Internship Technical Assessment

Data Analyst

TASKS: Data Extraction, Cleaning, EDA, Visualization, Pipelining, Documentation and Reporting

TIMELINE: 26/11/24 – 08/12/24

OVERVIEW:	**Business Need:** Smallholder Commercial Farmers are in need of live data updates on market prices on their commodities, to inform them on seasonal/monthly market price performance and ideal markets to sell to/in. **Data Analyst's Objective:** Analyse Market Price Data over the last 3 years (1st October 2021 – 30th September 2024), present these findings, insights and recommendations, along with a live weekly update on market prices, to adequately respond to the business case above.
DATA AND SCOPE:	**Data Source:** You will obtain data from the **KAMIS** website. **Geographical Scope:** Nairobi, Nyandarua, Nakuru, Meru and Kirinyaga Counties. **Commodities Scope:** Dry Maize, White Irish Potatoes, Dry Onions, Watermelons, Ripe and Unripe (Cooking) Bananas, Oranges, Mangoes, Kales and Regular Spinach.
QUESTIONS:	1. What are the seasonal (quarterly) and monthly trends? 2. Which are the best and worst times of the year to sell these commodities and why? (Be commodity-specific) 3. Which are the top 3 priced markets per county? 4. What is today's wholesale and retail price in each county? (Live data pipeline, only for the dashboard).

DELIVERABLES:	1. A neat, coherent, well documented, titled and subtitled Jupyter notebook of your analysis exercise, with markdowns on findings, insights and recommendations for your audience (farmers) where applicable. (.ipynb) 2. A link to your PDF presentation **summary** of your analysis exercise: Introduction, Findings, Insights and Recommendations. 3. A link to your published dashboard of this analysis exercise, with a tile on live-updated price of the previous day. Only the price will be live, the rest is past, static data.
TOOLS:	**Data Analysis:** Pythonic notebooks, preferably Jupyter. **Presentation Report:** We highly recommend Canva, but you can use what you are most comfortable with to get the job done. **Dashboard:** Tableau or Power BI.
EVALUATION CRITERIA :	**1. Jupyter Notebook (50%)** Correctness and completeness of data cleaning. Quality and accuracy of the analysis. Logical/Coherent flow of analysis. Use of appropriate visualizations. Clarity of documentation (comments and markdown cells). **2. Summary Report (20%)** Clarity and organization of insights. Relevance and feasibility of recommendations. Visual appeal and professional formatting. **3. Dashboard (30%)** Usability and interactivity. Clarity in communicating key metrics and insights. Aesthetic design and adherence to best practices.

Data Scientist / Machine Learning Engineer

TASKS: Data Extraction, Cleaning, EDA, Pre-processing, Model Building, Model Evaluation and Iteration, Model Deployment

TIMELINE: 26/11/24 - 08/12/24

OVERVIEW:	**Business Need:** Smallholder Commercial Wheat Farmers are in need a diagnostic tool to detect, diagnose and treat wheat crop pests and diseases to prevent yield reduction. **Data Scientist/Machine Learning Engineer's Objective:** Develop an image classification model to diagnose wheat crop pests and diseases, that can be deployed and used by smallholder farmers on a web platform.
DATA AND SCOPE:	**Data Source:** You will obtain data from **Kaggle**. **Pest and Disease Classes:** Pests: Select at least 1 pest class Diseases: Select at least 3 disease classes **NB:** Do not forget to train and test with the 'Healthy' class as well.
DELIVERABLES:	**Jupyter Notebook:** • Evidence of exploratory data analysis (EDA) on the provided dataset. • Explanation of pre-processing steps. • Model Development • Model Evaluation **Deployed Model:** • Model Deployment Code/Files • A Streamlit application script for deploying the trained model. • Features to include: upload image button, prediction results and a display of confidence scores or probabilities for each class. Please note that we require a working link to the deployed model (along with all necessary documentation and README files).

EVALUATION CRITERIA :

1. Data Handling & Pre-processing (20%)

Effective exploratory data analysis (Dataset characteristics like class distributions, insights into data quality).

Data Pre-processing: Use of appropriate techniques like resizing, normalization, and augmentation and handling of class imbalances.

2. Model Selection & Development (50%)

Model Selection process.

Appropriate use of libraries and model frameworks.

Appropriate model analysis and evaluation metrics.

Visualization of model performance.

Hyperparameter tuning to optimize parameters.

3. Deployment & Usability (20%)

A working deployment that can take an input image and provide predictions.

User-friendly interface.

4. Documentation & Communication (10%)

Clear, concise, and well-documented workflow in the notebooks.

Logical flow from problem definition to results and conclusions.

README files: Comprehensive setup instructions for running the solution.