Data Assignment 1 - `t54zheng` (20939203)

Task 2 - Basic Statistics

In [6]:
```
# imports

import pandas as pd
import matplotlib.pyplot as plt
import scipy.stats as stats
from datetime import timedelta
from math import sqrt

import warnings
warnings.filterwarnings('ignore')
```

In [2]:
```
# Import raw data
data_file = "djreturns.xlsx"

dj27 = pd.read_excel(data_file, sheet_name="dj27")
individual_dj27_returns = pd.read_excel(data_file, sheet_name="returns")
sp500_returns = pd.read_excel(data_file, sheet_name="sp500")
```

Task 2 - *Basic Statistics*

For each of the 27 stocks in `dj27`, and the market return `sp500_returns`, we want to find these statistics on their returns:

- The arithmetic mean
- standard deviation
- skewness
- kurtosis

In [3]:
```
# We need to group the data in individual_dj27_returns based on stock.
# Note that using PERMNO is a better idea since COMNAM can change.

# Let's show that our data actually has this issue:
duplicate_comnam_df = individual_dj27_returns[["PERMNO", "COMNAM"]].drop_duplicates().groupby("PERMNO").agg({'COMNAM': lambda x: list(x)})
duplicate_comnam_df
```

Out[3]:

PERMNO	COMNAM
10107	[MICROSOFT CORP]
10145	[HONEYWELL INTERNATIONAL INC]
11308	[COCA COLA CO]
12490	[INTERNATIONAL BUSINESS MACHS COR]
14008	[AMGEN INC]
14541	[CHEVRON CORP, CHEVRONTEXACO CORP, CHEVRON COR...
14593	[APPLE COMPUTER INC, APPLE INC]
18163	[PROCTER & GAMBLE CO]
18542	[CATERPILLAR INC]
19502	[WALGREEN CO, WALGREENS BOOTS ALLIANCE INC]
19561	[BOEING CO]
22111	[JOHNSON & JOHNSON]
22592	[MINNESOTA MINING & MFG CO, 3M CO]
22752	[MERCK & CO INC, MERCK & CO INC NEW]
26403	[DISNEY WALT CO]
43449	[MCDONALDS CORP]
47896	[CHASE MANHATTAN CORP NEW, J P MORGAN CHASE & ...
55976	[WAL MART STORES INC, WALMART INC]
57665	[NIKE INC]
59176	[AMERICAN EXPRESS CO]
59328	[INTEL CORP]
59459	[ST PAUL COS INC, ST PAUL TRAVELERS COS INC, T...
65875	[BELL ATLANTIC CORP, VERIZON COMMUNICATIONS INC]
66181	[HOME DEPOT INC]
76076	[CISCO SYSTEMS INC]
86868	[GOLDMAN SACHS GROUP INC]
92655	[UNITED HEALTHCARE CORP, UNITEDHEALTH GROUP INC]

We see many securities have multiple comnames as they have changed their company name over the period of the data, but PERMNO remains the same.

In [4]:
```
# So let's make a new dataframe for each PERMNO we have in dj27, and store them in a dict by PERMNO.
returns_dict = {} # permno -> dataframe(permno_returns)
```

```
permnos = dj27["PERMNO"]
for permno in permnos:
    returns_df = individual_dj27_returns[individual_dj27_returns["PERMNO"] == permno]
    returns_dict[permno] = returns_df
```

In [7]:

```
# Now that we have our data nicely organized, let's make a new dataframe to present our statistics
# We'll have every row describes the statistics for each return

stats_df = pd.DataFrame(columns=["permno", "Common Name(s)", "Mean (%)", "Standard Deviation (%)", "Skewness", "Kurtosis"])

# add using .loc[-1]
# First add the stats for the market portfolio
market_stats = {
    "permno": "market",
    "Common Name(s)": ["Market"],
    "Mean (%)": sp500_returns["SPRTRN"].mean(),
    "Standard Deviation (%)": sp500_returns["SPRTRN"].std(),
    "Skewness": stats.skew(sp500_returns["SPRTRN"]),
    "Kurtosis": stats.kurtosis(sp500_returns["SPRTRN"])
    }

# stats_df = stats_df.append(market_stats, ignore_index=True)
stats_df.loc[0] = [v for v in market_stats.values()]

# Now we'll add the rest of the securities from dj27
duplicate_comnam_dict = duplicate_comnam_df.to_dict()['COMNAM']

i = 1
for permno, df in returns_dict.items():
    permno_stats = {
        "permno": permno,
        "Common Name(s)": duplicate_comnam_dict[permno],
        "Mean (%)": df["RET"].mean(),
        "Standard Deviation (%)": df["RET"].std(),
        "Skewness": stats.skew(df["RET"]),
        "Kurtosis": stats.kurtosis(df["RET"])
    }
    # stats_df = stats_df.append(permno_stats, ignore_index=True)
    stats_df.loc[i] = [v for v in permno_stats.values()]
    i += 1

# Now we need to annualize the mean and standard deviation of the returns (currently monthly)
stats_df["Mean (%)"] = stats_df["Mean (%)"] * 12 # Annualize by multiplying by 12 (no compounding)
stats_df["Standard Deviation (%)"] = stats_df["Mean (%)"] * sqrt(12) # Annualizing stdev

# Format the results
stats_df["Mean (%)"] *= 100
stats_df["Standard Deviation (%)"] *= 100

# Round to 4 decimal places
stats_df = stats_df.round(4)
stats_df
```

Out[7]:

	permno	Common Name(s)	Mean (%)	Standard Deviation (%)	Skewness	Kurtosis
0	market	[Market]	6.5021	22.5238	-0.5312	1.0848
1	10107	[MICROSOFT CORP]	14.1157	48.8983	0.2160	3.3232
2	10145	[HONEYWELL INTERNATIONAL INC]	12.5522	43.4822	-0.1321	7.7921
3	11308	[COCA COLA CO]	7.5283	26.0787	-0.5014	1.1777
4	12490	[INTERNATIONAL BUSINESS MACHS COR]	6.5405	22.6570	0.4121	3.5562
5	14008	[AMGEN INC]	10.3270	35.7739	0.5216	1.8024
6	14541	[CHEVRON CORP, CHEVRONTEXACO CORP, CHEVRON COR...	10.7444	37.2197	0.4071	2.2700
7	14593	[APPLE COMPUTER INC, APPLE INC]	33.2421	115.1539	-0.6328	3.4377
8	18163	[PROCTER & GAMBLE CO]	9.3172	32.2759	-1.5059	9.1691
9	18542	[CATERPILLAR INC]	17.3336	60.0455	-0.0672	1.9528
10	19502	[WALGREEN CO, WALGREENS BOOTS ALLIANCE INC]	7.7165	26.7308	0.3525	0.6884
11	19561	[BOEING CO]	14.4704	50.1271	-0.3100	4.4419
12	22111	[JOHNSON & JOHNSON]	9.8556	34.1408	-0.1827	1.2589
13	22592	[MINNESOTA MINING & MFG CO, 3M CO]	10.4157	36.0809	-0.0241	0.6147
14	22752	[MERCK & CO INC, MERCK & CO INC NEW]	7.4162	25.6903	-0.2514	1.2391
15	26403	[DISNEY WALT CO]	12.0839	41.8598	0.0230	1.5061
16	43449	[MCDONALDS CORP]	13.0400	45.1720	-0.4927	2.3420
17	47896	[CHASE MANHATTAN CORP NEW, J P MORGAN CHASE & ...	12.4934	43.2783	-0.2393	1.2052
18	55976	[WAL MART STORES INC, WALMART INC]	6.9605	24.1118	-0.2615	1.0940
19	57665	[NIKE INC]	19.8069	68.6131	-0.1516	5.4126
20	59176	[AMERICAN EXPRESS CO]	11.4081	39.5188	2.7544	28.2196
21	59328	[INTEL CORP]	9.0218	31.2524	-0.5314	2.5265
22	59459	[ST PAUL COS INC, ST PAUL TRAVELERS COS INC, T...	12.4100	42.9895	1.4414	12.2851
23	65875	[BELL ATLANTIC CORP, VERIZON COMMUNICATIONS INC]	6.5369	22.6445	0.8691	5.4723
24	66181	[HOME DEPOT INC]	13.3259	46.1622	-0.1937	0.4112
25	76076	[CISCO SYSTEMS INC]	7.8689	27.2587	-0.1928	2.1937
26	86868	[GOLDMAN SACHS GROUP INC]	12.6296	43.7501	0.1206	0.7386
27	92655	[UNITED HEALTHCARE CORP, UNITEDHEALTH GROUP INC]	23.7253	82.1869	-0.6224	2.6404

In [8]:
```
# The market portfolio
stats_df[stats_df["permno"] == "market"]
```

Out[8]:

	permno	Common Name(s)	Mean (%)	Standard Deviation (%)	Skewness	Kurtosis
0	market	[Market]	6.5021	22.5238	-0.5312	1.0848

Statistics of the market portfolio

Here's a table comparing the skewness and kurtosis of the market returns to a normal distribution

	Normal Distribution	Market	Interpretation
Skewness	0	-0.5312	`Negative Skew`
Kurtosis	3	1.0848	`Low Kurtosis`

Skewness: Since the market's returns are negatively skewed, the returns are more heavily distributed towards the right, above zero. We can expect the market to return positive returns more often than if the market returns were perfectly normally distributed

Kurtosis: Since the market's returns have a low kurtosis, the returns lower but are more tightly concentrated towards the mean, indicating that market returns offer a lower amount of risk compared to if the market returns were perfectly normally distributed