Data Science (Eng) 774/874
Pre-block Assignment

March 1, 2021

1. Pre-reading

1) Study Chapters 1 and 2 of the prescribed textbook: Skiena, 2017, The Data Science Design Manual, Springer: Cham.

2) Do your own research on the CRoss Industry Standard Process for Data Mining (CRISP-DM) methodology. Hint: Focus on academic resources vs. popular articles.

2. Assignment

Answer the questions below and submit your work electronically on SunLearn by the due date at **23:55 on 14 March 2021.** This is an individual assignment. [33]

1) What do you understand by the term "data science"? [4]

2) What are the advantages of using the CRISP-DM methodology? [4]

3) Describe the different phases of the CRISP-DM methodology. [12]

4) If a positively skewed distribution has a median of 50, which of the following statements is true? [2]
   A) Mean is greater than 50
   B) Mean is less than 50
   C) Mode is less than 50
   D) Mode is greater than 50
   E) Both A and C
   F) Both B and D

5) Is standard deviation robust to outliers? Explain your answer. [2]

6) The correlation between two variables (Var1 and Var2) is 0.65. Now, after adding a value of 2 to all the values of Var1, the correlation coefficient will_____ ? [2]
   A) Increase
   B) Decrease
   C) None of the above

7) It is often said that "correlation does not imply causation". Explain and find a real-world example to motivate your answer. [4]

8) Compare the Pearson Correlation Coefficient and the Spearman Rank Correlation Coefficient. [3]