Group 12

PREDICTING INDIVIDUAL INCOME THROUGH HOUSEHOLD DEMOGRAPHIC STATISTICS: A DATA-DRIVEN APPROACH

110502528 HSUN-HAO CHANG
110502529 PO-SHEN CHEN
110502534 CHUN-YU CHEN
110502009 HUNG-YI HSU

data science final

Start Slide

Source of Dataset

Household Income and Expenditure Survey

Group 12

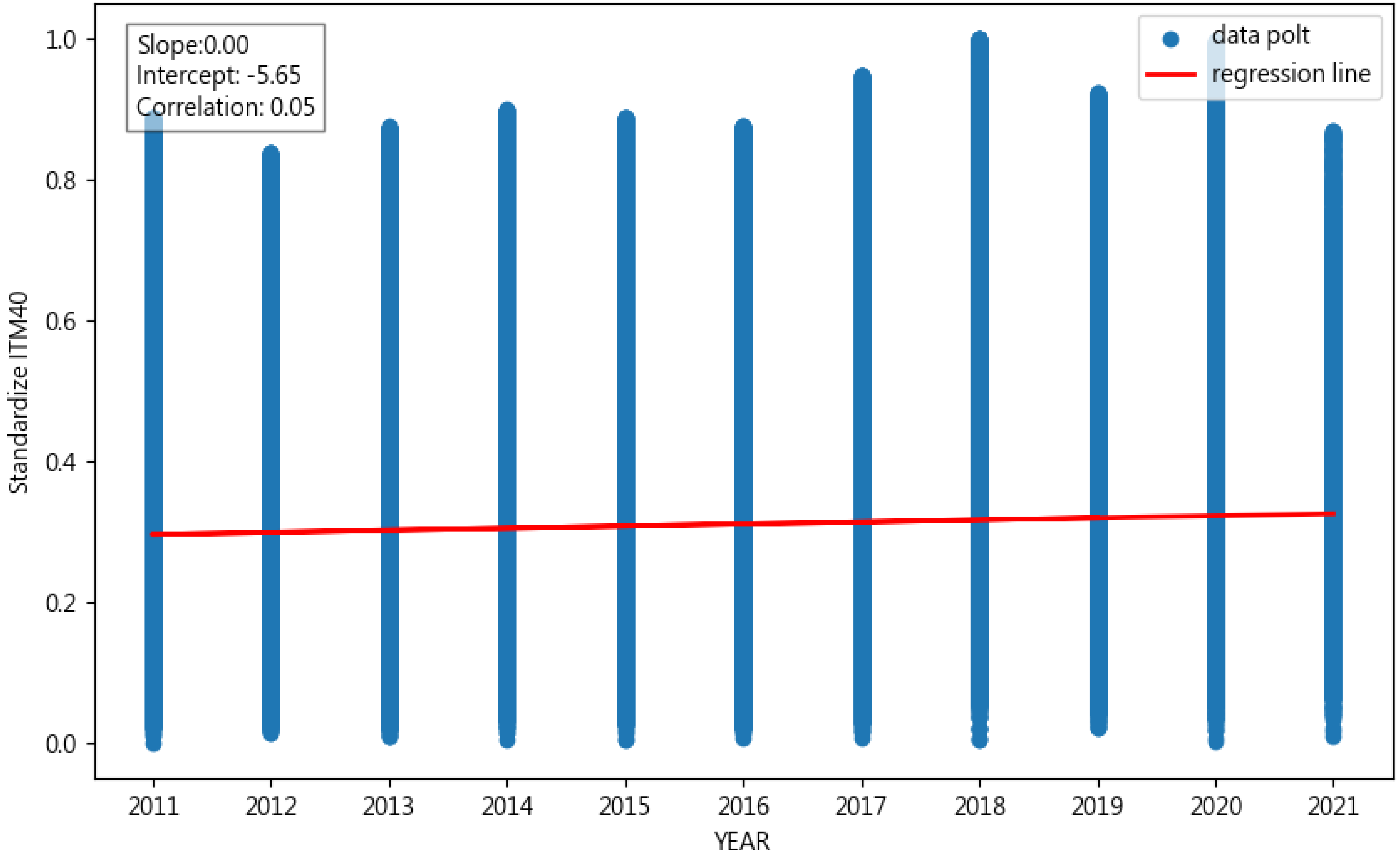
Dataset Visualization

YEAR

INCOME VS. YEARS

The Ideas

- Income rises with years.
- Growth rate slows with time.
- Correlation between income and years exists but not strong.
- Shows moderate correlation despite variability.



Group 12

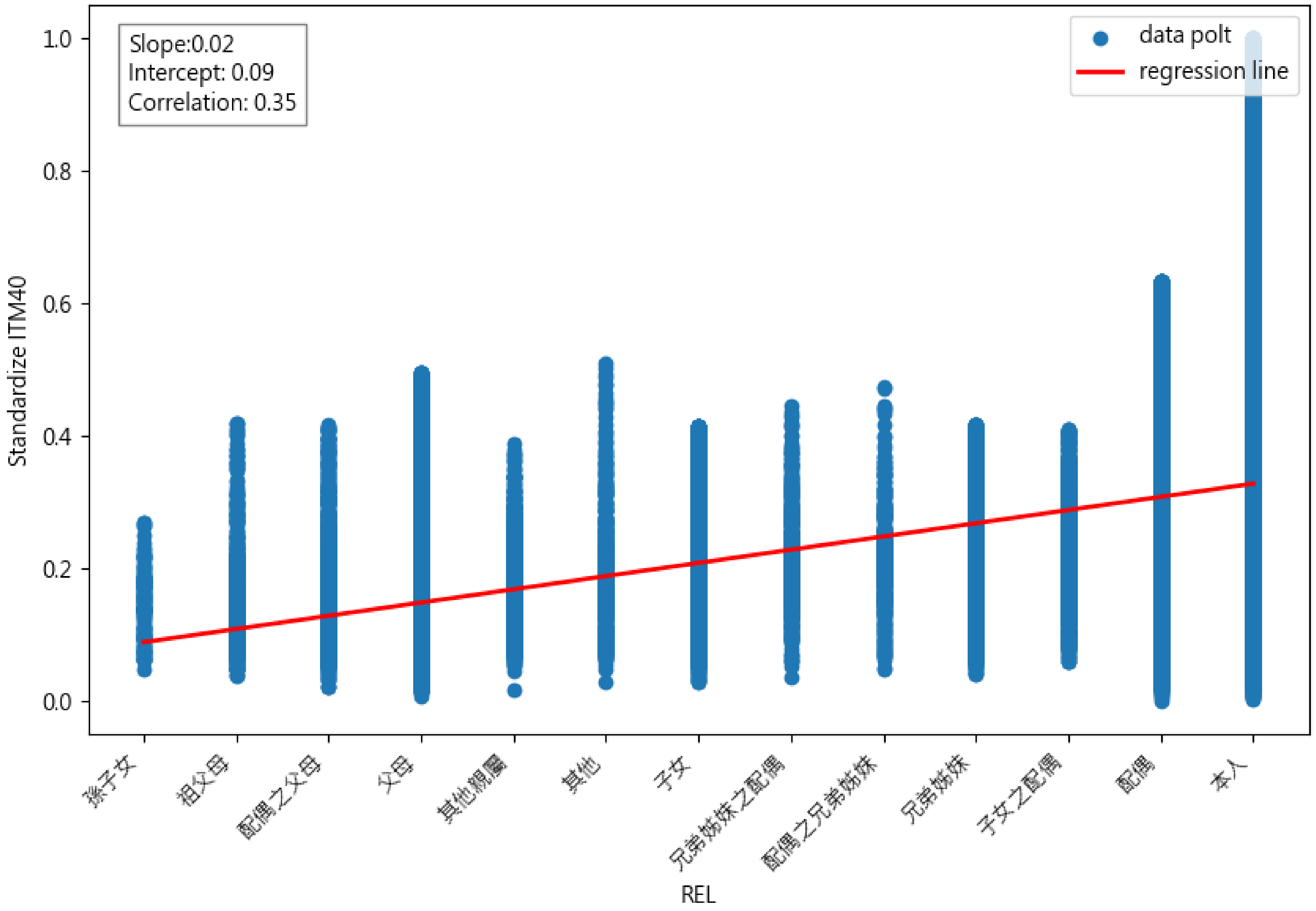
Dataset Visualization

REL

INCOME VS. FAMILY MEMBER TITLES

The Ideas

- Linear regression line fits well.
- Indicates correlation between head of household and income.
- Positive slope suggests closer relation = higher income.
- Relationship with Head of Household (REL) is significant in explaining income



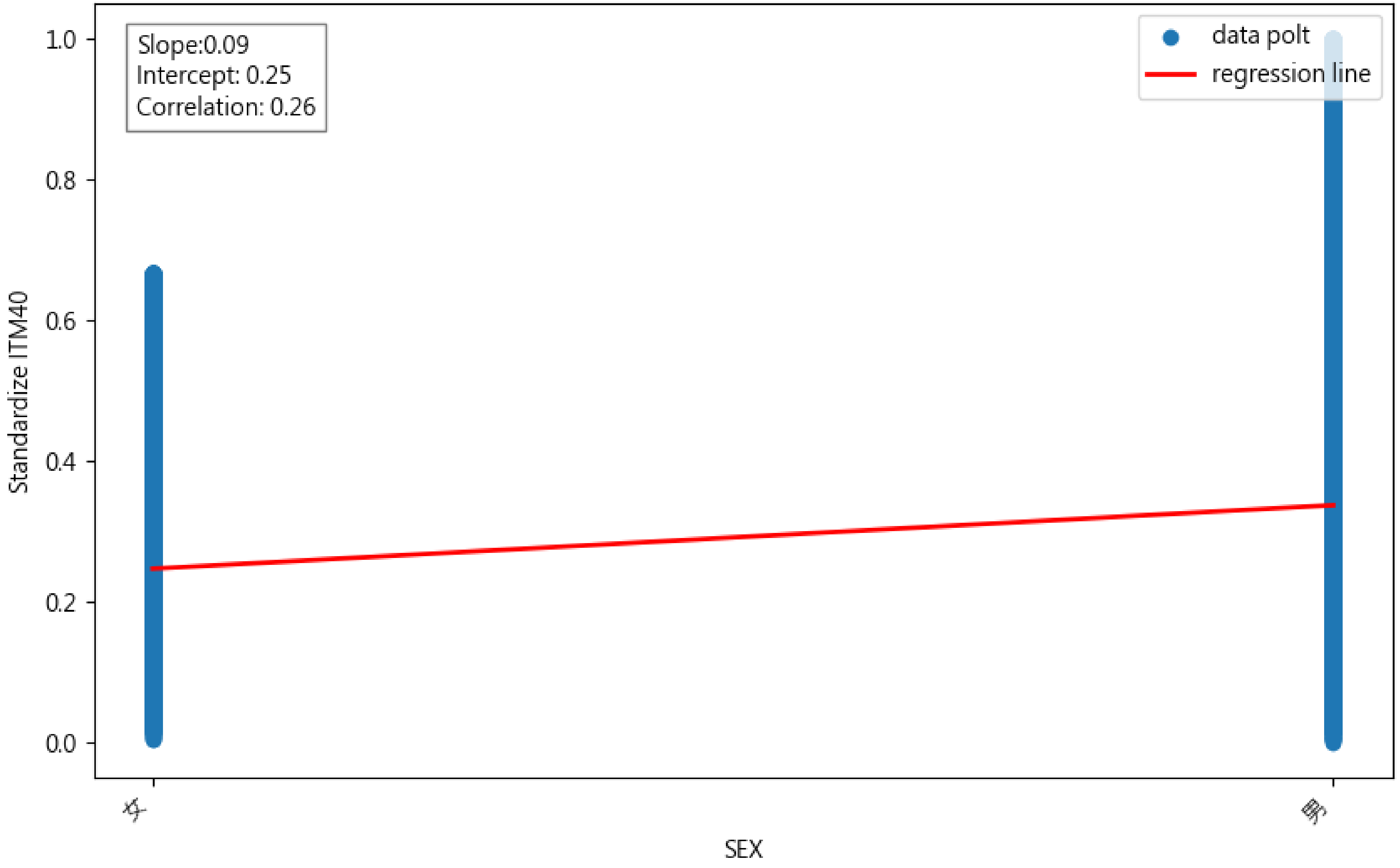
Dataset Visualization

SEX

INCOME VS. GENDER

The Ideas

- Linear regression line fits well.
- Men have higher incomes than women (positive slope).
- Low-income range: women’s are higher.
- High-income range: men’s are higher.
- Gender is a significant variable in explaining income differences.



Group 12

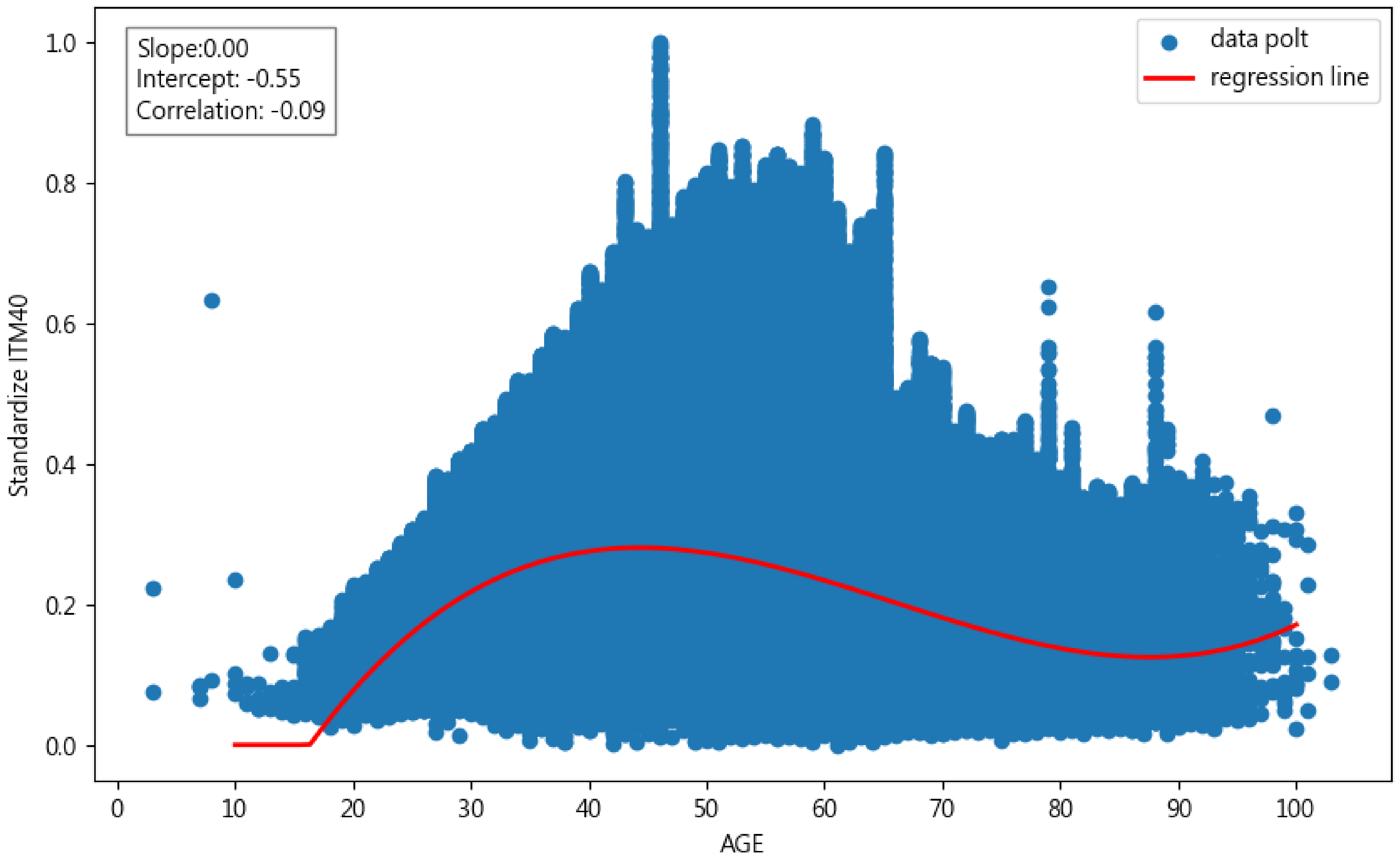
Dataset Visualization

AGE

INCOME VS. AGE

The Ideas

- Peak income: ages 50 to just before 60.
- Distribution resembles a normal curve.
- Steep rise before 50, quadratic pattern after 60.
- Cubic function used for better fitting the chart's distribution.



Group 12

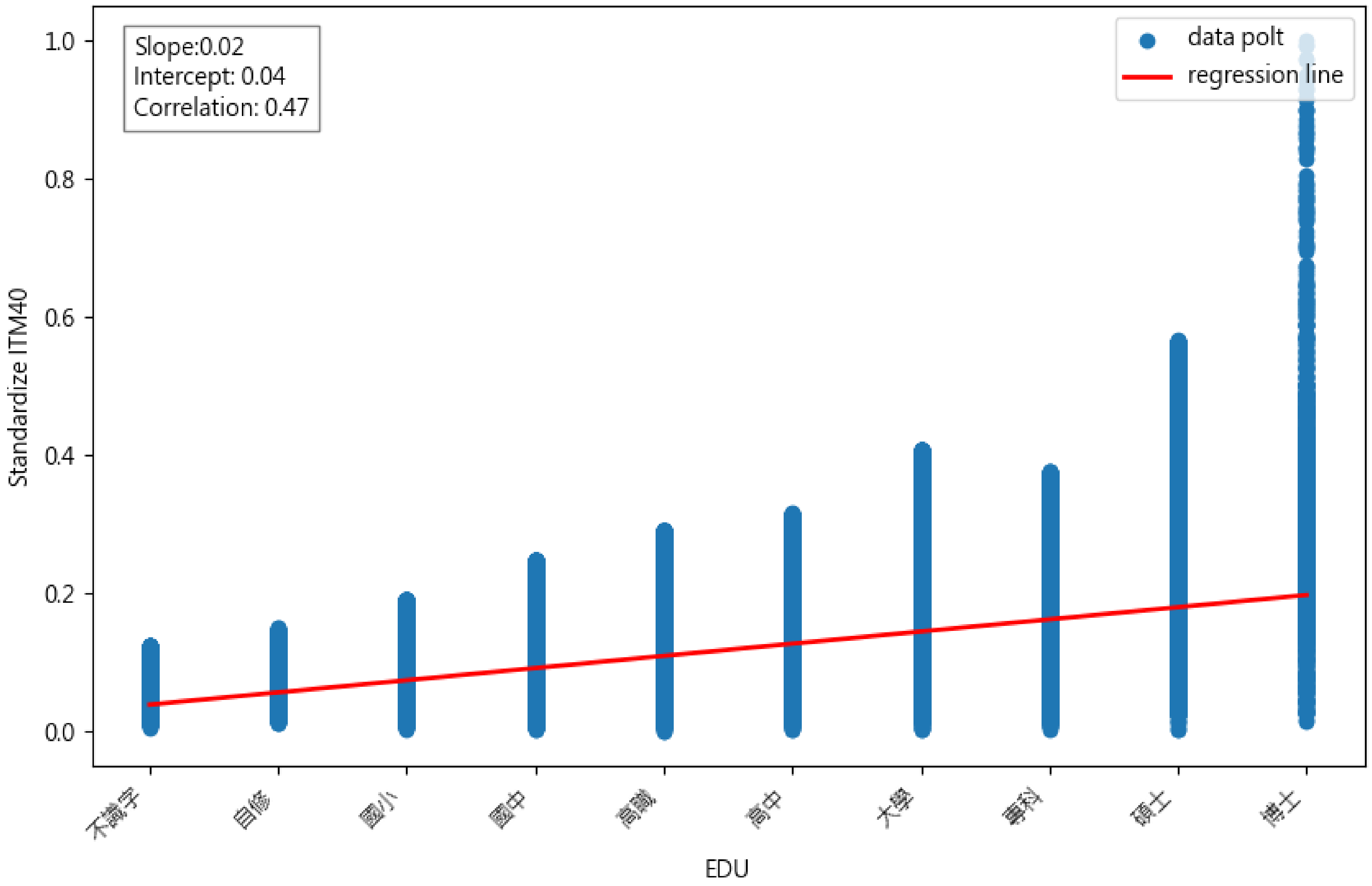
Dataset Visualization

EDU

INCOME VS. EDUCATION

The Ideas

- Higher education = higher income.
- Aligns with common knowledge.
- Education level is a significant differentiator and emphasizes the importance of further education pursuit.



Group 12

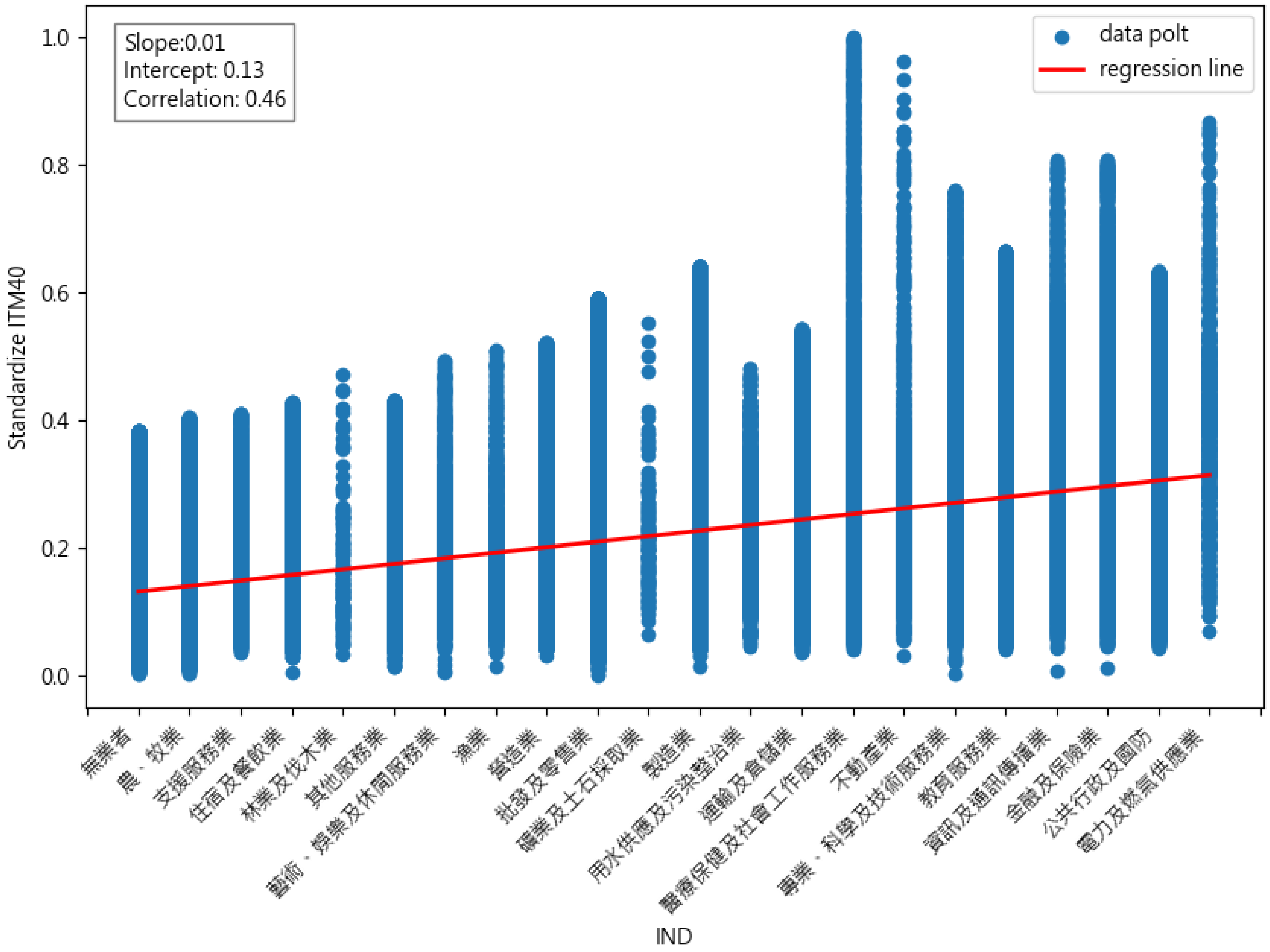
Dataset Visualization

IND

INCOME VS. INDUSTRIES

The Ideas

- Organized by average income per industry.
- Variation in income levels among industries.
- Some industries show broader income distributions.
- Healthcare, social assistance, and real estate industries notably wide-ranging.
- Challenges in predictions due to wide income ranges and potentially lower sample sizes in certain sectors.



Group 12

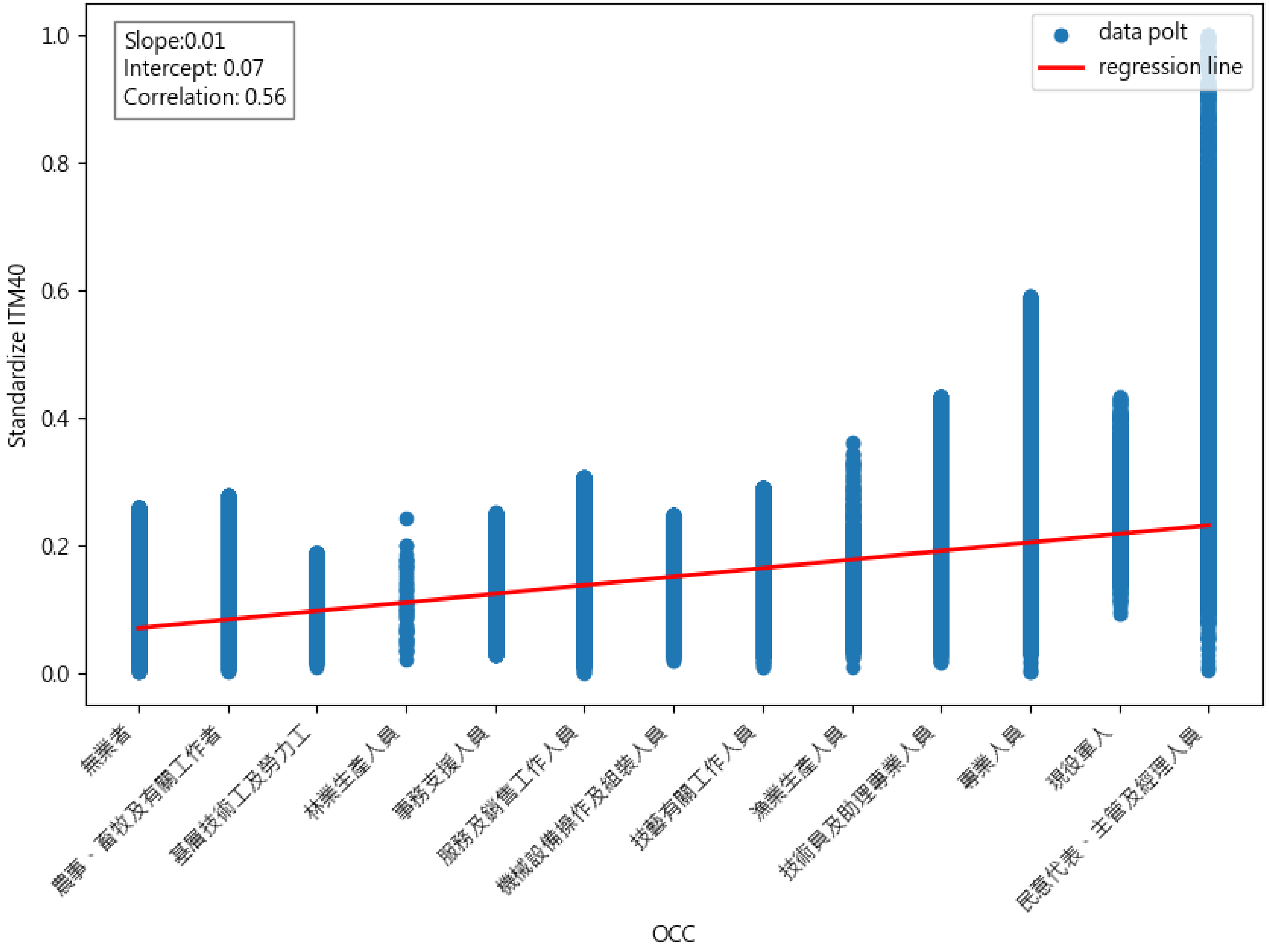
Dataset Visualization

OCC

INCOME VS. OCCUPATION

The Ideas

- Chart shows occupation vs. income.
- Ranks average incomes by occupation.
- Confirms higher incomes for representatives, supervisors, managers.
- Indicates lower salaries for frontline workers and production personnel.



Group 12

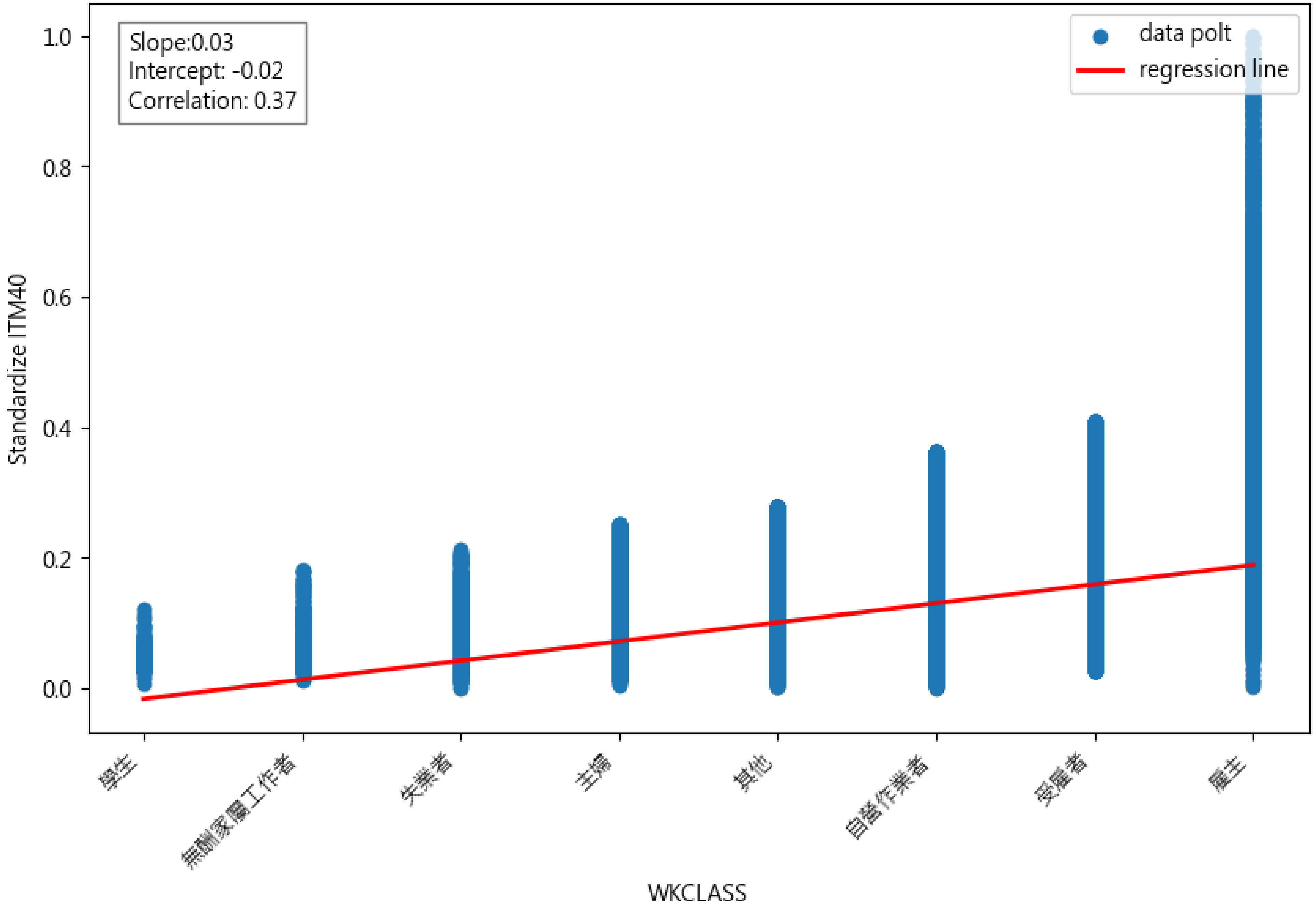
Dataset Visualization

WKClass

INCOME VS. EMPLOYMENT ROLE

The Ideas

- Categories: students, unpaid family workers, unemployed, homemakers, self-employed, employees, employers, others.
- Employer income notably highest.
- Significant margin above other categories, even doubling the second-highest category.



Group 12

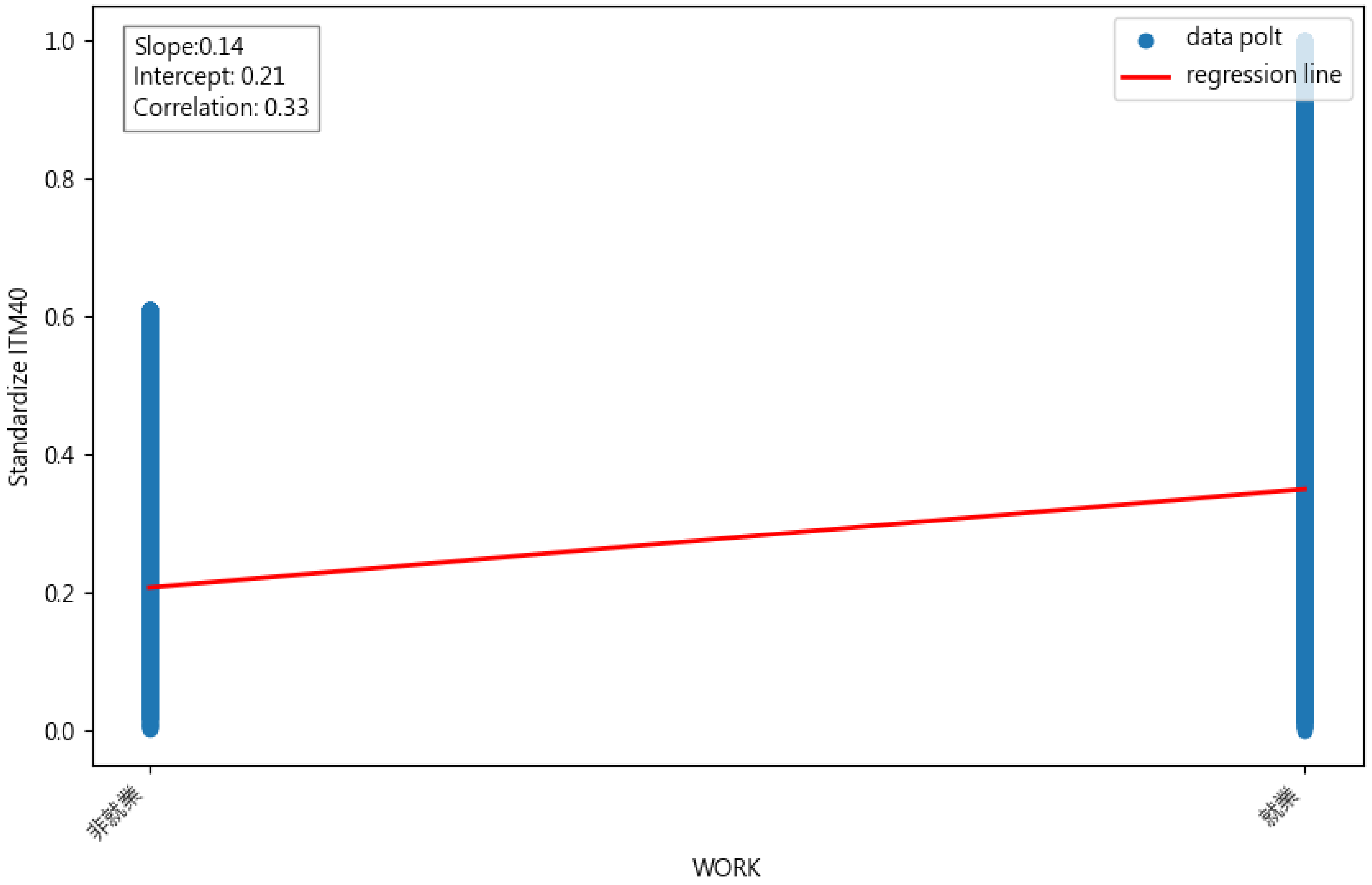
Dataset Visualization

WORK

INCOME VS. EMPLOYMENT STATUS

The Ideas

- Employed individuals tend to have higher incomes than unemployed individuals.



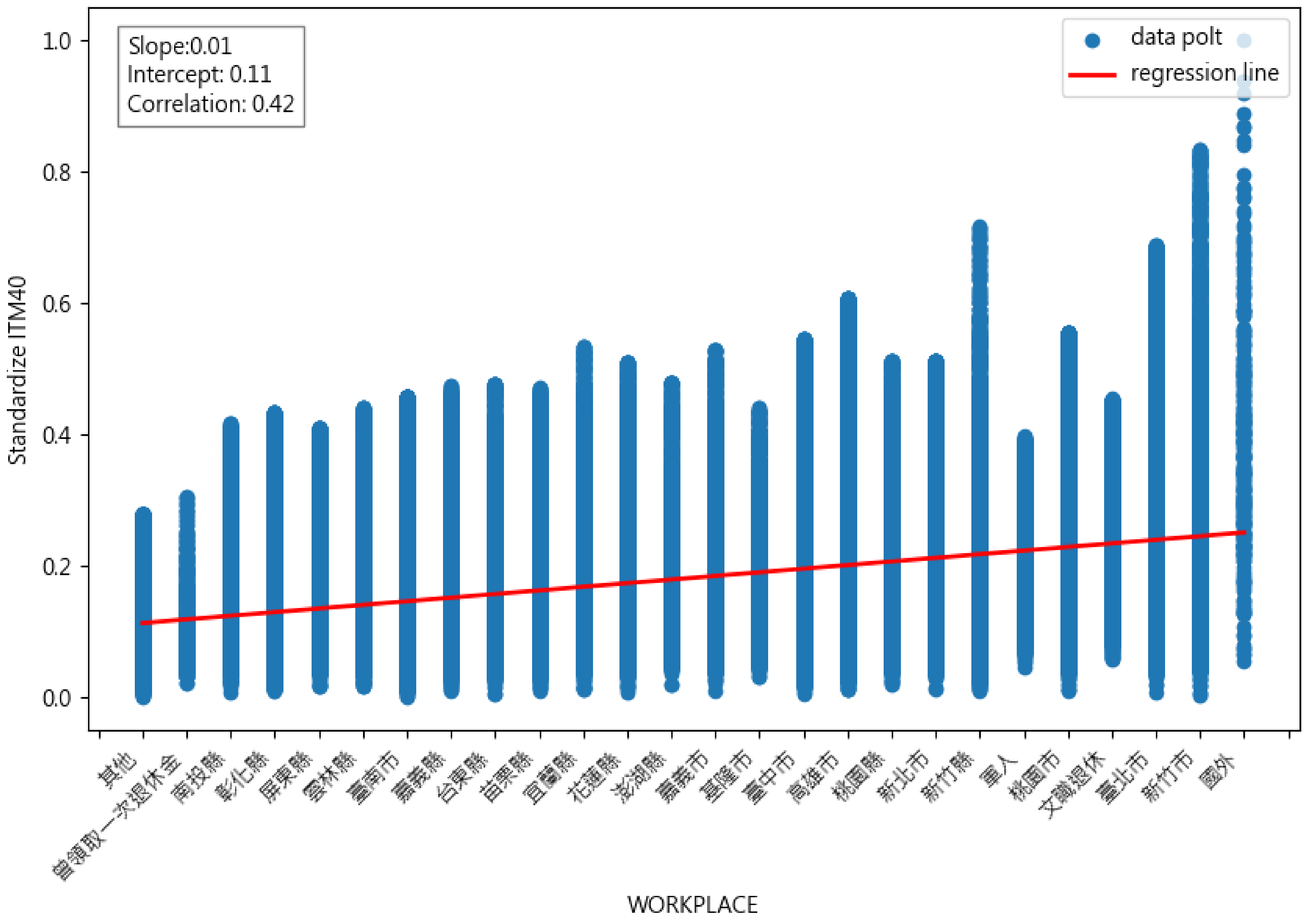
Dataset Visualization

WORKPlace

INCOME VS. WORK LOCATION IN TAIWAN

The Ideas

- Working abroad associated with highest income.
- Northern region and special municipalities also have relatively higher incomes.
- Peripheral areas in central and southern regions, along with those receiving one-time retirement pensions, tend to have lower incomes.



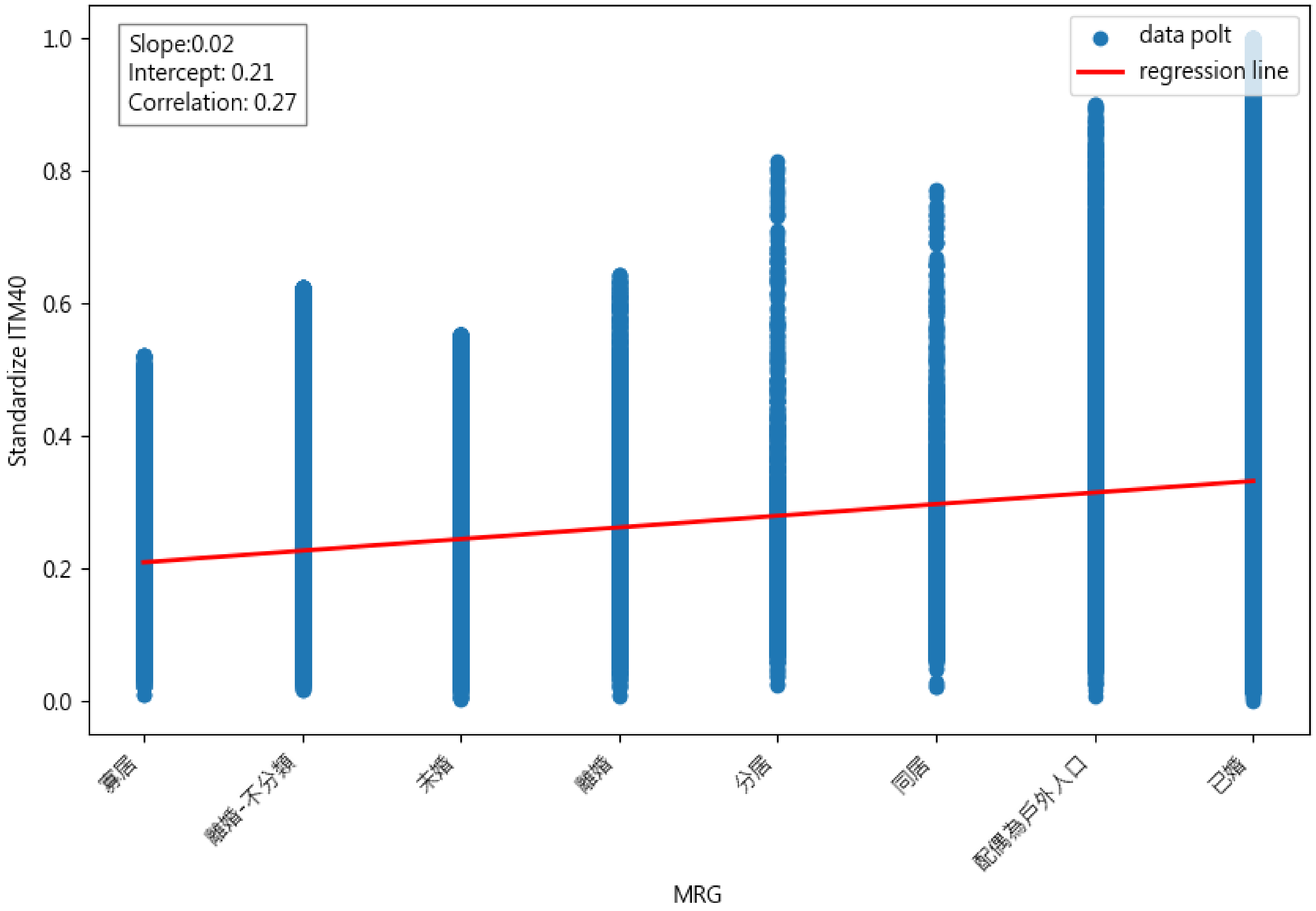
Dataset Visualization

MRG

INCOME VS. MARITAL STATUS

The Ideas

- Chart illustrates marital status vs. income.
- Generally, higher income for married individuals compared to unmarried.
- Suggests both partners having income sources contributes to higher overall income compared to reliance on a single income source for unmarried individuals.



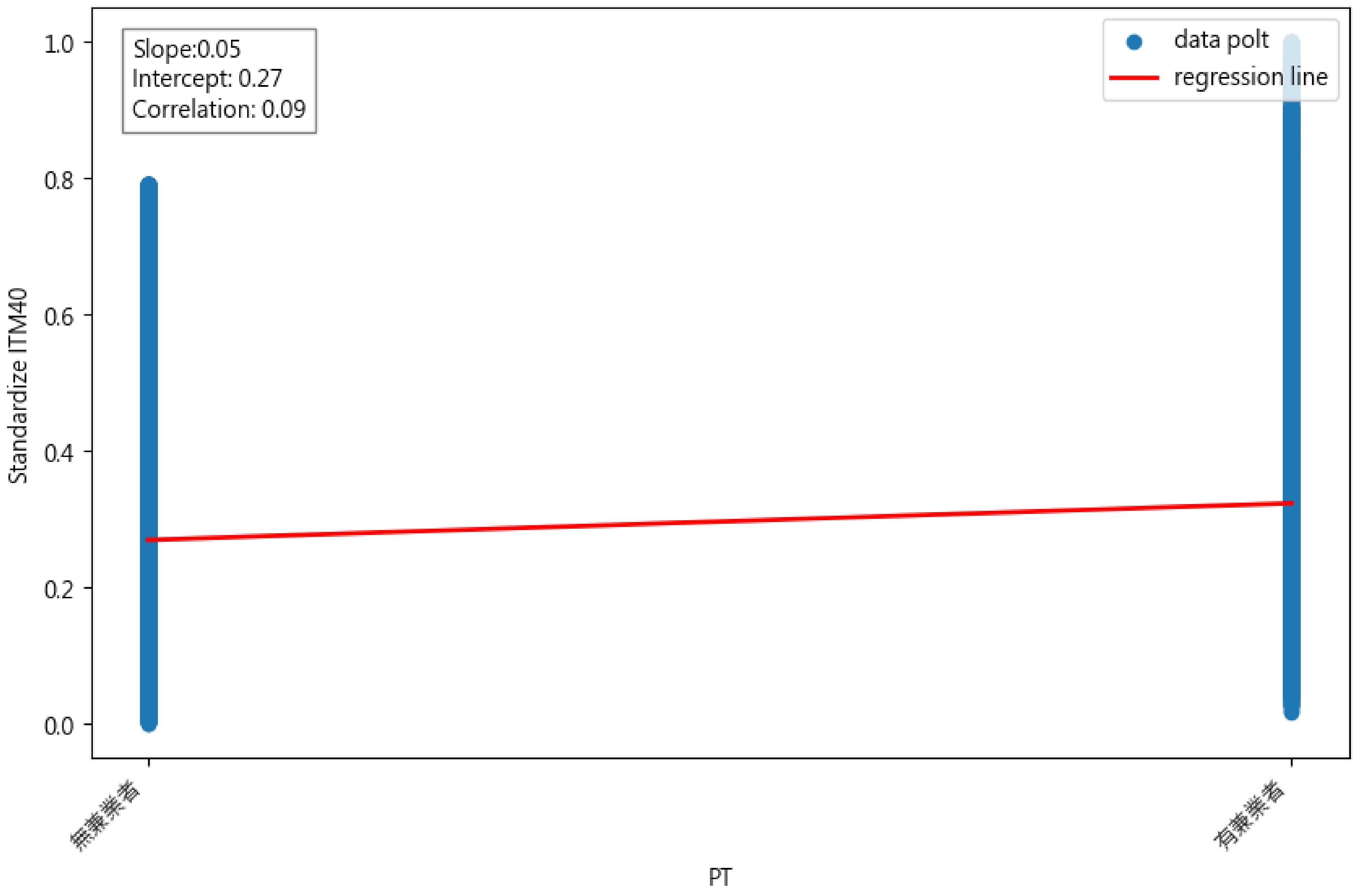
Dataset Visualization

PT

INCOME VS. RELATIONSHIP BETWEEN MULTIPLE JOBS (MOONLIGHTING)

The Ideas

- Individuals with multiple jobs have higher income compared to those without additional employment.



Methodology

tree Models	Svm Models
Neighbors Models	Linear Models
Neural Network Models	Cross Decomposition Models
Ensemble Models	Popular Models(on Kaggle)

Experiments Drop out Results

Table 1: MAE with One-Hot Encoding

Name	Test Loss	Train Loss
RANSAC	1.914991×10^{11}	2.076992×10^{10}
Linear	2.010680×10^{11}	3.853654×10^{-1}

Table 2: MAE without One-Hot Encoding

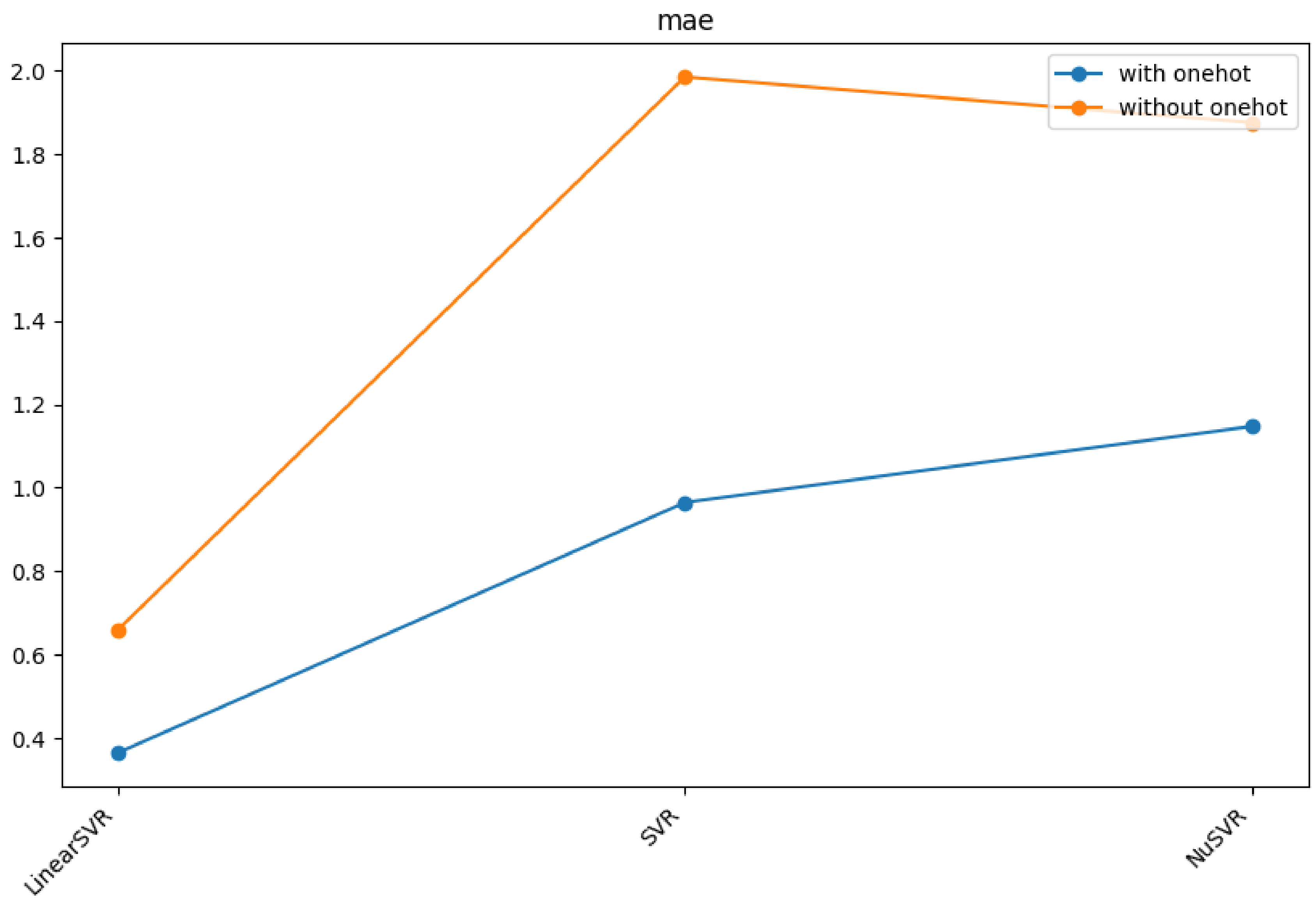
Name	Test Loss	Train Loss
SGD	3.377136×10^{14}	3.373692×10^{14}

Table 3: MSE with One-Hot Encoding

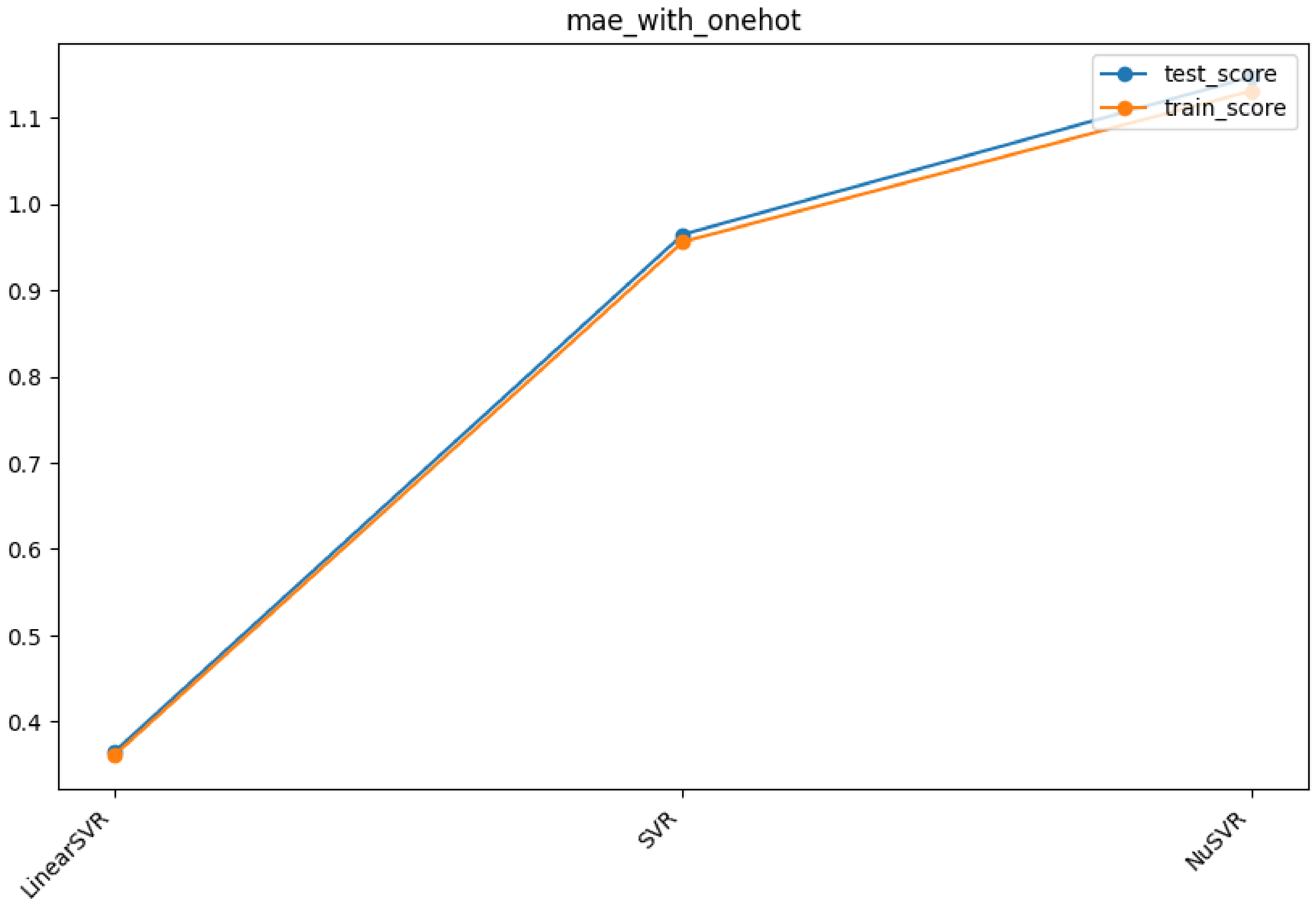
Name	Test Loss	Train Loss
Linear	9.477466×10^{22}	5.720158×10^{-1}
RANSAC	1.373472×10^{23}	5.160883×10^{21}

Experiments SVM

Group 12



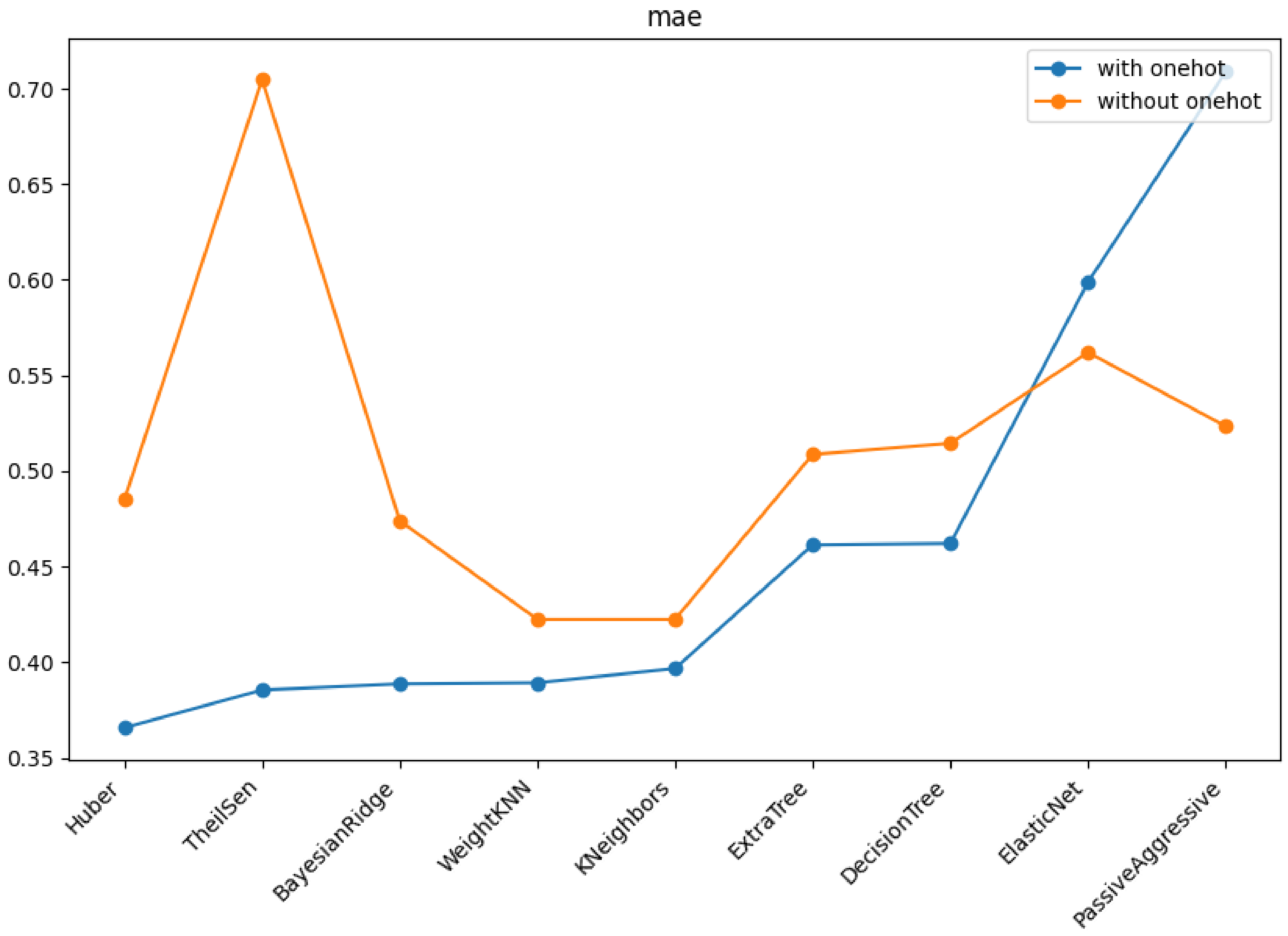
ONE-HOT OR NOT

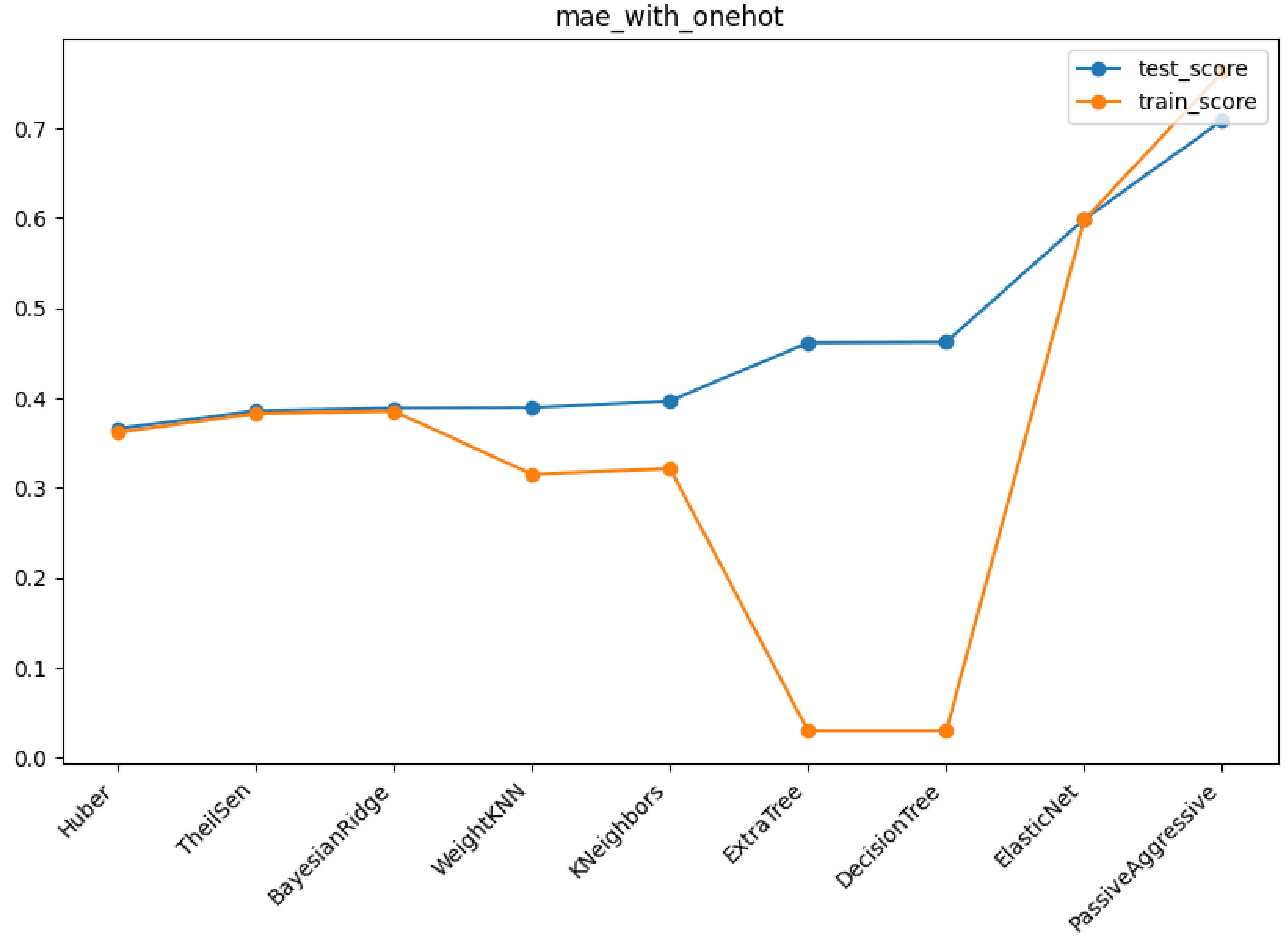


TRAIN VS TEST

Experiments Traditional models

Group 12



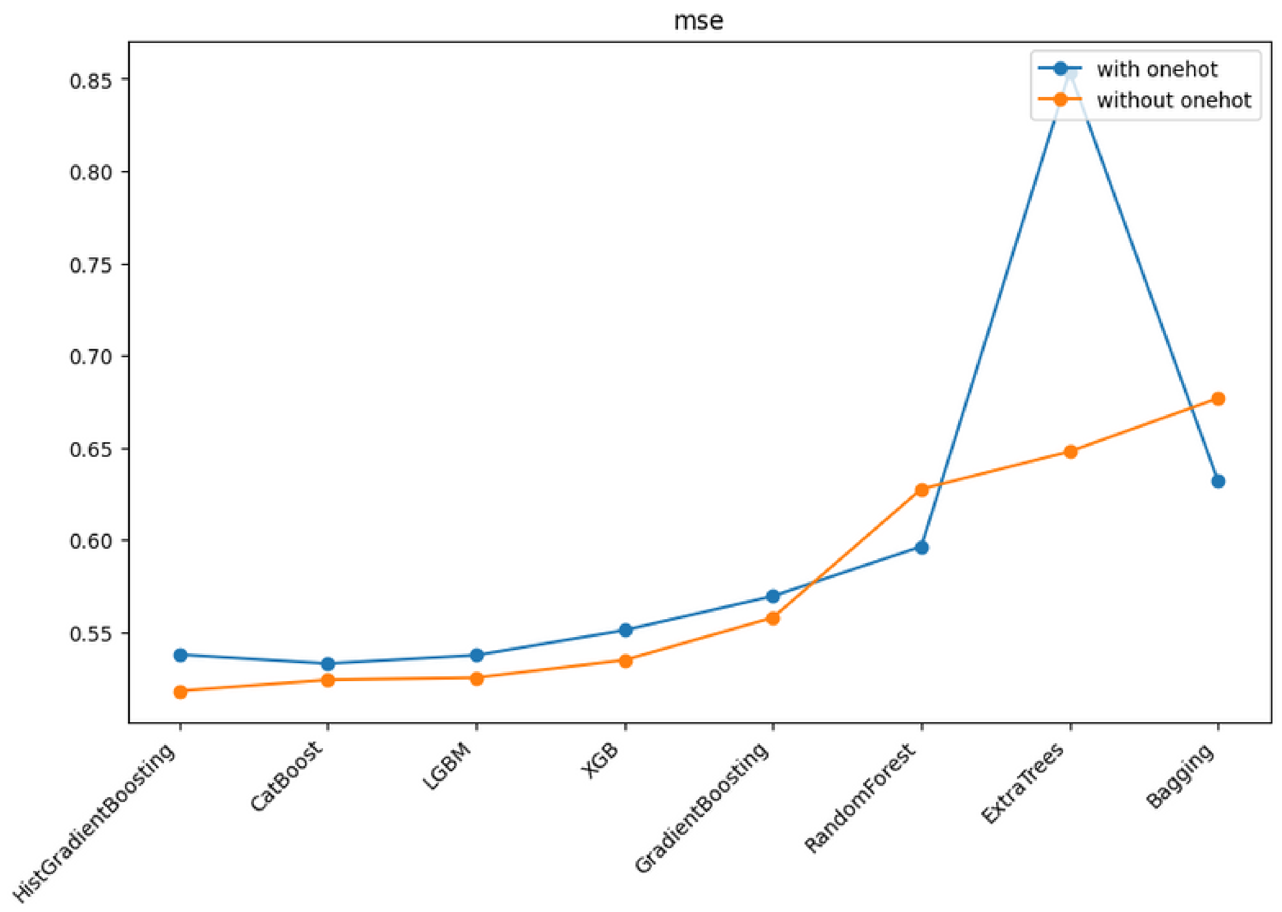


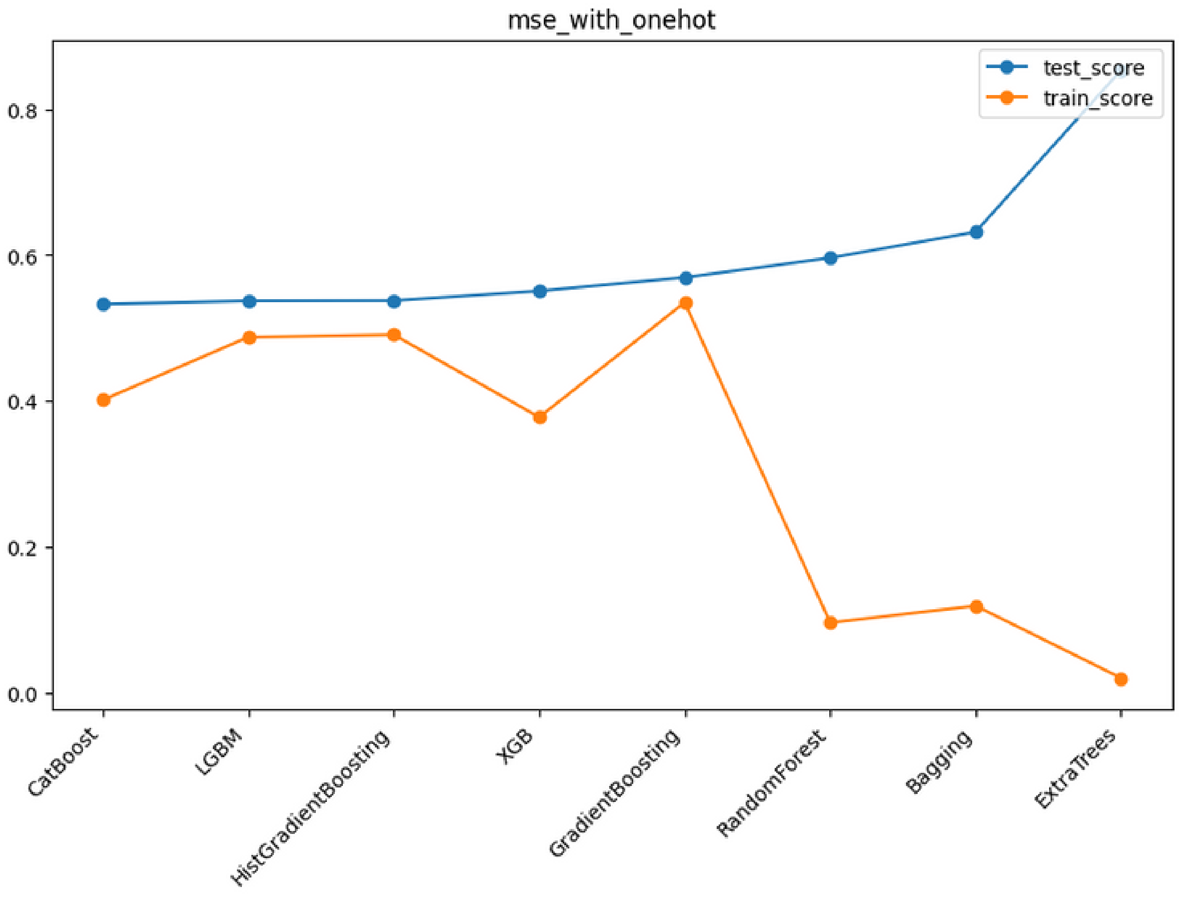
ONE-HOT OR NOT

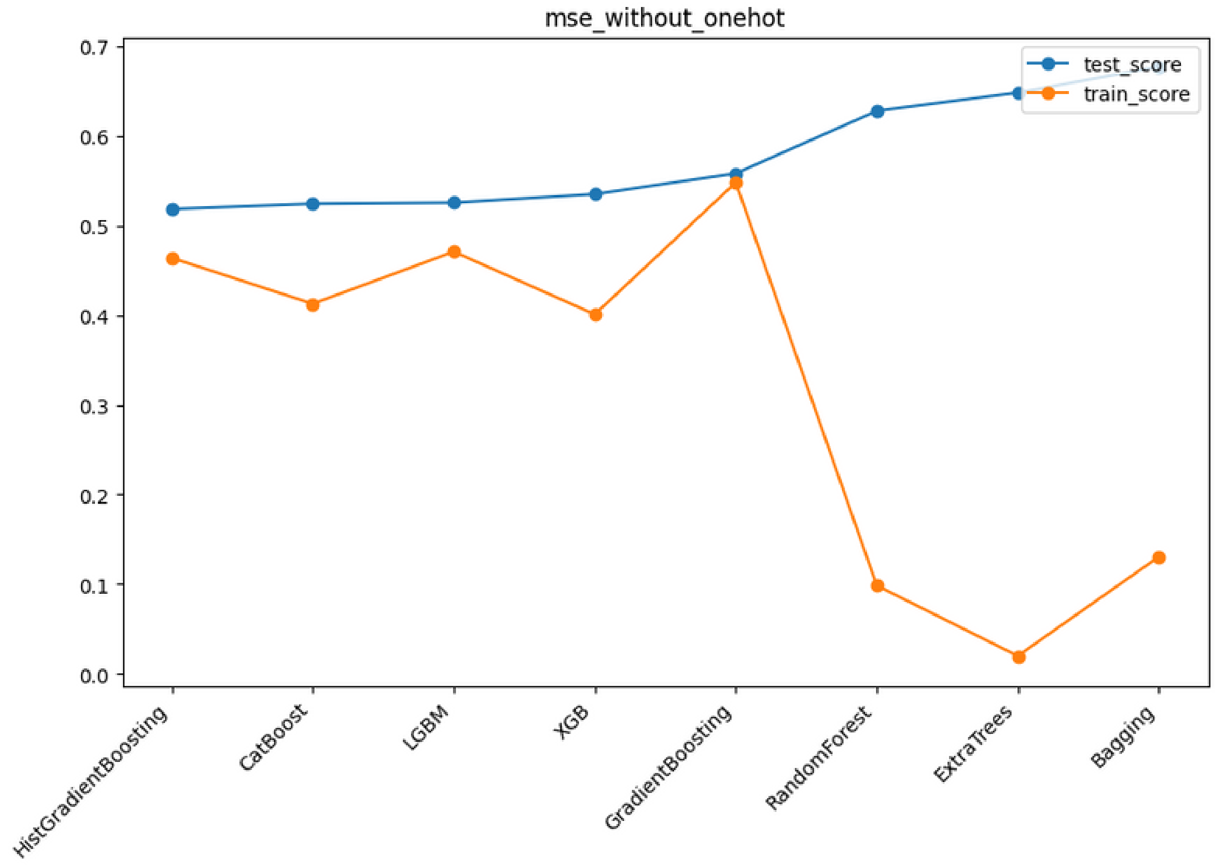
TRAIN VS TEST

Experiments Ensemble Models

Group 12







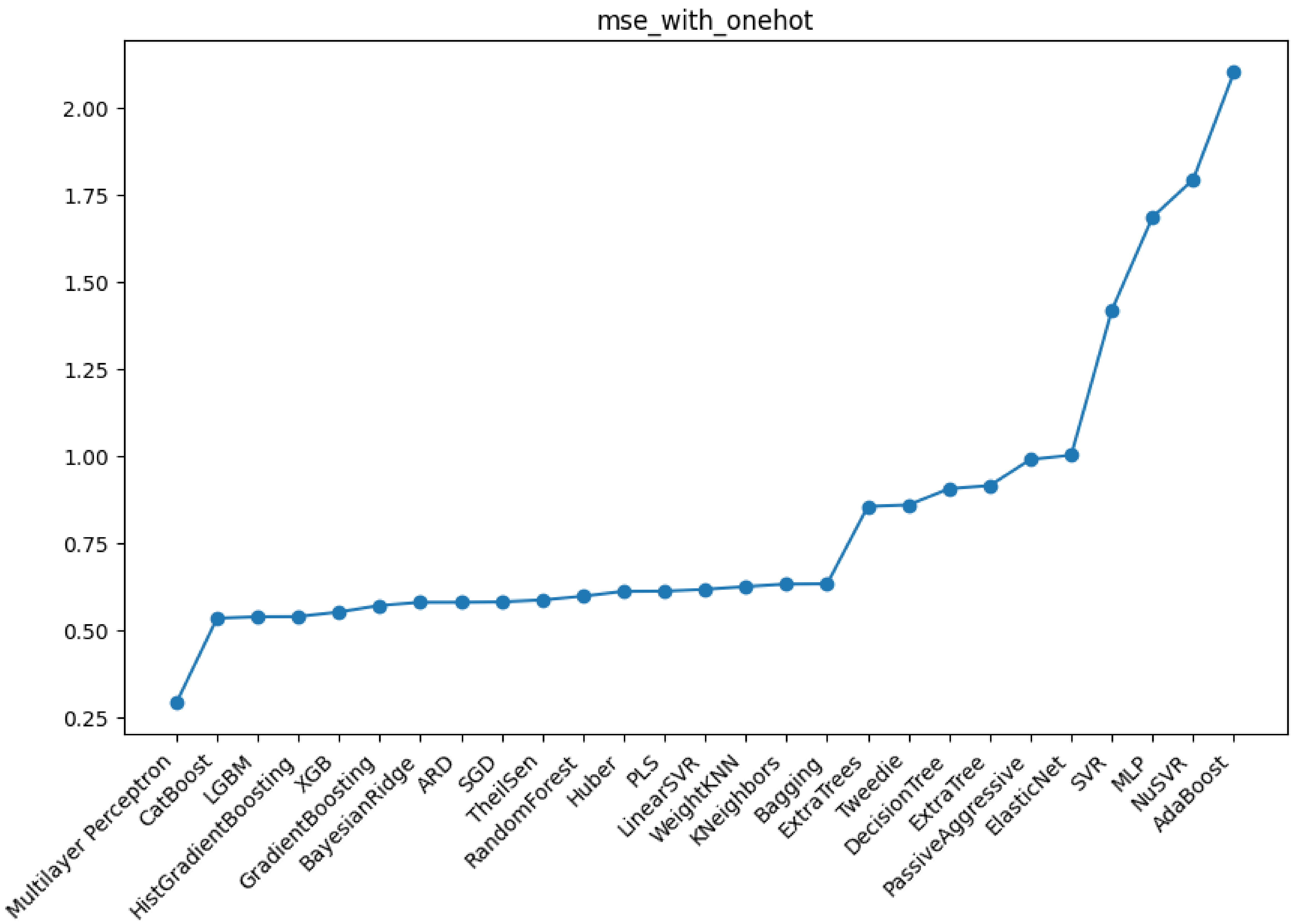
ONE-HOT OR NOT

TRAIN VS TEST

TRAIN VS TEST

Conclusion MSE Overall

Group 12



data science final

Group 12

THANK YOU

PREDICTING INDIVIDUAL INCOME THROUGH HOUSEHOLD DEMOGRAPHIC STATISTICS: A DATA-DRIVEN APPROACH

110502528 HSUN-HAO CHANG
110502529 PO-SHEN CHEN
110502534 CHUN-YU CHEN
110502009 HUNG-YI HSU

End Slide