pluto19v1-18v2.1-null_comparison

September 12, 2019

[1]:
```
from pyspark.sql.functions import mean, udf, col, round, isnan, when, count, lit
from pyspark.sql.types import DoubleType, StringType
from pyspark.context import SparkContext
from pyspark.sql.session import SparkSession
import pandas as pd
import numpy as np
import time
import matplotlib.pyplot as plt
import datetime
print(datetime.datetime.now())
%matplotlib inline

sc = SparkContext('local')
spark = SparkSession(sc)
```

```
2019-09-12 14:04:34.456160
```

0.1 import csv files into spark dataframes

Note: both files contain records from all 5 boroughs

[2]:
```
df1 = spark.read.csv('../data/pluto.csv', header=True)
df2 = spark.read.csv('../data/pluto_18v2_1.csv', header=True)
```

[3]:
```
df1 = df1.select([col(A).alias(A.lower()) for A in df1.schema.names])
df2 = df2.select([col(A).alias(A.lower()) for A in df2.schema.names])
```

[4]:
```
double_columns = ['bldgarea', 'facilfar',
                  'residfar', 'commfar', 'numbldgs', 'numfloors', 'bldgdepth',
                  'bldgfront', 'lotdepth', 'lotfront',
                  'exempttot', 'exemptland',  'assessland', 'assesstot',␣
 ↪'builtfar']
```

[5]:
```
df1 = df1.withColumn('exemptland', lit(None).cast(StringType()))
```

[6]:
```
cols = df2.columns
```

[7]:
```
df1 = df1.select(cols)
df2 = df2.select(cols)
```

0.2 Type Conversion

[8]:

```
for A in double_columns:
    df1 = df1.withColumn(A, round(col(A).cast(DoubleType()), 2))
    df2 = df2.withColumn(A, round(col(A).cast(DoubleType()), 2))
```

0.3 Count Null and 0

[9]:

```
null_1 = df1.select([(count(when(isnan(c) | col(c).isNull(), 1))\
                      + count(when(col(c)==0,1))).alias(c) for c in df1.
 ↪columns]).toPandas()
condo_null_1 = df1.filter(df1['lot'].rlike(r'^75'))\
                    .select([(count(when(isnan(c) | col(c).isNull(), 1))\
                      + count(when(col(c)==0,1))).alias(c) for c in df1.
 ↪columns]).toPandas()
```

[10]:

```
null_1
```

[10]:

```
   borough  block  lot    cd  ct2010  cb2010  schooldist  council  zipcode  \
0        0      1    1  3124    3611    3124        3842     3125    25347

   firecomp  ...  firm07_flag  pfirm15_flag  rpaddate  dcasdate  zoningdate  \
0      3862  ...       826394        795384         7         0           0

   landmkdate  basempdate  masdate  polidate  edesigdate
0           0           0   861089    861096           7

[1 rows x 96 columns]
```

[11]:

```
condo_null_1
```

[11]:

```
   borough  block  lot  cd  ct2010  cb2010  schooldist  council  zipcode  \
0        0      0    0  78      85      78          90       78      272

   firecomp  ...  firm07_flag  pfirm15_flag  rpaddate  dcasdate  zoningdate  \
0        90  ...        11834         11379         0         0           0

   landmkdate  basempdate  masdate  polidate  edesigdate
0           0           0    12490     12490           0

[1 rows x 96 columns]
```

[ ]:

```
null_2 = df2.select([(count(when(isnan(c) | col(c).isNull(), 1))\
                      + count(when(col(c)==0,1))).alias(c) for c in df2.
 ↪columns]).toPandas()
condo_null_2 = df2.filter(df2['lot'].rlike(r'^75'))\
                    .select([(count(when(isnan(c) | col(c).isNull(), 1))\
                      + count(when(col(c)==0,1))).alias(c) for c in df2.
 ↪columns]).toPandas()
```

[ ]:
```
null_all = pd.concat([null_2, null_1])
condo_null_all = pd.concat([condo_null_2, condo_null_1])

null_change = null_all.pct_change()
condo_null_change = condo_null_all.pct_change()
```

[ ]:
```
null_change.iloc[1,:].sort_values(ascending=False)[0:10]
```

[ ]:
```
condo_null_change.iloc[1,:].sort_values(ascending=False)[0:10]
```

[ ]:
```
null_all.index = ['18v2.1', '19v1']
null_all
```

[ ]:
```
condo_null_all.index = ['18v2.1', '19v1']
condo_null_all
```

[ ]:
```
plt.figure(figsize=(6, 30))

difference1 = null_all.iloc[1, :]-null_all.iloc[0, :]

plt.plot(difference1, range(96), label = '19v1-18v2.1', color = 'blue')
plt.vlines(0, 0, 96) #0 reference line

for i in range(96):
    if abs(difference1[i]) >= 10000:
        plt.text(x = difference1[i] , y = i - 0.15, s = '{}'.
 ↪format(difference1[i]), size = 10, color = 'blue')
    else:
        pass

plt.yticks(range(96), null_all.columns, rotation='horizontal')
plt.title('19v1-18v2.1 Null 0 Counts Comparison')
plt.legend()
plt.show()
```

[ ]:
```
plt.figure(figsize=(6, 30))

plt.plot(null_change.iloc[1,:], range(96), label = '19v1-18v2.1', color =␣
 ↪'blue')
plt.vlines(0, 0, 96) #0 reference line

for i in range(96):
    if abs(null_change.iloc[1,i]) <= 100:
        plt.text(x = null_change.iloc[1,i] , y = i - 0.15, s = '{}'.format(np.
 ↪round(null_change.iloc[1,i], 2)), size = 10, color = 'blue')
    else:
        pass

plt.yticks(range(96), null_all.columns, rotation='horizontal')
plt.title('19v1-18v2.1 Null 0 Counts Comparison pct difference')
plt.legend()
plt.savefig('19v1-18v2.1-Null0-Comparison-pct.png', bbox_inches='tight')
plt.show()
```

[ ]:

```
plt.figure(figsize=(6, 30))

plt.plot(condo_null_change.iloc[1,:], range(96), label = '19v1_18v2.1', color =␣
 ↪'blue')
plt.vlines(0, 0, 96) #0 reference line

for i in range(96):
    if abs(condo_null_change.iloc[1,i]) <= 100:
        plt.text(x = condo_null_change.iloc[1,i] , y = i - 0.15, s = '{}'.
 ↪format(np.round(condo_null_change.iloc[1,i], 2)), size = 10, color = 'blue')
    else:
        pass

plt.yticks(range(96), condo_null_change.columns, rotation='horizontal')
plt.title('CONDO 19v1-18v2.1 Null 0 Counts Comparison pct difference')
plt.legend()
plt.savefig('CONDO 19v1-18v2.1-Null0-Comparison-pct.png', bbox_inches='tight')
plt.show()
```

[ ]:

```
import os

os.system('jupyter nbconvert --to pdf pluto19v1-18v2.1-null_comparison.ipynb')
```