Vincent van Gogh Data Visualizations

Alyssa W. Zhang

2024-03-13

```
knitr::opts_chunk$set(echo = TRUE)

# Load libraries
library(ggplot2)
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(mgcv)
```

```
## Warning: package 'mgcv' was built under R version 4.3.2

## Loading required package: nlme

##
## Attaching package: 'nlme'

## The following object is masked from 'package:dplyr':
##
##     collapse

## This is mgcv 1.9-0. For overview type 'help("mgcv-package")'.
```

```
library(tidyr)
```

```
## Warning: package 'tidyr' was built under R version 4.3.2
```

```r
# Load CSV file output from Python program.
data <- read.csv("C:/Users/Alyss/PycharmProjects/pythonProject/van_gogh_output_new.csv")

# Add Artistic Period variable
data <- data %>%
  mutate(Artistic.Period = case_when(
    Artistic.Location %in% c("The Hague", "Scheveningen", "Nieuw-Amsterdam",
                             "Drente") ~ "Earliest Paintings (1881-83)",
    Artistic.Location %in% c("Nuenen", "Antwerp", "Amsterdam")
    ~ "Nuenen/Antwerp (1883-86)",
    Artistic.Location == "Paris" ~ "Paris (1886-88)",
    Artistic.Location == "Arles" ~ "Arles (1888-89)",
    Artistic.Location == "Saint-Remy" ~ "Saint-Remy (1889-90)",
    Artistic.Location == "Auvers-sur-Oise" ~ "Auvers-sur-Oise (1890)"
  ))

data$Artistic.Period <- factor(data$Artistic.Period, levels = c(
  "Earliest Paintings (1881-83)",
  "Nuenen/Antwerp (1883-86)",
  "Paris (1886-88)",
  "Arles (1888-89)",
  "Saint-Remy (1889-90)",
  "Auvers-sur-Oise (1890)"
))

# Define a function to convert time-related information from Origin to fractions
convert_to_fraction <- function(origin) {
  if (grepl("December", origin)) {
    return(23/24)
  } else if (grepl("November", origin)) {
    return(21/24)
  } else if (grepl("October", origin)) {
    return(19/24)
  } else if (grepl("September", origin)) {
    return(17/24)
  } else if (grepl("August", origin)) {
    return(15/24)
  } else if (grepl("July", origin)) {
    return(13/24)
  } else if (grepl("June", origin)) {
    return(11/24)
  } else if (grepl("May", origin)) {
    return(9/24)
  } else if (grepl("April", origin)) {
    return(7/24)
  } else if (grepl("March", origin)) {
    return(5/24)
  } else if (grepl("February", origin)) {
    return(3/24)
  } else if (grepl("January", origin)) {
    return(1/24)
  } else if (grepl("Autumn", origin)) {
    return (19/24)
```

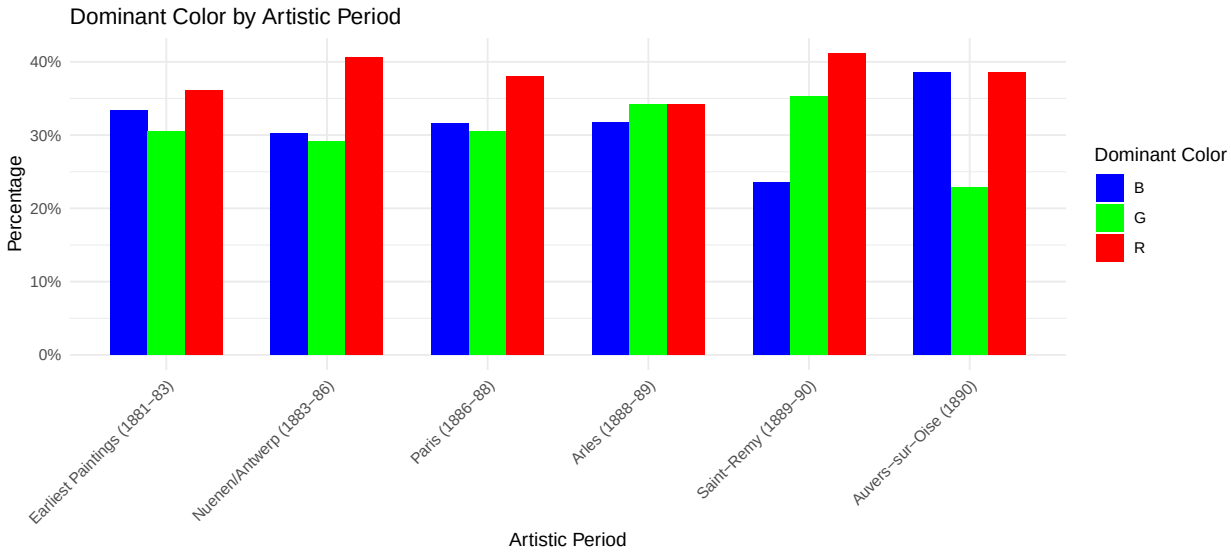
```
  } else if (grepl("Summer", origin)){
    return (13/24)
  } else if (grepl("Spring", origin)){
    return (7/24)
  } else if (grepl("Winter", origin)) {
    return (1/24)
  } else if (grepl("second half", origin)) {
    return (3/4)
  } else if (grepl("first half", origin)) {
    return (1/4)
  } else{
    return(1/2)
  }
}

# Apply convert_to_fraction to modify Creation Date
data$Creation.Date <- data$Creation.Date + sapply(data$Origin, convert_to_fraction)

# Separate data by time of ear mutilation
# (December 23, 1888, which approximately matches to Creation Date of 1889)
data <- mutate(data,
               Ear.Mutilation = ifelse(Creation.Date > 1889, "After", "Before"))
data$Ear.Mutilation <- factor(data$Ear.Mutilation)
```

Visualization for Dominant Color

```
# Calculate the percentage for each Dominant Color category within an Artistic Period
percentage_data <- data %>%
  group_by(Artistic.Period, Dominant.Color) %>%
  summarise(Frequency = n()) %>%
  group_by(Artistic.Period) %>%
  mutate(Percentage = Frequency / sum(Frequency) * 100)
```

```
## `summarise()` has grouped output by 'Artistic.Period'. You can override using
## the `.groups` argument.
```

```
# Plot the side-by-side bar chart with percentages
ggplot(percentage_data,
       aes(x = Artistic.Period, y = Percentage, fill = Dominant.Color)) +
  geom_bar(stat = "identity", position = "dodge", width = 0.7) +
  labs(title = "Dominant Color by Artistic Period",
       x = "Artistic Period",
       y = "Percentage",
       fill = "Dominant Color") +
  scale_fill_manual(values = c("R" = "red", "G" = "green", "B" = "blue")) +
  scale_y_continuous(labels = scales::percent_format(scale = 1)) +
  # Show percentages on the y-axis
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

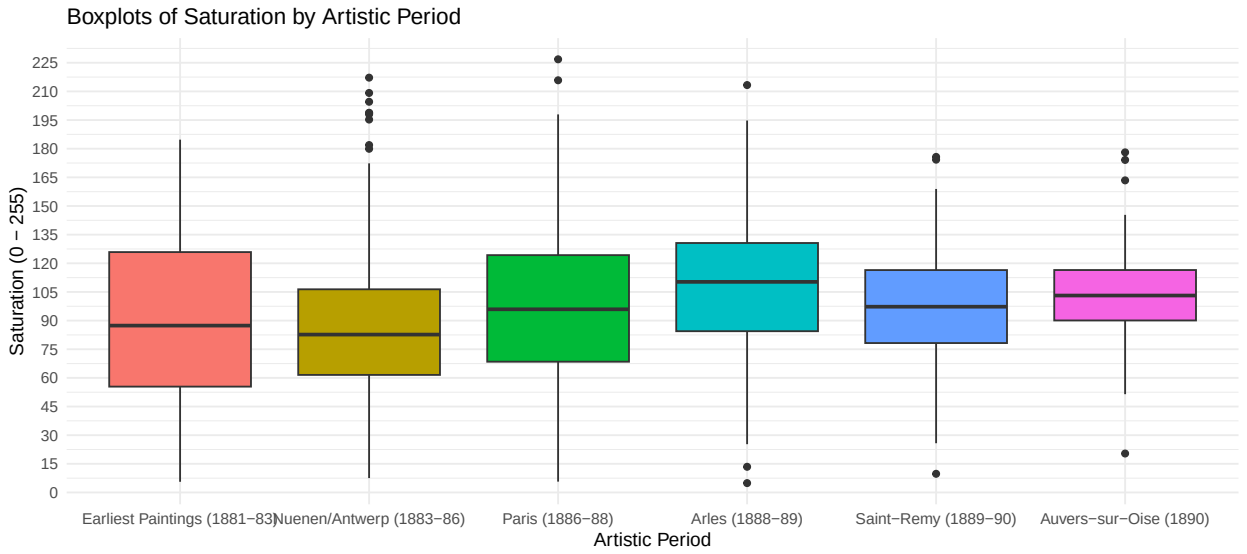



```
# Make a table displaying all percentage values from the bar chart
percentage_table <- percentage_data %>%
  select(Artistic.Period, Dominant.Color, Percentage) %>%
  spread(Dominant.Color, Percentage)

print(percentage_table)
```

```
## # A tibble: 6 x 4
## # Groups:   Artistic.Period [6]
##   Artistic.Period                B     G     R
##   <fct>                      <dbl> <dbl> <dbl>
## 1 Earliest Paintings (1881-83)  33.3  30.6  36.1
## 2 Nuenen/Antwerp (1883-86)      30.2  29.2  40.6
## 3 Paris (1886-88)               31.6  30.5  38.0
## 4 Arles (1888-89)               31.7  34.2  34.2
## 5 Saint-Remy (1889-90)          23.5  35.3  41.2
## 6 Auvers-sur-Oise (1890)        38.6  22.9  38.6
```

Visualizations for Saturation

```
# Plot the side-by-side boxplot
ggplot(data, aes(x = Artistic.Period, y = Saturation, fill = Artistic.Period)) +
  geom_boxplot() +
  labs(title = "Boxplots of Saturation by Artistic Period",
       x = "Artistic Period",
       y = "Saturation (0 - 255)") +
  scale_y_continuous(breaks = seq(0, 255, by = 15)) +
  guides(fill = "none") +
  theme_minimal()
```




```
# Make a table providing a numeric summary of the side-by-side box plot
saturation_summary_table <- data %>%
  group_by(Artistic.Period) %>%
  summarise(
    Min = min(Saturation, na.rm = TRUE),
    Q1 = quantile(Saturation, 0.25, na.rm = TRUE),
    Median = median(Saturation, na.rm = TRUE),
    Q3 = quantile(Saturation, 0.75, na.rm = TRUE),
    Max = max(Saturation, na.rm = TRUE),
    IQR = Q3 - Q1
  )

print(saturation_summary_table)
```

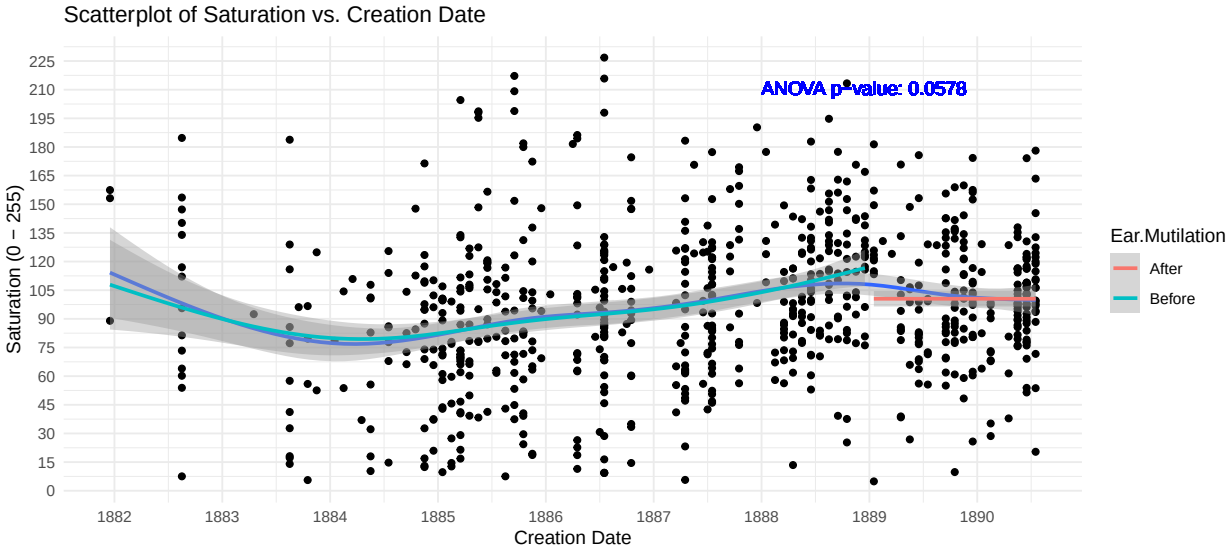
```
## # A tibble: 6 x 7
##   Artistic.Period                Min    Q1 Median    Q3   Max   IQR
##   <fct>                        <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl>
## 1 Earliest Paintings (1881-83)  5.59  55.4   87.4  126.  185.  70.4
## 2 Nuenen/Antwerp (1883-86)      7.53  61.6   82.7  106.  217.  44.8
## 3 Paris (1886-88)               5.68  68.5   95.9  124.  227.  55.8
## 4 Arles (1888-89)               4.91  84.4  110.   131.  213.  46.2
## 5 Saint-Remy (1889-90)          9.78  78.2   97.3  116.  176.  38.2
## 6 Auvers-sur-Oise (1890)       20.4   90.1  103.   116.  178.  26.4
```

```
# Plot the scatter plot with smooth curves and ANOVA
ggplot(data, aes(x = Creation.Date, y = Saturation)) +
  geom_point(method = 'gam') +
  geom_smooth(method ='gam') +
  geom_smooth(aes(color = Ear.Mutilation), method = 'gam') +
  labs(title = "Scatterplot of Saturation vs. Creation Date", x = "Creation Date",
       y = "Saturation (0 - 255)") +
  scale_x_continuous(breaks = seq(1881, 1890, by = 1), labels = as.character(seq(1881, 1890, by = 1))) +
  scale_y_continuous(breaks = seq(0, 255, by = 15)) +
  theme_minimal() +
```

```
  # Perform ANOVA and display p-value
  geom_text(aes(x = 1888, y = 200,
                label = paste("ANOVA p-value:",
                              round(anova(
                                lm(Saturation ~ Ear.Mutilation, data = filter(
                                  data, Creation.Date > (1889 - 1) &
                                    Creation.Date < 1889 + 1)))$`Pr(>F)`[1], 4))),
            hjust = 0, vjust = -1, color = "blue")
```

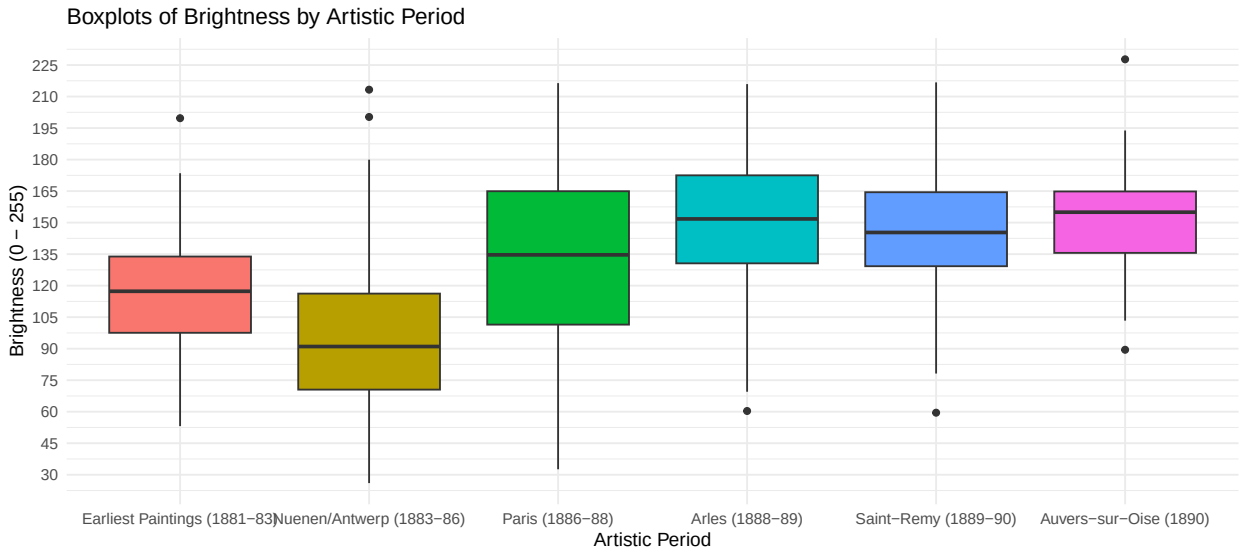
```
## Warning in geom_point(method = "gam"): Ignoring unknown parameters: `method`
```

```
## `geom_smooth()` using formula = 'y ~ s(x, bs = "cs")'
## `geom_smooth()` using formula = 'y ~ s(x, bs = "cs")'
```



Visualizations for Brightness

```
# Plot the side-by-side box plot
ggplot(data, aes(x = Artistic.Period, y = Brightness, fill = Artistic.Period)) +
  geom_boxplot() +
  labs(title = "Boxplots of Brightness by Artistic Period",
       x = "Artistic Period",
       y = "Brightness (0 - 255)") +
  scale_y_continuous(breaks = seq(0, 255, by = 15)) +
  guides(fill = "none") +
  theme_minimal()
```



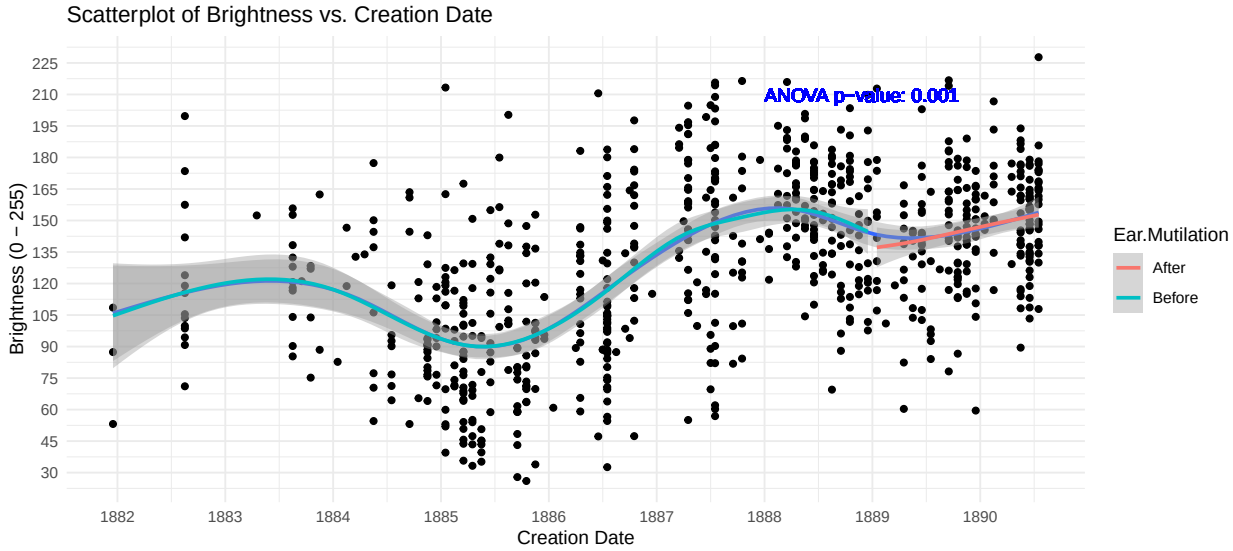

```
# Make a table providing a numeric summary of the side-by-side box plot
brightness_summary_table <- data %>%
  group_by(Artistic.Period) %>%
  summarise(
    Min = min(Brightness, na.rm = TRUE),
    Q1 = quantile(Brightness, 0.25, na.rm = TRUE),
    Median = median(Brightness, na.rm = TRUE),
    Q3 = quantile(Brightness, 0.75, na.rm = TRUE),
    Max = max(Brightness, na.rm = TRUE),
    IQR = Q3 - Q1
  )

print(brightness_summary_table)
```

```
## # A tibble: 6 x 7
##   Artistic.Period                Min    Q1 Median    Q3   Max   IQR
##   <fct>                        <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl>
## 1 Earliest Paintings (1881-83)  53.2  97.6  117.   134.  200.  36.3
## 2 Nuenen/Antwerp (1883-86)      26.0  70.5   91.0  116.  213.  45.7
## 3 Paris (1886-88)               32.6 101.   135.   165.  216.  63.5
## 4 Arles (1888-89)               60.3 131.   152.   173.  216.  41.9
## 5 Saint-Remy (1889-90)          59.5 129.   145.   164.  217.  35.2
## 6 Auvers-sur-Oise (1890)        89.5 136.   155.   165.  228.  29.2
```

```
# Plot the scatter plot with smooth curves and ANOVA
ggplot(data, aes(x = Creation.Date, y = Brightness)) +
  geom_point() +
  geom_smooth(method ='gam')+
  geom_smooth(aes(color = Ear.Mutilation), method ='gam') +
  labs(title = "Scatterplot of Brightness vs. Creation Date", x = "Creation Date",
       y = "Brightness (0 - 255)") +
  scale_x_continuous(breaks = seq(1881, 1890, by = 1), labels = as.character(seq(1881, 1890, by = 1))) +
  scale_y_continuous(breaks = seq(0, 255, by = 15)) +
  theme_minimal() +
```

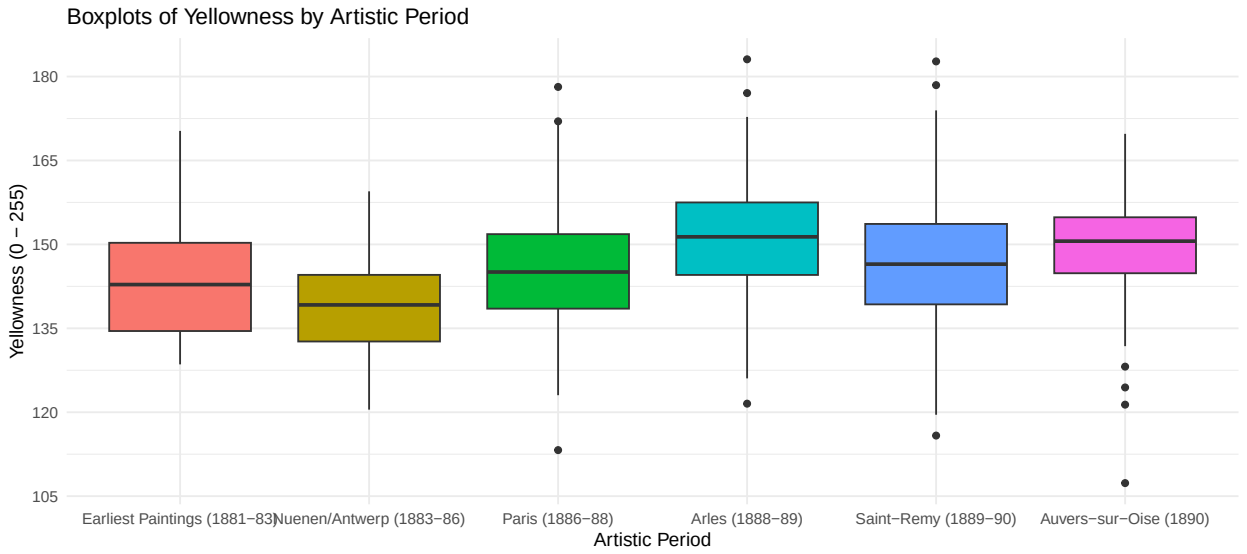
```
  # Perform ANOVA and display p-value
  geom_text(aes(x = 1888, y = 200,
                label = paste("ANOVA p-value:",
                              round(anova(
                                lm(Brightness ~ Ear.Mutilation, data = filter(
                                  data, Creation.Date > (1889 - 1) &
                                    Creation.Date < 1889 + 1)))$`Pr(>F)`[1], 4))),
            hjust = 0, vjust = -1, color = "blue")
```

```
## `geom_smooth()` using formula = 'y ~ s(x, bs = "cs")'
## `geom_smooth()` using formula = 'y ~ s(x, bs = "cs")'
```



Visualizations for Yellowness

```
# Plot the side-by-side box plot
ggplot(data, aes(x = Artistic.Period, y = Yellowness, fill = Artistic.Period)) +
  geom_boxplot() +
  labs(title = "Boxplots of Yellowness by Artistic Period",
       x = "Artistic Period",
       y = "Yellowness (0 - 255)") +
  scale_y_continuous(breaks = seq(0, 255, by = 15)) +
  guides(fill = "none") +
  theme_minimal()
```



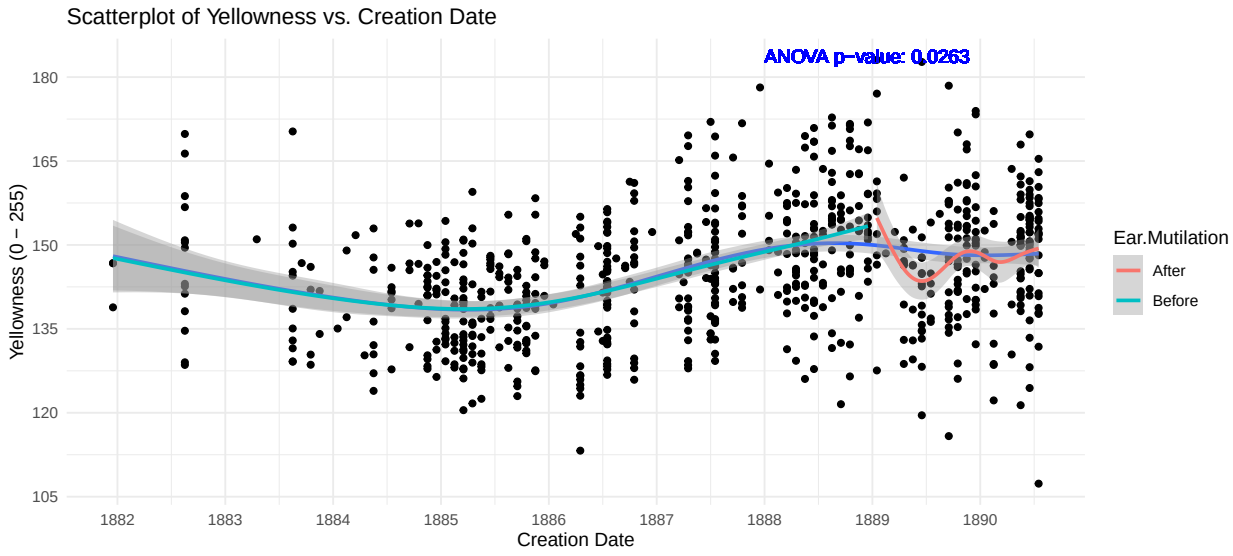

```
# Make a table providing a numeric summary of the side-by-side box plot
yellowness_summary_table <- data %>%
  group_by(Artistic.Period) %>%
  summarise(
    Min = min(Yellowness, na.rm = TRUE),
    Q1 = quantile(Yellowness, 0.25, na.rm = TRUE),
    Median = median(Yellowness, na.rm = TRUE),
    Q3 = quantile(Yellowness, 0.75, na.rm = TRUE),
    Max = max(Yellowness, na.rm = TRUE),
    IQR = Q3 - Q1
  )

print(yellowness_summary_table)
```

```
## # A tibble: 6 x 7
##   Artistic.Period                Min    Q1 Median    Q3   Max   IQR
##   <fct>                        <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl>
## 1 Earliest Paintings (1881-83)  129.  135.   143.  150.  170.  15.8
## 2 Nuenen/Antwerp (1883-86)      120.  133.   139.  145.  159.  11.9
## 3 Paris (1886-88)               113.  139.   145.  152.  178.  13.3
## 4 Arles (1888-89)               122.  145.   151.  157.  183.  13.0
## 5 Saint-Remy (1889-90)          116.  139.   146.  154.  183.  14.4
## 6 Auvers-sur-Oise (1890)        107.  145.   151.  155.  170.  10.0
```

```
# Plot with scatter plot with smooth curves and ANOVA
ggplot(data, aes(x = Creation.Date, y = Yellowness)) +
  geom_point() +
  geom_smooth(method='gam')+
  geom_smooth(aes(color=Ear.Mutilation), method='gam') +
  labs(title = "Scatterplot of Yellowness vs. Creation Date", x = "Creation Date",
       y = "Yellowness (0 - 255)") +
  scale_x_continuous(breaks = seq(1881, 1890, by = 1), labels = as.character(seq(1881, 1890, by = 1))) +
  scale_y_continuous(breaks = seq(0, 255, by = 15)) +
  theme_minimal() +
  # Perform ANOVA and display p-value
```

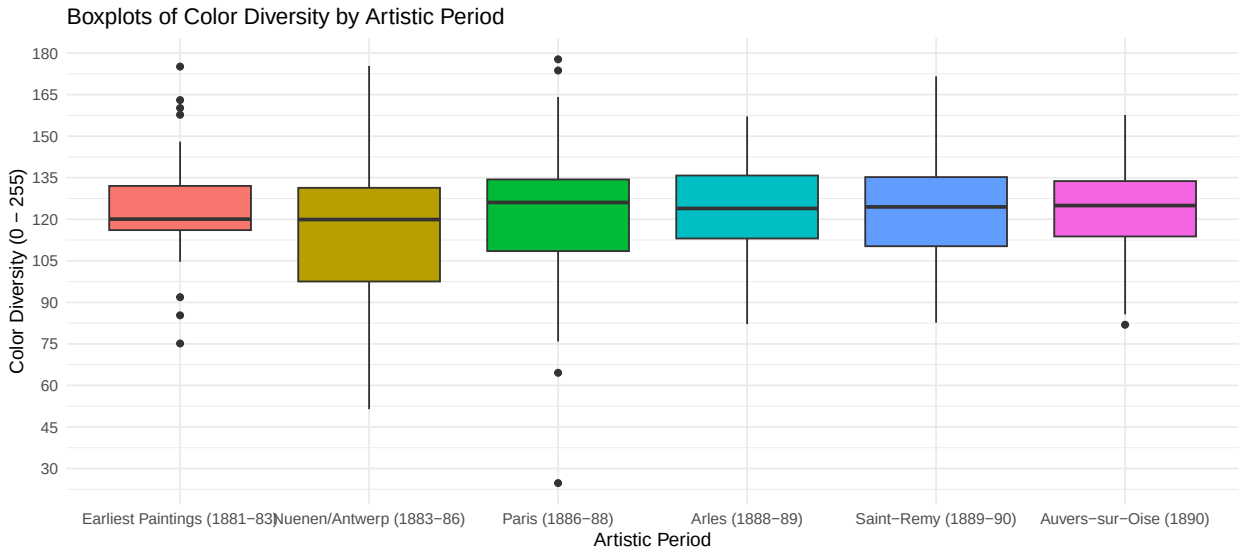
```
  geom_text(aes(x = 1888, y = 180,
                label = paste("ANOVA p-value:",
                              round(anova(
                                lm(Yellowness ~ Ear.Mutilation, data = filter(
                                  data, Creation.Date > (1889 - 1) &
                                    Creation.Date < 1889 + 1)))$`Pr(>F)`[1], 4))),
            hjust = 0, vjust = -1, color = "blue")
```

```
## `geom_smooth()` using formula = 'y ~ s(x, bs = "cs")'
## `geom_smooth()` using formula = 'y ~ s(x, bs = "cs")'
```



Visualizations for Color Diversity

```
# Plot the side-by-side box plot
ggplot(data, aes(x = Artistic.Period, y = Color.Diversity, fill = Artistic.Period)) +
  geom_boxplot() +
  labs(title = "Boxplots of Color Diversity by Artistic Period",
       x = "Artistic Period",
       y = "Color Diversity (0 - 255)") +
  scale_y_continuous(breaks = seq(0, 255, by = 15)) +
  guides(fill = "none") +
  theme_minimal()
```




```
# Make a table providing a numeric summary of the side-by-side box plot
color_diversity_summary_table <- data %>%
  group_by(Artistic.Period) %>%
  summarise(
    Min = min(Color.Diversity, na.rm = TRUE),
    Q1 = quantile(Color.Diversity, 0.25, na.rm = TRUE),
    Median = median(Color.Diversity, na.rm = TRUE),
    Q3 = quantile(Color.Diversity, 0.75, na.rm = TRUE),
    Max = max(Color.Diversity, na.rm = TRUE),
    IQR = Q3 - Q1
  )

print(color_diversity_summary_table)
```

```
## # A tibble: 6 x 7
##   Artistic.Period                Min    Q1 Median    Q3   Max   IQR
##   <fct>                        <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl>
## 1 Earliest Paintings (1881-83)  75.1 116.    120.  132.  175.  16.0
## 2 Nuenen/Antwerp (1883-86)      51.4  97.5   120.  131.  175.  33.8
## 3 Paris (1886-88)               24.7 109.    126.  134.  178.  25.9
## 4 Arles (1888-89)               82.2 113.    124.  136.  157.  22.7
## 5 Saint-Remy (1889-90)          82.7 110.    124.  135.  172.  25.0
## 6 Auvers-sur-Oise (1890)        81.9 114.    125.  134.  158.  20.0
```

```
# Plot the scatter plot with smooth curves and ANOVA
ggplot(data, aes(x = Creation.Date, y = Color.Diversity)) +
  geom_point() +
  geom_smooth() +
  geom_smooth(aes(color=Ear.Mutilation)) +
  labs(title = "Scatterplot of Color Diversity vs. Creation Date", x = "Creation Date",
       y = "Color Diversity (0 - 255)") +
  scale_x_continuous(breaks = seq(1881, 1890, by = 1), labels = as.character(seq(1881, 1890, by = 1))) +
  scale_y_continuous(breaks = seq(0, 255, by = 15)) +
  theme_minimal() +
```

```
  # Perform ANOVA and display p-value
  geom_text(aes(x = 1888, y = 170,
                label = paste("ANOVA p-value:",
                              round(anova(
                                lm(Color.Diversity ~ Ear.Mutilation, data = filter(
                                  data, Creation.Date > (1889 - 1) &
                                    Creation.Date < 1889 + 1)))$`Pr(>F)`[1], 4))),
            hjust = 0, vjust = -1, color = "blue")
```

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