# Group ID - MSc in Data Analytics

Author: N. IOLCHIN

e-mail: [sba23303@student.cct.ie](mailto:sba23303@student.cct.ie)

Student ID: sba23303

**Abstract**

*During the last 12 years Internet Access in Ireland’s regions significantly increase. However, verry little research has been carried out to determine its past, present, and future development. The aim of the study was therefore to investigate internet access penetration rate. People with internet access were studied with different machine learning algorithms, EDA and statistical methods. Results show that Internet Access penetration rate is on the high level over the country, however some regions are still in the development phase. This concludes that future development will be focused on services development and not on infrastructure.*

**Introduction**

The internet is now an essential part of our daily lives.

Use of the internet makes our life easy, fast, and simple. The internet helps us with facts and figures, information, and knowledge for personal, social, and economic development.

There is a big part of research about how internet access impacts economic, social, educational, health, and many other aspects of our lives, and just a few of them showing how internet access was developing.

This paper aims to determine the internet penetration rate in each region of Ireland it’s historical and future development.

A review of the existing research on this topic reveals that the studies are outdated, dating back 8-10 years, but even these studies only include Ireland as a single data point in a list of countries studied.

Main objective of current work is to find the Internet Access penetration rate in each region of Ireland. The significance of estimating the internet penetration rate is that it will provide insights into the overall development of the industry and its importance to the population of each region.

Analysing historical data will allow to:

* compare the rapid growth of connections for each region.
* identify the likelihood of continued development in future years.
* identify the likelihood of peak development in future years.
* assess the probability of a decline in the number of connections.
* predict further development for each region.

With a complete picture of historical and predicted data, we can predict when each region will reach a peak of internet penetration.

**Methodology**

1. Programming:

The selection of programming concepts in the design of programmatic solutions for data analytics is a crucial decision that can significantly impact the efficiency, maintainability, and scalability of the resulting software. In this context, paradigm and language selection are particularly important considerations.

1a. Paradigm Selection

The current project uses several common paradigms:

* Imperative Programming: Imperative programming focuses on explicitly stating the steps required to transform data from input to output. It provides fine-grained control over the execution process, making it suitable for tasks that require precise manipulation of individual data elements.
* Declarative Programming: Declarative programming focuses on specifying the desired outcome rather than the specific steps to achieve it. It emphasizes abstraction and data relationships, making it well-suited for tasks that involve complex data transformations and aggregations.
* Functional Programming: Functional programming emphasizes immutability and data composition. It breaks down problems into smaller, reusable functions, promoting modularity and reducing side effects, which can be challenging to manage in large-scale data analytics projects.

1b. Language Selection

Current project is using Python programming language. Python is a versatile, general-purpose language with a large and active community. Its extensive data science libraries, such as NumPy, Pandas, and SciPy, make it a good and popular choice for data analysis and machine learning tasks.

1. Statistics:

Descriptive statistical analyses are used in the project. They summarize and describe the characteristics of a dataset, provide a high-level overview of the data, helping to identify patterns, trends, and relationships.

2a. Measurements in this project are:

* Mean: The arithmetic average of all values in the dataset.
* Range: The difference between the largest and smallest values in the dataset.
* Standard Deviation: The square root of the variance.
* Histogram: A graphical representation of the frequency distribution.
* Boxplot: A visual representation of the quartiles and outliers in the dataset.
* Data Exploration: Identifying patterns, trends, and relationships in data.
* Data Cleaning: Identifying and correcting outliers or erroneous data points.

2b. The following statistical distributions applied:

Normal distribution is used to fill missing data. Using a normal distribution to fill in missing data is not a perfect solution. It assumes that the data is indeed normally distributed, and it may not be suitable for all types of data.

The binomial distribution is a valuable tool for estimating the probability of successfully reaching a specific level of development within a specific time frame, however, is only applicable to situations where the outcomes are discrete, meaning they can be counted. For this project binomial distribution is used to calculate the probability of successfully reaching a specific level of development in a specific period.

Poisson distribution is used to model the number of events that occur in a fixed interval of time or space. In this project Poisson distribution is used to predict probability of:

* Steady increase in the following specific years for people with Internet Access
* Sudden spike for same period
* Decrease

1. Machine Learning

Following key steps of a machine learning strategy are used:

* Problem definition
* Data collection
* Data exploration
* Model selection.
* Model training.
* Model evaluation.
* Deployment.
* Monitoring.

Range of classification and regression techniques used:

* Logistic Regression: Logistic regression is a statistical model that predicts the probability of a binary outcome (e.g., spam or not spam, cancer or not cancer). It is a linear model that uses the logistic sigmoid function to transform the output of the linear model into a probability.
* Support Vector Machines (SVMs):SVMs are a type of supervised learning algorithm that can be used for both classification and regression tasks. SVMs are particularly well-suited for classification problems that have a small number of features and a large number of data points.
* Ridge Regression: Ridge regression is a type of linear regression that is used to prevent overfitting in high-dimensional datasets. Ridge regression adds a penalty term to the linear regression objective function that penalizes the magnitude of the model's coefficients.
* Random Forests: Random forests are an ensemble learning method that combines multiple decision trees to improve classification accuracy. Random forests are robust to noise and overfitting, making them a powerful classification algorithm.
* Linear Regression: Linear regression is a statistical model that predicts a continuous numerical value based on a linear combination of one or more features. Linear regression is a simple and effective regression algorithm that can be used to model linear relationships between features and the target variable.
* Lasso Regression: Lasso regression is another type of linear regression that is used to prevent overfitting in high-dimensional datasets. Lasso regression adds an L1 penalty term to the linear regression objective function that forces some of the model's coefficients to be zero.

1. Data preparation and Visualization

EDA was performed to:

* Get an overview of the data
* Explore categorical variables
* Explore numerical variables
* Identify relationships between variables
* Develop hypotheses based on your findings

All graphs are using Tufte's Principles of Graphical Excellence.

**Data Collection and preprocessing**

The target population for this study is the population of the regions of Ireland with access to the internet. The recommended official statistics website (<https://www.cso.ie/>) contains data on the population of Ireland, including the data about internet access in each of the geographical regions and Ireland as a whole.

To cover the historical aspect of the industry's development, we will need a sample of data from as many years as possible. This will help us to assess development more accurately in each region.

While exploring CSO website was found that there are several sections that contain information about Internet Access: “People and Society” and “Census”

Next deeper exploration discovers that there are following relevant data:

* Number of Estimated populations per region for years from 1995 to 2023 separated in 2 parts: 1995 – 2017, 2011 – 2023.
* Percent of households that have access to Internet per region for years from 2007 to 2022, separated in 2 parts: 2007 – 2016, 2020 – 2022.

As you can see, there is no data containing the number of people who have access to the internet. There is only a percentage of households that have internet access. To transform the percentage of households into the number of people, we need to know the *average number of people per household* for all available years and the number of households in each region.

Thus, applying simple math methods, we can get desired data. These are math steps that will get us to the goal:

*Nr of Households= Estimated Population nr/ Average Nr of Persons Household*

*Nr of Households with Internet Access = Nr of Households \* % Households with Internet Access / 100*

*Nr of People with Internet Access = Nr of Households with Internet Access \* Average Nr of People Household*

*% of People with Internet Access = Nr of People with Internet Access/ Estimated Population nr\*100*

From these calculations it’s clear that *Average Nr of People Household* is a crucial value to have.

Additional exploration of the CSO website provided 1 file with the average number of people per household per county and 3 files with the number of households with internet access. Both sets of data were only for 3 separate years: 2011, 2016, and 2022.

First brief assessment of collected files reveals a couple of issues:

1. There are data only for 3 years with 5 years gap between them (2011, 2016, 2022)
2. Data are classified by county, not by regions.

There are several solutions that can be applied to the first issue, including statistical, machine learning, or math techniques to fill in the available gaps.

For the second issue, some additional research was conducted, and fortunately, the CSO website provides a classification for counties, cities, and regions.

Result of data collection phase is the following list of next steps:

**Step 1.** Estimated population data will be assessed by analysing 2 files:

* Number of Estimated Population for per regions and country for years 1995 – 2017
* Number of Estimated Population per all regions and country for years 2011 – 2023

**Step 2.** Percent of households that have access to Internet will be assessed by analysing 2 files:

* Percent of households that have access to Internet per all regions and country for years from 2007 to 2016
* Percent of households that have access to Internet per all regions and country for years from 2020 to 2022

**Step 3.** Number of households with internet access and internet access type will be assessed by analysing 3 files with using of additional file with Ireland Administrative regions classification:

* Number of households with internet access and internet access type per counties and country for 2011
* Number of households with internet access and internet access type per counties and country for 2016
* Number of households with internet access and internet access type per counties and country for 2022

**Step 4.** Average number of people per household will be assessed with using of additional file with Ireland Administrative regions classification:

* Average number of people per household per counties and country for 2011, 2016, 2022

**Step 5.** Aggregation of all cleaned and processed data into a resulting dataset. This dataset will be the basis for solving the tasks.

All data were uploaded to github and moodle.

**Data Processing**

**Step 1. Estimated population data processing (**Jupiter notebook file PEA.jpynb).

The first data processing block is two files that contain population information. I’ve removed irrelevant data on age groups and gender and combined the raw data files into one clean data file. I’ve also made several minor changes to the files during analysis, all of which had no impact on the accuracy of further measurements.

* renamed the “Midland” region in “Midlands”, to match the official classification of regions.
* replaced the “State” region with “Ireland” to make the data more readable and clearer.
* converted the population units from thousands to the simple number format, since we will be interested in the number of people in the future.
* dropped the “STATISTIC Label” column and moved its value to the name of “VALUE” column for both files. This small change will be helpful later when data frames will be aggregated.

There is no missing data in either dataset, which is good, as there will be no need to additional data processing.

By visualizing both datasets we can see that there is a strong correlation between year and population number. Data for Dublin region for the Ireland was plotted in a separate graph for better visibility, as their range of values a much higher than for other regions. From boxplot created we can see them at the top. Both graph types agree with Tufte principle 1: the representation of numbers, as physically measured on the surface of the graphic itself, should be directly proportional to the numerical quantities measured (Tufte, 2007, p.93).

A chart of a number of people

Description automatically generated with medium confidence

A graph of different colored lines

Description automatically generatedA graph of a number of people with numbers

Description automatically generated

**FIG.1 Data from 1995 to 2017**

A graph of a number of people

Description automatically generated

A graph of different colored lines

Description automatically generatedA graph with a line and a line

Description automatically generated with medium confidence

**FIG.2 Data from 2011 to 2023**

I’ve then combined the datasets into one and visualized it again. This revealed a problem with some regions, which required further investigation.

A graph of a number of people

Description automatically generated with medium confidenceA graph of different types of population

Description automatically generated with medium confidence

It turned out that the data are uneven only for overlapping years from 2011 to 2017. Further study of the statistics website revealed that Louth was moved from the Border region to the Mid-East region, and South Tipperary was moved from the Mid-West region to the South-East region.

One way to handle this issue would be to use an average for years with skewed data. However, this would introduce changes to the data at the first step and affect its integrity. Therefore, following the Tufts principles of data processing, I’ve decided to discard the file with older data from 1995 to 2017 and use only the file with more recent data from 2011 to 2022.

One more rational for this decision is that data in other files fill the range with more recent years.

As a result of this stage, we have one clean data file (PEA04\_c.csv).

**Step 2. Percent of households that have access to Internet data processing. (**Jupiter Notebook ICA.jpynb)

Next, I analysed files with percentage of households with Internet access.

Observations:

* there are data missing for 2017, 2018 and 2019.
* one of the files has excessive information.

A few minor changes were made. Such as changing the column names to lowercase and combined both data frames. The reason behind is to have same names of the columns in both files to easier future concatenation.

Check the visualisation of resulting data frame for each region and Ireland below.

A graph of a number of individuals

Description automatically generated with medium confidence

A group of graphs with numbers

Description automatically generatedFrom the graph, there is a clear gap in the data from 2017 to 2019. To address the missing data issue, Normal Distribution and Interpolation were used.

1. **Normal distribution** can be applied in case we assume that data are normally distributed.( ICA.ipynb)

The result of the first method is shown in the following graph:

A graph of different colored lines

Description automatically generated

The graph clearly shows that the missing data from 2017 to 2019 is erratic.

We can conclude that our data is not normally distributed.

1. **Interpolation method** can be an alternative as we clearly have linear correlation between year and % of households.

The next method is to use interpolation. You can find all the details of the file processing in Jupiter Notebook ICA.jpynb.

List of plots for each region after interpolation was applied. As you can see, interpolation gives excellent results.

A graph of a number of households

Description automatically generated

**Step 3. Number of households with internet access and internet access type data processing. Additional file with Ireland Administrative regions classification review and processing**. (SAP 2011 - 2022.ipynb)

Continuing to process the existing data, let's move on to the third stage and work with a group of 3 separate files with *Number of households with internet access and internet access type*.

The main issue with all the files is that the data is collected by county/city and not by region, as in the previous files. To address this issue, let's first look at the official classification of regions.

A map of ireland with different colored areas

Description automatically generated

For this data frame, I’ve used the method of filling empty rows with the previous value. This was done to make it easier to work with the file in the future and to ensure that each county/city has a according value in the region column.

A review of the Number of households with internet access and internet access type files shows that they don’t have any missing, which is good. The rest of the data processing involves creating a function to generate the appropriate region value for each city and region. There are some minor steps required:

1. Removing all excessive words as: “City”, “County”, “City and County”, etc.
2. Removing excessive spaces.

As soon as these steps are repetitive, a separate function was created.

Next, I’ve aggregated the data and combine all three datasets into one.

A graph of a number of blue squares

Description automatically generatedThe resulting data frame includes an additional column with internet connection type. This information would be useful for a wider picture of the development of internet access.

The graphs use clear and detailed labelling to defeat distortion and ambiguity, which is in line with Tufte rule 2 (Tufte, 2007, p.163).

The graph shows us that the available sample only covers three years. Given the nature of the data, the only way to fill in the missing data is to generate synthetic data. This would have a significant impact on the validity of the data and undermine the entire purpose of the work, so we must discard this data and move on to the next step.

**Step 4. Average number of people per household will be assessed with using of additional file with Ireland Administrative regions classification. (**F2002.ipynb)

The next short stage in data analysis and processing involves reviewing a file containing the average number of people per household. This additional data is necessary to calculate the number of people with internet access, which is our primary goal.

First, aggregating counties/cities in according to region, to have clear picture per region. To do this, approach from previous stage was applied. I.e., creating a function to generate the appropriate region value for each city and region. Same code was applied for the minor changes.

Now, let’s analyse our visualized result:

A graph of people in different colors

Description automatically generated

Taking in consideration all data analysis from the previous stages, the relevant time interval for us will be 2011 – 2022.

Like the file from the previous stage, this file will contain data for three separate years from our relevant time interval: 2011, 2016, and 2022. However, the nature of the data is slightly different. The standard deviation for the values is 0.08, which is about 3%. In this case, we can ignore this small difference and use interpolation to fill in the missing data.

By applying the same methods for interpolation of missing data as we used for the previous data frame, we obtain the following result.

A graph of the number of people in the united states

Description automatically generated

A graph of growth in the year

Description automatically generated with medium confidence

As you can see from the graphs, the plot is not smooth for every region. However, as we mentioned before, a small standard deviation will not have an impact on main data.

**Step 5. Aggregation of all cleaned and processed data into a resulting dataset. This dataset will be the basis for solving the tasks.** (Data Processing.ipynb.)

The final stage of data processing is to create the resulting data frame, that will be used to solve tasks of current project.

This data frame will aggregate data on estimated population, average number of persons per household and households with internet access. Then, using simple mathematical calculations, I will obtain the necessary columns with the number and percentage of the population with internet access.

Here are formulas to apply:

*Nr of Households= Estimated Population nr/ Average Nr of Persons Household*

*Nr of Households with Internet Access = Nr of Households \* % Households with Internet Access / 100*

*Nr of People with Internet Access = Nr of Households with Internet Access \* Average Nr of People Household*

*% of People with Internet Access = Nr of People with Internet Access/ Estimated Population nr\*100*

To make the data easier to read, we round the numbers to two decimal places. Our resulting file is now ready to be used to solve the tasks.

**Results**

Gaining insights with exploratory data analysis.

1. **Current penetration rate.** (Data overview.ipynb)

First and the most important is to assess current penetration rate. Here are the data for 2022 year.

A graph of people with internet access

Description automatically generated

For the country we have 94% of people already having internet which is strongly correlating with Western Europe data (93.7% according to Global internet penetration rate by region from https://www.statista.com/)

A screenshot of a graph

Description automatically generated

This means that Ireland has one of the highest percentages of Internet Access penetration rate.

Less % of population with Internet access are in Border (90%) and West (91%).

Dublin as a capital has the highest number (97%) of populations connected to the internet.

1. A graph of population growth

   Description automatically generated with medium confidence**Historical trends** (Jupiter notebookData overview.ipynb)

A graph of different numbers

Description automatically generated with medium confidence

A group of graphs showing the number of people with internet access

Description automatically generated

The first clear correlation between population number and the number of people with Internet access is a decreasing trend in 2014. This trend was evident across all regions and could be a factor behind the subsequent upsurge in Internet users in 2015.

Starting from 2016 we have stable increase up until 2021 for all the regions.

A significant increase in Internet access was observed in 2022 across nearly all regions, aligning with the trend of population growth.

Next graph showing the growth of internet penetration for each region in period from 2011 to 2022.

A graph of blue bars

Description automatically generated

We can observe that Border and South-East are the fastest growing regions.

1. **Future trends and probabilities with Machine learning and statistic methods and techniques.**

**3a. Machine learning part 1 and part 2** predicting Estimated number of population and Number of People with Internet Access in the next 5 years.

Following graph display how number of people with internet access trending during the time. It includes past (2011 – 2022) and future (2023 – 2027) data for better trend understanding. (ML Nr of population with IA.ipynb part 2)

A graph of different colored lines

Description automatically generated

Stable increasing trend is traceable for each region.

There is a visible drop for West region. It is related with lowest prediction score from machine learning algorithm for this specific region. Next stable linear growth prediction is possible because of hyperparameters tuning (using K-fold cross-validation and grid search technique). Estimated population and Year were the 2 independent variables for Number of people with internet access prediction. Meaning: Estimated population number was predicted beforehand.( ML Estimated population.ipynb – part 1)

While predicting Estimated population number, scores of predictions for regions wasn’t too high. Following graph shows the details.

A graph of blue rectangular bars

Description automatically generated with medium confidence

Using K-fold cross-validation and grid search technique I was able to increase rate of prediction for Border region from 0.2 to 0.44, and for Mid-West from 0.34 to 0.38.

I would like to note that such a low prediction rate might create a bias on further prediction of people with internet access.

**3b. Poisson distribution (stats\_poisson.ipynb)**

Continuing future predictions let’s have a look on the probably to have steady increase in the next 10 years taking as a base last 12 years development. For this Poisson distribution will be taken as a prediction method.

Following graph displays founds:

A graph of increasing numbers

Description automatically generated with medium confidence

As following from the graph, there is a very low probability of steady increase. Which is logical as current % od population connected to the internet is already high (94%)

Next 2 graphs will be presented together:

A graph of the number of states

Description automatically generated with medium confidence

The graphs demonstrate an evident inverse relationship, having very close percentage. Almost the same probability of decrease and sudden spike for future internet access trend means that this value is highly volatile and unpredictable. This can be caused by a number of factors, such as high levels of uncertainty or manipulation. A value with almost the same probability of decrease and sudden spike is often said to be "at risk" or "unstable." This means that there is a high probability that the value will experience a significant change, either upwards or downwards.

**3c. Binomial distribution. (Stats\_binom.ipynb)**

Based on the assumption that the predicted data for the number of people with internet access is accurate. let’s predict the peak of development for internet connectivity for each of the region. As a peak I will take 99% of population having internet access. Prediction will be based on the available historical data (2011 – 2022) and from the machine learning algorithm predicted data (2023 – 2027).

For this prediction I will use binomial distribution.

The binomial distribution is a discrete probability distribution that describes the number of successes in a sequence of independent experiments, each with two possible outcomes: "success" or "failure". Our expected result for success is rather region will achieve 99% of population to have internet access or not. The outcome is shown on the following graph.

A graph with blue dots

Description automatically generated

According to available trends we will have Border region and Dublin with high probability having 99% of people connected to the internet already this year.

In 2025 Not only Midlands, South-East and South-West will have a peak of internet connection, but the whole country as well.

Mid-west will get this point on 2026.

We don’t have West and Mid-East on the graph. This mean that these regions will be longer in the development state.

**3d. Part 3 of Machine Learning (ML Classification.ipynb)**

In this part I will opposite the classification Machine Learning Algorithm to Binomial Distribution to predict the year when Ireland will get 99% of people connected to the Internet.

Even though there were 2 models with same high score of prediction, results were different.

Logistic classifier gave us that there is no possibility for Ireland to get 99% of population connected to the Internet by 2030.

SVM model gave us that 2027 is the year when Ireland will get 99% of population connected to the Internet.

**Discussions/Conclusions**

Based on the findings, it can be concluded that nowadays republic of Ireland has a developed infrastructure of internet access as it’s penetration rate already 94%.

What does it mean for Irish population?

It means that they can expect for internet service providers to be more focussed on developing variety range of services to be more attractive to customers.

According to predicted trends regions West and Mid-East will be in the development stage for the next 5 years. Rest of the regions will be slowly increasing their % of population with internet access.

Prediction shows that in the following 5years there will be slow penetration internet rate increasing.

**References**

Environment International 130 (2019), *A comparison of linear regression, regularization, and machine learning algorithms to develop Europe-wide spatial models of fine particles and nitrogen dioxide*

Edward R. Tufte *The Visual Display of Quantitative Information* Second edition

Neil A. Weiss, 2017 *Introductory Statistics,* 10th edition

O'Donnell, Susan (2003) *"News Consumption in Ireland and the European Union: Traditional Media vs the Internet,"* Irish Communication Review: Vol. 9: Iss. 1, Article 4.,

Lisa Correa, Pietro Crocioni, 2012 *Can evidence of pricing power help market power assessment? Broadband Internet in Ireland and the Netherlands*

Aurélien Géron, O'Reilly Media Tukiainen, Janne (2004-02-06), *Access to Computer, Internet and Mobile Phone at Home in Finland, Ireland, Netherlands and Sweden*,

Mauro F. Guillén, Sandra L. Suárez 2001, *Developing the Internet: entrepreneurship and public policy in Ireland, Singapore, Argentina, and Spain*