STA610 Case Study 1 Team 4 Report

Cole Juracek (Programmer) Lauren Palazzo (Writer)
Lingyu Zhou (Checker + Coordinator) Fan Zhu (Presenter)

2021-03-21

Introduction

This study is an exploration of the relationship between the sale price per milligram of diverted pharmaceutical substances and other factors related to the sale, as reflected in self-reported data from the website StreetRx. StreetRx maintains a partnership with the Researched Abuse, Diversion, and Addiction-Related Surveillance System (RADARS), which aims for these data to inform public health measures such as health policy related to illicit drugs.

[insert something about our specific drug here]

Among the data available to us are location data pertaining to the city, state, and United States region of the sale. These variables present a natural grouping structure for the price data, which suggest that we employ a hierarchical model in our analysis. The benefits of a hierarchical model include the ability to "borrow information" from the data set as a whole to stabilize estimates of price for groups (e.g., states) with relatively small sample size, and to establish a estimated distribution of how price varies with respect to groups, which would in theory allow generalization to new groups not included in the original analysis.
Exploratory data analysis and data cleaning

Data cleaning

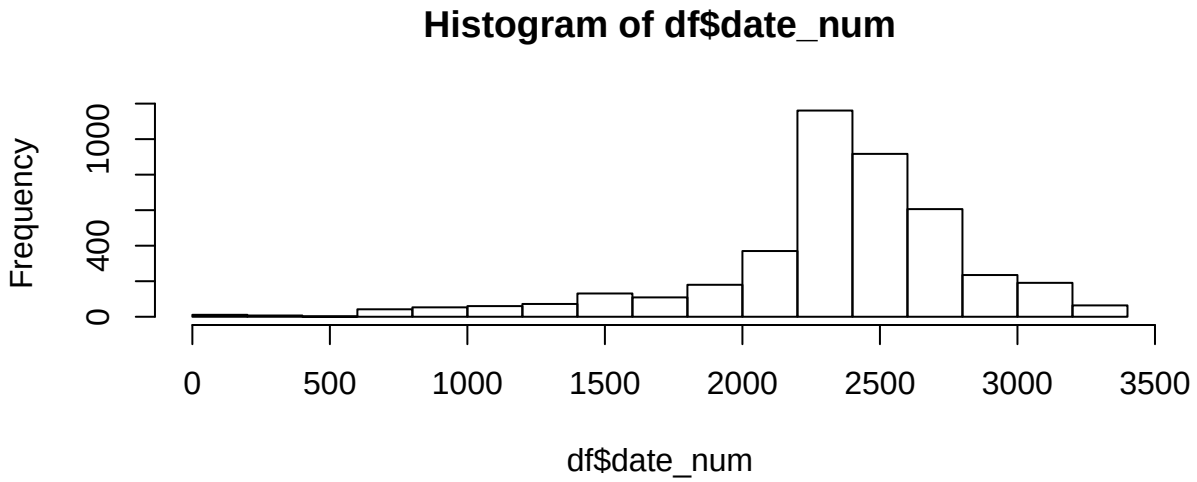
Group variables

We performed an initial examination of two grouping variables, `state` and `USA_region`. The state North Dakota had only one observation, and as such was not considered a well-defined "group" for which a variance could be estimated; thus this state's observation was removed. States labeled simply as "USA" were recoded to "Unknown". Region was also sometimes labeled as uknown or other; these observations coincided with the unknown states (see appendix Table X). (The variable `city` as a grouping variable is addressed later in the report.)

Date variable

We considered two different ways to engineer the date-of-sale feature. One intuitive and commonly-used approach would be to recode this as a continuous variables counting days from some start date, the intent of which would be to capture the influence of inflation for this price data that spans years. Another approach would be to consider whether there might be an effect of seasonality on drug prices.

For recoding date into a continuous variable, `date_num`, we filter out observations dated prior to the year 2010, as that is the year that the StreetRx web site was founded. For the seasonality variable, `date_quarter`, we extract the "quarter" of the year (numbered 1 through 4) as its own variable. Below we see the distributions of these variables:



Var1	Freq
1	1185
2	1236
3	937
4	853

Sources variable

As the variable `source` that describes the source of the price information is a user-input field, it is an inherently "messy" variable that calls for some recoding in order to make the levels of the variable more meaningful and interpretable, and to avoid overfitting. In the case of `source`, some user input is simply blank, while a small number of other entries indicate in some other way that the source is unknown or unclear. Since it is quite uncommon in the data for a source to be entered yet indicated as unknown, we decided to drop these unkown-source observations as being indicative of poor data quality. (If the user cannot identify their source of price information, why should we trust their price data?) We decided to distinguish between "Internet pharmacy" sources and other internet sources, since there may be a distinction between a source that is selling a drug versus perhaps just discussing a drug price in a forum, which may yield less legitimate information. Ultimately, we recoded the categorical variable "sources" into the following levels.

Level	Freq
Blank	1510
Internet	162
Internet Pharmacy	51
Personal	1611
Heard it	875

Bulk purchase variable

Our visualization of the bulk purchase variable indicates nothing unexpected that calls for data cleaning.

Purchase reason variable

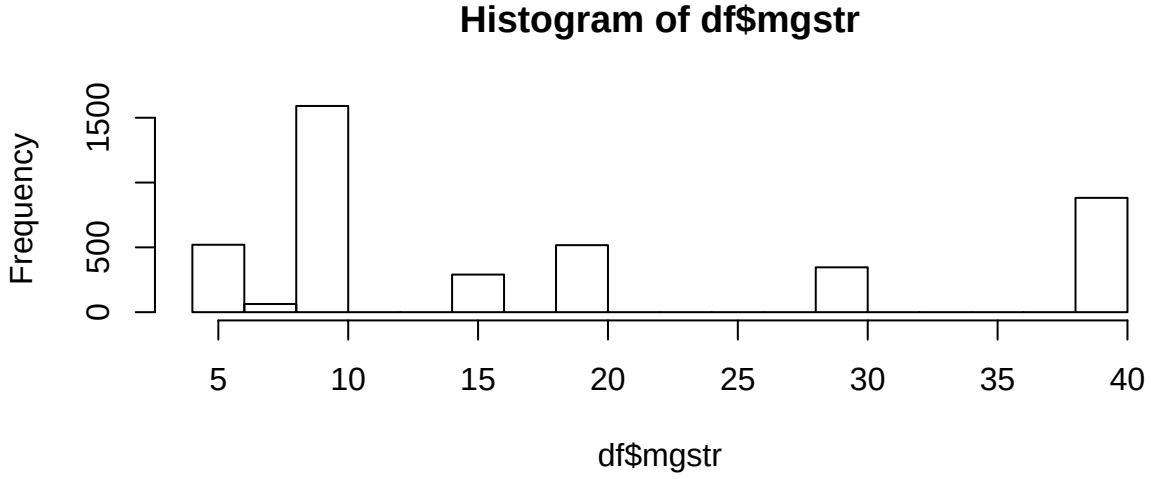
`Primary_Reason` for purchase is another user-input variable that calls for recoding. Similarly to the `source` variable, we retain the information about which entries were blank. We also try to distinguish

this from a different response level that is not blank, yet indicates the user did not provide a reason. We further still distinguished between these responses and a non-blank response of refusal to answer. We strove to avoid making presumptions of no difference between these similar-seeming response types, as they may indicate variations in data processing that hold some predictive power that we are not yet aware of. We ultimately recoded the categorical variable `Primary_Reason` into the following levels.

Level	Freq
Blank	2473
Left unanswered	780
Other or unknown	97
Self-treat	427
Enjoyment	104
Resell	18
Refuse to answer	310

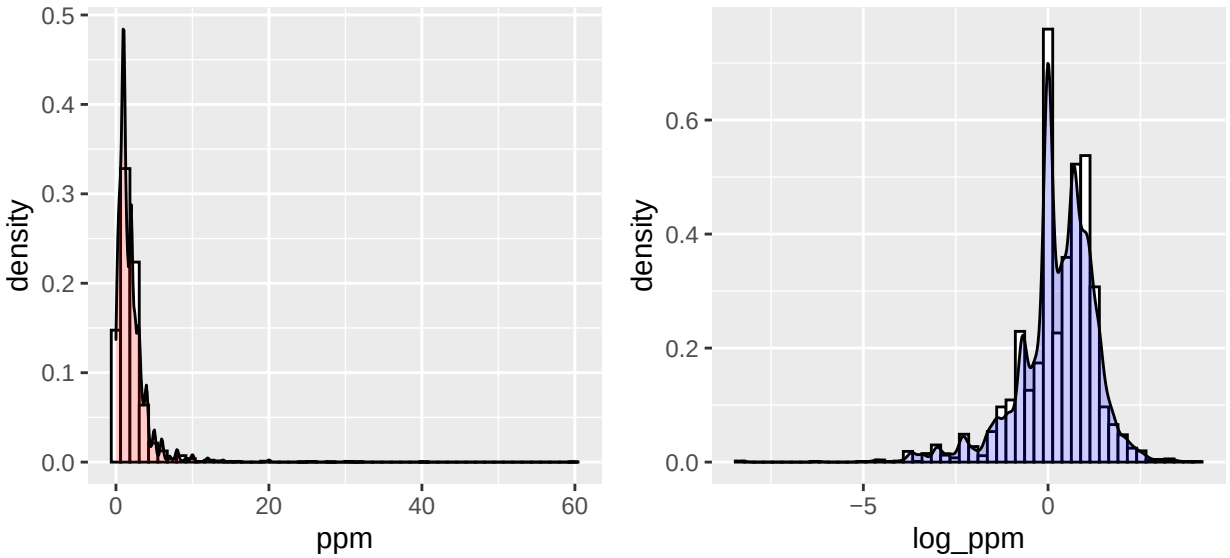
Drug strength variable

We renamed this variable `dose_per_mg` to make its meaning clearer to the reader. This is a continuous variable that nevertheless tends to take on only a few values, as depicted in the histogram. We considered a recoding of this variable into levels, but decided against it as the dose per milligram technically may take on any value and the generalizability of our model may be improved if we are able to consider dose levels that are in a realistic range yet not present in our current data.



Response variable: Price per milligram

We are considering fitting a hierarchical model with a response variable that has normally distributed errors, and so we would like our response variable, here `ppm`, or price per milligram, to be close to normally-distributed. The histogram of the `ppm` is extremely right skewed (not unexpected for data related to financial measures such as price, salary, earnings, etc.), while the distribution of the log(`ppm`) looks roughly normal. Thus, we will use `log_ppm` as the dependent variable in our model.



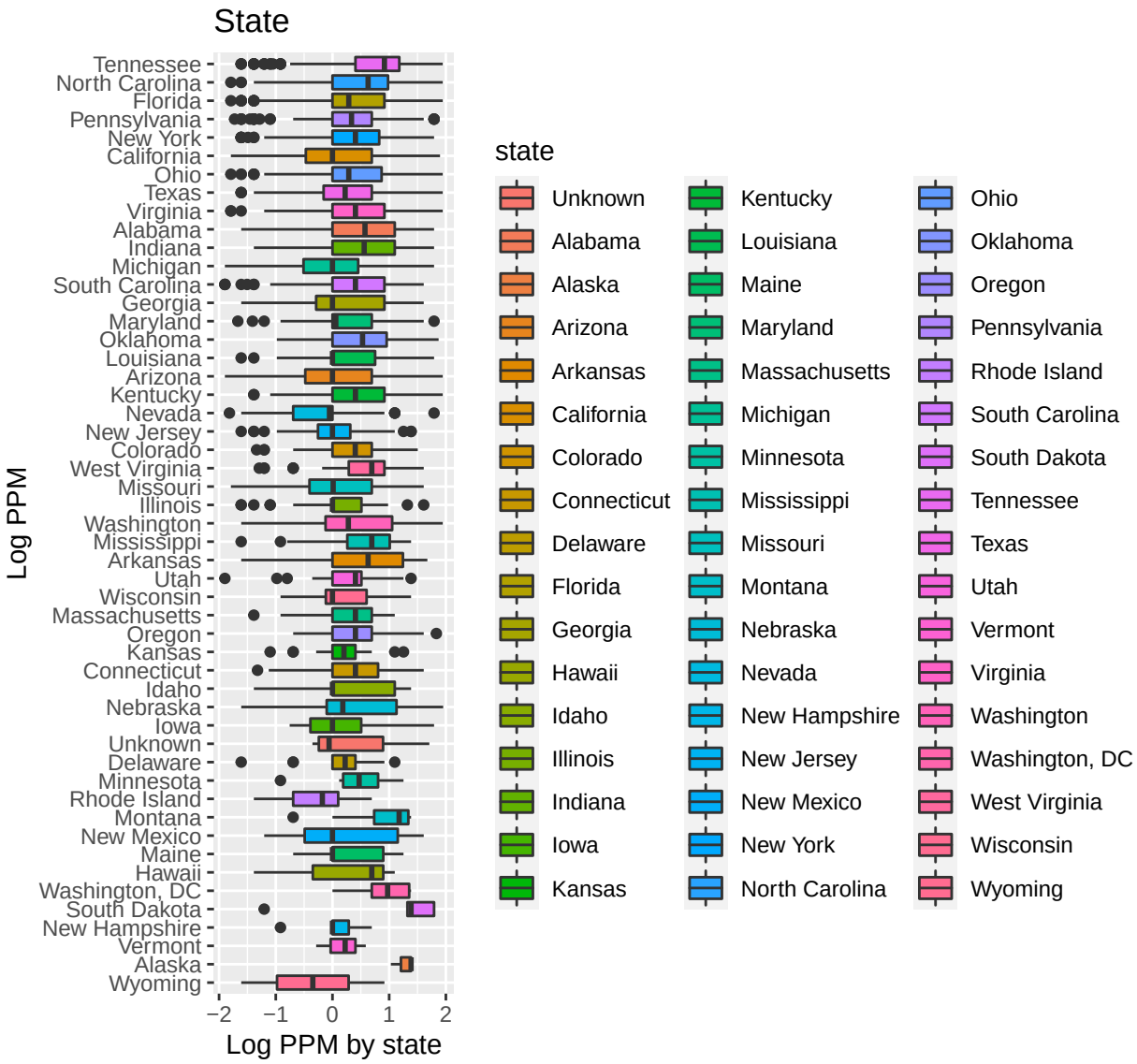
In order to cut off extreme price values that may skew the results for typical prices, we restrict the `log_ppm` distribution to more realistic and representative values:

Exploratory data analysis for random intercepts

We now take a closer look at our grouping variables with a focus on the question of whether random intercepts might be appropriate.

State

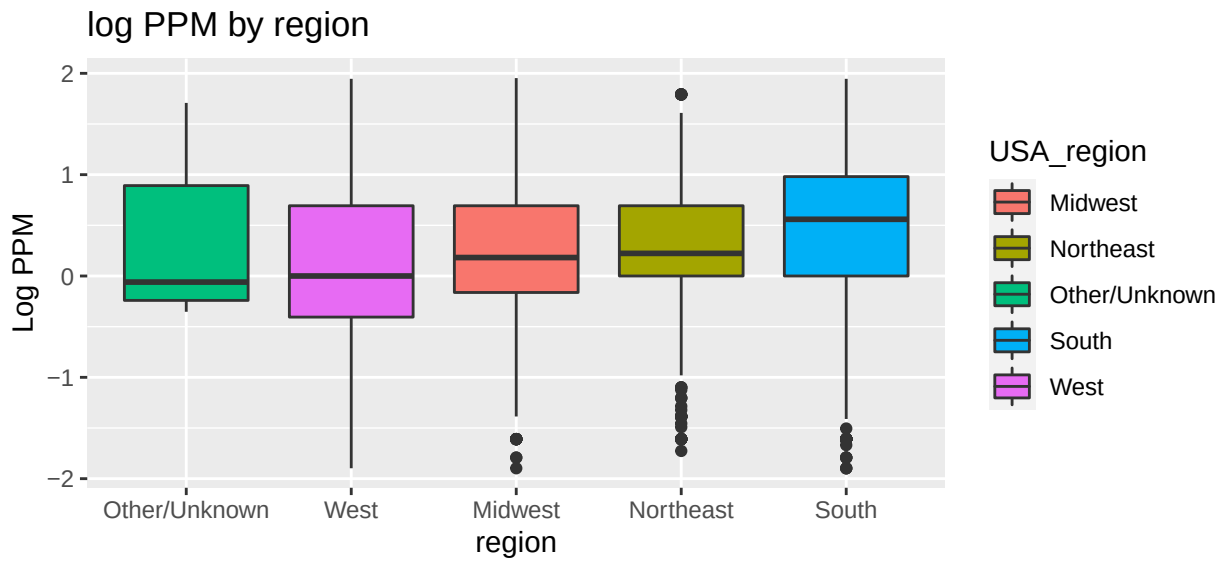
We show boxplots of `log_ppm` by `state`, arranged in the order of increasing sample size from the bottom to the top.



We see that some states with small sample sizes have distributions of `log_price` that seem unstable, with respect to how much they differ from other boxplots; this provides support for the idea that the stabilizing effect of a hierarchical model would benefit our analysis. The fact that some distributions differ from each other in a noteworthy way (e.g., boxplots hardly overlapping) suggests group-based differences in price that should be taken into account in modeling, as opposed to lumping all states' data together without distinction.

Region

Viewing a similar plot for the regions of the USA, we do see some variability in price among regions, but it is not as pronounced as for the states.



City

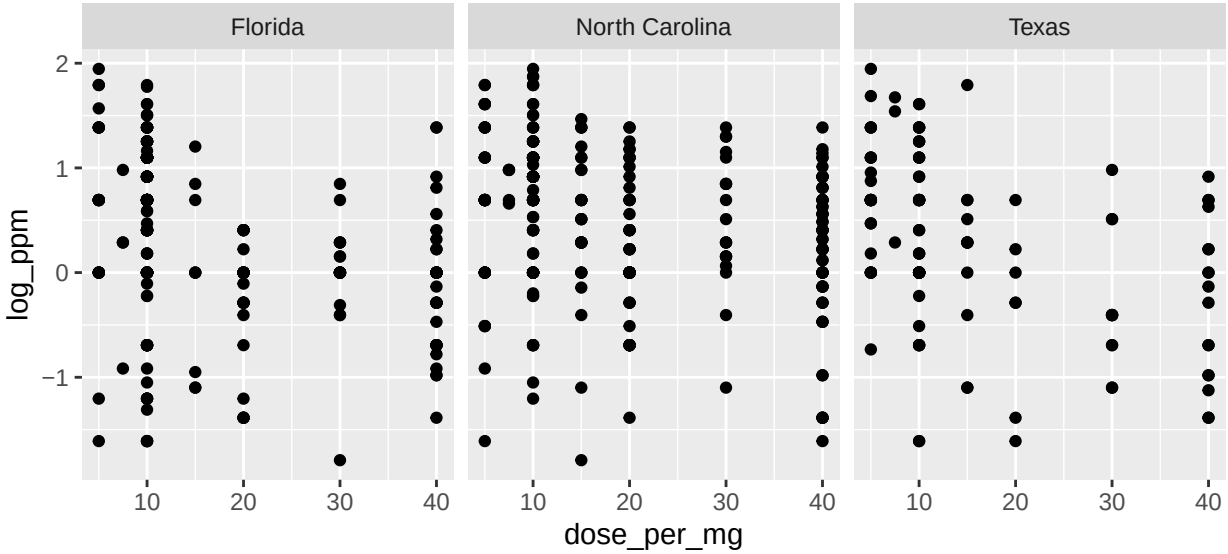
We find that there are many cities with a sample size of one, which makes `city` perhaps not the most natural choice for a grouping variable (please see appendix). We decided to focus on `state` as our primary candidate for a grouping variable moving forward.

Exploratory data analysis for random slopes

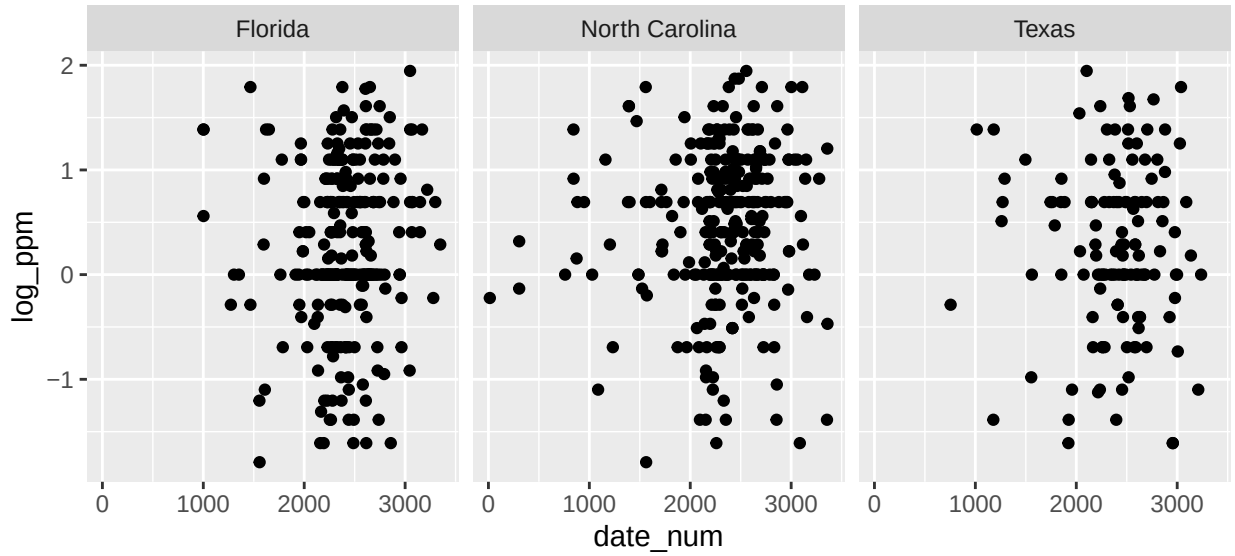
To consider whether we may want to include random slopes by state for our predictor variables, we examined plots of the predictor versus response among each state individually to determine whether it seemed that slopes fit to the group-level data would differ. (We present a sample of visualizations here rather than visualizations for every state, due to space constraints.)

For out continuous variables (continuous date and dosage strength), in by-state plots we often saw strikingly similar patterns in predictor vs. response, suggesting that random slopes may not be necessary for these variables.

dose_per_mg

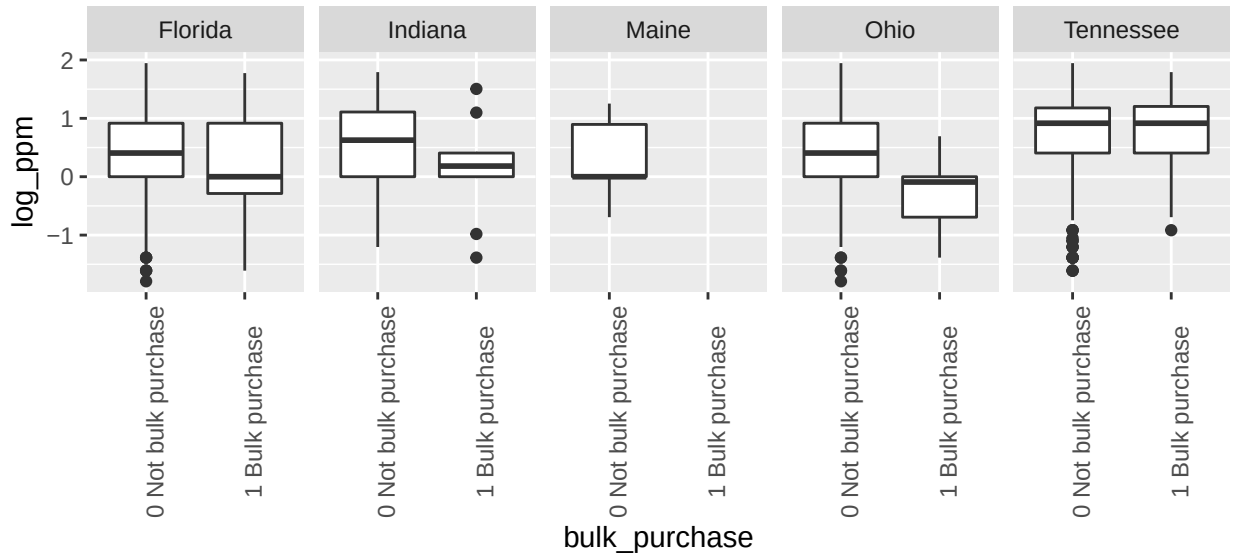


date_num

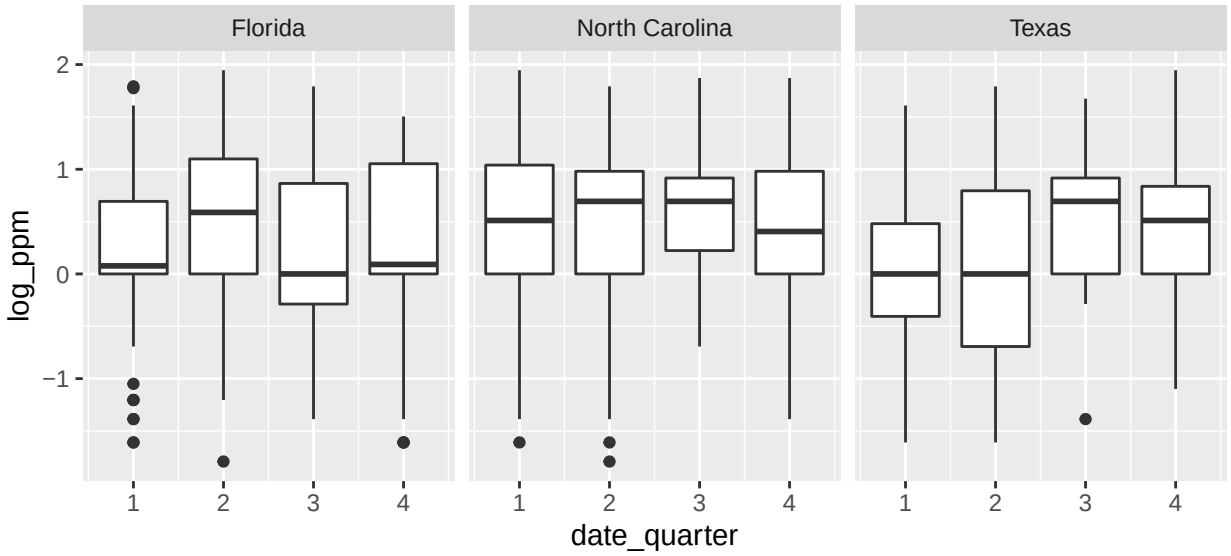


For categorical variables, we generally observed was there was indeed variation among the states, but it was not clear whether this variation had any well-defined structure. It was not clear whether our models would benefit from random slopes for these variables, so we decided to further investigate this in our modeling process.

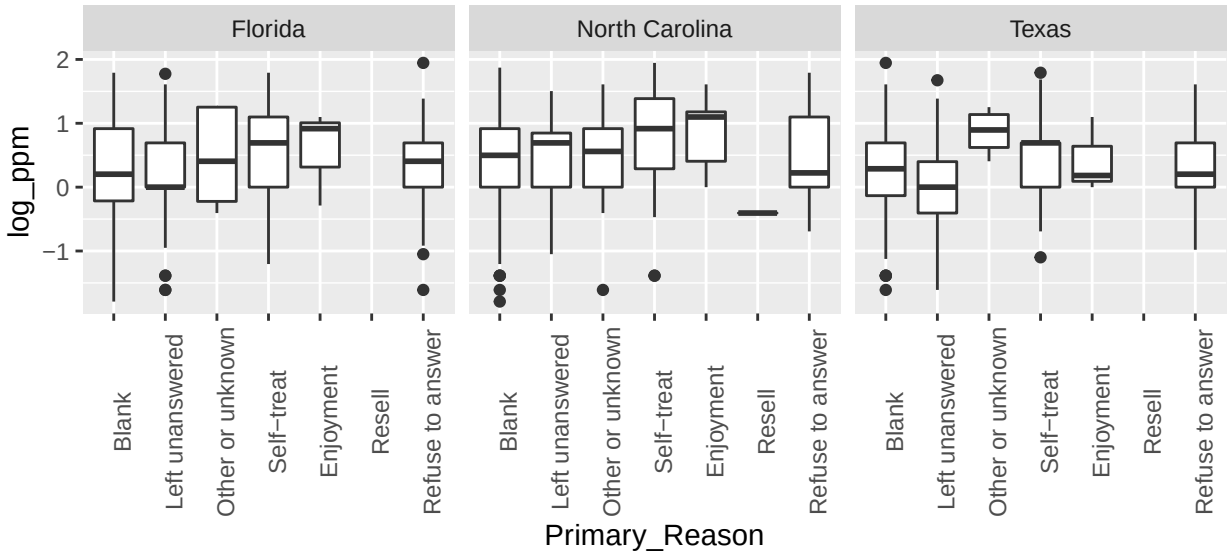
bulk_purchase

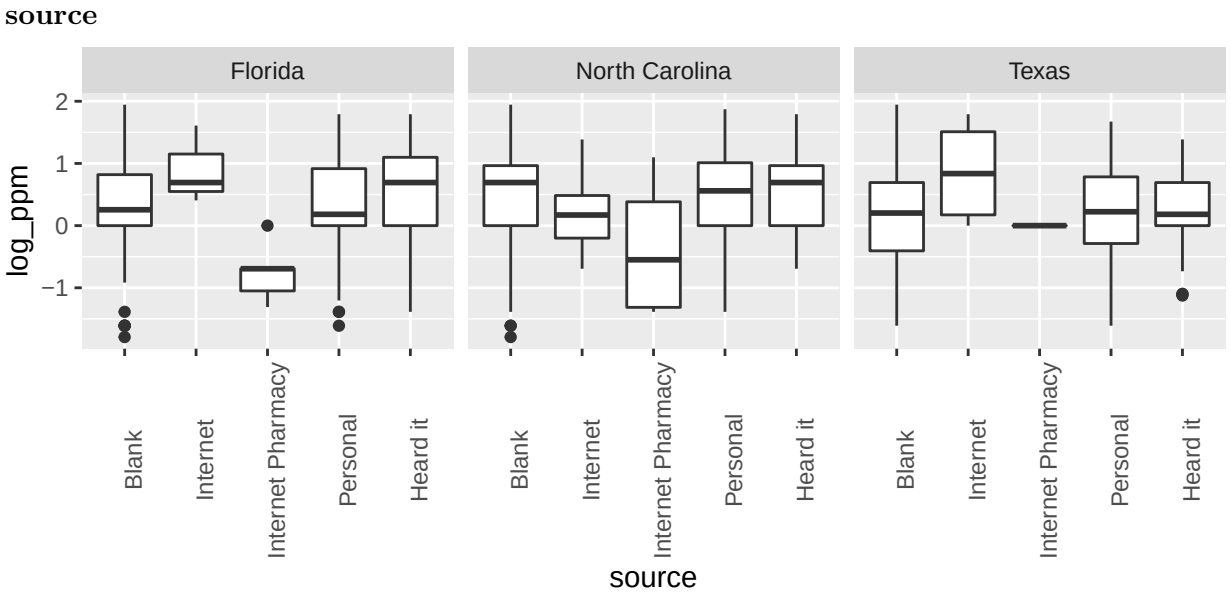


date_quarter



Primary_Reason





Model

Per the research question of interest, we are interested in exploring factors related to price per mg of oxymorphone. We are also specifically asked to control for location and how drug price changes with it. Cities seemed to be too noisy to use as a location predictor. On the flip side, region did not seem to be associated with a significant change in drug prices. Therefore, state provides the best granularity of accounting for location in this instance.

Linear regression models are usually a good standard approach to this, providing clear interpretations of how various predictors are associated with drug price. As seen in the EDA, however, many states have excessively small sample sizes. This produces corresponding $\hat{\beta}_{OLS}$ estimates with large variance. We would like to be able to *partially pool* information across states, letting us borrow information to help stabilize state estimates with small sample sizes.

This prompts a **mixed effects model**, where we instead model the state coefficients as having been drawn $N(0, \tau^2)$.

Model Selection

Based off of the results of the EDA, we want to initially fit a model with the following variables: source, dose_per_mg, bulk_purchase, Primary_Reason, date_num, and date_quarter. In terms of the research question, we would also like to group based on location. There are 3 options for this: region, state, or city.

In order to test whether keeping a variable in the model is "worth it", we may perform a likelihood ratio test for nested models. For adding k fixed effects, the LRT test statistic follows a $\chi^2(k)$ distribution, and for adding k random effects, the LRT test statistic follows a 50-50 mixture of $\chi^2(k)$ and $\chi^2(k+1)$ distribution. For non-nested models, we can use BIC as a rough indicator of which model is "better". BIC balances likelihood of model with how parsimonious it is.

First, let's decide what to group on for region:

Table 4: BIC for Models with Different Random Intercepts

Grouping	BIC
State	8643.014

Grouping	BIC
City	8769.939
Region	8788.861

State provides the best level of BIC, so we will choose this for our grouping variable.

Now let's investigate dropping various fixed effects and see results of our hypothesis tests:

The outline is as follows:

- Start with date_num, date_quarter, dose_per_mg, bulk_purchase, Primary_Reason, source
- Perform a LRT between competing models with and without the date variable. Observe a p-value of 0.494 and fail to reject the null (simpler model). Drop `date_num` from the subsequent model tests
- Perform a LRT between competing models with and without the quarter of the sale. Observe a p-value of 0.106 and fail to reject the null (simpler model). Drop `date_quarter` from the subsequent model tests
- Perform a LRT between competing models with and without dose per mg. Observe a p-value of ~0 and reject the null (simpler model). Decide on model keeping `dose_per_mg`
- Perform a LRT between competing models with and without the bulk purchase predictor. Observe a p-value of ~0 and reject the null (simpler model). Decide on model keeping `bulk_purchase`
- Perform a LRT between competing models with and without the primary reason predictor. Observe a p-value of 0.013 and reject the null (simpler model). Decide on keeping `primary_reason` in the model
- Perform a LRT between competing models with and without the source predictor. Observe a p-value of ~0 and reject the null (simpler model). Decide on keeping `source` in the model

Thus these results suggest we want to drop `date_num` and `date_quarter`, while keeping: dose_per_mg, bulk_purchase, primary_reason, source

Model Specification

Our final model is as specified:

$$log(y_{ij}) = \beta_{0,j} + \beta_1 x_{ij1} + \beta_2 x_{ij2} + \beta_3 x_{ij3} + \beta_4 x_{ij4} + \epsilon_{ij}$$

$$\beta_{0,j} = \beta_0 + b_{0,j}, b_{0,j} \sim N(0, \tau^2)$$

$$\epsilon_{ij} \perp b_{0,j} \overset{iid}{\sim} N(0, \sigma^2)$$

This linear model estimates the log price per miligram $log(y)$ of oxymorphone for an invidual i in state j (x_{ij}). The following predictors are defined as:

- x_1: Dosage strength in mg of the units purchased
- x_2: Bulk purchase, an indicator for whether 10+ units were purchased at once
- x_3: The primary reason for purchase. Factors include "blank", "left unanswered", "other or unknown", "self-treatment", "enjoyment", "resell", and "refuse to answer"
- x_4: Source of information (allows for reporting of transactions). Factors include "unkown", "internet", "internet pharmacy", "heard it", and "personal"

Model Diagnostics

Model diagnostics discussions here.

Results

Fixed Effects

	Estimate	Std.Error	t-value	P-value
(Intercept)	0.5204313	0.0385012	13.5172754	0.0000000
dose_per_mg	-0.0152224	0.0009142	-16.6507059	0.0000000
bulk_purchase1 Bulk purchase	-0.1716889	0.0307536	-5.5827296	0.0000000
Primary_ReasonLeft unanswered	0.0166237	0.0302059	0.5503465	0.5820818
Primary_ReasonOther or unknown	0.0289116	0.0756838	0.3820052	0.7024575
Primary_ReasonSelf-treat	0.1042256	0.0389198	2.6779570	0.0074073
Primary_ReasonEnjoyment	0.1691436	0.0743417	2.2752197	0.0228928
Primary_ReasonResell	-0.3148221	0.1686888	-1.8662895	0.0620009
Primary_ReasonRefuse to answer	0.0523060	0.0446617	1.1711600	0.2415345
sourceInternet	0.1361942	0.0627474	2.1705151	0.0299678
sourceInternet Pharmacy	-0.4135162	0.1078794	-3.8331344	0.0001265
sourcePersonal	0.0591284	0.0267121	2.2135460	0.0268600
sourceHeard it	0.1788115	0.0314385	5.6876587	0.0000000

dose_per_mg: Controlling for other variables, for every one milligram increase of dosage, the price per mg of oxymorphone increases by a multiplicative effect of exp(-0.0151868) = 0.9849279, about a 1.51% reduction.

bulk_purchase ("not bulk purchase" is the reference): Controlling for other variables, bulk purchase reduces the price per mg of oxymorphone by a multiplicative effect of exp(-0.1691121) = 0.8444142, about a 15.56% reduction.

source ("Source: Unknown" is the reference, other levels ommitted for brevity): Controlling for other variables, purchasing oxymorphone with Internet source tends to increase the price per mg by a multiplicative effect of exp(0.1339468) = 1.143332, about a 14.33% increase than the price per mg of oxymorphone for people with unknown drug source.

Primary Reason ("Blank" is the reference, other levels omitted for brevity): Controlling for other variables, purchasing oxymorphone with primary reason Self-treat tends to increase the price per mg by a multiplicative effect of exp(0.1034974) = 1.109043, about a 10.9% increase than the price per mg of oxymorphone for people with Blank primary reason.

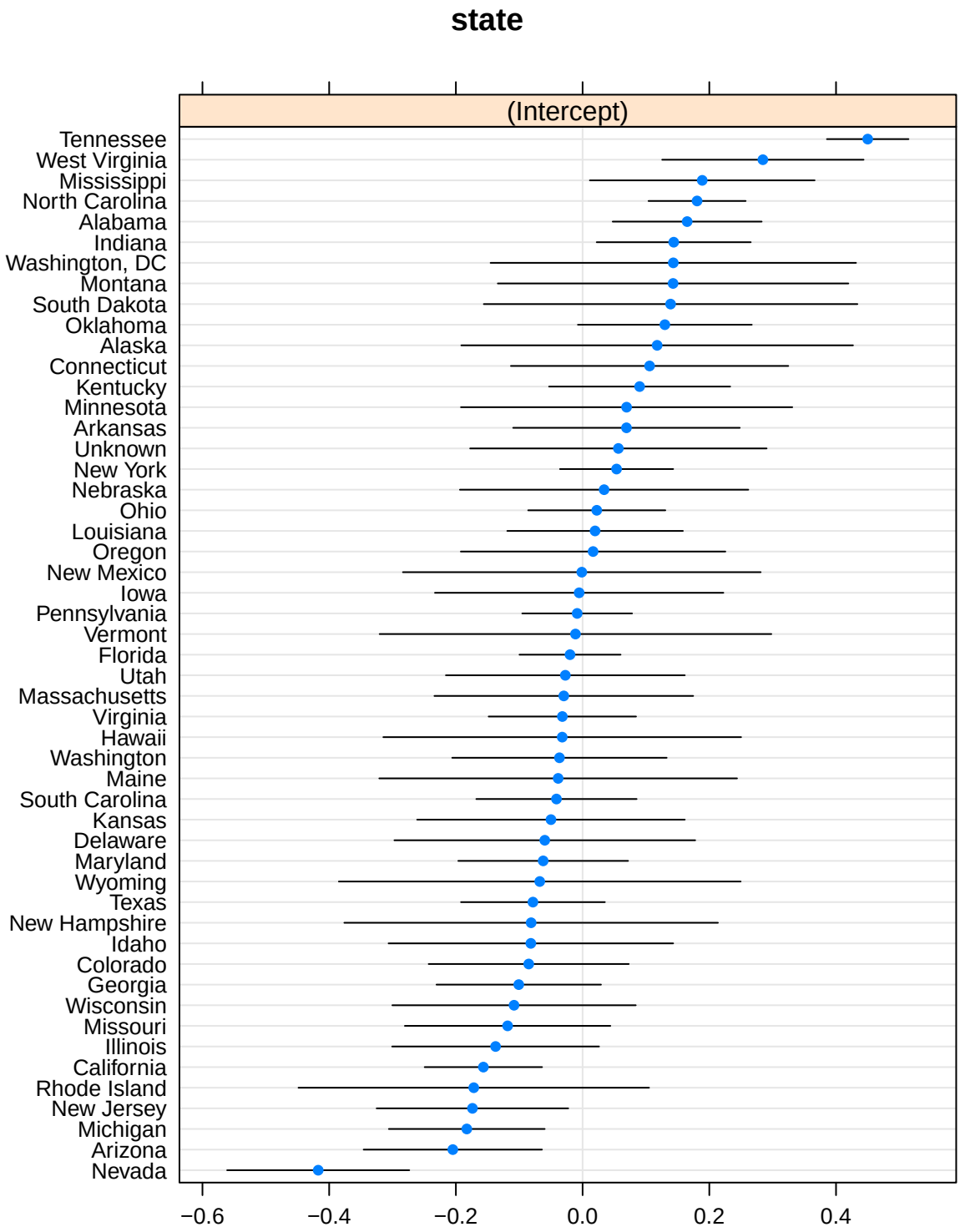
Random Effects

Group	Variance	Std.Dev
state (Intercept)	0.02292	0.1514
Residual	0.5006	0.70753

The within-state variance σ^2 is 0.5006, which is much larger than the across-state variance τ^2, 0.02292.

Intraclass Correlation
0.044

Intraclass correlation is 0.044, which means values are not very similar from the same state.



From the result of the random effects, we can see that the random intercept for Tennessee is the highest and it is significantly different from zero. The random intercept for Arizona is the lowest and this intercept is also significant.

Appendix

```
knitr::opts_chunk$set(echo = FALSE, cache=TRUE, fig.height = 3)
library(pacman)
pacman::p_load(tidyverse, lme4, gridExtra, grid, ggplot2, lattice, redres, stringr, influence.ME,
# devtools::install_github("goodekat/redres")
load("streetrx.rdata")
df <- streetrx %>% filter(api_temp == "oxymorphone", !is.na(ppm))
# cleaning state:
# table(df$state)
# remove levels corresponding to states with 0 observations in df
df$state <- droplevels(df$state)
# table(df$state)
# drop North Dakota - one observation
df <- df %>% filter(state != "North Dakota")
df$state <- droplevels(df$state)
# table(df$state)
# replace state "USA" with "Unknown"
df$state <- recode_factor(df$state, "USA" = "Unknown")
# table(df$state)
### PUT IN APPENDIX: the state-by-region table of coutns below
# table(df$state, df$USA_region)
# Change date to numeric (easier for interpretation)
# this sets
df$date_num <- as.numeric(as.Date(df$price_date, "%m/%d/%y")) - 14621 # sets 1/12/2010 (first pos
tempdf <- df %>% select(date_num, price_date)
# this filters out all data before the 1/1/2000, which probably shouldn't be included
df <- df %>% filter(date_num >= 0)
df$date_quarter <- as.numeric(substr(df$yq_pdate, 5, 5))
hist(df$date_num)
knitr::kable(table(df$date_quarter))
# recode sources
# remove unused levels
df$source <- droplevels(df$source)
source_df <- data.frame(table(df$source))
# levels(df$source)
levels(df$source)[1] <- "Blank"
source_df <- data.frame(table(df$source))
df$source <- recode_factor(df$source, "Blank" = "Blank",
                           "Internet" = "Internet",
                           "Internet Pharmacy" = "Internet Pharmacy",
                           "Personal" = "Personal",
                           "Heard it" = "Heard it",
                           "Idk" = "Unknown",
                           "w" = "Unknown",
                           "STREET PRICE" = "Heard it",
                           .default = "Internet") # everything else is a URL
source_df_2 <- data.frame(table(df$source))
# drop unknowns
df <- df %>% filter(df$source != "Unknown")
```

```r
df$source <- droplevels(df$source)
knitr::kable(table(df$source), col.names = c("Level", "Freq"))
#table(df$bulk_purchase, useNA = "always")
# looks fine
# remove unused levels
df$Primary_Reason <- droplevels(df$Primary_Reason)
reason_df <- data.frame(table(df$Primary_Reason))
# levels(df$Primary_Reason)
levels(df$Primary_Reason)[1] <- "Blank"
reason_df <- data.frame(table(df$Primary_Reason))
df$Primary_Reason <- recode_factor(df$Primary_Reason, "Blank" = "Blank",
                                   "0 Reporter did not answer the question" = "Left unanswered",
                                   "Other or unknown" = "Other or unknown",
                                   "9 To self-treat my pain" = "Self-treat",
                                   "3 To prevent or treat withdrawal" = "Self-treat",
                                   "10 To treat a medical condition other than pain" = "Self-trea
                                   "4 For enjoyment/to get high" = "Enjoyment",
                                   "5 To resell" = "Resell",
                                   "8 Prefer not to answer" = "Refuse to answer",
                                .default = "Other or unknown")
reason_df_2 <- data.frame(table(df$Primary_Reason))
#print(reason_df)
knitr::kable(reason_df_2, col.names = c("Level", "Freq"))
hist(df$mgstr)
# rename variable for ease of interpretability
df <- df %>% rename(dose_per_mg = mgstr)
# log transform ppm
df$log_ppm <- log(df$ppm)
hist_ppm <- ggplot(df, aes(x=ppm)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white", bins = 50)+
 geom_density(alpha=.2, fill="red")
hist_log_ppm <- ggplot(df, aes(x=log_ppm)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white", bins = 50)+
 geom_density(alpha=.2, fill="blue")
grid.arrange(hist_ppm, hist_log_ppm, ncol=2)
df <- df %>% filter(log_ppm > -2 & log_ppm < 2)
ggplot(df, aes(x=log_ppm, y=reorder(state, state, length), fill=state)) +
geom_boxplot() +
labs(title="State", x="Log PPM by state",y="Log PPM")
ggplot(df, aes(x=reorder(USA_region, USA_region, length), y=log_ppm, fill=USA_region)) +
geom_boxplot() +
labs(title="log PPM by region", x="region",y="Log PPM")
ggplot(data = df[df$state=="Florida" | df$state=="Texas" | df$state=="North Carolina",], aes(x =
  geom_point() +
  facet_grid(~ state)
ggplot(data = df[df$state=="Florida" | df$state=="Texas" | df$state=="North Carolina",], aes(x =
  geom_point() +
  facet_grid(~ state)
# categorical variables -- looking at South only
```

```r
set.seed(42)
states <- sample(df$state, 5)
ggplot(data = df %>% filter(state %in% states), aes(x = bulk_purchase, y = log_ppm)) +
  geom_boxplot() +
  facet_grid(~ state) +
  theme(axis.text.x = element_text(angle = 90))
ggplot(data = df[df$state=="Florida" | df$state=="Texas" | df$state=="North Carolina",], aes(x =
  geom_boxplot() +
  facet_grid(~ state)
ggplot(data = df[df$state=="Florida" | df$state=="Texas" | df$state=="North Carolina",], aes(x =
  geom_boxplot() +
  facet_grid(~ state) +
  theme(axis.text.x = element_text(angle = 90))
ggplot(data = df[df$state=="Florida" | df$state=="Texas" | df$state=="North Carolina",], aes(x =
  geom_boxplot() +
  facet_grid(~ state) +
  theme(axis.text.x = element_text(angle = 90))
# Fit 3 models grouping on different location specificity
mod_s <- lmer(data=df, log_ppm ~ (1 |state) + date_num + date_quarter + dose_per_mg + bulk_purcha
mod_c <- lmer(data=df, log_ppm ~ (1 |city) + date_num + date_quarter + dose_per_mg + bulk_purchas
mod_r <- lmer(data=df, log_ppm ~ (1 |USA_region) + date_num + date_quarter + dose_per_mg + bulk_p
BIC <- sapply(c(mod_s, mod_c, mod_r), BIC)
re_results <- data.frame('Grouping' = c('State', 'City', 'Region'), 'BIC' = BIC)
kable(re_results, caption = 'BIC for Models with Different Random Intercepts')
# Test date_num
drop_date_num <- anova(mod_s,
      lmer(data=df, log_ppm ~ (1 |state) + date_quarter + dose_per_mg + bulk_purchase + Primary_R
date_num_pval <- drop_date_num$`Pr(>Chisq)`[2]
# Test date_quarter
drop_quarter <- anova(lmer(data=df, log_ppm ~ (1 |state) + date_quarter + dose_per_mg +
                     bulk_purchase + Primary_Reason + source, REML=F),
      lmer(data=df, log_ppm ~ (1 |state) + dose_per_mg + bulk_purchase + Primary_Reason + source,
quarter_pval <- drop_quarter$`Pr(>Chisq)`[2]
# Test dose_per_mg
drop_dose <- anova(lmer(data=df, log_ppm ~ (1 |state) + dose_per_mg + bulk_purchase + Primary_Rea
      lmer(data=df, log_ppm ~ (1 |state) + bulk_purchase + Primary_Reason + source, REML=F))
dose_pval <- drop_dose$`Pr(>Chisq)`[2]
# Test bulk purchase
drop_bulk <- anova(lmer(data=df, log_ppm ~ (1 |state) + dose_per_mg + bulk_purchase + Primary_Rea
      lmer(data=df, log_ppm ~ (1 |state) + dose_per_mg + Primary_Reason + source, REML=F))
bulk_pval <- drop_bulk$`Pr(>Chisq)`[2]
# Test primary reason
drop_reason <- anova(lmer(data=df, log_ppm ~ (1 |state) + dose_per_mg + bulk_purchase + Primary_R
      lmer(data=df, log_ppm ~ (1 |state) + dose_per_mg + bulk_purchase + source, REML=F))
reason_pval <- drop_reason$`Pr(>Chisq)`[2]
# Test source
drop_source <- anova(lmer(data=df, log_ppm ~ (1 |state) + dose_per_mg + bulk_purchase + Primary_R
      lmer(data=df, log_ppm ~ (1 |state) + dose_per_mg + bulk_purchase + Primary_Reason, REML=F))
source_pval <- drop_source$`Pr(>Chisq)`[2]
```

```r
mod_final <- lmer(data=df, log_ppm ~ (1 |state) + dose_per_mg + bulk_purchase + Primary_Reason +
# resqq <- plot_resqq(mod1)
# ranefci <- plot_ranef(mod1)
# grid.arrange(resqq, ranefci, ncol = 2)
#knitr::kable(fixef(mod1))
#confint(mod1)[3:11,]
# extract coefficients
coefs <- data.frame(coef(summary(mod_final)))
# use normal distribution to approximate p-value
coefs$p.z <- 2 * (1 - pnorm(abs(coefs$t.value)))
knitr::kable(coefs, col.names = c("Estimate", "Std.Error", "t-value", "P-value"))
#VarCorr(mod_final)
Variance <- c(0.02292, 0.50060)
Std.Dev <- c(0.15140 , 0.70753)
Groups <- c("state (Intercept)", "Residual")
REvar <- cbind(Groups, Variance, Std.Dev)
knitr::kable(REvar, col.names = c("Group", "Variance", "Std.Dev"))
library(sjstats)
icc(mod_final)
ICC <- 0.044
knitr::kable(ICC, col.names = "Intraclass Correlation")
dotplot(ranef(mod_final, condVar=TRUE))$state
# Example visualization shown here are for all cities in the state of Florida.

ggplot(df[df$state=="Florida",], aes(x=log_ppm, y=reorder(city, city, length))) +
geom_boxplot() +
labs(title="log PPM by city (FL)", x="Log PPM",y="city")

# More EDA: slope and intercept distributions for predictors by state

# To get a better sense of how state-level groups and slopes are distributed, we created the foll

### Dose strength variable

group_int_list <- NULL
group_slope_list <- NULL
for(statej in unique(df$state)) {
    m <- lm(data = df[df$state==statej, ], log_ppm ~ dose_per_mg)
    group_int_list <- c(group_int_list, coef(m)[1])
    group_slope_list <- c(group_slope_list, coef(m)[2])
}
hist(group_int_list)
hist(group_slope_list)

### Continuous date variable

group_int_list <- NULL
group_slope_list <- NULL
for(statej in unique(df$state)) {
```

```
    m <- lm(data = df[df$state==statej, ], log_ppm ~ date_num)
    group_int_list <- c(group_int_list, coef(m)[1])
    group_slope_list <- c(group_slope_list, coef(m)[2])
}

hist(group_int_list)
hist(group_slope_list)
```