Python code for this project can be found in the CAT PCC folder on GitHub.

★ The aim of this vignette is to give a brief overview of some of the clustering approaches that I used to cluster data from the parts.cat.com website. I looked at the following parameters for each visitor:

- Average Quantity per Purchase.
- Median Hit Total.
- Median Page Views.
- Median Revenue per Session (in $US).
- Median Time Between Sessions (in seconds).
- Minimum Time Between Sessions (in seconds).
- Number of Sessions.
- Percent Sessions with Purchase.

★ We cluster based on these parameters using the `sklearn.cluster` module in Python. The clustering is unsupervised. One thing that we are particularly interested in is relationship between revenue and the median hit total. The hit total for a session acts as a metric for measuring the amount of interaction with the site. We commonly use this in place of 'time on site' due to the fact that time on site can by heavily inflated by users who enter the site and then do not interact with it or forget about it, yet still have the site pulled up.

★ The clustering methods that we use are as follows:

- k-Means clustering. Here we use three clusters. k-Means works best in settings where we want roughly equal cluster sizes.
- Birch clustering. This method is especially useful for large data sets with outliers.
- Ward hierarchical clustering. This is an agglomerative clustering method. The connectivity for clustering used the 10-nearest neighbors. We cluster into three and four clusters.

★ Speaking generally, the visitors seem to cluster into one of three clusters:

- **Cluster 1:** Visitors with little median time between sessions, and moderate to high amounts of median revenue per session. Hit number varies widely in this group. These visitors are what I would consider the prime revenue generators. They usually seem to generate substantial amounts of revenue and they have a fast turn around between sessions.

  Efforts could be made to retain these visitors and perpetuate their purchase habits long term. It is also quite likely that this cluster contains a significant amount of visitors who are parts dealers.

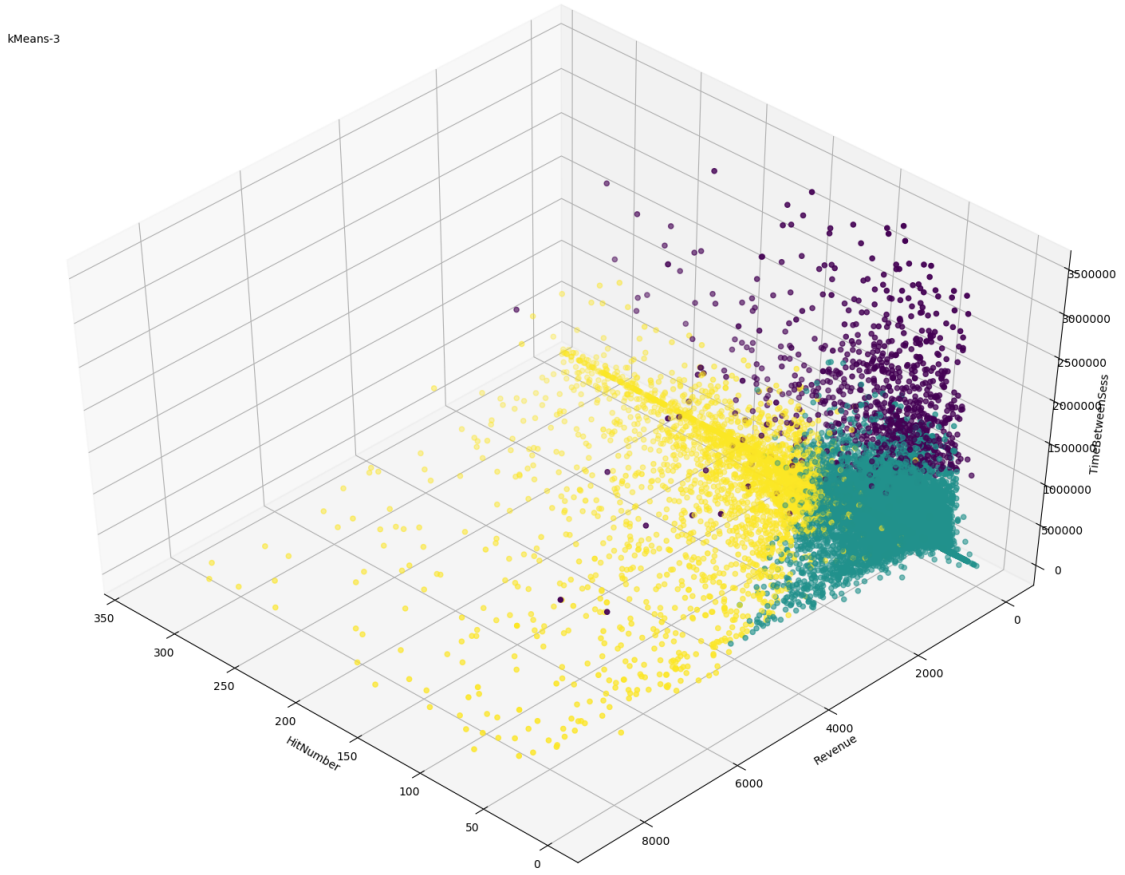
- **Cluster 2:** Visitors who produced relatively small amounts of revenue and occasionally had moderate amount of time between sessions. These visitors usually (except in Birch clustering) also have few median hit totals. Ward clustering with four clusters splits this cluster into two clusters, with median revenue per session being the distinguishing factor. This cluster contains visitors who do not interact with the site much or produce much revenue. Many of these visitors may be single time visitors who never purchase anything.

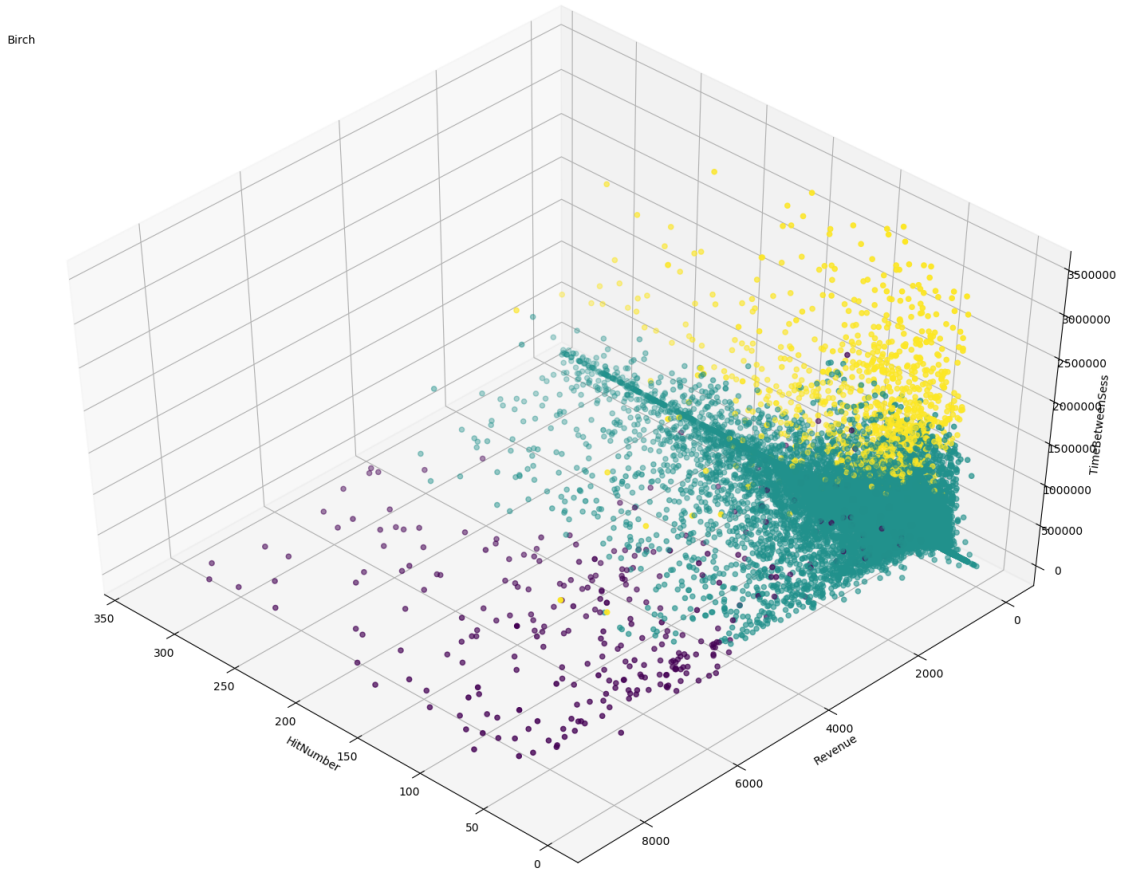
  I suggest that stronger efforts be made to look into why these visitors are reaching the site, then never purchasing or returning. Did they come by accident, or where they unable to complete their visit because the site was too frustrating to use?

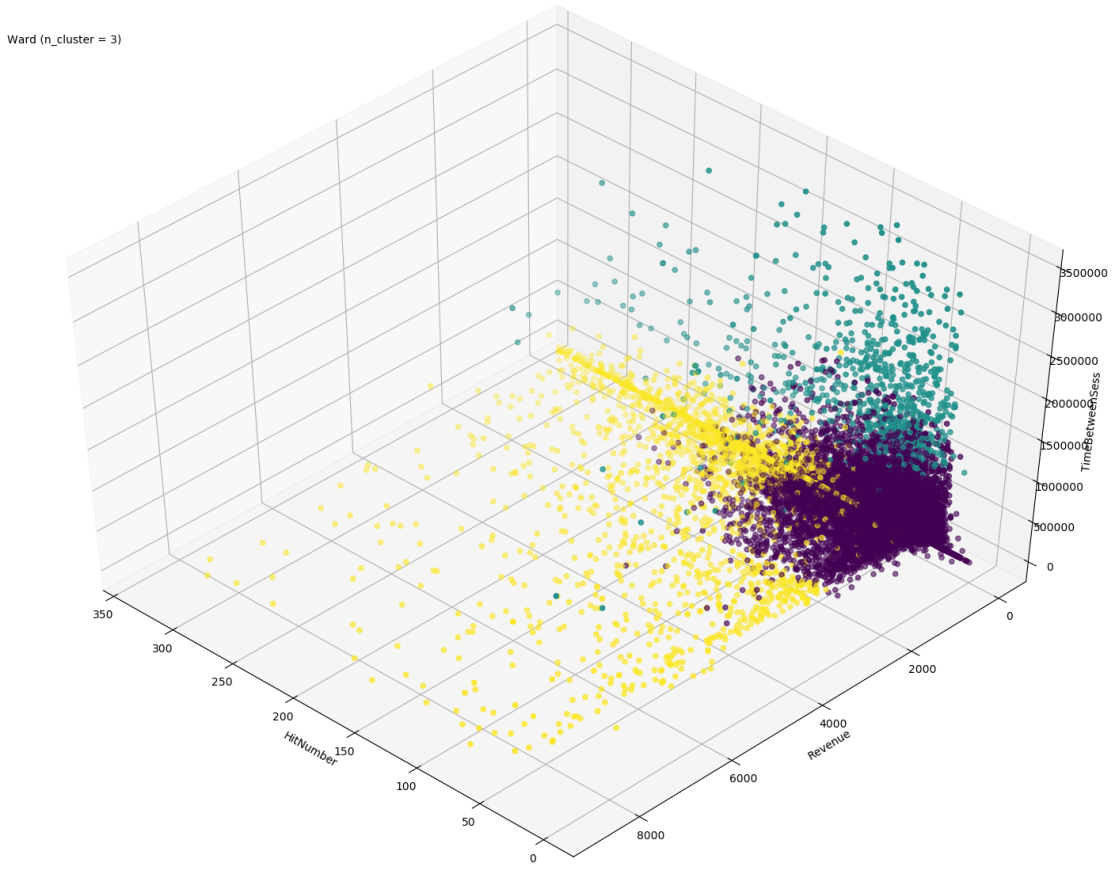
- **Cluster 3:** Visitors that produced very little revenue and usually had a large amount of time between sessions. The largest distinguishing factor for these visitors was the median time between sessions. Most of these visitors also produced total revenue less than $2000. These visitors may be individual buyers who own few machines and do not need to service their machines very often.

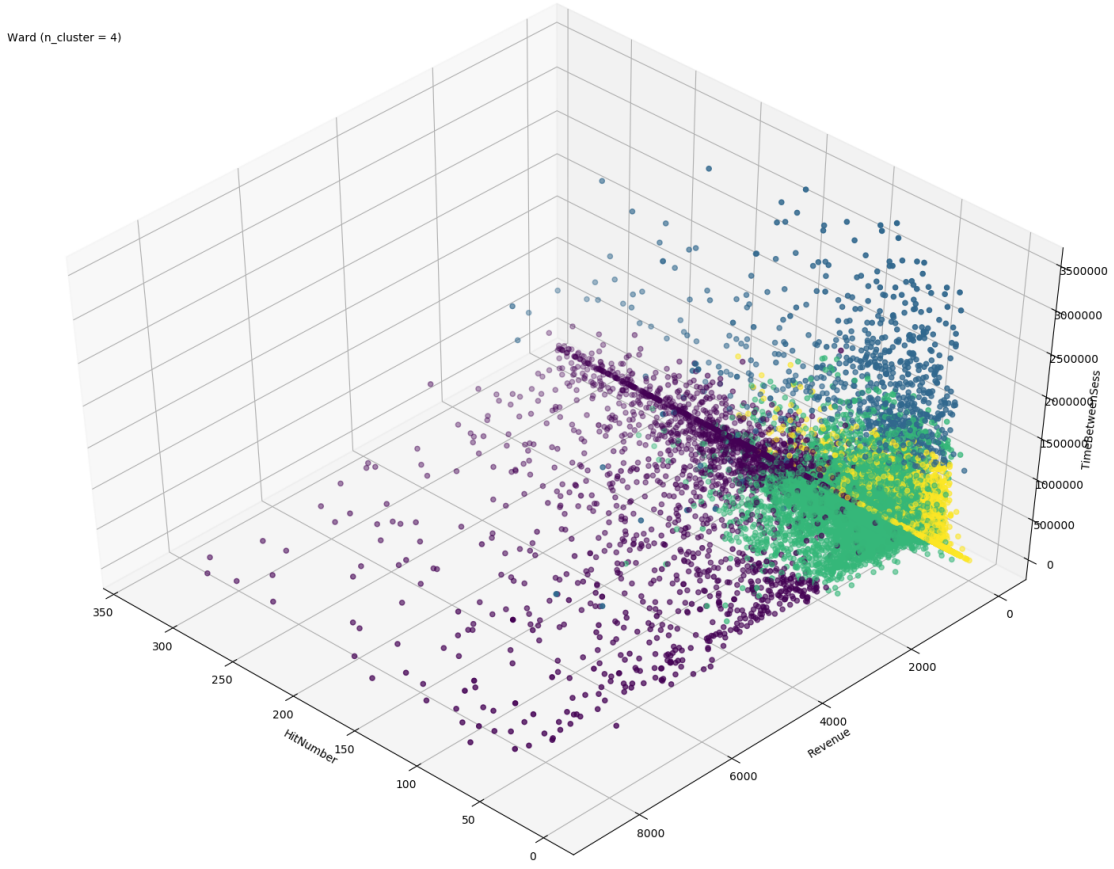
  I suggest that these visitors be targeted with incentives to come to the site more often and assess if they can use some of the many maintenance services that Caterpillar provides.

★ The plots of the four clustering approaches are below:


kMeans-3
HitNumber
Revenue
TimeBetweenSess
0
50
100
150
200
250
300
350
0
2000
4000
6000
8000
0
500000
1000000
1500000
2000000
2500000
3000000
3500000


Birch
HitNumber
Revenue
TimeBetweenSess
0
50
100
150
200
250
300
350
0
2000
4000
6000
8000
0
500000
1000000
1500000
2000000
2500000
3000000
3500000


Ward (n_cluster = 3)
HitNumber
Revenue
TimeBetweenSess


Ward (n_cluster = 4)
HitNumber
Revenue
TimeBetweenSess

★ Some final notes on principal components analysis (PCA) to be aware of:

- We found the first four principal components. These four PC explain about 81% of the variance in the data. Broken down by individual PC, the percentage of explained variance is as follows:

$$[27.2\%, \quad 23.9\%, \quad 16.9\%, \quad 12.6\%]$$

The first four PCs are as follows:

Average Quantity Per Purchase:	$\begin{pmatrix} 0.281 \end{pmatrix}$	$\begin{pmatrix} 0.017 \end{pmatrix}$	$\begin{pmatrix} 0.482 \end{pmatrix}$	$\begin{pmatrix} -0.039 \end{pmatrix}$
Median Hit Total Per Session:	$\begin{pmatrix} 0.589 \end{pmatrix}$	$\begin{pmatrix} 0.083 \end{pmatrix}$	$\begin{pmatrix} -0.368 \end{pmatrix}$	$\begin{pmatrix} 0.026 \end{pmatrix}$
Median Page View Per Session:	$\begin{pmatrix} 0.560 \end{pmatrix}$	$\begin{pmatrix} 0.081 \end{pmatrix}$	$\begin{pmatrix} -0.435 \end{pmatrix}$	$\begin{pmatrix} 0.0126 \end{pmatrix}$
Median Revenue Per Session:	$\begin{pmatrix} 0.359 \end{pmatrix}$	$\begin{pmatrix} 0.012 \end{pmatrix}$	$\begin{pmatrix} 0.464 \end{pmatrix}$	$\begin{pmatrix} -0.276 \end{pmatrix}$
Median Time Between Session:	$\begin{pmatrix} -0.084 \end{pmatrix}$	$\begin{pmatrix} 0.700 \end{pmatrix}$	$\begin{pmatrix} 0.045 \end{pmatrix}$	$\begin{pmatrix} 0.062 \end{pmatrix}$
Minimum Time Between Session:	$\begin{pmatrix} -0.071 \end{pmatrix}$	$\begin{pmatrix} 0.704 \end{pmatrix}$	$\begin{pmatrix} 0.020 \end{pmatrix}$	$\begin{pmatrix} -0.040 \end{pmatrix}$
Number of Sessions:	$\begin{pmatrix} 0.081 \end{pmatrix}$	$\begin{pmatrix} -0.016 \end{pmatrix}$	$\begin{pmatrix} 0.150 \end{pmatrix}$	$\begin{pmatrix} 0.956 \end{pmatrix}$
Percent Sessions With Purchase:	$\begin{pmatrix} 0.337 \end{pmatrix}$	$\begin{pmatrix} 0.019 \end{pmatrix}$	$\begin{pmatrix} 0.450 \end{pmatrix}$	$\begin{pmatrix} 0.037 \end{pmatrix}$

From these PCs, it seems that the largest amount of variation is explained by a differences in the hit and page view medians, as well as the revenue and purchase percentages. Secondarily, the gap times between sessions factor strongly in explaining the variation in the data.