UNIVERSITY OF KARACHI

Probability and Statistical Methods

BSCS-306

Name of Student: **MUHAMMAD AMAS**

Seat No: **B20102077**

Class Roll No: **36**

Section: **A**

Semester No: **2nd**

Submitted to: **Dr. Tahseen A. Jilani**

DEPARTMENT OF COMPUTER SCIENCE

UNIVERSITY OF KARACHI

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this

Input data into R

```
library(datasets)
###Following Data Set is about the weight of Chickens on different diets.

#if you wanna see data set so remove '#' from the bottom line
#data(ChickWeight)
View(ChickWeight)

#Showing upper few values of Data Set
head(ChickWeight)
```

```
##   weight Time Chick Diet
## 1     42    0     1    1
## 2     51    2     1    1
## 3     59    4     1    1
## 4     64    6     1    1
## 5     76    8     1    1
## 6     93   10     1    1
```

```
#Showing Last few values of Data Set
tail(ChickWeight)
```

```
##     weight Time Chick Diet
## 573    155   12    50    4
## 574    175   14    50    4
## 575    205   16    50    4
## 576    234   18    50    4
## 577    264   20    50    4
## 578    264   21    50    4
```

Data Summaries/ Description

```
#Will show the names of the columns
names(ChickWeight)
```

```
## [1] "weight" "Time"   "Chick"  "Diet"
```

```
# Dimension of data set.
dim(ChickWeight)
```

```
## [1] 578   4
```

```
# Giving a summarized view of data set.
summary(ChickWeight)
```

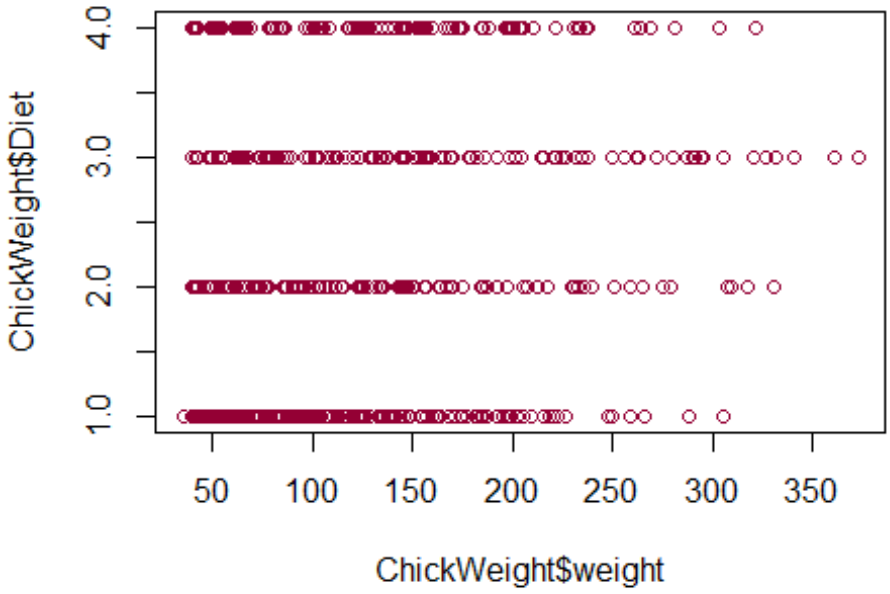
```
##      weight           Time           Chick     Diet
##  Min.   : 35.0   Min.   : 0.00   13     : 12   1:220
##  1st Qu.: 63.0   1st Qu.: 4.00   9      : 12   2:120
##  Median :103.0   Median :10.00   20     : 12   3:120
##  Mean   :121.8   Mean   :10.72   10     : 12   4:118
##  3rd Qu.:163.8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

```
# Describing one quantitative variable.
library(psych)

describe(ChickWeight)
```

```
##        vars   n   mean    sd median trimmed   mad min max range  skew kurt
osis
## weight    1 578 121.82 71.07    103  113.18 69.68  35 373   338  0.96
0.34
## Time      2 578  10.72  6.76     10   10.77  8.90   0  21    21 -0.02    -
1.26
## Chick*    3 578  26.26 14.00     26   26.27 17.79   1  50    49  0.00    -
1.19
## Diet*     4 578   2.24  1.16      2    2.17  1.48   1   4     3  0.31    -
1.39
##          se
## weight 2.96
## Time   0.28
## Chick* 0.58
## Diet*  0.05
```

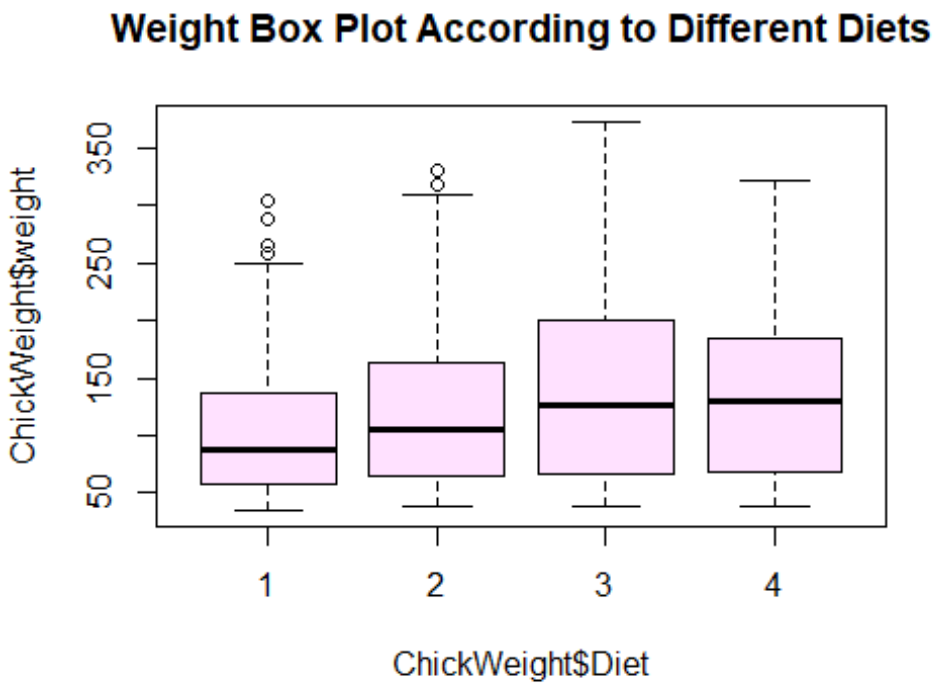
Data Plots/ Visualization

```
plot(ChickWeight$weight, ChickWeight$Diet, col="#940034")
```



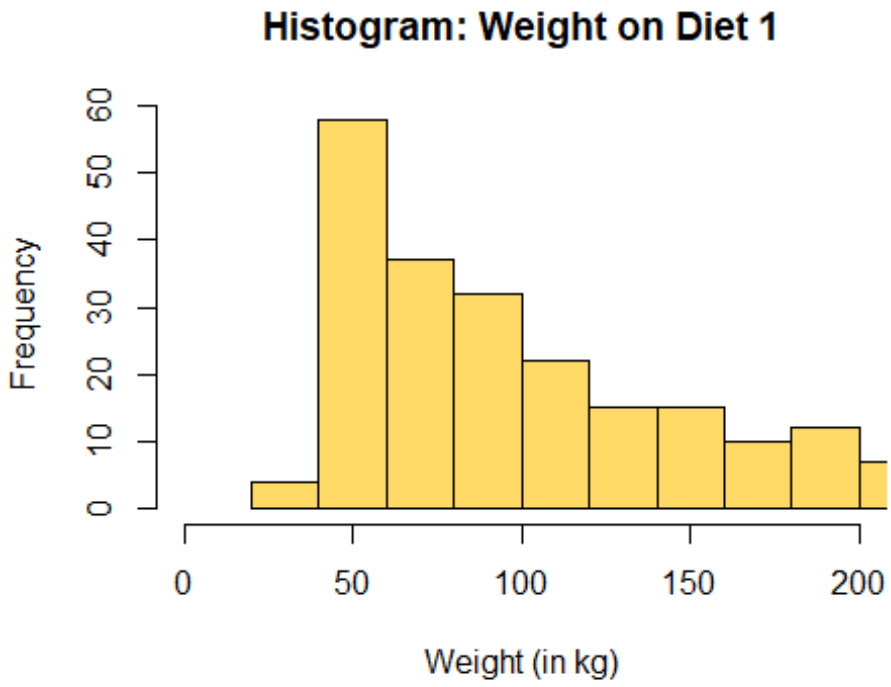

```
#it shows Scattered Plot of Chick weight with Chick diet
```

```
boxplot(ChickWeight$weight ~ ChickWeight$Diet, main="Weight Box Plot Accordin
g to Different Diets", col = "thistle1")
```

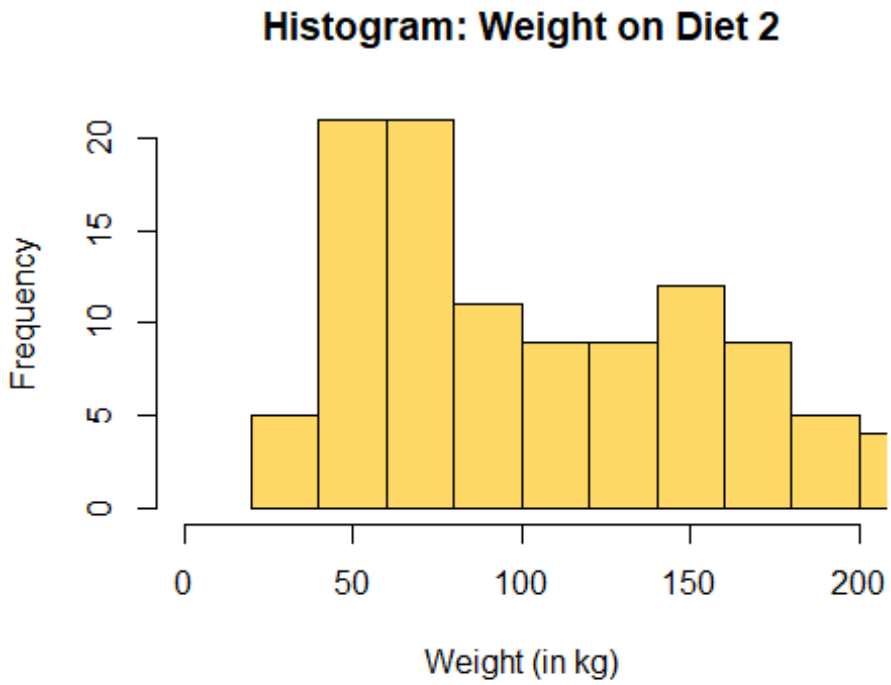



```
#Plot shows that chicks are gaining weight on Diet number 3
```

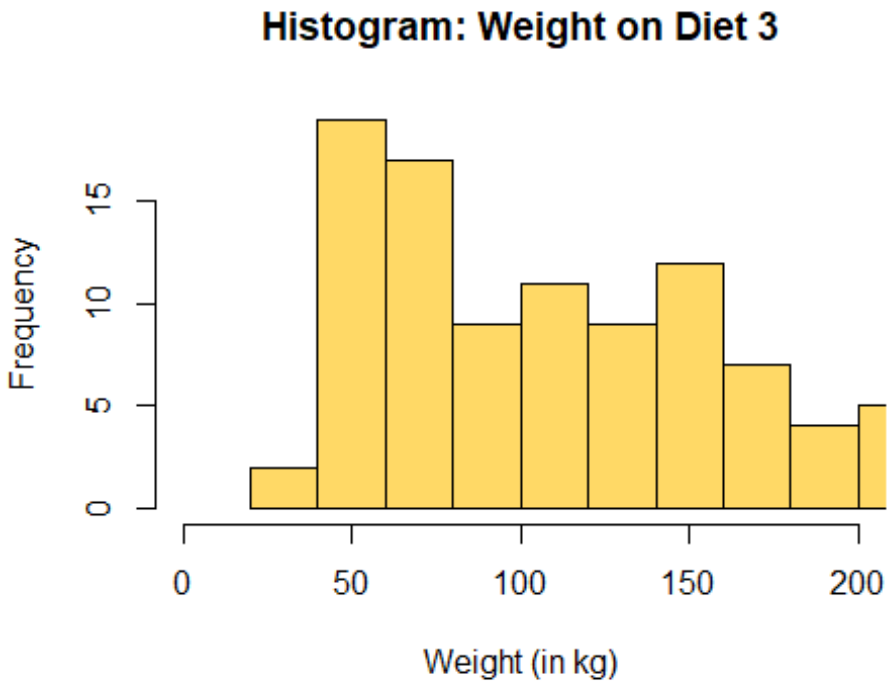
```r
##Histograms of Chick Weights according to their Diets.
hist(ChickWeight$weight [ChickWeight$Diet =="1"],
     col    = "#ffd966",
     xlim   = c(0,200),
     breaks = 15,
     main   = "Histogram: Weight on Diet 1",
     xlab   = "Weight (in kg)")
```



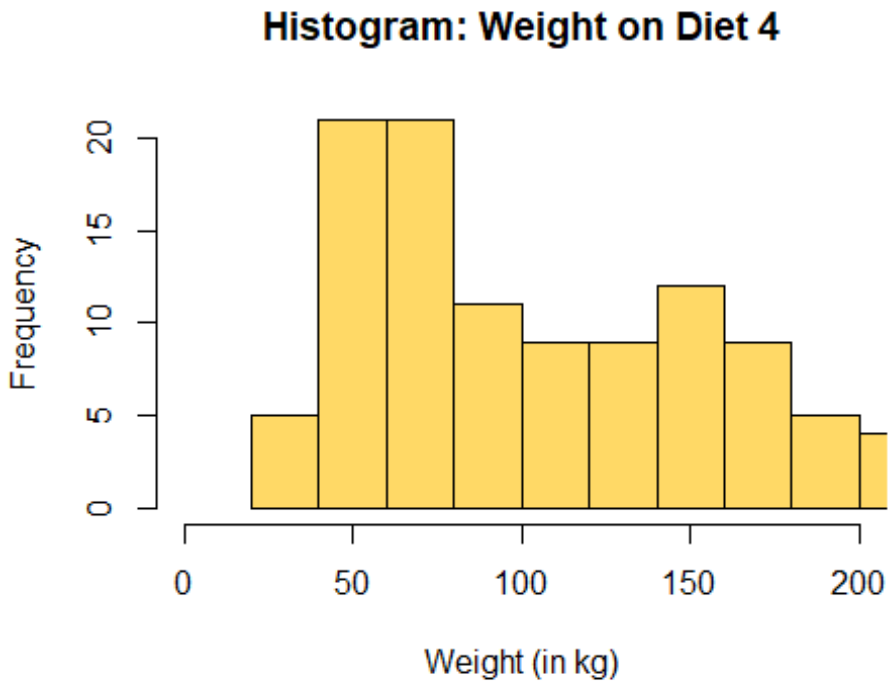
```r
hist(ChickWeight$weight [ChickWeight$Diet =="2"],
     col    = "#ffd966",
     xlim   = c(0,200),
     breaks = 15,
     main   = "Histogram: Weight on Diet 2",
     xlab   = "Weight (in kg)")
```



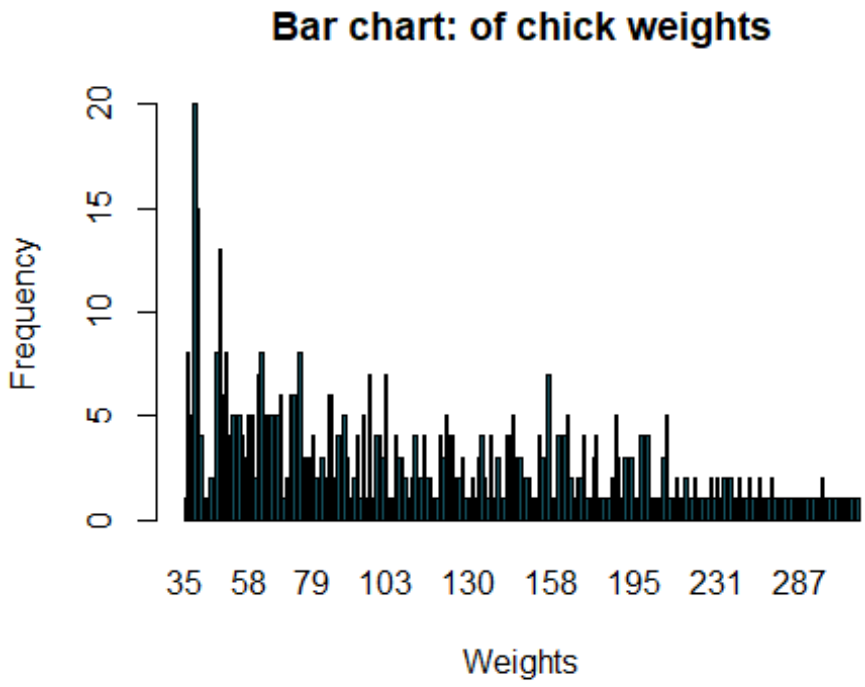
```r
hist(ChickWeight$weight [ChickWeight$Diet =="3"],
     col    = "#ffd966",
     xlim   = c(0,200),
     breaks = 15,
     main   = "Histogram: Weight on Diet 3",
     xlab   = "Weight (in kg)")
```



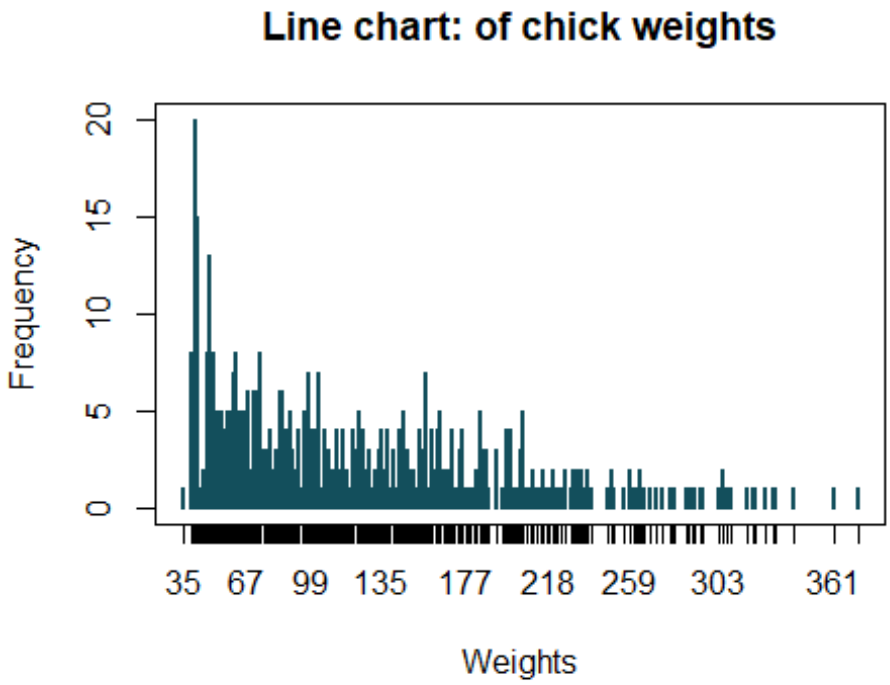
```r
hist(ChickWeight$weight [ChickWeight$Diet =="2"],
     col    = "#ffd966",
     xlim   = c(0,200),
     breaks = 15,
     main   = "Histogram: Weight on Diet 4",
     xlab   = "Weight (in kg)")
```



```r
Weight_of_chicks <- table(ChickWeight$weight)

#Bar plot of Chick Weights
barplot(Weight_of_chicks,
        main = "Bar chart: of chick weights",
        col = "#134f5c",
        xlab = "Weights",
        ylab = "Frequency")
```

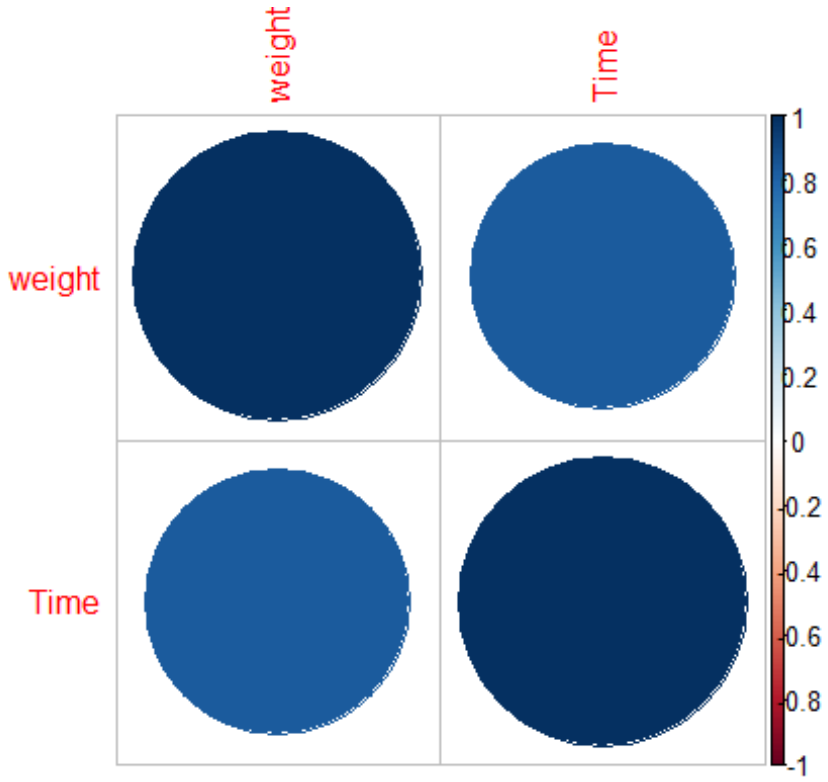


```r
#Line Charts of Chick Weights
plot(Weight_of_chicks,
     main = "Line chart: of chick weights",
     col = "#134f5c",
     xlab = "Weights",
     ylab = "Frequency")
```



Correlation

```
library(corrplot)

## corrplot 0.92 loaded

  cor(ChickWeight[, unlist(lapply(ChickWeight, is.numeric))])

##           weight      Time
## weight 1.0000000 0.8371017
## Time   0.8371017 1.0000000

  cor.mat.ChickWeight = cor(ChickWeight[, unlist(lapply(ChickWeight, is.numer
ic))])
corrplot(cor.mat.ChickWeight)
```




```
#From the plots we have concluded that the correlation is strong Positive.
```

Confidence Interval

```
library(Rmisc)

CI(ChickWeight$weight, ci = 0.95)

##    upper     mean    lower
## 127.6246 121.8183 116.0121

#Hence mean is lying in the confidence interval so we will accept the Null hy
pothesis.

CI(ChickWeight$Time, ci = 0.95)

##    upper     mean    lower
## 11.27012 10.71799 10.16586

#Hence mean is lying in the confidence interval so we will accept the Null hy
pothesis.
```

Hypothesis Testing

```
library(stats)

#one sample t test
#Question: Is the mean value of Weights from 121.8 or not?
t.test(ChickWeight$weight, mu=121.8)

##
##  One Sample t-test
##
## data:  ChickWeight$weight
## t = 0.0062036, df = 577, p-value = 0.9951
## alternative hypothesis: true mean is not equal to 121.8
## 95 percent confidence interval:
##  116.0121 127.6246
## sample estimates:
## mean of x
##  121.8183

#answer:First this is two tail test after this the mean value of weights diff
ers from 121.8,and p value is more than 0.05 so null hypothesis is accepted

#Is the mean value of Murder differ from 10.72 or not?
t.test(ChickWeight$Time, mu=10.72)

##
##  One Sample t-test
##
## data:  ChickWeight$Time
## t = -0.0071392, df = 577, p-value = 0.9943
## alternative hypothesis: true mean is not equal to 10.72
## 95 percent confidence interval:
```

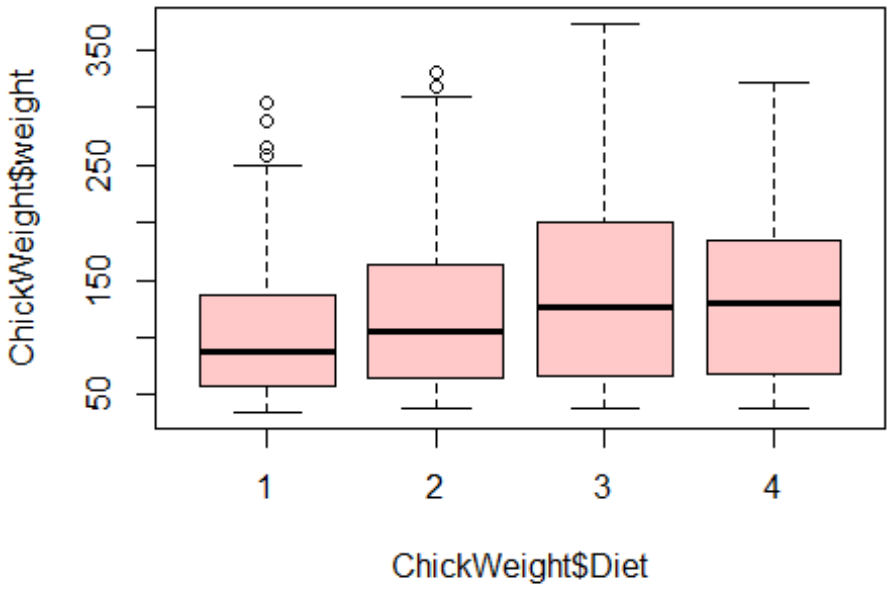
```
##  10.16586 11.27012
## sample estimates:
## mean of x
##  10.71799

#answer:Two tail test after this the mean value of time differs from 10.72,an
d p value is more than 0.05 so null hypothesis is accepted

#Two Sample Test
x <- rnorm(ChickWeight$weight)
y <- rnorm(ChickWeight$Diet)
t.test(x,y, var.equal = TRUE)

##
##  Two Sample t-test
##
## data:  x and y
## t = -0.41859, df = 1154, p-value = 0.6756
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.13766757  0.08925494
## sample estimates:
##   mean of x   mean of y
## -0.04587380 -0.02166748

#True difference in means is not equal to 0
```

Chi Square Test

```
#apply the Chi-Square test to see test of association/ independence.
chisq.test(ChickWeight$weight, ChickWeight$Diet)

## Warning in chisq.test(ChickWeight$weight, ChickWeight$Diet): Chi-squared
## approximation may be incorrect

##
##  Pearson's Chi-squared test
##
## data:  ChickWeight$weight and ChickWeight$Diet
## X-squared = 631.55, df = 633, p-value = 0.5088

#we have x-squared = 631.55,Since we get a p-Value greater than the significa
nce level of 0.05, we accept the null hypothesis and conclude that the two va
riables are in fact independent.
```

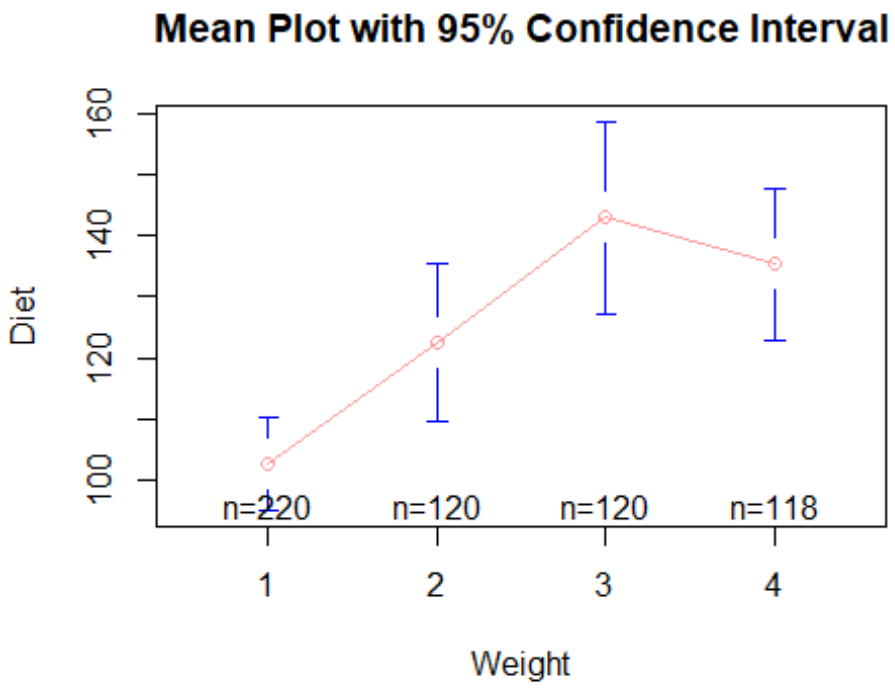
Analysis of Variance

```
#Question: Does the diet effect on weights?

boxplot(ChickWeight$weight ~ ChickWeight$Diet, col= "#ffc9c9")
```




```
model1 <- aov(ChickWeight$weight ~ ChickWeight$Diet)
summary(model1)

##                   Df  Sum Sq Mean Sq F value   Pr(>F)
## ChickWeight$Diet   3  155863   51954   10.81 6.43e-07 ***
## Residuals        574 2758693    4806
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

#it is observed that the F-statistic value is 10.81 and it is  significant as
the corresponding p-value is  smaller. Thus, it is wise to reject the null hy
pothesis of diets. In other words, the weights in diets does affect.

library(gplots)


## Attaching package: 'gplots'

## The following object is masked from 'package:stats':
##
##     lowess
```

```
plotmeans(ChickWeight$weight ~ ChickWeight$Diet, main="Mean Plot with 95% Con
fidence Interval", ylab = "Diet", xlab = "Weight", col = "#ffa09f")
```



Linear and Multiple Regression Models

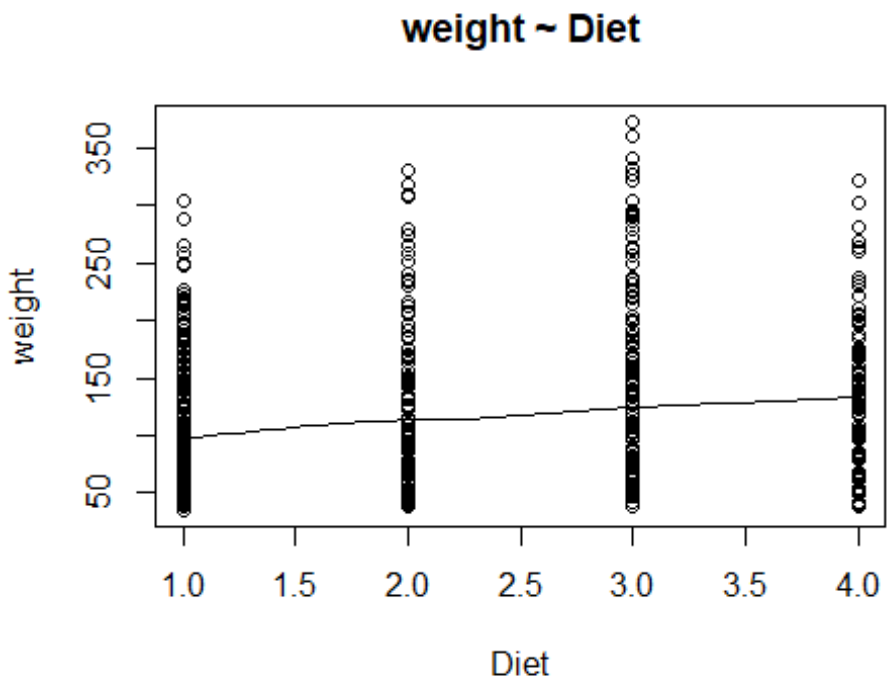
```
###Linear Regression line formula:

# This will give details of the model including the
# correlation, parameters (intercept and slope) along with P-value and Mean s
um of squares.

attach(ChickWeight)

fit.LR <- lm(weight~ Diet, data = ChickWeight)
summary(fit.LR)

##
## Call:
## lm(formula = weight ~ Diet, data = ChickWeight)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -103.95  -53.65  -13.64   40.38  230.05
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  102.645      4.674  21.961  < 2e-16 ***
```

```
## Diet2         19.971      7.867   2.538   0.0114 *
## Diet3         40.305      7.867   5.123 4.11e-07 ***
## Diet4         32.617      7.910   4.123 4.29e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 69.33 on 574 degrees of freedom
## Multiple R-squared:  0.05348,    Adjusted R-squared:  0.04853
## F-statistic: 10.81 on 3 and 574 DF,  p-value: 6.433e-07
```

#ANSWER:firstly,P value is less than 0.05 which mean intercept is significant while in Diets we see that p value is less than 0.05 so it will be accepted so in other word we can say that Diet has much significance impact on the Weight. Secondly, signs are positive which shows if one increases other greatly increases as well ,In Multiple R-squared is 0.04853 so for correlation we do square root of it so answer is 0.2202 which is positive so there is some corelation with Weight and Diet.

###PLOTS:

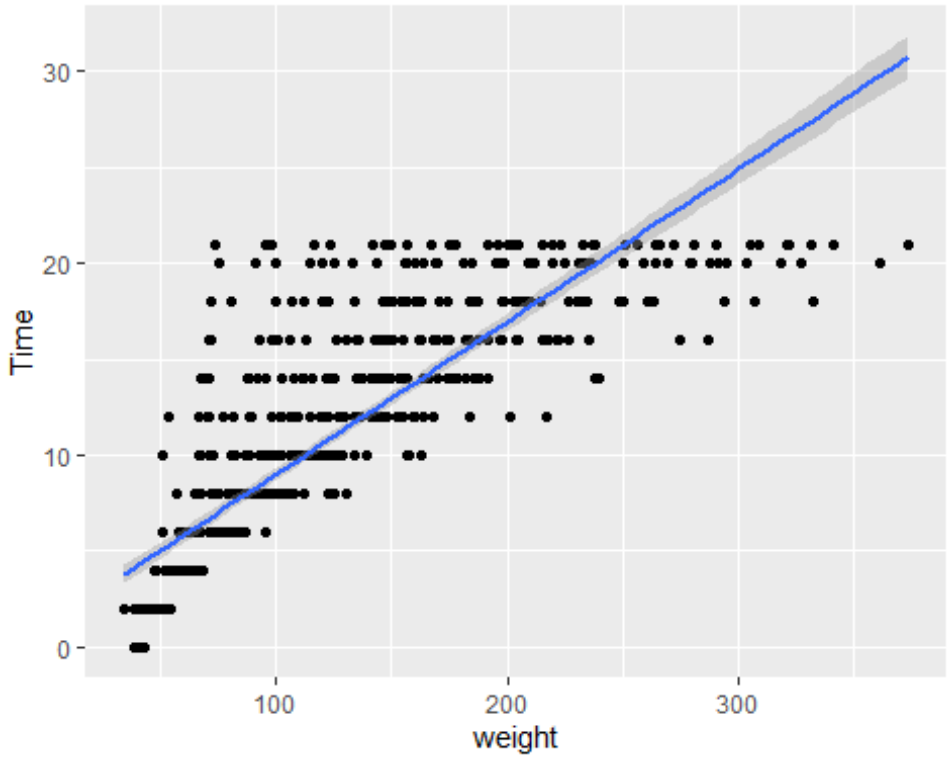
```
scatter.smooth(x=Diet, y = weight, main="weight ~ Diet")
```




```
library(ggplot2)
```

```
##
## Attaching package: 'ggplot2'
```

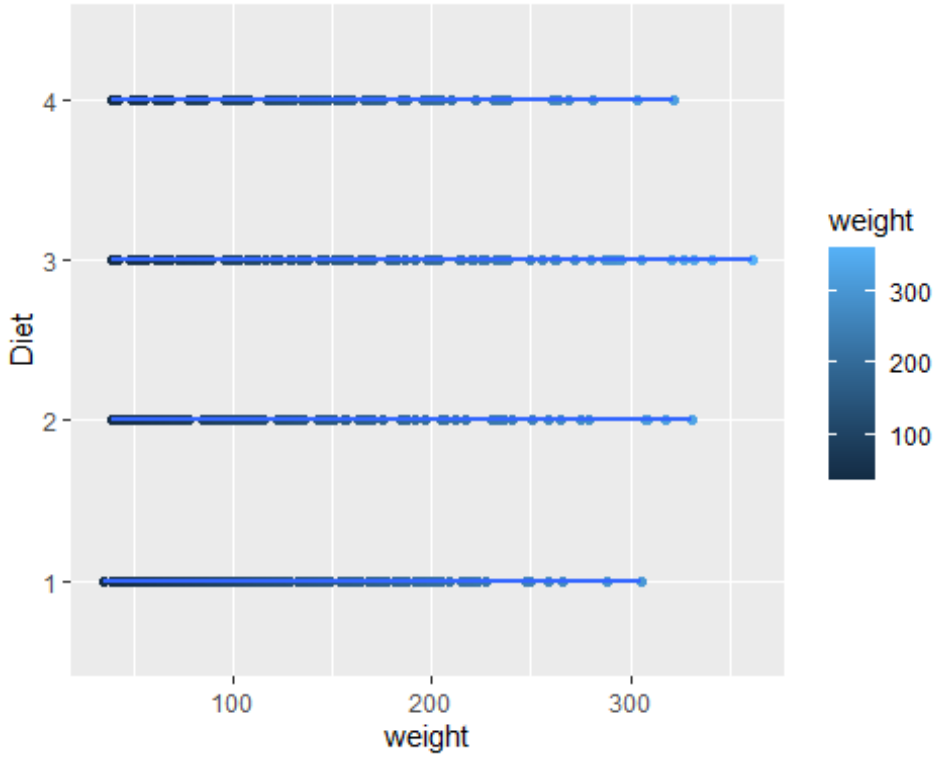
```
## The following objects are masked from 'package:psych':
##
##     %+%, alpha

ggplot(ChickWeight,aes(y=Time,x=weight))+geom_point()+geom_smooth(method="lm"
)

## `geom_smooth()` using formula 'y ~ x'
```




```
#plots show us:
#1)Linearity

#The relationship between the independent and dependent variable must be line
ar. We can test this visually with a scatter plot to see if the distribution
of data points could be described with a straight line.

#2)Independence of observations

#Because we only have one independent variable and one dependent variable fro
m the given data result so, we don't need to test for any hidden relationship
s among variables.

#3)Normality

#using the hist function we find from the above data the whether dependent va
riable follows normal distribution
```

```
###Multiple Regression line formula:
#--------------------------------

fit.MR <- lm(weight~ + Time + Diet  , data = ChickWeight)
summary(fit.MR)

##
## Call:
## lm(formula = weight ~ +Time + Diet, data = ChickWeight)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -136.851  -17.151   -2.595   15.033  141.816
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.9244     3.3607   3.251  0.00122 **
## Time          8.7505     0.2218  39.451  < 2e-16 ***
## Diet2        16.1661     4.0858   3.957 8.56e-05 ***
## Diet3        36.4994     4.0858   8.933  < 2e-16 ***
## Diet4        30.2335     4.1075   7.361 6.39e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 35.99 on 573 degrees of freedom
## Multiple R-squared:  0.7453, Adjusted R-squared:  0.7435
## F-statistic: 419.2 on 4 and 573 DF,  p-value: < 2.2e-16

#answer:P value is very low in intercept and in murder it has +ve sign so pos
itive correlation exists in it and taking sq.rt of Multiple R-squared: 0.7453
we get 0.694.

# beta0 = intercept of the regression line. which is 27.8983
# beta1 = slope of the time is 8.7152
# beta1 = slope of the Diet 2 is 16.1661
# beta1 = slope of the Diet 3 is 36.4994
# beta2 = slope of the Diet 4 is 30.2335

library(dplyr)

## Warning: package 'dplyr' was built under R version 4.1.2

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:plyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

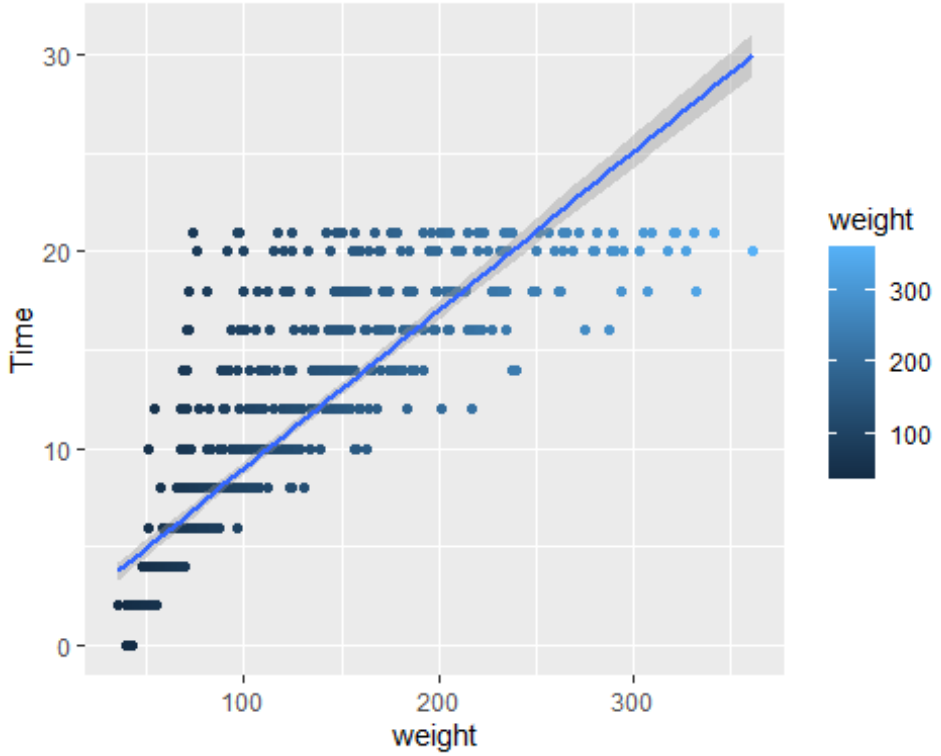
```
## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library(ggplot2)

ChickWeight %>%
  filter(weight < 373)%>%
  ggplot(aes(x=weight, y=Diet , col = weight))+ geom_point(alpha = 1)+
  geom_smooth(method = lm)

## `geom_smooth()` using formula 'y ~ x'
```




```
ChickWeight %>%
  filter(weight < 373)%>%
```

```
  ggplot(aes(x=weight, y=Time , col = weight))+ geom_point(alpha = 1)+
  geom_smooth(method = lm)
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
# Q:Is the overall regression model suitable?
# ----------------------------------------
anova(fit.LR)  # Test difference in slopes (joint F-test)
```

```
## Analysis of Variance Table
##
## Response: weight
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## Diet        3  155863   51954   10.81 6.433e-07 ***
## Residuals 574 2758693    4806
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# The F-statistic value will tell whether the result is good or not.

#answer: the f statistic value is 10.81 and P value is less than 0.05 so we c
an say that  by judging by f value result is good.
```

T-test formula for Intercept

$$t_{b_0} = \frac{b_0 - \beta_0}{s_{b_0}}$$

T-test formula for Slope

$$t_{b_1} = \frac{b_1 - \beta_1}{s_{b_1}}$$
