# R command to read the csv file ( data file)

> Data = read.csv("C:/Courses/BigData/hw1\_data.csv")

##what are the column names of the dataset?

> names(Data)

[1] "Ozone" "Solar.R" "Wind" "Temp" "Month" "Day"

#Extract the first 2 rows of the data frame and print them to the console. What does the output look like?

> Data[c(1,2),]

Ozone Solar.R Wind Temp Month Day

1 41 190 7.4 67 5 1

2 36 118 8.0 72 5 2

#How many observations (i.e. rows) are in this data frame?

> nrow(Data)

[1] 153

#Extract the last 2 rows of the data frame and print them to the console. What does the output look like?

> Data[c(152,153),]

Ozone Solar.R Wind Temp Month Day

152 18 131 8.0 76 9 29

153 20 223 11.5 68 9 30

#What is the value of Ozone in the 47th row?

> Data[47,]

Ozone Solar.R Wind Temp Month Day

47 21 191 14.9 77 6 16

Ans: 21

#How many missing values are in the Ozone column of this data frame?

> sum(is.na(Data$Ozone))

[1] 37

#What is the mean of the Ozone column in this dataset? Exclude missing values (coded as NA) from this calculation.

> mean(Data$Ozone,na.rm=TRUE)

[1] 42.12931

#Extract the subset of rows of the data frame where Ozone values are above 31 and Temp values are above 90. What is the mean of Solar.R in this subset?

> mean(Data$Solar.R[Data$Ozone>31 & Data$Temp>90],na.rm = TRUE)

[1] 212.8

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Ozone** | **Solar.R** | **Wind** | **Temp** | **Month** | **Day** |
| 97 | 267 | 6.3 | 92 | 7 | 8 |
| 97 | 272 | 5.7 | 92 | 7 | 9 |
| 76 | 203 | 9.7 | 97 | 8 | 28 |
| 118 | 225 | 2.3 | 94 | 8 | 29 |
| 84 | 237 | 6.3 | 96 | 8 | 30 |
| 85 | 188 | 6.3 | 94 | 8 | 31 |
| 96 | 167 | 6.9 | 91 | 9 | 1 |
| 78 | 197 | 5.1 | 92 | 9 | 2 |
| 73 | 183 | 2.8 | 93 | 9 | 3 |
| 91 | 189 | 4.6 | 93 | 9 | 4 |

#What is the mean of "Temp" when "Month" is equal to 6?

> mean(Data$Temp[Data$Month == 6],na.rm = TRUE)

[1] 79.1

#What was the maximum ozone value in the month of May (i.e. Month = 5)?

> max(Data$Ozone[Data$Month == 5],na.rm = TRUE)

[1] 115