Predicting Income

Group 3: Cathy Luy, Phuong Trang Tran, Salena Huynh, Brandon Yan, Calvin Zhang, Jennie Nguyen

How can different demographic and socioeconomic factors influence an individual's income?

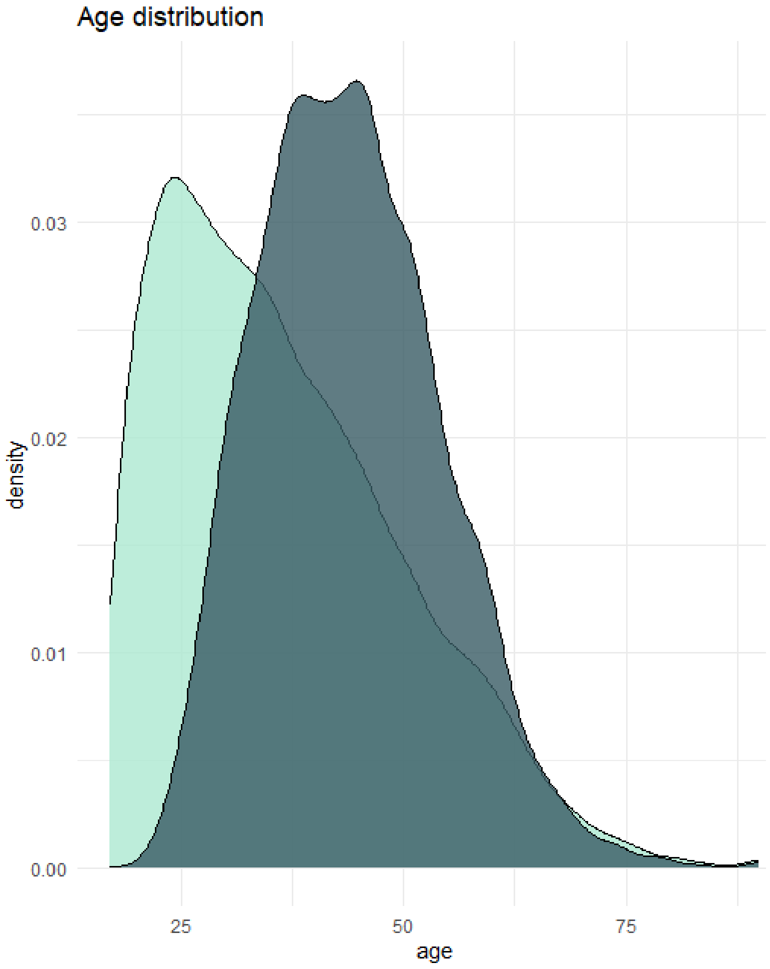
TABLE OF CONTENTS

Summary

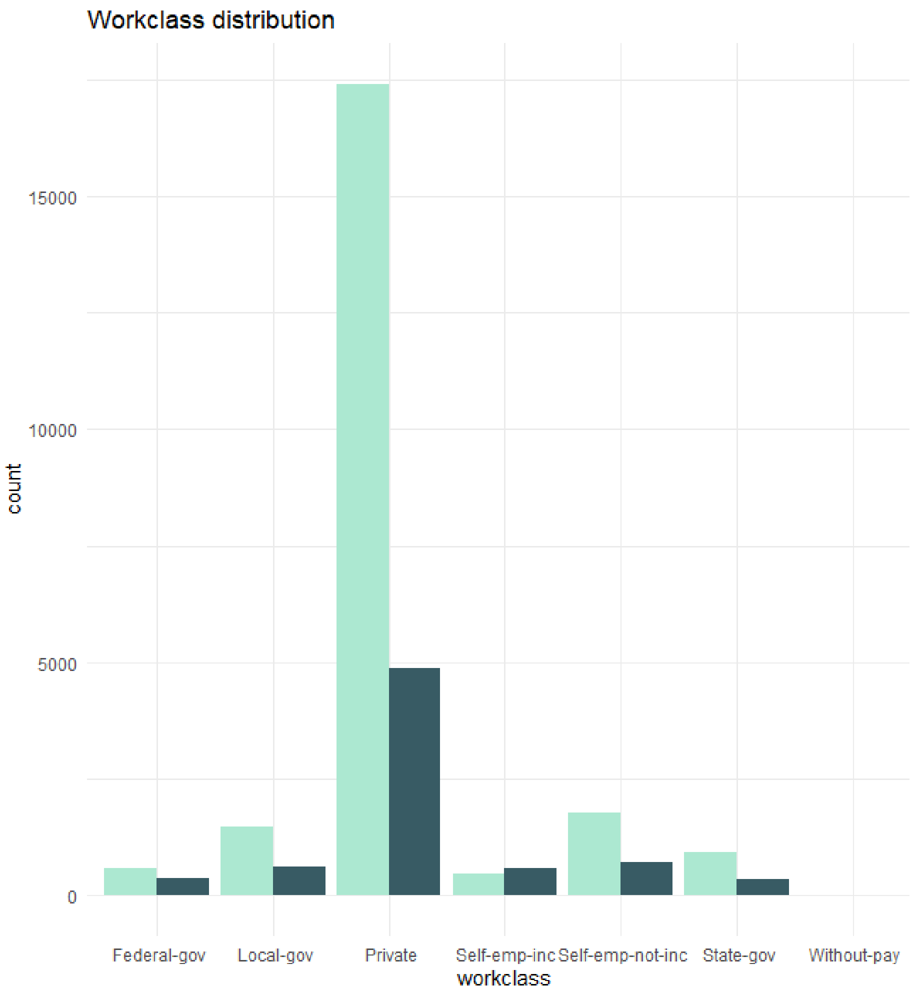
	age	workclass	education	education.num	marital.status	occupation	relationship	race	sex	hours.per.week	native.country	Income
1	39	State-gov	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	40	United-States	<=50K
2	50	Self-emp-not-inc	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	13	United-States	<=50K
3	38	Private	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	40	United-States	<=50K
4	53	Private	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	40	United-States	<=50K
5	28	Private	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	40	Cuba	<=50K
6	37	Private	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	40	United-States	<=50K
7	49	Private	9th	5	Married-spouse-absent	Other-service	Not-in-family	Black	Female	16	Jamaica	<=50K
8	52	Self-emp-not-inc	HS-grad	9	Married-civ-spouse	Exec-managerial	Husband	White	Male	45	United-States	>50K
9	31	Private	Masters	14	Never-married	Prof-specialty	Not-in-family	White	Female	50	United-States	>50K
10	42	Private	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	40	United-States	>50K

```
      age                    workclass            education      education.num                  marital.status              occupation
 Min.   :17.00   Federal-gov     :  943   HS-grad     :9840   Min.   : 1.00   Divorced             : 4214   Prof-specialty :4038
 1st Qu.:28.00   Local-gov       : 2067   Some-college:6678   1st Qu.: 9.00   Married-AF-spouse    :   21   Craft-repair   :4030
 Median :37.00   Private         :22286   Bachelors   :5044   Median :10.00   Married-civ-spouse   :14065   Exec-managerial:3992
 Mean   :38.44   Self-emp-inc    : 1074   Masters     :1627   Mean   :10.12   Married-spouse-absent:  370   Adm-clerical   :3721
 3rd Qu.:47.00   Self-emp-not-inc: 2499   Assoc-voc   :1307   3rd Qu.:13.00   Never-married        : 9726   Sales          :3584
 Max.   :90.00   State-gov       : 1279   11th        :1048   Max.   :16.00   Separated            :  939   Other-service  :3212
                 Without-pay     :   14   (Other)     :4618                   Widowed              :  827   (Other)        :7585
         relationship                    race            sex          hours.per.week        native.country     Income
  Husband        :12463   Amer-Indian-Eskimo:  286   Female: 9782   Min.   : 1.00   United-States:27504   <=50K:22654
  Not-in-family  : 7726   Asian-Pac-Islander:  895   Male  :20380   1st Qu.:40.00   Mexico       :  610   >50K : 7508
  Other-relative :  889   Black             : 2817                  Median :40.00   Philippines  :  188
  Own-child      : 4466   Other             :  231                  Mean   :40.93   Germany      :  128
  Unmarried      : 3212   White             :25933                  3rd Qu.:45.00   Puerto-Rico  :  109
  Wife           : 1406                                             Max.   :99.00   Canada       :  107
                                                                                   (Other)       : 1516
```

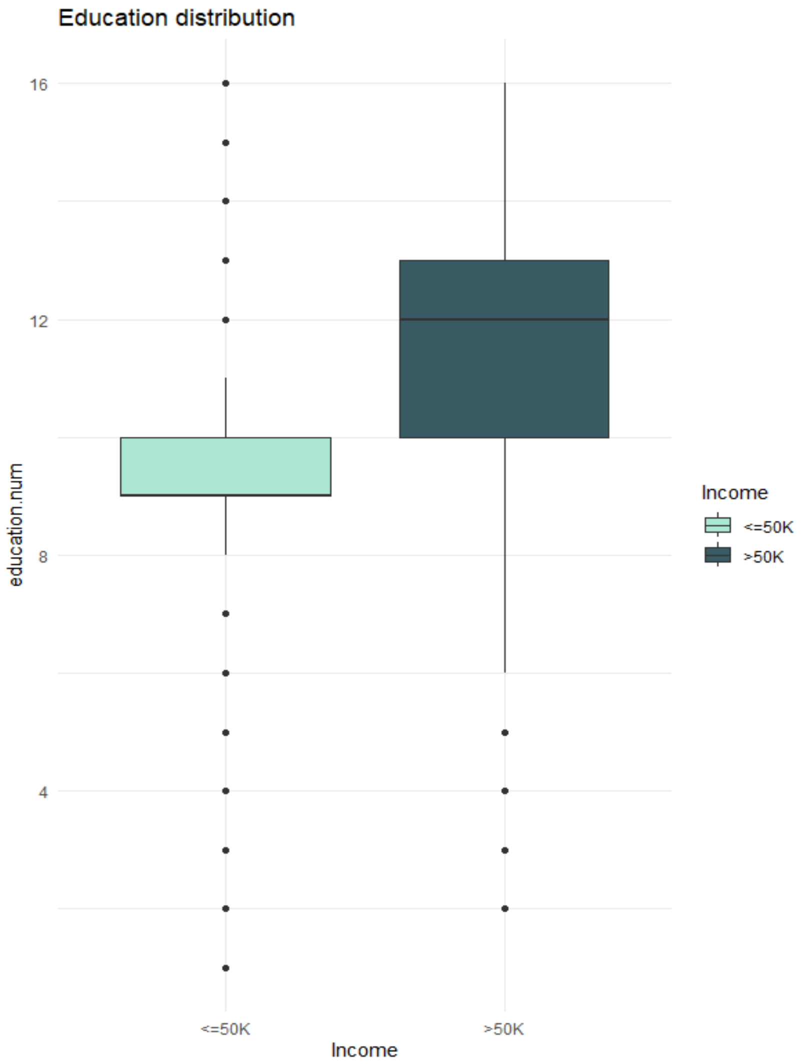
Summary



Age distribution
density
0.00
0.01
0.02
0.03
25
50
75
age



Workclass distribution
count
0
5000
10000
15000
Federal-gov
Local-gov
Private
Self-emp-inc
Self-emp-not-inc
State-gov
Without-pay
workclass



Education distribution
education.num
4
8
12
16
<=50K
>50K
Income
Income
<=50K
>50K

Logistic Regression

glm.fits=glm(Income~age+race+sex,data=data,family=binomial)

```
Call:
glm(formula = Income ~ age + race + sex, family = binomial, data = data)

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)    
(Intercept)            -4.499147   0.194968 -23.076  < 2e-16 ***
age                     0.041907   0.001084  38.657  < 2e-16 ***
race Asian-Pac-Islander 0.996292   0.203434   4.897 9.71e-07 ***
race Black              0.177631   0.196261   0.905    0.365    
race Other             -0.206242   0.300408  -0.687    0.492    
race White              0.863006   0.188009   4.590 4.43e-06 ***
sex Male                1.212370   0.036295  33.404  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 33851  on 30161  degrees of freedom
Residual deviance: 30485  on 30155  degrees of freedom
AIC: 30499

Number of Fisher Scoring iterations: 5
```

```
Call:
glm(formula = Income ~ age + race + sex, family = binomial, data = train)

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)    
(Intercept)             -4.509656   0.227008 -19.866  < 2e-16 ***
age                      0.043056   0.001259  34.209  < 2e-16 ***
race Asian-Pac-Islander  0.984789   0.236545   4.163 3.14e-05 ***
race Black               0.067566   0.229062   0.295 0.768020    
race Other              -0.296329   0.353763  -0.838 0.402228    
race White               0.825242   0.219210   3.765 0.000167 ***
sex Male                 1.218251   0.042073  28.955  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 25386  on 22620  degrees of freedom
Residual deviance: 22769  on 22614  degrees of freedom
AIC: 22783

Number of Fisher Scoring iterations: 5
```

```
        Income.num
glm.pred     0     1
       0 21244  7032
       1  1410   476
```

Accuracy: 72.01%

KNN Model:

```
train_features <- train[, c("age", "hours.per.week", "education.num")]
test_features <- test[, c("age", "hours.per.week", "education.num")]
```

Normalization

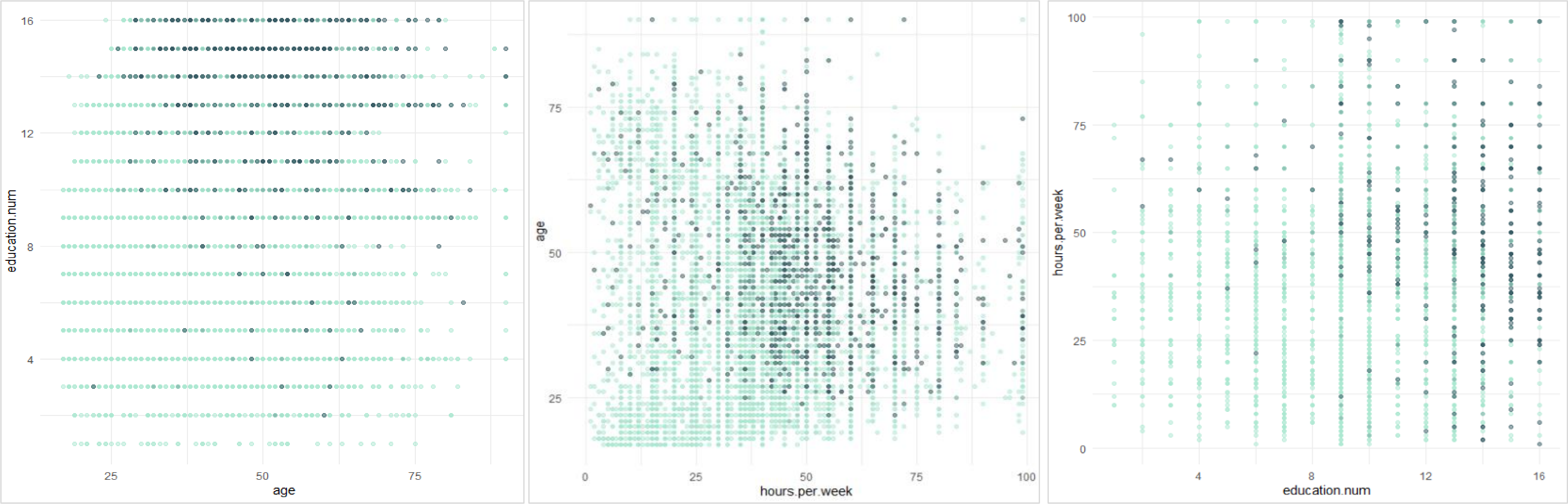
```
head(train_features)

##    age hours.per.week education.num
## 2   50             13            13
## 3   38             40             9
## 4   53             40             7
## 5   28             40            13
## 7   49             16             5
## 9   31             50            14

head(test_features)

##     age hours.per.week education.num
## 1    39             40            13
## 6    37             40            14
## 8    52             45             9
## 12   30             40            13
## 26   56             40            13
## 31   23             52            12
```

```
head(train_normalized)

##         age hours.per.week education.num
## 1 0.4520548      0.1224490     0.8000000
## 2 0.2876712      0.3979592     0.5333333
## 3 0.4931507      0.3979592     0.4000000
## 4 0.1506849      0.3979592     0.8000000
## 5 0.4383562      0.1530612     0.2666667
## 6 0.1917808      0.5000000     0.8666667

head(test_normalized)

##          age hours.per.week education.num
## 1 0.30136986      0.3979592     0.8000000
## 2 0.27397260      0.3979592     0.8666667
## 3 0.47945205      0.4489796     0.5333333
## 4 0.17808219      0.3979592     0.8000000
## 5 0.53424658      0.3979592     0.8000000
## 6 0.08219178      0.5204082     0.7333333
```

```
         Actual
Predicted    0    1
        0 5145 1137
        1  497  761
```

Accuracy: 78.33%

Which features are most predictive of higher income levels?



Tree Model

```
Call:
rpart(formula = Income.num ~ age + workclass + education + marital.status +
    occupation + relationship + hours.per.week + native.country +
    race + sex, data = train, method = "class")
  n= 24420

          CP nsplit rel error    xerror       xstd
1 0.12546721      0 1.0000000 1.0000000 0.01135538
2 0.01342168      2 0.7490656 0.7490656 0.01021201
3 0.01000000      5 0.7088005 0.7327557 0.01012442

Variable importance
  relationship marital.status      education     occupation            sex            age hours.per.week
            29             28             12             11              9              8              3
```

1 Relationship and Marital Status

2 Education

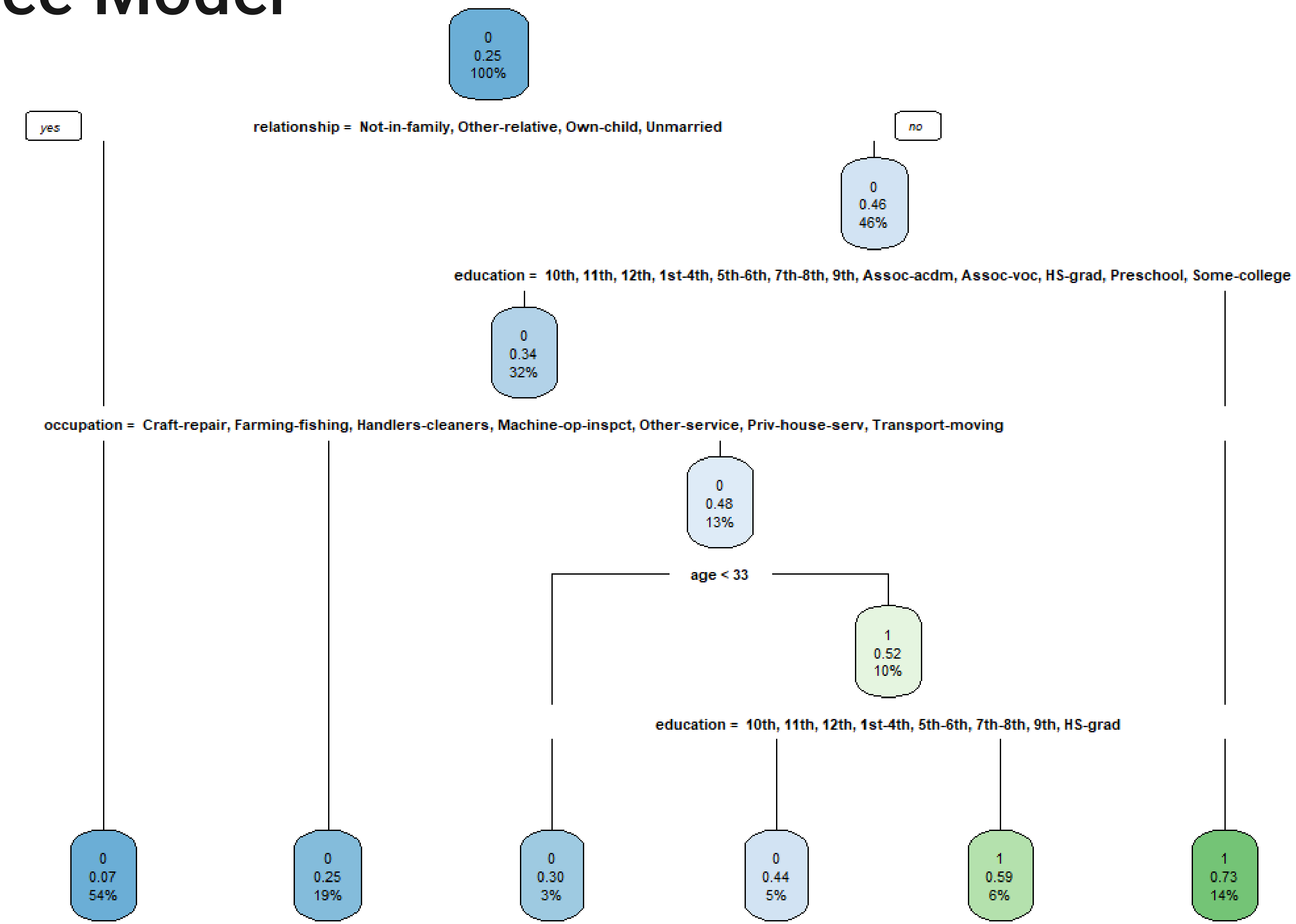
3 Occupation

4 Sex

5 Age

6 Hours per week

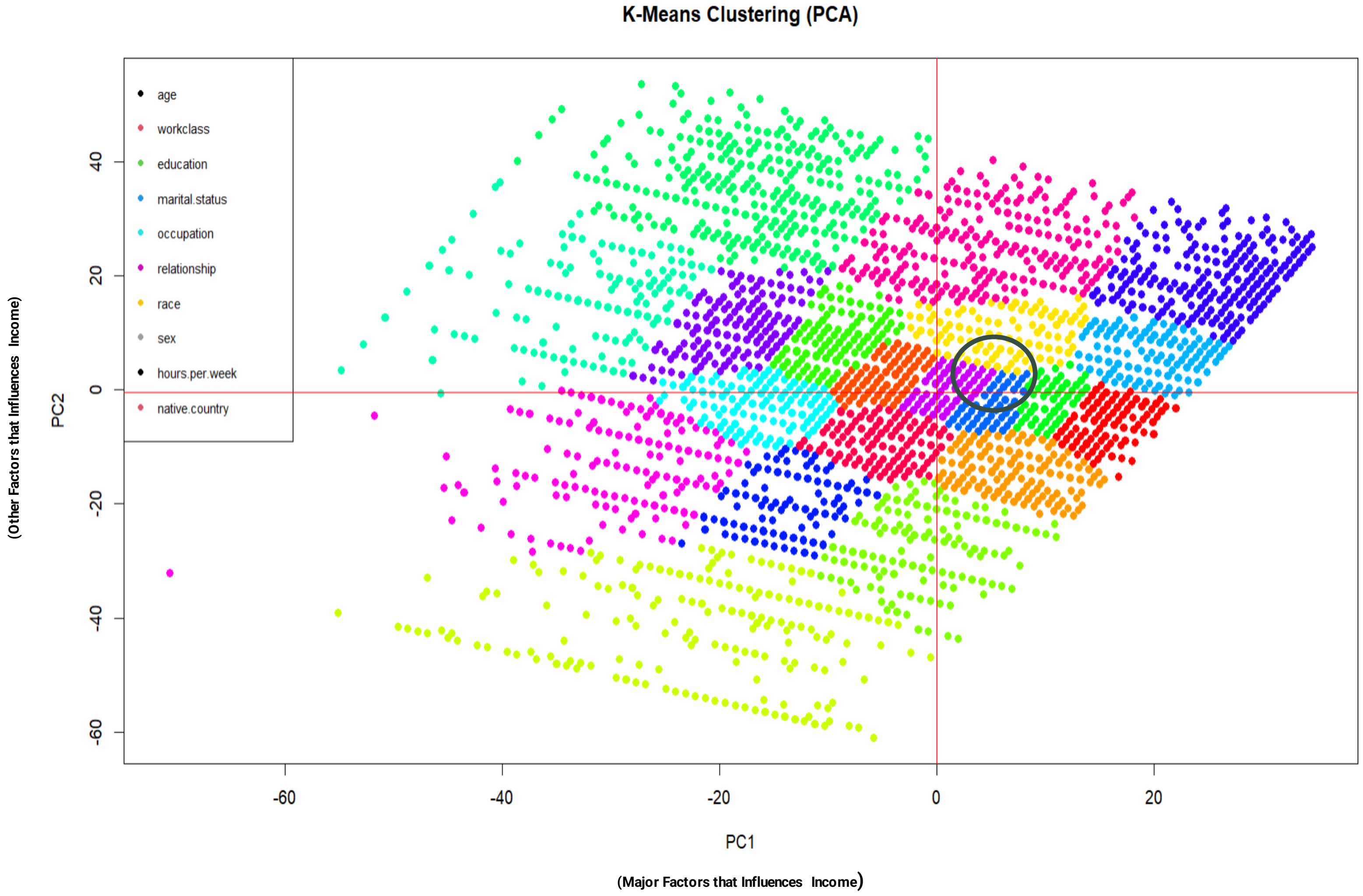
Tree Model



Key Insights

Root Node – Relationship
Education
Age
Occupation

Clustering



Top variables for PC1 (Accounts for 56.22% of variability):

1. Age
2. Hours per week
3. **Marital status:** Never-married, Married-civ-spouse
4. **Relationship:** Own-child
5. **Workclass:** Private, Self-emp-not-inc
6. **Sex:** Male
7. **Occupation:** Exec-managerial, Other-service

Top variables for PC2 (Accounts for 42.68% of variability):

1. Hours per week
2. Age
3. **Sex:** Male
4. **Occupation:** Other-service, Exec-managerial
5. **Marital status:** Married-civ-spouse, Widowed
6. **Education:** Bachelors
7. **Relationship:** Own-child, Unmarried

PCA Normalization Process

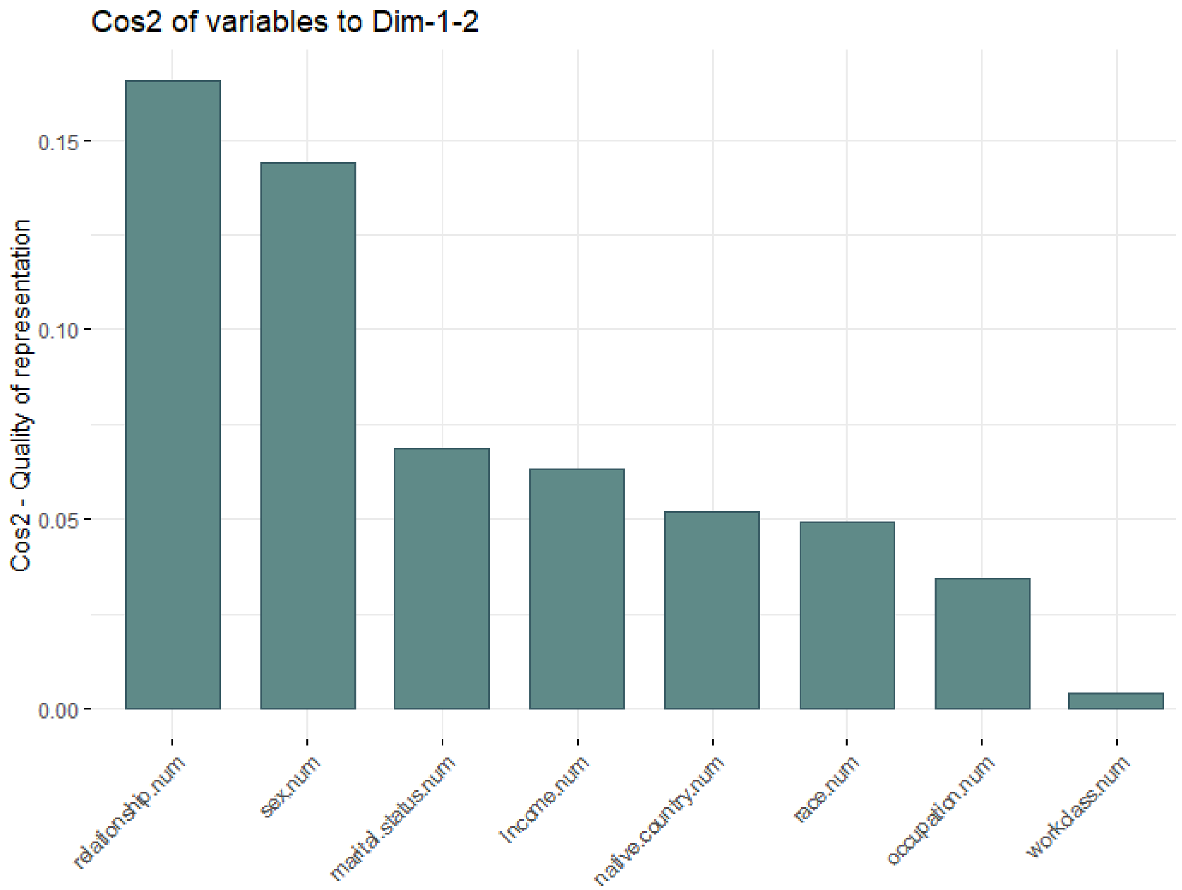
	workclass.num	marital.status.num	occupation.num	relationship.num
1	2.9359517	0.9478313	-1.4790301	-0.2612446
2	1.8876507	-0.3872683	-0.7345332	-0.8857223
3	-0.2089512	-1.7223678	-0.2382018	-0.2612446
4	-0.2089512	-0.3872683	-0.2382018	-0.8857223
5	-0.2089512	-0.3872683	0.7544608	2.2366662
6	-0.2089512	-0.3872683	-0.7345332	2.2366662

	race.num	sex.num	native.country.num	Income.num
1	0.3850415	0.6927947	0.2649196	-0.5756818
2	0.3850415	0.6927947	0.2649196	-0.5756818
3	0.3850415	0.6927947	0.2649196	-0.5756818
4	-2.0110019	0.6927947	0.2649196	-0.5756818
5	-2.0110019	-1.4433813	-5.3039463	-0.5756818
6	0.3850415	-1.4433813	0.2649196	-0.5756818

Importance of components:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
Standard deviation	0.6566579	0.3868625	0.3498713	0.3345691	0.30835196
Proportion of Variance	0.4128897	0.1433075	0.1172121	0.1071834	0.09104355
Cumulative Proportion	0.4128897	0.5561972	0.6734093	0.7805927	0.87163625

	Comp.6	Comp.7	Comp.8
Standard deviation	0.30385914	0.20426880	2.967518e-09
Proportion of Variance	0.08840979	0.03995396	8.432229e-18
Cumulative Proportion	0.96004604	1.00000000	1.000000e+00



THANK YOU