# Dataset Descriptions and Data Dictionaries

## 1 Auto dataset

Filename: auto.csv

Description: The dataset consists of data for automobiles from an auto insurance company. Each row represents an automobile with various attributes associated with it, including physical and performance characteristics. The dataset consists of 205 rows and 25 columns.

| Name | Description | Type | Details |
|---|---|---|---|
| Risk_Rating | The degree of riskiness for insuring the auto | Categorical | From 1 through to 3 in discrete levels. Higher value implies higher risk |
| Make | Make of the auto | Categorical | "Alfa-Romero", "Audi", "Bmw", "Chevrolet", "Dodge", "Honda", "Isuzu", "Jaguar", "Mazda", "Mercedes-Benz", "Mercury", "Mitsubishi", "Nissan", "Peugot", "Plymouth", "Porsche", "Renault", "Saab", "Subaru", "Toyota", "Volkswagen", "Volvo" |
| Fuel_Type | Fuel type | Categorical | “Diesel”, “Gas” |
| Aspiration | Engine aspiration | Categorical | “Std”, “Turbo” |
| Doors | Number of doors | Categorical | “Two”, “Four” |
| Body | Body style | Categorical | “Hardtop”, “Wagon”, “Sedan”, “Hatchback”, “Convertible” |
| Drive_Wheels | Drive system | Categorical | “4WD”, “FWD”, “RWD” |
| Engine_Loc | Engine location | Categorical | “Front”, “Rear” |
| Wheel_Base | Wheel base | Numerical | |
| Length | Length | Numerical | |
| Width | Width | Numerical | |
| Height | Height | Numerical | |
| Curb_Weight | Curb weight | Numerical | |
| Engine_Type | Engine type | Categorical | "DOHC", "DOHCV", "L", "OHC", "OHCF", "OHCV", "ROTOR" |
| Num_Of_Cylinders | Number of cylinders | Categorical | Eight", "Five", "Four", "Six", "Three", "Twelve", "Two" |
| Engine_Size | Engine size | Numerical | |
| Fuel_System | Fuel system | Categorical | "1BBL", "2BBL", "4BBL", "IDI", "MFI", "MPFI", "SPDI", "SPFI" |
| Bore | Bore | Numerical | |
| Stroke | Stroke | Numerical | |
| Compression_Ratio | Compression ratio | Numerical | |
| Horsepower | Horsepower | Numerical | |
| Peak_RPM | Peak RPM | Numerical | |
| City_MPG | MPG when driven in city | Numerical | |
| Highway_MPG | MPG when driven on highway | Numerical | |
| Price | Price | Numerical | |

# 2 Housing dataset

Filename: housing.csv

Description: The dataset contains data from a census survey with regards to housing in a region. Each row corresponds to a town, hence the values of each attribute is for each town. The dataset contains 506 rows and 14 columns.

| Name | Description | Type | Details |
|---|---|---|---|
| ID | Identification Number | Categorical | |
| Crime_Rate | Crime rate per capita | Categorical | "High", "Moderate", "Low" |
| Residential_Area | Percentage of residential area greater than 25000 square feet | Categorical | "Very High", "High", "Moderate", "Low", "Very Low" |
| Non_Retail | Percentage of business areas used for non-retail | Numerical | |
| River | Near bridge | Categorical | "Near", "Far" |
| Pollutants | Concentration of pollutants (parts per 10 million) | Numerical | |
| Rooms | Number of rooms per unit | Numerical | |
| Old_Housing | Percentage of units built before second world war | Numerical | |
| Distance_Office | Weighted distances to major office areas | Numerical | |
| Highway_Access | Accessibility to major highways | Categorical | "High", "Moderate", "Low" |
| Tax_Rate | Property tax rate per $10,000 | Numerical | |
| Student_Teacher_Ratio | Student to teacher ratio | Numerical | |
| Low_Income | Percentage of low income households | Numerical | |
| Price | Median price of homes (in $1000s) | Numerical | |

# 3 Wine dataset

Filename: wine.csv

Description: The dataset contains data related to the quality of wine. Each row of data contains physicochemical properties of a wine sample. The dataset contains 4898 rows and 13 columns.

| Name | Description | Type | Details |
|---|---|---|---|
| ID | Identification Number | Categorical | |
| Fixed_Acidity | Tartaric acid, measured in g/dm3 | Numerical | |
| Volatile_Acidity | Acetic acid, measured in g/dm3 | Numerical | |
| Citric_Acid | Measured in g/dm3 | Numerical | |
| Residual_Sugar | Level of residual sugar content | Categorical | "High", "Moderate", "Low" |
| Chlorides | Sodium chloride, measured in g/dm3 | Numerical | |
| Free_SO2 | Measured in mg/dm3 | Numerical | |
| Total_SO2 | Measured in mg/dm3 | Numerical | |
| Density | Measured in g/dm3 | Categorical | "<0.99", "0.99-1.00", ">1.00" |
| pH | pH value | Categorical | "Neutral", "Acidic", "Highly Acidic" |
| Sulphates | Potassium sulphate, measured in g/dm3 | Numerical | |
| Alcohol | Level of alcohol content | Categorical | 8% through 14% in discrete levels |
| Quality | Score based on sensory data | Categorical | 0 through 10 (0 is bad, 10 is good) |