Company Registration HMRC Report

miningAH

2025-01-17

Table of Contents

[Task Brief 3](#_Toc188140290)

[Chapter 1 - Introduction 4](#_Toc188140291)

[1.1. Data Source 4](#_Toc188140292)

[1.2. Technologies 4](#_Toc188140293)

[1.3. Methodology and report structure 4](#_Toc188140294)

[1.4. Assumptions 4](#_Toc188140295)

[Chapter 2 - Business / Contextual Understanding 5](#_Toc188140296)

[Company Registration 5](#_Toc188140297)

[UK postal codes 5](#_Toc188140298)

[Chapter 3 - Data Understanding 6](#_Toc188140299)

[3.1. Viewing the first and last five records of the dataset 6](#_Toc188140300)

[3.2. Viewing a glimpse of the dataset 7](#_Toc188140301)

[3.3. Manually exploring the dataset 8](#_Toc188140302)

[Chapter 4 - Data Preparation 9](#_Toc188140303)

[4.1. Feature Engineering - new attributes 9](#_Toc188140304)

[4.2. Fixing incorrect Data Types 9](#_Toc188140305)

[4.3. Checking for total duplicates 9](#_Toc188140306)

[4.4. Checking businesses that have multiple records 9](#_Toc188140307)

[4.5. Checking for missing data for each column 9](#_Toc188140308)

[4.6. Checking for incorrect record values 9](#_Toc188140309)

[4.6.1. Checking date values 10](#_Toc188140310)

[4.6.2. Checking Company Name values 10](#_Toc188140311)

[4.6.3. Checking postcode values 10](#_Toc188140312)

[Chapter 5 - Data Analysis 11](#_Toc188140313)

[5.1. Exploratory Data Analysis - Statistical summary of the dataset 11](#_Toc188140314)

[5.2. Exploratory Data Analysis - Visual investigation 11](#_Toc188140315)

[5.2.1. Viewing the statuses of organisations in the dataset (open, closed, in liquidation) 11](#_Toc188140316)

[5.3. Exploring task brief questions 13](#_Toc188140317)

[5.3.1. Question 1 - Which sectors show the greatest increase and decrease in turnover over the years available? 13](#_Toc188140318)

[5.3.2. Question 2 - What is the total turnover for “Farming” broken down by year? 15](#_Toc188140319)

[5.3.3. Question 3 - What is the LEAST common month to incorporate a business? Is this true for all Company Types? 16](#_Toc188140320)

[5.3.4. Question 4 - What other conclusions can be drawn from the data. We are particularly interested in trends over time, and comparisons between different geographic areas. 18](#_Toc188140321)

[Chapter 6 - Further Analysis 19](#_Toc188140322)

[6.1. Analysing businesses without a known location. 19](#_Toc188140323)

[6.2. Checking that all organisations in the dataset have updated their accounts in the last year. 19](#_Toc188140324)

[Chapter 7 - Conclusion 20](#_Toc188140325)

[References 21](#_Toc188140326)

# Task Brief

You are required to present an analysis of Company Registration data, along with turnover and business type.

The data is based upon Open-Source information. It has been anonymised and manipulated for the purposes of this exercise. Fictional turnover and Trade Type data has been added. The dataset is not representative of any real-life Company and no attempt should be made to identify real businesses through it.

All the necessary data is contained in the Excel Workbook, and you can use any software you wish to analyse it. However, it is expected that most people will use Excel, or an equivalent program like Google Sheets, or Open Office Calc.

We would like the presentation to cover your answers to the questions below, and the methods used to reach your conclusions.

1. Which sectors show the greatest increase and decrease in turnover over the years available?
2. What is the total turnover for “Farming” broken down by year?
3. What is the LEAST common month to incorporate a business? Is this true for all Company Types?
4. What other conclusions can be drawn from the data. We are particularly interested in trends over time, and comparisons between different geographic areas.

# Chapter 1 - Introduction

## 1.1. Data Source

See ‘The Task’ section of this report for details on the data source.

## 1.2. Technologies

Excel has been used to conduct feature engineering to gain useful geographical features from the postal code. R has been for all remaining data analysis activities including data pre-processing, visualizations, and analysis. The R-Studio IDE has been used as the environment for analysis efforts.

## 1.3. Methodology and report structure

The analysis methodology used for this task is based on the first four of the six stages of the CRISP-DM framework followed by a delivery / sharing stage. The five stages, which also give this report its structure, are as follows:

1. Business understanding - Exploring all unknown contextual concepts and issues.
2. Data understanding - Exploring the data for an initial understanding.
3. Data preparation - Cleansing the data from issues such as duplicates, incorrect values and so on.
4. Data Analysis - Analysing the data for insights and the questions mentioned in the task brief.
5. Insights sharing - Creating a presentation to share data insights discovered.

## 1.4. Assumptions

Due to the nature of this task, clarifying the dataset origins and testing its adequacy is not possible. Therefore, to be able to draw conclusions and insights from the data it is assumed that either:

* the dataset constitutes an entire population of records or a representative sample of one, or
* that the conclusions being incorrect to the real world is not an issue, as the dataset is to be dealt with in isolation.

# Chapter 2 - Business / Contextual Understanding

The following chapter describes the contextual research conducted to support the analysis phase of the task.

## Company Registration

A registered company is one that has been incorporated at Companies House (the UK’s national business registry).

An organisation would undergo this registration to become a limited company (Townley 2018).

“A limited company is a type of business structure that is incorporated (registered) at Companies House. It is a distinct legal entity that is separate from its owners (members), which means that it’s responsible for its own finances and debts. Companies are managed by one or more directors. They are appointed by members to run the business on their behalf, but it is common for directors to also be members. The owners of a limited company benefit from reduced financial responsibility for business debts. This is known as limited liability.” (Carpenter (2024)).

Company Registration data will therefore contain information about companies that have registered at Companies House. These are limited liability companies and therefore excludes sole traderships.

Note that “Once incorporation at Companies House has been accepted the Corporation Tax information will be sent to HMRC and will not be retained by Companies House.” (“Service Information”).

Finally, a lot of data is provided to Companies House on a yearly basis, including financial statements (“Companies House Accounts Guidance” (2024)).

## UK postal codes

UK postal codes were developed by the UK’s postal service, Royal Mail, to be able to deliver posts. The Royal Mail maintains a UK-wide system of postcodes.

Postcodes are alphanumeric references that help to identify geographical areas. They comprise of two parts, an outward code of between 2 to 4 characters and an inward code of three characters. The outward code references the postcode area and district of a particular geographical location. The inward code provides finer location identification (“Postal Geographies” (2023)).

It is important to note that UK postal codes always start with a letter and that a full postcode will always be between 5 and 7 characters long (Kurdi 2023).

# Chapter 3 - Data Understanding

The following chapter discusses an initial exploration of the dataset provided for the task at hand. The goal behind this stage was to gain a general understanding of the data by manually exploring the data, and to spot any quality issues which would need to be addressed in the subsequent stage of data preparation.

## 3.1. Viewing the first and last five records of the dataset

First 5 Rows of HMRC Data

|  | Company Name | Company Type | Postcode | CompanyCategory | Company Reference | CountryOfOrigin | Company Closed Date | IncorporationDate | Accounts Last Made Date | 2020 Turnover | 2021 Turnover | 2022 Turnover | 2023 Turnover | 2024 Turnover | Region | Country | Latitude | Longitude |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Record 1 | Business-1000289937 | Retail | NG7 2RR | Private Limited Company | Active | United Kingdom |  | 2024-05-29 00:00:00 |  |  |  |  |  |  | East Midlands | England | 52.943642 | -1.174498 |
| Record 2 | Business-1000550345 | Other | KT15 1NJ | Private Limited Company | Active | United Kingdom |  | 2024-01-30 00:00:00 |  |  |  |  |  |  | South East | England | 51.366909 | -0.498953 |
| Record 3 | Business-1000639152 | Other | CH6 6HB | Private Limited Company | Active | United Kingdom |  | 1984-07-02 00:00:00 | 2023-12-31 00:00:00 | 482,884 | 333,183 | 252,728 | 304,433 | 368,388 | Wales | Wales | 53.266786 | -3.177996 |
| Record 4 | Business-1000825104 | Other | RH16 3US | Private Limited Company | Active | United Kingdom |  | 2023-10-03 00:00:00 |  |  |  |  |  |  | South East | England | 51.003064 | -0.085272 |
| Record 5 | Business-1001001047 | Other | NN17 2YE | Private Limited Company | Active | United Kingdom |  | 2008-06-09 00:00:00 | 2023-09-30 00:00:00 | 537,182 | 306,586 | 280,267 | 403,063 | 501,979 | East Midlands | England | 52.499828 | -0.705569 |

The first five records of the HMRC dataset indicates fairly regular and clean data however with many missing values. Some attributes represent categorical features and would therefore need to be assessed for any category related errors. For instance, when the same word, one with a capitalised starting letter and the lower cased, constitutes two separate categories when they should be one. The issue of missing data needs to be investigated and the values of records, for each attribute, checked for correctness.

Last 5 Rows of HMRC Data

|  | Company Name | Company Type | Postcode | CompanyCategory | Company Reference | CountryOfOrigin | Company Closed Date | IncorporationDate | Accounts Last Made Date | 2020 Turnover | 2021 Turnover | 2022 Turnover | 2023 Turnover | 2024 Turnover | Region | Country | Latitude | Longitude |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Record 1 | Business-1000289937 | Retail | NG7 2RR | Private Limited Company | Active | United Kingdom |  | 2024-05-29 00:00:00 |  |  |  |  |  |  | East Midlands | England | 52.943642 | -1.174498 |
| Record 2 | Business-1000550345 | Other | KT15 1NJ | Private Limited Company | Active | United Kingdom |  | 2024-01-30 00:00:00 |  |  |  |  |  |  | South East | England | 51.366909 | -0.498953 |
| Record 3 | Business-1000639152 | Other | CH6 6HB | Private Limited Company | Active | United Kingdom |  | 1984-07-02 00:00:00 | 2023-12-31 00:00:00 | 482,884 | 333,183 | 252,728 | 304,433 | 368,388 | Wales | Wales | 53.266786 | -3.177996 |
| Record 4 | Business-1000825104 | Other | RH16 3US | Private Limited Company | Active | United Kingdom |  | 2023-10-03 00:00:00 |  |  |  |  |  |  | South East | England | 51.003064 | -0.085272 |
| Record 5 | Business-1001001047 | Other | NN17 2YE | Private Limited Company | Active | United Kingdom |  | 2008-06-09 00:00:00 | 2023-09-30 00:00:00 | 537,182 | 306,586 | 280,267 | 403,063 | 501,979 | East Midlands | England | 52.499828 | -0.705569 |

The last five records of the HMRC dataset confirm that there are records in the dataset that are not appropriate for analysis. This is because these records represent summary records.

## 3.2. Viewing a glimpse of the dataset

## Rows: 50,039  
## Columns: 18  
## $ `Company Name` <chr> "Business-1000289937", "Business-1000550345"…  
## $ `Company Type` <chr> "Retail", "Other", "Other", "Other", "Other"…  
## $ Postcode <chr> "NG7 2RR", "KT15 1NJ", "CH6 6HB", "RH16 3US"…  
## $ CompanyCategory <chr> "Private Limited Company", "Private Limited …  
## $ `Company Reference` <chr> "Active", "Active", "Active", "Active", "Act…  
## $ CountryOfOrigin <chr> "United Kingdom", "United Kingdom", "United …  
## $ `Company Closed Date` <dttm> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,…  
## $ IncorporationDate <dttm> 2024-05-29, 2024-01-30, 1984-07-02, 2023-10…  
## $ `Accounts Last Made Date` <dttm> NA, NA, 2023-12-31, NA, 2023-09-30, 2023-08…  
## $ `2020 Turnover` <dbl> NA, NA, 482884, NA, 537182, 593638, 429487, …  
## $ `2021 Turnover` <dbl> NA, NA, 333183, NA, 306586, 265103, 260400, …  
## $ `2022 Turnover` <dbl> NA, NA, 252728, NA, 280267, 315057, 307066, …  
## $ `2023 Turnover` <dbl> NA, NA, 304433, NA, 403063, 325861, 515549, …  
## $ `2024 Turnover` <dbl> NA, NA, 368388, NA, 501979, 299272, 472806, …  
## $ Region <chr> "East Midlands", "South East", "Wales", "Sou…  
## $ Country <chr> "England", "England", "Wales", "England", "E…  
## $ Latitude <chr> "52.943642", "51.366909", "53.266786", "51.0…  
## $ Longitude <chr> "-1.174498", "-0.498953", "-3.177996", "-0.0…

Viewing a quick glimpse of the dataset reveals that the data has 18 attributes and a little over 50 thousand records. It is important for analysis reasons to note that the turnover for the years 2020 to 2024 is provided within columns. This should mean that in general only a single record should exist for each company name.

## 3.3. Manually exploring the dataset

Manually exploring the entire dataset through changing column sorting and filtering revealed the following:

* Incorrect postcode values exist, for instance they might contain only numbers, be too short, or not start with a letter.
* Outliers and extreme values are not an issue within the dataset.

# Chapter 4 - Data Preparation

The following chapter outlines the data preparation activities undertaken to ensure that the data is of sufficient enough quality to move ahead with analysis efforts.

## 4.1. Feature Engineering - new attributes

To begin with, to undertake geographic analysis, the postcode attribute from the original dataset was used to generate the following four features: Country, Region, Longitude, and Latitude. The changes were made in Excel and involved calling the postcodes.io API (“Postcode & Geolocation API for the UK”) with each postcode and then extracting the relevant attribute values from the JSON formatted response.

## 4.2. Fixing incorrect Data Types

The data types of several attributes in the dataset were initially incorrect. To fix this issue the following changes were made:

|  |  |  |
| --- | --- | --- |
| Attribute | Old data type | New data type |
| Incorporation Date | POSIXct | Date |
| Company Closed Date | POSIXct | Date |
| Accounts Last Made Date | POSIXct | Date |
| Longitude | Char | Double |
| Latitude | Char | Double |
| Region | Char | Factor |
| Country | Char | Factor |
| CountryOfOrigin | Char | Factor |
| CompanyCategory | Char | Factor |
| Company Type | Char | Factor |
| Postcode | Char | Factor |

## 4.3. Checking for total duplicates

To ensure statistical figures would be accurate the existence of duplicates was investigated. A total of 35 rows were found to be total duplicates (same values across all attributes). This issue was resolved by filtering the dataset to ensure that only distinct values remained.

## 4.4. Checking businesses that have multiple records

Like duplicate checking, the issue of businesses having multiple records associated with themselves was explored. This is because the existence of multiple records could be problematic if they contain contradicting information, or one is new whilst the other outdated. The results showed that only one single business had two records and that these records were not problematic and so were left in the dataset.

## 4.5. Checking for missing data for each column

Many columns contain missing values. for some of these attributes the presence of missing values is expected, for instance 49,132 records have the company closed date as empty because they are still active and not closed. However, for many others no logical reason can be determined as to why they would be missing. In such a case of large amounts of missing values, imputing values or removing records is not useful, hence the dataset will be used with these limitations in mind.

## 4.6. Checking for incorrect record values

Checking attributes for appropriate values revealed the following points:

* Attributes with fewer distinct values such as Company Type, Company Category, Company Reference, and Country of origin are free from value related issues.
* Region, Country, longitude and latitude, due to being computed values engineered in excel, also do not have data value issues.
* The Turnover columns (2020 to 2024) also show to be free from value errors, determined through manually sorting.
* The remaining 5 attributes require deeper checks to ensure values are correct. Due to having too many distinct values a formatting-based check will be required to ensure data values are as expected. Sections 4.6.1 to 4.6.3 will discuss these checks further.

### 4.6.1. Checking date values

A check to ensure that that the Incorporated Date, Accounts Last Made Date, and Company Closed Date adhered to the ‘YYYY-MM-DD’ format was used to ensure that all date values were correct. Ignoring missing values, all records (for each of the attributes) adhered to the specified date format. Therefore, no value related issues exist for these attributes.

### 4.6.2. Checking Company Name values

To fix the company name value issues found during data understanding stage, company names beginning with the word ‘summary’ were removed from the dataset as they represented summary records and would therefore effect analysis values if left in. After removing these records, the remaining records were of the same format (“Business-” followed by a string of numbers) and thus do not have any remaining issues.

### 4.6.3. Checking postcode values

To fix the postcode value related issues found during data understanding stage a format check was used to find all records that start with other than a letter, only has alphanumeric values, and has a length of 4 to 7. After checking over the 40 records returned, the postcode values for each were removed due to being incorrect and would have a negative impact on analysis efforts.

# Chapter 5 - Data Analysis

This chapter details the data analysis efforts conducted for the task at hand, beginning with a general Exploratory Data Analysis stage, followed by exploration of the task-based questions.

## 5.1. Exploratory Data Analysis - Statistical summary of the dataset

A screenshot of a computer

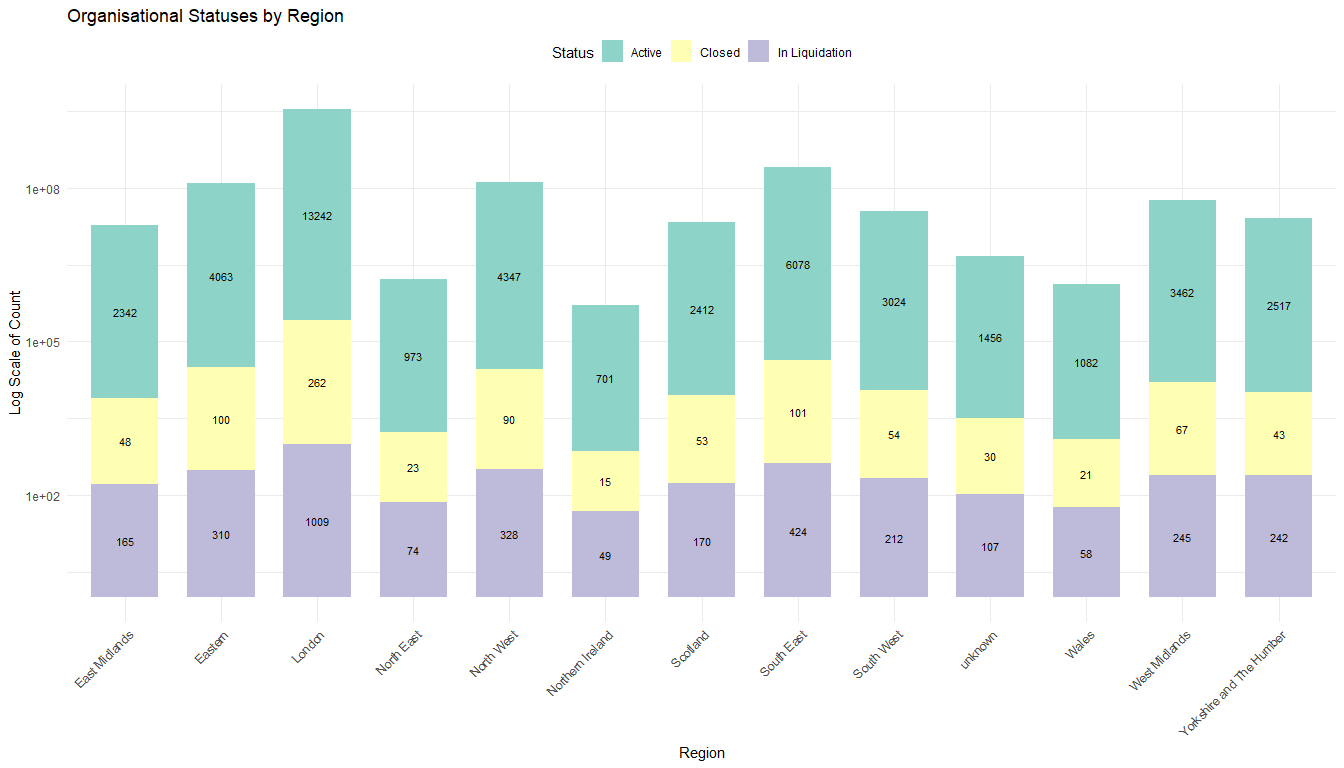
Description automatically generated

Viewing a statistical summary of the data we can learn that:

* The earliest incorporation date within the dataset is February 1902 and the latest is November 2024.
* 23 records don’t have an incorporation date which shouldn’t be missing because the dataset is about company registration.
* The mean turnover is the highest in 2020. This falls to the lowest average turnover of 590,573 in 2021. Since then, the turnover has been increasing year on year with the mean turnover falling to 765,641.
* The max turnover values for each year (2020 - 2024) are around 3.8 million showing that extreme outliers are not affecting the differences between mean values. The same concept of outliers and extreme values also applies to minimum turnover values.

## 5.2. Exploratory Data Analysis - Visual investigation

### 5.2.1. Viewing the statuses of organisations in the dataset (open, closed, in liquidation)



Viewing a visual summary of the data we can learn that:

* Most of the records in the dataset are open businesses, however in every region there are a handful of organisations that are closed or in liquidation. These findings are important to note when considering which records you want to keep when doing calculations / answering questions using the data.

## 5.3. Exploring task brief questions

### 5.3.1. Question 1 - Which sectors show the greatest increase and decrease in turnover over the years available?

#### Start vs current date

A graph showing a number of green squares

Description automatically generated

#### Using percentage change

The turnover for each Company Type, across all the years, can be seen below.

A graph of different colored squares

Description automatically generated with medium confidence

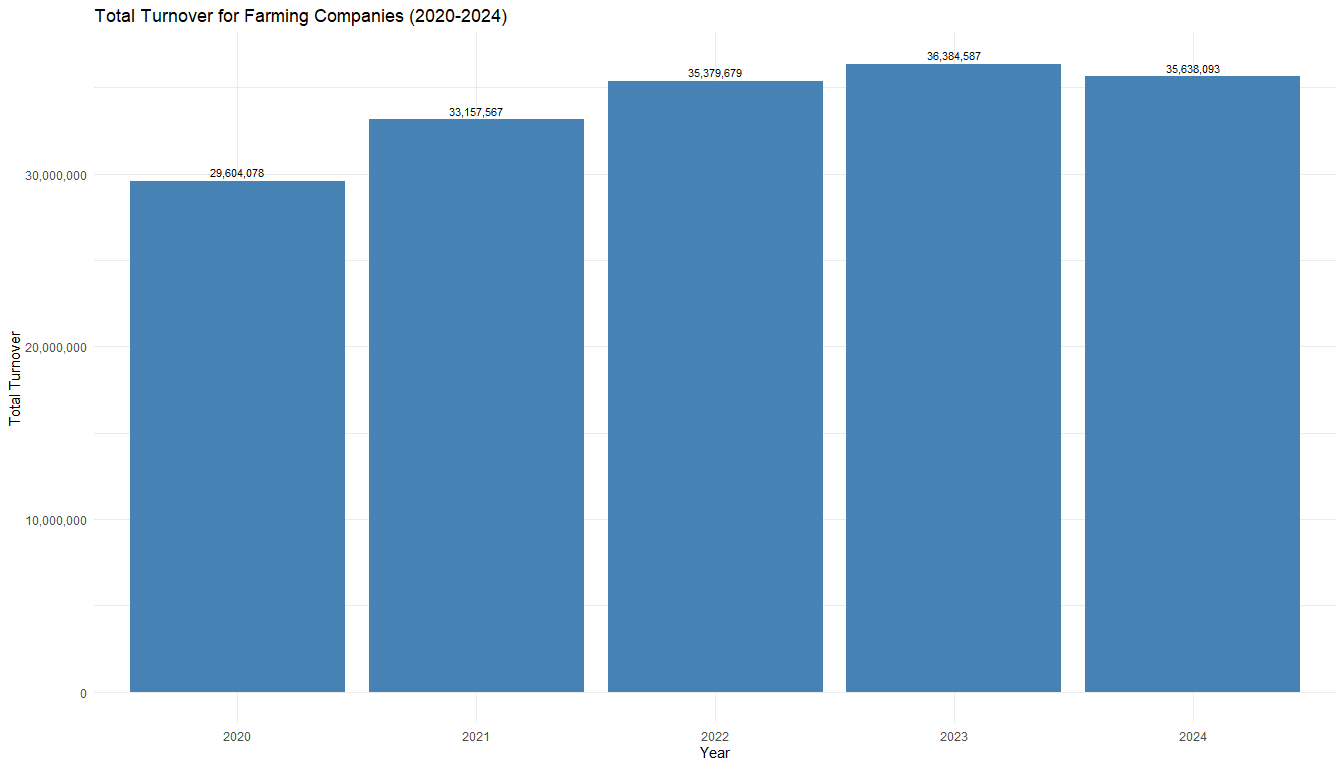
The size of the bars indicates the general increase and decrease over the years for each Sector however looking at percentage change will reveal exact statistical values. These values can be seen in the table below and the subsequent plot.

| Company Type | Year | PercentageChangeFromPrevious |
| --- | --- | --- |
| Transport and Logistics | 2022 | 58.7776927 |
| Construction | 2022 | 57.4471738 |
| Retail | 2022 | 57.3909926 |
| Consultancy | 2021 | 42.3841098 |
| Other | 2023 | 32.7744519 |
| Healthcare | 2021 | 18.6733218 |
| Consultancy | 2023 | 17.7246940 |
| Other | 2024 | 16.4861998 |
| Maintenance and Repair | 2022 | 13.5548778 |
| Consultancy | 2024 | 12.1010164 |
| Farming | 2021 | 12.0033767 |
| Other | 2022 | 11.3258706 |
| Manufacturing | 2021 | 11.0867077 |
| Maintenance and Repair | 2021 | 11.0095857 |
| Manufacturing | 2022 | 10.6873905 |
| Consultancy | 2022 | 10.3454716 |
| Healthcare | 2022 | 9.7590569 |
| Education | 2022 | 9.7064078 |
| Farming | 2022 | 6.7016739 |
| Retail | 2023 | 4.5999741 |
| Healthcare | 2023 | 4.1639773 |
| Transport and Logistics | 2023 | 4.0973181 |
| Manufacturing | 2023 | 3.9240219 |
| Construction | 2023 | 3.5322820 |
| Maintenance and Repair | 2023 | 2.8893317 |
| Farming | 2023 | 2.8403536 |
| Construction | 2024 | 0.4139964 |
| Maintenance and Repair | 2024 | 0.3649825 |
| Retail | 2024 | 0.0489977 |
| Healthcare | 2024 | -0.2436264 |
| Transport and Logistics | 2024 | -0.7082902 |
| Manufacturing | 2024 | -0.9732180 |
| Farming | 2024 | -2.0516764 |
| Education | 2021 | -11.2465039 |
| Education | 2024 | -16.8183146 |
| Education | 2023 | -17.3566393 |
| Construction | 2021 | -20.9258023 |
| Retail | 2021 | -20.9727069 |
| Transport and Logistics | 2021 | -21.9598605 |
| Other | 2021 | -30.8495061 |

From the percentage change values above, it can be determined that the ‘transport and logistics’ sector has experienced the greatest increase and decrease in turnover over the five years. A decrease in turnover between 2020 and 2021 was followed by a large increase the following year (2021 to 2022). This period of growth is likely to fall in line with the return to normality after covid-19 and can also be seen in other sectors such as retail and construction.

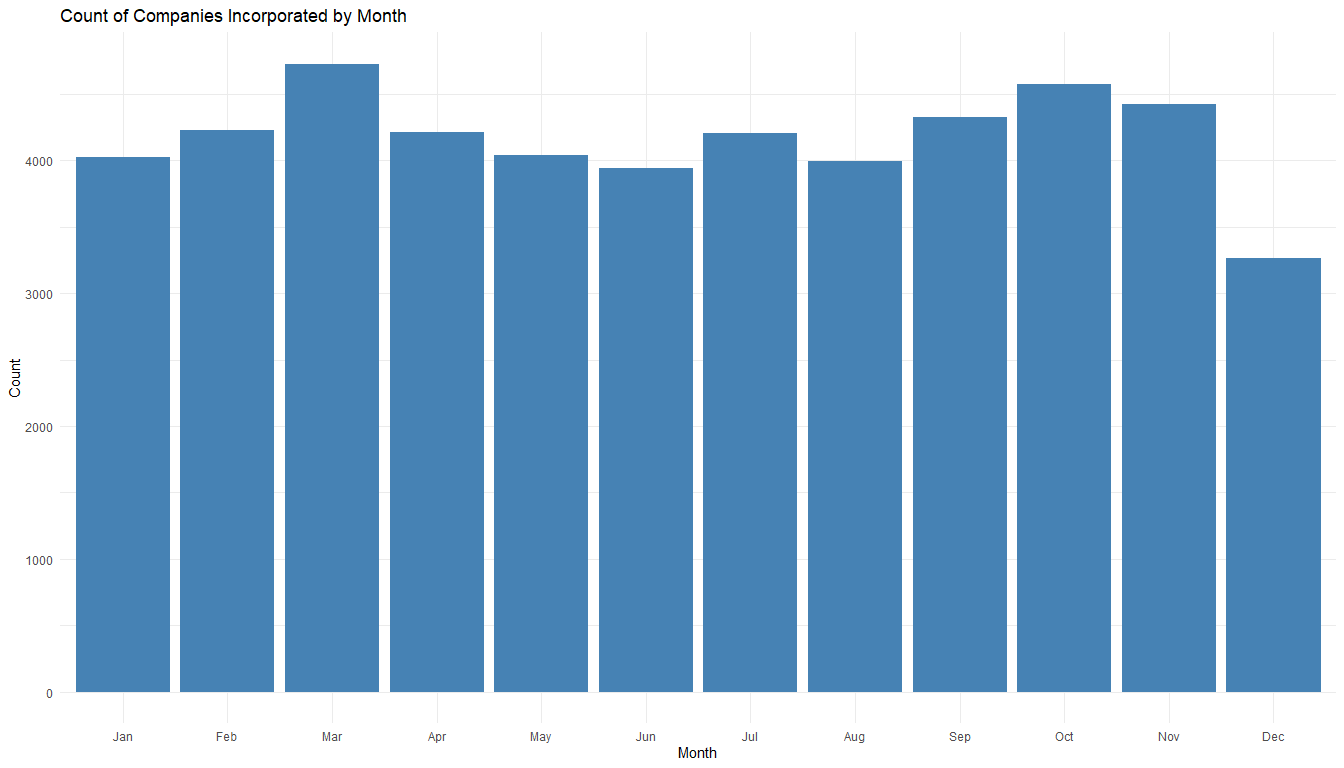
### 5.3.2. Question 2 - What is the total turnover for “Farming” broken down by year?

The total turnover for farming, broken down by years, can be seen in the following bar chart. The level of turnover has been increasing since 2020, however slightly less amounts are showing for 2024 than the year prior.

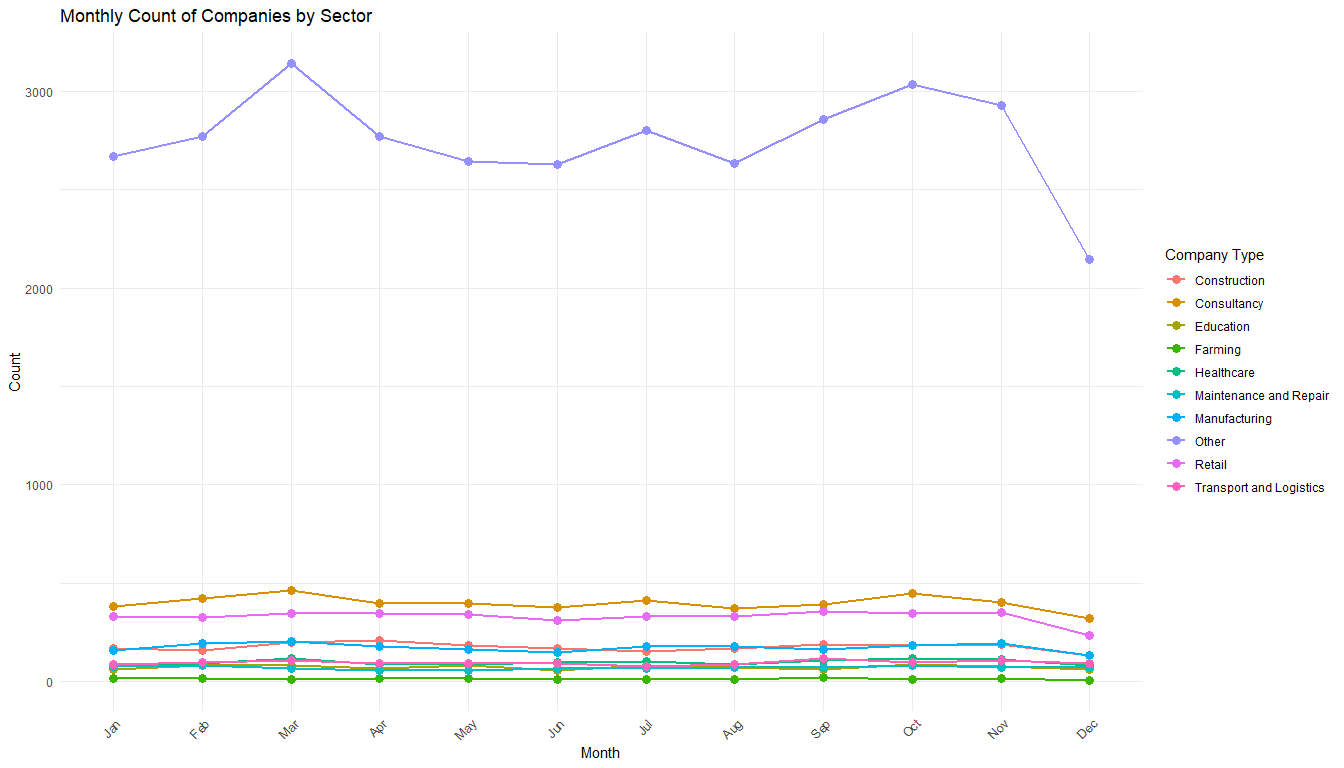


### 5.3.3. Question 3 - What is the LEAST common month to incorporate a business? Is this true for all Company Types?

The least common month to incorporate a business across all the years that the data spans is December.



This trend applies to most sectors too, with only 3 sectors having lower numbers of incorporation outside of December.



A screenshot of a graph

Description automatically generated

### 5.3.4. Question 4 - What other conclusions can be drawn from the data. We are particularly interested in trends over time, and comparisons between different geographic areas.

#### Total turnover by year and region.

A graph of different colored lines

Description automatically generated

Total Turnover for Each Year by Region with Totals

| Region | 2020 | 2021 | 2022 | 2023 | 2024 | Total |
| --- | --- | --- | --- | --- | --- | --- |
| East Midlands | 1,072,333,850 | 898,844,179 | 1,126,811,321 | 1,290,383,461 | 1,385,262,614 | 5,773,635,425 |
| Eastern | 2,011,059,707 | 1,706,150,022 | 2,089,658,496 | 2,400,312,481 | 2,559,448,994 | 10,766,629,700 |
| London | 6,466,402,162 | 5,518,289,834 | 6,785,214,262 | 7,672,100,457 | 8,195,802,802 | 34,637,809,517 |
| North East | 453,502,607 | 386,416,440 | 507,189,846 | 579,162,721 | 621,122,040 | 2,547,393,654 |
| North West | 2,047,115,092 | 1,764,843,752 | 2,169,755,523 | 2,492,133,511 | 2,637,484,220 | 11,111,332,098 |
| Northern Ireland | 363,328,014 | 311,667,083 | 374,774,221 | 425,392,369 | 451,493,997 | 1,926,655,684 |
| Scotland | 1,153,183,744 | 978,230,192 | 1,212,015,145 | 1,375,409,439 | 1,467,502,003 | 6,186,340,523 |
| South East | 2,944,113,722 | 2,504,924,183 | 3,042,593,350 | 3,458,348,624 | 3,689,882,615 | 15,639,862,494 |
| South West | 1,442,905,422 | 1,224,143,556 | 1,501,022,848 | 1,700,295,536 | 1,824,324,438 | 7,692,691,800 |
| Wales | 514,847,696 | 445,999,085 | 551,745,400 | 626,743,966 | 666,228,721 | 2,805,564,868 |
| West Midlands | 1,650,897,338 | 1,433,103,637 | 1,732,029,851 | 1,978,800,670 | 2,109,914,945 | 8,904,746,441 |
| Yorkshire and The Humber | 1,272,609,724 | 1,099,995,881 | 1,372,057,966 | 1,559,901,636 | 1,648,132,130 | 6,952,697,337 |
| unknown | 678,528,387 | 597,963,189 | 751,095,640 | 863,782,997 | 923,581,379 | 3,814,951,592 |
| Total | 22,070,827,465 | 18,870,571,033 | 23,215,963,869 | 26,422,767,868 | 28,180,180,898 | 118,760,311,133 |

#### Total turnover by year and Company Type

A graph of different colored lines

Description automatically generated

Total Turnover for Each Year by Company Type with Totals

| Company Type | 2020 | 2021 | 2022 | 2023 | 2024 | Total |
| --- | --- | --- | --- | --- | --- | --- |
| Construction | 1,351,898,374 | 1,069,002,793 | 1,683,114,685 | 1,742,567,042 | 1,749,781,207 | 7,596,364,101 |
| Consultancy | 1,216,780,870 | 1,732,502,610 | 1,911,738,176 | 2,250,587,918 | 2,522,931,930 | 9,634,541,504 |
| Education | 486,004,297 | 431,345,805 | 473,213,988 | 391,079,943 | 325,306,888 | 2,106,950,921 |
| Farming | 29,604,078 | 33,157,567 | 35,379,679 | 36,384,587 | 35,638,093 | 170,164,004 |
| Healthcare | 680,346,680 | 807,390,005 | 886,183,655 | 923,084,141 | 920,835,264 | 4,217,839,745 |
| Maintenance and Repair | 210,570,121 | 233,753,019 | 265,437,955 | 273,107,338 | 274,104,132 | 1,256,972,565 |
| Manufacturing | 3,760,266,476 | 4,177,156,228 | 4,623,585,227 | 4,805,015,723 | 4,758,252,444 | 22,124,276,098 |
| Other | 9,472,529,651 | 6,550,301,039 | 7,292,179,659 | 9,682,151,576 | 11,278,370,425 | 44,275,532,350 |
| Retail | 4,153,868,080 | 3,282,689,501 | 5,166,657,589 | 5,404,322,500 | 5,406,970,496 | 23,414,508,166 |
| Transport and Logistics | 708,958,838 | 553,272,466 | 878,473,256 | 914,467,100 | 907,990,019 | 3,963,161,679 |
| Total | 22,070,827,465 | 18,870,571,033 | 23,215,963,869 | 26,422,767,868 | 28,180,180,898 | 118,760,311,133 |

#### The status of companies within each region (open, closed, in liquidation)

Refer to section 5.2.1. - Viewing the statuses of organisation’s in the dataset (open, closed, in liquidation).

#### The number of companies within each sector per region

A graph with different colored squares

Description automatically generated

#### The Share of Yearly Turnover by Company Type

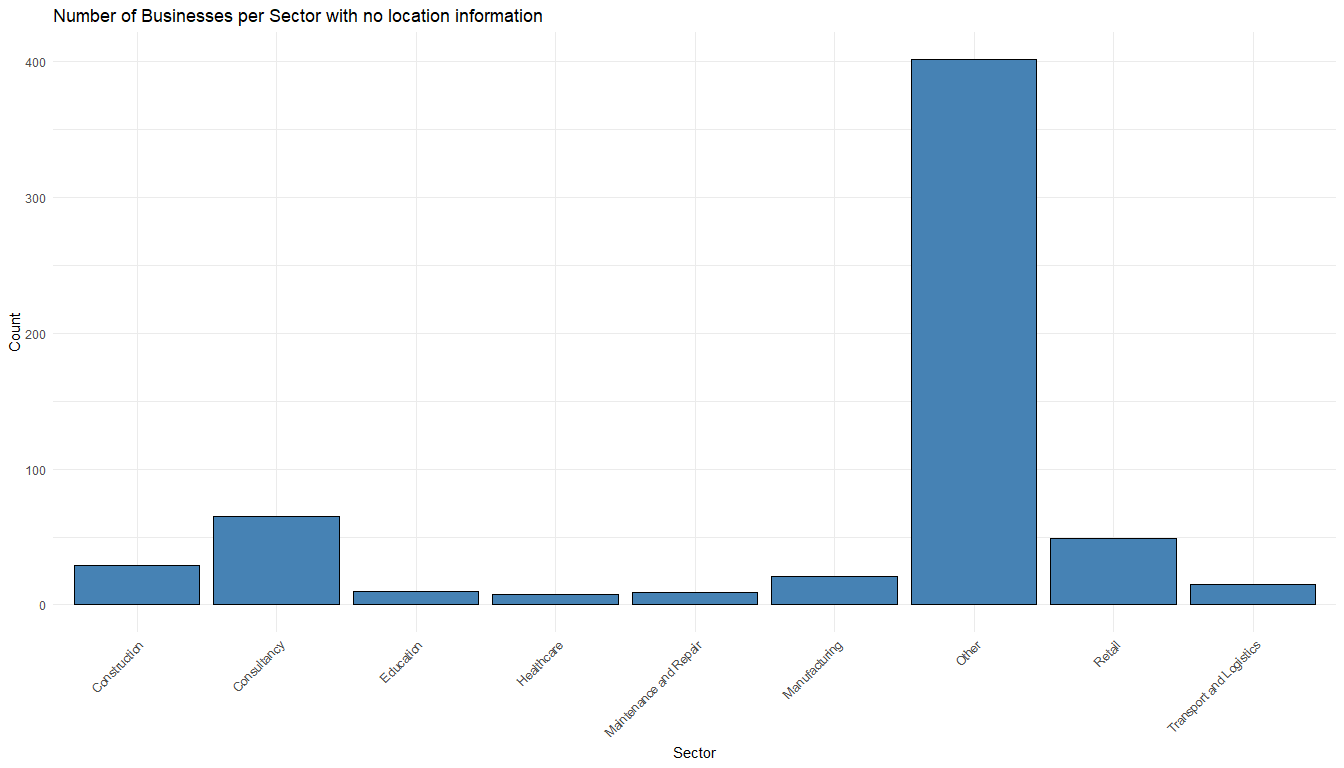
A graph with different colored squares

Description automatically generated with medium confidence

# Chapter 6 - Further Data Analysis

## 6.1. Analysing businesses without a known location.

Many businesses in the dataset have no identifiable location that links them to a geographical location. For organisations within sectors such as maintenance and repair this does not seem to be correct and could be a subject of further investigation, one such example is to check whether the accounts for currently open organisations are up to date.



## 6.2. Checking that all organisations in the dataset have updated their accounts in the last year.

Checking all open organisations for whether their accounts have been updated in the last year revealed that all the companies are up to date with their accounts.

# Chapter 7 – Insights sharing

## 7.1. Introduction

From the brief the deliverable for this task is to “to present an analysis of Company Registration data, along with turnover and business type.” To do so the presentation needs to cover your ***answers*** to the 4 questions mentioned in the task brief, including the ***methods*** used to reach your conclusions.

## 7.2. Answers & Methods

|  |  |  |
| --- | --- | --- |
| Question | Answer | Methods |
| Q1) Which sectors show the greatest increase and decrease in turnover over the years available? | The industries that show the greatest increase / decrease in turnover over the available years are the Transport and logistics, Construction, and Retail sectors. | Method 1: 2024 vs 2020 difference   1. Filtered the long formatted into a new data frame that only has 2020 and 2024 turnover data 2. Calculated the differences between the turnovers in 2024 and 2020 (2024 values - 2020 values) 3. Created a bar plot to visual the difference   Method 2: Annual percentage changes in turnover.   1. created a data frame from the long formatted dataset. grouping by company type and turnover year and summarising by the turnover sum 2. Calculated percentage change by sector and year 3. Formatted the output table - kable table 4. Create a faceted Line Plot 5. Then filtered again so that the data frame doesn't include any companies whose type is "other" 6. Calculated the sum turnover per sector per year 7. Created a stacked bar plot to view the results |
| Q2) What is the total turnover for “Farming” broken down by year? | The total turnover for the farming sector that has broken down by year, is shown in the bar chart in section 5.3.2. | 1. Filtered the non-long formatted dataset into a new dataframe that will only include any companies whose type is "farming" 2. Calculated the total turnover for each year 3. Created a bar plot to visualise these results |
| Q3) What is the LEAST common month to incorporate a business? Is this true for all Company Types? | The answer overall is December (6 of 9 sectors – 66.67%).  However, this is not true for all company types, namely education, maintenance and repair, and transport and logistics have a month outside of December | Part 1   1. Extracted into a new column the month from the incorporation date column (using the non-long format of the dataset) 2. Counted the occurrences of each month for each sector 3. Created a bar plot to view the results   Part 2   1. Filtered out the company type of other from the table 2. Identified the lowest value per company type via a new column. if the value is the lowest in that sector the column should take the value of islowest else other. 3. Created a scatter plot with facets for each sector and highlight lowest values |
| Q4) What other conclusions can be drawn from the data. We are particularly interested in trends over time, and comparisons between different geographic. |  | Insight 1 - The Share of Yearly Turnover by Company Type   * 1. using the long format dataset to create a new data frame in which: the data is grouped by company type and turnover year AND summarised by the turnover sum   2. Calculated the percentage contribution, for each company type, within each year.   3. Created a stacked bar chart adding the percentages within the stacked bar segments   Insight 2 - The number of companies within each sector per region   * + 1. Using the (non-long format) dataset to create a new data frame in which: the data is grouped by region and company type AND summarised by count     2. Created a stacked Bar Chart with a log scale for better visibility |

## 7.3. Presentation Plan

What needs to be in the presentation?

* Process & tools used to conduct the task
* Presentation = walking through each stage of the process
  + Business understanding
  + Developing data understanding - Screenshots of EDA stuff on slides
  + Data preparation
  + Data Analysis
  + Insights sharing

Create a visual that shows each of these stages and the activities done within each. This can be the content of the second slide – process and tools.

Plan for the 8 slides:

1. Introduction slide
2. Process and tools slide
3. Question 1 slide (give the answer & the method)
4. Question 2 slide (give the answer & the method)
5. Question 3 slide (give the answer & the method)
6. Question 4 slide (give the answer & the method)
7. Question 4 slide (give the answer & the method)
8. Conclusion slide (appendix - links to this report and the R-Markdown file)

# Appendix

* Presentation link:
* R-Markdown file link: [R-markdownFile\_Anas](https://github.com/miningAH/R-Markdown-company-registration-task/blob/main/hmrc.Rmd)

# References

Carpenter, John. 2024. “Sole Trader and Limited Company – What’s the Difference?” <https://www.1stformations.co.uk/blog/sole-trader-or-limited-company/#:~:text=A%20limited%20company%20is%20a,by%20one%20or%20more%20directors>.

“Companies House Accounts Guidance.” 2024. <https://www.gov.uk/government/publications/life-of-a-company-annual-requirements/life-of-a-company-part-1-accounts>.

Kurdi, Doaa. 2023. “The UK Postcode Format.” <https://ideal-postcodes.co.uk/guides/uk-postcode-format>.

“Postal Geographies.” 2023. <https://www.ons.gov.uk/methodology/geography/ukgeographies/postalgeography>.

“Postcode & Geolocation API for the UK.” <https://api.postcodes.io/>.

“Service Information.” <https://resources.companieshouse.gov.uk/serviceInformation.shtml#:~:text=Once%20incorporation%20at%20Companies%20House,be%20retained%20by%20Companies%20House>.

Townley, Gary. 2018. “What Does It Mean to Be a ’Limited’ Company?” <https://companieshouse.blog.gov.uk/2018/05/15/what-does-it-mean-to-be-a-limited-company/>.