Assignment 10 Solutions

Your name and student ID

Today's date

- Due date: Dec 2, 5:00pm (make sure to provide enough time for Gradescope submission to be uploaded if you include large image files).
- Please do not change the format of the assignment. This makes grading less efficient.
- Submission process: Please submit a PDF of your assignment to Gradescope. You must tell Gradescope which questions are on which pages. If you can't see it properly on Gradescope, open the PDF in a PDF viewer at the same time so you can make the selections accurately. Not selecting, or selecting inaccurately will result in points being deducted since this makes grading much less efficient.

Helpful hint:

- Knit your file early and often to minimize knitting errors! If you copy and paste code, you are bound to get an error that is hard to diagnose. Hand-writing code is the way to smooth knitting! We recommend knitting your file each time after you write a few sentences/add a new code chunk, so you can detect the source of knitting errors more easily.

Voting during the 1992 election

In the spirit of the upcoming 2020 presidential election, I thought it would be interesting to consider some historical data on voting patterns across US counties.

This code loads in the data frame `counties`:

```
load("../Data/A10_counties.sav")
```

These data are from the 1992 election and looks at the percent of votes cast (for each county) for the `democrat` (Bill Clinton), `republican` (George Bush), and independent presidential nominees (Ross `Perot`).

Ideally, if you were interested in voting patterns, you might look at the relationship between individual characteristics and whether each individual voted Democrat or Republican. However, data like that is often hard to come by. The `counties` data provide data on 3141 counties. Use `View()` to examine these data briefly and read the labels corresponding to the variables. Note that Alaska is not included and that two other counties with populations = 0 have also been excluded.

As discussed in class we have the entire population (not just a sample), so strictly speaking we don't need to perform statistical inference. However, we might pretend this is a sample so that we can apply the techniques of inference and gain competence creating and interpeting a linear model.

Question 1 [2 marks]. Looking only at California, plot the relationship between the % of votes cast for the Democratic candidate (`democrat`) and the population density of the county (`pop.density`).

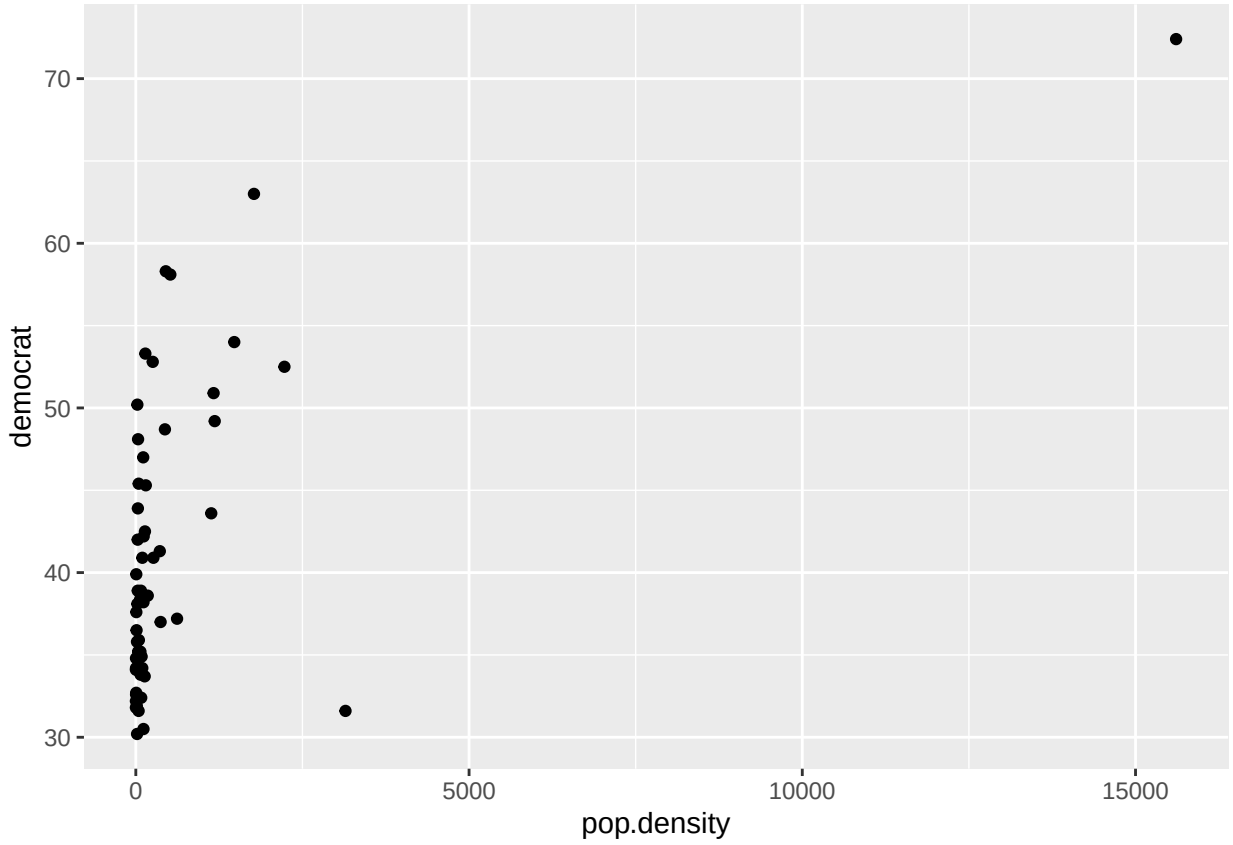
```
# Your code here.

# Solution
library(ggplot2)
library(dplyr)
```

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

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

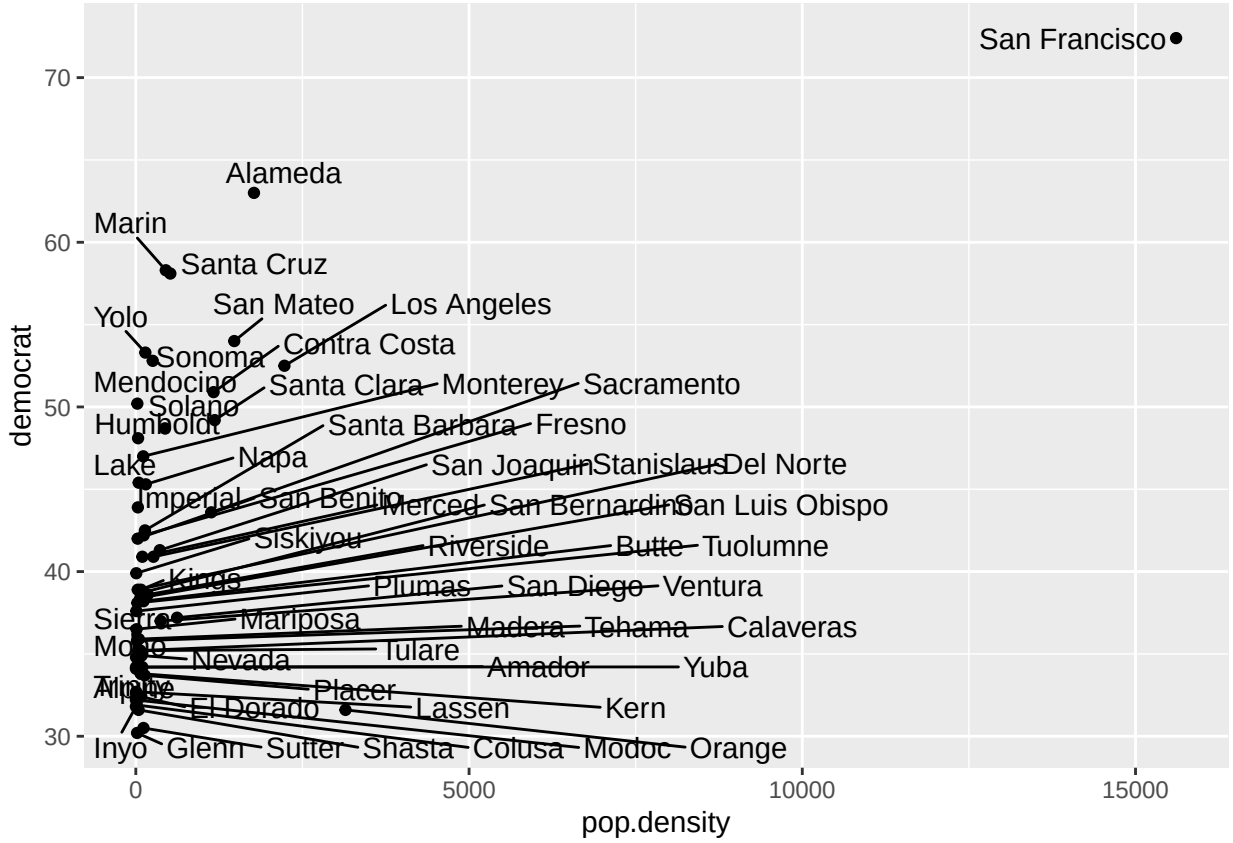
```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
ggplot(counties %>% filter(state == "CA"), aes(x = pop.density, y = democrat)) +
  geom_point()
```



Question 2 [1 mark]. The above plot you made does not look very good for question 1. The distribution of population density is skewed right, with a few counties having much higher densities than the majority of counties. To see which counties these are, we will use `geom_text_repel` from the library `ggrepel`. The template for using this function is: `geom_text_repel(aes(label = your_labelling_var))`. You will want to set the labeling variable to be the variable in the dataset containing the county names.

```
library(ggrepel)

# Your code goes here.

# Solution
ggplot(counties %>% filter(state == "CA"), aes(x = pop.density, y = democrat)) +
  geom_point() +
  geom_text_repel(aes(label = county))
```



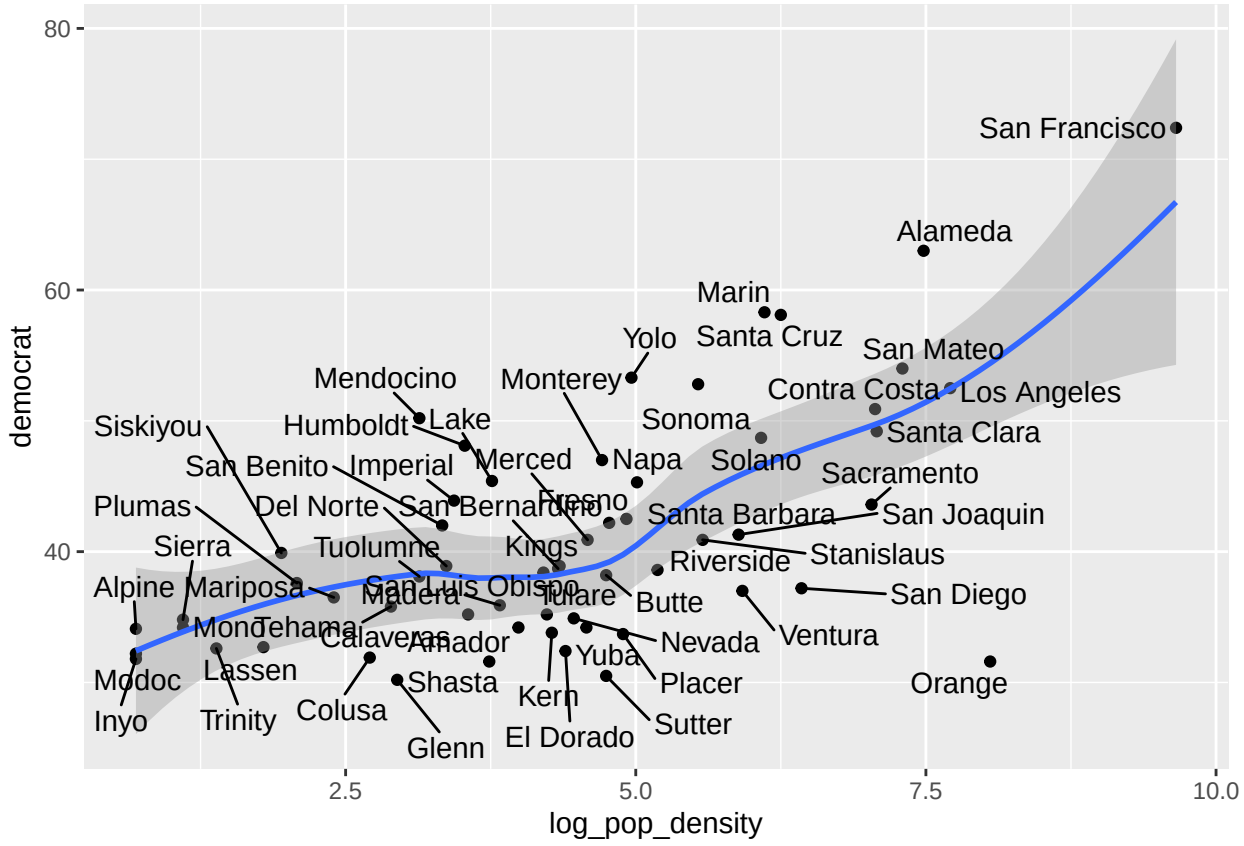
The current issue with these data is that San Francisco (as you can now hopefully point out) has a much higher population density than other counties, and that generally there is a large right skew in the distribution of the population density variable.

If we tried and fit a linear model to these data, it would not fit well– because the relationship between population density and the response variable is not linear. However, this is the perfect situation to try transforming the x variable.

Question 3. [2 marks] Try adding a log-transformed version of population density to the data frame and remake your plot using this new variable. Keep the population labels. Also add a smoothed fitted line:

```
# Your code goes here.

# Solution
counties_CA <- counties %>% filter(state == "CA") %>%
  mutate(log_pop_density = log(pop.density),
         log_pop2 = log_pop_density * log_pop_density)

ggplot(counties_CA, aes(x = log_pop_density, y = democrat)) +
  geom_point() +
  geom_smooth() +
  geom_text_repel(aes(label = county))
```

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



Question 4 [4 marks]. Describe the relationship between the (logged) population density and the response variable in terms of the shape, direction, strength, and outliers. These are concepts from Chapter 3. Calculate the correlation to comment on one of these aspects.

<Your description here.>

Solution:

Direction: There is a positive association between logged population density and the % of votes cast for the democratic candidate is positive. Shape: Roughly linear, or slightly curved. Outliers: No real large outliers, though SF and Orange County are a bit further out from the rest of the points. Strength:

```
counties_CA %>% summarise(cor = cor(democrat, log_pop_density))
```

```
##         cor
## 1 0.6381187
```

The correlation between the variables is 64%, indicating a moderate positive association.

Question 5 [4 marks] Run a linear model regression of the % votes cast for the democratic candidate as a function of the population density. Make sure you get the order of variables right in the `lm()` function! Use a `broom()` command to show the slope and intercept estimates. Interpret the association between the logged population density and the response variable. (You can `View()` the data frame to make sure you are getting your units right by checking the descriptions in the labels for each variable). Report and interpret the r-squared for the model.

```
# Your code here.

# Solution
lm_CA <- lm(formula = democrat ~ log_pop_density, data = counties_CA)

library(broom)
tidy(lm_CA)
```

```
## # A tibble: 2 x 5
##   term            estimate std.error statistic  p.value
##   <chr>              <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)        28.4      2.22      12.8  3.10e-18
## 2 log_pop_density     2.88     0.464      6.20 7.12e- 8
```

```
glance(lm_CA)
```

```
## # A tibble: 1 x 11
##   r.squared adj.r.squared sigma statistic p.value    df logLik   AIC   BIC
##       <dbl>         <dbl> <dbl>     <dbl>   <dbl> <int>  <dbl> <dbl> <dbl>
## 1     0.407         0.397  6.95      38.5 7.12e-8     2  -194.  393.  400.
## # ... with 2 more variables: deviance <dbl>, df.residual <int>
```

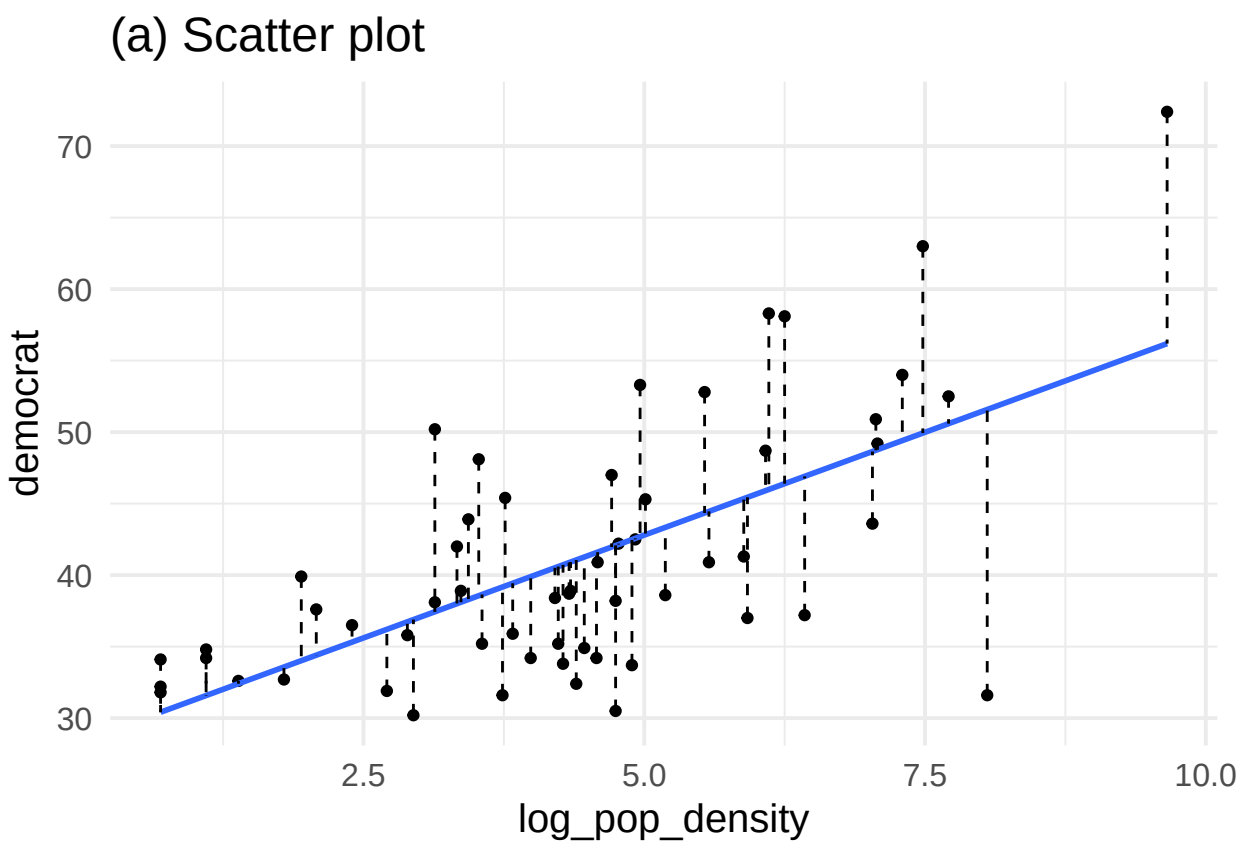
<Your solution goes here.>

Solution: A one unit change in the logged population density is associated with a 2.88 (where population density was the 1992 population per square-mile) percentage point increase in the percent of votes cast for the democratic candidate.

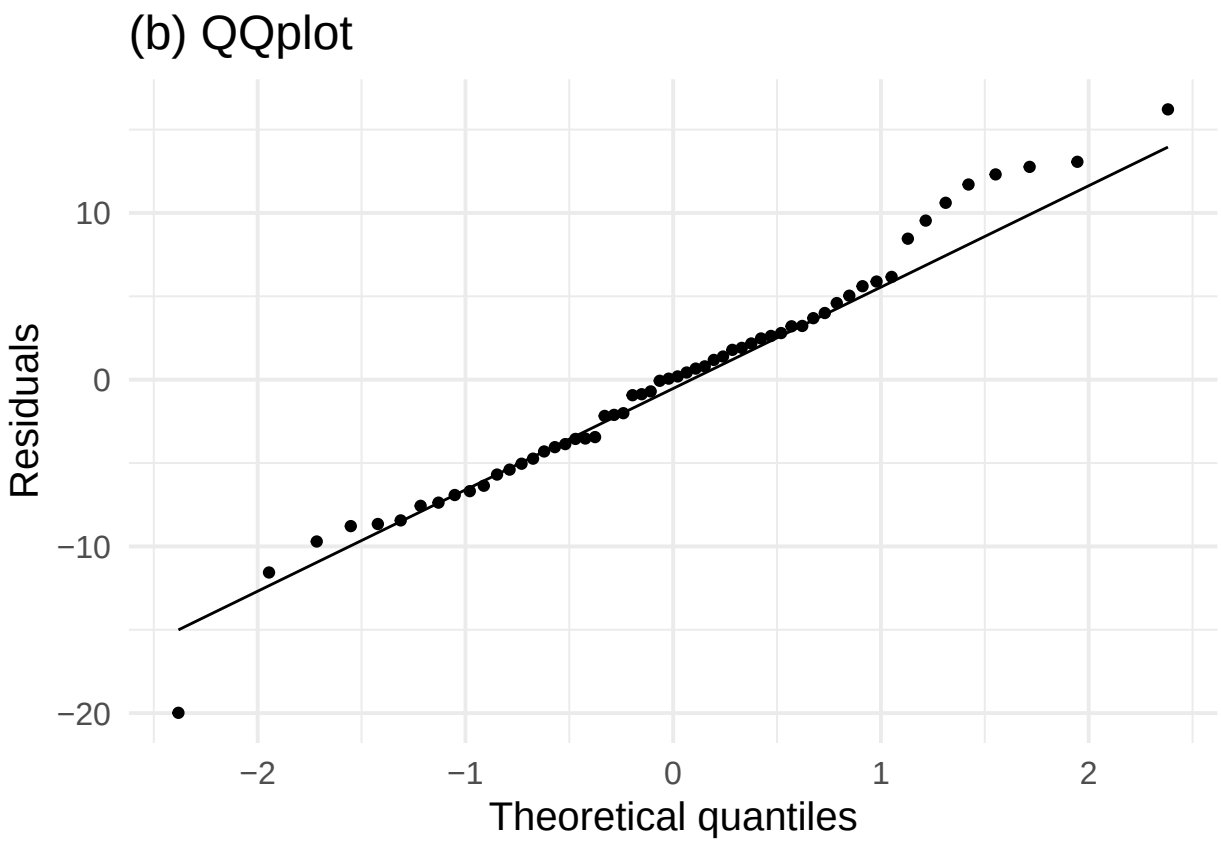
The r-squared is 0.41, implying that 41% of the variation in percentage votes casts is explained by the logged population density.

Question 6 [4 marks]. Using the code learned in class, that was also shown in last week's lab, make the four plots to examine the assumptions.

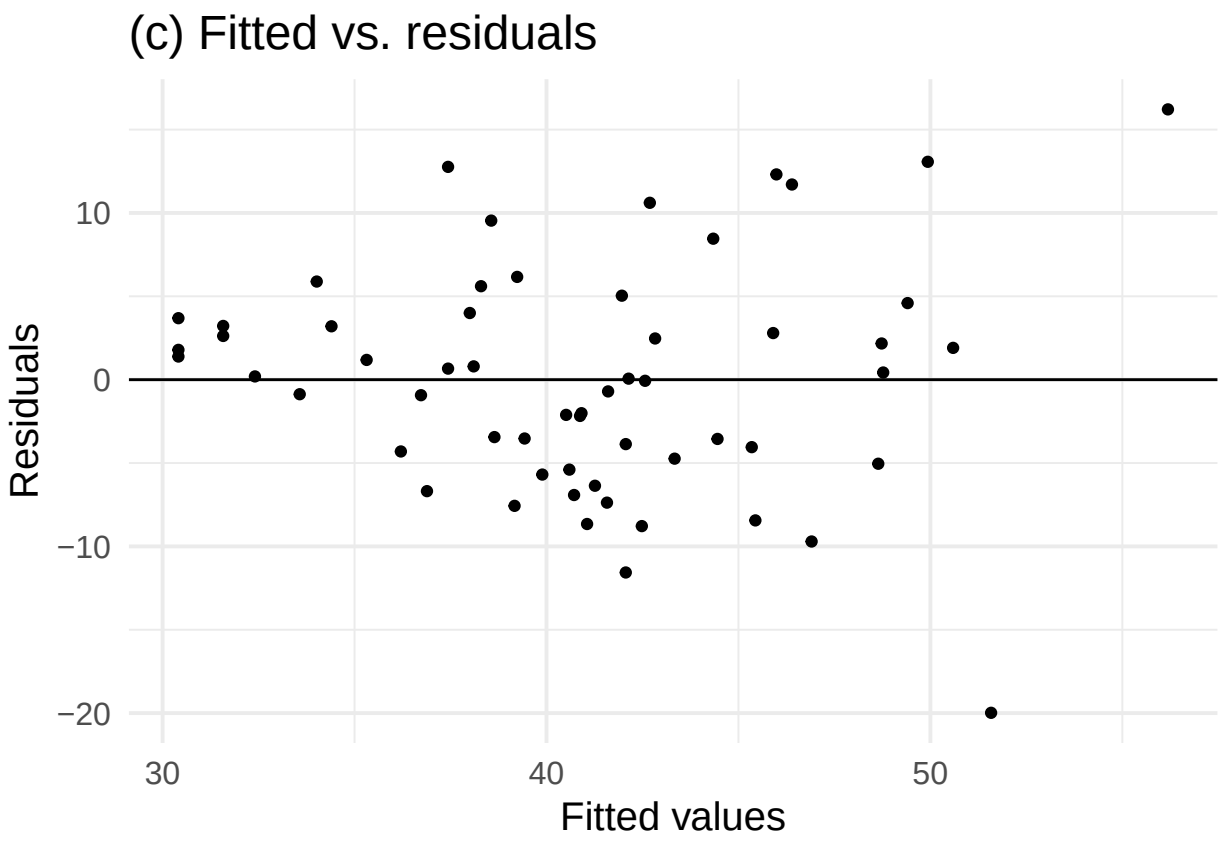
```
#Your code goes here.

#Solution
CA_augment <- augment(lm_CA)

# scatter plot
ggplot(CA_augment, aes(y = democrat, x = log_pop_density)) +
  geom_point() +
  geom_smooth(method = "lm", se = F) +
  geom_segment(aes(xend = log_pop_density, yend = .fitted), lty = 2) +
  theme_minimal(base_size = 15) +
  labs(title = "(a) Scatter plot")
```



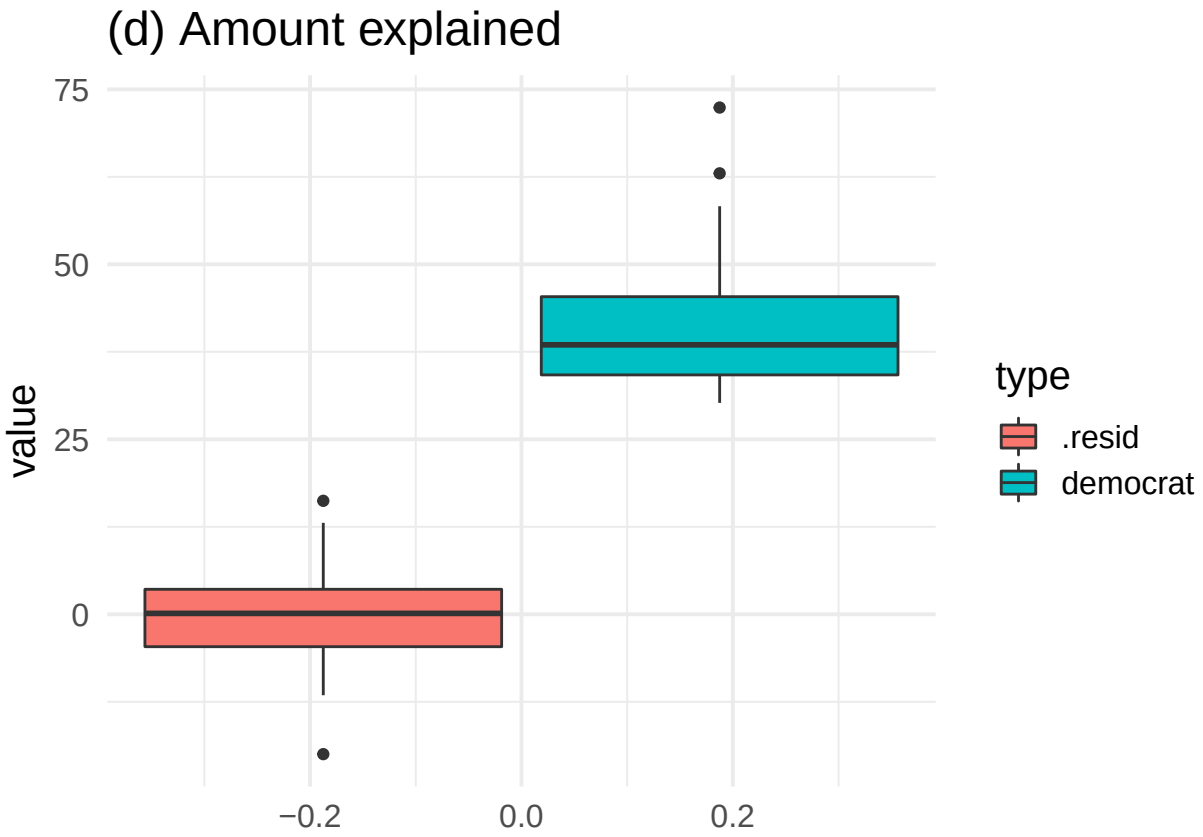

```
# QQ plot
ggplot(CA_augment, aes(sample = .resid)) +
  geom_qq() +
  geom_qq_line() +
  theme_minimal(base_size = 15) +
  labs(y = "Residuals", x = "Theoretical quantiles", title = "(b) QQplot")
```




```
## Fitted vs. residuals
ggplot(CA_augment, aes(y = .resid, x = .fitted)) +
  geom_point() +
  theme_minimal(base_size = 15) +
  geom_hline(aes(yintercept = 0)) +
  labs(y = "Residuals", x = "Fitted values", title = "(c) Fitted vs. residuals")
```




```
## Amount explained
library(tidyr)
bolt_gather <- CA_augment %>% select(democrat, .resid) %>%
  gather(key = "type", value = "value", democrat, .resid)

ggplot(bolt_gather, aes(y = value)) +
  geom_boxplot(aes(fill = type)) +
  theme_minimal(base_size = 15) +
  labs(title = "(d) Amount explained")
```



Question 7. [4 marks] Comment on each of the plots and conclude about which assumptions appear violated vs. not violated. Don't forget to comment on the one assumption that cannot be investigated using plots.

.

Solution: - There is a violation of the assumption that the standard deviation of the response variable are identical for all values of the explantory variable. We see here that as the log population density increases, the residuals become larger. - The relationship between X and Y is approximately linear, though there is a lot of variation around the line of best fit. (This points to the fact that though population density is predictive of the response variable, there are other factors that are not included in our model that have further predictive power.) - The QQ plot looks ok. (May be interpreted as a problem with the largest residuals. - We can't check the assumption that the points are independent using this plot.) Here that corresponds to the counties being independent of one another. This model treats them as independent units. [More sophisticaled models can take the spatial relationships between the counties into account (to account for the fact that counties closer to each other may be more similar.)]

Part 2: Choose your own adventure [10 marks for completion]

Using the `counties` data frame pick a state (California if you still want to, or another one) and investigate the relationship between voting patterns and an explanatory variable of your choice. You can keep using `democrat` as the response variable if you want to further your investigation of predictors of this response, or you can change it to `republican`, or `Perot` (the independent candidate). Set your explanatory variable to one of:

- `college`
- `crime`
- `white`
- `black`
- `farm`
- `age6574` or `age75`, or a `mutate`d variable based on age combining across these two age variables.
- `income`

Your task:

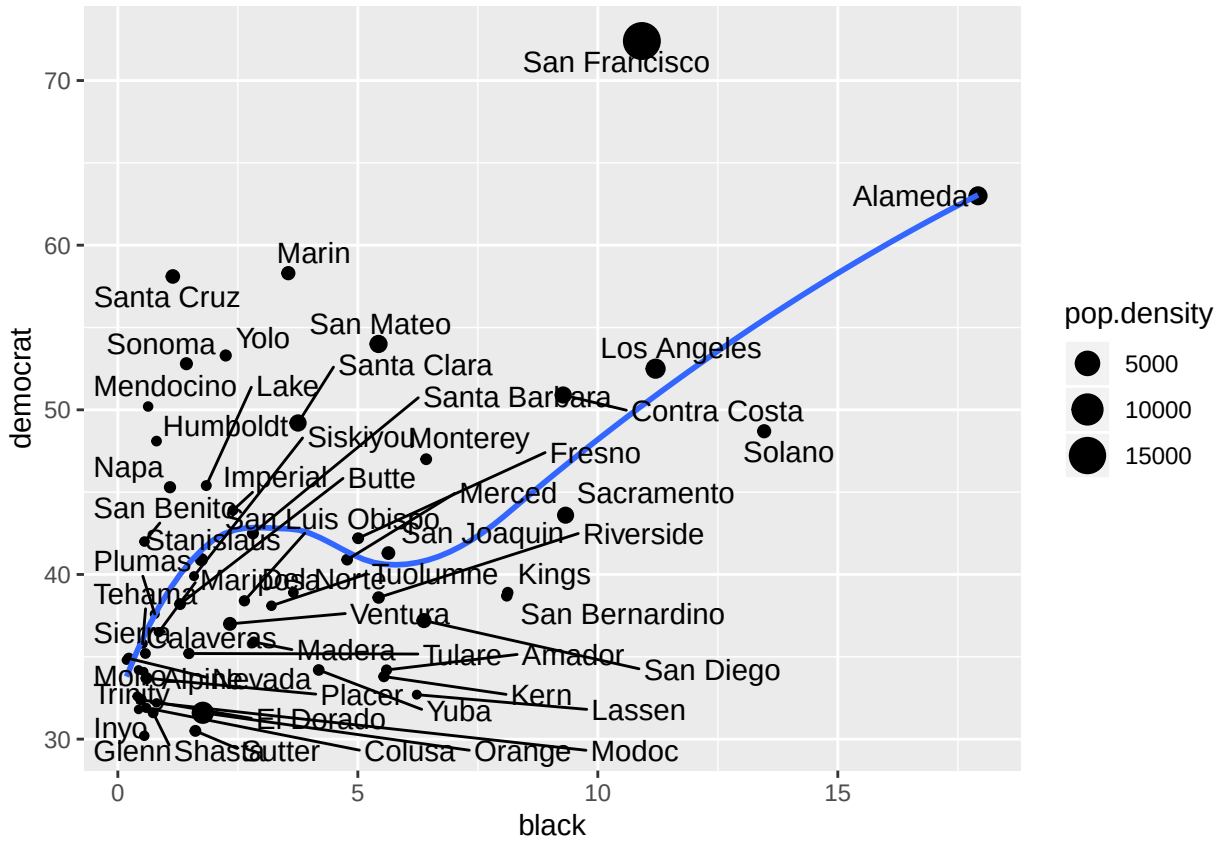
- Make a scatter plot of the relationship, with labels for the counties and a smoothed fitted line added to the plot. Link the size of the points to population density.
- Comment on the relationship in terms of form, direction, strength, and outliers.
- Even if the relationship is clearly not linear, make the residual plots anyway and comment on whether you can catch the non-linearity or a violation of one of the other model assumptions using the plots.
- 10 marks for the full question with marks deducted for insufficient examination in any of these steps.

Solutions:

Some of the plots they might come up with:

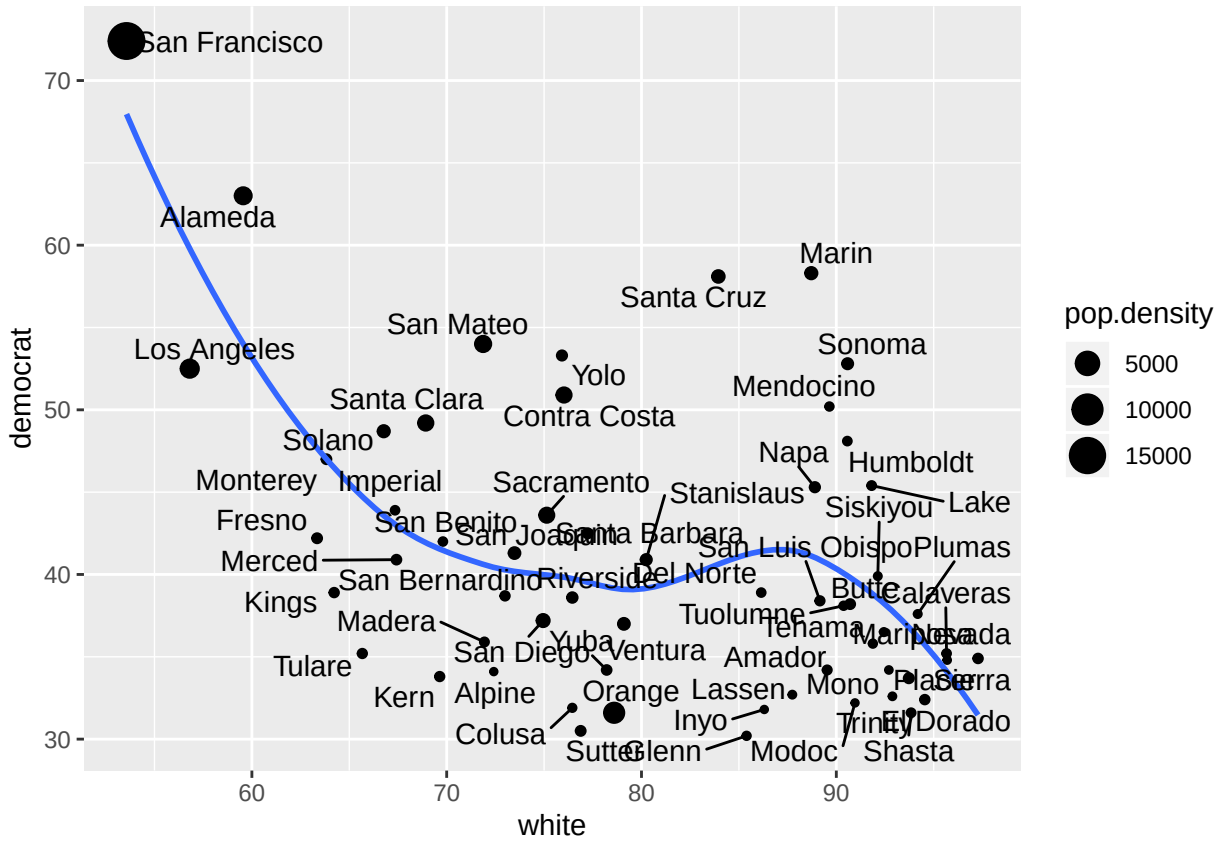
```
ggplot(counties_CA, aes(x = black, y = democrat)) +
  geom_point(aes(size = pop.density)) +
  geom_smooth(se = F) +
  geom_text_repel(aes(label = county))
```

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



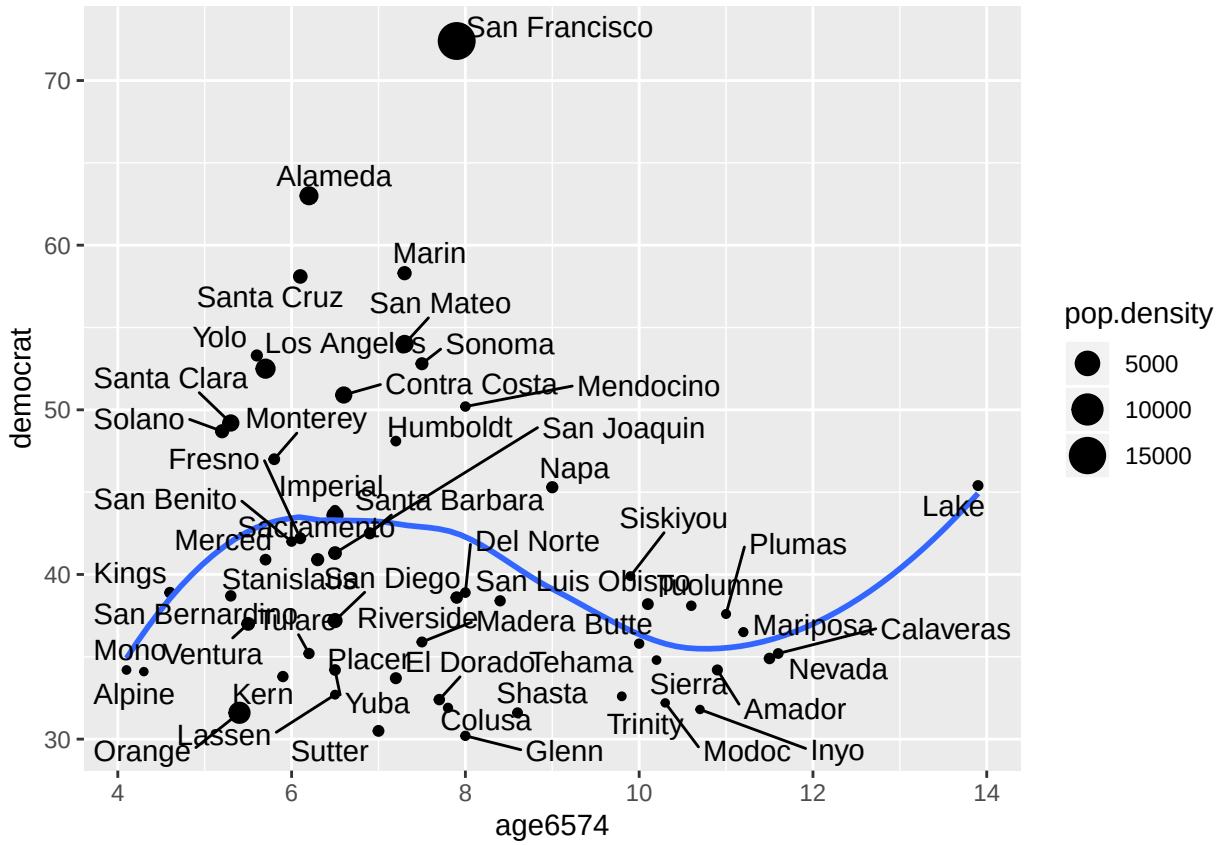

```
ggplot(counties_CA, aes(x = white, y = democrat)) +
  geom_point(aes(size = pop.density)) +
  geom_smooth(se = F) +
  geom_text_repel(aes(label = county))
```

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



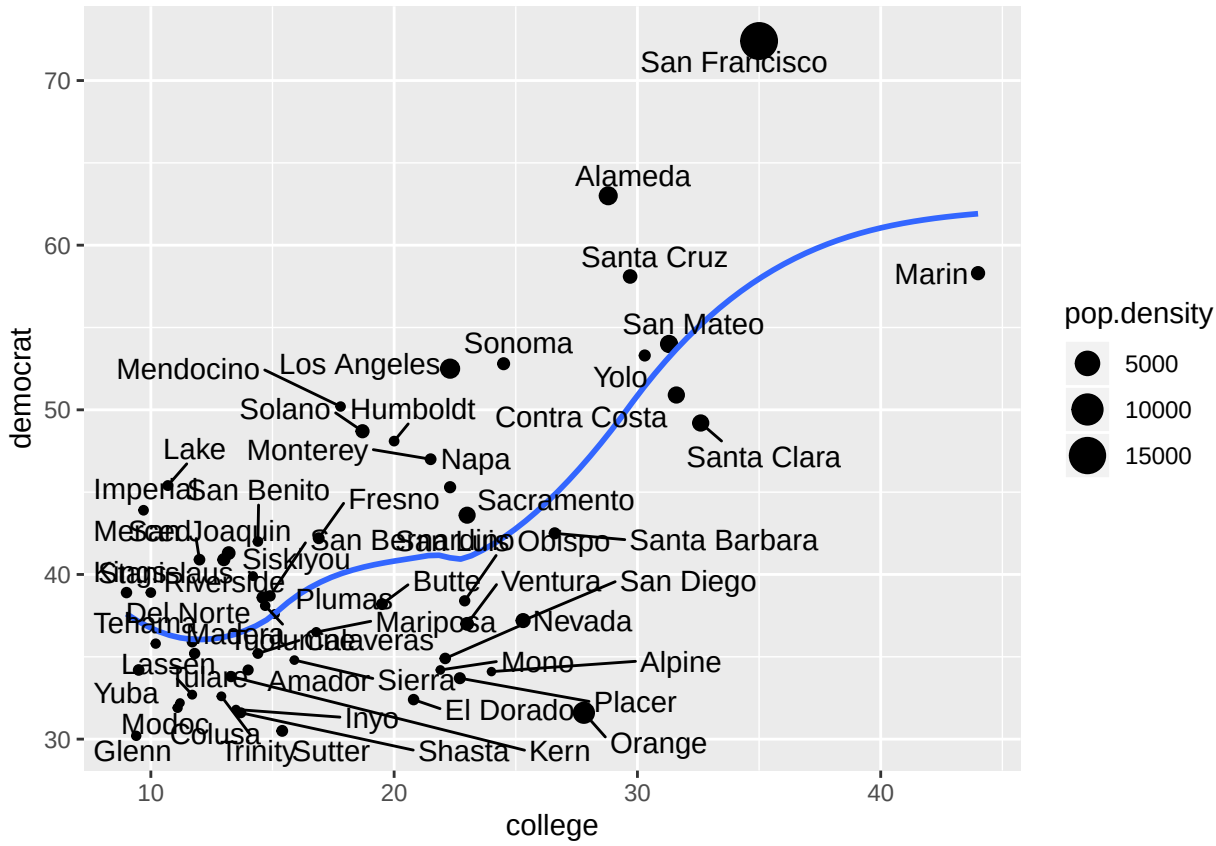

```
ggplot(counties_CA, aes(x = age6574, y = democrat)) +
  geom_point(aes(size = pop.density)) +
  geom_smooth(se = F) +
  geom_text_repel(aes(label = county))
```

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



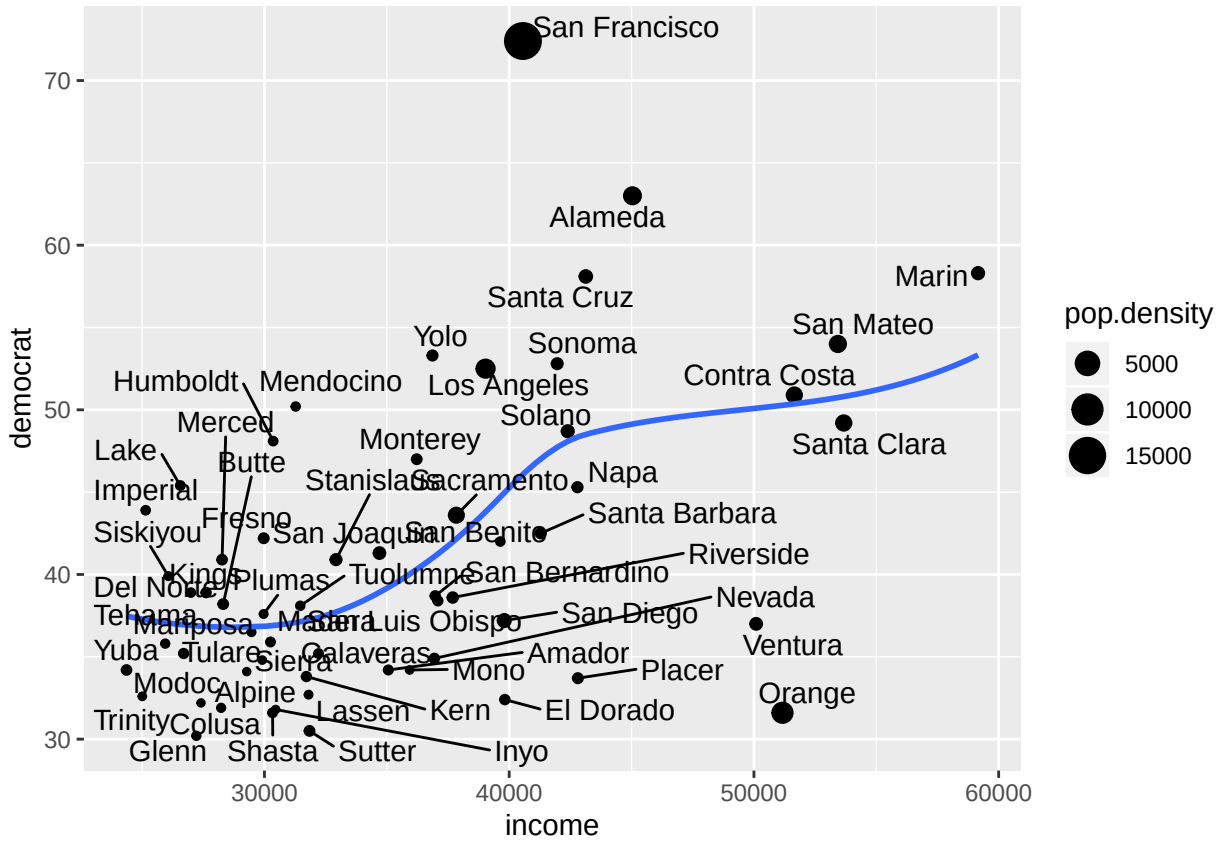

```
ggplot(counties_CA, aes(x = college, y = democrat)) +
  geom_point(aes(size = pop.density)) +
  geom_smooth(se = F) +
  geom_text_repel(aes(label = county))
```

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



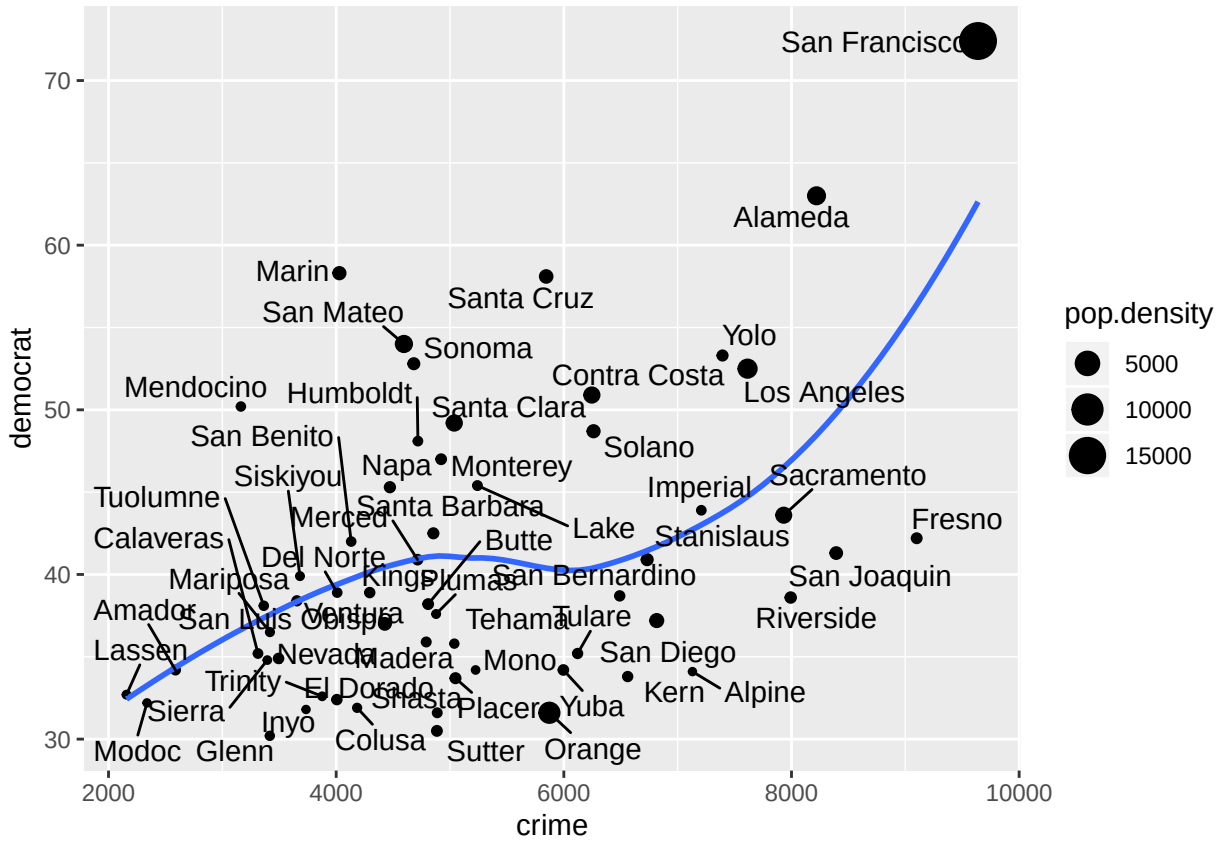

```
ggplot(counties_CA, aes(x = income, y = democrat)) +
  geom_point(aes(size = pop.density)) +
  geom_smooth(se = F) +
  geom_text_repel(aes(label = county))
```

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



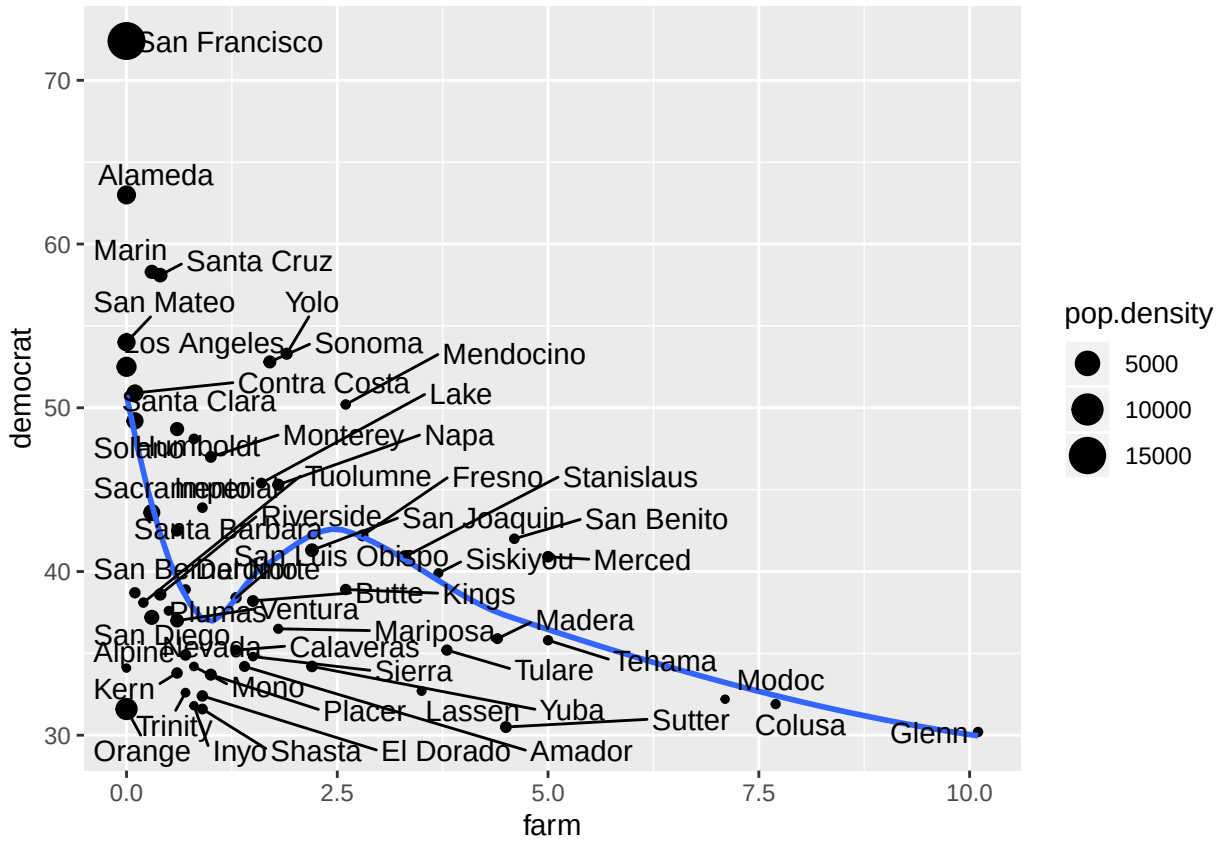

```
ggplot(counties_CA, aes(x = crime, y = democrat)) +
  geom_point(aes(size = pop.density)) +
  geom_smooth(se = F) +
  geom_text_repel(aes(label = county))
```

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




```
ggplot(counties_CA, aes(x = farm, y = democrat)) +
  geom_point(aes(size = pop.density)) +
  geom_smooth(se = F) +
  geom_text_repel(aes(label = county))
```

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


San Francisco
Alameda
Marin
Santa Cruz
San Mateo
Yolo
Los Angeles
Sonoma
Mendocino
Contra Costa
Lake
Santa Clara
Monterey
Napa
Tuolumne
Fresno
Stanislaus
Riverside
San Joaquin
San Benito
Santa Barbara
San Luis Obispo
Siskiyou
Merced
Butte
Kings
Ventura
Madera
Mariposa
San Diego
Calaveras
Tehama
Sierra
Tulare
Modoc
Kern
Mono
Placer
Lassen
Yuba
Sutter
Colusa
Glenn
Trinity
Orange
Inyo
Shasta
El Dorado
Amador
democrat
farm
pop.density
5000
10000
15000
70
60
50
40
30
0.0
2.5
5.0
7.5
10.0