Ejercicios Prácticos

Basado en el tutorial de: Relferreira https://github.com/relferreira (https://github.com/relferreira)

Tarea 1: Machine Learning con Spark

1. Genera un nuevo cluster llamado Ejercicio

**2. Carga un CSV en una Tabla de Spark llamada Diamonds utilizando la siguiente URL como origen de datos '/databricks-datasets/Rdatasets/data-001/csv/ggplot2/diamonds.csv' **

```
DROP TABLE IF EXISTS Diamonds;

CREATE TABLE Diamonds
USING csv
OPTIONS(path '/databricks-datasets/Rdatasets/data-001/csv/ggplot2/diamonds.csv'
,header=True)
```

OK

**3.Muestra los datos da toda la tabla con python. PAra ello tendrás que cargar los datos en un DataFrame **

```
SELECT * from Diamonds
```

	_c0	carat	cut	color	clarity	depth
1	1	0.23	Ideal	E	SI2	61.5
2	2	0.21	Premium	E	SI1	59.8

3	3	0.23	Good	E	VS1	56.9
4	4	0.29	Premium	I	VS2	62.4
5	5	0.31	Good	J	SI2	63.3
6	6	0.24	Very Good	J	VVS2	62.8
7	7	0.24	Very Good	I	VVS1	62.3

Truncated results, showing first 1000 rows.

```
%python
diamonds = spark.read.csv('/databricks-datasets/Rdatasets/data-
001/csv/ggplot2/diamonds.csv',header="true",inferSchema="true")

diamonds.write.format("delta").save("/delta/diamonds")
```

```
  AnalysisException: dbfs:/delta/diamonds already exists.
```

```
%python
display(diamonds)
```

	_c0	carat	cut	color	clarity	depth
1	1	0.23	Ideal	E	SI2	61.5
2	2	0.21	Premium	E	SI1	59.8
3	3	0.23	Good	E	VS1	56.9
4	4	0.29	Premium	I	VS2	62.4
5	5	0.31	Good	J	SI2	63.3
6	6	0.24	Very Good	J	VVS2	62.8
7	7	0.24	Very Good	I	VVS1	62.3

Truncated results, showing first 1000 rows.

**4.Con una consulta de SQL muestra el corte y el precio medio agrupado por el corte del diamante **

```
Select cut,avg(price) as price from Diamonds group by cut order by cut
```

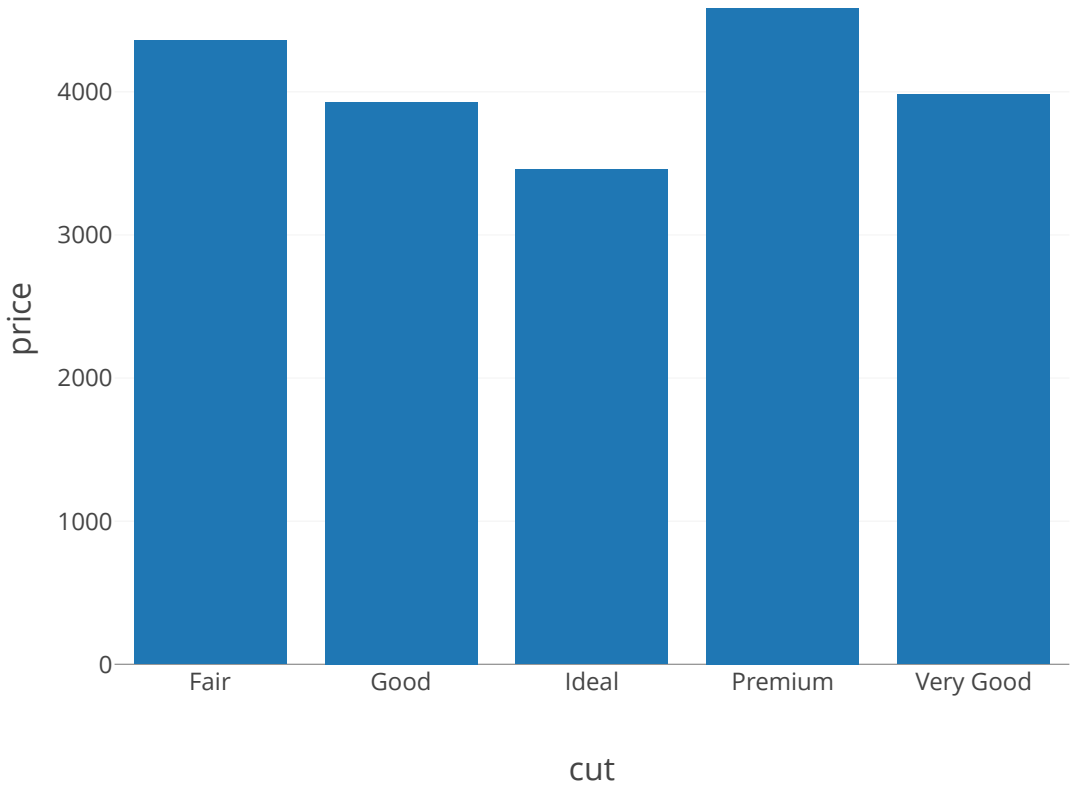
	cut	price
1	Fair	4358.757763975155
2	Good	3928.864451691806
3	Ideal	3457.541970210199
4	Premium	4584.2577042999055

5	Very Good	3981.7598907465654

Showing all 5 rows.

5.**Genera una grafica con la consulta anterior**

```
Select cut,avg(price) as price from Diamonds group by cut order by cut
```



Tarea 2: Apache Spark

1.**Crea un DataFrame con el siguiente textfile "/databricks-datasets/samples/docs/README.md"**

```
%python

textFile = spark.read.text("/databricks-datasets/samples/docs/README.md")
```

2.**Accede a la informacion del documento anterior**

```
%python

display(textFile)
```

	value
1	Welcome to the Spark documentation!
2	
3	This readme will walk you through navigating and building the Spark documentation, which is inclu
4	here with the Spark source code. You can also find documentation specific to release versions of
5	Spark at http://spark.apache.org/documentation.html.
6	
7	Read on to learn more about viewing documentation in plain text (i.e., markdown) or building the

Showing all 65 rows.

3.**Cuenta las lineas con la palabra "documentation"**

```
%python
condicion = textFile.value.contains("documentation")
textFile_fil = textFile.filter(condicion)
textFile_fil.count()
```

```
Out[4]: 10
```

Tarea 3: Databricks Datasets y Dataframes

1.**Importa el csv de "caso.csv" de la carpeta del escritorio**

2.**Importa el csv de "/databricks-datasets/samples/population-vs-price/data_geo.csv. Elimina valores faltantes"**

```
%python

dataFile = spark.read.csv("/databricks-datasets/samples/population-vs-price/data_geo.csv", header = "true", inferSchema="true")

dataFile = dataFile.dropna()

display(dataFile)
```

	2014 rank	City	State	State Code	2014 Popu
1	101	Birmingham	Alabama	AL	212247
2	125	Huntsville	Alabama	AL	188226
3	122	Mobile	Alabama	AL	194675
4	114	Montgomery	Alabama	AL	200481
5	6	Phoenix	Arizona	AZ	1537058
6	33	Tucson	Arizona	AZ	527972
7	119	Little Rock	Arkansas	AR	197706

Showing all 109 rows.

3.**Mediante SQL muestra el precion medio por estado. Para ello, previamente, tendrás que crear una tabla temporal"**

```
%python
dataFile.createOrReplaceTempView("Tabla_1")
```

```
select State, "2015 median sales price" from Tabla_1
```

	State	2015 median sales price
1	Alabama	2015 median sales price
2	Alabama	2015 median sales price
3	Alabama	2015 median sales price
4	Alabama	2015 median sales price
5	Arizona	2015 median sales price

6	Arizona	2015 median sales price
7	Arkansas	2015 median sales price

Showing all 109 rows.

Tarea 4: Machine Learning con Spark

Basado en el tutorial de: Relferreira https://github.com/relferreira (https://github.com/relferreira)

1.**Visualiza los datos del dataset de covid y analizalos"**

```
select * from caso
```

	date	state	city	place_type	c
1	2020-09-15T00:00:00.000+0000	AP	null	state	4
2	2020-09-14T00:00:00.000+0000	AP	null	state	4
3	2020-09-13T00:00:00.000+0000	AP	null	state	4
4	2020-09-12T00:00:00.000+0000	AP	null	state	4
5	2020-09-11T00:00:00.000+0000	AP	null	state	4
6	2020-09-10T00:00:00.000+0000	AP	null	state	4
7	2020-09-09T00:00:00.000+0000	AP	null	state	4

Truncated results, showing first 1000 rows.

```
describe caso
```

	col_name	data_type	comment
1	date	timestamp	null
2	state	string	null
3	city	string	null
4	place_type	string	null
5	confirmed	int	null
6	deaths	int	null
7	order_for_place	int	null

Showing all 12 rows.

2.**Visualiza los datos de date, state y deaths del estado MG, RJ o SP y del palce_type de State"**

```
select date,state, deaths from caso where state in ("MG","RJ","SP") and
place_type="state"
```

	date	state	deaths
1	2020-09-15T00:00:00.000+0000	MG	6328
2	2020-09-14T00:00:00.000+0000	MG	6286
3	2020-09-13T00:00:00.000+0000	MG	6276
4	2020-09-12T00:00:00.000+0000	MG	6200
5	2020-09-11T00:00:00.000+0000	MG	6114
6	2020-09-10T00:00:00.000+0000	MG	6009
7	2020-09-09T00:00:00.000+0000	MG	5935

Showing all 578 rows.

3.**Importa las librerias de pandas, pystan y fbprophet**

4.**Utiliza fbprophet para predecir el numero de muertes para un periodo de 30 dias del estado de MG. Para ello te recomiendo que hagas una query con los datos que vas a utilizar, despues entrenes el modelo con fit y utilices predict para la prediccion**