Cheatography

Pyspark Cheat Sheet

by mitcht via cheatography.com/50563/cs/14121/

Read / Write .csv

```
df =
(sqlContext.read.format("com.databricks.spark.csv")\
             .op tio n("h ead er", " tru e")\
             .op tio n("i nfe rsc hem a", " tru e")\
             .op tio n("m ode ", " DRO PMA LFO RME -
D")\
             .lo ad( " hdf s:/ /fi le.c sv "))
df.wri te.m od e(' ove rwr ite').
                  opt ion ("he ade r", " tru e").
                  csv ("fi le: //f ile nam e.c sv")
```

Meta Data

```
df.pri ntS chema()              df.count()
len(df.co lumns)                df.columns
df.dtypes
```

Arrange

```
df.withColumnRenamed("col","newcol")
df.ord erB y([ 'var1', 'var2'], ascending = [True,
False])
df.ord erB y(d f.var1, df.var 2.d esc())
```

Filter

```
df.filter(df.var > 10000)
df.sel ect ('c ol1', 'col2')
df[col list] # collist = ['var1', ...]
df.head(5) / df.show(5, trunca te= True) /
df.take(5)
df.dro p(' var')
df.dis tinct()
df.dro pDu pli cates()
df.dro pna (su bse t=' var') / df.na.d rop()
df.isN otN ull()
df.var.is in( " lev el1 ", " lev el2 ")
df.var.li ke( " str ing ")
df.var.st art swi th( " m") / df.var.en dsw ith -
("m")
df.sam ple (Fals e, 0.5, 12345)
```

with replac ement *fraction* *seed*

Useful Functions

Write Functions / UDF

```
from pyspark.sql.functions import udf
F1 = udf(lambda x: '-1' if condition else x,
String Type()) # NB return type
df = df.wit hCo lum n(' new var', F1(df[ 'in -
var']))
```

Applying Functions

```
df.select('val').map(lambda x: x*2)
```

Summarise

```
df.crosstab('col1', 'col2') # pair-wise count
df.gro upb y(' var ').f un ction()
df.gro upb y(' var ').a gg ({'val' : 'mean'})
```

Join

```
df3 = df1.join(df2, [df1.var1 == df2.var1,
df1.var2 == df2.var2], 'left')
xtra = df1.se lec t(' var ').s ub tra ct( df2.se -
lec t(' var')) # anti-join
```

Method Chaining

```
df
            .se lec t("c ol1 " ,"co l2", " col 3",
...)
            .fi lte r(d f.col1 > 30 )
            .show()
```

New Variable / Column

```
df.withColumn('varnew', df.var / 2.0)
```

To SQL / Pandas

```
df.registerAsTable('df_tbl')
sqlCon tex t.s ql( 'select var from df_tbl ').s -
how(5)
df.toP andas()
```

```
.descr ibe ('o pti ona l_va r')   .count()
.show()                           .filln a(v alue)
.min() / .max()                   .mean()
.stdev() / .varia nce()           .subst r(1,3)
F.when (df.var > 30, " Y").o th erw ise ("N")
df.var.al ias ('n ewvar') # used to rename something
df.cache() / df.unp ers ist ()    .cast( 'Do uble')
.repla ce(10, 20) # ????          .na.fi ll(0, subset = 'var' )
time()
from pyspar k.sql import functions as F
```

By **mitcht**
cheatography.com/mitcht/

Not published yet.
Last updated 29th December, 2017.