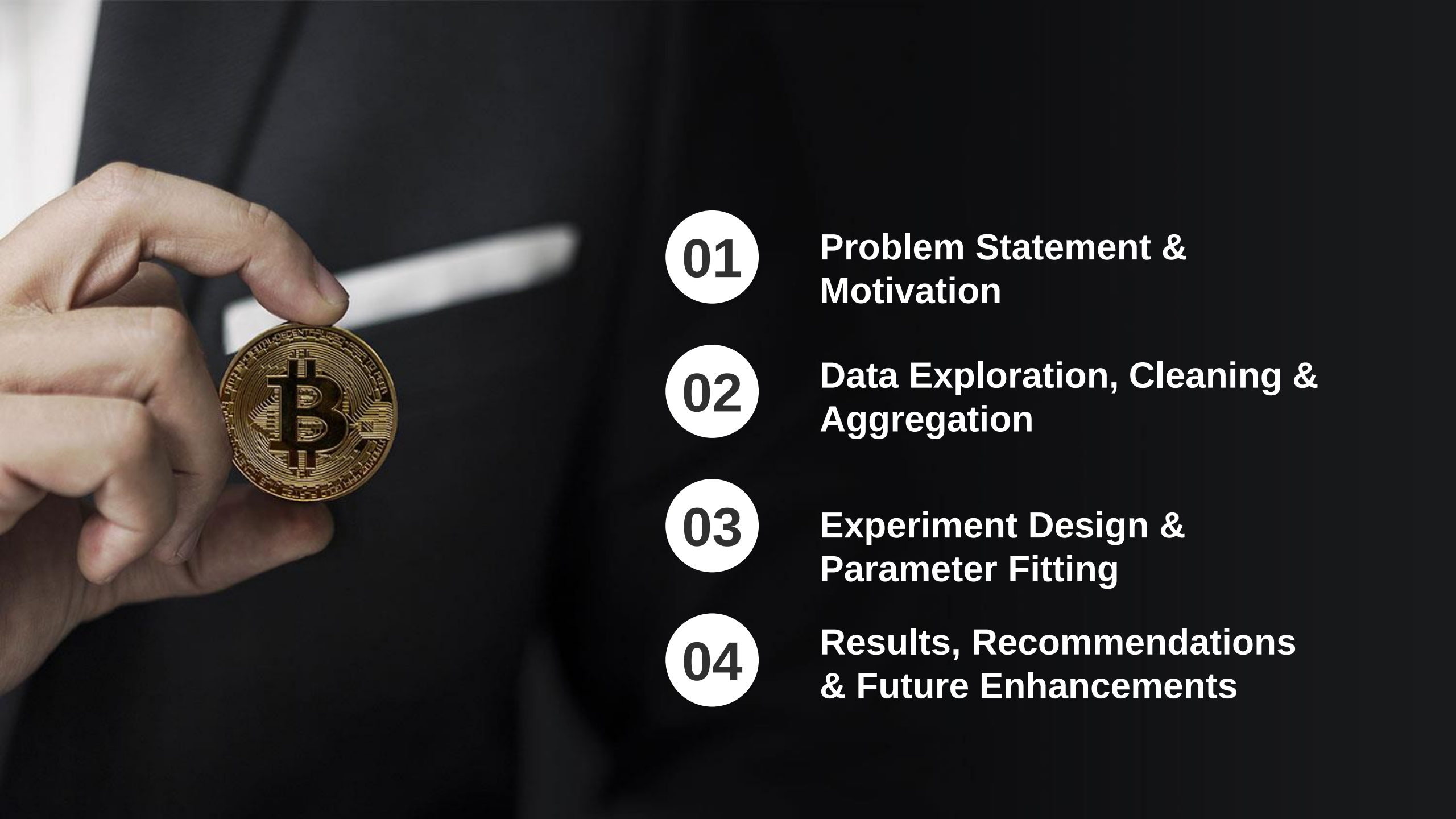
ST 502

Prediction of the direction of Bitcoin Price using “Web Search” Data

Presented By:
Group 9: Pengfei Zhao, Rohit Khurana, Haixin Wang

Under the guidance of:
Dr Ana-Maria Staicu
Stephanie Chen



01 Problem Statement & Motivation

02 Data Exploration, Cleaning & Aggregation

03 Experiment Design & Parameter Fitting

04 Results, Recommendations & Future Enhancements

Problem Statement & Motivation

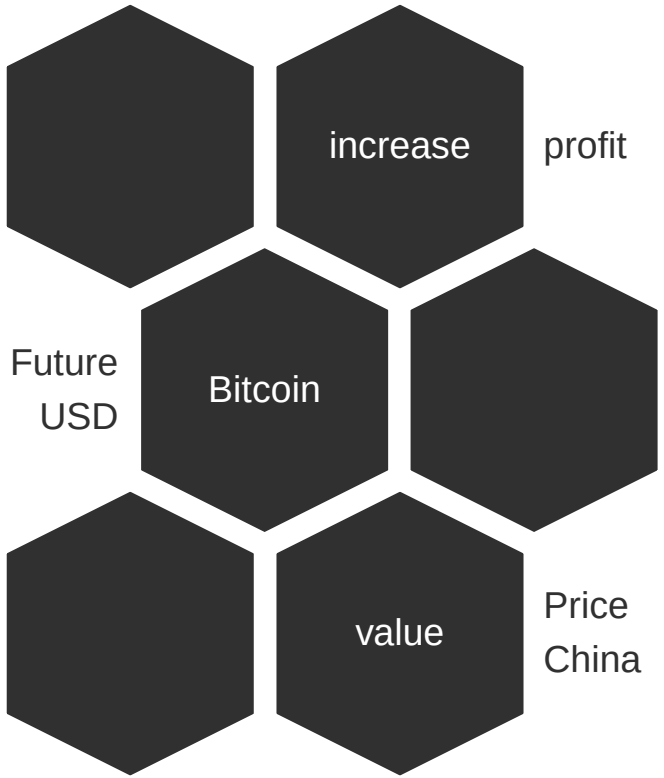
Motivation & Problem Statement

- One of Bitcoin’s uniqueness's is that its price highly relies on **people’s opinion and attitude** instead of institutionalized money regulation.

- How to predict direction for tomorrow’s bitcoin price using public sentiment & Emotions?

Data Exploration, Cleaning & Aggregation

Association Rule for Bitcoin



https://www.sas.com/content/dam/SAS/support/en/sas-global-forum-proceedings/2018/3601-2018.pdf

Data Exploration & Cleaning

- **Bitcoin Prices (daily prices)**
- *Last 5 years*
  - Date
  - Opening Price
  - Highest Price
  - Lowest Price
  - Closing Price
  - Volume
  - Market Cap

https://data.bitcoinity.org/markets/volume/30d?c=e&t=b

- **Emotions/ Web Search Data** (weekly)
- *Last 5 years*
  - Date
  - “String”
    - Increase
    - Future
    - Bitcoin
    - USD
    - Bitcoin Price
    - Normalized Data (0-100)
    - Global Data

https://trends.google.com/trends/explore?date=all&q=bitcoin,increase,future

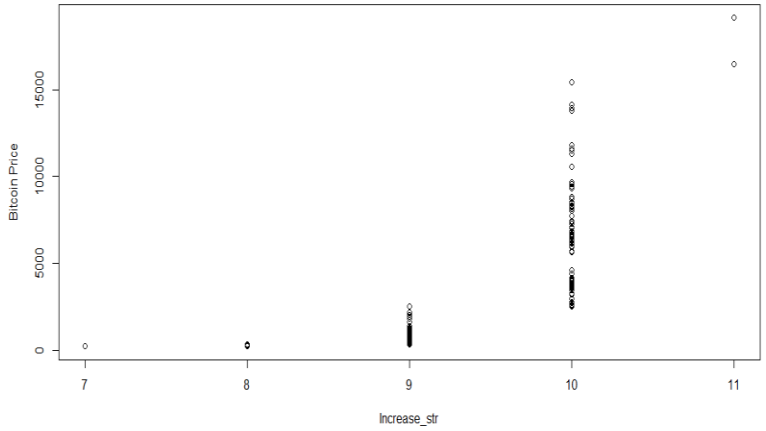
Data Aggregation

- Date Merged based upon the date
- Bitcoin Price data for the day (T) has been merged with Web Search Data of T-1.

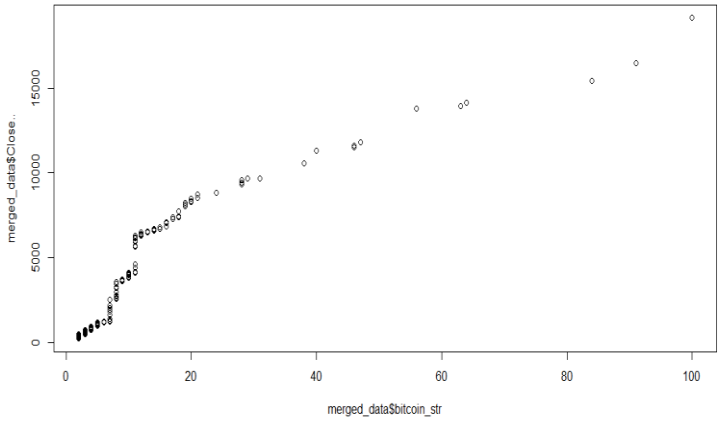
	Date	Open.	High	Low	Close..	Volume	Market.Cap	bitcoin_str	usd_str	bitcoin_price_str
1	1/1/2017	963.66	1003.08	958.70	998.33	147775008	16050407461	7	23	1.0
2	1/10/2016	448.24	448.31	440.35	447.99	35995900	6750440386	3	21	0.5
3	1/11/2015	274.61	279.64	265.04	265.66	18200800	3643307731	3	17	0.0
4	1/13/2019	3658.87	3674.76	3544.93	3552.95	4681302466	62106461671	8	26	2.0

Association of Bitcoin Price Various strings

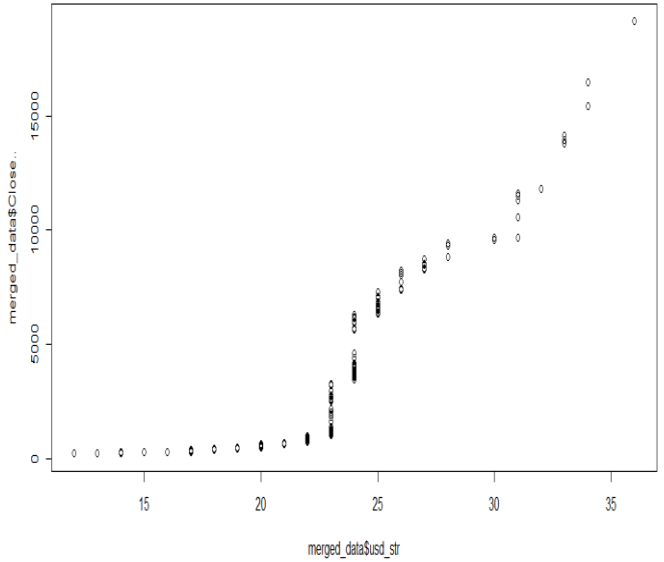
Vs Increase_Str



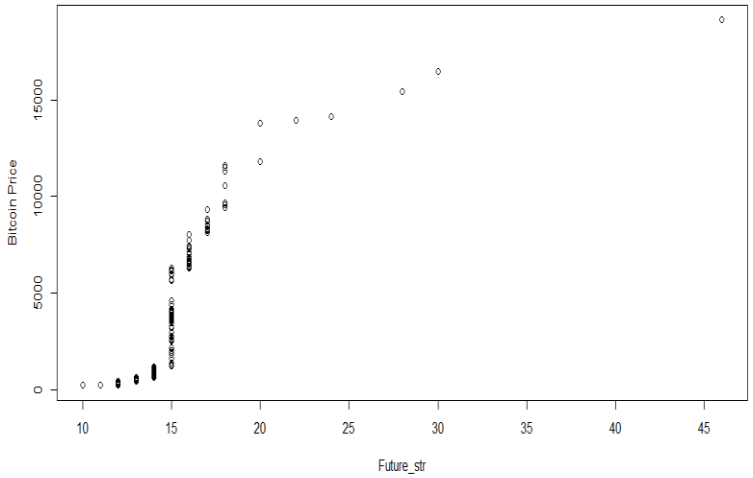
Vs Bitcoin_Str



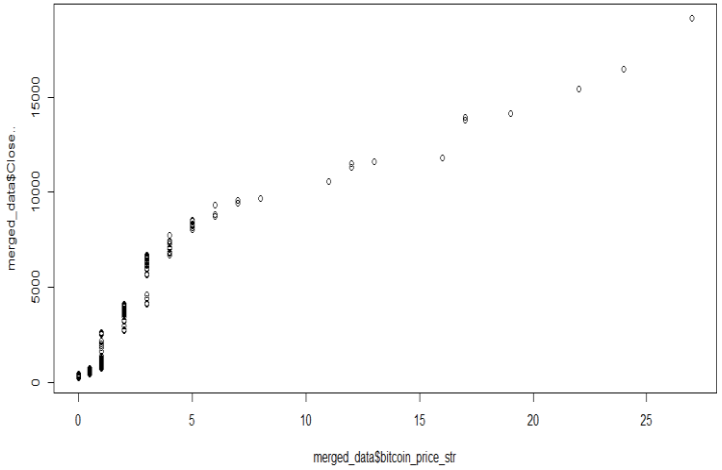
Vs USD_Str



Vs Future_Str



Vs Bitcoin_Price_Str



Merged Data

- Date column has been dropped
- **‘Direction’** Column has been added based upon the returns/ change of price

```
> head(new)
     Open.     High      Low   Close..      Volume   Market.Cap bitcoin_str usd_str bitcoin_price_str
1   963.66  1003.08   958.70    998.33   147775008  16050407461          7      23               1.0
2   448.24   448.31   440.35    447.99    35995900   6750440386          3      21               0.5
3   274.61   279.64   265.04    265.66    18200800   3643307731          3      17               0.0
4  3658.87  3674.76  3544.93   3552.95  4681302466  62106461671          8      26               2.0
5 14370.80 14511.80 13268.00 13772.00 11084099584 231413491364         56      32              17.0
6   818.14   823.31   812.87    821.80    71013600  13234840657          4      23               1.0
```

Experiment Design & Parameter Fitting

Linear Regression

- Y – Closing Price
- X1 – Bitcoin_str
- X2 - USD_str
- X3 - Bitcoin_price_str

```
> summary(lmod)

Call:
lm(formula = as.numeric(Close..) ~ bitcoin_str + usd_str + bitcoin_price_str,
    data = new)

Residuals:
    Min      1Q  Median      3Q     Max
-7003.4 -1194.7  -400.8   955.7  6344.1

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       -3667.645    678.472  -5.406 1.48e-07 ***
bitcoin_str           8.752     62.985   0.139   0.8896
usd_str             250.932     33.430   7.506 1.01e-12 ***
bitcoin_price_str   527.754    227.572   2.319   0.0212 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1815 on 256 degrees of freedom
Multiple R-squared:  0.7411,    Adjusted R-squared:  0.738
F-statistic: 244.2 on 3 and 256 DF,  p-value: < 2.2e-16
```

Logistics Regression

Y – Direction (1 – up, 0-down)

X1 – Bitcoin_str

X2 - USD_str

X3 - Bitcoin_price_str

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.10486    0.97032   2.169  0.03006 *
X1          -0.09282    0.10318  -0.900  0.36831
X2          -0.16832    0.05166  -3.258  0.00112 **
X3           1.21774    0.42725   2.850  0.00437 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 354.31  on 258  degrees of freedom
Residual deviance: 289.03  on 255  degrees of freedom
  (1 observation deleted due to missingness)
AIC: 297.03
```

Y – Direction (1 – up, 0-down)

X1 - USD_str

X2 - Bitcoin_price_str

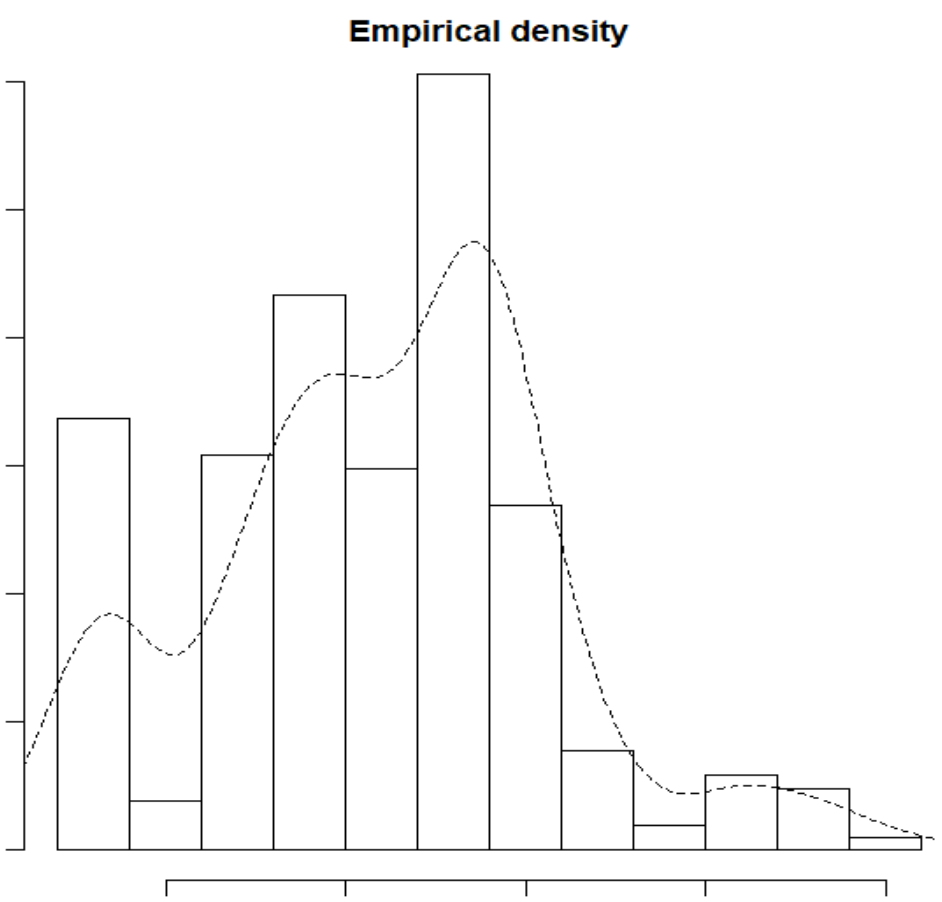
```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.79847    0.89888   2.001   0.0454 *
X1          -0.16008    0.05042  -3.175   0.0015 **
X2           0.86921    0.16366   5.311 1.09e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 354.31  on 258  degrees of freedom
Residual deviance: 289.84  on 256  degrees of freedom
  (1 observation deleted due to missingness)
AIC: 295.84
```

- *Beta0 = 1.79, Beta1 = -0.16, Beta2 = 0.87*

Distribution Fitting of Independent Variables

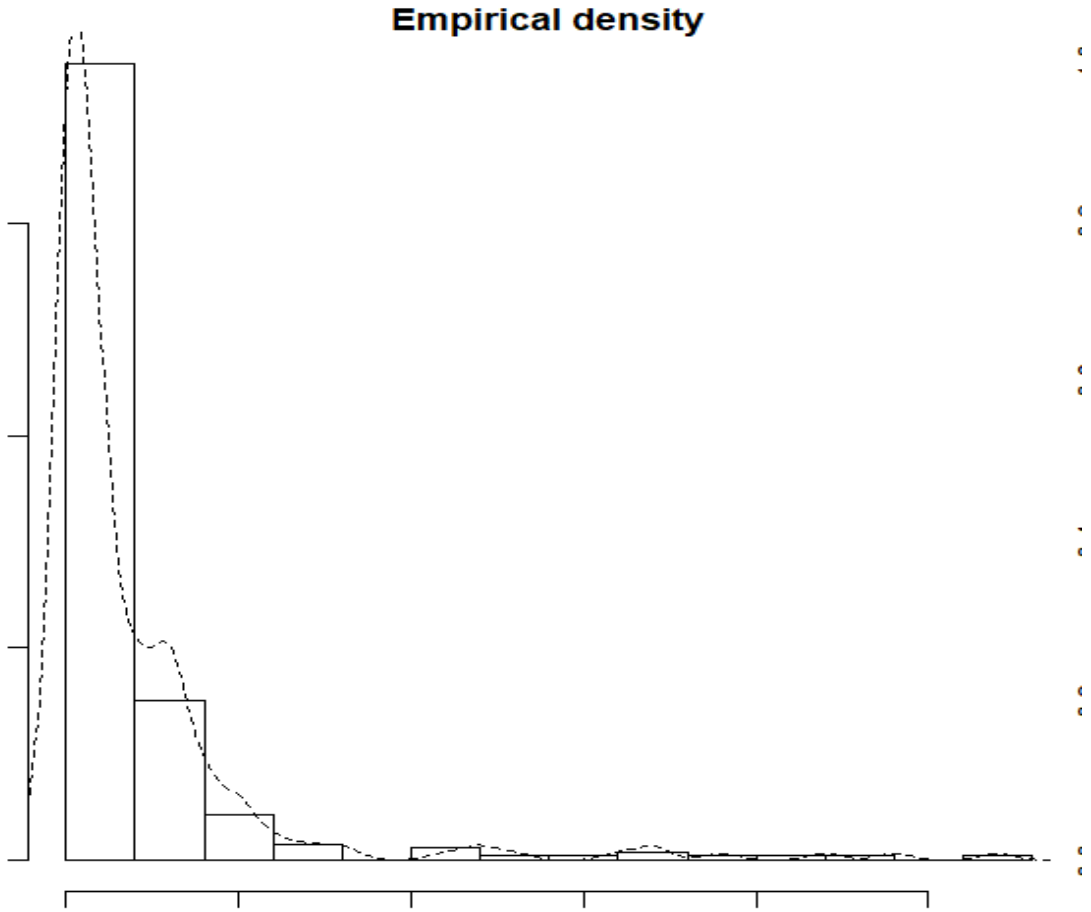
By MLE Parameters

Usd_Str – **Normal Distribution**



AD test, p-value: 0.06

Bitcoin_price_Str – **Beta Distribution**



AD Test, Beta p-value: 0.01

Results, Recommendations & Future Enhancements

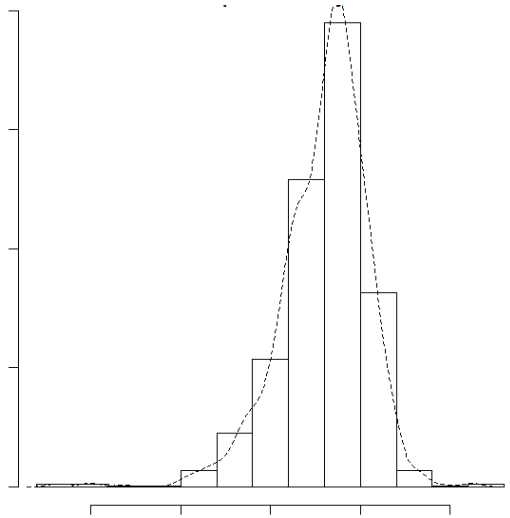
Sampling from the Existing Data (Non-Parametric BS)

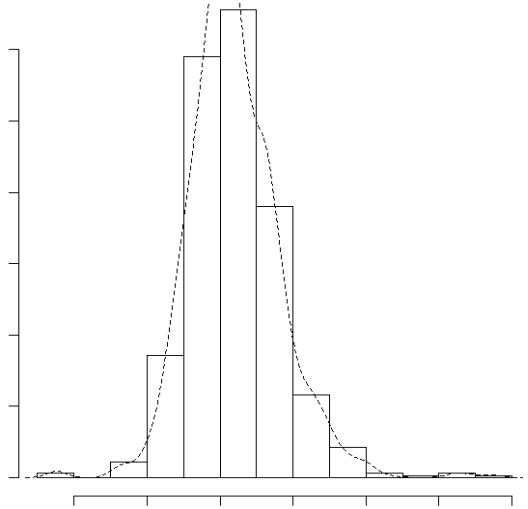
Beta0.hat from MC

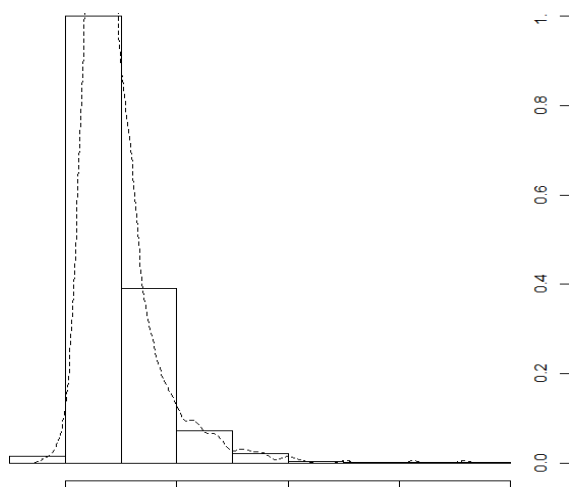
Beta1.hat from MC

Beta2.hat from MC

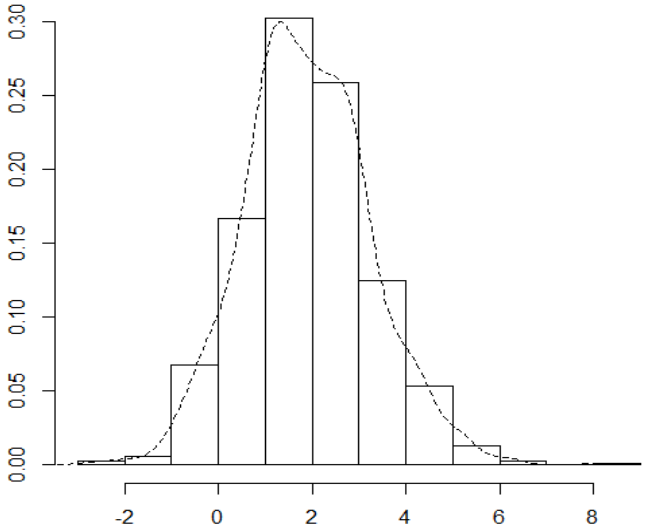
n=100

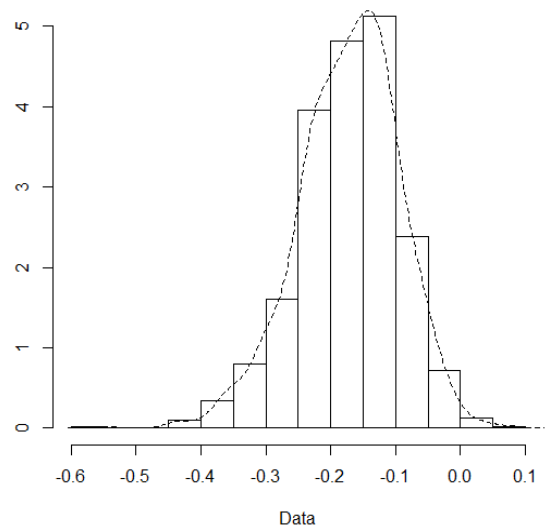


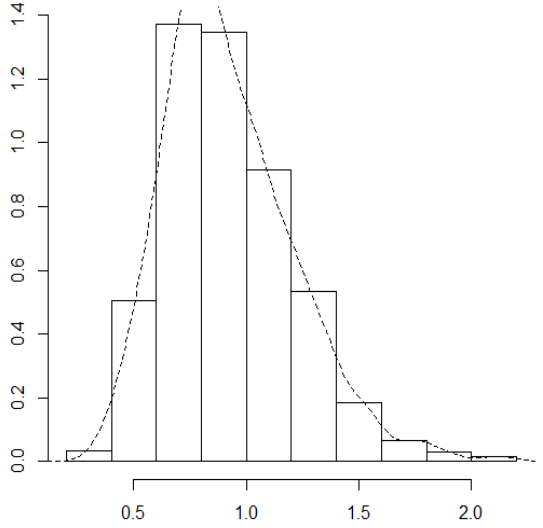




n=200

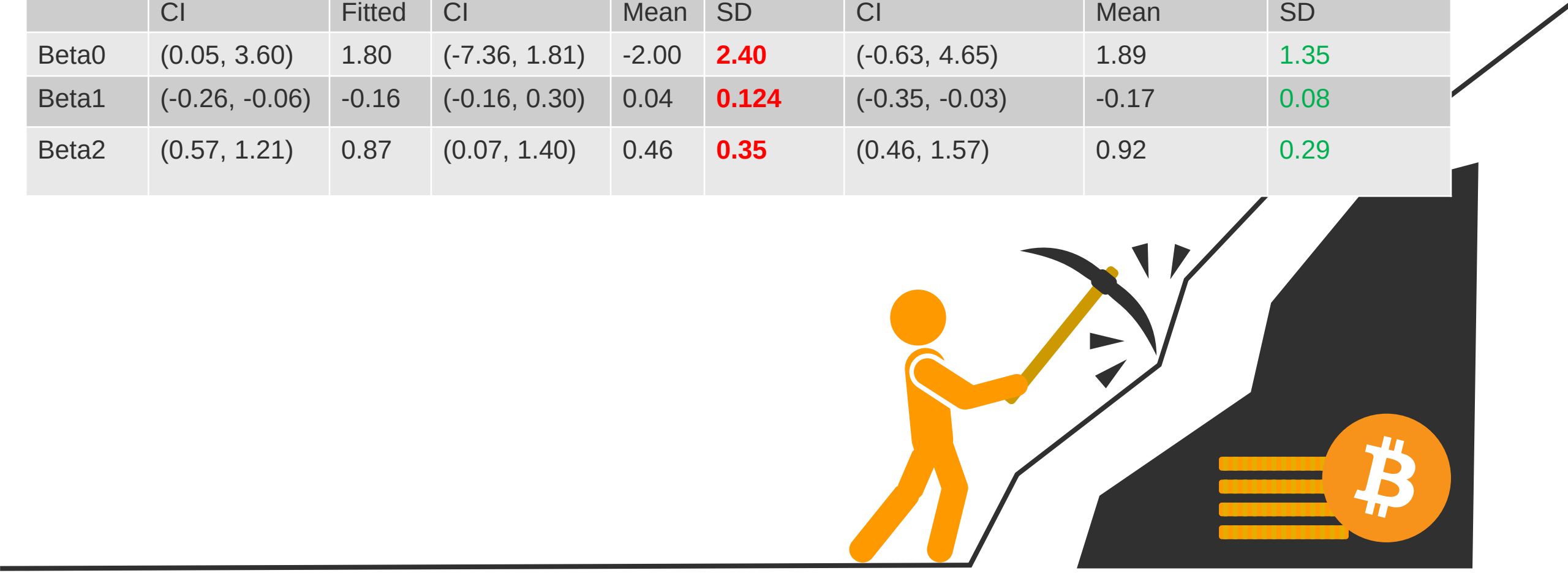






Sampling from the Existing Data (Non-Parametric BS)

Parameters	Fitted		n=100			n=200		
	CI	Fitted	CI	Mean	SD	CI	Mean	SD
Beta0	(0.05, 3.60)	1.80	(-7.36, 1.81)	-2.00	**2.40**	(-0.63, 4.65)	1.89	1.35
Beta1	(-0.26, -0.06)	-0.16	(-0.16, 0.30)	0.04	**0.124**	(-0.35, -0.03)	-0.17	0.08
Beta2	(0.57, 1.21)	0.87	(0.07, 1.40)	0.46	**0.35**	(0.46, 1.57)	0.92	0.29



Conclusion

- **Learnings**
  - Usage of Logistic Regression & Non-Parametric BS for prediction of bitcoin price movement and its dependencies on Web Search Data
  - With large sample size, the average length of 95% CI of the slopes becomes smaller and standard error decreases
- **Business Outcome**
  - In order to predict the bitcoin price, count of following strings from the Web Search Data can be use as per logistic regression model
    - USD
    - Bitcoin Price
- **Limitations**
  - Limited Data, i.e. weekly data
  - Could have back tested to ascertain the claim
  - Distribution Fitting to lognormal / transformed data might give better result
  - Association of Price to the Google trends
  - Usage of ARIMA for predicting the price

References

Data Reference

- https://cran.r-project.org/web/packages/gtrendsR/gtrendsR.pdf
- https://data.bitcoinity.org/markets/volume/30d?c=e&t=b
- https://trends.google.com/trends/explore?date=all&q=bitcoin,increase,future
- https://www.google.com/trends/correlate/search?e=bitcoin&t=weekly&p=us

Business Reference

- https://www.sas.com/content/dam/SAS/support/en/sas-global-forum-proceedings/2018/3601-2018.pdf

- https://www.reddit.com/r/CryptoCurrency/comments/aaacft/google_trends_vs_bitcoin_price_interesting/

- https://www.chepicap.com/en/news/3336/can-you-predict-price-raises-by-looking-at-search-trends-check-out-this-chart.html

- https://www.google.com/trends/correlate/whitepaper.pdf

Thank you

Question?