**Abalone**

This dataset has 4177 observations with 9 features.

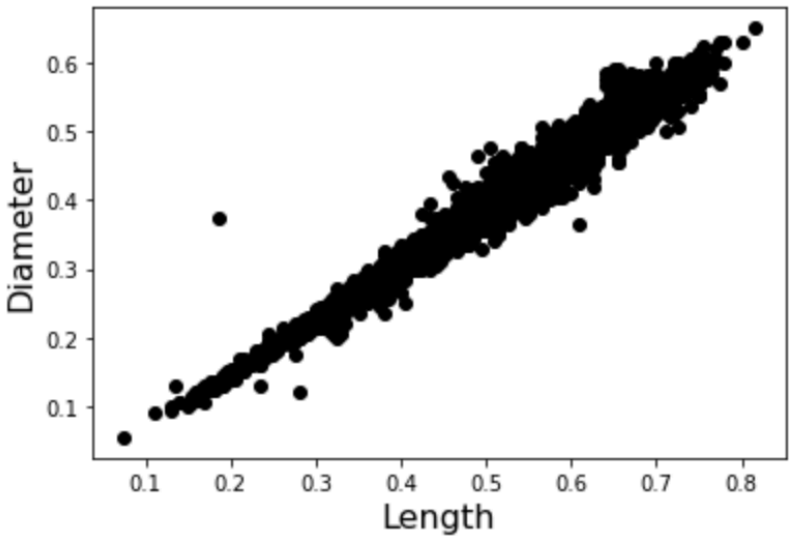
Columns = ['Sex', 'Length', 'Diameter', 'Height', 'Whole weight', 'Shucked weight',
'Viscera weight', 'Shell weight', ‘age']

No missing values in the dataset

Numerical features = ['Length', 'Diameter', 'Height', 'Whole weight', 'Shucked weight’, 'Viscera weight', 'Shell weight', 'age']
Categorical features = ['Sex']

**Exercise 1**

slope = 0.8154606917560964
intercept = -0.019413705519977176
correlation coefficient = 0.98681158




Interpretation of the slope: If the length of the abalone increases by 1 mm, then the model predicts that the diameter will increase by approximately 0.815 mm.
Interpretation of the intercept: If the length of the abalone is 0, then the model predicts that the diameter is approximately -0.019 mm.

The extreme correlation coefficient = 0.987 indicates a strong linear relationship between diameter and length of abalone where a change in one variable is accompanied by a perfectly consistent change in the other. For this relationship, all of the data points fall on a line. Positive coefficient indicates that when the value of length increases, the value of diameter also tends to increase. Positive relationship produces an upward slope on a scatterplot.

Let us consider the ratios:

```
count  4177.000000
mean      1.291880
std       0.059065
min       0.493333
25%       1.257732
50%       1.288462
75%       1.321839
max       2.333333
```

The ratios of length to diameter are similar for all samples. This is shown by the standard deviation being small.

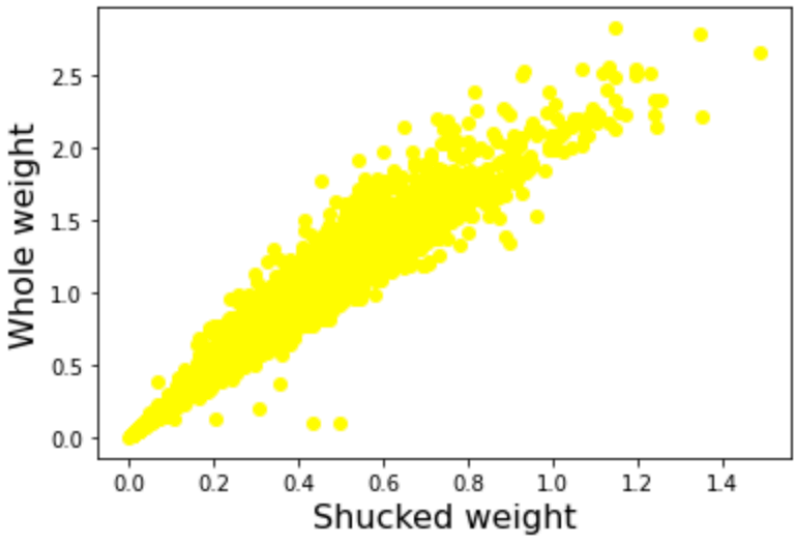
From the description of the data we know that length and diameter are measured perpendicular to each other. As abalones grow by adding new layers to their shell both the length and diameter have to increase but the proportion of them remains the same.

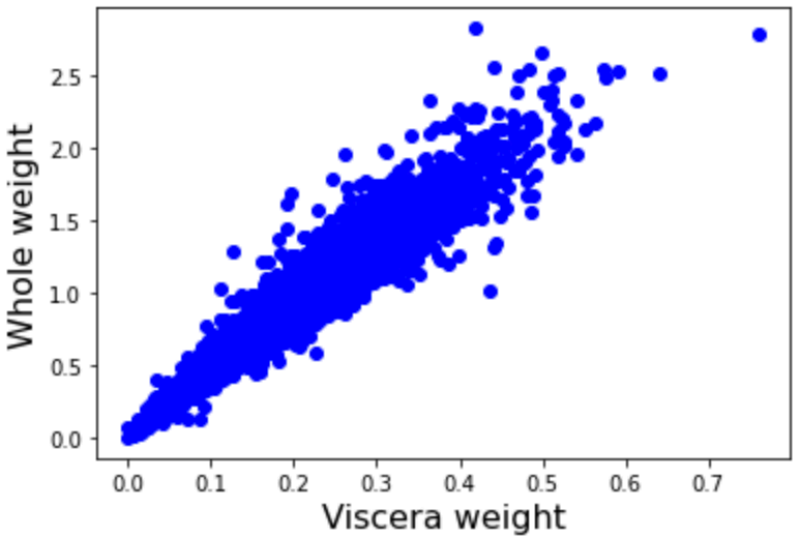
**Exercise 2**

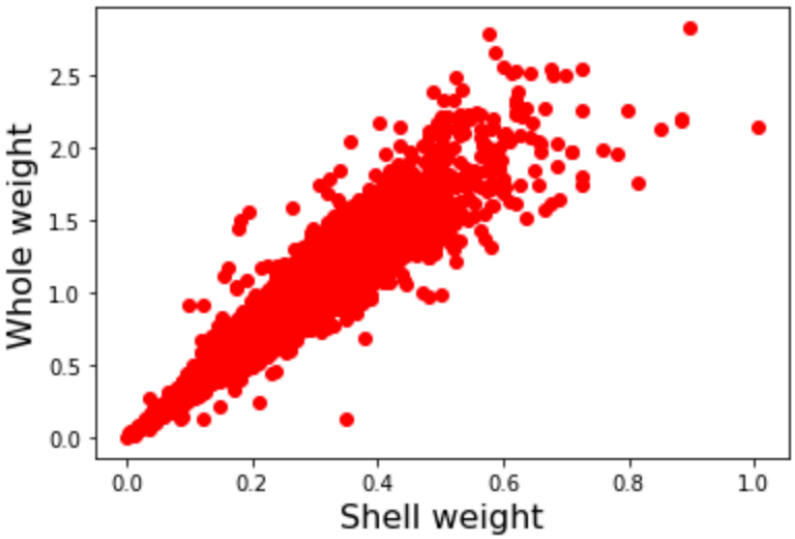
slope = [0.93656021 1.11164951 1.25296164]
intercept = -0.00782983143730065

correlation coefficients:

| | Shucked weight | Viscera weight | Shell weight | Whole weight |
|---|---|---|---|---|
| Shucked weight | 1.000000 | 0.931961 | 0.882617 | 0.969405 |
| Viscera weight | 0.931961 | 1.000000 | 0.907656 | 0.966375 |
| Shell weight | 0.882617 | 0.907656 | 1.000000 | 0.955355 |
| Whole weight | 0.969405 | 0.966375 | 0.955355 | 1.000000 |





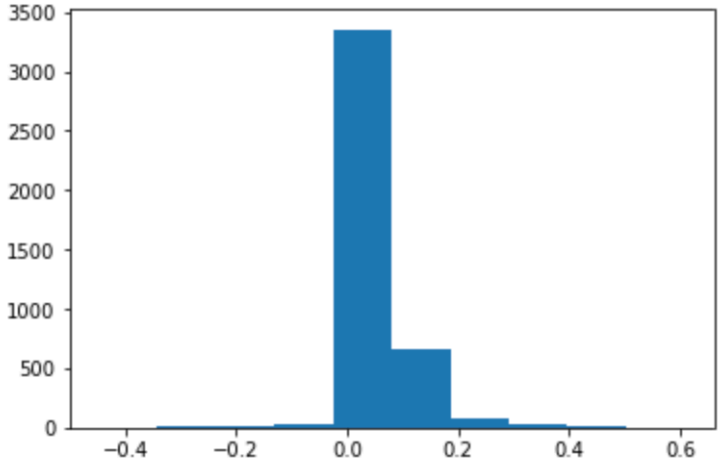

Interpretation of the slope:

1. If the Shucked weight of the abalone increases by 1 gram, then the model predicts that the whole weight will increase by approximately 0.936 gram.
2. If the Viscera weight of the abalone increases by 1 gram, then the model predicts that the whole weight will increase by approximately 1.112 gram.
3. If the Shell weight of the abalone increases by 1 gram, then the model predicts that the whole weight will increase by approximately 1.253 gram.

Interpretation of the intercept: If the weights of different parts of the abalone are equal to 0, then the model predicts that the whole weight is approximately -0.00783 gram.

Whole weight seems to be highly correlated with other weight predictors and seems to be the sum of Shucked weight, Viscera weight and Shell weight. Let's check this assumption.

The histogram below illustrates the differences between the whole weight and the sum of weights of different parts. We can see that most of these differences is approximately 0. However, some samples show a small deviation from zero and what is quite strange, we can see that for some samples the weight difference is negative. This could be considered as an error in data.



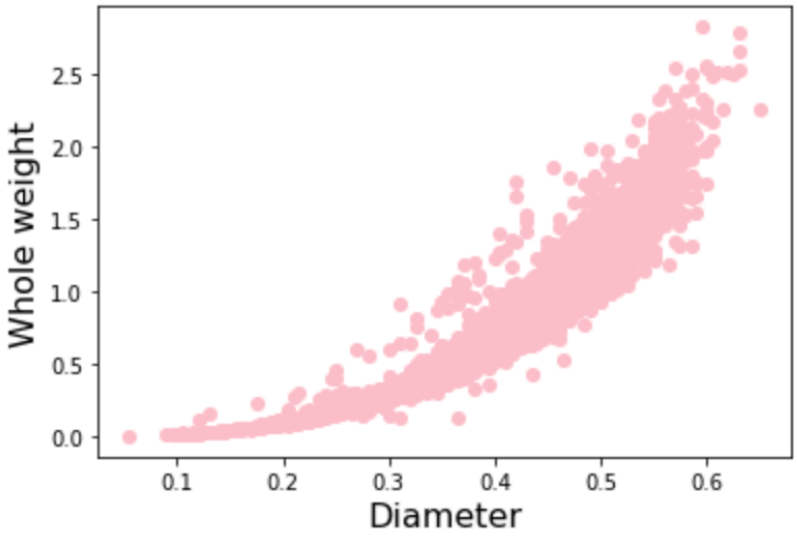
```
count  4177.000000
mean      0.049950
std       0.058072
min      -0.447500
25%       0.018000
50%       0.037000
75%       0.068000
max       0.608000
Number of weight difference observations that are negative: 155
```

I have also made a boolean array of every specimen, where corresponding element is True if the whole weight is equal to the sum of the weights of different parts, and False otherwise. From this, I found only 6 elements which are True.

To conclude, we are unable to say that the Whole weight is the sum of Shucked weight, Viscera weight and Shell weight because the data is invalid.

**Exercise 3**

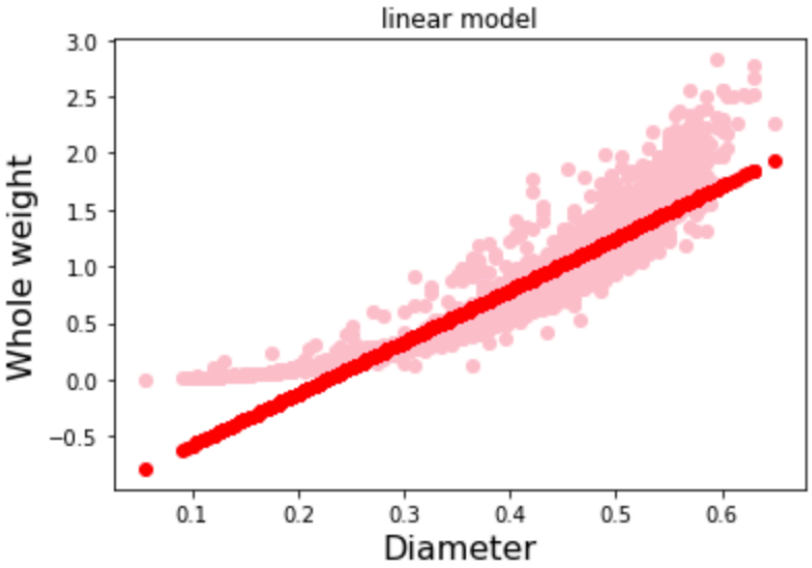
The scatter plot illustrates the relationship between Diameter and Whole weight.
Whole weight = function (Diameter)




**Exercise 3a**

A simple linear model, whole weight = a * diameter + b

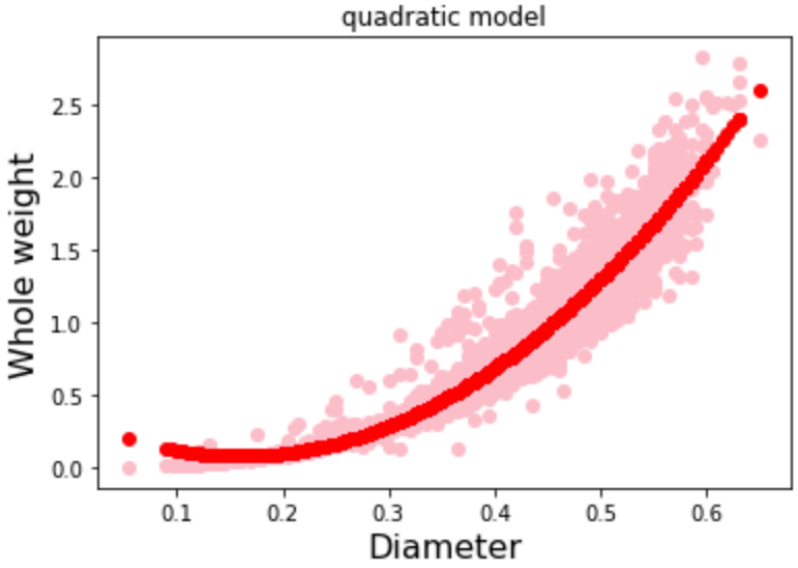
r2 = 0.856461592183965
correlation coefficient = 0.925452101507131




**Exercise 3b**

A quadratic model, whole weight = a * diameter + b * diameter^2 + c

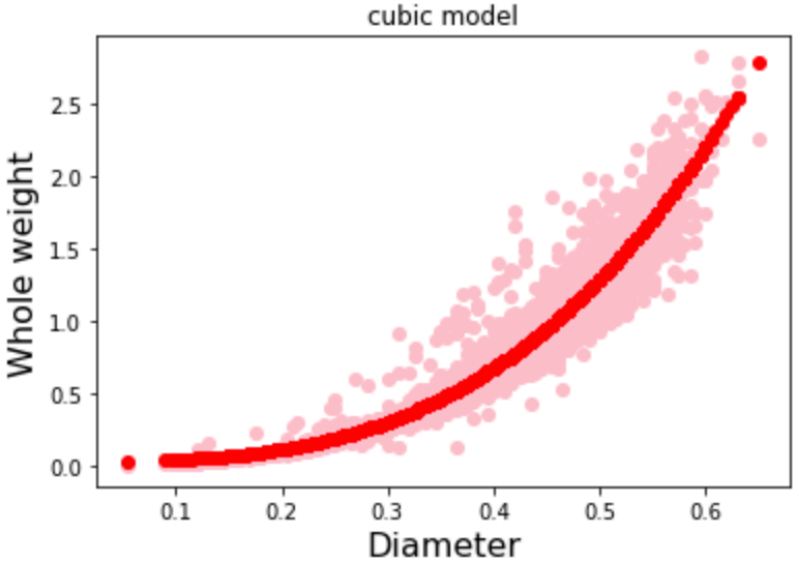
r2 = 0.9267621513288571
correlation coefficient = 0.962684866053714




**Exercise 3c**

A cubic model without lower order or constant terms, whole weight = a * diameter^3

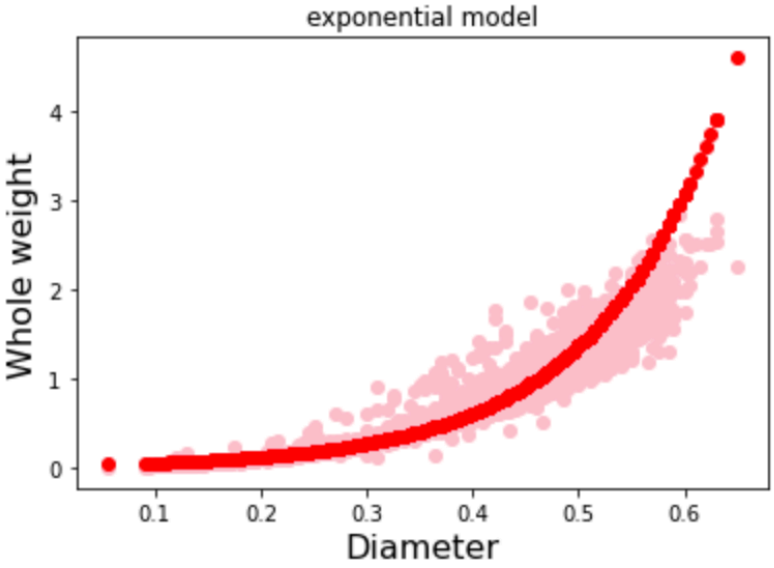
r2 = 0.9275205320032817
correlation coefficient = 0.9630786738388935




**Exercise 3d**

An exponential model, log (whole weight) = a * diameter + b

r2 = 0.9283555266562875
correlation coefficient = 0.9635120791439449




We can see the plot looks similar to the exponential function with a monotonically increasing relationship, and, as expected, as the diameter of shell increases we expect the whole weight to increase as abalone adds new growth rings to its shell. So we determine that the exponential model is the most effective according to correlation coefficient and scatter plot.

**Exercise 4**

First of all, we recode the class values (M and F -> 1, I -> 0).
For all next steps the data is split into training set (80%) and test set(20%), random_state = 0.

**Exercise 4a**
sex = function (length)

score = 0.7583732057416268

**Exercise 4b**
sex = function (whole weight)

score = 0.7834928229665071

**Exercise 4c**
sex = function (rings)

score = 0.7715311004784688

**Exercise 4d**
sex = function (length, whole weight, rings)

score = 0.8098086124401914

**Adult**

This dataset has 48842 observations with 15 features.

*Features:*

*workclass:* Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked. Individual work category.

*education*: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool. Individual's highest education degree.

*education-num:* Individual's number of degrees.

*marital-status*: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse. Individual marital status.

*occupation:* Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces. Individual's occupation.

*relationship*: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried. Individual's relation in a family.

*race:* White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black. Race of Individual.

*sex*: Female, Male.

*native-country:* United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands. Individual's native country.

*age:* continuous. Age of an individual.

*fnlwgt:* final weight, continuous. The weights on the CPS files are controlled to independent estimates of the civilian noninstitutional population of the US. These are prepared monthly for us by Population Division here at the Census Bureau.

*capital-gain:* continuous.

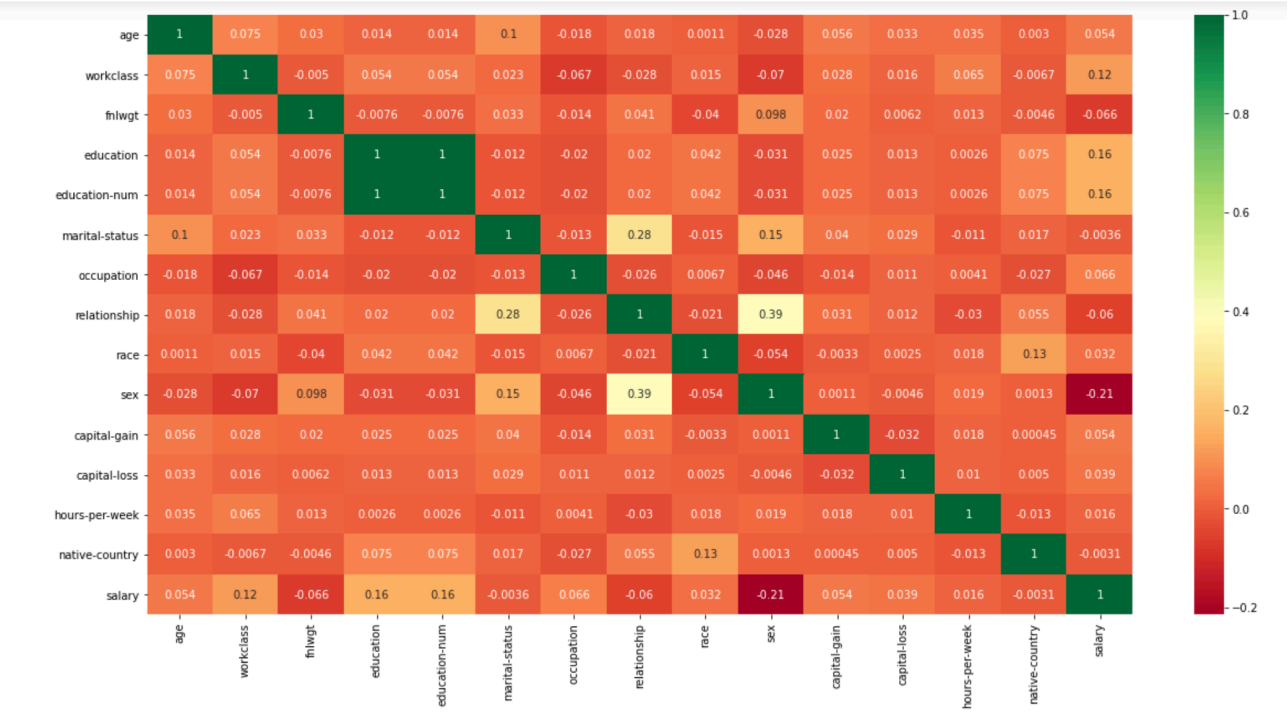
*capital-loss:* continuous.

*hours-per-week:* continuous.Individual's working hour per week.

*salary*: >50K, <=50K.

**Exercise 5**

Correlation matrix for attributes:




According to the matrix, we can see that the correlation coefficient between education and education-num is equal to 1. Hence, we can eliminate education-num because it is just a numeric representation of the attribute education.

**Exercise 5a**

First, we evaluate on the data using the whole set of attributes. Initial accuracy:

| | | |
|---|---|---|
| Correctly Classified Instances | 41348 | 84.6566 % |
| Incorrectly Classified Instances | 7494 | 15.3434 % |

For the next step, I run training and evaluation iteratively, removing a single feature in each iteration. Then, we check evaluation metrics against the initial accuracy.
The goal of this technique is to see which of the features don't significantly affect the evaluation.

accuracy without age:

| | | |
|---|---|---|
| Correctly Classified Instances | 41352 | 84.6648 % |
| Incorrectly Classified Instances | 7490 | 15.3352 % |

accuracy without workclass:

| | | |
|---|---|---|
| Correctly Classified Instances | 41252 | 84.4601 % |
| Incorrectly Classified Instances | 7590 | 15.5399 % |

accuracy without fnlwgt:

| | | |
|---|---|---|
| Correctly Classified Instances | 41259 | 84.4744 % |
| Incorrectly Classified Instances | 7583 | 15.5256 % |

accuracy without education:

| | | |
|---|---|---|
| Correctly Classified Instances | 41276 | 84.5092 % |
| Incorrectly Classified Instances | 7566 | 15.4908 % |

accuracy without marital-status:

| | | |
|---|---|---|
| Correctly Classified Instances | 41220 | 84.3946 % |
| Incorrectly Classified Instances | 7622 | 15.6054 % |

accuracy without occupation:

| | | |
|---|---|---|
| Correctly Classified Instances | 39448 | 80.7666 % |
| Incorrectly Classified Instances | 9394 | 19.2334 % |

accuracy without relationship:

| | | |
|---|---|---|
| Correctly Classified Instances | 38598 | 79.0262 % |
| Incorrectly Classified Instances | 10244 | 20.9738 % |

accuracy without race:

| | | |
|---|---|---|
| Correctly Classified Instances | 41362 | 84.6853 % |
| Incorrectly Classified Instances | 7480 | 15.3147 % |

accuracy without capital gain:

| | | |
|---|---|---|
| Correctly Classified Instances | 41348 | 84.6566 % |
| Incorrectly Classified Instances | 7494 | 15.3434 % |

accuracy without capital loss:

| | | |
|---|---|---|
| Correctly Classified Instances | 41353 | 84.6669 % |
| Incorrectly Classified Instances | 7489 | 15.3331 % |

accuracy without hours-per-week:

| | | |
|---|---|---|
| Correctly Classified Instances | 41033 | 84.0117 % |
| Incorrectly Classified Instances | 7809 | 15.9883 % |

accuracy without native country:

| | | |
|---|---|---|
| Correctly Classified Instances | 41299 | 84.5563 % |
| Incorrectly Classified Instances | 7543 | 15.4437 % |

accuracy without salary:

| | | |
|---|---|---|
| Correctly Classified Instances | 41312 | 84.5829 % |
| Incorrectly Classified Instances | 7530 | 15.4171 % |

The differences in accuracy are more than 1% for these attributes:
relationship (the difference is 5.6304%)
occupation (the difference is 3.89%)

Also it is worthwhile to mention that the next most significant features are hours-per-week(the difference is 0.6449%) and marital-status(the difference is 0.262%).

Every other feature (except relationship and occupation) can be removed without affecting the accuracy by more than 1%. This is because these attributes are not of much importance in terms of the contribution with gender. In addition to that, some attributes could be replaced by other attributes that contain the similar information. For instance, “age” has much in common with “marital-status”, “workclass” presents almost the same information as “occupation”.

Similar results are achieved using Weka InfoGainAttributeEval, evaluating the worth of an attribute by measuring the information gain with respect to the class “sex”.

```
=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 9 sex):
        Information Gain Ranking Filter

Ranked attributes:
 0.39185   7 relationship
 0.16185   5 marital-status
 0.12851   6 occupation
 0.05443  12 hours-per-week
 0.03669  14 class
 0.01419   2 workclass
 0.01159   1 age
 0.00891   8 race
 0.00789  10 capital-gain
 0.00656   4 education
 0.00418  11 capital-loss
 0.00265  13 native-country
 0.00115   3 fnlwgt

Selected attributes: 7,5,6,12,14,2,1,8,10,4,11,13,3 : 13
```

I evaluate the model accuracy when taking different combinations of these 4 attributes (relationship, marital-status, occupation, hours-per-week). This leads to the following results:

features = [marital-status, occupation, relationship, hours-per-week]

| | | |
|---|---|---|
| Correctly Classified Instances | 40954 | 83.85 % |
| Incorrectly Classified Instances | 7888 | 16.15 % |

Trying to decrease the amount of features:

features = [marital-status, occupation, relationship]
Correctly Classified Instances 40528 82.9778 %
Incorrectly Classified Instances 8314 17.0222 %

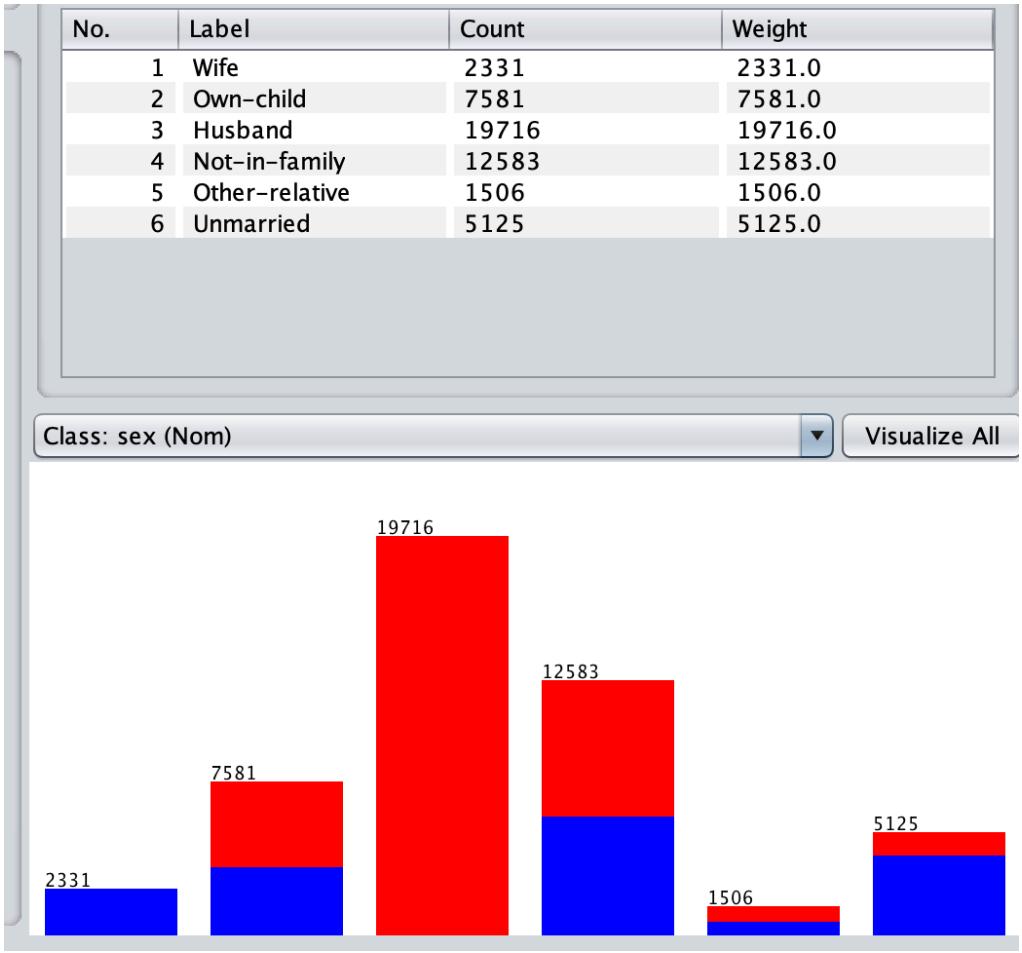
features = [occupation, relationship, hours-per-week]
Correctly Classified Instances 40836 83.6084 %
Incorrectly Classified Instances 8006 16.3916 %

features = [marital-status, relationship, hours-per-week]
Correctly Classified Instances 38983 79.8145 %
Incorrectly Classified Instances 9859 20.1855 %

features = [marital-status, occupation, hours-per-week]
Correctly Classified Instances 38200 78.2114 %
Incorrectly Classified Instances 10642 21.7886 %

The best combination of accuracy (83.6084%) and simplicity (only 3 attributes) is achieved using the features occupation, relationship, hours-per-week.

**Exercise 5b**



Relationship-status is helpful, because the correlation coefficient between sex and relationship is equal to 0.39 (the highest value with respect to sex). The data is correlated, because a husband is always male and a wife is always female.

**Exercise 5c**

native-country=Holand-Netherlands

The weight for native-country = Holand-Netherlands is the highest one due to the fact that the data consists of only one person from Holland-Netherlands and they are female. Therefore, classifier might always expect a person from Holand-Netherlands to be female.