NYCEDC Multimodal Analysis Project - Project Plan

1. Project Overview

1.1 Background

The NYCEDC Multimodal Analysis Project is a data-driven initiative aimed at analyzing and visualizing economic trends in life sciences, technology, and related industries. The project leverages structured and unstructured data sources to provide insights into research activity, federal funding allocation, patent trends, clinical trials, and startup investments.

1.2 Objective

The primary objective is to develop a fully automated multimodal data analysis system that enables NYCEDC to track, compare, and analyze industry trends over time. The final deliverable, AARDVARK, will be an interactive dashboard embedded in NYCEDC's internal website, providing policymakers with real-time data for strategic decision-making.

1.3 Success Criteria

- Functional AARDVARK Dashboard: A web-based, interactive visualization system for NYCEDC.
- Automated Data Pipeline: Fully automated scripts that refresh insights periodically.
- Comprehensive Data Integration: Merging multiple datasets into a unified analytical framework.
- Validated Scoring Model: A scoring mechanism that ranks cities and subsectors accurately.
- Stakeholder Satisfaction: Positive feedback from NYCEDC and industry stakeholders.

2. Scope of Work

2.1 Key Innovation Stages Covered

Stage	**What We're Tracking**	**Data Sources**
Early-Stage Innovation	Research & Publications	PubMed
Mid-Stage Innovation	Federal Funding & Patents	NIH RePORTER, USPTO
Late-Stage Innovation	Clinical Trials & Startups	ClinicalTrials.gov, PitchBook

2.2 Data Sources & Integration

- Research Trends: PubMed API for tracking publication impact.
- Federal Funding: NIH RePORTER for tracking grant allocations.
- Patent Activity: USPTO database for intellectual property trends.
- Clinical Trials: ClinicalTrials.gov for evaluating ongoing trials.
- Startup & VC Data: PitchBook for analyzing venture capital trends.

Additional Potential Data Sources:

- Data.gov
- USA.gov Data and Statistics
- Federal Reserve Data
- Bureau of Labor Statistics
- California Open Data Portal
- New York Open Data
- NOAA and NASA Open Data
- UCI Machine Learning Repository
- World Bank Open Datasets
- Columbia Library Datasets

3. Code Review & Findings

3.1 Key Observations

- Modular Structure: The codebase is well-structured, with separate scripts for each analytical component.
- Customizable Control File: The Control.xlsx file allows flexible updates to keyword filters and scoring parameters.
- Automated Data Processing Pipelines: Scripts fetch and process data from APIs (PubMed, NIH, USPTO, ClinicalTrials.gov, PitchBook).
- Scoring Model & Dashboard Integration: Outputs are designed to integrate into Looker Studio and Power BI.

3.2 Potential Issues & Areas for Improvement

- Data Processing Bottlenecks: Some scripts, particularly PubMed and NIH Funding analysis, may take longer than expected due to API rate limits and large dataset sizes.
- Keyword Classification Uncertainties: NIH funding and PitchBook startup classification might benefit from refined NLP models to improve accuracy.
- Manual Mapping in PitchBook Data: Some startup records could require manual sector tagging, which may need further automation.
- Score Merging & Normalization Considerations: Cross-dataset weighting might need review to ensure fair representation of each dataset.
- Visualization Tool Review: Explore open-source alternatives to Power BI/Tableau (see below).

4. Work Breakdown Structure (WBS)

Week	Task	Owner	Syllabus Mapping
1-2	Code Review & Dataset Exploration	Sahil & Zhiyi	Data Exploration & Initial EDA
3-4	Project Plan Formulation and Visualization Tool Proposals	Sahil & Zhiyi	Data Cleaning & Integration
5	Exploratory Data Analysis - Descriptive Analysis	TBD	Exploratory Data Analysis
6	Dollar-Per Insight Metric Computation		Advanced EDA & Derived Metrics
7	Draft Scoring Model Refinement		Model Development & Evaluation
8	Open-Source Dashboard Research & Prototype		Visualization Techniques & Dashboard Development
9	Final Dashboard & Report Preparation		Final Report & Presentation Drafting
10	NYCEDC Stakeholder Presentation		Capstone Final Presentation

5. Risk Assessment & Mitigation

Risk Factor	Potential Issue
Data Access Issues	API rate limits, restricted access (PitchBook)
Keyword Classification	Misclassification in NIH

Errors	funding and startup sectors
Large Data Processing Bottlenecks	NIH and USPTO datasets exceed 2GB
Timeline Constraints	Delays in data merging and testing

6. Dollar-Per Insight Analysis Focus

A key analytical goal is to compute 'dollar-per-insight' metrics to quantify economic impact per unit of investment. Specific metrics include:
• Number of patents per million dollars of NIH funding related to a given keyword.
• Number of startups per million dollars of NIH funding tied to research in specific domains.
This provides actionable economic efficiency measures for NYCEDC policymakers.

7. Conclusion

This project plan provides a structured roadmap for delivering an automated multimodal analysis system for NYCEDC. The primary focus is on automation, scalability, and real-time insights to support economic decision-making. By following this plan, we aim to enhance NYC’s ability to compete with other innovation hubs like Boston and the Bay Area while maximizing the impact of economic policies.

Next Steps:

1. Validate existing data and scoring models.
2. Finalize data integration and automation pipelines.
3. Develop interactive visualization dashboard.
4. Prepare for NYCEDC stakeholder presentation.

Prepared by Sahil Chavan