Part 1 - Data Cleaning

I construct the value-weighted market return using CRSP data, replicating the market return time series available in Kenneth French website. Also, I calculate the equal-weighted market return, and the lagged total market capitalization.

The head and tail of the cleaned data is shown in Table 1. Prior to any calculations, I clean the data as per the Ken French procedure listed on their website. Variables are defined as such:
- *SHRCD* = Share Code
- *EXCHCD* = Exchange Code (1 = NYSE, 2 = AMEX, 3 = NASDAQ)
- *DLRET* = Delisting Return in decimal
- *SHROUT* = Number of publicly held shares in thousands

Table 1. Cleaned data table.

```
> clean_dt
      Year Month Stock_lag_MV  Stock_Ew_Ret Stock_Vw_Ret
   1: 1926     1           NA  0.0240510467   0.00000000
   2: 1926     2     26573.29 -0.0554100480  -0.03338926
   3: 1926     3     25625.48 -0.0995962807  -0.06557717
   4: 1926     4     23564.32  0.0333359454   0.03752979
   5: 1926     5     24857.02  0.0009532739   0.01250893
  ---
1124: 2019     8  29555730.77 -0.0528323729  -0.02406683
1125: 2019     9  28769217.39  0.0217109105   0.01614717
1126: 2019    10  29137508.49  0.0020050146   0.02213428
1127: 2019    11  29667769.79  0.0400275411   0.03986245
1128: 2019    12  30785042.93  0.0531645199   0.02908009
```

1. Universe of stocks: Following Ken French procedure, the sample is filtered to show only the common shares (share codes 10 and 11) as well as to securities traded only on the NYSE, ASE, and NASDAQ (exchange codes 1,2,3, respectively)

2. Missing returns: The cumulative dividend total return (thereafter referred to as return) is calculated with the follow equation where h is holding period return and d is delisting return. In both the holding period returns (HRET) and delisting returns (DLRET), NA's are present. I apply the following set of equations below to calculate the cumulative dividend return. The data set is then subset to show only observations with a value in the RET column.

$$r_{i,t} = \begin{cases} r^h_{i,t} & \text{if } r^d_{i,t} \text{ missing} \\ r^d_{i,t} & \text{if } r^h_{i,t} \text{ missing} \\ (1+r^h_{i,t})(1+r^d_{i,t}) - 1 & \text{if both not missing} \end{cases}$$

3. Delisting return calculation: The value obtained from CRSP is already in decimal. Thus, there is no need to divide by 100. However, I did convert the returns to the numeric type.

4. Market Capitalization calculation: For the negative price per share, the negative signs is an indication of the bid/ask average for a trading month range when the closing price is not available. Compare to the full dataset's return median of 0% and mean of 0.0111%, the return median is 0% and the mean is 0.0096% for negative price per share. Therefore, it is suitable to be included in the analysis. The market capitalization is calculated by multiplying the shares outstanding in thousands (SHROUT) by the price per share (PRC) and divided by 1000 to obtain this in millions. This is further filtered for market caps greater than zero.

5. (Definition) Portfolio weights: For the equal-weighted portfolio, the average of the returns at a certain date is taken. Whereas for the value-weighted returns are calculated using the equations below. I first took the first lag of the individual securities market cap (by PERMNO key) and the total market cap on that particular day (by date key). I then divided the lagged individual securities market cap by the total market cap to obtain the weight at t-1 for a particular time t.

- Market Portfolio weights:

$$w_{i,t}^{mkt} = \frac{me_{i,t-1}}{\sum_i me_{i,t-1}}$$

- Market Portfolio

$$R^m kt_t = \sum_i w_{i,t}^{mkt} r_{i,t} = \sum_i \frac{me_{i,t-1}}{\sum_j me_{j,t-1}} r_{i,t}$$

6. Sample period: Lagged one-period of the respective stock is used at each date. This resulted in NA's, which will be omitted in the calculation of value-weighted returns. Reliable data on the companies did not become available until 1926 as per CRSP documentations. Therefore, 1926-2019 monthly data are used.

7. Missing Closing Price: As per CRSP, the price at close reflects either closing price or bid/ask average. If none of them are available, PRC is set to zero. In this case, I chose to omit the data. By filtering out the NA's in PRC, this also implicitly filters out the NAs in MKTCAP as well.

Part 2 - Fama-French Statistics

Using the risk-free rate from French's website, I *report the following moments of the market excess returns for both time series: annualized return, annualized volatility, annualized Sharpe ratio, skewness, and excess kurtosis.*

The summary statistics are shown in Table 2 below. I report the following five statistics, annualized mean return in excess of the risk-free rate, annualized standard deviation, annualized Sharpe ratio, skewness and excess kurtosis. In column 1, I report the statistics for the replicated value-weighted market portfolio of stocks calculated from Question 1. In column 2, I report the statistics for the value-weighted market portfolio of stocks from Ken French's website.

Table 2. Summary Statistics

	Replication	French's
Annualized Mean	0.0800	0.0800
Annualized Standard Deviation	0.1841	0.1844
Annualized Sharpe Ratio	0.4347	0.4339
Excess Skewness	0.2003	0.1763
Kurtosis	7.9145	7.8159

From question 1, I take the value-weighted time series and compare it with the time series data obtained from French's website. The statistics are computed as follows:

1. Sample period: monthly returns (not annualized) from July 1926 to December 2019

2. Excess skewness: I calculate it directly from the monthly time series data without annualization with the following formula for uni-variate data, using the full data. Where $\bar{Y}$ is the sample average.

$$Fisher - PearsonCoefficient = \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})^3/N}{s^3}$$

3. Kurtosis: I calculate it directly from the monthly time series data without annualization with the following formula for uni-variate data, using the full data. Minus three to accounts for the kurtosis inherent in the normal distribution.

$$Kurtosis = \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})^4/N}{s^4} - 3$$

4. Annualized mean:

$$\hat{\mu_A} = 12 * \hat{\mu_M}$$

5. Annualized standard deviation:

$$\hat{\sigma_A} = \sqrt{12} * \hat{\sigma_M}$$

6. Sharpe ratio: Where σ_M is the standard deviation of the portfolio's monthly excess return.

$$SR_M = \frac{Rp_M - Rf_M}{\sigma_M}$$

$$SR_A = \sqrt{12} * SR_M$$

Part 3 Correlation calculation

I report the time-series correlation between the replicated value-weighted market portfolio and the value-weighted market portfolio of stocks from Ken French's time series as well as their maximum absolute difference. Their correlation is **0.99998075** with a maximum absolute difference of 0.00460727.

I limit the sample to between July 1926 and December 2019. The maximum absolute difference is economically negligible as that 0.46% could be attributed to how the data was cleaned or to the difference in reporting rounding error from Ken French's data set. Everything was reported up to two decimal places in French's time series including the risk-free rate, whereas I used more than 8 decimal places in the calculation of excess returns. Since the maximum absolute difference is smaller than 0.01, and the correlation between the replicated and Ken-French's data is almost 1, it is therefore to use the replicated value-weighted portfolio as the one Ken French had constructed.

Summary statistics for the difference in bps:

```
> summary(joint$diff)*100
      Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
-0.2425954 -0.0092557 -0.0004659  0.0001997  0.0068620  0.4607273
```