**Assignment 1:**

**Provide all the code and results from R and report the answers in a document along with your interpretation.**

1. **Data Camp: [40 marks]**

Complete the first three chapters to the Introduction to R in Data Camp:

1. **Vectors in R: [25 marks]**

**Regression The following are a sample of observations on incoming solar radiation at a greenhouse.**

**12.3 11.4 7.1 9.2 10.3 10.8 9.5 11.8**

1. **Assign the object to a vector called *solar.radiation***

*> #vector creation*

*> solar.radiation=c(12.3, 11.4, 7.1, 9.2, 10.3, 10.8, 9.5, 11.8)*

1. **Find the mean, median and variance.**

*> #Mean of solar.radiation*

*> mean\_solar.radiation = mean (solar.radiation)*

*> mean\_solar.radiation*

*[1] 10.3*

*>*

*> #Median of solar.radiation*

*> median\_solar.radiation = median (solar.radiation)*

*> median\_solar.radiation*

*[1] 10.55*

*>*

*> #variance of solar.radiation*

*> var\_solar.radiation = var(solar.radiation)*

*> var\_solar.radiation*

*[1] 2.828571*

1. **+ Add 10 to each observation and assign the resulting vector to *sr10*. Find the mean, median and variance of *sr10*. Which statistics change and by how much?**

> # +10 = sr10

> sr10= solar.radiation + 10

> sr10

[1] 22.3 21.4 17.1 19.2 20.3 20.8 19.5 21.8

>

> #Mean of sr10

> mean\_sr10= mean (sr10)

> mean\_sr10

[1] 20.3

>

> #Median of sr10

> median\_sr10 = median (sr10)

> median\_sr10

[1] 20.55

>

> #variance of sr10

> var\_sr10=var(sr10)

> var\_sr10

[1] 2.828571

> #different of Solar.radiation to sr10

> mean\_sr10 - mean\_solar.radiation

[1] 10

> median\_sr10 - median\_solar.radiation

[1] 10

(> var\_sr10 - var\_solar.radiation

[1] -2.220446e-15 ????)

> 2.828571 - 2.828571

[1] 0

1. **Multiple each observation by -2 and assign the resulting vector to *sr2*. Find the mean, median and variance of *sr2*. Which statistics change and by how much?**

> #solar.radiation x -2 (sr2)

> sr2= solar.radiation \* -2

> sr2

[1] -24.6 -22.8 -14.2 -18.4 -20.6 -21.6 -19.0 -23.6

>

> #Mean of sr2

> mean\_sr2 = mean(sr2)

> mean\_sr2

[1] -20.6

>

> #Median of sr2

> median\_sr2 = median(sr2)

> median\_sr2

[1] -21.1

>

> #Variance of sr2

> var\_sr2= var(sr2)

> var\_sr2

[1] 11.31429

>

> #different of Solar.radiation to sr2

> mean\_solar.radiation - mean\_sr2

[1] 30.9

> median\_solar.radiation - median\_sr2

[1] 31.65

> var\_solar.radiation - var\_sr2

[1] -8.485714

1. **Plot a histogram of each of *solar.radiation*, *sr10* and *sr2***.

*hist\_sola = hist(solar.radiation, main="Solar Radiation", col= c("blue"))*

Chart, histogram

Description automatically generated

*hist\_sr10 = hist(sr10, main="Solar Radiation + 10", col= c("grey"))*

Chart, histogram

Description automatically generated

*hist\_sr2 = hist(sr2, main="Solar Radiation x -2", col= c("yellow"))*

Chart, histogram

Description automatically generated

1. **Descriptive Statistics - Old faithful data: [35 marks]**

The Melanoma data frame has data on 205 patients in Denmark with malignant melanoma.The data is stored in R in the package MASS, in a dataset called ‘Melanoma’.

1. Load the package "MASS ". (hint: use library() commands to install package and then load it. Use install.packages() if you have never installed the package). Then load in the data "Melanoma" in R (hint: use the data() command and help() to understand the data).
2. Describe the data, referring to the number of variables and sample size?

The Melanoma data are data collected from 205 melanoma patients which includes:

* Age in year
* Sex: male or female
* Time: how long they have survived for in days
* Status: If they are dead or alive and if dead did, they did from the melanoma
* Year: when the operation was done
* Thickness: the thickness of the tumour in mm
* Ulcer: where or not if they had ulcer.

1. What type of variables are present in the data? Refer to each variable.

* time: int
* status: Categorical (ordinal)
* sex: Categorical
* age: Numerical (continuous)
* year: Numerical (continuous)
* thickness: Numerical (continuous)
* ulcer: Categorical

1. Make suitable univariate plots for each of the variables in the data set and interpret the results.
2. For all numerical variables, report appropriate descriptive statistics for each of to describe the centrality and spread of the data. Explain the choice of statistic used.