Lab 12 - Statistics, Coordinates, Facets, and Themes

Johnathan Hsu

November 21st, 2017

Complete the following exercises below. Knit together the PDF document and commit both the Lab 12 RMDfile and the PDF document to Git. Push the changes to GitHub so both documents are visible in your public GitHub repository.

1. Choose one or more graphics you created for Lab 11 and either experiment with the underlying statistical layer if it already has one (i.e. if you made a histogram experiement with different bin widths) or add a separate statistical layer to your plot (i.e. a smooothing curve). Choose something you think will offer meaningful insight and describe why you made the choice you did. What additional information does this provide viewers of your graphic?

I chose the stata theme because I the loess curve is what I want to show. There are significant differences before and after US v. Booker in 2005

```
# library and data importing
library(extrafont)
```

```
## Registering fonts with R
```

```
library(ggplot2)
library(readr)
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------
```

```
## ✓ tibble  1.3.4     ✓ dplyr   0.7.4
## ✓ tidyr   0.7.2     ✓ stringr 1.2.0
## ✓ purrr   0.2.4     ✓ forcats 0.2.0
```

```
## -- Conflicts ------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggthemes)
data <- read_csv("~/monitoring-federal-criminal-sentences/clean_data/merged_data/96-15.csv")
```

```
## Parsed with column specification:
## cols(
##   USSCIDN = col_integer(),
##   MONRACE = col_integer(),
##   YEAR = col_integer(),
##   MONSEX = col_integer(),
##   CITIZEN = col_integer(),
##   EDUCATN = col_integer(),
##   AGE = col_integer(),
##   STATMIN = col_integer(),
##   STATMAX = col_integer(),
##   DISPOSIT = col_integer(),
##   TOTPRISN = col_integer(),
##   XFOLSOR = col_integer(),
##   XCRHISSR = col_integer(),
##   DISTRICT = col_integer(),
##   MONCIRC = col_integer()
```

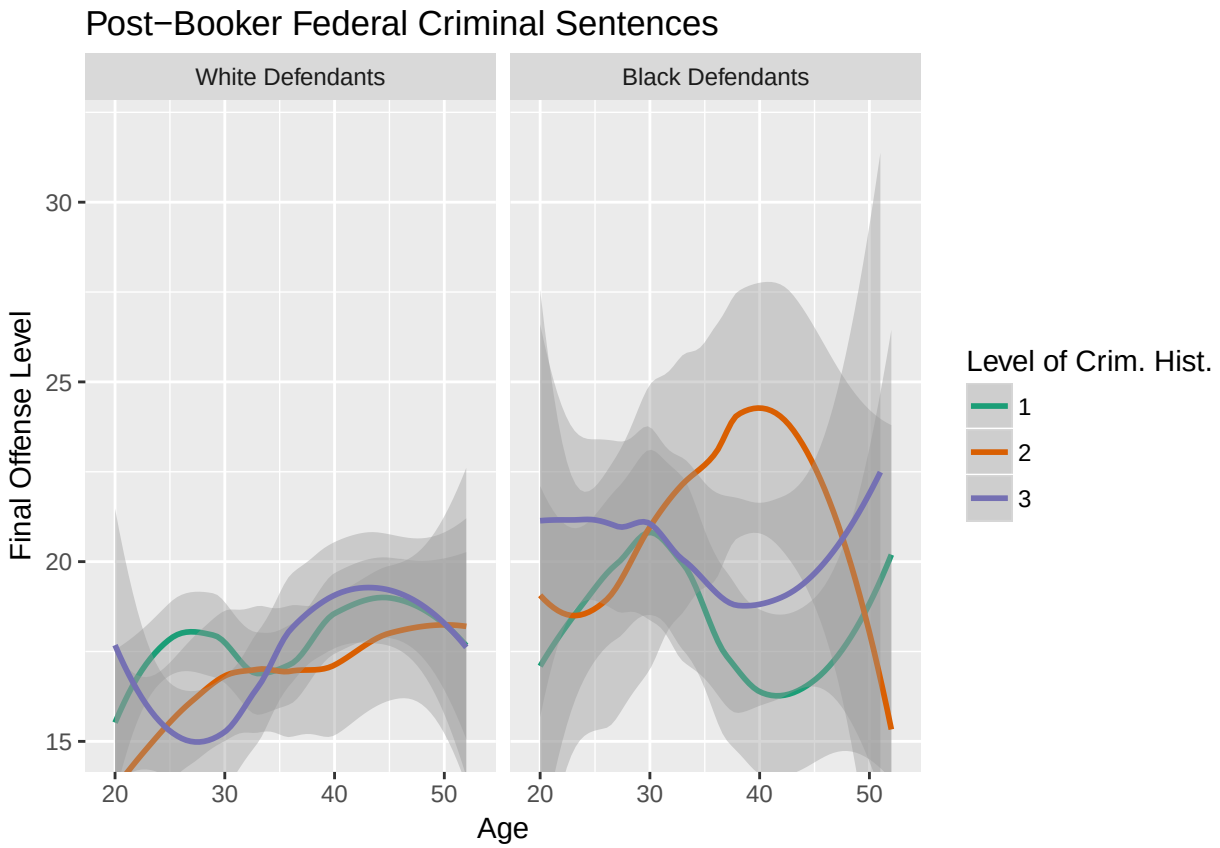
```
## )
```

```
## Warning in rbind(names(probs), probs_f): number of columns of result is not
## a multiple of vector length (arg 1)
```

```
## Warning: 47338 parsing failures.
## row # A tibble: 5 x 5 col     row   col           expected actual expected   <int> <chr>            <chr>  <
## ... ..... ... ..... ... ..... ... ..... ... ..... ... ..... ... ..... ... ..... ... ..... ... ..... ... .....
## See problems(...) for more details.
```

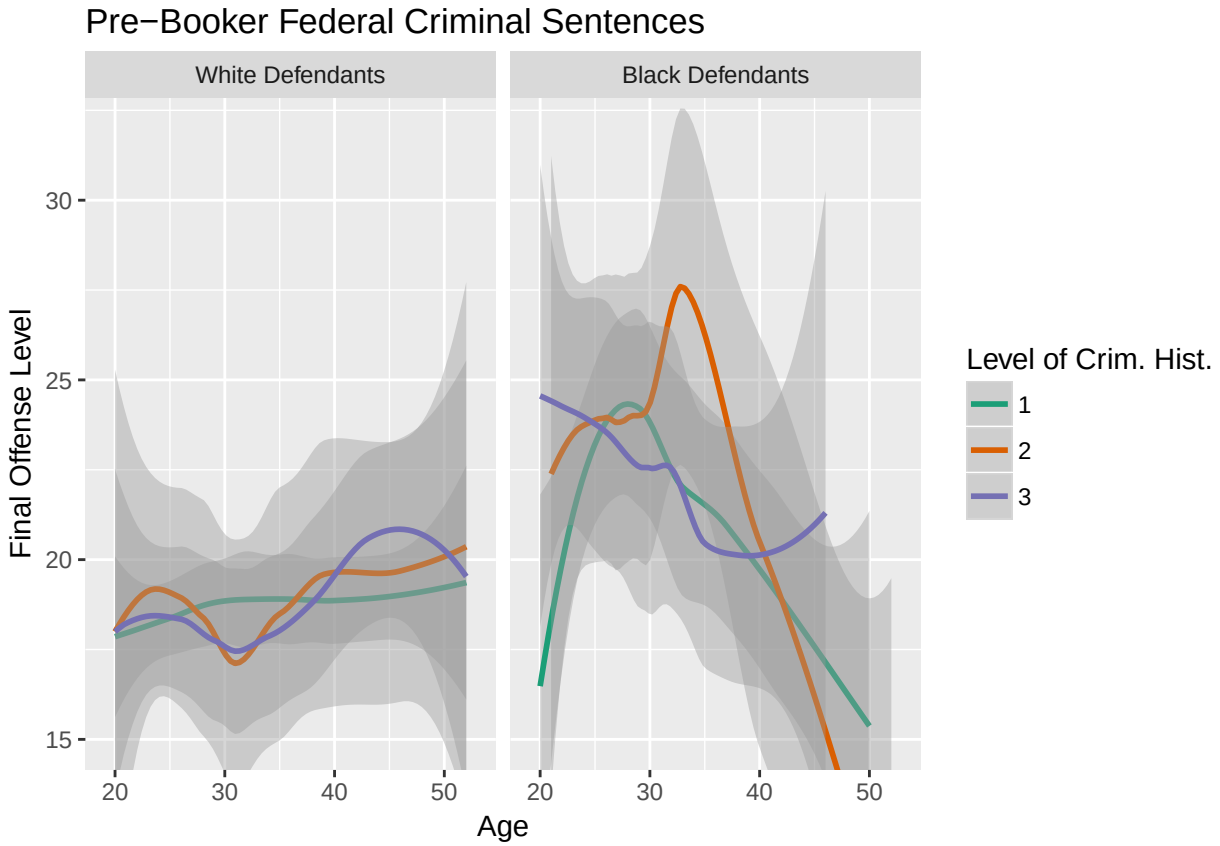
```
# using prebooker
data_sample <- sample_n(data, 6000)

data_prebooker <- data_sample %>%
  filter(MONRACE <= 2) %>%
  filter(XCRHISSR <= 3) %>%
  filter(YEAR < 2005) %>%
  filter(AGE < 53 & AGE > 19)

data_postbooker  <- data_sample %>%
  filter(MONRACE <= 2) %>%
  filter(XCRHISSR <= 3) %>%
  filter(YEAR > 2005) %>%
  filter(AGE < 53 & AGE > 19)

race_labels <- c('1' = "White Defendants",
                 '2' = "Black Defendants",
                 '3' = "Native American",
                 '4' = "Asian/Pacific Islander",
                 '5' = "Other")

post_booker_scatter <- ggplot(data_postbooker, aes(x = AGE, y = XFOLSOR, col = as.factor(XCRHISSR))) +
                               labs(title = "Post-Booker Federal Criminal Sentences", y = "Final Offens
                               scale_color_brewer(name = "Level of Crim. Hist.", type = "qual", palette
                               guides(guide_legend(title = "Race")) +
                               coord_cartesian(xlim = c(19,53), ylim = c(15,32)) +
                               facet_grid(.~ MONRACE, labeller = as_labeller(race_labels)) +
                               geom_smooth(method = 'loess')

pre_booker_scatter <- ggplot(data_prebooker, aes(x = AGE, y = XFOLSOR, col = as.factor(XCRHISSR))) +
                               labs(title = "Pre-Booker Federal Criminal Sentences", y = "Final Offense
                               scale_color_brewer(name = "Level of Crim. Hist.", type = "qual", palette
                               guides(guide_legend(title = "Race")) +
                               coord_cartesian(xlim = c(19,53), ylim = c(15,32)) +
                               facet_grid(.~ MONRACE, labeller = as_labeller(race_labels)) +
                               geom_smooth(method = 'loess')



post_booker_scatter
```

```
## Warning: Removed 11 rows containing non-finite values (stat_smooth).
```




```
pre_booker_scatter
```

```
## Warning: Removed 6 rows containing non-finite values (stat_smooth).
```



2. With the same or a different plot created in Lab 11, experiment with zooming in on specific areas of your graphic and changing the aspect ratio. Are their any benefits/drawbacks with either or both of these approaches for the visualizations you've created? What are they?

Done above - The benefit is that I ignore the data points and only focus on the loess curve close-up and really observe the differences. The flaw is that I don't get to see any peripheral data that might be outliers.

3. Try facetting a plot you have made by another categorical variable in your data (this can even be as simple as Male/Female). What is the difference between `facet_wrap()` and `facet_grid()`? How might facetting be useful in data visualization?

Done above with race.

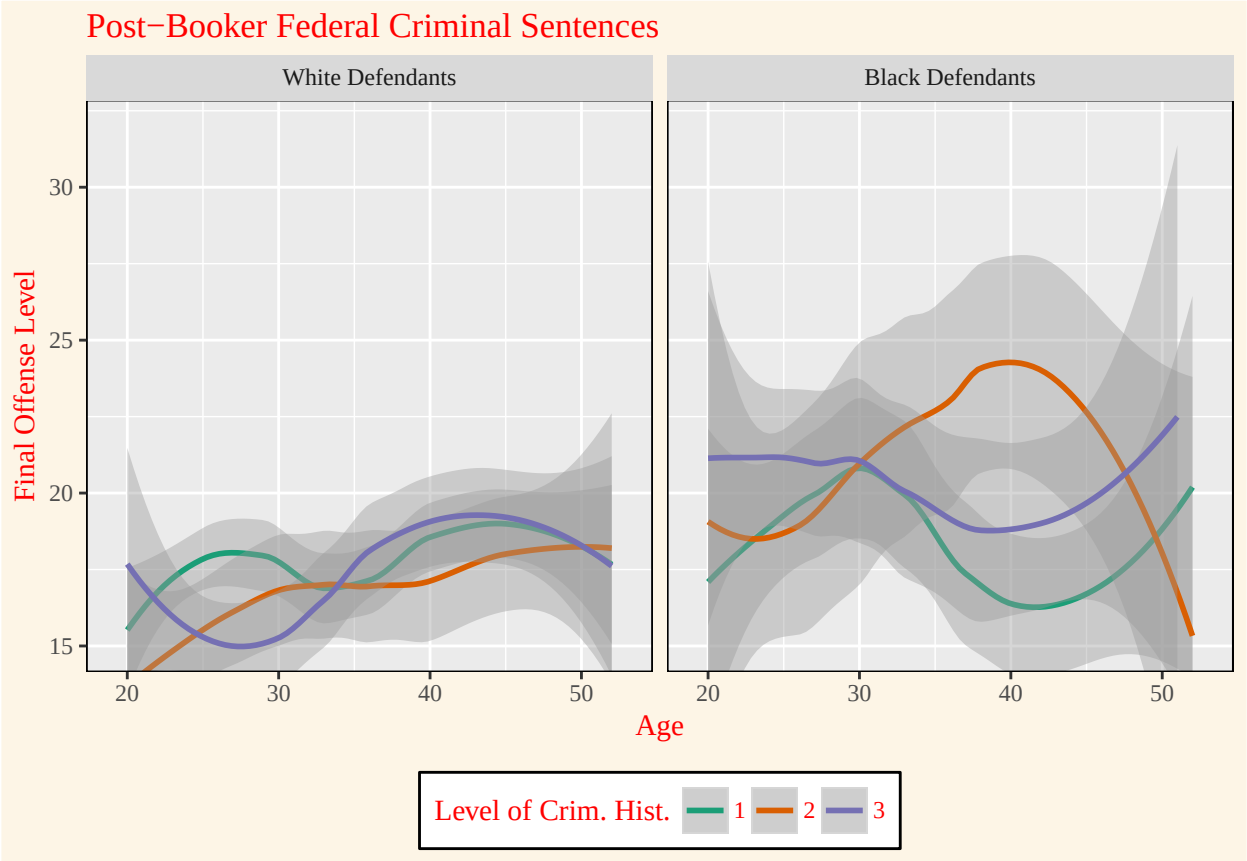
4. Use the `theme()` layer to change the appearance of a plot of your choice including the

- plot, axes, and legend titles
- axes tick marks
- text size
- legend position

Done above with stata. I'll change it on my own below to make manual modifications.

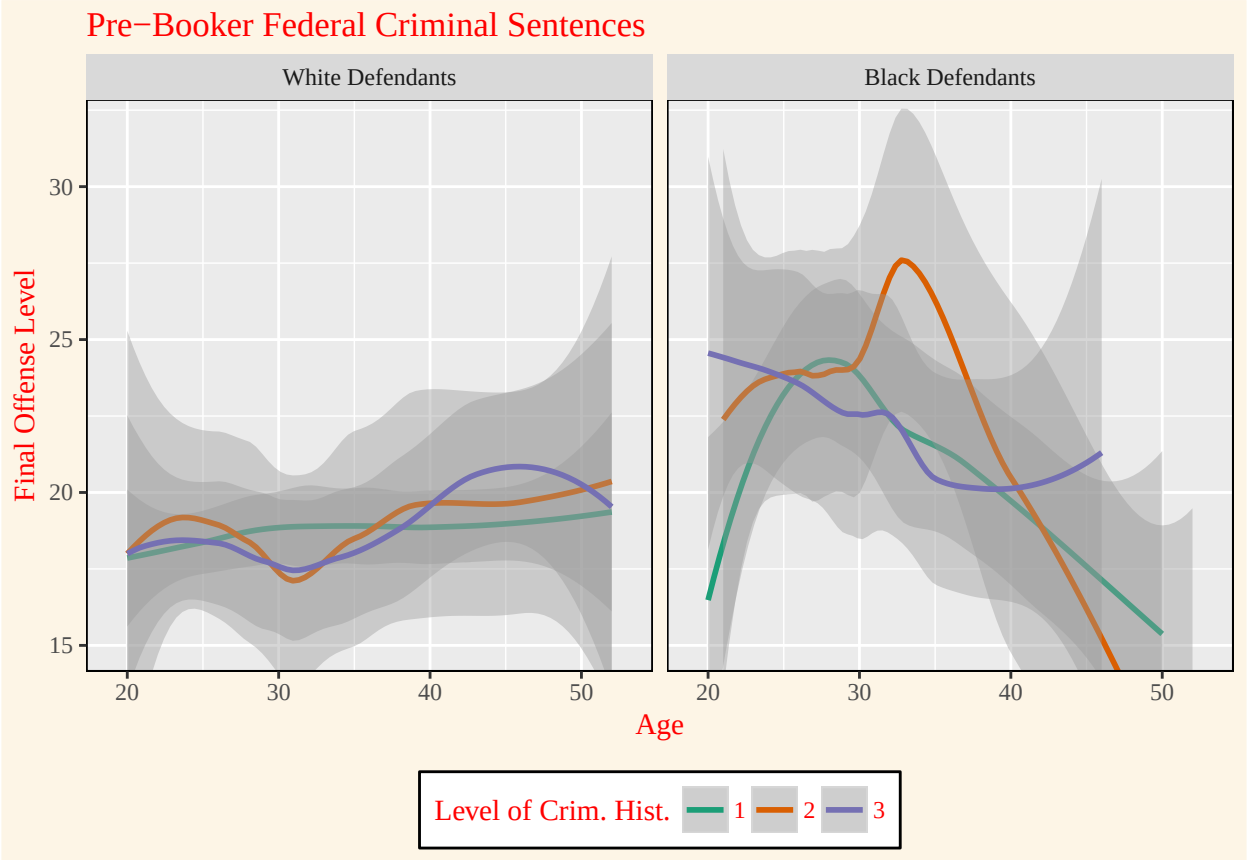
```
# creating theme
mytheme <- theme(text = element_text(family = "serif", colour = "red1"),
                 plot.background = element_rect(fill = "oldlace"),
                 legend.background = element_rect(colour = "black"),
                 legend.position = "bottom",
                 panel.border = element_rect(colour = "black", fill = NA))


post_booker_scatter + mytheme
```

```
## Warning: Removed 11 rows containing non-finite values (stat_smooth).
```




```
pre_booker_scatter + mytheme
```

```
## Warning: Removed 6 rows containing non-finite values (stat_smooth).
```

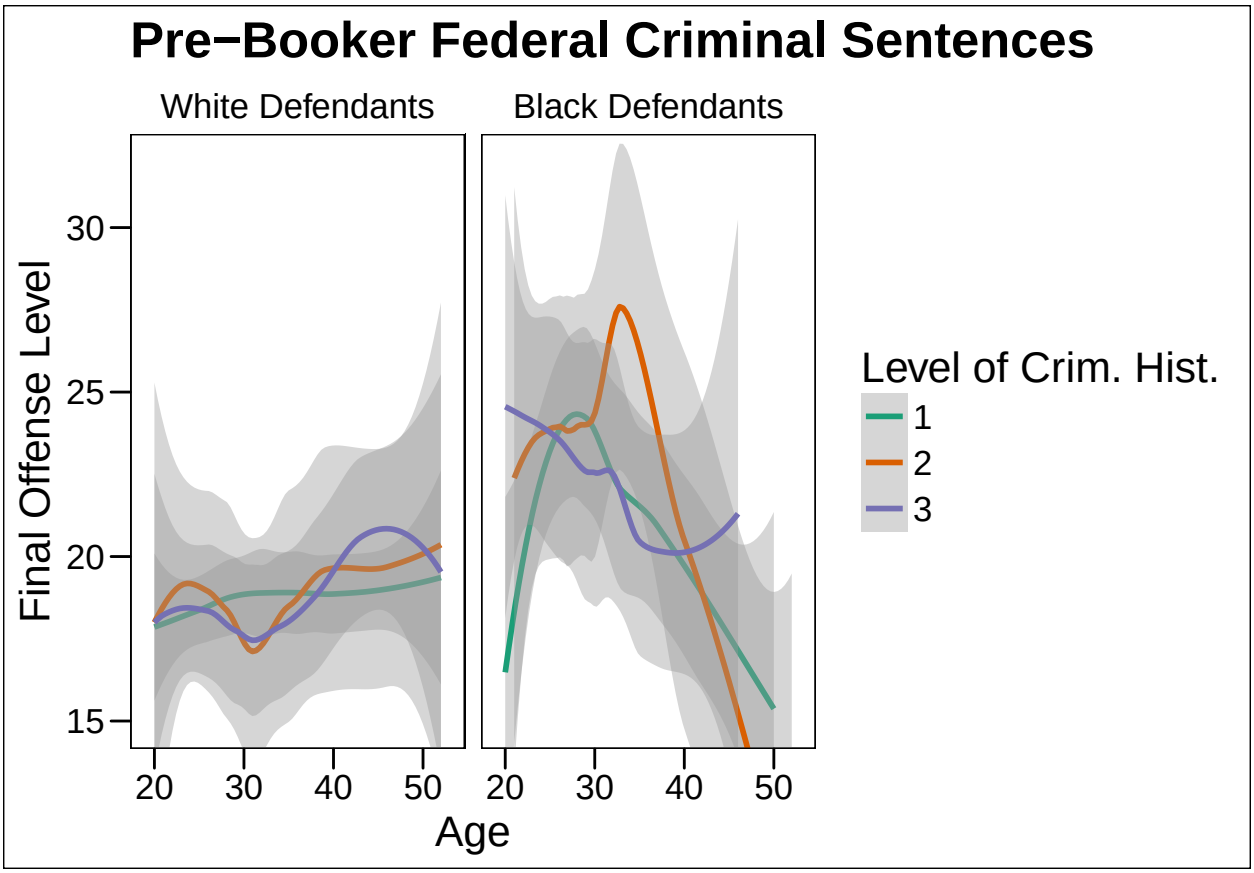


5. Create three versions of a graphic of your choice using different built-in themes or a theme created from `ggthemes`. Which ones do you think are best for presenting in an academic journal? A poster session? What are the qualities of the themes that you choice that you think make them more appropriate for presentation?

In my opinion STATA is the best option because it looks the cleanest. I will apply 3 differnet things with postbooker.

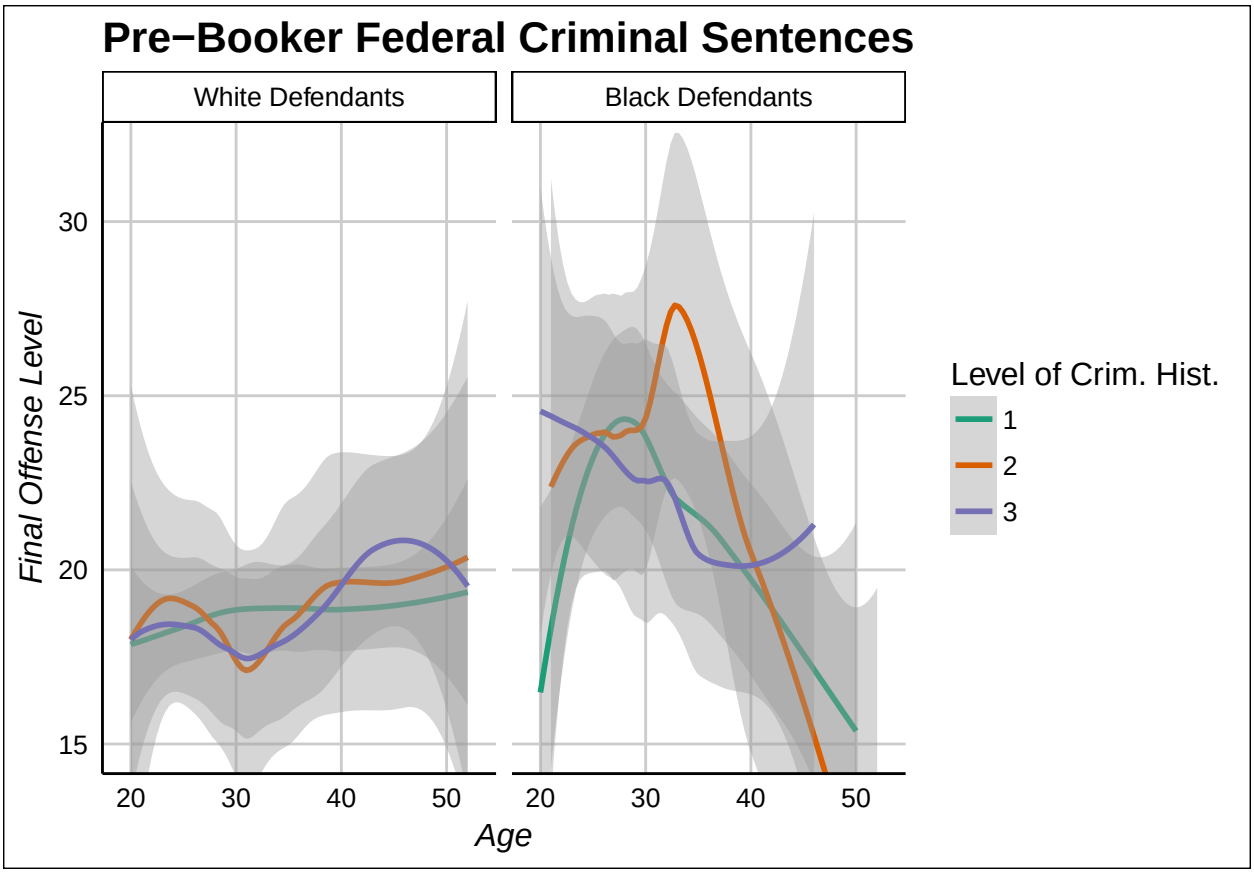
```
pre_booker_scatter + theme_base()
```

```
## Warning: Removed 6 rows containing non-finite values (stat_smooth).
```



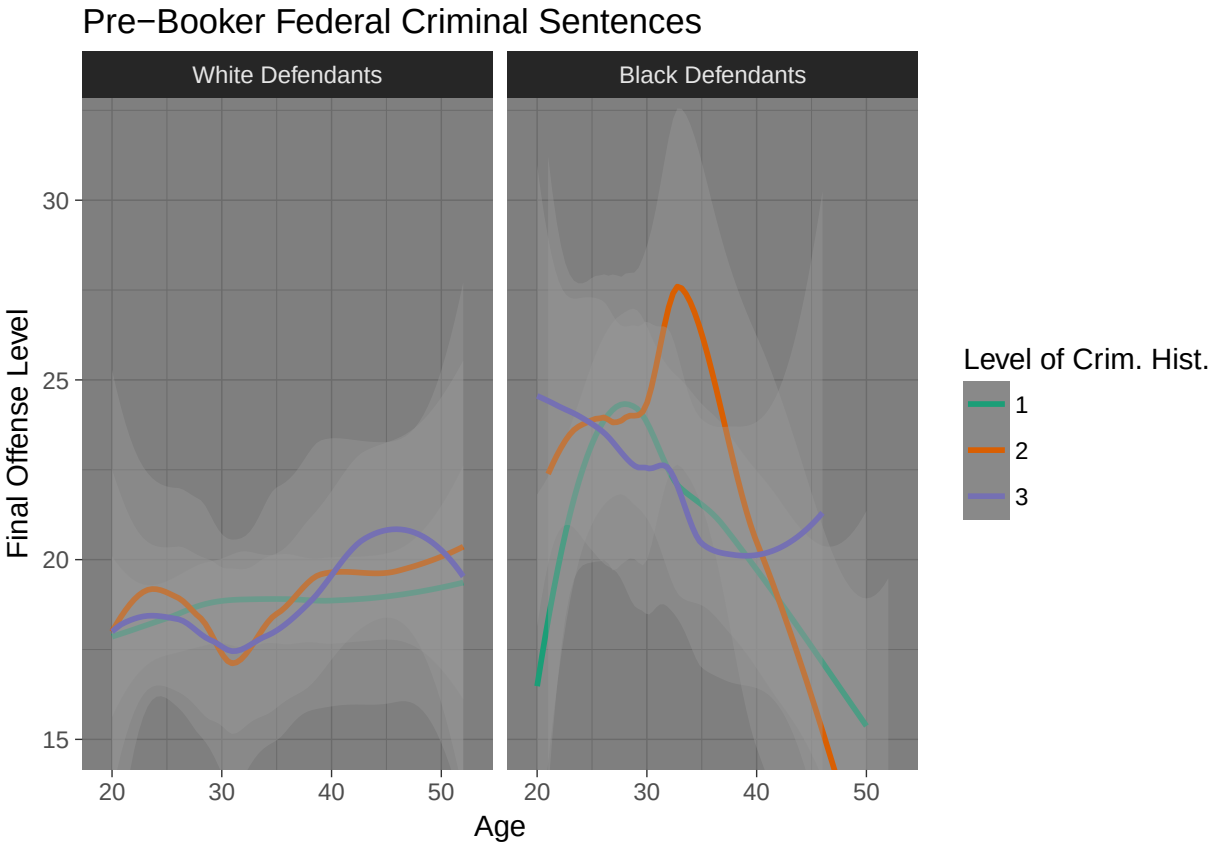

```
pre_booker_scatter + theme_gdocs()
```

```
## Warning: Removed 6 rows containing non-finite values (stat_smooth).
```




```
pre_booker_scatter + theme_dark()
```

```
## Warning: Removed 6 rows containing non-finite values (stat_smooth).
```


Pre−Booker Federal Criminal Sentences
White Defendants
Black Defendants
Final Offense Level
30
25
20
15
20
30
40
50
Age
Level of Crim. Hist.
1
2
3