DSC520 Final Project - Part III

MICHAEL ERSEVIM

Bellevue University, Fall 2021

Introduction

Like many companies, Waste Management (WM) is looking to streamline its business processes in an attempt to reduce internal costs while improving the customer experience. Specifically, that experience is being able to request a price for a certain service that we, WM can't provide without relying on the assistance of a third party, whose costs we don't know upfront. Being able to estimate these costs quickly and accurately allows us to service a customer more quickly, improving their satisfaction, as well as reducing the manual workload of a client service representative.

The problem statement

The issue to be solved is one of predicting a cost that a third-party hauler (TPH) will charge our company for a service on our behalf. When a client requests a service that is in one of the locations where we can't provide the service with our own trucks, this service must be passed along to a TPH. This TPH will then quote us a cost to do that service, and then we will get back to the client with a marked-up cost (aka 'price') to do that service. This introduces a long delay between requesting the service and receiving a price quote for that service.

Addressing the issue

By utilizing predictive modeling, the costs (and subsequent prices) could be predicted and quoted to the customer instantly, allowing them to make an informed decision before authorizing the service. Ideally, the eventual costs of the procured service will be close to the modeled estimate.

The downside is that this introduces risk to WM. If the model estimates a TPH cost that is too low, we will lose money by having to honor the quote which was based on a cost that was not able to be procured or negotiated in the open market. On the other hand, if the model is biased too high, the resultant price may drive away customers or to utilize other trash service providers.

The balance of minimizing the variance of the estimates and not biasing the model too high or low puts this problem squarely in the hands of a data scientist.

Analysis

The main dataset is proprietary WM client data saved as a '.csv' file. It will be anonymized and only represent a reasonable subset (<2% of total database records) for practicality and proprietary concerns.

What we ultimately want to get to is a reasonable cost for a Haul (H) and a disposal (DSP) for each county in a state.

First we start by importing and shaping the raw data, making dates into dates and characters into factors for the regression later on.

```
## Warning: package 'ggplot2' was built under R version 4.0.5
```

```
## Warning: package 'ggm' was built under R version 4.0.5
```

```
## Warning: package 'readxl' was built under R version 4.0.5
```

```
## Warning: package 'psych' was built under R version 4.0.5
```

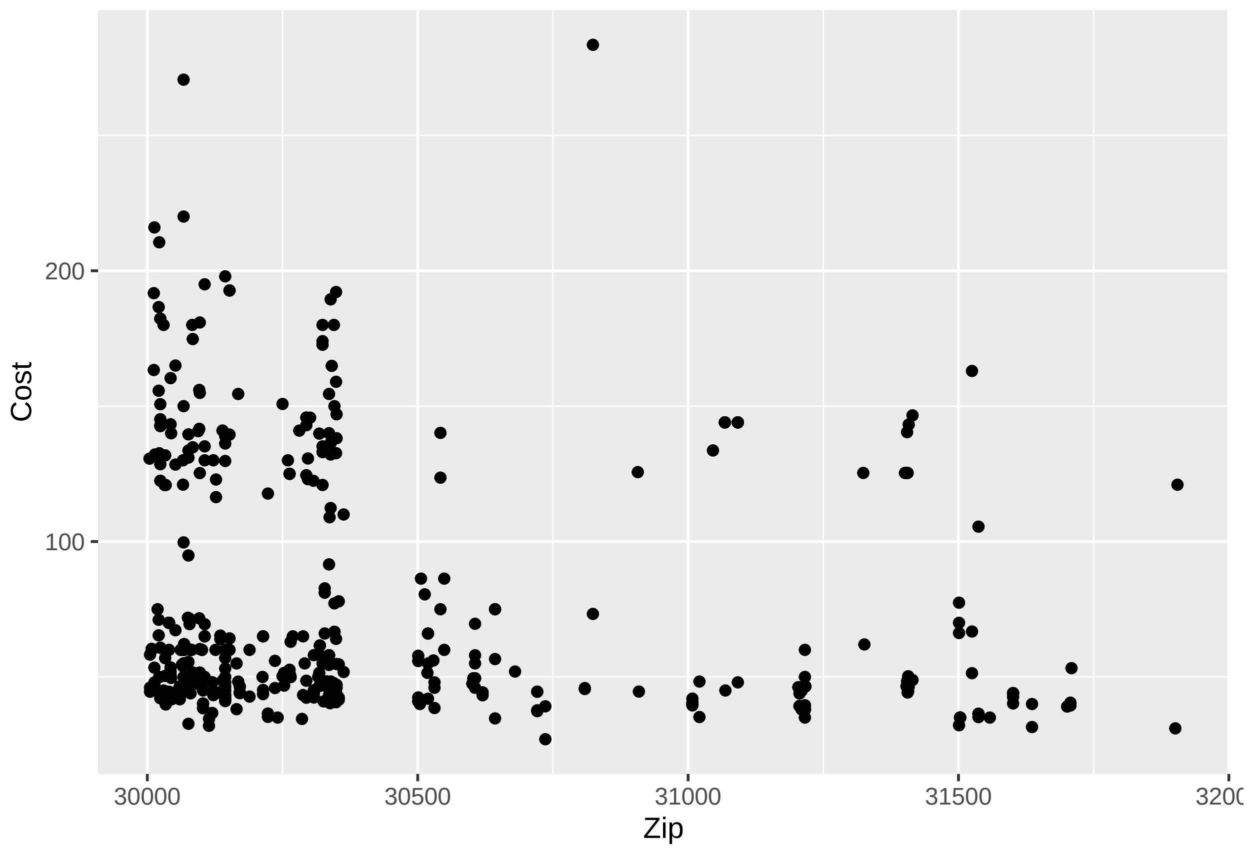
```
##
## Attaching package: 'psych'
```

```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
## Warning: package 'useful' was built under R version 4.0.5
```

```
##   StartEffDate          EndEffDate              Cost             City
##  Min.   :2013-08-01   Min.   :2021-01-07   Min.   :  10.03   Length:10000
##  1st Qu.:2021-02-25   1st Qu.:2021-03-31   1st Qu.:  64.00   Class :character
##  Median :2021-05-10   Median :2021-06-10   Median : 163.60   Mode  :character
##  Mean   :2021-03-20   Mean   :2021-06-07   Mean   : 213.00
##  3rd Qu.:2021-07-16   3rd Qu.:2021-08-20   3rd Qu.: 300.00
##  Max.   :2021-11-01   Max.   :2021-11-30   Max.   :2025.00
##                       NA's   :3297
##       County         State           Zip         Svc.Type    Mat
##  Cook     : 238   GA     : 787   Min.   :  659   DSP:5000   C&D: 375
##  Maricopa : 190   TX     : 747   1st Qu.:28373   H  :5000   MXR:   2
##  Harris   : 184   OH     : 567   Median :40509              SSR:  82
##  Jefferson: 182   IL     : 540   Mean   :45250              T  :9418
##  Fulton   : 153   FL     : 527   3rd Qu.:64622              WD : 123
##  Dallas   : 149   NC     : 476   Max.   :99701
##  (Other)  :8904   (Other):6356
##  Sched          TempOrPerm   Container       Size
##  OC :5000   Permanent:4406   CMP:2816   YRDS-30:7004
##  SOC:5000   Seasonal :  14   OT :7184   YRDS-40:1496
##             Temporary:5580              YRDS-20: 422
##                                         YRDS-34: 347
##                                         YRDS-35: 274
##                                         YRDS-42: 241
##                                         (Other): 216
```

Next, we'll pick a state (GA in this case) to look at the costs of DSP across zip codes.

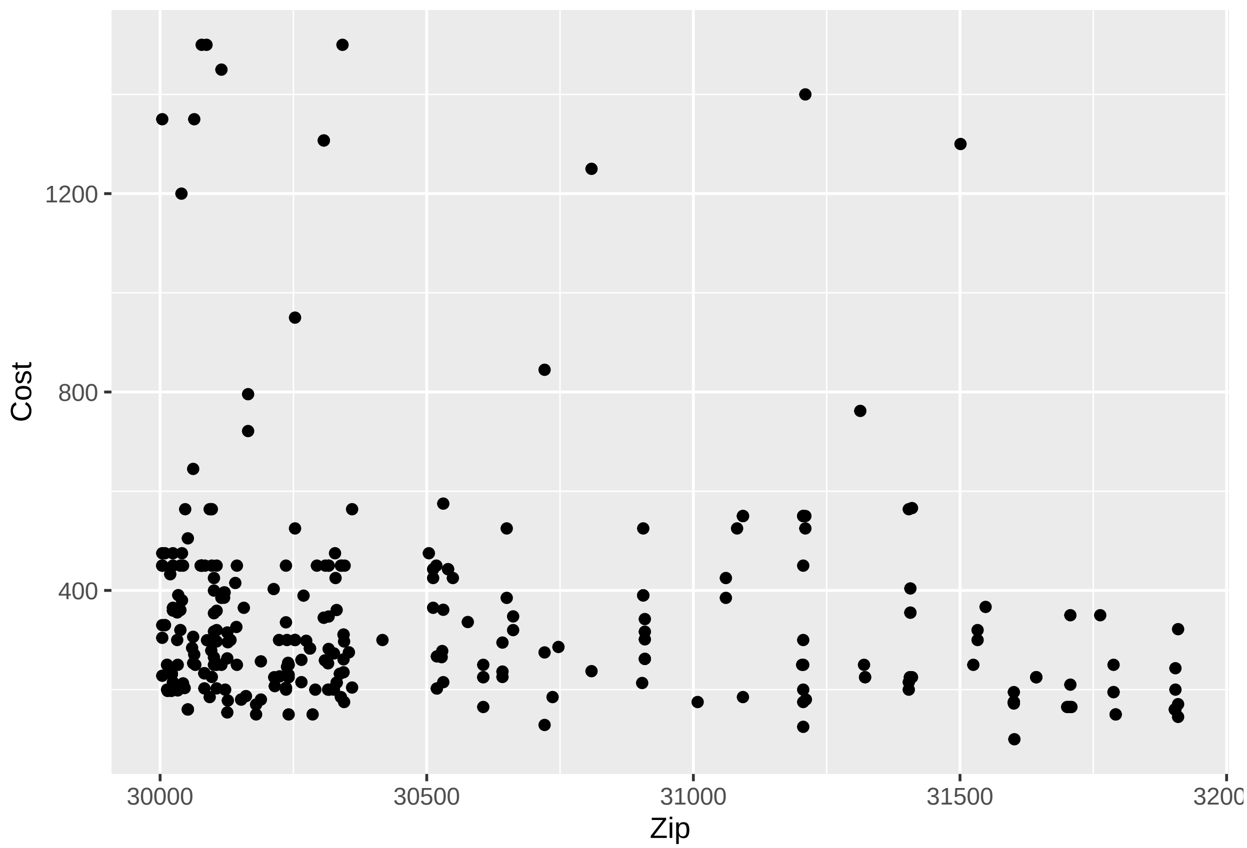


We see above that there are groupings or strata of costs. This is due to data entry errors. Many times in our industry, DSP costs are for 3 tons of material.

THe cost is supposed to be entered at the unit rate, not extended for 3 tons. The layer above the lower clusters are most likely these errors.

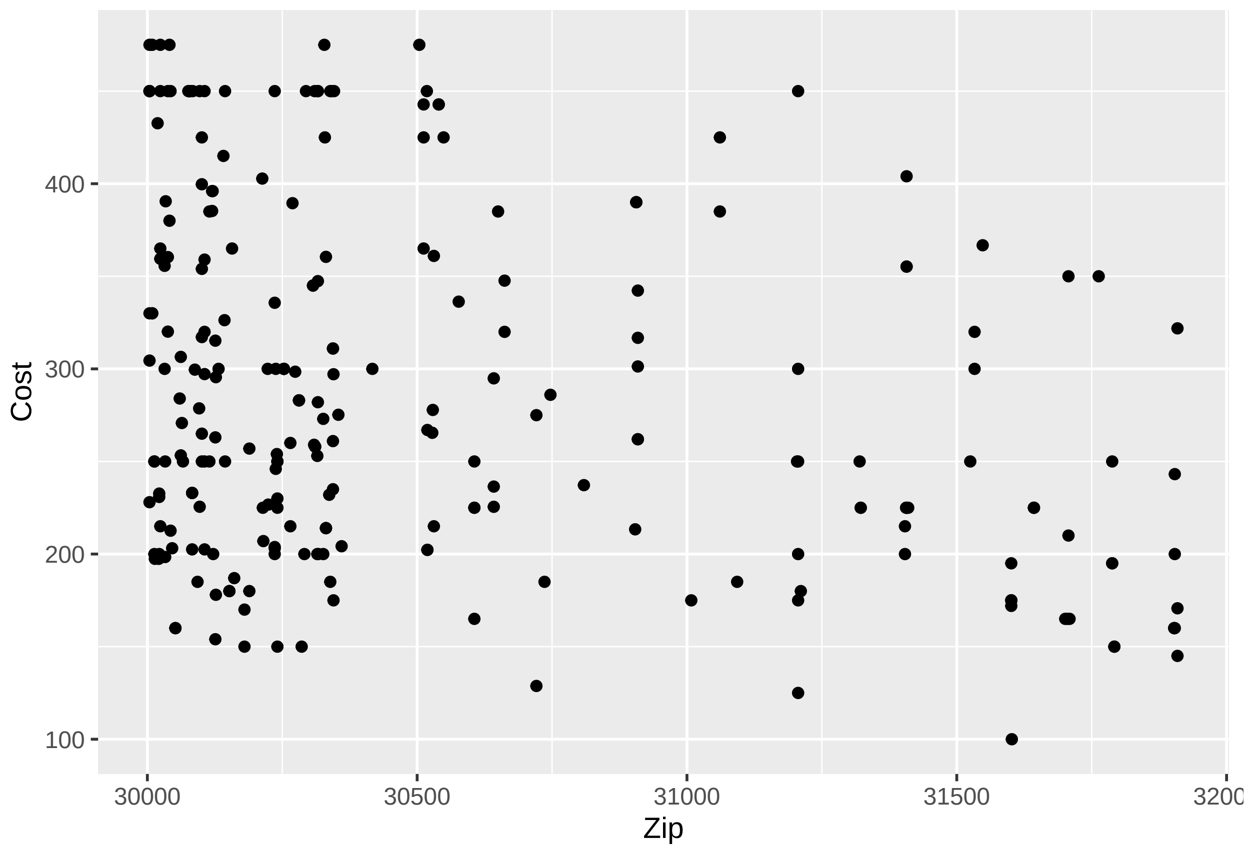
If you imagine the upper layer divided by 3, you see they'd fall in line with the bottom layer. For now, we will pick a reasonable cutoff for DSP, say, $90.

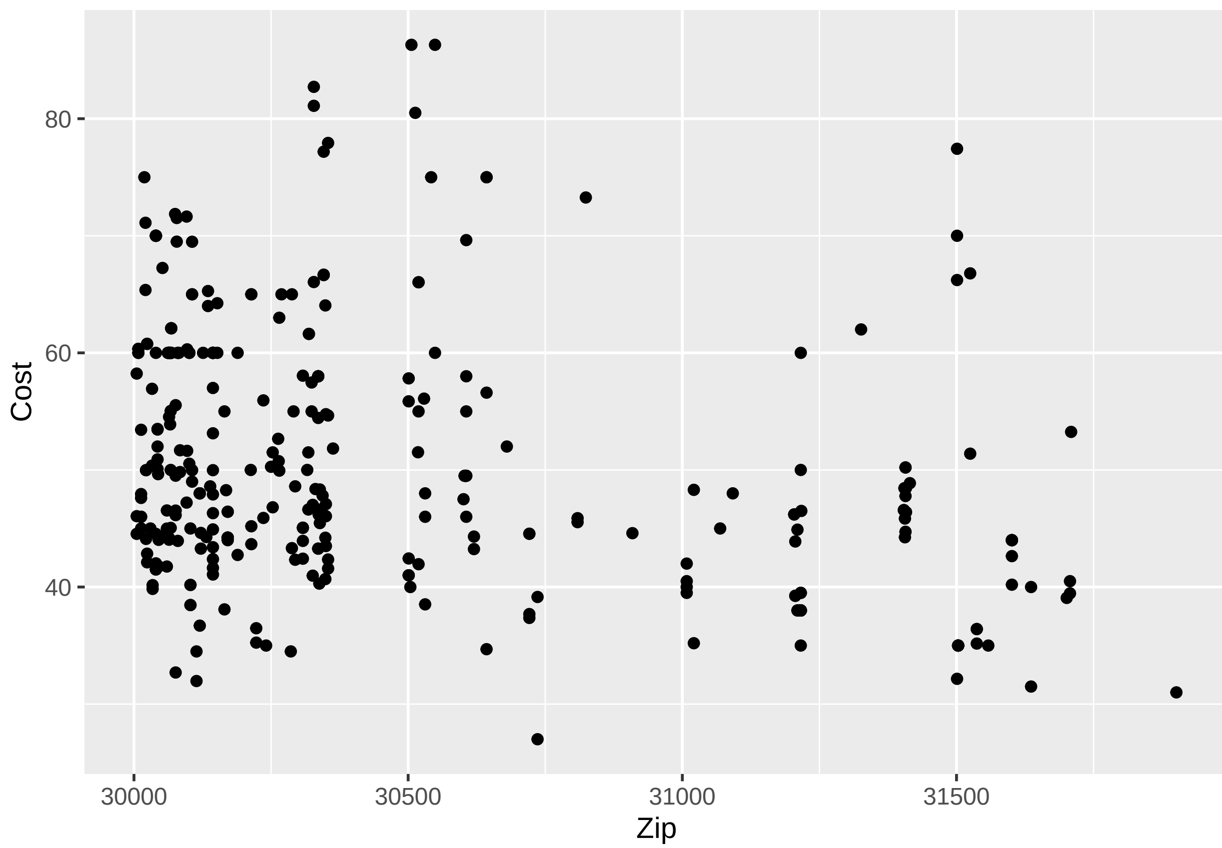
Next, we define and look at Haul (H) costs



Once again, there is a layer around $150 - $450. Other errors, like sometime including DSP costs with the Haul field creates some higher, unreasonable data points. We can select a cutoff of approximately $500.

We can re-run the graphs after defining the cutoff amounts and re-check the graphs. Looks good:


Cost
400
300
200
100
30000
30500
31000
31500
3200
Zip



Now we can run a very simple GLM to calculate relativities of costs from a base cost, across counties. Of course, we can add material type, container size, container type, schedule type, etc, - however, for this heuristic, we will simply look at and calculate the county relativities.

From the summary of the GA file, we use the most common county (highest counts) as the base or anchor. It will by definition have a relativity factor of 1.000. From the summary, we see the county with the most data is 'Fulton'. It's base cost will be 'e' raised to the model coefficient.

```
##   StartEffDate          EndEffDate               Cost           City
##  Min.   :2021-01-04   Min.   :2021-01-11   Min.   :100.0   Length:259
##  1st Qu.:2021-03-31   1st Qu.:2021-04-03   1st Qu.:205.6   Class :character
##  Median :2021-05-25   Median :2021-05-28   Median :261.0   Mode  :character
##  Mean   :2021-05-26   Mean   :2021-06-04   Mean   :285.9
##  3rd Qu.:2021-07-19   3rd Qu.:2021-07-27   3rd Qu.:357.3
##  Max.   :2021-10-19   Max.   :2021-10-30   Max.   :475.0
##
##       County        State           Zip         Svc.Type   Mat       Sched
##  Fulton  : 45   GA     :259   Min.   :30004   DSP:  0   C&D:  0   OC :259
##  Cobb    : 26   AK     :  0   1st Qu.:30101   H  :259   MXR:  0   SOC:  0
##  De Kalb : 23   AL     :  0   Median :30309             SSR:  0
##  Gwinnett: 19   AR     :  0   Mean   :30544             T  :259
##  Chatham : 11   AZ     :  0   3rd Qu.:30905             WD :  0
##  Muscogee:  9   CA     :  0   Max.   :31909
##  (Other) :126   (Other):  0
##      TempOrPerm  Container      Size
##  Permanent:  1   CMP:  0   YRDS-30:259
##  Seasonal :  0   OT :259   UNK    :  0
```

```
##  Temporary:258             YRDS-10:  0
##                            YRDS-12:  0
##                            YRDS-13:  0
##                            YRDS-14:  0
##                            (Other):  0
```

Now we run the GLM and get the coefficient, and the county relativities. We simply add the county estimates to the intercept term before exponentiating to get the estimated cost for a Haul in that county.

```
##
## Call:
## glm(formula = Cost ~ County, family = Gamma(link = "log"), data = dfgah)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.5927  -0.1857   0.0000   0.1016   0.6955
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       5.76292    0.04226 136.355  < 2e-16 ***
## CountyBaldwin     0.24097    0.20488   1.176 0.240919
## CountyBartow      0.20943    0.16906   1.239 0.216839
## CountyBibb       -0.27708    0.10878  -2.547 0.011603 *
## CountyBrooks     -0.34682    0.16906  -2.051 0.041503 *
## CountyCamden      0.14179    0.28665   0.495 0.621384
## CountyCarroll    -0.68774    0.20488  -3.357 0.000941 ***
## CountyCatoosa    -0.54256    0.28665  -1.893 0.059811 .
## CountyChatham    -0.19545    0.09536  -2.050 0.041689 *
## CountyChattooga -0.10693    0.28665  -0.373 0.709523
## CountyCherokee   -0.17197    0.14792  -1.163 0.246381
## CountyClarke     -0.38648    0.14792  -2.613 0.009655 **
## CountyClayton    -0.12943    0.10878  -1.190 0.235510
## CountyCobb       -0.09571    0.06984  -1.370 0.172062
## CountyCoffee     -0.02635    0.20488  -0.129 0.897812
## CountyColquitt   -0.40006    0.16906  -2.366 0.018901 *
## CountyColumbia   -0.29397    0.28665  -1.026 0.306326
## CountyCoweta     -0.29275    0.20488  -1.429 0.154580
## CountyDe Kalb    -0.02706    0.07267  -0.372 0.709989
## CountyDougherty -0.35799    0.14792  -2.420 0.016397 *
## CountyDouglas    -0.46460    0.28665  -1.621 0.106613
## CountyEvans      -0.05914    0.28665  -0.206 0.836765
## CountyFayette    -0.15042    0.16906  -0.890 0.374639
## CountyFloyd      -0.53181    0.28665  -1.855 0.065009 .
## CountyForsyth     0.29504    0.20488   1.440 0.151400
## CountyFranklin    0.04770    0.20488   0.233 0.816141
## CountyGilmer      0.33020    0.20488   1.612 0.108590
## CountyGlynn      -0.24146    0.28665  -0.842 0.400587
## CountyGreene     -0.23238    0.16906  -1.375 0.170783
## CountyGwinnett    0.07957    0.07757   1.026 0.306219
## CountyHabersham -0.09996    0.20488  -0.488 0.626165
## CountyHall        0.40040    0.28665   1.397 0.163992
## CountyHenry      -0.08788    0.12322  -0.713 0.476556
## CountyHouston    -0.54256    0.28665  -1.893 0.059811 .
## CountyJackson     0.09905    0.20488   0.483 0.629298
```

```
## CountyLee         0.09502    0.28665   0.331 0.740631
## CountyLiberty   -0.24146    0.28665  -0.842 0.400587
## CountyLowndes   -0.65495    0.12322  -5.315 2.79e-07 ***
## CountyMorgan     0.19033    0.28665   0.664 0.507464
## CountyMuscogee  -0.50963    0.10353  -4.923 1.76e-06 ***
## CountyNewton    -0.47763    0.28665  -1.666 0.097202 .
## CountyPaulding   0.12319    0.16906   0.729 0.467042
## CountyPeach     -0.59813    0.28665  -2.087 0.038170 *
## CountyPickens    0.02490    0.28665   0.087 0.930865
## CountyRichmond   0.00112    0.10353   0.011 0.991379
## CountyRockdale  -0.31045    0.16906  -1.836 0.067766 .
## CountySpalding  -0.18942    0.20488  -0.925 0.356322
## CountyStephens   0.05512    0.28665   0.192 0.847718
## CountySumter    -0.65697    0.28665  -2.292 0.022937 *
## CountyThomas    -0.75228    0.20488  -3.672 0.000308 ***
## CountyTroup     -0.34017    0.12322  -2.761 0.006296 **
## CountyUnion      0.25551    0.16906   1.511 0.132239
## CountyUpson     -0.75228    0.28665  -2.624 0.009340 **
## CountyWalton    -0.68774    0.16906  -4.068 6.78e-05 ***
## CountyWhite     -0.18138    0.28665  -0.633 0.527607
## CountyWhitfield -0.45524    0.20488  -2.222 0.027389 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.08038131)
##
##     Null deviance: 29.057  on 258  degrees of freedom
## Residual deviance: 16.322  on 203  degrees of freedom
## AIC: 3036.4
##
## Number of Fisher Scoring iterations: 4
```

If we want to only use model estimates for counties with asterisks, even better, since they indicate significance. Let's estimate the costs for both the base (Fulton) and Walton county:

```
## [1] "Base: 318.27634027549"
```

```
## [1] "Walton 160.000989565648"
```

As we can see, the cost for a Haul in Walton are nearly half of what Fulton county could expect. Note that Fulton does not show up in the list since its estimate would be '0', since it is being used as the reference or 'base' county.

We can repeat or loop this process for every state and its collection of counties, or at least those counties which are significant or credible. Tables of costs can be built and uploaded into automated systems and results can be compared to costs that are ultimately procured in the open marketplace.

Monitoring the performance of the cost estimates is crucial for building trust and confidence in the model. It is also great for learning important feedback for continually refining and tweaking the model.

Implications

The implications of implementing an automated system to predict costs for customers is huge. For other systems with similar anticipated speed-ups, customer satisfaction has been shown to jump by over 20 points.

It also helps us to reallocate resources that have been over-taxed by repetitive tasks and hand-offs to other departments. The biggest time reduction, however, is that of not having to contact a vendor to procure a cost. This step alone can save DAYS in the process of gathering costs and ultimately, computing a price.

Limitations

The most visible limitations to this process is that of coverage. Reducing the scope of modeling to areas (zips, counties, etc.) where we have a credible amount of data makes sense. This is because we will have the most data in the areas we are most likely to have repeat business.

Concluding Remarks

The issue WM faces regarding cost predictions is one ultimately of expediency. The ability to predict garbage service costs quickly and accurately is vital to better serving our customers and enhancing the bottom line. The ability to take raw data and ultimately build implementable business solutions is a top goal for data science in not just the waste industry, but for all industries.