ATTRITION ANALYSIS

Data Analysis Presentation
Tianchu Shu, OAE Intern
August 2018

CONTENT

- Background
- Project Goals
- Padding Model
- Candidate Attrition Model
- Recommendations

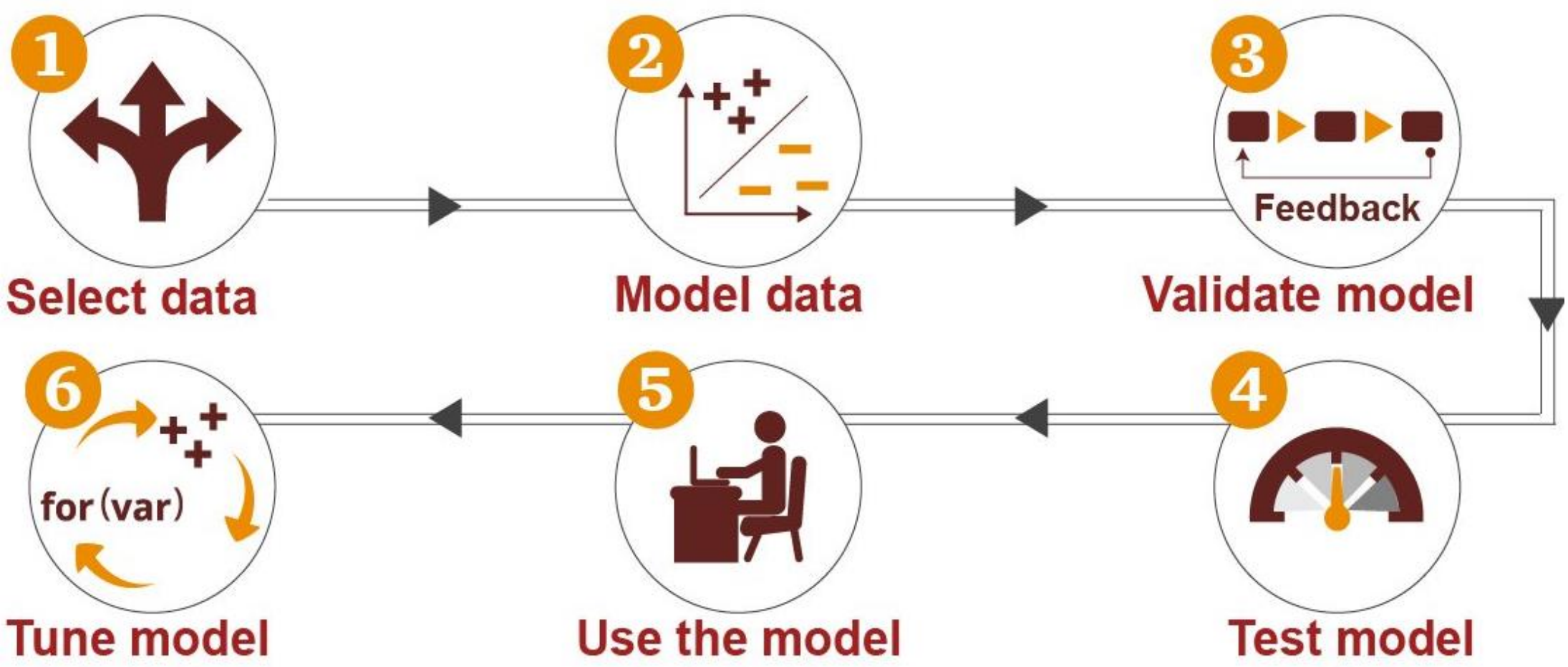
BACKGROUND

Peace Corps

PROJECT GOALS

Peace Corps

METHODOLOGY



1. PADDING MODEL

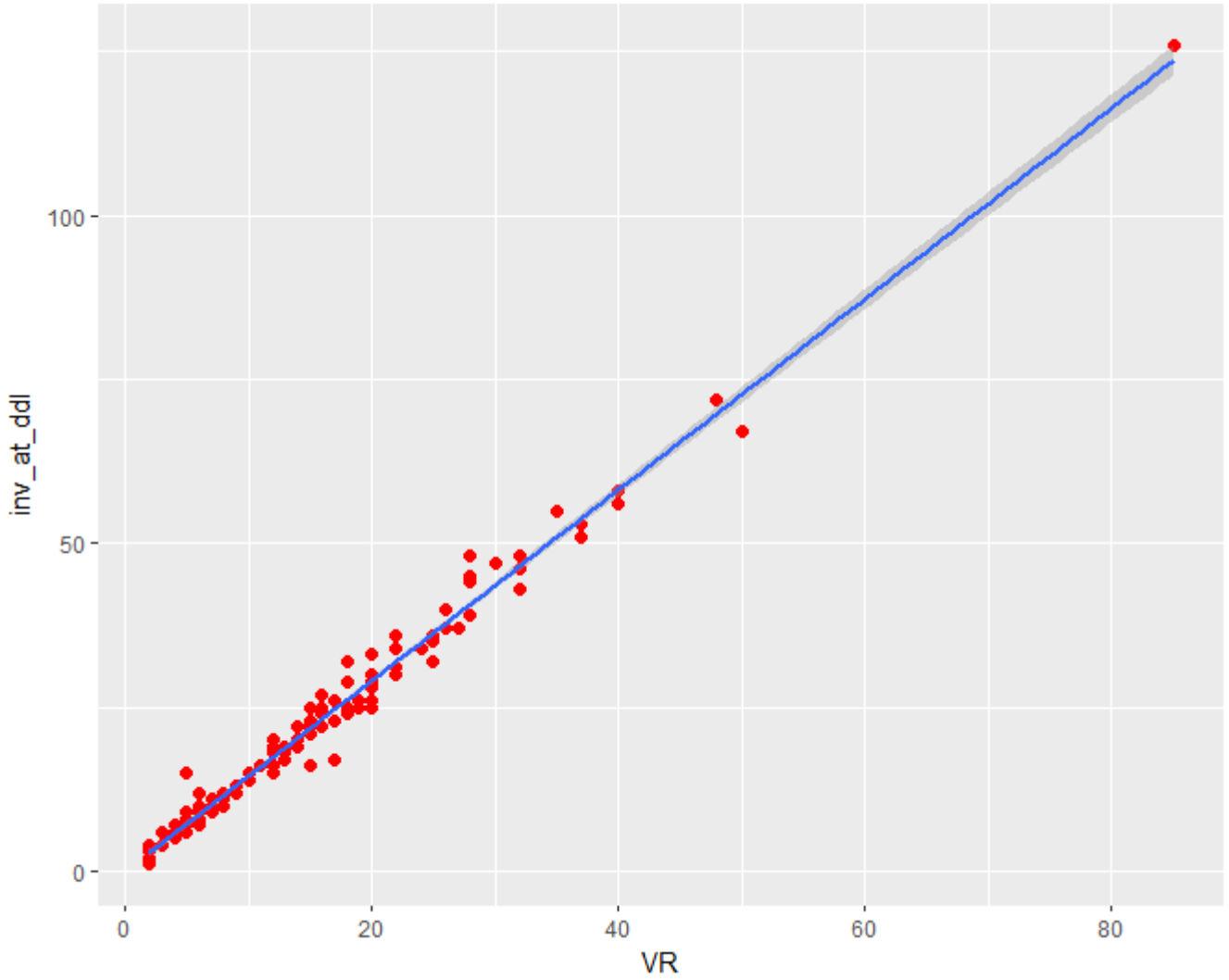
Peace Corps

DATA

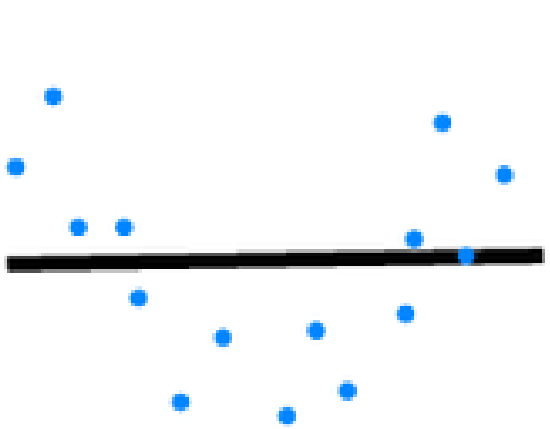
530 observations in total

Variable Name	Interpret
inv_at_ddl	The number of people invited at deadline.
VR	The number of volunteer requested by each project.
enter_on_duty	The number of people that entered on duty at the deadline.
AA	Assignment Area. Different project has different coefficient.
Post	Countries where the project is located.

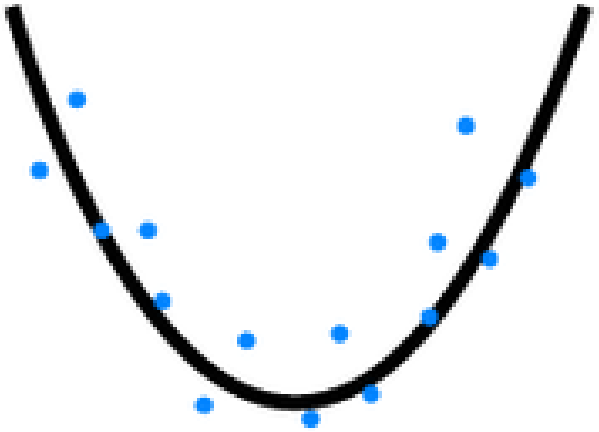
METHODOGY



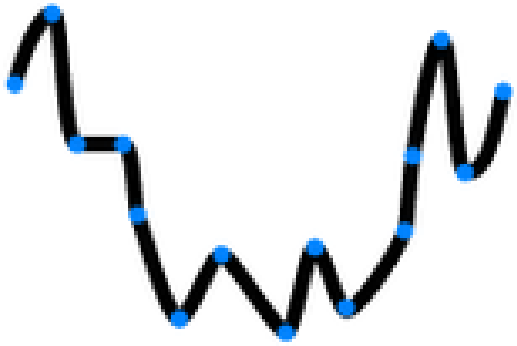
WHAT IS OVERFITTING



Underfitting



Desired

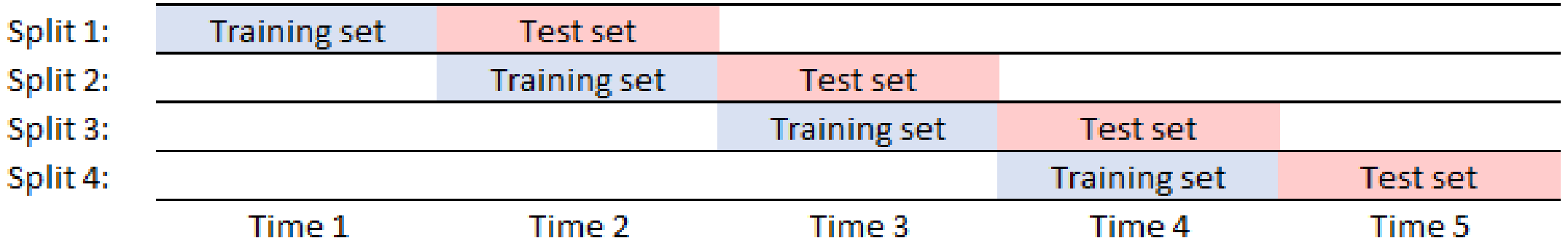


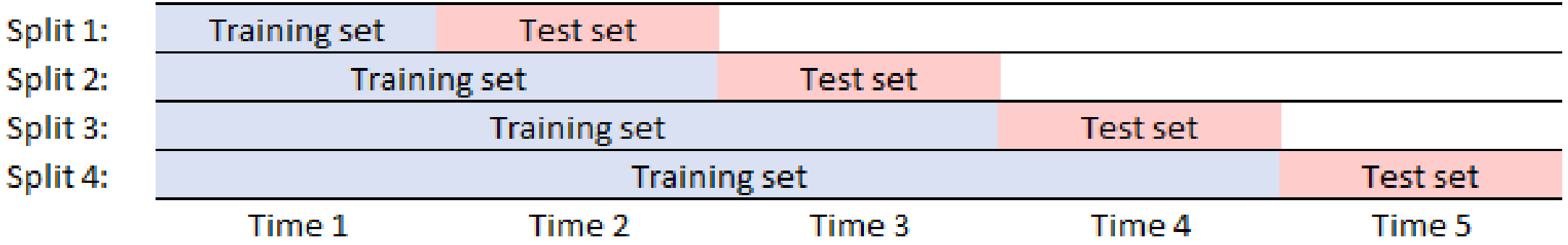
Overfitting

SOLUTIONS

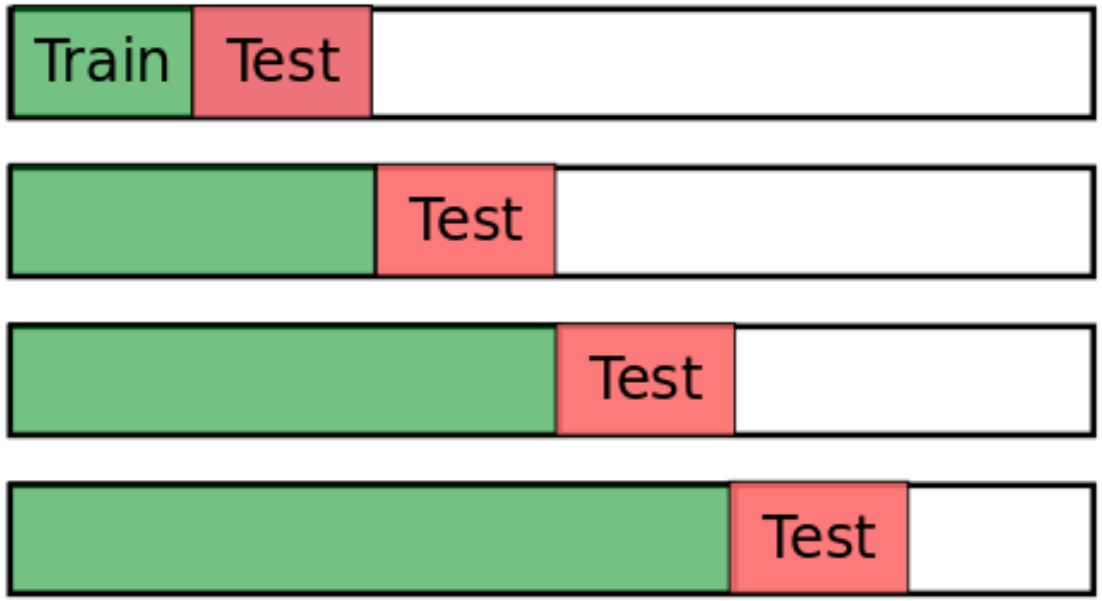
Peace Corps

CROSS VALIDATION





CROSS VALIDATION

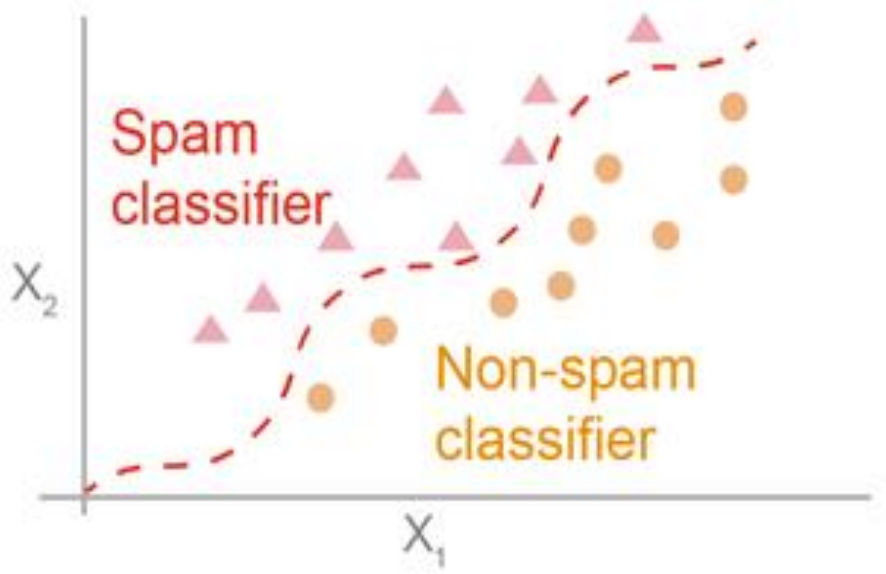


Training Subset	Validation
16Q3-17Q3 data	17Q4-18Q3 data
(281 observations)	(249 observations)

Peace Corps

REGRESSION

- Regression maps the behavior of a dependent variable relative to one or more independent variables.



Advantages	Use cases
Regression is useful for identifying continuous (not necessarily distinct) relationships between variables.	Traffic flow analysis, email filtering

ASSUMPTION

- The number of people enter on duty ideally would excess the number of requested volunteer by up to 6%.

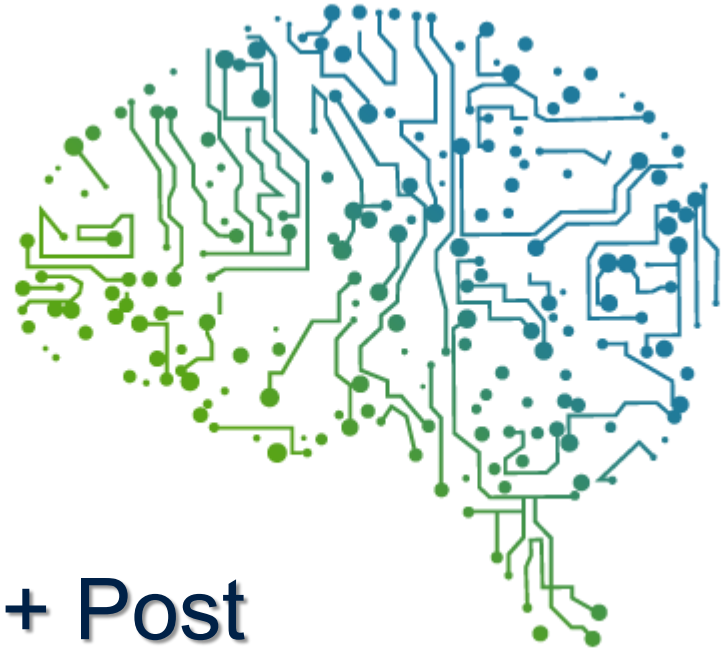
- Every post will have a fixed effect which is unique to that post; Could be caused by politics, culture and etc.

Peace Corps

RESULTS

Peace Corps

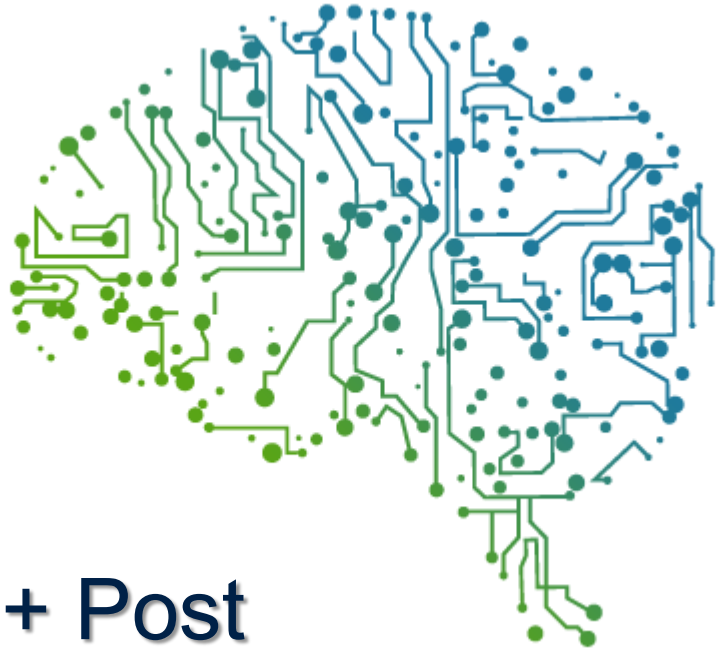
EQUATION



- inv_at_ddl = -1.11 + 1.11 * VR +
  0.31 * enter_on_duty+ AA + Post

- *82.18% of the predicted inv_at_ddl number is within ±2 of the actual inv_at_ddl number in the testing set, 62.35% of the predicted inv_at_ddl number is within ±1 of the actual inv_at_ddl number in the testing set; the model is doing a good job*

EQUATION



- inv_at_ddl = -1.11 + 1.11 * VR +
  0.31 * (1.06 * VR) + AA + Post

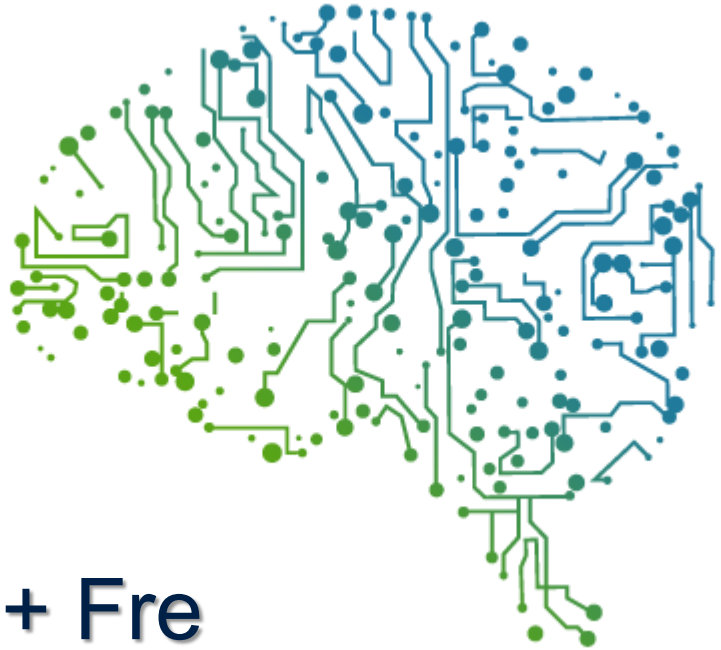
- 82.18% of the predicted inv_at_ddl number is within ±2 of the actual inv_at_ddl number in the testing set, 62.35% of the predicted inv_at_ddl number is within ±1 of the actual inv_at_ddl number in the testing set; the model is doing a good job

OLD MODEL

- Inv_deadline = -3.56 + 1.48 * VR + AA +1.32 * Fre

- *Frequency/Total Candidates. Present the medical friendliness in different countries. 54.17% -- 99.97%.*

USING FREQ

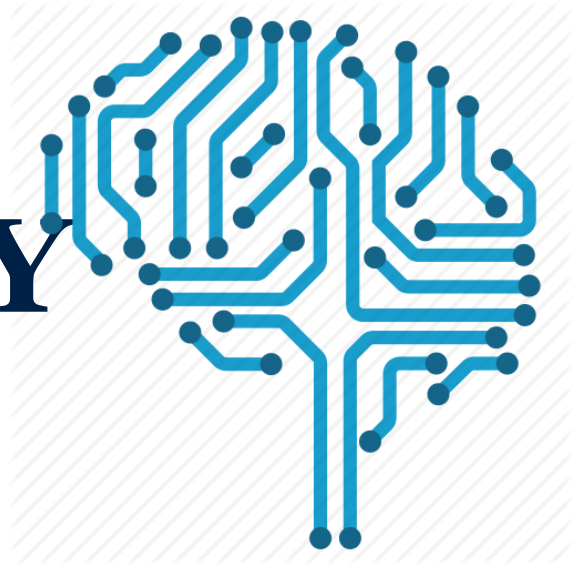


- inv_at_ddl = -0.29 + 1.08 * VR +
  0.35 * (1.06 * VR) + AA + Fre

- *80.72% of the predicted inv_at_ddl number is within ±2 of the actual inv_at_ddl number in the testing set; 68.27% of the predicted inv_at_ddl number is within ±1 of the actual inv_at_ddl number in the testing set.*

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

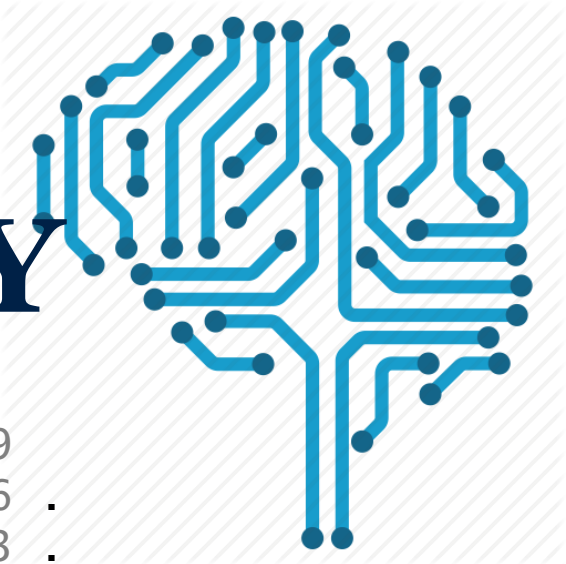
NEW MODEL SUMMARY



```
Call:
lm(formula = inv_at_ddl ~ VR + enter_on_duty + AA + Post, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-5.9012 -0.9274 -0.0431  0.8414 18.1609

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    -1.109293   0.980370  -1.132 0.258454
VR                              1.112639   0.044319  25.105  < 2e-16 ***
enter_on_duty                   0.312186   0.045186   6.909  1.7e-11 ***
AA103                           0.428515   1.100347   0.389 0.697141
AA104                          -0.906548   0.901443  -1.006 0.315125
AA110                           0.094942   0.827311   0.115 0.908688
AA114                           0.994895   0.926962   1.073 0.283726
AA117                           0.261140   0.752013   0.347 0.728565
AA124                          -0.381187   1.108742  -0.344 0.731158
AA131                           1.228301   1.365875   0.899 0.368992
AA134                           1.560345   2.145817   0.727 0.467513
```

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

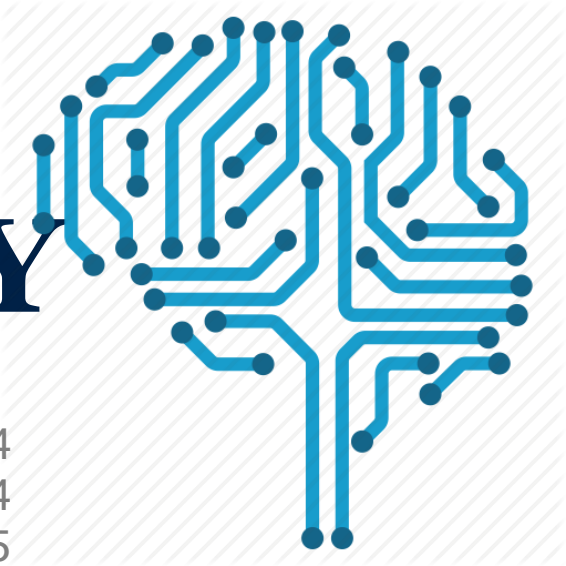
NEW MODEL SUMMARY



AA140	1.070995	0.800587	1.338	0.181659	
AA144	1.708552	0.924087	1.849	0.065136	.
AA145	1.642656	0.919458	1.787	0.074693	.
AA154	1.348467	0.771085	1.749	0.081018	.
AA155	0.254828	0.751006	0.339	0.734532	
AA162	0.864327	0.785451	1.100	0.271745	
AA164	0.720078	0.815721	0.883	0.377849	
AA170	0.774331	0.824067	0.940	0.347911	
AA171	0.857050	0.768284	1.116	0.265223	
AA172	0.800494	1.003335	0.798	0.425393	
AA173	0.813382	0.902409	0.901	0.367894	
AA175	0.784430	0.869444	0.902	0.367429	
AA177	1.131433	1.390893	0.813	0.416392	
AA191	0.737883	0.823102	0.896	0.370490	
AA199	1.278937	2.106993	0.607	0.544164	
PostArmenia	-0.593467	1.114016	-0.533	0.594489	
PostBelize	-0.450633	1.043815	-0.432	0.666157	
PostBenin	0.283855	0.962893	0.295	0.768289	
PostBotswana	-1.419037	0.799290	-1.775	0.076521	.
PostBurkina Faso	0.588917	1.046816	0.563	0.574005	

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

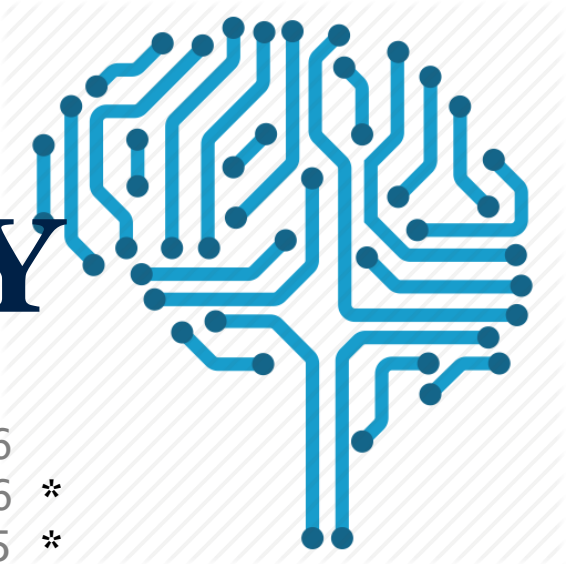
NEW MODEL SUMMARY



PostCambodia	0.239502	0.975576	0.245	0.806184	
PostCameroon	0.026980	0.857498	0.031	0.974914	
PostColombia	-0.080318	0.947239	-0.085	0.932465	
PostComoros	-0.570094	1.316649	-0.433	0.665234	
PostCosta Rica	-1.316865	1.105679	-1.191	0.234289	
PostDominican Republic	-1.057480	0.878087	-1.204	0.229115	
PostEast Timor	-0.062303	1.005695	-0.062	0.950631	
PostEastern Caribbean	-1.266334	1.533953	-0.826	0.409510	
PostEcuador	-1.579019	0.944591	-1.672	0.095299	.
PostEthiopia	2.098079	0.966404	2.171	0.030459	*
PostFederated States of Micronesia	1.067934	2.036305	0.524	0.600229	
PostFiji	-5.791836	1.594059	-3.633	0.000312	***
PostGeorgia	0.021996	1.136789	0.019	0.984572	
PostGhana	0.856583	0.882288	0.971	0.332144	
PostGuatemala	1.240006	0.884099	1.403	0.161446	
PostGuinea	-0.782440	0.896535	-0.873	0.383278	
PostGuyana	0.416594	0.909684	0.458	0.647209	
PostIndonesia	-2.527831	1.574711	-1.605	0.109146	
PostJamaica	-1.144229	0.981165	-1.166	0.244162	
PostKosovo	0.624249	1.344553	0.464	0.642675	

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

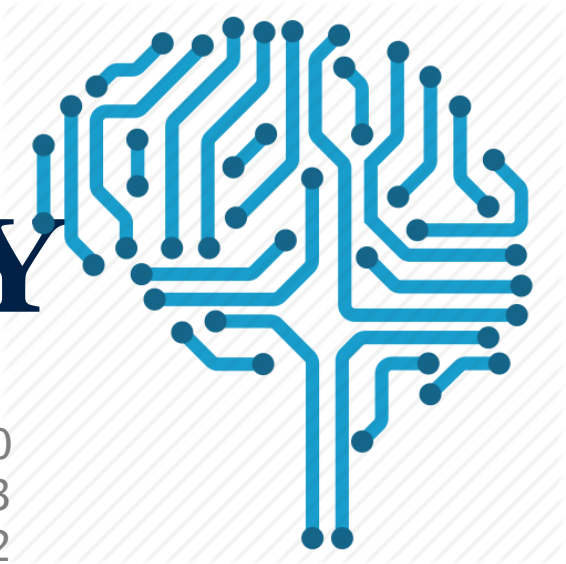
NEW MODEL SUMMARY



PostKyrgyz Republic	0.471673	1.547114	0.305	0.760606	
PostLesotho	-2.305907	0.976237	-2.362	0.018606	*
PostLiberia	2.656137	1.087034	2.443	0.014935	*
PostMacedonia	-0.234346	0.943504	-0.248	0.803956	
PostMadagascar	1.255431	1.011776	1.241	0.215328	
PostMalawi	3.508045	0.969790	3.617	0.000332	***
PostMexico	0.254670	0.988755	0.258	0.796861	
PostMoldova	0.690748	0.973718	0.709	0.478454	
PostMongolia	-0.043951	0.961150	-0.046	0.963548	
PostMorocco	-2.486787	1.294598	-1.921	0.055385	.
PostMozambique	-1.760810	0.941743	-1.870	0.062179	.
PostMyanmar (Burma)	0.455518	1.119106	0.407	0.684177	
PostNamibia	-2.116103	1.054406	-2.007	0.045364	*
PostNepal	0.434634	1.233318	0.352	0.724698	
PostNicaragua	-0.240450	0.827172	-0.291	0.771425	
PostPanama	-0.559756	0.868124	-0.645	0.519397	
PostParaguay	-0.621906	0.807306	-0.770	0.441503	
PostPeru	-0.696292	0.909446	-0.766	0.444308	
PostPhilippines	-0.229474	0.917703	-0.250	0.802663	
PostRwanda	-1.698509	1.114895	-1.523	0.128353	

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

NEW MODEL SUMMARY



PostSamoa	-0.798551	1.201040	-0.665	0.506470	
PostSenegal	0.939578	0.888294	1.058	0.290753	
PostSierra Leone	1.215578	0.932386	1.304	0.193002	
PostSouth Africa	-0.535653	0.919673	-0.582	0.560567	
PostSwaziland	-0.412118	1.530951	-0.269	0.787908	
PostTanzania	-1.025551	0.839878	-1.221	0.222707	
PostThailand	-0.084263	1.020392	-0.083	0.934224	
PostThe Gambia	0.003689	0.932160	0.004	0.996844	
PostThe Peoples Republic of China	2.677556	1.264171	2.118	0.034727	*
PostTogo	1.262407	0.963488	1.310	0.190790	
PostTonga	-1.196617	1.315810	-0.909	0.363625	
PostUganda	0.200107	0.884718	0.226	0.821165	
PostUkraine	2.853387	0.917875	3.109	0.002000	**
PostVanuatu	0.216257	0.982926	0.220	0.825962	
PostZambia	0.828064	0.880458	0.940	0.347477	

Residual standard error: 1.908 on 444 degrees of freedom
Multiple R-squared: 0.9873, Adjusted R-squared: 0.9849
F-statistic: 405.9 on 85 and 444 DF, p-value: < 2.2e-16

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

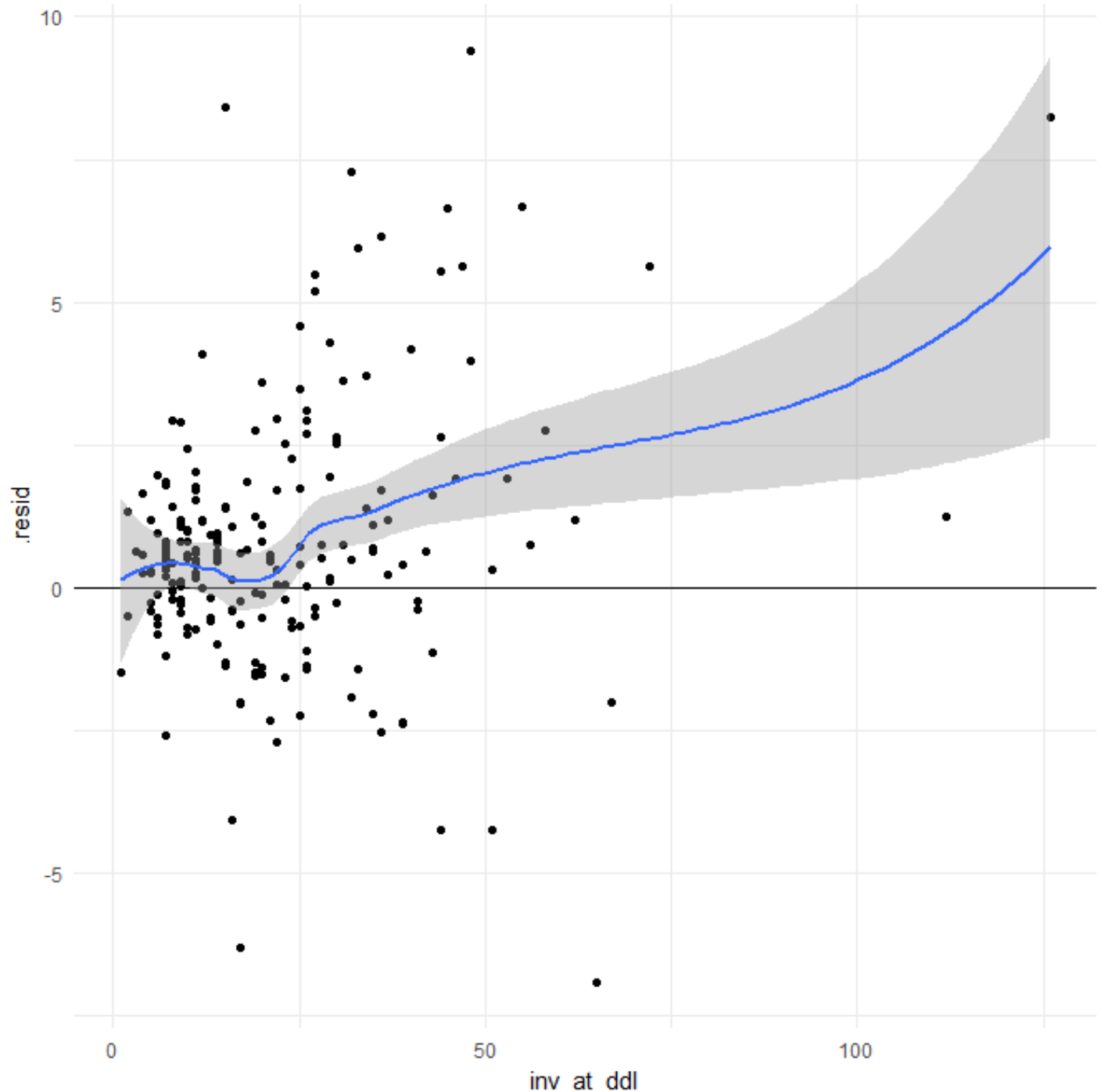
OLD METHOD SUMMARY

```
Call:
lm(formula = inv_at_ddl ~ VR + enter_on_duty + AA + freq, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-7.4132 -1.0951 -0.0521  0.8557 21.0633

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -0.286221   1.097849  -0.261   0.7944
VR             1.075240   0.046388  23.179  < 2e-16 ***
enter_on_duty  0.347984   0.046939   7.414 5.25e-13 ***
AA103         -0.287519   1.208555  -0.238   0.8121
AA104         -1.752015   0.959509  -1.826   0.0685 .
AA110         -0.372266   0.886602  -0.420   0.6748
AA114          0.006907   0.973155   0.007   0.9943
AA117         -0.041841   0.800299  -0.052   0.9583
AA124         -1.103345   1.141381  -0.967   0.3342
AA131         -0.026009   1.442986  -0.018   0.9856
AA134          0.886382   2.278338   0.389   0.6974
```

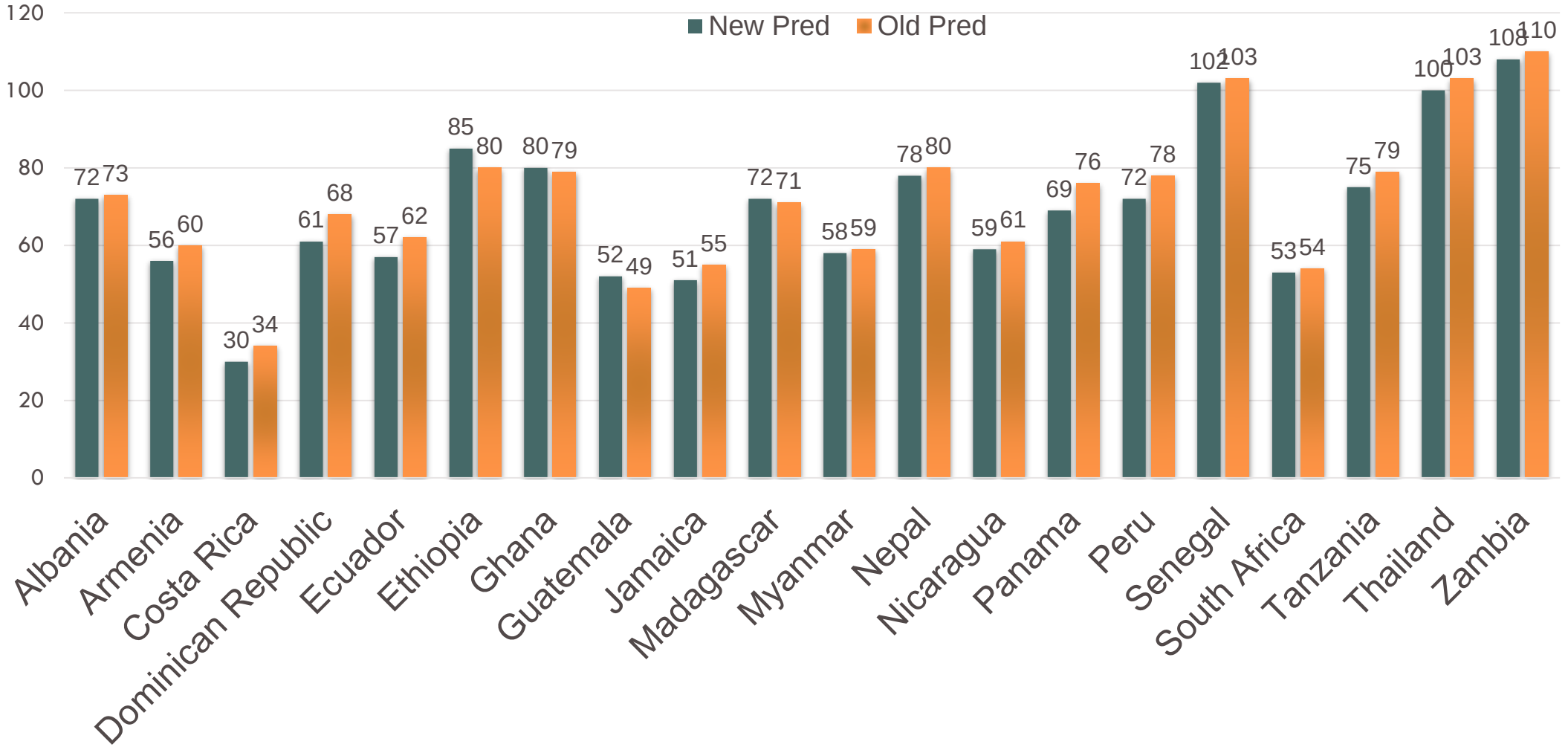
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

OLD METHOD SUMMARY

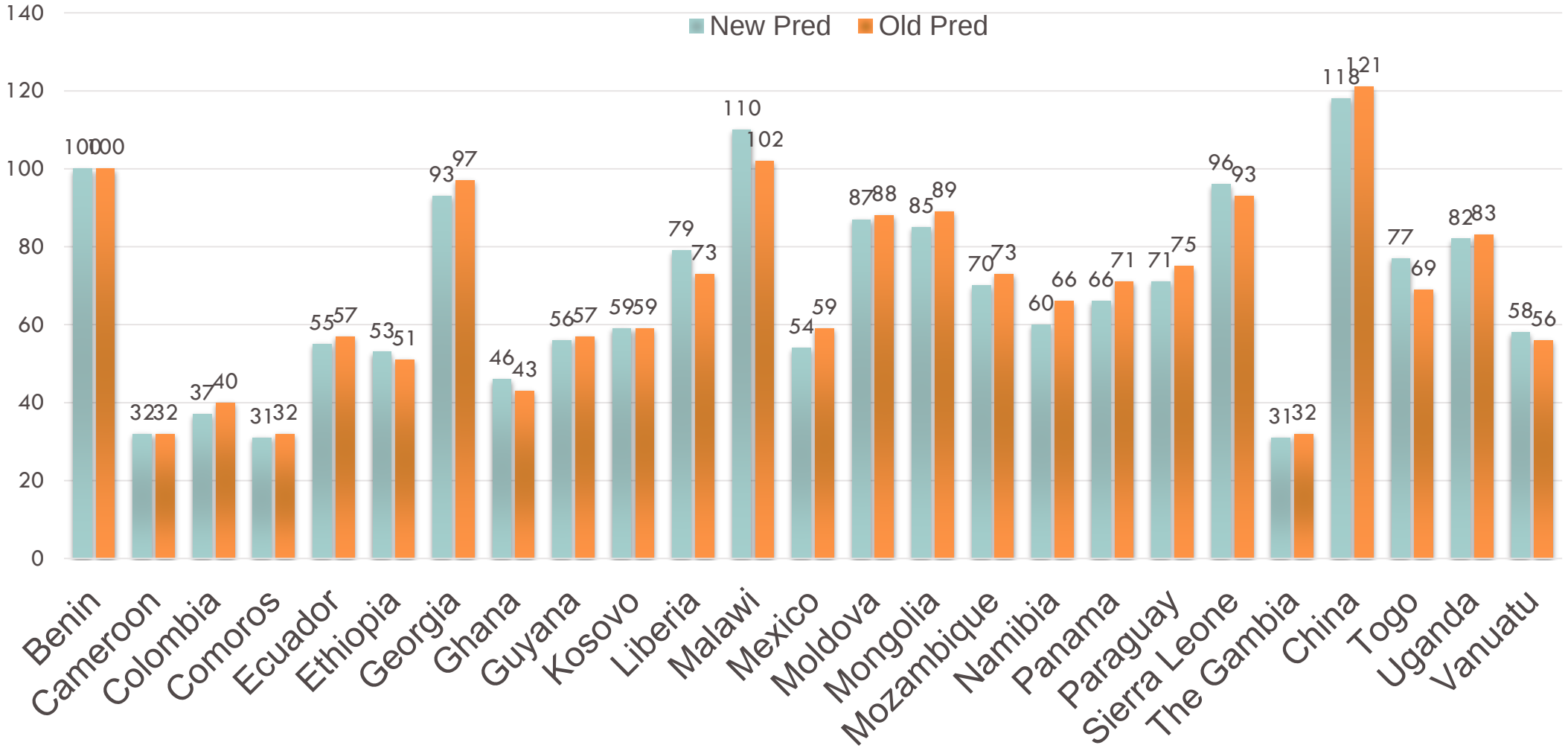
```
AA140          0.243015   0.833744   0.291   0.7708
AA144          0.927519   0.977589   0.949   0.3432
AA145          0.915638   0.940267   0.974   0.3306
AA154          0.446496   0.798813   0.559   0.5764
AA155         -0.435821   0.774715  -0.563   0.5740
AA162          0.424088   0.818037   0.518   0.6044
AA164         -0.962299   0.816892  -1.178   0.2394
AA170         -0.113358   0.837284  -0.135   0.8924
AA171          0.157174   0.784330   0.200   0.8413
AA172          1.228264   1.054136   1.165   0.2445
AA173         -0.074782   0.913723  -0.082   0.9348
AA175          0.076305   0.884132   0.086   0.9313
AA177          0.565252   1.440766   0.392   0.6950
AA191          0.298839   0.859870   0.348   0.7283
AA199          1.470553   2.278741   0.645   0.5190
freq          -0.158467   0.939279  -0.169   0.8661
---
Residual standard error: 2.161 on 503 degrees of freedom
Multiple R-squared:  0.9815, Adjusted R-squared:  0.9806
F-statistic:  1028 on 26 and 503 DF,  p-value: < 2.2e-16
```


10
5
0
-5
.resid
0
50
100
inv_at_ddl

COMPARISON (FY19 Q2)



COMPARISON (FY19 Q3)



2. CANDIDATE ATTRITION

Peace Corps

DATA

17,580 observations in total

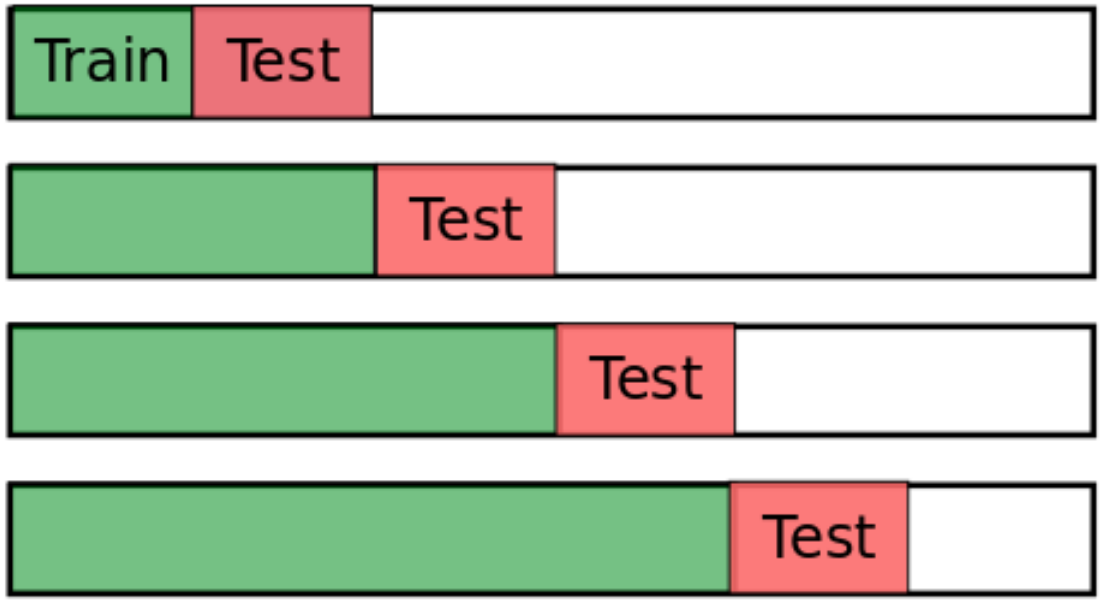
Variable Name	Interpret
EOD	A dummy variable indicating whether the candidate entered on duty (1) or not (0)
State	The state where the candidate comes from
Sex	The candidate's gender
Diversity	The candidate's ethnicity background: White, Black, Asian, Pacific, Indian, Hispanic or Latino, unknown, two or more races
Married_DP	The candidate's marital status: Yes/No
Serving_Spouse	Whether the candidate will serve with spouse: Yes, No

DATA

17,580 observations in total

Variable Name	Interpret
Med_Sort	The candidate's medical sort information: medical pending, nomination cleared, nomination validation required
Have_you_been_arres	The candidate's arresting information: Yes, No and unknown
Degree_Type	The candidate's degree type: associate, bachelor, master, doctorate and no degree.
Language_Level	The candidate's language level: F1, F2, F9, S1, S2, S9, RL, other and None
IRT_score	The IRT Score the candidate got ranging from 0 to 30
Agebin	The candidate's age at invited date in bins

CROSS VALIDATION



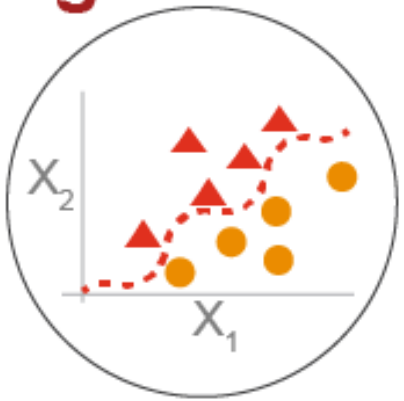
Training Subset	Validation
FY15-17 data (11,660 observations)	FY18 data (5,920 observations)

- The candidate EOD rate for the whole dataset is 61.2%, and for the training set is 60.9%.

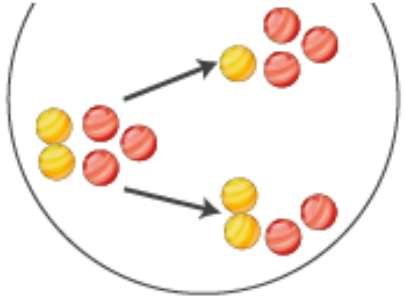
Peace Corps

MACHINE LEARNING

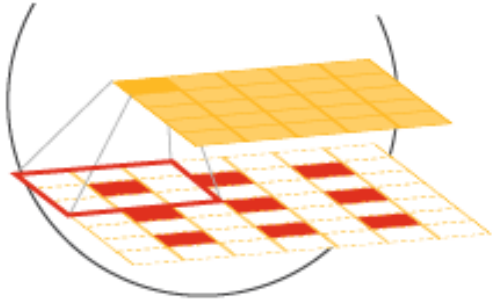
Regression



Naive Bayes classification



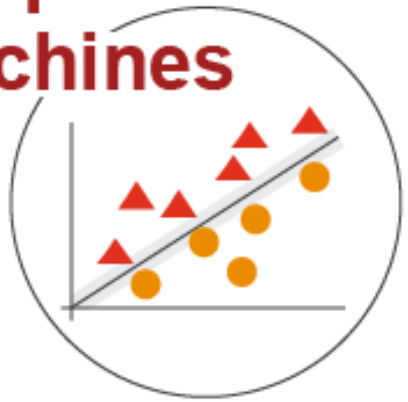
K Nearest Neighbor



Random forest

Decision trees

Support vector machines



Peace Corps

VALIDATION METHOD

Peace Corps

PERFORMANCE METRICS

- Accuracy
- Precision
- Recall
- F1
- ROC-AUC(Area under curve)

CONFUSION METRIC

		Actual	
		Positive	Negative
Predicted	Positive	**True Positive**	**False Positive**
	Negative	**False Negative**	**True Negative**

Peace Corps

PRECISION

		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$= \frac{True\ Positive}{Total\ Predicted\ Positive}$$

Peace Corps

RECALL

		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$= \frac{True\ Positive}{Total\ Actual\ Positive}$$

Peace Corps

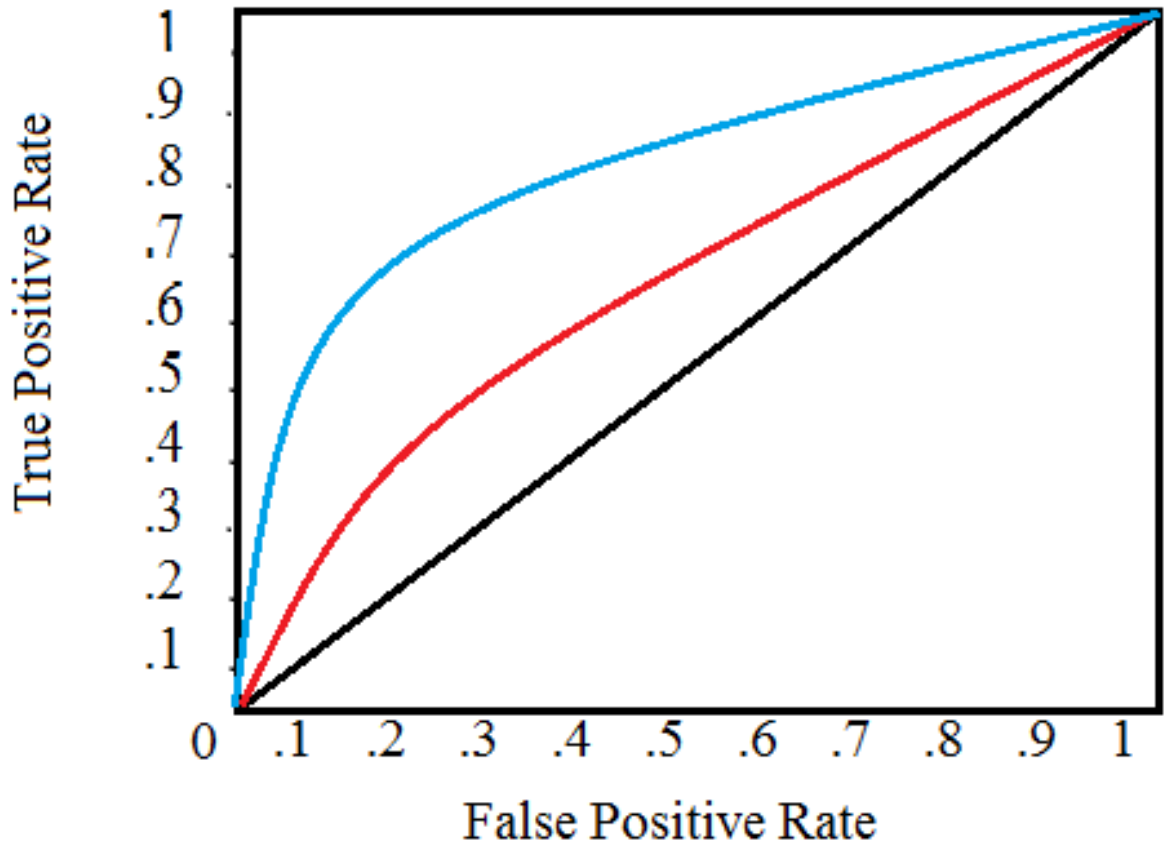
F1

- F1 score: single metric that combines recall and precision using the harmonic mean

$$F1 = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

Peace Corps

ROC-AUC



Peace Corps

RESULTS

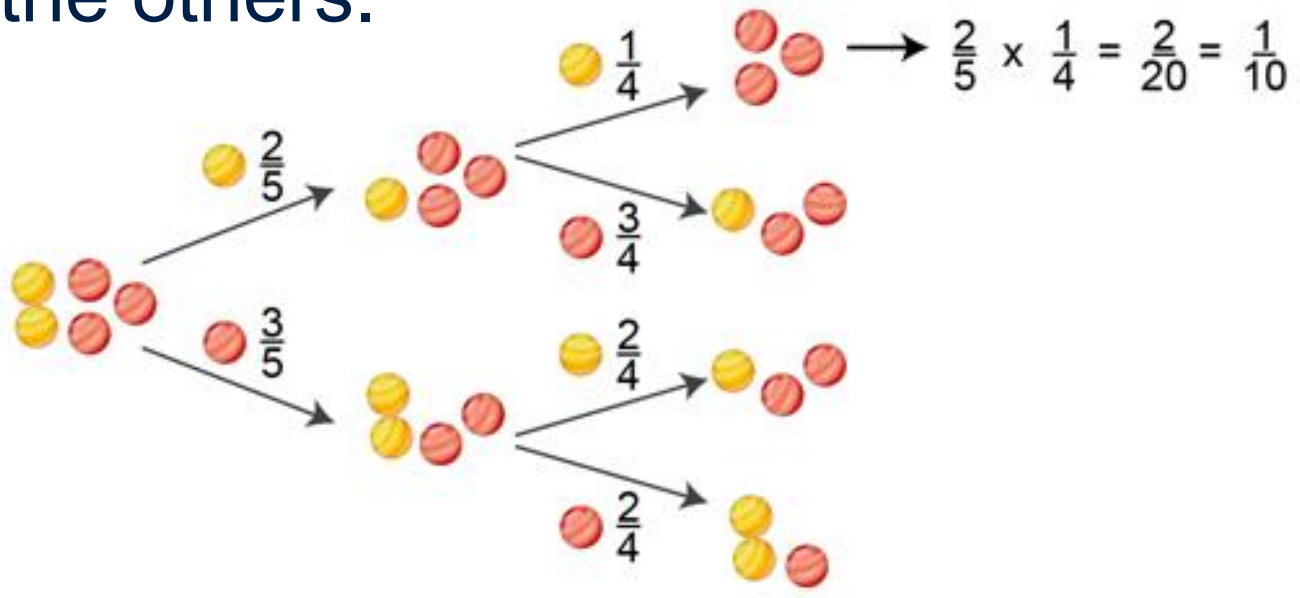
Peace Corps

RESULT

	KNN k=10	Random Forest	SVM	Naïve Bayes	Logistic	Decision Tree
Precision	0.66	0.70	0.65	0.63	0.70	0.67
Recall	0.61	0.42	0.67	0.96	0.43	0.80
F1	0.63	0.52	0.66	0.76	0.53	0.73
AUC-ROC	0.54	0.56	0.55	0.61	0.57	0.61
Running Time	25.54 secs	2.75 mins	2.98 mins	2.37 secs	0.24 secs	0.69 secs

NAÏVE BAYES

- Naïve Bayes compute probabilities, given tree branches of possible conditions. Each individual feature is “naive” or conditionally independent of, and therefore does not influence, the others.



Advantages	Use cases
Naive Bayes methods allow the quick classification of relevant items in small data sets that have distinct features.	Sentiment analysis, consumer segmentation

Source: Rod Pierce, et al., *MathIsFun*, 2014

FINDINGS

Peace Corps

IMPORTANT FEATURES

Variable Name	Interpret
EOD	A dummy variable indicating whether the candidate entered on duty (1) or not (0)
IRT_score	The IRT Score the candidate got ranging from 0 to 30
Agebin	The candidate's age at invited date in bins

- The baseline probability of candidate EOD rate which is 61.2% for the whole dataset, and 60.9% for the training set. Intuitively, I chose 0.61 as the threshold to determine if the candidate will enter on duty.

MORE FOR NAÏVE BAYES

	Naïve Bayes
Precision	0.63
Recall	0.96
F1	0.76
AUC-ROC	0.61
Running Time	2.37 secs

Prediction	Actual 0	Actual 1
0	215	146
1	2050	3509

Peace Corps

MORE FOR NAÏVE BAYES

```
Naive Bayes Classifier for Discrete Predictors
Call:
naiveBayes.default(x = X, y = Y, laplace = laplace, threshold = 0.61, eps = 1, ..3)

A-priori probabilities:
Y
        0         1
0.3878271 0.6121729

Conditional probabilities:
   Agebin
Y       (15,25]    (25,30]    (30,35]    (35,50]    (50,81]
  0 0.64854170 0.17162538 0.04763301 0.04733988 0.08486003
  1 0.74524009 0.14934522 0.03324974 0.02897743 0.04318752

   IRT
Y          (-1,5]       (5,10]      (10,15]      (15,20]      (20,25]      (25,30]
  0 0.0002930832 0.0008792497 0.0218347011 0.1636869871 0.6716002345 0.1417057444
  1 0.0002786033 0.0009286776 0.0187592868 0.1514673105 0.6819279346 0.1466381872
```

3. ET ANALYSIS

Peace Corps

DATA

6,843 observations in total

Variable Name	Interpret
ETR	A dummy variable indicating whether the candidate ET-Resignation (1) or not (0)
Admin_sep	A dummy variable indicating whether the candidate Resig in lieu of Admin Sep/ ET-Admin Sep(1) or not (0)
Sex	The candidate's gender
Diversity	The candidate's ethnicity background: White, Black, Asian, Pacific, Indian, Hispanic or Latino, unknown, two or more races
Married_DP	The candidate's marital status: Yes/No
Serving_Spouse	Whether the candidate will serve with spouse: Yes, No

FINDINGS

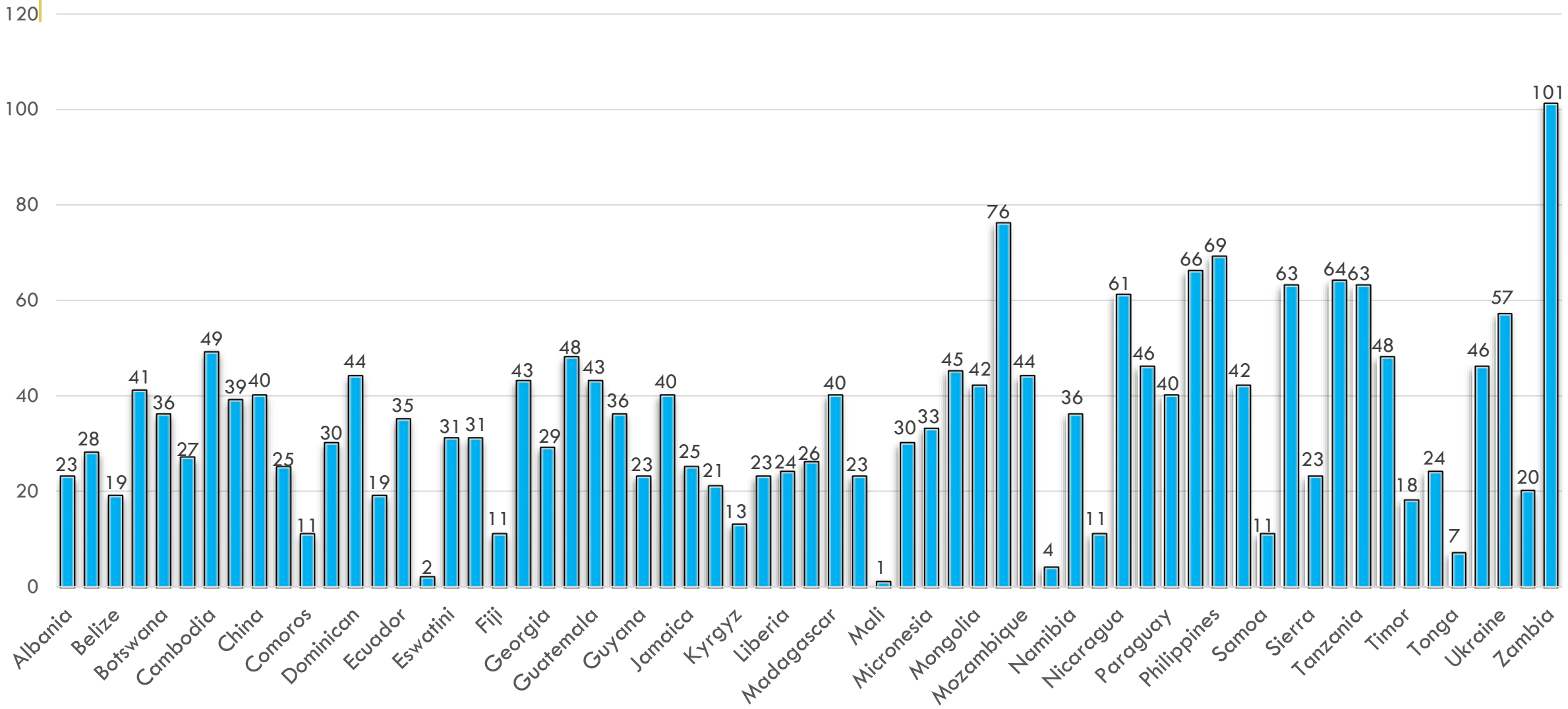
Peace Corps

FACTS

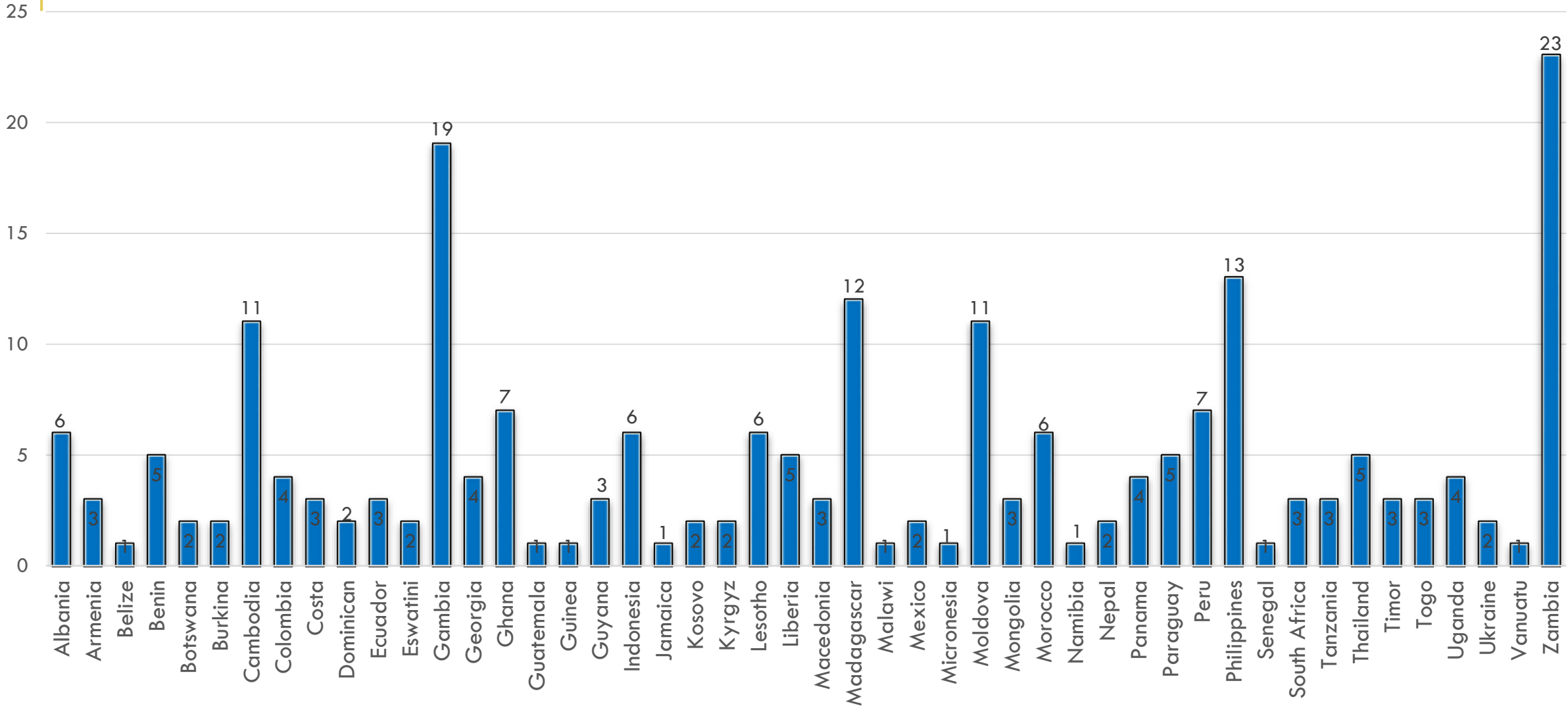
- 2,189 out of 6,843 volunteers ET-resignation
- 932 of them have application data such as gender, degree type etc

- 220 out of 6,843 volunteers resigned in lieu of Administrative Separation or ET-Administrative Separation.
- 116 of them have application data such as gender, degree type etc

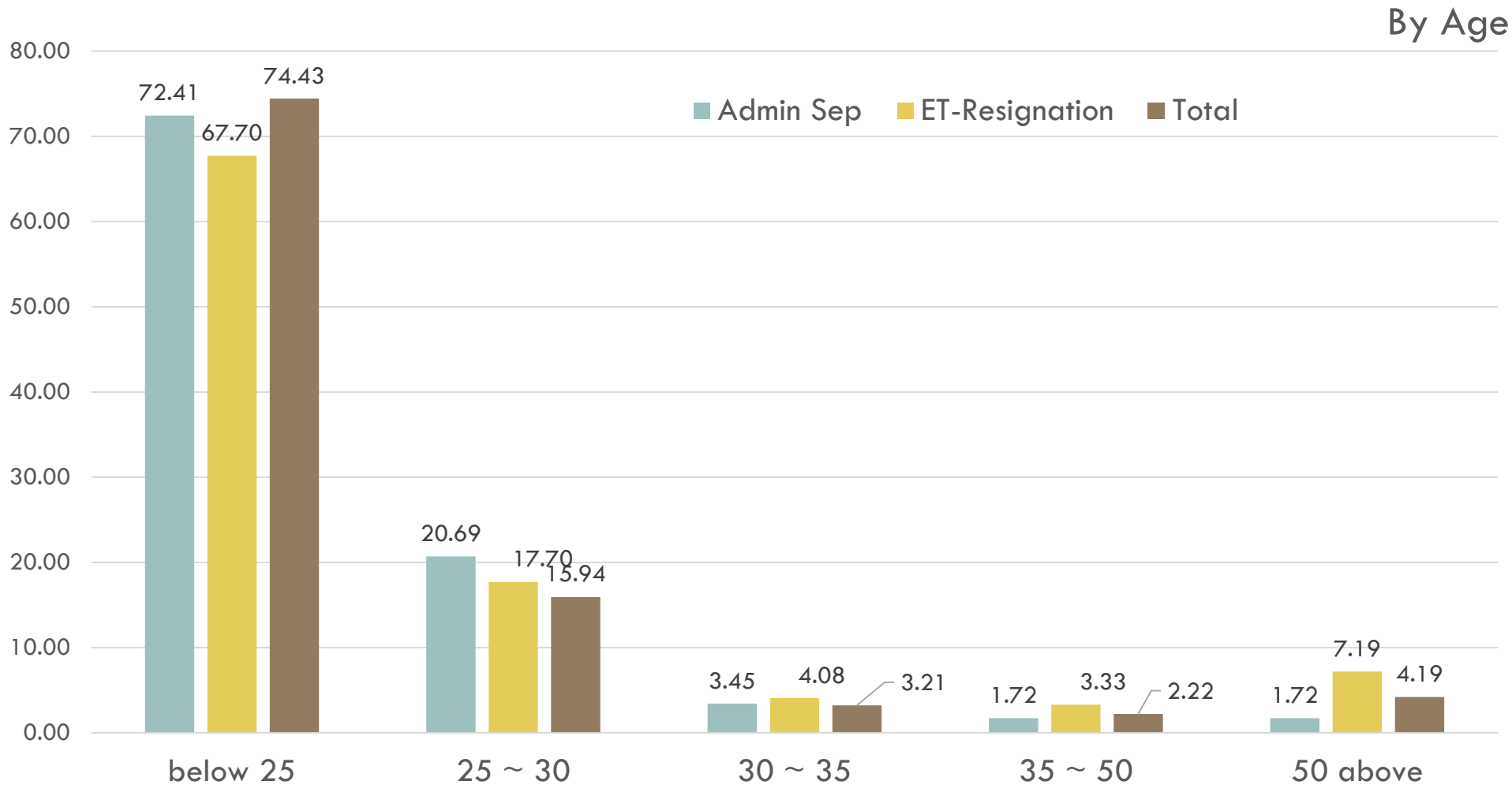
ET-RESIGNATION



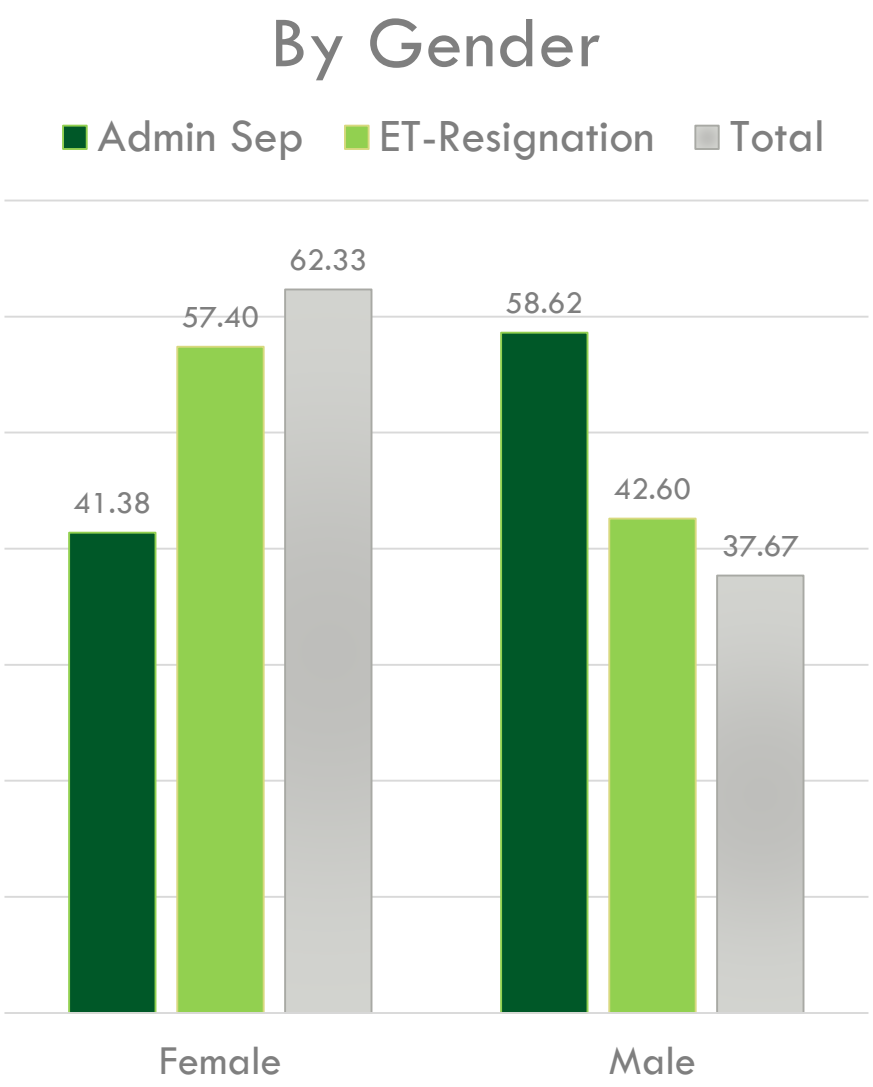
ADMIN SEP

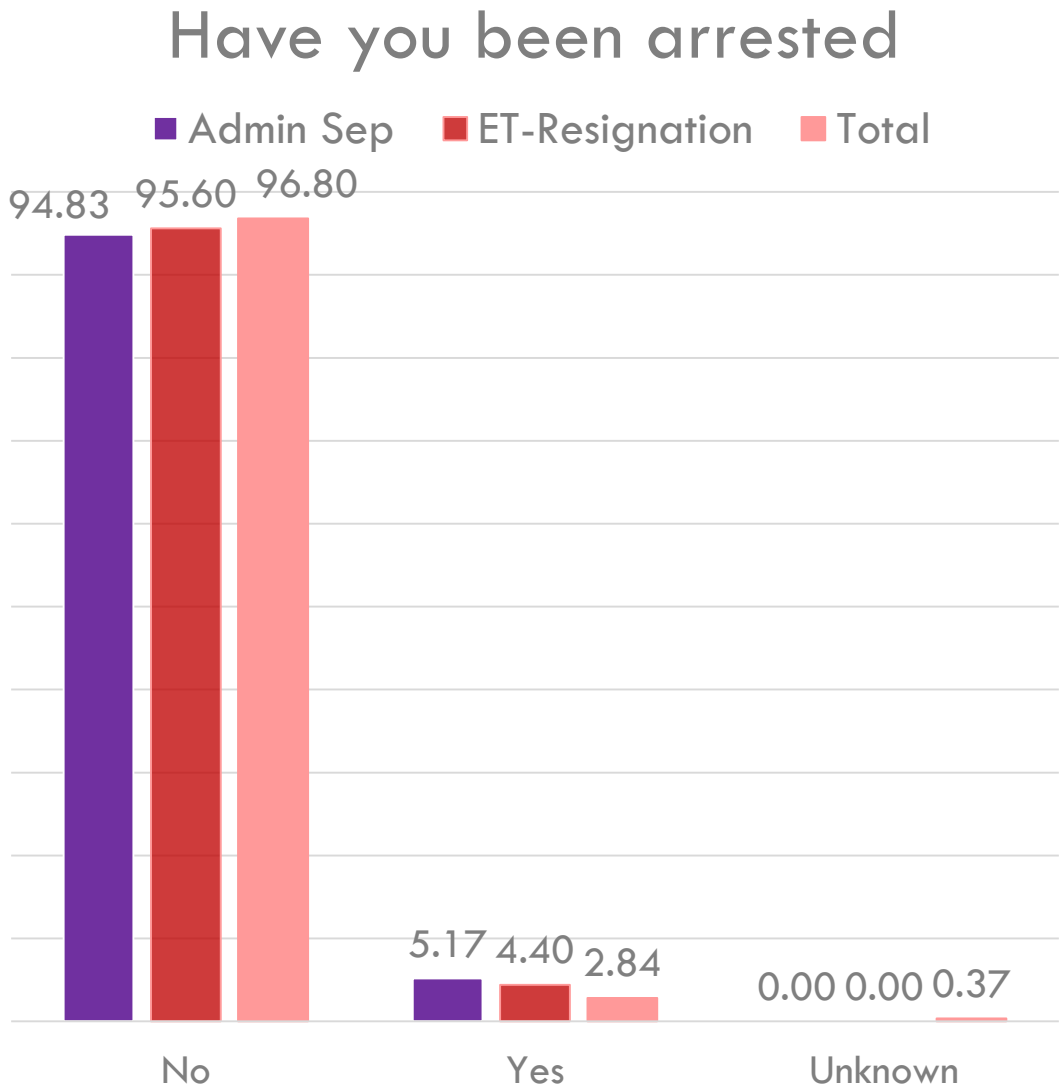

25
20
15
10
5
0
Albania 6
Armenia 3
Belize 1
Benin 5
Botswana 2
Burkina 2
Cambodia 11
Colombia 4
Costa 3
Dominican 2
Ecuador 3
Eswatini 2
Gambia 19
Georgia 4
Ghana 7
Guatemala 1
Guinea 1
Guyana 3
Indonesia 6
Jamaica 1
Kosovo 2
Kyrgyz 2
Lesotho 6
Liberia 5
Macedonia 3
Madagascar 12
Malawi 1
Mexico 2
Micronesia 1
Moldova 11
Mongolia 3
Morocco 6
Namibia 1
Nepal 2
Panama 4
Paraguay 5
Peru 7
Philippines 13
Senegal 1
South Africa 3
Tanzania 3
Thailand 5
Timor 3
Togo 3
Uganda 4
Ukraine 2
Vanuatu 1
Zambia 23

PERCENTAGE COMPARE



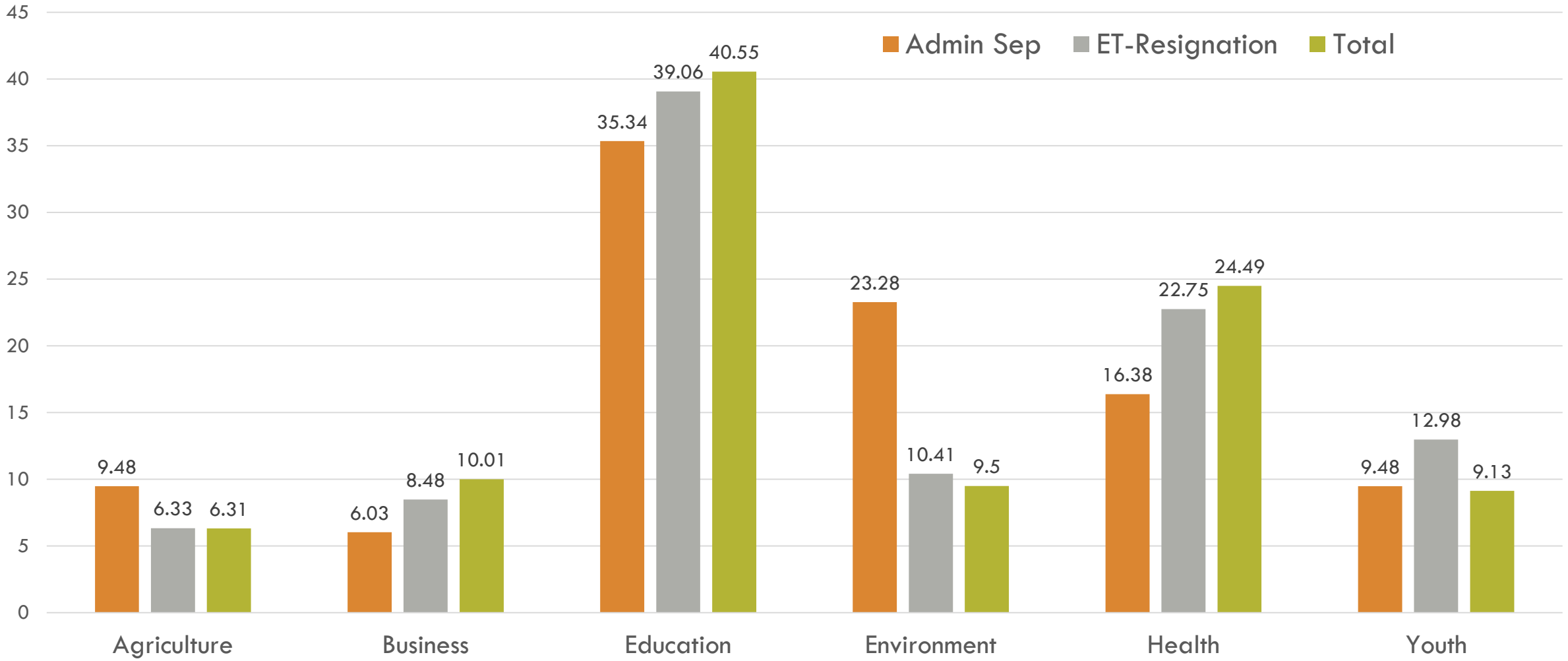
PERCENTAGE COMPARE



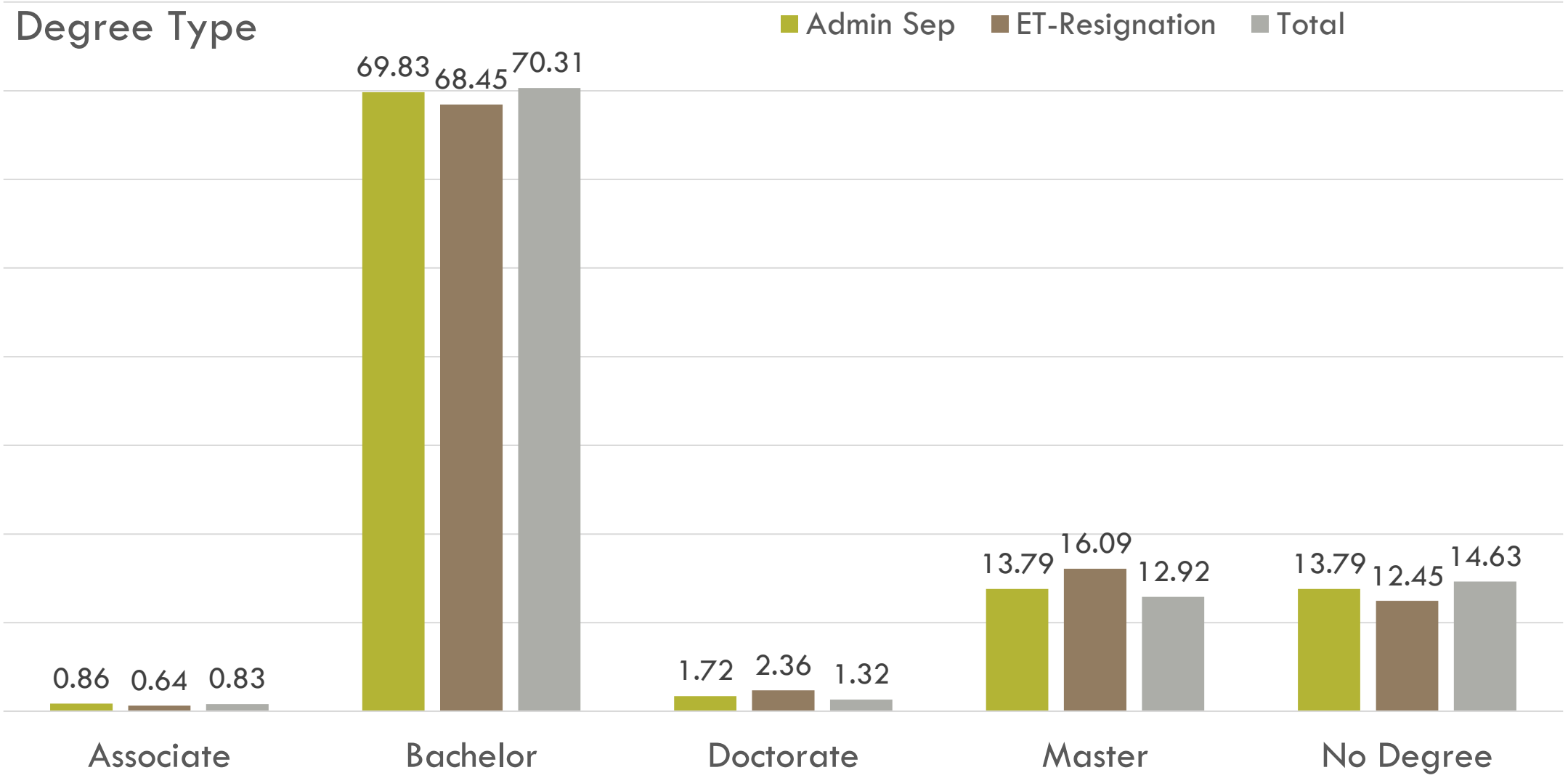


PERCENTAGE COMPARE

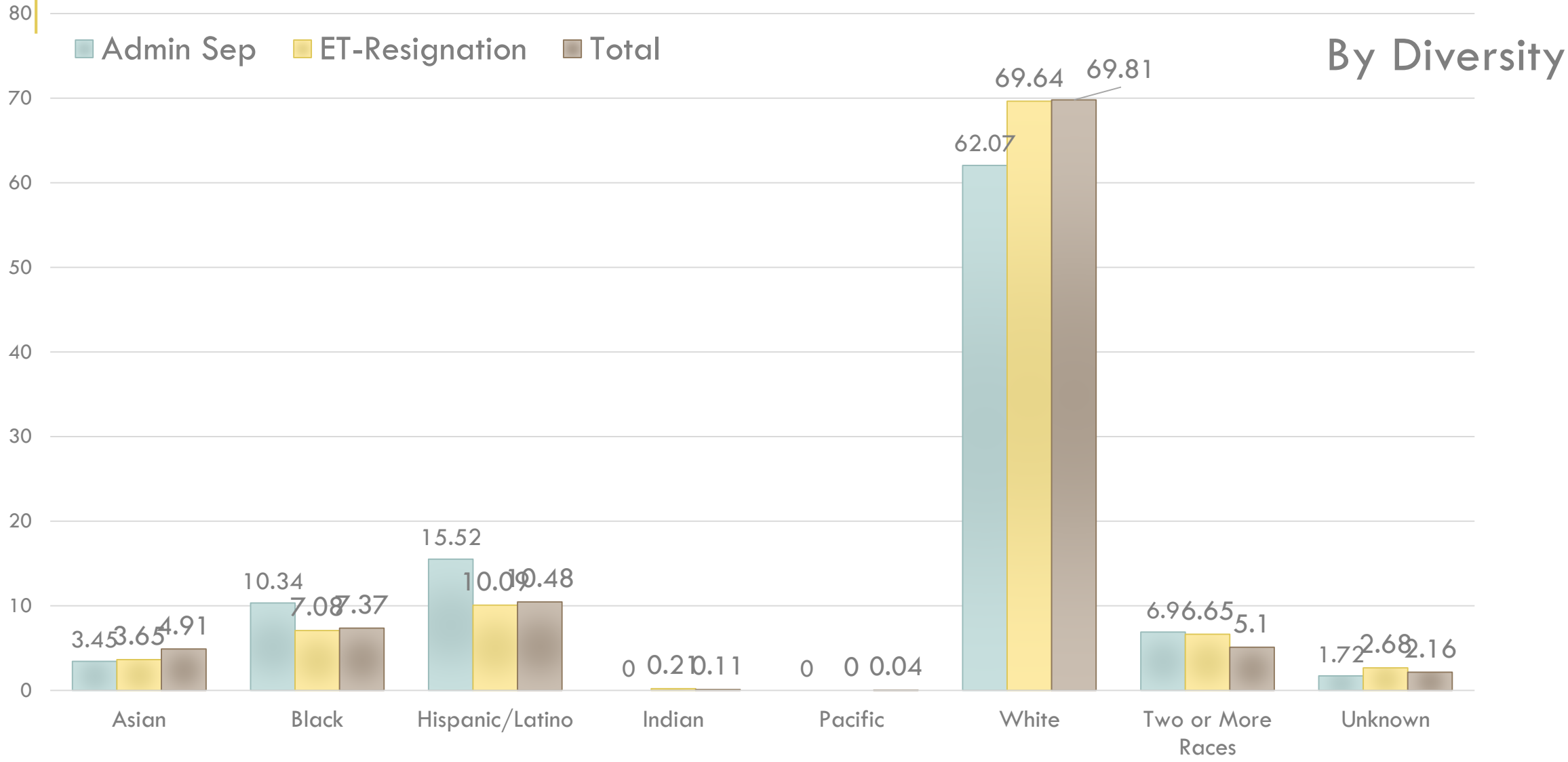
By Sector



PERCENTAGE COMPARE



PERCENTAGE COMPARE



QUESTIONS?

Peace Corps

THANK YOU ☺

Peace Corps