UK Price Paid and Energy Performance Study

Data Warehousing and Design Technique

Lappeenranta–Lahti University of Technology LUT

Master's Programme in Business Analytics, Engineering Science

March 2023

Author: Thanh Tran (000285359)

Nghia Nguyen (000275466)

Mohsin Qureshi (000341950)

1. Introduction

1.1. Background information

Our project focuses on defining a conception of data warehousing and designing an ETL (Extract-Transform-Load) service to practice the OLTP and OLAP processes on a business problem in the area of UK real estate's valuation in a consideration of energy impacts of the properties using open data provided by the UK government. Some useful insights can also be derived with business intelligence solution in order to resolve three strategic business questions as mentioned in the next section.

The objectives of the project are a well-designed data warehouse that is organized and highly facilitates the operations and maintenance, and an ETL solution that can be reused at anytime to automatically update data, or they can be set as a cron job to schedule the execution of data update without manual work.

The code base can be accessed from https://github.com/DarvinciVincent/BI_UK_realestates_pricepaid_energyperformance.

The concrete result of this project is a data set that is enriched with all data sources using the implemented ETL service and is up to date as the availability of data on the sources. The dataset has been uploaded to Kaggle network and that can be access via https://kaggle.com/datasets/b83009d6f5a7264b4c37614c60a29881da0024a7a301f948bda6a755431fbfb6 or with API command "kaggle datasets download -d tndt1902/uk-ppd-rre-epc-geo-details".

1.2. Solving problems

The business area of this project is real estates and there are 3 strategic questions that are defined as follows:

- Real estate valuation and market trends
- Energy performance of properties
- Correlation between real estates' value and market behaviours and the energy performance

1.3. Data Sources

Three official data sources and archives have been used for this project to ensure the reliability of the source data.

1.3.1. UK price paid

This data is a land registration data provided by the department for housing and communities with the coverage of UK and Wales. The data source provides information on all property sales in England and Wales that have either been sold for value or lodged for registration. The data is updated on the 20th working day of each month. The data can be accessed from the following link:

https://www.gov.uk/government/statistical-data-sets/price-paid-data-downloads

1.3.2. Energy performance certificate

This data source provides access to all energy performance certificates (EPCs) for building register in England and Wales. The data source is maintained by the department for Levelling Up, Housing and Communities (DLUHC) of UK. The data can be accessed from the following link:

https://epc.opendatacommunities.org/

1.3.3. Unique property_reference_number

This open data hub contains a frequently updated data on unique property reference numbers enriched with respective geometries in British National Grid and Latitude and Longitude. This data provides the opportunity to enrich the data visualization with map view or linking data to any further data set if the project has a chance to expand. The data can be accessed from the following link:

https://osdatahub.os.uk/downloads/open/OpenUPRN

2. OLTP processes & OLAP processes

For this project, we apply two concepts of OLAP (Online Analytical Processing) and OLTP (Online Transaction Processing) which are two different types of information processes used to process, update, and analyze data. The ETL service is a term that we use interchangeably to mention these two processes in the sense of practical solution. The solution is discussed in detailed in the next section.

To briefly reveal the term mentioned about the two processes, OLTP system is, in general, designed to handle real-time transactional processing, which involves recording and updating transactions as they occur. Particularly in this project, the OLTP system is optimized for data insertion, extraction, transformation, loading, deletion, and are used in applications bridge and enrich the backend interface.

The OLAP system is designed to support complex data analysis and reporting tasks according to interchangeable needs. The system is optimized for queries that require aggregation, consolidation, slicing and dicing of data.

The fundamental difference between OLTP and OLAP systems lies in their data structure and usage. OLTP system uses a normalized data structure, which is optimized for efficient data insertion, deletion, and modification. In contrast, OLAP system uses a denormalized data structure, which is optimized for efficient querying and reporting. This means that

OLTP systems are optimized for write-intensive workloads, while OLAP systems are optimized for read-intensive workloads.

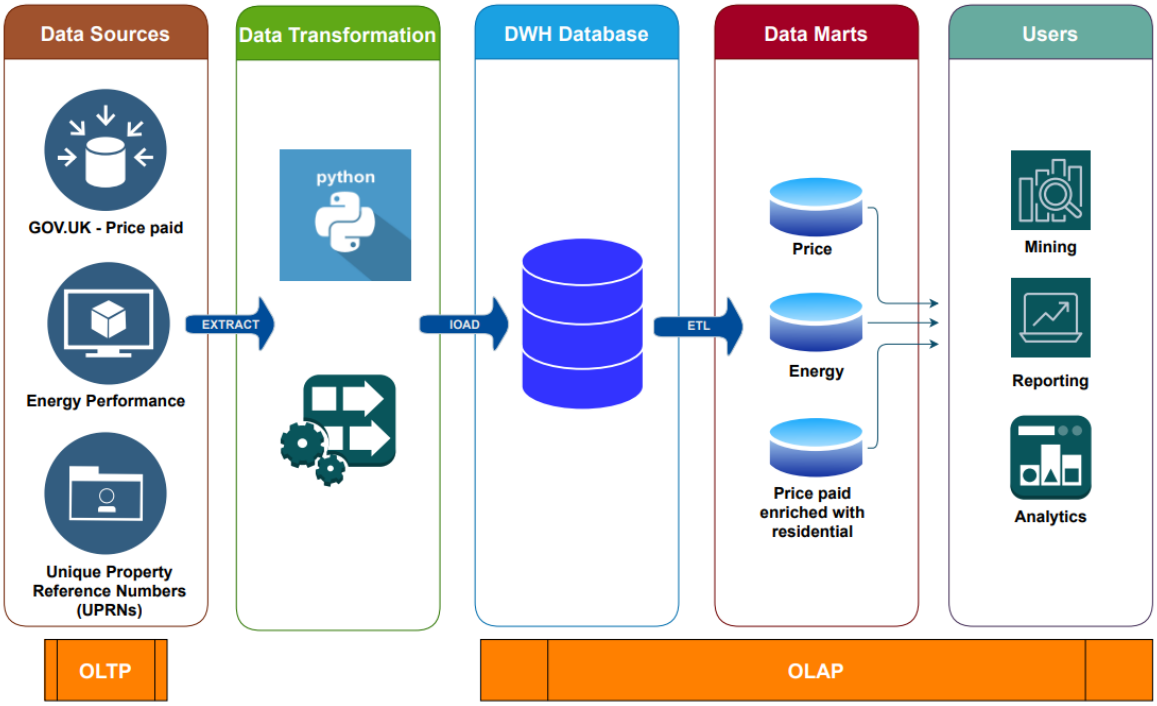


Figure 1. OLAP system

In this particular scenario, the datasets containing information related to price paid, energy performance, and unique property reference numbers were obtained from the relevant data source. To carry out the necessary data transformation tasks, a Pythonic and SQL solution has been implemented. The resultant transformed dataset was then loaded into a database, and based on the business strategies and analysis requirement of stakeholders in place, it would be partitioned into various data marts. Ultimately, these data marts are utilized for data mining, analytics, or decision-making, based on specific requirements and objectives.

This process involved retrieving and processing data from different sources, transforming it into a usable format, and then storing it in a centralized database. The utilization of Python programming language together with SQL facilitates the transformation of the data into the desired format. Once the data was organized and available in a centralized location, it could be partitioned into different data marts as per the business strategy. These data marts can then be utilized for conducting an analysis of various aspects related to real estate. Specifically, the analysis will focus on:

- Real estate value: The data marts will be utilized to conduct an analysis on the value of different types of properties, based on factors such as property type, category, county, and mechanical ventilation. This analysis will provide valuable insights into the real estate market, which can be utilized by various stakeholders including buyers, sellers, and real estate agents.
- Energy performance trends: The data marts will also be used to analyze the energy performance trends of different types of properties. This analysis will provide insights into the energy efficiency of different types of properties and help identify areas for improvement.
- Correlation between real estate value and energy performance: The data marts will be utilized to investigate the correlation between the energy performance rating of properties and their respective values. This analysis will provide insights into the relationship between energy efficiency and property values and can help identify opportunities for improving energy efficiency in the real estate market.

Overall, the use of data marts in this analysis will enable a detailed examination of the real estate market and its relationship with energy performance. The insights generated from this analysis can be utilized by various stakeholders to make informed decisions and improve the energy efficiency of the real estate market.

3. Data warehousing & ETL service

We provide solution for the ETL service in the open-source code which can be accessed via the link provided in the introduction part.

To design the solution for data warehousing and ETL service, we considered the data warehouse as having three data layers namely raw, transformed and analytics and data updating can be conducted automatically as data flow from an input to an output through a system of pipelines.

For each data source, we can have an extractor to extract the data and load it to the raw layer or section of the data warehouse. Then, any further transformation can be done to alter or enrich data according to specific needs. The transformed data can then be stored in the

transformed layer. Finally, upon requests, the data can then further be filtered and ordered, and then finally loaded to the analytics layer or in the other words, a data mart, to be ready for delivery or as backend interface of other services or applications. The figure aims at summarise and reflect the implementation of the technical solution in terms of data flow.

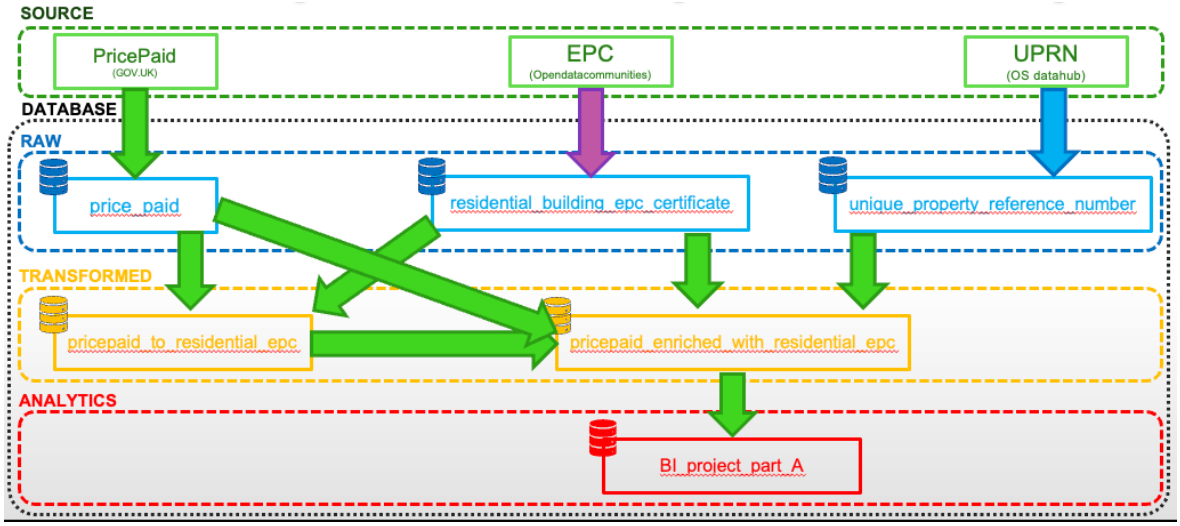


Figure 2 ETL service

The name shown in the data tables or data marts in the above figure are applied the same as in the source code provided.

We have implemented 3 pipeline systems as for the data sources namely UK_PPD (green arrows), UK_EPC (purple arrow) and UK_UPRN (blue arrow). There is an extractor in each of the pipeline system to extract raw data from source to the raw data layer and load it to the database as in the blue box. The UK_PPD pipeline system is the key pipeline in this project and we focus on enriching the price paid data with energy data. Therefore, there are multiple transforming step in that system. The second step of the UK_PPD pipeline system is to link price paid records with records extracted from EPC source. Therefore, a mapping table has been generated by altering and adjusting address variables in both price_paid and residential_building_epc_certificate table to get the linking keys. The mapping table pricepaid_to_residential_epc has only two fields representing one id field from the price_paid and one id field from the residential_building_epc_certificate. Obviously, a second transforming stage is to utilize the mapping table to map pricepaid records with epc records. Additionally in this step, we also link the resulting data with the unique_property_reference_number using the common field named “uprn” presenting in both EPC and UPRN. At this stage, I have achieved a full linked data of price paid enriched

with energy performance and geometry details. Upon requests from analytics part, we can then filter, adjust and load a partition of the full linked data to the analytics layer to make it available and ready for downstream operations.

4. Facts & ER models

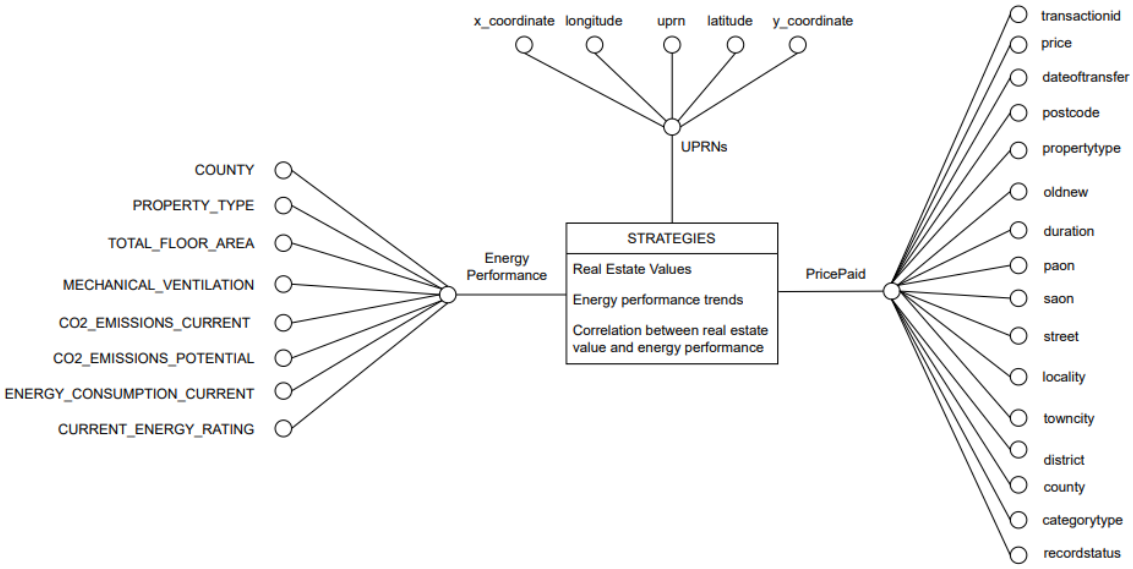


Figure 3.DFM

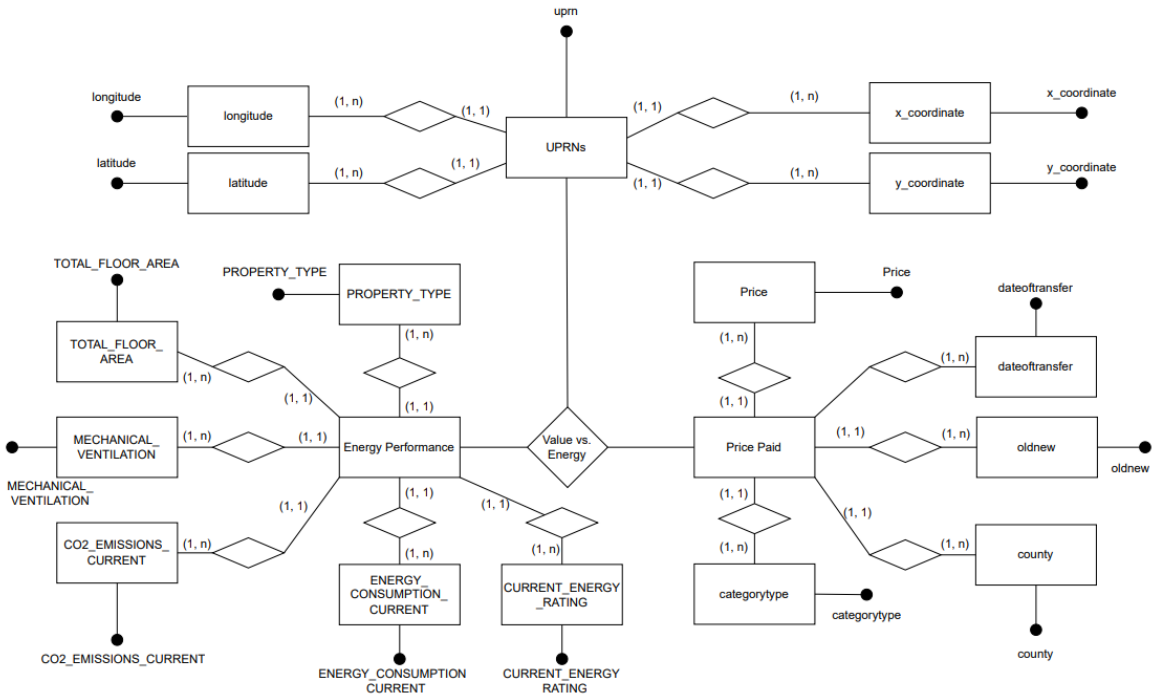


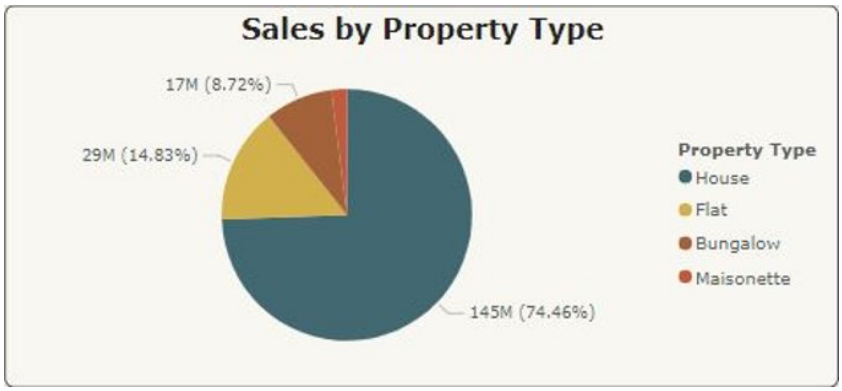
Figure 4. ER model

5. Business Intelligence

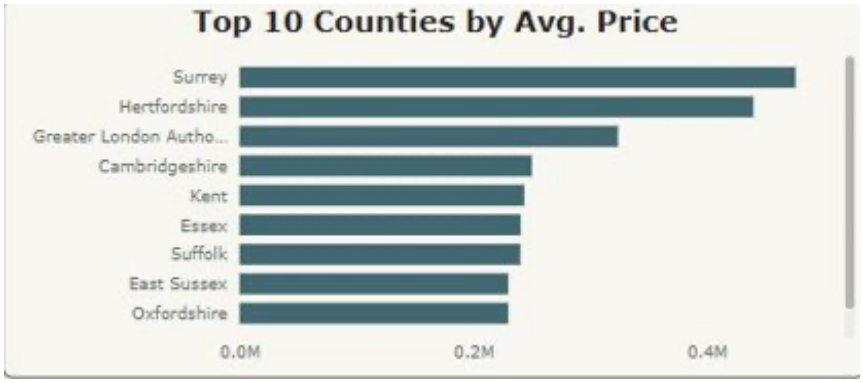
Q1- Real estate valuation and market trends:

In the business Intelligence part. We can comfortably answer some of the following potential business questions from the dashboard regarding real estate valuation and market trends.

-Most selling property type?

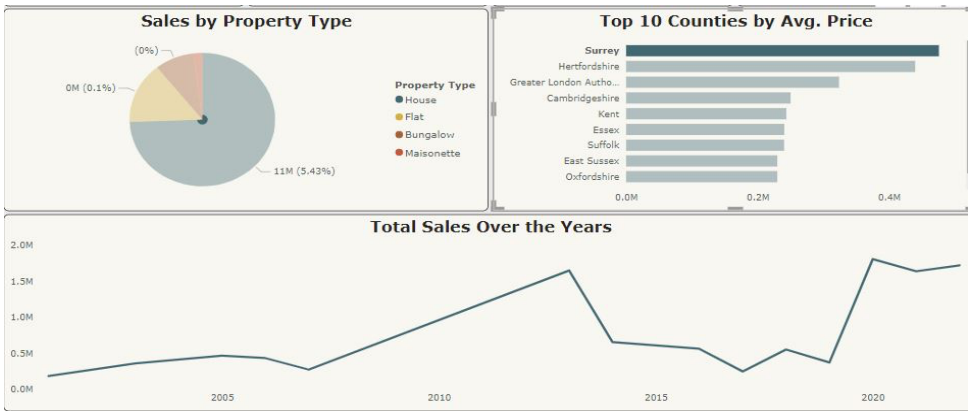


-What is the Average property price and floor area?



-Most Expensive and Least area to buy a property with respect to a county?

-Total sales over the years with respect to a county.



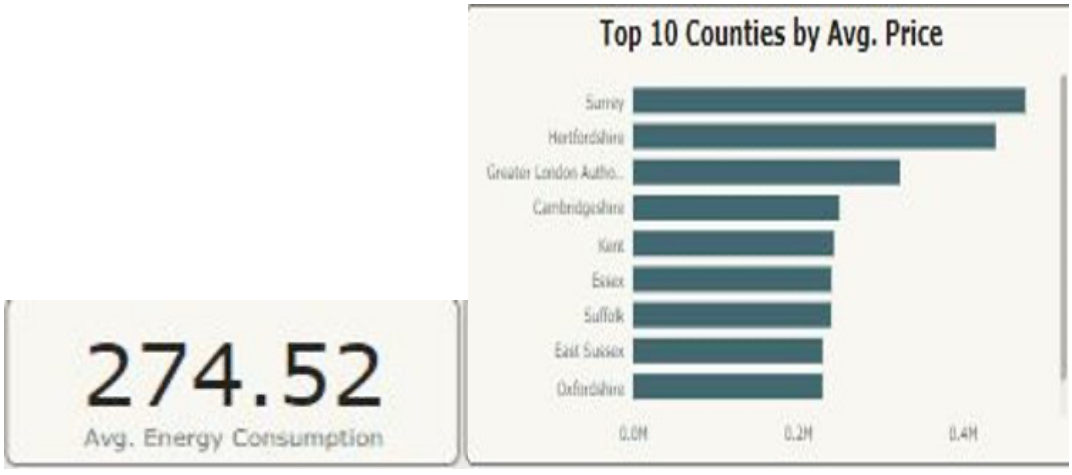
Q2-Energy performance of properties:

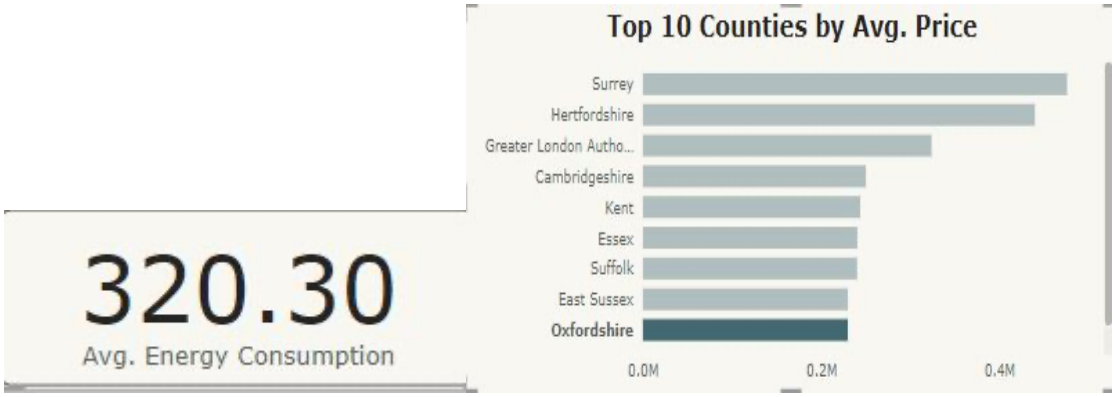
With the help of the dashboard. We can come up with the following potential business questions.

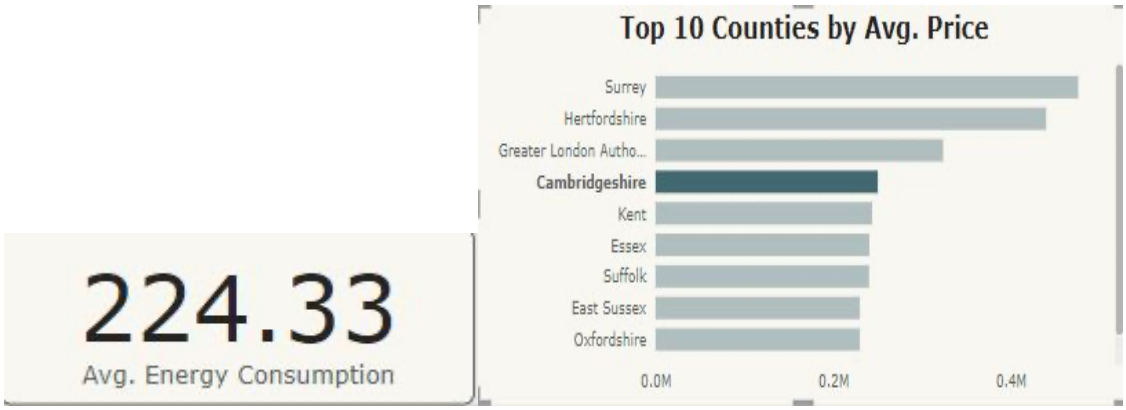
- What is the Average Energy consumption of all the counties.

-The county with the highest energy consumption

-The county with the Lowest energy consumption

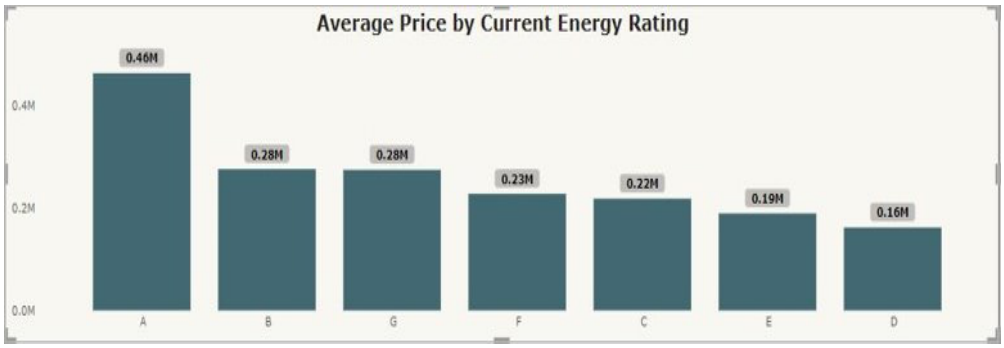




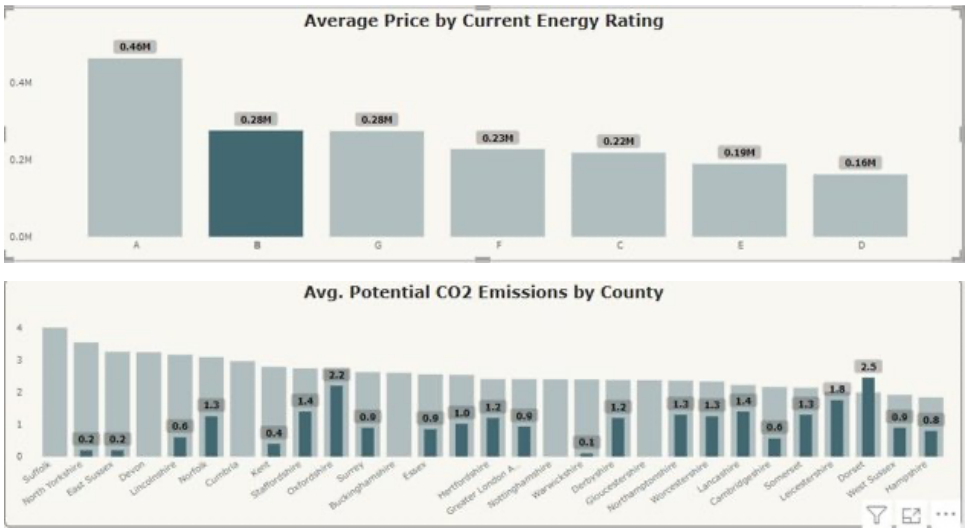


Q3-Correlation between real estate value and energy performance:

In this Part we can tell the energy performance of properties according to its real estate value.

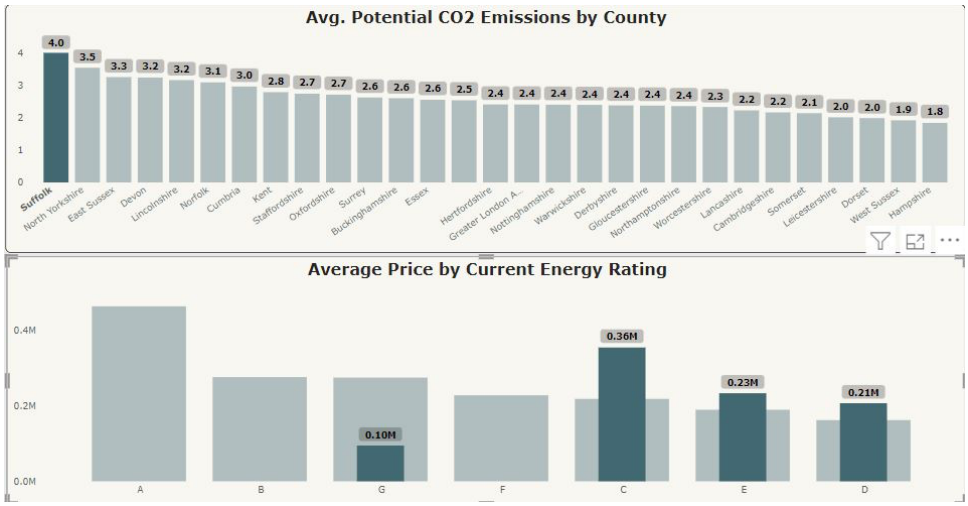


Such as

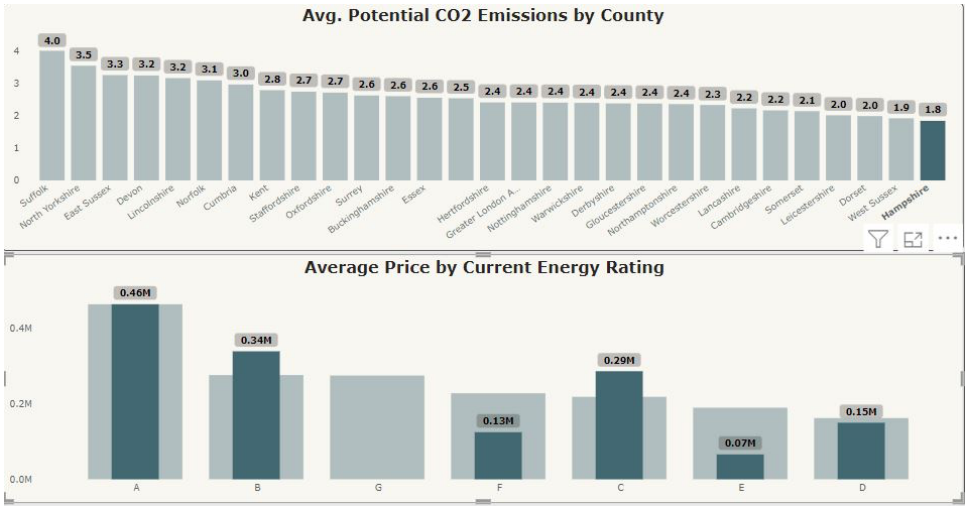


Another potential business question is to show the co-relation between average price by current energy rating with the highest and lowest co2 emissions by county.

Highest:



Lowest:



6. Conclusion

In conclusion, the design and the implementation of the technical solution works well for our defined problems, all the strategic questions can be answered with some extensions and the prospects of extra functionality that the model can offer. The design is comfortable and optimal for internal stakeholders (as in the team members) to collaborate. We consider this project as successful and the design of the technical solution is valid for solving this type of problem.