

# Project Report: Correlation Analysis between Vehicle sales and Co2 Emission in USA from 1985 until 2015

## Introduction

This project helps to investigate between Number of vehicles sold in USA from 1982 to 2015 and the increase in carbon level emission due to it in the environment. The emission of Co2 impact various factors like greenhouse effect, trapping more heat in the Earth's atmosphere, leading to global temperature rise which can lead to severe weather events such as hurricanes, droughts, heatwaves, and heavy rainfall. This report will help you to analyse how the sales of vehicles leads to the increase in carbon emissions.

## Methods

### Data Source

1. Vehicle Sales data of USA: The data is taken from Kaggle, this dataset is a open source MIT licensed. The link for dataset is [Vehicle sales data](#).
2. Co2 Emission data of USA: The data is taken from kaggle, this dataset is a open source dataset. The link for dataset is [Co\\_2\\_emission](#)

### Data Pipeline

The data pipeline has the following components:

### Extraction

The data extraction process involves of extracting data from kaggle. The python code is used to download the data from the kaggle. Here the extracted data is preprocessed and stored into a sqllite database in form of tables.

### Transformation

The data transformation process includes of following steps:

1. Filtering out necessary columns: The Vehicle sales data consisted of various other columns. Hence the unnecessary columns where filtered out from the dataset.
2. Checking the datatypes: Checking datatypes of columns before storing them it finally into database is the most essential as it may cause errors during the interpretations with data.
3. Renaming columns: The columns are renamed such that they are more interpretable.

4. Handling Missing values: The missing values can effect the quality of data and make it unusable. Hence, to ensure the usability of data the missing and null values were dropped.

All of the above transformation steps ensure that the data is clean and ready for analysis.

## Results and Limitations

Limitations:

1. Since the carbon emission data is for the entire country we can't do the analysis for different states in the country.
2. Since the Co<sub>2</sub> is not only emitted with the use of vehicles but also from the other things like factories and other different development ongoing projects.

Vehicle Data Analysis

```
In [15]: import sqlite3
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np
import seaborn as sns
```

```
In [16]: conn = sqlite3.connect("../data\\data.sqlite")
```

```
In [17]: vehicle_data = pd.read_sql_query("SELECT * FROM Vehicle_sales_data", conn)
```

```
In [18]: vehicle_data
```

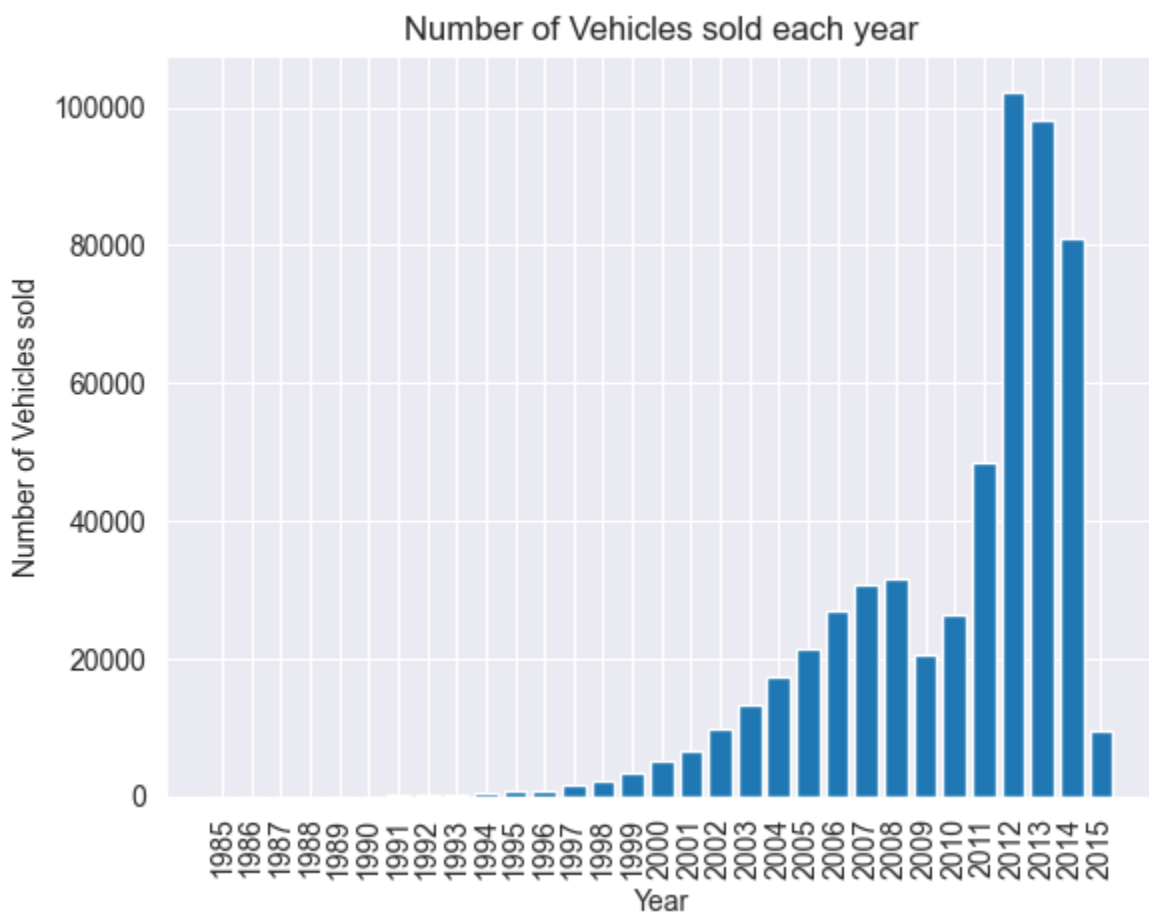
```
Out[18]:
```

	Year	Number of Vehicles sold
0	1985	10
1	1986	11
2	1987	8
3	1988	11
4	1989	20
5	1990	49
6	1991	67
7	1992	132
8	1993	205
9	1994	392
10	1995	711
11	1996	851
12	1997	1546
13	1998	2149
14	1999	3363
15	2000	5227
16	2001	6468
17	2002	9715

18	2003	13281
19	2004	17342
20	2005	21394
21	2006	26913
22	2007	30845
23	2008	31502
24	2009	20594
25	2010	26485
26	2011	48548
27	2012	102315
28	2013	98168
29	2014	81070
30	2015	9437

```
In [27]: plt.figure()
plt.bar( vehicle_data['Year'],vehicle_data['Number of Vehicles sold'])
plt.xticks(vehicle_data['Year'],rotation=90)
plt.xlabel('Year')
plt.ylabel('Number of Vehicles sold')
plt.title('Number of Vehicles sold each year')
```

```
Out[27]: Text(0.5, 1.0, 'Number of Vehicles sold each year')
```



Observation

From this graph we can conclude that in year 2012 the number of vehicle sold were more. We can also observe that the trend increased from year 1995 and went up till 2008 and then had an down trend till 2010 and rise up with a increasing trend untill 2012

## Co\_2 Emission analysis

```
In [20]: co2_data = pd.read_sql_query("SELECT * FROM co2_emission", conn)
```

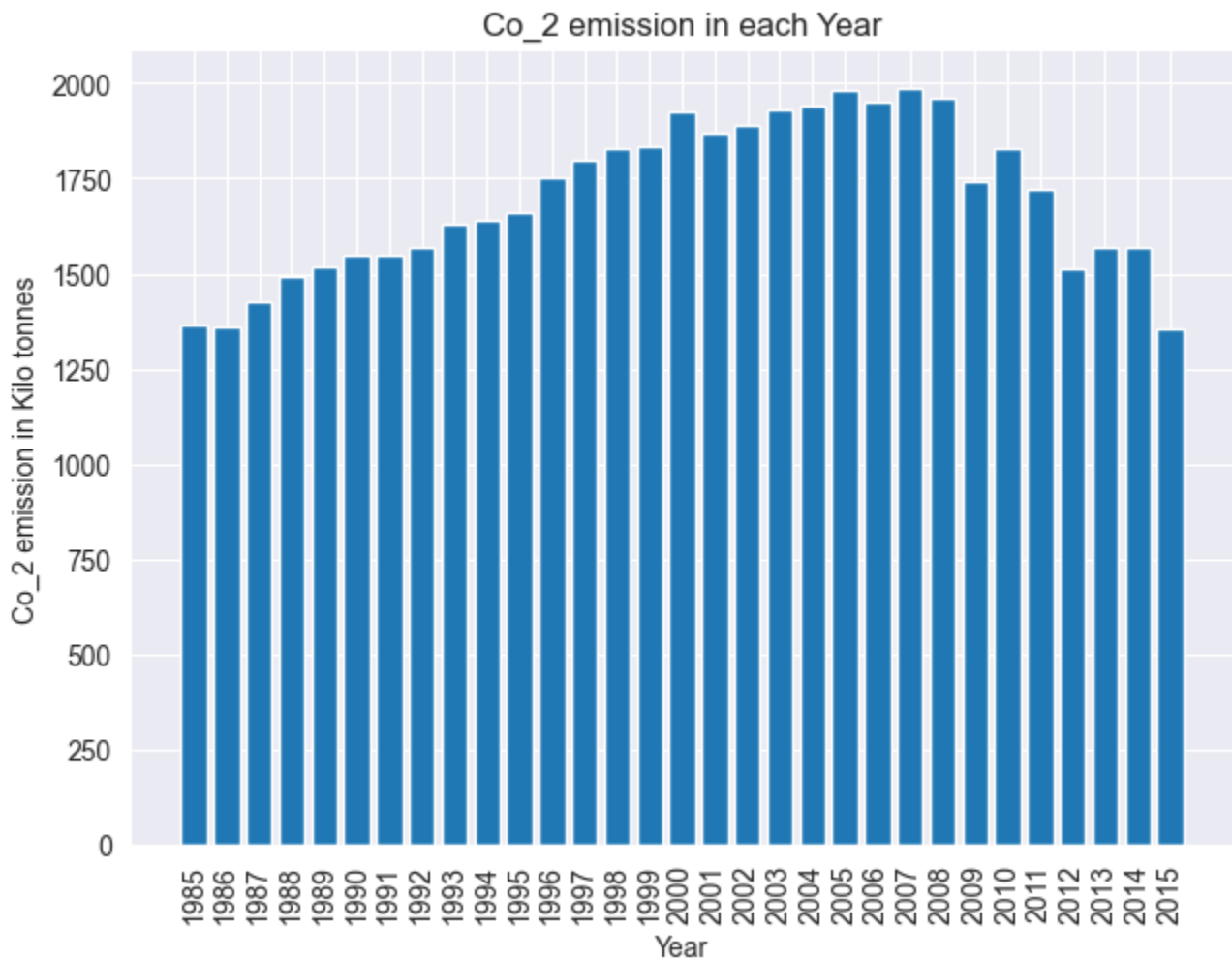
```
In [21]: co2_data
```

```
Out[21]:
```

	Year	Co2_emission
0	1985	1367.400
1	1986	1357.485
2	1987	1426.752
3	1988	1491.808
4	1989	1518.843
5	1990	1547.564
6	1991	1548.180
7	1992	1569.565
8	1993	1632.519
9	1994	1638.781
10	1995	1660.743
11	1996	1752.410
12	1997	1797.044
13	1998	1828.183
14	1999	1836.415
15	2000	1927.377
16	2001	1869.828
17	2002	1889.879
18	2003	1930.972
19	2004	1943.069
20	2005	1983.828
21	2006	1953.697
22	2007	1987.287
23	2008	1959.386
24	2009	1740.875
25	2010	1827.619
26	2011	1722.709
27	2012	1511.238
28	2013	1571.350
29	2014	1569.136
30	2015	1352.967

```
In [26]: plt.figure()
plt.bar(co2_data['Year'], co2_data['Co2_emission'])
plt.xticks(co2_data['Year'], rotation=90)
plt.tight_layout()
plt.xlabel('Year')
plt.ylabel('Co_2 emission in Kilo tonnes')
plt.title('Co_2 emission in each Year')
```

```
Out[26]: Text(0.5, 1.0, 'Co_2 emission in each Year')
```

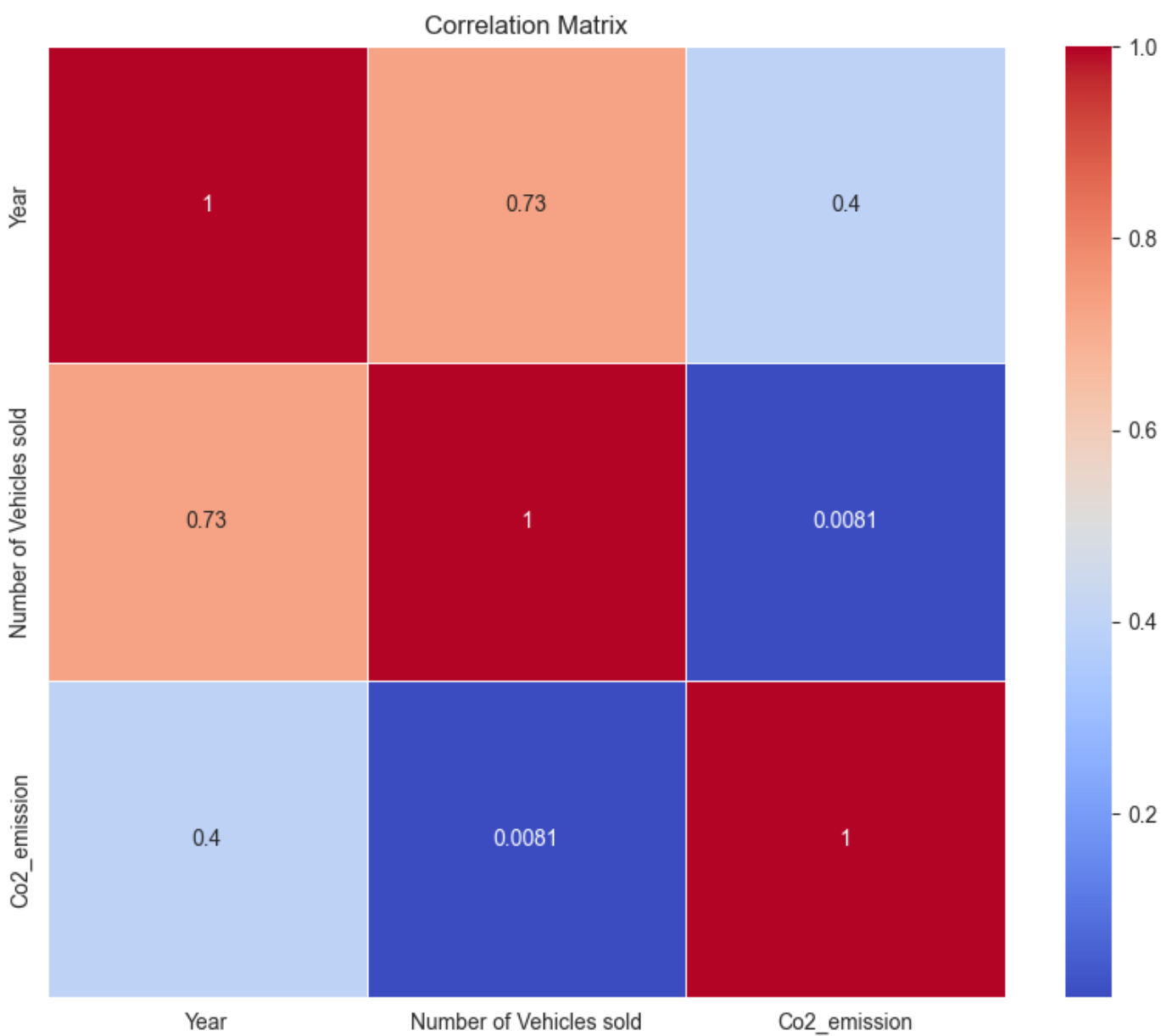


### Correlation Analysis

```
In [23]: merged_data = pd.merge(vehicle_data, co2_data, on='Year')
```

```
In [24]: correlation_matrix = merged_data.corr(numeric_only=True)
```

```
In [25]: plt.figure(figsize=(10, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', linewidths=0.5)
plt.title('Correlation Matrix')
plt.show()
```



## Observation

Based on the above observations we can state that the CO<sub>2</sub> emission is not only dependent on sales of vehicles but there are also other unknown factors which also lead to the increase or decrease in the carbon emission.

In [ ]: