DAT-A-VENGERS

1.Total credits given to you for the whole game is $**75,000**.
2.On every new step you do, there is a prticular amount of credits that would be deducted from the total. *Every click may take some time due to server issues so be patient after each click*.
3. The Dataset is available to you only for initial **5 minutes** so make the most of it
4. Always before performing any function , Check the credit value it would.
5. Read On

So lets get to work !!!

Options you are getting are:-

NULL VALUES

1. Column Wise NULL (CREDITS COST:- $**1500**)
-It will give you a list of total no. of columns with True(if there is a null element in the column) OR False(if the column is full).

2.Number Of Null in Each Column (CREDITS COST:- $**300**)
-You have to choose the column in which you want to check the total number of null elements

3. Number of Columns with NULL (CREDITS COST:- $**800**)
-This will give the total number of columns which have atleast one empty data.

NORMALIZATION

Here, you have to select a column name and a way of normalistion
.

$$S = \sqrt{\frac{\Sigma(X - \overline{X})^2}{N}}$$

where S = the standard deviation of a sample,
Σ means "sum of,"
X = each value in the data set,
$\overline{X}$ = mean of all values in the data set,
N = number of values in the data set.

1ST Way (CREDITS COST:- $400)

Every element (x) will be replaced by **(x - Mean) / Deviation.**

Where “Mean” is the mean of the data of whole column.

“Deviation” is the standard deviation of the data of the whole column.

2ND Way (CREDITS COST:- $400)

Every element(x) will be replaced by **(x - Mean)**

Where Mean is the mean of the data of whole column.

3RD Way (CREDITS COST:- $400)

Every element(x) will be replaced by **x / Deviation.**

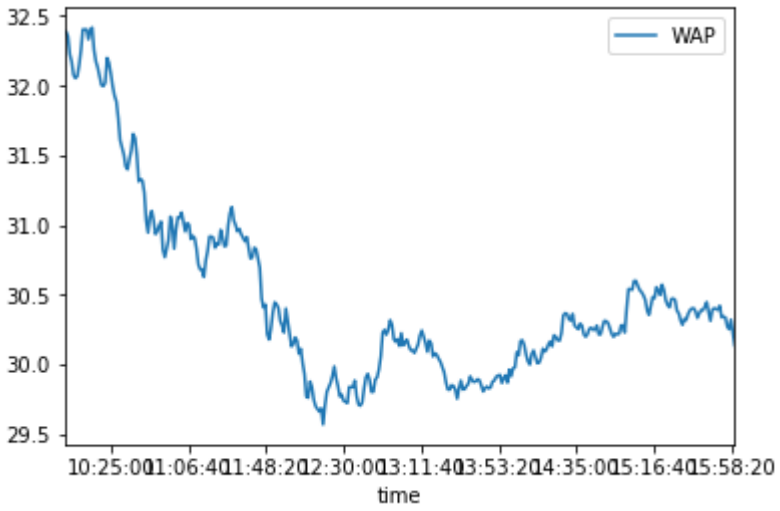
Where Deviation is the standard deviation of the data.

DATA VISUALIZATION

Here you have to select a graph type and column name

LINE (CREDITS COST:- $300)

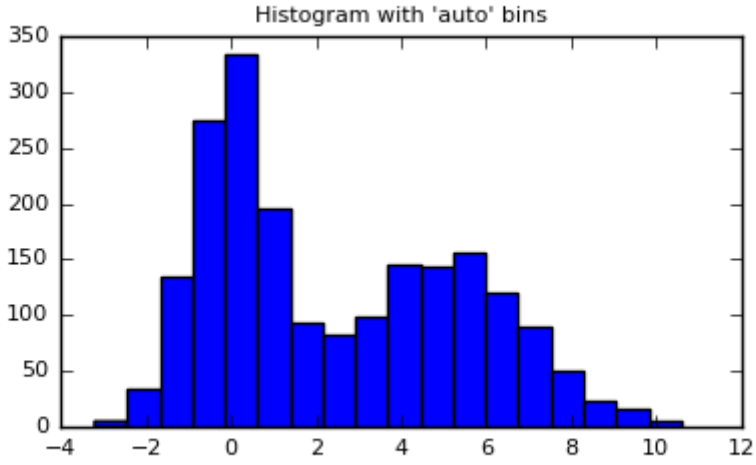
-You will get the graph of values of column vs entry number.



HISTOGRAM (CREDITS:- $250)

-It will give you the the graph between column entry and frequency.

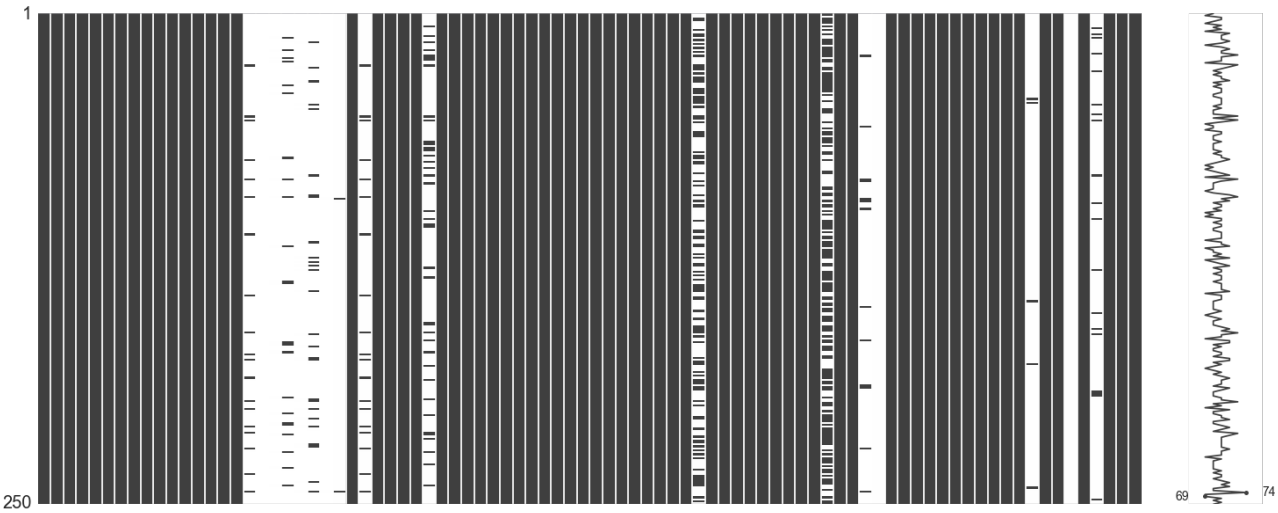
It would look like..



MISSING NUMBER VISUALISATION

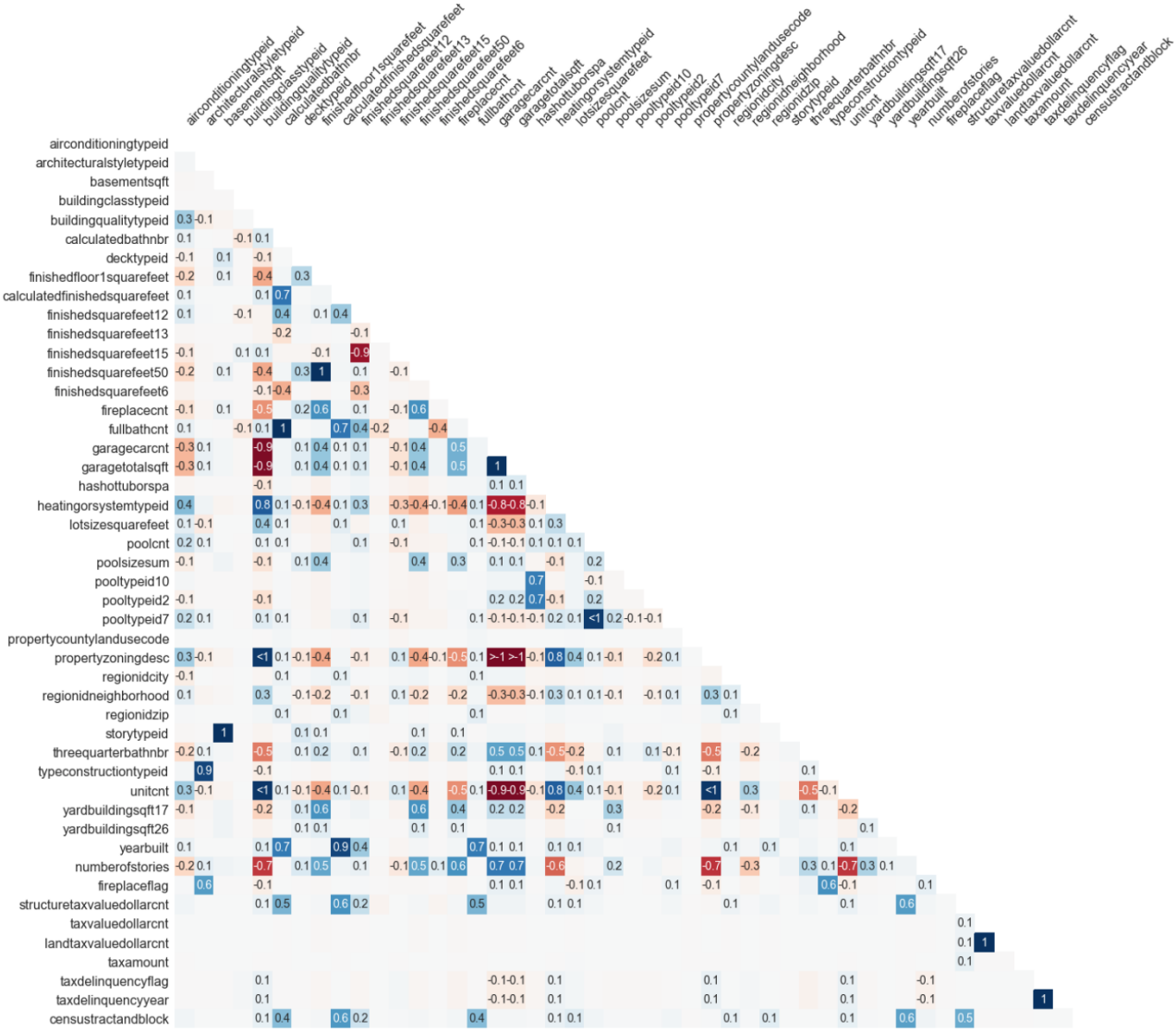
Matrix (CREDITS:- $600)

- The nullity matrix gives you a data-dense display which lets you quickly visually pick out the missing data patterns in the dataset.



HEATMAP (CREDITS:- $1000)

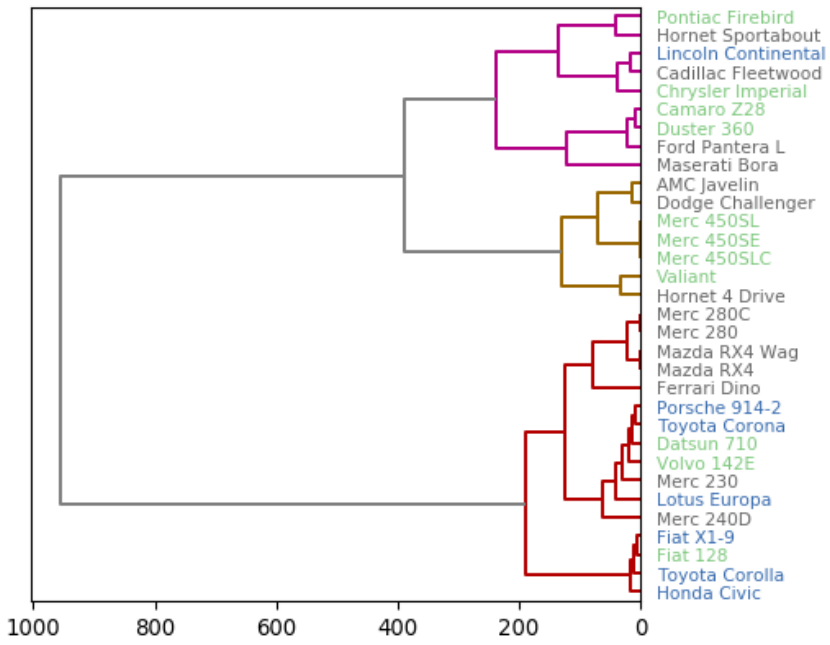
- This map describes the degree of nullity relationship between the different features. The range of this nullity correlation is from -1 to 1 ($-1 \leq R \leq 1$). Features with no missing value are excluded in the *heatmap*. If the nullity correlation is very close to zero ($-0.05 < R < 0.05$), no value will be displayed. Also, a perfect positive nullity correlation (R=1) indicates when the first feature and the second feature both have corresponding missing values while a perfect negative nullity correlation (R=-1) means that one of the features is missing and the second is not missing.



DENDROGRAM (CREDITS:- $700)

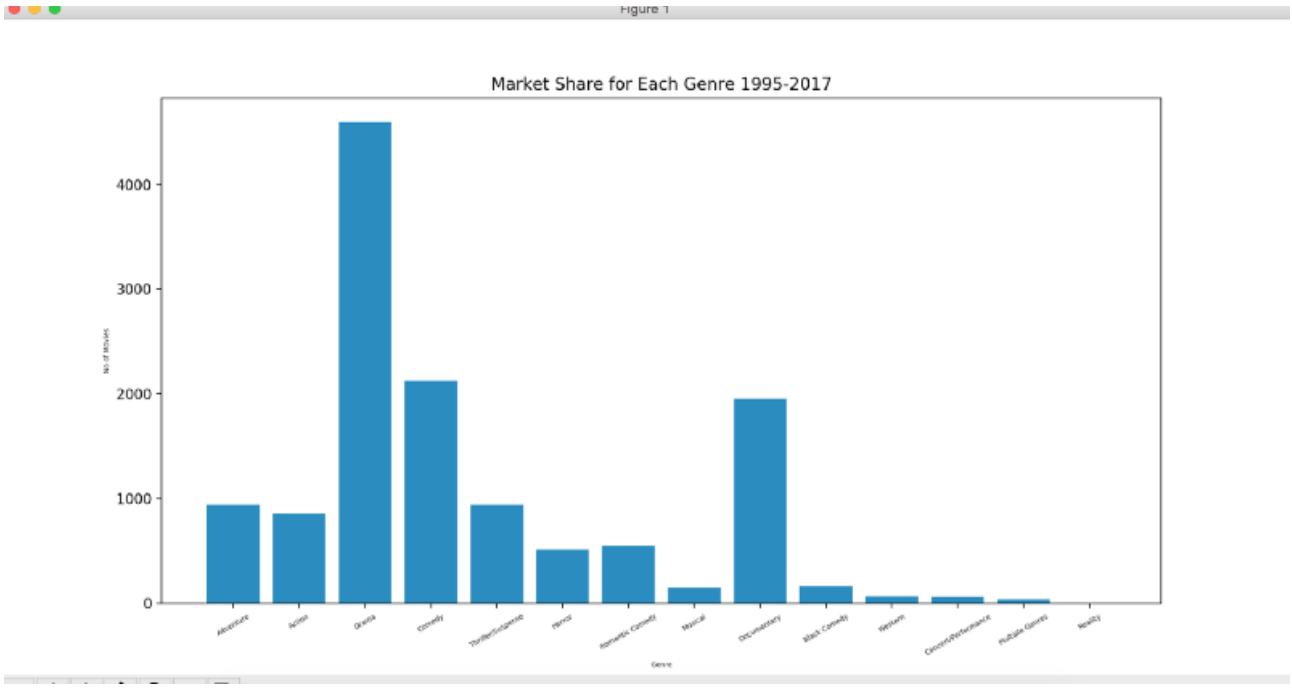
-Its a hierarchical clustering analysis used to analyse the order of dependencies of partially filled columns

The sample graph wold look like the below.



BAR (CREDITS:- $900)

-Its a bar graph of all the cols vs the fraction of non empty data in that col.



FILL NULL VALUES

Here you have to select a column name and Way of filling Null Values.

Mean (CREDITS:- $300)

Every Null space will be filled by **Mean**

Where Mean is the mean of the data of whole column.

Zero (CREDITS:- $300)

The null elements will be filled by **Zero**

Standard Deviation (CREDITS:- $300)

Every null element will be replaced by **Standard Deviation.** Where Deviation is the standard deviation of the data.

DROP COLUMN (CREDITS:- $700)

- It drops the required column from the dataset

LINEAR REGRESSION (CREDITS:- $5 per training example and after **3** linear regression $**3000 per call)**

- This is the function to call when you want to train the model.

- It displays the Training Accuracy for the Training Set

- Everytime you have to pick the no of data you want (one row is one data).

- REMINDER: The First 3 Clicks are "***Free of Cost***"

CHECKPOINT (CREDITS:- $2000)

- Everytime you do a crucial step, create a Checkpoint so that your work is saved and you can revert back

REVERT (CREDITS:- $500)

- It takes you back to the previous Checkpoint.

-Remember you can revert back to only two Checkpoints

TEST ACCURACY (CREDITS : $0)

- *******************CAUTION********************
TEST ACCURACY CAN BE SEEN ONLY THREE TIMES
****

- It gives the accuracy of your model on our test set (test data).

- This would be the most important criteria for your marking.

- Your model is trained on our Test Set and we see if your predictions matches the predicttions

GLOSSARY

- **MODEL :** The Linear Regression Algorithm works on the Training Set
- **TRAINING ACCURACY** : It calculates the Mean Sqaured Error for the Predictions on the Training Set and checks it with the original values present.
- **TEST ACCURACY :** It calculates the Mean Squared Error for the Predictions on the Test Set.

ENJOY YOUR DAY AND REMEMBER,

"WHAT YOU ARE DOING RIGHT NOW IS BEING DONE BY ONLY 8 % OF THE WORLD POPULATION,

SO, YES! WELCOME TO THE LEAGUE OF DATA SCIENTISTS (THE COOLEST WORK OF THE 21ST CENTURY)"