Data Exploration & Data Preparation

John Kelleher and Brian Mac Namee

- The work you do in todays lab will help you to complete the first assignment for the module.
- You do not have to submit anything today but if you implement todays lab following the instructions then the program you write will be directly relevant to what you will need to submit for the assignment. (In particular, pay attention to the files your program should read and write from).
- I have provided a dataset for you to test you code on. This is not the dataset that will be used for the assignment but it is similar. You can download this test dataset from Labs directory of the module page on webcourses. The test dataset is stored in the file DataSet.txt and there is also a file called FeatureDescriptions.txt which gives you some information about the dataset.

- Your task in this lab is to write a Python program that can take a dataset in a comma separated text file as input and generate the two tables in a data quality report (one for the continuous features and one for the categorical features) that contain the descriptive statistics for the features in the dataset.
- The rest of these slides describe the expected File IO structure of you program and the two tables you need to generate in the report.

- Your program should expect that the dataset is in a file called ’DataSet.txt’ that is stored in a directory called ’data’ that is a subdirectory of the directory your program is run from (in other words the path to the dataset file should be ’./data/DataSet.txt’)
- Your program should output the report containing into a text file with you student number as the name and the file should be stored in a directory called ’reports’ that is a subdirectory of the directory you run your program from (in other word the program should write the report to ’./reports/studentnumber.txt’)

- For **continuous features** the report should include the minimum, 1st quartile, mean, median, 3rd quartile, maximum and standard deviation statistics as well as the number of instances, the percentage of missing values and the number of distinct values present for the feature (or **cardinality**).
- For **categorical features** the report should contain the frequency count and proportions for the two most frequent categories along with the percentage of missing instances and the number of levels the feature exhibits (or **cardinality**).

- For continuous features the report should include:
    - minimum
    - 1^{st} quartile
    - mean
    - median
    - 3^{rd} quartile
    - maximum
    - standard deviation
    - the number of instances
    - the percentage of missing values
    - the number of distinct values present for the feature (or **cardinality**)

Table 1: The structure of a data quality report for continuous features.

		%			1^{st}			3^{rd}		**Std.**
Feature	**Count**	**Miss**	**Card**	**Min**	**Quart**	**Mean**	**Median**	**Quart**	**Max**	**Dev.**
—-	—-	—-	—-	—-	—-	—-	—-	—-	—-	—-
—-	—-	—-	—-	—-	—-	—-	—-	—-	—-	—-
—-	—-	—-	—-	—-	—-	—-	—-	—-	—-	—-
—-	—-	—-	—-	—-	—-	—-	—-	—-	—-	—-
—-	—-	—-	—-	—-	—-	—-	—-	—-	—-	—-
—-	—-	—-	—-	—-	—-	—-	—-	—-	—-	—-
—-	—-	—-	—-	—-	—-	—-	—-	—-	—-	—-
—-	—-	—-	—-	—-	—-	—-	—-	—-	—-	—-

- For categorical features the report should contain
    - mode frequency
    - mode percentage
    - 2^{nd} mode frequency
    - 2^{nd} mode percentage
    - the percentage of missing instances
    - the number of levels the feature exhibits (or **cardinality**)

Table 2: The structure of a data quality report for categorical features.

								2^{nd}	2^{nd}
		%			**Mode**	**Mode**	2^{nd}	**Mode**	**Mode**
Feature	**Count**	**Missing**	**Cardinality**	**Mode**	**Count**	**%**	**Mode**	**Count**	**%**
—-	—-	—-	—-	—-	—-	—-	—-	—-	—-
—-	—-	—-	—-	—-	—-	—-	—-	—-	—-
—-	—-	—-	—-	—-	—-	—-	—-	—-	—-
—-	—-	—-	—-	—-	—-	—-	—-	—-	—-
—-	—-	—-	—-	—-	—-	—-	—-	—-	—-
—-	—-	—-	—-	—-	—-	—-	—-	—-	—-
—-	—-	—-	—-	—-	—-	—-	—-	—-	—-
—-	—-	—-	—-	—-	—-	—-	—-	—-	—-