

A mannequin in a dark suit and blue striped tie stands with arms crossed. To its right is a large, 3D red arrow pointing upwards. The background is a blurred outdoor scene with green foliage and a white building.

MUTUAL FUNDS, YIELDS AND RISK.

**An exploratory data analysis and hypothesis
test.**

The level of risk in a mutual fund depends on what it invests in. Usually, the higher the potential returns, the higher the risk will be. For example, stocks are generally riskier than bonds, so an equity fund tends to be riskier than a fixed income fund.

Authored by
Sai Kiran Anagani and Shashank Reddy

Table of Contents

Introduction.....	2
Code source:.....	2
Executive Summary	3
Dataset Preparation	4
Initialize the Jupyter notebook and install dependencies	4
Data import.....	4
Exploratory analysis	5
Average Mutual fund risk takers.....	5
Top 10 mutual fund categories.....	6
Sector diversity	7
Fund Diversity	8
Growth rate in the top sectors	9
Fee, ratings and returns	10
.....	10
Risk vs Yield.....	11
Hypothesis testing:	12
Results:.....	14
ANOVA:	15
CONCLUSION	16

Introduction.

A mutual fund is a company that pools money from many investors and invests the money in securities such as stocks, bonds, and short-term debt. The combined holdings of the mutual fund are known as its portfolio.

The level of risk in a mutual fund depends on what it invests in. Usually, the higher the potential returns, the higher the risk will be. For example, stocks are generally riskier than bonds, so an equity fund tends to be riskier than a fixed income fund.

Some specialty mutual funds focus on certain kinds of investments, such as emerging markets, to try to earn a higher return. These kinds of funds also tend to have a greater risk of a larger drop in value.

Code source:

The source of this project is available on Github at https://github.com/anaganisk/mutual_funds

Executive Summary

The general belief among people that mutual funds tend to be riskier and are not safe to invest. In this project herein, we will focus on yield returns and rejecting the null hypothesis that the mutual funds are risky. We will do this by the following:

Part 1: Collection of the Dataset - Data is the backbone of any sound decision making. In this part, we will focus on the data sources.

Part 2: Exploratory Data Analysis (EDA) - an approach to analyzing data sets to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task.

Part 3: Hypothesis Testing - In theory, methods, and practice of testing a hypothesis by comparing it with the null hypothesis. The null hypothesis is only rejected if its probability falls below a predetermined significance level, in which case the hypothesis being tested is said to have that level of significance. In our case our Hypothesis Testing case would be:

H₀ : Null Hypothesis: Mutual funds are risky and does not give much returns.

Average rate of return (μ) \leq ~2 %

Alternate Hypothesis you are trying to prove if Mutual funds are safe to invest.

H_a : Alternate Hypothesis : Mutual funds are safe and give good returns.

Average rate of return (μ) $>$ 2%

Dataset Preparation

Our dataset consists for European stock markets scraped from Morningstar stock rating agency. You can download the scraped data from here.

https://raw.githubusercontent.com/anaganisk/mutual_funds/main/Morningstar%20-%20European%20Mutual%20Funds.csv

Technologies to be used

Jupyter, an interactive python playground.

Python statistical tools like sklearn and scipy.

Python visualization tools like matplotlib and seaborn.

Initialize the Jupyter notebook and install dependencies

```
!pip install sklearn
import pandas as pd
import numpy as np
import matplotlib as mpl
import matplotlib.pyplot as plt
import seaborn as sns
from scipy import stats
from sklearn.model_selection import train_test_split
from scipy import stats
from statsmodels.stats.weightstats import ztest as ztest
```

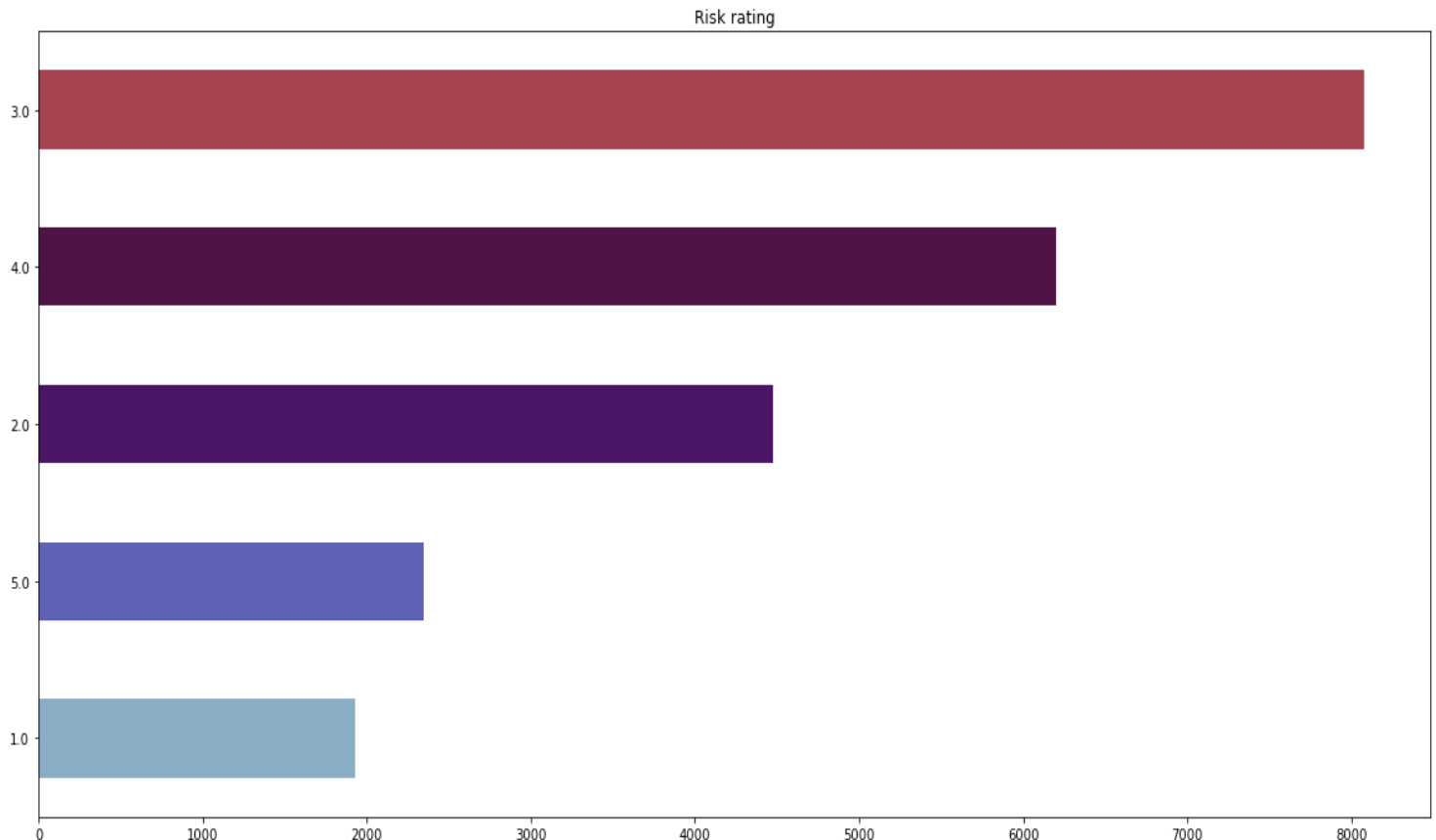
Data import

```
df = pd.read_csv("./Morningstar - European Mutual Funds.csv")
# drop rows with empty data
df.dropna(subset=['rating', 'risk_rating', 'roic', 'fund_trailing_return_ytd'],
inplace=True)
df.info()
<class 'pandas.core.frame.DataFrame'> Int64Index: 23034 entries, 0 to
57602 Columns: 132 entries, ticker to quarters_down dtypes: float64(108),
int64(2), object(22) memory usage: 23.4+ MB
```

Exploratory analysis

Average Mutual fund risk takers

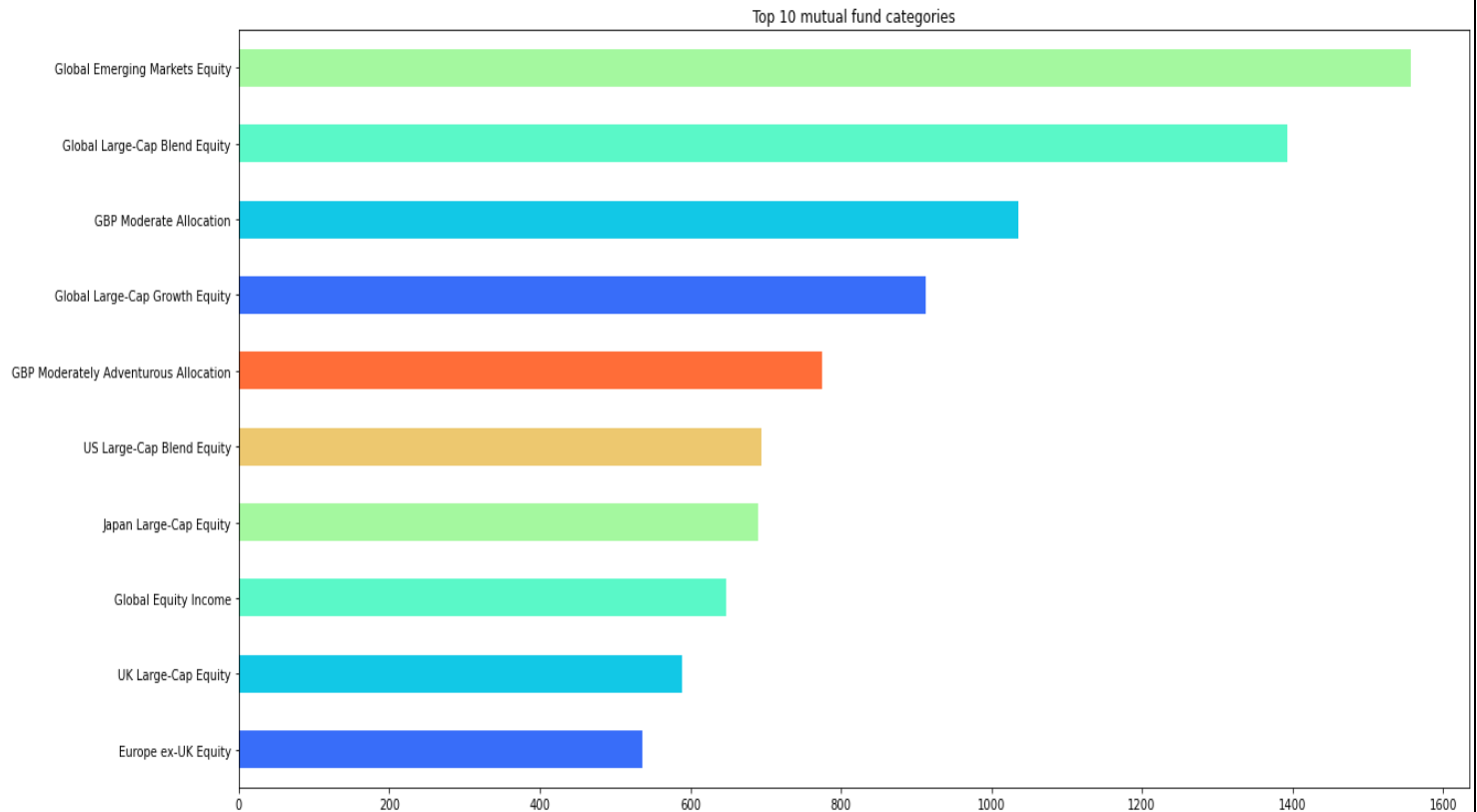
```
df.risk_rating.value_counts().iloc[:10].sort_values().plot(kind="barh", title = "Risk rating", figsize=(20,10), color = sns.color_palette("twilight"))
```



We can see that many mutual funds companies balance risk moderately, followed some companies taking higher risks. To boost adoption by people mutual funds, try to be more profitable, according to the general notion being, higher risk means better profits. But it also makes sense that **many like to play safe with moderation.**

Top 10 mutual fund categories.

```
df['category'].value_counts()[:10].sort_values().plot(kind="barh", title = "Top 10 mutual fund categories", figsize=(20,10), color = sns.color_palette("rainbow"))
```



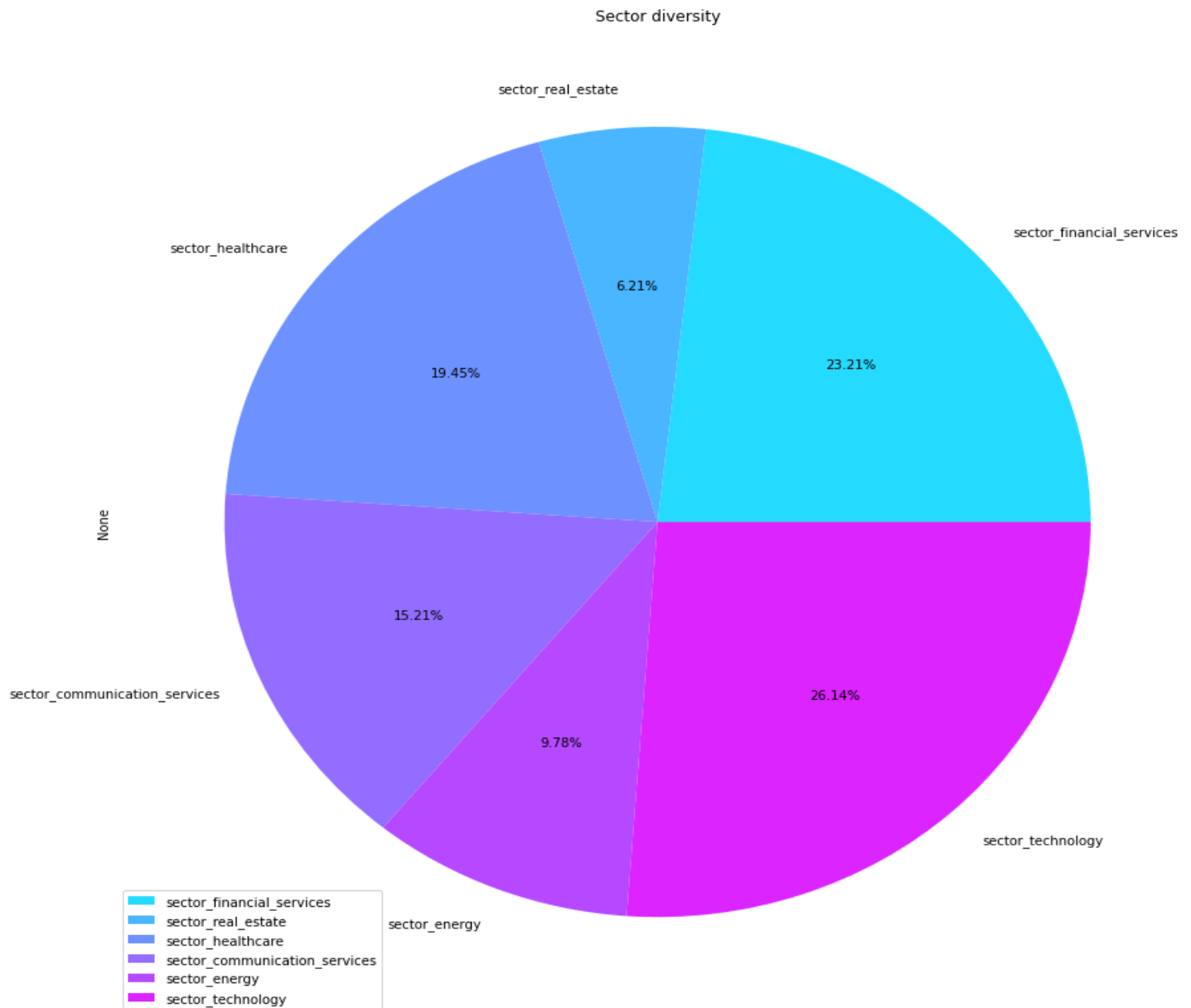
From the page of GEME

The Global Emerging Markets Equity Strategy seeks attractive long-term, risk-adjusted returns by investing in emerging market equities

From the categories show we see that **a large number of mutual funds are of equity types.**

Sector diversity

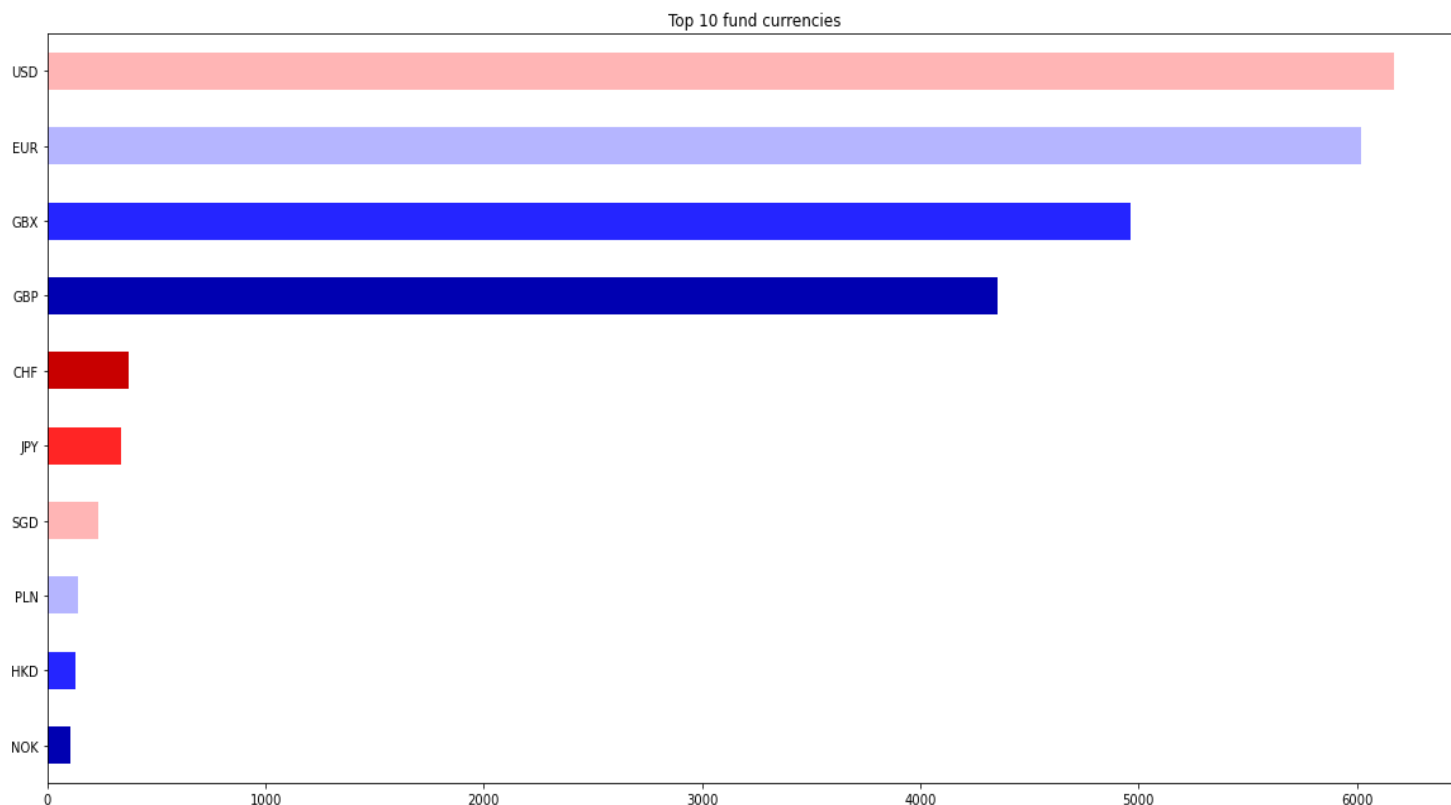
```
y = means.plot(kind='pie', figsize=(15,15), colors=sns.color_palette("cool"), legend=True, autopct='%1.2f%%', shadow=False, startangle=0, title="Sector diversity")
```



From the diagram we can notice the major part of mutual funds are invested into **technology, finance and healthcare**. Which makes sense due to the fact that that those sectors are growing at a quick pace. **Technology currently being heavily invested** in to make more profitable ventures **by optimizing business processes**.

Fund Diversity

```
df.nav_per_share_currency.value_counts().iloc[:10].sort_values().plot(kind="barh", title = "Top 10 fund currencies", figsize=(20,10), color = sns.color_palette("seismic"))
```



The USD seems to be the major fund currency, which could possibly be since, **technology sector takes the lions share in the sector diversity. Major tech giants are based in USA.** Investments made in USA are made in USD, hence it seems to be dominant.

Growth rate in the top sectors

```
fig, ax = plt.subplots(nrows=2, ncols=3, figsize=(25, 12))
```

```
sns.regplot(data=df, x='sector_financial_services', y='fund_trailing_return_ytd', truncate=True, ax=ax[0,0])
```

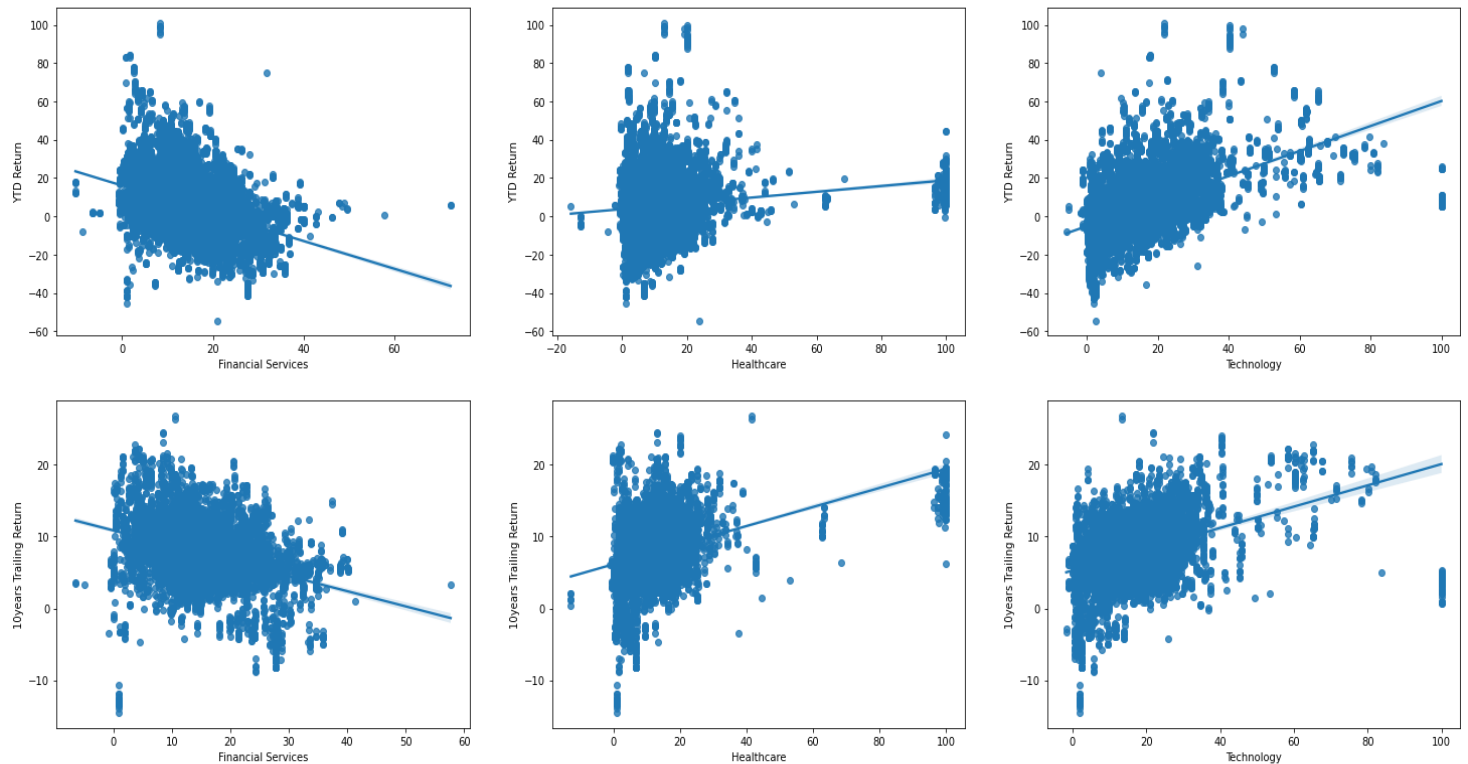
```
sns.regplot(data=df, x='sector_healthcare', y='fund_trailing_return_ytd', truncate=True, ax=ax[0,1])
```

```
sns.regplot(data=df, x='sector_technology', y='fund_trailing_return_ytd', truncate=True, ax=ax[0,2])
```

```
sns.regplot(data=df, x='sector_financial_services', y='fund_trailing_return_10years', truncate=True, ax=ax[1,0])
```

```
sns.regplot(data=df, x='sector_healthcare', y='fund_trailing_return_10years', truncate=True, ax=ax[1,1])
```

```
sns.regplot(data=df, x='sector_technology', y='fund_trailing_return_10years', truncate=True, ax=ax[1,2])
```

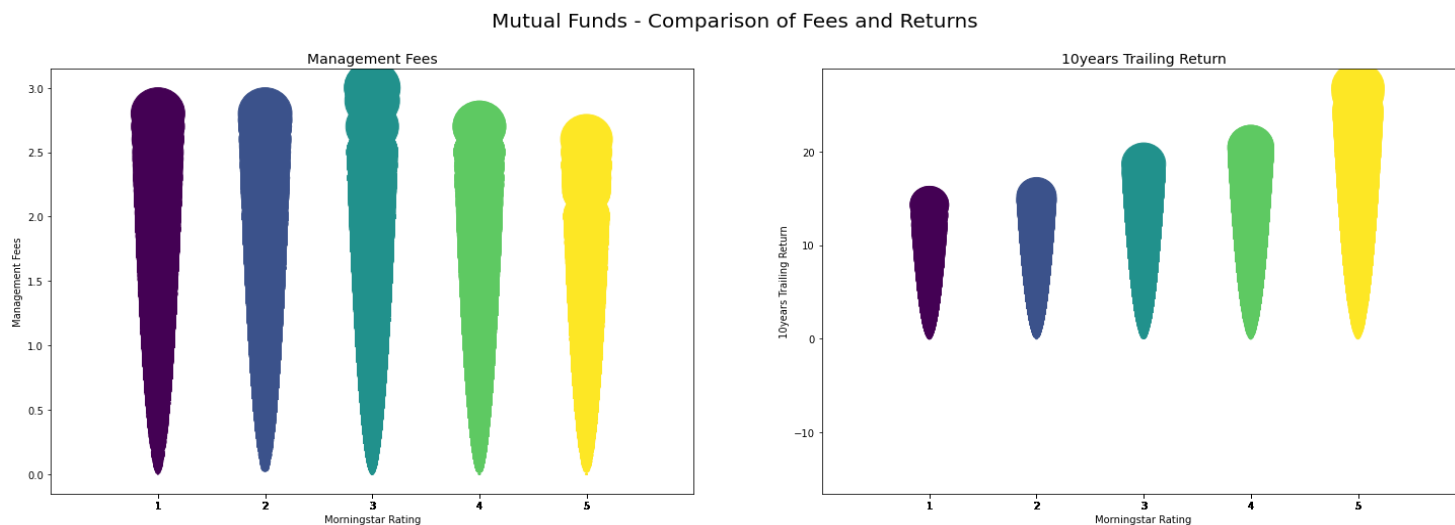


Here we are realizing the trends of **Year To Date** returns and **10 year trailing** returns, where we see similar patterns of growing yields in technology and healthcare sector while, **financial sectors are dropping in yield with a negative relationship**.

The rise to the health care could be attributed both to advancements in bio-medical industry, increase in number of sickness and also to the recent COVID vaccine race.

Fee, ratings and returns

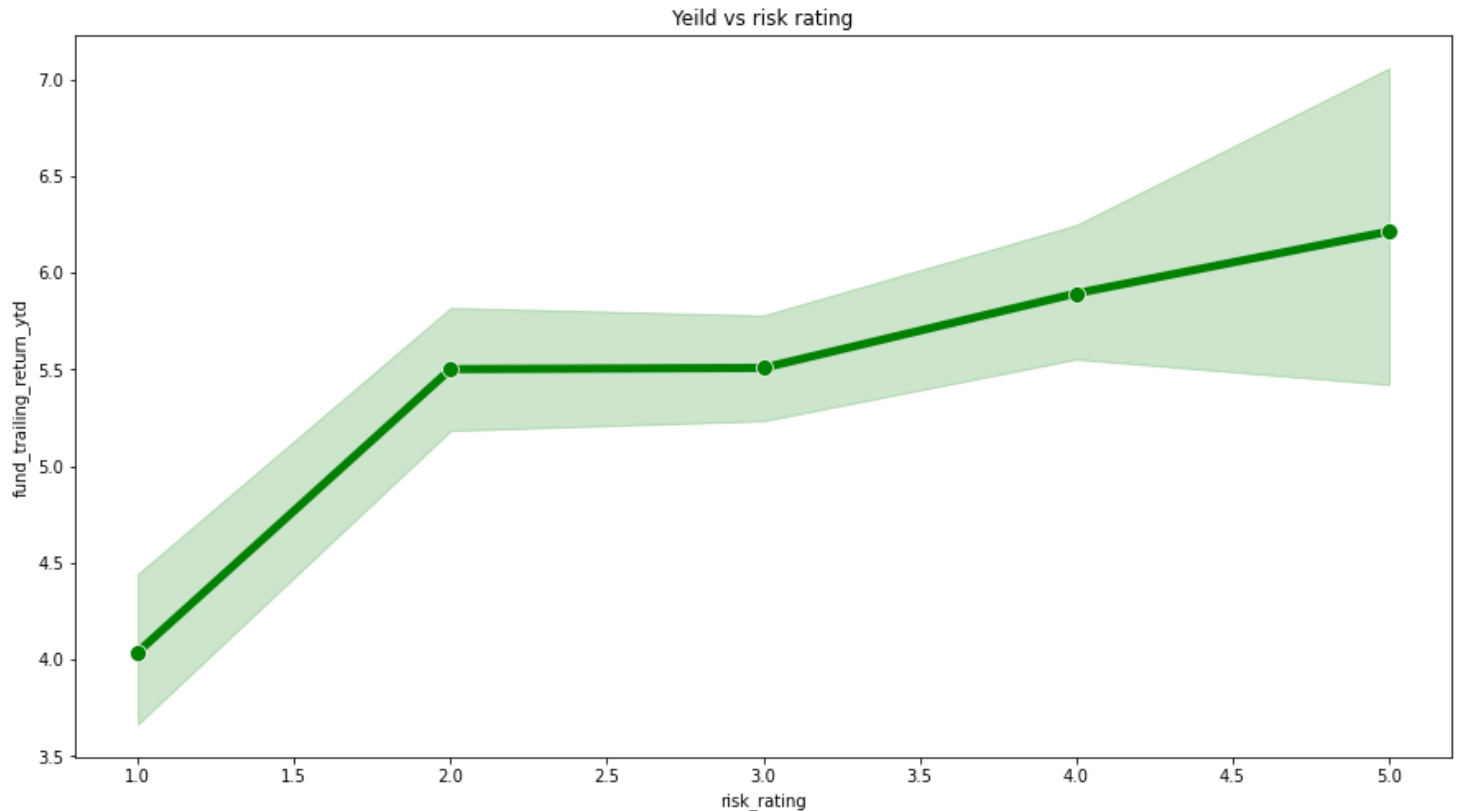
```
fig, ax = plt.subplots(nrows=1, ncols=2, figsize=(25, 8))
fig.suptitle('Mutual Funds - Comparison of Fees and Returns', fontsize=20)
ax[0].scatter(df['rating'], df['management_fees'], s=1e3*df['management_fees'], c=df['rating'])
ax[0].set_xticks(df['rating'])
ax[0].margins(x=0.25)
ax[0].set_xlabel('Morningstar Rating')
ax[0].set_ylabel('Management Fees')
ax[0].set_title("Management Fees", fontsize=14)
ax[1].scatter(df['rating'], df['fund_trailing_return_10years'], s=1e2*df['fund_trailing_return_10years'],
c=df['rating'])
ax[1].set_xticks(df['rating'])
ax[1].margins(x=0.25)
ax[1].set_xlabel('Morningstar Rating')
ax[1].set_ylabel('10years Trailing Return')
ax[1].set_title("10years Trailing Return", fontsize=14)
```



We can see there is not much relations with fee's charged for management and the rating rating of the fund, but we can see that as the social rating of the fund increases the yield is higher. Which makes it obvious that social ratings can help us invest in better yielding funds.

Risk vs Yield

```
p=sns.lineplot(x='risk_rating', y='fund_trailing_return_ytd', data=df, marker='o',  
markersize=10, color='green', linewidth=5)  
plt.title('Yeild vs risk rating')  
plt.gcf().set_size_inches(15, 8)
```



Based on the dataset we acquired we can clearly see that the increase in risk increases the yield too. **It has a positive relationship.**

Hypothesis testing:

H0 : Null Hypothesis: Mutual funds are risky and does not give much returns.

Average rate of return (μ) $\leq \sim 2\%$

Alternate Hypothesis you are trying to prove if Mutual funds are safe to invest.

Ha : Alternate Hypothesis : Mutual funds are safe and give good returns.

Average rate of return (μ) $> 2\%$

We are splitting the data randomly into **3 samples** with each sample consisting of **20% of the data, and perform z-test.**

```
from scipy import stats
from statsmodels.stats import weightstats as stests

df_sample_1 = df.sample(frac=0.20, random_state=29)
df_sample_2 = df.sample(frac=0.20, random_state=16)
df_sample_3 = df.sample(frac=0.20, random_state=3)

test_avg = 0.2

print("H0 : Null Hypothesis: Mutual funds are risky and does not give much returns.")
print("Ha : Alternate Hypothesis : Mutual funds are safe and give good returns.")

print("Sample 1")
ztest , pval = stests.ztest(df_sample_1['fund_trailing_return_ytd'], value=test_avg,
alternative='larger')
print(f"z-test: {ztest}, p-value: {pval}")
if pval<0.05:
    print("reject null hypothesis")
else:
    print("accept null hypothesis")
```

```
print("Sample 2")
ztest , pval = stests.ztest(df_sample_2['fund_trailing_return_ytd'], value=test_avg,
alternative='larger')
print(f"z-test: {ztest}, p-value: {pval}")
if pval<0.05:
    print("reject null hypothesis")
else:
    print("accept null hypothesis")

print("Sample 2")
ztest , pval = stests.ztest(df_sample_3['fund_trailing_return_ytd'], value=test_avg,
alternative='larger')
print(f"z-test: {ztest}, p-value: {pval}")
if pval<0.05:
    print("reject null hypothesis")
else:
    print("accept null hypothesis")
```

Results:

H₀ : Null Hypothesis: Mutual funds are risky and does not give much returns.

H_a : Alternate Hypothesis : Mutual funds are safe and give good returns.

Sample 1

z-test: 26.452833089968486, p-value: 1.6924704636798308e-154

reject null hypothesis

Sample 2

z-test: 27.469349374937803, p-value: 2.0405464578927426e-166

reject null hypothesis

Sample 2

z-test: 26.850330754754864, p-value: 4.1799529620670183e-159

reject null hypothesis

Based on the results of z-test from the samples we can reject null hypothesis and say Mutual funds are safe and give good returns.

ANOVA:

We are performing a one-way ANOVA test to determine the viability of our samples, so that we can say for sure our test is good.

```
F, p = stats.f_oneway(df_sample_1['fund_trailing_return_ytd'],  
df_sample_2['fund_trailing_return_ytd'], df_sample_3['fund_trailing_return_ytd'])  
print(p)  
if p<0.05:  
    print("reject null hypothesis - The samples are not similar")  
else:  
    print("accept null hypothesis - The samples are similar")
```

0.8047305068356624

accept null hypothesis - The samples are similar

Since the value is much higher than 0.05 we accept null hypothesis that the samples are similar and can say with confidence that **results are good**.

CONCLUSION

In this project, we deep-dived and performed an exhaustive analysis in support of our goal in finding supportive evidence to reject the notion that the Mutual funds are not safe.

We gathered data for dispersed sources in support of our EDA (Exploratory Data Analysis) and subsequently and most importantly the Hypothesis Testing to reject null Hypothesis.

We visualized risks, fees, types of funds, sector diversity currency, which all project a growth pattern.

The results of the Hypothesis testing were overwhelming in favor of rejecting our null Hypothesis and in support of Alternate Hypothesis that Mutual funds are safe and give good returns. In all 3 samples the probability was less than the Significant value of 0.05 in rejecting the null Hypothesis. In closing,

Mutual funds are safe and give good returns.