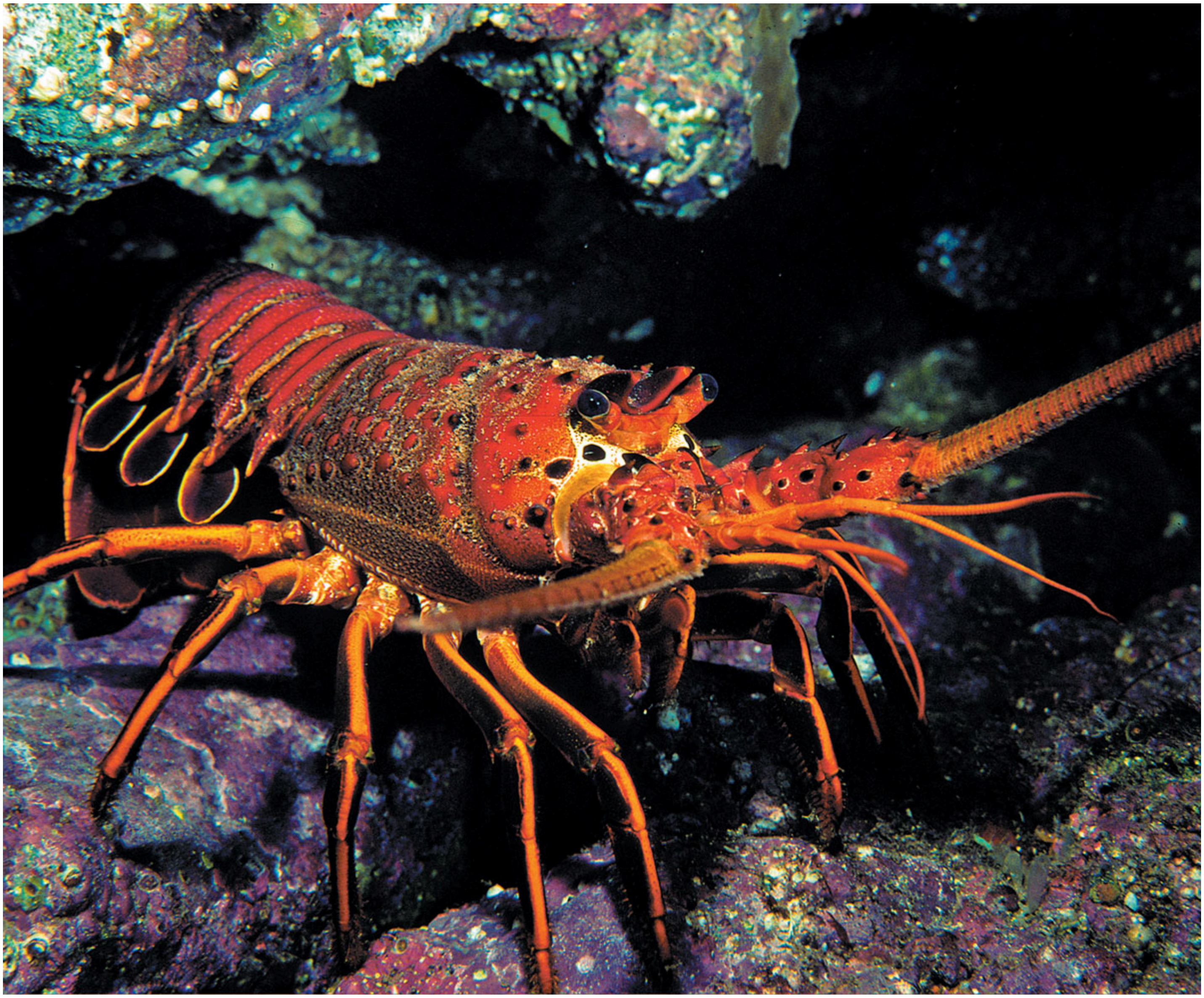
Assignment 1: California Spiny Lobster Abundance (*Panulirus Interruptus*)

Assessing the Impact of Marine Protected Areas (MPAs) at 5 Reef Sites in Santa Barbara County

Carmen Hoyt

1/8/2025 (Due 1/25/25)



Assignment instructions:

- Working with partners to troubleshoot code and concepts is encouraged! If you work with a partner, please list their name next to yours at the top of your assignment so Annie and I can easily see who collaborated.
- All written responses must be written independently (**in your own words**).
- Please follow the question prompts carefully and include only the information each question asks in your submitted responses.
- Submit both your knitted document and the associated `RMarkdown` or `Quarto` file.
- Your knitted presentation should meet the quality you'd submit to research colleagues or feel confident sharing publicly. Refer to the rubric for details about presentation standards.

Assignment submission: Carmen Hoyt

```r
# Load packages
library(tidyverse)
library(here)
library(janitor)
library(estimatr)
library(performance)
library(jtools)
library(gt)
library(gtsummary)
library(MASS) ## NOTE: The `select()` function is masked. Use: `dplyr::select()`
##
library(interactions)
library(ggridges)
library(beeswarm)
```

DATA SOURCE:

Reed D. 2019. SBC LTER: Reef: Abundance, size and fishing effort for California Spiny Lobster (Panulirus interruptus), ongoing since 2012. Environmental Data Initiative. https://doi.org/10.6073/pasta/a593a675d644fdefb736750b291579a0 (https://doi.org/10.6073/pasta/a593a675d644fdefb736750b291579a0). Dataset accessed 11/17/2019.

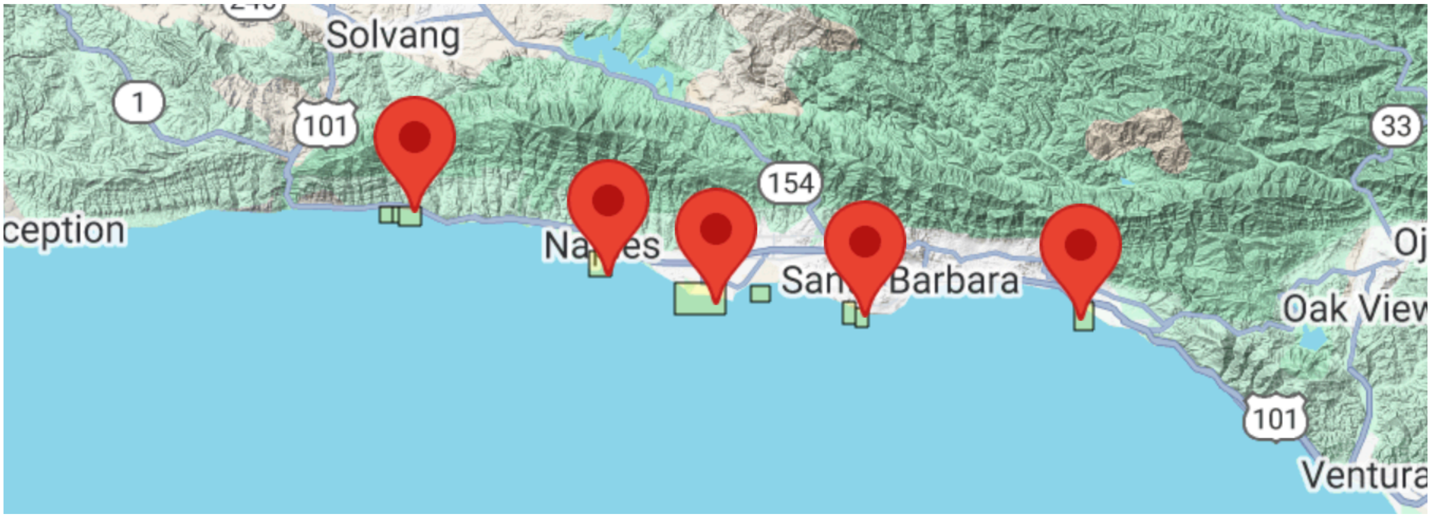
Introduction

You're about to dive into some deep data collected from five reef sites in Santa Barbara County, all about the abundance of California spiny lobsters! 🦞 Data was gathered by divers annually from 2012 to 2018 across Naples, Mohawk, Isla Vista, Carpinteria, and Arroyo Quemado reefs.

Why lobsters? Well, this sample provides an opportunity to evaluate the impact of Marine Protected Areas (MPAs) established on January 1, 2012 (Reed, 2019). Of these five reefs, Naples, and Isla Vista are MPAs, while the other three are not protected (non-MPAs). Comparing lobster health between these protected and non-protected areas gives us the chance to study how commercial and recreational fishing might impact these ecosystems.

We will consider the MPA sites the `treatment` group and use regression methods to explore whether protecting these reefs really makes a difference compared to non-MPA sites (our control group). In this assignment, we'll think deeply about which causal inference assumptions hold up under the research design and identify where they fall short.

Let's break it down step by step and see what the data reveals! 📊



Step 1: Anticipating potential sources of selection bias

a. Do the control sites (Arroyo Quemado, Carpenteria, and Mohawk) provide a strong counterfactual for our treatment sites (Naples, Isla Vista)? Write a paragraph making a case for why this comparison is centris paribus or whether selection bias is likely (be specific!).

The Isla Vista research site appears to be much larger than the other sites, which can influence abundance measures (by inflating counts) if not properly accounted for. Additionally, the Naples and Isla Vista sites are closer together, potentially introducing some bias in location/habitat similarity as compared to the control group, which is spread out over more coastline.

Step 2: Read & wrangle data

a. Read in the raw data. Name the data.frame (`df`) `rawdata`

b. Use the function `clean_names()` from the `janitor` package

```
# HINT: check for coding of missing values (`na = "-99999"`)
# Load data
rawdata <- read_csv(here("data", "spiny_abundance_sb_18.csv"), na = c("-99999"))
%>%
    clean_names()
```

c. Create a new `df` named `tidyata`. Using the variable `site` (reef location) create a new variable `reef` as a `factor` and add the following labels in the order listed (i.e., re-order the `levels`):

```
"Arroyo Quemado", "Carpenteria", "Mohawk", "Isla Vista",  "Naples"
```

```
tidydata <- rawdata %>%
    mutate(reef = factor(site))

levels(tidydata$reef) = c("Arroyo Quemado", "Carpenteria", "Isla Vista", "Mohaw
k", "Naples")
```

Create new `df` named `spiny_counts`

d. Create a new variable `counts` to allow for an analysis of lobster counts where the unit-level of observation is the total number of observed lobsters per `site`, `year` and `transect`.

- Create a variable `mean_size` from the variable `size_mm`
- NOTE: The variable `counts` should have values which are integers (whole numbers).
- Make sure to account for missing cases (`na`)!

e. Create a new variable `mpa` with levels `MPA` and `non_MPA`. For our regression analysis create a numerical variable `treat` where MPA sites are coded `1` and non_MPA sites are coded `0`

```
#HINT(d): Use `group_by()` & `summarize()` to provide the total number of lobster
s observed at each site-year-transect row-observation.

#HINT(e): Use `case_when()` to create the 3 new variable columns

spiny_counts <- tidydata %>%
    group_by(site, year, transect) %>%
    summarize(counts = sum(count, na.rm = TRUE),
           mean_size = mean(size_mm, na.rm = TRUE)) %>%
    mutate(mpa = case_when(
        site == "IVEE" ~ "MPA",
        site == "NAPL" ~ "MPA",
        site == "AQUE" ~ "non_MPA",
        site == "CARP" ~ "non_MPA",
        site == "MOHK" ~ "non_MPA")) %>%
    mutate(treat = case_when(
        mpa == "MPA" ~ 1,
        mpa == "non_MPA" ~ 0)) %>%
    ungroup()
```

> NOTE: This step is crucial to the analysis. Check with a friend or come to TA/instructor office hours to make sure the counts are coded correctly!

Step 3: Explore & visualize data

a. Take a look at the data! Get familiar with the data in each `df` format (`tidydata` , `spiny_counts`)

b. We will focus on the variables `count` , `year` , `site` , and `treat` (`mpa`) to model lobster abundance. Create the following 4 plots using a different method each time from the 6 options provided. Add a layer (`geom`) to each of the plots including informative descriptive statistics (you choose; e.g., mean, median, SD, quartiles, range). Make sure each plot dimension is clearly labeled (e.g., axes, groups).

- Density plot (https://r-charts.com/distribution/density-plot-group-ggplot2)
- Ridge plot (https://r-charts.com/distribution/ggridges/)
- Jitter plot (https://ggplot2.tidyverse.org/reference/geom_jitter.html)
- Violin plot (https://r-charts.com/distribution/violin-plot-group-ggplot2)
- Histogram (https://r-charts.com/distribution/histogram-density-ggplot2/)
- Beeswarm (https://r-charts.com/distribution/beeswarm/)

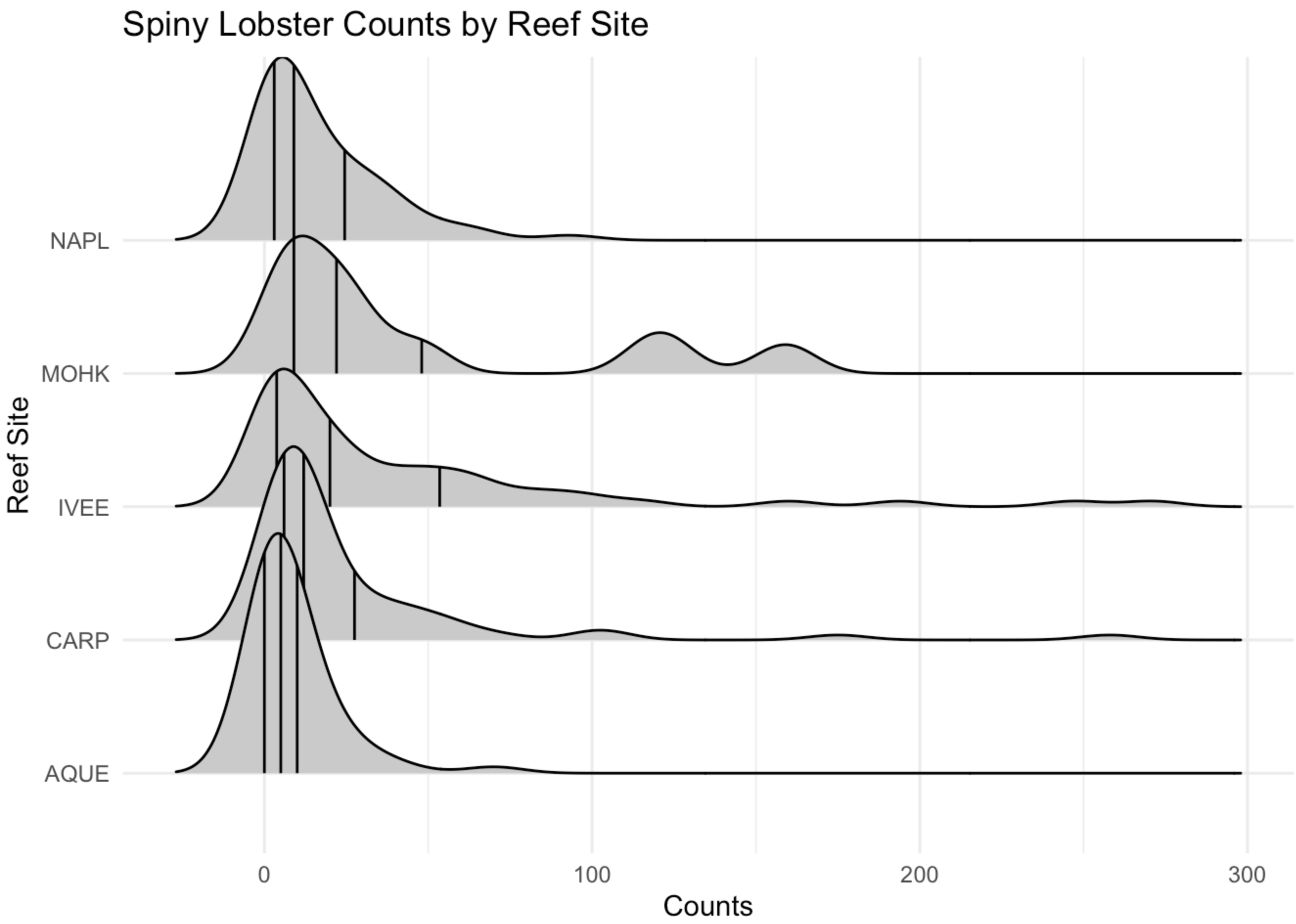
Create plots displaying the distribution of lobster **counts**:

1. grouped by reef site
2. grouped by MPA status
3. grouped by year

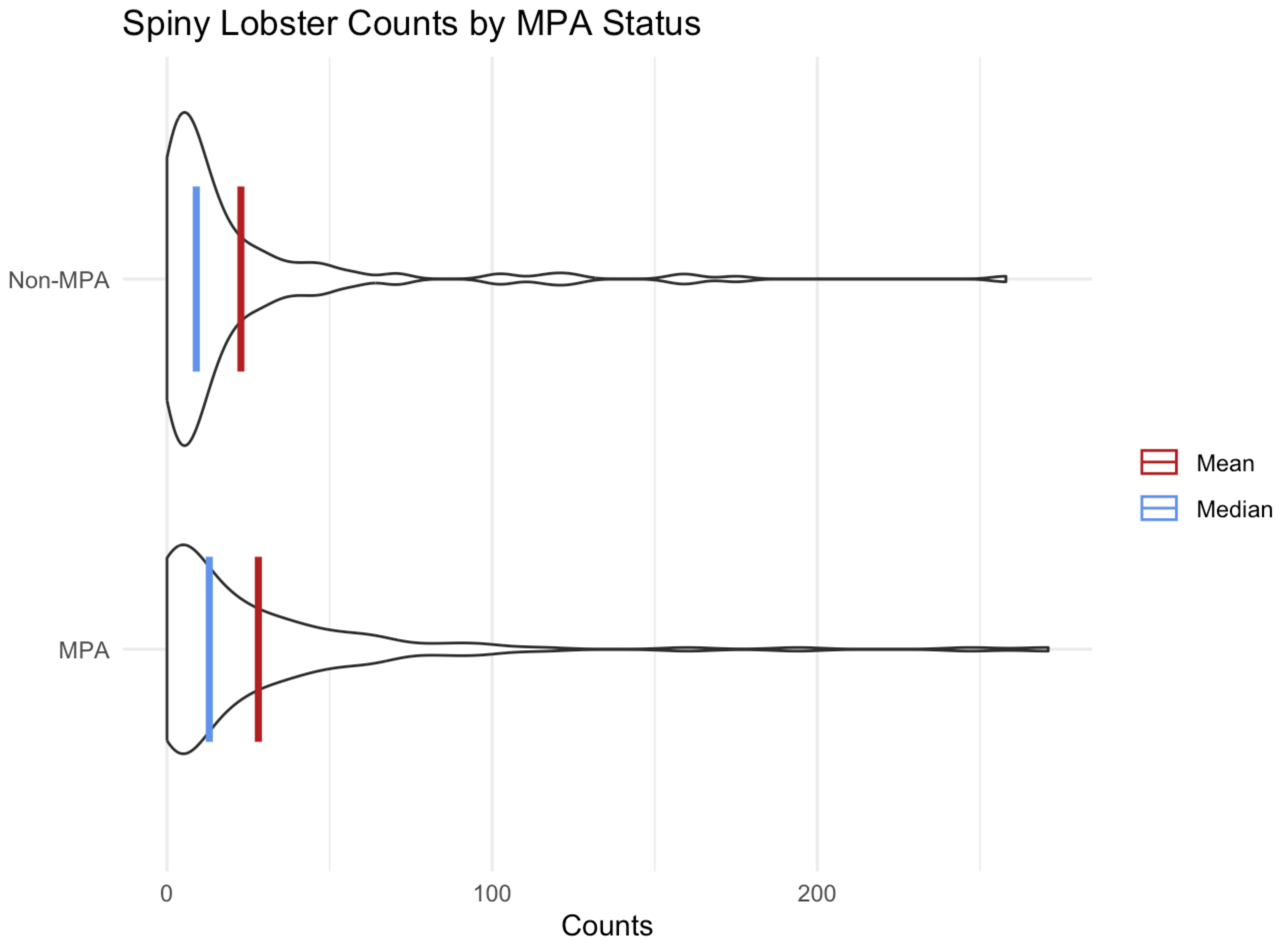
Create a plot of lobster **size** :

4. You choose the grouping variable(s)!

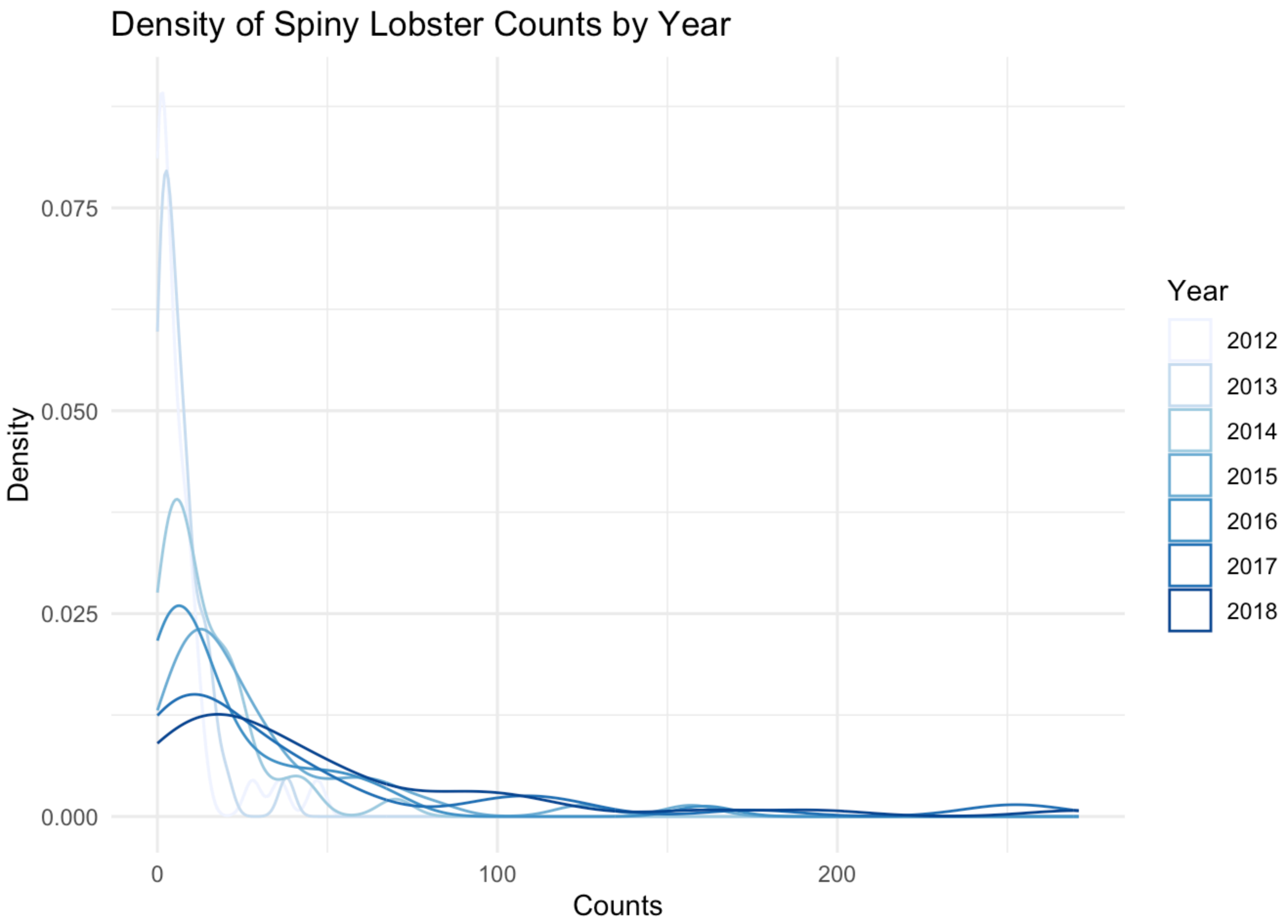
```
# plot 1:
spiny_counts %>%
ggplot(aes(x = counts, y = site)) +
    geom_density_ridges(quantile_lines = TRUE, fill = "gray80") +
    labs(x = "Counts",
         y = "Reef Site",
         title = "Spiny Lobster Counts by Reef Site",
         fill = "Site") +
    theme_minimal()
```



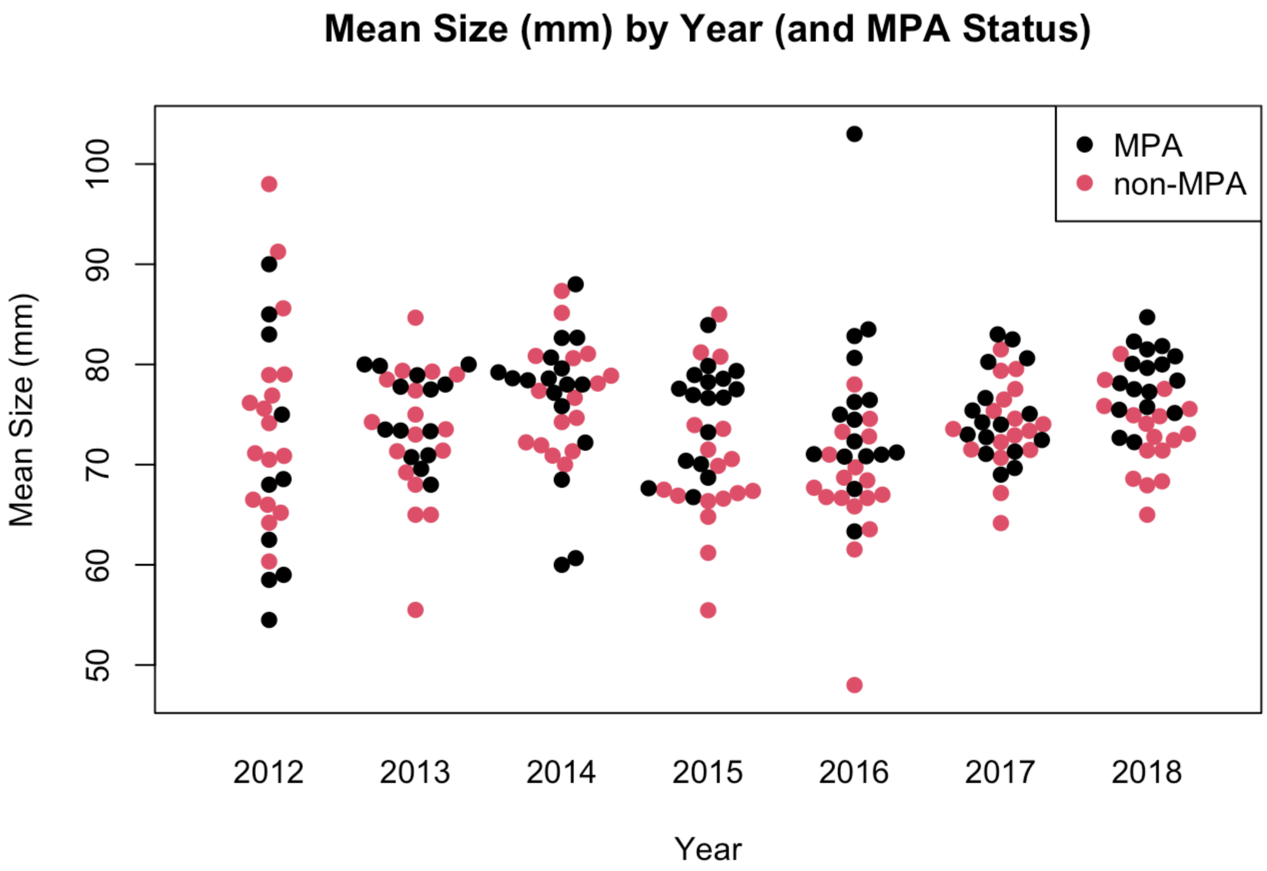
```
# plot 2:
spiny_counts %>%
    ggplot(aes(x = mpa, y = counts)) +
    geom_violin() +
    stat_summary(fun = "mean",
               geom = "crossbar",
               width = 0.5,
               aes(colour = "Mean")) +
    stat_summary(fun = "median",
               geom = "crossbar",
               width = 0.5,
               aes(color = "Median")) +
    scale_colour_manual(values = c("firebrick", "cornflowerblue"),
                     name = "") +
    scale_x_discrete(breaks=c("MPA","non_MPA"),
        labels=c("MPA", "Non-MPA")) +
    labs(x = "",
         y = "Counts",
         title = "Spiny Lobster Counts by MPA Status",
         fill = "") +
    coord_flip() +
    theme_minimal()
```



```
# plot 3:
spiny_counts %>%
    ggplot(aes(x = counts, color = factor(year))) +
    geom_density() +
    labs(x = "Counts",
         y = "Density",
         color = "Year",
         title = "Density of Spiny Lobster Counts by Year") +
    scale_color_brewer(palette = "Blues") +
    theme_minimal()
```



```
# plot 4:
beeswarm(mean_size ~ year, data = spiny_counts,
         pch = 19,
         pwcol = as.factor(mpa),
         xlab = "Year",
         ylab = "Mean Size (mm)",
         main = "Mean Size (mm) by Year (and MPA Status)")
legend("topright", legend = c("MPA", "non-MPA"),
       col = 1:2, pch = 19)
```



c. Compare means of the outcome by treatment group. Using the `tbl_summary()` function from the package `gt_summary` (https://www.danieldsjoberg.com/gtsummary/articles/tbl_summary.html)

```
# USE: gt_summary::tbl_summary()
# Compare the means of the counts by treatment group
spiny_counts %>%
    #ungroup() %>%
    dplyr::select(treat, counts) %>%
    tbl_summary(
        by = treat,
        statistic = list(all_continuous() ~ "{mean} ({sd})")) %>%
    modify_header(label ~ "**Variable**") %>%
    modify_spanning_header(c("stat_1", "stat_2") ~ "**Treatment**")
```

	Treatment	
Variable	**0** N = 133[1]	**1** N = 119[1]
counts	23 (39)	28 (44)

[1] Mean (SD)

Step 4: OLS regression- building intuition

a. Start with a simple OLS estimator of lobster counts regressed on treatment. Use the function `summ()` from the `jtools` (https://jtools.jacob-long.com/) package to print the OLS output

b. Interpret the intercept & predictor coefficients *in your own words*. Use full sentences and write your interpretation of the regression results to be as clear as possible to a non-academic audience.

```
# NOTE: We will not evaluate/interpret model fit in this assignment (e.g., R-squa
re)

m1_ols <- lm(counts ~ treat, spiny_counts)

summ(m1_ols, model.fit = FALSE)
```

Observations	252
Dependent variable	counts
Type	OLS linear regression

	Est.	S.E.	t val.	p
(Intercept)	22.73	3.57	6.36	0.00
treat	5.36	5.20	1.03	0.30

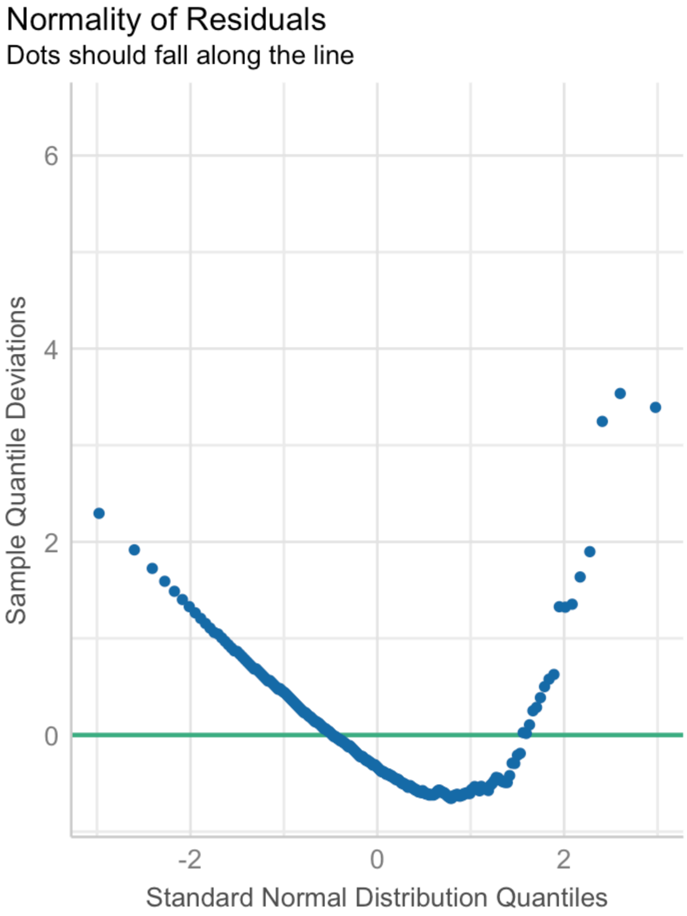
Standard errors: OLS

The ‘intercept’ coefficient is the value of lobster counts when the treatment is 0 (non-MPA); thus, there are an average of 22.73 lobster observations for the non-MPA group. Knowing this, we determine from the ‘treat’ coefficient that there are an average of 5.36 *more* lobster observations when the treatment is 1 (MPAs); or about 28 lobster observations for the MPA group.

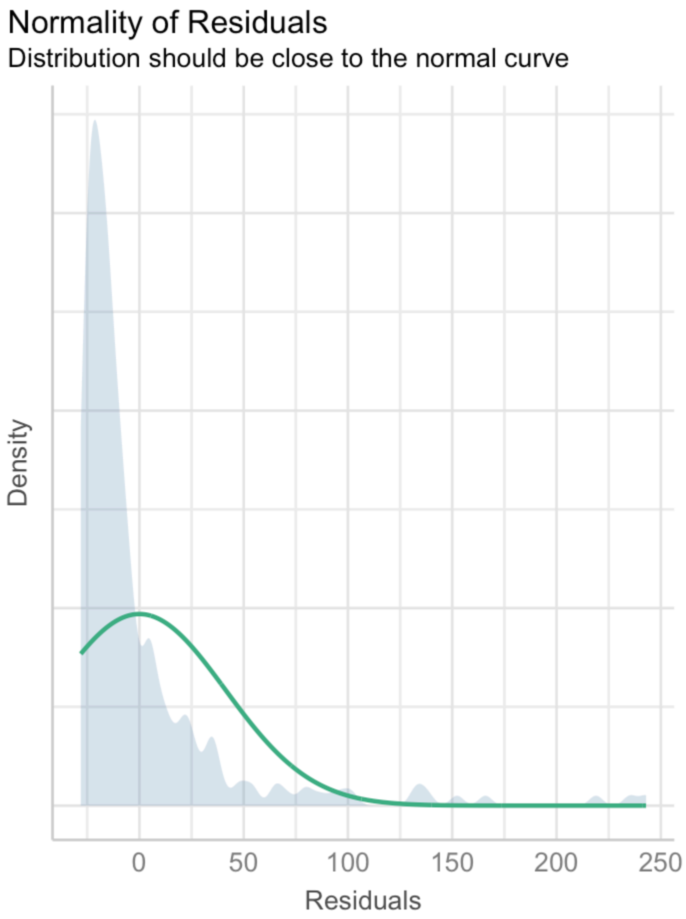
c. Check the model assumptions using the `check_model` function from the `performance` package

d. Explain the results of the 4 diagnostic plots. Why are we getting this result?

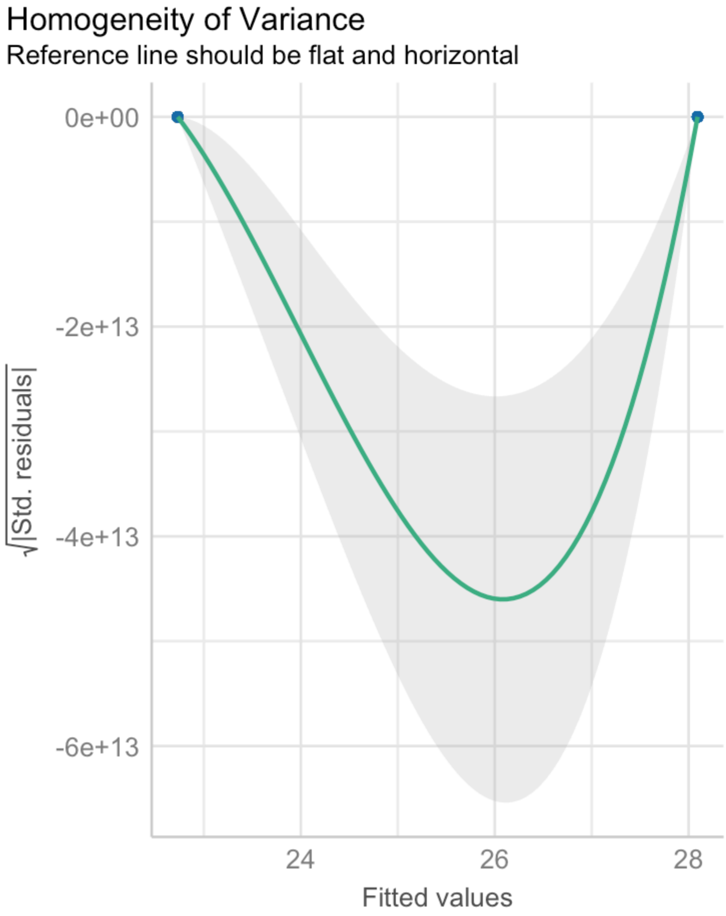
```
check_model(m1_ols,  check = "qq" )
```



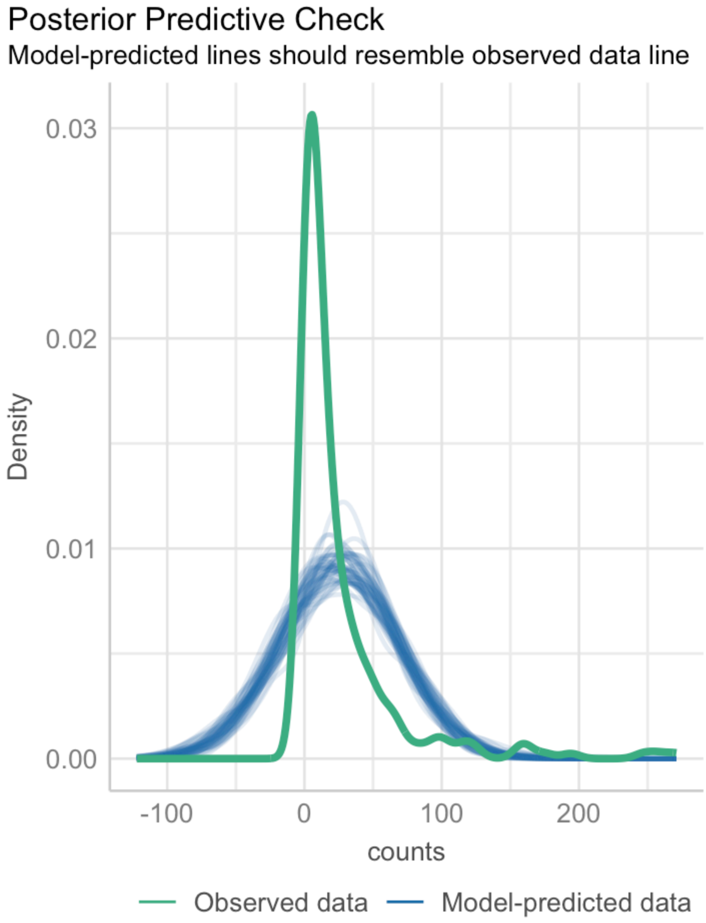
```
check_model(m1_ols, check = "normality")
```



```
check_model(m1_ols, check = "homogeneity")
```



```
check_model(m1_ols, check = "pp_check")
```



The four diagnostic plots are showing us that are data are not normal (right-skewed). Therefore, we are violating an assumption of OLS and can deduce that this model might not be the best fit.

Step 5: Fitting GLMs

a. Estimate a Poisson regression model using the `glm()` function

b. Interpret the predictor coefficient in your own words. Use full sentences and write your interpretation of the results to be as clear as possible to a non-academic audience.

c. Explain the statistical concept of dispersion and overdispersion in the context of this model.

d. Compare results with previous model, explain change in the significance of the treatment effect

```
#HINT1: Incidence Ratio Rate (IRR): Exponentiation of beta returns coefficient wh
ich is interpreted as the 'percent change' for a one unit increase in the predict
or

#HINT2: For the second glm() argument `family` use the following specification op
tion `family = poisson(link = "log")`

m2_pois <- glm(counts~treat,
               family = poisson(link = "log"),
               data = spiny_counts)

summ(m2_pois, model.fit = FALSE)
```

Observations	252
Dependent variable	counts
Type	Generalized linear model
Family	poisson
Link	log

	Est.	S.E.	z val.	p
(Intercept)	3.12	0.02	171.74	0.00
treat	0.21	0.03	8.44	0.00

Standard errors: MLE

```
exp(0.21)-1 # model estimates 23% increase in lobster counts
```

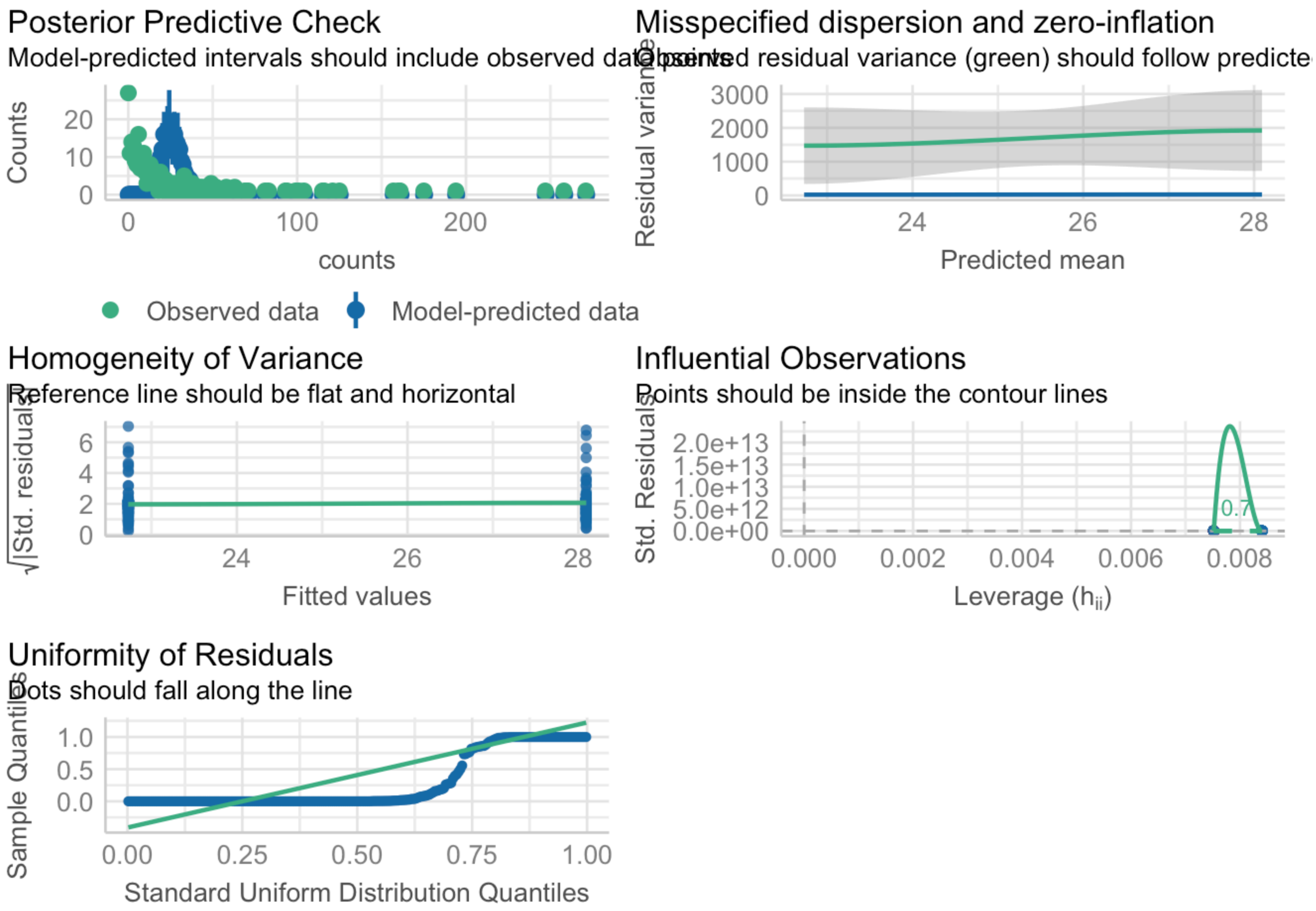
```
## [1] 0.2336781
```

The predictor coefficient is best interpreted once it is converted to a percent change. To achieve this, you must first exponentiate the coefficient (since it is on a log scale) and then subtract 1. The model estimates a 23% *increase* in lobster counts in the MPA treatment group vs. the non-MPA treatment group.

e. Check the model assumptions. Explain results.

f. Conduct tests for over-dispersion & zero-inflation. Explain results.

```
check_model(m2_pois)
```



```
check_overdispersion(m2_pois)
```

```
## # Overdispersion test
##
##        dispersion ratio =    67.033
##   Pearson's Chi-Squared = 16758.289
##                 p-value =   < 0.001
```

```
check_zeroinflation(m2_pois)
```

```
## # Check for zero-inflation
##
##    Observed zeros: 27
##   Predicted zeros: 0
##             Ratio: 0.00
```

Overdispersion was detected, meaning variance of the response variable (in this case `counts`) is significantly larger than the mean. This violates the poisson assumption that variance is proportional to the mean, so a poisson model may not be the best fit for the data. This overdispersion could be a result of zero-inflation (an excess of 0 lobster counts) as it was detected in the model.

g. Fit a negative binomial model using the function glm.nb() from the package `MASS` and check model diagnostics

h. In 1-2 sentences explain rationale for fitting this GLM model.

i. Interpret the treatment estimate result in your own words. Compare with results from the previous model.

```
# NOTE: The `glm.nb()` function does not require a `family` argument

m3_nb <- glm.nb(counts~treat,
                data = spiny_counts)

summ(m3_nb, model.fit = FALSE)
```

Observations	252
Dependent variable	counts
Type	Generalized linear model
Family	Negative Binomial(0.55)
Link	log

	Est.	**S.E.**	**z val.**	**p**
(Intercept)	3.12	0.12	26.40	0.00
treat	0.21	0.17	1.23	0.22

Standard errors: MLE

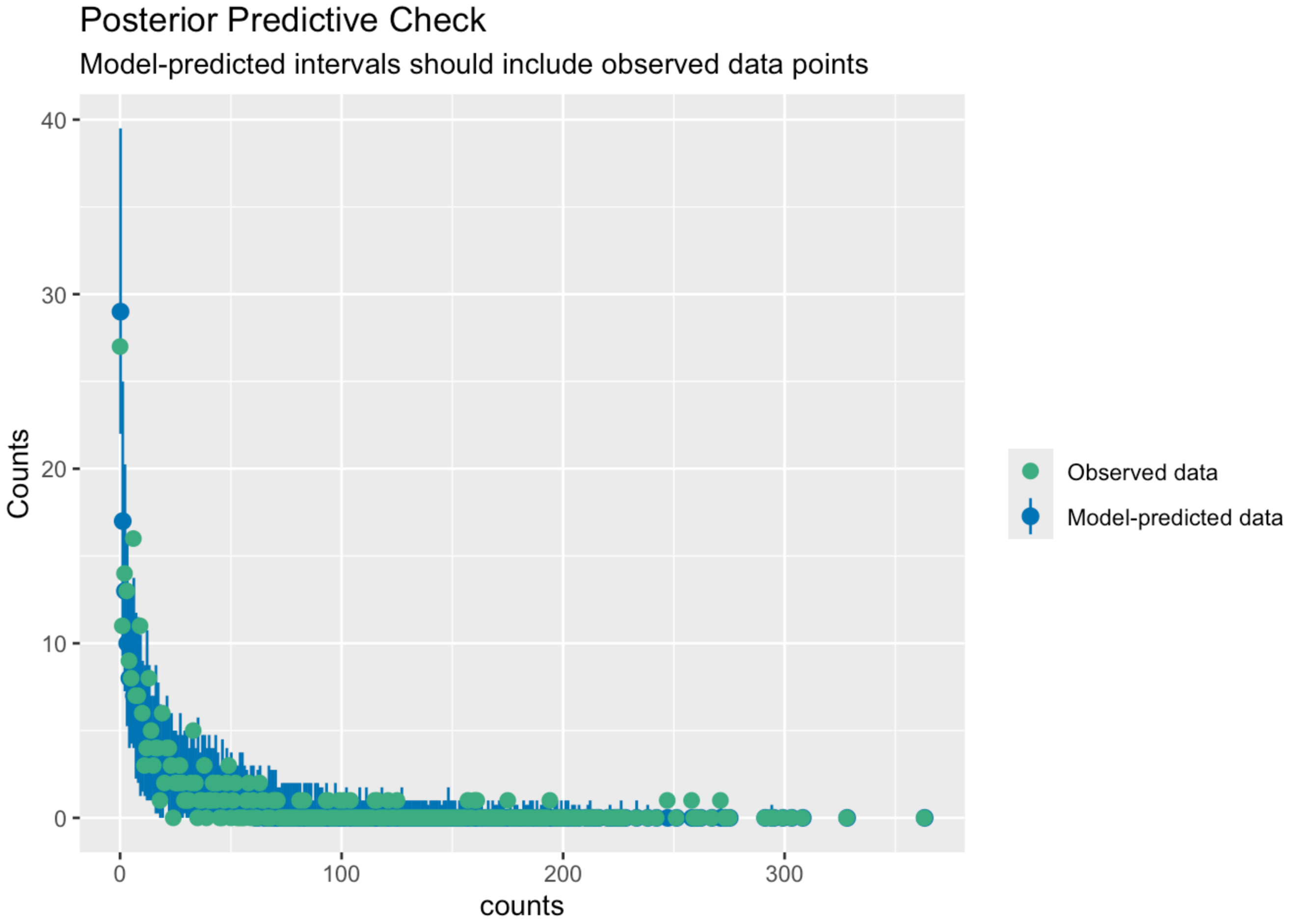
```
check_overdispersion(m3_nb)
```

```
## # Overdispersion test
##
##  dispersion ratio = 1.398
##           p-value = 0.088
```

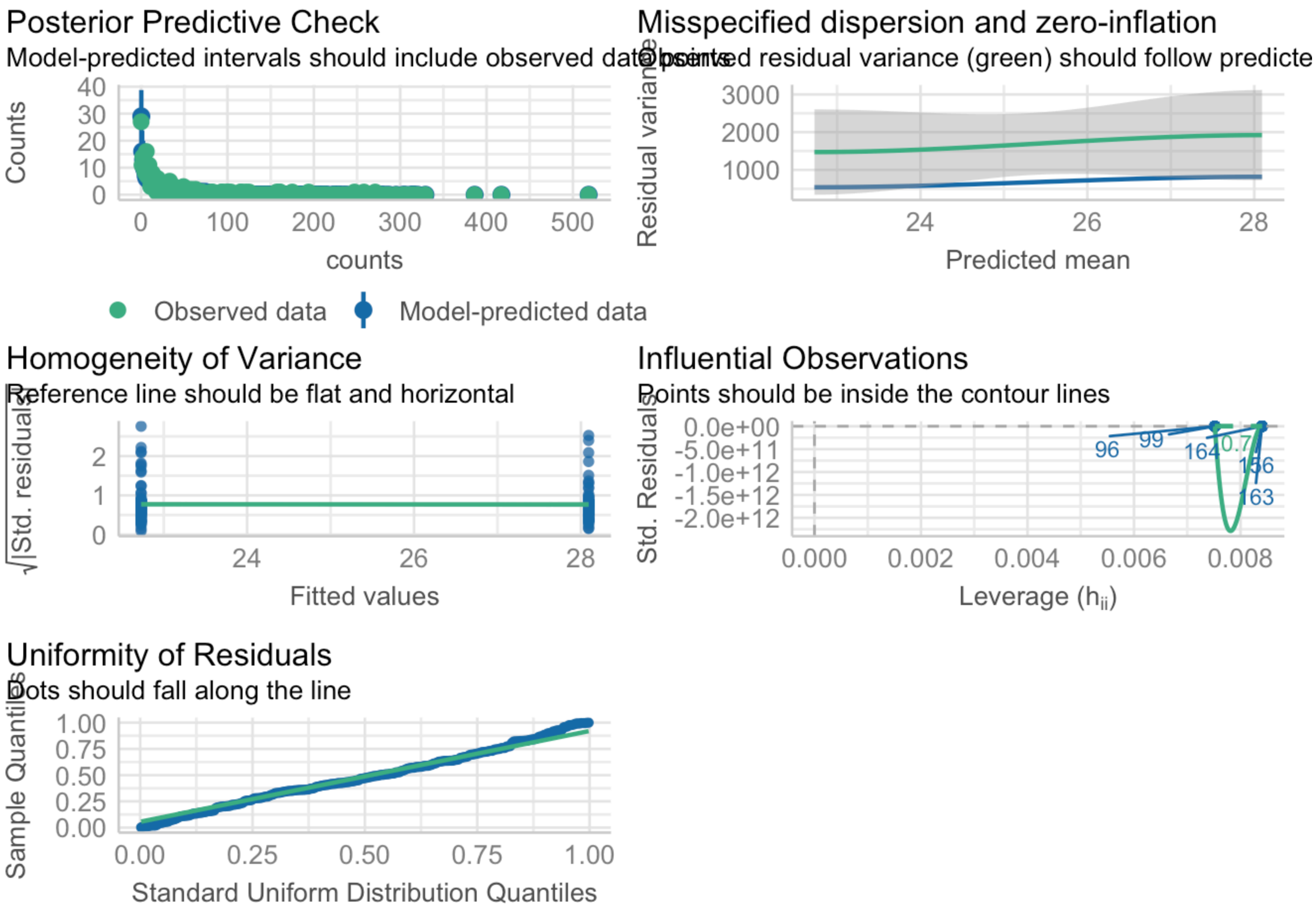
```
check_zeroinflation(m3_nb)
```

```
## # Check for zero-inflation
##
##    Observed zeros: 27
##   Predicted zeros: 30
##             Ratio: 1.12
```

```
check_predictions(m3_nb)
```



```
check_model(m3_nb)
```



A negative binomial model was fit to account for the overdispersion detected in the poisson model. The coefficients are the same, but the z-value for the predictor decreased while the p-value increased. This indicates that the predictor (treatment, or MPA status) might not have as significant of an impact on the response (lobster counts) in this model. Additionally, zero-inflation is still present.

Step 6: Compare models

a. Use the `export_summ()` function from the `jtools` package to look at the three regression models you fit side-by-side.

c. Write a short paragraph comparing the results. Is the treatment effect `robust` or stable across the model specifications.

```
export_summs(m1_ols, m2_pois, m3_nb, robust = "HC2",
             model.names = c("OLS","Poisson", "NB"),
             statistics = "none")
```

	OLS	Poisson	NB
(Intercept)	22.73 ***	3.12 ***	3.12 ***
	(3.34)	(0.15)	(0.15)

treat	5.36	0.21	0.21
	(5.24)	(0.21)	(0.21)

Standard errors are heteroskedasticity robust. *** p < 0.001; ** p < 0.01; * p < 0.05.

To understand whether or not the treatment effect is `robust`, we must first calculate percent change for the ols model. Since the coefficient told us that the are on average 5.36 more lobsters in MPAs than non-MPAs (where there are an average of 22.73 lobsters), we can divide 5.36/22.73 and get 23.5%. Upon interpreting both the Poisson and NB models, we found the percent change to be 23.4%. I would conclude that treatment effect is stable across the model specifications.

Step 7: Building intuition - fixed effects

a. Create new `df` with the `year` variable converted to a factor

b. Run the following OLS model using `lm()`

- Use the following specification for the outcome `log(counts+1)`
- Estimate fixed effects for `year`
- Include an interaction term between variables `treat` and `year`

c. Take a look at the regression output. Each coefficient provides a comparison or the difference in means for a specific sub-group in the data. Informally, describe the what the model has estimated at a conceptual level (NOTE: you do not have to interpret coefficients individually)

d. Explain why the main effect for treatment is negative? *Does this result make sense?

```
ff_counts <- spiny_counts %>%
    mutate(year=as_factor(year))

m5_fixedeffs <- glm.nb(
    counts~
        treat +
        year +
        treat*year,
    data = ff_counts)

summ(m5_fixedeffs, model.fit = FALSE)
```

Observations	252
Dependent variable	counts
Type	Generalized linear model
Family	Negative Binomial(0.8129)
Link	log

	Est.	S.E.	z val.	p
(Intercept)	2.35	0.26	8.89	0.00
treat	-1.72	0.42	-4.12	0.00
year2013	-0.35	0.38	-0.93	0.35
year2014	0.08	0.37	0.21	0.84
year2015	0.86	0.37	2.32	0.02
year2016	0.90	0.37	2.43	0.01
year2017	1.56	0.37	4.25	0.00
year2018	1.04	0.37	2.81	0.00
treat:year2013	1.52	0.57	2.66	0.01
treat:year2014	2.14	0.56	3.80	0.00
treat:year2015	2.12	0.56	3.79	0.00
treat:year2016	1.40	0.56	2.50	0.01
treat:year2017	1.55	0.56	2.77	0.01
treat:year2018	2.62	0.56	4.69	0.00

Standard errors: MLE

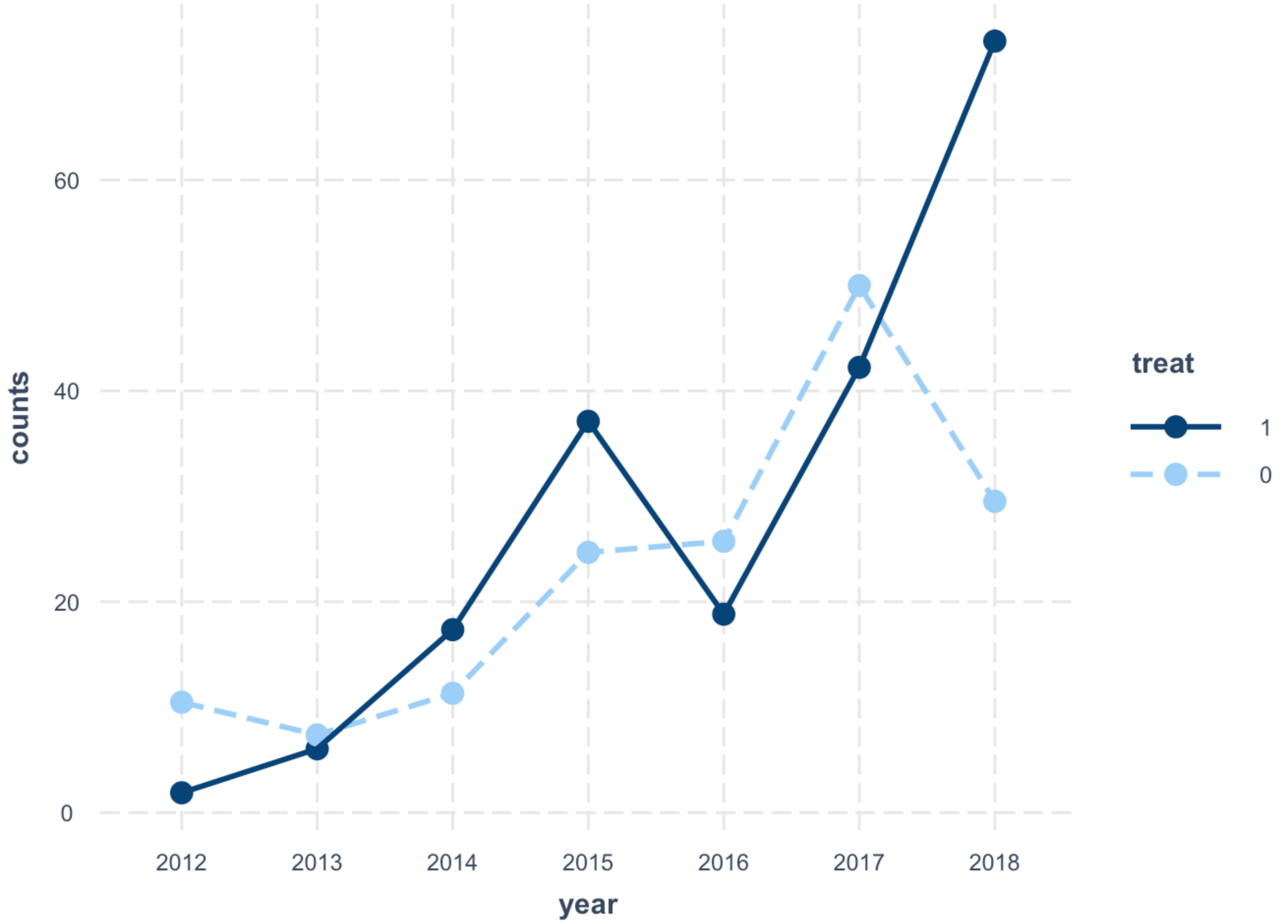
Conceptually, the model estimates mean lobster counts for each treatment group for each year accounting for an interaction between year and treatment. The main effect for treatment is negative because there were less lobsters predicted in the treatment group (1, MPA) than the control group (0, non-MPA) for the reference year 2012. This result makes sense because it will adjust on a year by year basis as determined by the coefficients produced for each year and treatment-year interaction.

e. Look at the model predictions: Use the `interact_plot()` function from package `interactions` to plot mean predictions by year and treatment status.

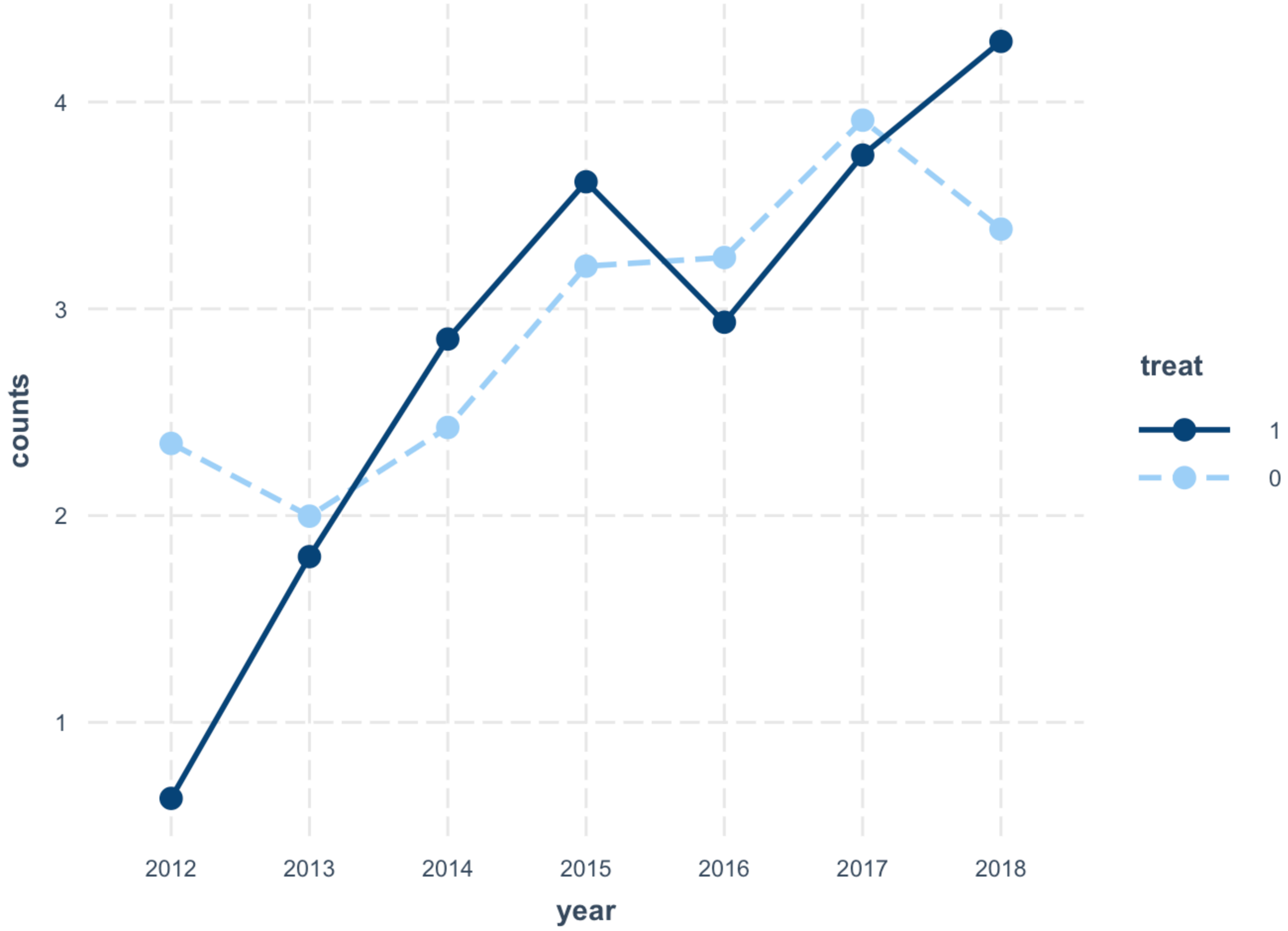
f. Re-evaluate your responses (c) and (b) above.

```
# Hint 1: Group counts by `year` and `mpa` and calculate the `mean_count`
# Hint 2: Convert variable `year` to a factor

interact_plot(m5_fixedeffs, pred = year, modx = treat,
              outcome.scale = "response")
```



```
interact_plot(m5_fixedeffs, pred = year, modx = treat,
              outcome.scale = "link")
```



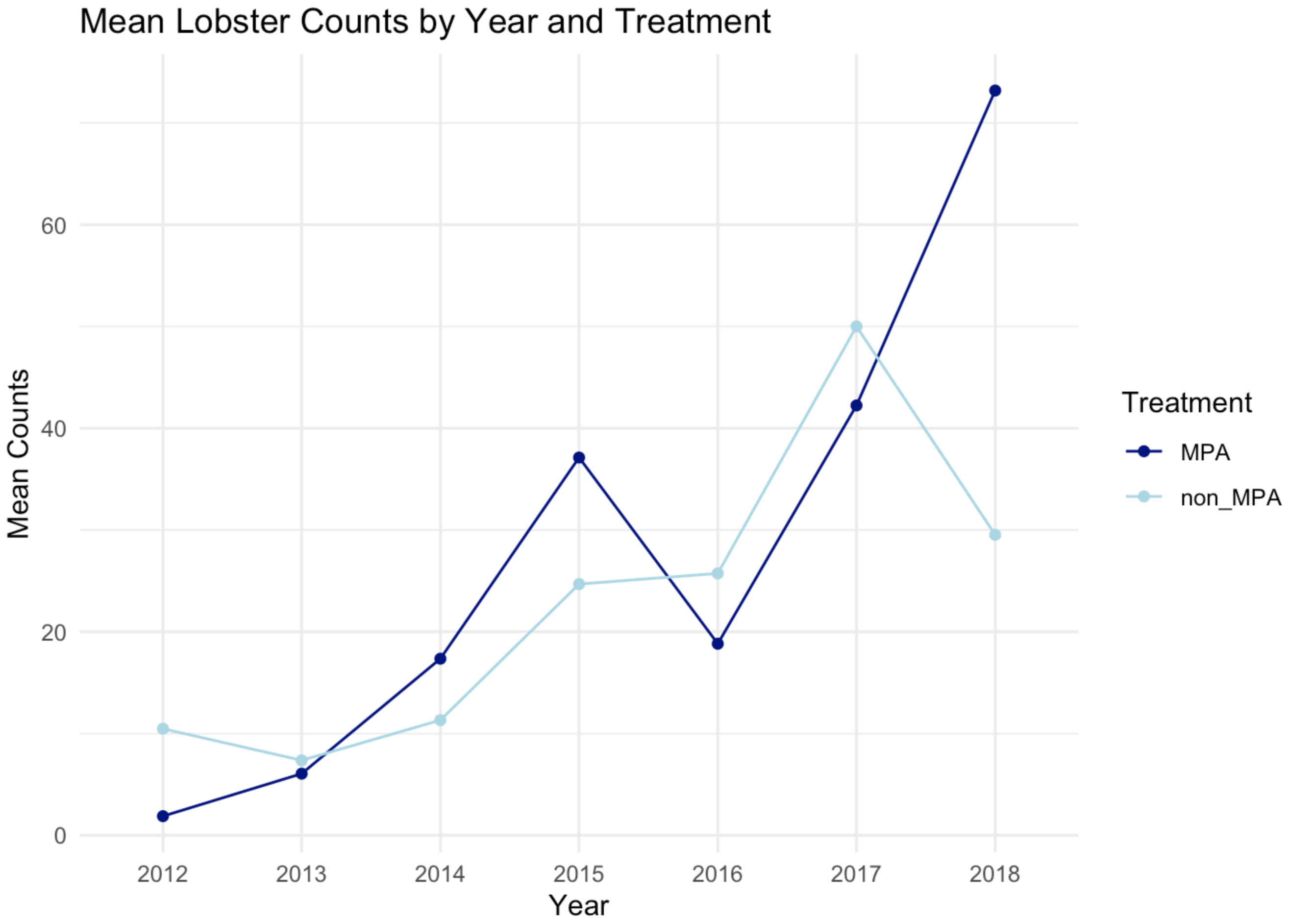
g. Using `ggplot()` create a plot in same style as the previous `interaction plot`, but displaying the original scale of the outcome variable (lobster counts). This type of plot is commonly used to show how the treatment effect changes across discrete time points (i.e., panel data).

The plot should have… - `year` on the x-axis - `counts` on the y-axis - `mpa` as the grouping variable

```
# Hint 1: Group counts by `year` and `mpa` and calculate the `mean_count`
# Hint 2: Convert variable `year` to a factor

plot_counts <- spiny_counts %>%
    mutate(year = as.factor(year)) %>%
    group_by(year, mpa) %>%
    summarise(mean_counts = mean(counts))

plot_counts %>%
    ggplot(aes(year, mean_counts, group = mpa)) +
    geom_line(aes(color = mpa)) +
    geom_point(aes(color = mpa)) +
    scale_linetype_manual(values = c("solid", "dashed")) +
    scale_color_manual(values = c("navyblue", "lightblue")) +
    labs(
        x = "Year",
        y = "Mean Counts",
        title = "Mean Lobster Counts by Year and Treatment",
        color = "Treatment"
    ) +
    theme_minimal()
```



Step 8: Reconsider causal identification assumptions

a. Discuss whether you think `spillover effects` are likely in this research context (see Glossary of terms; https://docs.google.com/document/d/1RIudsVcYhWGpqC-Uftk9UTz3PIq6stVyEpT44EPNgpE/edit?usp=sharing (https://docs.google.com/document/d/1RIudsVcYhWGpqC-Uftk9UTz3PIq6stVyEpT44EPNgpE/edit?usp=sharing))

There is certainly a possibility for spillover effects in this research context because the MPA/non-MPA boundaries are not physical and lobsters are very mobile. Thus, there is nothing stopping lobsters from the MPA from moving outside the MPA, especially if the MPA is doing what it is designed to do and the numbers of lobsters are increasing.

b. Explain why spillover is an issue for the identification of causal effects

Spillover is an issues for the identification of causal effects because it can muddle the difference between the control and treatment means, essentially making it difficult to identify the true impact of the treatment.

c. How does spillover relate to impact in this research setting?

In this research setting, spillover can raise the mean control (non-MPA) lobster counts. This makes it difficult to assess the true impact of the MPA treatment since the control group is artificially inflated to some unknown degree.

d. Discuss the following causal inference assumptions in the context of the MPA treatment effect estimator. Evaluate if each of the assumption are reasonable:

```
1)  SUTVA: Stable Unit Treatment Value assumption
    The SUTVA assumption is likely violated in this context, as the control group
(non–MPA) is likely indirectly affected by spillover from the treatment (MPA) gro
up.
2)  Excludability assumption
    It is unlikely that the excludability assumption is violated in this context,
as the models proved to be robust and thus the treatment effect was stable. If th
ere was another mechanism at play, outside of the proposed mechanisms, it would l
ikely manifest as differences in the models.
```

EXTRA CREDIT

> Use the recent lobster abundance data with observations collected up until 2024 (`lobster_sbchannel_24.csv`) to run an analysis evaluating the effect of MPA status on lobster counts using the same focal variables.

a. Create a new script for the analysis on the updated data
b. Run at least 3 regression models & assess model diagnostics
c. Compare and contrast results with the analysis from the 2012-2018 data sample (~ 2 paragraphs)