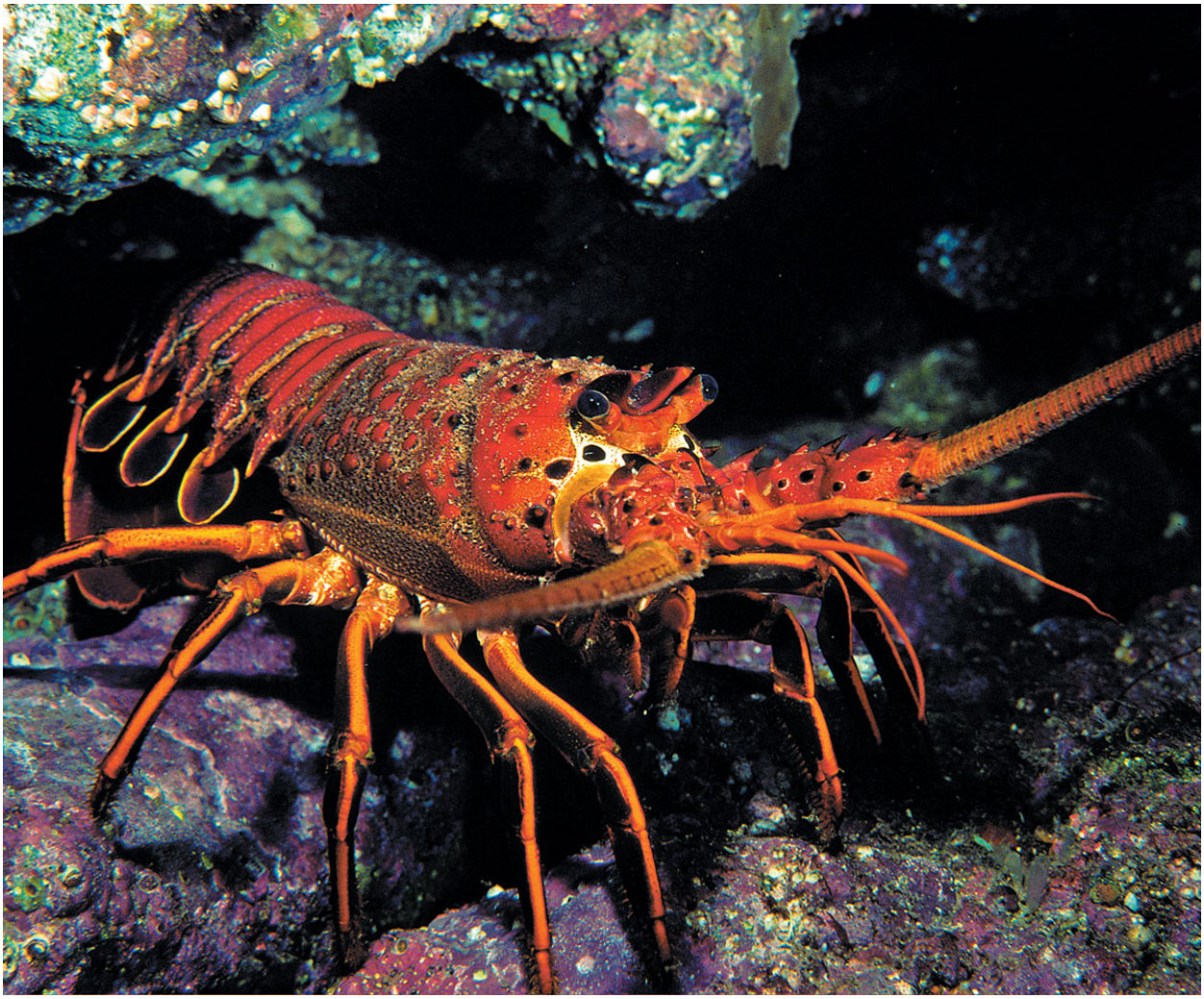
Assignment 1

California Spiny Lobster (*Panulirus Interruptus*): Assessing the Impact of Marine Protected Areas (MPAs) at 5 Reef Sites in Santa Barbara County

EDS 241 / ESM 244 (**Due: 1/17**)

1/8/26



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 (YOUR NAME): Sofia Sarak

```
# Load all necessary libraries
library(tidyverse)
library(here)
library(janitor)
library(estimatr)
library(performance)
library(jtools)
library(gt)
library(gtsummary)
library(interactions)
library(ggridges)
library(geomtextpath)
```

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. Data accessed 11/17/2019.

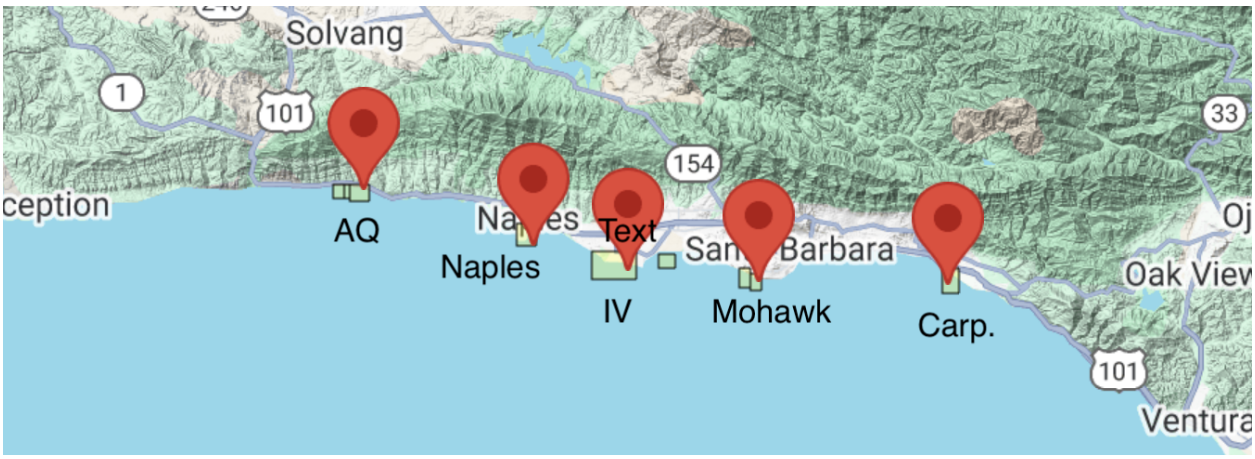
Introduction

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 ceteris paribus or whether selection bias is likely (be specific!).

This comparison is not ceteris paribus, but the control sites are relatively good options for the purposes of this evaluation. True ceteris paribus would be taking the *same* exact site and applying both the control and treatment to it, at the same time. Since this is impossible, selecting control sites that are spatially close to the treatment sites (increasing the odds that they are affected by similar currents, temperatures, human activity, and species biodiversity) are a good option. Despite this, selection bias must still be considered, since there are a number of variables *not* controlled for when comparing outside of ceteris paribus.

REVISION: It is important to highlight the selection bias that is present in this sample. Most notably, the treatment (MPA status) was not assigned randomly. The two sites that were assigned MPA status – Naples and Isla Vista – are treated in this study because the California Department of Fish and Wildlife felt that these areas needed protection. Specifically, that they contained sensitive ecosystems such as rocky reefs and kelp beds, and are even regarded as ecological hotspots. Because of this intentional selection of MPA sites, it is possible that the other three control sites may not have the same ecological diversity (or other conditions) as our treatment sites, potentially affecting our results. Thus, it is important to consider the effects of selection bias in our analysis.

Step 2: Read & wrangle data **a.** Read in the raw data from the "data" folder named `spiny_abundance_sb_18.csv`. Name the data.frame `rawdata`

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

```
# HINT: check for coding of missing values (`na = "-99999"`)

# Read in data from csv
rawdata <- read_csv(here("week1/data/spiny_abundance_sb_18.csv")) %>%

  # Put column names into lower snake case
  clean_names() %>%

  # Can use mutate (and reassign to original column name) in order to access column and replace NAs
  mutate(size_mm = na_if(size_mm, -99999))
```

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"`

```
# Pipe from original data
tidydata <- rawdata %>%

  # Add new column`reef`, with labels based on `site` column
  mutate(reef = factor(case_when(
    site == "AQUE" ~ "Arroyo Quemado",
    site == "CARP" ~ "Carpenteria",
    site == "MOHK" ~ "Mohawk",
    site == "IVEE" ~ "Isla Vista",
    site == "NAPL" ~ "Naples"))) %>%

  # Re-order factor levels to match provided order
  mutate(reef = factor(reef, levels = c("Arroyo Quemado", "Carpenteria", "Mohawk", "Isla Vista",  "Napl
```

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 lobsters observed at each sit

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

spiny_counts <- tidydata %>%

  # Dropping NAs in mean_size
  drop_na(size_mm) %>%

  # Create rows with all site-year-transect combinations
  group_by(reef, year, transect) %>%

  # Add `counts` and `mean_size` columns
  summarize(counts = sum(count),  ## REVISION: Changed n() to sum(count),
                                  # because we want sum of counts not # of rows
            mean_size_mm = mean(size_mm, na.rm = TRUE)) %>%

  # Add `mpa` variable
  mutate(mpa = case_when(
    reef %in% c("Naples", "Isla Vista") ~ "MPA",
    TRUE ~ "non_MPA"),

  # Add `treatment` variable based on `mpa`
  treat = case_when(
    mpa == "MPA" ~ 1,
    TRUE ~ 0))
```

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).

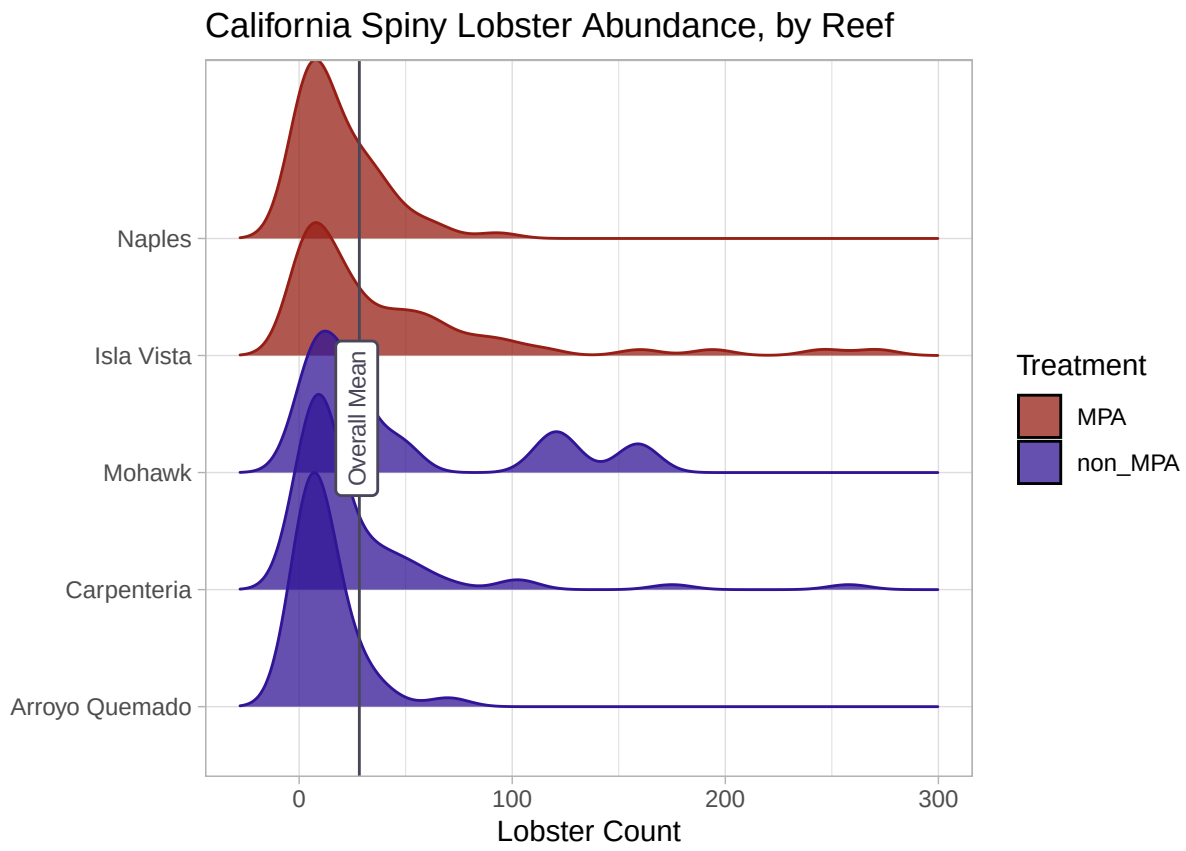
- Density plot
- Ridge plot
- Jitter plot
- Violin plot
- Histogram
- Beeswarm

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

```
# Save plot colors
mpa <- "#311596"
non_mpa <- "#961e15"
stat_label <- "#4a4659"
neutral_green <- "#46594a"
```

1) grouped by reef site

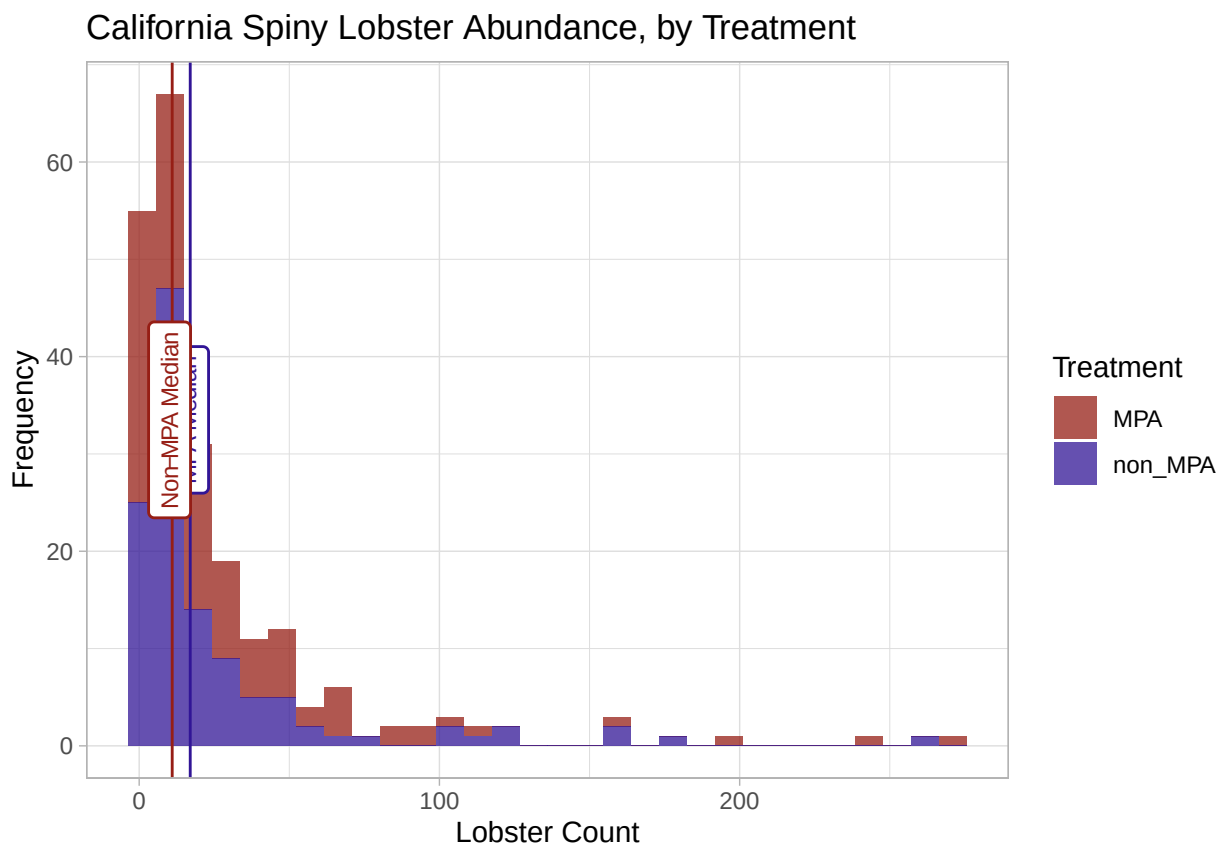
```
# Plot 1: Ridge Plot
ridge_reef <- ggplot(data = spiny_counts, aes(x = counts, y = reef,
                                fill = mpa, color = mpa)) +

  # Add ridge plot layer
 geom_density_ridges(scale = 2, alpha = 0.75) +

  # Add line of mean value
  geom_labelvline(xintercept = mean(spiny_counts$counts),
                  color = stat_label, label = "Overall Mean",
                  size = 3) +

  # Plot labels
  labs(title = "California Spiny Lobster Abundance, by Reef",
       x = "Lobster Count",
       y = " ",
       fill = "Treatment") +

  # Custom colors and remove redundant legends
  scale_color_manual(values = c(non_mpa, mpa), guide = FALSE) +
  scale_fill_manual(values = c(non_mpa, mpa)) +

theme_light()

ridge_reef
```



2) grouped by MPA status

```
# Median by treatment

median_mpa <- median(spiny_counts[spiny_counts$mpa == "MPA",]$counts)
median_non_mpa <- median(spiny_counts[spiny_counts$mpa == "non_MPA",]$counts)

# Plot 2: Histogram

hist_mpa <- ggplot(data = spiny_counts, aes(x = counts, fill = mpa)) +

  # Create histogram
  geom_histogram(alpha = 0.75) +

 # Add median value line for both treatments
  geom_labelvline(xintercept = median_mpa,
                  color = mpa, label = "MPA Median",
                  size = 3) +

  geom_labelvline(xintercept = median_non_mpa,
                  color = non_mpa, label = "Non-MPA Median",
                  size = 3) +

  # Plot labels
  labs(title = "California Spiny Lobster Abundance, by Treatment",
       x = "Lobster Count",
       y = "Frequency",
```

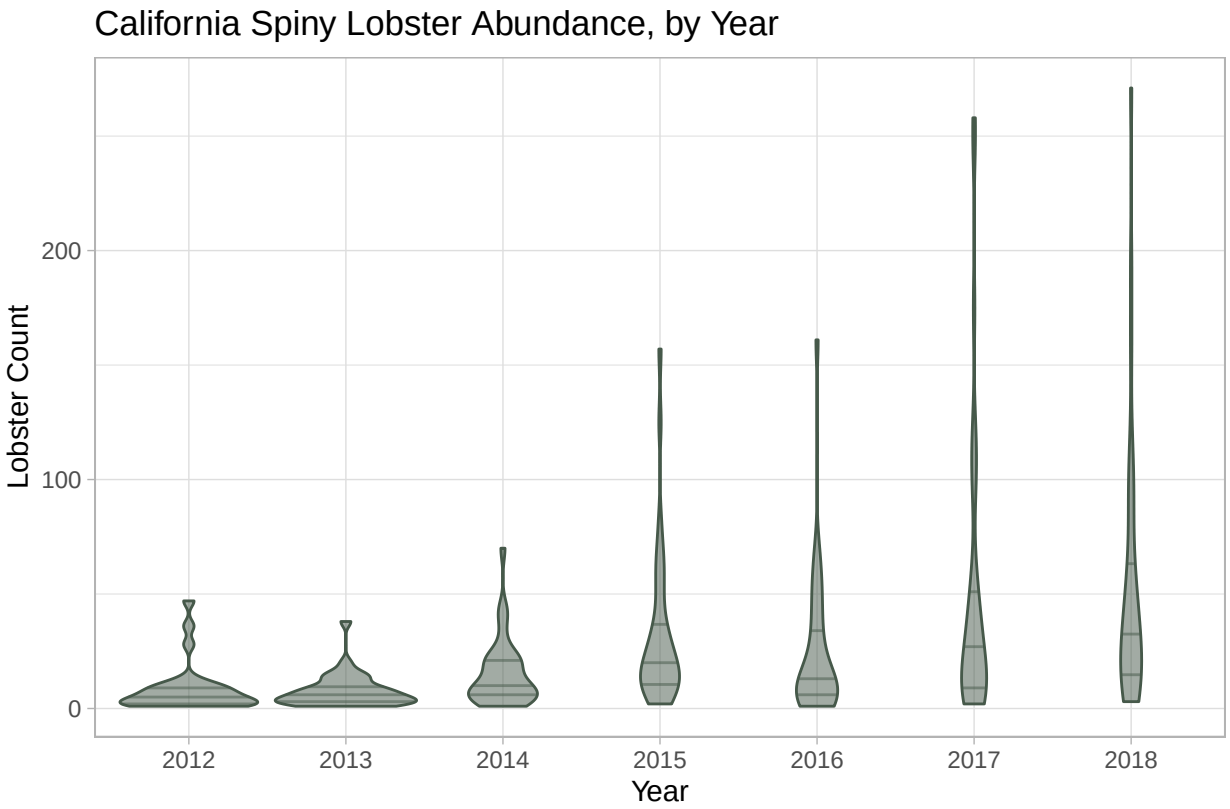
```
       fill = "Treatment") +

  # Custom colors and remove redundant legends
  scale_color_manual(values = c(non_mpa, mpa), guide = FALSE) +
  scale_fill_manual(values = c(non_mpa, mpa)) +

theme_light()

hist_mpa
```



3) grouped by year

```
# Plot 3: Violin Plot
violin_year <- ggplot(data = spiny_counts, aes(x = as.factor(year), y = counts)) +

  # Add violin plot layer
  geom_violin(draw_quantiles = c(0.25,0.5, 0.75), # Plot IQR
              color = neutral_green,
              fill = neutral_green,
              alpha = 0.5) +

  # Add title and labels
  labs(title = "California Spiny Lobster Abundance, by Year",
       x = "Year",
       y = "Lobster Count",
       caption = "The 25th, 50th, and 75th quantiles, by year, are represented within each violin plot.
```

```
  # Attempt to position caption on the left
  theme(plot.caption = element_text(hjust = 0)) +

  theme_light()

violin_year
```



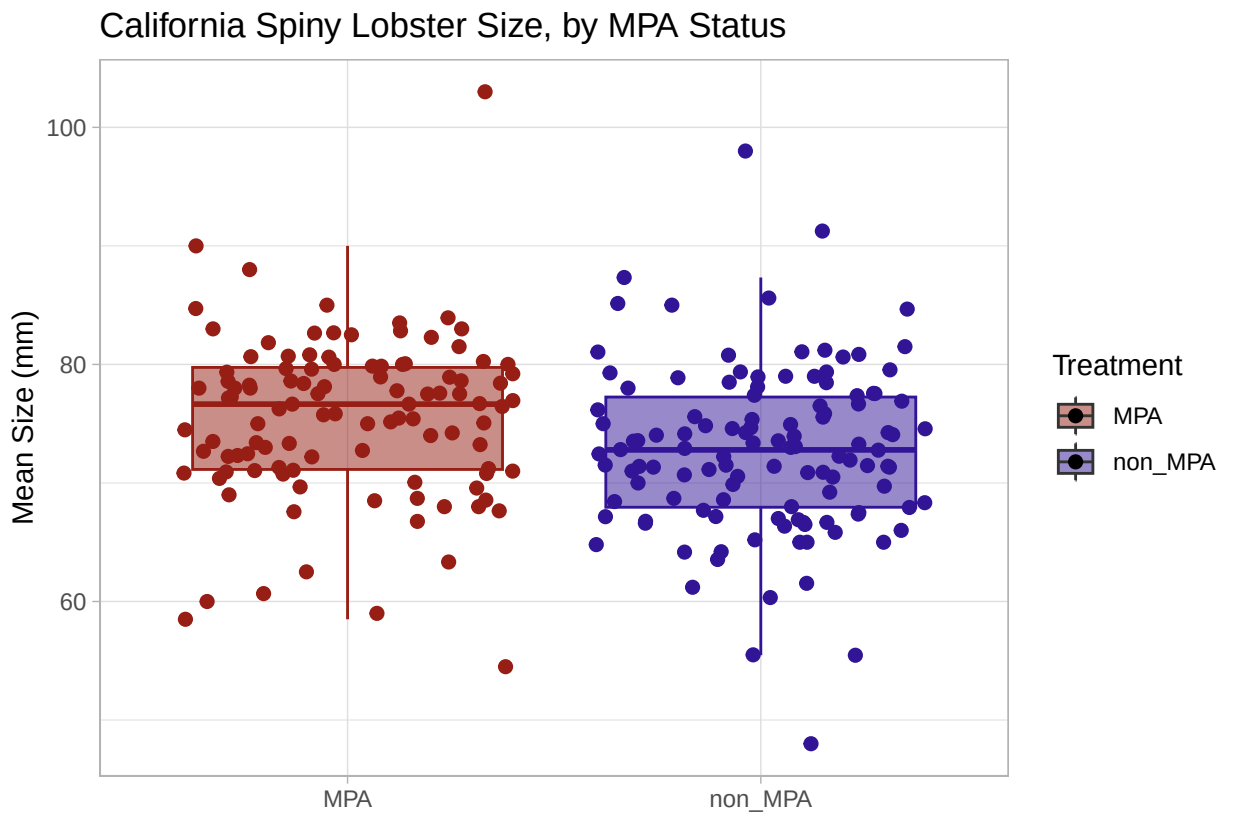
The 25th, 50th, and 75th quantiles, by year, are represented within each violin plot.

Create a plot of lobster **size** :

4) You choose the grouping variable(s)!

```
# Plot 4: Jitter Plot, by Treatment
jitter_size <- ggplot(data = spiny_counts, aes(x = mpa, y = mean_size_mm,
                                color = mpa, fill = mpa)) +

  # Add boxplot layer first so it plots below
  geom_boxplot(alpha = 0.5,
               outlier.shape = NA) + # Remove outliers so they are not double plotted

  # Add jitter layer on top
  geom_jitter(size = 2) +

  # Custom colors and remove redundant legends
  scale_color_manual(values = c(non_mpa, mpa), guide = FALSE) +
  scale_fill_manual(values = c(non_mpa, mpa)) +

  # Add title and labels
```

```
  labs(title = "California Spiny Lobster Size, by MPA Status",
       x = " ",
       y = "Mean Size (mm)",
       fill = "Treatment") +

  theme_light()

# Boxplots are plotted alongside the jitter plot to offer information on the respective medians and IQR

jitter_size
```



c. Compare means of the outcome by treatment group. Using the `tbl_summary()` function from the package `gt_summary`

```
# USE: gt_summary::tbl_summary()

# Create summary table
lobster_summary <- spiny_counts %>%
  gtsummary::tbl_summary(

    # Group by mpa
    by = mpa,

    # Variables to include

    include = c(counts, mean_size_mm),
```

Spiny Lobster Counts and Sizes
By Treatment and Control Groups

Variable	MPA[1]	Non-MPA[1]
counts	31	26
mean_size_mm	76	73

[1] Mean

Note: Data from Reed 2019 (SBC LTER, Environmental Data Initiative).

```
    # Display the mean and standard deviation
    statistic = list(all_continuous() ~ "{mean}"))


# Apply gt customizations
lobster_gt <- lobster_summary %>%

  as_gt() %>%

  # Add title and subtitle
  tab_header(
   title = "Spiny Lobster Counts and Sizes",
   subtitle = "By Treatment and Control Groups"
 ) %>%

  # Column labels
cols_label(
  label = "Variable",
  stat_1 = "MPA",
  stat_2 = "Non-MPA"# Could change control and treatment labels using "stat_1" and "stat_2"
) %>%

  # Add a source note
tab_source_note(
  source_note = "Note: Data from Reed 2019 (SBC LTER, Environmental Data Initiative)."
)

lobster_gt
```

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` 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-square)

# OLS estimator of lobster counts regressed on treatment
m1_ols <- lm(data = spiny_counts, formula = counts ~ treat)
```

```
# Print OLS output
summ(m1_ols, model.fit = FALSE)
```

Observations	225
Dependent variable	counts
Type	OLS linear regression

	Est.	S.E.	t val.	p
(Intercept)	25.61	3.92	6.53	0.00
treat	5.59	5.69	0.98	0.33

Standard errors: OLS

REVISION:

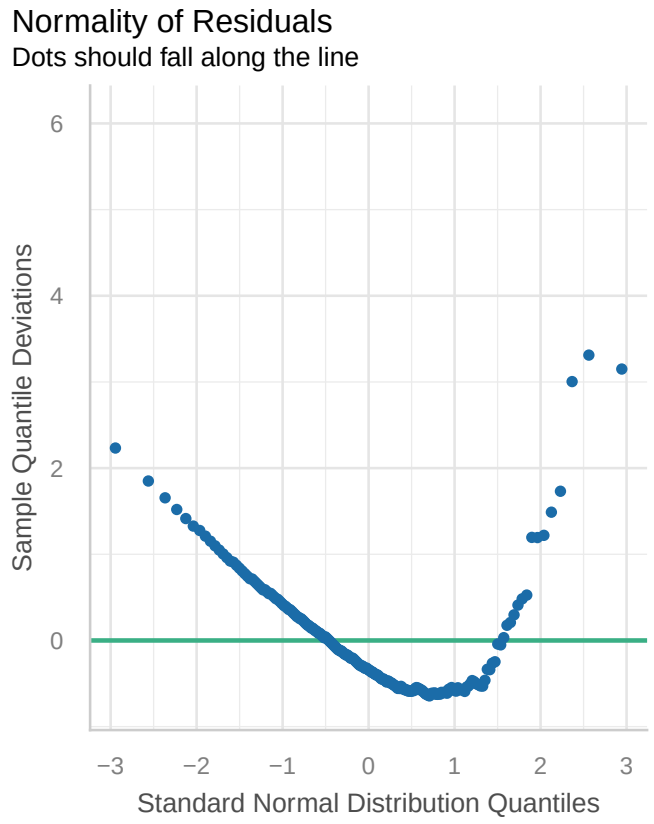
Intercept: Non-MPA sites in this study have an average lobster count of ~25.6.

treat: With the addition of the MPA treatment, lobster count increases by ~5.6, on average.

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

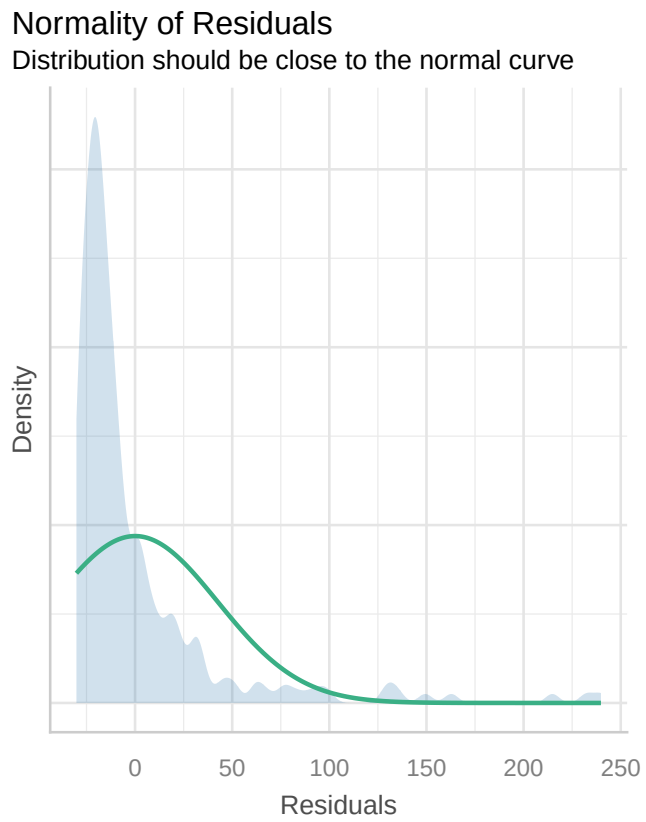
d. Explain the results of the 4 diagnostic plots. Why are we getting this result?

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



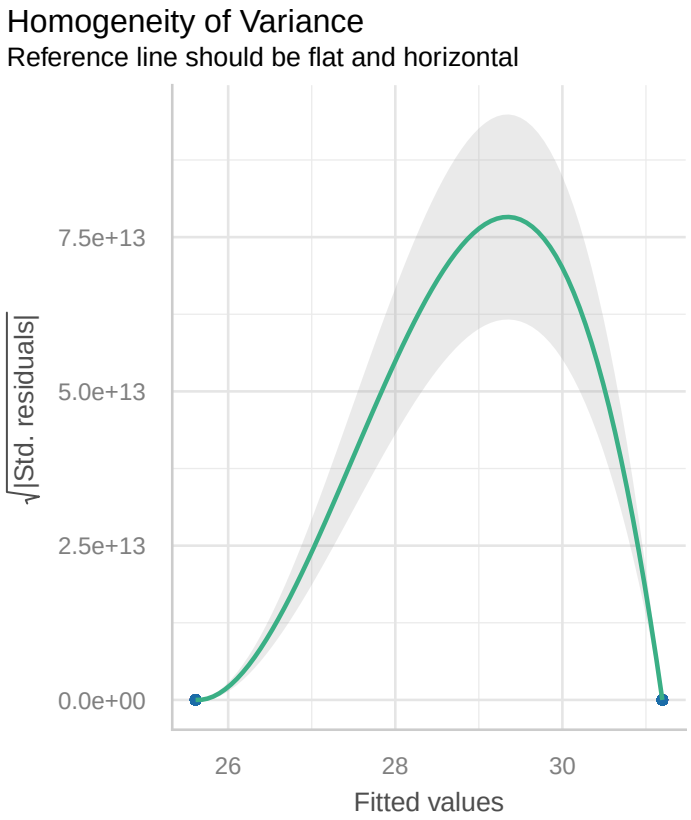
The data points should follow the horizontal line. Because they don't, we know that this data does not follow a normal distribution (which as an assumption of OLS regression).

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



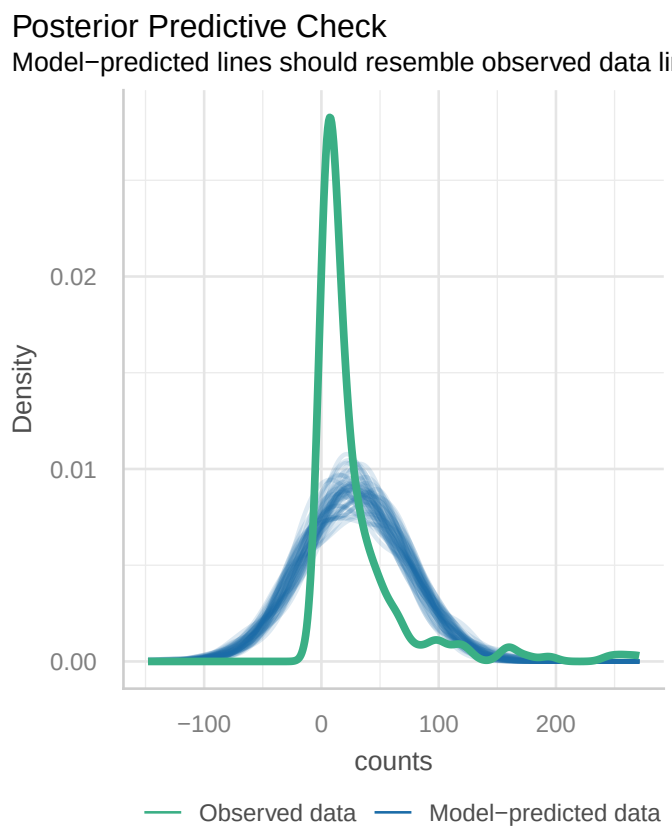
Our data does not follow the distribution of the normal curve: it both exceeds and falls short of it at different points of the residuals axis. This suggests that the distribution of our data is non-normal.

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



The reference line in this plot is not horizontal, showing our data's spread around the mean is not equal across groups.

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



In this plot, model-predicted data does not match the distribution of our data. If our data was normally-distributed, it would mimic the shape of the simulated curves.

Ultimately, what all four of these plots confirm is that our data does not follow a normal distribution, and therefore an OLS regression is not an appropriate model (as that is one of its core assumptions).

Step 5: Fitting GLMs **a.** Estimate a Poisson regression model using the `glm()` function

```
#HINT1: Incidence Ratio Rate (IRR): Exponentiation of beta returns coefficient which is interpreted as

#HINT2: For the second glm() argument `family` use the following specification option `family = poisson

# Poisson regression model
m2_pois <- glm(data = spiny_counts,
               formula = counts ~ treat,
               family = poisson(link = "log"))

# Print output
summ(m2_pois, model.fit = FALSE)
```

Observations	225
Dependent variable	counts
Type	Generalized linear model
Family	poisson
Link	log

	Est.	S.E.	z val.	p
(Intercept)	3.24	0.02	178.28	0.00
treat	0.20	0.03	7.86	0.00

Standard errors: MLE

```
print(paste("The intercept as a real, non-log value is", round(exp(coef(m2_pois)[1]), 2)))
```

```
## [1] "The intercept as a real, non-log value is 25.61"
```

```
print(paste("The treat coefficient as percent change is", (round(exp(coef(m2_pois)[2]), 2) - 1) * 100,
```

```
## [1] "The treat coefficient as percent change is 22 %."
```

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.

REVISION:

Intercept: Non-MPA sites have an average lobster count of ~25.6. treat: The creation of an MPA (treatment) increases the lobster count by an average of 22%.

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

A Poisson model assumes that mean = variance. Overdispersion in this case would occur if the variance is greater than the mean. Dispersion in general refers to the distribution and variance of the data.

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

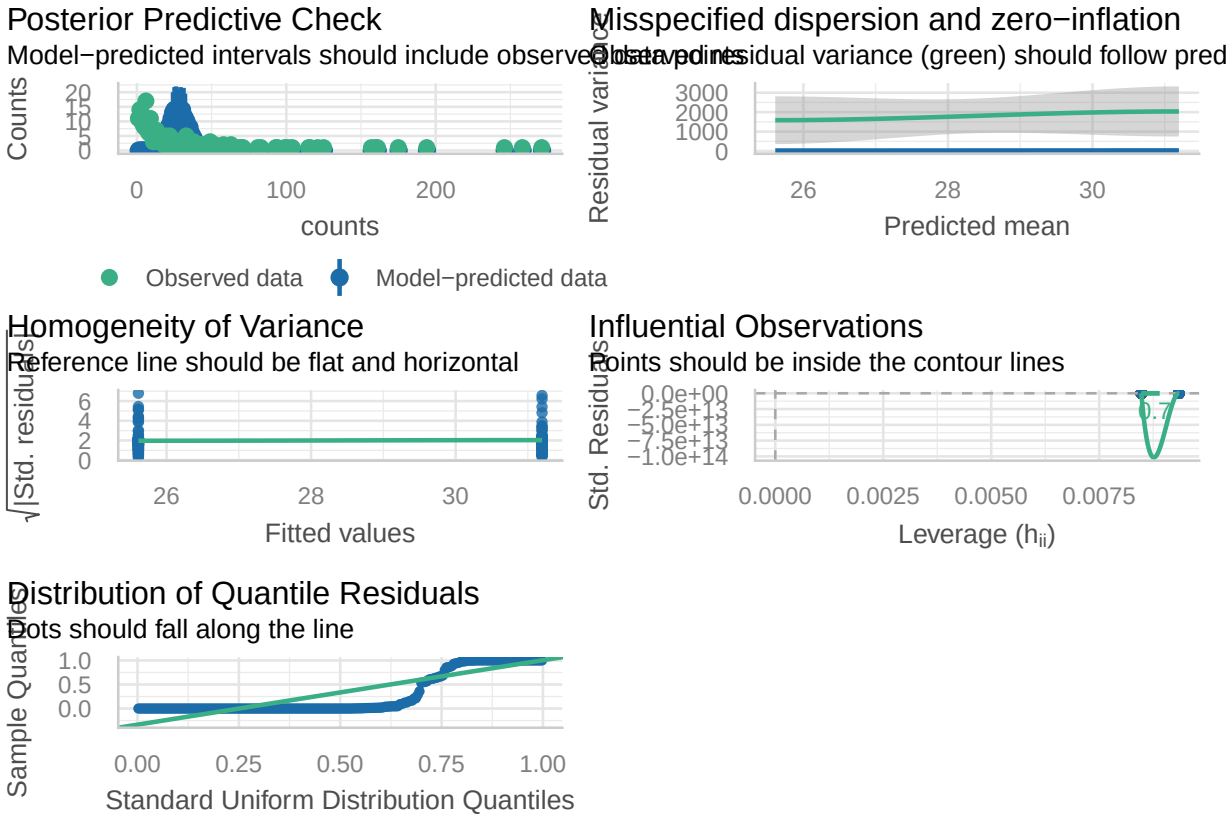
Between the OLS and Poisson models, the p-value of the treat coefficient went from 0.03 to 0.00. Although both of these values are indicative of statistical significance, the Poisson model is significant at a 99% confidence level, whereas OLS regression is only at a 95% confidence level.

e. Check the model assumptions. Explain results.

These data partially fit Poisson assumptions because they contain count information. However, characteristics of the data, like overdispersion, can sometimes lead to over-significant results, which may be what is taking place here. Since we don't know if our data's mean is equal to its variance, we cannot be certain that these results are accurate.

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

```
check_model(m2_pois)
```




```
check_overdispersion(m2_pois)
```

```
## # Overdispersion test
##
##        dispersion ratio =    64.090
##   Pearson's Chi-Squared = 14292.093
##                 p-value =   < 0.001
```

```
check_zeroinflation(m2_pois)
```

```
## Model has no observed zeros in the response variable.
```

```
## NULL
```

Based on our checks, we can see that simulated data based on model predictions does not match the same distributions as our data. Additionally, the dispersion ratio is significantly greater than 1 (a ratio of 1 being what we would expect if variance and mean were equal). These results suggest that our data does not follow a Poisson distribution, and that the model may not be appropriate.

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.

The negative binomial model has a dispersion parameter, something that the Poisson model doesn't have. This allows it to be more flexible and handle overdisperion, which is something that we observed in our data.

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

```
library(MASS) ## NOTE: The `select()` function is masked. Use: `dplyr::select()` ##
```

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

# Apply negative binomial model
m3_nb <- glm.nb(data = spiny_counts, formula = counts ~ treat)

# Print output
summ(m3_nb)
```

Observations	225
Dependent variable	counts
Type	Generalized linear model
Family	Negative Binomial(0.7805)
Link	log

$\chi^2(223)$	1.66
p	0.20
Pseudo-R^2 (Cragg-Uhler)	0.01
Pseudo-R^2 (McFadden)	0.00
AIC	1956.68
BIC	1966.93

	Est.	S.E.	z val.	p
(Intercept)	3.24	0.11	30.66	0.00
treat	0.20	0.15	1.29	0.20

Standard errors: MLE

```
print(paste("The intercept as a real, non-log value is", round(exp(coef(m3_nb)[1]), 2)))
```

```
## [1] "The intercept as a real, non-log value is 25.61"
```

```
print(paste("The treat coefficient as percent change is", (round(exp(coef(m3_nb)[2]), 2) - 1) * 100, "%
```

```
## [1] "The treat coefficient as percent change is 22 %."
```

REVISION:

Intercept: At Non-MPA sites, the lobster count on average is ~25.6. treat: The creation of an MPA increases lobster count by ~22%, on average.

Comparison: These results are incredibly similar to the Poisson model, with the Poisson model having a slightly smaller p-value. The intercept terms also match that of the OLS regression, but it is more difficult to compare the coefficient estimate across models as Poisson and Negative-Binomial produce results in units of percent change, whereas our OLS intercept directly represents lobster count.

```
check_overdispersion(m3_nb)
```

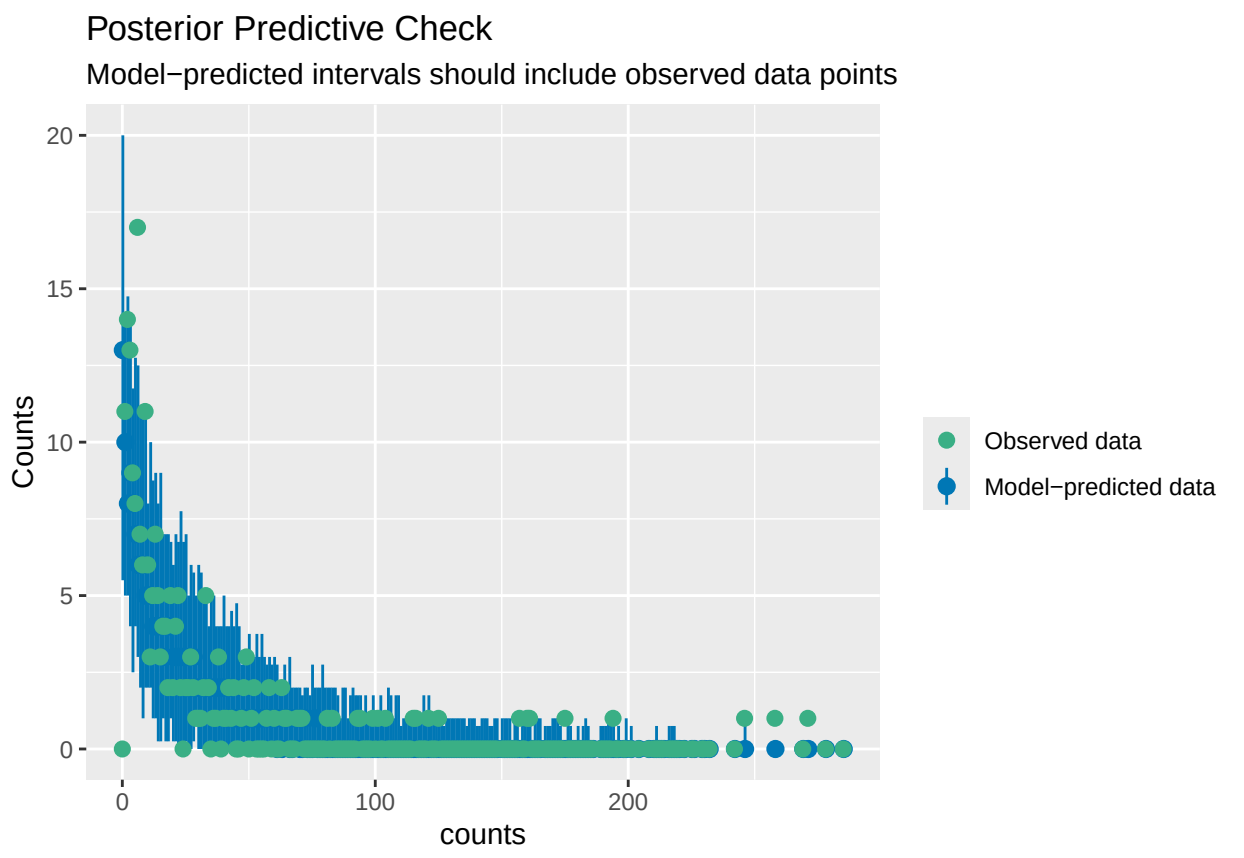
```
## # Overdispersion test
##
##  dispersion ratio =   1.681
##            p-value = < 0.001
```

```
check_zeroinflation(m3_nb)
```

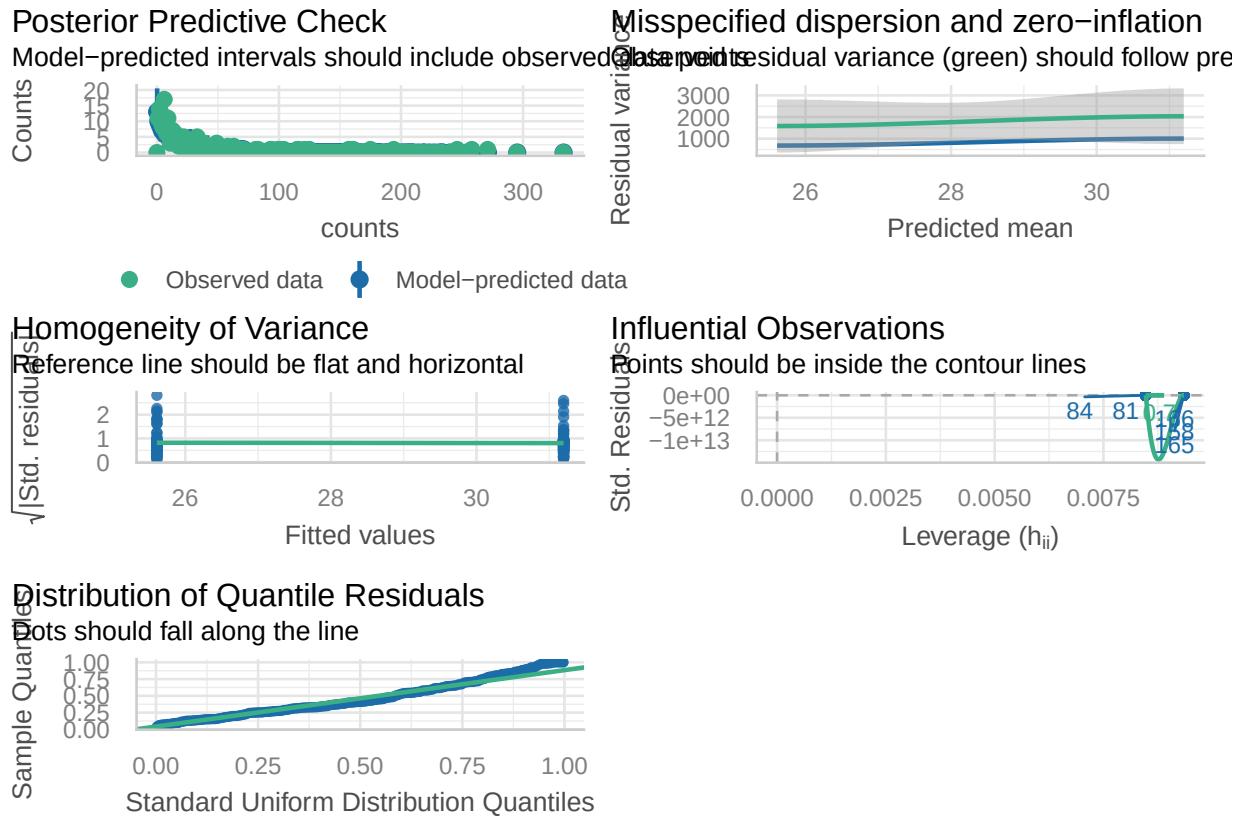
```
## Model has no observed zeros in the response variable.
```

```
## NULL
```

```
check_predictions(m3_nb)
```




```
check_model(m3_nb)
```



The Negative Binomial model appears to match our data the best compared to OLS and Poisson, and it even handles the overdispersion in our data (the dispersion ratio is now very close to 1).

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.

```
# Compare model outputs
export_summs(m1_ols, m2_pois, m3_nb,
             model.names = c("OLS","Poisson", "NB"),
             statistics = "none")
```

	OLS	Poisson	NB
(Intercept)	25.61 ***	3.24 ***	3.24 ***
	(3.92)	(0.02)	(0.11)
treat	5.59	0.20 ***	0.20
	(5.69)	(0.03)	(0.15)

*** p < 0.001; ** p < 0.01; * p < 0.05.

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

Like was mentioned before, intercept estimates are very similar across all three of the models. The treatment term in the OLS model is a lot larger than that of the Poisson and NB models (5.67 compared to 1.38). The mismatch of the data's distribution to the assumptions of the model may have produced these skewed results.

Despite differences in estimates, the treatment effect is robust across the model specifications. We know this because in all models the effect produces statistically significant coefficient estimates, suggesting that MPAs have an observable effect on lobster count across these study sites.

REVISION: In terms of performance, models differ in their statistical significance and p-values. Specifically, the Poisson model has the most significant coefficient (significant with level alpha = 0.001). This is to be expected, as the Poisson model is known to overestimate significance since it does not consider over-dispersion in its model and results. When it incorrectly assumes that variance and mean are equal, it underestimates standard error.

The OLS and NB models are both statistically significant at a 95% confidence level. This is important to note because it highlights that models can produce significant results even if the underlying data distribution does not match model assumptions.

Step 7: Building intuition - fixed effects **a.** Create new `df` with the `year` variable converted to a factor

```
# Year as a factor
ff_counts <- spiny_counts %>%
    mutate(year=as_factor(year))
```

b. Run the following negative binomial model using `glm.nb()`

- Add fixed effects for `year` (i.e., dummy coefficients)
- Include an interaction term between variables `treat` & `year` (`treat*year`)

```
# Run fixed effects model
m5_fixedeffs <- glm.nb(
    counts ~
        treat +
        year +
        treat*year,
    data = ff_counts)

# Print outputs
summ(m5_fixedeffs, model.fit = FALSE)
```

Observations	225
Dependent variable	counts
Type	Generalized linear model
Family	Negative Binomial(1.1567)
Link	log

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)

REVISION: Generally, treatment has an increasing effect on lobster counts across all years. It is difficult to discern an exact pattern that emerges in yearly increments, though the effect does increase from 1.18 to 2.09 across the 5 year period. This suggests that MPAs get more and more effective with each year.

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

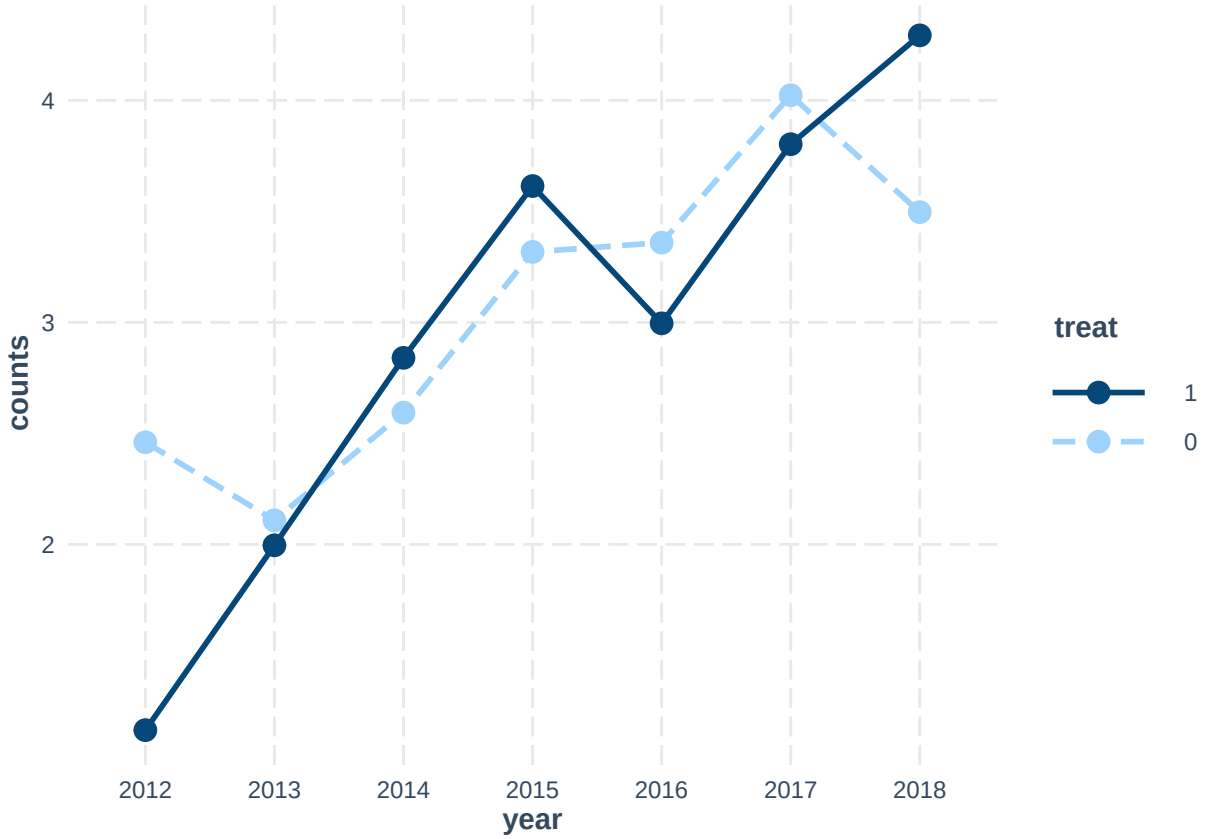
	Est.	S.E.	z val.	p
(Intercept)	2.46	0.24	10.41	0.00
treat	-1.30	0.42	-3.11	0.00
year2013	-0.35	0.34	-1.04	0.30
year2014	0.13	0.34	0.39	0.69
year2015	0.86	0.33	2.60	0.01
year2016	0.90	0.33	2.73	0.01
year2017	1.56	0.33	4.76	0.00
year2018	1.04	0.33	3.15	0.00
treat:year2013	1.18	0.55	2.15	0.03
treat:year2014	1.54	0.54	2.88	0.00
treat:year2015	1.59	0.53	3.02	0.00
treat:year2016	0.93	0.53	1.75	0.08
treat:year2017	1.08	0.53	2.03	0.04
treat:year2018	2.09	0.53	3.97	0.00

Standard errors: MLE

This result does make sense, because you would still be adding on the year term and the interaction term coefficients, which in most cases would make the treatment effect ultimately positive. An exception to this is 2013, but since that was the first year after the MPAs were enacted, it is not completely shocking that the treatment effect is negative.

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

```
# Plot model predictions
interact_plot(m5_fixedeffs, pred = year, modx = treat,
              outcome.scale = "link") # NOTE: y-axis on log-scale
```




```
# HINT: Change `outcome.scale` to "response" to convert y-axis scale to counts
```

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

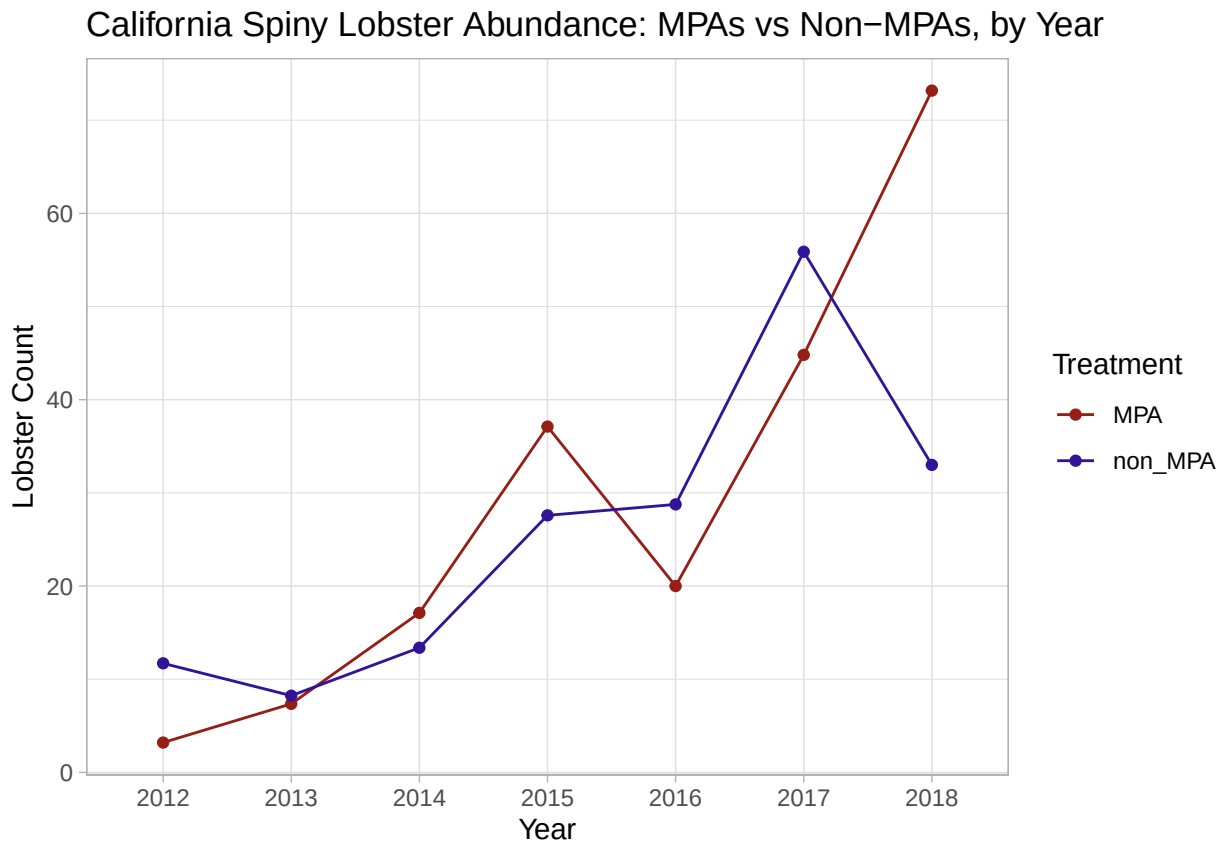
My responses to (c) and (b) still hold. As the plot shows, the areas that are now MPAs started out with lower lobster counts, explaining why in 2013, being in the treatment group correlated with a smaller lobster count than that of the control group. There is a positive trend in both the treatment and control groups, and in 2016 lobster counts at MPA sites actually dip below that of the non-MPA sites (but then regain superiority in the following year).

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

# Create df with mean counts by year and mpa
plot_counts <- ff_counts %>%
  group_by(year, mpa) %>%
  summarize(mean_count = mean(counts, na.rm = TRUE))

# Plot
plot_counts %>%
  ggplot(aes(x = year, y = mean_count, color = mpa, group = mpa)) + # Grouping my mpa ensures that geom
```

```
  # Plot line and point geometries to mimic look of previous plot
  geom_line() +
  geom_point() +

  # Add title and labels
  labs(title = "California Spiny Lobster Abundance: MPAs vs Non-MPAs, by Year",
       x = "Year",
       y = "Lobster Count",
       color = "Treatment") +

  # Custom colors
  scale_color_manual(values = c(non_mpa, mpa)) +

theme_light()
```



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)

Yes, spillover effects are very likely in this case. Because MPAs have no physical boundary between them, spiny lobsters can easily travel in between treatment and control sites, affecting results depending on their location during time of sampling.

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

The spillover effect is an issue because it reduces the difference between the control and treatment results. It prohibits the control from being a true control, since the treatment *does* have an effect on it. Without an accurate baseline, the impact of the treatment is muddled.

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

Because there is no exact control/baseline for measuring MPA effectiveness, the perceived impact of MPAs may be hindered. This means results are less impactful than what they might actually be.

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 (**REVISION**)

SUTVA includes two conditions:

1. No interference: One unit's treatment does not affect another unit's outcome.

This assumption suggests that MPA status (treatment) of a site would not effect have an effect on another sites lobster count. The sites are not directly next to each other, though they are quite close. Additionally, the two MPA/treatement sites – Isla Vista and Naples – are located right next to one another. It is possible that ecological health of one area *could* have an impact on the health (or spiny lobster count) of another, as the ocean's systems are very interconnected. Therefore, I do not believe that this assumption is reasonable to assume in this study.

2. No hidden variation: The treatment is implemented consistently for all units (i.e., all units receive the same treatment)

Both the Isla Vista and Naples sites were determined MPAs at the same time, but it is difficult to guarantee that each MPA gets respected/implemented in the same way. Since there are no physical barriers keeping people from fishing/damaging the areas, it is impossible to know whether to know if designation is working in the same exact way at both sites. However, out of the two SUTVA assumptions, I feel as though that this one is more reasonable to assume.

2) Exogeneity assumption (**REVISION**)

The exogeneity assumption suggests that our variable of interest (MPA status) is unrelated to unobserved factors (ecosystem charactertistics, for example) that could influence lobster abundance. In this case, the exogeneity assumption is not met. As alluded to in Question 1, selection bias that stems from the assignment of MPA could potentially link to the ecology and habitats that our sites contain. It is possible, for example, that Isla Vista and Naples are perfect ecosystems for the California spiny lobster (and our control sites are not), which is *why* they were made MPA sites. In this case, the suitability of habitat affects our variable of interest, meaning the exogeneity assumption is not met.
Random assignment is a method of meeting the exogeneity assumption, which was not utilized in this study.

EXTRA CREDIT

Use the recent lobster abundance data with observations collected up until 2024 (`extracredit_sblobstrs24.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)