JANUARY 2, 2022

INVESTIGATING VARIATION IN BROADBAND ACCESS BY REGION ACROSS THE UK

CODE FIRST: GIRLS NANODEGREE GROUP PROJECT

Team: Fardowsa Abdulkadir, Kim Clements, Isabelle Felsner, Yehjoon Kim, Lydia Maisander, Chinwendu Nwaozuzu

Contents

INTRODUCTION

Access to the internet is now considered a vital part of modern life, with the recent pandemic highlighting the importance of reliable connections as people have adapted to new ways of working and studying from home. The UK's broadband network provides this critical infrastructure and there has been heavy investment over recent years in rolling out faster fibre links across the country. From our own personal experiences, we knew that access is not equal with specific areas receiving higher/lower quality connections. We wanted to investigate this further and consider possible causes and potential solutions with regard to improving equality of access. Our broad project aims were as follows:

- Provide an overview picture of disparities in internet access across the country.
- Investigate the relationship between internet infrastructure/ accessibility and socio-economic factors in communities across the UK.
- Produce analysis and visualisations that convey our findings in a clear manner, providing useful insights to the topic.

BACKGROUND

We view our project as being relevant to audiences of two areas/industries, Government & Telecommunications, as these are responsible for providing the infrastructure. Our findings should be used to provide an overview picture of disparities across the UK and insight into possible links to socio-economic factors. This will allow the audience to better understand the current situation and consider causes and potential solutions with regard to improving equality of access.

With this in mind, we have refined our project objectives to answering the following three questions:

1. **What is the difference of infrastructure between Rural and Urban areas?**
   The logistics of installing fibre can be affected by the physical geography of locations, do we see this reflected in disparities between urban and rural areas? Can we identify particular areas where a Telecommunications company can invest into schemes in certain areas?
2. **Can we identify trends over time?**
   How has the rollout of fibre connections progressed over time? Is this correlated to trends in income?
3. **What is the nature of the relationship between internet infrastructure and income at this moment in time?**
   Is there a link between less economically advantaged areas and less developed infrastructure?

Background to Broadband Connection Infrastructure

Throughout this report we will refer to three different 'types' of broadband connection currently in operation across the UK, a brief summary of these is provided below for those unfamiliar with the topic

- FFBB (Full Fibre Broadband) - The fastest and most recently rolled out option
- UFBB (UltraFast Broadband) - older & slower
- SFBB (Superfast Broadband) - oldest and slowest connections

STEPS SPECIFICATIONS

Data Analysis Process

Defining the questions: The objective of the project was to find a correlation between internet infrastructure availability and standards of living across the UK. To this end, pertinent questions such as: "Does lack of access to internet infrastructure translate into a lower income bracket?"; "How exactly does availability (or lack of) define income earnings?"; "Is there a disparity in access in the rural and urban areas?"; "Urban areas potentially attract higher income brackets; is this in any way tied to internet accessibility?". From these, data was sourced individually by team members and based on the availability of data found, we were able to refine our objective to our three core questions.

Data gathering: Data was sourced via the internet and we found some sites (third party data collections) that gave access to simple csv/xlsx and API based data. The two main sources of our data were the Nomis and Ofcom websites which provided different datasets relating to internet access coverage and regional income information. We had to find a way to merge these different datasets to form a correlation and settled on using regions and district counties across the United Kingdom as both datasets gave detailed information on each region pertaining to its subject. From there, we were able to identify values that would form the foundation in relating these two topics.

Data cleaning: Once we had identified suitable datasets, we began data cleaning. In order to do this, we extracted the data which we identified as relevant and removed unreliable outliers. Next, we aligned the datasets by creating uniformity in string data, merged multiple sources to allow clearer analysis, removed NaN values and sorted the data into a relevant order.

Data analysis & Visualisation: Following our initial inspection, we conducted in-depth analysis involving a number of strategies, such as using linear regression analysis to identify trends over time periods and bivariate analysis to find the relationship between variables such as rural/urban classification and download speeds.

Data Sources

Internet Coverage Data: The data on internet coverage was taken from the annual Ofcom Connected Nations Report, which assesses the availability and coverage across areas of the UK. This dataset was found by investigating which research bodies provide annual data on the internet infrastructure available in different areas.

The Ofcom report offers data on different types of broadband. Q1 and Q3 used the 2021 data measuring availability of Full fibre broadband (FFBB), Superfast broadband (SFBB), and Ultrafast broadband (UFBB). Q2 used data from 2016 till 2021. As FFBB was still considered new generation technology in 2016, it was not measured individually in the dataset. Both SFBB and UFBB were considered, but after initial visualisations it became apparent that SFBB was already very widespread in 2016, and therefore UFBB was used to capture the rollout of new broadband over time. Ofcom provided the data for each Fixed Local Authority/Unitary Authority area. This format was chosen based on availability of other datasets.

Rural/Urban Classification: In order to answer the question on rural/urban differences we took our dataset from the 2011 Census Urban/Rural Classification data, available online at: https://www.gov.uk/government/statistics/2011-rural-urban-classification-lookup-tables-for-all-geographies

(Rural Urban Classification 2011 lookup tables for local authority areas)

This dataset covers local authorities across England and the metric we were interested in was the classification data that identifies areas as urban/rural according to six categories, listed from most urban to most rural these are:

- Urban with Major Conurbation
- Urban with Minor Conurbation
- Urban with City and Town Urban with Significant Rural
- Largely Rural
- Mainly Rural

Income API: The dataset for the income was obtained from the Nomis website. This website offered the options of querying available datasets to one's specification. Since we needed to find a correlation between income and internet infrastructure, and identify trends, it was necessary to get datasets that would show the median value of gross annual earnings of full-time workers across regions of the United Kingdom over the past five years, to coincide with the Ofcom's data. The median value was chosen over the mean for earnings data because it is less influenced by extreme earnings.

Implementation and Execution

Development Approach and Team Member Roles

For this project, it was essential that we recognised and highlighted our team's strengths and weaknesses before we began the project development process. This allowed for us to choose how to investigate each question whilst being able to showcase our strengths within the team, and gain deeper understanding in areas we were uncertain in. It was also vital to divide the project team into 3 subgroups so that a concerted effort would be made for each respective question.

The first step in the analysis process was followed to form the questions within the larger team. The next steps, gathering the data, cleaning the data and visualising the data, were completed by the smaller sub-teams for each separate question. This allowed us to keep our Jupyter Notebook in a clear and concise format, with each question separated for readability. The last step, in-depth analysis was completed by the larger team together, to see and validate if the objectives were being met, or if modifications to our processes were required. As a team we reflected on our separate interpretations of the data which facilitated in making meaningful insights and conclusions.

Tools and Libraries

For this project, tools and libraries that were touched upon in CFG classes were utilised. Some libraries were found from further exploration based on how we wanted our data to be presented. The below lists libraries, their used applications, and the questions/sections of the code in which they were employed.

Library	Method	Sections (divided by Questions)
Pandas	DataFrames, csv and JSON files	Question 1, 2, 3
Matplotlib	Data visualisation	Question 1, 2, 3
NumPy	Machine learning formulas	Question 2
Requests	API requests, using JSON	Question 2
Folium	Geospatial data visualisation	Question 1
SciKit-Learn	Supervised machine learning, predictive analysis models	Question 2

Implementation Process

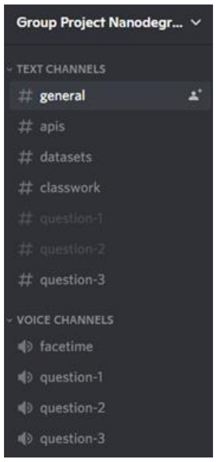


Figure 1 Project Discord

Agile development: The project was completed by following the agile development framework. We nominated a project lead and divided roles within the project team. Each sub-team assembled multiple times per week to complete the current sprint activities and resolve coding issues. The larger team would then meet to identify developments and discuss limitations found. This allowed for us to fully review and make changes to our outlook in a timely manner. The project was mostly controlled via Discord, due to its cross-functionality and its ability to make calls in separate channels simultaneously. Larger group calls were also made during the weekly sprint meetings, in the facetime channel. The project files were distributed through discord and GitHub, with code reviews made across both platforms.

Achievements: This project allowed for veritable accomplishments and insights into areas of data analysis we previously lacked in and considered our weaknesses. We discerned how readily available open-source data is and how detailed research is required in real-world analysis. The team were successful in understanding how to compare different datasets for correlations and make conclusions based on gathered reliable data. We effectively communicated as a team and managed the project with agile processes.

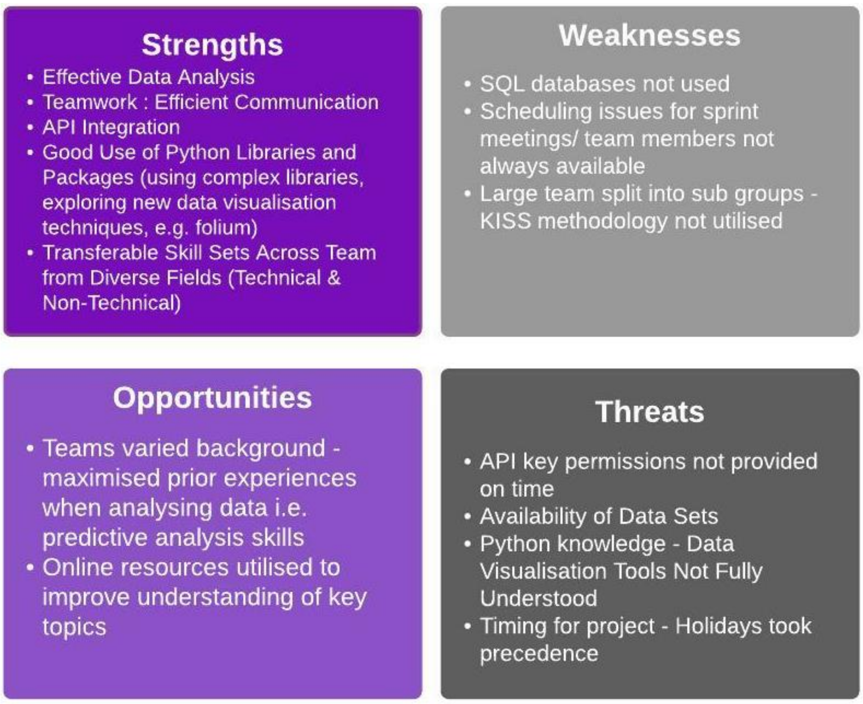
Challenges: As a team, one of the main challenges we faced was overcoming scheduling issues and making time for the project whilst balancing our other pursuits. To aid our team, we used the Discord to keep in contact, and speak to team members when needing help. Our decision to create sub-groups also aided our project and allowed for more flexible meeting schedules.

Decision to Change Something: Whilst completing the project, there were many instances where we found chosen resources were not available. For question 1, we initially planned to analyse data for Great Britain. The UK government data source seemed the most viable option, but it only contained data for England. This meant we had to modify our goal and focus our analysis on England districts that had a rural and urban classification. For question 2, we originally planned to analyse trends across Great Britain over a time period of 10 years. Initially we chose NOMIS data from 2010-2021, but Ofcom data was only openly available from 2016. Therefore, we chose to request only data encompassing 2016-2021.

Implementation Challenges: For question 1, we initially had issues visualising the map of England required for analysis. This was solved by using the folium library and highlighting the key areas for analysis. For question 2, we initially wanted to create a linear regression model represented by a graph increasing over time. Our model had a poor performance due to low correlations and visualisations were not possible. Due to this, we could not complete any predictive analysis.

SWOT Analysis

An analysis was completed during a retrospective meeting at the end of the project to help the team reflect on what we have learnt.



RESULT REPORTING

Question 1: What is the difference of infrastructure between Rural and Urban areas?

Availability of different broadband connections in Urban/Rural areas

In this section we checked broadband coverage data for the UK which shows speeds and type of broadband for each area, in conjunction with Rural/Urban classification data which indicates the designation for each area in England. By merging these two datasets, we were able to identify disparities between broadband availability across different types of location (rural or urban).

This graph depicts that SFBB (SuperFast Broadband), the oldest type of broadband, has achieved good availability overall. However, we can see that mainly rural areas have a gradual reduction in availability. UFBB (UltraFast Broadband) is broadly available among 'Urban with Major Conurbation', 'Urban with Minor Conurbation' and 'Urban with City and Town' areas (approximately 70-80% availability), yet there is a significant reduction in availability in 'Urban with Significant Rural', 'Largely Rural' and 'Mainly Rural' areas (approximately 25-50% availability). Full Fibre, the newest type of connection, has only attained around 25% availability in the most urbanised areas and gradually declines in more rural areas.

All connection types show a general trend of availability decreasing as regions become more rural, this is particularly evident in the UFBB availability line. We can see a major disparity of around 50% in the availability of UFBB between the most urban areas and most rural areas.

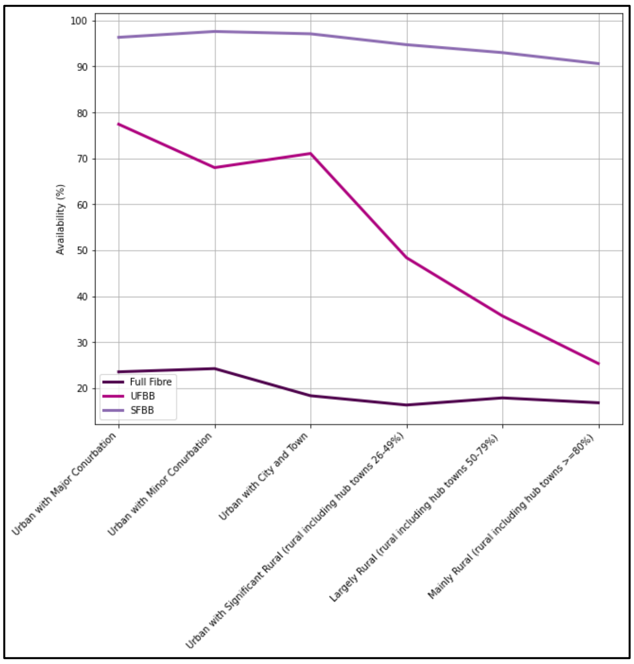


Figure 2 Line Graph - Data Availability vs Classifications

Areas with low-speed connections

In this section we consider how low download speeds can affect people's internet experience. One thing highlighted during the pandemic was the importance of high-quality connections particularly for people working from home and students who need to access online learning. Video conferencing is a particularly important tool for these users with various platforms recommending minimum download speeds around 8MBps. We identified local authority areas that have a significant percentage of premises (>=1%) that do not meet the minimum required download speed of 8MBps. The threshold for significance was defined after considering that overall, the percentages were relatively low.

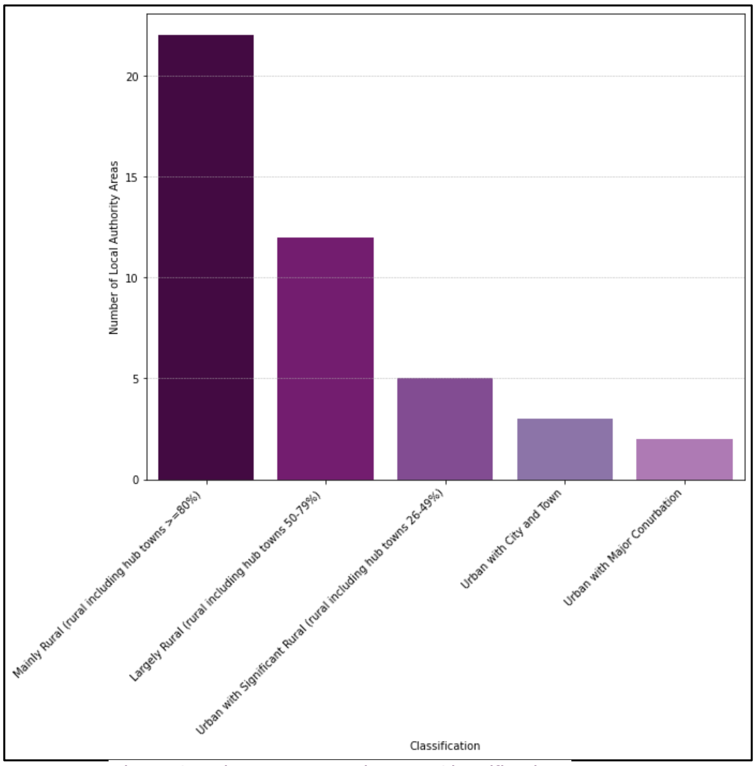


Figure 3 - Histogram - Regions vs Classification

This bar chart shows the number of Local Authority Areas that do not meet minimum speed requirements of 8MBps for each Rural/Urban classification. We see clear correlation to area type with most rural areas having the highest count of inadequate speeds.

'Mainly Rural' areas have more than 20 Local Authorities which do not meet the speed requirements. More urban areas have significantly smaller numbers of Local Authorities not meeting the requirements. This shows that the infrastructure in rural Local Authorities is less developed and users in these areas are at a disadvantage.

This plotted map shows the Local Authority Areas where the minimum requirement does not meet 8MBps in a significant number of premises. This can be used by telecommunications companies to highlight areas which would benefit from infrastructure investment schemes that would improve speeds for users who are currently affected.

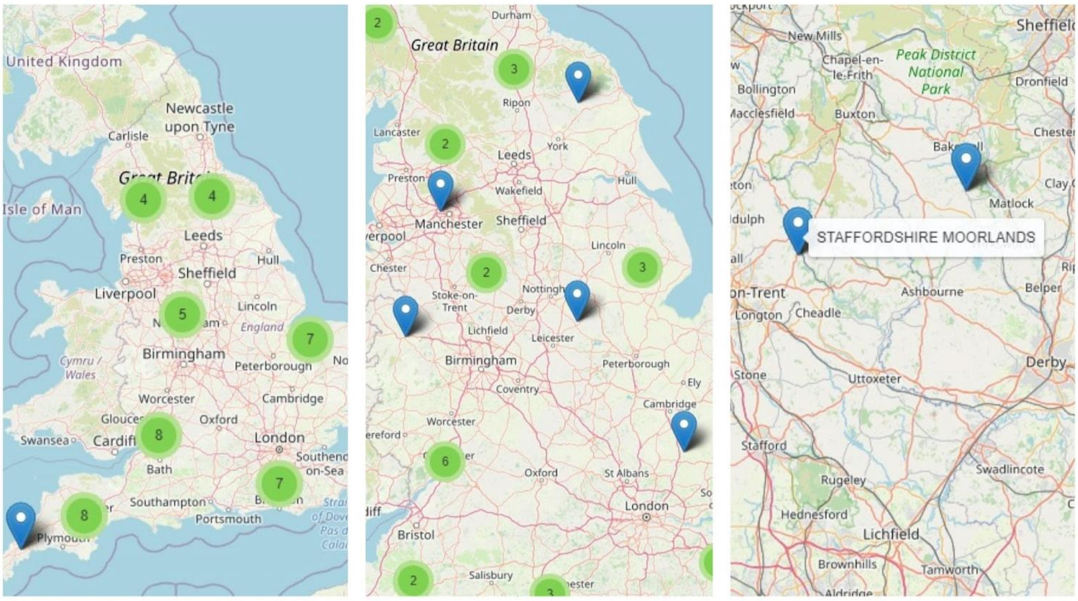


Figure 4 Visualisation Map for areas of low speed

Question 2: Can we identify trends over time?

The second aim of this report was to investigate trends over time between the availability of internet different internet speeds in an area and the income of an area. After incomplete data was removed, there were 302 areas in 11 regions across the UK.

Predicting internet availability from income over time

The relationship between Income and UFBB over time has been shown in Figure 5. Each datapoint represents an area and the year is indicated by colour.

This scatterplot shows that there is no discernible trend of income predicting UFBB availability. Most areas were at 0% of properties with UFBB availability in 2016. From 2017 onwards, there are many more areas with higher levels of availability, however in 2020 and 2021, there are still many areas which have very low levels of UFBB availability. There are less areas with a higher level of income, but these still have a similar patterning to the lower income areas. Correlation analyses and a linear regression confirmed the lack of significant relationship, however, due to brevity will not be reported here.

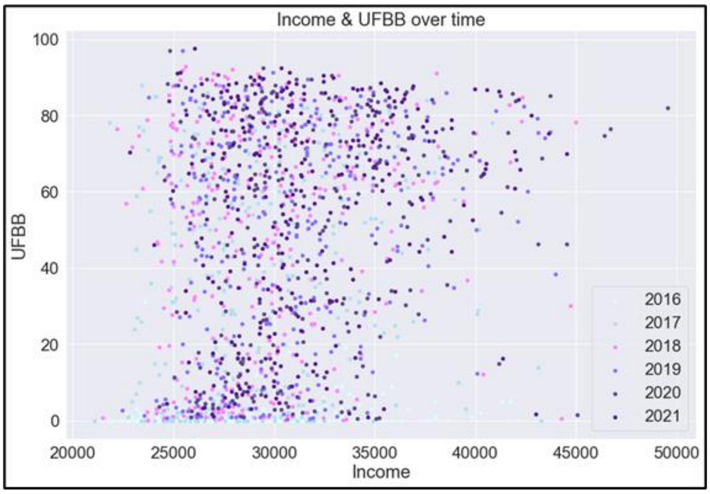


Figure 5 - Scatterplot showing the relationship between Income and UFBB, with the colours indicating the year

We believe that the reason for these findings is that broadband companies have continuously invested in the rollout of faster broadband options, no matter the income of an area. This makes sense when considering that one element of very high-income areas is that there will be a lot of single-household properties, while lower income areas are more likely to have denser living conditions, with a higher number of households in a smaller area. Therefore, it could still be beneficial for telecommunications companies to invest in lower income areas. Additionally, a relationship between income and internet availability would be based on the assumption that there is a link between higher income and more money available to pay for more expensive broadband packages. This does not take into consideration the willingness of consumers to spend their money on faster broadband, even if they may have a lower income.

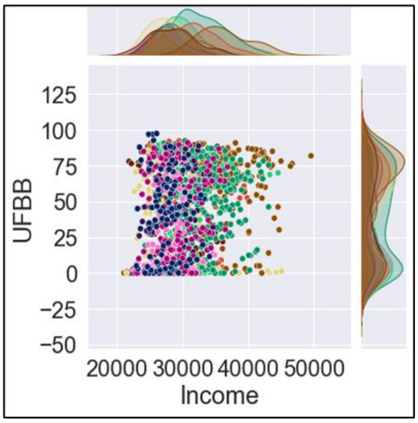


Figure 6 - Jointplot showing the relationship between Income and UFBB by area.

Investigating trends over time by region

As there was no significant relationship between UFBB and Income over time, we proceeded the analysis by looking at trends for both variables over time grouped by region. This can be seen in Figure 6.

As the Figure shows, there is no difference in the patterning of income predicting UFBB for different regions. There are some areas with higher levels of income, however, these do not necessarily have more properties with access to faster broadband, again confirming the findings above. Each region was also analysed individually, however, none of them showed a significant patterning.

Investigating the development of income and UFBB over time

As is evident from Figure 7, the graph shows that over time, income has increased for all regions. Most regions have remained in the same order of income as they were in 2016. London, Scotland and the South East have expressed the steepest increase in income from 2016 to 2021 with all having differences around £4000. On the other hand, the North East and Yorkshire showed the smallest increase, both being below £2000 increase. The data also clearly shows a decrease in income from 2020 to 2021 for most regions, which can be explained by the ongoing COVID-19 pandemic.

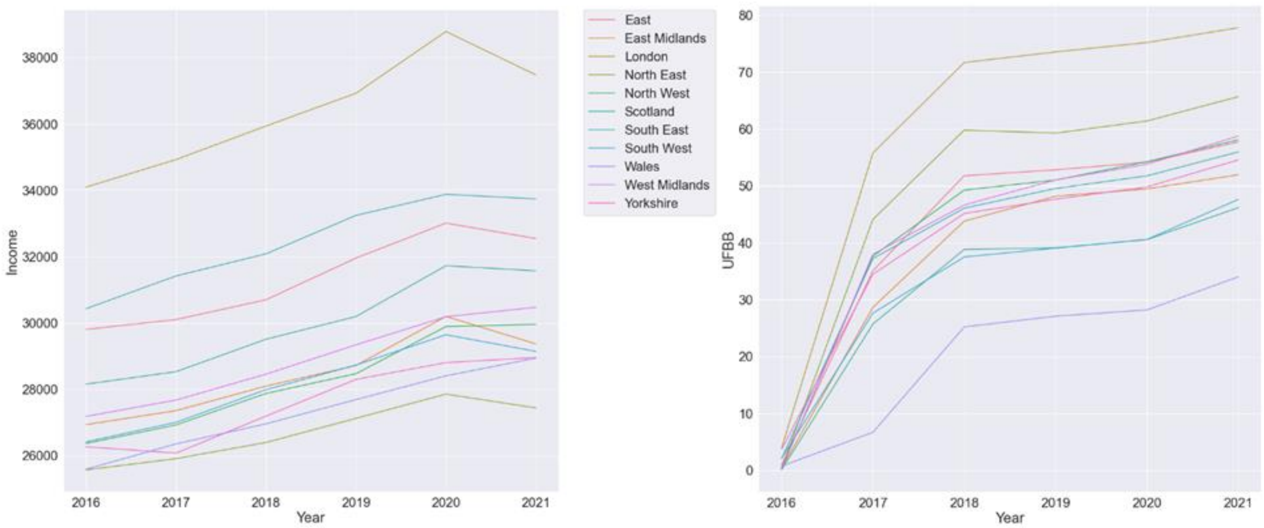


Figure 7 Line graph showing the trends of UFBB per region over time, Figure 8 Line graph showing the trends of income per region over time

Figure 7 shows the development of UFBB over time. In 2016, all regions had an average of 0-5% of properties with access to UFBB. Between 2016 and 2017, the steepest roll out of UFBB availability happened, except for in Wales, who remained below 10% in 2017, then had their steepest increase from 2017 till 2018, but have always remained below all other regions. In 2021, they are the only region which is still below 35% on average. The quick increase in rollout from 2016 continued till 2018, then the rollout slowed till 2020, with all regions increasing less than 5% over the time of two years. In 2020, a stronger increasing trend can be seen again, and all areas are still on an upwards tendency in 2021. London is the region with the highest availability of UFBB, they increased to nearly 60% by 2017, a value which many other regions are only just approaching in 2021. One factor that could explain this is that London is a city in which a large amount of people live in very dense conditions. This makes it both easier to place infrastructure which reaches many people and is also lucrative for telecommunications companies to provide more coverage.

Question 3: Is there a relationship between Income and Internet availability currently?

Here, we attempt to find the correlation, if any exists between internet infrastructure and income at this current time, so we would be working with 2021 data.

A glance at the scatterplots show mapping that appears to have a random movement which would denote that there is no relationship between internet availability and income. However, taking each plot individually and further examination of the regions on their own, show a slight relationship amongst these entities although not strong enough to authoritatively draw a conclusion on the existence of a relationship.

For the SFBB, East Midlands and Scotland show a moderate positive nonlinear relationship. In the East, South-west, Wales, London, and Yorkshire, the graph shows a random pattern denoting that there is no relationship in these areas whilst the remaining regions on this graph have an almost straight-line movement indicating that irrespective of the income in these regions, the internet availability remained the same.

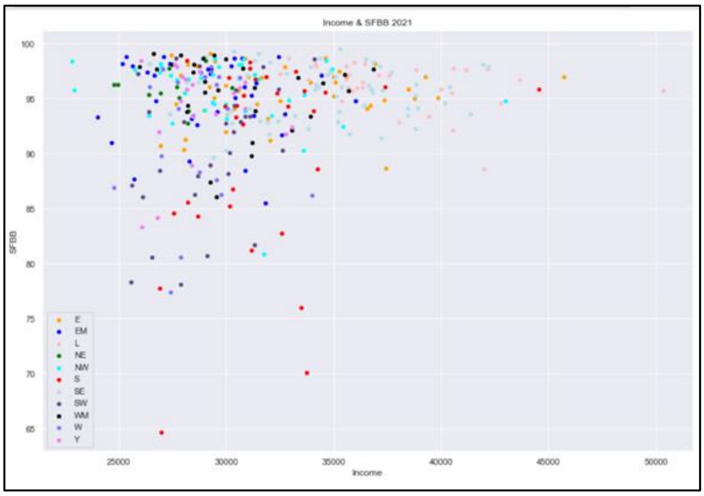


Figure 9 Scatter Graph - Income vs SFBB

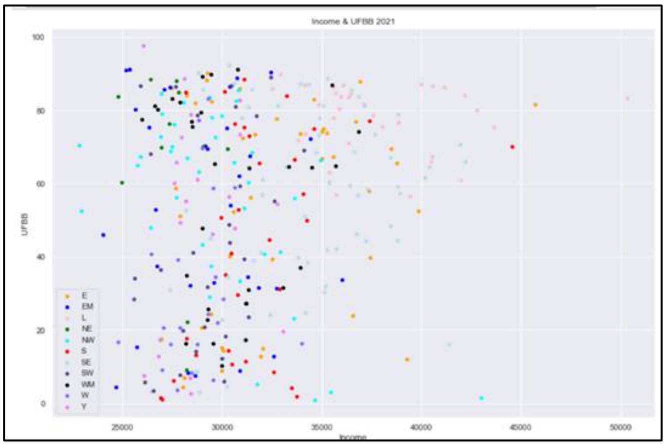


Figure 10 Scatter Graph - Income vs UFBB

For the UFBB, there was no relationship found as all regions described in a random movement excepting Wales which had a weak linear movement with a lot of outliers.

FFBB: Had the same pattern as the UFBB with most regions displaying random patterns. Regions like the East Midlands had moderate linear patterns whilst southwest and Wales displayed a very weak positive linear pattern.

Note: A positive linear pattern means that, as the income increases, so does the availability of internet infrastructure

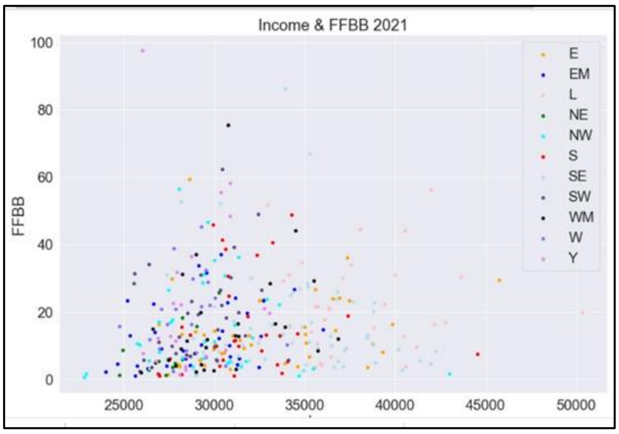


Figure 11 Scatter Graph - Income vs FFBB

Taking a quick look at the correlation matrix, all relationships between income and the different types of internet infrastructure had coefficients less than 0.2 depicting a weak relationship or entirely uncorrelated.

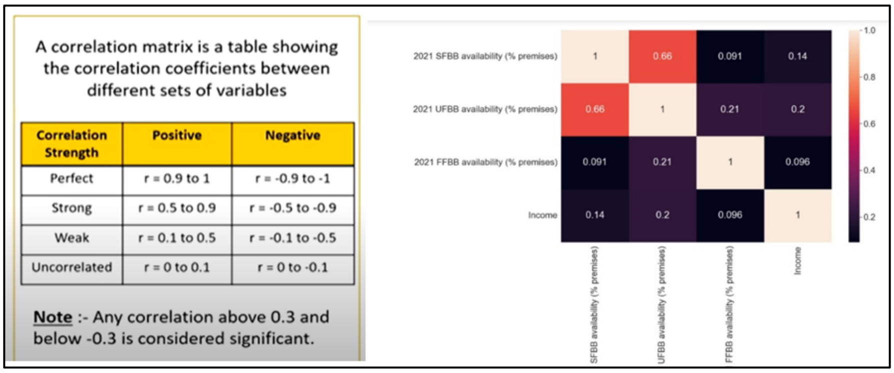


Figure 12 Correlation Matrix

CONCLUSION

This report set out to investigate the relationship between internet infrastructure and the properties of an area, particularly whether the area is classified as a rural or an urban area, as well as the average income of an area. The findings showed that the availability of internet infrastructure differed between the types of broadband.

- In 2021, most areas have a high coverage of SFBB, regardless of whether it is rural or urban. Additionally, no patterning with the income of an area can be identified. There were some trends visible for certain regions, however the correlation coefficient was nonsignificant.
- In 2021, very few areas have FFBB coverage, regardless of whether it is rural or urban. Again, income did not have an influence.
- In 2021, UFBB showed a clear difference in availability between rural and urban areas, with urban areas having a much higher percentage of premises who had UFBB available to them. Income was not a predictor of UFBB coverage.
- Over time, the UFBB coverage in an area could not be predicted by income. Both income and UFBB have been continuously trending upwards since 2016.
- A clear trend between rural and urban areas can be seen, when looking at internet speeds. A significant number of rural areas only have access to inadequate speeds.

The findings of this report showed that while income could not be used to identify any specific areas for further development, however, a clear need to increase the availability of faster broadband speeds in more rural areas became apparent. Particularly in times were working from home becomes more common, this is a topic that should be explored. We recommend that telecommunication companies should explore this further by weighing up the cost of expanding into more rural areas, and the amount of people that live there, keeping in mind that this report was done with a percentage of properties in an area and therefore giving no information about the total number of customers potentially willing to pay for faster speeds.