DATA SET DESCRIPTION

Francesco Biondi

1 Myopia Study

The *Myopia Dataset*[1] includes a set of variables collected from children and adolescents in order to study factors associated with the onset of myopia. The goal is to explore whether genetic or behavioral aspects (such as reading habits or use of glasses) are related to the presence or severity of myopia.

Name	Type	Description
ID	Numerical	Unique subject identifier (1 to 618)
STUDYYEAR	Numerical	Year in which the subject entered the study
MYOPIC	Categorical	1 if subject developed myopia, 0 otherwise
AGE	Numerical	Age at first visit (in years)
GENDER	Categorical	0 = male, 1 = female
SPHEQ	Numerical	Spherical equivalent refraction (diopters)
AL	Numerical	Axial length of the eye (in mm)
ACD	Numerical	Anterior chamber depth (in mm)
LT	Numerical	Lens thickness (in mm)
VCD	Numerical	Vitreous chamber depth (in mm)
SPORTHR	Numerical	Weekly hours spent in sports/outdoor activities
READHR	Numerical	Weekly hours spent reading for pleasure
COMPHR	Numerical	Weekly hours spent on computer
STUDYHR	Numerical	Weekly hours spent studying
TVHR	Numerical	Weekly hours spent watching television
DIOPTERHR	Numerical	Composite index of near-work activity
MOMMY	Categorical	1 if subject's mother is myopic, 0 otherwise
DADMY	Categorical	1 if subject's father is myopic, 0 otherwise

Table 1: List of attributes for `myopia.csv`

[1] Link to the dataset: kaggle.com/datasets/mscgeorges/myopia-study

1.1 Research Questions

1. Do myopic children spend more than 3 hours per week reading for pleasure?
2. Is the presence of myopia independent of the child's gender?
3. Is the time spent by myopic children on different activities equal to that of non-myopic children?
4. Is there a difference in the proportion of myopic children between those with myopic parents and those without?
5. Is there a correlation between the refraction error and time spent on near visual activities?

1.2 Possible Applications

The findings of this analysis may help understanding whether behavioral factors—such as time spent reading, studying, or using digital devices—or genetic predispositions, such as parental history of myopia, are more strongly associated with the onset of myopia in children and adolescents.

Understanding these relationships could inform the design of early screening strategies and targeted interventions. For example, identifying behavioral risk factors may support the development of educational programs aimed at reducing excessive near-work activities, while evidence of genetic influence may encourage earlier ophthalmologic evaluations for at-risk individuals.

2 House Prices

The *House Prices Dataset*[2] contains records of house sales in King County, USA, which includes Seattle and surrounding areas. The data was collected over a period of one year and includes a wide range of structural, spatial, and geographical features for each property. The main objective is to investigate how these features influence the final sale price of a property.

Name	Type	Description
ID	Numerical	Unique identifier for each property
DATE	Date	Date the house was sold
PRICE	Numerical	Sale price of the house (target variable)
BEDROOMS	Numerical	Number of bedrooms
BATHROOMS	Numerical	Number of bathrooms (including half baths)
SQFT_LIVING	Numerical	Interior living space (in square feet)
SQFT_LOT	Numerical	Lot size (in square feet)
FLOORS	Numerical	Number of floors
WATERFRONT	Categorical	1 if the house has waterfront view, 0 otherwise
VIEW	Numerical	Index indicating quality of view (0–4)
CONDITION	Numerical	Condition of the house (1 = poor, 5 = excellent)
GRADE	Numerical	Construction and design grade (1–13)
SQFT_ABOVE	Numerical	Square footage above the basement
SQFT_BASEMENT	Numerical	Square footage of the basement
YR_BUILT	Numerical	Year the house was built
YR_RENOVATED	Numerical	Year of renovation (0 if never renovated)
ZIPCODE	Numerical	ZIP code of the property
LAT	Numerical	Latitude coordinate
LONG	Numerical	Longitude coordinate
SQFT_LIVING15	Numerical	Living space of 15 nearest neighbors
SQFT_LOT15	Numerical	Lot area of 15 nearest neighbors

Table 2: List of attributes for `house_price_seattle.csv`

[2] Link to the dataset: www.kaggle.com/datasets/harlfoxem/housesalesprediction

2.1 Research Questions

1. Can house prices be accuratly predicted using only one or two key features?
2. Which features have the highest impact on the predicted sale price?
3. How does the condition or quality of a house affect its market value?

2.2 Possible Applications

Understanding the main factors that influence house prices can be really useful for various stakeholders in the real estate market.

Real estate agents can use these insights to better estimate property values based on key structural and locational features. Homeowners and investors may rely on simplified predictive models to assess the potential value of renovations or to evaluate the pricing of comparable properties. Moreover, identifying the most influential features can support data-driven decisions in urban planning and construction.

3 Heart Disease

The *Heart Disease Dataset*[3] contains clinical and demographic data for a group of patients undergoing cardiac assessment. The dataset includes medical measurements with the goal of predicting the presence or absence of heart disease.

It is composed of a series of variables related to cardiovascular health, such as blood pressure, cholesterol levels, and exercise-induced symptoms, along with personal factors like age and sex.

Name	Type	Description
AGE	Numerical	Age of the patient (in years)
SEX	Categorical	1 = male, 0 = female
CP	Categorical	Chest pain type (4 possible values)
TRESTBPS	Numerical	Resting blood pressure (in mm Hg)
CHOL	Numerical	Serum cholesterol (in mg/dl)
FBS	Categorical	Fasting blood sugar ¿ 120 mg/dl (1 = true, 0 = false)
RESTECG	Categorical	Resting ECG results
THALACH	Numerical	Maximum heart rate achieved
EXANG	Categorical	Exercise-induced angina (1 = yes, 0 = no)
OLDPEAK	Numerical	ST depression induced by exercise
SLOPE	Categorical	Slope of the peak exercise ST segment
CA	Numerical	Number of major vessels (0–3)
THAL	Categorical	3 = normal, 6 = fixed defect, 7 = reversible defect
TARGET	Categorical	1 = presence of heart disease, 0 = absence

Table 3: List of attributes for `heart.csv`

[3] Link to the dataset: https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset

3.1 Research Questions

1. Can heart disease be predicted accurately using basic clinical measurements such as cholesterol, blood pressure, and heart rate?

2. Does the predictive performance of classification models differ between males and females?

3. Are ST depression and maximum heart rate sufficient to detect heart disease patterns?

3.2 Possible Applications

The analysis of this dataset provides valuable insights into the early detection of heart disease based on simple, non-invasive clinical measurements. These results can assist healthcare professionals in identifying high-risk patients through easily accessible clinical information.

Moreover, understanding which clinical features most strongly differ between affected and unaffected individuals can inform more targeted screening programs and preventive strategies.