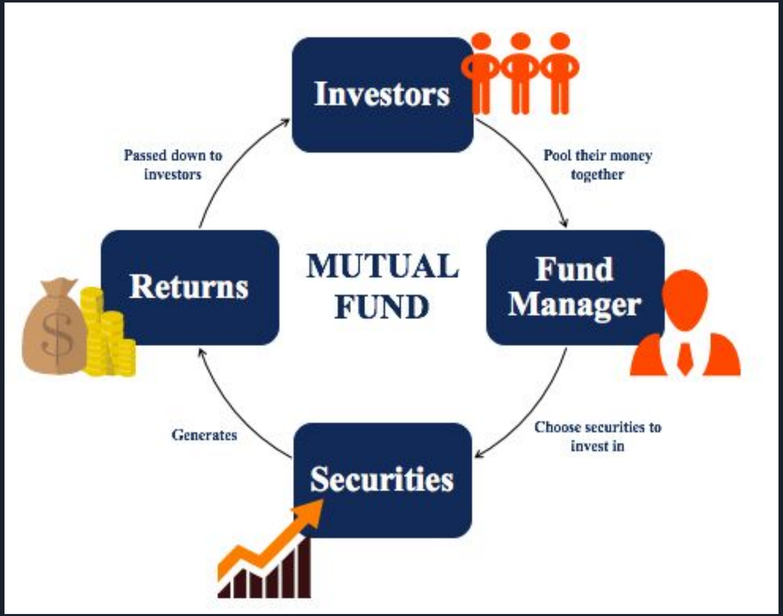
10 Year Mutual Fund Returns as an Indicator for Retirement

Luke Martin, Ben Kowalski, Ben Jaffe, Hayden Vaughn, Jack Boydell

Introduction/Overview

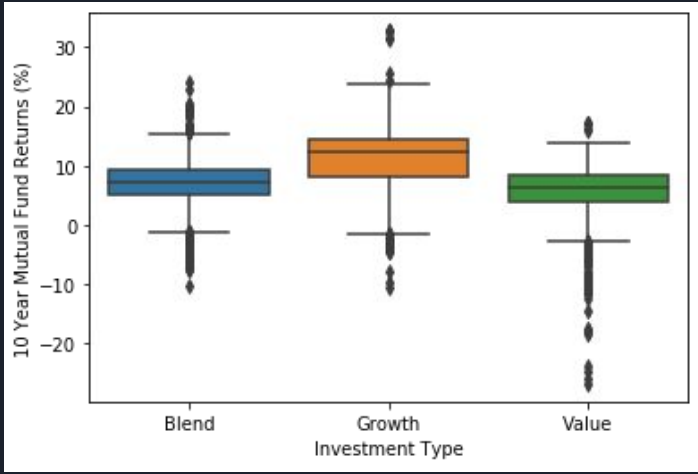
- Picking mutual funds for the purpose of retirement is a common topic of discussion among investors.
- The funds were broken up by investment strategy into 3 stratified samples.
- A pearson correlation heatmap was used within each sample to determine which sectors of the economy were most correlated with the 10 year return of mutual funds.
- These highly correlated sectors were combined with the size of fund to train a multiple linear regression model to predict the fund return over 10 years using a testing set.

What is a Mutual Fund?



- Mutual funds are companies that pool many investors money, and invest in several different securities across the market such as bonds and stocks.
- There are many different types of strategies for mutual funds, but the main mutual funds this project focuses on are the strategies of growth, value and blend.
- Typically, mutual funds are popular for retirement funds due to the low volatility.
- Because of this, mutual funds are much safer than investing in a single stock, and therefore many investors use mutual funds for their retirement.

The Three Types Investigated



- A value-oriented fund focuses primarily on stocks that are considered better value than given criteria.
- A growth-oriented fund focuses primarily on stocks that are predicted to grow at a rate faster than that of the overall market.
- Lastly, a blend investment strategy is a mix between both value and growth.

Project Goal

- The goal of this project is to determine, for the purpose of retirement planning, which factors, specifically the percentage makeup of mutual fund assets from different sectors of the economy, are most important in predicting the 10 year return on mutual funds grouped by 3 different investment strategies (growth, value, blend).
- The purpose of grouping by investment strategy type is to be able to observe any differences between groups and avoid any influence that the investment type parameter might have on a larger model if the three were not separated.

Stratified Sampling

Based on initial thoughts of investment type influence and exploratory analysis, we decided to take a stratified sample by investment type to create three avenues for separate analysis.

```
In [207]: n_sample = 3000 #how many of each of three investment types we want
          full_sample_df = mutual_funds.groupby('investment_type').apply(lambda x: x.sample(n_sample, random_state=1))
          full_sample_df.head(3001)

          #splitting up full dataframe sample by three investment types
          mutual_funds_Value = full_sample_df[full_sample_df['investment_type'] == 'Value']
          mutual_funds_Growth = full_sample_df[full_sample_df['investment_type'] == 'Growth']
          mutual_funds_Blend = full_sample_df[full_sample_df['investment_type'] == 'Blend']
```

Training/Testing Split

[**from** sklearn.model_selection **import** train_test_split]

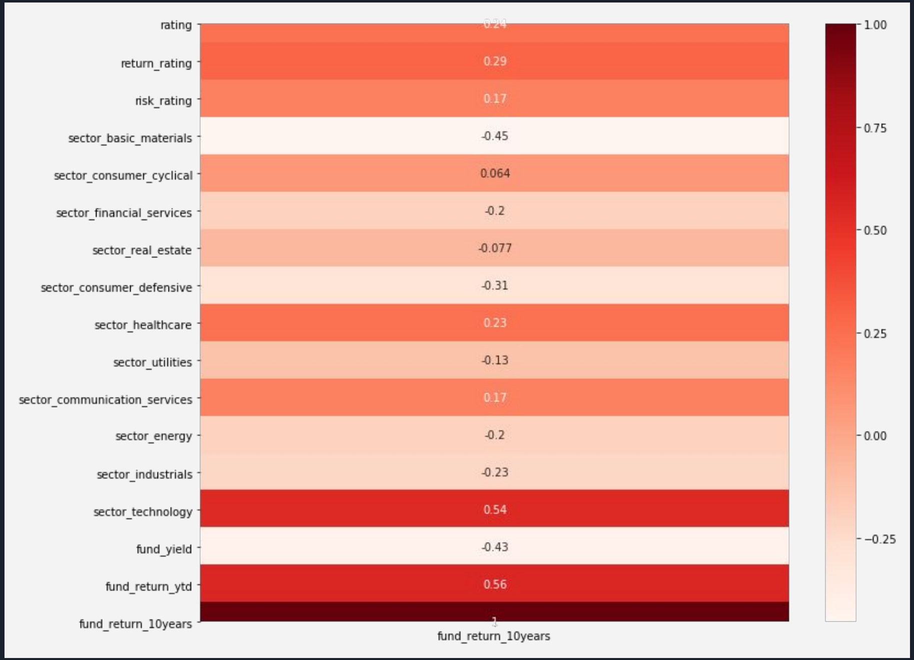
- Training (fitting of regression models) and testing (evaluating model performance) sets created using sklearn.model_selection package

```
In [313]: # splitting each stratified sample data frame into testing/training sets
          test_set_proportion = 0.25

          # random state set to 1 for reproducibility
          Value_train, Value_test = train_test_split(mutual_funds_Value, test_size = test_set_proportion, random_state=1)
          Growth_train, Growth_test = train_test_split(mutual_funds_Growth, test_size = test_set_proportion, random_state=1)
          Blend_train, Blend_test = train_test_split(mutual_funds_Blend, test_size = test_set_proportion, random_state=1)
```

Growth Implementation

Feature selection, particularly the use of Pearson Correlation, is effective in helping choose features to include in further model creation and analysis and is performed as a part of EDA visualization.

Seaborn Correlation Heatmap



Creating correlation heatmap selected for 10 year returns

```
In [256]: #Using Pearson Correlation
          plt.figure(figsize=(10,5)) #(changed from (12,10))
          cor = mutual_funds_Growth.corr()
          response_column = pd.DataFrame(cor['fund_return_10years'])
          sns.heatmap(response_column, annot=True, cmap=plt.cm.Reds)
          plt.show()
```

Selecting features with a correlation greater than 0.4 (absolute value)

```
In [257]: #Correlation with output variable
          cor_target = abs(cor["fund_return_10years"]) # absolute value
          #Selecting highly correlated features
          relevant_features = cor_target[cor_target>0.4]
          relevant_features

Out[257]: sector_basic_materials     0.453395
          sector_technology          0.544251
          fund_yield                 0.426123
          fund_return_ytd            0.557175
          fund_return_10years        1.000000
          Name: fund_return_10years, dtype: float64
```

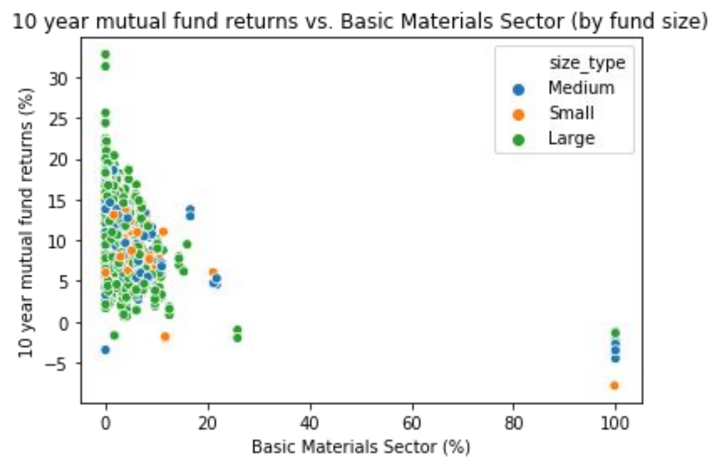
Growth Regression Models

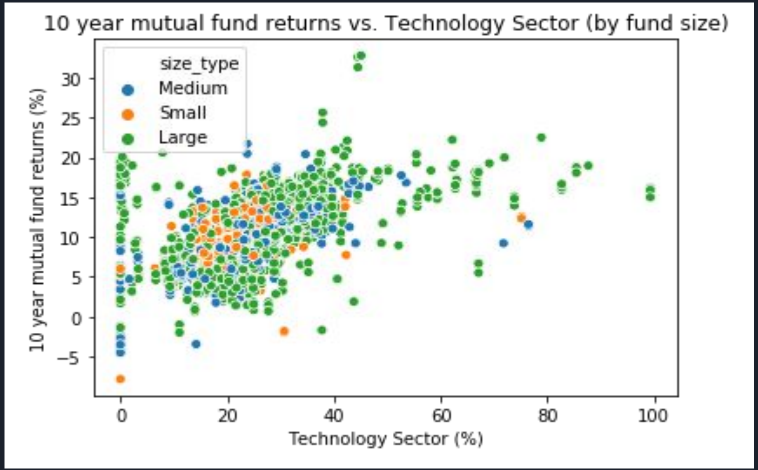
10 year mutual fund returns vs. basic materials sector + size type

```
In [280]: #multiple linear regression: basic materials + size type (small, medium, large)
          basic_materials_with_size = ols('fund_return_10years ~ sector_basic_materials + size_type + 0', data=Growth_train).fit(
          print(basic_materials_with_size.summary())
```

10 year mutual fund returns vs. technology sector + size type

```
In [284]: #multiple linear regression: technology + size type (small, medium, large)
          technology_with_size = ols('fund_return_10years ~ sector_technology + size_type + 0', data=Growth_train).fit()
          print(technology_with_size.summary())
```





Growth Results/Analysis

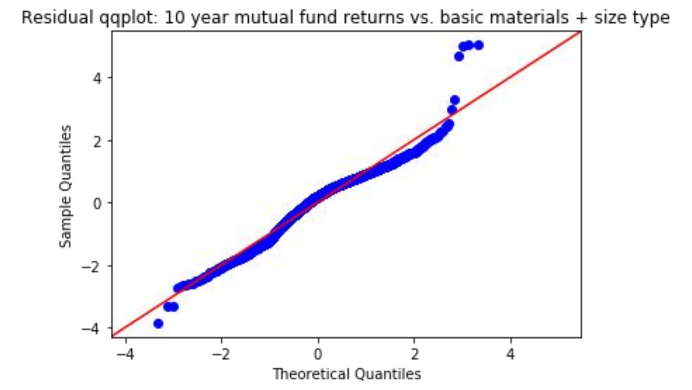
- Multiple linear regression fitted on correlated sectors while including size_type (small, med, large)

Basic Materials

OLS Regression Results

Dep. Variable:	fund_return_10years	R-squared:	0.212
Model:	OLS	Adj. R-squared:	0.211
Method:	Least Squares	F-statistic:	201.3
Date:	Wed, 05 May 2021	Prob (F-statistic):	1.29e-115
Time:	15:54:25	Log-Likelihood:	-6359.7
No. Observations:	2250	AIC:	1.273e+04
Df Residuals:	2246	BIC:	1.275e+04
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
size_type[Large]	12.2166	0.116	105.613	0.000	11.990	12.443
size_type[Medium]	12.2866	0.171	72.002	0.000	11.952	12.621
size_type[Small]	11.3698	0.226	50.241	0.000	10.926	11.814
sector_basic_materials	-0.1982	0.008	-24.188	0.000	-0.214	-0.182

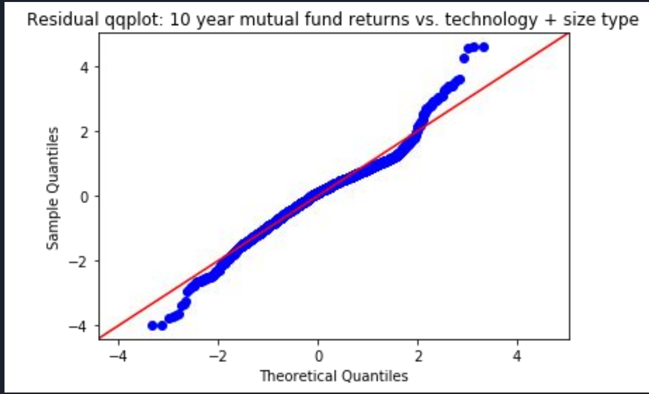


Technology

OLS Regression Results

Dep. Variable:	fund_return_10years	R-squared:	0.302
Model:	OLS	Adj. R-squared:	0.301
Method:	Least Squares	F-statistic:	323.8
Date:	Wed, 05 May 2021	Prob (F-statistic):	1.10e-174
Time:	16:00:11	Log-Likelihood:	-6223.3
No. Observations:	2250	AIC:	1.245e+04
Df Residuals:	2246	BIC:	1.248e+04
Df Model:	3		
Covariance Type:	nonrobust		

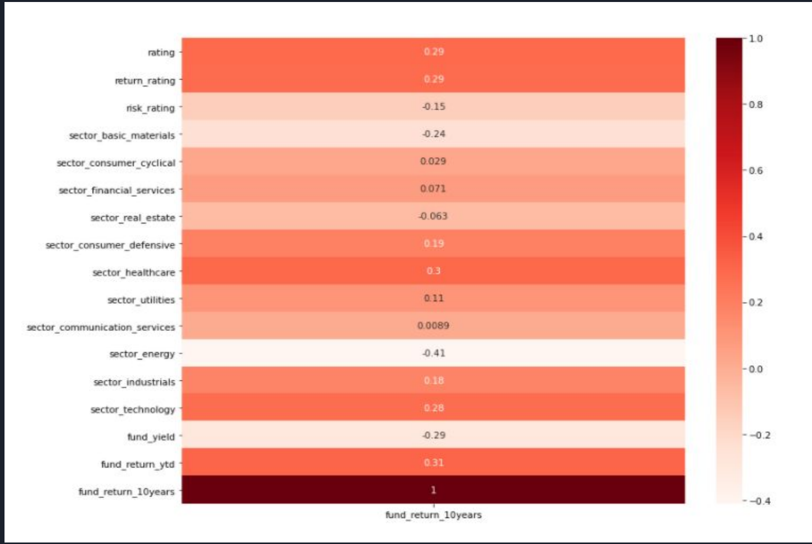
	coef	std err	t	P>|t|	[0.025	0.975]
size_type[Large]	6.0163	0.213	28.217	0.000	5.598	6.434
size_type[Medium]	6.1428	0.223	27.555	0.000	5.706	6.580
size_type[Small]	6.3123	0.256	24.688	0.000	5.811	6.814
sector_technology	0.2013	0.007	30.823	0.000	0.188	0.214



Qqplots: It can be observed that both extremes begin to stray away from the normality line in both models. This suggests that the distributions for each are slightly light tailed and potentially peaked in the middle in comparison to a normal distribution

Value Implementation

Seaborn Correlation Heatmap



It was determined that the healthcare and energy sectors have the strongest relationship to our response variable (~0.3 and ~-0.41, respectively).

Creating correlation heatmap selected for 10 year returns

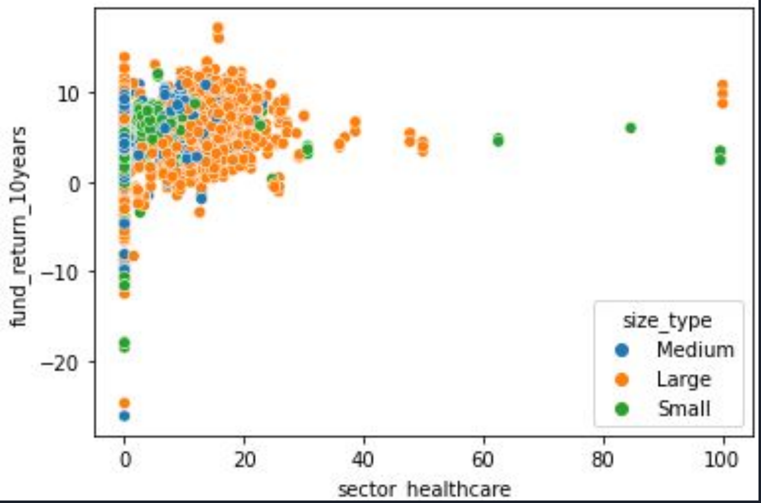
```
In [33]: #Using Pearson Correlation
         plt.figure(figsize=(12,10)) #(changed from (12,10))
         cor = mutual_funds_Value.corr()
         response_column = pd.DataFrame(cor['fund_return_10years'])
         sns.heatmap(response_column, annot=True, cmap=plt.cm.Reds)
         plt.show()
```

Selecting features with a correlation greater than 0.295

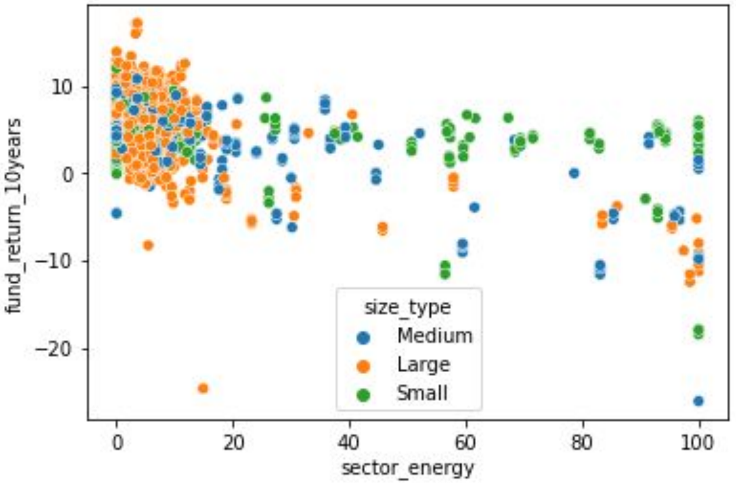
```
In [34]: #Correlation with output variable
         cor_target = abs(cor["fund_return_10years"]) # absolute value
         #Selecting highly correlated features
         relevant_features = cor_target[cor_target>0.295]
         relevant_features

Out[34]: sector_healthcare      0.295136
         sector_energy          0.410647
         fund_return_ytd        0.308930
         fund_return_10years    1.000000
         Name: fund_return_10years, dtype: float64
```

Value Implementation cont.

```
Intercept             4.657186
sector_healthcare     0.116333
dtype: float64
```




```
Intercept        6.549497
sector_energy   -0.069828
dtype: float64
```



Value Models

Linear model for 10 year return as a function of sector_healthcare and fund size

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     fund_return_10years   R-squared:                       0.086
Model:                             OLS   Adj. R-squared:                  0.085
Method:                  Least Squares   F-statistic:                     70.57
Date:                 Wed, 28 Apr 2021   Prob (F-statistic):           1.30e-43
Time:                         23:01:58   Log-Likelihood:                -6006.6
No. Observations:                 2250   AIC:                         1.202e+04
Df Residuals:                     2246   BIC:                         1.204e+04
Df Model:                            3
Covariance Type:             nonrobust
=====================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
size_type[Large]      5.1578      0.150     34.433      0.000       4.864       5.452
size_type[Medium]     4.7281      0.167     28.298      0.000       4.400       5.056
size_type[Small]      4.3282      0.171     25.268      0.000       3.992       4.664
sector_healthcare     0.0941      0.009     10.926      0.000       0.077       0.111
```

Linear model for 10 year return as a function of sector_energy and fund size

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     fund_return_10years   R-squared:                       0.159
Model:                             OLS   Adj. R-squared:                  0.157
Method:                  Least Squares   F-statistic:                     141.1
Date:                 Wed, 05 May 2021   Prob (F-statistic):           9.63e-84
Time:                         22:38:26   Log-Likelihood:                -5913.7
No. Observations:                 2250   AIC:                         1.184e+04
Df Residuals:                     2246   BIC:                         1.186e+04
Df Model:                            3
Covariance Type:             nonrobust
=====================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
size_type[Large]      6.7726      0.094     72.116      0.000       6.588       6.957
size_type[Medium]     5.8895      0.158     37.251      0.000       5.579       6.200
size_type[Small]      6.1907      0.179     34.659      0.000       5.840       6.541
sector_energy        -0.0637      0.004    -17.970      0.000      -0.071      -0.057
```

```python
#initial linear regression model testing
practice_model1 = ols('fund_return_10years ~ sector_energy + size_type + 0', data=Value_train).fit()
print(practice_model1.summary())

sns.scatterplot(x= 'sector_energy', y= 'fund_return_10years', data=mutual_funds_Value, hue='size_type')
```

Combined Value Model

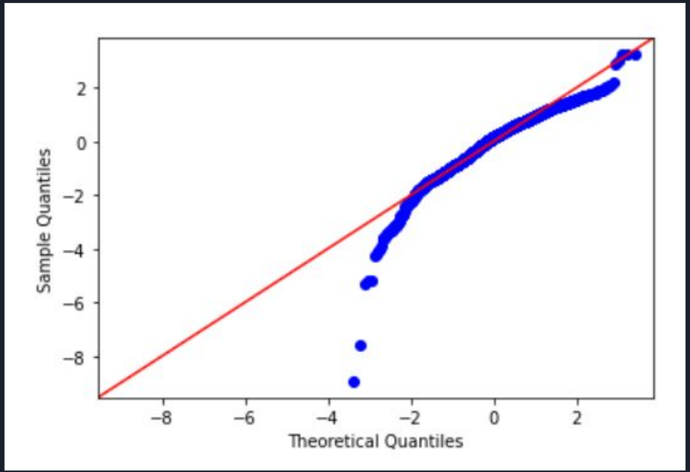
The two individual value models were combined into one better model for 10 year return as a function of sector_energy, sector_healthcare, and fund size.

```
#initial linear regression model testing
practice_model3 = ols('fund_return_10years ~ sector_healthcare + sector_energy + size_type + 0',data=Value_train).fit()
print(practice_model2.summary())
```

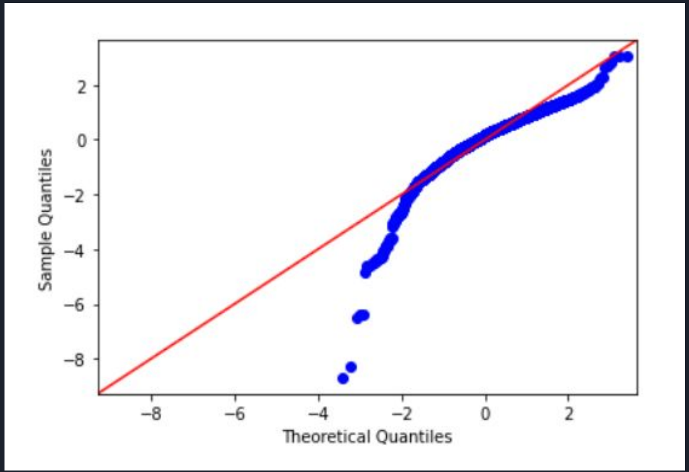
```
                            OLS Regression Results
==============================================================================
Dep. Variable:     fund_return_10years   R-squared:                       0.185
Model:                             OLS   Adj. R-squared:                  0.184
Method:                  Least Squares   F-statistic:                     127.8
Date:                 Wed, 05 May 2021   Prob (F-statistic):           2.09e-98
Time:                         22:27:50   Log-Likelihood:                -5877.1
No. Observations:                 2250   AIC:                         1.176e+04
Df Residuals:                     2245   BIC:                         1.179e+04
Df Model:                            4
Covariance Type:             nonrobust
=====================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
size_type[Large]      5.7930      0.147     39.524      0.000       5.506       6.080
size_type[Medium]     5.4509      0.164     33.296      0.000       5.130       5.772
size_type[Small]      5.7486      0.183     31.392      0.000       5.390       6.108
sector_energy        -0.0585      0.004    -16.545      0.000      -0.065      -0.052
sector_healthcare     0.0710      0.008      8.611      0.000       0.055       0.087
```

Value Results

QQplot for residuals of linear regression model (sector_energy) versus (fund_return_10years)



QQplot for residuals of linear regression model (sector_healthcare) versus (fund_return_10years)

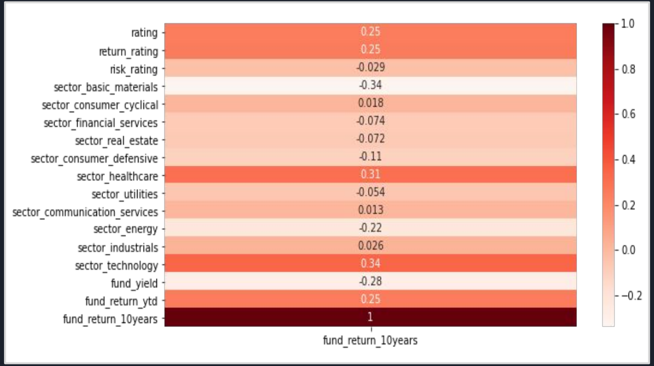


<u>Discussion:</u>

The qqplot is skewed left meaning that most of the data is distributed on the right side with a long “tail” of data extending out to the left. A normal distribution would be along the red line.

Blend Implementation

After splitting up the mutual funds by type, we ran a pearson correlation to determine which sectors had the most effect on the ten year return of the blend fund type mutual funds. To better visualize the correlations, we created a heat map, pictured left, of the correlations, where the darker colors indicate a higher positive correlation.



To better isolate the most correlated sectors, we created a table that listed the correlations between each sector and the blend type funds' ten year returns. Then, we selected the sectors with a correlation coefficient of at least (+/-) 0.3. For blend type funds, basic materials, healthcare, and technology were the most correlated to the ten year returns of the funds.

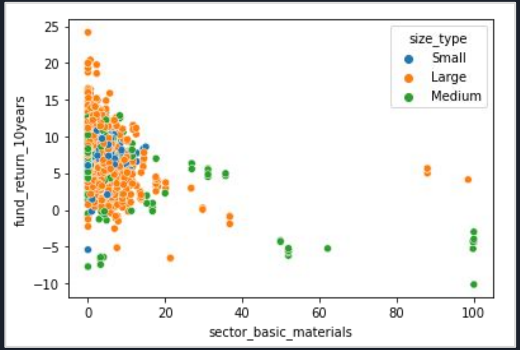
```
In [62]: #Correlation with output variable
         cor_target = abs(cor["fund_return_10years"]) # absolute value
         #Selecting highly correlated features
         relevant_features = cor_target[cor_target>0.3]
         relevant_features

Out[62]: sector_basic_materials    0.338678
         sector_healthcare         0.307685
         sector_technology         0.343153
         fund_return_10years       1.000000
         Name: fund_return_10years, dtype: float64
```

Blend Implementation

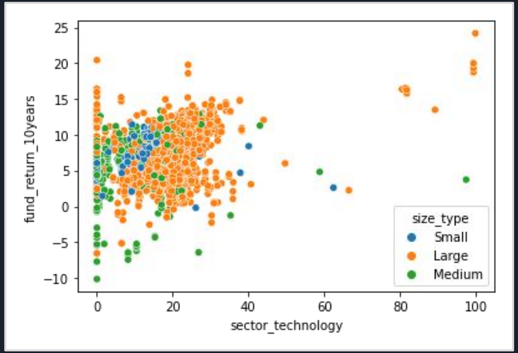
Before running linear regression, we plotted scatter plots with sector on the x-axis, ten year fund return on the y-axis, and colored the points by fund size to confirm there was a linear relationship between sector and fund return.

All three scatter plots are pictured on the right. Healthcare and technology had positive linear trends, whereas basic materials had a negative linear trends.

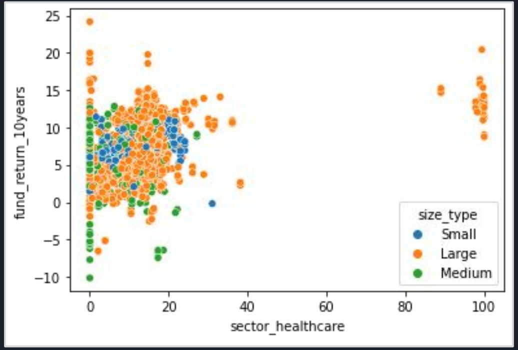
sector_basic_materials vs. fund_return_10years



sector_technology vs. fund_return_10years



sector_healthcare vs. fund_return_10years



Blend Regression Models: Healthcare

10 year mutual fund returns vs. healthcare sector + size type

```
practice_model_healthcare = ols('fund_return_10years ~ sector_healthcare + size_type + 0', data=Blend_train).fit()
print(practice_model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     fund_return_10years   R-squared:                       0.100
Model:                             OLS   Adj. R-squared:                  0.099
Method:                  Least Squares   F-statistic:                     82.94
Date:                 Thu, 06 May 2021   Prob (F-statistic):           6.90e-51
Time:                         14:04:21   Log-Likelihood:                -5824.9
No. Observations:                 2250   AIC:                         1.166e+04
Df Residuals:                     2246   BIC:                         1.168e+04
Df Model:                            3
Covariance Type:             nonrobust
=====================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
size_type[Large]      5.9111      0.124     47.695      0.000       5.668       6.154
size_type[Medium]     5.6280      0.193     29.209      0.000       5.250       6.006
size_type[Small]      6.3704      0.236     26.956      0.000       5.907       6.834
sector_healthcare     0.1029      0.007     14.756      0.000       0.089       0.117
==============================================================================
Omnibus:                      103.129   Durbin-Watson:                   1.980
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              340.980
Skew:                          -0.064   Prob(JB):                     9.06e-75
Kurtosis:                       4.903   Cond. No.                         56.9
==============================================================================
```

Blend Regression Models: Basic Materials

10 year mutual fund returns vs. basic materials sector + size type

```
practice_model_basic = ols('fund_return_10years ~ sector_basic_materials + size_type + 0', data=Blend_train).fit()
print(practice_model_basic.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     fund_return_10years   R-squared:                       0.120
Model:                             OLS   Adj. R-squared:                  0.119
Method:                  Least Squares   F-statistic:                     102.1
Date:                 Thu, 06 May 2021   Prob (F-statistic):           5.66e-62
Time:                         14:10:40   Log-Likelihood:                -5799.2
No. Observations:                 2250   AIC:                         1.161e+04
Df Residuals:                     2246   BIC:                         1.163e+04
Df Model:                            3
Covariance Type:             nonrobust
==========================================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------
size_type[Large]           8.0651      0.088     91.963      0.000       7.893       8.237
size_type[Medium]          7.1218      0.194     36.722      0.000       6.742       7.502
size_type[Small]           8.1346      0.228     35.749      0.000       7.688       8.581
sector_basic_materials    -0.1643      0.010    -16.571      0.000      -0.184      -0.145
==============================================================================
Omnibus:                       64.086   Durbin-Watson:                   1.961
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              158.503
Skew:                           0.002   Prob(JB):                     3.82e-35
Kurtosis:                       4.300   Cond. No.                         27.8
==============================================================================
```

Blend Regression Models: Technology

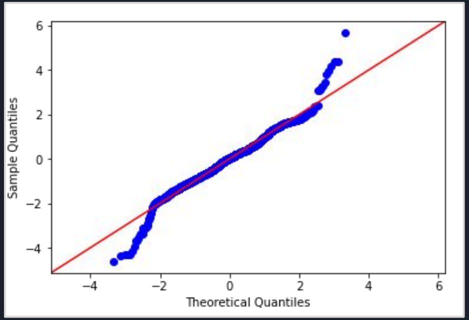
10 year mutual fund returns vs. technology sector + size type

```
practice_model_technology = ols('fund_return_10years ~ sector_technology + size_type + 0', data=Blend_train).fit()
print(practice_model_technology.summary())
```

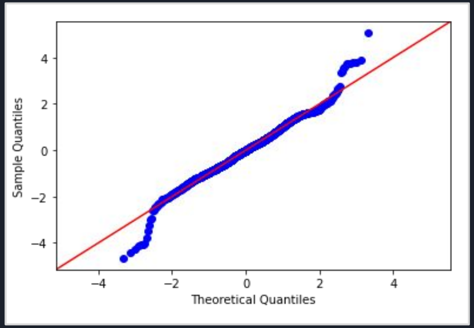
```
                            OLS Regression Results
==============================================================================
Dep. Variable:     fund_return_10years   R-squared:                       0.122
Model:                             OLS   Adj. R-squared:                  0.121
Method:                  Least Squares   F-statistic:                     104.3
Date:                 Thu, 06 May 2021   Prob (F-statistic):           3.15e-63
Time:                         14:16:11   Log-Likelihood:                -5796.3
No. Observations:                 2250   AIC:                         1.160e+04
Df Residuals:                     2246   BIC:                         1.162e+04
Df Model:                            3
Covariance Type:             nonrobust
=====================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
size_type[Large]      4.7506      0.172     27.562      0.000       4.413       5.089
size_type[Medium]     5.1004      0.198     25.794      0.000       4.713       5.488
size_type[Small]      5.8074      0.242     23.964      0.000       5.332       6.283
sector_technology     0.1343      0.008     16.767      0.000       0.119       0.150
==============================================================================
Omnibus:                      154.323   Durbin-Watson:                   1.989
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              381.110
Skew:                          -0.399   Prob(JB):                     1.75e-83
Kurtosis:                       4.852   Cond. No.                         76.3
==============================================================================
```

Blend Results

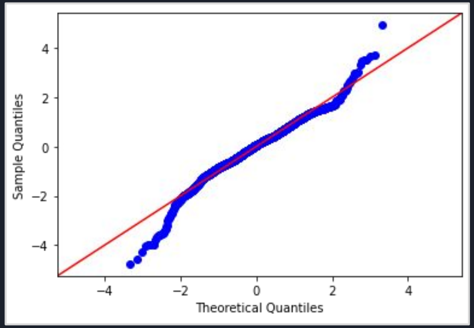
QQplot for residuals of linear regression model (sector_healthcare) versus (fund_return_10years)



QQplot for residuals of linear regression model (sector_basic_materials) versus (fund_return_10years)



QQplot for residuals of linear regression model (sector_technology) versus (fund_return_10years)



In all three QQ plots, the points stay close the line, meaning that the models' predicted data were very similar to the actual data. However, the models strayed from normal distribution in the cases where the funds were made up of very large or very small percentages of the healthcare, basic materials, or technology sectors.

Conclusion/Reflection:

General Takeaways:

The technology and healthcare sectors were shown to have significant positive relationships with 10 year returns for two of the investment types while technology claimed the highest correlation of any sector when paired with 10 year mutual fund returns in the Growth strata.

Another takeaway from this research is that when evaluating a mutual fund for the purpose of retirement, it is very important to separate the funds by investment type(Growth, Value, Blend). This proved to be essential since each investment type had different sectors and different fund sizes that were most highly correlated to the success of the fund over 10 years.

Surprisingly there doesn't appear to be a mutual fund size that provides the best overall 10 year returns. The optimal fund size was dependent on the investment strategy as well as the sector of the economy.

Further Work/Research:

Based on the visuals we created that illustrate various sectors of the economy plotted against 10 year mutual fund returns having gaps in percentage asset diversification, binning of these numeric variables into low, medium, and high categories might present an interesting train of analysis. For similar reasons clustering or other classification algorithms could be applied to the data we worked with.