**Week 1**

**Assignment 1**

**Worth 30 points**

1. Identify data types:
2. **(5 points)** Login to LionPath to complete the following:
   * **(2 points)** Identify 5 different data points that would be structured data.
   * **(2 points)** Identify 5 different data points that would be unstructured data.
   * Note: You can choose other pages than the main page to identify the data. Put the data into a **table**.
   * **(1 point)** Identify the data points as (interval, nominal, ordinal, or ratio)
3. **(5 points)** Answer the following questions:
   * **(2 points)** Capturing 5000 records from Expedia travel website, which includes zip codes, to determine where people from specific zip codes go on vacation. Is this a population or sample?
   * **(1.5 points)** What would need to be done to the previous question for the correct answer to be the opposite of what you answered. Meaning how would you make it a population or sample, depending on what you answered.
   * **(1.5 points)** An organization wants to know what consumers think about changing the color of their product. They have 100000 followers on social media. If they send a survey to all 100000 followers would this be a population or a sample?
4. This week you work for a Real Estate company and have exported the data from your ERP system to provide analysis for a meeting you have this week. Your boss wants you to provide analysis of the data as outlined below. Please use the Excel file (Week1\_House\_Price.xlsx) to complete the following.
5. **(5 points)** Identify missing and invalid data
   * In R, use functions to find the missing information. Indicate the column and the row numbers that have the missing (NA) values
6. **(7.5 points)** Subset the data into the different colleges and universities
   * Select 5 college universities and subset the data
   * Create different dataframes in R and create csv files for each dataframe
     + Name the dataframes appropriately
     + Make sure to document what you are doing in the script using the # symbol
   * For each subset, provide the count of sales and the total amount of sales in R, using functions
7. **(7.5 points)** Transform numeric variables
   * In R –
     + Import Excel file into dataframe – HousePrices
     + Bin Sale\_amount into bins (0, .33, .66, 1) use SaleBins for the vector name
     + Add the bin information to the HousePrices dataframe using the cut function
       - Label the bins with the cutoff amounts in the Bins (i.e. if bins were 10000, 20000, 30000 then label the bins 10k, 20k, 30k)
     + Create bins using the cut function for High, Medium and Low
       - Add the factors to the HousePrices dataframe