Open in Colab
(https://colab.research.google.com/github/shun-lin/deep-synthetic-feature-engineering/blob/master/EDA_and_Synthetic_Feature_Engineering_Technique_1.ipynb)

In [0]:
```
#Import Packages and Datasets
```

In [0]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import featuretools as ft
```

In [0]:
```
from google.colab import files
uploaded = files.upload()
```

Choose Files No file chosen

Upload widget is only available when the cell has been executed in the current browser session. Please rerun this cell to enable.

```
Saving news_score.1csv.csv to news_score.1csv.csv
```

In [38]:
```
import io
#Fall 2018 Team 1 Data
df = pd.read_csv(io.BytesIO(uploaded['news_score.1csv.csv']))
df['ID'] = df.index
df.head()
```

Out[38]:

	author	contents	description	publisher	
0	Bitcoinist.net	real time prices vires numeris bitcoin ethereu...	israel finance ministry bank israel considerin...	Bitcoinist.com	http://bitcoinist.com/kosher-crypto·
1	Michelle Fox	var postloadfunctions var foresee enabled var ...	bitcoin may still drop rally back year early b...	CNBC	https://www.cnbc.com/2018/02/01/
2	Scott Scanlon	core cryptocurrency networks miners people use...	core cryptocurrency networks miners people use...	Youbrandinc.com	https://www.youbrandinc.com/crytocu
3	Bruce Kleinman	demons digital gold part already done please r...	demons digital gold part	Hackernoon.com	https://hackernoon.com/remediatio
4	Jason Murphy	email password remember feb systems underpinni...	systems underpinning bitcoin truly revolutiona...	Crikey.com.au	https://www.crikey.com.au/2018/02/0

5 rows × 26 columns

In [37]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7054 entries, 0 to 7053
Data columns (total 25 columns):
author                  7054 non-null object
contents                7054 non-null object
description             7054 non-null object
publisher               7054 non-null object
source_url              7054 non-null object
title                   7054 non-null object
date                    7054 non-null object
time                    7054 non-null object
Open                    7054 non-null float64
High                    7054 non-null float64
Low                     7054 non-null float64
Close                   7054 non-null float64
Volume_(BTC)            7054 non-null float64
Volume_(Currency)       7054 non-null float64
Weighted_Price          7054 non-null float64
Average                 7054 non-null float64
Volatility              7054 non-null float64
SD                      7054 non-null float64
publisherLabel          7054 non-null float64
Mark                    7054 non-null float64
publisher_L             7054 non-null float64
author_L                7054 non-null float64
score_sentiment         7054 non-null float64
magnitude_sentiment     7054 non-null float64
tfidf                   7054 non-null float64
dtypes: float64(17), object(8)
memory usage: 1.3+ MB
```

In [0]:
```
df.describe()
```

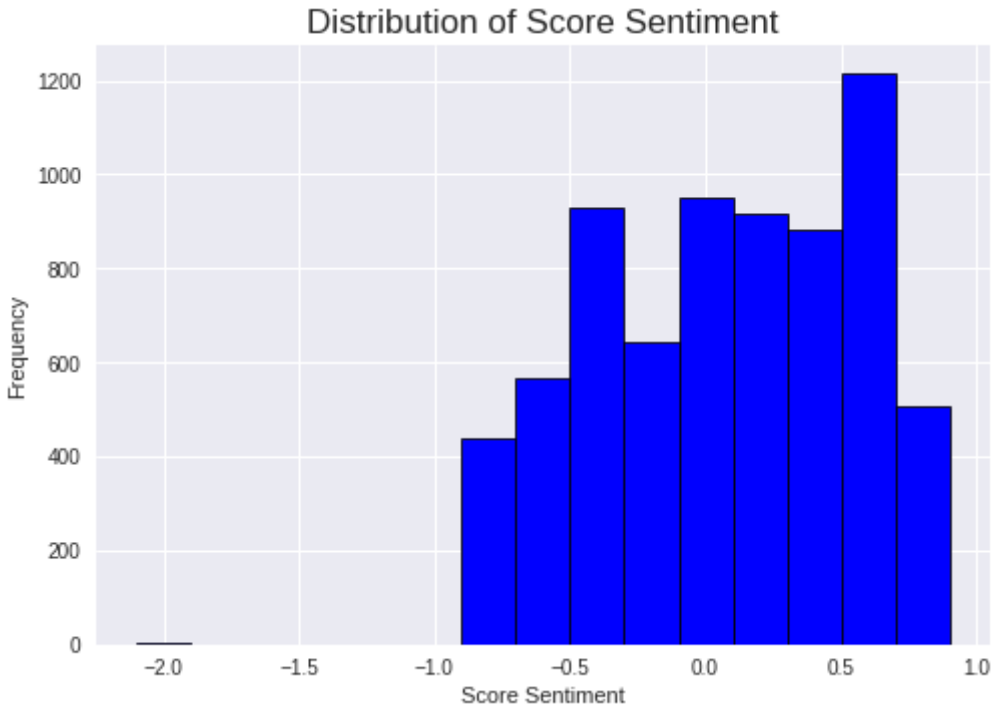
Out[14]:

	Open	High	Low	Close	Volume_(BTC)	Volume_(Currency)	Weig
count	7054.000000	7054.000000	7054.000000	7054.000000	7054.000000	7054.000000	7
mean	8182.139135	8190.604925	8173.207231	8182.071572	25.751414	199728.969814	8
std	682.239671	678.133716	686.505293	682.132607	16.902330	110347.430735	
min	6806.927451	6826.417535	6786.712438	6807.273597	9.265880	74305.946820	6
25%	7905.440896	7917.342931	7892.443951	7905.416028	13.935292	121295.129200	7
50%	8310.115639	8313.456646	8306.793611	8310.097187	18.984345	154043.483800	8
75%	8617.131229	8620.755771	8613.696326	8617.348722	35.615632	297975.902000	8
max	9435.828417	9441.771257	9429.061993	9435.448514	63.421083	432929.362400	9

In [0]:
```
#EDA
```

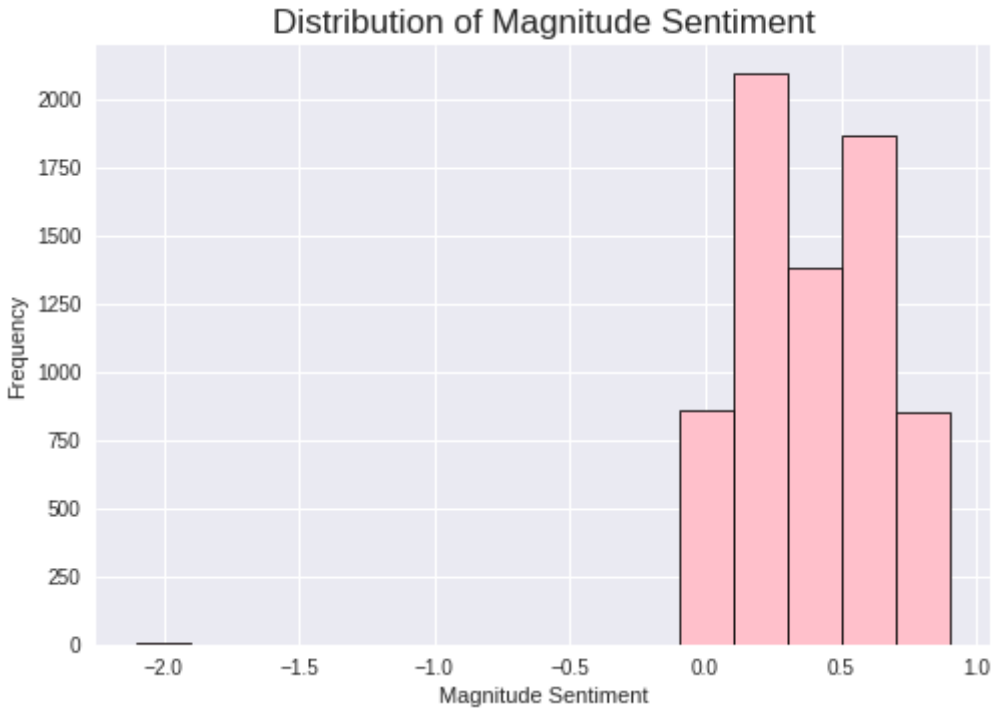
In [0]:
```
plt.hist(df['score_sentiment'], bins = np.arange(-2.1, 1, 0.2), color = 'bl
plt.title('Distribution of Score Sentiment', fontsize = 17)
plt.xlabel('Score Sentiment')
plt.ylabel('Frequency')
```

Out[10]:
```
Text(0, 0.5, 'Frequency')
```



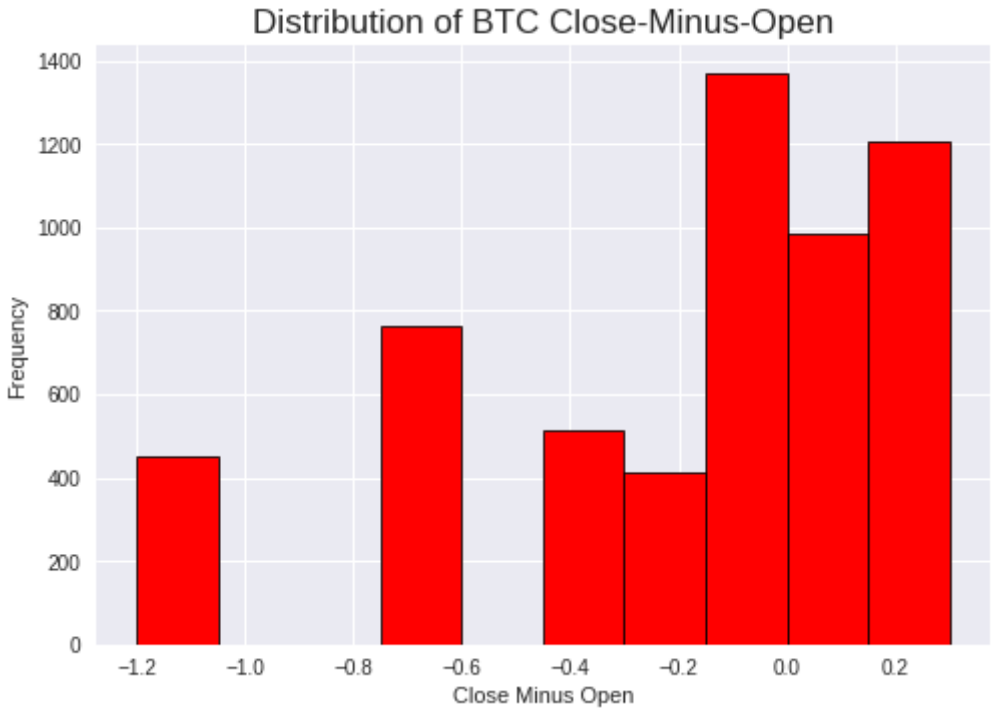
In [0]:
```
plt.hist(df['magnitude_sentiment'], bins = np.arange(-2.1, 1, 0.2), color =
plt.title('Distribution of Magnitude Sentiment', fontsize = 17)
plt.xlabel('Magnitude Sentiment')
plt.ylabel('Frequency')
```

Out[11]:
```
Text(0, 0.5, 'Frequency')
```



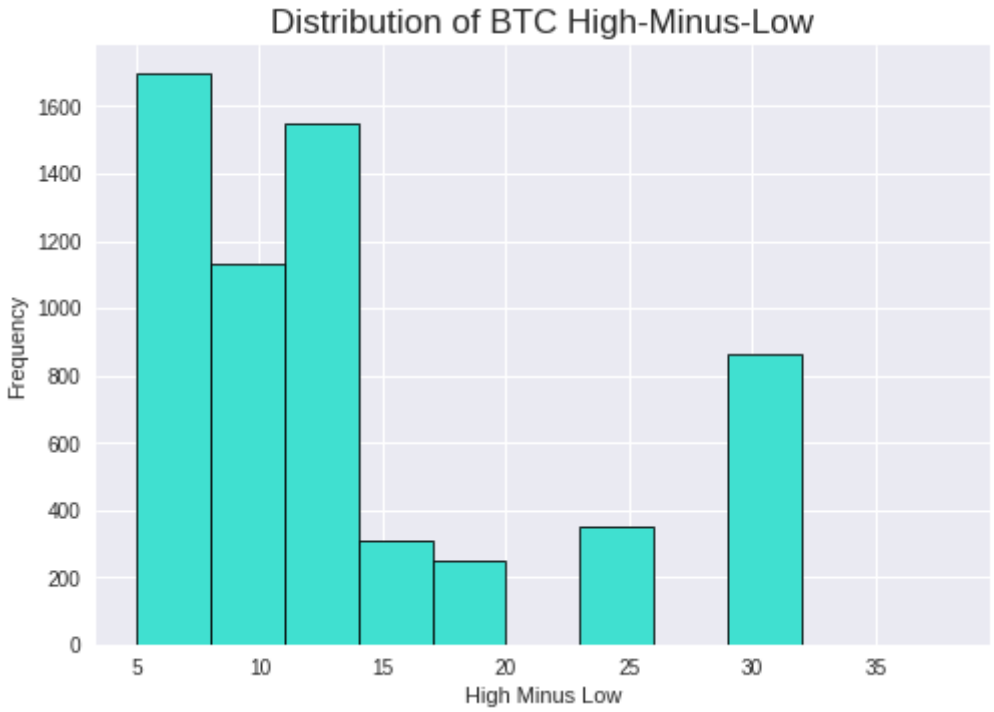
In [0]:
```
plt.hist(df['Close'] - df['Open'], bins = np.arange(-1.2, 0.4, 0.15), color
plt.title('Distribution of BTC Close-Minus-Open', fontsize = 17)
plt.xlabel('Close Minus Open')
plt.ylabel('Frequency')
```

Out[12]: Text(0, 0.5, 'Frequency')



In [0]:
```
plt.hist(df['High'] - df['Low'], bins = np.arange(5, 40, 3), color = 'turqu
plt.title('Distribution of BTC High-Minus-Low', fontsize = 17)
plt.xlabel('High Minus Low')
plt.ylabel('Frequency')
```

Out[13]:
```
Text(0, 0.5, 'Frequency')
```



In [0]:
```
#Apply DFS Techniques
```

In [0]:
```
es = ft.EntitySet(id = 'df')
```

In [52]:
```
z = dict()
for x in df['title']:
  if x in z:
    z[x] += 1
  else:
    z[x] = 1
for x in z:
  if z[x] == 2:
     print(x)
```

```
Cryptocurrency prices plunge as regulators clamp down
'The Mother Of All Bubbles And Biggest Bubble in Human History Comes Down
Crashing'
Ripple's XRP tumbles to its lowest price in months
OSC approves Canada’s first blockchain ETF
Ethereum sinks to its lowest level of 2018
Record highs, record heists: where is cryptocurrency heading?
Blockchain: an explainer
Japan's regulator urged Coincheck to fix flaws before $530 million cyber
theft
CRYPTO INSIDER: Bitcoin stages a comeback
Bitcoin’s huge arbitrage play just vanished as Korea bubble pops
Bitcoin train goes off the rails, plunging below $8,500
Cryptocurrency tumble erases over $100bn from digital currency market
Teen Arrested for Creating Malware That Steals Cryptocurrency Wallet Pass
words
Bitcoin leads cryptocurrency carnage, crashing below $8,000 for 1st time
since November
Cryptocurrency Traders Lose $115 Billion in 24 Hours as Bitcoin Bloodbath
```

In [57]:
```
df[df['title'] == "'The Mother Of All Bubbles And Biggest Bubble in Human H
```

Out[57]:
```
11     3.0
194    3.0
Name: publisher_L, dtype: float64
```

In [0]:
```
es = es.entity_from_dataframe(entity_id = 'df', dataframe = df,
                              index = 'ID', time_index = 'date')
```

In [48]:
```
df.iloc[0:10,5:15]
```

Out[48]:

	title	date	time	Open	High	Low	Close	Volume_(BT
0	Kosher Crypto BitCoen Is Setting a Course for ...	2/2/18	0:00:08	8547.864403	8562.224306	8533.223049	8547.647486	35.61563
1	Bitcoin near bottom, will rally to $20,000 thi...	2/2/18	0:02:00	8547.864403	8562.224306	8533.223049	8547.647486	35.61563
2	Brain Genius Submerges His Bitcoin Mining Rig ...	2/2/18	0:03:08	8547.864403	8562.224306	8533.223049	8547.647486	35.61563
3	Remediation, wherefore art thou?	2/2/18	0:18:34	8547.864403	8562.224306	8533.223049	8547.647486	35.61563
4	Cryptotragedy: what if bitcoin's greatest stre...	2/2/18	0:25:09	8547.864403	8562.224306	8533.223049	8547.647486	35.61563
5	Wolf Of Wall Street Says Bitcoin Could Hit $50...	2/2/18	0:49:21	8547.864403	8562.224306	8533.223049	8547.647486	35.61563
6	More bad news pushes bitcoin's value below $9000	2/2/18	1:03:11	8547.864403	8562.224306	8533.223049	8547.647486	35.61563
7	Bitcoin drops below key $9,000 level on Coinbase	2/2/18	1:40:17	8547.864403	8562.224306	8533.223049	8547.647486	35.61563
8	Cryptocurrency prices plunge as regulators cla...	2/2/18	1:42:56	8547.864403	8562.224306	8533.223049	8547.647486	35.61563
9	Bitcoin price plunges as India announces ban o...	2/2/18	2:02:58	8547.864403	8562.224306	8533.223049	8547.647486	35.61563

In [59]:
```
features, feature_names = ft.dfs(entityset=es, target_entity='df',
                                 max_depth = 2)

features.head()
```

Out[59]:

ID	author	publisher	Open	High	Low	Close	Volume_(BT
0	Bitcoinist.net	Bitcoinist.com	8547.864403	8562.224306	8533.223049	8547.647486	35.6156
1	Michelle Fox	CNBC	8547.864403	8562.224306	8533.223049	8547.647486	35.6156
2	Scott Scanlon	Youbrandinc.com	8547.864403	8562.224306	8533.223049	8547.647486	35.6156
3	Bruce Kleinman	Hackernoon.com	8547.864403	8562.224306	8533.223049	8547.647486	35.6156
4	Jason Murphy	Crikey.com.au	8547.864403	8562.224306	8533.223049	8547.647486	35.6156

5 rows × 35 columns