In [2]:
```
import pandas as pd
import os
print(os.getcwd())
```

```
C:\Users\spark\data science projects\Global Warming Project
```

In [3]:
```
os.chdir('C:\\Users\\spark\\Documents\\Data Science Project\\Global Warming\Data')
```

In [4]:
```
#The objective of this project is to use outer join, inner join, set addition, and then also use spark linear regres
#to come up with statistical analysis of global warming, while also using mathematical formulas
#such as taylor series.
#In terms of prediction, we will utilize linear regression spark to make a prediction of the temperature.
#And then we will also make a prediction using taylor series.
#we will also utilize staitics to prove the error
```

In [5]:
```
CO2Concentration = pd.read_csv('CO2Concentration.csv')
Methaneconcentration = pd.read_csv('Methane concentration.csv')
NO2Concentration= pd.read_csv('NO2Concentration.csv')
TemperatureIncrease = pd.read_csv('Temperature Increase.csv', parse_dates = ["Day"])
```

In [6]:
```
TemperatureIncrease.dtypes
```

Out[6]:
```
Entity                          object
Code                            object
Day                     datetime64[ns]
temperature_anomaly            float64
dtype: object
```

In [7]:
```
TemperatureIncrease
```

Out[7]:

	Entity	Code	Day	temperature_anomaly
0	Northern Hemisphere	NaN	1880-01-15	-0.35
1	Northern Hemisphere	NaN	1880-02-15	-0.51
2	Northern Hemisphere	NaN	1880-03-15	-0.23
3	Northern Hemisphere	NaN	1880-04-15	-0.30
4	Northern Hemisphere	NaN	1880-05-15	-0.06
...	...	...	...	...
5107	World	OWID_WRL	2021-08-15	0.82
5108	World	OWID_WRL	2021-09-15	0.92
5109	World	OWID_WRL	2021-10-15	1.00
5110	World	OWID_WRL	2021-11-15	0.93
5111	World	OWID_WRL	2021-12-15	0.86

5112 rows × 4 columns

In [8]:
```
TemperatureIncreasetest = TemperatureIncrease
```

In [9]:
```
c = TemperatureIncreasetest.Day
```

In [10]:
```
TemperatureIncreasetest.Day = c.dt.year
#Convert into years first.
```

In [11]:
```
print(TemperatureIncreasetest)
#Select only rows with value "Northern Hemisphere"
```

```
                 Entity      Code   Day  temperature_anomaly
0     Northern Hemisphere       NaN  1880                -0.35
1     Northern Hemisphere       NaN  1880                -0.51
2     Northern Hemisphere       NaN  1880                -0.23
3     Northern Hemisphere       NaN  1880                -0.30
4     Northern Hemisphere       NaN  1880                -0.06
...                   ...       ...   ...                  ...
5107                World  OWID_WRL  2021                 0.82
5108                World  OWID_WRL  2021                 0.92
5109                World  OWID_WRL  2021                 1.00
5110                World  OWID_WRL  2021                 0.93
5111                World  OWID_WRL  2021                 0.86

[5112 rows x 4 columns]
```

In [12]:
```
#TemperatureIncreasetest['avg_points_rebounds'] = TemperatureIncreasetest[['temperature_anomaly']].mean(axis=1)
```

In [13]:
```
TemperatureIncreasetest[0:13]
```

Out[13]:

	Entity	Code	Day	temperature_anomaly
0	Northern Hemisphere	NaN	1880	-0.35
1	Northern Hemisphere	NaN	1880	-0.51
2	Northern Hemisphere	NaN	1880	-0.23
3	Northern Hemisphere	NaN	1880	-0.30
4	Northern Hemisphere	NaN	1880	-0.06
5	Northern Hemisphere	NaN	1880	-0.16
6	Northern Hemisphere	NaN	1880	-0.18
7	Northern Hemisphere	NaN	1880	-0.26
8	Northern Hemisphere	NaN	1880	-0.23
9	Northern Hemisphere	NaN	1880	-0.32
10	Northern Hemisphere	NaN	1880	-0.43
11	Northern Hemisphere	NaN	1880	-0.40
12	Northern Hemisphere	NaN	1881	-0.30

In [14]:
```
TemperatureIncreasetest = pd.DataFrame(data=TemperatureIncreasetest)
cars_groups = TemperatureIncreasetest.groupby(TemperatureIncreasetest['Day'])
```

In [15]:
```
graph =cars_groups.mean()
```

In [16]:
```
graph
```

Out[16]:

	temperature_anomaly
Day	
1880	-0.161944
1881	-0.081667
1882	-0.108611
1883	-0.171944
1884	-0.285278
...	...
2017	0.923611
2018	0.849444
2019	0.982222
2020	1.022778
2021	0.850278

142 rows × 1 columns

In [17]:
```
import seaborn as sns
```

In [18]:
```
graph['temperature_anomaly']
```

Out[18]:
```
Day
1880   -0.161944
1881   -0.081667
1882   -0.108611
1883   -0.171944
1884   -0.285278
          ...
2017    0.923611
2018    0.849444
2019    0.982222
2020    1.022778
2021    0.850278
Name: temperature_anomaly, Length: 142, dtype: float64
```

In [19]:
```
graph
```

Out[19]:

	temperature_anomaly
Day	
1880	-0.161944
1881	-0.081667
1882	-0.108611
1883	-0.171944
1884	-0.285278
...	...
2017	0.923611
2018	0.849444
2019	0.982222
2020	1.022778
2021	0.850278

142 rows × 1 columns

In [20]:
```
CO2Concentration
```

Out[20]:

	Entity	Code	Year	CO2 concentrations (NOAA, 2018)
0	World	OWID_WRL	1	276.70
1	World	OWID_WRL	30	277.90
2	World	OWID_WRL	56	277.40
3	World	OWID_WRL	104	277.50
4	World	OWID_WRL	136	278.10
...	...	...	...	...
218	World	OWID_WRL	2014	398.65
219	World	OWID_WRL	2015	400.83
220	World	OWID_WRL	2016	404.24
221	World	OWID_WRL	2017	406.55
222	World	OWID_WRL	2018	408.52

223 rows × 4 columns

In [21]:
```
Methaneconcentration
```

Out[21]:

	Entity	Code	Year	CH4 concentration (EEA & NOAA (2019))
0	World	OWID_WRL	1750	719.01
1	World	OWID_WRL	1755	719.97
2	World	OWID_WRL	1760	720.93
3	World	OWID_WRL	1765	723.71
4	World	OWID_WRL	1770	726.50
...	...	...	...	...
82	World	OWID_WRL	2014	1824.40
83	World	OWID_WRL	2015	1834.63
84	World	OWID_WRL	2016	1842.40
85	World	OWID_WRL	2017	1849.63
86	World	OWID_WRL	2018	1857.62

87 rows × 4 columns

In [22]:
```
NO2Concentration
```

Out[22]:

	Entity	Code	Year	N2O concentrations (annual average) (EEA, 2019)
0	World	OWID_WRL	1750	270.00
1	World	OWID_WRL	1755	270.30
2	World	OWID_WRL	1760	270.60
3	World	OWID_WRL	1765	270.90
4	World	OWID_WRL	1770	271.20
...	...	...	...	...
80	World	OWID_WRL	2012	325.58
81	World	OWID_WRL	2013	326.53
82	World	OWID_WRL	2014	327.61
83	World	OWID_WRL	2015	328.51
84	World	OWID_WRL	2016	329.29

85 rows × 4 columns

In [23]:
```
data = pd.merge(NO2Concentration, Methaneconcentration, how='left', on ='Year')
```

In [24]:
```
data
```

Out[24]:

	Entity_x	Code_x	Year	N2O concentrations (annual average) (EEA, 2019)	Entity_y	Code_y	CH4 concentration (EEA & NOAA (2019))
0	World	OWID_WRL	1750	270.00	World	OWID_WRL	719.01
1	World	OWID_WRL	1755	270.30	World	OWID_WRL	719.97
2	World	OWID_WRL	1760	270.60	World	OWID_WRL	720.93
3	World	OWID_WRL	1765	270.90	World	OWID_WRL	723.71
4	World	OWID_WRL	1770	271.20	World	OWID_WRL	726.50
...	...	...	...	...	...	...	...
80	World	OWID_WRL	2012	325.58	World	OWID_WRL	1810.33
81	World	OWID_WRL	2013	326.53	World	OWID_WRL	1815.44
82	World	OWID_WRL	2014	327.61	World	OWID_WRL	1824.40
83	World	OWID_WRL	2015	328.51	World	OWID_WRL	1834.63
84	World	OWID_WRL	2016	329.29	World	OWID_WRL	1842.40

85 rows × 7 columns

In [25]:
```
data =data.drop(['Entity_y', 'Code_y','Entity_y'], axis=1)
```

In [26]:
```
data = pd.merge(data, Methaneconcentration, how='left', on ='Year')
```

In [27]:
```
data
```

Out[27]:

	Entity_x	Code_x	Year	N2O concentrations (annual average) (EEA, 2019)	CH4 concentration (EEA & NOAA (2019))_x	Entity	Code	CH4 concentration (EEA & NOAA (2019))_y
0	World	OWID_WRL	1750	270.00	719.01	World	OWID_WRL	719.01
1	World	OWID_WRL	1755	270.30	719.97	World	OWID_WRL	719.97
2	World	OWID_WRL	1760	270.60	720.93	World	OWID_WRL	720.93
3	World	OWID_WRL	1765	270.90	723.71	World	OWID_WRL	723.71
4	World	OWID_WRL	1770	271.20	726.50	World	OWID_WRL	726.50
...	...	...	...	...	...	...	...	...
80	World	OWID_WRL	2012	325.58	1810.33	World	OWID_WRL	1810.33
81	World	OWID_WRL	2013	326.53	1815.44	World	OWID_WRL	1815.44
82	World	OWID_WRL	2014	327.61	1824.40	World	OWID_WRL	1824.40
83	World	OWID_WRL	2015	328.51	1834.63	World	OWID_WRL	1834.63
84	World	OWID_WRL	2016	329.29	1842.40	World	OWID_WRL	1842.40

85 rows × 8 columns

In [28]:
```
TemperatureIncreasetest = TemperatureIncrease
```

In [29]:
```
TemperatureIncreasetest
```

Out[29]:

	Entity	Code	Day	temperature_anomaly
0	Northern Hemisphere	NaN	1880	-0.35
1	Northern Hemisphere	NaN	1880	-0.51
2	Northern Hemisphere	NaN	1880	-0.23
3	Northern Hemisphere	NaN	1880	-0.30
4	Northern Hemisphere	NaN	1880	-0.06
...	...	...	...	...
5107	World	OWID_WRL	2021	0.82
5108	World	OWID_WRL	2021	0.92
5109	World	OWID_WRL	2021	1.00
5110	World	OWID_WRL	2021	0.93
5111	World	OWID_WRL	2021	0.86

5112 rows × 4 columns

In [30]:
```
TemperatureIncreasetest = pd.DataFrame(data=TemperatureIncreasetest)
cars_groups = TemperatureIncreasetest.groupby(TemperatureIncreasetest['Day'])
```

In [31]:
```
graph =cars_groups.mean()
```

In [32]:
```
graph=graph.reset_index()
```

In [33]:
```
graph
```

Out[33]:

	Day	temperature_anomaly
0	1880	-0.161944
1	1881	-0.081667
2	1882	-0.108611
3	1883	-0.171944
4	1884	-0.285278
...	...	...
137	2017	0.923611
138	2018	0.849444
139	2019	0.982222
140	2020	1.022778
141	2021	0.850278

142 rows × 2 columns

In [34]:
```
graph = graph.rename(columns={'Day': 'Year'})
```

In [35]:
```
graph=graph.drop(columns ='avg_points_rebounds')
```

```
---------------------------------------------------------------------------
KeyError                                  Traceback (most recent call last)
~\AppData\Local\Temp/ipykernel_32448/4288715824.py in <module>
----> 1 graph=graph.drop(columns ='avg_points_rebounds')

C:\ProgramData\Anaconda3\lib\site-packages\pandas\util\_decorators.py in wrapper(*args, **kwargs)
    309                     stacklevel=stacklevel,
    310                 )
--> 311             return func(*args, **kwargs)
    312 
    313         return wrapper

C:\ProgramData\Anaconda3\lib\site-packages\pandas\core\frame.py in drop(self, labels, axis, index, columns, level, inpl
ace, errors)
   4904                 weight  1.0     0.8
   4905         """
-> 4906         return super().drop(
   4907             labels=labels,
   4908             axis=axis,

C:\ProgramData\Anaconda3\lib\site-packages\pandas\core\generic.py in drop(self, labels, axis, index, columns, level, in
place, errors)
   4148         for axis, labels in axes.items():
   4149             if labels is not None:
-> 4150                 obj = obj._drop_axis(labels, axis, level=level, errors=errors)
   4151 
   4152         if inplace:

C:\ProgramData\Anaconda3\lib\site-packages\pandas\core\generic.py in _drop_axis(self, labels, axis, level, errors)
   4183                 new_axis = axis.drop(labels, level=level, errors=errors)
   4184             else:
-> 4185                 new_axis = axis.drop(labels, errors=errors)
   4186             result = self.reindex(**{axis_name: new_axis})
   4187 

C:\ProgramData\Anaconda3\lib\site-packages\pandas\core\indexes\base.py in drop(self, labels, errors)
   6015         if mask.any():
   6016             if errors != "ignore":
-> 6017                 raise KeyError(f"{labels[mask]} not found in axis")
```

```
   6018             indexer = indexer[~mask]
   6019         return self.delete(indexer)

KeyError: "['avg_points_rebounds'] not found in axis"
```

In [36]:
```
graph
```

Out[36]:

	Year	temperature_anomaly
0	1880	-0.161944
1	1881	-0.081667
2	1882	-0.108611
3	1883	-0.171944
4	1884	-0.285278
...	...	...
137	2017	0.923611
138	2018	0.849444
139	2019	0.982222
140	2020	1.022778
141	2021	0.850278

142 rows × 2 columns

In [37]:
```
graph2 = pd.merge(graph, data, how='left', on ='Year')
```

In [38]:
```
graph2 = graph2.dropna()
```

In [39]:

```
graph2 #This one (from thsi one, now apply the spark linear regression model.)
```

Out[39]:

	Year	temperature_anomaly	Entity_x	Code_x	N2O concentrations (annual average) (EEA, 2019)	CH4 concentration (EEA & NOAA (2019))_x	Entity	Code	CH4 concentration (EEA & NOAA (2019))_y
0	1880	-0.161944	World	OWID_WRL	278.20	847.48	World	OWID_WRL	847.48
5	1885	-0.333333	World	OWID_WRL	278.70	857.35	World	OWID_WRL	857.35
10	1890	-0.347500	World	OWID_WRL	279.10	867.22	World	OWID_WRL	867.22
15	1895	-0.224722	World	OWID_WRL	279.50	878.76	World	OWID_WRL	878.76
20	1900	-0.081667	World	OWID_WRL	279.80	890.30	World	OWID_WRL	890.30
25	1905	-0.254722	World	OWID_WRL	280.30	912.07	World	OWID_WRL	912.07
30	1910	-0.430556	World	OWID_WRL	281.00	935.46	World	OWID_WRL	935.46
35	1915	-0.136389	World	OWID_WRL	281.80	961.48	World	OWID_WRL	961.48
40	1920	-0.271667	World	OWID_WRL	282.90	990.23	World	OWID_WRL	990.23
45	1925	-0.216111	World	OWID_WRL	284.00	1020.20	World	OWID_WRL	1020.20
50	1930	-0.150556	World	OWID_WRL	285.00	1049.05	World	OWID_WRL	1049.05
55	1935	-0.193056	World	OWID_WRL	285.90	1076.54	World	OWID_WRL	1076.54
60	1940	0.133333	World	OWID_WRL	286.70	1102.40	World	OWID_WRL	1102.40
65	1945	0.095556	World	OWID_WRL	287.80	1128.83	World	OWID_WRL	1128.83
70	1950	-0.176667	World	OWID_WRL	289.00	1161.73	World	OWID_WRL	1161.73
75	1955	-0.146944	World	OWID_WRL	290.10	1207.03	World	OWID_WRL	1207.03
80	1960	-0.025000	World	OWID_WRL	291.40	1262.97	World	OWID_WRL	1262.97
85	1965	-0.105833	World	OWID_WRL	292.90	1328.47	World	OWID_WRL	1328.47
90	1970	0.025833	World	OWID_WRL	294.90	1403.19	World	OWID_WRL	1403.19
95	1975	-0.014722	World	OWID_WRL	297.40	1483.57	World	OWID_WRL	1483.57
98	1978	0.068056	World	OWID_WRL	298.82	1532.77	World	OWID_WRL	1532.77
99	1979	0.166667	World	OWID_WRL	300.04	1549.53	World	OWID_WRL	1549.53

	Year	temperature_anomaly	Entity_x	Code_x	N2O concentrations (annual average) (EEA, 2019)	CH4 concentration (EEA & NOAA (2019))_x	Entity	Code	CH4 concentration (EEA & NOAA (2019))_y
100	1980	0.258889	World	OWID_WRL	300.65	1566.28	World	OWID_WRL	1566.28
101	1981	0.321667	World	OWID_WRL	301.23	1583.48	World	OWID_WRL	1583.48
102	1982	0.142500	World	OWID_WRL	303.56	1600.69	World	OWID_WRL	1600.69
103	1983	0.315833	World	OWID_WRL	303.78	1617.89	World	OWID_WRL	1617.89
104	1984	0.157222	World	OWID_WRL	304.02	1635.09	World	OWID_WRL	1635.09
105	1985	0.116667	World	OWID_WRL	304.54	1652.29	World	OWID_WRL	1652.29
106	1986	0.182500	World	OWID_WRL	305.37	1669.49	World	OWID_WRL	1669.49
107	1987	0.325556	World	OWID_WRL	305.55	1680.66	World	OWID_WRL	1680.66
108	1988	0.389444	World	OWID_WRL	306.49	1698.83	World	OWID_WRL	1698.83
109	1989	0.271111	World	OWID_WRL	307.48	1710.52	World	OWID_WRL	1710.52
110	1990	0.449722	World	OWID_WRL	308.78	1709.33	World	OWID_WRL	1709.33
111	1991	0.405556	World	OWID_WRL	309.57	1729.07	World	OWID_WRL	1729.07
112	1992	0.221944	World	OWID_WRL	310.00	1731.05	World	OWID_WRL	1731.05
113	1993	0.234444	World	OWID_WRL	310.25	1735.65	World	OWID_WRL	1735.65
114	1994	0.317778	World	OWID_WRL	310.98	1741.66	World	OWID_WRL	1741.66
115	1995	0.447222	World	OWID_WRL	311.78	1747.10	World	OWID_WRL	1747.10
116	1996	0.327222	World	OWID_WRL	312.81	1749.86	World	OWID_WRL	1749.86
117	1997	0.465000	World	OWID_WRL	313.53	1753.94	World	OWID_WRL	1753.94
118	1998	0.611389	World	OWID_WRL	314.20	1762.43	World	OWID_WRL	1762.43
119	1999	0.383889	World	OWID_WRL	315.15	1772.33	World	OWID_WRL	1772.33
120	2000	0.394722	World	OWID_WRL	316.14	1774.07	World	OWID_WRL	1774.07
121	2001	0.537778	World	OWID_WRL	316.89	1772.95	World	OWID_WRL	1772.95
122	2002	0.629167	World	OWID_WRL	317.47	1773.14	World	OWID_WRL	1773.14
123	2003	0.620278	World	OWID_WRL	318.21	1777.41	World	OWID_WRL	1777.41

	Year	temperature_anomaly	Entity_x	Code_x	N2O concentrations (annual average) (EEA, 2019)	CH4 concentration (EEA & NOAA (2019))_x	Entity	Code	CH4 concentration (EEA & NOAA (2019))_y
124	2004	0.536667	World	OWID_WRL	318.93	1775.44	World	OWID_WRL	1775.44
125	2005	0.678056	World	OWID_WRL	319.60	1774.55	World	OWID_WRL	1774.55
126	2006	0.638611	World	OWID_WRL	320.37	1776.40	World	OWID_WRL	1776.40
127	2007	0.663889	World	OWID_WRL	321.14	1781.75	World	OWID_WRL	1781.75
128	2008	0.545000	World	OWID_WRL	322.11	1789.94	World	OWID_WRL	1789.94
129	2009	0.658889	World	OWID_WRL	322.88	1793.63	World	OWID_WRL	1793.63
130	2010	0.723056	World	OWID_WRL	323.70	1796.84	World	OWID_WRL	1796.84
131	2011	0.607500	World	OWID_WRL	324.61	1803.42	World	OWID_WRL	1803.42
132	2012	0.648056	World	OWID_WRL	325.58	1810.33	World	OWID_WRL	1810.33
133	2013	0.677500	World	OWID_WRL	326.53	1815.44	World	OWID_WRL	1815.44
134	2014	0.745833	World	OWID_WRL	327.61	1824.40	World	OWID_WRL	1824.40
135	2015	0.901111	World	OWID_WRL	328.51	1834.63	World	OWID_WRL	1834.63
136	2016	1.019444	World	OWID_WRL	329.29	1842.40	World	OWID_WRL	1842.40

In [40]:

```
CO2Concentration2 = pd.read_csv('CO2Concentration.csv')
```

In [41]:

```
graph3 = pd.merge(graph2, CO2Concentration, how='left', on ='Year')
```

```
C:\Users\spark\AppData\Local\Temp/ipykernel_32448/2934077127.py:1: FutureWarning: Passing 'suffixes' which cause duplic
ate columns {'Entity_x', 'Code_x'} in the result is deprecated and will raise a MergeError in a future version.
  graph3 = pd.merge(graph2, CO2Concentration, how='left', on ='Year')
```

In [42]:
```
graph3 = graph3.dropna()
```

In [43]:
```
graph3
```

Out[43]:

	Year	temperature_anomaly	Entity_x	Code_x	N2O concentrations (annual average) (EEA, 2019)	CH4 concentration (EEA & NOAA (2019))_x	Entity_x	Code_x	CH4 concentration (EEA & NOAA (2019))_y	Entity_y	Code_y	con (N
0	1880	-0.161944	World	OWID_WRL	278.20	847.48	World	OWID_WRL	847.48	World	OWID_WRL	
2	1890	-0.347500	World	OWID_WRL	279.10	867.22	World	OWID_WRL	867.22	World	OWID_WRL	
4	1900	-0.081667	World	OWID_WRL	279.80	890.30	World	OWID_WRL	890.30	World	OWID_WRL	
5	1905	-0.254722	World	OWID_WRL	280.30	912.07	World	OWID_WRL	912.07	World	OWID_WRL	
6	1910	-0.430556	World	OWID_WRL	281.00	935.46	World	OWID_WRL	935.46	World	OWID_WRL	
8	1920	-0.271667	World	OWID_WRL	282.90	990.23	World	OWID_WRL	990.23	World	OWID_WRL	
9	1925	-0.216111	World	OWID_WRL	284.00	1020.20	World	OWID_WRL	1020.20	World	OWID_WRL	
11	1935	-0.193056	World	OWID_WRL	285.90	1076.54	World	OWID_WRL	1076.54	World	OWID_WRL	
12	1940	0.133333	World	OWID_WRL	286.70	1102.40	World	OWID_WRL	1102.40	World	OWID_WRL	
13	1945	0.095556	World	OWID_WRL	287.80	1128.83	World	OWID_WRL	1128.83	World	OWID_WRL	
14	1950	-0.176667	World	OWID_WRL	289.00	1161.73	World	OWID_WRL	1161.73	World	OWID_WRL	
15	1955	-0.146944	World	OWID_WRL	290.10	1207.03	World	OWID_WRL	1207.03	World	OWID_WRL	
16	1960	-0.025000	World	OWID_WRL	291.40	1262.97	World	OWID_WRL	1262.97	World	OWID_WRL	
17	1965	-0.105833	World	OWID_WRL	292.90	1328.47	World	OWID_WRL	1328.47	World	OWID_WRL	
18	1970	0.025833	World	OWID_WRL	294.90	1403.19	World	OWID_WRL	1403.19	World	OWID_WRL	
19	1975	-0.014722	World	OWID_WRL	297.40	1483.57	World	OWID_WRL	1483.57	World	OWID_WRL	
20	1978	0.068056	World	OWID_WRL	298.82	1532.77	World	OWID_WRL	1532.77	World	OWID_WRL	
21	1979	0.166667	World	OWID_WRL	300.04	1549.53	World	OWID_WRL	1549.53	World	OWID_WRL	
22	1980	0.258889	World	OWID_WRL	300.65	1566.28	World	OWID_WRL	1566.28	World	OWID_WRL	
23	1981	0.321667	World	OWID_WRL	301.23	1583.48	World	OWID_WRL	1583.48	World	OWID_WRL	
24	1982	0.142500	World	OWID_WRL	303.56	1600.69	World	OWID_WRL	1600.69	World	OWID_WRL	
25	1983	0.315833	World	OWID_WRL	303.78	1617.89	World	OWID_WRL	1617.89	World	OWID_WRL	

	Year	temperature_anomaly	Entity_x	Code_x	N2O concentrations (annual average) (EEA, 2019)	CH4 concentration (EEA & NOAA (2019))_x	Entity_x	Code_x	CH4 concentration (EEA & NOAA (2019))_y	Entity_y	Code_y	con (N
26	1984	0.157222	World	OWID_WRL	304.02	1635.09	World	OWID_WRL	1635.09	World	OWID_WRL	
27	1985	0.116667	World	OWID_WRL	304.54	1652.29	World	OWID_WRL	1652.29	World	OWID_WRL	
28	1986	0.182500	World	OWID_WRL	305.37	1669.49	World	OWID_WRL	1669.49	World	OWID_WRL	
29	1987	0.325556	World	OWID_WRL	305.55	1680.66	World	OWID_WRL	1680.66	World	OWID_WRL	
30	1988	0.389444	World	OWID_WRL	306.49	1698.83	World	OWID_WRL	1698.83	World	OWID_WRL	
31	1989	0.271111	World	OWID_WRL	307.48	1710.52	World	OWID_WRL	1710.52	World	OWID_WRL	
32	1990	0.449722	World	OWID_WRL	308.78	1709.33	World	OWID_WRL	1709.33	World	OWID_WRL	
33	1991	0.405556	World	OWID_WRL	309.57	1729.07	World	OWID_WRL	1729.07	World	OWID_WRL	
34	1992	0.221944	World	OWID_WRL	310.00	1731.05	World	OWID_WRL	1731.05	World	OWID_WRL	
35	1993	0.234444	World	OWID_WRL	310.25	1735.65	World	OWID_WRL	1735.65	World	OWID_WRL	
36	1994	0.317778	World	OWID_WRL	310.98	1741.66	World	OWID_WRL	1741.66	World	OWID_WRL	
37	1995	0.447222	World	OWID_WRL	311.78	1747.10	World	OWID_WRL	1747.10	World	OWID_WRL	
38	1996	0.327222	World	OWID_WRL	312.81	1749.86	World	OWID_WRL	1749.86	World	OWID_WRL	
39	1997	0.465000	World	OWID_WRL	313.53	1753.94	World	OWID_WRL	1753.94	World	OWID_WRL	
40	1998	0.611389	World	OWID_WRL	314.20	1762.43	World	OWID_WRL	1762.43	World	OWID_WRL	
41	1999	0.383889	World	OWID_WRL	315.15	1772.33	World	OWID_WRL	1772.33	World	OWID_WRL	
42	2000	0.394722	World	OWID_WRL	316.14	1774.07	World	OWID_WRL	1774.07	World	OWID_WRL	
43	2001	0.537778	World	OWID_WRL	316.89	1772.95	World	OWID_WRL	1772.95	World	OWID_WRL	
44	2002	0.629167	World	OWID_WRL	317.47	1773.14	World	OWID_WRL	1773.14	World	OWID_WRL	
45	2003	0.620278	World	OWID_WRL	318.21	1777.41	World	OWID_WRL	1777.41	World	OWID_WRL	
46	2004	0.536667	World	OWID_WRL	318.93	1775.44	World	OWID_WRL	1775.44	World	OWID_WRL	
47	2005	0.678056	World	OWID_WRL	319.60	1774.55	World	OWID_WRL	1774.55	World	OWID_WRL	
48	2006	0.638611	World	OWID_WRL	320.37	1776.40	World	OWID_WRL	1776.40	World	OWID_WRL	
49	2007	0.663889	World	OWID_WRL	321.14	1781.75	World	OWID_WRL	1781.75	World	OWID_WRL	

	Year	temperature_anomaly	Entity_x	Code_x	N2O concentrations (annual average) (EEA, 2019)	CH4 concentration (EEA & NOAA (2019))_x	Entity_x	Code_x	CH4 concentration (EEA & NOAA (2019))_y	Entity_y	Code_y	con (N
50	2008	0.545000	World	OWID_WRL	322.11	1789.94	World	OWID_WRL	1789.94	World	OWID_WRL	
51	2009	0.658889	World	OWID_WRL	322.88	1793.63	World	OWID_WRL	1793.63	World	OWID_WRL	
52	2010	0.723056	World	OWID_WRL	323.70	1796.84	World	OWID_WRL	1796.84	World	OWID_WRL	
53	2011	0.607500	World	OWID_WRL	324.61	1803.42	World	OWID_WRL	1803.42	World	OWID_WRL	
54	2012	0.648056	World	OWID_WRL	325.58	1810.33	World	OWID_WRL	1810.33	World	OWID_WRL	
55	2013	0.677500	World	OWID_WRL	326.53	1815.44	World	OWID_WRL	1815.44	World	OWID_WRL	
56	2014	0.745833	World	OWID_WRL	327.61	1824.40	World	OWID_WRL	1824.40	World	OWID_WRL	
57	2015	0.901111	World	OWID_WRL	328.51	1834.63	World	OWID_WRL	1834.63	World	OWID_WRL	
58	2016	1.019444	World	OWID_WRL	329.29	1842.40	World	OWID_WRL	1842.40	World	OWID_WRL	

In [44]:
```
graph3=graph3.drop(['Entity_x','CH4 concentration (EEA & NOAA (2019))_y', 'Entity_y','Code_y'], axis=1)
```

In [45]:
```
graph3.columns
```

Out[45]:
```
Index(['Year', 'temperature_anomaly', 'Code_x',
       'N2O concentrations (annual average) (EEA, 2019)',
       'CH4 concentration (EEA & NOAA (2019))_x', 'Code_x',
       'CO2 concentrations (NOAA, 2018)'],
      dtype='object')
```

In [46]:
```
graph3.drop(['Code_x'],axis=1) #apostrophe.
```

Out[46]:

	Year	temperature_anomaly	N2O concentrations (annual average) (EEA, 2019)	CH4 concentration (EEA & NOAA (2019))_x	CO2 concentrations (NOAA, 2018)
0	1880	-0.161944	278.20	847.48	287.77
2	1890	-0.347500	279.10	867.22	290.92
4	1900	-0.081667	279.80	890.30	294.22
5	1905	-0.254722	280.30	912.07	299.02
6	1910	-0.430556	281.00	935.46	297.87
8	1920	-0.271667	282.90	990.23	301.88
9	1925	-0.216111	284.00	1020.20	304.84
11	1935	-0.193056	285.90	1076.54	306.32
12	1940	0.133333	286.70	1102.40	310.38
13	1945	0.095556	287.80	1128.83	310.94
14	1950	-0.176667	289.00	1161.73	312.83
15	1955	-0.146944	290.10	1207.03	314.71
16	1960	-0.025000	291.40	1262.97	316.91
17	1965	-0.105833	292.90	1328.47	320.04
18	1970	0.025833	294.90	1403.19	325.68
19	1975	-0.014722	297.40	1483.57	331.11
20	1978	0.068056	298.82	1532.77	335.40
21	1979	0.166667	300.04	1549.53	336.84
22	1980	0.258889	300.65	1566.28	338.75
23	1981	0.321667	301.23	1583.48	340.11
24	1982	0.142500	303.56	1600.69	341.45
25	1983	0.315833	303.78	1617.89	343.05
26	1984	0.157222	304.02	1635.09	344.65

	Year	temperature_anomaly	N2O concentrations (annual average) (EEA, 2019)	CH4 concentration (EEA & NOAA (2019))_x	CO2 concentrations (NOAA, 2018)
27	1985	0.116667	304.54	1652.29	346.12
28	1986	0.182500	305.37	1669.49	347.42
29	1987	0.325556	305.55	1680.66	349.19
30	1988	0.389444	306.49	1698.83	351.57
31	1989	0.271111	307.48	1710.52	353.12
32	1990	0.449722	308.78	1709.33	354.39
33	1991	0.405556	309.57	1729.07	355.61
34	1992	0.221944	310.00	1731.05	356.45
35	1993	0.234444	310.25	1735.65	357.10
36	1994	0.317778	310.98	1741.66	358.83
37	1995	0.447222	311.78	1747.10	360.82
38	1996	0.327222	312.81	1749.86	362.61
39	1997	0.465000	313.53	1753.94	363.73
40	1998	0.611389	314.20	1762.43	366.70
41	1999	0.383889	315.15	1772.33	368.38
42	2000	0.394722	316.14	1774.07	369.55
43	2001	0.537778	316.89	1772.95	371.14
44	2002	0.629167	317.47	1773.14	373.28
45	2003	0.620278	318.21	1777.41	375.80
46	2004	0.536667	318.93	1775.44	377.52
47	2005	0.678056	319.60	1774.55	379.80
48	2006	0.638611	320.37	1776.40	381.90
49	2007	0.663889	321.14	1781.75	383.79
50	2008	0.545000	322.11	1789.94	385.60
51	2009	0.658889	322.88	1793.63	387.43

	Year	temperature_anomaly	N2O concentrations (annual average) (EEA, 2019)	CH4 concentration (EEA & NOAA (2019))_x	CO2 concentrations (NOAA, 2018)
52	2010	0.723056	323.70	1796.84	389.90
53	2011	0.607500	324.61	1803.42	391.65
54	2012	0.648056	325.58	1810.33	393.85
55	2013	0.677500	326.53	1815.44	396.52
56	2014	0.745833	327.61	1824.40	398.65
57	2015	0.901111	328.51	1834.63	400.83
58	2016	1.019444	329.29	1842.40	404.24

In [47]:

```
globalwarming_data = graph3
```

In [48]:

```
globalwarming_data.columns
```

Out[48]:

```
Index(['Year', 'temperature_anomaly', 'Code_x',
       'N2O concentrations (annual average) (EEA, 2019)',
       'CH4 concentration (EEA & NOAA (2019))_x', 'Code_x',
       'CO2 concentrations (NOAA, 2018)'],
      dtype='object')
```

In [49]:
```
globalwarming_data.rename(columns = {'N2O concentrations (annual average) (EEA, 2019)':'N20 concentration yearly',
                                     'CH4 concentration (EEA & NOAA (2019))_x':'CH4 concentration yearly',
                                     'CO2 concentrations (NOAA, 2018)':'C02 concenration yearly'
                                    }, inplace = True)

```

```
C:\ProgramData\Anaconda3\lib\site-packages\pandas\core\frame.py:5039: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy)
  return super().rename(
```

In [50]:
```
globalwarming_data
```

Out[50]:

	Year	temperature_anomaly	Code_x	N20 concentration yearly	CH4 concentration yearly	Code_x	C02 concenration yearly
0	1880	-0.161944	OWID_WRL	278.20	847.48	OWID_WRL	287.77
2	1890	-0.347500	OWID_WRL	279.10	867.22	OWID_WRL	290.92
4	1900	-0.081667	OWID_WRL	279.80	890.30	OWID_WRL	294.22
5	1905	-0.254722	OWID_WRL	280.30	912.07	OWID_WRL	299.02
6	1910	-0.430556	OWID_WRL	281.00	935.46	OWID_WRL	297.87
8	1920	-0.271667	OWID_WRL	282.90	990.23	OWID_WRL	301.88
9	1925	-0.216111	OWID_WRL	284.00	1020.20	OWID_WRL	304.84
11	1935	-0.193056	OWID_WRL	285.90	1076.54	OWID_WRL	306.32
12	1940	0.133333	OWID_WRL	286.70	1102.40	OWID_WRL	310.38
13	1945	0.095556	OWID_WRL	287.80	1128.83	OWID_WRL	310.94
14	1950	-0.176667	OWID_WRL	289.00	1161.73	OWID_WRL	312.83
15	1955	-0.146944	OWID_WRL	290.10	1207.03	OWID_WRL	314.71
16	1960	-0.025000	OWID_WRL	291.40	1262.97	OWID_WRL	316.91
17	1965	-0.105833	OWID_WRL	292.90	1328.47	OWID_WRL	320.04
18	1970	0.025833	OWID_WRL	294.90	1403.19	OWID_WRL	325.68
19	1975	-0.014722	OWID_WRL	297.40	1483.57	OWID_WRL	331.11
20	1978	0.068056	OWID_WRL	298.82	1532.77	OWID_WRL	335.40
21	1979	0.166667	OWID_WRL	300.04	1549.53	OWID_WRL	336.84
22	1980	0.258889	OWID_WRL	300.65	1566.28	OWID_WRL	338.75
23	1981	0.321667	OWID_WRL	301.23	1583.48	OWID_WRL	340.11
24	1982	0.142500	OWID_WRL	303.56	1600.69	OWID_WRL	341.45
25	1983	0.315833	OWID_WRL	303.78	1617.89	OWID_WRL	343.05
26	1984	0.157222	OWID_WRL	304.02	1635.09	OWID_WRL	344.65
27	1985	0.116667	OWID_WRL	304.54	1652.29	OWID_WRL	346.12

	Year	temperature_anomaly	Code_x	N20 concentration yearly	CH4 concentration yearly	Code_x	C02 concenration yearly
28	1986	0.182500	OWID_WRL	305.37	1669.49	OWID_WRL	347.42
29	1987	0.325556	OWID_WRL	305.55	1680.66	OWID_WRL	349.19
30	1988	0.389444	OWID_WRL	306.49	1698.83	OWID_WRL	351.57
31	1989	0.271111	OWID_WRL	307.48	1710.52	OWID_WRL	353.12
32	1990	0.449722	OWID_WRL	308.78	1709.33	OWID_WRL	354.39
33	1991	0.405556	OWID_WRL	309.57	1729.07	OWID_WRL	355.61
34	1992	0.221944	OWID_WRL	310.00	1731.05	OWID_WRL	356.45
35	1993	0.234444	OWID_WRL	310.25	1735.65	OWID_WRL	357.10
36	1994	0.317778	OWID_WRL	310.98	1741.66	OWID_WRL	358.83
37	1995	0.447222	OWID_WRL	311.78	1747.10	OWID_WRL	360.82
38	1996	0.327222	OWID_WRL	312.81	1749.86	OWID_WRL	362.61
39	1997	0.465000	OWID_WRL	313.53	1753.94	OWID_WRL	363.73
40	1998	0.611389	OWID_WRL	314.20	1762.43	OWID_WRL	366.70
41	1999	0.383889	OWID_WRL	315.15	1772.33	OWID_WRL	368.38
42	2000	0.394722	OWID_WRL	316.14	1774.07	OWID_WRL	369.55
43	2001	0.537778	OWID_WRL	316.89	1772.95	OWID_WRL	371.14
44	2002	0.629167	OWID_WRL	317.47	1773.14	OWID_WRL	373.28
45	2003	0.620278	OWID_WRL	318.21	1777.41	OWID_WRL	375.80
46	2004	0.536667	OWID_WRL	318.93	1775.44	OWID_WRL	377.52
47	2005	0.678056	OWID_WRL	319.60	1774.55	OWID_WRL	379.80
48	2006	0.638611	OWID_WRL	320.37	1776.40	OWID_WRL	381.90
49	2007	0.663889	OWID_WRL	321.14	1781.75	OWID_WRL	383.79
50	2008	0.545000	OWID_WRL	322.11	1789.94	OWID_WRL	385.60
51	2009	0.658889	OWID_WRL	322.88	1793.63	OWID_WRL	387.43
52	2010	0.723056	OWID_WRL	323.70	1796.84	OWID_WRL	389.90
53	2011	0.607500	OWID_WRL	324.61	1803.42	OWID_WRL	391.65

	Year	temperature_anomaly	Code_x	N20 concentration yearly	CH4 concentration yearly	Code_x	C02 concenration yearly
54	2012	0.648056	OWID_WRL	325.58	1810.33	OWID_WRL	393.85
55	2013	0.677500	OWID_WRL	326.53	1815.44	OWID_WRL	396.52
56	2014	0.745833	OWID_WRL	327.61	1824.40	OWID_WRL	398.65
57	2015	0.901111	OWID_WRL	328.51	1834.63	OWID_WRL	400.83
58	2016	1.019444	OWID_WRL	329.29	1842.40	OWID_WRL	404.24

In [51]:

```
globalwarming_data.head() #So we can make a predicion of what the temperature given the input values of year and
#carbon dioxide.
#For exmaple, if the carbon dioxide production was reduced by 20 percent, what would be the expected temperature?

```

Out[51]:

	Year	temperature_anomaly	Code_x	N20 concentration yearly	CH4 concentration yearly	Code_x	C02 concenration yearly
0	1880	-0.161944	OWID_WRL	278.2	847.48	OWID_WRL	287.77
2	1890	-0.347500	OWID_WRL	279.1	867.22	OWID_WRL	290.92
4	1900	-0.081667	OWID_WRL	279.8	890.30	OWID_WRL	294.22
5	1905	-0.254722	OWID_WRL	280.3	912.07	OWID_WRL	299.02
6	1910	-0.430556	OWID_WRL	281.0	935.46	OWID_WRL	297.87

In [52]:

```
globalwarming_data.columns
```

Out[52]:

```
Index(['Year', 'temperature_anomaly', 'Code_x', 'N20 concentration yearly',
       'CH4 concentration yearly', 'Code_x', 'C02 concenration yearly'],
      dtype='object')
```

In [53]:

```
import findspark

findspark.init()

import pyspark
```

In [54]:
```
from pyspark.sql import SparkSession
```

In [55]:
```
spark = SparkSession.builder.appName('GlobalWarmingProject').getOrCreate()
```

In [56]:
```
from pyspark.ml.regression import LinearRegression
```

In [57]:
```
globalwarming_data = globalwarming_data[['Year','temperature_anomaly','N2O concentration yearly','CH4 concentration
                                        ,'CO2 concenration yearly']]
```

In [58]:
```
globalwarming_data =globalwarming_data.dropna()
```

In [59]:

```
globalwarming_data
```

Out[59]:

	Year	temperature_anomaly	N20 concentration yearly	CH4 concentration yearly	C02 concenration yearly
0	1880	-0.161944	278.20	847.48	287.77
2	1890	-0.347500	279.10	867.22	290.92
4	1900	-0.081667	279.80	890.30	294.22
5	1905	-0.254722	280.30	912.07	299.02
6	1910	-0.430556	281.00	935.46	297.87
8	1920	-0.271667	282.90	990.23	301.88
9	1925	-0.216111	284.00	1020.20	304.84
11	1935	-0.193056	285.90	1076.54	306.32
12	1940	0.133333	286.70	1102.40	310.38
13	1945	0.095556	287.80	1128.83	310.94
14	1950	-0.176667	289.00	1161.73	312.83
15	1955	-0.146944	290.10	1207.03	314.71
16	1960	-0.025000	291.40	1262.97	316.91
17	1965	-0.105833	292.90	1328.47	320.04
18	1970	0.025833	294.90	1403.19	325.68
19	1975	-0.014722	297.40	1483.57	331.11
20	1978	0.068056	298.82	1532.77	335.40
21	1979	0.166667	300.04	1549.53	336.84
22	1980	0.258889	300.65	1566.28	338.75
23	1981	0.321667	301.23	1583.48	340.11
24	1982	0.142500	303.56	1600.69	341.45
25	1983	0.315833	303.78	1617.89	343.05
26	1984	0.157222	304.02	1635.09	344.65
27	1985	0.116667	304.54	1652.29	346.12

	Year	temperature_anomaly	N20 concentration yearly	CH4 concentration yearly	C02 concenration yearly
28	1986	0.182500	305.37	1669.49	347.42
29	1987	0.325556	305.55	1680.66	349.19
30	1988	0.389444	306.49	1698.83	351.57
31	1989	0.271111	307.48	1710.52	353.12
32	1990	0.449722	308.78	1709.33	354.39
33	1991	0.405556	309.57	1729.07	355.61
34	1992	0.221944	310.00	1731.05	356.45
35	1993	0.234444	310.25	1735.65	357.10
36	1994	0.317778	310.98	1741.66	358.83
37	1995	0.447222	311.78	1747.10	360.82
38	1996	0.327222	312.81	1749.86	362.61
39	1997	0.465000	313.53	1753.94	363.73
40	1998	0.611389	314.20	1762.43	366.70
41	1999	0.383889	315.15	1772.33	368.38
42	2000	0.394722	316.14	1774.07	369.55
43	2001	0.537778	316.89	1772.95	371.14
44	2002	0.629167	317.47	1773.14	373.28
45	2003	0.620278	318.21	1777.41	375.80
46	2004	0.536667	318.93	1775.44	377.52
47	2005	0.678056	319.60	1774.55	379.80
48	2006	0.638611	320.37	1776.40	381.90
49	2007	0.663889	321.14	1781.75	383.79
50	2008	0.545000	322.11	1789.94	385.60
51	2009	0.658889	322.88	1793.63	387.43
52	2010	0.723056	323.70	1796.84	389.90
53	2011	0.607500	324.61	1803.42	391.65

	Year	temperature_anomaly	N20 concentration yearly	CH4 concentration yearly	C02 concenration yearly
54	2012	0.648056	325.58	1810.33	393.85
55	2013	0.677500	326.53	1815.44	396.52
56	2014	0.745833	327.61	1824.40	398.65
57	2015	0.901111	328.51	1834.63	400.83
58	2016	1.019444	329.29	1842.40	404.24

In [60]:

```
df = pd.DataFrame(data=list(globalwarming_data))

```

In [61]:

```
from pyspark.ml.linalg import Vectors
from pyspark.ml.feature import VectorAssembler
```

In [62]:

```python
from pyspark.sql import SparkSession
#Create PySpark SparkSession
spark = SparkSession.builder \
    .master("local[1]") \
    .appName("SparkByExamples.com") \
    .getOrCreate()
#Create PySpark DataFrame from Pandas
sparkGlobalwarming=spark.createDataFrame(globalwarming_data)
sparkGlobalwarming.printSchema()
sparkGlobalwarming.show()
```

```
root
 |-- Year: long (nullable = true)
 |-- temperature_anomaly: double (nullable = true)
 |-- N20 concentration yearly: double (nullable = true)
 |-- CH4 concentration yearly: double (nullable = true)
 |-- C02 concenration yearly: double (nullable = true)

+----+--------------------+------------------------+------------------------+-----------------------+
|Year| temperature_anomaly|N20 concentration yearly|CH4 concentration yearly|C02 concenration yearly|
+----+--------------------+------------------------+------------------------+-----------------------+
|1880|-0.16194444444444445|                   278.2|                  847.48|                 287.77|
|1890|             -0.3475|                   279.1|                  867.22|                 290.92|
|1900|-0.08166666666666667|                   279.8|                   890.3|                 294.22|
|1905|-0.25472222222222224|                   280.3|                  912.07|                 299.02|
|1910| -0.4305555555555556|                   281.0|                  935.46|                 297.87|
|1920|-0.27166666666666667|                   282.9|                  990.23|                 301.88|
|1925|-0.21611111111111111|                   284.0|                  1020.2|                 304.84|
|1935|-0.19305555555555556|                   285.9|                 1076.54|                 306.32|
|1940| 0.13333333333333333|                   286.7|                  1102.4|                 310.38|
|1945| 0.09555555555555556|                   287.8|                 1128.83|                 310.94|
|1950|-0.17666666666666667|                   289.0|                 1161.73|                 312.83|
|1955|-0.14694444444444443|                   290.1|                 1207.03|                 314.71|
|1960|              -0.025|                   291.4|                 1262.97|                 316.91|
|1965|-0.10583333333333333|                   292.9|                 1328.47|                 320.04|
|1970|0.025833333333333333|                   294.9|                 1403.19|                 325.68|
|1975|-0.01472222222222...|                   297.4|                 1483.57|                 331.11|
|1978| 0.06805555555555556|                  298.82|                 1532.77|                  335.4|
|1979| 0.16666666666666666|                  300.04|                 1549.53|                 336.84|
```

```
|1980|  0.2588888888888889|                 300.65|                1566.28|               338.75|
|1981| 0.32166666666666666|                 301.23|                1583.48|               340.11|
+----+--------------------+-----------------------+-----------------------+----------------------+
only showing top 20 rows
```

In [63]:
```
sparkGlobalwarming.columns
```

Out[63]:
```
['Year',
 'temperature_anomaly',
 'N20 concentration yearly',
 'CH4 concentration yearly',
 'C02 concenration yearly']
```

In [64]:
```
sparkGlobalwarming.count()
```

Out[64]:
```
55
```

In [65]:
```
sparkGlobalwarming.dtypes
```

Out[65]:
```
[('Year', 'bigint'),
 ('temperature_anomaly', 'double'),
 ('N20 concentration yearly', 'double'),
 ('CH4 concentration yearly', 'double'),
 ('C02 concenration yearly', 'double')]
```

In [66]:
```
assembler = VectorAssembler(inputCols = ['Year','N20 concentration yearly',
                                        'CH4 concentration yearly',
                                       'C02 concenration yearly' ],outputCol='features')
```

In [67]:

```
globalwarming_data
```

Out[67]:

	Year	temperature_anomaly	N20 concentration yearly	CH4 concentration yearly	C02 concenration yearly
0	1880	-0.161944	278.20	847.48	287.77
2	1890	-0.347500	279.10	867.22	290.92
4	1900	-0.081667	279.80	890.30	294.22
5	1905	-0.254722	280.30	912.07	299.02
6	1910	-0.430556	281.00	935.46	297.87
8	1920	-0.271667	282.90	990.23	301.88
9	1925	-0.216111	284.00	1020.20	304.84
11	1935	-0.193056	285.90	1076.54	306.32
12	1940	0.133333	286.70	1102.40	310.38
13	1945	0.095556	287.80	1128.83	310.94
14	1950	-0.176667	289.00	1161.73	312.83
15	1955	-0.146944	290.10	1207.03	314.71
16	1960	-0.025000	291.40	1262.97	316.91
17	1965	-0.105833	292.90	1328.47	320.04
18	1970	0.025833	294.90	1403.19	325.68
19	1975	-0.014722	297.40	1483.57	331.11
20	1978	0.068056	298.82	1532.77	335.40
21	1979	0.166667	300.04	1549.53	336.84
22	1980	0.258889	300.65	1566.28	338.75
23	1981	0.321667	301.23	1583.48	340.11
24	1982	0.142500	303.56	1600.69	341.45
25	1983	0.315833	303.78	1617.89	343.05
26	1984	0.157222	304.02	1635.09	344.65
27	1985	0.116667	304.54	1652.29	346.12

	Year	temperature_anomaly	N20 concentration yearly	CH4 concentration yearly	C02 concenration yearly
28	1986	0.182500	305.37	1669.49	347.42
29	1987	0.325556	305.55	1680.66	349.19
30	1988	0.389444	306.49	1698.83	351.57
31	1989	0.271111	307.48	1710.52	353.12
32	1990	0.449722	308.78	1709.33	354.39
33	1991	0.405556	309.57	1729.07	355.61
34	1992	0.221944	310.00	1731.05	356.45
35	1993	0.234444	310.25	1735.65	357.10
36	1994	0.317778	310.98	1741.66	358.83
37	1995	0.447222	311.78	1747.10	360.82
38	1996	0.327222	312.81	1749.86	362.61
39	1997	0.465000	313.53	1753.94	363.73
40	1998	0.611389	314.20	1762.43	366.70
41	1999	0.383889	315.15	1772.33	368.38
42	2000	0.394722	316.14	1774.07	369.55
43	2001	0.537778	316.89	1772.95	371.14
44	2002	0.629167	317.47	1773.14	373.28
45	2003	0.620278	318.21	1777.41	375.80
46	2004	0.536667	318.93	1775.44	377.52
47	2005	0.678056	319.60	1774.55	379.80
48	2006	0.638611	320.37	1776.40	381.90
49	2007	0.663889	321.14	1781.75	383.79
50	2008	0.545000	322.11	1789.94	385.60
51	2009	0.658889	322.88	1793.63	387.43
52	2010	0.723056	323.70	1796.84	389.90
53	2011	0.607500	324.61	1803.42	391.65

	Year	temperature_anomaly	N20 concentration yearly	CH4 concentration yearly	C02 concenration yearly
54	2012	0.648056	325.58	1810.33	393.85
55	2013	0.677500	326.53	1815.44	396.52
56	2014	0.745833	327.61	1824.40	398.65
57	2015	0.901111	328.51	1834.63	400.83
58	2016	1.019444	329.29	1842.40	404.24

In [68]:

```
sparkGlobalwarming.show()
```

```
+----+--------------------+------------------------+------------------------+-----------------------+
|Year| temperature_anomaly|N20 concentration yearly|CH4 concentration yearly|C02 concenration yearly|
+----+--------------------+------------------------+------------------------+-----------------------+
|1880|-0.16194444444444445|                   278.2|                  847.48|                 287.77|
|1890|             -0.3475|                   279.1|                  867.22|                 290.92|
|1900|-0.08166666666666667|                   279.8|                   890.3|                 294.22|
|1905|-0.25472222222222224|                   280.3|                  912.07|                 299.02|
|1910| -0.4305555555555556|                   281.0|                  935.46|                 297.87|
|1920|-0.27166666666666667|                   282.9|                  990.23|                 301.88|
|1925|-0.21611111111111111|                   284.0|                  1020.2|                 304.84|
|1935|-0.19305555555555556|                   285.9|                 1076.54|                 306.32|
|1940| 0.13333333333333333|                   286.7|                  1102.4|                 310.38|
|1945| 0.09555555555555556|                   287.8|                 1128.83|                 310.94|
|1950|-0.17666666666666667|                   289.0|                 1161.73|                 312.83|
|1955|-0.14694444444444443|                   290.1|                 1207.03|                 314.71|
|1960|              -0.025|                   291.4|                 1262.97|                 316.91|
|1965|-0.10583333333333333|                   292.9|                 1328.47|                 320.04|
|1970|0.025833333333333333|                   294.9|                 1403.19|                 325.68|
|1975|-0.01472222222222...|                   297.4|                 1483.57|                 331.11|
|1978| 0.06805555555555556|                  298.82|                 1532.77|                  335.4|
|1979| 0.16666666666666666|                  300.04|                 1549.53|                 336.84|
|1980|  0.2588888888888889|                  300.65|                 1566.28|                 338.75|
|1981| 0.32166666666666666|                  301.23|                 1583.48|                 340.11|
+----+--------------------+------------------------+------------------------+-----------------------+
only showing top 20 rows
```

In [69]:
```
sparkGlobalwarming.columns
```

Out[69]:
```
['Year',
 'temperature_anomaly',
 'N20 concentration yearly',
 'CH4 concentration yearly',
 'C02 concenration yearly']
```

In [82]:
```
output = assembler.transform(sparkGlobalwarming)

```

In [83]:
```
output.show() #Challenge of this project was when there was NA that caused some issues which was fixed.
```

```
+----+--------------------+-----------------------+-----------------------+-----------------------+---------------
-----+
|Year| temperature_anomaly|N20 concentration yearly|CH4 concentration yearly|C02 concenration yearly|            fea
tures|
+----+--------------------+-----------------------+-----------------------+-----------------------+---------------
-----+
|1880|-0.16194444444444445|                  278.2|                 847.48|                 287.77|[1880.0,278.2,8
47...|
|1890|             -0.3475|                  279.1|                 867.22|                 290.92|[1890.0,279.1,8
67...|
|1900|-0.08166666666666667|                  279.8|                  890.3|                 294.22|[1900.0,279.8,8
90...|
|1905|-0.25472222222222224|                  280.3|                 912.07|                 299.02|[1905.0,280.3,9
12...|
|1910| -0.4305555555555556|                  281.0|                 935.46|                 297.87|[1910.0,281.0,9
35...|
|1920|-0.27166666666666667|                  282.9|                 990.23|                 301.88|[1920.0,282.9,9
90...|
|1925|-0.21611111111111111|                  284.0|                 1020.2|                 304.84|[1925.0,284.0,1
02...|
|1935|-0.19305555555555556|                  285.9|                1076.54|                 306.32|[1935.0,285.9,1
07...|
|1940| 0.13333333333333333|                  286.7|                 1102.4|                 310.38|[1940.0,286.7,1
10...|
|1945| 0.09555555555555556|                  287.8|                1128.83|                 310.94|[1945.0,287.8,1
12...|
|1950|-0.17666666666666667|                  289.0|                1161.73|                 312.83|[1950.0,289.0,1
16...|
|1955|-0.14694444444444443|                  290.1|                1207.03|                 314.71|[1955.0,290.1,1
20...|
|1960|              -0.025|                  291.4|                1262.97|                 316.91|[1960.0,291.4,1
26...|
|1965|-0.10583333333333333|                  292.9|                1328.47|                 320.04|[1965.0,292.9,1
32...|
|1970|0.025833333333333333|                  294.9|                1403.19|                 325.68|[1970.0,294.9,1
40...|
|1975|-0.01472222222222...|                  297.4|                1483.57|                 331.11|[1975.0,297.4,1
48...|
|1978| 0.06805555555555556|                 298.82|                1532.77|                  335.4|[1978.0,298.82,
```

```
15...|
|1979| 0.16666666666666666|                 300.04|                 1549.53|                 336.84|[1979.0,300.04,
15...|
|1980|  0.2588888888888889|                 300.65|                 1566.28|                 338.75|[1980.0,300.65,
15...|
|1981| 0.32166666666666666|                 301.23|                 1583.48|                 340.11|[1981.0,301.23,
15...|
+----+--------------------+-----------------------+------------------------+-----------------------+---------------
-----+
only showing top 20 rows
```

In [84]:
```
output.count()
```

Out[84]: 55

In [85]:
```
final_data = output.select('features','temperature_anomaly')
```

In [86]:
```
final_data.count()
```

Out[86]: 55

In [87]:

```
final_data.show()
```

```
|[1905.0,280.5,912...| 0.2547222222222224|
|[1910.0,281.0,935...| -0.4305555555555556|
|[1920.0,282.9,990...|-0.2716666666666667|
|[1925.0,284.0,102...|-0.2161111111111111|
|[1935.0,285.9,107...|-0.1930555555555556|
|[1940.0,286.7,110...| 0.1333333333333333|
|[1945.0,287.8,112...| 0.0955555555555556|
|[1950.0,289.0,116...|-0.1766666666666667|
|[1955.0,290.1,120...|-0.1469444444444443|
|[1960.0,291.4,126...|             -0.025|
|[1965.0,292.9,132...|-0.1058333333333333|
|[1970.0,294.9,140...|0.0258333333333333|
|[1975.0,297.4,148...|-0.01472222222222...|
|[1978.0,298.82,15...| 0.0680555555555556|
|[1979.0,300.04,15...| 0.1666666666666666|
|[1980.0,300.65,15...|  0.2588888888888889|
|[1981.0,301.23,15...| 0.3216666666666666|
+--------------------+--------------------+
only showing top 20 rows
```

In [88]:

```
train_data, test_data = final_data.randomSplit([0.7,0.3])
```

In [89]:

```
train_data.describe().show()
```

```
+-------+------------------+
|summary|temperature_anomaly|
+-------+------------------+
|  count|                39|
|   mean|0.31376780626780626|
| stddev|0.35175754298948464|
|    min|-0.4305555555555556|
|    max| 1.0194444444444446|
+-------+------------------+
```

In [90]:
```
test_data.describe().show()
```

```
+-------+--------------------+
|summary| temperature_anomaly|
+-------+--------------------+
|  count|                  16|
|   mean| 0.21232638888888888|
| stddev|  0.3125180632639578|
|    min|-0.25472222222222224|
|    max|  0.7230555555555556|
+-------+--------------------+
```

In [91]:
```
from pyspark.ml.regression import LinearRegression
```

In [92]:
```
lr = LinearRegression(labelCol = 'temperature_anomaly')
```

In [93]:
```
lr_model = lr.fit(train_data)

```

In [94]:
```
test_results = lr_model.evaluate(test_data)
```

In [95]:
```
test_results.rootMeanSquaredError
```

Out[95]: 0.1398338699218393

In [96]:
```
test_results.r2

```

Out[96]: 0.7864482035925086

In [97]:
```
unlabeled_data = test_data.select('features')
```

In [98]:
```
predictions = lr_model.transform(unlabeled_data)
```

In [99]:
```
predictions.show()
```

```
+--------------------+--------------------+
|            features|          prediction|
+--------------------+--------------------+
|[1905.0,280.3,912...|-0.17454252553710248|
|[1940.0,286.7,110...|-0.19095483251770595|
|[1945.0,287.8,112...|-0.22269712701778666|
|[1950.0,289.0,116...|-0.22196737048810178|
|[1955.0,290.1,120...|-0.21061878593053862|
|[1960.0,291.4,126...|-0.18955573033147033|
|[1970.0,294.9,140...|-0.04978758667968819|
|[1975.0,297.4,148...|  0.0328454593779437|
|[1979.0,300.04,15...| 0.11633918359325257|
|[1983.0,303.78,16...| 0.18578792716349923|
|[1996.0,312.81,17...|  0.4117597565057096|
|[2000.0,316.14,17...|  0.4779089582750924|
|[2002.0,317.47,17...|  0.5162193596788853|
|[2008.0,322.11,17...|  0.6546051624206264|
|[2009.0,322.88,17...|  0.6732327282080988|
|[2010.0,323.7,179...|  0.7049996335091375|
+--------------------+--------------------+
```

In [100]:
```
test_data.show()
```

```
+--------------------+--------------------+
|            features| temperature_anomaly|
+--------------------+--------------------+
|[1905.0,280.3,912...|-0.25472222222222224|
|[1940.0,286.7,110...| 0.13333333333333333|
|[1945.0,287.8,112...| 0.09555555555555556|
|[1950.0,289.0,116...|-0.17666666666666667|
|[1955.0,290.1,120...|-0.14694444444444443|
|[1960.0,291.4,126...|              -0.025|
|[1970.0,294.9,140...|0.025833333333333333|
|[1975.0,297.4,148...|-0.01472222222222...|
|[1979.0,300.04,15...| 0.16666666666666666|
|[1983.0,303.78,16...|  0.3158333333333333|
|[1996.0,312.81,17...|  0.3272222222222222|
|[2000.0,316.14,17...| 0.39472222222222225|
|[2002.0,317.47,17...|  0.6291666666666668|
|[2008.0,322.11,17...|               0.545|
|[2009.0,322.88,17...|  0.6588888888888889|
|[2010.0,323.7,179...|  0.7230555555555556|
+--------------------+--------------------+
```

```
# We want to see the percentage difference of test data and predicted model.
```

we are going to use interpolating polynomials through matlab to predict the input values in future

#https://www.n2olevels.org/ (https://www.n2olevels.org/) #Interpolating polynomials: canonical form, Newton’s polynomial

In [97]:
```
#So we are going to use Newtons' Interpolating polynomial. And we will predict ax^3+bx^2+cx+d
#using 4 points. The equation for Newton's interpolating polynomial is:
#P2 = c0 + C1(x-x1) + C2(x-x1)(x-x2)+c3(x-x1)(x-x2)(x-x3)+c4(x-x1)(x-x2)(x-x3)(X-x4)
#def func(x,y):
 #   x =[2018,2019,2020,2021]
 #   y = [330.9,332.4,333.2,334.6]

```

<!-- # Computation is a bit tricky and for now let's just assume that by 2030, the production increased by 5 percent. What would happen by then --

In [98]:
```
#So in 334.6 is what we have.
#Then 351.33 in 2030.
```

In [106]:
```
CO2Concentration = pd.read_csv('Globalandoverallpredction.csv')


# Keep names all consistent
#assembler = VectorAssembler(inputCols = ['Year','N2O concentrations (annual average) (EEA, 2019)',
                #                         'CH4 concentration (EEA & NOAA (2019))_x',
                 #                      'CO2 concentrations (NOAA, 2018)' ],outputCol='features')

```

In [107]:
```
CO2Concentration.columns

```

Out[107]:
```
Index(['Year', 'N02 concentration ', 'CH4 concentration ',
       'CO2 concentration'],
      dtype='object')
```

In [108]:
```

CO2Concentration.rename(columns = {'N02 concentration ':'N20 concentration yearly',
                                   'CH4 concentration ':'CH4 concentration yearly'
                    ,'CO2 concentration': 'C02 concentration yearly'}, inplace = True)


```

In [109]:

```
CO2Concentration
```

Out[109]:

	Year	N20 concentration yearly	CH4 concentration yearly	C02 concentration yearly
0	2018	330.9	1858	408.52
1	2018	330.9	1859	408.52
2	2018	330.9	1860	408.52
3	2018	330.9	1861	408.52
4	2018	330.9	1862	408.52
...	...	...	...	...
85	2030	343.0	1943	420.00
86	2030	343.0	1944	420.00
87	2030	343.0	1945	420.00
88	2030	343.0	1946	420.00
89	2030	343.0	1947	420.00

90 rows × 4 columns

In [110]:

```
from pyspark.sql import SparkSession
#Create PySpark SparkSession
spark = SparkSession.builder \
    .master("local[1]") \
    .appName("SparkByExamples.com") \
    .getOrCreate()
#Create PySpark DataFrame from Pandas
sparkPredictedFeatures=spark.createDataFrame(CO2Concentration )
sparkPredictedFeatures.printSchema()
sparkPredictedFeatures.show()

```

```
root
 |-- Year: long (nullable = true)
 |-- N20 concentration yearly: double (nullable = true)
 |-- CH4 concentration yearly: long (nullable = true)
 |-- C02 concentration yearly: double (nullable = true)

+----+------------------------+------------------------+------------------------+
|Year|N20 concentration yearly|CH4 concentration yearly|C02 concentration yearly|
+----+------------------------+------------------------+------------------------+
|2018|                   330.9|                    1858|                  408.52|
|2018|                   330.9|                    1859|                  408.52|
|2018|                   330.9|                    1860|                  408.52|
|2018|                   330.9|                    1861|                  408.52|
|2018|                   330.9|                    1862|                  408.52|
|2018|                   330.9|                    1863|                  408.52|
|2018|                   330.9|                    1864|                  408.52|
|2019|                   332.4|                    1865|                   409.0|
|2019|                   332.4|                    1866|                   409.0|
|2019|                   332.4|                    1867|                   409.0|
|2019|                   332.4|                    1868|                   409.0|
|2019|                   332.4|                    1869|                   409.0|
|2019|                   332.4|                    1870|                   409.0|
|2019|                   332.4|                    1871|                   409.0|
|2020|                   333.2|                    1872|                   410.0|
|2020|                   333.2|                    1873|                   410.0|
|2020|                   333.2|                    1874|                   410.0|
|2020|                   333.2|                    1875|                   410.0|
|2020|                   333.2|                    1876|                   410.0|
|2020|                   333.2|                    1877|                   410.0|
```

```
+----+----------------------+----------------------+----------------------+
only showing top 20 rows
```

In [111]:

```
sparkPredictedFeatures.columns
```

Out[111]:

```
['Year',
 'N20 concentration yearly',
 'CH4 concentration yearly',
 'C02 concentration yearly']
```

In [112]:

```
sparkPredictedFeatures.show()
```

```
+----+----------------------+----------------------+----------------------+
|Year|N20 concentration yearly|CH4 concentration yearly|C02 concentration yearly|
+----+----------------------+----------------------+----------------------+
|2018|                 330.9|                  1858|                408.52|
|2018|                 330.9|                  1859|                408.52|
|2018|                 330.9|                  1860|                408.52|
|2018|                 330.9|                  1861|                408.52|
|2018|                 330.9|                  1862|                408.52|
|2018|                 330.9|                  1863|                408.52|
|2018|                 330.9|                  1864|                408.52|
|2019|                 332.4|                  1865|                 409.0|
|2019|                 332.4|                  1866|                 409.0|
|2019|                 332.4|                  1867|                 409.0|
|2019|                 332.4|                  1868|                 409.0|
|2019|                 332.4|                  1869|                 409.0|
|2019|                 332.4|                  1870|                 409.0|
|2019|                 332.4|                  1871|                 409.0|
|2020|                 333.2|                  1872|                 410.0|
|2020|                 333.2|                  1873|                 410.0|
|2020|                 333.2|                  1874|                 410.0|
|2020|                 333.2|                  1875|                 410.0|
|2020|                 333.2|                  1876|                 410.0|
|2020|                 333.2|                  1877|                 410.0|
+----+----------------------+----------------------+----------------------+
only showing top 20 rows
```

In [113]:
```
predictedassembler = VectorAssembler(inputCols = ['Year',
                                    'N20 concentration yearly','CH4 concentration yearly',
                                   'C02 concentration yearly'],outputCol='features')
```

In [114]:
```
predictedoutput = predictedassembler.transform(sparkPredictedFeatures)
```

In [115]:
```
predictedoutput.show()
```

```
+----+------------------------+------------------------+------------------------+--------------------+
|Year|N20 concentration yearly|CH4 concentration yearly|C02 concentration yearly|            features|
+----+------------------------+------------------------+------------------------+--------------------+
|2018|                   330.9|                    1858|                  408.52|[2018.0,330.9,185...|
|2018|                   330.9|                    1859|                  408.52|[2018.0,330.9,185...|
|2018|                   330.9|                    1860|                  408.52|[2018.0,330.9,186...|
|2018|                   330.9|                    1861|                  408.52|[2018.0,330.9,186...|
|2018|                   330.9|                    1862|                  408.52|[2018.0,330.9,186...|
|2018|                   330.9|                    1863|                  408.52|[2018.0,330.9,186...|
|2018|                   330.9|                    1864|                  408.52|[2018.0,330.9,186...|
|2019|                   332.4|                    1865|                   409.0|[2019.0,332.4,186...|
|2019|                   332.4|                    1866|                   409.0|[2019.0,332.4,186...|
|2019|                   332.4|                    1867|                   409.0|[2019.0,332.4,186...|
|2019|                   332.4|                    1868|                   409.0|[2019.0,332.4,186...|
|2019|                   332.4|                    1869|                   409.0|[2019.0,332.4,186...|
|2019|                   332.4|                    1870|                   409.0|[2019.0,332.4,187...|
|2019|                   332.4|                    1871|                   409.0|[2019.0,332.4,187...|
|2020|                   333.2|                    1872|                   410.0|[2020.0,333.2,187...|
|2020|                   333.2|                    1873|                   410.0|[2020.0,333.2,187...|
|2020|                   333.2|                    1874|                   410.0|[2020.0,333.2,187...|
|2020|                   333.2|                    1875|                   410.0|[2020.0,333.2,187...|
|2020|                   333.2|                    1876|                   410.0|[2020.0,333.2,187...|
|2020|                   333.2|                    1877|                   410.0|[2020.0,333.2,187...|
+----+------------------------+------------------------+------------------------+--------------------+
only showing top 20 rows
```

In [116]:
```
predicted_unlabeled_data = predictedoutput.select('features')

```

In [117]:

```
predicted_unlabeled_data.show()
```

```
+--------------------+
|            features|
+--------------------+
|[2018.0,330.9,185...|
|[2018.0,330.9,185...|
|[2018.0,330.9,186...|
|[2018.0,330.9,186...|
|[2018.0,330.9,186...|
|[2018.0,330.9,186...|
|[2018.0,330.9,186...|
|[2019.0,332.4,186...|
|[2019.0,332.4,186...|
|[2019.0,332.4,186...|
|[2019.0,332.4,186...|
|[2019.0,332.4,186...|
|[2019.0,332.4,187...|
|[2019.0,332.4,187...|
|[2020.0,333.2,187...|
|[2020.0,333.2,187...|
|[2020.0,333.2,187...|
|[2020.0,333.2,187...|
|[2020.0,333.2,187...|
|[2020.0,333.2,187...|
+--------------------+
only showing top 20 rows
```

In [111]:
```
predicted_unlabeled_data.show()
```

```
+--------------------+
|            features|
+--------------------+
|[2018.0,330.9,185...|
|[2018.0,330.9,185...|
|[2018.0,330.9,186...|
|[2018.0,330.9,186...|
|[2018.0,330.9,186...|
|[2018.0,330.9,186...|
|[2018.0,330.9,186...|
|[2019.0,332.4,186...|
|[2019.0,332.4,186...|
|[2019.0,332.4,186...|
|[2019.0,332.4,186...|
|[2019.0,332.4,186...|
|[2019.0,332.4,187...|
|[2019.0,332.4,187...|
|[2020.0,333.2,187...|
|[2020.0,333.2,187...|
|[2020.0,333.2,187...|
|[2020.0,333.2,187...|
|[2020.0,333.2,187...|
|[2020.0,333.2,187...|
+--------------------+
only showing top 20 rows
```

In [118]:
```
predicted_test_results = lr_model.transform(predicted_unlabeled_data)



```

In [119]:
```
final_data.show()
```

```
+--------------------+--------------------+
|            features| temperature_anomaly|
+--------------------+--------------------+
|[1880.0,278.2,847...|-0.16194444444444445|
|[1890.0,279.1,867...|             -0.3475|
|[1900.0,279.8,890...|-0.08166666666666667|
|[1905.0,280.3,912...|-0.25472222222222224|
|[1910.0,281.0,935...| -0.4305555555555556|
|[1920.0,282.9,990...|-0.27166666666666667|
|[1925.0,284.0,102...|-0.21611111111111111|
|[1935.0,285.9,107...|-0.19305555555555556|
|[1940.0,286.7,110...| 0.13333333333333333|
|[1945.0,287.8,112...| 0.09555555555555556|
|[1950.0,289.0,116...|-0.17666666666666667|
|[1955.0,290.1,120...|-0.14694444444444443|
|[1960.0,291.4,126...|              -0.025|
|[1965.0,292.9,132...|-0.10583333333333333|
|[1970.0,294.9,140...|0.025833333333333333|
|[1975.0,297.4,148...|-0.01472222222222...|
|[1978.0,298.82,15...| 0.06805555555555556|
|[1979.0,300.04,15...| 0.16666666666666666|
|[1980.0,300.65,15...|  0.2588888888888889|
|[1981.0,301.23,15...| 0.32166666666666666|
+--------------------+--------------------+
only showing top 20 rows
```

In [120]:
```
predicted_test_results.show()
```

```
+--------------------+------------------+
|            features|        prediction|
+--------------------+------------------+
|[2018.0,330.9,185...|0.9415629388324405|
|[2018.0,330.9,185...|0.9422421625653179|
|[2018.0,330.9,186...|0.9429213862981918|
|[2018.0,330.9,186...|0.9436006100310657|
|[2018.0,330.9,186...|0.9442798337639431|
|[2018.0,330.9,186...| 0.944959057496817|
|[2018.0,330.9,186...|0.9456382812296944|
|[2019.0,332.4,186...|0.9137684119390741|
|[2019.0,332.4,186...| 0.914447635671948|
|[2019.0,332.4,186...|0.9151268594048254|
|[2019.0,332.4,186...|0.9158060831376993|
|[2019.0,332.4,186...|0.9164853068705767|
|[2019.0,332.4,187...|0.9171645306034506|
|[2019.0,332.4,187...|0.9178437543363245|
|[2020.0,333.2,187...|0.9148805893020899|
|[2020.0,333.2,187...|0.9155598130349638|
|[2020.0,333.2,187...|0.916239036767841 2|
|[2020.0,333.2,187...|0.9169182605007151|
|[2020.0,333.2,187...| 0.917597484233589|
|[2020.0,333.2,187...|0.9182767079664664|
+--------------------+------------------+
only showing top 20 rows
```

In [121]:
```
print(predicted_test_results.collect()[70])
#Here, we notice that in 2028 when each inputs, CO2, NH4, N20 increase approximately 10 percent,
#the temperature anmoly is 0.88.
```

```
Row(features=DenseVector([2028.0, 341.0, 1928.0, 418.0]), prediction=0.8898167847745153)
```

In [116]:
```
predicted_test_analyze = lr_model.evaluate(test_data)
```

In [117]:
```
predicted_test_analyze.r2
```

Out[117]: 0.8434081054485316

In [118]:
```
predicted_test_results.show()
```

```
+--------------------+------------------+
|            features|        prediction|
+--------------------+------------------+
|[2018.0,330.9,185...|0.8944759544037861|
|[2018.0,330.9,185...|0.8944422203472202|
|[2018.0,330.9,186...|0.8944084862906534|
|[2018.0,330.9,186...|0.8943747522340875|
|[2018.0,330.9,186...|0.8943410181775207|
|[2018.0,330.9,186...|0.8943072841209547|
|[2018.0,330.9,186...|0.8942735500643879|
|[2019.0,332.4,186...| 0.888188908671764|
|[2019.0,332.4,186...|0.8881551746151972|
|[2019.0,332.4,186...|0.8881214405586313|
|[2019.0,332.4,186...|0.8880877065020645|
|[2019.0,332.4,186...|0.8880539724454986|
|[2019.0,332.4,187...|0.8880202383889317|
|[2019.0,332.4,187...|0.8879865043323658|
|[2020.0,333.2,187...|0.8956385404352298|
|[2020.0,333.2,187...| 0.895604806378663|
|[2020.0,333.2,187...|0.8955710723220971|
|[2020.0,333.2,187...|0.8955373382655303|
|[2020.0,333.2,187...|0.8955036042089644|
|[2020.0,333.2,187...|0.8954698701523975|
+--------------------+------------------+
only showing top 20 rows
```

In [119]:
```
predicted_test_results.show()
```

```
+--------------------+------------------+
|            features|        prediction|
+--------------------+------------------+
|[2018.0,330.9,185...|0.8944759544037861|
|[2018.0,330.9,185...|0.8944422203472202|
|[2018.0,330.9,186...|0.8944084862906534|
|[2018.0,330.9,186...|0.8943747522340875|
|[2018.0,330.9,186...|0.8943410181775207|
|[2018.0,330.9,186...|0.8943072841209547|
|[2018.0,330.9,186...|0.8942735500643879|
|[2019.0,332.4,186...| 0.888188908671764|
|[2019.0,332.4,186...|0.8881551746151972|
|[2019.0,332.4,186...|0.8881214405586313|
|[2019.0,332.4,186...|0.8880877065020645|
|[2019.0,332.4,186...|0.8880539724454986|
|[2019.0,332.4,187...|0.8880202383889317|
|[2019.0,332.4,187...|0.8879865043323658|
|[2020.0,333.2,187...|0.8956385404352298|
|[2020.0,333.2,187...| 0.895604806378663|
|[2020.0,333.2,187...|0.8955710723220971|
|[2020.0,333.2,187...|0.8955373382655303|
|[2020.0,333.2,187...|0.8955036042089644|
|[2020.0,333.2,187...|0.8954698701523975|
+--------------------+------------------+
only showing top 20 rows
```

In [120]:

```
predictions.show() #This is the predicted value based on the ML model.
```

```
+--------------------+--------------------+
|            features|          prediction|
+--------------------+--------------------+
|[1880.0,278.2,847...|-0.44482479708784783|
|[1900.0,279.8,890...| -0.3405412661342071|
|[1920.0,282.9,990...|-0.23593215931649691|
|[1935.0,285.9,107...|-0.18206010718973342|
|[1940.0,286.7,110...|-0.12693124099950026|
|[1950.0,289.0,116...|-0.10087879810324463|
|[1970.0,294.9,140...| 0.04312387452194777|
|[1975.0,297.4,148...| 0.09894779365692674|
|[1981.0,301.23,15...| 0.19180346621372735|
|[1984.0,304.02,16...| 0.23041165646101103|
|[1988.0,306.49,16...|  0.3057461978045328|
|[1992.0,310.0,173...| 0.34445592603244535|
|[1993.0,310.25,17...| 0.35243773238228915|
|[1995.0,311.78,17...| 0.39153715972143033|
|[2002.0,317.47,17...|  0.5180427033893338|
|[2006.0,320.37,17...|  0.6145108684392957|
|[2007.0,321.14,17...|  0.6344347528504812|
|[2008.0,322.11,17...|  0.6512729277362812|
|[2015.0,328.51,18...|  0.8081715050815204|
+--------------------+--------------------+
```