# Nobel Prize Winners - Analysis

E10

Matteus Kalda, Karl Gregor Urmet

1. Introduction

► The aim of this project is to explore patterns in Nobel Prize awards and relate them to broader country-level development indicators.

► **Business goals:**
Analyze distribution of laureates by time, category, gender, region and age.

► Model age at award based on category, decade, gender and region.

► Explore relationship between country HDI and Nobel output per capita (1990+)
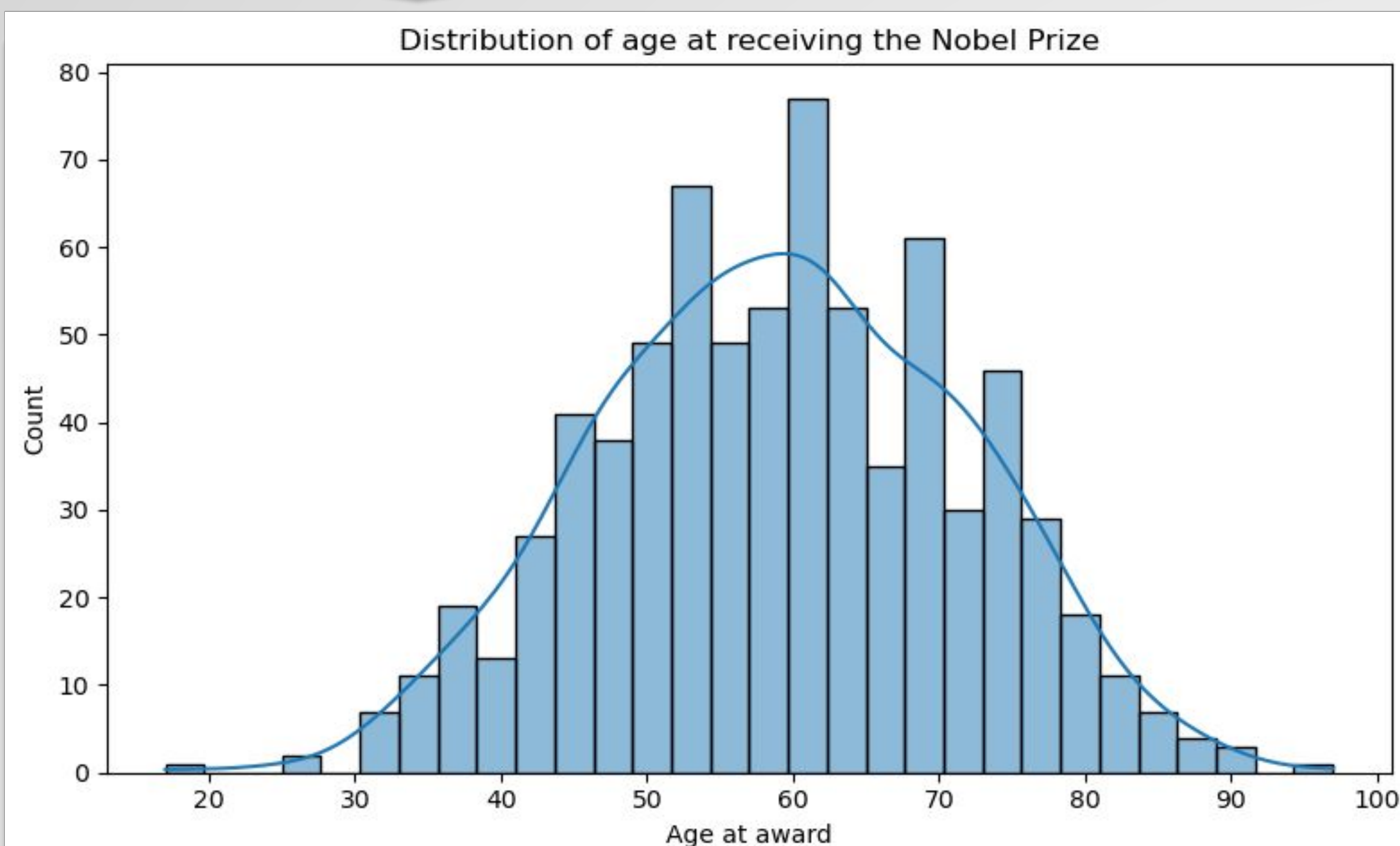
► Age at receiving the Nobel Prize, all laureates 1901−2023. The distribution is roughly bell-shaped with a peak around the late 50s.

2. Datasets

► We are using two datasets from Kaggle: Nobel Laureates [1901-2025] and Human Development Index Dataset [1990-2022].

► The first provides laureate level information. The second provides country level HDI and population measures.

Reference:
- https://www.kaggle.com/datasets/ahmeduzaki/nobel-prize-winners-dataset-1901-2025
- https://www.kaggle.com/datasets/lucasyukioimafuko/human-development-index-hdr-dataset-1990-2022

3. Used Methods

► Data cleaning and data-quality checks for the Nobel dataset.

► **Feature engineering** - birth_year, age_at_award, decade, region.

► **Exploratory data analysis** - age distribution and trends by decade comparisons by category, sex and region.

► **Statistical modelling** - OLS regression for age at award using category, sex, region and time.

► **Scikit-learn experiment** - regression models for age at award (LinearRegression, Ridge, Lasso) train/test split, MSE and R² comparison with the OLS model.
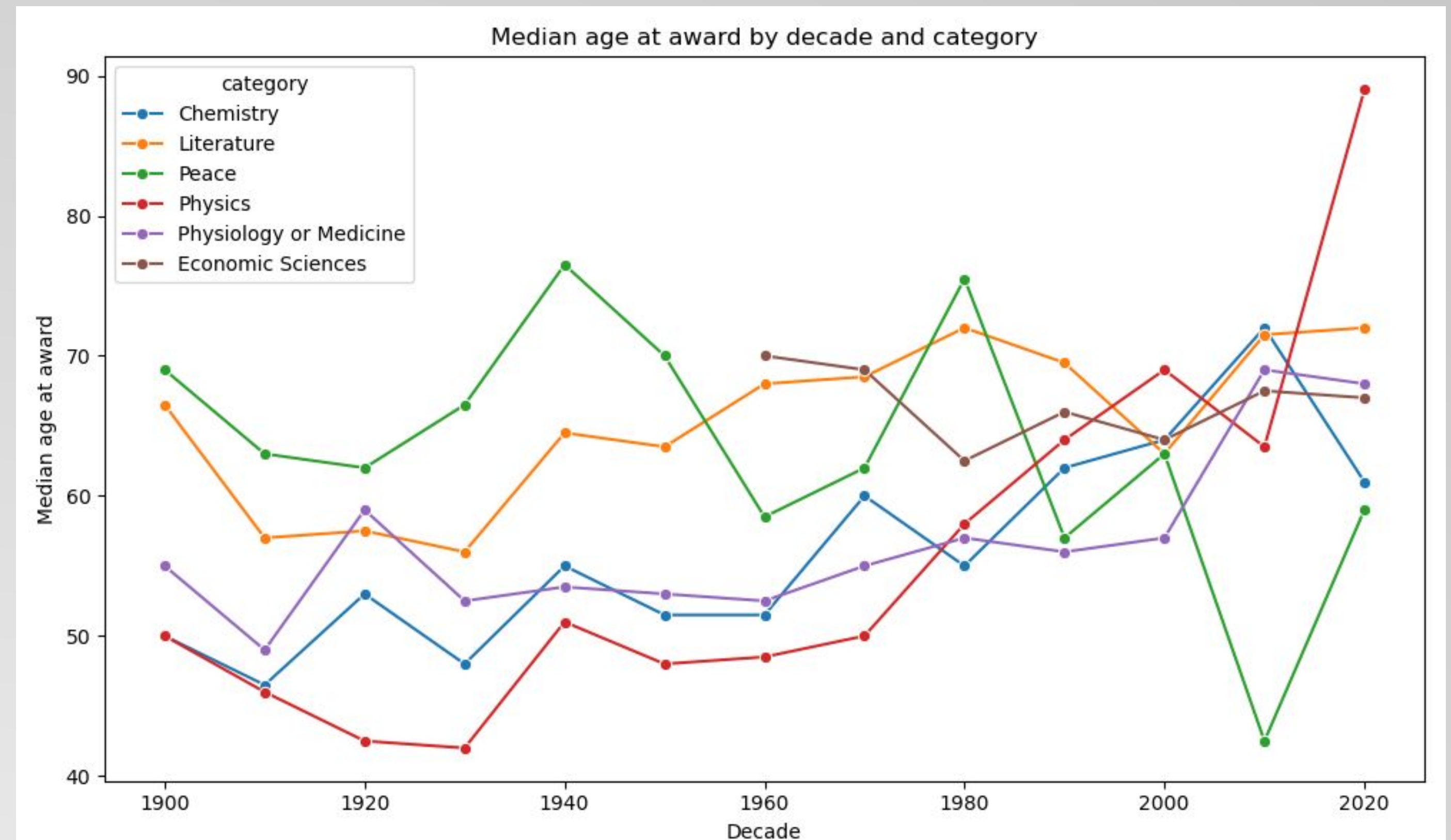
► Median age at award for Nobel Prize winners across six categories from 1900 to 2020, showing variation by field and decade.

4. Approach

► **Data cleaning** - removed obvious errors, dealt with missing birth dates and checked that HDI values stay within [0, 1].

► **Harmonising countries** - mapped birth and affiliation countries to a common list of country names and world regions, then joined them with the HDI dataset.

► **Nobel-level analysis** - studied the distribution of age at award, compared categories, sexes and regions, modelled age at award using a linear regression model.

► **Country-level analysis** - aggregated laureates by country and decade, computed Nobel laureates per million inhabitants, linked these outputs to HDI and population data to test the HDI−Nobel relationship.
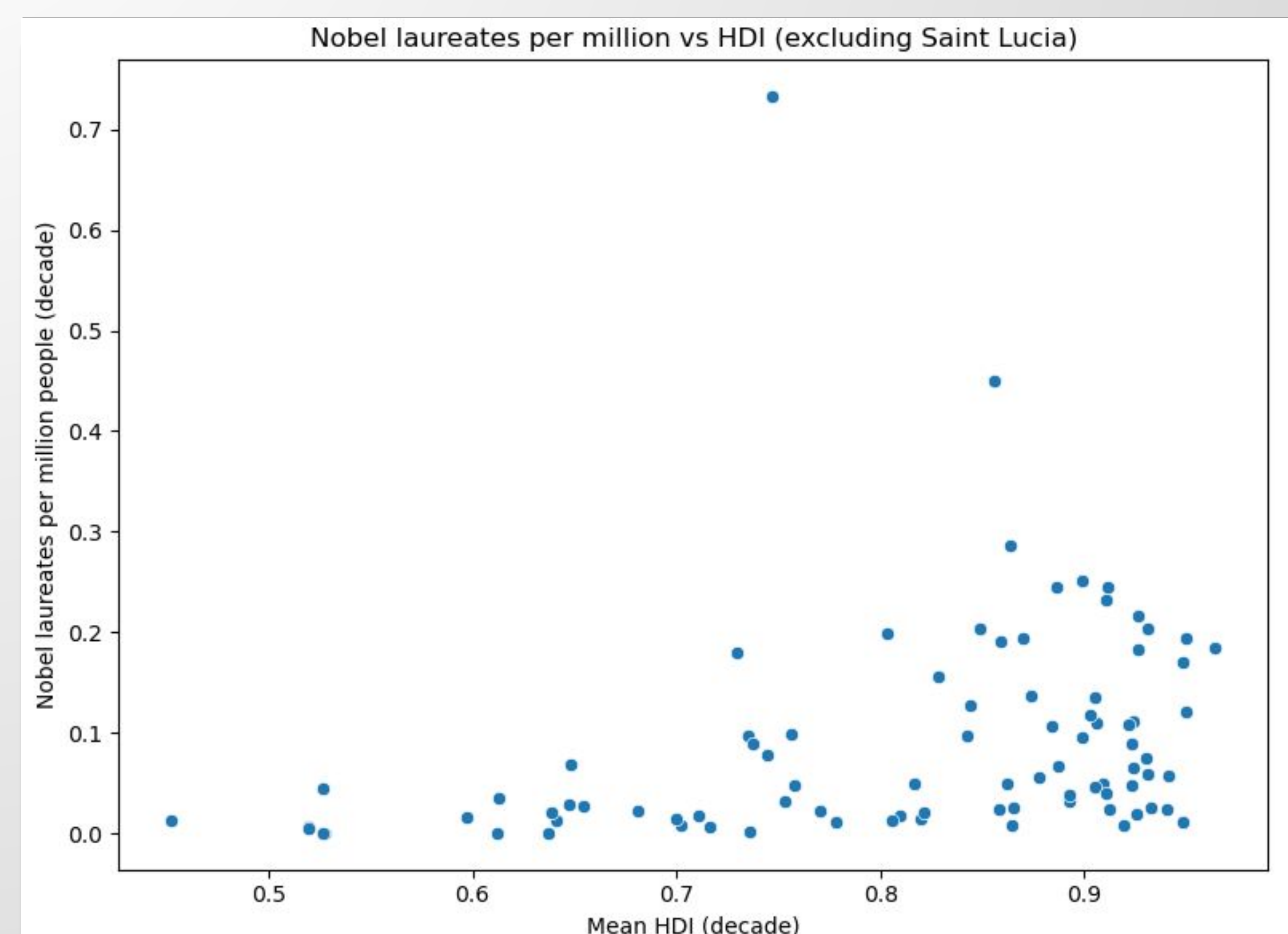
► Age at receiving the Nobel Prize, all laureates 1901−2023. The distribution is roughly bell-shaped with a peak around the late 50s.

5. Results and Conclusions

► Our main conclusion is that we can describe several clear patterns (e.g. older laureates over time, strong differences between fields), but we cannot build a practically useful model that would accurately predict a future laureate's age from these simple features alone.

GitHub: https://github.com/KarlGm4n/nobel-dataset