The Impact of Sectoral Price Fluctuations on Dining Costs (Personal Service): Consumer Price Index Analysis Based on the Distinction of COVID-19 Periods

1. Project Overview

1.1 Background:

This research reports the findings of an individual study conducted on the topic of "The Relationship between COVID-19 and Inflation." The study concluded that the rise in commodity prices, particularly in raw food materials, is not correlated with an increase in dining-out costs.

The research focused on various sectors, including food and non-alcoholic beverages, alcoholic beverages and tobacco, recreation and culture, and food and accommodation, using the Consumer Price Index as a central measure. The study divided the COVID-19 period into pre-, during, and post-phases, employing a two-way ANOVA analysis. The results indicated no significant correlation between inflation and price increases in each sector, with the most substantial inflation observed after the COVID-19 period (May 2022 to October 2023).

However, this finding contradicts common sense, we know there is a close relationship between the rise in dining-out costs and the increase in raw material prices. Therefore, the project was undertaken with the question of whether there is truly no correlation between the increase in food prices and the rise in dining-out costs.

1.2 Data Collection and Preprocessing

The data was collected through the KOSIS OpenAPI,
"품목별 소비자물가지수(품목성질별: 2020=100)."

Description of the Collected Data:
The data is broadly categorized into goods and services. Goods are further classified into three subcategories: agricultural and fishery products, industrial products, and electricity, gas, and water. Services are classified into rent, public services, and personal services.

For goods, the increase in price indices for agricultural & marine products, industrial products, and electricity, gas, and water all contribute to the overall increase.

Specifically addressing personal services, further categorization is made into dining and non-dining. Examining the data, it is evident that dining has the most significant impact on the overall increase in the price index for personal services. Therefore, in this study, personal services are considered synonymous with dining.

When collecting data through the openAPI, the collected price index data is in object format, requiring conversion to numeric values using pandas for subsequent statistical calculations. Additionally, the names of the fields, which should be the titles of each column, are collected as object data in the "field" column. Using pandas, this information is transformed into a dataframe with fields and corresponding price indices.

2. Data analysis

2.1 Data grouping and statistical analysis

```
df.groupby('분야').mean()
```

✓ 0.0s

분야	물가지수
Agricultural & marine products	102.015222
Electricity and Gas and Water	105.283000
Industrial products	102.986519
Personal services	101.623481

```
df.groupby('분야').count()
```

✓ 0.0s

분야	물가지수
Agricultural & marine products	27
Electricity and Gas and Water	27
Industrial products	27
Personal services	27

```
df.groupby('분야').max()
```

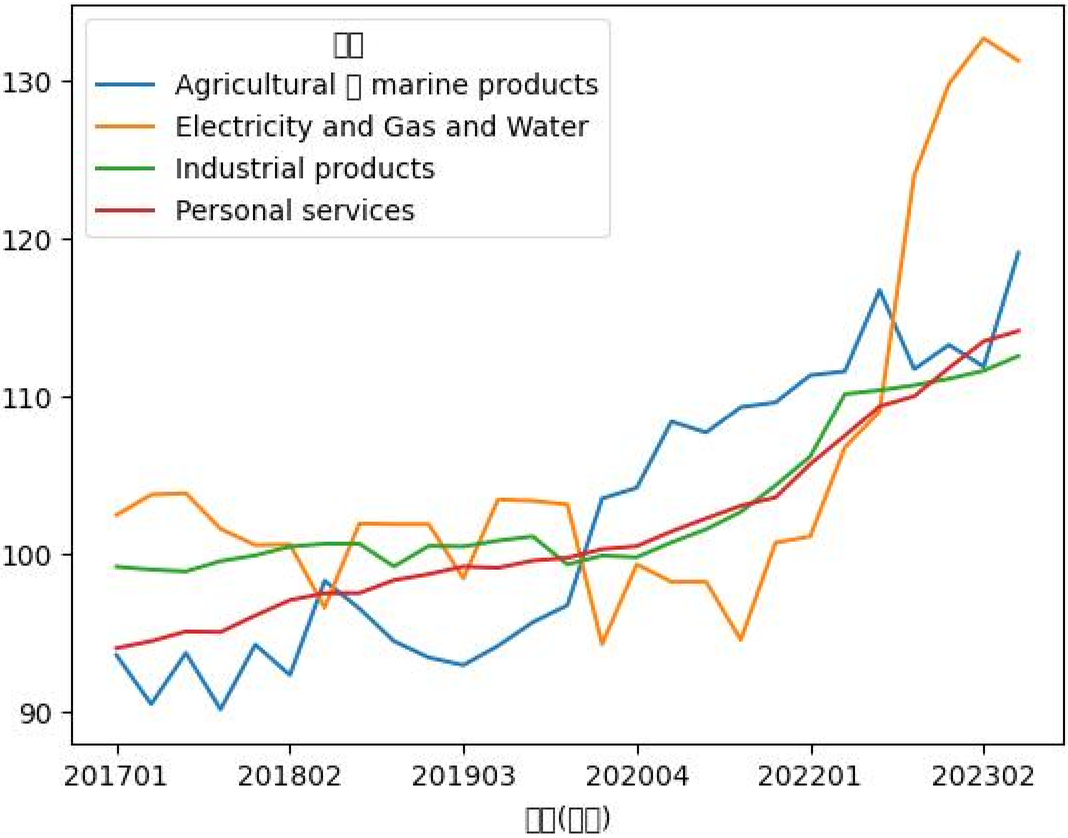
✓ 0.0s

분야	물가지수
Agricultural & marine products	119.06
Electricity and Gas and Water	132.65
Industrial products	112.53
Personal services	114.12

After extracting the data using the OpenAPI and retaining only the necessary information, as the fields are now represented as columns, I used groupby in this state to divide the data into groups and uncover statistical insights.

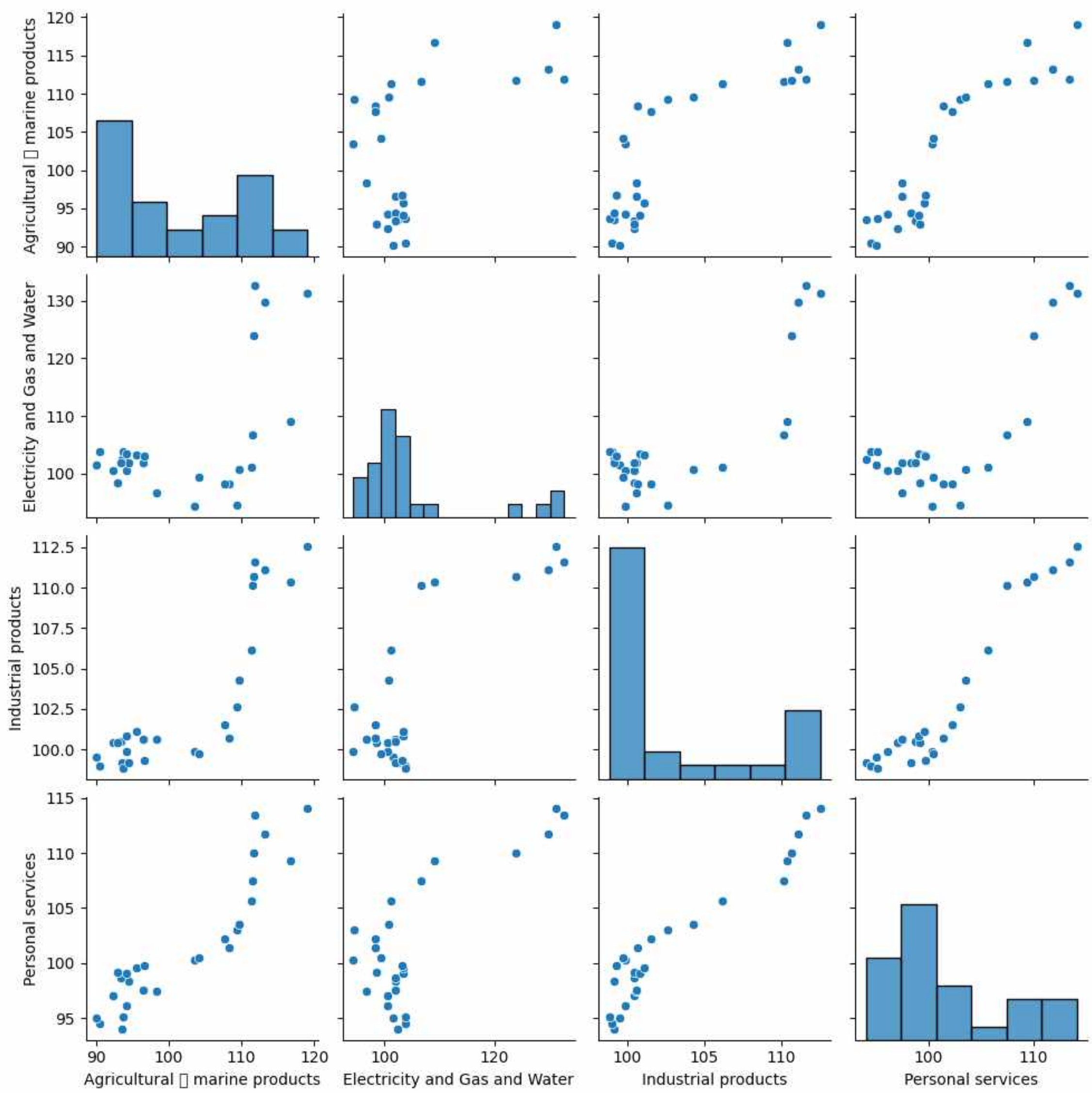
2.2 Visualization

Graph 1



I used Matplotlib to draw graphs of the DataFrame. The x-axis represents the period (quarter), and the y-axis represents the price index. I drew graphs for each field with different colors. It is noticeable that all of them depict an upward trend

Graph 2



I used Seaborn to create a graph that allows me to examine the partial correlations between the dependent variable and multiple independent variables.

3. Result interpretation and application direction design

3.1 Model training and evaluation

I will perform machine learning using the multiple regression approach. The independent variables are defined as follows:

```
x = ndf[['Agricultural & marine products','Electricity and Gas and Water','Industrial products']] #독립변수
y=ndf['Personal services'] #종속변수
```

```
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(x,y,test_size=0.3,random_state=10)
```

```
print('traindata: ',x_train.shape)
print('testdata : ',x_test.shape)
```

```
traindata:  (18, 3)
testdata :  (9, 3)
```

```
from sklearn.linear_model import LinearRegression
```

```
lr = LinearRegression()
```

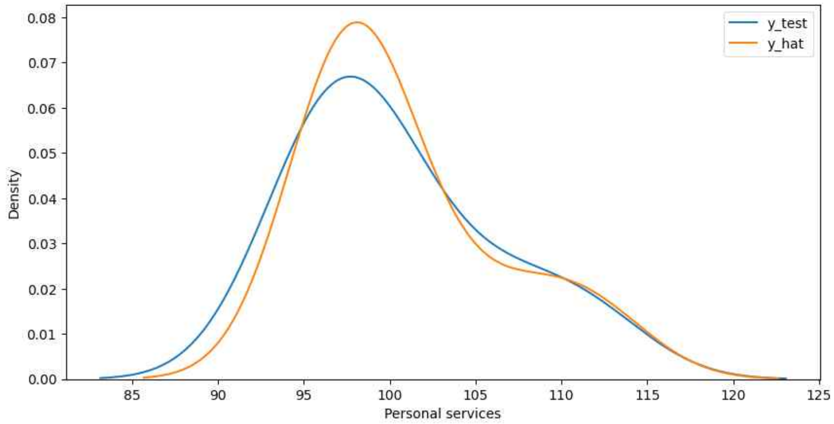
```
lr.fit(x_train, y_train)
```

```
r_square = lr.score(x_test,y_test)
print(r_square)
```

```
0.9566727500121981
```

I used sklearn to split the data into training and testing sets.

And also, I use skelarn to conduct multiple regression and evaluate its performance.
The machine learning model shows good performance.



In multiple regression, it is essential to check for multicollinearity. Multicollinearity occurs when there is a strong correlation between independent variables, which can make it challenging to interpret the results of regression analysis. Multicollinearity can be assessed using a correlation matrix or VIF (Variance Inflation Factor). For instance, a VIF generally exceeding 10 indicates a high likelihood of multicollinearity.

Correlation Matrix:

The correlation between 'Agricultural & marine products' and 'Electricity and Gas and Water' is 0.53.
The correlation between 'Agricultural & marine products' and 'Industrial products' is 0.85.
The correlation between 'Electricity and Gas and Water' and 'Industrial products' is 0.82.
VIF:

VIF for 'Agricultural & marine products' is 4.88, indicating a moderate level.
VIF for 'Electricity and Gas and Water' is 4.09, indicating a moderate level.
VIF for 'Industrial products' is 10.45, relatively high compared to other variables.
Given the higher VIF for 'Industrial products,' its presence may lead to inaccuracies in regression coefficients, a decrease in statistical significance, and a potential decline in model accuracy. Therefore, it is recommended to remove 'Industrial products' and reevaluate the regression model.

New Variables:
x_new = ndf[['Agricultural & marine products', 'Electricity and Gas and Water']] # New independent variables
y_new = ndf['Personal services'] # New dependent variable

```
# 독립 변수 열 선택
independent_variable_columns = ['Agricultural & marine products', 'Electricity and Gas and Water']

# 데이터프레임에서 필요한 열만 추출
data = ndf[independent_variable_columns]

# 상수 열 추가
X = sm.add_constant(data)

# VIF 계산
vif_data = pd.DataFrame()
vif_data["Variable"] = X.columns
vif_data["VIF"] = [variance_inflation_factor(X.values, i) for i in range(X.shape[1])]

# 결과 출력
print("VIF:")
print(vif_data)
```

```
VIF:
                        Variable         VIF
0                          const  152.317590
1  Agricultural & marine products    1.393964
2   Electricity and Gas and Water    1.393964
```

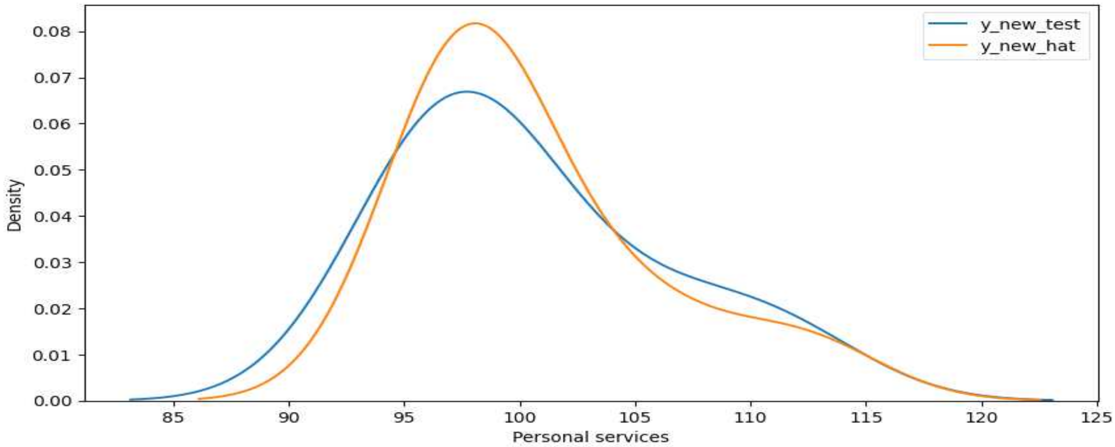
The VIF values for 'Agricultural & marine products' and 'Electricity and Gas and Water' variables are both 1.39. Typically, these low values suggest that there is no significant issue with multicollinearity.
The VIF for the constant term is relatively high at 152.32. However, high VIF for the constant term is not a major concern.
In the current results, there doesn't seem to be a significant multicollinearity issue among the selected independent variables.

3.2 Interpretation of results

3.2.1 Interpretation of model evaluation results

```
r_new_square = new_lr.score(x_new_test,y_new_test)
print(r_new_square)
```

```
0.9325743137617126
```



Performance analysis of new multiple regression model

In the multiple regression model for the two variables, the regression coefficients for each variable are statistically significant, and the model's coefficient of determination is high. This indicates that the selected two variables, 'Agricultural & marine products' and 'Electricity and Gas and Water,' have meaningful predictive power for the dependent variable 'Personal services,' and their relationship contributes to the model.

In other words, the model demonstrates good performance.

3.2.2 Interpretation of Data Analysis Results

The fluctuations in the prices of 'Agricultural & marine products' and 'Electricity and Gas and Water' have statistically significant effects on dining costs ('Personal services').

In my personal research investigating the correlation between COVID-19 and inflation, with a specific focus on the impact on 'Personal services'—namely, dining expenses, I initially concluded that there was no statistically significant association between 'Agricultural & marine products' and 'Personal services.' However, through data analysis and machine learning-based multiple regression analysis, a different picture emerged. The results revealed that 'Agricultural & marine products' and 'Electricity and Gas and Water' variables exhibit statistically significant predictive power concerning 'Personal services.' The regression coefficients between these variables were statistically significant, and the model's coefficient of determination was notably high.

The evaluation of multicollinearity indicated that there was not a significant multicollinearity issue among the chosen variables. The Variance Inflation Factors (VIF) for 'Agricultural & marine products' and 'Electricity and Gas and Water'

variables were generally low, suggesting no substantial multicollinearity concerns.
In the multiple regression model for these two variables, the regression coefficients for each variable were statistically significant, implying that the selected variables possess meaningful predictive power for 'Personal services.'

In summary, these comprehensive results indicate that the fluctuations in prices for 'Agricultural & marine products' and 'Electricity and Gas and Water' significantly influence dining expenses ('Personal services') in a statistically meaningful manner.

This indicates that prior research might have been incorrect, emphasizing the need to recognize potential inaccuracies in statistical experimental results. It underscores the importance of conducting analyses in various ways to gain a more comprehensive understanding.

3.3 Application direction design
Application direction and future plans based on project results

Consideration of Other External Variables:
Expanding the project results by incorporating various external variables such as economic conditions, demographic data, climate, etc., can lead to the development of a more comprehensive model. Analyzing the interplay of these factors can provide deeper insights into the dynamics affecting the dining-out industry.

Continuous Monitoring and Updates, Utilization of Artificial Intelligence:
Introducing artificial intelligence technologies can enhance predictive models and strengthen decision-making using real-time data. AI-driven insights can contribute to more agile and responsive marketing and operational strategies in the restaurant industry.

Creating a platform for uploading prices in the wholesale market and utilizing natural language processing and machine learning for real-time web crawling to update data. The platform integrates various factors and employs multiple regression for price prediction.

3. Upload

3.1 Repository URL

https://github.com/yoonseongan/statstat

I have uploaded the prior research on GitHub(물가지수이원배치법분석).