Before you turn this problem in, make sure everything runs as expected. First, **restart the kernel** (in the menubar, select Kernel$\rightarrow$Restart) and then **run all cells** (in the menubar, select Cell$\rightarrow$Run All).

Make sure you fill in any place that says `YOUR CODE HERE` or "YOUR ANSWER HERE", as well as your name and collaborators below:

In [1]:

```
NAME = "Ayush Koirala"
ID = "st122802"
```

Machine Learning Lab 04: Multinomial Logistic Regression

Generalized Linear Models

From lecture, we know that members of the exponential family distributions can be written in the form $$p(y;\eta) = b(y)e^{(\eta^\top T(y)-a(\eta))},$$ where

- η is the natural parameter or canonical paramter of the distribution,
- $T(y)$ is the sufficient statistic (we normally use $T(y) = y$),
- $b(y)$ is an arbitrary scalar function of y, and
- $a(\eta)$ is the log partition function. We use $e^{a(\eta)}$ just to normalize the distribution to have a sum or integral of 1.

Each choice of T, a, and b defines a family (set) of distributions parameterized by η.

If we can write $p(y \mid \mathbf{x} ; \theta)$ as a member of the exponential family of distributions with parameters $\mathbf{\eta}$ with $\eta_i = \theta^\top_i \mathbf{x}$, we obtain a *generalized linear model* that can be optimized using the maximum likelihood principle.

The GLM for the Gaussian distribution with natural parameter η being the mean of the Gaussian gives us ordinary linear regression.

The Bernoulli distribution with parameter ϕ can be written as an exponential distribution with natural parmeter $\eta = \log \frac{\phi}{1-\phi}$. The GLM for this distribution is logistic regression.

When we write the multinomial distribution with paremeters $\phi_i > 0$ for classes $i \in 1..K$ with the constraint that $$\sum_{i=1}^{K} \phi_i = 1$$ as a member of the exponential family, the resulting GLM is called *multinomial logistic regression*. The parameters $\phi_1, \ldots, \phi_K$ are written in terms of θ as $$\phi_i = p(y = i \mid \mathbf{x}; \theta) = \frac{e^{\theta^\top_i \mathbf{x}}}{\sum_{j=1}^{K}e^{\theta^\top_j \mathbf{x}}}. $$

Optimizing a Multinomial Regression Model

In multinomial regression, we have

1. Data are pairs $\mathbf{x}^{(i)}, y^{(i)}$ with $\mathbf{x}^{(i)} \in \mathbb{R}^n$ and $y \in 1..K$.
2. The hypothesis is a vector-valued function $$\mathbf{h}_\theta(\mathbf{x}) = \begin{bmatrix} p(y = 1 \mid \mathbf{x} ; \theta) \\
$$p(y = 2 \mid \mathbf{x} ; \theta) \\
\vdots \\
p(y = K \mid \mathbf{x} ; \theta) \end{bmatrix},$$

where $$p(y = i \mid \mathbf{x}) = \phi_i = p(y = i \mid \mathbf{x}; \theta) = \frac{e^{\theta^\top_i \mathbf{x}}}{\sum_{j=1}^{K}e^{\theta^\top_j \mathbf{x}}}. $$

We need a cost function and a way to minimize that cost function. As usual, we try to find the parameters maximizing the likelihood or log likelihood function, or equivalently, minimizing the negative log likelihood function:
$$\theta^* = \text{argmax}_\theta {\cal L}(\theta) = \text{argmax}_\theta \ell(\theta) = \text{argmin}_\theta J(\theta),$$

where $$\begin{eqnarray} J(\theta) & = & - \ell(\theta) \\ & = & - \sum_{i=1}^m \log p(y^{(i)} \mid \textbf{x}^{(i)} ; \theta). \end{eqnarray}$$

Now that we know what is $J(\theta)$, let's try to find its minimimum by taking the derivatives with respect to an arbitrary parameter θ_{kl}, the l-th element of the parameter vector θ_k for class k. Before we start, let's define a variable a_k as the linear activation for class k in the softmax function: $$ a_k = \theta_k^\top \mathbf{x}^{(i)}, $$ and rewrite the softmax more conveniently as $$ \phi_k = \frac{e^{a_k}}{\sum_{j=1}^K e^{a_j}}. $$ That makes it a little easier to compute the gradient: $$\begin{eqnarray} \frac{\partial J}{\partial \theta_{kl}} & = & - \sum_{i=1}^m \frac{1}{\phi_{y^{(i)}}} \frac{\partial \phi_{y^{(i)}}}{\partial \theta_{kl}}. \\ \end{eqnarray}$$ Using the chain rule, we have $$\frac{\partial \phi_{y^{(i)}}}{\partial \theta_{kl}} = \sum_{j=1}^K \frac{\partial \phi_{y^{(i)}}}{\partial a_j} \frac{\partial a_j}{\partial \theta_{kl}}$$ The second factor is easy: $$ \frac{\partial a_j}{\partial \theta_{kl}} = \delta(k=j)x^{(i)}_l. $$ For the first factor, we have $$\begin{eqnarray} \frac{\partial \phi_{y^{(i)}}}{\partial a_j} & = & \frac{ \left[\delta(y^{(i)}=j)e^{a_j} \sum_{c=1}^K e^{a_c} \right] - e^{a_j} e^{a_j} }{\left[\sum_{c=1}^K e^{a_c} \right]^2} \\ & = & \delta(y^{(i)}=j) \phi_j - \phi_j^2 \end{eqnarray}$$

Substituting what we've derived into the definition above, we obtain $$ \frac{\partial J}{\theta_{kl}} = - \sum_{i=1}^m \sum_{j=1}^K (\delta(y^{(i)}=j) - \phi_j) \frac{\partial a_j}{\partial \theta_{kl}}. $$

There are two ways to do the calculation. In deep neural networks with multinomial outputs, we want to first calculate the $\frac{\partial J}{\partial a_j}$ terms then use them to calculate $\frac{\partial J}{\partial \theta_{kl}}$.

However, if we only have the "single layer" model described up till now, we note that $$\frac{\partial a_j}{\partial \theta_{kl}} = \delta(j=k) x^{(i)}_l,$$ so we can simplify as follows: $$\begin{eqnarray} \frac{\partial J}{\theta_{kl}} & = & - \sum_{i=1}^m \sum_{j=1}^K (\delta(y^{(i)}=j) - \phi_j) \frac{\partial a_j}{\partial \theta_{kl}} \\ & = & - \sum_{i=1}^m \sum_{j=1}^K (\delta(y^{(i)}=j) - \phi_j) \delta(j=k) x^{(i)}_l \\ & = & - \sum_{i=1}^m (\delta(y^{(i)}=k) - \phi_k) x^{(i)}_l \\ \end{eqnarray}$$

Put It Together

OK! Now we have all 4 criteria for our multinomial regression model:

1. Data are pairs $\mathbf{x}^{(i)}, y^{(i)}$ with $\mathbf{x}^{(i)} \in \mathbb{R}^n$ and $y \in 1..K$.
2. The hypothesis is a vector-valued function $$\mathbf{h}_\theta(\mathbf{x}) = \begin{bmatrix} p(y = 1 \mid \mathbf{x} ; \theta) \

$$p(y = 2 \mid \mathbf{x} ; \theta) \\
\vdots \\
p(y = K \mid \mathbf{x} ; \theta) \end{bmatrix},$$

   where $$p(y = i \mid \mathbf{x}) = \phi_i = p(y = i \mid \mathbf{x}; \theta) = \frac{e^{\theta^\top_i \mathbf{x}}}{\sum_{j=1}^{K}e^{\theta^\top_j \mathbf{x}}}. $$
3. The cost function is $$J(\theta) = - \sum_{i=1}^m \log p(y^{(i)} \mid \textbf{x}^{(i)})$$
4. The optimization algorithm is gradient descent on $J(\theta)$ with the update rule $$\theta_{kl}^{(n+1)} \leftarrow \theta_{kl}^{(n)} - \alpha \sum_{i=1}^m (\delta(y^{(i)}=k) - \phi_k) x^{(i)}_l.$$

Multinomial Regression Example

The following example of multinomial logistic regression is from Kaggle (https://www.kaggle.com/saksham219/softmax-regression-for-iris-classification).

The data set is the famous Iris dataset from the UCI machine learning repository (https://archive.ics.uci.edu/ml/datasets/iris).

The data contain 50 samples from each of three classes. Each class refers to a particular species of the iris plant. The data include four independent variables:

1. Sepal length in cm
2. Sepal width in cm
3. Petal length in cm
4. Petal width in cm

The target takes on one of three classes:

1. Iris Setosa
2. Iris Versicolour
3. Iris Virginica

To predict the target value, we use multinomial logistic regression for $k=3$ classes i.e. $y \in \{ 1, 2, 3 \}$.

Given $\mathbf{x}$, we would like to predict a probability distribution over the three outcomes for y, i.e., $\phi_1 = p(y=1 \mid \mathbf{x})$, $\phi_2 = p(y=2 \mid \mathbf{x})$, and $\phi_3 = p(y=3 \mid \mathbf{x})$.

In [2]:

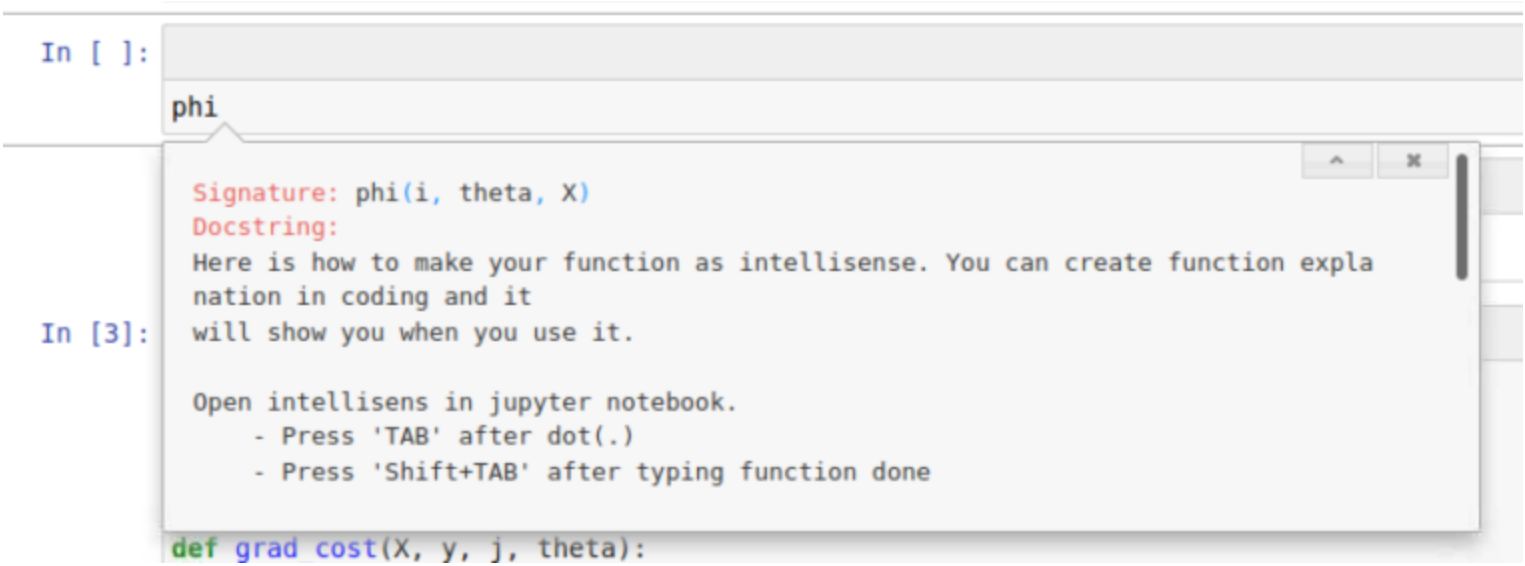
```
# importing libraries
import numpy as np
import pandas as pd
import random
import math
```

The `phi` function returns ϕ_i for input patterns $\mathtt{X}$ and parameters θ.

In [3]:

```
def phi(i, theta, X, num_class):
    """
    Here is how to make documentation for your function show up in intellisense.
    Explanation you put here will be shown when you use it.

    To get intellisense in your Jupyter notebook:
        - Press 'TAB' after typing a dot (.) to see methods and attributes
        - Press 'Shift+TAB' after typing a function name to see its documentation

    The `phi` function returns phi_i = h_theta(x) for input patterns X and parameters theta.

    Inputs:
        i=index of phi

        X=input dataset

        theta=parameters

    Returns:
        phi_i
    """
    mat_theta = np.matrix(theta[i])
    mat_x = np.matrix(X)
    num = math.exp(np.dot(mat_theta, mat_x.T))
    den = 0
    for j in range(0,num_class):
        mat_theta_j = np.matrix(theta[j])
        den = den + math.exp(np.dot(mat_theta_j, mat_x.T))
    phi_i = num / den
    return phi_i
```

Tips for using intellisense: Shift+TAB



The `grad_cost` function gives the gradient of the cost for data $\mathtt{X}, \mathbf{y}$ for class $j\in 1..k$.

In [4]:

```
def indicator(i, j):
    '''
    Check whether i is equal to j

    Return:
        1 when i=j, otherwise 0
    '''
    if i == j: return 1
    else: return 0


def grad_cost(X, y, j, theta, num_class):
    '''
    Compute the gradient of the cost function for data X, y for parameters of
    output for class j in 1..k
    '''
    m, n = X.shape
    sum = np.array([0 for i in range(0,n)])
    for i in range(0, m):
        p = indicator(y[i], j) - phi(j, theta, X.loc[i], num_class)
        sum = sum + (X.loc[i] * p)
    grad = -sum / m
    return grad

def gradient_descent(X, y, theta, alpha, iters, num_class):
    '''
    Perform iters iterations of gradient descent: theta_new = theta_old - alpha * cost
    '''
    n = X.shape[1]
    for iter in range(iters):
        dtheta = np.zeros((num_class, n))
        for j in range(0, num_class):
            dtheta[j,:] = grad_cost(X, y, j, theta, num_class)
        theta = theta - alpha * dtheta
    return theta

def h(X, theta, num_class):
    '''
    Hypothesis function: h_theta(X) = theta * X
    '''
    X = np.matrix(X)
    h_matrix = np.empty((num_class,1))
    den = 0
    for j in range(0, num_class):
        den = den + math.exp(np.dot(theta[j], X.T))
```

```
    for i in range(0,num_class):
        h_matrix[i] = math.exp(np.dot(theta[i], X.T))
    h_matrix = h_matrix / den
    return h_matrix
```

Exercise 1.1 (5 points)

Create a function to load **data** from **Iris.csv** using the Pandas library and extract y from the data.

You can use the Pandas 10 minute guide (https://pandas.pydata.org/pandas-docs/stable/user_guide/10min.html) to learn how to use pandas.

In [34]:

```
def load_data(file_name, drop_label, y_label, is_print=False):
    # 1. Load csv file
    data = pd.read_csv(file_name)
    if is_print:
        print(data.head())
    # 2. remove 'Id' column from data
    if drop_label is not None:
        data = data.drop([drop_label],axis=1)
        if is_print:
            print(data.head())
    # 3. Extract y_label column as y from data
    y = data[y_label]
    # 4. get index of y-column
    y_index = data.columns.get_loc(y_label)
    print(y_index)
    # 5. Extrack X features from data
    X = data.iloc[:,0:y_index]
    # YOUR CODE HERE
    #raise NotImplementedError()
    return X, y
```

In [35]:

```
X, y = load_data('Iris.csv', 'Id', 'Species', True)
print(X.head())
print(y[:5])

# Test function: Do not remove
# tips: this is how to create dataset using pandas
d_ex = {'ID':     [  1,   2,   3,    4,    5,    6,    7],
        'Grade':  [3.5, 2.5, 3.0, 3.75, 2.83, 3.95, 2.68],
        'Type':   ['A', 'B', 'C',  'A',  'C',  'A',  'B']
        }
df = pd.DataFrame (d_ex, columns = ['ID','Grade', 'Type'])
df.to_csv('out.csv', index=False)

Xtest, ytest = load_data('out.csv', 'ID', 'Type')
assert len(Xtest.columns) == 1, 'number of X_columns incorrect (1)'
assert ytest.name == 'Type', 'Extract y_column is incorrect (1)'
assert ytest.shape == (7,), 'number of y is incorrect (1)'
assert 'Grade' in Xtest.columns, 'Incorrect columns in X (1)'
Xtest, ytest = load_data('out.csv', None, 'Type')
assert len(Xtest.columns) == 2, 'number of X_columns incorrect (2)'
assert ytest.name == 'Type', 'Extract y_column is incorrect (2)'
assert ytest.shape == (7,), 'number of y is incorrect (2)'
assert 'Grade' in Xtest.columns and 'ID' in Xtest.columns, 'Incorrect columns in X (2)'
import os
os.remove('out.csv')

assert len(X.columns) == 4, 'number of X_columns incorrect (3)'
assert 'SepalWidthCm' in X.columns and 'Id' not in X.columns and 'Species' not in X.columns, 'Incorrect columns in X (3)'
assert y.name == 'Species', 'Extract y_column is incorrect (3)'
assert y.shape == (150,), 'number of y is incorrect (3)'

print("success!")
# End Test function
```

```
   Id  SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm      Species
0   1            5.1           3.5            1.4           0.2  Iris-setosa
1   2            4.9           3.0            1.4           0.2  Iris-setosa
2   3            4.7           3.2            1.3           0.2  Iris-setosa
3   4            4.6           3.1            1.5           0.2  Iris-setosa
4   5            5.0           3.6            1.4           0.2  Iris-setosa
   SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm      Species
0            5.1           3.5            1.4           0.2  Iris-setosa
1            4.9           3.0            1.4           0.2  Iris-setosa
2            4.7           3.2            1.3           0.2  Iris-setosa
3            4.6           3.1            1.5           0.2  Iris-setosa
4            5.0           3.6            1.4           0.2  Iris-setosa
4
   SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm
0            5.1           3.5            1.4           0.2
1            4.9           3.0            1.4           0.2
2            4.7           3.2            1.3           0.2
3            4.6           3.1            1.5           0.2
4            5.0           3.6            1.4           0.2
0    Iris-setosa
1    Iris-setosa
2    Iris-setosa
3    Iris-setosa
4    Iris-setosa
Name: Species, dtype: object
1
2
success!
```

In [36]:

```
print(X.shape)
```

```
(150, 4)
```

Expected result: \ SepalLengthCm SepalWidthCm PetalLengthCm PetalWidthCm \ 0 5.1 3.5 1.4 0.2\ 1 4.9 3.0 1.4 0.2\ 2 4.7 3.2 1.3 0.2\ 3 4.6 3.1 1.5 0.2\ 4 5.0 3.6 1.4 0.2\ 0 Iris-setosa\ 1 Iris-setosa\ 2 Iris-setosa\ 3 Iris-setosa\ 4 Iris-setosa\ Name: Species, dtype: object

Exercise 1.2 (10 points)

Partition data into training and test sets

- No need to use random.seed function!
- Ensure that the train set is 70% and the test set is 30% of the data.
- Encode the labels in the y attribute to be integers in the range 0..k-1.

Hint:

In [44]:

```
import random
def partition(X, y, percent_train):
    m, n = X.shape

 # 1. create index list
    idx = np.arange(0, m)

    # 2. shuffle index
    random.shuffle(idx)
    training = int(round(m*percent_train))

    # 3. Create train/test index
    train_idx = idx[0:training]
    test_idx = idx[training:]

    # 4. Separate X_Train, y_train, X_test, y_test
    X_train = X.iloc[train_idx,:]
    X_test = X.iloc[test_idx,:]

    y_train = y[train_idx]
    y_test = y[test_idx]

    # 5. Get y_labels_name from y using pandas.unique function
    y_labels_name = pd.unique(y)

    # 6. Change y_labels_name into string number and put into y_labels_new
    for i in range(len(y_labels_name)):
        y_train[y_train.index[y_train == y_labels_name[i]]]=i
        y_test[y_test.index[y_test == y_labels_name[i]]]=i

    # 7. Drop shuffle index columns
    # - pandas.reset_index() and pandas.drop(...) might be help
    X_train = X_train.reset_index()
    X_train = X_train.drop(['index'],axis=1)

    X_test = X_test.reset_index()
    X_test = X_test.drop(['index'],axis=1)

    y_train = y_train.reset_index()
    y_train = y_train.iloc[:,1]

    y_test = y_test.reset_index()
    y_test = y_test.iloc[:,1]

    y_labels_new = pd.unique(y_train)
```

```
#     y_labels_name = None
#     y_labels_new = None

    # YOUR CODE HERE
    #raise NotImplementedError()

    return idx, X_train, y_train, X_test, y_test, y_labels_name, y_labels_new
```

In [45]:

```
percent_train = 0.7
idx, X_train, y_train, X_test, y_test, y_labels_name, y_labels_new = partition(X, y, percent_train)
print('X_train.shape', X_train.shape)
print('X_test.shape', X_test.shape)
print('y_train.shape', y_train.shape)
print('y_test.shape', y_test.shape)
print('y_labels_name: ', y_labels_name)
print('y_labels_new: ', y_labels_new)
print(X_train.head())
print(y_train.head())

# Test function: Do not remove
assert len(y_labels_name) == 3 and len(y_labels_new) == 3, 'number of y uniques are incorrect'
assert X_train.shape == (105, 4), 'Size of X_train is incorrect'
assert X_test.shape == (45, 4), 'Size of x_test is incorrect'
assert y_train.shape == (105, ), 'Size of y_train is incorrect'
assert y_test.shape == (45, ), 'Size of y_test is incorrect'
assert 'Iris-setosa' in y_labels_name and 'Iris-virginica' in y_labels_name and \
        'Iris-versicolor' in y_labels_name, 'y unique data incorrect'
assert min(y_labels_new) == 0 and max(y_labels_new) < 3, 'label indices are incorrect'

print("success!")
# End Test function
```

```
X_train.shape (105, 4)
X_test.shape (45, 4)
y_train.shape (105,)
y_test.shape (45,)
y_labels_name:  ['Iris-setosa' 'Iris-versicolor' 'Iris-virginica']
y_labels_new:  [1 0 2]
   SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm
0            5.8           2.6            4.0           1.2
1            5.0           3.4            1.5           0.2
2            5.5           2.3            4.0           1.3
3            4.9           3.1            1.5           0.1
4            5.0           3.3            1.4           0.2
0    1
1    0
2    1
3    0
4    0
Name: Species, dtype: object
success!
```

Expected result: (*or similar*)\ X_train.shape (105, 4)\ X_test.shape (45, 4)\ y_train.shape (105,)\ y_test.shape (45,)\ y_labels_name: ['Iris-setosa' 'Iris-versicolor' 'Iris-virginica'] \ y_labels_new: [0, 1, 2]

SepalLengthCm SepalWidthCm PetalLengthCm PetalWidthCm\ 0 6.4 2.8 5.6 2.2\ 1 6.7 3.3 5.7 2.1\ 2 4.6 3.4 1.4 0.3\ 3 5.1 3.8 1.5 0.3\ 4 5.0 2.3 3.3 1.0\ Species\ 0 2\ 1 2\ 2 0\ 3 0\ 4 1

Exercise 1.3 (5 points)

Train your classification model using the `gradient_descent` function already provided. You might also play around with the gradient descent function to see if you can speed it up!

In [46]:

```
# num_class is the number of unique labels
num_class = len(y_labels_name)

if (X_train.shape[1] == X.shape[1]):
    X_train.insert(0, "intercept", 1)

# Reset m and n for training data
r, c = X_train.shape

# Initialize theta for each class
theta_initial = np.ones((num_class, c))

alpha = .05
iterations = 200

theta = gradient_descent(X_train, y_train, theta_initial, alpha, iterations, num_class)
# Logistic regression
# YOUR CODE HERE
#raise NotImplementedError()
```

In [47]:

```
print(theta)
print(theta.shape)

# Test function: Do not remove
assert theta.shape == (3, 5), 'Size of theta is incorrect'

print("success!")
# End Test function
```

```
[[ 1.16712636  1.32531861  1.81783852 -0.18915766  0.44533094]
 [ 1.08059326  1.17041205  0.74423293  1.18339319  0.81675929]
 [ 0.75228038  0.50426934  0.43792855  2.00576448  1.73790977]]
(3, 5)
success!
```

Expected result: (*or similar*)\ [[1.17632192 1.32360047 1.83204165 -0.20224445 0.44039155]\ [1.10140069 1.13537321 0.74833178 1.21907866 0.82567377]\ [0.72227738 0.54102632 0.41962657 1.98316579 1.73393467]]\ \ (3, 5)

Exercise 1.4 (5 points)

Let's get your model to make predictions on the test data.

In [48]:

```
# Prediction on test data

if (X_test.shape[1] == X.shape[1]):
    X_test.insert(0, "intercept", 1)

# Reset m and n for test data
r,c = X_test.shape

y_pred = []
for index,row in X_test.iterrows(): # get a row of X_test data
    # calculate y_hat using hypothesis function
    y_hat = h(row, theta, num_class)
    # find the index (integer value) of maximum value in y_hat and input back to prediction
    prediction = int(np.argmax(y_hat, axis = 0))
    # YOUR CODE HERE
    #raise NotImplementedError()
    # collect the result
    y_pred.append(prediction)
```

In [49]:

```
print(len(y_pred))
print(y_pred[:7])
print(type(y_pred[0]))

# Test function: Do not remove
assert len(y_pred) == 45, 'Size of y_pred is incorrect'
assert isinstance(y_pred[0], int) and isinstance(y_pred[15], int) and isinstance(y_pred[17], int), 'prediction type is incorrect'
assert max(y_pred) < 3 and min(y_pred) >= 0, 'wrong index of y_pred'

print("success!")
# End Test function
```

```
45
[0, 1, 2, 0, 1, 1, 0]
<class 'int'>
success!
```

Expected result: (*or similar*)\ 45 \ [2, 0, 2, 0, 0, 0, 2] \

<class 'int'>

Exercise 1.5 (5 points)

Estimate accuracy of model on test data
$$\text{accuracy} = \frac{\text{number of correct test predictions}}{m_{\text{test}}}$$

In [50]:

```
def calc_accuracy(y_test, y_pred):
    accuracy = (y_pred == y_test).astype(int).sum() / y_test.shape[0]
    # YOUR CODE HERE
    #raise NotImplementedError()
    return accuracy
```

In [51]:

```
accuracy = calc_accuracy(y_test, y_pred)
print('Accuracy: %.4f' % accuracy)

# Test function: Do not remove
assert isinstance(accuracy, float), 'accuracy should be floating point'
assert accuracy >= 0.8, 'Did you train the data?'

print("success!")
# End Test function
```

```
Accuracy: 0.9778
success!
```

Expected result: should be at least 0.8!

On your own in lab

We will do the following in lab:

1. Write a function to obtain the cost for particular $\mathtt{X}$, $\mathbf{y}$, and θ.
2. Plot the training set and test cost as training goes on and find the best value for the number of iterations and learning rate.
3. Make 2D scatter plots showing the predicted and actual class of each item in the training set, plotting two features at a time. Comment on the cause of the errors you observe. If you obtain perfect test set accuracy, re-run the train/test split and rerun the optimization until you observe some mistaken predictions on the test set.

Exercise 2.1 (15 points)

1. Write a function to obtain the cost for particular $\mathtt{X}$, $\mathbf{y}$, and θ. Name your function `my_J()` and implement

$$ J_j = -\delta(y, j)\log{\phi_j} $$

In [52]:

```
def my_J(theta, X, y, j, num_class):
    cost = -indicator(y,j)*(np.log(phi(j, theta, X, num_class)))

    # YOUR CODE HERE
    #raise NotImplementedError()
    return cost
```

In [53]:

```
# Test function: Do not remove
m, n = X_train.shape
test_theta = np.ones((3, n))
cost = my_J(test_theta, X_train.loc[10], y_train[10], 0, 3)
assert isinstance(cost, float), 'cost should be floating point'

print("success!")
# End Test function
```

```
success!
```

1. Implement `my_grad_cost` using your `my_J` function

In [55]:

```
def my_grad_cost(X, y, j, theta, num_class):
    cost = 0
    sum = np.array([0 for i in range(0,n)])
    for i in range(0, m):
        p = indicator(y[i], j) - phi(j, theta,X.loc[i], num_class)
        sum = sum + (X.loc[i] * p)
        # Add all the costs for each i
        cost = cost + my_J(theta, X.loc[i], y[i], j, num_class)
    grad = -sum/m

    return grad, cost
```

In [56]:

```
# Test function: Do not remove
m, n = X_train.shape
test_theta = np.ones((3, n))
grad, cost = my_grad_cost(X_train, y_train, 0, test_theta, num_class)
print(grad)
print(cost)
assert isinstance(cost, float), 'cost should be floating point'
assert isinstance(grad['intercept'], float) and \
        isinstance(grad['SepalLengthCm'], float) and \
        isinstance(grad['SepalWidthCm'], float) and \
        isinstance(grad['PetalLengthCm'], float) and \
        isinstance(grad['PetalWidthCm'], float) , 'grad should be floating point'
print("success!")
# End Test function
```

```
intercept        0.009524
SepalLengthCm    0.325714
SepalWidthCm    -0.094286
PetalLengthCm    0.797143
PetalWidthCm     0.330794
dtype: float64
37.352817814715735
success!
```

Expect result: (*or similar*)\ intercept 0.009524\ SepalLengthCm 0.316825\ SepalWidthCm -0.091429\ PetalLengthCm 0.780000\ PetalWidthCm 0.329524\ dtype: float64\ 37.352817814715735

1. Implement `my_gradient_descent` using your `my_grad_cost` function

In [57]:

```
def my_gradient_descent(X, y, theta, alpha, iters, num_class):
    cost_arr = []
    cost_arr = []
    for iter in range(iters):
        cost = 0
        for j in range(0, len(y.unique())):
            grad, cost_obtained = my_grad_cost(X, y, j, theta, num_class)
            theta[j] = theta[j] - alpha * grad
            # Add the costs
            cost = cost + cost_obtained
    # Get cost for every iter
        cost_arr.append(cost)
    return theta, cost_arr
    # YOUR CODE HERE
    #raise NotImplementedError()
    #return theta, cost_arr
```

In [58]:

```
# Test function: Do not remove
m, n = X_train.shape
test_theta = np.ones((3, n))
theta, cost = my_gradient_descent(X_train, y_train, theta_initial, 0.001, 5, 3)
print(theta)
print(cost)
print("success!")
# End Test function
```

```
[[0.99996878 0.9984716  1.00052021 0.99608766 0.99837065]
 [0.99995019 0.99982418 0.99939372 1.00055538 1.00010879]
 [1.0000756  1.00167113 1.00006987 1.00333289 1.0015125 ]]
[115.20703774834786, 115.0462527072857, 114.88936280214679, 114.7362104410744, 114.58664437027522]
success!
```

Expected result: (*or similar*)\ [[1.00001186 0.99618853 1.00183642 0.9889817 0.99528923]\ [1.00009697 1.0011823 0.99883395 1.00316763 1.00083055]\ [0.99987915 1.00255606 0.99929351 1.00779768 1.00386218]]\ [114.00099216453735, 113.89036233839263, 113.78163144339288, 113.67472269747496, 113.56956268162737]\ 37.352817814715735

Exercise 2.2 (20 points)

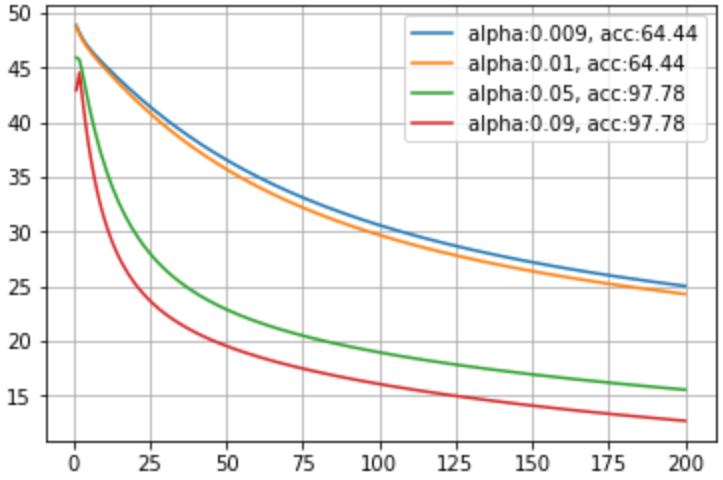
1. Plot the training set and test cost as training goes on and find the best value for the number of iterations and learning rate.
2. Make 2D scatter plots showing the predicted and actual class of each item in the training set, plotting two features at a time. Comment on the cause of the errors you observe. If you obtain perfect test set accuracy, re-run the train/test split and rerun the optimization until you observe some mistaken predictions on the test set.

In [59]:

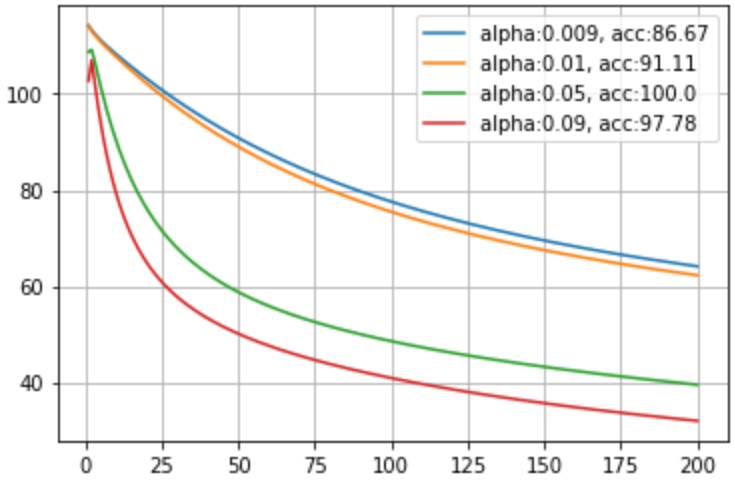
```
import matplotlib.pyplot as plt
```

In [62]:

```
theta_arr = []
cost_arr = []
accuracy_arr = []

# design your own learning rate and num iterations
alpha_arr = np.array([0.009, 0.01, 0.05, 0.09])
iterations_arr = np.array([200, 200, 200, 200])

# YOUR CODE HERE
m, n = X_test.shape
fig = plt.figure()
ax = plt.axes()
plt.grid(axis='both')
for i in range(0, len(alpha_arr)):
    theta_initial = np.ones((num_class, n))
    theta, cost = my_gradient_descent(X_train, y_train, theta_initial, alpha_arr[i], iterations_arr[i], num_class)

    y_pred = []
    for index,row in X_test.iterrows():
        h_matrix = h(row, theta, num_class)
        prediction = int(np.where(h_matrix == h_matrix.max())[0])
        y_pred.append(prediction)

    correct = (y_pred == y_test).value_counts()[True]
    accuracy = correct/m
    plt.plot(range(1,iterations_arr[i]+1), cost, label='alpha:'+str(alpha_arr[i]) +', acc:' + str(np.round(accuracy,4)*100))
    accuracy_arr.append(accuracy)

plt.legend()
plt.show()
```



Expected result: (*Yours doesn't have to be the same!*)



On your own to take home

We see that the Iris dataset is pretty easy. Depending on the train/test split, we get 95-100% accuracy.

Find a more interesting multi-class classification problem on Kaggle (Tell the reference), clean the dataset to obtain numerical input features without missing values, split the data into test and train, and experiment with multinomial logistic regression.

Write a brief report on your experiments and results. As always, turn in a Jupyter notebook by email to the instructor and TA.

I took the dataset from https://www.kaggle.com/datasets/deepu1109/star-dataset (https://www.kaggle.com/datasets/deepu1109/star-dataset) The dataset is about classify stars by plotting its features based on that graph.This is a dataset consisting of several features of stars.and there are 240 data.

In [65]:

```
def load_data(file_name, drop_label, is_print=False):

    data = pd.read_csv(file_name)

    if is_print:
        print(data.head())

    if drop_label is not None:
        data = data.drop(column=[drop_label], axis=1)
        if is_print:
            print(data.head())

    return data
```

In [66]:

```
data = load_data('Stars.csv', drop_label=None, is_print=True)
```

```
   Temperature         L       R    A_M Color Spectral_Class  Type
0         3068  0.002400  0.1700  16.12   Red              M     0
1         3042  0.000500  0.1542  16.60   Red              M     0
2         2600  0.000300  0.1020  18.70   Red              M     0
3         2800  0.000200  0.1600  16.65   Red              M     0
4         1939  0.000138  0.1030  20.06   Red              M     0
```

In [67]:

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 240 entries, 0 to 239
Data columns (total 7 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Temperature     240 non-null    int64
 1   L               240 non-null    float64
 2   R               240 non-null    float64
 3   A_M             240 non-null    float64
 4   Color           240 non-null    object
 5   Spectral_Class  240 non-null    object
 6   Type            240 non-null    int64
dtypes: float64(3), int64(2), object(2)
memory usage: 13.2+ KB
```

In [68]:

```
data.isnull().sum()
```

Out[68]:

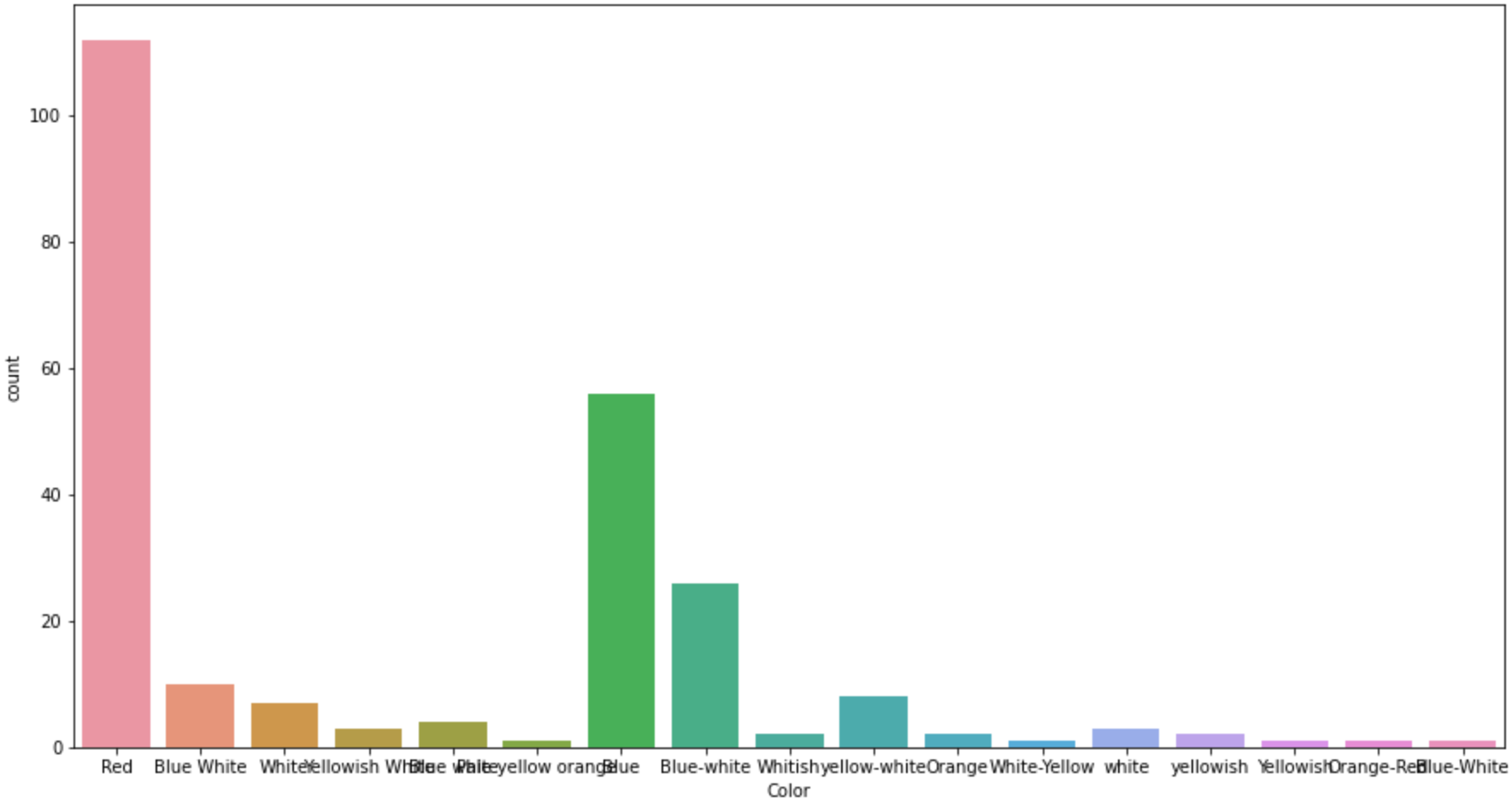
```
Temperature       0
L                 0
R                 0
A_M               0
Color             0
Spectral_Class    0
Type              0
dtype: int64
```

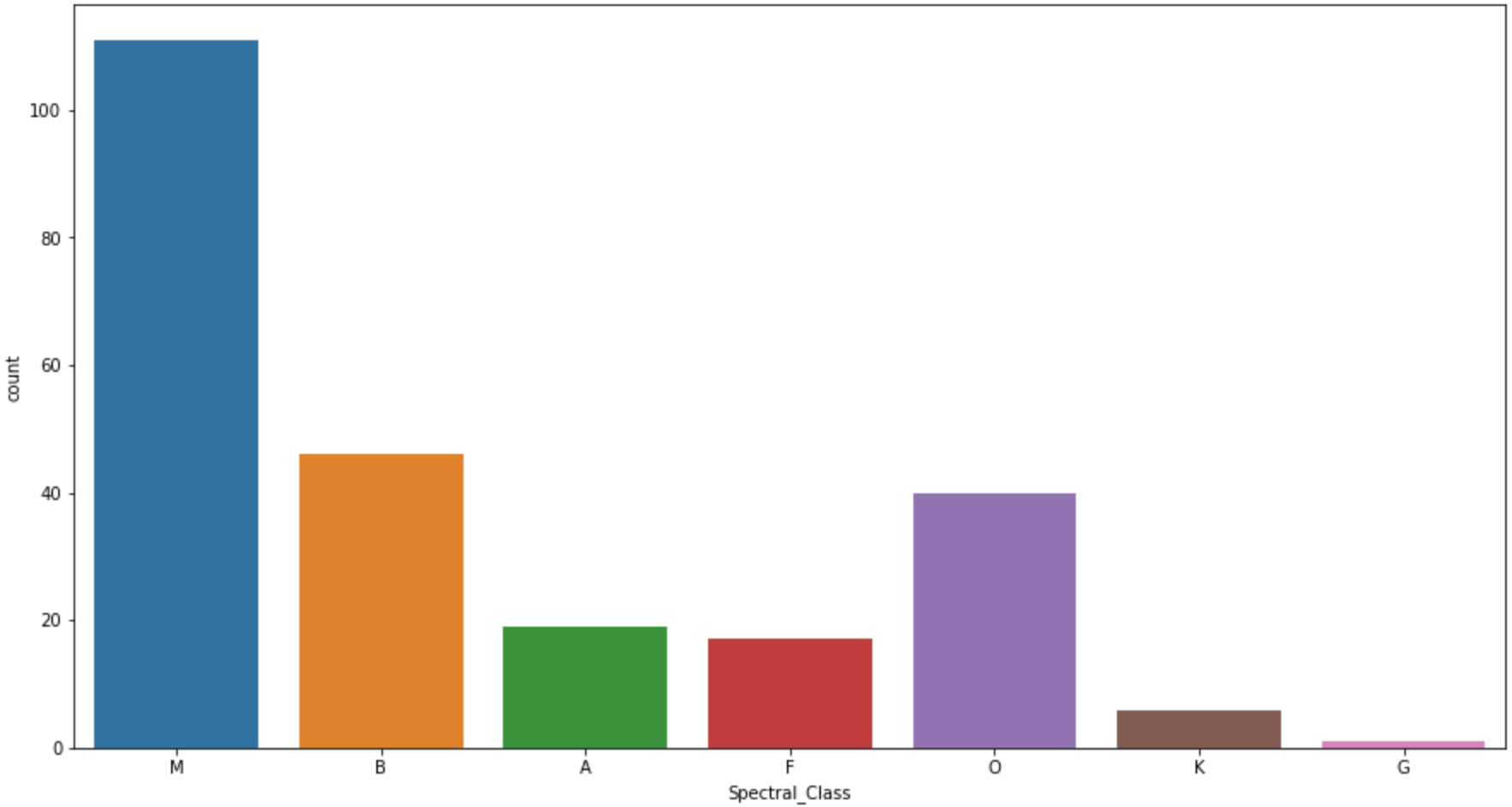
In [70]:

```
#lets mark the object and integer float64 seperately.
dis_col = data.select_dtypes(include=['object']).columns
#continuous column
con_col = data.select_dtypes(include=['int64', 'float64']).columns
```

In [75]:

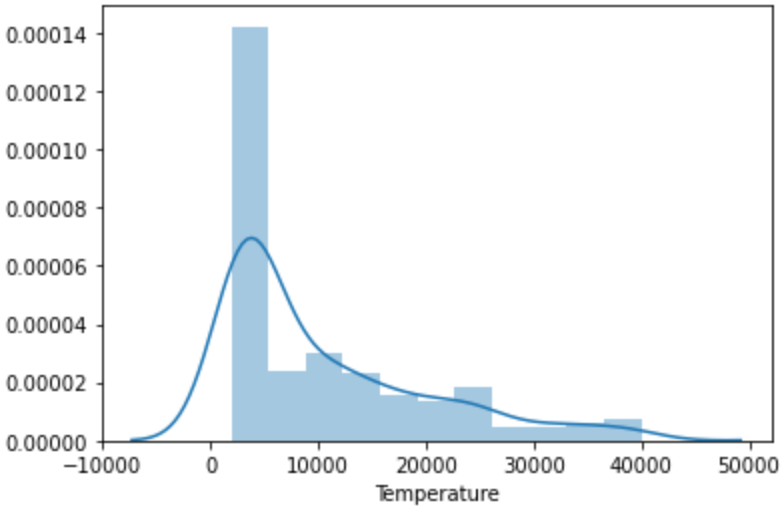
```
import seaborn as sns
for dis in dis_col:
    plt.figure(figsize=(15, 8))
    sns.countplot(x=dis,data=data)
    plt.show()
```

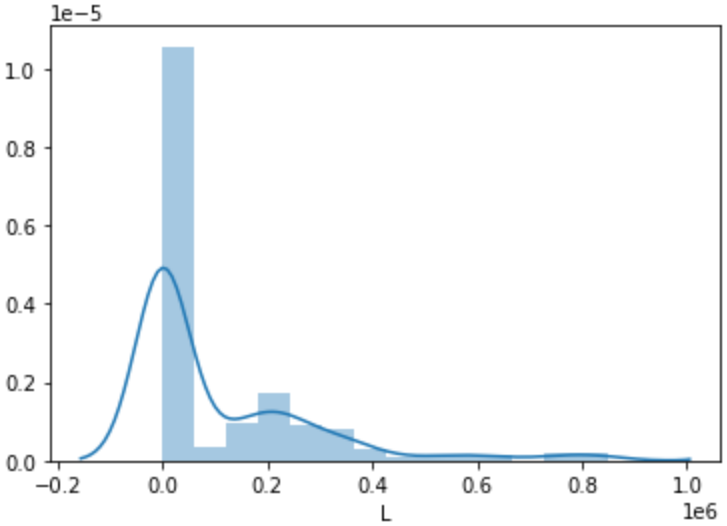

count
0
20
40
60
80
100
Red
Blue
White
Yellowish
Blue-white
Whitish
yellow-white
Orange
White-Yellow
white
yellowish
Yellowish
Orange-Red
Blue-White
Color

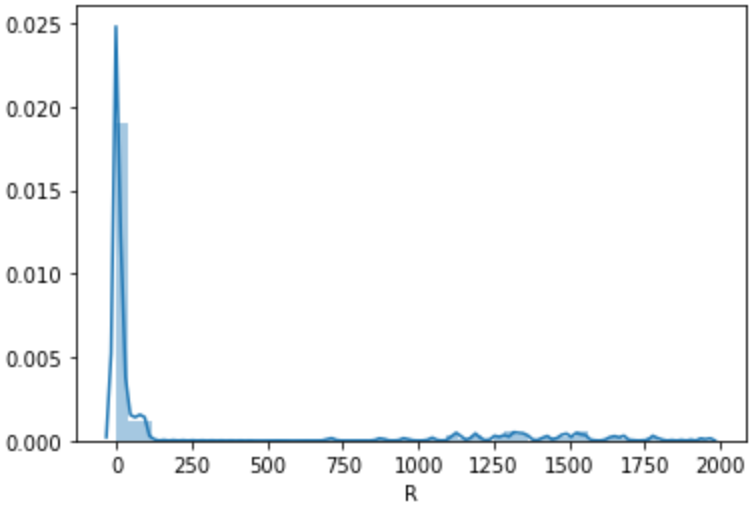

count
100
80
60
40
20
0
M
B
A
F
O
K
G
Spectral_Class

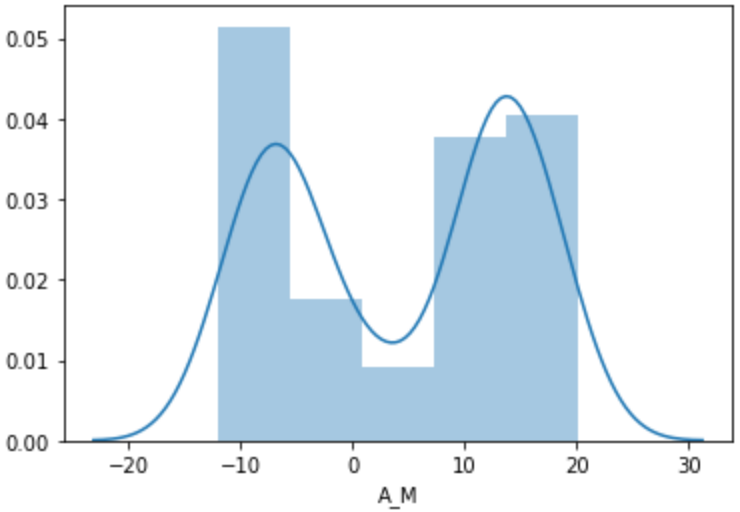
In [78]:

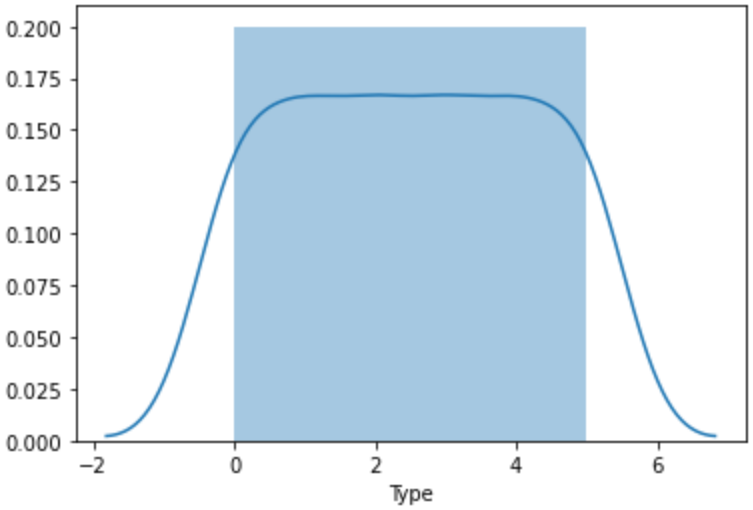
```
for col in con_col:
    sns.distplot(data[col])
    plt.show()
```


0.00014
0.00012
0.00010
0.00008
0.00006
0.00004
0.00002
0.00000
−10000
0
10000
20000
30000
40000
50000
Temperature


1e−5
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
1e6
L

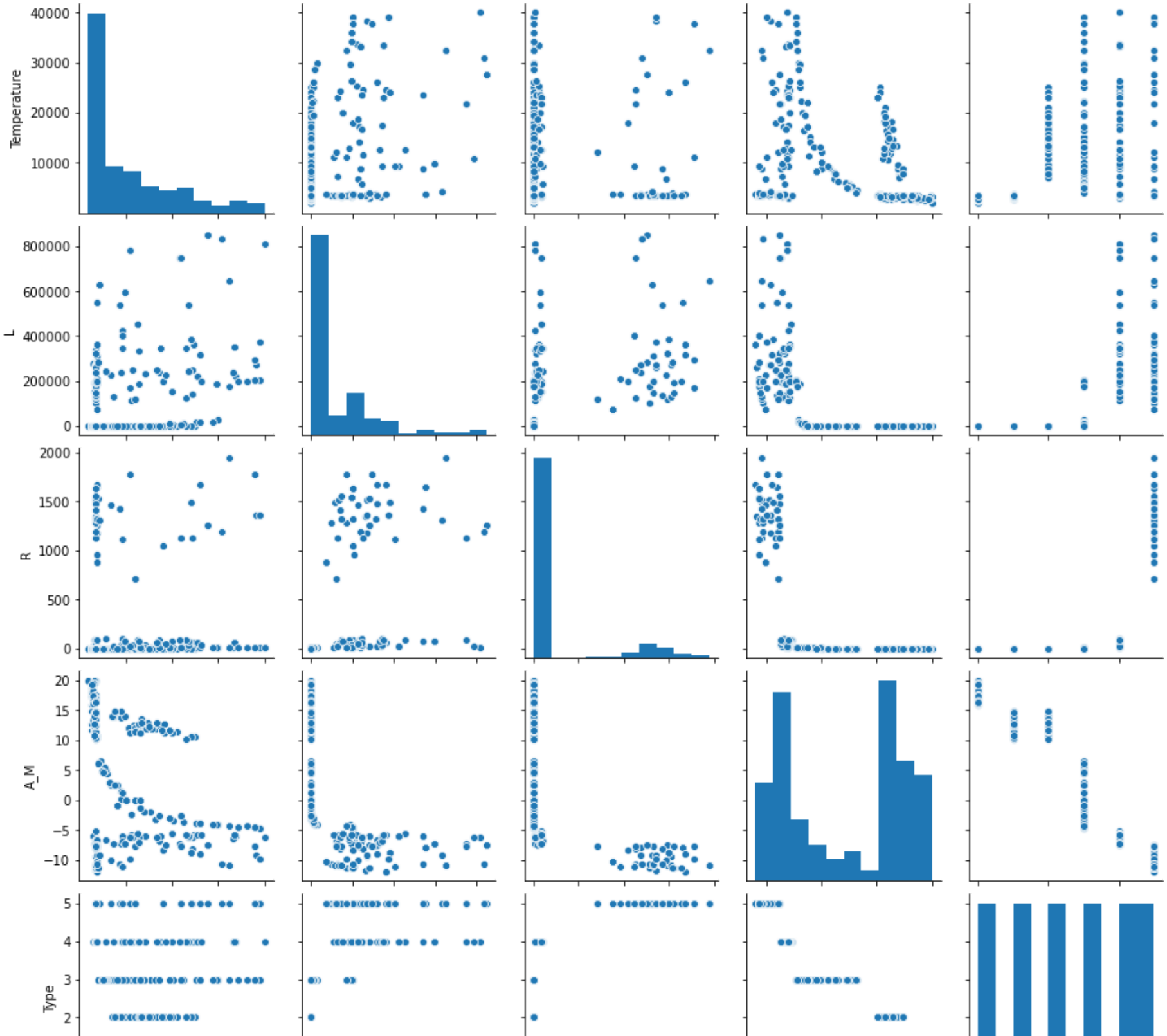

0.025
0.020
0.015
0.010
0.005
0.000
0
250
500
750
1000
1250
1500
1750
2000
R

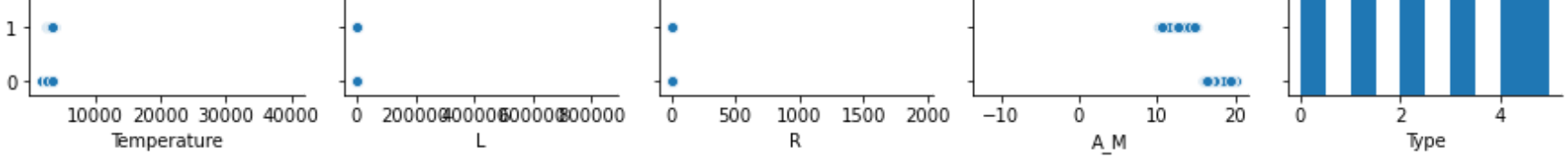

0.05
0.04
0.03
0.02
0.01
0.00
−20
−10
0
10
20
30
A_M


0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
−2
0
2
4
6
Type

In [79]:

```
import seaborn as sns
import matplotlib.pyplot as plt
sns.pairplot(data)
plt.show()
```


Temperature
L
R
A_M
Type
40000
30000
20000
10000
800000
600000
400000
200000
0
2000
1500
1000
500
0
20
15
10
5
0
−5
−10
5
4
3
2

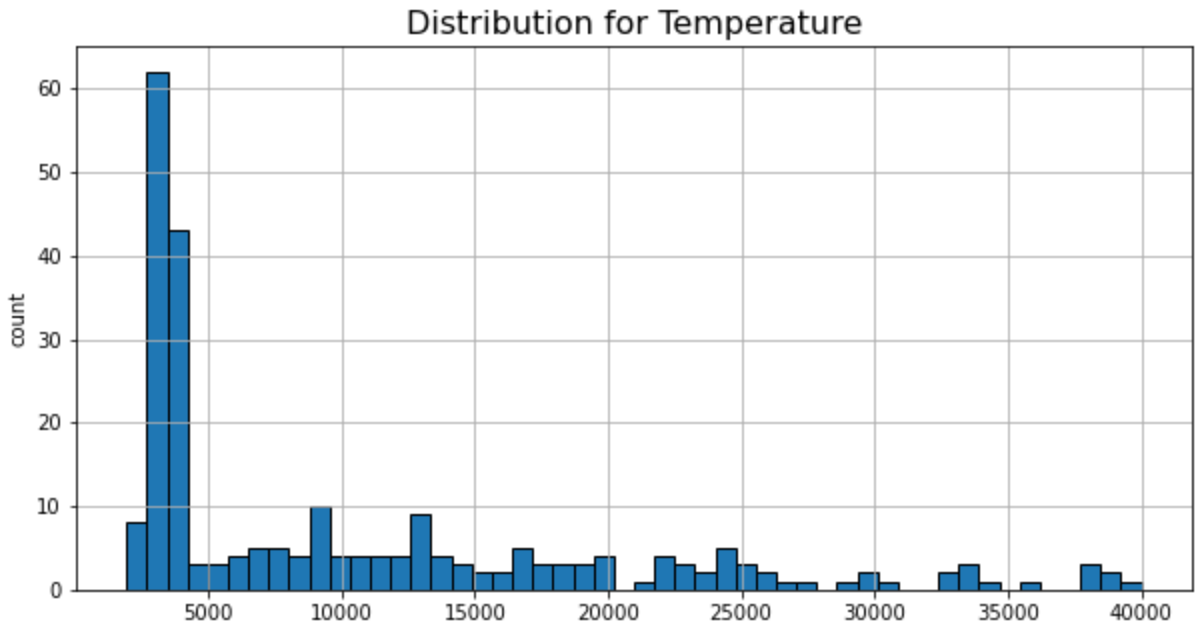


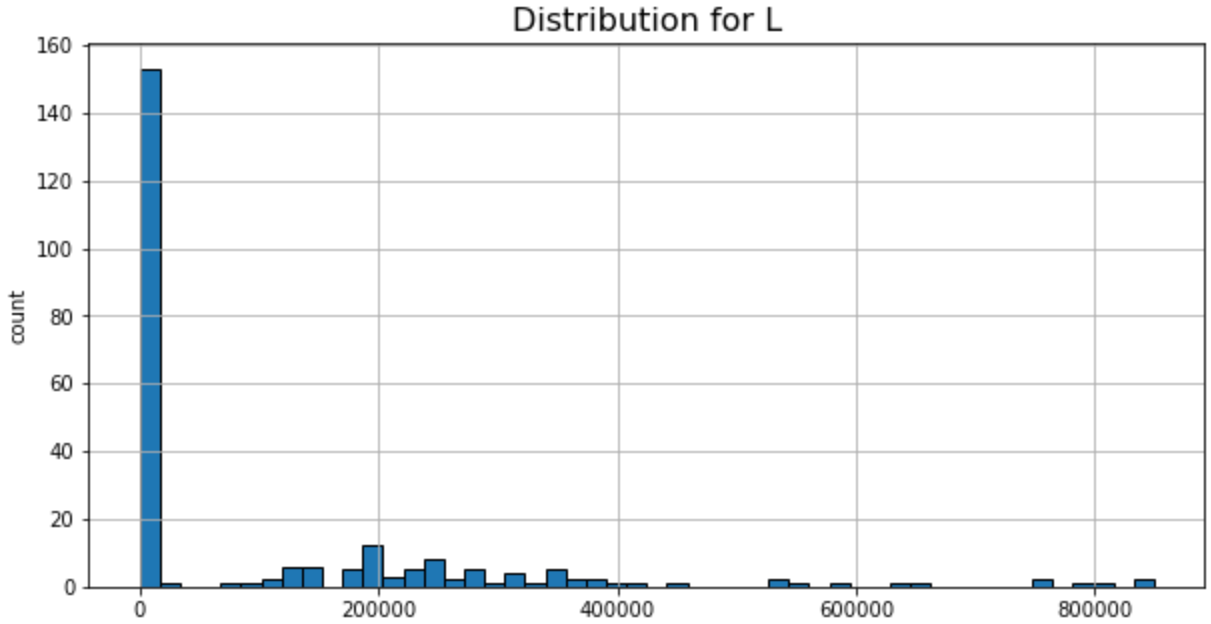
In [80]:

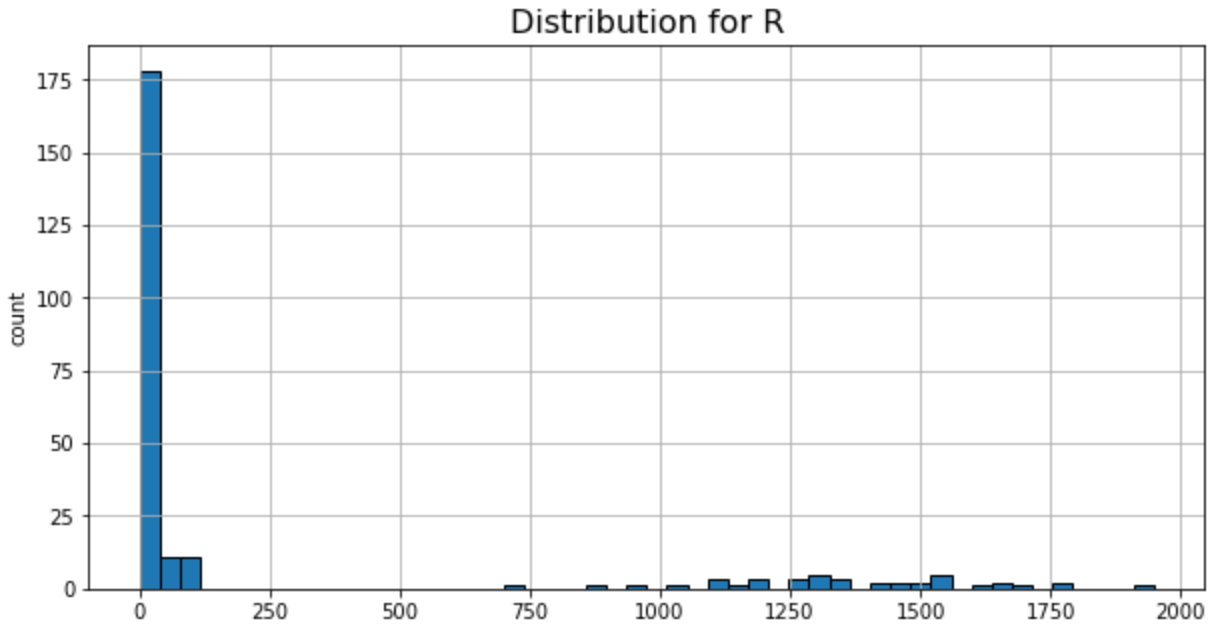
```
def histplot(data, name):

    plt.figure(figsize=(10,5))
    plt.grid()
    plt.hist(data[name], edgecolor='black', bins=50)
    plt.title(f"Distribution for {name}", size=16)
    plt.ylabel('count')
    plt.show()
```

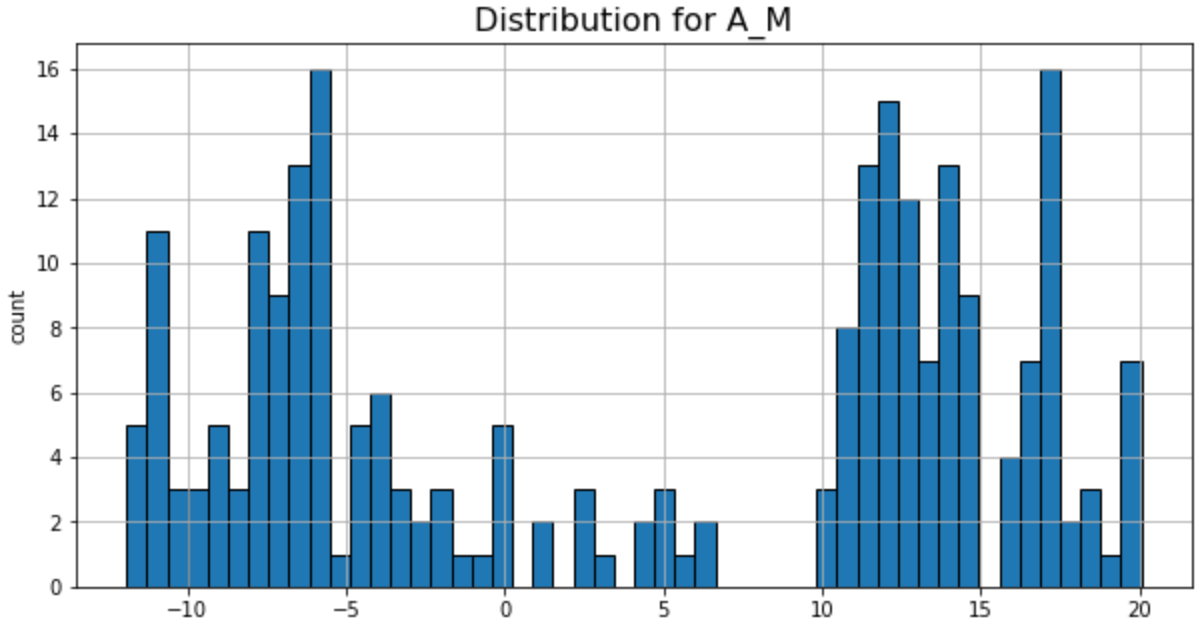
In [81]:

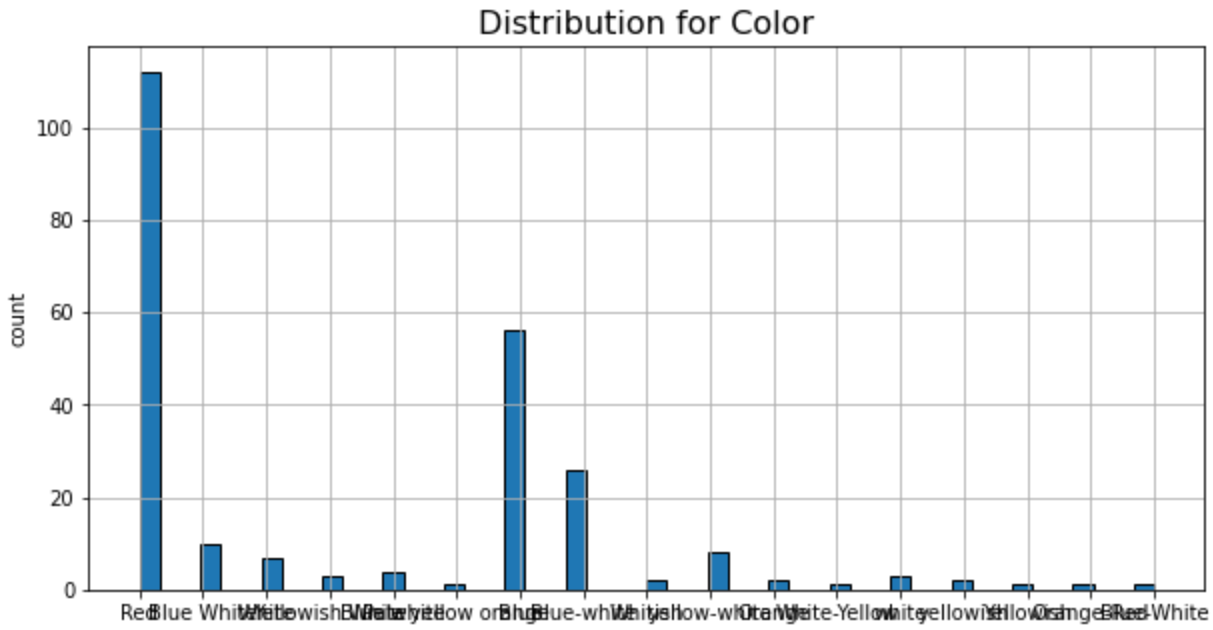
```
column_names = list(data.columns)
for name in column_names:
    histplot(data, name)
```

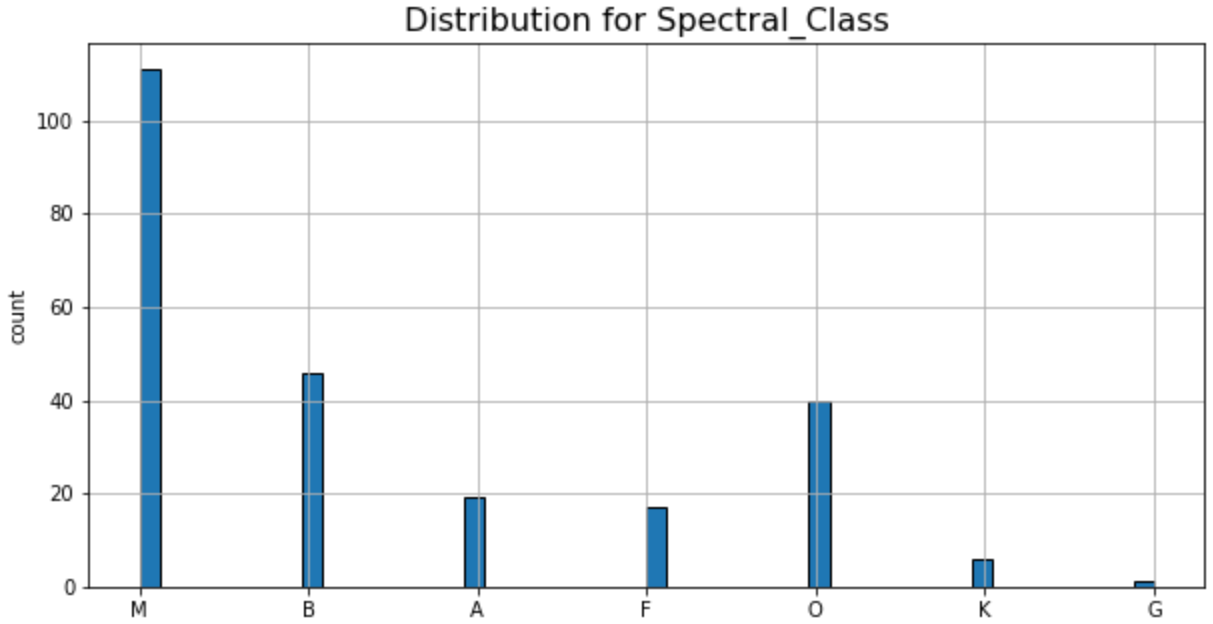

Distribution for Temperature
count
0
10
20
30
40
50
60
5000
10000
15000
20000
25000
30000
35000
40000

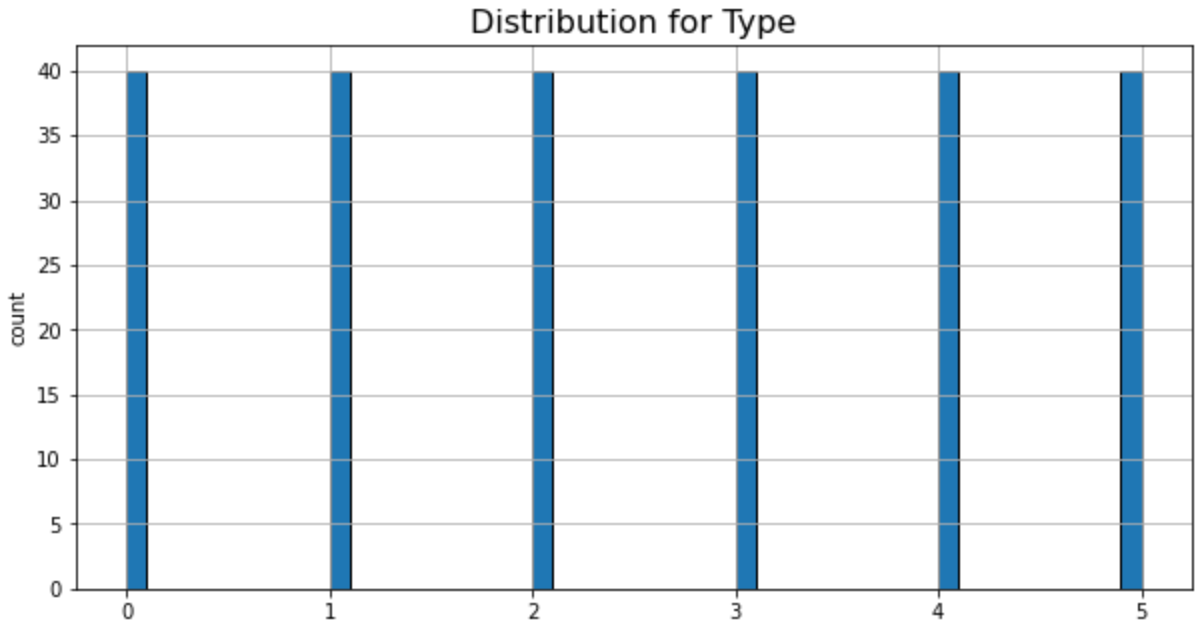

Distribution for L
count
0
20
40
60
80
100
120
140
160
0
200000
400000
600000
800000


Distribution for R
count
0
25
50
75
100
125
150
175
0
250
500
750
1000
1250
1500
1750
2000


Distribution for A_M
count
0
2
4
6
8
10
12
14
16
−10
−5
0
5
10
15
20


Distribution for Color
count
0
20
40
60
80
100
Red
Blue
Blue-white
White-Yellow
Orange
White


Distribution for Spectral_Class
count
0
20
40
60
80
100
M
B
A
F
O
K
G



Here, are two object value first, we will change categorial value to integers.

In [83]:

```
column_name = data['Color'].unique()
column_new = np.arange(len(column_name))
for i in range(len(column_name)):
    data['Color'] = data['Color'].replace(column_name[i], column_new[i])
```

In [84]:

```
print(data['Color'].unique())
print(len(data['Color'].unique()))
```

```
[ 0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16]
17
```

In [85]:

```
column_name = data['Spectral_Class'].unique()
column_new = np.arange(len(column_name))
for i in range(len(column_name)):
    data['Spectral_Class'] = data['Spectral_Class'].replace(column_name[i], column_new[i])
print(data['Spectral_Class'].unique())
print(len(data['Spectral_Class'].unique()))
```

```
[0 1 2 3 4 5 6]
7
```

In [86]:

```
#selecting X and y
y = data['Type']
y_index = data.columns.get_loc('Type')
X = data.iloc[:, :y_index]
print('X shape: ', X.shape)
print('Y shape: ', y.shape)
```

```
X shape:  (240, 6)
Y shape:  (240,)
```

In [87]:

```
def data_Norm(data):
    means = np.mean(data, axis=0)
    stds = np.std(data, axis=0)
    data_norm = (data - means) / stds
    return data_norm
X = data_Norm(X)
print(X.head())
```

```
   Temperature         L         R       A_M     Color  Spectral_Class
0    -0.779382 -0.598624 -0.459210  1.116745 -0.892015       -0.841613
1    -0.782110 -0.598624 -0.459241  1.162414 -0.892015       -0.841613
2    -0.828477 -0.598624 -0.459342  1.362213 -0.892015       -0.841613
3    -0.807496 -0.598624 -0.459229  1.167171 -0.892015       -0.841613
4    -0.897819 -0.598624 -0.459340  1.491607 -0.892015       -0.841613
```

In [88]:

```
def partition(X, y, percent_train):
    m, n = X.shape

 # 1. create index list
    idx = np.arange(0, m)

    # 2. shuffle index
    random.shuffle(idx)
    training = int(round(m*percent_train))

    # 3. Create train/test index
    train_idx = idx[0:training]
    test_idx = idx[training:]

    # 4. Separate X_Train, y_train, X_test, y_test
    X_train = X.iloc[train_idx,:]
    X_test = X.iloc[test_idx,:]

    y_train = y[train_idx]
    y_test = y[test_idx]

    # 5. Get y_labels_name from y using pandas.unique function
    y_labels_name = pd.unique(y)

    # 6. Change y_labels_name into string number and put into y_labels_new
    for i in range(len(y_labels_name)):
        y_train[y_train.index[y_train == y_labels_name[i]]]=i
        y_test[y_test.index[y_test == y_labels_name[i]]]=i

    # 7. Drop shuffle index columns
    # - pandas.reset_index() and pandas.drop(...) might be help
    X_train = X_train.reset_index()
    X_train = X_train.drop(['index'],axis=1)

    X_test = X_test.reset_index()
    X_test = X_test.drop(['index'],axis=1)

    y_train = y_train.reset_index()
    y_train = y_train.iloc[:,1]

    y_test = y_test.reset_index()
    y_test = y_test.iloc[:,1]

    y_labels_new = pd.unique(y_train)
```

```
#     y_labels_name = None
#     y_labels_new = None

    # YOUR CODE HERE
    #raise NotImplementedError()

    return idx, X_train, y_train, X_test, y_test, y_labels_name, y_labels_new
```

In [89]:

```
percent_train = 0.7
idx, X_train, y_train, X_test, y_test, y_labels_name, y_labels_new = partition(X, y, percent_train)
print('X_train.shape', X_train.shape)
print('X_test.shape', X_test.shape)
print('y_train.shape', y_train.shape)
print('y_test.shape', y_test.shape)
print('y_labels_name: ', y_labels_name)
print('y_labels_new: ', y_labels_new)
```

```
X_train.shape (168, 6)
X_test.shape (72, 6)
y_train.shape (168,)
y_test.shape (72,)
y_labels_name:  [0 1 2 3 4 5]
y_labels_new:  [2 3 5 0 1 4]
```

In [90]:

```
print(X_train.head())
```

```
   Temperature         L         R       A_M     Color  Spectral_Class
0     0.444221 -0.598624 -0.459522  0.809435  2.327287        0.989086
1     1.528721 -0.517644 -0.448068 -0.795619  0.985911       -0.231380
2    -0.723573  0.741724  1.846317 -1.167627 -0.892015       -0.841613
3    -0.897819 -0.598624 -0.459340  1.491607 -0.892015       -0.841613
4    -0.759555 -0.598624 -0.458648  0.603927 -0.892015       -0.841613
```

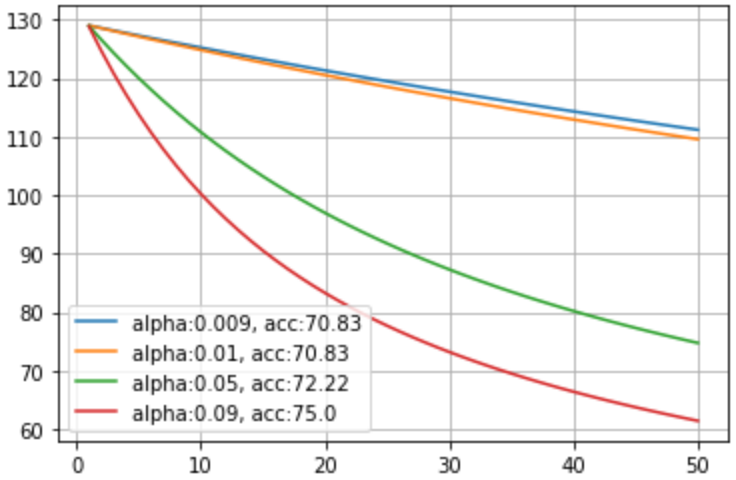
In [100]:

```
def phi(i, theta, X, num_class):
    mat_theta = np.matrix(theta[i])
    mat_x = np.matrix(X)
    num = math.exp(np.dot(mat_theta, mat_x.T))
    den = 0
    for j in range(0,num_class):
        mat_theta_j = np.matrix(theta[j])
        den = den + math.exp(np.dot(mat_theta_j, mat_x.T))
    phi_i = num / den
    return phi_i

def indicator(i, j):

    if i == j:
        return 1
    else:
        return 0

def grad_cost(X, y, j, theta, num_class):

    m, n = X.shape
    sum = np.array([0 for i in range(0,n)])
    for i in range(0, m):
        p = indicator(y[i], j) - phi(j, theta, X.loc[i], num_class)
        sum = sum + (X.loc[i] * p)
    grad = -sum / m
    return grad


def gradient_descent(X, y, theta, alpha, iters, num_class):

    n = X.shape[1]
    for iter in range(iters):
        dtheta = np.zeros((num_class, n))
        for j in range(0, num_class):
            dtheta[j,:] = grad_cost(X, y, j, theta, num_class)
            theta = theta - alpha * dtheta
        if iter % 10 == 0:
            print(f"Cost at iteration {iter}", cost)

    return theta

def h(X, theta, num_class):

    X = np.matrix(X)
```

```
    h_matrix = np.empty((num_class,1))
    den = 0
    for j in range(0, num_class):
        den = den + math.exp(np.dot(theta[j], X.T))
    for i in range(0,num_class):
        h_matrix[i] = math.exp(np.dot(theta[i], X.T))
    h_matrix = h_matrix / den
    return h_matrix

def my_J(theta, X, y, j, num_class):
    cost = -(indicator(y,j))*(np.log(phi(j, theta, X, num_class)))
    # YOUR CODE HERE
    #raise NotImplementedError()
    return cost

def my_grad_cost(X, y, j, theta, num_class):
    cost = 0
    sum = np.array([0 for i in range(0,n)])
    for i in range(0, m):
        p = indicator(y[i], j) - phi(j, theta,X.loc[i], num_class)
        sum = sum + (X.loc[i] * p)
        # Add all the costs for each i
        cost = cost + my_J(theta, X.loc[i], y[i], j, num_class)
    grad = -sum/m

    return grad, cost

def my_gradient_descent(X, y, theta, alpha, iters, num_class):
    cost_arr = []
    cost_arr = []
    for iter in range(iters):
        cost = 0
        for j in range(0, len(y.unique())):
            grad, cost_obtained = my_grad_cost(X, y, j, theta, num_class)
            theta[j] = theta[j] - alpha * grad
            # Add the costs
            cost = cost + cost_obtained
    # Get cost for every iter
        cost_arr.append(cost)
    return theta, cost_arr

def calc_accuracy(y_test, y_pred):
    accuracy = (y_pred == y_test).astype(int).sum() / y_test.shape[0]
    # YOUR CODE HERE
    #raise NotImplementedError()
    return accuracy
```

In [103]:

```
theta_arr = []
cost_arr = []
accuracy_arr = []

# design your own learning rate and num iterations
alpha_arr = np.array([0.009, 0.01, 0.05, 0.09])
iterations_arr = np.array([50, 50, 50, 50])

# YOUR CODE HERE
m, n = X_test.shape
num_class = len(y_labels_name)

fig = plt.figure()
ax = plt.axes()
plt.grid(axis='both')
for i in range(0, len(alpha_arr)):
    theta_initial = np.ones((num_class, n))
    theta, cost = my_gradient_descent(X_train, y_train, theta_initial, alpha_arr[i], iterations_arr[i], num_class)

    y_pred = []
    for index,row in X_test.iterrows():
        h_matrix = h(row, theta, num_class)
        prediction = int(np.where(h_matrix == h_matrix.max())[0])
        y_pred.append(prediction)

    correct = (y_pred == y_test).value_counts()[True]
    accuracy = correct/m
    plt.plot(range(1,iterations_arr[i]+1), cost, label='alpha:'+str(alpha_arr[i]) +', acc:' + str(np.round(accuracy,4)*100))
    accuracy_arr.append(accuracy)

plt.legend()
plt.show()
```



Conclusion:

I have choosen star classification data.In this data set has two categorical columns, Color and Spectral Class, but there is null values. These two columns' categorical (object) values were converted to integer values (0, 1, 2, 3, ..).Different plots were made for each column to show the distribution. After normalizing the data, it was divided into a training set and a test set with a test size with 30%. i have selected 4 alpha values (0.009, 0.01, 0.05, and 0.09) were used to train the training set over a total of 50 iterations. Only 50 iterations were performed in order to close the loop more quickly, but the theta value was still not at its ideal. The graph shows that the cost is still high and that at least 1000 iterations are required.

In [ ]: