**Data Science and Machine Learning Projects**

Projects are listed into three levels. We need to make all possible projects to GUI using Django.

**Beginner Projects**

1. **Weather Data Analysis**

**Objective:** Analyze historical weather data to identify patterns, trends, and possibly predict future weather conditions.

**Data Sources:** NOAA (National Oceanic and Atmospheric Administration), Weather.com API, Kaggle datasets.

**Key Concepts:** Data cleaning, time series analysis, data visualization.

**Tools:** Python (Pandas, Matplotlib/Seaborn, SciPy for statistical tests).

**Potential Analyses:** Average temperature changes over years, precipitation trends, seasonality in weather, extreme weather event frequency.  
  
**Estimated Time:** 2 - 3 weeks.

1. **Movie Ratings and Reviews Analysis**

**Objective:** Determine factors that contribute to a movie's success or failure using ratings and reviews.

**Data Sources:** IMDb, Rotten Tomatoes API, Kaggle datasets.

**Key Concepts:** Sentiment analysis, data visualization, correlation analysis.

**Tools:** Python (NLTK or spaCy for NLP, Pandas for data manipulation, Matplotlib/Seaborn for visualization).

**Potential Analyses:** Sentiment score of reviews, correlation between budget and success, impact of release timing on ratings.

**Estimated Time:** 3 - 4 weeks.

1. **E-commerce Product Reviews Analysis**

**Objective:** Analyze customer reviews to identify product strengths, weaknesses, and predict customer satisfaction.

**Data Sources:** Amazon Product Review Dataset, company-specific datasets.

**Key Concepts:** Natural language processing (NLP), sentiment analysis, topic modeling.

**Tools:** Python (NLTK, spaCy for NLP; Pandas for data manipulation; Matplotlib/Seaborn for visualization).

**Potential Analyses:** Most common complaints/praises, sentiment trends over time, impact of reviews on sales.

**Estimated Time:** 3 - 4 weeks.

**Intermediate Projects**

1. **Loan Default Prediction**

**Objective:** Predict the likelihood of loan default based on borrower characteristics.

**Data Sources:** Public financial datasets, Kaggle competitions.

**Key Concepts:** Classification models, feature engineering, model evaluation metrics (AUC-ROC, precision, recall).

**Tools:** Python (Scikit-learn for machine learning models, Pandas for data manipulation).

**Potential Models:** Logistic regression, random forest, gradient boosting.  
  
**Estimated Time:** 4-6 weeks.

1. **Healthcare Data Analysis for Disease Prediction**

**Objective:** Use patient data to predict the likelihood of diseases such as diabetes or heart disease.

**Data Sources**: UCI Machine Learning Repository, Kaggle datasets.

**Key Concepts:** Classification algorithms, data preprocessing, dealing with imbalanced data.

**Tools:** Python (Scikit-learn, TensorFlow or PyTorch for more complex models, Pandas for data manipulation).

**Potential Models:** Decision trees, SVM, neural networks.

**Estimated Time:** 5-7 weeks.

1. **Image Classification for Retail Products**

**Objective:** Classify retail products from images for inventory management.

**Data Sources:** Retail dataset images, Google Images (with appropriate permissions).

**Key Concepts:** Computer vision, convolutional neural networks (CNN), transfer learning.

**Tools:** Python (TensorFlow/Keras, PyTorch for deep learning; OpenCV for image preprocessing).

**Potential Models:** Pre-trained models (VGG16, ResNet) fine-tuned on your dataset.  
  
**Estimated Time:** 6 -8 weeks.

1. **Predicting Flight Delays**

**Objective:** Use historical flight data to predict delays based on factors like weather, airport traffic, and time of day.

**Data Sources:** Bureau of Transportation Statistics, OpenFlights API, weather APIs.

**Key Concepts:** Regression models, time series analysis, feature engineering.

**Tools:** Python (Scikit-learn for regression models, Pandas for data manipulation, Matplotlib/Seaborn for visualization).

**Potential Models:** Linear regression, random forest regressor, LSTM for time series prediction.  
  
**Estimated Time:** 5 - 6 weeks.

1. **Dynamic Pricing Models for E-Commerce**

**Objective:** Develop models to adjust product prices in real-time based on demand, availability, and customer behavior.

**Data Sources:** Company's sales data, competitor pricing (if available).

**Key Concepts:** Price elasticity modeling, demand forecasting, machine learning.

**Tools:** Python (Scikit-learn for predictive modeling, Pandas for data manipulation).

**Potential Models:** Regression models, time series forecasting models (ARIMA, Prophet).  
  
**Estimated Time:** 8 -10 weeks.

**Advanced Projects**

1. **Natural Language Generation for News Articles**

**Objective:** Develop a model that generates coherent and contextually relevant news articles or content based on input data or headlines.

**Data Sources:** News article datasets, web scraping from news websites.

**Key Concepts:** Natural language processing (NLP), natural language generation (NLG), deep learning.

**Tools:** Python (TensorFlow with the Transformer model, GPT-2/GPT-3 models if accessible, NLTK/spaCy for text preprocessing).

**Approach:** Fine-tune

**Estimated Time:** 10 -14 weeks.