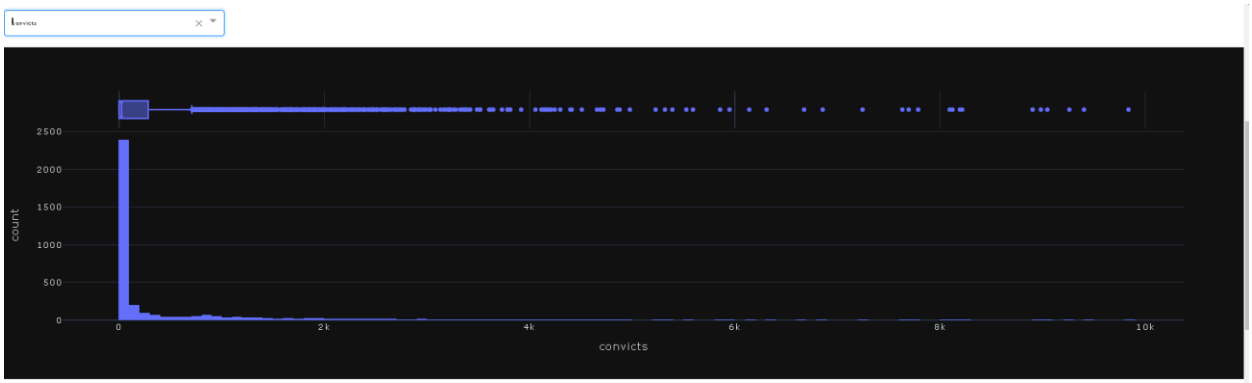
DASH APP

Distribution Tab:

This is the first tab of the EDA DASH APP. It consists of two charts

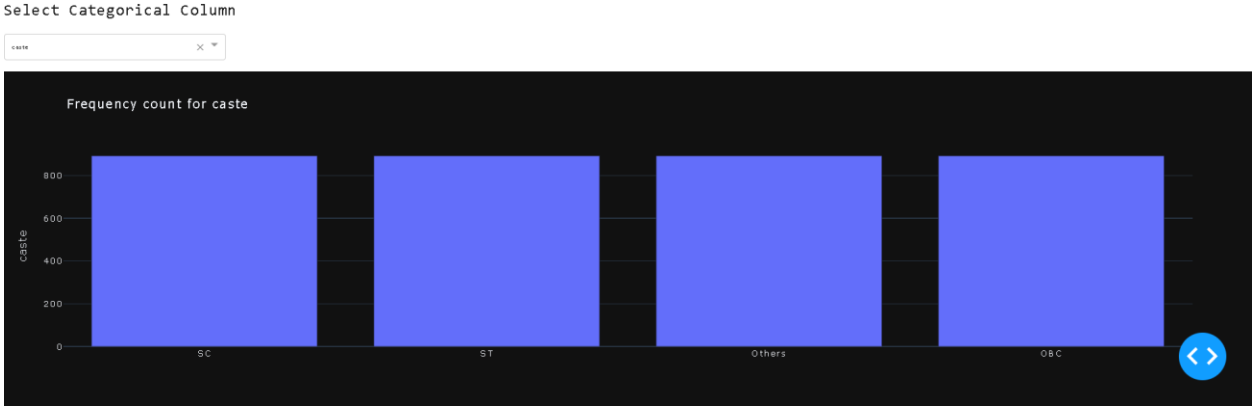
1. Distribution along with Box and Whisker Plot

The chart would take all the numerical columns which the CSV files has, and you could look at the distribution along with the outliers that selected feature has.



2. Frequency Plot for Categorical variables

This plot looks at the frequency of different categories which a categorical feature has. If the ML problem is a classification problem, then make sure that the dataset is balanced otherwise fix class imbalance issue by oversampling, under sampling etc.



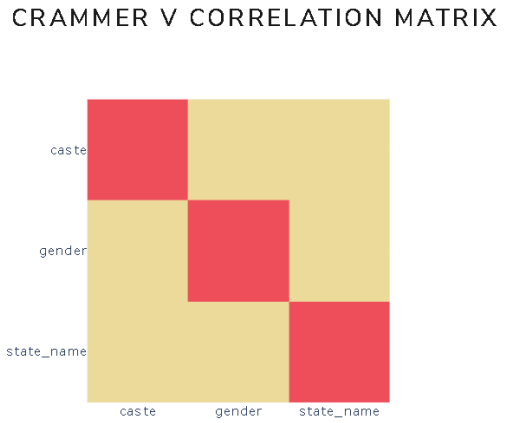
Correlation Tab:

This Tab might take a while to load. You won't see any loading bar for this. If the browser icon states updating, then wait for 5-10 mins. This tab lets you look how various features are correlated. By looking at these insights keep the relevant features and drop unnecessary features. In Machine Learning there is a saying "Garbage in is Garbage out ", so keep relevant ones.

Graphs in Correlation Tab:

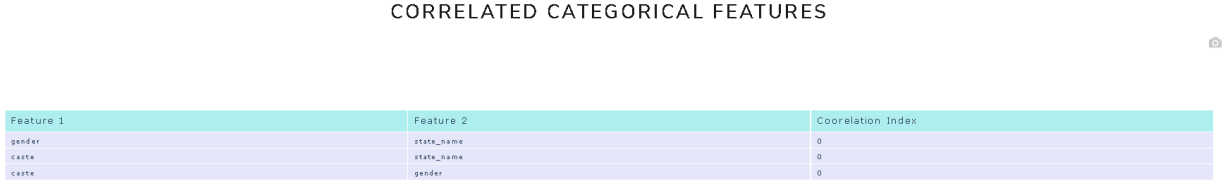
1. Crammer' V Correlation Matrix

It runs Crammer V test on Categorical variables and generates a heatmap. This value lies from 0-1. A higher value tells you that the association between 2 variables is high.



2. Correlated categorical Features

This table tells you the same thing which the above heatmap does. It would let you look at the correlation value sorted in descending order. High value (>0.5) means features have high correlation. For building good models, you need to have high correlation between target and features and low correlation among features.



3. Select the Target Categorical Variable to look for correlation

If you would like to filter the above table with respect to any categorical variables, you could do that from the dropdown.

SELECT THE TARGET CATEGORICAL VARIABLE TO LOOK FOR CORRELATION

caste

Feature 1	Feature 2	Coorelation Index
caste	state_name	0
caste	gender	0

4. Correlated Ratio Plot

This table let's you look at the association between a Categorical and Numerical Feature. The value ranges from 0-1. High value (>0.5) means features have high correlation. For building good models, you need to have high correlation between target and features and low correlation among features.

CORRELATED RATIO PLOT(ETA) TABLE

Categorical_Feature	Numerical_Feature	Coorelation Index
state_name	is_state	1
state_name	under_trial	0.57
state_name	convicts	0.54
state_name	detenues	0.48
state_name	others	0.44
gender	convicts	0.4
gender	under_trial	0.35
gender	detenues	0.18
caste	under_trial	0.13
caste	convicts	0.13
state_name	year	0.11
caste	detenues	0.07

5. Check Correlation between Numerical and Categorical Variables

You could filter the above table by any Categorical/Numerical feature from the dropdown.

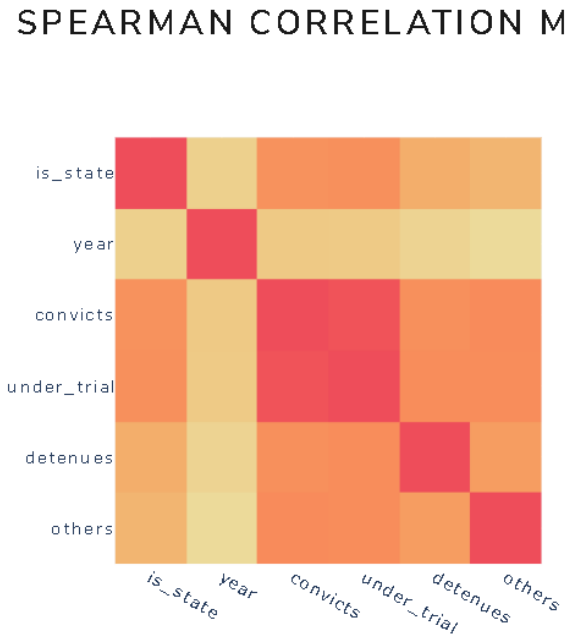
CHECK CORRELATION BETWEEN NUMERICAL AND CATEGORICAL VARIABLESS

caste

Categorical_Feature	Numerical_Feature	Coorelation Index
caste	under_trial	0.13
caste	convicts	0.13
caste	detenues	0.07
caste	others	0.05
caste	year	0
caste	is_state	0

6. Spearman Correlation Matrix

It runs Spearman test on Numerical variables and generates a heatmap. This value lies from -1 - 1. A higher absolute value tells you that the monotonic relationship between 2 numerical variables is high.



7. Correlation Features

This table tells you the same thing which the above heatmap does. It would let you look at the absolute correlation value sorted in descending order. High value (>0.5) means features have high correlation. For building good models, you need to have high correlation between target and features and low correlation among features.

CORRELATED FEATURES

Feature 1	Feature 2	Coorelation Index
convicts	under_trial	0.95
convicts	others	0.53
under_trial	detenues	0.52
under_trial	others	0.51
is_state	under_trial	0.49
convicts	detenues	0.49
is_state	convicts	0.48
detenues	others	0.4
is_state	detenues	0.28
is_state	others	0.22
year	others	0.07
year	convicts	0.06

8. Select the Target Numerical Variable to look for Correlation

If you would like to filter the above table with respect to any Numerical variables, you could do that from the dropdown.

SELECT THE TARGET NUMERICAL VARIABLE TO LOOK FOR CORRELATION

is_state

Feature 1	Feature 2	Coorelation Index
is_state	under_trial	0.49
is_state	convicts	0.48
is_state	detenues	0.28
is_state	others	0.22
is_state	year	0.01

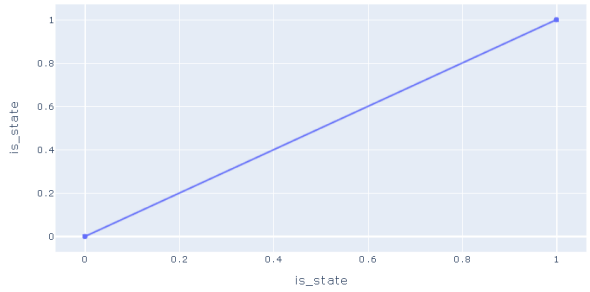
9. Regression and Scatterplot between features

Scatterplot is probably the best way to look at the relationship between 2 Numerical Variables. You could select two Numerical Columns from the Dropdown and look at the Scatterplot and fit an Ordinary Least Square regression model to look the line of best fit. The R square on the table right to it tells you about the linear relationship amongst the variables. It could be a case that The Spearman test tells you that the correlation is low, but by looking at scatterplot you find any Non-Linear relation which a model like Decision Tree might capture. This might help you in model selection.

REGRESSION AND SCATTERPLOT BETWEEN FEATURES

is_state

is_state



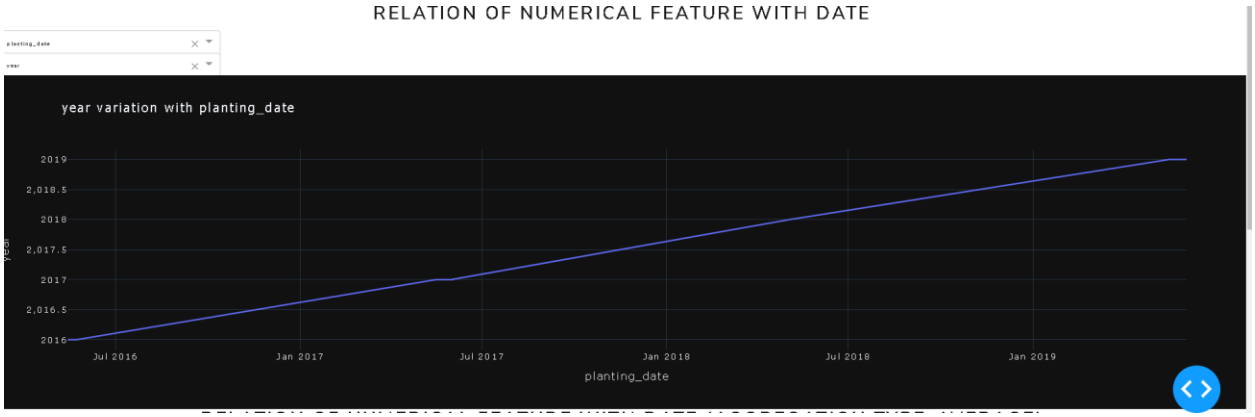
Dep. Variable:	y	R-squared:	1
Model:	OLS	Adj. R-squared:	1
Method:	Least Squares	F-statistic:	340000000000001e
Date:	Fri, 08 Jan 2021	Prob (F-statistic):	0
Time:	19:03:41	Log-Likelihood:	115770
No. Observations:	3568	AIC:	-231500
Df Residuals:	3566	BIC:	-231500
Df Model:	1	null	null
Covariance Type:	nonrobust	null	null

Timeline Tab:

This Tab let's you look how a Numerical column varies with the time. If you find that the dashboard is not able to capture this date column, then you need to format your date column. The algorithm behinds want date column to be of form month-day-year or day-month-year.

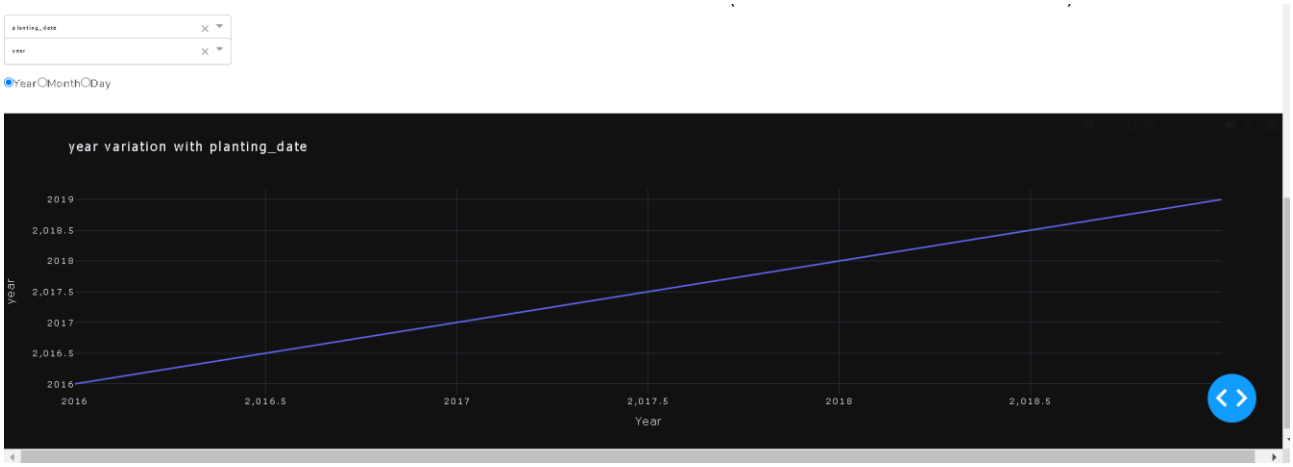
1. Relationship of Numerical feature with Date

This graph lets you look at the trend, seasonality. The two dropdowns would let you select the date column and Numerical Column.



2. Relation of Numerical Feature with Date.

If you would like to see how values are aggregated yearly, monthly or daily basis you could see using this graph. The two dropdowns would let you select the date column and Numerical Column. Use the radio button to look at different graphs.



The App is designed that only one user can use this application. If the app crashes just try loading the app again. If you want to try this again with new dataset, go to the home page from Explore Tab or reload the application.