**Assignment 6 - Spark RDDs, DataFrames, DataSets**

**Submission Deadline: *7h May 2019***

***Question 1: In this question use the spark Data-Frames to input the Bank dataset (given to you in previous assignments) and answer the following queries.***

1. For each job type, print the total number of employees. Also print the min, max and average account balance for each job type.
2. *Find out if the Single people in this dataset are more educated then married.*

***Question 2: Modify the code of K-means discussed in the class to cluster the bank dataset on following five attributes***

1. *Age*
2. *Job*
3. *Education*
4. *Marital*
5. *Balance*

*Find the type of each attribute and use the appropriate distance measure for each type to cluster and then sum over all distances.*

*Run your algorithm for various values of K and for different values of convergence show the results in your report.*

***Question 3. Use the K-mean, k-Medoid functions given in spark ml library to cluster the bank dataset using 7 attributes.***

*Run your algorithm for various values of K and for different values of convergence show the results in your report.*