# Exploratory Data Analysis
# On ChickWeight Dataset

1/30/2020

## Data

Lets look at the `ChickWeight` Dataset.

```
str(ChickWeight)
```

```
## Classes 'nfnGroupedData', 'nfGroupedData', 'groupedData' and 'data.frame':   578 obs. of 4 variables:
## $ weight: num 42 51 59 64 76 93 106 125 149 171 ...
## $ Time : num 0 2 4 6 8 10 12 14 16 18 ...
## $ Chick : Ord.factor w/ 50 levels "18"<"16"<"15"<..: 15 15 15 15 15 15 15 15 15 15 ...
## $ Diet : Factor w/ 4 levels "1","2","3","4": 1 1 1 1 1 1 1 1 1 1 ...
## - attr(*, "formula")=Class 'formula'  language weight ~ Time | Chick
##   .. ..- attr(*, ".Environment")=<environment: R_EmptyEnv>
## - attr(*, "outer")=Class 'formula' language ~Diet
##   .. ..- attr(*, ".Environment")=<environment: R_EmptyEnv>
## - attr(*, "labels")=List of 2
##   ..$ x: chr "Time"
##   ..$ y: chr "Body weight"
## - attr(*, "units")=List of 2
##   ..$ x: chr "(days)"
##   ..$ y: chr "(gm)"
```

The `ChickWeight` data frame has 578 observation and 4 variables from an experiment on the effect of diet on early growth of chicks.The body weights of the chicks were measured at birth and every second day thereafter until day 20. They were also measured on day 21. There were four groups on chicks on different protein diets. We want to know whether these diets have changed the chicken's weights after 20 days. Lets find out this by doing exploratory data analysis on our dataset.

## Summary

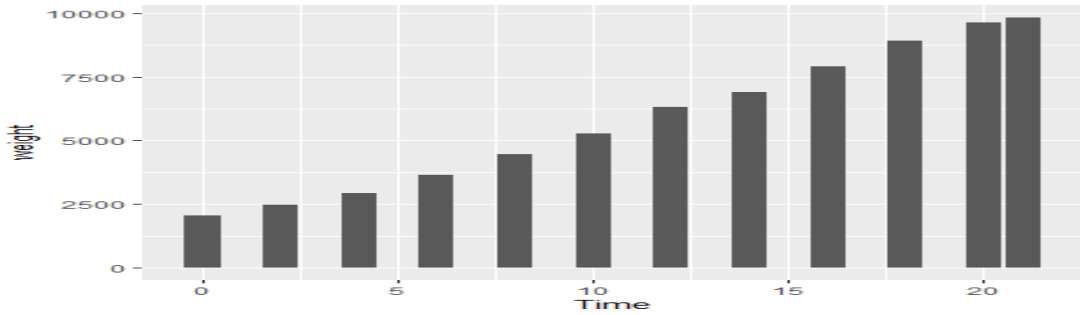
Let's look at the summary of the data

```
summary(ChickWeight)
```

```
##      weight           Time           Chick     Diet
##  Min.   : 35.0   Min.   : 0.00   13     : 12   1:220
##  1st Qu.: 63.0   1st Qu.: 4.00   9      : 12   2:120
##  Median :103.0   Median :10.00   20     : 12   3:120
##  Mean   :121.8   Mean   :10.72   10     : 12   4:118
##  3rd Qu.:163.8   3rd Qu.:16.00   17     : 12
##  Max.   :373.0   Max.   :21.00   19     : 12
##                                  (Other):506
```

From the summary of the data, we can see that there are no missing values in the data.It shows that Chick and Diet are categorical variables, hence displaying the number of counts per category. It shows that weight (in gms) and Time are continuous variables, so gives sensible summary statistics.

```
library(ggplot2)
ggplot(ChickWeight, aes(Time,weight)) + geom_col()
```




This plot shows the weight of the chickens are increasing from day 0 day 21. But we want to know the effect of all the different diets on the growth of the chickens.So, lets start our analysis.

## Analysis on Weight of the chickens

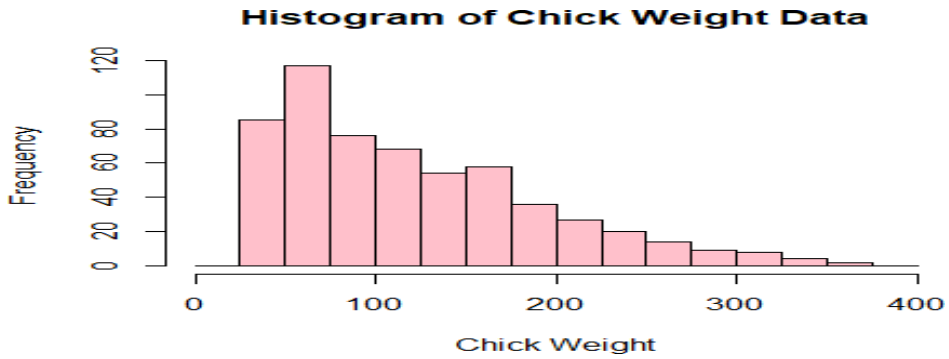
```
library(pastecs)
stat.desc(ChickWeight$weight, desc = F)

## nbr.val nbr.null    nbr.na      min      max    range      sum
##      578        0         0       35      373      338    70411
```

As we want to find the effect of our diet on the chickens, we are doing our analysis on the weight of the chickens as it is a continous variable. So,if we observe the basic statisctics of the weight of the chickens. The range of the weight of the chickens are in between 35 gm and 373 gm.

```
library(pastecs)
stat.desc(ChickWeight$weight, basic=F)

##       median         mean      SE.mean CI.mean.0.95          var      std.dev
##  103.0000000  121.8183391    2.9562038    5.8062322 5051.2234413   71.0719596
##     coef.var
##    0.5834258
```

From the above we found that standard deviation is 71 gm approximately. So, mostly the weight of the chickens fall between (121 - 71) 50 gm and (121 + 71) 192 gm. And as the standard deviation is less than the mean value, the more data is closer to the mean value i.e.more number of chicken weights are near to 121 gm. Lets plot the density of the distribution of our data for better understanding.

```
library(ggplot2)
hist(ChickWeight$weight,xlab = "Chick Weight", breaks = seq(0, 400, 25),
main = "Histogram of Chick Weight Data", col = 'pink')
```




From the above plot, we can say that the distribution is positively skewed i.e. more values clustered towards the left

tail of the distribution and this shows that mean of the weights (121 gm) is higher than median of the weights (103 gm) and it is clearly seen that more data is located near the mean value 121 gm.

#Calculating the mean weight of the chickens for all the four diets for further analysis

```
Mean_weight_at_day20 <-aggregate(formula = weight ~ Diet, FUN = mean, subset = Time %in% c(20
), data = ChickWeight)
Mean_weight_at_day20
```

```
##   Diet   weight
## 1    1 170.4118
## 2    2 205.6000
## 3    3 258.9000
## 4    4 233.8889
```

From the above, we can see that mean weight of the chickens are highest for diet-3, this tells us that the effect of diet-3 is more when compared to other diets on the chickens. And diet -1 is the very les s effective in the growth of the chickens.

```
library(Rmisc)

CI <- with(Mean_weight_at_day20, mean(Mean_weight_at_day20$weight) + c(-1, 1) * 1.96 * sd(Mean
_weight_at_day20$weight))
CI
```

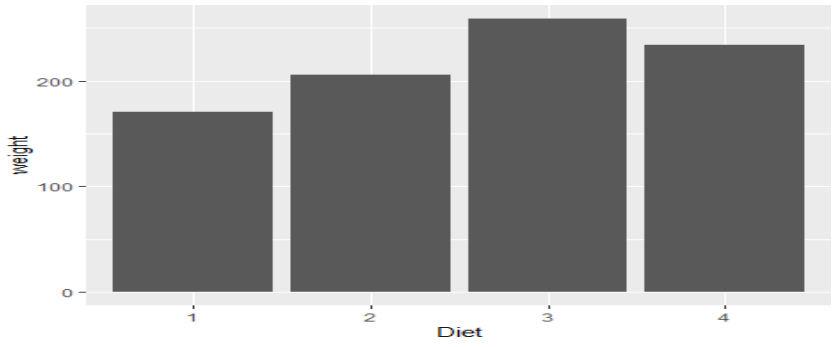
```
## [1] 142.6419 291.7584
```

```
with(Mean_weight_at_day20, sum(Mean_weight_at_day20$weight >= CI[1] & Mean_weight_at_day20$wei
ght <= CI[2]) / length(Mean_weight_at_day20$weight))
```

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

So, from the above results we can tell that the 95% of the mean of the chickens for every diet lies between the intervals 142 gm and 291 gm approximately. As our dataset is not having any outliers, when we are checking whether the data lies between the intervals of the confidence intervals it's giving that 100% of our data lies between the confidence intervals.

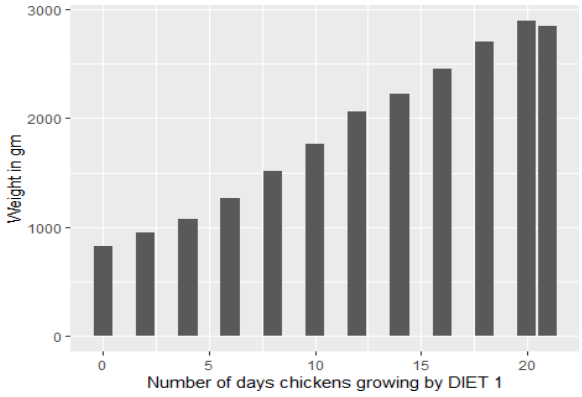
```
library(ggplot2)
ggplot(Mean_weight_at_day20, aes(Diet,weight)) + geom_col()
```

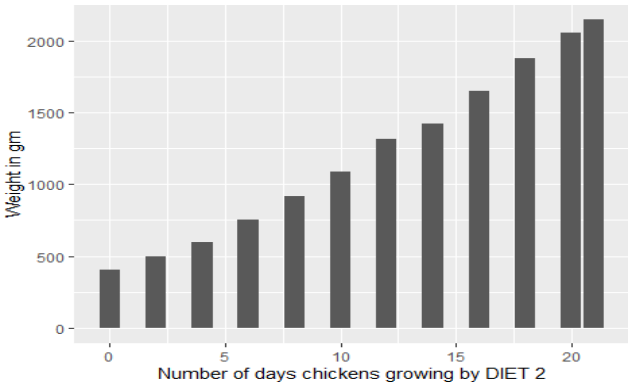



The plot clearly shows that the diet-3 is most effective in increasing the growth of chickens when compared to other diets and the least effective one is diet-1.

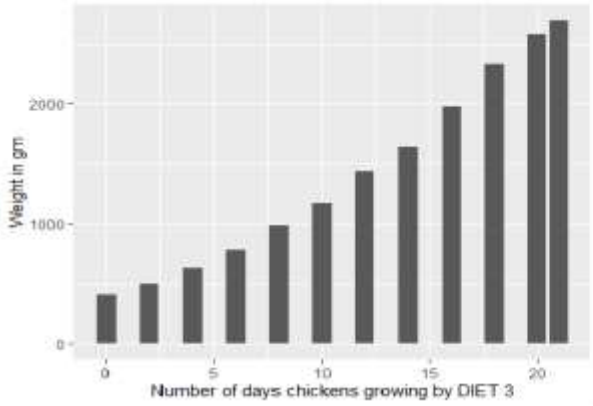
## Conclusion

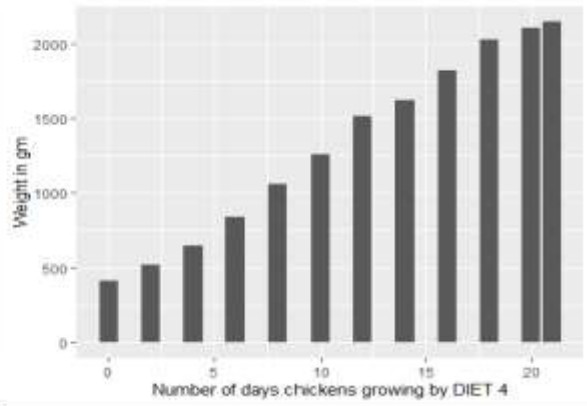
```
library(ggplot2)
library(dplyr)

chick_we_1 <- filter(select(ChickWeight,weight,Time,Diet), Diet == 1)
ggplot(chick_we_1, aes(Time,weight)) + geom_col() + labs(y="Weight in gm", x = "Number of days
chickens growing by DIET 1")

chick_we_2 <- filter(select(ChickWeight,weight,Time,Diet), Diet == 2)
ggplot(chick_we_2, aes(Time,weight)) + geom_col() + labs(y="Weight in gm", x = "Number of days
chickens growing by DIET 2")
```




```
chick_we_3 <- filter(select(ChickWeight,weight,Time,Diet), Diet == 3)
ggplot(chick_we_3, aes(Time,weight)) + geom_col() + labs(y="Weight in gm", x = "Number of days
chickens growing by DIET 3")

chick_we_4 <- filter(select(ChickWeight,weight,Time,Diet), Diet == 4)
ggplot(chick_we_4, aes(Time,weight)) + geom_col() + labs(y="Weight in gm", x = "Number of days
chickens growing by DIET 4")
```




From the above plots, we can conclude that growth of the chickens has been increased by using diet -3 and then by diet -4 , and then by diet -2 and then by diet -1. If our dataset is huge, we can obtain some linear relationship between the growth of the chickens and Number of days the chickens are growing.