pubh7450-assign0

```
library(ggplot2)
```

#1.load data

```
chicks <- read.csv('C:/Users/leonz/Desktop/UMN/Fall 2021/PubH7430 - Statistical methods for correlated c
head(chicks)
```

```
##   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
```

```
chicks$Chick <- as.factor(chicks$Chick)
chicks$Diet <- as.factor(chicks$Diet)
```

#2.

```
class(chicks$Chick)
```

```
## [1] "factor"
```

```
length(unique(chicks$Chick))
```

```
## [1] 50
```

There are 50 chicks

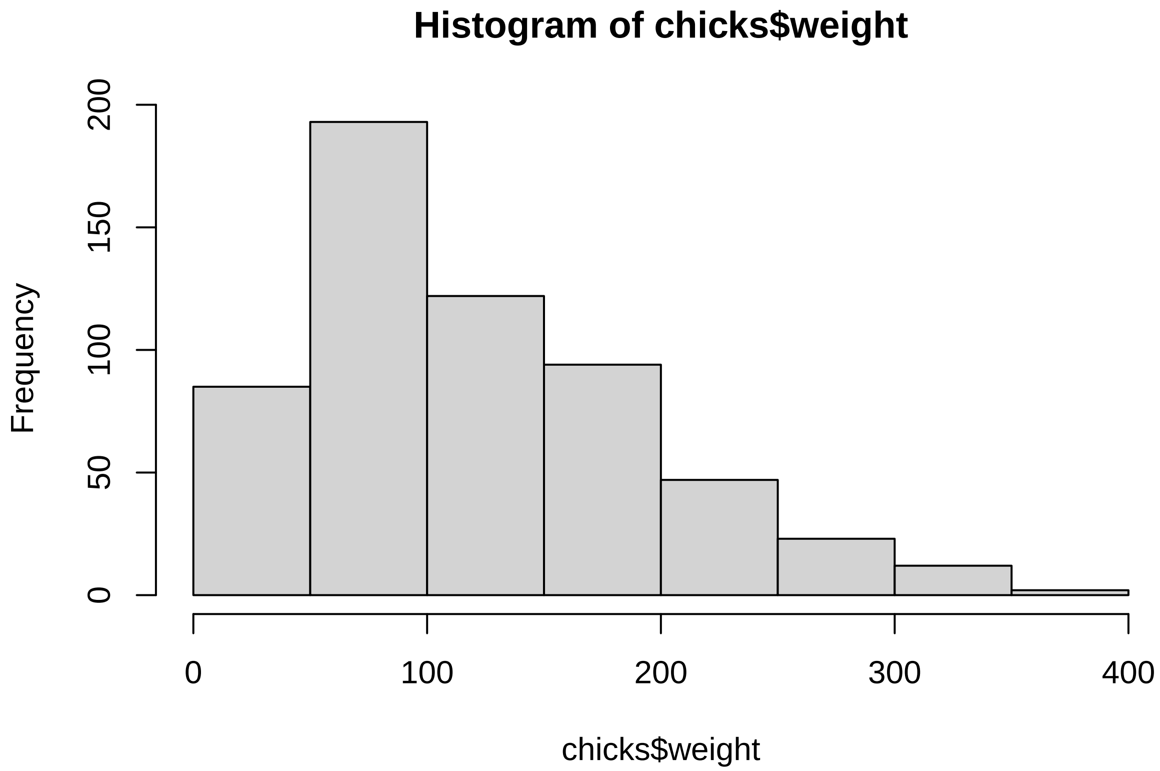
```
summary(chicks$Chick)
```

```
##  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
## 12 12 12 12 12 12 12 11 12 12 12 12 12 12  8  7 12  2 12 12 12 12 12 12 12 12
## 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50
## 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 10 12 12 12 12 12 12
```

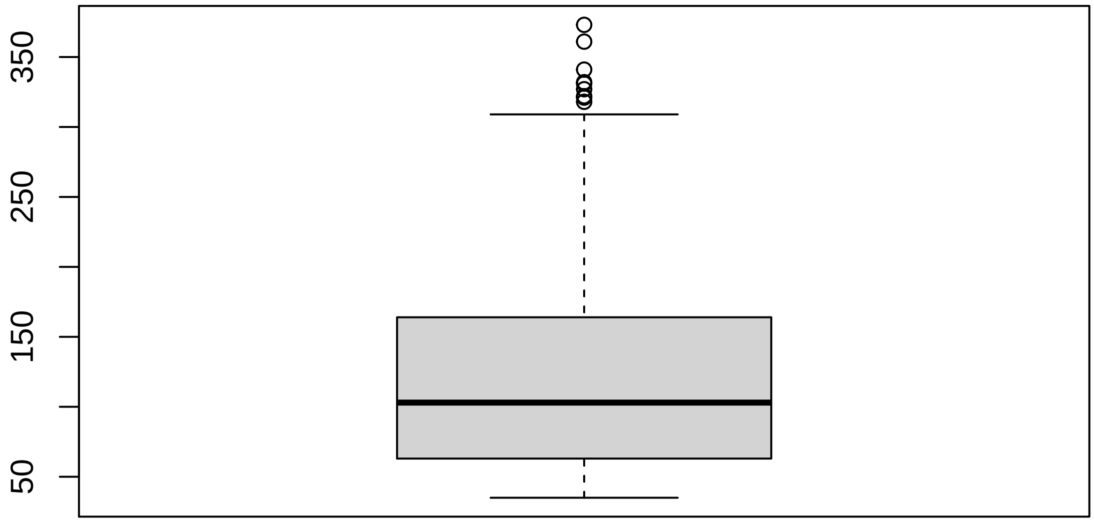
For most of the chicks, there are 12 repeated measures. However, there are some missing measures. For example, chick 15 only has 8 measures, chick 18 only has 2 measures.

#3. outliers

```
hist(chicks$weight)
```



```
boxplot(chicks$weight)
```



```
summary(chicks$weight)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    35.0    63.0   103.0   121.8   163.8   373.0
```

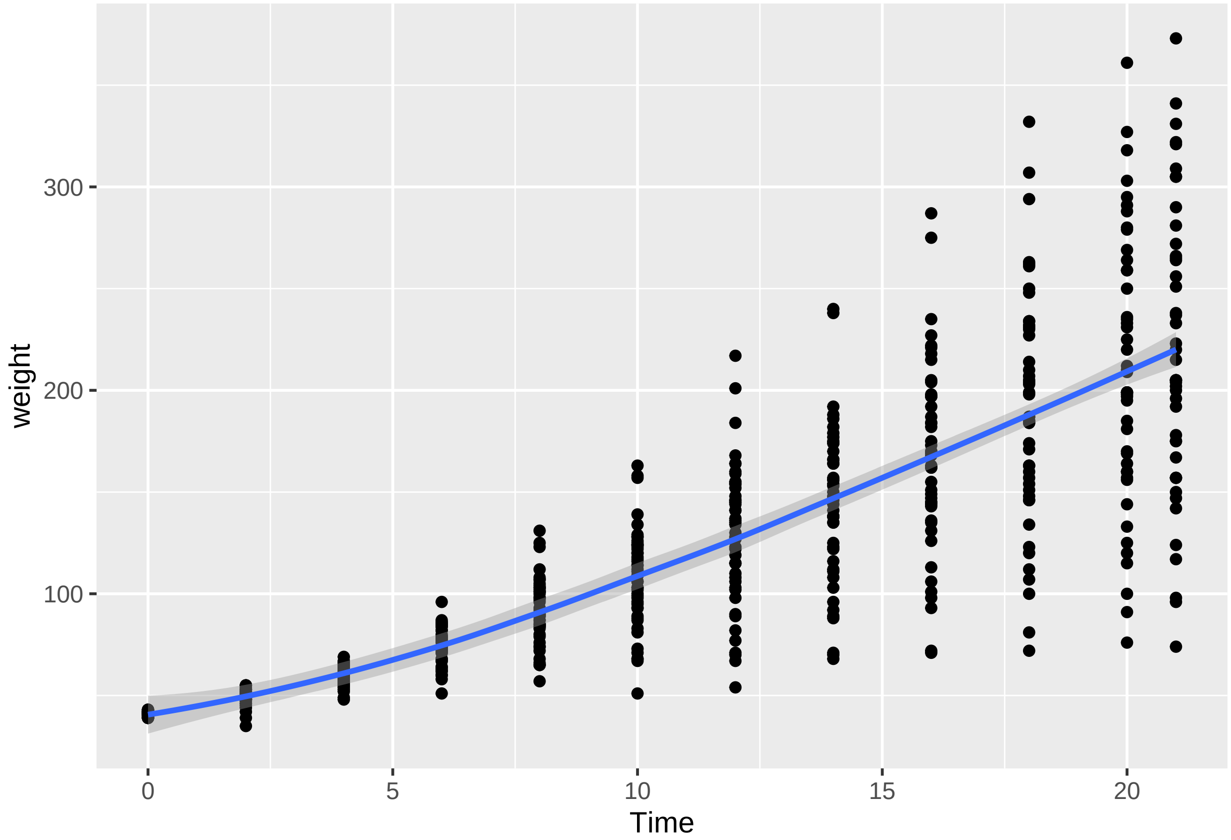
Seems like there are some outliers present in this data

#4. weight change overtime

Overall

```
ggplot(data = chicks, aes(x = Time, y = weight)) +
  geom_point() +
  geom_smooth()
```

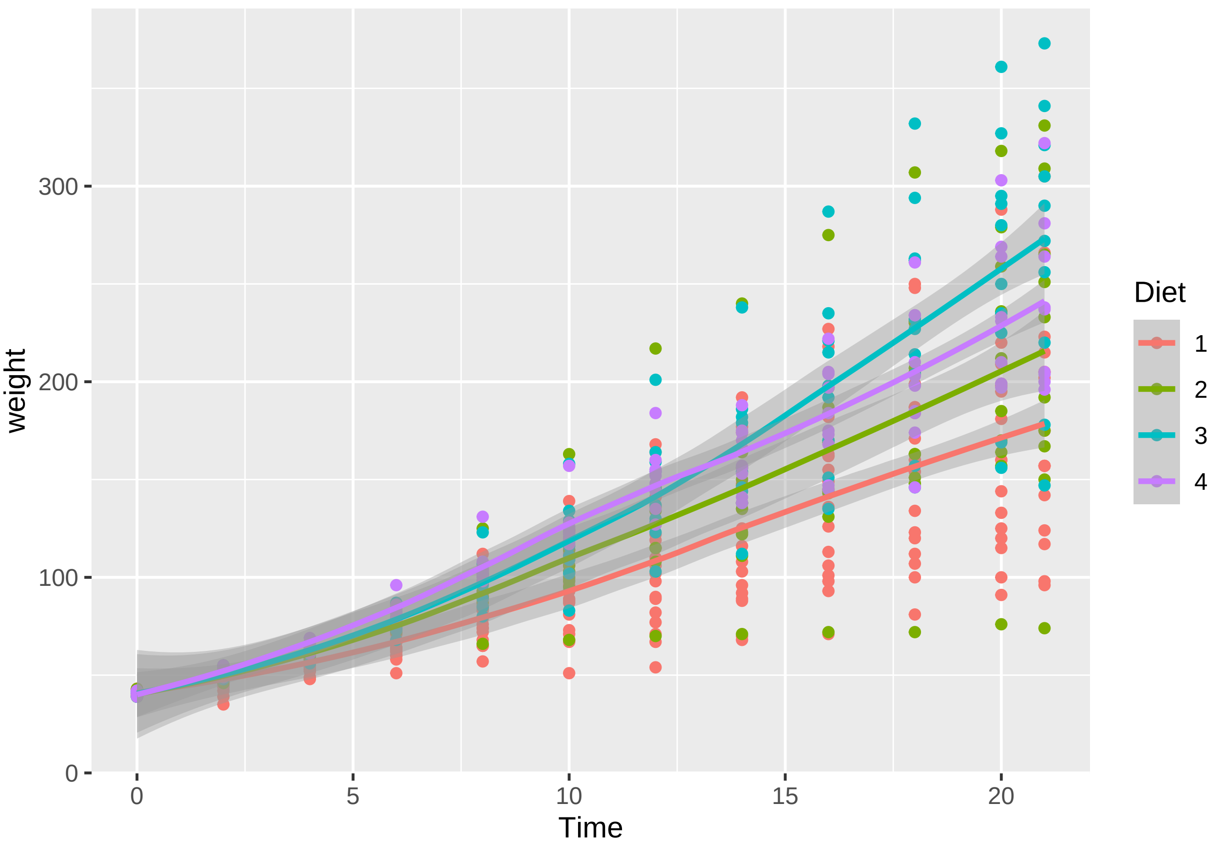
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Overall, weight increases as time increases

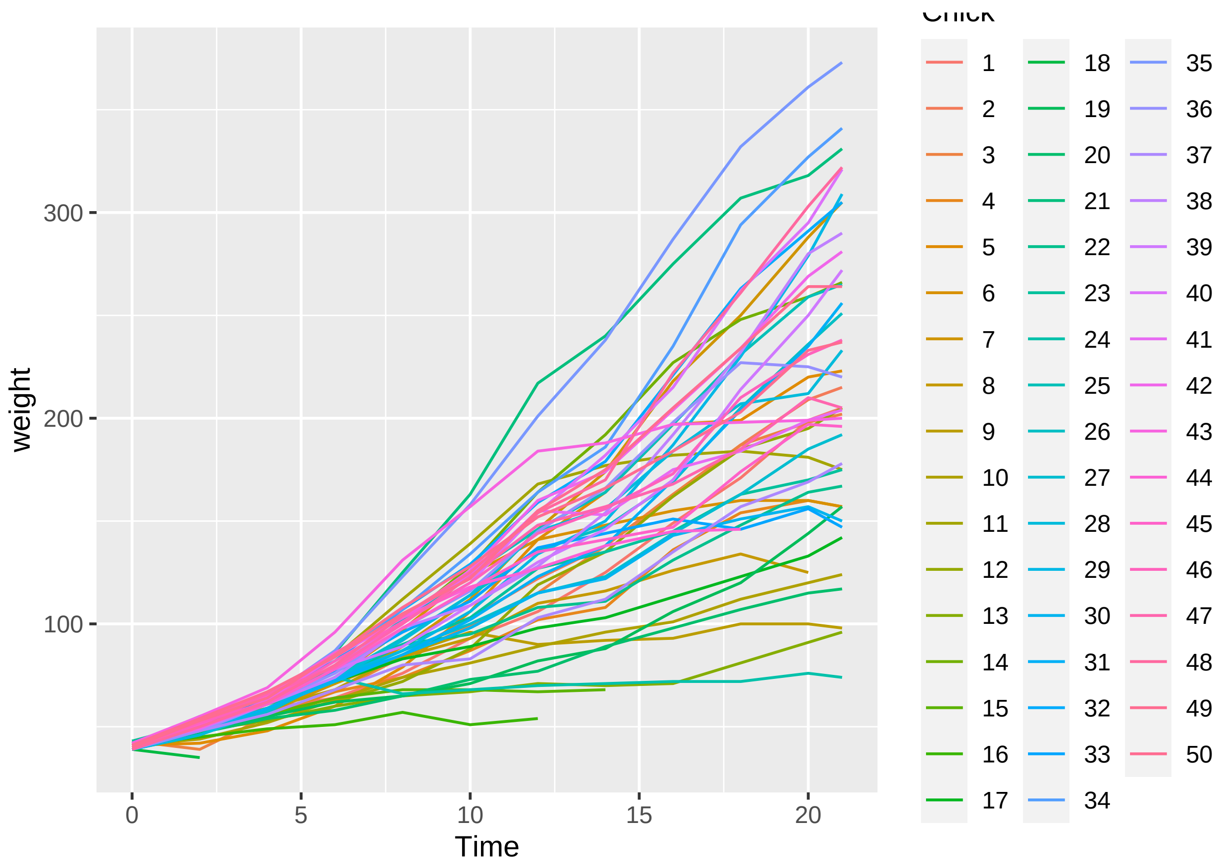
```
ggplot(data = chicks, aes(x = Time, y = weight, color = Diet)) +
  geom_point() +
  geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



##individual

```
gg.base <- ggplot(chicks, aes(x = Time, y = weight))
gg.idline <- gg.base + geom_line(aes(color = Chick, group = Chick))
gg.idline
```



#5. regression

If we assume each row is independent

```
lm1 <- lm(weight ~ Time, data = chicks)
summary(lm1)
```

```
##
## Call:
## lm(formula = weight ~ Time, data = chicks)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -138.331  -14.536    0.926   13.533  160.669
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  27.4674     3.0365   9.046   <2e-16 ***
## Time          8.8030     0.2397  36.725   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 38.91 on 576 degrees of freedom
## Multiple R-squared:  0.7007, Adjusted R-squared:  0.7002
## F-statistic:  1349 on 1 and 576 DF,  p-value: < 2.2e-16
```

```
lm2 <- lm(weight ~ Diet, data = chicks)
summary(lm2)
```

```
##
## Call:
## lm(formula = weight ~ Diet, data = chicks)
##
## 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
```

```
lm3 <- lm(weight ~ Time + Diet, data = chicks)
summary(lm3)
```

```
##
## Call:
## lm(formula = weight ~ Time + Diet, data = chicks)
##
## 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
```

```
anova(lm1,lm3)
```

```
## Analysis of Variance Table
##
## Model 1: weight ~ Time
## Model 2: weight ~ Time + Diet
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    576 872212
## 2    573 742336  3    129876 33.417 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Seems like, if we assume each row is independent, variable Time and Diet are both statistically significant with p-values less than 0.05.