At this stage of our project we will try to determine the linear regression & significance between different variables in every of our scenarios. Also we will try to represent our data combined with the interval of confidence in a single plot.

At first, we have tried to get results for the scenario of:

2011: Births Ratio ~ New House Starts Ratio (aggregated quartiles, divided with population 2011)

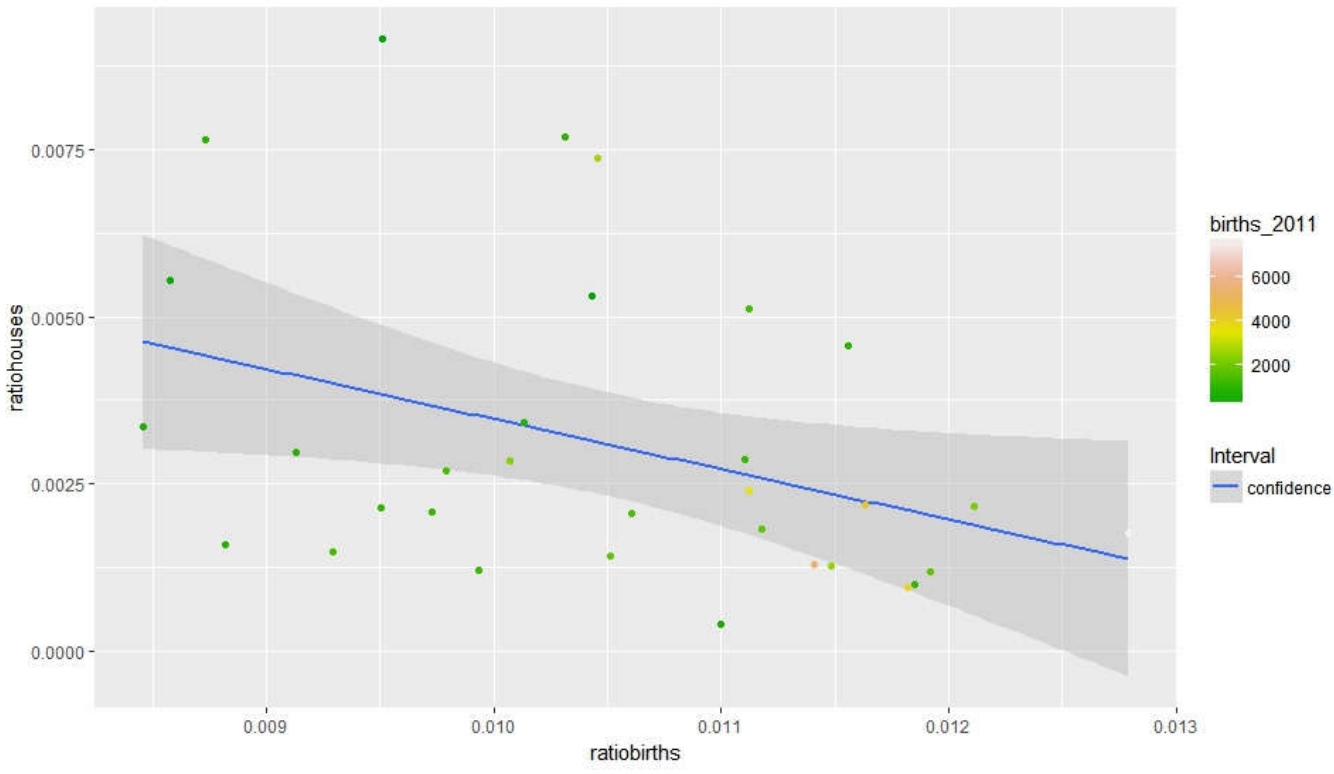
However, the significance seems to be low.

```
Call:
lm(formula = ratiobirths ~ ratiohouses, data = test)

Residuals:
Min         1Q     Median         3Q        Max
-0.0019917 -0.0009174  0.0002464  0.0007987  0.0020400

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.0110886  0.0003248  34.136   <2e-16 ***
ratiohouses -0.1891242  0.0850545  -2.224   0.0339 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.001075 on 30 degrees of freedom
Multiple R-squared:  0.1415,       Adjusted R-squared:  0.1129
F-statistic: 4.944 on 1 and 30 DF,  p-value: 0.03385
```



Also, the same variables for 2013, seem to have also low significance.

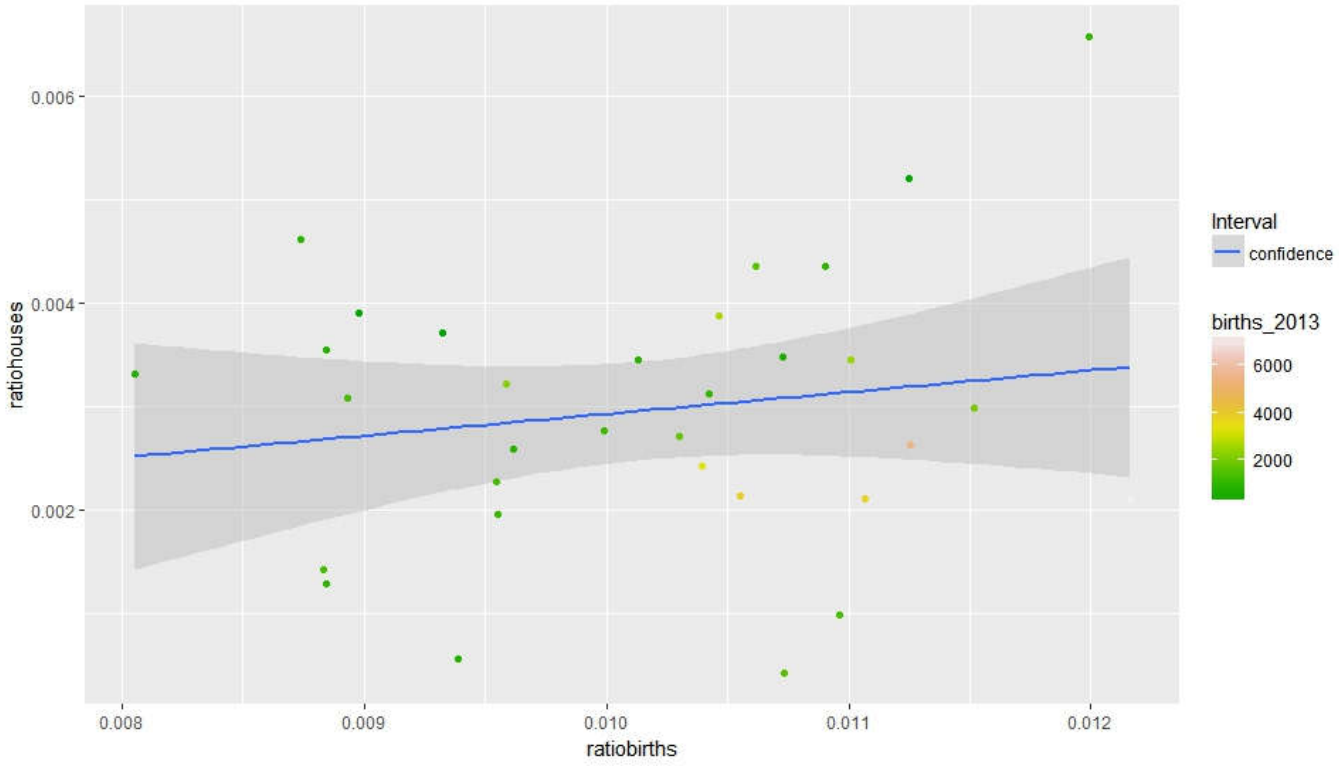
2013: Births Ratio ~ New House Starts Ratio (aggregated quartiles, divided with population of 2013)

```
Call:
lm(formula = ratiobirths ~ ratiohouses, data = test)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0021394 -0.0006743  0.0001916  0.0008032  0.0021240

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.0097650  0.0004554  21.443   <2e-16 ***
ratiohouses 0.1286290  0.1408104   0.913    0.368  null
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.001033 on 30 degrees of freedom
Multiple R-squared:  0.02706,   Adjusted R-squared:  -0.005369
F-statistic: 0.8345 on 1 and 30 DF,  p-value: 0.3683
```



But why there is a serious difference in the slope of the two linear models between years 2011, 2013?

- From our data, we can see that there was a rise in house building starts in 2013 compared to 2011. This might be the reason that the slope has become positive.

2013

ratiohouses	ratiobirths	pop2011	ratiobirths2011
0.0034614876	0.011009821	222460	0.011723456
0.0038871863	0.010462816	253650	0.010624877
0.0019692149	0.009553702	116200	0.009939759
0.0033162976	0.008052243	88930	0.008377375
0.0026381652	0.011254257	477940	0.011633259
0.0034906396	0.010725429	51500	0.010951456

2011

ratiohouses2011
0.0012991100
0.0074906367
0.0012220310
0.0039284606
0.0013181571
0.0004077670

Regarding our second scenario, the results are presented below:

2011: Cannabis new clients presenting at specialist drug treatment services ratio ~Number of degree holders ratio (divided with population of 2011)

For our bad luck, slope has low significance

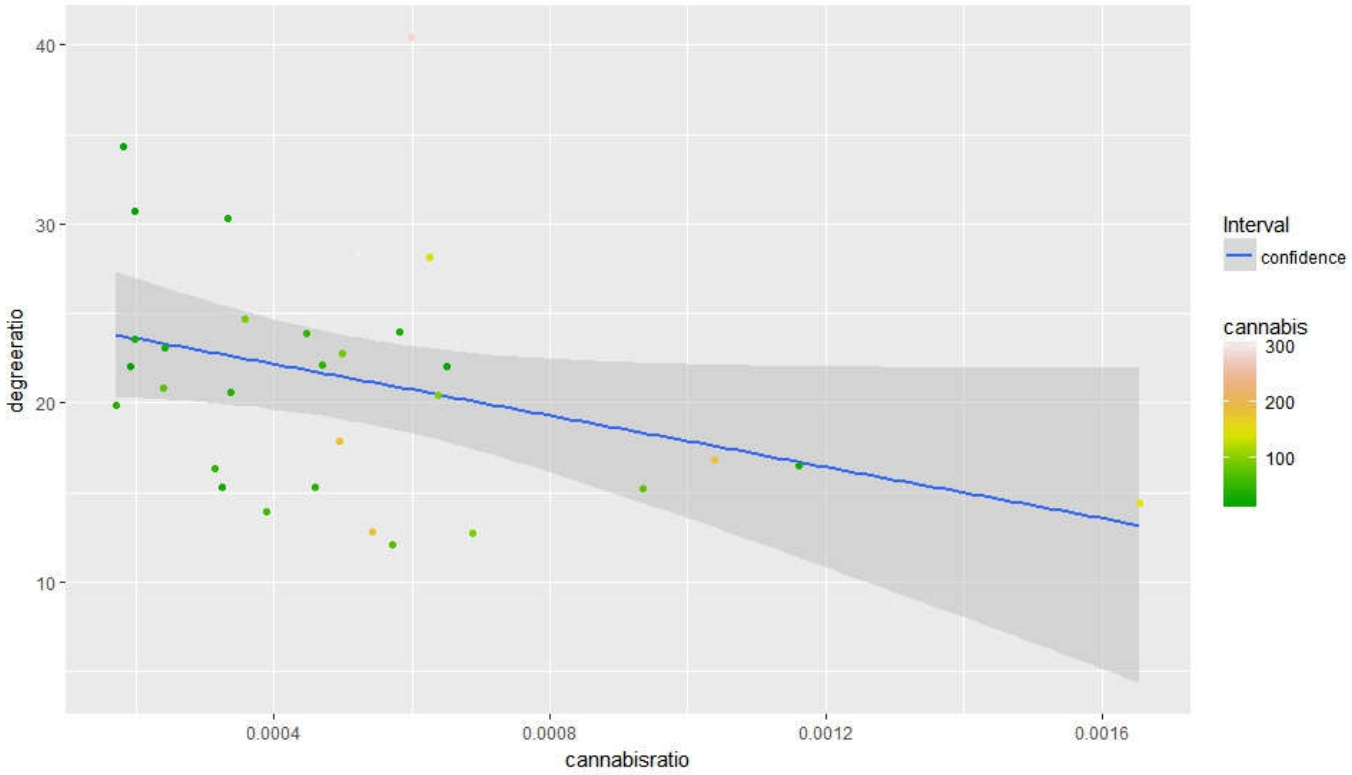
```
Call:
lm(formula = cannabisratio ~ degreeratio, data = test)

Residuals:
      Min        1Q    Median        3Q       Max
-0.036962 -0.022061 -0.007906  0.011441  0.102306

Coefficients:
             Estimate Std. Error t value  Pr(>|t|)
(Intercept)  0.0870389  0.0186883   4.657 0.0000657 ***
degreeratio -0.0016528  0.0008366  -1.976    0.0578 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.03089 on 29 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.1186,    Adjusted R-squared:  0.08824
F-statistic: 3.903 on 1 and 29 DF,  p-value: 0.05778
```

The plot of cannabis client ratio – Number of degree holders ratio.



To sum up, we realize that both of our previous scenarios might have been good ideas to examined, but in the end, **they didn't have a serious significance in their linear models.**

For this reason we had to find this kind of scenario, which it could provide us two variables that can have a serious significance in a linear model.

Finally, we have decided to use as dependent variable the "Total Labour Costs per Head" , as we believe that is a serious indicator which can have an effect on many other variables.

We have tried several combinations with other variables, in order to find strong significance.

Below you can find a list of them*:

Total Labour Costs per Head with:	**Degree Holders ratio**
	Employment/Unemployment ratio
	Total Earnings median
	Smokers percent
	Pay Gap Ratio

In the annex of this document you could find more analysis on each variable

After several trials we have achieved a high significance with the variable:

Business Births per year (confirmed with VAT & PAYE registrations)

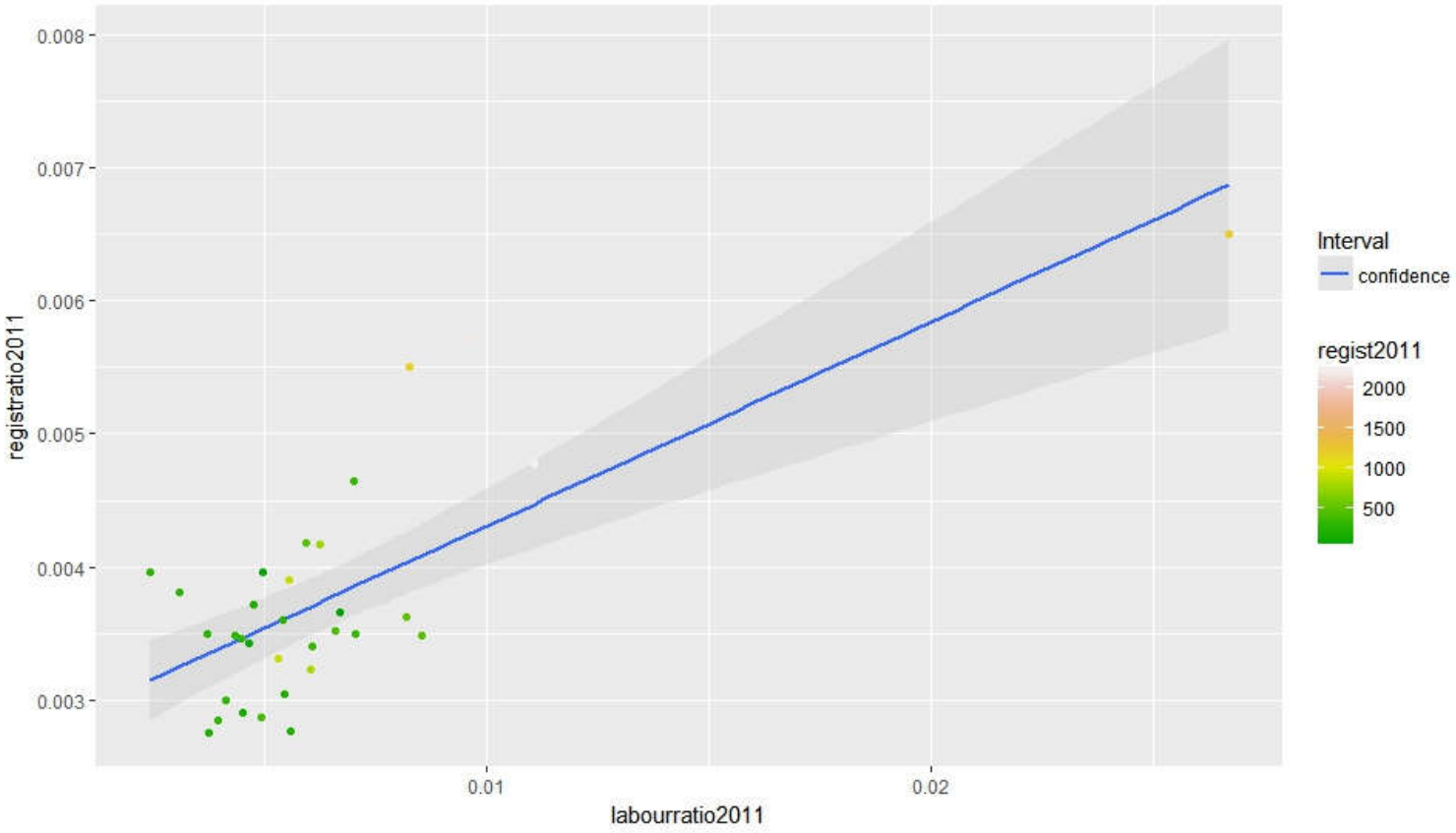
<u>2011</u>: Total Labour Costs per Head ~ Business Births (divided with population of 2011)

```
Call:
lm(formula = labourratio2011 ~ registratio2011, data = test)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0047121 -0.0016446  0.0004222  0.0012095  0.0106267

Coefficients:
                 Estimate Std. Error t value   Pr(>|t|)
(Intercept)     -0.006862   0.002288  -2.998    0.00541 **
registratio2011  3.529811   0.593758   5.945 0.00000163 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.002858 on 30 degrees of freedom
Multiple R-squared:  0.5409,   Adjusted R-squared:  0.5256
F-statistic: 35.34 on 1 and 30 DF,  p-value: 0.000001627
```



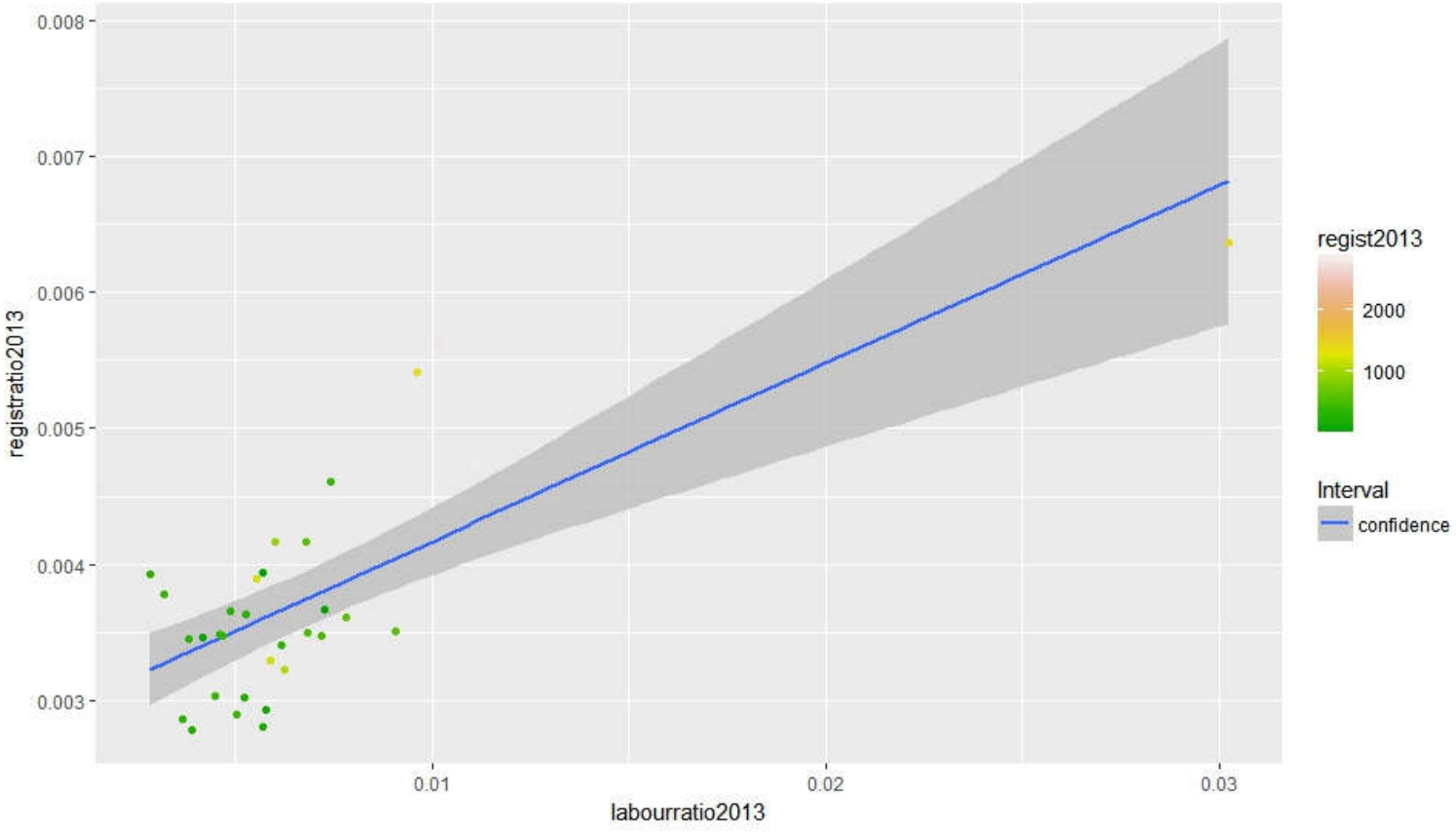
2013: Total Labour Costs per Head ~ Business Births (divided with population of 2013)

```
Call:
lm(formula = labourratio2013 ~ registratio2013, data = test)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0047400 -0.0017672  0.0003785  0.0015150  0.0124289

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -0.009070   0.002657  -3.414  0.00186 **
registratio2013  4.226564   0.693581   6.094 1.07e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.00321 on 30 degrees of freedom
Multiple R-squared:  0.5531,   Adjusted R-squared:  0.5382
F-statistic: 37.13 on 1 and 30 DF,  p-value: 1.073e-06
```



Findings:

From the final two linear models we can assume that the labour costs, on different territories of Scotland, are related positively with the number of new companies.

Annex: Total Labour Costs per Head trials with…

---

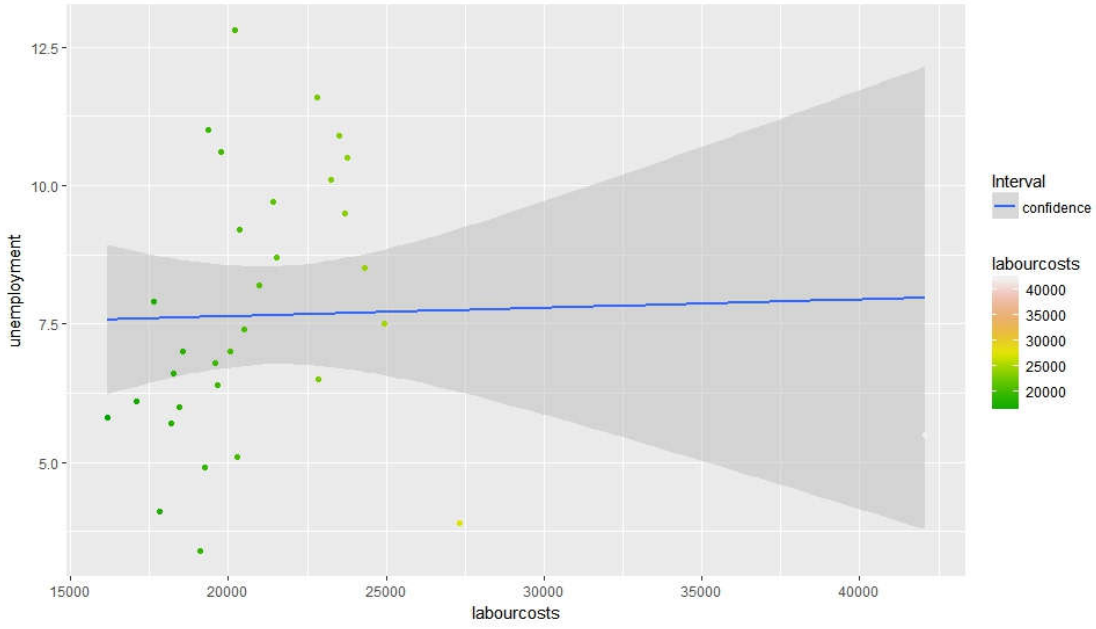
1) Unemployment

```
Call:
lm(formula = labourcosts ~ unemployment, data = test)

Residuals:
   Min     1Q Median     3Q    Max
 -5064  -2311  -1149   1614  20846

Coefficients:
             Estimate Std. Error t value     Pr(>|t|)
(Intercept)  20918.49    2765.81   7.563 0.0000000196 ***
unemployment    54.42     345.17   0.158        0.876
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4636 on 30 degrees of freedom
Multiple R-squared:  0.0008277, Adjusted R-squared:  -0.03248
F-statistic: 0.02485 on 1 and 30 DF,  p-value: 0.8758
```



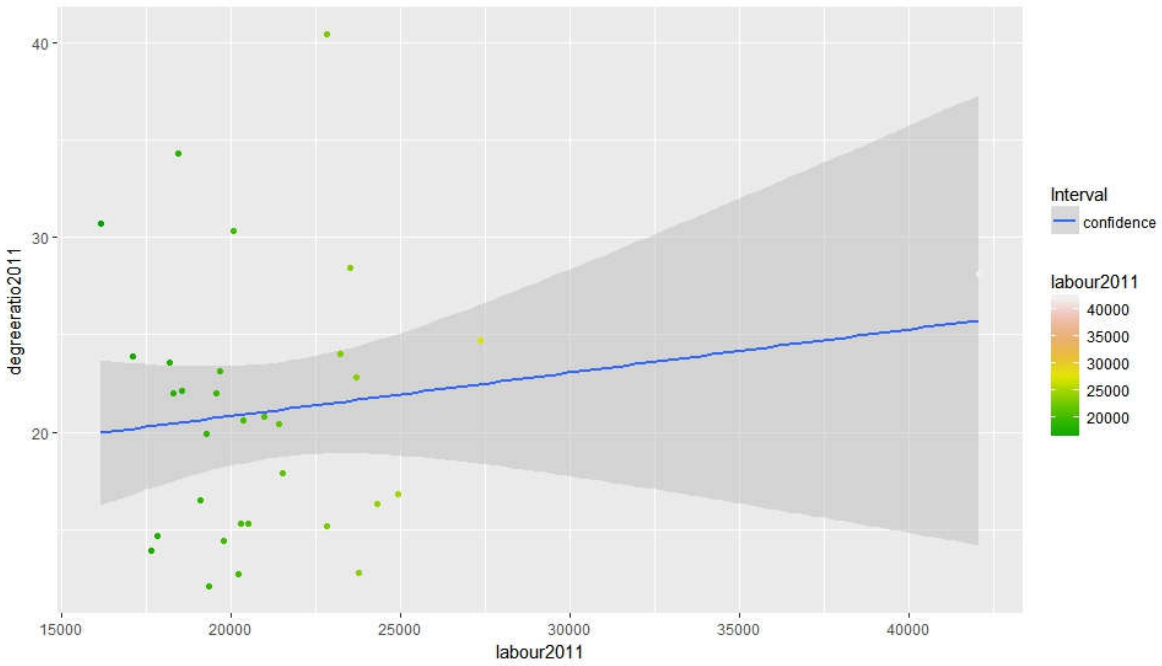
2) Degree Holders Ratio

```
Call:
lm(formula = labour2011 ~ degreeratio2011, data = test)

Residuals:
    Min      1Q  Median      3Q     Max
-6143.2 -2362.4  -669.1  1486.4 20016.4

Coefficients:
                Estimate Std. Error t value     Pr(>|t|)
(Intercept)      19176.6     2706.8   7.085 0.0000000706 ***
degreeratio2011    102.2      122.3   0.836         0.41
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4584 on 30 degrees of freedom
Multiple R-squared:  0.02275,    Adjusted R-squared:  -0.009824
F-statistic: 0.6984 on 1 and 30 DF,  p-value: 0.4099
```



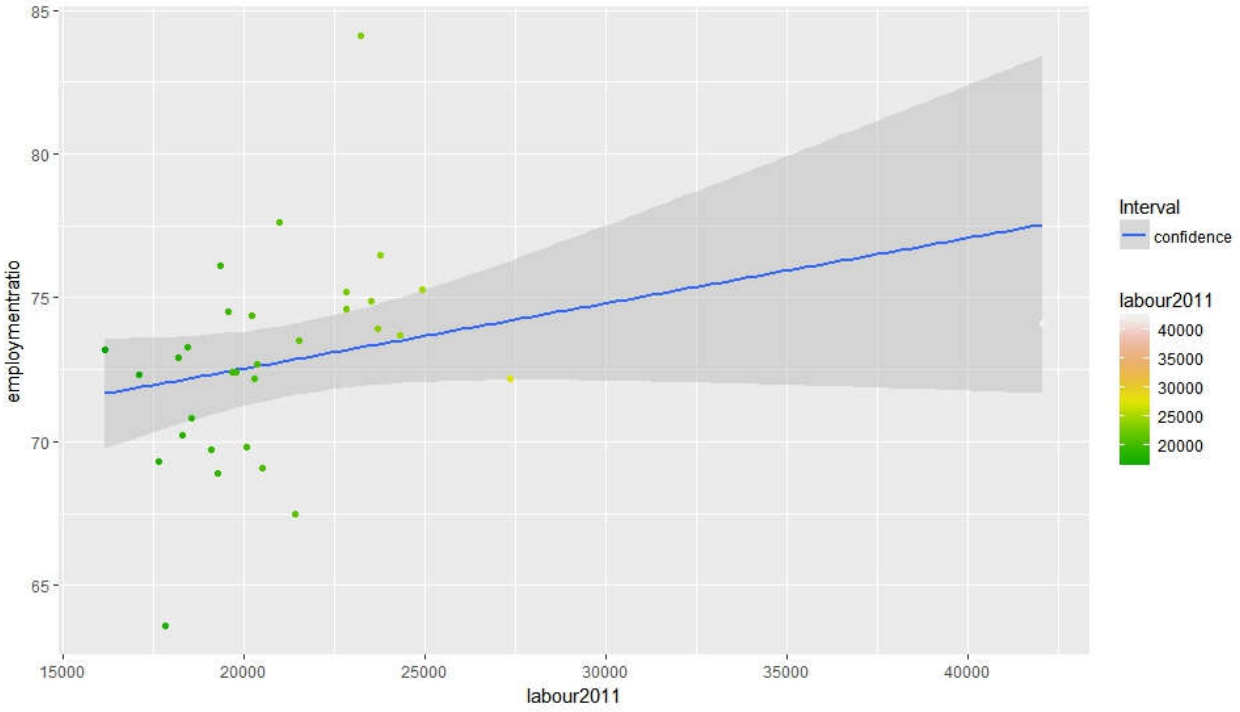
3) Employment Ratio

```
Call:
lm(formula = labour2011 ~ employmentratio, data = test)

Residuals:
    Min      1Q  Median      3Q     Max
-5300.1 -2199.8  -871.5   877.6 20255.3

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      -6065.2    16386.5  -0.370    0.714
employmentratio    376.2      224.7   1.674    0.105

Residual standard error: 4435 on 30 degrees of freedom
Multiple R-squared:  0.08543,    Adjusted R-squared:  0.05495
F-statistic: 2.802 on 1 and 30 DF,  p-value: 0.1045
```



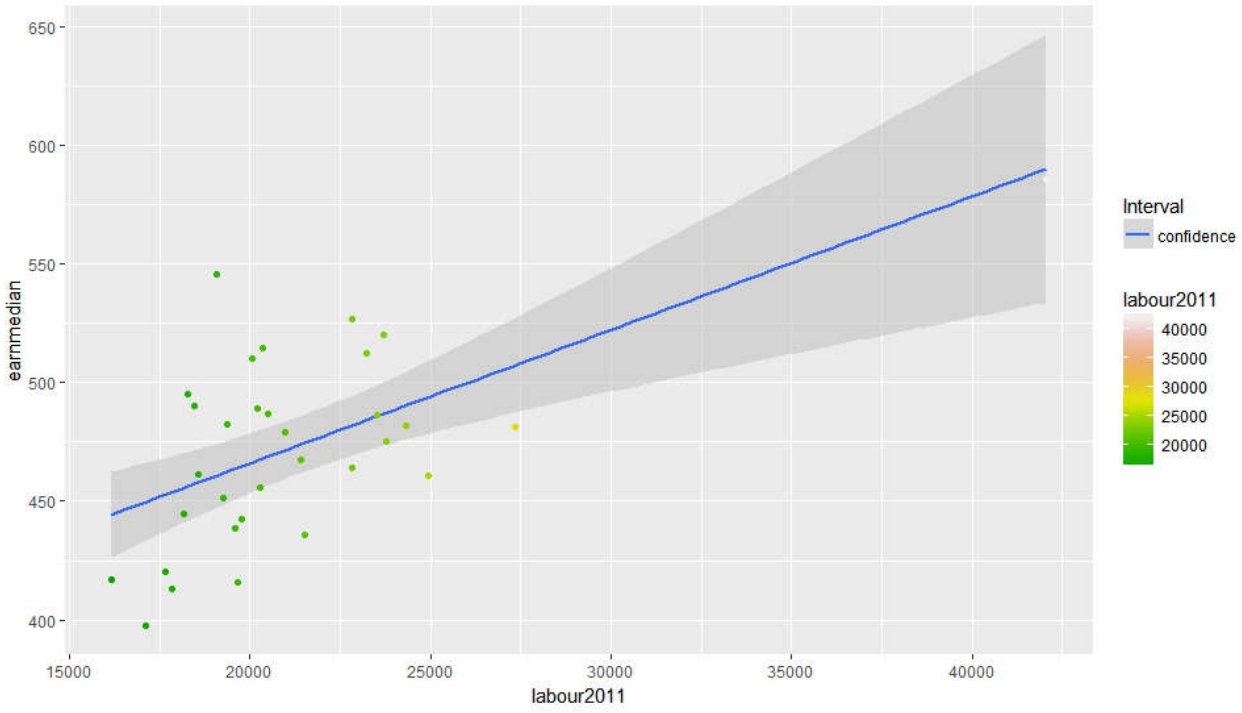
4) Total Earning Ratio

```
Call:
lm(formula = labour2011 ~ earnmedian, data = test)

Residuals:
    Min      1Q  Median      3Q     Max
-7158.8 -2003.9  -316.6  1506.5 13081.9

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -10924.82    7492.93  -1.458 0.155226
earnmedian      68.14      15.77   4.321 0.000157 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3641 on 30 degrees of freedom
Multiple R-squared: 0.3837 , Adjusted R-squared:  0.3631
F-statistic: 18.67 on 1 and 30 DF,  p-value: 0.0001568
```



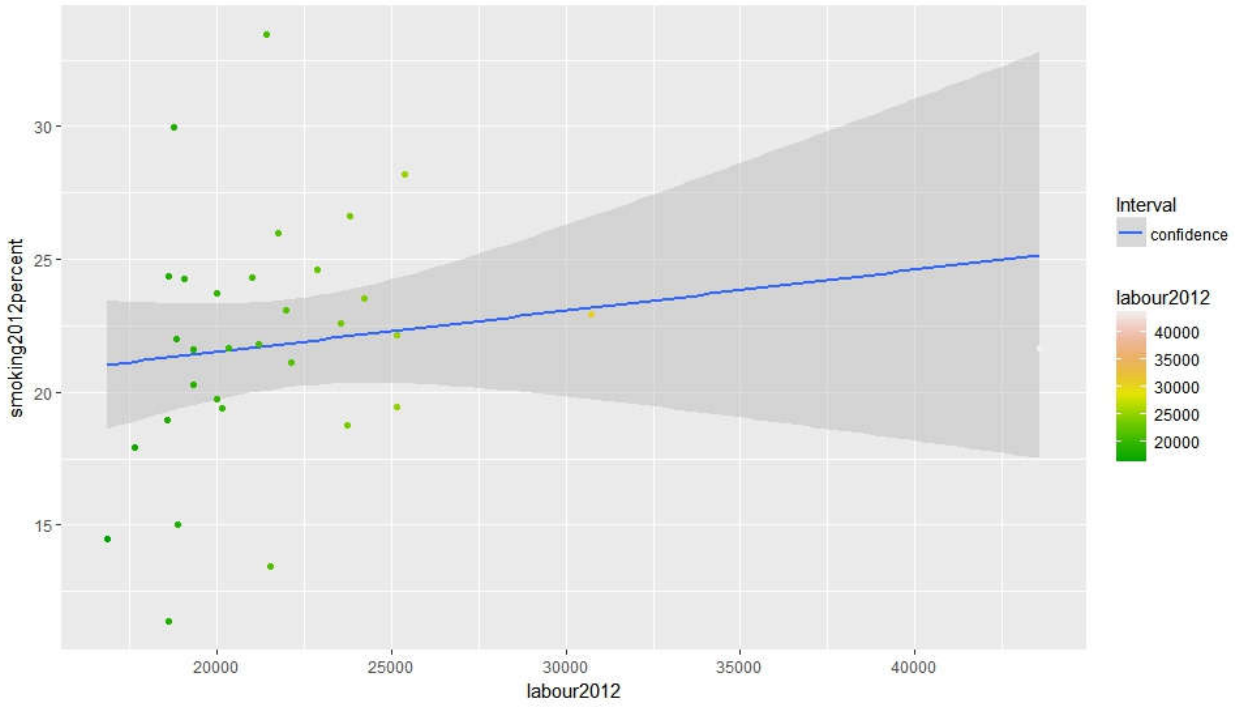
5) **Smokers percent**

```
Call:
lm(formula = labour2012 ~ smoking2012percent, data = test)

Residuals:
   Min     1Q Median     3Q    Max
 -4680  -2624  -1441   1093  21617

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         18174.5     4276.3   4.250 0.000191 ***
smoking2012percent    175.2      191.8   0.913 0.368263
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4893 on 30 degrees of freedom
Multiple R-squared:  0.02706,    Adjusted R-squared:  -0.005368
F-statistic: 0.8345 on 1 and 30 DF,  p-value: 0.3683
```



6) Pay Gap Ratio

```
Call:
lm(formula = labour2011 ~ paygap2010ratio, data = test)

Residuals:
   Min     1Q Median     3Q    Max
 -4702  -2734  -1317   2095  19035

Coefficients:
                Estimate Std. Error t value          Pr(>|t|)
(Intercept)      20346.7     1334.1  15.251 0.000000000000036
***
paygap2010ratio    176.5      125.2   1.409             0.171
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4676 on 25 degrees of freedom
  (5 observations deleted due to missingness)
Multiple R-squared:  0.07358,    Adjusted R-squared:  0.03652
F-statistic: 1.985 on 1 and 25 DF,  p-value: 0.1711
```