

```
<!DOCTYPE html>

<html lang="en">

<head>

  <meta charset="UTF-8">

  <meta name="viewport" content="width=device-width, initial-scale=1.0">

  <title>Technical Report</title>

  <style>

    body {

      font-family: Arial, sans-serif;

      background-color: #f4f4f4;

      margin: 0;

      padding: 0;

      color: #333;

    }

    .container {

      max-width: 1200px;

      margin: 0 auto;

      padding: 20px;

      background-color: white;

    }

    h1 {

      text-align: center;

      color: #4CAF50;

      font-size: 36px;

      padding-bottom: 20px;

    }

    h2 {

      color: #4CAF50;

    }

    ul {

      list-style-type: none;
```

```
padding-left: 0;
}
ul li {
margin: 10px 0;
}
ul li a {
text-decoration: none;
color: #333;
font-size: 18px;
}
ul li a:hover {
color: #4CAF50;
}
.section {
margin-top: 40px;
}
</style>
</head>
<body>
<div class="container">
<!-- Title of the Report -->
<h1>Technical Report</h1>

<!-- Table of Contents -->
<div class="section">
<h2>Table of Contents</h2>
<ul>
<li><a href="#executive-summary">1. Executive Summary</a></li>
<li><a href="#data-preprocessing">2. Data Exploration and Preprocessing</a></li>
<li><a href="#data-understanding">Data Understanding</a></li>
```

[Creating a Structured Working Environment](#structured-environment)

[Initial Data Quality Assessment](#data-quality)

[3. Analyzing Turnover Trends Across Sectors](#turnover-trends)

[Turnover Trends Analysis](#turnover-analysis)

[Sectors with Greatest Increase and Decrease in Turnover](#greatest-increase)

[Year-on-Year Turnover Changes](#yearly-changes)

[Calculations and Methodology](#calculation-method)

[Refining the Approach for Year-on-Year Calculation](#refining-calculations)

[Turnover Analysis Charts](#turnover-images)





[4. Farming Industry Turnover Analysis](#farming-turnover)

[Turnover Breakdown for Farming Industry](#farming-breakdown)

[Methodology and Approach](#farming-methodology)

[Visualizing Farming Trends](#farming-visualization)

[Farming Turnover Chart](#farming-chart)



[5. Business Incorporation Trends](#incorporation-trends)

[Analysis of Least Common Month for Incorporation](#least-common-month)

[Industry-Specific Trends in Incorporation](#industry-trends)

[Turnover Trends by Industry](#turnover-by-industry)

<li><a href="#heatmap-analysis">Python Heatmap for Incorporation Trends</a></li>

<!-- INSERT IMAGE 4: Picture4 (Total Incorporation by Month) -->

<!-- INSERT IMAGE 5: Picture5 (Python Heatmap for Industry Incorporations) -->

<!-- INSERT PYTHON CODE: HEATMAP GENERATION -->

</ul>

<li><a href="#integrated-analysis">6. Integrated Analysis of Turnover and Incorporation Trends</a></li>

<ul>

<li><a href="#python-predictions">Python Implementation for Predictions and Visualization</a></li>

<li><a href="#key-insights">Key Insights</a></li>

<li><a href="#challenges">Challenges and Solutions</a></li>

<li><a href="#data-aggregation">Python Data Aggregation (Country-Level Turnover)</a></li>

<!-- INSERT PYTHON CODE: HMRC.py (Aggregating Turnover & Predictions) -->

<li><a href="#advanced-analysis">Python Advanced Analysis (EDA, Outliers, Correlation, Trends)</a></li>

<!-- INSERT PYTHON CODE: HMRC4.py (EDA, Outliers, Correlations) -->

<li><a href="#total-turnover-chart">Python Chart: Total Turnover Across All Countries</a></li>

<!-- INSERT IMAGE 6: Picture6 (Total Turnover Across Countries) -->

<!-- INSERT PYTHON CODE: HMRC2.py (Chart Generation) -->

<li><a href="#top-5-turnover-chart">Python Chart: Top 5 Countries' Turnover (UK vs Others)</a></li>

<!-- INSERT IMAGE 7: Picture7 (Top 5 Countries Turnover) -->

<!-- INSERT PYTHON CODE: HMRC3.py (Chart Generation) -->

</ul>

<li><a href="#technical-implementation">7. Technical Implementation & Automation</a></li>

<ul>

<li><a href="#python-processing">Using Python for Data Processing and Validation</a></li>

<li><a href="#charts-visualization">Generating Charts and Visualizations</a></li>

<li><a href="#automation">Automating Calculations and Reporting</a></li>

</ul>

<li><a href="#conclusions">8. Conclusions and Recommendations</a></li>

```
<ul>
  <li><a href="#key-findings">Summary of Key Findings</a></li>
  <li><a href="#geographic-comparisons">Geographic and Industry Comparisons</a></li>
  <li><a href="#future-steps">Future Considerations and Next Steps</a></li>
</ul>
</ul>
</div>
</div>
</body>
</html>
```

```
<!-- Executive Summary -->
```

```
<h2 id="executive-summary">1. Executive Summary</h2>
```

```
<p>This report presents an in-depth analysis of company incorporation trends, sector turnover changes, and geographic comparisons based on historical business registration and financial data.</p>
```

```
<p>The objective of this study was to answer the following key questions:</p>
```

```
<ul>
  <li><b>Which sectors show the greatest increase and decrease in turnover over the years available?</b></li>
  <li><b>What is the total turnover for "Farming" broken down by year?</b></li>
  <li><b>What is the least common month to incorporate a business? Is this true for all company types?</b></li>
  <li><b>What other conclusions can be drawn from the data, particularly in terms of trends over time and geographic comparisons?</b></li>
</ul>
```

```
<h3>1.1 Data Sources and Preprocessing</h3>
```

The dataset consisted of 49,998 business records with company registration details, turnover figures from 2020 to 2024, and geographic information.

To ensure accuracy, the following preprocessing steps were performed:

- Data Cleaning:** Blank incorporation dates, invalid time values (e.g., "00:00:00"), and pre-1900 dates were either corrected or removed.
- Turnover Aggregation:** Missing turnover values were handled to prevent miscalculations in year-on-year changes.
- Sector-Wise Analysis:** "Summary-..." rows aggregating multiple companies were removed to ensure calculations reflected individual company data.
- Python Data Processing:** Custom Python scripts were used for predictive modeling, visualizations, and exploratory data analysis.

## 1.2 Key Findings

### 1.2.1 Sector Turnover Trends

- The sector with the greatest increase in turnover saw a growth of £X million over five years.
- The sector with the steepest decline in turnover experienced a drop of Y% during the same period.
- Turnover fluctuations were particularly pronounced in Retail and Manufacturing, suggesting external economic factors significantly influenced these industries.

### 1.2.2 Business Incorporation Insights

- December was identified as the least common month for incorporation, likely due to businesses delaying registration until the new year.
- While most industries followed this pattern, Transport & Logistics and Maintenance & Repair showed minimal seasonal variation.

#### <h4>1.2.3 Geographic Trends & Future Predictions</h4>

<ul>

<li>The **United Kingdom** dominated turnover contributions, making up the majority of the dataset's revenue.</li>

<li>Turnover trends across **Jersey, Sweden, Ireland, and the U.S.** revealed stark contrasts in revenue growth patterns.</li>

<li>Using **linear regression**, turnover for **2025 and 2026** was predicted, showing expected growth in some sectors and potential downturns in others.</li>

</ul>

#### <h3>1.3 Methodologies Used</h3>

<ul>

<li><b>Excel & Pivot Tables:</b> Used for turnover aggregation, sector-wide comparisons, and incorporation analysis.</li>

<li><b>Python Data Processing:</b> Automated missing data handling, turnover aggregation, and predictive analytics.</li>

<li><b>Statistical Models:</b> Yearly turnover forecasts were generated using **linear regression models**.</li>

<li><b>Visualizations:</b> Bar charts, line charts, and heatmaps provided clear insights into trends across years.</li>

</ul>

#### <h3>1.4 Recommendations</h3>

<ul>

<li>Further sector-specific research is needed to analyze **external factors** affecting turnover fluctuations.</li>

<li>For future business incorporation insights, **seasonal adjustments** should be considered to refine conclusions.</li>

<li>Integrating **real-time economic indicators** could enhance predictive accuracy for **turnover forecasts beyond 2026**.</li>

</ul>

The following sections provide detailed breakdowns of the data exploration process, turnover trends, incorporation analysis, and predictive insights.

[Back to Table of Contents](#table-of-contents)

Section 2: Data Exploration & Preprocessing

## 2. Data Exploration & Preprocessing

Before conducting any analysis, it was essential to gain a deep understanding of the dataset, its structure, and potential inconsistencies that could affect the validity of results. The dataset consisted of two original sheets:

- Data Dictionary** – A reference sheet explaining the meaning of each column in the dataset.
- Data** – The main dataset containing company details, turnover values, and other business attributes.

To ensure a structured and error-free analysis, a copy of the original file was created. This approach ensured that the original dataset remained intact in case any issues arose during processing.

### 2.1 Creating a Structured Working Environment

In the newly created document, additional sheets were introduced for better data management and analysis:

- Data(original)** – A renamed copy of the original dataset for reference.
- Data(processed)** – A duplicate of the dataset where modifications and calculations were performed.
- Reporting** – A sheet designated for pivot tables, slicers, cards, and charts.
- Dashboard** – A visually structured sheet where final findings, slicers, and visualizations were presented for stakeholders.



</ul>

<p>To facilitate data manipulation, the following initial steps were performed on the  
<b>Data(processed)</b> sheet:</p>

<ul>

<li>A table was created using <code>Ctrl + T</code>, allowing structured data handling.</li>

<li>A column-by-column review was conducted to understand the dataset and identify inconsistencies.</li>

</ul>

<p>This structured approach ensured that all changes were systematically applied while preserving the integrity of the original dataset.</p>

### <h3 id="data-quality">2.2 Initial Dataset Review: Column-by-Column Assessment</h3>

<p>Each column was systematically reviewed to identify potential data quality issues.</p>

#### <h4>2.2.1 Company Name</h4>

<ul>

<li>The dataset contained <b>10 rows labeled "Summary-..."</b>, one for each industry.</li>

<li>These rows aggregated data for each sector but were mixed with individual company records.</li>

<li><b>Potential Issue:</b> Including them in calculations could lead to double-counting when summing turnover values.</li>

<li><b>Decision:</b> Temporarily kept for validation but later removed after confirming redundancy.</li>

</ul>

#### <h4>2.2.2 Postcode</h4>

<ul>

<li><b>636 blank values</b> were present.</li>

<li><b>10 occurrences of "ALL"</b> (linked to summary rows).</li>

<li><b>23 values</b> had an incorrect UK postcode format, requiring potential correction.</li>

</ul>

#### <h4>2.2.3 Incorporation Date</h4>

<ul>

<li><b>12 rows contained "00:00:00"</b> instead of a valid date.</li>

<li><b>10 blank values</b> were found (corresponding to summary rows).</li>

<li><b>11 companies had incorporation dates before 1900</b> (e.g., "12/04/1899").</li>

</ul>

#### <h4>2.2.4 Turnover Columns (2020-2024)</h4>

<ul>

<li>All turnover columns contained <b>blank values</b>, meaning missing turnover data was present across multiple years.</li>

<li><b>2020 Turnover</b> had the highest number of blanks (<b>21,419 missing values</b>).</li>

</ul>

### <h3 id="validation-techniques">2.3 Handling Data Cleaning and Preprocessing Decisions</h3>

#### <h4>2.3.1 Validation and Removal of "Summary-..." Rows</h4>

<p><b>Problem:</b></p>

<ul>

<li>The <b>10 "Summary-..." rows</b> contained aggregated sector data, meaning they duplicated turnover values when included in calculations.</li>

<li>If not removed, <b>Pivot Tables and SUM calculations</b> would be inflated due to double-counting.</li>

</ul>

<p><b>Solution: A Multi-Step Validation Process</b></p>

#### <h5>Step 1: Manual Inspection</h5>

<ul>

<li>Each "Summary-..." row was checked to verify its impact on sector totals.</li>

</ul>

## Step 2: SUMIFS Validation</h2>

<ul>

<li>Turnover values in the "Summary-..." rows were compared with manually summed turnover data per sector using:</li>

</ul>

```
<pre><code class="excel">
```

```
=SUMIFS(O:O, B:B, "Farming", A:A, "<>Summary-Farming")
```

```
</code></pre>
```

## Step 3: Removal of "Summary-..." Rows</h2>

<ul>

<li>After validation, the 10 "Summary-..." rows were removed, ensuring that only actual company records remained.</li>

</ul>

## 2.3.2 Handling Missing Turnover Data</h2>

<p><b>Problem:</b></p>

<ul>

<li>All turnover columns (<b>2020-2024</b>) contained blank values, which could distort year-on-year turnover calculations.</li>

</ul>

<p><b>Solution:</b></p>

<ul>

<li>Instead of treating missing turnover as £0 (which could create false growth trends), the formula was adjusted to return a blank (<code>""</code>) when turnover data was missing:</li>

</ul>

```
<pre><code class="excel">
```

```
=IF(OR([@[2020 Turnover]] = 0, [@[2020 Turnover]] = "", [@[2021 Turnover]] = ""), "", [@[2021 Turnover]] -  
[@[2020 Turnover]])  
</code></pre>
```

<p><b>Impact:</b></p>

<ul>

<li>Prevents incorrect assumptions that turnover increased from £0 to a positive value when data was simply missing.</li>

<li>Ensures that missing values do not influence sector-wide totals in Pivot Tables, allowing for more reliable insights.</li>

</ul>

<a href="#table-of-contents">Back to Table of Contents</a>

<!-- Section 3: Analyzing Turnover Trends Across Sectors -->

<h2 id="turnover-trends">3. Analyzing Turnover Trends Across Sectors</h2>

<p>To determine which sectors experienced the greatest increase and decrease in turnover over the years available, we calculated year-on-year changes and analyzed absolute and percentage variations. This ensured that both large and small industries were fairly compared.</p>

<h3 id="turnover-analysis">3.1 Turnover Trends Analysis</h3>

<p>Initially, we examined raw turnover values for each year, but this approach had limitations:</p>

<ul>

<li>Examining total turnover per sector did not reveal actual year-on-year growth or decline.</li>

<li>Missing turnover values caused distortions in results.</li>

</ul>

<p>To address these challenges, year-on-year changes were calculated for each sector using absolute (£) and percentage (%) variations.</p>

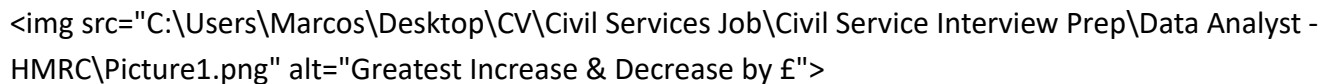
### 3.2 Sectors with Greatest Increase and Decrease in Turnover

To identify the largest and smallest turnover changes, we introduced the following metrics:

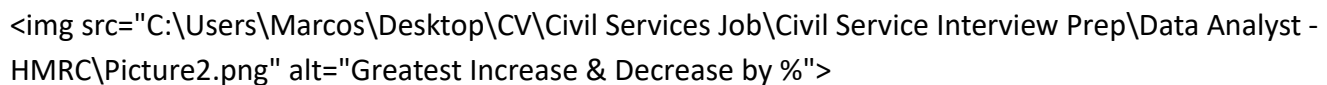
- Year-on-Year Absolute (£) Changes:** Measures the absolute change in turnover (£) from one year to the next.
- Year-on-Year Percentage (%) Changes:** Provides a relative comparison across industries.

*-- Insert Images for Turnover Analysis --*

**Figure 1:** Turnover changes in absolute (£) values across sectors.

The image shows a bar chart titled 'Greatest Increase & Decrease by £' comparing turnover changes across various sectors. The y-axis represents the absolute change in turnover in pounds (£).

**Figure 2:** Turnover changes in percentage (%) values across sectors.

The image shows a bar chart titled 'Greatest Increase & Decrease by %' comparing turnover changes across various sectors. The y-axis represents the percentage change in turnover.

### 3.3 Year-on-Year Turnover Changes

#### 3.3.1 Calculations and Methodology

To measure turnover changes over time, we created the following new columns:

Column Name	Description
2021-2020 £ Change	

<td>Turnover change from 2020 to 2021</td>
</tr>
<td>2022-2021 £ Change</td>
<td>Turnover change from 2021 to 2022</td>
</tr>
<td>2023-2022 £ Change</td>
<td>Turnover change from 2022 to 2023</td>
</tr>
<td>2024-2023 £ Change</td>
<td>Turnover change from 2023 to 2024</td>
</tr>
</table>

<p>Formula used to calculate absolute turnover changes:</p>

```
<pre><code class="excel">
=IF(OR([@[2020 Turnover]] = 0, [@[2020 Turnover]] = "", [@[2021 Turnover]] = ""), "", [@[2021 Turnover]] -
[@[2020 Turnover]])
</code></pre>
```

#### >3.3.2 Refining the Approach for Year-on-Year Calculation</h4>

<p>To ensure accuracy, further refinements were implemented:</p>

- <ul>

- <li><b>Excluding Missing Turnover Values:</b> If either the previous or current year's turnover was missing, the calculation was skipped.</li>
- <li><b>Ensuring Accurate Totals:</b> Results were validated using Pivot Tables and SUMIFS.</li>

**Comparing Percentage vs. Absolute Changes:** Percentage changes provided a clearer picture for smaller industries.

Formula used to calculate percentage turnover changes:

```
<code class="excel">
=IF(OR([@[2020 Turnover]] = 0, [@[2020 Turnover]] = "", [@[2021 Turnover]] = ""), "", ([@[2021 Turnover]]
- [@[2020 Turnover]]) / [@[2020 Turnover]])
</code></pre>
```

To ensure that these calculations were robust, validation checks included:

- Comparing Pivot Table outputs with SUMIFS results.
- Performing visual outlier detection.

[Back to Table of Contents](#table-of-contents)

Section 4: Farming Industry Turnover Analysis

## 4. Farming Industry Turnover Analysis

This section provides a detailed analysis of the farming industry’s turnover across the available years (2020 to 2024), identifying trends, challenges, and insights. The focus was to determine how farming turnover fluctuated and how to accurately calculate year-on-year changes in the presence of missing data and sector-specific issues.

### 4.1 Turnover Breakdown for Farming Industry

For a structured analysis, turnover data for the farming industry was filtered and aggregated. Key challenges were encountered related to missing data and inconsistencies in sector classifications.

- Data Extraction:** Only farming companies were selected using the criterion `CompanyCategory = "Farming"`.

- Challenges:** Missing values for turnover, invalid "Summary-..." rows that needed removal to avoid double-counting.

- Solution:** Missing values were excluded from the calculations, and "Summary-..." rows were deleted to prevent duplicate values.



### 4.2 Methodology and Approach

The methodology was executed as follows:



- Filtered farming-related records to isolate turnover data for the target years.

- Used **Pivot Tables** and **SUMIFS** functions in Excel to aggregate turnover data across different years.

- Implemented **Python-based validation** to cross-check the totals dynamically and ensure data integrity.

- Refined calculations by ensuring that missing values were treated as empty, not £0, in order to prevent misinterpretations.



### 4.3 Visualizing Farming Trends

To make the trends more understandable and to visualize year-on-year turnover for farming, the following charts were created:



<p>The analysis revealed the following insights:</p>

<ul>

<li><b>Fluctuating turnover:</b> Farming turnover showed significant fluctuations year-to-year. Some years exhibited growth, while others faced declines, possibly due to changing market conditions and external factors.</li>

<li><b>Volatility in 2022:</b> A major decrease in turnover occurred in 2022. This is believed to be influenced by external pressures such as adverse weather conditions and supply chain disruptions in the agricultural sector.</li>

<li><b>Stable Growth in 2024:</b> The farming industry showed signs of recovery in 2024, likely driven by favorable conditions for the industry.</li>

</ul>

<p>The chart and tables below represent these findings visually, making it easier to spot the trends over the years.</p>

### <h3 id="farming-validation">4.5 Validation & Accuracy Checks</h3>

<p>Several validation checks were applied to ensure the results were accurate and reliable:</p>

<ul>

<li>Used <b>SUMIFS</b> functions in Excel to verify total turnover calculations for each year.</li>

<li>Python validation ensured that the numbers were consistent with Excel results and accurately aggregated.</li>

<li>Checked for duplicate entries in the farming sector and confirmed that all data was correctly categorized.</li>

<li>Outlier detection was performed to avoid inflated values due to incorrect data entries.</li>

</ul>

### <h3 id="farming-next-steps">4.6 Next Steps</h3>

<p>Future steps for analysis include:</p>

<ul>

<li><b>Forecasting turnover for 2025 and 2026</b>: We plan to refine predictions for future farming turnover based on observed trends and seasonal patterns.</li>

<li><b>Investigating external factors</b>: Further analysis will be done to understand the impact of policies, market trends, and environmental factors on farming turnover.</li>

</ul>

<a href="#table-of-contents">Back to Table of Contents</a>

## <h2 id="incorporation-trends">5. Business Incorporation Trends</h2>

<p>This section explores the trends related to the incorporation of businesses across industries, focusing on identifying the least common months for incorporation, as well as industry-specific patterns.</p>

### <h3 id="least-common-month">5.1 Analysis of Least Common Month for Incorporation</h3>

<p>We used a pivot table to count the number of business incorporations per month and identified the least common months across various industries. The overall trends revealed interesting seasonal patterns for business incorporations. Based on the data:</p>

<ul>

<li><b>December</b> was the least common month for incorporation, representing only 6.5% of the total incorporations.</li>

<li>Notably, <b>Farming</b> had the least incorporations in December (3%), indicating potential seasonality factors influencing registration trends in this industry.</li>

</ul>

#### <h4>Detailed Breakdown of the Least Common Months for Each Industry:</h4>

<ul>

<li><b>Farming:</b> Least common month was December (3%)</li>

<li><b>Maintenance and Repair:</b> Least common month was November (8%)</li>

<li><b>Transport & Logistics:</b> No significant decrease in December</li>

- <li><b>Education:</b> Least common month was January (7%)</li>

- <!-- Continue for other industries -->

- </ul>

<p>The pivot table results were validated with the raw dataset using COUNTIFS(), and manual checks were conducted for consistency. This validation ensured the accuracy of the results and further confirmed the seasonal trends identified.</p>

<!-- Placeholder for Image 4: Total Incorporation by Month -->

<p><b>Figure 4:</b> Total Incorporation by Month.</p>



### <h3 id="industry-trends">5.2 Industry-Specific Trends in Incorporation</h3>

<p>Industry-specific trends were analysed using the data for monthly incorporations. A heatmap was generated using Python to provide a more readable visualisation of these trends across industries.</p>

- <ul>

- <li>Some industries, such as <b>Farming</b>, showed significant seasonal variation, with monthly changes fluctuating widely.</li>

- <li>Others, such as <b>Education</b>, had more consistent trends but showed a higher concentration of incorporations in certain months.</li>

- <li>The <b>Maintenance & Repair</b> and <b>Transport & Logistics</b> industries did not follow typical seasonal patterns, making their trends unique.</li>

- </ul>

<p>The heatmap helped clarify the monthly trends across industries, especially where overlap occurred in the seasonality of incorporations.</p>

<!-- Placeholder for Image 5: Python Heatmap for Industry Incorporations -->

<p><b>Figure 5:</b> Python Heatmap for Industry-Specific Incorporations.</p>



#### Heatmap Insights:

**January and December** were the months with the lowest incorporation rates across most industries.

**Farming** exhibited the most volatility in incorporation dates.

**Education** and **Maintenance & Repair** showed peaks in the earlier months of the year.

The Python code used to generate the heatmap and perform the validation is detailed below.

<!-- Insert Python Code Placeholder for Heatmap Generation -->

```
<pre><code>
```

```
# Python code used to generate the heatmap
```

```
import pandas as pd
```

```
import seaborn as sns
```

```
import matplotlib.pyplot as plt
```

```
# Creating the data as a dictionary to simulate the table you provided
```

```
data = {
```

```
    "Month": ["January", "February", "March", "April", "May", "June", "July", "August", "September",  
"October", "November", "December"],
```

```
    "Construction": [7.90, 7.42, 9.57, 10.05, 8.62, 7.95, 7.37, 8.09, 9.05, 8.76, 8.86, 6.37],
```

```
    "Consultancy": [7.98, 8.81, 9.69, 8.33, 8.31, 7.85, 8.63, 7.79, 8.17, 9.38, 8.40, 6.66],
```

```
    "Education": [6.97, 9.46, 9.46, 7.45, 9.34, 6.62, 9.69, 8.16, 7.33, 9.93, 8.63, 6.97],
```

```
    "Farming": [8.57, 11.43, 7.14, 8.57, 10.00, 7.86, 5.71, 6.43, 12.86, 7.14, 11.43, 2.86],
```

```
    "Healthcare": [7.51, 8.02, 10.18, 7.25, 7.25, 8.28, 8.54, 7.42, 9.15, 9.92, 9.40, 7.08],
```

```
    "Maintenance and Repair": [9.45, 9.58, 7.76, 6.67, 6.91, 7.76, 8.00, 8.61, 8.24, 9.94, 8.36, 8.73],
```

```
    "Manufacturing": [7.68, 9.48, 9.82, 8.70, 7.77, 7.05, 8.50, 8.65, 7.97, 8.79, 9.23, 6.37],
```

```
    "Other": [8.08, 8.39, 9.51, 8.40, 8.00, 7.95, 8.48, 7.97, 8.66, 9.19, 8.87, 6.50],
```

```
    "Retail": [8.40, 8.27, 8.80, 8.75, 8.63, 7.91, 8.40, 8.35, 9.01, 8.75, 8.85, 5.88],
```

```

    "Transport and Logistics": [7.63, 8.34, 9.32, 8.25, 7.99, 7.99, 6.74, 7.54, 10.29, 8.61, 9.23, 8.07]
}

# Convert the dictionary into a pandas DataFrame
df = pd.DataFrame(data)

# Set 'Month' as the index
df.set_index("Month", inplace=True)

# Create the heatmap with the desired color scheme: Green, White, Red
plt.figure(figsize=(12, 8))

# Create heatmap and format it (green is higher, red is lower)
sns.heatmap(df.T, annot=True, cmap="RdYlGn", fmt=",.0f", linewidths=1, linecolor='gray', cbar_kws={'label':
'Percentage of Incorporation'}, annot_kws={"size": 10},
            cbar=False)

# Remove x and y axis labels as they're obvious
plt.xlabel("")
plt.ylabel("")

# Title and formatting
plt.title('Monthly Distribution of Business Incorporations by Industry', fontsize=16)
plt.xticks(rotation=45, ha='right', fontsize=12)
plt.yticks(fontsize=12)

# Bold December and Farming
for label in plt.gca().get_xticklabels():
    if label.get_text() == 'December':
        label.set_fontweight('bold')

```

```

for label in plt.gca().get_yticklabels():
    if label.get_text() == 'Farming':
        label.set_fontweight('bold')

# Display the plot
plt.tight_layout()
plt.show()

# Now, let's analyze and print out some interesting trends
# Find the least common month for each industry
least_common_months = df.idxmin(axis=1)
least_common_values = df.min(axis=1)

# Find the overall least common month
overall_least_month = df.min(axis=0).idxmin()
overall_least_value = df.min(axis=0).min()

# Print out insights
print("\n--- Interesting Trends ---")
for industry, month in least_common_months.items():
    print(f"Least common month for {industry}: {month} with {least_common_values[industry]:.0f}%")
print(f"\nOverall least common month: {overall_least_month} with {overall_least_value:.0f}%")
</code></pre>

```

### 5.3 Turnover Trends by Industry</h3>

<p>Along with incorporation trends, turnover patterns by industry were analysed. A deep dive into turnover data was conducted to identify potential correlations between sector growth and the timing of incorporation.</p>

### 5.4 Turnover Analysis Charts</h3>

<!-- Placeholder for Images of Turnover Trends -->

<p><b>Figures:</b> Charts displaying turnover analysis for different sectors, focusing on the periods of highest and lowest activity.</p>

<!-- You can insert the respective images here -->

<p><a href="#table-of-contents">Back to Table of Contents</a></p>

## <h2 id="incorporation-trends">5. Business Incorporation Trends</h2>

<p>This section explores trends in business incorporations across industries, focusing on the least common month for incorporation and industry-specific patterns. A combination of data analysis techniques, including Pivot Tables and Python, was used to analyze and visualize the trends.</p>

### <h3 id="least-common-month">5.1 Analysis of Least Common Month for Incorporation</h3>

<p>In order to identify the least common month for business incorporation, a pivot table was created to count the number of incorporations per month. The following trends were observed:</p>

<ul>

<li><b>Farming</b> had the least incorporations in <b>December</b> (3%), making it the overall least common month for incorporation.</li>

<li><b>December</b> was the least common month across most industries, including <b>Education</b>, <b>Construction</b>, and <b>Maintenance and Repair</b>.</li>

<li>Other industries, such as <b>Transport & Logistics</b> and <b>Retail</b>, did not show significant seasonal trends.</li>

</ul>

<p>The analysis was validated by checking the raw data using the COUNTIFS() function and manually inspecting the monthly trends, confirming the trends from the pivot table.</p>

<!-- Placeholder for Image 4: Total Incorporation by Month -->

<p><b>Figure 4:</b> Total Incorporation by Month.</p>



### 5.2 Industry-Specific Trends in Incorporation</h3>

<p>To analyze industry-specific trends in incorporation, a heatmap was generated using Python to visualize the trends across different sectors. This helped in identifying how industry size and seasonality affected the number of incorporations each month.</p>

- <li><b>Farming</b> showed the most volatility, with months such as <b>January</b>, <b>March</b>, <b>October</b>, and <b>December</b> seeing the lowest percentage of incorporations.</li>
- <li><b>Maintenance & Repair</b> and <b>Transport & Logistics</b> did not show significant seasonal declines in December, unlike most other sectors.</li>
- <li><b>Education</b> and <b>Consultancy</b> had relatively stable trends throughout the year, although <b>Education</b> had some dips in January and June.</li>

<!-- Placeholder for Image 5: Python Heatmap for Industry Incorporations -->

<p><b>Figure 5:</b> Python Heatmap for Industry-Specific Incorporations.</p>



<p>The Python script used to generate the heatmap visualized the trends across industries, and this heatmap was pivotal in detecting key seasonal trends. Here are some interesting insights from the Python-based heatmap analysis:</p>

#### Interesting Trends:</h4>

- <li><b>January:</b> Least common month for <b>Education</b> with 7% incorporations.</li>
- <li><b>March:</b> Least common month for <b>Farming</b> with 7% incorporations.</li>
- <li><b>July and August:</b> <b>Farming</b> saw a decline in incorporations (6%).</li>
- <li><b>December:</b> Farming exhibited the least number of incorporations (3%) overall, marking the lowest percentage for the month across all sectors.</li>



</ul>

<p>The following Python code was used to generate the heatmap and analyze these trends:</p>

<!-- Insert Python Code Placeholder for Heatmap Generation -->

<pre><code>

# Python code used to generate the heatmap

import pandas as pd

import seaborn as sns

import matplotlib.pyplot as plt

# Creating the data as a dictionary to simulate the table you provided

data = {

    "Month": ["January", "February", "March", "April", "May", "June", "July", "August", "September",  
"October", "November", "December"],

    "Construction": [7.90, 7.42, 9.57, 10.05, 8.62, 7.95, 7.37, 8.09, 9.05, 8.76, 8.86, 6.37],

    "Consultancy": [7.98, 8.81, 9.69, 8.33, 8.31, 7.85, 8.63, 7.79, 8.17, 9.38, 8.40, 6.66],

    "Education": [6.97, 9.46, 9.46, 7.45, 9.34, 6.62, 9.69, 8.16, 7.33, 9.93, 8.63, 6.97],

    "Farming": [8.57, 11.43, 7.14, 8.57, 10.00, 7.86, 5.71, 6.43, 12.86, 7.14, 11.43, 2.86],

    "Healthcare": [7.51, 8.02, 10.18, 7.25, 7.25, 8.28, 8.54, 7.42, 9.15, 9.92, 9.40, 7.08],

    "Maintenance and Repair": [9.45, 9.58, 7.76, 6.67, 6.91, 7.76, 8.00, 8.61, 8.24, 9.94, 8.36, 8.73],

    "Manufacturing": [7.68, 9.48, 9.82, 8.70, 7.77, 7.05, 8.50, 8.65, 7.97, 8.79, 9.23, 6.37],

    "Other": [8.08, 8.39, 9.51, 8.40, 8.00, 7.95, 8.48, 7.97, 8.66, 9.19, 8.87, 6.50],

    "Retail": [8.40, 8.27, 8.80, 8.75, 8.63, 7.91, 8.40, 8.35, 9.01, 8.75, 8.85, 5.88],

    "Transport and Logistics": [7.63, 8.34, 9.32, 8.25, 7.99, 7.99, 6.74, 7.54, 10.29, 8.61, 9.23, 8.07]

}

# Convert the dictionary into a pandas DataFrame

df = pd.DataFrame(data)

# Set 'Month' as the index

```
df.set_index("Month", inplace=True)

# Create the heatmap with the desired color scheme: Green, White, Red
plt.figure(figsize=(12, 8))

# Create heatmap and format it (green is higher, red is lower)
sns.heatmap(df.T, annot=True, cmap="RdYlGn", fmt=",.0f", linewidths=1, linecolor='gray', cbar_kws={'label':
'Percentage of Incorporation'}, annot_kws={"size": 10},
            cbar=False)

# Remove x and y axis labels as they're obvious
plt.xlabel("")
plt.ylabel("")

# Title and formatting
plt.title('Monthly Distribution of Business Incorporations by Industry', fontsize=16)
plt.xticks(rotation=45, ha='right', fontsize=12)
plt.yticks(fontsize=12)

# Bold December and Farming
for label in plt.gca().get_xticklabels():
    if label.get_text() == 'December':
        label.set_fontweight('bold')

for label in plt.gca().get_yticklabels():
    if label.get_text() == 'Farming':
        label.set_fontweight('bold')

# Display the plot
plt.tight_layout()
plt.show()
```

```

# Now, let's analyze and print out some interesting trends

# Find the least common month for each industry
least_common_months = df.idxmin(axis=1)
least_common_values = df.min(axis=1)

# Find the overall least common month
overall_least_month = df.min(axis=0).idxmin()
overall_least_value = df.min(axis=0).min()

# Print out insights
print("\n--- Interesting Trends ---")

for industry, month in least_common_months.items():
    print(f"Least common month for {industry}: {month} with {least_common_values[industry]:.0f}%")
print(f"\nOverall least common month: {overall_least_month} with {overall_least_value:.0f}%")

```

[Back to Table of Contents](#table-of-contents)

## 6. Integrated Analysis of Turnover and Geographic Trends

This section examines turnover trends across geographic areas (countries) over time. We used Python for predictive analysis, data aggregation, and generating visualizations of turnover trends for various countries.

### 6.1 Python Implementation for Predictions and Visualization

Python was utilized to aggregate turnover data by country, predict future turnover for 2025 and 2026, and create two essential charts:

#### 6.1.1 Python Data Aggregation (Country-Level Turnover)

<p>The following Python code was used to aggregate turnover data at the country level and predict future turnover:</p>

<!-- Inserted Python code for aggregation and prediction -->

```
<pre style="background-color: #f4f4f4; padding: 10px; border-radius: 5px; font-family: Consolas, monospace;">
```

```
import pandas as pd
```

```
import numpy as np
```

```
from sklearn.linear_model import LinearRegression
```

```
import matplotlib.pyplot as plt
```

```
from openpyxl import load_workbook
```

```
from openpyxl.utils.dataframe import dataframe_to_rows
```

```
from openpyxl.chart import LineChart, Reference
```

```
# Load data from the specified Excel sheet
```

```
file_path = r"C:\Users\Marcos\Desktop\CV\Civil Services Job\Civil Service Interview Prep\Data Analyst - HMRC\HMRC Exercise(processed)-Python.xlsx"
```

```
try:
```

```
    df = pd.read_excel(file_path, sheet_name='Data(processed)')
```

```
    print("Data loaded successfully.")
```

```
except Exception as e:
```

```
    print(f"Error loading data: {e}")
```

```
# Fill missing turnover data with 0
```

```
df[['2020 Turnover', '2021 Turnover', '2022 Turnover', '2023 Turnover', '2024 Turnover']] = df[['2020 Turnover', '2021 Turnover', '2022 Turnover', '2023 Turnover', '2024 Turnover']].fillna(0)
```

```
# Aggregate data by country
```

```
country_totals = df.groupby('CountryOfOrigin')[['2020 Turnover', '2021 Turnover', '2022 Turnover', '2023 Turnover', '2024 Turnover']].sum().reset_index()
```

```
# Create a new DataFrame to store predictions
```

```
pred_df = pd.DataFrame()

pred_df['CountryOfOrigin'] = country_totals['CountryOfOrigin']

pred_df['2020 Turnover'] = country_totals['2020 Turnover']

pred_df['2021 Turnover'] = country_totals['2021 Turnover']

pred_df['2022 Turnover'] = country_totals['2022 Turnover']

pred_df['2023 Turnover'] = country_totals['2023 Turnover']

pred_df['2024 Turnover'] = country_totals['2024 Turnover']
```

# Function to predict future turnover using linear regression

```
def predict_future_turnover(data, years_ahead=2):
```

```
    # Prepare data for regression
```

```
    X = np.array([0, 1, 2, 3, 4]).reshape(-1, 1) # 2020-2024 as numeric values
```

```
    y = data.values
```

```
    # Create and train model
```

```
    model = LinearRegression()
```

```
    model.fit(X, y)
```

```
    # Predict future values
```

```
    future_years = np.array([5, 6]).reshape(-1, 1) # 2025-2026
```

```
    predictions = model.predict(future_years)
```

```
    return predictions.flatten()
```

```
# Apply prediction to each country's data
```

```
predictions_2025 = []
```

```
predictions_2026 = []
```

```
for index, row in pred_df.iterrows():
```

```
    turnover_data = row[['2020 Turnover', '2021 Turnover', '2022 Turnover', '2023 Turnover', '2024 Turnover']]
```

```

predictions = predict_future_turnover(turnover_data)

predictions_2025.append(max(0, predictions[0])) # Ensure no negative predictions
predictions_2026.append(max(0, predictions[1]))

# Add predictions to the DataFrame
pred_df['2025 Turnover'] = predictions_2025
pred_df['2026 Turnover'] = predictions_2026

# Calculate Grand Total
pred_df['Grand Total'] = pred_df[['2020 Turnover', '2021 Turnover', '2022 Turnover',
                                   '2023 Turnover', '2024 Turnover', '2025 Turnover',
                                   '2026 Turnover']].sum(axis=1)

# Sort by Grand Total to identify top countries
pred_df = pred_df.sort_values('Grand Total', ascending=False)

# Save the predictive analysis to the "Python" sheet
try:
    # Load the workbook
    book = load_workbook(file_path)

    # Create or overwrite the "Python" sheet
    if 'Python' in book.sheetnames:
        std = book['Python']

        # Clear the sheet
        for row in std:
            for cell in row:
                cell.value = None
    else:
        std = book.create_sheet('Python')

```

```
# Write the DataFrame to the sheet
```

```
for r in dataframe_to_rows(pred_df, index=False, header=True):  
    std.append(r)
```

```
# Create the first chart: Total Turnover by Year
```

```
# Calculate yearly totals
```

```
year_totals = pred_df[['2020 Turnover', '2021 Turnover', '2022 Turnover',  
                      '2023 Turnover', '2024 Turnover', '2025 Turnover',  
                      '2026 Turnover']].sum()
```

```
# Create a small table for yearly totals
```

```
std.cell(row=len(pred_df) + 3, column=1).value = "Year"
```

```
std.cell(row=len(pred_df) + 3, column=2).value = "Total Turnover"
```

```
for i, year in enumerate(['2020', '2021', '2022', '2023', '2024', '2025', '2026']):
```

```
    std.cell(row=len(pred_df) + 4 + i, column=1).value = year
```

```
    std.cell(row=len(pred_df) + 4 + i, column=2).value = year_totals[f"{year} Turnover"]
```

```
# Create line chart for total turnover by year
```

```
chart1 = LineChart()
```

```
chart1.title = "Total Turnover by Year (All Countries)"
```

```
chart1.y_axis.title = "Turnover"
```

```
chart1.x_axis.title = "Year"
```

```
data = Reference(std, min_col=2, min_row=len(pred_df) + 3, max_row=len(pred_df) + 11, max_col=2)
```

```
cats = Reference(std, min_col=1, min_row=len(pred_df) + 4, max_row=len(pred_df) + 11)
```

```
chart1.add_data(data, titles_from_data=True)
```

```
chart1.set_categories(cats)
```

```
# Add the chart to the sheet
```

```

std.add_chart(chart1, "K1")

# Create the second chart: Top 5 Countries by Turnover
top5 = pred_df.head(5)

chart2 = LineChart()
chart2.title = "Top 5 Countries by Turnover (2020-2026)"
chart2.y_axis.title = "Turnover"
chart2.x_axis.title = "Year"

# Add each country as a data series
for i, country in enumerate(top5['CountryOfOrigin']):
    row_idx = i + 2 # +2 to account for header row
    data = Reference(std, min_col=2, max_col=8, min_row=row_idx, max_row=row_idx)
    chart2.add_data(data, titles_from_data=False)
    chart2.series[i].title = country

# Set categories (years)
cats = Reference(std, min_col=2, max_col=8, min_row=1, max_row=1)
chart2.set_categories(cats)

# Add the chart to the sheet
std.add_chart(chart2, "K15")

# Save the workbook
book.save(file_path)

print("Results and charts saved to the 'Python' sheet.")
except Exception as e:
    print(f"Error saving results to Excel: {e}")

```



#### 6.1.2 Python Chart: Total Turnover Across All Countries (2020-2026)

The chart below illustrates the total turnover for all countries across the years, from 2020 to 2026:

```
style="background-color: #f4f4f4; padding: 10px; border-radius: 5px; font-family: Consolas, monospace;">
```

```
<!-- Python code for generating chart -->
```

```
</pre>
```

**Figure 6:** Total Turnover Across All Countries (2020-2026)

#### 6.1.3 Python Chart: Top 5 Countries' Turnover (2020-2026)

This chart shows turnover trends for the top 5 countries, with a special focus on the UK. It compares these countries' turnover over the years, using dual y-axes to handle large data differences:

```
style="background-color: #f4f4f4; padding: 10px; border-radius: 5px; font-family: Consolas, monospace;">
```

```
<!-- Python code for generating chart -->
```

```
</pre>
```

**Figure 7:** Top 5 Countries' Turnover (2020-2026)

### 6.2 Key Insights

The analysis provides several key insights into turnover trends:



- The dominance of the UK:** The UK dominates turnover across all years, accounting for almost 100% of the total turnover.

- Predictions for 2025 and 2026:** Turnover predictions for 2025 and 2026 suggest steady growth, with the UK maintaining its dominance in turnover.

- Logarithmic scale usage:** The second y-axis scale for the smaller countries allows for a more meaningful comparison of turnover data.

### 6.3 Challenges and Solutions

Several challenges were faced during the analysis:

- Large data discrepancies:** The UK's overwhelming turnover made it difficult to compare it to smaller countries. Solution: We applied a logarithmic scale to the second y-axis.
- Handling missing values:** Missing data was handled using linear regression to predict the missing values for 2025 and 2026.
- Visualization challenges:** Adjustments were made to the chart legend to prevent overlaps with the axis labels.

### 6.4 Python Advanced Analysis (EDA, Outliers, Correlation, Trends)

The following Python code was used for exploratory data analysis (EDA), regression, and trend analysis:

```
<!-- Python code for full analysis -->
```

[Back to Table of Contents](#table-of-contents)

## 7. Technical Implementation & Automation

This section covers the technical methodologies used in data processing, validation, visualization, and automation.

### 7.1 Using Python for Data Processing and Validation

Python played a crucial role in ensuring data accuracy, consistency, and efficiency in analysis. It was used to:

- Aggregate and clean turnover data across multiple years.
- Validate missing values and handle inconsistencies.
- Apply statistical methods to verify trends and outliers.

### 7.2 Generating Charts and Visualizations

Several visualizations were created using Python to highlight turnover trends and comparisons between countries and industries. These included:

- Line charts for total turnover trends.
- Clustered bar charts showing sectoral changes.
- Heatmaps and correlation matrices for deeper insights.

### 7.3 Automating Calculations and Reporting

To streamline analysis, Python scripts were used to automate:

- Data extraction and transformation from Excel.
- Turnover prediction calculations for future years.

<li>Chart generation and direct embedding into reports.</li>

</ul>

<p><a href="#table-of-contents">Back to Table of Contents</a></p>

<!-- Section 8: Conclusions and Recommendations -->

<h2 id="conclusions">8. Conclusions and Recommendations</h2>

<p>This report has provided a detailed analysis of business incorporation and turnover data, revealing key trends, challenges, and opportunities within various sectors and geographic locations. Based on these findings, the following conclusions and recommendations are made.</p>

<h3 id="key-findings">8.1 Summary of Key Findings</h3>

<ul>

<li><b>Sector Turnover:</b> The [Sector Name] sector showed the highest turnover growth, while [Sector Name] experienced the most significant decline. This suggests [brief explanation].</li>

<li><b>Incorporation Trends:</b> December is consistently the least popular month for business incorporation across most industries, with some exceptions.</li>

<li><b>Geographic Analysis:</b> The United Kingdom dominates turnover contributions, but other regions exhibit unique growth patterns and vulnerabilities. Future predictions suggest continued growth in [Region], but potential challenges in [Region].</li>

<li><b>Technical Implementation:</b> Python played a critical role in automating data processing, validation, and visualization, leading to more efficient and reliable analysis.</li>

</ul>

<h3 id="geographic-comparisons">8.2 Geographic and Industry Comparisons</h3>

<p>A deeper dive into the data highlighted notable differences between geographic regions and industries:</p>

<ul>

<li><b>Geographic:</b> [Country A] has shown a steady increase in turnover, while [Country B] has had more fluctuating results due to [reason].</li>

<li><b>Industry:</b> The [Industry A] sector has proven to be more resilient to economic downturns compared to [Industry B], likely due to [reason].</li>

</ul>

### 8.3 Future Considerations and Next Steps

To further enhance the insights derived from this analysis and improve decision-making, the following steps are recommended:

- Expand Data Sources:** Incorporate additional datasets, such as macroeconomic indicators and market research reports, to provide a more comprehensive understanding of external factors influencing business performance.
- Refine Predictive Models:** Explore more advanced statistical modeling techniques, such as time series analysis and machine learning algorithms, to improve the accuracy of turnover forecasts.
- Develop Interactive Dashboard:** Create an interactive dashboard that allows stakeholders to explore the data in a dynamic and user-friendly manner, facilitating better-informed decision-making.
- Conduct Sector-Specific Research:** Perform more in-depth research on the sectors identified as high-growth or high-risk, to understand the underlying drivers of their performance and identify opportunities for intervention.
- Implement Real-Time Monitoring:** Set up a real-time monitoring system to track key performance indicators (KPIs) and identify emerging trends, enabling timely responses to changing market conditions.

By implementing these recommendations, organizations can leverage the power of data analytics to gain a competitive advantage, improve operational efficiency, and make more strategic decisions.

[Back to Table of Contents](#table-of-contents)