|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| WELCOME: UKM FTSM | | | | | | |
|  | | | GROUP [ 17 ] | |  | |
|  | |  | | | |  |
|  |  | | |  | | |
|  | Team member:  |  |  | | --- | --- | | Matric number | Name | | A197547 | Wu yuntian | | A191629 | WANG ZEDONG | | A184982 | ZHAO YANHAO | | A 206300 | Muhammad Zahran | | | |  | | |
|  | 9TH APRIL 2024 Date of submission —TTTC2453 MACHINE LEARNING—DR. WANDEEP KAUR | | |  | | |

**LAB FORMAT**

FONT: CAMBRIA (12

PARAGRAPH: FOLLOW AS SHOWN IN FIG 1

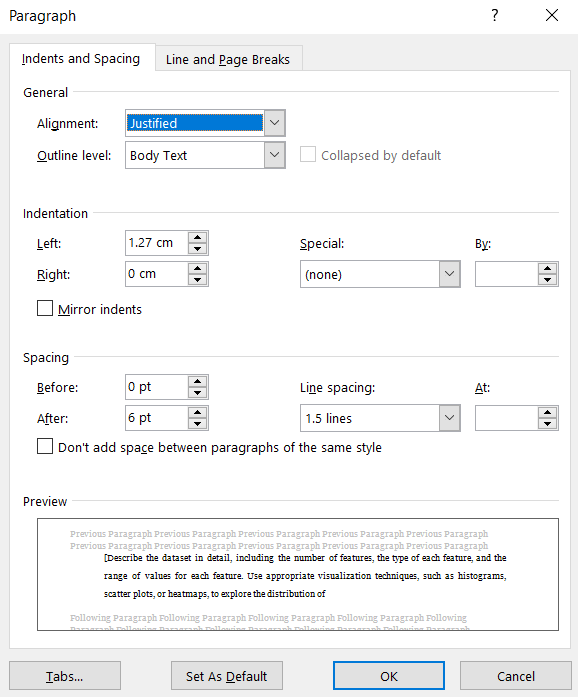


Figure 1: Paragraph Spacing

Content of report should cover the following:

1. **Data Exploration and Visualization**

[Describe the dataset in detail, including the number of features, the type of each feature, and the range of values for each feature. Use appropriate visualization techniques, such as histograms, scatter plots, or heatmaps, to explore the distribution of each feature and its relationship with the target variable. Identify any outliers or missing values and discuss how you handled them.]

1. **Data Preprocessing and Feature Engineering**

[Describe the steps you took to preprocess the dataset, such as splitting the data into training and testing sets, scaling or normalizing the features, and handling missing values and outliers. Also describe any feature engineering you performed, such as creating new features based on domain knowledge or interaction terms between existing features.]

1. **Model Evaluation**

[Provide explanation on results obtained for model evaluation.]

1. **Discussion**

[Answer each discussion question here]

1. **Conclusion**
2. **References**

[List any external sources, such as papers or blog posts, that you used in your analysis. Also cite any Python libraries or functions you used, such as scikit-learn or pandas.]

**IMPORTANT**

1. Name your lab report as:

TC2453\_Wxx\_Gxx.pdf

1. Place your PDF file + code file into a folder and zip it
2. Name the folder as:

Gxx.zip

1. Upload to UKMFolio
2. Only ONE member of the group will submit.

1 group = 1 submission ONLY

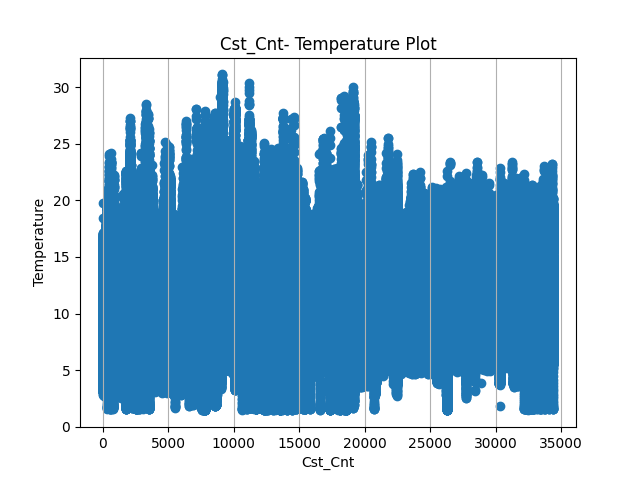
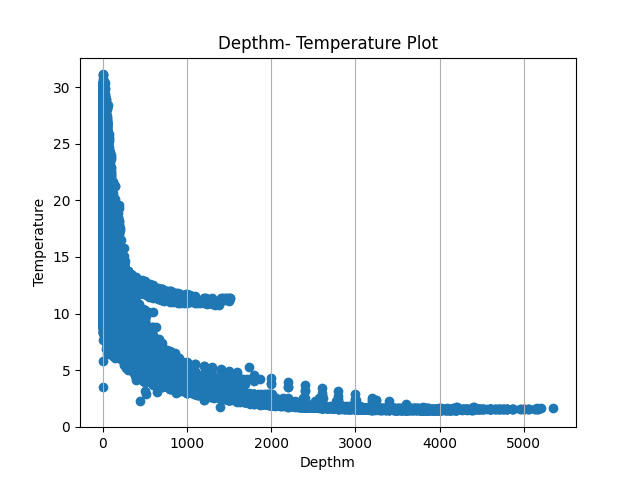
1. **Data Exploration and Visualization**

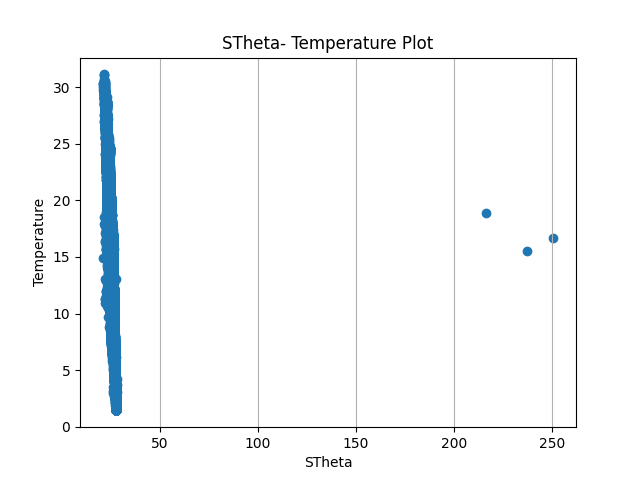
Bottle.csv is a csv file of oceanographic data, measured and complied by [CALCOFI](https://calcofi.org/data/oceanographic-data/bottle-database/).

Bottle csv contains 864863 rows , 74 columns : 5 int64 cols, 65 float64 cols, 4 object cols. The temperature cols used in model is “T\_degC” of index 6. The others cols of dtype int and float are used to create linear regression model with temperature and generate model attributes.

The Pearson correlation of coefficient (r), indicate how strong the relationship between cols and temperature. To reduce cols numbers, only cols with rvalue meet r>0.8 and r < -0.8 are selected in further model , also cols with little data are excluded.

A screenshot of a computer screen

Description automatically generated  A graph of a temperature plot

Description automatically generated with medium confidence

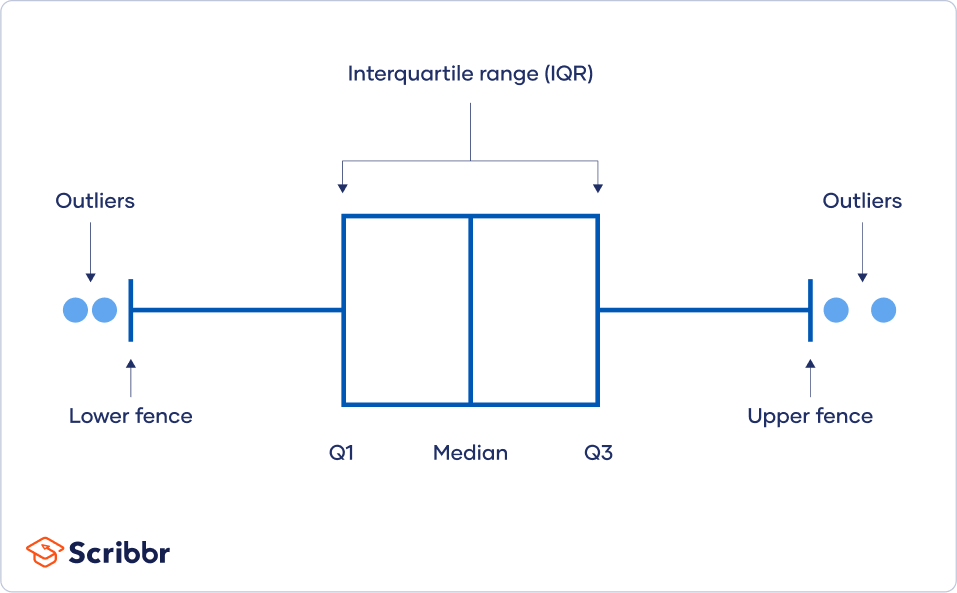
A screenshot of a computer code

Description automatically generated

1. **Data Preprocessing and Feature Engineering**

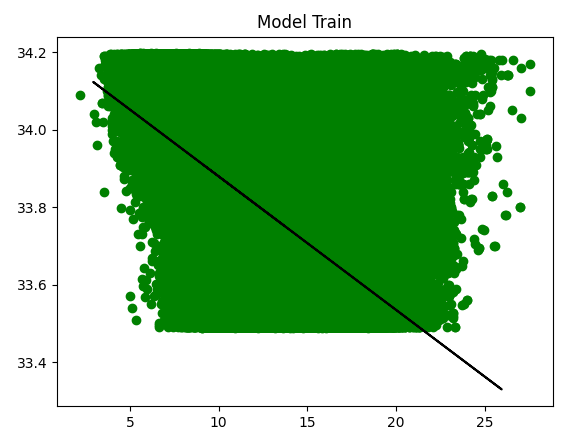
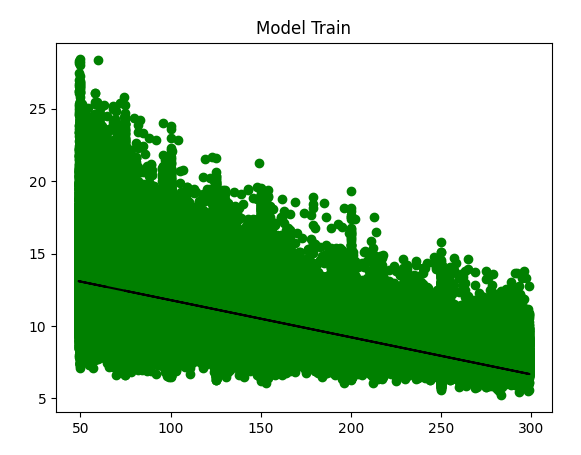
The First step to process data is dropna() and drop\_duplicate() to minimize impact on data accuracy.

Next ,drop outlier using 1.5 \* IQR . The IQR is the difference between the 75th and 25th .  low outliers are below Q1 – 1.5 \* IQR  and high outliers are above  Q3 – 1.5 \* IQR

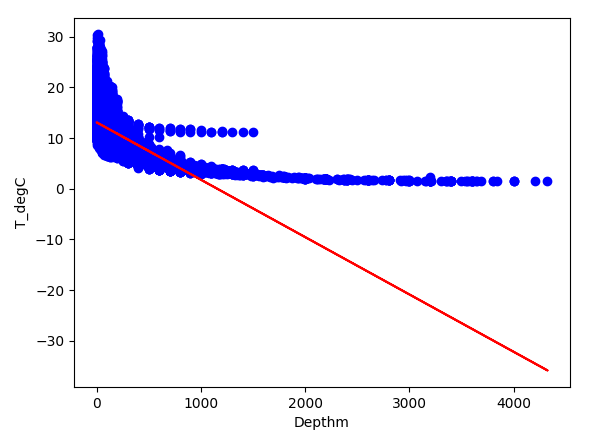
 A white text with black numbers

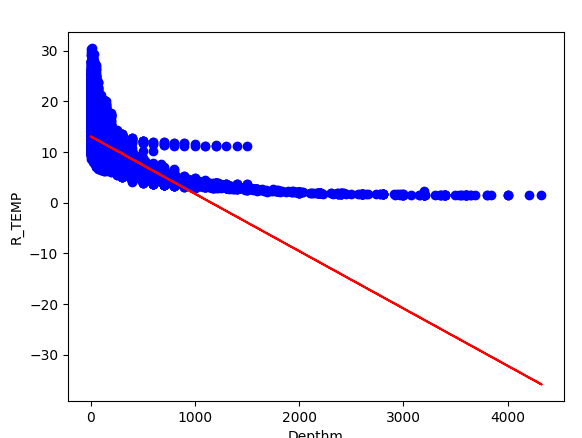
Description automatically generated

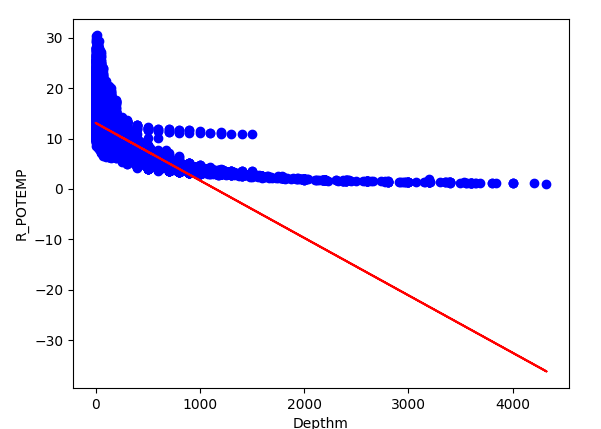
1. **Model Evaluation**

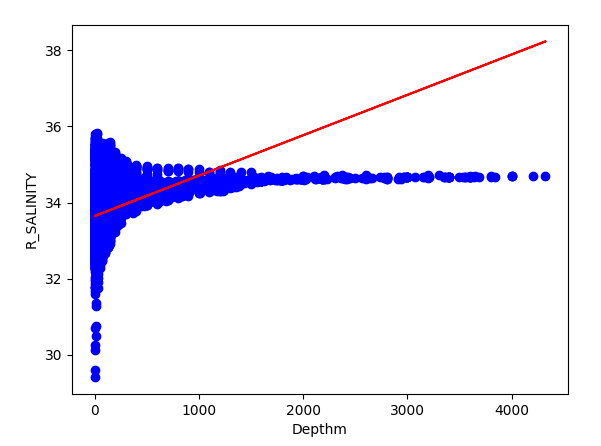


1. **Model Evaluation**









Variable: T\_degC

Coefficients: -0.011330306528971584

Intercept: 13.117745156728345

Mean absolute error: 2.11

Mean sum of squares (MSE): 7.73

R2-score: 0.55

Variable: R\_TEMP

Coefficients: -0.011330306528971584

Intercept: 13.117745156728345

Mean absolute error: 2.11

Mean sum of squares (MSE): 7.73

R2-score: 0.55

Variable: R\_POTEMP

Coefficients: -0.01140823639773838

Intercept: 13.114147829025693

Mean absolute error: 2.11

Mean sum of squares (MSE): 7.74

R2-score: 0.56

Variable: R\_SALINITY

Coefficients: 0.0010622144927140177

Intercept: 33.64492090248298

Mean absolute error: 0.26

Mean sum of squares (MSE): 0.13

R2-score: 0.39

1. **Visualization**

This picture shows the relationship between Salinity and Temperature. The scatter plot shows the distribution of salinity and temperature in the training data, with the blue scatter representing the actual observations. However, the red line is the fitted line obtained by predicting the test data by the linear regression model. Overall, the graph shows a general trend relationship between salinity and temperat

1. **Discussion**

**i)**

**Which environmental variables have the most significant**

**impact on sea surface temperature according to the**

**regression model, and how confident are you in these**

**findings? Provide statistical evidence to support your**

**conclusions.**

**R value which is closest to 1/-1**

**ii)**

**What is the overall performance of the linear regression**

**model in predicting sea surface temperature? Discuss key**

**performance metrics and interpret the model's ability to**

**accurately capture variations in sea surface temperature.**

**Additionally, identify any limitations or areas for**

**improvement.**

**Model evaluation , test th emodel , an d xmpare the output**

**iii)**

**Identify any outliers in the dataset that may influence the**

**linear regression model's predictions. Assess the impact of**

**these outliers on the model's coefficients and discuss**

**potential strategies for handling outliers. How might**

**addressing outliers enhance the model's reliability?**

**Data preprocess ,**

1. **Conclusion**

**Jack.**

1. **References**