Kaodi Okafor | Nhi Nguyen | Meet Patel | Pradyumn Pottapatri | Anesia McCarthy

Final Project Report

I. Web Scraping

For web scraping, our team used the python language to gather all the information we needed into dataframes. The dataframes were then transferred into SQL where our queries were later built.

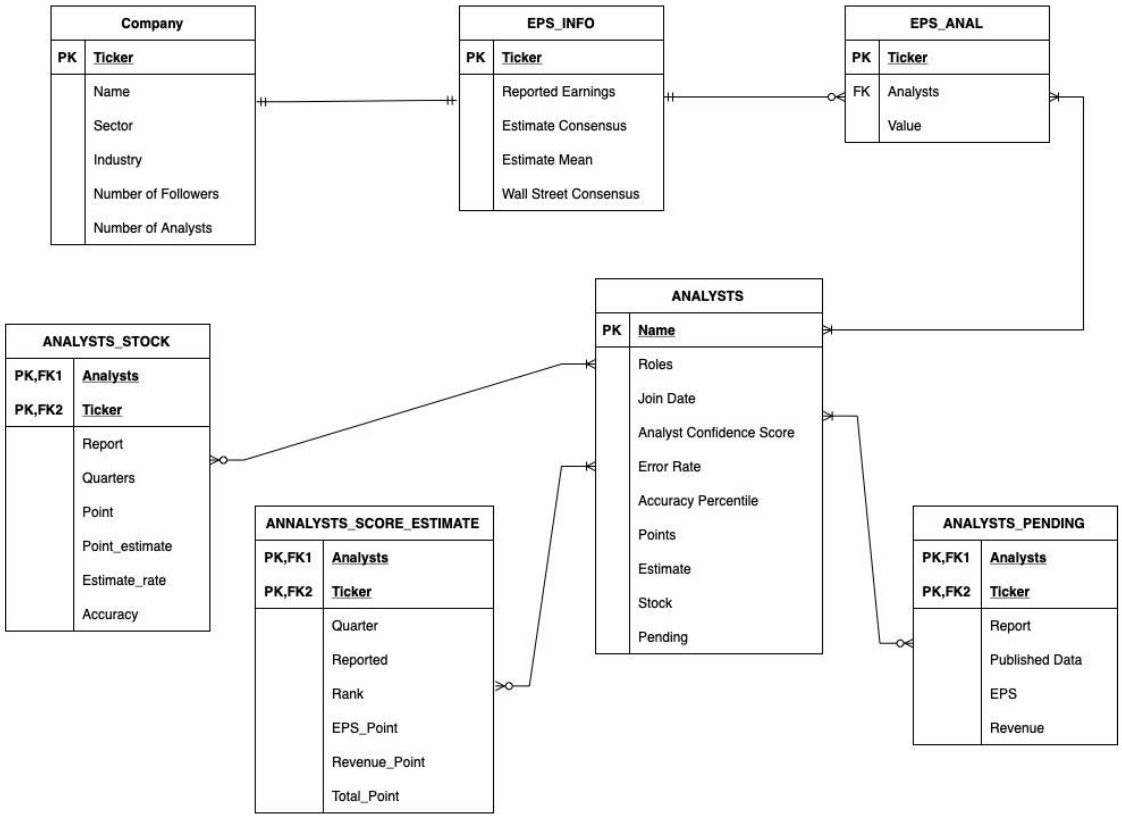
(a) When it came to part a, the major problem was scraping the data for one year. It was a challenge trying to figure out which year had the most data for all the tickers we selected as we wanted to have enough information at our disposal. The tickers we selected were [300 – 350] from excel sheet (GNC - GPS).

(b) There was not much of a challenge when it came to providing the earnings per share information. Since most of the web scraping was done in part a, it was pretty straight forward to grab the html information. However, there were instances when some of the elements were not present such as wall street consensus for a few companies but after they were identified, they were added empty blanks for the same. As for the EPS for the analysts, their names were not always shown completely, for which the ‘show more’ button had to be located and clicked properly. It took some time to understand the functioning of the show more buttons and how it is changing dynamically. After we understood the functioning, we decided to use 'sendkeys()' to click on the ‘show more’ button first.

(c) To extract the analyst's name for part c, the analyst username had to be lower case for all to extract information. The same problem that occurred in part b when it came to the analysts also occurred here. The names of the analysts were not showing correctly. The ‘show more’ button had to be retrieved. In the case where an analyst didn’t have some of the required elements ‘try’ and ‘except’ statements were used.

II. Database Model

After scraping all data from the Estimize website, we build a schema in SQL named EPS and import all required data. The schema is illustrated by the figure below.



As shown in the figure, we have total seven tables represent our database:

- EPS_INFO: includes Ticker as the primary key and the analysis/estimate numbers
- Company: includes information about the company
- EPS_ANAL: shows which analysts do the estimation for given Ticker
- Analyst: information about all analysts in our data
- Analysts_Stock: shows which stock covered that analysts estimate
- Anlaysts_Score_Estimate: shows information about ticker's score estimates for given analysts.
- Analysts_Peding: all pending estimation for given analysts.

The reason that we choose SQL to build our database is that it is easier for scientists to manipulate and extract the required information. Also, with organized structure, it is easy to show relationships between different information/data included in our database.

III. Analysis

- EPS_INFO

  EPS_INFO table has 204 rows (51 tickers x 4 quarters). Summary of EPS_INFO table shown below:

```
EPS_INFO.isnull().sum()
```

```
Ticker                0
Reported_Earnings     4
Estimate_Consensus    19
Estimate_Mean         19
dtype: int64
```

	Reported_Earnings	Estimate_Consensus	Estimate_Mean
count	200.000000	185.000000	185.000000
mean	0.755000	0.880108	0.909730
std	1.108205	1.020985	1.038898
min	-5.600000	-1.860000	-1.850000
25%	0.190000	0.250000	0.260000
50%	0.590000	0.630000	0.620000
75%	1.212500	1.230000	1.230000
max	5.080000	4.780000	5.110000

  As we can see from the summary tables, Reported Earnings has 4 null values while both Estimate Consensus and Estimate Mean have 19 null values in total. The distribution of Estimate Consensus and Estimate Mean are very similar to each other, but Reported Earning's distribution is more skewed to the left, which lead to the difference between Estimate Consensus and Reported Earnings being significant.

- EPS_ANAL

  EPS_ANAL table has 3302 rows with no null values for all columns. Summary of EPS_ANAL table shown below:

```
EPS_ANAL.isnull().sum()
```

```
Ticker     0
Analyst    0
Value      0
dtype: int64
```

	Value
count	3302.000000
mean	1.094079
std	1.065662
min	-2.050000
25%	0.390000
50%	0.810000
75%	1.570000
max	10.000000

  The distribution of EPS_ANAL is skewed right.

- Company

  The Company table has 51 rows which includes information about the company of our 51 tickers; thus, the Company table has no null value. All of the companies are from the Specialty Retail industry and Consumer Discretionary sector. Summary of can be found below.

```
Company.isnull().sum()
```

```
Ticker       0
Name         0
Sector       0
Industry     0
Followers    0
Analysts     0
dtype: int64
```

	Followers	Analysts
count	51.000000	51.000000
mean	162.274510	254.784314
std	232.488028	311.981494
min	6.000000	5.000000
25%	37.500000	52.000000
50%	62.000000	146.000000
75%	251.000000	331.000000
max	1398.000000	1807.000000

  As shown in the 'describe' table, the minimum number of followers and analysts is 6 and 5, respectively. Those stats belong to Murphy USA, Inc. (MUSA). The company has the highest number of followers and analysts (1398 and 1807 respectively) is The Home Depot, Inc. (HD).

- ANALYSTS

  The analysts table has information of 276 analysts where four of them have no Analyst_Confidence_Score. An example of the analysts table is shown below.

	name	roles	Join_Date	Analyst_Confidence_Score	error_rate	Accuracy_Percentile	points	estimate	stocks	pending
0	Bill	Non Professional Financials Professional Services	['Jul 2014']	8.4	16.5%	59%	25,430	13.6	1,870	35
1	Sentinel	Financial Professional Independent Independent...	['Aug 2018']	6.6	-	-	-	-	-	0
2	Frederick Tremblay	Financial Professional Buy Side Asset Manager	['Apr 2014']	7.9	-	-	-	-	-	0
3	Analyst_2855865	Non Professional Health Care Biotechnology	['Feb 2015']	6.2	-	-	-	-	-	0
4	Analyst_4146227	Non Professional Other Other	['Apr 2019']	8.0	-	-	-	-	-	0
5	Buck_446488	Financial Professional Sell Side Broker	['Oct 2018']	4.1	19.4%	4%	-36	-18	2	1
6	Buck_138258	Financial Professional Sell Side Broker	['Aug 2015']	6.4	-	-	-	-	-	0
7	Stephanie Benjamin	Financial Professional Sell Side Broker	['Nov 2017']	4.5	35.8%	3%	-302	-33.6	10	8
8	Proximilar AI	Financial Professional Independent Other	['Jan 2017']	8.6	-	-	-	-	-	0
9	Michael Ward, CFA	Financial Professional Sell Side Broker	['Aug 2016']	3.8	64.3%	14%	-68	-27.2	3	0

As we can see from the table, analysts who have no stock estimate (stocks column has no value) will result in no information for error_rate, accuracy_percentile, points, and points/estimate. Subsequently, there are a lot of null values in our analysts table since 194 analysts have no stocks estimate. Nevertheless, the average of analysts' confidence score is 6.8 (maximum is 9.6 and minimum is 1.3).

- ANALYSTS_STOCK

  The analysts_stock table has 3279 rows with no null values. Summary of analysts_stock table is illustrated below.

```
anal_stock.isnull().sum()
```

```
Analysts           0
Ticker             0
Report             0
Quarters           0
Points             0
Points_Estimate    0
Estimate_rate      0
Accuracy           0
dtype: int64
```

	Quarters	Points	Points_Estimate	Estimate_rate	Accuracy
count	3279.000000	3279.000000	3279.000000	3279.000000	3279.000000
mean	6.934736	70.655078	8.489600	14.212168	49.535529
std	6.603428	119.826485	12.985847	14.030320	18.989927
min	0.000000	-298.000000	-50.000000	0.000000	0.000000
25%	2.000000	3.000000	1.050000	4.400000	38.000000
50%	4.000000	33.000000	10.000000	8.900000	54.000000
75%	10.000000	105.000000	17.000000	19.800000	63.000000
max	33.000000	665.000000	50.000000	100.000000	100.000000

  The highest Accuracy is 100% and the lowest is 0% while the average Accuracy of our analysts is about 49.54%.

- ANALYSTS_SCORE_ESTIMATE

  Analysts_score_estimate table has 3567 rows. Summary of the table is shown below.

```
anal_score.isnull().sum()
```

```
Analysts          0
Ticker            0
Quarter           0
Reported          0
Rank              0
EPS_Points       13
Revenue_Points  442
Total_Points      0
dtype: int64
```

	EPS_Points	Revenue_Points	Total_Points
count	3554.000000	3125.000000	3567.000000
mean	1.967642	1.574720	3.340062
std	12.494359	13.156083	20.406130
min	-25.000000	-25.000000	-50.000000
25%	-6.000000	-6.000000	-9.000000
50%	-2.000000	-2.000000	5.000000
75%	12.000000	12.000000	19.000000
max	25.000000	25.000000	50.000000

  From the summary tables, EPS_Points column has 13 null values, and Revenue_Points has 442 null values. Moreover, the highest Total_points that analysts get is 50 while the

minimum Total_points is -50. EPS_Points and Revenue_points have almost the same distribution because their 0, 25, 50, 75, and 100 percentiles take the same values.

- ANALYSTS_PENDING

  Analysts_pending table has 461 rows. Summary information of this table can be found below.

```
anal_pen.isnull().sum()
```

```
Analyst             0
Ticker              0
Reports             0
Published Data      0
EPS                 2
Revenue           121
dtype: int64
```

	EPS	Revenue
count	4.590000e+02	340.000000
mean	-4.357298e+15	11225.242794
std	6.593773e+16	25893.661732
min	-1.000000e+18	12.810000
25%	5.000000e-01	713.467500
50%	1.200000e+00	1822.000000
75%	2.515000e+00	7683.500000
max	1.370000e+02	150174.000000

  As we can see from the summary table, EPS column has 2 null values, and Revenue has 121 null values. All other columns do not include any null value. The maximum EPS from our dataset is 137 while the minimum is -1. Similarly, the maximum of Revenue is 150174, and the minimum of Revenue is 340.

IV. Regression

We combined the files using the concat and merge methods for combining the data between all the excel sheets.

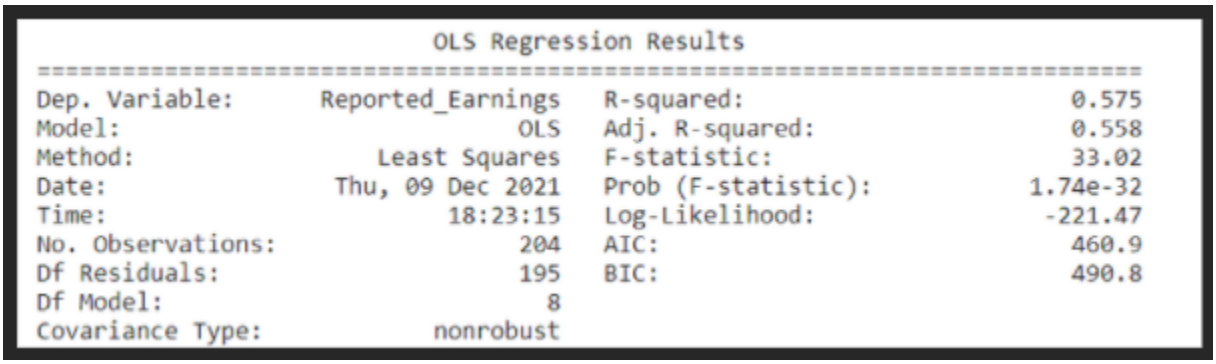
Steps for implementing the Regression Model:

- Importing the data
- Removing NaN and merging files
- Labelling the categorical columns accordingly
- Fitting a Multiple Linear Regression model and Ordinary Least Squares for fitting regression model
- Calculating the R2 Score and Mean Squared Error

Insights and Analysis:

We implemented a Multiple Linear Regression Model and OLS Regression model using the Wall Street Consensus, Estimize Mean and Estimize consensus. From the Multiple Linear regression model, we have received an R2 score of 48% which specifies that the MLR is able to successfully estimate the actual values for about 48% of the time. And there is a 42% error rate between the actual and the predicted value of the regression from the mean squared error.

From the OLS Regression model, we see that R2 score is 57% for which the model is able to correctly predict the values about 57% of the time.



OLS Regression Results

Dep. Variable:	Reported_Earnings	R-squared:	0.575
Model:	OLS	Adj. R-squared:	0.558
Method:	Least Squares	F-statistic:	33.02
Date:	Thu, 09 Dec 2021	Prob (F-statistic):	1.74e-32
Time:	18:23:15	Log-Likelihood:	-221.47
No. Observations:	204	AIC:	460.9
Df Residuals:	195	BIC:	490.8
Df Model:	8		
Covariance Type:	nonrobust		

Since we have extremely diverse estimates for predicting the values of the reported earnings, we can create a new estimate using the wall street consensus, estimate consensus and estimate mean which would lead to a new estimate and it will be helpful to better analyse the regression parameters.