Part b

Question 2

Splitting data set into 2 training and testing splits- 80/20 split and a 70/30 split.

Cmd 7

```
1  (training, test) = df.randomSplit([0.8, 0.2])
```

▸ training: pyspark.sql.dataframe.DataFrame = [movieId: integer, rating: integer ... 1 more fields]
▸ test: pyspark.sql.dataframe.DataFrame = [movieId: integer, rating: integer ... 1 more fields]

Command took 0.05 seconds -- by vishal.kanna@mail.utoronto.ca at 6/16/2021, 9:45:38 AM on firstCluster

Cmd 8

```
1  (training2, test2) = df.randomSplit([0.7, 0.3])
```

▸ training2: pyspark.sql.dataframe.DataFrame = [movieId: integer, rating: integer ... 1 more fields]
▸ test2: pyspark.sql.dataframe.DataFrame = [movieId: integer, rating: integer ... 1 more fields]

Command took 0.07 seconds -- by vishal.kanna@mail.utoronto.ca at 6/16/2021, 10:04:56 AM on firstCluster

Cmd 9

The ALS model being used train and test-

Cmd 9

```
1  als = ALS(maxIter=5, regParam=0.05, rank= 15, userCol= "userId", itemCol= "movieId", ratingCol="rating", coldStartStrategy = "drop")
```

Command took 0.17 seconds -- by vishal.kanna@mail.utoronto.ca at 6/22/2021, 3:15:42 PM on firstCluster

Cmd 10

```
1  eval = RegressionEvaluator(metricName= "rmse", labelCol= "rating",predictionCol= "prediction")
2  eval2 = RegressionEvaluator(metricName= "mse", labelCol= "rating",predictionCol= "prediction")
3  eval3 = RegressionEvaluator(metricName= "mae", labelCol= "rating",predictionCol= "prediction")
```

Command took 0.08 seconds -- by vishal.kanna@mail.utoronto.ca at 6/22/2021, 3:16:21 PM on firstCluster

Model 1-

Cmd 11

```
1  model = als.fit(training)
```

▸ (5) Spark Jobs

Command took 16.48 seconds -- by vishal.kanna@mail.utoronto.ca at 6/18/2021, 11:01:54 AM on firstCluster

Model 2-

Cmd 15

```
1  model2 = als.fit(training2)
```

▸ (5) Spark Jobs

Command took 13.18 seconds -- by vishal.kanna@mail.utoronto.ca at 6/18/2021, 11:02:50 AM on firstCluster