**Data Mining – Project 1 – Summer 2016**

Name: Uday Sagar Panjala

UID: U00771807

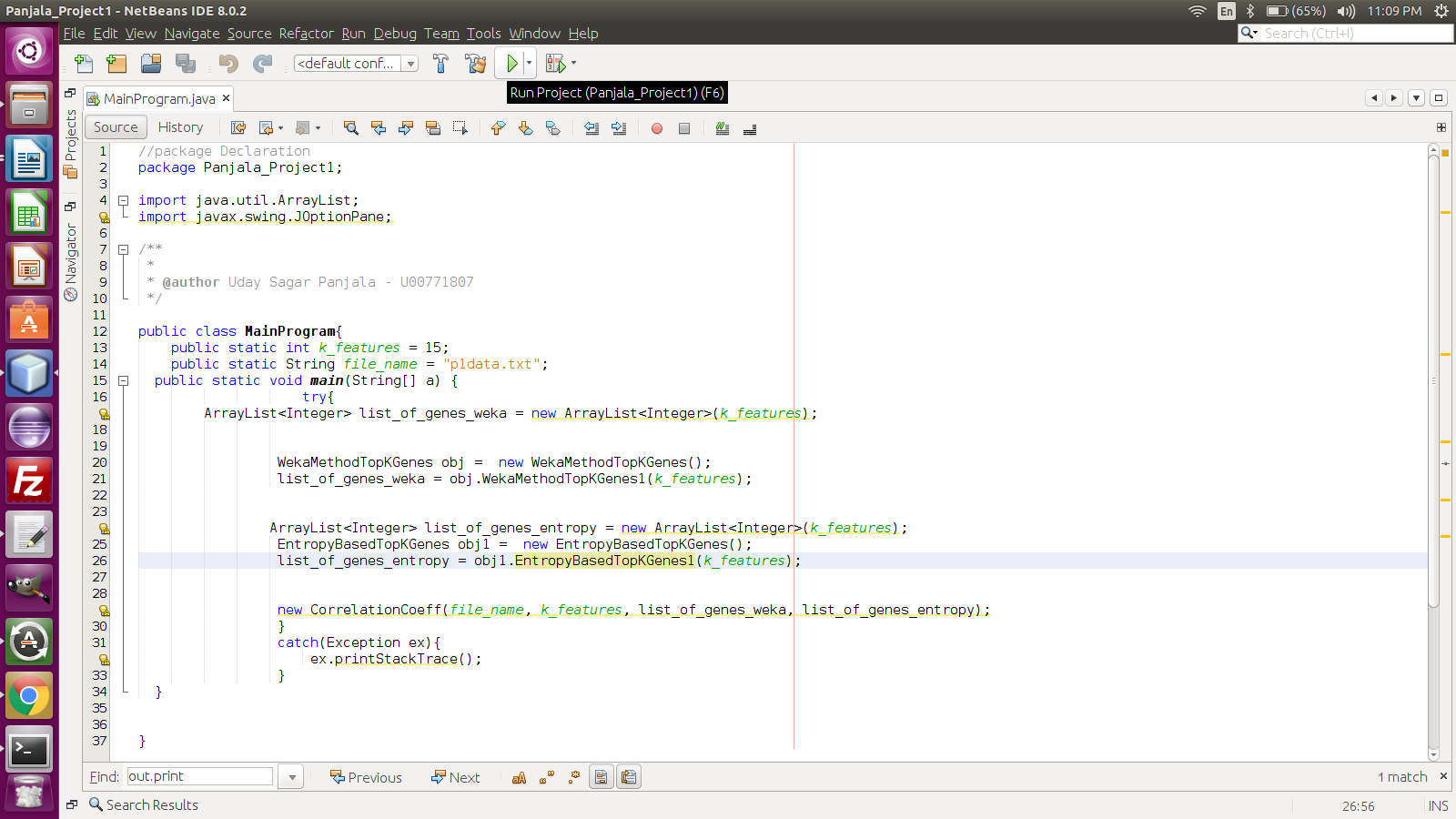
Course No. 7720-01

Project 1

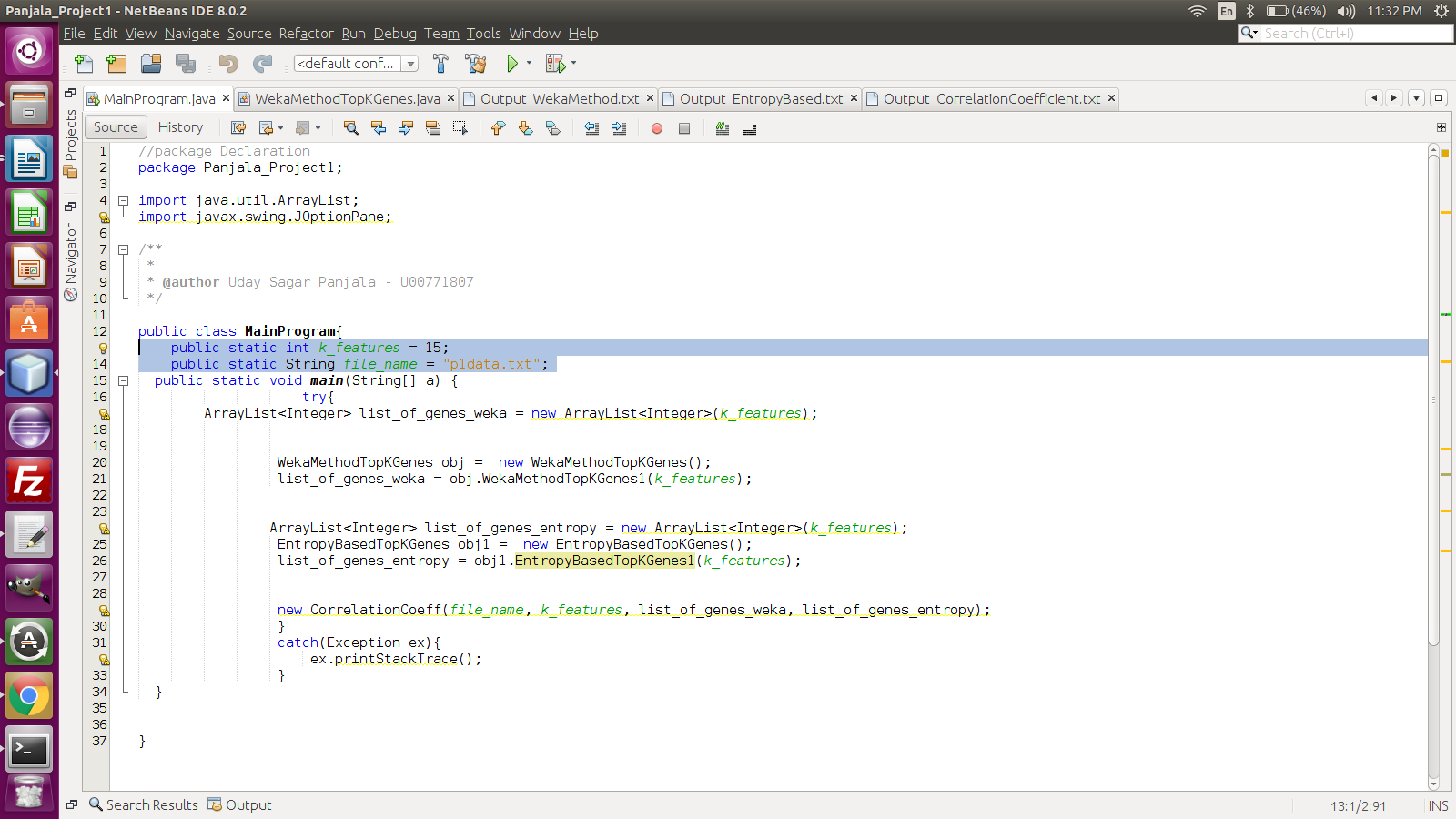
**Instructions:**

1. Please open the project via **NetBeans.** The project name is Panjala\_Project1 (lastName\_Project1).

2. The main file is named as MainProgram.java and could be run by clicking on the “run” button as shown in the screenshot below.

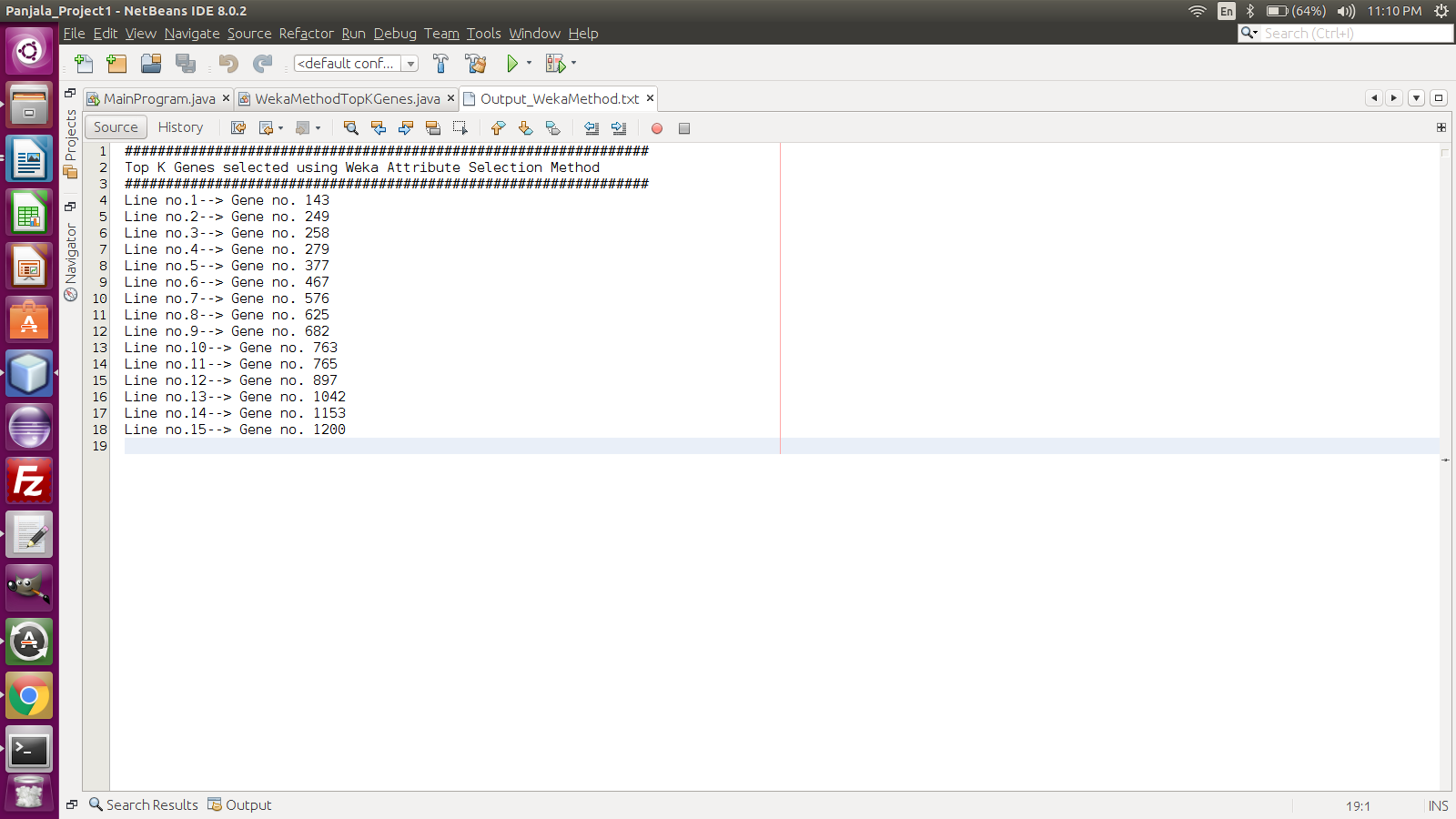


The value of the top K features and the input dataset file can be changed in the above file as highlighted in figure below



3. The main program triggers all the three tasks.

4. The first task is to calculate the top K features based on weka. A call to weka is made internally to the program and can be seen in “**WekaMethodTopKGenes.java**”. The output is saved in the file, “Output\_WekaMethod.txt”. It is shown below.

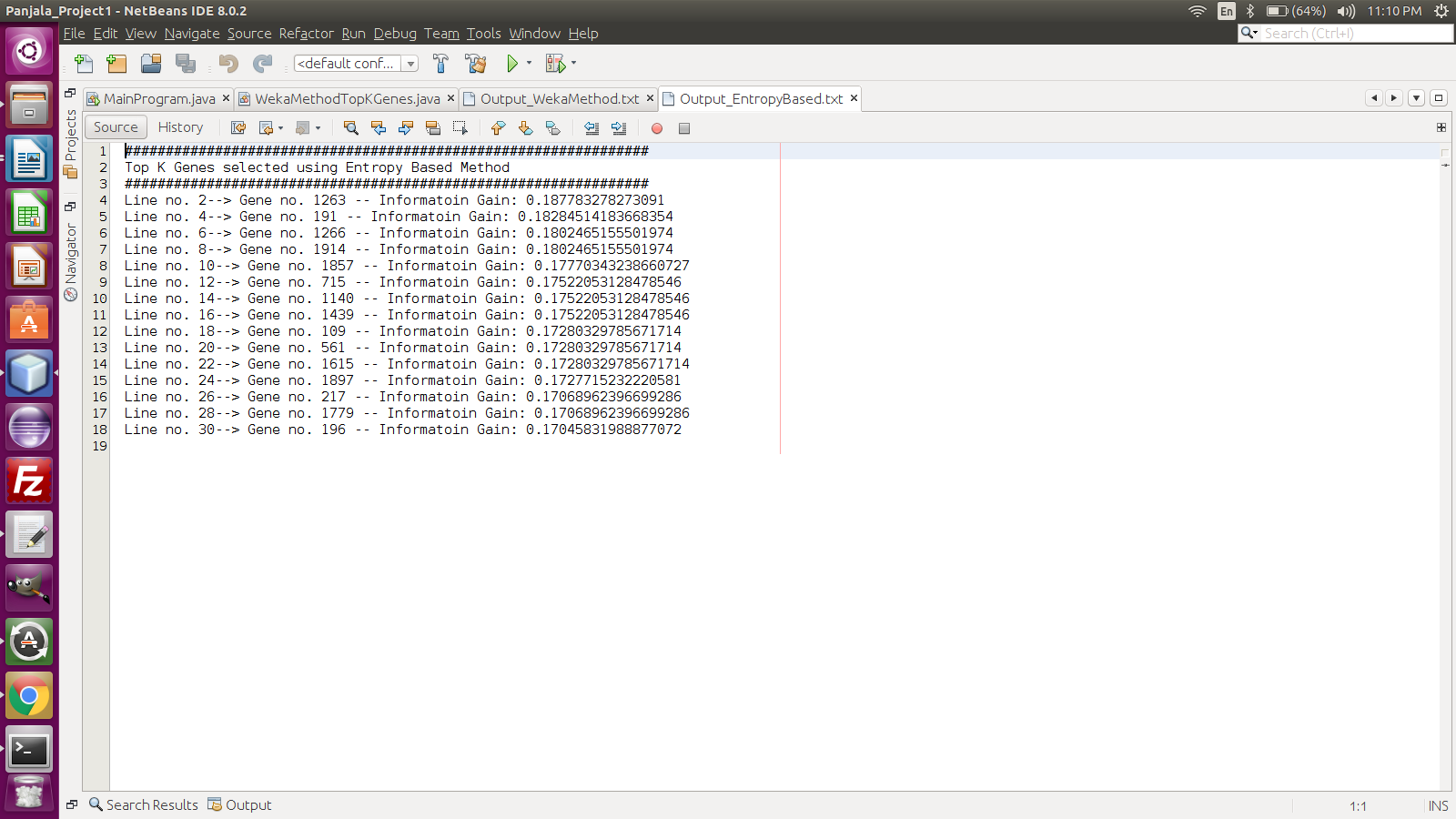


5. The second task is to find the top K genes present in the 62 cases each having around 2000 genes with the class given. The information gain is calculated for each gene based on two split methods. In the first split method, a split is made in the half which prepares two bins. A second split is made in the left bin. For the second split method, the second split is made in right bin.

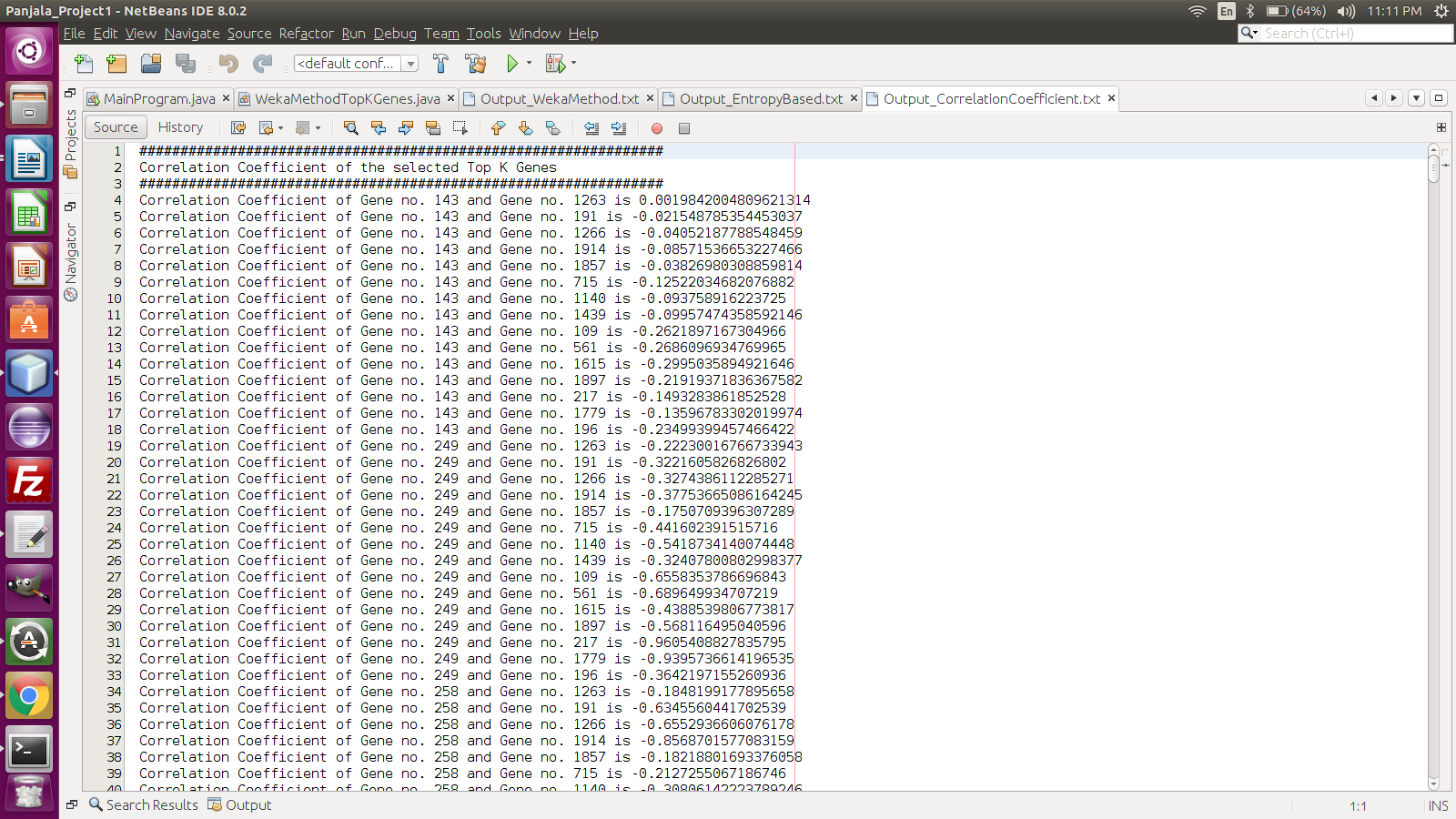
6. The entropy of all the genes is calculated which is used to get the split information which is inturn used to calculate the information gains of each split using the formula as shown below.

*Information Gain = Entropy – Information Split*

7. The entropy based top K genes are outputted in “Output\_EntropyBased.txt” file. It is shown below.



8. The correlation coefficients are calculated between the top K Genes selected using weka and top K Genes using entropy based information gain. We get a total of K \* K combinations. The output is saved in “Output\_CorrelationCoefficient.txt” and is shown below.



9. The correlation coefficients values here are mostly negative which indicates that both variables move in the opposite direction. If one variable increases, the other variable decreases with the same magnitude, and vice versa.

10. For example, suppose a study is conducted to assess the relationship between outside temperature and heating bills. The study concludes that there is a negative correlation between the prices of heating bills and the outdoor temperature. The correlation coefficient is calculated to be -0.96. This strong negative correlation signifies that as the temperature decreases outside, the prices of heating bills increase and vice versa.

**Citations:**

http://www.investopedia.com/ask/answers/032515/what-does-it-mean-if-correlation-coefficient-positive-negative-or-zero.asp