Assessment 01

Introduction to assessment:
You are a data analyst and a small fictional Ice Cream Shop in Rome has asked you do analyse their sales data.
The dataset (`IceCreamSales.csv`) that you have has the next columns:

- product_id: Unique identifier for the product being sold. Ice Cream Type
- date: The specific date when sales data was recorded.
- sales_amount: The amount of sales or revenue generated on the given date.
- temperature: The recorded temperature on the date in Celcius.
- rainfall: The amount of rainfall.
- weather_condition: A description of the prevailing weather conditions on the so we have "sunny" and "not sunny".

Each correct answer will be counted as 1 point and your result will have to match with the results on the pic below. If yes reward yourself 1 point. This assessment is also testing your independent research skills, therefore you are allowed to use Google to find the answers.

Part I

1. Import the data and show the first 5 rows.

	product_id	date	sales_amount	temperature	rainfall	weather_condition
0	2	2023-01-01	86.320006	10	7	not sunny
1	3	2023-01-02	194.868345	17	8	sunny
2	2	2023-01-03	196.927862	15	2	sunny
3	1	2023-01-04	69.325892	12	6	not sunny
4	1	2023-01-05	45.123280	6	0	not sunny

2. Check for missing values for all columns

```
product_id           0
date                 0
sales_amount         2
temperature          0
rainfall             0
weather_condition    0
dtype: int64
```

3. Show where are these nulls

	product_id	date	sales_amount	temperature	rainfall	weather_condition
19	1	2023-01-20	NaN	17	4	sunny
24	2	2023-01-25	NaN	6	6	not sunny

4. Show the descriptive analysis of all values. Focus on sales.

	product_id	sales_amount	temperature	rainfall
count	83.000000	81.000000	83.00000	83.000000
mean	1.927711	95.184456	8.73494	5.144578
std	0.808237	51.995662	5.11355	3.112260
min	1.000000	40.858484	0.00000	0.000000
25%	1.000000	63.683101	5.00000	2.500000
50%	2.000000	78.053205	8.00000	6.000000
75%	3.000000	95.804563	13.00000	7.500000
max	3.000000	285.781178	17.00000	10.000000

5. Show the average sales for each product and average sales for each type of weather condition.
You will get two different outputs.
Hint: GroupBy concept: https://www.geeksforgeeks.org/python-pandas-dataframe-groupby/

```
weather_condition
not sunny     73.954307
sunny        196.785882
Name: sales_amount, dtype: float64

product_id
1     90.692164
2     88.447025
3    108.472978
Name: sales_amount, dtype: float64
```

6. Drop ID Columns and check the columns of df

```
Index(['date', 'sales_amount', 'temperature', 'rainfall',
'weather_condition'], dtype='object')
```

7. Fill the missing values with the mean of the sales with according to the group based on weather condition.Check wether they have really filled with realtive mean?
Hint: GroupBY + Transform : **https://www.statology.org/pandas-groupby-transform/**

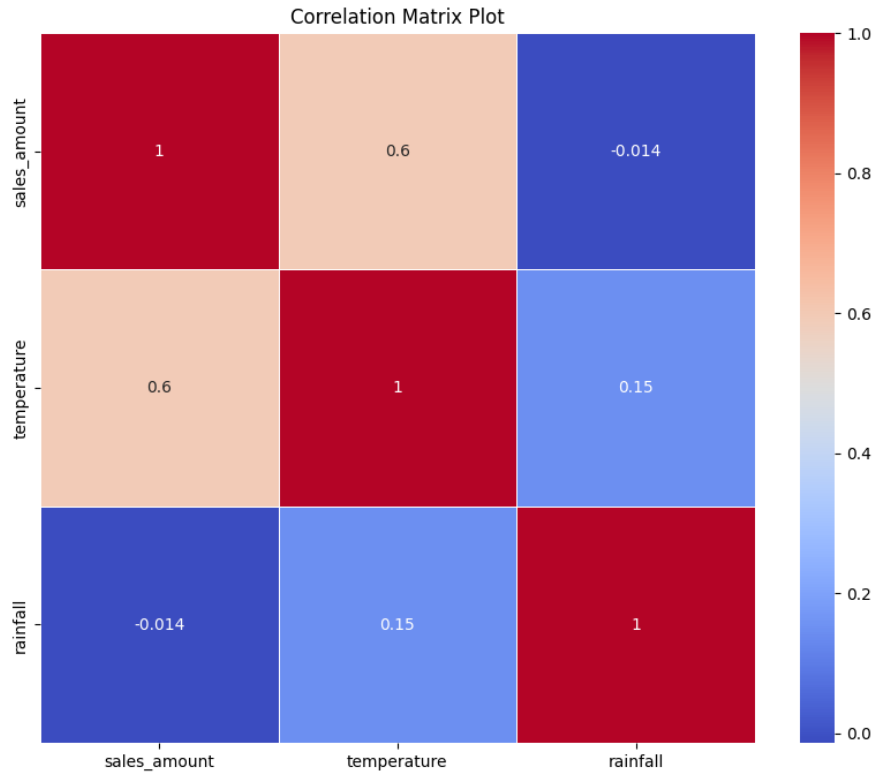
	date	sales_amount	temperature	rainfall	weather_condition
19	2023-01-20	196.785882	17	4	sunny
24	2023-01-25	73.954307	6	6	not sunny

Part II

You will need these libraries

```
import numpy as np #For numeric operations
import pandas as pd #For operations on datasets
import matplotlib.pyplot as plt #For visualization
import seaborn as sns #For visualization
```

8. Plot the correlation plot of the continuous variables.



9. Get dummy variables (1/0) for the weather_condition. Because model doesn't understand the text

Hint: https://pandas.pydata.org/docs/reference/api/pandas.get_dummies.html

	date	sales_amount	temperature	rainfall	weather_condition_sunny
0	2023-01-01	86.320006	10	7	0
1	2023-01-02	194.868345	17	8	1
2	2023-01-03	196.927862	15	2	1
3	2023-01-04	69.325892	12	6	0
4	2023-01-05	45.123280	6	0	0

10. Find a correlation for CONTINUOUS and CATEGORICAL VARIABLE. Sales and weather condition.

Import this libraries

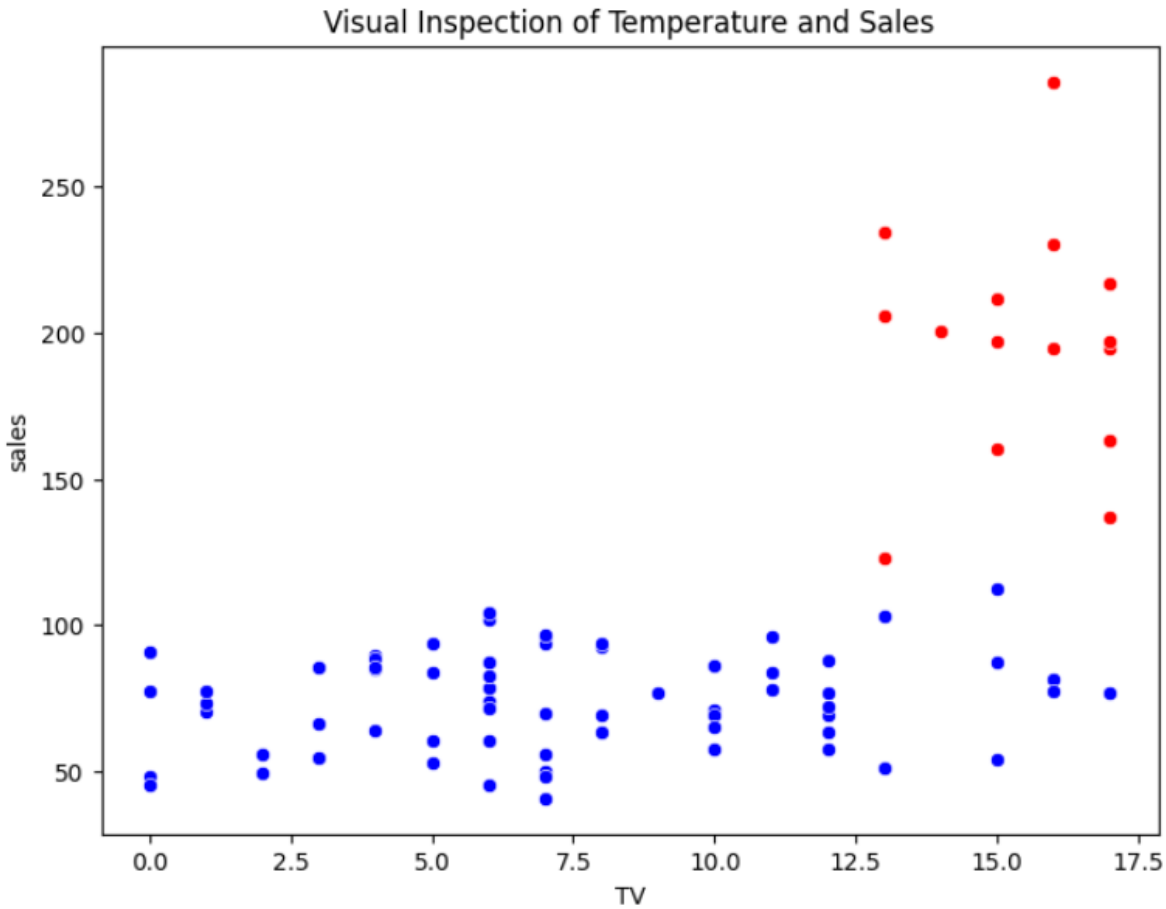
```
from scipy import stats
```

Hint:

https://colab.research.google.com/corgiredirector?site=https%3A%2F%2Fwww.statology.org%2Fcorrelation-between-continuous-categorical-variables%2F

```
(0.9038109438185752, 1.3504179636882083e-31)
```

11. Visual Inspection: Analyse the relationship between TEMPERATURE and SALES colored by WEATHER_CONDITION.



Part III Modeling

12. Split into train and test. Train will have the columns

```
from sklearn.model_selection import train_test_split
```

Check the next `print(len(X_train), len(y_train),len(X_test),len(y_test) )`

66 66 17 17 **Do you have the same results?**

13. Run the linear regression model of statsmodels with X_train and y_train

```
import statsmodels.api as sm
```

```
                                 OLS Regression Results
=======================================================================================
Dep. Variable:           sales_amount   R-squared (uncentered):                  0.901
Model:                            OLS   Adj. R-squared (uncentered):             0.896
Method:                 Least Squares   F-statistic:                             191.1
Date:                Tue, 24 Oct 2023   Prob (F-statistic):                   1.41e-31
Time:                        10:29:45   Log-Likelihood:                         -327.81
No. Observations:                  66   AIC:                                      661.6
Df Residuals:                      63   BIC:                                      668.2
Df Model:                           3
Covariance Type:            nonrobust
==========================================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------
temperature                4.7498      0.854      5.561      0.000       3.043       6.456
rainfall                   5.0585      1.208      4.186      0.000       2.644       7.473
weather_condition_sunny  103.8516     13.918      7.462      0.000      76.039     131.664
==============================================================================
Omnibus:                        0.563   Durbin-Watson:                   1.688
Prob(Omnibus):                  0.755   Jarque-Bera (JB):                0.130
Skew:                           0.002   Prob(JB):                        0.937
Kurtosis:                       3.218   Cond. No.                         35.4
==============================================================================
```

14. Referring to the results of OLS/Linear regression model above.

-What is the Adjusted R-Squared of your model? Is it good/bad? What does it indicate.

-Which variables (coefficents) are significant