DSC 423
Assignment 1

Based on Prerequisites and Modules 1 and 2

Your submission must include your name and student ID. Your submission must include the honor statement: "I have completed this work independently. The solutions given are entirely my own work." You submission must be submitted as a PDF.

1. Short Essay (10 pts.) For each of these questions, your audience are persons that are not experts in statistics. Write with complete sentences and paragraphs. Cite any references that you use.
    a. (5 pts.) Imagine you fit a regression model to a dataset and find that R-squared = 0.69. Is this a good regression model or not? If you cannot tell, what additional information do you need? Explain.
    b. (5 pts.) Research and then explain the “regression fallacy”. Provide at least one example.
2. Short Essay (5 pts.) Consider the following two scenarios. A) take a simple random sample of 100 graduate students at DePaul university and b) take a simple random sample of 100 graduate students studying Data Science. For each sample you record the amount spent on textbooks used for classes. Which sample do you expect to have the smaller standard deviation? Explain your answer.
3. Empirical rule (10 pts.) The 222 students enrolled in online-learning courses offered by a college ranged from 18 to 64 years of age. The mean age was 28 with standard deviation equal to 4. Use the 68-95-99.7 rule to answer the following questions:
    a. (5 pts.) Compute the percentage of students that are between 24 and 32 years old. Show your work.
    b. (5 pts.) Compute the percentage of students that are older than 36 years. Show your work.
4. Z-scores (5 pts.) Monthly sale figures for a particular e-retailer tend to be normally distributed with mean equal to 150 thousand dollars and a standard deviation of 35 thousand dollars. Use the normal distribution to determine the top 1% monthly sale figure (a.k.a. 99th percentile)? Show your work.
5. Hypothesis Testing (10 pts.) A network provider investigated the number of blocked intrusions to its network, and found that there were, on average, 45 blocked intrusions per day. After a change in firewall settings, the mean number of intrusions during the next 35 days was 42 with a standard deviation equal to 15.5. Perform a hypothesis test to determine if the change in firewall settings reduced the number of intrusions. Show your work.
6. QUASAR (10 pts.) -- A quasar is a distant celestial object (at least four billion light-years away) that provides a powerful source of radio energy. The Astronomical Journal (July 1995) reported on a study of 90 quasars detected by a deep space survey. The survey enabled astronomers to measure several different quantitative characteristics of each quasar, including:

    X1 - Redshift
    X2 - Line Flux
    X3 - Line Luminosity
    X4 - AB1450 Magnitude

X5 - Absolute Magnitude
Y1 - Rest frame Equivalent Width

a. (5 pts.) Use R to perform a regression analysis on the QUASAR dataset (found on the D2L). For each of the explanatory variables create a regression model and copy/paste it into your submission.
b. (5 pts.) Evaluate your models. For each discuss how well they predict the dependent variable. Your description should begin by reporting basic facts about your model; but should also include an analysis of the findings. What is the best model? Assume your audience is a fellow DSC423 student