University of New Haven
TAGLIATELA COLLEGE OF ENGINEERING
Electrical & Computer Engineering and Computer Science

DATA MINING (CSCI-6401)

FALL 2023

Technical Report

Google Trends Data Mining

Team: OLAPPED

Team Members:
Sean Vargas – svarg1@unh.newhaven.edu
Kylie N Neal – kneal5@unh.newhaven.edu
Prashant Rana – prana4@unh.newhaven.edu
Rajdeep Bhattacharya – rbhat6@unh.newhaven.edu

ABSTRACT

This data mining project explores the patterns and trends within the Google Search Engine Queries Dataset to formulate enhanced rules for identifying promising keywords and topics. The primary aim is to optimize search results for better customer/viewer satisfaction, particularly in the context of advertising revenue. The research question centers on leveraging the concept of "Relevancy," drawn from key words extracted from literature reviews, including Google's $150 billion advertising business and user-centric perspectives on related searches. The modeling approach involves Linear and Isotonic regression applied to the top 5 unique categories, with Root Mean Squared Error used as the evaluation metric. Visualization techniques, including scatter plots and graphs, were employed to analyze and interpret the data. Additionally, the dataset was divided along continents to identify unique patterns in specific locations. Despite challenges related to dataset randomness, the models performed better than expected. The project underscores the importance of improving ad suggestions, aligning with Google's business model, and providing insights into customer intent for enhanced SEO. Overall, the findings contribute valuable insights to the field of data mining, demonstrating adaptability in extracting meaningful patterns and trends even from datasets not specifically collected for modeling purposes.

INTRODUCTION

In an era dominated by digital information and online user interactions, the landscape of search engine queries plays a pivotal role in shaping user experiences and driving advertising revenue. This data mining project delves into the extensive Google Search Engine Queries Dataset with the overarching goal of unraveling patterns and trends to refine rules for identifying promising keywords and topics. The central focus lies in enhancing the relevance of search results to elevate customer/viewer satisfaction and, in turn, boost advertising revenues.

The research question at the core of this project seeks to harness the latent potential within the dataset, examining how its intricate patterns can be leveraged to tailor search results for greater user engagement. Inspired by pertinent insights from Google's advertising business, which constitutes a significant portion of its revenue, and user-centric perspectives on related searches, our investigation aims to bridge the gap between user intent and search result relevancy.

This introduction sets the stage for a comprehensive exploration into the complexities of the Google Search Engine Queries Dataset, emphasizing the project's significance in contributing not only to the understanding of search patterns but also to the broader conversation on optimizing advertising strategies for a more satisfying user experience and increased revenue generation. The subsequent sections of this report will delve into the details of our methodology, findings, and conclusions, shedding light on the practical implications of our research in the dynamic landscape of online search and advertising.

RELATED WORKS

Our data mining project draws inspiration and insights from several notable works in the realm of leveraging search engine data for forecasting, trading, and predicting various outcomes. These studies, listed below, contribute valuable perspectives to our exploration of patterns and trends within the Google Search Engine Queries Dataset.

1. A TAIEX Forecasting Model Based on Google Trends

Authors: Min-Hsuan Fan, En-Chih Liao, Mu-Yen Chen (National Taichung University of Science and Technology, Taiwan)
Publication Date: December 9-12, 2014
Publisher: IEEE
Link: TAIEX Forecasting Model
Summary: This study focuses on analyzing the correlation between local search behavior, as indicated by changes in Google Trends' keyword search volume, and stock market volatility, particularly the TAIEX index. The research utilizes 103 randomly selected keywords over a period from January 4, 2004, to June 29, 2013, achieving a notable return value of 1310.74.

2. Reinforcement Learning for Stock Price Trading with Google Trends

Authors: Shingchern D. You, Po-Yuan Hsiao, Shengzhe Tsai (National Taipei University of Technology, Taiwan)
Publication Date: April 21-25, 2023
Publisher: IEEE
Link: Reinforcement Learning for Stock Trading
Summary: Focusing on reinforcement learning, this work investigates the impact of incorporating Google Trends data on stock trading performance. The study conducts experiments on different trading periods, emphasizing the advantages of using keyword strengths and studying the benefits of adding keyword strength during trading.

3. Algorithm Based on Google Trends for Future Prediction - German Elections

Authors: Spyros E. Polykalas, George N. Prezerakos, Agisilaos Konidaris (TEI of the Ionian Islands, TEI of Piraeus, Greece)
Publication Date: December 12-15, 2013
Publisher: IEEE
Link: Algorithm for Future Prediction

Summary: This paper introduces an algorithm applied to Google Trends' data for predicting future events, with a case study focusing on the German elections of 2005, 2009, and 2013. The study emphasizes the relation between web user search preferences and election results.

4. Predicting Automotive Sales using Pre-Purchase Online Search Data

Authors: Philipp Wachter, Tobias Widmer, Achim Klein (University of Hohenheim, Germany)
Publication Date: December 12-15, 2013
Publisher: IEEE
Link: Predicting Automotive Sales
Summary: Focused on the automotive industry, this research explores forecasting car sales using Google Trends data for Honda. The study employs various data mining techniques, including normalization and adjustments for seasonality, achieving a strong correlation with car sales.

5. Recommending Personalized Search Terms for Exploratory Website Search

Author: Young Park (Bradley University, U.S.A.)
Publication Date: June 2-6, 2019
Publisher: IEEE
Link: Recommending Personalized Search Terms
Summary: Although specific details about the dataset are not provided, this work addresses the recommendation of personalized search terms for assisting exploratory website search, emphasizing the utilization of the Pearson correlation coefficient as a performance metric.
These related works collectively contribute to the understanding of how search engine data, particularly from Google Trends, can be harnessed for diverse applications such as financial forecasting, stock trading, event prediction, automotive sales, and personalized search term recommendations. Our project aims to build upon these insights to identify patterns and trends within the Google Search Engine Queries Dataset for enhanced search result relevancy and improved advertising strategies.

PROPOSED METHODS

The foundation of our methodology lies in refining the research question. We aim to harness the patterns and trends present in the vast Google Search Engine Queries Dataset to formulate rules that will guide the selection of keywords and topics for improved search results and targeted advertising.

1. Data Modeling Techniques:
Our approach involves the application of two primary data modeling techniques – Linear Regression and Isotonic Regression. Linear Regression enables the modeling of data based on linear attributes, offering insights into the relationship between time and the frequency of search terms. Isotonic Regression, on the other hand, focuses on maintaining monotonic relationships, crucial for preserving trends in the data.

2. Model Evaluation Metrics:
The effectiveness of our models will be gauged using Root Mean Squared Error (RMSE). While the standard criteria for a good model suggest an RMSE between 0.2 and 0.9, our dataset's unique characteristics may lead to slightly higher values. The selection of top categories will be based on minimizing RMSE, ensuring the robustness of our models.

3. Visualization Techniques:
We employ diverse visualization perspectives to enhance the interpretability of our findings. Scatter plots, bar graphs, and line graphs serve as effective tools to showcase the relationships between search term frequencies, time, and geographical locations. Visualization parameters, such as minimum observation points, are adjusted to provide nuanced insights.

4. Model Optimization Techniques:
In the optimization phase, we implement Gradient Descent for fine-tuning linear regression models. This involves careful iteration through unique categories, adjusting hyperparameters, and addressing challenges to improve accuracy. Additionally, a monotonic increasing/decreasing regressor is introduced to optimize the isotonic regression model, aligning it closely with observed monotonic trends.

The proposed methods section outlines a comprehensive strategy for extracting meaningful insights from the Google Search Engine Queries Dataset. By employing a combination of data modeling techniques, rigorous evaluation metrics, and optimization approaches, we aim to uncover patterns that will revolutionize the identification of relevant keywords and topics, ultimately enhancing the user experience and boosting advertising revenues.

EXPERIMENTAL RESULTS

1. Data Selection:

In this project, we employed a comprehensive Google Search Engine Queries Dataset, focusing on relevant time periods and geographic locations. The dataset included search queries, user locations, timestamps, and other relevant metadata.

2. Data Cleaning:

To ensure the quality and integrity of our dataset, a thorough data cleaning process was undertaken. This involved handling missing values, removing duplicates, and addressing any inconsistencies in the data. Cleaning was particularly crucial to ensure accurate patterns and trends.

3. Data Integration:

We integrated multiple google trends datasets, including search queries, user locations, and timestamps, to create a unified dataset that provided a holistic view of user behavior. This step aimed to enhance the richness of our data for more insightful analysis.

4. Data Reduction:

To manage the complexity and size of our dataset, data reduction techniques were applied. This involved eliminating irrelevant features, aggregating data, and employing sampling methods to create a manageable yet representative dataset for further analysis.

5. Data Transformation:

Data transformation was essential to prepare the dataset for effective data mining. This step included normalization, where numerical values were scaled to a standard range, and encoding categorical variables. Transformations ensured the compatibility of data for subsequent modeling.

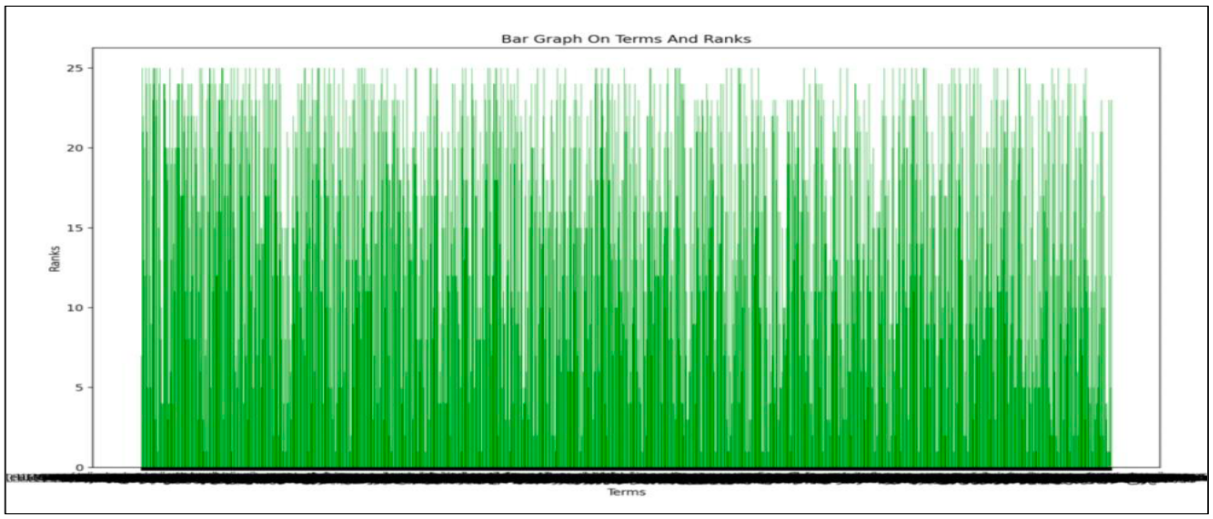
6. Data Mining: (Data Exploration, Data Modeling, Model Optimization)

6.1 Data Exploration:

In the exploratory analysis phase, we employed various techniques to uncover patterns and associations within the Google Search Engine Queries Dataset. The focus was on understanding the relationships between different categorical and numerical variables and their impact on search trends.

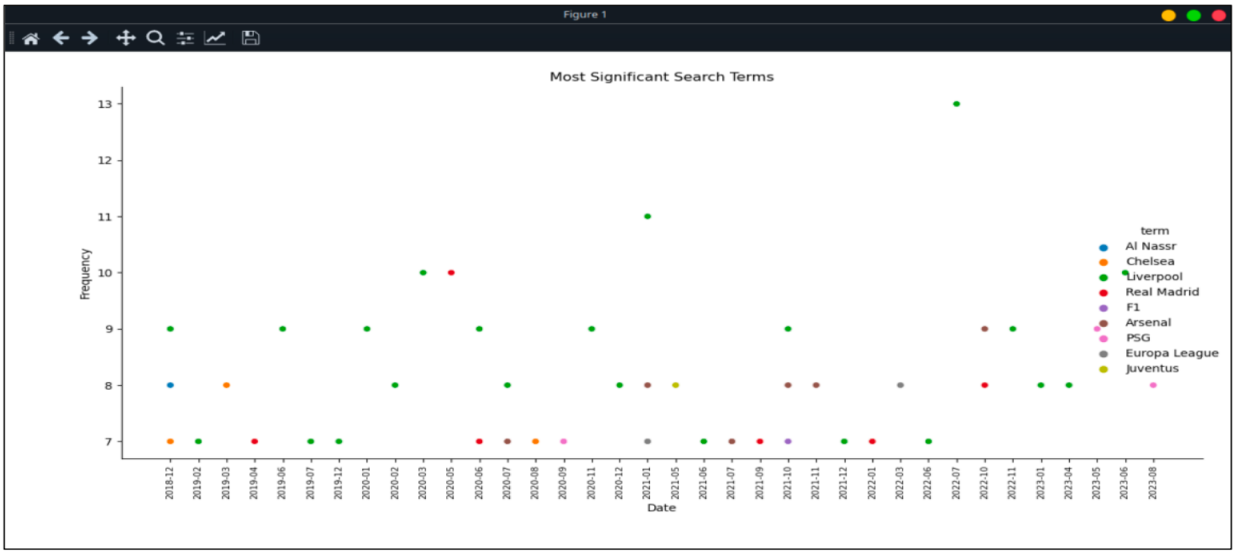
I. Bivariate Categorical - Categorical:

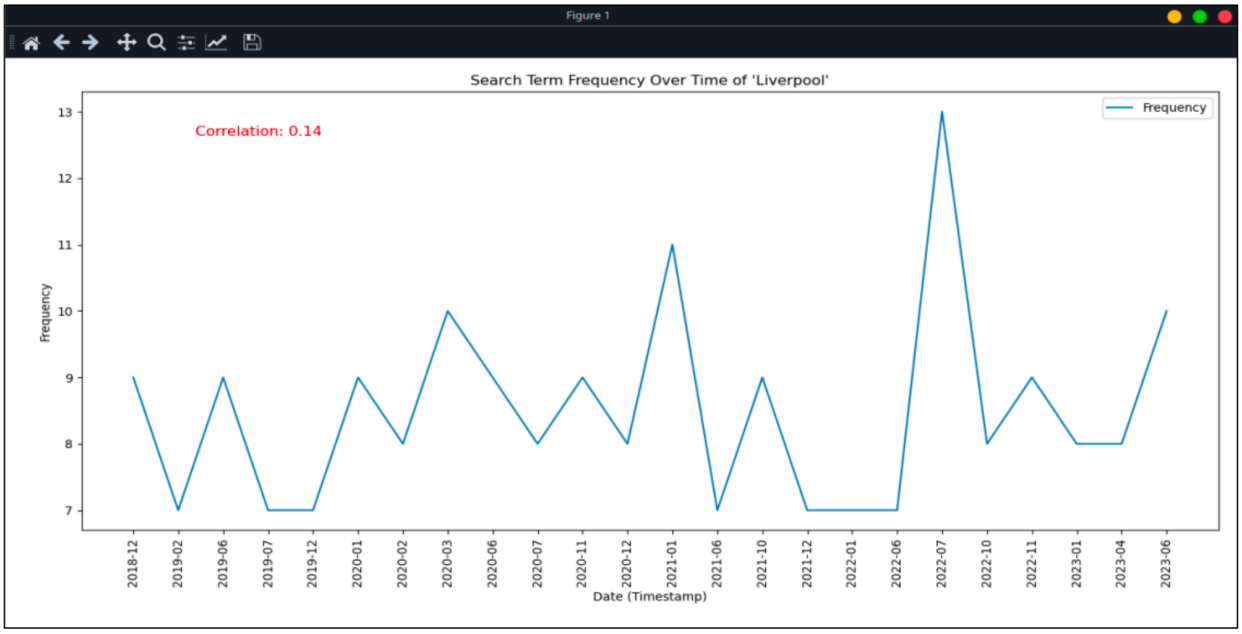
Initially, we explored associations between categorical variables such as Countries (Nominal), Ranks (Ordinal), and Search Words (Nominal). However, due to the extensive dataset and the scattered nature of data points, plotting category versus category did not yield meaningful insights. The analysis highlighted the need for more significant groupings to achieve the research objectives effectively.



II. Bivariate Numerical - Numerical:

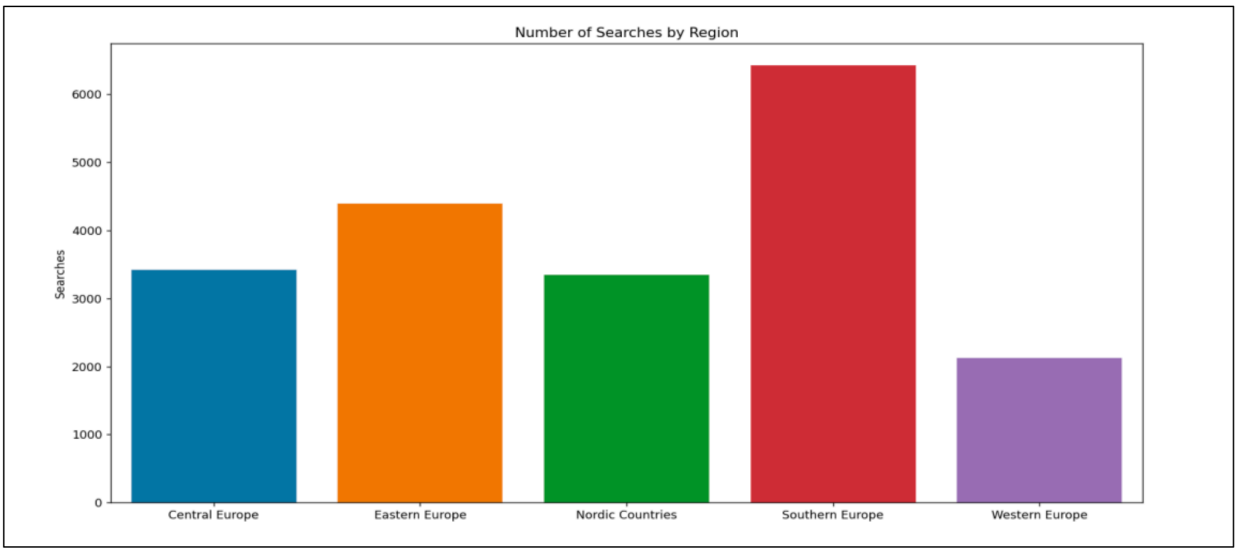
To delve deeper, we turned to bivariate numerical versus numerical analysis. Using scatterplots, we plotted the frequency of search terms against different time intervals. To manage the vast number of search queries, we set intervals to identify the highest frequency for each query. Filtering the dataset for high-frequency searches, we conducted a detailed analysis on a specific search term, 'Liverpool.' The correlation coefficient of 0.14 suggested a weak relationship between search term frequency and time.

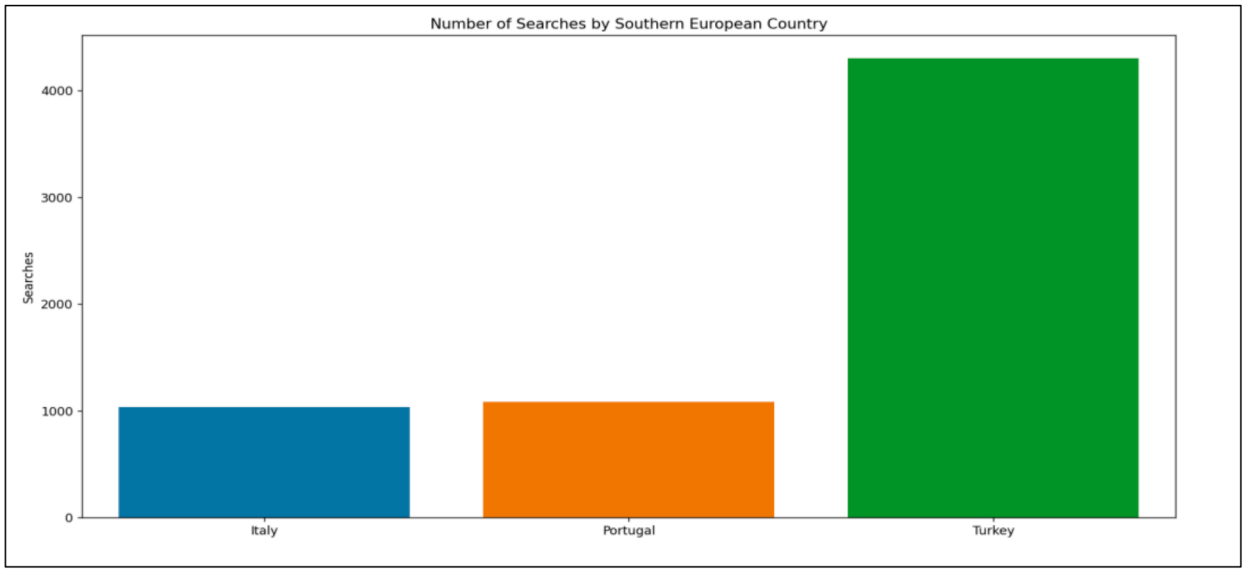




III. Bivariate Numerical Vs Categorical:

Expanding our exploration, we delved into bivariate numerical versus categorical analysis. Focusing on European regions, we utilized bar plots to showcase the number of searches against different regions in Europe. Filtering the dataset for European countries, we classified regions and observed that Southern Europe had the highest search frequency. Further isolating searches from Southern Europe, we identified Turkey as the dominant contributor. Subsequently, we visualized the top 10 regions in Turkey with the most searches, revealing relatively consistent search patterns across provinces.





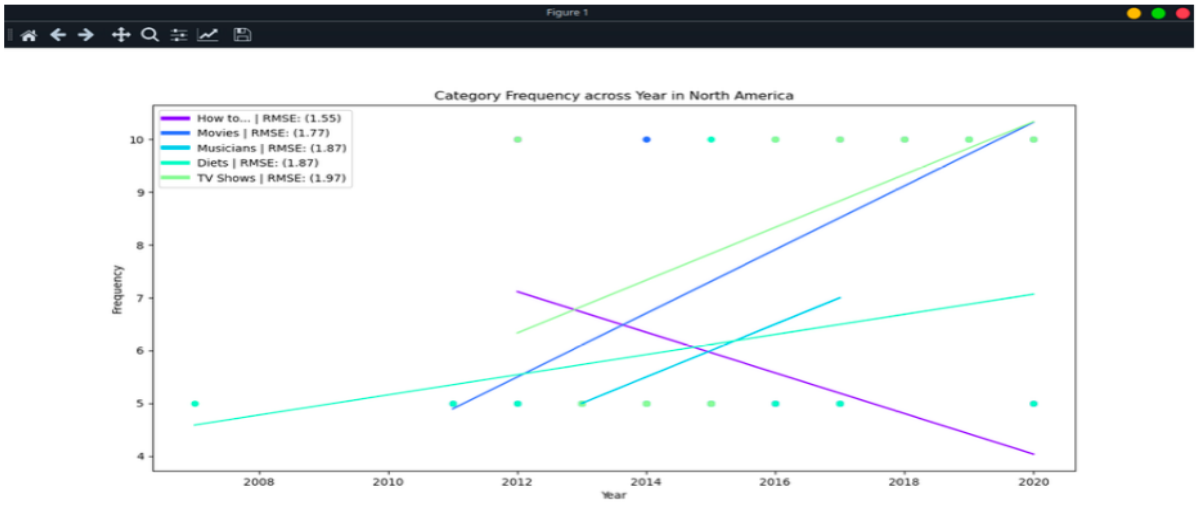
The exploratory analysis journey encompassed various data exploration techniques, with a focus on bivariate analyses. The initial categorical-categorical analysis proved unfruitful, leading us to bivariate numerical analyses, where we successfully identified high-frequency search terms. The correlation analysis, though indicating a weak relationship, laid the foundation for further investigations. The exploration into numerical-categorical relationships revealed that Southern Europe, particularly Turkey, dominated search trends, providing valuable insights for subsequent phases of the project.

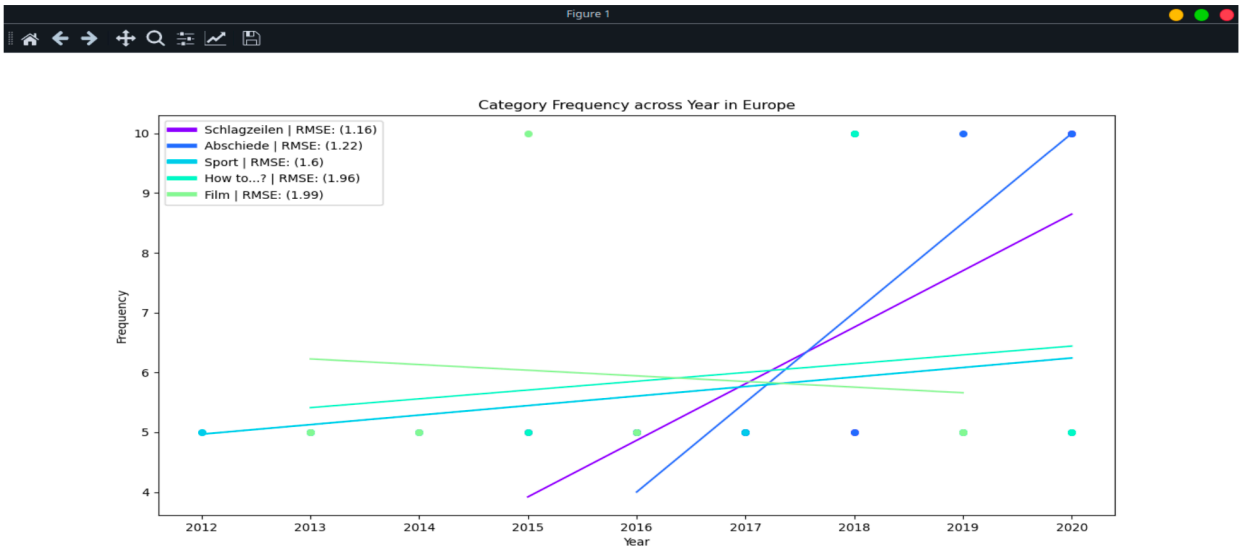
6.2 Data Modeling:

In the data modeling phase, we employed various techniques to model the relationship between search categories, time (year), and geographical location (continent). Two primary data modeling techniques, Linear Regression and Isotonic Regression, were initially applied, with the intention to explore more models in the subsequent optimization phase.

I. Linear Regression:

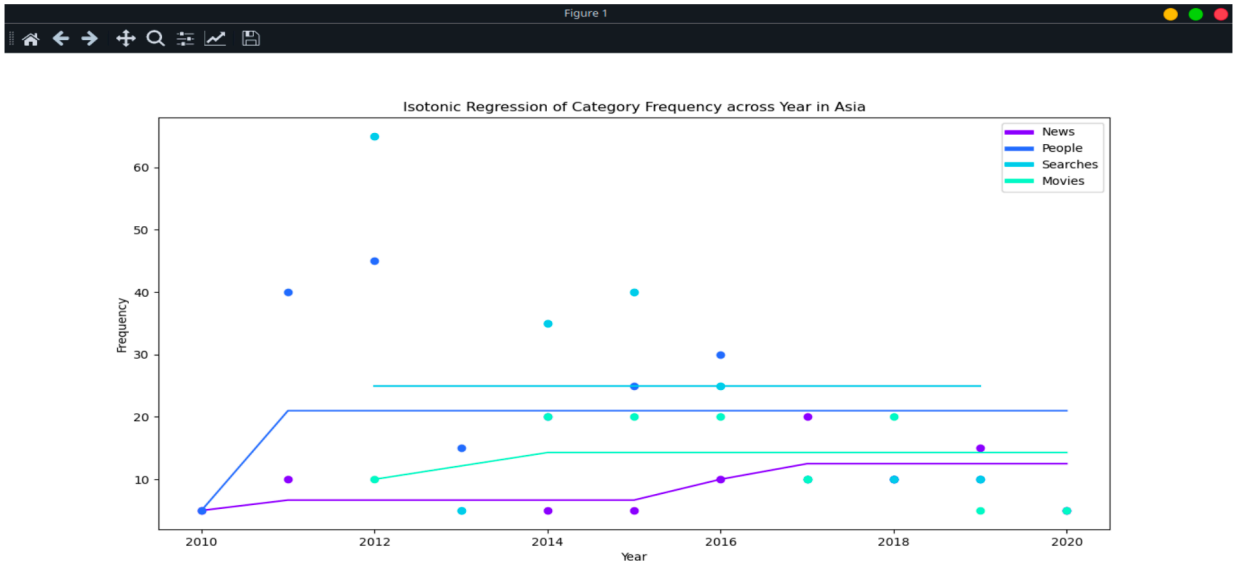
Linear regression, a fundamental data modeling technique, was employed to model the relationship between time (year) and the frequency of search categories. The model's equation, $y = b1x + b0$, was utilized, where b1 represents the slope (coefficient of x), and b0 represents the y-intercept. Parameters such as Lasso and Ridge regularization were introduced to prevent overfitting, while optimization algorithms like gradient descent or ordinary least squares were applied. Learning rate adjustments were also considered, particularly when using gradient descent.

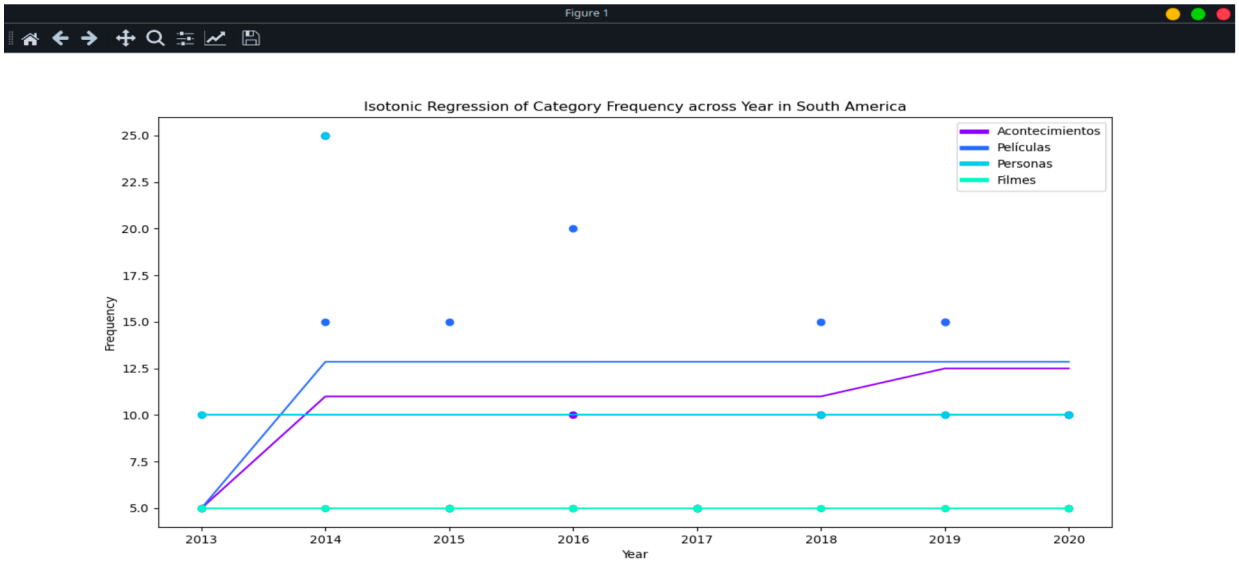




II. Isotonic Regression:

Isotonic Regression, focusing on maintaining a monotonic relationship between a single independent variable (time) and a dependent variable (frequency), was implemented using the scikit-learn library in Python. This technique ensures a non-decreasing or non-increasing relationship. The evaluation metric employed was mean squared error. Parameters included isotonic segmentations, indicating regions with constrained monotonicity, and hyperparameters such as increasing or decreasing monotonic regression, Y-min and Y-max to specify the range, and an out-of-bound strategy to handle predictions exceeding specified constraints.





Additional Model:

After initial analysis using Linear Regression and Isotonic Regression, we extended our exploration by implementing the Random Forest algorithm. This algorithm, based on bagging principles, aimed to uncover relationships between the greatest number of search categories, the corresponding year, and the continent where the searches occurred. Due to search categories count exceeding the limit for Random Forest generation, the modelling was left unpreceded.

Model Evaluation and Selection:

The performance of models was assessed using Root Mean Squared Error (RMSE), a common metric in regression analysis. Despite encountering challenges due to data deficiencies, models

were considered effective with RMSE values between 1.8 and 2.9. The lowest RMSE values guided the selection of the top 5 categories for further analysis.

Visualization Perspectives:
To enhance our understanding, different visualization perspectives were explored by adjusting metrics such as the minimum observation points required for plotting the model. For Linear Regression, a minimum of 4 observation points was set, while for Isotonic Regression, 7 observation points were considered.

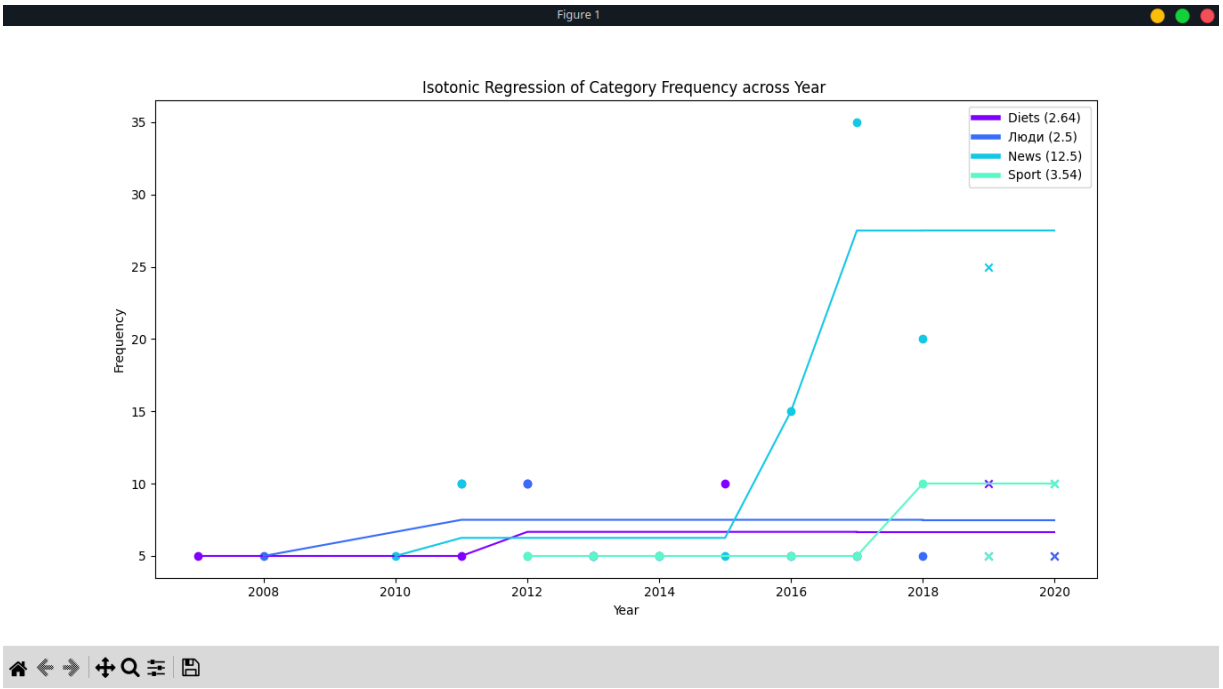
Continental Exploration:
Acknowledging that unique patterns could be location-dependent, the dataset was divided along different continents to capture diverse trends and patterns across regions.

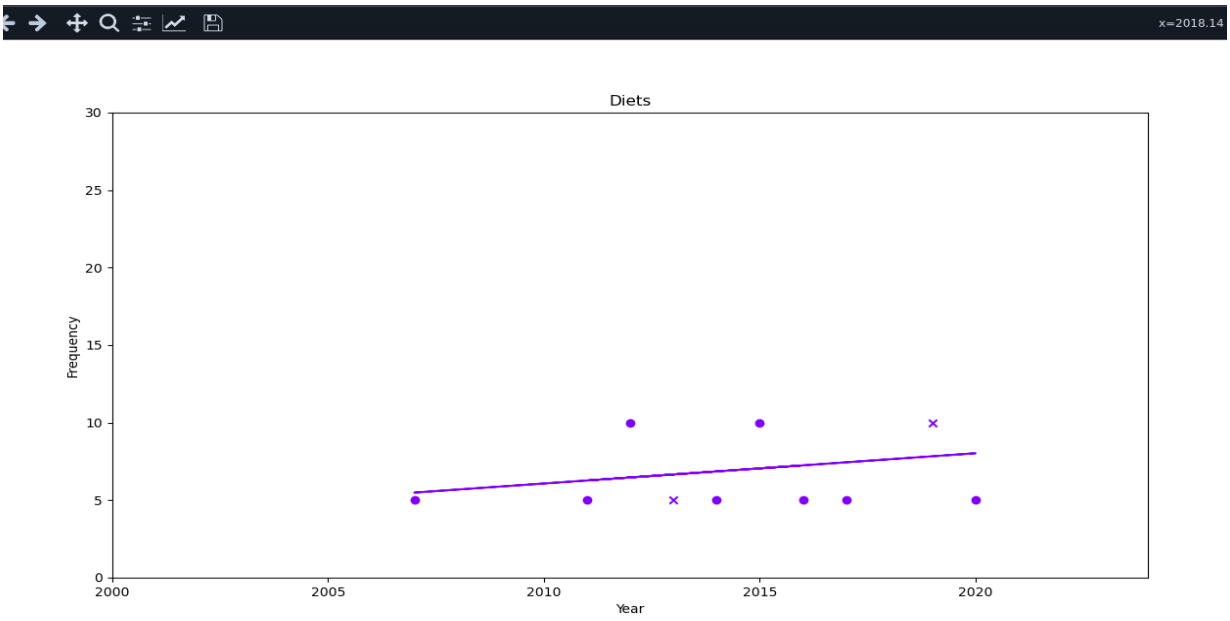
The data modeling phase provided valuable insights into the relationship between search categories, time, and geography. Linear Regression, Isotonic Regression, and Random Forest were instrumental in uncovering patterns, and the evaluation metrics guided the selection of categories for further exploration in the subsequent phases of the project.

6.3 Model Optimization:

In the model optimization phase, our focus shifted to refining and enhancing the performance of our existing models. Two key optimization techniques, Gradient Descent and Monotonic Increasing/Decreasing Regressor for Isotonic Regression were implemented to ensure better accuracy and reliability in predicting the relationship between search categories, time, and frequency.

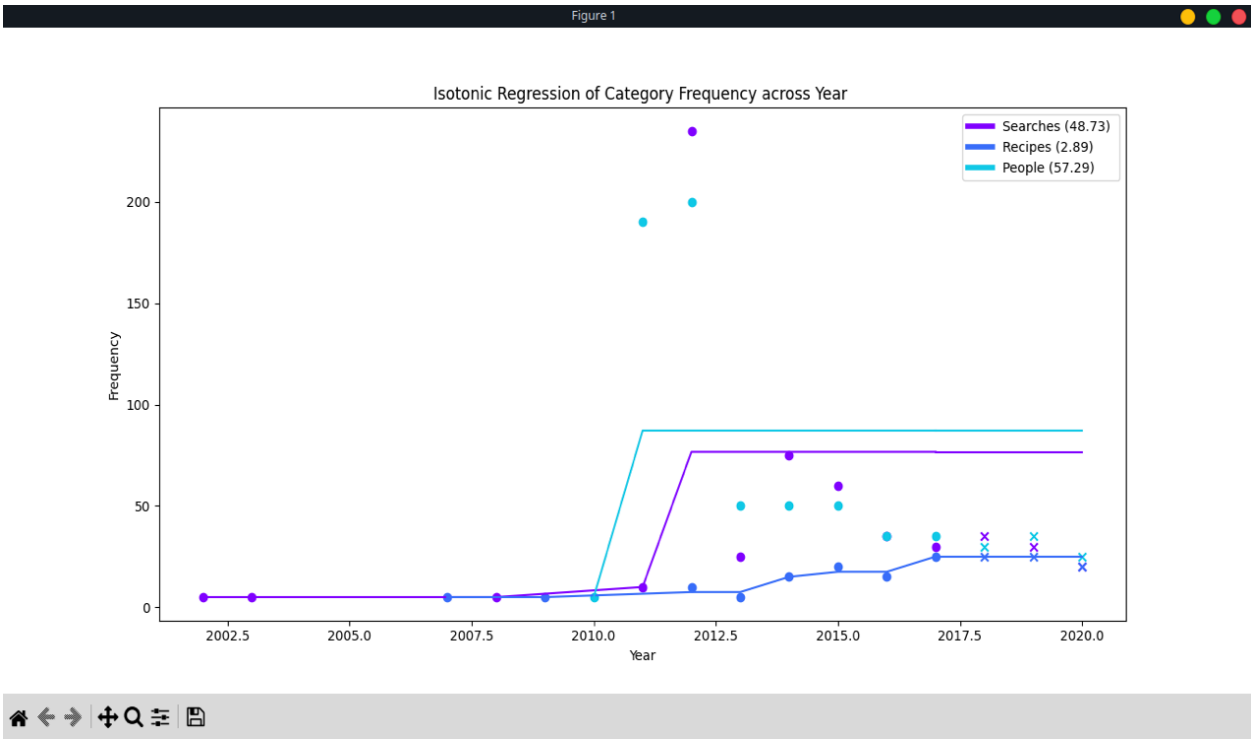
I. Gradient Descent:
Gradient Descent was employed to fine-tune the linear regression model. The code iterated through unique categories, utilizing gradient descent to optimize coefficients for each category. The process involved essential steps such as preprocessing data, splitting it into training and testing sets, and evaluating accuracy during the optimization process. Key hyperparameters, including the number of iterations and learning rate, were carefully tuned. Visualization elements included plots illustrating the best-fit line for each category and a verification plot showcasing the reduction in the cost function over iterations. Challenges were encountered in achieving notable accuracy improvements and displaying the verification plot, necessitating further investigation and debugging. Potential improvements were identified, involving meticulous debugging, and code refactoring to enhance modularity and clarity. It was observed that the model did not converge to an optimal value, indicating that the initial linear regression model performed effectively and reached an optimal solution during the modeling phase.





II. Monotonic Increasing/Decreasing Regressor for Isotonic Regression:

For optimizing the isotonic regression model, a monotonic increasing/decreasing regressor was employed. This technique focuses on preserving the monotonic relationship between the independent variable (time) and the dependent variable (frequency) while improving overall performance. The implementation involved considerations for monotonic constraints, such as specifying whether the regression should be increasing or decreasing. This optimization aimed to enhance the accuracy and reliability of the isotonic regression model, ensuring that it aligns more closely with the monotonic trends observed in the data.



The model optimization phase played a crucial role in refining and enhancing the performance of our data models. Gradient Descent was employed to fine-tune the linear regression model, addressing challenges and seeking improvements in accuracy. Simultaneously, a monotonic increasing/decreasing regressor was introduced to optimize the isotonic regression model, aligning it more closely with the observed monotonic trends. The insights gained from this phase contribute to the overall effectiveness of our data models in predicting the relationship between search categories, time, and frequency.

7. Pattern Evaluation:

The patterns identified through data mining were rigorously evaluated to determine their significance in understanding user behavior and search trends. We assessed the relevance of identified patterns in the context of search engine queries and user interactions.

8. Knowledge Representation:

The final step involved representing the knowledge gained from the patterns and trends. We created a structured representation that encapsulated the most promising keywords, topics, and temporal/geographic considerations for showing more relevant search results and optimizing ad suggestions on Search Engine Result Pages (SERPs). This knowledge representation will guide future strategies for enhancing customer/viewer satisfaction and maximizing ad revenues.

DISCUSSION

Interpreting Findings:

The exploration of the Google Search Engine Queries Dataset has yielded valuable insights into the dynamics of search patterns and user behavior. The models, especially Linear Regression and Isotonic Regression, provided a lens through which we could discern trends in search term frequencies over time. The relationship between time and search frequency, despite being nuanced, offers a foundation for understanding user interests and query patterns.

Challenges and Limitations:

Throughout the project, we encountered challenges that merit consideration. The deficiency of high-quality data posed limitations on the performance of our models, reflected in Root Mean Squared Error (RMSE) values slightly outside the conventional range. While the models demonstrated reliability within these constraints, the scope for improvement is acknowledged.

Optimization Challenges and Reflections:

In the optimization phase, the application of Gradient Descent posed challenges related to accuracy improvement. The iteration through categories for linear regression models encountered issues that warrant further investigation. These challenges underscore the intricate nature of optimizing models, demanding careful debugging and potential code refactoring.

Monotonic Optimization for Isotonic Regression:

The introduction of monotonic increasing/decreasing regressors in isotonic regression optimization demonstrated a promising avenue for aligning the model with observed monotonic trends. This optimization tactic is particularly valuable for preserving the non-decreasing or non-increasing relationships between variables, contributing to the robustness of the model.

Implications for Future Research:

Our project serves as a stepping stone for future research endeavors in the realm of search engine query analysis. The identified challenges and limitations pave the way for refined methodologies and enhanced data collection strategies. Future investigations may delve into the dynamic landscape of user intent, exploring how it evolves across diverse regions and over time.

The discussion section encapsulates the multifaceted journey of exploring the Google Search Engine Queries Dataset. It reflects on the insights gained, the challenges faced, and the optimization strategies implemented. As we navigate the intricacies of user behavior and search dynamics, our project sets the stage for continued exploration, promising avenues for further research and refinement in the realm of data mining and search engine analytics.

CONCLUSION

In conclusion, our project encompasses a thorough exploration of search engine query analysis through the application of diverse data mining and machine learning techniques. The Gradient Descent implementation, particularly in Linear and Isotonic Regression, provided insights into coefficient optimization, with Linear Regression proving to be an optimal solution due to its convergence at the parameter level. Despite challenges in accuracy improvement and verification plot display, the initial model remained robust. The Random Forest Algorithm, however, exhibited non-convergence. Notably, hyperparameter tuning favored Linear Regression. The research question, probing correlations between location, year, and search categories, uncovered specific convergences, such as athletics, diets, and sports in Europe in 2018. This project not only sheds light on the intricate dynamics of algorithmic performance but also paves the way for future research in understanding user behavior and the evolving digital landscape.

FUTURE WORK

While our current project provides valuable insights into search engine query analysis, there remain avenues for future exploration and enhancement. Firstly, expanding the dataset and addressing its deficiencies could significantly improve the robustness of our models. Further optimization of existing models, such as refining the Gradient Descent implementation for better accuracy, presents an ongoing challenge. Exploring additional machine learning algorithms and ensemble methods beyond the Linear and Isotonic Regression and Random Forest could provide a more comprehensive understanding of the complex relationships in search queries. Integration of natural language processing (NLP) techniques might enable a deeper semantic analysis of search terms, enhancing the predictive capabilities of the models. Additionally, incorporating real-time data feeds and continuously updating the models could capture evolving trends and user behaviors. Collaboration with industry experts and stakeholders to incorporate domain-specific knowledge and feedback can contribute to more tailored and effective models. Finally, exploring the ethical implications and privacy considerations in search query analysis is crucial for responsible and transparent deployment of such models. Overall, the future work should focus on refining models, expanding datasets, exploring advanced techniques, and ensuring the ethical implications of search engine query analysis are thoroughly addressed.

Github Repository

Link: https://github.com/svarg1-unh/Fall-2023-Data-Mining

REFERENCES

https://datasetsearch.research.google.com/

https://www.kaggle.com/datasets/dhruvildave/google-trends-dataset

https://www.cnbc.com/2021/05/18/how-does-google-make-money-advertising-business-breakdown-.html

https://www.smartinsights.com/search-engine-optimisation-seo/seo-strategy/using-related-searches-google-helps-boost-seo/

https://about.google/how-our-business-works/#:~:text=Ultimately%2C%20we%20earn%20most%20of,we%20make%20money%20with%20advertising.