Taller 2 - Gestión de datos

Calidad del aire - Estación La Flora Santiago de Cali

Ronald Fernando Rodríguez Barbosa
Maestría en Ingeniería de Sistemas y Computación
Maestría en Analítica para la Inteligencia de Negocios
Pontificia Universidad Javeriana

18 de Mayo de 2019

Introducción

La **calidad del aire**, se define como la cantidad general de polución presente en un area y como la pureza promedio atmosférica en relación a las medidas de descarga tomadas de una fuente de polución (Gooch 2007). La contaminación del aire, representa un importante riesgo medioambiental para la salud, bien sea en los países desarrollados o en los países en desarrollo ya que es evidenciado en casos en morbilidad por trastornos cerebrovasculares, cánceres de pulmón y neumopatías crónicas y agudas. Por lo tanto, cuanto más bajos sean los niveles de contaminación del aire, mejor será la salud cardiovascular y respiratoria de la población a largo y a corto plazo (OMS 2016).

De los 23 países de América Latina y el Caribe, 18 tienen sus propias regulaciones vigentes en la actualidad relacionadas con la calidad del aire, que son de acceso público en los sitios web oficiales (Morantes et al. 2016). La trazabilidad para tales regulaciones se establece para los contaminantes de criterio (PM10, PM2.5, SO2, NO2, O-3, CO), utilizando como referencia la secuencia histórica de estándares de la Agencia de Protección Ambiental de los Estados Unidos (USEPA 2019) y los valores de referencia de La Organización Mundial de la Salud (OMS-WHO 2019).

En Colombia, el Sistema de Vigilancia de la Calidad del Aire de Cali (SVCAC 2019), opera bajo la coordinación y administración del Departamento Administrativo de Gestión del Medio Ambiente (DAGMA 2019),Grupo de Calidad del Aire. El Sistema de Vigilancia de Calidad del Aire de Santiago de Cali SVCASC fue acreditado en la norma NTC-ISO/IEC 17025 del año 2005 por el IDEAM a través de la Resolución 1328 del 23 de junio de 2018. El SVCASC actualmente funciona con nueve (9) estaciones: La estación La Flora, ubicada en el barrio La Flora en la zona norte; La estación ubicada en el barrio obrero y la estaciuón Ermita ubicada en el barrio San Pedro ambas en la zona centro; la estación transitoria EDB - Navarro ubicada en el barrio Poblado en la zona oriente; la estación base Aérea, ubicada en el acuaparque de la Caña en zona nororiente; la estación Pance ubicadada en la Zona Rural; la estación Univalle ubicada en el barrio Meléndez en la zona sur; la estación compartir ubicada en el Barrio Compartir en la zona oriente y la estación Cañaveralejo ubicada en la estación SITM del MIO en la zona suroccidente.

El presente trabajo, realizará un análisis de la captura de datos de la estación *La Flora*, la cual es una estación automática que reporta información horaria al centro de control del DAGMA. Esta estación, mide los niveles de Material Particulado Menor a 10 micrómetros (PM10), Dióxido de Azufre (SO2), Dióxido de Nitrógeno (NO2), Monóxido de Carbono (CO), Ozono (O3) y variables meteorológicas como velocidad del viento, dirección del viento, temperatura, humedad, radiación solar y precipitación. El proceso de preparación y análisis de los datos seguirá la siguiente estructura:

1. ***Carga y exploración***: Se incluirá el archivo de datos y se identificarán las características del conjunto de datos
2. ***Limpieza de datos***: A partir de la exploración, se definirán los procedimientos para la limpieza de datos para facilitar el análisis
3. ***Creación de la vista minable***: Se establecerán los conjuntos de datos finales y su correspondiente análisis
4. ***Conclusiones e infografía***: Se compilará el conocimiento adquirido y las cifras de interés adquiridas.

1. Carga y exploración

Tablas de resumen

Para los procedimientos de carga y exploración, se emplearán las herramientas RStudio y RapidMiner. Inicialmente, se realiza la carga de archivos con el fin de resumir las características generales de los datos.

```
datos_base_flora<-read.csv(file="data/dataCAFloraPrep.csv",
                 header = TRUE,sep = ",", dec = ".",fileEncoding = "latin1")
str(datos_base_flora,vec.len=0)
```

```
## 'data.frame':    84629 obs. of  12 variables:
##  $ Fecha...Hora             : Factor w/ 37937 levels "1/1/11 1:00 AM",..: NULL ...
##  $ PM10..ug.m3.             : num  NULL ...
##  $ SO2...ug.m3.             : num  NULL ...
##  $ NO2...ug.m3.             : num  NULL ...
##  $ CO...ug.m3.              : num  NULL ...
##  $ O3...ug.m3.              : num  NULL ...
##  $ Vel.Viento...m.s.        : num  NULL ...
##  $ Dir.Viento..Grados.      : num  NULL ...
##  $ Temperatura..CÂ..        : num  NULL ...
##  $ Humedad....              : num  NULL ...
##  $ Radiacion.Solar..Watt.M2.: num  NULL ...
##  $ Lluvia..mm.              : int  NULL ...
```

El conjunto de datos contiene un total 84.629 observaciones con 12 variables. La descripción de las variables se relaciona a continuación:

Variable	Tipo de variable	Descripción
Fecha & Hora	Fecha	Fecha y hora de la captura del senso de polutantes
PM10 (ug/m3)	Contínua	Concentración de Material Particulado Menor a 10 micrómetros
SO2 (ug/m3)	Contínua	Concentración de Dióxido de Azufre
NO2 (ug/m3)	Contínua	Concentración de Dióxido de Nitrógeno
CO (ug/m3)	Contínua	Concentración de Monóxido de Carbono
O3 (ug/m3)	Contínua	Concentración de Ozono
Vel Viento (m/s)	Contínua	Velocidad del viento en metros por segundo
Dir Viento (Grados)	Contínua	Dirección del viento
Temperatura (C°)	Contínua	Temperatura en grados celsius
Humedad (%)	Contínua	Porcentaje de humedad
Radiacion Solar	Contínua	Radiación Solar (Watt/M2)
Lluvia (mm)	Contínua	Cantidad de precipitaciones

Resumen de frecuencia de valores en las variables

Con el fin de identificar los posibles valores en las variables, la frecuencia de dichos valores de forma general y detectar valores ausentes se realiza la siguiente exploración mediante la exposición y resumen. Con el fin de facilitar su visualización se realiza una partición por cada 3 variables, utilizando la función summary del lenguaje R.

En las siguientes listas, se puede apreciar una cantidad significativa de registros con ausencia de valor numérico o valores en cero, con diferentes magnitudes de frecuencia entre las diferentes variables. Según los boletines emitidos por la SVCASC, dichos datos faltantes pueden estar relacionados a las anomalías que se dan en las estaciones de monitoreo, tales como: Fallas en los equipos, falta de energía eléctrica en la zona, hurto de

equipos o cableado, mantenimiento o cambio de equipos y la inclusión o exclusión de algunos contaminantes o variables meteorológicas. Por otra parte, se puede evidenciar que existe más de un registro con misma fecha y hora, lo que puede sugerir un proceso de limpieza en el que se excluirían dichos registros duplicados. Este procedimiento, se verá con más detalle en la sección 2.

```
summary(datos_base_flora[1:3])
```

```
##           Fecha...Hora    PM10..ug.m3.    SO2...ug.m3.
##  1/1/11 1:00 AM :    4   Min.   :  0.1   Min.   :  0.03
##  1/1/11 10:00 AM:    4   1st Qu.: 17.1   1st Qu.:  7.62
##  1/1/11 11:00 AM:    4   Median : 31.0   Median : 12.34
##  1/1/11 12:00 PM:    4   Mean   : 34.8   Mean   : 19.05
##  1/1/11 2:00 AM :    4   3rd Qu.: 47.2   3rd Qu.: 21.04
##  1/1/11 3:00 AM :    4   Max.   :245.0   Max.   :334.60
##  (Other)        :84605   NA's   :17551   NA's   :64788
```

```
summary(datos_base_flora[4:6])
```

```
##   NO2...ug.m3.     CO...ug.m3.       O3...ug.m3.
##  Min.   :  0.02   Min.   :  11.23   Min.   :  0.00
##  1st Qu.: 17.39   1st Qu.: 999.77   1st Qu.:  3.96
##  Median : 23.98   Median :1408.64   Median :  8.82
##  Mean   : 26.16   Mean   :1497.34   Mean   : 22.69
##  3rd Qu.: 32.46   3rd Qu.:1902.66   3rd Qu.: 31.11
##  Max.   :124.50   Max.   :4970.61   Max.   :231.40
##  NA's   :58116    NA's   :69992     NA's   :53828
```

```
summary(datos_base_flora[7:9])
```

```
##  Vel.Viento...m.s. Dir.Viento..Grados. Temperatura..CÂ..
##  Min.   :0.00      Min.   :  0.0       Min.   :16.20
##  1st Qu.:0.30      1st Qu.: 75.3       1st Qu.:22.20
##  Median :0.60      Median :155.3       Median :24.20
##  Mean   :0.82      Mean   :165.5       Mean   :24.75
##  3rd Qu.:1.10      3rd Qu.:258.3       3rd Qu.:27.30
##  Max.   :7.40      Max.   :360.0       Max.   :34.80
##  NA's   :39021     NA's   :39021       NA's   :39022
```

```
summary(datos_base_flora[10:12])
```

```
##   Humedad....    Radiacion.Solar..Watt.M2.  Lluvia..mm.
##  Min.   : 23.4   Min.   :  0.0             Min.   :  0.00
##  1st Qu.: 57.5   1st Qu.:  0.0             1st Qu.:  0.00
##  Median : 71.7   Median :  5.1             Median :  0.00
##  Mean   : 70.6   Mean   :178.2             Mean   :  0.24
##  3rd Qu.: 82.9   3rd Qu.:345.0             3rd Qu.:  0.00
##  Max.   :100.4   Max.   :992.4             Max.   :115.00
##  NA's   :39021   NA's   :26964             NA's   :7023
```

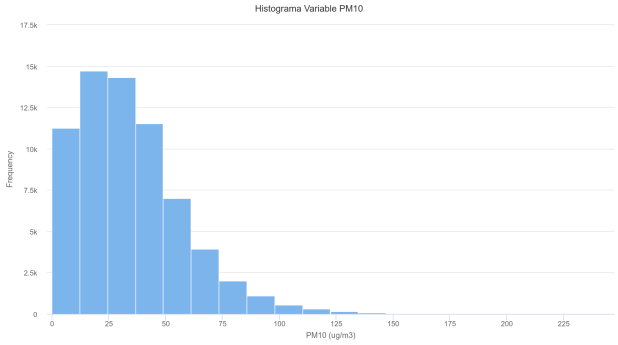
Métricas de tendencia central, forma y dispersión

Explicar librerira psych, dplyr y los valores que se están calculando Explicar librerira psych, dplyr y los valores que se están calculando Explicar librerira psych, dplyr y los valores que se están calculando Explicar librerira psych, dplyr y los valores que se están calculando Explicar librerira psych, dplyr y los valores que se están calculando Archivo 1_visualización_exploración.rmp

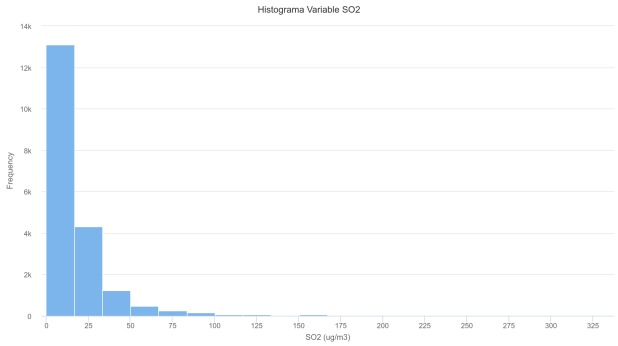
```
library(psych)
library(dplyr)
estadisticas <- describe(datos_base_flora[,2:12],na.rm = TRUE, interp=FALSE,skew = TRUE,
            ranges = TRUE,trim=.1,check=TRUE,fast=NULL,quant=c(.25,.50,.75),IQR=TRUE)
estadisticas <- estadisticas%>%mutate(Q0.75 + (1.5* IQR))
estadisticas <- estadisticas%>%mutate(Q0.25 - (1.5* IQR))
estadisticas
```

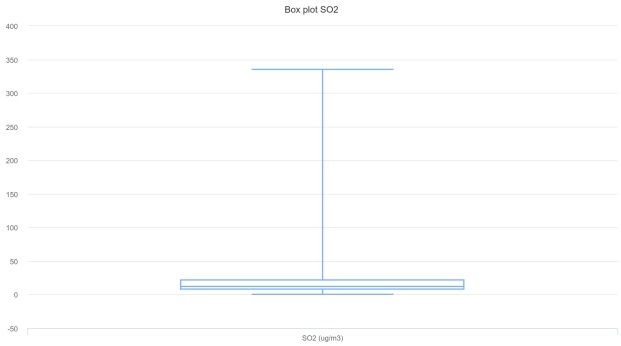
```
##    vars     n          mean           sd  median      trimmed         mad
## 1     1 67078   34.7997376  23.8538055   31.00   32.2670971  22.090740
## 2     2 19841   19.0547089  23.4678521   12.34   14.4286789   8.465646
## 3     3 26513   26.1596504  12.6160812   23.98   24.9010518  10.852632
## 4     4 14637 1497.3424896 720.2868315 1408.64 1450.4953736 656.851104
## 5     5 30801   22.6939064  29.0319482    8.82   16.6626167  10.704372
## 6     6 45608    0.8248224   0.7186832    0.60    0.7128015   0.593040
## 7     7 45608  165.4732240 101.4770316  155.30  162.6492929 137.288760
## 8     8 45607   24.7470388   3.1275970   24.20   24.6022912   3.558240
## 9     9 45608   70.6005723  14.9864314   71.70   70.7135798  18.680760
## 10   10 57665  178.2396098 246.4868363    5.10  136.6000368   7.561260
## 11   11 77606    0.2395691   2.2602299    0.00    0.0000000   0.000000
##       min     max   range       skew    kurtosis          se      IQR
## 1    0.10  245.00  244.90  1.3031292   3.1569537 0.092101695  30.1000
## 2    0.03  334.60  334.57  4.7681813  33.6002347 0.166606356  13.4200
## 3    0.02  124.50  124.48  1.2231264   2.6750929 0.077480968  15.0700
## 4   11.23 4970.61 4959.38  0.7151988   0.8073156 5.953597064 902.8900
## 5    0.00  231.40  231.40  1.8226874   3.0355610 0.165422192  27.1471
## 6    0.00    7.40    7.40  1.8316292   4.8376458 0.003365247   0.8000
## 7    0.00  360.00  360.00  0.1795216  -1.1707887 0.475168060 183.0000
## 8   16.20   34.80   18.60  0.3579858  -0.8436607 0.014645191   5.1000
## 9   23.40  100.40   77.00 -0.0980337  -1.0294851 0.070174240  25.4000
## 10   0.00  992.40  992.40  1.0880605  -0.3023481 1.026449488 345.0000
## 11   0.00  115.00  115.00 18.0165404 461.5626109 0.008113439   0.0000
##         Q0.25    Q0.5   Q0.75 Q0.75 + (1.5 * IQR) Q0.25 - (1.5 * IQR)
## 1   17.100000   31.00   47.20            92.35000           -28.05000
## 2    7.620000   12.34   21.04            41.17000           -12.51000
## 3   17.390000   23.98   32.46            55.06500            -5.21500
## 4  999.770000 1408.64 1902.66          3256.99500          -354.56500
## 5    3.962895    8.82   31.11            71.83066           -36.75776
## 6    0.300000    0.60    1.10             2.30000            -0.90000
## 7   75.300000  155.30  258.30           532.80000          -199.20000
## 8   22.200000   24.20   27.30            34.95000            14.55000
## 9   57.500000   71.70   82.90           121.00000            19.40000
## 10   0.000000    5.10  345.00           862.50000          -517.50000
## 11   0.000000    0.00    0.00             0.00000             0.00000
```

bla bla bla PM10

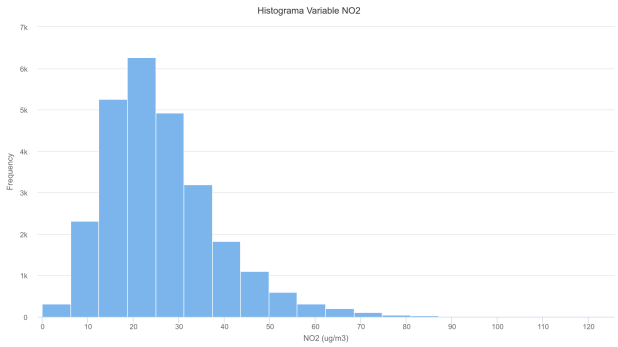


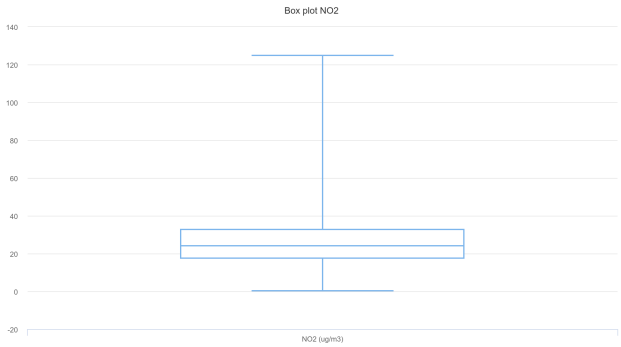
bla bla bla SO2



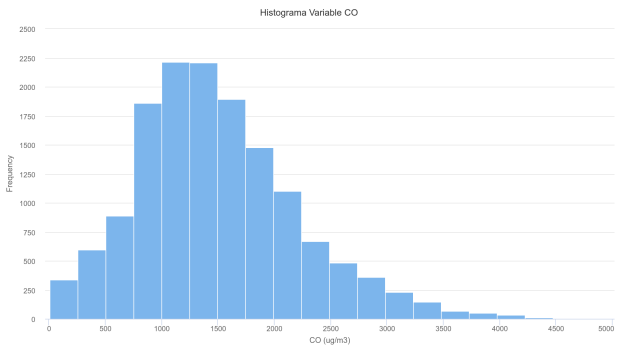


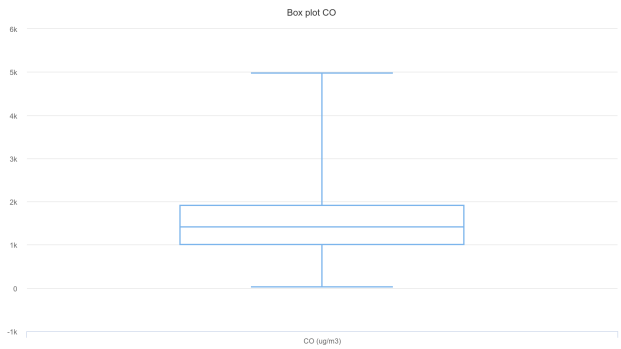
bla bla bla NO2



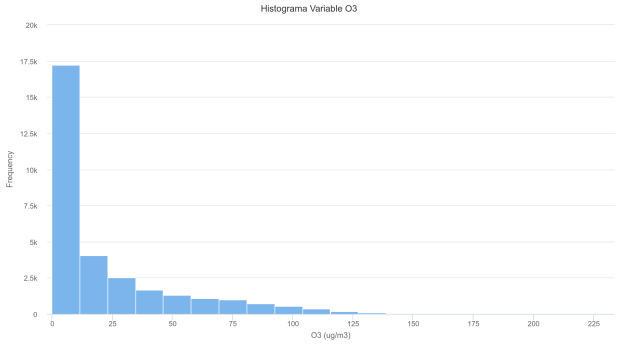


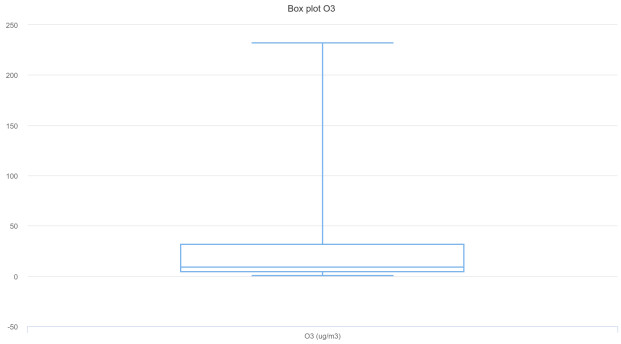
bla bla bla CO



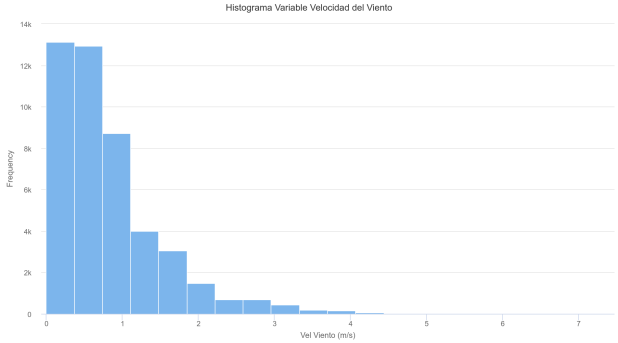


bla bla bla O3

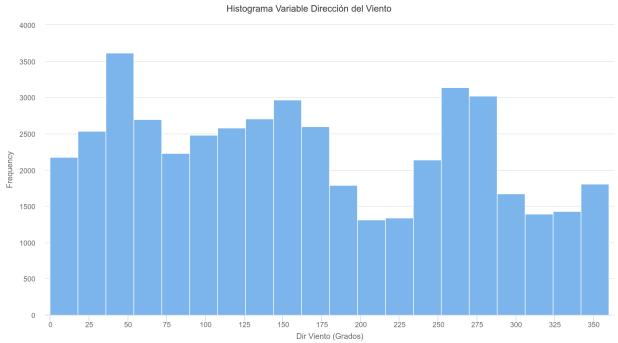


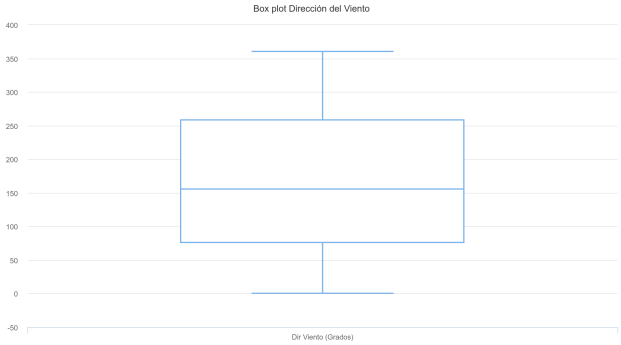


bla bla bla Velocidad del viento

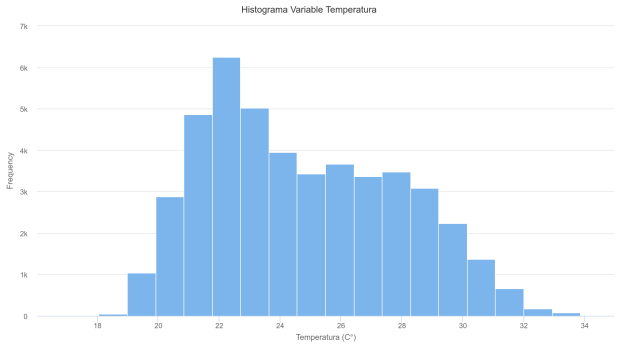


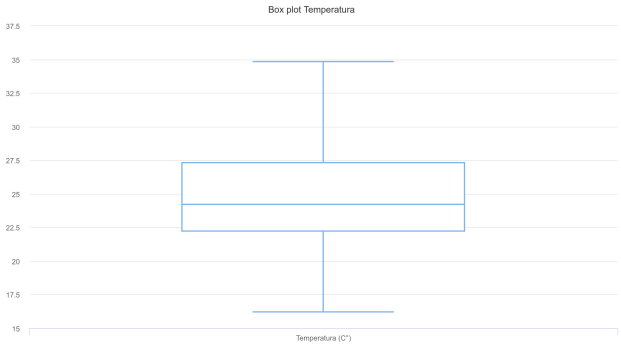
bla bla bla Dirección del viento



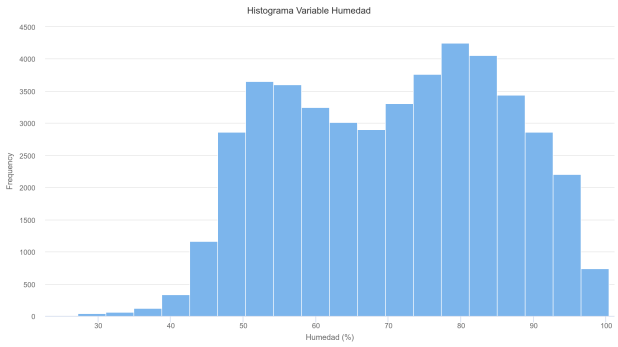


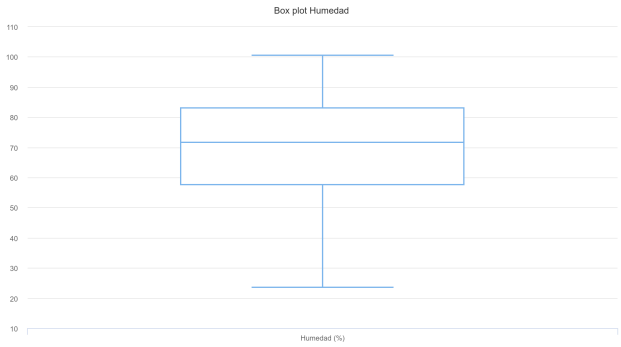
bla bla bla Temperatura



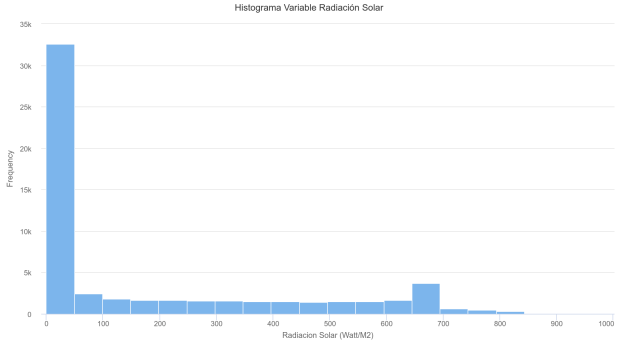


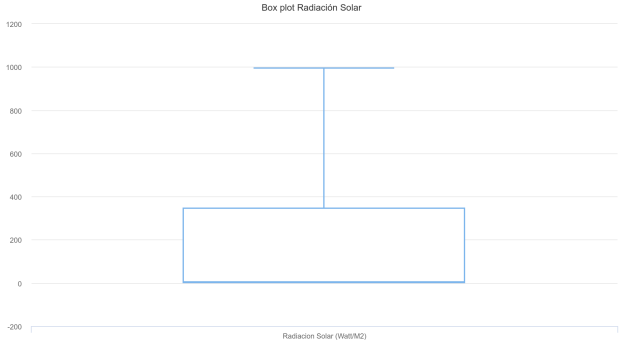
bla bla bla Humedad



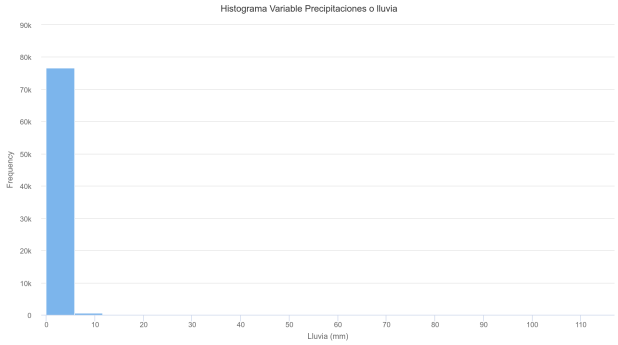


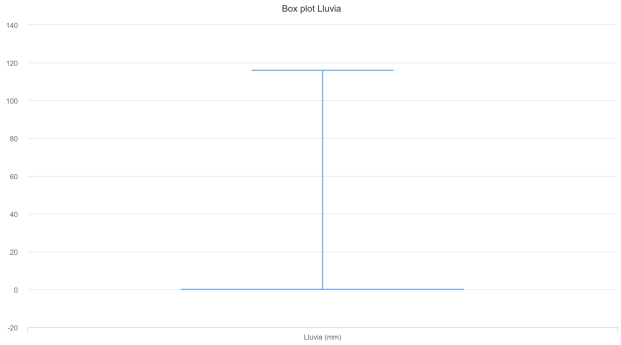
bla bla bla RSolar





bla bla bla Lluvia





Análisis de Correlación

bla bla bla correlación

Attributes	PM10 ...	SO2 ...	NO2...	CO ...	O3 ...	Vel Viento...	Dir Viento...	Temper...	Humedad...	Radiacion Solar...	Lluvia ...
PM10 (ug/m3)	1	0.165	0.466	0.233	-0.036	-0.162	-0.126	-0.018	0.007	0.097	-0.068
SO2 (ug/m3)	0.165	1	0.275	-0.041	0.160	0.110	-0.088	0.236	-0.207	0.338	-0.033
NO2 (ug/m3)	0.466	0.275	1	0.346	-0.075	-0.181	-0.111	-0.066	0.041	0.024	-0.044
CO (ug/m3)	0.233	-0.041	0.346	1	-0.242	-0.120	-0.014	-0.120	0.078	-0.040	-0.050
O3 (ug/m3)	-0.036	0.160	-0.075	-0.242	1	0.408	0.103	0.724	-0.659	0.551	-0.034
Vel Viento (m/s)	-0.162	0.110	-0.181	-0.120	0.408	1	0.079	0.487	-0.469	0.349	0.043
Dir Viento (Grados)	-0.126	-0.088	-0.111	-0.014	0.103	0.079	1	0.051	-0.078	-0.177	0.013
Temperatura (C°)	-0.018	0.236	-0.066	-0.120	0.724	0.487	0.051	1	-0.957	0.601	-0.152
Humedad (%)	0.007	-0.207	0.041	0.078	-0.659	-0.469	-0.078	-0.957	1	-0.523	0.180
Radiacion Solar (Watt/M2)	0.097	0.338	0.024	-0.040	0.551	0.349	-0.177	0.601	-0.523	1	-0.071
Lluvia (mm)	-0.068	-0.033	-0.044	-0.050	-0.034	0.043	0.013	-0.152	0.180	-0.071	1

2. Limpieza de datos

Archivo 2_limpieza_datos.rmp

Reconocimiento y tratamiento de atributos con valores únicos o distintos

El campo fecha y hora corresponde a un campo único, sin embargo existen registros duplicados con dicho campo y se evidencian valores diferentes entre los conjuntos resultantes

Conjunto A con 37937 observaciones

Fecha & Hora	Date time	0	Earliest date May 6, 2010 8:00 AM	Latest date Jan 1, 2019 12:00 PM	Duration 3162d 4h 0m 0s
PM10 (ug/m3)	Real	7910	Min 0.100	Max 234.400	Average 40.387
SO2 (ug/m3)	Real	31815	Min 0.030	Max 334.600	Average 15.696
NO2 (ug/m3)	Real	27485	Min 0.240	Max 124.500	Average 26.217
CO (ug/m3)	Real	32809	Min 11.230	Max 4516.310	Average 1575.293
O3 (ug/m3)	Real	24923	Min 0	Max 192.900	Average 9.348
Vel Viento (m/s)	Real	19176	Min 0	Max 3.400	Average 0.564
Dir Viento (Grados)	Real	19176	Min 0	Max 360	Average 154.783
Temperatura (C°)	Real	19177	Min 16.200	Max 31	Average 22.739
Humedad (%)	Real	19177	Min 37.900	Max 100.400	Average 80.165
Radiacion Solar (Watt/M2	Real	13155	Min 0	Max 918.500	Average 132.789
Lluvia (mm	Integer	3178	Min 0	Max 98.550	Average 0.335

Conjunto B con 46692 observaciones

Fecha & Hora	Date time	0	Earliest date May 7, 2010 1:00 AM	Latest date Dec 31, 2018 12:00 PM	Duration 3160d 11h 0m 0s
PM10 (ug/m3)	Real	9641	Min 0.100	Max 245	Average 30.272
SO2 (ug/m3)	Real	32973	Min 0.030	Max 334.600	Average 20.554
NO2 (ug/m3)	Real	30631	Min 0.020	Max 124.500	Average 26.122
CO (ug/m3)	Real	37183	Min 11.230	Max 4970.610	Average 1455.305
O3 (ug/m3)	Real	28905	Min 0	Max 231.400	Average 32.458
Vel Viento (m/s)	Real	19845	Min 0	Max 7.400	Average 1.007
Dir Viento (Grados)	Real	19845	Min 0	Max 360	Average 172.944
Temperatura (C°)	Real	19845	Min 16.200	Max 34.800	Average 26.150
Humedad (%)	Real	19844	Min 23.400	Max 100.400	Average 63.917
Radiacion Solar (Watt/M2	Real	13809	Min 0	Max 992.400	Average 212.493
Lluvia (mm	Integer	3845	Min 0	Max 115.820	Average 0.212

Reconocimiento y tratamiento de atributos con valores faltantes

Reconocimiento y tratamiento de atributos con valores atípicos

Reconocimiento y tratamiento de registros atípicos

Reconocimiento y tratamiento de atributos redundantes

3. Creación de la vista minable

Generación de variables derivadas tipo 1 y 2

Normalización de al menos un atributo

Discretización de al menos un atributo

Numerización 1 a n de al menos un atributo

4. Conclusiones e Infografía

Referencias

DAGMA, Departamento Administrativo de Gestión del Medio Ambiente. 2019. "Sitio Oficial - Departamento Administrativo de Gestión Del Medio Ambiente." http://www.cali.gov.co/dagma/.

Gooch, Jan W., ed. 2007. "Ambient Air Quality." In *Encyclopedic Dictionary of Polymers*, 48–48. New York,

NY: Springer New York. doi:10.1007/978-0-387-30160-0_522.

Morantes, Giobertti, Narciso Perez, Rafael Santana, and Gladys Rincon. 2016. "A REVIEW OF THE REGULATORY INSTRUMENTS FOR AIR QUALITY AND ATMOSPHERIC MONITORING SYSTEMS: LATIN AMERICA AND THE CARIBBEAN." *INTERCIENCIA* 41 (4): 235–42.

OMS, Organizacion Mundial de la Salud. 2016. "Calidad Del Aire Ambiente Y Salud." http://origin.who.int/mediacentre/factsheets/fs313/es/.

OMS-WHO, Organizacion Mundial de la Salud. 2019. "Sitio Oficial - Organización Mundial de La Salud." https://www.who.int/es/home/.

SVCAC, Sistema de Vigilancia de Calidad del Aire de Cali. 2019. "Sitio Oficial - Sistema de Vigilancia de Calidad Del Aire de Cali." http://www.cali.gov.co/dagma/publicaciones/38365/sistema_de_vigilancia_de_calidad_del_aire_de_cali_svcac/.

USEPA, Agencia de Protección Ambiental de los Estados Unidos. 2019. "Sitio Oficial - Agencia de Protección Ambiental de Los Estados Unidos." https://www.epa.gov/.