Big Data Analytics

N**ote:** Please submit the IPython Notebook (.ipynb file). Submissions in screenshots / Word documents or any other format will **NOT** be evaluated.

1.
Using SparkSession and 'data.csv', print all the distinct countries in ascending order with '**an**' in their name.
Output:

```
+------------------+
|           Country|
+------------------+
|            Canada|
|   Channel Islands|
|European Community|
|           Finland|
|            France|
|           Germany|
|           Iceland|
|             Japan|
|           Lebanon|
|         Lithuania|
|       Netherlands|
|            Poland|
|       Switzerland|
+------------------+
```

2.
Using SparkSession and '*data.csv*', show the InvoiceNo, StockCode, and Description for the highest unit price.
Output:

```
+---------+---------+-----------+
|InvoiceNo|StockCode|Description|
+---------+---------+-----------+
|  C556445|        M|     Manual|
+---------+---------+-----------+
```

3.
Using SparkSession and the file *fakefriends-header.csv,* Show each name's total number of friends. Order the results by name in alphabetical order.
Output:

```
+-------+------------+
|   name|sum(friends)|
+-------+------------+
|    Ben|        4888|
|Beverly|        6128|
|  Brunt|        4805|
|   Data|        7192|
| Deanna|        3479|
|  Dukat|        5317|
|   Elim|        2541|
|   Ezri|        4236|
| Geordi|        4728|
| Gowron|        2602|
+-------+------------+
only showing top 10 rows
```

4.
Using SparkSession and the file *ContainsNull.csv,* explain the significance of *how* and *thresh* arguments in *drop()* function.

5.
Using SparkSession and the file *ContainsNull.csv,* fill the null sales values with the minimum sales value.
Output:

```
----+-----+-----+
|  Id| Name|Sales|
+----+-----+-----+
|emp1| John|345.0|
|emp2| null|345.0|
|emp3| null|345.0|
|emp4|Cindy|456.0|
+----+-----+-----+
```

6.
Using SparkSession and the file *appl_stock.csv,* show the unique trade years in descending order with the output column name as shown below.
Output:

```
+----+
|year|
+----+
|2016|
|2015|
|2014|
|2013|
|2012|
|2011|
|2010|
+----+
```

7.
Using SparkSession and the file *appl_stock.csv,* show the average trade volume for each year with the output column names and values as shown below.
Output:

```
+----+---------------+
|year|Final Avg Volume|
+----+---------------+
|2010|  149,826,316.67|
|2011|  123,074,741.67|
|2012|  131,964,204.40|
|2013|  101,608,700.00|
|2014|   63,152,730.56|
|2015|   51,837,886.90|
|2016|   38,415,362.30|
+----+---------------+
```