

rawDF: org.apache.spark.sql.DataFrame = [Index: timestamp, Open: double ... 6 more fields]

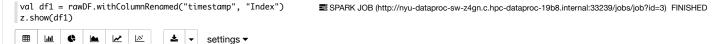
Took 49 sec. Last updated by Is6211\_nyu\_edu at December 04 2023, 11:15:16 PM.

#### 1. Column Description Infer

FINISHED

First I convert the column name to more readable names since the database lack column description.

Took 3 sec. Last updated by Is6211\_nyu\_edu at December 04 2023, 11:15:19 PM.



Index ~	Open ~	High ~	Low ~	Close	Volume_(BTC) ≡	
2011-12-31 07:52:00.0	4.39	4.39	4.39	4.39	0.45558087	
2011-12-31 07:53:00.0	4.39	4.39	4.39	4.39	0.555046181987448	
2011-12-31 07:54:00.0	4.39	4.39	4.39	4.39	0.654511493974895	
2011-12-31 07:55:00.0	4.39	4.39	4.39	4.39	0.753976805962343	
2011-12-31 07:56:00.0	4.39	4.39	4.39	4.39	0.853442117949791	
2011-12-31 07:57:00.0	4.39	4.39	4.39	4.39	0.952907429937239	
2011-12-31 07:58:00.0	4.39	4.39	4.39	4.39	1.05237274192469	
2011-12-31 07:59:00.0	4.39	4.39	4.39	4.39	1.15183805391213	

Output is truncated to 1000 rows. Learn more about zeppelin.spark.maxResult

df1: org.apache.spark.sql.DataFrame = [Index: timestamp, Open: double ... 6 more fields]

Took 1 sec. Last updated by Is6211\_nyu\_edu at December 04 2023, 11:15:20 PM.

import org.apache.spark.sql.functions.{min, max}
import org.apache.spark.sql.Row

SPARK JOB FINISHED

# projectinprofiting maxValue: Double) = df1.agg(min("Close"), max("Close")).head

import org.apache.spark.sql.functions.{min, max}

import org.apache.spark.sql.Row
minValue: Double = 1.5
maxValue: Double = 19665.75

Took 9 sec. Last updated by Is6211\_nyu\_edu at December 04 2023, 11:15:29 PM.

### From here we can infer the currency is US dollar

**FINISHED** 

Took 0 sec. Last updated by Is6211\_nyu\_edu at December 04 2023, 11:15:29 PM.

val df2 = df1.withColumnRenamed("Volume\_(BTC)", "btc\_volume").withColumnRenamed("Volume\_(Currency)", "usd\_volume")

FINISHED

df2: org.apache.spark.sql.DataFrame = [Index: timestamp, Open: double ... 6 more fields]

Took 1 sec. Last updated by Is6211\_nyu\_edu at December 04 2023, 11:15:30 PM.

#### z.show(df2)

SPARK JOB (http://nyu-dataproc-sw-z4gn.c.hpc-dataproc-19b8.internal:33239/jobs/job?id=6) FINISHED



Index ~	Open ~	High ~	Low	Close	btc_volume ≡
2011-12-31 07:52:00.0	4.39	4.39	4.39	4.39	0.45558087
2011-12-31 07:53:00.0	4.39	4.39	4.39	4.39	0.555046181987448
2011-12-31 07:54:00.0	4.39	4.39	4.39	4.39	0.654511493974895
2011-12-31 07:55:00.0	4.39	4.39	4.39	4.39	0.753976805962343
2011-12-31 07:56:00.0	4.39	4.39	4.39	4.39	0.853442117949791
2011-12-31 07:57:00.0	4.39	4.39	4.39	4.39	0.952907429937239
2011-12-31 07:58:00.0	4.39	4.39	4.39	4.39	1.05237274192469
2011-12-31 07:59:00.0	4.39	4.39	4.39	4.39	1.15183805391213
0044 40 04 00 00 00 0	4.00	4.00	4.00	4.00	4 05400000500050

## Output is truncated to 1000 rows. Learn more about zeppelin.spark.maxResult

**FINISHED** 

Took 0 sec. Last updated by Is6211\_nyu\_edu at December 04 2023, 11:15:30 PM.

#### 2. Datatype check

Took 0 sec. Last updated by Is6211\_nyu\_edu at December 04 2023, 11:15:30 PM.

df2.printSchema FINISHED

root

I-- Index: timestamp (nullable = true)

I-- Open: double (nullable = true)

I-- High: double (nullable = true)

I-- Low: double (nullable = true)

|-- Close: double (nullable = true)

I-- btc\_volume: double (nullable = true)
I-- usd\_volume: double (nullable = true)

I-- Weighted\_Price: double (nullable = true)

Took 0 sec. Last updated by Is6211\_nyu\_edu at December 04 2023, 11:15:31 PM.

The data type is fine. The 'infer' option has already convert the String type timestamp to timestamp data type

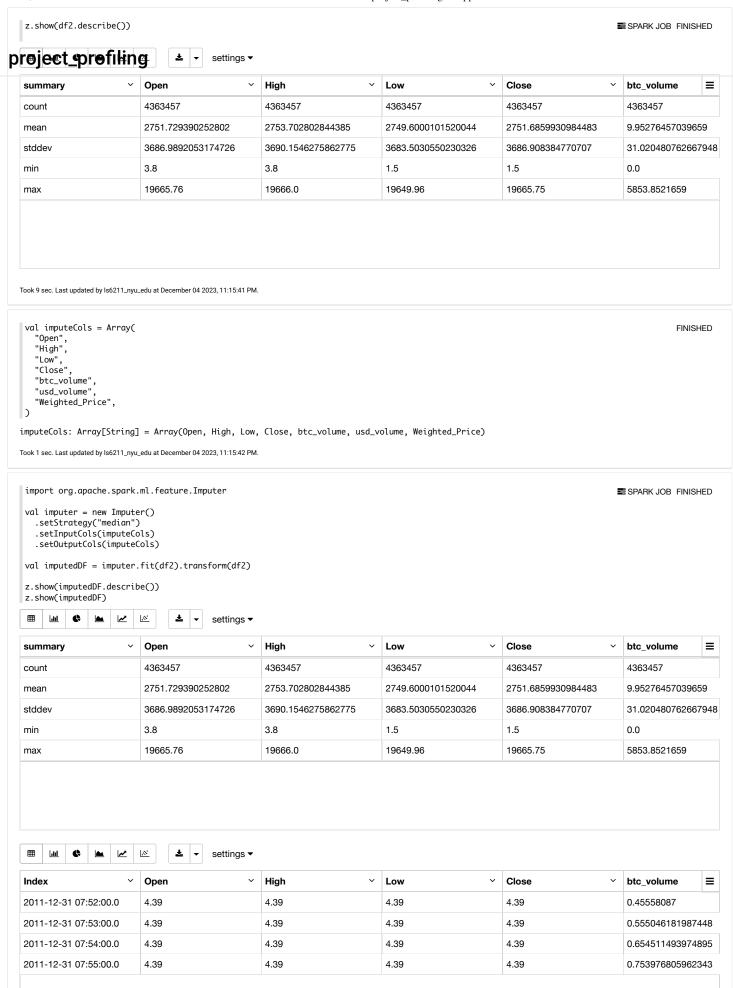
FINISHED

Took 0 sec. Last updated by Is6211\_nyu\_edu at December 04 2023, 11:15:31 PM.

## 3. statistical check

FINISHED

Took 0 sec. Last updated by Is6211\_nyu\_edu at December 04 2023, 11:15:32 PM.



	2011-12-31 07:56:00.0	4.39	4.39	4.39	4.39	0.853442117949791	
	2011-12-31 07:57:00.0	4.39	4.39	4.39	4.39	0.952907429937239	
p	roject_profiling	<b>g</b> i.39	4.39	4.39	4.39	1.05237274192469	
	2011-12-31 07:59:00.0	4.39	4.39	4.39	4.39	1.15183805391213	

#### Output is truncated to 1000 rows. Learn more about zeppelin.spark.maxResult

×

import org.apache.spark.ml.feature.Imputer

imputer: org.apache.spark.ml.feature.Imputer = imputer\_eb11547106f6

imputedDF: org.apache.spark.sql.DataFrame = [Index: timestamp, Open: double ... 6 more fields]

Took 27 sec. Last updated by Is6211\_nyu\_edu at December 04 2023, 11:16:09 PM.

## 4. Getting rid of extreme or abnormal values

FINISHED

Took 0 sec. Last updated by Is6211\_nyu\_edu at December 04 2023, 11:16:09 PM.

imputedDF.filter(\$"Weighted\_Price" <= 0).count</pre>

res7: Long = 0

Took 8 sec. Last updated by Is6211\_nyu\_edu at December 04 2023, 11:16:17 PM.

val valueDF = imputedDF.withColumn("PctChange",(\$"Close"- \$"Open")/\$"Open")

FINISHED

valueDF: org.apache.spark.sql.DataFrame = [Index: timestamp, Open: double ... 7 more fields]

settinas -

4.39

Took 1 sec. Last updated by Is6211\_nyu\_edu at December 04 2023, 11:16:18 PM.

4.39

<u>₩</u> | Ø

z.show(valueDF)

2011-12-31 07:59:00.0

z.show(valueDF.describe())

SPARK JOB FINISHED

1.15183805391213

SPARK JOB FINISHED

Open ~	High ~	Low	Close	btc_volume ~	usd_volu <b>==</b> 6
4.39	4.39	4.39	4.39	0.45558087	2.00000019
4.39	4.39	4.39	4.39	0.555046181987448	2.436652738
4.39	4.39	4.39	4.39	0.654511493974895	2.873305458
4.39	4.39	4.39	4.39	0.753976805962343	3.309958178
4.39	4.39	4.39	4.39	0.853442117949791	3.746610897
4.39	4.39	4.39	4.39	0.952907429937239	4.183263617
4.39	4.39	4.39	4.39	1.05237274192469	4.619916337
	4.39 4.39 4.39 4.39 4.39 4.39	4.39     4.39       4.39     4.39       4.39     4.39       4.39     4.39       4.39     4.39       4.39     4.39       4.39     4.39	4.39       4.39       4.39         4.39       4.39       4.39         4.39       4.39       4.39         4.39       4.39       4.39         4.39       4.39       4.39         4.39       4.39       4.39	4.39       4.39       4.39       4.39         4.39       4.39       4.39       4.39         4.39       4.39       4.39       4.39         4.39       4.39       4.39       4.39         4.39       4.39       4.39       4.39         4.39       4.39       4.39       4.39         4.39       4.39       4.39       4.39	4.39       4.39       4.39       0.45558087         4.39       4.39       0.555046181987448         4.39       4.39       0.555046181987448         4.39       4.39       0.654511493974895         4.39       4.39       0.753976805962343         4.39       4.39       0.853442117949791         4.39       4.39       0.952907429937239

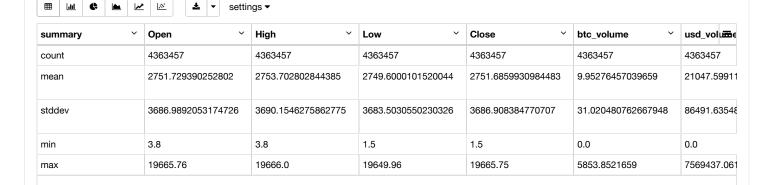
4.39

4.39

## Output is truncated to 1000 rows. Learn more about zeppelin.spark.maxResult

×

5.056569056



Took 12 sec. Last updated by Is6211\_nyu\_edu at December 04 2023, 11:16:30 PM.

The percentage of price change with one minute is reasonable.

FINISHED

# project1%profiting less than 100% decrease. Standard deviation is 0.2%.

Took 0 sec. Last updated by Is6211\_nyu\_edu at December 04 2023, 11:16:30 PM.

import org.apache.spark.sql.expressions.Window

FINISHED

val w = Window.orderBy("Index")

import org.apache.spark.sql.expressions.Window

w: org.apache.spark.sql.expressions.WindowSpec = org.apache.spark.sql.expressions.WindowSpec@2fcb4b80

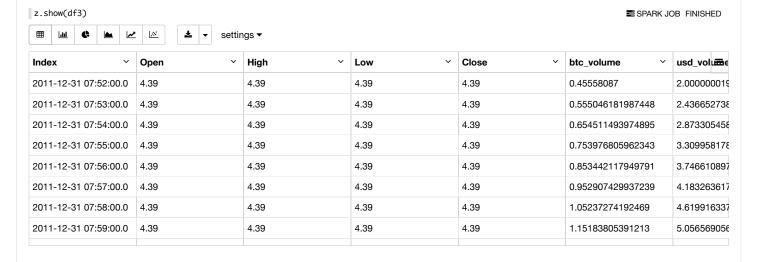
Took 1 sec. Last updated by Is6211\_nyu\_edu at December 04 2023, 11:16:31 PM.

val df3 = imputedDF.withColumn("return", (col("Close") - lag("Close", 1).over(w)) / lag("close", 1).over(w))

**FINISHED** 

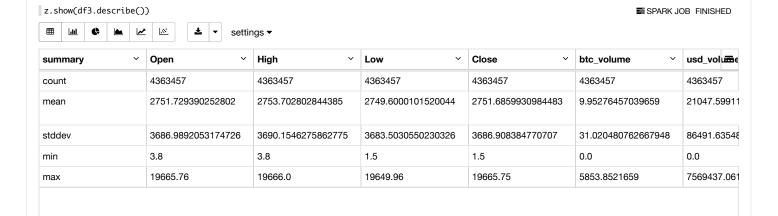
df3: org.apache.spark.sql.DataFrame = [Index: timestamp, Open: double ... 7 more fields]

Took 0 sec. Last updated by Is6211\_nyu\_edu at December 04 2023, 11:16:31 PM



#### Output is truncated to 1000 rows. Learn more about zeppelin.spark.maxResult

Took 29 sec. Last updated by Is6211\_nyu\_edu at December 04 2023, 11:17:00 PM.



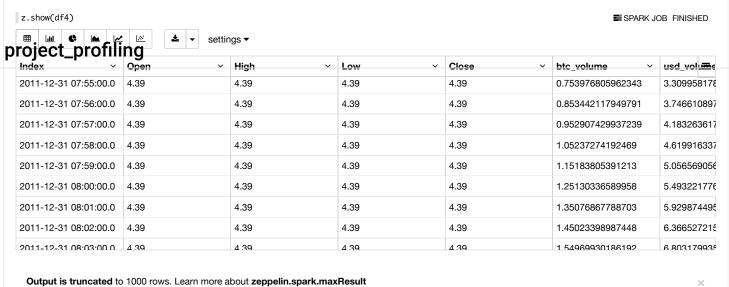
Took 24 sec. Last updated by Is6211\_nyu\_edu at December 04 2023, 11:17:24 PM

val df4 = df3.na.drop("any")

**FINISHED** 

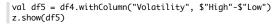
df4: org.apache.spark.sql.DataFrame = [Index: timestamp, Open: double ... 7 more fields]

Took 0 sec. Last updated by Is6211\_nyu\_edu at December 04 2023, 11:17:25 PM.



#### Output is truncated to 1000 rows. Learn more about zeppelin.spark.maxResult

Took 23 sec. Last updated by Is6211\_nyu\_edu at December 04 2023, 11:17:48 PM.



SPARK JOB FINISHED

▦	<u>lılıl</u>	¢	<u></u>	<u>~</u>	<u> M</u>	Ŧ	•	settings ▼

Index ~	Open v	High ~	Low	Close	btc_volume ~	usd_volume Ξ
11:32:00.0						
2011-12-31 11:33:00.0	4.39	4.39	4.39	4.39	22.4374148192259	98.5002510564019
2011-12-31 11:34:00.0	4.39	4.39	4.39	4.39	22.5368801312134	98.9369037760268
2011-12-31 11:35:00.0	4.39	4.39	4.39	4.39	22.6363454432008	99.3735564956517
2011-12-31 11:36:00.0	4.39	4.39	4.39	4.39	22.7358107551883	99.8102092152766
2011-12-31	4.39	4.39	4.39	4.39	22.8352760671757	100.246861934901

#### Output is truncated to 1000 rows. Learn more about zeppelin.spark.maxResult

df5: org.apache.spark.sql.DataFrame = [Index: timestamp, Open: double ... 8 more fields]

Took 27 sec. Last updated by Is6211\_nyu\_edu at December 04 2023, 11:27:53 PM. (outdated)

## z.show(df5.describe()) settings ▼

SPARK JOB FINISHED

ıry	∨ Open ∨	High ~	Low	Close	btc_volume ~	usd_volume ~	Weight
	4363456	4363456	4363456	4363456	4363456	4363456	436345
	2751.7300198774374	2753.703432921279	2749.600639288636	2751.6866227131372	9.952766746924485	21047.603940597306	2751.66
	3686.989393221721	3690.154815717843	3683.5032426696835	3686.9085726680382	31.0204839840625	86491.64481331731	3686.88
	3.8	3.8	1.5	1.5	0.0	0.0	3.8
	19665.76	19666.0	19649.96	19665.75	5853.8521659	7569437.0613	19663.2

Took 24 sec. Last updated by Is6211\_nyu\_edu at December 04 2023, 11:29:09 PM. (outdated)

localhost:33618/#/notebook/2JK5Z6EHM