**Problem:**

Predict the product that a customer will most likely end up when shopping for an insurance policy.

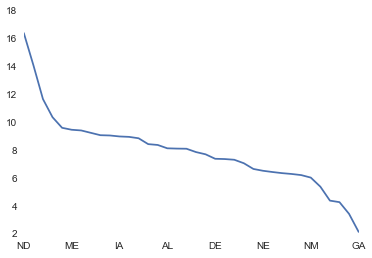
**Known:**

* We know the customer history. Every time he shops/calls to find out the price of the product type he is looking for.
* We know what options he is requesting information on in the product during every interaction.
* We know when he buys it and the final set of set of options purchased.
* We know that the product cost is a function of the product options plus the customer characteristics.

**Exploratory Data Analysis:**

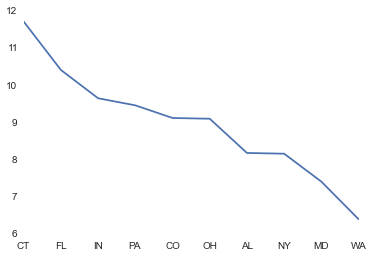
Questions Asked:

1. Regional States / Zip codes
   1. What region/states contribute to largest customer inflow? Average customers per Zip code Location – gives a good indication of density of customer traffic.



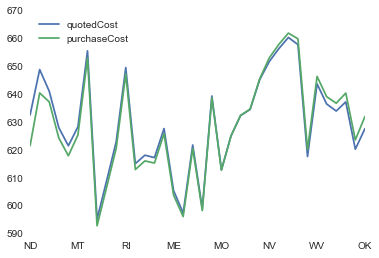
Interesting thing is for the set of states greater than the mean number of customers contributed, we see the following divergence for the number of customers per location.

|  | **No. of Customers** | **No. of Location** | **Customers/Location** |
| --- | --- | --- | --- |
| **CT** | 1549 | 132 | 11.73 |
| **FL** | 7789 | 747 | 10.43 |
| **IN** | 2078 | 215 | 9.67 |
| **PA** | 4823 | 509 | 9.48 |
| **CO** | 1911 | 209 | 9.14 |
| **OH** | 3531 | 387 | 9.12 |
| **AL** | 1918 | 234 | 8.20 |
| **NY** | 7250 | 886 | 8.18 |
| **MD** | 2274 | 306 | 7.43 |
| **WA** | 1932 | 301 | 6.42 |

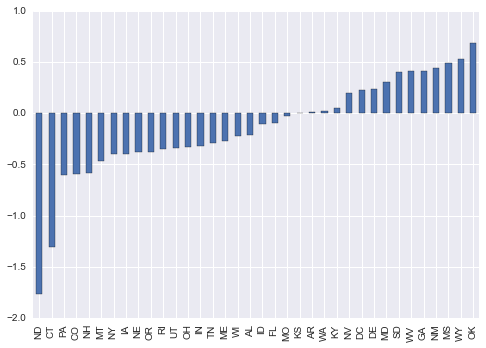
****

**Recommendation:** Would be to good effort see why there is such a huge divergence in density of customer traffic? There may be opportunity for growth by deploying more marketing dollars in areas with low density of customer traffic and higher divergences.

* 1. What is the mean quoted / purchase price per customer in different states?

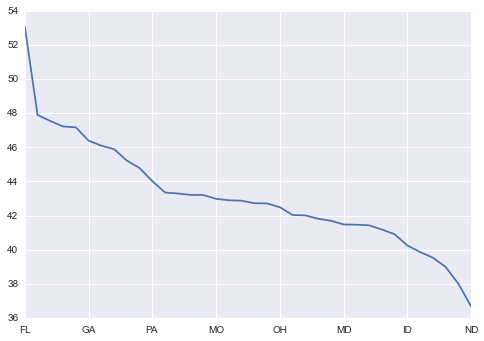


We can see that the people in ND seem to typically buy their policies a little below the quoted price at the time of interactions. Whereas folks in ME / MO typically tend to purchase the policy pretty much about the same price at the time of initial interactions. The chart below would give a better snapshot into the sensitivity of folks in different states.



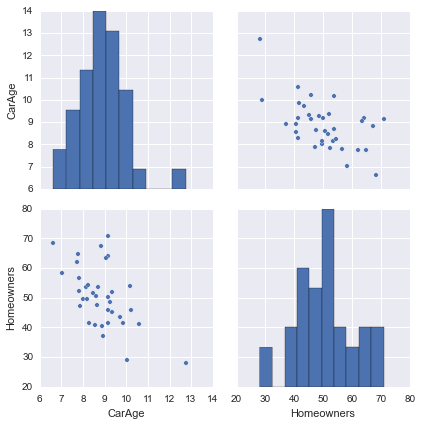
**Recommendation:** Folks in CT, ND seem to be sensitive to price. So a product option on higher price range than what is being offered today would not be a good idea for those states. However, the folks in WY, OK do not seem to be as affected by the price since they seem to come back and end up purchasing products at higher price points. Those markets could be potential test bed for new product options at higher prices, to gauge the willingness to pay from the consumers.

1. People Persona
   1. Average Oldest



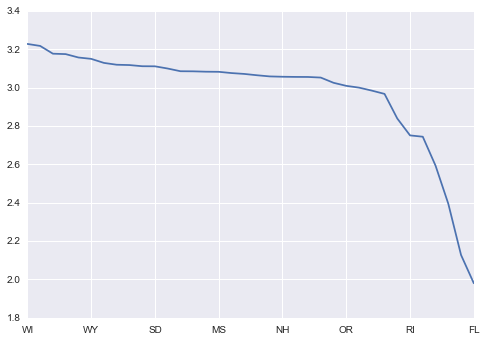
Florida seems to have the oldest population and North Dakota the youngest. Car insurance would probably be a better sell with states on the higher end of the age range given the tendency of car ownership seems to be declining with younger population.

* 1. Home Ownership vs Car Age



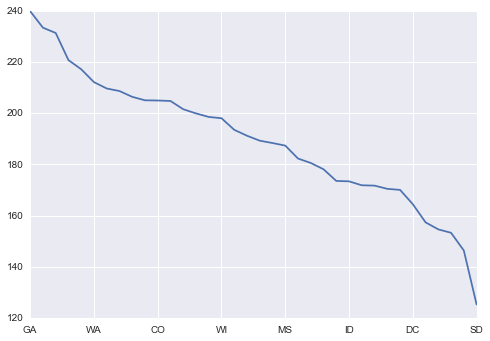
This one is an interesting info graphic. It kind of seems like the higher the concentration of homeowners, the lower the age of the car. That is surprising. One likely explanation is that homeowners tend to buy newer cars. It would be recommended to target homeowners when selling insurance products since they probably are more inclined to buy with newer cars in the lot.

* 1. Risk Factor



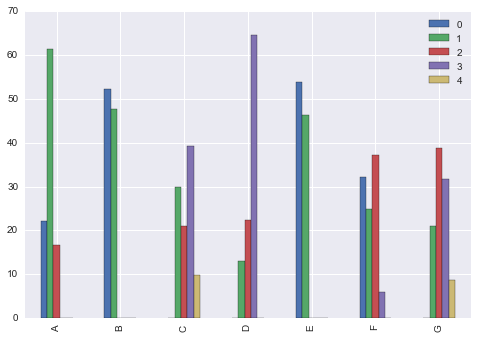
Florida also seems to be way less risky than a lot of the other states. However the cost of their auto insurance is still among the top 5 purchase costs for customers. There may be a way to lower the cost of products to consumers from such less risky states.

* 1. Average Interaction Span to Sale:



People from Georgia seem to talk a lot before making an actual purchase. Call center traffic from there is probably going to need a few extra resources to provide satisfactory services.

1. Distribution of Product Options and their Attributes.



The above distribution of product options actually purchased by consumers, reflect their inclination to the final choices. Clearly there are some preferences. For e.g. in options C and G, clearly the choice 4 is not really preferred by consumers. Maybe that can be dropped from the availability and the resources can be deployed towards experimenting with some other alternative.

**Model Building:**

**Model Assumptions:**

* If risk factor is unknown, substituting it with mean = 2.5 risk factor of the sample.
* If Previous C product option was NA, substituting it with 0
* If previous client enrollment duration is NA, substituting with 0
* If car value is NA, substituting with NA

**Variable Definitions/Transformations:**

* Total Customer Interactions/touch points
  + This is the total number of times the customer has interacted with the business ending up with a product buy.
* Time Interval
  + Time Interval between interactions
* Total Time to Sale
  + Total time taken from first interaction to actual sale

**Model approach: 1**

* Split 30% of data into test set.
* Mark each product option as a dependent/target variable
* For each product option
  + Build a classifier
    - Since this involves a lot of categorical variables, I went with a random forests / decision tree classifier to start with. There may be better results with other algorithms.
  + Get the classification scores on the training set.
  + Use the model as a predictor on the test set.
* Use the test scores across the 7 options to get the likelihood of this being a good predictor of consumer preference.
* Once we have the predicted 7 options, the product choice per consumer is just a combination of the predicted options.
  + That should help us determine the product a customer is likely to choose.
  + Could filter the ones with record type = 1 (Purchase point), with mean time interval between transactions / time to sale.
  + Could help explore when folks with mean shopping touch points/time to sale of the overall sample are likely to buy.

**Model Results:**

* Accuracy of the Classifier:

|  |  |  |
| --- | --- | --- |
| **Modelled Variable** | **Training Set** | **Test Set** |
| A | 95.39% | 95.50% |
| B | 92.44% | 92.52% |
| C | 91.75% | 91.26% |
| D | 92.84% | 92.56% |
| E | 94.21% | 94.62% |
| F | 93.03% | 93.35% |
| G | 83.97% | 83.29% |

For the Training set, I used a 10 fold cross validation and the accuracy is the mean of all the 10 scores. The training set was 70% of the original data set.