period.Here's how it's determined:
  + If the Rate is 100% or above (almost or fully meeting the expected amount), the AllotedAmount is the full ExpectedMonthlyFee. Carry over excess amount (PrePayment) to the following month.
  + Anything below 100% is not consumed thus carry over.
* **Benefit\_Allotment :** This translates the AllotmentRate into a benefit category:
  + **"Full":** Rate is 100% o (fully meeting expectation). Eligible for FULL Benefit for that given Month
  + **"None":** Rate is less than 100% (paying less than the expected amount). Not Eligible for Any Benefit for that given Month.
* **EndBalance** : This is the remaining balance after subtracting the AllotedAmount from the StartBalance for a month. It represents how much the franchisee still has in surplus (Balance Carried Forward)
* **ActualRate :** This is a ratio calculated by dividing the AllotedAmount for a month by the expected amount. It represent the actual to meeting the expected collection fee for that particular month

***Hint:*** *Use a single Python function with nested loops. You can structure the output as either one row per month for each center or one row per center with monthly data as separate columns—choose the approach that best works for you.*

**Step 4:** Data Analysis, Visualization and Reporting **(30 points)**

Using the resulting DataFrame from Step 3.3, create a report in the form of an interactive dashboard (using Power BI) to address the following business questions:

1. What is the average collection rate (ActualRate) for each month?
2. Which county performed the best in terms of average collection rate (ActualRate) each month?
3. Which pilot program achieves the highest Actual collection rate? Does this performance vary by PilotPackage or CenterType?

**C: Submission**

Prepare a single .zip file containing the following:

1. An Excel file with all the expected outputs, organized into different tabs:
   * + **Step\_1 & Step\_2 Output**
     + **Wrong Account Payments**
     + **Step\_3 Output**
2. A Jupyter Notebook with all your Python scripts and code that documents the data cleaning process. Ensure your code includes clear comments for reproducibility during the assessment scoring.
3. A Power BI file with your report/dashboard. Would be checking on your DAX too.

* Name your submission .zip file as: **"SirName\_FirstName\_Data\_Designs\_Technical\_Assessment\_2024"**
* You have **48 hours** from receiving this test to complete the assignment and submit the zip file to **---------**. You may submit it as soon as you're finished. Be sure to comment about the time it took to complete each section.

**Enjoy the task !!!🥳🥳**