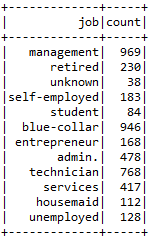
# Assignment – 6

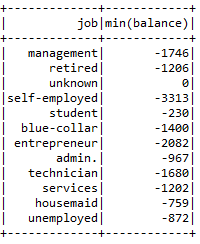
## Big Data Analytics

### Question 1.

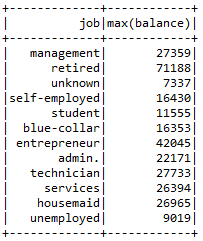
1. **Total Number of Employees**



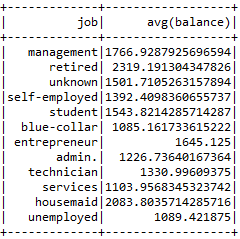
1. **Minimum Balance for each Job Type**



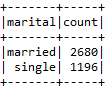
1. **Maximum Balance for each Job Type**



1. **Average Balance for each Job Type**



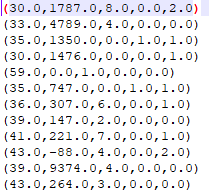
1. **Find out if the Single people in this dataset are more educated then married.**



### Question 2.

In this part, we first had to deal with the categorical data such as {Job, Education, Marital}. For this, I have implemented ***string indexer***. The tutorial for this was found from this link, <https://towardsdatascience.com/feature-encoding-with-spark-2-3-0-part-1-9ede45562740>

The *Encoded Features Data* is stored in the file shared alongside, also afterwards, extracted the features in requirement to perform KMeans is also shown in the *Attributes Data* file.



As we increase the values of K, from K=2, where the error is perfect. As we increase the value of K, the error initially increases and then reaches a substantial value at K=5.

### Question 3.

We have implemented the K-means function given in the *Spark ML Library* with varying number of clusters and calculated, “***Within Set Sum of Squared Errors***”

1. Cluster - 2



1. Cluster - 3



1. Cluster – 4



1. Cluster – 6



1. Cluster – 8



We can see that as the number of clusters increase initially, the error value also increases. However, given that we have selected only 7 random, yet not categorical attributes to work on. The error reduces and obtains a certain low and stable value.

Also, below image showcases the parsed data after *vectors.dense* is applied on the data-set,

