Wage Prediction Hackathon

Reem Mokhtar
Jonna Pander
Kemal Alaeddinoglu

Contents

EDA

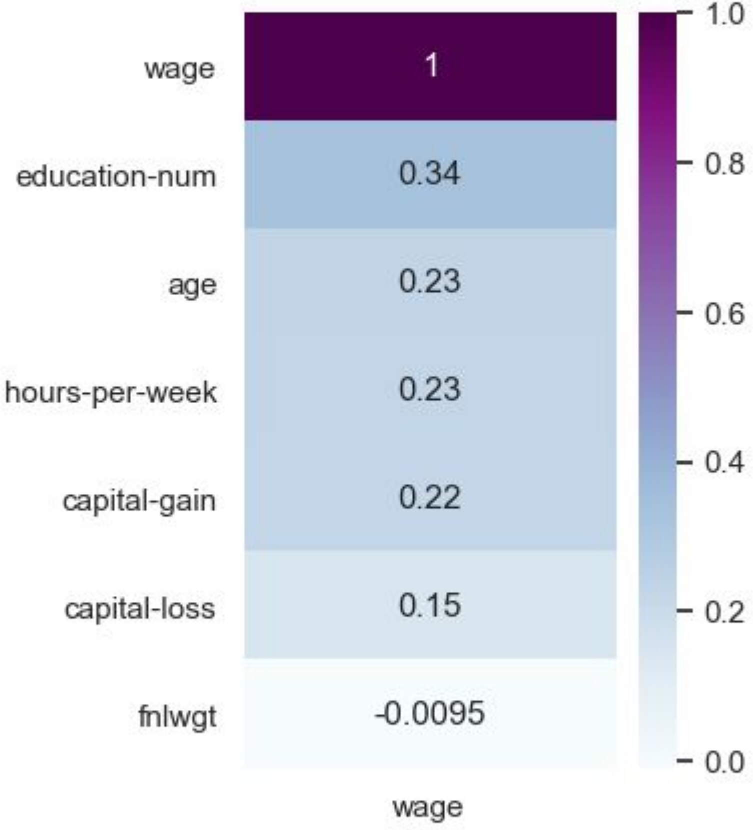
Large Train Sample data - (32561 entries and 14 columns)

7 columns are categorical

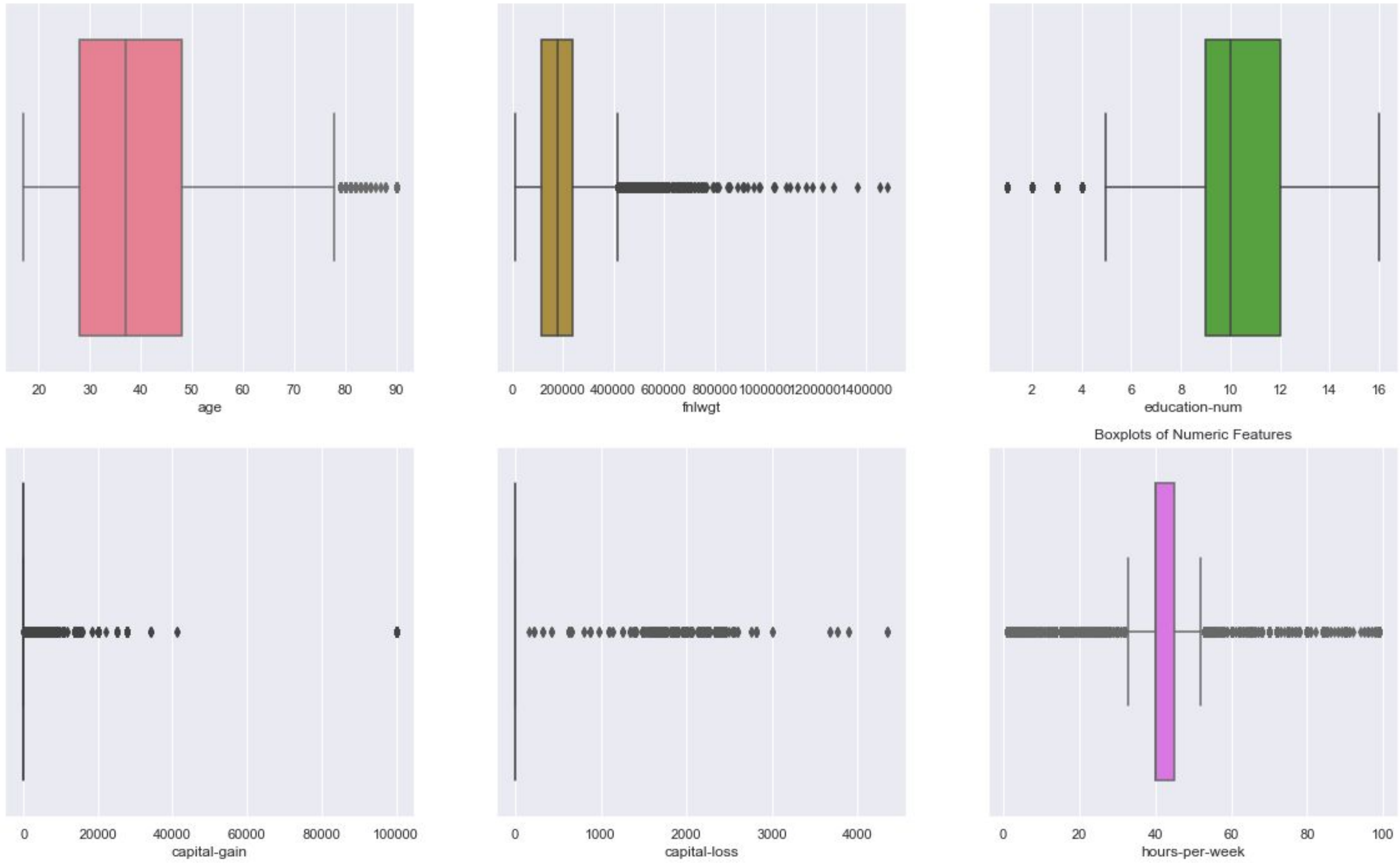
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32561 entries, 0 to 32560
Data columns (total 14 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   age             32561 non-null  int64
 1   workclass       32561 non-null  object
 2   fnlwgt          32561 non-null  int64
 3   education       32561 non-null  object
 4   education-num   32561 non-null  int64
 5   marital-status  32561 non-null  object
 6   occupation      32561 non-null  object
 7   relationship    32561 non-null  object
 8   sex             32561 non-null  object
 9   capital-gain    32561 non-null  int64
 10  capital-loss    32561 non-null  int64
 11  hours-per-week  32561 non-null  int64
 12  native-country  32561 non-null  object
 13  wage            32561 non-null  object
dtypes: int64(6), object(8)
memory usage: 3.5+ MB
```

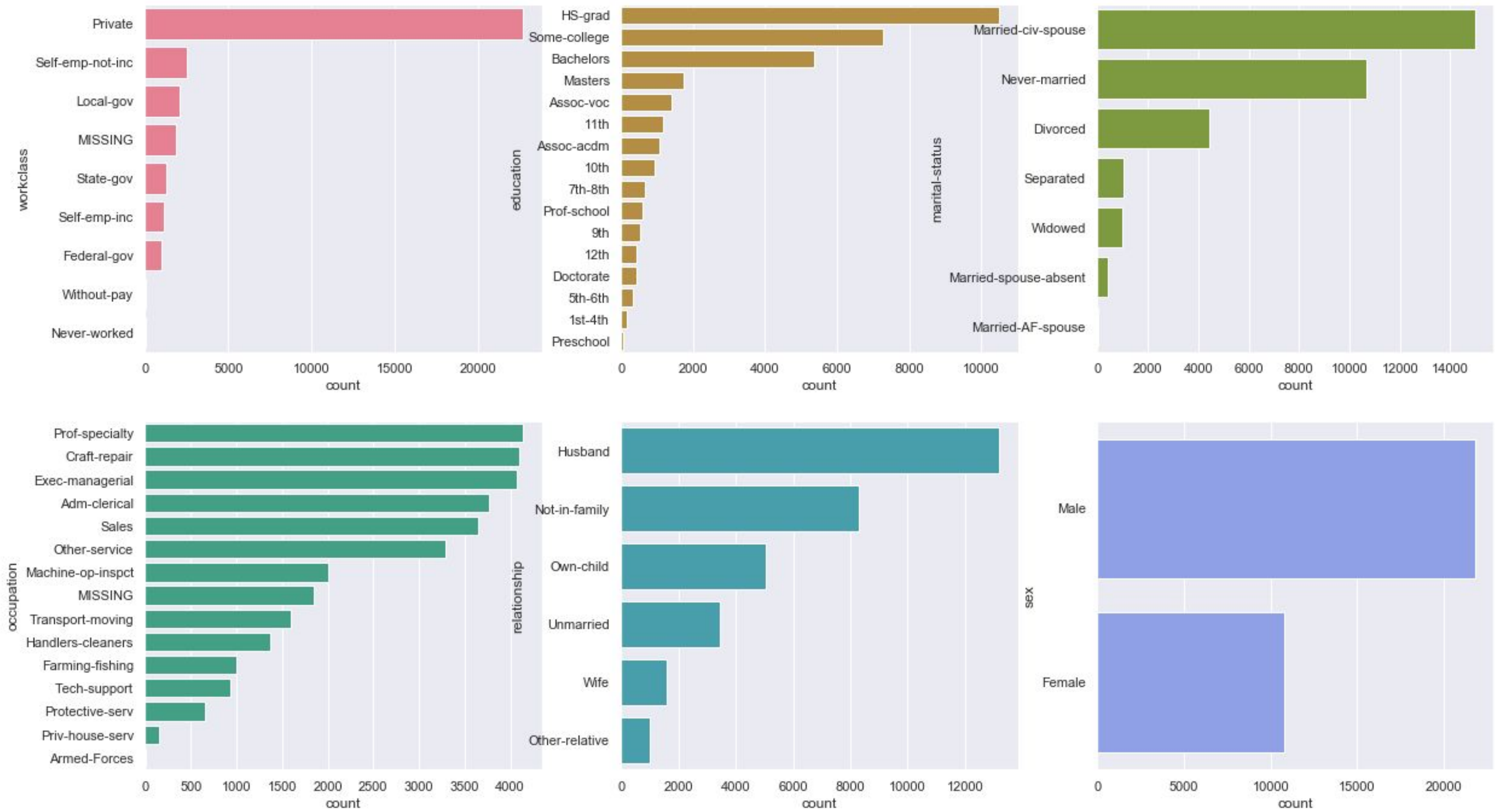
EDA

Not a lot of strong correlation



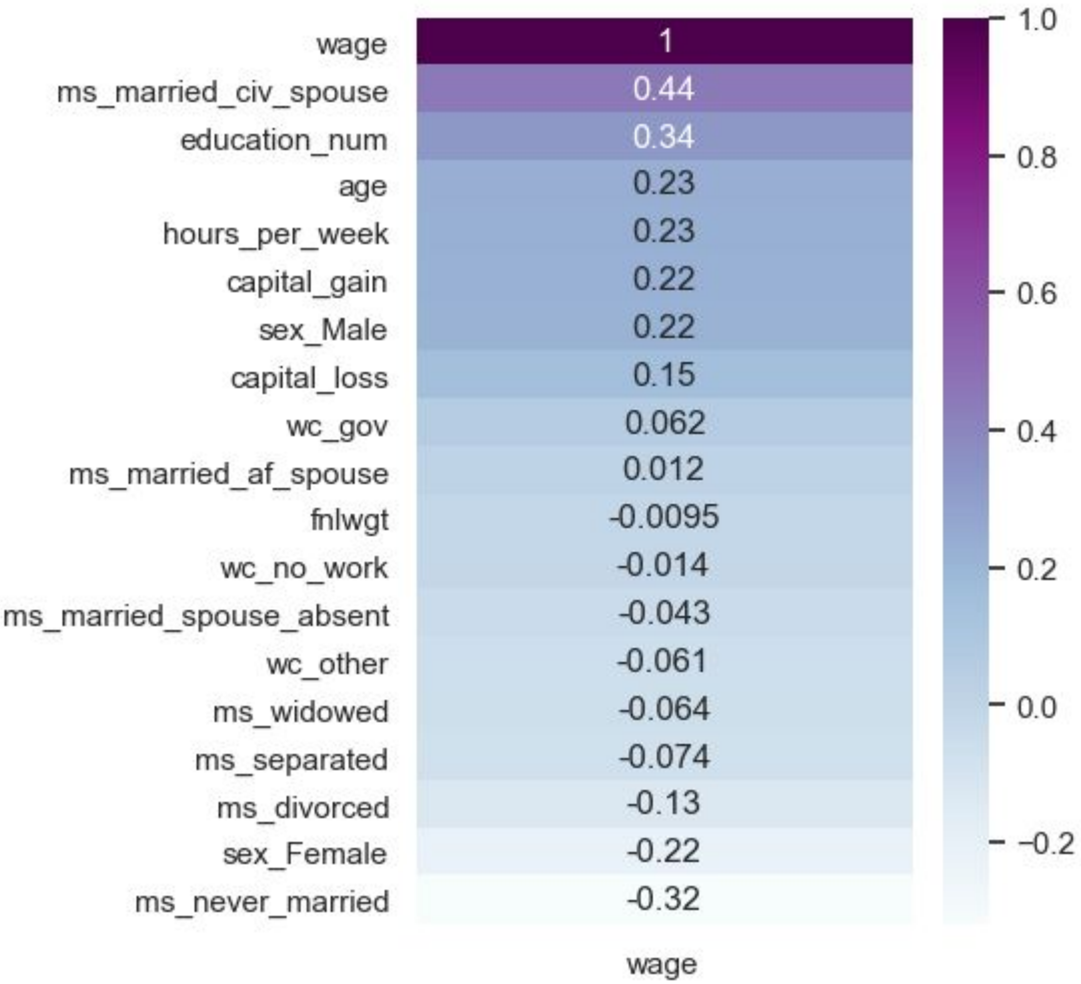
Numeric Features




workclass
Private
Self-emp-not-inc
Local-gov
MISSING
State-gov
Self-emp-inc
Federal-gov
Without-pay
Never-worked
0
5000
10000
15000
20000
count
education
HS-grad
Some-college
Bachelors
Masters
Assoc-voc
11th
Assoc-acdm
10th
7th-8th
Prof-school
9th
12th
Doctorate
5th-6th
1st-4th
Preschool
0
2000
4000
6000
8000
10000
count
marital-status
Married-civ-spouse
Never-married
Divorced
Separated
Widowed
Married-spouse-absent
Married-AF-spouse
0
2000
4000
6000
8000
10000
12000
14000
count
occupation
Prof-specialty
Craft-repair
Exec-managerial
Adm-clerical
Sales
Other-service
Machine-op-inspct
MISSING
Transport-moving
Handlers-cleaners
Farming-fishing
Tech-support
Protective-serv
Priv-house-serv
Armed-Forces
0
500
1000
1500
2000
2500
3000
3500
4000
count
relationship
Husband
Not-in-family
Own-child
Unmarried
Wife
Other-relative
0
2000
4000
6000
8000
10000
12000
count
sex
Male
Female
0
5000
10000
15000
20000
count

Features

1. age
2. education_num
3. capital_gain
4. hours_per_week
5. wc_gov
6. wc_no_work
7. ms_married_af_spouse
8. ms_married_civ_spouse
9. ms_married_spouse_absent
10. ms_never_married
11. ms_separated
12. ms_widowed
13. sex_Male



Scores

Models	train score	test score	Mean CV_score	f1score_train	f1score_test
Random Forest	0.8453	0.8462	0.8450	0.6076	0.5979
Logistic Regression	0.8393	0.8393	0.8392	0.6256	0.6177
KNN	0.8491	0.837	0.8359	0.6553	0.6182
Adaboost	0.9322	0.8259	0.8258	0.8576	0.6155
SVC	0.8453	0.8448	0.8448	0.6256	0.6114

Model

RANDOM FOREST

'ccp_alpha': 0.001,
'max_features': 'sqrt',
'max_leaf_nodes': 30,
'min_impurity_decrease': 0.0,
'min_samples_leaf': 10,
'min_samples_split': 4,
'min_weight_fraction_leaf': 0.0,
'n_estimators': 100

Round 1

1. age
2. education-num
3. marital-status
4. sex

[166]:

	train	test	Mean_CV_Score	f1score_train	f1score_test
RForest_Kemal	0.822569	0.821988	0.82099	0.542534	0.522908
KNN_Kemal	0.830247	0.817382	0.820858	0.611975	0.569498
SVC_Reem	0.819454	0.819224	0.819454	0.548942	0.532804
LOGREG_Jonna	0.818533	0.819122	0.818094	0.560934	0.544235
Adaboost_Reem	0.845516	0.811854	0.81182	0.656387	0.568139
BagginClassifier_Jonna	0.844419	0.80561	0.811337	0.651943	0.559499

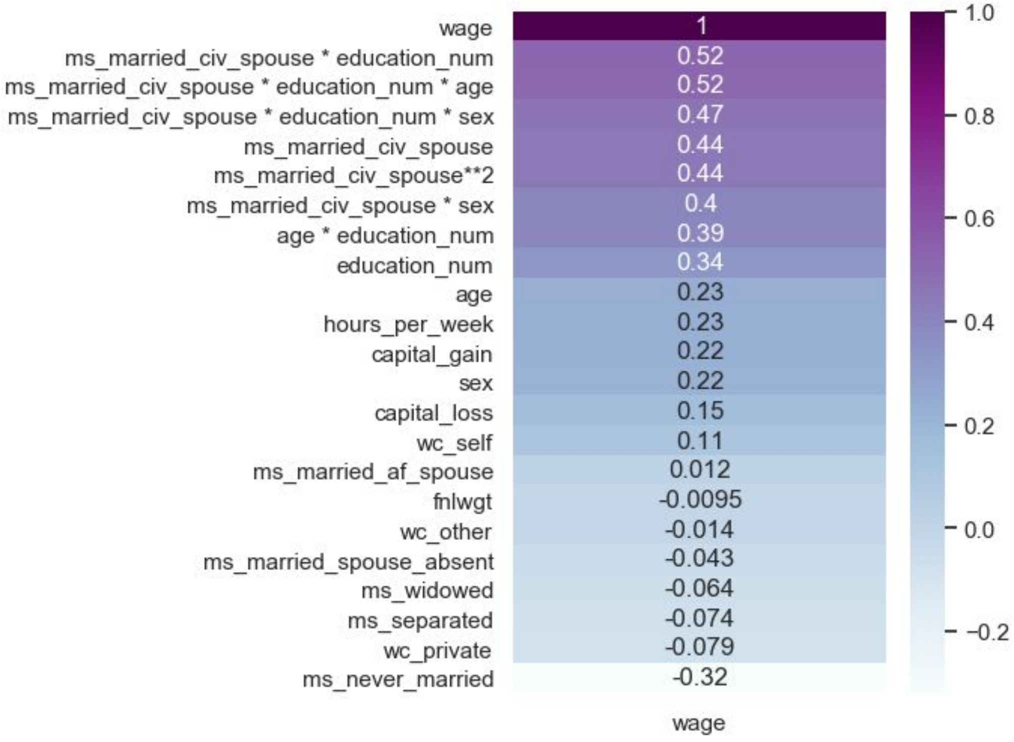
Round 2

1. age
2. education-num
3. marital-status
4. sex
5. workclass
6. hours per week
7. capital gains

	train	test	Mean_CV_Score	f1score_train	f1score_test
SVC_Reem	0.845779	0.844406	0.844858	0.629102	0.612443
LOGREG_Jonna	0.839681	0.839595	0.839461	0.626686	0.617151
KNN_Kemal	0.850298	0.836728	0.837706	0.660362	0.618512
Adaboost_Reem	0.932169	0.82772	0.826957	0.857669	0.617587

Round 3

1. age
2. education-num
3. marital-status
4. sex
5. workclass
6. hours per week
7. capital gains
8. relationship
9. ms_married_civ_spouse^2
10. ms_married_civ_spouse * education_num
11. age * education_num



Round 3

1. age
2. education-num
3. marital-status
4. sex
5. workclass
6. hours per week
7. capital gains
8. relationship
9. ms_married_civ_spouse^2
10. ms_married_civ_spouse * education_num
11. age * education_num

	train	test	Mean_CV_Score	f1score_train	f1score_test
SVC_Reem	0.845384	0.844815	0.844814	0.625664	0.611481
LOGREG_Jonna	0.840383	0.837854	0.840471	0.631632	0.61516
KNN_Kemal	0.849158	0.837752	0.835951	0.655373	0.618348
Adaboost_Reem	0.932257	0.82598	0.825816	0.857696	0.615559

Round 2 - Detour

1. age
2. education_num
3. capital_gain
4. hours_per_week
5. wc_gov
6. wc_no_work
7. ms_married_af_spouse
8. ms_married_civ_spouse
9. ms_married_spouse_abse
10. ms_never_married
11. ms_separated
12. ms_widowed
13. sex_Male

	train	test	Mean_CV_Score	f1score_train	f1score_test
RForest_Kemal	0.845384	0.846248	0.845077	0.60766	0.597966
LOGREG_Jonna	0.83933	0.83939	0.839286	0.625639	0.617783