ANOVA Lab Report
Xinyi (Anny) Cui

1 Introduction

This paper explores the top 2000 companies in 2017 as listed on Forbes.com and investigates whether there is a difference in the sales between sectors to gauge the relative strengths of different economic sectors. The analysis section includes the data exploration and inference explanation. The conclusion introduces the relative strengths of different economic sectors

2 Analysis

2.1 Data Exploration

X	Rank	Company	Country	Sales	Profits	Assets	Market.Value	Sector	Industry
N/A	1	ICBC	China	151.4	42.0	3473.2	229.8	Financials	Major Banks

Table 1: The example of data format

Before we do the analysis of variance, we need to do the data exploration to see whether we need to re-express the data. After creating the boxplot (see Figure 1) and densityplot (see Figure 2) of sales based on sectors, all data of sales based on sectors approaching to the bottom and have high-end points. We then need to construct a linear model of log mean values and log standard deviation values of sales based on sectors for more checking on what kind of transformation we need to do. The value of slope is 1.0143 which means we need to do log transformation on the data of sales. The new boxplot (see Figure 3) of log sales based on sectors is better normally distributed. There exists a difference in the sales between sectors. The materials sector has the lowest mean value of sales, 13.67, and the Energy sector has the highest mean value of sales, 30.11.
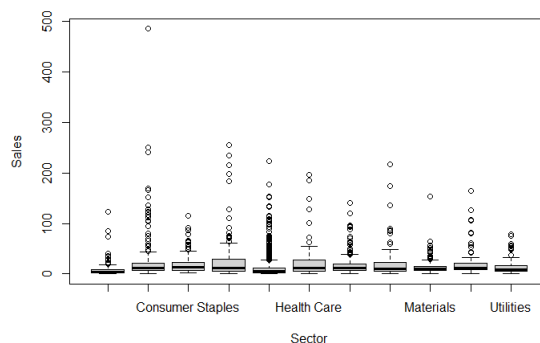

Figure 1: Boxplot of sales based on sectors. The data is approaching to the bottom and has points at high end.
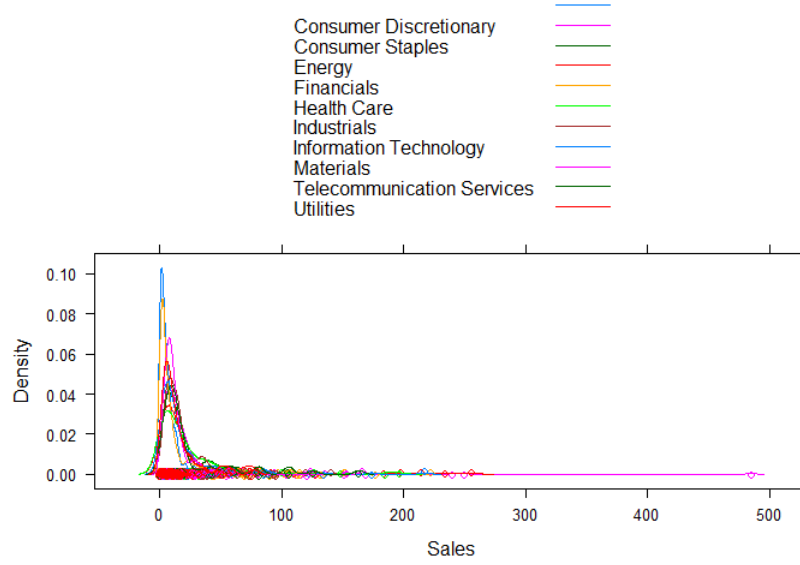
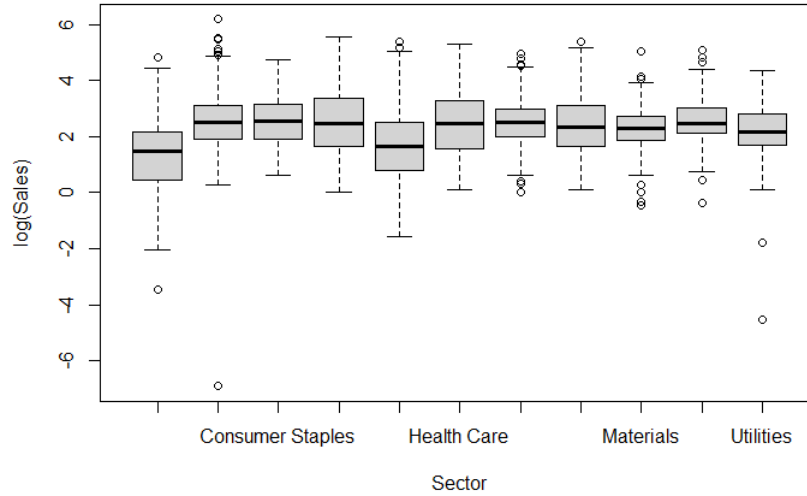Figure 2: Densityplot of sales based on sectors. The data is left-skewed.

Figure 3: Boxplot of log sales based on sectors. The data is better normally distributed.

2.2 ANOVA Analysis

After doing the analysis of variance, the p-value is $< 2^{-16}$ which is almost 0, and the F-statistic is 27.87, seems a good value. There is statistically significant evidence that the different sectors do not all have the same mean value of log sales. The log sales of different sectors statistically significantly differ from others. This helps us infer that the values of sales differs between the ten sectors. The graph of assessment (see Figure 4) shows that the variances do not seem to be hugely different from one another.
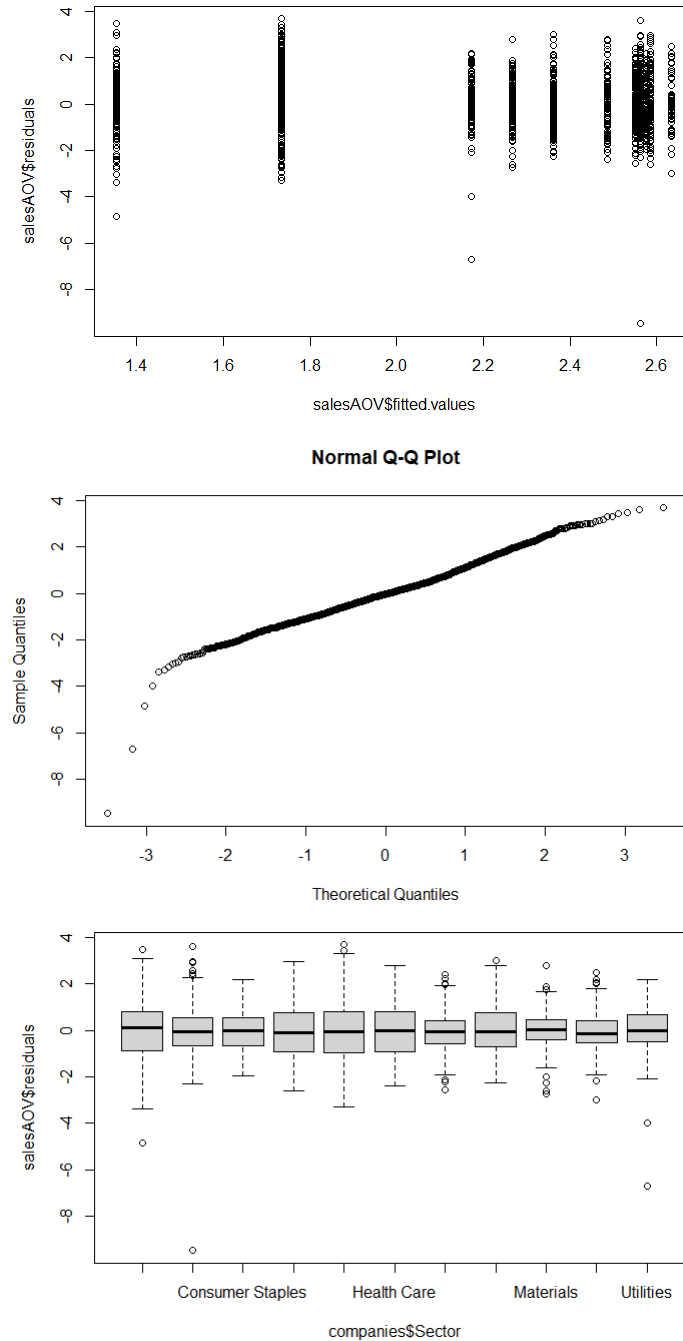
2

Figure 4: The graphs of ANOVA assessment. The first is fitted values v.s. residuals. The second is the normal quantile-quantile graph. The third one is the boxplot of residuals.

2.3 Tukey HSD Analysis

Only the differences between Financials and Consumer Discretionary (p-adjust = 0.0000166), Materials and Consumer Discretionary (p-adjust = 0.0022314), Financials and Energy (p-adjust = 0.0000046), Industrials and Energy (p-adjust = 0.0424419),

Materials and Energy (p-adjust = 0.0001544), Utilities and Energy (p-adjust = 0.0044539) are statistically meaningful because their p-values are smaller than 0.05. From those comparisons, the differences between Energy with most other sectors (Financials, Industrials, Materials, and Utilities) are statistically meaningful.

3   Conclusion

The sector Energy has the highest coefficient and the sector Financials has the lowest coefficient statistically (see Table 2). Also, the sector Energy has the highest average value of sales and the sector Financials has the lowest average value of sale statistically relatively (see Table 3). Thus, the sector Energy is relatively stronger than the sectors of Financials, Industrials, Materials, and Utilities. the sector Consumer Discretionary is relatively stronger than the sector of FInancials and Materials. In short, the sector Energy is the best performed sector in the top 2000 companies in 2017 as listed on Forbes.com.

Sectors	Converted Coefficient (m)
Consumer Discretionary	3.349858
Consumer Staples	3.395915
Energy	3.427806
Financials	1.460619
Health Care	3.104434
Industrials	3.305856
Information Technology	2.736672
Materials	2.493331
Telecommunication Services	3.595779
Utilities	2.269057

Table 2: The values of slopes for each sector after converting the coefficient back. The formula form is $Sales = 3.874202 \times m^{sector}$.

Sectors	SDs	means	medians	Number
Consumer Discretionary	46.32438	25.415194	12.50	237
Consumer Staples	21.01018	20.291304	13.00	115
Energy	48.30296	30.108182	12.05	110
Financials	24.70040	14.028383	5.30	583
Health Care	35.48458	24.146154	11.60	91
Industrials	20.59047	19.003828	12.10	209
Information Technology	30.26140	19.596154	10.15	130
Materials	15.36732	13.671040	9.75	174
Telecommunication Services	34.65686	26.116868	12.00	53
Utilities	15.74264	14.578980	8.80	101

Table 3: The table of sales values of standard deviation, average, and median for each sector. The last column displays the number of each sector in the Forbes list.