**Project Plan: Dayton, Ohio Demographic Database with Automated Dashboard**

**An Ouroboros Analytics Initiative**

**1. Project Initiation**

* **Define Objectives**: Clearly outline the goals of your demographic database project.
* **Stakeholder Identification**: Identify who will benefit from this project (e.g., potential employers, community planners).

**2. Storage Setup**

* **Google Cloud Storage**
  + **Create Storage Buckets**: Set up buckets for raw and processed data.
  + **Configuration**: Ensure proper configuration and access permissions.

**3. Data Gathering (Ongoing)**

* **Web Scraping**
  + **Tools**: Python (BeautifulSoup, Scrapy), Selenium for dynamic content.
  + **Data Sources**: Websites with relevant demographic information (local government sites, community portals).
  + **Output**: Raw data files (CSV, JSON).
* **APIs**
  + **Tools**: Requests library in Python.
  + **Data Sources**: U.S. Census Bureau API, other demographic data APIs.
  + **Output**: Raw data files (CSV, JSON).
* **Public Datasets**
  + **Sources**: Kaggle, government databases, academic studies.
  + **Tools**: Direct download or API.
  + **Output**: Raw data files (CSV, JSON).

**4. Data Cleaning and Preprocessing**

* **Handling Missing Values**
  + **Tools**: Pandas, NumPy.
  + **Techniques**: Imputation (mean, median, mode), deletion.
* **Removing Duplicates**
  + **Tools**: Pandas.
  + **Techniques**: Identify and drop duplicates.
* **Data Transformation**
  + **Tools**: Pandas, NumPy.
  + **Techniques**: Format conversion, handling categorical variables, normalization.

**5. Data Analysis**

* **Descriptive Analysis**
  + **Tools**: Pandas, Matplotlib, Seaborn.
  + **Output**: Summary statistics, basic visualizations.
* **Exploratory Data Analysis (EDA)**
  + **Tools**: Pandas, Matplotlib, Seaborn.
  + **Output**: Detailed visualizations (histograms, scatter plots, box plots).
* **Advanced Analysis**
  + **Tools**: Scikit-learn for machine learning, Statsmodels for statistical analysis.
  + **Output**: Regression models, clustering results, time series analysis.

**6. ETL Processes**

* **Extract**
  + **Tools**: Google Cloud Storage, Cloud Functions.
  + **Techniques**: Automated data extraction scripts.
* **Transform**
  + **Tools**: Cloud Dataflow.
  + **Techniques**: Batch and stream data processing.
* **Load**
  + **Tools**: BigQuery.
  + **Techniques**: Load cleaned and transformed data into BigQuery tables.

**7. Automation**

* **Scheduling and Automation**
  + **Tools**: Apache Airflow (Cloud Composer), Cloud Functions.
  + **Techniques**: Schedule and automate ETL processes, monitor job execution.

**8. Dashboard Creation and Updates**

* **Dashboard Development**
  + **Tools**: Looker, Google Data Studio.
  + **Techniques**: Design interactive dashboards to visualize demographic data.
* **Automated Updates**
  + **Tools**: Looker API, Google Data Studio integration.
  + **Techniques**: Automate dashboard updates based on data changes in BigQuery.

**9. Documentation and Presentation**

* **Documentation**
  + \*\*Create detailed documentation of your data sources, methods, and results.
  + **Tools**: Jupyter Notebooks, Google Docs.
* **Presentation**
  + \*\*Create visual presentations of your findings and analysis.
  + **Tools**: Looker, Google Data Studio, Jupyter Notebooks.

**12-Week Detailed Timeline**

**Week 1-2: Project Initiation & Storage Setup**

* **Define Objectives**: 2 hours - Outline project goals and scope.
* **Stakeholder Identification**: 1 hour - Identify potential beneficiaries.
* **Google Cloud Storage Setup**: 15 hours - Set up storage buckets, configure settings, ensure proper access.

**Week 3-4: Data Gathering & Initial Cleaning**

* **Web Scraping Setup**: 7 hours - Research data sources, set up BeautifulSoup/Scrapy.
* **API Integration**: 7 hours - Set up connections to U.S. Census Bureau and other APIs.
* **Public Datasets**: 3 hours - Locate and download relevant datasets.
* **Initial Data Cleaning**: 3 hours - Handle missing values and remove duplicates.

**Week 5-6: Data Preprocessing**

* **Data Transformation**: 10 hours - Format conversion, handle categorical variables.
* **Normalization**: 5 hours - Normalize numerical values.
* **Documentation**: 3 hours - Document data preprocessing steps.
* **Review and Adjust**: 2 hours - Review preprocessing, make necessary adjustments.

**Week 7-8: Data Analysis**

* **Descriptive Analysis**: 4 hours - Calculate summary statistics.
* **Visualizations**: 6 hours - Create visualizations using Matplotlib/Seaborn.
* **Exploratory Data Analysis (EDA)**: 6 hours - Identify patterns and relationships.
* **Advanced Analysis Setup**: 4 hours - Set up regression models, clustering.

**Week 9-10: ETL Processes**

* **Extract**: 6 hours - Automate data extraction scripts, set up Google Cloud Storage.
* **Transform**: 6 hours - Use Cloud Dataflow for processing data.
* **Load**: 6 hours - Load data into BigQuery.
* **Automation Setup**: 2 hours - Set up initial automation with Cloud Functions.

**Week 11-12: Dashboard Creation, Updates, and Final Review**

* **Dashboard Development**: 10 hours - Design and build dashboards using Looker or Google Data Studio.
* **Integration with BigQuery**: 5 hours - Connect dashboards to BigQuery.
* **Automate Updates**: 5 hours - Set up automated updates using Looker API or Google Data Studio integration.
* **Documentation**: 10 hours - Compile detailed project documentation.
* **Presentation**: 6 hours - Finalize visualizations, create presentation slides.
* **Review & Polish**: 4 hours - Review entire project, make final adjustments.

**Weekly Breakdown**

| **Week** | **Activity** | **Hours** |
| --- | --- | --- |
| 1-2 | Project Initiation & Storage Setup | 20 |
| 3-4 | Data Gathering & Initial Cleaning | 20 |
| 5-6 | Data Preprocessing | 20 |
| 7-8 | Data Analysis | 20 |
| 9-10 | ETL Processes | 20 |
| 11-12 | Dashboard Creation, Updates, and Final Review | 20 |