

# Data Analysis in RStudio

## Setup

1. Install R. Digital Ocean (<https://www.digitalocean.com/community/tutorials/how-to-install-r-on-ubuntu-20-04>) has a nice guide on this (and many other things Linux-related, they are one of my preferred documentation sources)
2. Install RStudio. The RStudio Downloads Page (<https://www.rstudio.com/products/rstudio/download/>) has a Debian package you can download. Install with `sudo dpkg -i <pkgname>.deb`
3. Add data.table with Tools->Install Packages..., enter `data.table`
4. Install texlive full with `sudo apt install texlive-full` to enable PDF output

## Datasets

Load profiles come from <http://wzy.ece.iastate.edu/Testsystem.html> (<http://wzy.ece.iastate.edu/Testsystem.html>)

## Analysis

### Load Data

```
feeder_a <- read.csv("feeder_a_data.csv")
names(feeder_a)
```

```
## [1] "Time"      "Year"      "Month"     "Day.of.Week"
## [5] "Hour"      "Elapsed.Days" "Elapsed.Hours" "Total.Power"
## [9] "Bus.1001"  "Bus.1002"  "Bus.1003"  "Bus.1004"
## [13] "Bus.1005"  "Bus.1006"  "Bus.1007"  "Bus.1008"
## [17] "Bus.1009"  "Bus.1010"  "Bus.1011"  "Bus.1012"
## [21] "Bus.1013"  "Bus.1014"  "Bus.1015"  "Bus.1016"
## [25] "Bus.1017"
```

There's a lot of columns here from the different smart meters. Let's just take the total power

```
fa <- feeder_a[,c("Month", "Day.of.Week", "Hour", "Elapsed.Days", "Total.Power")]
```

Take a quick look at the loaded data. You can also use the "Environment" tab, but this won't show up in the notebook

```
head(fa)
```

	Month <int>	Day.of.Week <int>	Hour <int>	Elapsed.Days <dbl>	Total.Power <dbl>
1	1	1	1	0.00	88.82

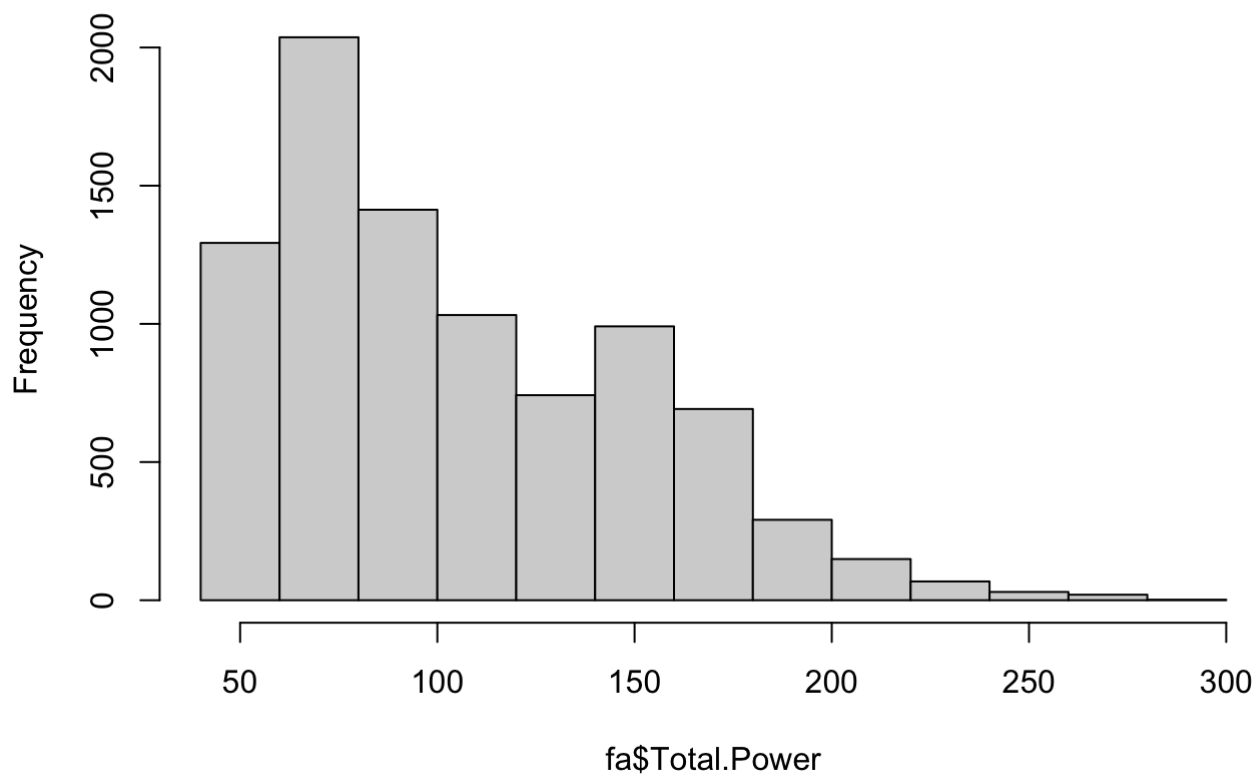
	Month <int>	Day.of.Week <int>	Hour <int>	Elapsed.Days <dbl>	Total.Power <dbl>
2	1	1	2	0.04	81.83
3	1	1	3	0.08	78.36
4	1	1	4	0.13	81.32
5	1	1	5	0.17	78.71
6	1	1	6	0.21	76.73

6 rows

## Visualize Data

```
hist(fa$Total.Power)
```

**Histogram of fa\$Total.Power**



How

to do arithmetic with data frame columns

```
fa$b12 <- feeder_a$Bus.1001 + feeder_a$Bus.1002
```

```
mean(fa$Total.Power)
```

```
## [1] 105.8696
```

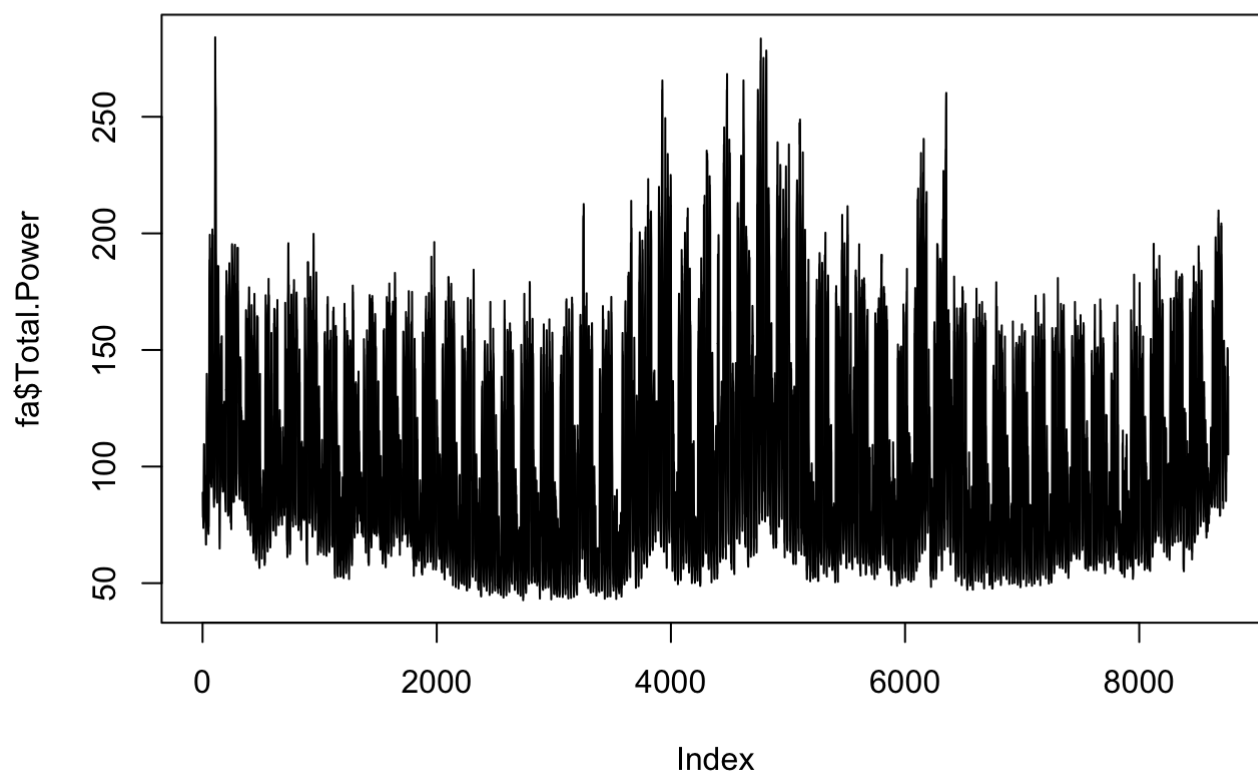
There's a convenient command to print summary statistics for each column

```
summary(fa)
```

```
##      Month      Day.of.Week      Hour      Elapsed.Days
## Min.      : 1.000    Min.      :1.000    Min.      : 0.00    Min.      :  0.00
## 1st Qu.: 4.000    1st Qu.:2.000    1st Qu.: 5.75    1st Qu.: 91.24
## Median : 7.000    Median :4.000    Median :11.50    Median :182.48
## Mean   : 6.526    Mean   :3.992    Mean   :11.50    Mean   :182.48
## 3rd Qu.:10.000    3rd Qu.:6.000    3rd Qu.:17.25    3rd Qu.:273.72
## Max.   :12.000    Max.   :7.000    Max.   :23.00    Max.   :364.96
##
##      Total.Power      b12
## Min.      : 42.68    Min.      :0
## 1st Qu.: 68.07    1st Qu.:0
## Median : 94.11    Median :0
## Mean   :105.87    Mean   :0
## 3rd Qu.:141.18    3rd Qu.:0
## Max.   :284.09    Max.   :0
##                      NA's      :1
```

Plot the power over time. Yikes! There's a lot of data here? Just how long is it?

```
plot(fa$Total.Power,type="l")
```



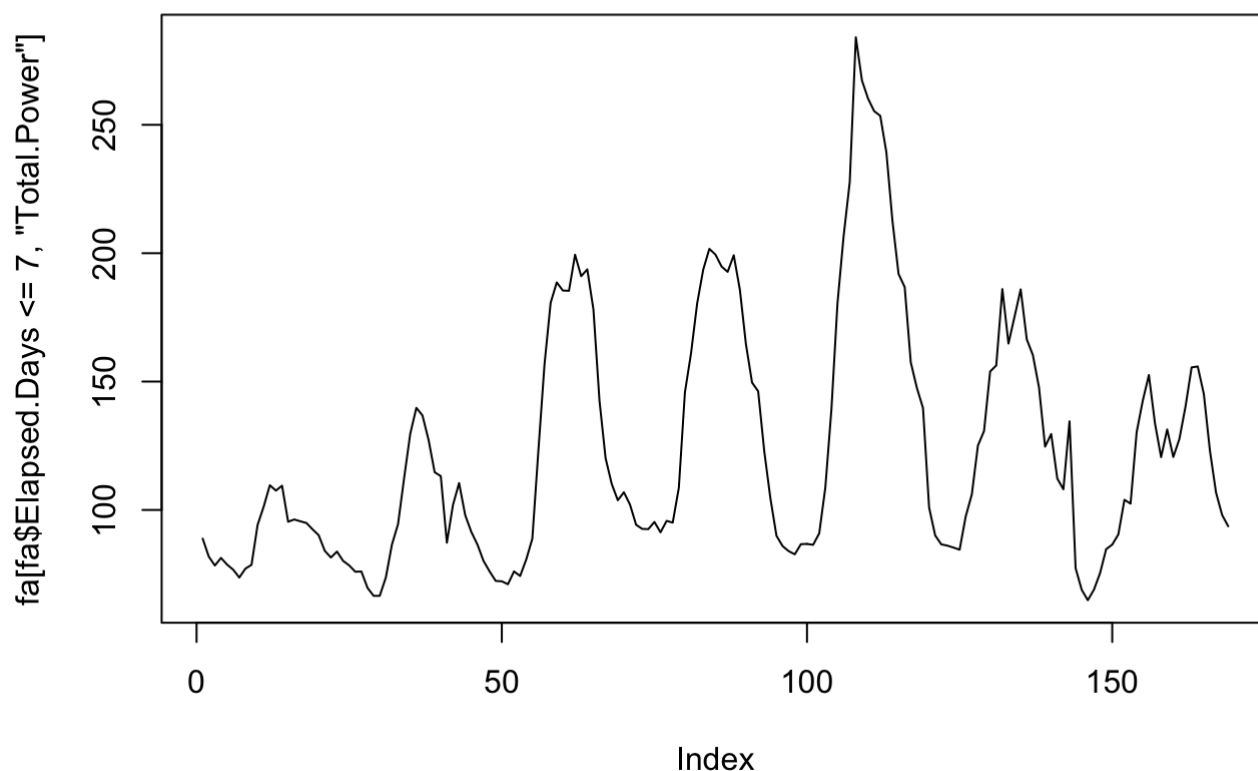
whole year! That's too much!

```
max(fa$Elapsed.Days)
```

```
## [1] 364.96
```

We can plot just the first week by taking a subset of rows

```
plot(fa[fa$Elapsed.Days<=7,"Total.Power"],type="l")
```



```
names(fa)
```

```
## [1] "Month"          "Day.of.Week"    "Hour"           "Elapsed.Days"   "Total.Power"
## [6] "b12"
```

## How to Replace Excel Pivot Tables with R Data Tables

```
require(data.table)
```

```
## Loading required package: data.table
```

```
ta <- as.data.table(fa)
```

Let's look at how much power is used per day of week (1 = Sunday). People are a little lazy on Mondays

```
ta[,mean(Total.Power),by=Day.of.Week]
```

**Day.of.Week**  
<int>

**V1**  
<dbl>

<b>Day.of.Week</b> <int>	<b>V1</b> <dbl>
1	78.96709
2	106.49592
3	114.98106
4	118.26026
5	117.86812
6	115.44796
7	89.56215

7 rows

Now let's look at power vs month. It looks like air conditioning loads cause a peak in July

```
ta[,mean(Total.Power),by=Month]
```

<b>Month</b> <int>	<b>V1</b> <dbl>
1	115.28683
2	104.92824
3	102.92249
4	87.79007
5	90.53629
6	116.84221
7	132.60573
8	108.71017
9	107.07769
10	93.32519

1-10 of 12 rows

Previous **1** 2 Next

Finally, let's look at power vs time of day. The peak is at noon, though in some regions/seasons it is common to have two peaks in mornings/evenings

```
ta[,mean(Total.Power),by=Hour]
```

<b>Hour</b> <int>	<b>V1</b> <dbl>
1	63.41608

Hour <int>	V1 <dbl>
2	61.39458
3	60.51874
4	60.75890
5	63.33058
6	70.42581
7	93.31123
8	116.95967
9	138.12444
10	149.16318
1-10 of 24 rows	
Previous 1 2 3 Next	