Georgia Institute of Technology
School of Business

MGT 6203: Data Analytics in Business

Spring 2024

Team 6 Final Project Report:

Optimizing Retirement Investments for Long-Term Sustainability

Team Members:

Member 1 : Austin Franks | Edx username: austin_franks22 | GitHub username: Austin-Franks
Member 2 : Elson Dauti | Edx username: elsondauti | GitHub username: edauti91
Member 3 : Peter Chu | Edx username: pychu_2 | GitHub username: peterychu
Member 4 : Kenichi (Ken) Sugimoto | Edx username: sptd | GitHub username: etlfs
Member 5 : Suleman Shehzad | Edx username: sulemanshehzad

*__Note__: Suleman Shehzad withdrew from the course in March.

Project Repository: Team 6 on GitHub

Background/Problem Statement

Retirement is one of the largest lifestyle changes that an individual undergoes. An average American citizen is expected to live 20 years beyond their retirement date where they must find a way to provide for themselves without employment. Retirement can be a celebration of the beginning of a new chapter in life, but also can be stressful if one is not financially prepared for it. The median wealth for an individual of retirement age is $52,000 (Poterba, 2014) leaving them heavily reliant upon programs like Social Security to cover expenses. Such a low amount of wealth can be attributed to various factors such as high cost of living, low compensation, lack of steady employment, poor financial planning, or lack of optimal financial investing. This project hones in on optimizing financial investing to improve quality of life in retirement.

Not only does a larger amount of wealth make for a more comfortable retirement, but also it has been identified that it can help in extending the life expectancy of an individual. The richest man in the United States is expected to live 15 years longer than the poorest, while this gap is 10 years for women (Chetty et al., 2014). Chetty goes on to highlight how having more wealth gives access to better healthcare, healthier diets, and better fitness among other factors as shown in *Figure 1*.

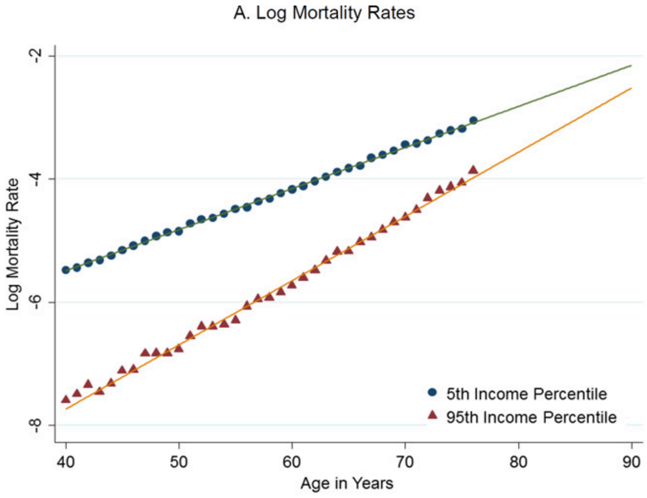


Figure 1: Mortality Rates Based Upon Income (Chetty et al.)

A powerful tool for meeting one's financial retirement goals is the stock market. It has historically had positive returns over a long time, and there are many different kinds of stocks for a person to invest into. For example, there are stocks that generate higher returns than average, but are also riskier. There are also stocks that have smaller returns than the market average, but have basically guaranteed returns. These many "risky" and "safe" stocks enable investors to construct unique stock portfolios to achieve their retirement goals.

Research Questions:

Given a retirement age goal of 67 for an individual (year 2042), the analysis put together aims to answer the question: How does one construct the optimal 401 (k) investment strategy to achieve a comfortable retirement that is not reliant upon social programs? This question can be broken down into three smaller questions that can all be addressed to provide a holistic answer.

1. Is it possible to predict life expectancy given an individual's personal information?
2. What are the expected annual expenditures during retirement?

What is the best investment strategy to achieve these goals? (In this project best is considered the investment strategy with highest returns and acceptable standard deviation.

Initial Hypothesis:

The group's initial hypothesis is life expectancy will be predictable with a linear model. They should expect to annually spend $150,000 annually on expenditures. A heavy large cap value stock investment strategy is the ideal way to accrue wealth while minimizing risk.

Methodology

The first question to be tackled is the life expectancy of retirees. The University of Wisconsin Population Health Institute aggregates data at the county level detailing metrics such a percentage of smokers, life expectancy, obesity rate, binge drinking rate, flu vaccination rate, and income. Given this data, a linear regression of life expectancy versus various other predictor variables is constructed to give future retirees a data supported estimate of life expectancy. To estimate annual retirement expenditures, an analysis of CPI and current retirement spending conditions will be used to extrapolate out into the future.

With the wealth goal generated from the first two phases of the workflow, we can begin to create investment options. 9 investment funds with various investment strategies and returns were selected to construct 4 different investment portfolios. Each fund has at least 20 years of history to ensure that there is a sufficient amount of data to provide meaningful insights. A factor regression was used to characterize the attributes of each investment fund. Additionally, Sharpe and Treynor ratios will be employed to identify the risk vs. reward of each fund.

Finally, the portfolio with the highest returns is selected and used to calculate the necessary amount of money needed at retirement to ensure a retirement life without the need to reinvest or rely upon programs like social security.

Datasets/Cleaning and initial insights

The first data set to be analyzed was the life expectancy data from the University of Wisconsin. It contains 720 different columns within the data. To get the data down to a reasonable size that could have a reasonable amount of regressions to be analyzed, it was narrowed down to 9 columns that the team predicted would be applicable to a life expectancy at the individual level. Some stats like infancy death rates would not apply to an adult that is already in the workforce. The University provided a data dictionary that helped to explain some of the column names.

Oftentimes one piece of data such as flu vaccinations had multiple columns including percent of population, a numerator, a denominator, and a handful of confidence level columns. For this analysis, the percent of population was the most effective column type to be used. Some additional cleaning involved steps were that the second row in the data was not actual data and about 50 rows did not have life expectancy estimates. Given that life expectancy is the outcome variable, those rows were removed.

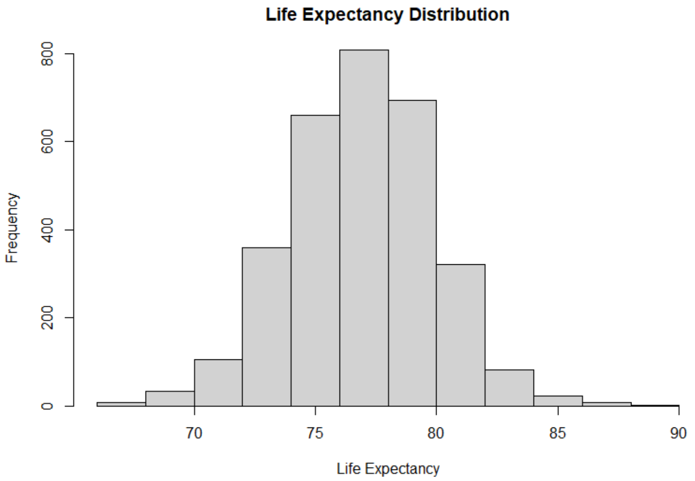


Figure 2: Life Expectancy Histogram

There were about 20 outlier points in the life expectancy data, as shown in *Figure 2*, as well with those counties having a life expectancy over 90 or under 67 that were removed. These counties did not have a large enough population to instill confidence in the validity of the numbers. Figure 2 above depicts life expectancy having a normal distribution.

To predict future retirement expenses, we utilized two datasets: one detailing 20 years of monthly inflation rates (CPI) and another tracking average annual expenses for U.S. citizens. The inflation rate was distilled to a fixed annual increase by averaging monthly rates per year and then determining the median increase across years. For average expenses, we calculated yearly percentage growth to find a consistent growth factor. Using the 2022 expense data and combining the fixed inflation and expense growth rates, we estimated the average expenses for 2042.

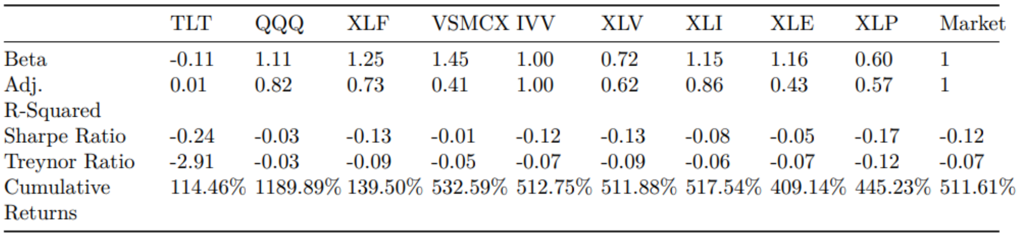
To find out how we can achieve our financial retirement goal, we used 9 assets. The individual 9 assets are as follows:

1. TLT - iShares 20 Plus Year Treasury Bond ETF
2. QQQ - Invesco Nasdaq 100 composite index ETF
3. XLE - Energy Select Sector SPDR Fund
4. VSMCX - Invesco Small Cap Value Fund Class C
5. IVV - Vanguard S&P 500 ETF
6. XLP - Consumer Staples Select Sector SPDR Fund
7. XLV - Health Care Select Sector SPDR Fund
8. XLI - Industrial Select Sector SPDR Fund
9. XLF - Financial Select Sector SPDR Fund

They have been carefully selected to have diverse & extensive coverage of various industry sectors as well as asset characteristics, such as growth, large cap vs small cap, bond vs stock.

For each asset we chose to use monthly data starting from January 2004 to December 2023 (20 years exact) as we believe that 240 data points (20 years in months) is sufficient enough for our intended research and analysis. The raw data was pulled from Yahoo Finance and included attributes such as the data, the monthly adjusted close, and any dividends paid out. This raw data was then transformed to include each month's simple return to calculate various metrics such as each asset's Beta Value. For the risk free rate we chose to use 1 month constant maturityT-bills and the data was collected from the Federal Reserve Bank of St. Louis

Table of calculated statistical metric



	TLT	QQQ	XLF	VSMCX	IVV	XLV	XLI	XLE	XLP	Market
Beta	-0.11	1.11	1.25	1.45	1.00	0.72	1.15	1.16	0.60	1
Adj. R-Squared	0.01	0.82	0.73	0.41	1.00	0.62	0.86	0.43	0.57	1
Sharpe Ratio	-0.24	-0.03	-0.13	-0.01	-0.12	-0.13	-0.08	-0.05	-0.17	-0.12
Treynor Ratio	-2.91	-0.03	-0.09	-0.05	-0.07	-0.09	-0.06	-0.07	-0.12	-0.07
Cumulative Returns	114.46%	1189.89%	139.50%	532.59%	512.75%	511.88%	517.54%	409.14%	445.23%	511.61%

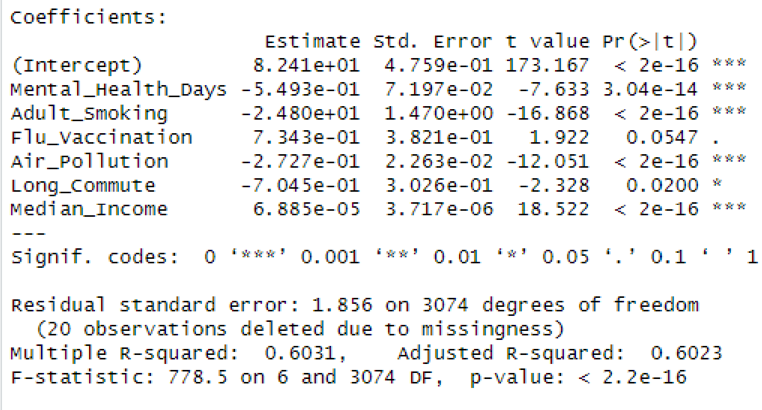
Figure 3: Statistical metrics of each asset

From these calculated metrics, we can choose how to build various portfolios for further analysis. Some initial insights, *Figure 3*, show that on average, most of our assets performed better than the market due to having Beta values greater than 1, but had higher risks as they had Sharpe Ratios higher than the market Sharpe Ratio. Another initial insight is that all of the assets have negative Sharpe and Treynor Ratios. This suggests that the risk free asset we used, 1 month maturity T-Bills, may be worthwhile to investigate. Using this information, we can perform more analysis on portfolios with different asset weights to best answer our research question.

For the factor regression, we have selected 5 factors as follows:
MKT (market), SMB (size), HML (value), RMW (profitability), CMA (investment)
Also we collect the risk free rate of returns which is the yield of short term treasury bills.
Data source is Fama French 5 factor models hosted on Dartmouth university server. (URL in the reference section at the end of this document)
For reference, this website also describes each factor in detail.

Life Expectancy Regression

To predict an individual's life expectancy, a linear regression was performed on the life expectancy data from the University of Wisconsin. In *figure 4*, all predictor variables besides a flu vaccination are significant at the 5% level. R-squared values indicate that the model can moderately predict the life expectancy based on these predictor variables. The sign of each coefficient makes logical sense with negative events like poor mental health days, smoking, heavy air pollution, and a long commute negatively affecting one's life span. On the other hand, getting a flu vaccine and having a higher income leads to a longer life.

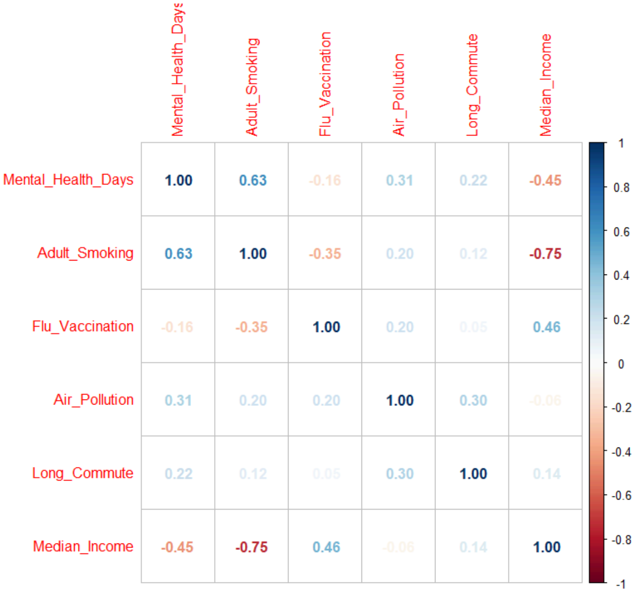


```
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)         8.241e+01  4.759e-01 173.167  < 2e-16 ***
Mental_Health_Days -5.493e-01  7.197e-02  -7.633 3.04e-14 ***
Adult_Smoking      -2.480e+01  1.470e+00 -16.868  < 2e-16 ***
Flu_Vaccination     7.343e-01  3.821e-01   1.922   0.0547 .  
Air_Pollution      -2.727e-01  2.263e-02 -12.051  < 2e-16 ***
Long_Commute       -7.045e-01  3.026e-01  -2.328   0.0200 *  
Median_Income       6.885e-05  3.717e-06  18.522  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.856 on 3074 degrees of freedom
  (20 observations deleted due to missingness)
Multiple R-squared:  0.6031,	Adjusted R-squared:  0.6023 
F-statistic: 778.5 on 6 and 3074 DF,  p-value: < 2.2e-16
```

Figure 4: Life Expectancy Regression

One important check when performing a linear regression is a multicollinearity test. Below is a correlation matrix between all of the predictor variables. The strongest correlations were poor mental health days/smoking and median income/smoking. A potential explanation is that lower income households are more likely to have individuals that smoke. Additionally, smoking can cause anxiety which can lead to having more poor mental health days.



	Mental_Health_Days	Adult_Smoking	Flu_Vaccination	Air_Pollution	Long_Commute	Median_Income
Mental_Health_Days	1.00	0.63	-0.16	0.31	0.22	-0.45
Adult_Smoking	0.63	1.00	-0.35	0.20	0.12	-0.75
Flu_Vaccination	-0.16	-0.35	1.00	0.20	0.05	0.46
Air_Pollution	0.31	0.20	0.20	1.00	0.30	-0.06
Long_Commute	0.22	0.12	0.05	0.30	1.00	0.14
Median_Income	-0.45	-0.75	0.46	-0.06	0.14	1.00

Figure 5: Correlation Matrix

To dive deeper into the matter, a variance inflation factor calculation was run on the predictor variables. VIF values for all variables lie within the 1-5 range. According to literature, a VIF within this window indicates a moderate correlation between predictors, but not strong enough to justify removing any from the linear regression model. It can be concluded that the life expectancy model detailed in *Figure 3* is a valid model.

Predicted Annual Expenses & Annual Inflation Rate

To calculate how much a person needs to invest to live off passive income if they decide to retire in 2042, we need to calculate the average annual expenses that an individual incurs.

For this project, we have chosen to consider the entire U.S. population collectively without differentiating by state-specific factors, despite the significant demographic and economic differences between states. This approach allows us to include all relevant expense categories (such as rent, car or health insurance, medical expenses, groceries, etc.) to provide a comprehensive annual calculation of average expenses across the nation.

This first step is accompanied by annual inflation rates to provide a better analysis. To calculate the annual inflation rate, multipliers were set in the model to see how much the price has changed month by month. Then a moving average was used to calculate the annual percentage change to smooth out short-term fluctuations and highlight longer-term trends in inflation, and then the mean was calculated.

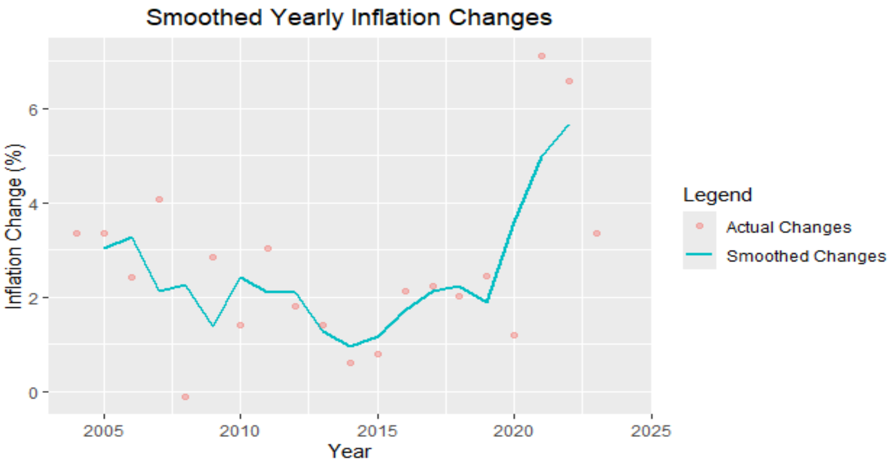


Figure 6: Smoothed Yearly Inflation

This was used to start the predictions for the average expense adjusted by the mean smoothed inflation change as shown above in *Figure 6*. A recursive calculation was used for each year from 2023 to 2042 as the chosen endpoint for our individual deciding to retire.

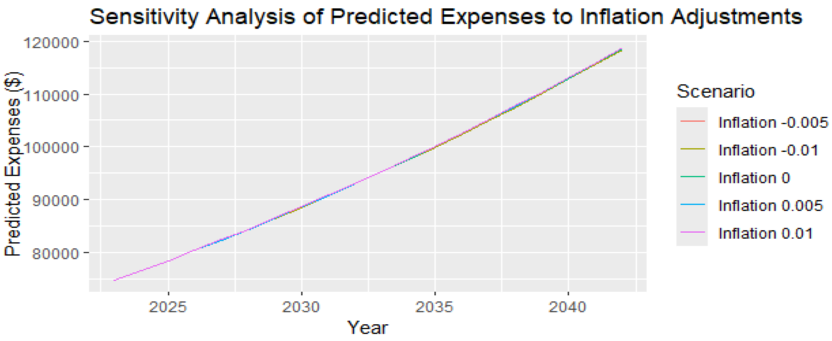


Figure 7: Sensitivity Analysis

At the end of our model, after deciding on a fixed annual percentage increase for the inflation rate as the median percentage, and after obtaining a median increase in expenses that an individual must face with data collected from 2004 to 2023, we used the median percentage obtained from the inflation increase as a fixed multiplier. This was multiplied by the annual percentage increase in expenses based on the last figure of 2023 as the starting point for our prediction of the average expenses in 2042 resulting in a predicted expense for 2042 of $118,553. There will continue to be inflation beyond 2042, but an individual can counteract this by purchasing treasury bonds.

Factor Regressions

Continuing off our initial analysis of the individual 9 assets, we first visualize the cumulative returns of individual factors for the past 20 years. As in *Figure 8*, we observe the market factor to be the dominant winner, followed by the profitability factor.

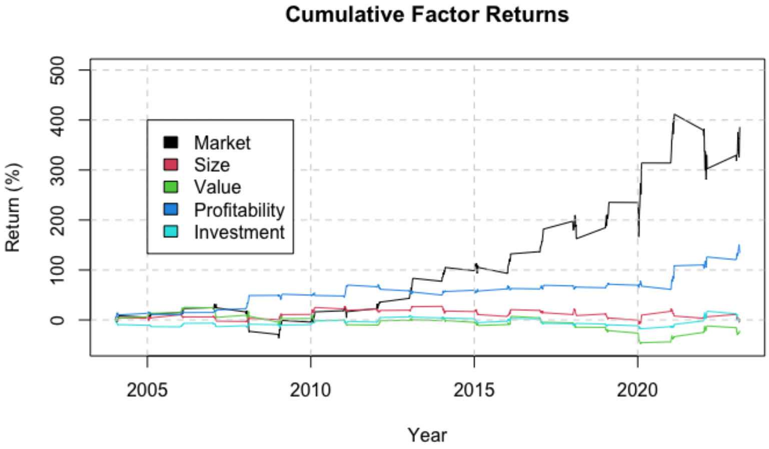
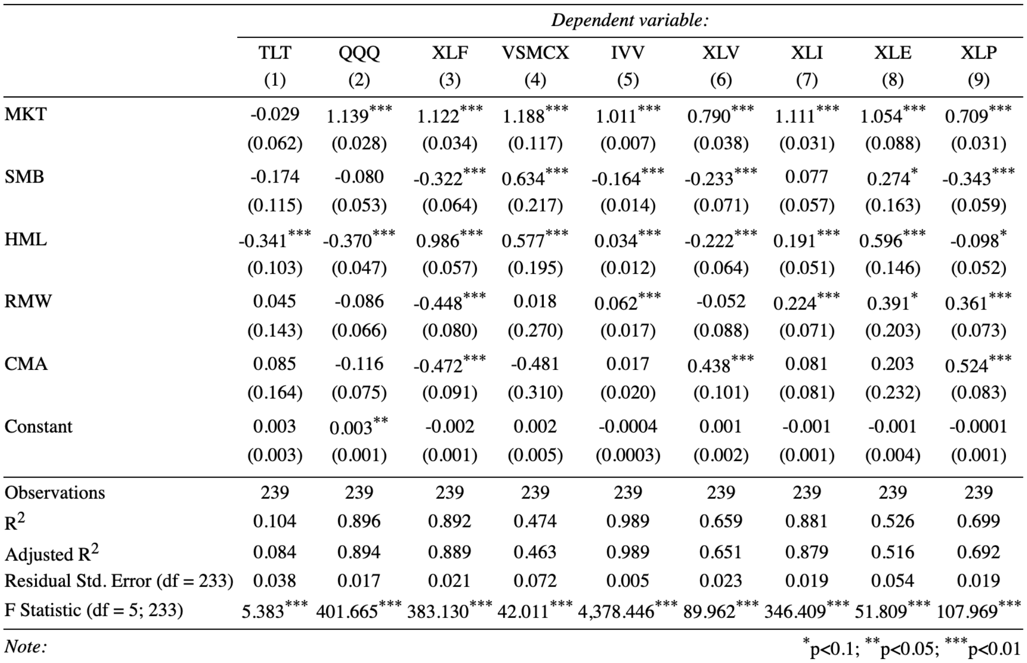


Figure 8: Factor Cumulative Returns

For factor regression, we conduct 9 separate linear regressions using R lm() functions, where we regress the monthly returns of individual assets minus the risk free rate of returns to the monthly returns of the above five factors. The results of the regressions are summarized into a table below generated using the Stargazer package in R.

As shown in *Figure 9*, the market factor and the value factor proves to be significant for most assets. Some of the interesting observations are 1. Small cap fund (VSMCX) has higher beta to the market return, indicating it is riskier but can have more returns, which confirms intuition that smaller stocks may collapse more easily but when they blossom, their returns can be extraordinary. All the giants (e.g. Apple, Microsoft) were once smaller stocks. So it is a worthy consideration to allocate some of the capital to those small stocks via VSMCX fund. Another interesting observation is how XLP (consumer staples ETF) has lower beta coefficient to the CMA (investment aggressiveness) factor. This is because the consumer staple industry is “secular” (as in, relatively immune to the macro economic condition compared to other industry sectors) in nature. Another interesting observation is TLT’s market beta coefficient is slightly negative, which makes sense as TLT is essentially a “bond” ETF which historically has been known to have inverse correlation to the equity market.



	Dependent variable:								
	TLT	QQQ	XLF	VSMCX	IVV	XLV	XLI	XLE	XLP
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
MKT	-0.029	1.139***	1.122***	1.188***	1.011***	0.790***	1.111***	1.054***	0.709***
	(0.062)	(0.028)	(0.034)	(0.117)	(0.007)	(0.038)	(0.031)	(0.088)	(0.031)
SMB	-0.174	-0.080	-0.322***	0.634***	-0.164***	-0.233***	0.077	0.274*	-0.343***
	(0.115)	(0.053)	(0.064)	(0.217)	(0.014)	(0.071)	(0.057)	(0.163)	(0.059)
HML	-0.341***	-0.370***	0.986***	0.577***	0.034***	-0.222***	0.191***	0.596***	-0.098*
	(0.103)	(0.047)	(0.057)	(0.195)	(0.012)	(0.064)	(0.051)	(0.146)	(0.052)
RMW	0.045	-0.086	-0.448***	0.018	0.062***	-0.052	0.224***	0.391*	0.361***
	(0.143)	(0.066)	(0.080)	(0.270)	(0.017)	(0.088)	(0.071)	(0.203)	(0.073)
CMA	0.085	-0.116	-0.472***	-0.481	0.017	0.438***	0.081	0.203	0.524***
	(0.164)	(0.075)	(0.091)	(0.310)	(0.020)	(0.101)	(0.081)	(0.232)	(0.083)
Constant	0.003	0.003**	-0.002	0.002	-0.0004	0.001	-0.001	-0.001	-0.0001
	(0.003)	(0.001)	(0.001)	(0.005)	(0.0003)	(0.002)	(0.001)	(0.004)	(0.001)
Observations	239	239	239	239	239	239	239	239	239
R^2	0.104	0.896	0.892	0.474	0.989	0.659	0.881	0.526	0.699
Adjusted R^2	0.084	0.894	0.889	0.463	0.989	0.651	0.879	0.516	0.692
Residual Std. Error (df = 233)	0.038	0.017	0.021	0.072	0.005	0.023	0.019	0.054	0.019
F Statistic (df = 5; 233)	5.383***	401.665***	383.130***	42.011***	4,378.446***	89.962***	346.409***	51.809***	107.969***
Note:	*p<0.1; **p<0.05; ***p<0.01								

Figure 9: Factor Regression Summary

Retirement Portfolio Selection

Incorporating the factor regression analysis, four distinct portfolios were created to determine the best portfolio for our retirement investment. Each portfolio represents its own investment thesis as below.

Portfolio 1 | aka “Diversified (by asset class)” portfolio. This combines what is considered to be the safest asset (TLT) and the riskiest asset (VSMCX) for diversification purposes.

Portfolio 2 | aka “Equal weighted” portfolio. Every asset (IVV, VSMCX, TLT, QQQ, XLP, XLE, XLF, XLI, XLV) is equally weighted in terms of notional dollar value.

Portfolio 3 | aka “Market” portfolio. We allocate IVV 50% and QQQ 50%. This combines the two most popular ETFs, effectively mimicking the average citizen's 401k portfolio.

Portfolio 4 | XLE 50% and XLP 50%. This is another approach to combining what is considered to be the riskiest industry sector (XLE = energy) and the safest industry sector (XLP = consumer staples) for diversification.

As shown in *Figure 10*, we plot the cumulative returns of those portfolios below. We observe the market portfolio performs the best (exceeding 800%) while others perform barely in a 425 ~ 535% range. It is worth noting that for the period of 2004 to 2015, the market portfolio was the least performant and even goes under water (reaching -15%) in 2008. Because the retirement investment considers the longer term result, the overall outperformance of the market portfolio still makes it the most viable choice.

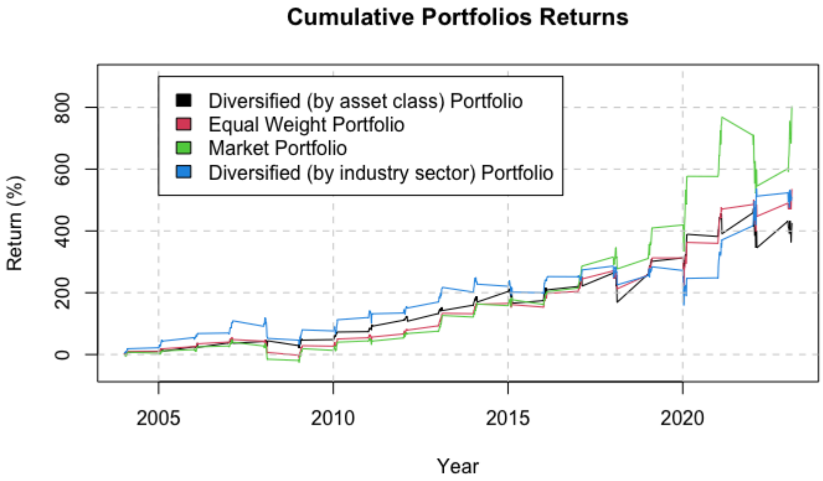
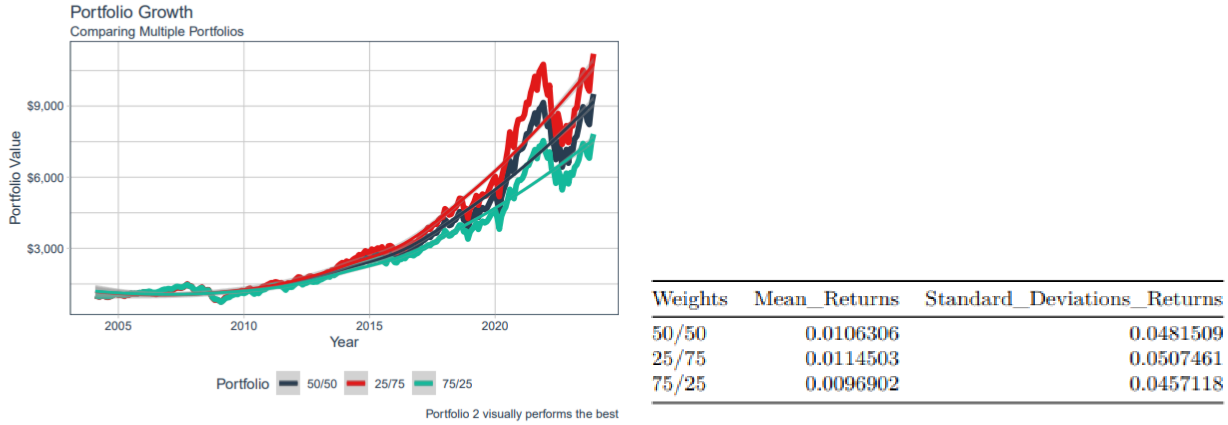


Figure 10: Cumulative Portfolio Returns

In the next section, we further analyze this market portfolio by exploring different weights of underlying assets and also explore a more elaborate scenario where we assume new cash flow gets injected into the portfolio on a regular basis, which more realistically represents how a typical retirement account is managed (i.e. people contribute little by little per paycheck over time).

Further Portfolio Analysis

Now that we have our best performing portfolio from our initial portfolio list which consists of 50% IVV and 50% QQQ stocks, we can change the weights to see if we can achieve even higher monthly returns. To test this we use our original portfolio weights of 50% IVV / 50 QQQ%, 25% IVV / 75% QQQ and 75% IVV / 25% QQQ weights as well.



Weights	Mean_Returns	Standard_Deviations_Returns
50/50	0.0106306	0.0481509
25/75	0.0114503	0.0507461
75/25	0.0096902	0.0457118

Figure 11: Portfolio Growth

Creating these 3 portfolios as shown in *Figure 11* and getting their average return, standard deviation, and plotting them we can see that a 25/75 split generates slightly higher returns. Thus, we will use this monthly return rate of 0.0114 when calculating the necessary amount of money needed at retirement so that when we invest in our portfolio, it will cover all annual expenses without the need to invest again. Basically, without touching the portfolio after investing the necessary amount, it will accrue enough returns on its own to cover all annual expenses until predicted death. For this project, we will assume that the annual expenses are equally spread out over the year. Thus, the monthly expenses will be $11,8553 / 12 = $9,879. Below is a plot showing the remaining balance of our best portfolio over the years past retirement.

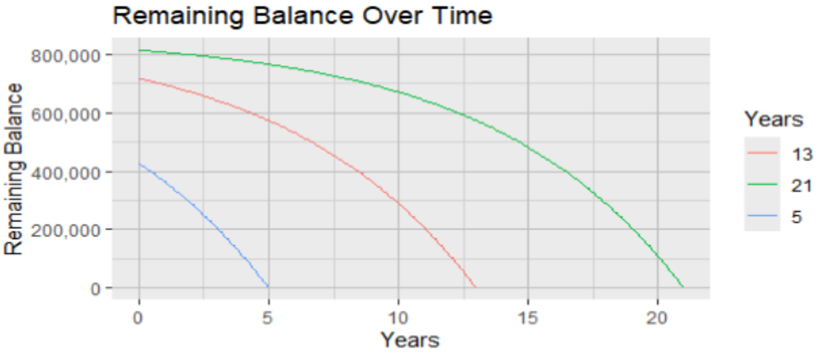


Figure 12: Remaining Balance

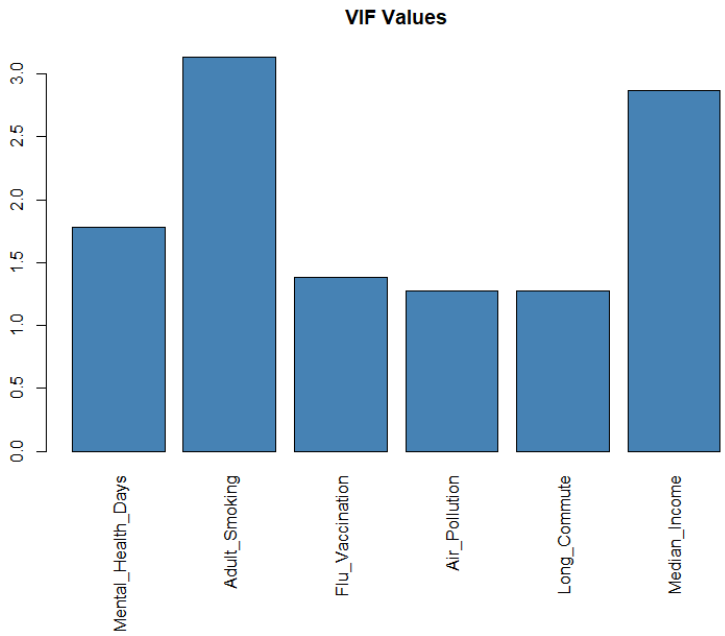
As we can see in *Figure 12*, the longer a person is expected to live past retirement, the more money they will need at retirement. However, once this goal is reached and invested, it will cover all annual expenses until their predicted death. How a person achieves this goal can be answered similarly to how we calculated this goal. A person has X years before they hit retirement age, and thus can use our portfolio investment methodology to calculate how to achieve such a goal.

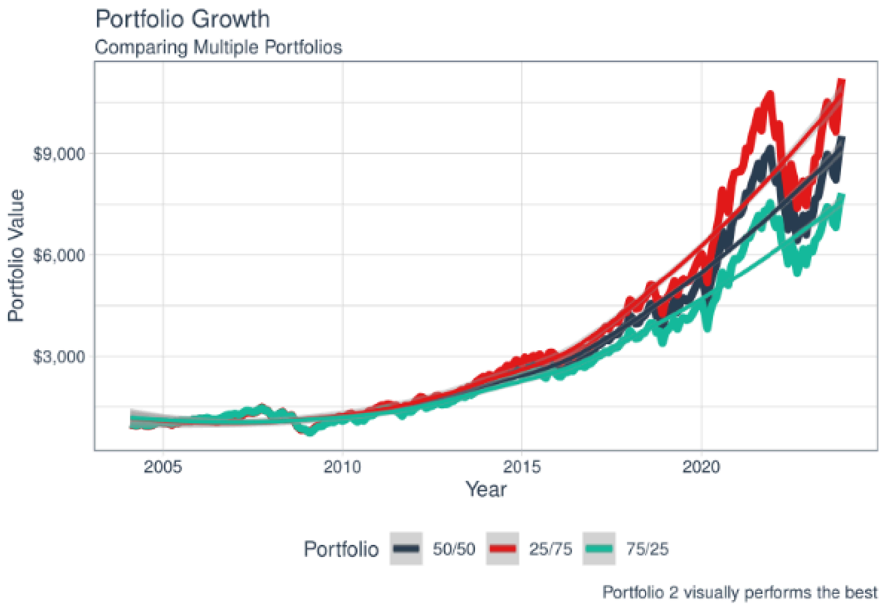
Conclusion

From our analysis we were able to answer our main research question and three sub-research questions. Our first sub-research question aimed to find the life expectancy of an individual past retirement given that they do survive to retirement age. We were able to answer this question by using quality life expectancy data from the University of Wisconsin and were able to construct a predictive life expectancy model with multiple factors and an acceptable R-Squared value. Our second sub-research question which asked what the expected average annual expenses were during retirement predicted annual expenditures of $118,553 per year. Finally, our third sub-research question asked what the best investment strategy is to cover all annual expenses until predicted death. This was solved by taking the average return rate from the best performing portfolio from 2004 to 2024 and using it to solve for a present value needed to invest in order to cover all annual expenses. The Fama French factor-based regression shows this best performing portfolio is predominantly linked to the market beta factor. After finding answers to all three of our sub-research questions, we were able to solve our main research question; a highly optimal strategy for assuring a successful retirement is building a portfolio centered upon IVV/QQQ fund that will feasibly allow an individual to not be reliant upon social programs.

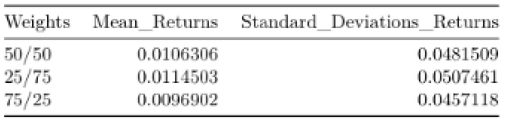
This model predicts an exact answer for how much wealth an individual would need. An additional more advanced topic would be to examine the amount of error involved with the model to make sure that an individual is prepared for a worst case scenario.

Appendix





Portfolio 2 visually performs the best



Weights	Mean_Returns	Standard_Deviations_Returns
50/50	0.0106306	0.0481509
25/75	0.0114503	0.0507461
75/25	0.0096902	0.0457118

[Reference/Citation]

- Raj Chetty, PhD et al. “The Association Between Income and Life Expectancy in the United States, 2001-2014” https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4866586/ Accessed 10 Mar. 2024
- James Poterba et al. “The Composition and Drawdown of Wealth in Retirement” The Composition and Drawdown of Wealth in Retirement - PMC (nih.gov) Accessed 10 Mar. 2024
- Yahoo Finance Historical Prices https://finance.yahoo.com/lookup?s=HISTORY Accessed 12th March, 2024
- Stargazer R software https://www.rdocumentation.org/packages/stargazer/versions/5.2.3/topics/stargazer Accessed 16th March, 2024
- Factor returns data & model details http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html#Research Accessed 16th March 2024

[R code related resources we referenced]

- How to plot multiple columns in R plot https://www.geeksforgeeks.org/how-to-plot-all-the-columns-of-a-dataframe-in-r/
- How to configure legends in R plot https://www.geeksforgeeks.org/add-legend-to-plot-in-r/
- How to customize x/y dimension axis labels in R plot https://r-charts.com/base-r/axes/#google_vignette
- How to configure title/subtitle/labels in R plot http://www.sthda.com/english/wiki/add-titles-to-a-plot-in-r-software
- How to configure grid in R plot https://stat.ethz.ch/R-manual/R-devel/library/graphics/html/grid.html
- Stargazer R software that helps summary many models into one table https://www.rdocumentation.org/packages/stargazer/versions/5.2.3/topics/stargazer https://stackoverflow.com/questions/76452883/stargazer-throws-error-when-more-than-five-models-used
- 1 Month Maturity T-Bills rate from the federal reserve bank of St. Louis https://fred.stlouisfed.org/series/DGS1MO#0