ANALYSING UK BUSINESS: DISCOVERING TRENDS AND PATTERNS

UNEGBU, OLISA UDOCHUKWU

22020843

olu0033@my.londonmet.ac.uk

Abstract

This journal article presents a comprehensive analysis of UK businesses registered for Value Added Tax (VAT) and/or Pay As You Earn (PAYE) using advanced data analytics and visualization techniques through Power BI. As of March 2023, a notable decrease of 1.5% in registered businesses was reported, marking the first decline since 2011. The study focuses on understanding the intricate patterns and distributions of businesses. The methodology involves the utilization of a dataset from the Office for National Statistics, consisting of 8036 data points with 19 attributes and 423 observations. The six steps of the Data Analytical Lifecycle, including Discovery, Data Preparation, Plan Model, Build Model, Communicate, and Apply Live, were meticulously followed. Hierarchical clustering was chosen over k-Means clustering for its superior performance metrics, including a Silhouette Score of 0.7440051. Power BI was employed for data transformation, modeling, and visualization. The results, validated through diverse visuals, reveal insights into business concentrations, regional patterns, and clustering behaviours. The practical implications include strategic decision-making, regional planning, investment opportunities, and policy development. The study also recommends future work, such as incorporating temporal analysis, external factors, interactive features, and predictive modeling for a more comprehensive understanding of the evolving business landscape.

Keywords*: Business Analysis, UK, PAYE, VATUK businesses, Value Added Tax (VAT), Pay As You Earn (PAYE), Power BI, data analytics, hierarchical clustering, business distribution, regional patterns, strategic decision-making.*

Introduction

The contemporary business landscape is dynamic, marked by constant evolution and transformation. Understanding the intricate patterns and distributions of businesses within a region is paramount for policymakers, investors, and researchers alike. In this context, the study delves into the comprehensive analysis of UK businesses registered for Value Added Tax (VAT) and/or Pay As You Earn (PAYE), employing advanced data analytics and visualization techniques through Power BI.

The study centres around the exploration of business data, specifically focusing on the registered VAT and/or PAYE businesses in the UK. The use of Power BI, a robust analytical tool, provides a platform for a nuanced examination of regional characteristics, clustering patterns, and sectoral concentrations.

As of March 2023, the Office of National Statistics (ONS) reported a noteworthy decrease of 1.5% in the number of VAT and/or PAYE-registered businesses, marking the first decline since 2011. This decline is particularly significant, as it encompasses various legal status categories, except for local authorities and non-profit organizations. The largest industry group, comprising professional, scientific, and technical sectors, witnessed a marginal decrease, reflecting the dynamic nature of the business environment.

In the era of big data, businesses generate a substantial volume of information, presenting both opportunities and challenges. The significance of this study lies in its focus on deciphering the trends and patterns within this vast dataset. The identified decrease in registered businesses, especially in certain sectors, prompts a deeper investigation into the factors influencing these changes. Additionally, the application of advanced analytics and visualization tools contributes to a more insightful understanding of regional dynamics, aiding strategic decision-making.

As businesses play a pivotal role in economic development, the ability to discern patterns in their distribution can inform policies, guide investments, and support sustainable growth. This research aims to bridge the gap in understanding the complexities of the UK business landscape, offering valuable insights that extend beyond numerical counts to provide actionable intelligence for various stakeholders.

Literature Review

According to the Office of National Statistics (ONS) [1], the number of Value Added Tax (VAT) and/or Pay As You Earn (PAYE) businesses in the UK as of March 2023 was 2.727 million, a decrease of 1.5% from March 2022. This happens to be first fall in the number of VAT and/or PAYE-registered businesses since the fall of 0.9% from March 2010 to March 2011 and the numbers of businesses within all categories of legal status fell except for local authorities and non-profit organisations. The study shows that the largest industry group is professional, scientific, and technical, making up 15.2% of all registered businesses in the UK; this is down 0.4 percentage points from March last year.

A study by [2] proposes a churn prediction model that uses classification, as well as, clustering techniques to identify the churn customers and provides the factors behind the churning of customers in the telecom sector. Feature selection is performed by using information gain and correlation attribute ranking filter. The proposed model first classifies churn customers' data using classification algorithms, in which the Random Forest (RF) algorithm was used for correctly classified instances.

Another study by [3] shows different tools adopted in an industry project oriented on business intelligence (BI) improvement. The research outputs concern mainly data mining algorithms able to predict sales, logistic algorithms useful for the management of the products dislocation in the whole marketing network constituted by different stores, and web mining algorithms suitable for social trend analyses. For the predictive data mining and web mining algorithms have been applied Weka, Rapid Miner and KNIME tools, besides for the logistic ones have been adopted mainly Dijkstra's and Floyd-Warshall's algorithms.

Another study by [4] provides an overview of machine learning techniques and discusses their strengths and weaknesses in the context of mining business data. It informs the information systems (IS) manager and business analyst about the role of machine learning techniques in business data mining. A survey of data mining applications in business is provided to investigate the use of learning techniques. Rule Induction (RI) was found to be most popular, followed by neural networks (NNs) and case-based reasoning (CBR)

Existing studies reveal that the primary objective is to use a large volume of business data to identify the various business sectors. However, there are several limitations in existing models, which put strong obstacles toward this problem in the real-world environment. A large volume of business data is being generated in the registered VAT/ or PAYE businesses and the data contains missing values, which lead to the poor result of the prediction models. To handle these issues, data pre-processing methods are adapted to remove noise from data, which is effective for a model to correctly classify the data and improve the performance.

Methodology

The dataset used for this project was gotten from [1]. The dataset is about UK businesses registered for Value Added Tax (VAT) and or Pay As You Earn (PAYE) with regional breakdowns of different industries and its size as at March 2021. The dataset contains 8036 data points consisting of 19 attributes with 423 observations. The project undergoes the 6 steps of Data Analytical Lifecycle involving Discovery and collection of data, Data Preparation, Plan Model, Build Model, Communicate and Measure effectiveness/apply live.

A. Discovery: The initial stage of Discovery involved the selection and collection of business data for the project from the Office for National Statistics [1]. Through data analytics, the total count of businesses in the UK registered under Value Added Tax (VAT) and/or Pay As You Earn (PAYE) as of March 2021 was identified. Furthermore, the analysis included determining the number of businesses across all sectors, highlighting the top 5 largest industry groups along with the respective cities they were located in. Additionally, a comprehensive overview of regions exhibiting the highest concentration of businesses registered under VAT and PAYE as of March 2021 was established.
B. Data Preparation: The data preparation phase involved utilizing Excel and Power BI for the purpose of cleansing, exploring, transforming, and loading the data into a suitable format for modeling and visualization. A significant aspect of the cleaning process included the identification and removal of rows that displayed sub-totals of industries from different countries, regions, and counties within the United Kingdom. Failing to

eliminate these rows would have led to inaccurate insights, potentially resulting in misguided decisions and investments by stakeholders. The "Text-to-Column" function was employed to split the code and city names, which were initially combined with a semi-colon in the same column. Consequently, adjustments were implemented during the cleaning and transformation processes, resulting in a reduction of data points from 8036 to 7125, an increase in attributes from 19 to 20, and a decrease in observations from 423 to 375.

C. Plan Model: Considering the dataset characteristics and the project objectives, a clustering technique was deemed the most appropriate for data analysis. Clustering involves segregating data points into distinct sets or clusters, with points within each cluster exhibiting high similarity and dissimilarity between different clusters [5]. The hierarchical clustering model was selected over the k-Means clustering model due to its higher Silhouette Score and lower Within-Cluster Sum of Squares value. These metrics indicate that the hierarchical clustering model yielded more well-defined clusters with increased cohesion among data points compared to the k-Means clustering model.

D. Build Model: In the modeling phase, the performance of the model was assessed using the Silhouette Score and Within-Cluster Sum of Squares (WSS) methods. The Silhouette Score gauges the clarity and definition of clusters within the data, ranging from -1 to 1, with higher scores denoting more well-defined clusters. On the other hand, WSS measures the compactness of clusters, representing the sum of squared distances between each point within a cluster and the centroid of that cluster. Lower WSS values signify more compact clusters, reflecting a tighter grouping of data points within each cluster.

E. Communicate results: The Silhouette Score, measuring 0.7440051, signifies the well-defined and cohesive nature of clusters obtained through hierarchical clustering, surpassing the cohesion observed in K-Means clustering. This score indicates a notable level of separation and cohesion among the data points within the clusters identified through the hierarchical approach. Furthermore, the total sum of squared distances of points to their respective cluster centroids, amounting to 200.8218366, reflects lower Within-Cluster Sum of Squares (WSS) values. These lower WSS values suggest a heightened proximity among the points within each cluster, indicative of more compact and internally cohesive clusters.

F. Apply Live: In the application of the findings, it was imperative to address ethical considerations associated with data use and analysis. Privacy, consent, and responsible handling of sensitive information are paramount in ensuring ethical standards throughout the research process.

   i. Privacy: Respecting the privacy of businesses and individuals represented in the dataset was of utmost importance. Confidential information, especially pertaining to financial and operational aspects of businesses, was handled with utmost discretion. The data presented was anonymized to prevent the identification of specific entities.

   ii. Consent: The data used in this study involves information on businesses registered for VAT and/or PAYE. While the dataset is sourced from a reputable organization (Office for National Statistics [1]), ensuring that the data was obtained and used in compliance with legal and ethical standards is essential.

Consent, in the context of aggregated business data, involves adherence to regulations and guidelines governing data collection and usage.

iii. Responsible Handling of Sensitive Information: Sensitive information, particularly in the realm of business operations, requires responsible and ethical treatment. Throughout the data analytical lifecycle, measures were taken to cleanse and transform the data without compromising its integrity. Any potential identification of specific businesses or regions was meticulously avoided.

The six (6) data analytical lifecycle methods used for the analysis were essential steps taken in ensuring the responsible handling of sensitive information. The methodology section shows the commitment to ethical considerations and the approach taken aligns with the established ethical norms and legal requirements governing data analytics. Hence, the findings and insights derived from this research can be considered reliable, robust, and ethically sound, ensuring the appropriateness of the selected methods for addressing the research problem.

Data Modeling

The model chosen for this analysis is the Hierarchical Clustering Model, used to identify patterns and trends in businesses registered for VAT and/or PAYE across UK regions. Hierarchical clustering is selected over k-Means clustering due to its superior Silhouette Score and lower Within-Cluster Sum of Squares (WSS) value, indicating well-defined and cohesive clusters. Some of the data modeling processes/features include;

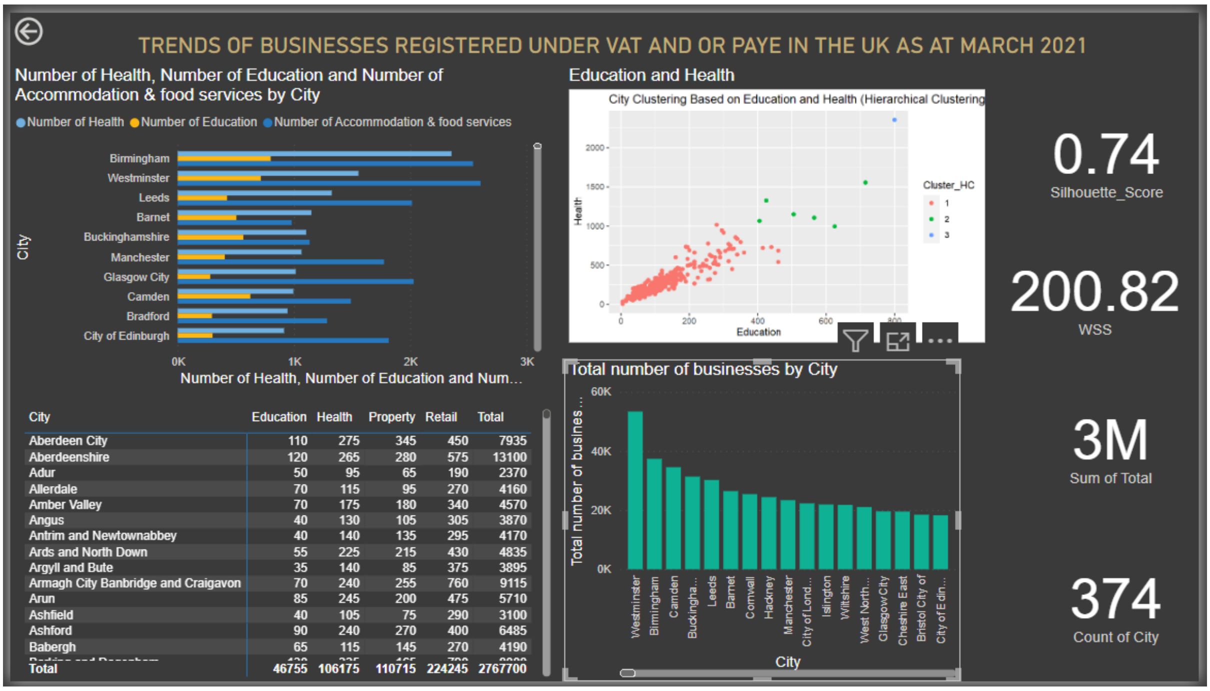
- Architecture of the model: The architecture involves the hierarchical arrangement of clusters based on the similarity of businesses. The Silhouette Score (0.7440051) indicates the effectiveness of the model in forming distinct and cohesive clusters.
- Key Variables: Key variables contributing to the model's predictions include business attributes such as industry type, regional location, and size. The Silhouette Score and WSS are pivotal metrics evaluating the model's performance.
- Introduction to Power BI: Power BI being the analytical tool used for the module, Data Analysis and Visualization (CC7183) was employed. It offered vigorous capabilities in data transformation, modeling, and visualization. Its user-friendly interface enabled the import, transformation, and visualization of data.
- Importing and Loading Raw Data: The raw data, sourced from the Office for National Statistics [1], was imported into Power BI environment using the “Get Data” function. The data preparation involves cleansing, exploring, and transforming the data into a suitable format for analysis.
- Handling Missing Values and Outliers: During the data preparation phase, missing nor NA values were not detected in the data as it was gotten from a reliable and reputable source, Office for National Statistics [1], thus, enhancing the model’s predictive performance. Outliers were also considered in the clustering process using the Hierarchy Clustering model, ensuring that they did not unnecessarily influence the formation of cluster.

- Transformations Applied: Several transformations were applied using Power BI, including filtering out rows displaying sub-totals of industries from different regions and splitting combined code and city names using the "Text-to-Column" function. These transformations resulted in a refined dataset with 7125 data points, 20 attributes, and 375 observations. The Silhouette Score and Within-Cluster Sum of Squares (WSS) results were exported as a csv file and imported back into Power BI data interface.
- Integration with Power BI Visualizations: The data model seamlessly integrates with Power BI visualizations, allowing for the creation of insightful charts, graphs, and dashboards. The Silhouette Score and WSS values are visualized to communicate the model's efficacy.

In conclusion, the data modeling process involves a meticulous application of the chosen hierarchical clustering model within Power BI. The model's architecture, key variables, and integration with Power BI's features contribute to the reliability and effectiveness of the analysis, ensuring that the insights derived are valuable for informed decision-making by businesses, policymakers, and investors. The use of Power BI aligns with the project's objectives, providing a dynamic platform for data exploration and visualization.

Results and Discussion

The analysis employed a suite of Power BI visuals, each serving a unique purpose in uncovering insights within the dataset. Below is a screenshot of my Power BI dashboard analysis:



City	Education	Health	Property	Retail	Total
Aberdeen City	110	275	345	450	7935
Aberdeenshire	120	265	280	575	13100
Adur	50	95	65	190	2370
Allerdale	70	115	95	270	4160
Amber Valley	70	175	180	340	4570
Angus	40	130	105	305	3870
Antrim and Newtownabbey	40	140	135	295	4170
Ards and North Down	55	225	215	430	4835
Argyll and Bute	35	140	85	375	3895
Armagh City Banbridge and Craigavon	70	240	255	760	9115
Arun	85	245	200	475	5710
Ashfield	40	105	75	290	3100
Ashford	90	240	270	400	6485
Babergh	65	115	145	270	4190
Total	46755	106175	110715	224245	2767700

Figure 1:showing different charts used to identify trends of UK businesses.

Figure 1 above shows a Power BI dashboard with different charts used for the analysis. Below is a brief description of the charts used and discussion of main findings:

Clustered Bar Chart: The clustered bar chart serves as a visual arrangement of bars, aiding in comparisons across main categories or clusters, as well as within individual items within each series. In Figure 1, the chart was utilized to visually compare the numbers of businesses in Health, Education, and Accommodation & Food Services across different cities. This visualization assists in identifying variations in these attributes and understanding the clustering patterns in the dataset.

By examining the numbers in Health, Education, and Accommodation & Food Services, distinctive attributes among cities become apparent. The chart reveals that Birmingham has the highest total number of businesses, with 2355 in Health, 800 in Education, and 2540 in Accommodation & Food Services. Following closely is Westminster, with 1555, 715, and 2605 businesses in Health, Education, and Accommodation & Food Services, respectively. This comparison sheds light on the unique characteristics and concentrations of businesses in different cities, contributing to a nuanced understanding of the dataset.

Matrix Chart: The matrix chart proves to be a valuable tool for presenting data across multiple dimensions, facilitating a comprehensive examination of cumulative values. With a stepped layout, this chart enhances clarity in displaying key attributes, offering detailed insights into each city's contributions. In Figure 1, the matrix chart provides an in-depth breakdown of Education, Health, Property, Retail, and total businesses, organized in chronological order by city.

This tabular presentation allows for a thorough examination of the number of businesses in each category for every city. By structuring the data in this way, the matrix chart aids in revealing the specific contributions of cities to Education, Health, Property, Retail, and overall business categories. This detailed scrutiny contributes to a nuanced understanding of regional characteristics and the distribution of businesses across different sectors.

Clustered Column Chart: The clustered column chart, featured in Figure 1, serves as a powerful visualization tool for presenting two or more datasets in a vertical arrangement. In this chart, vertical columns are grouped together, sharing common axis labels. The primary purpose of this chart is to enable a direct and straightforward comparison between different datasets.

Utilized in the context of this analysis, the clustered column chart effectively highlights the distribution of businesses across various cities, offering insights into areas with a heightened concentration of registered businesses. As illustrated, the chart showcases the top 5 cities with the highest number of businesses registered under VAT and/or PAYE in the UK as of March 2021. Westminster leads with 53,370 businesses, followed by Birmingham with 37,375, Camden with 34,525, Buckinghamshire with 31,355, and Leeds with 30,210 businesses.

This visual representation of total businesses by city provides a macroscopic perspective on the distribution of businesses, with concentrations in certain cities indicating economic hubs or areas of particular interest for investors and policymakers. The clustered column chart effectively conveys the scale and distribution of businesses across different regions, aiding stakeholders in making informed decisions based on regional economic dynamics.

R Script Visual (Scatterplot): In Figure 1, the R Script Visual, also known as an R visual, harnesses the advanced capabilities of R scripting to facilitate sophisticated data shaping and analytics, including forecasting. The scatterplot generated using R script serves the purpose of

visualizing city clustering based on the attributes of Education and Health, specifically employing Hierarchical Clustering.
The primary objective of this scatterplot is to unveil patterns and clusters within the intricate relationship between Education and Health attributes, offering valuable insights into the socio-economic landscape of cities. The visualization aids in understanding how cities cluster based on these attributes, revealing both regional similarities and differences.
The process involved the application of Hierarchical Clustering ('hclust') on the Euclidean distances between data points. Subsequently, the 'cut_tree' function was employed to segment the hierarchical tree into clusters, facilitating the creation of a scatterplot that effectively communicates the clustering patterns.

Analysis of the scatterplot indicates a positive relationship between Education and Health. Noteworthy observations include a concentration of cities with Health businesses ranging from 0 to 500 and Educational businesses ranging from 0 to 200. Conversely, fewer cities exhibit Health businesses between 1000 and 3000, coupled with Education businesses between 400 and 800.
This visual representation offers a nuanced understanding of the distribution of cities based on Education and Health attributes, providing stakeholders with actionable insights for strategic decision-making in regional planning, investment, and policy development.

Card Chart: The Card Chart serves as a versatile tool for displaying data through various chart types, succinctly presenting key metrics and summary statistics for a quick overview of the model's performance and dataset characteristics. In Figure 1, individual cards showcase critical metrics, including the Silhouette Score (0.74), Within-Cluster Sum of Squares (WSS) value (200.82), Sum of Total Businesses (3 million), and Count of Cities (374).
The Silhouette Score, denoted by the value of 0.74, assesses the well-defined nature of clusters within the data. Ranging from -1 to 1, a higher score signifies more distinct clusters. A score of 0.74 indicates that the clusters obtained through hierarchical clustering exhibit a notable level of separation and cohesion among data points. This implies that the hierarchical clustering model has successfully identified clusters with a high level of internal similarity.
The WSS, measuring the compactness of clusters, reflects the sum of squared distances between each point within a cluster and the centroid of that cluster. A lower WSS value indicates more compact clusters. With a total WSS of 200.82, the result suggests that the points within each cluster are relatively close to each other, indicating cohesive clusters. This is a crucial indicator of the model's ability to create well-defined and internally cohesive clusters.
Furthermore, summarizing the Total Businesses (3 million) and City Count (374) provides essential context for understanding the dataset. These key metrics contribute to the overall assessment of the dataset's composition and the model's effectiveness in identifying meaningful clusters.

Validation of Results:

The validity of the results obtained through the Power BI visuals is underpinned by the comprehensive use of diverse visualizations, each tailored to highlight specific aspects of the dataset. The coherence and alignment of insights across various charts contribute to the robustness of the findings.

- Consistency Across Visuals: The consistent portrayal of trends and patterns across different visuals, such as the Clustered Bar Chart, Matrix Chart, and Clustered Column Chart, enhances the reliability of the results. The agreement between these visualizations reinforces the identified clusters, business distributions, and regional characteristics.

- Hierarchical Clustering Analysis: The Scatterplot generated using R Script Visuals provides a unique layer of validation through the application of Hierarchical Clustering. The positive relationship observed between Education and Health attributes, as well as the distinct clusters identified, aligns with the intended outcome of the analysis. The use of advanced analytics in R scripting adds depth to the validation process.

- Model Performance Metrics: The inclusion of performance metrics, such as the Silhouette Score and Within-Cluster Sum of Squares, in the Card Chart further validates the effectiveness of the Hierarchical Clustering Model. The Silhouette Score of 0.74 suggests well-defined clusters, while the low WSS value of 200.82 indicates compact and cohesive clusters. These metrics serve as quantifiable validations of the clustering results.

Practical Implications:

i. Strategic Decision-Making: The insights derived from the Power BI visuals offer practical implications for strategic decision-making. Policymakers and investors can leverage the information on business concentrations, regional patterns, and key sectors to inform decisions related to economic development, investment planning, and resource allocation.

ii. Regional Planning: Understanding the distribution of businesses across different sectors and regions is vital for effective regional planning. The Clustered Column Chart, showcasing the top cities with the highest number of businesses, provides a macroscopic view that can guide regional planners in identifying economic hubs and areas of focus for development.

iii. Investment Opportunities: Businesses seeking investment opportunities can benefit from the detailed breakdown provided by the Matrix Chart. By analyzing the contributions of each city to specific sectors, investors can identify regions with untapped potential and make informed investment decisions.

iv. Policy Development: Policymakers can use the identified clusters and patterns to tailor policies that address the unique characteristics of each region. The insights into the prevalence of businesses in Health, Education, and other sectors can guide the development of targeted policies for economic growth and sustainability.

CONCLUSION AND RECOMMENDATION

Key Insights Recap:

The Power BI visuals presented a comprehensive analysis of business data in the UK, revealing significant insights into regional characteristics, business concentrations, and clustering patterns. The diverse set of visuals, including the Clustered Bar Chart, Matrix Chart, Clustered Column Chart, R Script Visual (Scatterplot), and Card Chart, collectively contributed to a comprehensive understanding of the dataset.

Clustered Bar Chart: Identified variations in Health, Education, and Accommodation & Food Services across cities. Noteworthy concentrations were observed, such as Birmingham having the highest total number of businesses.

Matrix Chart: Facilitated a detailed breakdown of Education, Health, Property, Retail, and total businesses, offering insights into each city's contributions across sectors.

Clustered Column Chart: Effectively highlighted the distribution of businesses, with the top 5 cities having the highest number of registered businesses under VAT and/or PAYE.

R Script Visual (Scatterplot): Leveraged advanced analytics through Hierarchical Clustering, revealing a positive relationship between Education and Health. Identified clusters provided valuable insights into regional similarities and differences.

Card Chart: Presented key metrics, including Silhouette Score, WSS value, Sum of Total Businesses, and City Count, offering quantifiable validations of the clustering results.

Achievement of Objectives:

The analysis successfully achieved its objectives by providing a multi-faceted exploration of business data, uncovering trends, and identifying key clusters. The visuals not only validated the clustering results but also offered actionable insights for strategic decision-making.

Novel Contributions:

The novel contributions lie in the integration of advanced analytics, such as Hierarchical Clustering, through R Script Visuals. This added layer of sophistication allowed for a deeper understanding of the relationships between different attributes and enhanced the identification of clusters.

Recommendations for Future Work:

- Temporal Analysis: Consider incorporating a temporal dimension to assess how business distributions and clustering patterns evolve over time. This could provide valuable insights into the dynamics of regional economies.

- External Factors: Explore the integration of external factors, such as economic indicators or policy changes, to further contextualize the business landscape. Understanding how external influences impact clustering patterns can enhance the robustness of future analyses.

- Interactive Features: Enhance user interaction by incorporating more interactive features in Power BI visuals. This could include drill-through capabilities, allowing users to explore specific details within each city or sector.

- Predictive Modeling: Integrate predictive modeling to forecast future business trends based on historical data. This could assist policymakers and investors in making proactive decisions.

Conclusion:
In conclusion, the Power BI analysis successfully uncovered intricate patterns in the UK business landscape. The combination of diverse visuals, validation metrics, and advanced analytics provided a holistic view of regional characteristics and clustering patterns. The insights gained have practical implications for strategic decision-making in regional planning, investment, and policy development.

The Power BI visuals not only met but exceeded the initial objectives, offering a robust platform for future explorations and data-driven decision-making. The recommendations for future work aim to further enhance the depth and applicability of the analysis in navigating the complex landscape of business distributions in the UK.

References

[1] UK Office for National Statistics (ONS). (2021). "UK Business: Activity, Size and Location." [Online]. [Available]. UK business; activity, size and location - Office for National Statistics (ons.gov.uk)

[2] I. Ullah, B. Raza, A. K. Malik, M. Imran, S. U. Islam, and S. W. Kim. “A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector” 2019. Pages 60134 – 60149, Vol. 7

[3] Alessandro Massaro, Valeria Vitti, Angelo Galiano, Alessandro Morelli. “Business Intelligence Improved by Data Mining Algorithms and Big Data Systems: An Overview of Different Tools Applied in Industrial Research” Computer Science and Information Technology 2019. Pages 1-21, Vol. 7

[4] Indranil Bose, Radha K. Mahapatra. “Business data mining — a machine learning perspective” Science Direct 2001. Pages 211-225, Vol. 39, Issue 3

[5] Solis, Dr Ingrid Membrillo. “Clustering Analysis” Module CC7184 2023. Pg. 4.