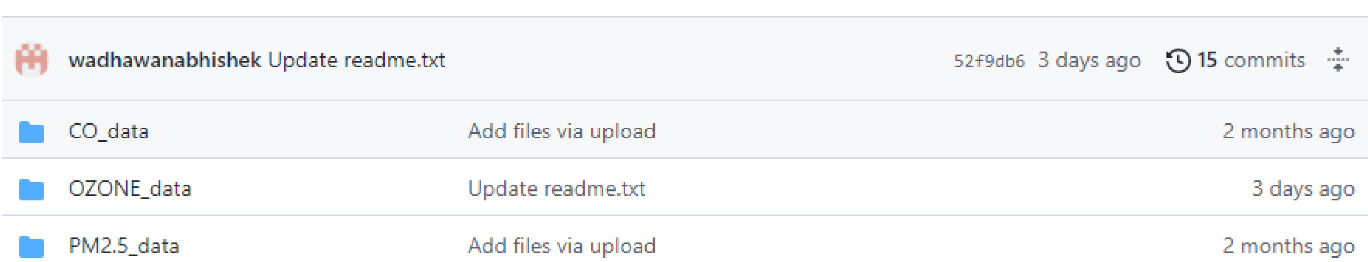
In [47]:
```
import plotly.io as pio
pio.kaleido.scope.default_format = "png"
```

1. Data Collection

- The data consists concentration of pollutants - Carbon Monixide (CO),PM2.5 and Ozone for different California counries for the years 2010 to 2021.
- The raw data contains the columns:
  - **Date** : The date the observation was taken.
  - **Source** : Source of the observation.
  - **Site ID** : Site ID for the particular site in the California county.
  - **POC** : This is the "Parameter Occurrence Code" used to distinguish different instruments that measure the same parameter at the same site.
  - **Daily Mean Pollutant Concentration** : Mean concentration for the pollutant for the date.
  - **UNITS** : Units of measurement for the pollutant concentration.
  - **DAILY_AQI_VALUE** : Air Quality Index value for the date.
  - **Site Name** : Site Name within the County.
  - **DAILY_OBS_COUNT** : The number of values that comprise the daily data set.
  - **PERCENT_COMPLETE** : The percentage of required observations (or scheduled days) made for the given assessment time period.
  - **AQS_PARAMETER_CODE** : The AQS code corresponding to the parameter measured by the monitor
  - **AQS_PARAMETER_DESC** : description assigned in AQS to the parameter measured by the monitor. Parameters may be pollutants or non-pollutants (e.g., wind speed).
  - **CBSA_CODE** : The code of the core based statistical area (metropolitan area) where the monitoring site is located.
  - **CBSA_NAME**: The short version of the OMB-assigned title for the core-based statistical area (CBSA).
  - **STATE_CODE** : Code for the State in USA.
  - **STATE** : State Name
  - **COUNTY_CODE** : County code within the State.
  - **COUNTY** : County name.
  - **SITE_LATITUDE** : Latitude for the site.
  - **SITE_LONGITUDE** :Longiture fir the site.

- All the data pertaining to the pollutants has been stored in a public github repository : https://github.com/wadhawanabhishek/CrowdDoing

- Each pollutant information has been stored in a separate folder within the repository as follows:-



2. Data Transformation

The data stored in the github repository is transformed for further analysis.

- Data is transformed from a daily level of pollutant concentration to a monthly level.
- Along with pollutant concentration, the AQI levels are also calculated
- Only the data corresponding to years fed to function is transformed an further analysed.

In [48]:

```
import pandas as pd
import re
import requests
from bs4 import BeautifulSoup
import plotly.express as px

# Currently we have data from 2010 - 2021

class Pollutant:
    #

    def __init__(self,pollutant:str,start_year:int,end_year:int):
        self.__pollutant = pollutant.upper()
        self.__start_year = start_year
        self.__end_year = end_year
        self.__master_list = []
        self.__pollutant_list = ["CO","PM2.5","OZONE"]
        self.__units = None

        assert self.__pollutant in self.__pollutant_list ,"Not a valid Pollutant"
        assert isinstance(start_year,int),"Start Year is not Integer!"
        assert isinstance(end_year,int),"End Year is not Integer!"


    @property
    def pollutant(self):
        return self.__pollutant

    def __get_data(self):

        #

        try:
            #
            year_regex = re.compile(r'\d\d\d\d')
            github_url = 'https://github.com/wadhawanabhishek/CrowdDoing/tree/main/'+self.
            result = requests.get(github_url)
            soup = BeautifulSoup(result.text, 'html.parser')
            csvfiles = soup.find_all(title=re.compile("\.csv$"))
        except:
            #
            print("Resuouce not Found!")

        filename = [ ]
        for i in csvfiles:
            #
            filename.append(i.extract().get_text())

        years=[]
        for file in filename:
            year = year_regex.search(file)
            years.append(year.group())
        years =[int(i) for i in years]

        # 2011-2016
        if (self.__start_year < min(years)) or (self.__start_year > max(years)):
            #
```

```python
            print("Invalid Year Range. The Start Year Does not Exist")

        if (self.__end_year > max(years)) or (self.__end_year < self.__start_year) or (sel
            #
            print("Invalid Year Range. The End Year Does not Exist")

        new_lst = [i for i in range(self.__start_year,self.__end_year+1)]

        check =  all(item in years for item in new_lst)

        up_file=[]

        if check == True:
            #
            for file in filename:
                for yr in new_lst:
                    if str(yr) in file:
                        up_file.append(file)
        else:
            raise Exception("The data for the given years not present")

        github_url = github_url.replace("github.com",'raw.githubusercontent.com')
        github_url = github_url.replace("tree/",'')

        appended_data =[]

        for f in up_file:
            url = github_url +'/'+ f
            data = pd.read_csv(url)
            appended_data.append(data)

        final_df = pd.concat(appended_data)
        self.__units = final_df['UNITS'].unique()[0]
        return final_df

    def __feature_extraction(self,data_df):

        df = data_df
        df = df.iloc[:, [0,17,2,4,6,15]]
        df = df[df['STATE']=='California']
        return df

    def __get_transformed_data(self,f_df):
        initial_df= f_df
        drop_cols = ['Site ID','STATE']
        for col in drop_cols:
            initial_df = initial_df.drop(col,axis = 1)
        initial_df = initial_df.groupby(['COUNTY','Date']).sum()
        initial_df=initial_df.reset_index(['Date','COUNTY'])
        initial_df['Date']= pd.to_datetime(initial_df['Date'],format='%m/%d/%Y')
        initial_df['Year']= pd.to_datetime(initial_df['Date']).dt.to_period('Y')
        initial_df['Month']= pd.to_datetime(initial_df['Date']).dt.to_period('M')
        initial_df = initial_df.drop("Date",axis = 1)
        cols = [initial_df.columns]
        final_df = initial_df.groupby(['COUNTY','Year','Month']).mean()
        final_df = final_df.reset_index(['COUNTY','Year','Month'])
        pollutant_col = "Monthly Avg "+ self.__pollutant+ " Concentration"
        final_df = final_df.rename(columns={final_df.columns[3]:pollutant_col,"DAILY_AQI_V
        final_df['Pollutant']= self.__pollutant
        final_df['Units'] = self.__units
        return final_df

    def run(self):
        data_df = self.__get_data()
        feature_df = self.__feature_extraction(data_df)
```

```
        trans_df = self.__get_transformed_data(feature_df)
        return trans_df
```

In [49]:
```
df1 = Pollutant("co",2010,2021).run()
df2 = Pollutant("pm2.5",2010,2021).run()
df3 = Pollutant("ozone",2010,2021).run()
```

In [50]:
```
df1.head()
```

Out[50]:

	COUNTY	Year	Month	Monthly Avg CO Concentration	Monthly_Avg_AQI_VALUE	Pollutant	Units
0	Alameda	2010	2010-01	2.645161	30.451613	CO	ppm
1	Alameda	2010	2010-02	2.071429	23.892857	CO	ppm
2	Alameda	2010	2010-03	1.754839	20.322581	CO	ppm
3	Alameda	2010	2010-04	1.316667	15.533333	CO	ppm
4	Alameda	2010	2010-05	0.993548	11.000000	CO	ppm

In [51]:
```
df2.head()
```

Out[51]:

	COUNTY	Year	Month	Monthly Avg PM2.5 Concentration	Monthly_Avg_AQI_VALUE	Pollutant	Units
0	Alameda	2010	2010-01	62.574194	226.096774	PM2.5	ug/m3 LC
1	Alameda	2010	2010-02	41.082143	165.857143	PM2.5	ug/m3 LC
2	Alameda	2010	2010-03	35.087097	145.677419	PM2.5	ug/m3 LC
3	Alameda	2010	2010-04	31.866667	132.100000	PM2.5	ug/m3 LC
4	Alameda	2010	2010-05	29.545161	123.096774	PM2.5	ug/m3 LC

In [52]:
```
df3.head()
```

Out[52]:

	COUNTY	Year	Month	Monthly Avg OZONE Concentration	Monthly_Avg_AQI_VALUE	Pollutant	Units
0	Alameda	2010	2010-01	0.075323	69.709677	OZONE	ppm
1	Alameda	2010	2010-02	0.106214	98.642857	OZONE	ppm
2	Alameda	2010	2010-03	0.140323	129.774194	OZONE	ppm
3	Alameda	2010	2010-04	0.147800	136.766667	OZONE	ppm
4	Alameda	2010	2010-05	0.136323	126.354839	OZONE	ppm

3. Concentration Trends for Different Counties Over the Years

- After the data is transformed to the desired level, the pollutant concentration and AQI values are plotted for different years for different counties.
- This is done to analyse whether the pollutant concentrations follow a particluar pattern in different counties and to understand the trend of pollutant concentration in different counties within California.

- Further, The year in which the pollutant levels were maximum for most of the counties is analysed.
- Here the function takes two arguments -
    1. The transformed data for the pollutant analysed
    2. conc / aqi : Whether the pollutant concentration needs to be plotted or the AQI levels.

In [53]:

```
from plotly.subplots import make_subplots
from math import ceil
import plotly.graph_objects as go

class Plot_Map():

    def __init__(self,df,calc_type:str):
        self.df = df
        self.typ = calc_type.lower()
        self.__pollutant = None
        self.units = None

        assert self.typ in ['conc','aqi'], "Not a valid Calculation Type!"

    def _transform_data(self):

        self.__pollutant = self.df.Pollutant.unique().tolist()[0]
        units = self.df.Units.unique().tolist()[0]
        self.units = units
        trans_df = self.df.drop(['Month','Pollutant','Units'],axis=1)
        trans_df= trans_df.groupby(['COUNTY','Year']).mean()
        trans_df.reset_index(['COUNTY','Year'],inplace=True)
        trans_df['Year'] = trans_df.Year.apply(lambda x : str(x))
        cols = trans_df.columns.tolist()
        trans_df[cols[2]] = trans_df[cols[2]].round(2)
        trans_df[cols[3]] = trans_df[cols[3]].round(2)
        pollutant_col = "Yearly Avg "+ self.__pollutant+ " Concentration"+"("+units+")"
        aqi_col = "Yearly Avg "+ self.__pollutant+" AQI VALUE"
        trans_df = trans_df.rename(columns={trans_df.columns[2]:pollutant_col,trans_df.co

        if self.typ == 'conc':
            trans_df.drop(trans_df.columns[3],axis=1,inplace=True)
        else:
            trans_df.drop(trans_df.columns[2],axis=1,inplace=True)

        return trans_df

    def __getmap(self,trans_df):

        trans_df = trans_df

        counties = trans_df.COUNTY.unique()
        # print(len(counties))
#         cols = df_x.columns
        rows = ceil(len(counties)/4)
        colus = 4
        fig = make_subplots(rows=rows, cols=colus,subplot_titles=counties)
        fig['layout'].update(height=3400, width=1800)
        fig['layout'].update(title = self.__pollutant+" Data Trend")
        r = 1
        c = 1


        for county in counties:
            cdf = trans_df[trans_df['COUNTY']==county]
#             col = cdf.columns
            fig.add_trace(go.Scatter(x=cdf['Year'], y=cdf.iloc[:,2] ,name=county),row=r, c
            fig.update_xaxes(title_text = "Year")
```

```
            fig.update_yaxes(title_text = "Pollutant Concentration "+"("+self.units+")")

            c+=1
            if c > 4:
                r+=1
                c=1

        fig.show()
#           return fig
    def __getmap_analysis(self,trans_df):
        trans_df = trans_df
        cols = trans_df.columns.tolist()
        dic= dict(trans_df.groupby("COUNTY")[cols[2]].max())
        lst =[]
        for county in dic.keys():
            val = trans_df[(trans_df["COUNTY"]==county) & (trans_df[cols[2]]==dic[county])
            data = {"County": county,"Concentration":dic[county],"Year":val}
            lst.append(data)

        d = pd.DataFrame(lst)
        d= d.explode(['Year'])
        d_n = pd.DataFrame(d['Year'].value_counts()).reset_index()
        d_n= d_n.rename(columns = {"index":"Year","Year":"Count"})
        f = px.bar(d_n, x= "Year",y ="Count",text = "Count",text_auto=True)
        f['layout'].update(title = self.__pollutant+" Maximum Pollution Years")
        f.update_yaxes(title_text = "No. of Counties")
        f.show()

    def run(self):
        trans_data = self._transform_data()
        self.__getmap(trans_data)
        self.__getmap_analysis(trans_data)
        return trans_data
```

In [54]:

```
df4 = Plot_Map(df1,'conc').run()
```

In [55]:

```
df5 = Plot_Map(df2,'conc').run()
```

In [56]:

```
df6 = Plot_Map(df3,'conc').run()
```

4. Regression Analysis

- Regression Analysis (OLS method) is done to examine the relationship between Year and the pollutant concentration levels.

- The relation between different years and pollutant concentration is examined for different counties. For each county, the regression line slope value,R squared value, P-value and the regression line equation is calculated. Further, the significance of each relationship is checked by analysing the calculated P-value at a 95% confidence level.

In [57]:

```
from sklearn import linear_model
import plotly.graph_objects as go
import warnings
from scipy import stats
import numpy as np
from plotly.subplots import make_subplots
warnings.filterwarnings("ignore")

class regression_analysis(Plot_Map):

    def __init__(self,df,calc_type):
        self.__pollutant = None

        super().__init__(df,calc_type)

    def __r_trans_df(self):

        df =  super()._transform_data()
        cols = df.columns
        self.__pollutant = cols[2].split()[2]
        df['Year'] = pd.to_numeric(df['Year'])
        return df

    def __reg_analysis(self,initial_df):

        df = initial_df
        cols = df.columns
        counties_list = df.COUNTY.unique().tolist()
        appended_data =[]

        for i,county in enumerate(counties_list):

            df2 = df[df['COUNTY']==county]

            ## Regression Analysis
            X = df2['Year'].values.tolist()
            y = df2[cols[2]].values.tolist()

            X_f = np.array(X, dtype=np.float32)
            y_f = np.array(y, dtype=np.float32)

            slope, intercept, r_value, p_value, std_err = stats.linregress(X_f,y_f)
            line = str(round(slope,4)) + ' * '+'Year '+ '+ ' + str(round(intercept,4))

            sig=lambda p_value: True if p_value <= 0.05 else False


            data = {'County':county,'Slope':slope,'R-Squared Value':r_value**2,'P-Value':p
                    'P-Value less than 0.05?':sig(p_value),
                    'Line-Equation': line}

            data_df = pd.DataFrame(data,index = [i])
            appended_data.append(data_df)

            final_df = pd.concat(appended_data)

        self.__annotations = final_df['Line-Equation'].values.tolist()
```

```python
        return final_df

    def __getmap(self,trans_df):

        map_df = trans_df
        map_df['Year'] = pd.to_numeric(map_df['Year'])
        cols = map_df.columns.tolist()

        fig = px.scatter(map_df, x="Year", y=cols[2], color="COUNTY",trendline='ols',trenc

        fig.update_layout(
            title = self.__pollutant+" Data " + "Regression Analysis",
            updatemenus=[
                {
                    "buttons": [
                        {
                            "label": m,
                            "method": "update",
                            "args": [
                                {
                                    "visible": [
                                        True if m == "All" else t.name == m for t in fig.c
                                    ]

                                }
                            ],
                        }
                        for m in ["All"] + map_df["COUNTY"].unique().tolist()
                    ]
                }
            ]
        )


        fig.show()

    def run(self):
        initial_df = self.__r_trans_df()
        final_df = self.__reg_analysis(initial_df)
#         f_df = final_df[['County','R-Squared_value']]
#         f_df = f_df.groupby(['County']).max()
#         f_df.reset_index(inplace=True)
        final_df['Pollutant']  = self.__pollutant
        self.__getmap(initial_df)
        return final_df
```

4.1 Regression Analysis for Carbon Monoxide Pollutant

In [58]:

```python
df7 = regression_analysis(df1,'conc').run()
df7
```

Out[58]:

	County	Slope	R-Squared Value	P-Value	P-Value less than 0.05?	Line-Equation	Pollutant
0	Alameda	0.181014	0.773692	1.622657e-04	True	0.181 * Year + -362.8829	CO
1	Butte	-0.005175	0.125143	2.592817e-01	False	-0.0052 * Year + 10.8099	CO
2	Contra Costa	0.026888	0.428582	2.088022e-02	True	0.0269 * Year + -52.9505	CO
3	Fresno	-0.074406	0.436578	1.932940e-02	True	-0.0744 * Year + 151.4761	CO
4	Humboldt	-0.021434	0.266949	8.544420e-02	False	-0.0214 * Year + 43.6885	CO
5	Imperial	-0.122622	0.789959	1.108163e-04	True	-0.1226 * Year + 248.0612	CO
6	Inyo	0.007143	0.474084	1.304062e-01	False	0.0071 * Year + -14.2862	CO
7	Kern	-0.024023	0.381131	4.300232e-02	True	-0.024 * Year + 48.9327	CO
8	Los Angeles	-0.321958	0.946043	1.151702e-07	True	-0.322 * Year + 656.4047	CO
9	Madera	-0.012000	0.450000	2.151700e-01	False	-0.012 * Year + 24.452	CO
10	Marin	-0.005804	0.455913	1.597737e-02	True	-0.0058 * Year + 12.1367	CO
11	Monterey	-0.018706	0.786473	1.205383e-04	True	-0.0187 * Year + 38.065	CO
12	Napa	-0.018497	0.467606	1.419980e-02	True	-0.0185 * Year + 37.8172	CO
13	Orange	0.011783	0.012210	7.324470e-01	False	0.0118 * Year + -21.6632	CO
14	Riverside	-0.090175	0.597405	3.201102e-03	True	-0.0902 * Year + 183.9449	CO

	County	Slope	R-Squared Value	P-Value	P-Value less than 0.05?	Line-Equation	Pollutant
15	Sacramento	-0.129720	0.810342	6.583694e-05	True	-0.1297 * Year + 262.9862	CO
16	San Bernardino	0.043811	0.162288	1.941439e-01	False	0.0438 * Year + -85.7373	CO
17	San Diego	-0.107867	0.438314	1.900591e-02	True	-0.1079 * Year + 218.9637	CO
18	San Francisco	0.000140	0.000051	9.824136e-01	False	0.0001 * Year + 0.1564	CO
19	San Joaquin	0.007028	0.156639	2.028339e-01	False	0.007 * Year + -13.7657	CO
20	San Mateo	-0.016678	0.779575	1.418100e-04	True	-0.0167 * Year + 34.1577	CO
21	Santa Barbara	-0.090559	0.521985	7.952537e-03	True	-0.0906 * Year + 183.6775	CO
22	Santa Clara	0.039161	0.445220	1.776415e-02	True	0.0392 * Year + -77.837	CO
23	Santa Cruz	NaN	0.000000	NaN	False	nan * Year + nan	CO
24	Solano	-0.008077	0.241107	1.050098e-01	False	-0.0081 * Year + 16.8199	CO
25	Sonoma	-0.012587	0.473343	1.339056e-02	True	-0.0126 * Year + 25.8133	CO
26	Stanislaus	-0.051993	0.739006	3.371726e-04	True	-0.052 * Year + 105.3911	CO
27	Sutter	0.030000	0.566751	1.419174e-01	False	0.03 * Year + -60.302	CO

4.2 Regression Analysis for PM2.5 Pollutant

In [59]:

```
df8 = regression_analysis(df2,'conc').run()
df8
```

Out[59]:

	County	Slope	R-Squared Value	P-Value	P-Value less than 0.05?	Line-Equation	Pollutant
0	Alameda	2.086616	0.490191	0.016441	True	2.0866 * Year + -4165.0157	PM2.5
1	Alpine	1.158000	0.031944	0.734760	False	1.158 * Year + -2316.3115	PM2.5
2	Butte	0.642417	0.066192	0.445018	False	0.6424 * Year + -1262.7563	PM2.5
3	Calaveras	0.206515	0.109433	0.320385	False	0.2065 * Year + -407.4393	PM2.5
4	Colusa	0.544700	0.179519	0.194082	False	0.5447 * Year + -1081.8609	PM2.5
5	Contra Costa	1.036348	0.540636	0.009924	True	1.0363 * Year + -2072.5521	PM2.5
6	Del Norte	0.552797	0.575770	0.006777	True	0.5528 * Year + -1108.8905	PM2.5
7	El Dorado	0.518571	0.539057	0.010090	True	0.5186 * Year + -1039.8428	PM2.5
8	Fresno	3.834933	0.361976	0.050205	False	3.8349 * Year + -7643.3185	PM2.5
9	Glenn	0.219466	0.136211	0.264001	False	0.2195 * Year + -433.1729	PM2.5
10	Humboldt	-0.437170	0.206439	0.160343	False	-0.4372 * Year + 891.8041	PM2.5
11	Imperial	-0.913825	0.218636	0.146992	False	-0.9138 * Year + 1875.8925	PM2.5
12	Inyo	2.940467	0.525770	0.011572	True	2.9405 * Year + -5907.2062	PM2.5
13	Kern	-0.352737	0.014288	0.726295	False	-0.3527 * Year + 778.1667	PM2.5
14	Kings	1.116308	0.223537	0.141931	False	1.1163 * Year + -2223.3193	PM2.5
15	Lake	0.333077	0.414838	0.032451	True	0.3331 * Year + -666.7082	PM2.5
16	Los Angeles	2.288524	0.121861	0.292697	False	2.2885 * Year + -4458.0526	PM2.5
17	Madera	-1.041656	0.559775	0.008090	True	-1.0417 * Year + 2117.1682	PM2.5

	County	Slope	R-Squared Value	P-Value	P-Value less than 0.05?	Line-Equation	Pollutant
18	Marin	-0.298137	0.349405	0.055474	False	-0.2981 * Year + 611.202	PM2.5
19	Mariposa	0.631662	0.239007	0.127010	False	0.6317 * Year + -1259.7653	PM2.5
20	Mendocino	0.453411	0.207226	0.159447	False	0.4534 * Year + -898.6228	PM2.5
21	Merced	0.421328	0.159051	0.224374	False	0.4213 * Year + -832.0609	PM2.5
22	Mono	3.248558	0.437510	0.026659	True	3.2486 * Year + -6535.9886	PM2.5
23	Monterey	0.788425	0.328144	0.065478	False	0.7884 * Year + -1573.0969	PM2.5
24	Napa	-0.208031	0.117828	0.301373	False	-0.208 * Year + 429.1586	PM2.5
25	Nevada	1.252029	0.367868	0.047887	True	1.252 * Year + -2510.6853	PM2.5
26	Orange	0.293925	0.058200	0.474834	False	0.2939 * Year + -564.2686	PM2.5
27	Placer	3.495447	0.608348	0.004633	True	3.4954 * Year + -7020.5749	PM2.5
28	Plumas	2.796569	0.313464	0.073273	False	2.7966 * Year + -5604.4559	PM2.5
29	Riverside	-2.587156	0.292193	0.086014	False	-2.5872 * Year + 5346.4703	PM2.5
30	Sacramento	2.837951	0.655459	0.002530	True	2.838 * Year + -5673.3534	PM2.5
31	San Benito	0.113732	0.083108	0.389953	False	0.1137 * Year + -222.523	PM2.5
32	San Bernardino	4.123091	0.687333	0.001605	True	4.1231 * Year + -8249.0527	PM2.5
33	San Diego	-0.059639	0.000311	0.958950	False	-0.0596 * Year + 188.5169	PM2.5
34	San Francisco	-0.113725	0.101034	0.340817	False	-0.1137 * Year + 237.9259	PM2.5
35	San Joaquin	0.394599	0.024553	0.645440	False	0.3946 * Year + -757.2746	PM2.5
36	San Luis Obispo	0.080614	0.002660	0.880302	False	0.0806 * Year + -133.742	PM2.5
37	San Mateo	-0.210274	0.215366	0.150464	False	-0.2103 * Year + 432.1707	PM2.5
38	Santa Barbara	1.086489	0.446352	0.024648	True	1.0865 * Year + -2166.991	PM2.5
39	Santa Clara	0.339132	0.057519	0.477511	False	0.3391 * Year + -655.5384	PM2.5
40	Santa Cruz	0.656682	0.432541	0.027848	True	0.6567 * Year + -1312.8166	PM2.5
41	Shasta	-0.017537	0.000550	0.945430	False	-0.0175 * Year + 44.3541	PM2.5

	County	Slope	R-Squared Value	P-Value	P-Value less than 0.05?	Line-Equation	Pollutant
42	Siskiyou	1.341475	0.460917	0.021612	True	1.3415 * Year + -2694.4367	PM2.5
43	Solano	0.728164	0.315640	0.072068	False	0.7282 * Year + -1450.2304	PM2.5
44	Sonoma	-0.106148	0.093961	0.359235	False	-0.1061 * Year + 221.2794	PM2.5
45	Stanislaus	-0.188151	0.027286	0.627407	False	-0.1882 * Year + 406.6262	PM2.5
46	Sutter	0.299599	0.098113	0.348280	False	0.2996 * Year + -584.2546	PM2.5
47	Tehama	0.127190	0.016799	0.704080	False	0.1272 * Year + -248.9353	PM2.5
48	Trinity	1.614166	0.305647	0.077746	False	1.6142 * Year + -3240.7739	PM2.5
49	Tulare	0.641695	0.134237	0.267759	False	0.6417 * Year + -1252.8405	PM2.5
50	Ventura	0.529346	0.085929	0.381661	False	0.5293 * Year + -1022.2254	PM2.5
51	Yolo	0.182076	0.079343	0.401388	False	0.1821 * Year + -356.5348	PM2.5

4.3 Regression Analysis for Ozone Pollutant

In [60]:

```
df9 = regression_analysis(df3,'conc').run()
df9
```

Out[60]:

	County	Slope	R-Squared Value	P-Value	P-Value less than 0.05?	Line-Equation	Pollutant
0	Alameda	0.005420	0.697124	0.000726	True	0.0054 * Year + -10.7857	OZONE
1	Amador	0.000000	0.000000	1.000000	False	0.0 * Year + 0.04	OZONE
2	Butte	-0.000420	0.094406	0.331313	False	-0.0004 * Year + 0.929	OZONE
3	Calaveras	0.000105	0.017165	0.684849	False	0.0001 * Year + -0.1706	OZONE
4	Colusa	-0.000315	0.154482	0.206256	False	-0.0003 * Year + 0.6734	OZONE
5	Contra Costa	0.003706	0.601327	0.003040	True	0.0037 * Year + -7.3434	OZONE
6	El Dorado	-0.000315	0.011560	0.739444	False	-0.0003 * Year + 0.7317	OZONE
7	Fresno	0.004965	0.553692	0.005518	True	0.005 * Year + -9.7087	OZONE
8	Glenn	0.000000	0.000000	1.000000	False	0.0 * Year + 0.04	OZONE
9	Humboldt	-0.000699	0.055208	0.462273	False	-0.0007 * Year + 1.4528	OZONE
10	Imperial	-0.002832	0.328504	0.051402	False	-0.0028 * Year + 5.8591	OZONE
11	Inyo	0.008147	0.855695	0.000016	True	0.0081 * Year + -16.3358	OZONE
12	Kern	0.001049	0.024585	0.626508	False	0.001 * Year + -1.7342	OZONE
13	Kings	0.000734	0.264336	0.087259	False	0.0007 * Year + -1.3941	OZONE
14	Lake	-0.000944	0.566434	0.004733	True	-0.0009 * Year + 1.9402	OZONE
15	Los Angeles	0.002133	0.098192	0.321297	False	0.0021 * Year + -3.7163	OZONE
16	Madera	-0.001189	0.047367	0.496820	False	-0.0012 * Year + 2.4927	OZONE
17	Marin	0.000000	0.000000	1.000000	False	0.0 * Year + 0.03	OZONE
18	Mariposa	-0.004336	0.363258	0.038056	True	-0.0043 * Year + 8.8285	OZONE
19	Mendocino	0.000699	0.419580	0.022752	True	0.0007 * Year + -1.3811	OZONE
20	Merced	-0.000350	0.065559	0.421817	False	-0.0003 * Year + 0.7514	OZONE
21	Monterey	0.000490	0.128497	0.252545	False	0.0005 * Year + -0.8833	OZONE
22	Napa	0.000734	0.264336	0.087259	False	0.0007 * Year + -1.4457	OZONE
23	Nevada	-0.003636	0.756364	0.000237	True	-0.0036 * Year + 7.3941	OZONE
24	Orange	-0.002308	0.368487	0.036336	True	-0.0023 * Year + 4.7978	OZONE
25	Placer	0.001748	0.007458	0.789568	False	0.0017 * Year + -3.3936	OZONE
26	Riverside	0.001364	0.056677	0.456202	False	0.0014 * Year + -2.0892	OZONE
27	Sacramento	-0.008007	0.584260	0.003791	True	-0.008 * Year + 16.3839	OZONE
28	San Benito	0.000000	0.000000	1.000000	False	0.0 * Year + 0.08	OZONE
29	San Bernardino	-0.000140	0.000743	0.933004	False	-0.0001 * Year + 0.8736	OZONE

	County	Slope	R-Squared Value	P-Value	P-Value less than 0.05?	Line-Equation	Pollutant
30	San Diego	-0.012308	0.770873	0.000173	True	-0.0123 * Year + 25.1911	OZONE
31	San Francisco	0.000000	0.000000	1.000000	False	0.0 * Year + 0.03	OZONE
32	San Joaquin	-0.000245	0.029371	0.594334	False	-0.0002 * Year + 0.5691	OZONE
33	San Luis Obispo	-0.001503	0.250259	0.097653	False	-0.0015 * Year + 3.3011	OZONE
34	San Mateo	0.000000	0.000000	1.000000	False	0.0 * Year + 0.03	OZONE
35	Santa Barbara	-0.016573	0.681139	0.000948	True	-0.0166 * Year + 33.8204	OZONE
36	Santa Clara	0.000769	0.026721	0.611716	False	0.0008 * Year + -1.4187	OZONE
37	Santa Cruz	-0.002308	0.387223	0.030712	True	-0.0023 * Year + 4.6928	OZONE
38	Shasta	-0.000420	0.068659	0.410667	False	-0.0004 * Year + 0.994	OZONE
39	Siskiyou	0.000245	0.038073	0.543370	False	0.0002 * Year + -0.4558	OZONE
40	Solano	0.001329	0.540959	0.006408	True	0.0013 * Year + -2.5813	OZONE
41	Sonoma	-0.000699	0.029138	0.595823	False	-0.0007 * Year + 1.4594	OZONE
42	Stanislaus	0.000664	0.280497	0.076575	False	0.0007 * Year + -1.2565	OZONE
43	Sutter	0.000315	0.020474	0.657312	False	0.0003 * Year + -0.5734	OZONE
44	Tehama	0.000455	0.101299	0.313348	False	0.0005 * Year + -0.8453	OZONE
45	Tulare	-0.001678	0.402797	0.026625	True	-0.0017 * Year + 3.5627	OZONE
46	Tuolumne	-0.000315	0.062937	0.431579	False	-0.0003 * Year + 0.6817	OZONE
47	Ventura	-0.002063	0.593723	0.003358	True	-0.0021 * Year + 4.3853	OZONE
48	Yolo	-0.000350	0.058275	0.449729	False	-0.0003 * Year + 0.7797	OZONE

5. Clustering The Counties

- The Counties are now clustered based on their regression slope values
- The counties are then further divided on the bases of their clusters into different pollution trends - Low, Medium, Increasing, Heavily Increasing

In [61]:

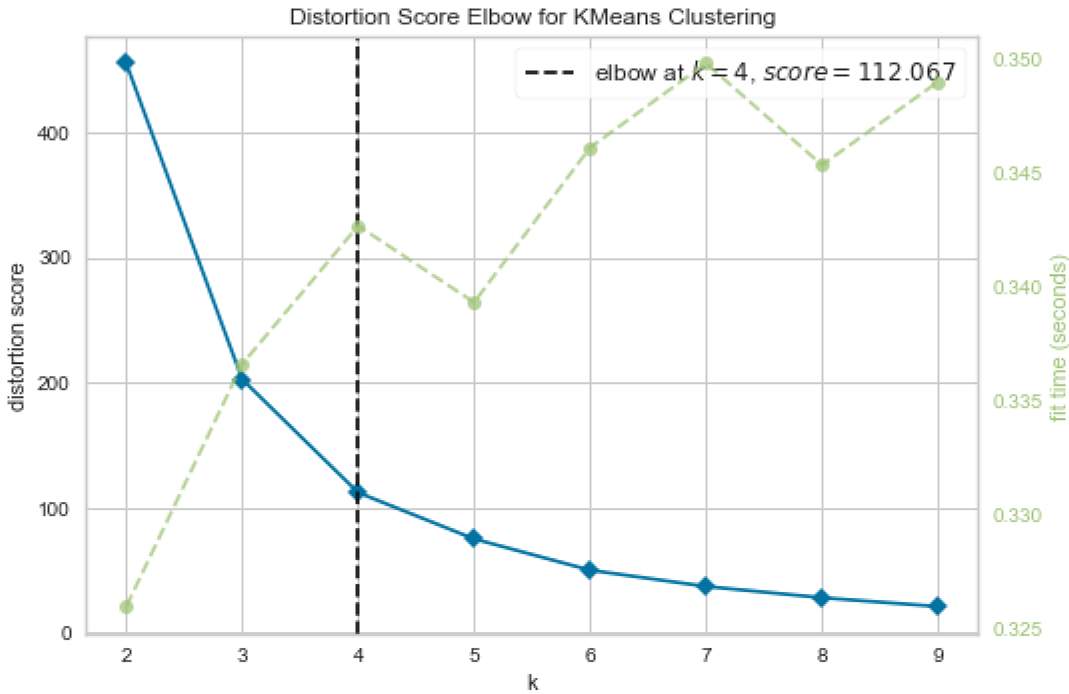
```
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt
from yellowbrick.cluster import KElbowVisualizer
class Clustering_data:

    def __init__(self,regression_df,clusters = 4):
        self.df = regression_df
        self.f_clusters = clusters


    def __cluster_number_analysis(self):
        self.df.sort_values('Slope',ignore_index=True,inplace=True)
        x1 = np.array(self.df.index.values)
        x2 = np.array(self.df['Slope'].values)
        X = np.array(list(zip(x1, x2))).reshape(len(x1), 2)
```

```
        X=np.nan_to_num(X)

#         distortions=[]
#         for i in range(1, 11):
#             km = KMeans(
#                 n_clusters=i, init='k-means++',
#                 n_init=10, max_iter=300,
#                 tol=1e-04, random_state=0
#             )
#             km.fit(X)
#             distortions.append(km.inertia_)

#         # plot
#         plt.plot(range(1, 11), distortions, marker='o')
#         plt.xlabel('Number of clusters')
#         plt.ylabel('Distortion')
#         plt.title("Determining the number of clusters")
#         plt.show()

        model = KMeans()
        visualizer = KElbowVisualizer(
            model, k=(2,10))

        visualizer.fit(X)        # Fit the data to the visualizer
        visualizer.poof()

        return X

    def __assigning_clusters(self,X):
        km = KMeans(
        n_clusters=self.f_clusters, init='k-means++',
        n_init=10, max_iter=300,
        tol=1e-04, random_state=0)
        y_km = km.fit_predict(X)

        self.df['Cluster']=y_km
        self.df['group'] = self.df['Cluster'].ne(self.df['Cluster'].shift()).cumsum()

        mapping = {1:'Low', 2:'Medium', 3:'Increasing',4:'Heavily Increasing'}

        self.df['Pollution Trend']= self.df['group'].apply(lambda x : mapping[x])
        self.df.drop('group',axis=1,inplace=True)

    def __plot_clusters(self):
        fig = px.scatter(self.df,x=self.df.index.values,y='Slope',color='Cluster',hover_na
                         dict(x = "County Index", Slope = "Slope of Regression Line"))
        title = 'Clustering for '+ self.df['Pollutant'].unique()[0]+' data'
        fig.update_layout(title=title)
        fig.show()

    def run(self):
        X = self.__cluster_number_analysis()
        self.__assigning_clusters(X)
        self.__plot_clusters()
        final_df = self.df
        return final_df
```

5.1 Clustering Counties for Carbon Monoxide Pollutant

In [62]:

```
f1_df = Clustering_data(df7).run()
f1_df
```



Out[62]:

	County	Slope	R-Squared Value	P-Value	P-Value less than 0.05?	Line-Equation	Pollutant	Cluster	Pollution Trend
0	Los Angeles	-0.321958	0.946043	1.151702e-07	True	-0.322 * Year + 656.4047	CO	1	Low
1	Sacramento	-0.129720	0.810342	6.583694e-05	True	-0.1297 * Year + 262.9862	CO	1	Low

	County	Slope	R-Squared Value	P-Value	P-Value less than 0.05?	Line-Equation	Pollutant	Cluster	Pollution Trend
2	Imperial	-0.122622	0.789959	1.108163e-04	True	-0.1226 * Year + 248.0612	CO	1	Low
3	San Diego	-0.107867	0.438314	1.900591e-02	True	-0.1079 * Year + 218.9637	CO	1	Low
4	Santa Barbara	-0.090559	0.521985	7.952537e-03	True	-0.0906 * Year + 183.6775	CO	1	Low
5	Riverside	-0.090175	0.597405	3.201102e-03	True	-0.0902 * Year + 183.9449	CO	1	Low
6	Fresno	-0.074406	0.436578	1.932940e-02	True	-0.0744 * Year + 151.4761	CO	1	Low
7	Stanislaus	-0.051993	0.739006	3.371726e-04	True	-0.052 * Year + 105.3911	CO	3	Medium
8	Kern	-0.024023	0.381131	4.300232e-02	True	-0.024 * Year + 48.9327	CO	3	Medium
9	Humboldt	-0.021434	0.266949	8.544420e-02	False	-0.0214 * Year + 43.6885	CO	3	Medium
10	Monterey	-0.018706	0.786473	1.205383e-04	True	-0.0187 * Year + 38.065	CO	3	Medium
11	Napa	-0.018497	0.467606	1.419980e-02	True	-0.0185 * Year + 37.8172	CO	3	Medium
12	San Mateo	-0.016678	0.779575	1.418100e-04	True	-0.0167 * Year + 34.1577	CO	3	Medium
13	Sonoma	-0.012587	0.473343	1.339056e-02	True	-0.0126 * Year + 25.8133	CO	3	Medium
14	Madera	-0.012000	0.450000	2.151700e-01	False	-0.012 * Year + 24.452	CO	0	Increasing
15	Solano	-0.008077	0.241107	1.050098e-01	False	-0.0081 * Year + 16.8199	CO	0	Increasing
16	Marin	-0.005804	0.455913	1.597737e-02	True	-0.0058 * Year + 12.1367	CO	0	Increasing
17	Butte	-0.005175	0.125143	2.592817e-01	False	-0.0052 * Year + 10.8099	CO	0	Increasing
18	San Francisco	0.000140	0.000051	9.824136e-01	False	0.0001 * Year + 0.1564	CO	0	Increasing
19	San Joaquin	0.007028	0.156639	2.028339e-01	False	0.007 * Year + -13.7657	CO	0	Increasing
20	Inyo	0.007143	0.474084	1.304062e-01	False	0.0071 * Year + -14.2862	CO	0	Increasing
21	Orange	0.011783	0.012210	7.324470e-01	False	0.0118 * Year + -21.6632	CO	2	Heavily Increasing
22	Contra Costa	0.026888	0.428582	2.088022e-02	True	0.0269 * Year + -52.9505	CO	2	Heavily Increasing
23	Sutter	0.030000	0.566751	1.419174e-01	False	0.03 * Year + -60.302	CO	2	Heavily Increasing

	County	Slope	R-Squared Value	P-Value	P-Value less than 0.05?	Line-Equation	Pollutant	Cluster	Pollution Trend
24	Santa Clara	0.039161	0.445220	1.776415e-02	True	0.0392 * Year + -77.837	CO	2	Heavily Increasing
25	San Bernardino	0.043811	0.162288	1.941439e-01	False	0.0438 * Year + -85.7373	CO	2	Heavily Increasing
26	Alameda	0.181014	0.773692	1.622657e-04	True	0.181 * Year + -362.8829	CO	2	Heavily Increasing
27	Santa Cruz	NaN	0.000000	NaN	False	nan * Year + nan	CO	2	Heavily Increasing

5.1.1 Checking for Outliers

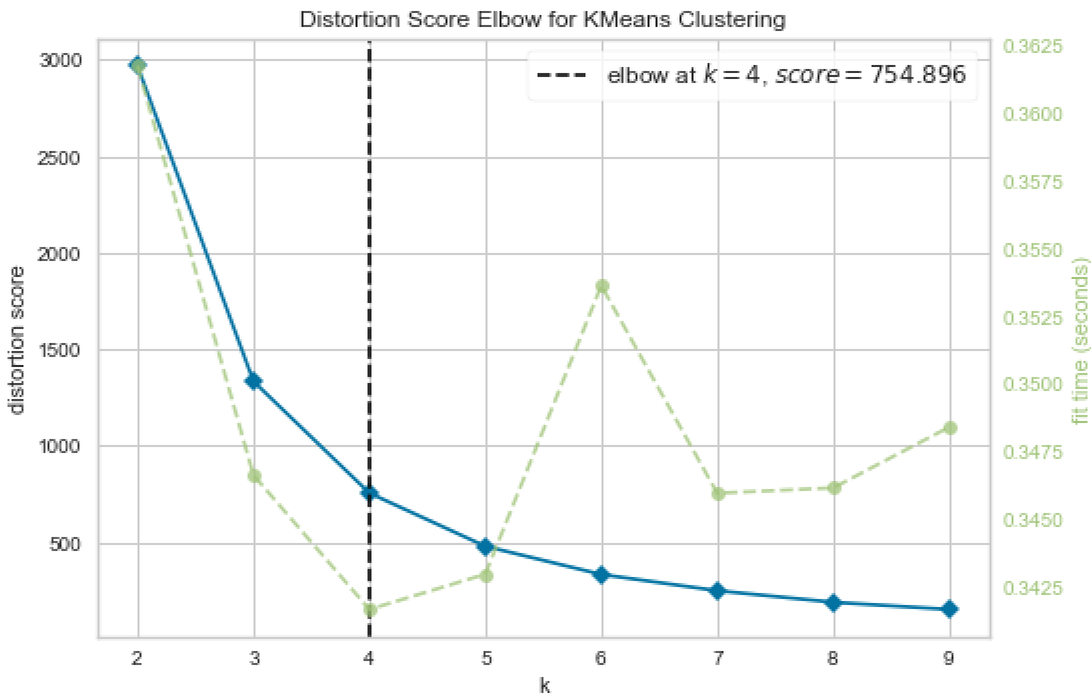
In [63]:

```
fig = px.box(f1_df, y="Slope",hover_name="County",title="CO Outliers")

fig.add_annotation(x=0.05, y=0.18, #Q1
            text="Alameda",
            font=dict(size=12),
            showarrow=False,
            )
fig.add_annotation(x=0.06, y=-0.32, #Q1
            text="Los Angeles",
            font=dict(size=12),
            showarrow=False,
            )

fig.show()
```

- For Carbon Monoxide, it is quite evident that the counties Alamaeda and Los Angeles are the outliers. Alamaeda had an increasing trend for pollutant concentration over the years while the county of Los Angeles had a decreasing trend for pollutant concentration over the years.

5.2 Clustering Counties for PM2.5 Pollutant

In [64]:
```
f2_df=Clustering_data(df8).run()
f2_df
```



Out[64]:

	County	Slope	R-Squared Value	P-Value	P-Value less than 0.05?	Line-Equation	Pollutant	Cluster	Pollution Trend
0	Riverside	-2.587156	0.292193	0.086014	False	-2.5872 * Year + 5346.4703	PM2.5	0	Low
1	Madera	-1.041656	0.559775	0.008090	True	-1.0417 * Year + 2117.1682	PM2.5	0	Low
2	Imperial	-0.913825	0.218636	0.146992	False	-0.9138 * Year + 1875.8925	PM2.5	0	Low
3	Humboldt	-0.437170	0.206439	0.160343	False	-0.4372 * Year + 891.8041	PM2.5	0	Low
4	Kern	-0.352737	0.014288	0.726295	False	-0.3527 * Year + 778.1667	PM2.5	0	Low
5	Marin	-0.298137	0.349405	0.055474	False	-0.2981 * Year + 611.202	PM2.5	0	Low
6	San Mateo	-0.210274	0.215366	0.150464	False	-0.2103 * Year + 432.1707	PM2.5	0	Low
7	Napa	-0.208031	0.117828	0.301373	False	-0.208 * Year + 429.1586	PM2.5	0	Low
8	Stanislaus	-0.188151	0.027286	0.627407	False	-0.1882 * Year + 406.6262	PM2.5	0	Low
9	San Francisco	-0.113725	0.101034	0.340817	False	-0.1137 * Year + 237.9259	PM2.5	0	Low
10	Sonoma	-0.106148	0.093961	0.359235	False	-0.1061 * Year + 221.2794	PM2.5	0	Low
11	San Diego	-0.059639	0.000311	0.958950	False	-0.0596 * Year + 188.5169	PM2.5	0	Low
12	Shasta	-0.017537	0.000550	0.945430	False	-0.0175 * Year + 44.3541	PM2.5	0	Low
13	San Luis Obispo	0.080614	0.002660	0.880302	False	0.0806 * Year + -133.742	PM2.5	2	Medium
14	San Benito	0.113732	0.083108	0.389953	False	0.1137 * Year + -222.523	PM2.5	2	Medium
15	Tehama	0.127190	0.016799	0.704080	False	0.1272 * Year + -248.9353	PM2.5	2	Medium
16	Yolo	0.182076	0.079343	0.401388	False	0.1821 * Year + -356.5348	PM2.5	2	Medium
17	Calaveras	0.206515	0.109433	0.320385	False	0.2065 * Year + -407.4393	PM2.5	2	Medium
18	Glenn	0.219466	0.136211	0.264001	False	0.2195 * Year + -433.1729	PM2.5	2	Medium
19	Orange	0.293925	0.058200	0.474834	False	0.2939 * Year + -564.2686	PM2.5	2	Medium
20	Sutter	0.299599	0.098113	0.348280	False	0.2996 * Year + -584.2546	PM2.5	2	Medium

	County	Slope	R-Squared Value	P-Value	P-Value less than 0.05?	Line-Equation	Pollutant	Cluster	Pollution Trend
21	Lake	0.333077	0.414838	0.032451	True	0.3331 * Year + -666.7082	PM2.5	2	Medium
22	Santa Clara	0.339132	0.057519	0.477511	False	0.3391 * Year + -655.5384	PM2.5	2	Medium
23	San Joaquin	0.394599	0.024553	0.645440	False	0.3946 * Year + -757.2746	PM2.5	2	Medium
24	Merced	0.421328	0.159051	0.224374	False	0.4213 * Year + -832.0609	PM2.5	2	Medium
25	Mendocino	0.453411	0.207226	0.159447	False	0.4534 * Year + -898.6228	PM2.5	2	Medium
26	El Dorado	0.518571	0.539057	0.010090	True	0.5186 * Year + -1039.8428	PM2.5	2	Medium
27	Ventura	0.529346	0.085929	0.381661	False	0.5293 * Year + -1022.2254	PM2.5	1	Increasing
28	Colusa	0.544700	0.179519	0.194082	False	0.5447 * Year + -1081.8609	PM2.5	1	Increasing
29	Del Norte	0.552797	0.575770	0.006777	True	0.5528 * Year + -1108.8905	PM2.5	1	Increasing
30	Mariposa	0.631662	0.239007	0.127010	False	0.6317 * Year + -1259.7653	PM2.5	1	Increasing
31	Tulare	0.641695	0.134237	0.267759	False	0.6417 * Year + -1252.8405	PM2.5	1	Increasing
32	Butte	0.642417	0.066192	0.445018	False	0.6424 * Year + -1262.7563	PM2.5	1	Increasing
33	Santa Cruz	0.656682	0.432541	0.027848	True	0.6567 * Year + -1312.8166	PM2.5	1	Increasing
34	Solano	0.728164	0.315640	0.072068	False	0.7282 * Year + -1450.2304	PM2.5	1	Increasing
35	Monterey	0.788425	0.328144	0.065478	False	0.7884 * Year + -1573.0969	PM2.5	1	Increasing
36	Contra Costa	1.036348	0.540636	0.009924	True	1.0363 * Year + -2072.5521	PM2.5	1	Increasing
37	Santa Barbara	1.086489	0.446352	0.024648	True	1.0865 * Year + -2166.991	PM2.5	1	Increasing
38	Kings	1.116308	0.223537	0.141931	False	1.1163 * Year + -2223.3193	PM2.5	1	Increasing
39	Alpine	1.158000	0.031944	0.734760	False	1.158 * Year + -2316.3115	PM2.5	1	Increasing
40	Nevada	1.252029	0.367868	0.047887	True	1.252 * Year + -2510.6853	PM2.5	3	Heavily Increasing
41	Siskiyou	1.341475	0.460917	0.021612	True	1.3415 * Year + -2694.4367	PM2.5	3	Heavily Increasing
42	Trinity	1.614166	0.305647	0.077746	False	1.6142 * Year + -3240.7739	PM2.5	3	Heavily Increasing

	County	Slope	R-Squared Value	P-Value	P-Value less than 0.05?	Line-Equation	Pollutant	Cluster	Pollution Trend
43	Alameda	2.086616	0.490191	0.016441	True	2.0866 * Year + -4165.0157	PM2.5	3	Heavily Increasing
44	Los Angeles	2.288524	0.121861	0.292697	False	2.2885 * Year + -4458.0526	PM2.5	3	Heavily Increasing
45	Plumas	2.796569	0.313464	0.073273	False	2.7966 * Year + -5604.4559	PM2.5	3	Heavily Increasing
46	Sacramento	2.837951	0.655459	0.002530	True	2.838 * Year + -5673.3534	PM2.5	3	Heavily Increasing
47	Inyo	2.940467	0.525770	0.011572	True	2.9405 * Year + -5907.2062	PM2.5	3	Heavily Increasing
48	Mono	3.248558	0.437510	0.026659	True	3.2486 * Year + -6535.9886	PM2.5	3	Heavily Increasing
49	Placer	3.495447	0.608348	0.004633	True	3.4954 * Year + -7020.5749	PM2.5	3	Heavily Increasing
50	Fresno	3.834933	0.361976	0.050205	False	3.8349 * Year + -7643.3185	PM2.5	3	Heavily Increasing
51	San Bernardino	4.123091	0.687333	0.001605	True	4.1231 * Year + -8249.0527	PM2.5	3	Heavily Increasing

5.2.1 Checking for Outliers

In [65]:

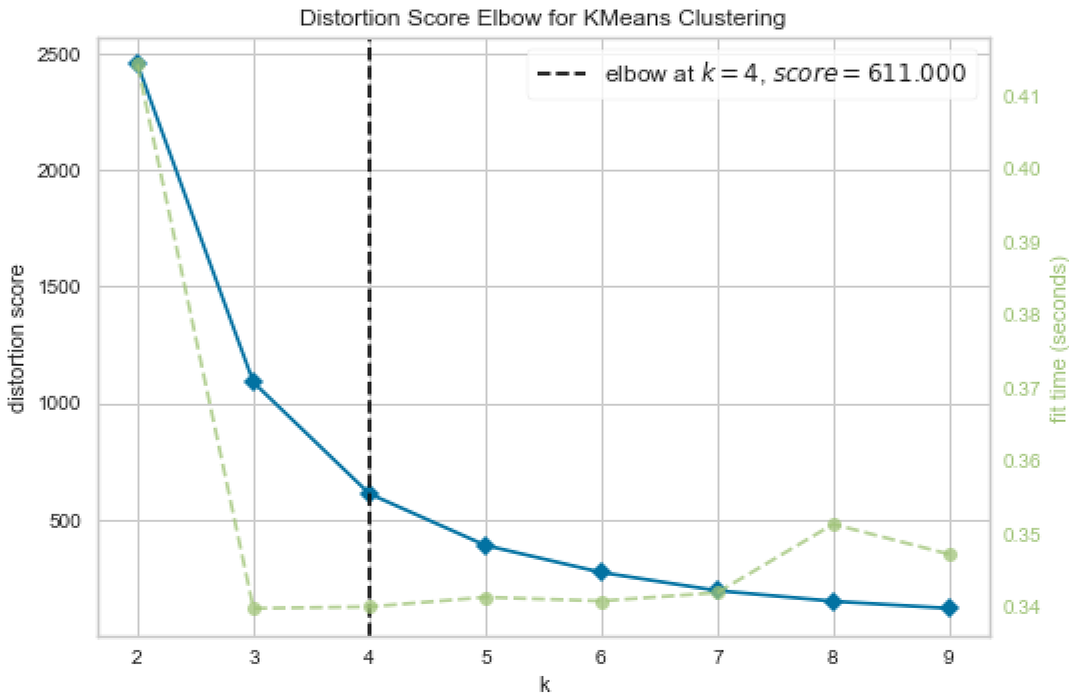
```
q1=f2_df["Slope"].quantile(0.25)

q3=f2_df["Slope"].quantile(0.75)

IQR=q3-q1

outliers = f2_df[((f2_df["Slope"]<(q1-1.5*IQR)) | (f2_df["Slope"]>(q3+1.5*IQR)))]
fig = px.box(f2_df, y="Slope",hover_name="County",title="PM2.5 Outliers")
outliers_annotations = outliers[["County","Slope"]].sort_values(by ="Slope",ascending=Fals
outliers_lst = outliers_annotations.values.tolist()

for county, slope in outliers_lst:
    fig.add_annotation(x=0.05, y=slope, #Q1
            text=county,
            font=dict(size=10),
            showarrow=False,
            )
fig['layout'].update(height = 800,width =800)
fig.show()
```

- From the outlier analysis of PM2.5 pollutant concentrations, it is evident that the counties - Riverside and San Bernardino are the extreme outliers

5.3 Clustering Counties for Ozone Pollutant

In [66]:

```
f3_df = Clustering_data(df9).run()
f3_df
```



Out[66]:

	County	Slope	R-Squared Value	P-Value	P-Value less than 0.05?	Line-Equation	Pollutant	Cluster	Pollution Trend
0	Santa Barbara	-0.016573	0.681139	0.000948	True	-0.0166 * Year + 33.8204	OZONE	1	Low
1	San Diego	-0.012308	0.770873	0.000173	True	-0.0123 * Year + 25.1911	OZONE	1	Low

	County	Slope	R-Squared Value	P-Value	P-Value less than 0.05?	Line-Equation	Pollutant	Cluster	Pollution Trend
2	Sacramento	-0.008007	0.584260	0.003791	True	-0.008 * Year + 16.3839	OZONE	1	Low
3	Mariposa	-0.004336	0.363258	0.038056	True	-0.0043 * Year + 8.8285	OZONE	1	Low
4	Nevada	-0.003636	0.756364	0.000237	True	-0.0036 * Year + 7.3941	OZONE	1	Low
5	Imperial	-0.002832	0.328504	0.051402	False	-0.0028 * Year + 5.8591	OZONE	1	Low
6	Orange	-0.002308	0.368487	0.036336	True	-0.0023 * Year + 4.7978	OZONE	1	Low
7	Santa Cruz	-0.002308	0.387223	0.030712	True	-0.0023 * Year + 4.6928	OZONE	1	Low
8	Ventura	-0.002063	0.593723	0.003358	True	-0.0021 * Year + 4.3853	OZONE	1	Low
9	Tulare	-0.001678	0.402797	0.026625	True	-0.0017 * Year + 3.5627	OZONE	1	Low
10	San Luis Obispo	-0.001503	0.250259	0.097653	False	-0.0015 * Year + 3.3011	OZONE	1	Low
11	Madera	-0.001189	0.047367	0.496820	False	-0.0012 * Year + 2.4927	OZONE	1	Low
12	Lake	-0.000944	0.566434	0.004733	True	-0.0009 * Year + 1.9402	OZONE	1	Low
13	Humboldt	-0.000699	0.055208	0.462273	False	-0.0007 * Year + 1.4528	OZONE	3	Medium
14	Sonoma	-0.000699	0.029138	0.595823	False	-0.0007 * Year + 1.4594	OZONE	3	Medium
15	Butte	-0.000420	0.094406	0.331313	False	-0.0004 * Year + 0.929	OZONE	3	Medium
16	Shasta	-0.000420	0.068659	0.410667	False	-0.0004 * Year + 0.994	OZONE	3	Medium
17	Merced	-0.000350	0.065559	0.421817	False	-0.0003 * Year + 0.7514	OZONE	3	Medium
18	Yolo	-0.000350	0.058275	0.449729	False	-0.0003 * Year + 0.7797	OZONE	3	Medium
19	Tuolumne	-0.000315	0.062937	0.431579	False	-0.0003 * Year + 0.6817	OZONE	3	Medium
20	Colusa	-0.000315	0.154482	0.206256	False	-0.0003 * Year + 0.6734	OZONE	3	Medium
21	El Dorado	-0.000315	0.011560	0.739444	False	-0.0003 * Year + 0.7317	OZONE	3	Medium
22	San Joaquin	-0.000245	0.029371	0.594334	False	-0.0002 * Year + 0.5691	OZONE	3	Medium
23	San Bernardino	-0.000140	0.000743	0.933004	False	-0.0001 * Year + 0.8736	OZONE	3	Medium
24	San Mateo	0.000000	0.000000	1.000000	False	0.0 * Year + 0.03	OZONE	3	Medium

	County	Slope	R-Squared Value	P-Value	P-Value less than 0.05?	Line-Equation	Pollutant	Cluster	Pollution Trend
25	Marin	0.000000	0.000000	1.000000	False	0.0 * Year + 0.03	OZONE	2	Increasing
26	San Francisco	0.000000	0.000000	1.000000	False	0.0 * Year + 0.03	OZONE	2	Increasing
27	Glenn	0.000000	0.000000	1.000000	False	0.0 * Year + 0.04	OZONE	2	Increasing
28	Amador	0.000000	0.000000	1.000000	False	0.0 * Year + 0.04	OZONE	2	Increasing
29	San Benito	0.000000	0.000000	1.000000	False	0.0 * Year + 0.08	OZONE	2	Increasing
30	Calaveras	0.000105	0.017165	0.684849	False	0.0001 * Year + -0.1706	OZONE	2	Increasing
31	Siskiyou	0.000245	0.038073	0.543370	False	0.0002 * Year + -0.4558	OZONE	2	Increasing
32	Sutter	0.000315	0.020474	0.657312	False	0.0003 * Year + -0.5734	OZONE	2	Increasing
33	Tehama	0.000455	0.101299	0.313348	False	0.0005 * Year + -0.8453	OZONE	2	Increasing
34	Monterey	0.000490	0.128497	0.252545	False	0.0005 * Year + -0.8833	OZONE	2	Increasing
35	Stanislaus	0.000664	0.280497	0.076575	False	0.0007 * Year + -1.2565	OZONE	2	Increasing
36	Mendocino	0.000699	0.419580	0.022752	True	0.0007 * Year + -1.3811	OZONE	2	Increasing
37	Napa	0.000734	0.264336	0.087259	False	0.0007 * Year + -1.4457	OZONE	0	Heavily Increasing
38	Kings	0.000734	0.264336	0.087259	False	0.0007 * Year + -1.3941	OZONE	0	Heavily Increasing
39	Santa Clara	0.000769	0.026721	0.611716	False	0.0008 * Year + -1.4187	OZONE	0	Heavily Increasing
40	Kern	0.001049	0.024585	0.626508	False	0.001 * Year + -1.7342	OZONE	0	Heavily Increasing
41	Solano	0.001329	0.540959	0.006408	True	0.0013 * Year + -2.5813	OZONE	0	Heavily Increasing
42	Riverside	0.001364	0.056677	0.456202	False	0.0014 * Year + -2.0892	OZONE	0	Heavily Increasing
43	Placer	0.001748	0.007458	0.789568	False	0.0017 * Year + -3.3936	OZONE	0	Heavily Increasing
44	Los Angeles	0.002133	0.098192	0.321297	False	0.0021 * Year + -3.7163	OZONE	0	Heavily Increasing
45	Contra Costa	0.003706	0.601327	0.003040	True	0.0037 * Year + -7.3434	OZONE	0	Heavily Increasing
46	Fresno	0.004965	0.553692	0.005518	True	0.005 * Year + -9.7087	OZONE	0	Heavily Increasing
47	Alameda	0.005420	0.697124	0.000726	True	0.0054 * Year + -10.7857	OZONE	0	Heavily Increasing
48	Inyo	0.008147	0.855695	0.000016	True	0.0081 * Year + -16.3358	OZONE	0	Heavily Increasing

5.3.1 Checking for Outliers

In [67]:

```
q1=f3_df["Slope"].quantile(0.25)

q3=f3_df["Slope"].quantile(0.75)

IQR=q3-q1

outliers = f3_df[((f3_df["Slope"]<(q1-1.5*IQR)) | (f3_df["Slope"]>(q3+1.5*IQR)))]
fig = px.box(f3_df, y="Slope",hover_name="County",title = "Ozone Outliers")
outliers_annotations = outliers[["County","Slope"]].sort_values(by ="Slope",ascending=Fals
outliers_lst = outliers_annotations.values.tolist()

for county, slope in outliers_lst:
    fig.add_annotation(x=0.06, y=slope, #Q1
            text=county,
            font=dict(size=10),
            showarrow=False,
            )
fig['layout'].update(height = 800,width =800)
fig.show()
```

- From the outlier analysis of Ozone pollutant concentrations, it is evident that the counties - Santa Barbara and Inyo are the extreme outliers

5.4 Analysing Pollution Trends

- Analysing which Counties had the Pollution Trend as **Heavily Increasing** for all three pollutants:

In [68]:

```
##Getting counties with Heavily Increasing pollution trend for all 3 pollutants

co_counties_high = set(f1_df[f1_df['Pollution Trend']=='Heavily Increasing']['County'].val
pm_counties_high = set(f2_df[f2_df['Pollution Trend']=='Heavily Increasing']['County'].val
ozone_counties_high = set(f3_df[f3_df['Pollution Trend']=='Heavily Increasing']['County'].

common_counties_high = co_counties_high.intersection(pm_counties_high,ozone_counties_high)
common_counties_high
```

Out[68]:

```
{'Alameda'}
```

- Analysing which Counties had the Pollution Trend as **Low** for all three pollutants:

In [69]:

```
##Getting counties with low pollution trend for all 3 pollutants

co_counties_low = set(f1_df[f1_df['Pollution Trend']=='Low']['County'].values.tolist())
pm_counties_low = set(f2_df[f2_df['Pollution Trend']=='Low']['County'].values.tolist())
ozone_counties_low = set(f3_df[f3_df['Pollution Trend']=='Low']['County'].values.tolist())

common_counties_low = co_counties_low.intersection(pm_counties_low,ozone_counties_low)
common_counties_low
```

Out[69]:

```
{'Imperial', 'San Diego'}
```

In [ ]:

```
```