~\Desktop\class\ml\data processing pipeline\data processing pipeline.py

```python
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd

data=pd.read_csv(r'C:\Users\Admin\Desktop\class\ml\simple linear regression pipeline\Salary_Data.csv')

x=data.iloc[:,:-1]
y=data.iloc[:,-1]

from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.20,random_state=0)

x_train=x_train.values.reshape(-1,1) #to convert values into array
x_test=x_test.values.reshape(-1,1)

from sklearn.linear_model import LinearRegression

regressor=LinearRegression()
regressor.fit(x_train,y_train)

y_predict=regressor.predict(x_test)

plt.scatter(x_test, y_test, color='red')
plt.plot(x_train,regressor.predict(x_train))
plt.title('salary vs experience (test set)')
plt.xlabel('years of experience')
plt.ylabel('salary')

m_slope=regressor.coef_ #for slope (m)
print(m_slope)

c_intercept=regressor.intercept_ #for constant (c)
print(c_intercept)

y_15=m_slope*15+c_intercept #y^
print(y_15)

comparsion=pd.DataFrame({'actual':y_test,'predicted':y_predict})
print(comparsion)

data.mean()

data.std()

data['Salary'].mean()

data.median()


data['Salary'].median()

```

```
data.mode()

data['Salary'].mode()

data.var()

data['Salary'].var()

from scipy.stats import variation #coff variation

variation(data.values)

variation(data['Salary'])

data.corr()

data['Salary'].corr(data['YearsExperience'])

data.skew()

data['Salary'].skew()

import scipy.stats as stats

data.apply(stats.zscore)

y_mean=np.mean(y)
stats.zscore(data['Salary'])
ssr=np.sum((y_predict-y_mean)**2)
print(ssr)

y=y[0:6]
sse=np.sum((y-y_predict)**2)
print(sse)


mean_total=np.mean(data.values)
sst=np.sum((data.values-mean_total)**2)
print(sst)



rsquare=1-(ssr/sst)
print(rsquare)

import pickle
filename = 'linear_regression_model.pkl'
with open(filename, 'wb') as file:
    pickle.dump(regressor, file)
print("Model has been pickled and saved as linear_regression_model.pkl")

import os
print(os.getcwd())

```