UCD Michael Smurfit Graduate Business School

A data-driven approach to locate electric car charging stations to support strategic power network expansion

Shweta Soni, B.Sc., and Siva Thirumavalavan, B.Tech.

A Capstone submitted to University College Dublin in part fulfilment of the requirements of the degree of M.Sc. in Business Analytics

Michael Smurfit Graduate School of Business, University College Dublin

September 2021

Supervisor: Professor Annunziata Esposito Amideo

Head of School: Professor Anthony Brabazon

Dedication

It is our genuine gratefulness and warmest regard that we dedicate this work to KPMG Ireland, our academic professors, our family, friends, and everyone working towards making this world a better place.

Table of Contents

List of Figures **v**

List of Tables **vi**

List of important abbreviations **vii**

Executive Summary **viii**

Chapter 1 - Introduction **1**

1.1 History of Electric vehicles 1

1.2 The current scenario 2

1.2.1 World Stats 2

1.2.2 Ireland Stats 5

1.3 Why are people hesitant to purchase electric vehicles? 6

1.4 Business Objective 7

Chapter 2 - Literature Review **9**

2.1 Unsupervised learning - clustering 9

2.1.1 Hierarchical 10

2.1.2 K means 11

2.1.3 ANN 12

2.1.4 Evolutionary algorithms 12

2.1.5 OPTIC algorithm 13

2.1.6 Mixed models in network 14

2.1.7 Spectral Clustering 14

2.2 Time Series Forecasting 18

2.2.1 ARIMA 18

2.2.2 Hybrid Model 20

Chapter 3 - Approach **21**

Chapter 4 - Success criteria **22**

4.1 Success metrics for EDA 22

4.2 Success metrics for clustering 22

4.3 Success metrics for dashboard 23

Chapter 5 - Exploratory Data Analysis 24

5.1 Data description 24

5.2 Data transformations 27

5.3 Assumptions 29

5.4 External Data Sources 30

Chapter 6 - Grouping the households 32

6.1 Clustering 32

6.2 Profiling 34

6.3 Summary 38

Chapter 7 - Dashboard 40

Chapter 8 - Proposed action 44

Conclusion 48

References 50

List of Figures

FIGURE 1: EV SALES TREND 2
FIGURE 2: PRIVATE CHARGERS- COUNTRY-WISE 3
FIGURE 3: PUBLIC CHARGERS- COUNTRY-WISE 4
FIGURE 4: PURCHASE CONCERNS FOR EVS 7
FIGURE 5: CURRENT CHARGE POINT MAP 8
FIGURE 6: CLUSTERING REVIEW 10
FIGURE 7: SEAI DATA 24
FIGURE 8: BER RATING IMPUTATION 29
FIGURE 9: DATA TRANSFORMATION 31
FIGURE 10: K MEANS SCREE PLOT 32
FIGURE 11: APPROVED RATE VS HOUSE VINTAGE 33
FIGURE 12: APPROVAL RATE VS TOTAL MEASURES 34
FIGURE 13: SNIPPET OF DASHBOARD -FIRST VIEW 40
FIGURE 14: SNIPPET OF THE DASHBOARD - COUNTY SELECTION 41
FIGURE 15: SNIPPET OF THE DASHBOARD - BER CHART 41
FIGURE 16: SNIPPET OF THE DASHBOARD - CLAIMS INFO 42
FIGURE 16: SNIPPET OF THE DASHBOARD - HS FUEL 42
FIGURE 18: CLUSTER ASSOCIATION WITH THE COUNTIES 44
FIGURE 19: SCOPE OF EACH COUNTY 46

List of Tables

TABLE 1: SUMMARY OF CLUSTERING LITERATURE 17
TABLE 2: ADVANTAGES AND LIMITATIONS OF THE CHOSEN CLUSTERING APPROACH ... 21
TABLE 3: SEAI DATA DESCRIPTION 27
TABLE 4: DERIVED COLUMNS 28
TABLE 5: CENSUS DATA DESCRIPTION 30
TABLE 6: CLUSTER ANALYSIS: HOUSE VINTAGE 35
TABLE 7: CLUSTER ANALYSIS - BER RATING 36
TABLE 8: CLUSTER ANALYSIS - HS FUEL 36
TABLE 9: SCHEME COUNT 36
TABLE 10: CLUSTER ANALYSIS - MEASURES 37
TABLE 11: CLUSTER ANALYSIS - SUBMISSIONS 37
TABLE 12: CLUSTER ANALYSIS - APPROVAL RATE 38
TABLE 13: CLUSTER PROFILES 39
TABLE 14: COUNTY-WISE PROPORTION OF HOUSE PROFILES 45

List of important abbreviations

Abbreviation	Expansion
ABT	Analytics Base Table
ACF	Autocorrelation function
ANN	Artificial Neural Network
ARIMA	Autoregressive Integrated Moving Average
BER	Building Energy Rating
BEV	Battery Powered Electric Vehicle
CAGR	Compound Annual Growth Rate
CNN	Convoluted Neural Network
EDA	Exploratory Data Analysis
ESB	Electricity Supply Board
EV	Electric Vehicles
MAPE	Mean Absolute Percentage Error
OPTIC	Ordering Points by Identifying Cluster structure
PACF	Partial Autocorrelation Function
PCA	Principal Component Analysis
PHEV	Plug-in Hybrid Electric Vehicle
SEAI	Sustainable Energy Authority of Ireland
TS	Time Series

Executive Summary

Our planet is currently facing a major climate crisis and the consumption of non-renewable resources is only making it more difficult. Essential steps towards saving our mother earth are necessary. One such step is the investment, development, and adoption of electric vehicles.

Vehicle exhaust fumes are one of the leading causes of air pollution. About one out of every five deaths is caused by air pollution, and vehicle emissions play a substantial role in this. Electric vehicles eliminate exhaust fumes by vehicles thereby preventing air pollution. It makes use of electricity as the source of fuel which reduces the level of poisonous gas leading to a much better quality of air. The cost of purchase of an EV is higher than a traditional car, however, there is no cost associated with petrol/ diesel. The only cost associated is the electricity charging cost. The maintenance cost is also very less than compared to the other cars. So, in the longer run, one is profitable owning and driving an EV. EV's deliver a smoother driving experience along with being safer than a traditional vehicle.

According to *Climate Action Plan 2019,* Ireland has a target of switching to 55% renewable power, commitment to deliver full BusConnects programme for all cities, retrofit plan for 4.5lakh homes and at least 0.5 million EVs on road with additional charging infrastructure by 2030 under the Project Ireland 2040 plan. In March 2021, the minister of Ireland stated that there were almost 30k EVs under taxation in Ireland having 50% of each BEV and PHEV by end of February 2021. Currently, the grant available for purchasing BEVs or PHEVs is 5,000 Euro for vehicles priced at 20,000 Euro or more. Based on a few estimates on charging an electric car at night and its usage, it offers a saving of approx. 1,400 Euro per year when compared to petrol prices.

Our project aims to help ESB accomplish its target by finding potential areas where charging points can be installed. The major objectives of our project are to provide insights to identify the potential geographical areas that need additional charging points by studying the current installation of household-based Better Energy communities' scheme, find factors correlated with energy consumption and identify similar households by performing clustering and present the analysis on a dashboard.

Chapter 1 - Introduction

1.1 History of Electric vehicles

The history of electric vehicles dates to the 1800s (*History of the electric vehicle*). When electricity was discovered, few people recognised its potential and began designing proof-of-concept vehicles. The entire history of electric vehicle development can be divided into four periods:

- the ‘Pre-electric car era' from the 1700s to the 1880s
- the ‘Golden Age' from 1881 to the 1920s
- the ‘Dark Ages' from the 1920s to 1969 and
- the ‘Renaissance' from the 1970s to the present

Contributions in each era laid the groundwork for future improvements.

In the ‘pre-electric car era', Anyos Jedlik designed a motor and a small electric vehicle toy in 1828, which led to Thomas Davenport developing another type of car powered by batteries in 1834. Franz Kravogl created an electric bicycle in 1867.

In the 'Golden Age' of 1881, Faure focused on increasing the power of batteries patented in 1859. ‘Electrobat’, the first commercial electric automobile, was invented in 1894, and electric taxis began to appear in 1897. Porsche developed the P1 in 1898 and the first electric hybrid in 1901.

This was followed by the 'Dark Age' and is so-called because the discovery of vast reserves of fossil fuels inspired people to opt for non-electric cars, resulting in relatively little progress during this period. AMC and Sonotone Corp. combined forces in 1959, resulting in the invention of a self-charging battery-powered vehicle.

Moving on to the ‘Renaissance' era, in 1985 a small one-person Sinclair C5 was unveiled, and in 1990, GM began mass-production of electric cars after gaining positive feedback for the Impact Electric Concept Car. In 2004, Tesla Motors was established, which resulted in the launch of the Tesla Roadster 2004. The Chevy Bolt, GM's first plug-in hybrid, was launched in 2010. Following that, other major automakers, such as Nissan and BMW, began manufacturing electric vehicles.

1.2 The current scenario

1.2.1 World Stats

In 2014, plug-in hybrids accounted for 1% of all vehicles registered in Norway. There were 25,710 pure battery electric vehicles among the 26,886 plug-in cars. In 2017, 5% of all cars in Norway were electric, rising to 10% in April 2020. According to the latest study on the electric vehicle industry from Allied Market Research, the global electric vehicle market was estimated at $162.34 billion in 2019 and is expected to hit $802.81 billion by 2027, at a CAGR of 22.6 %. Tesla and Nissan have had a competitive advantage in recent years. As Tesla Model 3 was released in 2018, it became the first plug-in vehicle to sell 1 lakh units in a single year. It replaced the Nissan Leaf as the world's best-selling plug-in vehicle in early 2020. (Allied Market Research - Global Opportunity Analysis and Industry Forecast, 2020–2027)

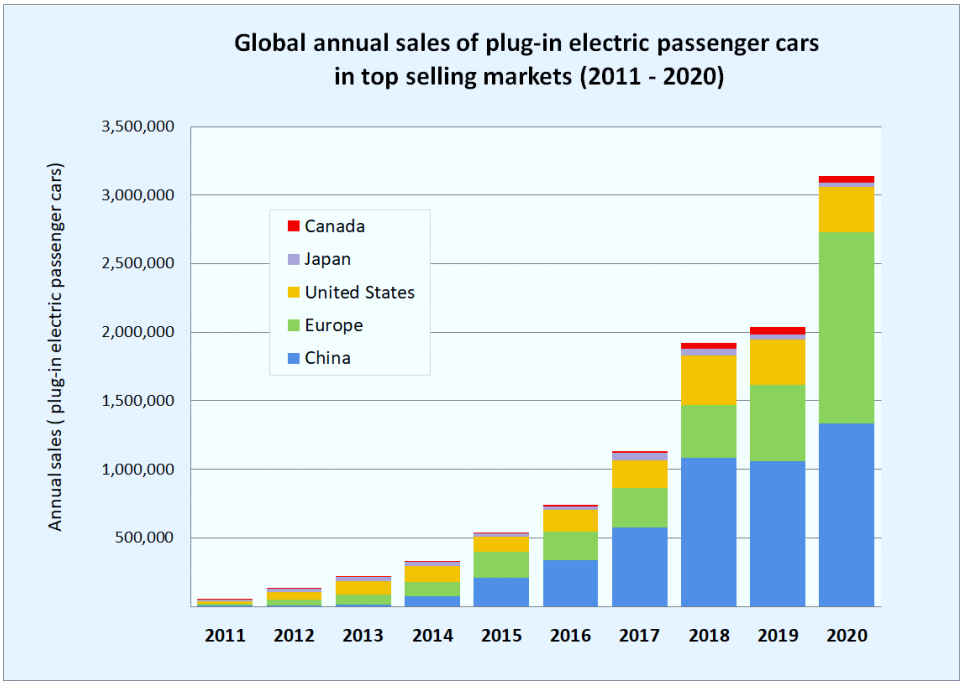

Figure 1: EV sales trend

https://commons.wikimedia.org/w/index.php?curid=47854741

According to figure 1 sales of plug-in electric vehicles have increased since 2011. In 2020, turnover would have surpassed 3 million, with a significant boost from China and Europe. About 90% of global revenues are centred in China, Europe, and the

United States. As compared to 2019, Europe's purchases have almost doubled. After being in the lead for the past few years, China seems to enter a slowing period. Germany is expected to be Europe's primary market for plug-in hybrid cars. It is influenced by factors such as regulations and incentives.

“Most charging is done at home and work, yet deploying publicly accessible charging points is outpacing electric vehicle sales” (Global EV Outlook 2020 – Analysis - IEA, 2021)

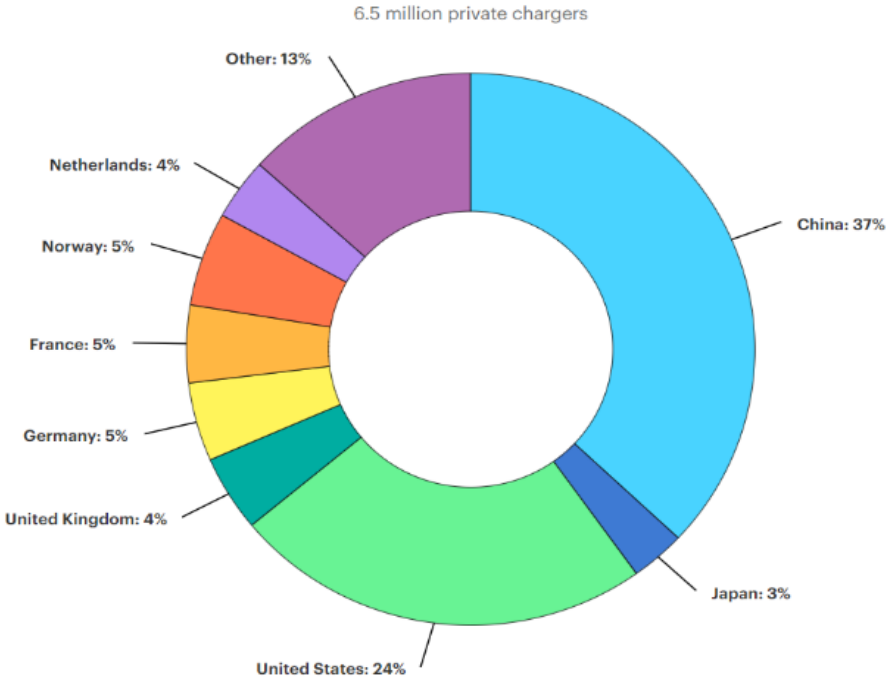

Figure 2: Private chargers- Country-wise

https://www.iea.org/reports/global-ev-outlook-2020

The expansion for EV charging by building and developing its infrastructure continues. A report mentions that in 2019, there were about 7.3 million chargers worldwide. 89% of these chargers were private and the rest 11-12% public. The dominance of private chargers is due to their accessibility, policies, and cost-efficacy.

“Electrifying heavy-duty trucks and air- and seaport operations offer opportunities for cost and emission savings. Environmental and sustainability objectives drive electric

vehicle policy support at all governance levels" (Global EV Outlook 2020 – Analysis - IEA, 2021)

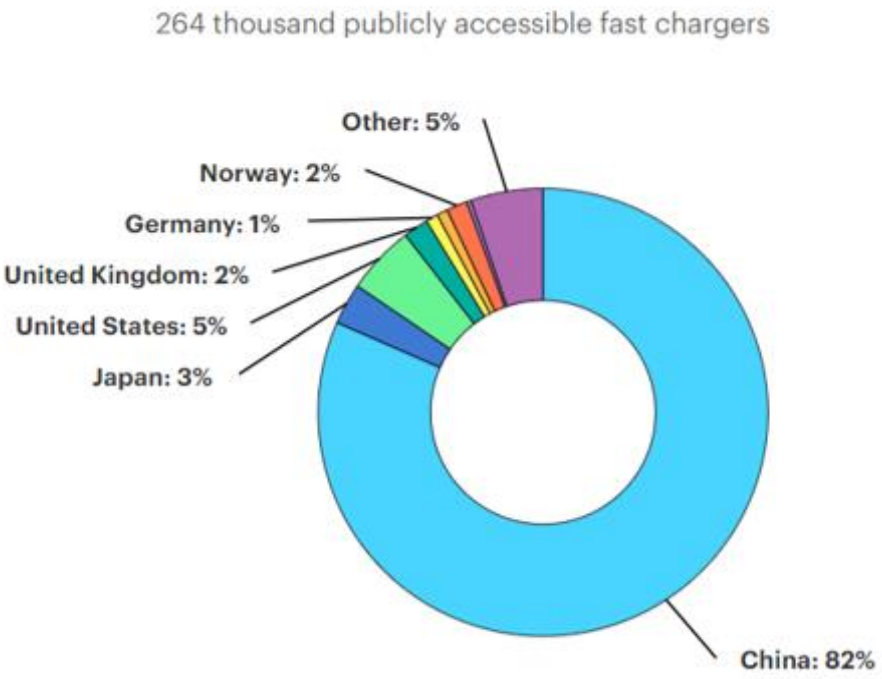

Figure 3: Public chargers- Country-wise

https://www.iea.org/reports/global-ev-outlook-2020

The benefits of electric vehicles are multiple. A few being environment friendlies reduces the air pollution by a huge extent and hence extremely beneficial in densely populated areas. It also has an overall reduction in GHG emissions, zero tailpipe emissions and provides better efficacy than other engine vehicles. The electrification of buses, trains, and trucks Urban areas. By 2050, 17 countries have announced a 100% zero-emission target.

The demand for vehicles with a longer duration of batteries and BEV relative to plug-in hybrid EVs is increasing.

1.2.2 Ireland Stats

In March 2021, the minister of Ireland stated that there were almost 30k EVs under taxation in Ireland having 50% of each BEV and PHEV by end of February 2021. Ireland aims to end fossil fuel cars sales by 2030. The government has proposed around 9.36 million EVs on Irish roads by 2030 via one-third of current vehicles. The most popular car in Ireland in 2020 was Nissan Leaf. The launch of Renault ZOE resulted in Renault being the bestselling overall EV manufacturer in Ireland in 2020.

The government of Ireland supports the purchase of EVs to promote a cleaner and greener environment. It has facilitated multiple grants. Currently, the grant available for purchasing BEVs or PHEVs is 5K for vehicles priced at 20k or more. From 1st July 2021, the grant for PHEVs would be reduced to 2.5k. To avail of this grant, there is a criterion such as purchasing the car from a verified & approved dealer. When home chargers are considered, a grant of 600 Euro is available which helps with the purchase and installation of it.

Driving electric cars is more profitable than driving any other commercial vehicle. Based on a few calculations on charging an electric car at night time and its usability, when compared with petrol costs, it's a saving of 1361 Euro a year. The additional cost of purchasing an E-car can be recovered in 3 years. The other benefits of going ahead with electric cars are Vehicle Registration Tax is zero, benefits in the kind scheme and the motor tax for BEV is 120 Euro a year which is the lowest rate. Servicing cost of an electric vehicle is also very less when compared with commercial vehicles.

There are approximately 1200 public charging points in Ireland of which 900 belong to the Republic of Ireland. The challenge of accessing the fast-charging points placed at every 50kms on specific routes is that they tend to be out of order. The charging time varies if the charging point is fast or slow. Fast usually takes around 30 minutes to charge 80% of the battery whereas the slow ones take about 3-4 hours to charge completely. The current fee for charging is through two modes- Pay as you go or membership. The pay as you go costs 26.8c per kWh whereas the membership is 4.60Euro monthly subscription along with 23c per kWh.

The government is taking charge to improvise the current charging grid as per the *Climate Action Plan of 2019*. A few points of the proposition are:

- To address the freight emissions, supporting & expanding the charging network and the other fuel network for alternative vehicles.
- Following the act of Transport'93, the Climate Action Fund mentions the inclusion of over 90 high powered chargers at important locations on the national road network. By encouraging people to use fast chargers, installing new 50 and replacing about the existing 250 standard chargers.
- Aims for installation of at least 1 charging point for new non-residential buildings with over 10 parking spaces by 1st Jan 2025. The existing non-residential buildings with over 20 parking spaces also require a minimum number of charging points after consulting such that it does not compromise on the demand. Hence, developing appropriate rules and guidelines for the same.
- Develop a network capable of offering support to approximately 8lakh EVs by 2030 to stay ahead of the curve of demand.
- Post 2030, no new non-zero emission vehicles are to be sold to sell at least 5lakh EVs with 75% of them being BEVs.

Ireland has a target of switching to 55% renewable power, commitment to deliver full BusConnects programme for all cities, retrofit plan for 4.5lakh homes and at least 0.5 million EVs on road with additional charging infrastructure by 2030 under the Project Ireland 2040 plan.

1.3 Why are people hesitant to purchase electric vehicles?

According to the findings of KPMG's Global Automotive Executive Survey (2019), one of the key barriers to purchase intention of electric cars is the negative charging experience, which is mostly attributed to a lack of charging points, particularly when the consumer does not have the option to charge the vehicle at their home. People are wary and sceptical of transitioning to fully electric technology as a result of this. A well-connected charging point network would fix consumers' unpleasant charging experiences, thus encouraging the purchasing of electric cars. Furthermore, electric

vehicles are perceived as a precursor for autonomous vehicles and are an important component in infrastructure planning for the future.

This poses a challenge to the ESB's electric grid network in terms of determining the best location for these charge points.

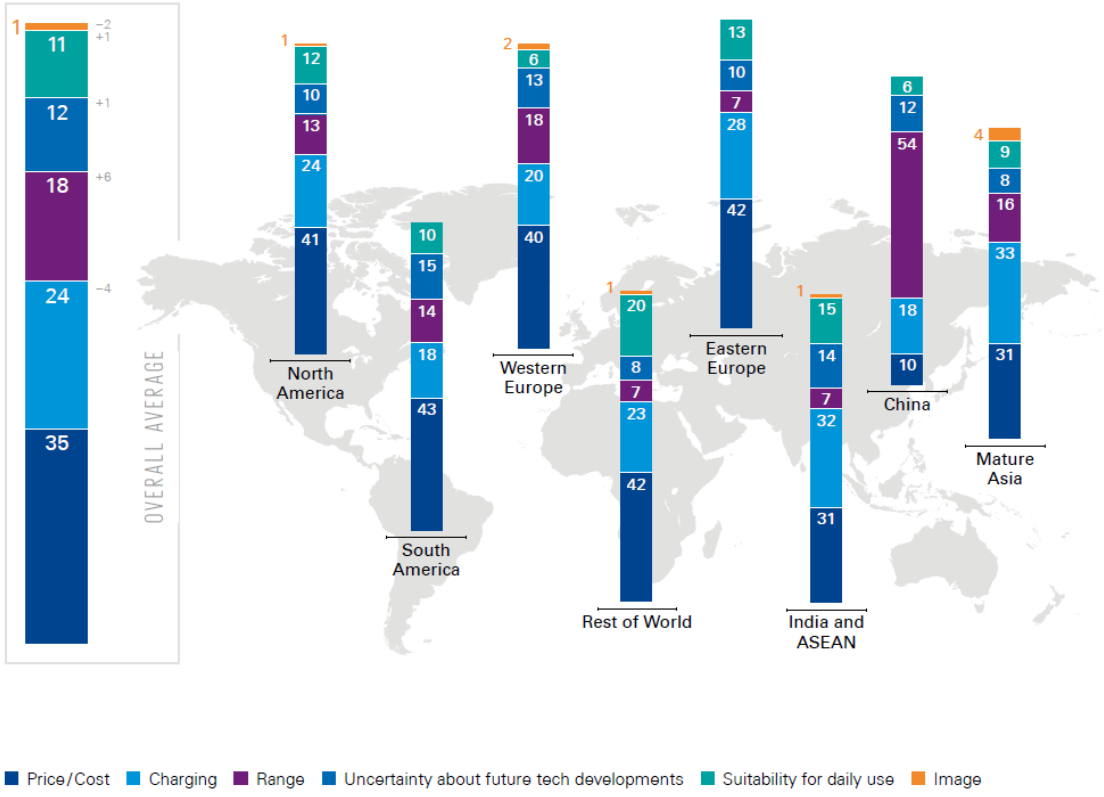

Figure 4: Purchase concerns for EVs

https://automotive-institute.kpmg.de/GAES2019/ecosystem-value/data-supremacy

1.4 Business Objective

To facilitate national electric car travel, ESB has installed Fast charging points along interurban roads. Fast charging can be accomplished with either a 3-phase, 63A AC (44kW) or a 120A, 400V DC power supply (50kW). A 50kW DC quick charging point will charge an electric vehicle up to 80% in 20-30 minutes. Our project aims to support this initiative by finding affluent areas where charging points can be installed.

To arrive at this, we must identify the factors correlated with purchasing electric vehicles and the hidden trends among various geographies in electricity consumption. We aim to summarise the findings using a dashboard. The outcome of this will be actionable insights that will assist the Electricity Supply Board of Ireland (ESB) to make an informed decision about determining the location of charging stations. As the type of charging station is influenced by resource constraints (which we do not have access to), we limit our objective to identifying the regions to be upgraded, whereas the business stakeholders can decide the type of stations to install.

The major objectives of our project are:

- to identify variables correlated with the purchase of an electric vehicle
- to group households and geographies that have identical power usage patterns
- to graphically indicate the places that require additional charging outlets, as well as other actionable insights that assist the climate action plan

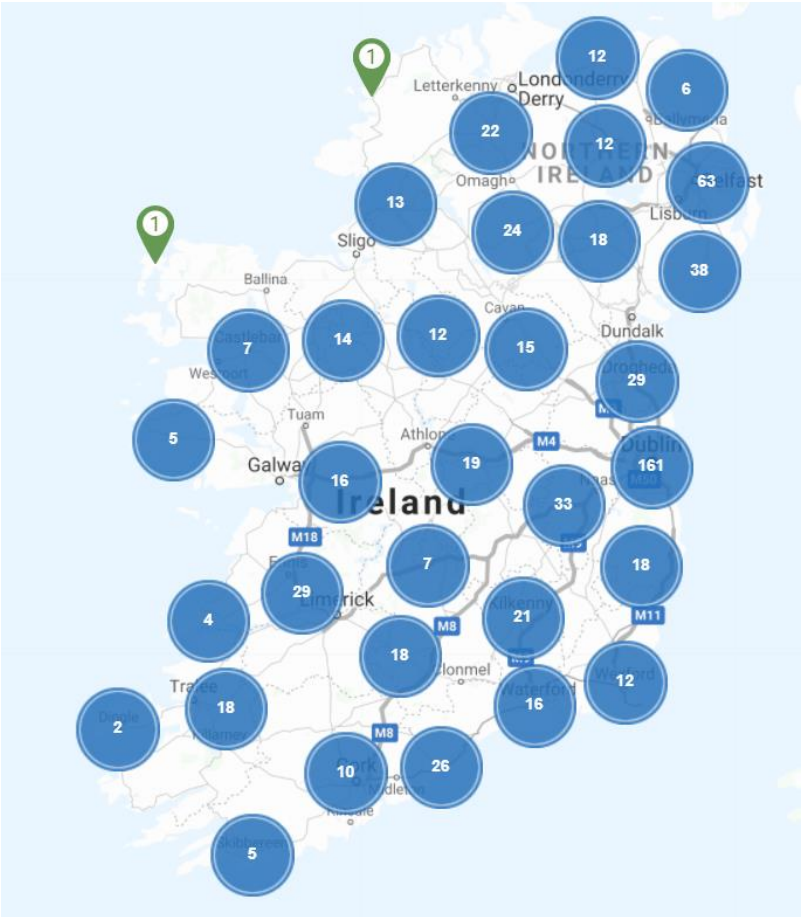

Figure 5: Current Charge Point Map

https://esb.ie/ecars/charge-point-map

Chapter 2 - Literature Review

2.1 Unsupervised learning - clustering

Clustering is an unsupervised learning method for classifying data of similar types. Its application varies in different disciplines and is the most widely used technique for performing exploratory data analysis. With the adoption of different assumptions and application of novel concepts, it internally has multiple classification methods.

Jain, A.K. et al. (1999) summarises various clustering methods taking statistical patterns into perspective. It starts from explaining fundamental concepts in clustering to advising on the adoption & selection of clustering techniques by explaining the taxonomy and cross-cutting themes along with the progress made so far in clustering, lastly describing the applications. Subject to the application of data makes the entire process of clustering tough. That leads to the use and exploration of different algorithms. Knowledge of the domain of application is the key to clustering, as it helps in the reasoning of grouping after clustering. The knowledge-based algorithm offers a link in the business understanding and available dynamic features. Four steps of clustering, mainly pattern representation, similarity computation, grouping process and cluster representation, are discussed in this review.

Pattern representation is identifying the pattern in the data. It is easily identifiable with small data sets, but with a larger dataset, the computation costs increase and become difficult to spot. To cater to this, many feature extraction methods have got a distinct combination to use as a source of measurement to represent patterns. Similarity computation is the identifying and measuring similarity between two patterns. The representation of patterns in a correct form is important to extract meaningful information. Rather than looking at similarity between two patterns, dissimilarity with distance as a metric is widely used. The grouping process can be divided into 2 schemes, namely hierarchical and partitional.

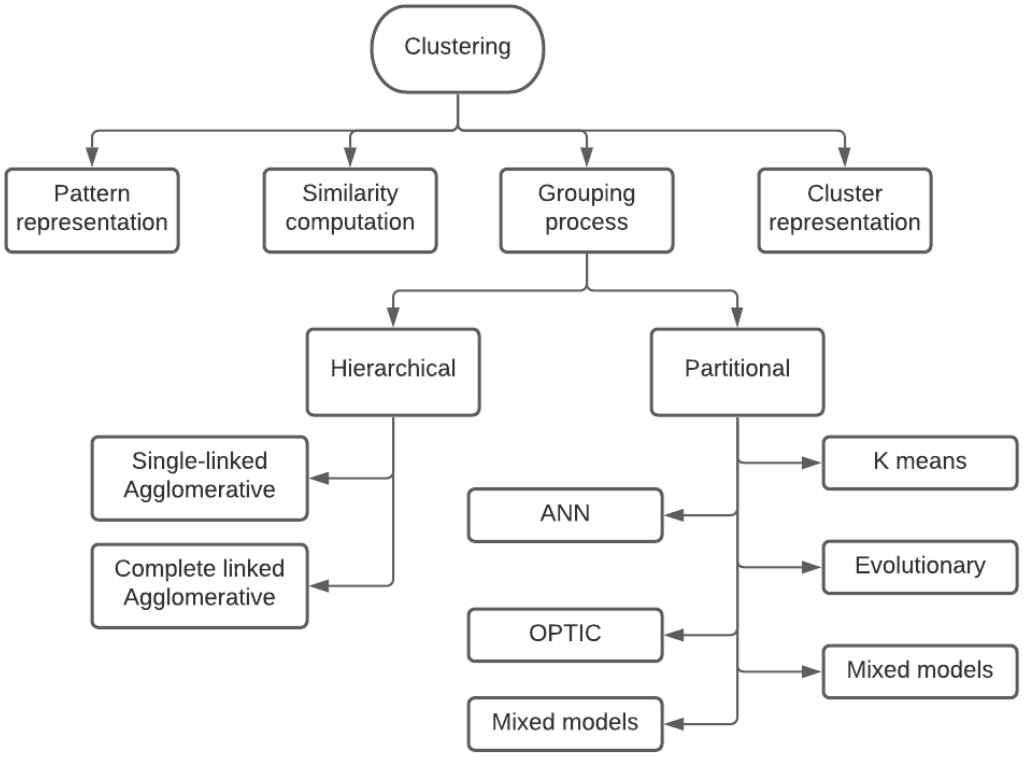

Figure 6: Clustering review

2.1.1 Hierarchical

Hierarchical algorithm groups similar data points in a way that towards the end, there is a hierarchy represented by a dendrogram. The most popular algorithms here are single-link and complete-link. Single-link algorithms work based on having a minimum distance between two clusters of the distances of all pairs of patterns drawn from two clusters. The chaining effect is identified by such algorithms. For complete-link algorithms, the maximum distance between two clusters for pairwise patterns. Tightly bound clusters are identified by such algorithms. In both the algorithms, the two clusters are merged to form a larger cluster based on minimum distance criteria. The hierarchical schemes are usually expensive and multipurpose, whereas the partitional schemes are less expensive.

With an increasing demand for data sources and information, the segregating and grouping of relevant information together is important. With this, demand for a consistent and more reliable clustering algorithm increases with its capability to

explore data at diverse stages of granularity. *Zhao, Y. et al. (2005)* discusses such solutions built by clustering algorithms and different criterion functions and merging schemes used by partitional and agglomerative algorithms. It also presents constrained agglomerative algorithms, which focus on improving the quality of clustering solutions. From the experiments conducted, constrained agglomerative methods led to better hierarchical solutions when compared to partitional and agglomerative methods alone. Improvements of neighbourhoods lead to the enhancement of overall clustering solutions by imposing partitional cluster constraints. Hence, starting with purer neighbourhoods leads to clustering solutions with improved quality of the constrained agglomerative schemes.

2.1.2 K means

The nearest neighbour clustering is clustering based on the proximity between the data points based on a threshold. *Jain, A.K. et al. (1999)* discusses a Fuzzy clustering that uses membership function to identify the association between patterns in each cluster. For a few applications, where clustering is more important and not partitioning of data, which has the flexibility to overlap, fuzzy and functional clustering is more suited. For fuzzy clustering, representation of clusters in a way where the decision-maker understands is important as there is no general approach to it. It differs as per the objective of the business.

Selecting the right k value is critical in K means algorithm as it directly impacts the model performance. Elbow is one of the most well-known techniques for determining the appropriate value of k and improving model performance. We iterate over several k values. The model randomly initializes k clusters for each iteration and computes the sum of the squares of the points and their average distance. When k =1, the within-cluster sum of the squares is high, and it decreases as the value of k increases. We plot a graph between each k-value and their within-cluster sum of the squares. The best value for k is found at the point where the graph abruptly declines.

Following that, k means (points) are chosen at random and k clusters are formed by linking each point with its nearest mean. The new mean is determined by the centroid of each new cluster. This process is repeated until convergence is achieved.

For large datasets, the K-means algorithm, and its neural implementation- the Kohonen net are implemented successfully. K-means is easy to implement and compute. It again faces some challenges when the complexity and size of the datasets is concerned. Additional algorithms, such as leader and neural network, such as the ART network, can group and cluster large data sets.

2.1.3 ANN

For finding the optimal solution of a problem, Jain, A.K. et al. (1999) discusses a partitional algorithm that uses maximising the squared error criterion function. Partitional algorithms divide the data set into distinct sets, creating a single partition. The algorithms here run under multiple iterations before arriving at the final solution. However, different algorithms have failed in achieving so and it is not computationally possible on large datasets. The ANN-based clustering scheme can automatically normalise the data and extract important features for clustering. Some features of the ANNs that are important in pattern clustering are:

- Quantitative features only
- Inherently parallel and distributed processing architectures
- may learn their interconnection weights adaptively
- Can act as pattern normalisers and feature selectors by a selection of weights

Mixture resolving and mode-seeking algorithms run on the assumption that the patterns to identify the similarity between data points in a dataset for clustering are drawn from one of the several distributions to finally glean the parameters and their number. MLE, EM and non-parametric techniques such as density-based clustering are some common techniques used.

2.1.4 Evolutionary algorithms

The major focus of clustering is to keep the similar data points in a cluster close, whereas the dissimilar ones are apart. To arrive at the most optimal solution towards the end, clustering can be considered as a particular NP-hard grouping problem. Hence, with a curiosity of exploring the best algorithm for such problems, an evolutionary algorithm is widely used as it takes a decent amount of computation and offers near to the most optimal solution. When the value of k is provided by the

business experts as per the business requirement, there is an ease of looking for a suitable algorithm as the person working on it will then look for algorithms for fixed k such as k-means, expectation maximisation and SOM algorithms. The algorithm is extensively discussed in *Hruschka, E.R. et al. (2009).*

Evolutionary algorithms use probabilistic rules to process the clusters built and provide a better solution. The closer the data points are in a cluster, the higher is the probability of being sampled. Hence, the evolutionary search is very dependable on the existing cluster solution to make it more computationally efficient ultimately giving a much better cluster partition than a traditional random approach. Evolutionary algorithms work for problems where k is pre-defined by the business experts as well as when k is not pre-defined. For the case where it is not pre-defined, it provides the best number of clusters, i.e., k and the partitions in sections of the space where they are more probable to be found. A few of them even make use of k-means for local search techniques to estimate the value of k.

A couple of disadvantages of evolutionary algorithms is that they cannot handle the missing values in a dataset automatically. If the number of missing values is low in proportion, it computes pair-wise similarity measures on known values. They also cannot distinguish between relevant and irrelevant features for the clustering process. K-means fine-tunes rough partitions laid out by evolutionary search thus collaboration between the two making the results stronger. This collaboration results in minimising the variances of clusters that do not stick at a suboptimal centroid. Research has also shown that evolutionary algorithms outperform, estimating the unknown number of fuzzy clusters as well from a theoretical and experimental perspective.

2.1.5 OPTIC algorithm

Ankerst, M. et al. (1999) introduces a new algorithm that creates augmented ordering containing information of a broad range of parameter settings of the data points offering a density-based clustering structure. It performs automatic and interactive cluster analysis. It easily extracts information such as representative points, arbitrary shaped clusters, and intrinsic clustering structures. To represent the order and making it more interactive and offering additional insights, it has been recommended to represent graphically for medium-sized data sets whereas for large datasets, a

visualisation technique. The other advantage of using this OPTIC algorithm is that it does not limit itself to one global parameter setting.

2.1.6 Mixed models in network

With the interactive and strong topology of the system present in biology, physics, and other areas, understanding its structure is a constantly growing process that involves a lot of research. In *Newman, M. and Leicht, E., (2007),* a technique of clubbing similar patterns of connection in a class made of nodes in a large-scale network data is described. It makes use of the machinery of probabilistic mixture models and expectation-maximisation algorithms which detect types of structure in networks. This method helps them determine the properties of each class. The algorithm developed can determine the structures present in the network which could include bipartite, fuzzy, overlapping, etc. The paper also discusses the application in real-world and computer-generated networks. It hence provides flexibility to explore and identify the type having no expected structure in mind.

2.1.7 Spectral Clustering

Donath and Hoffman (1973) proposed the first graph partitions based on eigenvectors of the adjacency matrix, which led to the development of spectral clustering. The similarity matrix, which is supplied as an input, is a quantitative estimate of the relative similarity of each pair of points in the dataset. The similarity matrix may be described as asymmetric matrix A given an enumerated collection of data points, where A_{ij} indicates a measure of similarity between data points with indices i and j. In general, spectral clustering is accomplished by employing a conventional clustering algorithm on relevant eigenvectors of a Laplacian in the given symmetric matrix A. There are several methods to define a Laplacian, each with its mathematical connotation, and therefore the clustering will have its interpretation. *Von Luxburg, U. (2007)* goes through several Laplacian methods in depth.

Authors	Journal	Topic	Main Findings	Further Research
Jain, A.K., Murty, M.N., Flynn, P.J.	Data clustering: A review - 1999	Types and steps for clustering	Knowledge of the domain of application is the key to clustering as it helps in the reasoning of grouping after clustering. Four steps of clustering mainly pattern representation, similarity computation, grouping process and cluster representation are discussed in this paper. The ANN-based clustering scheme can automatically normalise the data and extract important features for clustering.	More research on parallelization of clustering techniques for the success of cluster analysis on large-scale datasets
Ankerst, M., Breunig, M. M., Kriegel, H. Sander, J.	OPTICS – 1999	OPTIC algorithm, its function, and advantages	It creates augmented ordering containing information of a broad range of parameter settings of the data points offering a density-based clustering structure. It performs automatic and interactive cluster analysis. It easily	Research on the application of high dimensional spaces for the OPTICS algorithm to ensure its feasibility. With the database updates, managing the

			extracts information such as representative points, arbitrary shaped clusters, and intrinsic clustering structure	cluster-order and research possibility on effects of trade-off on a limited amount of accuracy for a large gain in efficiency
Zhao, Y., Karypis, G. Fayyad, U.	Hierarchical Clustering Algorithms for Document Datasets - 2005	Discusses clustering algorithms and different criterion functions and merging schemes used by partitional and agglomerative algorithms	Constrained agglomerative methods led to better hierarchical solutions when compared to partitional and agglomerative methods alone	NA
Newman, M. E. J. Leicht, E. A.	Mixture models and exploratory analysis in networks - 2007	Clustering of similar nodes in a network.	A technique that makes the use of the machinery of probabilistic mixture models and expectation-maximisation algorithm which detect types of structure in networks is introduced. This method helps them determine the properties of each class. The algorithm	NA

			developed can determine the structures present in the network which could include bipartite, fuzzy, etc.	
Von Luxburg, U.	Statistics and Computing - 2007	A tutorial on spectral clustering	The results produced by spectral clustering frequently surpass the results provided by traditional techniques. spectral clustering is relatively easy to implement and can be solved rapidly using basic linear algebra methods.	NA
Hruschka, E.R., Campello, R.J.G.B., Freitas, A.A., de Carvalho, A.C.P.L.F.	A survey of evolutionary algorithms for clustering -2009	What are evolutionary algorithms and their application in clustering	Evolutionary algorithms work for problems where k is pre-defined by the business experts as well as when k is not pre-defined. The evolutionary search is very dependable on the existing cluster solution to make it more computationally efficient ultimately giving a much better cluster partition than a traditional random approach.	Research on detailed theoretical analyses in terms of time complexity for evolution clustering, multi-objective clustering, the combination of evolutionary approaches with traditional hierarchical clustering algorithms

Table 1: Summary of Clustering literature

2.2 Time Series Forecasting

Over the last few decades, time series analysis and forecasting have been important study fields. De Gooijer, J.G. and Hyndman, R.J. (2005) examines the advancements in time series forecasting during the last several decades, highlighting all the unique approaches and research undertaken during this period. For our research, we are particularly interested in linear and hybrid models since they are more closely related to real-world problems. As the accuracy of time series forecasting is critical to many decision processes, research to improve the efficacy of forecasting models has never stopped. The purpose of time series prediction is to forecast a future value based on current and previous data samples. Time series prediction seeks to find a function f (x) such that the anticipated value of the time series at a future point in time, is unbiased and consistent. Mathematically,

$$\hat{x}(t + \Delta t) = f(x(t - a), x(t - b), x(t - c), \dots) \quad (1)$$

Where: $\hat{x}$ is the predicted value of a discrete-time series t

Based on the estimator function f(x) time series can be classified as linear and non-linear.

2.2.1 ARIMA

Linear time series are represented as Autoregressive or Moving Average models, which are coupled to form the ARIMA process. The Auto-Regressive Integrated Moving Average (ARIMA) is a popular forecasting approach. Rao, J.N.K., Box, G.E.P. and Jenkins, G.M. (1972) devised a practical method for improving the ARIMA model, which is frequently used in time series analysis and forecasting applications. A non-seasonal time series can be described as a combination of past values and previous errors, where the predicted value of a variable is a linear function of various past observations and random errors. The model is denoted as ARIMA (p,d,q) and is given by the equation,

$$y_t = \theta_0 + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p}$$

$$+\varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q} \quad (2)$$

Where: t = time

y_t = actual value at time t

ε_t = random error at time t with a mean of zero and a variance of σ^2

ϕ_i (i = 1, 2, …, p) and θ_j (j = 1, 2, ..., q) are the model parameters

p is the order of autoregressive terms, d is the order of differencing needed for stationarity, and q is the order of moving average. They take integer values and are referred to as model orders.

The framework of this method is composed of three steps:

- **Model Identification** - The essential principle behind model identification is that a time series generated by an ARIMA process should exhibit theoretical autocorrelation characteristics. It entails calculating the model orders (p, d, and q) to capture the data's key properties. To determine the order, graphical techniques such as the series plot, Autocorrelation function (ACF), and Partial Autocorrelation Function (PACF) are utilised.
- **Parameter estimation -** At this step, we apply computing methods to determine the best parameter values. The parameters are calculated such that the total amount of error is minimised. This is feasible using a nonlinear optimisation approach.
- **Model Verification -** This step is to validate the model assumptions regarding the errors. Several diagnostic statistics and residual plots can be used to assess the quality of fit of the model. If the model is infeasible, an alternative model should be identified, followed by the parameter estimation and model verification stages.

This three-step model construction method is often repeated several times until a satisfactory model is found. The final model chosen can then be utilised to make predictions.

2.2.2 Hybrid Model

Although ARIMA can model a broad spectrum of time series problems, it cannot be employed in every forecasting scenario. ARIMA's major drawback is the model's presumed linear structure. Though it may solve complex real-world problems after linear approximation, it does not always yield accurate solutions. To mitigate this drawback, Zhang, G.P. (2003) proposed a hybrid technique that integrates both ARIMA and ANN models as a novel way to leverage the distinctive strengths of ARIMA and ANN models in linear and nonlinear modelling. The suggested hybrid model technique comprises two stages.

- The time series problem is analysed using an ARIMA model in the first stage. As the ARIMA model can capture only the linear component of the model, the residuals of the model will contain the nonlinearity information
- In the second stage, a neural network model is employed to analyse the ARIMA model's residuals

This hybrid model takes advantage of the distinct strengths and efficacy of both the ARIMA and the ANN models in identifying distinct patterns in data. The combination technique can be a useful strategy to enhance forecasting performance for complicated problems with both linear and nonlinear correlation patterns. Experiment results using actual data sets show that the hybrid algorithm can be an effective option for improving predicting accuracy over either of the individual models.

Chapter 3 - Approach

The ultimate focus of our research is to provide insights to identify areas that require additional charging stations. To do this, we use exploratory data analysis on the shared data to discover the parameters associated with energy consumption, cluster similar households in the network and discover the trend for each cluster. We propose conducting the EDA in Excel, clustering in R and creating a dashboard in PowerBI. We intend to present our findings in a report to KPMG, highlighting important trends and opportunities. Based on research, the clustering approach we suggest adopting is the K-means based clustering scheme. We have compiled the advantages, limitations and workarounds of the method chosen as shown in Table 2.

Advantages	Limitations	Solution
• Can scale for large datasets – new data points are allocated to clusters based on the closest distance	• The results are affected by the initial points selected	• we may alleviate this by running the model for several initial points and seeding the one that produced the best results
• Guaranteed convergence	• The results are prone to outliers	• Before applying the model, outliers should be treated
• We have the flexibility of determining the number of clusters	• It uses quantitative features only	• this does not affect our use case as we intend to use only the numeric variables in our clustering
• It is relatively easier to implement and explain	• The distance-based similarity algorithm converges to a constant value as the number of dimensions increases	• . We can get around this by reducing the dimensionality using PCA

Table 2: Advantages and limitations of the chosen clustering approach

Chapter 4 - Success criteria

4.1 Success metrics for EDA

The most fundamental phase in our project is exploratory data analysis. It assesses our ability to comprehend the knowledge and the objective of the analysis. To assess its success, a few criteria such as the utilisation of external data, communication with stakeholders, principles that are aligned between us and the stakeholders on which analysis is created, and input from stakeholders can be considered.

4.2 Success metrics for clustering

Based on our research of *U. Maulik and S. Bandyopadhyay (2002)*, we determined that the following metrics are the best to consider when evaluating our clusters.

Clusters are evaluated based on the distance between points of cluster i.e., similarity and dissimilarity. By the nature and purpose of clustering, it is considered to work well if similar observations are clustered together, and dissimilar observations are separated. We intend to evaluate clustering with the Silhouette coefficient and Dunn's Index.

Silhouette coefficient (s) is calculated based on two scores for each data point, one score is the mean distance between a data point and other points in the same cluster and the other score is the mean distance between a data point and other points in the next nearest cluster. The formula is given as:

$$s = \frac{b - a}{max(a, b)} \tag{7}$$

Where: a is the mean intra-cluster distance

b is the mean nearest-cluster distance

S lies between -1 and +1, where -1 is for incorrect clustering, 0 is for overlapping clusters and +1 is for dense clustering. The higher the coefficient value, the greater quality is of the clusters.

Dunn's Index (D) is the ratio of minimum inter-cluster distance to the maximum cluster size. The higher the value of Dunn's index, the better is the clustering i.e., the clusters are well apart from other clusters and each cluster is densely compact.

$$D = \frac{\min\left(inter\ cluster\ distance\right)}{\max\left(intra\ cluster\ distance\right)} \tag{8}$$

4.3 Success metrics for dashboard

A few metrics for dashboard success are usage rate, leveraging dashboards for insights and decision making by the stakeholders, peer review and response time. These metrics are credible only when implementing and maintaining a dashboard for over a period greater than 3 months. Since this requires a much larger audience to track and goes beyond the scope of the project, we feel we can base our success for the dashboard currently on the feedback we receive from the stakeholders.

Chapter 5 - Exploratory Data Analysis

5.1 Data description

For our study, we were provided with a dataset of SEAI's various programmes for the past 10 years. The data was detailed at each household level (represented by the UUID). The data comprises dwelling information, such as the various measures installed in the house, the house's area code and county information, and system information, such as the various schemes each household applied for, the date of application, the status of the application and the energy ratings. The raw data had 914,268 rows and 42 columns.

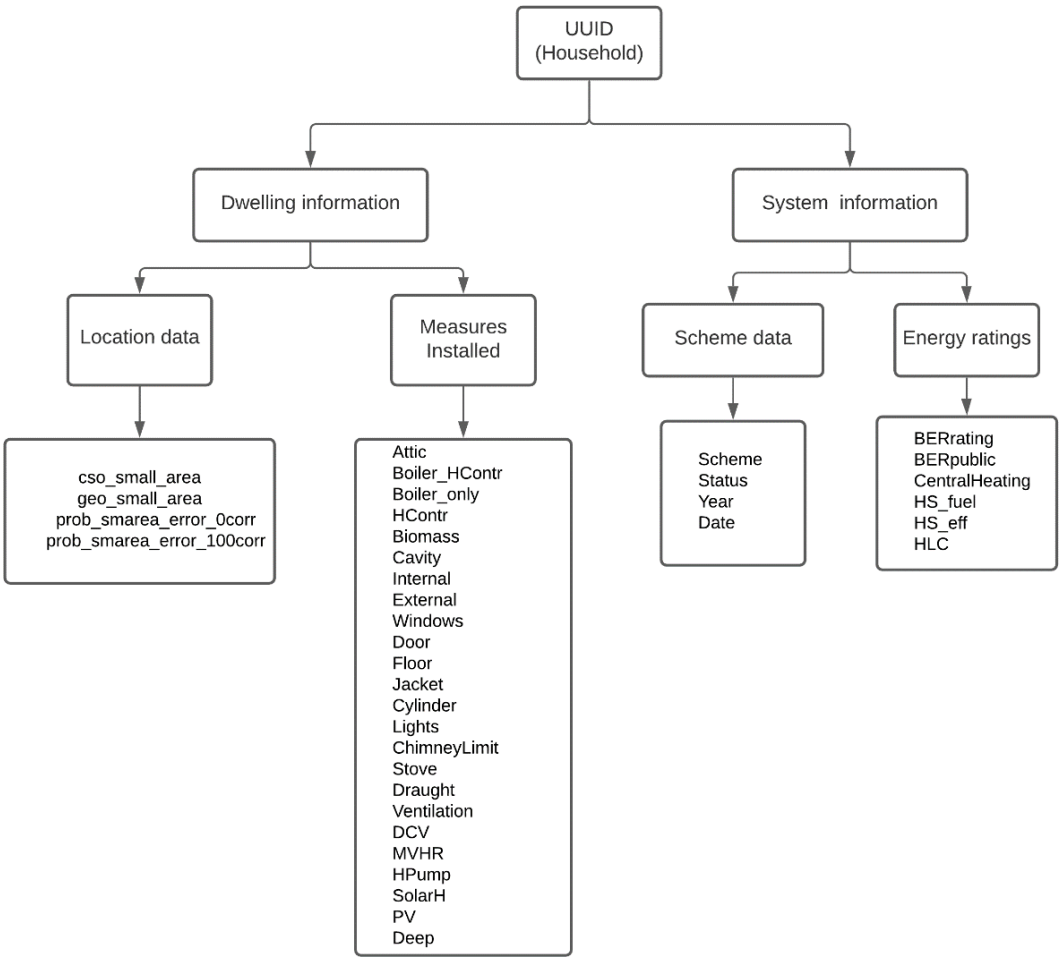

Figure 7: SEAI data

Column Name	Description
UUID	Universally unique identifier - represents unique household
cso_small_area	Small area codes in CSO format.
geo_small_area	Small area codes in GeoDirectory format.
prob_smarea_error_0corr	Variable calculated by Geo-coding programme used to match Eircode/Small areas to address fields within BER database. It represents a probability of the algorithm assigning a wrong Eircode/Small Area code to a given record. 0corr represents the assumption of perfect independence between various geocoding algorithms/programmes used
prob_smarea_error_100corr	Variable calculated by Geo-coding programme used to match Eircode/Small areas to address fields within BER database. It represents a probability of the algorithm assigning a wrong Eircode/Small Area code to a given record. 100corr represents the assumption of 100% correlation between different geocoding algorithms/programmes used.
Scheme	BEAB (Area based Better Energy Communities scheme), BEC (Better Energy Communities scheme), BEF (Better Energy Finance pilot scheme), BEH (Better Energy homes scheme), BEH_Cparty (Better Energy homes scheme, counterparty application), Caranua (Warmer homes scheme - project Caranua), DRF (Deep Retrofit pilot programme), NGR (Non-grant residential), WAW (Warmth and Wellbeing scheme), WHS (Warmer homes scheme).
Status	Ongoing (Application in progress), Cancelled (Application Cancelled), Completed (Application completed and indicated measures installed)
Year	Calendar year used for reporting

Date	Date the application was finalized (e.g., the grant paid, or measures installed). For Ongoing and Cancelled applications date represents the application received date.
Attic	Measure – attic insulation installed
Boiler_HContr	Measure – New heating boiler and heating controls installed
Boiler_only	Measure – New heating boiler installed
HContr	Measure – Heating controls installed
Biomass	Measure – Biomass boiler installed
Cavity	Measure – Cavity wall insulation installed
Internal	Measure – Internal wall insulation installed
External	Measure – External wall insulation installed
Windows	Measure – New energy efficient windows installed
Door	Measure – New Energy-efficient doors installed
Floor	Measure – Floor insulation installed
Jacket	Measure – Lagging jacket for hot water cylinder installed
Cylinder	Measure – New energy-efficient hot water cylinder installed
Lights	Measure – New energy-efficient light bulbs installed
ChimneyLimit	Measure – Chimney limiter installed
Stove	Measure – Stove with external air intake installed as secondary heating system
Draught	Measure – Draught stripping installed
Ventilation	Measure – Air vent installed
DCV	Measure – Demand controlled ventilation installed
MVHR	Measure – Mechanical ventilation with heat recovery installed
HPump	Measure – Heat pump installed
SolarH	Measure – Solar hot water heating system installed
PV	Measure – PV system installed
Deep	Measure – Unknown number of measures similar to Deep retrofit installed
BERrating	BER rating achieved after measure installation

BERpublic	"Yes" = A given BER is the most current BER available, "No" = A given BER is not the most current BER available
CentralHeating	The dwelling has a central heating system
HS_fuel	Fuel used for the main heating system
HS_eff	The energy efficiency of the main heating system
HLC	Heat loss coefficient/indicator in W/Km2
County26	County in the format of main counties
County52	County in the format of main counties plus Dublin 1,2,3, etc.
Measures	Measures installed in a string format

Table 3: SEAI Data Description

5.2 Data transformations

The raw data is at each application level. Each row, therefore, represents a single instance of an application made by the home. As a result, a single house may have numerous entries in the system, each corresponding to the submission date. For our analysis, we aggregated the data at the household level, combining all applications made by a single household. The aggregation is detailed below:

- The location information for all entries in the household remains the same. Hence, we consider the unique value of this information to be the value after aggregation.
- To aggregate the measures data, we introduced a new column called *totalMeasures*, which is the count of all the measures installed in each house.
- The scheme-related information is the crucial data for our analysis. Using the date column, we added three new columns: *min_date* (the earliest date on which the house made an application), *max_date* (the latest date on which the house made an application), and *houseVintage* (the difference in months between *min_date* and *max_date*).
- Using the Scheme column, we computed the number of distinct schemes applied by a household and labelled it schemeCount. We also included a schemeConcat reference column, which contains a list of the schemes applied by a household.

- We aggregated the Status column by creating total_submissions (the total number of applications filed by a house), *approvedClaims* (the total number of approved applications), *rejectedClaims* (the total number of rejected applications), and *pendingClaims* (the total number of pending applications).
- Energy ratings of a household may vary over time. As a practical assumption, the ratings of the household on their most recent application as captured by the SEAI system are considered the final rating of the household

Derived Columns	**Description**
totalMeasures	Total count of measures installed in a house
min_date	the earliest date on which the house made an application
max_date	the latest date on which the house made an application
houseVintage	the difference in months between *min_date* and *max_date*
schemeCount	the number of distinct schemes applied by a household
schemeConcat	concatenation of the schemes applied by a household
total_submissions	the total number of applications filed by a house
approvedClaims	the total number of approved applications
rejectedClaims	the total number of rejected applications
pendingClaims	the total number of pending applications
approvalRate	Percentage of applications approved

Table 4: Derived Columns

5.3 Assumptions

After transforming the provided data, we observed 198,444 households (39 per cent) that lacked energy rating information. Following further investigation, we discovered that these households can be divided into two groups: exempt and non-exempt.

For our study, we require a household's energy ratings to determine its link with the scheme approvals and to determine whether a household is likely to purchase an electric car. The BER rating of a building indicates how energy efficient it is. The BER cert will give the building an energy rating ranging from A1 to G, with A being the most efficient and G being the least efficient. Households may be exempted from energy ratings under certain conditions.

When an application is accepted, the SEAI system records the energy rating of the residence. For 26 per cent of the households, the energy rating columns were not available even with an accepted application. We categorise these as the exempted households present in our data. These households are not useful for our study and can be removed from the data.

We classify the remaining households that do not have an energy rating as non-exempt. We approximated the energy ratings of these households as we did not want to lose a lot of information. We used a county-level distribution imputation approach, wherein we imputed the energy ratings of non-exempted households based on the proportion of registered energy ratings in that county. This method allows us to prevent skewness in our imputation.

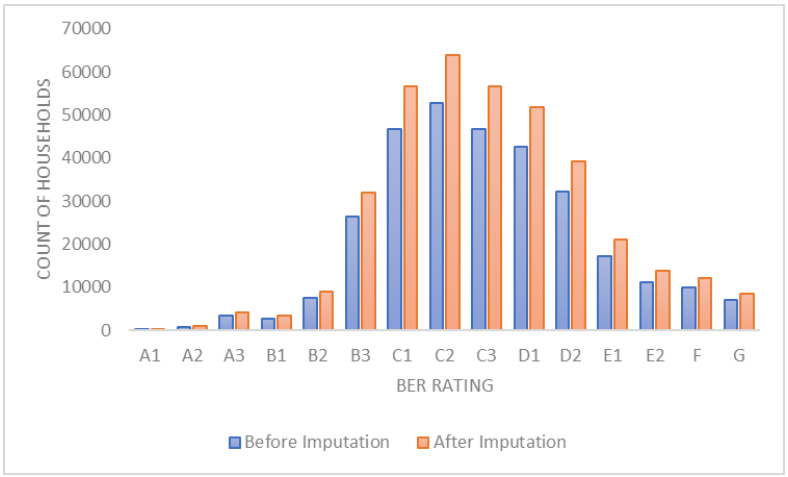

Figure 8: BER rating imputation

Figure 10 shows that the distribution of ratings before and after imputation is identical. Hence it is proved that there is no skewness caused by our imputation.

5.4 External Data Sources

We merged 2016 Ireland county-level census data with SEAI data on the county column. The census data comprises population information, deprivation score, unemployment rate, the total number of households, households without a car, and people with disabilities. These columns will help us better assess each county's potential and develop county-specific strategies.

Columns	**Description**
Total.Population.2016	Total population in the county as per 2016
Population.Change..2011.16	Percentage change in population from the previous census report
Unemployment.rate.Male.2016	Percentage of unemployed males
Unemployment.rate.Female.2016	Percentage of unemployed females
No..of.Households.2016	Number of households in the county
Households.without.a.Car.2016	Number of households without a car
Persons.with.a.Disability.2016	Population with disability
Deprivation.Score.2016	Poverty index of the county

Table 5: Census data description

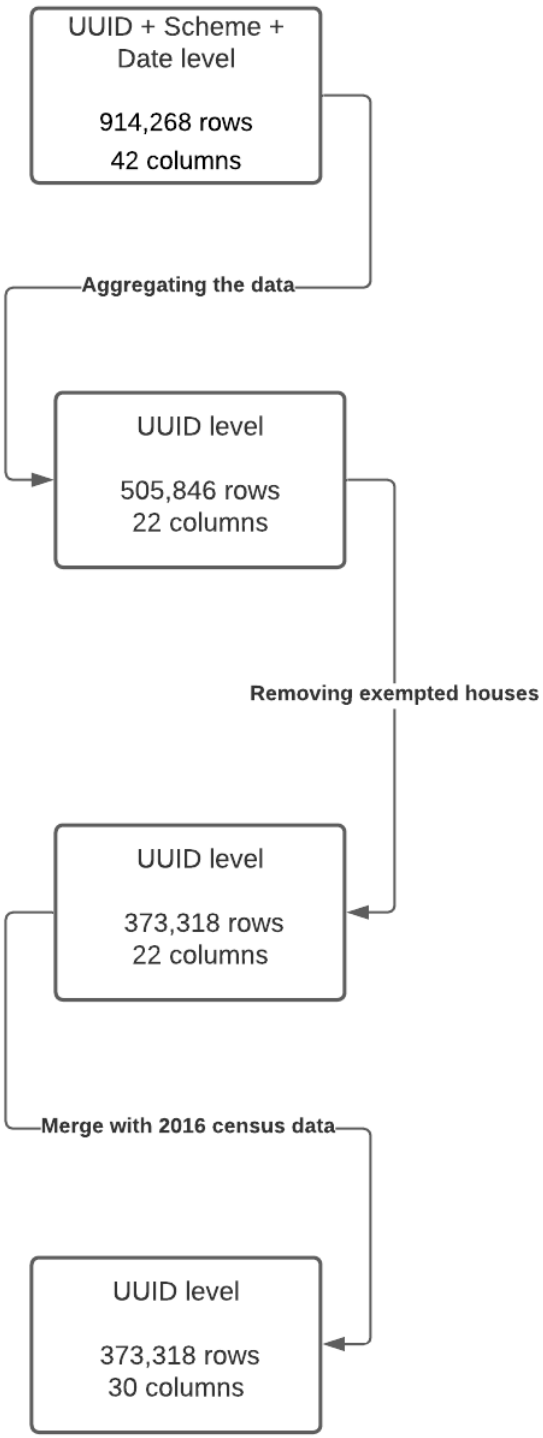

Figure 9: Data transformation

The final data set we considered for our clustering analysis consisted of 373,318 rows and 30 columns.

Chapter 6 - Grouping the households

6.1 Clustering

We employed a clustering algorithm to seek households that displayed similar behaviour. Our final analytic base table had both categorical and numerical data, and we used k-means as stated in the literature review. The K-means algorithm requires only numeric data to be fed. The numeric variables in our analytic base table were scheme count, total measures, total submissions, rejected claims, approved claims, pending claims, house vintage and approval rate. Variables such as total submissions rejected claims, approved claims and pending claims are highly dependent on each other and hence we went forward with only keeping total submissions amongst these.

Hence, a final subset of the table having UUID, scheme count, total measures, total submissions, house vintage and approval rate was created.

An important aspect of clustering is the scaling of data given highly ranged variables such as house vintage and total submissions. Scaling of data aids in bringing variables to a consistent scale, removing dominance of a specific variable due to high range values, and perform meaningful analysis. After scaling, the next step was to identify the optimal number of clusters. A Scree plot was created to assist in determining the number of clusters. The plot provided us with a total within clusters sum of squares vs the number of clusters graph, and the goal is to choose the number of clusters corresponding to an elbow break in the curve.

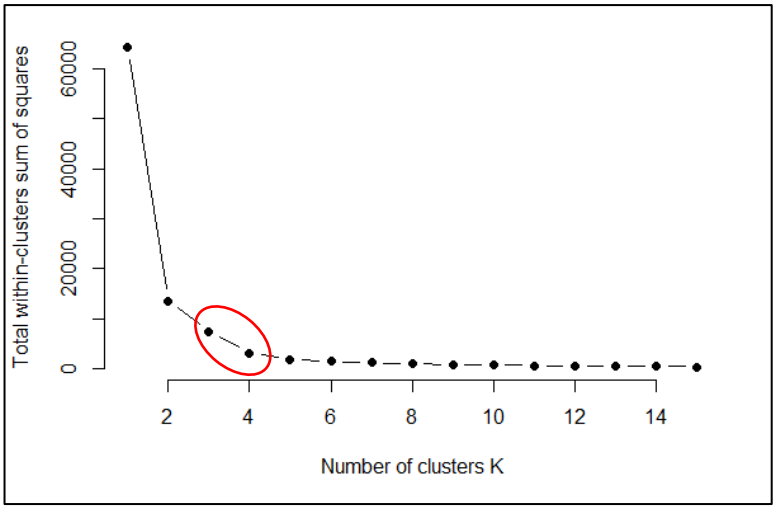

Figure 10: K means scree plot

From figure 12, we can observe a good break at 4. We decided to perform two iterations i.e., taking 3 and 4 as the number of clusters for k-means.

For the first iteration, we ran k means with 4 clusters. The observed compactness was 95.2%. This score was a good indication of clusters and so we began profiling. However, when we profiled the data, we discovered that two of the four clusters had similar behaviour for all variables except one, total submission. Having a distinct cluster based just on one characteristic is not significant. Hence, we decided to analyse the profiles with only 3 clusters.

The observed compactness was 88.5% which was about 6% lower than the previous iteration. The Dunn Index was 0.22, which is low but justified given that our cluster 2 is a combination of two clusters from the prior iteration, resulting in a larger cluster size. The average silhouette score is 0.88, which is satisfactory.

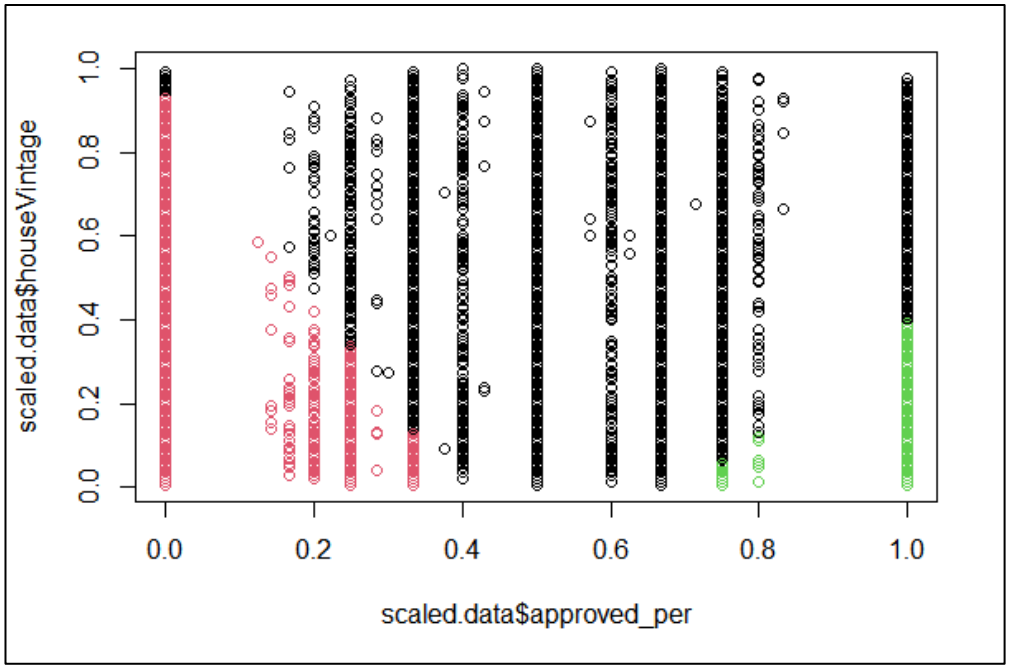

Figure 11: Approved Rate vs House Vintage

Figure 13 shows the cluster distribution when looked at House vintage vs approval rate for a household. The majority of the households observed here have a distinct & unique value for the combination of vintage and approval rate leading to a straight line

observed. Cluster 2, which is coloured black, has the most observations, spanning the bulk of the graph with its diverse distribution.

Figure 12: Approval rate vs Total measures

Figure 14 depicts the approval rate distribution with the total number of measures placed in a home. We can see a good differentiation in the clusters as indicated by the colour code.

6.2 Profiling

Once we got the clusters, we started profiling by observing the distribution of households in each cluster and its variable value. The three clusters were profiled with the following variables: Household vintage, BER Rating, HF Fuel, Scheme count, total measures, total submissions, and approval rate.

- **House Vintage:** The house vintage, which was calculated in months as previously described in the data preparation phase, was further classified as new, semi, and old. New were the households having house vintage <=6 months, semi were households having vintage greater than 6 months and <= 24 months, and >24 months house vintage was labelled old.

Table 6 below shows the row-wise population distribution of households across clusters vs house vintage. The Grand total is the distribution of households irrespective of the cluster they belong to. The general trend of the households observed is 84.6% are new i.e., the households which have made applications for installation in their first 6 months. 6.2% of them belong to semi and 9.2% belong to old.

Now, having such a generalised trend, the expected distribution in each cluster would be the same but we observe a huge difference. To make the visual impact easy from the given table, we have highlighted the cell having a huge gap in the distribution value and grand total value of the same column. Due to these differing gaps, they define that cluster. Hence, to profile the clusters as per the observed distribution difference when compared with the grand total, cluster 1 is defined by the old and semi households, cluster 2 and 3 by new households. It is interesting to observe that in cluster 1, the value of old is around 5 times the grand total whereas the value of semi is around 3 times.

HouseVintage	new	old	semi
1	30.5%	49.8%	19.7%
2	92.7%	3.4%	3.9%
3	91.3%	3.2%	5.5%
Grand Total	**84.6%**	**9.2%**	**6.2%**

Table 6: Cluster analysis: House Vintage

- **BER rating:** The BER rating values were A1, A2, A3, B1, B2, B3, C1, C2, C3, D1, D2, E1, E2, F, G. The values A's, B's, C's, D's and E's were clubbed to A, B, C, D, E respectively for the analysis purpose. Table 7 below provides the household distribution cluster wise. As observed, the cluster 1 values differ a lot as compared to the general grand total distribution. From what we see, cluster 1 is households with high-efficiency houses since rating A, B, C are much more efficient than the rest ratings. Clusters 2 and 3 are following the general trend and hence there is not much to distinguish.

BER	A	B	C	D	E	F	G
1	1.61%	17.90%	49.89%	21.28%	6.31%	1.75%	1.26%
2	1.37%	10.77%	47.17%	24.91%	9.84%	3.51%	2.44%
3	1.38%	11.75%	47.08%	24.51%	9.48%	3.37%	2.44%
Grand Total	**1.40%**	**11.85%**	**47.50%**	**24.37%**	**9.32%**	**3.26%**	**2.29%**

Table 7: Cluster analysis - BER rating

- **HS Fuel:** The table below shows the household distribution cluster and HS fuel wise. While the general trend of the households has oil and gas as their major fuel, the cluster wise distribution differs. As observed, cluster 1 can be defined by electricity, oil and wood as the fuel and cluster 2 can be defined by solid fuel. These are exceptional distribution patterns than the general trend. The distribution of cluster 3 aligns with the distribution of the total.

HS Fuel	Electricity	Gas(LPG)	Oil	Solid	Wood
1	6.46%	43.39%	47.92%	1.96%	0.27%
2	4.78%	43.96%	43.54%	7.54%	0.18%
3	5.21%	43.01%	44.90%	6.69%	0.20%
Grand Total	**5.07%**	**43.72%**	**44.34%**	**6.68%**	**0.19%**

Table 8: Cluster Analysis - HS Fuel

- **Scheme count:** About 89% of the households have only 1 scheme that they have applied to, but this distribution differs when observed in clusters. In cluster 1, the majority of the households have applied for 1 and 2 schemes, whereas clusters 2 and 3 have the majority of households for only 1 scheme. Hence, to define, cluster 1 has households applying for multiple schemes i.e., >1, cluster 2 follows the trend for general households and cluster 3 households apply for a single scheme.

Scheme Count	1	2	3	4
1	55.47%	40.71%	3.76%	0.06%
2	92.82%	7.03%	0.15%	0.00%
3	97.61%	2.34%	0.05%	0.00%
Grand Total	**88.97%**	**10.43%**	**0.59%**	**0.01%**

Table 9: Scheme count

- **Measures:** Referring to table 10, around 40% of households have 2 measures and about 29% of households have more than 3 measures. Cluster 1 has 76.40% of households applying for more than 3 measures, cluster 2 has a majority of

households applying for at least 1 and mostly 2 measures. Cluster 3 has 19.30% of households that have no measures at all.

Measures	0	=1	=2	>=3
1	0.00%	1.76%	21.85%	76.40%
2	0.00%	32.01%	45.64%	22.35%
3	19.30%	29.56%	28.63%	22.51%
Grand Total	**3.55%**	**27.73%**	**39.51%**	**29.21%**

Table 10: Cluster Analysis - Measures

- **Submissions:** Submissions are the total number of applications made by the household (refer to the definition in the data description). 76.73% of households have only 1 submission and the rest have >1. Cluster 1, however, has no household with a single submission. All the households have made more than one submission. Cluster 2 and 3 have more than 85% of households having made only a single submission. Around 11% of the households in cluster 2 have made 2 submissions whereas cluster 3 has around 9% households.

Submissions	1	2	>2
1	0.00%	72.67%	27.33%
2	88.39%	10.95%	0.65%
3	85.76%	8.82%	5.42%
Grand Total	**76.73%**	**18.37%**	**4.90%**

Table 11: Cluster Analysis - Submissions

- **Approval Rate:** For each household, the approval rate was calculated as (Approved submissions/ Total submissions). This rate had distinct values and hence approval rate of 100% was labelled as Complete approval, 0% was labelled as Complete rejection and any rate between 0% and 100% was grouped as semi-approved. Table 12 shows the distribution of these labels with the cluster-wise distribution.

Approval rate	Complete approval	Complete rejection	Semi
1	23.51%	0.06%	76.43%
2	99.93%	0.00%	0.07%
3	0.00%	96.27%	3.73%
Grand Total	**71.90%**	**17.70%**	**10.40%**

Table 12: Cluster Analysis - Approval rate

From table 12, we observe that around 72% of households have complete approval on submitting. However, when observed cluster wise, cluster 1 has only about 23.51% completely approved households while 76.43% of households have a mix of approval and rejections on their submissions. Cluster 2 has 99.9% households having complete approval on their submissions and the contrary, cluster 3 has 96.27% households having complete rejection on their submissions.

6.3 Summary

Cluster 1 can be labelled as 'Potential households' since the households have been associated with making multiple applications for more than 6 months since their first application. They are extremely efficient households that utilise electricity, oil, and wood as fuel, which differs from the average household trend. Since the majority of the households have made multiple submissions, they have multiple schemes and measures, the highlighting approval rate is semi-approved. We believe that because of their interest in installing electrical appliances and their continual attempt to improve house efficiency, these are the households that are most likely to purchase an electric car.

Cluster 2 can be labelled as 'Motivated households' as the approval rate of the majority of these households is completely approved. They have also made predominantly 1 submission as they are only associated for at the most 6 months since the first submission. They outstand the overall household trend for fuel and show a keen interest in solid fuel, however, follow the trend for the number of schemes as the overall. Hence, the aim for us should be pushing a few of these households to cluster 1 i.e., to potential households. These households, based on their behaviour, show a great interest in increasing efficiency, and the majority of households' experiences have been positive so far since they have complete approval on submissions. We

believe that if a new scheme is introduced, these households will be more inclined to purchase an electric vehicle.

Cluster 3 can be classified as "non-ideal households" since it contains the bulk of the households with applications that were completely rejected. As a result, their desire to apply for another scheme/measure would be minimal. As they mostly have made one submission and the scheme, they have applied for is, by implication, one, their overall experience has not been particularly favourable. As a result, their desire to purchase an electric vehicle will be diminished.

Cluster	**1**	**2**	**3**
House Vintage	Old and semi	New	New
Efficiency	Very Efficient	In line with the household trend	In line with the household trend
HF Fuel	Electricity, Oil and Wood	Solid	In line with the household trend
Scheme Count	Multiple (>1)	In-line with household trend	Single
Measures	Many Measures	At least 1, Mostly 2	Mostly 0
Submission	At least 2 submissions	At max 2, predominantly 1	Predominantly 1
Approval Rate	Semi-approved	Complete Approval	Complete Rejections
EV purchase	Potential households	Motivated Households	Non-ideal households

Table 13: Cluster Profiles

Chapter 7 - Dashboard

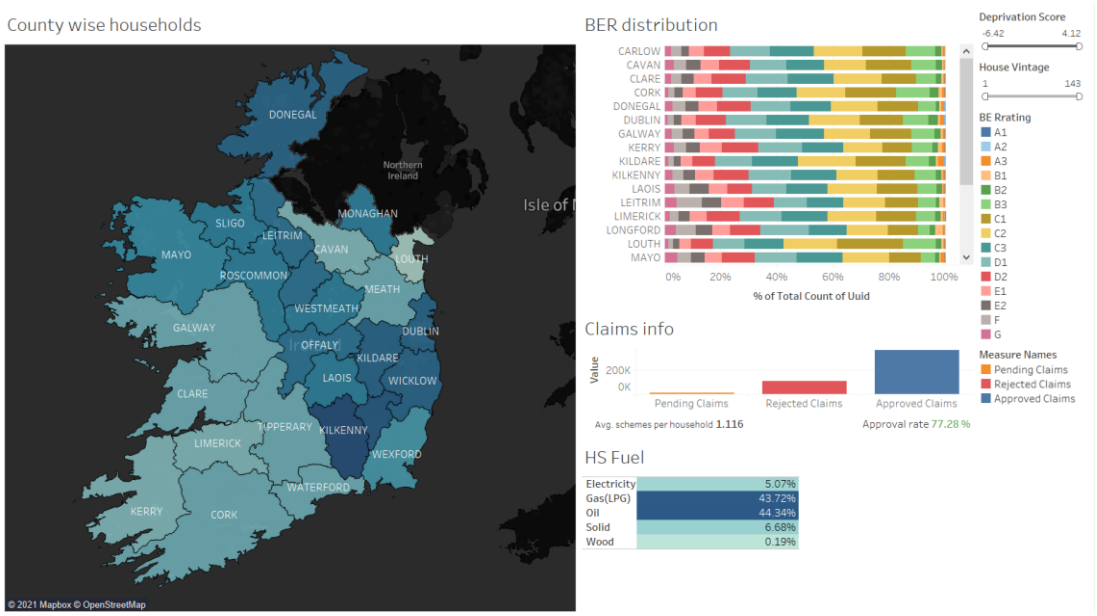

Figure 13: Snippet of Dashboard -first view

This is a snapshot from the first dashboard we created for general data analysis. This dashboard is interactive, allowing users to see the distribution of households by county, BER, HS fuel, and the status of claims. We've included the Deprivation score and house vintage filters. The colour shading of the counties in the map above corresponds to the county's Scope index. In this context, Scope Index is calculated as

$$Scope\ Index = 1 - \frac{No\ of\ households\ that\ made\ an\ application}{Total\ number\ of\ households} \qquad (9)$$

The darker the shade, the higher the scope, and vice versa.

Let's have a look at a few examples of how the dashboard appears when we choose and apply a few filters.

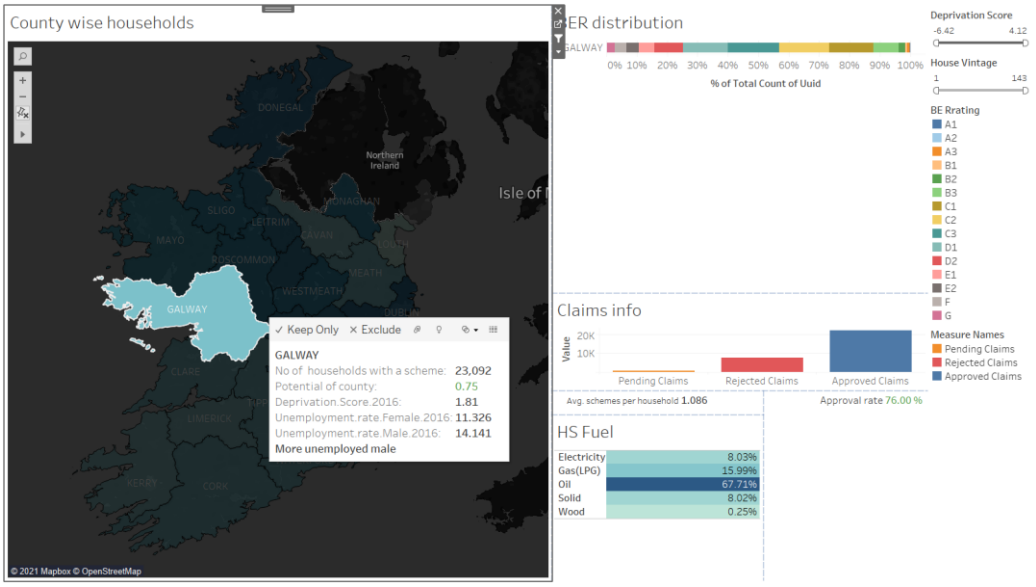

Figure 14: Snippet of the dashboard - County selection

Galway was chosen as the example county in figure 14. On the tooltip of the map, one can see the following information: the number of households in that county, the county's scope index, the deprivation score, the unemployment rate of male and female, and the last statement is a comparison of the unemployment rate of males and females in that county. We also see that because it is an interactive dashboard, the county selection dynamically changes the other charts on the screen.

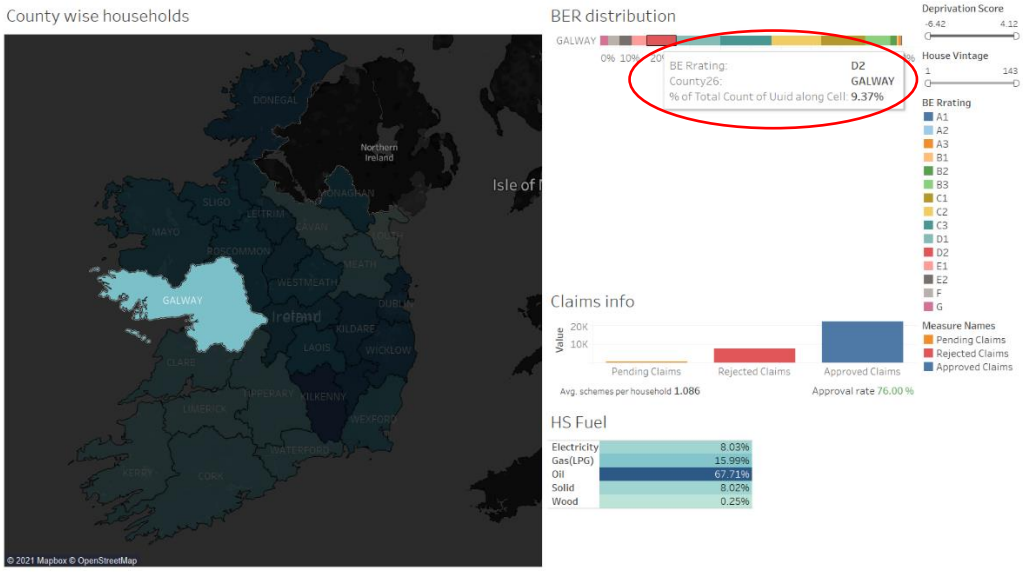

Figure 15: Snippet of the dashboard - BER chart

In figure 15, the tooltip of the BER distribution chart displays the BER Rating for that county, as well as the number of households in percentage that fall under that rating.

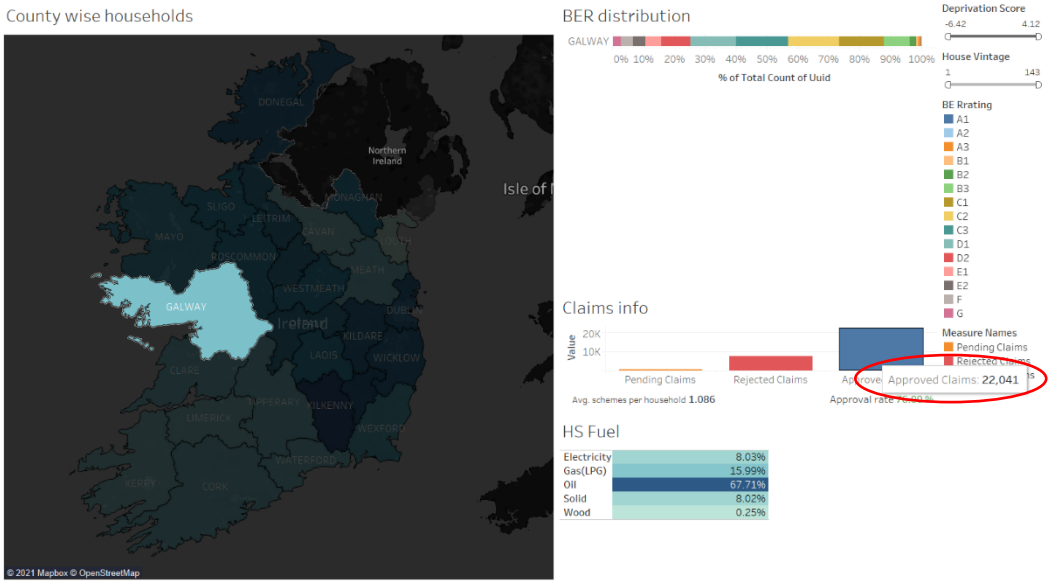

Figure 16: Snippet of the dashboard - Claims info

The claims info chart displays the distribution of the status of the application. The tooltip on approved claims in figure 16 provides the number of claims denied in that county.

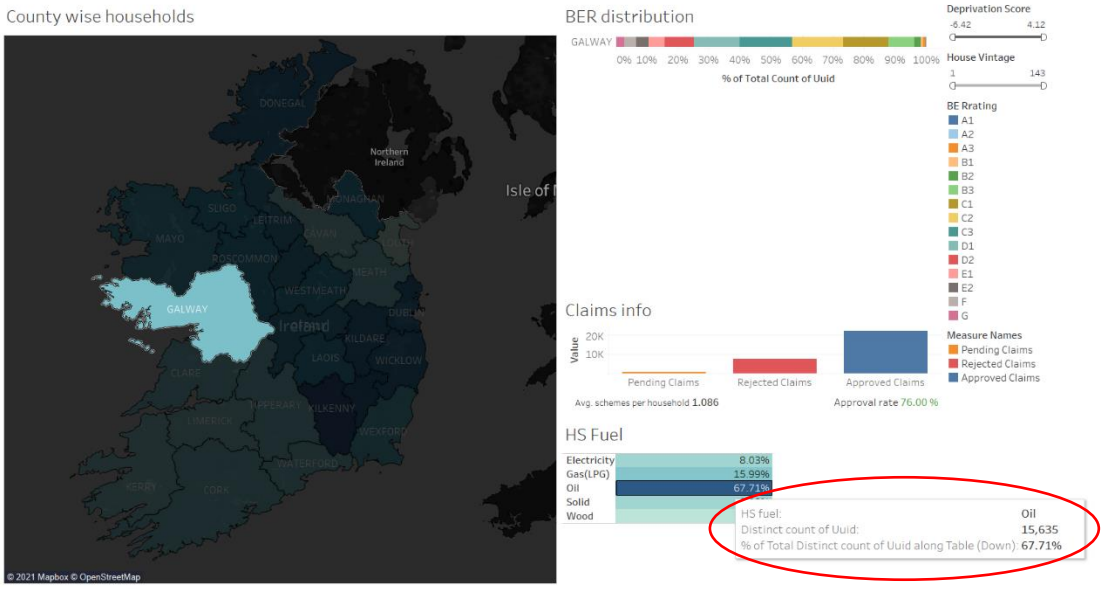

Figure 17: Snippet of the dashboard - HS Fuel

Figure 16 also depicts information such as the average schemes per household in the county. For this example, households in Galway had 1.08 schemes on average, with a 76 per cent approval rate on submission. The displayed tooltip of the HS Fuel chart is for Oil. It provides data such as the number of homes that use oil as a fuel, as well as the proportion of those that do. In addition to this, filters of deprivation score (counties with >0 deprivation score) or households in counties with vintage ranging from 7 months to 14 months, etc. can also be applied. Since all the filters are linked to the charts on the dashboard, they dynamically update and interact with one another. This dashboard conveys the overall behaviour of homes across Ireland in simple charts and helps us to cross-compare counties.

This dashboard was useful for our meeting with the stakeholders to explain the overall behaviour of homes across Ireland and develop business insight from it.

Chapter 8 - Proposed action

Looking at the entire landscape, we have about 13% of potential households, 69% of motivated households, and 18% of non-ideal residences. This indicates a good probability for the adoption of electric cars, as 1 and 2 account for 82 per cent of the population. While ESB intends to improve the entire charging grid by 2030, we aim to give guidance on which counties should be upgraded first.

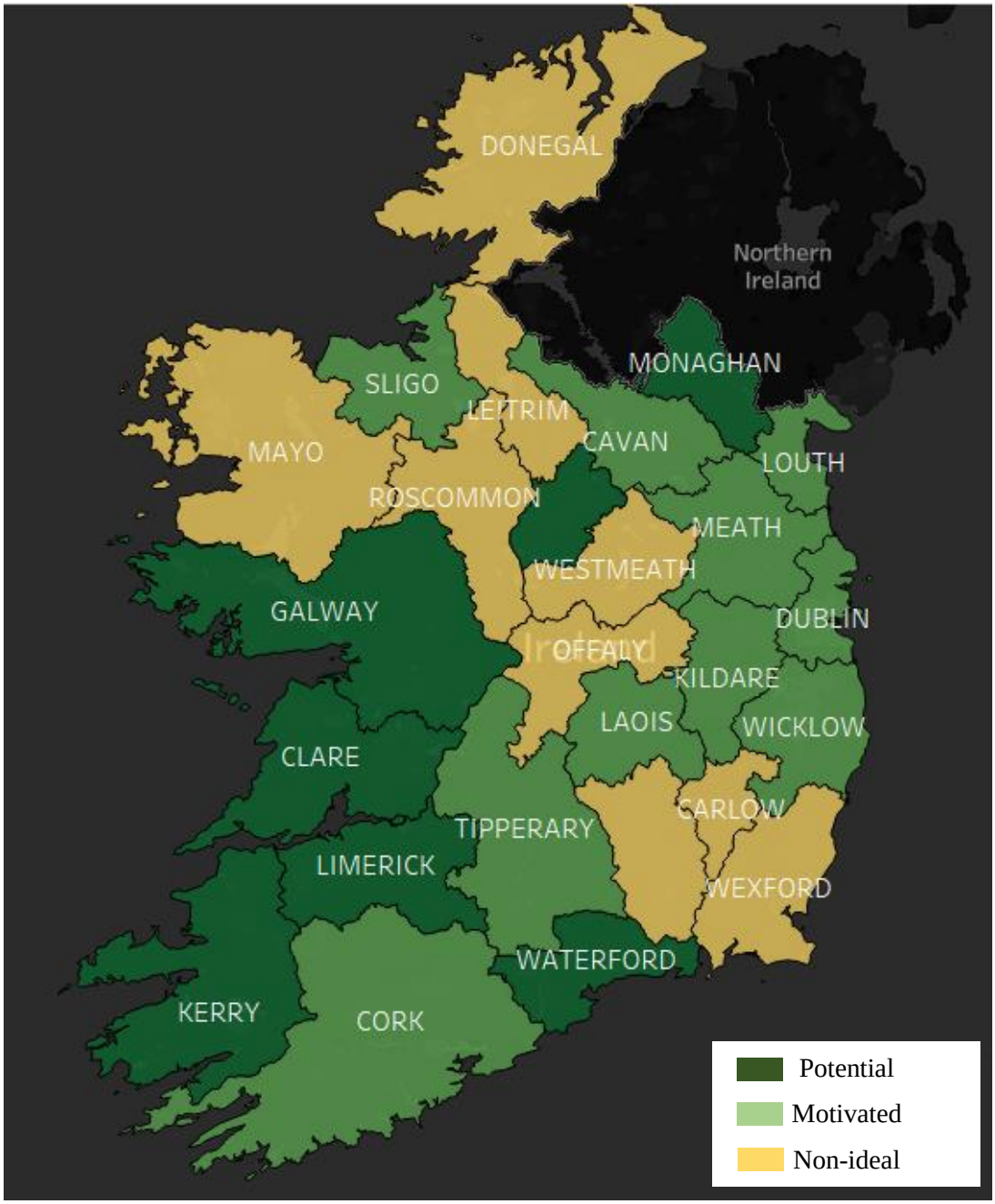

Figure 18: Cluster association with the counties

Looking at the absolute number of clusters for each county will not aid our analysis because the present electric grid is already designed proportionally for each county's population. As a result, we want to compare the proportion of our clusters in each county. This will assist us in identifying counties with homes that are embracing electric energy, necessitating the installation of additional power grids.

County	Potentials	Motivated	Non-Ideal
Carlow	11%	64%	25%
Cavan	12%	68%	20%
Clare	16%	66%	18%
Cork	14%	70%	16%
Donegal	10%	64%	26%
Dublin	12%	72%	16%
Galway	14%	67%	19%
Kerry	13%	67%	20%
Kildare	11%	70%	19%
Kilkenny	13%	65%	23%
Laois	11%	71%	17%
Leitrim	8%	67%	25%
Limerick	15%	68%	17%
Longford	13%	65%	22%
Louth	12%	70%	18%
Mayo	12%	63%	25%
Meath	11%	71%	18%
Monaghan	13%	65%	22%
Offaly	11%	68%	21%
Roscommon	11%	67%	23%
Sligo	10%	69%	21%
Tipperary	13%	70%	17%
Waterford	16%	66%	19%
Westmeath	12%	65%	23%
Wexford	12%	68%	20%
Wicklow	11%	69%	20%

Table 14: County-wise proportion of house profiles

We can infer from figure 18 that the counties of Kerry, Waterford, Limerick, Clare, Galway, Longford, and Monaghan have a comparatively high share of potential households (when compared to the general trend). This suggests an increasing trend in the use of sustainable energy in those counties, which would presumably be correlated with the purchase of an electric vehicle. As a result, these counties should be given priority when it comes to grid upgrades.

Motivated households are concentrated in the counties of Cavan, Cork, Dublin, Kildare, Laois, Louth, Meath, Sligo, Tipperary, and Wicklow. These are the households that are new to scheme claims (low count of installed measures and low house vintage) but are highly interested as their applications were successful. They follow the course of potential households and may eventually attain their status. In the event of a new electric car scheme, they are more inclined to purchase one. When such a scheme is introduced, these homes should be frequently checked for their applications, and the corresponding grid improvements should be facilitated.

When compared to the general proportion, the counties of Carlow, Donegal, Kilkenny, Leitrim, Mayo, Offaly, Roscommon, Westmeath, and Wexford have a high proportion of non-ideal households. This means that these counties may not expect a future increase in grid demand.

To further increase the quality of recommendations, we calculated the Scope Index of each county. Scope Index (9) of a county can be defined as the percentage of households in the county that never made an application. These are the households that require more awareness and better incentives to apply to a scheme. The denominator is the number of households in each county as reported in the 2016 census. Figure 14 represents the scope of each county.

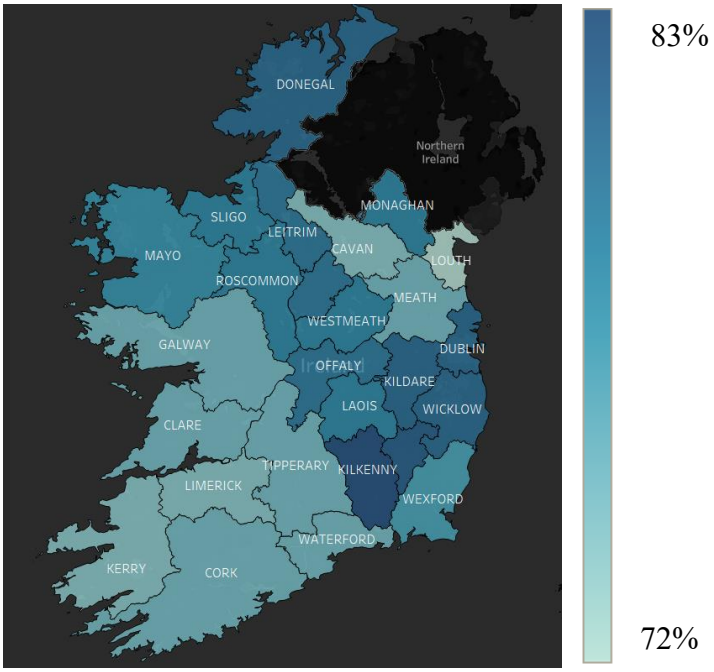

Figure 19: Scope of each county

From figure 19, we can infer that the above-listed counties have a relatively high scope index. They might lack the awareness of schemes and that's why don't have a lot of applications from those counties. Creating awareness measures and tailored schemes might motivate more houses from these counties to apply, pushing them to cluster 2.

Conclusion

The objective of this paper is to help the ESB identify areas where extra charging stations are needed. We examined SEAI scheme data and established indicators associated with the purchasing of electric cars. We used several data transformations to fine-tune the granularity of the data, and we also used 2016 census report information to improve the recommendations. We grouped the households based on their housing and scheme information, and then built profiles aligning with the project's scope.

The insights gained from this study assisted us in identifying identical households in Ireland in terms of electric measure adoption and determining the causes for their behaviour. We also investigated the distribution of these profiles across counties, linking them to county characteristics and developing strategies for each county based on the distribution.

We only had scheme-related information for each household in the data we worked with. While this allowed us to identify trends in different counties and develop personalized plans for each of them, we could provide a lot more insights if we could analyse the energy consumption history in each of them.

We can estimate future power requirements at the most granular level by correlating the demographic characteristics of the home with its consumption behaviour. This would allow us to better understand the behaviour patterns of each kind of household, identify the elements that contribute to a certain consumption trend, and assess how effective the SEAI schemes were in encouraging renewable energy adoption. While forecasting was part of the project's initial scope, the data provided was insufficient for a forecasting model.

We also had no way of knowing whether a household owns an electric car. This information would be extremely beneficial to the project's scope. It also assists us in determining each county's potential - some counties may have achieved an early saturation, whilst others may have a huge EV purchasing potential. We can map the factors that influence the purchase of an electric car and develop tailored schemes for each house type to encourage the adoption of electric vehicles.

Another area where we can improve is that if we have visibility into the current charging grid infrastructure, its capacity, and supply, we can model it as a facility location optimisation problem, using the forecasted power demand of the house profiles as the demand constraint, and identify the optimal location of the new charging stations. This will significantly assist ESB in building/upgrading grids capable of fulfilling future demand.

References

Department of Communications, Climate Action & Environment (2019). Climate Action Plan 2019. [online] Government of Ireland. Government of Ireland. Available at: https://www.gov.ie/en/publication/ccb2e0-the-climate-action-plan-2019/ [Accessed 23 Apr. 2021].

Wikipedia Contributors (2019). History of the electric vehicle. [online] Wikipedia. Available at: https://en.wikipedia.org/wiki/History_of_the_electric_vehicle.

Energy.gov. 2021. Electric Vehicle Benefits. [online] Available at: <https://www.energy.gov/eere/electricvehicles/electric-vehicle-benefits> [Accessed 7 June 2021].

McFadden, C., Bergan, B., Lang, F., Wendorf, M. and Trosper, J., 2021. A Brief History and Evolution of Electric Cars. [online] Interestingengineering.com. Available at: <https://interestingengineering.com/a-brief-history-and-evolution-of-electric-cars> [Accessed 23 May 2021].

Money Guide Ireland. 2021. Electric Cars in Ireland – Some Facts and Figures. [online] Available at: <https://www.moneyguideireland.com/electric-cars-facts-figures.html> [Accessed 23 May 2021].

Sandyfordmotorcentre.com. 2021. The Advantages of Electric Cars | Sandyford Motor Centre. [online] Available at: <https://www.sandyfordmotorcentre.com/article/the-advantages-of-electric-cars#:~:text=A%20large%20benefit%20of%20electric,fuels%20polluting%20the%20earth's%20air.> [Accessed 7 June 2021].

Statista. 2021. Global plug-in electric light vehicle sales 2020 | Statista. [online] Available at: <https://www.statista.com/statistics/665774/global-sales-of-plug-in-light-vehicles/> [Accessed 23 May 2021].

IEA (2020). Global EV Outlook 2020 – Analysis. [online] IEA. Available at: https://www.iea.org/reports/global-ev-outlook-2020.

automotive-institute.kpmg.de. (n.d.). KPMG's Automotive Institute Publication Platform. [online] Available at: https://automotive-institute.kpmg.de/GAES2019/?m=0 [Accessed 8 Jun. 2021].

Jain, A.K., Murty, M.N. & Flynn, P.J. 1999.Data clustering: A review., ACM Computing Surveys, vol. 31, no. 3, pp. 264-323.

Zhao, Y., Karypis, G. and Fayyad, U., 2005. Hierarchical Clustering Algorithms for Document Datasets. Data Mining and Knowledge Discovery, 10(2), pp.141-168.

von Luxburg, U. (2007). A tutorial on spectral clustering. Statistics and Computing, [online] 17(4), pp.395–416. Available at: https://arxiv.org/pdf/0711.0189.pdf [Accessed 23 Dec. 2019].

Hruschka, E.R., Campello, R.J.G.B., Freitas, A.A. & de Carvalho, A.C.P.L.F. 2009.A survey of evolutionary algorithms for clustering. IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews, vol. 39, no. 2, pp. 133-155.

Ankerst, M., Breunig, M., Kriegel, H. and Sander, J., 1999. OPTICS. ACM SIGMOD Record, 28(2), pp.49-60.

U. Maulik and S. Bandyopadhyay, "Performance evaluation of some clustering algorithms and validity indices," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 12, pp. 1650-1654, Dec. 2002, DOI: 10.1109/TPAMI.2002.1114856.

Newman, M. and Leicht, E., 2007. Mixture models and exploratory analysis in networks. Proceedings of the National Academy of Sciences, 104(23), pp.9564-9569.

De Gooijer, J.G. and Hyndman, R.J. (2005). 25 Years of IIF Time Series Forecasting: A Selective Review. SSRN Electronic Journal.

Rao, J.N.K., Box, G.E.P. and Jenkins, G.M. (1972). Time Series Analysis Forecasting and Control. Econometrica, 40(5), p.970.

Holt, C.C. (2004). Forecasting seasonals and trends by exponentially weighted moving averages. International Journal of Forecasting, 20(1), pp.5–10.

Zhang, G.Peter. (2003). Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing, [online] 50, pp.159–175. Available at: https://www.sciencedirect.com/science/article/pii/S0925231201007020.

Hyndman, R.J. and Khandakar, Y. (2008). Automatic Time Series Forecasting: The forecast Package for R. Journal of Statistical Software, 27(3).