COMP 551 - APPLIED MACHINE LEARNING

Mini-project 1

Hong Kun Tian
McGill University
260866355
hong.tian@mail.mcgill.ca

Lane Hunter-Leiski
McGill University
260661836
lane.hunter-leiski
@mail.mcgill.ca

Murat Polat
McGill University
260738186
murat.polat@mail.mcgill.ca

February 11, 2020

Abstract

Logistic Regression and Naive Bayes in machine learning are two important classifiers. The aim of this project to predict the outcome from four different datasets using Naive Bayes and Logistic Regression, and to evaluate their performance. K-fold Cross Validation, a powerful tool to evaluate different machine learning models, was applied to all four datasets. Several tools such as pandas and numpy were used for data analysis and processing. Accuracy of both methods in all four datasets was measured. Later discussed in Results section, Logistic Regression was found to be a more accurate method for the datasets used while Naive Bayes was found to be the faster method to train.

1 Introduction

The first dataset that is used in the project is data gathered from the ionosphere. The data is gathered by transmitting power and measuring the signals received. The task for this dataset is to predict whether there exists a structure in the ionosphere. A "good" outcome is a returned signal while a "bad" outcome is absence of a return signal, e.g. the transmitted power passing through the ionosphere.

The second dataset used is data gathered from adults. The task for this dataset is to predict whether an adult has an annual income of over 50000 dollars by evaluating the attributes.

The third dataset used is a set of mushroom records. The attributes in this dataset are either binary, or categorical, and the task for this dataset is to predict whether the mushroom is definitely edible or definitely poisonous.

The fourth and the last dataset used is a set of bank marketing data. The dataset has 20 input variables, and the task is to predict whether a client subscribed a term deposit by evaluating the given variables.

The dataset sizes vary from 351 instances to 48842 instances, and those datasets were used to train and evaluate the performance of Logistic Regression and Naive Bayes. To evaluate the performance of Naive Bayes, the algorithm was written from scratch.

2 Datasets

The first dataset, Ionosphere, consists of 351 non-null instances and 34 attributes. The second dataset, Adult, consists of 48842 non-null instances and 15 instances. The third dataset, Mushroom, consists of 8124 non-null instances and 23 instances. Finally, the fourth dataset, Bank, has 41188 entries and 21 instances. One instance in each dataset contains the attribute to be predicted. The data in each dataset was split into a training and a test set.

On each dataset, 20% of the instances were used for testing while the remaining 80% were used for training. Train_test_split method from scikit-learn was used for this process. The instances were chosen in random to minimize the bias, By using the "describe()" method in numpy

package, we gathered statistics such as the mean, standard deviation and the minimum value as well as the maximum. Histograms for each attribute were then gathered.

In the first dataset, the attribute for "good" and "bad" was mapped to 1 and 0 for convenience. Moreover, attribute 1 consists only of the value 0, thus it was discarded.

In the second dataset, the attribute class has 4 categories; however, values ">50K" and ">50K." were mapped to 1 while the values "<=50K" and "<=50K." were mapped to 0. The attributes were then split into two: Categorical, and Numerical. 91.7% of the values for the attribute capital-gain and 95.3% of the values for the attribute capital-loss feature zero while 89.7% of the values for the attribute country feature United States. We have considered omitting those attributes. However, doing so decreases the performance of the algorithm.

In the third dataset, the values for the attribute edibility, p and e, were mapped to 1 and 0 respectively. Moreover, as the dataset attribute values are categorical, we converted each attribute to binary sub-attributes to gather a numerical binary dataset. The modified dataset has 96 attributes.

In the fourth dataset, the values for the attribute y, yes and no, were mapped to 1 and 0 respectively. Moreover, similar to the third dataset, we converted each attribute to binary sub-attributes to gather a numerical binary dataset. The modified dataset has 54 attributes.

3 Results

As the primary goal of the project, we examined how Logistic Regression performed against Gaussian Naive Bayes. We performed 5-fold cross validation on the following experiment. We opted to use scikit-learn's train_test_split method to split our datasets into training and testing sets. We originally wrote our own method, but chose to use scikit-learn's because it could easily do stratified splits whereas the split method we wrote from scratch could not. This decision was made on the basis that scikit-learn's splits would yield more consistent and robust results as we gauge performance on the test set. For our experiment, Logistic Regression is run with $\alpha = 0.02$ and $\epsilon = 0.01$. We examine how differing learning rates affect accuracies later.

Presented below are the average training and validation accuracies gathered from fitting Logistic Regression and Naive Bayes using k-fold cross validation on the training set.

	LR	NB
Training accuracy	0.9241	0.7768
Validation accuracy	0.8571	0.7571

Table 1: Ionosphere

.	LR	NB
Training accuracy	0.8278	0.7522
Validation accuracy	0.8273	0.7522

Table 2: Adult

	LR	NB
Training accuracy	0.9948	0.9008
Validation accuracy	0.9935	0.9007

Table 3: Mushroom

	LR	NB
Training accuracy	0.8976	0.8873
Validation accuracy	0.8973	0.8873

Table 4: Bank

The most prominent observation is that Naive Bayes fits significantly faster than Logistic Regression on large datasets. This is especially apparent when fitting on the adult dataset with over 30000 instances of training data using full batch gradient descent. This is to be expected given that gradient descent on large datasets is expensive to compute and will impact the time it takes to fit data, whereas Naive Bayes computes its parameters explicitly instead of iteratively adjusting them.

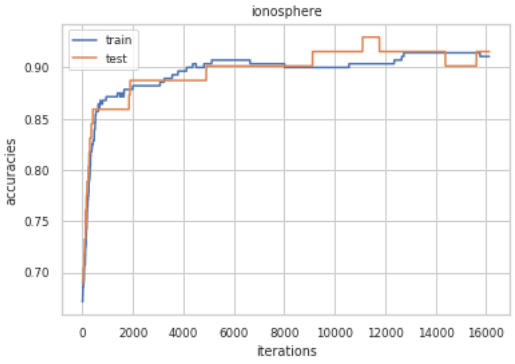
Additionally, we observe better accuracies across the board for Logistic Regression in comparison to Naive Bayes, for both average training accuracies as well as validations accuracies. This is indicative of the high bias, low variance that Naive Bayes suffers from. Also, the poor performance of Naive Bayes can also be attributed to the high number features and how they may not be completely independent of each other given the class label we wish to predict.

Now that we have established the differences between Logistic Regression and Naive Bayes,

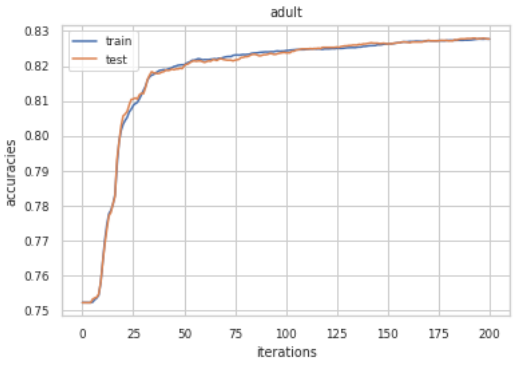
we investigate different learning rates for gradient descent applied to Logistic Regression. In our experiment, we set the threshold for change in cost function as a termination criteria as $\epsilon_{cost} = 0.0001$. The specific implementation for gradient descent we used for Logistic Regression did not employ any optimization techniques, and thus should be taken into account when considering sources of error. Since our learning rate remains constant throughout, it maybe be hard to land close to the true local minimum.

We ran the experiment with 5-fold cross validation on the training set for each learning rate in order to generate more robust results. For each learning rate, we took the average of the five validation accuracies. The learning rate with the highest average validation accuracy was deemed the best. The learning rates that we tested for each are $[0.02, 0.04, 0.06, 0.08, 0.1, 0.2, 0.4, 0.6, 0.8]$.

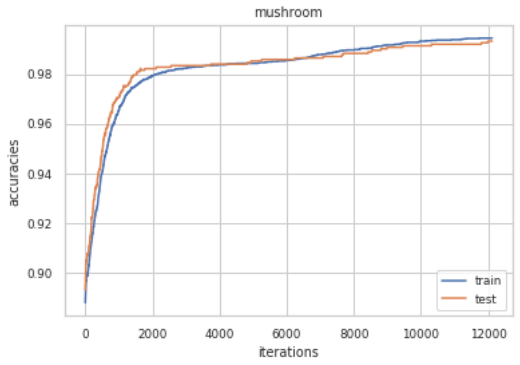
Presented below are the graphs for the four datasets plotting the training and test set accuracies over the number of iterations gradient descent took. The learning rate used is the best one found in the aforementioned experiment using the training set.



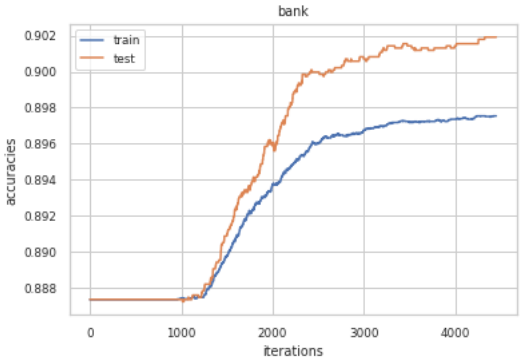
learning_rate = 0.02



learning_rate = 0.8



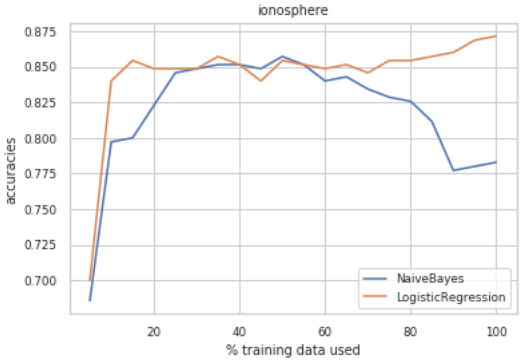
learning_rate = 0.02

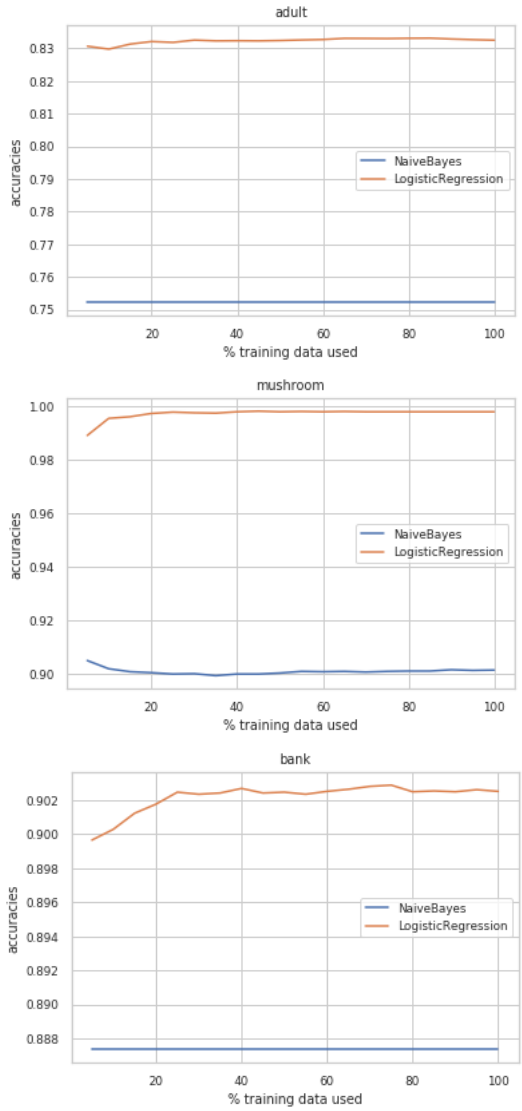


learning_rate = 0.02

One particular trend that we noticed while running the experiment is that the smaller the learning rate, the longer it takes to converge at a local minimum, whereas if the learning is too high, we're not guaranteed to converge and may oscillate too much when performing gradient descents.

Finally, our last experiment consisted of comparing how the Logistic Regression and Naive Bayes models performed when fit on varying training sizes. We performed 5-fold cross validation on the entirety of the dataset. For each fold, we take the $[5, 10, 15, 20, ..., 90, 95, 100]$ % of the dataset and save the accuracies for both the training subset and testing set. At the end of the experiment, we average out all the values for each percentage and plot the results. The following are the results from the four datasets.





Upon inspection of these graphs, it is apparent that an increase in percentage of training data does seem to help improve both models' performance in terms of accuracy. However, one key observation to notice would be that Naive Bayes' performance gains quickly plateaus as the quantity of instances increases. We can observe that for the ionosphere dataset where the number of instances is small, an increase in percentage of training data has Naive Bayes improve just as quickly as Logistic Regression. With that said, the improvements become almost nil on bigger datasets as can be seen by looking at Naive Bayes' accuracies for the adult dataset, mushroom dataset, and bank dataset. Logistic Regression has a similar trend where an increase in percentage of training data only improves the test accuracies by so much. In brief, there is diminishing return in improvement of testing accuracies as we increase the amount of training data.

4 Discussion and Conclusions

We have gathered several key takeaways from the differences between Naive Bayes and Logistic Regression approaches. We have seen for all datasets that Logistic Regression approach was more accurate, while the Naive Bayes approach was significantly faster.

As our four datasets varied greatly in size, we were able to compare both approaches in speed and performance. By looking at the benchmarks, we see that the difference in accuracy reduces as the dataset size increases.

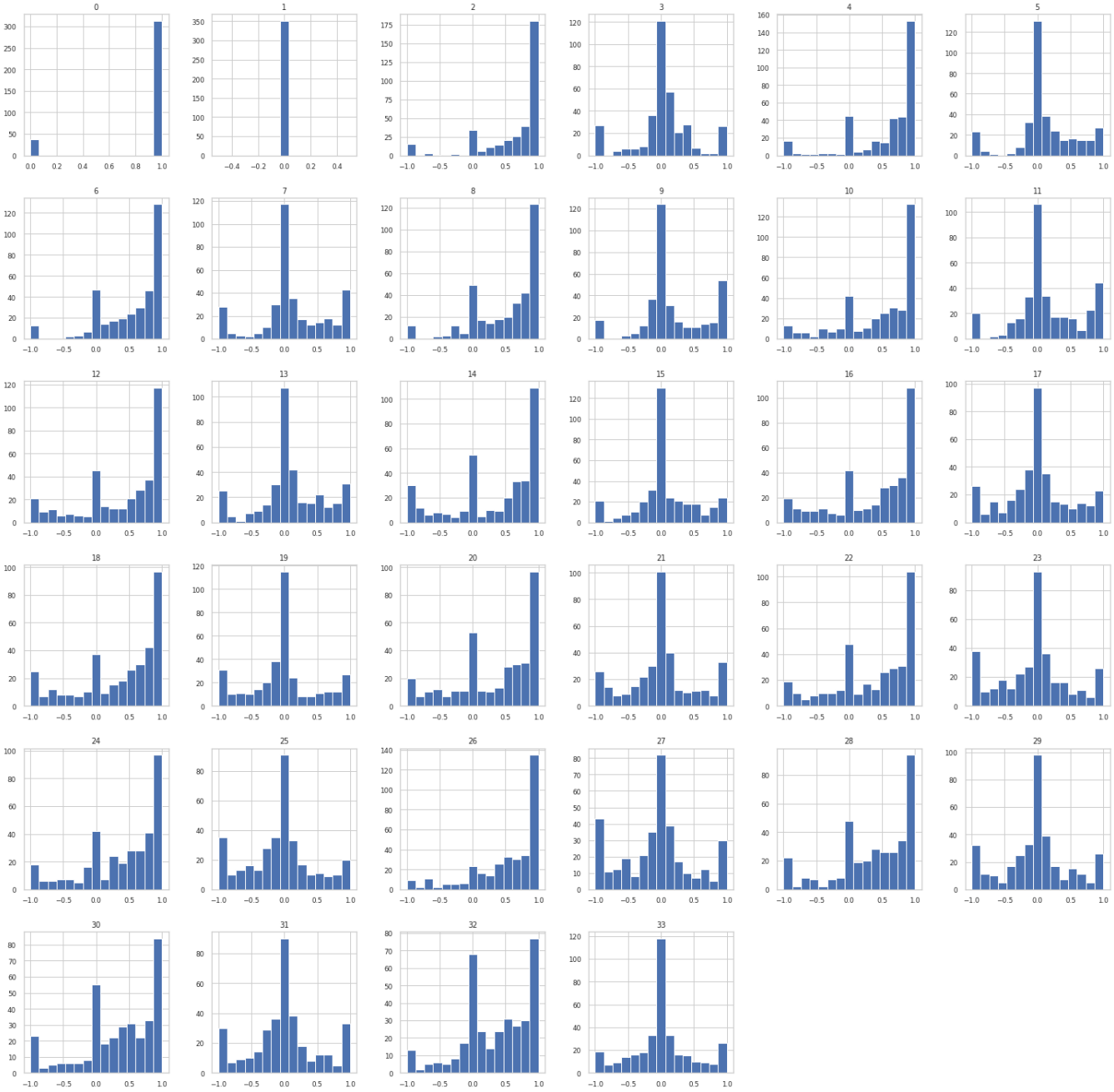
We have run the algorithm after removing attributes that are heavily weighed towards a value, expecting an increase in accuracy. However, we have actually seen decreased accuracy. The decrease in accuracy is greater when using Logistic Regression. This is expected, as Logistic Regression works with less information when the attributes are removed.

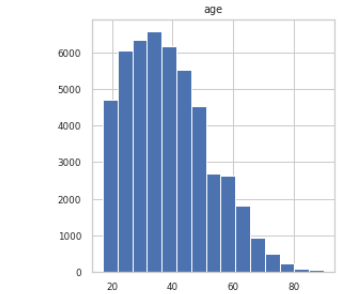
One approach to possibly increase the accuracy of the algorithm is regularization. Moreover, implementing an optimization technique for gradient descent in Logistic Regression will increase the accuracy.

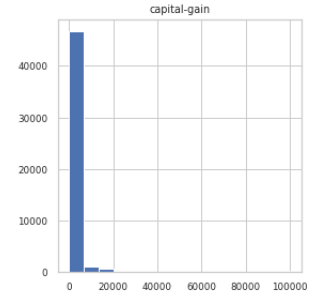
5 Statement of Contributions

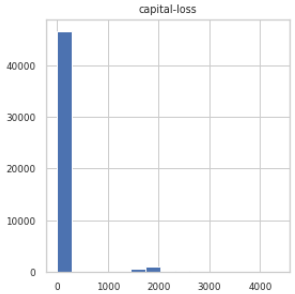
Hong Kun worked on for data processing, Logistic Regression, Naive Bayes implementation, and experiments code implementation, and wrote the Results section. Murat worked on data processing and Naive Bayes implementation and wrote the Abstract, Introduction, Datasets and Discussion&Conclusion sections.

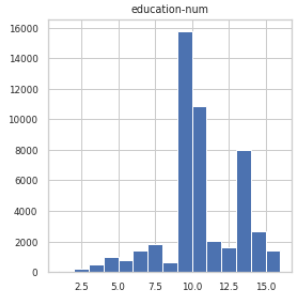
6 Appendix

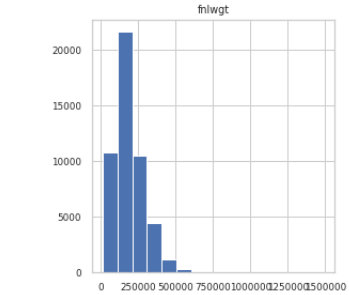

0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33

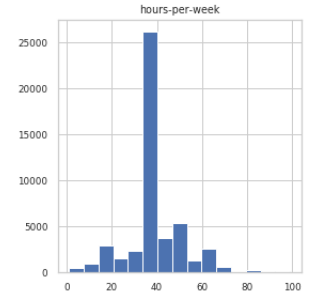

age

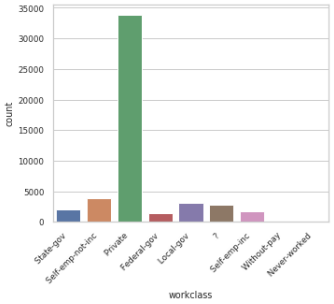

capital-gain

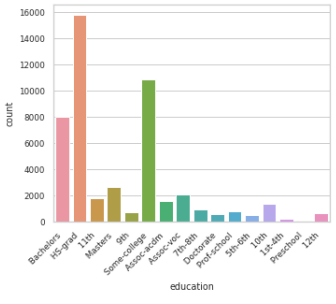

capital-loss

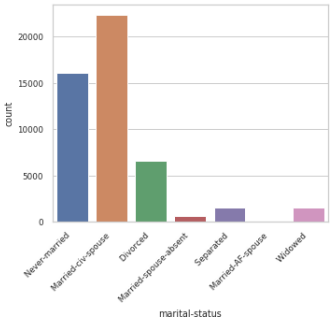

education-num

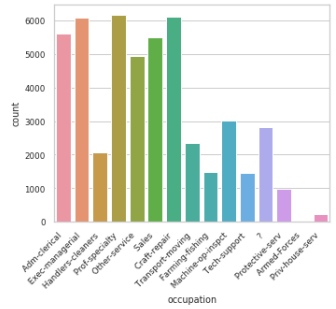

fnlwgt

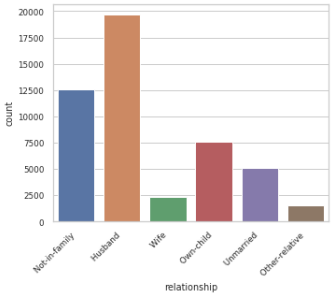

hours-per-week

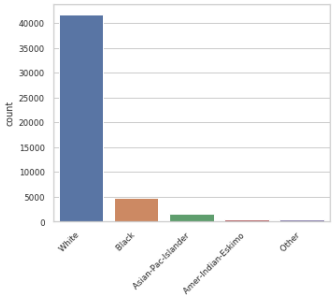

count
State-gov
Self-emp-not-inc
Private
Federal-gov
Local-gov
?
Self-emp-inc
Without-pay
Never-worked
workclass

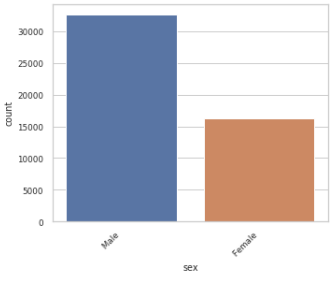

count
Bachelors
HS-grad
11th
Masters
9th
Some-college
Assoc-acdm
Assoc-voc
7th-8th
Doctorate
Prof-school
5th-6th
10th
1st-4th
Preschool
12th
education

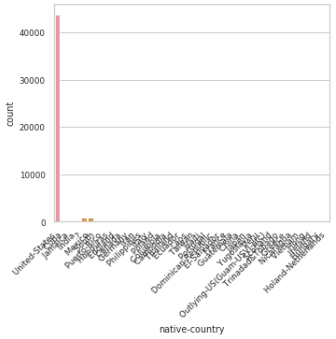

count
Never-married
Married-civ-spouse
Divorced
Married-spouse-absent
Separated
Married-AF-spouse
Widowed
marital-status


count
Adm-clerical
Exec-managerial
Handlers-cleaners
Prof-specialty
Other-service
Sales
Craft-repair
Transport-moving
Farming-fishing
Machine-op-inspct
Tech-support
?
Protective-serv
Armed-Forces
Priv-house-serv
occupation


count
Not-in-family
Husband
Wife
Own-child
Unmarried
Other-relative
relationship


count
White
Black
Asian-Pac-Islander
Amer-Indian-Eskimo
Other
race


count
Male
Female
sex


count
native-country