# Michael Ersevim – Final Project, DSC530

**Bellevue University, Fall 2021**

## Introduction/Background

I work for Waste Management (WM), which collects trash for millions of residential and commercial clients across the United States. There are physical locations all over the country where we are responsible for picking up the trash, however, for many reasons, there are plenty of locations where it is unfeasible for us to pick up the trash and we sub-contract out the work. In fact, almost 30% of the services for our commercial clients are done on our behalf by a third-party hauler.

**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.

However, 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.

## Statistical/Hypothetical questions

* How much does a particular service cost in a particular zip code?
* How much variance is there in the marketplace, all things being equal regarding a service?
* How much does geography influence the costs of picking up trash?
* What variables are most strongly correlated with costs?
* What trends are happening in the marketplace?
* What commonplace errors are contained in the data?
* What is the appropriate length of time of a window should data be included?

In order to answer most of these questions, several steps must be taken. First of all, the business understanding is that service to a customer should comprise several variables:

1) The size of the dumpster in cubic yards? (2, 4, 6, 8 being most common sizes)

2) The type of waste material (Trash, Cardboard, Food waste, Metal, etc)

3) The frequency (in times per week) the trash is picked up

4) The State/zip code in which the service is being provided

5) The type of cost of the services (this is the dependent variable to model)

## Data

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.

## Importing and cleaning the data

The first step in importing the data is to run the SQL separately and save the output as a .csv file for easy importing into a Panda data frame.

The **Select** statement:

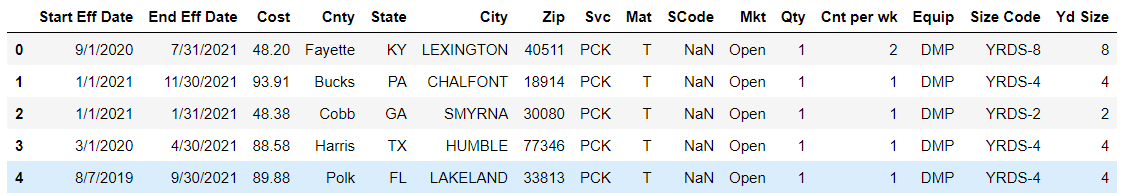
cast([effectiveDT] as date) as 'Start Eff Date' ,cast([lastEffectiveDT] as date) as 'End Eff Date', ,[specCost] as Cost ,[locationCountyName] as Cnty ,[locationStateProvinceCode] as State, ,[locationCityName] as City ,[locationPostalCode] as Zip ,[serviceComponentTypeCode] as Svc, [materialTypeCode] as Mat ,[sensitivityCode] as SCode ,[specIsClosedMarketType] as Mkt, ,[quantity] as Qty ,[serviceCountPerWeek] as 'Cnt per wk' ,[equipmentTypeCode] as Equip, [equipmentSizeCode] as 'Size Code' ,[sizeMeasurement] as 'Yd Size'

Here are **Where** statements for filtering the fields from the WM ACORN database:

Where [lastEffectiveDT] > '2021-07-01' and [sensitivityCode] is Null and [vendorType] <> 'WM' and [specIsClosedMarketType] = 'Open' and [serviceComponentTypeCode] = 'PCK’ and [equipmentTypeCode] = 'DMP' and [specCost] > 5 and [locationStateProvinceCode] not in ('AB', 'BC', 'MB', 'NB', 'NL', 'ON', 'PE', 'QC', 'SK', 'NT', 'NU', 'YT') and [sizeMeasurement] in ('2', '4', '6', '8') and [serviceCountPerWeek] in ('1', '2', '3', '4', '5', '6', '7')

## Data import profile and characteristics

Below is a picture of the top of the .csv file that gets imported into Python as a Panda data frame:

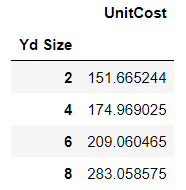
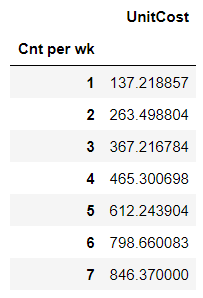


Depending on the date ranges I am interested in, it will yield different row counts as output. I purposely chose a fairly recent window of dates in order to keep the file size small for working with as a heuristic.

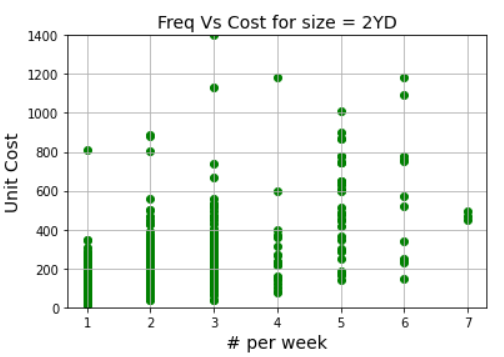
Each row will have the date range, cost of service, the geographic information of the service location, and some explanatory variables: cost may depend on material type, how many pickups per week (‘Cnt per wk’) and the size (‘Yd size’ = capacity in cubic yards) of the dumpster being emptied. Some of these explanatory values have already been filtered for certain common values as found above in the Where statements. This reduces certain outlier types and constrains the output to a very focused dataset.

## Outcome of EDA

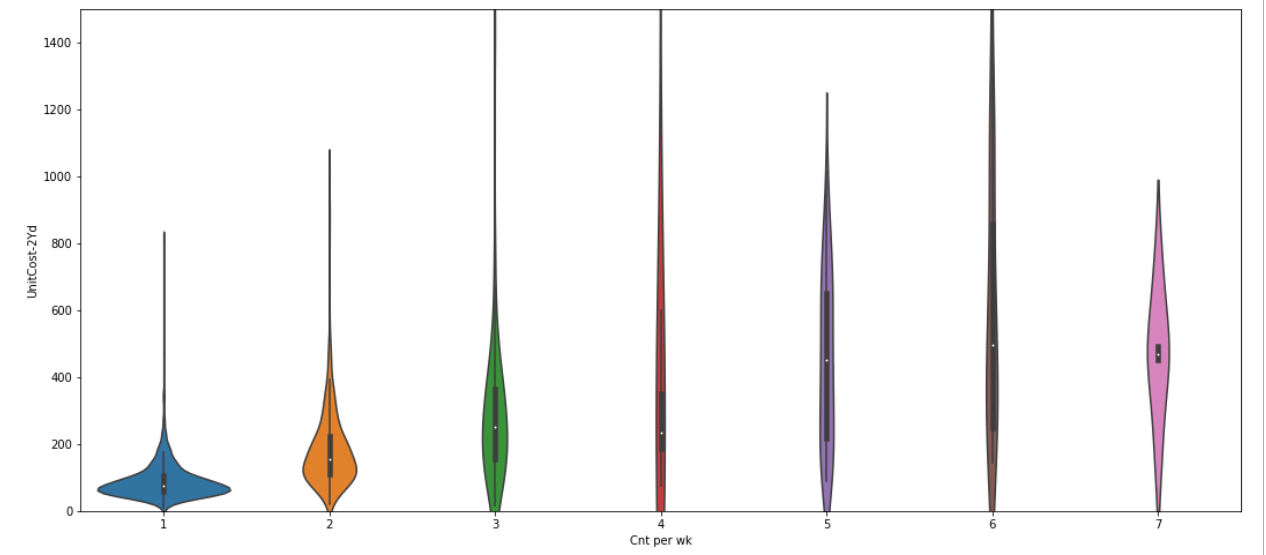
Here are some high-level, univariate looks at the unit cost variable (cost/Qty) by average cost by size of container, and by times per week collected.

Below, we see the distribution of costs for a ‘2-Yard’ container across times per week collected. Many of the values are compressed below the $500 amount, however, you can see the expected upward trend as collections per week increases. Much of the variance around cost is likely attributable to material type and zip code.



The violin plot below better shows the obscured amount of cost densities in the $80 - $450 range of unit cost. Notice that 1 and 2 times per week are quite pronounced, whereas the higher frequencies are skinnier (more uniformly distributed) across the cost range.



## Opportunities and variables missing from analysis

* What is the incremental cost to a TPH for their trash truck picking up trash in a particular client’s zip code? WM could never know what other stops or in what order they do their routes.
* While it is not that hard to calculate the straight-line distance from a zip code centroid to a landfill, transfer station, or other waste facility (when the Lat/Long is known), it is harder to know the **route** a hauler will take, or if they have to utilize a different landfill (or facility) than the one closest to them.
* Price fixing amongst competitors is not legal, but likely happens from time to time. Trying to decipher what the fair market price of a service is when these costs may be routinely manipulated is difficult.