Color_Analysis

yiqb

2024-12-09

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(dslabs)
library(readr)
```

```
df <- read_csv("rgb_colored.csv")
```

```
## Rows: 15402 Columns: 16
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (9): gender, masterCategory, subCategory, articleType, baseColour, seaso...
## dbl (7): R, G, B, color_Hasler, color_HS, id, year
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
summary(df)
```

```
##        R                G                 B           color_Hasler
##  Min.   :  1.0   Min.   :  0.9004   Min.   :  1.00   Min.   : 0.000
##  1st Qu.:104.5   1st Qu.: 95.2743   1st Qu.: 98.19   1st Qu.: 3.536
##  Median :164.3   Median :150.5841   Median :149.63   Median : 8.436
##  Mean   :156.3   Mean   :145.9185   Mean   :147.13   Mean   :13.270
##  3rd Qu.:212.0   3rd Qu.:198.5109   3rd Qu.:196.98   3rd Qu.:18.837
##  Max.   :255.0   Max.   :254.9746   Max.   :254.97   Max.   :72.808
##  NA's   :589     NA's   :589        NA's   :589      NA's   :589
##     color_HS            id           gender          masterCategory
```

```
##  Min.   :0.0000   Min.   : 1163   Length:15402       Length:15402
##  1st Qu.:0.0479   1st Qu.:10377   Class :character   Class :character
##  Median :0.1146   Median :19892   Mode  :character   Mode  :character
##  Mean   :0.1752   Mean   :23195
##  3rd Qu.:0.2494   3rd Qu.:34027
##  Max.   :0.9346   Max.   :60000
##  NA's   :589
##  subCategory        articleType         baseColour           season
##  Length:15402       Length:15402       Length:15402       Length:15402
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##       year         usage           productDisplayName     link
##  Min.   :2008   Length:15402       Length:15402       Length:15402
##  1st Qu.:2011   Class :character   Class :character   Class :character
##  Median :2011   Mode  :character   Mode  :character   Mode  :character
##  Mean   :2011
##  3rd Qu.:2012
##  Max.   :2018
##  NA's   :1
```

```
baseColour_freq <- table(df$baseColour)
filtered_baseColours <- names(baseColour_freq[baseColour_freq >= 50])


df <- df[df$baseColour %in% filtered_baseColours, ]

filtered_baseColour_freq <- table(df$baseColour)
print("Filtered Base Colour Frequencies:")
```

```
## [1] "Filtered Base Colour Frequencies:"
```

```
print(filtered_baseColour_freq)
```

```
##
##     Beige     Black      Blue     Brown  Charcoal     Cream     Green      Grey
##       184      2022      2449       387        60       180      1270      1101
##  Lavender   Magenta    Maroon     Multi Navy Blue Off White     Olive    Orange
##        71        73       256       134       807       117       115       273
##     Peach      Pink    Purple       Red      Teal     White    Yellow
##       101       780       833      1191        54      2239       492
```

```
season_to_quarter <- function(season) {
  if (is.na(season)) {
    return(NA)
  }
  if (season == "Winter") {
    return(0)
  } else if (season == "Spring") {
```

```
    return(0.25)
  } else if (season == "Summer") {
    return(0.5)
  } else if (season == "Fall") {
    return(0.75)
  } else {
    return(NA)  # Handle any unexpected season values
  }
}


df$years <- df$year + sapply(df$season, season_to_quarter)

year_counts <- df %>%
  group_by(years) %>%
  summarise(count = n()) %>%
  arrange(years)

df <- df[df$years %in% year_counts$years, ]

print("Filtered Years Distribution:")
```

```
## [1] "Filtered Years Distribution:"
```

```
print(table(df$years))
```

```
##
## 2008.25 2009.75    2010 2010.25  2010.5 2010.75    2011 2011.25  2011.5 2011.75
##       1       3      57       1      15     327      43      47    2768    5537
##    2012 2012.25  2012.5 2012.75    2013 2013.25  2013.5 2013.75    2014 2014.25
##      44      12    5691     300       6      33      78      10       1       3
##  2014.5    2015  2015.5 2015.75    2016  2016.5    2017  2017.5    2018
##       5      14      45      11      58      73       1       3       1
```

```
color_data <- data.frame(
  color = c("Beige", "Black", "Blue", "Brown", "Cream", "Green", "Grey", "Maroon", "Multi",
            "Navy Blue", "Off White", "Olive", "Orange", "Peach", "Pink", "Purple", "Red",
            "White", "Yellow"),
  R = c(245, 0, 0, 139, 255, 0, 169, 128, 160, 0, 255, 128, 255, 255, 255, 128, 255, 255, 255),
  G = c(245, 0, 0, 69, 253, 255, 169, 0, 174, 0, 255, 128, 165, 189, 192, 0, 0, 255, 255),
  B = c(220, 0, 255, 19, 115, 0, 250, 0, 255, 0, 81, 0, 171, 70, 254, 128, 0, 0, 0)
)

df$RGB <- sapply(df$baseColour, function(base_colour) {
  matching_row <- color_data[color_data$color == base_colour, ]

  if (nrow(matching_row) > 0) {
    paste(matching_row$R, matching_row$G, matching_row$B, sep = ",")
  } else {
    NA
  }
})
```

```r
print(head(df))
```

```
## # A tibble: 6 x 18
##       R     G     B color_Hasler color_HS    id gender masterCategory
##   <dbl> <dbl> <dbl>        <dbl>    <dbl> <dbl> <chr>  <chr>
## 1 135.  135.  151.          4.85   0.0648 15970 Men    Apparel
## 2 128.  147.  154.          7.54   0.102  53759 Men    Apparel
## 3 185.  170.  179.          4.35   0.0565  1855 Men    Apparel
## 4 121.  135.  108.          7.34   0.106  30805 Men    Apparel
## 5 127.   78.9 100.         14.5    0.189  26960 Women  Apparel
## 6  75.1  48.7  61.5         7.93   0.104  12369 Men    Apparel
## # i 10 more variables: subCategory <chr>, articleType <chr>, baseColour <chr>,
## #   season <chr>, year <dbl>, usage <chr>, productDisplayName <chr>,
## #   link <chr>, years <dbl>, RGB <chr>
```

```r
gender_stats <- df %>%
  group_by(gender) %>%
  summarise(count = n(), .groups = "drop") %>%
  mutate(percentage = (count / sum(count)) * 100)

print("Gender Ratio:")
```

```
## [1] "Gender Ratio:"
```

```r
print(gender_stats)
```

```
## # A tibble: 5 x 3
##   gender count percentage
##   <chr>  <int>      <dbl>
## 1 Boys     631      4.15
## 2 Girls    341      2.25
## 3 Men     8756     57.6
## 4 Unisex    84      0.553
## 5 Women   5377     35.4
```

```r
base_color_stats <- df %>%
  group_by(baseColour) %>%
  summarise(count = n(), .groups = "drop") %>%
  arrange(desc(count))

print("Base Colour Distribution:")
```

```
## [1] "Base Colour Distribution:"
```

```r
print(base_color_stats)
```

```
## # A tibble: 23 x 2
##    baseColour count
```

```
##    <chr>      <int>
##  1 Blue        2449
##  2 White       2239
##  3 Black       2022
##  4 Green       1270
##  5 Red         1191
##  6 Grey        1101
##  7 Purple       833
##  8 Navy Blue    807
##  9 Pink         780
## 10 Yellow       492
## # i 13 more rows
```

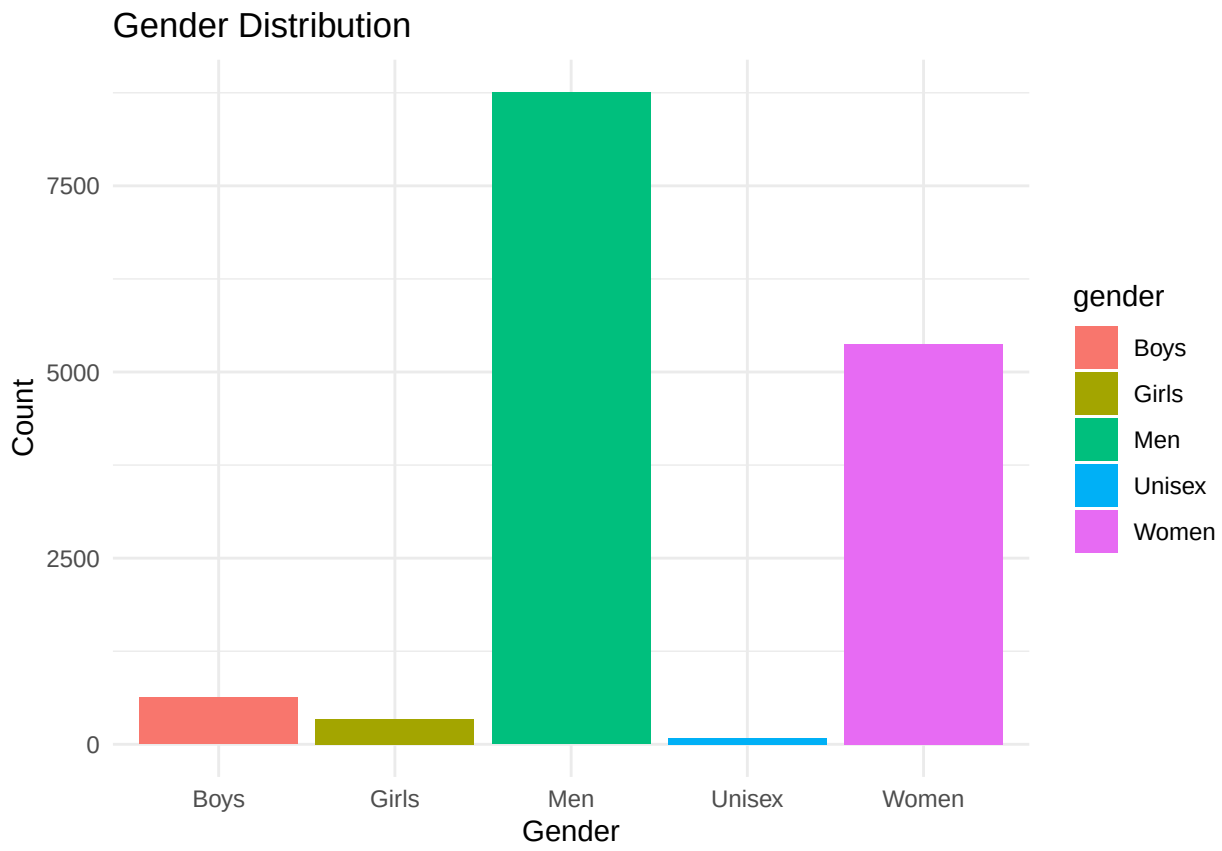
```
yearly_stats <- df %>%
  group_by(years) %>%
  summarise(count = n(), .groups = "drop")%>%
  arrange(years)

print("Yearly Distribution (Filtered):")
```

```
## [1] "Yearly Distribution (Filtered):"
```

```
print(yearly_stats)
```

```
## # A tibble: 30 x 2
##    years count
##    <dbl> <int>
##  1 2008.     1
##  2 2010.     3
##  3 2010     57
##  4 2010.     1
##  5 2010.    15
##  6 2011.   327
##  7 2011     43
##  8 2011.    47
##  9 2012.  2768
## 10 2012.  5537
## # i 20 more rows
```

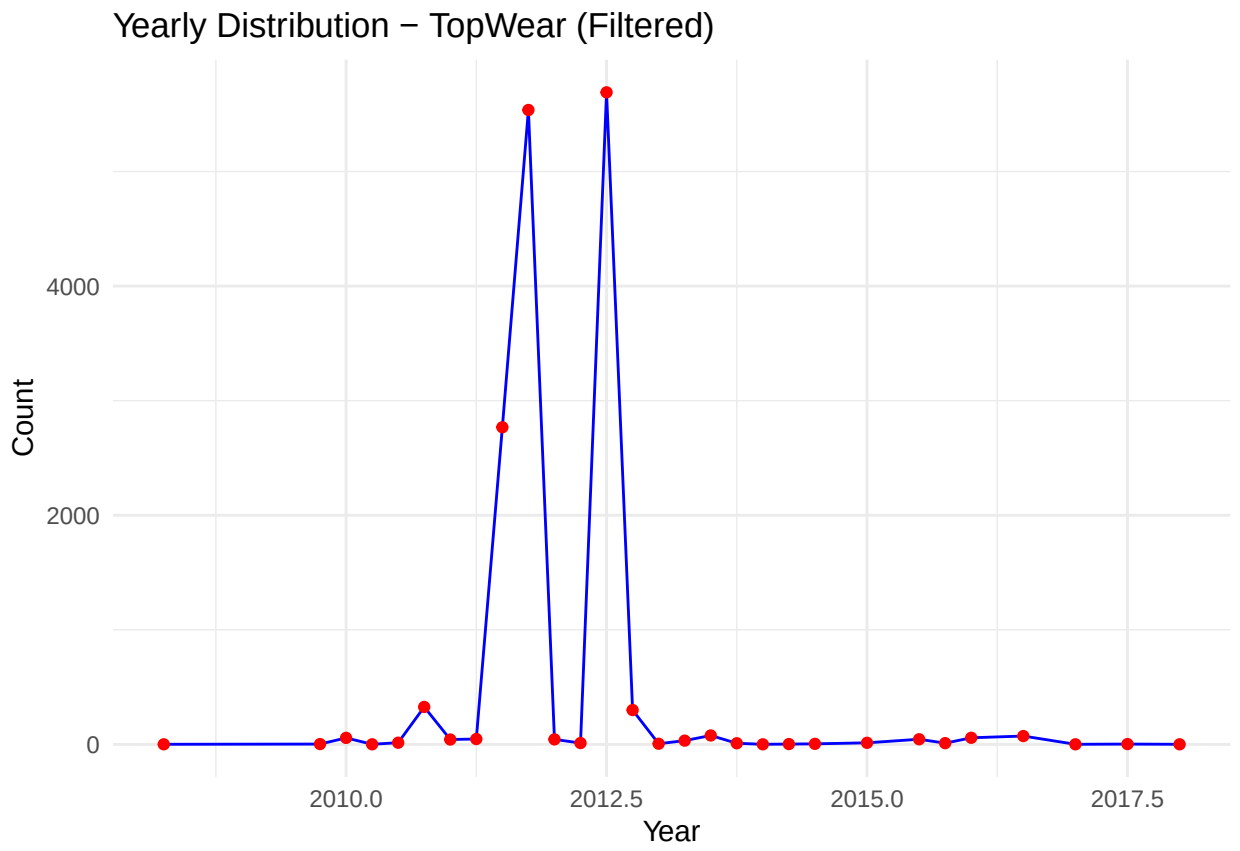
```
ggplot(gender_stats, aes(x = gender, y = count, fill = gender)) +
  geom_bar(stat = "identity") +
  labs(
    title = "Gender Distribution",
    x = "Gender",
    y = "Count"
  ) +
  theme_minimal()
```




```
ggplot(yearly_stats, aes(x = years, y = count)) +
  geom_line(group = 1, color = "blue") +
  geom_point(color = "red") +
  labs(
    title = "Yearly Distribution - TopWear (Filtered)",
    x = "Year",
    y = "Count"
  ) +
  theme_minimal()
```

```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_line()`).
```

```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_point()`).
```




```r
# Compute Hasler and HS metrics
df$color_Hasler_premade <- apply(df, 1, function(row) {
  # Extract RGB values
  r <- as.numeric(row["R"])
  g <- as.numeric(row["G"])
  b <- as.numeric(row["B"])

  rg <- abs(r - g)
  yb <- abs(0.5 * (r + g) - b)


  sqrt(rg^2 + yb^2) + 0.3 * mean(c(rg, yb))
})

df$color_HS_premade <- apply(df, 1, function(row) {

  r <- as.numeric(row["R"])
  g <- as.numeric(row["G"])
  b <- as.numeric(row["B"])

  # Calculate rg, yb channels
  rg <- abs(r - g)
  yb <- abs(0.5 * (r + g) - b)

  # Compute HS's colorfulness
  sqrt(mean(c(rg^2, yb^2)))
```

```
})


print(head(df))
```

```
## # A tibble: 6 x 20
##       R     G     B color_Hasler color_HS    id gender masterCategory
##   <dbl> <dbl> <dbl>        <dbl>    <dbl> <dbl> <chr>  <chr>
## 1 135.  135.  151.          4.85   0.0648 15970 Men    Apparel
## 2 128.  147.  154.          7.54   0.102  53759 Men    Apparel
## 3 185.  170.  179.          4.35   0.0565  1855 Men    Apparel
## 4 121.  135.  108.          7.34   0.106  30805 Men    Apparel
## 5 127.   78.9 100.         14.5    0.189  26960 Women  Apparel
## 6  75.1  48.7  61.5         7.93   0.104  12369 Men    Apparel
## # i 12 more variables: subCategory <chr>, articleType <chr>, baseColour <chr>,
## #   season <chr>, year <dbl>, usage <chr>, productDisplayName <chr>,
## #   link <chr>, years <dbl>, RGB <chr>, color_Hasler_premade <dbl>,
## #   color_HS_premade <dbl>
```

```
# write.csv(df, "filtered_with_colorfulness.csv", row.names = FALSE)
```

```
summary(df)
```

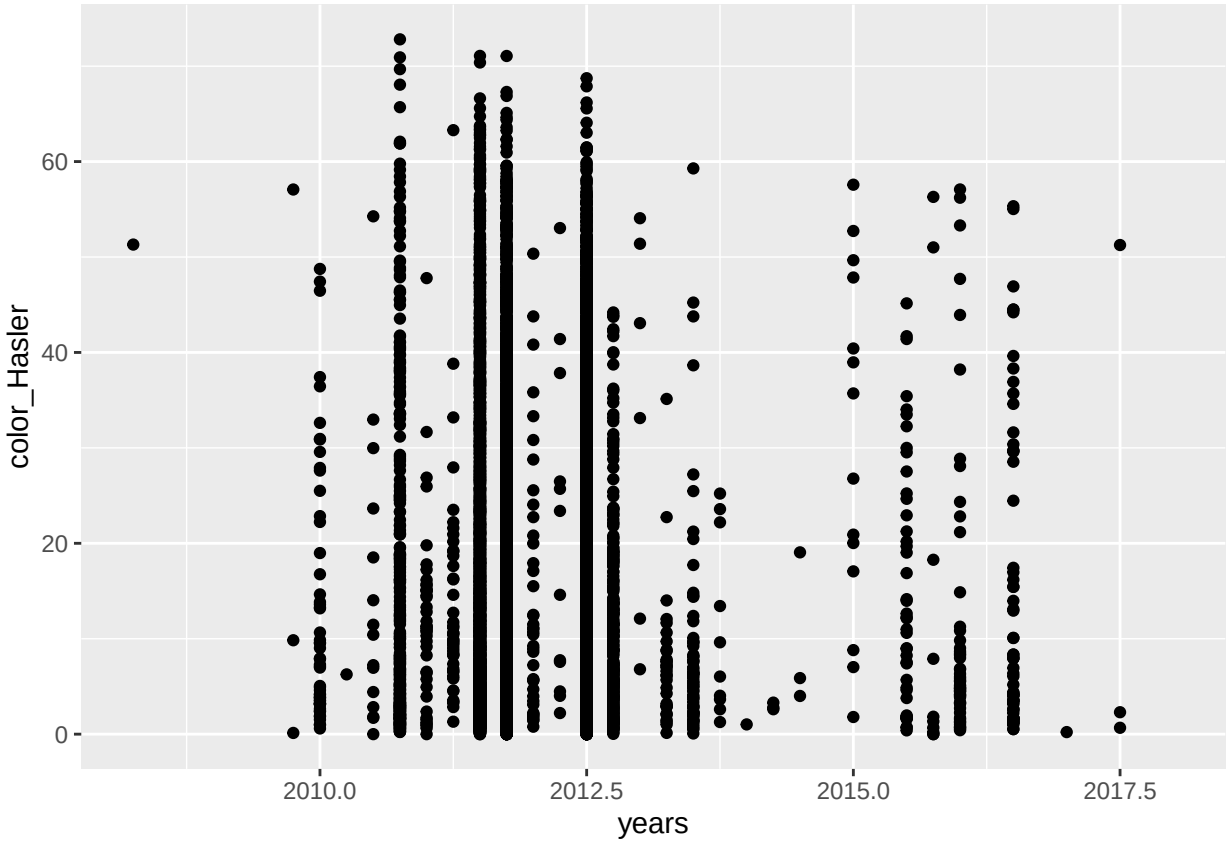
```
##        R                G                 B           color_Hasler
##  Min.   :  1.0   Min.   :  0.9004   Min.   :  1.00   Min.   : 0.000
##  1st Qu.:103.9   1st Qu.: 95.1355   1st Qu.: 98.68   1st Qu.: 3.515
##  Median :164.3   Median :150.5376   Median :149.95   Median : 8.371
##  Mean   :156.2   Mean   :145.9089   Mean   :147.43   Mean   :13.186
##  3rd Qu.:212.2   3rd Qu.:198.6761   3rd Qu.:197.29   3rd Qu.:18.622
##  Max.   :255.0   Max.   :254.9746   Max.   :254.97   Max.   :72.808
##  NA's   :579     NA's   :579        NA's   :579      NA's   :579
##     color_HS            id           gender          masterCategory
##  Min.   :0.0000   Min.   : 1163   Length:15189       Length:15189
##  1st Qu.:0.0477   1st Qu.:10186   Class :character   Class :character
##  Median :0.1136   Median :19790   Mode  :character   Mode  :character
##  Mean   :0.1741   Mean   :23025
##  3rd Qu.:0.2472   3rd Qu.:33943
##  Max.   :0.9346   Max.   :60000
##  NA's   :579
##  subCategory        articleType         baseColour           season
##  Length:15189       Length:15189       Length:15189       Length:15189
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##       year         usage           productDisplayName     link
##  Min.   :2008   Length:15189       Length:15189       Length:15189
##  1st Qu.:2011   Class :character   Class :character   Class :character
##  Median :2011   Mode  :character   Mode  :character   Mode  :character
##  Mean   :2011
```

```
##  3rd Qu.:2012
##  Max.   :2018
##  NA's   :1
##      years          RGB            color_Hasler_premade color_HS_premade
##  Min.   :2008   Length:15189       Min.   :  0.00       Min.   :  0.000
##  1st Qu.:2012   Class :character   1st Qu.: 14.02       1st Qu.:  8.286
##  Median :2012   Mode  :character   Median : 33.28       Median : 19.730
##  Mean   :2012                      Mean   : 52.44       Mean   : 31.079
##  3rd Qu.:2012                      3rd Qu.: 73.95       3rd Qu.: 43.892
##  Max.   :2018                      Max.   :290.29       Max.   :171.611
##  NA's   :1                         NA's   :579          NA's   :579
```

```
colorfulness_stats <- df %>%
  group_by(gender) %>%
  summarise(
    avg_color_Hasler = mean(color_Hasler, na.rm = TRUE),
    avg_color_HS = mean(color_HS, na.rm = TRUE)
  )
```

```
ggplot(df,
       aes(x=years,
           y=color_Hasler)) +
  geom_point()
```

```
## Warning: Removed 580 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



```r
m1 <- lm(color_Hasler~years, df)
m2 <- lm(color_HS~years, df)

summary(m1)
```

```
##
## Call:
## lm(formula = color_Hasler ~ years, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.237  -9.671  -4.811   5.423  59.578
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  88.36684  320.29946   0.276    0.783
## years        -0.03737    0.15919  -0.235    0.814
##
## Residual standard error: 13.09 on 14607 degrees of freedom
##   (580 observations deleted due to missingness)
## Multiple R-squared:  3.772e-06,  Adjusted R-squared:  -6.469e-05
## F-statistic: 0.0551 on 1 and 14607 DF,  p-value: 0.8144
```

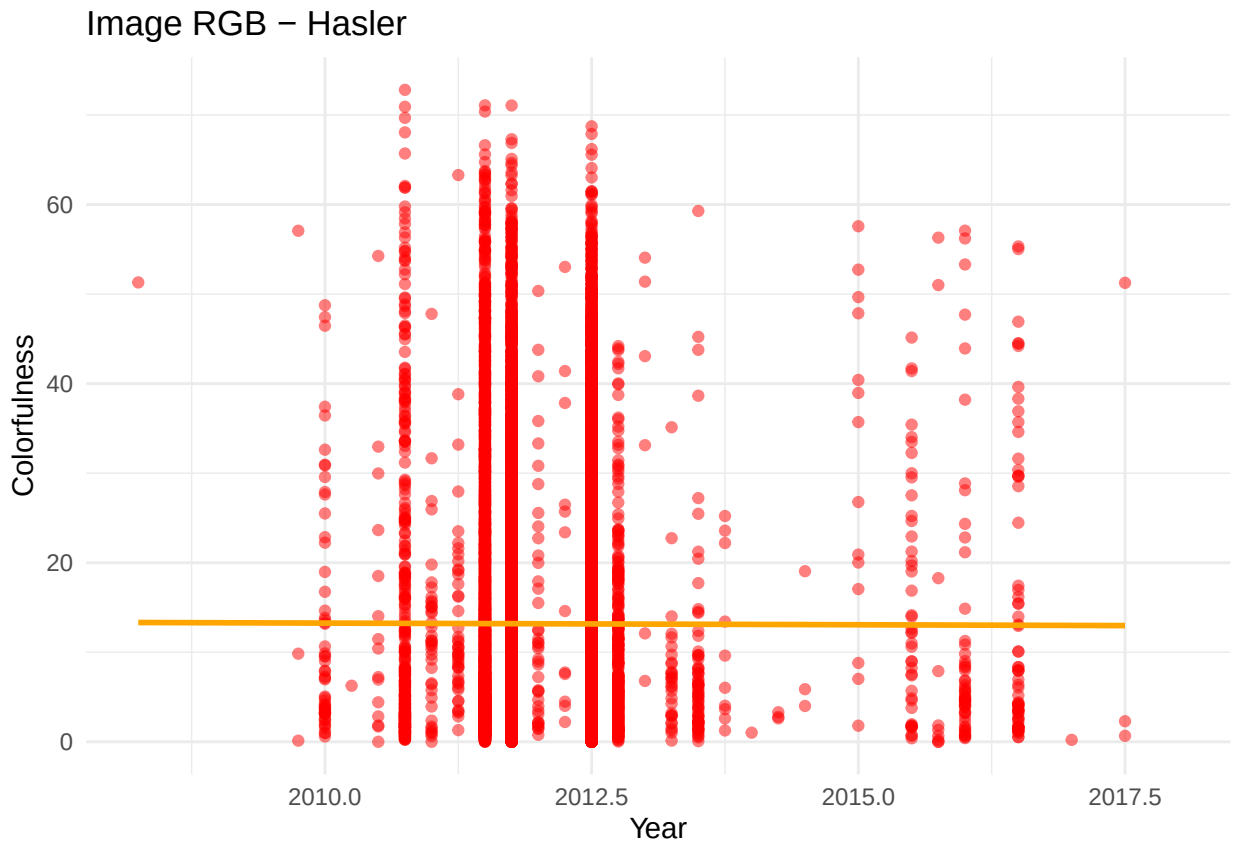
```r
summary(m2)
```

```
##
## Call:
## lm(formula = color_HS ~ years, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.17563 -0.12637 -0.06061  0.07308  0.76113
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.338797   4.118122   0.568    0.570
## years       -0.001076   0.002047  -0.526    0.599
##
## Residual standard error: 0.1683 on 14607 degrees of freedom
##   (580 observations deleted due to missingness)
## Multiple R-squared:  1.892e-05,  Adjusted R-squared:  -4.954e-05
## F-statistic: 0.2763 on 1 and 14607 DF,  p-value: 0.5991
```

```r
ggplot(df, aes(x = years, y = color_Hasler)) +
  geom_point(color = "red", alpha = 0.5) +  # Scatter plot of points
  geom_smooth(method = "lm", color = "orange", se = FALSE) +  # Linear regression line
  labs(title = "Image RGB - Hasler", x = "Year", y = "Colorfulness") +
  theme_minimal()
```

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

## Warning: Removed 580 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

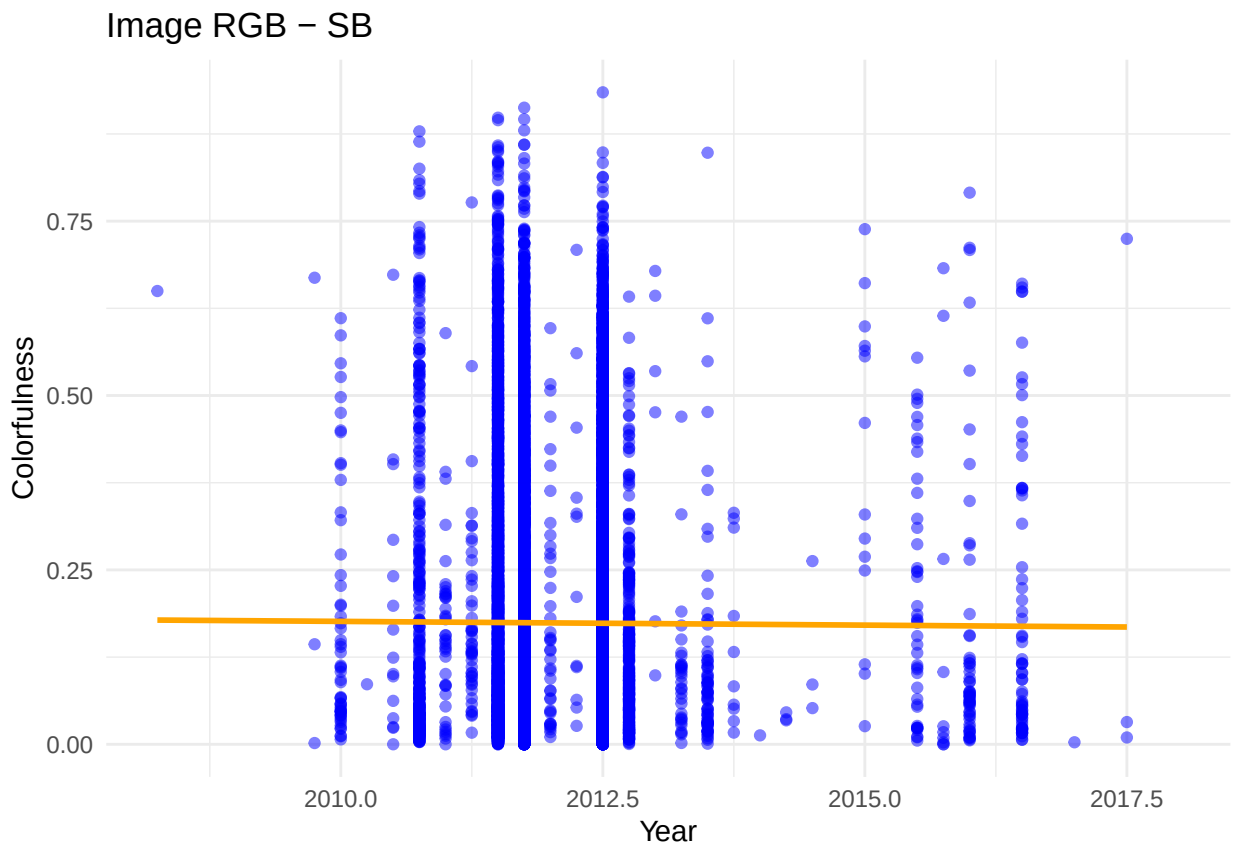
```
## Warning: Removed 580 rows containing missing values or values outside the scale range
## (`geom_point()`).
```




```
ggplot(df, aes(x = years, y = color_HS)) +
  geom_point(color = "blue", alpha = 0.5) +  # Scatter plot of points
  geom_smooth(method = "lm", color = "orange", se = FALSE) +  # Linear regression line
  labs(title = "Image RGB - SB", x = "Year", y = "Colorfulness") +
  theme_minimal()
```

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

```
## Warning: Removed 580 rows containing non-finite outside the scale range
## (`stat_smooth()`).
## Removed 580 rows containing missing values or values outside the scale range
## (`geom_point()`).
```




```
# Assuming you created a 'year_group' or similar categorization
aov_result <- aov(color_Hasler ~ years, data = df)
summary(aov_result)
```

```
##                Df  Sum Sq Mean Sq F value Pr(>F)
## years           1       9    9.44   0.055  0.814
## Residuals   14607 2502781  171.34
## 580 observations deleted due to missingness
```

```
cor.test(df$years, df$color_Hasler)
```

```
##
##  Pearson's product-moment correlation
##
## data:  df$years and df$color_Hasler
## t = -0.23473, df = 14607, p-value = 0.8144
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.01815764  0.01427428
## sample estimates:
##          cor
## -0.001942192
```

```r
cor.test(df$years, df$color_Hasler, method = "spearman")
```

```
## Warning in cor.test.default(df$years, df$color_Hasler, method = "spearman"):
## Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  df$years and df$color_Hasler
## S = 5.2069e+11, p-value = 0.809
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##          rho
## -0.002000114
```

```r
m3 <- lm(color_Hasler_premade~years, df)
m4 <- lm(color_HS_premade~years, df)

summary(m3)
```

```
##
## Call:
## lm(formula = color_Hasler_premade ~ years, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -52.59 -38.42 -19.17  21.48 237.71
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  287.3534  1274.7401   0.225    0.822
## years         -0.1168     0.6336  -0.184    0.854
##
## Residual standard error: 52.1 on 14607 degrees of freedom
##   (580 observations deleted due to missingness)
## Multiple R-squared:  2.325e-06,  Adjusted R-squared:  -6.613e-05
## F-statistic: 0.03396 on 1 and 14607 DF,  p-value: 0.8538
```

```r
summary(m4)
```

```
##
## Call:
## lm(formula = color_HS_premade ~ years, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -31.20 -22.80 -11.34  12.78 140.43
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 208.28264  754.95308   0.276    0.783
```

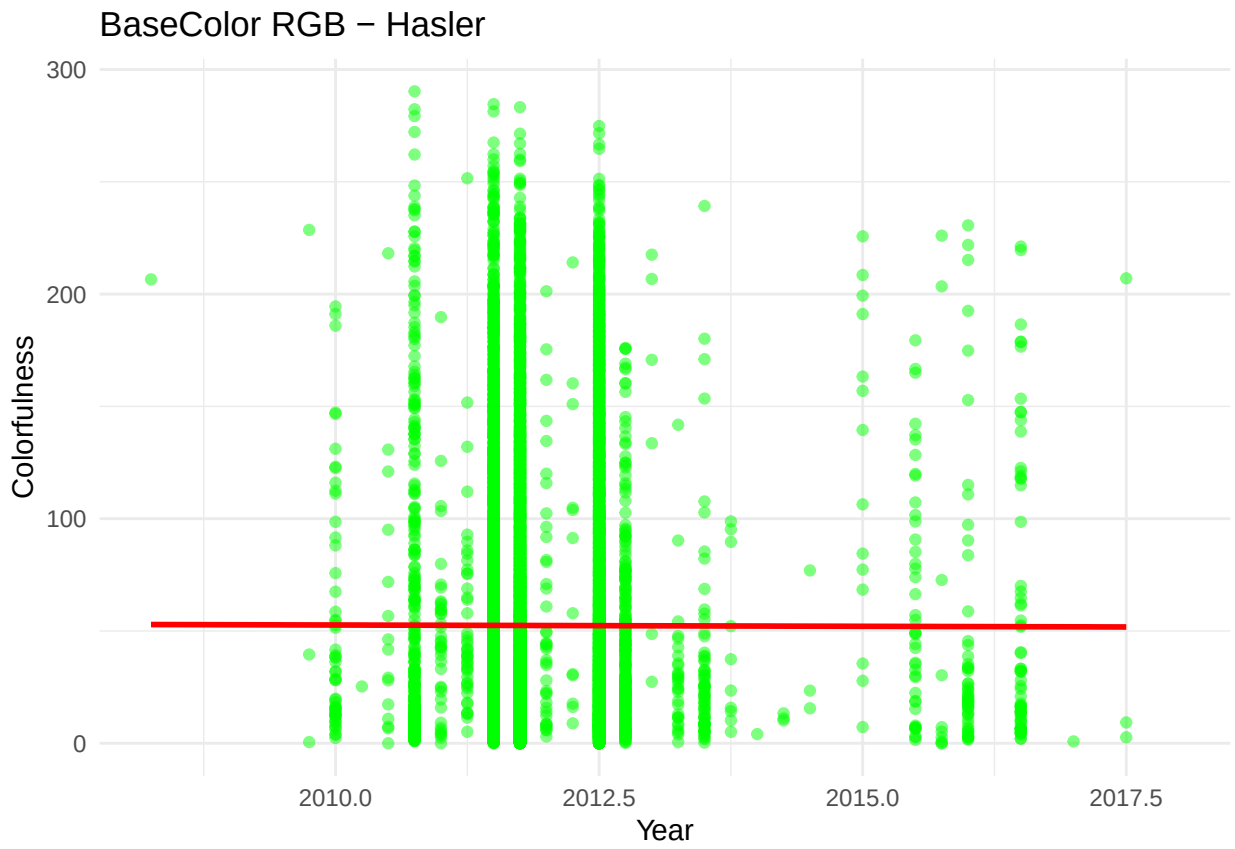
```
## years        -0.08808    0.37522  -0.235    0.814
##
## Residual standard error: 30.85 on 14607 degrees of freedom
##   (580 observations deleted due to missingness)
## Multiple R-squared:  3.772e-06,  Adjusted R-squared:  -6.469e-05
## F-statistic: 0.0551 on 1 and 14607 DF,  p-value: 0.8144
```

```r
library(ggplot2)

# Plotting for color_Hasler vs. year
ggplot(df, aes(x = years, y = color_Hasler_premade)) +
  geom_point(color = "green", alpha = 0.5) +  # Scatter plot of points
  geom_smooth(method = "lm", color = "red", se = FALSE) +  # Linear regression line
  labs(title = "BaseColor RGB - Hasler", x = "Year", y = "Colorfulness") +
  theme_minimal()
```

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

```
## Warning: Removed 580 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

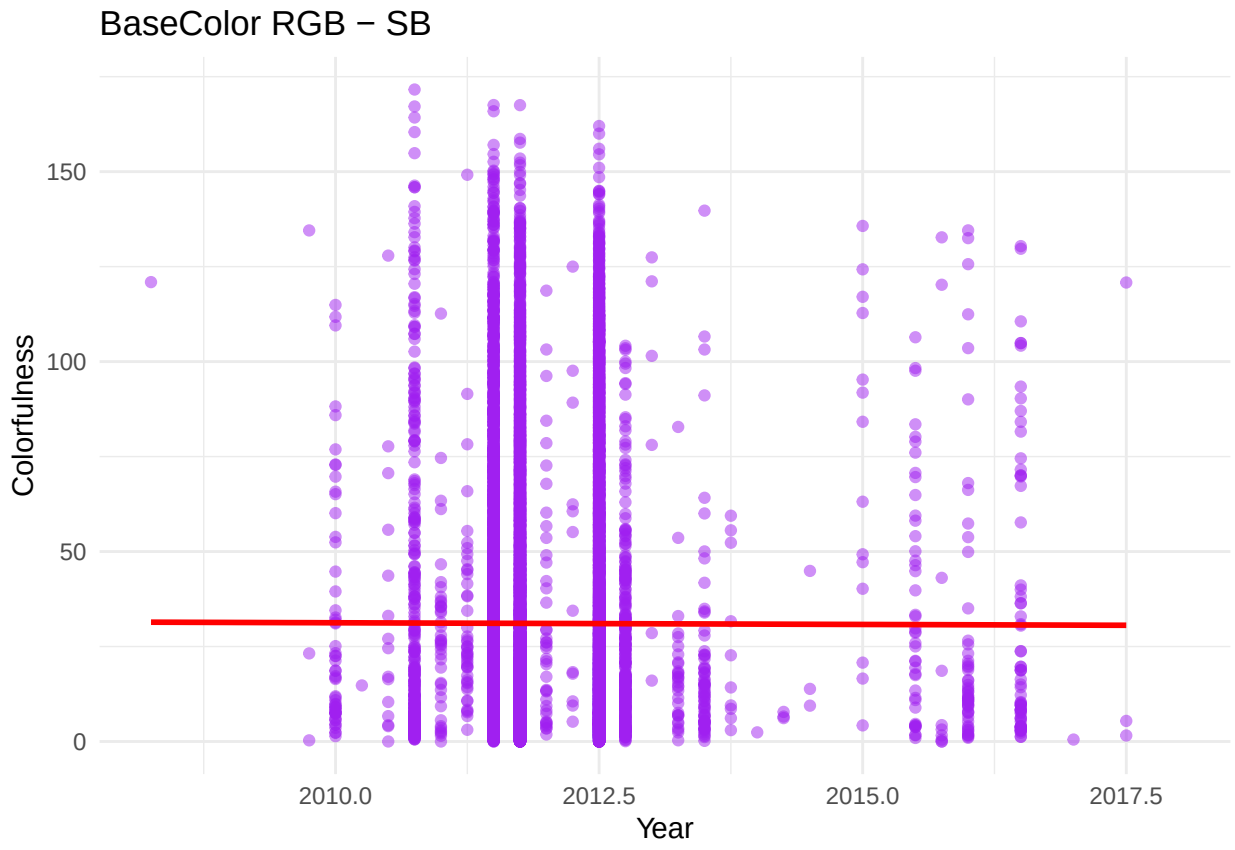
```
## Warning: Removed 580 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



```r
# Plotting for color_HS vs. year
ggplot(df, aes(x = years, y = color_HS_premade)) +
  geom_point(color = "purple", alpha = 0.5) +  # Scatter plot of points
  geom_smooth(method = "lm", color = "red", se = FALSE) +  # Linear regression line
  labs(title = "BaseColor RGB - SB", x = "Year", y = "Colorfulness") +
  theme_minimal()
```

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

```
## Warning: Removed 580 rows containing non-finite outside the scale range
## (`stat_smooth()`).
## Removed 580 rows containing missing values or values outside the scale range
## (`geom_point()`).
```




```r
# Fit the linear model
m5 <- lm(color_Hasler ~ years + gender, df)
summary(m5)
```

```
##
## Call:
## lm(formula = color_Hasler ~ years + gender, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -23.445  -9.201  -4.561   5.222  60.594
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  659.0505   320.6290   2.055   0.0398 *
## years         -0.3196     0.1593  -2.006   0.0449 *
## genderGirls    1.5509     0.8868   1.749   0.0803 .
## genderMen     -4.2266     0.5455  -7.749 9.89e-15 ***
## genderUnisex   7.2441     1.5445   4.690 2.75e-06 ***
## genderWomen   -1.3744     0.5550  -2.476   0.0133 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.97 on 14603 degrees of freedom
##   (580 observations deleted due to missingness)
## Multiple R-squared:  0.01845,    Adjusted R-squared:  0.01811
## F-statistic:  54.9 on 5 and 14603 DF,  p-value: < 2.2e-16
```

```r
unique(df$gender)
```

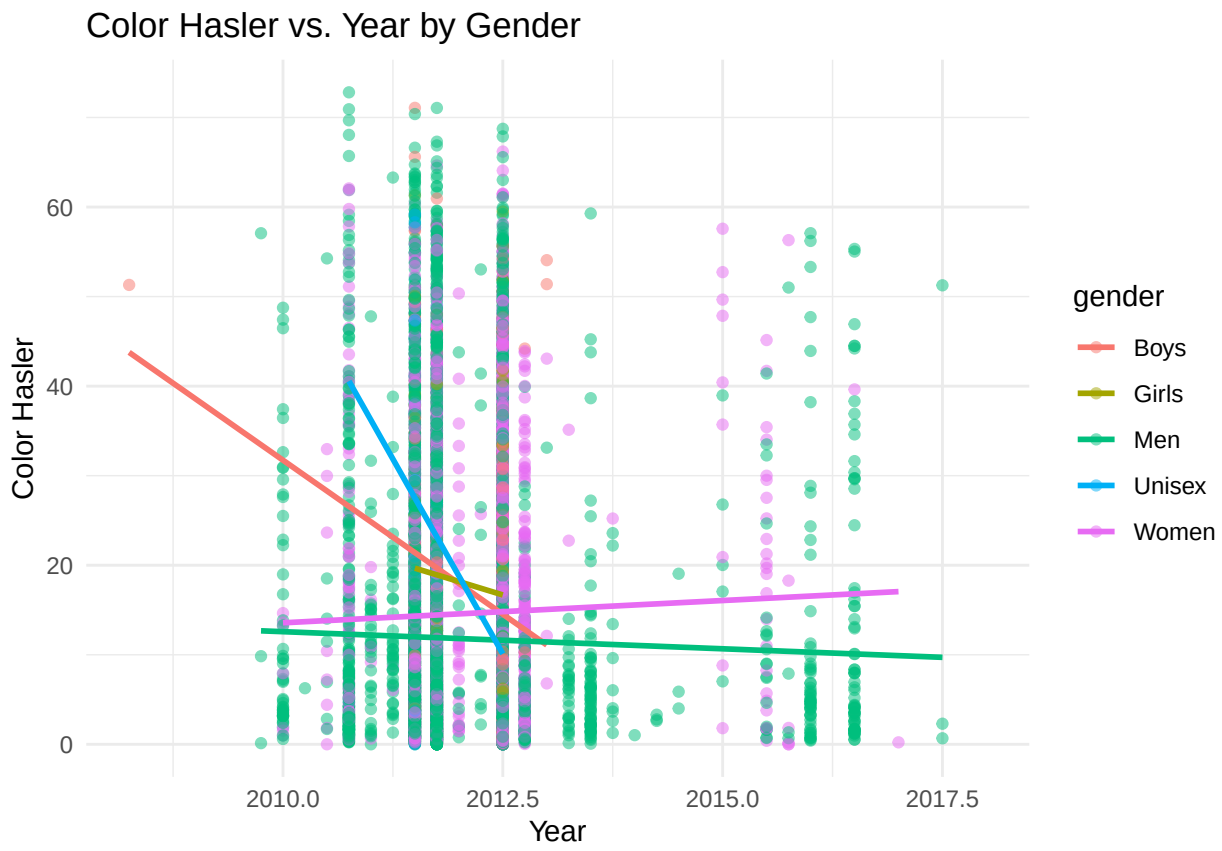
```
## [1] "Men"    "Women"  "Girls"  "Boys"   "Unisex"
```

```r
ggplot(df, aes(x = years, y = color_Hasler, color = gender)) +
  geom_point(alpha = 0.5) +  # Scatter plot for all genders
  geom_smooth(method = "lm", aes(group = gender), se = FALSE) +  # Separate regression lines for each g
  labs(title = "Color Hasler vs. Year by Gender", x = "Year", y = "Color Hasler") +
  theme_minimal()
```

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

```
## Warning: Removed 580 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

```
## Warning: Removed 580 rows containing missing values or values outside the scale range
## (`geom_point()`).
```

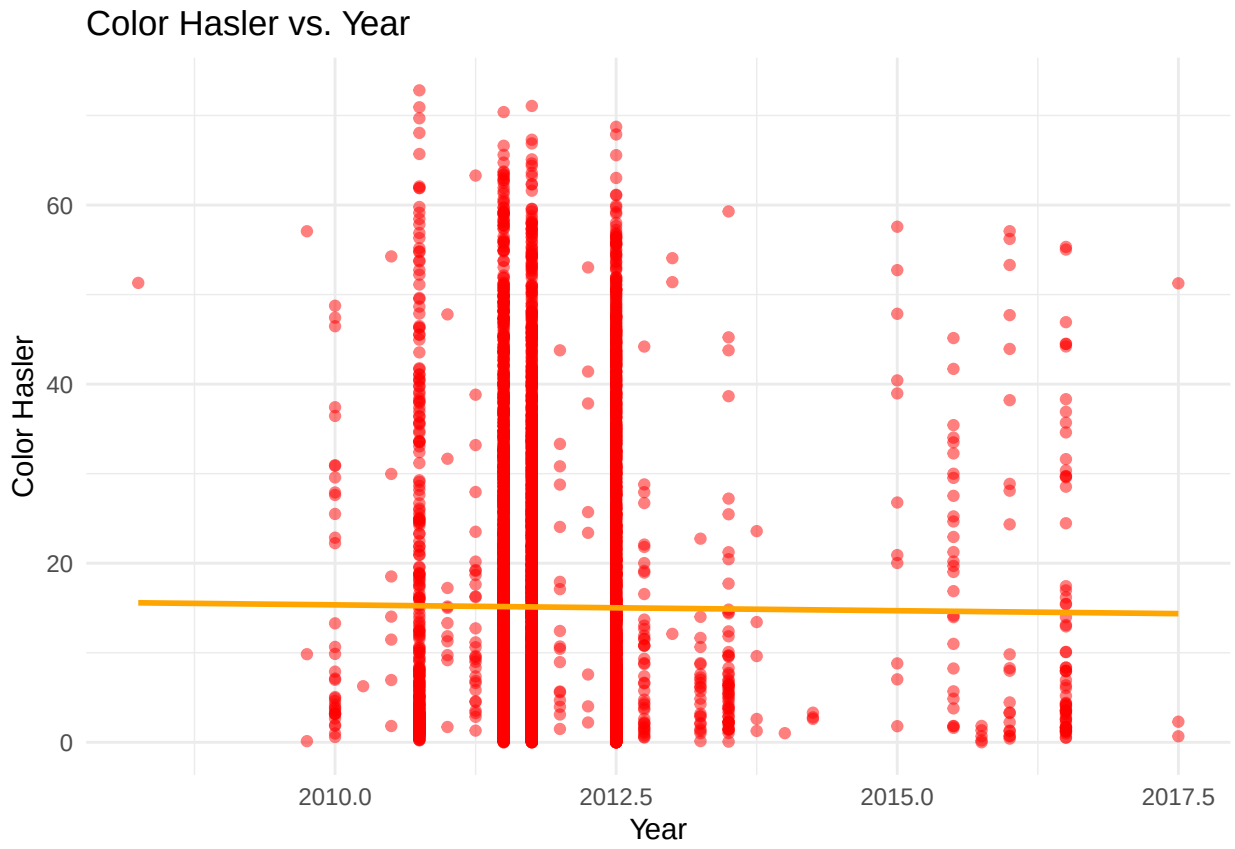



```
tshirt_df <- df[df$articleType == "Tshirts", ]

ggplot(tshirt_df, aes(x = years, y = color_Hasler)) +
  geom_point(color = "red", alpha = 0.5) +  # Scatter plot of points
  geom_smooth(method = "lm", color = "orange", se = FALSE) +  # Linear regression line
  labs(title = "Color Hasler vs. Year", x = "Year", y = "Color Hasler") +
  theme_minimal()
```

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

```
## Warning: Removed 284 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

```
## Warning: Removed 284 rows containing missing values or values outside the scale range
## (`geom_point()`).
```




```
m5 <- lm(color_Hasler ~ years , tshirt_df)
summary(m5)
```

```
##
## Call:
## lm(formula = color_Hasler ~ years, data = tshirt_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.253 -11.402  -5.610   7.658  57.557
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 278.2279   452.4770   0.615    0.539
## years        -0.1308     0.2249  -0.582    0.561
##
## Residual standard error: 14.66 on 6713 degrees of freedom
##   (284 observations deleted due to missingness)
## Multiple R-squared:  5.038e-05,  Adjusted R-squared:  -9.858e-05
## F-statistic: 0.3382 on 1 and 6713 DF,  p-value: 0.5609
```