Linear Regression Assignment

support@intellipaat.com

+91-7022374614

US: 1-800-216-8930 (Toll-Free)

Problem Statement:

1. You have access to the salary information of several employees along with their Years of Experience. Using Linear regression analysis in machine learning, create a linear regression model can predict the salary of an employee based on the years of experience.
2. House prices can be an ever changing trend, but it does change based on certain parameters. You are provided with housing data that has information on various houses and their prices. Use the data at hand to predict the prices of the house using linear regression in machine learning.

Dataset Information:

1. **Data.csv -** This dataset contains two columns with 30 entries each for employee years of experience and their salary.

Column Name	Description
YearsExperience	The column contains 30 entries of the employee's years of experience.
Salary	The salary column contains 30 entries of their respective salary for their years of experience.

2. **Housing.csv** - The dataset is considerably larger and contains the following columns in the data. The dataset contains more than 20,000 entries for information about the houses, prices and various other parameters.

Column Name	Description
id	The id column contains a separate id for the houses in the data.
date	The data contains the time series in which all the houses' respective dates have been mentioned.
price	The price column lists the price of the house.
bedrooms	The number of bedrooms in the house.
bathrooms	The number of bathrooms in the house.
sqft_living	The area of the living room.
sqft_lot	The area of the lot.
floors	Number of floors in the house.

waterfront	If the house has a waterfront or not
view	If the house has a viewfront or not.
condition	Condition of the house represented in various categories.
grade	The grade of the house in various categories.
sqft_above	The area above.
sqft_basement	The basement area.
yr_built	In which year the house was built.
yr_renovated	In which year the house was renovated.
zipcode	The zipcode of the house.
lat	The latitude information of the house.
long	The longitude information of the house.
sqft_lot15	The average square footage of the 15 closest houses.
Sqft_basement_15	The average square footage of the 15 closest houses.

Explore the datasets, and perform EDA on both the datasets before starting the following exercise.

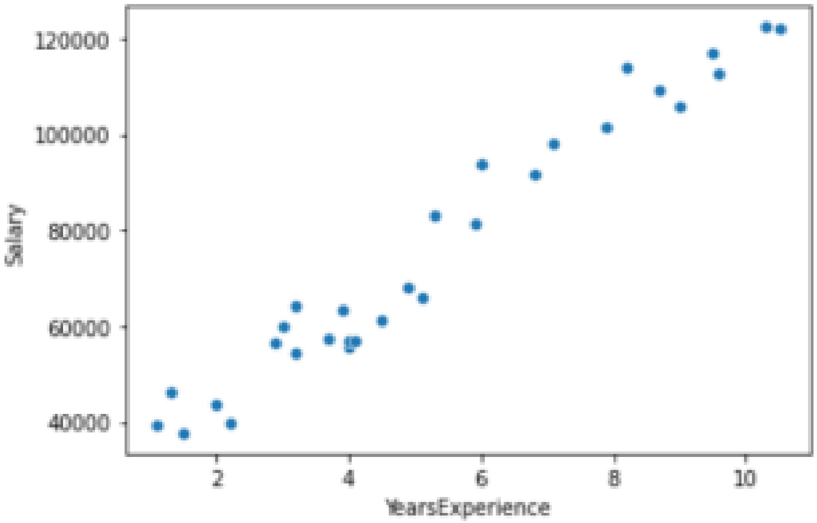
Use the data.csv for the questions mentioned below

1. How many employees having more than 5 years experience are earning more than 60000?
   a. 41
   b. 12
   c. 21
   [✓ d. 14]

2. How many employees are earning between 50000-80000?
   a. 14
   [✓ b. 12]
   c. 10
   d. 13

3. The scatter plot in the following image shows the relationship between the “YearsExperience” and “Salary” columns. What possible inferences can be drawn from the

plot?



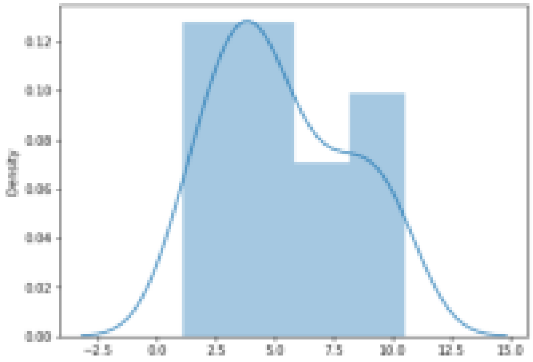
[✓ a. The plot shows a positive correlation between the ‘YearsExperience” and “Salary” column.]

b. The plot shows no significant relationship between the “YearExperience” and “Salary” column.

c. The plot shows a negative correlation between the “YearsExperience” and “Salary” column.

d. None of the above.

4. The distribution plot of the column “YearsExperience” is shown in the image below, what possible inferences can be drawn from the plot.



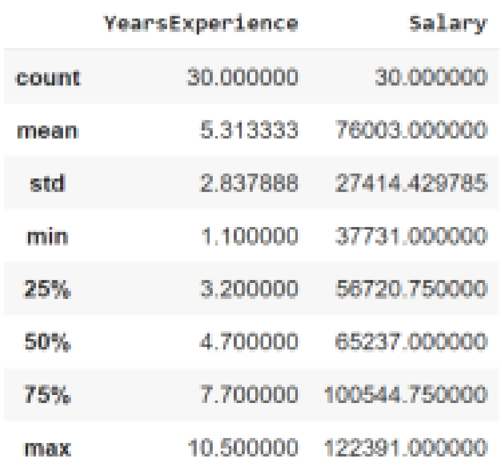
a. “YearsExperience” data is normally distributed.

[✓ b. “YearsExperience” data is positively skewed.]

c. “YearsExperience” data is negatively skewed.

d. None of the above.

5. What all inferences can be drawn from the table shown below:



	YearsExperience	Salary
count	30.000000	30.000000
mean	5.313333	76003.000000
std	2.837888	27414.429785
min	1.100000	37731.000000
25%	3.200000	56720.750000
50%	4.700000	65237.000000
75%	7.700000	100544.750000
max	10.500000	122391.000000

[✓ a. The range of the “YearsExperience” and “Salary” data is (9.4 , 84660)]
b. The range of the “YearsExperience” and “Salary” data is (4.7 , 65237)
c. The range of the “YearsExperience” and “Salary” data is (10.5, 122391)
d. The range of the “YearsExperience” and “Salary” data is (7.7 ,100544)

6. To split the dataset into training and testing data, if we use the following code. X = data['YearsExperience']

y = data['Salary']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,

random_state=0) What does it mean when we write the test size as 0.2?

a. The testing data will be 2% accurate.
b. The testing data will have 80% samples from the total population.
c. The testing data will have 2% samples from the total population.
[✓ d. The training data will consist of 80% of the samples from the total population.]

7. In the above example code, we have taken the random state as 0, if we change the random state as 42, what does it mean for our training and testing data?
a. The shape of the training data will become (42,)
b. The shape of the training data will become (42,2)
[✓ c. The random state does not have any effect on the shape of the data.]
d. The random state will increase the efficiency of the model by 42%.

8. If the r2 score calculated in the above example is 0.98 , change the sample size of the training and testing set in the ratio 60:40, and build a linear regression model again. After plotting the best fit line on the test data, calculate the r2_score for the new model.

a. 0.98
[✓ b. 0.96]
c. 1.0

d. 0.0

9. If while fitting the model with training and testing data, you get the following error `ValueError: Expected 2D array, got 1D array instead:` What could be the issue with the data, and how can you solve it?

[✓ a. Reshape the data to a two dimensional array]
b. Reshape the data to two arrays of 1-D each.
c. Both A and B
d. None of the above

The exercise after this contains questions that are based on the housing dataset.

10. How many houses have a waterfront?

a. 21000
b. 21450
[✓ c. 163]
d. 173

11. How many houses have 2 floors?

a. 2692
[✓ b. 8241]
c. 10680
d. 161

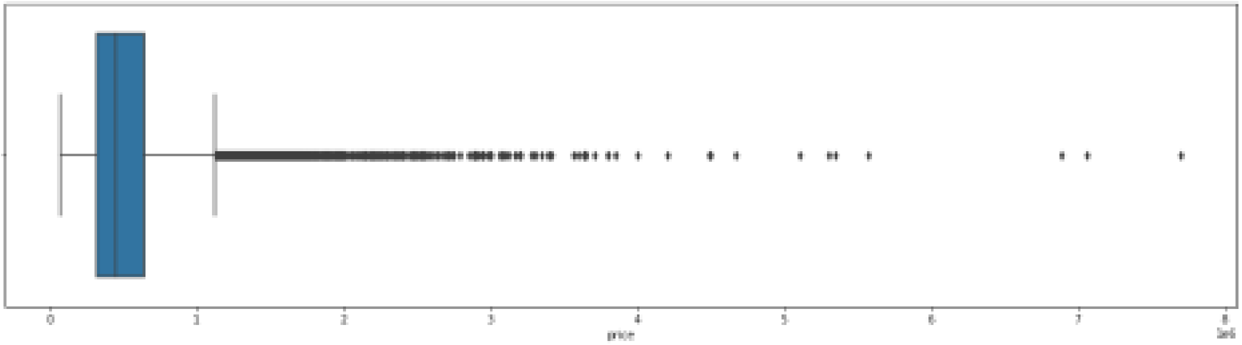
12. How many houses built before 1960 have a waterfront?

[✓ a. 80]
b. 7309
c. 90
d. 92

13. What is the price of the most expensive house having more than 4 bathrooms?

[✓ a. 7700000]
b. 187000
c. 290000
d. 399000

14. The image shown below shows the boxplot of the price column from the housing dataset. What inferences can you make from the plot?



a. The price column is normally distributed.
b. There might be high chances of price data having null values.
c. There is a presence of outliers in the price data. ✓
d. There is no presence of outliers in the price data.

15. For instance, if the ‘price’ column consists of outliers, how can you make the data clean and remove the redundancies?
a. Calculate the IQR range and drop the values outside the range. ✓
b. Calculate the p-value and remove the values less than 0.05.
c. Calculate the correlation coefficient of the price column and remove the values less than the correlation coefficient.
d. Calculate the Z-score of the price column and remove the values less than the z-score.

16. What are the various parameters that can be used to determine the dependent variables in the housing data to determine the price of the house?

a. Correlation coefficients ✓
b. Z-score
c. IQR Range
d. Range of the Features

17. If we get the r2 score as 0.38, what inferences can we make about the model and its efficiency?
a. The model is 38% accurate, and shows poor efficiency. ✓✗
b. The model is showing 0.38% discrepancies in the outcomes.
c. Low difference between observed and fitted values.
d. High difference between observed and fitted values. ✓

18. If the metrics show that the p-value for the grade column is 0.092, what all inferences can we make about the grade column?
a. Significant in presence of other variables.
b. Highly significant in presence of other variables
c. insignificance in presence of other variables ✓
d. None of the above

19. If the Variance Inflation Factor value for a feature is considerably higher than the other features, what can we say about that column/feature?

[✓ a. High multicollinearity]
b. Low multicollinearity
c. Both A and B
d. None of the above