Answers to Q1

A1.

* 1. What are the characteristics of data?  
     Good quality data is precise, consistent and reliable. False data is worse than incomplete data or no data at all. The data should be relevant to the problem.
  2. What is the kind of data that she is analyzing?  
     We need to understand the nature of data we are analyzing before deciding the technique for data analysis. Answers to questions like - “Whether the data is numerical or not? Is it labeled or not?” can help determine the data analysis technique.
  3. Is the data sufficient? Is it clean and ready for analysis?  
     The analysis always becomes more accurate when there is more data to analyze. The data should be free from empty parameters, outliers, errors, etc.
  4. How is she estimating a firm’s performance? What data analysis technique is she using to find correlation between R&D and a firm's performance?  
     It is difficult to give a number to a firm’s performance especially because so many parameters such as total assets, book value, share value, ebitda, etc. reflect it.  
     How is correlation determined between R&D and a firm's performance? Have all other factors been kept constant for this analysis?
  5. Are there any patterns in data that may introduce bias?  
     Some firms may not have high correlation between R&D expenditure and performance because of the nature of the industry it operates in and vice versa. This can introduce bias and error in results. Similarly, other patterns in data can lead to heavy bias.

1. Populated Tobin’s Q metric for the samples in the dataset. Please refer to the code and the output file.
2. Calculated correlation between R&D expenditure and Tobin’s Q metric. Correlation between R&D expenditure and tq came out to be: -0.01. Thus, the results suggest no correlation between them.
3. Since there are patterns in data like common operating industry of firms and time frame of analysis, it is better to find correlation between R&D expenditure and firm’s performance by grouping based on these parameters. As a result of data grouping and analysis, the variable ‘**corr\_by\_time\_industry**’ states correlation segregated by financial year and industry. The results suggest that there is indeed a huge difference in correlation based on industry of operation and financial year. For example, firms with industry code **337910** and **423450** showed high correlation of **1.0** and **0.97** respectively between R&D expenditure and performance in year 2010 whereas firms with industry code **423690** and **333242** showed very poor correlation of **0.33** and **0.19** respectively in the same financial year.  
   Note: Some of the correlation values are “not interpretable” and internally represent NaN because the values do not vary.  
   Formula for calculating correlation:   
     
    **cor(i,j) = cov(i,j)/[stdev(i)\*stdev(j)]**

If the values of the ith or jth variable do not vary, then the respective standard deviation will be zero and so will the denominator of the fraction. Thus, the correlation will be NaN.

Additional note: Please refer to the code and output file attached.