Applied Data Science Lab

Issued by WorldQuant University

To earn this badge, I have completed eight end-to-end, applied data science projects. In each project, accessed data from files, SQL and NoSQL databases, and APIs. I have demonstrated my ability to explore and clean data and create functions and ETL pipelines to prepare training sets. I have built machine learning models for supervised and unsupervised learning tasks and have created visualizations to explain data characteristics and model predictions for non-technical audiences.

Skills

- **API Design**
- **Data Science**
- **Data Visualization**
- **Machine Learning**
- **MongoDB**
- **Python (Programming Language)**
- **SQL**
- **Statistics**

Earning Criteria

- I completed eight projects. Each project consists of four self-paced lessons, followed by an assignment that is programmatically graded. For each assessment, students must score 90% or better.
- 1. HOUSING IN MEXICO: I used a dataset of 21,000 properties to determine if real estate prices are influenced more by property size or location. Imported and cleaned

data from a CSV file, built data visualizations, and examined the relationship between two variables using correlation.

- 2. APARTMENT SALES IN BUENOS AIRES: I built a linear regression model to predict apartment prices in Argentina. I created a data pipeline to impute missing values and encoded categorical features, and improved model performance by reducing overfitting.
- 3. AIR QUALITY IN NAIROBI: I built an ARMA time-series model to predict particulate matter levels in Kenya. I extracted data from a MongoDB database using Pymongo, and improved model performance through hyperparameter tuning.
- 4. EARTHQUAKE DAMAGE IN NEPAL: I built logistic regression and decision tree models to predict earthquake damage to buildings. Extracted data from an SQLite database, and revealed the biases in data that can lead to discrimination.
- 5. BANKRUPTCY IN POLAND: Built random forest and gradient boosting models to predict whether a company will go bankrupt. Navigated the Linux command line, addressed imbalanced data through resampling, and considered the impact of performance metrics precision and recall.
- 6. CUSTOMER SEGMENTATION IN THE US: Built a k-means model to cluster US consumers into groups. Used principal component analysis (PCA) for data visualization and created an interactive dashboard with Plotly Dash.
- 7. A/B TESTING AT WORLDQUANT UNIVERSITY: Conducted a chi-square test to determine if sending an email can increase program enrollment at WQU. Built custom Python classes to implement an ETL process and created an interactive data application following a three-tiered design pattern.
- 8. VOLATILITY FORECASTING IN INDIA: Created a GARCH time series model to predict asset volatility. Acquired stock data through an API, cleaned and stored it in an SQLite database, and built my own API to serve model predictions.