```
# Step 1: Import Packages and classes
import sqlite3
import pandas as pd
import numpy as np
from sklearn.linear_model import LinearRegression
import matplotlib.pyplot as plt
from matplotlib import style
%matplotlib inline


# Step 2:
# Provide data from the tables

con = sqlite3.connect('cba_log_rv .sqlite')

# log_rv Table: log realised variance (RV) of CBA
# This is the dependent variable
rv_data = pd.read_sql_query('SELECT * FROM log_rv',con)
print(rv_data)

            date    log_rv
0     2003-01-07 -0.195388
1     2003-01-08 -0.779210
2     2003-01-09 -0.196713
3     2003-01-10  0.067592
4     2003-01-13 -0.838226
...          ...       ...
4694  2021-08-16 -0.271400
4695  2021-08-17 -0.413196
4696  2021-08-18 -0.151205
4697  2021-08-19 -0.295768
4698  2021-08-20 -0.661106

[4699 rows x 2 columns]

# log_rv feature: log RV based features
f_data = pd.read_sql_query('SELECT * FROM log_rv_feature', con)
print(f_data)

            date  log_rv_lag1  log_rv_avg5  log_rv_avg22
log_rv_avg253  \
0     2003-01-07    -0.741694    -0.449644     -0.515572      -
0.367086
1     2003-01-08    -0.195388    -0.545231     -0.521978      -
0.365305
2     2003-01-09    -0.779210    -0.463997     -0.551932      -
0.365947
3     2003-01-10    -0.196713    -0.436124     -0.567596      -
0.363599
4     2003-01-13     0.067592    -0.369083     -0.543945      -
0.359683
```

```
...          ...          ...          ...          ...             ..
.
4694  2021-08-16     0.017612    -0.159335    -0.630536      -
0.297995
4695  2021-08-17    -0.271400    -0.149591    -0.598307      -
0.299752
4696  2021-08-18    -0.413196    -0.058005    -0.583753      -
0.300944
4697  2021-08-19    -0.151205    -0.162366    -0.578148      -
0.302293
4698  2021-08-20    -0.295768    -0.222792    -0.581853      -
0.304542

      log_rv_up
0             0
1             0
2             0
3             0
4             0
...         ...
4694          0
4695          0
4696          0
4697          0
4698          0

[4699 rows x 6 columns]
```

```
# rng_feature: Range based features
rng_data = pd.read_sql_query('SELECT * FROM rng_feature', con)
print(rng_data)
```

```
            date  rng_lag1  rng_avg5  rng_avg22  rng_avg253  rng_up
0     2003-01-07  0.323813  0.607873   0.553536    0.703900       0
1     2003-01-08  0.685980  0.524683   0.544489    0.705035       0
2     2003-01-09  0.398439  0.523289   0.535820    0.704763       0
3     2003-01-10  0.795818  0.557333   0.537353    0.705806       0
4     2003-01-13  0.614600  0.563730   0.537039    0.706781       0
...          ...       ...       ...        ...         ...     ...
4694  2021-08-16  0.755281  0.755477   0.545684    0.663862       0
4695  2021-08-17  0.788250  0.744859   0.556607    0.663746       0
4696  2021-08-18  0.633679  0.782351   0.563163    0.664030       0
4697  2021-08-19  0.659298  0.684340   0.565064    0.662888       0
4698  2021-08-20  0.788937  0.725089   0.574249    0.662476       0

[4699 rows x 6 columns]
```

```
# qtl_rng_feature: Quantile range based features
qtl_data = pd.read_sql_query('SELECT * FROM qtl_rng_feature', con)
print(qtl_data)
```

```
            date  qtl_rng_lag1  qtl_rng_avg5  qtl_rng_avg22
qtl_rng_avg253  \
0     2003-01-07      0.269568      0.320090       0.304471
0.309980
1     2003-01-08      0.361912      0.321091       0.307479
0.310409
2     2003-01-09      0.271133      0.320783       0.303049
0.310296
3     2003-01-10      0.286407      0.311396       0.297146
0.310441
4     2003-01-13      0.434706      0.324745       0.304952
0.311234
...          ...           ...           ...            ...
...
4694  2021-08-16      0.341765      0.312585       0.261822
0.312503
4695  2021-08-17      0.262366      0.311847       0.265213
0.312003
4696  2021-08-18      0.314158      0.332025       0.268340
0.311909
4697  2021-08-19      0.339382      0.330446       0.268275
0.311960
4698  2021-08-20      0.281104      0.307755       0.267951
0.311475

      qtl_rng_up
0              0
1              0
2              0
3              0
4              0
...          ...
4694           0
4695           0
4696           0
4697           0
4698           0

[4699 rows x 6 columns]
```

```
# Lets join the qtl_rng_feature & log_rv_feature & rng_data

join_data0 = pd.merge(qtl_data, f_data, on='date')
print(join_data0)
```

```
            date  qtl_rng_lag1  qtl_rng_avg5  qtl_rng_avg22
qtl_rng_avg253  \
0     2003-01-07      0.269568      0.320090       0.304471
0.309980
1     2003-01-08      0.361912      0.321091       0.307479
```

```
0.310409
2     2003-01-09      0.271133      0.320783       0.303049
0.310296
3     2003-01-10      0.286407      0.311396       0.297146
0.310441
4     2003-01-13      0.434706      0.324745       0.304952
0.311234
...          ...           ...           ...            ...
...
4694  2021-08-16      0.341765      0.312585       0.261822
0.312503
4695  2021-08-17      0.262366      0.311847       0.265213
0.312003
4696  2021-08-18      0.314158      0.332025       0.268340
0.311909
4697  2021-08-19      0.339382      0.330446       0.268275
0.311960
4698  2021-08-20      0.281104      0.307755       0.267951
0.311475

      qtl_rng_up  log_rv_lag1  log_rv_avg5  log_rv_avg22
log_rv_avg253  \
0              0    -0.741694    -0.449644     -0.515572      -
0.367086
1              0    -0.195388    -0.545231     -0.521978      -
0.365305
2              0    -0.779210    -0.463997     -0.551932      -
0.365947
3              0    -0.196713    -0.436124     -0.567596      -
0.363599
4              0     0.067592    -0.369083     -0.543945      -
0.359683
...          ...           ...          ...           ...            ..
.
4694           0     0.017612    -0.159335     -0.630536      -
0.297995
4695           0    -0.271400    -0.149591     -0.598307      -
0.299752
4696           0    -0.413196    -0.058005     -0.583753      -
0.300944
4697           0    -0.151205    -0.162366     -0.578148      -
0.302293
4698           0    -0.295768    -0.222792     -0.581853      -
0.304542

      log_rv_up
0             0
1             0
2             0
3             0
```

```
4            0
...        ...
4694         0
4695         0
4696         0
4697         0
4698         0

[4699 rows x 11 columns]
```

```
# joining the join_data0 to log_rv

join_data1 = pd.merge(rv_data, join_data0, on='date')
print(join_data1)
```

```
            date    log_rv  qtl_rng_lag1  qtl_rng_avg5  qtl_rng_avg22
\
0     2003-01-07 -0.195388      0.269568      0.320090       0.304471

1     2003-01-08 -0.779210      0.361912      0.321091       0.307479

2     2003-01-09 -0.196713      0.271133      0.320783       0.303049

3     2003-01-10  0.067592      0.286407      0.311396       0.297146

4     2003-01-13 -0.838226      0.434706      0.324745       0.304952

...          ...       ...           ...           ...            ...

4694  2021-08-16 -0.271400      0.341765      0.312585       0.261822

4695  2021-08-17 -0.413196      0.262366      0.311847       0.265213

4696  2021-08-18 -0.151205      0.314158      0.332025       0.268340

4697  2021-08-19 -0.295768      0.339382      0.330446       0.268275

4698  2021-08-20 -0.661106      0.281104      0.307755       0.267951


      qtl_rng_avg253  qtl_rng_up  log_rv_lag1  log_rv_avg5
log_rv_avg22  \
0           0.309980           0    -0.741694    -0.449644     -
0.515572
1           0.310409           0    -0.195388    -0.545231     -
0.521978
2           0.310296           0    -0.779210    -0.463997     -
0.551932
3           0.310441           0    -0.196713    -0.436124     -
```

```
0.567596
4           0.311234           0     0.067592    -0.369083     -
0.543945
...              ...         ...          ...          ...           .
..
4694        0.312503           0     0.017612    -0.159335     -
0.630536
4695        0.312003           0    -0.271400    -0.149591     -
0.598307
4696        0.311909           0    -0.413196    -0.058005     -
0.583753
4697        0.311960           0    -0.151205    -0.162366     -
0.578148
4698        0.311475           0    -0.295768    -0.222792     -
0.581853

      log_rv_avg253  log_rv_up
0         -0.367086          0
1         -0.365305          0
2         -0.365947          0
3         -0.363599          0
4         -0.359683          0
...             ...        ...
4694      -0.297995          0
4695      -0.299752          0
4696      -0.300944          0
4697      -0.302293          0
4698      -0.304542          0

[4699 rows x 12 columns]
```

```
# Remove date

f = join_data1.drop(['date'], axis=1)
print(f)
```

```
        log_rv  qtl_rng_lag1  qtl_rng_avg5  qtl_rng_avg22
qtl_rng_avg253  \
0    -0.195388      0.269568      0.320090       0.304471
0.309980
1    -0.779210      0.361912      0.321091       0.307479
0.310409
2    -0.196713      0.271133      0.320783       0.303049
0.310296
3     0.067592      0.286407      0.311396       0.297146
0.310441
4    -0.838226      0.434706      0.324745       0.304952
0.311234
...        ...           ...           ...            ...
...
```

```
4694 -0.271400      0.341765      0.312585       0.261822
0.312503
4695 -0.413196      0.262366      0.311847       0.265213
0.312003
4696 -0.151205      0.314158      0.332025       0.268340
0.311909
4697 -0.295768      0.339382      0.330446       0.268275
0.311960
4698 -0.661106      0.281104      0.307755       0.267951
0.311475

      qtl_rng_up  log_rv_lag1  log_rv_avg5  log_rv_avg22
log_rv_avg253  \
0              0    -0.741694    -0.449644     -0.515572      -
0.367086
1              0    -0.195388    -0.545231     -0.521978      -
0.365305
2              0    -0.779210    -0.463997     -0.551932      -
0.365947
3              0    -0.196713    -0.436124     -0.567596      -
0.363599
4              0     0.067592    -0.369083     -0.543945      -
0.359683
...          ...          ...          ...           ...            ..
.
4694           0     0.017612    -0.159335     -0.630536      -
0.297995
4695           0    -0.271400    -0.149591     -0.598307      -
0.299752
4696           0    -0.413196    -0.058005     -0.583753      -
0.300944
4697           0    -0.151205    -0.162366     -0.578148      -
0.302293
4698           0    -0.295768    -0.222792     -0.581853      -
0.304542

      log_rv_up
0             0
1             0
2             0
3             0
4             0
...         ...
4694          0
4695          0
4696          0
4697          0
4698          0

[4699 rows x 11 columns]
```

```
# Step 3:
# Relationship between variables
# Creating models and fit


def clean_data(nodes):
    nodes = nodes.stack().reset_index()
    nodes.columns = ['variable_1','variable_2','r']
    nodes = nodes.loc[nodes['variable_1'] != nodes['variable_2'],:]
    nodes['abs_r'] = np.abs(nodes['r'])
    nodes = nodes.sort_values('abs_r', ascending = False)
    return(nodes)

nodes =
f.select_dtypes(include=['float64','int']).corr(method='pearson')
clean_data(nodes).head(10)
```

	variable_1	variable_2	r	abs_r
103	log_rv_avg253	qtl_rng_avg253	0.981233	0.981233
53	qtl_rng_avg253	log_rv_avg253	0.981233	0.981233
41	qtl_rng_avg22	log_rv_avg22	0.946594	0.946594
91	log_rv_avg22	qtl_rng_avg22	0.946594	0.946594
29	qtl_rng_avg5	log_rv_avg5	0.920140	0.920140
79	log_rv_avg5	qtl_rng_avg5	0.920140	0.920140
13	qtl_rng_lag1	qtl_rng_avg5	0.893746	0.893746
23	qtl_rng_avg5	qtl_rng_lag1	0.893746	0.893746
95	log_rv_avg22	log_rv_avg5	0.893307	0.893307
85	log_rv_avg5	log_rv_avg22	0.893307	0.893307

```
x = f[['log_rv_lag1','log_rv_avg5','log_rv_avg22','log_rv_avg253']]
y = f['log_rv']

model = LinearRegression().fit(x,y)


# Step 4:
# Get results

result = model.score(x, y)# obtaining R squared
print('coefficient of determination:', result)
print('intercept:', model.intercept_)
print('slope:', model.coef_)

coefficient of determination: 0.6061669518002699
intercept: -0.012652404429932207
slope: [0.2497617  0.4176194  0.23483427 0.05839917]

# Step 5:
# Predict Response
pred = model.intercept_ + np.sum(model0.coef_ * x, axis=1)
```

```
print('predicted response:', pred, sep='\n')
```

```
predicted response:
0      -0.528191
1      -0.433064
2      -0.552027
3      -0.398443
4      -0.298649
          ...
4694   -0.240269
4695   -0.300918
4696   -0.294737
4697   -0.271648
4698   -0.333990
Length: 4699, dtype: float64
```