Jack Wong
Daniel Fu
STAT27410 - Bayesian Final Project Proposal (Revised)

0 Cover Letter

1. Included derivation of the annual stock return variable

   In Figure 1 we derive the annual stock returns for firm *i* by calculating each weekly growth rates r_{it}

   $$r_{it} = \frac{PX\_LAST_{it}}{PX\_LAST_{i,t-1}}$$

   … and compounding over all weeks in the year.

   $$a_i = \prod_{t \in year} r_{it} - 1$$

2. Explained that outperformance rate denominator is not exactly 10 (only consider up to 9 full calendar years worth of data), but switched to a 0.1 bin width for ease of interpretation
3. Clarified volatility: We measure volatility using the standard deviation of weekly returns because the underlying price data are weekly and weekly returns provide many observations per firm, yielding a more stable estimate of risk than annual volatility (which would be based on only ~9 data points). This measure is directly comparable to annual risk because weekly volatility can be annualized by multiplying weekly vol by square root of 52 weeks
4. Replaced scatterplots with side-by-side boxplots categorized by whether the firm-year beat the SPY, for both P/E ratio and dividend yield
5. Added boxplots similar to those outlined in (4), faceted by GICS sector
6. Condensed logistic regression models into one section and updated model specification to include time subscripts
7. Added logistic regressions fitted within each sector, and a mixed-effects logistic regression model with firm-level random intercepts and slopes, with analysis of results

1 Introduction

The primary question of interest is whether individual firms outperform the market over time. To this end, we use the SPDR S&P 500 ETF Trust (SPY) as a benchmark and define outperformance at the annual horizon. Specifically, we study (i) the probability that a firm's annual return exceeds SPY and (ii) whether differences across stocks, specifically in outperformance can be explained by firm characteristics such as valuation and income measures using price-to-earnings (P/E) ratio and dividend yield as proxies.

This distinction is hard to make in practice since returns are noisy, especially over short horizons (we only have access to 10 years worth of data, and only 9 full calendar years), so even large wins can be driven by luck rather than persistent skill. Focusing on annual returns helps smooth out some of this volatility and gives a more economically meaningful measure of performance. At the same time, valuation and dividend measures provide predictors that may help explain why some firms appear to beat the market more often than others.

Bayesian methods are especially useful here because the data are limited at the firm level: each firm only has about ten annual observations. By treating a firm's probability of beating the market as unknown and combining the data with reasonable prior beliefs, Bayesian inference allows us to express uncertainty directly and reduces the risk of over-interpreting extreme outcomes that may simply be noise.

Table 1: Data Dictionary

Variable	**Description**
ticker	Unique trading symbol used to identify the firm's equity security
name	Legal name of the firm
gics_sector_name	Global Industry Classification Standard (GICS) sector of the firm
mkt cap	Market capitalization of the firm
date	Observation date for the reported variables
PX_LAST	Last traded (closing) price of the firm's equity on the observation date
EQY_DVD_YLD_IND	Equity dividend yield (dividends per share divided by current share price) measured at date
PE_RATIO	Price-to-earnings ratio (share price divided by earnings per share) measured at observation date. For prediction we use lagged fundamentals (start of year t or end of year t-1) to avoid look ahead bias

2 Exploratory Data Analysis

In Figure 1 we derive the annual stock returns for firm *i* by calculating each weekly growth rates r_{it}

$$r_{it} = \frac{PX\_LAST_{it}}{PX\_LAST_{i,t-1}}$$

… and compounding over all weeks in the year.

$$a_i = \prod_{t \in year} r_{it} - 1$$

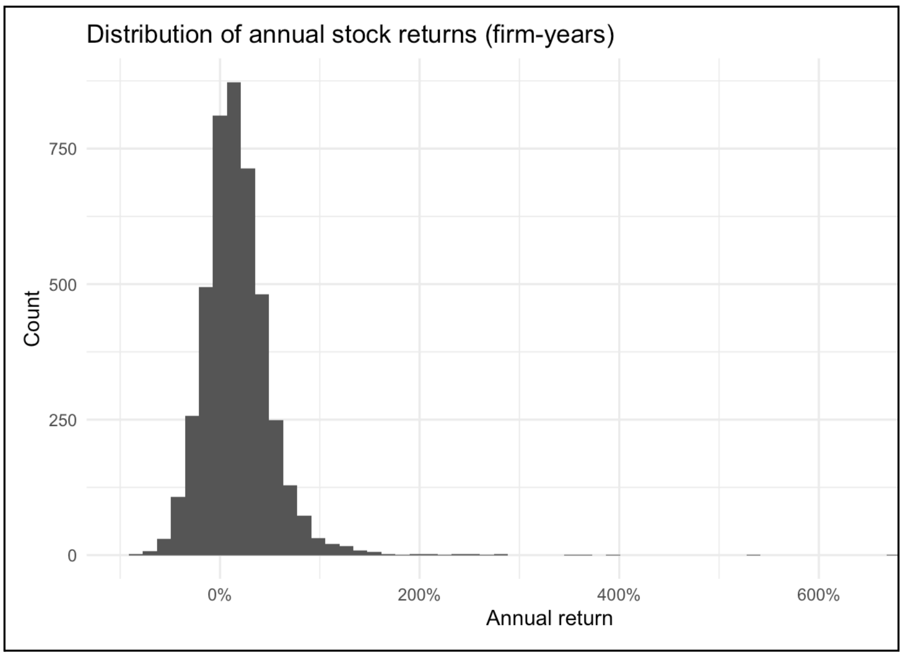

Figure 1: Distribution of Annual Stock Returns

Figure 2 displays the distribution of firms' outperformance rates relative to SPY over the sample period as a bar plot. For each firm, the outperformance rate is calculated as the fraction of years in which the firm's annual return exceeded that of SPY. We consider only full calendar years and firms with at least 8 full calendar years' worth of data. For ease of interpretation, we bin by increments of 10% to approximate one-year increments.

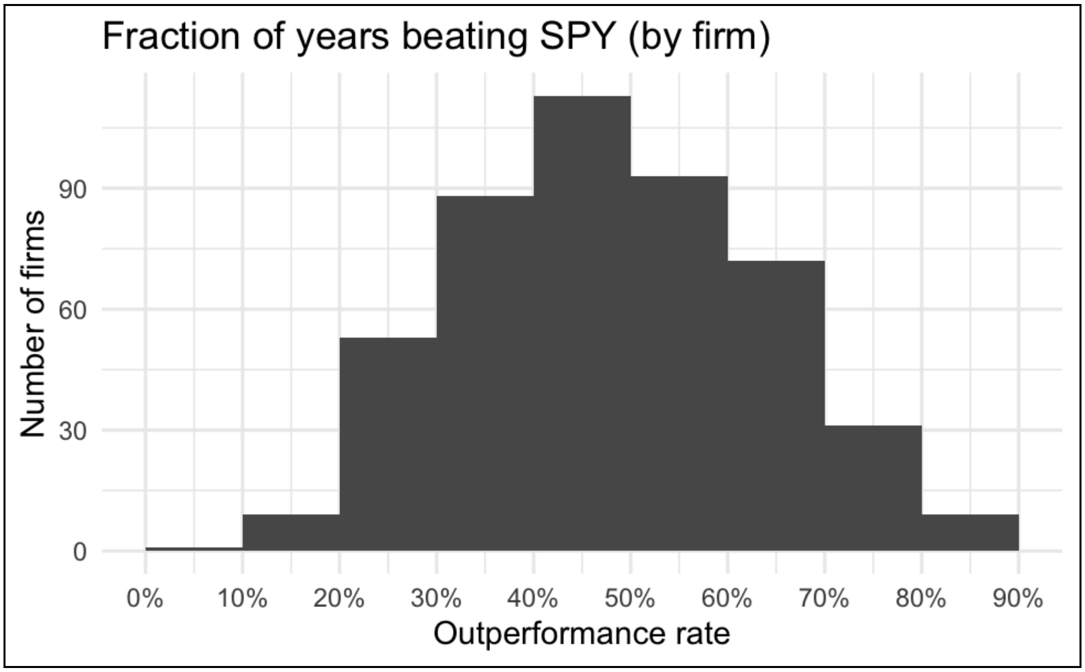

Figure 2: Distribution of Outperformance Rates

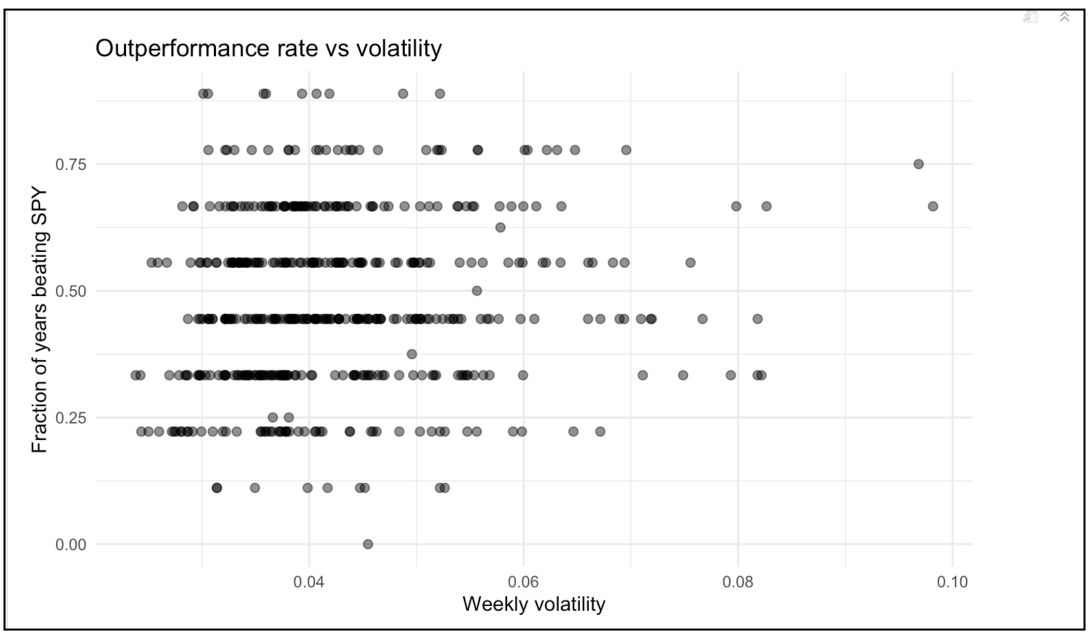

Figure 3: Volatility vs Outperformance

In our initial data set, when we look at weekly or even monthly data data, stock performance varies a lot from firm to firm, but much of this variation appears to be noisy rather than meaningful.

We measure volatility using the standard deviation of weekly returns because the underlying price data are weekly and weekly returns provide many observations per firm, yielding a more stable estimate of risk than annual volatility (which would be based on only ~9 data points). This measure is directly comparable to annual risk because weekly volatility can be annualized

Additionally, annual returns are very spread out and skewed by a small number of huge winners, which makes it clear that extreme outcomes play an outsized role and that it makes more sense to look at performance over longer horizons. Most firms beat SPY in about half of the years in the sample, and only a small handful do so consistently. This pattern suggests that many "top performers" are likely the result of luck rather than persistent skill. We also see little connection between how often a firm beats the market and how volatile its weekly returns are, implying that higher outperformance is not simply compensation for taking more risk. Overall, these patterns caution against taking raw performance rankings at face value and point to the value of a Bayesian approach that explicitly accounts for uncertainty when estimating which firms truly outperform the market.

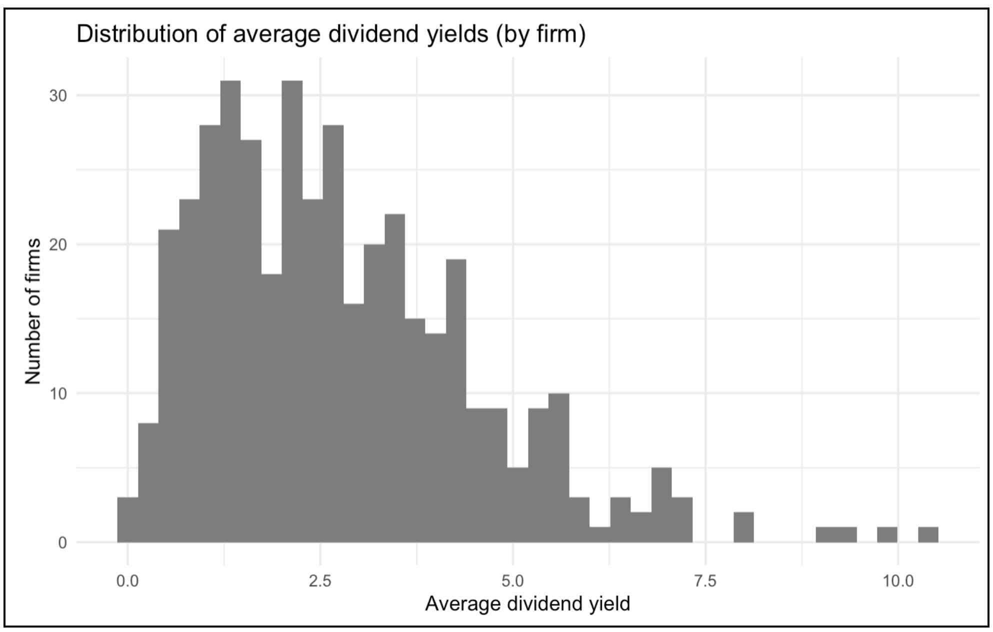

Figure 4: Distribution of Average Dividend Yield

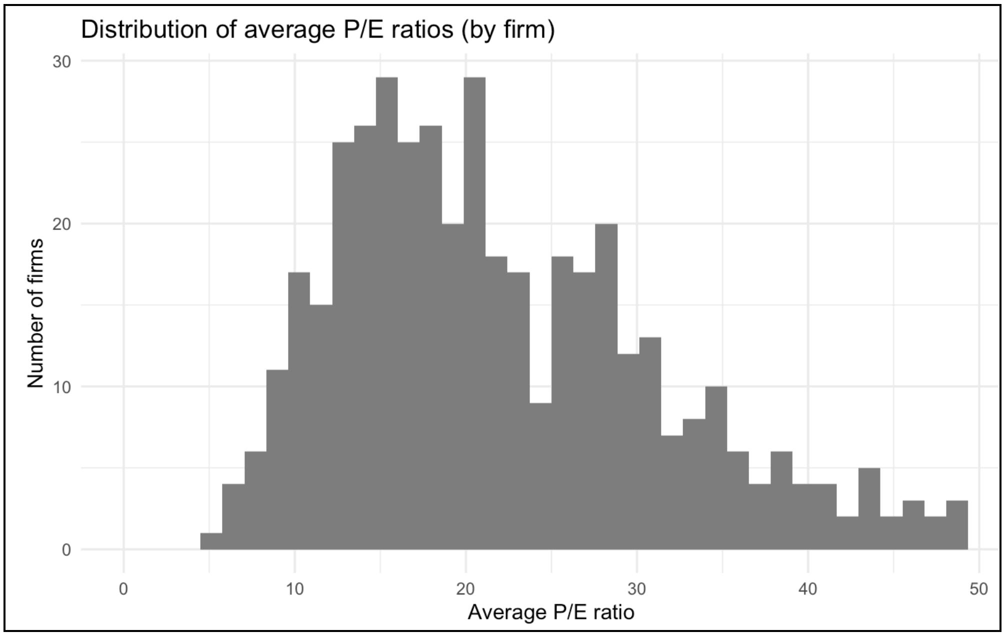

Figure 5: Distribution of Average P/E Ratio

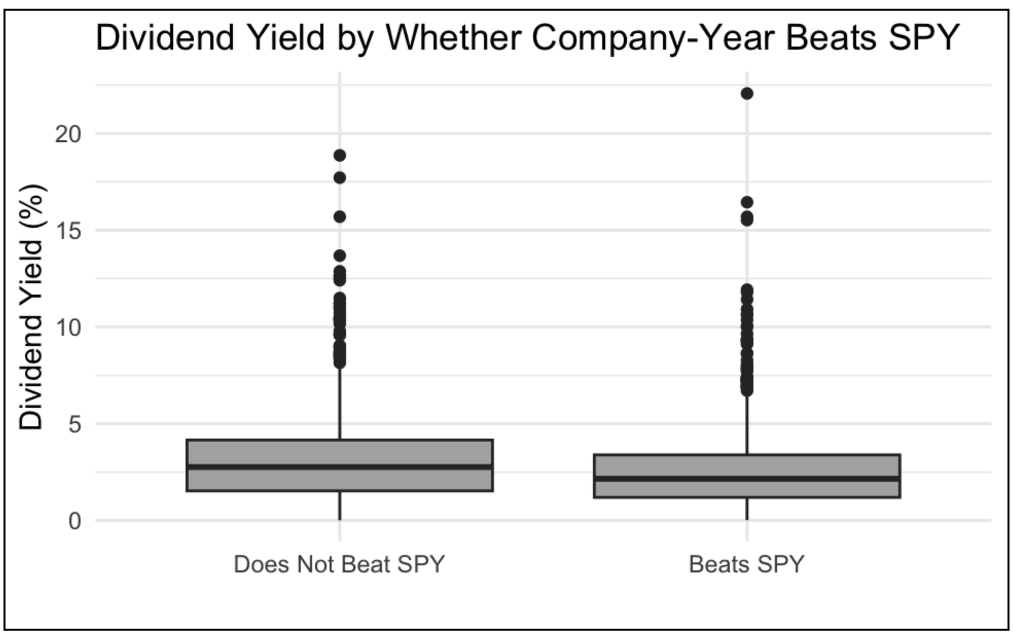

Figure 6: Dividend Yield vs Outperformance

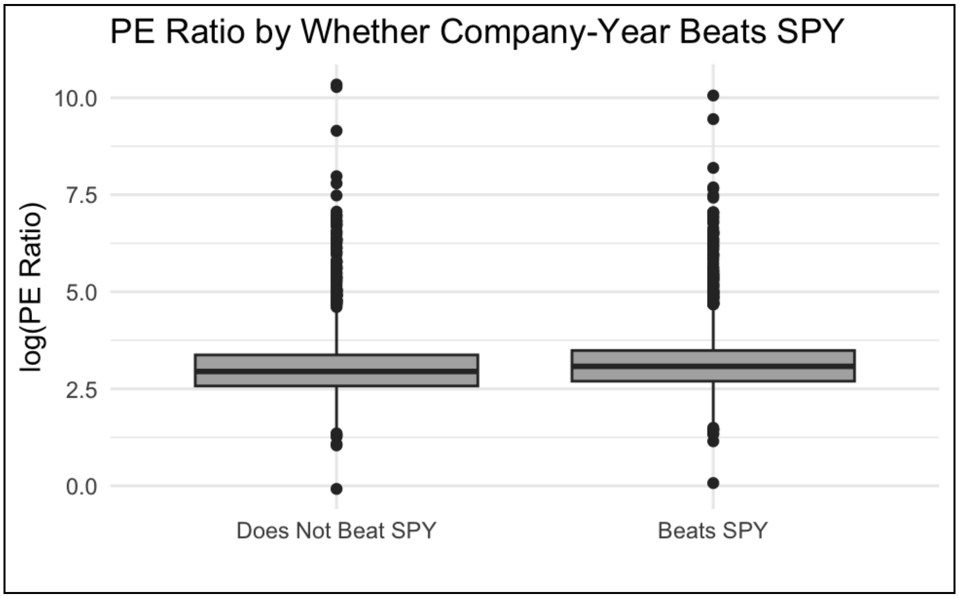

Figure 7: P/E Ratio vs Outperformance Rate

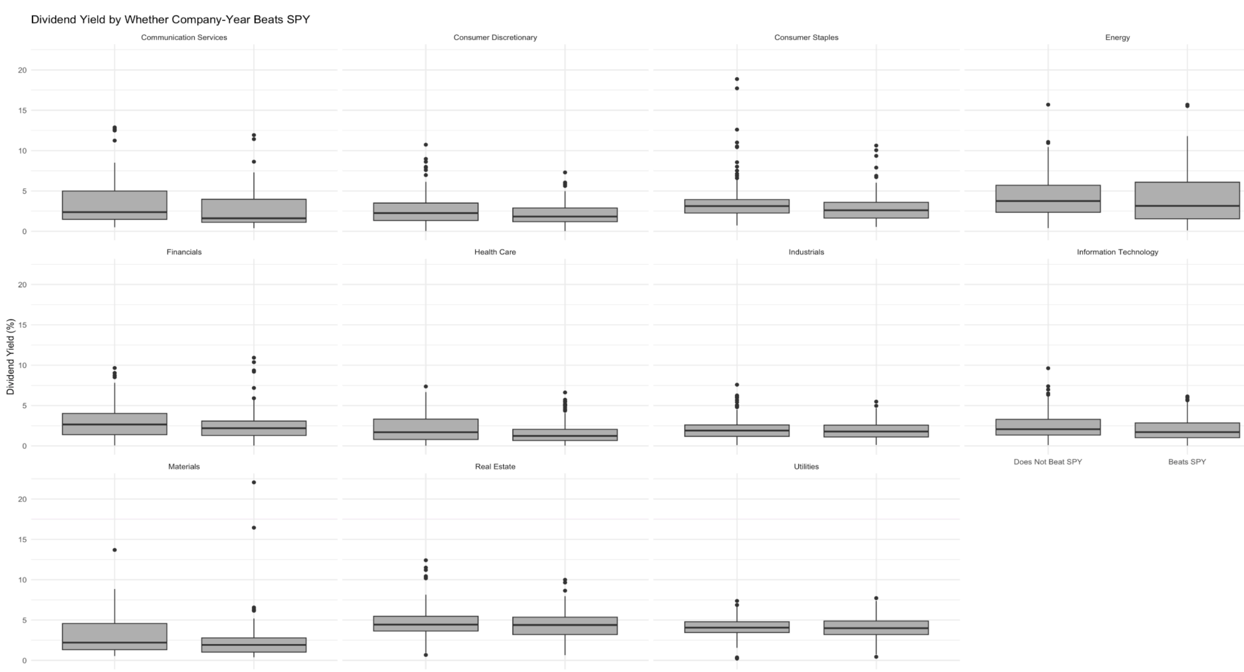

Figure 8: Dividend Yield vs Outperformance, faceted by GICS sector

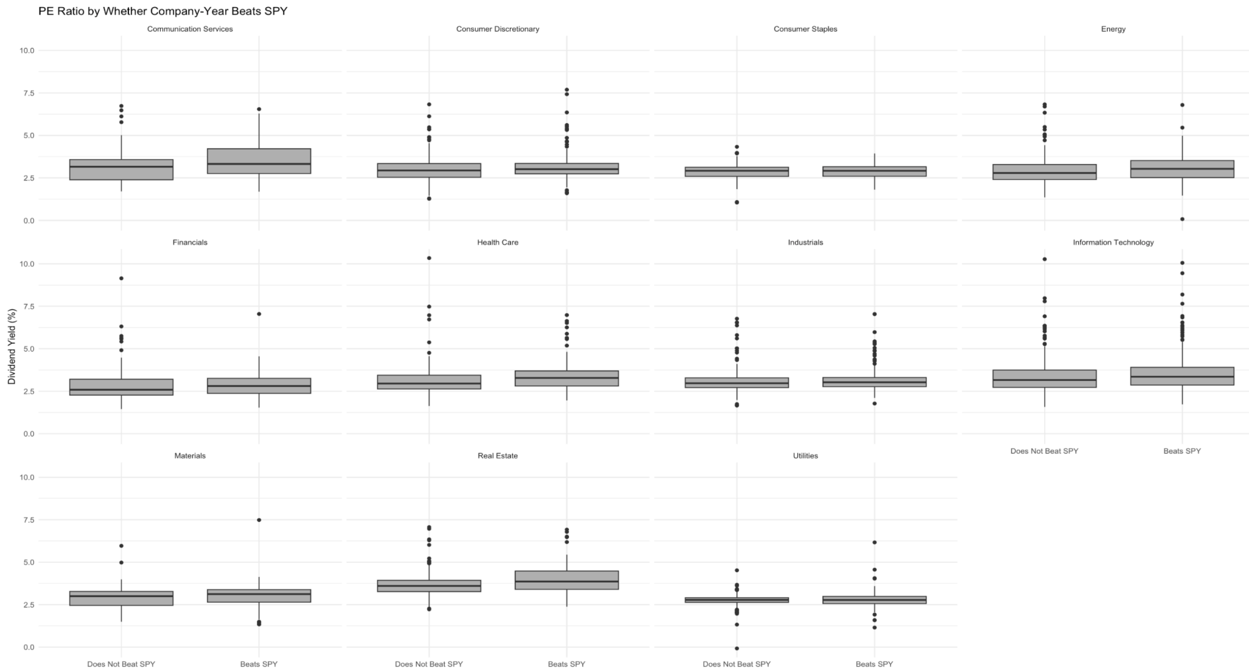

Figure 9: P/E Ratio vs Outperformance Rate, faceted by GICS sector

When we expand the exploratory analysis in terms of risk, valuation, and dividend policies, these differences do not translate into clear differences in how often they beat the market. P/E ratios and dividend yields vary widely across firms and have long right tails, which tells us there is real dispersion in fundamentals. When we plot these characteristics by firm-year, categorized by whether or not the firm-year beat the SPY (Figs. 6, 7), we see that firm-years which beat the SPY see lower median dividend yields and higher PE ratios versus firm-years that did not beat the SPY. This gap appears consistent even when only comparing firm-years in a shared sector. However, the differences appear small and noisy. Higher volatility does not reliably lead to more frequent outperformance, suggesting that simply taking on more risk does not mechanically increase the chances of beating the market.

What stands out most across all the plots is how tightly outperformance rates cluster around 50%. Since we only observe a small number of annual outcomes per firm (9 full calendar years), this discreteness makes it easy for firms to look unusually good or bad just by chance. Overall, the exploratory data analysis suggests that apparent differences in outperformance across firms are largely driven by noise rather than persistent skill, and that simple, realized comparisons can be very misleading without a framework that explicitly accounts for uncertainty. Thus, these findings motivate the use of a Bayesian framework.

3 Proposed Frequentist Analysis

3.1

To study market outperformance at the firm level, we have adopted a binary approach to define if a firm did/did not outperform SPY in a given year. Using this approach, we assume that the outcomes follow a Bernoulli distribution where the probability of *firm i* beats the market benchmark in a given year *t* from 0 to 1. Additionally, we treat each firm's performance in different years as independent observations. This means that, once returns are aggregated to the annual level, how a firm performs in one year is assumed not to directly determine how it performs in the next. This is a reasonable simplifying assumption because annual returns already smooth out short-term momentum, reversals, and other high-frequency dynamics that operate at the weekly or monthly level.

$Y_{it} \sim \text{Bernoulli}(p_{it})$, where p_{it} := Probability that firm i outperformed SPY in year t

$$logit(p_{it}) = \beta_0 + \beta_1 log(PE_{it}) + \beta_2 DivYield_{it}$$

We use log P/E ratio as some companies have extremely high P/E ratios which can dominate our regression and skew our results.

For each firm, we calculate its outperformance rate as the fraction of years in which it outperformed the market. To keep comparisons fair, we only include firms with at least eight years of data. We measure uncertainty using standard 95% confidence intervals based on a normal approximation.

Ticker <chr>	n_years <int>	y <int>	p_hat <dbl>	se <dbl>	ci_low <dbl>	ci_high <dbl>
AVGO	10	9	0.9	0.09486833	0.7140581	1
CDNS	10	9	0.9	0.09486833	0.7140581	1
CTAS	10	9	0.9	0.09486833	0.7140581	1
FICO	10	9	0.9	0.09486833	0.7140581	1
SNPS	10	9	0.9	0.09486833	0.7140581	1
AJG	10	8	0.8	0.12649111	0.5520774	1
APH	10	8	0.8	0.12649111	0.5520774	1
COST	10	8	0.8	0.12649111	0.5520774	1
FI	10	8	0.8	0.12649111	0.5520774	1
INTU	10	8	0.8	0.12649111	0.5520774	1

Figure 10: Top 10 by frequentist probability to beat the S&P

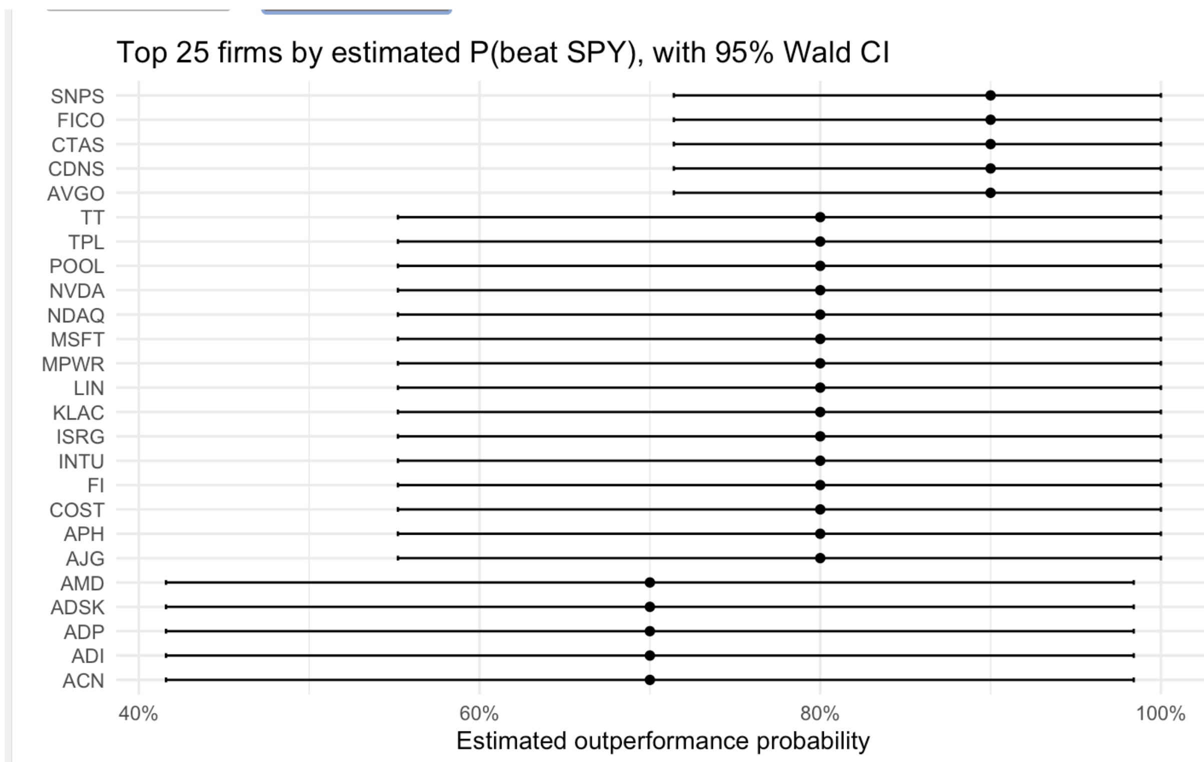

Figure 11: Top 25 by frequentist probability to beat the S&P

While the estimated outperformance rates vary quite a bit across firms, these differences are often hard to interpret. Because each firm is observed for only a small number of years, the confidence intervals are wide and overlap substantially. This means that many firms that appear to be strong performers may not be meaningfully different from one another once sampling noise is taken into account. As a result, rankings such as the ones above based on historical outperformance can be unstable and misleading. This limitation motivates the Bayesian to better separate true differences in performance from random variation.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.17522    0.19108   0.917    0.359
log_pe       0.03057    0.05524   0.553    0.580
div_yield   -0.15169    0.01974  -7.684 1.55e-14 ***
```

Figure 12: Logistic Regression (Model 1) Results

Based on figure 10: Dividend yield is strongly and robustly related to outperformance, while valuation (log P/E) is not.

Starting with the coefficients, the estimate on log(P/E) is small and statistically insignificant . This implies that, conditional on dividend yield, higher or lower valuation multiples do not meaningfully change the probability that a firm outperforms the market in a given year. In other words, once you account for dividends, being cheap/expensive on a P/E basis does not translate into higher odds of beating SPY in your sample.

By contrast, dividend yield has a large, negative, and highly significant coefficient. Interpreted directly on the log‑odds scale, higher dividend yield is associated with a lower probability of beating SPY in a given year. This is consistent with the idea that high‑dividend firms tend to be more mature, slower‑growing, and less likely to deliver market‑beating total returns, even if they provide steady income.

term	estimate	std.error
<chr>	<dbl>	<dbl>
(Intercept)	1.191511	0.19107879
log_pe	1.031042	0.05524396
div_yield	0.859258	0.01974176

Figure 13: Odds Ratio Table

In the context of an odds ratio a one unit increase in log(P/E) changes the odds of beating SPY by a factor of 1.03, which is inline with our previous estimate that it has no effect

On the other hand, a one unit increase in dividend yield changes the odds of .86 which highlights that higher dividend yield reduces odds of outperformance

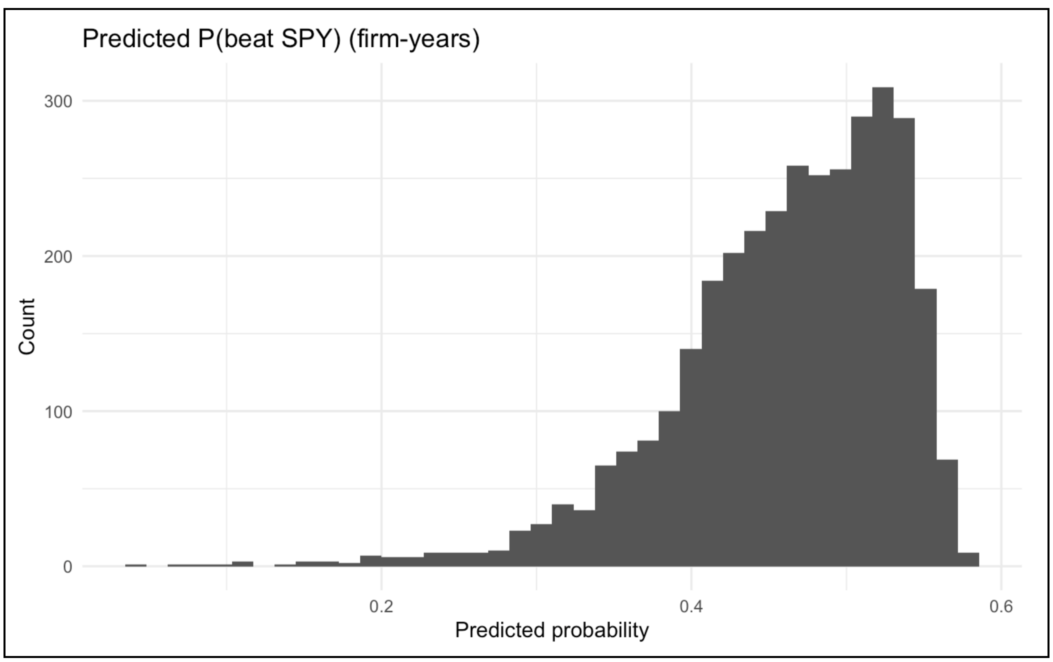

Figure 14: Predicted Probabilities based on Logistic Regression

The model assigns most firm-year observations predicted probabilities between roughly 35% and 55% which suggest that there is not a strong correlation between the two value characteristics and outperformance.

3.2

We further fit separate logistic regression models within each sector. For a firm *i* in sector *j*, the probability that it beats the SPY in year *t* is modelled as:

$$logit(p_{ijt}) = \beta_{0,j} + \beta_{1,j} log(PE_{ijt}) + \beta_{2,j} DivYield_{ijt}$$

	Health Care	Information Technology	Consumer Staples	Industrials	Utilities	Financials	Materials	Real Estate	Energy
(Intercept)	-0.846	1.524	0.684	-0.559	0.512	-2.651	0.231	0.005	-0.381
	(1.388)	(0.741)	(1.268)	(0.571)	(0.566)	(1.001)	(0.726)	(0.780)	(0.806)
log_pe	0.215	-0.304	-0.262	0.049	0.022	0.791	0.000	0.277	0.091
	(0.403)	(0.212)	(0.373)	(0.153)	(0.168)	(0.293)	(0.214)	(0.215)	(0.229)
div_yield	-0.044	-0.279	-0.181	-0.012	-0.184	-0.038	-0.074	-0.189	-0.077
	(0.097)	(0.089)	(0.090)	(0.052)	(0.060)	(0.096)	(0.081)	(0.091)	(0.073)
Num.Obs.	94	283	306	189	603	321	577	318	198
AIC	131.3	387.2	395.5	257.9	826.0	427.6	803.8	421.0	272.9
BIC	138.9	398.2	406.7	267.7	839.2	438.9	816.9	432.3	282.8
Log.Lik.	-62.641	-190.622	-194.752	-125.971	-410.016	-210.811	-398.899	-207.510	-133.45
RMSE	0.49	0.49	0.47	0.49	0.49	0.48	0.50	0.48	0.49

Figure 15: Firm-level Model Results

The firm-level model results are consistent with the Model 1 results. Dividend yield effects are consistently negative and significant, although there is heterogeneity in magnitude. P/E ratio effects are inconsistent in sign and magnitude and consistently have high standard errors, indicating that P/E ratio effects are insignificant.

3.3

We fit a mixed effects model with firm-level random intercepts and slopes. For a firm i, the probability that it beats the SPY in year t is modelled as:

$$logit(p_{it}) = \beta_0 + \beta_1 log(PE_{it}) + \beta_2 DivYield_{it} + \alpha_{0,i} + \alpha_{1,i} log(PE_{it}) + \alpha_{2,i} DivYield_{it}$$

```
Random effects:
 Groups Name        Variance Std.Dev. Corr
 Ticker (Intercept) 0.289481 0.53803
        log_pe      0.012842 0.11332  -1.00
        div_yield   0.007972 0.08928  -1.00  1.00
Number of obs: 3400, groups:  Ticker, 406

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.24602    0.20147   1.221    0.222
log_pe       0.02221    0.05797   0.383    0.702
div_yield   -0.17073    0.02119  -8.057 7.84e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
          (Intr) log_pe
log_pe    -0.949
div_yield -0.565  0.335
```

Figure 16: Mixed-effect Model Results

Looking at the fixed effects, we again see that dividend yield has a negative and significant effect, with a small and insignificant P/E ratio effect. However, looking at the random effects, it can be seen that we have achieved a singular fit (|correlation coefficient| = 1). This indicates the model is overparametrized, especially given our sparse annual data.