Effective Data Visualization for Actuaries

Jordan Bonner & Brian A. Fannin

Antitrust Notice

- The Casualty Actuarial Society is committed to adhering strictly to the letter and spirit of the antitrust laws. Seminars conducted under the auspices of the CAS are designed solely to provide a forum for the expression of various points of view on topics described in the programs or agendas for such meetings.

- Under no circumstances shall CAS seminars be used as a means for competing companies or firms to reach any understanding – expressed or implied – that restricts competition or in any way impairs the ability of members to exercise independent business judgment regarding matters affecting competition.

- It is the responsibility of all seminar participants to be aware of antitrust regulations, to prevent any written or verbal discussions that appear to violate these laws, and to adhere in every respect to the CAS antitrust compliance policy.

Effective Data Visualization for Actuaries

Jordan Bonner & Brian A. Fannin

How is visualization useful?

Visualization is a tool which facilitates communication with the less numerate.

ACTUARY

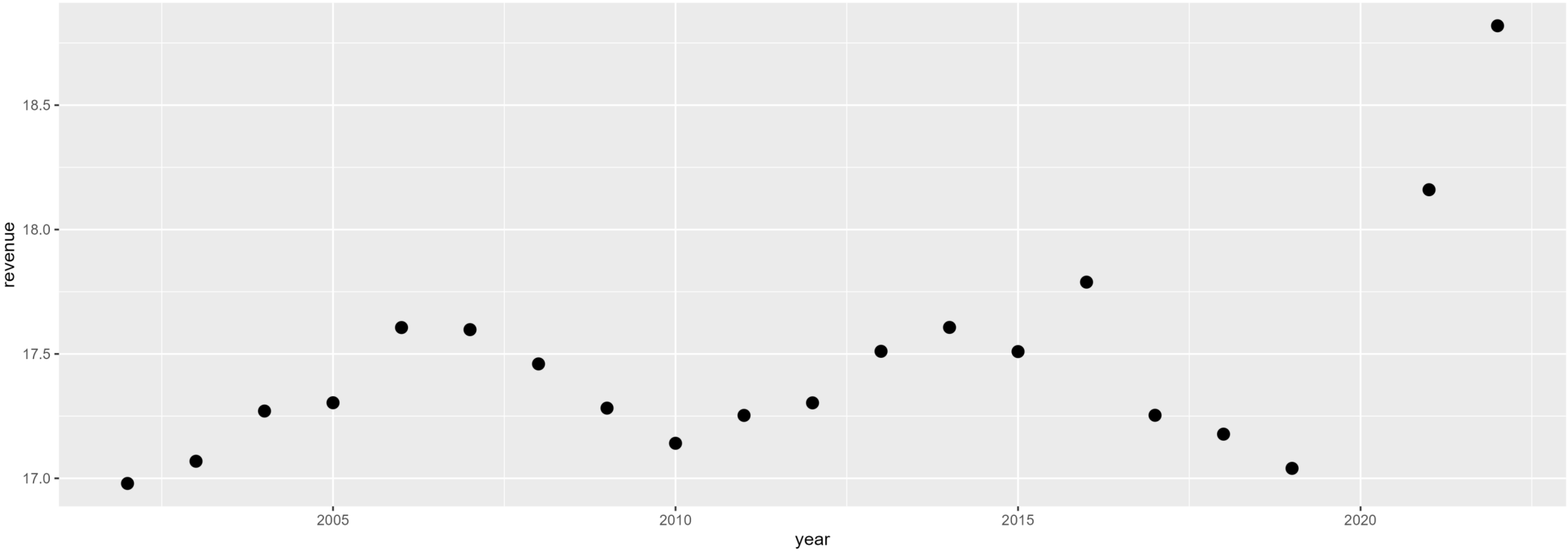
STAKEHOLDER

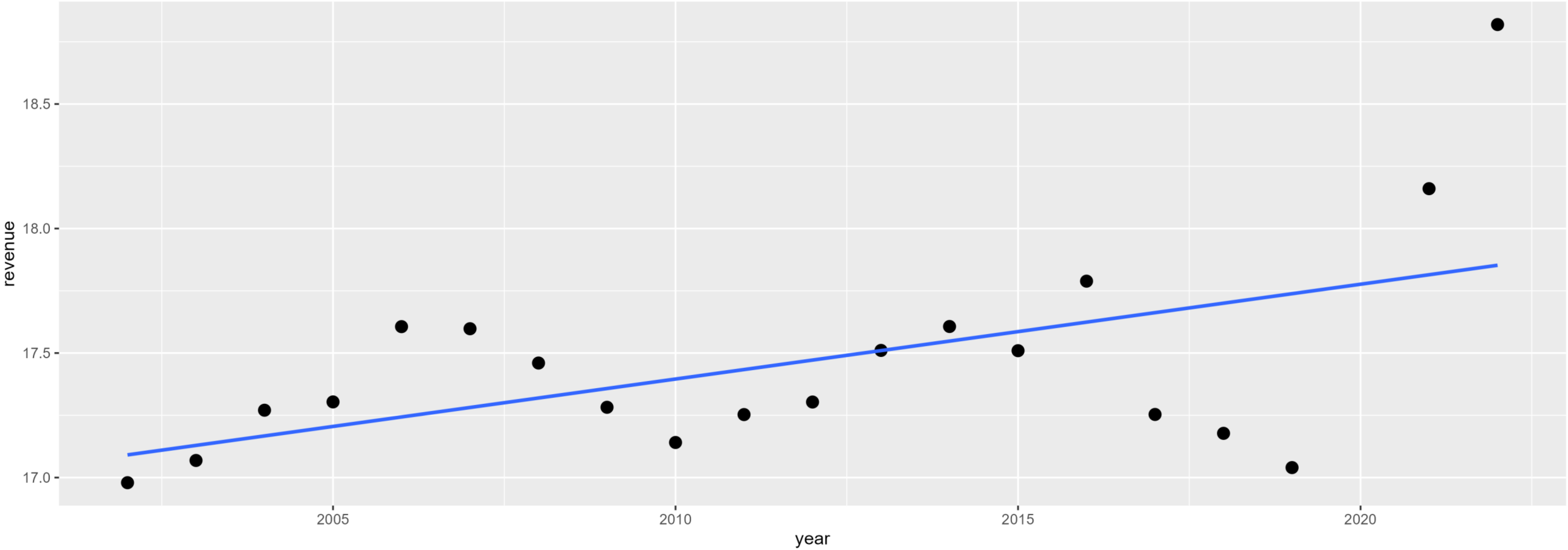
Describe how linear regression works using only equations or numbers.

$$\widehat{y_i} = \beta_0 + \sum_{j=0}^{p} \beta_j X_{ij} + \epsilon_{ij}$$

$$\epsilon_{ij} \sim N(0, \sigma^2)$$

$$\hat{\beta} = \operatorname{argmin} \sum_{j=0}^{p} (y_i - \widehat{y_i})^2$$

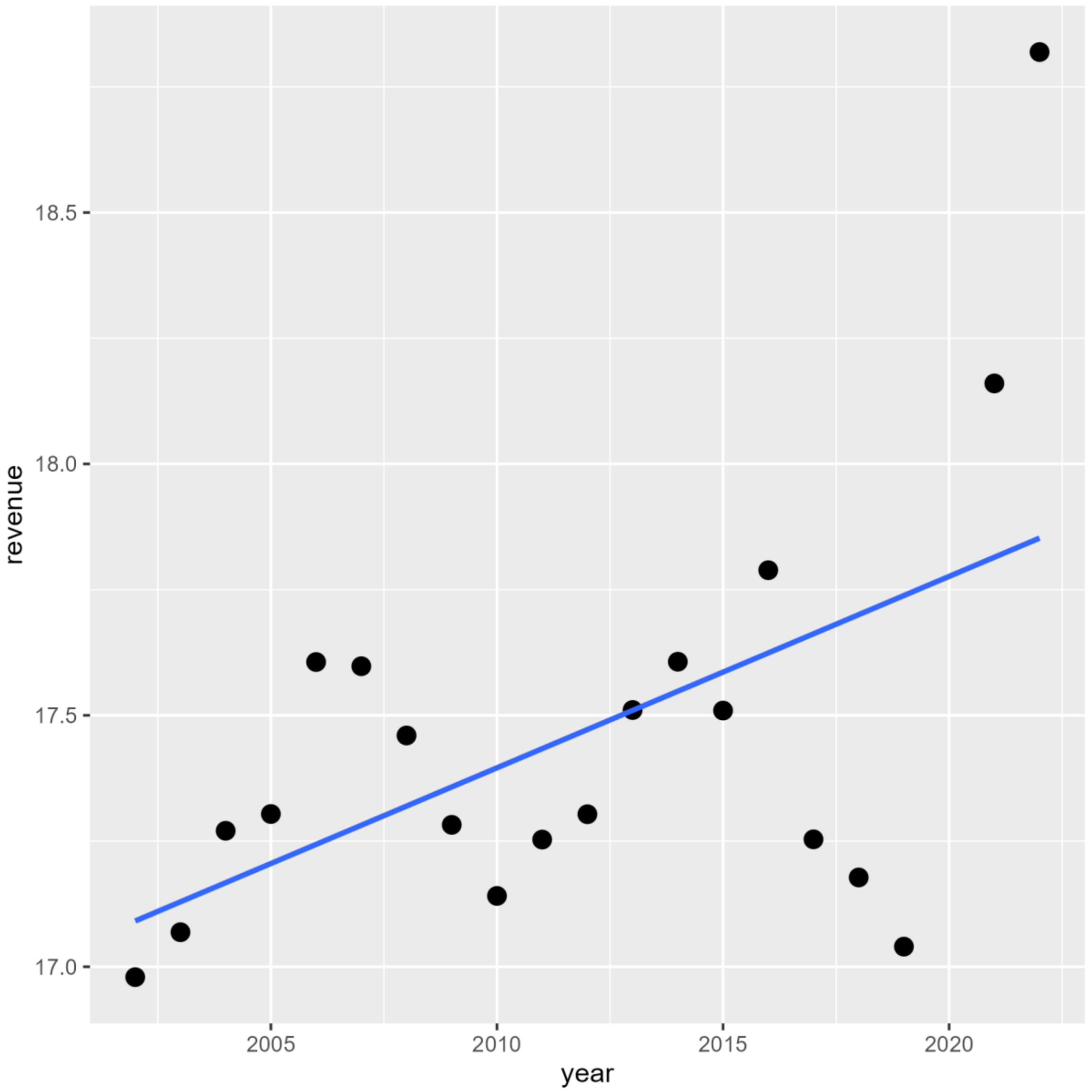

revenue
18.5
18.0
17.5
17.0
2005
2010
2015
2020
year


18.5
18.0
17.5
17.0
revenue
2005
2010
2015
2020
year

$$\widehat{y_i} = \beta_0 + \sum_{j=0}^{p} \beta_j X_{ij} + \epsilon_{ij}$$

$$\epsilon_{ij} \sim N(0, \sigma^2)$$

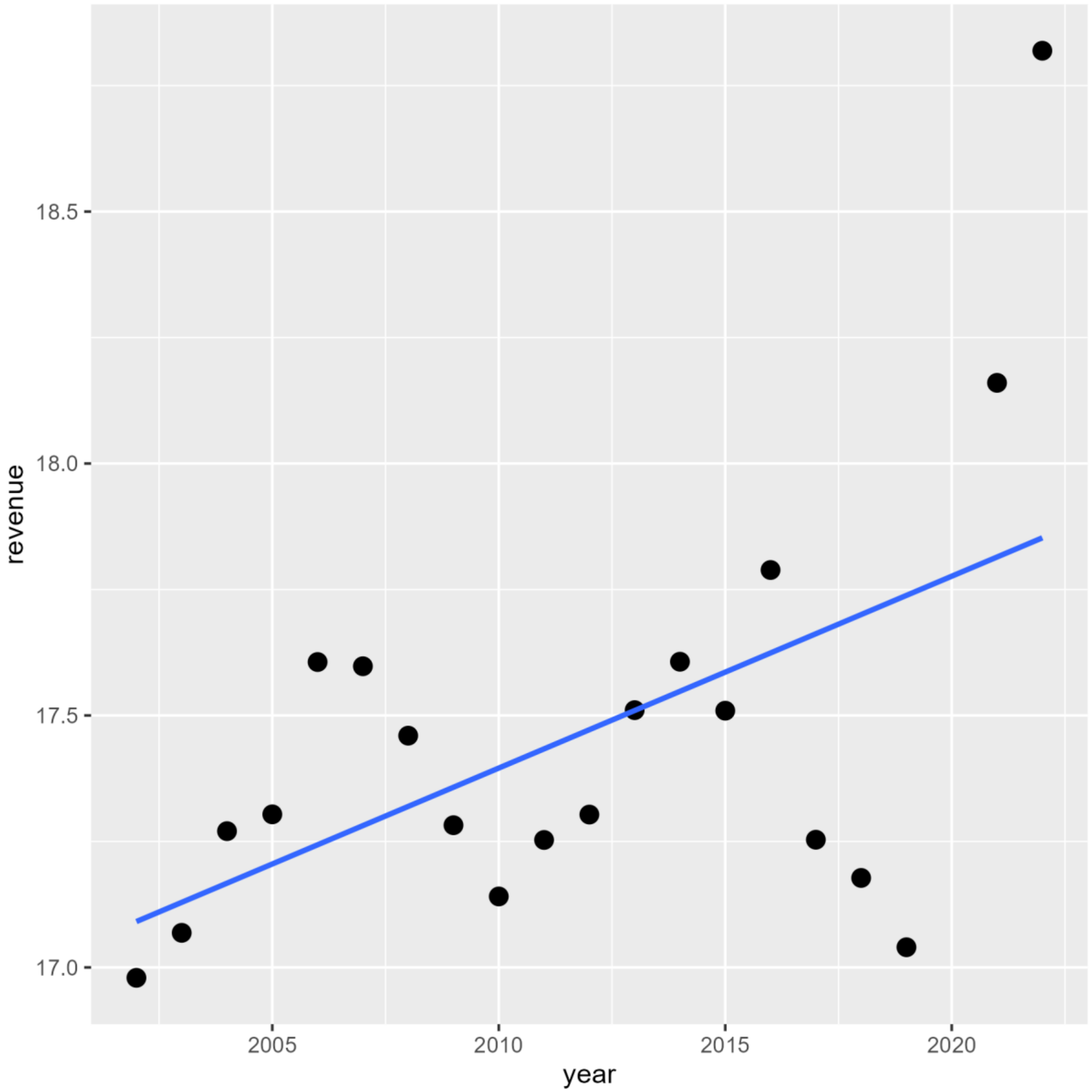
$$\hat{\beta} = \operatorname{argmin} \sum_{j=0}^{p} (y_i - \widehat{y_i})^2$$

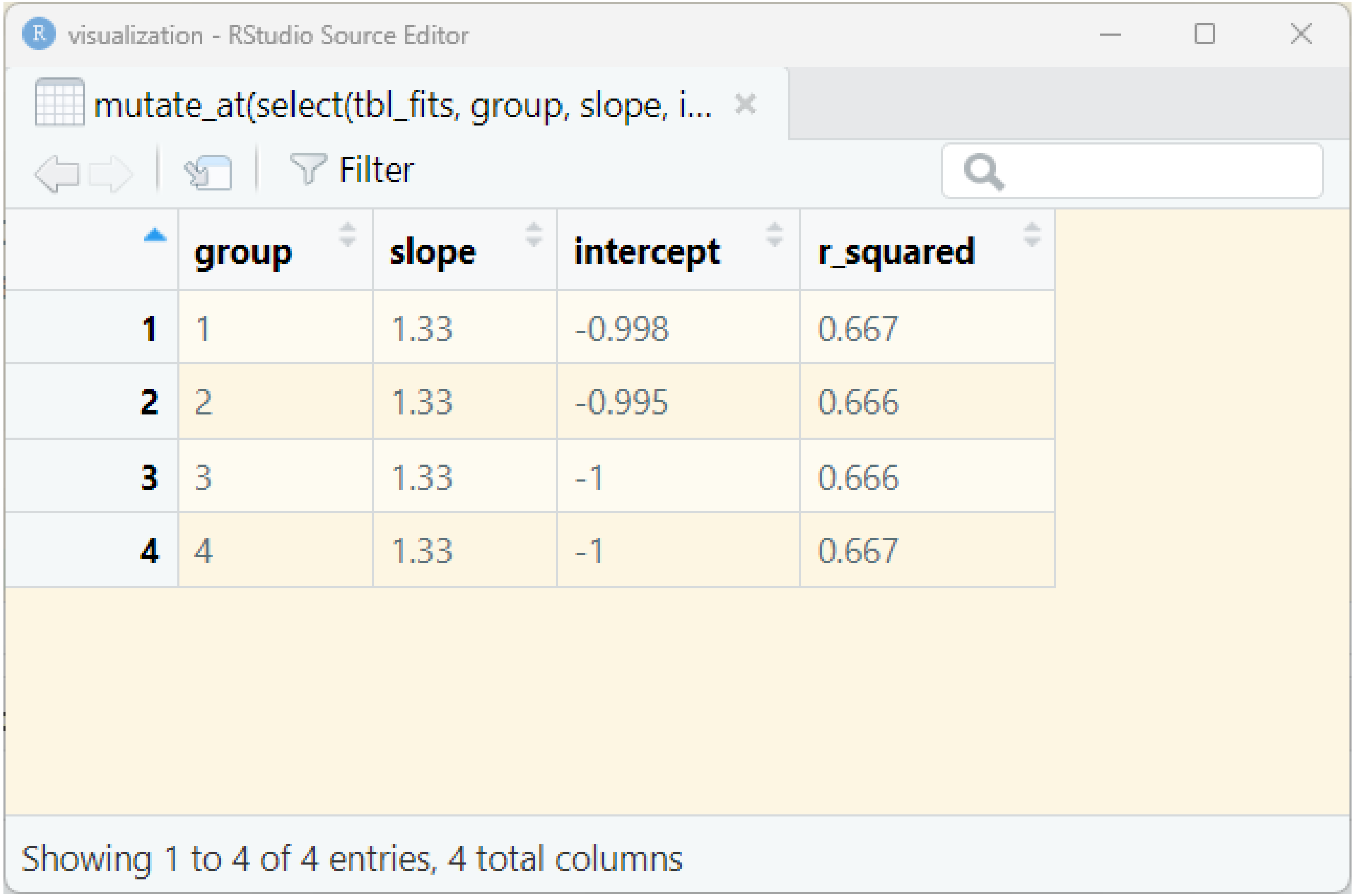


Visualization is a tool which facilitates communication with the less numerate.

In particular, it is an indispensable aid for actuaries who are trying to learn or interpret statistical models.

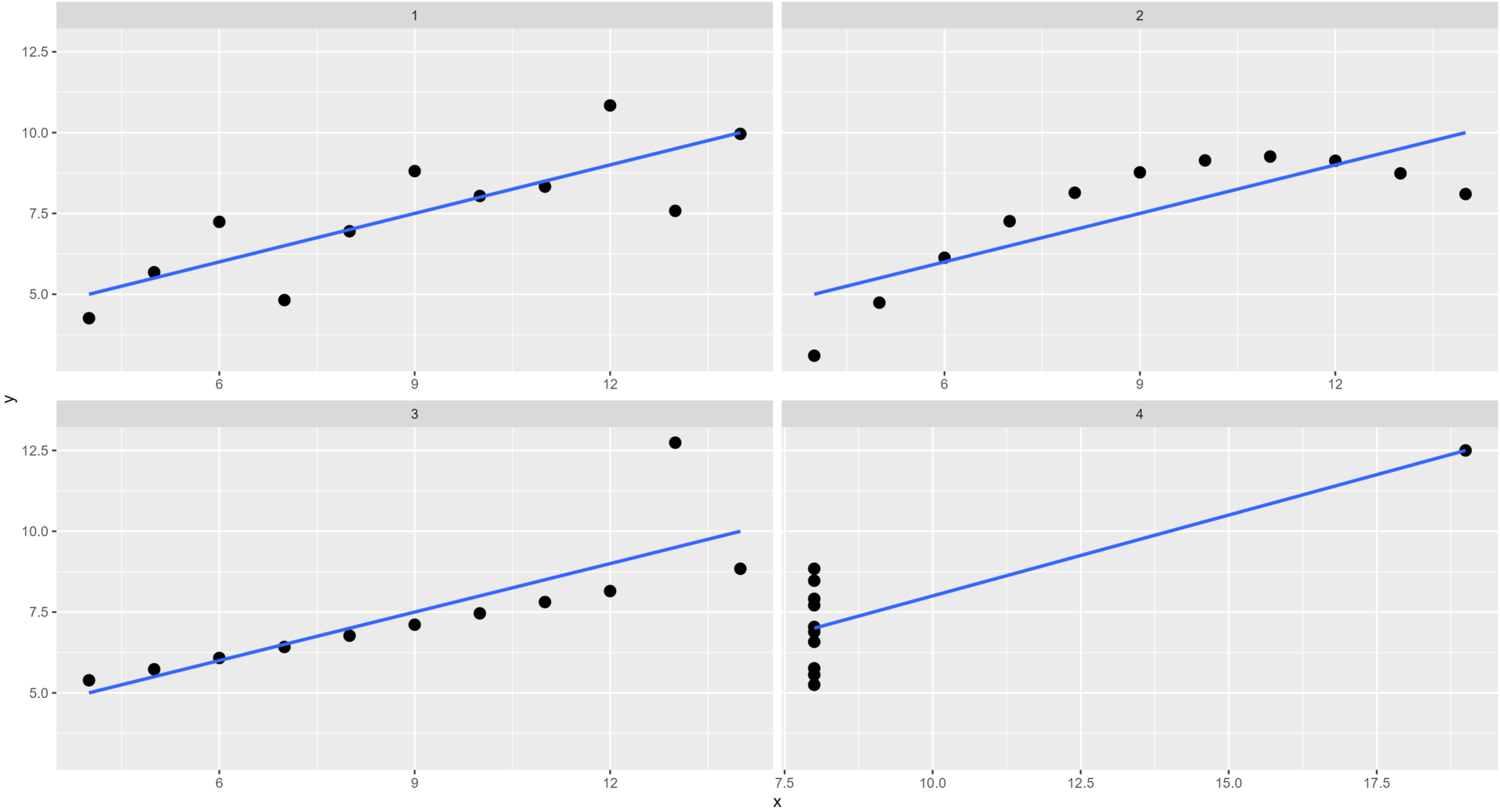
```
tbl_wide |>
    ggplot(aes(year, revenue)) +
    geom_point() +
    geom_smooth(
        method = lm,
        se = FALSE)
```

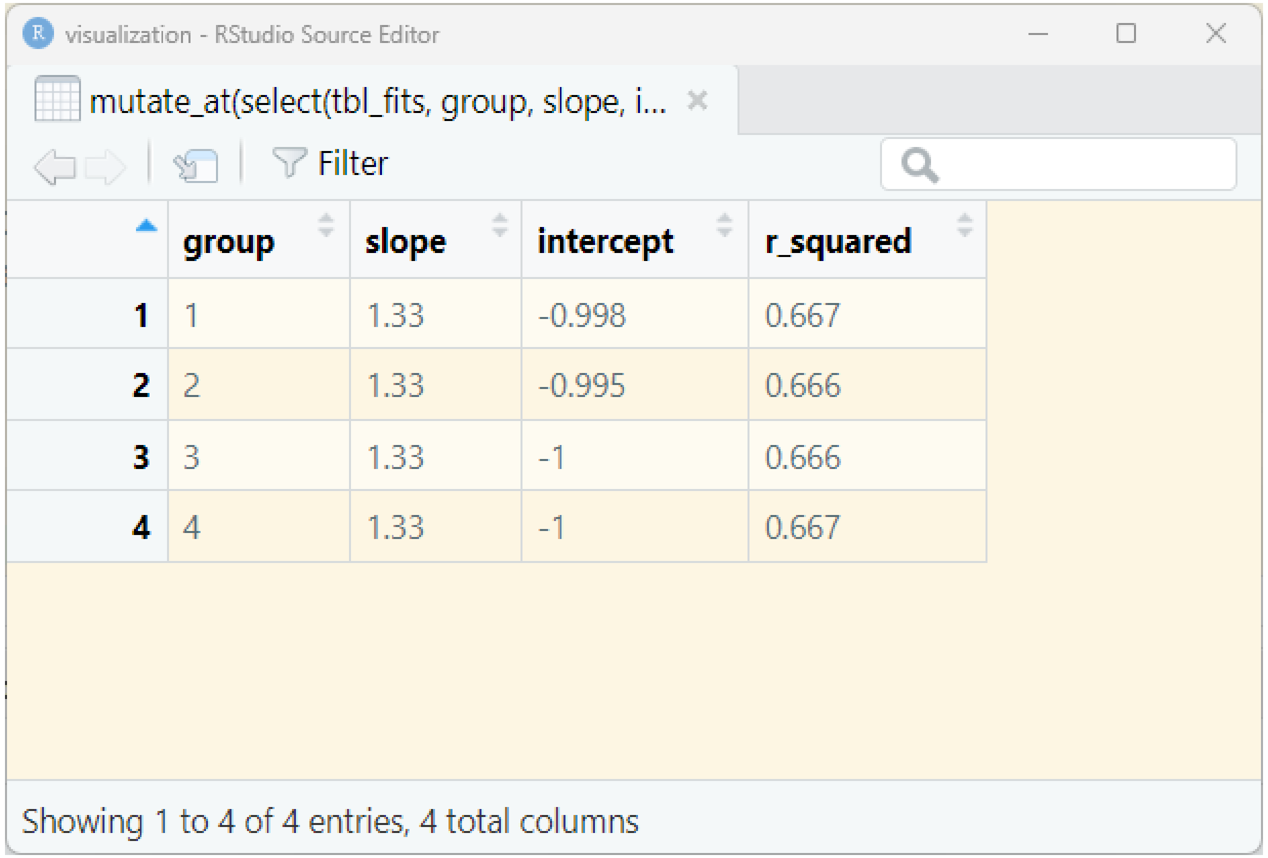



visualization - RStudio Source Editor
mutate_at(select(tbl_fits, group, slope, i...
Filter

	group	slope	intercept	r_squared
1	1	1.33	-0.998	0.667
2	2	1.33	-0.995	0.666
3	3	1.33	-1	0.666
4	4	1.33	-1	0.667

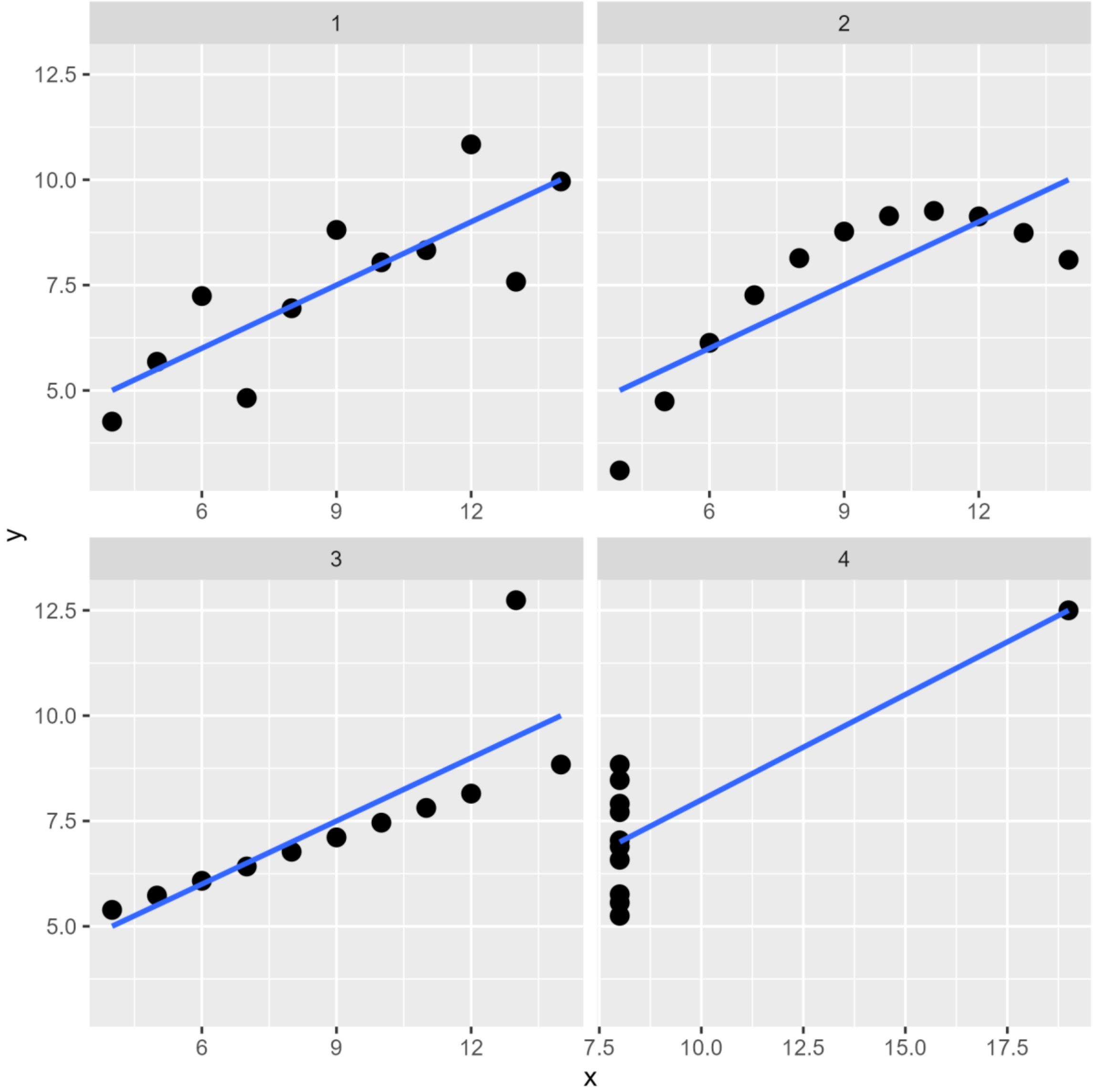
Showing 1 to 4 of 4 entries, 4 total columns


1
2
3
4
12.5
10.0
7.5
5.0
12.5
10.0
7.5
5.0
6
9
12
6
9
12
6
9
12
7.5
10.0
12.5
15.0
17.5
y
x

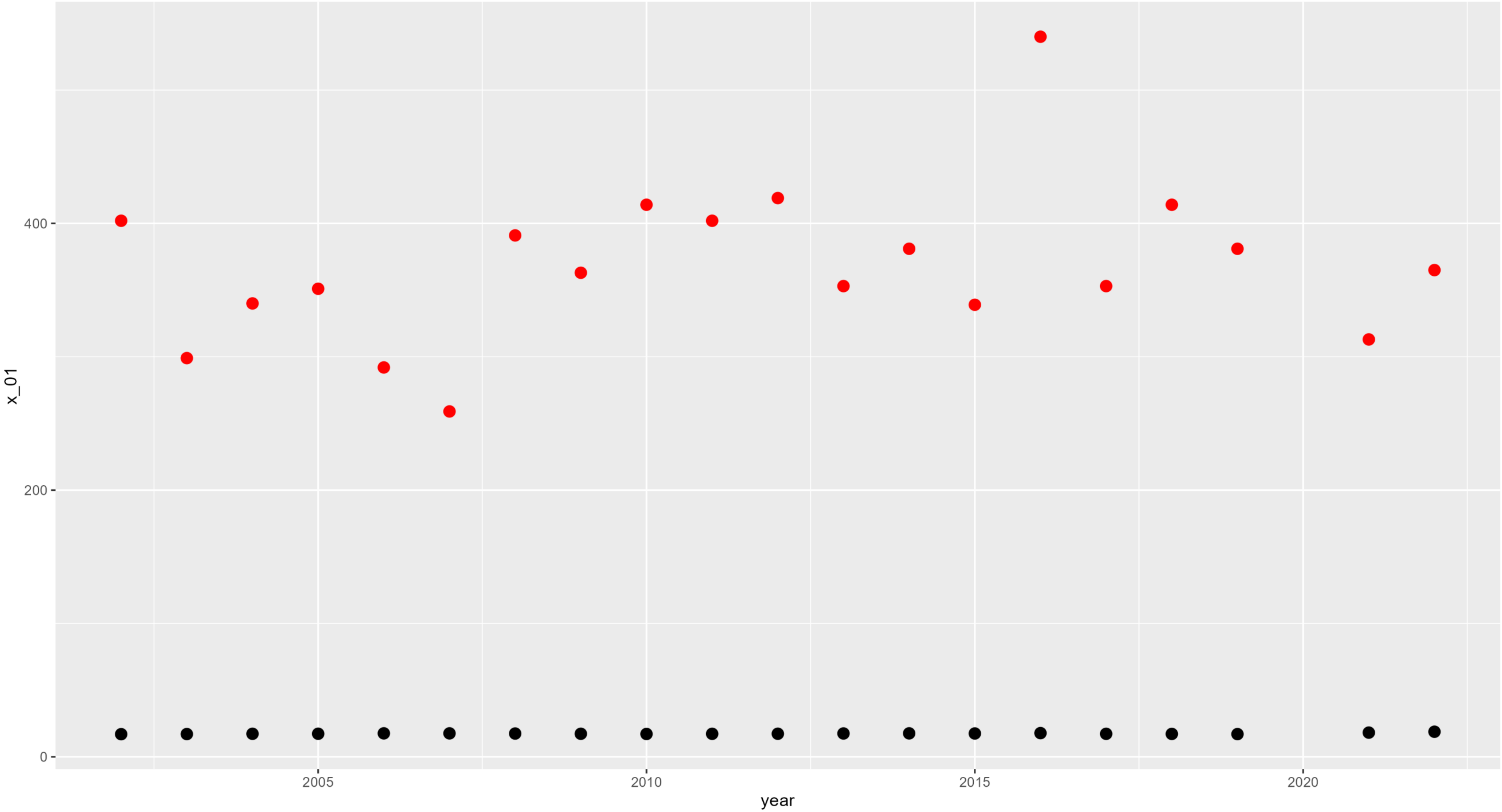

visualization - RStudio Source Editor
mutate_at(select(tbl_fits, group, slope, i...
Filter

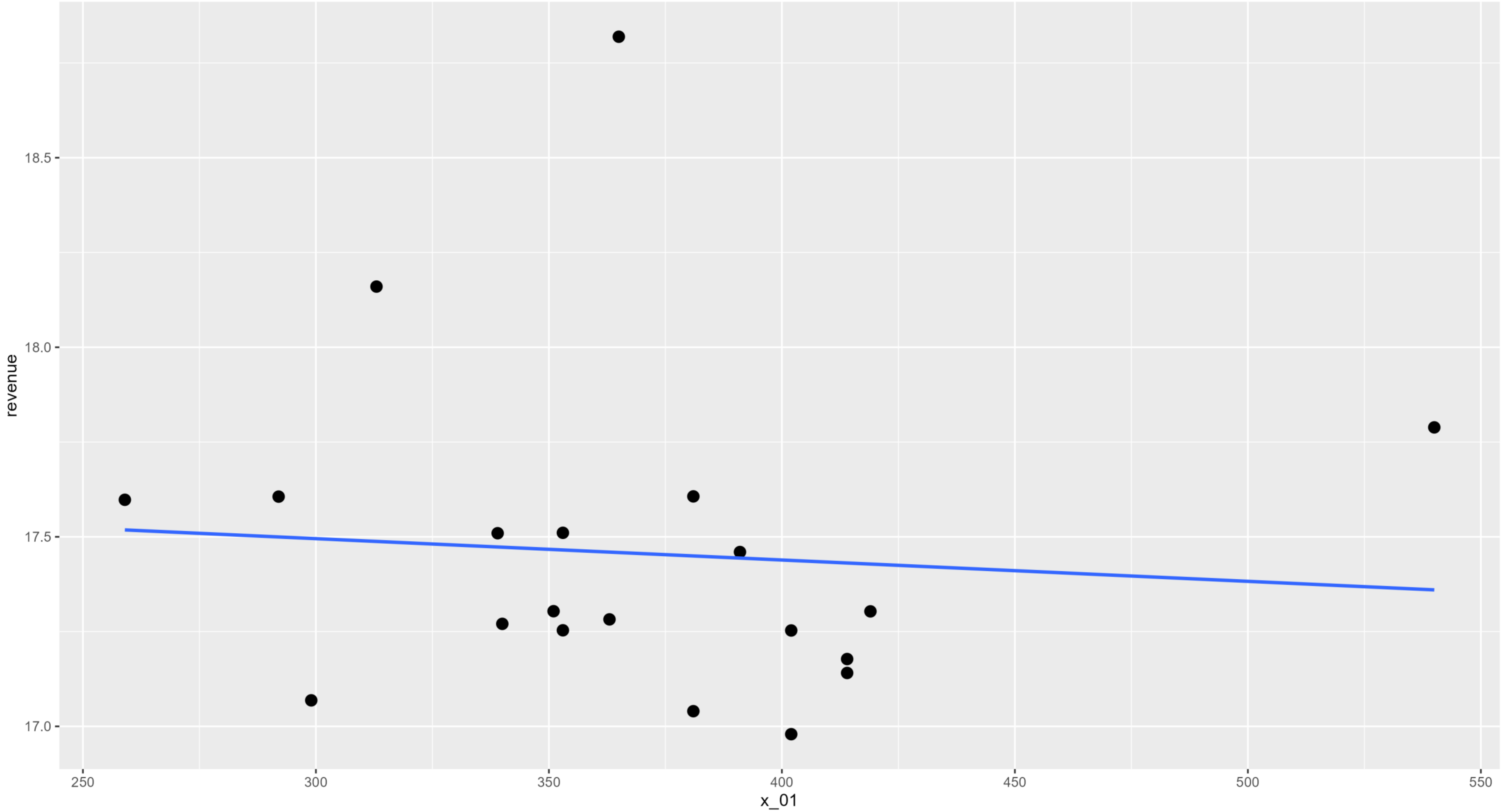
	group	slope	intercept	r_squared
1	1	1.33	-0.998	0.667
2	2	1.33	-0.995	0.666
3	3	1.33	-1	0.666
4	4	1.33	-1	0.667

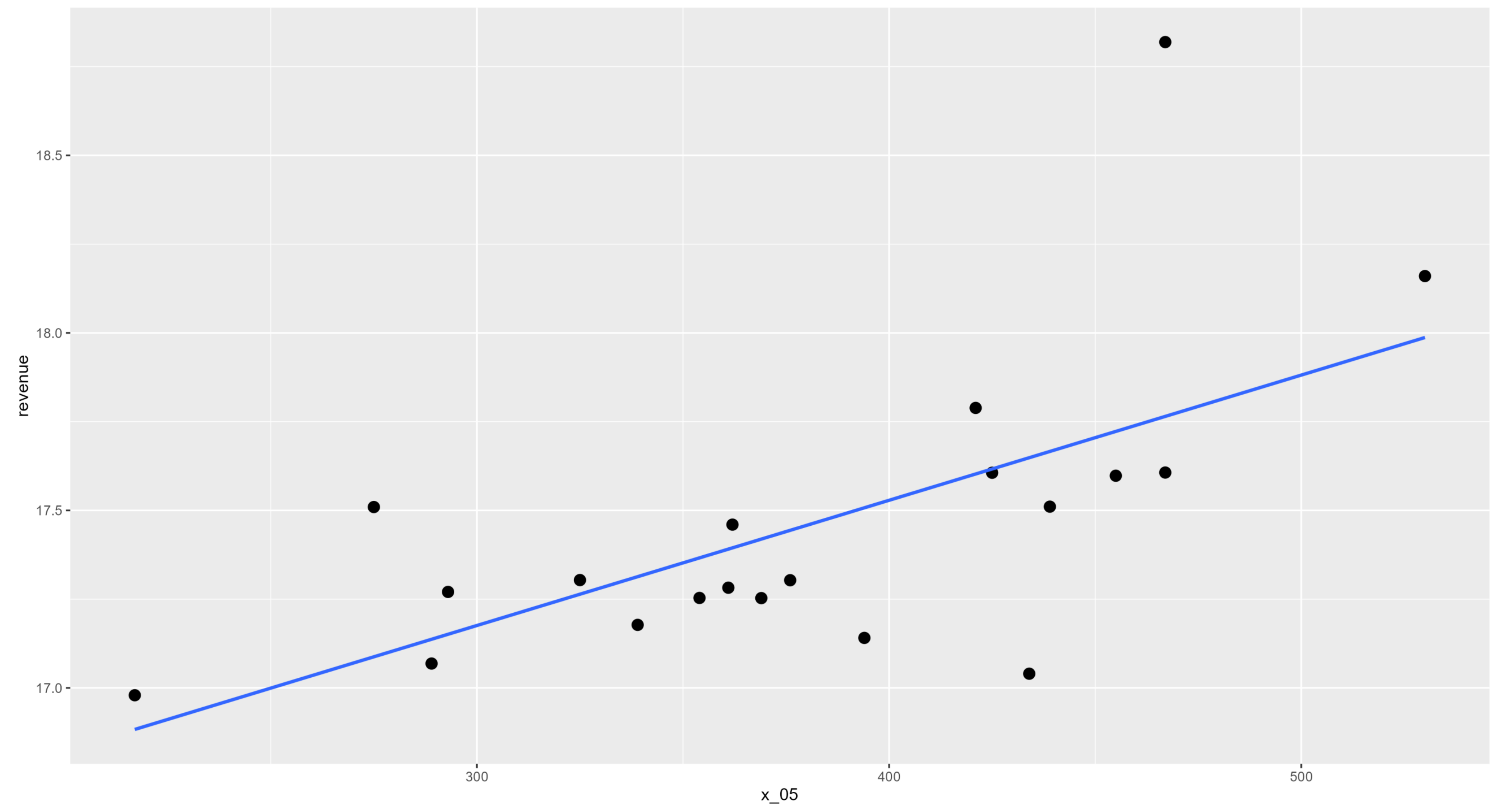
Showing 1 to 4 of 4 entries, 4 total columns

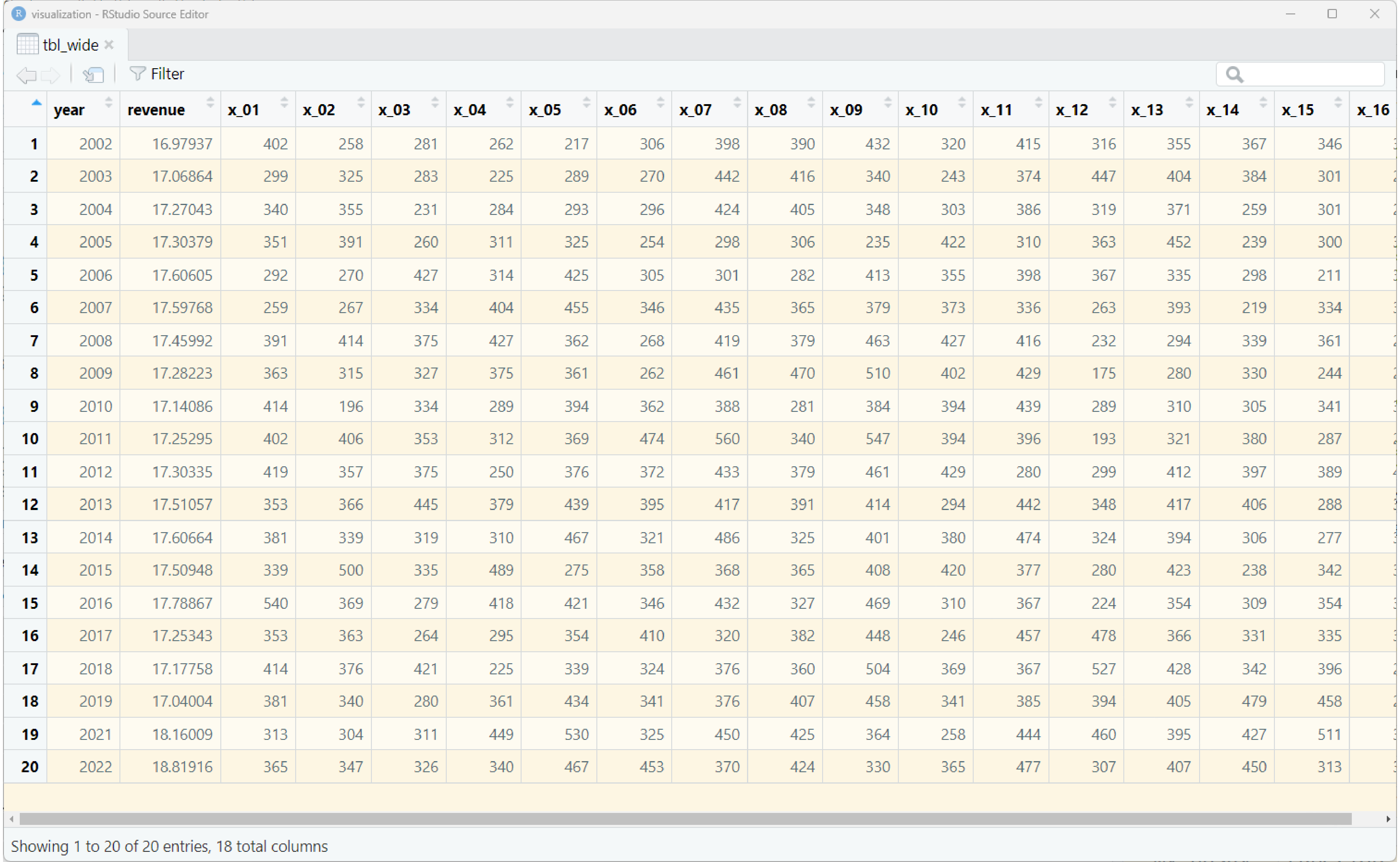


You have been given predictive data by your broker. There are 16 columns to work with.


x_01
400
200
0
2005
2010
2015
2020
year

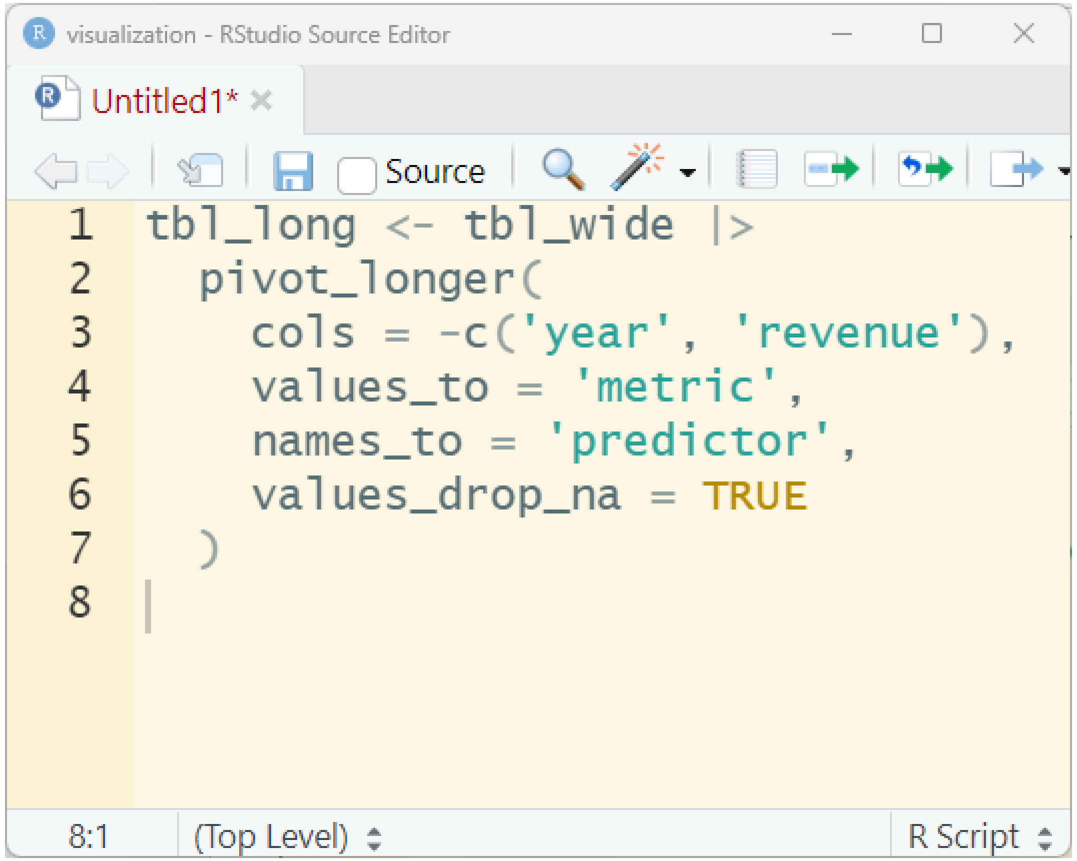

18.5
18.0
17.5
17.0
revenue
250
300
350
400
450
500
550
x_01


18.5
18.0
17.5
17.0
revenue
300
400
500
x_05

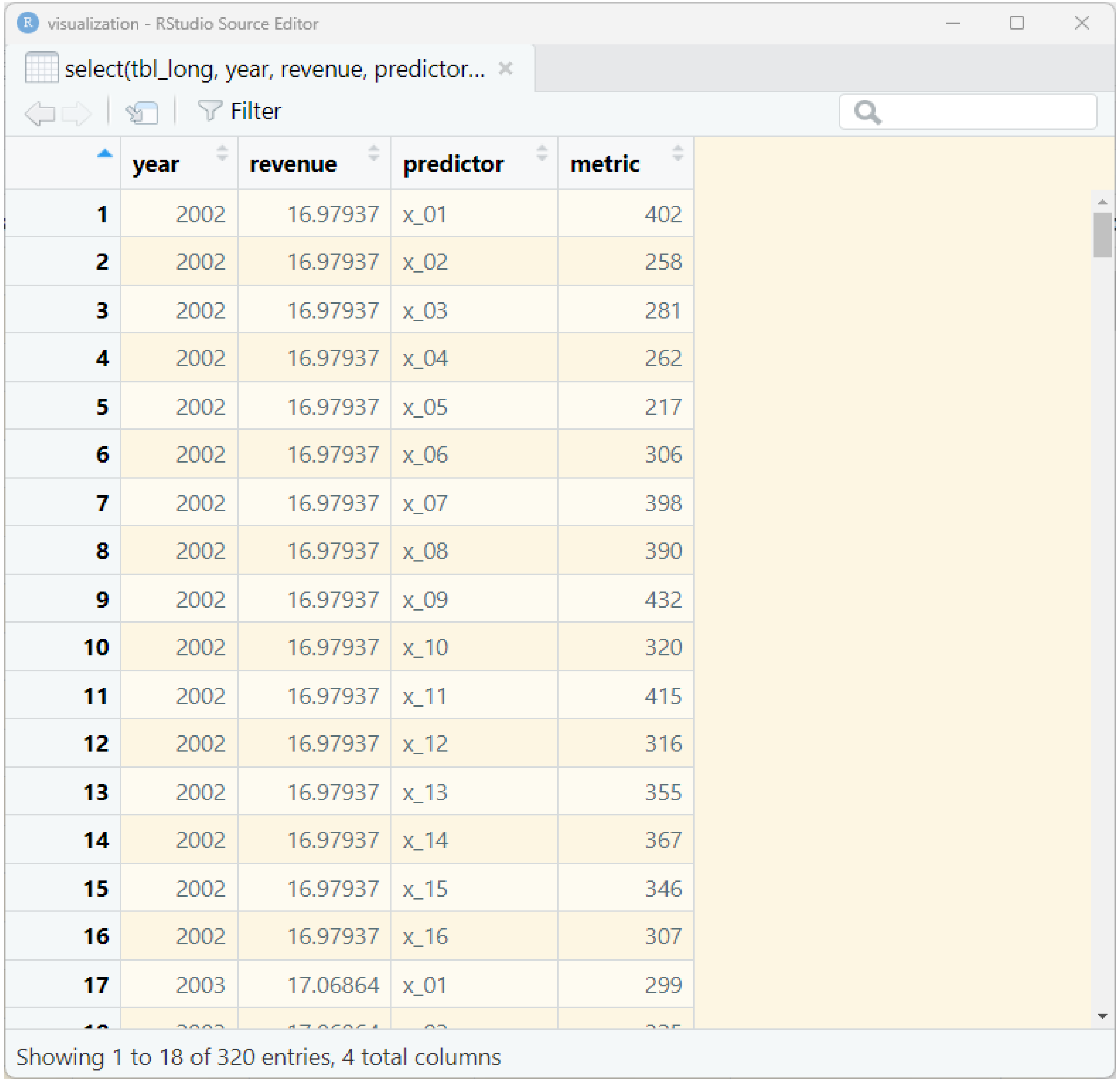


	year	revenue	x_01	x_02	x_03	x_04	x_05	x_06	x_07	x_08	x_09	x_10	x_11	x_12	x_13	x_14	x_15	x_16
1	2002	16.97937	402	258	281	262	217	306	398	390	432	320	415	316	355	367	346	
2	2003	17.06864	299	325	283	225	289	270	442	416	340	243	374	447	404	384	301	
3	2004	17.27043	340	355	231	284	293	296	424	405	348	303	386	319	371	259	301	
4	2005	17.30379	351	391	260	311	325	254	298	306	235	422	310	363	452	239	300	
5	2006	17.60605	292	270	427	314	425	305	301	282	413	355	398	367	335	298	211	
6	2007	17.59768	259	267	334	404	455	346	435	365	379	373	336	263	393	219	334	
7	2008	17.45992	391	414	375	427	362	268	419	379	463	427	416	232	294	339	361	
8	2009	17.28223	363	315	327	375	361	262	461	470	510	402	429	175	280	330	244	
9	2010	17.14086	414	196	334	289	394	362	388	281	384	394	439	289	310	305	341	
10	2011	17.25295	402	406	353	312	369	474	560	340	547	394	396	193	321	380	287	
11	2012	17.30335	419	357	375	250	376	372	433	379	461	429	280	299	412	397	389	
12	2013	17.51057	353	366	445	379	439	395	417	391	414	294	442	348	417	406	288	
13	2014	17.60664	381	339	319	310	467	321	486	325	401	380	474	324	394	306	277	
14	2015	17.50948	339	500	335	489	275	358	368	365	408	420	377	280	423	238	342	
15	2016	17.78867	540	369	279	418	421	346	432	327	469	310	367	224	354	309	354	
16	2017	17.25343	353	363	264	295	354	410	320	382	448	246	457	478	366	331	335	
17	2018	17.17758	414	376	421	225	339	324	376	360	504	369	367	527	428	342	396	
18	2019	17.04004	381	340	280	361	434	341	376	407	458	341	385	394	405	479	458	
19	2021	18.16009	313	304	311	449	530	325	450	425	364	258	444	460	395	427	511	
20	2022	18.81916	365	347	326	340	467	453	370	424	330	365	477	307	407	450	313	

Showing 1 to 20 of 20 entries, 18 total columns

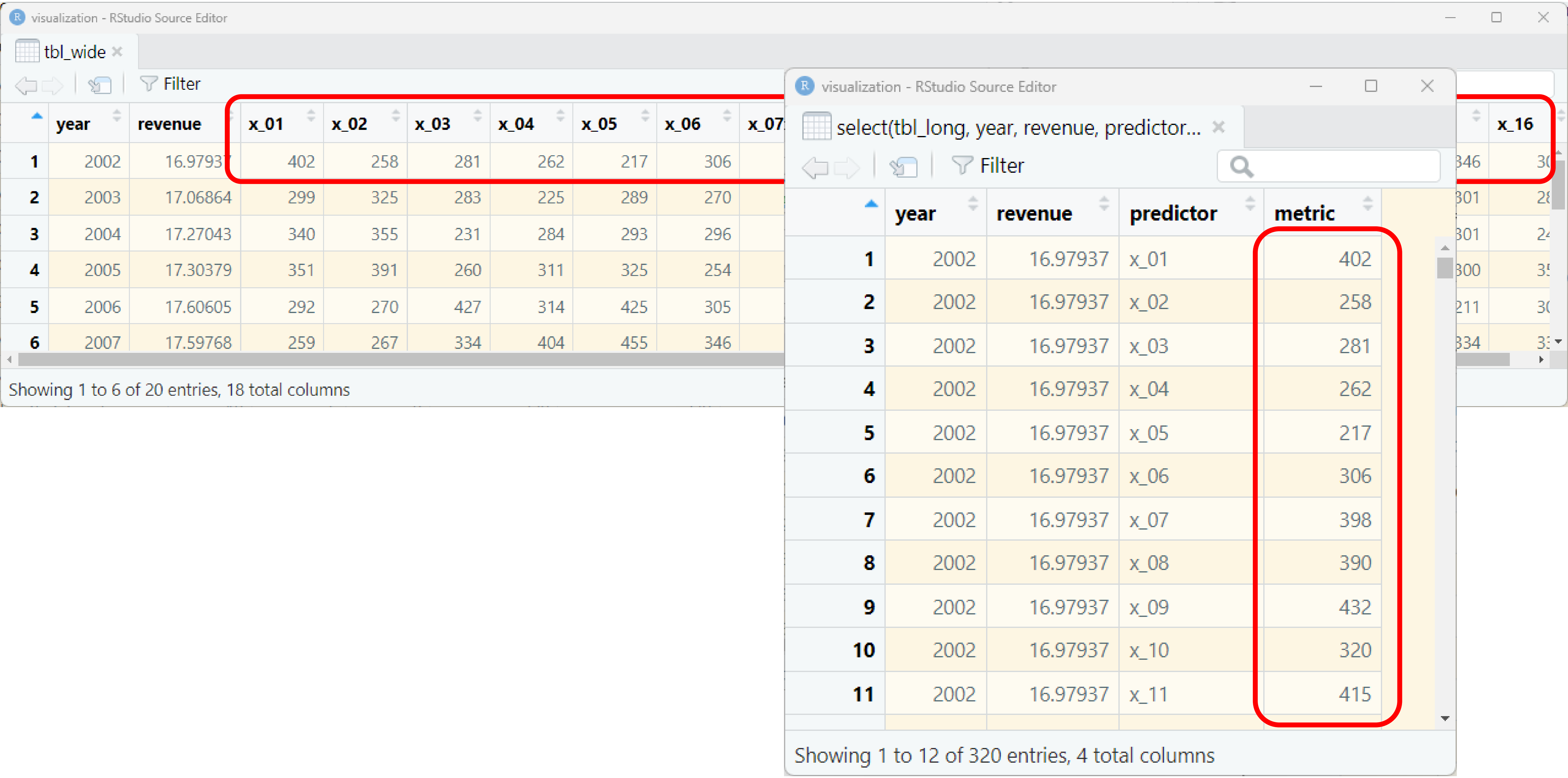


```r
tbl_long <- tbl_wide |>
  pivot_longer(
    cols = -c('year', 'revenue'),
    values_to = 'metric',
    names_to = 'predictor',
    values_drop_na = TRUE
  )

```


visualization - RStudio Source Editor
select(tbl_long, year, revenue, predictor...
Filter

	year	revenue	predictor	metric
1	2002	16.97937	x_01	402
2	2002	16.97937	x_02	258
3	2002	16.97937	x_03	281
4	2002	16.97937	x_04	262
5	2002	16.97937	x_05	217
6	2002	16.97937	x_06	306
7	2002	16.97937	x_07	398
8	2002	16.97937	x_08	390
9	2002	16.97937	x_09	432
10	2002	16.97937	x_10	320
11	2002	16.97937	x_11	415
12	2002	16.97937	x_12	316
13	2002	16.97937	x_13	355
14	2002	16.97937	x_14	367
15	2002	16.97937	x_15	346
16	2002	16.97937	x_16	307
17	2003	17.06864	x_01	299

Showing 1 to 18 of 320 entries, 4 total columns



visualization - RStudio Source Editor
tbl_wide
Filter

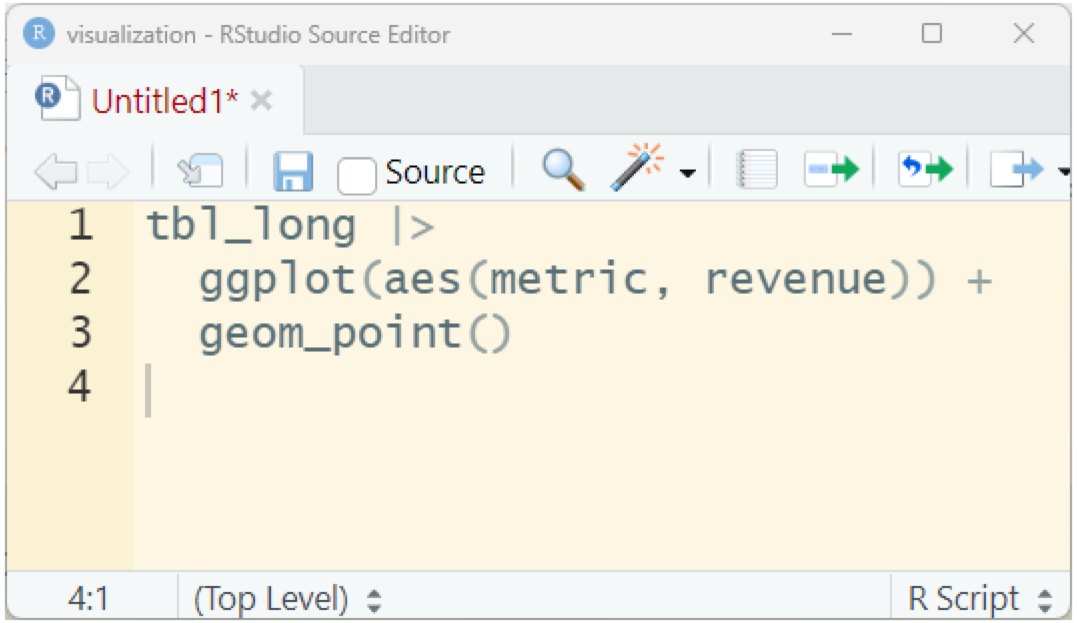
	year	revenue	x_01	x_02	x_03	x_04	x_05	x_06	x_07		x_16
1	2002	16.97937	402	258	281	262	217	306		346	30
2	2003	17.06864	299	325	283	225	289	270		301	28
3	2004	17.27043	340	355	231	284	293	296		301	24
4	2005	17.30379	351	391	260	311	325	254		300	35
5	2006	17.60605	292	270	427	314	425	305		211	30
6	2007	17.59768	259	267	334	404	455	346		334	33

Showing 1 to 6 of 20 entries, 18 total columns

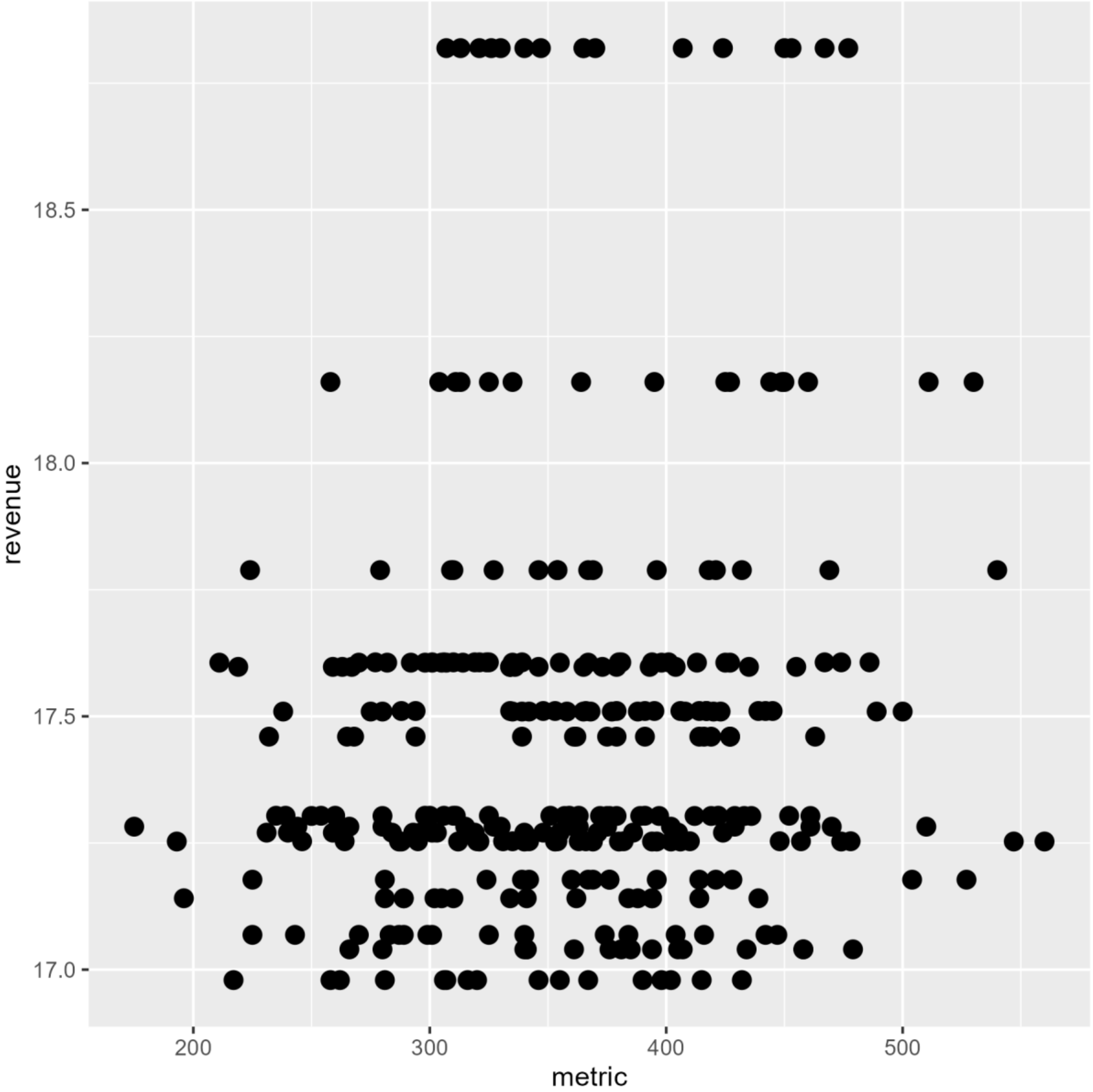
visualization - RStudio Source Editor
select(tbl_long, year, revenue, predictor...
Filter

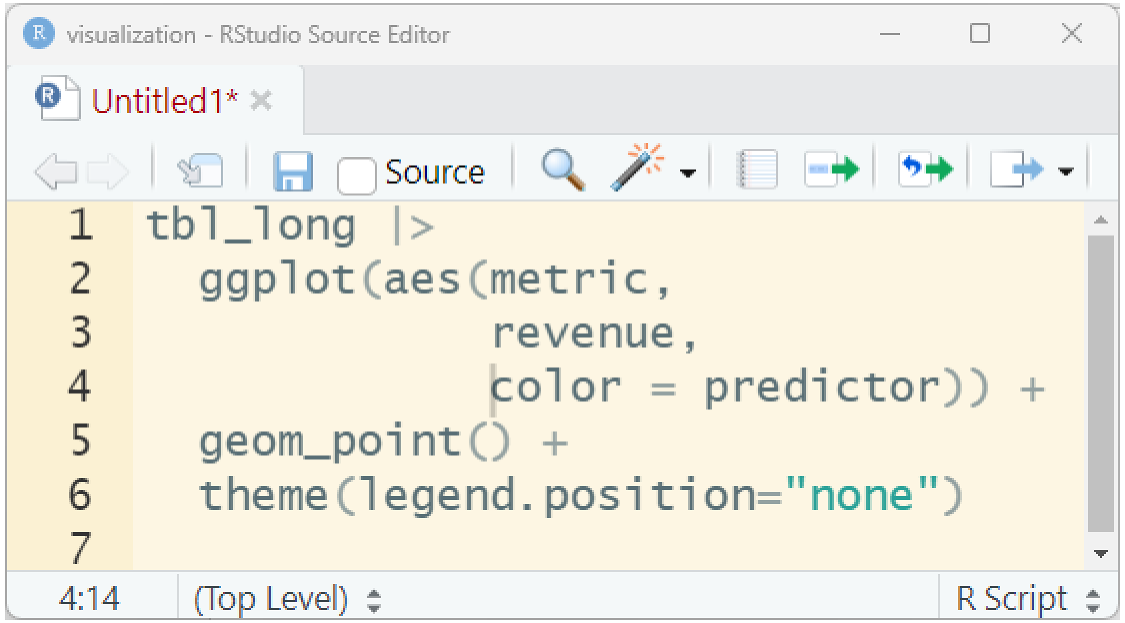
	year	revenue	predictor	metric
1	2002	16.97937	x_01	402
2	2002	16.97937	x_02	258
3	2002	16.97937	x_03	281
4	2002	16.97937	x_04	262
5	2002	16.97937	x_05	217
6	2002	16.97937	x_06	306
7	2002	16.97937	x_07	398
8	2002	16.97937	x_08	390
9	2002	16.97937	x_09	432
10	2002	16.97937	x_10	320
11	2002	16.97937	x_11	415

Showing 1 to 12 of 320 entries, 4 total columns

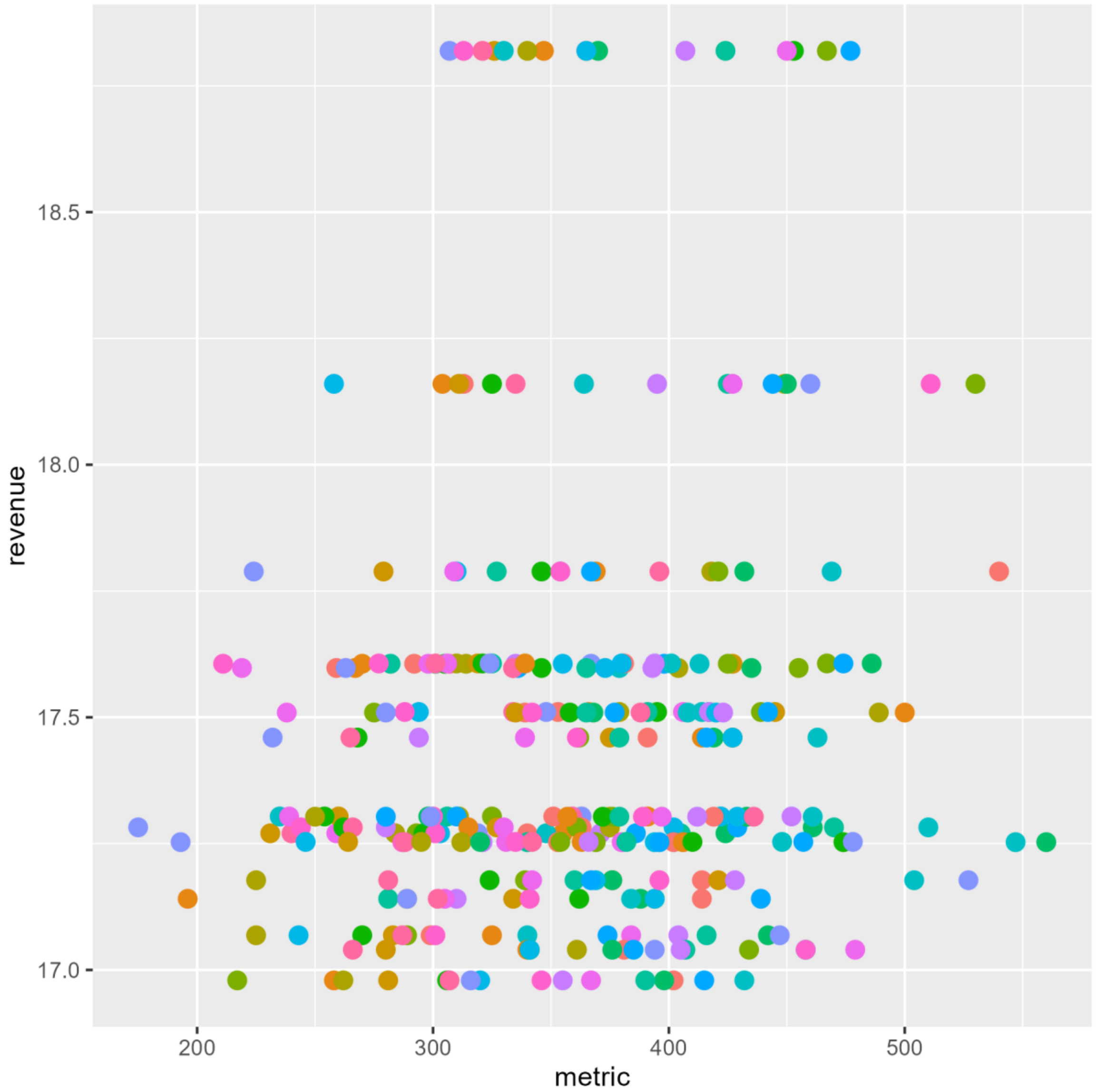


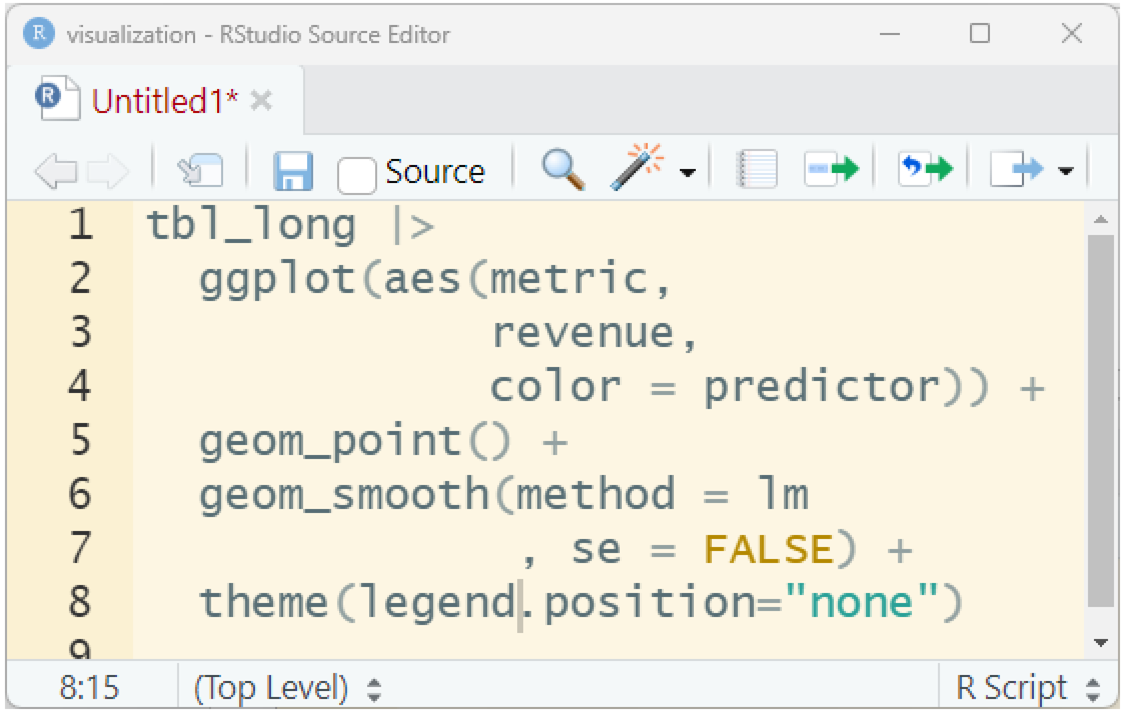
```
tbl_long |>
  ggplot(aes(metric, revenue)) +
  geom_point()

```



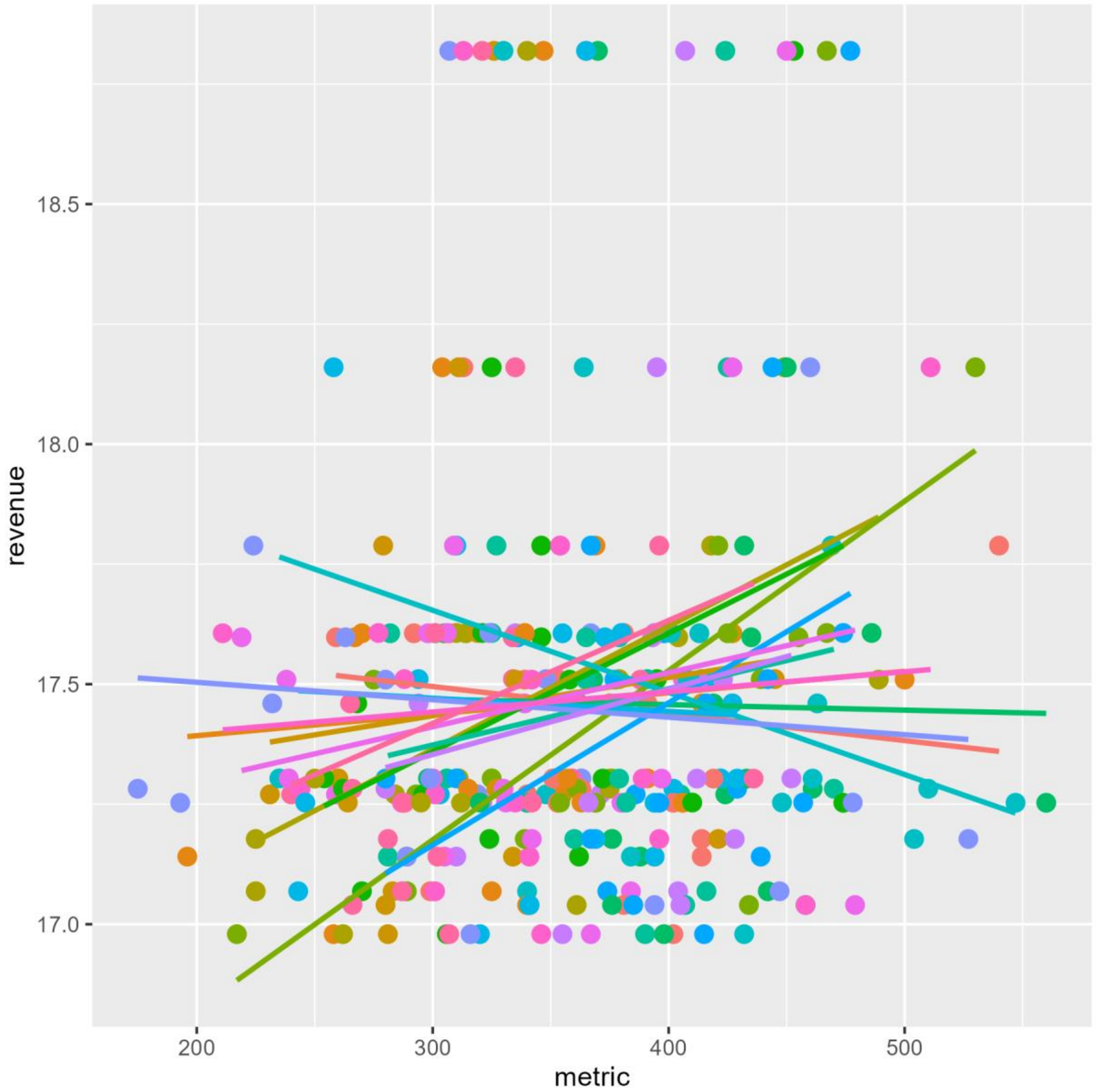


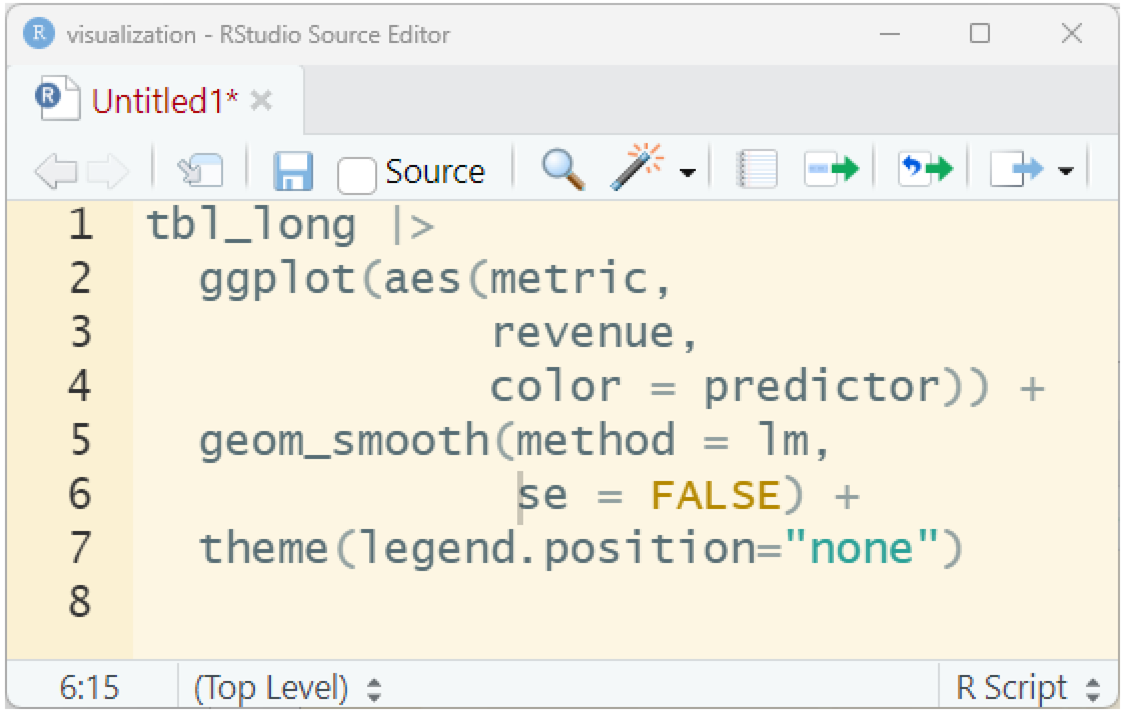
```
tbl_long |>
  ggplot(aes(metric,
             revenue,
             color = predictor)) +
  geom_point() +
  theme(legend.position="none")
```



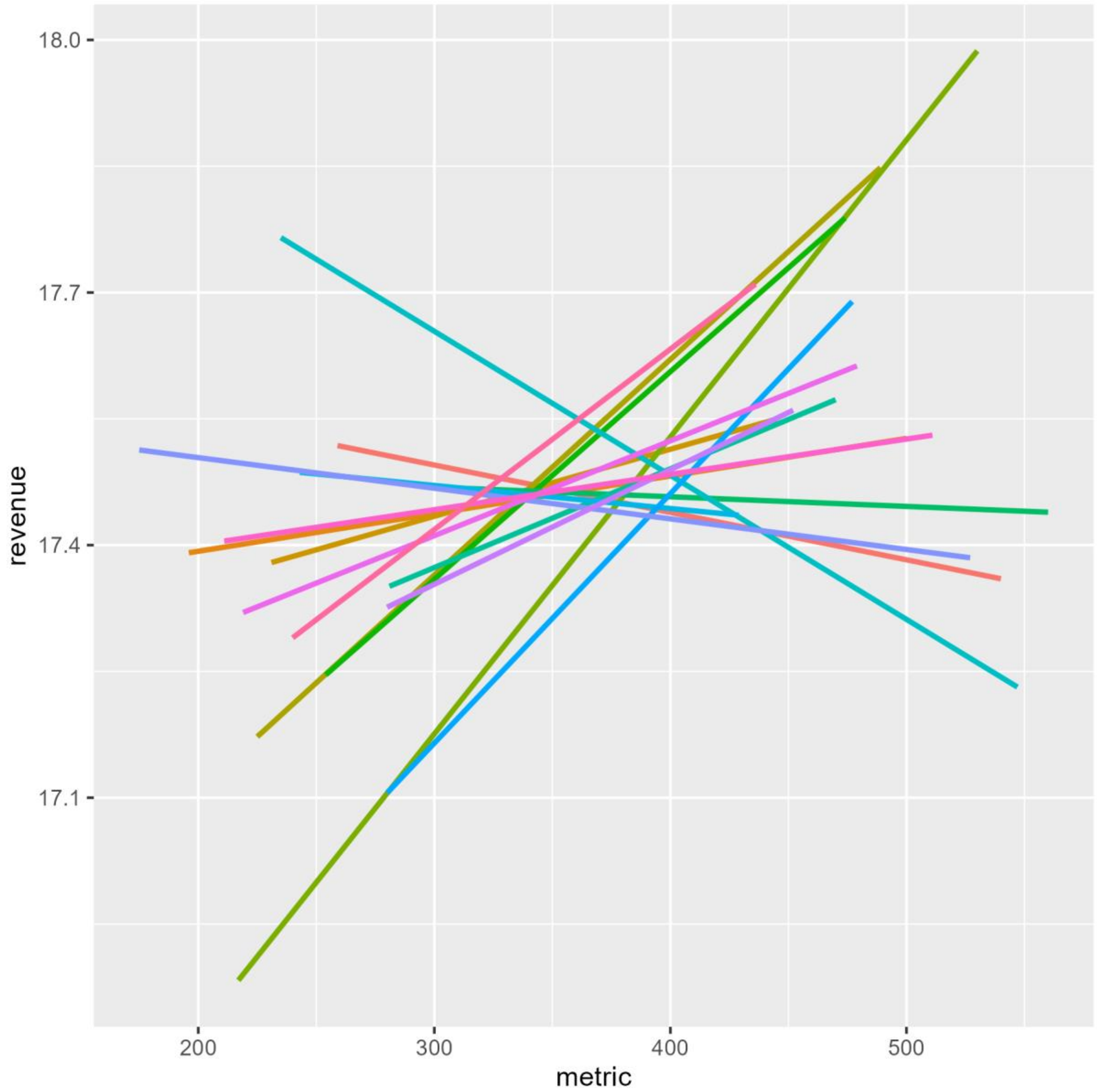


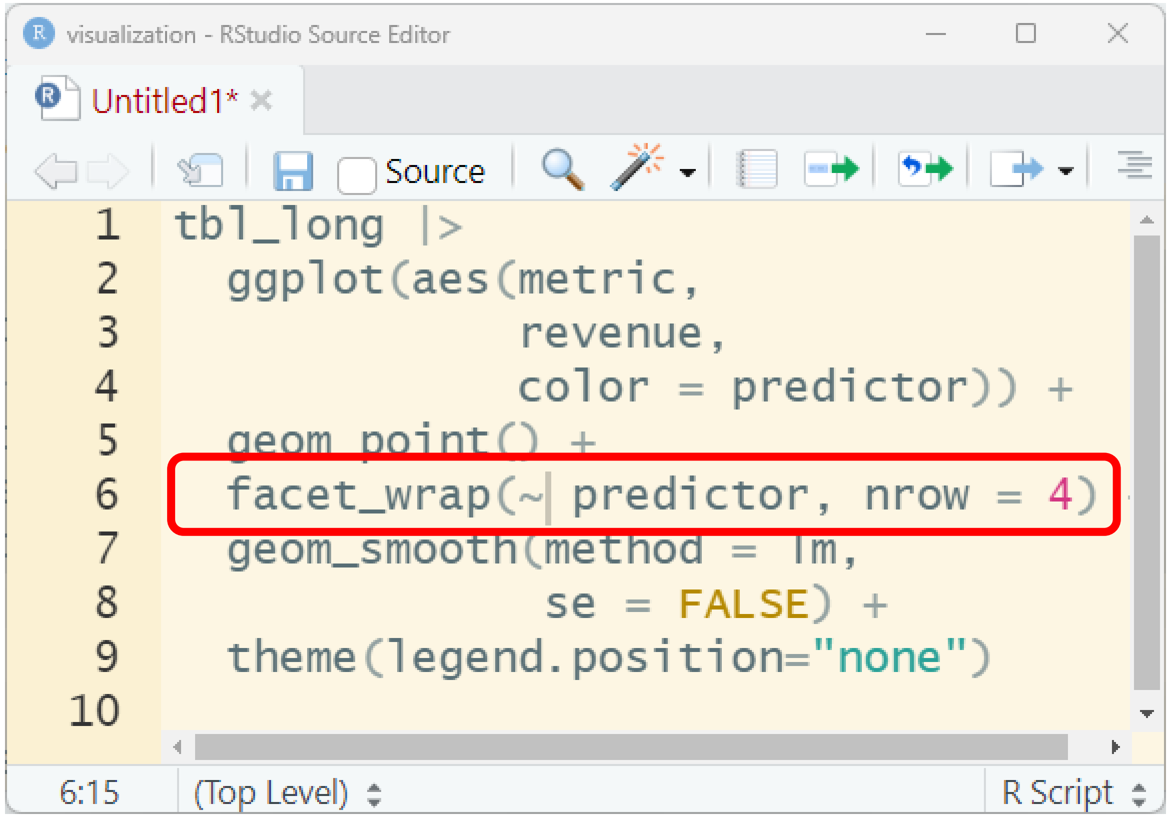
```r
tbl_long |>
  ggplot(aes(metric,
             revenue,
             color = predictor)) +
  geom_point() +
  geom_smooth(method = lm
              , se = FALSE) +
  theme(legend.position="none")
```



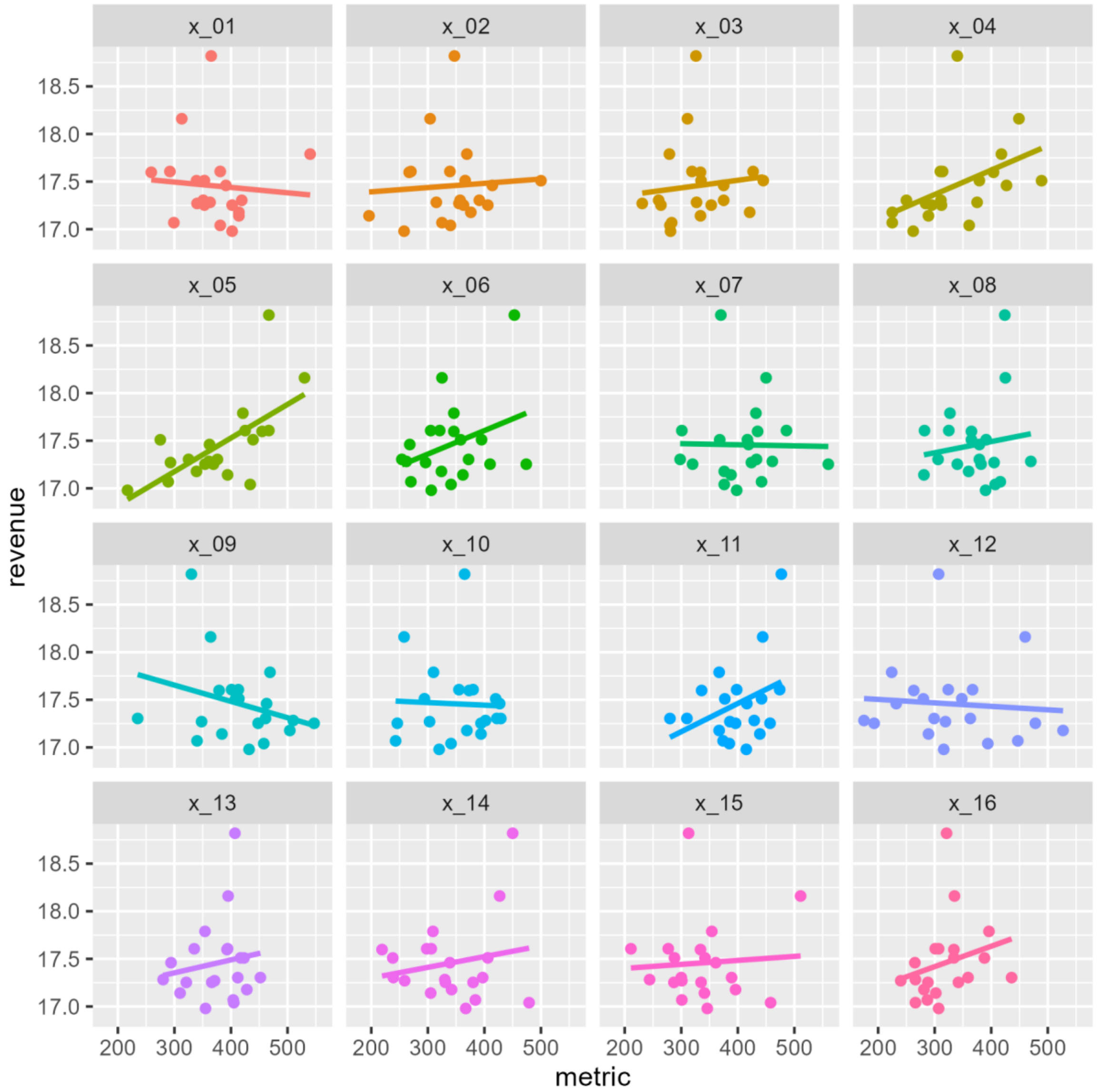


```
tbl_long |>
  ggplot(aes(metric,
             revenue,
             color = predictor)) +
  geom_smooth(method = lm,
              se = FALSE) +
  theme(legend.position="none")
```

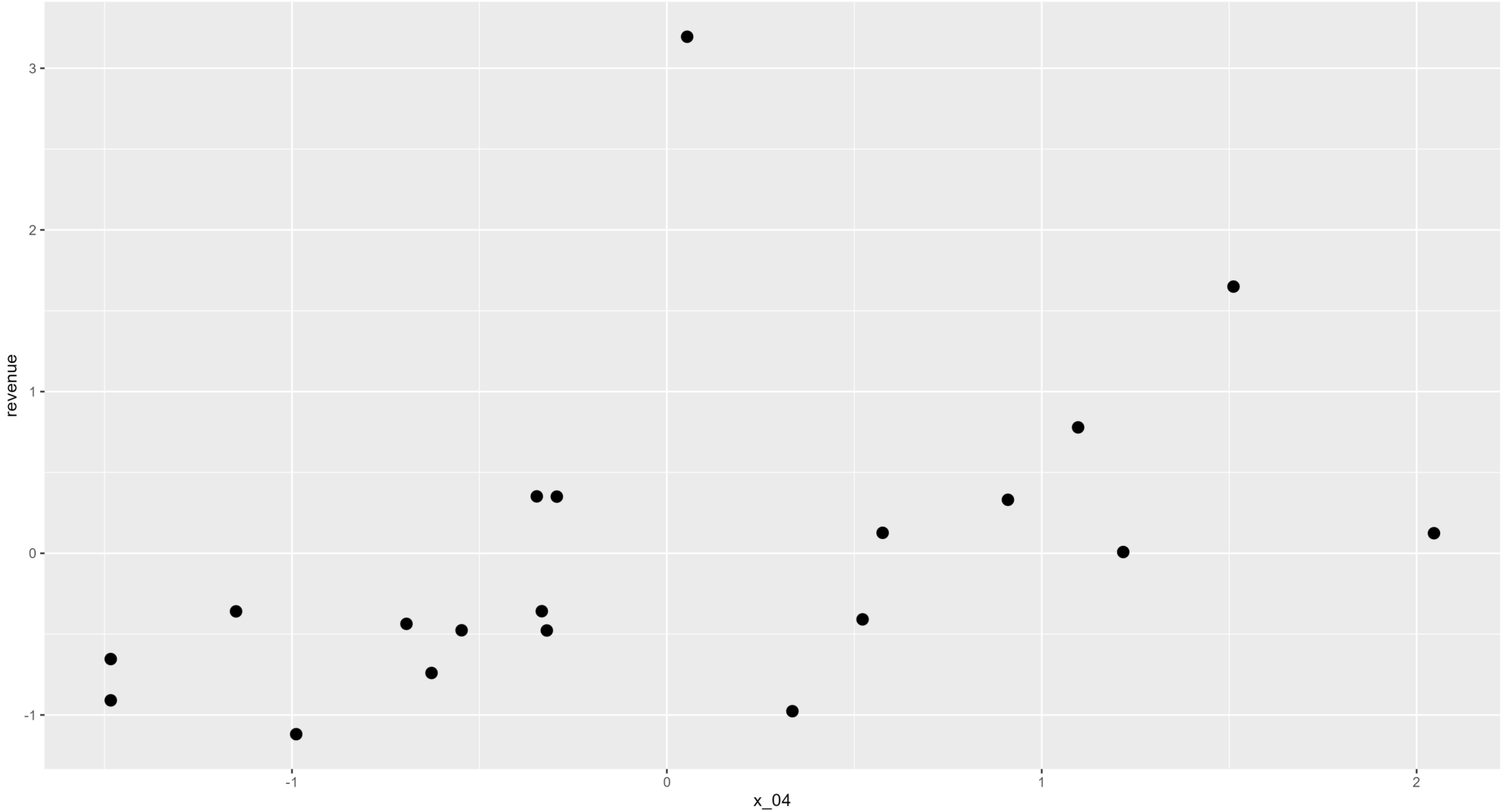


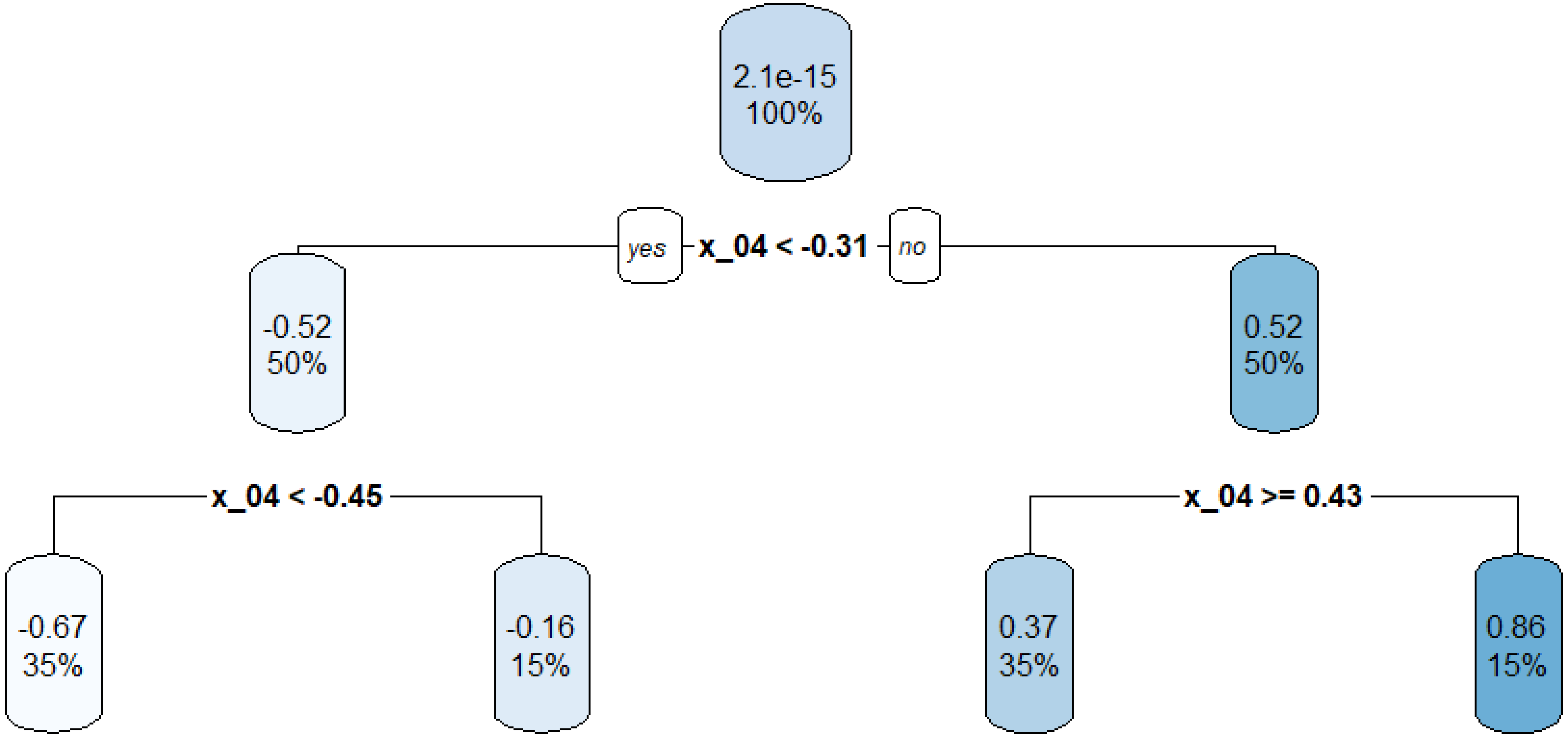


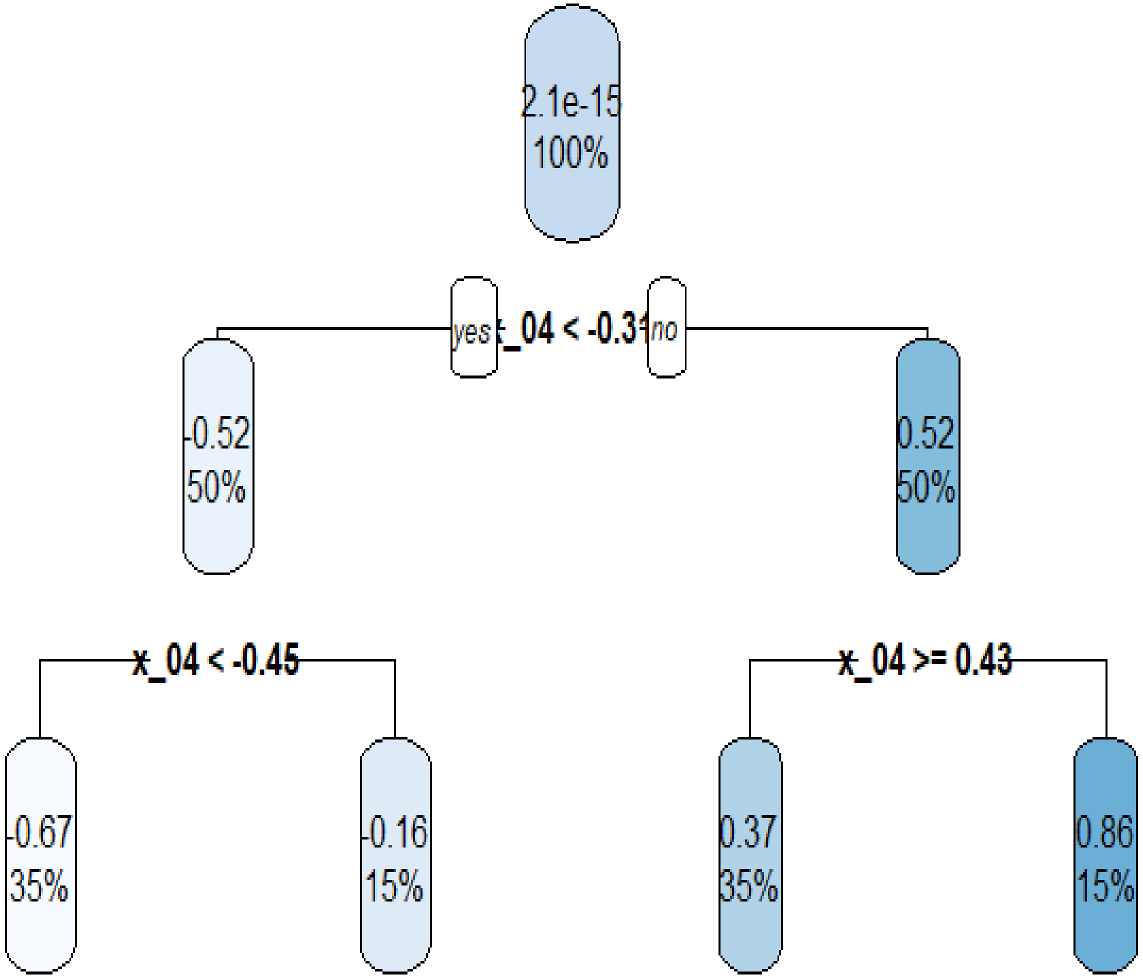
```
tbl_long |>
  ggplot(aes(metric,
             revenue,
             color = predictor)) +
  geom_point() +
  facet_wrap(~ predictor, nrow = 4)
  geom_smooth(method = lm,
              se = FALSE) +
  theme(legend.position="none")
```

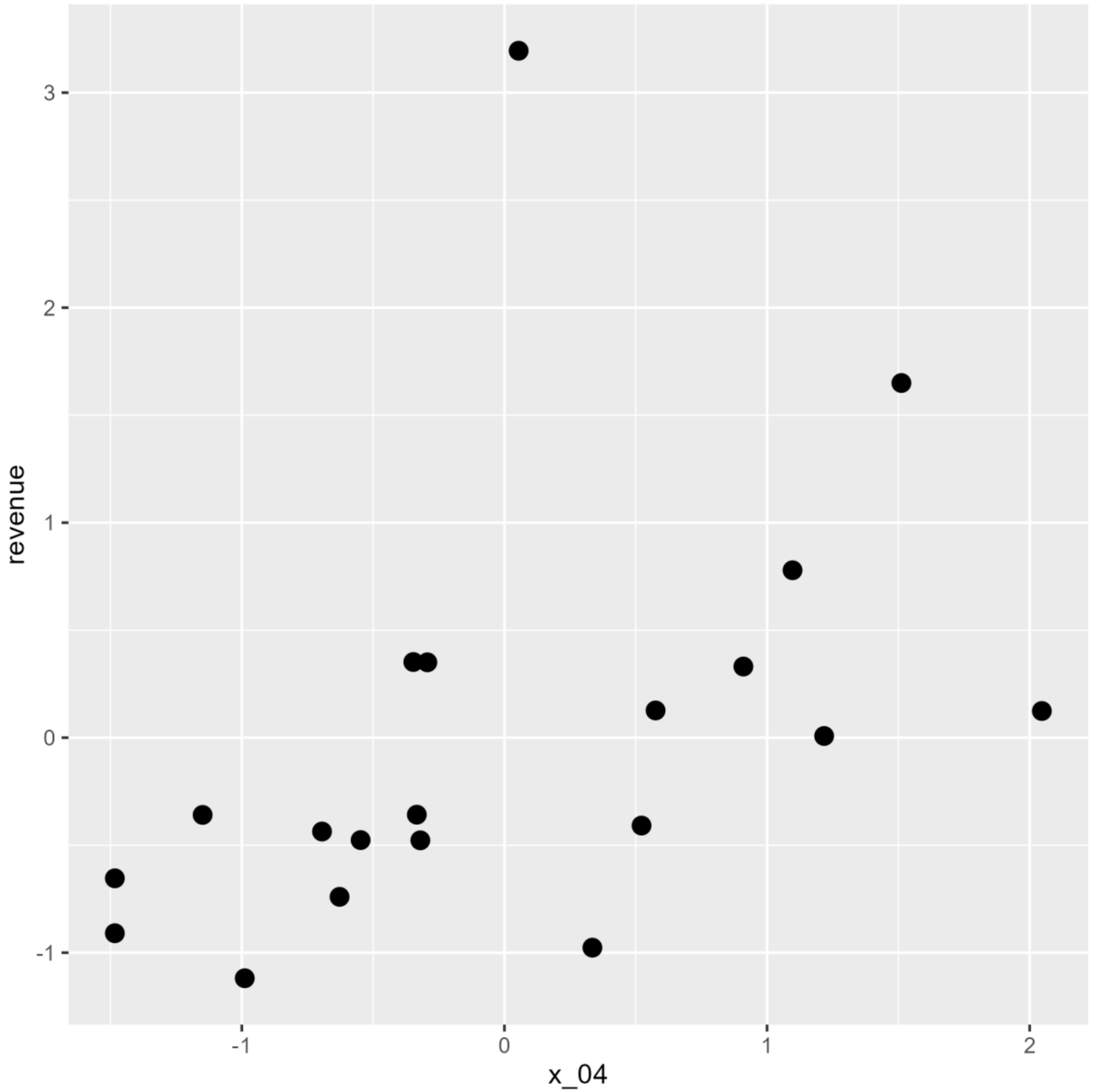


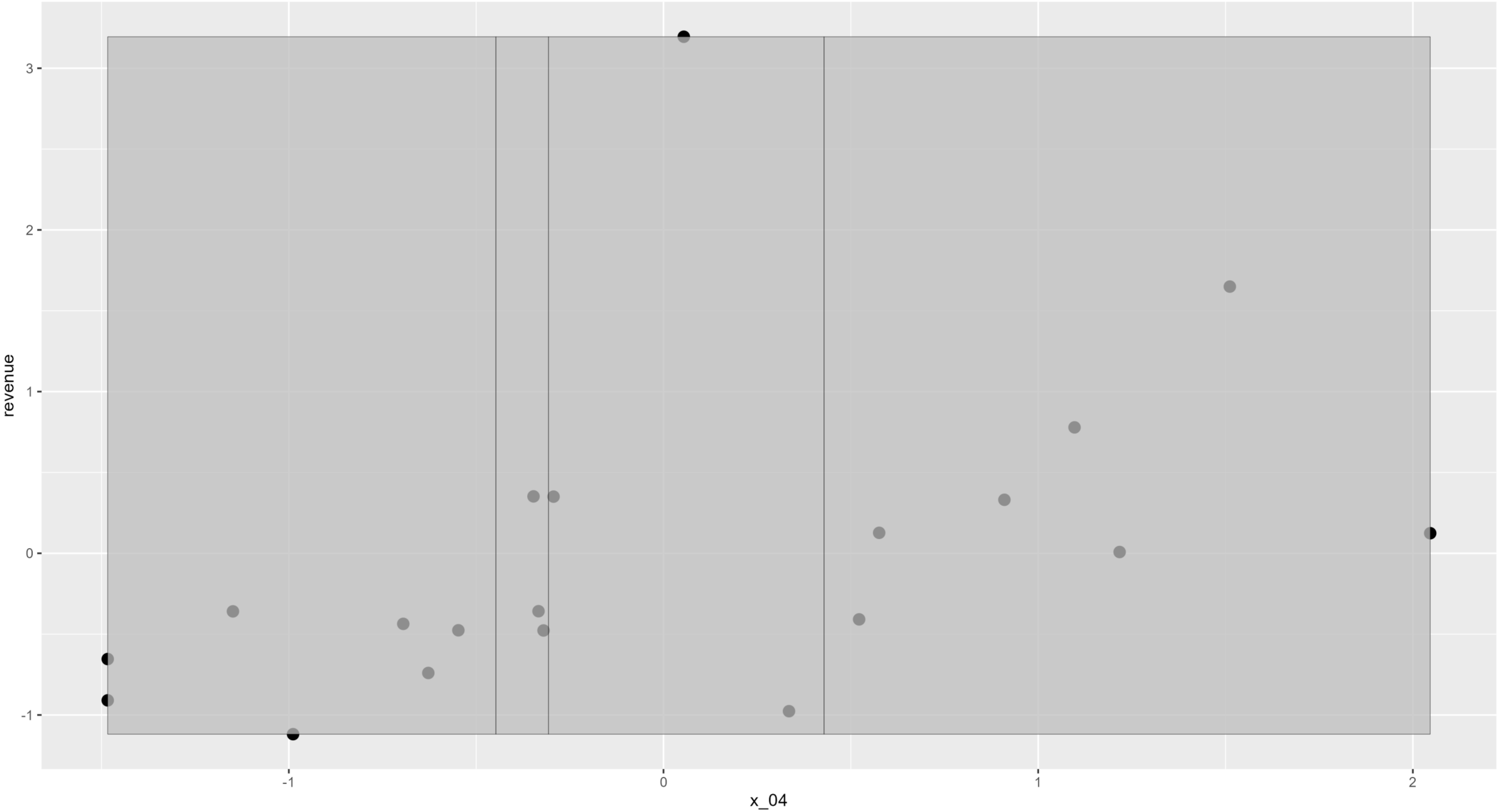
Non-linearity

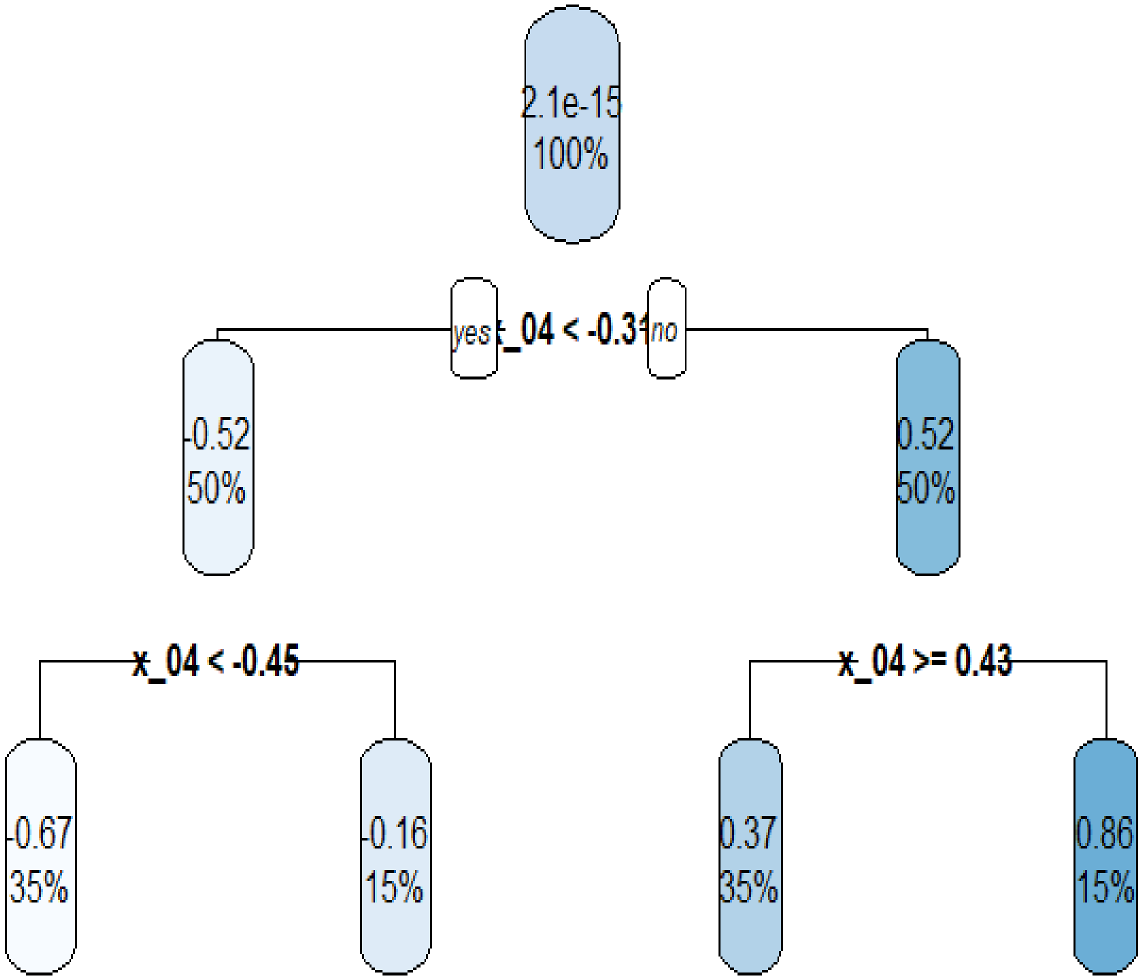

revenue
3
2
1
0
-1
-1
0
1
2
x_04

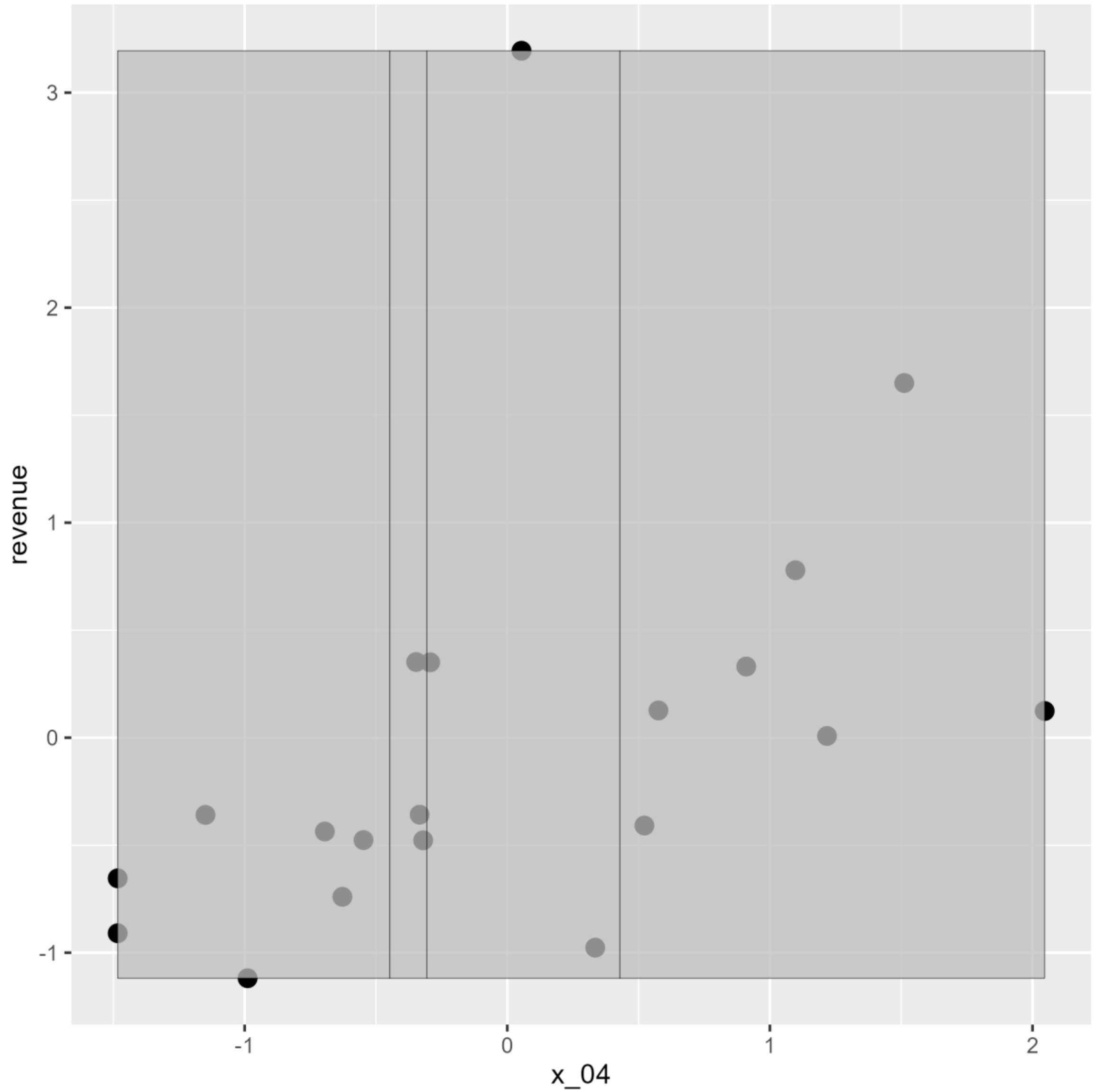

2.1e-15
100%
yes
x_04 < -0.31
no
-0.52
50%
0.52
50%
x_04 < -0.45
x_04 >= 0.43
-0.67
35%
-0.16
15%
0.37
35%
0.86
15%

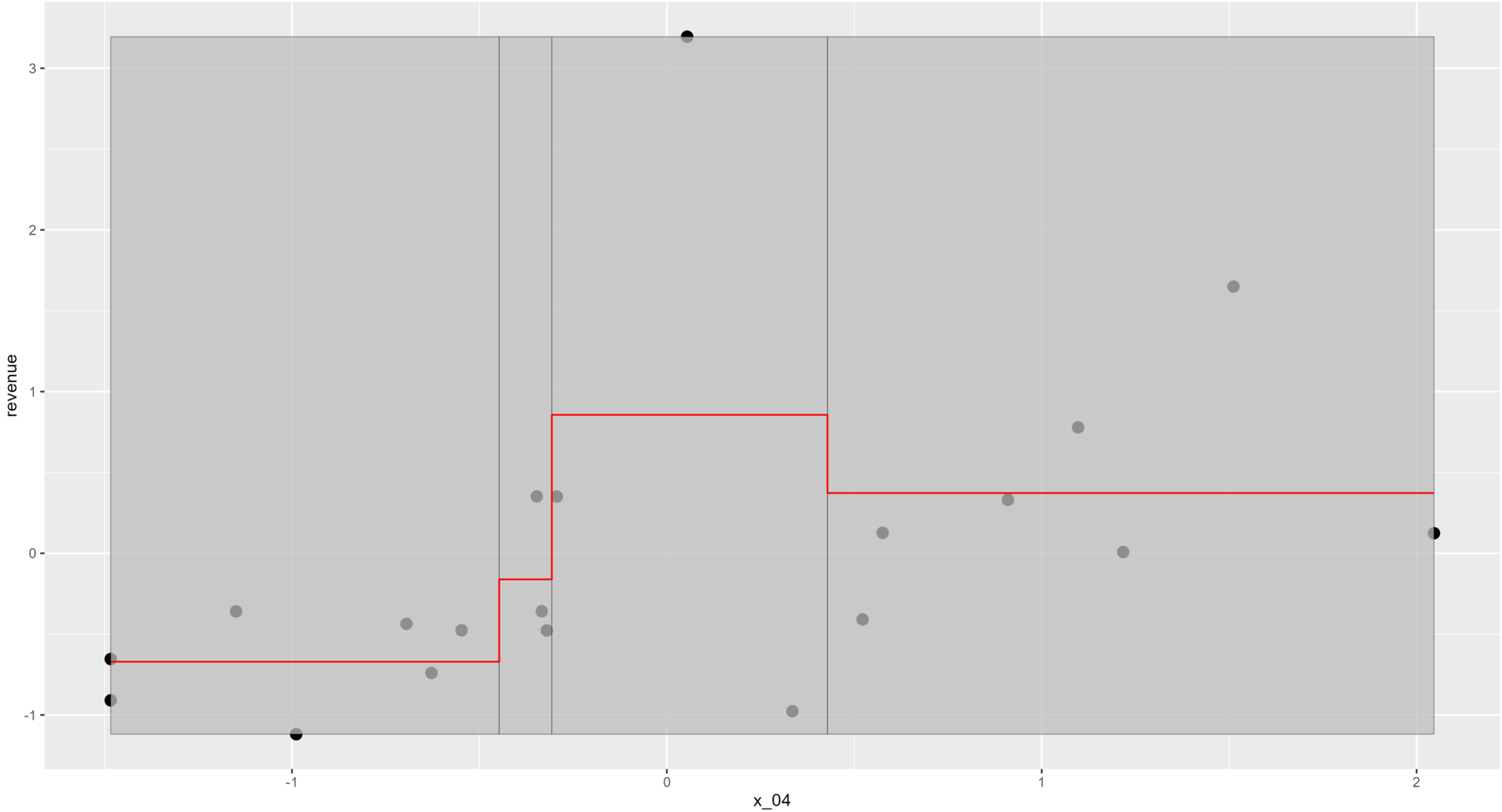

2.1e-15
100%
yes
x_04 < -0.31
no
-0.52
50%
0.52
50%
x_04 < -0.45
x_04 >= 0.43
-0.67
35%
-0.16
15%
0.37
35%
0.86
15%


revenue
3
2
1
0
-1
x_04
-1
0
1
2

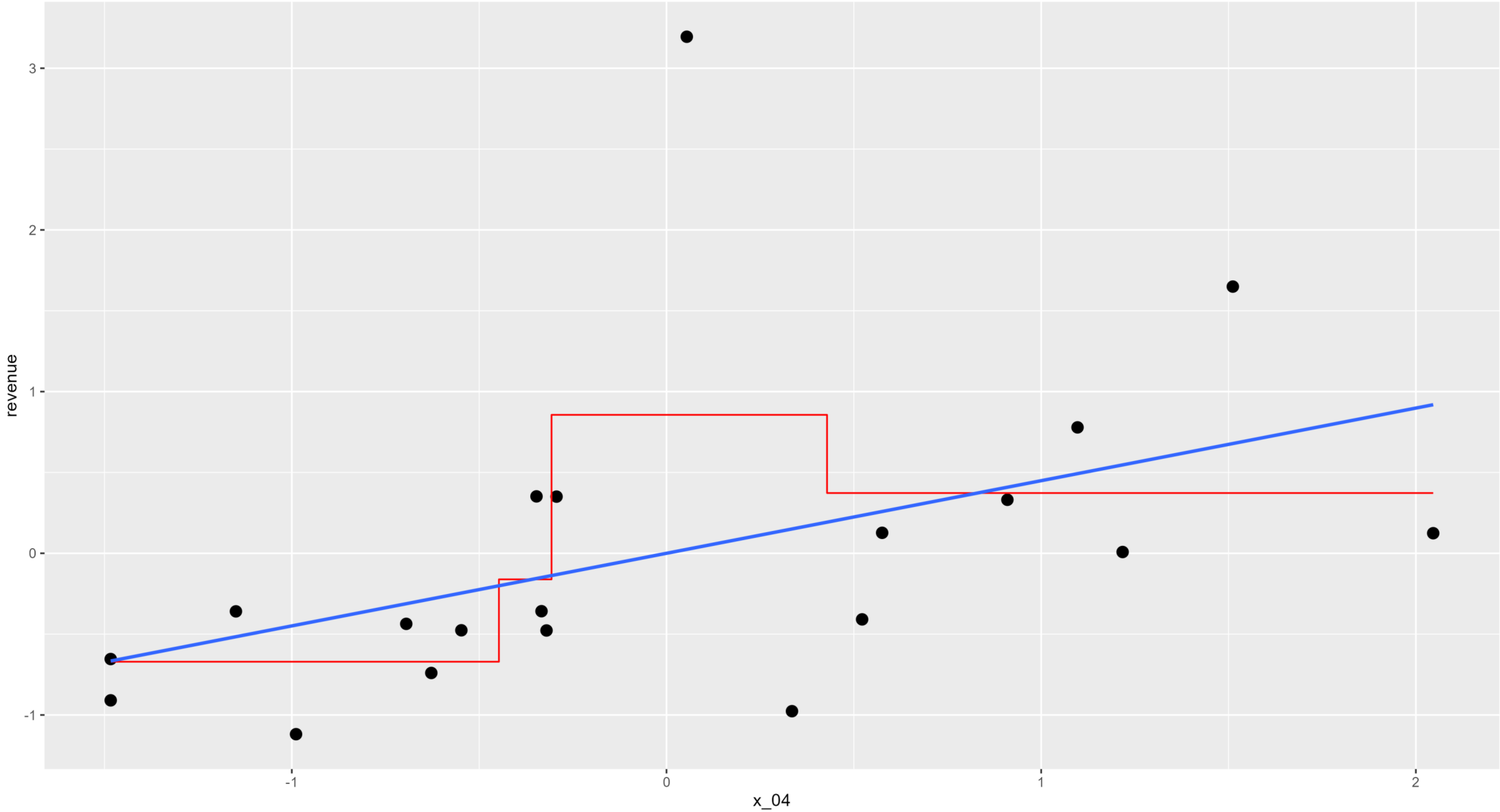

revenue
3
2
1
0
-1
-1
0
1
2
x_04


2.1e-15
100%
yes
x_04 < -0.31
no
-0.52
50%
0.52
50%
x_04 < -0.45
x_04 >= 0.43
-0.67
35%
-0.16
15%
0.37
35%
0.86
15%

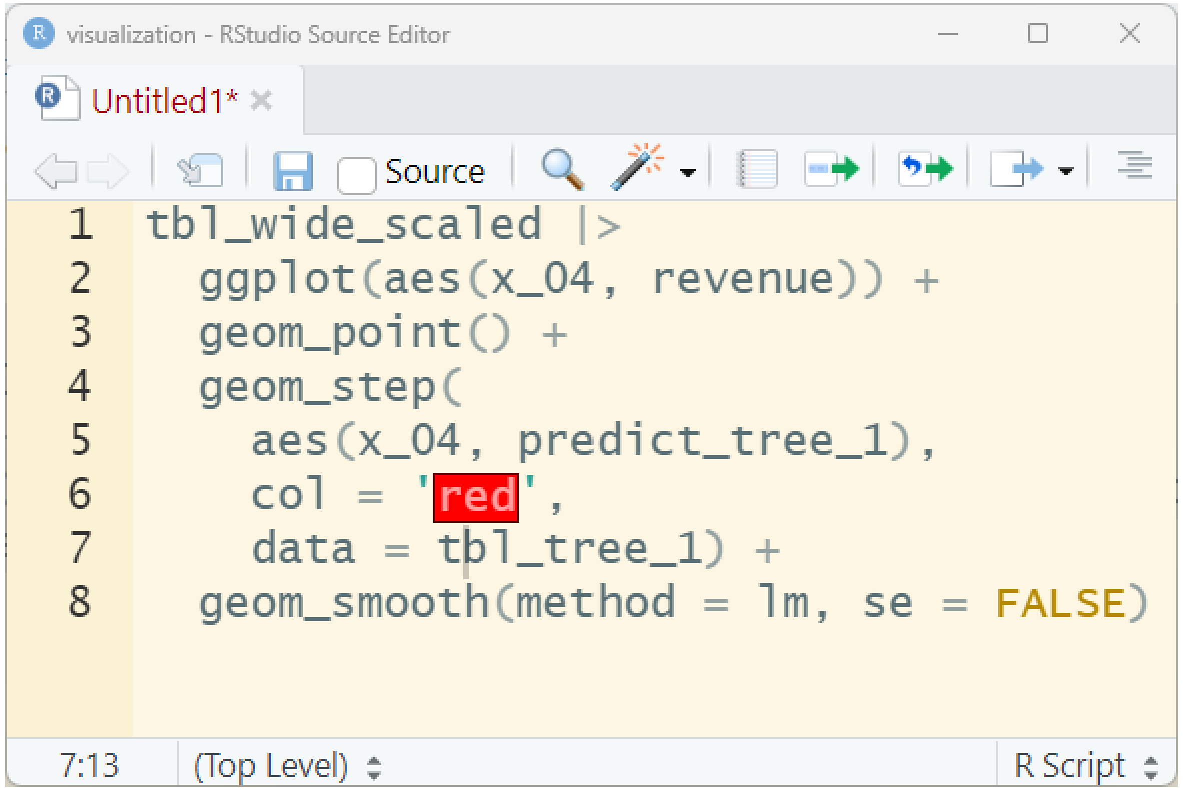

revenue
3
2
1
0
-1
x_04
-1
0
1
2



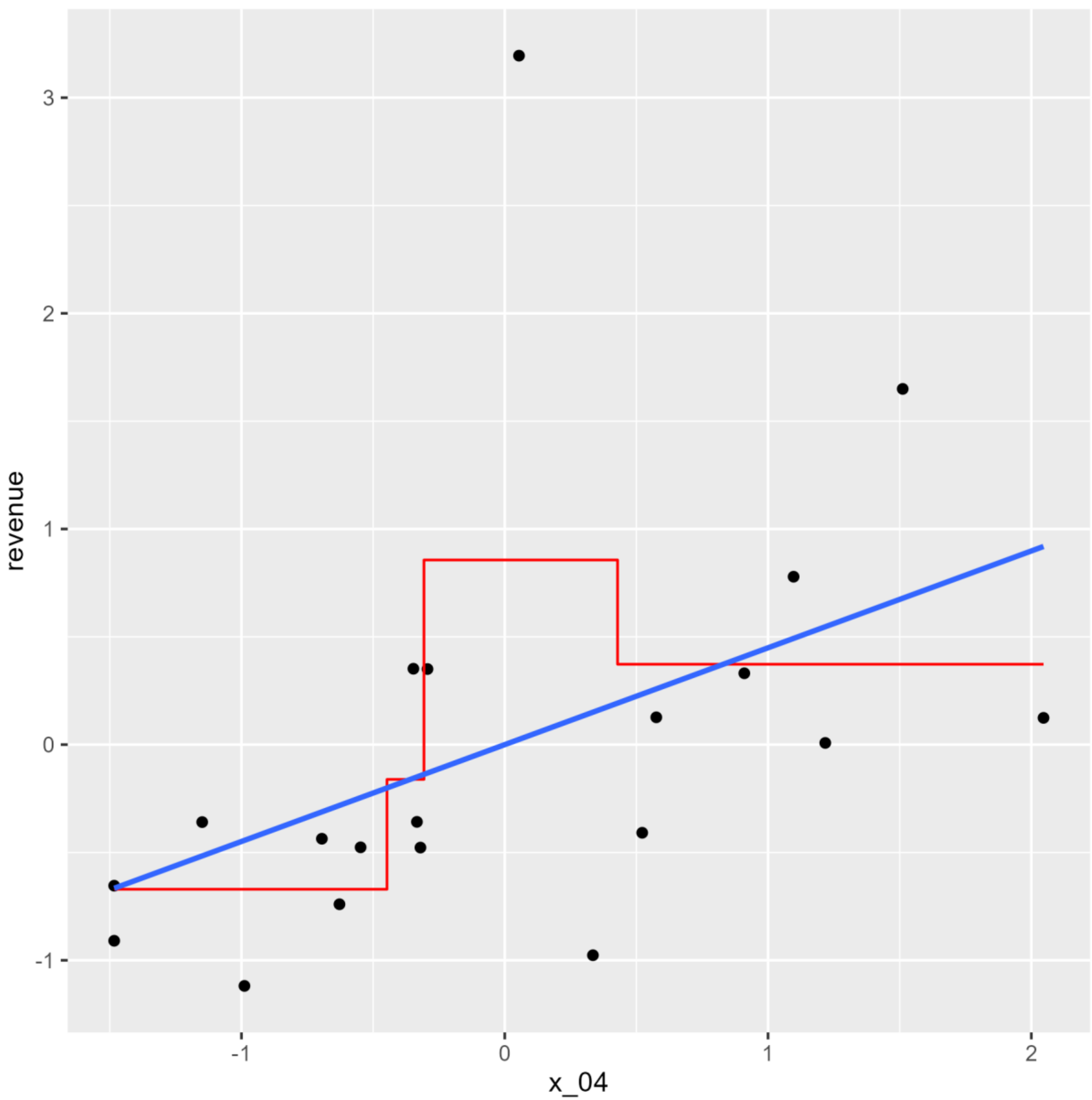
revenue
3
2
1
0
-1
x_04
-1
0
1
2

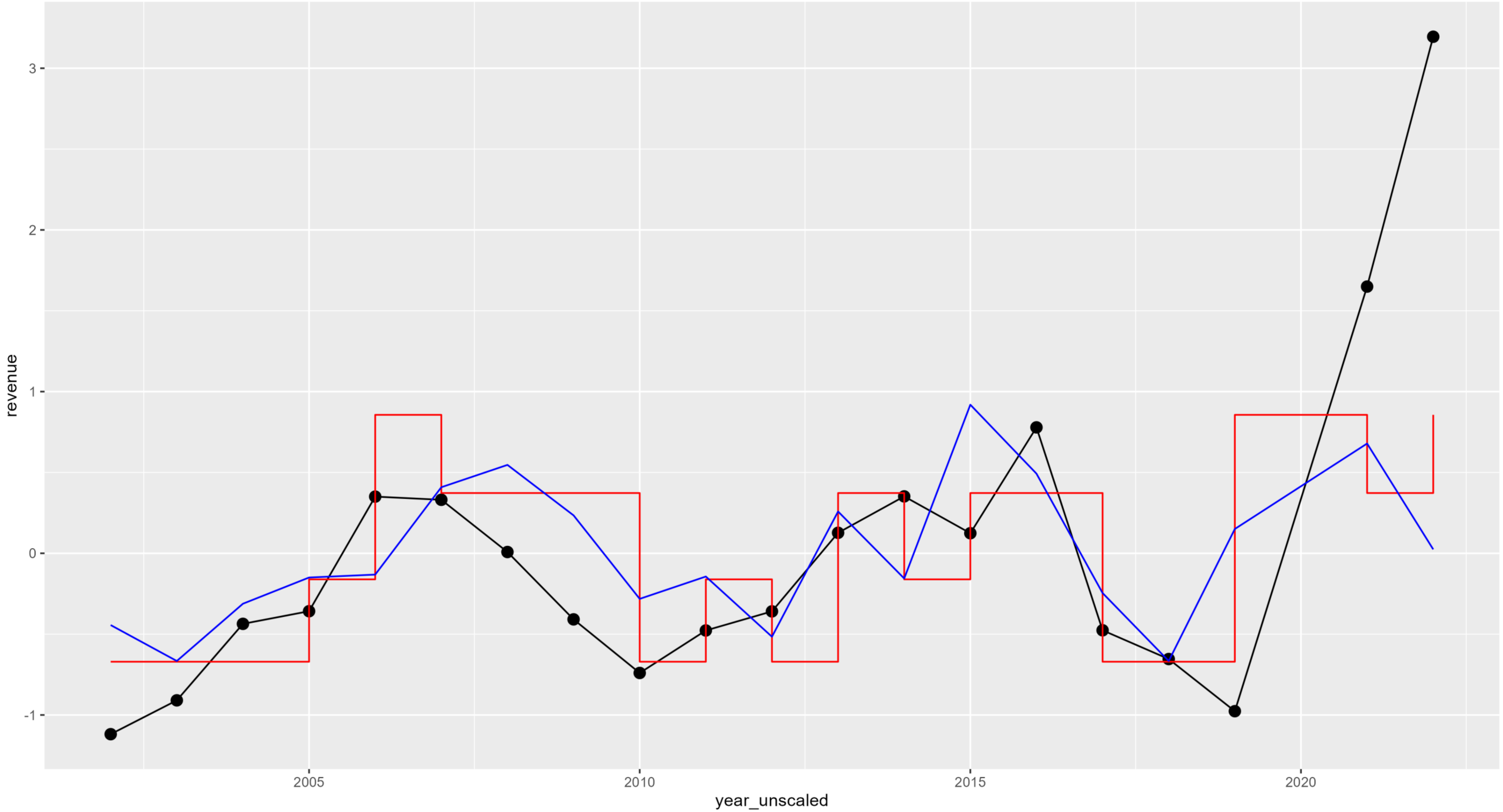


3
2
1
0
-1
revenue
-1
0
1
2
x_04

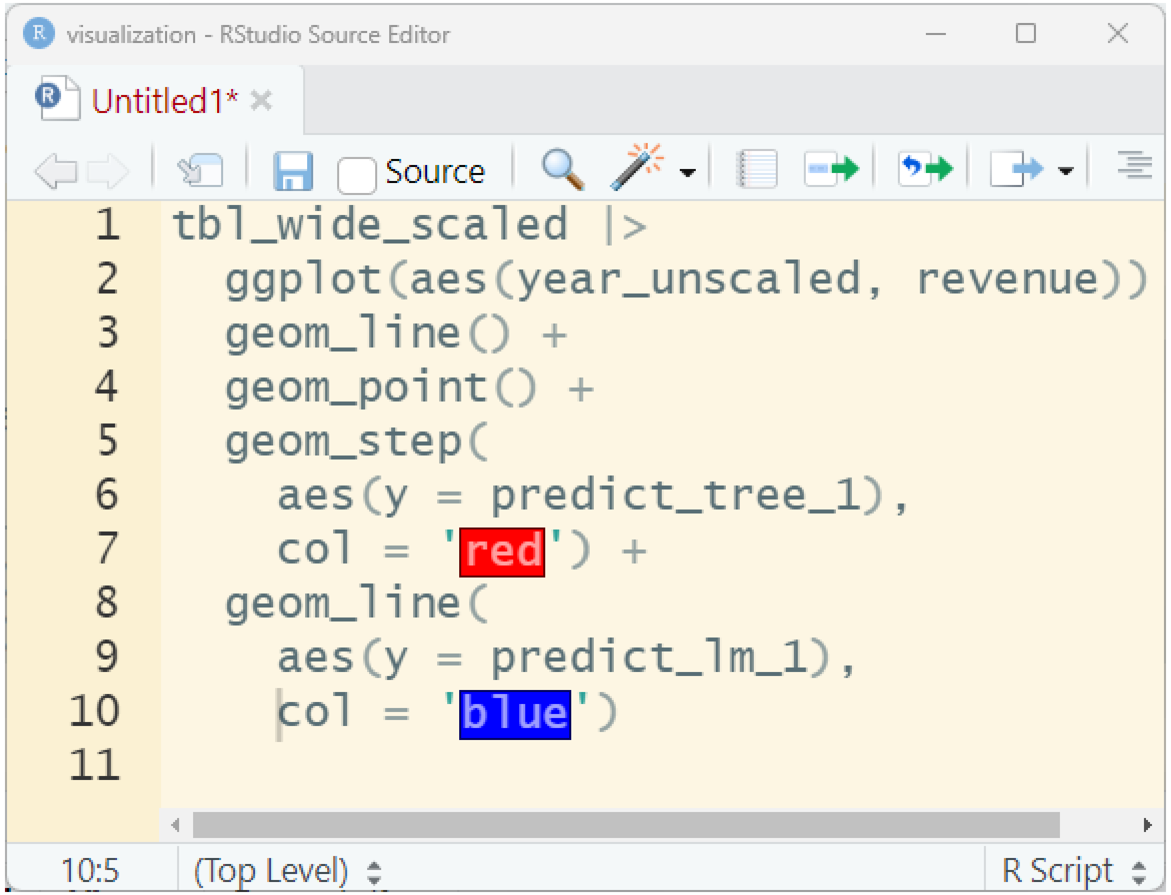


```r
tbl_wide_scaled |>
  ggplot(aes(x_04, revenue)) +
  geom_point() +
  geom_step(
    aes(x_04, predict_tree_1),
    col = 'red',
    data = tbl_tree_1) +
  geom_smooth(method = lm, se = FALSE)
```

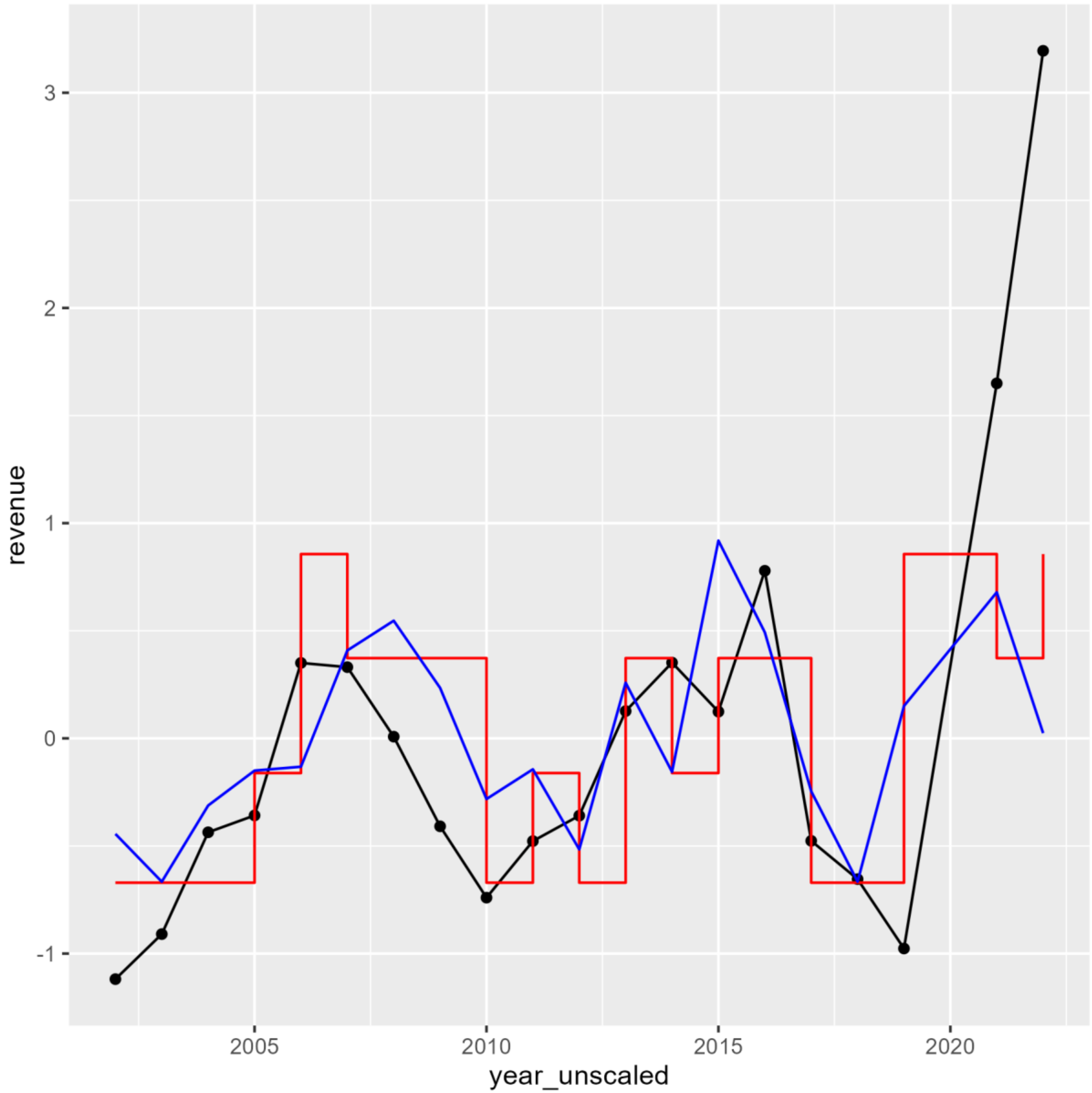


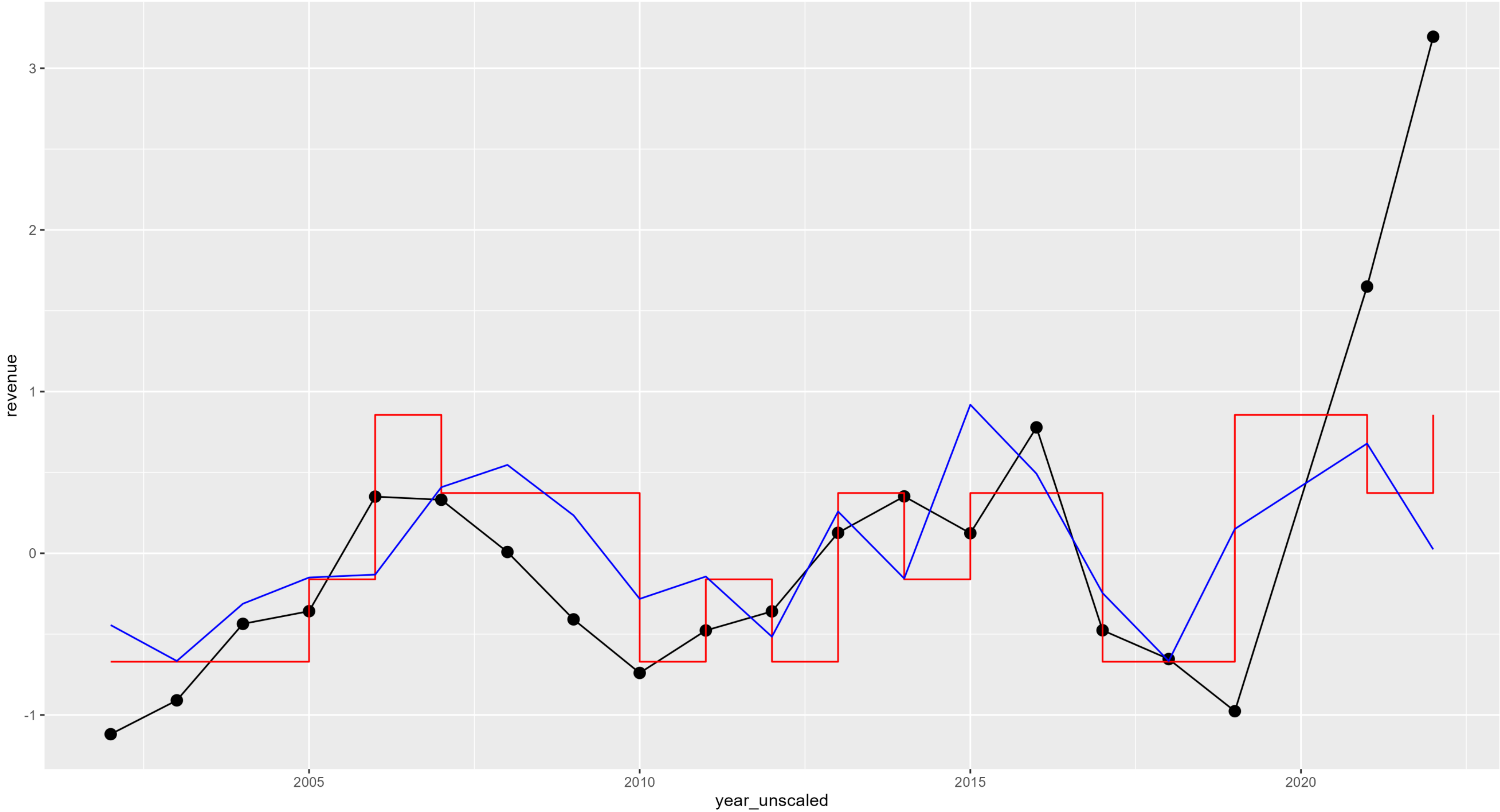


revenue
3
2
1
0
-1
2005
2010
2015
2020
year_unscaled



```
tbl_wide_scaled |>
  ggplot(aes(year_unscaled, revenue))
  geom_line() +
  geom_point() +
  geom_step(
    aes(y = predict_tree_1),
    col = 'red') +
  geom_line(
    aes(y = predict_lm_1),
    col = 'blue')

```




revenue
3
2
1
0
-1
2005
2010
2015
2020
year_unscaled

Clarity Matters

Product	25th Percentile	Median	Mean	75th Percentile	90th Percentile
A	14,738	23,047	24,222	27,995	35,049
B	46,333	59,952	60,119	66,669	72,812
C	15,038	22,852	23,831	28,062	31,854
D	7,333	17,956	24,383	30,177	46,827

Product	25th Percentile	Median	Mean	75th Percentile	90th Percentile
A	14,738	23,047	24,222	27,995	35,049
B	46,333	59,952	60,119	66,669	72,812
C	15,038	22,852	23,831	28,062	31,854
D	7,333	17,956	24,383	30,177	46,827

Twenty data points.

Product	25th Percentile	Median	Mean	75th Percentile	90th Percentile
A	14,738	23,047	24,222	27,995	35,049
B	46,333	59,952	60,119	66,669	72,812
C	15,038	22,852	23,831	28,062	31,854
D	7,333	17,956	24,383	30,177	46,827

Twenty data points.
Information overload.

Product	25th Percentile	Median	Mean	75th Percentile	90th Percentile
A	14,738	23,047	24,222	27,995	35,049
B	46,333	59,952	60,119	66,669	72,812
C	15,038	22,852	23,831	28,062	31,854
D	7,333	17,956	24,383	30,177	46,827

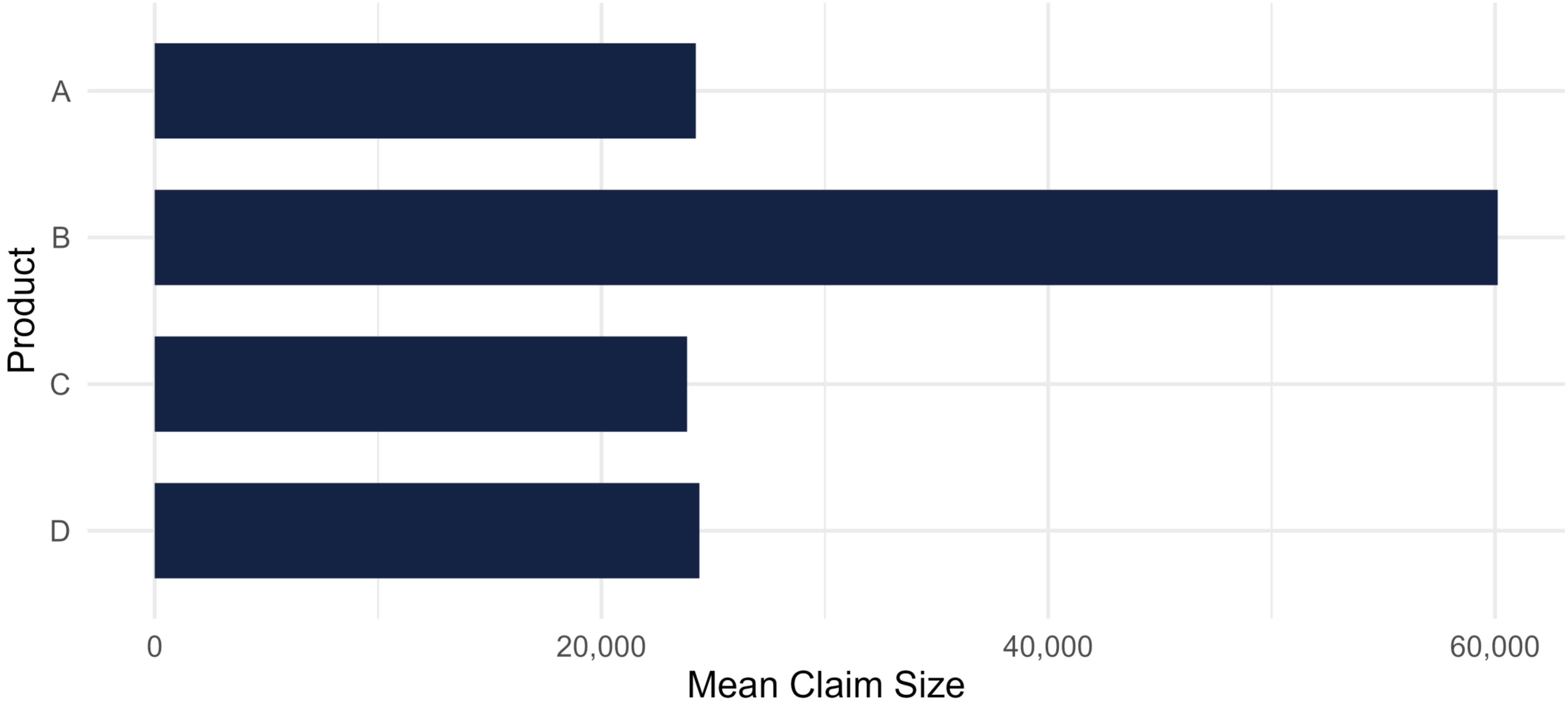
Twenty data points.

Information overload.

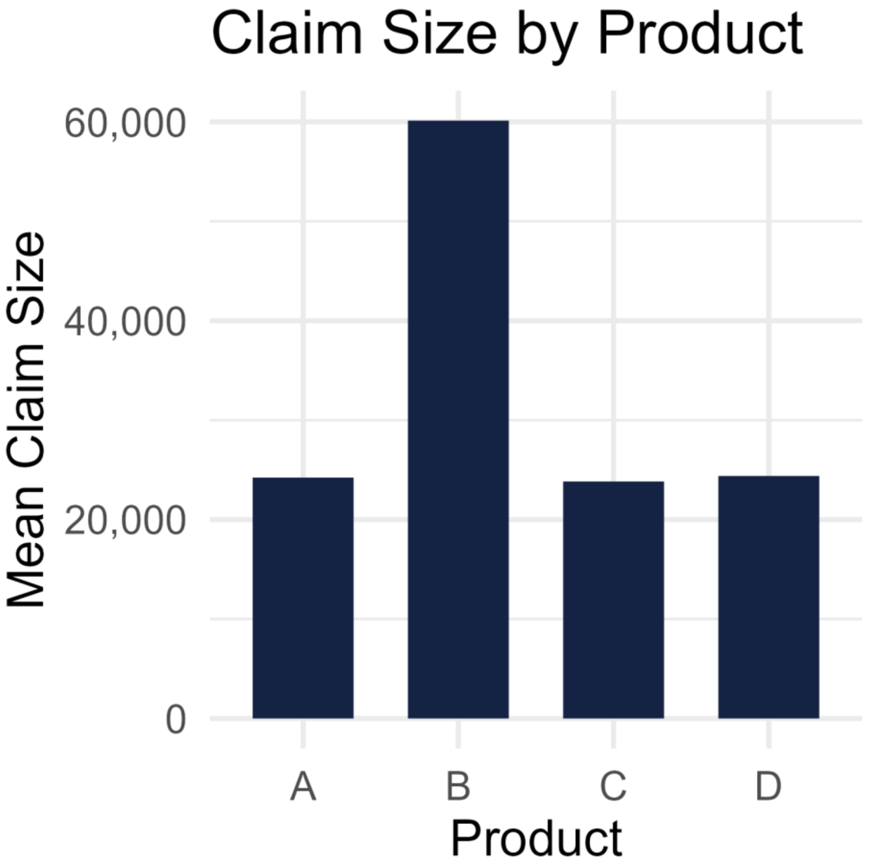
Especially for non-actuaries.

Severity estimate
of $60,000 for Product B
and $24,000 for all others

Claim Size by Product



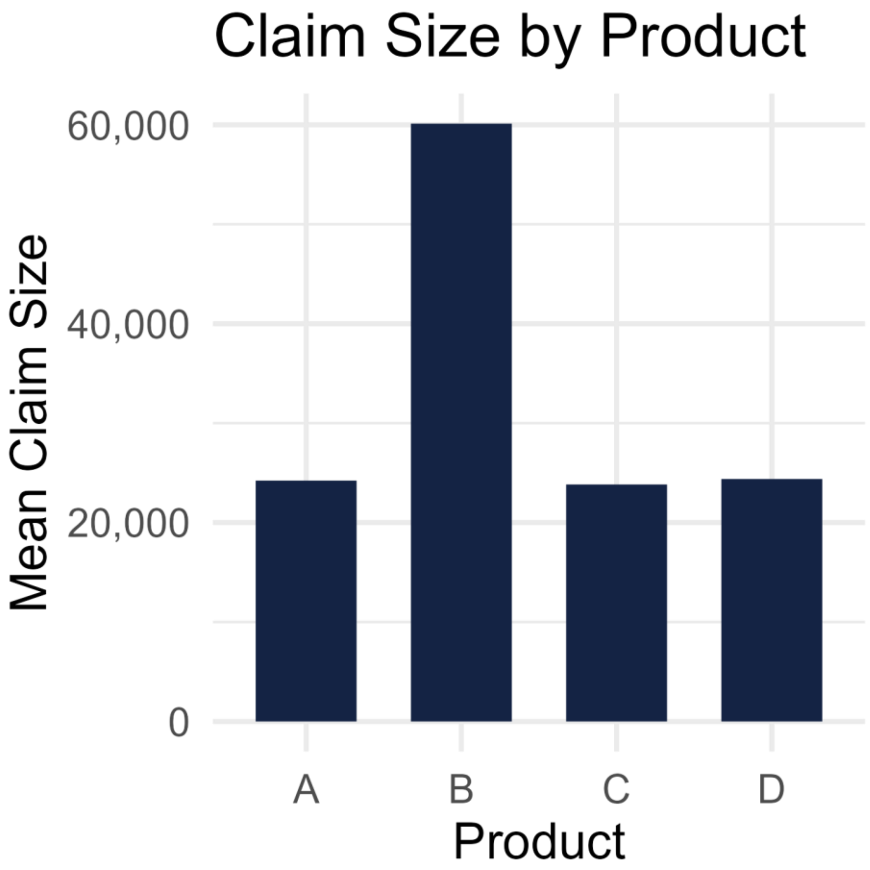
Severity estimate
of $60,000 for Product B
and $24,000 for all others



2 data points.

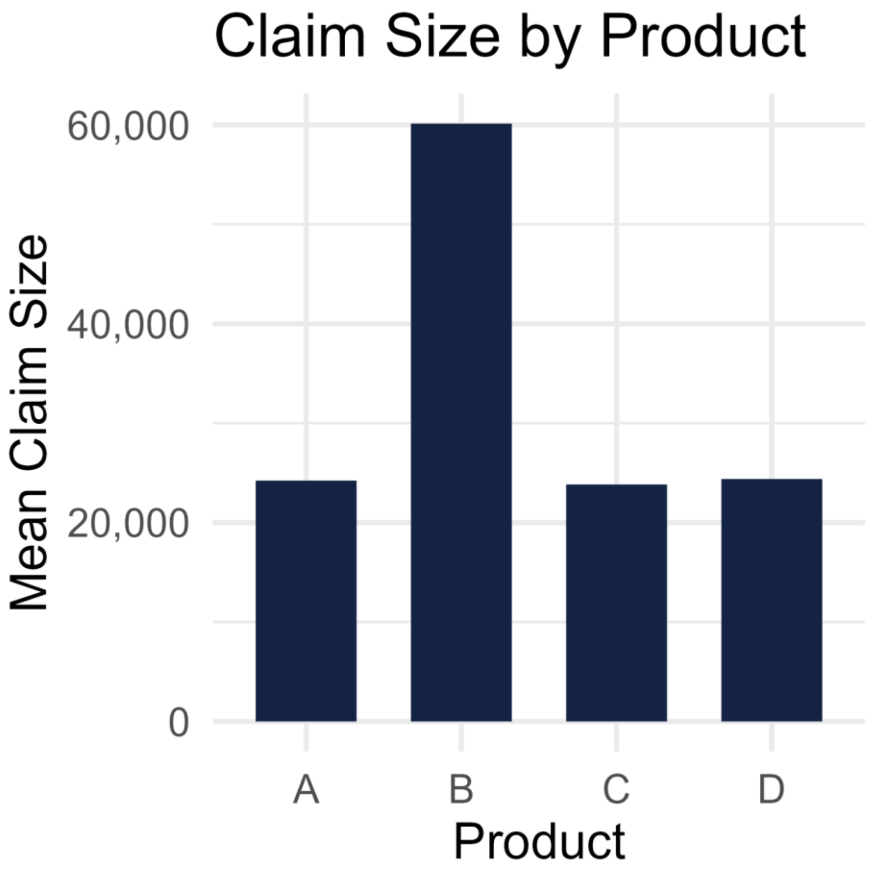
4 data points.

Product	25th %ile	Median	Mean	75th %ile	90th %ile
A	14,738	23,047	24,222	27,995	35,049
B	46,333	59,952	60,119	66,669	72,812
C	15,038	22,852	23,831	28,062	31,854
D	7,333	17,956	24,383	30,177	46,827

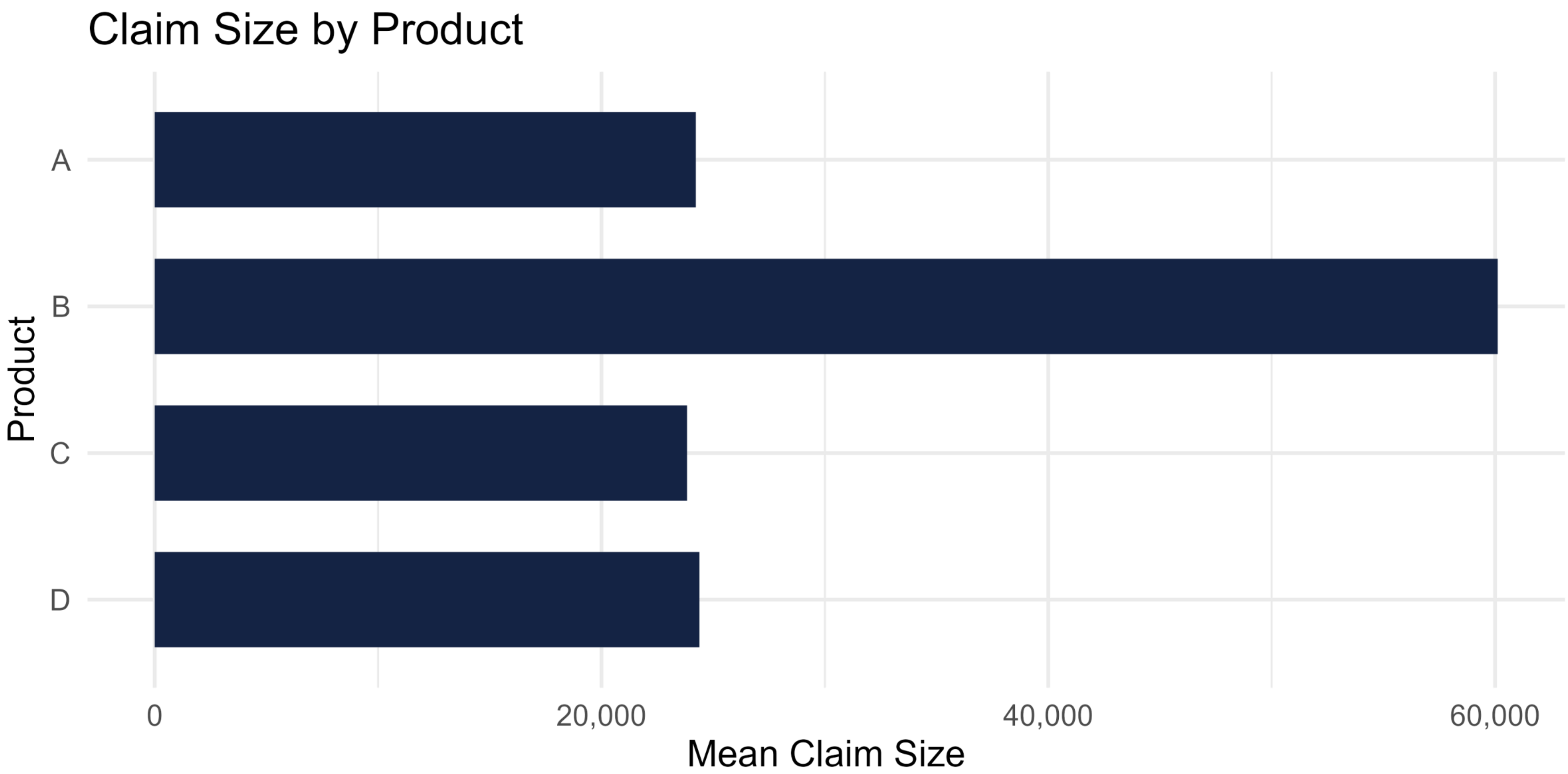


Which product might require more supporting capital?

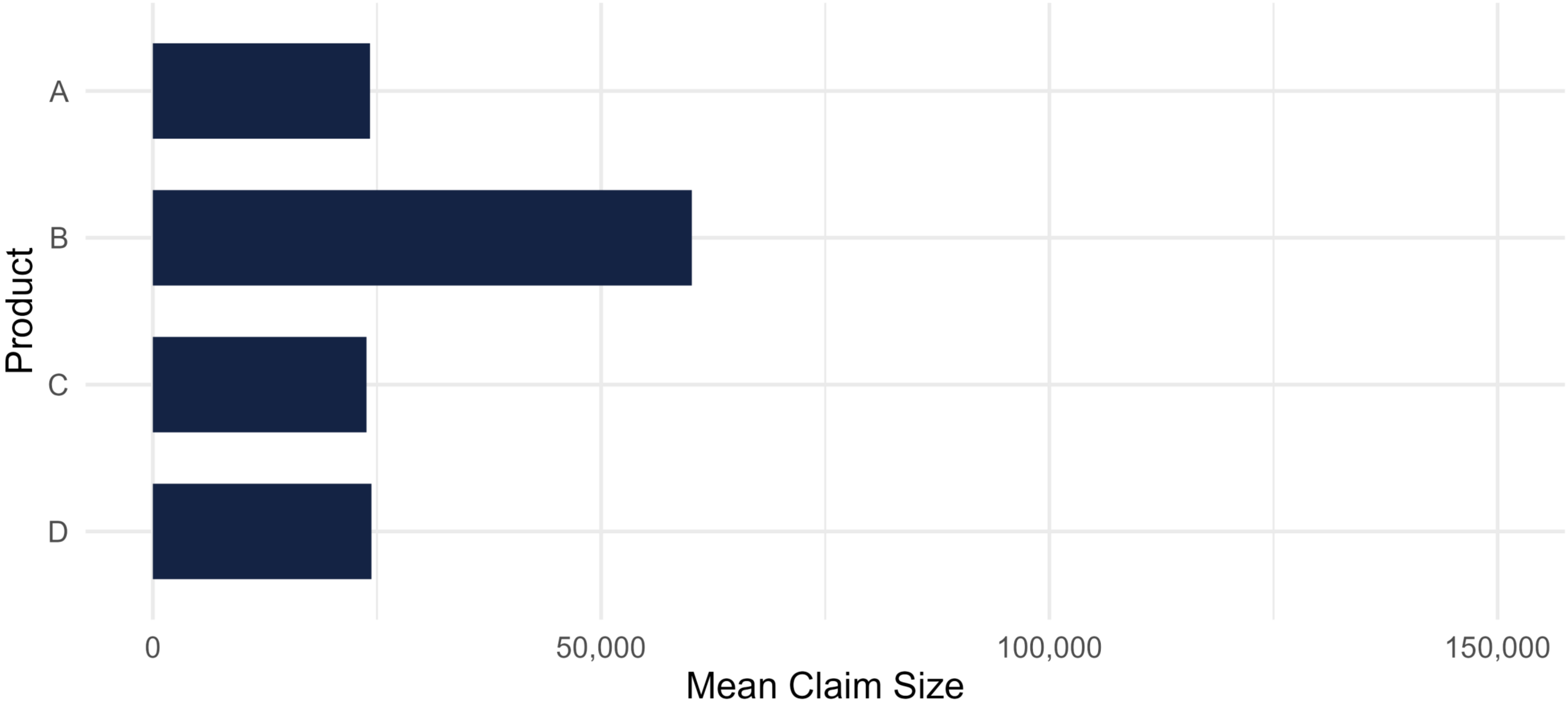
Product	25th %ile	Median	Mean	75th %ile	90th %ile
A	14,738	23,047	24,222	27,995	35,049
B	46,333	59,952	60,119	66,669	72,812
C	15,038	22,852	23,831	28,062	31,854
D	7,333	17,956	24,383	30,177	46,827

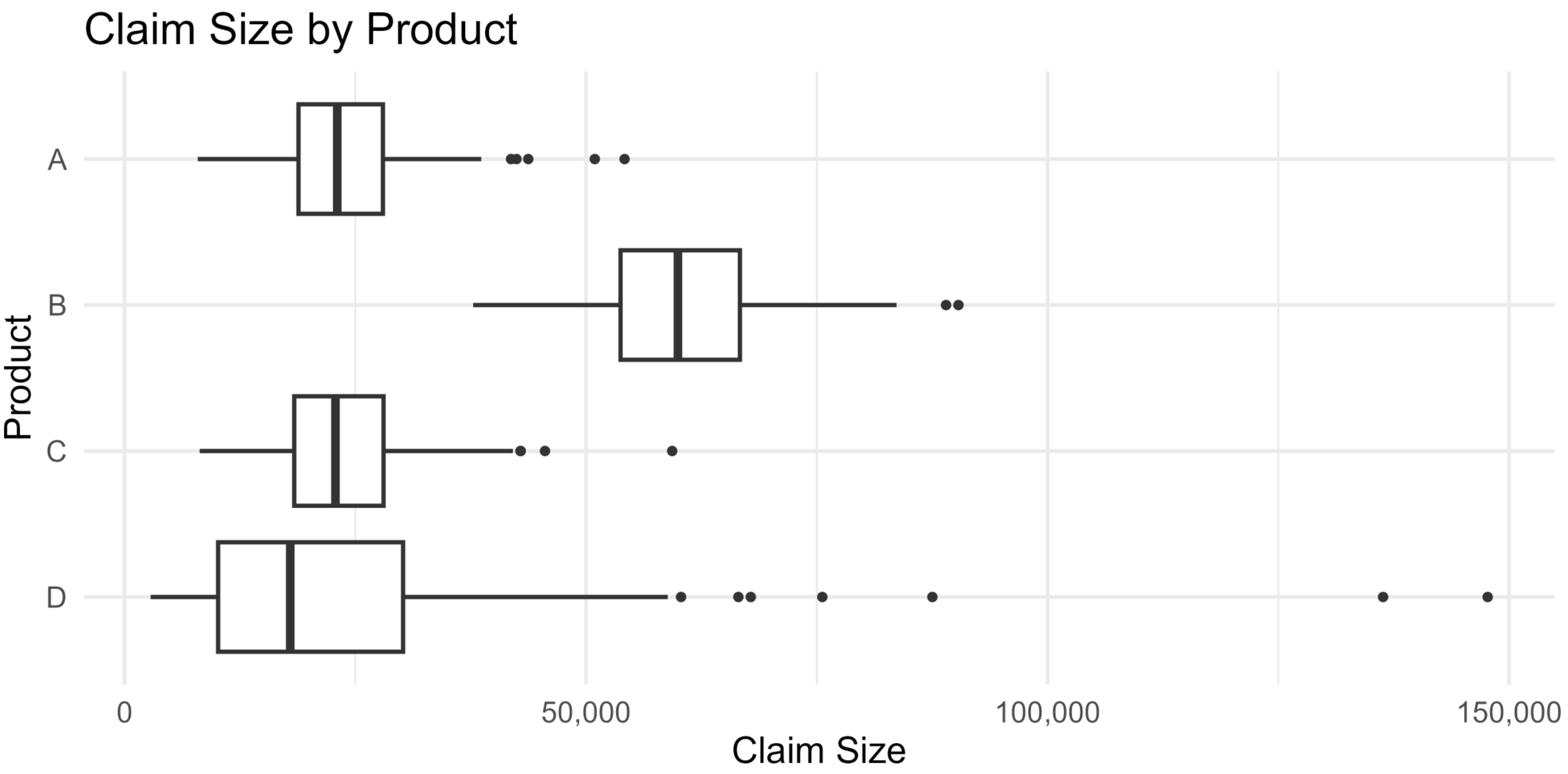


It isn't immediately clear.


Claim Size by Product
Product
A
B
C
D
0
20,000
40,000
60,000
Mean Claim Size

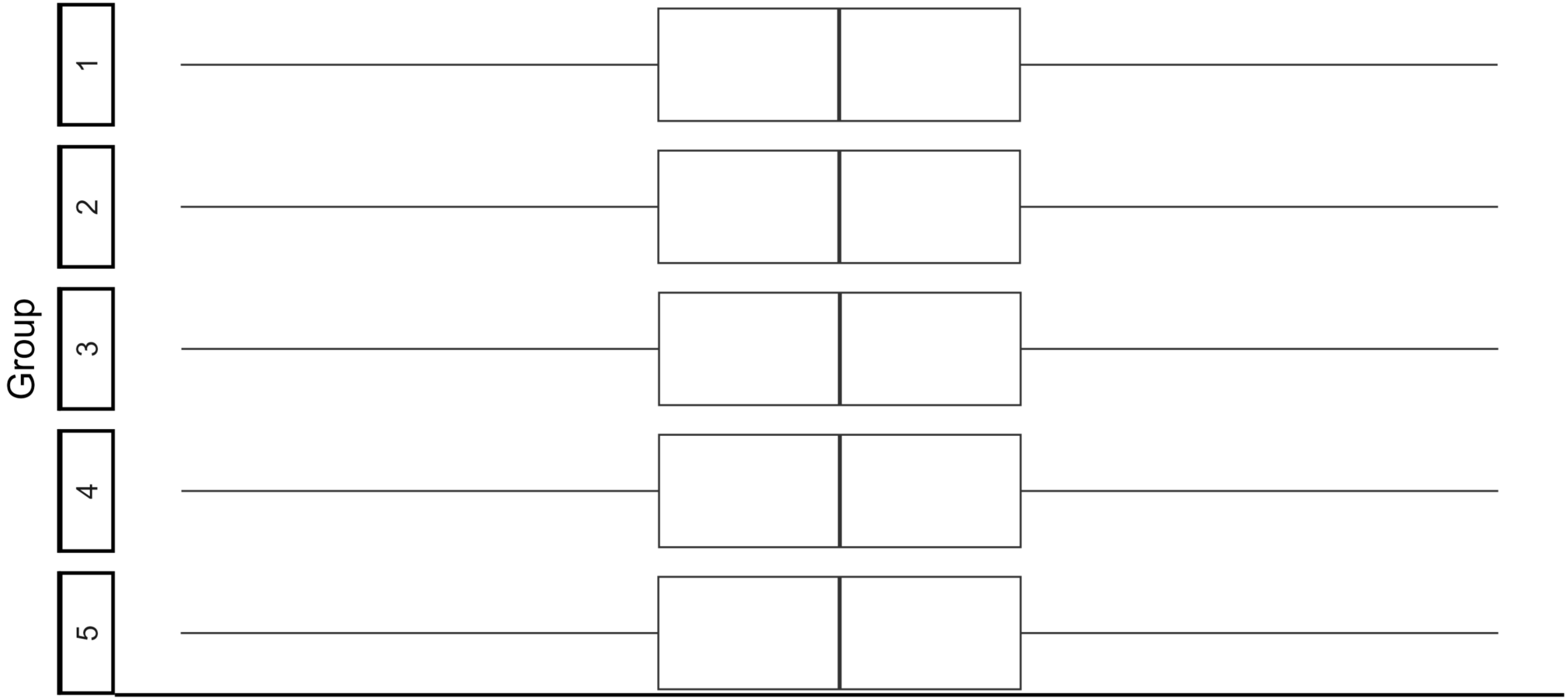
Claim Size by Product



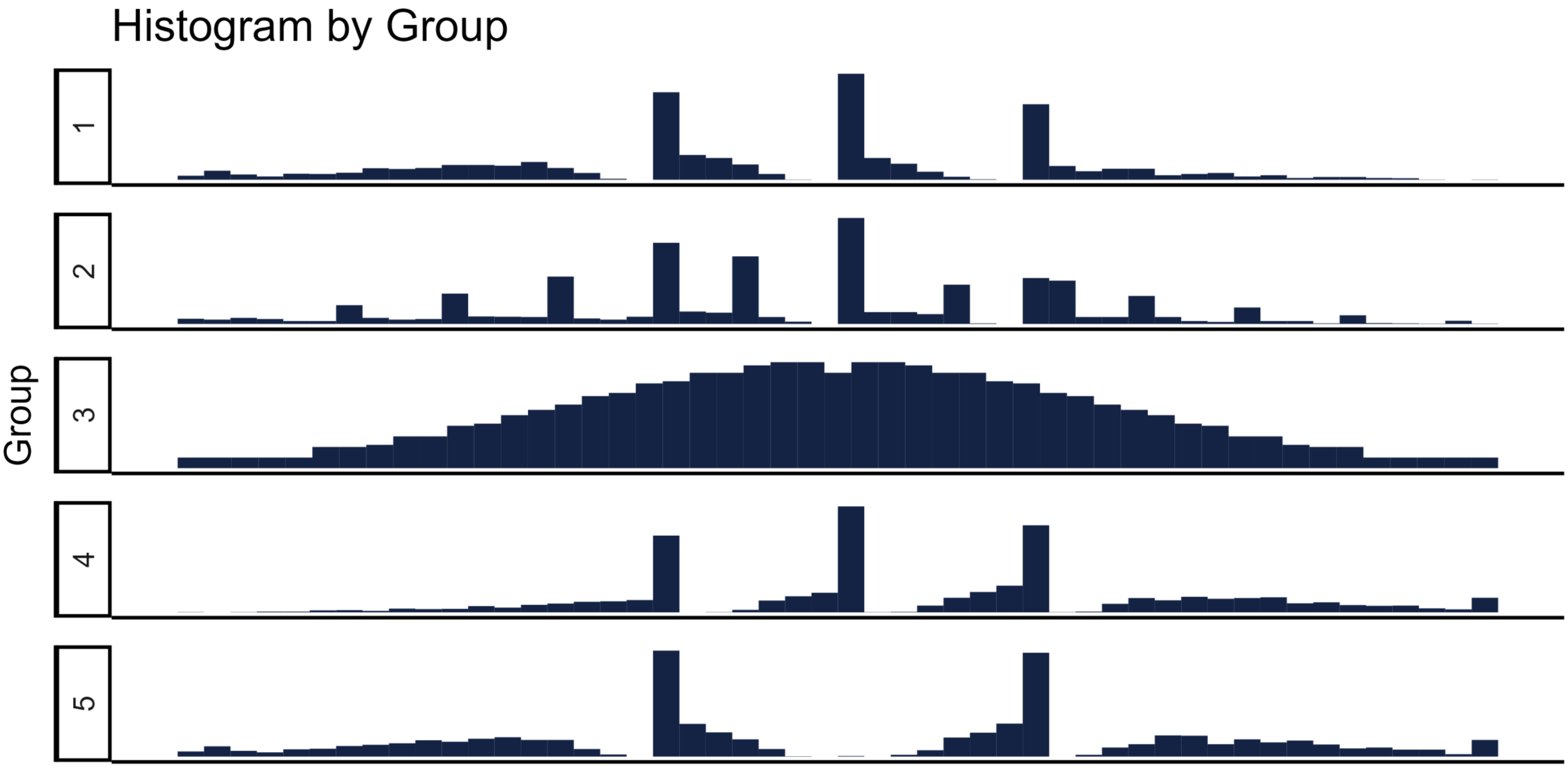


Claim Size by Product

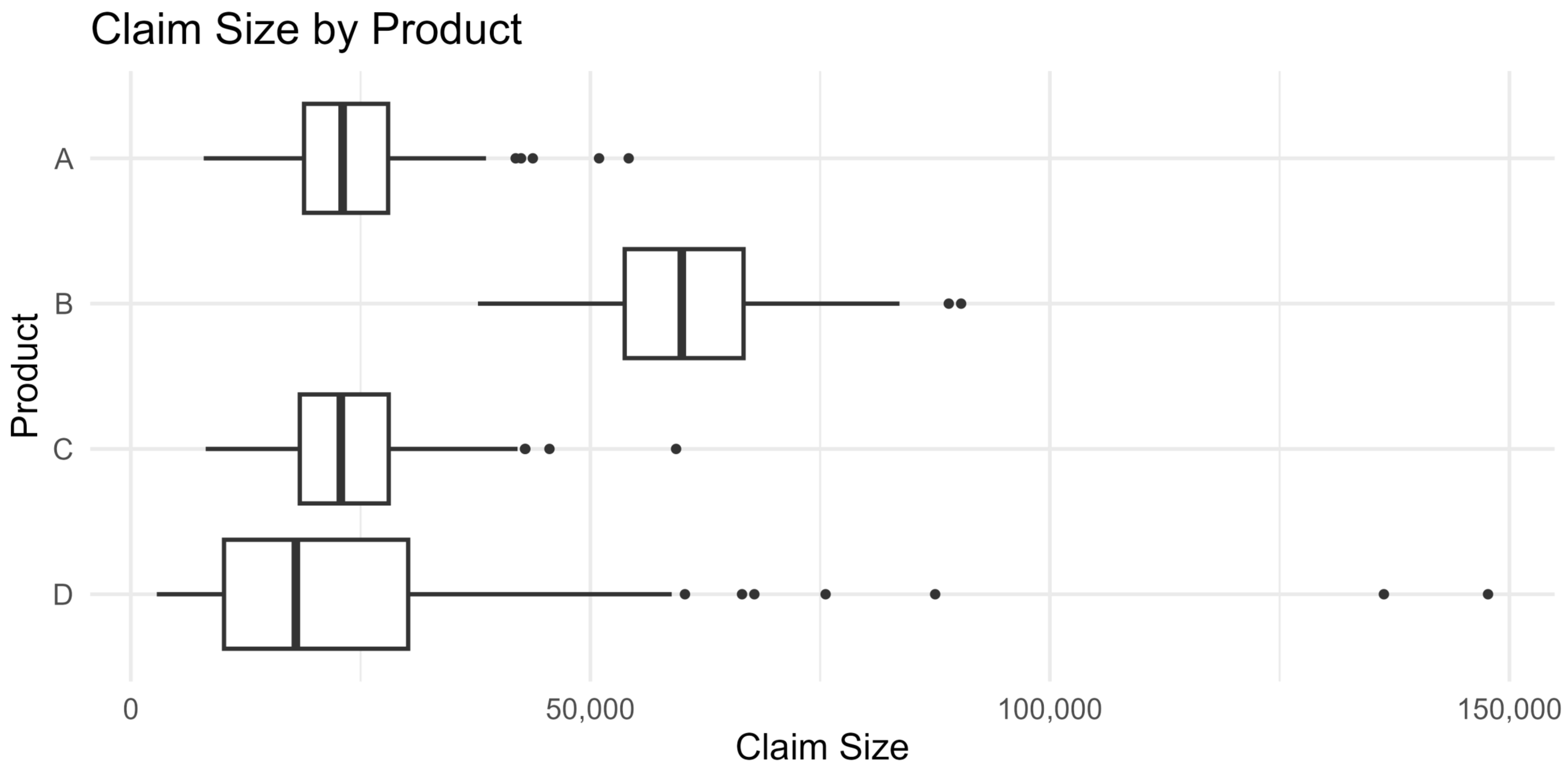
Boxplot by Group

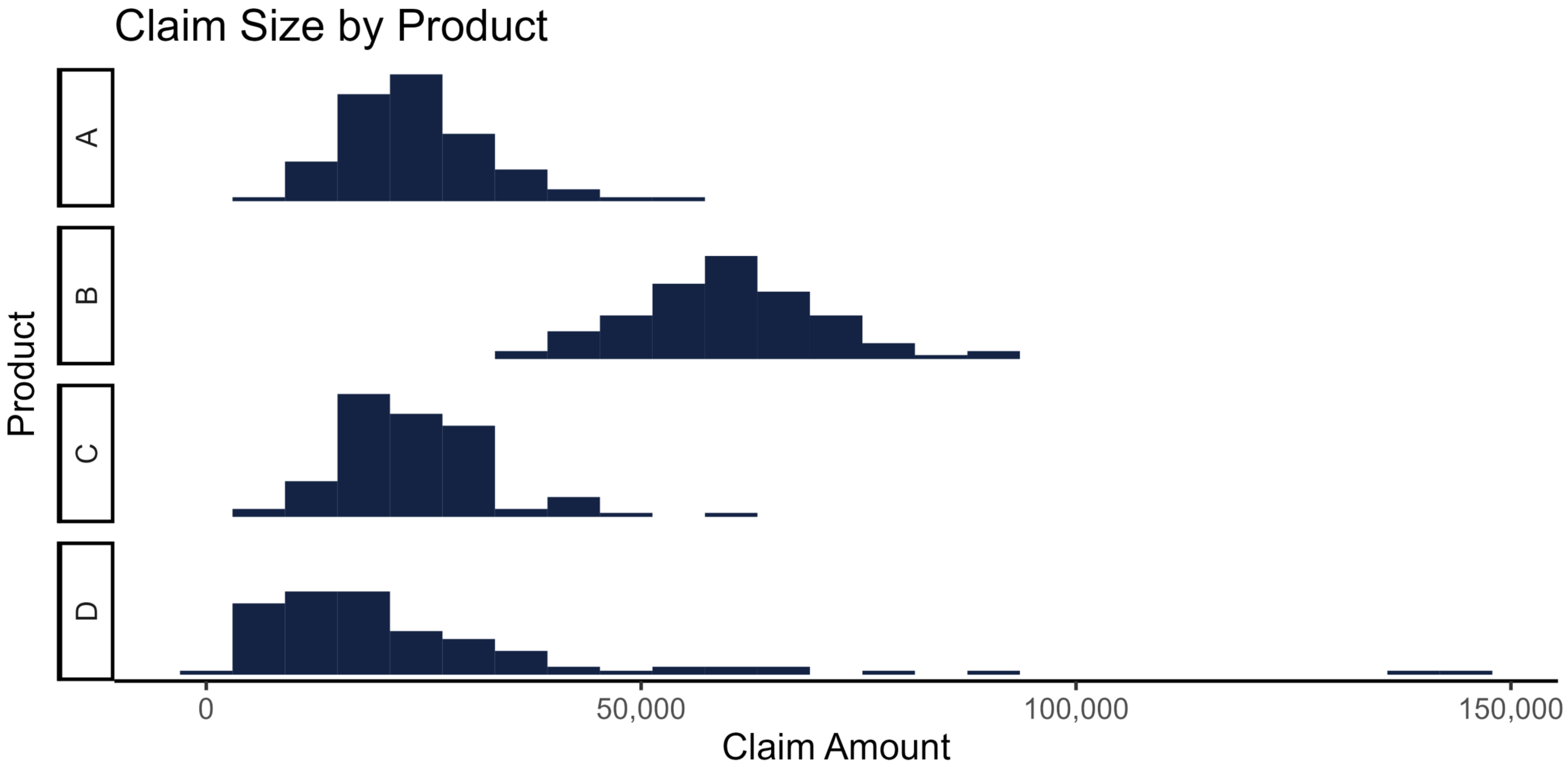


Matejka, J., & Fitzmaurice , G. (2017). Same Stats, Different Graphs… *Autodesk Research*.

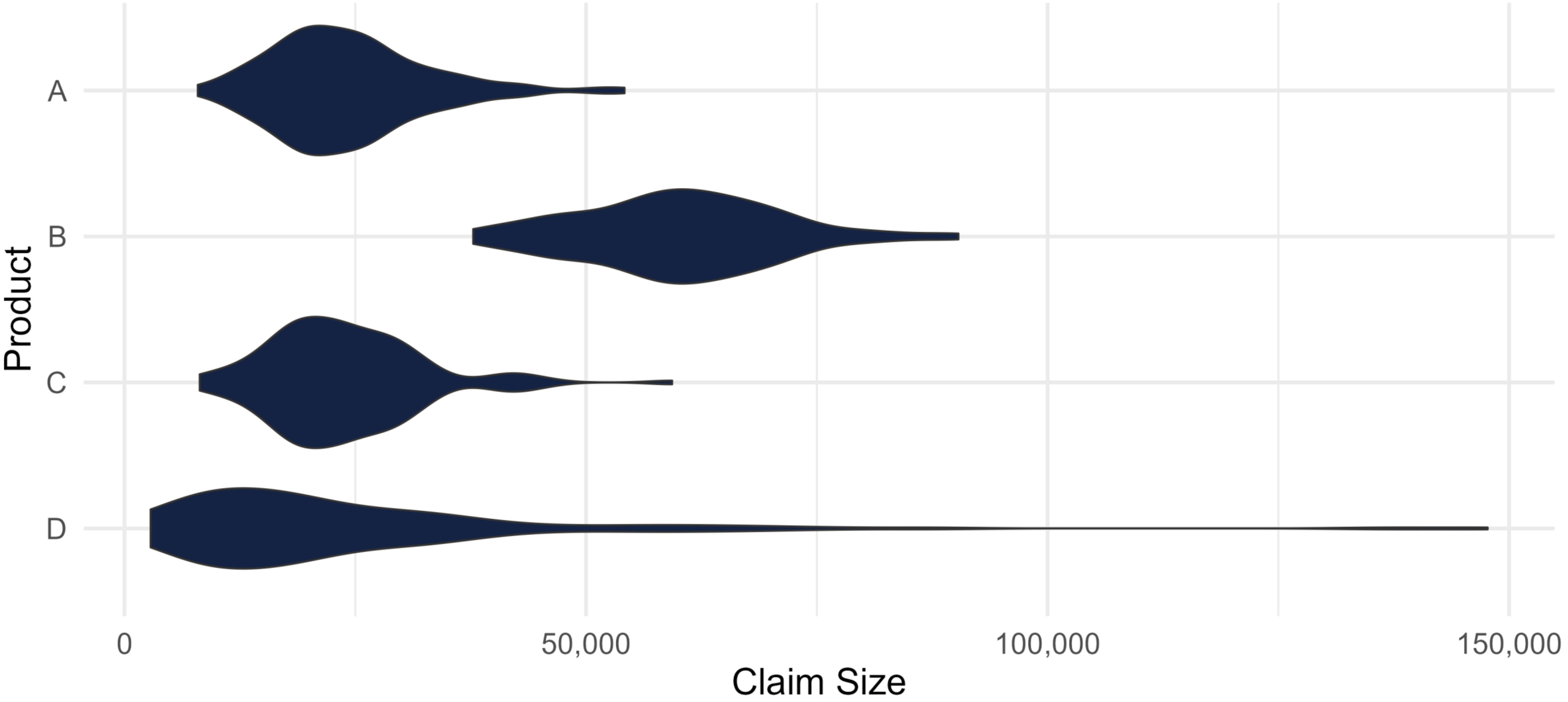


Understanding the distribution matters.

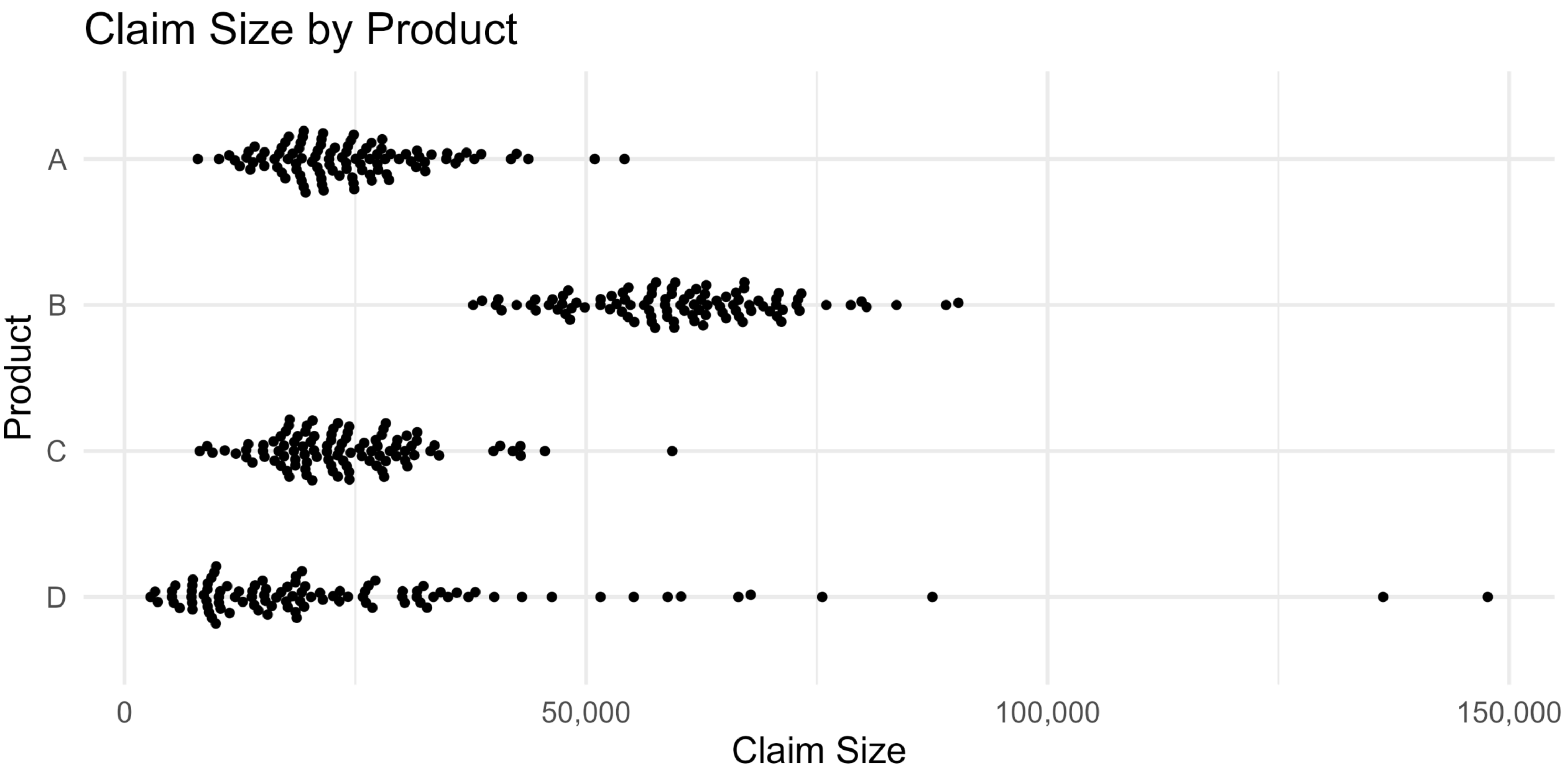

Claim Size by Product
Product
A
B
C
D
0
50,000
100,000
150,000
Claim Size

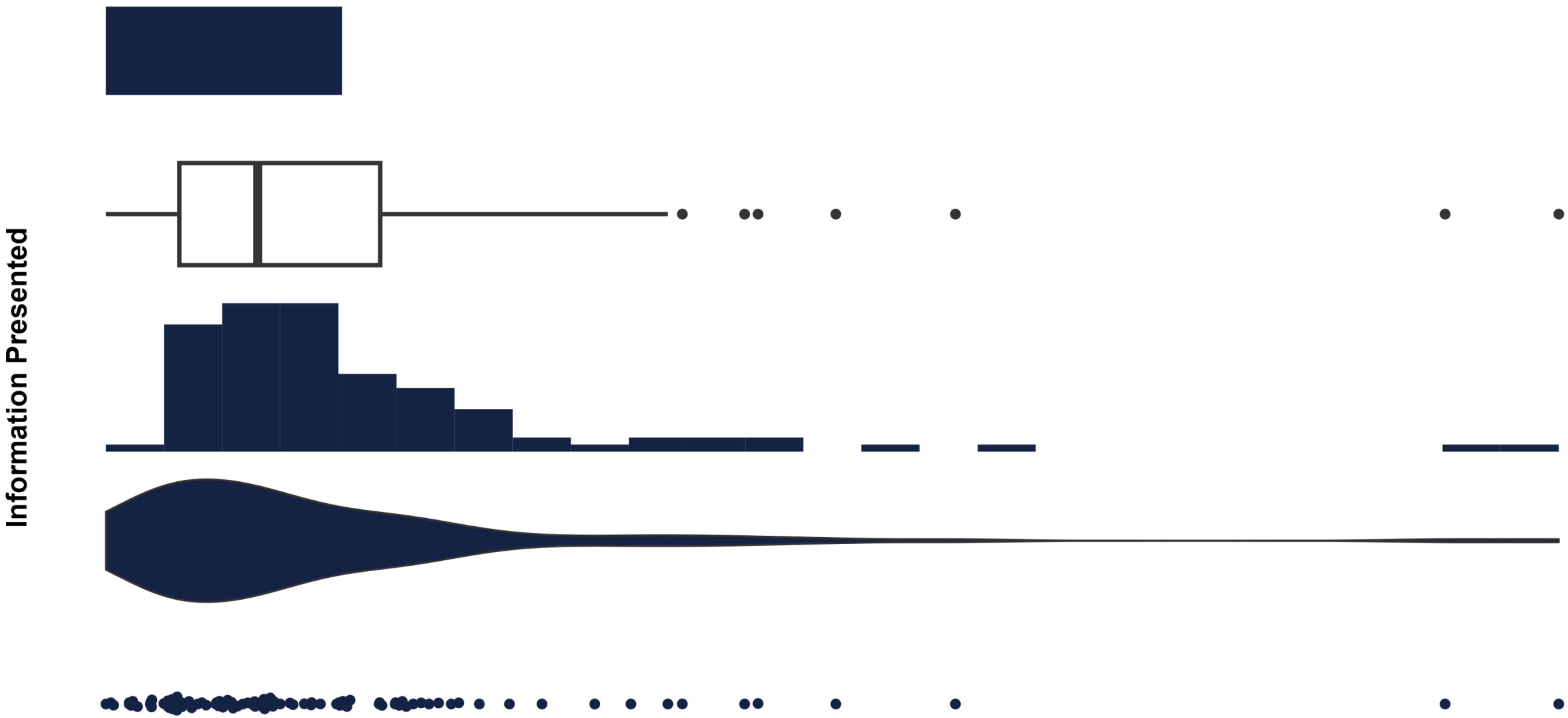


Claim Size by Product



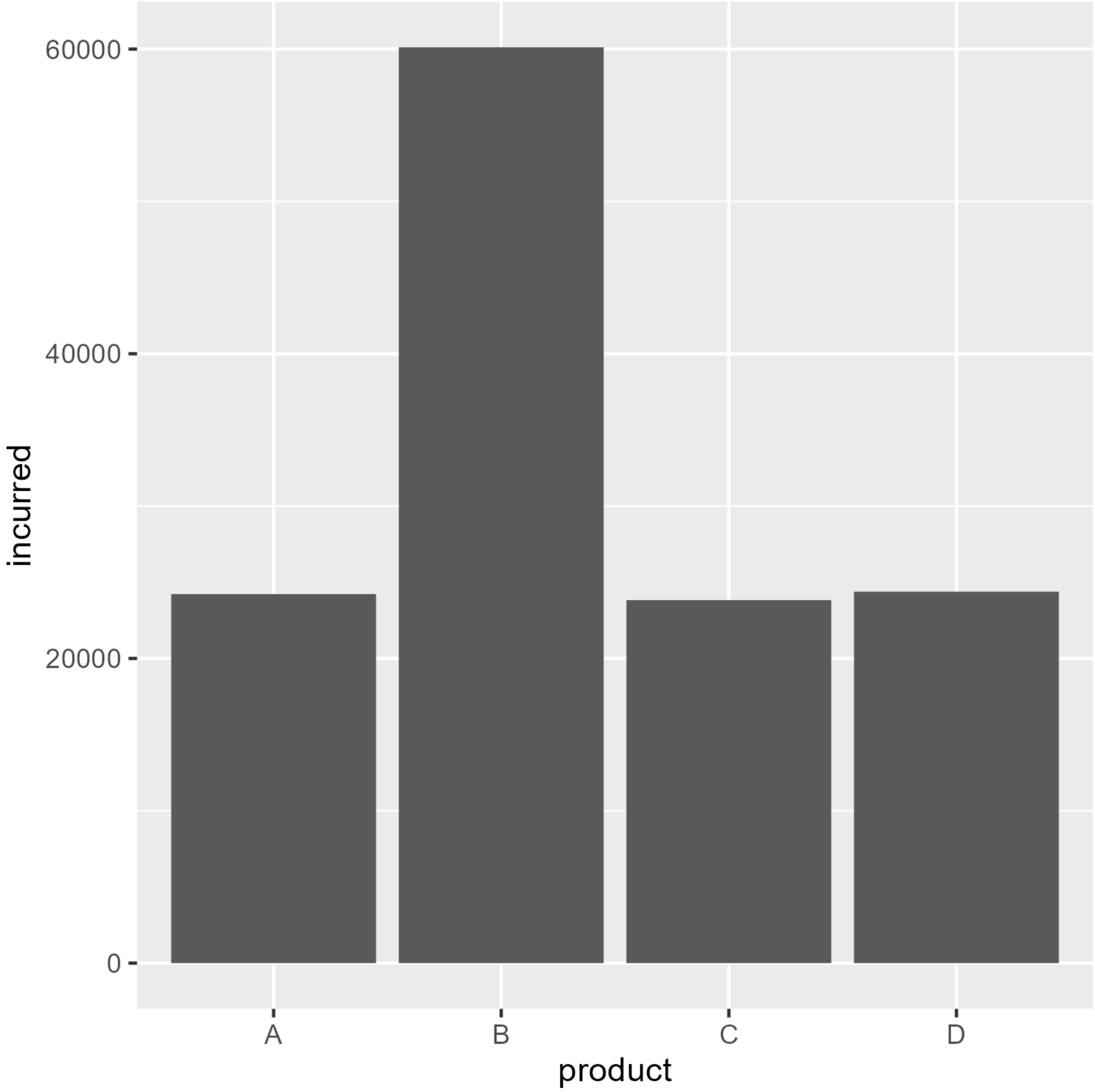
Claim Size by Product


Claim Size by Product
Product
A
B
C
D
0
50,000
100,000
150,000
Claim Size

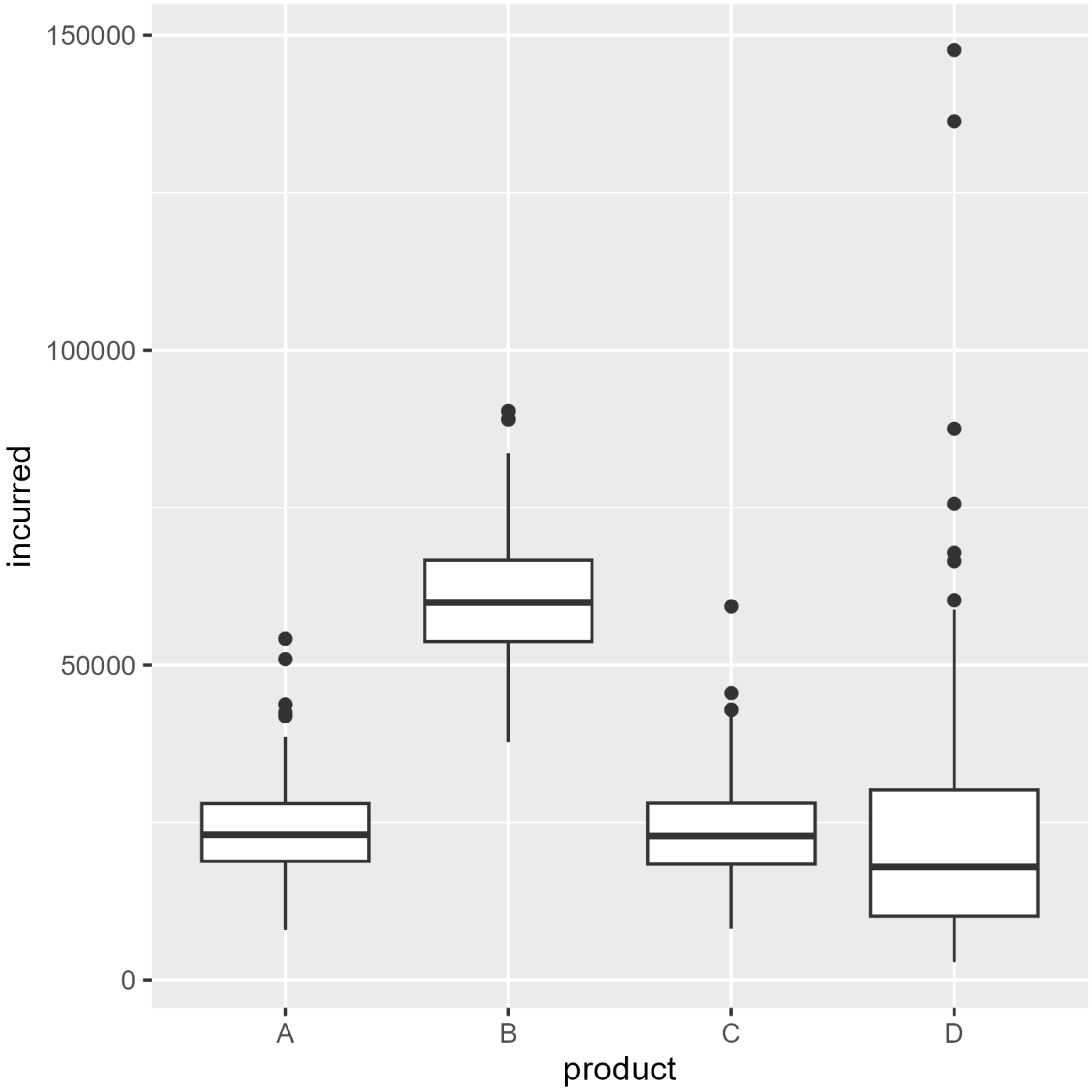


Consider relevance.

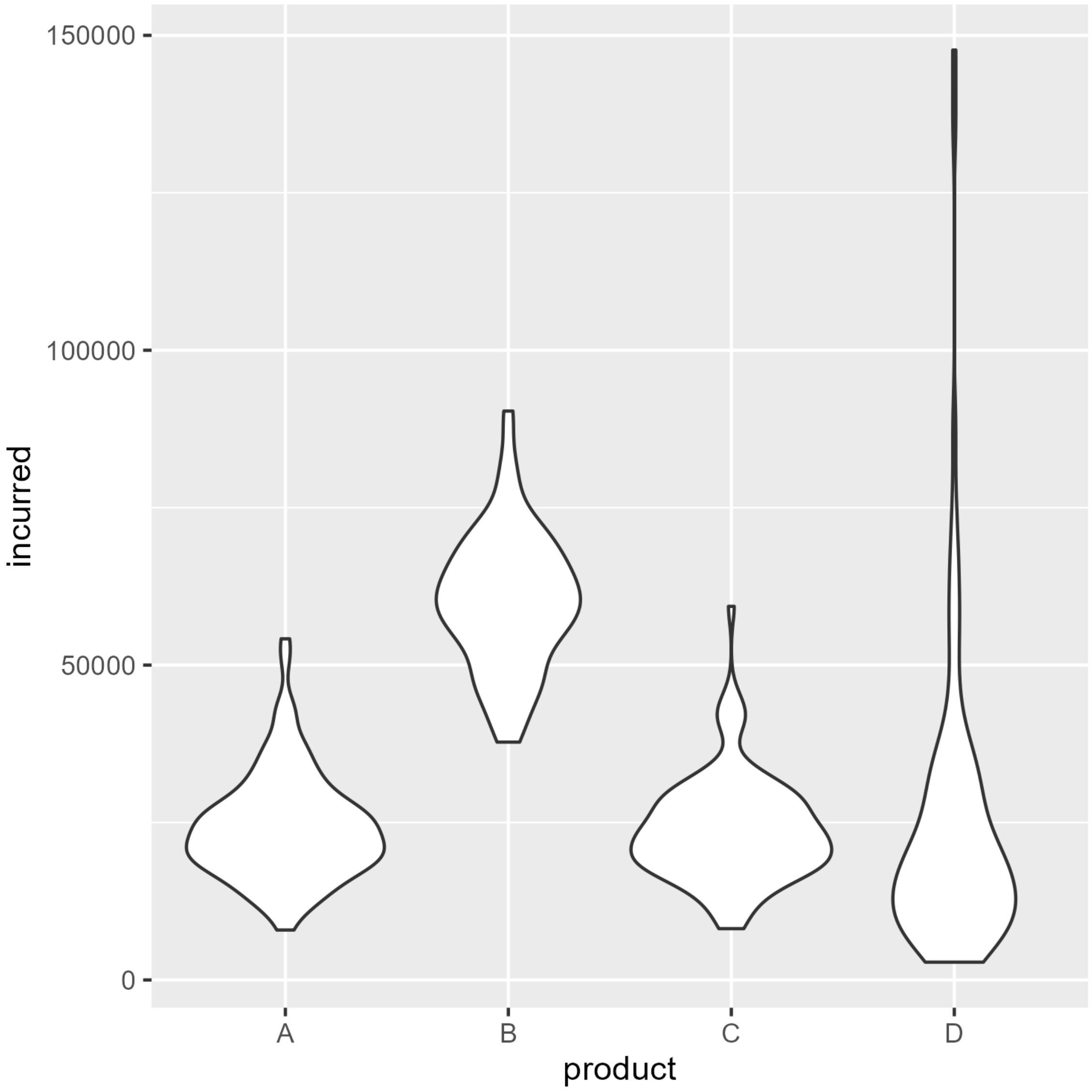
```
tbl_claims |>
  ggplot(aes(product, incurred)) +
  geom_bar(stat = "summary",
           fun = "mean")
```



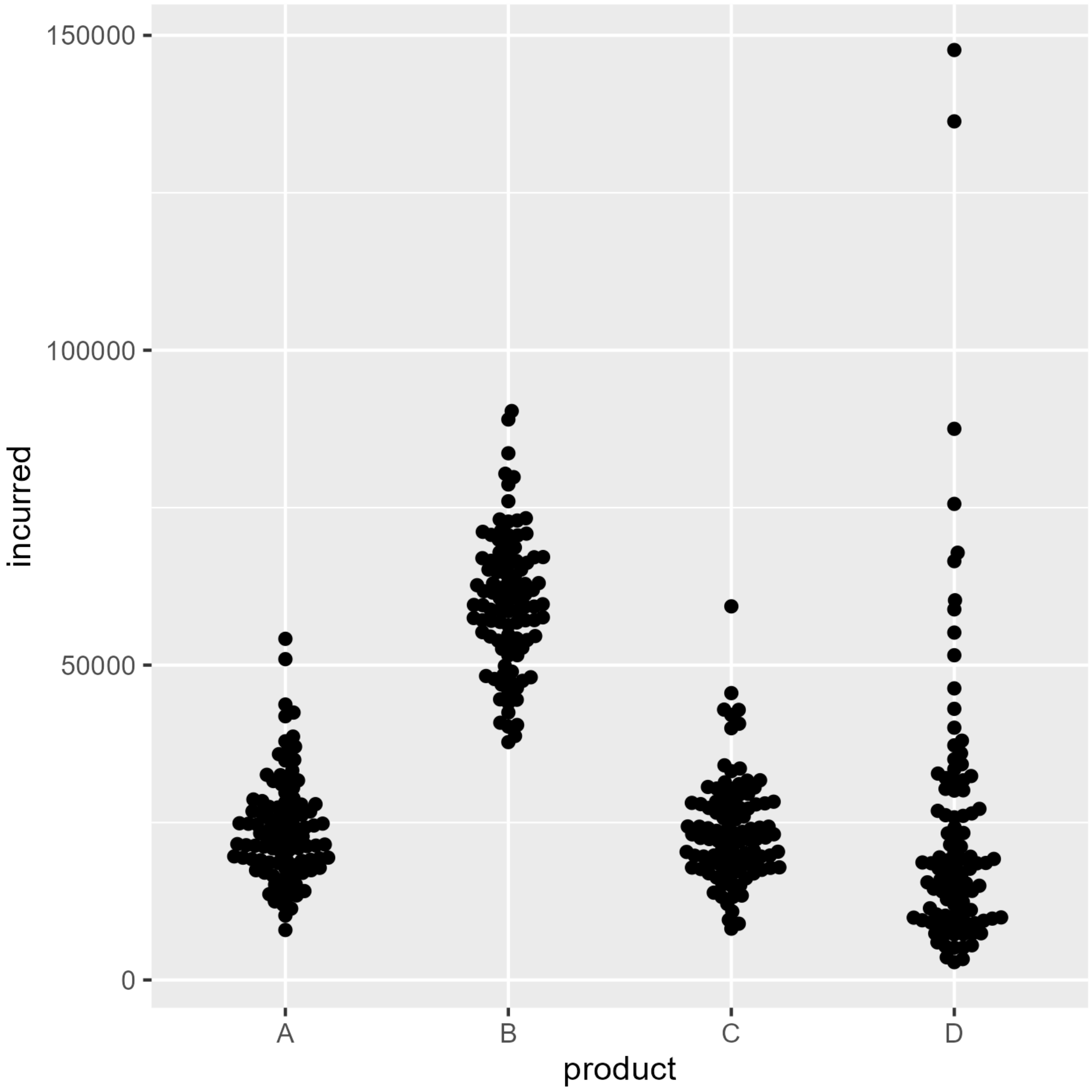
```
tbl_claims |>
  ggplot(aes(product, incurred)) +
  geom_boxplot()
```

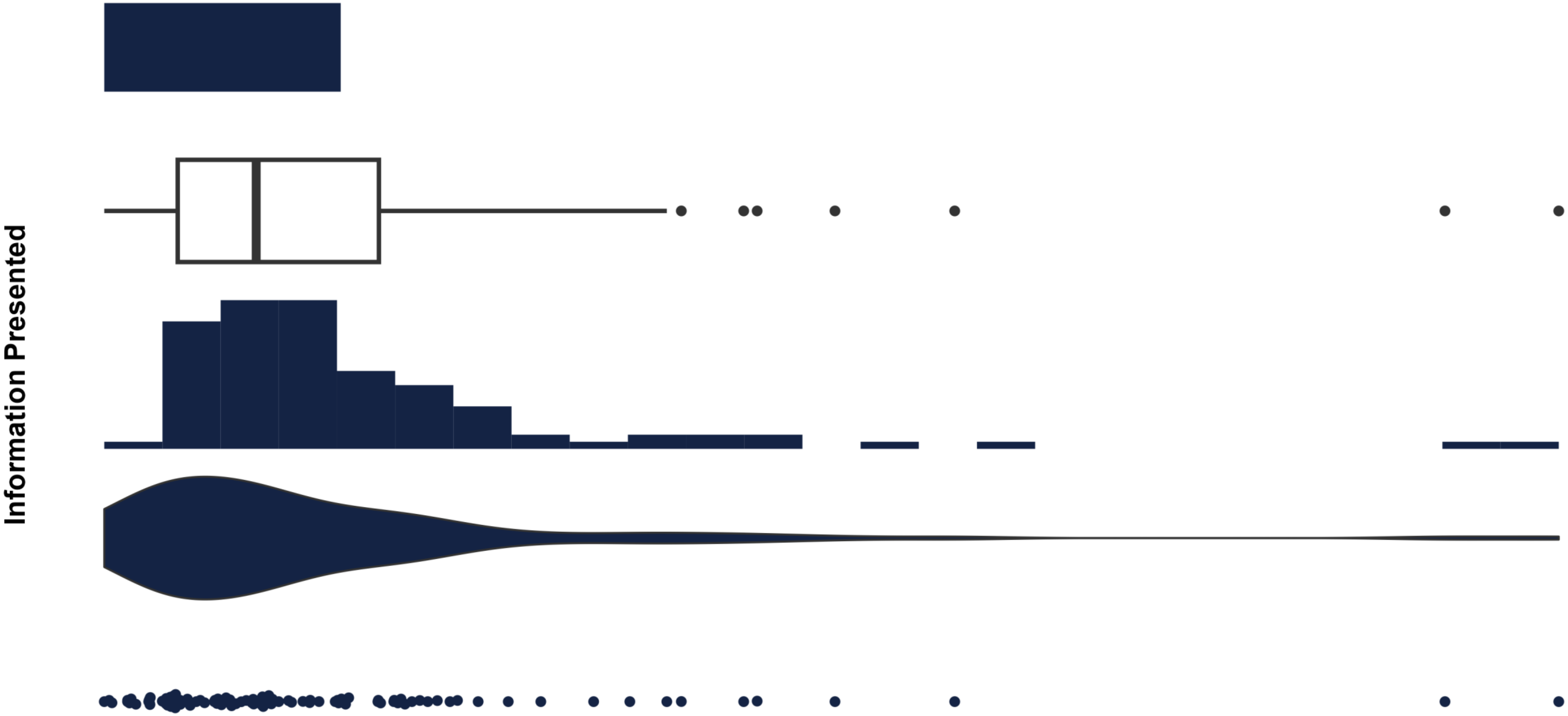


```
tbl_claims |>
  ggplot(aes(product, incurred)) +
  geom_violin()
```



```
library(ggbeeswarm)
tbl_claims |>
  ggplot(aes(product, incurred)) +
  geom_beeswarm()
```

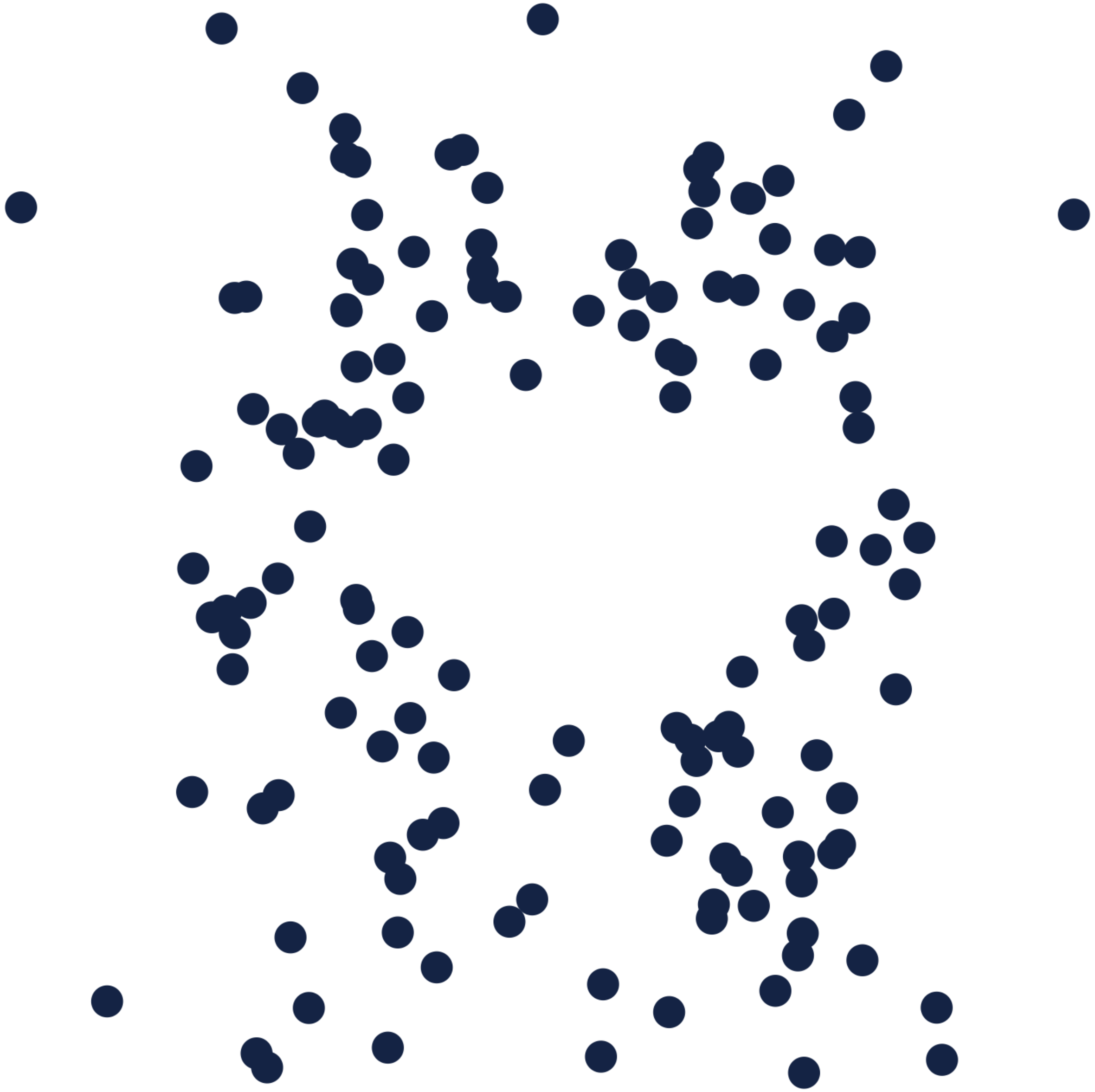


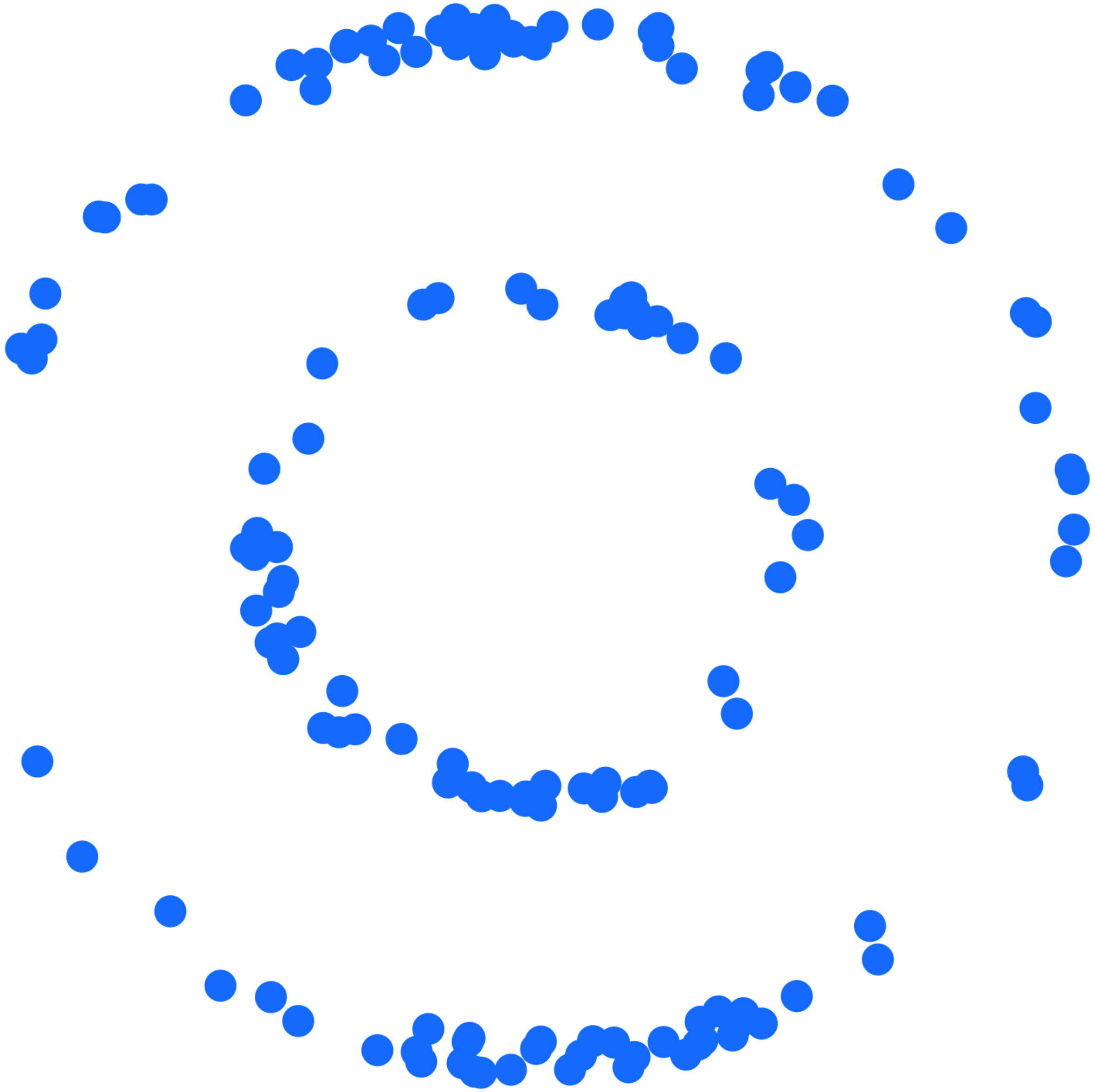


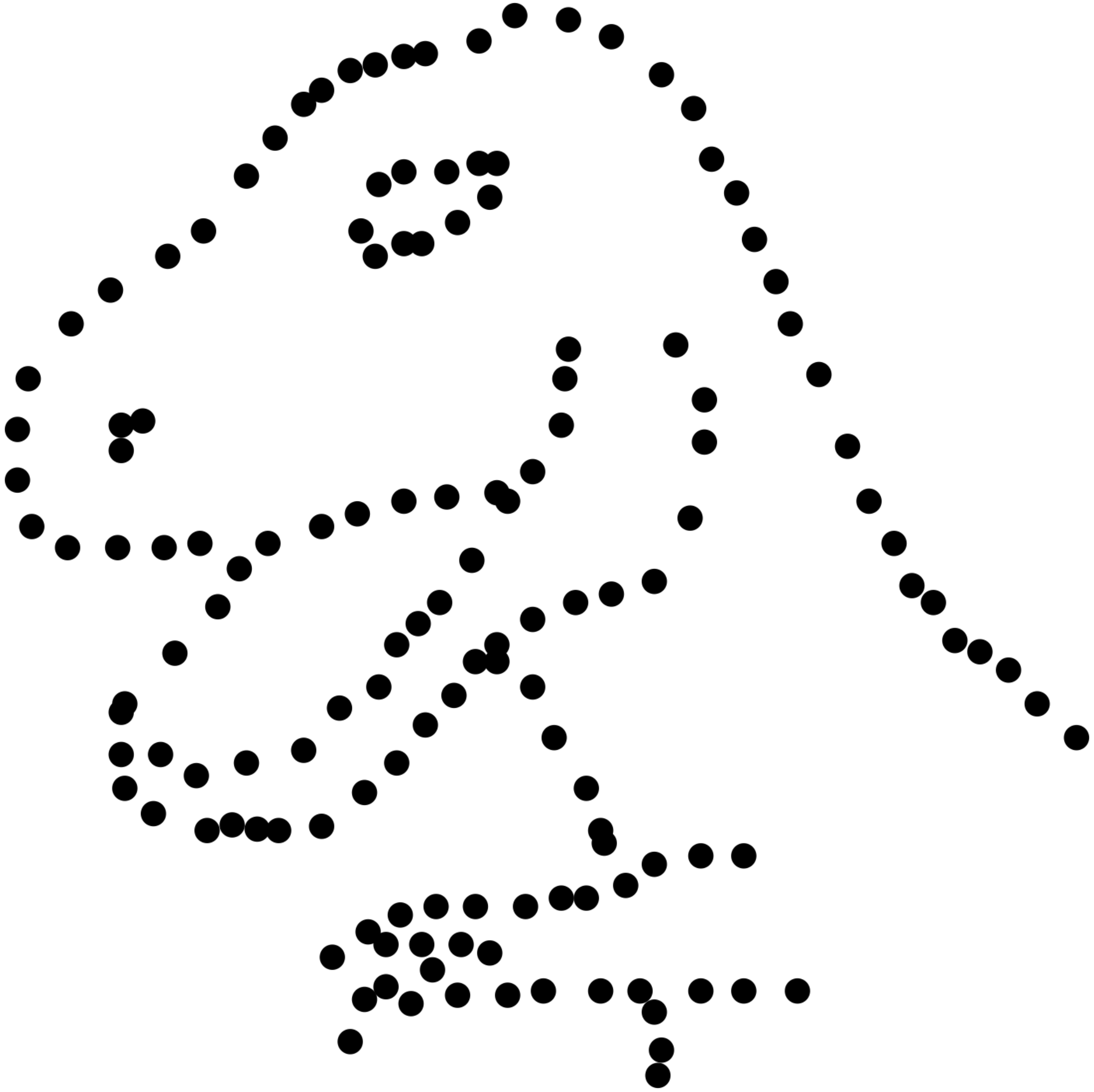
Consider relevance.

Dataset	Mean of x	Mean of y	Std. Dev. of x	Std. Dev. of y	Correlation of x, y
A	54.26	47.83	16.76	26.93	-0.0641
B	54.26	47.83	16.76	26.93	-0.0641
C	54.26	47.83	16.76	26.93	-0.0641
D	54.26	47.83	16.76	26.93	-0.0641

Matejka, J., & Fitzmaurice , G. (2017). Same Stats, Different Graphs… *Autodesk Research*.







Wrapping up

- Visualization is useful for all of us
- Clarity matters
- Keep learning and experimenting!

Thank you!

Any questions?