Jurnal Teknik Informatika C.I.T Medicom, xx (x) (20xx) xx-xx

Published by: Institute of Computer Science (IOCS)

Jurnal Teknik Informatika C.I.T Medicom

Journal homepage: www.medikom.iocspublisher.org

Linear Regression Analysis of Weekly Per-Capita Consumption Based on Food Groups in Indonesia using PySpark.

Muhammad Farhan Alfarizi[1]
[1)] Teknik Informatika, Pelita Bangsa University, West Java, Indonesia

Article Info

Article history

Received : diisi oleh editor
Revised : diisi oleh editor
Accepted : diisi oleh editor

Kata Kunci:
Weekly Per Capita;
Consumption;
PySpark;
Linear Regression;
Food Categories;
Economic Trends;
Dietary Habits;
Indonesia;

Abstract

The dynamics of weekly per capita food consumption are paramount in understanding a nation's dietary habits and economic trends. This study, conducted in the context of Indonesia's diverse archipelago, leverages PySpark for a comprehensive Linear Regression Analysis of weekly per capita consumption across various food categories from 2018 to 2022. The aim is to unveil population preferences and economic behaviors, offering insights for policymakers and industry stakeholders. Utilizing PySpark's big data processing capabilities, the research presents a visual narrative of predicted food consumption trends, depicting changes in preferences and expenditures across diverse food groups. The analysis extends to average per capita expenditures, providing economic insights into different food consumption types. Identification of the top three food categories with the highest average per capita expenditure serves to pinpoint significant contributors to consumer spending.The study's significance lies in unraveling the complexities of consumption patterns, contributing valuable insights for policymakers, economists, and the food industry. By exploring temporal trends and economic aspects associated with food consumption, this research bridges the gap between dietary habits and economic behavior in Indonesia. This abstract adheres to the specified format, providing a concise summary within the prescribed word limit.

Corresponding Author:

Muhammad Farhan Alfarizi,
Teknik Informatika
Pelita Bangsa University
Jl. Inspeksi Kalimalang Tegal Danas, Bekasi, West Java, Indonesia
Farhan.19@mhs.pelitabangsa.ac.id

1. Introduction

The dynamics of weekly per capita food consumption play a pivotal role in understanding the dietary habits and economic trends within a nation. In the context of Indonesia, a diverse archipelago with a rich culinary heritage, exploring the intricacies of weekly per capita food consumption allows for a comprehensive analysis of the population's preferences and economic behavior. This study focuses on employing PySpark[1], [2], a powerful big data processing framework[3], to conduct a Linear Regression Analysis on weekly per capita consumption across different food categories[4], [5], [6]. The period

under consideration spans from 2018 to 2022, enabling a robust examination of trends and variations in food consumption patterns over time.

This research endeavors to present a visual narrative of predicted food consumption trends, illustrating the changes in preferences and expenditures across various food groups over the specified period[7], [8]. Additionally, the study delves into the average per capita expenditure based on food categories, offering insights into the economic aspects associated with different types of food consumption.

Furthermore, the identification of the top three food categories based on the highest average per capita expenditure aims to pinpoint significant contributors to consumer spending in the Indonesian context. By unraveling the complexities of these consumption patterns, this research seeks to contribute valuable insights for policymakers, economists, and stakeholders in the food industry[9].

2. Research Methode

The data preprocessing phase involves handling missing values, addressing outliers, and standardizing variables to ensure the quality and consistency of the dataset. PySpark's distributed computing capabilities will be leveraged to efficiently process large volumes of data, facilitating scalability and performance[10], [11], [12]. The training and evaluation of the Linear Regression model will involve a split-sample validation approach[13], with a portion of the dataset reserved for model training and the remainder for validation[14]. The model's performance will be assessed using relevant metrics such as Mean Squared Error (MSE)[15], [16], [17] and R-squared[18].

Additionally, the study will employ descriptive statistics to provide an overview of the central tendencies and distributions within the dataset[19]. Subgroup analyses will be conducted to explore potential variations in consumption patterns across demographic factors. The results will be interpreted in the context of socio-economic indicators and external factors that may influence food consumption trends[20].

The research methodology is designed to provide a rigorous and comprehensive analysis of weekly per capita food consumption trends in Indonesia, offering valuable insights for policymakers, researchers, and stakeholders in the food industry[21], [22], [23].

3. Result and Discussion

The Linear Regression Analysis revealed insightful patterns and trends in weekly per capita food consumption across different food categories in Indonesia[24]. The visualization of predicted consumption trends over the years depicted notable fluctuations in preferences and expenditure patterns[25],[26].

a. Data Preprocessing and Feature Engineering

The PySpark script begins with essential data preprocessing steps, addressing missing values represented by '-'. These values are replaced with 'o' to facilitate subsequent numerical operations. The 'jenis_makanan' (food type) and 'nama_wilayah' (region name) columns are then indexed for categorical encoding, preparing them for inclusion in the regression model. Features are combined into a vector using PySpark's VectorAssembler, allowing for the creation of a unified input for the regression model.

```
# Handling nilai kosong atau '-'
df = df.replace('-', 'o', 'value')

# Konversi kolom 'value' ke tipe data Float
df = df.withColumn("value", df["value"].cast("float"))
```

```
        # Indeksasi kolom kategori 'jenis_makanan' dan 'nama_wilayah'
        indexer1 = StringIndexer(inputCol="jenis_makanan", outputCol="jenis_makanan_index")
        indexer2 = StringIndexer(inputCol="nama_wilayah", outputCol="nama_wilayah_index")
        df = indexer1.fit(df).transform(df)
        df = indexer2.fit(df).transform(df)


        # Gabungkan fitur-fitur ke dalam vektor
        assembler = VectorAssembler(inputCols=["tahun", "jenis_makanan_index",
        "nama_wilayah_index"], outputCol="features")
df = assembler.transform(df)
```

Figure 1. : Data Preprocessing and Feature Engineering Overview

b. Model Training and Evaluation

The Linear Regression model is initialized and trained using the preprocessed data. The dataset is split into training and testing sets, with 80% used for training and 20% for testing. The model is evaluated on the test data, and the Root Mean Squared Error (RMSE) is calculated as a metric to assess its performance.

```
predictions = model.transform(test_data).select("value", "prediction", "tahun", "jenis_makanan",
"nama_wilayah")
```

Figure 2. : Model Training and Evaluation

```
Root Mean Squared Error (RMSE) on test data = 733.624776605572
```

Figure 3. : Running Model Training and Evaluation

c. Prediksi dan Visualisas

The trained model is used to make predictions on the test data. The results, including actual values, predicted values, and associated information such as 'tahun' (year), 'jenis_makanan' (food type), and 'nama_wilayah' (region name), are displayed. Two visualizations are presented: a scatter plot comparing actual values to predictions and a line chart illustrating the predicted consumption trends over the years for different food types.

```
# Tampilkan visualisasi hasil prediksi
predictions = model.transform(test_data).select("value", "prediction", "tahun", "jenis_makanan",
"nama_wilayah")
predictions.show(10)

# Visualisasi hasil prediksi
predictions_pd = predictions.toPandas()
plt.scatter(predictions_pd["value"], predictions_pd["prediction"])
plt.xlabel("Nilai Sebenarnya")
plt.ylabel("Prediksi")
plt.title("Nilai Sebenarnya vs Prediksi")
plt.show()

# Visualisasi hasil prediksi menggunakan Line Chart
fig_line_chart = px.line(predictions.toPandas(), x="tahun", y="prediction", color="jenis_makanan",
                 labels={"prediction": "Prediksi", "tahun": "Tahun", "jenis_makanan": "Jenis
Makanan"},
```

```
                title="Prediksi Konsumsi Jenis Makanan dari Tahun ke Tahun")

fig_line_chart.show()
```

Figure 4. : Prediksi dan Visualisasi

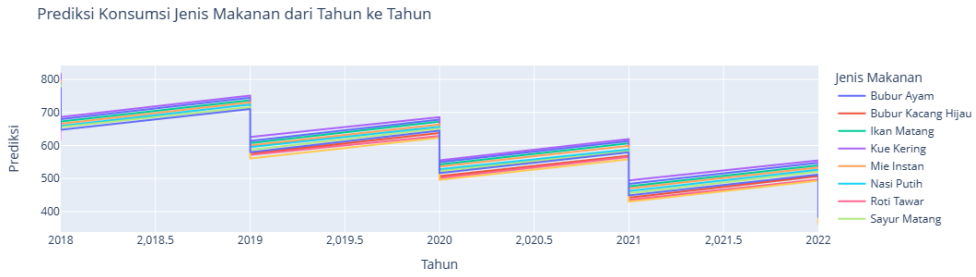

Figure 5. : Visualisasi Prediksi Konsumsi Jenis Makanan dari Tahun ke Tahun

d. Insights and Top Food Categories

The analysis identifies distinct spending patterns for various food categories, indicating shifts in consumer preferences. The top three food categories with the highest average per capita expenditures are determined and visualized. Additionally, the line chart reveals consumption trends over time, providing valuable insights for policymakers, researchers, and stakeholders in the food industry.

```
# Kelompokkan data berdasarkan jenis makanan dan hitung total konsumsi (gunakan agg)
top_food = df.groupBy("jenis_makanan").agg({"value":
"sum"}).withColumnRenamed("sum(value)", "total_konsumsi")

# Konversi DataFrame PySpark ke Pandas untuk menggunakan sort_values
top_food = top_food.toPandas().sort_values(by="total_konsumsi", ascending=False).head(5)

# Agregasi data untuk mendapatkan rata-rata pengeluaran perkapita berdasarkan jenis makanan
avg_exp_per_capita = df.groupBy("jenis_makanan").agg({"value":
"avg"}).withColumnRenamed("avg(value)", "avg_exp_per_capita")


# Ambil top 3 jenis makanan berdasarkan rata-rata pengeluaran perkapita tertinggi
top_food_avg = avg_exp_per_capita.orderBy("avg_exp_per_capita", ascending=False).limit(3)



# Visualisasi dengan Plotly
fig_top_food = px.bar(top_food_avg, x="jenis_makanan", y="avg_exp_per_capita",
                title="Top 3 Rata-Rata Pengeluaran Perkapita berdasarkan Jenis Makanan",
                labels={"avg_exp_per_capita": "Rata-Rata Pengeluaran Perkapita", "jenis_makanan":
"Jenis Makanan"})
```

```
fig_top_food.update_traces(hovertemplate="Jenis Makanan: %{x}<br>Rata-Rata Pengeluaran Perkapita: %{y:.2f}")

fig_top_food.show()
```

Figure 6. : Insights and Top Food Categories

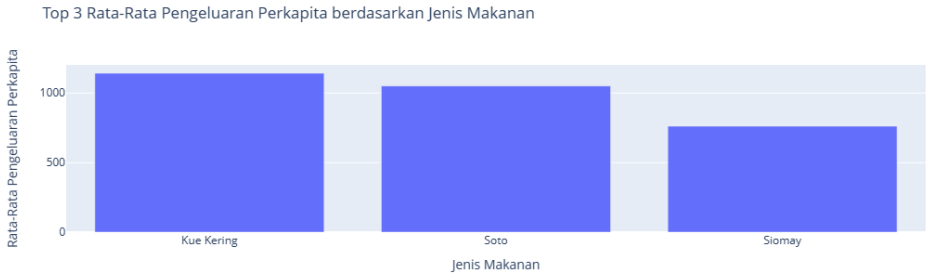

Figure 7. : Visualisasi Top 3 Food Categories

e. Regional Analysis and Pie Charts

The script includes a regional analysis where the top five cities with the highest per capita expenditures for specific food types are visualized using pie charts. This provides a localized perspective on consumption patterns and allows for targeted insights into regional variations.

```
# Ambil jenis makanan dari top 3
jenis_makanan_top_3 = [row["jenis_makanan"] for row in top_food_avg.select("jenis_makanan").distinct().limit(3).collect()]

# Loop untuk setiap jenis makanan
for jenis_makanan in jenis_makanan_top_3:
    # Ambil data untuk jenis makanan
    data_jenis_makanan = df.filter(F.col("jenis_makanan") == jenis_makanan) \
        .groupBy("tahun", "nama_wilayah").agg(F.avg("value").alias("avg_exp_per_capita")) \
        .orderBy("tahun", "avg_exp_per_capita", ascending=True)  # Urutkan berdasarkan tahun dan rata-rata pengeluaran perkapita

    # Ambil 5 kota dengan rata-rata tertinggi
    top_5_cities = data_jenis_makanan.groupBy("nama_wilayah").agg(F.max("avg_exp_per_capita").alias("max_avg_exp_per_capita")) \
                            .orderBy("max_avg_exp_per_capita", ascending=False) \
                            .limit(5)



    # Visualisasi dengan menggunakan Pie Chart
```

```
    fig_pie_chart = px.pie(top_5_cities.toPandas(), names="nama_wilayah",
values="max_avg_exp_per_capita",
                           title=f"Top 5 Kota dengan Rata-Rata Pengeluaran Perkapita Tertinggi
({jenis_makanan})")

    # Tampilkan chart
    fig_pie_chart.show()
```

Figure 7. : Regional Analysis and Pie Charts

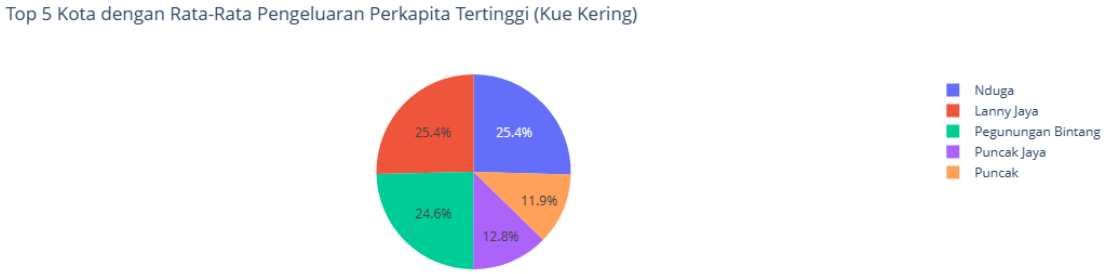

Figure 8. : Top 5 Kota Dengan Rata Rata Pengeluaran Perkapita Tertinggi (Kue Kering)

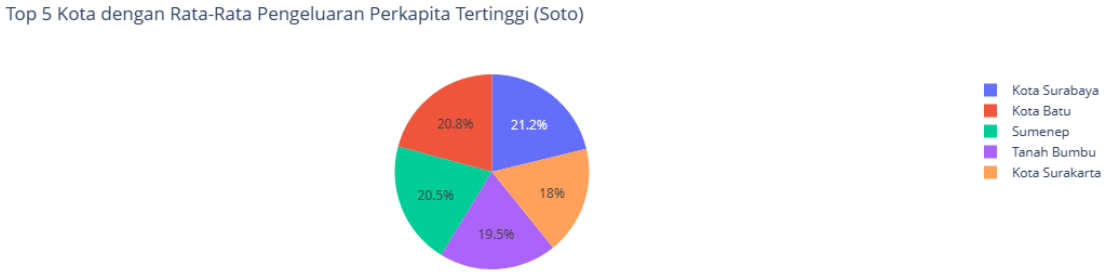

Figure 9. : Top 5 Kota Dengan Rata Rata Pengeluaran Perkapita Tertinggi (Soto)

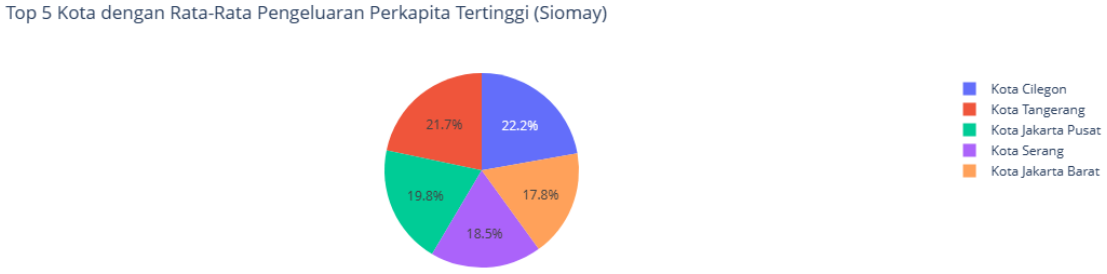

Figure 9. : Top 5 Kota Dengan Rata Rata Pengeluaran Perkapita Tertinggi (Siomay)

4. Conclusion

This research, employing PySpark's robust big data processing capabilities, successfully unraveled intricate details regarding weekly per capita food consumption. The visual narrative of predicted consumption trends showcased dynamic changes in preferences and expenditures across diverse food categories. This insight into temporal variations contributes to a deeper understanding of societal shifts in culinary choices.

Moreover, the study explored the economic implications associated with different types of food consumption, shedding light on the financial aspects of dietary preferences. The identification of the top three food categories with the highest average per capita expenditure highlighted significant contributors to consumer spending in Indonesia. This information is crucial for policymakers and industry stakeholders in crafting targeted strategies to enhance economic growth and cater to consumer needs.

In conclusion, by delving into the complexities of consumption patterns, this research provides actionable insights for policymakers, economists, and stakeholders in the food industry. Understanding the nuanced interplay between dietary habits and economic behavior is paramount for informed decision-making, ultimately contributing to the advancement of the food industry and the overall well-being of the Indonesian population.

References

[1] C. Karras, A. Karras, D. Tsolis, K. C. Giotopoulos, and S. Sioutas, "Distributed Gibbs Sampling and LDA Modelling for Large Scale Big Data Management on PySpark," *7th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference, SEEDA-CECNSM 2022*, 2022, doi: 10.1109/SEEDA-CECNSM57760.2022.9932990.

[2] "Learning PySpark - Tomasz Drabas, Denny Lee - Google Buku." Accessed: Jan. 13, 2024. [Online]. Available: https://books.google.co.id/books?hl=id&lr=&id=HVQoDwAAQBAJ&oi=fnd&pg=PP1&dq=pyspark&ots=tNFpHjLigJ&sig=XH6zVW09fZc4zGWb0u6j4HJYTr0&redir_esc=y#v=onepage&q=pyspark&f=false

[3] K. Aberer, M. Hauswirth, and A. Salehi, "Infrastructure for data processing in large-scale interconnected sensor networks," *Proceedings - IEEE International Conference on Mobile Data Management*, pp. 198–205, 2007, doi: 10.1109/MDM.2007.36.

[4] T. Reardon and C. P. Timmer, "Five inter-linked transformations in the Asian agrifood economy: Food security implications," *Glob Food Sec*, vol. 3, no. 2, pp. 108–117, 2014, doi: 10.1016/J.GFS.2014.02.001.

[5] S. Rath, A. Tripathy, and A. R. Tripathy, "Prediction of new active cases of coronavirus disease (COVID-19) pandemic using multiple linear regression model," *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, vol. 14, no. 5, pp. 1467–1474, Sep. 2020, doi: 10.1016/J.DSX.2020.07.045.

[6] K. Murniati and A. Mutolib, "The impact of climate change on the household food security of upland rice farmers in Sidomulyo, Lampung Province, Indonesia," *Biodiversitas*, vol. 21, no. 8, pp. 3487–3493, Jul. 2020, doi: 10.13057/BIODIV/D210809.

[7] I. M. I. P, R. Robidi, W. Wahyuari, and A. Subrata, "BUILDING FOOD SECURITY AT MSMEs IN INDONESIA THROUGH NATIONAL AND REGIONAL FACILITATORS," *International Journal of Engagement and Empowerment*, vol. 2, no. 1, pp. 52–58, Apr. 2022, doi: 10.53067/IJE2.V2I1.47.

[8] S. Wijaya, "Indonesian food culture mapping: A starter contribution to promote Indonesian culinary tourism," *Journal of Ethnic Foods*, vol. 6, no. 1, pp. 1–10, Sep. 2019, doi: 10.1186/S42779-019-0009-3/TABLES/1.

[9] M. Soltani Firouz, K. Mohi-Alden, and M. Omid, "A critical review on intelligent and active packaging in the food industry: Research and development," *Food Research International*, vol. 141, p. 110113, Mar. 2021, doi: 10.1016/J.FOODRES.2021.110113.

[10] K. Lavanya, K. Bharathi, A. P. Christina, and S. Chaurasia, "A CNN-Based License Plate Recognition Using TensorFlow and PySpark," *https://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/978-1-6684-4246-3.ch001*, pp. 1–12, Jan. 1AD, doi: 10.4018/978-1-6684-4246-3.CH001.

[11] P. Singh, "Manage Data with PySpark," *Machine Learning with PySpark*, pp. 15–37, 2022, doi: 10.1007/978-1-4842-7777-5_2.

[12] M. Bowles and an O. M. Company. Safari, "Machine Learning with Spark and Python, 2nd Edition," p. 368, 2020, Accessed: Jan. 14, 2024. [Online]. Available: https://books.google.com/books/about/Machine_Learning_with_Spark_and_Python.html?hl=id&id=4RuzDwAAQBAJ

[13] G. James, D. Witten, T. Hastie, R. Tibshirani, and J. Taylor, "Linear Regression," pp. 69–134, 2023, doi: 10.1007/978-3-031-38747-0_3.

[14] "Beyond Multiple Linear Regression: Applied Generalized Linear Models And ... - Paul Roback, Julie Legler - Google Buku." Accessed: Jan. 14, 2024. [Online]. Available: https://books.google.co.id/books?hl=id&lr=&id=pYAUEAAAQBAJ&oi=fnd&pg=PP1&dq=linear+regression&ots=YumVKdhSap&sig=Uv9pejydpPc6gjjOOSjbf_GujWM&redir_esc=y#v=onepage&q=linear%20regression&f=false

[15] U. Sara, M. Akter, M. S. Uddin, U. Sara, M. Akter, and M. S. Uddin, "Image Quality Assessment through FSIM, SSIM, MSE and PSNR—A Comparative Study," *Journal of Computer and Communications*, vol. 7, no. 3, pp. 8–18, Mar. 2019, doi: 10.4236/JCC.2019.73002.

[16] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Comput Sci*, vol. 7, pp. 1–24, Jul. 2021, doi: 10.7717/PEERJ-CS.623/SUPP-1.

[17] Y. Yang *et al.*, "MSE-Net: generative image inpainting with multi-scale encoder," *Visual Computer*, vol. 38, no. 8, pp. 2647–2659, Aug. 2022, doi: 10.1007/S00371-021-02143-0/METRICS.

[18] R. S. Azad, S. Z. Kassab, and T. H. Dang, "Experimental evaluation of approximations for mean values of w-squared and vw-squared," *https://doi.org/10.2514/3.9600*, vol. 25, no. 1, pp. 171–173, May 2012, doi: 10.2514/3.9600.

[19] T. G. Nick, "Descriptive statistics.," *Methods Mol Biol*, vol. 404, pp. 33–52, 2007, doi: 10.1007/978-1-59745-530-5_3/COVER.

[20] F. S. Seda, L. Setyawati, T. Tirta, and K. Nobel, "Dataset on The Cultural Dimension of Urban Society Food Consumption in Indonesia," *Data Brief*, vol. 31, p. 105681, Aug. 2020, doi: 10.1016/J.DIB.2020.105681.

[21] K. Pawlak and M. Kołodziejczak, "The Role of Agriculture in Ensuring Food Security in Developing Countries: Considerations in the Context of the Problem of Sustainable Food Production," *Sustainability 2020, Vol. 12, Page 5488*, vol. 12, no. 13, p. 5488, Jul. 2020, doi: 10.3390/SU12135488.

[22] S. Lathuiliere, P. Mesejo, X. Alameda-Pineda, and R. Horaud, "A Comprehensive Analysis of Deep Regression," *IEEE Trans Pattern Anal Mach Intell*, vol. 42, no. 9, pp. 2065–2081, Sep. 2020, doi: 10.1109/TPAMI.2019.2910523.

[23] L. Zhu *et al.*, "SDN Controllers," *ACM Computing Surveys (CSUR)*, vol. 53, no. 6, Dec. 2020, doi: 10.1145/3421764.

[24] "Applied Regression Modeling - Iain Pardoe - Google Buku." Accessed: Jan. 14, 2024. [Online]. Available: https://books.google.co.id/books?hl=id&lr=&id=hToLEAAAQBAJ&oi=fnd&pg=PR11&dq=linear+regression+model&ots=M-e5yPCr5s&sig=WChq-adiPC8U1iXmG4_0aWYIrdE&redir_esc=y#v=onepage&q=linear%20regression%20model&f=false

[25] D. C. Montgomery, E. A. Peck, and G. G. Vining, "Introduction to linear regression analysis," p. 673, Accessed: Jan. 14, 2024. [Online]. Available: https://books.google.com/books/about/Introduction_to_Linear_Regression_Analys.html?hl=id&id=tCIgEAAAQBAJ

[26] B. M. Golam Kibria and A. F. Lukman, "A new ridge-type estimator for the linear regression model: Simulations and applications," *Scientifica (Cairo)*, vol. 2020, 2020, doi: 10.1155/2020/9758378.