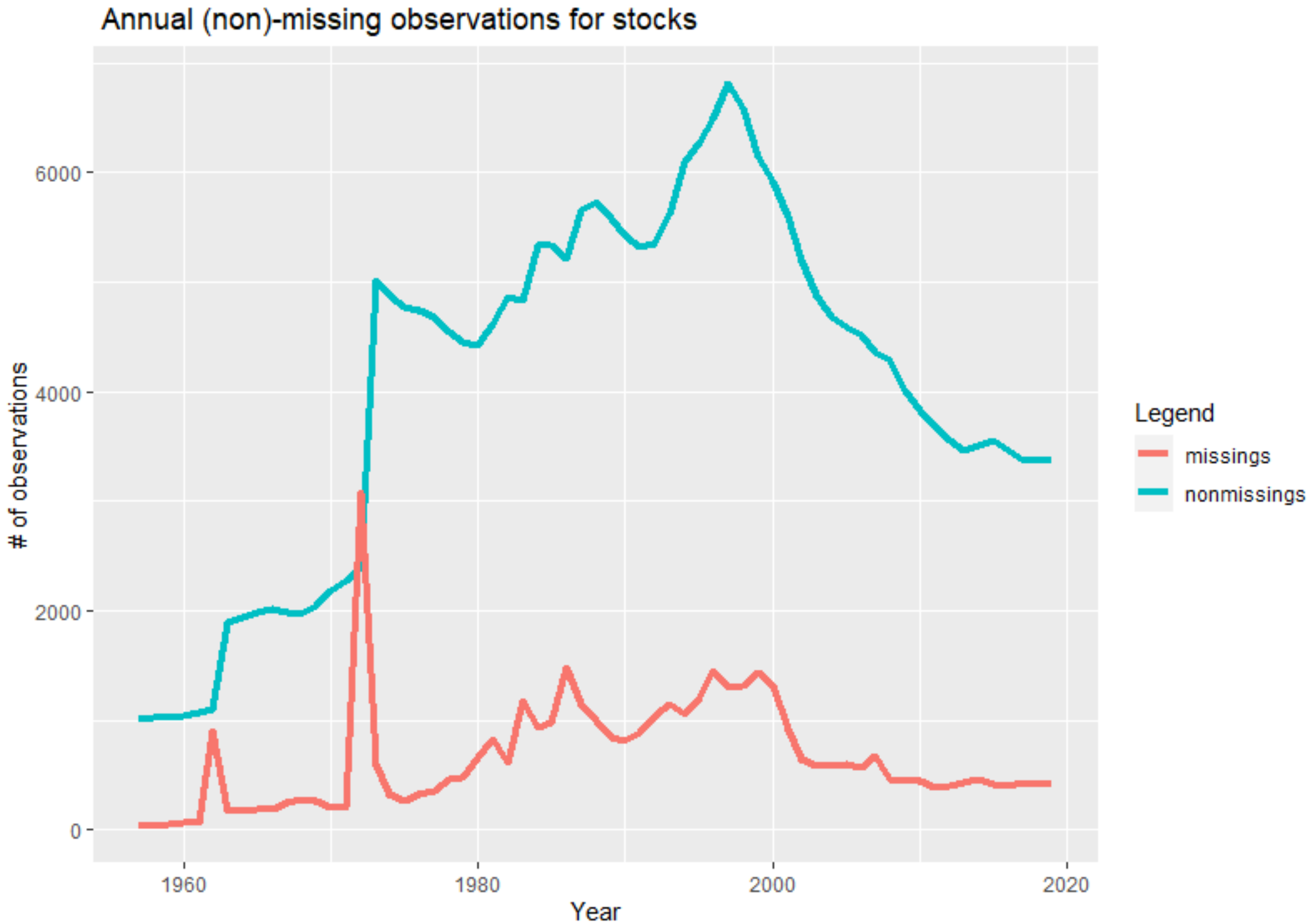
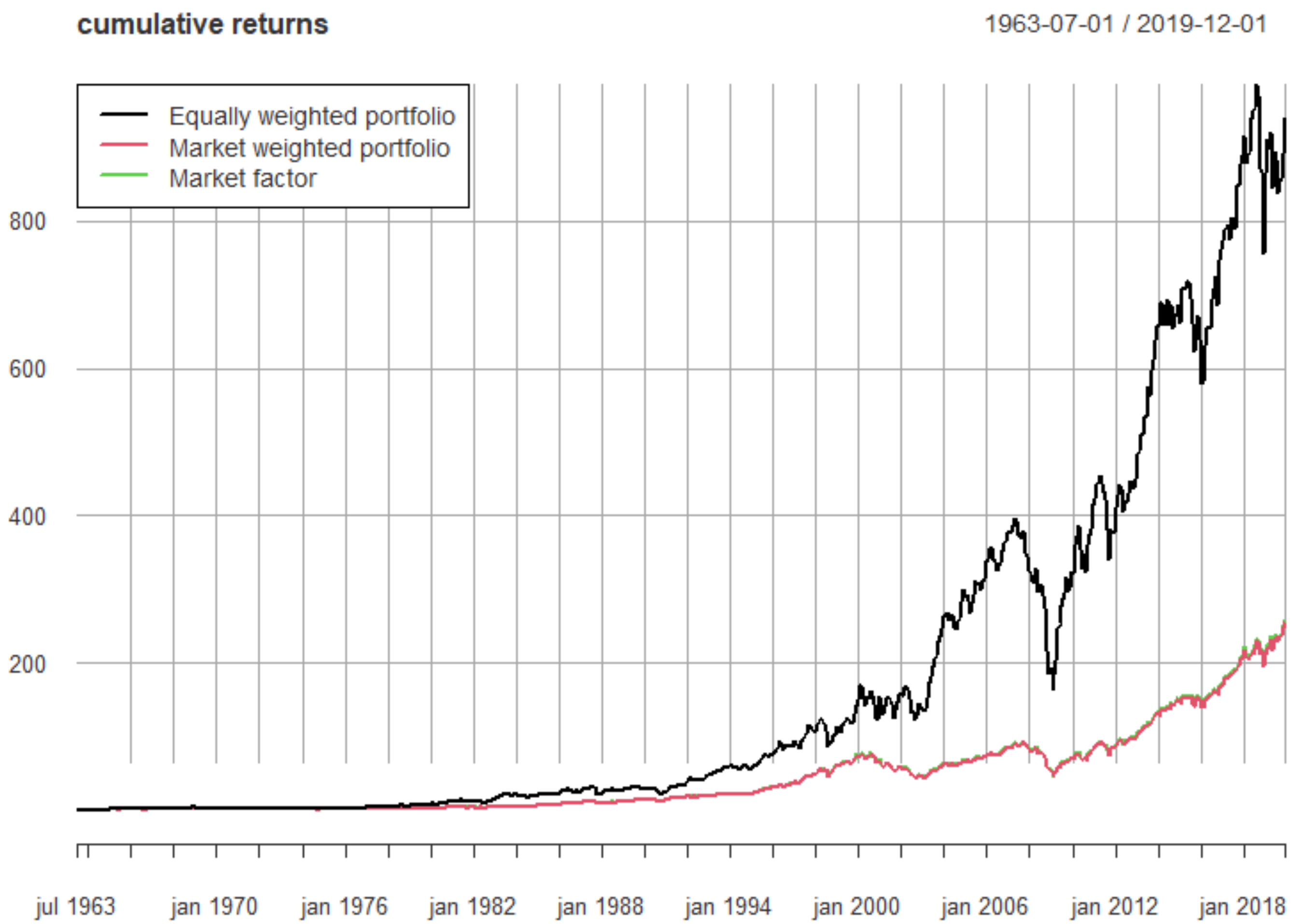
Q1 – Summary Statistics



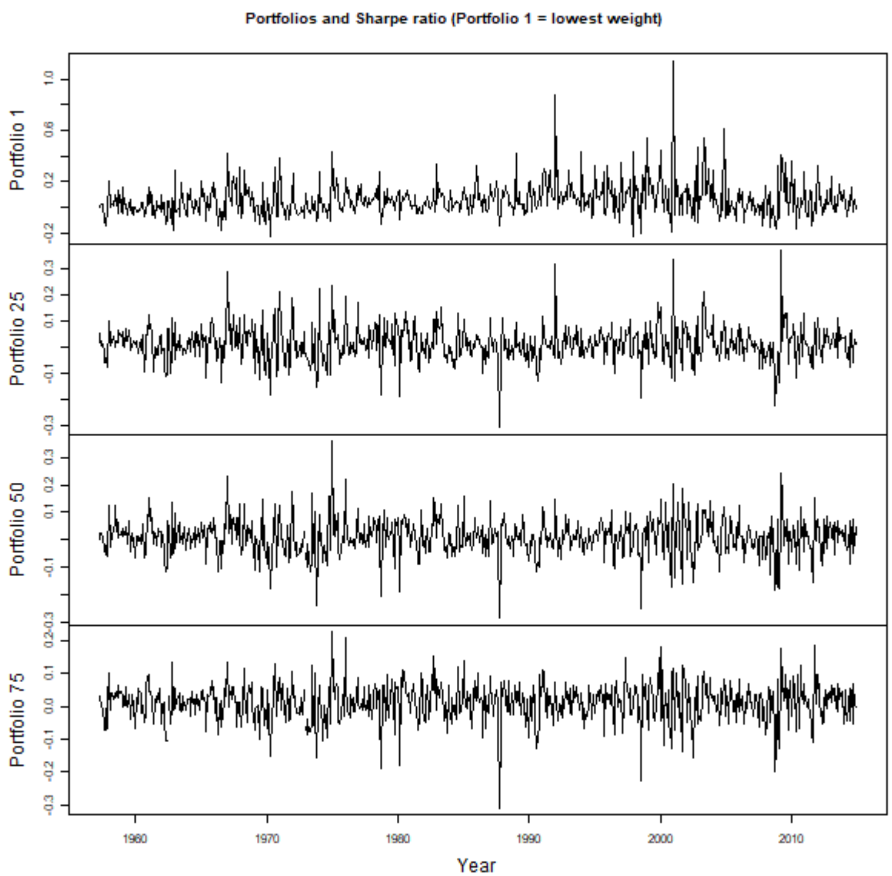
μ Monthly Return	σ Monthly Returns	μ Log Market Cap	σ Log Market Cap
0.0117	0.1471	4.6588	1.9134

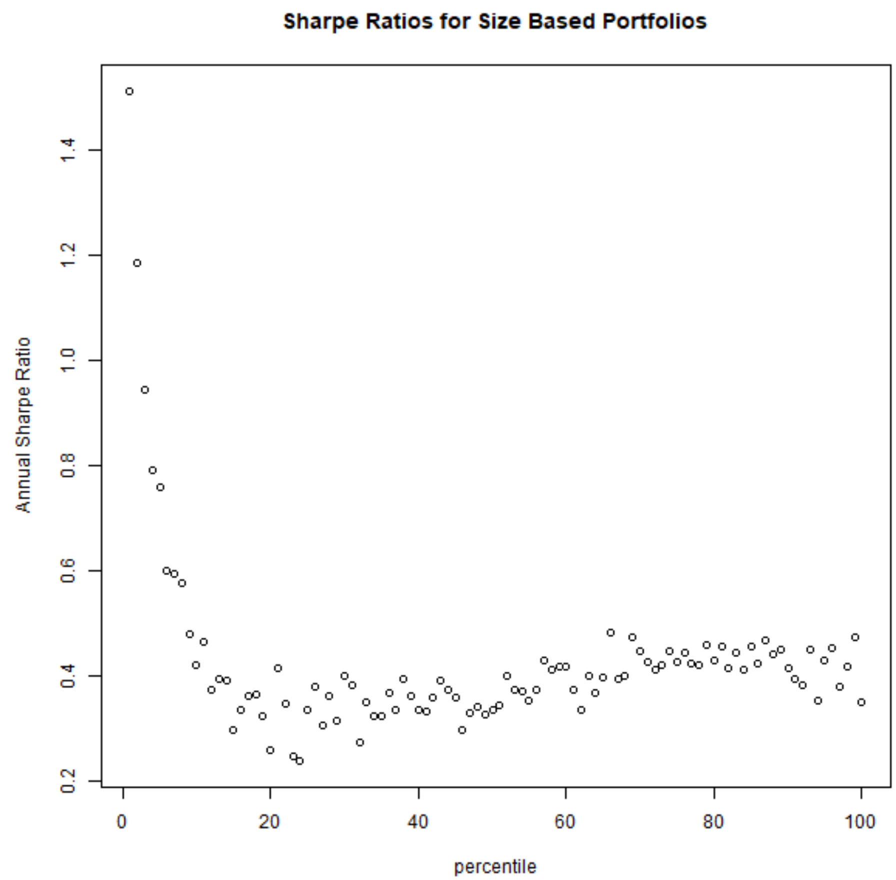
Q2 – Cumulative Returns



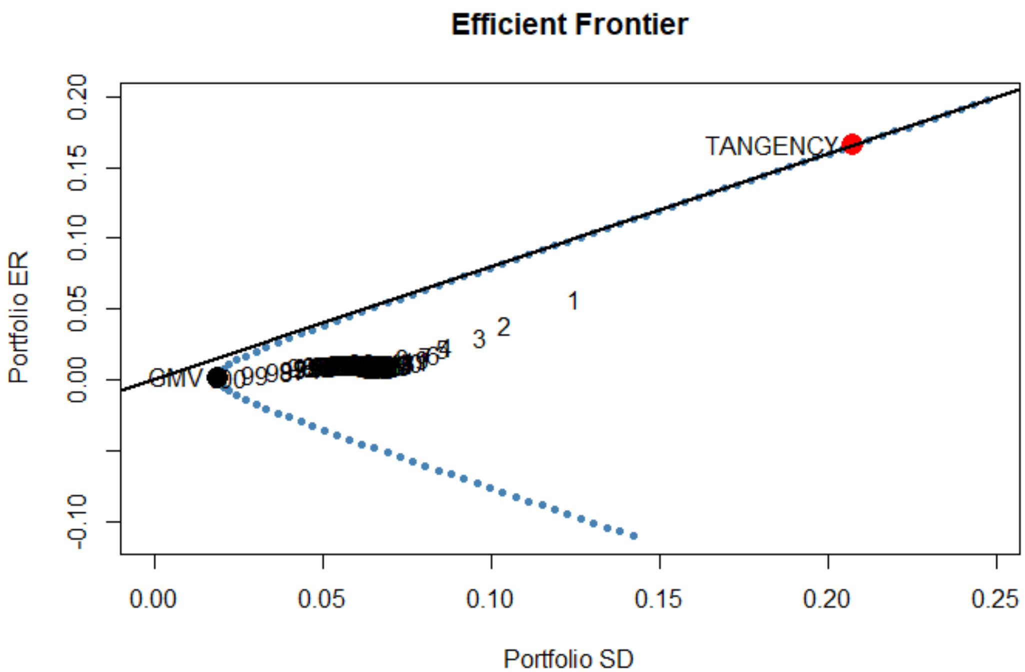
Value weighted portfolio and market factor cumulative returns are almost identical. This follows from the fact that the weights in the market factor should be the same as the weights in the value weighted portfolio (composition of portfolios are the same).

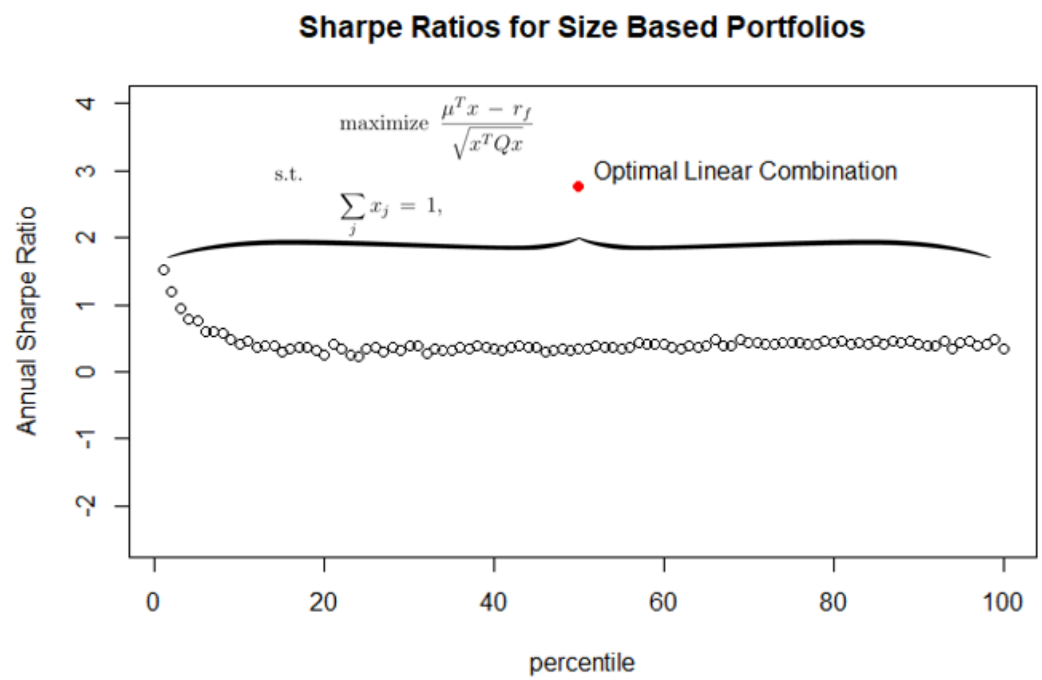
Q3 – Sharpe Ratios for 100 Size Based Portfolios





Q4 – Finding The Tangency Portfolio



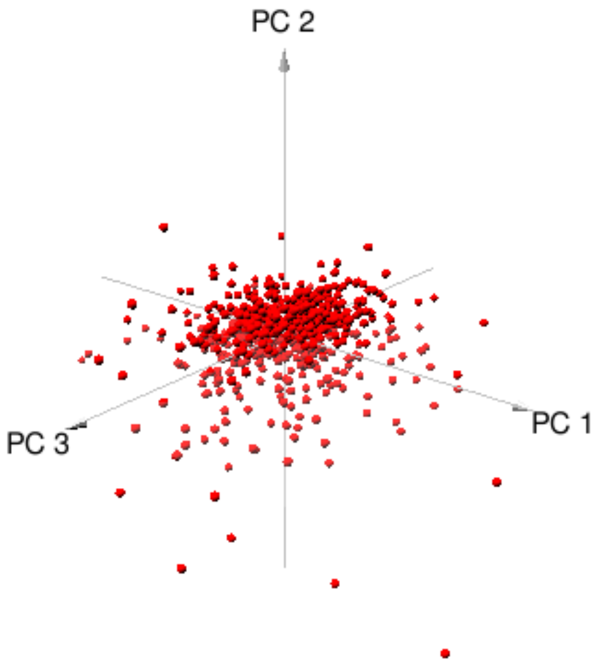


Sharpe Ratio Comparison			
Period	Optimal Weighted	Market Portfolio	Equally Weighted
In Sample	2.7617	0.3945	0.4852
Out-Of-Sample	4.2462	0.8571	0.4053

- Negative weights (short positions) allowed.
- Portfolio with smallest cap stocks performed best, which is in line with the small cap premium captured in the SMB factor of the Fama-French 5 factor model.

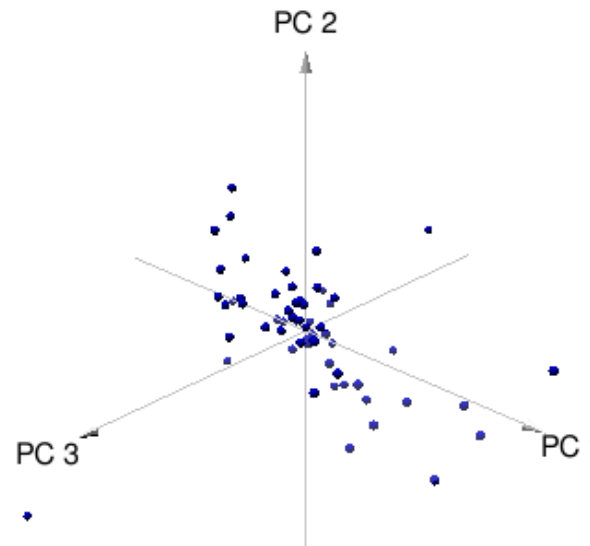
Q5 – Principal Component Analysis on 100 Factors

In-Sample



Principal Component Analysis			
PC Importance Measure	PC1	PC2	PC3
Standard Deviation	9.2068	2.9926	1.3117
Variance Explained	0.7521	0.0795	0.0153
Cumulative Proportion	0.7521	0.8315	0.8468

Out-Of-Sample



Principal Component Analysis			
PC Importance Measure	PC1	PC2	PC3
Standard Deviation	8.8800	2.7529	2.1240
Variance Explained	0.6627	0.0640	0.0379
Cumulative Proportion	0.6627	0.7264	0.7644

Q6 – Annual Sharpe Ratios for First 3 PC's

Principal Component Sharpe Ratio Comparison			
Period	PC1	PC2	PC3
In Sample	3.0065	2.0610	1.4752
Out-Of-Sample	2.6757	1.1421	0.0765

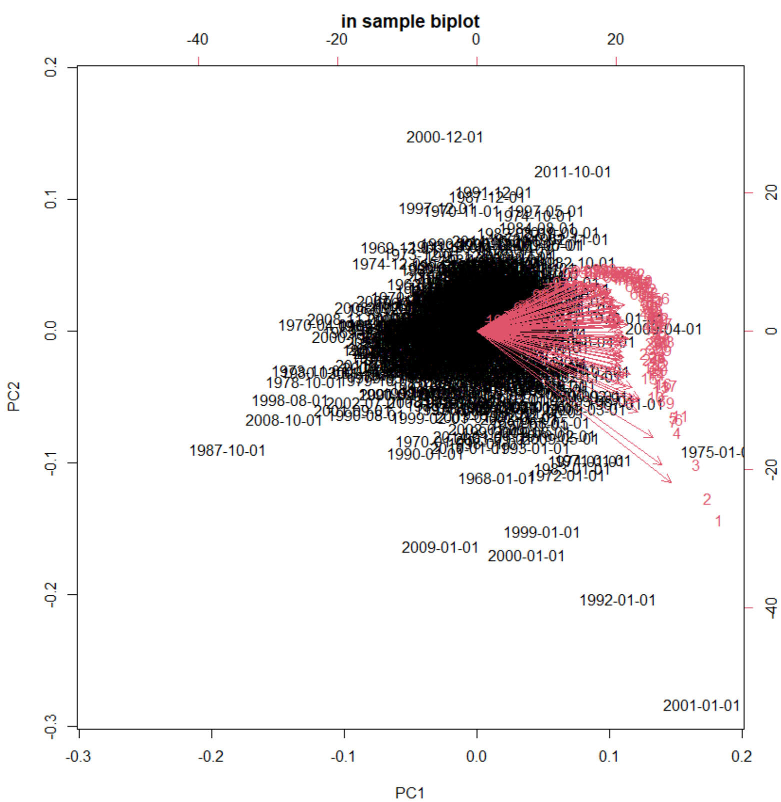
PC1 > PC2 > PC3; this performance order can again be related to the SMB factor of Fama and French, as the PC's are ordered from smallest (PC1) to largest (PC3) in terms of the stocks that they include. This comparison shows that holding small cap stocks yield higher Sharpe Ratios for both in and out-of-sample period, revealing the well-known size premium in stock returns. We can also see the presence of Markowitz stability issues (weights very skewed and extreme), as the weight allocated to PC3 will generate a very poor OOS Sharpe Ratio.

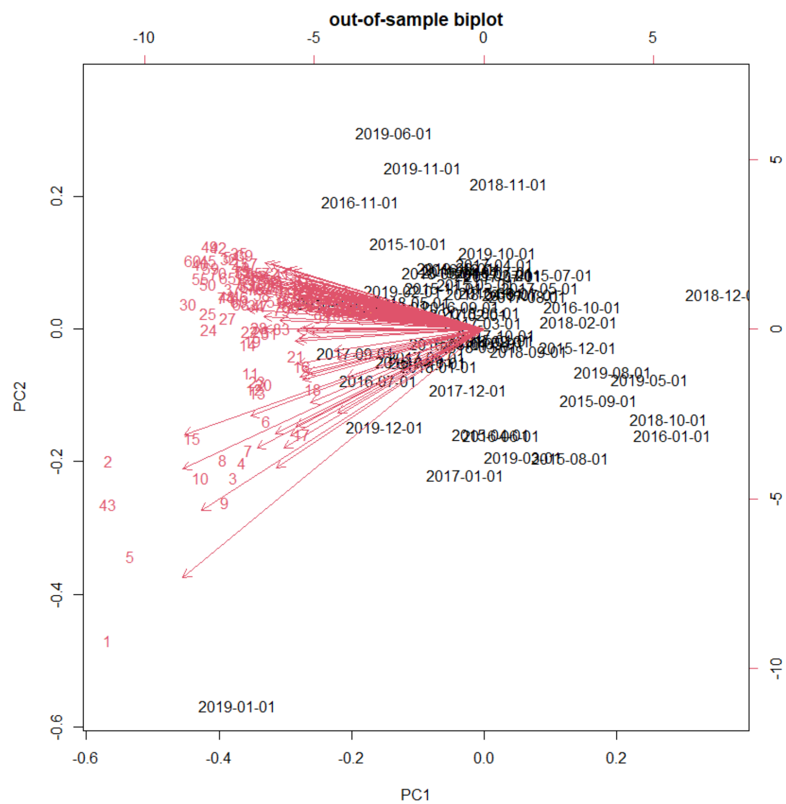
Q7 – Correlation Matrix for IS and OOS PC's

Principal Component Correlation Matrix						
PC-period	PC1-IS	PC2-IS	PC3-IS	PC1-OOS	PC2-OOS	PC3-OOS
PC1-IS	-	-0.6519	-0.2163	-0.6658	-0.4008	-0.1984
PC2-IS	-0.6519	-	-0.0062	0.4847	0.7894	0.2489
PC3-IS	-0.2163	-0.0062	-	-0.0112	-0.2611	-0.3755
PC1-OOS	-0.6658	0.4847	-0.0112	-	0.3967	0.1684
PC2-OOS	-0.4008	0.7894	-0.2611	0.3967	-	-0.0029
PC3-OOS	-0.1984	0.2489	-0.3755	0.1684	-0.0029	-

*Correlation matrix was constructed on periods in which X matrices overlapped

PC1 has a strong negative correlation with the other PC's in-sample, indicating that it captures a lot of variation in the data. We observe a mildly negative correlation between PC2 and PC3 following from the fact that their explanatory power is much weaker than PC1's. From the biplot we see that all PC's > PC3 are very highly correlated and thus have little individual explanatory power. PC1-IS and PC1-OOS have a strong negative correlation, showing persistent explanatory power for the size premium in the cross-section of stock returns.





Q8 – Missing Values and Imputation of Cross-sectional Medians

```
for (i in 1:754){
  check[i] <- (df2$monthlabel[i] == nafilling[i,1])
  df2$shrcd[is.na(df2$shrcd[c(check[i]==TRUE)])] <- nafilling[i,2]
  df2$exchcd[is.na(df2$exchcd[c(check[i]==TRUE)])] <- nafilling[i,3]
  df2$cfacpr[is.na(df2$cfacpr[c(check[i]==TRUE)])] <- nafilling[i,4]
  df2$cfacshr[is.na(df2$cfacshr[c(check[i]==TRUE)])] <- nafilling[i,5]
  df2$shrout[is.na(df2$shrout[c(check[i]==TRUE)])] <- nafilling[i,6]
  df2$prc[is.na(df2$prc[c(check[i]==TRUE)])] <- nafilling[i,7]
  df2$vol[is.na(df2$vol[c(check[i]==TRUE)])] <- nafilling[i,8]
  df2$retx[is.na(df2$retx[c(check[i]==TRUE)])] <- nafilling[i,9]
  df2$retadj.1mn[is.na(df2$retadj.1mn[c(check[i]==TRUE)])] <- nafilling[i,10]
  df2$ME[is.na(df2$ME[c(check[i]==TRUE)])] <- nafilling[i,11]
  df2$port.weight[is.na(df2$port.weight[c(check[i]==TRUE)])] <- nafilling[i,12]
  df2$ln_marketcap[is.na(df2$ln_marketcap[c(check[i]==TRUE)])] <- nafilling[i,13]
}
```

- The loop above was used to fill in N/A values with cross-sectional medians for each characteristic of the base data set.
- Month numbers were matched [i] between data sets to impute the cross-sectional median for that particular month.
- The predictors were afterwards scaled to the [-1,1] interval (also for the additional features, which we imported in question 10 - see Appendix for detailed list).

Q9 – Comparison of Summary Statistics

Summary statistics for question 1

μ Monthly Return	σ Monthly Returns	μ Log Market Cap	σ Log Market Cap
0.0117	0.1471	4.6588	1.9134

Summary statistics for question 9

μ Monthly Return	σ Monthly Returns	μ Log Market Cap	σ Log Market Cap
0	0.8231	4.5442	1.8890

- This sample is representative as the numbers are similar. We decided to not scale log(marketcap) before the comparison, as the log-transform already 'scaled' this variable. However, we do not think that an average time-series return of 0% for stocks makes a lot of sense, as stock prices generally have an upward drift (according to Ito's lemma).

Q10 – Predicting Stock Returns: Model Selection

- *Metric for model selection = highest R-squared:*

$$R^2 = 1 - \frac{\sum_{(i,t)}(r_{i,t} - f(r_{i,t}))^2}{\sum_{(i,t)} r_{i,t}^2},$$

- This metric has more 'individual meaning' than RMSE, it tells you how much variation in the response variable is captured by the estimated model.

- ***Tested models***

(i) LASSO (3 models)

(ii) Ridge (3 models)

(iii) Partial Least Squares

(iv) Principal Component Regression

(v) Gradient Boosting Machine (tree-based)

- We obtained the highest R2 for our models by tuning the hyperparameters. Regular cross-validation for hyperparameter tuning was not possible, as in time-series the order of the data matters (which would be violated in CV). This could be solved by using Time-series Cross-Validation *if* our data set would have been smaller.

Q11 – Three Best Machine Learning Models

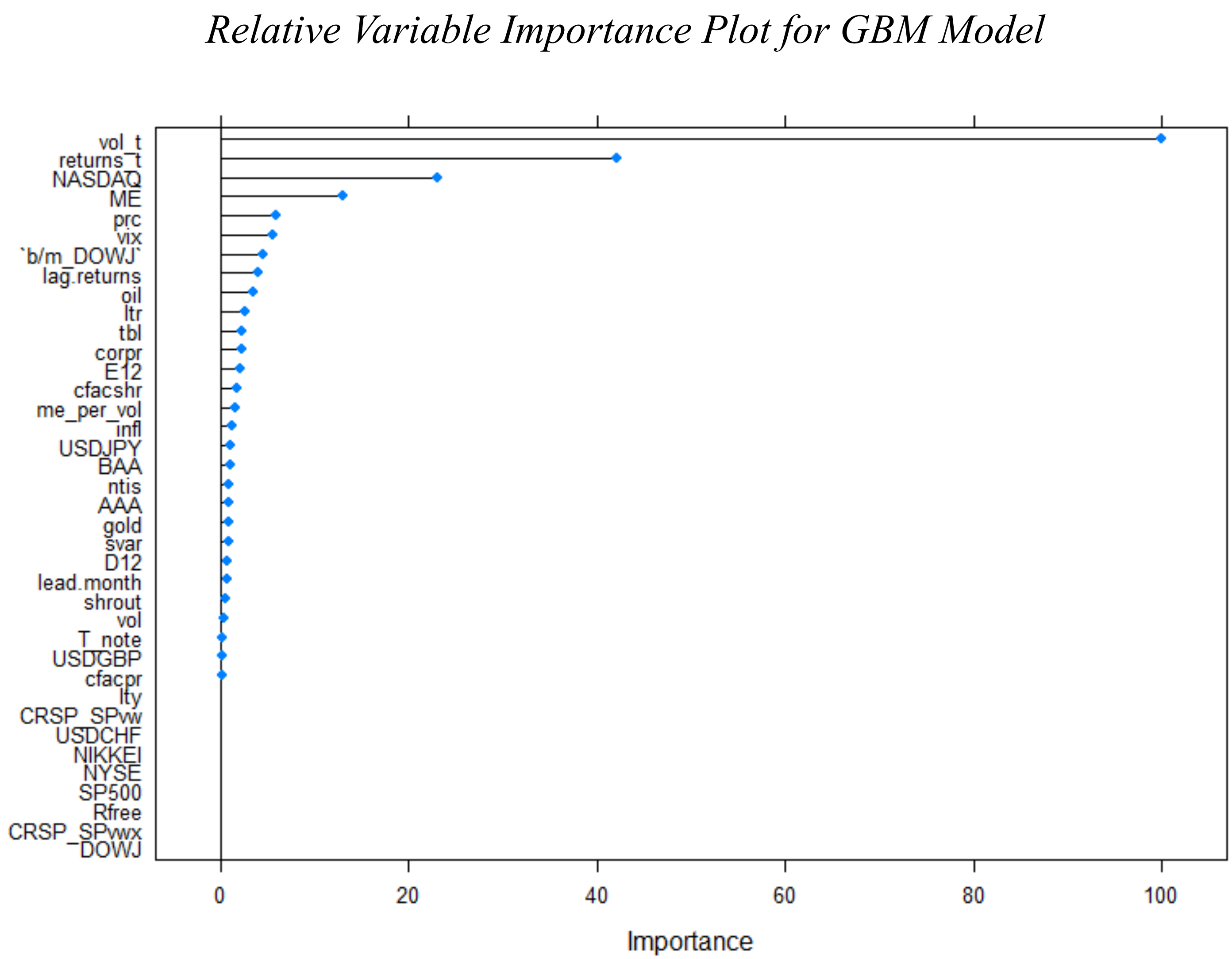
Performance Of Machine Learning Models					
ML Model	Tuned Parameters	RMSE Training	R-Squared Training	RMSE Validation	R-Squared Validation
LASSO	lambda = 0.012	0.2609	0.0139	0.1657	0.0074
Ridge	lambda = 0.07	0.2049	0.0333	0.1655	0.0105
PLS	ncomp = 1	0.2084	0.0000	0.1664	0.0000
PCR	ncomp = 1	0.2084	0.0000	0.1664	0.0000
GBM	n.trees = 400 interaction depth = 2 shrinkage = 0.01 min.obs.in.node = 5	0.2027	0.0540	0.1651	0.0149

- ***3 best models based on validation R-squared***

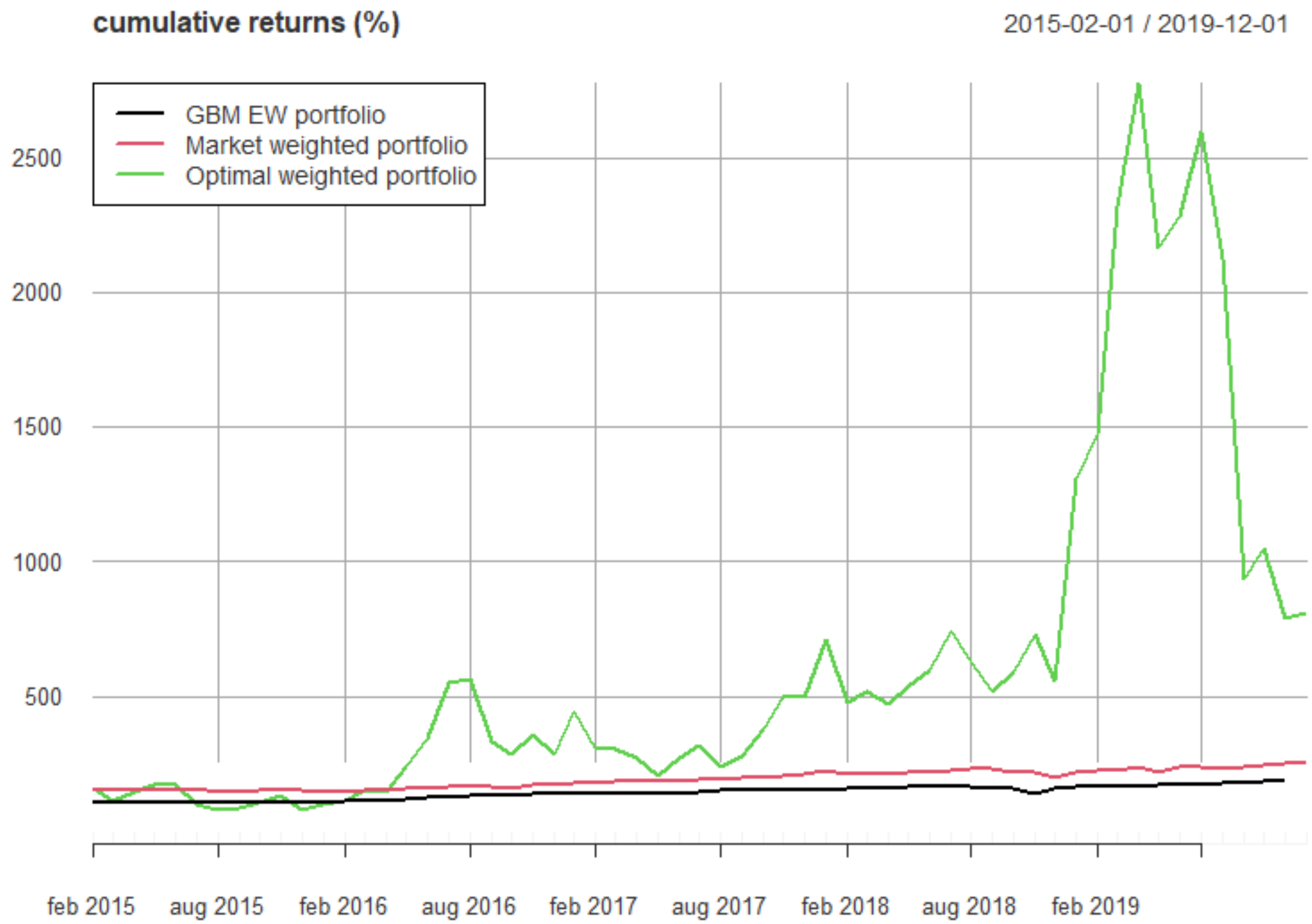
1) Gradient Boosting Machine
2) LASSO
3) Ridge regression

, ordered from best to worst performance.

Q12 – Most Important Features for Predicting Stock Returns



Q13 – Out-Of-Sample Portfolio Performance



The market portfolio outperformed our equally weighted GBM portfolio in terms of absolute cumulative returns. However, based on annual Sharpe Ratios, the GBM portfolio outperformed the market (see graph in Appendix). The tangency portfolio outperformed GBM and the market on both measures.

Q14 – Fama-French Regressions

	Portfolio 1	*Portfolio 2*
	GBM EW	OW
smb	0.4851***	0.7621
	(0.0980)	(1.8887)
hml	0.0610	−0.5019
	(0.1072)	(2.0660)
rmw	0.0877	0.3798
	(0.1647)	(3.1735)
cma	−0.0245	1.0997
	(0.1796)	(3.4598)
MarketFactor	0.5817***	4.0835**
	(0.0656)	(1.2650)
alpha	0.0060**	0.0456
	(0.0016)	(0.0419)
F-test	33.86	2.73
Adj. R^2	0.7391	0.1298
p-value	0.0000	0.0288
N	59	59

Note: Standard errors in parentheses

* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

OW = Optimally Weighted (= tangency portfolio)
EW = Equally Weighted
GBM = Gradient Boosting Machine

Steps for portfolio creation

(i) Select stocks with time-series average return of > 1%.

(ii) Equally weight all stocks in cross-sectional dimension every month; sum of weights = 1.

Findings

(i) OW portfolio has a high, but statistically insignificant alpha.
(ii) GBM portfolio generates a statistically significant alpha at the 5% level.
(iii) Both portfolios load heavily on small cap stocks (high SMB). OW has a market beta >1, likely caused by short positions which allow for overweighting in certain assets.

(iv) Existence of small cap premium (SMB) again important, GBM EW portfolio automatically loaded heavily on small cap stocks (statistically significant).

(v) The variation in returns of the GBM EW portfolio is for a large part explained by FF-5 (Adj. R2 = 0.7391).

Q14 – Bonus Question: Power of Rejecting α = 0 vs. α > 0

(i) Hypothesis

- $Test \quad H_0: \alpha \leq 0 \quad vs. \quad H_1: \alpha > 0. \qquad ; significance\ level = 1\%$

(ii) Test statistic

- P-value

(iii) Rejection region

- P-value < 0.01

(iv) Value

- Linear model 1 (GBM): P-value = 0.0041 < 0.01 ➔ Reject H_0: $\alpha \leq 0$
- Linear model 2 (OW): P-value = 0.1403 > 0.01 ➔ Fail to reject H_0: $\alpha \leq 0$

(v) Conclusion

- For the optimal weighted portfolio (OW), we fail to reject the null of a zero intercept against the alternative of a positive alpha. For our GBM portfolio (GBM) we have sufficient statistical power to reject the null, which means that we have constructed a portfolio that generates statistically significant positive alpha.

Appendix

Features added to predict stock returns (before Q10)

Category	*Factors*
Macroeconomic indicators	DOWJ, VIX, NYSE, NIKKEI, NASDAQ, SP500, Crude Oil, Gold, S&P500, NASDAQ, T-note (13w)
Major fiat currencies pairs	USD/CHF, USD/GBP, USD/JPY
Factors from Gu et al. (2019)	D12, E12, b/m_DOWJ, tbl, AAA, BAA, lty, ntis, Rfree, infl, ltr, corpr, svar, csp, CRSP_SPvw, CRSP_SPvwx

Gu, S., Kelly, B. T., & Xiu, D. (2019). Empirical Asset Pricing Via Machine Learning. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3159577

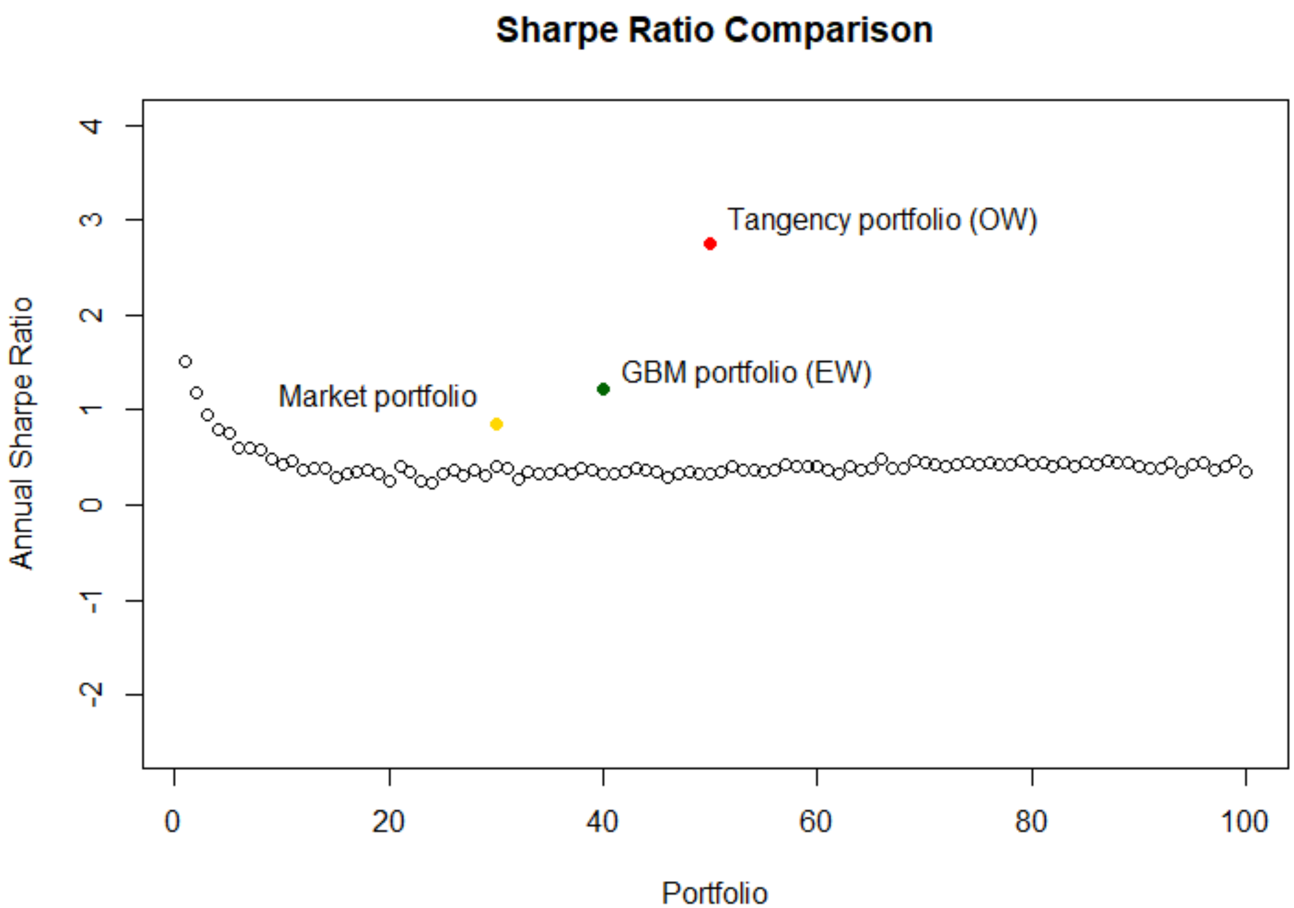
Assumptions and clarifications for specific questions

Q1: Missing values for stocks were found in a loop that counted permno observations per year (<12 obs. = missing).
Q2: Cumulative return plot starts at the date when Fama-French 5 factor data was available (1963).
Q3: Annual Sharpe Ratios calculated as: (Annualized excess return / annualized volatility).
Q4: We used the tangency.portfolio() function in R to obtain the weights that optimized the annual Sharpe Ratio.
Q5: We interpreted PC = ‘portfolio’, as there should be 100 factors according to the assignment.
Q10: Our training data starts in 1990, to avoid missing values from earlier years. This cut-off also made the data set smaller and allowed for more efficient and faster model testing.

Appendix

Q12: We used a variable importance plot to identify which features were most important for predicting stock returns.
Q13: There was nothing mentioned about portfolio weights, so we used a monthly equal weighting scheme.
Q14: We assumed that we had to plot the 3 portfolios for the test period (last 5 years of data).

Extra graph for the Q13 portfolio comparison



The transparent dots are the annual Sharpe Ratios for the 100 size-based portfolios (Q3). We believe that we achieved the ultimate goal of this assignment; identifying a ‘good’ model for predicting stock returns, and using predictions to form investable portfolios that outperform the market (in terms of Sharpe Ratio, and a positive, statistically significant alpha shown in Q14). This ultimate goal is a quote from the Canvas page.