

Inferential Statistics

Case Study-2



Diet Case Study

✓ Problem Statement

✓ Business Context

From children to adults to the elderly, the importance of a balanced diet can't be emphasized enough for any age group for a healthy lifestyle. A proper, well-balanced meal plan helps to attain ideal body weight and reduce the risk of chronic diseases like diabetes, cardiovascular and other types of cancer.

The Health Company, which provides various diet plans for weight loss, conducted a market test experiment to test three different kinds of diets (A, B, C). Each of the volunteers was given one of the three diet plans and asked to follow the diet for 6 weeks.

✓ Objective

In order to understand the effectiveness of each of the different diets for weight loss across various age groups, the executives of the company reached out to you, a data scientist at the company. The weights before starting the diet and the weight 6 weeks after following the diet were recorded for 78 volunteers across different age groups who were provided with either of the three diet plans. You have been asked to perform a statistical analysis to find evidence of whether the mean weight losses with respect to the three diet plans and different age groups are significantly different.

Note: Consider a 5% significance level for the analysis.

✓ Data Dictionary

The *diet.csv* file contains the following information:

1. diet: Type of the diet
2. preweight: Weight (in kg) before following the diet plan
3. weight6weeks: Weight (in kg) after 6 weeks of following the diet plan
4. age_group: Age group of the participants when they started following the diet plan

✓ Installing and importing the necessary libraries

```
# # Installing the libraries with the specified version.  
# !pip install numpy==1.25.2 pandas==2.0.3 matplotlib==3.7.1 seaborn==0.13.1 scipy==1.11.4 s
```

Note: After running the above cell, kindly restart the notebook kernel/runtime (depending on whether Jupyter Notebook/Google Colab is being used) and run all cells sequentially from the next cell.

```
# Libraries to help with reading and manipulating data  
import numpy as np  
import pandas as pd
```

```
# Libraries to help with data visualization
```

```
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline

# Library to help with statistical analysis
import scipy.stats as stats
from statsmodels.formula.api import ols      # For n-way ANOVA
from statsmodels.stats.anova import _get_covariance, anova_lm # For n-way ANOVA
```

✓ Loading the dataset

```
# uncomment the below code cell to mount your google drive
# from google.colab import drive
# drive.mount('/content/drive')

data = pd.read_csv('diet.csv')

# copying data to another variable to avoid any changes to original data
df = data.copy()
```

✓ Data Overview

The initial steps to get an overview of any dataset is to:

- observe the first few rows of the dataset, to check whether the dataset has been loaded properly or not
- get information about the number of rows and columns in the dataset
- find out the data types of the columns to ensure that data is stored in the preferred format and the value of each property is as expected.
- check the statistical summary of the dataset to get an overview of the numerical columns of the data

✓ Displaying the first few rows of the dataset

```
# looking at head (5 observations)
df.head()
```



	diet	preweight	weight6weeks	age
0	B	60	60.0	45
1	B	103	103.0	38
2	A	58	54.2	31
3	A	60	54.0	18
4	A	64	63.3	35

- The dataset contains the diet type, pre-weight and weight after 6 weeks for one person in each row.

✓ Checking the shape of the dataset

```
df.shape
```



```
(78, 4)
```

- The dataset contains 78 rows and 4 columns

✓ Checking the data types of the columns for the dataset

```
df.info()
```




```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 78 entries, 0 to 77
Data columns (total 4 columns):
#   Column          Non-Null Count  Dtype
---  -
0   diet            78 non-null    object
1   preweight       78 non-null    int64
2   weight6weeks    78 non-null    float64
3   age            78 non-null    int64
dtypes: float64(1), int64(2), object(1)
memory usage: 2.6+ KB
```

- The diet column is categorical. While, preweight, weight6weeks and age are numerical.
- There are no missing values in the dataset

✓ Statistical summary of the dataset

```
df.describe()
```




	preweight	weight6weeks	age
count	78.000000	78.000000	78.000000
mean	72.525641	68.680769	34.025641
std	8.723344	8.924504	8.543205
min	58.000000	53.000000	18.000000
25%	66.000000	61.850000	28.250000
50%	72.000000	68.950000	34.000000
75%	78.000000	73.825000	39.000000
max	103.000000	103.000000	54.000000

- The average weight before any diet plan is 72.53 kg, while the average weight after 6 weeks of a diet plan is 68.68 kg. Overall, taking a diet has reduced the average weight.
- The variation in the weights is very similar before and after taking a diet.

✓ Exploratory Data Analysis

✓ Value counts of different types of diet

```
df.diet.value_counts()
```



```
diet
B    27
C    27
A    24
Name: count, dtype: int64
```

- As mentioned in the problem statement, there are three types of diet plans: A, B, and C
- In the sample of 78 people, 27 each took diet plans B and C, while 24 took diet plan A

✓ Creating a new column 'weight_loss' and 'age_category'

To assess the loss in weight after following a diet plan and considering different age groups, we shall create a new column `weightloss` by subtracting the weight before starting the diet from the

weight after completing the diet. Additionally, we will add a new column `age_category` based on the age groups defined as 18-25, 25-40, and 40+ using the provided age categorization function.

```
# calculate the weight-loss
df['weight_loss'] = df['weight6weeks'] - df['preweight']

# Defining the age groups
def categorize_age(age):
    if age >= 18 and age < 25:
        return "18-25"
    elif age >= 25 and age < 40:
        return "25-40"
    else:
        return "40+"

# Add the 'age' column to the DataFrame
df['age_category'] = df['age'].apply(categorize_age)

df.head()
```



	diet	preweight	weight6weeks	age	weight_loss	age_category
0	B	60	60.0	45	0.0	40+
1	B	103	103.0	38	0.0	25-40
2	A	58	54.2	31	-3.8	25-40
3	A	60	54.0	18	-6.0	18-25
4	A	64	63.3	35	-0.7	25-40

- A negative value of `weightloss` indicates that the diet was effective in decreasing the weight of the person after 6 weeks

✓ Value counts of different types of age category

```
df.age_category.value_counts()
```



```
age_category
25-40      47
40+        19
18-25      12
Name: count, dtype: int64
```

- Out of a sample of 78 people, 12 are in the 18-25 age category, 47 are in the 25-40 category, and 19 fall into the 40+ category.

✓ Check for Missing Values

```
df.isnull().sum()
```

```
⇒ diet          0
   preweight     0
   weight6weeks  0
   age           0
   weight_loss   0
   age_category  0
   dtype: int64
```

- There are no missing values present in the dataset.

✓ Check for duplicate Values

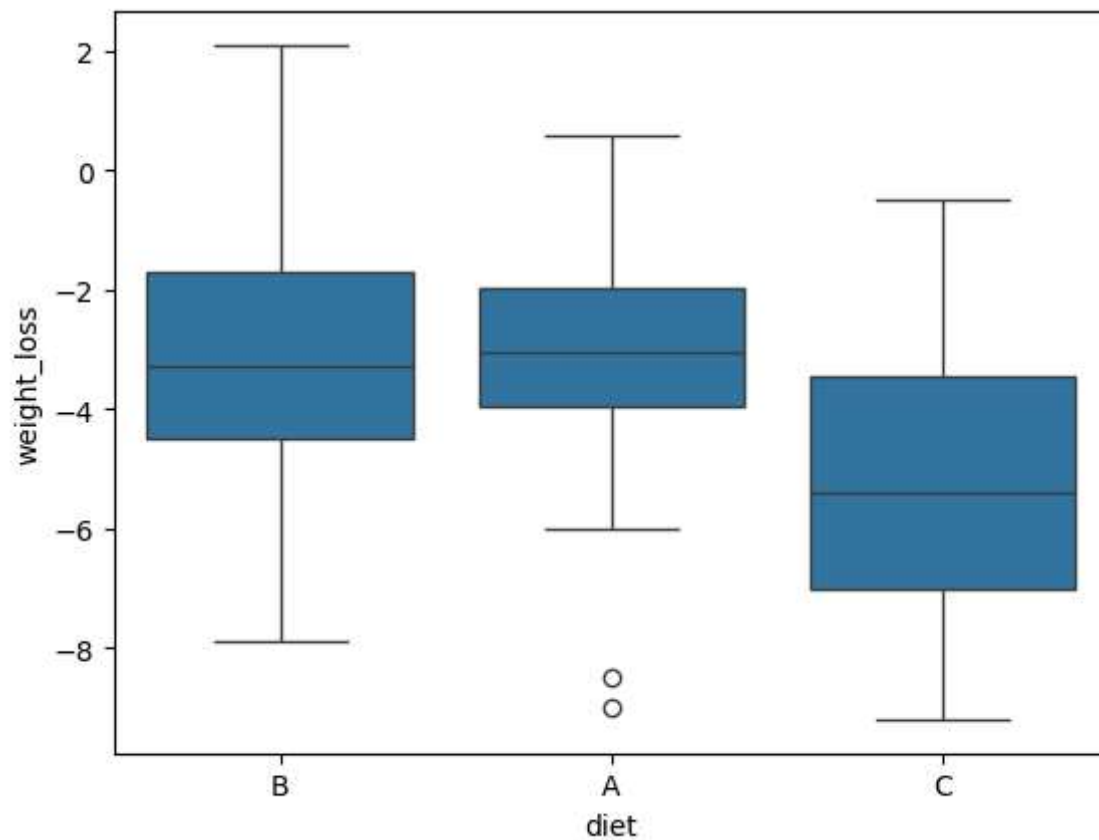
```
df.duplicated().sum()
```

```
⇒ 0
```

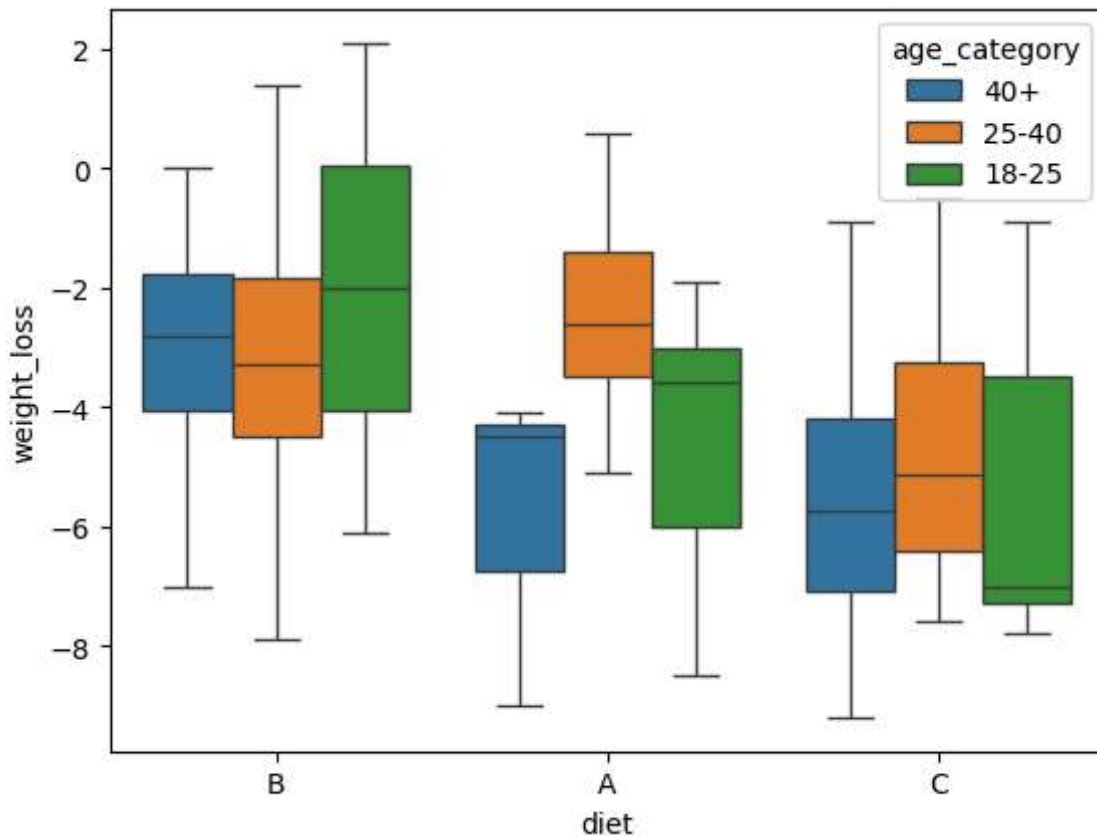
- There are no duplicate values present in the dataset.

✓ Visualize data

```
# visual analysis of the weightloss for the three diet-plans
sns.boxplot(x="diet", y="weight_loss", data = df)
plt.show()
```



```
# Visual analysis of the weight loss for the three diet-plans with hue based on 'age' column
sns.boxplot(x="diet", y="weight_loss", hue="age_category", data=df)
plt.show()
```

- The distribution of weight loss seem to differ among the three groups.
- Diet plan C seems to impact the highest weight loss.
- The median weight loss seems to be very close for the plans A and B, but the variation is higher in the weight-losses by diet-plan B as compared to A.
- Is the osberved difference in weight loss among the three groups significant enough to conclude the same about the three diet plans? To determine this, we will test the difference using a statistical test.

✓ Hypothesis Testing

✓ Hypothesis - 1

✓ Step 1: Define null and alternative hypotheses for One-Way ANONA

The null and alternative hypotheses can be formulated as:

H_0 : The mean weight losses with respect to each **diet category** is equal.

H_a : At least one of the mean weight losses with respect to the three **diet category** is

different.

✓ Step 2: Select Appropriate test

This is a problem, concerning three population means. One-way ANOVA is an appropriate test here provided normality and equality of variance assumptions are verified.

One-way ANOVA test

In a one-way ANOVA test, we compare the means from several populations to test if there is any significance difference between them. The results from an ANOVA test are most reliable when the assumptions of normality and equality of variances are satisfied.

- For testing of normality, Shapiro-Wilk's test is applied to the response variable.
- For equality of variance, Levene test is applied to the response variable.

Shapiro-Wilk's test

We will test the null hypothesis

H_0 : The weight losses follow a normal distribution

against the alternative hypothesis

H_a : The weight losses do not follow a normal distribution

```
# Assumption 1: Normality
# Use the shapiro function for the scipy.stats library for this test

# find the p-value
w, p_value = stats.shapiro(df['weight_loss'])
print('The p-value is', p_value)
```

⇒ The p-value is 0.8019888997077942

Since p-value of the test is larger than the 5% significance level, we fail to reject the null hypothesis that the response follows the normal distribution.

Levene's test

We will test the null hypothesis

H_0 : All the population variances are equal

against the alternative hypothesis

H_a : At least one variance is different from the rest

```
#Assumption 2: Homogeneity of Variance
# use levene function from scipy.stats library for this test

# find the p-value
statistic, p_value = stats.levene(df[df['diet']=='A']['weight_loss'],
                                   df[df['diet']=='B']['weight_loss'],
                                   df[df['diet']=='C']['weight_loss'])
print('The p-value is', p_value)
```

➡ The p-value is 0.5376731304274011

Since the p-value is large than the 5% significance level, we fail to reject the null hypothesis of homogeneity of variances.

✓ Step 3: Decide the significance level

As given in the problem statement, we select $\alpha = 0.05$

✓ Step 4: Collect and prepare data

```
# create separate variables to store the weightlosses with respect to the three diet-plans
weightloss_diet_A = df[df['diet']=='A']['weight_loss']
weightloss_diet_B = df[df['diet']=='B']['weight_loss']
weightloss_diet_C = df[df['diet']=='C']['weight_loss']
```

✓ Step 5: Calculate the p-value

- We will use the `f_oneway()` function from the `scipy.stats` library to perform a one-way ANOVA test.
- The `f_oneway()` function takes the sample observations from the different groups and returns the test statistic and the p-value for the test.
 - The sample observations are the values of weight losses with respect to the three diet-plans.

```
# import the required function
from scipy.stats import f_oneway

# find the p-value
test_stat, p_value = f_oneway(weightloss_diet_A, weightloss_diet_B, weightloss_diet_C)
print('The p-value is ', p_value)
```

⇒ The p-value is 0.0032290142385893524

✓ Step 6: Compare the p-value with α

```
# print the conclusion based on p-value
if p_value < 0.05:
    print(f'As the p-value {p_value} is less than the level of significance, we reject the r
else:
    print(f'As the p-value {p_value} is greater than the level of significance, we fail to r
```

⇒ As the p-value 0.0032290142385893524 is less than the level of significance, we reject t



✓ Step 7: Draw inference

Since the p-value is less than the level of significance (5%), we reject the null hypothesis. Hence, we have enough statistical evidence to say that at least one of the mean weight losses with respect to the three diet-plans is different.

✓ Hypothesis - 2

✓ Step 1: Define null and alternative hypotheses for One-Way ANOVA

The null and alternative hypotheses can be formulated as:

H_0 : The mean weight losses with respect to each **age category** is equal.

H_a : At least one of the mean weight losses with respect to the three **age category** is different.

✓ Step 2: Select Appropriate test

This is a problem, concerning three population means. One-way ANOVA is an appropriate test here provided normality and equality of variance assumptions are verified.

One-way ANOVA test

In a one-way ANOVA test, we compare the means from several populations to test if there is any significance difference between them. The results from an ANOVA test are most reliable when the assumptions of normality and equality of variances are satisfied.

- For testing of normality, Shapiro-Wilk's test is applied to the response variable.
- For equality of variance, Levene test is applied to the response variable.
- We've conducted thorough evaluation for test on `weight_loss` and our findings indicate that we cannot reject the null hypothesis for either of these assessments.

Levene's test

We will test the null hypothesis

H_0 : All the population variances are equal

against the alternative hypothesis

H_a : At least one variance is different from the rest

```
#Assumption 2: Homogeneity of Variance
# use levene function from scipy.stats library for this test

# find the p-value
statistic, p_value = stats.levene(df[df['age_category']=='40+']['weight_loss'],
                                  df[df['age_category']=='25-40']['weight_loss'],
                                  df[df['age_category']=='18-25']['weight_loss'])
print('The p-value is', p_value)
```

→ The p-value is 0.12538330776005183

Since the p-value is large than the 5% significance level, we fail to reject the null hypothesis of homogeneity of variances.

✓ Step 3: Decide the significance level

As given in the problem statement, we select $\alpha = 0.05$

✓ Step 4: Collect and prepare data

```
# create separate variables to store the weightlosses with respect to the three diet-plans
weightloss_Elderly = df[df['age_category']=='40+']['weight_loss']
weightloss_Middle_aged = df[df['age_category']=='25-40']['weight_loss']
weightloss_Young = df[df['age_category']=='18-25']['weight_loss']
```

✓ Step 5: Calculate the p-value

- We will use the `f_oneway()` function from the `scipy.stats` library to perform a one-way ANOVA test.
- The `f_oneway()` function takes the sample observations from the different groups and returns the test statistic and the p-value for the test.
 - The sample observations are the values of weight losses with respect to the three age categories.

```
# import the required function
from scipy.stats import f_oneway

# find the p-value
test_stat, p_value = f_oneway(weightloss_Elderly, weightloss_Middle_aged, weightloss_Young)
print('The p-value is ', p_value)
```

➡ The p-value is 0.05544168556142372

✓ Step 6: Compare the p-value with α

```
# print the conclusion based on p-value
if p_value < 0.05:
    print(f'As the p-value {p_value} is less than the level of significance, we reject the r
else:
    print(f'As the p-value {p_value} is greater than the level of significance, we fail to r

➡ As the p-value 0.05544168556142372 is greater than the level of significance, we fail to
```



✓ Step 7: Draw inference

Since the p-value is greater than the level of significance (5%), we fail to reject the null hypothesis. Hence, we have do not enough statistical evidence to conclude that the mean weight losses is different for at least one age category.

✓ Hypothesis - 3

✓ Step 1: Define null and alternative hypotheses for Two-Way ANONA

The null and alternative hypotheses can be formulated as:

H_0 : The effect of diet on weight_loss does not depend on the effect of the age_category variable (a.k.a. no interaction effect)

H_a : There is an interaction effect between diet and age_category on weight_loss .

✓ Step 2: Select Appropriate test

This is a problem, concerning the effect of two independent variables on a dependent variable. **Two-way ANOVA test** is an appropriate test here.

Following are the assumptions of the Two-way ANOVA test:

- The populations from which the samples are obtained must be normally distributed.
- Sampling is done correctly. Observations for within and between groups must be independent.
- The variances among populations must be equal (homoscedastic).
- The dependent data must be measured at an interval scale.

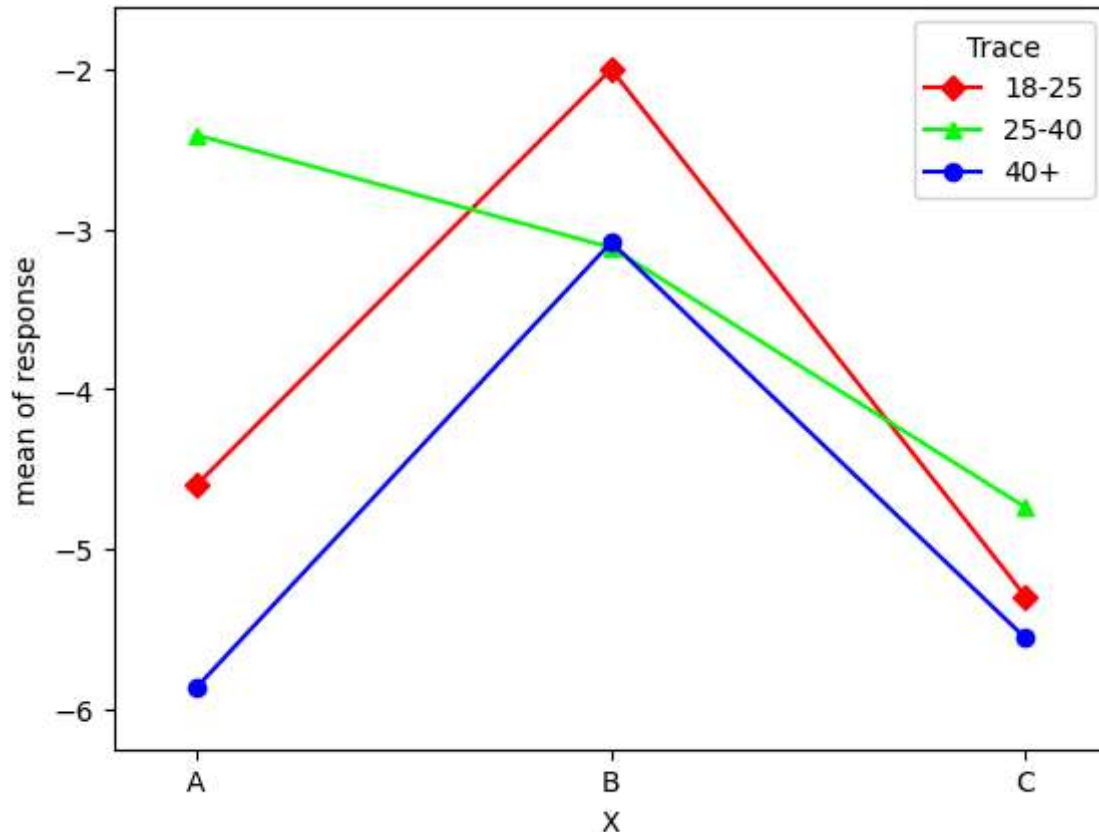
✓ Step 3: Decide the significance level

As given in the problem statement, we select $\alpha = 0.05$

✓ Step 4: Check for Interaction Effect

- We will now analyse the effect of both the diet and age_category on the weight_loss variable.

```
from statsmodels.graphics.factorplots import interaction_plot
interaction_plot(np.array(df['diet']), np.array(df['age_category']), np.array(df['weight_loss'])
```



- We can see that there is some sort of interaction between the diet and age_category.

✓ Step 5: Calculate the p-value

```
formula = 'weight_loss ~ C(diet) + C(age_category) + C(diet):C(age_category)'
model = ols(formula, df).fit()
aov_table = anova_lm(model)
(aov_table)
```



	df	sum_sq	mean_sq	F	PR(>F)
C(diet)	2.0	71.093689	35.546845	6.399140	0.002822
C(age_category)	2.0	17.498000	8.749000	1.574994	0.214359
C(diet):C(age_category)	4.0	29.390330	7.347582	1.322711	0.270226
Residual	69.0	383.290930	5.554941	NaN	NaN

✓ Step 6: Compare the p-value with α


```
row_name = 'C(diet):C(age_category)'
p_value_interaction = aov_table.loc[row_name, 'PR(>F)']

# print the conclusion based on p-value
if p_value_interaction < 0.05:
    print(f'As the p-value {p_value_interaction} is less than the level of significance, we
else:
    print(f'As the p-value {p_value_interaction} is greater than the level of significance,
```