Calculations of Value at Risk for the Portfolio of 5 S&P 500 Stocks Using 5-Dimensional Copula Functions

Assanali Kakenov

April 28, 2024

Capstone Project
Supervisor: Dongming Wei
Second Reader: Francesco Rocciolo

Abstract

The purpose of this project is to compare copula estimations of Value at Risk (VaR) for a portfolio of 5 S&P 500 stocks to historical, normal distribution, and Monte Carlo methods employing dependence measures and ARIMA-GARCH time series models. This study will provide interpretations of financial data between 2019-2024 in a scope of 5 equations: Gaussian, Clayton, t-Copula, Gumbel, and Frank copulas. Correlations between closing prices of the largest 30 S&P 500 companies by market capitalization were calculated, and the portfolio was constructed by selecting 5 stocks with the least average correlation. The Markowitz portfolio optimization model was utilized to estimate the weights of the assets in the portfolio. Log returns, skewness, kurtosis, Shapiro-Wilk, and ADF were measured to describe stationarity and normality of the data. Data autocorrelation was assessed using ACF and PACF for volatility before ARIMA-GARCH modeling. All methodology was followed by appropriate hypothesis tests. Finally, 5-dimensional copulas were used for the VaR estimations for different confidence intervals. While AIC and BIC showed that t-copula was the best fit, the Clayton copula passed the goodness-of-fit test with the largest p-value. Subsequently, the Clayton copula generated VaR estimations closest to the historical data. The method used in this study can be extended for more than five assets without theoretical obstacles.

Keywords: VaR, portfolio, 5-dimensional copula, Markowitz, dependence, time series, ARIMA-GARCH, marginal distribution.

1 Introduction

The quantitative risk management is a discipline that applies the scientific method to build risk-balanced investment strategies. A significant aim of this field is to maximize the profit of an investment portfolio, while appropriately adjusting the measure of risk depending on the strategy [3], [9]. This is achieved by studying returns and variance of assets, as in the Markowitz portfolio optimization model, to identify investments that bring the largest profits for the lowest quantity of risk and assign a weight to them in the portfolio [8], [14].

The risk is also often reduced by diversification. By calculating the correlations between asset returns, risk managers determine how to hedge against similar market dynamics of several assets. To study the risk of a portfolio over a particular period of time, they calculate the *Value at Risk* (VaR) either solely from the empirical data or also by fitting them into appropriate distributions [10]. Copula functions can be used for the VaR estimations to model the joint distribution of several risk factors, which is much more accurate than the historical VaR or calculations based on the normal distribution [9], [12]. Additionally, the historical VaR does not capture the drastic events occurring in the stock market because it assumes that all trading days are equally weighted [10]. On the other hand, the copula functions help model the dependence structure between several assets with linear and nonlinear correlations without assuming symmetricity [4], [7], [11], [13]. Evidently, the quantitative risk management borrows concepts from probability and statistics that need statistical tests to show the reliability of the results.

2 Theory

2.1 Value at Risk

VaR is one of the most accessible risk measures in portfolio management. It estimates the maximum possible loss of a portfolio value over a particular time period. VaR is a single variable statistic defined for specified confidence levels [10].

Definition 1. Let a random variable L be the loss of a portfolio and γ be a given confidence level. Then, the VaR is represented as

$$P[L \geq VaR_{\gamma}] = 1 - \gamma \tag{1}$$

If one-day 95% VaR of a portfolio is \$400, then the portfolio manager potentially loses \$400 or less 95% of time on a given day [10]. Although the 95% confidence level has become a conventional metric for VaR calculations, this study aims to calculate 90%, 99%, and 99.5% VaRs to locate both the broad and precise estimations.

There is a range of approaches to VaR calculations, such as historical data, normal distribution, and Monte Carlo estimations. The historical VaR is based only on the past returns of the portfolio, and the normal VaR intends to fit the returns to a normal distribution. The Monte Carlo method is particularly useful for asset prices that follow nonlinear paths and do not fit simple distributions. All three approaches are incorporated to this study.

2.2 Markowitz Portfolio Optimization Model

One of the modern approaches for diversification of a portfolio is the Markowitz portfolio optimization model. The purpose of this model is to identify optimal values for asset weights by maximizing their expected returns and minimizing risks [8], [14]. It was first constructed by the Nobel Prize winner Harry Markowitz in 1952.

The Markowitz model is a quadratic programming problem, where the optimal solution is the weights $x_j, j = 1, 2, ...$ on the investment portfolio consisting of assets with returns $R_j, j = 1, 2, ...$ [14]. Per Vanderbei [14], they are obtained by solving

$$\begin{array}{lll}
\text{Minimize} & -\sum_j x_j \mathbb{E} R_j + \mu \mathbb{E}\left(\sum_j x_j \tilde{R}_j\right)^2 & \\
\text{subject to} & \sum_j x_j = 1 & \\
& x_j \geq 0 & j = 1, 2, ...
\end{array}$$

where

R_j - the return of the jth asset in the portfolio,

μ - the parameter of variance,

$\mathbb{E}R_j$ - expected return of a single asset, such as a stock,

$\mathbb{E}(\sum_j x_j \tilde{R}_j)^2$ - the variance of the return of the entire portfolio.

The variance is an estimation of risk, as higher variance reflects volatile performance of a company on the stock market and lower variance means a gradual change in prices. Higher value of the parameter of variance indicates low risk tolerance and vice versa. Due to the computational complexity of the problem, a portfolio of only 5 assets are considered in this project.

2.3 Copula Functions

Copula functions are a mathematical tool applied to finance for the study of dependence between assets, markets, risk factors, and other measures [4], [7], [12]. Copulas have become widespread in the field due to the need to arrive at more accurate estimations in the increasingly volatile behavior of the financial markets [4].

Copula functions are best described by the Sklar's theorem.

Theorem 1. The m-dimensional copula is a function denoted by C such that $[0,1]^m \to [0,1]$. The function satisfies the following three properties:

$$\begin{aligned}
&(1) \quad C(1,\ldots,1,a_n,1,\ldots,1) = a_n \quad \text{for every } n \le m \text{ and } \forall a_n \in [0,1];\\
&(2) \quad C(a_1,\ldots,a_m) = 0 \quad \text{if } a_n = 0 \text{ for any } n \le m;\\
&(3) \quad C \text{ is } m\text{-increasing.}
\end{aligned}$$

The theorem suggests that a copula function represents a multivariate distribution as the marginal distributions of this function. Thus, for m-dimensional copula, the joint distribution function is

$$F(x_1,\ldots,x_m) = C(F_1(x_1),\ldots,F_m(x_m);\theta) \tag{2}$$

where
θ - the dependence parameter measuring the dependence between the marginals [11].

The function in (2) is also called a dependence function. If the marginal distributions $F_1(x_1),\ldots,F_m(x_m)$ are continuous, then the copula C is unique; otherwise, it is unique only for $\mathrm{Ran}(F_1) \times \ldots \times \mathrm{Ran}(F_m)$, where Ran stands for *range*, or it is uniquely determined. If marginal distributions are not all continuous, then the copula is not unique, but it still can be expressed as in (2). The dependence then can be modeled without knowing the exact form of the distribution. Therefore, copulas are useful when the form of the joint distribution is overwhelming to identify or unknown [4], [11].

Definition 2. For a m-variate joint cumulative distribution function $F(y_1,\ldots,y_m)$, which has univariate marginal CDFs denoted by $F_1,\ldots,F_m$ in the range $[0,1]$, the joint CDF is bounded by the Fréchet-Hoeffding bounds F_L and F_U:

$$\begin{aligned}
F_L(y_1,\ldots,y_m) &= \max\left[\sum_{j=1}^{m} F_j - m + 1, 0\right]\\
F_U(y_1,\ldots,y_m) &= \min[F_1,\ldots,F_m]
\end{aligned}$$

such that

$$F_L(y_1,\ldots,y_m) \le F(y_1,\ldots,y_m) \le F_U(y_1,\ldots,y_m).$$

The upper bound is a CDF in all cases, and the lower bound is a CDF for $m = 2$. F_U can be a CDF for $m > 2$ when the joint marginal distributions involving $F_1,\ldots,F_m$ are specified [11].

2.4 Types of Copula Functions

This study requires application of different types of copulas due to the dynamics of the financial markets that often cannot be described by a normal distribution. Different copula functions are able to capture different dependence structures. The two major copula families are elliptical and Archimedean copulas [11].

Elliptical copulas are functions derived from the elliptical distribution by using the Sklar's theorem. In most cases, they do not have a closed-form expression, which complicates the copula calculations. An elliptical distribution generalizes the multivariate normal distribution. Due to the invariant property, an elliptical copula's symmetric dependence is defined by the correlation matrix associated with the data about stock returns. This property simplifies the measure of the dependence. Some of the most frequently used elliptical copulas are Gaussian and t-copula functions [4], [7], [11].

Definition 3. A Gaussian copula is represented by

$$C_{\mathrm{Gauss}_\Sigma}(u_1,u_2,\ldots,u_m) = \Phi_\Sigma\left\{\Phi^{-1}(u_1),\Phi^{-1}(u_2),\ldots,\Phi^{-1}(u_m)\right\} \tag{3}$$

where Σ - the correlation matrix with any element in $diag(\Sigma) = 1$,
Φ_Σ - the multivariate Gaussian distribution function,
Φ^{-1} - the inverse standard normal CDF.

Applying Sklar's theorem, the Gaussian copula produces the standard joint normal distribution if and only if its marginal distribution functions are standard normal [11].

Definition 4. A t-copula is represented by

$$C_{\text{t-Student}\Sigma}(u_1,u_2,\ldots,u_m;\nu) = T_\nu\left\{t_\nu^{-1}(u_1),t_\nu^{-1}(u_2),\ldots,t_\nu^{-1}(u_m)\right\} \tag{4}$$

where

ν - the degrees of freedom,

T_ν - the multivariate t-copula distribution function,

t_ν^{-1} - the inverse of the t-copula CDF.

The Archimedean family of copula functions includes Gumbel, Clayton, and Frank copulas. These functions can be expressed in closed form and are constructed using generators [13]. Archimedean copulas can capture asymmetric dependence between the stock returns. The general form of such functions is formulated by

$$C(u_1, u_2, \ldots, u_m) = \psi^{-1}\left(\psi(u_1) + \psi(u_2) + \ldots + \psi(u_m)\right) \tag{5}$$

where

u_m - the marginal uniform distribution function,

ψ - the Archimedean generator.

Theorem 2. The general form of any Archimedean function is a copula if and only if ψ^{-1} is completely monotonic on $[0, \infty)$ [13].

Vice versa, from the property of complete monotonicity, one can arrive at the general form of the Archimedean functions.

Definition 5. The following are the formulas of Gumbel, Clayton, and Frank copulas:

$$C_{\text{Gumbel}}(u_1, \ldots, u_m) = \exp\left(-\left(\sum_{i=1}^{m}(-\log u_i)^\theta\right)^{\frac{1}{\theta}}\right) \qquad \theta > 1. \tag{6}$$

$$C_{\text{Clayton}}(u_1, \ldots, u_m) = \max\left(\sum_{i=1}^{m} u_i^{-\theta} - m + 1, 0\right)^{-\frac{1}{\theta}} \qquad \theta > 0. \tag{7}$$

$$C_{\text{Frank}}(u_1, \ldots u_m) = -\frac{1}{\theta}\log\left(1 + \frac{\prod_{i=1}^{m}\left(e^{-\theta u_i} - 1\right)}{e^{-\theta} - 1}\right) \qquad \theta > 0. \tag{8}$$

where θ - the parameter characterizing tail dependence [11], [13].

The Gumbel copula is able to capture the upper positive tail dependence. In contrast, the Clayton copula is used for left tail dependencies. The Frank copula function can capture dependencies on both ends that could be symmetric or asymmetric [11]. The functions used for this study and their generators are summarized in the Table 1.

Table 1: Formulas for Archimedean Copulas

Generator	5-Dimensional Formula
$\phi(t;\theta) = (-\ln(t))^\theta$	$C_{\text{Gumbel}}(u;\theta) = \exp\left\{-\left[(-\ln u_1)^\theta + (-\ln u_2)^\theta + (-\ln u_3)^\theta + (-\ln u_4)^\theta + (-\ln u_5)^\theta\right]^{\frac{1}{\theta}}\right\}$
$\phi(t;\theta) = \frac{t-\frac{1}{\theta}}{\theta}$	$C_{\text{Clayton}}(u;\theta) = \max\left(u_1^{-\theta} + u_2^{-\theta} + u_3^{-\theta} + u_4^{-\theta} + u_5^{-\theta} - 4, 0\right)^{-\frac{1}{\theta}}$
$\phi(t;\theta) = \ln\left(\frac{e^{-\theta t}-1}{e^{-\theta}-1}\right)$	$C_{\text{Frank}}(u;\theta) = -\frac{1}{\theta}\log\left(1 + \frac{(e^{-\theta u_1}-1)(e^{-\theta u_2}-1)(e^{-\theta u_3}-1)(e^{-\theta u_4}-1)(e^{-\theta u_5}-1)}{e^{-\theta}-1}\right)$

2.5 Pair-Copula Construction

The pair-copula construction allows for the fragmentation of a multivariate copula into $n(n+1)/2$ bivariate copulas [1]. The tree in Figure 1 represents the fragmentation process.

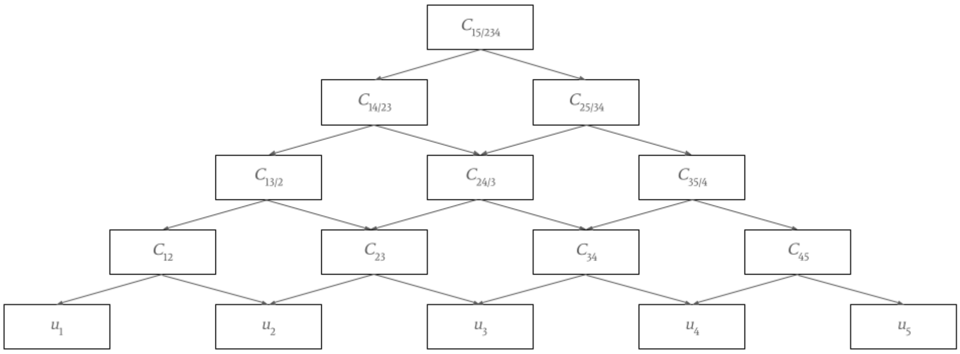

Figure 1: 5-Dimensional Pair-Copula Construction.

The fragmentation in Figure 1 helps to relax a high dimensional equation into several equations amendable to manual solution. The equations below demonstrate that every joint distribution function characterizes both the marginal behavior and the dependence structure [1], [2].

In equation form, the fragmented copula is given by

$$\begin{aligned} C(u_1, u_2, u_3, u_4, u_5) = C(u_1) \cdot C(u_2) \cdot C(u_3) \cdot C(u_4) \cdot C(u_5) \\ \cdot C_{12}(F(u_1), F(u_2)) \cdot C_{23}(F(u_2), F(u_3)) \cdot C_{34}(F(u_3), F(u_4)) \cdot C_{45}(F(u_4), F(u_5)) \\ \cdot C_{13|2}(F(u_1|u_2), F(u_3|u_2)) \cdot C_{24|3}(F(u_2|u_3), F(u_4|u_3)) \cdot C_{35|4}(F(u_3|u_4), F(u_5|u_4)) \\ \cdot C_{14|23}(F(u_1|u_2, u_3), F(u_4|u_2, u_3)) \cdot C_{25|34}(F(u_2|u_3, u_4), F(u_5|u_3, u_4)) \\ \cdot C_{15|234}(F(u_1|u_2, u_4, u_3), F(u_5|u_2, u_4, u_3)) \end{aligned}$$

where the joint distributions are calculated by

$$\begin{aligned} F(u_1|u_2) &= \frac{dC_{12}(u_1, u_2)}{du_2} \\ F(u_3|u_2) &= \frac{dC_{23}(u_2, u_3)}{du_2} \\ F(u_2|u_3) &= \frac{dC_{23}(u_2, u_3)}{du_3} \\ F(u_4|u_3) &= \frac{dC_{34}(u_3, u_4)}{du_3} \\ F(u_1|u_2, u_3) &= \frac{dC_{13|2}(F(u_1|u_2), F(u_3|u_2))}{dF(u_3|u_2)} \\ F(u_4|u_2, u_3) &= \frac{dC_{24|3}(F(u_4|u_3), F(u_2|u_3))}{dF(u_2|u_3)} \\ F(u_2|u_3, u_4) &= \frac{dC_{24|3}(F(u_4|u_3), F(u_2|u_3))}{dF(u_2|u_3)} \\ F(u_1|u_2, u_4, u_3) &= \frac{dC_{14|23}(F(u_2|u_3, u_4), F(u_4|u_3), F(u_1, u_3))}{dF(u_1|u_3)} \\ F(u_5|u_2, u_4, u_3) &= \frac{dC_{25|34}(F(u_2|u_3, u_4), F(u_4|u_3), F(u_5, u_3))}{dF(u_5|u_3)} \end{aligned}$$

2.6 Measures of Dependence

One of the basic measures of dependence is the *Pearson correlation coefficient*, denoted by ρ, that takes values in the interval $[-1, 1]$. It estimates linear dependence assuming normal distribution and absence of outliers in the data. The Pearson correlation can be suitable for elliptical distributions that generalize normal distributions [11], [12], [13]. Its formula is given by

$$\rho(X, Y) = \frac{Cov(X, Y)}{\sigma_X \sigma_Y} \tag{9}$$

where

X, Y - the random variables,

$Cov(X, Y)$ - the covariance of X and Y,

σ - the standard deviation of X and Y.

There are also scale-invariant measures implying that they are unchanged for strictly increasing random variables. Therefore, these measures can be expressed in the form of copula functions of random variables [7], [11]. This project applies two scale-invariant measures of dependence: Kendall's tau and Spearman's rho. $X_1, X_2, \ldots$ and $Y_1, Y_2, \ldots$ used in these measures of dependence are vectors that form a joint distribution as described by the pair-copula construction.

The *Kendall's tau* represents concordance between vectors of random variables. Let (x_i, y_i) and (x_j, y_j) be vectors of (X, Y). Then, these vectors are concordant if $(x_i - x_j)(y_i - y_j) > 0$ and discordant if $(x_i - x_j)(y_i - y_j) < 0$ [7], [11]. The value of Kendall's tau τ gives the difference between the probability of concordance and the probability of discordance:

$$\tau = \tau_{X,Y} = P\left[(X_1 - X_2)(Y_1 - Y_2) > 0\right] - P\left[(X_1 - X_2)(Y_1 - Y_2) < 0\right] \tag{10}$$

Theorem 3. Suppose (X_1, Y_1) and (X_2, Y_2) are two independent vectors of continuous random variables X and Y that have joint distribution functions and C_1 and C_2 are copulas of these vectors, respectively. With joint distributions given by $C_1(F_1(x), F_2(y))$ and $C_2(F_1(x), F_2(y))$, where F_1 and F_2 are marginal distributions, Kendall's tau is represented by

$$\tau_{C_1,C_2} = 4 \int_0^1 \int_0^1 C_2(u_1, u_2) dC_1(u_1, u_2) - 1 \tag{11}$$

The Spearman's rho for independent random vectors (X_1, Y_1), (X_2, Y_2), and (X_3, Y_3) is represented by

$$\rho = \rho_{X,Y} = 3(P\left[(X_1 - X_2)(Y_1 - Y_3) > 0\right] - P\left[(X_1 - X_2)(Y_1 - Y_3) < 0\right]) \tag{12}$$

It is also given by

$$\rho_C = 12 \int_0^1 \int_0^1 C(u_1, u_2) du_1 du_2 - 3 \tag{13}$$

where

C - the copula of continuous random variables X and Y [4], [7], [11].

In addition to these measures of dependence, it is essential to apply *tail dependence* that characterizes simultaneous drastic market dynamics of multiple assets. Lower tail dependence associates such movements with extremely low values, while upper tail dependence symbolizes extremely high values [7], [12]. The formulas for lower and upper tail dependencies are as follows:

$$\lambda_L = \lim_{t \to 0^+} \frac{C(t,t)}{1-t} \tag{14}$$

$$\lambda_U = 2 - \lim_{t \to 1^-} \frac{1 - C(t,t)}{1-t} \tag{15}$$

2.7 Time Series Analysis

Definition 6. A *time series model* is a collection of random variables $(X_t)_{t \in \mathbb{Z}}$ indexed by integers and defined on a probability space $(\Omega, \mathcal{F}, P)$, where Ω is a sample space, $\mathcal{F}$ is a σ-algebra, and P is a probability measure [9].

The concept of stationarity is significant in the time series analysis. A stationary time series possesses properties that do not depend on time at which the time series was produced. There are two common types of time series considered in this study: strictly and weakly stationary. These types observe that a time series behaves similarly in any epoch, meaning that changes in the mean, variance or covariance contradict stationarity [9].

Definition 7. The time series $(X_t)_{t \in \mathbb{Z}}$ is *strictly stationary* if, for all $t_1, \ldots, t_n, k \in \mathbb{Z}$, and $\forall n \in \mathbb{N}$,

$$(X_{t_1}, \ldots, X_{t_n}) \stackrel{\mathrm{d}}{=} (X_{t_1+1}, \ldots, X_{t_n+k}). \tag{16}$$

Definition 8. The time series $(X_t)_{t\in\mathbb{Z}}$ is *weakly stationary* if the first two moments exist and satisfy

$$\mu(t) = \mu, \qquad t \in \mathbb{Z} \tag{17}$$

$$\gamma(t,s) = \gamma(t+k, s+k), \quad t,s,k \in \mathbb{Z}. \tag{18}$$

where
$\mu(t)$ - the mean of the time series at time t,
$\gamma(t,s)$ - the covariance of t and s [9].
From Equation (18), one can deduce

$$\gamma(t-s,0) = \gamma(t,s) = \gamma(s,t) = \gamma(s-t,0)$$

which means that the covariance between X_t and X_s depends only on the absolute difference $|x-t|$ called *lag* [9]. Therefore, the autocorrelation function of a weakly stationary time series model can be expressed as

$$\gamma(h) := \gamma(h,0), \quad \forall h \in \mathbb{Z} \tag{19}$$

where the function contains only one variable.

Definition 9. The *autocorrelation function* (ACF) $\rho(h)$ of a weakly time series model is

$$\rho(h) = \rho(X_h, X_0) = \frac{\gamma(h)}{\gamma(0)}, \quad \forall h \in \mathbb{Z}. \tag{20}$$

ACF defines the similarity between an observation at one time and an observation at some previous time. In the formula above, the function is calculated at lag h. The y-axis has a domain $[-1,1]$ for the correlation coefficients, and the x-axis takes the lag values. The *partial autocorrelation function* (PACF) at a lag k allows to avoid the impact of shorter lags on the correlation [7], [9]. The Section 3 includes the autocorrelation plots resulting from the ACF and PACF calculations. To determine whether the financial data is stationary or non-stationary, the *Augmented Dickey-Fuller test* (ADF), with a *unit root* as a characteristic for non-stationarity and the following two hypotheses was used:
H_0 - the unit root is present in the time series. The data is non-stationary and time dependent.
H_1 - the unit root is absent in the time series. The data is stationary and time independent [9].

Definition 10. $(X_t)_{t\in\mathbb{Z}}$ is a *white noise* time series model, also denoted by $(\varepsilon_t)_{t\in\mathbb{Z}}$, if it is weakly stationary and has the autocorrelation function

$$\rho(h) = \begin{cases} 1, & \text{if } h = 0, \\ 0, & \text{if } h \neq 0. \end{cases} \tag{21}$$

At this stage, the ARIMA and GARCH models can be defined to provide the means for returns analysis. Given that the input financial data used in this study is the closing prices of each stock in the portfolio, the return series of each company needed to be analyzed for stationarity and volatility.

Definition 11. Suppose $(\varepsilon_t)_{t\in\mathbb{Z}}$ is a mean zero white noise with variance σ_ε^2. Then, $(X_t)_{t\in\mathbb{Z}}$ is a zero mean *Autoregressive Moving Average* ARMA(p,q) with p as the number of lag observations in the model (lag order), q as the volume of the moving average (order of moving average), and weakly stationarity expressed as

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}, \quad \forall t \in \mathbb{Z} \tag{22}$$

where,
X_t - the data used for the ARMA model,
ϕ_i - the parameters as AR coefficients,
θ_i - the parameters as MA coefficients [9].
For the *Autoregressive Integrated Moving Average* ARIMA(p,d,q), d is the degree of differencing of raw observations. It is the ARMA(p,q) differenced d times. Let ∇ be the *difference operator* such that $\nabla X_t = X_t - X_{t-1}$. After differencing d times, the time series becomes

$$\nabla^d X_t = \begin{cases} \nabla X_t, & \text{if } d = 1, \\ \nabla^{d-1}(\nabla X_t) = \nabla^{d-1}(X_t - X_{t-1}), & \text{if } d > 1. \end{cases} \tag{23}$$

For $d > 1$, ARIMA models are non-stationary, and they are useful to fit the non-stationary data into a stationary ARMA model [9]. In this project, the daily log returns of the stocks suggest a non-stationary ARIMA$(p, 1, q)$ model, but, in fact, the ARMA(p, q) model will be followed.

Definition 12. For a mean zero strict white noise, where there is a time series with independent and identically distributed random variables with variance 1, $(X_t)_{t\in\mathbb{Z}}$ is a *Generalized Autoregressive Conditional Heteroskedasticity* GARCH(p, q) given it is strictly stationary and

$$X_t = \sigma_t Z_t, \tag{24}$$

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i X_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2, \tag{25}$$

where,

Z_t - the strict white noise,

$(\sigma_t)_{t\in\mathbb{Z}}$ - the positive valued time series model,

$\alpha_0 > 0$, $\alpha_i \geq 0$ for $i = 1, \ldots, p$,

$\beta \geq 0$ for $j = 1, \ldots, q$ [9].

Only GARCH$(1, 1)$ models, for which $p, q = 1$, will be used in this study. It is the simplest yet widely used GARCH model to capture volatility of a time series data [9].

To select appropriate ARIMA and GARCH models, they can be compared to the real data distribution. This project uses the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) to evaluate the inadequacy of a model. Both criteria punish the model for excessive parameters and reward for high log-likelihood values, the BIC being stricter with comparatively larger sample sizes. The smaller the positive values or the larger the negative values of AIC and BIC, the more suitable is the model for time series analysis [7], [9].

$$\text{AIC} = -2\ln(\boldsymbol{\theta}_{\text{obs}}) + 2p \tag{26}$$

$$\text{BIC} = -2\ln(\boldsymbol{\theta}_{\text{obs}}) + p\ln n \tag{27}$$

where,

$\boldsymbol{\theta}_{\text{obs}}$ - the maximum log-likelihood estimator,

p - the number of parameters,

n - the sample size.

3 Applications and Results

3.1 Stock Selection

To select the companies of the portfolio, the correlations between each of the 30 largest companies by market capitalization from the Standard & Poor's 500 index were calculated based on all closing prices between 2019-2024. The data about market capitalization of the companies were borrowed from Yahoo Finance. Market capitalization is the number of common shares of a company outstanding times the share price. On average, the least correlated stocks were Meta (META), Amazon (AMZN), Bank of America (BAC), Salesforce (CRM), and Exxon Mobil (XOM). Their correlation coefficients are given in Table 2. This method was used to select the largest companies that follow least similar market dynamics, and therefore enhance diversification.

The weights of the stocks in the portfolio were estimated using the Markowitz portfolio optimization [8], [14]. The floor for a single weight was set to be 5% so that the portfolio is not excessively formalized. The risk that was considered in the Markowitz model is purely dependent on the past numerical data; meanwhile, financial risk also comprises of currency, interest rate, legal, political, and other factors that are not immediately priced in the share on a particular time interval. The weights of each company in the portfolio are given in Table 2. It can be seen that it abundantly favors CRM because its share price was less volatile between 2019-2024 than other chosen stocks (see Figure 2).

The stock selection process in this project reflects a scientific approach rather than pragmatism that is common in asset management. The goal is to estimate the risk associated not only with the returns, variance, or dependence structures, but also with arguably random selection of investments. The main reason why only 5 companies were chosen for the study is that as the dimensions of a copula increase,

Ticker	Correlation	Weight in Portfolio
META	0.479963	5.00%
AMZN	0.507493	5.00%
BAC	0.509304	5.00%
CRM	0.513795	70.83%
XOM	0.552480	14.17%

Table 2: Correlation Coefficients and Markowitz Portfolio Optimization

the estimations become less accurate; therefore, calculating VaR for a well-diversified portfolio demands more sophisticated mathematical approaches.

Before proceeding with modeling the financial data, fundamental statistics for each asset need to be measured. The skewness indicates how different the data are from a symmetrical distribution, or how different are the mean, median, and mode from each other. Negative skewness for META implies that its left tail is fatter than the opposite tail and that the company oversaw insignificant positive returns and drastic losses on the stock market. High kurtosis values for all stocks indicate the non-normality of the log returns of the stocks. That is, fat tails prevail in the distributions, and there is a relatively higher probability of drastic price changes. The visuals for the log returns are provided in Figure 2. The near zero p-values for the Shapiro-Wilk normality test also suggest that the data follows a distribution considerably different from the normal [3]. Therefore, we fail to reject the H_0.

Ticker	Mean	Variance	Skewness	Kurtosis	Shapiro-Wilk test results
META	0.001147	0.000755	-0.547751	18.311363	$< 2.2 \times 10^{-16}$
AMZN	0.000787	0.000492	0.090031	4.251938	$< 2.2 \times 10^{-16}$
BAC	0.000588	0.000509	0.421410	10.325752	$< 2.2 \times 10^{-16}$
CRM	0.000818	0.000586	0.780385	13.799002	$< 2.2 \times 10^{-16}$
XOM	0.000724	0.000466	0.058826	4.660252	$< 2.2 \times 10^{-16}$

Table 3: Stocks Statistics

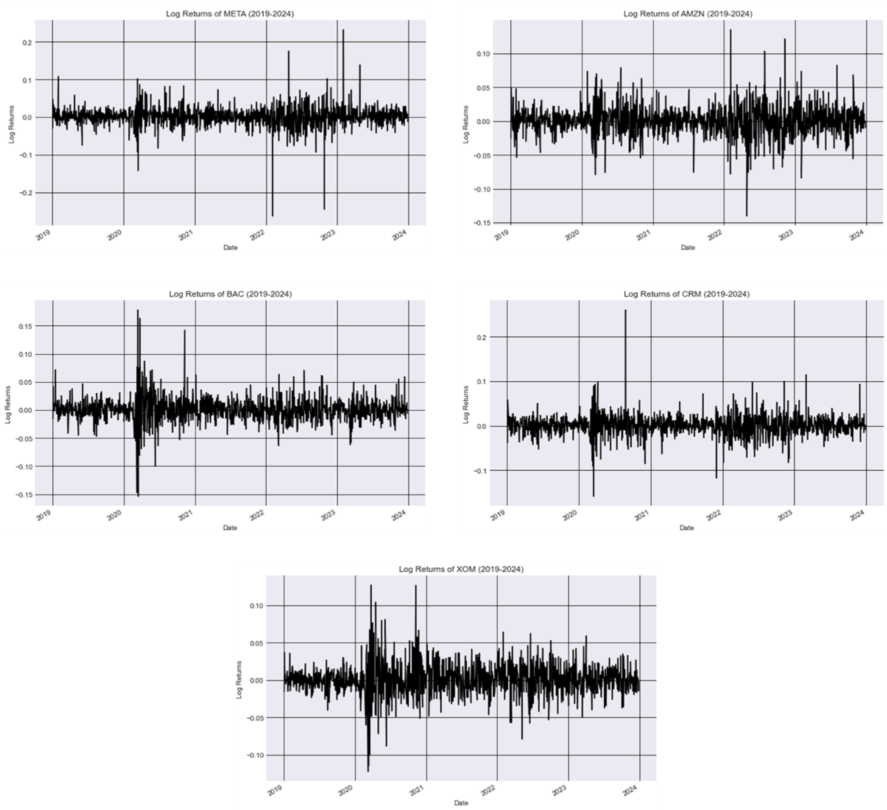

Figure 2: Log Returns of the Selected Companies

3.2 Value at Risk from Historical Data

The copula estimations of VaR need to be compared to the metrics from the real-world data to validate the approach of this study. First, to calculate the VaR from the historicals, the closing prices from the Yahoo Finance for each stock for the given period were parsed. Second, the covariance matrix was calculated using the log returns. It is then decomposed into two matrices, one being the lower triangular matrix and the other its transpose, the upper triangular matrix. Their product $\Sigma = LL^T$ is known as the Cholesky decomposition [6]:

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} & \sigma_{15} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} & \sigma_{24} & \sigma_{25} \\ \sigma_{13} & \sigma_{23} & \sigma_{33} & \sigma_{34} & \sigma_{35} \\ \sigma_{14} & \sigma_{24} & \sigma_{34} & \sigma_{44} & \sigma_{45} \\ \sigma_{15} & \sigma_{25} & \sigma_{35} & \sigma_{45} & \sigma_{55} \end{bmatrix} = \begin{bmatrix} l_{11} & 0 & 0 & 0 & 0 \\ l_{21} & l_{22} & 0 & 0 & 0 \\ l_{31} & l_{32} & l_{33} & 0 & 0 \\ l_{41} & l_{42} & l_{43} & l_{44} & 0 \\ l_{51} & l_{52} & l_{53} & l_{54} & l_{55} \end{bmatrix} \begin{bmatrix} l_{11} & l_{21} & l_{31} & l_{41} & l_{51} \\ 0 & l_{22} & l_{32} & l_{42} & l_{52} \\ 0 & 0 & l_{33} & l_{43} & l_{53} \\ 0 & 0 & 0 & l_{44} & l_{54} \\ 0 & 0 & 0 & 0 & l_{55} \end{bmatrix}$$

Finally, before the historical and normal VaR calculations, the portfolio returns are expressed as

$$R = \sum_{i=1}^{n} w_i R_i \tag{28}$$

The Python code that output the values for historical, normal distribution, and Monte Carlo VaRs with 100,000 simulations is given below. For the latter, we choose random variables from a standard normal distribution to calculate the returns. The results are provided in Table 4.

```
historical_var = np.percentile(portfolio_returns, 100 * (1 - np.array(confidence_levels)))
normal_var = norm.ppf(1 - np.array(confidence_levels), loc=np.mean(portfolio_returns), scale=np.std(portfolio_returns))

np.random.seed(42)
simulations = 100000

monte_carlo_returns = np.dot(np.random.normal(0, 1, size=(simulations, len(stocks))), cholesky_matrix.T)
monte_carlo_portfolio = np.dot(monte_carlo_returns, weights)
monte_carlo_var = np.percentile(monte_carlo_portfolio, 100 * (1 - np.array(confidence_levels)))
```

Confidence Level	Historical	Normal Distribution	Monte Carlo
90%	0.180747	0.210086	0.217826
99%	0.463261	0.387987	0.395013
99.5%	0.542217	0.430467	0.435532

Table 4: VaR Results for the Portfolio

3.3 Determining Marginal Distributions

Figures 3, 4, 5, 6, and 7 provide insights into autocorrelation of the time series based on the log returns. Significant spikes in PACF of all companies signifies autocorrelation at the indicated lags. Somewhat gradually decaying ACF's and the corresponding chaotic behavior in PACF's for BAC and XOM implies a relatively high autocorrelation common for the autoregressive processes. The appropriate ARIMA and GARCH models, supported by the AIC and Ljung-Box test with a lag of 20 that checks for nonzero values in the autocorrelations, are provided in Table 5.

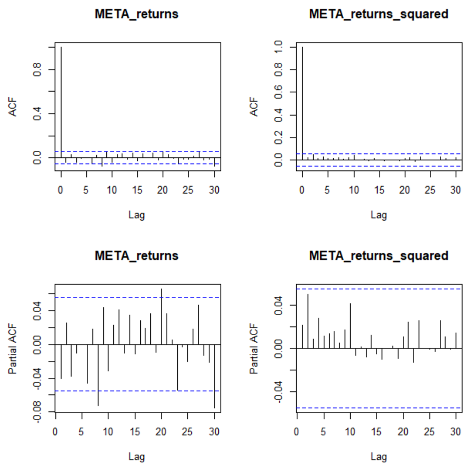

Figure 3: ACF and PACF Plots for META

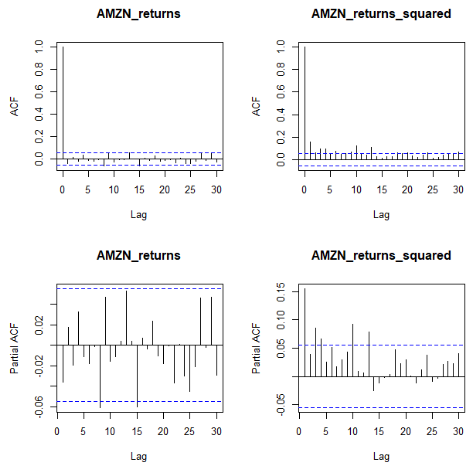

Figure 4: ACF and PACF Plots for AMZN

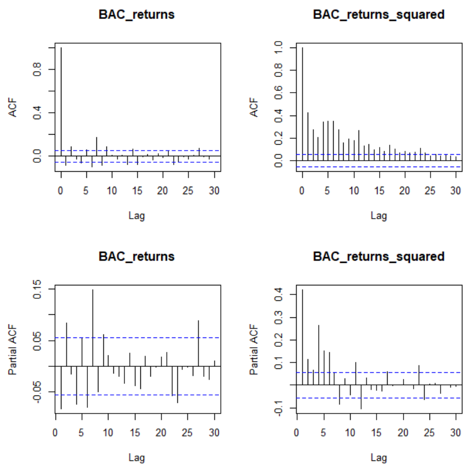

Figure 5: ACF and PACF Plots for BAC

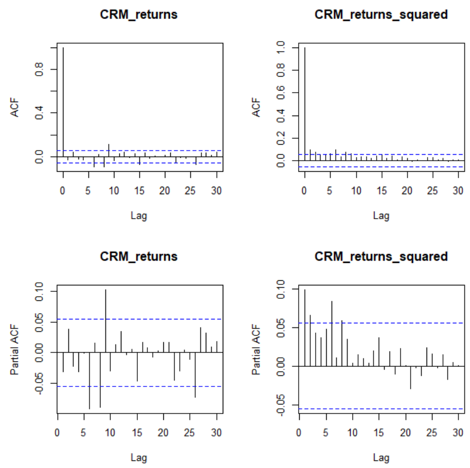

Figure 6: ACF and PACF Plots for CRM

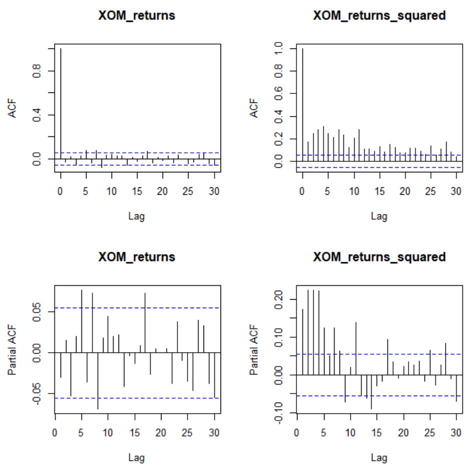

Figure 7: ACF and PACF Plots for XOM

Combining the non-normality of the distribution, non-stationarity from the ADF, and the autocorrelation in the data, the appropriate ARIMA and GARCH models were selected and tested. The results from the Ljung-Box test in Table 5 suggest that the standardized residuals for all models are serially independent. They are then remodeled into pseudo-observations (Figure 8) to determine the parameters of the copulas.

Ticker	ARIMA(p, d, q)	GARCH(p, q)	Ljung-Box test results
AMZN	(2,0,4)	(1,1)	0.9151
META	(4,0,4)	(1,1)	0.5501
BAC	(4,0,4)	(1,1)	0.8990
CRM	(3,0,4)	(1,1)	0.7636
XOM	(4,0,4)	(1,1)	0.3329

Table 5: ARIMA and GARCH Models Selected by AIC

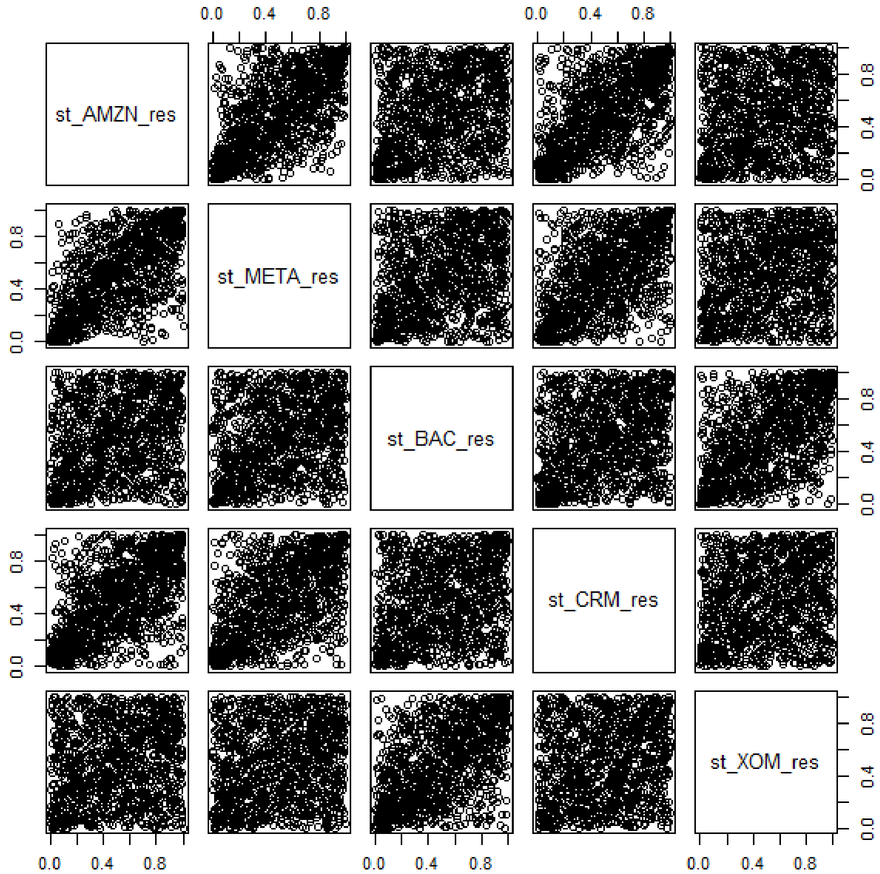

Figure 8: Pseudo-Observations

3.4 Copula VaR Estimations

The parameters given in Table 6 represent the dependence structure among the log returns of all stocks. For the Gaussian and t-copula, parameter values closer to 1 indicate strong positive linear correlation. For the portfolio of this project, this correlation is relatively weak. The parameter of the Clayton copula represents the degree of lower tail dependence, which captures significant losses. The parameter of 0.46689 implies that such losses do not correlate well for the selected companies; however, the market downturn due to the pandemic in 2020 likely contributed to the value. In contrast, the Gumbel copula parameter studies upper tail dependence, and the Frank copula parameter grasps the general dependence [11].

The log-likelihood measures the likelihood of the data fitting in a distribution for a particular copula. The highest value was observed for the t-Copula, which, in comparison to the Gaussian, also controls the volume of the tails in the distribution. However, the best goodness-of-fit test p-value was observed for the Clayton copula. The AIC and BIC values correlate with the mentioned metrics quite well.

Copula	Parameters	Log-Likelihood	Goodness-of-Fit test results	AIC	BIC
Gaussian	0.36402	561.9	0.04545	-1121.735	-1116.600
Clayton	0.46689	543.6	0.08333	-1085.231	-1080.095
t-Copula	0.38149	656.6	0.04845	-1309.185	-1298.914
Gumbel	1.28150	457.6	0.03846	-913.2029	-908.067
Frank	2.23970	504.1	0.03125	-1006.229	-1001.093

Table 6: Copula Model Results

In Table 7, it can be seen that VaR estimations closest to historical observations were produced by the Clayton copula. Noticeable deviation of the Gumbel and Frank copulas may suggest that the upper tail and general dependence are unsuitable for the VaR calculations, which is reasonable considering that these are estimations of potential losses. The R code used for the calculations is based on Imanbayeva (2021).

Method	90%	99%	99.5%
Historical	**0.1807**	**0.4633**	**0.5422**
Gaussian	0.2297	0.3450	0.3640
Clayton	**0.2709**	**0.3914**	**0.4669**
t-Copula	0.2275	0.3675	0.3815
Gumbel	1.1150	1.2700	1.2810
Frank	1.826	2.1710	2.2400

Table 7: VaR Estimations at Given Confidence Levels

4 Conclusion

This Capstone project aimed at estimating the VaR of a portfolio whose stocks were weighted based on the strategy of minimizing risks for the respective returns of each asset. The period for VaR was chosen to be 2019-2024. 5-dimensional copulas were used to compare the results with historical values, normal, and Monte Carlo approximations and to determine the estimations closest to the past values.

Copula functions were used to calculate VaR by combining marginal distributions of the investments. ARIMA and GARCH models were incorporated to determine the stationary and volatility of the data to determine these marginal distributions. Shapiro-Wilk, Ljung-Box, and goodness-of-fit tests, AIC, and BIC were utilized to check the suitability of the selected methods and models. As a result, it was found that Clayton copulas, the functions that better capture the left tail of a distribution, provided the most appropriate estimations.

For future research, more than 5 companies with diverse market dynamics can be selected for the portfolio to discover how Clayton copulas will estimate the VaR and if there will be a better approximation provided by other copula functions. In addition, a well-diversified portfolio, in combination with suitable copulas and techniques that will provide accurate estimations, can be studied to reflect the real-world asset and risk management strategies.

5 Acknowledgements

I want to use the chance to express my sincere gratitude to my Capstone supervisor, Professor Dongming Wei, for offering me the opportunity to work on a topic I am passionate about and for providing assistance throughout this semester. I also want to thank Professor Francesco Rocciolo, who kindly agreed to be my second reader, for his valuable opinion and enthusiasm about the project. Finally, I am grateful for Professor Tolga Etgu's involvement and feedback during the meetings and to my friends who attended my Capstone defense on April 19, 2024.

6 References

[1] Aas, K. et al. (2009) 'Pair-copula constructions of multiple dependence', *Insurance: Mathematics and Economics,* 44(2), pp. 182–198. doi:10.1016/j.insmatheco.2007.02.001.

[2] Aas, K. (2016) "Pair-copula constructions for financial applications: A Review," *Econometrics*, vol. 4, no. 4, p. 43. doi:10.3390/econometrics4040043

[3] Alexander, C. (2010) *Quantitative Methods in Finance.* Chichester: Wiley.

[4] Benzi, S. (2024) "Introduction to Copulas in Finance. Financial Econometrics Course."

[5] Imanbayeva, S. (2021) "Application of Copulas in the Portfolio Risk Management of Kazakhstan's Mining Companies," Capstone project, Nazarbayev University.

[6] Moriya, B. (2023) "Markowitz portfolio optimization," Kaggle, https://www.kaggle.com/code/bhavinmoriya/markowitz-portfolio-optimization (accessed Apr. 17, 2024).

[7] Joe, H. (2022) *Dependence Modeling with Copulas.* Boca Raton: CRC Press.

[8] Markowitz, H. (1952) "Portfolio Selection," *The Journal of Finance*, vol. 7, no. 1, pp. 77–91.

[9] McNeil, A. J., Frey, R., and Embrechts, P. (2015) *Quantitative Risk Management: Concepts, Techniques and Tools.* Princeton, NJ: Princeton University Press.

[10] Miller, M. B. (2019) "Market Risk: Value at Risk," in *Quantitative Financial Risk Management,* Hoboken, New Jersey: Wiley Finance Series, pp. 51–71.

[11] Nelsen, R. B. (2006) *An introduction to copulas,* 2nd ed. Springer.

[12] Sampid, M. G. and Hasim, H. M. (2018) "Estimating value-at-risk using a multivariate copula-based volatility model: Evidence from European Banks," *International Economics,* vol. 156, pp. 175–192. doi:10.1016/j.inteco.2018.03.001

[13] Trivedi, P. K. and Zimmer, D. M. (2007) *Copula Modeling: An Introduction for Practitioners.* Boston - Delft: Foundations and Trends in Econometrics.

[14] Vanderbei, R. J. (2020) *Linear Programming: Foundations and Extensions.* Cham: Springer International Publishing.

7 R Code

```
# Load required libraries
library("readxl")
library("e1071")
library("car")
library("tseries")
library("copula")
library("VineCopula")
library("aTSA")
library("forecast")
library("rugarch")
library("gofCopula")
library("corrplot")
library("RColorBrewer")
library("univariateML")
library("MASS")
library(ggplot2)
```

```
library(tidyr)
library(dplyr)

# Read portfolio data from Excel
data = read_excel("C:/Users/User/Downloads/portfolio_data.xlsx")

# Calculate mean of each column
sapply(data, mean, na.rm=TRUE)

# Calculate log returns for each stock
logAMZN = diff(log(data$ClosingPrice_AMZN), lag=1)
logMETA = diff(log(data$ClosingPrice_META), lag=1)
logBAC = diff(log(data$ClosingPrice_BAC), lag=1)
logCRM = diff(log(data$ClosingPrice_CRM), lag=1)
logXOM = diff(log(data$ClosingPrice_XOM), lag=1)

# Perform Shapiro-Wilk tests for normality
shapiro.test(logAMZN)
shapiro.test(logMETA)
shapiro.test(logBAC)
shapiro.test(logCRM)
shapiro.test(logXOM)

# Plot ACF and PACF for each stock
par(mfrow=c(2,2))
acf(logAMZN, main="AMZN_returns")
acf(logAMZN^2, main="AMZN_returns_squared")
pacf(logAMZN, main="AMZN_returns")
pacf(logAMZN^2, main="AMZN_returns_squared")

# Perform Ljung-Box test for autocorrelation
Box.test(logAMZN, lag=20, type=c("Ljung-Box"))
Box.test(logMETA, lag=20, type=c("Ljung-Box"))
Box.test(logBAC, lag=20, type=c("Ljung-Box"))
Box.test(logCRM, lag=20, type=c("Ljung-Box"))
Box.test(logXOM, lag=20, type=c("Ljung-Box"))

# Perform ADF test for stationarity
adf.test(logAMZN)
adf.test(logMETA)
adf.test(logBAC)
adf.test(logCRM)
adf.test(logXOM)

# Fit ARIMA models and select best orders based on AIC
META_f.aic = Inf
META_f.order = c(0,0,0)
for (p in 1:4) for (d in 0:2) for (q in 1:4) {
  META_curr.aic = AIC(arima(META, order=c(p,d,q), optim.control=list(maxit=1000)))
  if (META_curr.aic < META_f.aic) {
    META_f.aic = META_curr.aic
    META_f.order = c(p,d,q)
    META_f.arima = arima(META, order=META_f.order)
  }
}

# Plot ACF and PACF of ARIMA residuals
par(mfrow=c(2,2))
```

```
acf(resid(META_f.arima))
acf(resid(META_f.arima)^2)
pacf(resid(META_f.arima))
pacf(resid(META_f.arima)^2)

# Fit GARCH models
META_garch = garch(META, trace=FALSE)
summary(META_garch)

# Plot ACF and PACF of GARCH residuals
par(mfrow=c(2,2))
acf(META_res)
acf(META_res^2)
pacf(META_res)
pacf(META_res^2)

# Standardize residuals and calculate mean and standard deviation
sA = sd(META_res)
st_META_res = META_res/sA
mean(st_META_res)
sd(st_META_res)

# Similar steps for other tickers
...

# Fit copula models and find VaR estimations
fitcop_norm = fitCopula(normalCopula(dim=5), data=pseudo, method="mpl")
summary(fitcop_norm)
tau(normalCopula(param=0.36402))

fitcop_cl = fitCopula(claytonCopula(dim=5), data=pseudo, method="mpl")
summary(fitcop_cl)
tau(claytonCopula(param=0.46689))

fitcop_t = fitCopula(tCopula(dim=5), data=pseudo, method="mpl")
summary(fitcop_t)
tau(tCopula(param=0.38149))

fitcop_g = fitCopula(gumbelCopula(dim=5), data=pseudo, method="mpl")
summary(fitcop_g)
tau(gumbelCopula(param=1.28150))

fitcop_frank = fitCopula(frankCopula(dim=5), data=pseudo, method="mpl")
summary(fitcop_frank)
tau(frankCopula(param=2.23970))

# Calculate log-likelihood, AIC, and BIC of fitted copula models
cat("Log-likelihood_of_Gaussian_copula:", logLik(fitcop_norm), "\n")
cat("AIC_of_Gaussian_copula:", AIC(fitcop_norm), "\n")
cat("BIC_of_Gaussian_copula:", BIC(fitcop_norm), "\n")

# Goodness-of-fit test for copula models
gofCopula(normalCopula(param=0.36402, dim=5), as.matrix(mg), N=10, method="Sn")
gofCopula(claytonCopula(param=0.46689, dim=5), as.matrix(mg), N=5, method="Sn")
gofCopula(tCopula(param=0.38149, dim=5), as.matrix(mg), N=12, method="SnB")
gofCopula(gumbelCopula(param=1.28150, dim=5), as.matrix(mg), N=12, method="Sn")
gofCopula(frankCopula(param=2.23970, dim=5), as.matrix(mg), N=15, method="Sn")
```

```
# Fit marginal distributions
mod_Gauss_META = fitdistr(st_META_res, "normal")
mod_Gauss_AMZN = fitdistr(st_AMZN_res, "normal")
mod_Gauss_BAC = fitdistr(st_BAC_res, "normal")
mod_Gauss_CRM = fitdistr(st_CRM_res, "normal")
mod_Gauss_XOM = fitdistr(st_XOM_res, "normal")
mod_t_META = fitdistr(st_META_res, "t")
mod_t_AMZN = fitdistr(st_AMZN_res, "t")
mod_t_BAC = fitdistr(st_BAC_res, "t")
mod_t_CRM = fitdistr(st_CRM_res, "t")
mod_t_XOM = fitdistr(st_XOM_res, "t")

# Calculate AIC and BIC for marginal distributions
cat("AIC_of_normal_distribution_of_META:", AIC(mod_Gauss_META), "\n")
cat("BIC_of_normal_distribution_of_META:", BIC(mod_Gauss_META), "\n")
cat("AIC_of_t-student_marginal_distribution_of_META:", AIC(mod_t_META), "\n")
cat("BIC_of_t-student_marginal_distribution_of_META:", BIC(mod_t_META), "\n")

# Similar steps for other tickers
...
```