Abstract

The census data obtained from UCI Machine Learning Repository is analyzed for pattern recognition to predict the income falling into one of two classes – '<= $50K' or > '$50K'. A model ensembling technique is used to combine predictions from Generalized Linear Model, Random Forests and Stochastic Gradient Boosting Model to achieve an accuracy of 86.5 % on the test set scaling up the accuracy rates of these individual prediction models.

Aim

The goal of this project is to predict whether income exceeds $50K/yr based on census data downloaded from UCI Machine Learning Repository.

[Complete information regarding the dataset can be found at http://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.names]

Overview of dataset

Below I examine the descriptive statistics for all the numeric variables included in the complete dataset in Table 1.

Table 1: Overview of all numeric variables in the data

	Observations	Mean	Standard deviation	Median	Min	Max
age	45222	38.55	13.22	37	17	90
fnlwgt	45222	189734.73	105639.2	178316	13492	1490400
educationnum	45222	10.12	2.55	10	1	16
capitalgain	45222	1101.43	7506.43	0	0	99999
capitalloss	45222	88.6	404.96	0	0	4356
hoursperweek	45222	40.94	12.01	40	1	99

For Income class <= 50 K

	Observations	Mean	Standard deviation	Median	Min	Max
age	34014	36.75	13.56	34	17	90
fnlwgt	34014	190175.21	106653.66	178952.5	13492	1490400
educationnum	34014	9.63	2.42	9	1	16
capitalgain	34014	149.02	927.45	0	0	41310
capitalloss	34014	54.03	312.22	0	0	4356
hoursperweek	34014	39.37	11.97	40	1	99

For income class > 50 K

	Observations	Mean	Standard deviation	Median	Min	Max
age	11208	44.01	10.34	43	19	90
fnlwgt	11208	188397.97	102492.12	176775.5	13769	1226583
educationnum	11208	11.6	2.37	12	1	16
capitalgain	11208	3991.79	14616.54	0	0	99999
capitalloss	11208	193.49	592.64	0	0	3683
hoursperweek	11208	45.69	10.8	40	1	99

As can be seen in Table 1, the median age in the dataset is 37 years with standard deviation of 13.22. So, most of the subjects included in the dataset will be eligible as working professionals. Subjects included in the dataset have varying levels of educational qualifications. The median number of hours subjects work in a week are 40 with standard deviation of 12. Age can be considered as one of the important factors determining the classification of a subject into a specific income class. The median age for subjects falling in higher income class is 43 whereas it is 34 for subjects falling in lower income class. The 2 variables "capital-gain" and "capital-loss" look highly variable within both income classes.

Table 2 presents the percentage distribution of subjects by different workclasses such as working for a private company, federal government, state government, self employed etc. within each income class.

Table 2: Percentage distribution within different workclasses in both income classes

	<=50K	>50K
Federal-gov	2.51955077	4.89828694
Local-gov	6.42382548	8.16381156
Private	76.6037514	64.69486081
Self-emp-inc	2.15793497	8.13704497
Self-emp-not-inc	8.04668666	9.44860814
State-gov	4.19239137	4.63954318
Without-pay	0.05585935	0.0178444

As can be seen in Table 2, within both income classes, most of the subjects are employed in private companies; 76% for less than 50K category and 65% for above 50K category.

Table 3 represents percentage distribution within different education levels of subjects within both income classes.

Table 3: Percentage distribution within different education levels in both income classes

	<=50K	>50K
Highly-qualified	16.63433	44.34333
Less-qualified	83.36567	55.65667

To view the percentage distribution of subjects in both income class by education levels, I create 2 categories – "Highly-qualified" and "Less-qualified". "Highly-qualified" are those who have obtained either Bachelors or Masters or Doctorate; all others are classified as "Less-qualified".

As can be seen from Table 3, within <=50K category, 16.63% of subjects have an academic degree like Bachelors or Masters or Doctorate. In contrary, 44.34% of subjects earning more than 50K income have academic degrees (Bachelors or Masters or Doctorate). This shows education level as an important predictor for earnings and hence, classification of a subject in either <= 50K or > 50K category.

Table 4 represents percentage distribution of subjects by marital status within both income classes.

Table 4: Percentage distribution by marital status in both income classes

	<=50K	>50K
Divorced	16.58728759	5.84403997
Married-AF-spouse	0.05291939	0.12491078
Married-civ-spouse	33.78314812	85.33190578
Married-spouse-absent	1.46410302	0.48179872
Never-married	40.85670606	6.2544611
Separated	3.85723526	0.88329764
Widowed	3.39860058	1.07958601

As can be seen from Table 4, there is a significant difference in marital status of subjects within both income classes. 16% of subjects falling under lower income class are divorced whereas 5.84 % of subjects falling under higher income class got divorced. The major difference can be seen for subjects recognized as Married-civ-spouse. 85.3 % of subjects having income are married-civ-spouse whereas this number is quite low for lower income class, 33.7 %. Another major difference can be seen in Never-married subjects. 40.8 % of subjects having lower income were never married whereas only 6.2 % of subjects with higher income never got married.

I classify the subjects within both income classes by their occupations. The 2 occupations - "Exec-managerial" and "Prof-specialty" have been categorized as "executive-skilled" occupations while all others have been classified as “others”. Table 5 represents percentage distribution of subjects by occupation within both income classes.

Table 5: Percentage distribution by occupation in both income classes

	<=50K	>50K
executive-skilled	18.87752	49.70557
others	81.12248	50.29443

As can be seen from Table 5, we can see a significant difference in type of occupation between higher income and lower income subjects. 49.7 % of subjects in higher income class are skilled executives whereas this number is only 18.9 % for lower income class, which explains the classification of subjects into higher income or lower income class.

Table 6 represents percentage distribution of subjects by relationship within both income classes.

Table 6: Percentage distribution by occupation in both income classes

	<=50K	>50K
Husband	29.8671135	75.901142
Not-in-family	30.7932028	10.9564597
Other-relative	3.8190157	0.4461099
Own-child	19.1715176	0.9368308
Unmarried	13.188687	2.6945039
Wife	3.1604633	9.0649536

A significant difference can be observed in relationship status of subjects in both income classes. 29.86 & of subjects in lower income class are husbands in contrary to 75.90 % of subjects in higher income class. Similarly, the percentage of subjects as wives is also higher in higher income class (9.06 %) in comparison to 3.16 % in lower income class.

Table 7 represents percentage distribution of subjects by race within both income classes.

Table 7: Percentage distribution by race in both income classes

	<=50K	>50K
Amer-Indian-Eskimo	1.123067	0.4728765
Asian-Pac-Islander	2.7459281	3.2922912
Black	10.860234	4.764454
Other	0.9055095	0.4014989
White	84.3652614	91.0688794

As can be seen from Table 7, most of the subjects in both income classes belong to white race; 84.36 % within lower income class and 91.07 % within higher income class. So, no definite conclusion can be derived regarding to role of race in predicting the income class of a subject.

Table 8 represents percentage distribution of subjects by sex within both income classes.

Table 8: Percentage distribution by sex in both income classes

	<=50K	>50K
Female	38.296	14.89115
Male	61.704	85.10885

As can be observed from Table 8, most of the subjects (85.1 %) in higher income class are males whereas this difference in number of males and females is not that huge in lower income class where 38.29 % are females and 61.70 % are males.

Since most of the subjects in the dataset have their native country as "United States", I create 2 groups for this variable as "US" and "Other" and try to observe any difference, if any, between the percentage distribution of subjects by native country within both income classes.

Table 9: Percentage distribution by native country in both income classes

	<=50K	>50K
Other	9.319692	6.780871
US	90.68031	93.21913

There isn't much variation observed in native country within both income classes since most of the subjects in both classes have their native country as United States. So, this variable doesn't seem to be an effective predictor of the target variable.

Exploratory Data Analysis

Figure 1: Density plot distribution of “age” in both income classes

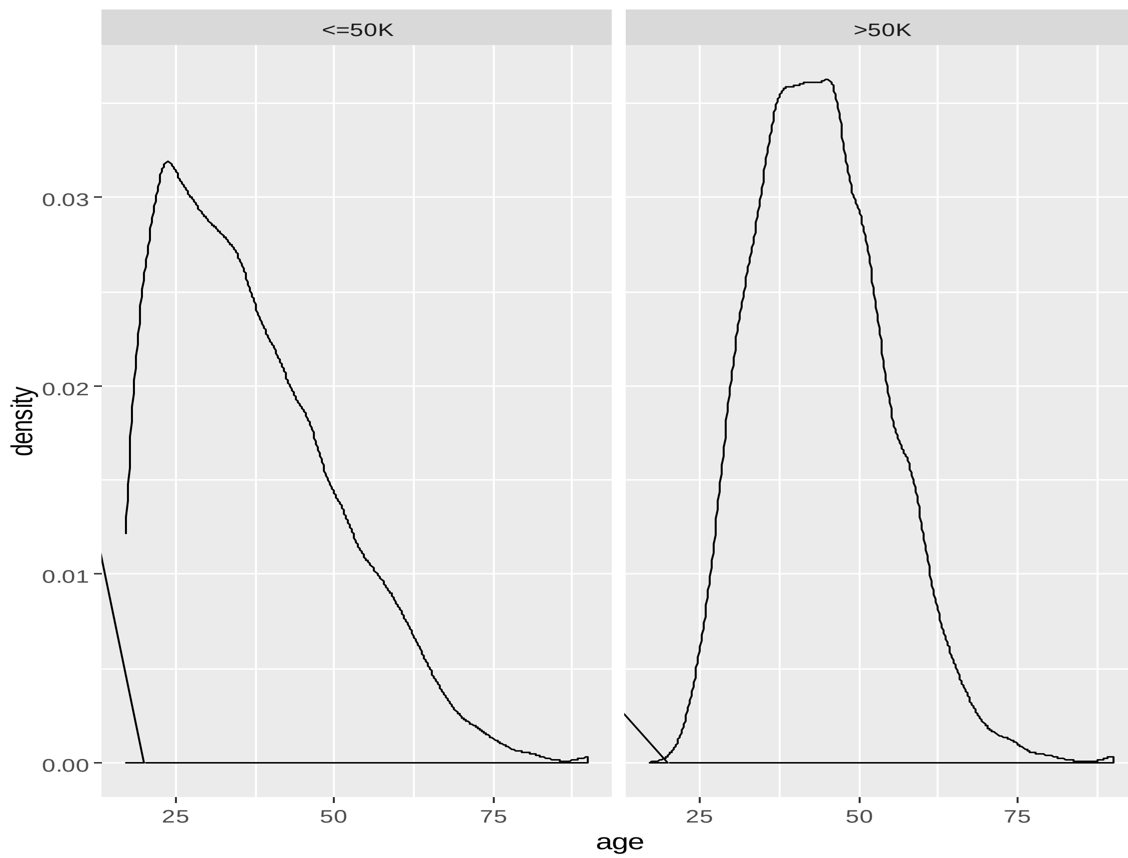


Figure 2: Density plot distribution of "fnlwgt" in both income classes

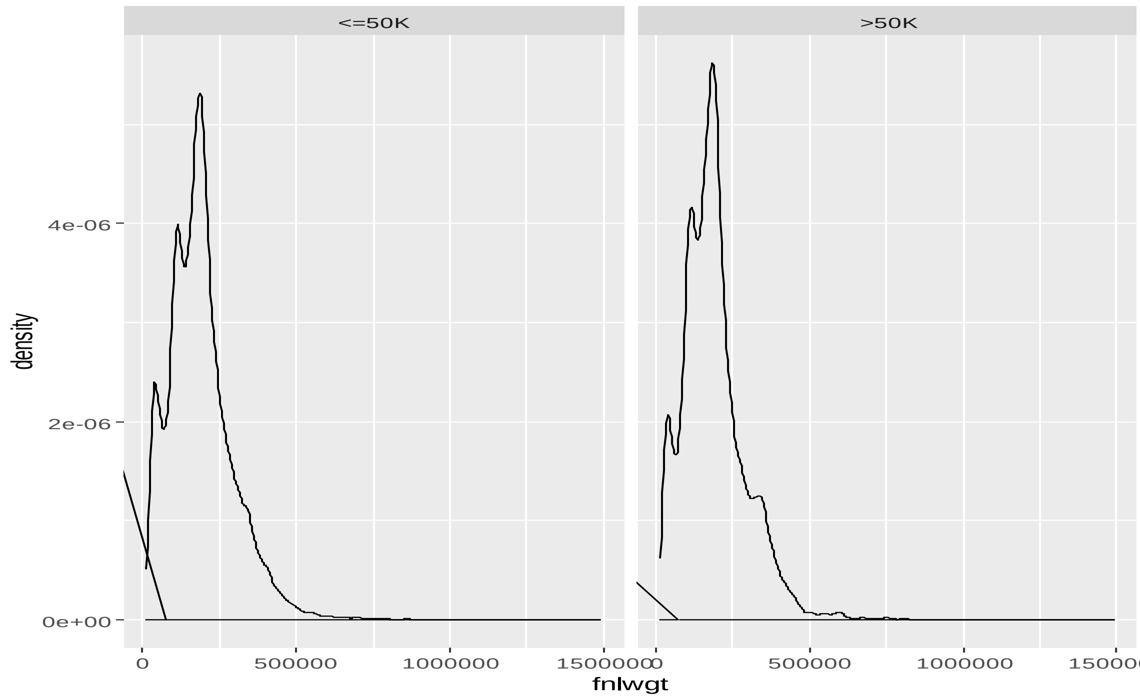


Figure 3: Density plot distribution of "educationnum" in both income classes

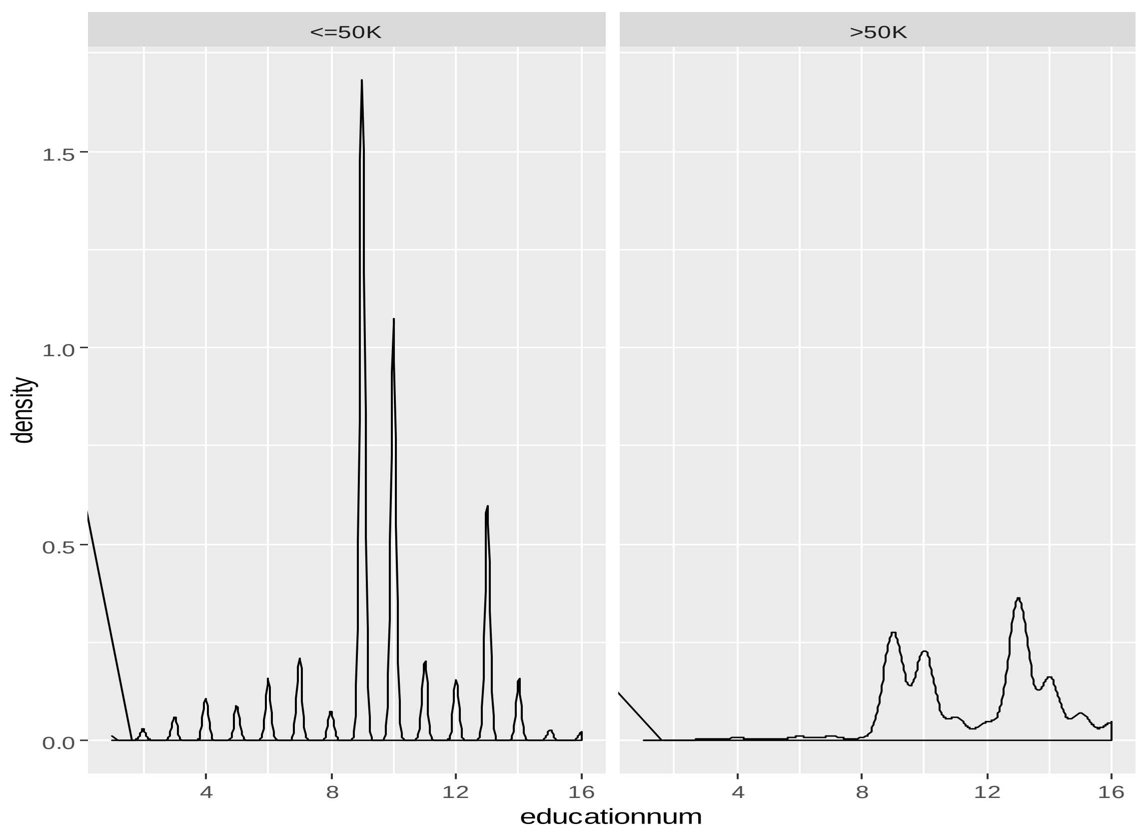


Figure 4: Histograms showing "capitalgain" variation in both income classes

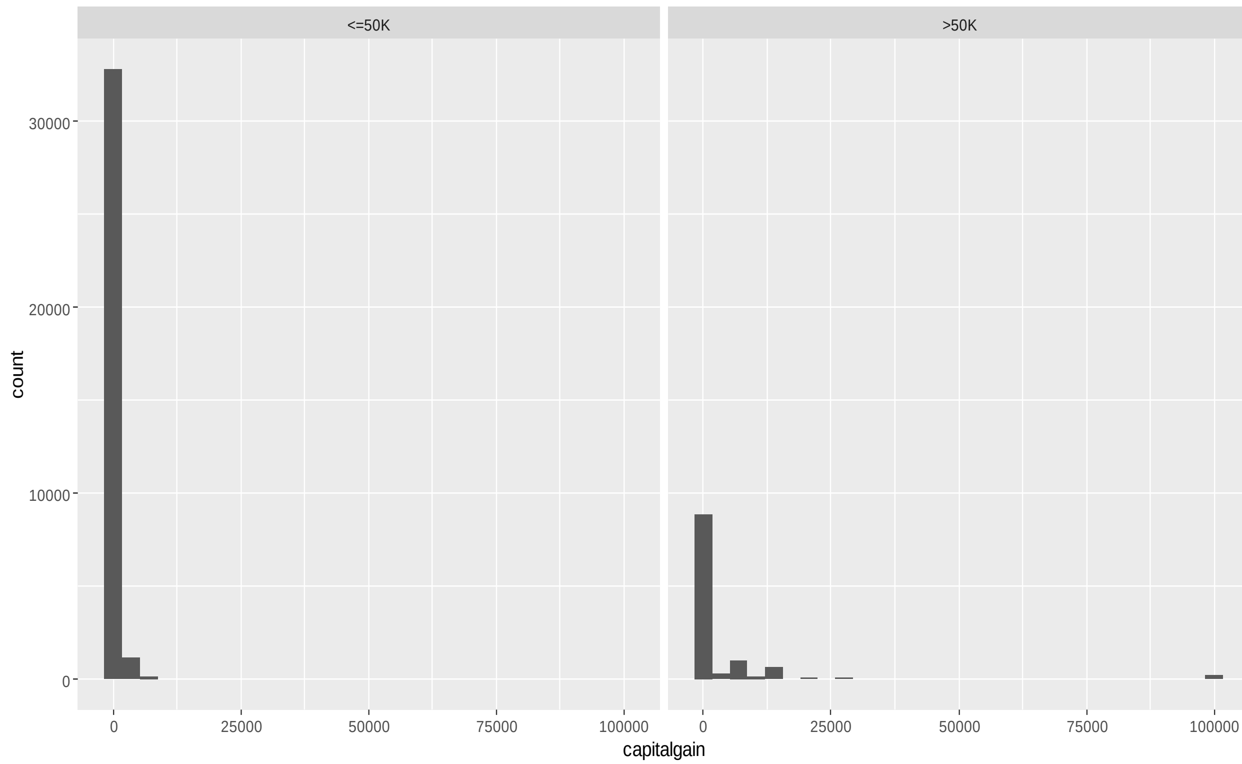


Figure 5: Histograms showing "capitalloss" variation in both income classes

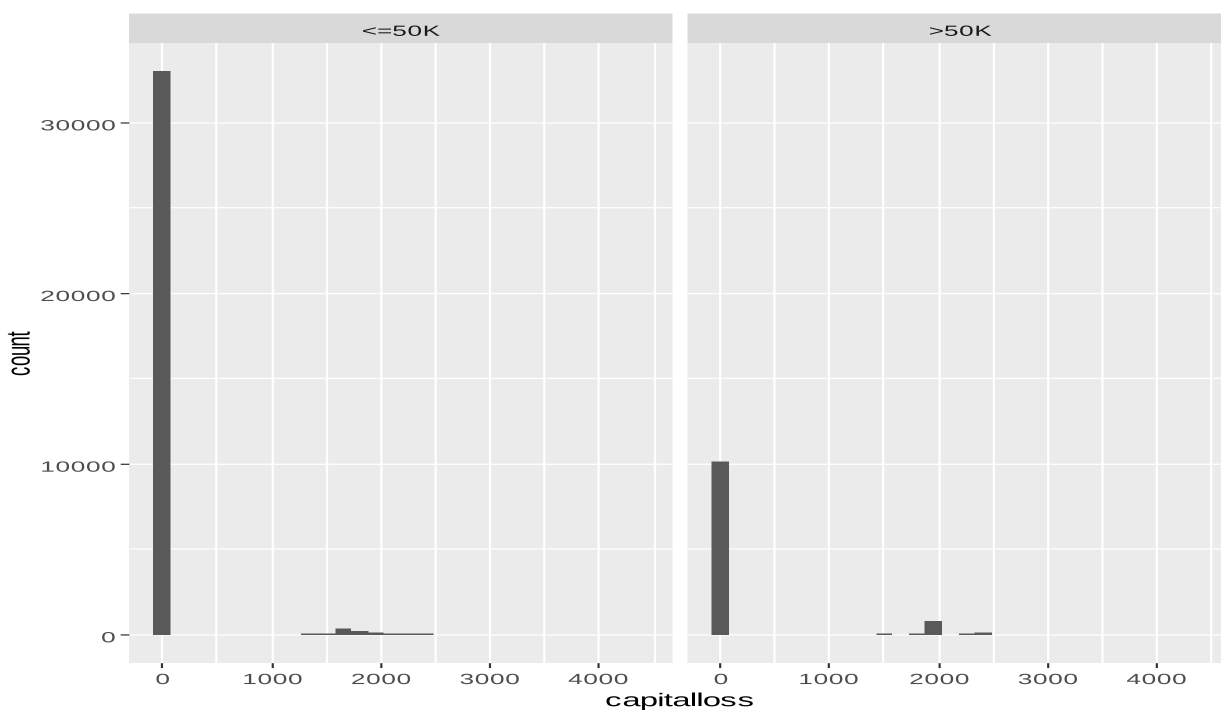


Figure 6: Density plot distribution of “hoursperweek” in both income classes

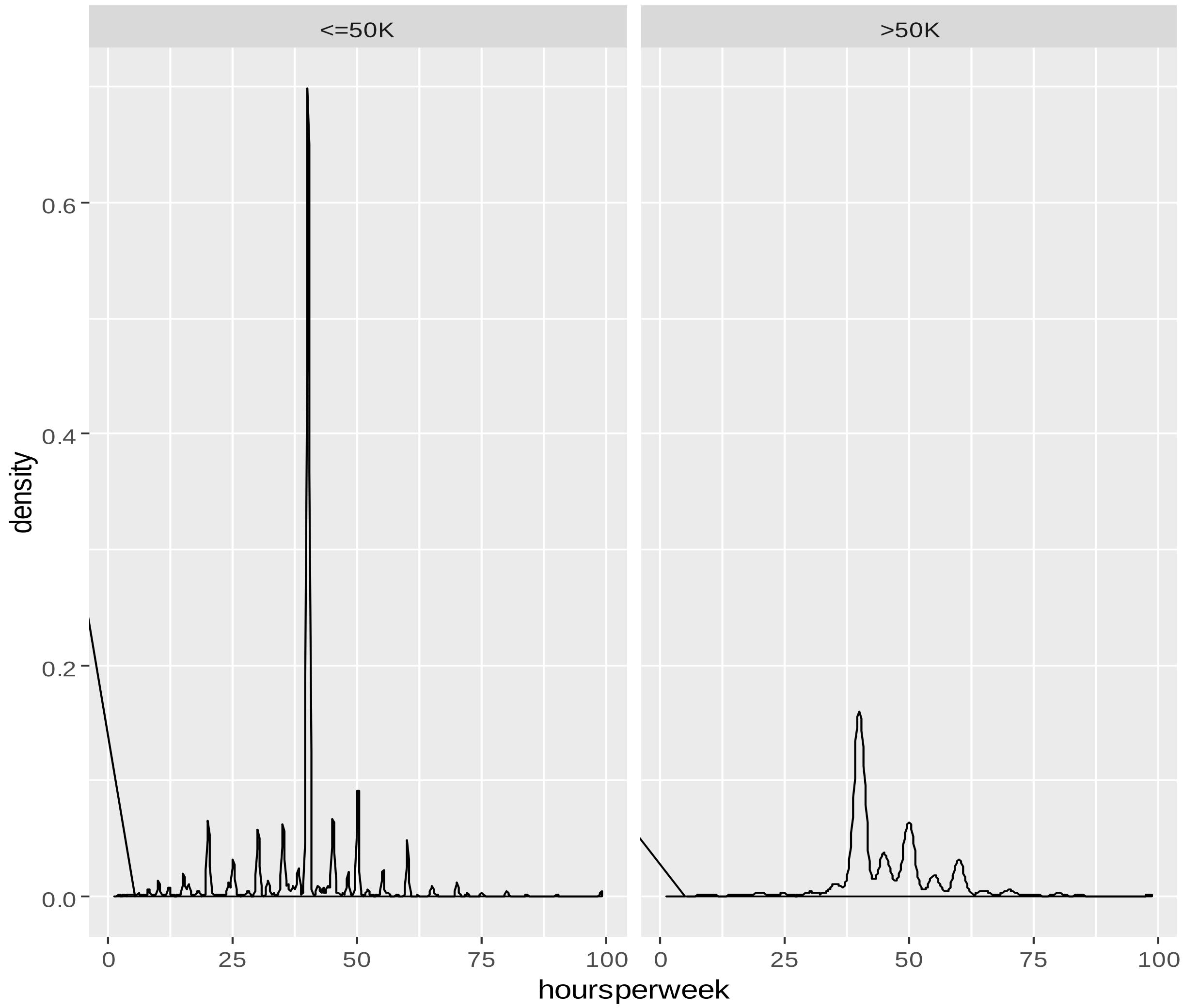


Figure 7: Distribution by "workclass" in both income classes

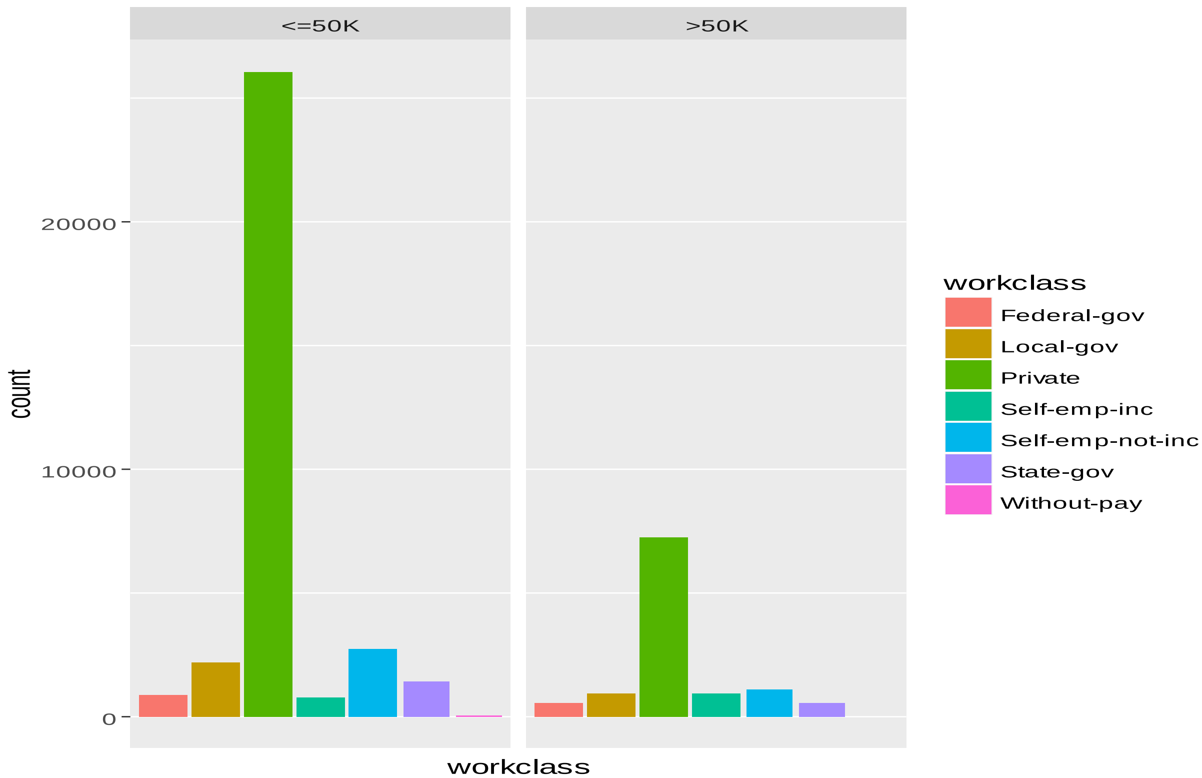


Figure 8: Distribution by "education" in both income classes

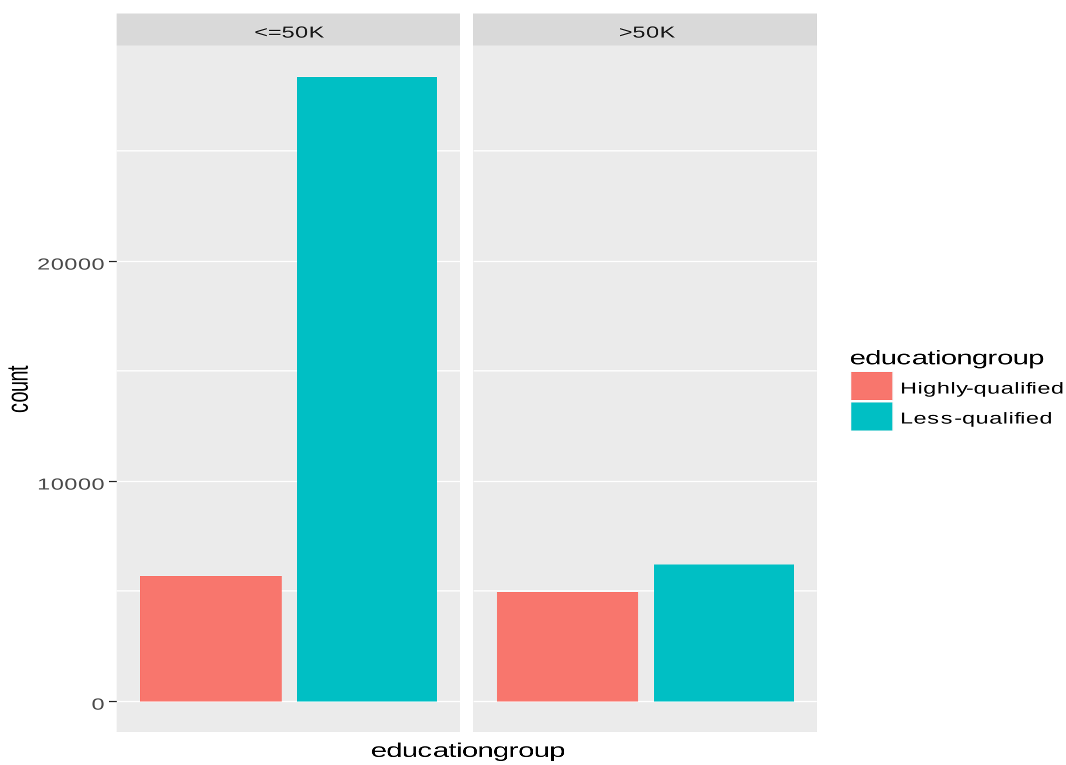


Figure 9: Distribution by "maritalstatus" in both income classes

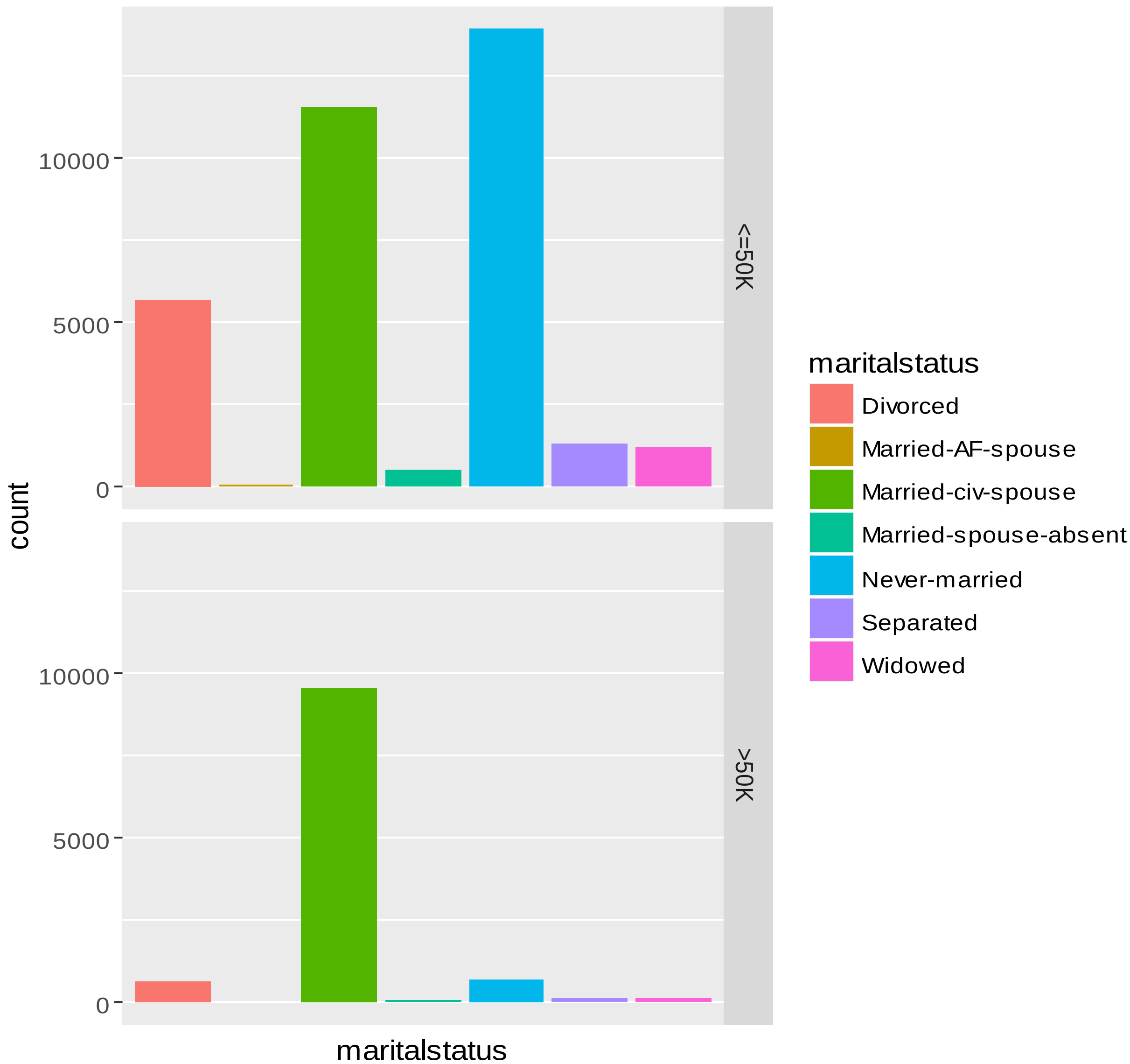


Figure 10: Distribution by “occupation” in both income classes

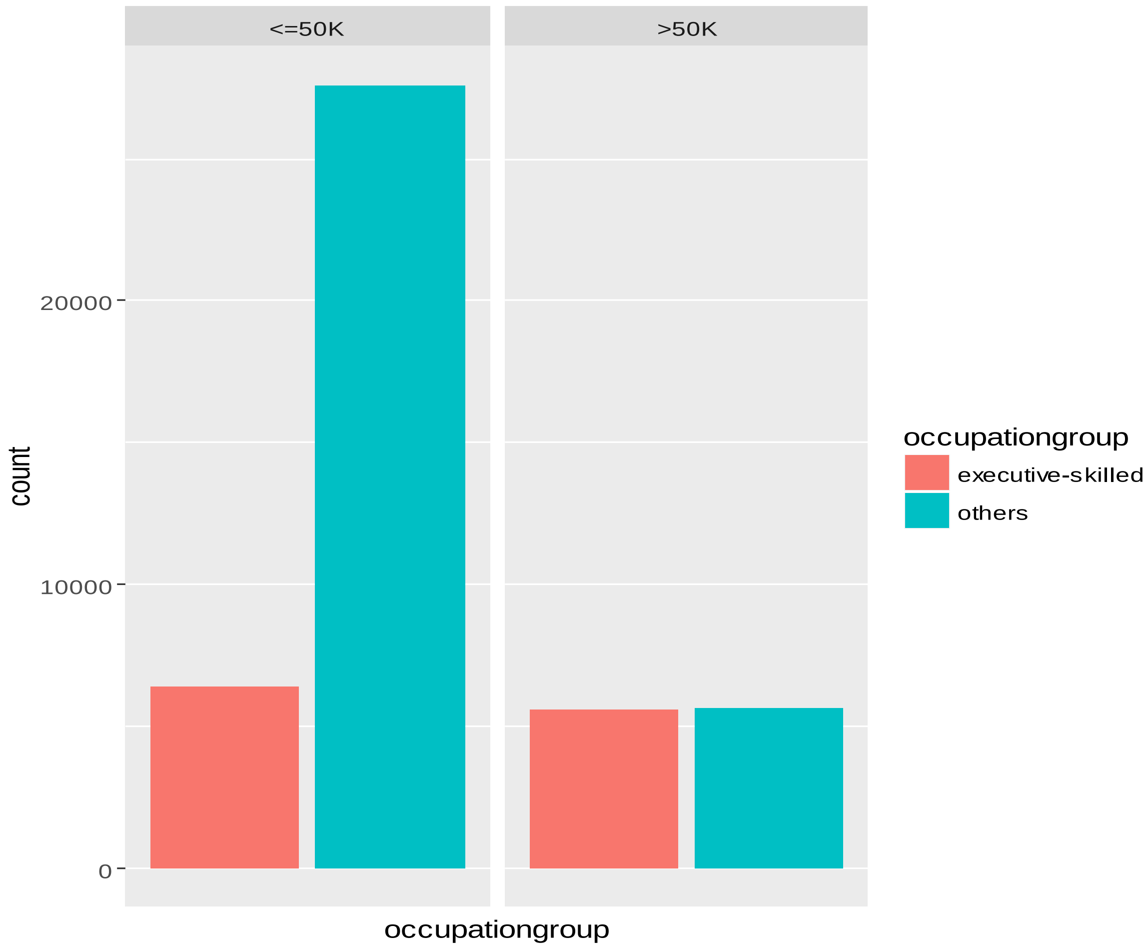


Figure 11: Distribution by "relationship" in both income classes

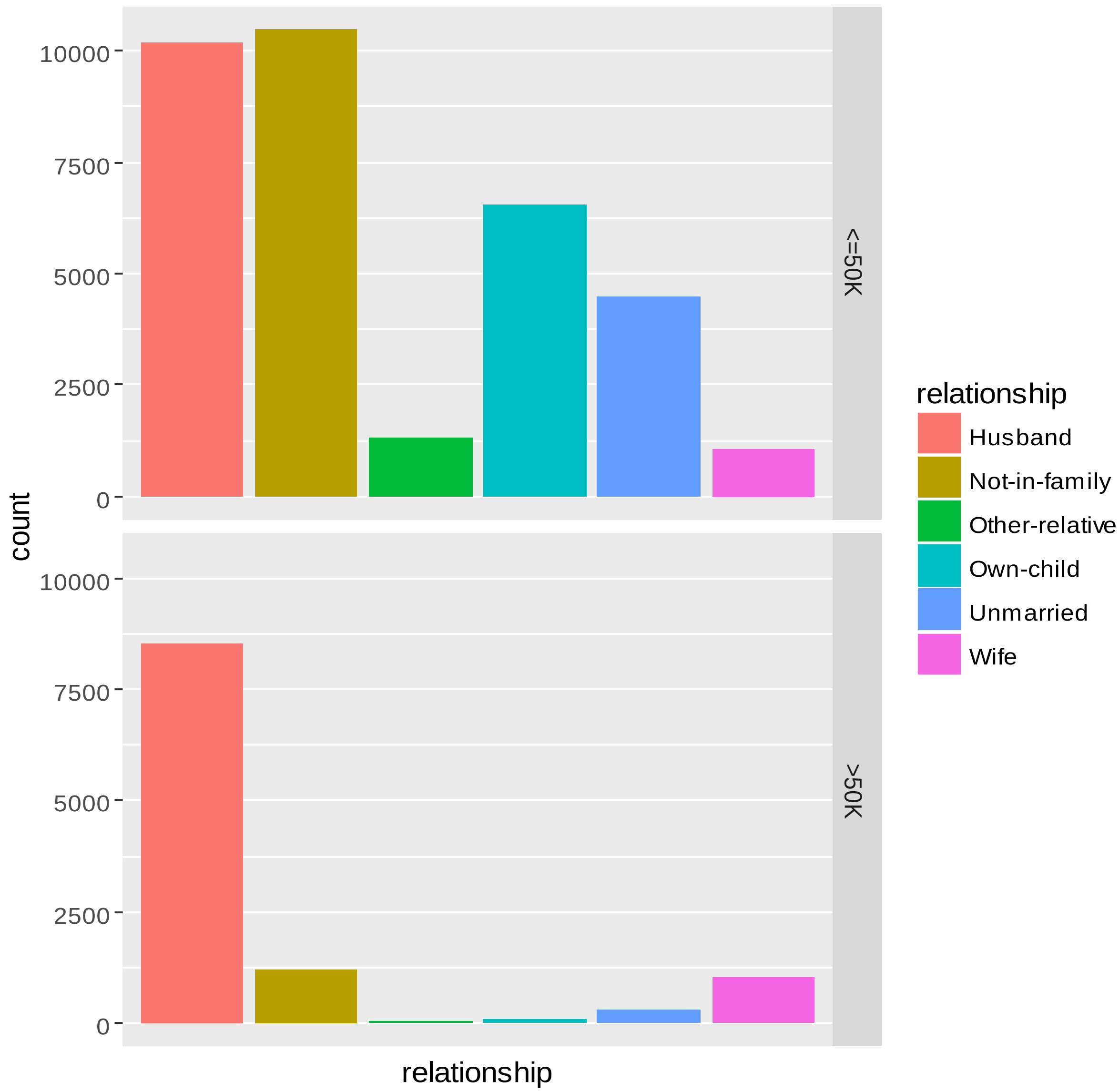


Figure 12: Distribution by “race” in both income classes

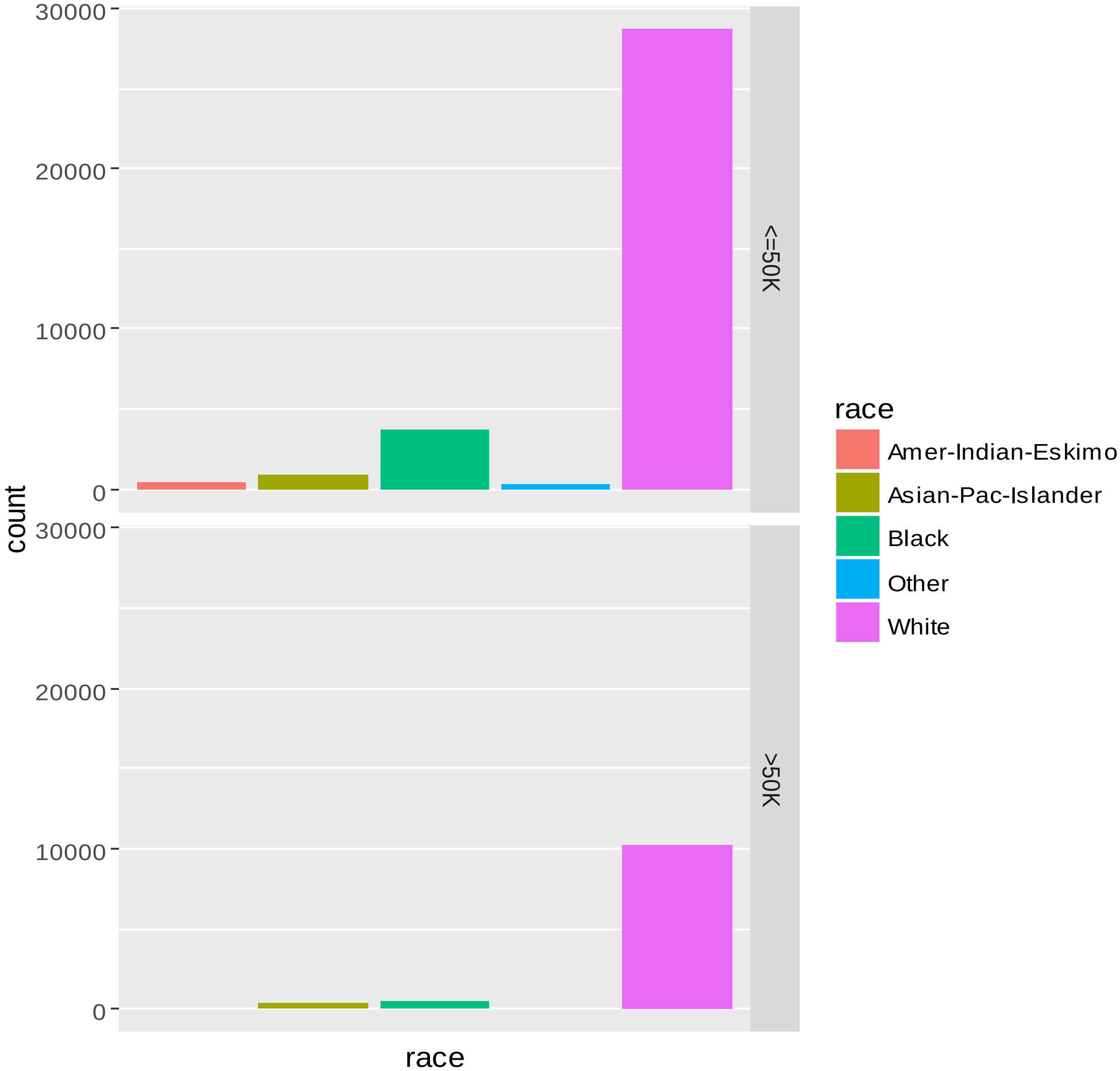


Figure 13: Distribution by "sex" in both income classes

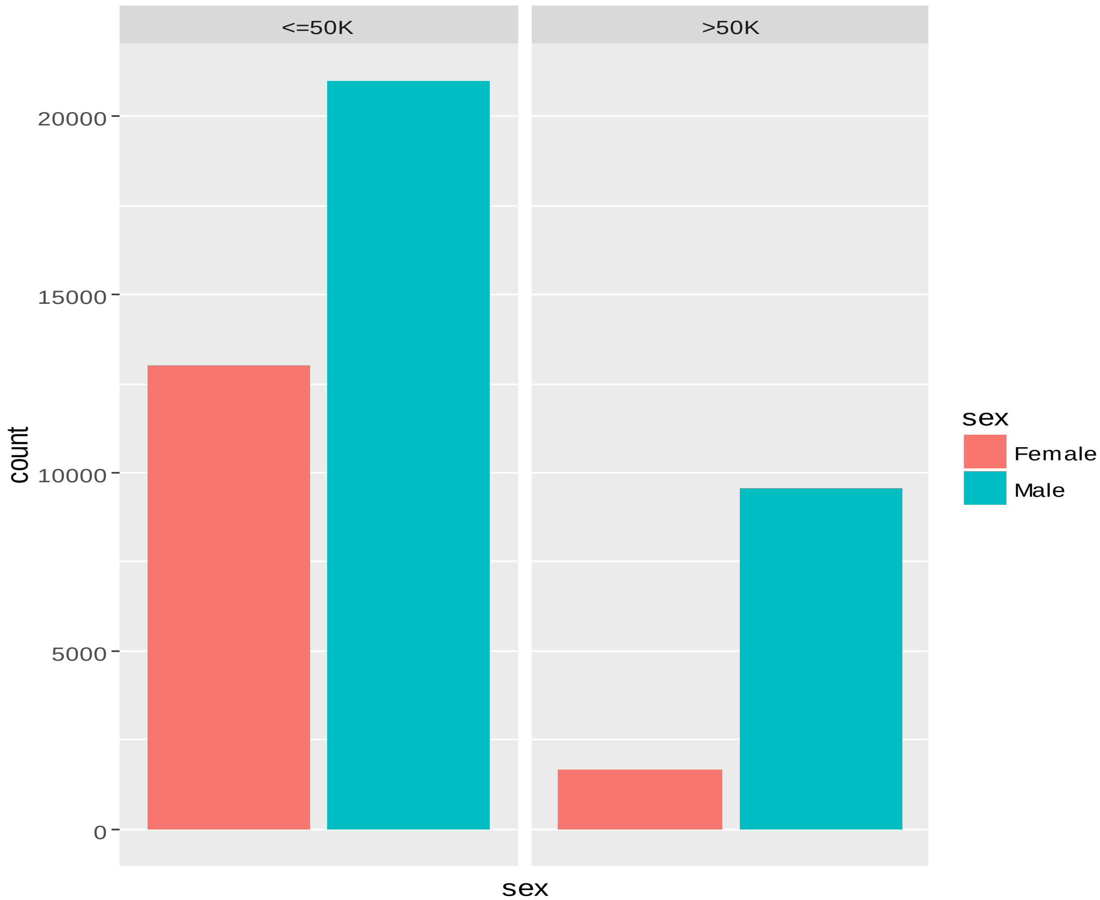
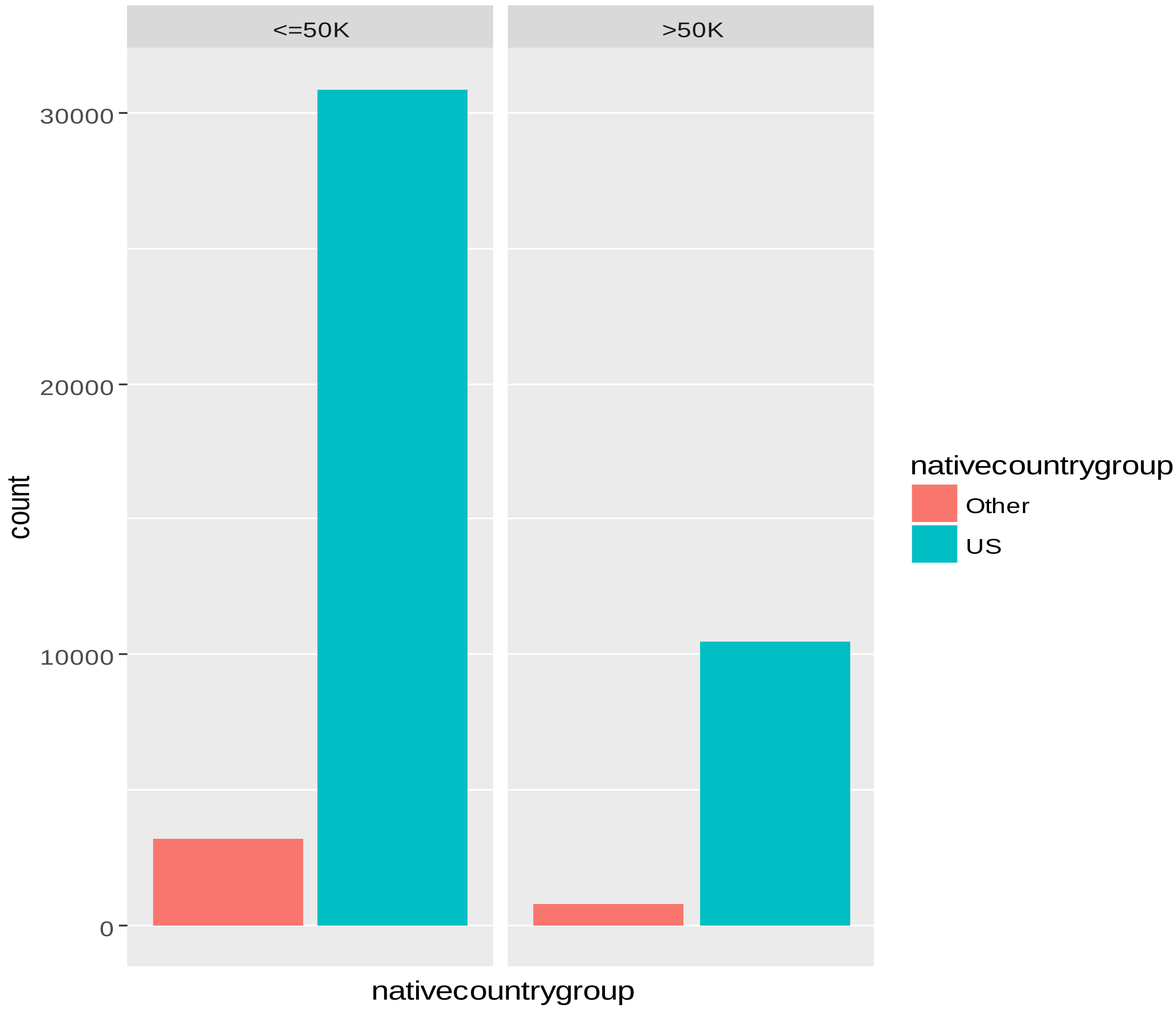


Figure 14: Distribution by “nativecountry” in both income classes



Methodology

The dataset is checked for missing values and observations with missing data are removed from the dataset.

All the variables in the dataset are analyzed in an explorative manner, through summary and descriptive statistics and several data distribution plots which can be viewed in ‘Exploratory Data Analysis’ section above.

The ‘caret’ package is used for developing prediction algorithm to fulfill the desired goal.

The dataset is checked for the zero variance predictors as they exhibit zero variability and thus will not exert any influence on the variation in target variable values. Hence, columns having a unique value are removed from the dataset.

In next step, training and test sets are created from “file” dataset including 70% of total rows randomly in the training set and rest 30% in test set. Since it is clear from Figures 4 and 5 in Exploratory Data Analysis section above that ‘capitalgain’ and ‘capitalloss’ are highly skewed variables, they are pre-processed by centering and scaling transformation to standardize both these variables in training as well as test set.

The training set is split into a training and validation set randomly to avoid the problem of model overfitting.

Three different prediction models – Generalized Linear Model, Random Forests and Gradient Boosting Model are used individually to predict target outcome which is a classification variable. To scale up the prediction accuracy, predictions from all these models are used to ensemble the models. Final prediction from the ensemble model is chosen as the outcome predicted by minimum two out of three predictions in every sample in the testing set. This model ensembling technique is more reliable and delivers a higher accuracy rate of 86.5 % on the testing set.