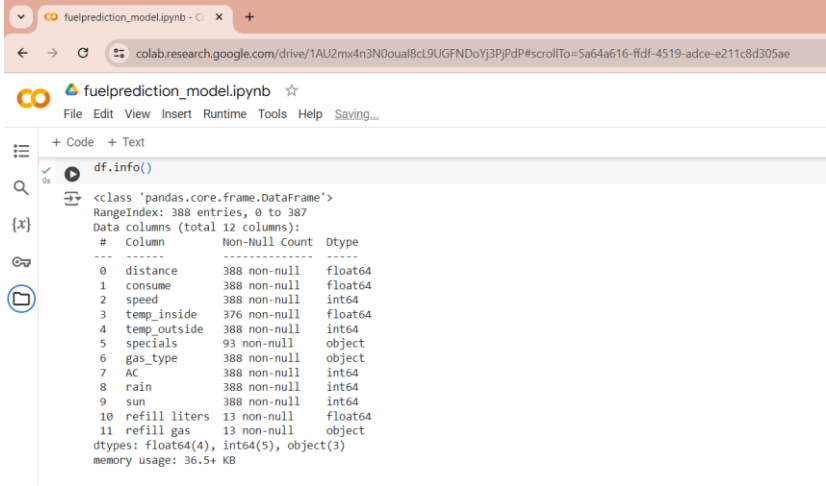



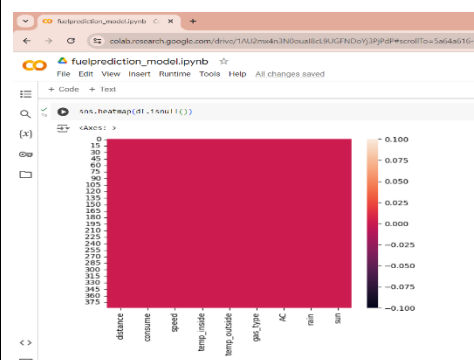
## Data Collection and Preprocessing Phase

Date	9 July 2024
Team ID	739652
Project Title	Trip-Based Modelling of Fuel Consumption in Modern Fleet Vehicles Using Machine Learning
Maximum Marks	6 Marks

### Data Exploration and Preprocessing Template

Identifies data sources, assesses quality issues like missing values and duplicates, and implements resolution plans to ensure accurate and reliable analysis.

Section	Description
Data Overview	<p>388 rows x 12 columns, dtypes: float64(4),int64(5),object(3)</p>  <pre> df.info() &lt;class 'pandas.core.frame.DataFrame'&gt; RangeIndex: 388 entries, 0 to 387 Data columns (total 12 columns): #   Column              Non-Null Count  Dtype   ---  - 0   distance            388 non-null    float64 1   consume             388 non-null    float64 2   speed               388 non-null    int64    3   temp_inside         376 non-null    float64 4   temp_outside        388 non-null    int64    5   specials            93 non-null     object   6   gas_type            388 non-null    object   7   AC                  388 non-null    int64    8   rain                388 non-null    int64    9   sun                 388 non-null    int64    10  refill liters       13 non-null     float64 11  refill gas          13 non-null     object   dtypes: float64(4), int64(5), object(3) memory usage: 36.5+ KB </pre>
Univariate Analysis	Exploration of individual of accuracy_score,mean_squared_error,r2_score,mean_absolute_error
Bivariate Analysis	Relationships between two variables (correlation, scatter plots)

																																																																																																							
Multivariate Analysis																																																																																																							
Outliers and Anomalies and null values	<p>We do not have any outliers and anomalies in our dataset except that we have null values so we have handle the null values and drop them in the way as follows</p> 																																																																																																						
Data Preprocessing Code Screenshots																																																																																																							
Loading Data	<pre>[7] df = pd.read_csv('measurements.csv') print(df.head())</pre> <table><tr><th></th><th>distance</th><th>consume</th><th>speed</th><th>temp_inside</th><th>temp_outside</th><th>specials</th><th>gas_type</th><th>AC</th><th>\</th></tr><tr><td>0</td><td>28.0</td><td>5.0</td><td>26</td><td>21.5</td><td>12</td><td>NaN</td><td>E10</td><td>0</td><td></td></tr><tr><td>1</td><td>12.0</td><td>4.2</td><td>30</td><td>21.5</td><td>13</td><td>NaN</td><td>E10</td><td>0</td><td></td></tr><tr><td>2</td><td>11.2</td><td>5.5</td><td>38</td><td>21.5</td><td>15</td><td>NaN</td><td>E10</td><td>0</td><td></td></tr><tr><td>3</td><td>12.9</td><td>3.9</td><td>36</td><td>21.5</td><td>14</td><td>NaN</td><td>E10</td><td>0</td><td></td></tr><tr><td>4</td><td>18.5</td><td>4.5</td><td>46</td><td>21.5</td><td>15</td><td>NaN</td><td>E10</td><td>0</td><td></td></tr></table> <table><tr><th></th><th>rain</th><th>sun</th><th>refill</th><th>liters</th><th>refill</th><th>gas</th></tr><tr><td>0</td><td>0</td><td>0</td><td></td><td>45.0</td><td></td><td>E10</td></tr><tr><td>1</td><td>0</td><td>0</td><td></td><td>NaN</td><td></td><td>NaN</td></tr><tr><td>2</td><td>0</td><td>0</td><td></td><td>NaN</td><td></td><td>NaN</td></tr><tr><td>3</td><td>0</td><td>0</td><td></td><td>NaN</td><td></td><td>NaN</td></tr><tr><td>4</td><td>0</td><td>0</td><td></td><td>NaN</td><td></td><td>NaN</td></tr></table>		distance	consume	speed	temp_inside	temp_outside	specials	gas_type	AC	\	0	28.0	5.0	26	21.5	12	NaN	E10	0		1	12.0	4.2	30	21.5	13	NaN	E10	0		2	11.2	5.5	38	21.5	15	NaN	E10	0		3	12.9	3.9	36	21.5	14	NaN	E10	0		4	18.5	4.5	46	21.5	15	NaN	E10	0			rain	sun	refill	liters	refill	gas	0	0	0		45.0		E10	1	0	0		NaN		NaN	2	0	0		NaN		NaN	3	0	0		NaN		NaN	4	0	0		NaN		NaN
	distance	consume	speed	temp_inside	temp_outside	specials	gas_type	AC	\																																																																																														
0	28.0	5.0	26	21.5	12	NaN	E10	0																																																																																															
1	12.0	4.2	30	21.5	13	NaN	E10	0																																																																																															
2	11.2	5.5	38	21.5	15	NaN	E10	0																																																																																															
3	12.9	3.9	36	21.5	14	NaN	E10	0																																																																																															
4	18.5	4.5	46	21.5	15	NaN	E10	0																																																																																															
	rain	sun	refill	liters	refill	gas																																																																																																	
0	0	0		45.0		E10																																																																																																	
1	0	0		NaN		NaN																																																																																																	
2	0	0		NaN		NaN																																																																																																	
3	0	0		NaN		NaN																																																																																																	
4	0	0		NaN		NaN																																																																																																	

## Handling Missing Data

```
df.isnull()
```

	distance	consume	speed	temp_inside	temp_outside	specials	gas_type	AC	rain	sun	refill	liters	refill	gas
0	False	False	False	False	False	True	False	False	False	False	False	False	False	False
1	False	False	False	False	False	True	False	False	False	False	True	True	True	True
2	False	False	False	False	False	True	False	False	False	False	True	True	True	True
3	False	False	False	False	False	True	False	False	False	False	True	True	True	True
4	False	False	False	False	False	True	False	False	False	False	True	True	True	True
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
383	False	False	False	False	False	True	False	False	False	False	True	True	True	True
384	False	False	False	False	False	False	False	False	False	False	True	True	True	True
385	False	False	False	False	False	True	False	False	False	False	True	True	True	True
386	False	False	False	False	False	False	False	False	False	False	True	True	True	True
387	False	False	False	False	False	False	False	False	False	False	True	True	True	True

388 rows × 12 columns

## Data Transformation

### Splitting the data :

```
[24] x=x.values
     y=y.values

#Splitting Data Into Train And Test

[26] x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.3,random_state=42)
```

## Feature Engineering

### modifying existing ones

```
[20] #seperating independent and dependent variables

[21] from sklearn.model_selection import train_test_split
     from sklearn.linear_model import LinearRegression

[22] x=df.drop(['consume','gas_type'],axis=1)
     y=df['consume']

[23] x.columns
     Index(['distance', 'speed', 'temp_inside', 'temp_outside', 'AC', 'rain',
           'sun'],
           dtype='object')

[24] x=x.values
     y=y.values
```

## Save Processed Data

### Code to save the cleaned and processed data for future use.

```
import pickle
pickle.dump(dt,open('fuel2.pkl','wb'))
```