Spark Assignment – iNeuron

Student: Akilesh Vishnu Mohan Raj

Email: akileshvishnu10@gmail.com

1. Download the data from the given URL :

https://www.kaggle.com/datasets/kimjihoo/coronavirusdataset

Sol) The Data Sets has been downloaded from the Kaggle Site.

4. Collect your data as a pyspark dataframe and perform different operations:

Note: Consider only three files for creating a dataframe among all case, region and TimeProvince

Sol) The below Datasets have been moved to HDFS and read as Spark Dataframes:

a. Read the data, show it and Count the number of records.

```
>>>
>>> CaseDf = spark.read.option("header",True).option("inferSchema",True).csv("/input_data/Case.csv")
ionDf = spark.read.option("header",True).option("inferSchema",True).csv("/input_data/Region.csv")
TimeProvinceDf = spark.read.option("header",True).option("inferSchema",True).csv("/input_data/TimeProvince.csv")>>> RegionDf = spark.read.option("header",True).option("inferSchema",True).csv("/input_data/Region.csv")
>>> TimeProvinceDf = spark.read.option("header",True).option("inferSchema",True).csv("/input_data/TimeProvince.csv")
>>>
>>>
>>> CaseDf.show()
+-------+--------+--------------+-----+------------------+---------+--------+---------+
| case_id|province|          city|group|      infection_case|confirmed| latitude| longitude|
+-------+--------+--------------+-----+------------------+---------+--------+---------+
| 1000001|   Seoul|     Yongsan-gu| true|       Itaewon Clubs|      139|37.538621|126.992652|
| 1000002|   Seoul|      Gwanak-gu| true|            Richway|      119| 37.48208|126.901384|
| 1000003|   Seoul|        Guro-gu| true| Guro-gu Call Center|       95|37.508163|126.884387|
| 1000004|   Seoul|  Yangcheon-gu| true|Yangcheon Table T...|       43|37.546061|126.874209|
```

```
>>> CaseDf.count()
174
>>> CaseDf.select(count(" case_id")).show()
+---------------+
|count( case_id)|
+---------------+
|            174|
+---------------+
```

b. Describe the data with a describe function.

```
>>> CaseDf.describe()
DataFrame[summary: string,  case_id: string, province: string, city: string, infection_case: string, confirmed: string, latitude: string, longitude: string]
>>>
```

c. If there is any duplicate value drop it.

```
>>> CaseDf.dropDuplicates()
DataFrame[ case_id: int, province: string, city: string, group: boolean, infe
>>> RegionDf.dropDuplicates()
TimeProvinceDf.dropDuplicates()DataFrame[code: int, province: string, city: s
ty_count: int, academy_ratio: double, elderly_population_ratio: double, elde
>>> TimeProvinceDf.dropDuplicates()
DataFrame[date: string, time: int, province: string, confirmed: int, released
>>>
>>>
>>>
```

d. Use limit function for showcasing a limited number of records.

```
>>> CaseDf.limit(3).show()
meProvinceDf.limit(3).show()+--------+--------+----------+-----+------------------
| case_id|province|      city|group|     infection_case|confirmed| latitude| longit
+--------+--------+----------+-----+-------------------+---------+---------+------
| 1000001|   Seoul|Yongsan-gu| true|      Itaewon Clubs|      139|37.538621|126.992
| 1000002|   Seoul| Gwanak-gu| true|            Richway|      119| 37.48208|126.901
| 1000003|   Seoul|   Guro-gu| true|Guro-gu Call Center|       95|37.508163|126.884
+--------+--------+----------+-----+-------------------+---------+---------+------

>>> RegionDf.limit(3).show()
+-----+--------+-----------+---------+----------+----------------------+----------
--+
| code|province|       city| latitude| longitude|elementary_school_count|kindergart
nt|
+-----+--------+-----------+---------+----------+----------------------+----------
--+
|10000|   Seoul|      Seoul|37.566953|126.977977|                   607|
39|
|10010|   Seoul| Gangnam-gu|37.518421|127.047222|                    33|
88|
|10020|   Seoul|Gangdong-gu|37.530492|127.123837|                    27|
23|
+-----+--------+-----------+---------+----------+----------------------+----------
--+

>>> TimeProvinceDf.limit(3).show()
+----------+----+--------+---------+--------+--------+
|      date|time|province|confirmed|released|deceased|
```

e. If you find the column name is not suitable, change the column name.[optional]

```
>>> CaseDf.withColumnRenamed(" case_id", "case_id").show()
+-------+--------+--------------+-----+--------------------+---------+---------+----------+
|case_id|province|          city|group|      infection_case|confirmed| latitude| longitude|
+-------+--------+--------------+-----+--------------------+---------+---------+----------+
|1000001|   Seoul|     Yongsan-gu| true|       Itaewon Clubs|      139|37.538621|126.992652|
|1000002|   Seoul|      Gwanak-gu| true|             Richway|      119| 37.48208|126.901384|
|1000003|   Seoul|        Guro-gu| true| Guro-gu Call Center|       95|37.508163|126.884387|
|1000004|   Seoul|   Yangcheon-gu| true|Yangcheon Table T...|       43|37.546061|126.874209|
|1000005|   Seoul|      Dobong-gu| true|     Day Care Center|       43|37.679422|127.044374|
|1000006|   Seoul|        Guro-gu| true|Manmin Central Ch...|       41|37.481059|126.894343|
|1000007|   Seoul|from other city| true|SMR Newly Planted...|       36|        -|         -|
|1000008|   Seoul|  Dongdaemun-gu| true|       Dongan Church|       17|37.592888|127.056766|
|1000009|   Seoul|from other city| true|Coupang Logistics...|       25|        -|         -|
|1000010|   Seoul|      Gwanak-gu| true|     Wangsung Church|       30|37.481735|126.930121|
|1000011|   Seoul|   Eunpyeong-gu| true|Eunpyeong St. Mar...|       14| 37.63369|  126.9165|
|1000012|   Seoul|   Seongdong-gu| true|    Seongdong-gu APT|       13| 37.55713|  127.0403|
|1000013|   Seoul|      Jongno-gu| true|Jongno Community ...|       10| 37.57681|   127.006|
|1000014|   Seoul|     Gangnam-gu| true|Samsung Medical C...|        7| 37.48825| 127.08559|
|1000015|   Seoul|        Jung-gu| true|Jung-gu Fashion C...|        7|37.562405|126.984377|
|1000016|   Seoul|   Seodaemun-gu| true|  Yeonana News Class|        5|37.558147|126.943799|
|1000017|   Seoul|      Jongno-gu| true|Korea Campus Crus...|        7|37.594782|126.968022|
|1000018|   Seoul|     Gangnam-gu| true|Gangnam Yeoksam-d...|        6|        -|         -|
|1000019|   Seoul|from other city| true|Daejeon door-to-d...|        1|        -|         -|
|1000020|   Seoul|   Geumcheon-gu| true|Geumcheon-gu rice...|        6|        -|         -|
+-------+--------+--------------+-----+--------------------+---------+---------+----------+
only showing top 20 rows
```

f. Select the subset of the columns.

```
>>> CaseDf.select(" case_id", "city", "confirmed").show()
+--------+--------------+---------+
| case_id|          city|confirmed|
+--------+--------------+---------+
| 1000001|     Yongsan-gu|      139|
| 1000002|      Gwanak-gu|      119|
| 1000003|        Guro-gu|       95|
| 1000004|   Yangcheon-gu|       43|
| 1000005|      Dobong-gu|       43|
| 1000006|        Guro-gu|       41|
| 1000007|from other city|       36|
| 1000008|  Dongdaemun-gu|       17|
| 1000009|from other city|       25|
| 1000010|      Gwanak-gu|       30|
| 1000011|   Eunpyeong-gu|       14|
| 1000012|   Seongdong-gu|       13|
| 1000013|      Jongno-gu|       10|
```

g. If there is any null value, fill it with any random value or drop it.

```
>>> CaseDf.fillna("0").show(5)
+--------+--------+-----------+-----+--------------------+---------+---------+----------
| case_id|province|       city|group|      infection_case|confirmed| latitude| longitude
+--------+--------+-----------+-----+--------------------+---------+---------+----------
| 1000001|   Seoul| Yongsan-gu| true|       Itaewon Clubs|      139|37.538621|126.992652
| 1000002|   Seoul|  Gwanak-gu| true|             Richway|      119| 37.48208|126.901384
| 1000003|   Seoul|    Guro-gu| true| Guro-gu Call Center|       95|37.508163|126.884387
| 1000004|   Seoul|Yangcheon-gu| true|Yangcheon Table T...|       43|37.546061|126.874209
| 1000005|   Seoul|  Dobong-gu| true|     Day Care Center|       43|37.679422|127.044374
+--------+--------+-----------+-----+--------------------+---------+---------+----------
only showing top 5 rows
```

```
>>> CaseDf.dropna().show(10)
+--------+--------+--------------+-----+--------------------+---------+---------+----------+
| case_id|province|          city|group|      infection_case|confirmed| latitude| longitude|
+--------+--------+--------------+-----+--------------------+---------+---------+----------+
| 1000001|   Seoul|     Yongsan-gu| true|       Itaewon Clubs|      139|37.538621|126.992652|
| 1000002|   Seoul|      Gwanak-gu| true|             Richway|      119| 37.48208|126.901384|
| 1000003|   Seoul|        Guro-gu| true| Guro-gu Call Center|       95|37.508163|126.884387|
| 1000004|   Seoul|   Yangcheon-gu| true|Yangcheon Table T...|       43|37.546061|126.874209|
| 1000005|   Seoul|      Dobong-gu| true|     Day Care Center|       43|37.679422|127.044374|
| 1000006|   Seoul|        Guro-gu| true|Manmin Central Ch...|       41|37.481059|126.894343|
| 1000007|   Seoul|from other city| true|SMR Newly Planted...|       36|        -|         -|
| 1000008|   Seoul|  Dongdaemun-gu| true|       Dongan Church|       17|37.592888|127.056766|
| 1000009|   Seoul|from other city| true|Coupang Logistics...|       25|        -|         -|
| 1000010|   Seoul|      Gwanak-gu| true|     Wangsung Church|       30|37.481735|126.930121|
+--------+--------+--------------+-----+--------------------+---------+---------+----------+
only showing top 10 rows
```

h. Filter the data based on different columns or variables and do the best analysis.

```
>>> CaseDf.select(avg("confirmed").alias("confirmed_cases")).show()
+-----------------+
|  confirmed_cases|
+-----------------+
|65.48850574712644|
+-----------------+

>>> CaseDf.select(sum("confirmed").alias("Total_Cases")).show()
+-----------+
|Total_Cases|
+-----------+
|      11395|
+-----------+
```

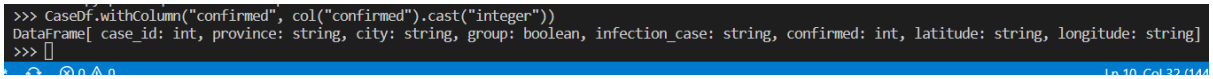
```
>>> CaseDf.filter("confirmed > 100").show()
+--------+----------------+--------------+-----+--------------------+---------+---------+----------+
| case_id|        province|          city|group|      infection_case|confirmed| latitude| longitude|
+--------+----------------+--------------+-----+--------------------+---------+---------+----------+
| 1000001|           Seoul|     Yongsan-gu| true|       Itaewon Clubs|      139|37.538621|126.992652|
| 1000002|           Seoul|      Gwanak-gu| true|             Richway|      119| 37.48208|126.901384|
| 1000036|           Seoul|             -|false|     overseas inflow|      298|        -|         -|
| 1000037|           Seoul|             -|false|contact with patient|      162|        -|         -|
| 1200001|           Daegu|         Nam-gu| true|  Shincheonji Church|     4511| 35.84008|  128.5667|
| 1200002|           Daegu|   Dalseong-gun| true|Second Mi-Ju Hosp...|      196|35.857375|128.466651|
| 1200003|           Daegu|         Seo-gu| true|Hansarang Convale...|      124|35.885592|128.556649|
| 1200004|           Daegu|   Dalseong-gun| true|Daesil Convalesce...|      101|35.857393|128.466653|
| 1200009|           Daegu|             -|false|contact with patient|      917|        -|         -|
| 1200010|           Daegu|             -|false|                 etc|      747|        -|         -|
| 2000020|      Gyeonggi-do|             -|false|     overseas inflow|      305|        -|         -|
| 4100001|Chungcheongnam-do|     Cheonan-si| true|gym facility in C...|      103| 36.81503|  127.1139|
| 6000001| Gyeongsangbuk-do|from other city| true|  Shincheonji Church|      566|        -|         -|
| 6000002| Gyeongsangbuk-do|   Cheongdo-gun| true|Cheongdo Daenam H...|      119| 35.64887|  128.7368|
| 6000012| Gyeongsangbuk-do|             -|false|contact with patient|      190|        -|         -|
| 6000013| Gyeongsangbuk-do|             -|false|                 etc|      133|        -|         -|
+--------+----------------+--------------+-----+--------------------+---------+---------+----------+

>>> CaseDf.filter("confirmed > 100").count().show()
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
AttributeError: 'int' object has no attribute 'show'
>>> CaseDf.filter("confirmed > 100").count()
16
```

i.Sort the number of confirmed cases. Confirmed column is there in the dataset. Check with descending sort also.

```
>>> CaseDf.orderBy(CaseDf["confirmed"].desc()).show()
+--------+----------------+--------------+-----+--------------------+---------+---------+----------+
| case_id|        province|          city|group|      infection_case|confirmed| latitude| longitude|
+--------+----------------+--------------+-----+--------------------+---------+---------+----------+
| 1200001|           Daegu|         Nam-gu| true|  Shincheonji Church|     4511| 35.84008|  128.5667|
| 1200009|           Daegu|             -|false|contact with patient|      917|        -|         -|
| 1200010|           Daegu|             -|false|                 etc|      747|        -|         -|
| 6000001| Gyeongsangbuk-do|from other city| true|  Shincheonji Church|      566|        -|         -|
| 2000020|      Gyeonggi-do|             -|false|     overseas inflow|      305|        -|         -|
| 1000036|           Seoul|             -|false|     overseas inflow|      298|        -|         -|
| 1200002|           Daegu|   Dalseong-gun| true|Second Mi-Ju Hosp...|      196|35.857375|128.466651|
| 6000012| Gyeongsangbuk-do|             -|false|contact with patient|      190|        -|         -|
| 1000037|           Seoul|             -|false|contact with patient|      162|        -|         -|
| 1000001|           Seoul|     Yongsan-gu| true|       Itaewon Clubs|      139|37.538621|126.992652|
| 6000013| Gyeongsangbuk-do|             -|false|                 etc|      133|        -|         -|
| 1200003|           Daegu|         Seo-gu| true|Hansarang Convale...|      124|35.885592|128.556649|
| 1000002|           Seoul|      Gwanak-gu| true|             Richway|      119| 37.48208|126.901384|
| 6000002| Gyeongsangbuk-do|   Cheongdo-gun| true|Cheongdo Daenam H...|      119| 35.64887|  128.7368|
| 4100001|Chungcheongnam-do|     Cheonan-si| true|gym facility in C...|      103| 36.81503|  127.1139|
| 1200004|           Daegu|   Dalseong-gun| true|Daesil Convalesce...|      101|35.857393|128.466653|
| 1000038|           Seoul|             -|false|                 etc|      100|        -|         -|
| 1000003|           Seoul|       Guro-gu| true| Guro-gu Call Center|       95|37.508163|126.884387|
| 2000022|      Gyeonggi-do|             -|false|                 etc|       84|        -|         -|
| 1400005|         Incheon|             -|false|     overseas inflow|       68|        -|         -|
+--------+----------------+--------------+-----+--------------------+---------+---------+----------+
```

j. In case of any wrong data type, cast that data type from integer to string or string to integer.



k. Use group by on top of province and city column and agg it with sum of confirmed cases.

df.groupBy(["province","city"]).agg(function.sum("confirmed")

(Sol)

CaseDf.groupBy("province","city").agg(sum("confirmed").alias("total_confirmed"))

l. For joins we will need one more file.you can use region file. User different different join methods.

(Sol)

CaseDf.join(RegionDf, ['province','city'],how='left').select(CaseDf.caseID,CaseDf.province,CaseDf.city)

CaseDf.join(RegionDf, ['province','city'],how='right').select(CaseDf.caseID,CaseDf.province,CasDf.city)

CaseDf.join(RegionDf, ['province','city'],how='inner').select(CaeDf.caseID,CaseDf.province,CaseDf.city)

5. If you want, you can also use SQL with data frames. Let us try to run some SQL on the cases table.

- *from pyspark.sql.functions import **
- *from pyspark.sql.functions import col*
- *CaseDf.createOrReplaceTempView("cases")*
- *spark.sql(" select * from cases limit 5").show()*
- *df = spark.sql("select province,confirmed from cases").show(3)*
- *spark.sql("select province, sum(confirmed) as sum_salary from cases group by province").show()*
- *spark.sql("select province, rank() over(partition by province order by confirmed desc) as rank_salary from cases").show()*
- *CaseDf.alias("cases1").join(CaseDf.alias("cases2") , col("cases1.city") == col("cases2.city"), "inner").select(col("cases1.city"), col("cases2.latitude"), col("cases2.longitude")).show(100)*
- *CaseDf.select(col("case_id") , col("infection_case"), upperCaseUDF(col("infection_case"))).show()*

- *def upperCase(in_str):*
  - *out_str = in_str.upper()*
  - *return out_str*

- *upperCaseUDF = udf(lambda z : upperCase(z) , StringType())*

- *CaseDf.select(col("case_id") , col("infection_case"), upperCaseUDF(col("infection_case"))).show()*

```
df = spark.sql("select province,confirmed from cases").show>>>
>>> spark.sql(" select * from cases limit 5").show()
(3)

spark.sql("select province, sum(confirmed) as sum_salary from cases group by province").show()

spark.sql("select province, rank() over(partition by province order by confirmed desc) as rank_salary from cases").show()

CaseDf.alias("cases1").join(CaseDf.alias("cases2") , col("cases1.city") == col("cases2.city"), "inner").select(col("cases1.city"), co
(100)
+--------+--------+-----------+-----+--------------------+---------+---------+----------+
| case_id|province|       city|group|      infection_case|confirmed| latitude| longitude|
+--------+--------+-----------+-----+--------------------+---------+---------+----------+
| 1000001|   Seoul| Yongsan-gu| true|       Itaewon Clubs|      139|37.538621|126.992652|
| 1000002|   Seoul|  Gwanak-gu| true|             Richway|      119| 37.48208|126.901384|
| 1000003|   Seoul|    Guro-gu| true| Guro-gu Call Center|       95|37.508163|126.884387|
| 1000004|   Seoul|Yangcheon-gu| true|Yangcheon Table T...|       43|37.546061|126.874209|
| 1000005|   Seoul|  Dobong-gu| true|     Day Care Center|       43|37.679422|127.044374|
+--------+--------+-----------+-----+--------------------+---------+---------+----------+

>>>
>>> df = spark.sql("select province,confirmed from cases").show(3)
+--------+---------+
|province|confirmed|
+--------+---------+
|   Seoul|      139|
|   Seoul|      119|
|   Seoul|       95|
+--------+---------+
only showing top 3 rows

>>>
>>> spark.sql("select province, sum(confirmed) as sum_salary from cases group by province").show()
+----------------+----------+
|        province|sum_salary|
```

```
>>>
>>> spark.sql("select province, sum(confirmed) as sum_salary from cases group by province").show()
+----------------+----------+
|        province|sum_salary|
+----------------+----------+
|          Sejong|        49|
|           Ulsan|        51|
|Chungcheongbuk-do|        60|
|      Gangwon-do|        62|
|          Gwangju|        43|
| Gyeongsangbuk-do|      1324|
|           Daegu|      6680|
| Gyeongsangnam-do|       132|
|         Incheon|       202|
|          Jeju-do|        19|
|     Gyeonggi-do|      1000|
|           Busan|       156|
|         Daejeon|       131|
|           Seoul|      1280|
|Chungcheongnam-do|       158|
|     Jeollabuk-do|        23|
|     Jeollanam-do|        25|
+----------------+----------+

>>>
>>> spark.sql("select province, rank() over(partition by province order by confirmed desc) as rank_salary from cases").show()
+----------------+-----------+
|        province|rank_salary|
+----------------+-----------+
|           Busan|          1|
|           Busan|          2|
|           Busan|          3|
|           Busan|          4|
|           Busan|          5|
```

```
>>>
>>> CaseDf.alias("cases1").join(CaseDf.alias("cases2") , col("cases1.city") == col("cases2.city"), "inner").select(col("cases1.city"), col("cases2.latit
show(100)
+--------------+---------+----------+
|          city| latitude| longitude|
+--------------+---------+----------+
|    Yongsan-gu|37.538621|126.992652|
|     Gwanak-gu| 37.48208|126.901384|
|     Gwanak-gu| 37.48208|126.901384|
|       Guro-gu|37.508163|126.884387|
|       Guro-gu|37.508163|126.884387|
|       Guro-gu|37.508163|126.884387|
|  Yangcheon-gu|37.546061|126.874209|
|  Yangcheon-gu|37.546061|126.874209|
|     Dobong-gu|37.679422|127.044374|
|       Guro-gu|37.481059|126.894343|
|       Guro-gu|37.481059|126.894343|
|       Guro-gu|37.481059|126.894343|
|from other city|        -|         -|
|from other city|        -|         -|
|from other city|        -|         -|
|from other city|        -|         -|
|from other city|        -|         -|
|from other city|        -|         -|
|from other city|        -|         -|
|from other city|        -|         -|
|from other city|        -|         -|
|from other city|        -|         -|
|from other city|        -|         -|
|from other city|        -|         -|
```

6. Create Spark UDFs. Create function casehighlow()

- *def casehighlow(CaseDf):*

  return CaseDf.withColumn("Severity", when(col("confirmed") < 50, "Less").otherwise("High")).select("city", "Severity").show()

- *casehighlowUDF = casehighlow(CaseDf)*

```
>>> def casehighlow(CaseDf):
...     return CaseDf.withColumn("Severity", when(col("confirmed") < 50, "Less").otherwise("High")).select("city", "Severity").show()
...
>>>
>>> casehighlowUDF = casehighlow(CaseDf)
+--------------+--------+
|          city|Severity|
+--------------+--------+
|    Yongsan-gu|    High|
|     Gwanak-gu|    High|
|       Guro-gu|    High|
|  Yangcheon-gu|    Less|
|     Dobong-gu|    Less|
|       Guro-gu|    Less|
|from other city|    Less|
| Dongdaemun-gu|    Less|
|from other city|    Less|
|     Gwanak-gu|    Less|
|  Eunpyeong-gu|    Less|
|  Seongdong-gu|    Less|
|     Jongno-gu|    Less|
|    Gangnam-gu|    Less|
```