

COLLEGE OF COMPUTING TECHNOLOGY - DUBLIN
BACHELOR OF SCIENCE IN INFORMATION TECHNOLOGY

BIG DATA INTEGRATION

Assessment 3 Exploration of the Darts dataset using statistics

Adelo Vieira

Student Number: 2017279

Lecturer: Dr. Muhammad Iqbal

April 1, 2019

The aim of this assignment is to analyse a dataset provided by the *Wall Street Journal* (the *Darts* dataset) that contain the results of 4 experts investing in the stock market and the results of 4 random investors (Darts). In this sense, we want to use a series of statistical tools to verify if stock market professionals can **really** do better than simply throwing darts at pages of stock market listings.

Contents

1	The Darts dataset	1
1.1	Importing the Darts.xls dataset into R	1
2	Exploratory data analysis	2
2.1	Shape of distribution: Computing Skewness, kurtosis and visualizing the result using Histograms	2
2.2	Simple Linear Correlation	5
3	Descriptive Data Analysis	6
3.1	Central tendency: Mean, Media and Mode	6
3.2	Measures of Variation	8
3.2.1	Computing the Min, Max and Range	8
3.2.2	Computing the Quantiles and visualizing the result using Box Plots	9
3.2.3	Variance	9
3.2.4	Standard Deviation	11
3.2.5	Z-score	11
4	A little deeper analysis of the Dart Dataset	11
	Declaration	15
	Bibliography	16

List of Figures

1.1	The Darts Dataset	1
2.1	Our RStudio project	3
2.2	Quantiles	4
2.3	Histogram obtained for DJIA	5
2.4	Simple Linear Correlation	6
3.1	<i>Mean, Media and Mode</i> for Expert#1	7
3.2	<i>Mean, Media and Mode</i> for Expert#4	7
3.3	Range of values for Expert#1	8
3.4	Range of values for Dart#1	9
3.5	Quantiles	10
3.6	Variance for Expert#1, Dart#1 and DJIA	10
3.7	Standard Deviation for Expert#1, Dart#1 and DJIA	11
3.8	Z-score for Dart#1	11

4.1	Mean of the profit made by the 4 Experts and the one made by the 4 Darts for the whole dataset (from 1990 to 2002).	13
4.2	Mean by year	14
4.3	Mean by month in 1991	14
4.4	Mean by month in 2002	15

List of Snippets

1	Computing the Mean by YEAR of the 4 Experts, the Mean of the 4 Darts and Mean of the DJIA .	12
2	Making a BarChart of the total Mean by Year	13

1 The Darts dataset

« In 1988, the Wall Street Journal began a contest that was inspired by Burton Malkiel’s book *A Random Walk Down Wall Street*. In the book, the Princeton Professor theorized that “a blindfolded monkey throwing darts at a newspaper’s financial pages could select a portfolio that would do just as well as one carefully selected by experts.”
» [Investorhome.com]

The *Darts* dataset (Figure 1.1) is the results of this contest. The dataset contains the results (in terms of the percentage of profit or gain) of 4 stock market *experts*, the results of 4 *darts* and the *DJIA*¹.

	Month	Year	Expert #1	Expert #2	Expert #3	Expert #4	Dart #1	Dart #2	Dart #3	Dart #4	DJIA
1	Jul	1990	51.6	8.0	7.7	-16.7	53.8	-10.2	-8.6	-35.0	2.5
2	Aug	1990	56.7	37.8	27.8	-16.7	36.7	-3.7	-3.9	-22.0	11.5
3	Sep	1990	29.8	4.6	-9.4	-14.9	6.8	-9.8	-11.3	-42.9	-2.3
4	Oct	1990	-13.7	-18.2	-19.4	-28.6	44.4	-9.0	-20.3	-44.0	-9.2
5	Nov	1990	25.8	-39.8	-40.4	-96.9	12.9	-9.8	-31.4	-37.1	-8.5
6	Dec	1990	2.8	-17.3	-48.7	-69.9	-2.9	-24.1	-29.4	-53.4	-12.8
7	Jan	1991	8.9	0.6	-20.8	-29.3	-2.8	-23.0	-31.5	-32.7	-9.3
8	Feb	1991	5.0	-7.0	-14.1	-65.2	-2.7	-11.3	-40.0	-95.2	-0.8
9	Mar	1991	117.2	33.0	3.3	2.1	53.6	2.6	-1.6	-64.7	11.0
10	Apr	1991	74.2	47.7	4.9	-45.9	32.6	31.3	7.2	-26.3	15.8
11	May	1991	53.4	53.4	51.2	44.3	158.0	60.3	45.7	27.6	16.2
12	Jun	1991	86.9	83.1	58.5	39.3	28.3	27.3	14.8	14.0	17.3

Figure 1.1: The Darts Dataset.

1.1 Importing the Darts.xls dataset into R

We imported the dataset into *RStudio* using `read_excel()` from `library(readxl)` instead of using the GUI functionalities of *RStudio*:

```
1 darts = read_excel("/home/adelo/1-system/desktop/it_cct/3-Big_Data_Integration/6-5-6-CA3/Darts.xls")
```

¹The Jones Industrial Average

2 Exploratory data analysis

In this section we explore our dataset in order to summarize its main characteristics. We are going to perform some measures of the *Shape of distribution* and visualize the distribution of the data using Histograms.

2.1 Shape of distribution: Computing Skewness, kurtosis and visualizing the result using Histograms

In this part we first compute the *Skewness* and *Kurtosis*:

- *Skewness* is a method for quantifying the lack of symmetry in the distribution of a variable. [Iqbal (2019)]
 - *Skewness* value of zero indicates that the variable is distributed symmetrically. Positive number indicate asymmetry to the left, negative number indicates asymmetry to the right. [Iqbal (2019)]
- *Kurtosis* is a measure that gives indication in terms of the peak of the distribution.
 - Variables with a pronounced peak toward the mean have a high *Kurtosis* score and variables with a flat peak have a low *Kurtosis* score.

In Figure 2.1 we show the results of the *Skewness* and the *Kurtosis* for `darts$Expert#1`. We got a value of 1.558487 and 6.009875 for the *Skewness* and the *Kurtosis*, respectively.

From the *Skewness* we got we can conclude that the data for the `Expert #1` is not symmetry. The positive value of the *Skewness* indicates that this data is asymmetry to the left. This is confirmed by the corresponding histogram shown in Figure 2.2.

In Figures 2.2 and 2.3 we show the histograms obtained for `Expert #4`, `Dart #1`, `Dart #4` and `DJIA`, respectively. The captions of this Figures show the *Skewness* and *Kurtosis* for the corresponding data.

It is very important to notice that this kind of analysis are crucial to determine the appropriate approach in following stages. For example, we have to know the distribution of the dataset to determine which measures of the *central tendency* is more appropriate. When the data is skewed or not symmetry (like in our case) the median is generally considered to be the best representative of the central location of the data. [Statistics.laerd (a)]

In the following code snippet we show the code used to compute the *skewness*, *kurtosis* and to display a histogram for `darts$Expert#1`:

```
1 skewness(darts$`Expert #1`)
```

```

2 kurtosis(darts$`Expert #1`)
3 hist(darts$`Expert #1`,
4       border="grey",
5       col="blue"
6     )
7 abline(h=0, col="red")
8 box();

```

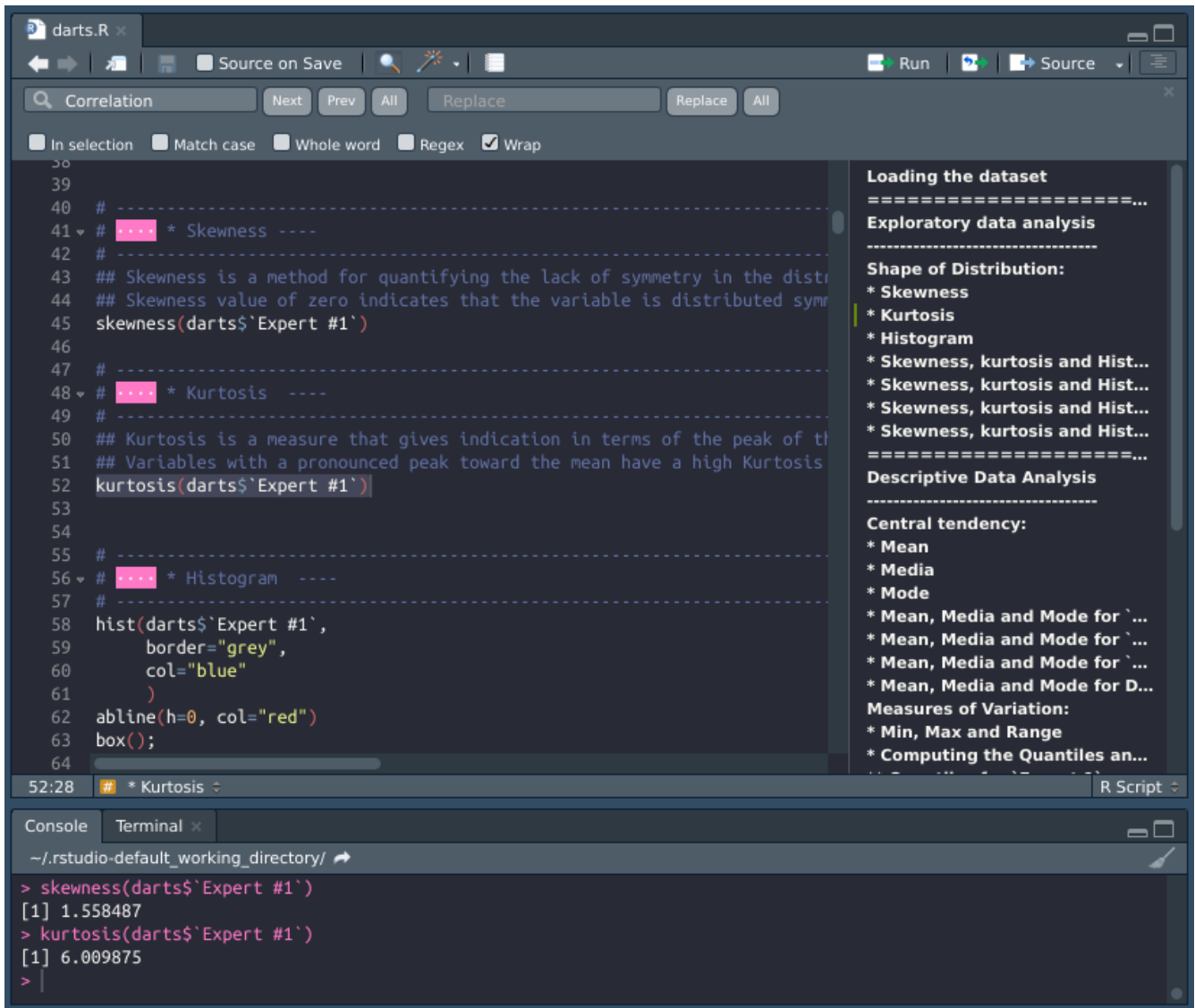


Figure 2.1: Our RStudio project.

Figure 2.2: Quantiles

Figure 2.3: Histogram obtained for DJIA.
 $Skewness = -0.06647257$, $kurtosis = 2.537426$

2.2 Simple Linear Correlation

We performed and plot the simple linear correlation of:

- Mean of *Experts* against Mean of *Darts*.
- Mean of *Experts* against *DJIA*.
- Mean of *Darts* against *DJIA*. (Figure 2.4)

The results shown a dispersed data for all the plots, so we can not say that there is relationship between *Experts* and *Darts* or *DJIA*.

We must also notice that all the *regression Lines* shown a positive slope. We could therefore say that as one variable increases the other should generally decreases. However, given that the data is skewed the previous assumption is questionable.

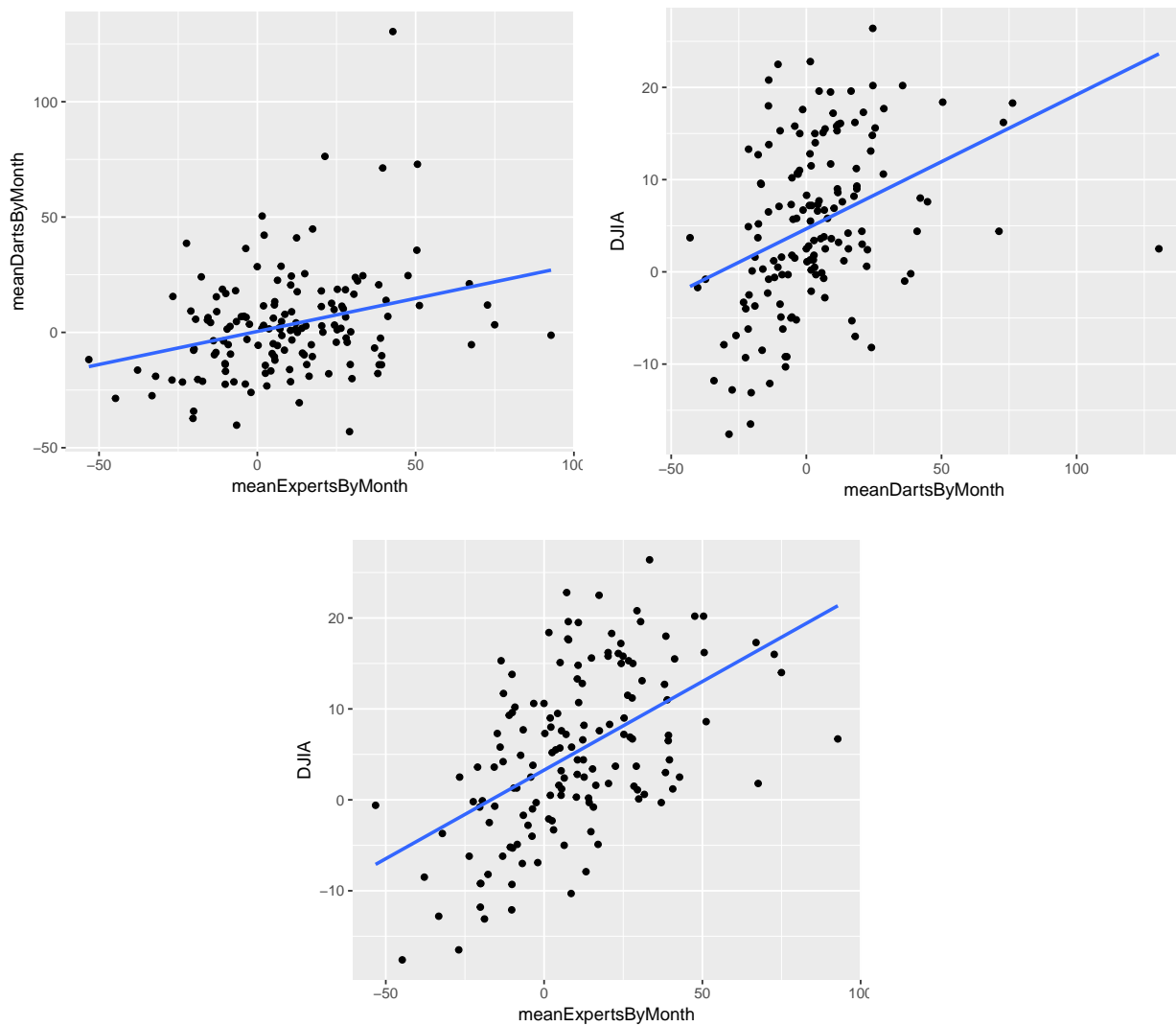


Figure 2.4: Simple Linear Correlation

3 Descriptive Data Analysis

In this section we are going to compute a series of measure in order to quantitatively describe some features of our dataset.

3.1 Central tendency: Mean, Media and Mode

We start computing the *Mean*, *Media* and *Mode* for some of the columns of our dataset. In Figures 3.1 and 3.4, we show the values obtained for `Expert#1` and `Expert#4`, respectively.

Notice that for `Expert#1` we obtained values quite different for the *Mean*, *Media* and *Mode*. This was expected

because of the asymmetry distribution of the **Expert#1** data represented for its histogram in Figure 2.2². In the other hand, notice that the data corresponding to **Expert#4** show very similar values for the *Mean*, *Media* and *Mode*, which is consistent with its corresponding symmetrical histogram in Figure 2.2.

```

153 # -----
154 # *** * Mean, Media and Mode for `Expert #1` ----
155 # -----
156 mean(darts$`Expert #1`)
157 median(darts$`Expert #1`)
158 mean(mfv(darts$`Expert #1`, method='mfv'))
159
160
159:1 # * Mean, Media and Mode for `Expert #1`
R Script

```

```

> mean(darts$`Expert #1`)
[1] 48.92993
> median(darts$`Expert #1`)
[1] 38.7
> mean(mfv(darts$`Expert #1`, method='mfv'))
[1] 36.48889
>

```

Figure 3.1: *Mean, Media and Mode* for Expert#1

```

153 # -----
154 # *** * Mean, Media and Mode for `Expert #4` ----
155 # -----
156 mean(darts$`Expert #4`)
157 median(darts$`Expert #4`)
158 mean(mfv(darts$`Expert #4`, method='mfv'))
159
160
158:43 # * Mean, Media and Mode for `Expert #4`
R Script

```

```

> mean(darts$`Expert #4`)
[1] -24.58367
> median(darts$`Expert #4`)
[1] -21.4
> mean(mfv(darts$`Expert #4`, method='mfv'))
[1] -21.4
>

```

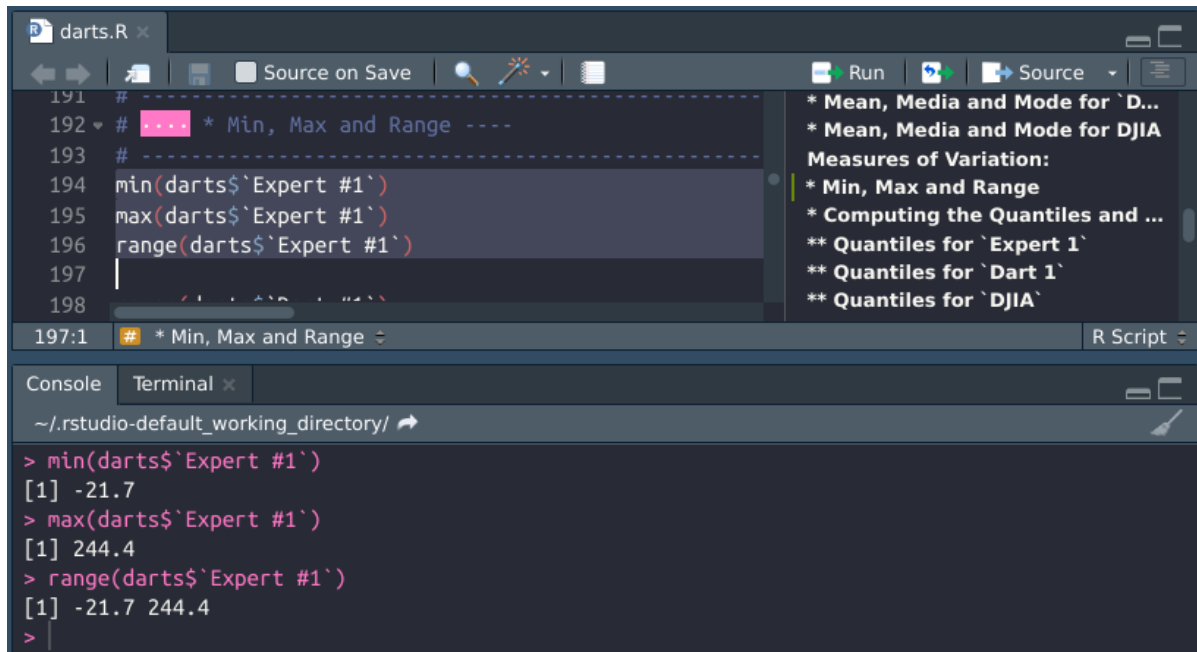
Figure 3.2: *Mean, Media and Mode* for Expert#4

²The more skewed the distribution, the greater the difference between the median and mean [Statistics.laerd (a)]

3.2 Measures of Variation

3.2.1 Computing the Min, Max and Range

We also calculated the ranges of our dataset to have an idea of the variation of the data. In Figure 3.5e we show the result for Expert#1 and Dart#1. Notice the big range of values with have for both columns but specially for Dart#1.



```
191 # -----
192 # *** * Min, Max and Range ----
193 # -----
194 min(darts$`Expert #1`)
195 max(darts$`Expert #1`)
196 range(darts$`Expert #1`)
197 |
198 |
197:1 # * Min, Max and Range
```

```
> min(darts$`Expert #1`)
[1] -21.7
> max(darts$`Expert #1`)
[1] 244.4
> range(darts$`Expert #1`)
[1] -21.7 244.4
>
```

Figure 3.3: Range of values for Expert#1

```
191 # -----
192 # *** * Min, Max and Range ----
193 # -----
194 min(darts$`Dart #1`)
195 max(darts$`Dart #1`)
196 range(darts$`Dart #1`)
197 |
198

197:1 # * Min, Max and Range

* Mean, Media and Mode for `D...
* Mean, Media and Mode for DJIA
Measures of Variation:
* Min, Max and Range
* Computing the Quantiles and ...
** Quantiles for `Expert 1`
** Quantiles for `Dart 1`
** Quantiles for `DJIA`

~/rstudio-default_working_directory/
> min(darts$`Dart #1`)
[1] -7
> max(darts$`Dart #1`)
[1] 591.4
> range(darts$`Dart #1`)
[1] -7.0 591.4
> |
```

Figure 3.4: Range of values for Dart#1

3.2.2 Computing the Quantiles and visualizing the result using Box Plots

Quartiles are another useful tool to measure of spread of the data.

The quartiles tell us about the spread of a data set by breaking the data set into quarters, just like the median breaks it in half. [Statistics.laerd (b)]

In Figure 3.5 we show can visualizing the results of the quartiles for some of the columns of the *Darts* dataset.

3.2.3 Variance

We also analysis the spread of the data computing the *Variance*. The results obtained for Expert#1, Dart#1 and DJIA are shown in Figures 3.6.

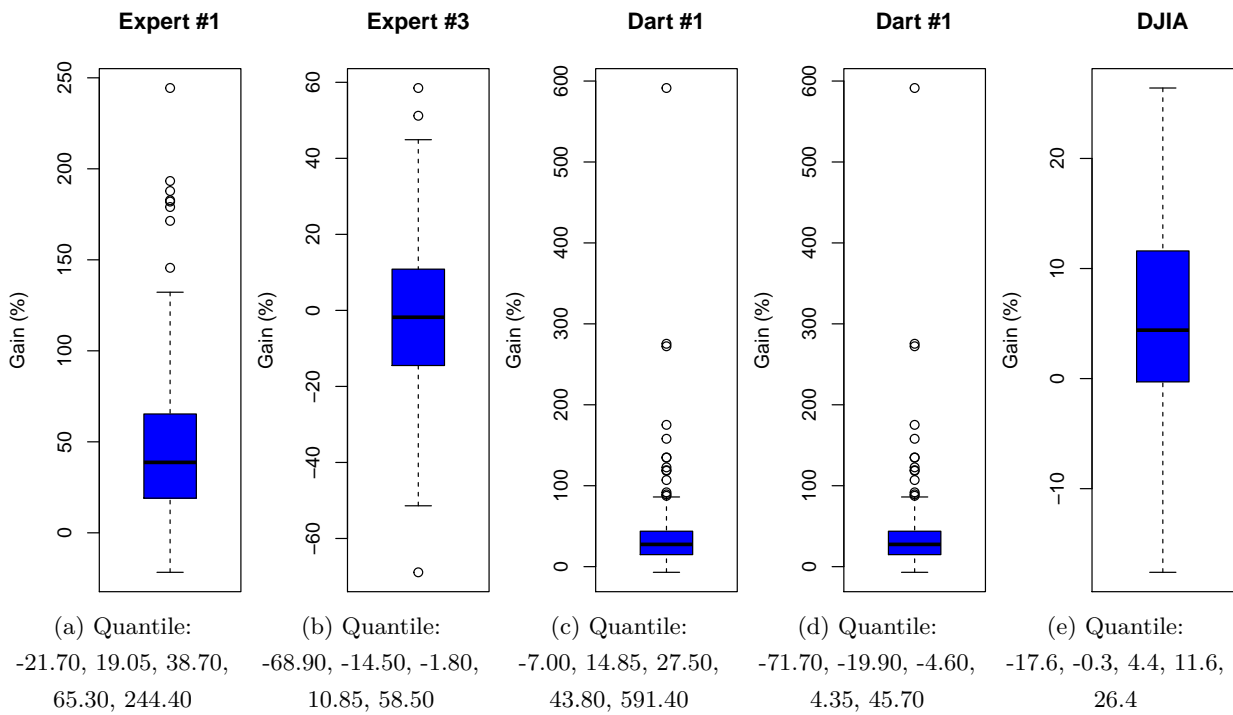


Figure 3.5: Quantiles

```

263 # -----
264 # * Variance ----
265 # -----
266 var(darts$`Expert #1`)
267 var(darts$`Dart #1`)
268 var(darts$DJIA)
269
270
271
264:3 # * Variance

```

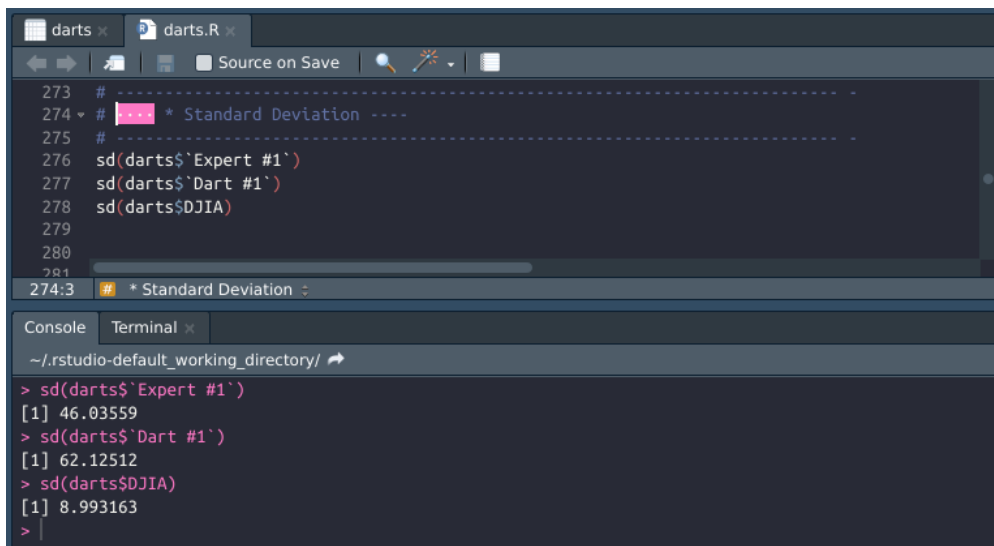
```

~/rstudio-default_working_directory/
> var(darts$`Expert #1`)
[1] 2119.276
> var(darts$`Dart #1`)
[1] 3859.531
> var(darts$DJIA)
[1] 80.87699
>

```

Figure 3.6: Variance for Dart#1, Dart#1 and DJIA

3.2.4 Standard Deviation

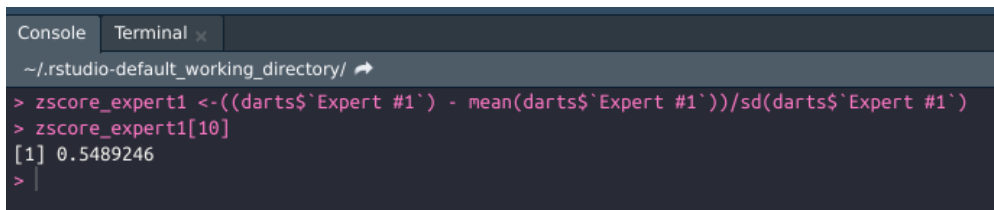


```
273 # -----
274 # * Standard Deviation ----
275 # -----
276 sd(darts$`Expert #1`)
277 sd(darts$`Dart #1`)
278 sd(darts$DJIA)
279
280
281
274:3 # * Standard Deviation :
```

```
~/rstudio-default_working_directory/
> sd(darts$`Expert #1`)
[1] 46.03559
> sd(darts$`Dart #1`)
[1] 62.12512
> sd(darts$DJIA)
[1] 8.993163
>
```

Figure 3.7: Standard Deviation for Dart#1, Dart#1 and DJIA

3.2.5 Z-score



```
~/rstudio-default_working_directory/
> zscore_expert1 <- ((darts$`Expert #1`) - mean(darts$`Expert #1`))/sd(darts$`Expert #1`)
> zscore_expert1[10]
[1] 0.5489246
>
```

Figure 3.8: Z-score for Dart#1

4 A little deeper analysis of the Dart Dataset

Here is where the most important part of our analysis begins. As explained at the beginning of this report, the *Dart* dataset was built in order to verify if stock market professionals can **really** do better than simply throwing darts at pages of stock market listings. Therefore, the most pertinent analysis we can do is to compare the profit made by the Experts and Darts to verify our hypothesis.

So, in order to know who got a better profit (Expert or Darts), we are going to start by computing the Mean of the profit made by the 4 Experts and the one made by the 4 Darts for the whole dataset. This is shown in Figure 4.1. From this figure we can see that Experts have clearly a better profit than Darts and the DJIA. However, we must notice that the Darts also made a positive profit and it is not so low compare to the one made for Experts but specially compare to the DJIA.

As was expected, *Experts* beat *Darts*, but we have to admit that *Darts* didn't show a bad result considering its random nature.

Now, let's go a little deeper into our analysis and study what happened year by year. To do so, we first calculated the Mean for the 4 *Experts* and the Mean for the 4 *Darts* year by year. We Snippet 1 to compute the Mean by year. Then we use the Snippet 2 to plot the Bar Chart presented in Figure 4.2.

From Figure 4.2 we can see:

- *Experts* have beaten *Darts* and *DJIA* almost every year.
- *Darts* have beaten *Experts* only in 1996 and 2001.
- Also notice that *Darts* have beaten the *DJIA* 5 times (1991, 1994, 1996, 2000, and 2001). So *Darts* have a better result than *DJIA* in 5 out of 13 years.
- The other important feature that we can see from this Chart is that the trend is consistent. In general we can say that if one year the *Experts* had an increase/decrease of the profit, *Darts* and *DJIA* also show an increase/decrease.

We can go even deeper into our analysis and Student in detail the results month by month in every year. In Figure 4.3 and 4.4 we show the results of every month in 1991 and 2002, respectively. In general the detailed results and trend for this years (month by month) just confirm the features we have mentioned in the previous analysis by year shown by the Chart in Figure 4.2.

Snippet 1: Computing the Mean by YEAR of the 4 Experts, the Mean of the 4 Darts and Mean of the DJIA

```
1 # -----
2 # Computing the Mean "by YEAR" of the 4 Experts, the Mean of the 4 Darts and Mean of the DJIA
3 # -----
4 meanExpertsByYear = c(1990:2002)
5 meanDartsByYear = c(1990:2002)
6 meanDJIAByYear = c(1990:2002)
7 Year = c(1990:2002)
8
9 j = 1
10 for (i in c(1990:2002)) {
11   meanExpert_1_ByYear = mean(darts$'Expert #1'[which(darts$Year==i)])
12   meanExpert_2_ByYear = mean(darts$'Expert #2'[which(darts$Year==i)])
13   meanExpert_3_ByYear = mean(darts$'Expert #3'[which(darts$Year==i)])
14   meanExpert_4_ByYear = mean(darts$'Expert #4'[which(darts$Year==i)])
15
16   meanDart_1_ByYear = mean(darts$'Dart #1'[which(darts$Year==i)])
17   meanDart_2_ByYear = mean(darts$'Dart #2'[which(darts$Year==i)])
18   meanDart_3_ByYear = mean(darts$'Dart #3'[which(darts$Year==i)])
19   meanDart_4_ByYear = mean(darts$'Dart #4'[which(darts$Year==i)])
20
21   meanExpertsByYear[j] = mean(c(meanExpert_1_ByYear,
22                               meanExpert_2_ByYear,
```

```

23         meanExpert_3_ByYear,
24         meanExpert_4_ByYear))
25
26     meanDartsByYear[j] = mean(c(meanDart_1_ByYear,
27                               meanDart_2_ByYear,
28                               meanDart_3_ByYear,
29                               meanDart_4_ByYear))
30
31     meanDJIAByYear[j] = mean(darts$DJIA[which(darts$Year==i)])
32     j = j+1
33 }
34 meanByYear <- data.frame(Year, meanExpertsByYear, meanDartsByYear, meanDJIAByYear)

```

Snippet 2: Making a BarChart of the total Mean by Year

```

1  # -----
2  # Making a BarChart of the total Mean by Year -----
3  # -----
4  cols <- c('red', 'blue', 'grey');
5  ylim <- c(min(meanByYear[c('meanExpertsByYear', 'meanDartsByYear', 'meanDJIAByYear')]*1.2, max(meanByYear[c('meanExpertsByYear', 'meanDartsByYear', 'meanDJIAByYear')]*1.2));
6  par(lwd=0);
7  barplot(
8     t(meanByYear[c('meanExpertsByYear', 'meanDartsByYear', 'meanDJIAByYear')]),
9     beside=T,
10    ylim=ylim,
11    border=cols,
12    col=cols,
13    names.arg=meanByYear$Year,
14    ylab='Gain (%)'
15 );
16 abline(h=0, col="yellow")
17 box();

```

Figure 4.1: Mean of the profit made by the 4 Experts and the one made by the 4 Darts for the whole dataset (from 1990 to 2002).

Figure 4.2: Mean “by YEAR” of the 4 Experts combined (Red), the the 4 Darts combined (Blue) and the DJIA (Grey).

Figure 4.3: Mean by month in 1991. Experts (Red), Darts (Blue) and DJIA (Grey)

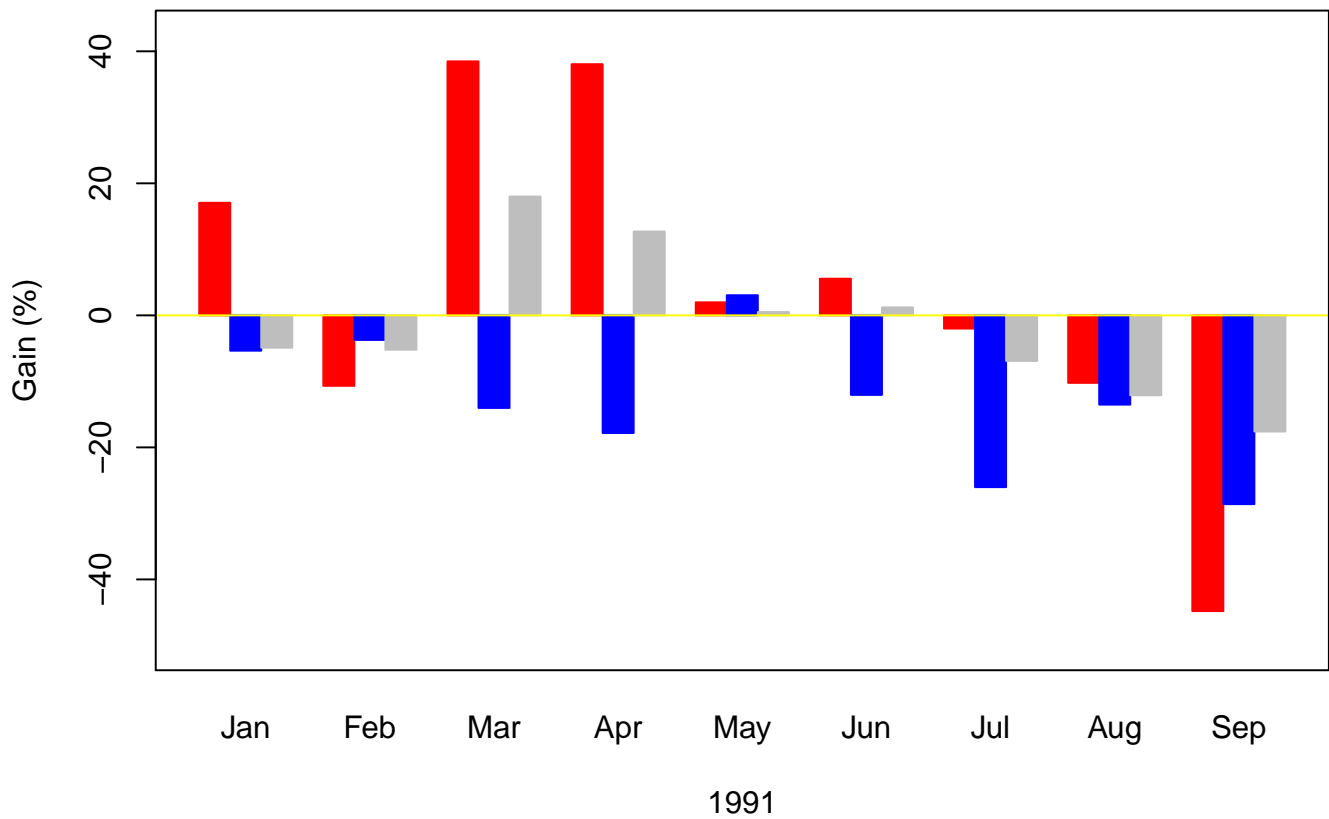


Figure 4.4: Mean by month in 2002. Experts (Red), Darts (Blue) and DJIA (Grey)

Declaration

I hereby declare that all of the work shown here is my own work.

Student's Name: Adelo Vieira

Student Number: 2017279

Date: April 1, 2019

Bibliography

Investorhome.com. The wall street journal dartboard contest. URL <http://www.investorhome.com/darts.htm>.

1

Dr. Muhammad Iqbal. *Lecture of the Big Data Integration course at CCT: Lecture 6 - Statistics*. 2019. 2

Statistics.laerd. Measures of central tendency, a. URL <https://statistics.laerd.com/statistical-guides/measures-central-tendency-mean-mode-median.php>. 2, 7

Statistics.laerd. Measures of spread, b. URL <https://statistics.laerd.com/statistical-guides/measures-of-spread-range-quartiles.php>. 9

Wikipedia.org. Descriptive statistics. URL https://en.wikipedia.org/wiki/Descriptive_statistics.

Wikipedia.org. *Exploratory data analysis*. wikipedia.org, 2004. URL https://en.wikipedia.org/wiki/Exploratory_data_analysis.