|  |  |  |
| --- | --- | --- |
| Linear Regression Analysis of Weekly Per-Capita Consumption Based on Food Groups in Indonesia using PySpark. | | |
|  | | |
| Muhammad Farhan Alfarizi1 | | |
| 1) Teknik Informatika, Pelita Bangsa University, West Java, Indonesia | | |
|  |  |  |
| Article Info |  | Abstract |
| *Article history*  Received : diisi oleh editor  Revised : diisi oleh editor  Accepted : diisi oleh editor |  | *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.* |
|  |  |
| *Kata Kunci:*  Weekly Per Capita;  Consumption;  PySpark;  Linear Regression;  Food Categories;  Economic Trends;  Dietary Habits;  Indonesia; |  |
| ***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 | | |
| *This is an open access article under the* [*CC BY-NC*](https://creativecommons.org/licenses/by-nc/4.0/) *license.* | | |

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, a powerful big data processing framework, to conduct a Linear Regression Analysis on weekly per capita consumption across different food categories. 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. 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.

1. **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. The training and evaluation of the Linear Regression model will involve a split-sample validation approach, with a portion of the dataset reserved for model training and the remainder for validation. The model's performance will be assessed using relevant metrics such as Mean Squared Error (MSE) and R-squared.

Additionally, the study will employ descriptive statistics to provide an overview of the central tendencies and distributions within the dataset. 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.

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.

1. **Result and Discussion**

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

1. Data Preprocessing and Feature Engineering

The PySpark script begins with essential data preprocessing steps, addressing missing values represented by '-'. These values are replaced with '0' 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('-', '0', '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

1. 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



Figure 3. : Running Model Training and Evaluation

1. 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

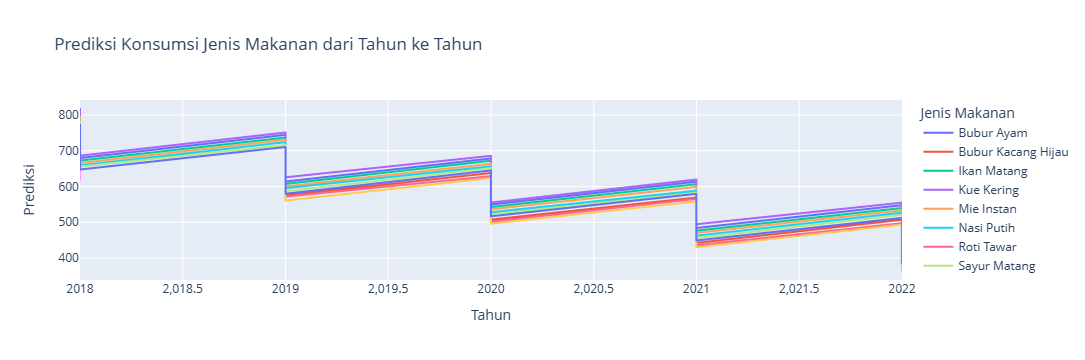


Figure 5. : Visualisasi Prediksi Konsumsi Jenis Makanan dari Tahun ke Tahun

1. 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() |

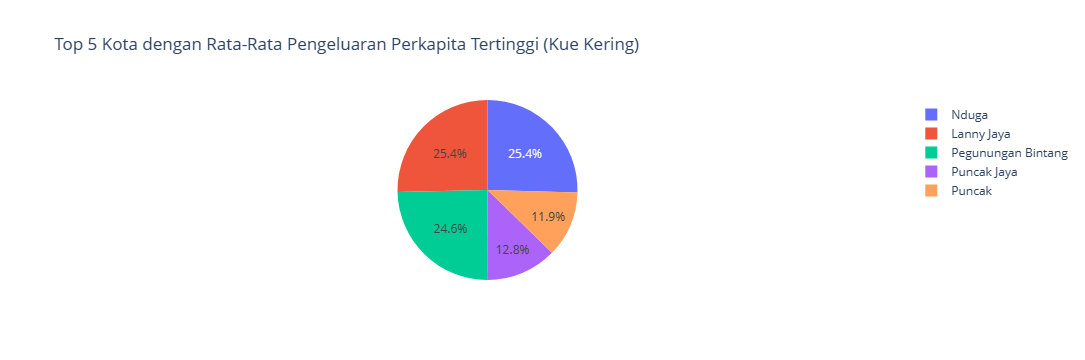
Figure 6. : Insights and Top Food Categories  

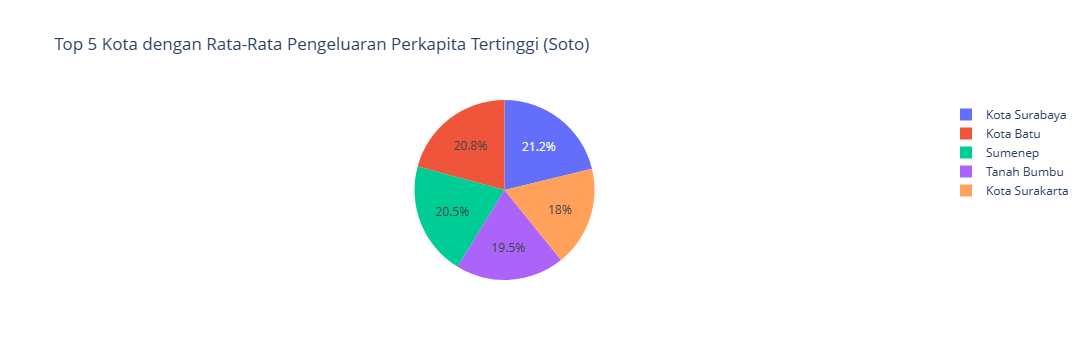

Figure 7. : Visualisasi Top 3 Food Categories

1. 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)

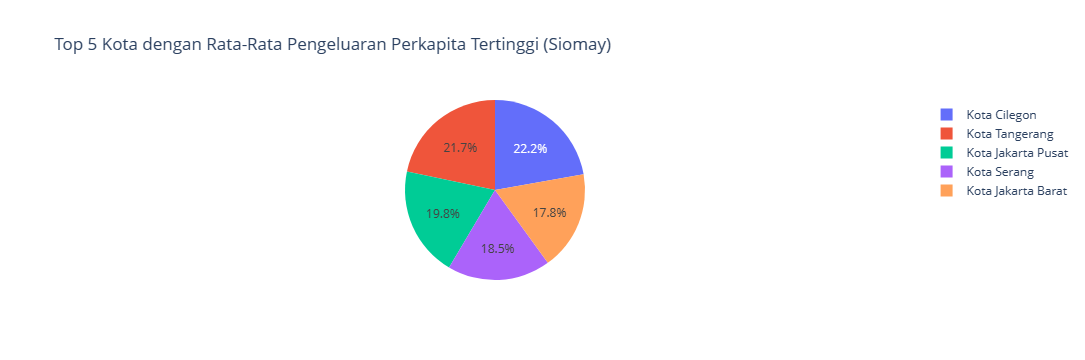


Figure 9. : Top 5 Kota Dengan Rata Rata Pengeluaran Perkapita Tertinggi (Siomay)

1. **Conclusionnya**

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**

(one blank single space line, 10 pt)

References must be written in the order they appear in the text, using Arabic numerals in square brackets, as shown in the example. References of at least 25 journal articles in the last 5 years, 80% of which must be sourced from international reputable journals and try the reference has a URL/Link. References must use a Reference Management application such as Mendeley or similar application and use IEEE style and 9pt font size.

[1] R. E. Caraka *et al.*, “Impact of COVID-19 large scale restriction on environment and economy in Indonesia,” *Glob. J. Environ. Sci. Manag.*, vol. 6, no. Special Issue (Covid-19), pp. 65–84, 2020.

[2] M. Machida *et al.*, “Incorrect use of face masks during the current COVID-19 pandemic among the general public in Japan,” *Int. J. Environ. Res. Public Health*, vol. 17, no. 18, p. 6484, 2020.

[3] W. Bourequat and H. Mourad, “Sentiment Analysis Approach for Analyzing iPhone Release using Support Vector Machine,” *Int. J. Adv. Data Inf. Syst.*, vol. 2, no. 1, pp. 36–44, 2021, doi: 10.25008/ijadis.v2i1.1216.

[4] P. Chen, R. Tong, B. Yu, and Y. Wang, “Reliable shortest path finding in stochastic time-dependent road network with spatial-temporal link correlations: A case study from Beijing,” *Expert Syst. Appl.*, vol. 147, 2020, doi: 10.1016/j.eswa.2020.113192.

[5] D.-H. Oh, J. Ahn, S. Lee, and H. Choi, “Measuring technical inefficiency and CO2 shadow price of Korean fossil-fuel generation companies using deterministic and stochastic approaches,” *Energy Environ.*, p. 0958305X20932547, 2020.

[6] Y. J. Han *et al.*, “Treatment of severe acute respiratory syndrome (SARS), Middle East respiratory syndrome (MERS), and coronavirus disease 2019 (COVID-19): a systematic review of in vitro, in vivo, and clinical trials,” *Theranostics*, vol. 11, no. 3, p. 1207, 2021.

[7] A. Hocine, M. S. Guellil, E. Dogan, S. Ghouali, and N. Kouaissah, “A fuzzy goal programming with interval target model and its application to the decision problem of renewable energy planning,” *Environ. Ecol. Stat.*, vol. 27, no. 3, pp. 527–547, 2020.

[8] Z. Zhai, J. Fernán Martínez, N. Lucas Martínez, and V. Hernández Díaz, “Applying case-based reasoning and a learning-based adaptation strategy to irrigation scheduling in grape farming,” *Comput. Electron. Agric.*, vol. 178, p. 105741, Nov. 2020, doi: 10.1016/j.compag.2020.105741.

[9] A. M. Catai, C. M. Pastre, M. F. de Godoy, E. da Silva, A. C. de Medeiros Takahashi, and L. C. M. Vanderlei, “Heart rate variability: are you using it properly? Standardisation checklist of procedures,” *Brazilian J. Phys. Ther.*, vol. 24, no. 2, pp. 91–102, 2020.

[10] B. M. F. Viana, L. T. Pereira, C. F. M. Toledo, S. R. dos Santos, and S. M. D. M. Maia, “Feasible–Infeasible Two-Population Genetic Algorithm to evolve dungeon levels with dependencies in barrier mechanics,” *Appl. Soft Comput.*, vol. 119, p. 108586, 2022.

[11] S. R. Jayashree, G. Dias, J. J. Andrew, S. Saha, F. Maurel, and S. Ferrari, “Multimodal Web Page Segmentation Using Self-organized Multi-objective Clustering,” *ACM Trans. Inf. Syst.*, vol. 40, no. 3, pp. 1–49, 2022.

[12] A. Mughaid *et al.*, “An intelligent cybersecurity system for detecting fake news in social media websites,” *Soft Comput.*, pp. 1–15, 2022.

[13] O. Nahli, A. Sanna, M. Bandini, and F. Boschetti, “Commerce Numérique: traffic signals for the crossroads between cultures,” *Int. J. Inf. Sci. Technol.*, vol. 5, no. 3, pp. 36–45, 2022.

[14] M. Kumar, J. Aggarwal, A. Rani, T. Stephan, A. Shankar, and S. Mirjalili, “Secure video communication using firefly optimization and visual cryptography,” *Artif. Intell. Rev.*, vol. 55, no. 4, pp. 2997–3017, 2022.

[15] M. K. Gupta *et al.*, “Detection and localization for watermarking technique using LSB encryption for DICOM Image,” *J. Discret. Math. Sci. Cryptogr.*, pp. 1–12, 2022.

[16] M. Salehi, G. De Borger, D. Hughes, and B. Crispo, “NemesisGuard: Mitigating interrupt latency side channel attacks with static binary rewriting,” *Comput. Networks*, p. 108744, 2022.

[17] G. Palumbo *et al.*, “A Structured Approach for the Design and Manufacturing of Titanium Cranial Prostheses via Sheet Metal Forming,” *Metals (Basel).*, vol. 12, no. 2, p. 293, 2022.

[18] J.-S. Park, J.-H. Kim, Y.-C. Jeong, H.-T. Kim, S.-K. Kim, and J.-M. Lee, “Effect of Corrugated Sheet Diameter on Structural Behavior under Cryogenic Temperature and Hydrodynamic Load,” *Metals (Basel).*, vol. 12, no. 3, p. 521, 2022.

[19] P. Stachelek, L. MacKenzie, D. Parker, and R. Pal, “Circularly polarised luminescence laser scanning confocal microscopy to study live cell chiral molecular interactions,” *Nat. Commun.*, vol. 13, no. 1, pp. 1–8, 2022.

[20] D. Mezad-Koursh *et al.*, “Response to: Comment on:‘Peripapillary hyperreflective ovoid mass-like structures—a novel entity as frequent cause of pseudopapilloedema in children,’” *Eye*, p. 1, 2022.

[21] P. Chen, L. Xu, and Z. Li, “Modeling methods of self-centering energy dissipation braces using OpenSees,” *J. Constr. Steel Res.*, vol. 193, p. 107267, 2022.

[22] D. Stanford, Z. Yang, and S. Yao, “Subleading weingartens,” *J. High Energy Phys.*, vol. 2022, no. 2, pp. 1–50, 2022.

[23] B. Joksimoski *et al.*, “Technological solutions for sign language recognition: a scoping review of research trends, challenges, and opportunities,” *IEEE Access*, 2022.

[24] J. K. Lê and T. Schmid, “The practice of innovating research methods,” *Organ. Res. Methods*, vol. 25, no. 2, pp. 308–336, 2022.

[25] J. Miller and R. Ulrich, “Optimizing Research Output: How Can Psychological Research Methods Be Improved?,” *Annu. Rev. Psychol.*, vol. 73, pp. 691–718, 2022.

[26] S. Hosseinpour, A. Gaudin, and O. A. Peters, “A critical analysis of research methods and experimental models to study biocompatibility of endodontic materials,” *Int. Endod. J.*, 2022.