Project Title	**Extracting Location Information from Transaction Description Data.**
Client	Kazilytics
Project Lead	Dennis Chang'ach
Project Duration	40 Hours
Document Status	In Review

Project Description:

Kazilytics aims to leverage transaction data from mobile money platforms and bank statements to deliver advanced credit analytics to its clients. This initiative will facilitate informed lending decisions by extracting critical insights from financial transactions.

Project Objectives:

Extracting location information from transaction descriptions to enable the analysis of the movement of potential borrowers.

Deliverables

An algorithm that:

- Extracts location names from transaction descriptions.
- Adds the extracted location names as a new column in a table of transaction records.
- Integrates with a location database containing coordinate information to enable plotting borrower movements on a map.

Scope of Work:

1. **Discovery phase:** Understanding the data stack/infrastructure.
2. **Data Collection** : Compiling a comprehensive list of Kenyan locations, including cities, towns, and other relevant geographical entities.
3. **Developing Location Extraction Algorithm:**
    a. Data cleaning & preprocessing of transaction data from the database

    b. Creating location extraction algorithm.
    c. Algorithm validation and refinement.

4. **Deployment and Testing**: Deploying the algorithm to enable extraction of location information from current transactions data and uploaded statements.

Project Timeline(Workplan)

Day 1 & 2: Orientation and Analysis:

- Familiarize with the company's workflows, current data models, and infrastructure.
- Gain access to the necessary databases and tables, such as Mpesa transactions and bank transactions.
- Align on the project delivery approach, i.e creating new tables in the database to store the new features.

Day 3, 4 & 5: Development Phase

- Develop the location extraction algorithm, validate and optimize..

Day 6 & 7: Deployment and Testing

- Deploy the location extraction algorithm.
- Conduct tests on the front-end dashboards to ensure the integration and functionality of the new features.

Implementation Approach

Combining the strengths of NLP techniques with domain-specific knowledge of Kenyan geography to effectively extract and validate location information from statement descriptions.

Data Collection:

1. **Building a location database:**
    - **Comprehensive List of Locations:** Compiling a list of Kenyan locations, including cities, towns, and other relevant geographical entities. Sourcing this information from government databases, Wikipedia, or other geographical data providers.
    - **Coordinate Information**: Collecting coordinates (latitude and longitude) for each location to enable visualization on a map.

2. **Collecting statements:**
    - Data Pipeline: Developing a pipeline to retrieve bank and M-Pesa statements from the database. Ensuring the pipeline is secure and compliant with data privacy regulations.

Developing Location Extraction Algorithm:

1. **Data Cleaning and Preprocessing**:
    - **Text Cleaning**: Cleaning transaction descriptions by removing unnecessary characters (e.g., numbers, special characters), correcting common misspellings, and standardizing text formats.
2. **Implementing Named Entity Recognition (NER) Models:**
    - **Pre-trained NER Models:** Using pre-trained models available in NLP libraries like SpaCy and NLTK, which are designed to recognize location entities (e.g., "GPE" for geopolitical entities).
    - **Custom NER Model:** Depending on the performance of the pre-trained model, fine-tuning a custom NER model using a labeled dataset of Kenyan locations. This dataset should include city names, towns, neighborhoods, etc.
3. **Validation and Refinement:**
    - Validation against known addresses: Validating the extracted locations against a list of known addresses.
    - Manual Review: Manually reviewing a subset of the extracted locations to fine-tune the model and improve accuracy.
    - Fuzzy Matching: Implementing fuzzy matching techniques to handle misspellings and variations in location names.

Deploying the Location Extraction Algorithm

- **Deployment**: Depending on the data stack, the deployment approach will vary. The goal is to deploy a script/algorithm that:
    - Adds the extracted location names as a new column in a table of transaction records.
    - Integrates with a location database containing coordinate information to enable plotting borrower movements on a map.
- **Testing**
    - Uploading New Statements: Uploading new bank and M-Pesa statements to the system and verifying that the algorithm correctly extracts location information.
    - Front-End Dashboards: Conducting tests on the front-end dashboards to ensure the integration and functionality of the new features. Ensure that the extracted locations and plotted movements

are displayed correctly and provide meaningful insights into borrower behavior.

Pricing and Services Breakdown

We estimate approximately 40 hours of work for the pilot project, billed at a rate of $30 per hour. The total estimated cost for the pilot project is **$1,200**.

During this phase, we will conduct a thorough analysis to identify any additional infrastructure or compute costs required for deploying the project. This includes evaluating the existing infrastructure and determining if additional resources are needed to support the deployment and operation of the new analytics features.

Summary of Costs:

- **Development and Implementation**: 40 hours at $30/hour = $1,200
- **Potential Additional Costs:** Any identified infrastructure or compute costs will be evaluated and discussed separately based on the findings during the discovery phase.

Total Estimated Cost: $1,200 (excluding any additional infrastructure/compute costs)