IE 6200 PROBABILITY AND STATISTICS
PROJECT REPORT
GROUP 14

ABILASH SIVAKUMAR,
MURALEEDHARAN RAJAGOPALAN,
SATHYA PRAKASH VETTUVAPALAYAM RAGHURAMAN

Contents

SECTION 1: OBJECTIVE

The objective of this project is to study the Pollution trends in each state across the US. In order to compare, the following has been considered: Vehicle Registrations, Population, Land Area and Per Capita Income data. Using these datasets collected from the US-governed websites, and a few other reputed dataset providers, the statistical analysis and inferential analysis were performed using R software to visualize and provide a series of postulates that has been highlighted in this report.

SECTION 2: SCOPE

Data of 5 different categories of US states for the year 2018, 2019, 2020 are compared.

Data-Description:
The main dataset used is the Pollution data from United State Environmental Protection Agency, and Vehicle Registration data from US Department of Transportation's Highway statistics & Census Bureau statistics. In addition, Population, Land Area and Per Capita income of each state are taken.

The mentioned datasets are combined into a single dataset and used for analyzing and inferring.

Pollution is taken as a main objective because it is the common problem for all not limited to human beings but to entire planet.

To monitor and control pollution, the US government is spending 1000 billion Dollars/annum in average. And the country's ultimate agenda is to reduce carbon emissions.

Information's on the factors considered.

1. Pollution data - Considering only Transport pollution, since it is the largest contributor among others. In order to get the transportation pollution, 27% of total pollution is taken. Among transportation pollution only light duty vehicles and medium and heavy-duty trucks are considered for its major contributors.

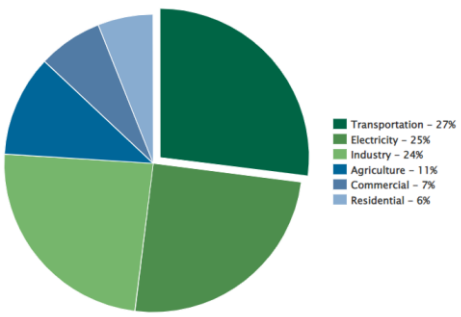


Fig.2.1 – U.S. Emissions by Sector

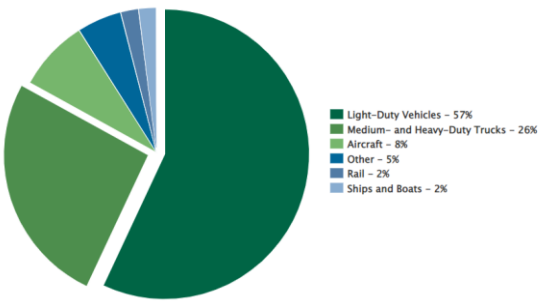


Fig.2.2 – U.S. Transportation Sector Emissions by Source

2. Transportation data – Both Private and Public Vehicle registrations are considered.

3. Population – From the total population, only people above the age 18 are filtered, since those are ones who are officially eligible to buy/register/drive a vehicle.

4. Land Area – Total land area in Square Kilometres is taken into consideration for this analysis.

5. Per Capita Income – The Per Capita Income of each state is collected from the US Economic data.

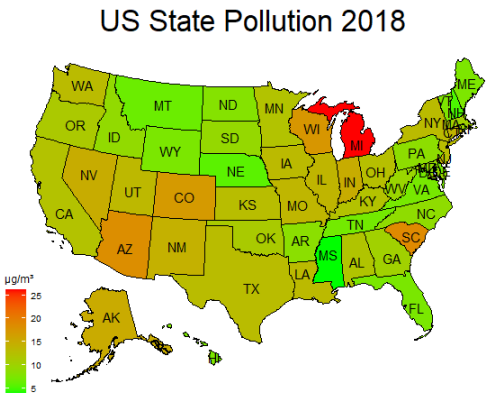
SUMMARY OF DATA:

```
      area          tot_pollution18  tot_pollution19  tot_pollution20
 Min.   :    177   Min.   : 3.987   Min.   : 6.038   Min.   : 6.353
 1st Qu.:  92980   1st Qu.: 8.277   1st Qu.: 8.396   1st Qu.: 7.946
 Median : 145746   Median :11.956   Median :11.451   Median :10.287
 Mean   : 192814   Mean   :11.417   Mean   :11.169   Mean   :10.996
 3rd Qu.: 218163   3rd Qu.:13.388   3rd Qu.:13.075   3rd Qu.:13.005
 Max.   :1723337   Max.   :26.215   Max.   :20.310   Max.   :24.794

     pop_18             pop_19             pop_20
 Min.   :  445330   Min.   :  446223   Min.   :  444752
 1st Qu.: 1379891   1st Qu.: 1390580   1st Qu.: 1420048
 Median : 3457394   Median : 3565203   Median : 3645221
 Mean   : 4977815   Mean   : 4990163   Mean   : 5063395
 3rd Qu.: 5730066   3rd Qu.: 5779897   3rd Qu.: 5736033
 Max.   :30622192   Max.   :30740509   Max.   :32223652

    trans18            trans19            trans20
 Min.   :  351933   Min.   :  350463   Min.   :  356537
 1st Qu.: 1834778   1st Qu.: 1863114   1st Qu.: 1850414
 Median : 3942875   Median : 3919157   Median : 4095442
 Mean   : 5364621   Mean   : 5421396   Mean   : 5410063
 3rd Qu.: 6123063   3rd Qu.: 6128237   3rd Qu.: 6126841
 Max.   :31022328   Max.   :31247270   Max.   :30398249
```

Fig.2.3

SECTION 3: CONTRIBUTION OF POLLUTION IN EACH STATE

Pollution of each state in the US in the years 2018, 2019 and 2020:



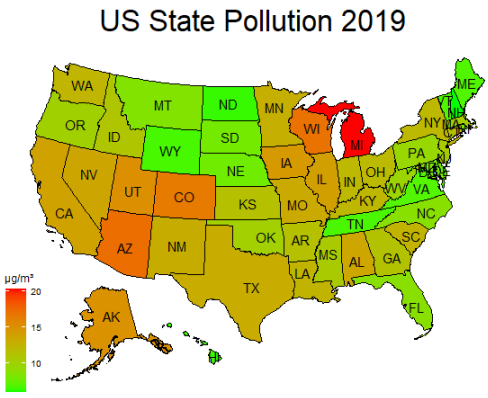


Fig.3.1

Fig.3.2

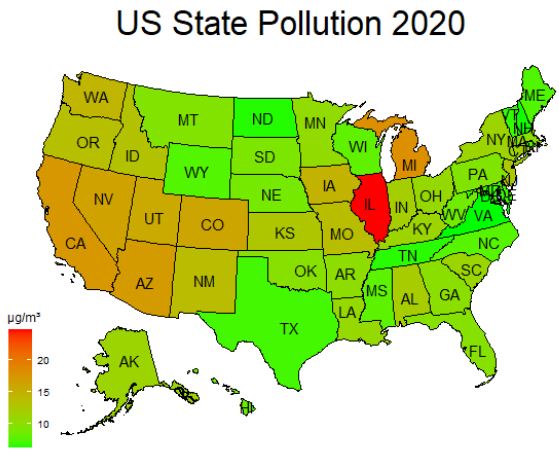


Fig.3.3

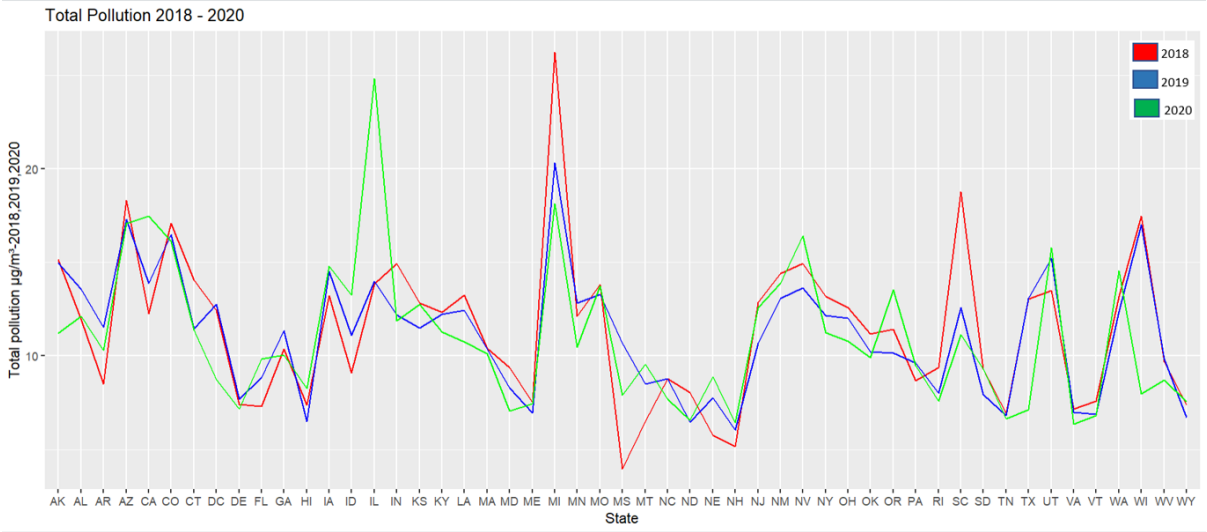


Fig.3.4

The pollutants are converted in μg/m³. It ranges from 3 to 26 μg/m³ in the year 2018, 6 to 20 μg/m³ in 2019 and 6 to 25 μg/m³ in 2020. The pollution data is a sum of all the pollutants involved and taken into consideration. Where the State MI(Michigan), emits the highest pollution of above 25 and 20 μg/m³ during the year 2018 and 2019 respectively. However in the year 2020, MI shows a decrease in pollution below 20 μg/m³ and in turn the State IL(Illinois) shows an increase in pollution in the year 2020 topping with the emmision units above 20 μg/m³. The attached multi line bar plot*(Fig.3.4)* will help in visualizing the pollution trends over the selected years. Along with it the heat maps*(Fig.3.1-3.3)* represents the average range of pollution in each state.

SECTION 4: CONTRIBUTION OF POLLUTANTS IN EACH YEAR

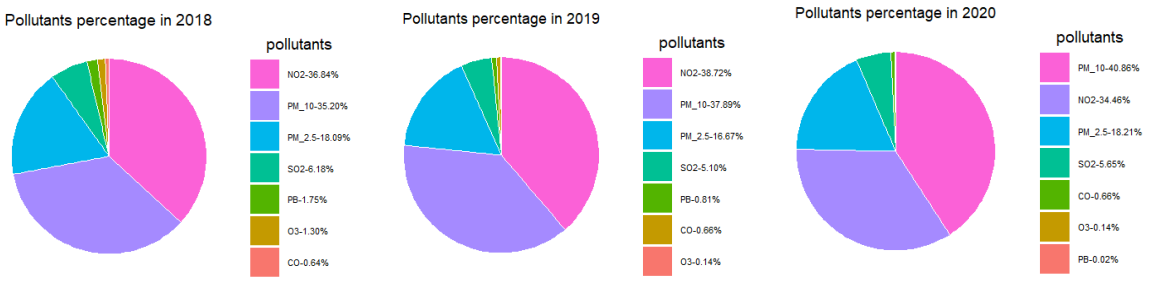


Fig.4 Percentage of Pollutants for the year 2018,19,20.

On considering the pollutants, percentage of NO2 tops in the year 2018 and 2019 with 36.84% and 38.72% followed by PM10. In the year 2020, PM10 contributes the highest pollution with 40.86% followed by NO2.

	Pollutants	Average in 2018	Average in 2019	Average in 2020	Max in 2018	Max in 2019	Max in 2020	Min in 2018	Min in 2019	Min in 2020
1	O3	0.14807	0.01612	0.01572	6.75	0.02	0.02	0.01	0.01	0.01
2	CO	0.07363	0.07378	0.07284	0.15	0.15	0.18	0.01	0.01	0.02
3	SO2	0.70600	0.07378	0.62107	5.13	3.69	6.31	0.08	0.01	0.05
4	NO2	4.20618	4.32442	3.78902	10.08	10.12	11.60	0.33	0.30	0.36
5	PB	0.19965	0.09084	0.00225	5.28	2.31	0.02	0.01	0.01	0.01
6	PM_10	4.01834	4.23211	4.49242	9.13	7.07	8.69	0.92	1.17	1.95
7	PM_2.5	2.06479	1.86218	2.00222	5.25	2.49	3.37	1.15	0.74	1.04

Table-1

On considering the average of pollutants in each year, PM10 shows an upward trend in all three years and NO2 shows a downward trend in the respective years.

SECTION 5: COMPARISION OF POLLUTION DIFFERENCE BETWEEN THE SUCCESSIVE YEARS

Fig.5.1-Pollution difference between 2018 and 2019:

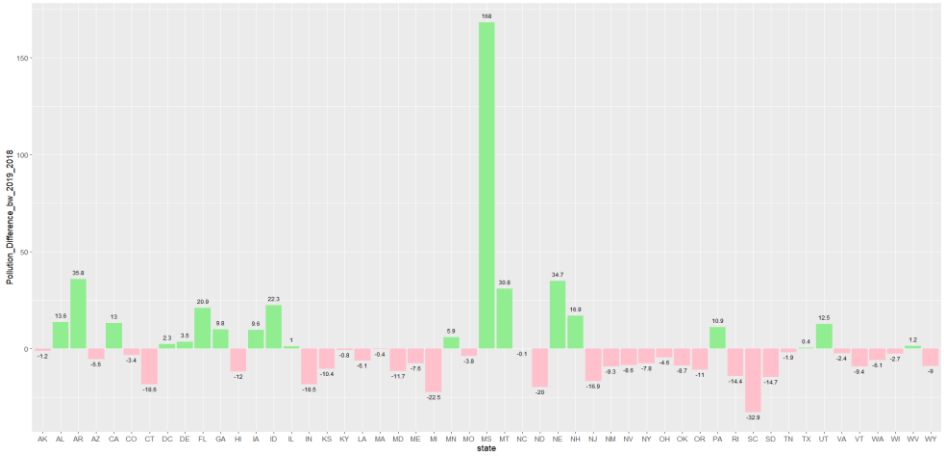
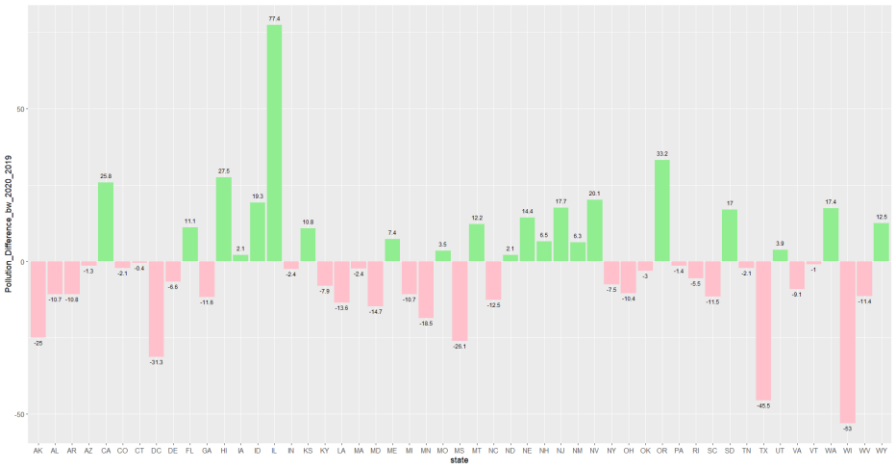


Fig.5.2-Pollution difference between 2019 and 2020:



The above two are the comparison of average pollution difference between all states for the years 2018&19 and 2019&20. On studying the comparison differences, there is a recordable change in the pollution emissions between the subjected years. In this comparison, IL has the highest difference in 2020 comparative with its preceding year. Pollution difference's mean of 2018-19 is 2.15%. Pollution difference's mean of 2019-20 is a downtrend of -0.42%.

SECTION 6: COMPARISION OF VEHICLE REGISTRATION BETWEEN THE SUCCESSIVE YEARS

Fig.6.1-Vehicle registration difference between 2018 and 2019:

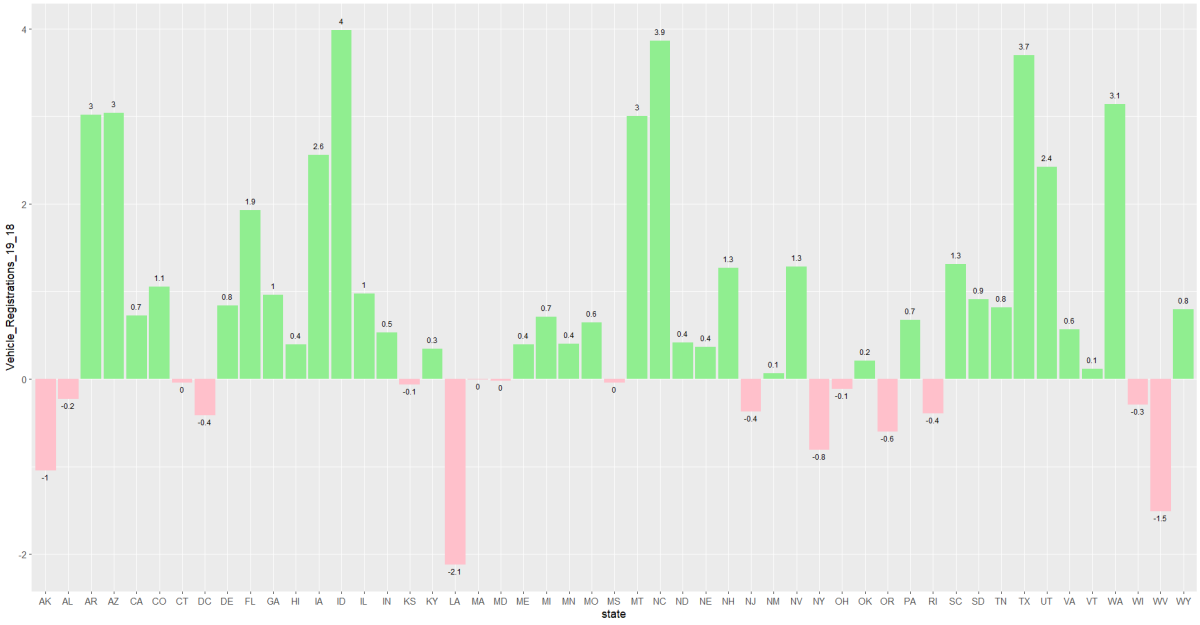
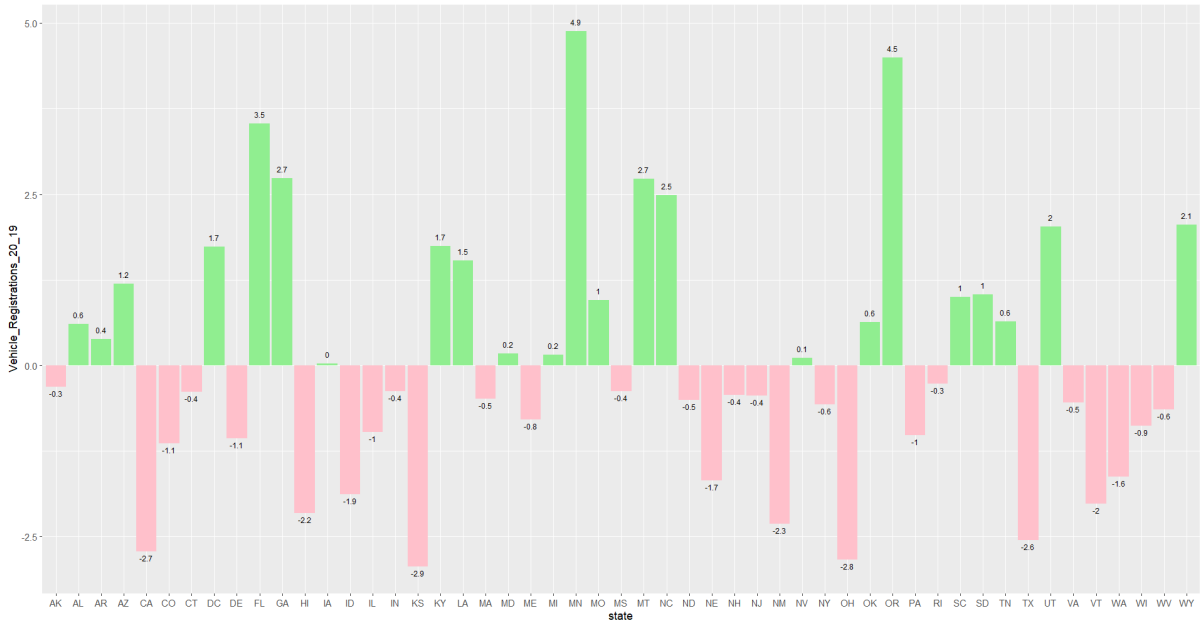


Fig.6.2-Vehicle registration difference between 2019 and 2020:



There is a significant increase in vehicle registrations in the States MN(Minnesota), ID(Idaho), NC (North Carolina), TX(Texas) and OR(Oregon) in the respective years. Vehicle registration difference's mean of 2018-19 is 1% and in the years 2019 and 2020 is 0.05%.

SECTION 7: CORRELATION BETWEEN POPULATION AND POLLUTION

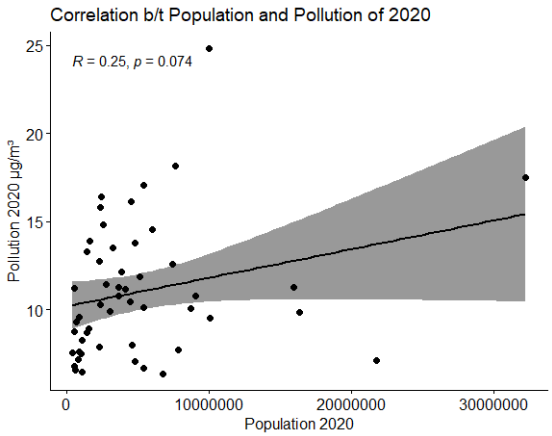


Fig.7.1

absolute values of r	Interpretation
0.90 - 1.00	Very high correlation
0.70 - 0.90	High correlation
0.50 - 0.70	Moderate correlation
0.30 - 0.50	Low correlation
0 - 0.30	Negligible or weak correlation

Fig.7.2

On running a correlation test between the population and pollution for the years 2018-20. The following are the obtained results.
Year 2018 - r=0.17, p=0.25
Year 2019 - r=0.25, p=0.078
Year 2020 - r=0.25, p=0.074
The result obtained is negligible or weak correlation between population and pollution. So, the common ideology that if the population increases, the pollution also increases does not satisfy here.

SECTION 8: CORRELATION BETWEEN POLLUTION AND AREA

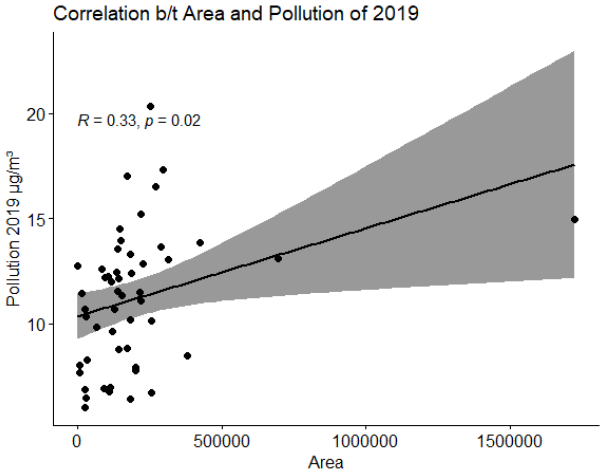


Fig.8.1

absolute values of r	Interpretation
0.90 - 1.00	Very high correlation
0.70 - 0.90	High correlation
0.50 - 0.70	Moderate correlation
0.30 - 0.50	Low correlation
0 - 0.30	Negligible or weak correlation

Fig.8.2

On running a correlation test between the pollution and area for the years 2018-20.
The following are the obtained results:
Year 2018 - r=0.23, p=0.1
Year 2019 - r=0.33, p=0.02
Year 2020 - r=0.15, p=0.28
The result obtained is Low correlation between pollution and area.

SECTION 9: CORRELATION BETWEEN POLLUTION AND VEHICLE REGISTRATION

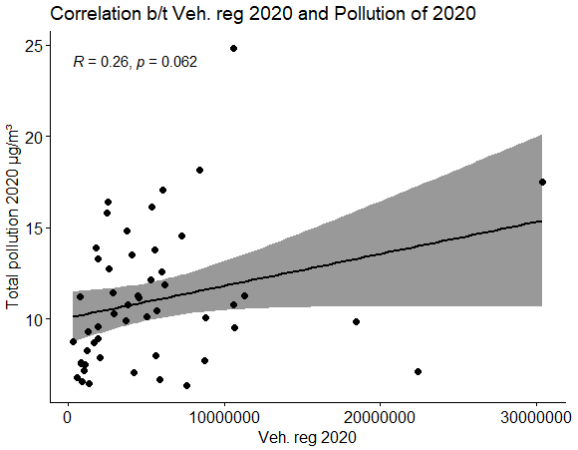


Fig.9.1

absolute values of r	Interpretation
0.90 - 1.00	Very high correlation
0.70 - 0.90	High correlation
0.50 - 0.70	Moderate correlation
0.30 - 0.50	Low correlation
0 - 0.30	Negligible or weak correlation

Fig.9.2

On running a correlation test between the pollution and vehicle registration for the years 2018-20.
The following are the obtained results:
Year 2018 - r=0.18, p=0.2
Year 2019 - r=0.26, p=0.061
Year 2020 - r=0.26, p=0.062
The result obtained is negligible or weak correlation between pollution and area.

SECTION 10: PROBABILITY DISTRIBUTION OF POLLUTION

The distribution of the graph must be determined before doing any probabilistic computations. Before obtaining the probability distribution type from Culley and Frey graph for pollution data 2018, bar graph is plotted to get an overview and then proceeding with the Culley and Frey graph.

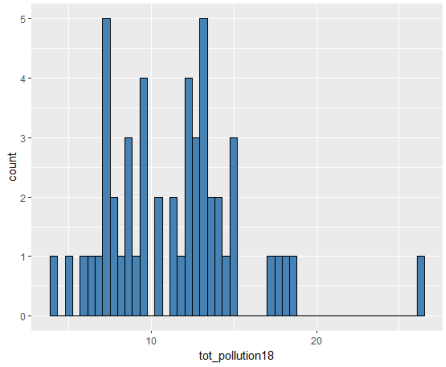


Fig.10.1

```
summary statistics
------
min:  3.986778   max:  26.21508
median:  11.95616
mean:  11.41666
estimated sd:  4.063467
estimated skewness:  0.9248911
estimated kurtosis:  5.194934
```

Fig.10.2

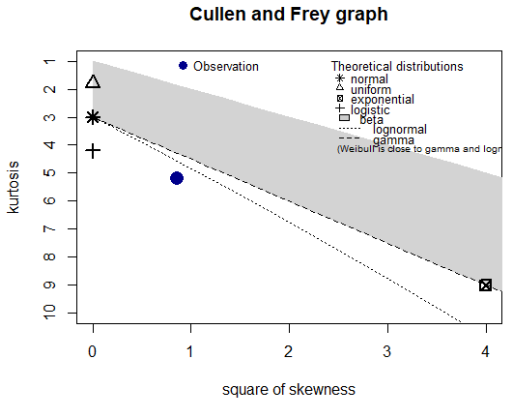


Fig.10.3

According to the above graph, this pattern is similar to the gamma, normal, and lognormal distributions.

With the goodness of fitness test, lognormal shows the best fit among others. Since it has the lowest AIC and BIC value.

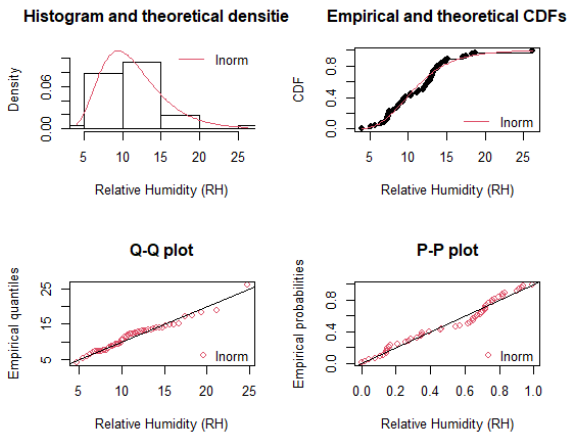


Fig.10.4

```
Fitting of the distribution ' lnorm ' by maximum likelihood
Parameters :
         estimate Std. Error
meanlog 2.3729923 0.05019728
sdlog   0.3584803 0.03549360
Loglikelihood:  -141.0685   AIC:  286.137   BIC:  290.0007
Correlation matrix:
        meanlog sdlog
meanlog       1     0
sdlog         0     1
```

Fig.10.5

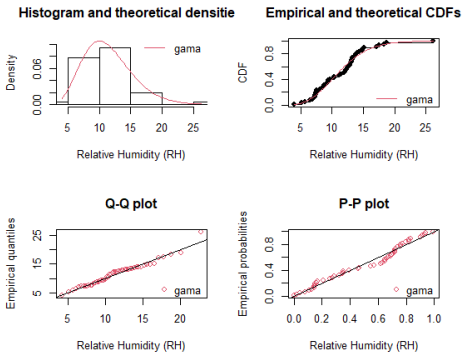


Fig.10.6

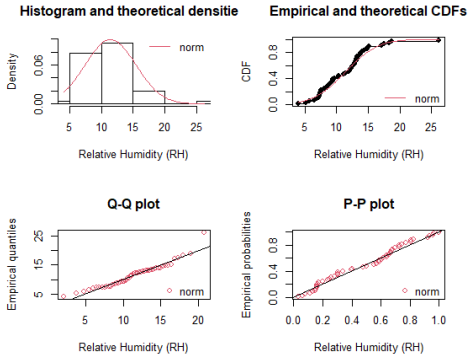


Fig.10.7

SECTION 11: HYPOTHESIS TESTING

The Null hypothesis has the pollution mean of 0.86, considering a confidence interval of 95% we selected to perform a two tailed test, since that region has a pollution difference between 2018-19 & 2019-20.

On determining the p-value using hypothesis testing the obtained p-value is 0.25 which is to the right of the average value which falls in the rejection region; hence we reject the null hypothesis and accept the alternate hypothesis.

Derived P value falls in the RR, so we reject H0 and the

H0: pollution mean = 0.86

H1: pollution mean ≠ 0.86

Two tailed test: Z-calc value= -0.699; P value=0.25

SECTION 12: CONFIDENCE INTERVAL

On running the confidence interval test on NO2 for 30 samples, the upper bound and lower bound values came as 4.9 and 3.61. To conclude, the Mean of sample NO2 lies between the bounded range with 95% confidence interval.

```
> lower.bound <- sample.mean - margin.error
> upper.bound <- sample.mean + margin.error
> print(c(lower.bound,upper.bound))
[1] 3.618030 4.908655
```

Fig-12.1

SECTION13: CONCLUSION

In order to reduce the carbon foot print, and to observe the current trend in automotive emissions, the pollution data for the years 2018,19 and 20 has been taken into consideration with specific focus to light duty vehicles and medium and heavy-duty trucks across the states of US. The following tests have been conducted to observe the trends and correlation among the data.

The general data distribution for the dataset was identified depending whether the data was continuous/discrete and the goodness of fit test as well. After which the correlation was identified and the hypothesis was postulated.

REFERENCES:

1. https://www.kaggle.com/datasets/alpacanonymous/us-pollution-20002021
2. https://www.epa.gov/outdoor-air-quality-data/download-daily-data
3. https://www.epa.gov/greenvehicles/fast-facts-transportation-greenhouse-gas-emissions
4. https://www.fhwa.dot.gov/policyinformation/statistics/2010/mv1.cfm
5. https://fred.stl ouisfed.org/release/tables?rid=110&eid=257197&od=2020-01-01#
6. https://www.educba.com/graphs-in-r/
7. https://bolt.mph.ufl.edu/6050-6052/unit-1/case-q-q/linear-relationships/
8. https://www.youtube.com/@statswithr602/videos
9. https://www.breeze-technologies.de/blog/air-pollution-how-to-convert-between-mgm3-%C2%B5gm3-ppm-ppb/
10. https://donortracker.org/united-states/climate

PROJECT CONTRIBUTION:

ABILASH SIVAKUMAR	MURALEEDHARAN RAJAGOPALAN	SATHYA PRAKSAH VETUVAPALAYAM RAGHURAMAN
34%	33%	33%