

Does white paper page count affect \$ raised via ICO?

Michael Rosenberg and Michael McCaffrey

11/27/2018

Introduction

Within the cryptocurrency markets, there has been a recent uptick in white paper length. Between Q1 2016 and Q4 2018, the word count has increased from around 3,000 words per paper to 9,000 words per paper. Many ICO projects slated to launch in 2019 look to be continuing this trend.

We are interested in seeing if this increase in length informs the amount of money raised by projects that held an ICO. In particular, does white paper length predict a higher \$ amount raised by close date?

Our intuition suggests that white paper length might, on one hand, be an indication of project complexity. On the other hand, white paper length may be a result of additional graphics, stylistic differences and general verbosity.

To analyze this question, we collected \$ amount raised at close date per cryptocurrency via Coindesk's ICO Tracker. we then manually looked up each cryptocurrency's white paper and identified the page count on those papers. Due to the time-intensive nature of that manual process, we decided to start by analyzing ICOs between January 2018 and July 2018. We will discuss the implications of this subsetting to this timeframe in our next steps.

Data Exploration

Within our dataset, there are 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. I may be able to do a more in-depth analysis when we have considered 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.

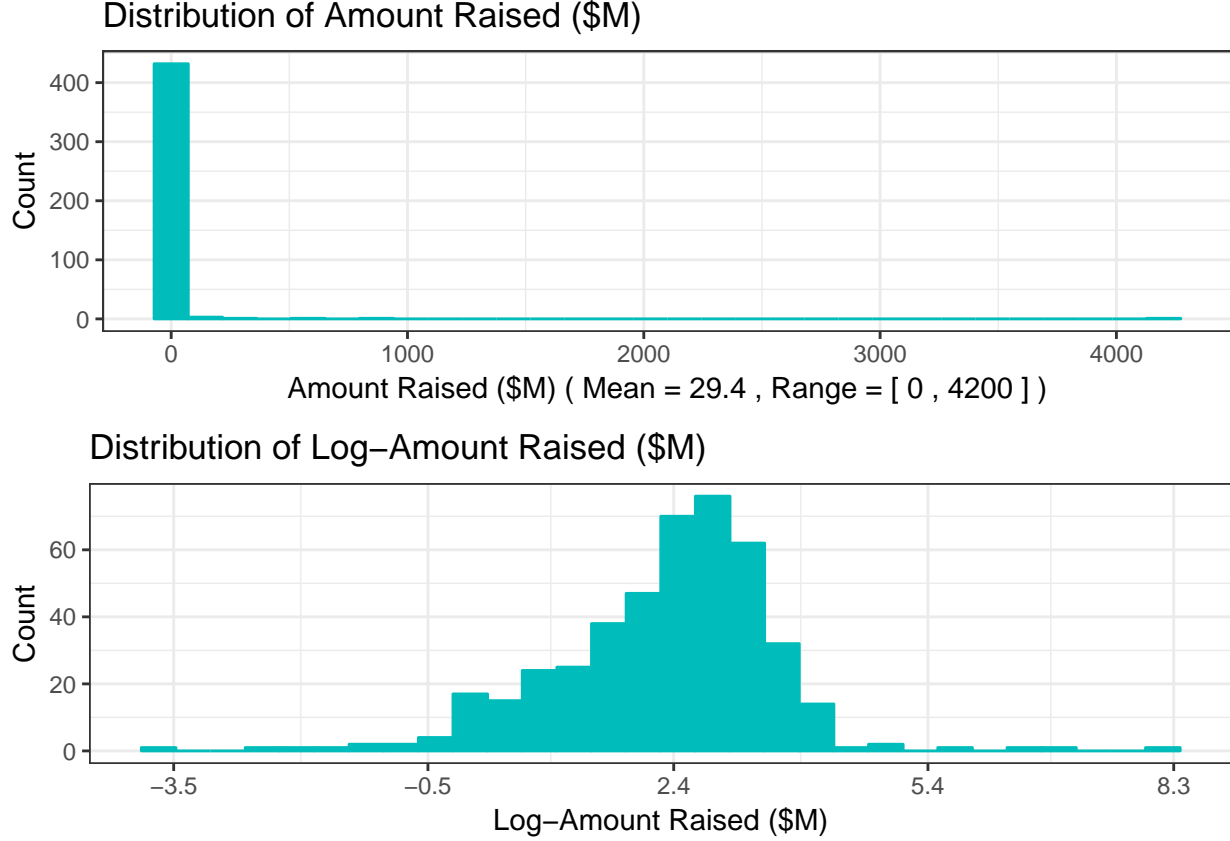


Figure 1: Distribution of Amount Raised in ICO (\$M). The regular amount raised is on the top, while the log-amount raised is on the bottom.

We see that the raw amount raised is very right-skewed (top). This is not uncommon for financial data; there are many projects whose ICOs 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 (bottom) 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 \$ amount raised in our methodology.

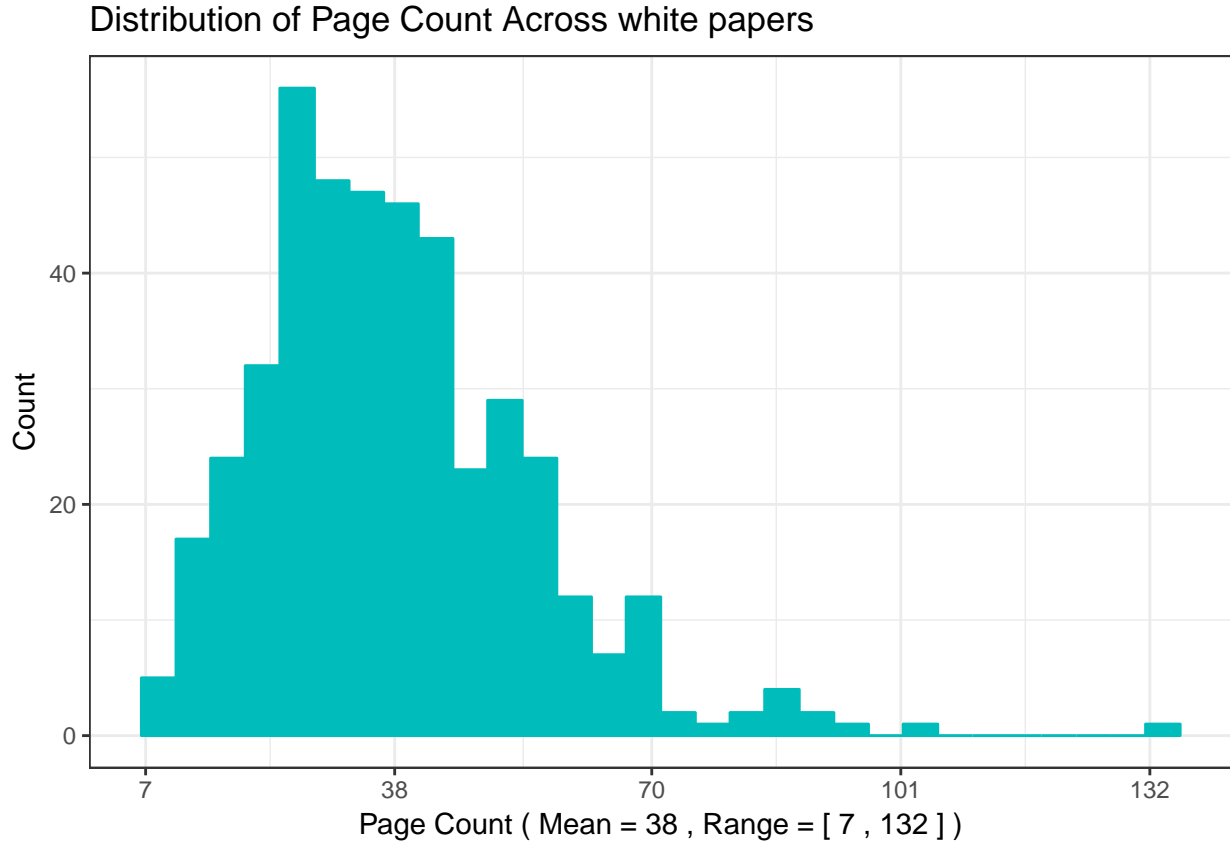


Figure 2: Distribution of Page Count per ICO white paper.

Like amount raised, page count is also a right-skewed variable. On average, white papers tend to be 38, but the longest white paper in our dataset is 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 it may be difficult with this dataset to make statistically meaningful statements about very long white papers.

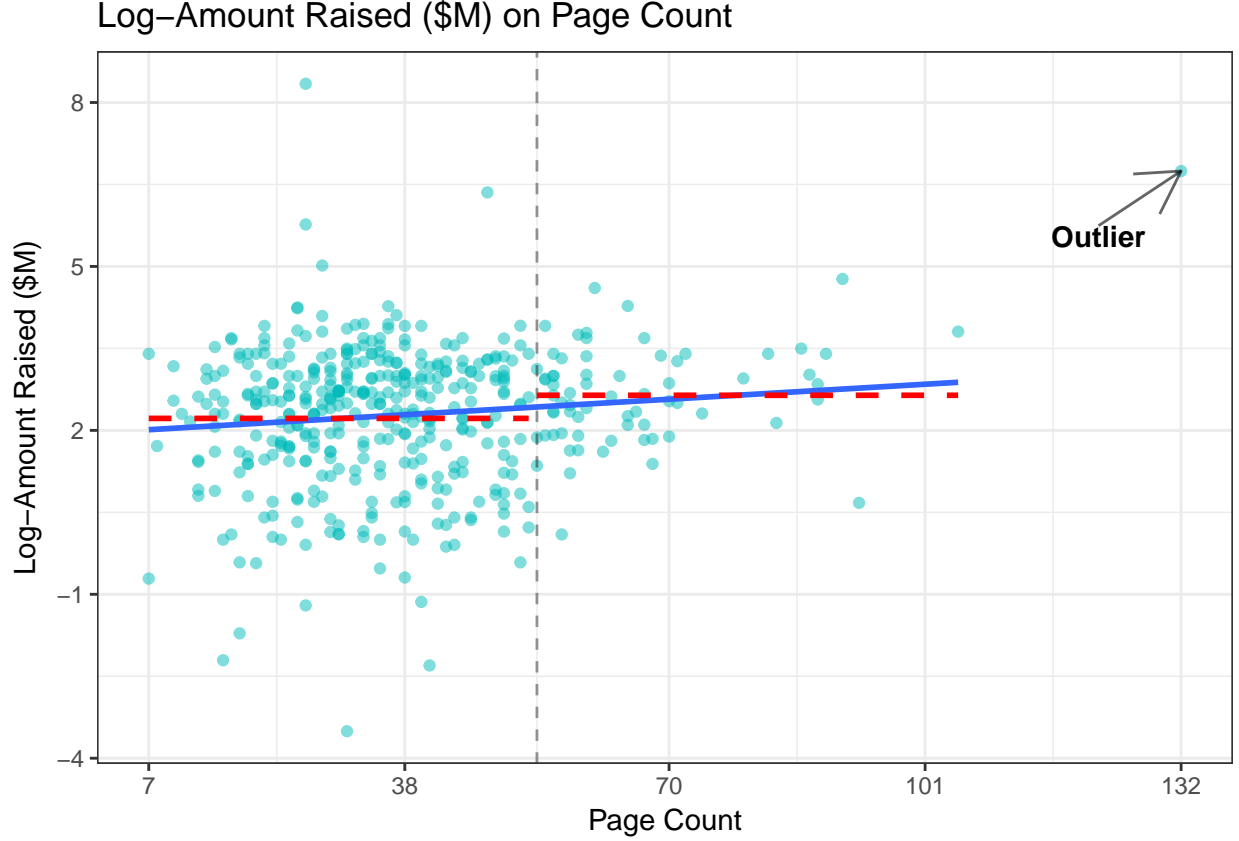


Figure 3: Log-Amount Raised in ICO (\$M) on Page Count (teal). We have removed the page count outlier (Page Count = 132) from analysis within the plot. The blue line represents the linear trend for the core ICO set. The dashed black line represents the page count of 54, while the red dashed lines represent 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 105 pages and the amount raised at the 132-page white paper is very high, we feel uncomfortable interpolating the page count effect within this gap. Thus, we are going to remove the 132-page white paper 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, our observations (teal) make it clear that 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 assumptions for statistically testing the relationship between log-amount raised and page count. For now, we leave this robustness check outside the scope of this analysis and consider it in our next steps.

When analyzing this relationship, we also noticed a clear conditional lift in log-amount raised at around the 54-page cutoff (black, dashed). If we just analyze mean amount raised pre-cutoff and post-cutoff (red, dashed), 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 ICO white papers in the top 16% for page count.

Methodology

For transparency on our analysis, we would like to give a deep dive into our methodology for predicting \$ amount raised in ICO. However, we realize that this description can be quite verbose. Thus, we will provide both a quick summary and a deep dive on our approach.

Quick Summary

Due to right-skewedness of \$ amount raised in ICO, we plan to predict log-amount raised to lighten the prediction problem for simple regression methods. We consider both a linear model and a cutoff model (see Figure 3) that control for month-based seasonality in log-amount raised while estimating a page count effect. We select the model that minimizes cross-validated root mean squared error (CV-RMSE) under a 5-fold simulation. Since we are predicting log-amount raised, we interpret our effects on \$ amount raised in ICO as multipliers rather than linear changes in money.

Deep Dive

Given the size of our data, we want to limit this initial analysis to an application of simple linear regression. Let Y_i be the amount raised in ICO (in \$M) for cryptocurrency 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). We will consider other controls in our next steps. Based on our data exploration, we will consider 2 models of interest:

1. A linear page count effect model:

$$\log(Y_i) \sim \text{Page_Count}_i + \sum_{t=1}^7 I(\text{Month_Of_Close}_i = t). \quad (1)$$

- Page_Count_i is the page count of cryptocurrency i 's white paper.
- 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.
- $I(\text{Month_Of_Close}_i = t)$ 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 account for the month-to-month variation expected in log-amount raised when we keep a given white paper's page count constant. These indicators may be representing the effects of speculative hype within the cryptocurrency business, which can vary over time.

2. A percentile effect model:

$$\log(Y_i) \sim I(\text{Percentile}(\text{Page_Count}_i) \geq 84\%) + \sum_{t=1}^7 I(\text{Month_Of_Close}_i = t). \quad (2)$$

- $\text{Percentile}(\text{Page_Count}_i) = x\%$ means $x\%$ of white papers have equal or fewer pages than cryptocurrency i 's white paper. Thus, $I(\text{Percentile}(\text{Page_Count}_i) \geq 84\%) = 1$ for white papers that have page counts in the top 16% of the page count distribution. This will generate a binary effect that creates a lift in log-amount raised for white papers in the top 16% for page count. This cutoff was inspired by our exploration in Figure 3.

Interpreting our regression for $\log(Y_i)$ is somewhat different when compared to a regression on Y_i . For a $\log(Y_i)$ regression, effects are represented 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:

$$f(\text{Page_Count}_i, \text{Month_of_Close}_i) = \beta_0 + \beta_1 \text{Page_Count}_i + \sum_{t=1}^7 \delta_t I(\text{Month_Of_Close}_i = t),$$

Where we predict $\widehat{\log(Y_i)} = f(\text{Page_Count}_i, \text{Month_of_Close}_i)$. 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. Thus, we can predict \widehat{Y}_i using $e^{f(\text{Page_Count}_i, \text{Month_of_Close}_i)}$.

This exponentiation makes all of the linear effects in $f(\text{Page_Count}_i, \text{Month_of_Close}_i)$ turn into multiplicative effects on Y_i . For instance, say I wanted to predict the effect on \$ amount raised in ICO when adding an additional page to a cryptocurrency's white paper. Our prediction for $\widehat{\log(Y_i)}$ will be

$$f(\text{Page_Count}_i + 1, \text{Month_Of_Close}_i) = \beta_1 + f(\text{Page_Count}_i, \text{Month_Of_Close}_i).$$

Thus, our prediction for \widehat{Y}_i will be

$$e^{\beta_1 + f(\text{Page_Count}_i, \text{Month_Of_Close}_i)} = e^{\beta_1} \cdot e^{f(\text{Page_Count}_i, \text{Month_Of_Close}_i)}.$$

Since we originally predicted amount raised with $e^{f(\text{Page_Count}_i, \text{Month_Of_Close}_i)}$, 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 in ICO 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 (\widehat{Y}_i) and their actual values in the dataset (Y_i). Cross-validated RMSE (CV-RMSE) indicates how well our model performs on average for data outside of our training sample (i.e. test data). This gives us a robustness check for how well our model generalizes to outside data. We calculate CV-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, \dots, M_F).
- Randomly partition D into K equally-sized folds (D_1, D_2, \dots, 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 (i.e. $D - D_k$).
 - * Predict amount raised (\widehat{Y}_i) using M_f on D_k .
 - * Calculate (on D_k):

$$\text{Given_RMSE} = \sqrt{\text{AVG}((Y - \widehat{Y})^2)}. \quad (3)$$

- * Add Given_RMSE to your $RMSE_Set$ (i.e. $RMSE_Set := RMSE_Set \cup \{\text{Given_RMSE}\}$).
- Get CV-RMSE for model M_f via $\text{CV-RMSE}_f = \text{AVG}(RMSE_Set)$.

Given the small size of our dataset and the 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 CV-RMSE.

Results

We see that the CV-RMSE for the linear and percentile effect models are 111.84 and 111.9 respectively. While these CV-RMSEs are very close, we will select the linear model (model 1 in the Methodology section) since its CV-RMSE is slightly smaller than the percentile model’s CV-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 \$ amount raised in ICO. This is severe underfitting of the fundraising process, and we 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%
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 in iCO (\$M) implied by the coefficient estimates.

We see that when we control for seasonality, increasing the length of a white paper 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 that there is a statistically significant chance that page count 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 among different cryptocurrencies, but white paper length might give an impression of complexity and due-diligence for an ICO that causes investors to provide more fundraising. On the other hand, there might be genuine content differences that are informing both the length of white papers and the general amount raised in ICO (e.g. new technological breakthroughs, ambitious designs). In this regard, we should further analyze the language content of these white papers in our next steps.

While there is varying statistical significance among our month indicators, their negative coefficients make it clear that there is a general decline in \$ amount raised via ICO post-February 2018. It may be the case that enthusiasm around cryptocurrency has declined over the year, which could be informing lower \$ amounts raised via ICO post-February.

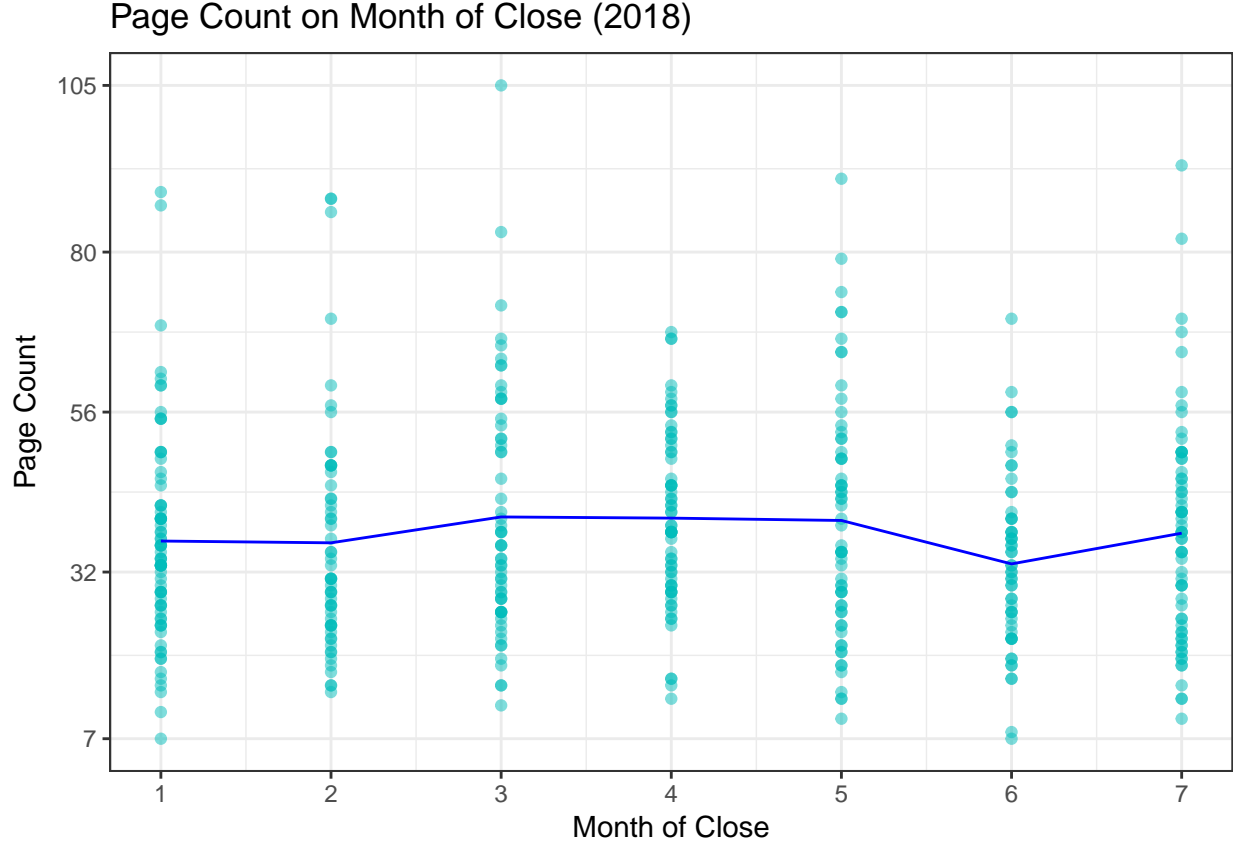


Figure 4: Page Count on Month of Close for ICOs in our modeling dataset (teal). The page count means per month of close are indicated by the blue line.

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 collinearity is not an issue in this context. Across the months in 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 multicollinearity when interpreting the effect of page count on \$ amount raised.

Next Steps

In this analysis, we identified a statistically significant relationship between a cryptocurrency’s white paper page count and \$ amount raised in ICO. In particular, our model suggests that an additional page to a white paper is predicted to increase \$ amount raised in ICO by around 1%. This relationship could potentially impact the way that analysts reflect on cryptocurrency white papers from a 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 white papers affect \$ amount raised in ICO.

1. As discussed in Figure 3, it is clear that the changing variance of log-amount raised with respect to page count could potentially violate our current statistical tests. As a followup, we should consider a weighted least squares regression model that can account for heteroscedasticity better than our current model.
2. CV-RMSE suggests that our current model is off on average by around \$112M per ICO. This is severe underfitting, and it suggests that we should consider a more feature-dense approach to predicting \$ amount raised. This will require us to think more deeply about the mechanisms that affect fundraising

- per ICO and collect features that will capture those mechanisms within our current modeling process.
3. If we want to consider a more feature-rich regression model, we would statistically benefit from introducing ICOs from prior years within our dataset. Given that we only have 438 cryptocurrencies within our final modeling dataset, we will lose statistical significance quickly if we overload features for modeling on this 2018 dataset. We can probably offset increased dimension to our model if we introduce the large number of ICOs that occurred in 2017 into our modeling dataset. On a more secondary note, we will also be able to control for more seasonal variations when we introduce earlier time points within our dataset.
 4. On a similar note, we should consider other model families besides linear models for performing this regression once we have more features. As we increase the size of our feature set, we could automate interaction identification by considering decision trees or kernel methods.
 5. Due to the use of different fonts, formats, and pictures within each white paper, it is possible that page count is a noisy measure of true white paper length. As a robustness check, we may be interested in modeling the effect of word count on \$ amount raised in ICO rather than the effect of page count.
 6. For a causal narrative, we are interested in spending more time mining the true mechanisms for how white paper length informs fundraising per ICO. In particular, we are interested in using natural language processing to see if the language content of the white papers informs the \$ amount raised in ICO to any degree. Since the language content is directly informing how long these white papers are, identifying this confound will present a more nuanced narrative on how communication on cryptocurrencies affects fundraising in ICO. If 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. white paper length) than on communicated content.