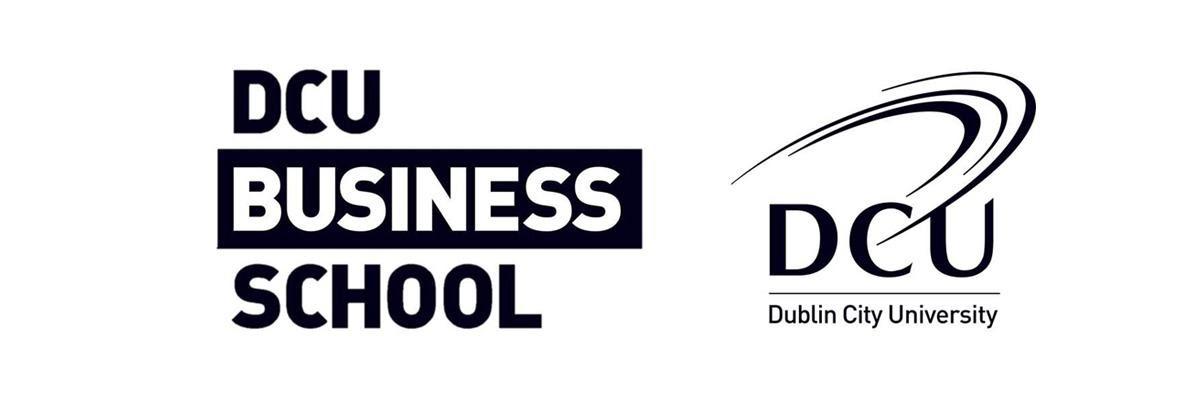
**E-Portfolio Report**



In the study program Global Business Germany (EBG4) and the module Professional Business Analytics Portfolio (MT412) at DCU Business School

Submitted by:

Lasse Willen

Student-ID: 22108556

Submitted to:

Dr. Mathieu Mercadier

Dublin, 13/04/2024

Table of Contents

[1. Grid Details 3](#_Toc163900887)

[1.1 Portfolio Link 3](#_Toc163900888)

[1.2 Repository Link 3](#_Toc163900889)

[2. Portfolio 3](#_Toc163900890)

[2.1 Project Number 3](#_Toc163900891)

[2.2 Tools number 3](#_Toc163900892)

[2.3 Data 3](#_Toc163900893)

[2.3.1 Data Quality 3](#_Toc163900894)

[2.3.2 Data Preprocessing 4](#_Toc163900895)

[2.4 Analytics 4](#_Toc163900896)

[2.4.1 Code Quality 4](#_Toc163900897)

[2.4.2 Code Length 4](#_Toc163900898)

[2.4.3 Analytics & 2.4.4 Machine Learning Diversity 4](#_Toc163900899)

[2.5 Visualization 6](#_Toc163900900)

[2.5.1 Visualizations Number 6](#_Toc163900901)

[2.5.2 Visual Type 6](#_Toc163900902)

[2.5.3 Report 7](#_Toc163900903)

[2.5.4 Report Interactivity 7](#_Toc163900904)

[3. Additional Pages 8](#_Toc163900905)

[3.1 Biography 8](#_Toc163900906)

[3.2 Resume 8](#_Toc163900907)

[3.3 Media Category 8](#_Toc163900908)

[3.4 Media Relevance 8](#_Toc163900909)

[3.5 Blogs 8](#_Toc163900910)

[3.6 Contact 9](#_Toc163900911)

[3.7 Optional 9](#_Toc163900912)

# 1. Grid Details

I have provided detailed information for every criterion of the report.

The code documentation is in the repository. E.g. Project5\_...\_.py

## 1.1 Portfolio Link

<https://lassewillen.github.io/Business_Analytics_Portfolio_Lasse_Willen/index.html>

## 1.2 Repository Link

<https://github.com/LasseWillen/Business_Analytics_Portfolio_Lasse_Willen>

# 2. Portfolio

## 2.1 Project Number

There are six projects available in the portfolio. 5 tools were used: Python, MS Excel, MS PowerPoint, SQL, and Power BI. Please note that I embedded the Microsoft applications as the Google versions due to compatibility reasons. The titles of the projects are:

1. Visualizing Second-Hand Car Purchases in Ireland
2. Heart Attack Diagnosis Analysis
3. Exploring University Class Dataset
4. G20 Energy and Sustainability Report
5. Vehicle Price Predictions and Clustering
6. Product Review Sentiment Analysis

## 2.2 Tools number

5 tools were used: Python, MS Excel, MS PowerPoint, SQL, and Power BI. Please note that I embedded the Microsoft applications as the Google versions due to compatibility reasons. The titles of the projects are:

* Visualizing Second-Hand Car Purchases in Ireland: PowerPoint
* Heart Attack Diagnosis Analysis: SQL
* Exploring University Class Dataset: Python, MS Excel
* G20 Energy and Sustainability Report: Power BI, MS PowerPoint
* Vehicle Price Predictions and Clustering: Python
* Product Review Sentiment Analysis

## 2.3 Data

### 2.3.1 Data Quality

Project 4 G20 Energy and Sustainability Report: The selection of the dataset was driven by our group's conviction that climate change and energy consumption are pressing global concerns. We deliberately opted for a clean dataset from the World Bank, ensuring the integrity and reliability of our analysis by utilizing a reputable and authoritative source. In our presentation, we meticulously detailed the columns employed, spotlighting crucial data that would inform our insights. Through DAX Code Column programming, we enriched the dataset by adding columns that differentiate between developed and emerging nations, as well as between renewable and non-renewable energy sources. This enrichment is pivotal for a nuanced analysis that recognizes the complexities within global energy dynamics. Additionally, our dataset poses no conflict with GDPR standards since it contains no personal data, thus eliminating concerns regarding data privacy. Our approach reflects a deep understanding of the dataset's significance, aligned with our commitment to excellence in communication and analytics in the realm of environmental sustainability.

The dataset is linked on the website and in the portfolio.

### 2.3.2 Data Preprocessing

Used in Project 5, Vehicle Price Predictions. The code demonstrates a comprehensive approach to data preprocessing and feature engineering for a car prices dataset, essential for ensuring data quality and model accuracy. Missing values are strategically handled by dropping incomplete rows and imputing the 'condition' field with its median, preserving data integrity. Text data normalization (converting to lowercase) avoids duplications and ensures consistency across entries. Categorical data, crucial for regression analysis, are transformed using OneHotEncoder, allowing models to interpret these variables correctly. Additionally, outliers in 'odometer' readings are identified using the IQR method and addressed by removing extreme values, mitigating their potential to skew predictions. These steps are encapsulated in a scikit-learn Pipeline, streamlining the workflow and preventing data leakage, thus enhancing the robustness and reliability of the predictive models derived from this data.

The code is available as a .ipynb and .py file in the repository. The dataset can be found using the link on the portfolio website as it would be too big to upload on GitHub.

## 2.4 Analytics

### 2.4.1 Code Quality

Used in Project 5, Vehicle Price Predictions. The code demonstrates excellent organization and readability, crucial attributes for effective software development in data science. It employs a structured format with consistent use of comments to guide the reader through each step, from data importation to model evaluation. Variable and function names are descriptively and clearly chosen, making the code intuitive to understand and follow. For instance, functions like `count\_outliers\_iqr` and variables such as `outlier\_counts` explicitly convey their purpose without ambiguity. The use of Python's `Pipeline` and `ColumnTransformer` from scikit-learn not only streamlines the workflow but also cleanly separates preprocessing steps from model training, thereby enhancing code readability and maintainability. These practices are not just about adhering to coding standards but also about facilitating collaboration and ensuring that the code can be efficiently updated and debugged, key aspects of high-quality software development in any data-driven field.

The code is available as a .ipynb and .py file in the repository. The dataset can be found using the link on the portfolio website as it would be too big to upload on GitHub.

### 2.4.2 Code Length

* Visualizing Second-Hand Car Purchases in Ireland: No Code
* Heart Attack Diagnosis Analysis: 51 lines excluding comments
* Exploring University Class Dataset: 50 lines excluding comments
* G20 Energy and Sustainability Report: No Code
* Vehicle Price Predictions and Clustering: 153 lines excluding comments
* Product Review Sentiment Analysis 101 lines excluding comments

Codes for all the projects using code are available in the GitHub repository as both .ipynb and .py files. The SQL code is available in a Word document.

### 2.4.3 Analytics & 2.4.4 Machine Learning Diversity

*Please note that I combined these 2 criteria for simplicity and clarity reasons*

Project: Vehicle Price Predictions:

Supervised Learning

With Linear Regression Algorithm for Price Predictions.

The code aims to predict car prices. Since Linear Regression was too inaccurate, I diverted to using a Random Forest Regression algorithm to improve the accuracy utilizing advanced statistical and machine learning techniques for enhanced accuracy. It meticulously preprocesses the data, normalizing numerical features and encoding categorical ones, to ensure optimal input quality for the RandomForest algorithm. This approach not only increases the model's predictive performance, as evidenced by evaluations using MSE and R² scores, but also aligns with best practices in data science to derive actionable insights effectively. The use of a sophisticated ensemble method further underscores a commitment to leveraging complex algorithms to improve prediction reliability and understanding of feature influences on pricing.

Unsupervised Learning

The code is designed to cluster car data into distinct groups based on vehicle characteristics, which helps in identifying patterns and insights about different types of cars. It uses advanced machine learning techniques by first dropping the target variable 'sellingprice' to focus purely on the features. Numerical features like 'year', 'odometer', and 'condition' are scaled to standardize their ranges, while categorical features such as 'make' and 'model' are transformed into one-hot vectors to properly encode their categorical nature for the KMeans clustering algorithm. The code employs a pipeline to streamline the preprocessing and clustering processes, ensuring data integrity and simplification of the workflow. After clustering, it utilizes TruncatedSVD for dimensionality reduction to visualize the data clusters in two dimensions, enhancing interpretability of the clustering results by providing a visual representation of how cars are grouped based on their features. This structured approach not only aids in efficient data analysis but also aligns with best practices in utilizing unsupervised learning for insightful data exploration.

Project: Product Review Sentiment Analysis

Deep Learning

The code integrates advanced Natural Language Processing (NLP) techniques and deep learning algorithms for sentiment analysis, specifically focusing on evaluating product reviews. Using the NLTK library, the script initially conducts basic text analysis, such as tokenization and part-of-speech tagging, to prepare the data. The core of the analysis involves using the Vader SentimentIntensityAnalyzer from NLTK to assess sentiment based on polarity scores, which quantify the emotional tone of the texts.

To further refine sentiment detection, the code employs a pre-trained deep learning model, Roberta, from the transformers library, optimized for sentiment analysis. This model, which has been trained on vast amounts of text data, including Twitter posts, provides a nuanced understanding of context and the interplay of words in sentences. The application of Roberta allows for a high-precision sentiment evaluation, making it highly effective in identifying subtle expressions of feelings.

This combination of traditional NLP techniques and cutting-edge deep learning enhances the ability to accurately gauge consumer sentiment, providing valuable insights into customer satisfaction and product reception. Such dual-method analysis is crucial for businesses to understand and predict consumer behavior and to refine product offerings accordingly.

The code is available as a .ipynb and .py file in the repository. The dataset can be found using the link on the portfolio website as it would be too big to upload on GitHub.

## 2.5 Visualization

### 2.5.1 Visualizations Number

* Visualizing Second-Hand Car Purchases in Ireland
* Tools: PowerPoint, Visualization Count: 13
* Interactive, embedded PowerPoint Presentation with 12 slides and 13 graphs that showcase car sales development
* Heart Attack Diagnosis Analysis
* Tools: None, only a screenshot of code
* Exploring University Class Dataset
* Tools: Excel, Python, PowerPoint, Visualization Count: 3 + 5 Excel tables +1 Excel visualization
* 2 interactive Excel sheets, 1 correlation matrix visual from Python, including code. One Report included in the page w/ 5 Excel Tables and 2 PPT visualizations, 1 correlation matrix visual from Excel
* G20 Energy and Sustainability Report
* Tools: PowerBI, PowerPoint, Visualization Count: 15 PowerBI, 1 PowerPoint
* This is the Power BI Report. Since I do not have a Power BI license, the Power BI Report is integrated into an embedded PowerPoint presentation. There are 4 PowerBI Slides with a total of 16 visualizations in the presentation and 8 slides in total.
* Vehicle Price Predictions and Clustering
* Tools: Python, Visualization Count: 1
* One Graphic made using Python that depicts car clustering
* Product Review Sentiment Analysis
* Tools: Python, Visualization Count: 1
* One Graphic made using Python that depicts the results of the Vader analysis

🡪 total number of tools used for visualizations: 4: Excel, Python, PowerPoint, PowerBI

🡪 total number of visualizations: 34

### 2.5.2 Visual Type

* Visualizing Second-Hand Car Purchases in Ireland
* The reason why every graph was chosen is to best describe the data presented. This applies to all graphs in the presentation. Every graph has a header, so it becomes clear when reviewing the presentation.
* Line graphs were chosen to showcase development over time, bar charts to showcase development over time in comparison to another metric and other characteristics of the dataset, stacked bar charts to give a good overview of how consumer preferences differ, as the most prominent part of the bar stands out, treemap to showcase relative data, supported by clear content support, e.g. USA flag behind USA datapoint, or color red standing for color red,
* Exploring University Class Dataset
* Reason for the correlation matrix (Python) was to first give an overview of the correlations in the dataset.
* Excel tables are shown in the report to showcase what data was used for the regression
* Excel correlation matrix is shown in the report to showcase the new correlations including the personality traits in the dataset
* PowerPoint Graphs that are pasted into the report are chosen to help with the storytelling in the report. These
* G20 Energy and Sustainability Report
* PowerBI visuals: The reason why every graph was chosen is to best describe the data presented. This applies to all graphs in the presentation. Every graph has a header, so it becomes clear when reviewing the presentation.
* Pie charts were used to compare average two categories, land cards to showcase location data, line graphs to show development, tables to show numbers, bar charts to showcase comparisons between more countries, the treemap diagram to show relative data, odometer to showcase percentage of goal reached, and scatterplot to show how average countries compare to the average.
* The reason for the PowerPoint visualization in the conclusion is to best depict the four focus points of the presentation
* Vehicle Price Predictions and Clustering
* The car clustering was shown using Python to give readers an overview of how a KMeans cluster can look and how the cars are distributed.
* Product Review Sentiment Analysis
* The visualization of the results from the Vader analysis was made to showcase how review scores correlate with the positive, negative, and neutral sentiments of the reviews. Bar chart with relative line was best suited to showcase this.

### 2.5.3 Report

Project: G20 Energy and Sustainability Report

The report is pasted into the PowerPoint Presentation as publishing it was not possible due to licensing reasons. As discussed in class, the PowerBI file is in the GitHub repository. The report focuses on various data points and tells a comprehensive story.

### 2.5.4 Report Interactivity

The report is highly interactive. Filters and data points can be selected via filters or the graphs themselves. The PowerPoint in which the report is, is embedded on the web page.

# 3. Additional Pages

## 3.1 Biography

295 words. I’ve focused on my academic journey and my career so far, also stating important values to me. Also, I’ve given a brief outlook of what I plan to achieve in my career.

## 3.2 Resume

The link for the Download is on the left side of the page with the contact info, not at the end of the biography due to design reasons. The sections in the resume are:

1. Header
2. Contact Info (on the right)
3. Objective
4. Education
5. Work Experience
6. Extracurricular Activities
7. Skills and Interests

## 3.3 Media Category

There are 3 media categories. There are 3 items included for every category:

1. Podcasts
2. Video
3. Website

## 3.4 Media Relevance

1. The podcasts dive into how Artificial Intelligence influences the industry and also showcase how companies and employees can be enabled by AI, namely Chat GPT. Also, one podcast gives a 30-minute business analyst course for beginners.
2. The websites give an introduction to the field of Business Analytics and what a typical profile of a data analyst looks like. Also, one website leads to a Master’s program in Business Analytics that I have applied for myself.
3. The videos give an outline on how one can become a data analyst and also educate the viewers in Python, going from visualization skills to a 6-hour course for an intermediate Python level.

All of the media is very much intertwined with business analytics.

## 3.5 Blogs

In total, there are 16 blog posts. Each of the posts is longer than 100 words.

I’ve structured the blogs into 5 parts:  
Firstly, four blogs delve into the topic of applications, since this is especially valuable for students such as myself at the moment.

Then, 4 posts give career guidance and blueprints for the first steps in data and business analytics.

Thirdly, I’ve picked three blog posts that cover skills in data analytics, such as machine learning skills and general data skills.

Fourthly, 4 posts deal with artificial intelligence, covering general information on the topic and moving along to prompt engineering and deep learning.

Lastly, one post delves into trends in business analytics.

## 3.6 Contact

The contact page is on my Resume, Biography, and More About Me pages of my portfolio since I think this makes more sense.

I’ve provided my email address, phone number, LinkedIn, Location, and also GitHub since I do not have X and believe GitHub is better fitting here than X.

## 3.7 Optional

The Page is called “More About Me” in my portfolio.

Section 1 is 205 words long and contains information on my volunteering work.

Section 2 is 207 words long and contains information on my hobbies.

I’ve added a representative picture for both sections to provide more depth.

Also, my contact information is also available on this page.