Creado por:

Isabel Maniega

Spark: SQL y Dataframe

Podemos encontrar el dataset en:

https://data.vermont.gov/Finance/Vermont-Vendor-Payments/786x-sbp3

In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from pyspark.sql import SparkSession
```

In [2]:
```
spark = SparkSession.builder.appName("PysparkExample") \
    .config("spark.some.config.option", "some-value") \
    .getOrCreate()
```

In [3]:
```
df = spark.read.csv("Vermont_Vendor_Payments.csv",
                    header="true", inferSchema=True)
df
```

Out[3]:
```
DataFrame[Quarter Ending: string, Department: string, UnitNo: int, Vendo
r Number: string, Vendor: string, City: string, State: string, DeptID De
scription: string, DeptID: string, Amount: string, Account: string, Acct
No: string, Fund Description: string, Fund: string]
```

In [4]:
```
# Modificamos la columna de amount a decimal:

df = df.withColumn("Amount", df["Amount"].cast("double"))
df
```

Out[4]:
```
DataFrame[Quarter Ending: string, Department: string, UnitNo: int, Vendo
r Number: string, Vendor: string, City: string, State: string, DeptID De
scription: string, DeptID: string, Amount: double, Account: string, Acct
No: string, Fund Description: string, Fund: string]
```

In [5]:
```
columns = df.columns
print("The column Names are:")

for i in columns:
    print(i)
```

```
The column Names are:
Quarter Ending
Department
UnitNo
Vendor Number
Vendor
City
State
DeptID Description
DeptID
Amount
Account
AcctNo
Fund Description
Fund
```

In [6]:
```
print("The total number of rows is:", df.count(),
      "\nThe total number columns is:", len(df.columns))
```

```
The total number of rows is: 1714538
The total number columns is: 14
```

In [7]:
```
# Visualización de las primeras filas del df:

df.show(5)
```

```
+--------------+--------------------+------+-------------+--------------
------+-----------+-----+--------------------+----------+---------+------
--------------+------+--------------------+-----+
|Quarter Ending|          Department|UnitNo|Vendor Number|
Vendor|       City|State| DeptID Description|    DeptID|   Amount|
Account|AcctNo|    Fund Description| Fund|
+--------------+--------------------+------+-------------+--------------
------+-----------+-----+--------------------+----------+---------+------
--------------+------+--------------------+-----+
|    12/31/2019|Vt Housing & Cons...|  9150|   0000002188|Vermont Housin
g &...| Montpelier|   VT|               Trust|9150120000|1075000.0|Transf
er Out - Co...|720010|Housing & Conserv...|90610|
|    12/31/2019|Vt Housing & Cons...|  9150|   0000375660|Wagner Develop
men...|Brattleboro|   VT|            VT REDI|9150293000|   4612.5|Other
Direct Gran...|552990|Housing & Conserv...|90610|
|    12/31/2019|Vt Housing & Cons...|  9150|   0000043371|Vermont Land T
rus...| Montpelier|   VT|               Trust|9150120000|112916.67|Other
Direct Gran...|552990|Housing & Conserv...|90610|
|    12/31/2019|Vt Housing & Cons...|  9150|   0000042844|University of
Ver...| Burlington|   VT|Farm Viability-VHCB|9150255000| 17152.74|Other
Direct Gran...|552990|Housing & Conserv...|90610|
|    12/31/2019|Vt Housing & Cons...|  9150|   0000160536|Lahar Stephani
e &...| Montpelier|   VT|Farm Viability-VHCB|9150255000|   4850.0|Other
Direct Gran...|552990|Housing & Conserv...|90610|
+--------------+--------------------+------+-------------+--------------
------+-----------+-----+--------------------+----------+---------+------
--------------+------+--------------------+-----+
only showing top 5 rows
```

In [8]:
```
# Con head nos muestra la primera línea:
df.head()
```

Out[8]:
```
Row(Quarter Ending='12/31/2019', Department='Vt Housing & Conserv Boar
d', UnitNo=9150, Vendor Number='0000002188', Vendor='Vermont Housing & C
onservation Board', City='Montpelier', State='VT', DeptID Description='T
rust', DeptID='9150120000', Amount=1075000.0, Account='Transfer Out - Co
mponent Units', AcctNo='720010', Fund Description='Housing & Conserv Tru
st Fund', Fund='90610')
```

In [9]:
```
df.describe().show()
```

```
+-------+--------------+--------------------+-----------------+--------
---------+--------------------+--------+-----------------+------------
------+--------------------+--------------------+--------------------+--
----------------+--------------------+--------------------+
|summary|Quarter Ending|          Department|           UnitNo|     Ven
dor Number|              Vendor|    City|            State|DeptID Descr
iption|              DeptID|              Amount|             Account|
AcctNo|    Fund Description|                Fund|
+-------+--------------+--------------------+-----------------+--------
---------+--------------------+--------+-----------------+------------
------+--------------------+--------------------+--------------------+--
----------------+--------------------+--------------------+
|  count|       1714538|             1714538|          1714538|
1714538|             1714538|  972215|          1714490|           1714
001|             1714538|             1714187|             1714538|
1714538|             1714536|             1714537|
|   mean|          null|                null| 4066.099494441068|105899.0
6434975739|                null|     0.0|1.5151515151515151|
null|4.0674150891768756E9|  185136.91537552894| 7.047635113583219E8| 532
221.4964487541|   517499.7797356828|   25998.45324564796|
| stddev|          null|                null|2330.9352198984125| 121984.
8001293792|                null|     0.0|10.605508422766931|
null| 2.330581000053294E9|1.4150774880904038E7| 5.672550213285482E8|3018
4.612746648232|   4461.381794650692|    19269.4350036219|
|    min|    03/31/2010|AOT Proprietary F...|             1100|
0000000002|"Jewett,Martin A ...|       0|                0|
"""Admin.|              CCV"""|         -2880183.34|            -294.00
|          -294.00|              507200|               10000|
|    max|    12/31/2019|  Women's Commission|             9150|
SINGLE|           xAd, Inc.|w Berlin|               ZZ|     Youth at Ri
sk|                 Seg|          6.10001E9|Youth Development...|
Water/Sewer|Youth Substance A...|Facilities Operat...|
+-------+--------------+--------------------+-----------------+--------
---------+--------------------+--------+-----------------+------------
------+--------------------+--------------------+--------------------+--
----------------+--------------------+--------------------+
```

Mostrar información

- SQL

In [10]:
```
df.createOrReplaceTempView("VermontVendor")
spark.sql(
'''
SELECT `Quarter Ending`, Department, Amount, State FROM VermontVendor
LIMIT 10
```

```
'''
).show()
```

```
+--------------+--------------------+---------+-----+
|Quarter Ending|          Department|   Amount|State|
+--------------+--------------------+---------+-----+
|    12/31/2019|Vt Housing & Cons...|1075000.0|   VT|
|    12/31/2019|Vt Housing & Cons...|   4612.5|   VT|
|    12/31/2019|Vt Housing & Cons...|112916.67|   VT|
|    12/31/2019|Vt Housing & Cons...| 17152.74|   VT|
|    12/31/2019|Vt Housing & Cons...|   4850.0|   VT|
|    12/31/2019|Vt Housing & Cons...|   1755.0|   VT|
|    12/31/2019|Vt Housing & Cons...| 26837.54|   VT|
|    12/31/2019|Vt Housing & Cons...| 30396.35|   VT|
|    12/31/2019|Vt Housing & Cons...|  5430.17|   VT|
|    12/31/2019|Vt Housing & Cons...|   1000.0|   VT|
+--------------+--------------------+---------+-----+
```

- Spark

In [11]:
```
df.select("Quarter Ending", "Department", "Amount", "State").show(10)
```

```
+--------------+--------------------+---------+-----+
|Quarter Ending|          Department|   Amount|State|
+--------------+--------------------+---------+-----+
|    12/31/2019|Vt Housing & Cons...|1075000.0|   VT|
|    12/31/2019|Vt Housing & Cons...|   4612.5|   VT|
|    12/31/2019|Vt Housing & Cons...|112916.67|   VT|
|    12/31/2019|Vt Housing & Cons...| 17152.74|   VT|
|    12/31/2019|Vt Housing & Cons...|   4850.0|   VT|
|    12/31/2019|Vt Housing & Cons...|   1755.0|   VT|
|    12/31/2019|Vt Housing & Cons...| 26837.54|   VT|
|    12/31/2019|Vt Housing & Cons...| 30396.35|   VT|
|    12/31/2019|Vt Housing & Cons...|  5430.17|   VT|
|    12/31/2019|Vt Housing & Cons...|   1000.0|   VT|
+--------------+--------------------+---------+-----+
only showing top 10 rows
```

2º Ejemplo

In [12]:
```
spark.sql(
'''
SELECT `Quarter Ending`, Department, Amount, State FROM VermontVendor
WHERE Department = 'Education'
LIMIT 10
'''
).show()
```

```
+--------------+----------+--------+-----+
|Quarter Ending|Department|  Amount|State|
+--------------+----------+--------+-----+
|    12/31/2012| Education|  302.12|   VT|
|    12/31/2012| Education|531548.0|   VT|
|    12/31/2012| Education| 14082.0|   VT|
|    12/31/2012| Education| 5337.66|   VT|
|    12/31/2012| Education|164436.0|   VT|
|    12/31/2012| Education|  8295.0|   VT|
|    12/31/2012| Education|   646.5|   VT|
|    12/31/2012| Education|    29.9|   VT|
|    12/31/2012| Education| 34159.0|   VT|
|    12/31/2012| Education|  2626.0|   VT|
+--------------+----------+--------+-----+
```

In [13]:
```
df.select("Quarter Ending", "Department", "Amount", "State") \
    .filter(df["Department"] == "Education").show(10)
```

```
+--------------+----------+--------+-----+
|Quarter Ending|Department|  Amount|State|
+--------------+----------+--------+-----+
|    12/31/2012| Education|  302.12|   VT|
|    12/31/2012| Education|531548.0|   VT|
|    12/31/2012| Education| 14082.0|   VT|
|    12/31/2012| Education| 5337.66|   VT|
|    12/31/2012| Education|164436.0|   VT|
|    12/31/2012| Education|  8295.0|   VT|
|    12/31/2012| Education|   646.5|   VT|
|    12/31/2012| Education|    29.9|   VT|
|    12/31/2012| Education| 34159.0|   VT|
|    12/31/2012| Education|  2626.0|   VT|
+--------------+----------+--------+-----+
only showing top 10 rows
```

Visualización

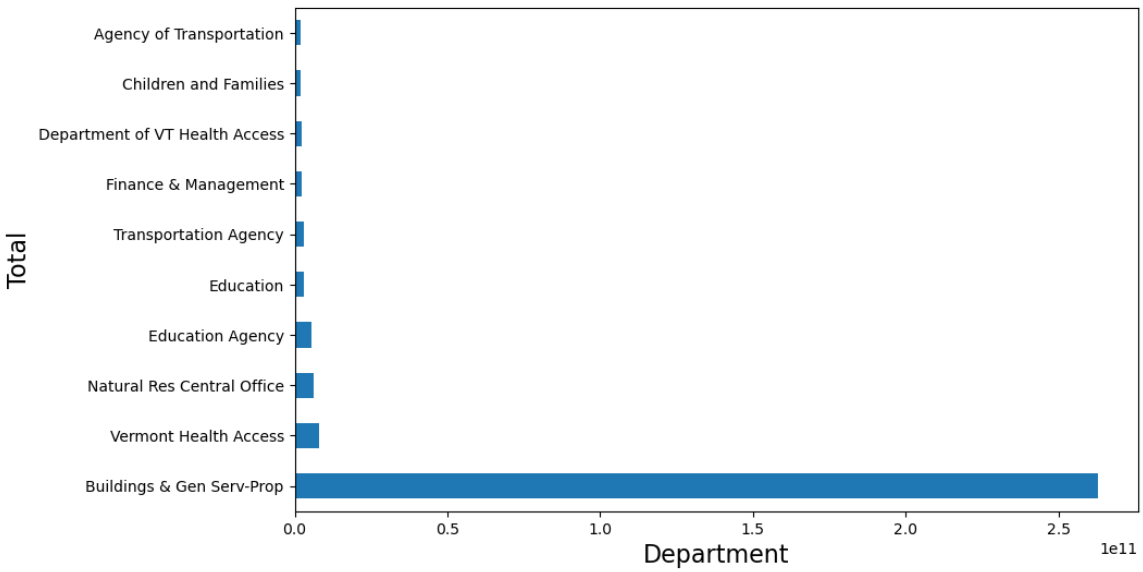
In [14]:
```
plot_df = spark.sql(
'''
SELECT Department, SUM(Amount) as Total FROM VermontVendor
GROUP BY Department
ORDER BY Total DESC
LIMIT 10
'''
).toPandas()
plot_df
```

Out[14]:

	Department	Total
0	Buildings & Gen Serv-Prop	2.628152e+11
1	Vermont Health Access	7.988020e+09
2	Natural Res Central Office	6.115936e+09
3	Education Agency	5.514746e+09
4	Education	3.166973e+09
5	Transportation Agency	2.971937e+09
6	Finance & Management	2.400270e+09
7	Department of VT Health Access	2.393175e+09
8	Children and Families	2.038478e+09
9	Agency of Transportation	1.920834e+09

In [15]:
```
fig, ax = plt.subplots(1,1, figsize=(10,6))
plot_df.plot(x='Department', y='Total',
             kind='barh', color='C0',
             ax=ax, legend=False)

ax.set_xlabel("Department", size=16)
ax.set_ylabel("Total", size=16)
plt.savefig("barplot.png")
plt.show()
```



In [16]:
```
import seaborn as sns
```

In [17]:
```
plot_df2 = spark.sql(
'''
SELECT Department, SUM(Amount) as Total FROM VermontVendor
GROUP BY Department
'''
).toPandas()
plot_df
```

Out[17]:

	Department	Total
0	Buildings & Gen Serv-Prop	2.628152e+11
1	Vermont Health Access	7.988020e+09
2	Natural Res Central Office	6.115936e+09
3	Education Agency	5.514746e+09
4	Education	3.166973e+09
5	Transportation Agency	2.971937e+09
6	Finance & Management	2.400270e+09
7	Department of VT Health Access	2.393175e+09
8	Children and Families	2.038478e+09
9	Agency of Transportation	1.920834e+09

In [18]:
```
plt.figure(figsize=(10,6))

sns.distplot(np.log(plot_df2["Total"]))

plt.title("Histogram of LOG Totals for all Departments in Dataset",
          size=16)
plt.ylabel("Density", size=16)
plt.xlabel("Log Total", size=16)
plt.savefig("distplot.png")
plt.show()
```

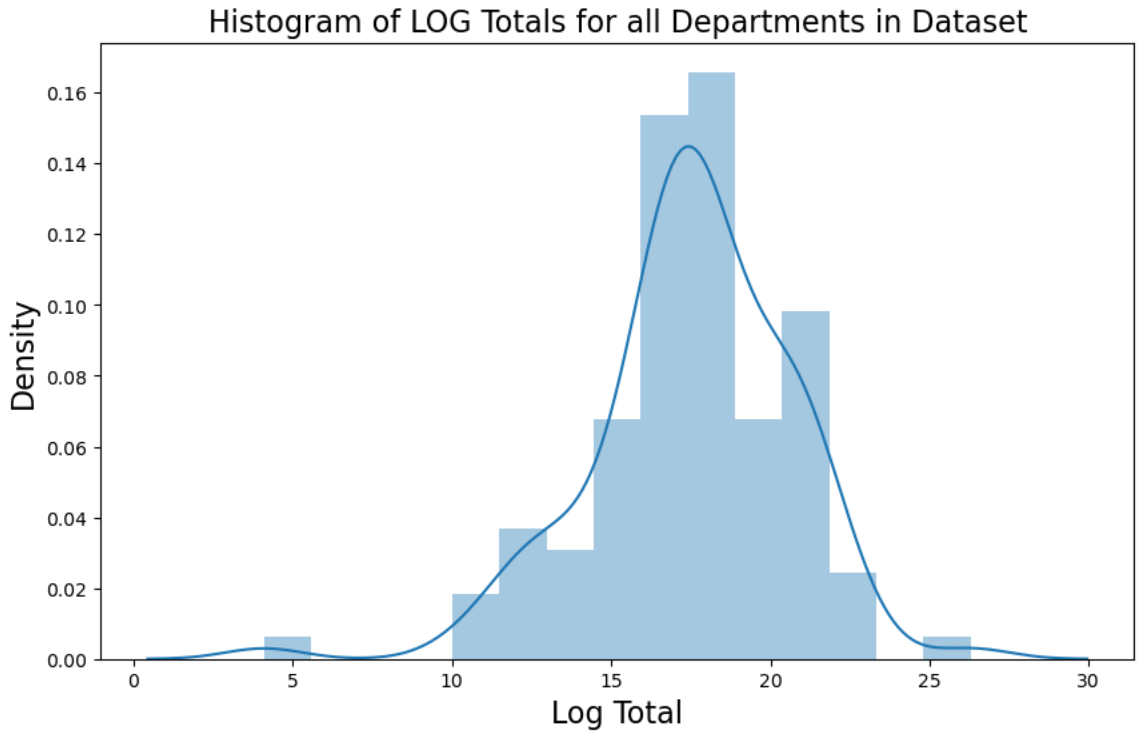
```
/tmp/ipykernel_7507/3885342777.py:3: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.1
4.0.

Please adapt your code to use either `displot` (a figure-level function
with
similar flexibility) or `histplot` (an axes-level function for histogram
s).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(np.log(plot_df2["Total"]))
```



Creado por:

Isabel Maniega