

E11: Data Analysis of a Chemical Processing Plant

GROUP 40:

HARDIK GOHIL - 22B1293

MADHAVA SRIRAM - 22B1233



Exploratory Data Analysis

I) IDENTIFYING MISSING VALUES:

- 1) The columns 'c82', 'c88', 'c96', 'c97', 'c102', 'c110', 'c113', 'c133', 'c149', 'c188', 'c189', 'c190', 'c206' have cells occupied by **#REF!** reference error values. This typically happens when a formula refers to a cell or range of cells that has been deleted or moved.
- 2) The columns 'c133', 'c149', 'c128', 'c130' have cells occupied by **#VALUE!** Error. It indicates that Excel cannot interpret one or more of the values used in the formula.

3) The columns 'c206','c207','c208','c209', 'c210', 'c211', 'c212', 'c213', 'c214', 'c215', 'c216', 'c217', 'c218', 'c219', 'c220', 'c221', 'c222' have cells occupied by #N/A error.

4) The columns 'c199', 'c202', 'c204', 'c212', 'c222', 'c223', 'c226', 'c229', 'c230', 'c231', 'c232', 'c233', 'c234' have empty cells.

II) CONVERTING ALL ERROR VALUES TO NaN:

1) Replacing all the error or empty values with **NaN** to be able to have a better view and understanding of all the missing data as one value 'NaN'

III) CREATED A COPY OF THE DATAFRAME SO AS TO NOT LOSE THE ACTUAL INFORMATION

IV) : DROPPING THE COLUMNS WHICH HAVE MORE THAN 50 PERCENT MISSING VALUES

1) The columns ['c188', 'c189', 'c190', 'c199', 'c202', 'c204', 'c206', 'c223', 'c226', 'c229', 'c231', 'c232', 'c233', 'c234'] have more than 50 percent missing values.

2) It would not be a good model to predict the remaining more than 50 % data based on the available less than 50 % data

V) : IDENTIFYING OUTLIERS IN DATA

1) Used the **Inter-Quartile Range** method to find the Upper Bound and Lower Bound for all the columns and made dataframes for UB and LB.

$Q1 = 25^{\text{TH}}$ PERCENTILE OF THE DATA

$Q3 = 75^{\text{TH}}$ PERCENTILE OF THE DATA

$IQR = Q3 - Q1$

$LB = Q1 - 1.5 * IQR$

$UB = Q3 + 1.5 * IQR$

2) Classified the cells which have values greater than Upper Bound and less than Lower Bound of that column as outliers in that column.

VI) : REPLACED OUTLIERS WITH NaN

1) Replaced the outliers with NaN so that they also have the same value as the other cells whose value is to be predicted.

Filling Missing Values

- 1) For each column except c1, c51, c52, c53, making the numeric values as Train data and the NaN values as test data. (Made sure that there are no other types of values in our data at this point).
- 2) Trained Polynomial regression model of degree 100 using the Train data i.e numeric values.
- 3) Used this model to predict the values in the test data i.e NaN values.
- 4) Assigned the predicted values to the corresponding cell in the copy dataframe.

COLLINEARITY DETECTION

- 1) Calculated the **Variance Inflation Factor (VIF)** for all the columns except c1, c51, c52, c53, c54.
- 2) Filtered out the columns who had $VIF > 20$.
- 3) $VIF > 20$ indicates a high level of collinearity and hence, removed all the columns with $VIF > 20$ from the dataframe.
- 4) This resulted in dropping of about 159 columns from the dataframe.
- 5) These columns were not very important for the prediction of vibrations.

DATA VISUALIZATION

- 1) Made a plot of all the problematic columns before and after predicting the values
- 2) This gave insight into the differences that were created due to outlier removal and polynomial regression
- 3) This helped us to verify whether the EDA has been performed correctly or not.
- 4) Most of the plots were as expected.

Creating ML model for c51, c52, c53, c54 using all columns

OLS Regression Results

```
=====
Dep. Variable:          c51    R-squared:                0.673
Model:                  OLS    Adj. R-squared:           0.655
Method:                 Least Squares    F-statistic:           37.00
Date:                  Mon, 13 Nov 2023    Prob (F-statistic):    3.25e-197
Time:                  14:01:12    Log-Likelihood:       -1909.1
No. Observations:      1025    AIC:                   3928.
Df Residuals:          970    BIC:                   4200.
Df Model:              54
Covariance Type:      nonrobust
=====
```

```
=====
Dep. Variable:          c52    R-squared:                0.831
Model:                  OLS    Adj. R-squared:           0.821
Method:                 Least Squares    F-statistic:           88.03
Date:                  Mon, 13 Nov 2023    Prob (F-statistic):    0.00
Time:                  14:01:12    Log-Likelihood:       -1362.2
No. Observations:      1025    AIC:                   2834.
Df Residuals:          970    BIC:                   3106.
Df Model:              54
Covariance Type:      nonrobust
=====
```

OLS Regression Results			
=====			
Dep. Variable:	c53	R-squared:	0.884
Model:	OLS	Adj. R-squared:	0.878
Method:	Least Squares	F-statistic:	137.5
Date:	Mon, 13 Nov 2023	Prob (F-statistic):	0.00
Time:	14:01:12	Log-Likelihood:	-2251.2
No. Observations:	1025	AIC:	4612.
Df Residuals:	970	BIC:	4884.
Df Model:	54		
Covariance Type:	nonrobust		
=====			

Dep. Variable:	c54	R-squared:	0.876
Model:	OLS	Adj. R-squared:	0.869
Method:	Least Squares	F-statistic:	127.2
Date:	Mon, 13 Nov 2023	Prob (F-statistic):	0.00
Time:	14:01:12	Log-Likelihood:	-2187.5
No. Observations:	1025	AIC:	4485.
Df Residuals:	970	BIC:	4756.
Df Model:	54		
Covariance Type:	nonrobust		
=====			

NOTE : These models were trained on data obtained after dropping the collinear columns

Creating ML model for c51, c52, c53, c54 using controllable parameters

OLS Regression Results

```
=====
Dep. Variable:          c51    R-squared (uncentered):      0.959
Model:                  OLS    Adj. R-squared (uncentered):    0.959
Method:                 Least Squares    F-statistic:          1249.
Date:                  Mon, 13 Nov 2023    Prob (F-statistic):      0.00
Time:                  14:01:12    Log-Likelihood:        -2150.2
No. Observations:      1025    AIC:                   4338.
Df Residuals:          1006    BIC:                   4432.
Df Model:              19
Covariance Type:       nonrobust
=====
```

```
=====
Dep. Variable:          c52    R-squared (uncentered):      0.983
Model:                  OLS    Adj. R-squared (uncentered):    0.983
Method:                 Least Squares    F-statistic:          3155.
Date:                  Mon, 13 Nov 2023    Prob (F-statistic):      0.00
Time:                  14:01:12    Log-Likelihood:        -1645.0
No. Observations:      1025    AIC:                   3328.
Df Residuals:          1006    BIC:                   3422.
Df Model:              19
Covariance Type:       nonrobust
=====
```

OLS Regression Results

```

=====
Dep. Variable:          c53      R-squared (uncentered):          0.948
Model:                  OLS      Adj. R-squared (uncentered):        0.947
Method:                 Least Squares      F-statistic:              964.1
Date:                  Mon, 13 Nov 2023      Prob (F-statistic):         0.00
Time:                  14:01:12      Log-Likelihood:           -2463.3
No. Observations:      1025      AIC:                      4965.
Df Residuals:          1006      BIC:                      5058.
Df Model:              19
Covariance Type:       nonrobust
=====

```

```

=====
Dep. Variable:          c54      R-squared (uncentered):          0.937
Model:                  OLS      Adj. R-squared (uncentered):        0.936
Method:                 Least Squares      F-statistic:              784.2
Date:                  Mon, 13 Nov 2023      Prob (F-statistic):         0.00
Time:                  14:01:12      Log-Likelihood:           -2485.1
No. Observations:      1025      AIC:                      5008.
Df Residuals:          1006      BIC:                      5102.
Df Model:              19
Covariance Type:       nonrobust
=====

```

c51

[illegible]

[illegible]

[illegible]

c54

[illegible]

Sort the p-values in ascending order for each model

c51

Variable	Coefficient	P-value
c161	2.075398e-02	6.633554e-26
c158	3.115210e-01	2.246108e-17
c39	1.300734e+01	1.539583e-12
c156	1.302331e-14	2.915443e-09
c27	6.476557e-05	2.267978e-04
c143	5.596057e-03	1.778786e-03
c28	6.952627e-02	9.093442e-03
c31	-6.505412e-02	1.565456e-02
c30	1.560065e+00	2.283270e-02
c32	3.001348e-01	2.601228e-02
c157	-9.015735e-02	1.148564e-01
c142	7.056150e-02	1.170523e-01
c26	-8.898137e-02	1.911171e-01
c33	2.200611e-01	4.276409e-01
c139	4.072722e-02	4.600851e-01
c160	-2.056615e-03	5.504104e-01
c29	3.519571e-02	6.021834e-01
c155	-5.159549e-03	7.538032e-01
c162	-4.314907e-04	8.839341e-01
c163	8.646758e-05	9.821910e-01

c52

Variable	Coefficient	P-value
15	c158	3.601033e-01 1.985080e-53
17	c161	1.253529e-02 2.160394e-25
3	c29	-3.679800e-01 2.073346e-18
2	c28	1.407381e-01 1.835697e-17
8	c39	8.674654e+00 1.341318e-14
0	c26	2.989787e-01 1.212993e-12
7	c33	1.079346e+00 2.837637e-10
13	c156	8.378545e-15 4.202238e-10
14	c157	2.072156e-01 3.972648e-09
9	c139	-1.835047e-01 6.302097e-08
1	c27	5.527932e-05 2.811771e-07
5	c31	7.815165e-02 2.205135e-06
6	c32	-3.818035e-01 3.893964e-06
11	c143	5.003893e-03 5.071776e-06
4	c30	1.893196e+00 6.634028e-06
12	c155	-3.974052e-02 8.164644e-05
19	c163	9.290505e-03 9.173242e-05
10	c142	-6.441022e-02 1.926699e-02
16	c160	3.737062e-03 7.587779e-02
18	c162	1.262280e-03 4.845180e-01

c53

Variable	Coefficient	P-value
12	c155	5.178470e-01 1.034787e-95
19	c163	5.646324e-02 1.468468e-25
15	c158	3.630598e-01 2.557733e-13
2	c28	2.412885e-01 3.867750e-11
1	c27	1.503051e-04 3.738285e-10
14	c157	4.201907e-01 7.492538e-08
11	c143	1.238280e-02 3.884922e-07
13	c156	-1.036167e-14 4.652165e-04
8	c39	-7.877056e+00 1.439561e-03
17	c161	6.285676e-03 1.589226e-02
4	c30	1.943844e+00 3.660463e-02
9	c139	-1.293465e-01 8.406650e-02
7	c33	5.991213e-01 1.117304e-01
0	c26	-1.315350e-01 1.545200e-01
3	c29	1.275863e-01 1.640119e-01
16	c160	6.297493e-03 1.780548e-01
18	c162	-4.820621e-03 2.296489e-01
6	c32	1.583405e-01 3.863826e-01
5	c31	2.285048e-03 9.500522e-01
10	c142	1.826447e-04 9.976136e-01

c54

Variable	Coefficient	P-value
12	c155	3.226652e-01 1.353274e-41
15	c158	6.217864e-01 4.050970e-33
19	c163	5.434100e-02 5.371011e-23
17	c161	1.691118e-02 3.021425e-10
2	c28	2.320472e-01 4.672364e-10
13	c156	-1.426820e-14 2.519920e-06
14	c157	3.472687e-01 1.284995e-05
7	c33	1.677560e+00 1.412415e-05
4	c30	3.923866e+00 3.830761e-05
8	c39	-1.017901e+01 5.695392e-05
1	c27	9.421859e-05 1.098802e-04
11	c143	8.362206e-03 7.604396e-04
5	c31	8.128869e-02 2.933691e-02
9	c139	-1.628563e-01 3.329400e-02
10	c142	-1.277989e-01 4.069818e-02
16	c160	8.693567e-03 6.885421e-02
6	c32	-2.080862e-01 2.651805e-01
18	c162	-2.352401e-03 5.659537e-01
3	c29	4.669651e-02 6.178847e-01
0	c26	3.237333e-02 7.314491e-01

Making the model for c241 sorting the p values in ascending order

	Variable	Coefficient	P-value
8	c20	0.086403	6.272224e-04
10	c22	-0.144736	4.062602e-09
11	c23	0.053989	2.237151e-02
12	c30	-0.603754	1.179912e-02
17	c42	0.358678	4.847498e-03
18	c44	-0.186047	3.084779e-04
21	c60	-0.246794	1.862429e-02
23	c62	-0.238910	3.316080e-02
25	c72	-0.152748	2.925802e-02
33	c137	-0.127683	7.348400e-04
39	c152	16.474173	1.336661e-04
53	c178	0.414866	8.746317e-03
59	c230	0.093507	2.802967e-02

13

OLS Regression Results

```
=====
Dep. Variable:          c241      R-squared (uncentered):      0.910
Model:                  OLS      Adj. R-squared (uncentered):    0.909
Method:                  Least Squares      F-statistic:          785.6
Date:                    Mon, 13 Nov 2023      Prob (F-statistic):      0.00
Time:                    14:01:13      Log-Likelihood:        -1088.9
No. Observations:        1025      AIC:                   2204.
Df Residuals:            1012      BIC:                   2268.
Df Model:                 13
Covariance Type:          nonrobust
=====
```