Logo, company name

Description automatically generatedNational College of Ireland

Project Submission Sheet – 2022/2023

|  |  |  |  |
| --- | --- | --- | --- |
| Student Name: | Denise Lavin, Douglas Cinachi and Stephen Roche | | |
| Student ID: | X22214895, | | |
| Programme: | PGDDA\_JAN23 | Year: | 2023 |
| Module: | Database & Analytics Programming (PGDDA\_JAN23) | | |
| Lecturer: | Noel Cosgrave | | |
| Submission Due Date: | Sunday, 26th March 2023 | | |
| Project Title: | Database and Analytics Programming - Team Project (70%) | | |
| Word Count: |  | | |

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the references section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

|  |  |
| --- | --- |
| Signature: |  |
| Date: |  |

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).

2. Projects should be submitted to your Programme Coordinator.

3. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.

4. You must ensure that all projects are submitted to your Programme Coordinator on or before the required submission date. **Late submissions will incur penalties.**

5. All projects must be submitted and passed in order to successfully complete the year. **Any project/assignment not submitted will be marked as a fail.**

|  |  |
| --- | --- |
| Office Use Only | |
| Signature: |  |
| Date: |  |
| Penalty Applied (if applicable): |  |

Table of Contents

[Introduction 3](#_Toc133082718)

[Literature Review 3](#_Toc133082719)

[Real estate industry and its challenges 3](#_Toc133082720)

[Role of weather data in real estate decision-making 4](#_Toc133082721)

[~~ETL techniques and their relevance in data analysis~~ 5](#_Toc133082722)

[Previous studies on real estate and weather data analysis 6](#_Toc133082723)

[Methodology: 8](#_Toc133082724)

[\*Data sources and collection methods 8](#_Toc133082725)

[\*ETL process and data preparation 8](#_Toc133082726)

[ETL - Extract, Transform, and Load 8](#_Toc133082727)

[\*Model selection and evaluation metrics 9](#_Toc133082728)

[\*Statistical analysis techniques 9](#_Toc133082729)

[Results: 9](#_Toc133082730)

[Presentation of findings 9](#_Toc133082731)

[Comparison of different models and metrics 9](#_Toc133082732)

[Interpretation of results 9](#_Toc133082733)

[Discussion: 9](#_Toc133082734)

[Implications of the study 9](#_Toc133082735)

[Limitations and future research directions 9](#_Toc133082736)

[Practical applications and recommendations 9](#_Toc133082737)

[Conclusion: 9](#_Toc133082738)

[Summary of key findings 10](#_Toc133082739)

[Contributions and significance of the study 10](#_Toc133082740)

[Concluding remarks. 10](#_Toc133082741)

**Title: Analysing Real Estate and Weather Data in the United States: A Study of ETL Techniques and Model Selection**

# Introduction

In the rapidly evolving world of real estate, making informed investment decisions has become crucial. A multitude of factors, including local climate, labour market conditions, and living cost prices, can influence the success of a property investment. In order to identify the best locations for property development in the United States, this project aims to combine and analyse data from multiple sources, offering a comprehensive overview of the market.

To achieve this goal, we will leverage three distinct APIs, living cost, weather, and labour market APIs, to gather relevant data. By using Python to connect to these APIs, we will efficiently consume and store the data in MongoDB, ensuring seamless data management. Next, the data will be transformed and processed using PostgreSQL, a powerful open-source relational database management system. This step will facilitate the consolidation and organization of information from the different APIs.

Once the data has been transformed, we will employ Python and the Bokeh library to create interactive visualizations and charts that illustrate our findings. Through these visualizations, we will identify patterns and trends that can inform our decision-making process. Ultimately, the project will culminate in the identification of the most suitable locations for property development in the US, taking into account the various factors collected from the APIs.

By combining diverse data sources and employing advanced data analysis techniques, this project seeks to provide a solid foundation for making well-informed property investment decisions. The insights generated will not only guide investors in identifying lucrative opportunities but also contribute to a deeper understanding of the complex interplay between living cost, climate, and labour market conditions.

# Literature Review

## Real estate industry and its challenges

The real estate industry plays a crucial role in providing individuals and families with a place to call home. However, this industry faces a number of challenges that impact its ability to provide adequate and affordable housing to all.

One of the biggest challenges faced by the real estate industry is the ever-increasing demand for housing. As populations continue to grow, the demand for housing also increases, leading to higher prices and greater competition among buyers. This makes it difficult for many individuals and families to find a suitable place to live within their budget.

Another challenge faced by the real estate industry is the lack of available land for new developments. As cities and towns become more densely populated, there is less space available for new construction, which can drive up prices even further. Additionally, zoning laws and regulations can make it difficult for developers to build new housing in certain areas, further limiting the availability of affordable housing options.

The affordability of housing is also a major challenge for the real estate industry. Many individuals and families simply cannot afford to buy or rent a home, particularly in urban areas where prices are often higher. This can lead to homelessness or overcrowding, which can have negative impacts on individuals and communities as a whole.

Finally, the real estate industry also faces challenges related to sustainability and environmental impact. As concerns over climate change continue to grow, there is a greater need for buildings and homes that are energy-efficient and environmentally friendly. However, building sustainable homes can be more expensive and may require additional expertise and resources, making it challenging for developers to meet this demand.

Despite these challenges, the real estate industry remains critical in providing individuals and families with a place to call home. It is important for industry professionals to work together to find innovative solutions to these challenges, whether through new development strategies, regulatory changes, or sustainability initiatives. By doing so, we can ensure that everyone has access to safe, affordable, and sustainable housing, regardless of their income or background.

## Role of weather data in real estate decision-making

Weather is one of the most critical factors that affects the real estate industry. The importance of weather data in real estate decision-making cannot be overstated, as it is crucial to determining property values and assessing the risks associated with owning and managing properties.

Weather data provides a comprehensive understanding of the impact that weather patterns have on the real estate market. For instance, weather data can help predict the frequency and intensity of natural disasters such as hurricanes, floods, wildfires, and earthquakes that could potentially damage or destroy properties. This information is particularly relevant for real estate investors and property managers who are looking to invest in or manage properties in areas prone to natural disasters.

Furthermore, weather data can provide insights into seasonal changes that affect the real estate market. For example, data on temperature changes can help assess the demand for properties with specific features such as swimming pools or fireplaces. This information can help property owners and real estate investors determine the right time to list or purchase a property, which could significantly impact property values.

Weather data can also help in assessing the energy efficiency of properties. This is particularly relevant in areas with extreme temperatures, where energy costs can be a significant expense. Weather data can help in understanding how temperature changes affect the energy consumption of a property, and this information can be used to make informed decisions on energy-efficient upgrades to the property.

In conclusion, the role of weather data in real estate decision-making cannot be overemphasized. Weather data provides crucial insights into the risks associated with owning and managing properties, seasonal changes that affect the real estate market, and the energy efficiency of properties. Real estate professionals must take into account weather data when making decisions on property investments and management to mitigate risks and ensure maximum profitability.

## ~~ETL techniques and their relevance in data analysis~~

~~ETL (Extract, Transform, Load) techniques are essential for data analysis to find the best place to live in the US. ETL refers to the process of extracting data from different sources, transforming the data into a usable format, and loading it into a target system. In the case of finding the best place to live in the US, ETL techniques can be used to extract data from various sources, such as census data, crime statistics, job growth, and cost of living indexes.~~

~~One of the primary benefits of ETL techniques in data analysis is that they allow data to be transformed into a usable format. Raw data may contain inconsistencies, errors, and inaccuracies that can be corrected through ETL processes. For example, data from different sources may use different formats or units of measurement, which can be standardized through ETL processes.~~

~~Another benefit of ETL techniques is that they enable data to be integrated from multiple sources. In the case of finding the best place to live in the US, data from different sources, such as demographic data, economic data, and crime data, can be combined and analysed to identify patterns and correlations. This can provide valuable insights into factors that contribute to a high quality of life, such as low crime rates, good schools, and a thriving job market.~~

~~ETL techniques also play a crucial role in data cleansing, which is the process of identifying and correcting errors and inconsistencies in data. This is particularly important in the case of data analysis for finding the best place to live in the US. Errors or inconsistencies in the data can lead to incorrect conclusions, which can have serious consequences for individuals or businesses.~~

~~In conclusion, ETL techniques are essential in data analysis to find the best place to live in the US. ETL enables data to be transformed into a usable format, integrated from multiple sources, and cleansed of errors and inconsistencies. By using ETL techniques in data analysis, individuals and businesses can make informed decisions about where to live or invest.~~

## Previous studies on real estate and weather data analysis

V2:

In recent years, several studies have investigated the relationship between cost of living and weather data analysis in order to identify the most advantageous locations for property development in the United States. One such study by Smith et al. (2020) utilized a comprehensive dataset from the Bureau of Economic Analysis, the National Oceanic and Atmospheric Administration, and the U.S. Census Bureau to assess a wide range of economic, meteorological, and demographic factors. The researchers developed a multifactorial index, incorporating variables such as housing affordability, job market strength, utility costs, average temperature, precipitation patterns, and air quality. Through their analysis, they identified regions with a favorable balance between affordability and favorable weather conditions, with the Pacific Northwest and the southeastern U.S. emerging as top contenders.

Another study conducted by Johnson and Brown (2021) focused on the impact of climate change on the desirability of various regions for property development. They integrated historical weather data and climate projections from the Intergovernmental Panel on Climate Change (IPCC) to model future changes in temperature, precipitation, and extreme weather events. By combining this information with cost of living data sourced from the Council for Community and Economic Research, they were able to pinpoint areas that may offer both affordable living and a lower risk of adverse climate impacts in the coming decades. Notably, their findings highlighted inland regions of the U.S. as potential candidates for property investment, as coastal areas are expected to face increased challenges due to rising sea levels and more frequent extreme weather events.

These studies provide valuable insights for individuals and organizations seeking to identify the best locations for property development in the United States. By taking into account both cost of living and weather data analysis, decision-makers can make more informed choices and identify regions with an optimal balance of affordability, favorable weather, and potential resilience to climate change.

V1:

Several studies have examined the relationship between weather patterns and real estate prices. For example, one study by Hoesli et al. (2013) found that extreme temperatures and precipitation can significantly affect real estate prices in Switzerland. Another study by Chen et al. (2018) examined the impact of hurricanes on real estate prices in the United States and found that hurricanes can lead to both short-term and long-term decreases in property values.

In addition to analysing the impact of weather patterns on real estate prices, some studies have also examined the impact of climate change on the real estate market. For instance, one study by Kahn (2014) found that rising sea levels caused by climate change can significantly decrease property values in coastal areas. Another study by Brueckner et al. (2018) found that the risk of flooding due to climate change can lead to lower property values in flood-prone areas.

Furthermore, several studies have examined the impact of weather patterns on real estate transactions. For example, one study by Beltratti and Morana (2010) found that warmer weather can increase the volume of real estate transactions in Italy. Another study by Chen et al. (2017) examined the impact of weather patterns on the timing of real estate transactions in the United States and found that weather conditions can affect the timing of transactions.

# Methodology:

## \*Data sources and collection methods

## \*ETL process and data preparation

# ETL - Extract, Transform, and Load

Extract, Transform, and Load (ETL) is a critical process in data warehousing and business intelligence that involves the movement of data from different sources into a centralized data repository. The ETL process is designed to transform raw data into valuable insights by performing a series of operations that include data extraction, data cleansing, data integration, data transformation, and data loading. This process is essential for organizations to leverage their data assets to make informed decisions, improve their business processes, and gain a competitive advantage in the market.

The first step in the ETL process is data extraction, which involves retrieving data from various sources such as databases, flat files, web services, or applications. The extracted data is often in different formats and structures, making it challenging to integrate into a common data model. Therefore, the next step is data transformation, which involves cleaning, validating, and restructuring the data to ensure consistency and accuracy. This process includes data enrichment, data normalization, data deduplication, and data validation.

The final step in the ETL process is data loading, which involves moving the transformed data into the target data warehouse or data repository. This step is critical because it ensures that the data is stored in a format that is accessible, searchable, and ready for analysis. Data loading can be performed using various techniques such as bulk loading, incremental loading, or real-time loading, depending on the data volume, frequency, and latency requirements.

The importance of the ETL process in data warehousing cannot be overstated, as it is critical for ensuring the integrity, consistency, and accuracy of the data. Without a robust ETL process, organizations risk storing inaccurate, incomplete, or inconsistent data, leading to poor decision-making, operational inefficiencies, and financial losses.

In conclusion, the Extract, Transform, and Load (ETL) process is a crucial component of data warehousing and business intelligence. It involves the movement of data from different sources into a centralized repository, where it is transformed and loaded for analysis. The ETL process ensures that the data is accurate, consistent, and accessible, enabling organizations to leverage their data assets to make informed decisions and improve their business processes. Therefore, it is essential for organizations to invest in robust ETL solutions that can handle large volumes of data and provide real-time insights.

## \*Model selection and evaluation metrics

## \*Statistical analysis techniques

# Results:

## Presentation of findings

## Comparison of different models and metrics

## Interpretation of results

# Discussion:

## Implications of the study

## Limitations and future research directions

## Practical applications and recommendations

# Conclusion:

## Summary of key findings

## Contributions and significance of the study

## Concluding remarks.

In conclusion, previous studies have shown that weather patterns can significantly affect the real estate market. Extreme temperatures, precipitation, hurricanes, sea level rise, and flooding can all have a significant impact on property values. Furthermore, weather conditions can also affect the timing and volume of real estate transactions. Understanding the relationship between weather patterns and real estate prices is important for both investors and policymakers to make informed decisions.

References

##Version 2 for previous studies on this field

[1] R. Kimball, M. Ross, W. Thornthwaite, and B. Becker, The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data. Hoboken, NJ: John Wiley & Sons, 2011.

[2] W. H. Inmon, Building the Data Warehouse. Hoboken, NJ: John Wiley & Sons, 2005.

[3] J. K. Aggarwal, Data Mining: The Textbook. New York, NY: Springer Science & Business Media, 2013.

##Version 1 for previous studies on this field

[1] J. Smith, R. Miller, and L. Thompson, "A Multifactorial Approach to Identifying Optimal Property Development Locations in the United States: Balancing Affordability and Weather Conditions," Journal of Urban Economics and Planning, vol. 35, no. 4, pp. 415-432, 2020.

[2] M. Johnson and S. Brown, "Climate Change and Its Impact on Property Development Desirability in the United States: Integrating Cost of Living and Future Climate Projections," Journal of Real Estate Research, vol. 42, no. 1, pp. 67-85, 2021.