Lab 03: Feature Selection

1. Missing Values Ratio

what to do when there are too many missing values, say, over 50%? In such situations, you can set a threshold value and use the missing values ratio method. If the percentage of missing values in a variable exceeds the threshold, you can drop the variable.

ID	season	holiday	workingday	weather	temp	atemp	humidity	windspeed	count
AB101	1.0	0.0	0.0	1.0	9.84	14.395	81.0	NaN	16
AB102	1.0	NaN	0.0	NaN	9.02	13.635	80.0	NaN	40
AB103	1.0	0.0	NaN	1.0	9.02	13.635	80.0	NaN	32
AB104	NaN	0.0	NaN	1.0	9.84	14.395	75.0	NaN	13
AB105	1.0	NaN	0.0	NaN	9.84	14.395	NaN	16.9979	1
AB106	1.0	0.0	NaN	2.0	9.84	12.880	75.0	NaN	1
AB107	1.0	0.0	0.0	1.0	9.02	13.635	80.0	NaN	2
AB108	1.0	NaN	0.0	1.0	8.20	12.880	86.0	NaN	3
AB109	NaN	0.0	0.0	NaN	9.84	14.395	NaN	NaN	8
AB110	1.0	0.0	0.0	1.0	13.12	17.425	76.0	NaN	14

2. Low Variance Filter:

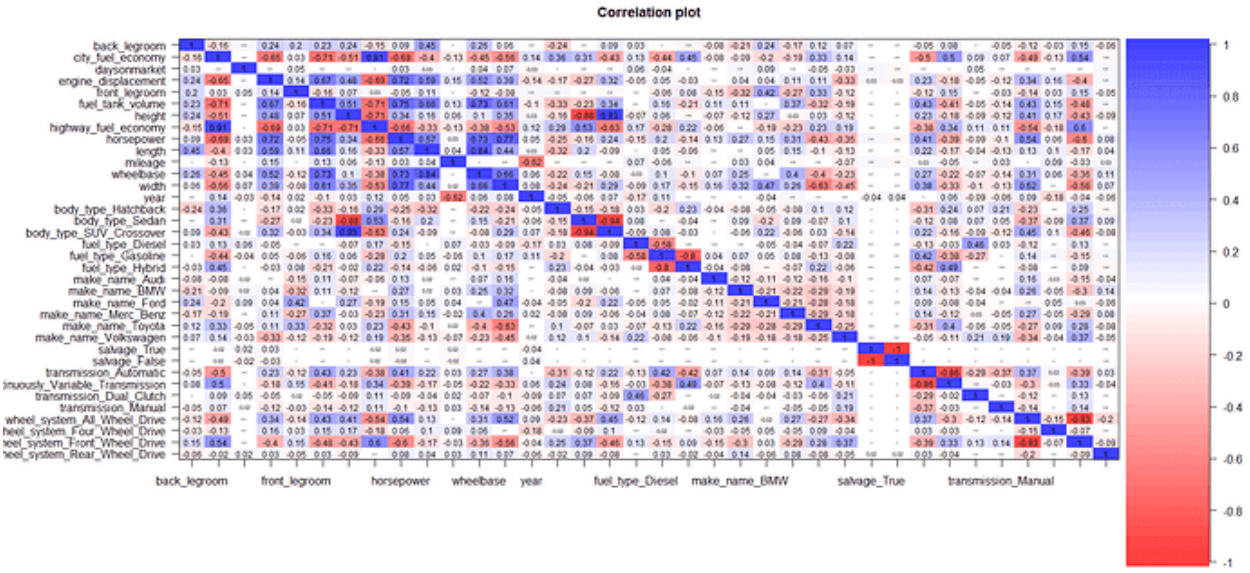
All the data columns with variance lower than the threshold value will be eliminated from the dataset

Note: Normalization is a must before implementing this dimensionality reduction technique

ID	season	holiday	workingday	weather	f5	temp	atemp	humidity	windspeed	count
AB101	1	0	0	1	7	9.84	14.395	81	0.0000	16
AB102	1	0	0	1	7	9.02	13.635	80	0.0000	40
AB103	1	0	0	1	7	9.02	13.635	80	0.0000	32
AB104	1	0	0	1	7	9.84	14.395	75	0.0000	13
AB105	1	0	0	1	7	9.84	14.395	75	0.0000	1
AB106	1	0	0	2	7	9.84	12.880	75	6.0032	1
AB107	1	0	0	1	7	9.02	13.635	80	0.0000	2
AB108	1	0	0	1	7	8.20	12.880	86	0.0000	3
AB109	1	0	0	1	7	9.84	14.395	75	0.0000	8
AB110	1	0	0	1	7	13.12	17.425	76	0.0000	14

3. High Correlation Filter

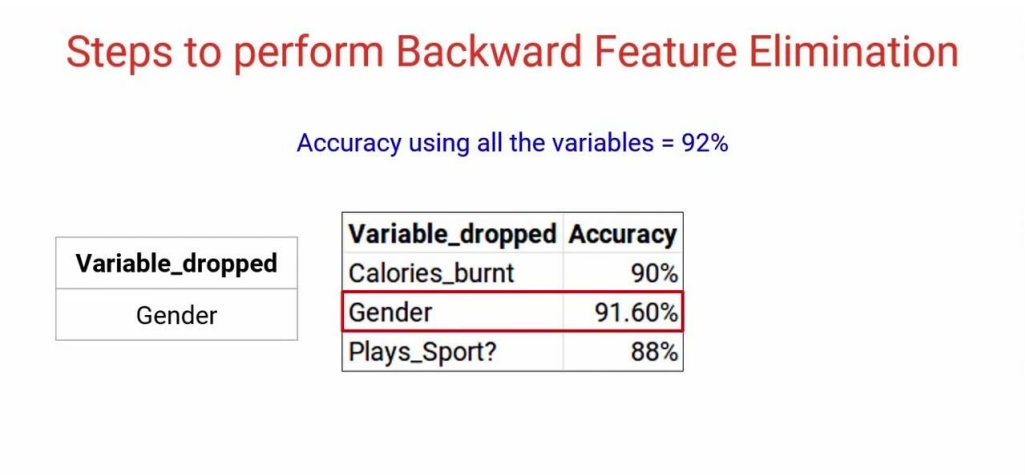
This dimensionality reduction algorithm tries to discard inputs that are very similar to others. In technical words, if there is a very high correlation between two input variables, we can safely drop one of them.



4. Backward Feature Elimination

In the backward feature elimination technique, you have to begin with all 'n' dimensions. Thus, at a given iteration,
Step 01: train a specific classification algorithm is trained on *n* input features
Step 02: Now, you have to remove one input feature at a time and train the same model on *n-1* input variables *n* times
Step 03: Similarly, you repeat the classification using *n-2* features, and this continues till no other variable can be removed

Each iteration (*k)* creates a model trained on n-k features having an error rate of *e(k)*. Then you must select the maximum bearable error rate to define the smallest number of features needed to reach that classification performance with the given ML algorithm.



5. Forward Feature Construction

The forward feature construction is the opposite of the backward feature elimination method. In the forward feature construction method, you begin with one feature and continue to progress by adding one feature at a time (this is the variable that results in the greatest boost in performance).

Lab Tasks

1. Load the dataset

```
Data = pd.read_csv('Data.csv')
```

2. Implement each of above reduction technique from stretch with proper comments, heading and visualization.
3. By applying each technique show all statistics and information about your analysis

- Use the threshold of your choice for the 1st three techniques
- Use appropriate plots for the last 2 feature selection techniques after selection
- Use co-relation matrix for the 3rd type of reduction technique