Nobel Prize Winners

Group: Group 5, E10
Team members: Matteus Kalda, Karl Gregor Urmet
Github repository: https://github.com/KarlGm4n/nobel-dataset
Kaggle datasets:
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

Business understanding

Identifying your business goals

Background

The Nobel Prize is a well known marker of scientific and cultural prestige. Since 1901 the prizes have been awarded in several categories to individuals and organisations. The Kaggle dataset "Nobel Laureates, 1901-Present" lists all laureates and includes year, category, prize, laureate_id, laureate_type, full_name, birth_date, birth_country, sex and organization_country.

The Human Development Index (HDI) summarises a country's development level using health, education and income. The Kaggle dataset "Human Development Index Dataset [1990-2022]" contains yearly HDI data in the file hdr_general.csv with columns such as iso3, country, year, hdi and pop_millions. In our project we mainly analyse the Nobel data and use HDI and population from pop_millions at the country level as context.

Business goals

1. Describe how Nobel laureates are distributed over time by category, gender, broad region and age at award.

2. Model the age at which laureates receive the Nobel Prize and see how it depends on category, award decade, gender and region.

3. Explore whether countries with higher HDI have more Nobel laureates per capita since around 1990.

Business success criteria

The project is successful if we

- create a small set of clear figures and tables that show key Nobel trends
- fit at least one interpretable regression model for age at award that reveals meaningful differences between groups
- obtain a sensible and honest result about the relationship between Nobel output per capita and HDI, even if this relationship is weak

---

Assessing your situation

Inventory of resources

We use two datasets: Nobel Laureates, 1901-Present and Human Development Index Dataset [1990-2022]. The first provides laureate level information. The second provides country level HDI and population measures through columns such as country, year, hdi and pop_millions. For analysis we use Python, pandas, numpy, matplotlib, seaborn, statsmodels or scikit-learn, Jupyter notebooks, laptops and GitHub. CRISP-DM materials and course notes are our main methodological resources.

Requirements, assumptions and constraints

We must deliver a reproducible analysis and written report by the deadline. A reader should be able to start from the raw CSV files in the repository and recreate our main results.

We assume the Nobel dataset correctly lists laureates and that the key variables are reliable. We assume the HDI dataset is based on official UNDP reports and that hdi and pop_millions are comparable across countries and years. Time is limited to roughly 30 hours per team member, so we focus on descriptive

analysis, one main age model and a simple country level comparison rather than complex modelling.

Risks and contingencies

Main risks are missing or inconsistent values for birth_date, sex or country in the Nobel data, country name mismatches between Nobel and HDI datasets, very low Nobel counts and noisy Nobel per capita values for many countries and no clear relationship between HDI and Nobel output per capita.

As contingencies we will restrict the age model to laureates with a valid birth year and report coverage, build a small country name mapping and group some countries into broad regions, aggregate Nobel counts by country and decade and normalise by decade level population using pop_millions and, if the HDI link is weak, present it briefly and emphasise the Nobel only analysis.

Terminology

Laureate - person or organisation that receives a Nobel Prize
Category - prize field such as Physics, Chemistry, Physiology or Medicine, Literature, Peace or Economics
Age at award - award year minus birth year
Nobel output per capita - country level Nobel count divided by population from pop_millions, for example per million people
HDI - Human Development Index, a value from 0 to 1 that combines health, education and income

Costs and benefits

We use open data and free tools, so there are no direct financial costs. The main cost is our time. The benefits are educational. We practise the CRISP-DM process, data cleaning, modelling and communication, and we produce a compact analysis of how Nobel Prize patterns relate to global human development.

Defining your data-mining goals

Data-mining goals

- Perform descriptive statistics and visualisations of Nobel laureates by year, decade, category, gender, region and age at award.

- Fit a regression model for age at award using category, decade, gender and region as predictors.

- Aggregate Nobel data by country and decade from 1990 onwards, compute decade level HDI and pop_millions averages, derive Nobel output per capita and explore the relationship between this output and HDI using correlations and a simple regression model.

Data-mining success criteria

Data mining is successful if

- exploratory analysis reveals clear patterns in the Nobel data

- the age at award model shows understandable differences between categories and across decades

- the country level analysis leads to an interpretable conclusion about how Nobel output per capita is related to HDI, or that it is only weakly related, with this uncertainty clearly explained

---

Data understanding

Gathering data

Outline data requirements

We work with two datasets.

From the Nobel dataset "Nobel Laureates, 1901-Present" we need for each laureate prize record

- year - award year
- category - prize category such as Physics or Literature
- prize, laureate_id and laureate_type
- full_name
- sex
- birth_date, birth_city and birth_country
- organization_name, organization_city and organization_country

These fields allow us to compute age at award, assign regions and study patterns by gender and geography.

From the HDI dataset "Human Development Index Dataset [1990-2022]" we need from hdr_general.csv

- iso3 - country code
- country - HDR country name
- year - year between 1990 and 2022
- hdi - Human Development Index value
- pop_millions - total population in millions

These fields allow us to compute decade level averages of hdi and population and later Nobel output per capita.

Verify data availability

The Nobel dataset contains about 1000 rows with year, category, prize, laureate_id, laureate_type, full_name, birth_date, birth_country, sex, organization_name, organization_country and death_date, so all key variables we need are available.

The HDI dataset provides a panel of country year observations in hdr_general.csv with iso3, country, year, hdi and pop_millions. The period 1990-2022 overlaps with the later part of the Nobel dataset. Both files are in CSV format and can be loaded into pandas.

Define selection criteria

For the Nobel dataset we plan to

- use all records from 1901 onwards for descriptive statistics and visualisations
- limit the age model to rows with a non missing birth year
- when linking to HDI, focus on laureates from 1990 onwards

We need a rule for assigning a country to each laureate. By default we will use birth_country. If birth_country is missing but organization_country is present, we may use organization_country instead and document this choice.

For the HDI dataset we plan to

- keep rows with year from 1990 up to the most recent year with good coverage
- drop rows where country, hdi or pop_millions is missing
- exclude entries that are not individual countries
- compute decade level averages of hdi and pop_millions for each country

---

Describing data

The Nobel dataset is a table where each row is a laureate prize record. The main numeric variable is year. We will derive age_at_award by subtracting the birth year from the award year. The main categorical variables are category, sex, birth_country and organization_country. The dataset has fewer than 1000 rows, so it is small enough for quick exploration but covers more than a century of awards.

The HDI dataset is a country year panel with iso3, country, year, hdi and pop_millions. From this dataset we will build a smaller table with one row per country and decade that contains average hdi and average pop_millions for that decade.

Exploring data

For the Nobel data we plan to explore

- counts of laureate records by decade and category
- the share of female laureates by decade and category
- distributions of age_at_award overall and by category and decade
- counts of laureates by broad region, where we map birth_country to regions such as Europe, North America, Asia and others

For the HDI data we plan to

- inspect the distribution of hdi values and how many countries fall into low, medium, high and very high human development groups
- look at how hdi changes over time for selected countries

- examine the range of pop_millions and how it changes over time
- check how many countries appear in both datasets after harmonising country names

---

Verifying data quality

For the Nobel dataset we will

- confirm that year values fall in a reasonable range and that category contains only official Nobel categories
- parse birth_date where possible and count missing or invalid dates
- compute age_at_award and flag clearly unrealistic ages such as below 18 or above 100
- look for obvious duplicate rows using laureate_id, year and category

If many birth years are missing we will report the share of missing ages and restrict the age model to the reliable subset.

For the HDI dataset we will

- check that hdi values fall between 0 and 1
- check that pop_millions is non negative and has plausible ranges
- look for duplicate iso3 and year pairs
- examine missing hdi or pop_millions values and decide whether we need to drop some countries or limit the time period
- harmonise a few problematic country names so that they match the Nobel country names as closely as possible

After these checks we expect to have two cleaned tables. One is a laureate level Nobel table with age_at_award. The other is a country decade table with Nobel counts, hdi, pop_millions and Nobel output per capita, which we will use in the modelling and visual analysis.

Planning your project

Planned tasks and hours

1. Data acquisition and cleaning

- Download Nobel and HDI datasets, load them into pandas, clean date formats and country names, check missing values and save processed CSV files in the repository

- Matteus 6 hours, Karl 6 hours

2. Nobel exploratory analysis and visualisation

- Produce descriptive statistics and plots of Nobel laureates by decade, category, gender, region and age_at_award

- Matteus 7 hours, Karl 7 hours

3. Age at award modelling

- Compute age_at_award, create features such as category, decade, gender and region and fit a simple regression model, then interpret the results

- Matteus 6 hours, Karl 6 hours

4. Country level Nobel and HDI analysis

- Aggregate Nobel data by country and decade from 1990 onwards, compute decade level averages of hdi and pop_millions, derive Nobel output per capita, merge the tables and run basic correlation and

regression analysis with a few plots

- Matteus 6 hours, Karl 6 hours

5. Final report and presentation

- Write and edit the report, choose final figures and tables, clean notebooks, make slides and create the poster

- Matteus 5 hours, Karl 5 hours

Each team member is planned to contribute 30 hours, so the total planned effort is 60 hours.