# Topic: Simple Linear Regression

**Instructions:**

Please share your answers filled in-line in the word document. Submit code separately wherever applicable.

Please ensure you update all the details:

**Name: Vishvash C Batch ID:** 23012024

**Topic: Simple Linear Regression**

**Guidelines:**

**1. An assignment submission is considered complete only when the correct and executable code(s) and documentation explaining the method and results are submitted. Failing to submit either of those will be considered an invalid submission and not a correct submission.**

**2. Ensure that you submit your assignments correctly and in full. Resubmission is not allowed.**

**3. Post the submission you can evaluate your work by referring to the keys provided. (will be available only post the submission).**

**Hints:**

1. **Business Problem**
   1. **What is the business objective?**
   2. **Are there any constraints?**
2. **Work on each feature of the dataset to create a data dictionary as displayed in the image below:**



**Make a table as shown above and provide information about the features such as its data type and its relevance to the model building. And if not relevant, provide reasons and a description of the feature.**

1. **Data Pre-processing.**
   1. **Data Cleaning, Feature Engineering, etc.**

**3.2. Outlier Treatment.**

1. **Exploratory Data Analysis (EDA):**

**4.1 Summary.**

**4.2 Univariate analysis.**

**4.3 Bivariate analysis.**

1. **Model Building:**
   1. **Perform Simple Linear Regression on the given datasets.**
   2. **Apply different transformations such as exponential, log, polynomial, etc. transformations and calculate RMSE values, R-Squared values, and the correlation coefficient for each model.**
   3. **Build the models and choose the best-fit model.**
   4. **Briefly explain the model output in the documentation.**
2. **Write about the benefits/impact of the solution - in what way does the business (client) benefit from the solution provided**

**Problem Statement: -**

1. A logistics company recorded the time taken for delivery and the time taken for the sorting of the items for delivery. Build a Simple Linear Regression model to find the relationship between delivery time and sorting time with the delivery time as the target variable. Apply necessary transformations and record the RMSE and correlation coefficient values for different models.



|  |  |  |  |
| --- | --- | --- | --- |
| Name of Feature | Description | Type | Relevance |
| Delivery Time | Time taken for delivery | Quantitative | Relevant |
| Sorting Time | Time taken for sorting | Quantitative | Relevant |

**Code:**

**Output:**

delv.head()

Out[24]:

Delivery\_Time Sorting\_Time

0 21.00 10

1 13.50 4

2 19.75 6

3 24.00 9

4 29.00 10

delv.describe()

Out[25]:

Delivery\_Time Sorting\_Time

count 21.000000 21.000000

mean 16.790952 6.190476

std 5.074901 2.542028

min 8.000000 2.000000

25% 13.500000 4.000000

50% 17.830000 6.000000

75% 19.750000 8.000000

max 29.000000 10.000000

Delivery\_Time Sorting\_Time

0 8.00 2.0

1 9.50 3.0

2 11.50 3.0

3 12.03 3.0

4 13.50 4.0

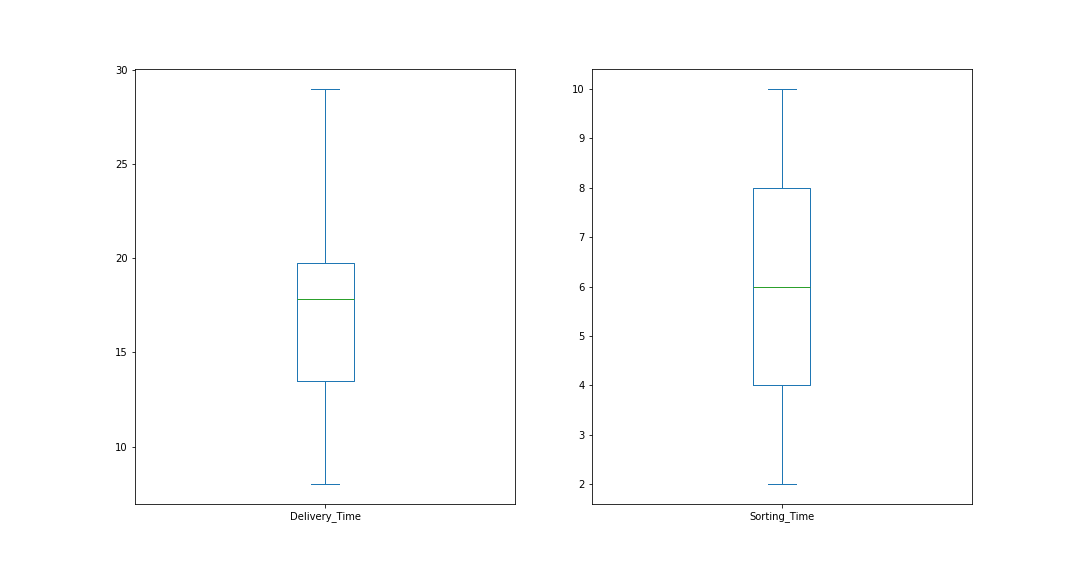
5 14.88 4.0

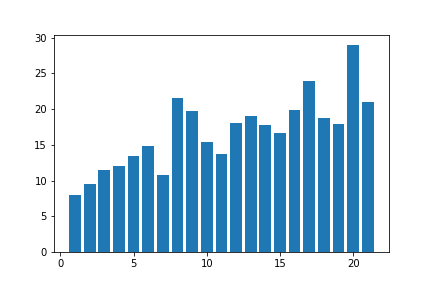
6 10.75 4.0

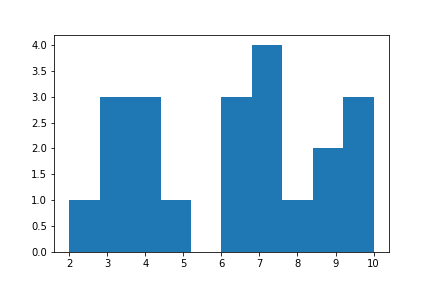
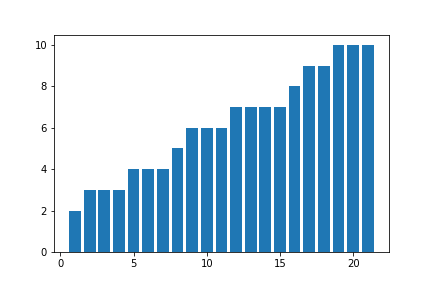
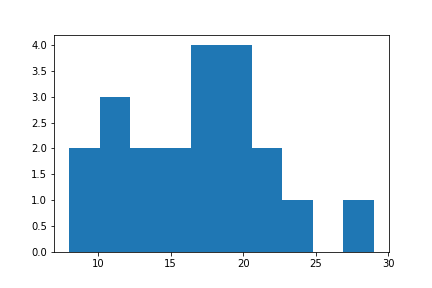
7 21.50 5.0

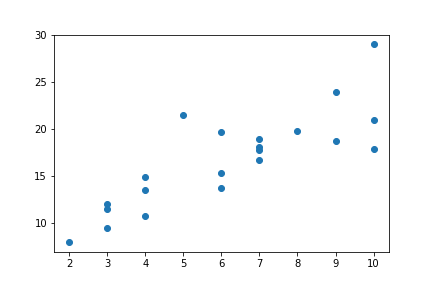
8 19.75 6.0

9 15.35 6.0







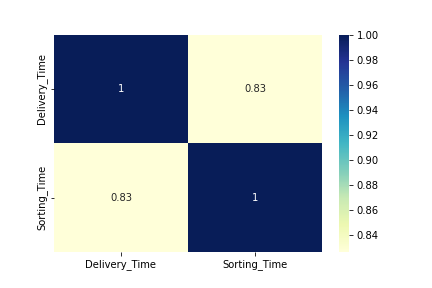


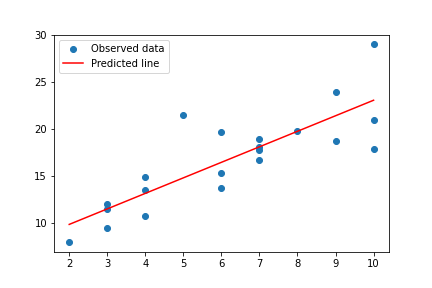
array([[1. , 0.82599726],

[0.82599726, 1. ]])

cov\_output

Out[107]: 10.655809523809523





<class 'statsmodels.iolib.summary.Summary'>

"""

OLS Regression Results

==============================================================================

Dep. Variable: Delivery\_Time R-squared: 0.682

Model: OLS Adj. R-squared: 0.666

Method: Least Squares F-statistic: 40.80

Date: Fri, 05 Apr 2024 Prob (F-statistic): 3.98e-06

Time: 12:09:29 Log-Likelihood: -51.357

No. Observations: 21 AIC: 106.7

Df Residuals: 19 BIC: 108.8

Df Model: 1

Covariance Type: nonrobust

================================================================================

coef std err t P>|t| [0.025 0.975]

--------------------------------------------------------------------------------

Intercept 6.5827 1.722 3.823 0.001 2.979 10.186

Sorting\_Time 1.6490 0.258 6.387 0.000 1.109 2.189

==============================================================================

Omnibus: 3.649 Durbin-Watson: 2.322

Prob(Omnibus): 0.161 Jarque-Bera (JB): 2.086

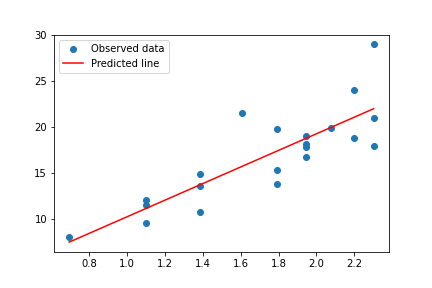
Skew: 0.750 Prob(JB): 0.352

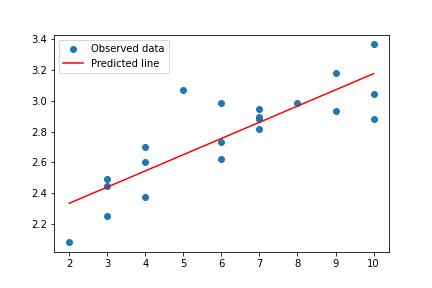
Kurtosis: 3.367 Cond. No. 18.3

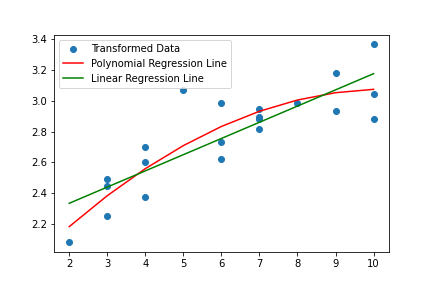
==============================================================================

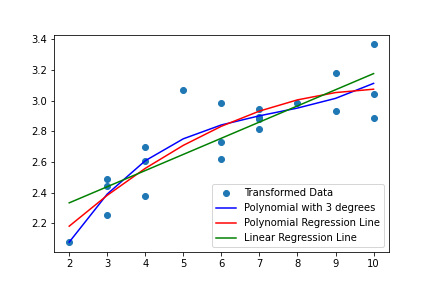
Notes:

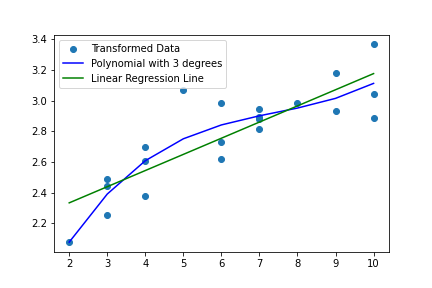
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.











|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Transformation | Y | X | R^2 | RMSE | Issue | Solve |
| Linear Regression | Delivery\_Time | Sorting\_Time | 0.682 | 2.79 | Heteroscedasticity | Transformation |
| Logarithmic Regression | Delivery\_Time | log(Sorting\_Time) | 0.695 | 2.73 | Heteroscedasticity | Transformation |
| Exponential Regression | log(Delivery\_Time) | Sorting\_Time | 0.711 | 2.94 | Non-linearity | Transformation |
| Quadratic Regression | log(Delivery\_Time) | Sorting\_Time, Sorting\_Time^2 | 0.765 | 2.79 | Non-linearity | Transformation |
| Polynomial regression with 3 degrees | log(Delivery\_Time) | Sorting\_Time, Sorting\_Time^2, Sorting\_Time^3 | 0.782 | 2.7 |  |  |

MODEL RMSE

0 SLR 2.791650

1 Log model 2.733171

2 Exp model 2.940250

3 Poly model 2.799042

4 Poly 3 deg 2.706757

final

Out[232]:

Pred\_Delivery\_Time Sorting\_Time

0 22.797240 10

1 13.892856 4

2 17.150048 6

3 20.827000 9

4 22.797240 10

5 17.150048 6

6 18.270140 7

7 11.222284 3

8 22.797240 10

9 20.827000 9

10 19.406879 8

11 13.892856 4

12 18.270140 7

13 11.222284 3

14 11.222284 3

15 13.892856 4

16 17.150048 6

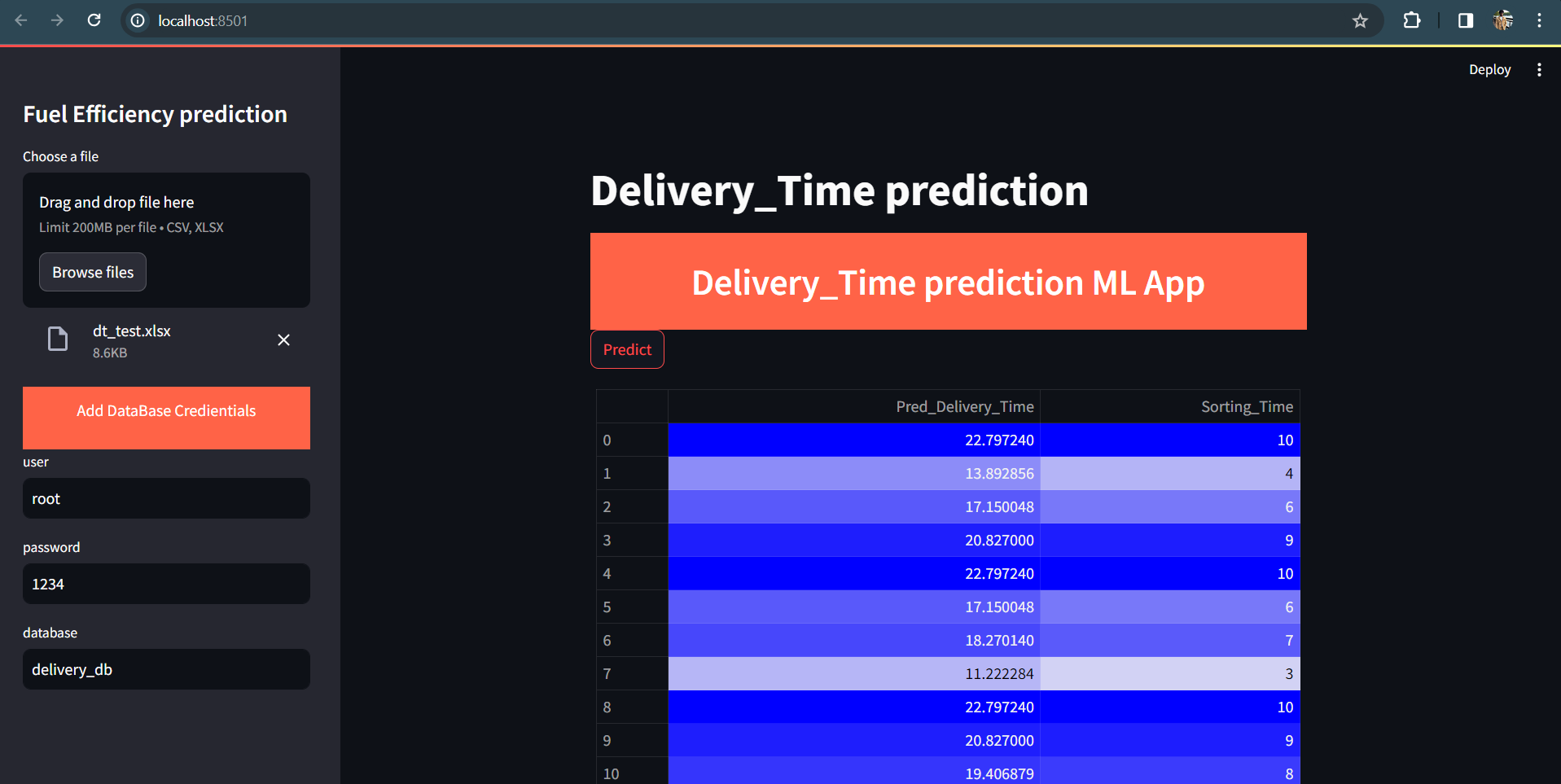
17 18.270140 7

18 7.501412 2

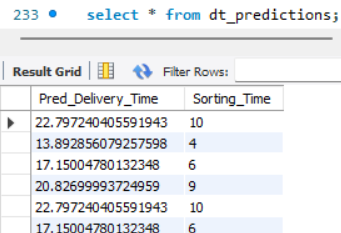
19 18.270140 7

20 15.779865 5

**Deployment of Delivery Time Prediction with Simple linear regression model using Streamlit**

****

**Saving the Predicted Results in MySQL for monitoring**

****