**Executive Summary**

Overview of Work

For this project, I applied the K-Means clustering algorithm to the U.S. Census ACS PUMS dataset, which was acquired from the Registry of Open Data on AWS. The data was provided in Turtle RDF format and that added an extra layer of complexity. I had to extract and preprocess it from RDF before doing the clustering, making sure that the information was ready to be analysed. I then used the K-Means algorithm to spot the patterns and group the data points that had similar information. Along the way, I normalised the data, I evaluated the results using different techniques such as the Elbow Method to find out the best value for K (the number of clusters that confirmed the initial assumptions), and I measured the validity of the K Means analysis using a different metric like the Silhouette Score. This was not about running an algorithm, but about overcoming the structural challenges of the data format to extract insights.

**CHALLENGES FACED**

There was plenty of friction on the way, as one would expect with an RDF dataset:

**Compiling the Data:** RDF files are arranged in a proprietary way – my single task here, after gathering a list of objects, was to extract the data and filter out all the other stuff. Time-consuming and awkward – translating RDF objects into numbers was one of the hardest aspects of the work.

**Preprocessing:** The biggest issue here was to get the data into an acceptable format to be used for clustering. Persisting a pivot, dealing with missing values and normalising the data was tedious. Corral wants to apply PCA to visualise clusters: another of the technical roadblocks were applying PCA to visualise clusters. The data set was not provided enough features enough times for this to be effective. This led to a recalibration of my thinking on how to proceed if PCA’s not an option.

**HOW I OVERCAME THEM**

**Working with RDF Data:** I’ve had to define my custom predicates of interest to filter just the triples that I need for me to coerce to a numeric value to which I can cluster.

**Iterative Testing and Refinement:** Therewere two important secrets to make this work: Breaking down the project into smaller tasks, and performing a test at each step to evaluate the progress. The Elbow Method can be used to determine the optimal number of clusters for an analysis. Iteratively, the parameters of the analysis were tweaked to produce better and better results.

**Data Normalization**: In order to apply K-Means clustering, I have to normalize the data using Standard Scaler, otherwise the results would be all over the place.

**INTERESTING FINDINGS**

With the technical challenges behind me, interesting patterns started to emerge:

Secondly, **Cluster Insights:** The K-Means algorithm made a good effort at separating the data into clusters, especially based on housing values, income adjustment and ratings (see the resulting three-cluster model below). We could pick up a story in each of the clusters and draw conclusions on how some socioeconomic factors seamed together. Thirdly, Evaluation Results: the Silhouette Score and the Davies-Bouldin Index indicated that a model with three clusters was the most appropriate for describing the data. It was very rewarding to see the clusters verify themselves with the precise evaluation metrics.

**Data-Driven Insights:** This would be obvious if anyone bothered to cluster the data, and sure enough there are stark differences when we lump people together into groupings defined by housing, income and other key variables, which should be significant for policy makers seeking to comprehend regional variation.

**CONCLUSION**

Along the way, I’ve spent a fair share of time staring at my screen, scratching my head and muttering Googly absurdities in a mad riddle-trying-to-be-solved manner. But most of the time was spent having fun, getting better at RDF data issues, K-Means clustering, lingo, visualisation, data-analysis, machine-learning, trial-and-error, and cracking riddles. The fun that commences once things finally begin to reveal themselves after the third or fourth iteration of trying is like crack cocaine, period.