The Joy of Plotting

Vincent D. Warmerdam `@koaning` `koaning.io`

this **vincent**.

who is this guy?

THE WIFI SUCKS HERE
WANT MY MONEY BACK.

It goes to rediculous lengths.

LinkedIn

The keyword bingo:

`python, javascript, R, dplyr, ggplot, ditto, scala, mongo, html5, bootstrap, git, sublime, d3, leaflet, sawk, pandas, feebas, numpy, scikit, nltk, crebase, juypter, onyx, lodash, mesos, docker, django, flask, neo4j, vulpix, selenium, node-webkit, hadoop, hoopa, impala, spark, azurill, ansible, hadoop, mapreduce.`

It goes to rediculous lengths.

LinkedIn

python, javascript, R, dplyr, ggplot, ditto, scala, mongo, html5, bootstrap, git, sublime, d3, leaflet, sawk, pandas, feebas, numpy, scikit, nltk, crebase, juypter, onyx, lodash, mesos, docker, django, flask, neo4j, vulpix, selenium, node-webkit, hadoop, hoopa, impala, spark, azurill, ansible, hadoop, mapreduce.

Can you recognize which of these are pokemon?

I'm also involved in conferences

Leftover Speaker Bio

GoDataDriven
proudly part of Xebia Group

… Vincent has a blog about less obvious aspects in the world of data science over at koaning.io and he’s known for giving free lectures in data science around Europe. He’s especially keen to give one in Lisbon, Ljubljana, Split or Belfast. Feel free to notify Vincent if you’ve got a venue in any of these places (or other awesome place to visit) as he’s **actually** bound to show up.

About today

- we'll pretend that we're consultants
- not enterprise kind, the kind that actually solves problems
- we are presenting to humans
- you can type along but attention is better
- the goal is to demo the ggplot functionality
- … maybe with an uncommon pattern

Chick Weight Case

- help out farmer Fred
- he gave chickens different food
- we want fat chickens
- A/B/C/D test

I've got data.

Now what?

I've got data.

Now what?

Why not just look at it?

What is the data like?

What do you think this code does?

```
> head(ChickWeight)
```

What is the data like?

What do you think this code does?

```
> 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
```

Summary of the data.

What do you think this code does?

```
> summary(ChickWeight)
```

Summary of the data.

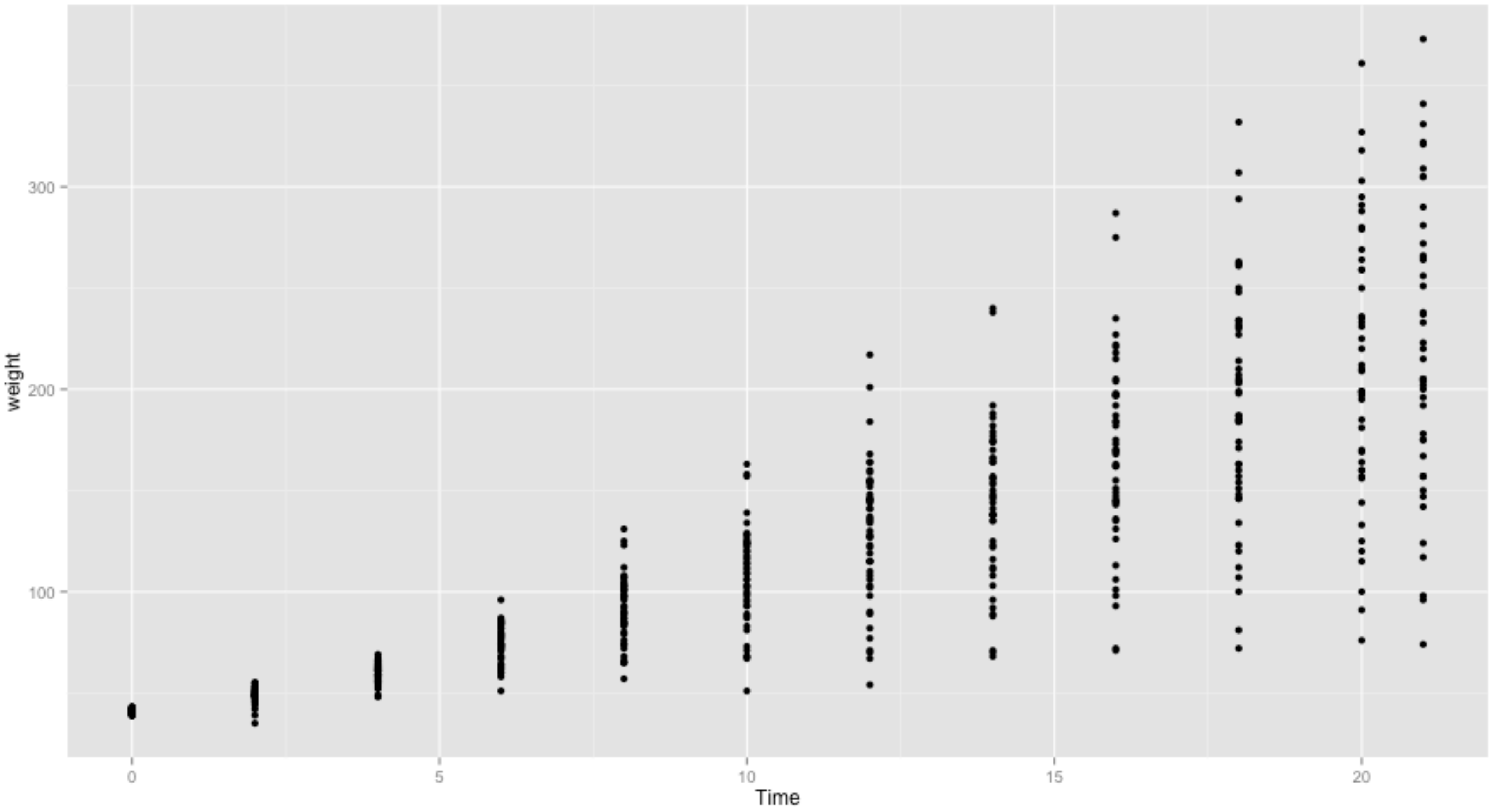
What do you think this code does?

```
> 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
```

Look at the data.

What do you thinkg this code does?

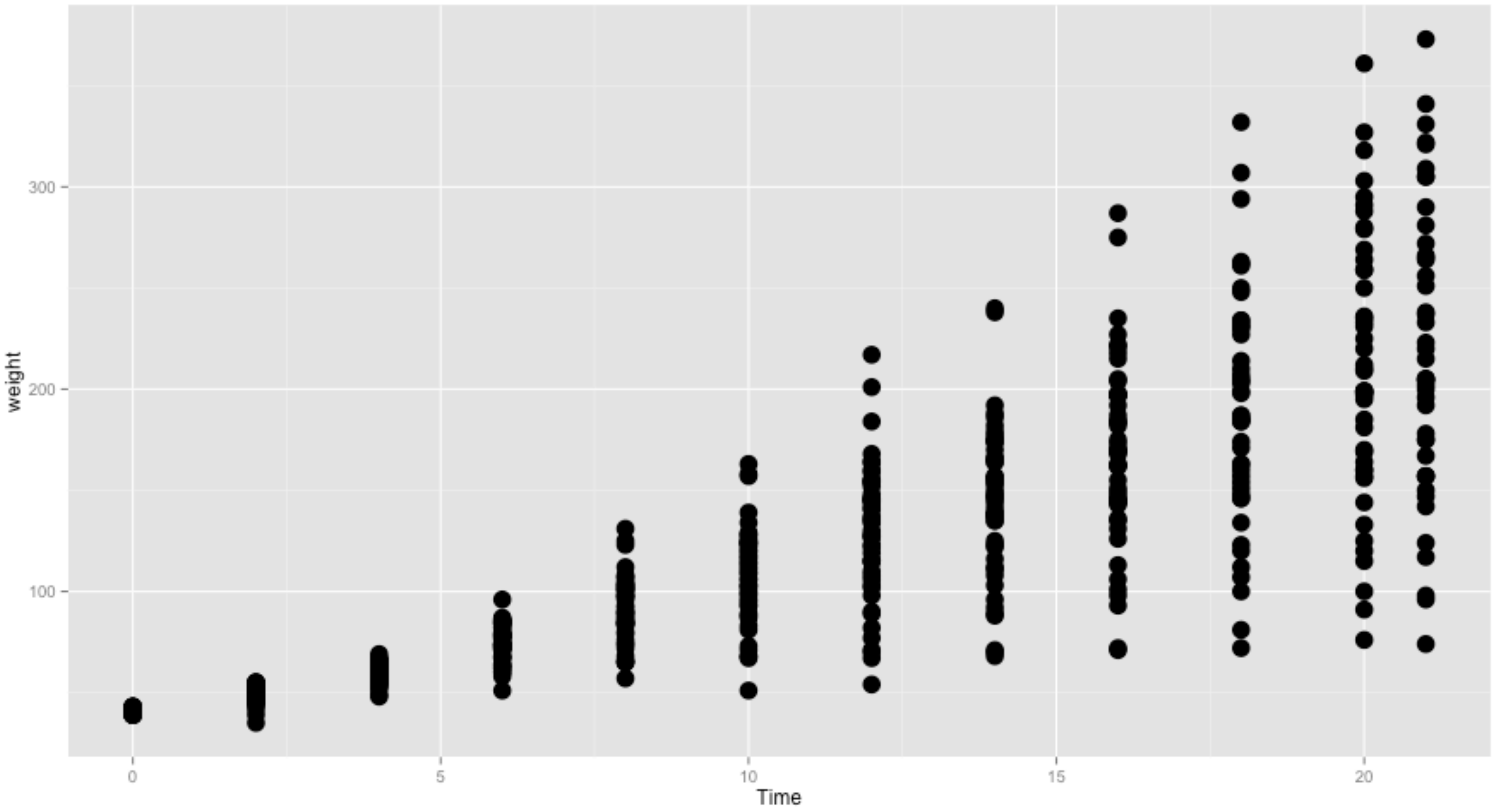
```
p <- ggplot()
p + geom_point(
  data=ChickWeight,
  aes(x=Time, y=weight)
)
```


300
200
100
weight
0
5
10
15
20
Time

Look at the data.

What do you thinkg this code does?

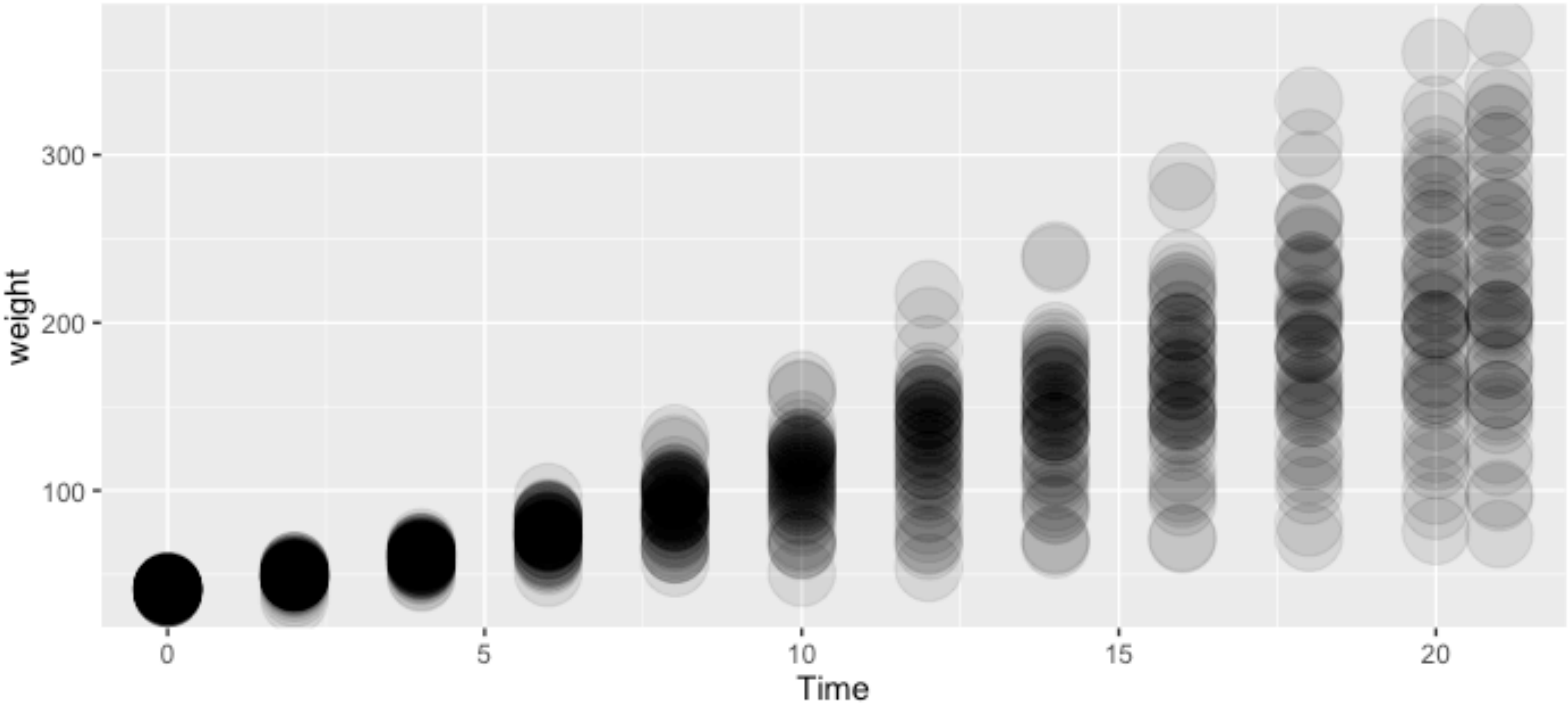
```
p <- ggplot()
p + geom_point(
  data=ChickWeight,
  aes(x=Time, y=weight),
  size = 5
)
```


300
200
100
weight
0
5
10
15
20
Time

Look at the data.

What do you thinkg this code does?

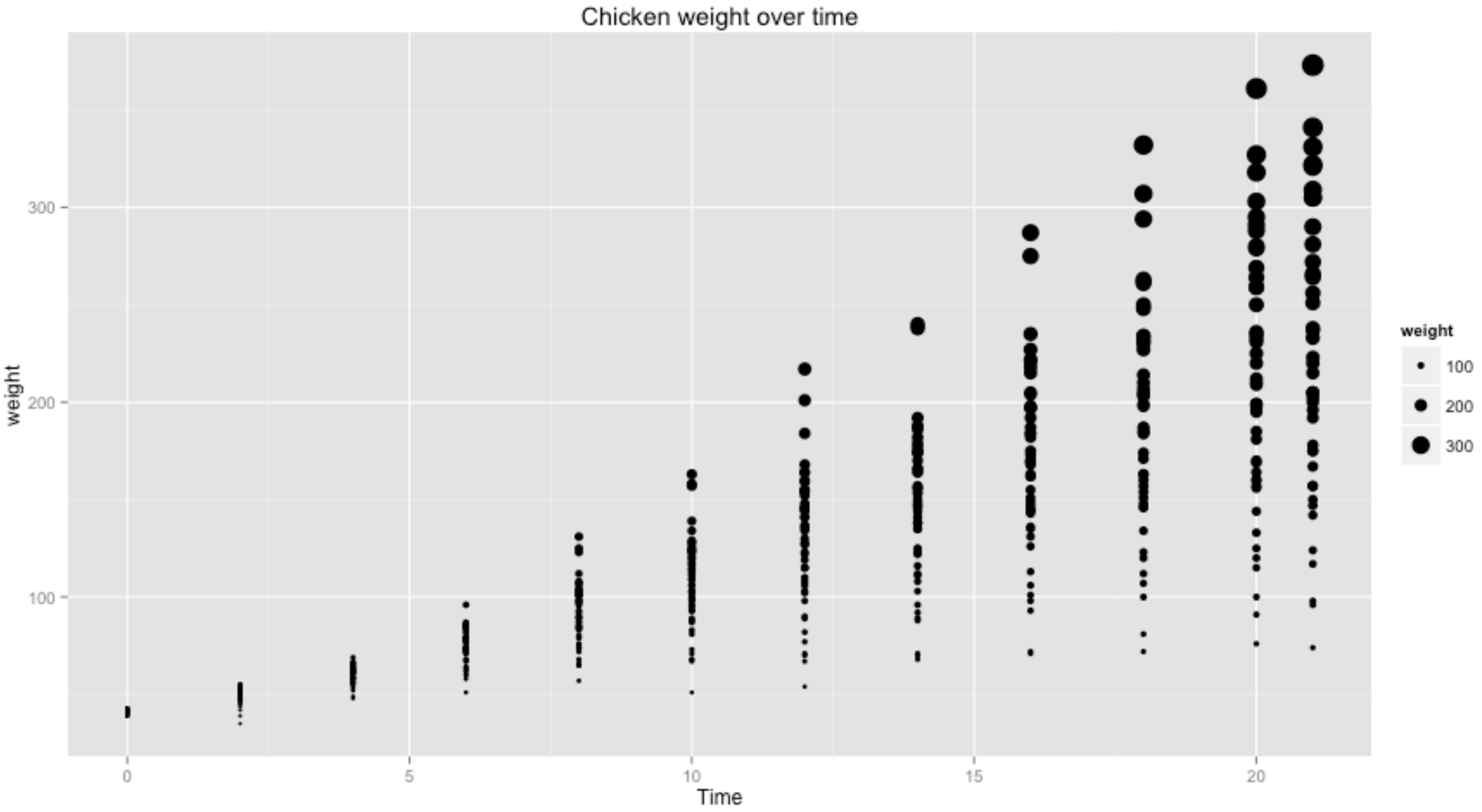
```
p <- ggplot()
p + geom_point(
  data=ChickWeight,
  aes(x=Time, y=weight),
  size = 10, alpha = 0.1
)
```


weight
300
200
100
0
5
10
15
20
Time

Look at the data.

What do you thinkg this code does?

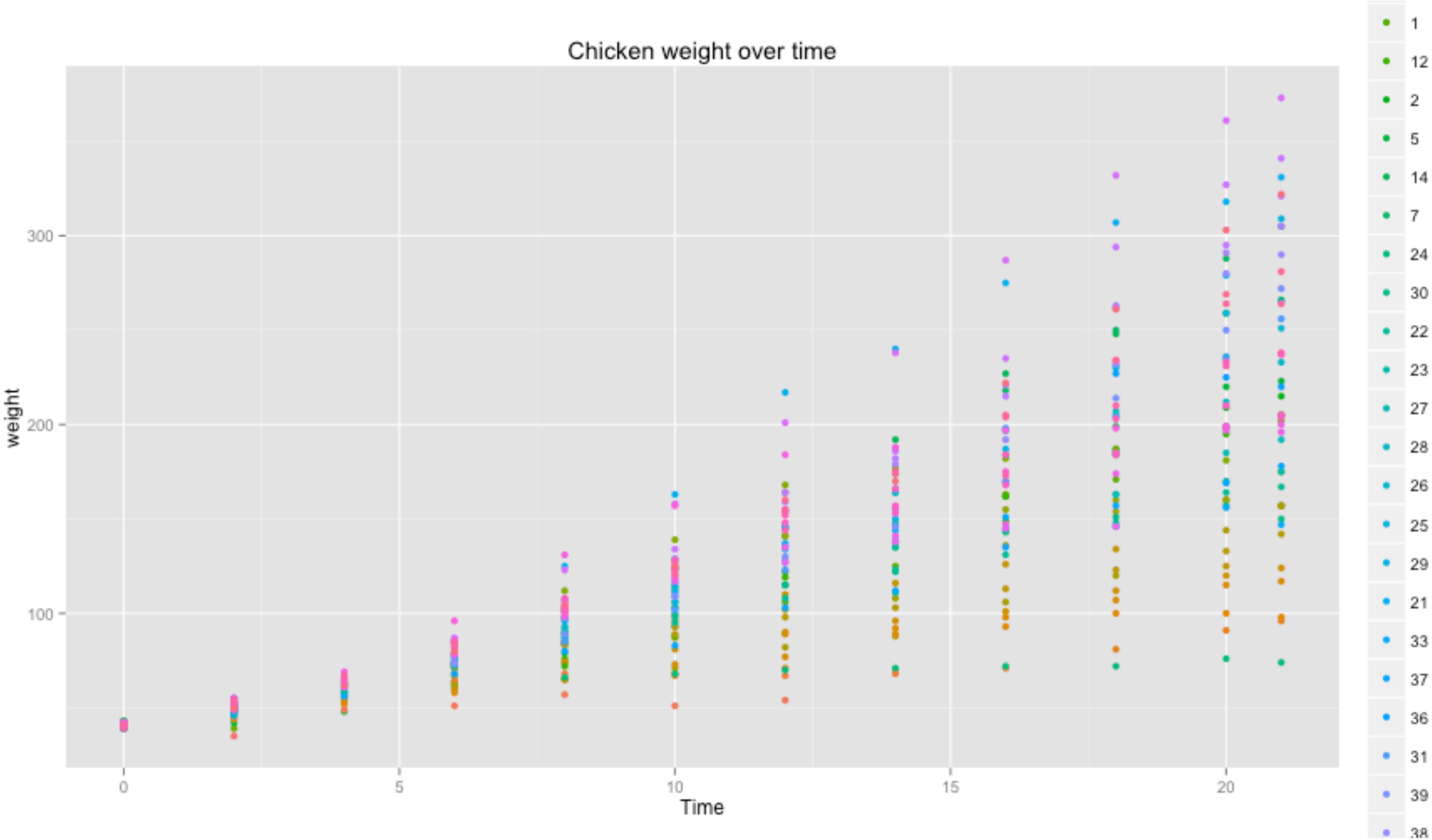
```
p <- ggplot()
p + geom_point(
  data=ChickWeight,
  aes(x=Time, y=weight, size=weight)
)
```


Chicken weight over time
weight
300
200
100
0
5
10
15
20
Time
weight
100
200
300

Look at the data.

What do you thinkg this code does?

```
p <- ggplot()
p = p + geom_point(
  data=ChickWeight,
  aes(x=Time, y=weight, colour=Chick)
)
p + ggtitle("Chicken weight over time")
```


Chicken weight over time
weight
300
200
100
0
5
10
15
20
Time
1
12
2
5
14
7
24
30
22
23
27
28
26
25
29
21
33
37
36
31
39
38

What makes the previous plot terrible?

Remove

to improve

(the **data-ink** ratio)

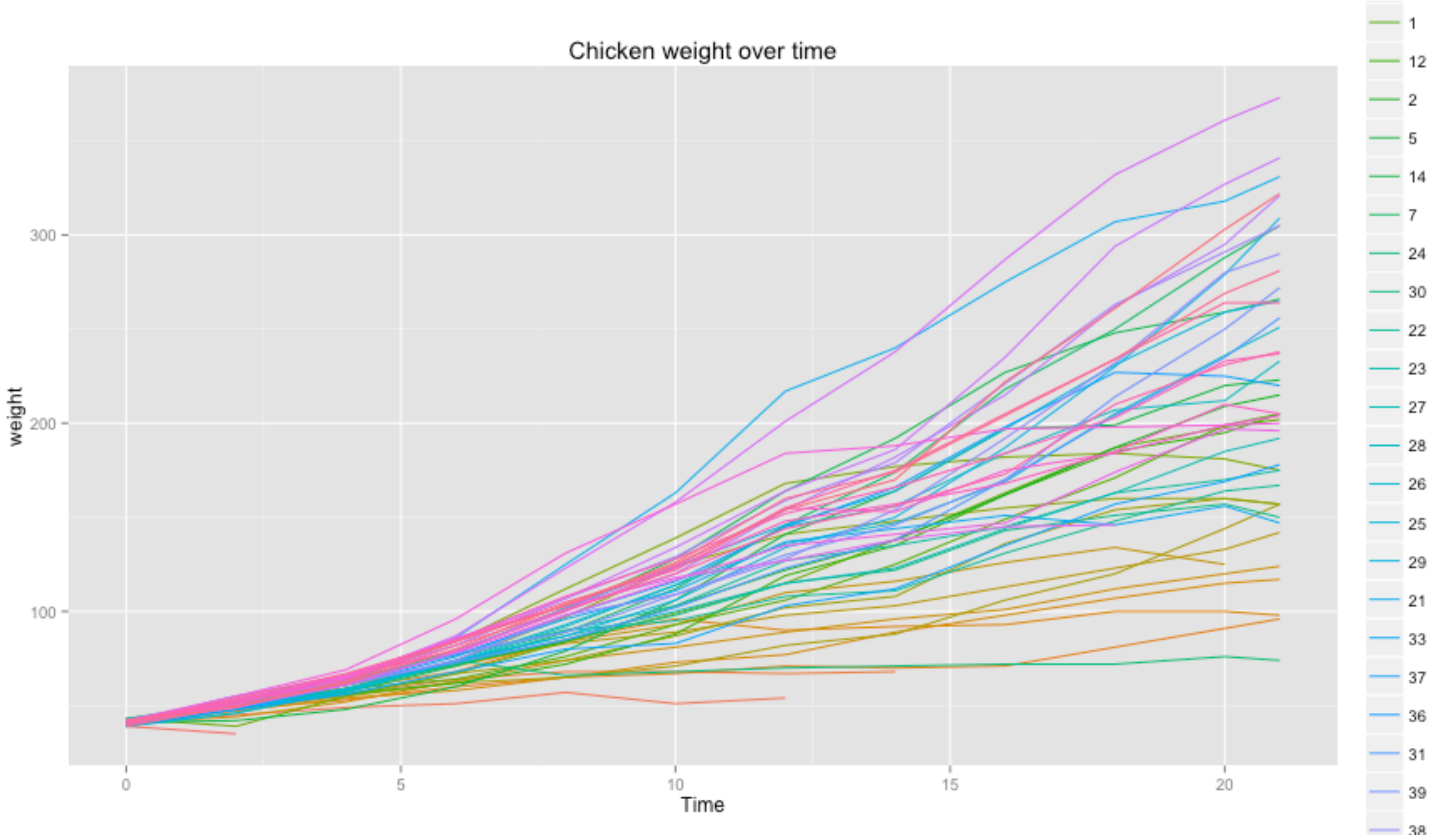
Created by **Darkhorse Analytics**

www.**darkhorseanalytics**.com

Look at the data.

What do you thinkg this code does?

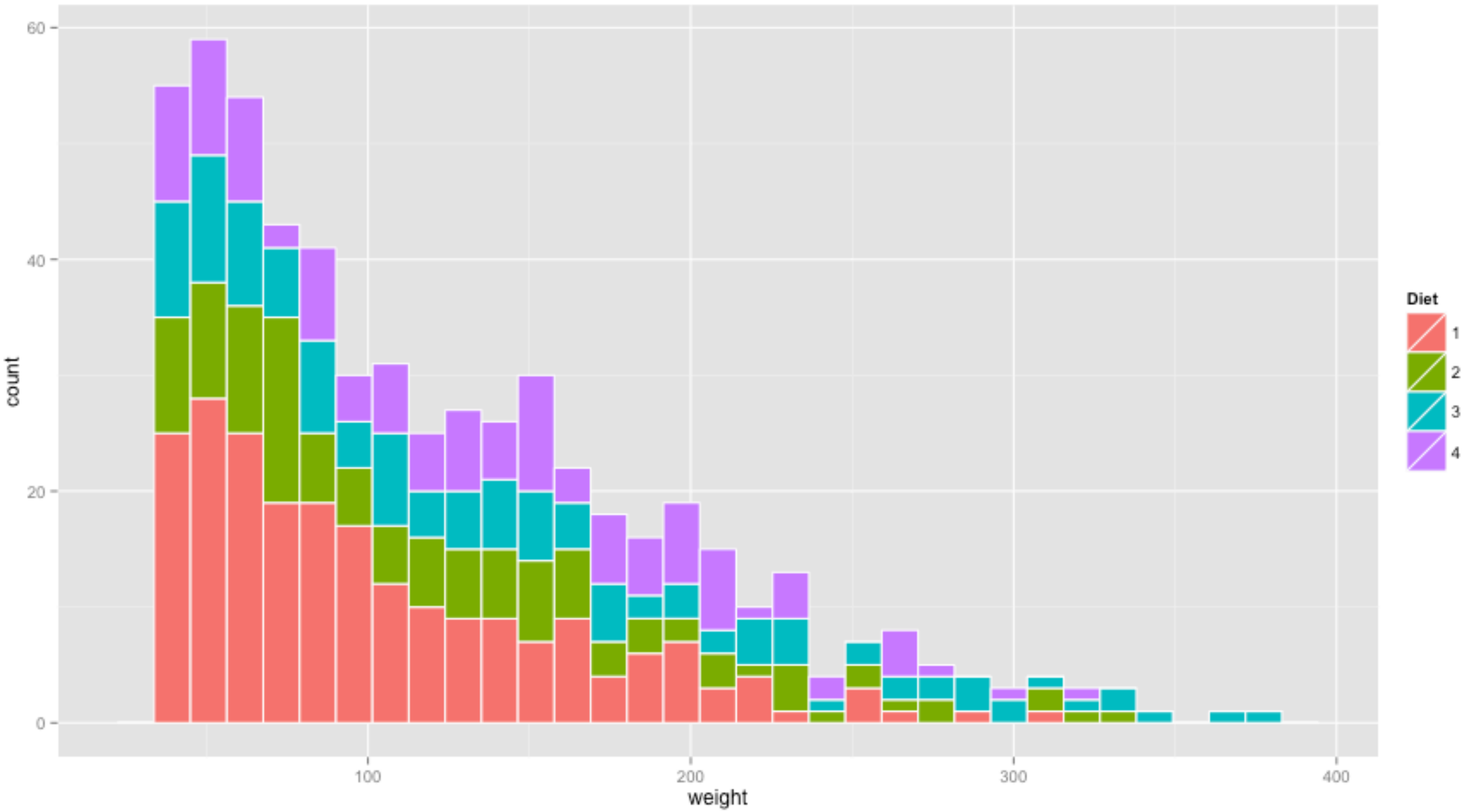
```
p <- ggplot()
p <- p + geom_line(
  data=ChickWeight,
  aes(x=Time, y=weight, colour=Chick)
)
p + ggtitle("Chicken weight over time")
```


Chicken weight over time
weight
300
200
100
Time
0
5
10
15
20
1
12
2
5
14
7
24
30
22
23
27
28
26
25
29
21
33
37
36
31
39
38

Look at the data.

What do you think this code does?

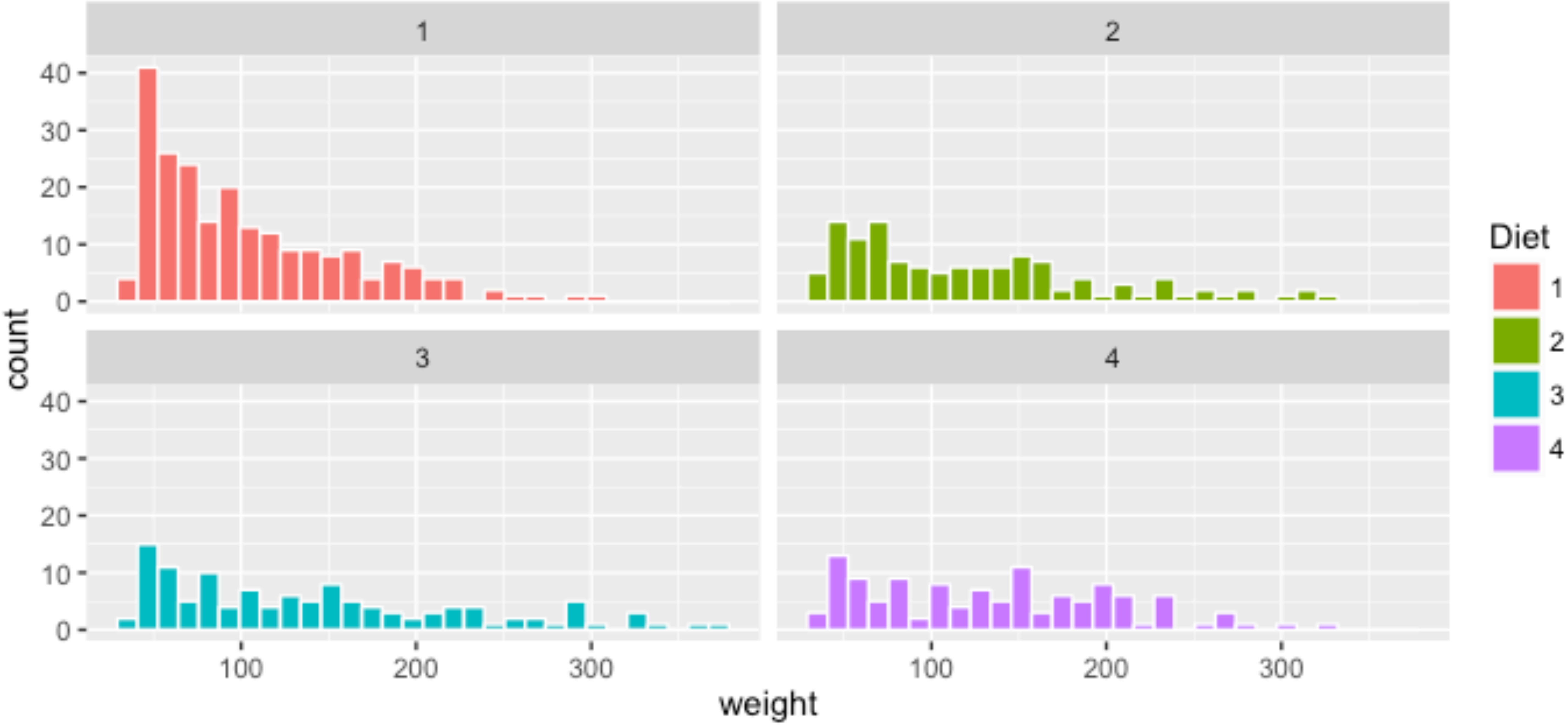
```
p <- ggplot()
p + geom_histogram(
  data=ChickWeight,
  aes(x=weight, fill=Diet),
  colour="white"
)
```


60
40
20
0
count
100
200
300
400
weight
Diet
1
2
3
4

Look at the data.

What do you think this code does?

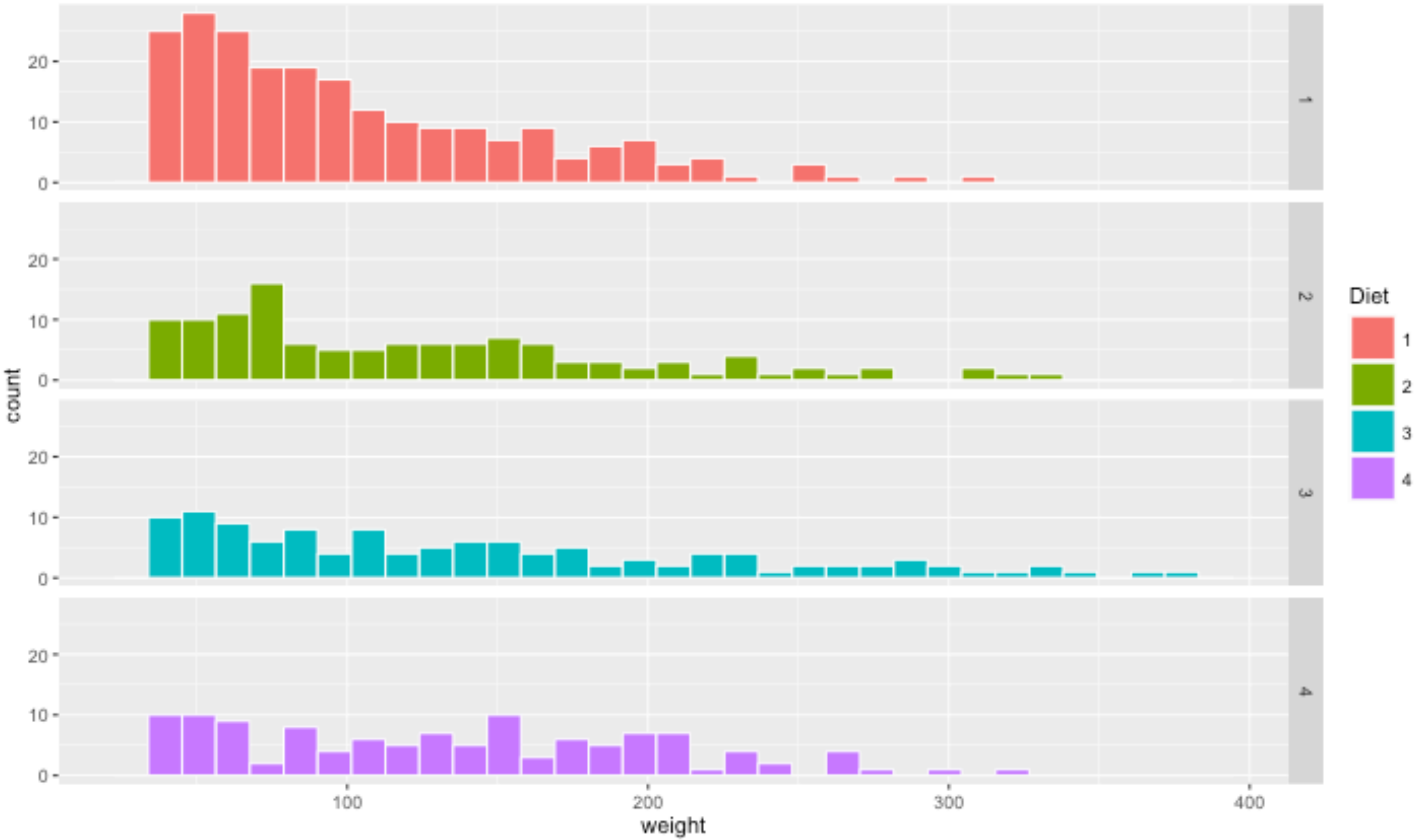
```
p <- ggplot()
p <- p + geom_histogram(
  data=ChickWeight,
  aes(x=weight, fill=Diet),
  colour="white"
)
p + facet_wrap(~ Diet)
```


1
2
3
4
count
40
30
20
10
0
100
200
300
weight
Diet
1
2
3
4

Look at the data.

What do you think this code does?

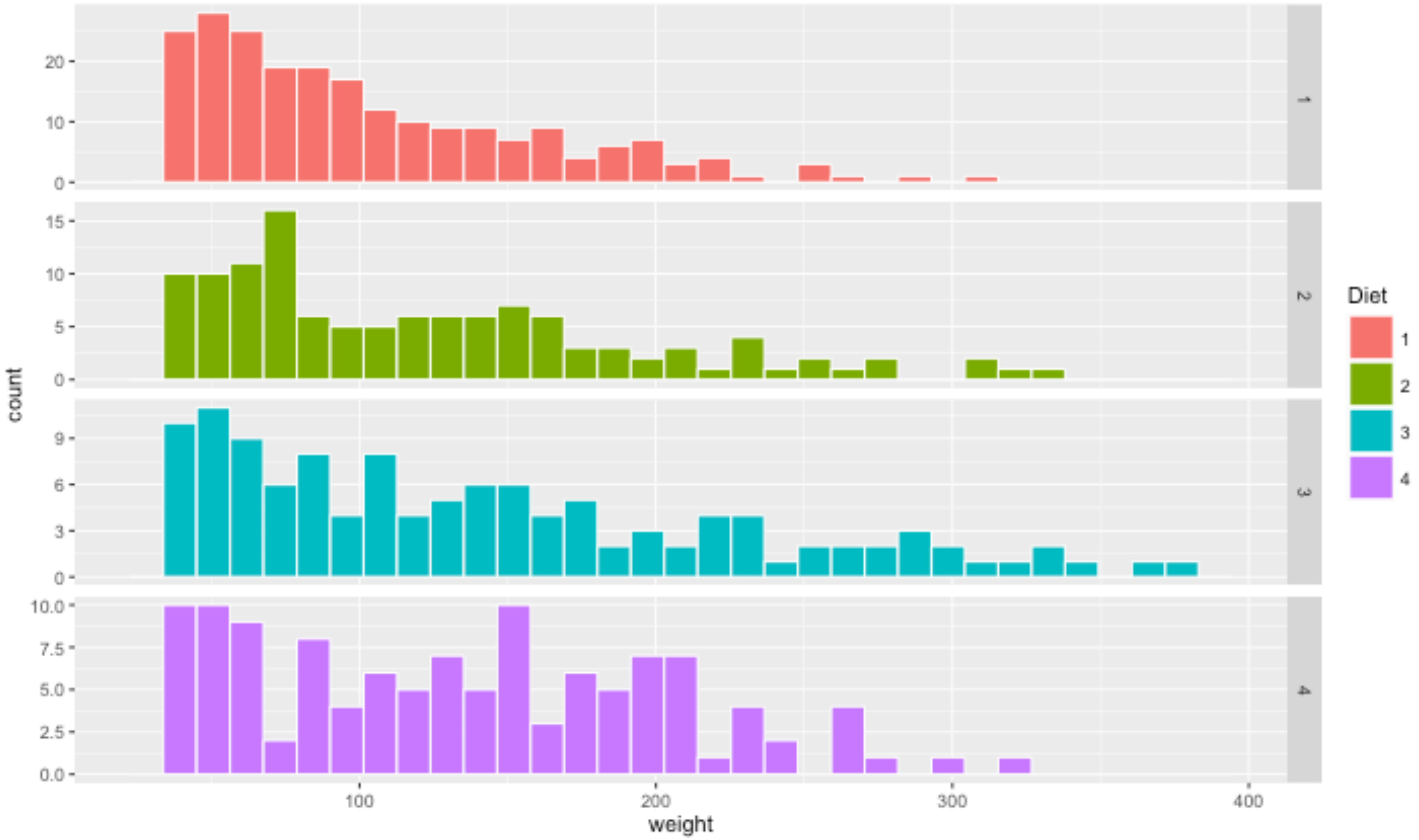
```
p <- ggplot()
p <- p + geom_histogram(
  data=ChickWeight,
  aes(x=weight, fill=Diet),
  colour="white"
)
p + facet_grid(Diet ~ .)
```


count
20
10
0
1
2
3
4
100
200
300
400
weight
Diet
1
2
3
4

Look at the data.

What do you think this code does?

```
p <- ggplot()
p <- p + geom_histogram(
  data=ChickWeight,
  aes(x=weight, fill=Diet),
  colour="white"
)
p + facet_grid(Diet ~ ., scales = "free_y")
```


count
20
10
0
15
10
5
0
9
6
3
0
10.0
7.5
5.0
2.5
0.0
1
2
3
4
100
200
300
400
weight
Diet
1
2
3
4

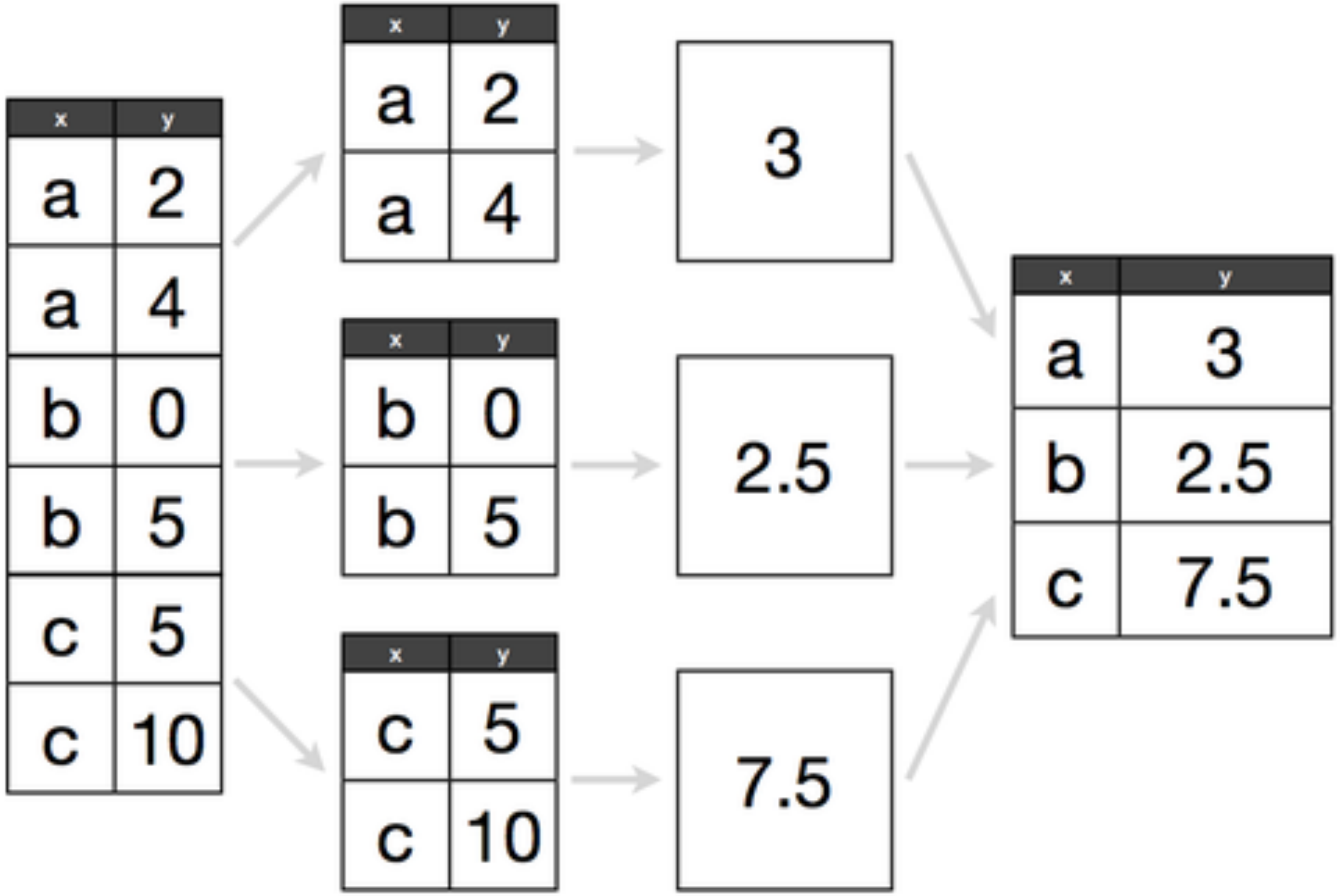
Next steps

It seems obvious that we want to calculate something per category.

In real life, we do this **all** the time.

The category in this case a chickens, but this is something that is similar for any A/B-test.

Split - Apply - Combine



x	y
a	2
a	4
b	0
b	5
c	5
c	10

x	y
a	2
a	4

x	y
b	0
b	5

x	y
c	5
c	10

x	y
a	3
b	2.5
c	7.5

Crunch the data.

What do you think this code does?

```
ChickWeight %>%
  group_by(Diet) %>%
  summarise(m = mean(weight))
```

Crunch the data.

```
  Diet        m
1    1 102.6455
2    2 122.6167
3    3 142.9500
4    4 135.2627
```

Crunch the data.

What do you think this code does?

```
ChickWeight %>%
  group_by(Diet, Time) %>%
  summarise(m = mean(weight))
```

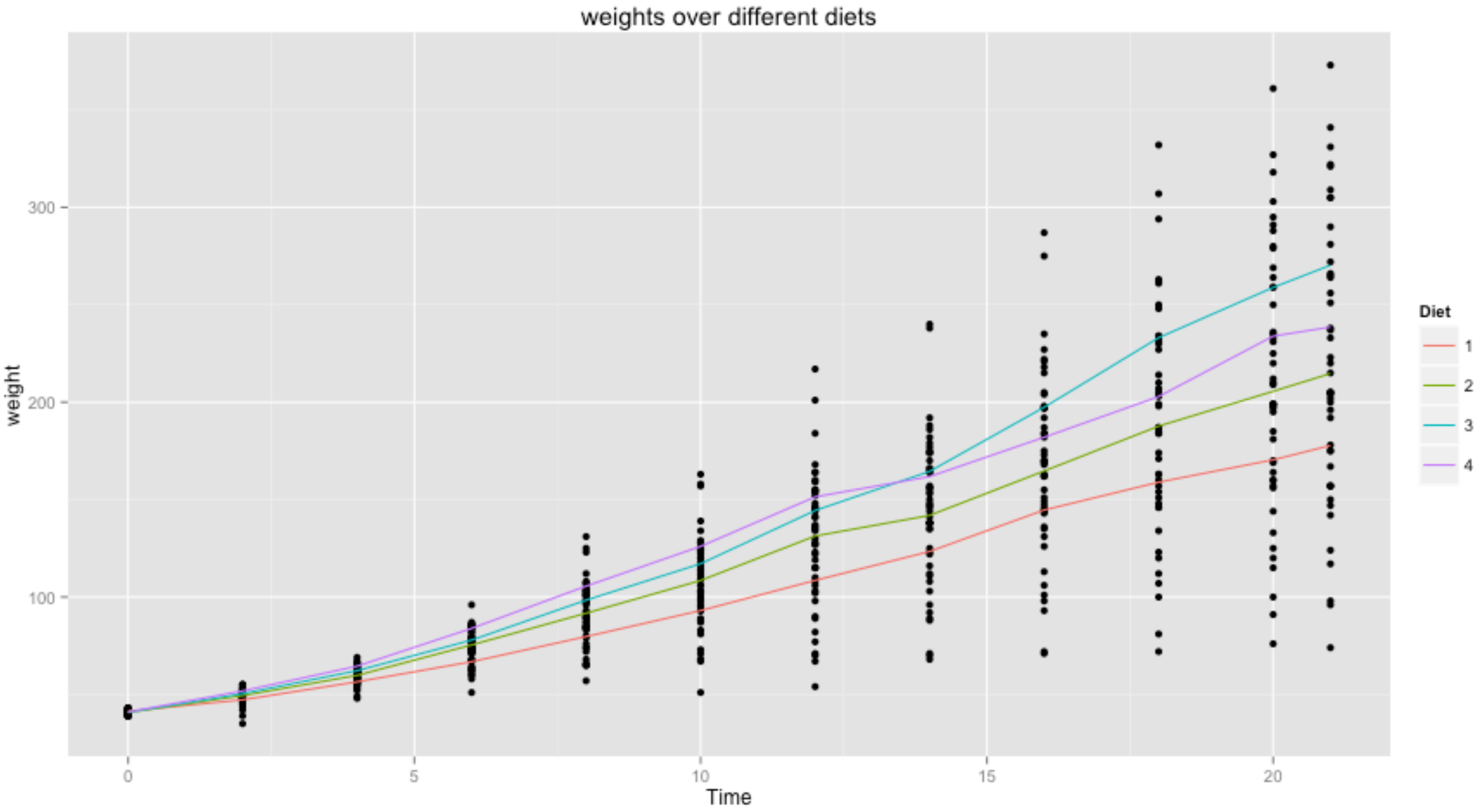
Crunch the data.

```
   Diet Time    weight
1     1    0  41.40000
2     2    0  40.70000
3     3    0  40.80000
4     4    0  41.00000
5     1    2  47.25000
6     2    2  49.40000
7     3    2  50.40000
8     4    2  51.80000
9     1    4  56.47368
10    2    4  59.80000
...
```

Combination

```
agg <- ChickWeight %>%
  group_by(Diet, Time) %>%
  summarise(m = mean(weight))

p <- ggplot()
p <- p + geom_point(data=ChickWeight, aes(Time, weight))
p <- p + geom_line(data=agg, aes(Time, m, colour=Diet))
p + ggtitle("weights over different diets")
```


weights over different diets
weight
300
200
100
Time
0
5
10
15
20
Diet
1
2
3
4

If there is time left:

- Monopoly Story
- Lego Story
- World of Warcraft story
- Heroes Dataset -> Practice!
- Cigarette Dataset -> Practice!

Time to Practice?

```
library(readr)
library(dplyr)
library(tidyr)
library(ggplot2)

heroes_df <- read_csv("http://koaning.io/theme/data/heroes.csv")

pltr <- heroes_df %>%
    na.omit() %>%
    group_by(role) %>%
    summarise(attack = mean(attack), hp = mean(hp)) %>%
    gather(key, value, -role)

ggplot() +
  geom_bar(data=pltr, aes(role, value), stat="identity") +
  facet_grid(key ~ ., scales = "free")
```

Time to Practice?

```
cig_df <- read_csv("http://koaning.io/theme/data/cigarette.csv")
```

What is the effect of taxes on cigarette consumption?

Explain it visually!

If you like what you've heard today, feel free to check the blog.

koaning.io, **@fishnets88**

I'm around the conference, ask me anything!