Ben Bauchwitz Case Study 1

Data Cleaning

```
## Step 1: filter out data on other drugs so we only have the drug we are studying
streetrx.m <- streetrx %>% filter(api_temp == "morphine")
```

```
## Step 3: fix noisy city data, including typos, nicknames, and alternate identifications by cross-refe
## fill in the missing values
streetrx.m$city <- as.character(streetrx.m$city)
streetrx.m[streetrx.m$city == "",]$city <- "Other/Unknown"

## convert names to upper case to enable matching across cases used
streetrx.m$city <- sapply(streetrx.m$city, toupper)
cities$city <- sapply(cities$city, toupper)
city.mapping$Old_name <- sapply(city.mapping$Old_name, toupper)
city.mapping$New_name <- sapply(city.mapping$New_name, toupper)

## find case-corrected city names that match the official database
city.intsct <- intersect(streetrx.m$city, cities$city)
city.intsct.df <- as.data.frame(do.call(rbind, as.list(city.intsct)))
city.intsct.df$Updated_city_name <- city.intsct.df$V1
streetrx.m <- merge(x = streetrx.m, y = city.intsct.df, by.x = "city", by.y = "V1", all.x = TRUE)

## For cities that are not in the official city database, check if they are in the dictionary of mispri
colnames(city.mapping)[2] = "Updated_city_name"
streetrx.m <- merge(x = streetrx.m, y = city.mapping, by.x = "city", by.y = "Old_name", all.x = TRUE)
streetrx.m <- streetrx.m %>% mutate(Updated_city_name.x = coalesce(Updated_city_name.x, Updated_city_na

## If the name still wasn't found, fill in with "Other/Unknown"
streetrx.m$Updated_city_name.x <- as.character(streetrx.m$Updated_city_name.x)
streetrx.m[is.na(streetrx.m$Updated_city_name.x),]$Updated_city_name.x <- "Other/Unknown"
streetrx.m$City_final = paste(streetrx.m$Updated_city_name.x, "-", streetrx.m$state)
streetrx.m$City_final <- as.factor(streetrx.m$City_final)
```

Exploratory Data Analysis

In this part, we explore distributions of the variables, potential relationships, and potential interactions that we might include in the model.

1. Response Variable

First, we examine the `ppm` variable (price per milligram), our primary outcome variable. Even though the values don't vary over many orders of magnitude, it has extreme right skew. Taking the log of these values appears to dramatically improve various indicators of normality while the data still fails a Shapiro-Wilk normality assessment this transform appears reasonable for the data.

2. Group Variables

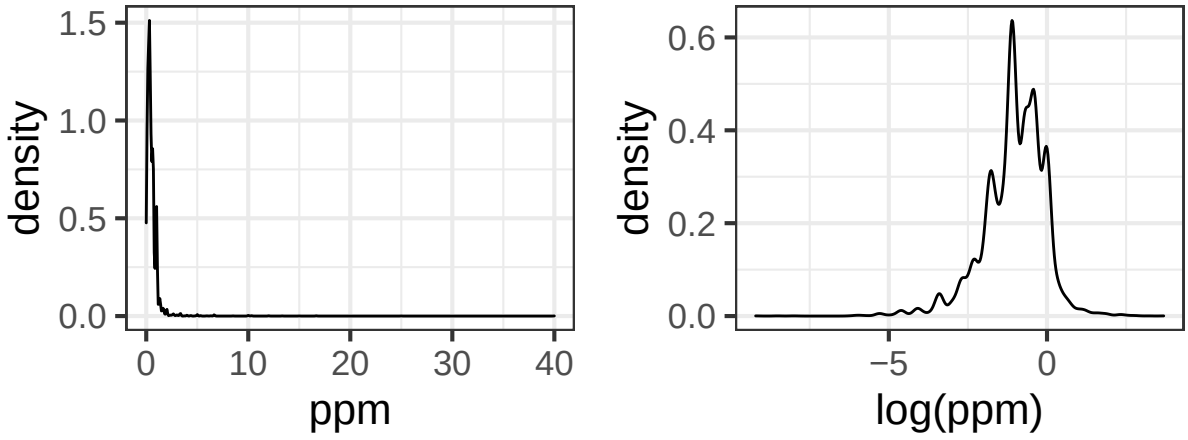


Figure 1: Log Transformation on Response

City There are 1654 unique levels of `City_final`, which indicates the unique cities recorded in our data set. To explore whether there is a relationship between the cities and the log price per milligram of morphine, we make a box plot of randomly chosen 25 cities. The log price per milligram of morphine seems to differ by cities.

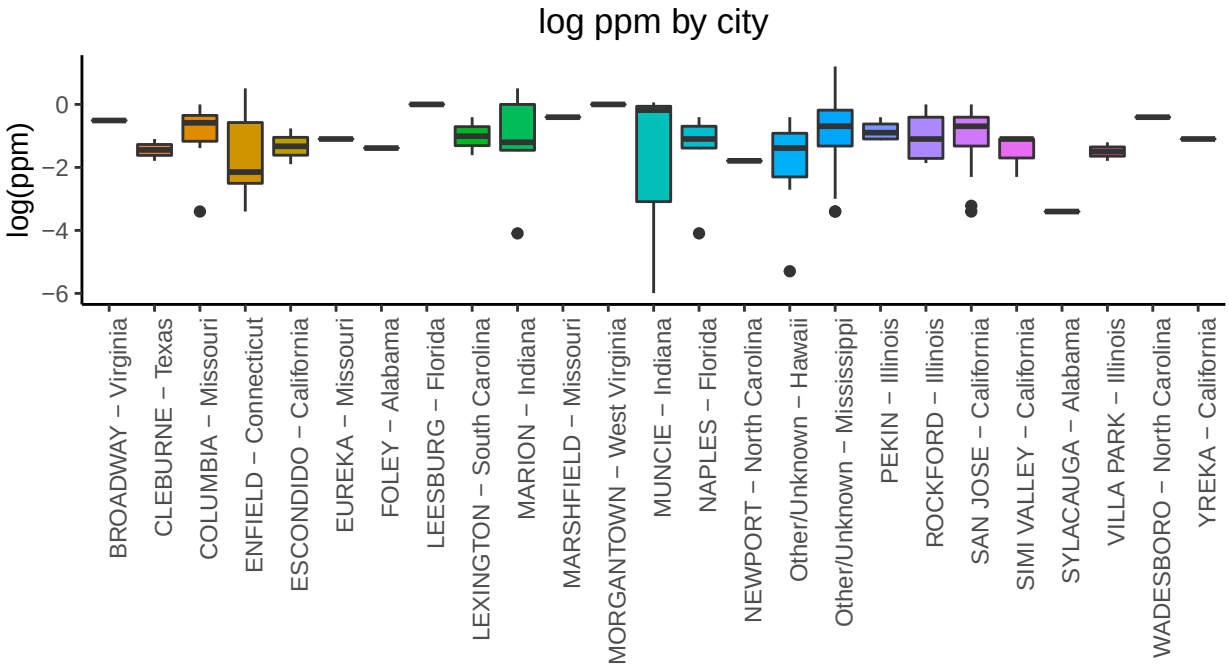


Figure 2: log ppm differs by cities

State Next, we explore the relationship between log ppm and state. We can observe that the log ppm differs by the states.

Region Finally, we observe that the log ppm also differs by USA regions, but such difference is not quite obvious.

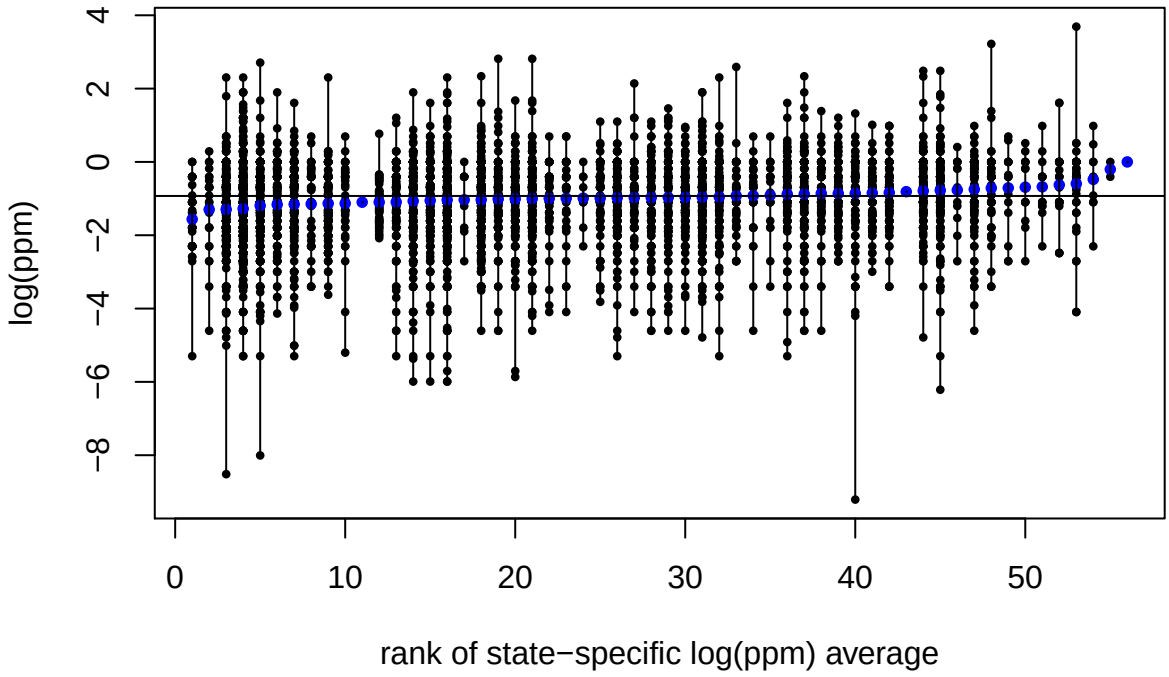
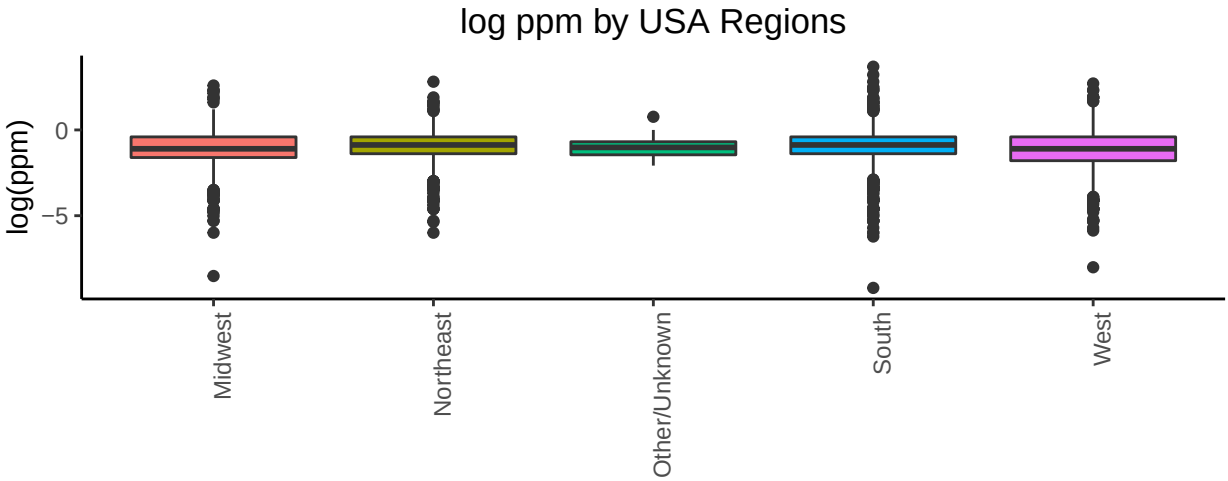


Figure 3: log ppm differs by state



However, to determine which one of these grouping variables should be included in the model, we need to perform formal tests and analyses. Preliminarily, we make three models with cities, states, and regions as the only predictors respectively, and perform ANOVA tests. The results show that log ppm differs significantly by states and regions, but not that much by cities.

We then build three random effects anova models with only cities, states, and regions as predictors respectively. The plots of random effects show that state might be the best to included individually as our grouping variables, it indicates reasonable amount of information to demonstrates the heterogeneity of log ppm across locations.**(See Appendix for random effect plots for cities and regions)** A series of nested tests for including multiply grouping variables will be performed in the Model section.

3. Fixed Effect Predictors

Days The first variable we consider is the linear temporal variable we created, which is number of days since Jan 1, 2010. This lets us encode time in a semi-continuous manner. Despite the discrete encoding, the

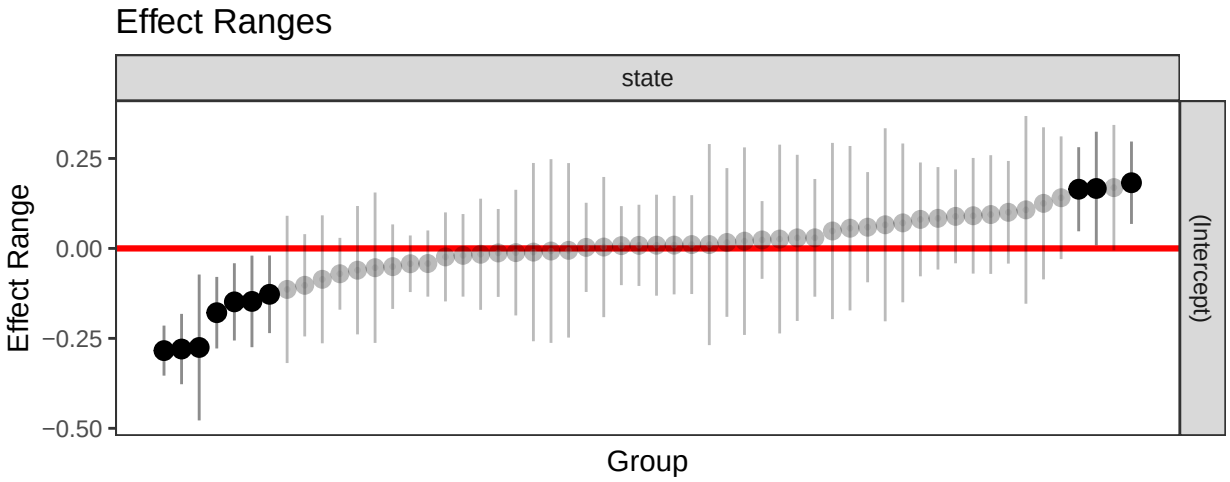


Figure 4: Distribution of State Random Effects

data actually functions quite well as a numeric predictor since there is high resolution relative to the overall time scale. The only apparent thing is that when we plot the price data against over time, the data is clearly sparser earlier in streetrx's history. However, with the appropriate parameter selection, this should not be a significant issue in the overall models we will create.

However, by only looking at the relationship between this variable and the log ppm, we do not observe obvious linear association.**(See Appendix Fig.7)**

We then consider the effects of days in each group, presumably by state. We randomly sample 8 states. In some states, the log ppm seems to change across the days elapsed. Such change may not be very obvious as the data points are sparse. Still, it might be worth trying to add random slopes of days by state to the model.

Dosage Strength Although `mgstr`, the dosage strength, is numeric, values are reported at discrete intervals, perhaps due to standard packaging sizes or users rounding off the values that they reported. All told, there are 16 unique values ranging from 1 to 200. We have two options: treat this as a categorical variable or accept the sparse numeric coding. If we do ultimately evaluate `mgstr` in our model, it does not seem that we would be interested in categorical relationships. In other words, we don't have any reason to believe there is something unique about individual package size levels. Instead, we are concerned with the general trend in how different volumes affect price. Since 16 points is more than adequate to fit a line in most applications and since the data we've collected has multiple volume sizes at each order of magnitude, it seems reasonable to continue with the numeric encoding.

There is a slightly negative linear relationship between dosage strength and log ppm. We would not consider a random slope by dosage strength by state.**(See Appendix Fig.14)**

Primary Reason (See Appendix Fig.8 and Fig.9) It seems that, in general, log ppm does not differ too much across different primary reasons for purchasing morphine.

With-in each state, the log ppm differs according to different primary purchase reasons. Therefore, it is possible to include random slopes of primary reasons by state in the model.

Source (See Appendix Fig.10 and Fig.11) It seems that in general, the log ppm varies according to the values of the sources class, which indicates the source of each transactions recorded in the data set.

By examining the associations between source class and log ppm in a random sample of states, we can see that such associations vary across states. It might be worth considering random slopes of sources class by states in the model.

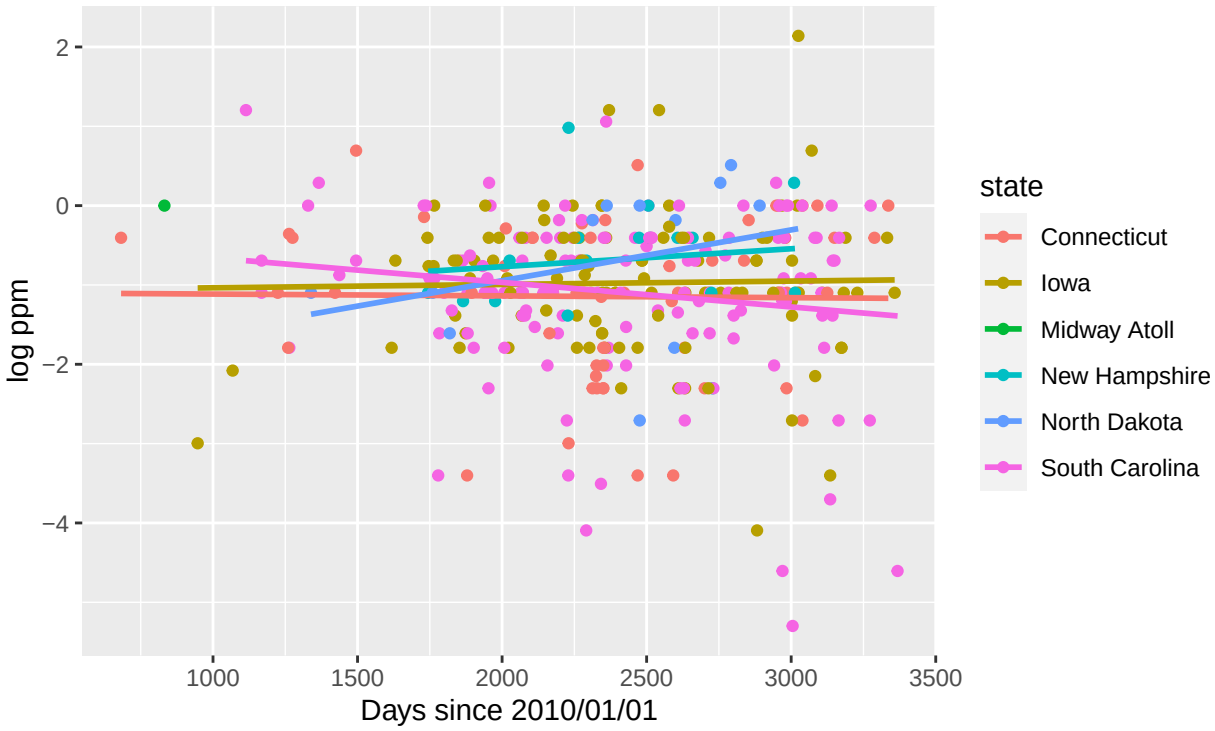


Figure 5: Change in log ppm by time in different states

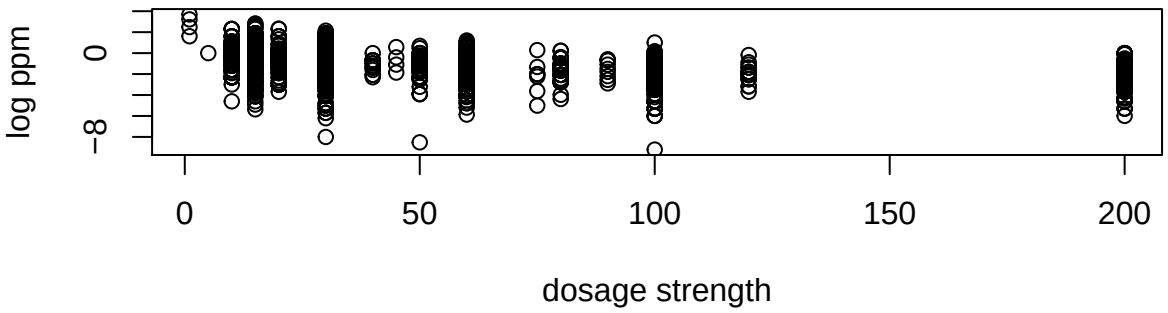
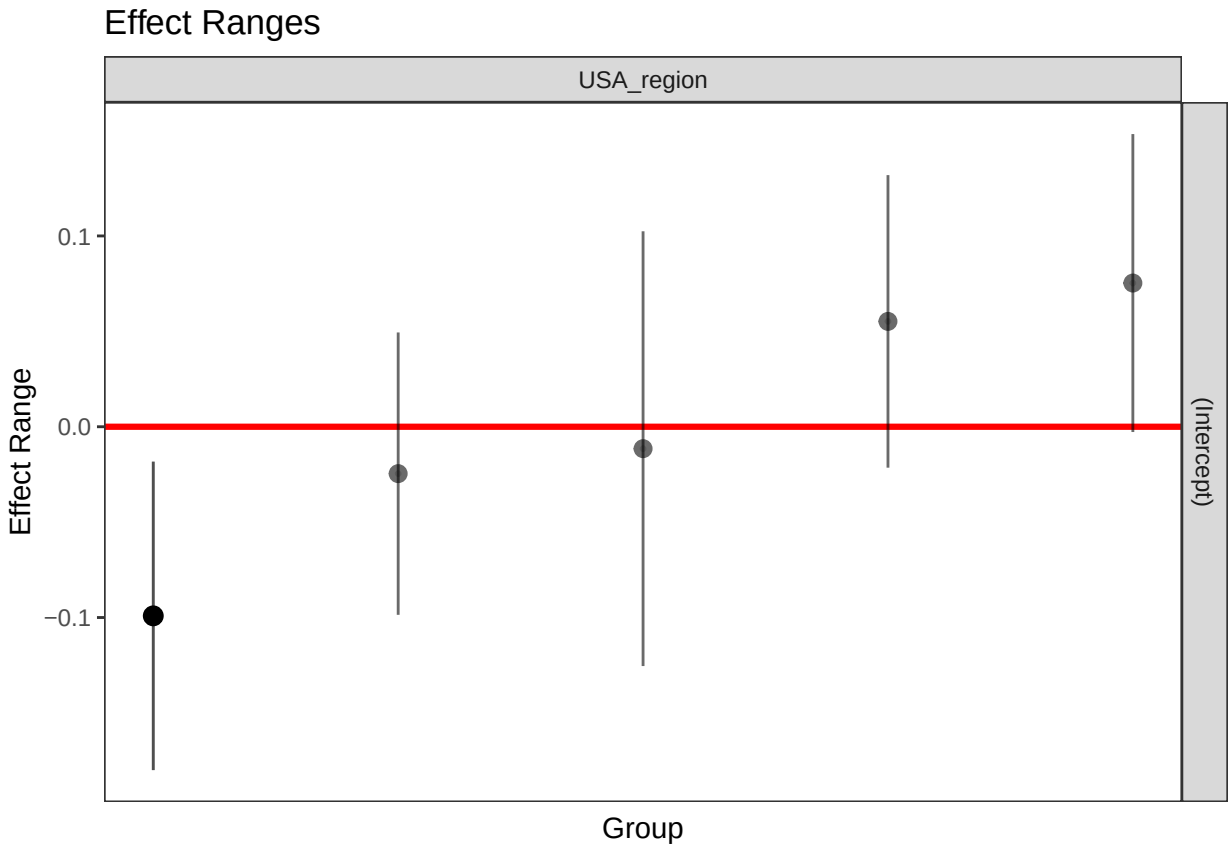
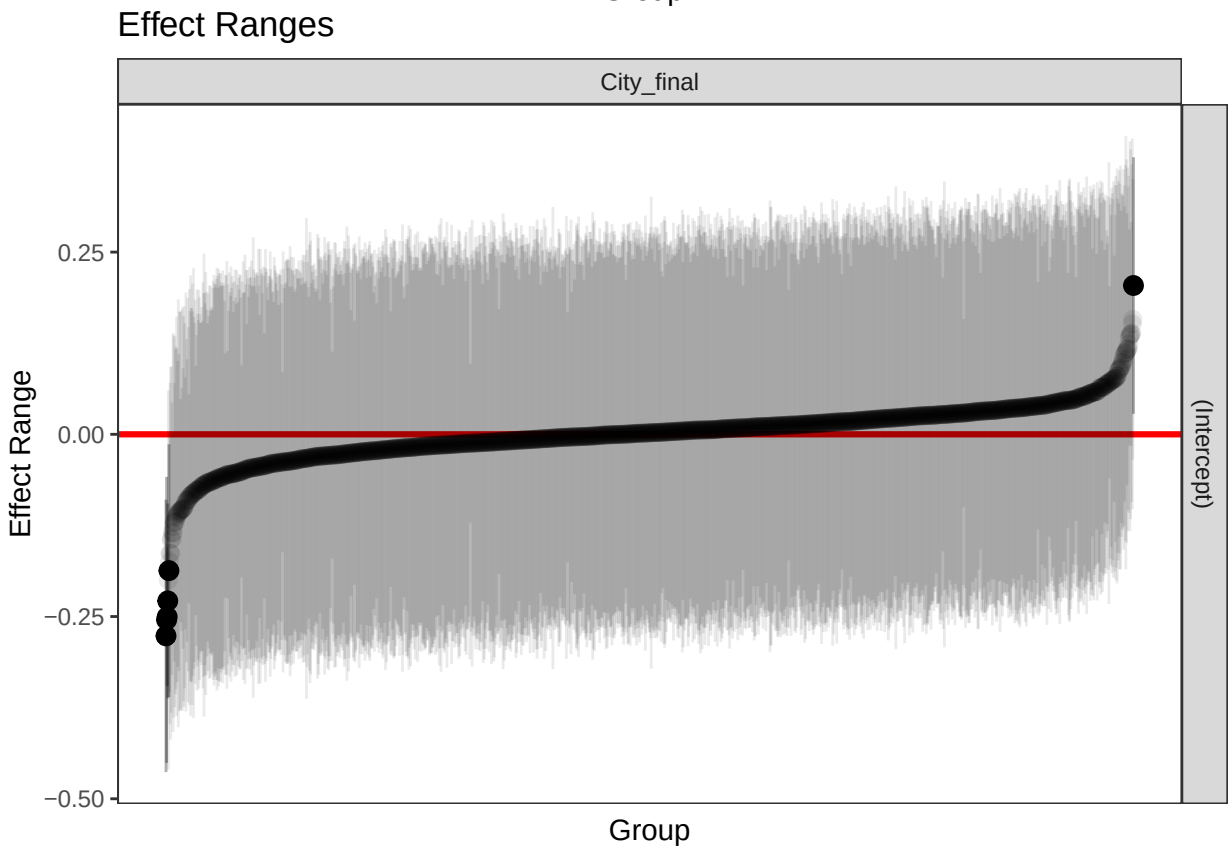


Figure 6: Slightly negative association between log ppm and dosage strength

Bulk (See Appendix Fig.12 and Fig.13) The indicator for purchasing 10 units at once, Bulk, does not seem to have an obvious relationship with log ppm. We will not include a random slope of bulk by state either as the plot does not show that such association vary by states.

Appendix


Effect Ranges
USA_region
0.1
0.0
−0.1
Effect Range
(Intercept)
Group


Effect Ranges
City_final
0.25
0.00
−0.25
−0.50
Effect Range
(Intercept)
Group

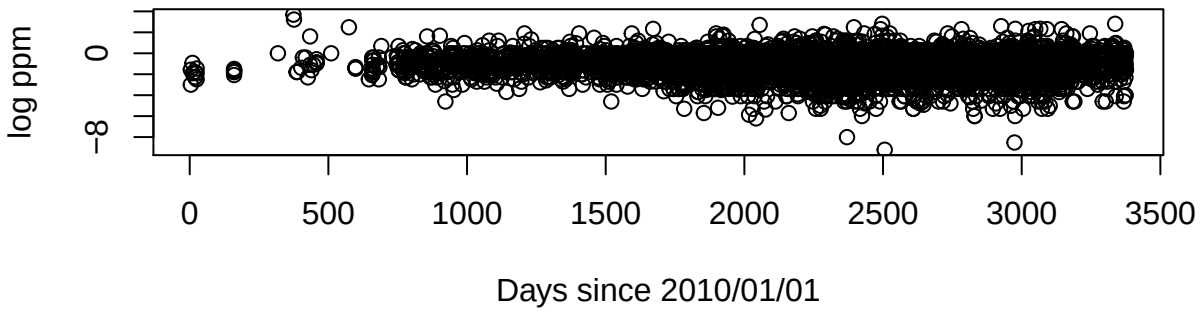


Figure 7: Relationship between log ppm and Days Elasped

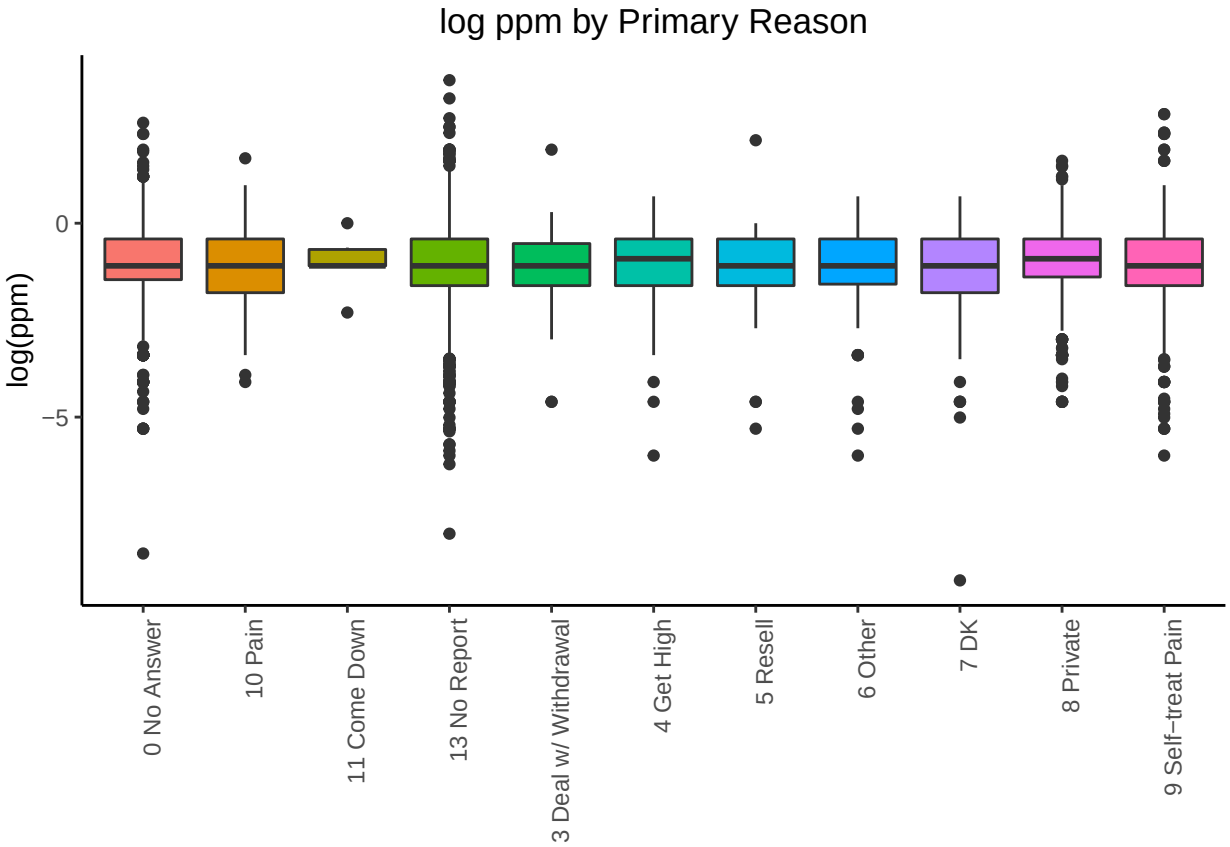


Figure 8: Relationship between log ppm and Primary Reasons

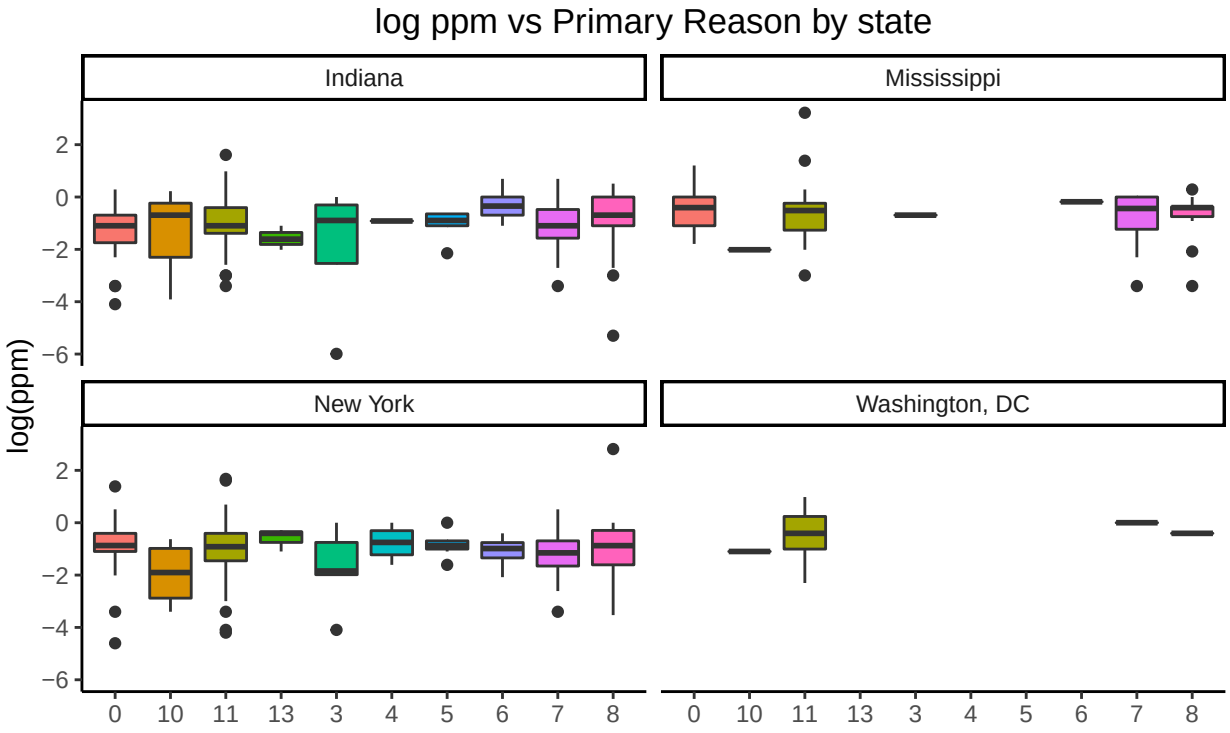


Figure 9: Analysis: random slopes of Primary Reasons by State

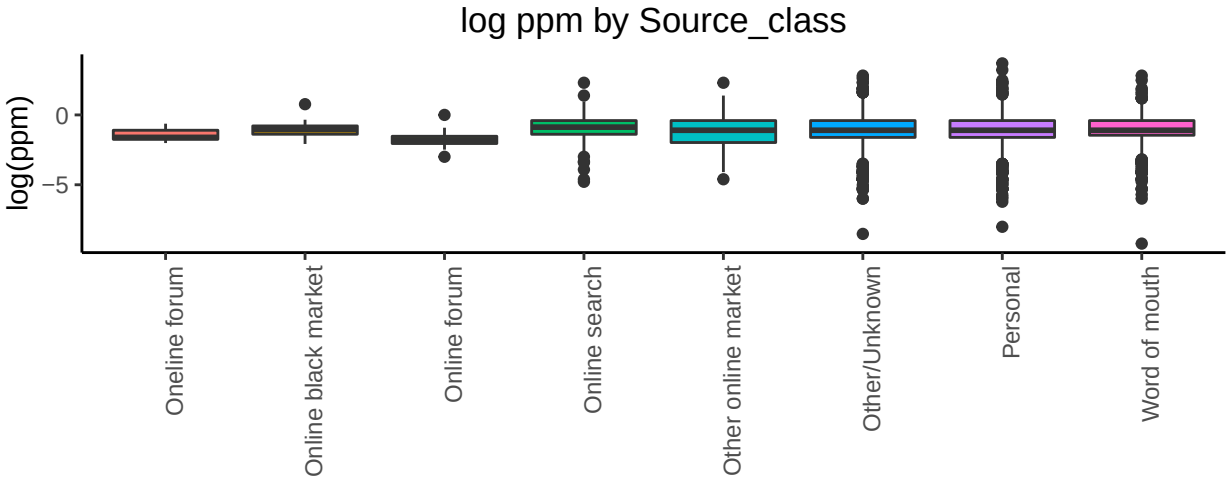


Figure 10: Relationship between log ppm and Source Class

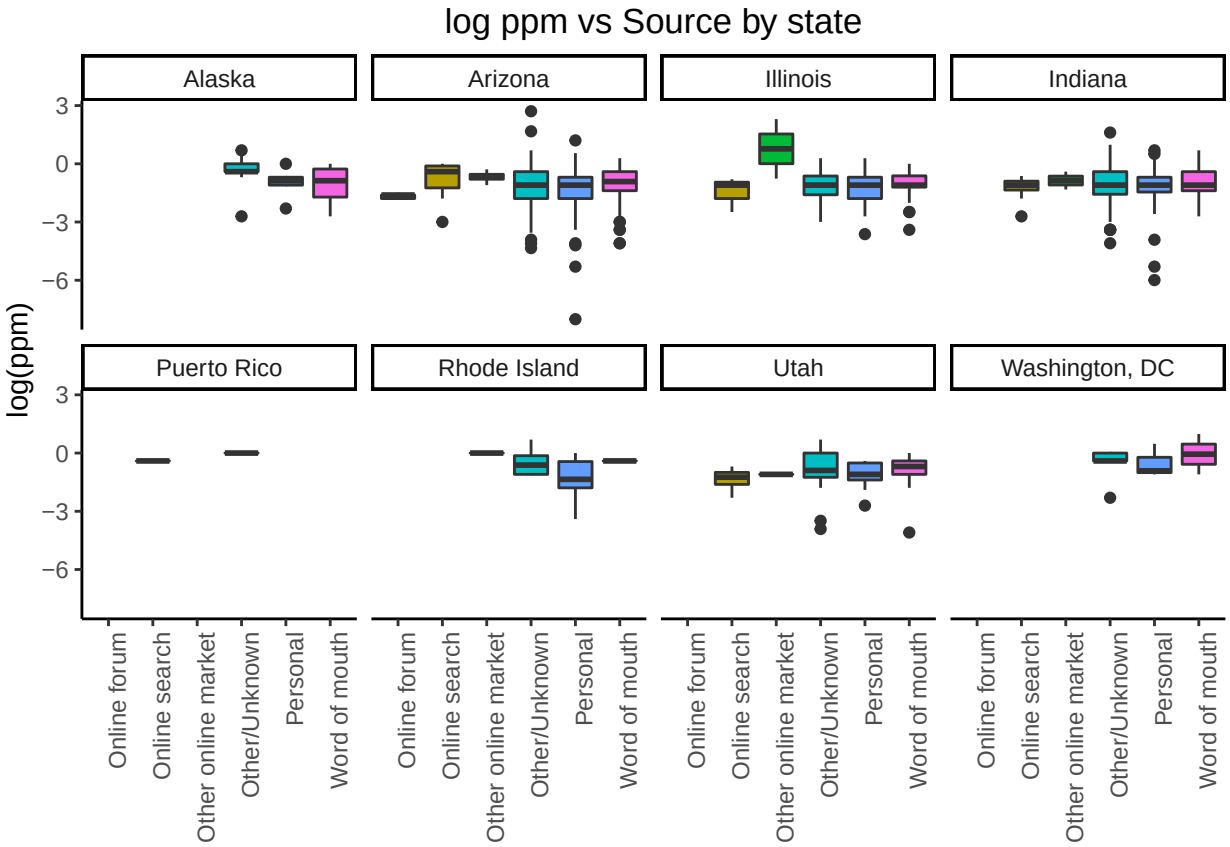


Figure 11: Analysis: random slopes of Source class by State

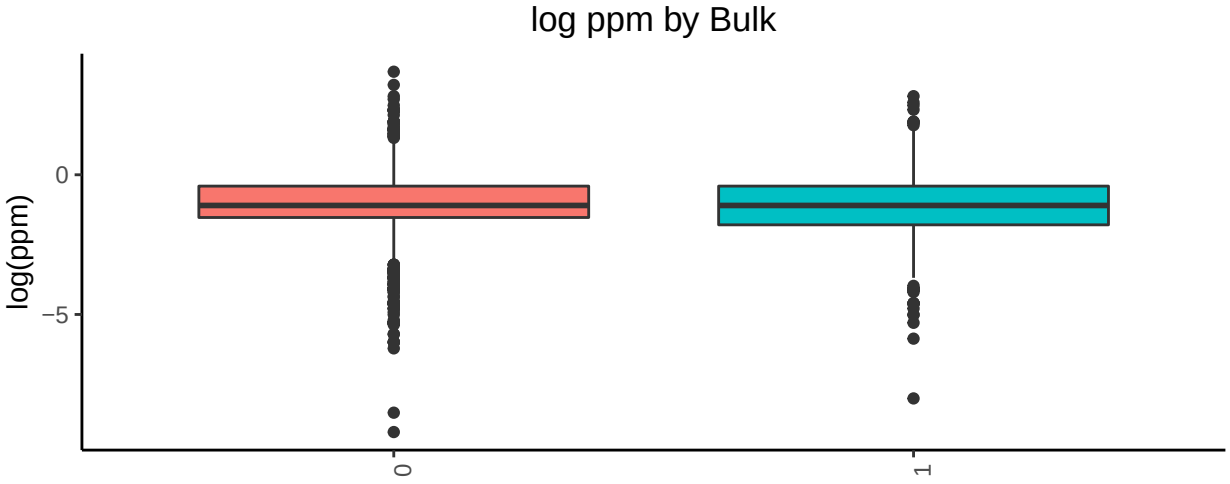


Figure 12: Relationship between log ppm and Bulk

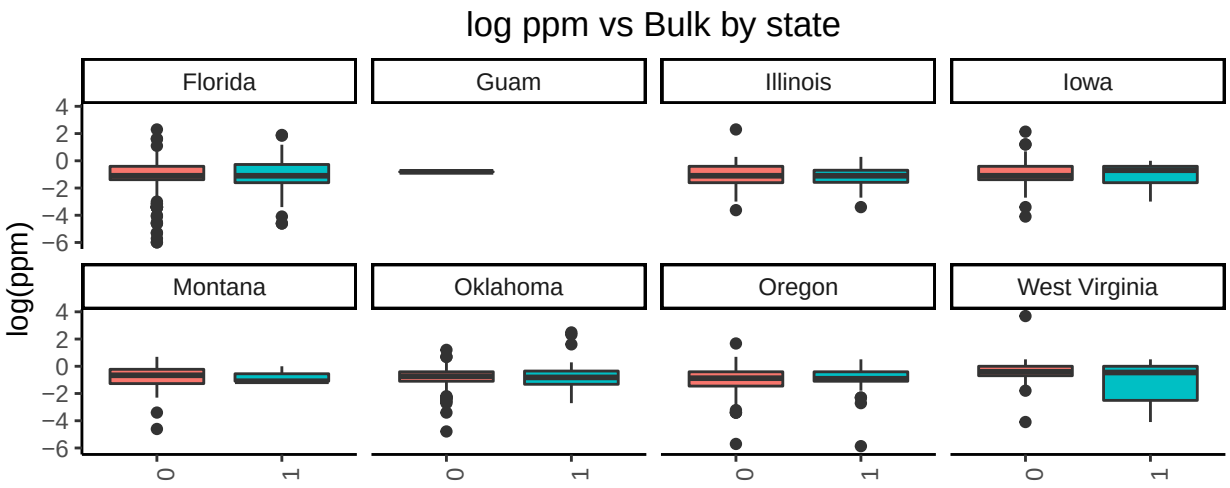


Figure 13: Analysis: random slopes of Bulk by State

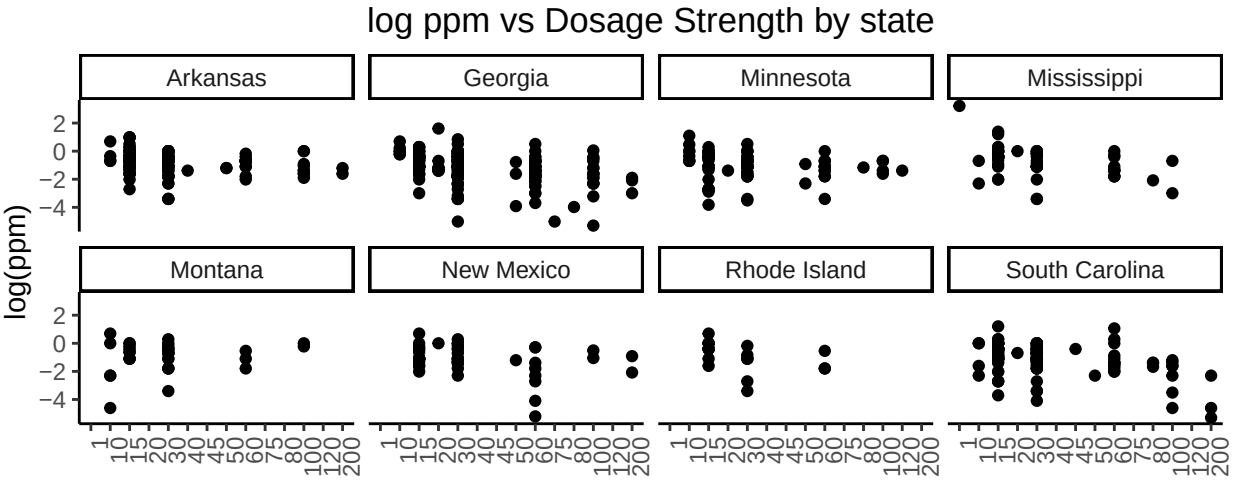


Figure 14: Analysis: random slopes of Dosage Strength by State