

Exploratory Data Analysis(EDA) - S&P 500 Returns

This R notebook conducts an observational study on a sample of S&P 500 returns from 1/1/2000 to 12/31/2025. The general purpose of this notebook is to gain familiarity with the syntax of R and showcase EDA relating to financial time-series data. I also provide commentary on the observations made throughout the notebook. The HTML was rendered from the corresponding R notebook script.

Given a daily price series P_t , the return series R_t is

$$R_t = \frac{P_t - P_{t-1}}{P_{t-1}}$$

```
# Load libraries
library(quantmod)

## Loading required package: xts
## Loading required package: zoo
##
## Attaching package: 'zoo'
## The following objects are masked from 'package:base':
##
##   as.Date, as.Date.numeric
## Loading required package: TTR
## Registered S3 method overwritten by 'quantmod':
##   method      from
##   as.zoo.data.frame zoo
library(moments)
library(Cairo)

# Download data
ohlcv = getSymbols("SPY", src="yahoo", from="2000-01-01", to="2025-12-31", auto.assign=FALSE)

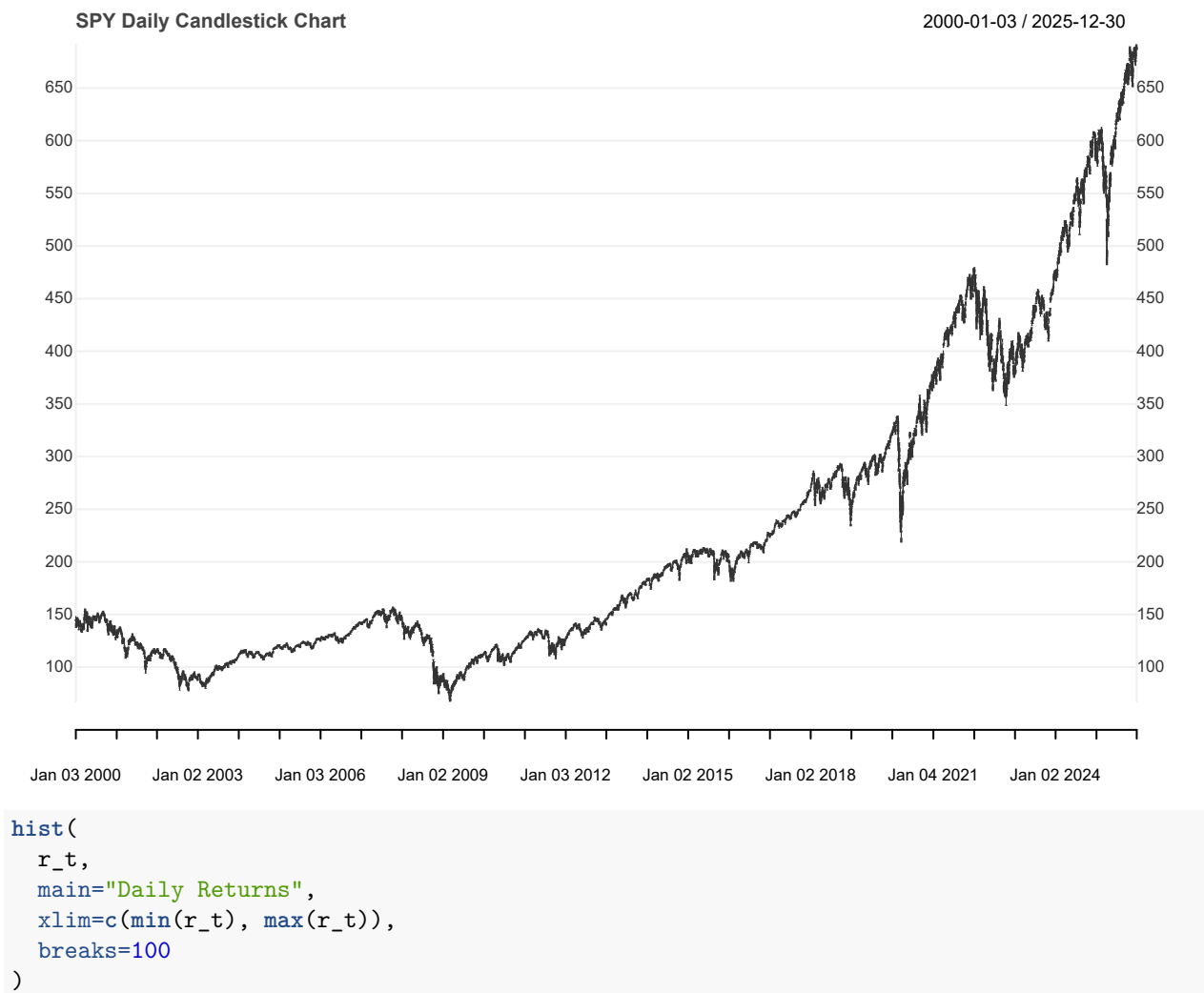
# Compute returns
r_t = dailyReturn(Cl(ohlcv), type="arithmetic")
mr_t = monthlyReturn(Cl(ohlcv), type="arithmetic")

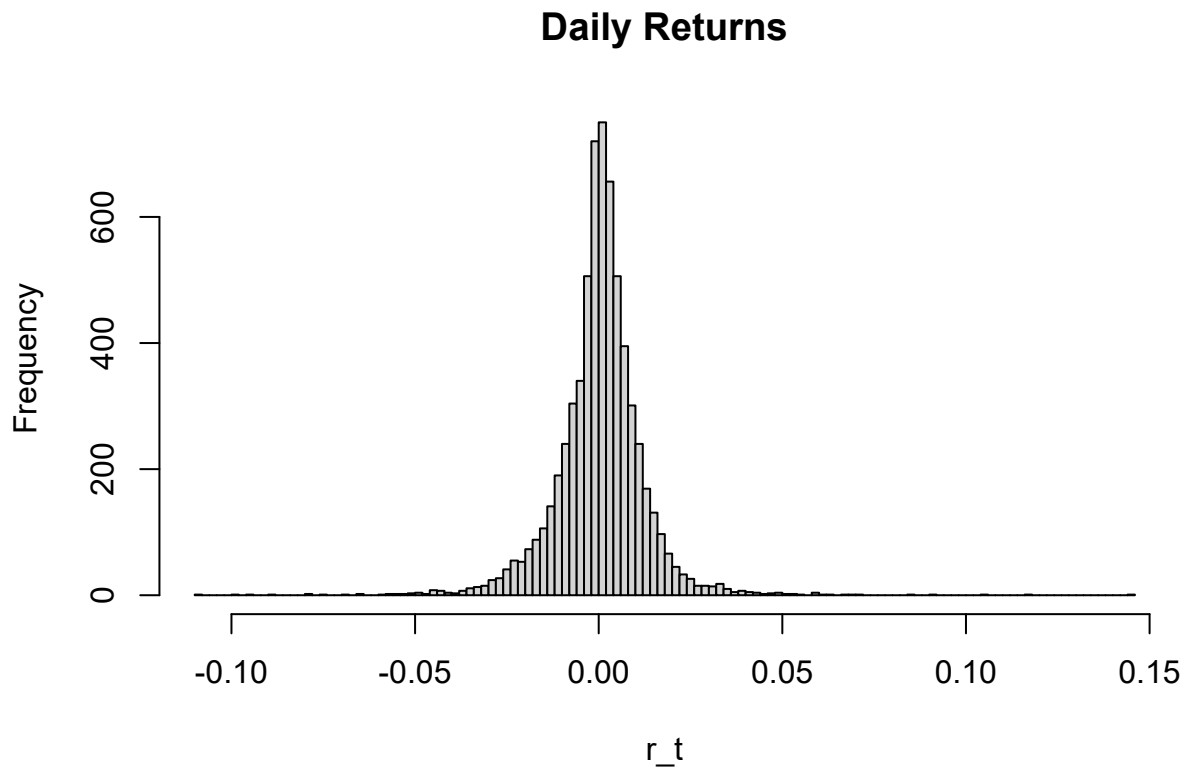
r_t = na.omit(r_t)
mr_t = na.omit(mr_t)
```

Plots

This section explores various types of plots for return and price.

```
# Create a daily candlestick chart
quantmod::chart_Series(ohlcv, name = "SPY Daily Candlestick Chart")
```

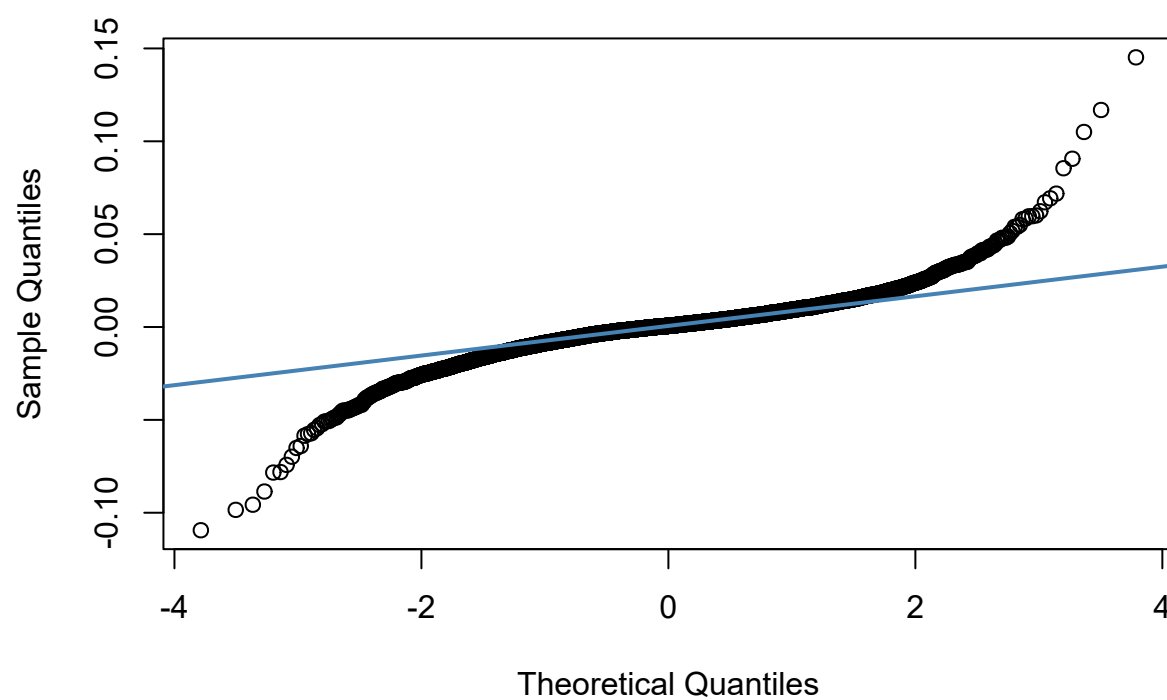




Visually inspecting the daily return distribution, it appears to follow a normal distribution, with fatter tails. Below is the Q-Q plot to further inspect tail behavior.

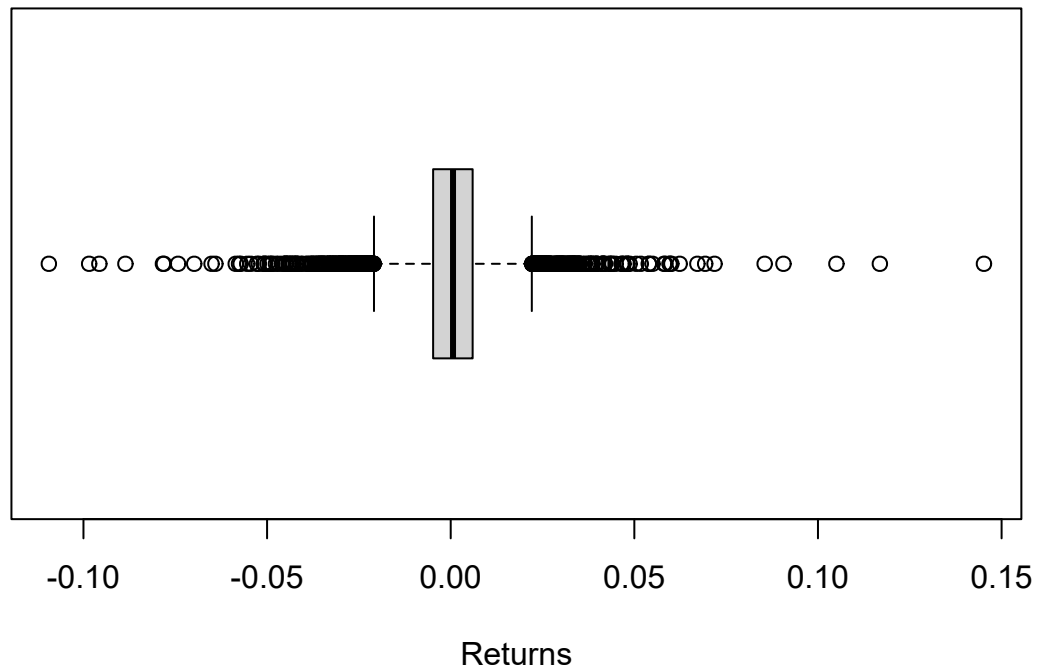
```
# Normal Q-Q plot  
qqnorm(r_t, main="Normal Q-Q Plot of Returns")  
qqline(r_t, col="steelblue", lwd=2)
```

Normal Q-Q Plot of Returns



```
# Box plot of daily returns  
boxplot(  
  r_t,  
  main="Box Plot of Returns",  
  xlab="Returns",  
  horizontal=TRUE  
)
```

Box Plot of Returns



From the box-plot, we can see there are many return outliers, further validating the fat tail behavior of the daily return distribution.

Numerical Summaries

This section computes numerical statistics for return data.

```
# Compute the summary statistics
var_95 = quantile(r_t, 0.05)

statistics = c(
  mean = mean(r_t),
  meadian = median(r_t),
  min = min(r_t),
  max = max(r_t),
  iqr = IQR(r_t),
  sd = sd(r_t),
  skew = skewness(r_t),
  kurt = kurtosis(r_t),
  VaR_95 = var_95,
  ES_95 = mean(r_t[r_t <= var_95])
)
statistics_df = data.frame(
  Statistic=names(statistics),
  Value=round(as.numeric(statistics), 6)
)
```

```
print(statistics_df, row.names=FALSE)
```

```
##           Statistic      Value
##           mean    0.000312
##           meadian  0.000631
##           min    -0.109424
##           max     0.145198
##           iqr     0.010759
##           sd      0.012239
## skew.daily.returns  0.043367
## kurt.daily.returns 14.945455
##           VaR_95.5% -0.019073
##           ES_95    -0.029342
```

The summary statistics give us numerical insights into the sample daily return distribution of the S&P 500.

The mean daily return was ~0.03% and the median daily return was ~0.06%. This suggests that, on average, the sample had a positive daily return.

The maximum one day return in the sample was ~14.5% and the minimum one day return in the sample was negative ~10.9%. There are large daily returns in the sample. The next section explores the frequency of large returns.

The sample return distribution has a slight positive skew. The high kurtosis validates our observations in the plotting section; the return distribution has fat tails.

Based on the historical sample, the estimated probability of a daily return below negative 1.91% is 5%.

Based on the historical sample, the average return of the smallest 5% of returns is negative 2.93%.

Frequency table of positive and negative returns

```
BOUND = 0.025
```

```
frequency_table = rbind(
  data.frame(
    Type="Pos. R_t",
    Count=sum(r_t > 0),
    Probability=mean(r_t > 0)
  ),
  data.frame(
    Type="Neg. R_t",
    Count=sum(r_t < 0),
    Probability=mean(r_t < 0)
  ),
  data.frame(
    Type="Large Pos. (> BOUND)",
    Count=sum(r_t > BOUND),
    Probability=mean(r_t > BOUND)
  ),
  data.frame(
    Type="Large Neg. (< -BOUND)",
    Count=sum(r_t < -BOUND),
    Probability=mean(r_t < -BOUND)
  )
)

print(frequency_table)
```

```
##              Type Count Probability
## 1          Pos. R_t 3536 0.54083818
## 2          Neg. R_t 2980 0.45579688
## 3 Large Pos. (> BOUND) 131 0.02003671
## 4 Large Neg. (< -BOUND) 162 0.02477822
```

We can see that ~54% of the daily returns were positive and ~46% of the daily returns were negative in the sample. We can also observe that ~2% of the daily returns were greater than the bound and ~2.5% of the daily returns were lower than the bound.

```
# Mean day of the week return
dow = weekdays(index(r_t))

aggregate(r_t, by=list(dow), mean)
```

```
##
## Friday      -0.0001947930
## Monday       0.0003603448
## Thursday     0.0002666920
## Tuesday      0.0006587070
## Wednesday    0.0004631526
```

```
month = months(index(mr_t))

aggregate(mr_t, by=list(month), mean)
```

```
##
## April        0.016761798
## August       0.003570860
## December     0.001688576
## February     -0.002420312
## January      0.001154267
## July         0.016598160
## June         -0.003852184
## March        0.007526238
## May          0.008823490
## November     0.024162936
## October      0.013733313
## September    -0.016250244
```

The section above aggregates the mean returns by the day of the week and month.

All days of the week had a positive mean return except Friday. The day with the largest mean return was Tuesday (~0.006%).

February, June, and September had negative mean returns, while the rest of the months had positive mean returns. November had the largest mean return (2.42%).