Does Whitepaper Page Count Affect ICO Valuation?

Michael Rosenberg and Michael McCaffrey

11/22/2018

Introduction

Within the cryptocurrency markets, there has been a recent uptick in whitepaper length. Between Q1 2016 and Q4 2018, the word count has increased from around 3000 words per paper to 9000 words per paper. Many ICO projects slated to launch in 2019 look to be continuing this trend.

We are interested in seeing if this length increase informs any ICO valuation processes. In particular, does whitepaper length predict higher ICO valuations by close date?

Our intuition suggests that whitepaper length might be an indication of ICO complexity. On one hand, this complexity might be a sign of innovative work, which would imply a higher **(FILL INFORMATION IN HERE)**

To analyze this question, we collected amount raised as close date per cryptocurrency via Coindesk's ICO Tracker. We then manually looked up each cryptocurrency's whitepaper and identified the page count on those papers. Due to the work-intensive nature of that manual process, we decided to solely analyze ICOs between January 2018 and July 2018. We will discuss the implications of this data subsetting in our next steps.

Data Exploration

Within our dataset, there are around 439 ICOs between January and July of this year. This dataset size is relatively small, which suggests that we may not be powered to see statistically significant results with a large feature set. We may be able to do a more in-depth analysis when we consider earlier years in our future work.

For our analysis, we will be predicting the amount raised in ICO (in millions) per cryptocurrency using page count. Let's take a look at our variables of interest.

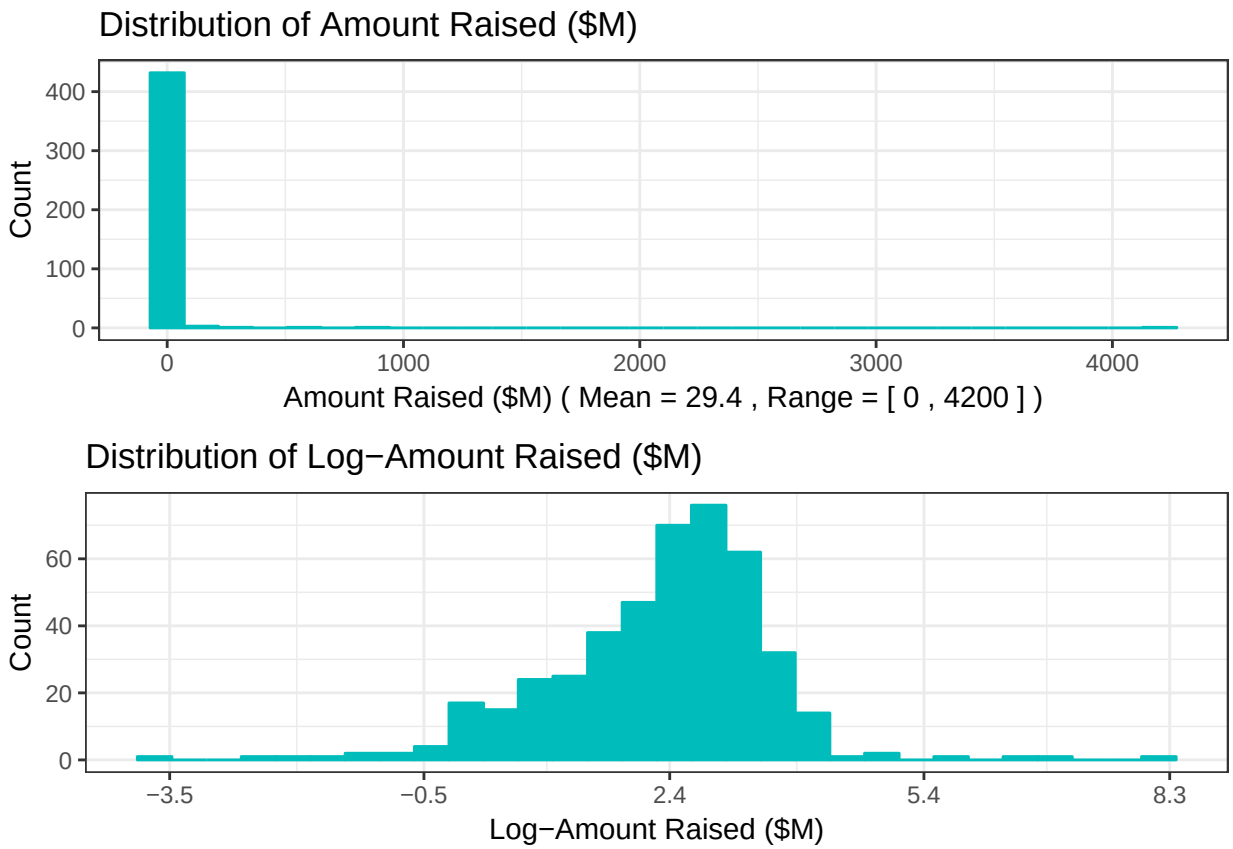

Figure 1: Distribution of Amount Raised in ICO ($M). The regular amount raised is on the top, while the log-amount raised is on the bottom.

We see that the raw amount raised is very right-skewed (top). This is typical of financial data; there are many ICOs that have raised relatively little by their close date and a handful of ICOs that have raised a huge amount of money. For reference, the median amount raised is around $12.2M while the max is around $4200M. While this is perfectly reasonable as a financial process, it is often difficult for simple predictive models to fit right-skewed variables. Because the natural logarithm of amount raised is much more normally distributed (which tends to be easier to predict with simple regression methods), we will aim to predict the log-transformed version of our amount raised in our methodology.

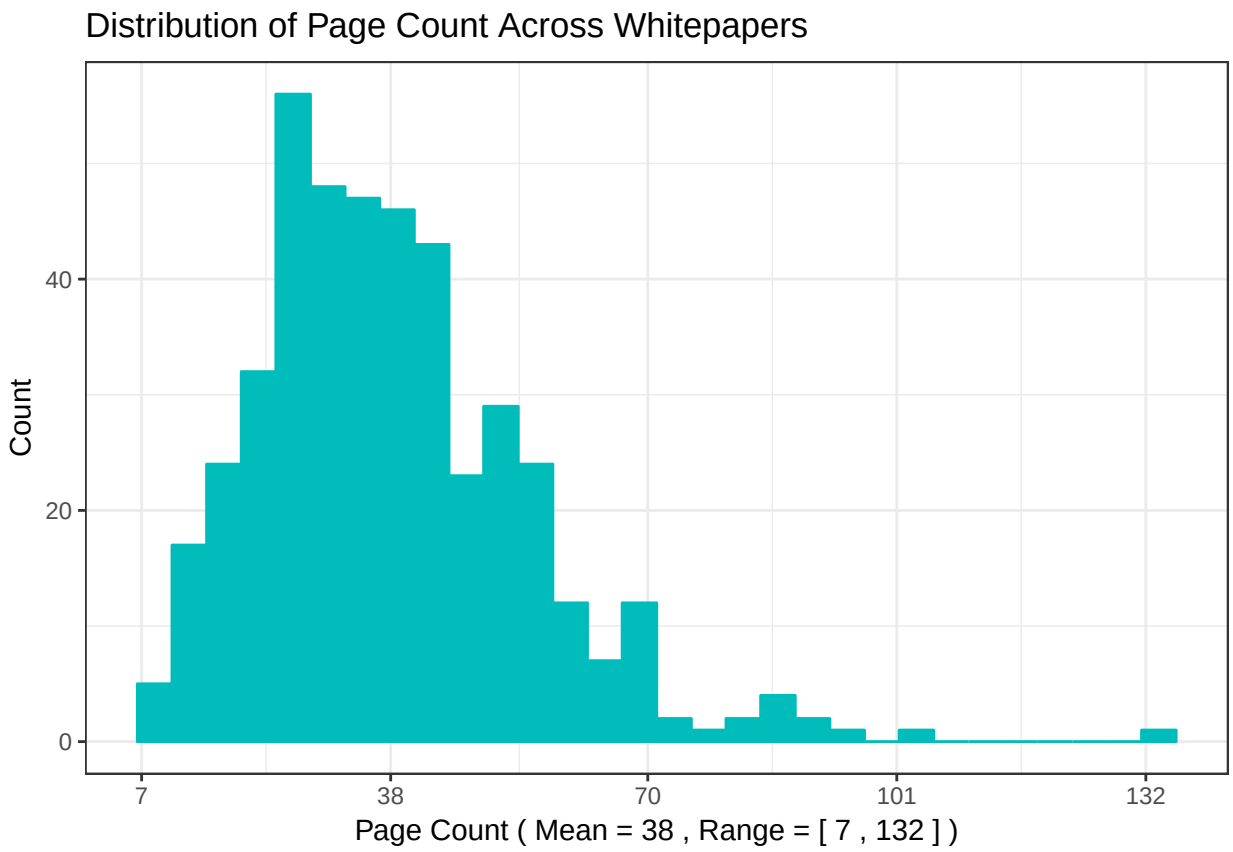

Figure 2: Distribution of Page Count per ICO whitepaper.

Like amount raised, page count is also a right-skewed variable. On average, ICO whitepapers tend to be around 38, but the longest whitepaper in our dataset is around 132 pages. Since most simple regression methods make no normality assumptions about explanatory variables, I am not too concerned about this. However, the sparsity of the page count distribution above 70 pages suggests that we may not currently be able to make statistically meaningful statements about very long whitepapers.

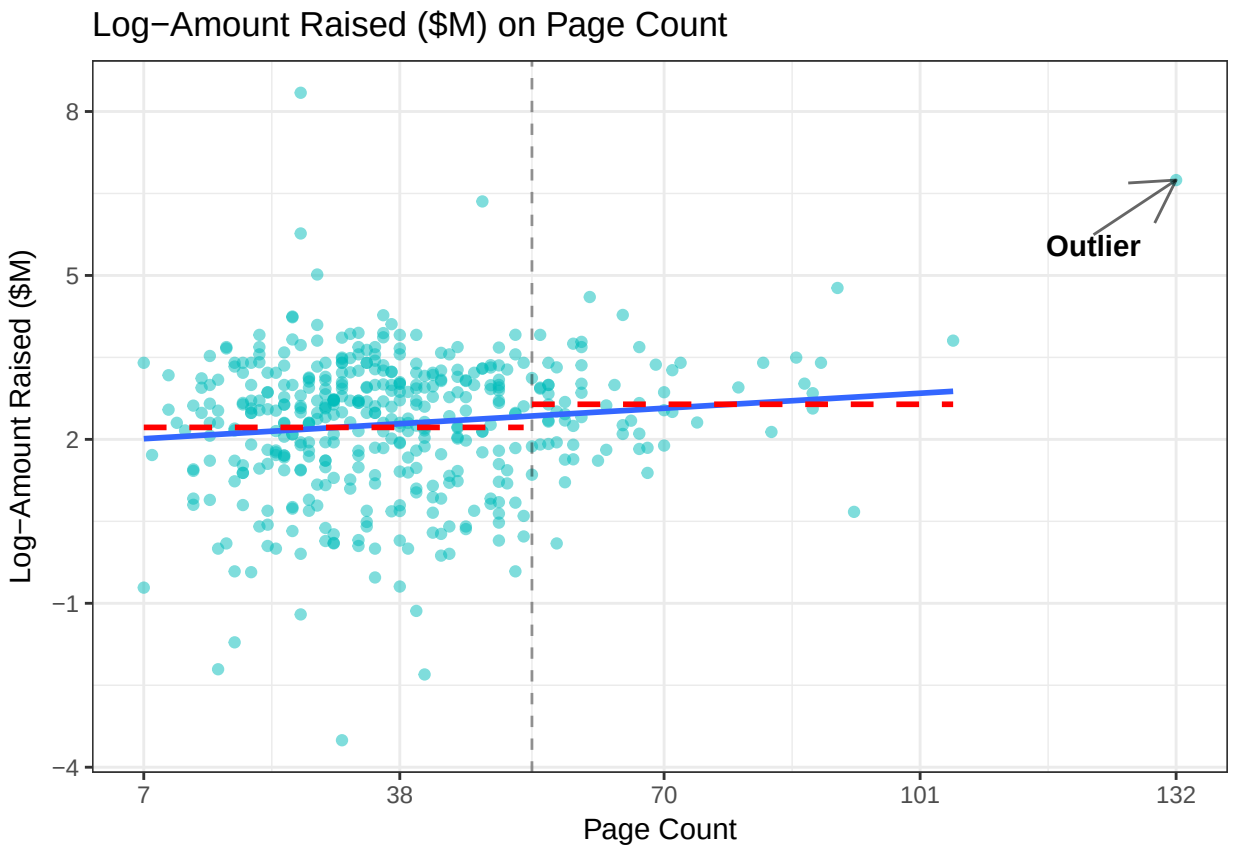

Figure 3: Log-Amount Raised ($M) on Page Count (teal). We have removed the page count outlier (Page Count = 132) from annotation within the plot. The blue line represents the linear trend for the core ICO set. The dashed black line represents the page count of 54, while the red dashed lines represent the mean log-Amount raised pre-cutoff and post-cutoff.

When plotting log-amount raised on page count, there is a very clear outlier at around 132 pages. Given that the second largest page count is only around 105 pages and the amount raised at the 132-page whitepaper is very high, I feel uncomfortable interpolating the page count effect within this gap. Thus, we are going to remove the 132-page whitepaper from our analysis.

By the linear trend (blue), we see that there is a key positive relationship between increased page counts and amount raised. However, it is also clear based on our observations (teal), the noise around the linear trend is non-constant. In particular, it looks like the variation of log-amount raised decreases after around 54 pages. This heteroscedasticity may violate some of the assumptions around statistically testing the relationship between log-amount raised and page count. Due to our timeline, we leave this robustness check outside the scope of this analysis. We will review the implications in our next steps.

When analyzing this relationship, we also noticed a clear conditional lift in log-amount raised at around the 54-page mark. If we just analyze mean amount raised pre-cutoff and post-cutoff, we measure around a 52.66% lift in amount raised. While this measurement will likely be dampened when controlling for other sources of variation (see methodology), this lift seems substantial enough to be considered as an alternative predictive hypothesis to a linear trend (blue). While the 54 number is relatively arbitrary, it corresponds with the 84th percentile of the page count distribution. Thus, we will consider a model that represents the page count effect on log-amount raised as a lift for cryptocurrencies with whitepapers in the top 16% for page count.

Methodology

Given the size of our data, we want to limit this initial analysis to using simple linear regression. Let Y_i be the amount raised (in \$M) for c ryptocurrency i. As discussed before, Y_i is very right-skewed, which is difficult to predict using simple regression methods. Based on Figure 1, we think it will be easier to generate a good fit on $\log(Y_i)$.

Given the manual process of gathering more data per cryptocurrency, we are planning to focus on the effect of page count on log-amount raised with only controls on seasonality (i.e. when the ICO closed in 2018). Based on our data exploration, we will consider 2 models of interest:

1. A linear page count effect model:

$$\log(Y_i) \sim Page\_Count_i + \sum_{t=1}^{7} I(Month\_Of\_Close_i == t). \tag{1}$$

   - $Page\_Count_i$ is the raw page count of cryptocurrency i's whitepaper.
   - $Month\_Of\_Close_i$ indicates the month of the ICO close date for cryptocurrency i. Since we are only considering ICOs between January and July of 2018, this will be an integer between 1 and 7.
   - $I(Month\_Of\_Close_i == t)$ is an indicator that equals 1 when cryptocurrency i closed in month t and 0 when said cryptocurrency closed in a month other than t. We sum 7 of these indicators for the 7 months featured in our dataset.
   - These indicator variables are meant to control for seasonality. They control for the variation in ICO amount raised from month to month, regardless of the page count of cryptocurrency i. These may be accounting for the effects of speculative hype within the cryptocurrency business, which can vary over time.
2. A percentile effect model:

$$\log(Y_i) \sim I(Percentile(Page\_Count_i) >= 84\%) + \sum_{t=1}^{7} I(Month\_Of\_Close_i == t). \tag{2}$$

   - $Percentile(Page\_Count_i) = x\%$ means $x\%$ of whitepapers have equal or fewer pages than cryptocurrency i's whitepaper. Thus, $I(Percentile(Page\_Count_i) >= 84\%) = 1$ for cryptocurrencies that have page counts in the top 16% of the page count distribution. This is essentially a binary effect that creates a lift in log-amount raised for cryptocurrencies in the top 16% for whitepaper length. This cutoff was inspired by our exploration in Figure 3.

Interpreting our regression for $\log(Y_i)$ is a bit different than for a regression on Y_i. Simply put, our effects are examined as multiplicative changes on Y_i rather than linear changes to Y_i. Say that for instance, we have fit $\log(Y_i)$ with the first model:

$$f(Page\_Count_i, Month\_of\_Close_i) = \beta_0 + \beta_1 Page\_Count_i + \sum_{t=1}^{7} \delta_t I(Month\_Of\_Close_i == t),$$

Where we predict $\widehat{\log(Y_i)} = f(Page\_Count_i, Month\_of\_Close_i)$. We can recover our amount raised prediction by exponentiating $\widehat{\log(Y_i)}$:

$$\widehat{Y_i} = e^{\widehat{\log(Y_i)}},$$

where e is the base of the natural logarithm. it's apparent then that we predict $\widehat{Y_i}$ using $e^{f(Page\_Count_i, Month\_of\_Close_i)}$.

This exponentiation makes all of the linear effects in $f(Page\_Count_i, Month\_of\_Close_i)$ turn into multiplicative effects on Y_i. For instance, say I wanted to predict the effect on amount raised when adding an additional page to a whitepaper. Our prediction for log-amount raised ($\widehat{\log(Y_i)}$) will be

$$f(Page\_Count_i + 1, Month\_Of\_Close_i) = \beta_1 + f(Page\_Count_i, Month\_Of\_Close_i).$$

Thus, our prediction for $\widehat{Y_i}$ will be

$$e^{\beta_1 + f(Page\_Count_i, Month\_Of\_Close_i)} = e^{\beta_1} \cdot e^{f(Page\_Count_i, Month\_Of\_Close_i)}.$$

Since we originally predicted amount raised with $e^{f(Page\_Count_i, Month\_Of\_Close_i)}$, we see that adding an additional page to a white paper is predicted to multiply amount raised by e^{β_1}. Thus, an additional page is predicted to increase amount raised by $(e^{\beta_1} - 1) \cdot 100\%$.

We will evaluate our models using cross-validated root mean-squared error (RMSE). RMSE is a metric that measures the average difference between our amount raised predictions ($\hat{Y}_i$) and their actual values in the dataset (Y_i). Cross-validated RMSE is designed to see how well our model performs on average for data outside of our training sample (i.e. test data). This gives us a robustness check on how well our model generalizes to outside data. We calculate cross-validated RMSE for each of our models the following way:

- Let D be our set of n datapoints ($|D| = n$). We are considering F models to evaluate ($M_1, M_2, ...M_F$).
- Randomly partition D into K equally-sized folds ($D_1, D_2, ..., D_K$). We will iterate through these folds as our test datasets.
- For each model type $f \in [F]$:
  - Set $RMSE\_Set = \emptyset$. We will store out-of-sample RMSE in this set for each fold.
  - For each fold $k \in [K]$:
    - Train model M_f using all data besides D_k ($D - D_k$).
    - Predict amount raised ($\hat{Y}_i$) using M_f on D_k.
    - Calculate (on D_k):

$$Given\_RMSE = \sqrt{\text{AVG}((Y - \hat{Y})^2)}. \tag{3}$$

    - Add $Given\_RMSE$ to your $RMSE\_Set$ (i.e. $RMSE\_Set := RMSE\_Set \cup \{Given\_RMSE\}$).
  - Get cross-validated RMSE for model M_f via $CV\_RMSE_f = \text{AVG}(RMSE\_Set)$.

Given the small size of our dataset and standard cross-validation practices, we have chosen $K = 5$ folds for our model evaluation process. We will choose to select and analyze the model that minimizes cross-validated RMSE.

Results

We see that the cross-validated RMSE for the linear and percentile effect models are 111.84 and 111.9 respectively. While these RMSEs are very close, I will select the linear model (model 1 in the Methodology section) since its cross-validated RMSE is slightly smaller than the percentile model's cross-validated RMSE.

That being said, this RMSE is concerning from a fit perspective. The linear model implies that, on average, our model is off by around $112M for each cryptocurrency's ICO. This is pretty severe underfitting of the valuation process, and I think it is worthwhile to consider a more feature-dense model in our next steps.

	Coefficient	Std. Error	P-Value	Percent Change
(Intercept)	2.539	0.186	0.000	NA
Page Count	0.010	0.003	0.005	1.005%
Month Of Close = 2 (February)	-0.113	0.205	0.583	-10.685%
Month Of Close = 3 (March)	-0.564	0.203	0.006	-43.107%
Month Of Close = 4 (April)	-1.228	0.197	0.000	-70.712%

	Coefficient	Std. Error	P-Value	Percent Change
Month Of Close = 5 (May)	-0.607	0.202	0.003	-45.502%
Month Of Close = 6 (June)	-0.821	0.209	0.000	-56.001%
Month Of Close = 7 (July)	-1.140	0.198	0.000	-68.018%

Table 1: The coefficient table from our selected regression. "Percent Change" is the expected percent change in amount raised ($M) implied by the coefficient estimates.

We see that when we control for seasonality, increasing the length of a whitepaper by 1 page is predicted to increase amount raised by around 1%. This is also very statistically significant, with a p-value below .01. This means there is a statistically significant likelihood that is having some effect on amount raised. That being said, there are still open questions on the narrative of the effect. On one end, page count might be simply a form of obfuscation; there might not be major differences in the qualities of different cryptocurrencies, but whitepaper length might give an impression of complexity and due-diligence for an ICO that causes investors to value it higher. On the other hand, there might be genuine content differences that is informing both the length of whitepapers and there general valuation (e.g. new technological breakthroughs, ambitious designs). In this regard, it will be important to further analyze the language content of these whitepapers in our next steps.

While there is varying statistical significance of our month indicators, their negative coefficients make it clear that there is a general decline in ICO valuation post-February. It may be the case that enthusiasm around cryptocurrency has declined over the year, which could be informing lower ICO valuations post-February.

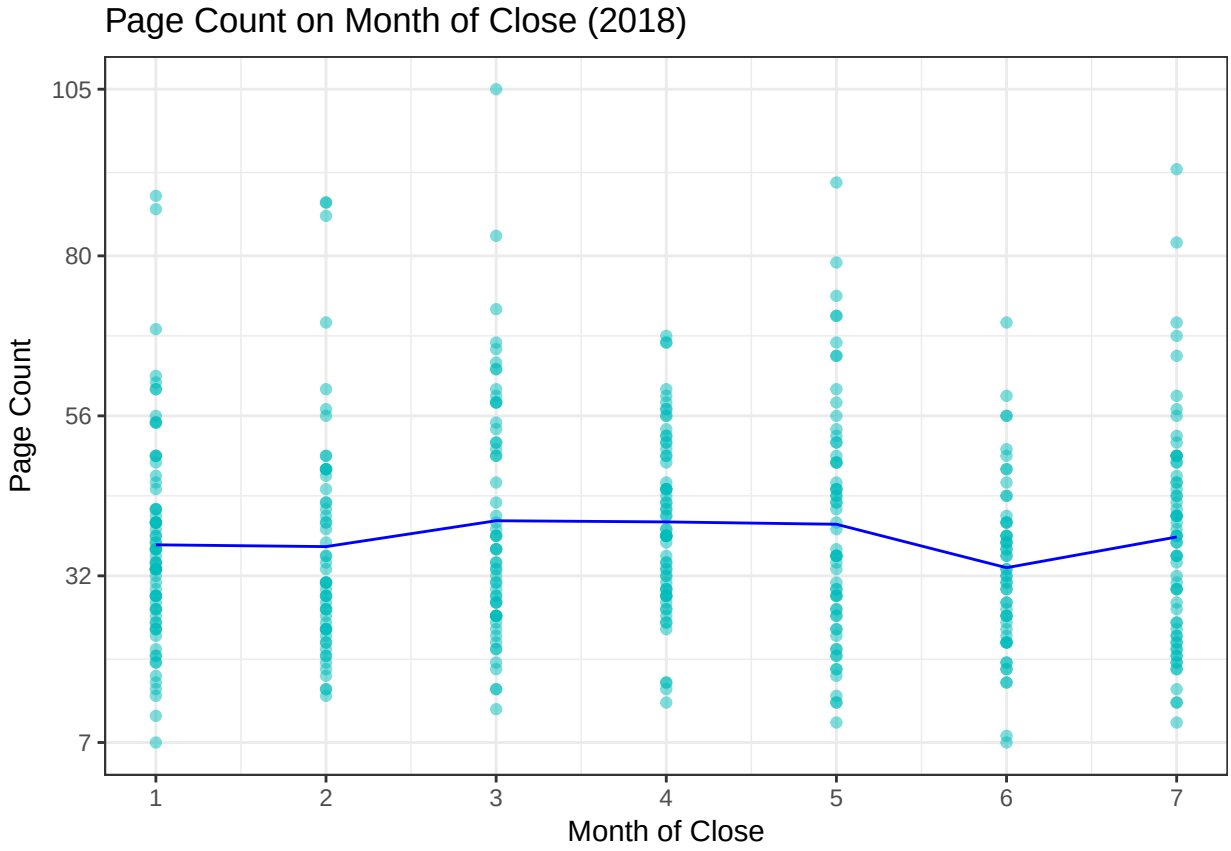

Figure 4: Page Count on Month of Close for ICOs in our modeling dataset (teal). The page count means per month of close are indicated by the blue line.

As a robustness check, we wanted to make sure there was little collinearity between page count and month

of close. If there was, it would make it difficult to interpret the page count effect on amount raised when controlling for seasonality. Thankfully, it looks like we will not have to worry substantially about this issue. Across the months of our dataset, page count hovers between 32 and 36 pages (blue line). Since this is very little variation in mean across months, we would argue that we do not need to be concerned about this multicollinearity when interpreting the effect of page count on amount raised.

Next Steps

In this analysis, we showed there is a statistically significant relationship between whitepaper page count and amount raised per ICO. In particular, our model suggests that an additional page to a whitepaper is predicted to increase ICO valuation by around 1%. This relationship has the potential to have a meaningful impact on how analysts reflect on cryptocurrency whitepapers from a very surface-level perspective. However, we have a few next steps in mind to improve the robustness of our current model and better understand the mechanisms of how whitepapers affect the valuation of ICOs.

1. As discussed in Figure 3, it is clear that the changing variance of log-amount raised with respect to page count has the potential to violate our current statistical tests. As a general next step, we may want to consider a weighted least squares regression model that can account for heteroscedasticity better than our current model.
2. The cross-validated RMSE suggests that our current model is off on average by around $112M per ICO. To me, this is severe underfitting, and it suggests that we should consider a more feature-dense approach to analyzing the valuation process. This will require us to think more deeply about the mechanisms that affect ICO valuation and collect features that will allow us to capture those mechanisms within our current modeling process.
3. If we are to consider a more feature-rich regression model, we would benefit from introducing ICOs from prior years within our dataset. Given that we only have around 438 cryptocurrencies within our final modeling dataset, we will lose statistical significance quickly if we overload features when modeling on this 2018 dataset. We will probably be able to offset increased dimension to our model if we introduce the large number of ICOs that occured in 2017. On a more secondary note, we will also be able to control for more seasonal variations when we introduces earlier timepoints within our dataset.
4. On a similar note, we may want to consider other model families besides linear models for performing this regression once we have more features. As we increase the feature set size, we may be able to automate interaction identification by considering decision trees or kernel methods.
5. Due to the use of different fonts, formats, and pictures within each whitepaper, it may be the case that page count is a noisy signal of true whitepaper length. As a robustness check, we may be interested in seeing the effect of word count on ICO valuation rather than page count.
6. From a causal narrative perspective, we are interested in spending more time analyzing what are the true mechanisms for how whitepaper length informs valuation. In particular, we are interested in using natural language processing to see if the language content of the whitepapers informs the valuation process to any degree. Since the language content is directly informing how long these whitepapers are, identifying this confound might give us a more nuanced sense of how communication on cryptocurrency affects eventual valuation at close. If the language itself is not presenting meaningful signal to amount raised, it could be the case that speculation on these cryptocurrencies is based more on perceived complexity (i.e. whitepaper length) than on communicated content.