TD 2

1.Ecrivez une fonction (udf) qui ajoute une colonne contenant des code postaux en une colonne contenant le numéro de département.

```
def new_departement (arg1:String)={
  arg1.substring(0,2)
}

def new_dep= udf(new_departement _)
```

2.Ajoutez une colonne avec le code postal utilisant cette fonction nouvellement codée.

```
val new_data = reader.withColumn("dep", new_dep(col("Code_commune_INSEE") ) )
new_data.show
```

```
(nombre de commune :,39201)+------------------+--------------------+-----------+--------------------+--------+--------------------+---+---
|Code_commune_INSEE|         Nom_commune|Code_postal|Libelle_acheminement| Ligne_5|     coordonnees_gps|dep|
+------------------+--------------------+-----------+--------------------+--------+--------------------+---+
|             90093|          SERMAMAGNY|      90300|          SERMAMAGNY|    null|47.687801557,6.83...| 90|
|             91093|   BOULLAY LES TROUX|      91470|   BOULLAY LES TROUX|    null|48.6753515056,2.0...| 91|
|             91100|            BOUVILLE|      91880|            BOUVILLE|    null|48.4326483441,2.2...| 91|
|             91129|               CERNY|      91590|               CERNY|    null|48.4859798517,2.3...| 91|
|             91184|COURDIMANCHE SUR ...|      91720|COURDIMANCHE SUR ...|    null|48.418031424,2.36...| 91|
|             91186|   COURSON MONTELOUP|      91680|   COURSON MONTELOUP|    null|48.5982232022,2.1...| 91|
|             91243|  FONTENAY LES BRIIS|      91640|  FONTENAY LES BRIIS|    null|48.6107631779,2.1...| 91|
|             91244| FONTENAY LE VICOMTE|      91540| FONTENAY LE VICOMTE|    null|48.5468386132,2.4...| 91|
|             91339|               LINAS|      91310|               LINAS|    null|48.6255061941,2.2...| 91|
|             91386|             MENNECY|      91540|             MENNECY|    null|48.5586242184,2.4...| 91|
|             91441|NAINVILLE LES ROCHES|      91750|NAINVILLE LES ROCHES|    null|48.5047296879,2.4...| 91|
|             91461|         OLLAINVILLE|      91340|         OLLAINVILLE|    null|48.6063091787,2.2...| 91|
|             91471|               ORSAY|      91400|               ORSAY| ORSIGNY|48.7004093953,2.1...| 91|
|             91661| VILLEBON SUR YVETTE|      91140| VILLEBON SUR YVETTE|    null|48.6956774396,2.2...| 91|
|             92060| LE PLESSIS ROBINSON|      92350| LE PLESSIS ROBINSON|ROBINSON|48.7804383642,2.2...| 92|
|             93001|       AUBERVILLIERS|      93300|       AUBERVILLIERS|    null|48.9121722626,2.3...| 93|
|             93049|   NEUILLY PLAISANCE|      93360|   NEUILLY PLAISANCE|    null|48.8643287852,2.5...| 93|
|             95181|COURCELLES SUR VI...|      95650|COURCELLES SUR VI...|    null|49.073025738,1.99...| 95|
|             95353|           MAFFLIERS|      95560|           MAFFLIERS|    null|49.083629147,2.31...| 95|
|             95365|    MAREIL EN FRANCE|      95850|    MAREIL EN FRANCE|    null|49.0700384798,2.4...| 95|
+------------------+--------------------+-----------+--------------------+--------+--------------------+---+
only showing top 20 rows
```

3.Quel département contient le plus de codes postaux ?

```
val tmp1 = new_data
   .groupBy(col("dep"))
   .agg(countDistinct("Code_postal"))
      val tmp2 = tmp1.agg(max("count(DISTINCT Code_postal)"))
val departement_max = tmp1
      .join(tmp2,tmp1("count(DISTINCT Code_postal)") ===
      tmp2("max(count(DISTINCT Code_postal))"),"inner")


departement_max.select(col="dep","count(DISTINCT Code_postal)" ).show()
```

```
+---+---------------------------+
|dep|count(DISTINCT Code_postal)|
+---+---------------------------+
| 59|                        250|
+---+---------------------------+
```

4.Combien de code postaux un département a-t-il en moyenne ?

```
val departement_mean = tmp1.agg(mean("count(DISTINCT Code_postal)"))
departement_mean.show()
```

```
+---------------------------------+
|avg(count(DISTINCT Code_postal))|
+---------------------------------+
|                64.1010101010101|
+---------------------------------+
```

5.Lisez le nouveau fichier. Que pouvez vous en dire ?

```
// Il ne contient pas de header. Il a une colonne département comme la colonne que nous avons ajouté au data du fichier 1.
// Il contient une ligne pour chaque département avec diverse information. Une intersection avec le premier fichier semble être possible.

val data_fichier_departement = spark.read
  .option("inferSchema", "true")
  .option("header", "false")
  .option("index","true")
  .option("delimiter",",")
  .format("csv")
  .load("D:/Document D/COURS ESGI/SPARK/departement.csv")
```

```
data_fichier_departement.show()
```

```
+---+---+--------------------+--------------------+--------------------+----------+
|_c0|_c1|                 _c2|                 _c3|                 _c4|       _c5|
+---+---+--------------------+--------------------+--------------------+----------+
|  1| 01|                 Ain|                 AIN|                 ain|      A500|
|  2| 02|               Aisne|               AISNE|               aisne|      A250|
|  3| 03|              Allier|              ALLIER|              allier|      A460|
|  5| 05|        Hautes-Alpes|        HAUTES-ALPES|        hautes-alpes|    H32412|
|  4| 04|Alpes-de-Haute-Pr...|ALPES-DE-HAUTE-PR...|alpes-de-haute-pr...|A412316152|
|  6| 06|     Alpes-Maritimes|     ALPES-MARITIMES|     alpes-maritimes| A41256352|
|  7| 07|             Ardèche|             ARDÈCHE|             ardeche|      A632|
|  8| 08|            Ardennes|            ARDENNES|            ardennes|     A6352|
|  9| 09|              Ariège|              ARIÈGE|              ariege|      A620|
| 10| 10|                Aube|                AUBE|                aube|      A100|
| 11| 11|                Aude|                AUDE|                aude|      A300|
| 12| 12|             Aveyron|             AVEYRON|             aveyron|      A165|
| 13| 13|    Bouches-du-Rhône|    BOUCHES-DU-RHÔNE|    bouches-du-rhone|     B2365|
| 14| 14|            Calvados|            CALVADOS|            calvados|     C4132|
| 15| 15|              Cantal|              CANTAL|              cantal|      C534|
| 16| 16|            Charente|            CHARENTE|            charente|      C653|
| 17| 17|   Charente-Maritime|   CHARENTE-MARITIME|   charente-maritime|  C6535635|
| 18| 18|                Cher|                CHER|                cher|      C600|
| 19| 19|             Corrèze|             CORRÈZE|             correze|      C620|
| 20| 2a|        Corse-du-sud|        CORSE-DU-SUD|        corse-du-sud|    C62323|
+---+---+--------------------+--------------------+--------------------+----------+
only showing top 20 rows
```

6.Joignez les deux fichiers. Quel type de jointure sera exécutée ? Pourquoi ?

// Inner join . Le but est de lié les deux tables par une même clefs pour garder les informations des deux tables.

```
val data_intersection=new_data.join(data_fichier_departement,
             data_fichier_departement("_c1")=== new_data("dep"),"inner")
data_intersection.show
```

Code_commune_INSEE	Nom_commune	Code_postal	Libelle_acheminement	Ligne_5	coordonnees_gps	dep	_c0	_c1	_c2	_c3	_c4	_c5
90093	SERMAMAGNY	90300	SERMAMAGNY	null	47.687801557,6.83...	90	91	90	Territoire de Bel...	TERRITOIRE DE BEL...	territoire-de-bel...	T636314163
91093	BOULLAY LES TROUX	91470	BOULLAY LES TROUX	null	48.6753515056,2.0...	91	92	91	Essonne	ESSONNE	essonne	E250
91100	BOUVILLE	91880	BOUVILLE	null	48.4326483441,2.2...	91	92	91	Essonne	ESSONNE	essonne	E250
91129	CERNY	91590	CERNY	null	48.4859798517,2.3...	91	92	91	Essonne	ESSONNE	essonne	E250
91184	COURDIMANCHE SUR ...	91720	COURDIMANCHE SUR ...	null	48.418031424,2.36...	91	92	91	Essonne	ESSONNE	essonne	E250
91186	COURSON MONTELOUP	91680	COURSON MONTELOUP	null	48.5982232022,2.1...	91	92	91	Essonne	ESSONNE	essonne	E250
91243	FONTENAY LES BRIIS	91640	FONTENAY LES BRIIS	null	48.6107631779,2.1...	91	92	91	Essonne	ESSONNE	essonne	E250
91244	FONTENAY LE VICOMTE	91540	FONTENAY LE VICOMTE	null	48.5468386132,2.4...	91	92	91	Essonne	ESSONNE	essonne	E250
91339	LINAS	91310	LINAS	null	48.6255061941,2.2...	91	92	91	Essonne	ESSONNE	essonne	E250
91386	MENNECY	91540	MENNECY	null	48.5586242184,2.4...	91	92	91	Essonne	ESSONNE	essonne	E250
91441	NAINVILLE LES ROCHES	91750	NAINVILLE LES ROCHES	null	48.5047296879,2.4...	91	92	91	Essonne	ESSONNE	essonne	E250
91461	OLLAINVILLE	91340	OLLAINVILLE	null	48.6063091787,2.2...	91	92	91	Essonne	ESSONNE	essonne	E250
91471	ORSAY	91400	ORSAY	ORSIGNY	48.7004093953,2.1...	91	92	91	Essonne	ESSONNE	essonne	E250
91661	VILLEBON SUR YVETTE	91140	VILLEBON SUR YVETTE	null	48.6956774396,2.2...	91	92	91	Essonne	ESSONNE	essonne	E250
92060	LE PLESSIS ROBINSON	92350	LE PLESSIS ROBINSON	ROBINSON	48.7804383642,2.2...	92	93	92	Hauts-de-Seine	HAUTS-DE-SEINE	hauts-de-seine	H32325
93001	AUBERVILLIERS	93300	AUBERVILLIERS	null	48.9121722626,2.3...	93	94	93	Seine-Saint-Denis	SEINE-SAINT-DENIS	seine-saint-denis	S525352
93049	NEUILLY PLAISANCE	93360	NEUILLY PLAISANCE	null	48.8643287852,2.5...	93	94	93	Seine-Saint-Denis	SEINE-SAINT-DENIS	seine-saint-denis	S525352
95181	COURCELLES SUR VI...	95650	COURCELLES SUR VI...	null	49.073025738,1.99...	95	96	95	Val-d'oise	VAL-D'OISE	val-doise	V432
95353	MAFFLIERS	95560	MAFFLIERS	null	49.083629147,2.31...	95	96	95	Val-d'oise	VAL-D'OISE	val-doise	V432
95365	MAREIL EN FRANCE	95850	MAREIL EN FRANCE	null	49.0700384798,2.4...	95	96	95	Val-d'oise	VAL-D'OISE	val-doise	V432

only showing top 20 rows

7.Écrivez le dataframe en fichier csv avec ces informations. Que remarquez vous ?

```
data_intersection.write.option("header","true").csv("C:/Users/33665/Desktop/
        Recherche_Alternance/scala.csv")
```

```
2020-01-22 16:33:18,425 - ERROR [Executor task launch worker for task 1018] org.apache.spark.executor.Executor - Exception in task 0.0 in stage 35.0 (TID 1018)
java.io.IOException: (null) entry in command string: null chmod 0644 C:\Users\33665\Desktop\Recherche_Alternance\scala.csv\_temporary\0\_temporary\attempt_2020012216331
```

Nous n'avons pas réussi a corrigé l'erreur.

8.Faites la même chose en utilisant le langage SQL plutôt que le DSL. Ecrivez ici la ou les requêtes utilisées.

```
data_fichier_departement.createOrReplaceTempView("data1")
new_data.createOrReplaceTempView("data2")
```

```
val sql_variable = spark.sql("""
  |SELECT *
  |FROM data1
  |INNER JOIN data2
  |ON data1._c1 = data2.dep
  """.stripMargin)
sql_variable.show()
```

_c0	_c1	_c2	_c3	_c4	_c5	Code_commune_INSEE	Nom_commune	Code_postal	Libelle_acheminement	Ligne_5	coordonnees_gps	dep
91	90	Territoire de Bel...	TERRITOIRE DE BEL...	territoire-de-bel...	T636314163	90093	SERMAMAGNY	90300	SERMAMAGNY	null	47.687801557,6.83...	90
92	91	Essonne	ESSONNE	essonne	E250	91093	BOULLAY LES TROUX	91470	BOULLAY LES TROUX	null	48.6753515056,2.0...	91
92	91	Essonne	ESSONNE	essonne	E250	91100	BOUVILLE	91880	BOUVILLE	null	48.4326483441,2.2...	91
92	91	Essonne	ESSONNE	essonne	E250	91129	CERNY	91590	CERNY	null	48.4859798517,2.3...	91
92	91	Essonne	ESSONNE	essonne	E250	91184	COURDIMANCHE SUR ...	91720	COURDIMANCHE SUR ...	null	48.418031424,2.36...	91
92	91	Essonne	ESSONNE	essonne	E250	91186	COURSON MONTELOUP	91680	COURSON MONTELOUP	null	48.5982232022,2.1...	91
92	91	Essonne	ESSONNE	essonne	E250	91243	FONTENAY LES BRIIS	91640	FONTENAY LES BRIIS	null	48.6107631779,2.1...	91
92	91	Essonne	ESSONNE	essonne	E250	91244	FONTENAY LE VICOMTE	91540	FONTENAY LE VICOMTE	null	48.5468386132,2.4...	91
92	91	Essonne	ESSONNE	essonne	E250	91339	LINAS	91310	LINAS	null	48.6255061941,2.2...	91
92	91	Essonne	ESSONNE	essonne	E250	91386	MENNECY	91540	MENNECY	null	48.5586242184,2.4...	91
92	91	Essonne	ESSONNE	essonne	E250	91441	NAINVILLE LES ROCHES	91750	NAINVILLE LES ROCHES	null	48.5047296879,2.4...	91
92	91	Essonne	ESSONNE	essonne	E250	91461	OLLAINVILLE	91340	OLLAINVILLE	null	48.6063091787,2.2...	91
92	91	Essonne	ESSONNE	essonne	E250	91471	ORSAY	91400	ORSAY	ORSIGNY	48.7004093953,2.1...	91
92	91	Essonne	ESSONNE	essonne	E250	91661	VILLEBON SUR YVETTE	91140	VILLEBON SUR YVETTE	null	48.6956774396,2.2...	91
93	92	Hauts-de-Seine	HAUTS-DE-SEINE	hauts-de-seine	H32325	92060	LE PLESSIS ROBINSON	92350	LE PLESSIS ROBINSON	ROBINSON	48.7804383642,2.2...	92
94	93	Seine-Saint-Denis	SEINE-SAINT-DENIS	seine-saint-denis	S525352	93001	AUBERVILLIERS	93300	AUBERVILLIERS	null	48.9121722626,2.3...	93
94	93	Seine-Saint-Denis	SEINE-SAINT-DENIS	seine-saint-denis	S525352	93049	NEUILLY PLAISANCE	93360	NEUILLY PLAISANCE	null	48.8643287852,2.5...	93
96	95	Val-d'oise	VAL-D'OISE	val-doise	V432	95181	COURCELLES SUR VI...	95650	COURCELLES SUR VI...	null	49.073025738,1.99...	95
96	95	Val-d'oise	VAL-D'OISE	val-doise	V432	95353	MAFFLIERS	95560	MAFFLIERS	null	49.083629147,2.31...	95
96	95	Val-d'oise	VAL-D'OISE	val-doise	V432	95365	MAREIL EN FRANCE	95850	MAREIL EN FRANCE	null	49.0700384798,2.4...	95

only showing top 20 rows