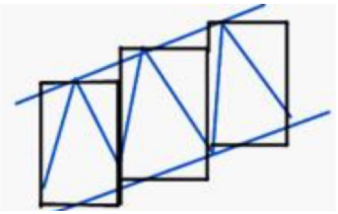
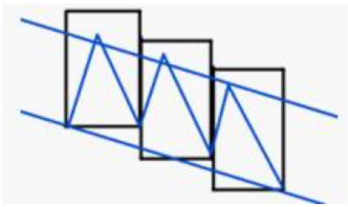
Big Data Analysis project (midterm)
Bitcoin price prediction

NCCU Executive Master of Computer Science
Directed by Prof. Hu Yu Chung
2018.04.23

106971002 林靖耀
106971008 余昊祥
106971017 王崇飛

Box theory

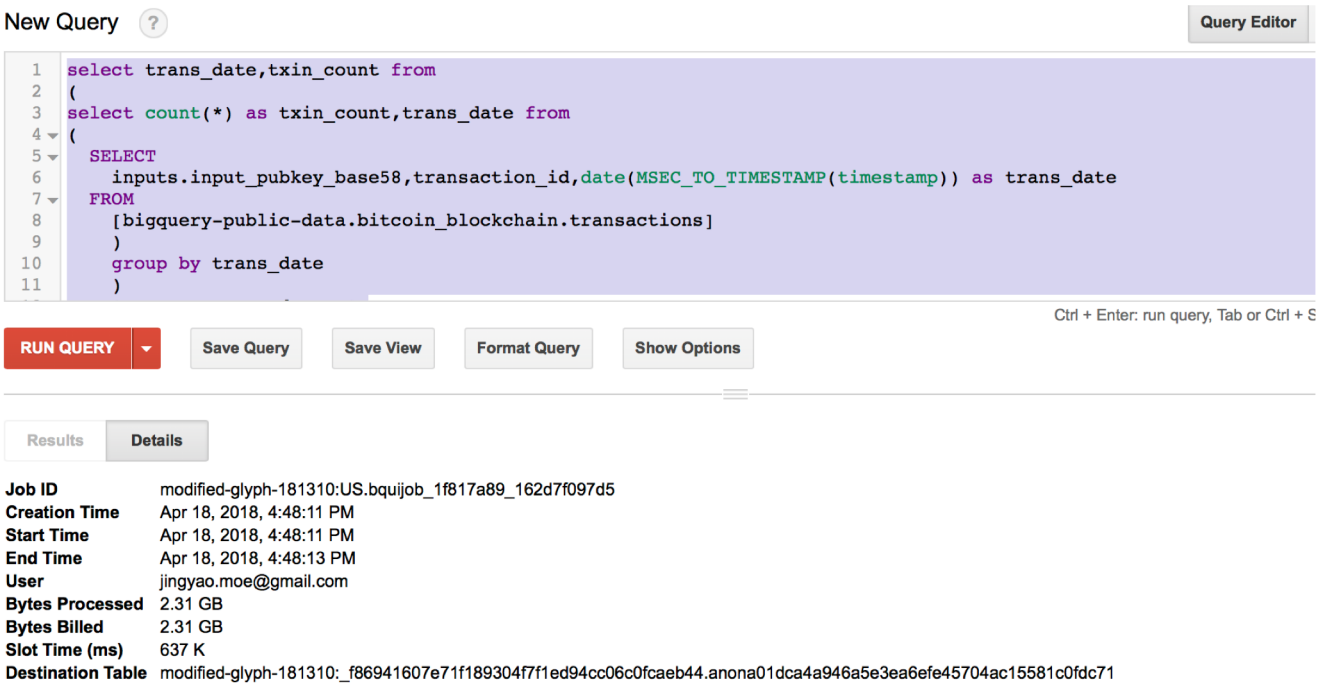




http://wiki.mbalib.com/zh-tw/股票箱理论

Big query to get transaction data

- Aggregate by date
  - Date
  - Count Of Transaction In
  - Count Of Transaction Out
  - Total satoshi in transcaction



Coindesk exchange history data

- Aggregate by date
  - Date
  - Open Price
  - Highest Price
  - Lowest Price
  - Close Price

https://www.coindesk.com/price/

Feature Engineering

- 挑選資料集欄位
  - 以交易量、交易金額及市場價格波動進行預測

trans_date	txin_#	txout_#	sum_satoshi	sum_BTC	Open	High	Low	Close
2013/4/28	41879	173840	5.79462E+13	579461.8822	135.3	135.98	132.1	134.21
2013/4/29	51776	166534	1.88956E+14	1889559.779	134.44	147.49	134	144.54
2013/4/30	47599	146931	1.00334E+14	1003343.628	144	146.93	134.05	139
2013/5/1	55327	151991	1.35072E+14	1350721.891	139	139.89	107.72	116.99
2013/5/2	55429	157671	9.07309E+13	907309.085	116.38	125.6	92.28	105.21
2013/5/3	51412	144723	8.09308E+13	809308.004	106.25	108.13	79.1	97.75
2013/5/4	50261	171517	5.04688E+13	504688.3552	98.1	115	92.5	112.5
2013/5/5	50332	153342	6.38183E+13	638182.8299	112.9	118.8	107.14	115.91
2013/5/6	53340	168310	8.58406E+13	858406.0152	115.98	124.66	106.64	112.3

- 特徵選取
  - 使用每日變化率取代純交易價/量

nextUpDow	pUpDown	ratiotxin	ratiotxout	ratiotxinout	ratiotxBTC	ratiooopen	ratiohigh	ratiolow	ratioclose
1	1	-0.164619165	-0.181481481	0.766917293	0.040888661	0	0.166666667	0	0.166666667
0	1	0.067647059	0.113122172	1.264705882	0.188267056	0.166666667	0.285714286	-0.833333333	0.285714286
1	0	0.104683196	0.130081301	1.201550388	0.203346836	0.285714286	0.333333333	6	0
0	1	-0.094763092	-0.14028777	0.741935484	-0.342087427	0	0.083333333	0.142857143	0.111111111
0	0	-0.05785124	-0.052301255	0.965217391	-0.506795192	0.111111111	-0.230769231	0.125	-0.1
1	0	0.029239766	0.061810155	1.162162162	0.366433891	-0.1	0	-0.111111111	0
0	1	0.034090909	0.018711019	0.976744186	0.433769828	0	0	0.125	0.111111111

Feature Engineering

- 特徵定義

nextUpDown	隔天漲(1)或不漲(0)
pUpDown	當天漲(1)或不漲(0)
ratiotxin	(當天 tx in數量-前一天 tx in數量) / 前一天 tx in 數量
rationtxout	(當天 tx out數量-前一天 tx out數量) / 前一天 tx out 數量
rationtxinout	(當天txin-out數量差-前一天txin-out數量差) / 前一天 txin-out數量差
ratiotxBTC	(當天總交易BTC – 前一天總交易BTC) / 前一天總交易BTC
ratioopen	(當天開盤美金價 – 前一天開盤美金價) / 前一天開盤美金價
ratiohigh	(當天最高美金價 – 前一天最高美金價) / 前一天最高美金價
ratiolow	(當天最低美金價 – 前一天最低高美金價) / 前一天最低美金價
ratioclose	(當天收盤美金價 – 前一天收盤高美金價) / 前一天收盤美金價

Demo

In [10]:
```
import findspark
findspark.init('/home/joadmin/spark-2.1.2-bin-hadoop2.7/')
import pyspark
from pyspark.sql import SparkSession
from pyspark.sql.functions import *
spark = SparkSession.builder.appName('BTC').config("spark.jars.packages").getOrCreate()

# Load csv file as RDD
train_path='/home/joadmin/BTC/train.csv'
test_path='/home/joadmin/BTC/test.csv'
# # 1. Data Loading and Parsing
train_df = spark.read.load(train_path,format="csv", delimiter=",", header=True)
test_df = spark.read.load(test_path,format="csv", delimiter=",", header=True)
```

啟用Spark並匯入train及test資料

In [11]:
```
train_df.show(5)
test_df.show(5)
train_df.printSchema()
test_df.printSchema()
print (train_df.count())
print (test_df.count())
```

確認資料已存入DataFrame並檢查Schema及數量

```
+----------+-------+-----------+-----------+-----------+-----------+----------+-----------+-----------+----------
+
|nextUpDown|pUpDown|   ratiotxin|  ratiotxout|ratiotxinout|  ratiotxBTC|  ratioopen|   ratiohigh|    ratiolow| ratioclose
|
+----------+-------+-----------+-----------+-----------+-----------+----------+-----------+-----------+----------
+
|         1|      1|-0.164619165|-0.181481481| 0.766917293| 0.040888661|         0| 0.166666667|          0|0.166666667
|
|         0|      1| 0.067647059| 0.113122172| 1.264705882| 0.188267056|0.166666667| 0.285714286|-0.833333333|0.285714286
|
|         1|      0| 0.104683196| 0.130081301| 1.201550388| 0.203346836|0.285714286| 0.333333333|          6|          0
|
|         0|      1|-0.094763092| -0.14028777| 0.741935484|-0.342087427|         0| 0.083333333| 0.142857143|0.111111111
|
|         0|      0| -0.05785124|-0.052301255| 0.965217391|-0.506795192|0.111111111|-0.230769231|       0.125|       -0.1
|
+----------+-------+-----------+-----------+-----------+-----------+----------+-----------+-----------+----------
+
only showing top 5 rows
```

```
+-------+-----------+-----------+-----------+-----------+-----------+-----------+-----------+-----------+
|pUpDown|   ratiotxin|  ratiotxout|ratiotxinout|  ratiotxBTC|   ratioopen|   ratiohigh|    ratiolow|  ratioclose|
+-------+-----------+-----------+-----------+-----------+-----------+-----------+-----------+-----------+
|      0|-0.107046262| 0.047880985| 1.137963572| 0.127524541|-0.068761444|-0.030899612|-0.050110695|-0.059844092|
|      1| 0.010685731|-0.132579048| 0.802054776|-0.324149475|-0.059844092|-0.047277492|-0.014182553| 0.032453311|
|      0|-0.165551614|-0.147926568| 0.862206851|-0.261481291| 0.032453311| 0.010765863|-0.022077649|-0.046020182|
|      0|-0.062542871|-0.019425988| 1.004565717|-0.188640768|-0.046020182|-0.061987018|-0.008153567|-0.009906655|
|      1| 0.287190496| 0.264165678| 1.252209757| 0.396136073|-0.009906655| 0.059353875|  0.00879235| 0.075796488|
+-------+-----------+-----------+-----------+-----------+-----------+-----------+-----------+-----------+
only showing top 5 rows
```

```
root
 |-- nextUpDown: string (nullable = true)
 |-- pUpDown: string (nullable = true)
 |-- ratiotxin: string (nullable = true)
 |-- ratiotxout: string (nullable = true)
 |-- ratiotxinout: string (nullable = true)
 |-- ratiotxBTC: string (nullable = true)
 |-- ratioopen: string (nullable = true)
 |-- ratiohigh: string (nullable = true)
 |-- ratiolow: string (nullable = true)
 |-- ratioclose: string (nullable = true)

root
 |-- pUpDown: string (nullable = true)
 |-- ratiotxin: string (nullable = true)
 |-- ratiotxout: string (nullable = true)
 |-- ratiotxinout: string (nullable = true)
 |-- ratiotxBTC: string (nullable = true)
 |-- ratioopen: string (nullable = true)
 |-- ratiohigh: string (nullable = true)
 |-- ratiolow: string (nullable = true)
 |-- ratioclose: string (nullable = true)

2695
50
```

Demo

整理train及test資料欄位
增加Mark識別並替test資料加上
預測欄位

```
In [13]: # train_df add col. "Mark" with value train
train_df = train_df.withColumn('Mark',lit('train'))
train_df.show(3)
# test_df append col. "nextUpDown" with value '0' and  col. "Mark" with value 'test'
test_df = (test_df.withColumn('nextUpDown',lit(0)).withColumn('Mark',lit('test')))
test_df.show(3)
# declare test_df has the same header(ordered) with train_df
test_df = test_df[train_df.columns]
test_df.show(3)
```

```
+----------+-------+------------+------------+------------+-----------+-----------+-----------+-----------+-----------+-
----+
|nextUpDown|pUpDown|   ratiotxin|  ratiotxout|ratiotxinout| ratiotxBTC|  ratioopen|  ratiohigh|    ratiolow| ratioclose|
Mark|
+----------+-------+------------+------------+------------+-----------+-----------+-----------+-----------+-----------+-
----+
|         1|      1|-0.164619165|-0.181481481| 0.766917293|0.040888661|          0|0.166666667|          0|0.166666667|t
rain|
|         0|      1| 0.067647059| 0.113122172| 1.264705882|0.188267056|0.166666667|0.285714286|-0.833333333|0.285714286|t
rain|
|         1|      0| 0.104683196| 0.130081301| 1.201550388|0.203346836|0.285714286|0.333333333|          6|          0|t
rain|
+----------+-------+------------+------------+------------+-----------+-----------+-----------+-----------+-----------+-
----+
only showing top 3 rows
+-------+------------+------------+------------+------------+-----------+-----------+-----------+-----------+------+-
---+----------+
|pUpDown|   ratiotxin|  ratiotxout|ratiotxinout|  ratiotxBTC|   ratioopen|   ratiohigh|    ratiolow|  ratioclose|ACTION|M
ark|nextUpDown|
+-------+------------+------------+------------+------------+-----------+-----------+-----------+-----------+------+-
---+----------+
|      0|-0.107046262| 0.047880985| 1.137963572| 0.127524541|-0.068761444|-0.030899612|-0.050110695|-0.059844092|     0|t
est|         0|
|      1| 0.010685731|-0.132579048| 0.802054776|-0.324149475|-0.059844092|-0.047277492|-0.014182553| 0.032453311|     0|t
est|         0|
|      0|-0.165551614|-0.147926568| 0.862206851|-0.261481291| 0.032453311| 0.010765863|-0.022077649|-0.046020182|     0|t
est|         0|
+-------+------------+------------+------------+------------+-----------+-----------+-----------+-----------+------+-
---+----------+
only showing top 3 rows
+----------+-------+------------+------------+------------+------------+-----------+-----------+-----------+----------
--+----+
|nextUpDown|pUpDown|   ratiotxin|  ratiotxout|ratiotxinout|  ratiotxBTC|   ratioopen|   ratiohigh|    ratiolow|  ratioclo
se|Mark|
+----------+-------+------------+------------+------------+------------+-----------+-----------+-----------+----------
--+----+
|         0|      0|-0.107046262| 0.047880985| 1.137963572| 0.127524541|-0.068761444|-0.030899612|-0.050110695|-0.0598440
92|test|
|         0|      1| 0.010685731|-0.132579048| 0.802054776|-0.324149475|-0.059844092|-0.047277492|-0.014182553| 0.0324533
11|test|
|         0|      0|-0.165551614|-0.147926568| 0.862206851|-0.261481291| 0.032453311| 0.010765863|-0.022077649|-0.0460201
82|test|
+----------+-------+------------+------------+------------+------------+-----------+-----------+-----------+----------
--+----+
only showing top 3 rows
```

合併Train及Test以便後續使用
mllib建立訓練模型

```
In [14]: ## Append Test data to Train data
df = train_df.unionAll(test_df)
print (df.count())
df.printSchema()
```

```
2745
root
 |-- nextUpDown: string (nullable = true)
 |-- pUpDown: string (nullable = true)
 |-- ratiotxin: string (nullable = true)
 |-- ratiotxout: string (nullable = true)
 |-- ratiotxinout: string (nullable = true)
 |-- ratiotxBTC: string (nullable = true)
 |-- ratioopen: string (nullable = true)
 |-- ratiohigh: string (nullable = true)
 |-- ratiolow: string (nullable = true)
 |-- ratioclose: string (nullable = true)
 |-- Mark: string (nullable = false)
```

Demo

In [31]:
```
# 4. Apply Models from ML/MLLIB
# ML is built based on DataFrame, while mllib is based on RDD.
# I'm going to fit the logistic, decision tree and random forest models from ML packages.

#Logistic Regression
from pyspark.ml.classification import LogisticRegression

# regPara: lasso regularisation parameter (L1)
lr = LogisticRegression(maxIter = 100, regParam = 0.05, labelCol='index').fit(train)
print (type(lr))
# Evaluate model based on auc ROC(default for binary classification)
from pyspark.ml.evaluation import BinaryClassificationEvaluator

def testModel(model, validate = validate):
    pred = model.transform(validate)
    evaluator = BinaryClassificationEvaluator(labelCol = 'index')
    #return evaluator.evaluate(prod)
    return evaluator.evaluate(pred)

print ('AUC ROC of Logistic Regression model is: '+str(testModel(lr)))
```

```
<class 'pyspark.ml.classification.LogisticRegressionModel'>
AUC ROC of Logistic Regression model is: 0.535529715762274
```

分別套用Logistic Regression、Decision Tree及Random Forest
測試結果準確率很低...

In [35]:
```
#Decision Tree and Random Forest
from pyspark.ml.classification import DecisionTreeClassifier, RandomForestClassifier

dt = DecisionTreeClassifier(maxDepth = 3, labelCol ='index').fit(train)
rf = RandomForestClassifier(numTrees = 100, labelCol = 'index').fit(train)

models = {'LogisticRegression':lr,
          'DecistionTree':dt,
          'RandomForest':rf}

#modelPerf = {k:testModel(v) for k,v in models.iteritems()}
modelPerf = {k:testModel(v) for k,v in models.items()}

print (modelPerf)
```

```
{'RandomForest': 0.5428202288667405, 'DecistionTree': 0.5257672836576491, 'LogisticRegression': 0.535529715762274}
```

Demo

In [96]:
```
prediction=[]
for i in range(100,200):
    rf = RandomForestClassifier(numTrees=i, labelCol = 'index').fit(train)
    prediction.append(testModel(rf))
#    print(i,prediction[i-100])
import numpy as np
print(prediction.index(np.max(prediction))+100,np.max(prediction))
```

```
171 0.559787480884
```

In [103]:
```
prediction=[]
for j in range(3,8):
    dt = DecisionTreeClassifier(maxDepth = j, labelCol ='index').fit(train)
    prediction.append(testModel(dt))
#    print(i,prediction[i-100])
import numpy as np
print(prediction.index(np.max(prediction))+3,np.max(prediction))
```

```
6 0.550862205347
```

調整參數找出最佳預測率Decision Tree及Random Forest
準確率仍過低...

預計改善方向：

- 依週期(180天、360天、720天)分別建模，觀察準確率變化
- 嘗試加入其他Feature

附件 – 參考資料

Bitcoin Core

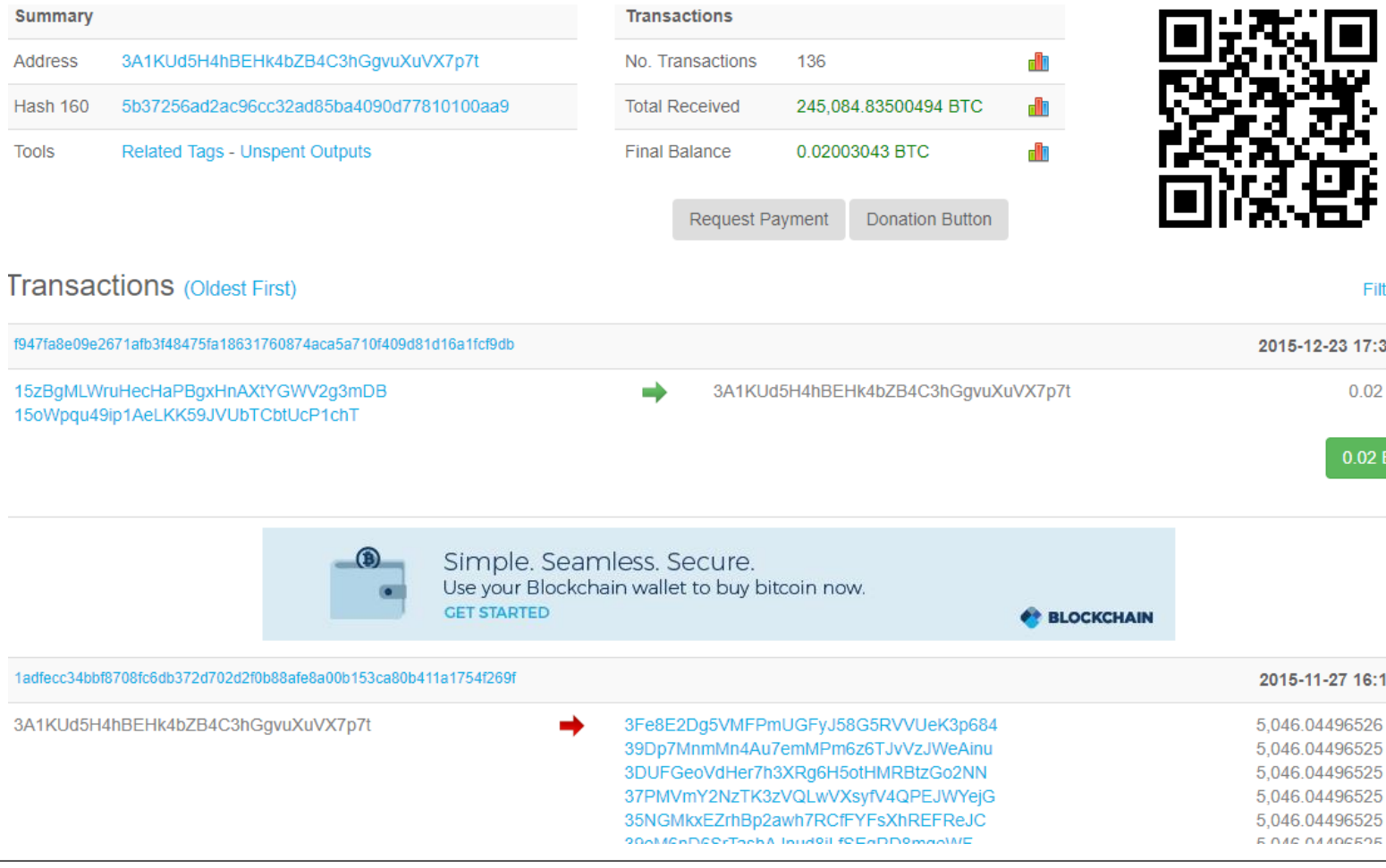
- http://blog.xuite.net/jason_kuso/kuso/549398872-Bitcoin+Core+註冊比特幣錢包+%28Bitcoin+Wallet%29+教學

Kaggle BTC101

- https://www.kaggle.com/ibadia/bitcoin-101-bitcoins-and-detailed-insights/data

Blockchain.info

- https://blockchain.info/address/3A1KUd5H4hBEHk4bZB4C3hGgvuXuVX7p7t



Summary	
Address	3A1KUd5H4hBEHk4bZB4C3hGgvuXuVX7p7t
Hash 160	5b37256ad2ac96cc32ad85ba4090d77810100aa9
Tools	Related Tags - Unspent Outputs

Transactions	
No. Transactions	136
Total Received	245,084.83500494 BTC
Final Balance	0.02003043 BTC

Request Payment | Donation Button

Transactions (Oldest First) Filter

f947fa8e09e2671afb3f48475fa18631760874aca5a710f409d81d16a1fcf9db 2015-12-23 17:30:03

15zBgMLWruHecHaPBgxHnAXtYGWV2g3mDB
15oWpqu49ip1AeLKK59JVUbTCbtUcP1chT → 3A1KUd5H4hBEHk4bZB4C3hGgvuXuVX7p7t 0.02 BTC

0.02 BTC

1adfecc34bbf8708fc6db372d702d2f0b88afe8a00b153ca80b411a1754f269f 2015-11-27 16:10:18

3A1KUd5H4hBEHk4bZB4C3hGgvuXuVX7p7t →

3Fe8E2Dg5VMFPmUGFyJ58G5RVVUeK3p684	5,046.04496526 BTC
39Dp7MnmMn4Au7emMPm6z6TJvVzJWeAinu	5,046.04496525 BTC
3DUFGeoVdHer7h3XRg6H5otHMRBtzGo2NN	5,046.04496525 BTC
37PMVmY2NzTK3zVQLwVXsyfV4QPEJWYejG	5,046.04496525 BTC
35NGMkxEZrhBp2awh7RCfFYFsXhREFReJC	5,046.04496525 BTC

Coindesk BTC Price

- https://www.coindesk.com/9-useful-bitcoin-data-resources/

Step for GCP Library in python

- https://cloud.google.com/bigquery/docs/reference/libraries#client-libraries-usage-python