
NIB
FINANCING
THE
FUTURE
INTERNSHIP
PRESENTATION
PHAT NGUYEN

AGENDA

1/ OVERVIEW & DATA PREP

2/ EXPLORATORY ANALYSIS

3/ MODEL BUILDINGS & EVALUATION

4/ CONCLUSIONS & RECOMMENDATIONS

01 OVERVIEW

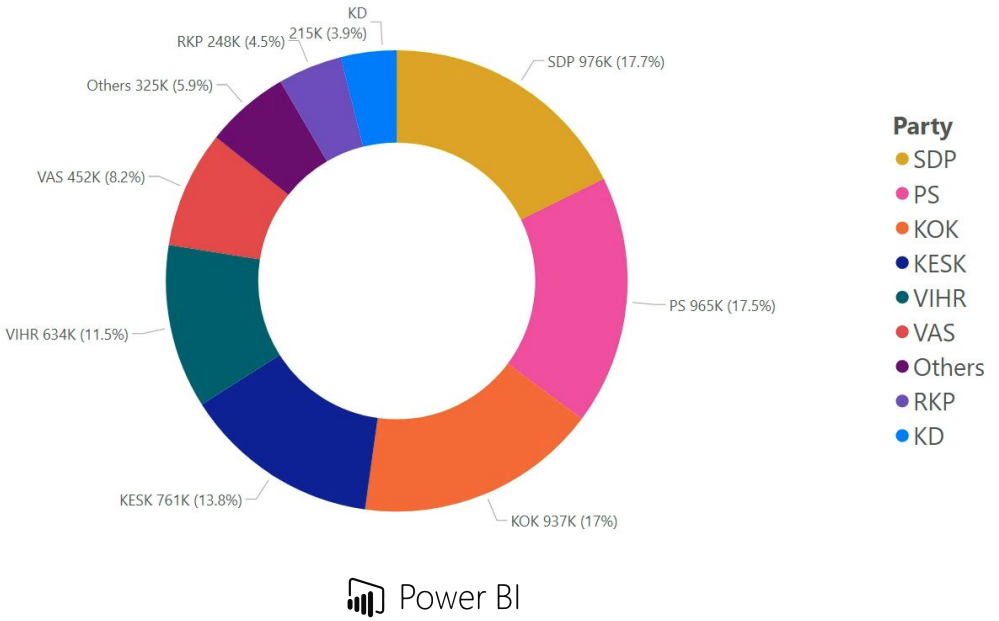
2 tables were provided processed and merged into 1 table via Excel, ***removing*** regional information and ***keeping*** municipality information.

- **Table 1 (403 cols x 8 rows)** population characteristics of 400 municipalities and regions of Finland including
  - Population
  - Degree of urbanization
  - % of foreign citizens
  - % of household living in detached/terrace homes
  - % of household living in rental homes
  - Employment rate
  - Unemployment rate
  - Workplace self-sufficiency

- **Table 2 (10 cols x 313 rows)** voting preference for 9 political parties (KESK, PS, KOK, SDP, VIHR, VAS, RKP, KD, Others) of 311 municipalities and the whole Finland.

02 EXPLORATORY ANALYSIS

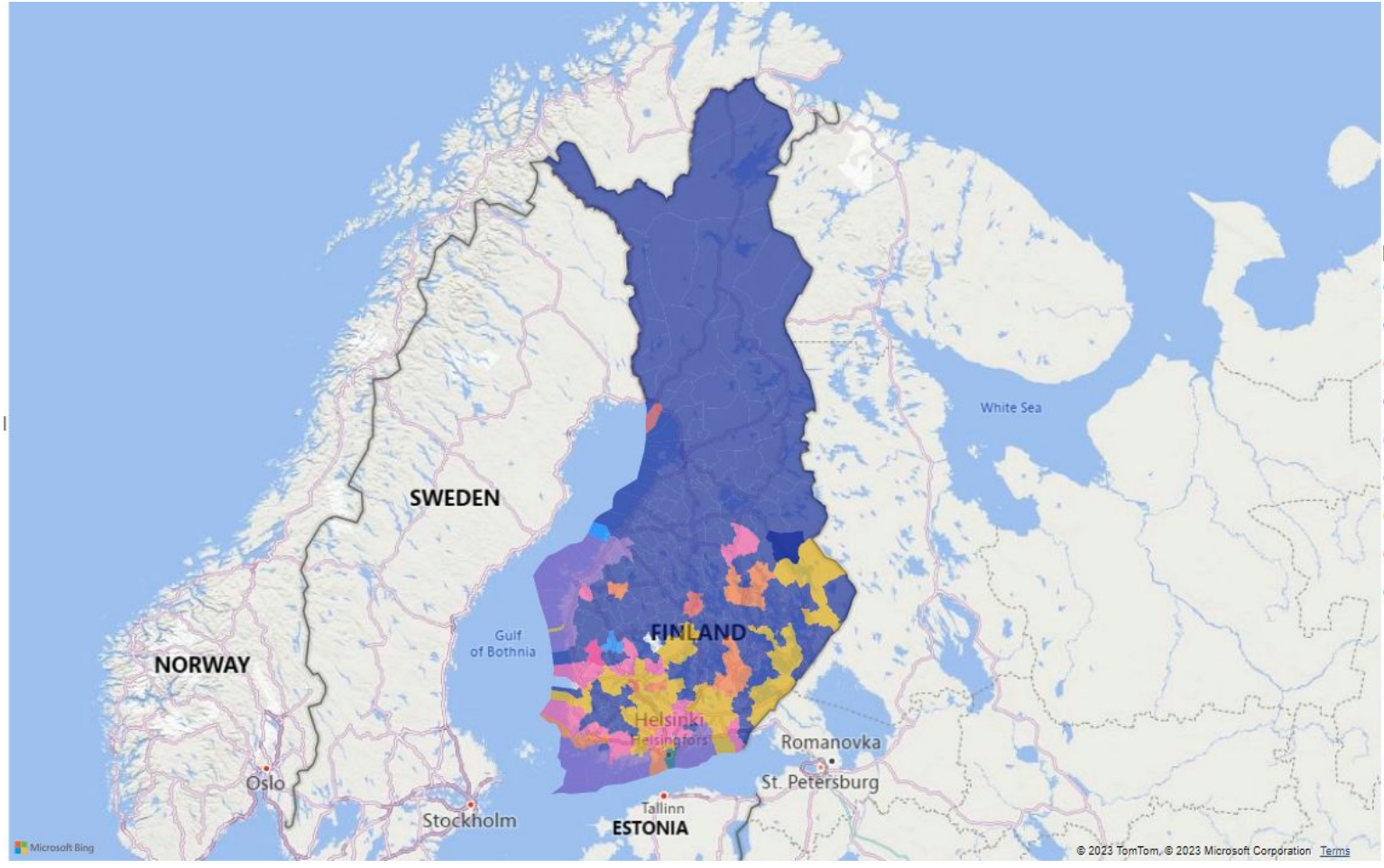
Political parties in Finland



Power BI

Municipalities by Dominating Party

Power BI

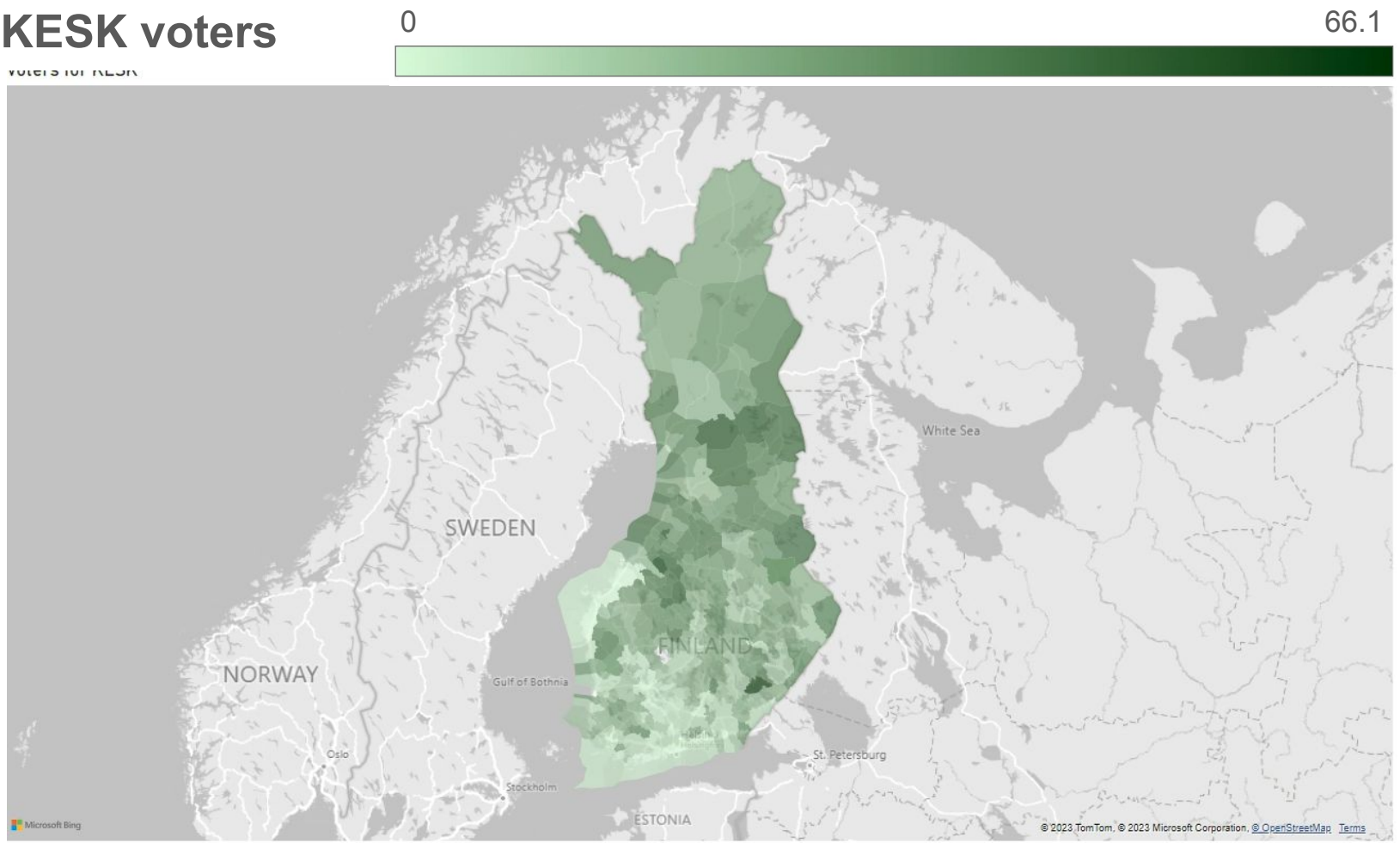


Dominating Party

- KD
- KESK
- KOK
- Others
- PS
- RKP
- SDP
- VAS
- VIHR

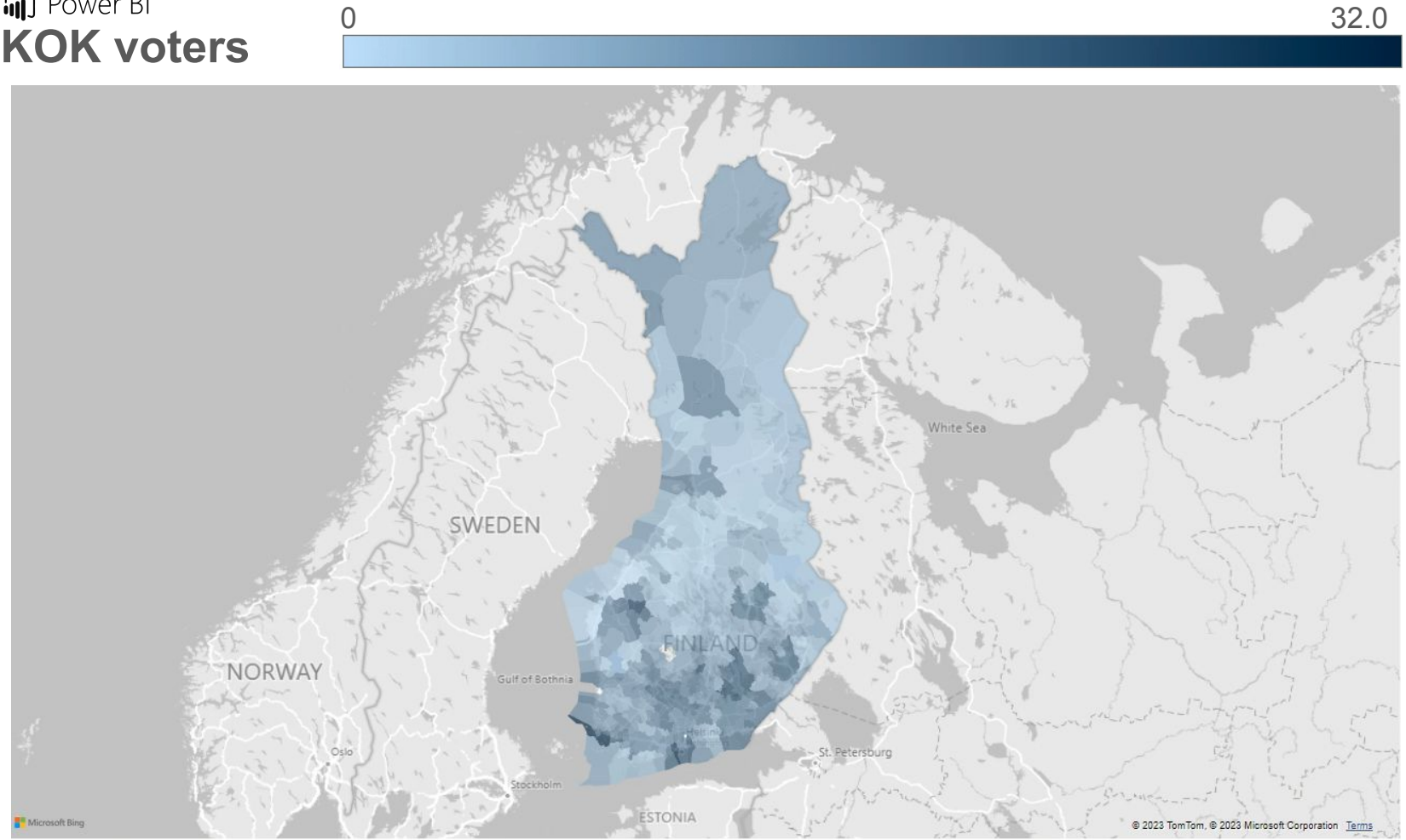
Power BI

KESK voters



Power BI

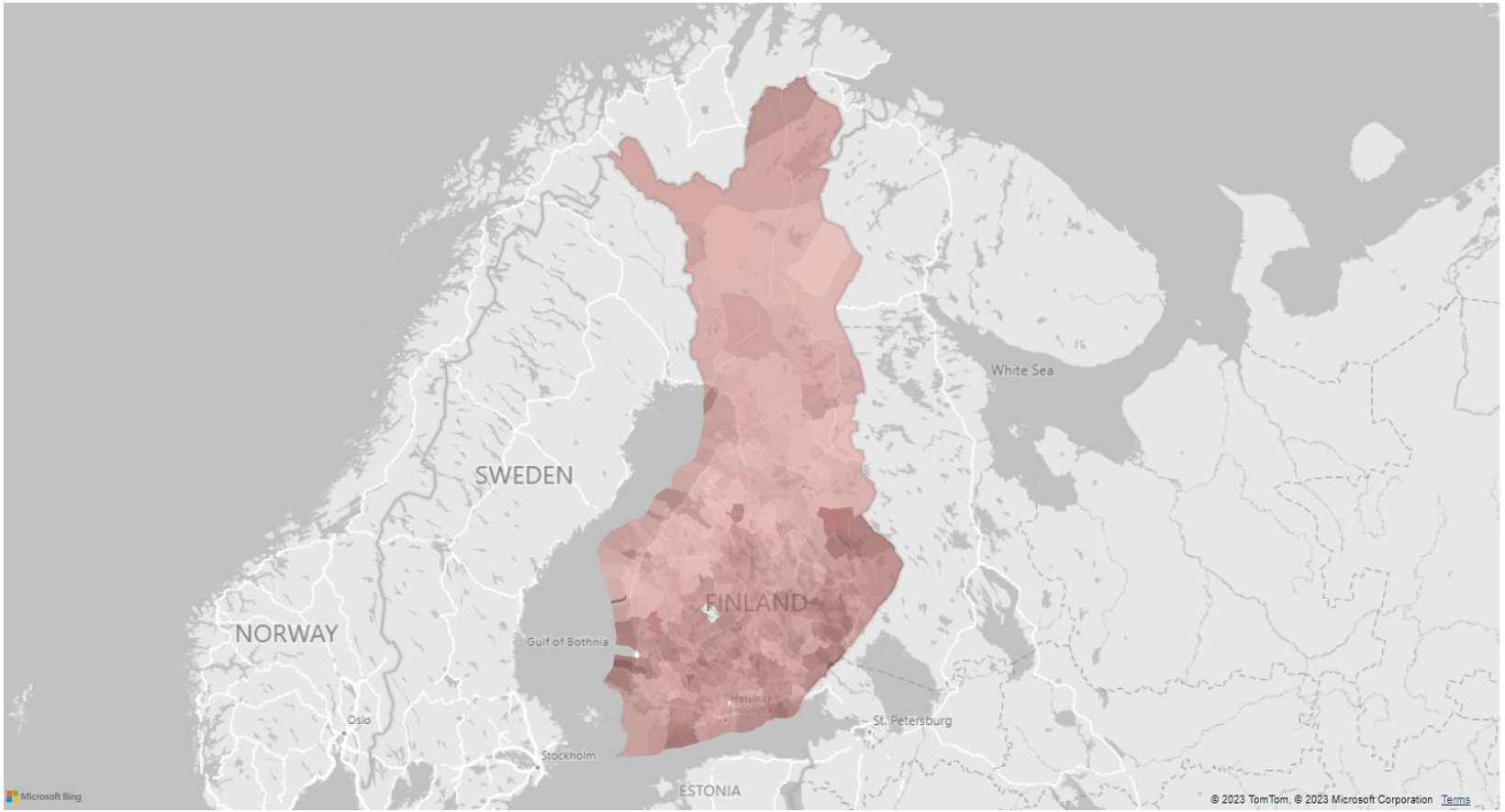
KOK voters



Power BI

SDP voters

0 40.8



Power BI

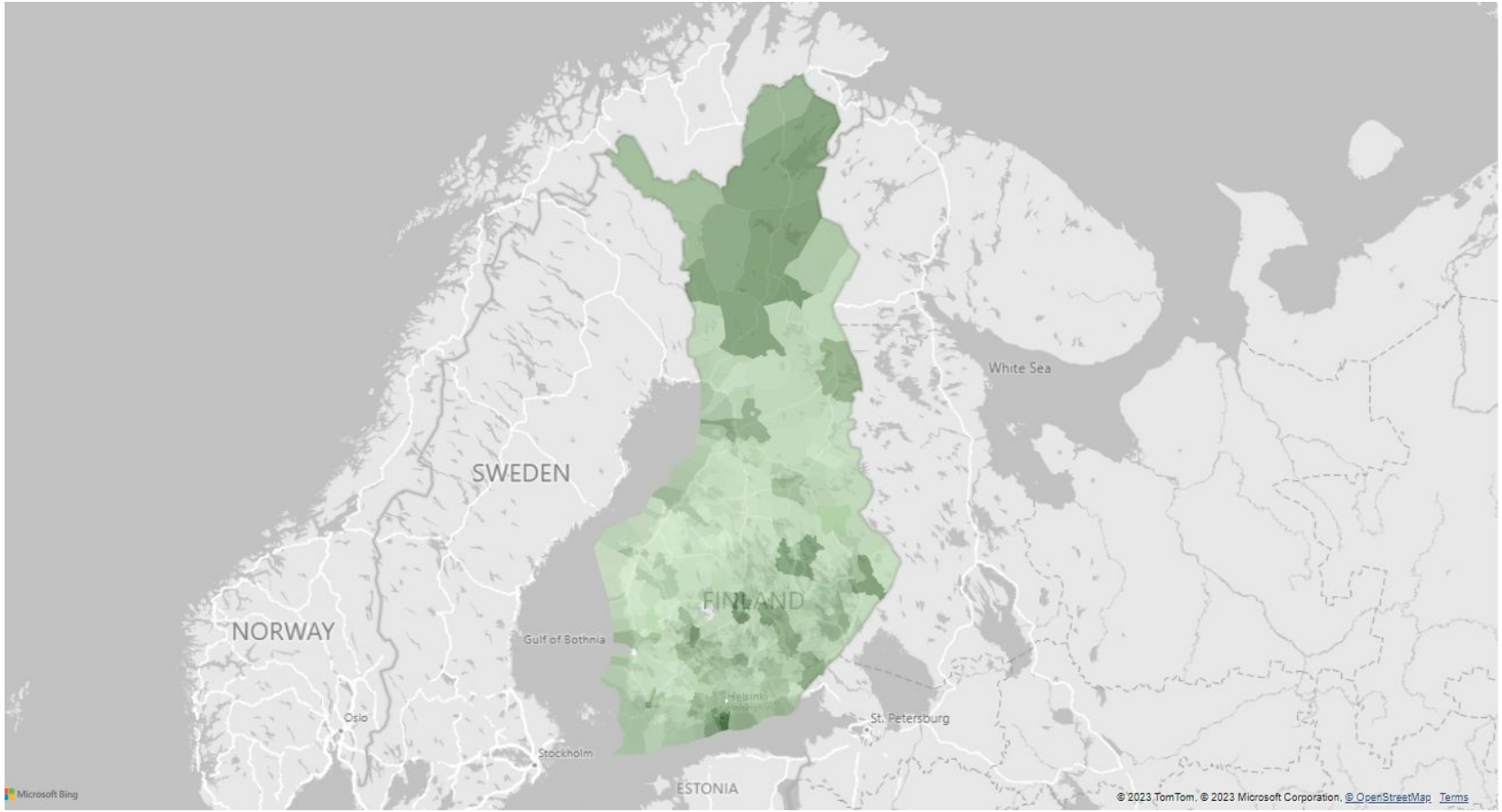
PS voters

0 42.7

Power BI

VIHR voters

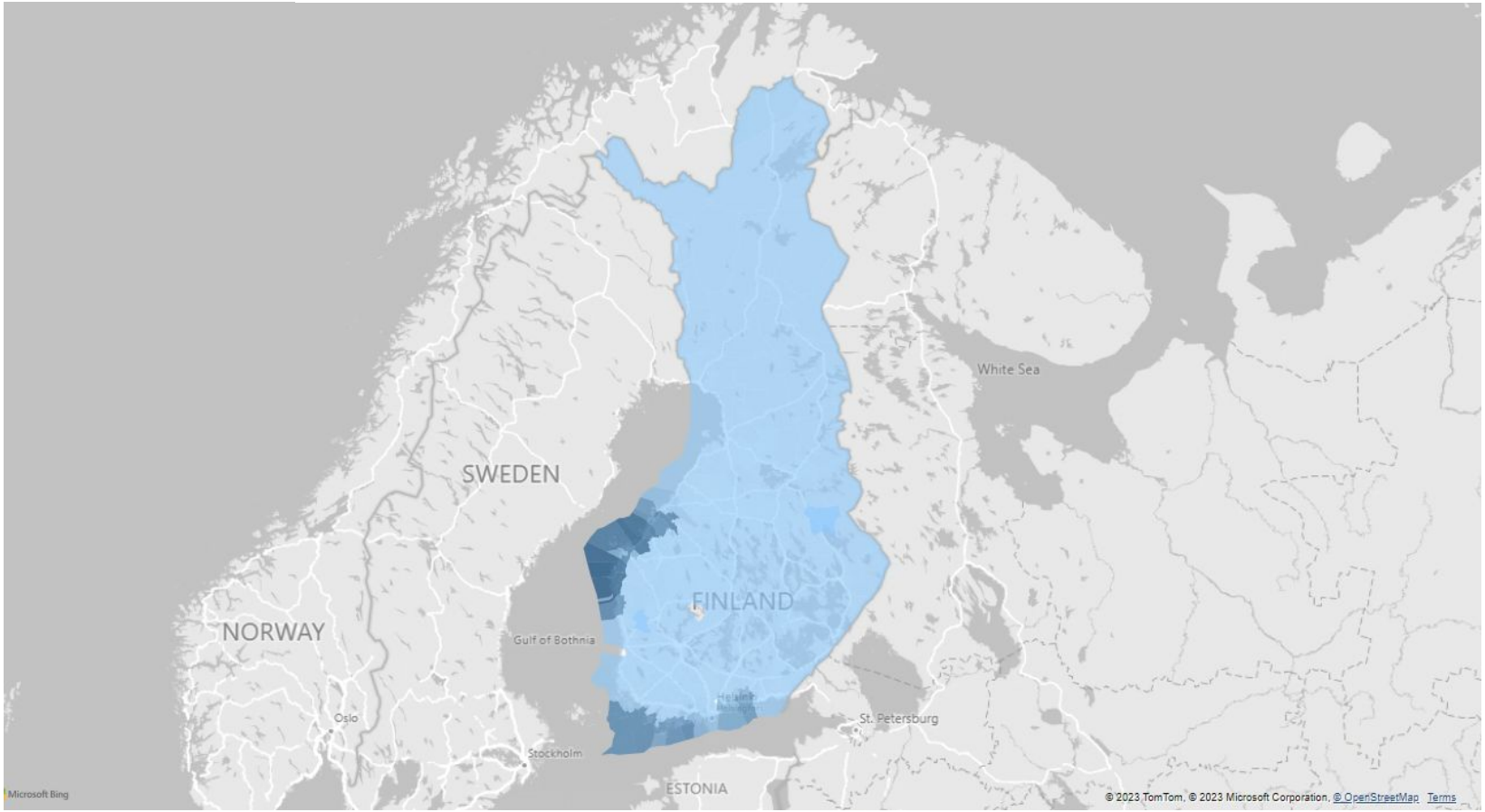
0 — 23.5



Power BI

RKP voters

0 92.6



Power BI

VAS voters

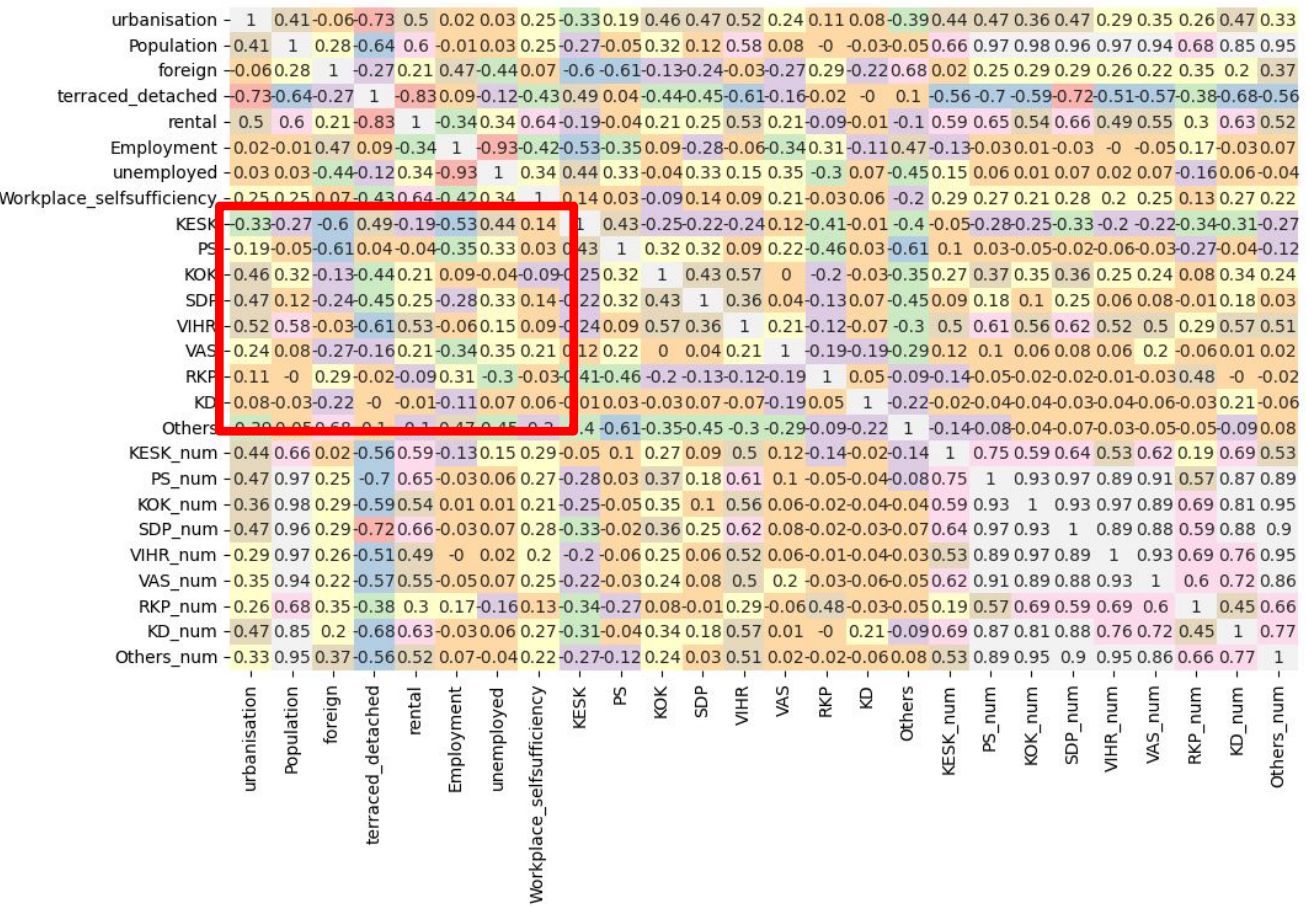
0 28.3

Power BI

KD voters

0 48.0

Correlation Heatmap



Correlation Method: Pearson (linear)

KESK x foreign (-0.6)
KESK x employment (-0.53)
KESK x terrace (+0.49)
KESK x unemployment (+0.43)

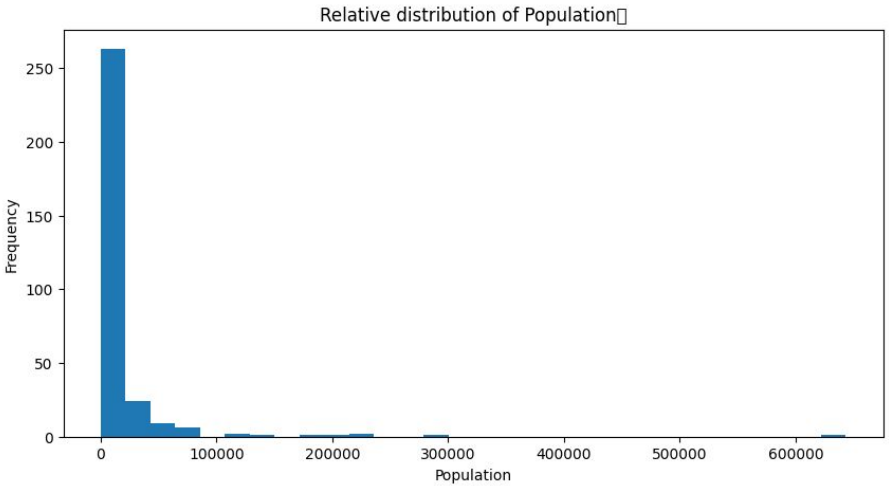
PS x foreign (-0.61)

SDP x urbanization (+0.47)
SDP x terrace (-0.45)

VIHR x urbanization (+0.52)
VIHR x population (+0.58)
VIHR x rental (+0.53)
VIHR x terrace (-0.61)

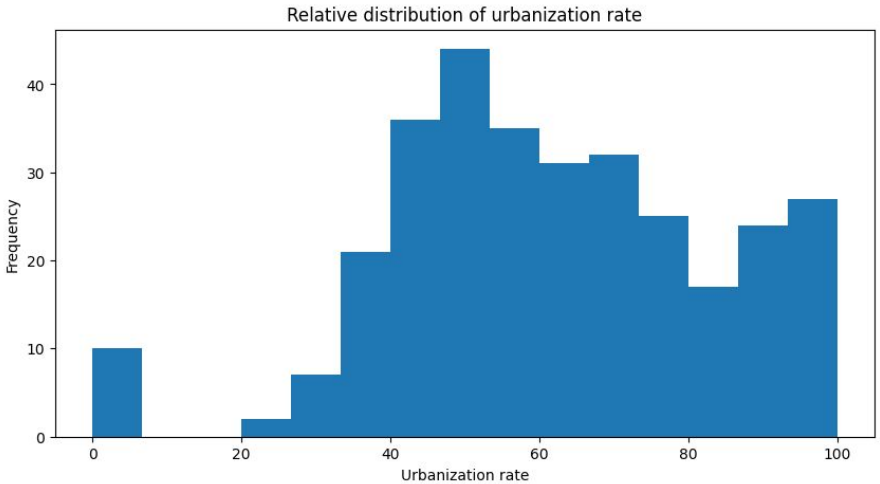
Nothing major for VAS, RKP, KD, Others.

Distribution of Population



Vast majority of municipalities have <100k population

Distribution of urbanization rate



Overall, Finland is pretty urbanized, with the majority of municipalities having urbanization rate of >40%.

python™

3A LINEAR REGRESSION

- We try to build a Linear Regression Model
- We split data into **train (70%) / test (30%)**
- We calculate R^2 score and Mean Absolute Error, with the results below.

python™

	KESK	PS	KOK	SDP	VIHR	VAS	RKP	KD	Others
R^2	0.66	0.32	0.24	0.4	0.64	0.1	0.14	-0.33	0.08
MAE	7.09	4.14	4.71	4.75	1.89	4.32	8.49	2.42	8.61

3A LINEAR REGRESSION

- We try to build another Linear Regression Model
- This time, we use **cross-validation (5 folds)** instead of train/test split
- We calculate R^2 score and Mean Absolute Error, with the results below.

python™

	KESK	PS	KOK	SDP	VIHR	VAS	RKP	KD	Others
R^2	0.67	0.34	0.37	0.41	0.53	0.09	-0.5	0.05	0.5
MAE	6.69	4.06	4.17	4.57	1.85	3.95	7.97	2.63	8.77

Train/Test	KESK	PS	KOK	SDP	VIHR	VAS	RKP	KD	Others
R^2	0.66	0.32	0.24	0.4	0.64	0.1	0.14	-0.33	0.08
MAE	7.09	4.14	4.71	4.75	1.89	4.32	8.49	2.42	8.61

Cross-Validation	KESK	PS	KOK	SDP	VIHR	VAS	RKP	KD	Others
R^2	0.67	0.34	0.37	0.41	0.53	0.09	-0.5	0.05	0.5
MAE	6.69	4.06	4.17	4.57	1.85	3.95	7.97	2.63	8.77

python™

3A LINEAR REGRESSION

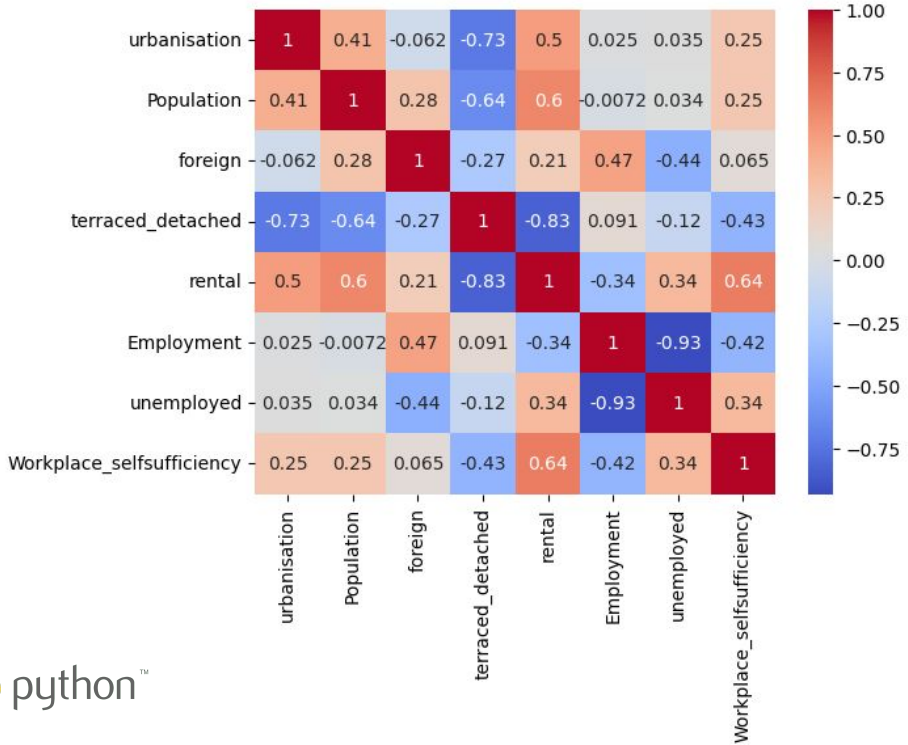
python™

Check for Multicollinearity by Variance Inflation Factor (VIF)

	feature	VIF
0	urbanisation	28.489891
1	Population	2.069472
2	foreign	4.667190
3	terraced_detached	174.643414
4	rental	55.773647
5	Employment	283.022299
6	unemployed	25.732745
7	Workplace_selfsufficiency	38.711634

VIF measures how much the variance of a regression coefficient is inflated due to collinearity □ Features in red are highly correlated with each other and are redundant to the model.

3A LINEAR REGRESSION



Correlation Method: Pearson (linear)

Employed x unemployed (-0.93)

Urbanization x terraced (-0.73)

Population x terrace (-0.64)

Rental x terrace (-0.83)

Rental x workplace_selfsufficiency (0.64)

python™

3B RANDOM FOREST REGRESSOR

- We try to build a Random Forest Regressor (n_estimator=100)
- We split data into **train (70%) / test (30%)**
- We calculate R^2 score and Mean Absolute Error, with the results below.

python™

	KESK	PS	KOK	SDP	VIHR	VAS	RKP	KD	Others
R^2	0.68	0.27	0.46	0.46	0.63	0.01	0.22	-0.22	0.51
MAE	6.63	3.96	3.82	4.61	1.89	4.22	6.72	2.2	3.44

3B RANDOM FOREST REGRESSOR

- We try to build a Random Forest Regressor (n_estimator=100)
- This time, we use **cross-validation (5 folds)** instead of train/test split
- We calculate R^2 score and Mean Absolute Error, with the results below.

python™

	KESK	PS	KOK	SDP	VIHR	VAS	RKP	KD	Others
R^2	0.68	0.46	0.46	0.42	0.54	0.05	0.13	-0.06	0.76
MAE	6.32	3.69	3.59	4.45	1.7	3.9	5.2	2.66	3.98

3B RANDOM FOREST REGRESSOR

- We perform hypertuning for some parameters of the DFG using GridSearch
- Parameter grid

```
param_grid = {'n_estimators': [50, 100, 200],
              'max_depth': [None, 5, 10,20],
              'min_samples_split': [2, 5, 10]}
```

- We calculate R^2 score and Mean Absolute Error, with the results below.

python™

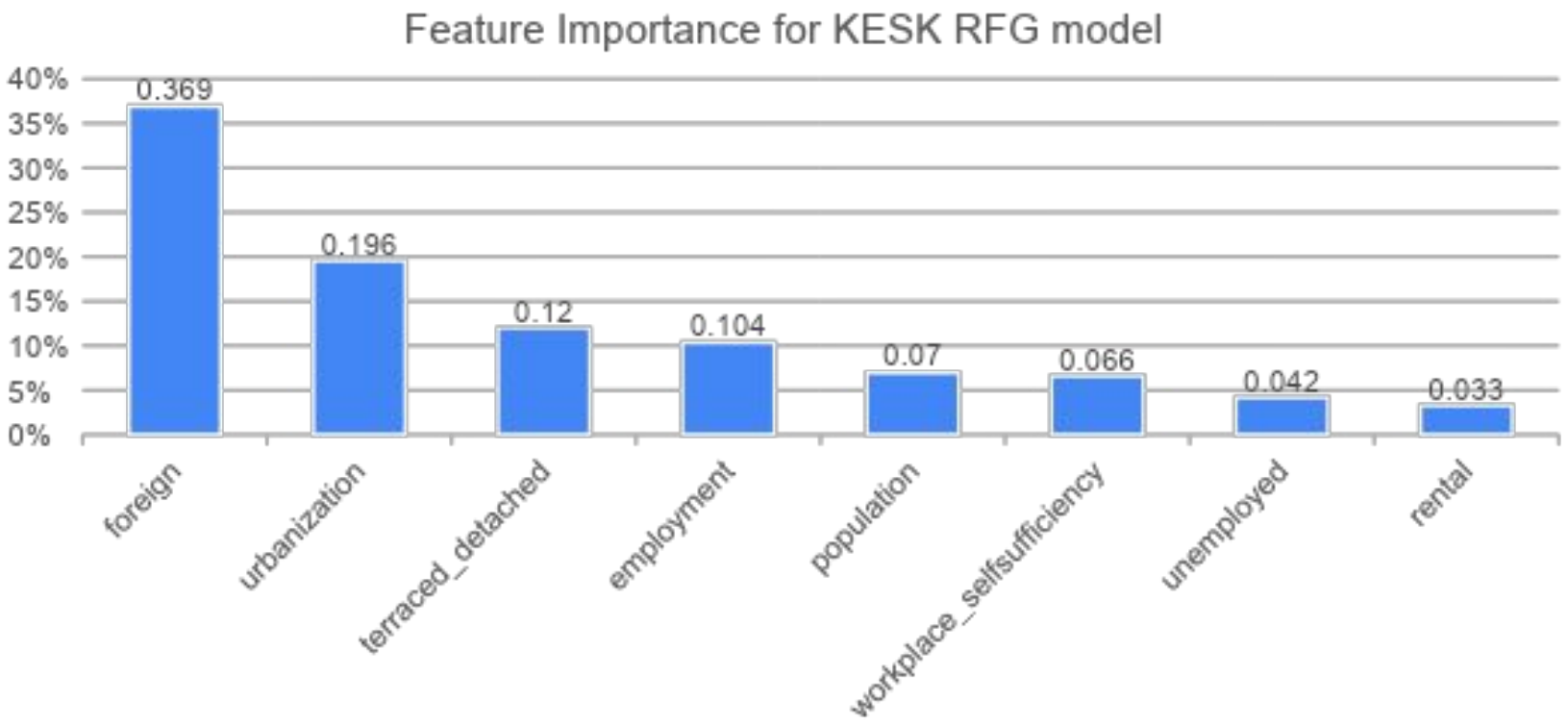
	KESK	PS	KOK	SDP	VIHR	VAS	RKP	KD	Others
R^2	0.68	0.47	0.47	0.43	0.53	0.08	0.13	0.01	0.78
MAE	6.37	3.69	3.56	4.46	1.67	3.97	5.19	2.69	3.89

	KESK	PS	KOK	SDP	VIHR	VAS	RKP	KD	Others
R^2	0.68	0.27	0.46	0.46	0.63	0.01	0.22	-0.22	0.51
MAE	6.63	3.96	3.82	4.61	1.89	4.22	6.72	2.2	3.44

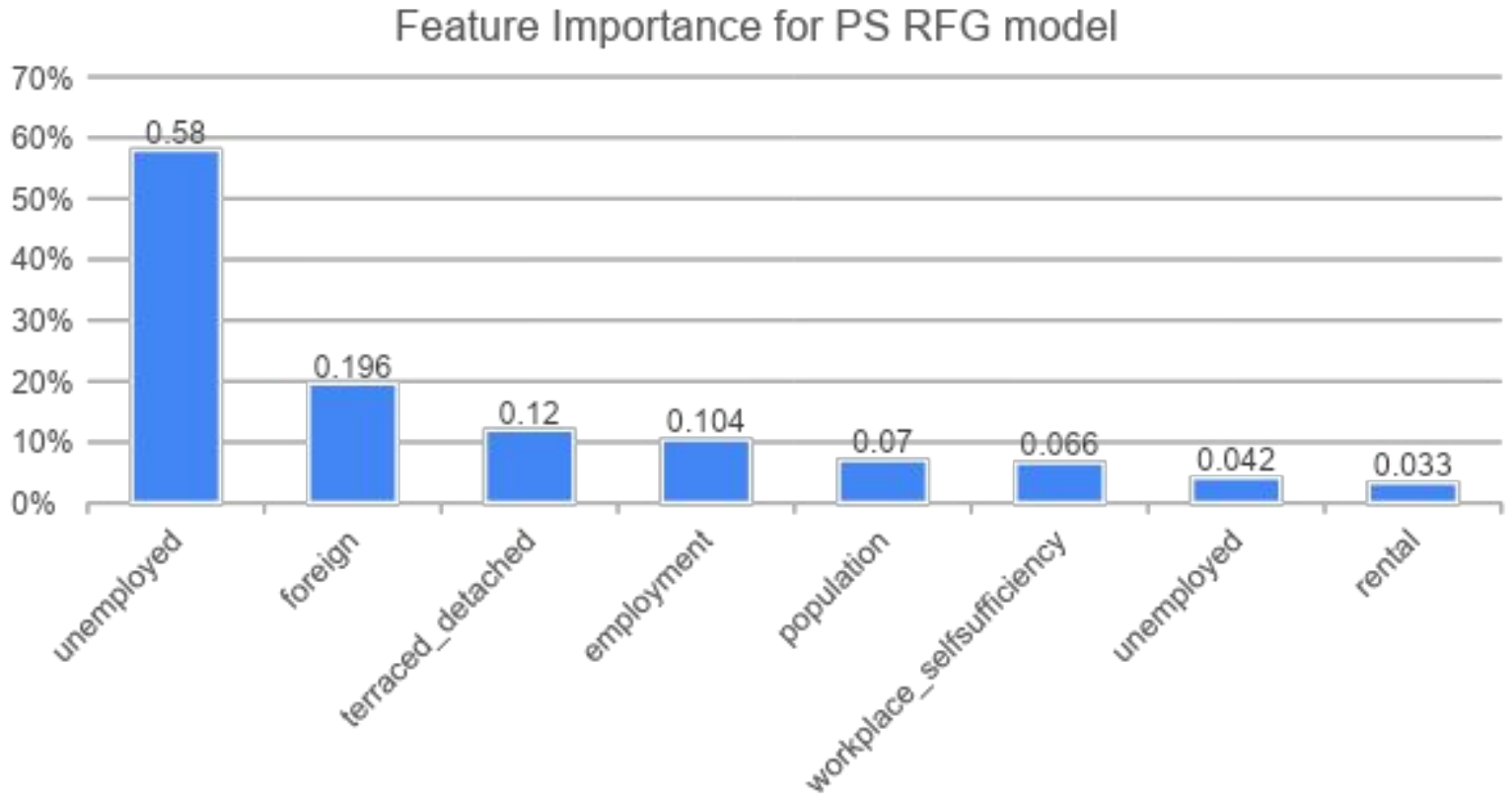
	KESK	PS	KOK	SDP	VIHR	VAS	RKP	KD	Others
R^2	0.68	0.46	0.46	0.42	0.54	0.05	0.13	-0.06	0.76
MAE	6.32	3.69	3.59	4.45	1.7	3.9	5.2	2.66	3.98

	KESK	PS	KOK	SDP	VIHR	VAS	RKP	KD	Others
R^2	0.68	0.47	0.47	0.43	0.53	0.08	0.13	0.01	0.78
MAE	6.37	3.69	3.56	4.46	1.67	3.97	5.19	2.69	3.89

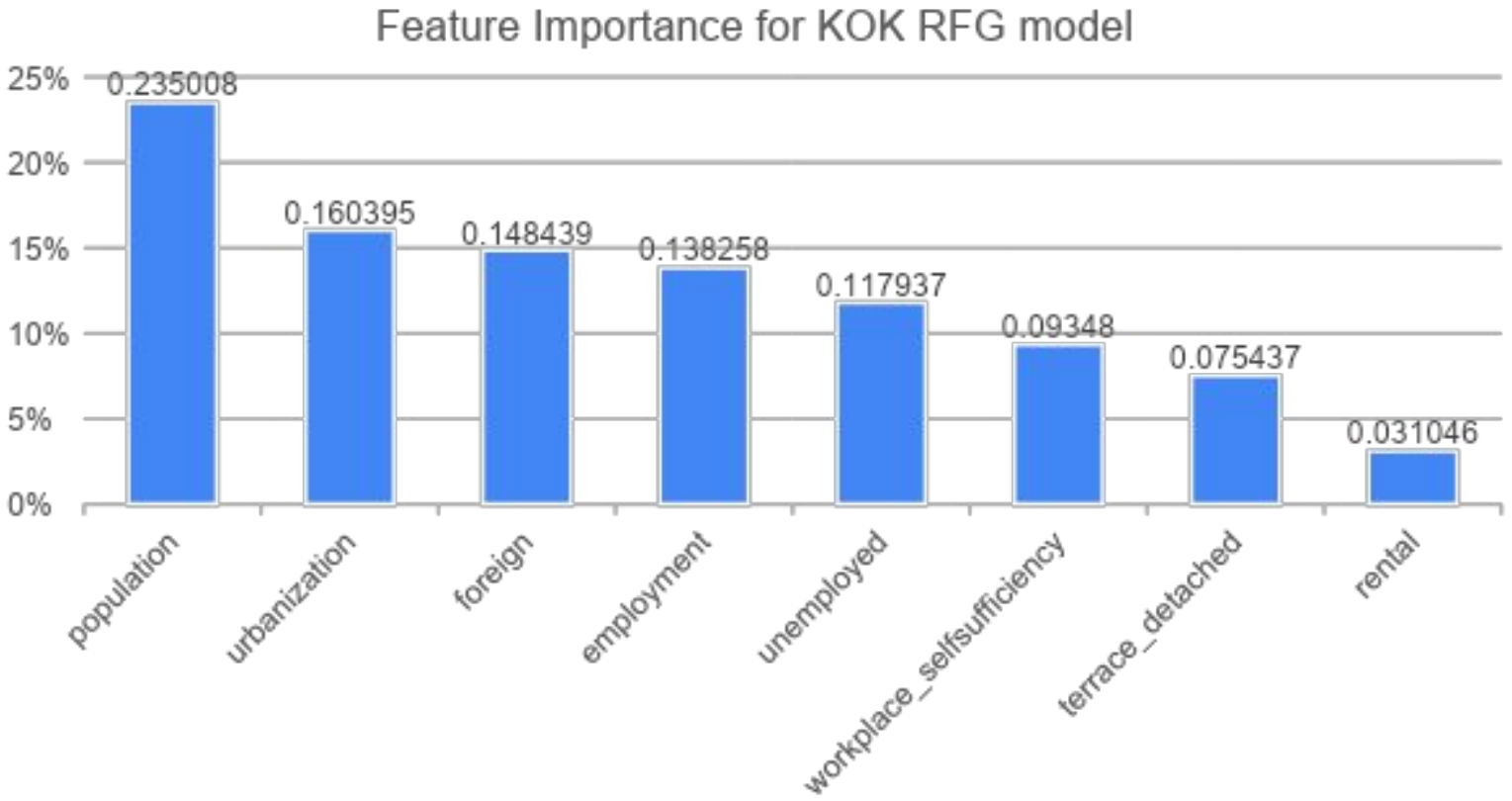
04 CONCLUSION



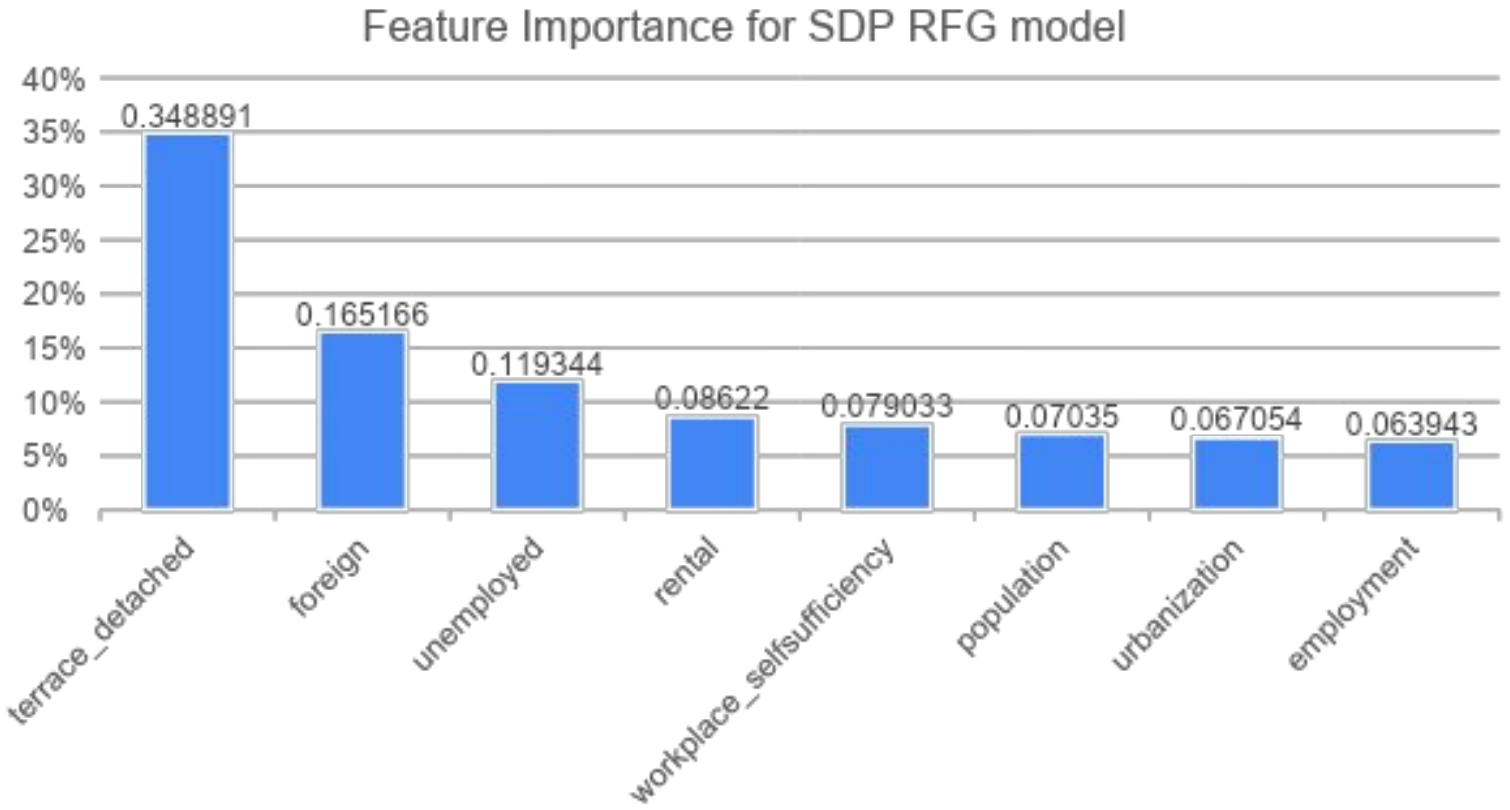
04 CONCLUSION



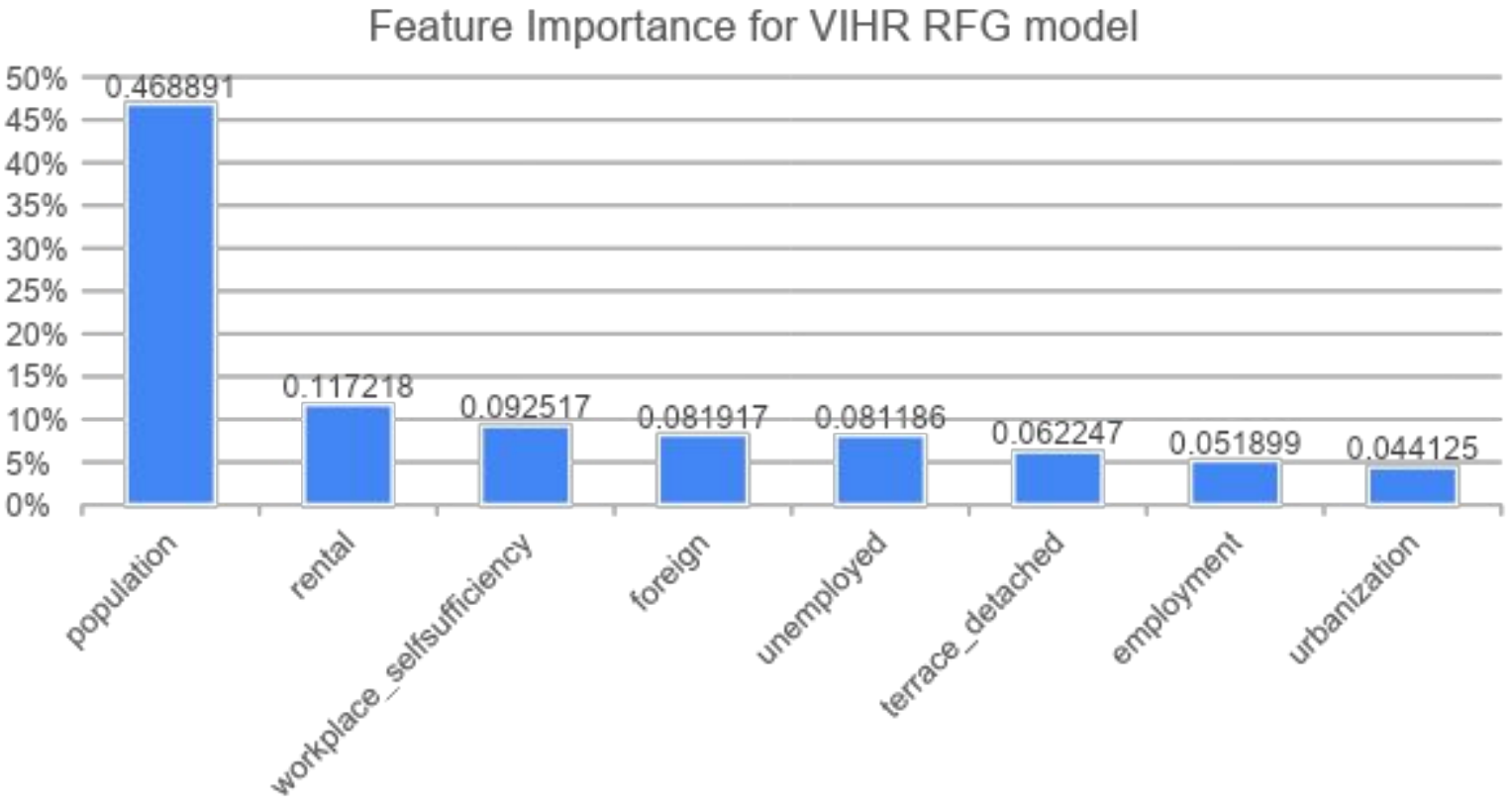
04 CONCLUSION



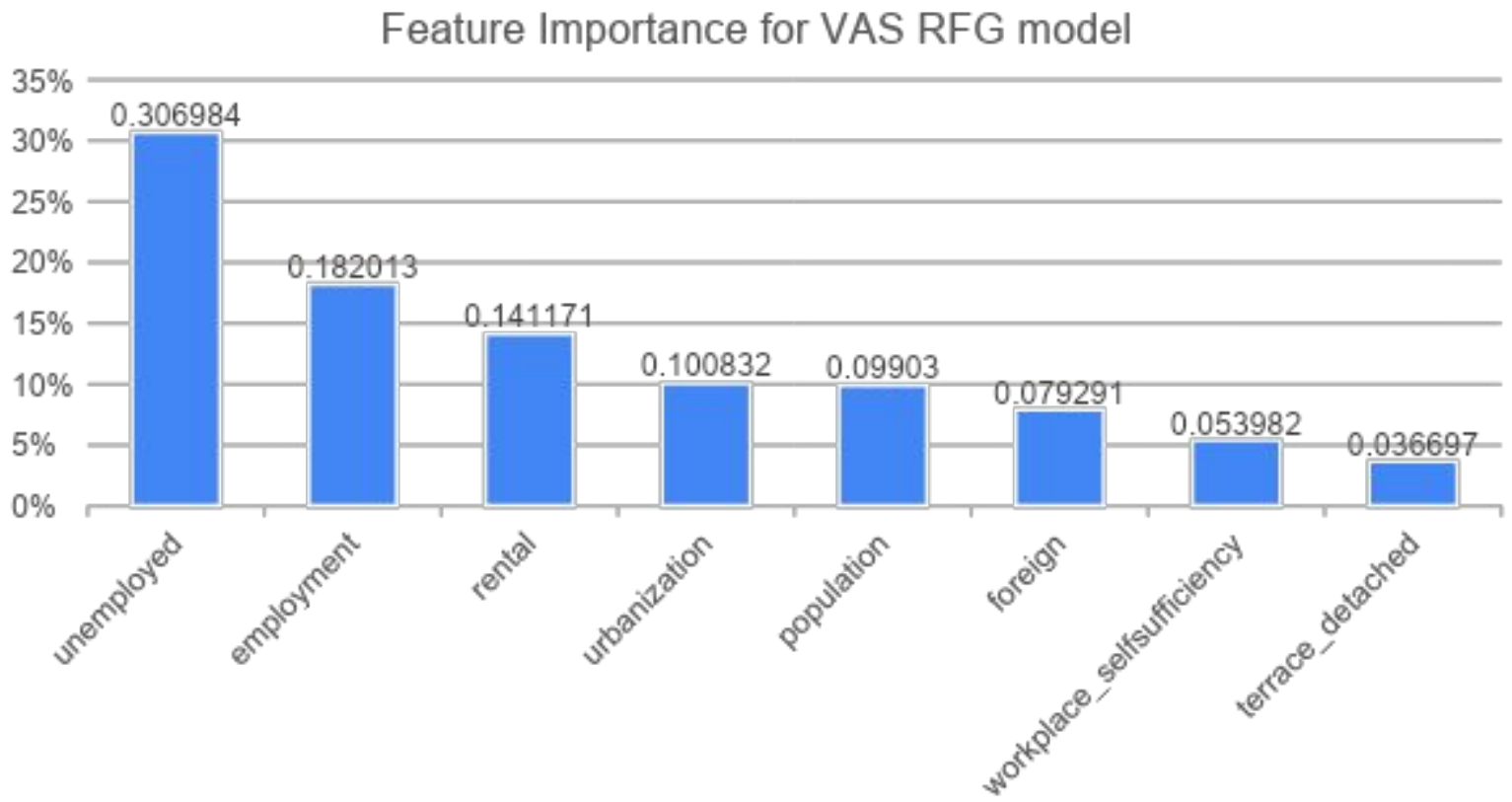
04 CONCLUSION



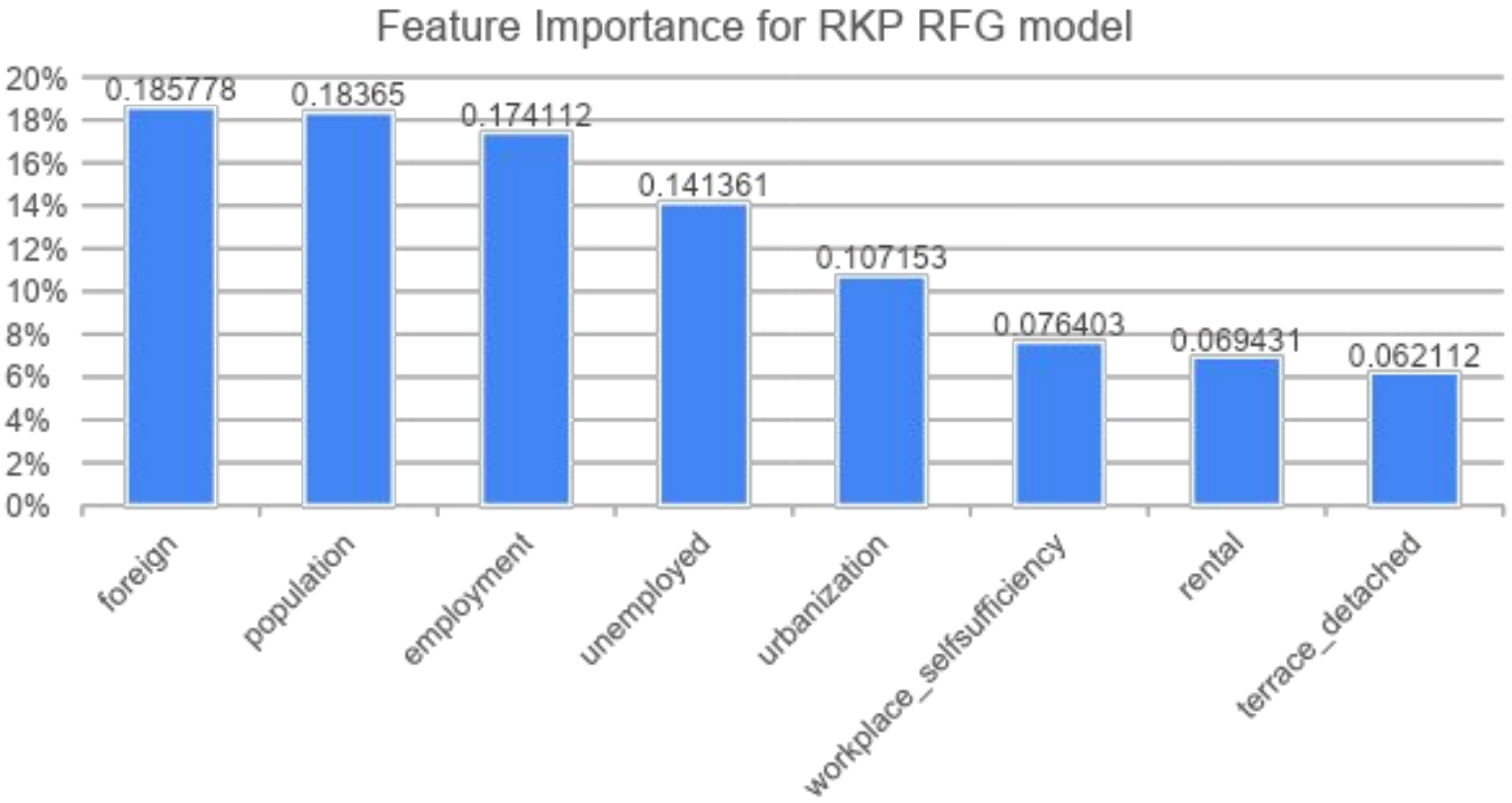
04 CONCLUSION



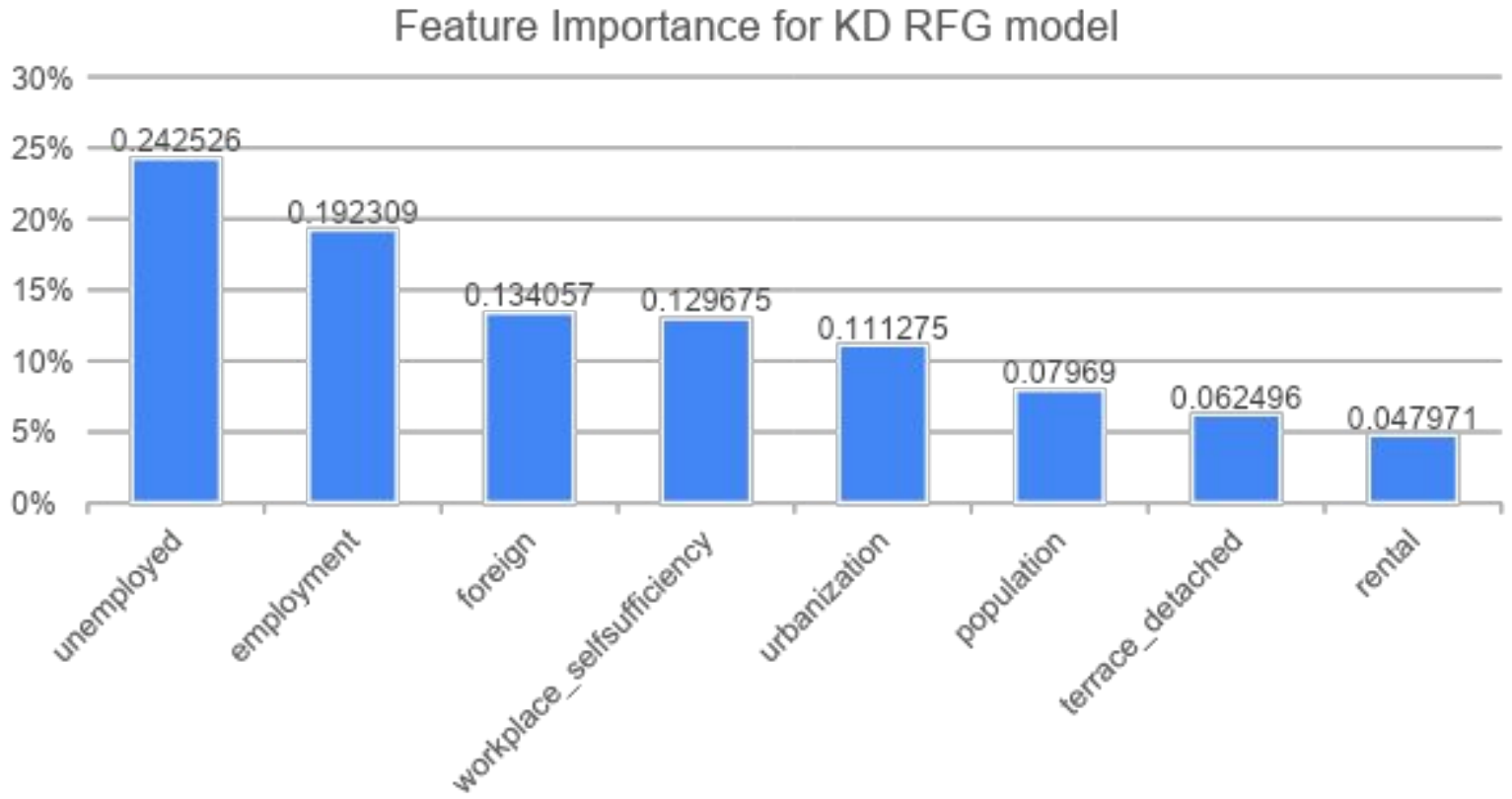
04 CONCLUSION



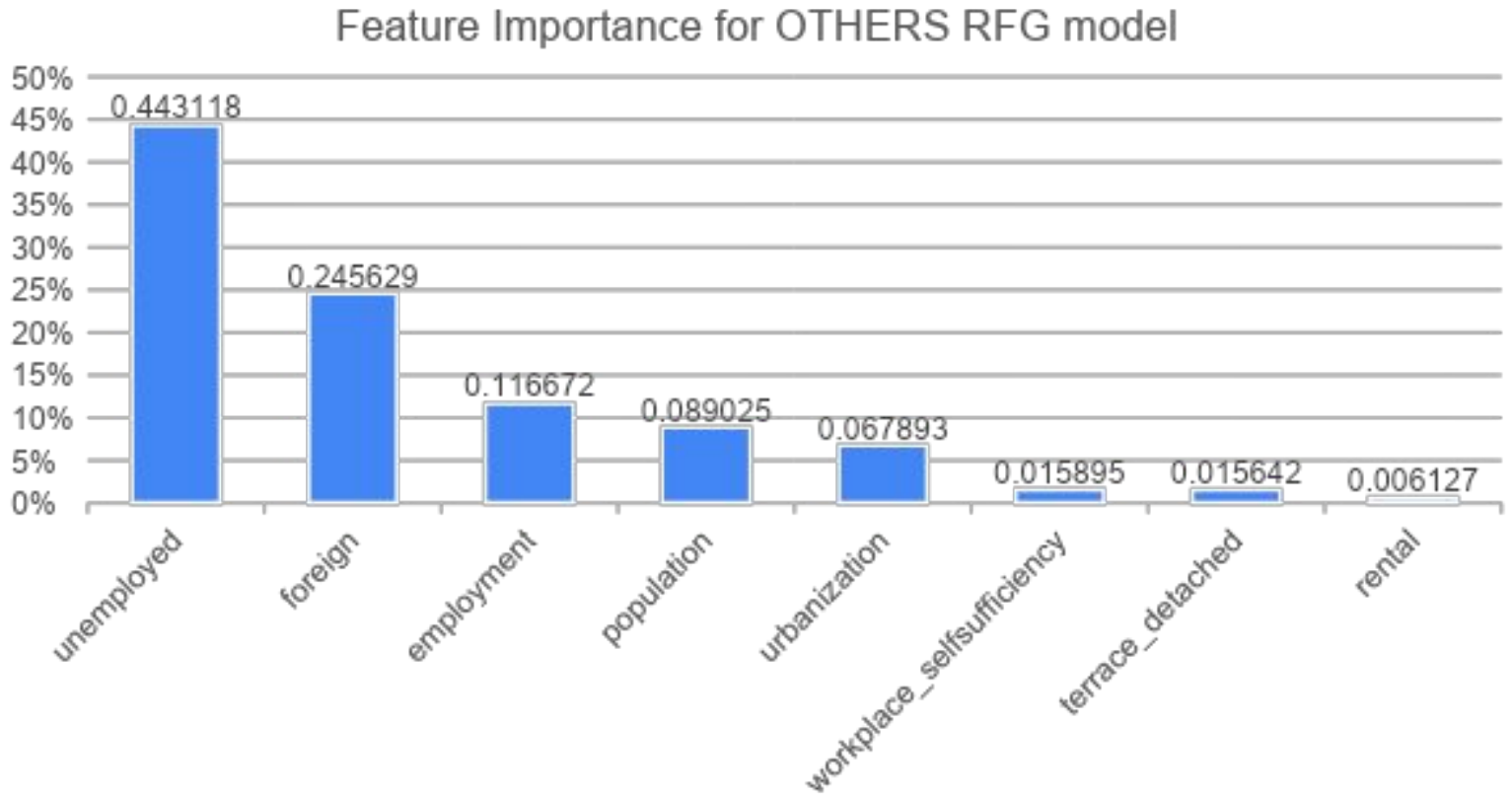
04 CONCLUSION



04 CONCLUSION



04 CONCLUSION



04 RECOMMENDATIONS

- The model can explain quite well **KESK**, **SDP**, **KOK**, **VIHR**, **PS**, and **Others**.
- Prediction for the remaining parties are not so well (**VAS**, **KD**, **RKP**).
- All factors seem to be contributing to the predictive models, with various degree of importance. Most important features overall include ***foreign*** and ***unemployment***.
- The model might be improved by incorporating more variables such as income, racial profile of the population, religion and degree of religiosity, social and economic class, educational level, regional characteristics, and gender profile, etc.


THANK
YOU