Technical University of Denmark

Trading ETFs

Student :
Visnukaran Kirubakaran,
Student ID: s224527

Teacher :

October 29, 2024

Contents

1 Descriptive analysis

a) Description of the data material

In this project, data from the BMI study is analyzed, which aims to establish an appropriate multiple linear regression model for BMI. The provided dataset contains 847 observations, focusing on the following variables:

- ID: The respondent's number (can be used for identification).
- BMI: The respondent's BMI (in kg/m^2).
- Age: The respondent's age (in years).
- Fastfood: Frequency of the respondent's visits to fast-food restaurants (days/year).

Additionally, a variable such as log(BMI), which is the logarithm of the BMI, is also added. This variable simplifies later calculations. This is a complete sample, as all values that are given are included. All the mentioned variables are quantitative, as the data is numerical, and can both be quantified and measured. Of the aforementioned variables, three are selected: log(BMI), age, and fastfood. For these three variables, we can create scatter plots where log(BMI) is plotted against age and fastfood to investigate the correlation between them.

On the following scatter plot of log(BMI) over age, it appears that there is no real correlation between the two factors. This can be observed from the scattered data.

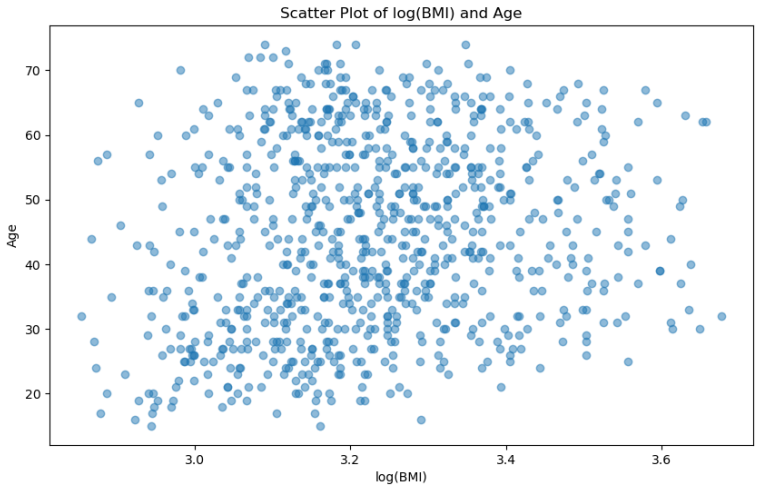

Figure 1: Scatterplot of log(BMI) and Age.

The same thing can be seen on the scatterplot of the log(BMI) and fastfood where it does not seem there is a correlation between BMI and fastfood, since the data is scattered.

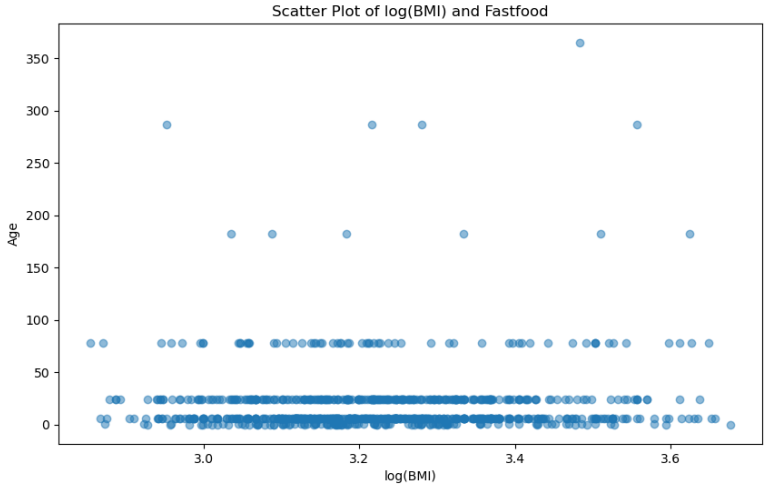

Figure 2: Scatterplot of log(BMI) and Fastfood.

Further, we undertake a graphical examination of each variable using histograms and box plots.

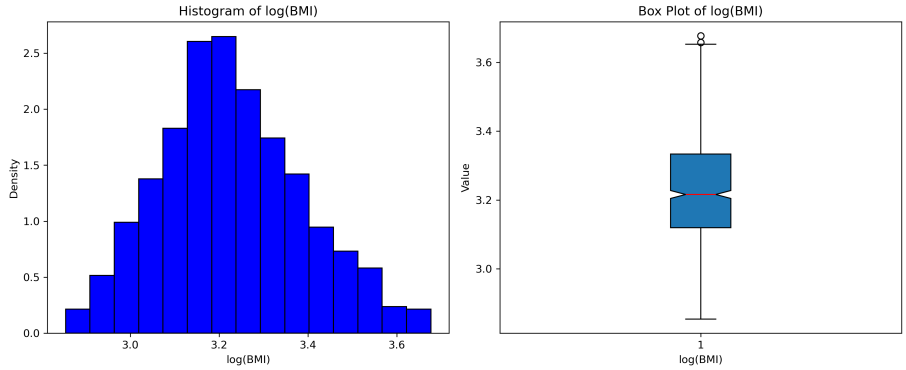

Figure 3: Histogram and boxplot of log(BMI).

The histogram for log(BMI) indicates a tendency towards a normal distribution with some skewness. The boxplot highlights a median slightly above 3.2, suggesting a slight skew towards lower values. Notably, a couple of outliers are present on the upper end, indicating some deviations from a typical normal distribution.

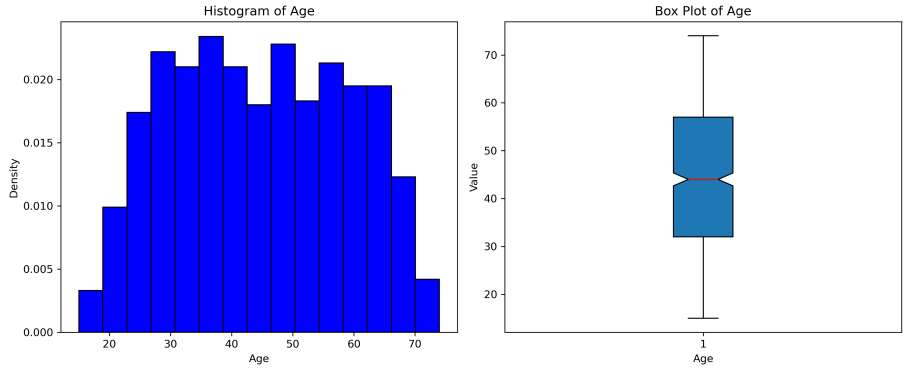

Figure 4: Histogram and boxplot of Age.

The histogram for age reveals a broad spread of values, indicative of high variability within the age data. The boxplot shows a median value around 45, nicely centered between the quartiles, reflecting a balanced distribution despite the wide range. The plot suggests a fairly symmetric distribution of age, although the range suggests varied participant ages.

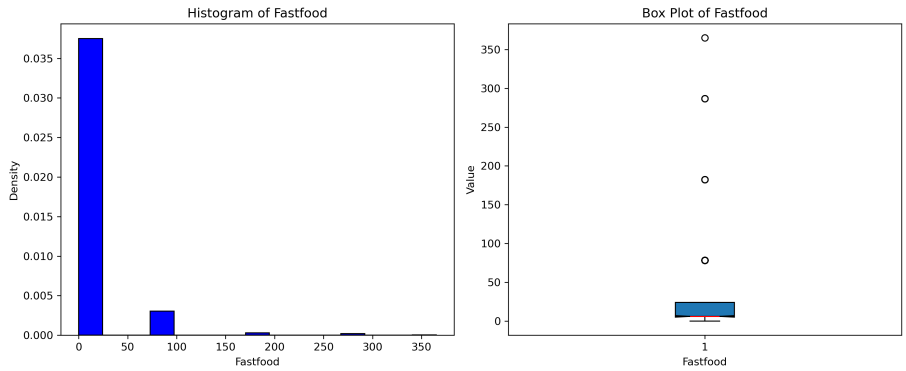

Figure 5: Histogram and boxplot of Fastfood.

The fastfood consumption histogram clearly shows a concentration of data points below 100 days, with rare occurrences extending beyond this range. The corresponding boxplot underscores this concentration, with the bulk of data points situated at lower consumption levels. The median, close to the lower quartile, and several high-range outliers highlight the uneven distribution across the dataset.

All relevant data can be seen here:

	bmi	age	fastfood	logbmi
count	847.000000	847.000000	847.000000	847.000000
mean	25.573024	44.622196	19.044628	3.228495
std	4.217671	14.532799	32.651239	0.160372
50%	24.930748	44.000000	6.000000	3.216102
25%	22.637770	32.000000	6.000000	3.119620
75%	28.039151	57.000000	24.000000	3.333602

Figure 6: Summary statistics for the dataset.

The table in Figuredisplays the summary statistics of our main variables. These statistics provide essential insights into the distribution characteristics of each variable, including their central tendencies and variabilities. The median values, along with the first and third quartiles, are particularly helpful in understanding the asymmetry and potential outliers in the data. This detailed statistical overview aids in ensuring the robustness and reliability of subsequent analyses.

b) Density histogram of the weekly returns from the ETF AGG

c) The weekly return over time for each of the four ETFs

d) Box plot of the weekly returns by ETF

e) Summary sizes for the four ETFs

2 Statistical analysis

f) Statistical models describing the weekly return for each of the four ETFs

g) Confidence intervals

h) Hypothesis test

i) Hypothesis Test: Comparison of Weekly Returns from VAW and AGG

j) The importance of statistical test

k) Correlation between ETFs