Project Plan
Writing the project plan cannot be done without spending some time researching your problem domain. You need to familiarize yourself with the topic and research potential ways of solving the problem so you already have a preliminary idea of what you can do in practice. After that, you likely know enough to write this project plan.

Project name: Educational Disparities in Access to Essential Services: A Distance-Based Analysis (2010–2012)

Project members: Yagub Hajiyev, Elchin Huseynov

Problem statement:
Access to essential services such as healthcare facilities, educational institutions, and commercial centers is a critical determinant of a household's quality of life. However, this access may not be distributed equally across different educational levels of household heads. This project investigates whether households led by individuals with lower educational attainment face longer average distances to critical places compared to those led by more educated individuals. While the topic may not have direct commercial applicability, it offers valuable experience in applying business data analytics and machine learning techniques to real-world social datasets, emphasizing the intersection between socioeconomics and accessibility analysis.

Objectives:
- Primary Objective: To quantify and model the relationship between the educational level of the household head and the average distance to essential services.
- Secondary Objective: To predict average distances based on educational level and potentially other socio-demographic features if available.

Expected Outcome:
- Successfully identify whether significant disparities exist.
- Build a basic predictive model (e.g., regression or classification) to forecast the average distance based on education level and possibly other variables.
- Model evaluation target: achieving a reasonable predictive performance, such as an R^2 score above 0.70 for regression models, or equivalent metrics for classification tasks. However, due to the simplicity of the data, performance benchmarks will be flexible.

Data:
Yes, the dataset is available. Public statistical databases (user-provided).
https://avaandmed.eesti.ee/datasets/leibkondade-keskmine-kaugus-olulistest-punktidest-leibkonnapea-haridustaseme-jargi-(20102012)

Methodology: The project will primarily use Supervised Machine Learning techniques.
Proposed Models and Methods:
- Regression Models:

    - Linear Regression: To predict average distances based on the educational level numerically.
    - Random Forest Regression: To capture non-linear relationships between education level and distance to services.
- Classification Models (if distance categories are created, e.g., "Short", "Medium", "Long" distances):
    - Logistic Regression: For binary or multiclass classification tasks.
    - Decision Trees / Random Forest Classifier: To model categorical outputs based on education levels.

Additional Techniques:

- Exploratory Data Analysis (EDA) and feature encoding will precede modeling.
- Cross-validation will be used to ensure the robustness of model performance.

Evaluation:

Machine learning outcomes will be evaluated using appropriate metrics based on the model type:

- For Regression Tasks:
    - R^2 Score (Coefficient of Determination): Target $R^2 > 0.70$ if achievable.
    - Root Mean Squared Error (RMSE): To measure average prediction error in distance units.
- For Classification Tasks (if applicable):
    - Accuracy Score: Preferably aiming above 80%.
    - Precision, Recall, and F1-Score: Especially if the data is imbalanced across classes.

Evaluation will be done through train/test splits and cross-validation to avoid overfitting and ensure generalizability.

Expected challenges:

Several challenges are anticipated:

- Limited Feature Set:
  The dataset may primarily consist of education level and distances without much auxiliary data (like income level, region, urban/rural distinction), which could limit model complexity and predictive power.
    - Mitigation: Focus on robust feature engineering and, if possible, integrate external socioeconomic data.
- Data Imbalance:
  Certain education levels may have very few entries compared to others.
    - Mitigation: Apply techniques like oversampling, undersampling, or stratified sampling.
- Non-linear Relationships:
  The relationship between education and distance might not be linear.

    - Mitigation: Use non-linear models (e.g., Random Forests) and apply transformations to the features if necessary.
- Overfitting Risk:
  Particularly if the dataset is small relative to model complexity.
    - Mitigation: Use cross-validation, regularization (e.g., Ridge/Lasso Regression), and keep the models interpretable.

Resources and tools:

Programming Language:

- Python 3.x

Libraries:

- Pandas: Data manipulation and preprocessing.
- NumPy: Numerical operations.
- Matplotlib / Seaborn: Data visualization.
- Scikit-learn: Machine learning models and evaluation metrics.
- Statsmodels: For statistical analysis if necessary.

Environment:

- Work will be done on a local machine (laptop/desktop).
- No additional cloud computing resources are currently required.
- If necessary later (e.g., for larger model training or GPU use), services like Google Colab or AWS Free Tier may be considered.

Questions for further guidance:

1. Feature Expansion:
   Would it be acceptable to integrate additional publicly available socioeconomic data (e.g., urbanization level, income brackets) into the dataset to enrich the feature set, or should the project strictly remain limited to the original data?
2. Model Scope:
   Considering the dataset's simplicity, would it be appropriate to attempt both regression and classification approaches to demonstrate multiple machine learning techniques, or should the project focus on refining a single modeling direction?
3. Evaluation Flexibility:
   If the dataset’s structure leads to naturally low variance in target values (e.g., distances clustered around similar values), would success criteria (e.g., R^2 threshold) be adjusted accordingly to reflect the data's real-world characteristics?