STA610 Case Study 1

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

13 October, 2021

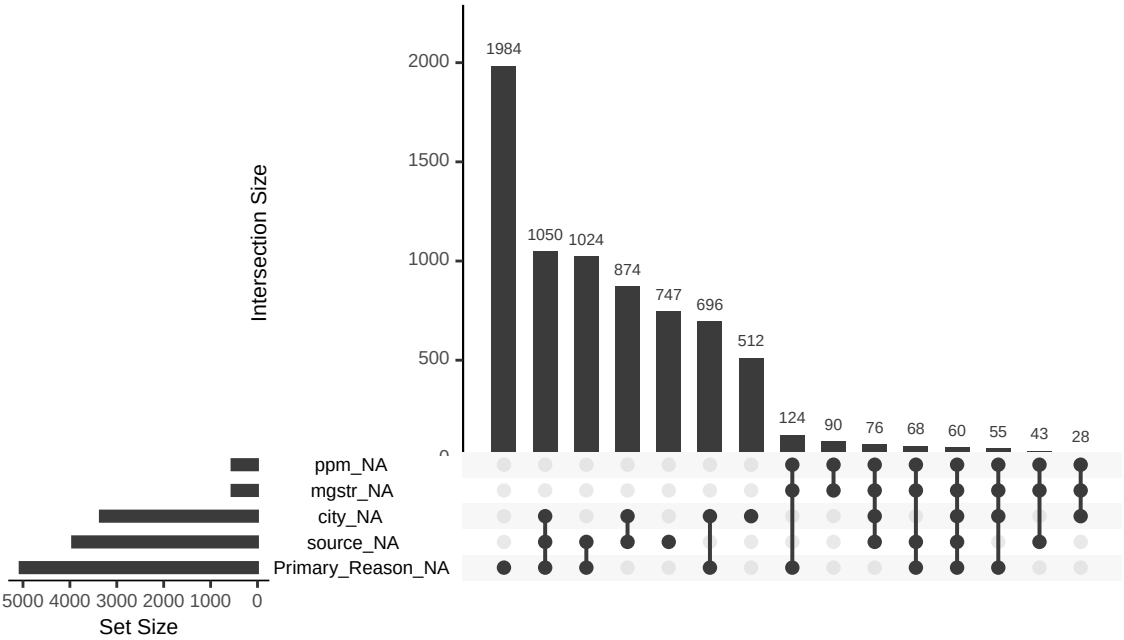
Introduction

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

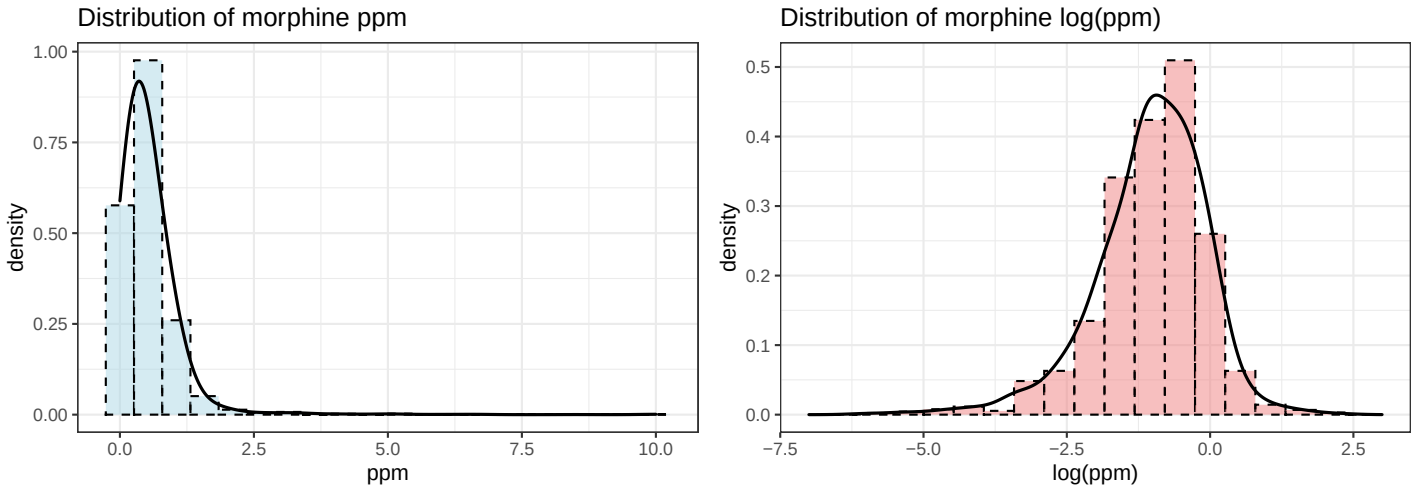
Missing Values



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 and "0 Reporter did not answer this question" in `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 removed other rows with missing values.

In addition, we think 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 observations we have is 5,582 now.

Response Variable: Price per milligram



Whether we fit a hierarchical model or linear regression, the response variable should be normally distributed. From the histogram on the left, the distribution of `ppm` is clearly right-skewed. Since `ppm` is strictly non-negative, a log transformation may be appropriate. The distribution of `log(ppm)` is given above, and appears closer to the desired normal.

Grouping Variable: city, state, and USA_region

Since we want to analyze the heterogeneity in pricing by location, we have three choices of grouping variable, `city`, `state`, and `USA_region`.

City There are 1642 unique `city` values, and many cities have small sample size (i.e. less than 5 observations). We decide not o use `city` as the grouping variable (see appendix).

```
## [1] 1642
```

```
## [1] 52
```

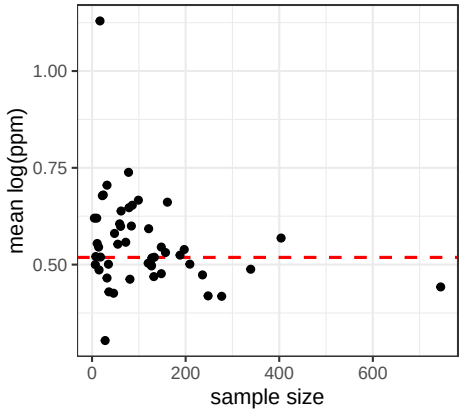
```
## [1] 5
```

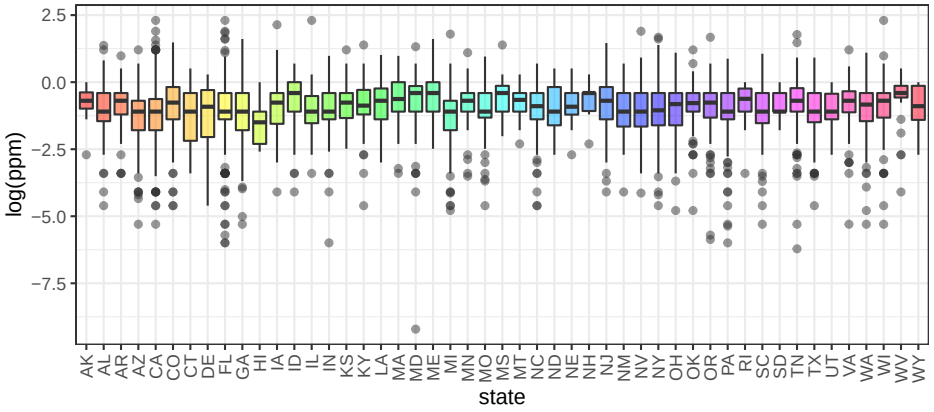
```
##  [1] Arizona        Alabama        Florida        New York       Texas
##  [6] Hawaii         Colorado       California     Pennsylvania   New Jersey
## [11] Washington, DC Oklahoma       Washington     Tennessee      Ohio
## [16] Oregon         Kentucky       North Carolina Nevada         Missouri
## [21] Illinois       Kansas         Michigan       Alaska         Georgia
## [26] Maryland       Minnesota      Arkansas       Wisconsin      Delaware
## [31] Montana        Iowa           Indiana        New Mexico     Massachusetts
## [36] Virginia       Louisiana      South Carolina Wyoming        Connecticut
## [41] Rhode Island   Maine          West Virginia  Utah           Nebraska
## [46] Vermont        Idaho          North Dakota   Mississippi    New Hampshire
## [51] South Dakota   Puerto Rico
## 52 Levels: Alabama Alaska Arizona Arkansas California Colorado ... Wyoming
```

State As for state, we examined the sample sizes in each group and decide to out filter Puerto Rico and Vermont, because they have less than 5 observations.

Table 1: 5 States with Smallest Sample Size

Puerto Rico	Vermont	North Dakota	South Dakota	Wyoming
1	3	5	7	8

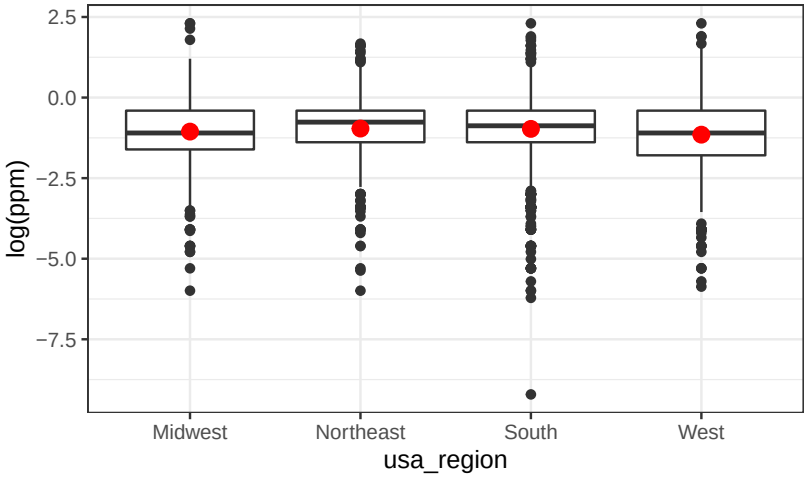




Then we inspect the state-level differences closer by plotting the group-level means agains the sample sizes. We observed that the within-state means for states with smaller sample sizes vary a lot, while the within-state means for states with higher sample sizes in general adhere more closely to the grand mean. This is conducive to the borrowing of information between states with a hierarchical model. From the above boxplot of `log(ppm)` against `state`, it is also evident that the `log(ppm)` distributions differ across states. This indicates the potential state-level differences in drug prices. Therefore, we decide to use state as our grouping variable at this stage.

Region From the boxplot we still see the `log(ppm)` distributions differ slightly across regions, though not that much as across states. We may also consider using region as the grouping variable.

	usa_region	n	mean
1	Midwest	1168	−1.056
2	Northeast	674	−0.962
3	South	1953	−0.972
4	West	1773	−1.151

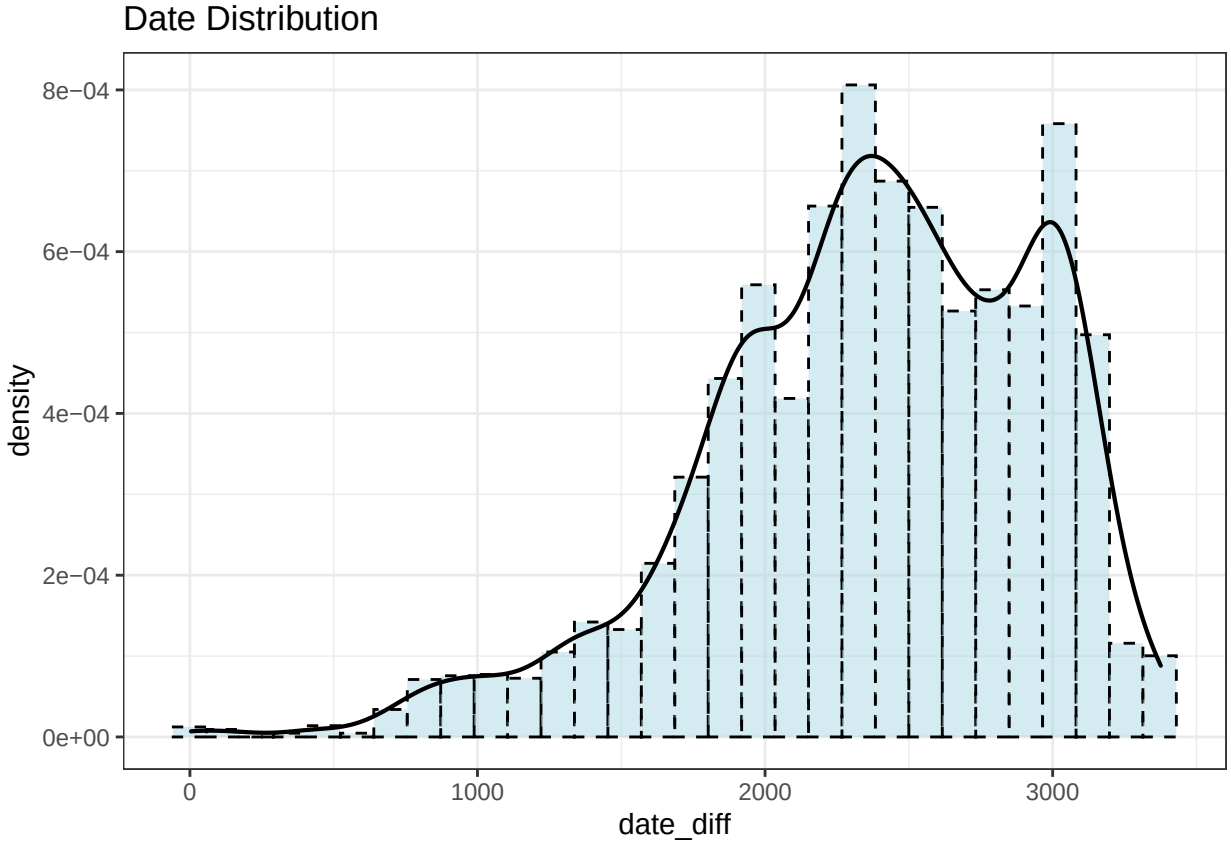


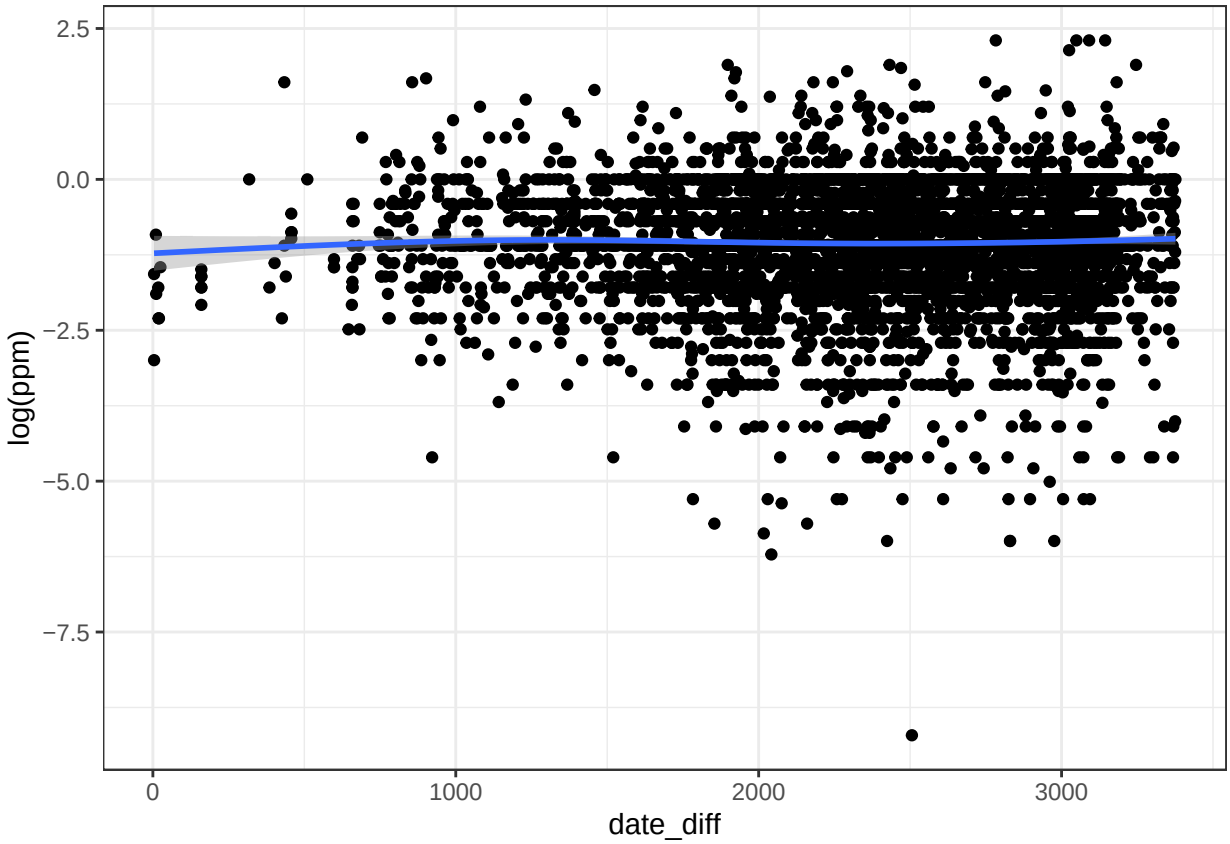
Date (price_date)

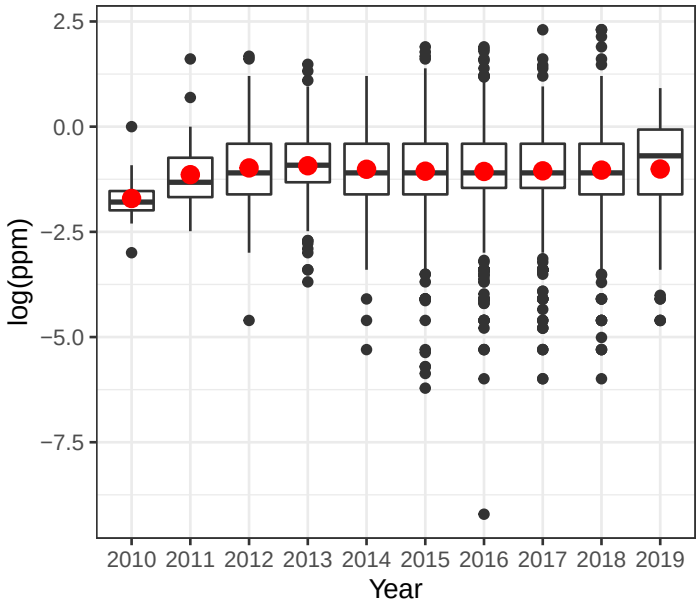
As for the `price_date`, we noticed some observations are prior to the establishment of StreetRx, which might be wrong inputs. We dropped the observations prior to 2010. For the rest observations, we came up with two ways of data cleaning on the date variable. The first choice is to choose a starting date and convert the feature as the date differences (`date_diff`) from that starting date. The second choice is to split this date variable

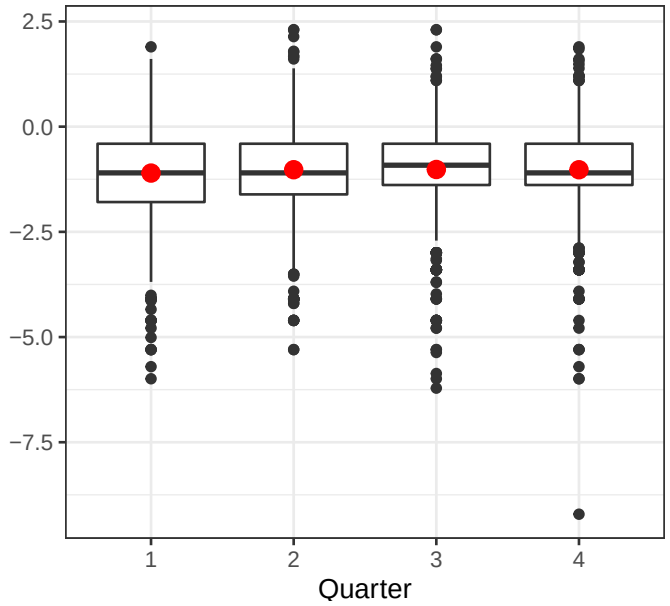
into two components, `year` and `quarter`, so that we can explore the trend of unit drug price over time and the seasonality.

Our visualizations suggested there is no clear trend that the log value of per milligram price of morphine varies along with `date_diff`. However, for different `year` and `quarter`, the `log(ppm)` value varies a little bit (see appendix).







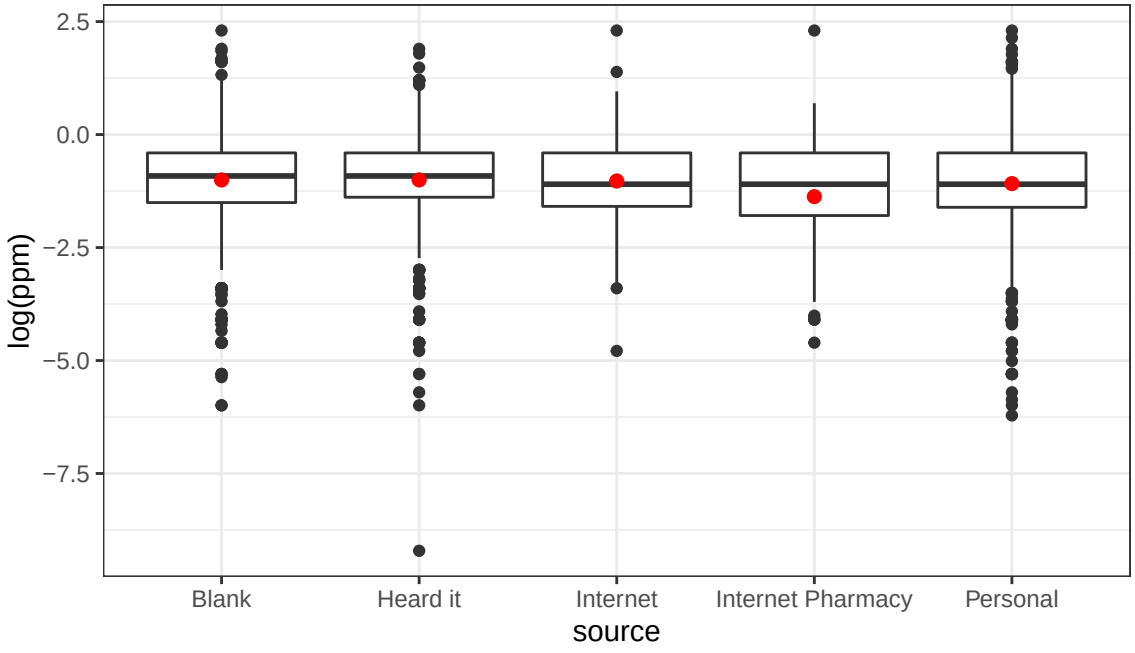


Bulk_purchase

There is no need to conduct any data cleaning on `bulk_purchase`. And from the boxplot (see appendix), there is a slight trend that the drug price may be lower if purchased in bulk. Therefore, `bulk_purchase` might be a potential predictor.

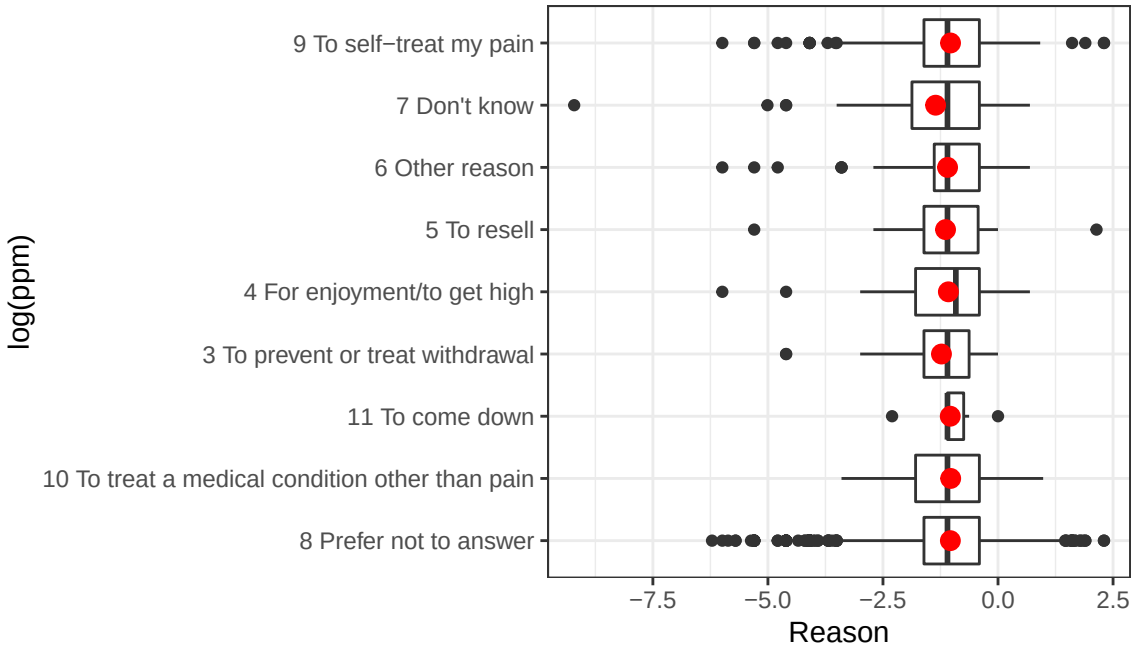
Source

We have recoded the missing value as "Blank" and the name of websites as "Internet". And we dropped the only observation whose `source` is "Drug Forum". From the boxplot, we see the `log(ppm)` value varies among different sources (see appendix).



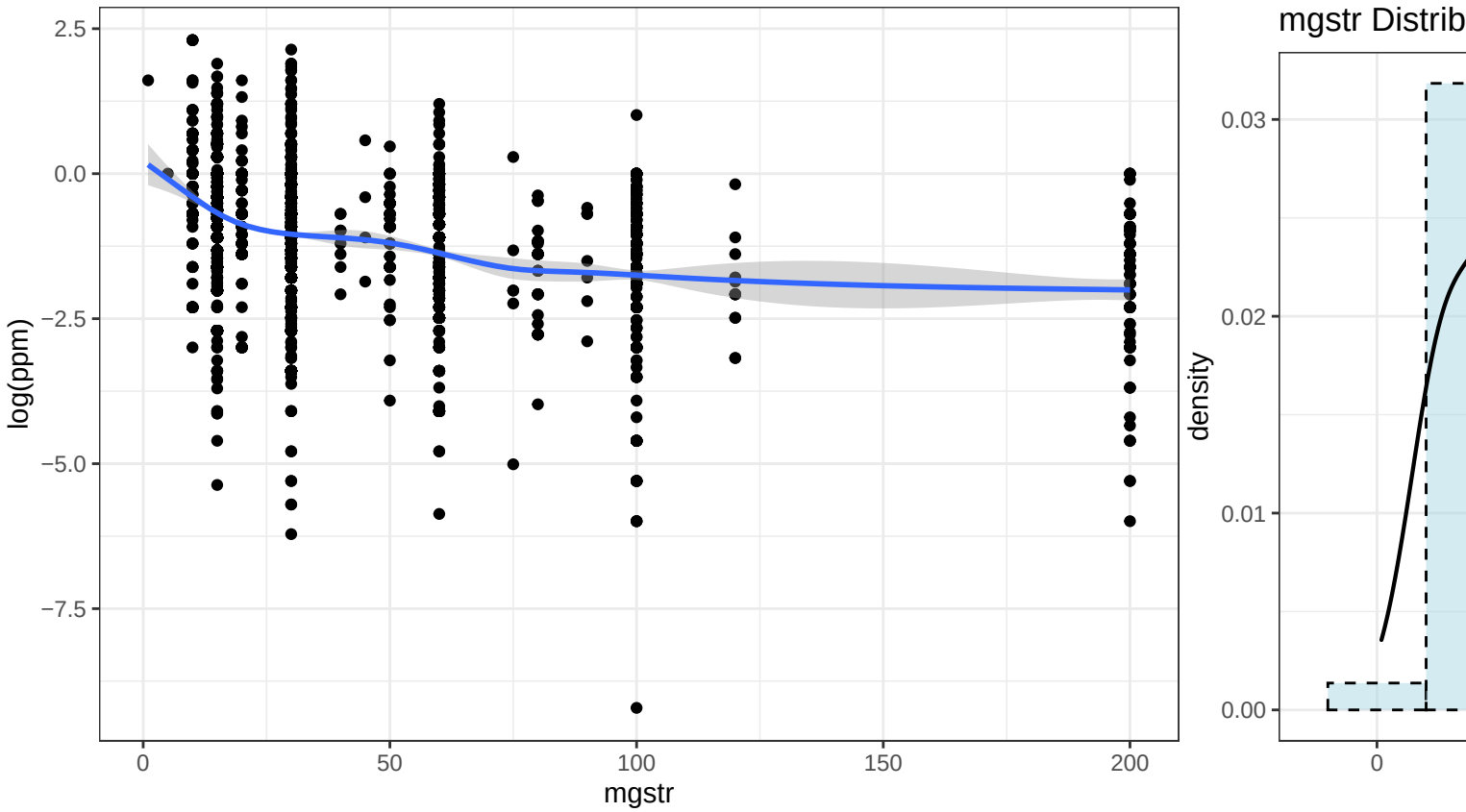
Primary Reason

For `primary_reason`, we have converted the empty cells and "0 Reporter did not answer this question" to "8 Prefer not to answer". The `log(ppm)` value varies a lot among different reasons for purchasing the morphine (see appendix).



Dosage Strength (mgstr)

From the scatter plot of `log(ppm)` against `mgstr`, there is a slight trend that the larger the dosage strength, the smaller the per milligram price. We have also noticed that `mgstr` only takes 16 discrete values. Therefore, we consider to label it into 4 levels ("low", "medium", "medium high", and "high") based on the 0.25, 0.5, and 0.75 quantiles of `mgstr`.



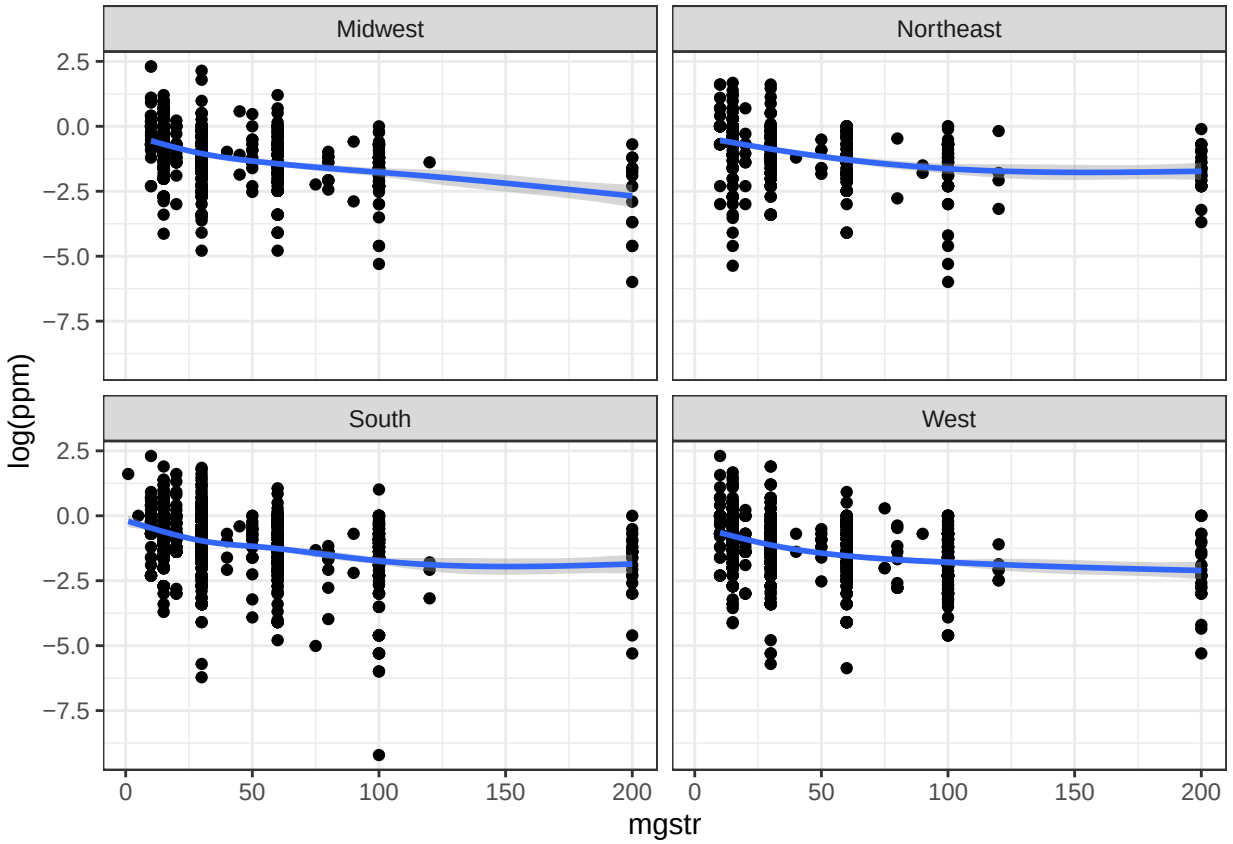
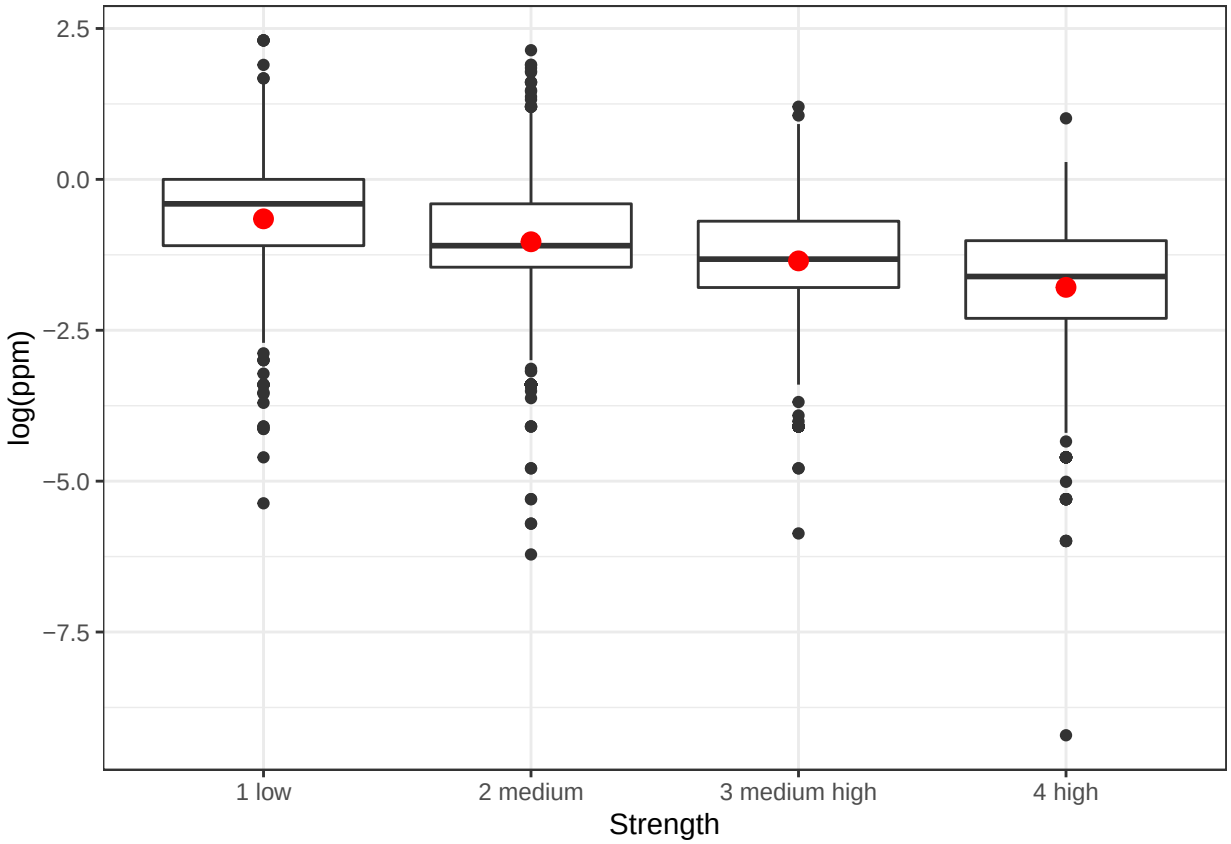


Table 2: Sample Size for mgstr Levels

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



From the boxplot, we see a more clear trend that the `log(ppm)` values decrease as the dosage strength increase.

Model

choose grouping variable

Grouping	BIC
State	14704.44
City	14755.31
Region	14738.32

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
```

```
## Data: morph_data
## Models:
## modelg: log(ppm) ~ quarter + source + mgstr2 + bulk_purchase + (1 | state)
## modelh: log(ppm) ~ quarter + primary_reason + mgstr2 + bulk_purchase + (1 | state)
##        npar   AIC   BIC  logLik deviance Chisq Df Pr(>Chisq)
## modelg   14 14549 14641 -7260.4    14521
## modelh   18 14562 14681 -7263.1    14526     0  4          1
```

```
## Backward reduced random-effect table:
##
##             Eliminated npar  logLik   AIC    LRT Df Pr(>Chisq)
## <none>                   32 -7246.4 14557
## (1 | state)          0   31 -7279.8 14622 66.805  1  2.997e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Backward reduced fixed-effect table:
## Degrees of freedom method: Satterthwaite
##
##                Eliminated Sum Sq Mean Sq NumDF  DenDF  F value    Pr(>F)
## date_diff               1   0.01   0.005     1 5559.0   0.0067 0.9345712
## year                    2  10.55   1.172     9 5556.2   1.4951 0.1433668
## primary_reason          3  11.40   1.424     8 5550.9   1.8132 0.0697607 .
## quarter                 0   7.60   2.533     3 5553.1   3.2157 0.0218925 *
## mgstr2                  0 676.70 225.567     3 5553.1 286.3423 < 2.2e-16 ***
## bulk_purchase           0  11.03  11.032     1 5559.5  14.0040 0.0001843 ***
## source                  0  15.18   3.796     4 5555.1   4.8189 0.0007032 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Model found:
## log(ppm) ~ quarter + mgstr2 + bulk_purchase + source + (1 | state)
```

```
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
```

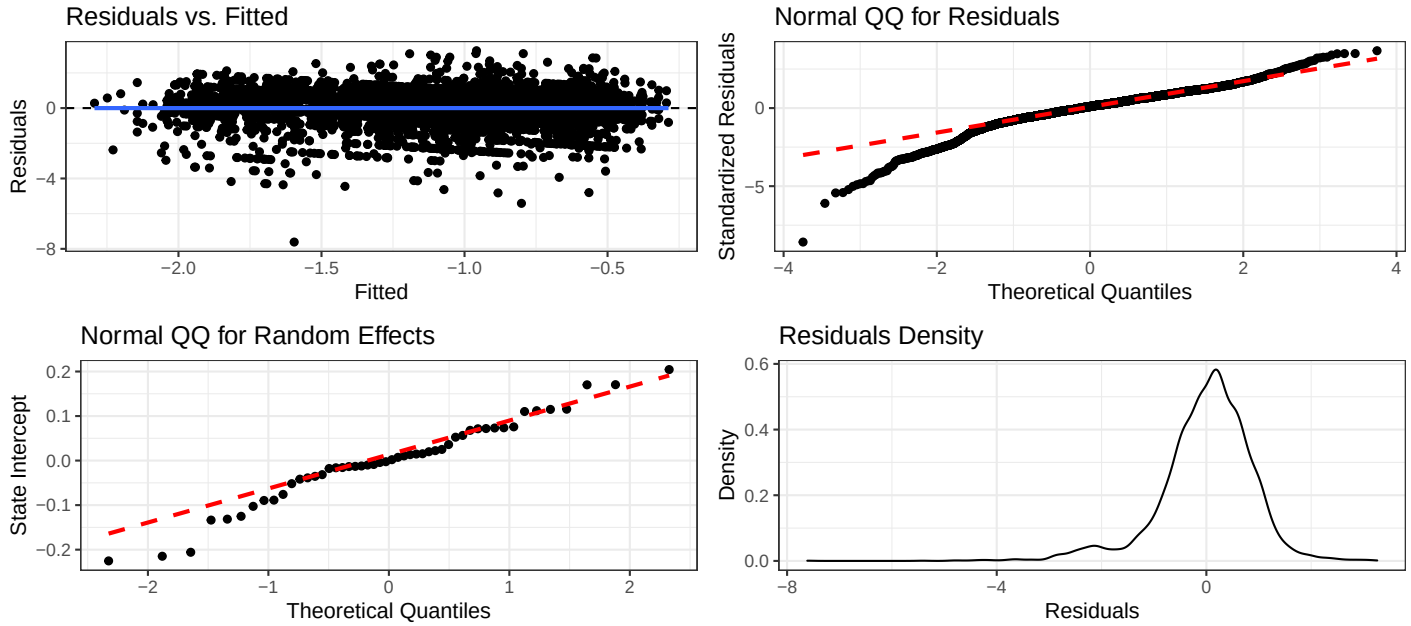
```
## Data: morph_data
## Models:
## modelg: log(ppm) ~ quarter + source + mgstr2 + bulk_purchase + (1 | state)
## modelggggggg: log(ppm) ~ quarter + source + mgstr2 + bulk_purchase + (1 | state) + source * mgstr
##              npar   AIC   BIC  logLik deviance  Chisq Df Pr(>Chisq)
## modelg      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

	Estimate	exp(Estimate)	Std. Error	df	t value	Pr(>|t|)
(Intercept)	-0.6335762	0.5306905	0.0393663	307.3301	-16.0943785	0.0000000
quarter2	0.0852053	1.0889406	0.0323359	5552.0953	2.6350079	0.0084369
quarter3	0.0842916	1.0879461	0.0333819	5556.5889	2.5250728	0.0115950
quarter4	0.0782277	1.0813689	0.0343152	5552.8820	2.2796776	0.0226646
sourceHeard it	0.0570002	1.0586561	0.0336924	5557.8673	1.6917812	0.0907438
sourceInternet	-0.0043480	0.9956614	0.0629524	5557.1909	-0.0690682	0.9449378
sourceInternet Pharmacy	-0.3209736	0.7254424	0.1023343	5550.3544	-3.1365197	0.0017186
sourcePersonal	-0.0403433	0.9604596	0.0283325	5559.1807	-1.4239227	0.1545250
mgstr22 medium	-0.3818074	0.6826265	0.0281207	5550.8745	-13.5774692	0.0000000
mgstr23 medium high	-0.7004080	0.4963827	0.0367354	5556.3478	-19.0662952	0.0000000
mgstr24 high	-1.1330801	0.3220398	0.0422091	5560.9614	-26.8444439	0.0000000
bulk_purchase1 Bulk purchase	-0.1116701	0.8943392	0.0298408	5559.5373	-3.7421928	0.0001843

	Estimate
τ^2	0.0154824
σ^2	0.7877548

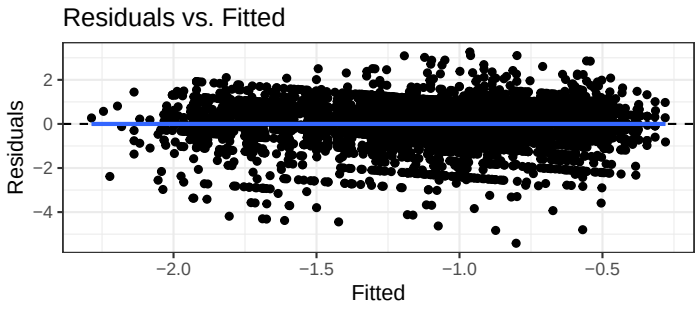
final model

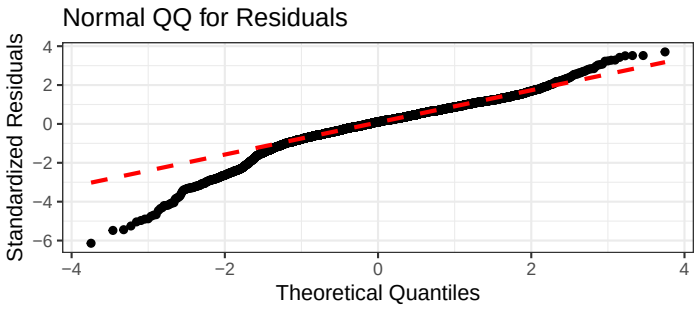


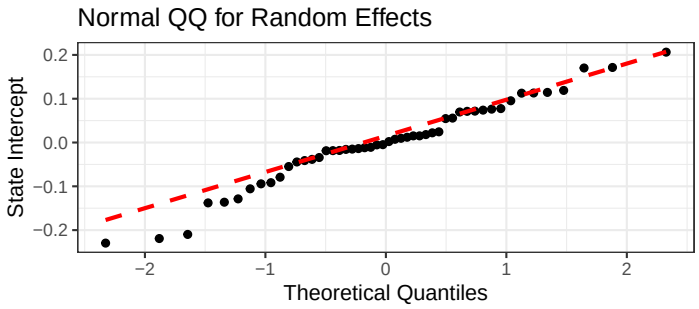
Remove the data point with the lowest residual.

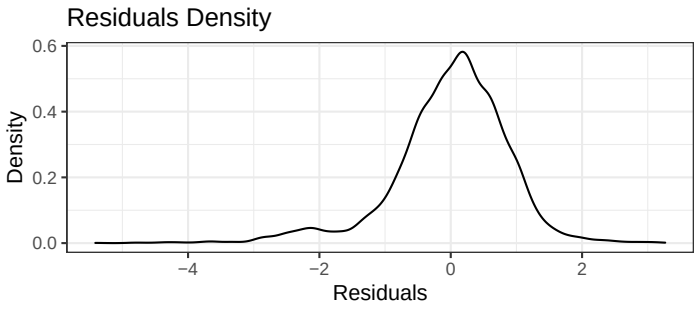
	Estimate	exp(Estimate)	Std. Error	df	t value	Pr(>|t|)
(Intercept)	-0.6346251	0.5301342	0.0393469	296.1819	-16.128986	0.0000000
quarter2	0.0853664	1.0891161	0.0321194	5550.5948	2.657785	0.0078881
quarter3	0.0840533	1.0876869	0.0331588	5555.0726	2.534876	0.0112759
quarter4	0.0844116	1.0880767	0.0340933	5551.1650	2.475899	0.0133197
sourceHeard it	0.0632824	1.0653277	0.0334751	5556.1197	1.890431	0.0587524
sourceInternet	-0.0041012	0.9959072	0.0625318	5555.5822	-0.065586	0.9477098
sourceInternet Pharmacy	-0.3226665	0.7242154	0.1016490	5548.9347	-3.174322	0.0015101
sourcePersonal	-0.0398430	0.9609403	0.0281434	5557.7473	-1.415712	0.1569159
mgstr22 medium	-0.3816419	0.6827395	0.0279323	5549.1951	-13.663103	0.0000000
mgstr23 medium high	-0.7000076	0.4965815	0.0364899	5554.7006	-19.183610	0.0000000
mgstr24 high	-1.1197461	0.3263627	0.0419556	5559.7107	-26.688860	0.0000000
bulk_purchase1 Bulk purchase	-0.1140900	0.8921776	0.0296429	5557.9880	-3.848810	0.0001200

	Estimate
τ^2	0.0160842
σ^2	0.7771652





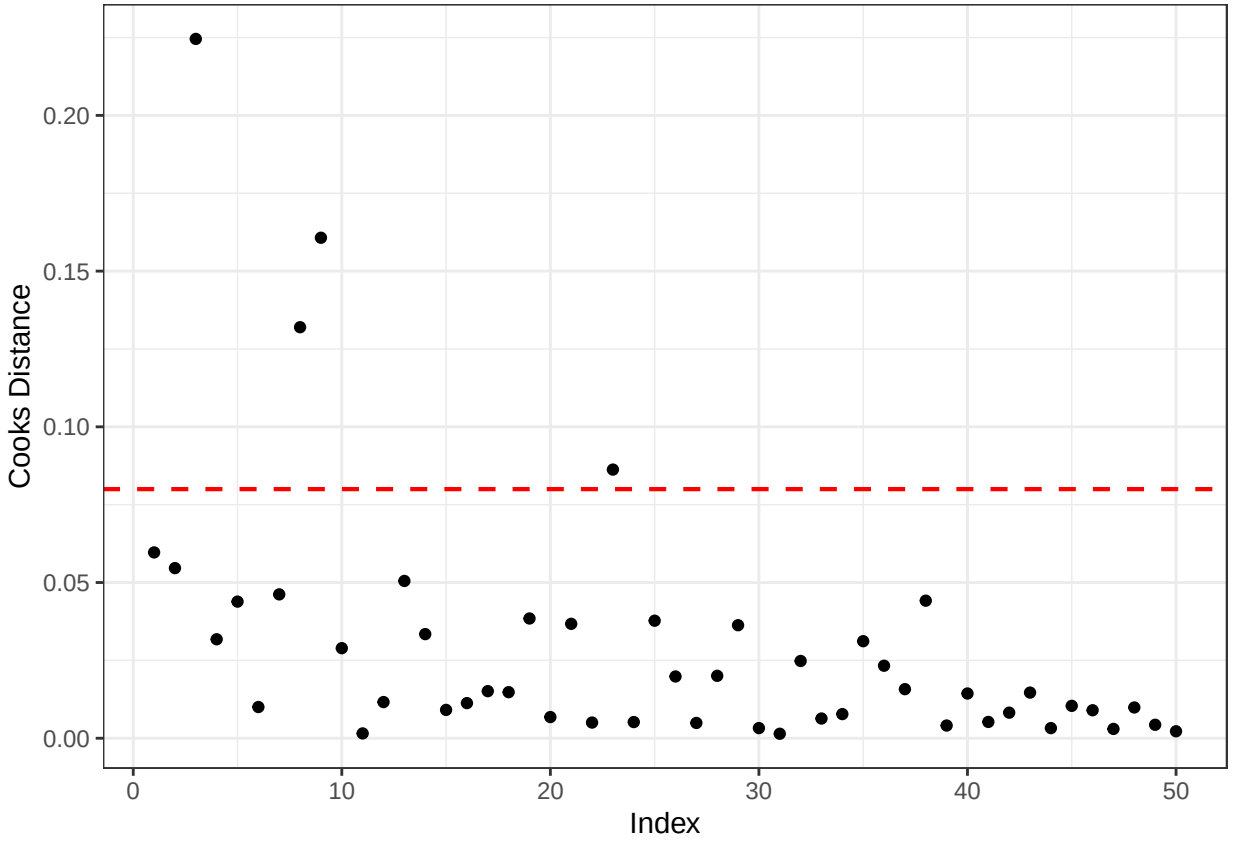




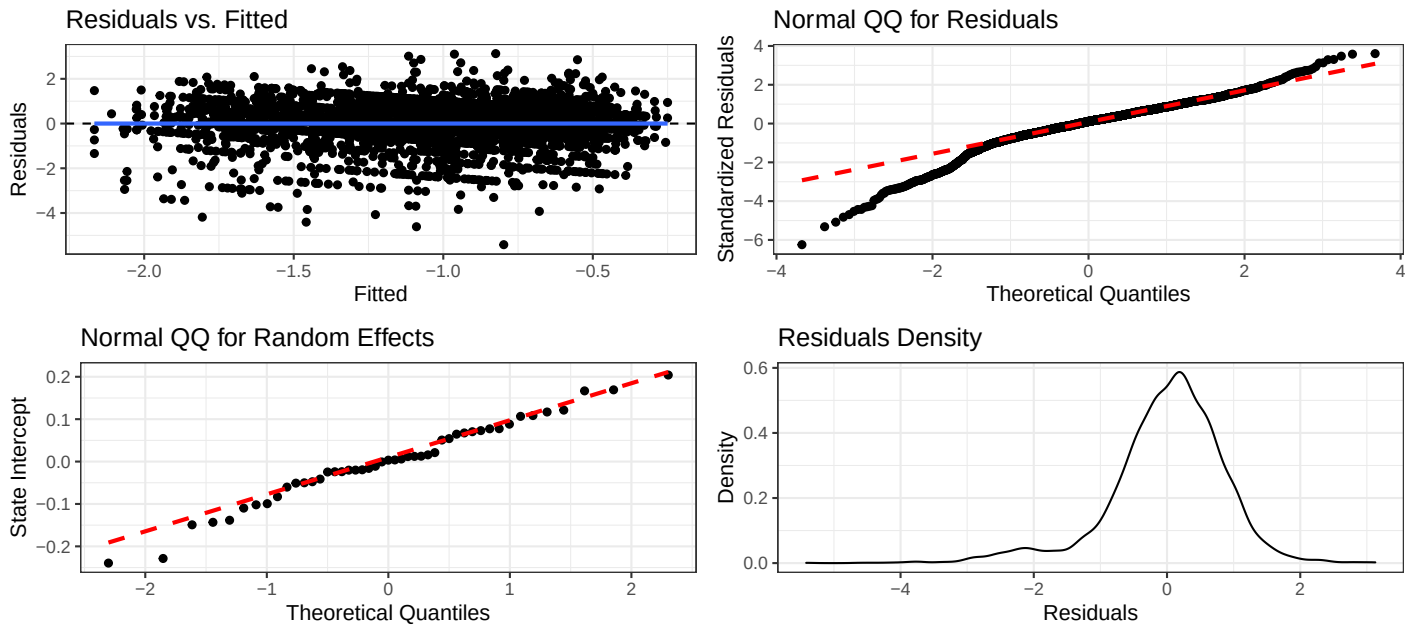
Cook's Distance
TRUE
TRUE
TRUE
TRUE

	Estimate	exp(Estimate)	Std. Error	df	t value	Pr(>|t|)
(Intercept)	-0.6312230	0.5319408	0.0432226	325.2158	-14.6040189	0.0000000
quarter2	0.0984455	1.1034543	0.0365911	4171.0812	2.6904214	0.0071646
quarter3	0.1004865	1.1057087	0.0374903	4174.7206	2.6803348	0.0073837
quarter4	0.1047844	1.1104711	0.0386918	4171.1112	2.7081769	0.0067930
sourceHeard it	0.0722337	1.0749065	0.0377578	4175.9755	1.9130788	0.0558063
sourceInternet	-0.0238741	0.9764086	0.0705837	4175.2761	-0.3382382	0.7352007
sourceInternet Pharmacy	-0.1924988	0.8248953	0.1186962	4170.4188	-1.6217767	0.1049268
sourcePersonal	-0.0591093	0.9426038	0.0320558	4177.7104	-1.8439498	0.0652612
mgstr22 medium	-0.3747454	0.6874643	0.0313241	4167.1767	-11.9634954	0.0000000
mgstr23 medium high	-0.7022359	0.4954763	0.0414718	4172.4074	-16.9328535	0.0000000
mgstr24 high	-1.0951019	0.3345055	0.0492032	4176.7239	-22.2567267	0.0000000
bulk_purchase1 Bulk purchase	-0.1416385	0.8679350	0.0338652	4177.9448	-4.1824239	0.0000294

Influence

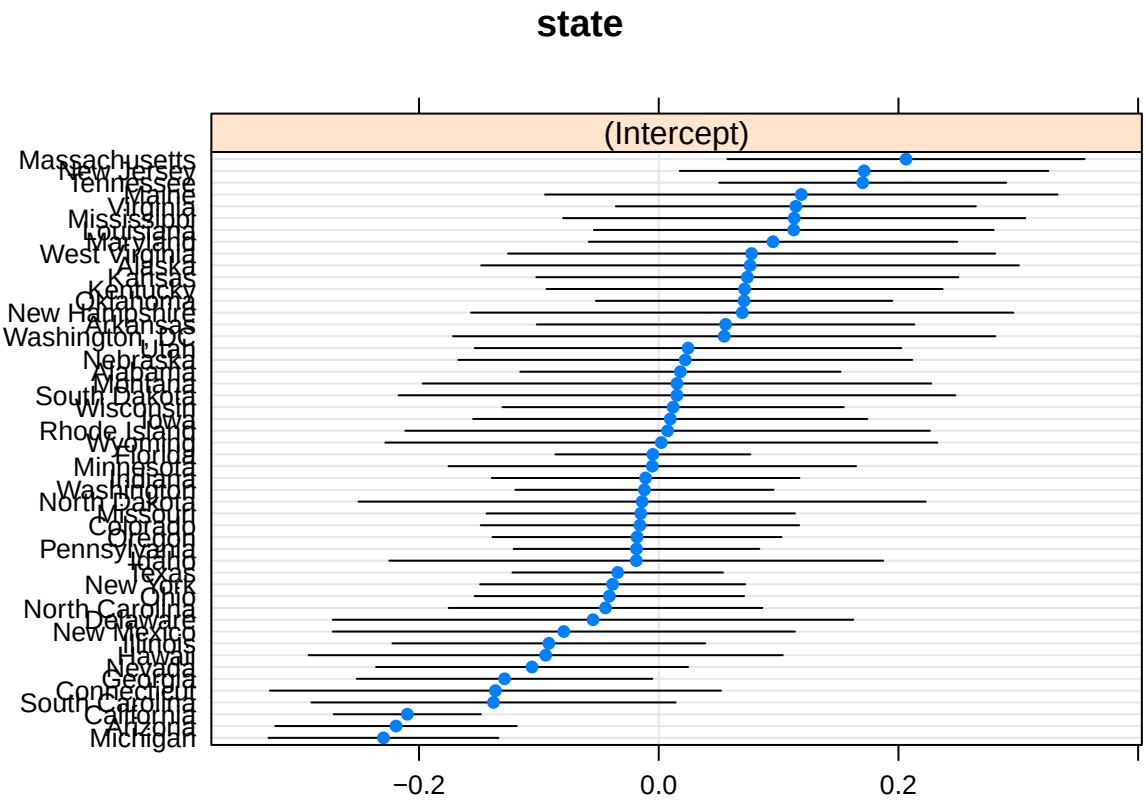


	Estimate
τ^2	0.0164183
σ^2	0.7521674



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

```
## $state
```



grpvar	term	grp	condval	condsd
state	(Intercept)	Alabama	0.0180621	0.0681530
state	(Intercept)	Alaska	0.0761966	0.1144620
state	(Intercept)	Arizona	-0.2192177	0.0513864
state	(Intercept)	Arkansas	0.0557536	0.0803693
state	(Intercept)	California	-0.2098077	0.0313387
state	(Intercept)	Colorado	-0.0157019	0.0677492
state	(Intercept)	Connecticut	-0.1362571	0.0960053
state	(Intercept)	Delaware	-0.0548465	0.1107876
state	(Intercept)	Florida	-0.0049114	0.0414968
state	(Intercept)	Georgia	-0.1286539	0.0629182
state	(Intercept)	Hawaii	-0.0943523	0.1009117
state	(Intercept)	Idaho	-0.0187574	0.1051288
state	(Intercept)	Illinois	-0.0917013	0.0665798
state	(Intercept)	Indiana	-0.0109387	0.0654690
state	(Intercept)	Iowa	0.0096280	0.0839329
state	(Intercept)	Kansas	0.0739183	0.0898259
state	(Intercept)	Kentucky	0.0716252	0.0843160
state	(Intercept)	Louisiana	0.1126790	0.0850980
state	(Intercept)	Maine	0.1189453	0.1090783
state	(Intercept)	Maryland	0.0953321	0.0784374
state	(Intercept)	Massachusetts	0.2063322	0.0760657
state	(Intercept)	Michigan	-0.2296611	0.0489520
state	(Intercept)	Minnesota	-0.0054051	0.0867296
state	(Intercept)	Mississippi	0.1130184	0.0983669
state	(Intercept)	Missouri	-0.0150272	0.0656503
state	(Intercept)	Montana	0.0152198	0.1082528
state	(Intercept)	Nebraska	0.0221270	0.0965797
state	(Intercept)	Nevada	-0.1057041	0.0663908
state	(Intercept)	New Hampshire	0.0696657	0.1154392
state	(Intercept)	New Jersey	0.1713437	0.0784374
state	(Intercept)	New Mexico	-0.0792024	0.0983669
state	(Intercept)	New York	-0.0384718	0.0563999
state	(Intercept)	North Carolina	-0.0444309	0.0667705
state	(Intercept)	North Dakota	-0.0138529	0.1207307
state	(Intercept)	Ohio	-0.0412173	0.0573466
state	(Intercept)	Oklahoma	0.0712079	0.0630790
state	(Intercept)	Oregon	-0.0181378	0.0615237
state	(Intercept)	Pennsylvania	-0.0185311	0.0522822
state	(Intercept)	Rhode Island	0.0074025	0.1116730
state	(Intercept)	South Carolina	-0.1378602	0.0775222
state	(Intercept)	South Dakota	0.0151629	0.1185281
state	(Intercept)	Tennessee	0.1701571	0.0610791
state	(Intercept)	Texas	-0.0342430	0.0447943
state	(Intercept)	Utah	0.0244102	0.0907733
state	(Intercept)	Virginia	0.1144124	0.0766384
state	(Intercept)	Washington	-0.0119301	0.0549568
state	(Intercept)	Washington, DC	0.0546268	0.1154392
state	(Intercept)	West Virginia	0.0774447	0.1036649
state	(Intercept)	Wisconsin	0.0120061	0.0726320
state	(Intercept)	Wyoming	0.0021434	0.1174711