**IT8701 Introduction to Programming for Data Science**

**Lab 4 – Data Munging with the pandas library**

**Lab 4 – Submission Questions**

# 6. Selection

pandas provides optimized data access methods, .at, .iat, .loc,.iloc and .ix for selecting subsets of a dataset. Practise these methods by completing the lab questions below.

## Subset columns

### Task 3: Create derived columns by boolean logic

|  |  |
| --- | --- |
|  | Pick specific columns with values > 500k Load the data from the ***singstats\_maritalstatus.xlsx*** file into a Pandas DataFrame object ***mydf*** . Set the index as the first column of the Excel file (i.e. “Variables”)  • Find out how many rows of the data in the 1980 column have values more than 500000 and how many rows by of the data have values less than or equal to 500000 by creating a derived column using boolean indexing on this column and using the count() method you have learnt in Section 4 of this lab.  Your output should look similar to that below. |

### Task 4: Select columns by regular expression

|  |  |
| --- | --- |
|  | Pick specific columns that start with characters 201 (filter) Load the data from the ***singstats\_maritalstatus.xlsx*** file into a Pandas DataFrame object ***mydf*** . Set the index as the first column of the Excel file (i.e. “Variables”)   * Select the columns that start with “201” and store them in a dataframe named df\_2010\_and\_after * Print out the values of df\_2010\_and\_after   Your output should look similar to that below. |

## Subset rows

### Task 3: Select row(s) by boolean indexing

|  |  |
| --- | --- |
|  | Select rows that have > 300mm of rainfall Load the data from the *rainfall-monthly-total.csv* file into a Pandas DataFrame object ***df\_rainfall***.  Set the index as the first column of the file (i.e. “month”)  Use boolean indexing to select only the rows of data that has more than 300 mm of rainfall and sort the resulting data in ascending order.  Display the twelve year/month with the most rainfall in a barchart using Matplotlib as shown. |
|  | |

## Drop missing values with dropna

### Task 1: Drop missing values with dropna

|  |  |
| --- | --- |
|  | Drop missing values with dropna() Load the data from the singstats\_maritalstatus.xlsx file into a Pandas DataFrame object mydf. Set the index as the first column of the Excel file (i.e. “Variables”) and consider values that have the value “-“ as being missing or invalid.   * Print out the first 10 rows of the dataset which should already reveal the columns or rows with missing or invalid data * Next, use the pandas ***dropna*** method to drop all the columns with missing data from mydf   Your output should look similar to that below |

**-- End of Lab --**