STA610 Case Study 1

Emily Gentles (Presenter) Weiyi Liu (Writer) Jack McCarthy (Programmer)
Qinzhe Wang (Coordinator & Checker)

13 October, 2021

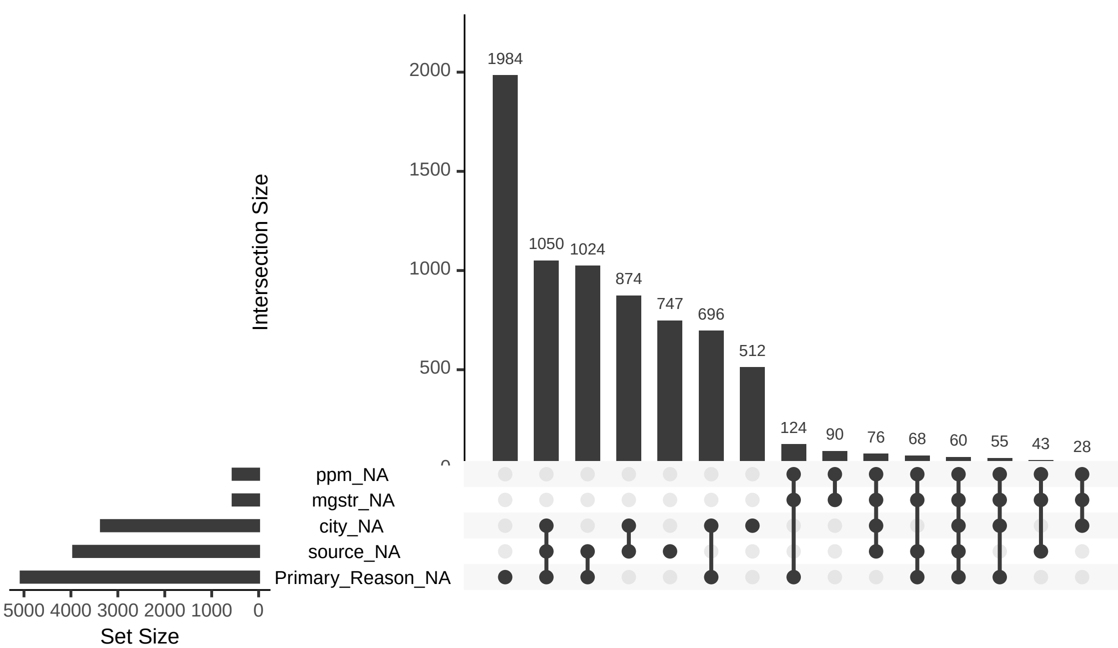
Introduction

Prescription opioid abuse plays an essential role in public health issues. The price of prescription opioids indicates the supply-demand relationship of drugs. This study case aims to explore the relationship between drugs' unit price and other factors. More specifically, our group's interest is to explore the factors related to the cost per milligram and the heterogeneity in the region. The dataset is provided by StreetRx, a reporting tool for people at large to anonymously report the price they paid or heard for diverted prescription drugs.

Our drug interest is Morphine. Morphine is used to "relieve moderate to severe pain and maybe habit-forming," especially with prolonged use (MedlinePlus).

EDA

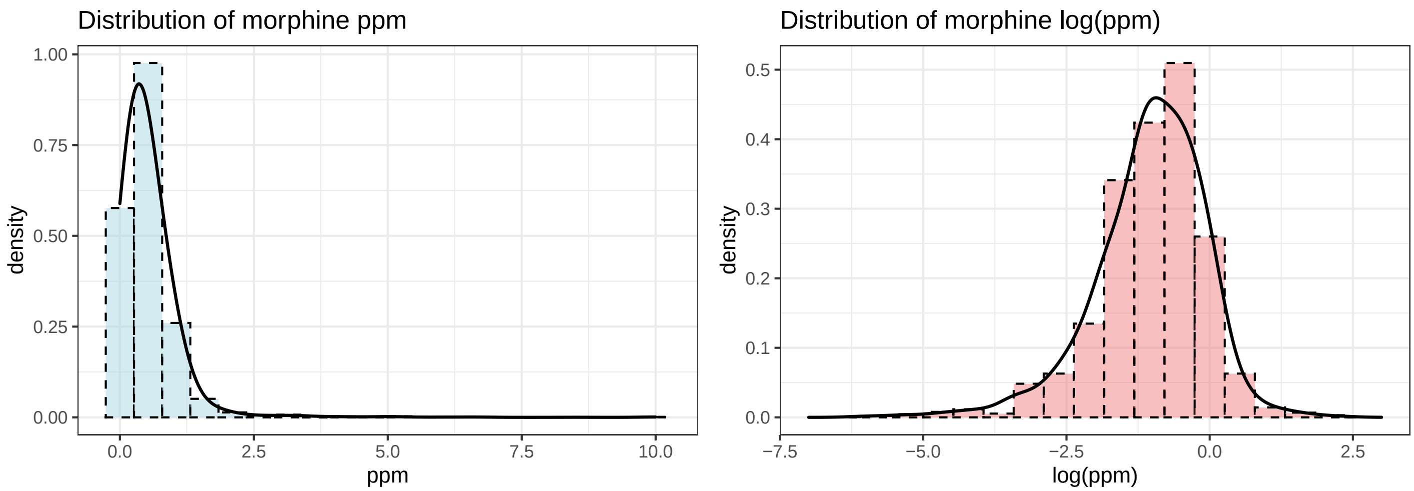
Missing Values



The dataset (Morphine) contains 9,268 observations with 13 variables. There are 13,443 empty cells (both the missing values and the blank). To maintain the statistical power and avoid bias, our group decided to recode the empty cells in the feature `Primary_Reason` (5061 in total) as "8 Prefer not to answer" and recode the empty cells in `source` (3942 in total) as "Blank" because of the high missing rates. Then, we remove rows containing missing values.

In addition, our group thinks there is no reason that the price per milligram can be a non-positive value or values greater than 10 (may because some people input the total price by mistake). The number of the ovservations we have is 5,582 now.

Price per milligram (Response variable)

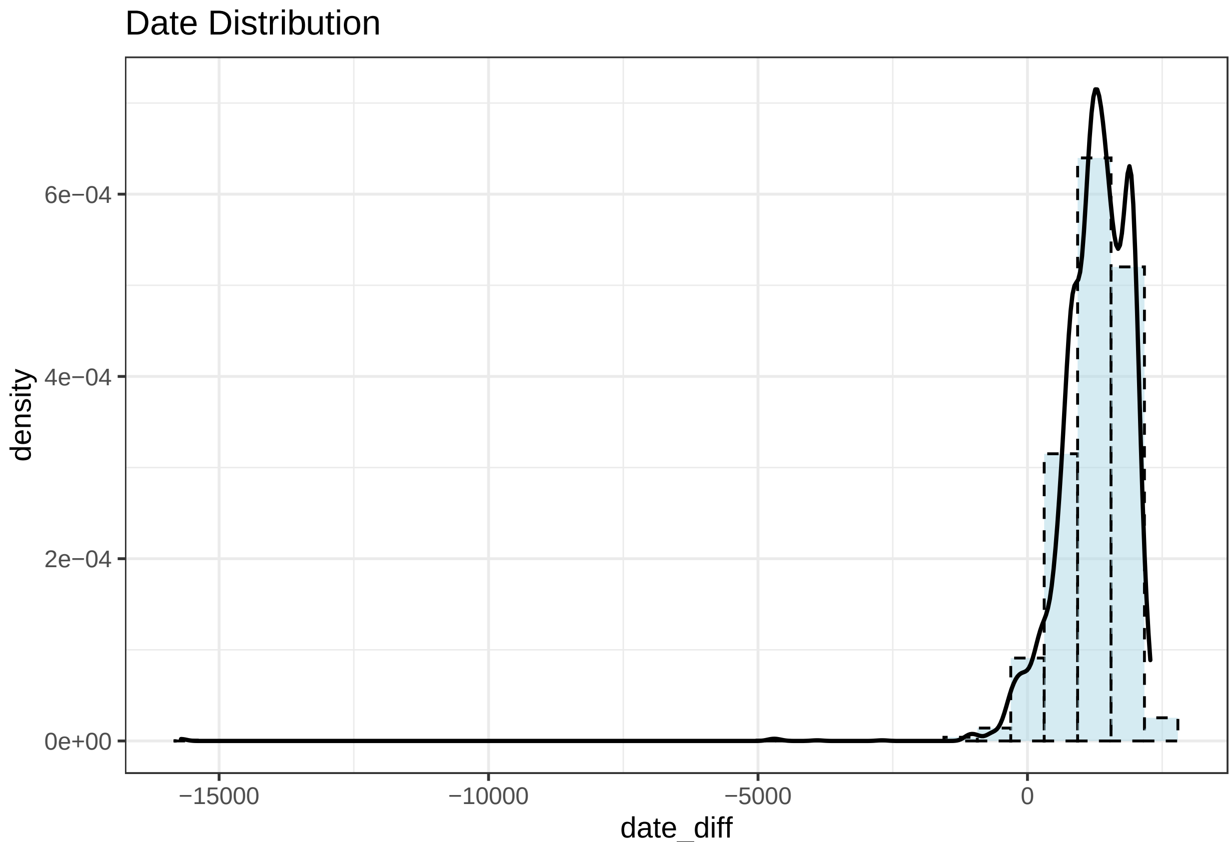


Whether in a hierarchical model or linear regression, the response variable has to be normally distributed. From the histogram on the left, the distribution of ppm is clearly right-skewed, and it is strictly nonnegative in value, so a log transformation may be appropriate. The distribution of log(ppm) is given above, and appears closer to the desired normal.

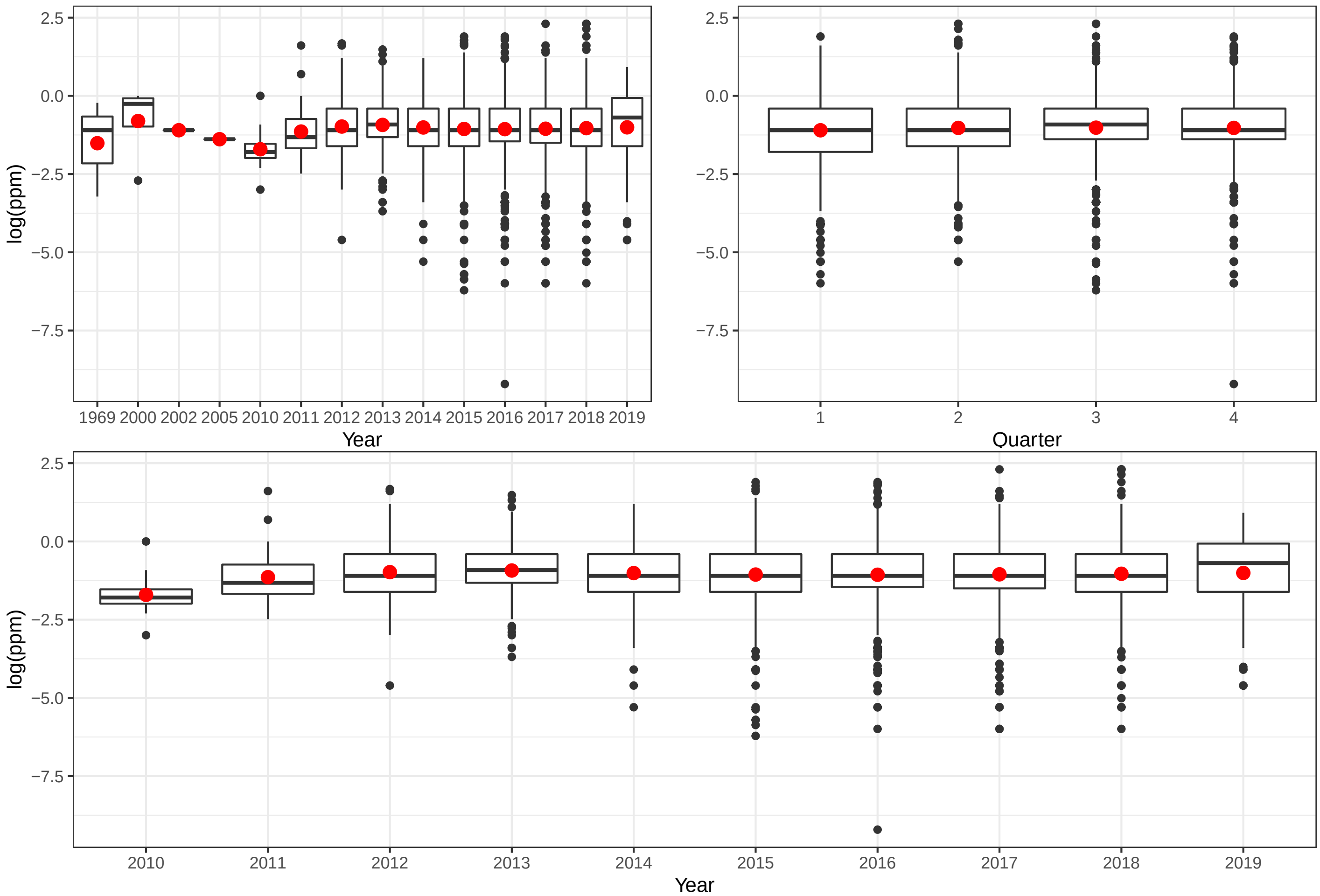
Date

Our group has considered three ways of data cleaning on the date variable. The first choice is to choose a starting date and convert the feature as the date differences from that starting date. The second choice is to split this date variable into two components, year and quarter, because we want the determine the relationship between drug unit price and seasons.

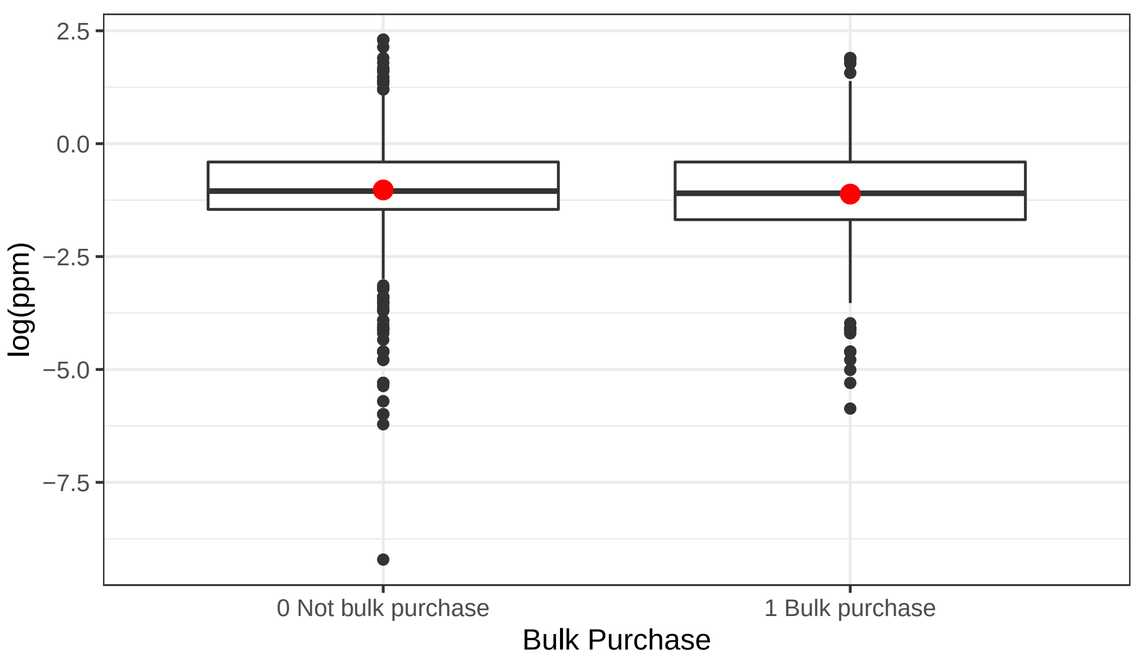
record `price_date` as a continuous variable counting days from some start date.


Date Distribution
density
6e−04
4e−04
2e−04
0e+00
−15000
−10000
−5000
0
date_diff

year & quarter vs.log(ppm)



bulk_purchase vs.log(ppm)

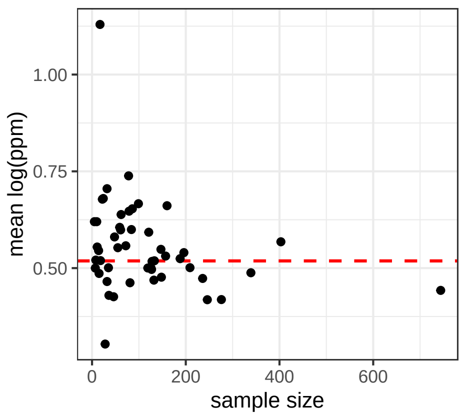


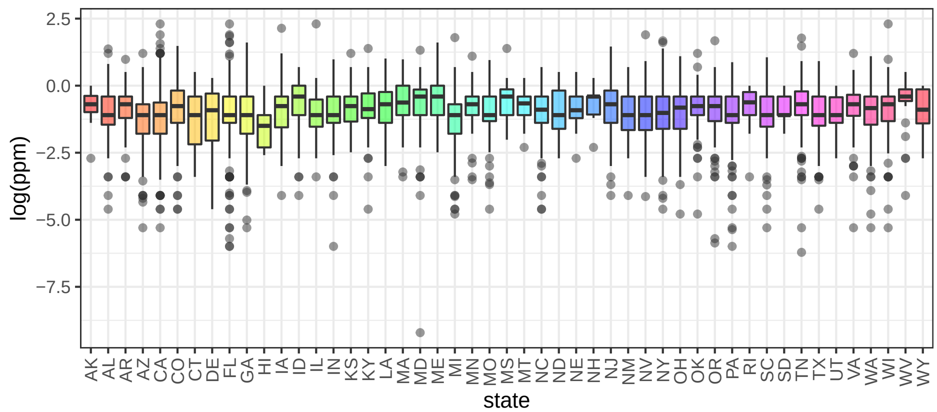
state vs. log(ppm)

We see that there are 4 states that have a sample size of 1, North Dakota, Vermont, Washington DC, and Wyoming, as well as 1 state that has a sample size of 2, Alaska. Due to the extremely small sample sizes we decided to remove these states form our dataset to avoid computational instability.

Table 1: 7 States with Smallest Sample Size

Puerto Rico	Vermont	North Dakota	South Dakota	Wyoming	New Hampshire	Washington, DC
1	3	5	7	8	10	10



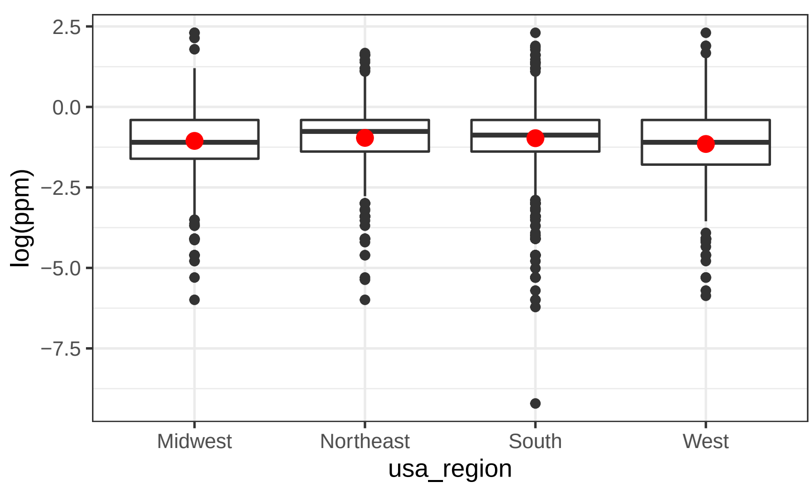


We observe that the within-state means for states with higher sample sizes in general adhere more closely to the grand mean. It is also evident that the log(ppm) distributions differ little as compared to the within-state variance. This is conducive to the borrowing of information between states.

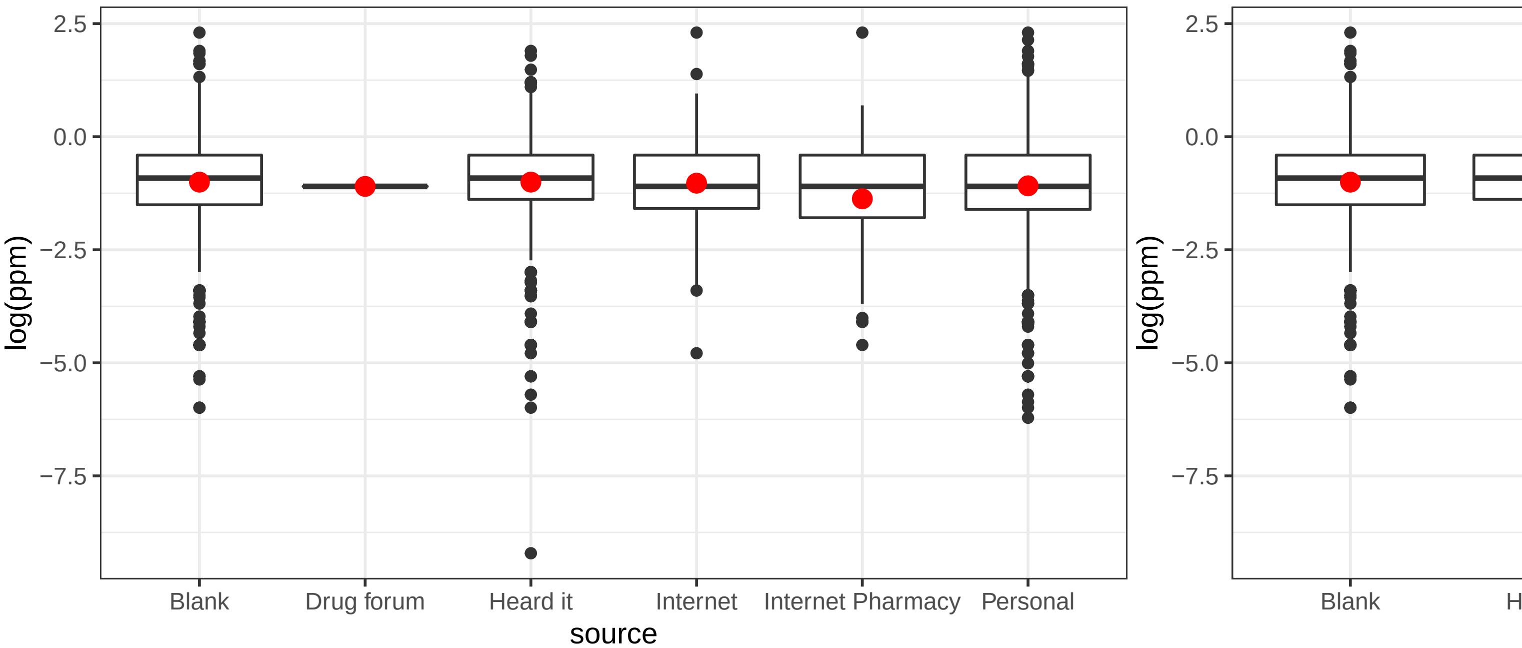
region vs. log(ppm)

We also have access to the broader region in which a purchase is made. This could be useful if we wanted to develop a simpler model that still captured variation by purchase location.

	usa_region	n	mean
1	Midwest	1167	−1.055
2	Northeast	673	−0.962
3	South	1949	−0.972
4	West	1770	−1.151

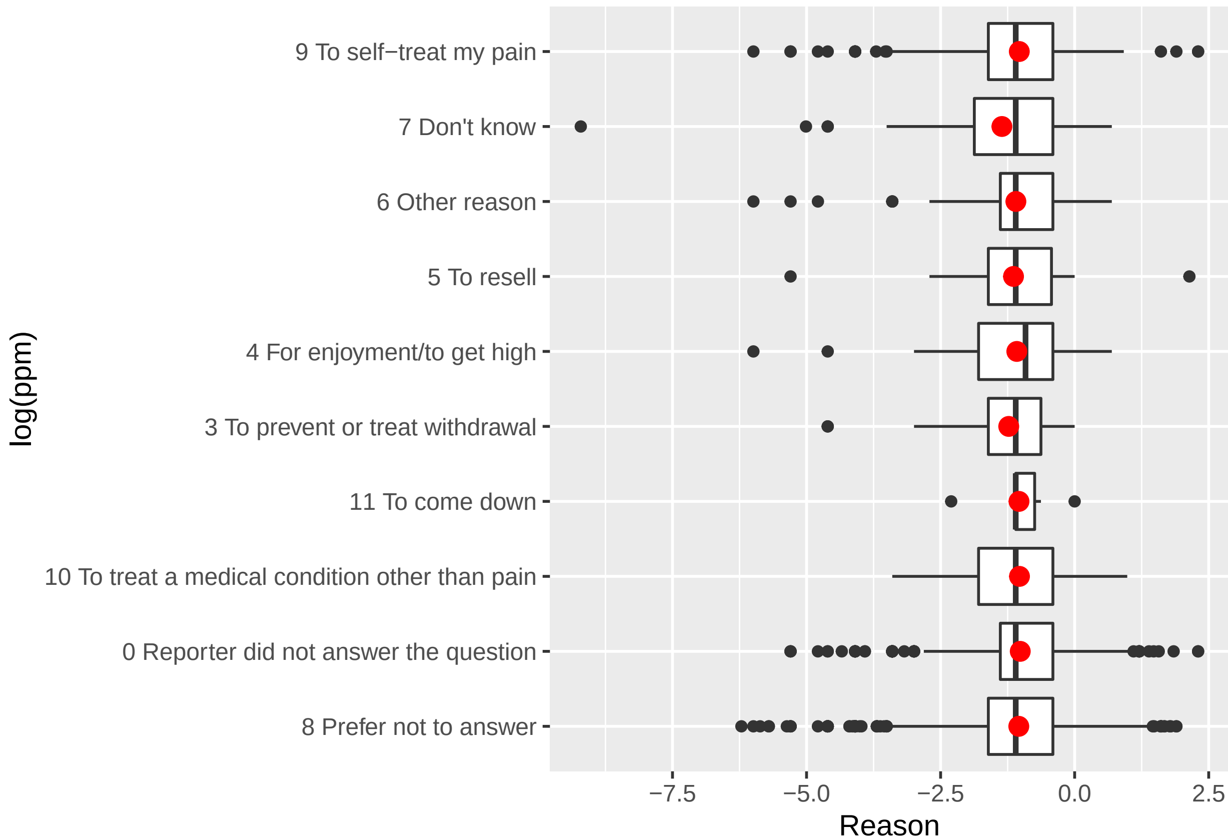


source vs. log(ppm)

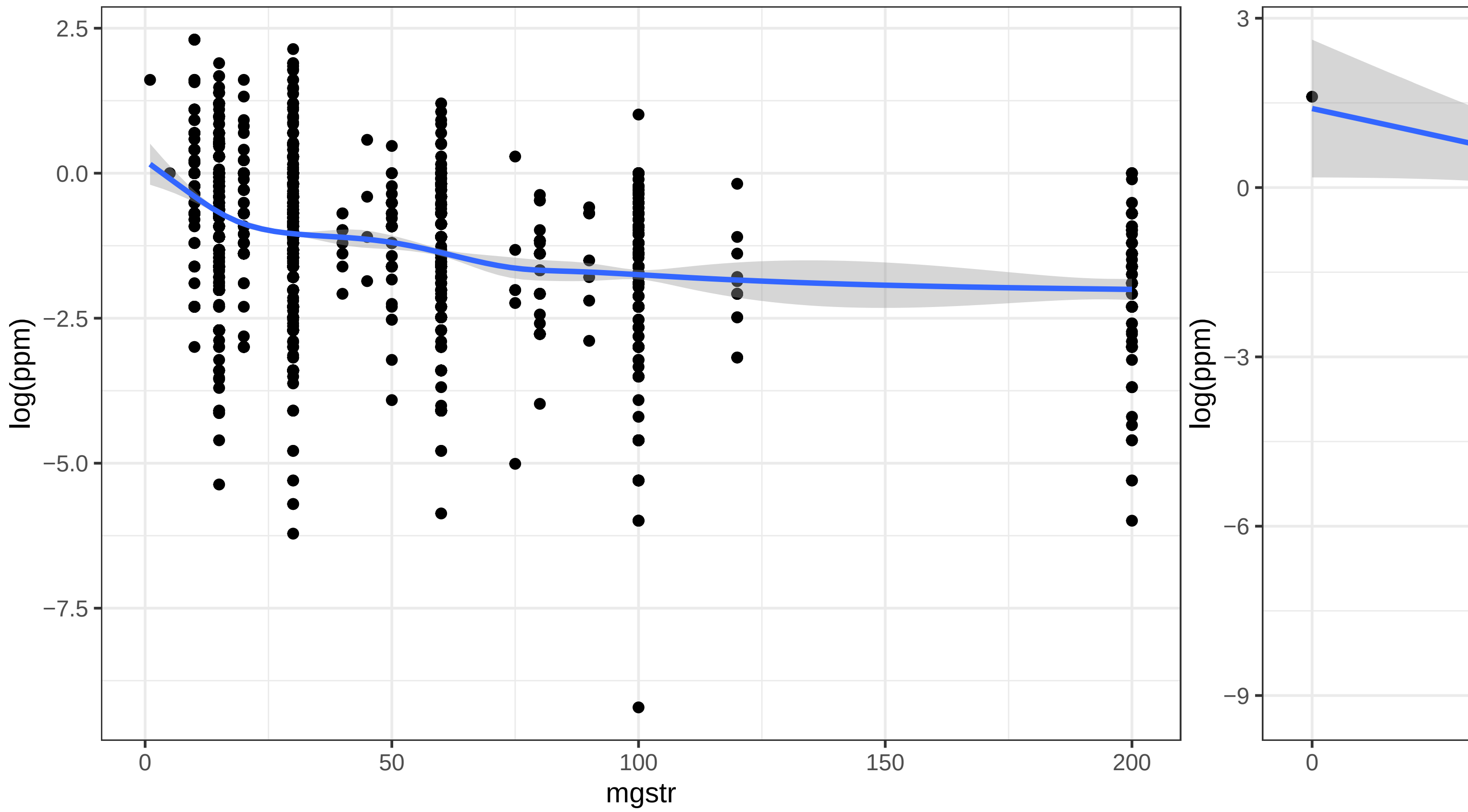



```
## # A tibble: 5 x 2
##   source                n
##   <chr>             <int>
## 1 Blank              1765
## 2 Heard it           1190
## 3 Internet            227
## 4 Internet Pharmacy    79
## 5 Personal           2307
```

Primary_Reason vs.log(ppm)



mgstr vs. log(ppm)



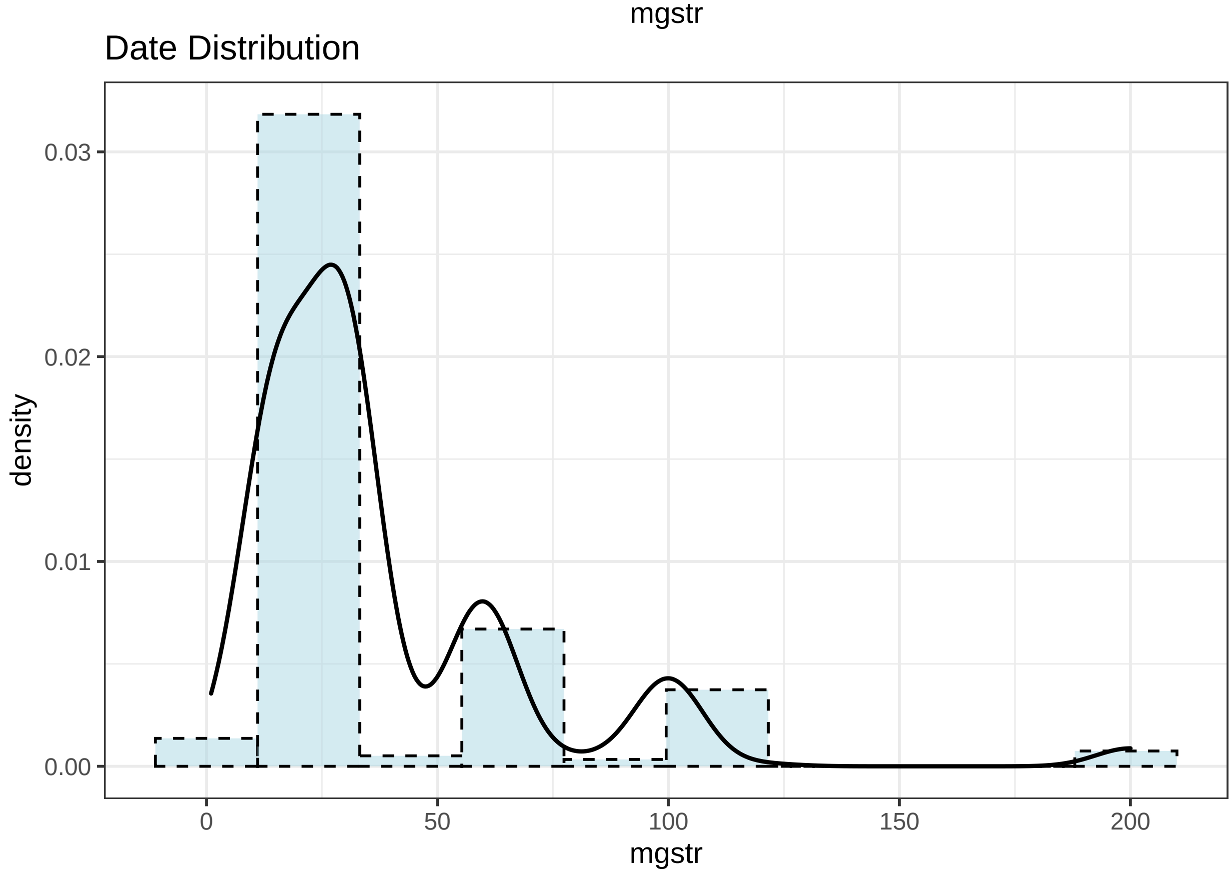
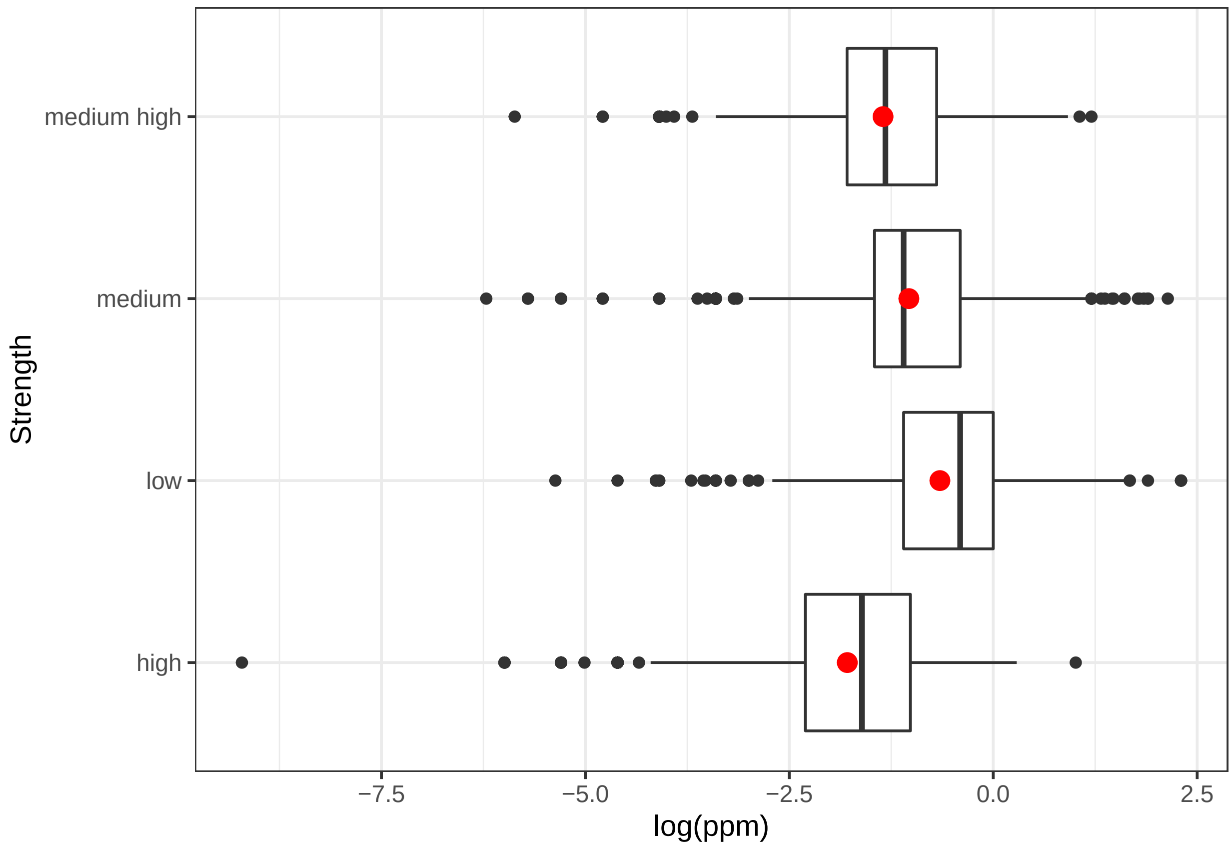


Table 2: Sample Size for mgstr Levels

1	5	10	15	20	30	40	45	50	60	75	80	90	100	120	200
1	1	166	1607	120	2192	8	4	51	819	6	34	7	446	14	92

	mgstr
25%	15
50%	30
75%	60



Model

choose grouping variable

Grouping	BIC
State	14712.37
City	14763.15
Region	14745.99

Choose `State` as our grouping variable

```
## Data: morph_data
## Models:
## model1: log(ppm) ~ (1 | city)
## model2: log(ppm) ~ (1 | city) + state
##        npar   AIC   BIC  logLik deviance  Chisq Df Pr(>Chisq)
```

```
## model1    3 15409 15428 -7701.3    15403
## model2   52 15356 15700 -7625.8    15252 150.92 49  2.585e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Data: morph_data
## Models:
## modela: log(ppm) ~ (1 | state)
## model3: log(ppm) ~ (1 | state) + usa_region
##        npar   AIC   BIC  logLik deviance  Chisq Df Pr(>Chisq)
## modela    3 15355 15375 -7674.4    15349
## model3    6 15354 15394 -7670.9    15342 7.1319  3    0.06781 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Data: morph_data
## Models:
## modela: log(ppm) ~ (1 | state)
## modelb: log(ppm) ~ mgstr2 + (1 | state)
##        npar   AIC   BIC  logLik deviance Chisq Df Pr(>Chisq)
## modela    3 15355 15375 -7674.4    15349
## modelb    6 14576 14616 -7281.9    14564   785  3  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Data: morph_data
## Models:
## modelb: log(ppm) ~ mgstr2 + (1 | state)
## modelc: log(ppm) ~ mgstr2 + bulk_purchase + (1 | state)
##        npar   AIC   BIC  logLik deviance  Chisq Df Pr(>Chisq)
## modelb    6 14576 14616 -7281.9    14564
## modelc    7 14564 14610 -7274.8    14550 14.247  1  0.0001603 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Data: morph_data
## Models:
## modelc: log(ppm) ~ mgstr2 + bulk_purchase + (1 | state)
## modeld: log(ppm) ~ year + mgstr2 + bulk_purchase + (1 | state)
##        npar   AIC   BIC  logLik deviance  Chisq Df Pr(>Chisq)
## modelc    7 14564 14610 -7274.8    14550
## modeld   16 14567 14673 -7267.6    14535 14.428  9     0.1079

## Data: morph_data
## Models:
## modelc: log(ppm) ~ mgstr2 + bulk_purchase + (1 | state)
## modele: log(ppm) ~ quarter + mgstr2 + bulk_purchase + (1 | state)
##        npar   AIC   BIC  logLik deviance Chisq Df Pr(>Chisq)
## modelc    7 14564 14610 -7274.8    14550
## modele   10 14560 14626 -7270.0    14540   9.7  3     0.0213 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Data: morph_data
## Models:
## modelc: log(ppm) ~ mgstr2 + bulk_purchase + (1 | state)
## modelf: log(ppm) ~ date_diff + mgstr2 + bulk_purchase + (1 | state)
##        npar   AIC   BIC  logLik deviance  Chisq Df Pr(>Chisq)
## modelc    7 14564 14610 -7274.8    14550
## modelf    8 14564 14617 -7273.9    14548 1.7621  1     0.1844

## Data: morph_data
## Models:
## modele: log(ppm) ~ quarter + mgstr2 + bulk_purchase + (1 | state)
## modelg: log(ppm) ~ quarter + source + mgstr2 + bulk_purchase + (1 | state)
##        npar   AIC   BIC  logLik deviance  Chisq Df Pr(>Chisq)
## modele   10 14560 14626 -7270.0    14540
## modelg   14 14549 14641 -7260.4    14521 19.237  4  0.0007061 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
modelg <- lmer(log(ppm) ~  quarter + source + mgstr2 + bulk_purchase + (1|state), data =
morph_data)
```

```
## Data: morph_data
## Models:
## modelg: log(ppm) ~ quarter + source + mgstr2 + bulk_purchase + (1 | state)
## modelgg: log(ppm) ~ quarter + source + mgstr2 + bulk_purchase + (1 | state) + quarter * bulk_purc
##         npar   AIC   BIC  logLik deviance  Chisq Df Pr(>Chisq)
## modelg    14 14549 14641 -7260.4    14521
## modelgg   17 14549 14662 -7257.6    14515 5.4786  3     0.1399

## Data: morph_data
## Models:
## modelg: log(ppm) ~ quarter + source + mgstr2 + bulk_purchase + (1 | state)
## modelggg: log(ppm) ~ quarter + source + mgstr2 + bulk_purchase + (1 | state) + quarter * mgstr2
##          npar   AIC   BIC  logLik deviance  Chisq Df Pr(>Chisq)
## modelg     14 14549 14641 -7260.4    14521
## modelggg   23 14558 14711 -7256.2    14512 8.3905  9     0.4953

## Data: morph_data
## Models:
## modelg: log(ppm) ~ quarter + source + mgstr2 + bulk_purchase + (1 | state)
## modelgggg: log(ppm) ~ quarter + source + mgstr2 + bulk_purchase + (1 | state) + bulk_purchase * m
##           npar   AIC   BIC  logLik deviance Chisq Df Pr(>Chisq)
## modelg      14 14549 14641 -7260.4    14521
## modelgggg   17 14549 14662 -7257.6    14515 5.402  3     0.1446

## Data: morph_data
## Models:
## modelg: log(ppm) ~ quarter + source + mgstr2 + bulk_purchase + (1 | state)
## modelggggg: log(ppm) ~ quarter + source + mgstr2 + bulk_purchase + (1 | state) + quarter * source
##            npar   AIC   BIC  logLik deviance  Chisq Df Pr(>Chisq)
```

```
## modelg       14 14549 14641 -7260.4    14521
## modelggggg   26 14542 14714 -7244.8    14490 31.137 12   0.001877 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Data: morph_data
## Models:
## modelg: log(ppm) ~ quarter + source + mgstr2 + bulk_purchase + (1 | state)
## modelgggggg: log(ppm) ~ quarter + source + mgstr2 + bulk_purchase + (1 | state) + bulk_purchase *
##             npar   AIC   BIC  logLik deviance  Chisq Df Pr(>Chisq)
## modelg        14 14549 14641 -7260.4    14521
## modelgggggg   18 14552 14672 -7258.2    14516 4.3824  4     0.3567

## Data: morph_data
## Models:
## modelg: log(ppm) ~ quarter + source + mgstr2 + bulk_purchase + (1 | state)
## modelggggggg: log(ppm) ~ quarter + source + mgstr2 + bulk_purchase + (1 | state) + source * mgstr
##              npar   AIC   BIC  logLik deviance  Chisq Df Pr(>Chisq)
## modelg         14 14549 14641 -7260.4    14521
## modelggggggg   26 14566 14738 -7257.1    14514 6.4869 12     0.8896
```

We now have `quarter`, `bulk_purchase`, `primary_reason` and `mgstr2` in our model, regarding `state` as the grouping variable.

Why/how did we choose the variables to put into the model? too many predictors?

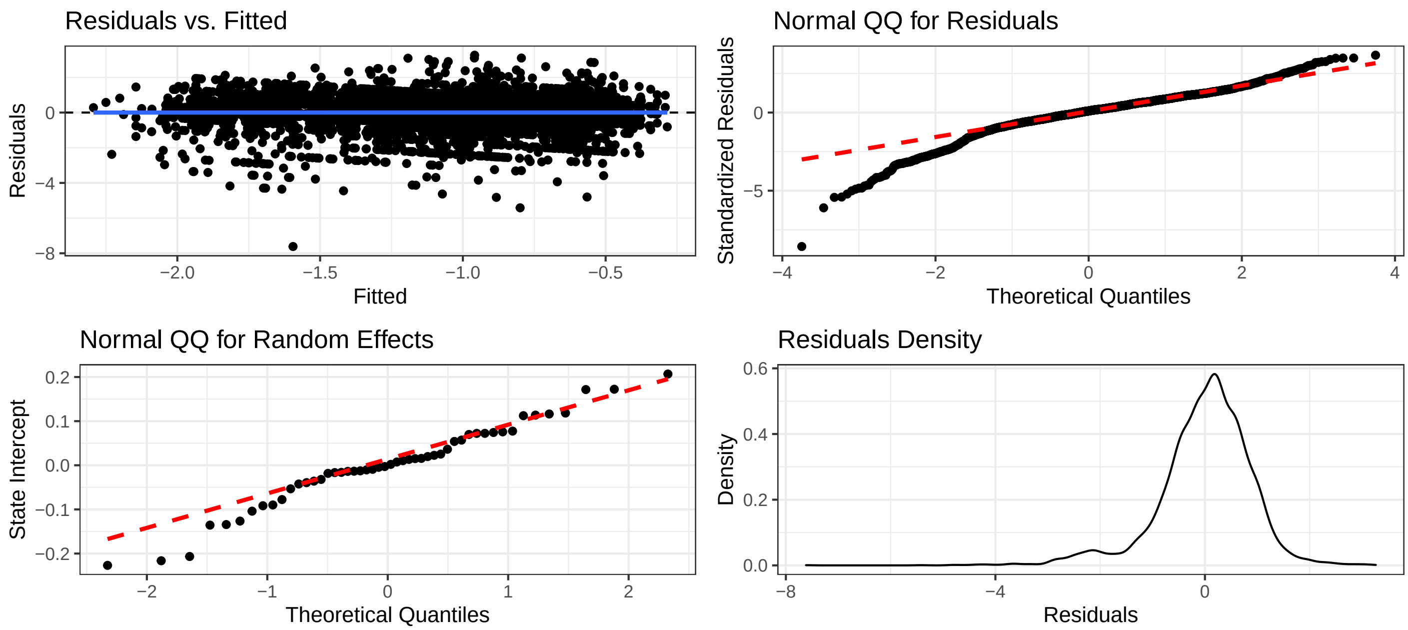
Unique values: state = 45 quarter = 4 bulk purchase = 2 primary reason = 10 mgstr = 14

only have 1831 observations

final model

	Estimate	exp(Estimate)	Std. Error	df	t value	Pr(>|t|)
(Intercept)	-1.7662127	0.1709793	0.0501878	695.5745	-35.1920702	0.0000000
quarter2	0.0851884	1.0889222	0.0323691	5540.6272	2.6317818	0.0085174
quarter3	0.0841888	1.0878343	0.0334164	5545.0209	2.5193848	0.0117839
quarter4	0.0781355	1.0812692	0.0343505	5541.3215	2.2746516	0.0229649
sourceHeard it	0.0570122	1.0586687	0.0337274	5546.2036	1.6903813	0.0910112
sourceInternet	-0.0044422	0.9955676	0.0630177	5545.5447	-0.0704915	0.9438050
sourceInternet Pharmacy	-0.3208306	0.7255462	0.1024391	5538.9433	-3.1319148	0.0017458
sourcePersonal	-0.0402830	0.9605176	0.0283620	5547.5892	-1.4203135	0.1555726
mgstr2low	1.1330601	3.1051441	0.0422532	5549.2455	26.8159608	0.0000000
mgstr2medium	0.7512885	2.1197295	0.0409249	5543.2562	18.3577321	0.0000000
mgstr2medium high	0.4326425	1.5413251	0.0472311	5538.5552	9.1601222	0.0000000
bulk_purchase1 Bulk purchase	-0.1116846	0.8943263	0.0298719	5547.8012	-3.7387862	0.0001868

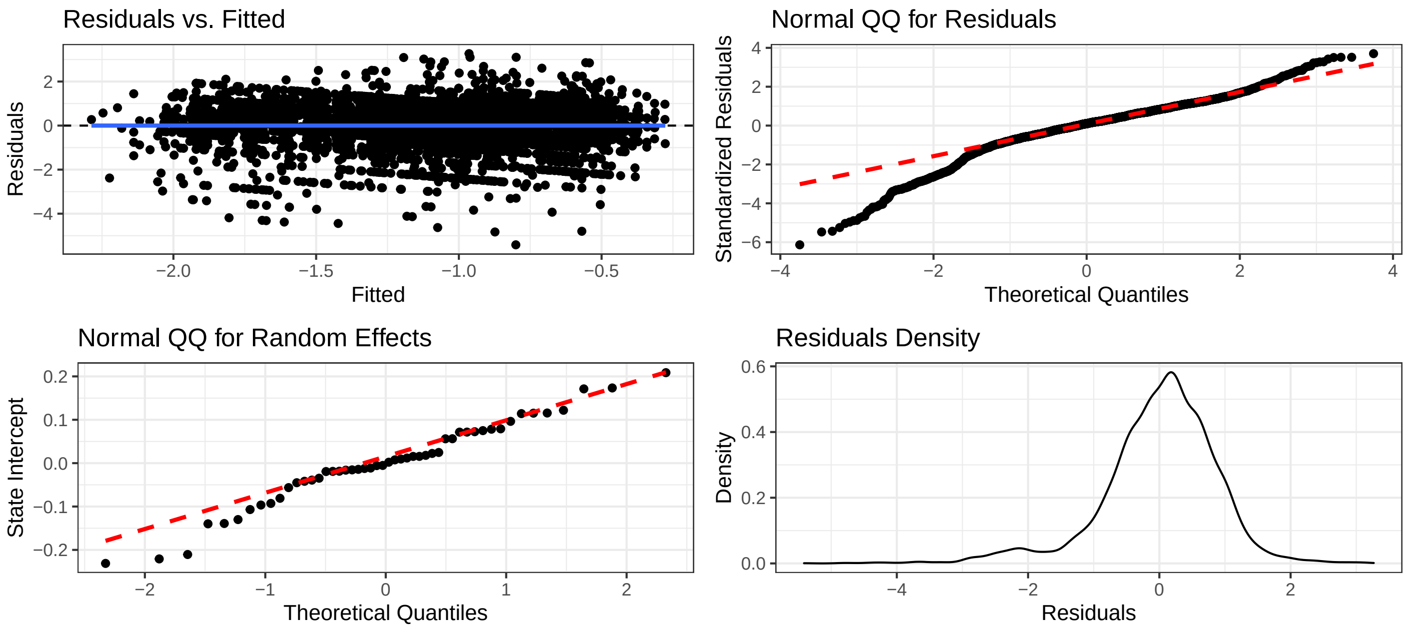
	Estimate
τ^2	0.0160650
σ^2	0.7893212



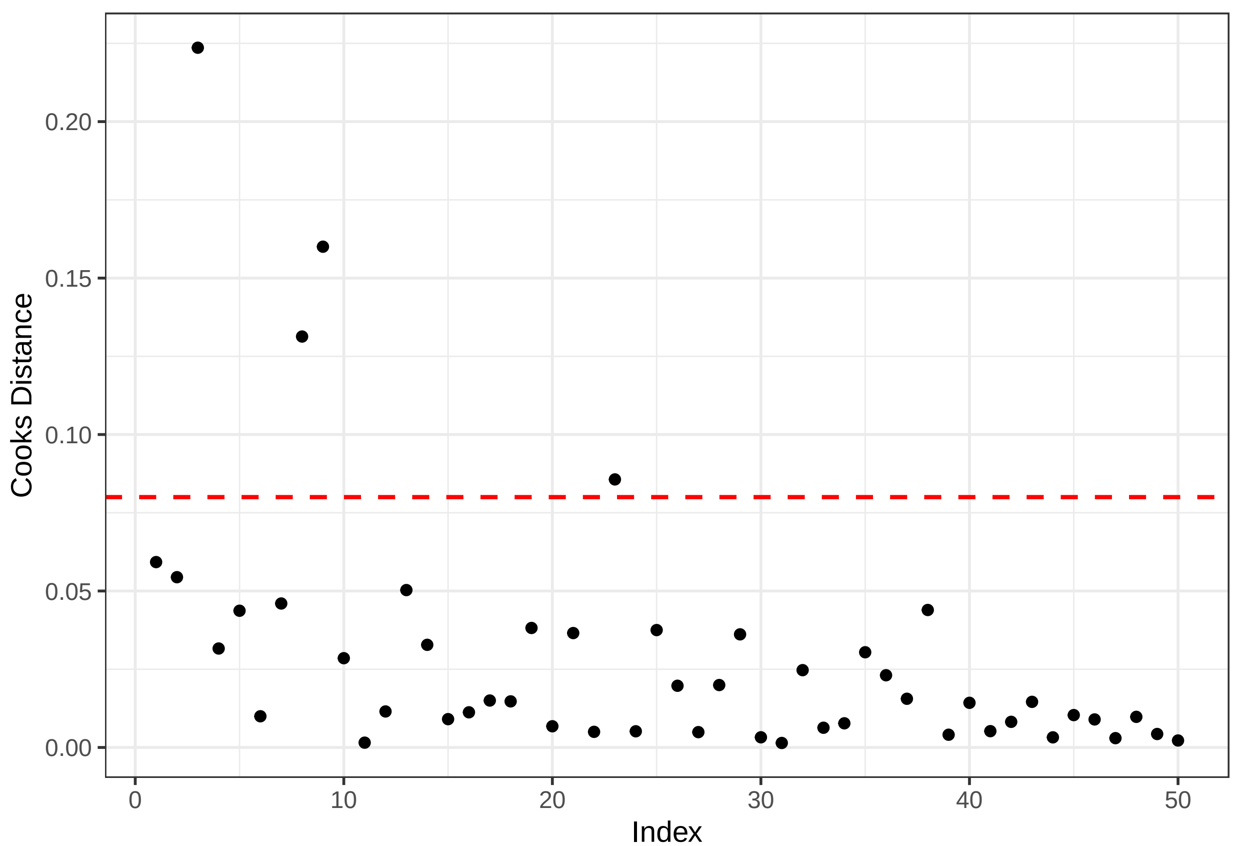
Remove the data point with the lowest residual.

	Estimate	exp(Estimate)	Std. Error	df	t value	Pr(>|t|)
(Intercept)	-1.7539294	0.1730925	0.0500673	668.4077	-35.0314697	0.0000000
quarter2	0.0853547	1.0891033	0.0321524	5539.1391	2.6546940	0.0079607
quarter3	0.0839552	1.0875801	0.0331930	5543.5182	2.5292998	0.0114565
quarter4	0.0843205	1.0879776	0.0341283	5539.6219	2.4706890	0.0135152
sourceHeard it	0.0632902	1.0653360	0.0335098	5544.4746	1.8887057	0.0589834
sourceInternet	-0.0041877	0.9958210	0.0625965	5543.9492	-0.0669005	0.9466633
sourceInternet Pharmacy	-0.3225594	0.7242929	0.1017530	5537.5340	-3.1700223	0.0015326
sourcePersonal	-0.0397818	0.9609991	0.0281727	5546.1674	-1.4120705	0.1579853
mgstr2low	1.1197175	3.0639884	0.0419994	5548.0026	26.6603389	0.0000000
mgstr2medium	0.7381139	2.0919862	0.0406796	5541.9141	18.1445782	0.0000000
mgstr2medium high	0.4197101	1.5215205	0.0469381	5536.8765	8.9417801	0.0000000
bulk_purchase1 Bulk purchase	-0.1141111	0.8921588	0.0296738	5546.2648	-3.8455220	0.0001216

	Estimate
τ^2	0.0166796
σ^2	0.7787108



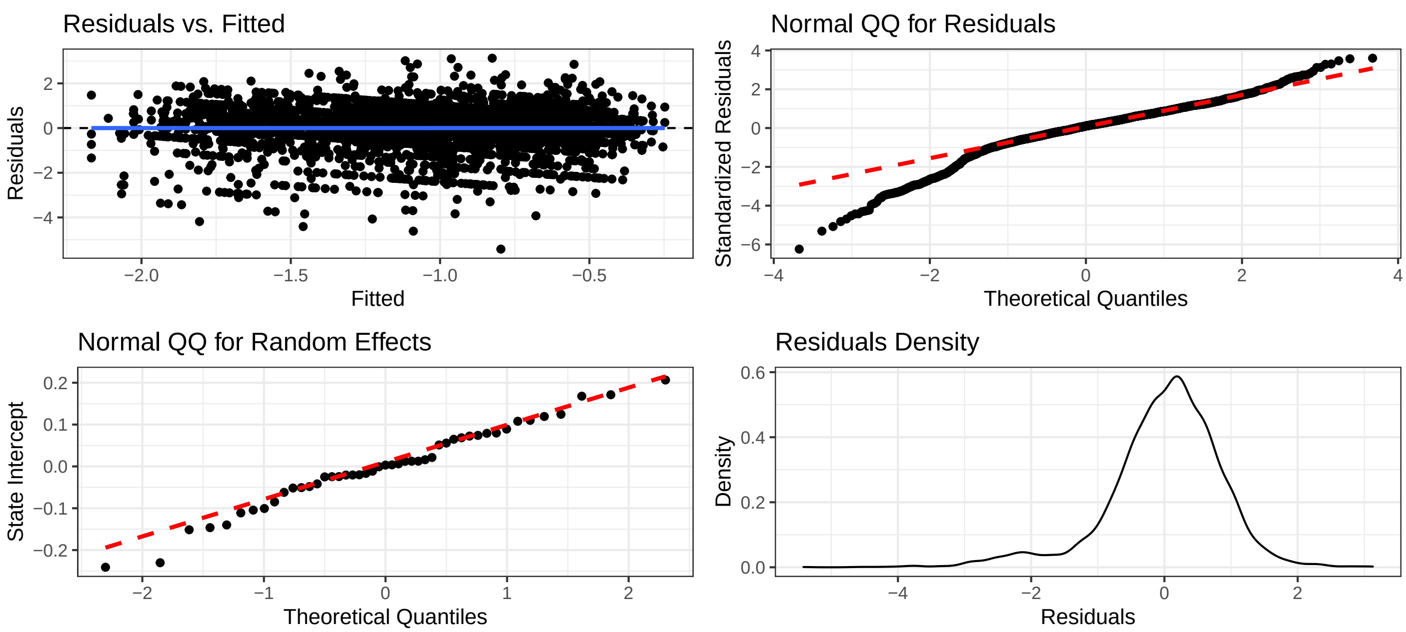
Influence



Cook's Distance
TRUE
TRUE
TRUE
TRUE

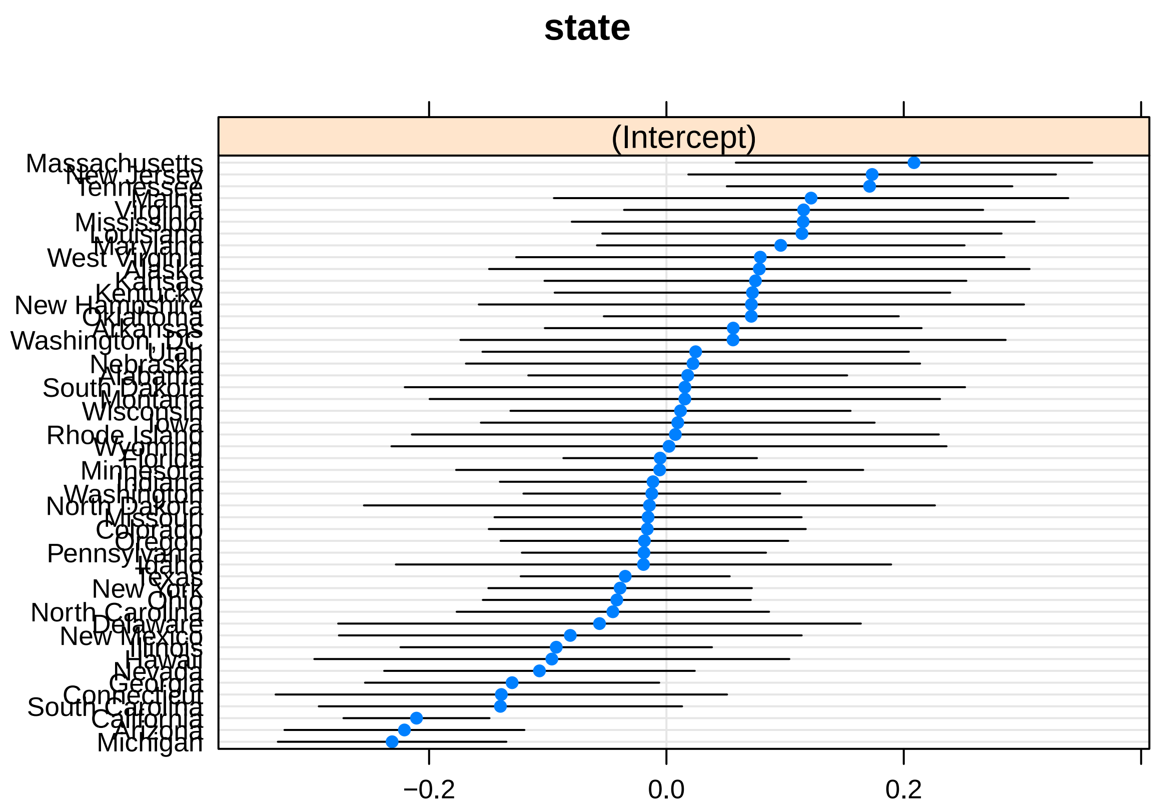
	Estimate	exp(Estimate)	Std. Error	df	t value	Pr(>|t|)
(Intercept)	-1.7259391	0.1780058	0.0572501	825.079	-30.1473548	0.0000000
quarter2	0.0984300	1.1034372	0.0366413	4159.584	2.6863155	0.0072532
quarter3	0.1003537	1.1055619	0.0375420	4163.137	2.6731026	0.0075446
quarter4	0.1046460	1.1103175	0.0387449	4159.537	2.7008985	0.0069433
sourceHeard it	0.0722513	1.0749254	0.0378101	4164.281	1.9108993	0.0560861
sourceInternet	-0.0239783	0.9763069	0.0706812	4163.606	-0.3392466	0.7344411
sourceInternet Pharmacy	-0.1923604	0.8250095	0.1188587	4158.969	-1.6183951	0.1056533
sourcePersonal	-0.0590381	0.9426708	0.0321003	4166.078	-1.8391761	0.0659604
mgstr2low	1.0951158	2.9895290	0.0492714	4165.124	22.2262036	0.0000000
mgstr2medium	0.7204225	2.0553014	0.0480325	4159.702	14.9986488	0.0000000
mgstr2medium high	0.3928732	1.4812306	0.0552091	4155.416	7.1160993	0.0000000
bulk_purchase1 Bulk purchase	-0.1416847	0.8678948	0.0339122	4166.169	-4.1779824	0.0000300

	Estimate
τ^2	0.0170776
σ^2	0.7541611



Does not change much, but the sample size decreases sharply -> decide not to remove these groups.

```
## $state
```


state
(Intercept)
Massachusetts
New Jersey
Tennessee
Maine
Virginia
Mississippi
Louisiana
Maryland
West Virginia
Alaska
Kansas
Kentucky
New Hampshire
Oklahoma
Arkansas
Washington DC
Utah
Nebraska
Alabama
South Dakota
Montana
Wisconsin
Iowa
Rhode Island
Wyoming
Florida
Minnesota
Indiana
Washington
North Dakota
Missouri
Colorado
Oregon
Pennsylvania
Idaho
Texas
New York
Ohio
North Carolina
Delaware
New Mexico
Illinois
Hawaii
Nevada
Georgia
Connecticut
South Carolina
California
Arizona
Michigan
−0.2
0.0
0.2