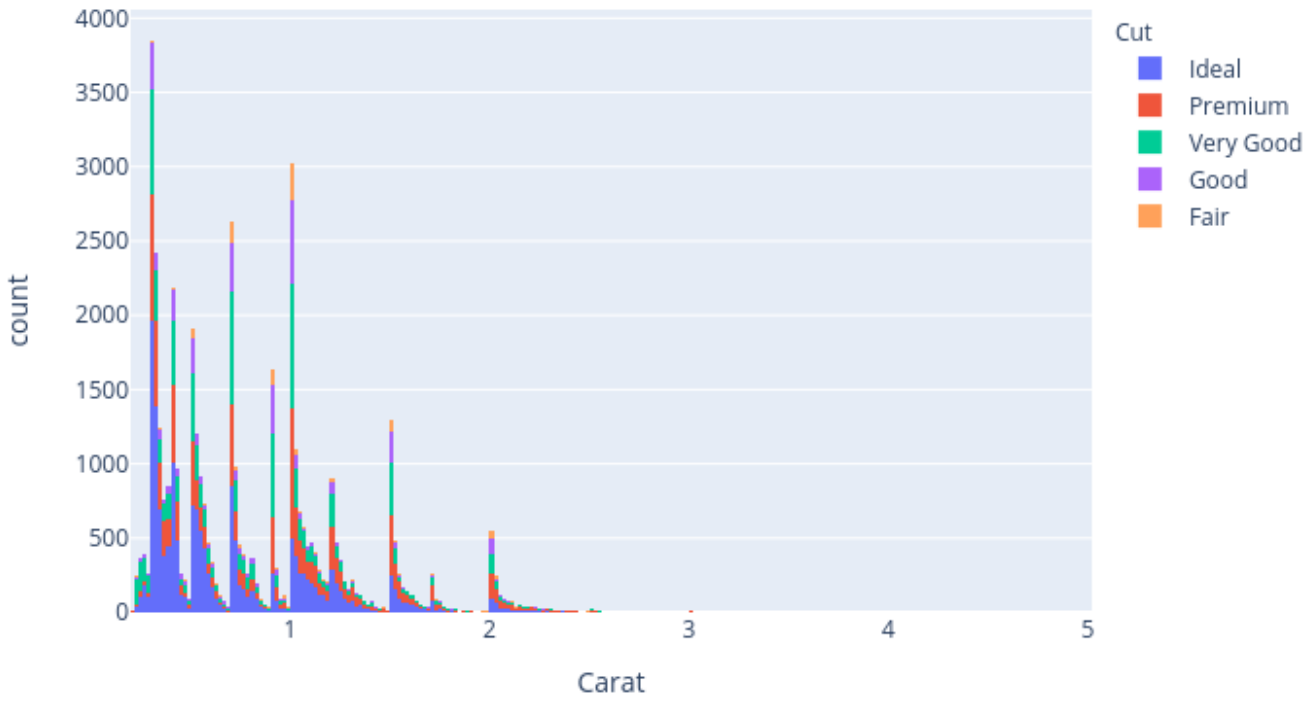
On the purpose of the project

Since the use of the model would be *validation* of the prices proposed by Krenk, just give an approximation to the price could have not been enough. If we give likely ranges for the values of the diamond, the validation can be done by verifying if the total cost is reasonable in terms of the computed ranges. Therefore some interval estimation is needed.

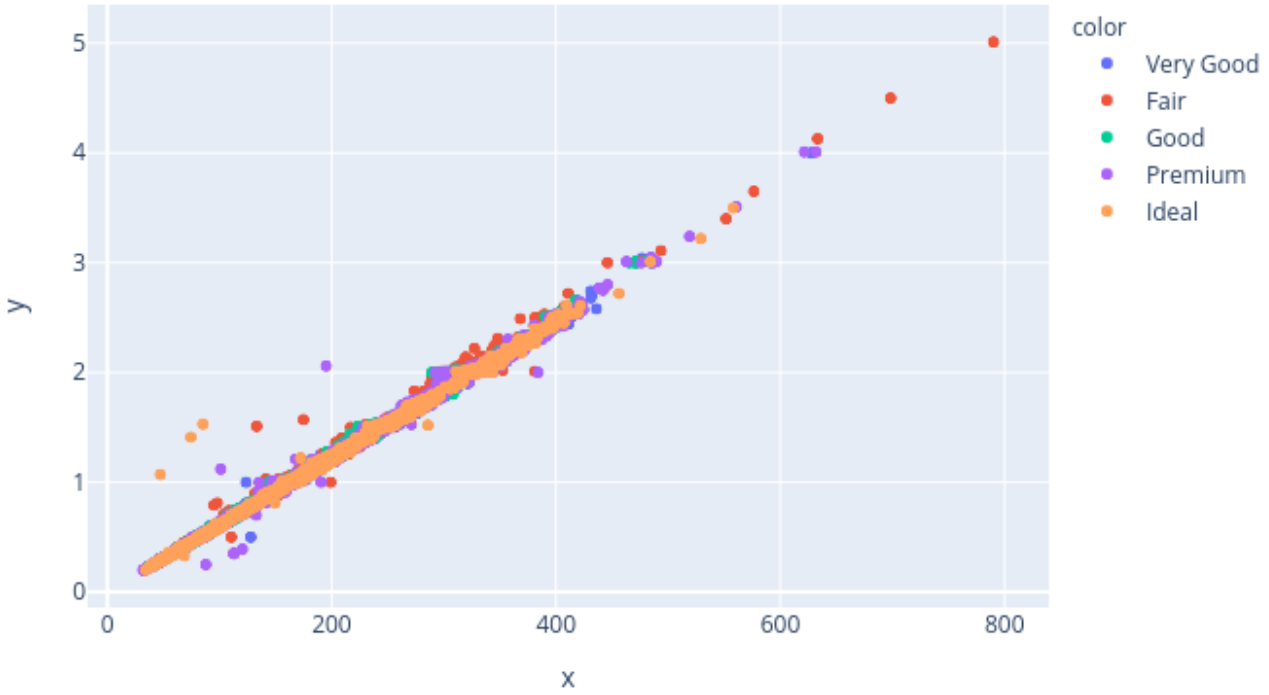
On preprocessing

The basic idea of the cleaning process was to get as much data available as possible for regression purposes. Some effort was put in correcting the categories corresponding to the **4C** labels. Most of them were *typos* where extra characters were added. Nonetheless I decided not to verify the impact on the particular classes of the stolen diamonds (increase in sample size). Those queries would have taken some time, and several of them were not performed since it would have taken to me significantly longer.

I think that the histogram of carats reflect some categories marked by "round values". Maybe stratified sampling considering the 4C's would be better for setting aside the test set, after *binning* the Carat variable. On each bin (after an appropriate shift) I guess the distribution of Carats is something like Weibull distributed.



The interpretation of the variables x, y and z as lengths for producing a volume was succesful. In fact, high correlation with Carats was detected. The correlation was used to clean the data of non-positive entries in those columns, yielding the next image. The same should have been done to the test set.



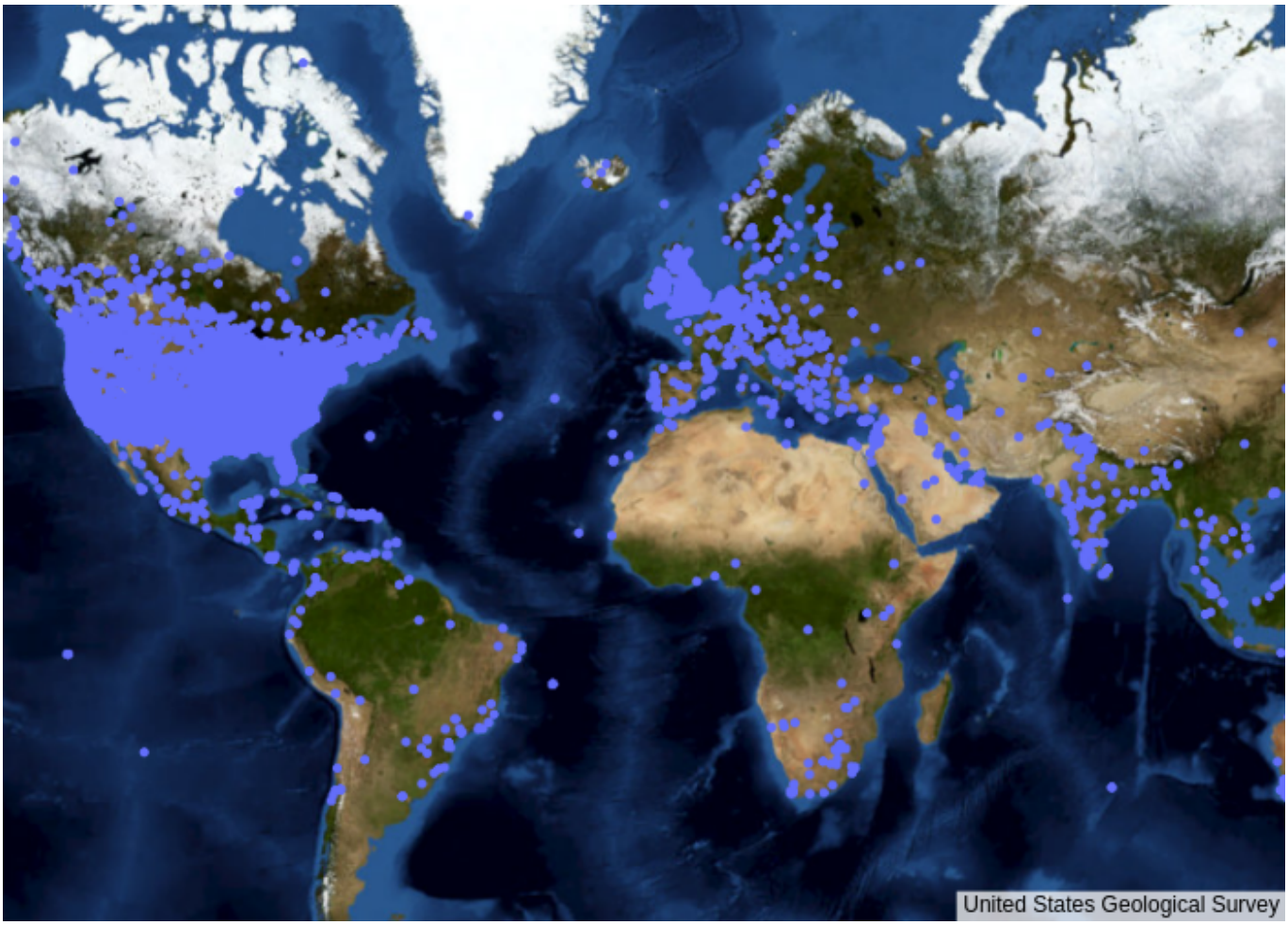
Some other feature engineering would have worked, for example, considering the measurements of `Depth` and `Table` and its realtion to `Cut`, since the cut is affected by the particular proportions of the diamond parts. Maybe the `Cut` is the result of categorizing via `Depth` and `Table` in a smart way.

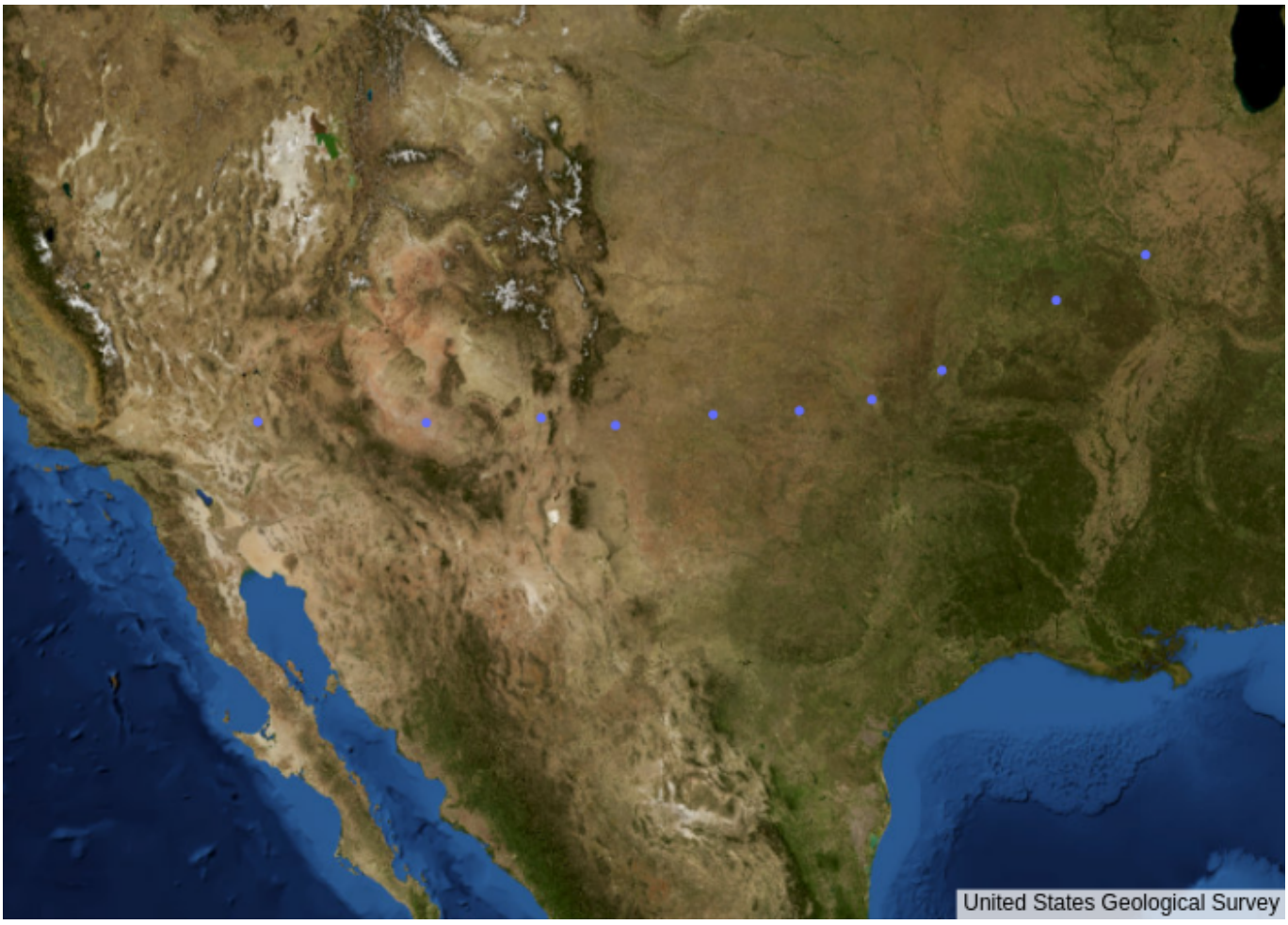
The benefits of identifying Categorical variables as cuts on continuous variables is that it's preferrable consider the continuous information instead of a crude approximation. On the other hand, `Color` and `Clarity` don't seem to have any relation whatsoever to the diamond size. But this was never confirmed. In conclusion, more feature engineering is needed.

On the exploration of data

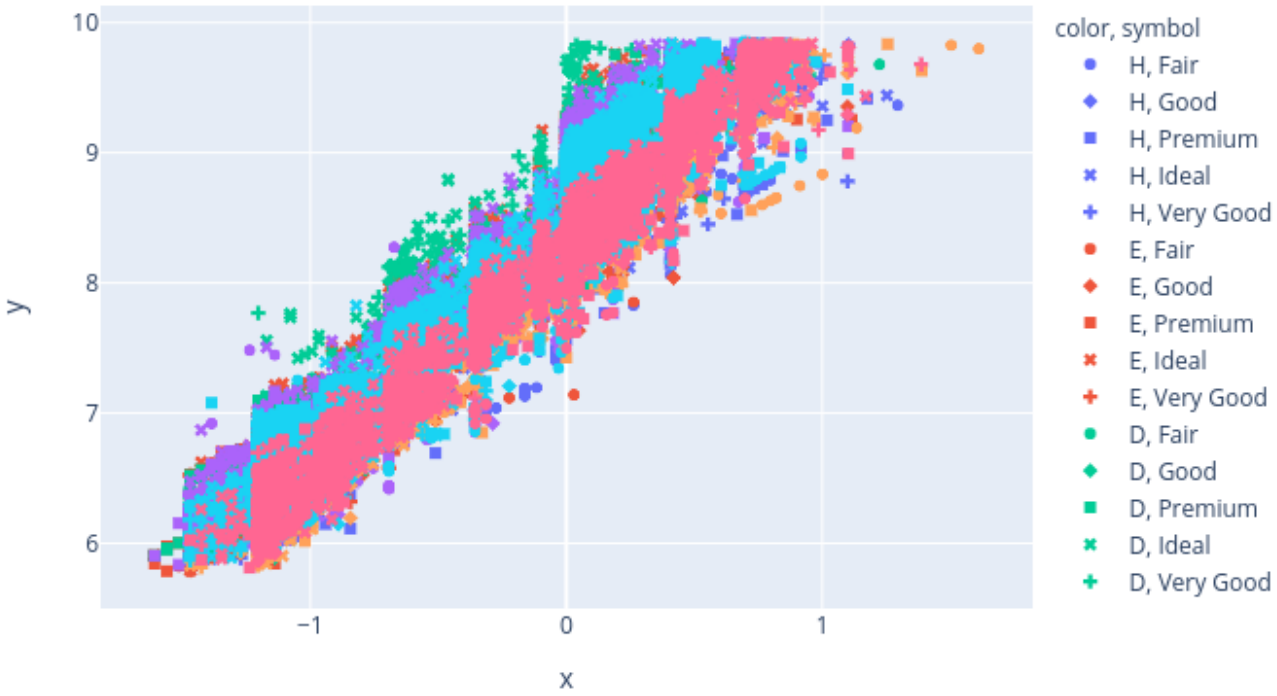
The most unexplored part of the data was the geographic part. The locations given in the database are shown below. The Gringotts diamonds had locations in the USA, almost tracking a curve. But without more information about what those locations means, it's har to say anything.

In the code I obtained a filtered database considering only diamonds in the USA. For ease of coding, I considered an enclosing rectangle of the USA, but maybe the country or continent information could have been valuable.


United States Geological Survey


United States Geological Survey

The principal observation was that the logarithm of Carat and the logarithm of price were highly correlated. Together with the other C's, it was the base for our estimation.



On models

The *Manual solution* is somewhat of a combination of a decision tree with a 1-NN Regressor. First, all data is classified according to their color, cut and clarity. Then the nearest diamond in such class was given as the prediction for the value. I would expect the result to be of high variance, because of the maximal division among clases in the tree part, and the selection of a single neighbor. This estimation was more or less robus to a change in the numeric feature variables. Those variables were scaled to be in the interval [0,1]. Sampling would be needed in order to produce likely ranges for the mean.

A natural idea to try was a Regression Tree of Random forest to perform the same task, but they were not used because of time. Other instance-based learning techniques could be useful such as nearest neighbors on its own. Of course parameters would have to be chosen by Cross-validation.

The linear model was easy to implement. After a brief test using all the possible variables in regression and having an enormous condition number (indicative of correlation), a small model with just the 4C was used. A variable was added but it didn't work so well. I think that a selection method would have been useful, such as a greedy algorithm to add variables one at a time, and using the Akaike information criterion for selecting the best model. Again, time was the enemy.

In the small linear model, al variables turned out to be significant, and the categorical variables responded as expected, since they had order and an increment in the category was correlated with increment in the diamond's price. The use of logarithms lead to a lognormal random variable, which was used to predict interval around the mean price.

Here I have supposed that the fair price of the diamonds is the mean price of similar diamonds. Maybe the best concept is the median (half the diamonds are priced higher and half lower). In that case we would be overestimating the price. The mean of the lognormal(mu,sigma) is exp(mu + \sigma^2/2), while the median is exp(mu).