# Chapter 8: Data Visualisation

Answers to Exercises

Brian Fogarty

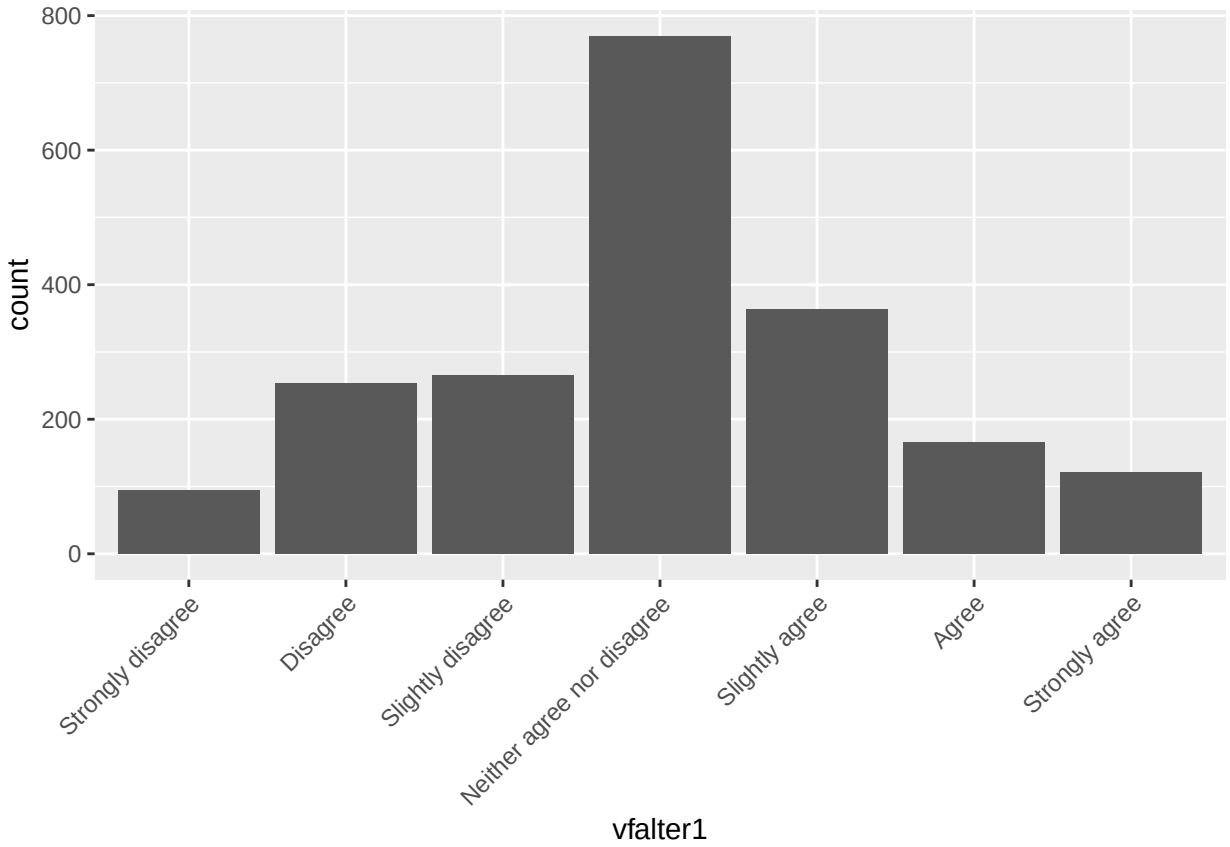
## Contents



## Exercise 1

a. We need to re-order the values of `vfalter` before plotting.

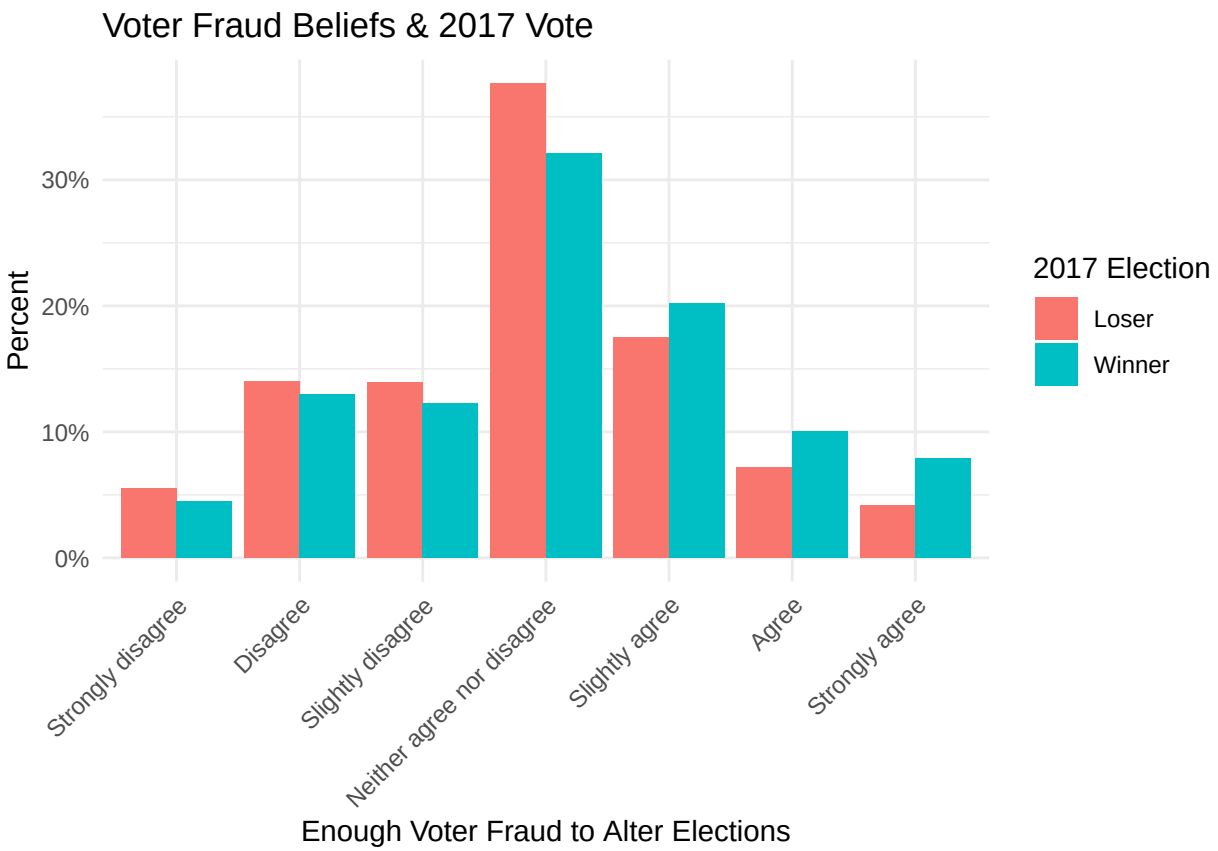
```
library(tidyverse)
vf_england <- read_csv("VF England.csv")

vf_england %>%
  mutate(vfalter1 = factor(vfalter,
           levels=c("Strongly disagree","Disagree",
                    "Slightly disagree","Neither agree nor disagree",
                    "Slightly agree","Agree","Strongly agree"))) %>%
ggplot() +
  geom_bar(mapping = aes(vfalter1)) +
  scale_x_discrete(guide = guide_axis(angle = 45))
```




We see that the modal response is "neither agree nor disagree" that there's enough voter fraud to alter UK elections.

b. To get the percentages on the $y$-axis, we need to load the `scales` package. We also need to remove missing values in `vote2017_dum`.

```
library(scales)

vf_england %>%
  mutate(vfalter1 = factor(vfalter,
          levels=c("Strongly disagree","Disagree",
                   "Slightly disagree","Neither agree nor disagree",
                   "Slightly agree","Agree","Strongly agree"))) %>%
  filter(!is.na(vote2017_dum)) %>%
ggplot() +
  geom_bar(mapping = aes(x = vfalter1, y = ..prop..,
                         group = vote2017_dum, fill = vote2017_dum),
                         stat = "count", position = "dodge") +
  labs(x = "Enough Voter Fraud to Alter Elections",
       title = "Voter Fraud Beliefs & 2017 Vote",
       fill = "2017 Election", y = "Percent") +
  scale_x_discrete(guide = guide_axis(angle = 45)) +
  scale_y_continuous(labels = percent_format()) +
  theme_minimal()
```

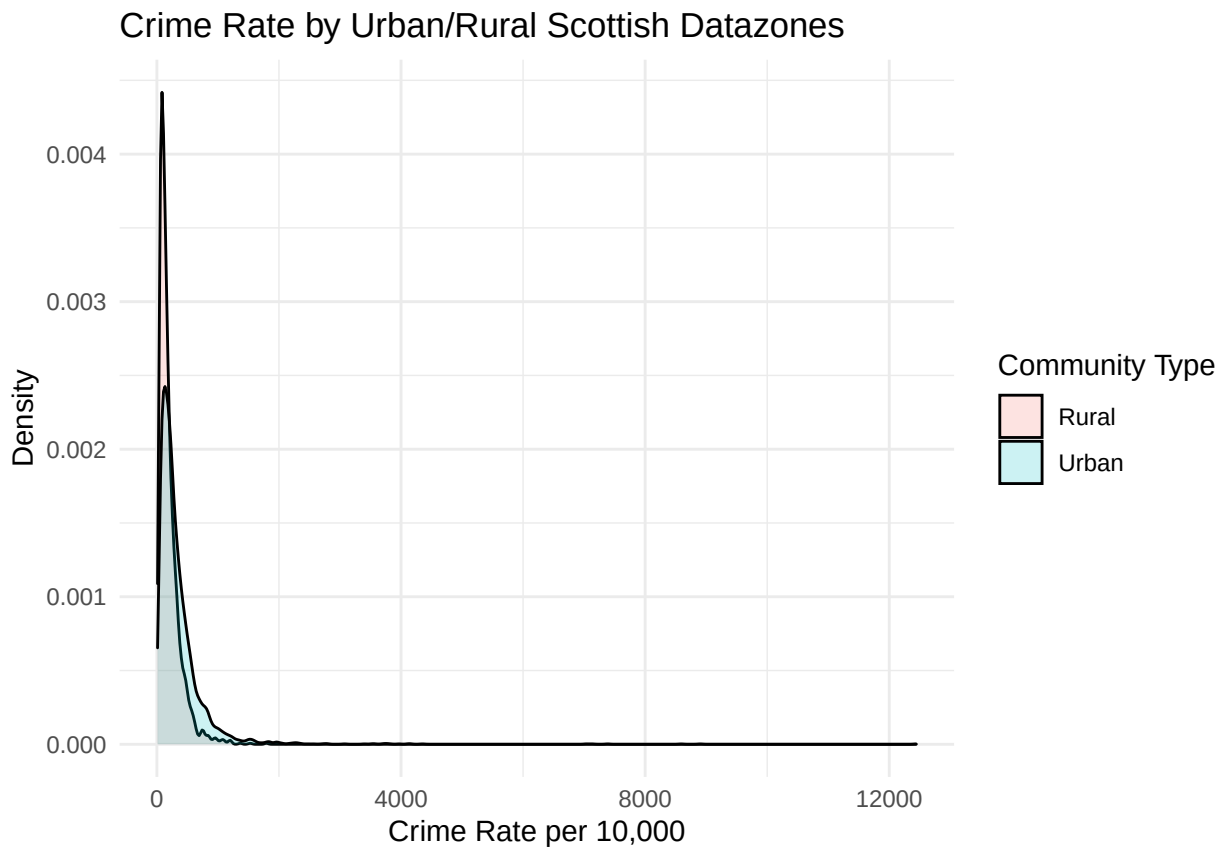



Overall, the bars are similar to what we saw in 'a.' However, we see that a higher percentage of "winners" in the 2017 election agree, to some extent, that there's enough voter fraud to alter UK elections than "losers" in the 2017 election.

## Exercise 2

a. We should re-label the numeric values of `urban` so they are clearer in the plot.

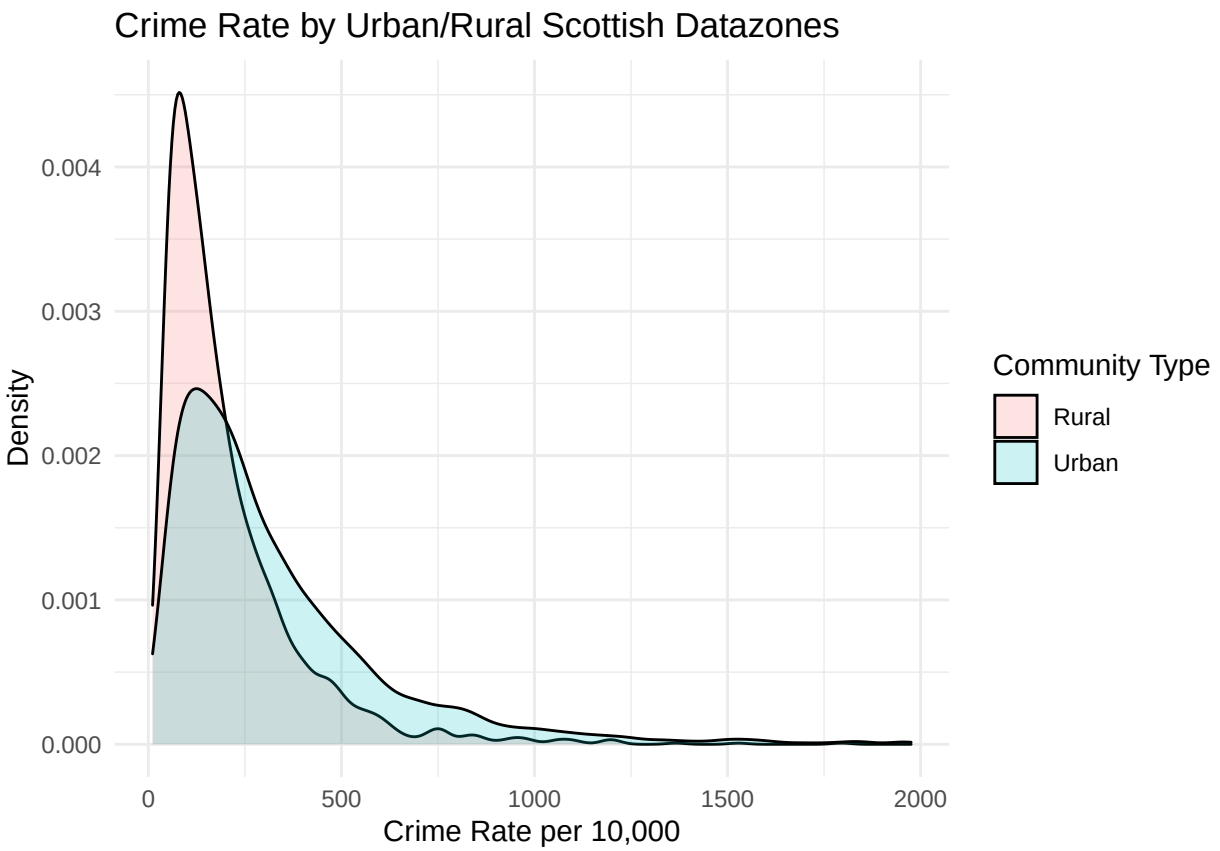
```
simd <- read_csv("simd2020.csv", na = "*")

simd %>%
  mutate(urban_fct = recode(urban, `1` = "Urban", `0` = "Rural")) %>%
  filter(!is.na(crime_rate)) %>%
ggplot() +
  geom_density(mapping = aes(crime_rate, fill = urban_fct), alpha = .2) +
  labs(x = "Crime Rate per 10,000",y = "Density",
       title = "Crime Rate by Urban/Rural Scottish Datazones",
       fill = "Community Type") +
  theme_minimal()
```




It appears that there are more rural datazones with very low crime rates than urban datazones. However, we cannot see much in this plot because of some extreme outliers in `crime_rate`.

b. We filter `crime_rate` to only plot values less than or equal to 2,000.

```
simd %>%
  mutate(urban_fct = recode(urban, `1` = "Urban", `0` = "Rural")) %>%
  filter(!is.na(crime_rate) & crime_rate <= 2000) %>%
ggplot() +
  geom_density(mapping = aes(crime_rate, fill = urban_fct), alpha = .2) +
  labs(x = "Crime Rate per 10,000",y = "Density",
       title = "Crime Rate by Urban/Rural Scottish Datazones",
       fill = "Community Type") +
  theme_minimal()
```

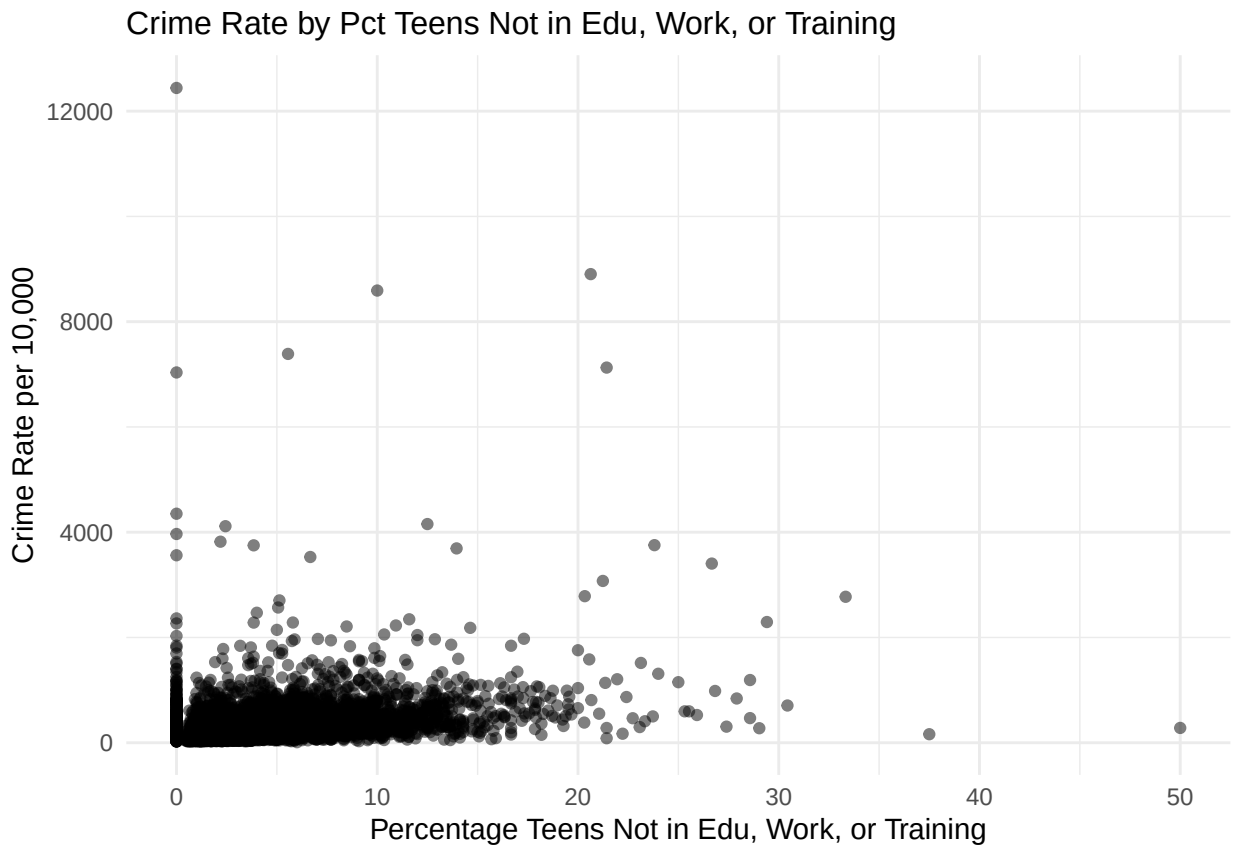



Filtering `crime_rate` to remove outliers makes it clearer to see that rural datazones have lower crime rates than urban datazones.

## Exercise 3

a. We need to multiple `not_participating` by 100 to make the variable more sensible as a percentage.

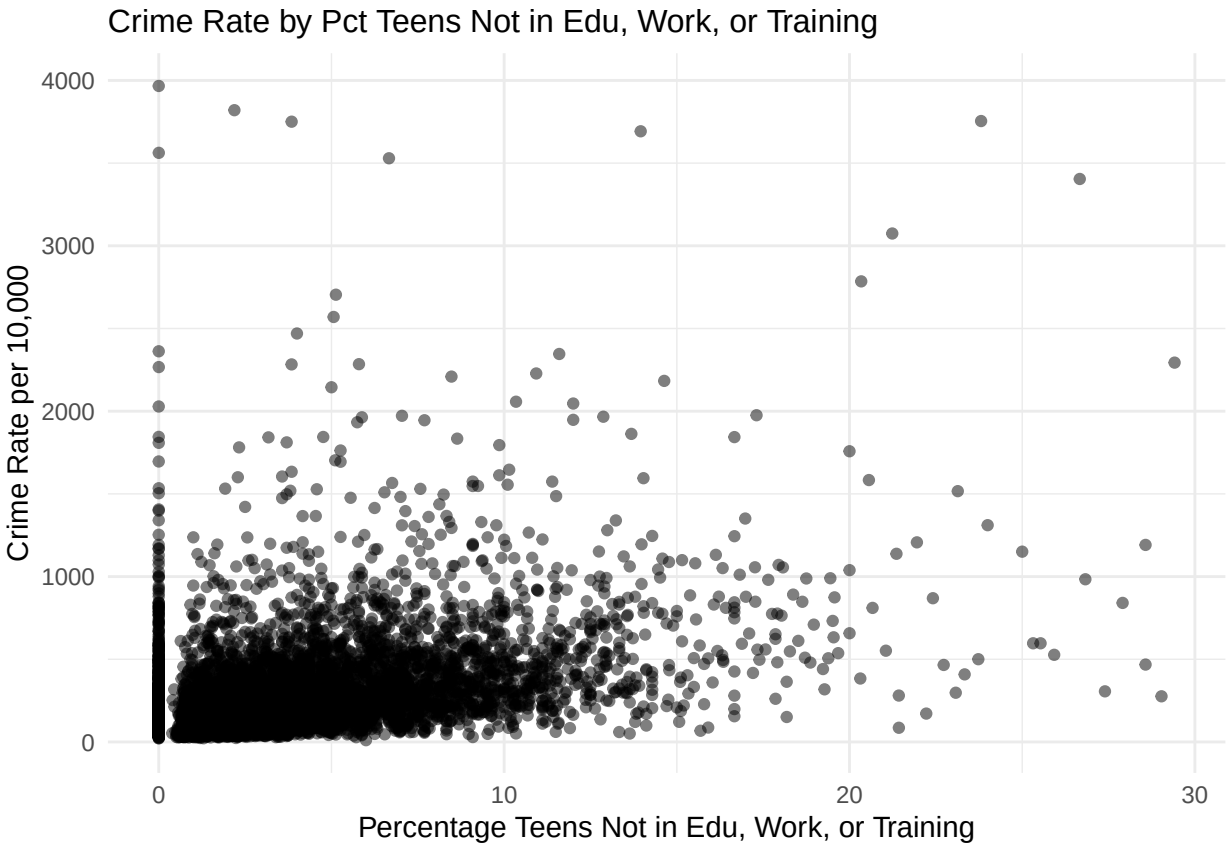
```
simd %>%
  mutate(not_participating = not_participating*100) %>%
  filter(!is.na(crime_rate) & !is.na(not_participating)) %>%
ggplot() +
  geom_point(mapping = aes(x = not_participating, y = crime_rate),
             position = "jitter", alpha = .5) +
  labs(x = "Percentage Teens Not in Edu, Work, or Training",
       y = "Crime Rate per 10,000",
       title = "Crime Rate by Pct Teens Not in Edu, Work, or Training") +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 12)
  )
```




There appears to be a slight positive relationship between `not_participating` and `crime_rate`, but the outliers (in both variables) make it difficult to observe.

b. We filter `crime_rate` and `not_participanting` to remove outliers.

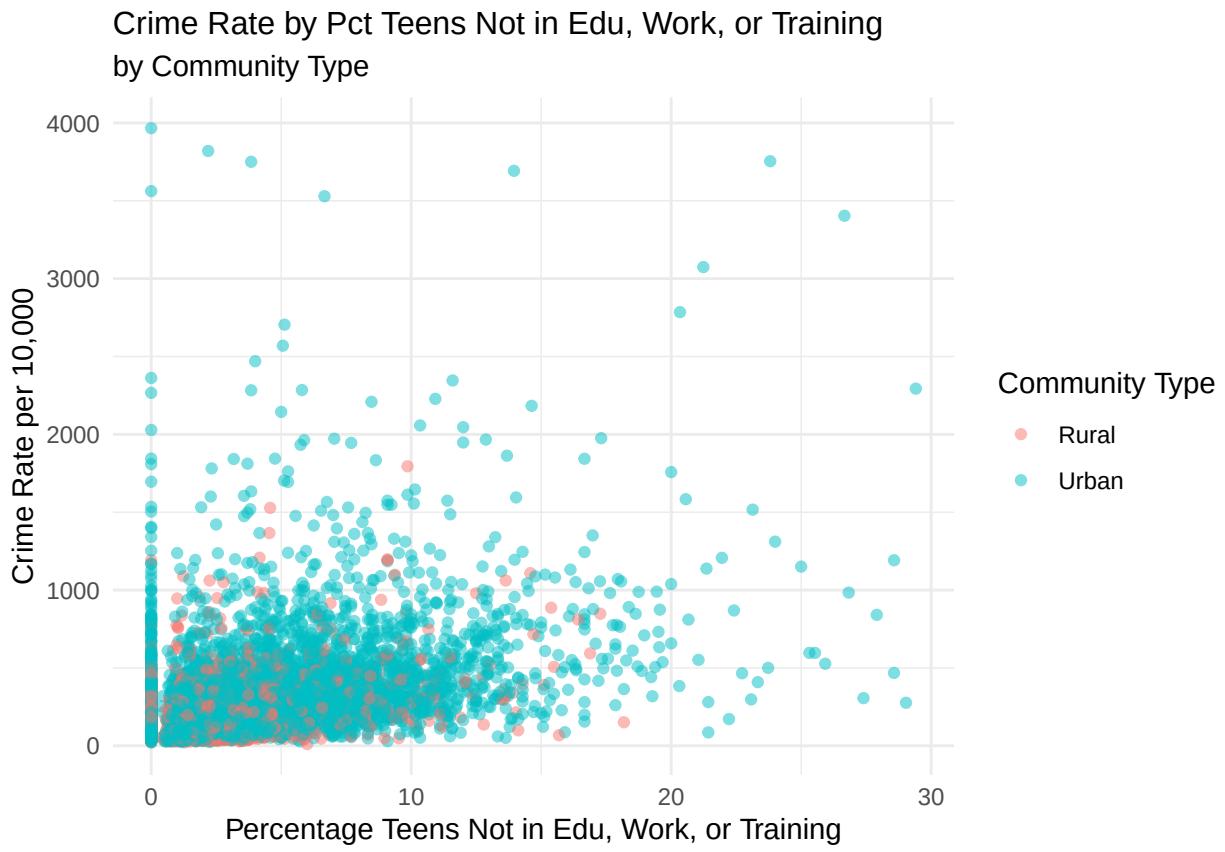
```
simd %>%
  mutate(not_participating = not_participating*100) %>%
  filter(!is.na(crime_rate) & !is.na(not_participating) &
           crime_rate <= 4000 & not_participating < 30) %>%
ggplot() +
  geom_point(mapping = aes(x = not_participating, y = crime_rate),
             position = "jitter", alpha = .5) +
  labs(x = "Percentage Teens Not in Edu, Work, or Training",
       y = "Crime Rate per 10,000",
       title = "Crime Rate by Pct Teens Not in Edu, Work, or Training") +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 12)
  )
```




Now it is easier to see the relationship between `not_participating` and `crime_rate`. Although most of the datazones are low on both variables, there does appear to be a weak positive relationship between `not_participating` and `crime_rate`. This means that datazones with higher percentages of teens that are not in education, work, or training have higher crime rates.

c. We add `urban` as the third variable.

```
simd %>%
  mutate(urban_fct = recode(urban, `1` = "Urban", `0` = "Rural"),
         not_participating = not_participating*100) %>%
  filter(!is.na(crime_rate) & !is.na(not_participating) &
         crime_rate <= 4000 & not_participating < 30) %>%
ggplot() +
  geom_point(mapping = aes(x = not_participating, y = crime_rate,
                           colour = urban_fct),
             position = "jitter", alpha = .5) +
  labs(x = "Percentage Teens Not in Edu, Work, or Training",
       y = "Crime Rate per 10,000",
       title = "Crime Rate by Pct Teens Not in Edu, Work, or Training",
       subtitle = "by Community Type", colour = "Community Type") +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 12)
  )
```

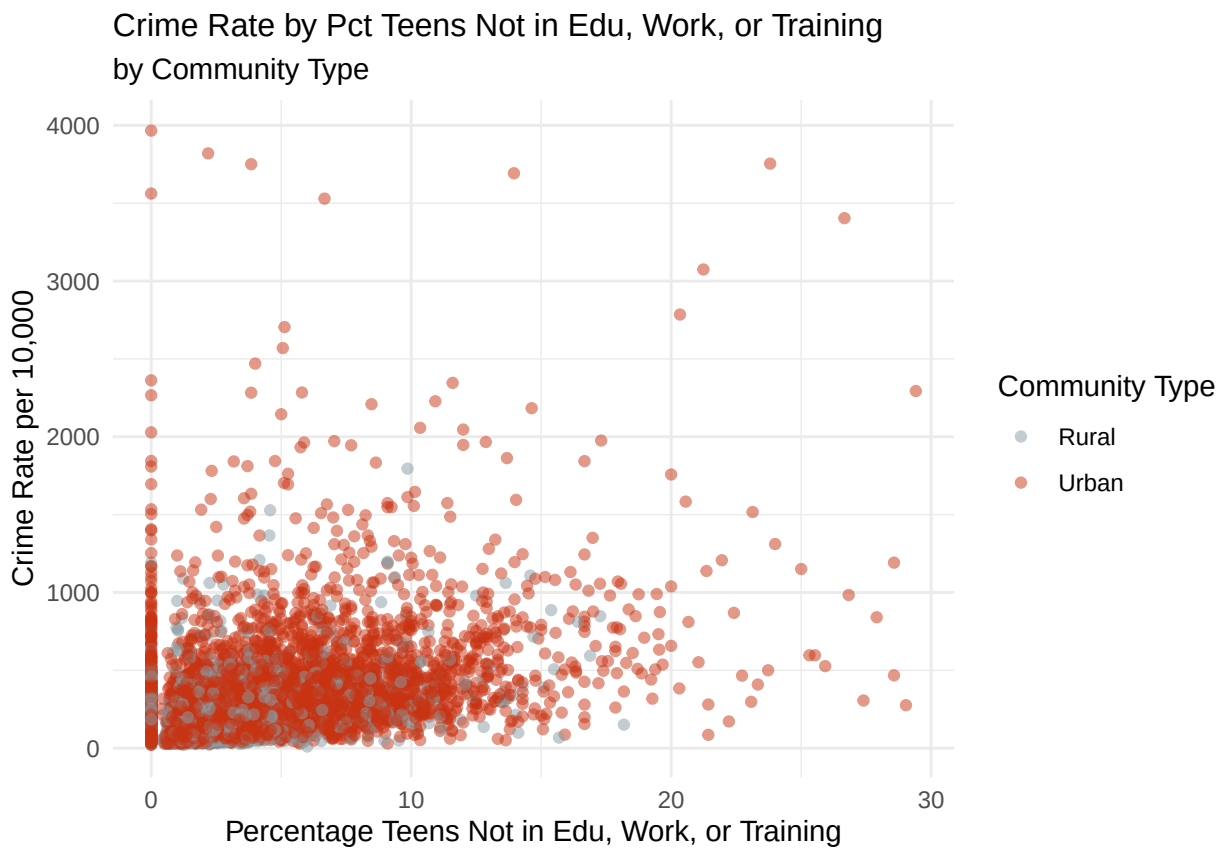



Overall, we see that datazones with higher percentages of teens that are not in education, work, or training and higher crime rates tend to be urban. It is also the case that urban datazones have the highest crime rates as well as highest percentages of teens there are not in education, work, or training, as separate measures. Because there are many more urban datazones overall, it is somewhat difficult to observe the rural datazones. Therefore, how the relationship differs by `urban` is not very clear.

d. The code below uses the `Royal1` palette from the `wesanderson` package to change the colours.

```
library(wesanderson)

simd %>%
  mutate(urban_fct = recode(urban, `1` = "Urban", `0` = "Rural"),
         not_participating = not_participating*100) %>%
  filter(!is.na(crime_rate) & !is.na(not_participating) &
         crime_rate <= 4000 & not_participating < 30) %>%
ggplot() +
  geom_point(mapping = aes(x = not_participating, y = crime_rate,
                           colour = urban_fct),
             position = "jitter", alpha = .5) +
  labs(x = "Percentage Teens Not in Edu, Work, or Training",
       y = "Crime Rate per 10,000",
       title = "Crime Rate by Pct Teens Not in Edu, Work, or Training",
       subtitle = "by Community Type", colour = "Community Type") +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 12)
```

```
  ) +
  scale_colour_manual(values = wes_palette("Royal1"))
```




e. We'll use the `facet_wrap()` function for the faceting. If we leave `colour = urban_fct` in the `aes()` argument, the plots will have different colours.

```
simd %>%
  mutate(urban_fct = recode(urban, `1` = "Urban", `0` = "Rural"),
         not_participating = not_participating*100) %>%
  filter(!is.na(crime_rate) & !is.na(not_participating) &
         crime_rate <= 4000 & not_participating < 30) %>%
ggplot() +
  geom_point(mapping = aes(x = not_participating, y = crime_rate,
                           colour = urban_fct),
             position = "jitter", alpha = .5) +
  labs(x = "Percentage Teens Not in Edu, Work, or Training",
       y = "Crime Rate per 10,000",
       title = "Crime Rate by Pct Teens Not in Edu, Work, or Training",
       subtitle = "by Community Type", colour = "Community Type") +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 12)
  ) +
  facet_wrap(~ urban_fct)
```


By plotting urban and rural datazones separately, we see that the weak positive relationship between `not_participanting` and `crime_rate` exists in both types of communities. The main difference is that urban datazones have larger outliers than rural datazones.