

# Case Study 1 Writeup

## Introduction

The goal of this project is to investigate the heterogeneity of price per milligram of morphine by location and its relationship with other factors in a data set collected from [streetrx.com](https://streetrx.com). StreetRx is a study that has been collecting citizen-reported street price data of diverted pharmaceutical substances, such as morphine since 2010. Since the price of morphine provides information about the demand, availability, and potential drug abuse, the investigation of this project could reveal insights of the morphine price over various locations and other indicators that are helpful in health surveillance for drug abuse and other health-related issues. To this end, we have built a hierarchical model on morphine price with location grouping variable through a data analysis and modeling process performed in the following orders. First, since the data set relies on crowdsourced and noisy free-text data, we performed various data cleaning processes to generate a data set ready for modeling. Second, we performed exploratory data analysis to understand the distributions of variables and their potential relationships. Third, we looked into the heterogeneity of the response variable by different scales of locations and chose the proper location variable as the grouping variable with random effect. Then, we iteratively added other predictors, including both grouping variables and covariates, by identifying logically sound relationships and then performing diagnostics on the resulting models to assess validity. Possible interactions among predictors, which could be either fixed effects or random slopes, were also investigated in this step. Finally, after the final model was built using the results from the previous steps, we checked the validness of the model assumptions, interpreted the model results, and drew the conclusion.

Here are the findings from the modeling process. In summary, dosage strength and unit of drugs purchased are the two most important factors affecting the price of Morphine. Location factor such as state has a limited random effect on the price of the Morphine.

## Data Cleaning

The original data set contains reports of transactions of various types of drugs. By inspecting the data set entries and the structure for each predictor, we found the several issues with noisy, corrupted, or poorly formatted data. Here is a summary of the issues and the corresponding solutions to clean the data set.

First, in the original data set, `api_temp` variable specifies the active ingredient of the drug for each observation, and we need to filter the entire data set to obtain the morphine transactions according to the drug type specified by this variable. However, there were ambiguous labels of the `api_temp` variable entries, For example, there were categories specified as `morphine/oxycodone` rather than `morphine` only. Since we did not want to include observations that were related to other drug types to interfere the morphine price modeling process, we only kept the entries that were exclusively stated as “morphine”. This filtering process resulted in a new data set with 9268 records.

Second, we reformatted the `date` data in the data set. There were two fields representing date information: `yq_pdate` and `price_date`. The first field coded the year and quarter as a pseudo-continuous range. However, it was not suitable for assessing the data on a true continuous scale because it portrayed quarters within the same year as being closer together than adjacent quarters from different years (i.e., fourth quarter 2019 was closer to third quarter 2019 than it was to first quarter 2020 in this coding scheme). Therefore, instead of using `yq_pdate`, we used the `price_date` to construct a true continuous scale for measuring the date of purchase. Using R’s string parsing methods we parsed the date into its individual *month*, *date*, and *year*, and then created a new field for each record that counted the *number of elapsed days since a reference date*. In this case, we set *January 1, 2010* as the starting date, since the Streetrx data collection began in 2010. In addition, upon inspecting the date data we found that 3 entries were from the 1960s and another 11 entries

were from the 2000s, which were years before the data collection was active. In this case, there was legitimate concern that the entered data might not be accurate if it was supplied long after the drug purchase event. Since a relatively small number of data were affected, we filtered out the 14 entries with dates prior to 2010, leaving us with 9264 observations.

Third, we assessed the reported data for city in which the drugs were purchased, and the data set contained **city aliases and ambiguous references to some cities** other than the formal names of cities. We saw that there were 1690 unique “city” values among the 9254 observations. However, when we inspected visually we could see almost immediately that there were numerous entries with **different listed names that clearly refer to the same city**. This was because users might have used different conventions when supplying city names. For example, users might have listed either “Fort Lauderdale” or “Ft Lauderdale” to refer to the same city, resulting in two different values. We saw a range of other common data discrepancies, such as using city nicknames (e.g., “Philly” for Philadelphia), airport codes (e.g., “ATL” for Atlanta), or using the abbreviation used by major sports franchises within the city (e.g., “JAX” for Jacksonville). Various other issues were observed frequently, such as users providing redundant state information (e.g., “Des Moines, IA” instead of just “Des Moines”), referring to a city by the specific neighborhood or borough (e.g., “Brooklyn” instead of “New York”) or including single character typos (e.g., “Hollywood” instead “Hollywood”). We employed the following two approaches to address this issue.

In most cases, the original city could be unambiguously identified and corrected. First, we imported US census data which defined the official city abbreviation used by each city and cross referenced it with the listed data. Fortunately, this process covered the majority of the entries, but there were still about 300 city values supplied that were not on the list. Therefore, we manually created a new dictionary mapping the noisy values to the corrected values, in the cases identified above where the data could be definitively determined. Then, we applied this mapping to the original data to correct those entries which were non-compliant. During this process we also identified legitimate names of unincorporated areas and townships which were excluded from the original census data, so that we would not be unnecessarily throwing out data from real cities.

In other cases, the original city could not be unambiguously identified from the data given. For example, some users listed their zip code, which often crossed city lines. Others listed their county, which included multiple cities, and others listed the general metropolitan areas (e.g., “Lehigh Valley” or “Dallas - Fort Worth”). We did not want to guess, so in these cases, we replaced any remaining city names with **Other/Unknown**.

One pitfall that we had to avoid was inappropriately aggregating city data that were not related. For example, the data set contained both “Hollywood, FL” and “Hollywood, CA”. If we ultimately built a hierarchical model with both state and city grouping variables, we did not want to mistakenly label data from those two places as being from different states but the same city. Therefore, we augmented our coding of the city name by appending the state as well so that each city was uniquely encoded, even if it shared its name with another city in a different state.

Then, we inspected data defining the **source** of the price transaction data. Most entries were labeled with “Personal” or “Heard it” as the source, but there were still over 50 unique entries, too many to do rigorous grouping on. However, when we inspected each unique entry, we saw some common themes. First, many users entered the specific URL for various web pages they searched, and several web pages were represented repeatedly with different URLs. But more importantly, there were clear patterns in the types of sources listed. In particular, we observed that each of the sources was one of the following: (A) personal, (B) word of mouth, (C) a web forum, (D) an online black market, (E) a legal online market, (F) a web search, or (G) other/unknown. In fact, we could very easily convert the raw data to these categories by using sub-string search to find particular keywords, such as “silkroad”, “bluelight”, “reddit”, “opiophile”, “forum”, “pharmacy” and various other terms. As a result, we bundled each source into one of those seven categories, making for much simpler and more informative grouping. This data cleaning step did not change the size of the data set.

Our final data cleaning step included many minor issue corrections and basic cleaning of other predictors. First, we noticed that the **state** field included both the entries “USA” and “”. We converted these values to “Other/Unknown” category since they did not provide any meaningful information. Second, we saw that **bulk\_purchase** was coded as a string, which wasn’t helpful, so we converted it to a factor (0 = False, 1 = True). Third, we observed that there was nothing listed under the **Primay\_Reason** field when the user did

not enter the reason for the purchase. We marked such entries with missing values as “Other/Unknown” as we have with the other fields.

In summary, after the data cleaning process, we obtained a data set with 8712 observations and 9 predictors, which are shown below.

- **ppm**: response variable, price of morphine per milligram, numeric value.
- **city**: cities, where the transaction took place, 1654 unique factors.
- **state**: states, where the transaction took place, 56 unique factors.
- **region**: regions, where the transaction took place, 5 unique factors.
- **primary reason**: reasons for purchasing morphine, 11 unique factors.
- **source**: sources of such transactions, 8 unique values.
- **days ellapsed**: number of days passed since 01/01/2010, which is the time when the data collection process started, discrete numerical value.
- **dosage strength**: dosage strength of the purchased morphine, discrete numeric value, ranging from 1 to 200.
- **bulk**: indicator for purchased 10+ units, factor with levels 0 and 1.

The data analysis and modeling were performed upon the cleaned data set with the above predictors.

## Model design

### Group Variable Selection

Before determining other predictors to be included in the model, we would like to first select the best location group variable which explained the heterogeneity of morphine price over that location level. We used two approaches to achieve this end. First, by inspecting the EDA figures, (**Figures Needed Here**), we could observe that the  $\log(\text{ppm})$  did not differ too much across different **region** categories, but there were some variations in  $\log(\text{ppm})$  across the sampled cities and sampled states. Therefore, we suspected that **region** might not be a good predictor explaining the heterogeneity across locations. As for **city** variable, the sample sizes within most cities were very small, so the **city** variable was probably not optimal to be included as the only location group variable in the model. The **state** variable seemed to be optimal for location group variables. In addition, we also would like to consider including more than one location group variables in the final model.

With these suspicions and ideas in mind, we built hierarchical models with random effects on every possible combination of the three location variables, which were **city**, **region**, and **state**. There were seven models in total, which were composed by 3 hierarchical models including each location variable only, three hierarchical models each including a unique pair of the three location variables, and finally the last hierarchical model including all three location variables. Then, we compared these seven models using BIC values as criterion based on the principle that, in each comparison, the two models differed in at most one variable. For example, we could compare the BIC score of the model of random effects in **city** and **region** with the BIC score of the model of random effect in **city**. But we could not compare the model of random effects in **city** and **region** with the model of random effects in **city** and **state**. This principle could also be observed in the model comparison results, which are shown below and ranked by BIC values.

1.  $(\text{state}) < [(\text{state}, \text{city}), (\text{state}, \text{region})] < (\text{state}, \text{city}, \text{region}) < (\text{city}, \text{region}).$
2.  $(\text{state}, \text{city}) < (\text{city}).$
3.  $(\text{state}, \text{region}) < (\text{region}).$

Here are some explanations of the notations.  $(\text{state})$  indicates the model with random effect in **state** only, without any other predictors.  $(\text{state}, \text{region})$  indicates the model with random effects in both **state** and **region**, no any other predictors involved. And if  $\text{model1} < \text{model2}$ , it means that the BIC score of model 1 is smaller than the BIC score of model2. The BIC scores of the models mentioned above are shown in the table below.

Table 1. The BIC score of the models with location group variables only.

Name	Degree of Freedom	BIC Score	Random Effects
re m1	5	24141	city+state+region
re m2	4	24192	city+region
re m3	4	24132	state+region
re m4	4	24133	city+state
re m5	3	24207	city
re m6	3	24124	state
re m7	3	24209	region

From the above results, it was easy to see that, according to the BIC scores, the model with random effect in **state** only was the best. Therefore, we chose to use **state** as the only group variable in locations with random effect.

### Selection Among Other Variables

Now, the base model only involves the location variable **state** with random effect. We built the rest part of the model using forward stepwise selection of the predictors.

Table 2. The BIC scores of the models with specified random effects and fixed effects.

Name	Degree of Freedom	BIC Score	Random Effects	Fixed Effect
base	3	24124	<b>state</b>	None
m1	4	23105	<b>state</b>	<b>dosage strength</b>
m2	4	23100	<b>state + dosage strength (by state)</b>	<b>dosage strength</b>
m3	5	23098	<b>state</b>	<b>dosage strength + bulk</b>
m4	6	23105	<b>state</b>	<b>dosage strength * bulk</b>
m5	12	23134	<b>state</b>	<b>dosage strength + source + bulk</b>
m6	47	23442	<b>state + source (by state)</b>	<b>dosage strength + source + bulk</b>
m7	19	23182	<b>state</b>	<b>dosage strength + source * bulk</b>
m8	15	23182	<b>state</b>	<b>dosage strength + primary reason + bulk</b>
m9	80	23140	<b>state + primary reason (by state)</b>	<b>dosage strength + primary reason + bulk</b>
m10	6	23104	<b>state</b>	<b>dosage strength + bulk + days ellapsed</b>
m11	8	23117	<b>state + days ellapsed (by state)</b>	<b>dosage strength + bulk + days ellapsed</b>

First, from the (**mgstr EDA graph**), we could see that there was a slightly negative relationship between the **log(ppm)** and the **dosage strength** variable. Since the **ppm** was log-transformed, a slightly negative coefficient between **log(ppm)** and **dosage strength** might be enlarged after exponentiation. Therefore, we considered the **dosage strength** (dosage strength) as a quite important predictor and added it to the model. By comparing with the base model, the model with the addition of **dosage strength** as a fixed effect had much lower BIC score. (Table 2. m1 vs base model) Therefore, we decided to include **dosage strength** in the model. Now, the model contained random effect of **state** and fixed effect of **dosage strength**.

From the **EDA plots**, **dosage strength** seemed to have different effects on the  $\log(\text{ppm})$  response variable across the sampled states. Therefore, we also considered adding random slopes of **dosage strength** by **state** to the model. Not to our surprise, the BIC score of the model decreased compared to the model without random slope of **dosage strength** by **state**. (Table 2 m2 vs m1) However, the model was not converged, so we did not include the random slope of **dosage strength** by **state**. Therefore, **dosage strength** was added to the model only as a fixed effect.

Second, from the **EDA Figure bulk**, the  $\log(\text{ppm})$  seemed not to vary across different bulk sizes, both in general and across sampled states. Therefore, we considered **bulk** as an unimportant variable. However, adding **bulk** to the model slightly lower the BIC score. (Table 2. m3 vs m2) We preferred not to add random slopes of **bulk** by **state** to the model since **bulk** variable did not have different effects on the  $\log(\text{ppm})$  across the sampled states. Thus, we decided to add **bulk** to the model as fixed effect only. At this point, the model included random effect of **state**, and fixed effects of **dosage strength** and **bulk**.

There was a potential interaction between **dosage strength** and **bulk** variable since such interactions might imply a potential discount in the drug sale. We added the interaction to the model, but the BIC score increased. (Table 2. m4 vs m3) Therefore, the interaction term was excluded from the model.

Third, we considered to add **source** and **primary reason** to the model. From the **EDA graphs**, within sampled states, the **source** predictor seemed to have different effects on  $\log(\text{ppm})$  across different states. Therefore, we considered to add random slope of **source** by **state** and fixed effect **source** to the model, however, the BIC score of the model increased a lot after adding either fixed effect of **source** or both random effect and fixed effect of **source** to the model. (Table 2. m5 vs m3, m6 vs m5) Therefore, the **source** variable was excluded from the model. We also excluded the **Primary Reason** from the model. From the (**Figure Primary**), we could see that the response variable,  $\log(\text{ppm})$ , did not change across different **Primary Reason** categories. And across the sampled states, **Primary Reason** did not have different effects on the  $\log(\text{ppm})$  response variable. Therefore, we would like to exclude **Primary Reason** from the final model. In addition, the model with **primary reason** as fixed effect or both random effect and fixed effect increased BIC values a lot, which further confirmed our decision not to include it in the model. (Table 2. m9 vs m8, m8 vs m3)

We also tried to include the interaction between **source** and **bulk** as fixed effect in the model, but this interaction increased the BIC score a lot, so it was finally excluded from the model. (Table 2. m7 vs m5)

Finally, based on the **EDA Figures**, it seemed that there was no obvious relationship between the  $\log(\text{ppm})$  and **days ellapsed**. However, we could observe slightly different slopes of **days ellapsed** of the sampled states. In the meantime, we thought that as the local economy changed over time, the price of the morphine might also change over time as well, which also made us believe that the **days ellapsed** should be added to the model. And since each state had different economy development histories, the **days ellapsed** variable might have different relationships with  $\log(\text{ppm})$  in different states, which led us to add random slope of **days ellapsed** by **state** to the model. Therefore, we added **days ellapsed** with random slope by **state** to the model along with its fixed effect. However, either adding fixed effect **days ellapsed** only or adding both random effect and fixed effect would increase the BIC score of the model. (Table 2. m11 vs m10, m10 vs m3) Therefore, we finally excluded this variable from the model.

### Model Summary

From the above reasoning and testing results, our final model could be summarised as the following formula.

$$y_{ij} = \beta_{0,j} \text{ State}_j + \beta_1 \text{ Bulk}_{i,j} + \beta_2 \text{ Dosage Strength}_{i,j} + \epsilon_{i,j}$$

$$\beta_{0,j} = \beta_0 + b_j; \quad b_j \sim \text{Normal}(0, \tau^2)$$

$$\epsilon_{i,j} \sim \text{Normal}(0, \sigma^2)$$

In the above formula,  $b_j$  represents the random effects of **state**, and  $\epsilon_{i,j}$  is the error term.  $i = 1, \dots, n_j$ , where  $n_j$  represents the number of observations in group  $j$ .  $j = 1, \dots, J$ , where  $J$  represents the number of groups.

We have neither information about nor strong belief regarding the random effect parameter  $\tau^2$  and error term variance  $\sigma^2$ , we would like apply a non-informational prior on these parameters. Under this condition,

the information from the data set has a quite large weight estimating the model parameters, which is similar to fit the model without priors. Therefore, considering the time cost of running the models, we decided to use the normal hierarchical model, which is specified as above.

### **Potential Limitations**

forward stepwise selection might miss the optimal model.