# **Part 1 – Data Preparation and Pre-processing [8 points]**

## 1) Describe the dataset.

### For example:

### What are the categories/domains of the dataset? Id, Title, Company, Date, Location, Area, Classification, Sub-classification, Requirement, Full Description, Lowest Salary, Highest Salary, Job Type.

### What is the dataset size of each variation?

### What is dataset structure/format?

### What are attributes/features of review data you are going to use?

### What are attributes/features of product data you are going to use?

### Which parts of the dataset will you use or all of them?

### [1-2 paragraphs, 3 points]

There are 13 categories for this dataset, they include: Id, Title, Company, Date, Location, Area, Classification, Sub-classification, Requirement, Full Description, Lowest Salary, Highest Salary, Job Type.

Originally there were 318,477 data entries with no duplicate entries. After removing the Id column, it was found that there were 8607 duplicate entries and after removing the Full Description column an additional 1384 duplicates were found, taking the total entries after removing duplicates to 308,486.

The categories that provide no use are Id and Full Description. These columns were dropped to allow for more effective and efficient data cleaning. An Average Salary column was added to provide a more thorough understanding of the Salary differences.

## 2) Describe the steps you used for data preparation and pre-processing.

### For example:

### How do you load the data using Pandas?

### How do you normalize the data?

### How do you clean the data?

### [2-3 paragraphs, 4 points]

To load the data, we must first import the pandas module into our jupyter notebook. From this, we can use the ‘read\_csv(file\_name)’ function with the input variable being ‘data.csv’. This will be saved to the variable ‘df’, short for data file.

Normalizing the data requires…

The cleaning process begins with…

## 3) What is your hypothesis (expectation) about the analysis outcome?

### [1-2 paragraphs, 1 point]

xxx