**Investment Strategy**

Data have been collected from a random sample of companies from two sectors, retail and manufacturing.

**Project aims**

Your study is to find which variables give the best measures of the investment potential in these sectors and draw conclusions as to which types of stock would provide the best investment. Your investigation should consider whether this differs depending on size and type of company. You should assess the validity of any predictions that you make.

Note that you can decide on what measure makes a good investment. Don’t be concerned if you have no background in finance. You are not being tested on whether the measure that you choose is in fact the best method (or in fact a method that is used by investment managers).

**Task**

Your task is to perform relevant statistical analyses of a sample data set to answer a given question about the population under investigation. This data set should be sufficient, and there is no need to supplement with additional data.

\_\_\_\_\_Solution\_\_\_

**Task 1: Correlation Analysis**

**Description:**

This task focuses on understanding the relationships between various financial variables by calculating the correlation matrix. The correlation matrix provides a numerical measure of the strength and direction of linear relationships between pairs of variables. A correlation plot is generated using the corrplot package to visualize these relationships. This analysis helps identify which variables exhibit strong correlations, which is crucial for understanding potential investment factors.

**Task 2: Boxplot of Market Price by Category**

**Description:**

In this task, we explore the distribution of market prices based on a hypothetical categorical variable named 'Category'. The 'Category' variable is randomly assigned to each observation, creating three categories: 'Large', 'Medium', and 'Small'. A boxplot is then created using ggplot2 to compare the market prices across these categories. This visualization aids in identifying potential trends or differences in market prices based on the assigned categories.

**Task 3: Scatter Plot for Validity Assessment**

**Description:**

The goal of this task is to assess the validity of predictions made on the relationship between total sales and total market capitalization. A scatter plot is created using ggplot2 to visualize the distribution of total market capitalization (TotMktCap) against the sum of total sales in 2017 and 2018 (TotalSales17 + TotalSales18). Additionally, a red linear regression line is added to the plot. The linear regression line helps assess the validity of predictionsand understand the overall trend between these two variables.

**Exploratory Data Analysis (EDA):**

The exploratory data analysis (EDA) section provides a summary of the dataset using the summary function. Histograms and a boxplot are created to visualize the distribution of selected variables (MktPrice, DivYield, PERatio). A scatter plot using ggplot2 is generated to visualize the relationship between total sales in 2017 and 2018.

**Data Preprocessing:**

The code includes data preprocessing steps such as checking and converting data types, handling missing values by replacing them with column means, and ensuring that all missing values have been successfully filled.

**Plot Saving:**

Finally, the generated plots (correlation plot, boxplot, and scatter plot) are saved as PNG files using the ggsave function, allowing for easy sharing and documentation of the visualizations.

**Code Description:**

The code provides insights into the relationships between variables and helps draw conclusions about investment potential.

**Reading Excel File:**

The code uses the read\_excel function from the readxl package to read an Excel file.

**Handling Missing Values:**

The code checks the data type of the DivYield column using the class function.

It converts the values in the DivYield column to numeric using as.numeric.

Missing values are then filled with the mean of the respective column using a loop.

**Exploratory Data Analysis (EDA):**

The summary function provides summary statistics for the data.

Histograms and a boxplot are created for selected columns (MktPrice, DivYield, PERatio) using base R plotting functions.

A scatter plot of TotalSales17 vs. TotalSales18 is created using ggplot2.

**Task 1: Correlation Analysis:**

The correlation matrix is calculated for selected columns.

The corrplot function from the corrplot package is used to create a correlation plot.

**Task 2: Boxplot of Market Price by Category:**

A hypothetical grouping variable named Category is created.

A boxplot of MktPrice by Category is created using ggplot2.

**Task 3: Scatter Plot for Validity Assessment:**

A scatter plot of TotalSales17 + TotalSales18 vs. TotMktCap is created using ggplot2.

A red linear regression line is added to assess the validity of predictions.

**Saving Plots:**

The ggsave function is used to save the correlation, boxplot, and scatter plot as PNG files.

**Further details**

**Task 1: Find which variables give the best measures of the investment potential**

Answer:

The correlation analysis is performed on the financial variables (MktPrice, TotMktCap, DivYield, PERatio, Beta, TotalSales17, TotalSales18, CapEmp, Dividend, MktBook, Ret17, Ret18). The correlation plot visually represents the strength and direction of relationships between these variables. High positive or negative correlations indicate variables that might be good measures of investment potential.

**Task 2: Draw conclusions on which types of stock would provide the best investment**

Answer:

A hypothetical grouping variable named 'Category' is created randomly with values ("Large", "Medium", "Small"). A boxplot is then created based on this hypothetical 'Category' variable, comparing the market price (MktPrice) across different categories. This aids in drawing conclusions about which types of stocks (Large, Medium, Small) might offer better investment potential.

**Task 3: Assess the validity of any predictions made**

Answer:

A scatter plot is created to visualize the relationship between total sales (TotalSales17 + TotalSales18) and total market capitalization (TotMktCap). Additionally, a linear regression line is added to assess the validity of predictions about this relationship. If the regression line aligns well with the data points, it suggests that the predictions are more valid.

**Exploratory Data Analysis (EDA)**

Answer:

Summary statistics are provided for the filled dataset (data\_filled). Histograms are created for MktPrice, DivYield, and a boxplot for PERatio. These visualizations offer insights into the distribution and central tendencies of these variables.

The generated plots (correlation plot, boxplot, and scatter plot) are saved as images: "correlation\_plot.png," "boxplot.png," and "scatter\_plot.png," respectively.

In summary, the code performs exploratory data analysis and addresses each task by providing visualizations and insights into the relationships between financial variables, the impact of hypothetical categories on market prices, and the validity of predictions made about the relationship between total sales and total market capitalization.