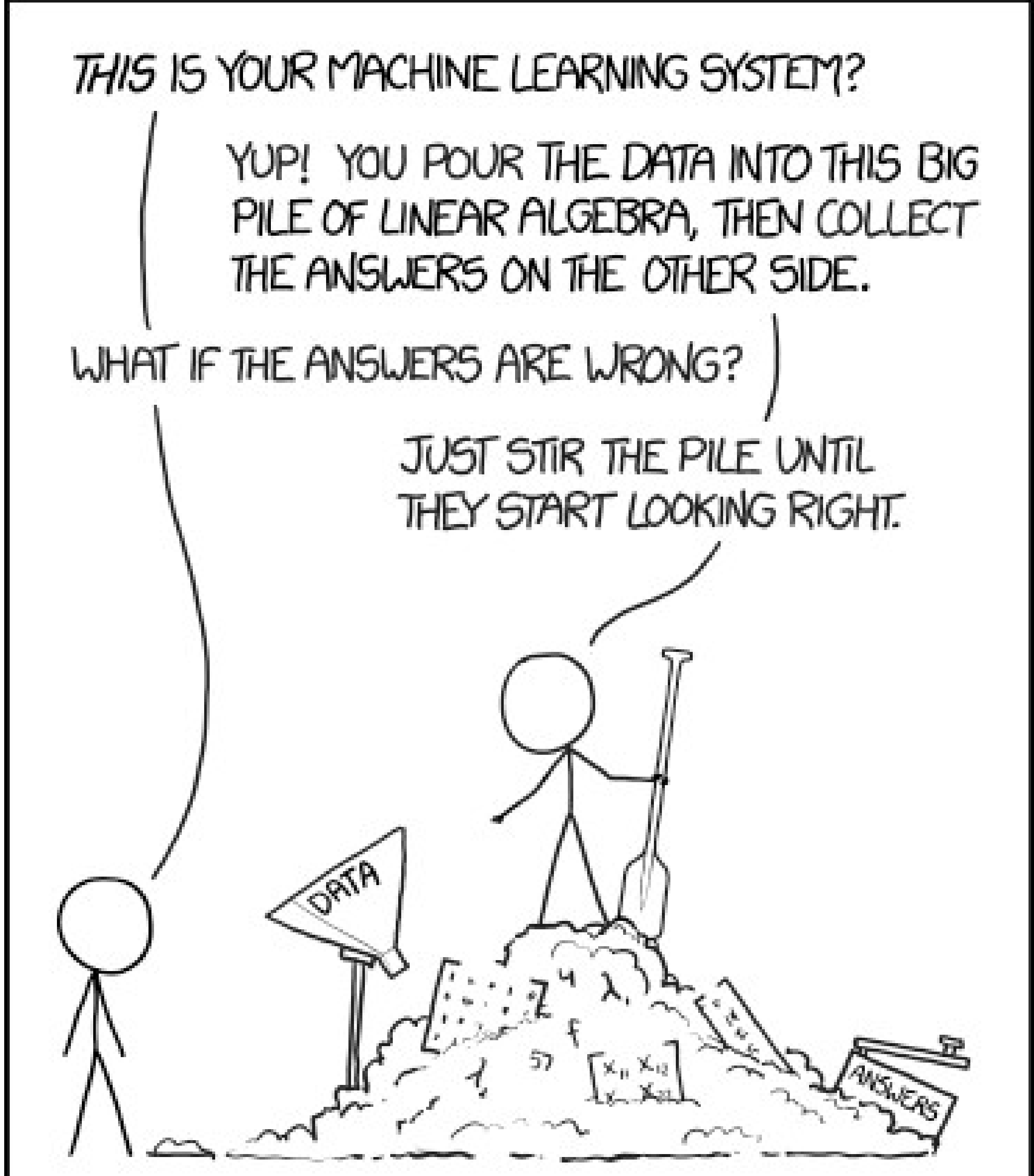
# *Spatial* Data Science

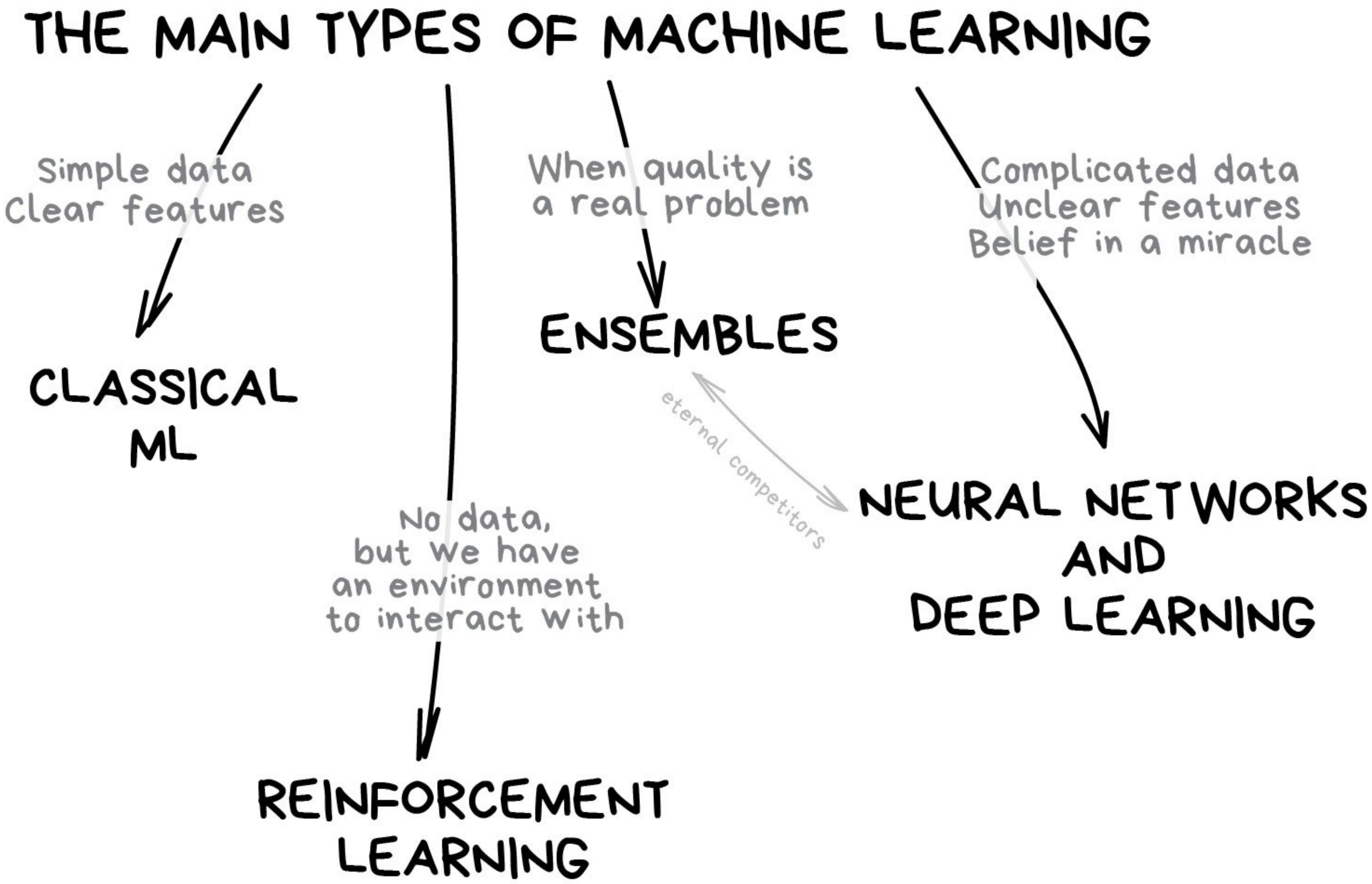
## Machine Learning for Everyone

Lecture 5

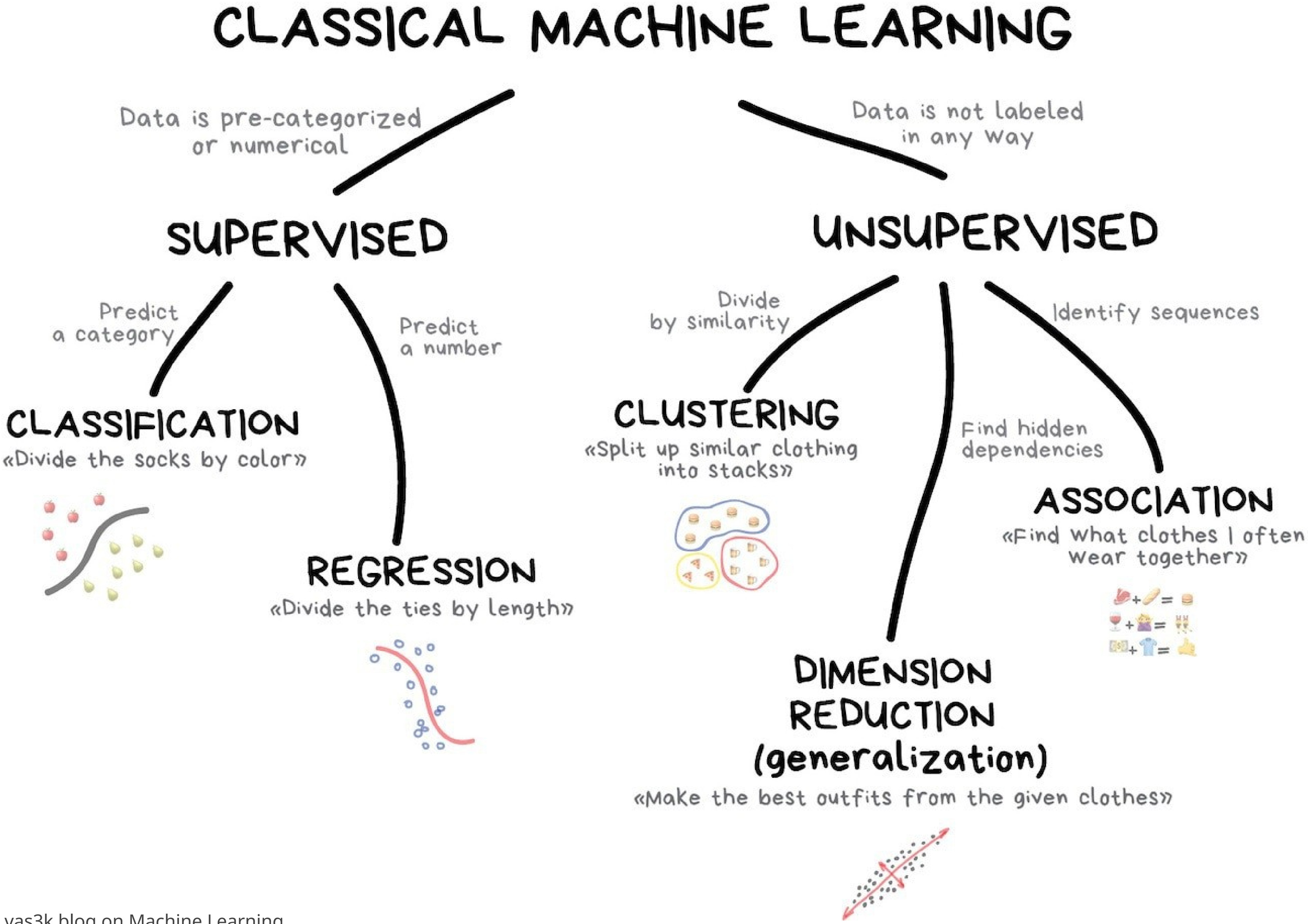
Theodoros Chatzivasileiadis


THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

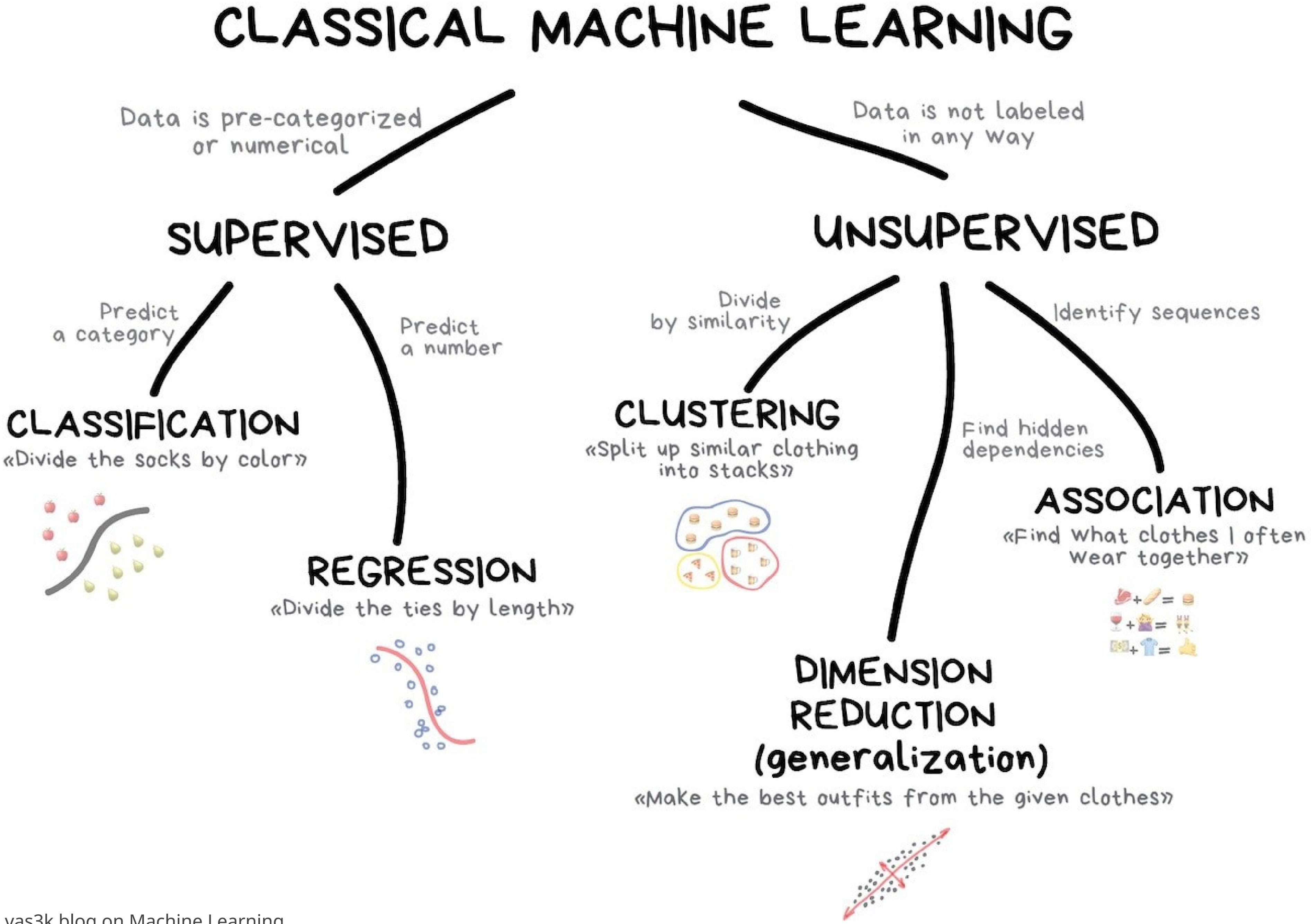
XKCD Creative Commons Attribution-NonCommercial 2.5 Licen


THE MAIN TYPES OF MACHINE LEARNING
Simple data
Clear features
CLASSICAL ML
No data, but we have an environment to interact with
REINFORCEMENT LEARNING
When quality is a real problem
ENSEMBLES
eternal competitors
Complicated data
Unclear features
Belief in a miracle
NEURAL NETWORKS AND DEEP LEARNING

nages taken from the vas3k blog on Machine Learning


CLASSICAL MACHINE LEARNING
Data is pre-categorized or numerical
Data is not labeled in any way
SUPERVISED
UNSUPERVISED
Predict a category
Predict a number
Divide by similarity
Find hidden dependencies
Identify sequences
CLASSIFICATION
«Divide the socks by color»
REGRESSION
«Divide the ties by length»
CLUSTERING
«Split up similar clothing into stacks»
ASSOCIATION
«Find what clothes I often wear together»
DIMENSION REDUCTION (generalization)
«Make the best outfits from the given clothes»

nages taken from the vas3k blog on Machine Learning


CLASSICAL MACHINE LEARNING
Data is pre-categorized or numerical
Data is not labeled in any way
SUPERVISED
UNSUPERVISED
Predict a category
Predict a number
Divide by similarity
Identify sequences
Find hidden dependencies
CLASSIFICATION
«Divide the socks by color»
REGRESSION
«Divide the ties by length»
CLUSTERING
«Split up similar clothing into stacks»
ASSOCIATION
«Find what clothes I often wear together»
DIMENSION REDUCTION (generalization)
«Make the best outfits from the given clothes»

nages taken from the vas3k blog on Machine Learning

# What is a regression model

**Regression analysis** is a statistical method that helps us understand and predict how different factors are related. By analyzing data, it shows how changes in one variable (like population density) might affect another (such as traffic flow or housing demand).

# What is a regression model

Regression analysis is a **statistical method** that helps us understand and predict how different factors are related. By analyzing data, it shows how changes in one variable (like population density) might affect another (such as traffic flow or housing demand).

# What is a regression model

Regression analysis is a statistical method that helps us **understand and predict** how different factors are related. By analyzing data, it shows how changes in one variable (like population density) might affect another (such as traffic flow or housing demand).

# What is a regression model

Regression analysis is a statistical method that helps us understand and predict **how different factors are related**.

# What is a regression model

Regression analysis is a statistical method that helps us understand and predict **how different factors are related**.

**By analyzing data, it shows how changes in one variable (like population density) might affect another (such as traffic flow or housing demand).**

This tool enables us to make **informed** decisions by forecasting outcomes and identifying key influences in urban design and planning.

- Regression analysis is a statistical method that helps us understand and predict **how different factors are related**.

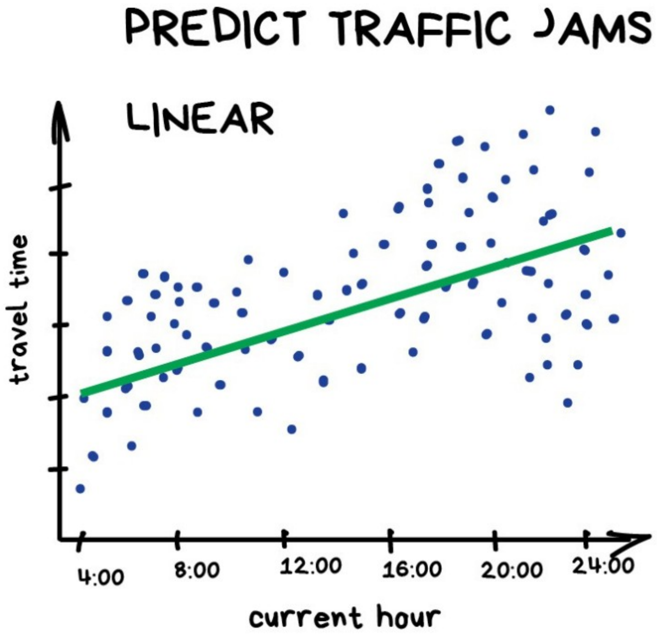
- Each regression model has **at least** two variables:

1. A Dependent Variable: Always referred as y or Y
2. An Independent Variable: Referred as x.

- **Dependent Variable**: This is the outcome you're interested in predicting or understanding. It's called "dependent" because its value depends on other factors. For architects and spatial planners, this could be variables like property values, energy consumption of a building, or the level of foot traffic in a public space.

- **Independent Variables**: These are the factors that you suspect have an influence on the dependent variable. They are "independent" because they are presumed to cause or explain changes in the dependent variable, **not the other way around.** Examples include the distance to public transportation, availability of green spaces, building materials used, or population density.

In essence, the regression model helps you analyze how changes in independent variables affect the dependent variable, enabling you to make data-driven decisions in your projects.



$$Y_i = \alpha_0 + \beta_1 * x_i + \varepsilon_i$$

When the line is straight — it's a linear regression, when it's curved – polynomial. If the model has more than 1 dependent variables it is a multivariate regression model

nages taken from the vas3k blog on Machine Learning

$$Y_i = \alpha_0 + \beta_1 * x_{1i} + \varepsilon_i$$

- $Y_i$ : The dependent Variable (Travel time)
- $\alpha_0$: The intercept, or constant
- $\beta_1$: The Coefficient of x on Y
- $x_{1i}$: The independent variable (Time of the day)
- $\varepsilon_i$: The error term or residual

# Response vs. Predictor Variables

$X = X_1, \dots, X_p$

$X_j = x_{1j}, \dots, x_{ij}, \dots, x_{nj}$

**Independent**

$Y = y_1, \dots, y_n$

outcome

**response** variable

dependent variable

***i*** observations

| TV budget | Radio budget | Newspaper budget | sales |
|---|---|---|---|
| 230.1 | 37.8 | 69.2 | 22.1 |
| 44.5 | 39.3 | 45.1 | 10.4 |
| 17.2 | 45.9 | 69.3 | 9.3 |
| 151.5 | 41.3 | 58.5 | 18.5 |
| 180.8 | 10.8 | 58.4 | 12.9 |

***p*** predictors

This example and all material are provided by the Data Science course at Harvard Unive

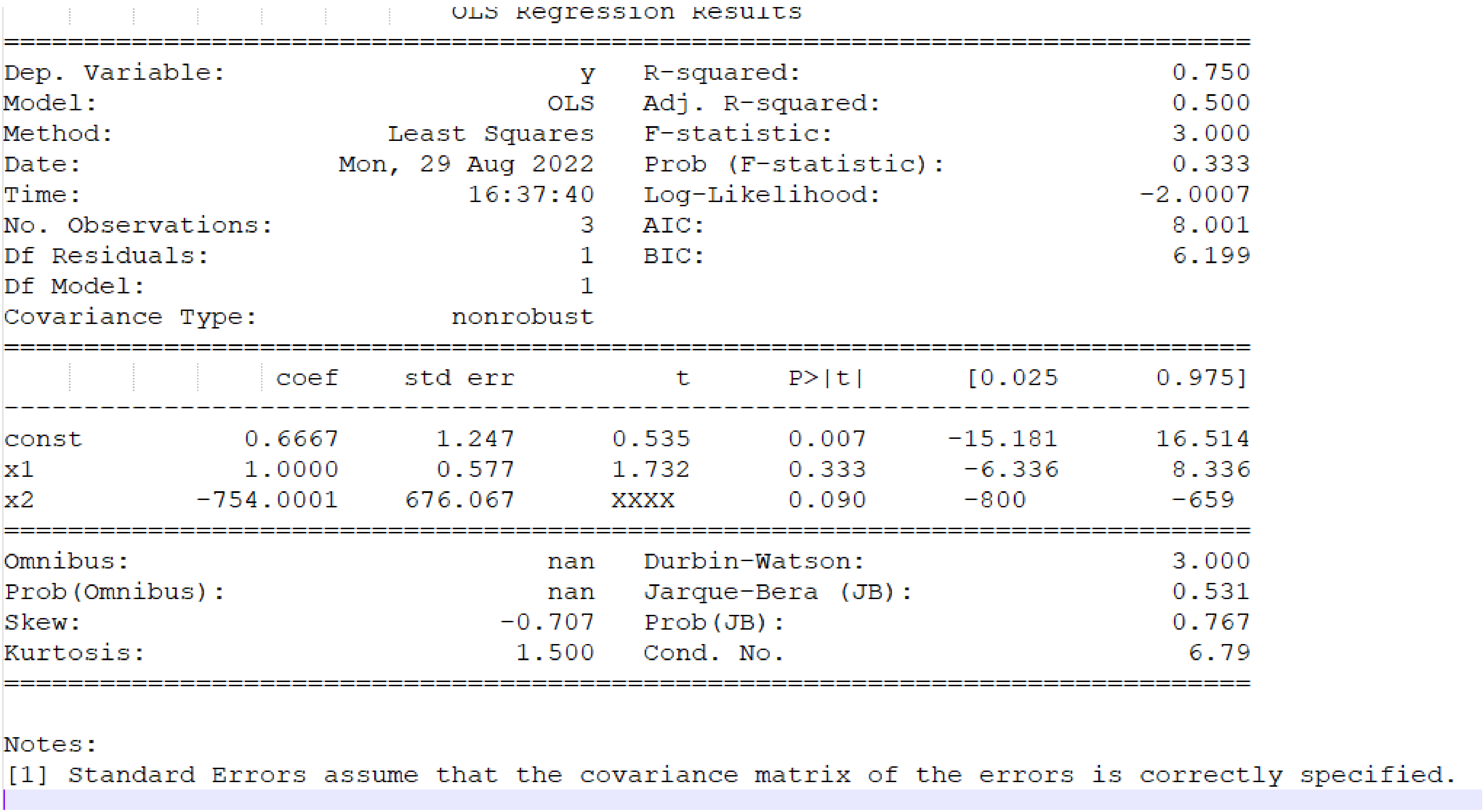
$$Y_i = \alpha_0 + \beta_1 * x_{1i} + \varepsilon_i$$

| TV budget | sales |
|---|---|
| 230.1 | 22.1 |
| 44.5 | 10.4 |
| 17.2 | 9.3 |
| 151.5 | 18.5 |
| 180.8 | 12.9 |

- Here what we are asking is: How much does TV budget (if any) does effect sales of a product.

$$Sales_i = \alpha_0 + \beta_1 * TV\_Budget_i + \varepsilon_i$$

IF Sales = 0 if TV_Budget =0, the $\alpha_0$ = 0. In our case however sales would not be 0 without advertisement, so we have an intercept. That is the number of sales **irrespective** of the independent variable.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.750
Model:                            OLS   Adj. R-squared:                  0.500
Method:                 Least Squares   F-statistic:                     3.000
Date:                Mon, 29 Aug 2022   Prob (F-statistic):              0.333
Time:                        16:37:40   Log-Likelihood:                -2.0007
No. Observations:                   3   AIC:                             8.001
Df Residuals:                       1   BIC:                             6.199
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.6667      1.247      0.535      0.007     -15.181      16.514
x1             1.0000      0.577      1.732      0.333      -6.336       8.336
x2          -754.0001    676.067      XXXX       0.090      -800        -659
==============================================================================
Omnibus:                          nan   Durbin-Watson:                   3.000
Prob(Omnibus):                    nan   Jarque-Bera (JB):                0.531
Skew:                          -0.707   Prob(JB):                        0.767
Kurtosis:                       1.500   Cond. No.                         6.79
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

# Lets go to an actual problem

What are the drivers of **population density in Amsterdam?**

We have a dataset available with these variables,

- **Urban Green Space (%)**
- **Public Transport Density (stations/km²)**
- **Proximity to City Center (km)**
- **Residential Zoning (%)**
- **Distance to Nearest Park (km)**
- **Employment Accessibility (jobs/km²)**

- **And of course Population Dencity.**

- **R-squared**: 0.71
- **F-statistic**: 26.43
- **Prob(F-statistic)**: 0.000
- **Log-Likelihood**: -156.3
- **Observations**: 150

| Variable | Coefficient | Std. Error | t-Statistic | P-value | 95% Confidence Interval |
|---|---|---|---|---|---|
| **Intercept** | 5.123 | 0.742 | 6.90 | 0.000 | (3.669, 6.577) |
| **Urban Green Space (%)** | 0.245 | 0.112 | 2.19 | 0.029 | (0.024, 0.467) |
| **Public Transport Density (stations/km²)** | 0.362 | 0.093 | 3.89 | 0.000 | (0.179, 0.545) |
| **Proximity to City Center (km)** | -0.432 | 0.091 | -4.75 | 0.000 | (-0.611, -0.253) |
| **Residential Zoning (%)** | 0.178 | 0.057 | 3.12 | 0.002 | (0.066, 0.290) |
| **Distance to Nearest Park (km)** | -0.219 | 0.083 | -2.64 | 0.009 | (-0.383, -0.055) |
| **Employment Accessibilit** | 0.487 | 0.145 | 3.36 | 0.001 | (0.199, 0.775) |

- **R-squared**: 0.62
- **F-statistic**: 26.43
- **Prob(F-statistic)**: 0.000
- **Observations**: 150

| **Variable** | **Coefficient** | **Std. Error** | **t-Statistic** | **P-value** | **95% Confidence Interval** |
|---|---|---|---|---|---|
| **Intercept** | 5.123 | 0.742 | 6.90 | 0.000 | (3.669, 6.577) |
| **Public Transport Density (stations/km$^2$)** | 0.362 | 0.093 | 3.89 | 0.000 | (0.179, 0.545) |
| **Proximity to City Center (km)** | -0.432 | 0.091 | -4.75 | 0.000 | (-0.611, -0.253) |
| **Residential Zoning (%)** | 0.178 | 0.057 | 3.12 | 0.002 | (0.066, 0.290) |
| **Distance to Nearest Park (km)** | -0.219 | 0.083 | -2.64 | 0.009 | (-0.383, -0.055) |

- **R-squared**: 0.02
- **F-statistic**: 0.43
- **Prob(F-statistic): 0.600**
- **Observations**: 150

| **Variable** | **Coefficient** | **Std. Error** | **t-Statistic** | **P-value** | **95% Confidence Interval** |
|---|---|---|---|---|---|
| **Intercept** | 5.123 | 0.742 | 6.90 | 0.000 | (3.669, 6.577) |
| **Employment Accessibility (jobs/km$^2$)** | 0.487 | 0.145 | 3.36 | 0.001 | (0.199, 0.775) |