DSA210 – Introduction to Data Science: Final Report

**Project Title:** Does Living in a City Make You Taller: Analyzing the Effect of Urbanization on Human Height

**Student Name:** İrem Çakmakcı

**Date:** 28/05/2025

**GitHub Repository:** <https://github.com/4meri/DSA210-Project>

1. **Motivation**

In this term project, I aimed to investigate how urbanization affected average human height by analyzing data of possible parameters such as income levels, daily supply of calories per person, healthcare access, child mortality rates etc. Motivation behind this study was to associate my collective interest in how urbanization has influenced quality of life through several aspects to a health-related parameter, average height, as an indicator of physical development.

1. **Data Source**
   1. **Data Collection**

All my datasets are publicly available, collected from Our World in Data (n.d.) which is a research and database platform on global development trends.

* 1. **Data Preparation**

Some years had missing height data, which were left unfilled, as individuals born after this year had not reached adult height (age 18), and interpolation or imputation would not have been meaningful. I also removed redundant columns and standardized year formatting. No extreme outliers were detected as I checked for unreasonable values using summary statistics and visualizations. I normalized variables for an honest comparison, if necessary.

1. **Data Analysis**
   1. **Exploratory Data Analysis (EDA)**

I used initial visualizations such as histograms, box plots and scatter plots to catch any initial patterns. A general trend in how urbanization parameters affect average height in a positive way was observed as an initial insight.

* 1. **Hypothesis Testing**

*Null Hypothesis:* Urbanization has no significant relationship with average human height.

*Alternative Hypothesis:* Urbanization tends to correlate positively with average human height.

I chose significance level (alpha) as 0.05 in my project. Various tests are used for hypothesis testing, for instance chi-square test is used to investigate daily calorie intake vs average height or analysis of GDP per capita vs average height is done using Pearson correlation coefficient.

* 1. **Machine Learning**

I applied linear regression for interpretability and random forest regression for nonlinearity. Each model was trained both for female and male height. I calculated R^2 and RMSE for both models which respectively measure how well the models explain the variance in height and how accurate our predictions are.

1. **Findings**

The results of the tests that I applied to some of the most important parameters:

1. GDP per capita: Its p-value is 0.0020 which is less than the significance level 0.05 and we reject the null hypothesis. Therefore, also with the correlation coefficient 0.49, we can conclude that there is a significant positive relationship, thus wealthier societies tend to be taller.
2. NOx emissions: P-value of this parameter is 0.0005, being less than our significance level 0.05, it further contributes to rejection of null hypothesis. We can observe a significant positive relationship, with correlation coefficient 0.54, which leads us to the fact that more emissions (indicating industrial activity and urbanization) correlate with taller height.
3. Daily calorie intake: Indicating nutrition, p-value of this parameter is 0.0132, which is again less than our significance level 0.05. Despite its relatively moderate significant correlation (coefficient of 0.40), we still conclude that it has a positive relationship.
4. Employment in agriculture: With correlation coefficient -0.26, we can say that correlation is weak negative and not that significant. P-value is 0.127, more than significance level, we fail to reject null hypothesis.
5. Factors such as vaccine coverage etc. don’t have significant relationship with average height.

Interpreting and using the ML results, we can conclude that the model explains 98.8% of the variance in male height and 86.8% in female height. Predictions of linear regression are highly accurate for male height with a mean prediction error of approximately 2.7 cm and, though not as high as male height, accurate for female height with an average prediction error of about 3.4 cm. Whereas Random Forest is better for female height prediction since the model explains 94.7% of the variance in average male height and performs a bit more variation in prediction errors with RMSE 5.6 cm but the corresponding female equivalences are 92.8% and 2.5 cm that are better than linear regression performance. All in all, since R^2 > 0.85 for both models, they show very strong explanatory power. Also urbanization-related parameters play significant roles in predicting average human height.

Aside from the individual evaluations, by analyzing the below graphs (Figure 1.1 and 1.2) that are generated using pieces of my code, we can clearly see that daily calorie intake, nutrition, GDP per capita and NOx emissions are most important features among factors affecting male and female height.

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir. Figure 1.1.

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir. Figure 1.2.

1. **Limitations and Future Work**

Missing height values for certain years/periods may cause bias. Even though many factors related to urbanization like GDP per capita and daily calorie intake are analyzed in order to investigate whether there is a relationship between urbanization and average human height, some other factors that are not included such as genetic diversity and prenatal care could possibly impact height.

All these factors can be further investigated by dividing the sample into two: city districts vs countryside. Also, it can be analyzed in terms of specific policies that affect nutrition or pollution levels. Besides, using studies that track individuals from birth to adulthood could isolate urbanization’s influence over time which can reveal a more concentrated research.

1. **Use of LLM’s**

In ML phase, I asked ChatGPT which ML methods that I can use in my analysis.

* My prompt: “Which ML methods are suitable for my project?”
* Output: Linear Regression and Random Forest Regressor

I used ChatGPT to gather the code pieces that I prepared to implement the tests into a .ipynb file in a meaningful way.

* My prompt: “Can you compound these code pieces into a single .ipynb filein such a way that it is obvious which code piece implements which test?”
* Output: Statistical\_Analysis\_Report.ipynb

I also used Gemini to convert resources of my datasets to reference list format.

* My prompt: “How should I cite this dataset in my reference list in APA style?”
* Output: Reference list entries of my datasets

Also, I used ChatGPT to combine my datasets into a final .csv file.

* My prompt: “Can you merge the datasets that I provide you to a single final dataset?”
* Output: Fully\_Cleaned\_Dataset.csv

1. **Reference List**

*Average Height by Year of Birth*. Data adapted from Baten and Blum (2015). Our World in Data. Retrieved from <https://ourworldindata.org/grapher/average-height-by-year-of-birth>

Bolt, J., & Van Zanden, J. L. (2023). *Maddison Project Database 2023* and Maddison Database 2010. Data compiled from multiple sources by World Bank (2025). With major processing by Our World in Data. Retrieved from <https://ourworldindata.org/grapher/global-average-gdp-per-capita-over-the-long-run>

*Dietary Composition by Country*. Our World in Data. Retrieved from <https://ourworldindata.org/grapher/dietary-composition-by-country>

Food and Agriculture Organization of the United Nations (2024). Data compiled from multiple sources including Harris et al. (2015), Floud et al. (2011), Jonsson (1998), Grigg (1995), Fogel (2004), Food and Agriculture Organization of the United Nations (2000), Food and Agriculture Organization of the United Nations (1949), and USDA Economic Research Service (ERS) (2015). With major processing by Our World in Data. Retrieved from <https://www.google.com/search?q=https://ourworldindata.org/grapher/daily-per-capita-calorie-supply>

Gapminder & UN Inter-agency Group for Child Mortality Estimation. (2025). *Child Mortality Rate*. Processed by Our World in Data. Retrieved from https://ourworldindata.org/grapher/child-mortality-rate

*Global Vaccination Coverage*. Our World in Data. Retrieved from <https://ourworldindata.org/grapher/global-vaccination-coverage>

*Long-Run Air Pollution*. (2021). Our World in Data. Retrieved from <https://ourworldindata.org/grapher/long-run-air-pollution>

Our World in Data based on International Labor Organization (via the World Bank) and historical sources – processed by Our World in Data. “Number of people employed in agriculture” [dataset]. Our World in Data based on International Labor Organization (via the World Bank) and historical sources [original data].

UNICEF, World Health Organization, & World Bank. (2024). *Share of Children Who Are Stunted*. Processed by Our World in Data. Retrieved from https://www.google.com/search?q=https://ourworldindata.org/grapher/share-of-children-who-are-stunted

United States Department for Agriculture (USDA) Economic Research Service – processed by Our World in Data. “Agricultural output” [dataset]. United States Department for Agriculture (USDA) Economic Research Service [original data]

WHO/UNICEF Joint Monitoring Programme for Water Supply, Sanitation and Hygiene (JMP). (2024). *Share of the Population Using Improved Sanitation Facilities*. Processed by Our World in Data. Retrieved from https://www.google.com/search?q=https://ourworldindata.org/grapher/share-of-population-using-improved-sanitation-facilities

World Health Organization. (2025). *Current Health Expenditure per Capita*. Processed by Our World in Data. Retrieved from <https://www.google.com/search?q=https://ourworldindata.org/grapher/current-health-expenditure-per-capita>

1. **Appendix**

Datasets without identifiable authors (average height by year of birth, dietary composition by country, global vaccination coverage and long-run air pollution) were sourced from <https://ourworldindata.org/>. All datasets including these, for which original authorship could not be specified, are cited in the Reference List section.