How Linear Regression with Interaction Helps Optimize Cosmetic Formulations

Juan David Marin

2025-03-05

Recently, I faced a challenge in the company I worked for: we needed to better understand the products we were manufacturing and formulating. Due to a lack of understanding, we encountered some formulation problems during manufacturing. One of them was that we were adding more raw materials than the formulation required. When I performed chemical analysis in the laboratory, I noticed a great opportunity to optimize both the process and the formulation based on chemical data.

It turned out that some formulations had a higher sun protection factor (SPF) than labeled because certain raw materials used in them enhanced the SPF claim. The physicochemical variables I considered were:

- Octinoxate and Zinc oxide: These are UV filters, the main raw materials responsible for the SPF value in the foundation formulation.
- Spf_in_vitro: This is the sun protection value of the formulations, depending on the amount of UV filters.
- Emulsifier (categorical: Natural, Synthetic), emulsifier_type (categorical: ionic, non-ionic), base_type (categorical: oil, water): These describe the formulation characteristics.
- Viscosity (continuous numerical): The viscosity of the makeup base or formulation.
- CSAT (categorical: dislike, like, so-so): Consumer satisfaction.
- Formula_price (continuous numerical): The cost of the formulation.

(The data was simulated, but it presents a good example of what I worked on.)

The questions of the investigation were:

1. *How is the relationship between UV filters and FPS in vitro:*

- How is the relationship of octinoxate and Zinc oxide against in vitro SPF?
- Is there any interaction between octinoxate and Zinc oxide that impacts the in vitro SPF?

2. *What is the impact of categorical variables with the in vitro SPF:*

- How do the categorical variables (emulsifier, emulsifier_type, base_type) influence the in vitro SPF?
- Is there an interaction between categorical variables and the UV filter octinoxate that affects the in vitro SPF?

3. *Which is the viscosity effect in the formulation of makeup foundations:*

- How does viscosity affect the in vitro SPF and customer satisfaction?
- Is there any interaction between viscosity and the rest of the variables that affect the in vitro SPF and customer satisfaction?

4. *Customer satisfaction:*

- How is the interaction between variables (octinoxate, viscosity, emulsifier, emulsifier_type, base_type) with customer satisfaction?
- Are there interactions among these variables that affect customer satisfaction?

5. *Cost formulation relationship:*

- How are the formulation variables related (octinoxate, viscosity, emulsifier, emulsifier_type, base_type) with customer satisfaction?
- Are there interactions among these variables that affect cost formulation?

6. *Formulation optimization:*

- Is it possible to reduce the UV filter amounts (octinoxate, Zinc oxide) while maintaining the in vitro SPF labeled, if we optimize the other variables?
- What is the optimal combination among variables that enhances the in vitro SPF and customer satisfaction while minimizing cost formulation?

Those were the questions. Maybe there could be more, but these helped me to understand the makeup foundation formulation and to solve the formulation overruns.

It is necessary to mention again that the data I am going to analyze were simulated.

This is what the data looks like:

```
foundation <- read.csv('SPF_formulaII.csv')
head(foundation)
```

```
  octinoxate zinc_oxide eficacia_octinoxate_hplc eficacia_zno_aa spf_in_vitro
1   5.527050   1.147532                 95.52040        97.92851    10.669123
2   2.859906   7.490659                 95.61372        84.46541    10.789527
3   2.474534   1.069638                 99.39909        85.36249     9.670812
4   4.583546   1.763400                 98.15057        83.88996    10.126165
5   5.676548   3.029486                 92.02090        99.35002    11.339992
6   3.750192   8.876121                 94.38860        82.25080    11.537090
  emulsifier emulsifier_type base_type viscosidad    CSAT formula_price
1    Natural           ionic       Oil   630.9090 dislike         2.743
2    Natural           ionic     Water   178.4680 dislike         3.090
3    Natural       non-ionic     Water   933.0352 dislike         1.688
4    Natural           ionic       Oil   769.2506 dislike         1.260
5   Sintetic       non-ionic       Oil  4604.7955   so_so        12.946
6   Sintetic       non-ionic       Oil  1152.6147 dislike         1.552
```

CORRELATION ANALYSIS

Let's check the normality of the variables.

```r
sapply(foundation[c('octinoxate', 'zinc_oxide', 'spf_in_vitro')], shapiro.test)
```

```
          octinoxate                    zinc_oxide
statistic 0.9620628                     0.9579528
p.value   4.665932e-07                  1.325683e-07
method    "Shapiro-Wilk normality test" "Shapiro-Wilk normality test"
data.name "X[[i]]"                      "X[[i]]"
          spf_in_vitro
statistic 0.9692902
p.value   5.18963e-06
method    "Shapiro-Wilk normality test"
data.name "X[[i]]"
```

As we see, the p-values of the variables for the normality test are lower than 0.05, which means they don't follow a normal distribution. So, for that reason, I used the Spearman method to evaluate the correlation.

```r
cor.test(x = foundation$octinoxate, y = foundation$spf_in_vitro, method = 'spearman')
```

```
    Spearman's rank correlation rho

data:  foundation$octinoxate and foundation$spf_in_vitro
S = 1283430, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.7147902
```

```r
cor.test(x = foundation$zinc_oxide, y = foundation$spf_in_vitro, method = 'spearman')
```

```
    Spearman's rank correlation rho

data:  foundation$zinc_oxide and foundation$spf_in_vitro
S = 1937240, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.5694974
```

Both variables have a positive and significant correlation against in vitro SPF (p-value $< 2.2e-16$), however, Octinoxate shows a better and stronger correlation (rho $= 0.71$) than Zinc Oxide (rho $= 0.57$). It means Octinoxate has a stronger impact against in vitro SPF than Zinc Oxide.

Now, I made sure that the correlation between Octinoxate and in vitro SPF was not influenced by Zinc Oxide (Causality)

```r
pcor.test(x = foundation$octinoxate,
          y = foundation$spf_in_vitro,
          z = foundation$zinc_oxide)
```

```
  estimate      p.value statistic   n gp  Method
1 0.876622 2.329231e-96  31.39665 300  1 pearson
```

The correlation between octinoxate and in vitro SPF is not masked by zinc oxide, as it remains very strong and significant (rho = 0.88, $p < 2.33e-96$) even after controlling for zinc oxide. This means that octinoxate has a direct and independent impact on in vitro SPF.

Statistical power

```r
pwr.r.test(n = dim(foundation)[1],r = 0.7147902, sig.level = 0.05)
```

```

     approximate correlation power calculation (arctangh transformation)

              n = 300
              r = 0.7147902
      sig.level = 0.05
          power = 1
    alternative = two.sided
```

```r
pwr.r.test(n = dim(foundation)[1],r = 0.5694974, sig.level = 0.05)
```

```

     approximate correlation power calculation (arctangh transformation)

              n = 300
              r = 0.5694974
      sig.level = 0.05
          power = 1
    alternative = two.sided
```

Both correlations have a statistical power of 1. It means that the sample size of 300 is more than enough to detect these relationships.

As Octinoxate is the variable that has a stronger relationship against in vitro SPF, this variable will be used in the next analysis.

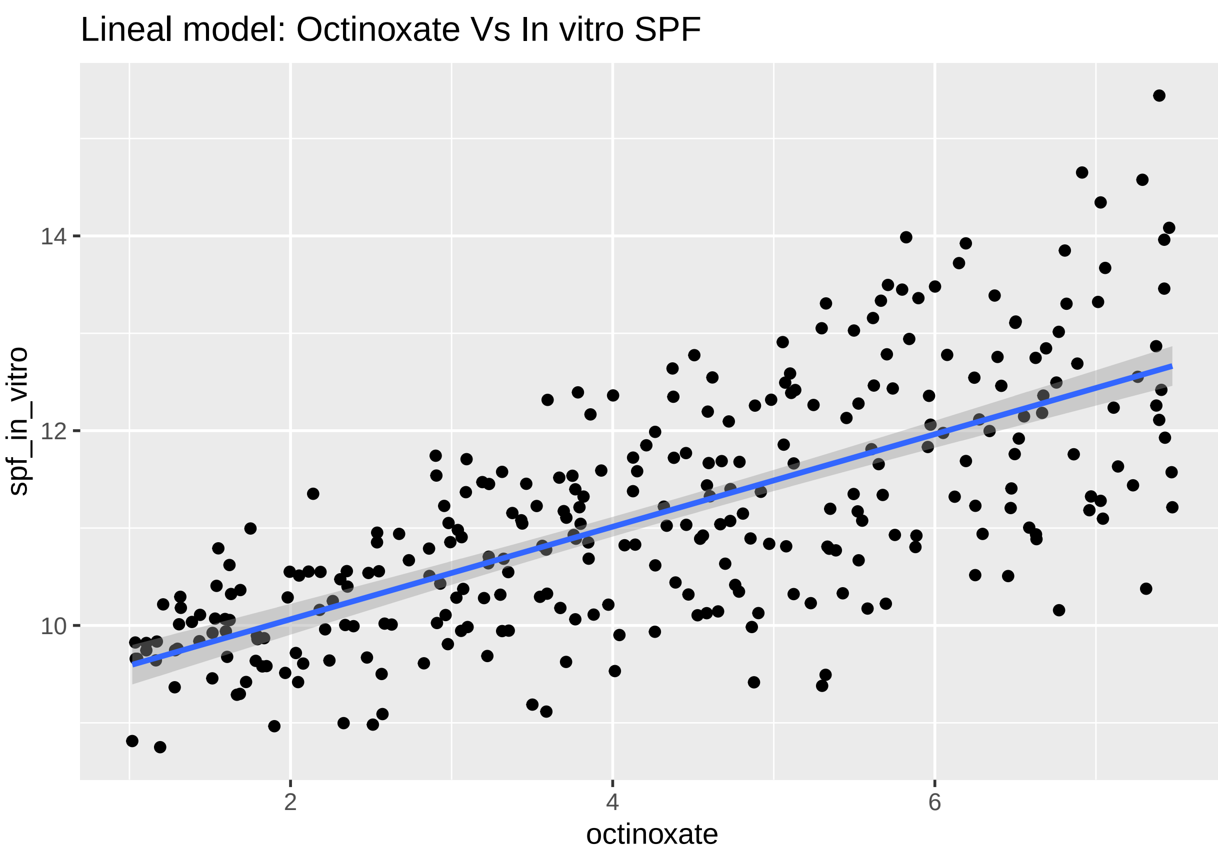
Let's perform a linear regression to figure out the interactions.

```r
octinoxate_lm <- lm(spf_in_vitro ~ octinoxate, data = foundation)
summary(octinoxate_lm)
```

```

Call:
lm(formula = spf_in_vitro ~ octinoxate, data = foundation)

Residuals:
     Min       1Q   Median       3Q      Max
-2.25132 -0.61720  0.02769  0.54339  2.81510

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.11385    0.12830   71.03   <2e-16 ***
```

```
octinoxate   0.47498    0.02778   17.10   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8817 on 298 degrees of freedom
Multiple R-squared:  0.4952,    Adjusted R-squared:  0.4935
F-statistic: 292.4 on 1 and 298 DF,  p-value: < 2.2e-16
```

```r
ggplot(data = foundation, aes(x = octinoxate, y = spf_in_vitro)) +
  geom_point()+
  geom_smooth(method="lm", formula = 'y ~ x') +
  ggtitle('Lineal model: Octinoxate Vs In vitro SPF')
```



The results of the model say that each time Octinoxate increases by 1%, the in vitro SPF increases by 0.475 units, and the relationship is pretty significant (p < 2e-16). Also, when Octinoxate has 0% concentration, the in vitro SPF in the formulas has an in vitro SPF of 9.11.

This relationship is highly significant, confirming that octinoxate is a key ingredient to achieve the desired level of sun protection in formulations.

Interactions

Let's review the possible interactions among continuous variables in order to figure out their relationships.

The code below is the linear model for the in vitro SPF according to the interaction between Octinoxate concentration and Zinc Oxide concentration.

```r
cont_cont <- lm(spf_in_vitro ~ octinoxate*zinc_oxide, data = foundation)
summary(cont_cont)
```

```
Call:
lm(formula = spf_in_vitro ~ octinoxate * zinc_oxide, data = foundation)

Residuals:
     Min       1Q   Median       3Q      Max
-0.77740 -0.22145 -0.00144  0.24563  0.90315

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            9.114819   0.124554  73.180  < 2e-16 ***
octinoxate             0.099451   0.027365   3.634 0.000329 ***
zinc_oxide            -0.002294   0.020492  -0.112 0.910957
octinoxate:zinc_oxide  0.069459   0.004534  15.319  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3632 on 296 degrees of freedom
Multiple R-squared:  0.9149,    Adjusted R-squared:  0.9141
F-statistic:  1061 on 3 and 296 DF,  p-value: < 2.2e-16
```

$In - vitro\ FPS = \beta_0 + \beta_1 * Conc\_Octinoxate + \beta_2 * Conc\_Zinc\_Oxide + \beta_3 * (Octinoxate * Zinc\_Oxide)$

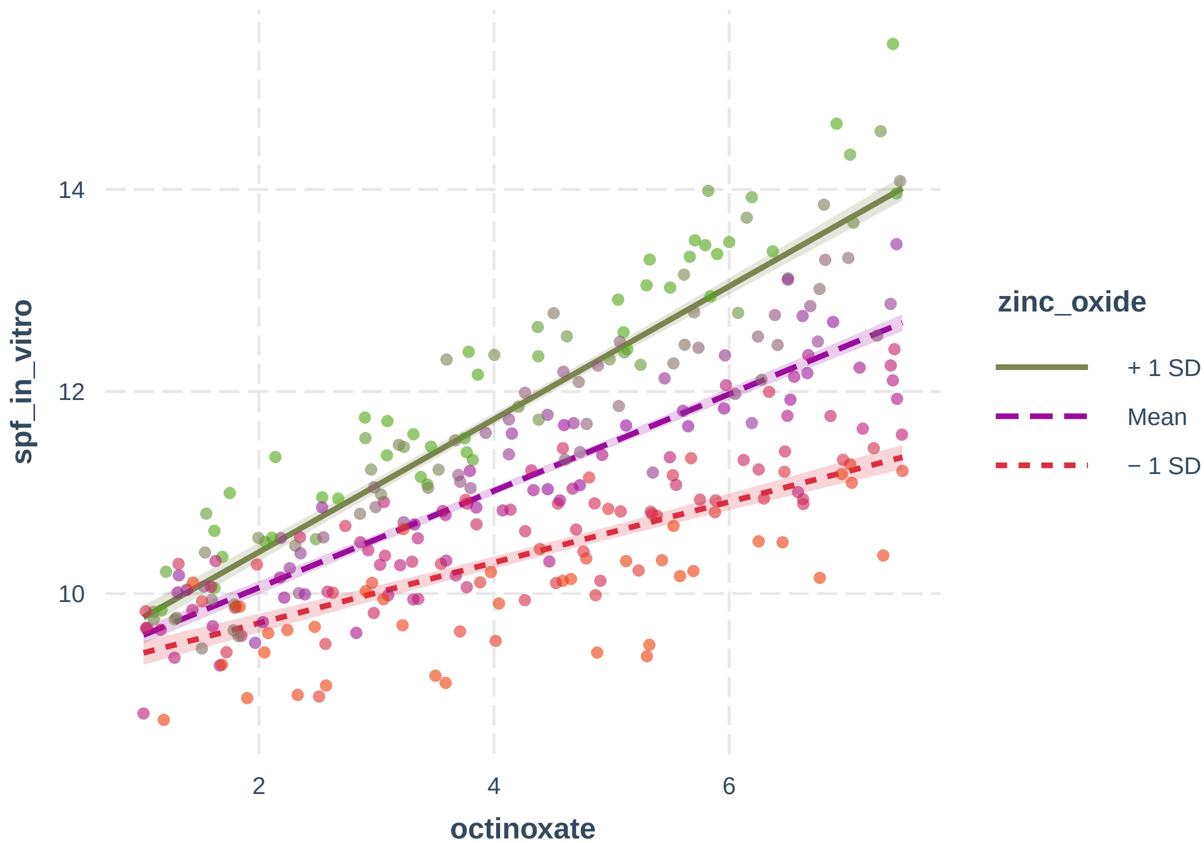
- β_0: This is the intercept or the predicted result when Octinoxate = 0 and Zinc Oxide = 0.
- β_1 $(octinoxate)$ This is the simple slope of Octinoxate for a one-unit change in Octinoxate.
- β_2 $(Zinc\_Oxide)$ This is the simple slope of Zinc Oxide for a one-unit change in Zinc Oxide.
- β_3 $(Octinoxate * Zinc\_Oxide)$ This is the interaction between Octinoxate and Zinc Oxide. It's the change in the slope of Octinoxate for each unit increase in Zinc Oxide.

The interaction between Octinoxate and Zinc Oxide is significant, and the combination has a synergistic effect on the in vitro SPF ($p <$ 2e-16). Octinoxate keeps the most important impact by itself in providing solar protection to a makeup foundation formulation, but Zinc Oxide doesn't have the same impact.

Using Octinoxate and Zinc Oxide together enhances the in vitro SPF. Depending on the context, we can consider using less Zinc Oxide without affecting the SPF value because its effect is not as important as Octinoxate's effect. However, we have to be careful when using this approach because Zinc Oxide provides the UVA protection to a sun protection factor formulation

Simple slopes analysis of interactions (octinoxate:zinc_oxide) allows us to figure out how the SPF is affected when Octinoxate's impact changes with different concentrations of Zinc Oxide. This is useful because it might help us optimize the amount of both ingredients required to reach the desired SPF without adding more than the formulation needs. This leads to reducing overcosts by minimizing the use of Zinc Oxide where it is not critical, speeding up formulation times by avoiding unnecessary testing, enhancing the development of new products by predicting the SPF without multiple tests, and reducing formulation and manufacturing times.

```r
interact_plot(cont_cont, pred = 'octinoxate', modx = 'zinc_oxide', plot.points = T, interval =  T ,colo
```



The interaction graphic showcases Octinoxate's effects on the in vitro SPF depending on the Zinc Oxide level. The more Zinc Oxide concentration, the more Octinoxate's effects. This approach allows us to understand the UV filters' behavior in order to get the optimal concentrations to achieve the desired SPF value, saving time and money by avoiding unnecessary testing, making the formulation and manufacturing process faster, and ensuring that the products have the labeled SPF level.

```
sim_slopes(cont_cont, pred = 'octinoxate', modx = 'zinc_oxide', cond.int =T)
```

```
JOHNSON-NEYMAN INTERVAL

When zinc_oxide is OUTSIDE the interval [-2.51, -0.59], the slope of
octinoxate is p < .05.

Note: The range of observed values of zinc_oxide is [1.00, 9.99]

SIMPLE SLOPES ANALYSIS

When zinc_oxide = 2.885854 (- 1 SD):

                               Est.   S.E.   t val.      p
---------------------------- ------ ------ -------- ------
Slope of octinoxate            0.30   0.02    18.27   0.00
Conditional intercept          9.11   0.08   120.95   0.00

When zinc_oxide = 5.460298 (Mean):
```

```
                              Est.   S.E.   t val.      p
--------------------------- ------ ------ -------- ------
Slope of octinoxate           0.48   0.01    41.83   0.00
Conditional intercept         9.10   0.05   172.21   0.00

When zinc_oxide = 8.034742 (+ 1 SD):

                              Est.   S.E.   t val.      p
--------------------------- ------ ------ -------- ------
Slope of octinoxate           0.66   0.02    40.40   0.00
Conditional intercept         9.10   0.07   122.85   0.00
```

As we noticed in the previous analysis above, Octinoxate has the main interaction with SPF, and it happens to be that Zinc Oxide increases this interaction if we raise its concentration in a formula. For instance, if we add 2.89% of Zinc Oxide into a formula, the in vitro SPF will increase by 0.33 units each time Octinoxate increases by 1% in the formula. This confirms the idea that we can optimize the SPF value by optimizing the amount of UV filters needed to reach the desired SPF, leading to optimization and saving formulation overrun.

These analyses have shown how the Octinoxate and Zinc Oxide interaction affects in vitro SPF, and how this interaction may allow us to optimize formulation processes, costs, and time, in general, to be more efficient. However, there are more variables that I was interested in analyzing in order to understand how they affect the formulations in general, formulation costs, and customer satisfaction. That is why I performed other interactions between numerical and categorical variables. Let's see them.

The first variable I wanted to check was the emulsifier type because they affect formulation stability and texture, which is very important when distributing the solar filters properly in order to optimize the in vitro SPF formulations.

I chose the ionic group as the reference group.

```r
foundation$emulsifier_type <- factor(foundation$emulsifier_type)
contrasts(foundation$emulsifier_type)
```

```
          non-ionic
ionic             0
non-ionic         1
```

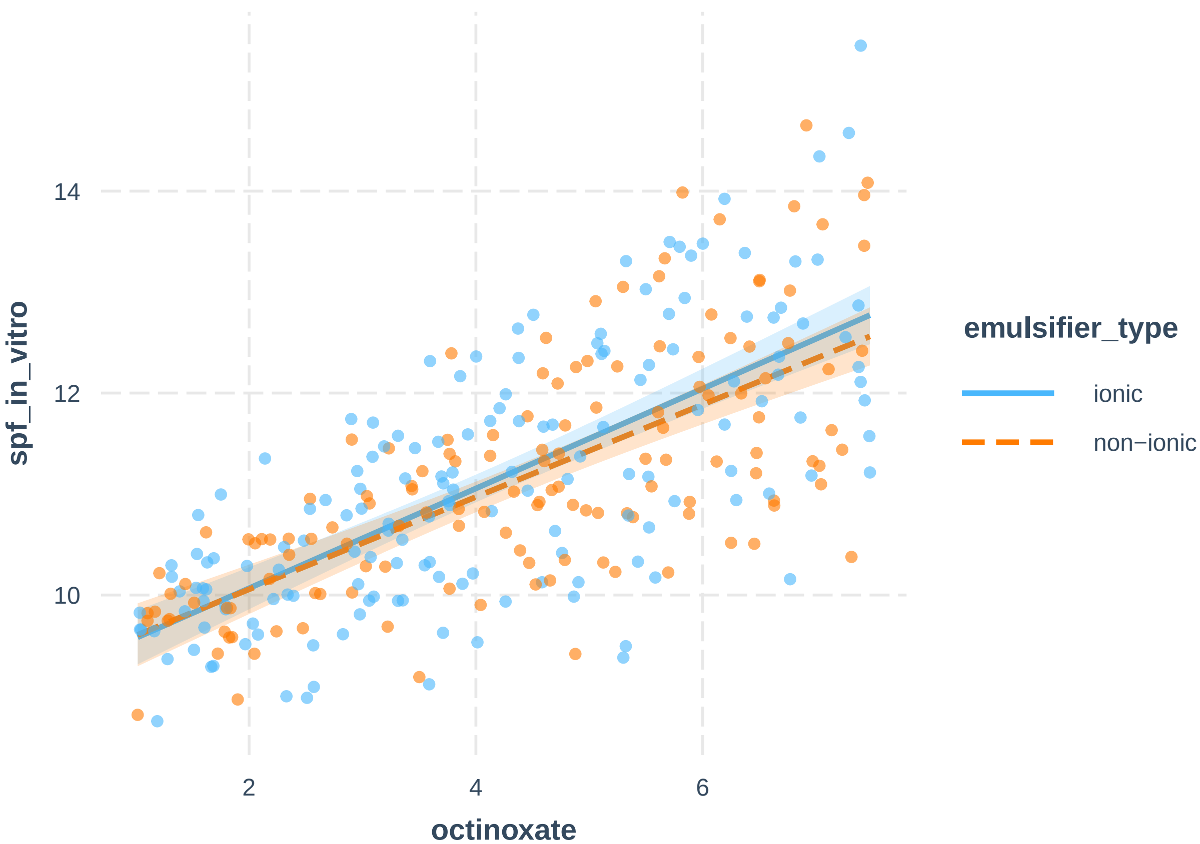
ionic = 0 and nonionic = 1, so ionic is the reference group

```r
cont_cat <- lm(spf_in_vitro~octinoxate*emulsifier_type, data = foundation)
summary(cont_cat)
```

```
Call:
lm(formula = spf_in_vitro ~ octinoxate * emulsifier_type, data = foundation)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.31752 -0.61591  0.02481  0.53477  2.70883 

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)                          9.07859    0.17054  53.235   <2e-16 ***
```

```
octinoxate                               0.49412    0.03809  12.973   <2e-16 ***
emulsifier_typenon-ionic                 0.06201    0.25993   0.239    0.812
octinoxate:emulsifier_typenon-ionic -0.03648    0.05601  -0.651    0.515
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8828 on 296 degrees of freedom
Multiple R-squared:  0.4974,    Adjusted R-squared:  0.4923
F-statistic: 97.63 on 3 and 296 DF,  p-value: < 2.2e-16
```

```
interact_plot(cont_cat, pred = octinoxate, modx = emulsifier_type, plot.points = T, interval = T)
```



The results of the model's analysis and graph indicate that the in vitro SPF is not affected by the emulsifier type (ionic, non-ionic) ($p = 0.812$) and does not interact significantly with Octinoxate either ($p = 0.515$). We might think that the emulsifier is not a raw material that affects the in vitro SPF directly, so we can proceed to figure out the rest of the variable interactions

So now, I will to check the interaction between Octinoxate and emulsifier (Natural or Synthetic)

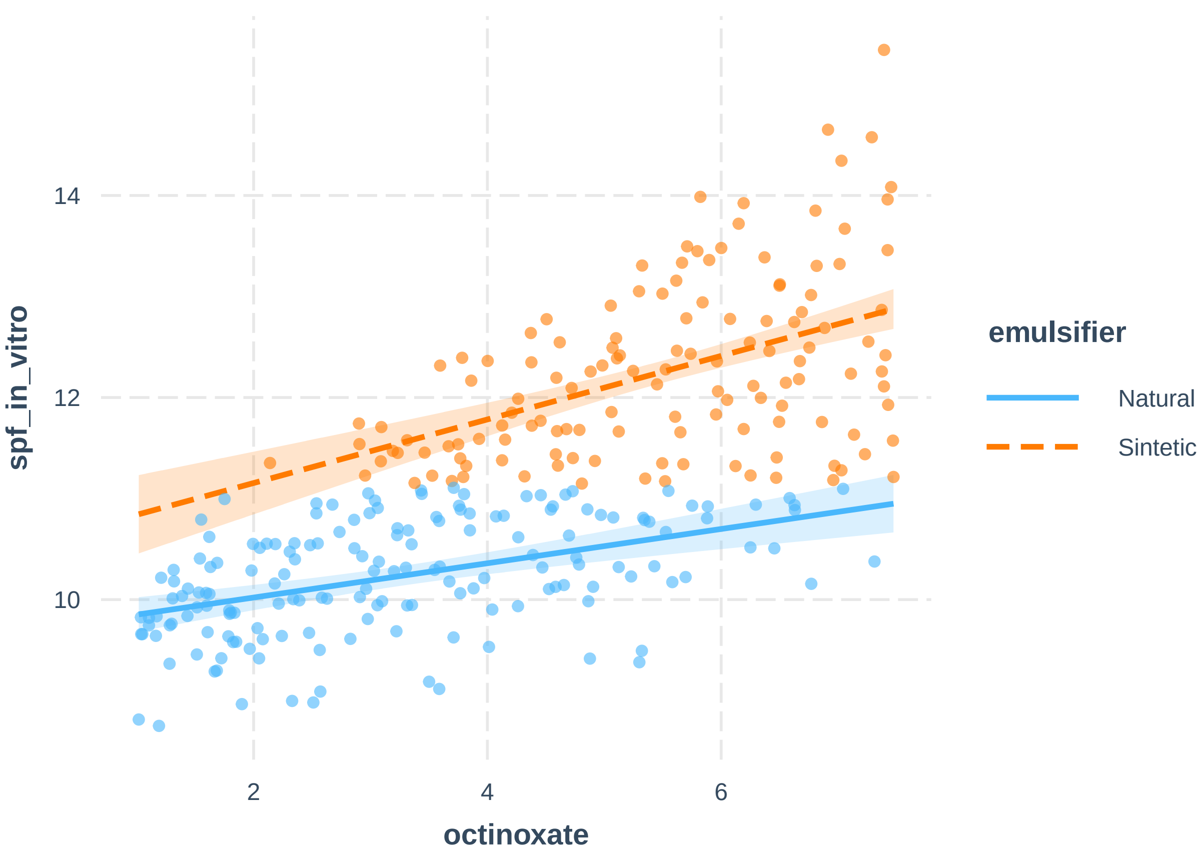
```
foundation$emulsifier <- factor(foundation$emulsifier)
contrasts(foundation$emulsifier)
```

```
         Sintetic
Natural         0
Sintetic        1
```

```r
cont_cat2 <- lm(spf_in_vitro~octinoxate*emulsifier, data = foundation)
summary(cont_cat2)
```

```
Call:
lm(formula = spf_in_vitro ~ octinoxate * emulsifier, data = foundation)

Residuals:
     Min       1Q   Median       3Q      Max
-1.66258 -0.41488 -0.01368  0.43893  2.58994

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                    9.68118    0.11509  84.121  < 2e-16 ***
octinoxate                     0.16963    0.03207   5.290 2.38e-07 ***
emulsifierSintetic             0.84344    0.26472   3.186  0.00160 **
octinoxate:emulsifierSintetic  0.14499    0.05298   2.736  0.00659 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6408 on 296 degrees of freedom
Multiple R-squared:  0.7352,    Adjusted R-squared:  0.7325
F-statistic: 273.9 on 3 and 296 DF,  p-value: < 2.2e-16
```

```r
interact_plot(cont_cat2, pred = octinoxate, modx = emulsifier, plot.points = T, interval = T)
```



$\beta_0 = 9.68118$: This is the intercept, or the predicted in vitro SPF when octinoxate = 0 in the reference group emulsifier, where Natural = 0. When octinoxate = 0 and emulsifier = Natural, the predicted in vitro SPF is 9.68.

β_1 $(octinoxate) = 0.16963$ 0.16963 is the simple slope of octinoxate for the reference group emulsifier, where Natural = 0. For each unit that octinoxate increases, the in vitro SPF increases by 0.17 ($p < 0.001$).

β_2 $(emulsifierSintetic) = 0.84344$ e0.84344 is the simple effect of the emulsifier or the difference in the in vitro SPF between emulsifier Natural and emulsifier Synthetic when octinoxate is 0. The way R does things implies that Synthetic is included in the equation. Changing to a synthetic emulsifier increases the in vitro SPF by 0.84 (p = 0.0016)

β_3 $(Octinoxate * emulsifier) = 0.14499 = 0.14499$ Interaction between octinoxate and emulsifier, is the difference in the simple slopes of octinoxate concentration for emulsifier Synthetic versus emulsifier Natural. The effect of octinoxate is greater when the emulsifier is synthetic. For each unit that octinoxate increases, the in vitro SPF increases by an additional 0.14 (p = 0.0066).

Model goodness: R^2 (0.7352): The model explains 73.52% of the variability in the in vitro SPF. Model p-value (< 2.2e-16): The model is significant.

```r
sim_slopes(cont_cat2, pred = 'octinoxate', modx = 'emulsifier')
```

```
SIMPLE SLOPES ANALYSIS

Slope of octinoxate when emulsifier = Natural:

  Est.   S.E.   t val.      p
------ ------ -------- ------
  0.17   0.03     5.29   0.00

Slope of octinoxate when emulsifier = Sintetic:

  Est.   S.E.   t val.      p
------ ------ -------- ------
  0.31   0.04     7.46   0.00
```

Simple Slopes:

- When emulsifier = Natural: The slope of octinoxate is 0.17 ($p < 0.001$). For each unit that octinoxate increases, the in vitro SPF increases by 0.17.
- When emulsifier = Synthetic: The slope of octinoxate is 0.31 ($p < 0.001$). For each unit that octinoxate increases, the in vitro SPF increases by 0.31.

These results showcase that the emulsifier (Natural vs. Synthetic) affects the in vitro SPF value. Synthetic emulsifiers increase the SPF value (coefficient = 0.84344, p = 0.0016) and also have a synergistic effect with Octinoxate (coefficient = 0.14499, p = 0.00659). This means that the effect of Octinoxate is higher in formulations with synthetic emulsifiers than with natural emulsifiers. These results allow us to be more efficient when choosing the emulsifier type or to reduce time by focusing on ingredient combinations to maximize the SPF.

Let's check weather the formula's base type, either oil or water, has any effect on in vitro SPF

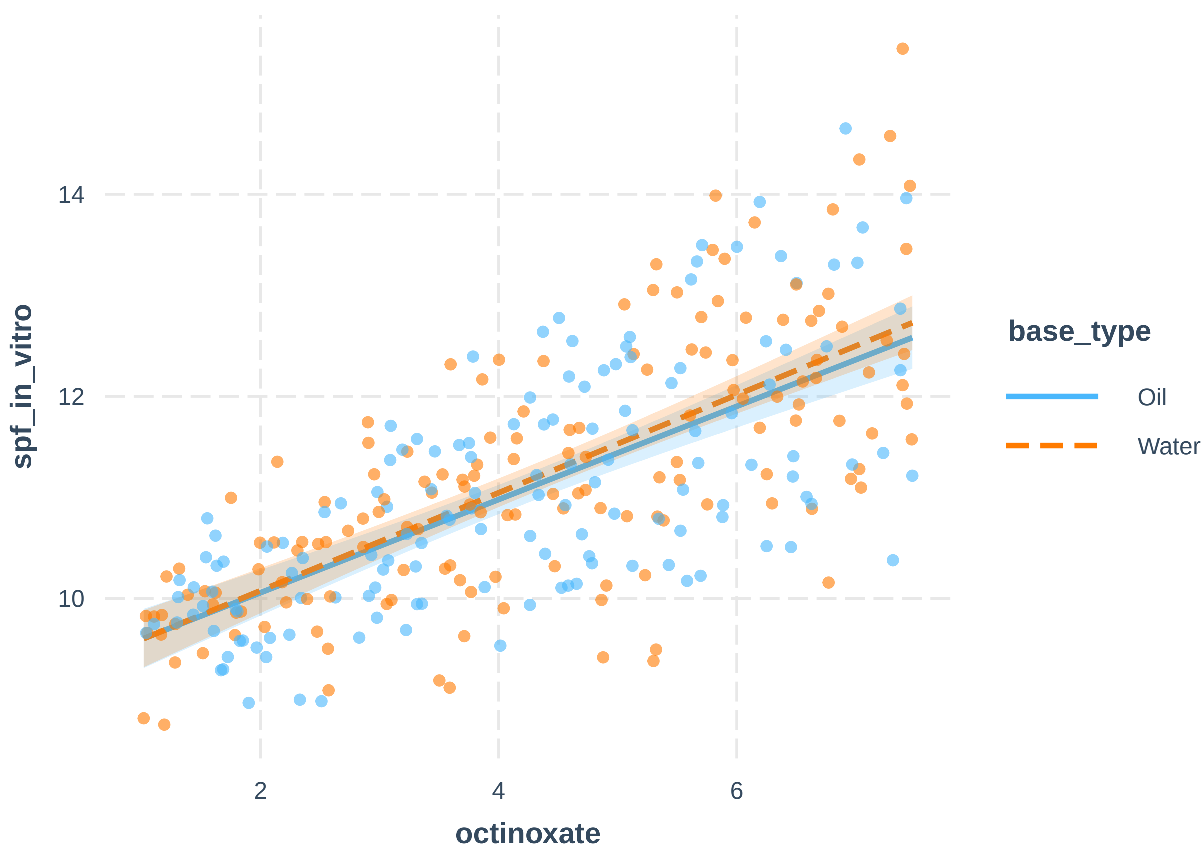
```r
foundation$base_type <- factor(foundation$base_type)
contrasts(foundation$base_type)
```

```
      Water
Oil       0
Water     1
```

```r
cont_cat3 <- lm(spf_in_vitro~octinoxate*base_type, data = foundation)
summary(cont_cat3)
```

```
Call:
lm(formula = spf_in_vitro ~ octinoxate * base_type, data = foundation)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.29506 -0.61480  0.01671  0.55100  2.75188 

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)                9.13391    0.18687  48.879   <2e-16 ***
octinoxate                 0.46113    0.04158  11.090   <2e-16 ***
base_typeWater            -0.02565    0.25800  -0.099    0.921    
octinoxate:base_typeWater  0.02316    0.05605   0.413    0.680    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8837 on 296 degrees of freedom
Multiple R-squared:  0.4964,    Adjusted R-squared:  0.4913 
F-statistic: 97.25 on 3 and 296 DF,  p-value: < 2.2e-16
```

```r
interact_plot(cont_cat3, pred = octinoxate, modx = base_type, plot.points = T, interval = T)
```



Based on the results, the type of base (Oil vs. Water) doesn't have a significant effect on the in vitro SPF ($p = 0.921$), it doesn't interact significantly with Octinoxate either ($p = 0.680$). It is not worth spending time on the interactions analysis since the SPF is not affected greatly by oil or water base formulations.

Interaction among categorical variables

Now, let's figure out how the interaction among categorical variables is.

- **FPS, EMULSIFIER (Natural, Synthetic), CSAT**

As we have seen so far, Octinoxate has a positive and significant effect on in vitro SPF, but emulsifier type (ionic or non-ionic) doesn't. However, the natural or synthetic base type shows a positive and significant effect on SPF values. Also, Octinoxate has a higher impact when it is used with synthetic emulsifiers.

Maybe the natural or synthetic base affects customer satisfaction, which can make a difference in the formula perceptions.

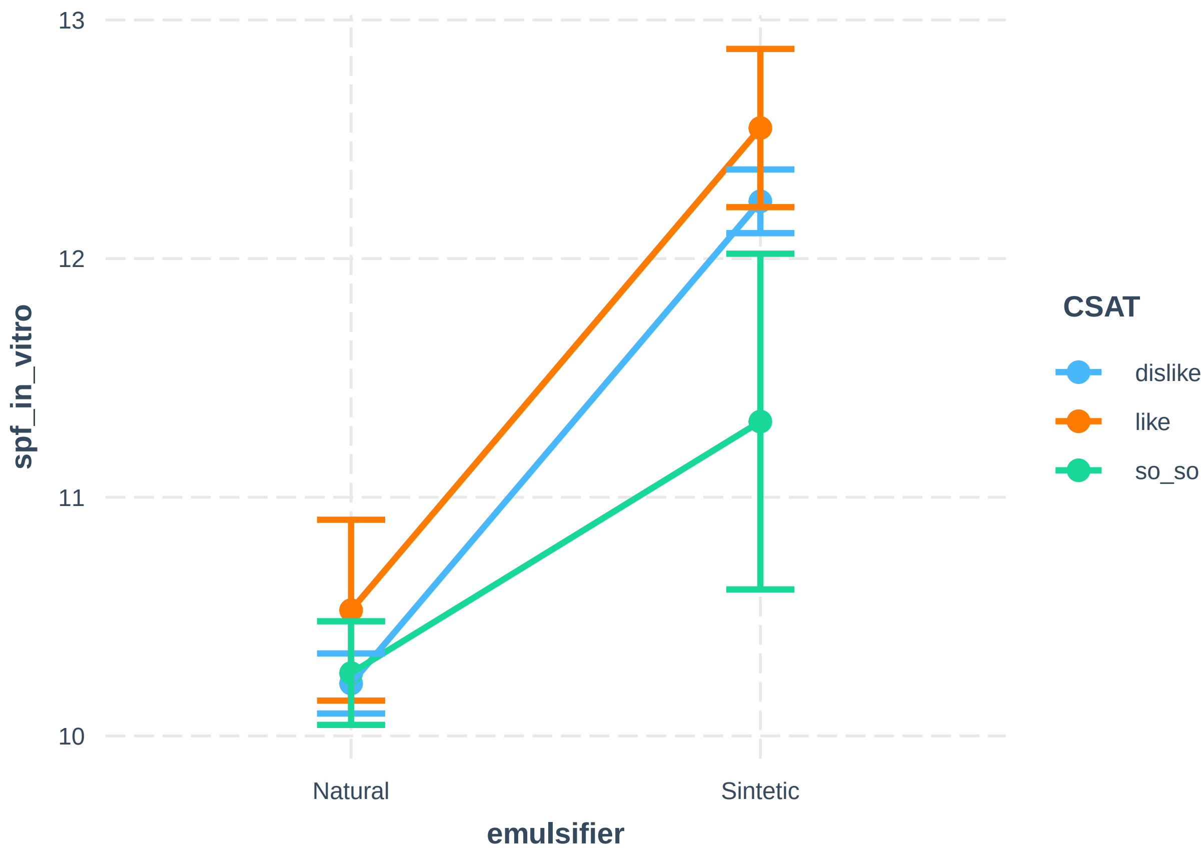
```
foundation$CSAT <- factor(foundation$CSAT)
foundation$CSAT <- relevel(foundation$CSAT, ref = 'dislike')
contrasts(foundation$CSAT)
```

```
        like so_so
dislike    0     0
like       1     0
so_so      0     1
```

```r
cat_cat <- lm(spf_in_vitro~emulsifier*CSAT, data = foundation)
summary(cat_cat)
```

```
Call:
lm(formula = spf_in_vitro ~ emulsifier * CSAT, data = foundation)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.4695 -0.5543 -0.0233  0.5155  3.2007 

Coefficients: (1 not defined because of singularities)
                             Estimate Std. Error t value Pr(>|t|)    
(Intercept)                  10.21963    0.06394 159.835   <2e-16 ***
emulsifierSintetic            2.02048    0.09323  21.672   <2e-16 ***
CSATlike                      0.30704    0.18164   1.690   0.0920 .  
CSATso_so                     0.04356    0.12750   0.342   0.7328    
emulsifierSintetic:CSATlike        NA         NA      NA       NA    
emulsifierSintetic:CSATso_so -0.96657    0.38550  -2.507   0.0127 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7149 on 295 degrees of freedom
Multiple R-squared:  0.6716,	Adjusted R-squared:  0.6671 
F-statistic: 150.8 on 4 and 295 DF,  p-value: < 2.2e-16
```

```r
cat_plot(cat_cat, pred = "emulsifier", modx = "CSAT", geom="line")
```



Those results shows how the SPF values change depending on the referenced category (dislike) using the Natural emulsifier group (The Natural emulsifier and dislike groups are the model's reference)

$CSAT : like$ (p = 0.0920): This result means that there isn't a statistical difference between the customers who says like from those who says dislike when they tested the Natural emulsifier formulation. So, there is no difference between 'like' and 'dislike' judgment when testing natural emulsifier formula.

$CSAT : so - so$ (p = 0.7328): There isn't a statistical difference for SPF between "so_so" and "dislike" when using Natural emulsifier either. So, there isn't a different perception between those customers who says "so_so" and "dislike" when using natural emulsifier

Using a synthetic emulsifier instead of a Natural one, the SPF value increases by 2.02.

Although synthetic emulsifiers usually increase SPF, if customer satisfaction is "so_so", this increase is partially negated.

Summarizing:

Formulations with Natural Emulsifier: There is no significant difference in SPF between "dislike", "like" and "so_so" ($p > 0.05$). SPF performance is consistent, but the formulation does not make a clear difference in customer perception.

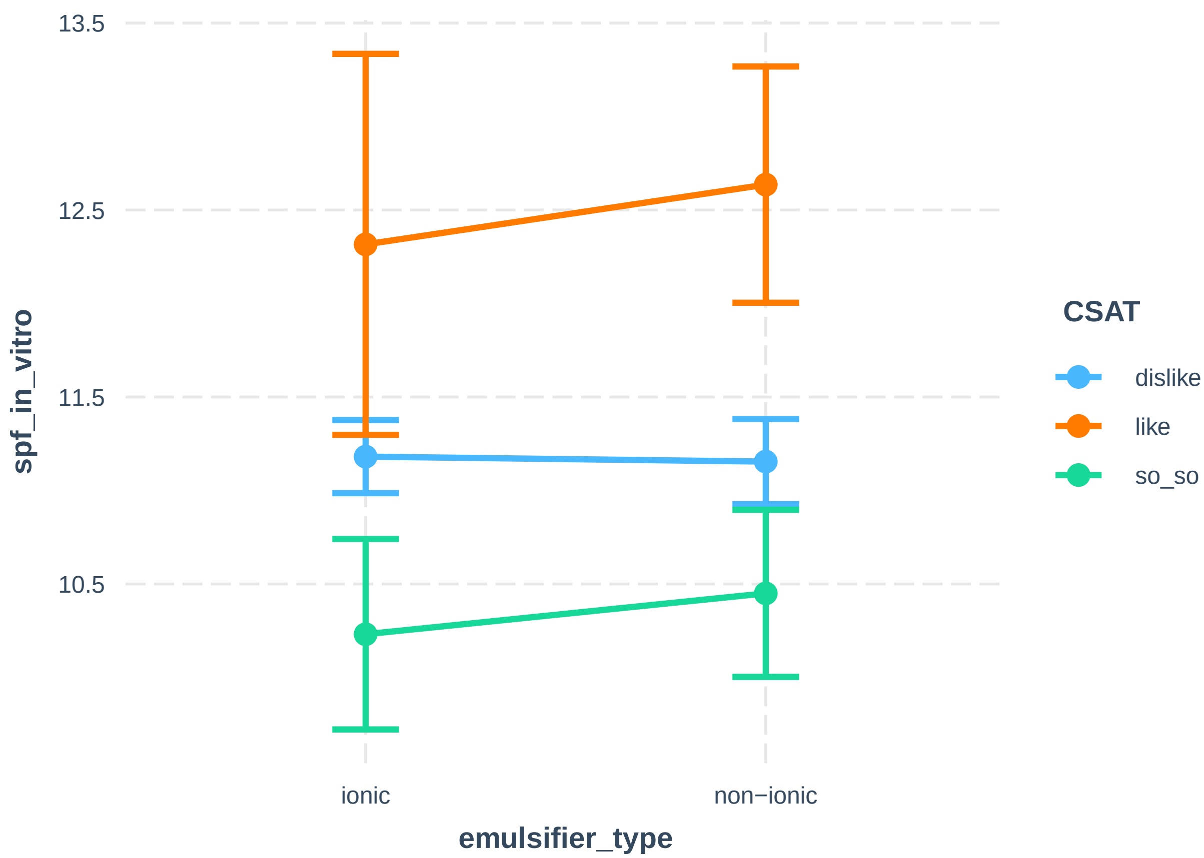
Formulations with Synthetic Emulsifier:There is no difference in SPF between "dislike" and "like" (p = 0.0920), but SPF decreases by 0.97 units in "so_so" (p = 0.0127). Synthetic formulas might have an inconsistent user experience, especially for those with a "so_so" perception.

- **FPS, EMULSIFIER (Ionic, No-ionic), CSAT**

```r
foundation$emulsifier_type <- factor(foundation$emulsifier_type)
foundation$emulsifier_type <- relevel(foundation$emulsifier_type, ref = 'ionic')

cat_cat2 <- lm(spf_in_vitro~emulsifier_type*CSAT, data = foundation)
summary(cat_cat2)
```

```
Call:
lm(formula = spf_in_vitro ~ emulsifier_type * CSAT, data = foundation)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.4310 -0.8714 -0.1119  0.6478  4.2596 

Coefficients:
                                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)                        11.18114    0.09925 112.656  < 2e-16 ***
emulsifier_typenon-ionic           -0.02643    0.15247  -0.173 0.862488    
CSATlike                            1.13523    0.52706   2.154 0.032060 *  
CSATso_so                          -0.94976    0.27719  -3.426 0.000699 ***
emulsifier_typenon-ionic:CSATlike   0.34598    0.62788   0.551 0.582038    
emulsifier_typenon-ionic:CSATso_so  0.24484    0.37651   0.650 0.516006    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.157 on 294 degrees of freedom
Multiple R-squared:  0.1419,    Adjusted R-squared:  0.1273 
F-statistic: 9.721 on 5 and 294 DF,  p-value: 1.317e-08
```

```r
cat_plot(cat_cat2, pred = "emulsifier_type", modx = "CSAT", geom="line")
```



Ionic Emulsifier: SPF increases by +1.14 for "like" ($p = 0.032$), the SPF decreases by -0.95 for "so_so" ($p = 0.0007$). SPF performance changes based on customer perception. When the product is liked ("like"), the SPF is better. When the perception is "so_so", the SPF goes down. Customer perception influences product performance, indicating a possible relationship between user experience and SPF efficacy in these formulas.

Non-Ionic Emulsifier: There are no significant differences in SPF, regardless of whether customers said "dislike", "like" or "so_so" ($p > 0.05$ across all interactions) Formulas with non-ionic emulsifier are consistent in their SPF, regardless of customer perception. Unlike ionic emulsifier, here satisfaction does not affect product performance. This could reflect a lack of measurable effect of customer satisfaction on SPF, which could be positive if product consistency is sought, or negative if better perception is expected to translate into better performance.

- **FPS, base type and CSAT**

```
contrasts(foundation$base_type)
```

```
      Water
Oil       0
Water     1
```

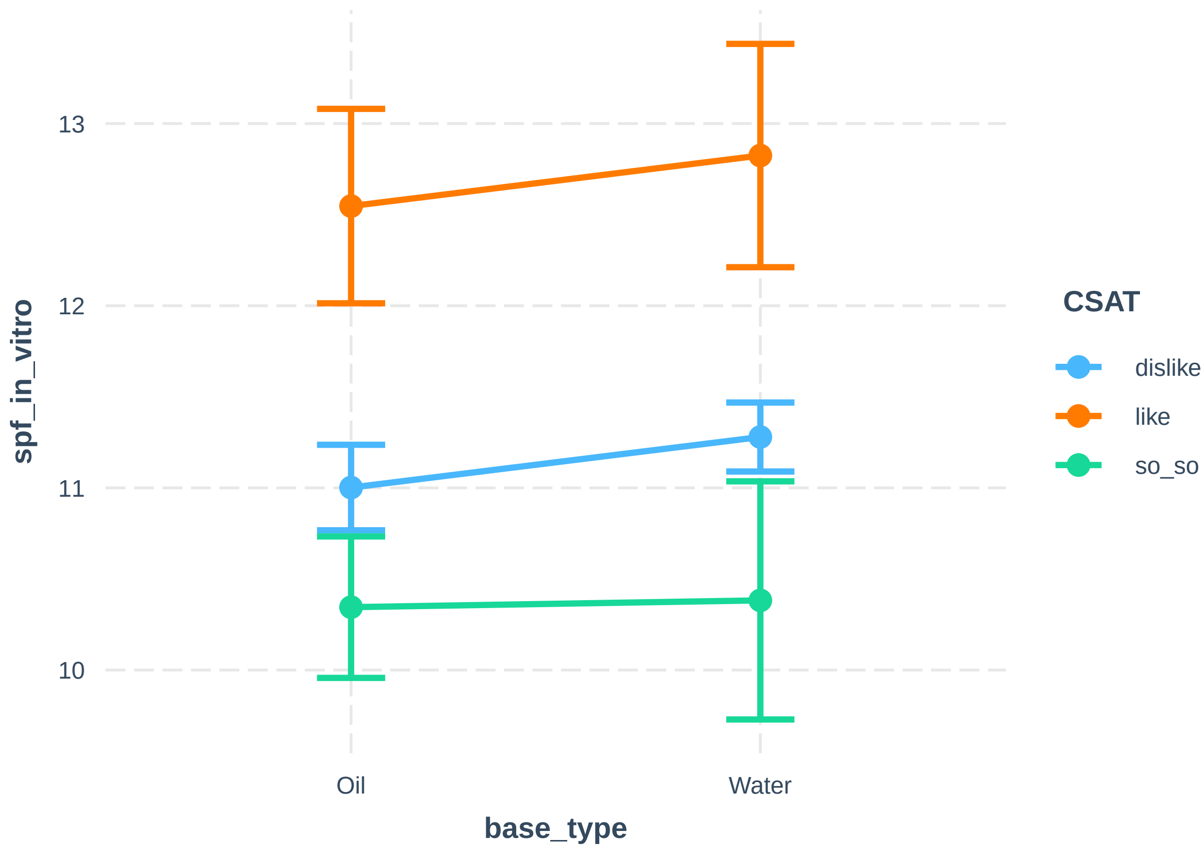
```
contrasts(foundation$CSAT)
```

```
        like so_so
dislike    0     0
like       1     0
so_so      0     1
```

```r
cat_cat_fps_base_type_CSAT <- lm(spf_in_vitro~base_type*CSAT, data = foundation)
summary(cat_cat_fps_base_type_CSAT)
```

```
Call:
lm(formula = spf_in_vitro ~ base_type * CSAT, data = foundation)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.5292 -0.8342 -0.0804  0.6518  4.1615 

Coefficients: (1 not defined because of singularities)
                         Estimate Std. Error t value Pr(>|t|)    
(Intercept)               11.0018     0.1193  92.220  < 2e-16 ***
base_typeWater             0.2774     0.1533   1.810  0.07128 .  
CSATlike                   1.5453     0.2963   5.216 3.45e-07 ***
CSATso_so                 -0.6567     0.2306  -2.848  0.00471 ** 
base_typeWater:CSATlike        NA         NA      NA       NA    
base_typeWater:CSATso_so  -0.2403     0.4156  -0.578  0.56357    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.15 on 295 degrees of freedom
Multiple R-squared:  0.1493,    Adjusted R-squared:  0.1377 
F-statistic: 12.94 on 4 and 295 DF,  p-value: 1.014e-09
```

```r
cat_plot(cat_cat_fps_base_type_CSAT, pred = "base_type", modx = "CSAT", geom="line")
```



“So, we can analyze the above results in a more or less technical way, as follows:

Reference categories: Base: oil, CSAT: dislike

$In - vitro\ \ FPS = 11.0018 + base_typeWater * 0.2774 + CSATlike * 1.5453 - CSATso_so * 0.6567 - base_typeWater : CSATso_so * 0.2403$

- CSATlike: When the base is ‘oil’ type and the customer satisfaction (CSAT) is ‘like’, the in vitro SPF is 1.55 higher than the reference group (base oil and CSAT dislike).
- CSATso_so: When the base is ‘oil’ type and the customer satisfaction (CSAT) is ‘so-so’, the in vitro SPF goes down 0.6567 compared to the reference group.
- base_typeWater: When the base is water type and the customer satisfaction is ‘dislike’, the SPF increases by 0.2774 units compared to the oil base type marked as ‘dislike’.
- There isn’t any case for the base_typeWater:CSATlike interaction in the data.
- The interaction base_typeWater:CSATso_so has a coefficient of -0.2403, which means that in base water formulations that are marked as ‘so-so’, the SPF decreases by 0.2403 units compared to those base oil formulations marked as ‘dislike’.

Those results can sometimes be overwhelming and hard to understand. However, we can come up with an easy interpretation. For instance, if we want to enhance customer satisfaction and optimize the formulation process, we should focus on oil-based formulations, mainly to achieve a ‘like’ perception from customers. According to the model coefficients, the ‘like’ group has 1.55 SPF units higher than the ‘dislike’ group. Water-based formulations don’t show a clear impact for those formulas with a ‘so-so’ customer perception, as they don’t have a significant change. That’s why we should prioritize oil-based formulations to improve their performance and effectiveness.

Viscosity approach

Viscosity is a key characteristic in foundation makeup formulations because:

- Regarding SPF, getting a proper viscosity ensures a uniform distribution of UV filters, enhancing the SPF level.
- Oil-based formulations tend to have a higher viscosity than water-based formulations. Either oil or water affects the viscosity, impacting stability and perception on the skin.
- It is very common for synthetic emulsifiers to increase viscosity more than natural ones.
- **Viscosity, Emulsifier (Natural, Synthetic), CSAT**

```r
contrasts(foundation$emulsifier)
```

```
         Sintetic
Natural         0
Sintetic        1
```

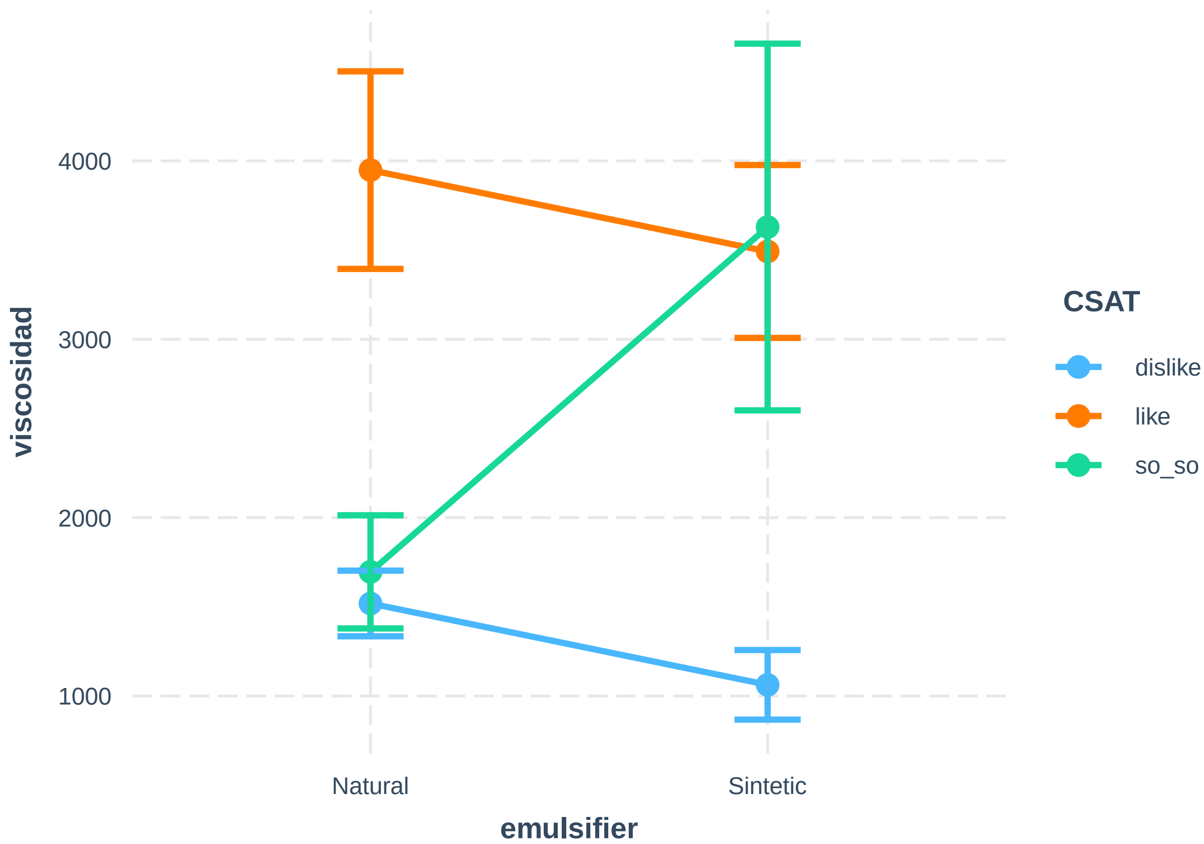
```r
contrasts(foundation$CSAT)
```

```
        like so_so
dislike    0     0
like       1     0
so_so      0     1
```

```r
cat_cat_visc <- lm(viscosidad~emulsifier*CSAT, data = foundation)
summary(cat_cat_visc)
```

```

Call:
lm(formula = viscosidad ~ emulsifier * CSAT, data = foundation)

Residuals:
    Min      1Q  Median      3Q     Max 
-1417.2  -749.5  -317.1   504.7  3473.8 

Coefficients: (1 not defined because of singularities)
                             Estimate Std. Error t value Pr(>|t|)    
(Intercept)                   1518.74      93.43  16.255  < 2e-16 ***
emulsifierSintetic            -456.16     136.24  -3.348 0.000919 ***
CSATlike                      2429.56     265.43   9.153  < 2e-16 ***
CSATso_so                      177.35     186.31   0.952 0.341927    
emulsifierSintetic:CSATlike        NA         NA      NA       NA    
emulsifierSintetic:CSATso_so  2389.42     563.33   4.242 2.97e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1045 on 295 degrees of freedom
Multiple R-squared:  0.2587,	Adjusted R-squared:  0.2486 
F-statistic: 25.74 on 4 and 295 DF,  p-value: < 2.2e-16
```

```
cat_plot(cat_cat_visc, pred = "emulsifier", modx = "CSAT", geom="line")
```



The coefficients of the model and graph show that formulas with natural emulsifiers and a "dislike" perception have a higher SPF than those with synthetic emulsifiers and the same "dislike" perception. Customers marked both types of emulsifier formulations (natural or synthetic) as "dislike" when they had lower viscosity values. Less viscosity, worse perception.

However, there is an interesting challenge with natural formulations that were marked as 'so-so' because their viscosity increases significantly, so they can be adjusted to enhance customer satisfaction.

$Intercept$, 1518.74 cP: This value represents the average viscosity for formulations made with a natural emulsifier and marked as "dislike."

$EmulsifierSynthetic$, -456.16 cP: When switching from a natural emulsifier to a synthetic one in formulas marked as "dislike," viscosity decreases by 456.16 cP.

$CSATlike$, 2429.56 cP: Viscosity shows a notable change for natural emulsifier formulations between those marked as "dislike" and "like." It seems that customers perceive formulas with higher viscosity made with a natural emulsifier more favorably.

$CSATso_so$, 177.35 cP: For natural emulsifier formulations, customers can't identify a viscosity difference between formulas marked as "so-so" and "dislike" because the viscosity change (177.35 cP) is not statistically significant.

$EmulsifierSynthetic : CSATso_so$, 2389.42 cP: Customers perceive a significant viscosity change in synthetic emulsifier formulations. When viscosity is low, they perceive them as "dislike," but if viscosity is higher, they perceive them as "so-so."

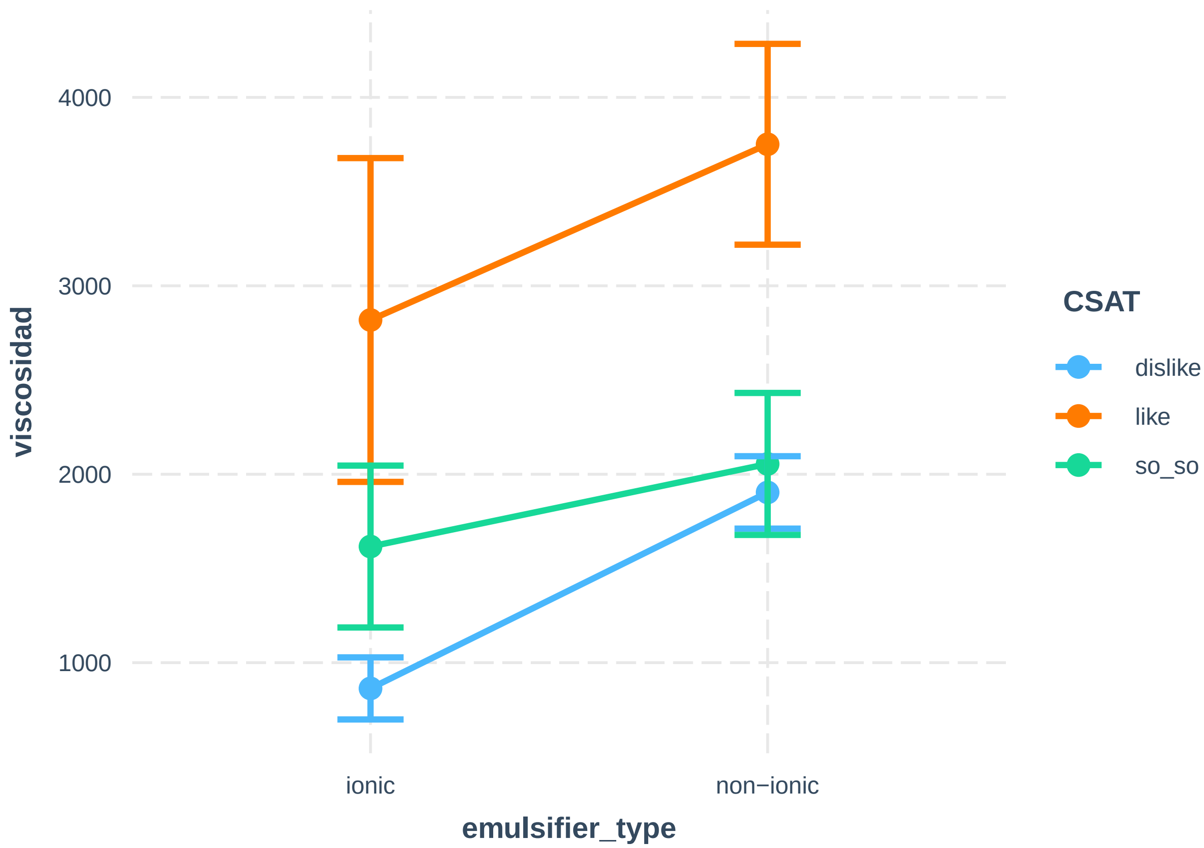
$EmulsifierSynthetic : CSATlike$, (NA): There is no available data for this in the dataset.

- **Viscosity, Emulsifier type (Oil, Water), CSAT**

```
cat_cat_visc_emulsifier_type <- lm(viscosidad ~ emulsifier_type*CSAT, data = foundation)
summary(cat_cat_visc_emulsifier_type)
```

```
Call:
lm(formula = viscosidad ~ emulsifier_type * CSAT, data = foundation)

Residuals:
    Min      1Q  Median      3Q     Max
-1680.5  -555.5  -195.9   214.0  3088.9

Coefficients:
                                   Estimate Std. Error t value Pr(>|t|)
(Intercept)                           863.4       83.7  10.316  < 2e-16 ***
emulsifier_typenon-ionic             1040.2      128.6   8.089 1.59e-14 ***
CSATlike                             1955.2      444.5   4.399 1.52e-05 ***
CSATso_so                             753.0      233.8   3.221  0.00142 **
emulsifier_typenon-ionic:CSATlike    -107.6      529.5  -0.203  0.83905
emulsifier_typenon-ionic:CSATso_so   -601.8      317.5  -1.895  0.05903 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 976.1 on 294 degrees of freedom
Multiple R-squared:  0.3549,    Adjusted R-squared:  0.3439
F-statistic: 32.35 on 5 and 294 DF,  p-value: < 2.2e-16
```

```
cat_plot(cat_cat_visc_emulsifier_type, pred= 'emulsifier_type', modx = 'CSAT', geom = 'line')
```



Regarding emulsifier type (ionic, non-ionic), the non-ionic formulations have the highest viscosity values.

Ionic emulsifier formulations: Customers prefer ionic emulsifier formulations with higher viscosity. Samples with higher viscosity are better rated ('like'), whereas samples with intermediate viscosity are perceived in a neutral way ('so-so').

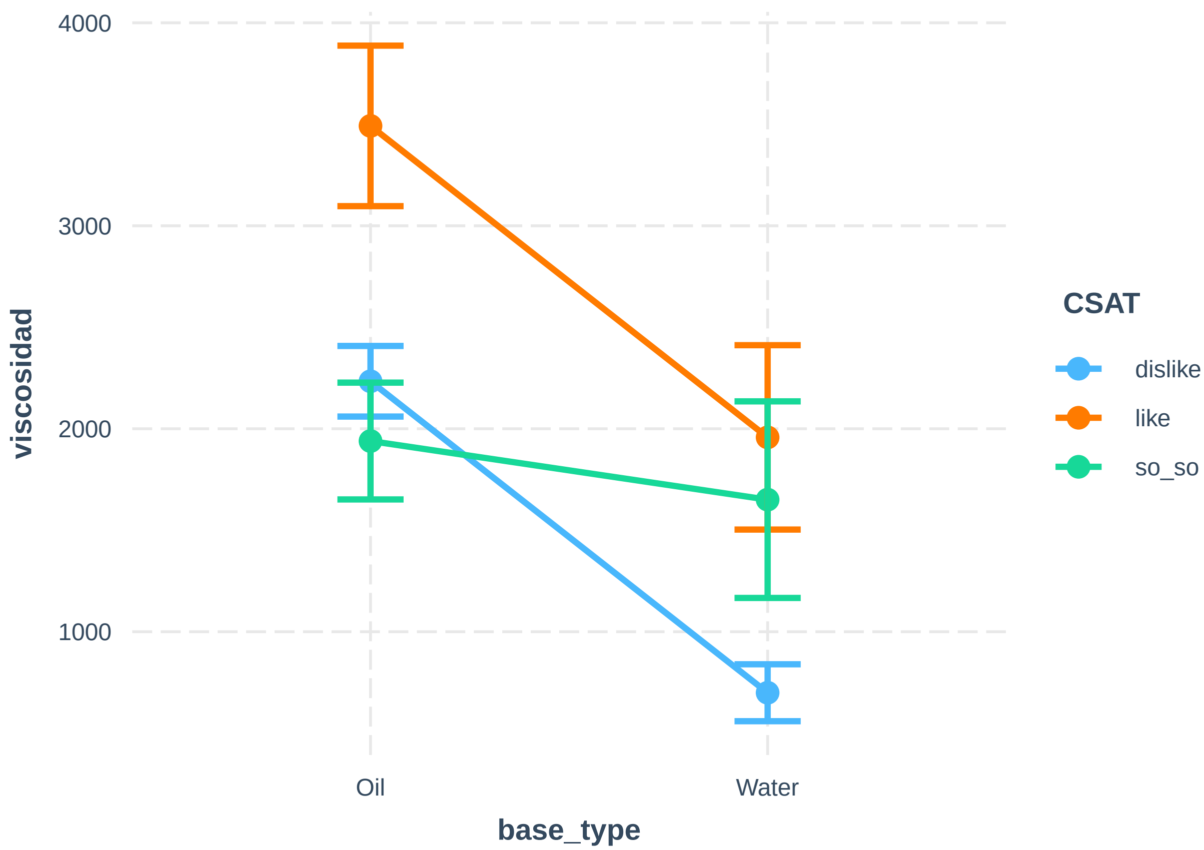
For non-ionic formulations, samples marked as 'like' have higher viscosity, but the interaction effect of emulsifier_type:CSATlike (-107.6 cP) is not significant ($p = 0.84$). This means that, even though formulations with higher viscosity are well rated, the increase in viscosity is the only factor that explains the customer experience.

```
cat_cat_visc_base_type <- lm(viscosidad~base_type*CSAT, data = foundation)
summary(cat_cat_visc_base_type)
```

```
Call:
lm(formula = viscosidad ~ base_type * CSAT, data = foundation)

Residuals:
     Min       1Q   Median       3Q      Max 
-1712.63  -445.55   -96.55   284.13  2758.51 

Coefficients: (1 not defined because of singularities)
                         Estimate Std. Error t value Pr(>|t|)    
(Intercept)               2234.03      88.39  25.275  < 2e-16 ***
base_typeWater           -1534.57     113.55 -13.514  < 2e-16 ***
CSATlike                  1258.10     219.50   5.732 2.46e-08 ***
```

```
CSATso_so                  -294.51     170.83  -1.724   0.0858 .
base_typeWater:CSATlike         NA         NA      NA       NA
base_typeWater:CSATso_so  1245.81     307.92   4.046 6.66e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 852.4 on 295 degrees of freedom
Multiple R-squared:  0.5064,    Adjusted R-squared:  0.4997
F-statistic: 75.66 on 4 and 295 DF,  p-value: < 2.2e-16
```

```
cat_plot(cat_cat_visc_base_type, pred ="base_type" , modx = "CSAT", geom="line")
```



Oil based formulations have the highest viscosity (2234 cP) and are generally preferred, because samples with higher viscosity are generally rated as 'like'.

The 'so-so' perception does not seem to be related to viscosity, as this group shows a slight reduction of 294.51 cP compared to 'dislike', but the difference is not statistically significant (p = 0.0858).

Water-based formulations have lower viscosity (-1535 cP) than oil based formulations. When customers perceive them as 'so-so', viscosity increases by 1246 cP, suggesting that very fluid formulations are less acceptable.

Costs

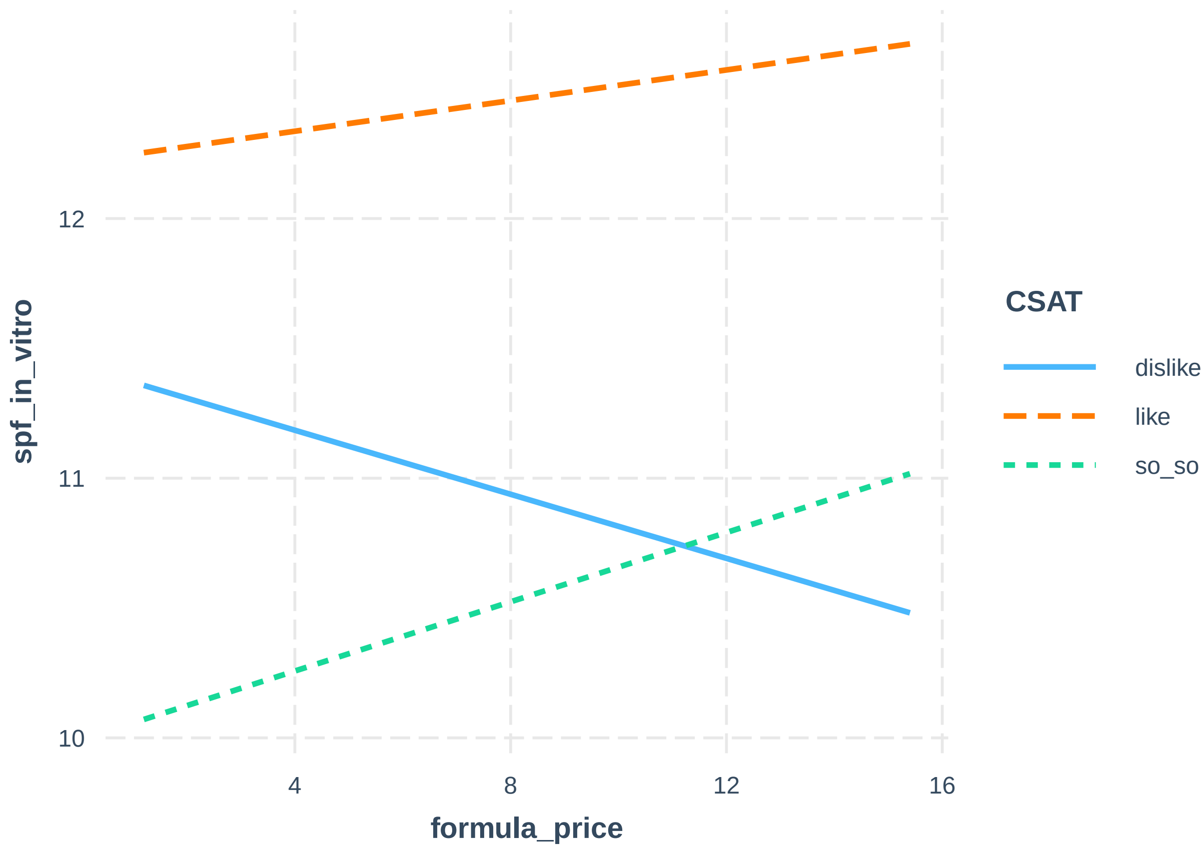
Analyzing formulation costs is interesting, especially how prices are affected by SPF and viscosity.

- **In vitro SPF, formula_price, Customer satisfaction**

```r
cat_cont_formula_price_CSAT <- lm(spf_in_vitro~formula_price*CSAT, data = foundation)
summary(cat_cont_formula_price_CSAT)
```

```
Call:
lm(formula = spf_in_vitro ~ formula_price * CSAT, data = foundation)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.5947 -0.8260 -0.0459  0.6315  4.1570 

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)             11.43151    0.11287 101.282  < 2e-16 ***
formula_price           -0.06169    0.02007  -3.073  0.00232 ** 
CSATlike                 0.78668    1.39124   0.565  0.57220    
CSATso_so               -1.44071    0.41474  -3.474  0.00059 ***
formula_price:CSATlike   0.09119    0.12366   0.737  0.46144    
formula_price:CSATso_so  0.12835    0.06927   1.853  0.06491 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.139 on 294 degrees of freedom
Multiple R-squared:  0.1695,	Adjusted R-squared:  0.1554 
F-statistic:    12 on 5 and 294 DF,  p-value: 1.384e-10
```

```r
interact_plot(cat_cont_formula_price_CSAT, pred ="formula_price", modx = "CSAT", geom="line")
```



Regarding customer perception, the results suggest that formulas marked as 'like' aren't negatively influenced by formula price in terms of in vitro SPF. This behavior differs from formulas marked as 'dislike,' where their in vitro SPF decreases as the formula price increases. However, since the interactions aren't statistically significant, there isn't strong evidence that formula price has a real effect on how customers perceive SPF.

- **Viscosity, formula_price, Customer satisfaction**

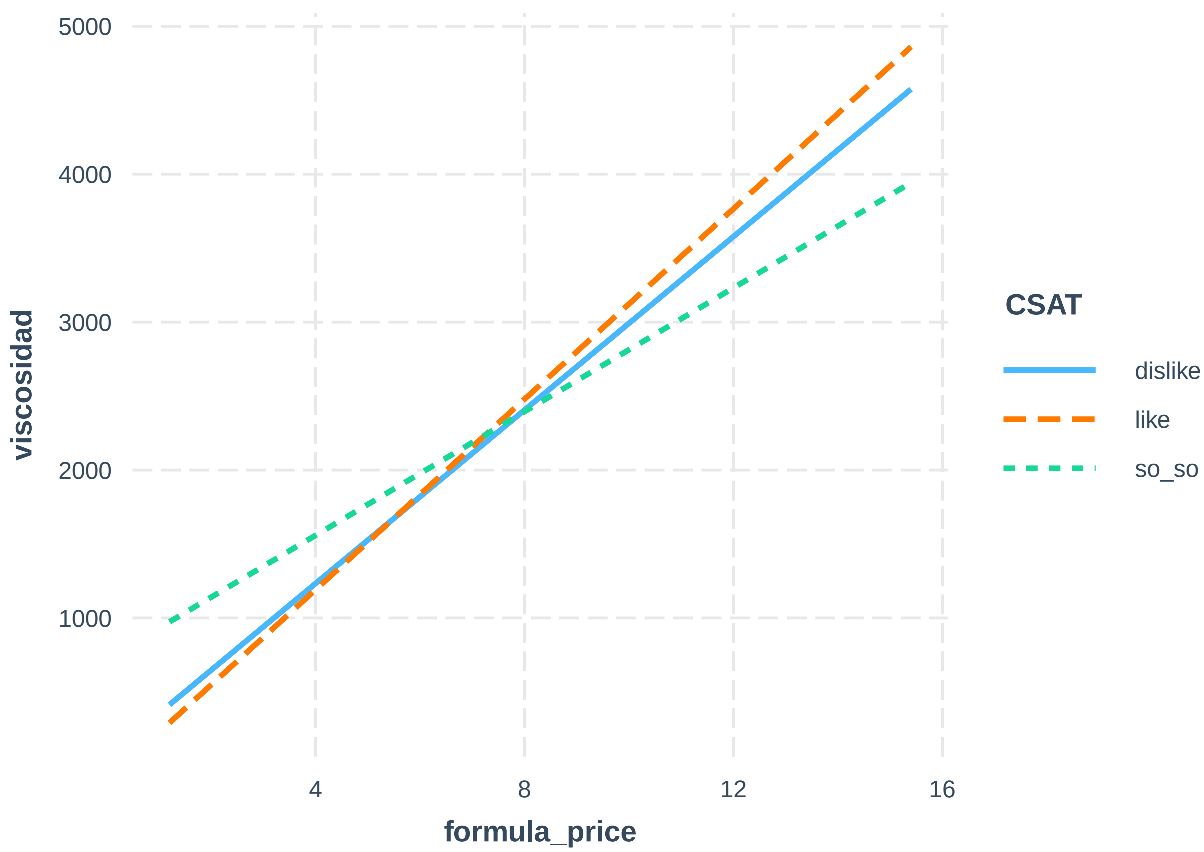
```
        like so_so
dislike    0     0
like       1     0
so_so      0     1
```

```
cat_cont_viscosidad_formula_price_CSAT <- lm(viscosidad~formula_price*CSAT, data = foundation)
summary(cat_cont_viscosidad_formula_price_CSAT)
```

```
Call:
lm(formula = viscosidad ~ formula_price * CSAT, data = foundation)

Residuals:
    Min      1Q  Median      3Q     Max
-900.69 -282.77  -35.88  255.23 1341.39

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)                61.757     40.734   1.516 0.130566    
formula_price             293.015      7.245  40.444  < 2e-16 ***
CSATlike                 -157.248    502.101  -0.313 0.754367    
CSATso_so                 660.912    149.681   4.415 1.42e-05 ***
formula_price:CSATlike     28.774     44.630   0.645 0.519607    
formula_price:CSATso_so   -83.976     25.001  -3.359 0.000886 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 410.9 on 294 degrees of freedom
Multiple R-squared:  0.8857,    Adjusted R-squared:  0.8837 
F-statistic: 455.5 on 5 and 294 DF,  p-value: < 2.2e-16
```

```r
interact_plot(cat_cont_viscosidad_formula_price_CSAT, pred ="formula_price", modx = "CSAT", geom="line")
```



The results show that more expensive formulations tend to have higher viscosity, but this relationship depends on customer perception:

- Formulas rated as 'dislike' show a stronger relationship between formula price and viscosity: higher price correlates with higher viscosity.
- Formulas rated as 'so-so' tend to have higher viscosity overall, but the relationship between price and viscosity is weaker compared to 'dislike' formulas.
- Formulas rated as 'like' do not show a clear relationship between price and viscosity. Although there is a slight increase in viscosity as the - - formula price increases, this effect is not statistically significant.
- This suggests that higher viscosity does not necessarily translate into better customer acceptance. In fact, the formulas perceived as 'so-so' tend to have the highest viscosity, which could indicate that excessive viscosity is not always favorable for customer satisfaction.

Conclusion:

This is a basic approach, but you can go further and do more.

This study has shown how we can use linear regression with interactions as a useful tool for those who need to understand or improve a formulation process or enhance formulations. Through the physicochemical analysis of variables, formulation characteristics, and customer perception, I identified key relationships that can improve foundation makeup formulations with SPF.

The results showed that:

- Octinoxate is the raw material that has the greatest in vitro sun protection factor impact. Its effect is enhanced when blended with Zinc Oxide. This interaction of UV filters allows us to adjust the desired or required SPF without negatively affecting product performance.
- The emulsifier affects viscosity and customer perception but not SPF directly. However, synthetic emulsifiers improve Octinoxate performance.
- Viscosity is a key factor in customer perception. Formulations with higher viscosity tend to be better rated by customers, although excessive viscosity could generate a worse perception rather than a positive one.
- Formula price doesn't seem to significantly affect customer perception in terms of SPF, although a relationship between price and viscosity is observed.

These findings could be applied in any industry to improve manufacturing or formulation processes, reducing unnecessary costs and enhancing product quality while avoiding multiple laboratory tests. The combination of chemistry, statistics, and data science helps companies make strategic decisions based on data, improving efficiency in general.

The main purpose of this study is to show a brief example of how we can apply a solution. This may give you an idea of how to tackle a challenge similar to this approach.

THANK YOU FOR READING. I HOPE IT HAS BEEN HELPFUL FOR YOU.