MA334 Individual Project, Spring 2022

Introduction to dataset

Forbes2000 is a list comprising of large companies working worldwide. The list is published on regular basis for the acknowledgement of the profitable companies with large number of employees. Other parameters considered for the completion of Forbes2000 selection are sales, assets and market value. All of these parameters are given as column names in the dataset used for analysis in next sections. Moreover the dataset also includes the company name, category of products provided and the countries they belong to originally. The example dataset we use here is given on this link. We have selected this dataset since it includes a reasonable number of rows and columns for the statistical analysis. Although we can include all the columns for our analysis yet we are going to keep only three categorical variable which will be helpful for the boxplot in the data visualization section of our report.

Overview of the Variables

Description of our example data set `Forbes2000` is given in the table below. The data set has the shape (2000, 8which basically illustrates that it has 2000 rows and 8 columns. The following table also shows the variable types.

Table 1: Description of dataset

Sr. #	Column name	Unit	Variable type
1	rank	Number	numeric
2	name	-	character
3	country	-	character
4	category		character
5	sales	USD (Billions)	numeric
6	profits	USD (Billions)	numeric
7	assets	USD (Billions)	numeric
8	marketvalue	USD (Billions)	numeric

Other summary statistics about the dataset are given below

```
#>       rank           name              country           category
#>  Min.   :   1.0   Length:2000        Length:2000        Length:2000
#>  1st Qu.: 500.8   Class :character   Class :character   Class :character
#>  Median :1000.5   Mode  :character   Mode  :character   Mode  :character
#>  Mean   :1000.5
#>  3rd Qu.:1500.2
#>  Max.   :2000.0
#>
#>      sales            profits             assets          marketvalue
#>  Min.   :  0.010   Min.   :-25.8300   Min.   :   0.270   Min.   :  0.02
#>  1st Qu.:  2.018   1st Qu.:  0.0800   1st Qu.:   4.025   1st Qu.:  2.72
```

```
#>  Median :  4.365   Median :  0.2000   Median :   9.345   Median :  5.15
#>  Mean   :  9.697   Mean   :  0.3811   Mean   :  34.042   Mean   : 11.88
#>  3rd Qu.:  9.547   3rd Qu.:  0.4400   3rd Qu.:  22.793   3rd Qu.: 10.60
#>  Max.   :256.330   Max.   : 20.9600   Max.   :1264.030   Max.   :328.54
#>                    NA's   :5
```

Boxplot and quartile check

Boxplot is an effective way to show quartile and median. For the ease of analysis we are selecting only 3 countries and 3 categories for boxplot. The quartiles (Q1 and Q3) are shown by edges of the boxplot for each country separately.

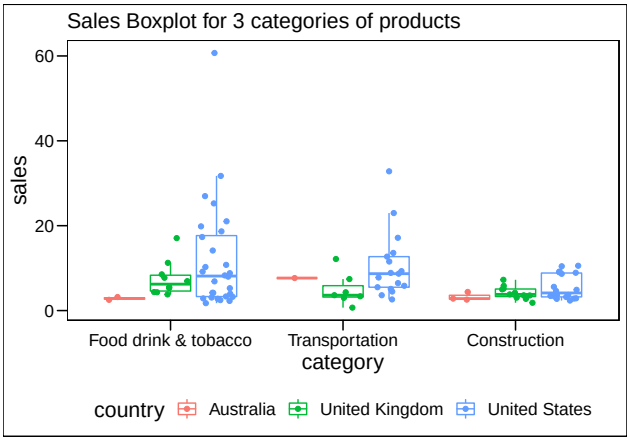

Figure 1: Sales evolution in 3 english Speaking countries

The boxplot in Fig. 1 shows that for 3 english speaking countries the mean value of sales are less than 20 billion USD while there are some spots with outliers. The outliers of sales are shown by points outside the boxplot. We conclude that for the category of Food and Drink the sales are much higher as shown by Q3 value in lowest boxpot. For the companies involved in construction the mean sales is less as compared to a company which involves Food and drink prodicts. One important observation for the above boxplots is the small value of variation of quartiles in Australia as compared to UK and USA. The underline reason behind small value of sales in all 3 categories for Australia can be represented with more analysis.

Skewness and kurtosis check

The skewness and kurtosis of the dataset is shown statistically by distribution Plots. For our dataset we are selecting 3 categories as above to plot the distribution of dataset. It helps us to determine if our data is normally distributed. The Fig. 2 shows that the sales are normally distrbuted with a mean value around zero. It also shows that out of 3 countries selected the none of the sales group is left of right skewed. Moreover the mean value can be checked with the help of `Shapiro wilk` test.

Shapiro Wilk Test

We will select the sales group with the 3 countries to run the shapiro wilk test. Before the Shapiro test we need the define the significance level for our tests since it is very basic principle of statistical tests to check our hypothesis according to the confidence level chosen. It is very common in the scientific community to select the 95% confidence level or an alpha value of 0.05 to check if there is significant difference between mean values of two observations. We have to check if the above plotted graph actually shows the correct results that sales are normally distributed so for this purpose we will perform the shapiro wilk test. It is

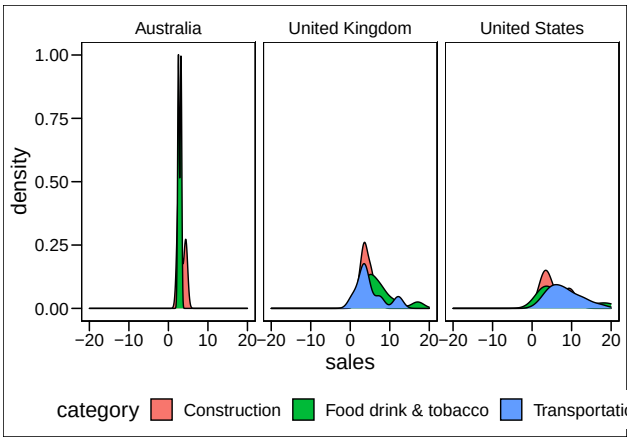

Figure 2: Normal Distrbution check of 'sales'

essential to define the null hypothesis since the output of the Shapiro wilk test provides p-value linked to acceptance or rejection of hypothesis. For the sake of simplicity we take confidence level = 95% and at the alpha value of 0.05 our hypothesis are as follows

- Null hypothesis: H_0 = The Sales for the 2 countries is normally distributed
- Alternate hypothesis: H_1 = The sales of the 2 countries is not normally distributed

The two hypothesis will be checked with the help of p-value. If the result of Shapiro wilk test gives us p-value < 0.05 which is less than $\alpha = 0.05$ we will reject our null hypothesis and accept the alternate hypothesis. We will conduct the Shapiro wilk test for 2 countries separately.

- For USA

```
#>           sales
#> statistic 0.4203573
#> p.value   1.209563e-43
#> method    "Shapiro-Wilk normality test"
#> data.name "X[[i]]"
```

- For UK

```
#>           sales
#> statistic 0.3558951
#> p.value   5.612034e-22
#> method    "Shapiro-Wilk normality test"
#> data.name "X[[i]]"
```

We observe that the p-value in all 3 countries is less than 0.05 hence we reject our null hypothesis that the sales in 2 countries is normally distributed.

Now we will the conduct the non-parametric test now since our filtered data is not normally distributed. The non-parametric test involve the Wilcoxon-rank sum test for our example since we are going to compare 2 countries for sales comparison. The other choice would have been t-test if the `sales` group had p-value $>$ α .

Wilcoxon-Rank sum test

We select UK and USA for our non parametric test as in both groups the sales are independent of each other. Our assumption revolves around p-value again which in turn is linked to null and alternate hypothesis. Our two hypothesis are;

- H_0 : Sales in UK are equal to the sales in USA.
- H_1 : Sales in UK are not equal to USA.

```
#>
#>  Wilcoxon rank sum test with continuity correction
#>
#> data:  wilcox$sales by wilcox$country
#> W = 55092, p-value = 0.1864
#> alternative hypothesis: true location shift is not equal to 0
```

The result of Wilcoxon sum test indicate that at 5% significance level p-value > 0.05 hence there exists no statistically significant difference between the sales of UK and USA. In other words we accept our null hypothesis H_0 and reject alternate hypothesis H_1.

The result of the test can be shown on boxplot

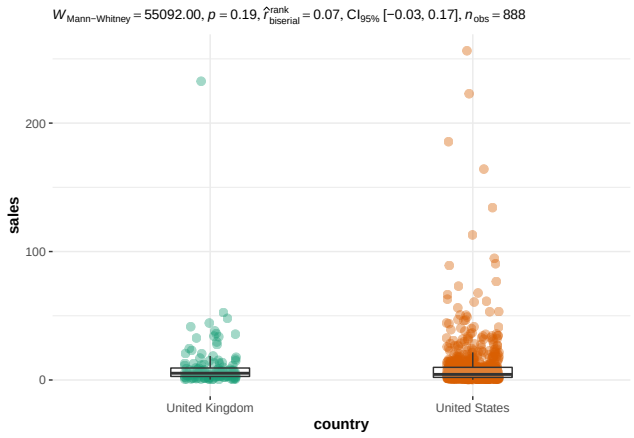

Figure 3: Wilcoxon test plotting

Corelation plot and Regression Analysis

The correlation plot analyses if there is any correlation between different variables in our dataset. The figure below shows that there is high correlation between market value and sales as compared to sales and profit. Similarly the companies which have high market value are most likely to get more profits according to correlation coefficient value of 0.55. The correlation plot also shows that almost all the variables are statistically significant to each other since we have high number of stars given below according to statistical labeling technique. One more aspect in the corelation plot is the multiple regression lines given on the bottom left quarter. It shows that the variables for which there is high correlation value (such as profits and market value with correlation value of 0.64) , there is very high chance of linear incremental trend.

```
#>
#> Call:
#> lm(formula = sales ~ marketvalue, data = df)
```

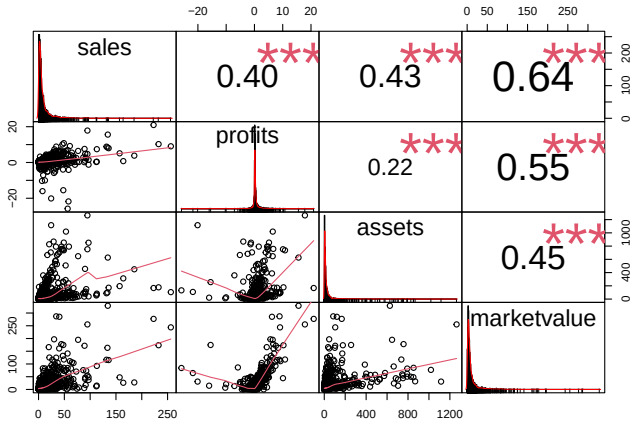

Figure 4: Corelation plot between numeric variables

```
#>
#> Residuals:
#>     Min      1Q  Median      3Q     Max
#> -105.44   -4.55   -2.64    1.17  168.46
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept)  4.08425    0.34318   11.90   <2e-16 ***
#> marketvalue  0.47255    0.01262   37.43   <2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 13.81 on 1998 degrees of freedom
#> Multiple R-squared:  0.4122, Adjusted R-squared:  0.4119
#> F-statistic:  1401 on 1 and 1998 DF,  p-value: < 2.2e-16
```

The p-value for the `market value` is very less than 0.05 which indicates that it gives us a reliable guess about the sales of the company. Multiple and adjusted R-squared for us are > 0.4 which means that `market value` can indicates 41% of sales generally.

Summary

`Forbes2000` dataset was chosen for the statistical and regression analysis for assignment. Out of many categorical and numerical variables we chose only one numerical (sales) and two categorical variables (UK and USa) for the statistical tests. It is reported by running various statistical tests that the sales is positively corelated with the market values of the companies given in the dataset. The result of shapiro wilk normality test indicates that the sales in UK, USA and Australia are not normally distributed. p-value is Wilcoxon sum test indicates that there exists significant difference between the sales of UK and USA for the companies included in `Forbes2000`. Correlation plot and Regression analysis indicates that there are many variables in the dataset which have high correlation with each other. Finally, the boxplot illustration shows that the sales in Australia are very less as compared to UK and USA. Furthermore, there are cases in all 3 countries when sales are much more than mean value as indicated by outliers on boxplot.

Appendix

```
knitr::opts_chunk$set(
  comment = "#>", echo = FALSE, out.width = "50%", out.height = "50%",fig.align="center"
)
```

```
# Importing libraries
library(readr)
library(tidyverse)
library(dplyr)
library(stats)
library(broom)
library(ggplot2)
library(ggpubr)
library(captioner)
library(PerformanceAnalytics)
library(ggstatsplot)
library(knitr)
```

```
table_captions <- captioner::captioner(prefix="Tab.")
figure_captions <- captioner::captioner(prefix="Fig.")

t.ref <- function(label){
  stringr::str_extract(table_captions(label), "[^:]*")
}

f.ref <- function(label){
  stringr::str_extract(figure_captions(label), "[^:]*")
}
```

```
df <- read_csv("~/Documents/R_data_Visualizations/Forbes2000.csv")

df <- df %>% select(-"...1")
```

```
attach(df)

head(round(prop.table(table(OFFICER_YEARS_ON_FORCE,OFFICER_GENDER),1)*100))
```

```
df %>% group_by(OFFICER_GENDER,OFFICER_YEARS_ON_FORCE) %>%
  summarise(number = n()) %>%
  pivot_wider(names_from = OFFICER_GENDER,values_from = number)
```

```
df %>%
 filter(category %in% c("Food drink & tobacco", "Transportation", "Construction")) %>%
  filter(country %in% c("United States","United Kingdom", "Australia")) %>%
  ggboxplot(x = "category", y = "sales", color="country",palette ="country", add = "jitter") +
 labs(subtitle = "Sales Boxplot for 3 categories of products") +
 ggthemes::theme_base()+ theme(legend.position = "bottom")
```

```
df %>%
 filter(category %in% c("Food drink & tobacco", "Transportation", "Construction")) %>%
  filter(country %in% c("United States","United Kingdom", "Australia")) %>%

 ggplot() +
  aes(x = sales, fill = category) +
  geom_density(adjust = 1L) +
  scale_fill_hue(direction = 1) +
  ggthemes::theme_base() +   facet_wrap(vars(country)) + theme(legend.position = "bottom")+
  xlim(-20, 20)
```

```
df %>%
filter(country %in% c("United States")) %>% select(sales) %>%
  sapply(.,shapiro.test)
```

```
df %>%
filter(country %in% c("United Kingdom")) %>% select(sales) %>%
  sapply(.,shapiro.test)
```

```
wilcox <- df %>%
filter(country %in% c("United Kingdom", "United States"))

wilc <-  wilcox.test(wilcox$sales ~ wilcox$country)

wilc
```

```
# plot with statistical results
ggbetweenstats( # independent samples
  data = wilcox,
  x = country,
  y = sales,
  plot.type = "box", # for boxplot
  type = "nonparametric", # for wilcoxon
  centrality.plotting = FALSE # remove median
)
```

```
df %>% select(-c("country","category","name","rank")) %>%
chart.Correlation(df, histogram = TRUE, method = "pearson")
```

```
model <- lm(sales ~ marketvalue, data=df)
summary(model)
```