`

**A picture containing clipart

Description automatically generated**

**CST8390 - Business Intelligence and Data Analytics**

**Lab 8 - *Regression***

**Name: Min Li - Id: 040930563**

**Due Date:** Week 11 in own lab sessions.

**Introduction**

The goal of this lab is to perform **linear regression** on housing file.

**Steps for Linear Regression:**

1. Open the housing.arff file (uploaded in **Brightspace**) in Notepad to read about the data. Fill in the following questions:
   1. Number of instances: 513. b. Number of attributes: 14. c. Attribute Information:

1. CRIM per capita crime rate by town

2. ZN proportion of residential land zoned for lots over 25,000 sq.ft.

3. INDUS proportion of non-retail business acres per town

4. CHAS Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)

5. NOX nitric oxides concentration (parts per 10 million)

6. RM average number of rooms per dwelling

7. AGE proportion of owner-occupied units built prior to 1940

8. DIS weighted distances to five Boston employment centres

9. RAD index of accessibility to radial highways

10. TAX full-value property-tax rate per $10,000

11. PTRATIO pupil-teacher ratio by town

12. B 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town

13. LSTAT % lower status of the population

14. MEDV Median value of owner-occupied homes in $1000's

1. Start **Weka** and open the file housing.arff. Find the following information from the preprocess tab. The **median** is the middle value of a sorted list, so **click** on the edit tab, and **sort** the columns and find the middle element ( How to get the those median numbers, I forgot it.)

1. Median House Value (class) x $1000: 21.1.
2. Median number of rooms per dwelling: 8.704.
3. Median per capita crime rate: 0.33169.
4. Click on the Classify tab and choose “LinearRegression” from Functions. Modify the algorithm parameters so that outputAdditionalStats is “**true**”. Ensure that “class” is set for what value is being computed. Run the algorithm to output the **weights** of the regression. (***Answer should be typed in. Snippet or screenshot not permitted.***)
   1. What is the linear regression **model** for this set?

Class =

-0.0914 \* CRIM +

0.0577 \* ZN +

-0.0931 \* INDUS +

2.8323 \* CHAS=1 +

-72.568 \* NOX +

2.5705 \* RM +

-1.2806 \* DIS +

0.2532 \* RAD +

-0.0132 \* TAX +

-0.7959 \* PTRATIO +

0.0094 \* B +

-0.6428 \* LSTAT +

65.9273

* 1. Which are the **two highest** factors which have a **positive influence** on the housing price?

CHAS=1, RM.

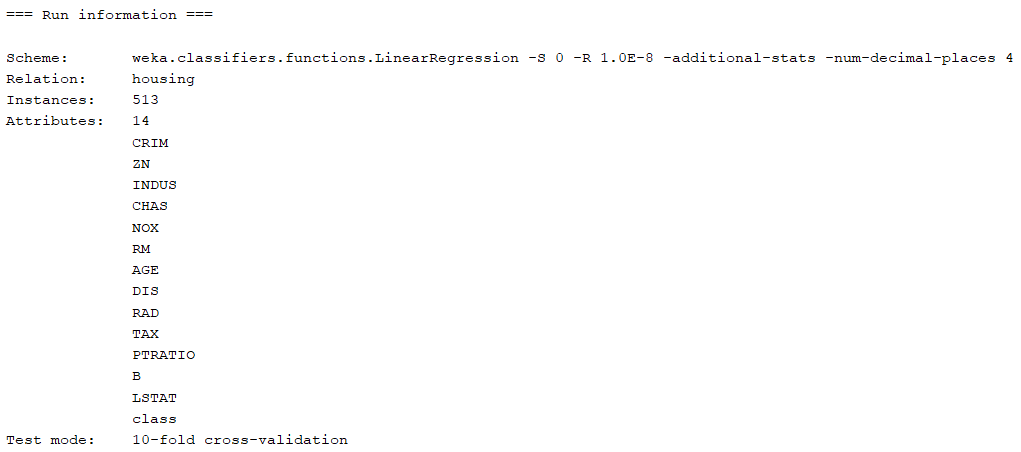
* 1. Which are the **two highest** factors that have a **negative influence** on housing price?

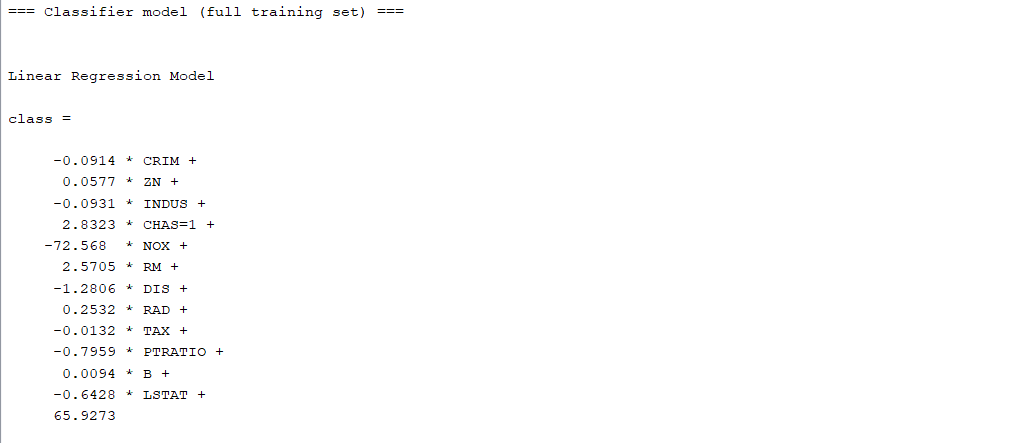
NOX, DIS.

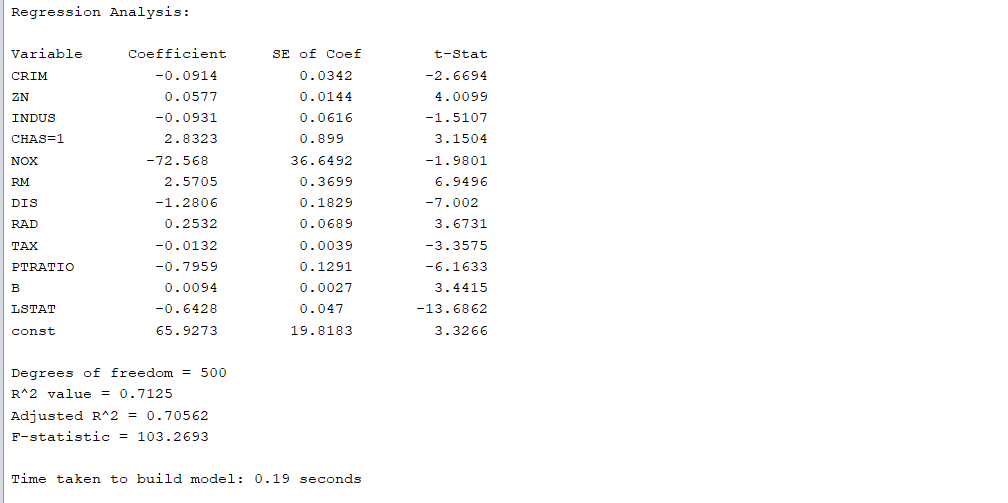
**REMEMBER:**

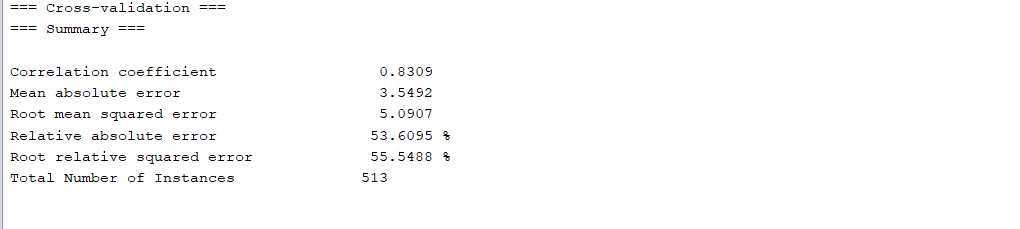
Show your **answers** to the lab professor when you are done.

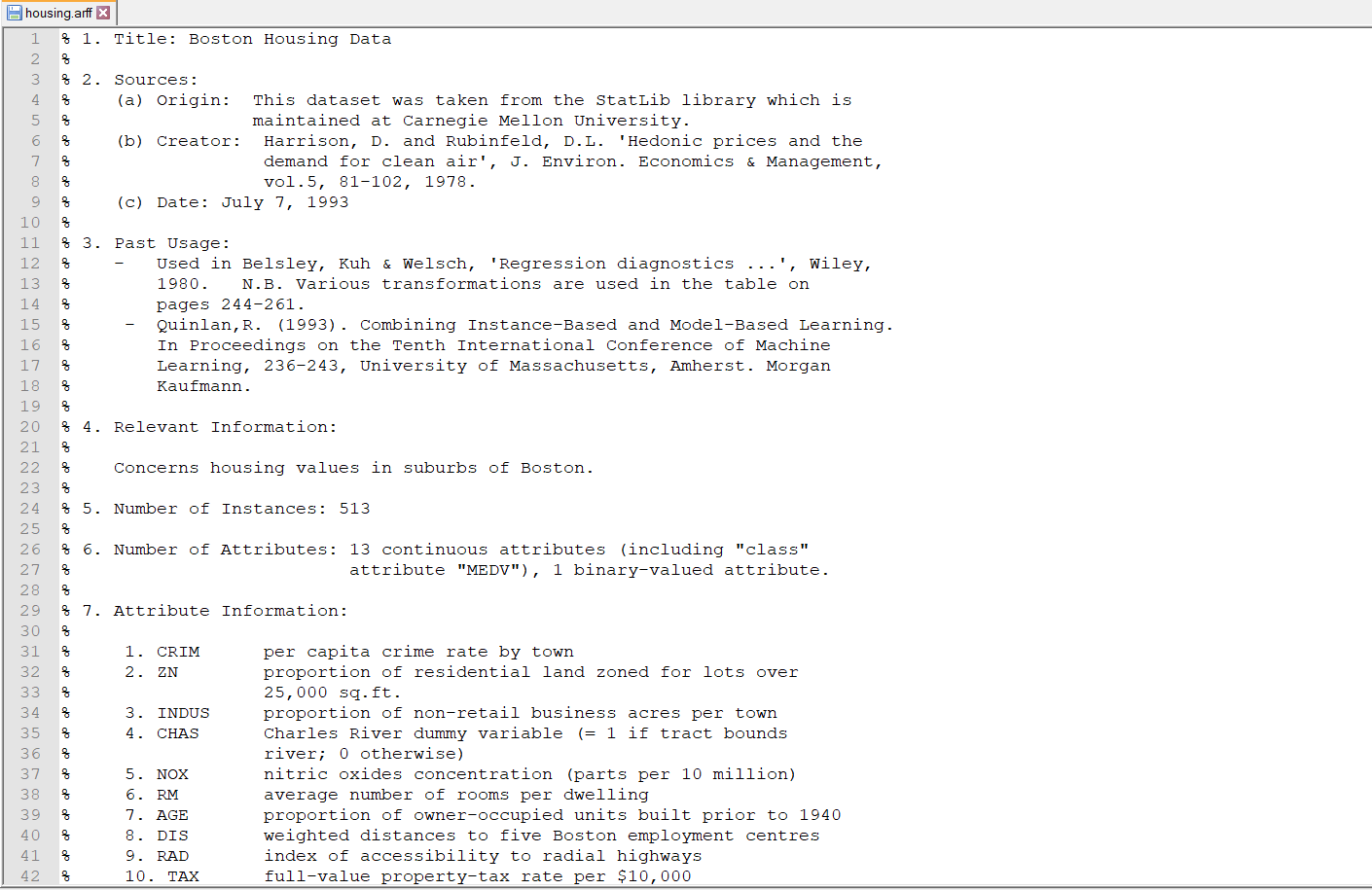
You should be ready with your results in the result pane and housing file opened in **Notepad++**.











***FOR YOUR ANALYSIS:***

***\* Option 1****: Explain what a* ***Regression*** *is**and where to use it.*

***\* Option 2****: Explain how to determine the* ***factors*** *and their* ***impact*** *(positive or negative) to the analysis.*

Option1: **Linear regression** is one of the most basic predictive process. As the name suggests, linear regression is used to find out linear relation between dependent variable and an independent variable. In a prediction, **dependent variable** means the variable which is dependent on other factors and **independent variable** refers to the mutually independent variables which effect the value of the target variables. In case of linear regression, the relation between dependent and independent variables are assumed to be linear. We can use Linear regression result to make predictions. For example, there is a **linear relationship**between miles driven and total paid for gas. Because this relationship is linear, if you spend less/more money — e.g. half vs full tank — you'll be able to drive fewer/more miles. And because that relationship is linear and you know how long is your drive from San Francisco to Las Vegas, using a **linear model**will help you **predict** how much you are going to budget for gas.

Ottawa, Mar 2020.