Tecnológico de Monterrey

Materia:
Análisis de Métodos Multivariados para Ciencia de Datos

Reto - Etapa 2

Profesores:
Blanca Rosa Ruiz Hernández
Monica Guadalupe Elizondo Amaya

Alumnos:
Gian Marco Innocenti A00834310
Andrés Villarreal González A00833915
Andrea Hernandez A00835225

5 de octubre de 2023, Monterrey, N.L.

Para nuestra solución propondremos un estudio donde se verá si la calidad del aire mejoró en el año 2020 donde se experimentó la pandemia de COVID 19 donde la movilidad de las personas se redujo drásticamente en comparación a un año donde la actividad se normalizo tanto en industria como en movilidad humana en el año 2021. Para esto se tomarán registros de 3 diferentes estaciones de recolección de datos en 2020 y 2021. Las estaciones escogidas fueron la Estación Centro, la Estación Noreste2, y la Estación Sureste3. También durante este análisis solo se tomarán en cuenta los contaminantes que se incluyen en el Índice de Aire y Salud y las variables meteorológicas proporcionadas por la estación de recolección.

Indica el número de variables y el tamaño del dataset

Para esta problemática utilizaremos un dataset con 18 variables y 52,462 registros. Por lo que el tamaño del dataset es de 52,462 x 18.

Describe las variables que consideramos importantes y su relevancia para cumplir con el objetivo del proyecto.

Para nuestra solución utilizaremos las siguiente variables:

- **date: Variable Categórica**
- **Año: Variable Categórica**
- **Estacion: Variable Categórica**
  - Esta variable representa la estación donde fue recolectada la información.
- **CO: Variable Categórica**
  - Esta variable representa el nivel del contaminante CO en el índice de calidad del aire y salud medido por el promedio móvil ponderado medido en la estación durante una hora. Es importante porque queremos descubrir el impacto de este contaminante en la calidad del aire.
- **NO2: Variable Categórica**
  - Esta variable representa el nivel del contaminante NO2 en el índice de calidad del aire y salud medido por el promedio móvil ponderado medido en la estación durante una hora. Es importante porque queremos descubrir el impacto de este contaminante en la calidad del aire.
- **O3: Variable Categórica**
  - Esta variable representa el nivel del contaminante O3 en el índice de calidad del aire y salud medido por el promedio móvil ponderado medido en la estación durante una hora. Es importante porque queremos descubrir el impacto de este contaminante en la calidad del aire.
- **PM10: Variable Categórica**
  - Esta variable representa el nivel del contaminante PM10 en el índice de calidad del aire y salud medido por el promedio móvil ponderado medido en la estación durante una hora. Es importante porque queremos descubrir el impacto de este contaminante en la calidad del aire.
- **PM2.5: Variable Categórica**
  - Esta variable representa el nivel del contaminante PM2.5 en el índice de calidad del aire y salud medido por el promedio móvil ponderado medido en la estación durante una hora. Es importante porque queremos descubrir el impacto de este contaminante en la calidad del aire.
- **SO2: Variable Categórica**
  - Esta variable representa el nivel del contaminante SO2 en el índice de calidad del aire y salud medido por el promedio móvil ponderado medido en la estación durante una hora. Es importante porque queremos descubrir el impacto de este contaminante en la calidad del aire.

- **CO3Concentración: Variable Numérica**
  - Esta variable representa el nivel del contaminante NO2 en el índice de calidad del aire y salud medido por la concentración medida en la estación durante una hora. Es importante porque queremos descubrir el impacto de este contaminante en la calidad del aire.
- **PRS: Variable Numérica**
  - Esta variable representa la presión atmosférica en la estación durante una hora. Es importante para el modelo ya que queremos descubrir que tanto este factor meteorológico impactó con la calidad del aire.
- **RAINF: Variable Numérica**
  - Esta variable representa la cantidad de precipitación en la estación durante una hora. Es importante para el modelo ya que queremos descubrir que tanto este factor meteorológico impactó con la calidad del aire.
- **RH: Variable Numérica**
  - Esta variable representa el porcentaje de humedad relativa en la estación durante una hora.Es importante para el modelo ya que queremos descubrir que tanto este factor meteorológico impactó con la calidad del aire.
- **SR: Variable Numérica**
  - .Esta variable representa la radiación solar en la estación durante una hora.Es importante para el modelo ya que queremos descubrir que tanto este factor meteorológico impactó con la calidad del aire.
- **TOUT: Variable Numérica**
  - Esta variable representa la temperatura promedio en la estación durante una hora.Es importante para el modelo ya que queremos descubrir que tanto este factor meteorológico impactó con la calidad del aire.
- **WSR: Variable Numérica**
  - Esta variable representa la velocidad del viento en la estación durante una hora.Es importante para el modelo ya que queremos descubrir que tanto este factor meteorológico impactó con la calidad del aire.
- **WDR: Variable Numérica**
  - Esta variable representa la dirección del viento en la estación durante una hora. Es importante para el modelo ya que queremos descubrir que tanto este factor meteorológico impactó con la calidad del aire.
- **Objetivo: Variable Numérica**
  - El valor del Índice de Calidad de Aire más alto medido cada hora en la estación. Es importante porque es la variable que trataremos de predecir.

Los diferentes valores que las siguientes variables pueden tomar para los diferentes años y estaciones se pueden ver en la siguiente Tabla1.

	Dataset	date	CO	NO2	O3	PM10	PM2.5	PRS	RAINF	RH	SO2	SR	TOUT	WSR	WDR
0	Centro_2020	[2020-01-01 00:00:00, 2020-01-01 01:00:00, 202...	[0.34-6.36]	[0.0-35.5]	[1.0-148.0]	[2.0-720.0]	[2.25-112.84]	[689.4-724.8]	[0.0-0.0]	[1.0-92.0]	[0.7-33.5]	[0.0-0.282]	[3.15-37.93]	[0.7-23.4]	[1.0-360.0]
1	Noreste2_2020	[2020-01-01 00:00:00, 2020-01-01 01:00:00, 202...	[0.15-16.93]	[0.0-42.7]	[1.0-125.0]	[2.0-763.0]	[2.09-140.45]	[692.1-715.4]	[0.0-0.48]	[1.0-90.0]	[0.5-192.1]	[0.0-0.878]	[1.89-38.03]	[0.6-29.2]	[1.0-359.0]
2	Sureste3_2020	[2020-01-01 00:00:00, 2020-01-01 01:00:00, 202...	[0.05-17.71]	[0.3-75.3]	[2.0-101.0]	[3.0-711.0]	[2.0-158.0]	[706.9-747.6]	[0.0-4.39]	[1.0-94.0]	[0.5-90.8]	[0.0-0.889]	[2.38-40.2]	[0.6-11.6]	[1.0-360.0]
3	Centro_2021	[2021-01-01 00:00:00, 2021-01-01 01:00:00, 202...	[0.05-14.6]	[0.1-90.4]	[1.0-135.0]	[2.0-634.0]	[2.17-144.66]	[700.7-722.7]	[0.0-0.0]	[1.0-92.0]	[0.7-61.9]	[0.0-1.007]	[-4.75-40.26]	[0.9-21.9]	[1.0-360.0]
4	Noreste2_2021	[2021-01-01 00:00:00, 2021-01-01 01:00:00, 202...	[0.05-22.38]	[0.0-80.1]	[2.0-139.0]	[2.0-718.0]	[2.05-160.57]	[680.0-713.3]	[0.0-22.38]	[1.0-90.0]	[0.5-58.5]	[0.0-0.774]	[-6.3-41.13]	[0.1-62.1]	[1.0-360.0]
5	Sureste3_2021	[2021-01-01 00:00:00, 2021-01-01 01:00:00, 202...	[0.11-4.15]	[0.0-77.9]	[1.0-129.0]	[2.0-613.0]	[2.0-131.0]	[680.0-745.7]	[0.0-2.1]	[3.0-94.0]	[0.5-175.5]	[0.0-0.803]	[-3.22-42.49]	[0.1-42.0]	[1.0-360.0]

Tabla 1. Posibles Valores Para las Variables

Resume el trabajo realizado en la preparación de datos: necesidad de unión de bases, limpieza de los datos, imputación (de haber sido necesaria), porcentaje de datos faltantes, datos atípicos, duplicados, etc

Durante este proceso primero identificamos los datos faltantes para cada estación y su respectiva columna. El desglose de los datos faltantes se puede ver en la siguiente Tabla2.

	Dataset	date	CO	NO2	O3	PM10	PM2.5	PRS	RAINF	RH	SO2	SR	TOUT	WSR	WDR	Total_Estacion	Porcentaje
0	Centro_2020	0	124	967	1450	380	1845	132	123	126	447	54	135	131	133	6047	0.021553
1	Noreste2_2020	0	108	3496	1784	253	1438	116	108	147	1325	33	109	781	1726	11424	0.040719
2	Sureste3_2020	0	918	136	2505	510	1456	128	126	137	1109	15	126	131	141	7438	0.026511
3	Centro_2021	0	654	1294	1064	288	834	208	198	200	1483	131	200	240	206	7000	0.02495
4	Noreste2_2021	0	315	592	517	400	481	899	313	598	1869	74	657	764	1480	8959	0.031933
5	Sureste3_2021	0	481	770	517	693	1311	505	484	727	1271	40	483	491	493	8266	0.029463
6	Total	0	2600	7255	7837	2524	7365	1988	1352	1935	7504	347	1710	2538	4179	49134	0.175128

Tabla 2. Valores Faltantes para cada Estación y sus Respectivas Variables

En esta tabla se puede apreciar que la estación con más datos faltantes es la Estación Noreste 2. También que en total en todo el dataset faltan 49134 datos lo que representa el 17% del dataset.

El siguiente paso de este proceso es identificar datos atípicos. Esto se hizo calculando el valor de Z con respecto a cada estación y su año. En otras palabras cada valor será evaluado sólo con los datos en su columna de su estación y de su año. Si se calculara en respecto a toda la población estaría sesgado por los datos recolectados en un lugar diferente o en otro año. La metodología implementada fue calcular Z.

$$i = columna$$

$$j = estación$$

$$Z = (x - \mu_{ij}) / \sigma_{ij}$$

Posteriormente un dato sería considerado como atípico si su valor Z es mayor a 2. Lo cual significa que ese valor se encuentra a más de dos desviaciones estándar de la media para su respectiva estación y año. El desglose de los datos atípicos se puede ver en la Tabla 3.

	Dataset	CO	NO2	O3	PM10	PM2.5	PRS	RAINF	RH	SO2	SR	TOUT	WSR	WDR	Total_Estacion	Porcentaje
0	Centro_2020	505	587	389	336	396	513	0	199	395	134	329	312	547	4642	0.016545
1	Noreste2_2020	336	80	373	286	380	487	94	247	206	632	347	156	830	4454	0.015875
2	Sureste3_2020	296	545	476	339	388	512	48	246	244	633	365	322	0	4414	0.015733
3	Centro_2021	777	461	391	321	438	471	0	209	304	691	370	336	627	5396	0.019233
4	Noreste2_2021	321	350	463	325	379	480	1	265	763	10	380	181	825	4743	0.016905
5	Sureste3_2021	225	486	331	364	410	346	16	243	288	649	426	350	183	4317	0.015387
6	Total	2460	2509	2423	1971	2391	2809	159	1409	2200	2749	2217	1657	3012	27966	0.099679

Tabla 3. Valores Atípicos para cada Estación y sus Respectivas Variables

En esta tabla se puede apreciar que el 9% de los datos en el dataset son atípicos.

Calcula medidas estadísticas

Variables Cuantitativas

Media:

Media Por Año

Year	PRS	RAINF	RH	SR	TOUT	WSR	WDR
2020	714.874189	0.002596	57.543063	0.176938	22.776099	5.464991	124.824219
2021	714.336249	0.001134	56.800864	0.144953	22.824300	7.687634	128.218224

Media por Estación

Estacion	PRS	RAINF	RH	SR	TOUT	WSR	WDR	Year
Centro_2020	711.649508	0.000000	55.891894	0.171184	22.545980	5.980757	123.467772	2020.0
Centro_2021	711.479000	0.000000	55.054875	0.191332	22.680522	7.077567	113.157451	2021.0
Noreste2_2020	701.551924	0.001789	56.591998	0.173775	22.288953	7.717388	79.352444	2020.0
Noreste2_2021	700.105564	0.002567	55.017200	0.087056	22.578841	9.003122	107.414659	2021.0
Sureste3_2020	731.421503	0.005999	60.145485	0.185855	23.493391	2.696769	171.652596	2020.0
Sureste3_2021	731.424184	0.000837	60.330517	0.156471	23.213537	6.982212	164.082561	2021.0

Mediana:

Mediana Por Año

Year	PRS	RAINF	RH	SR	TOUT	WSR	WDR
2020	711.3	0.0	59.0	0.165	23.12	3.9	103.0
2021	711.4	0.0	58.0	0.038	23.65	7.1	104.0

Mediana por Estación

Estacion	PRS	RAINF	RH	SR	TOUT	WSR	WDR	Year
Centro_2020	711.2	0.0	57.0	0.172	22.89	5.6	96.0	2020.0
Centro_2021	711.4	0.0	56.0	0.158	23.48	6.7	81.0	2021.0
Noreste2_2020	701.2	0.0	58.0	0.007	22.68	6.9	82.0	2020.0
Noreste2_2021	700.6	0.0	56.0	0.033	23.66	8.7	98.0	2021.0
Sureste3_2020	730.4	0.0	63.0	0.041	23.80	2.5	147.0	2020.0
Sureste3_2021	731.4	0.0	63.0	0.036	23.80	6.4	138.0	2021.0

Moda:

Moda Por Año

Year	PRS	RAINF	RH	SR	TOUT	WSR	WDR	Estacion	Year
2020	729.6	0.0	78.0	0.0	23.74	1.3	11.0	Centro_2020	2020.0
2021	699.0	0.0	63.0	0.0	25.16	10.2	10.0	Centro_2021	2021.0

Moda por Estación

Estacion	PRS	RAINF	RH	SR	TOUT	WSR	WDR	Estacion	Year
Centro_2020	711.2	0.0	58.0	0.170	23.03	0.9	78.0	Centro_2020	2020.0
Centro_2021	711.8	0.0	74.0	0.000	24.68	5.9	60.0	Centro_2021	2021.0
Noreste2_2020	700.4	0.0	77.0	0.000	20.44	1.3	11.0	Noreste2_2020	2020.0
Noreste2_2021	699.0	0.0	63.0	0.000	25.16	10.2	10.0	Noreste2_2021	2021.0
Sureste3_2020	729.6	0.0	81.0	0.036	23.31	2.3	123.0	Sureste3_2020	2020.0
Sureste3_2021	730.3	0.0	65.0	0.001	23.12	10.3	225.0	Sureste3_2021	2021.0

Medidas de dispersión: rango: máximo - mínimo, varianza, desviación estándar.

Rango:

Rango Por Año

Year	PRS	RAINF	RH	SR	TOUT	WSR	WDR
2020	58.2	4.39	93.0	0.889	38.31	28.6	359.0
2021	65.7	22.38	93.0	1.007	48.79	62.0	359.0

Rango por Estación

Estacion	PRS	RAINF	RH	SR	TOUT	WSR	WDR
Centro_2020	35.4	0.00	91.0	0.282	34.78	22.7	359.0
Centro_2021	22.0	0.00	91.0	1.007	45.01	21.0	359.0
Noreste2_2020	23.3	0.48	89.0	0.878	36.14	28.6	358.0
Noreste2_2021	33.3	22.38	89.0	0.774	47.43	62.0	359.0
Sureste3_2020	40.7	4.39	93.0	0.889	37.82	11.0	359.0
Sureste3_2021	65.7	2.10	91.0	0.803	45.71	41.9	359.0

Varianza:

Varianza Por Año

Varianza por Estación

Year	PRS	RAINF	RH	SR	TOUT	WSR	WDR
2020	167.858988	0.003084	383.527673	0.036387	41.894295	19.435902	9463.971076
2021	186.486330	0.019399	396.999305	0.045598	46.181681	21.479593	9518.584570

Estacion	PRS	RAINF	RH	SR	TOUT	WSR	WDR
Centro_2020	11.794414	0.000000	368.565659	0.000343	38.191538	11.000374	7793.539807
Centro_2021	10.729360	0.000000	366.518740	0.063706	42.364988	13.596635	8151.478325
Noreste2_2020	13.124200	0.000356	381.451376	0.061360	42.721619	32.435291	6938.864284
Noreste2_2021	26.343547	0.057340	403.361878	0.019847	47.474794	29.923329	9293.930994
Sureste3_2020	16.965703	0.008877	390.250363	0.047348	43.974855	1.873566	9398.956486
Sureste3_2021	19.697437	0.000857	402.517681	0.047615	48.483381	18.323128	9166.435370

Desviación Estándar:

Desviación Estándar Por Año

Year	PRS	RAINF	RH	SR	TOUT	WSR	WDR
2020	167.858988	0.003084	383.527673	0.036387	41.894295	19.435902	9463.971076
2021	186.486330	0.019399	396.999305	0.045598	46.181681	21.479593	9518.584570

Desviación Estándar por Estación

Estacion	PRS	RAINF	RH	SR	TOUT	WSR	WDR
Centro_2020	11.794414	0.000000	368.565659	0.000343	38.191538	11.000374	7793.539807
Centro_2021	10.729360	0.000000	366.518740	0.063706	42.364988	13.596635	8151.478325
Noreste2_2020	13.124200	0.000356	381.451376	0.061360	42.721619	32.435291	6938.864284
Noreste2_2021	26.343547	0.057340	403.361878	0.019847	47.474794	29.923329	9293.930994
Sureste3_2020	16.965703	0.008877	390.250363	0.047348	43.974855	1.873566	9398.956486
Sureste3_2021	19.697437	0.000857	402.517681	0.047615	48.483381	18.323128	9166.435370

Variables cualitativas

Tabla de distribución de frecuencia:

Tabla de Frecuencia Por Año

Resultados para el año: 2020

	1.0	2.0	3.0	4.0	5.0
CO	26255	2	0	0	0
NO2	25591	0	0	0	0
O3	24445	1723	58	0	0
PM10	12912	7896	4947	449	53
PM2.5	15582	8193	2320	162	0
SO2	18041	7231	0	0	0
O3Concentracion	24342	1779	106	6	0
Objetivo	6673	13109	5859	563	53

Resultados para el año: 2021

	1.0	2.0	3.0	4.0	5.0
CO	26204	0	1	0	0
NO2	26205	0	0	0	0
O3	24716	1446	43	0	0
PM10	11578	8267	5008	345	187
PM2.5	20601	4478	1064	62	0
SO2	21110	5095	0	0	0
O3Concentracion	24538	1584	82	1	0
Objetivo	9328	10391	5909	390	187

Tabla de Frecuencia por Estación

Resultados para la estación: Centro_2020

	1.0	2.0	3.0	4.0	5.0
CO	8753	0	0	0	0
NO2	8748	0	0	0	0
O3	7516	1186	51	0	0
PM10	5025	2089	1548	81	10
PM2.5	4556	3082	1034	81	0
SO2	8285	466	0	0	0
O3Concentracion	7487	1179	81	6	0
Objetivo	2826	3691	2072	154	10

Resultados para la estación: Centro_2021

	1.0	2.0	3.0	4.0	5.0
CO	8735	0	0	0	0
NO2	8735	0	0	0	0
O3	7892	819	24	0	0
PM10	3338	3075	2246	60	16
PM2.5	5814	2194	686	41	0
SO2	7696	1039	0	0	0
O3Concentracion	7825	863	47	0	0
Objetivo	2736	3561	2332	90	16

Resultados para la estación: Sureste3_2020

	1.0	2.0	3.0	4.0	5.0
CO	8751	1	0	0	0
NO2	8752	0	0	0	0
O3	8703	49	0	0	0
PM10	3365	3169	1898	295	25
PM2.5	5606	2498	624	24	0
SO2	7053	1686	0	0	0
O3Concentracion	8673	78	1	0	0
Objetivo	2566	3767	2097	297	25

Resultados para la estación: Sureste3_2021

	1.0	2.0	3.0	4.0	5.0
CO	8735	0	0	0	0
NO2	8735	0	0	0	0
O3	8347	372	16	0	0
PM10	3793	2719	1875	206	142
PM2.5	7163	1326	233	13	0
SO2	6751	1984	0	0	0
O3Concentracion	8274	435	26	0	0
Objetivo	3036	3457	1884	216	142

Resultados para la estación: Noreste2_2020

	1.0	2.0	3.0	4.0	5.0
CO	8751	1	0	0	0
NO2	8091	0	0	0	0
O3	8226	488	7	0	0
PM10	4522	2638	1501	73	18
PM2.5	5420	2613	662	57	0
SO2	2703	5079	0	0	0
O3Concentracion	8182	522	24	0	0
Objetivo	1281	5651	1690	112	18

Resultados para la estación: Noreste2_2021

	1.0	2.0	3.0	4.0	5.0
CO	8734	0	1	0	0
NO2	8735	0	0	0	0
O3	8477	255	3	0	0
PM10	4447	2493	1687	79	29
PM2.5	7624	958	145	8	0
SO2	6663	2072	0	0	0
O3Concentracion	8439	286	9	1	0
Objetivo	3556	3373	1693	84	29

Mediana (escala ordinal):

Mediana Por Año	Mediana por Estación

Year	CO	NO2	O3	PM10	PM2.5	SO2	O3Concentracion	Objetivo
2020	1.0	1.0	1.0	2.0	1.0	1.0	1.0	2.0
2021	1.0	1.0	1.0	2.0	1.0	1.0	1.0	2.0

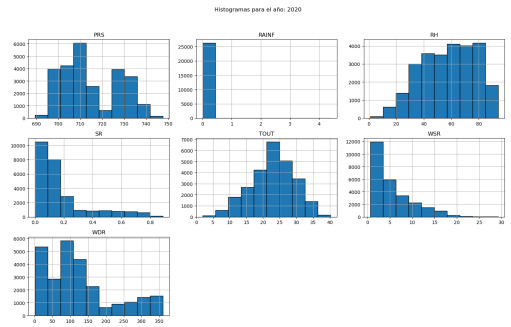
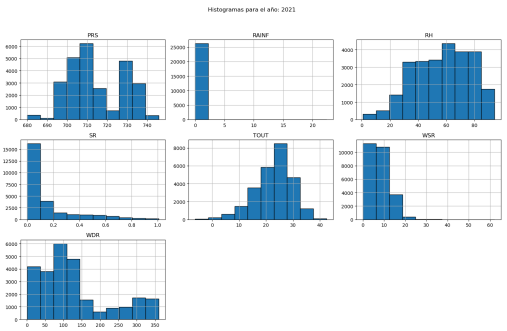
Estacion	CO	NO2	O3	PM10	PM2.5	SO2	O3Concentracion	Objetivo
Centro_2020	1.0	1.0	1.0	1.0	1.0	1.0	1.0	2.0
Centro_2021	1.0	1.0	1.0	2.0	1.0	1.0	1.0	2.0
Noreste2_2020	1.0	1.0	1.0	1.0	1.0	2.0	1.0	2.0
Noreste2_2021	1.0	1.0	1.0	1.0	1.0	1.0	1.0	2.0
Sureste3_2020	1.0	1.0	1.0	2.0	1.0	1.0	1.0	2.0
Sureste3_2021	1.0	1.0	1.0	2.0	1.0	1.0	1.0	2.0

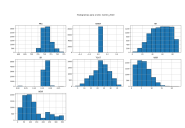
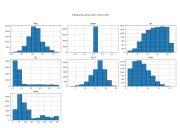
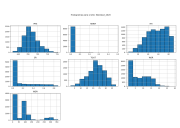
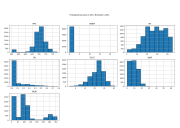
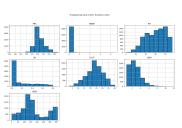
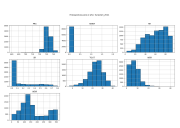
Explora los datos usando herramientas de visualización

Variables cuantitativas:

- Medidas de posición no-central: cuartiles, outlier (valores atípicos), boxplots

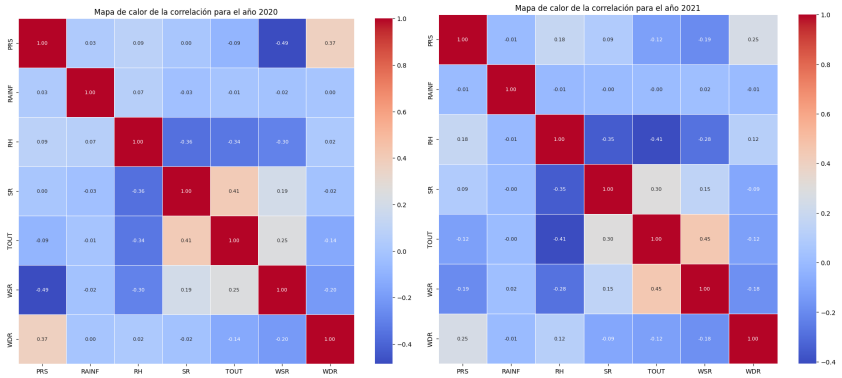
Análisis de distribución de los datos (Histogramas).

Histograma 2020	Histograma 2021
	
Podemos ver que para 2020 las variables de WSR y SR tienen un sesgo a la derecha muy claro. WDR y PRS tienen un sesgo más leve hacia la recha y TOUT es la variable que más se asemeja a una distribución normal.	Podemos ver que para 2021 las variables de WSR y SR tienen un sesgo a la derecha muy claro. WDR y PRS tienen un sesgo más leve hacia la recha y TOUT es la variable que más se asemeja a una distribución normal.

Histograma Centro 2020	Histograma Centro 2021	Histograma Noreste 2020	Histograma Noreste 2021	Histograma Sureste 2021	Histograma Sureste 2021
					
La variable WSR,WDR tiene sesgo a la izquierda.	La variable WSR y SR tienen sesgo a la izquierda.	La variable WSR y SR,WDR tienen sesgo a la izquierda.	La variable WSR y SR tienen sesgo a la izquierda.	La variable WSR y SR tienen sesgo a la izquierda.	La variable WSR y SR tienen sesgo a la izquierda.

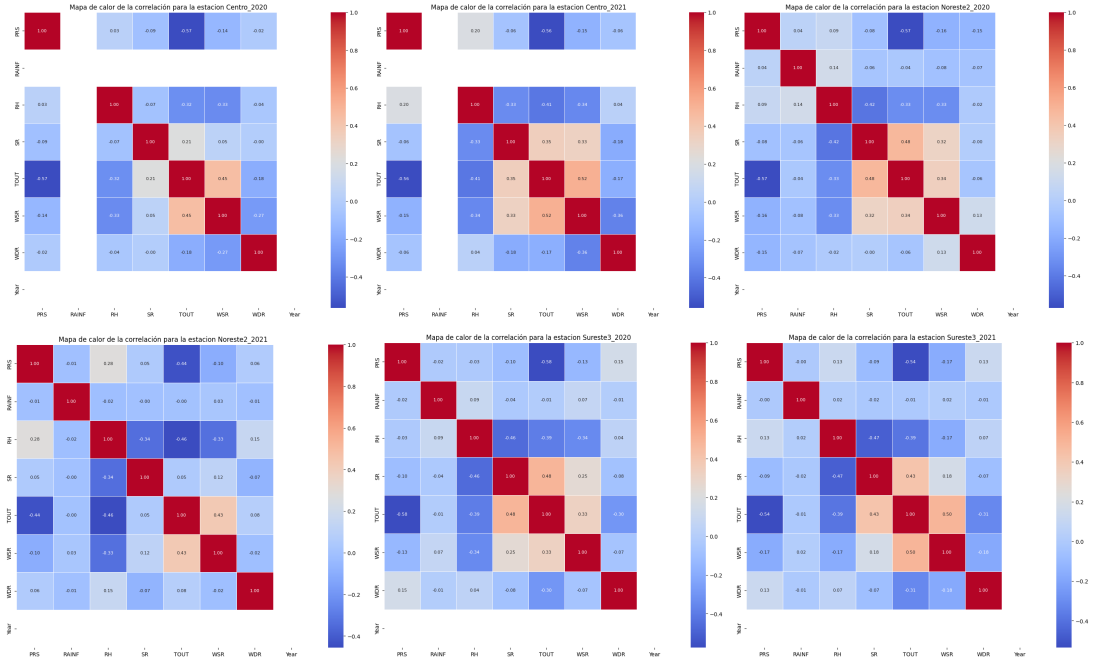
Análisis de correlación de los datos, mapa de calor

Mapas de calor por año



Podemos observar que en ambos años, todas nuestras variables tienen una correlación baja o moderada.

Mapas de calor por estación

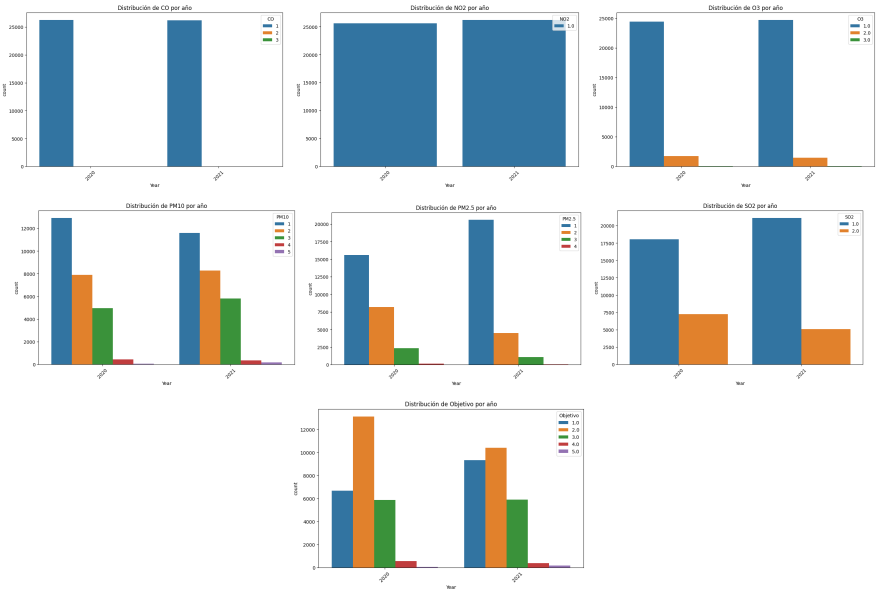


Podemos observar que en las seis estaciones, todas nuestras variables tienen una correlación baja o moderada.

Variables categóricas:

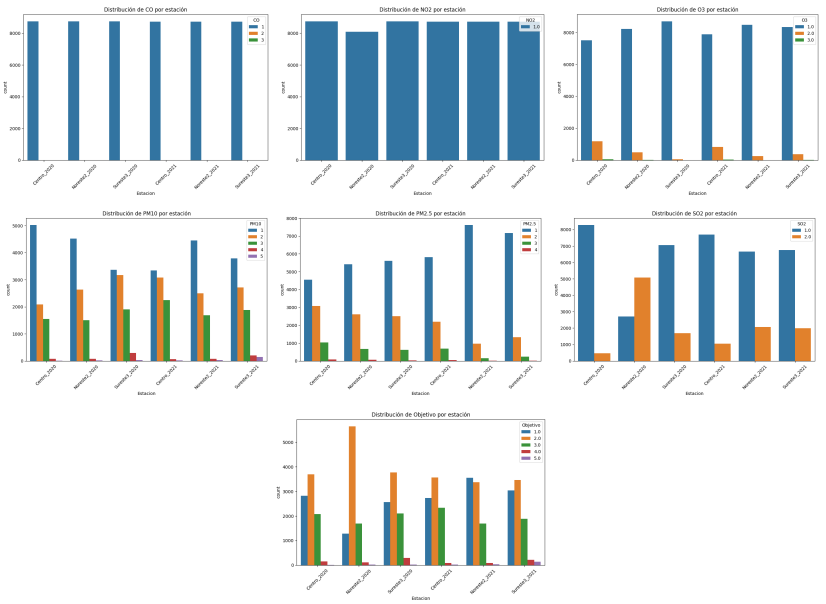
- Distribución de los datos (diagramas de barras, diagramas de pastel)

Histogramas por año



Podemos observar que 2021 presenta mejores resultados para todas las variables con excepción de PM10, también podemos observar que PM10 resulta ser la variable más problemática, presentando la mayoría de las malas calificaciones.

Histogramas por estación



Nuevamente observamos que la variable más problemática es PM10, además podemos notar un pico en Noreste2 de la variable SO2.
Nuestra variable objetivo nos muestra que la estación Noreste2 es la que tiene la mayor cantidad de

¨2¨, sin embargo tiene el menor número de las demás observaciones.

2) Verifica la calidad de los datos:

Después de analizar los datos podemos concluir que hay columnas como la variable Rain que está totalmente vacía, por lo que la mejor opción para el estudio es eliminar esta variable del análisis. Los datos faltantes para cada columna se calcularon de la siguiente manera. En el caso de que se encontrara un dato faltante en la posición xi se reemplazará por el promedio del valor anterior y el valor siguiente:

$$x_i = \frac{x_{i-1} + x_{i+1}}{2}$$

También al verificar los datos y sus distribuciones podemos concluir que ninguno sigue una distribución normal y que la mayoría de variables tienen un sesgo muy marcado a la izquierda o a la derecha.

Concluye con un breve resumen de lo que has logrado hasta ahora, qué falta por hacer, qué camino piensan seguir para lograrlo, cuáles son las preguntas que te han surgido, etc

Hasta el momento hemos logrado limpiar nuestra base de datos y determinar cuales variables no serán útiles para nuestro estudio como la variable Rain que esta llena de ceros. Hemos logrado convertir todas nuestras variables de la medida de los contaminantes de su medición por las estaciones a variables categóricas de su calificación en el índice de calidad de aire según la normativa del gobierno mexicano. También hemos analizado nuestras variables cualitativas y cuantitativas con sus determinadas medidas estadísticas visualmente para lograr entender cómo están distribuidos estos valores y cuantas variables están sesgadas. Como nuestro estudios se enfoca en encontrar si hay diferencia significativa entre la calidad de aire en 2020 y 2021 el análisis estadístico de media, mediana, y moda nos ayudaron a tener una hipótesis de que no hay una diferencia significativa entre los dos años dado que ambos tienen una frecuencia muy parecida en sus valores de la variable objetivo. Las preguntas que nos han surgido es si hay una diferencia significativa en la calidad del aire entre los años 2020 y 2021. Si ha habido un cambio significativo en los niveles de los contaminantes entre los años 2020 y 2021. Finalmente si las condiciones meteorológicas tienen un efecto en el nivel de los contaminantes. Los siguientes pasos a seguir es hacer un análisis de anova tanto por estación y por año para nuestra variable objetivo, hacer un análisis factorial para ver que factores afectan mas nuestra variable objetivo, y hacer un análisis de discriminante para poder clasificar la calidad de aire para una entrada.

Inserta una liga que dé acceso a tu documento de trabajo.

https://drive.google.com/file/d/1nL_FutUiIOpESg7gzDg4_1THi_pxkeo5/view?usp=sharing
https://drive.google.com/file/d/1nL_FutUiIOpESg7gzDg4_1THi_pxkeo5/view?usp=sharing