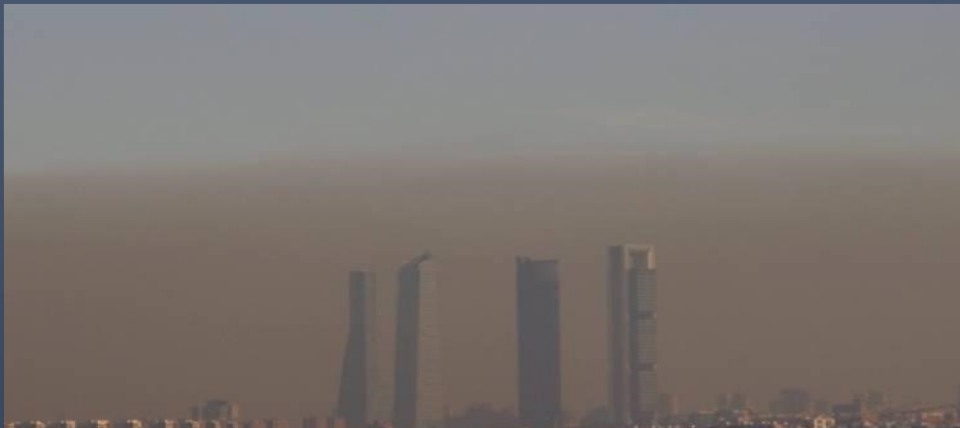


Memoria final TGI: Estudio sobre la contaminación en Madrid



REALIZADO POR:

DAVID BAUDET MORENO
SERGIO CALAHORRO UREÑA

ÍNDICE

INTRODUCCIÓN AL PROBLEMA	2
JUSTIFICACIÓN DE LA SOLUCIÓN ELEGIDA.....	2
DESCRIPCIÓN DE LOS CONJUNTOS DE DATOS ELEGIDOS.....	4
ESTUDIO SOBRE LAS CUESTIONES PLANTEADAS SOBRE LA INFORMACIÓN	5
DISEÑO DEL ESQUEMA.....	6
DESCRIPCIÓN DE LOS PROCESOS DE EXTRACCIÓN, TRANSFORMACIÓN Y CARGA.....	10
PLANTEAMIENTO DE CONSULTAS	11
CONCLUSIONES SOBRE EL TRABAJO Y AUTOEVALUACIÓN	14
BIBLIOGRAFÍA	16

INTRODUCCIÓN AL PROBLEMA

El objetivo de nuestro estudio es determinar cuál es la situación de Madrid en temas de polución, observando las mediciones tomadas en las diversas estaciones de control que hay repartidas por la ciudad, y cómo afecta el tráfico a la polución, ya que podemos observar últimamente en las noticias que en la ciudad se toman diversas medidas para combatir la polución que generan una gran controversia.

Algunas de estas medidas tomadas son, por ejemplo, las obras en Gran Vía para reducir carriles de circulación y ensanchar el acerado, la prohibición de circulación de vehículos que no dispongan de etiqueta ambiental de la DGT por la M-30 ni por el centro de la ciudad, el fomento de vehículos eléctricos (menos contaminantes que los convencionales), reducción de la velocidad máxima en algunos tramos de la ciudad...

En nuestro proyecto hemos realizado un estudio sobre la contaminación en la provincia de Madrid en sus principales estaciones de control, y la influencia que tiene el tráfico sobre la contaminación.

JUSTIFICACIÓN DE LA SOLUCIÓN ELEGIDA

La herramienta que hemos utilizado para el proyecto de prácticas es SAS (University Edition).

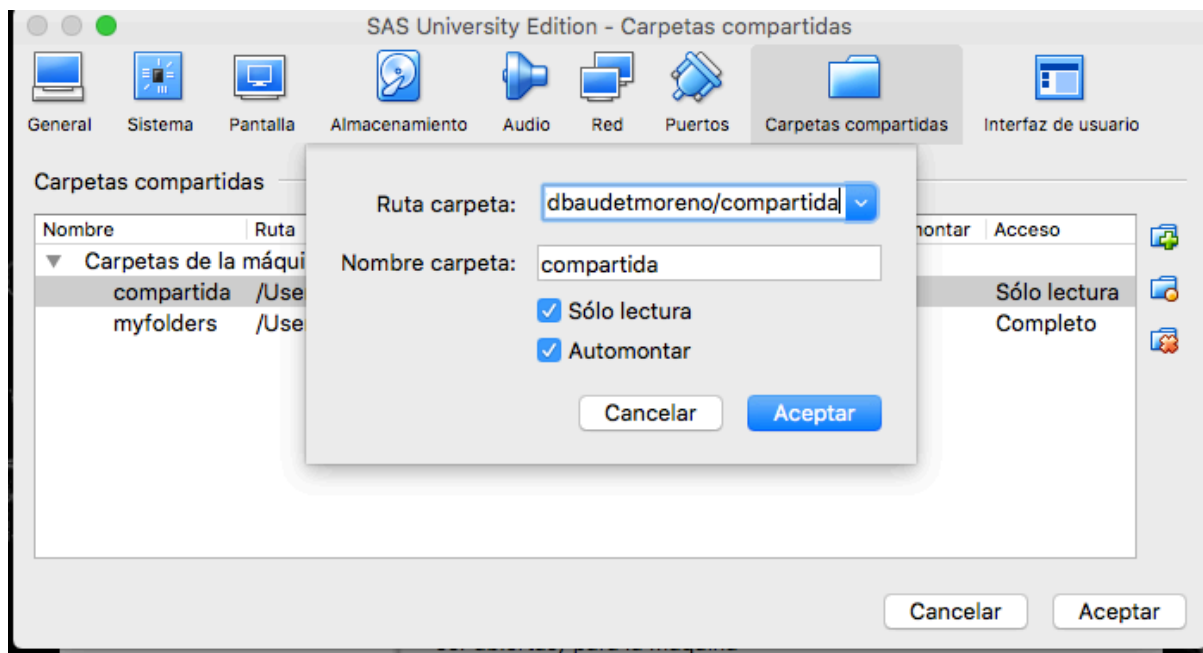
Nos hemos decidido por este software por su capacidad de gestionar datos, transformarlos, crear informes y gráficos de alto impacto visual junto a su capacidad de distribución a terceros. Además pudimos observar que existen gran cantidad de tutoriales tanto en la propia web del software como por parte de usuarios ajenos a éste, por lo que pensamos que de esta forma el aprendizaje sería más sencillo al ser un entorno nuevo para nosotros.

Sus características más importantes son:

- Potente modelo de explotación de Datos.
- Excelente herramienta ETL para la construcción de modelos de datos para su posterior explotación tanto con Analytics Pro como con Visual Analytics de SAS.
- Integración con otras soluciones de SAS.
- Interfaz gráfica de rápida asimilación por los usuarios e intuitiva que permite interactuar con el software de ordenadores Windows, Linux y Mac.
- Ofrecer capacidades de autoservicio a un nivel empresarial.

- Accesible también mediante código SAS. Ofrece un potente lenguaje de programación fácil de aprender y de usar, el cual permite análisis especializados y exploración más en profundidad.
- Proporciona herramientas de ámbito extenso que incluyen métodos estadísticos del estado del arte.
- Responder preguntas complejas de forma rápida y elevar la productividad gracias a su potencial analítico.

En la Wiki de la asignatura está toda la información de los pasos a seguir para la instalación del software. Una vez instalada la máquina virtual, hay que crear una carpeta llamada *myfolders* que debe añadirse como carpeta compartida para la máquina guest en VirtualBox y activar la opción de “Automontar”.



DESCRIPCIÓN DE LOS CONJUNTOS DE DATOS ELEGIDOS

Los conjuntos de datos elegidos para la realización del proyecto son:

- **Calidad del aire. Estaciones de control:**
 - Número de estación
 - Nombre de estación
 - Dirección
 - Longitud
 - Latitud
 - Altitud
 - Tipo de estación
 - Contaminante medido:
 - NO₂: dióxido de nitrógeno ($\mu\text{g}/\text{m}^3$)
 - SO₂: dióxido de azufre ($\mu\text{g}/\text{m}^3$)
 - CO: monóxido de carbono (mg/m^3)
 - PM₁₀: partículas < 10 μm ($\mu\text{g}/\text{m}^3$)
 - PM_{2,5}: partículas < 2,5 μm ($\mu\text{g}/\text{m}^3$)
 - O₃: ozono ($\mu\text{g}/\text{m}^3$)
 - BTX: benceno, tolueno y xilenos ($\mu\text{g}/\text{m}^3$)
 - HC: hidrocarburos (mg/m^3)
 - Sensores meteorológicos (UV, VV, DV, TMP, HR, PRB, RS, LL)
- **Calidad del aire. Tiempo real:**
 - Provincia
 - Municipio
 - Número de estación
 - Magnitud
 - Punto de muestreo
 - Año
 - Mes
 - Día
 - Dato medido por hora (01-24)
 - Validación del dato por hora
- **Calidad del aire. Datos diarios años 2001-2018:**
 - Provincia
 - Municipio
 - Número de estación
 - Magnitud
 - Punto de muestreo
 - Año
 - Mes
 - Dato medido por día (01-31)
 - Validación del dato por día

- **Air Pollution in Madrid (CO):**
 - Fecha
 - Valor medio de CO (mg/m³)
- **Estudio del parque circulante de la ciudad de Madrid (años 2013 y 2017):**
 - Sector
 - Subsector
 - Tecnología
 - Zona A (% de ocupación)
 - Zona B (% de ocupación)
 - Zona C (% de ocupación)
 - Zona D (% de ocupación)
 - Zona E (% de ocupación)
 - Total (% de tipos de vehículos)
- **Tráfico. Histórico de datos del tráfico desde 2013:**
 - Identificación única del punto de medida
 - Fecha (cada 15 minutos)
 - Intensidad (vehículos/hora)
 - Ocupación (% de vehículos)
 - Carga (grado de uso de la vía, de 0 a 100)
 - Tipo de punto de medida
 - Velocidad media (km/h)
 - Error
 - Periodo de integración (nº de muestras)

ESTUDIO SOBRE LAS CUESTIONES PLANTEADAS SOBRE LA INFORMACIÓN

Las consultas realizadas consistirán en conocer:

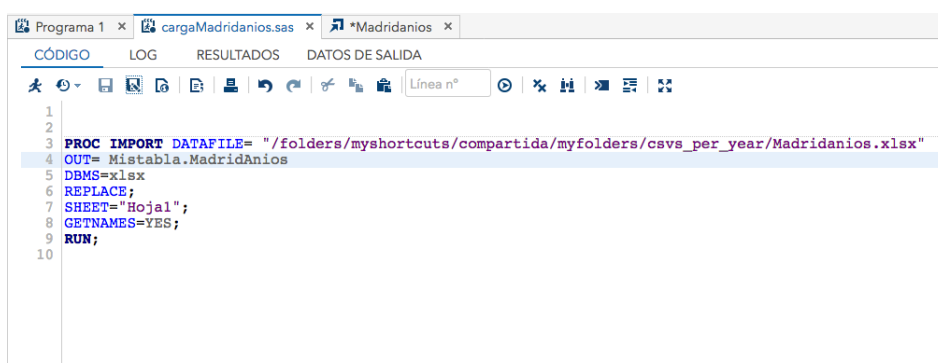
- Niveles de polución por años/meses/días/horas en las estaciones de control.
- Tipo de contaminante medido por las estaciones control, su posición (altitud, coordenadas), tipos de sensores meteorológicos de cada estación de control.
- Niveles de CO en Madrid por años/meses/días/horas.
- Tipo de vehículos por zona.

- Niveles de tráfico por zona.
- Tipo de tecnología de los vehículos.

DISEÑO DEL ESQUEMA

Por defecto, al crear tablas en SAS éstas quedan guardadas en el directorio de trabajo llamado 'WORK.IMPORT', que es un directorio utilizable sólo en la sesión de trabajo actual. De esta forma, cada vez que se cierra sesión en la máquina virtual todas las tablas creadas serán eliminadas, y al iniciar sesión aparecerá ese directorio completamente vacío, pues la sesión de trabajo acabaría de empezar, teniendo entonces que ejecutar los scripts de creación de tablas cada vez que se inicie sesión en la plataforma. Esto es bastante engorroso y tedioso.

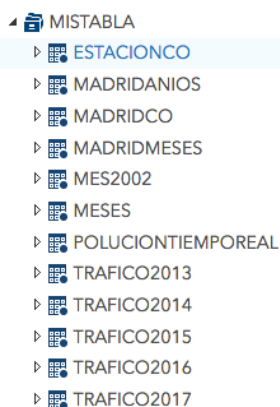
Para evitar tener que hacer esto siempre que se inicie sesión, cambiamos los scripts de creación de tablas de forma que ahora el directorio de trabajo donde se almacenan las tablas cambia a una librería llamada "MISTABLA". De esta forma, todas las tablas de la librería quedan almacenadas de forma persistente y estarán disponibles siempre que se inicie sesión sin necesidad de crearlas de nuevo con cada inicio de sesión en la plataforma. La carga de los archivos se haría de la siguiente forma para cada conjunto de datos:



```

Programa 1 x cargaMadridanos.sas x *Madridanos x
CÓDIGO LOG RESULTADOS DATOS DE SALIDA
1
2
3
4 PROC IMPORT DATAFILE= "/folders/myshortcuts/compartida/myfolders/csvs_per_year/Madridanos.xlsx"
5 OUT= Mistabla.MadridAnios
6 DBMS=xlsx
7 REPLACE;
8 SHEET="Hoja1";
9 GETNAMES=YES;
10 RUN;
  
```

Una vez comentado ésto, éstas serían todas las tablas que contienen los conjuntos de datos anteriormente mencionados.



Nótese que, aunque en la imagen anterior la mayoría de tablas se corresponden exactamente con los datasets mencionados previamente, algunas no lo harán así porque están filtradas (originalmente su tamaño era excesivo e imposibilitaba la representación de los datos), por lo que sólo contienen datos de forma parcial.

En la siguiente tabla se muestra el nivel de CO, (medido en mg/m³) por cada estación, cuyo código viene representado por:

- **28** (código de la comunidad de Madrid), que se repetirá en todas las estaciones
- **079** (código del municipio de Madrid), que se repetirá en todas las estaciones
- **0XX** (número de la estación), que variará dependiendo de la estación

	CO	station
1	0.3700000047683716	28079001
2	0.3400000035762787	28079035
3	0.2800000011920929	28079003
4	0.4699999988079071	28079004
5	0.38999998569488525	28079039
6	0.6299999952316284	28079006
7	0.2800000011920929	28079007
8	0.6700000166893005	28079009
9	0.4099999964237213	28079038
10	0.17000000178813934	28079011
11	0.3799999952316284	28079012
12	0.17000000178813934	28079040
13	0.18000000715255737	28079014
14	0.23999999463558197	28079015

En la siguiente tabla se muestran los niveles medidos por hora, siendo válidos los datos horarios que en la columna 'VXX' tengan un carácter 'V'.

	ESTACION	ANO	MES	DIA	H01	V01	H02
1	4	2018	12	2	12	V	12
2	4	2018	12	2	000.4	V	000.5
3	4	2018	12	2	15	V	30
4	4	2018	12	2	43	V	50
5	4	2018	12	2	65	V	97
6	8	2018	12	2	7	V	7
7	8	2018	12	2	000.6	V	000.6
8	8	2018	12	2	77	V	74
9	8	2018	12	2	84	V	80
10	8	2018	12	2	16	V	17
11	8	2018	12	2	26	V	29
12	8	2018	12	2	203	V	194
13	8	2018	12	2	06.74	V	06.56
14	8	2018	12	2	003.0	V	004.2

En la siguiente tabla se muestra el código de estación, siguiendo el formato mencionado antes, la técnica utilizada y magnitud del dato medido. Esta tabla es análoga a las de otros datasets para años/meses/días/horas, sólo cambiará que incorporarán más columnas y filas.

	ESTACI_N	MAGNITUD	T_CNICA	DATO_DIARIO	A_O
1	28079001	1	38	2	2
2	28079001	1	38	2	2
3	28079001	1	38	2	2
4	28079001	1	38	2	2
5	28079001	1	38	2	2
6	28079001	1	38	2	2
7	28079001	1	38	2	2
8	28079001	1	38	2	2
9	28079001	1	38	2	2
10	28079001	1	38	2	2
11	28079001	1	38	2	2
12	28079001	1	38	2	2
13	28079001	1	38	2	2
14	28079001	1	38	2	2

En la siguiente tabla se muestra el código de estación, siguiendo el formato mencionado antes, la técnica utilizada y magnitud del dato medido, incorporando estos datos al punto de muestreo, siguiendo el formato: códigoEstación_magnitud_técnica.

	PROVINCIA	MUNICIPIO	ESTACION	MAGNITUD	PUNTO_MUES...
1	28	79	4	1	28079004_1_38
2	28	79	4	6	28079004_6_48
3	28	79	4	7	28079004_7_8
4	28	79	4	8	28079004_8_8
5	28	79	4	12	28079004_12_8
6	28	79	8	1	28079008_1_38
7	28	79	8	6	28079008_6_48
8	28	79	8	7	28079008_7_8
9	28	79	8	8	28079008_8_8
10	28	79	8	9	28079008_9_47
11	28	79	8	10	28079008_10_47
12	28	79	8	12	28079008_12_8
13	28	79	8	14	28079008_14_6
14	28	79	8	20	28079008_20_59

En la siguiente tabla se muestran datos referentes al tráfico: identificador del punto de medida, fecha (cada 15 minutos), intensidad, ocupación, carga, tipo de punto de medida, velocidad media, datos de error y periodo de integración (número de muestras tomadas).

	identif	fecha	intensidad	ocupacion	carga	tipo
1	61081	20JAN13:20:15:00	336	1	9	E
2	61082	20JAN13:20:15:00	320	2	10	E
3	61083	20JAN13:20:15:00	704	10	51	E
4	61085	20JAN13:20:15:00	328	2	10	E
5	61086	20JAN13:20:15:00	236	1	7	E
6	61087	20JAN13:20:15:00	268	1	7	E
7	61088	20JAN13:20:15:00	264	2	8	E
8	61089	20JAN13:20:15:00	212	1	7	E
9	61090	20JAN13:20:15:00	288	1	9	E
10	61091	20JAN13:20:15:00	180	2	11	E
11	62085	20JAN13:20:15:00	484	7	29	E
12	62100	20JAN13:20:15:00	800	2	13	E
13	62501	20JAN13:20:15:00	240	1	14	E
14	62556	20JAN13:20:15:00	744	9	23	E

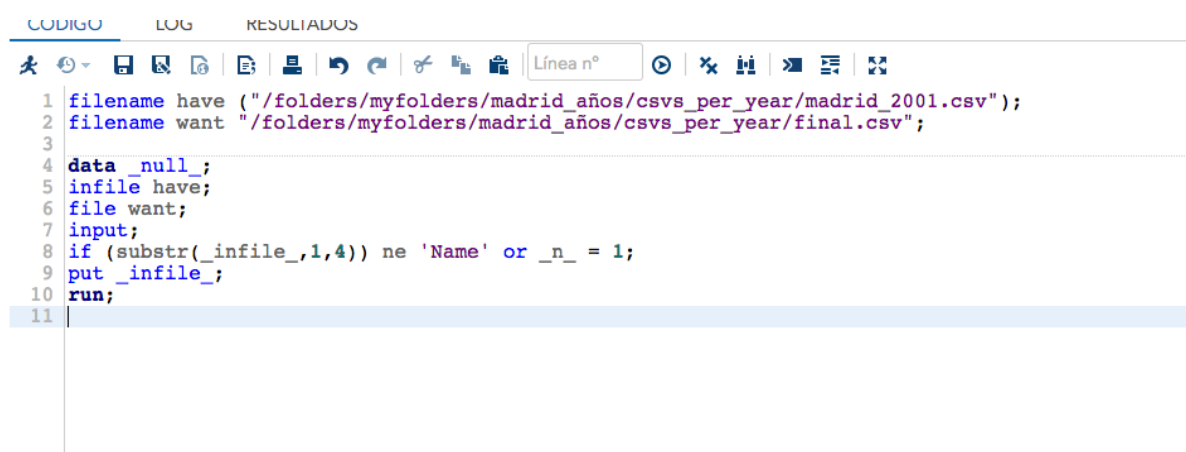
vmed	error	periodo_integracion
0	N	5
0	N	5
0	N	5
0	N	5
0	N	5
0	N	5
0	N	5
0	S	5
0	N	5
0	N	5
0	N	5
0	N	5
0	N	5
0	N	5
0	N	5
0	N	5

DESCRIPCIÓN DE LOS PROCESOS DE EXTRACCIÓN, TRANSFORMACIÓN Y CARGA

Muchos de los conjuntos de datos que hemos escogido para el proyecto estaban en formato TXT y CSV mientras que la mayoría de datasets estaban en formato XLS o XLSX, que son los archivos compatibles con Microsoft Excel y en los cuales los datos están separados en distintas columnas, sin necesidad de que haya algún tipo de carácter que sirva como separación entre ellos (como las comas ',' en los archivos CSV). Ya que trabajar con los datos separados en columnas es más cómodo para poder cargarlos en las tablas de BBDD en SAS, decidimos que antes de la carga era necesario un preprocesamiento de los archivos para poder convertir su formato original al formato deseado. De esta forma, podríamos trabajar con todos los conjuntos de datos con los mismos formatos.

Para llevar a cabo este preprocesamiento de los archivos utilizamos Microsoft Excel, cargando los archivos CSV y exportándolos a formato XLSX, habiendo eliminado previamente las filas redundantes para que la carga de los ficheros en SAS fuera lo más eficiente posible. De esta forma ya tendríamos archivos perfectamente utilizables para trabajar con ellos en SAS.

En la siguiente captura se muestra cómo combinar varios archivos CSV en uno sólo.



The screenshot shows a SAS code editor window with a toolbar at the top. The code is as follows:

```

1 filename have ("/folders/myfolders/madrid_años/csvs_per_year/madrid_2001.csv");
2 filename want "/folders/myfolders/madrid_años/csvs_per_year/final.csv";
3
4 data _null_;
5 infile have;
6 file want;
7 input;
8 if (substr(_infile_,1,4)) ne 'Name' or _n_ = 1;
9 put _infile_;
10 run;
11

```

El caso de los archivos TXT era más complejo porque, a diferencia de los archivos en formato CSV, éstos no tienen ningún tipo de carácter que indique el inicio o el fin de cada tipo de dato para su separación en columnas, de forma que en cada fila estaban juntos los datos de todos los campos (columnas) sin ningún tipo de separación.

Para ello tuvimos que buscar en la documentación asociada a cada dataset, para poder saber dónde empezaba y terminaba cada tipo de dato para poder separarlos en columnas, y a qué campo hacía referencia cada uno de los datos de la fila. También tuvimos que añadir todos los campos (nombres de las columnas) en la primera fila de los archivos resultantes, para que al cargarlos en las tablas de SAS apareciesen los nombres al inicio de cada columna.

A la hora de la extracción de los datos de las tablas nos dimos cuenta de las limitaciones que tiene la versión de SAS que hemos utilizado (University Edition), ya que con conjuntos de datos muy grandes el tiempo de las operaciones era excesivo y en muchos casos surgían

errores, sobre todo a la hora de representarlos. Por ello, hemos tenido que simplificar las tablas realizando un filtrado de la información y volcando esta información en distintas tablas, utilizando éstas para su representación (realización de gráficas y consultas).

PLANTEAMIENTO DE CONSULTAS

En este fragmento de código se muestra un ejemplo de cómo se guardarían las consultas en tablas (filtrado de datos) usando una sentencia SQL.

```
1 PROC SQL;
2 CREATE TABLE MISTABLA.meses AS
3 SELECT ESTACI__N , MAGNITUD , MES , D__A FROM MISTABLA.MADRIDMESES;
4 RUN;
5 QUIT;
6
7 PROC DATASETS NOLIST NODETAILS;
8 CONTENTS DATA=WORK.query OUT=WORK.details;
9 RUN;
10
11 PROC PRINT DATA=WORK.details;
12 RUN;
```

En este fragmento de código se muestra una sentencia SQL para calcular las medias de intensidades sobre la intensidad del tráfico en el año 2013.

```
1 PROC SQL;
2
3 SELECT SUM(intensidad)/1048575 as AÑO2013 FROM MISTABLA.TRAFICO2013;
4
5 RUN;
6 QUIT;
```

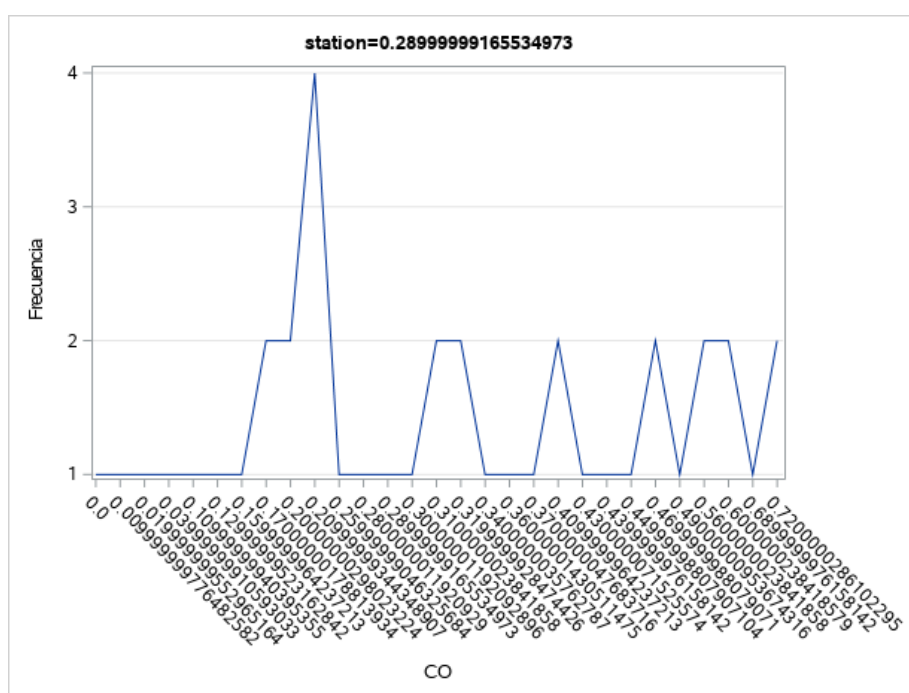
En este fragmento de código se muestra cómo se consigue mediante sentencias SQL la media de magnitudes medidas en cada estación de control para cada mes de 2002.

```
1 PROC SQL;
2
3 SELECT SUM(Magnitud)/1048575 as Enero2002 FROM MISTABLA.MADRIDMESES;
4 WHERE A_O = 2 AND MES = 1;
5
6
7
8 SELECT SUM(Magnitud)/1048575 as Febrero2002 FROM MISTABLA.MADRIDMESES;
9 WHERE A_O = 2 AND MES = 2;
10
11 SELECT SUM(Magnitud)/1048575 as Marzo2002 FROM MISTABLA.MADRIDMESES;
12 WHERE A_O = 2 AND MES = 3;
13
14 SELECT SUM(Magnitud)/1048575 as Abril2002 FROM MISTABLA.MADRIDMESES;
15 WHERE A_O = 2 AND MES = 4;
16
17 SELECT SUM(Magnitud)/1048575 as Mayo2002 FROM MISTABLA.MADRIDMESES;
18 WHERE A_O = 2 AND MES = 5;
19
20 SELECT SUM(Magnitud)/1048575 as Junio2002 FROM MISTABLA.MADRIDMESES;
21 WHERE A_O = 2 AND MES = 6;
22
23 SELECT SUM(Magnitud)/1048575 as Julio2002 FROM MISTABLA.MADRIDMESES;
24 WHERE A_O = 2 AND MES = 7;
```

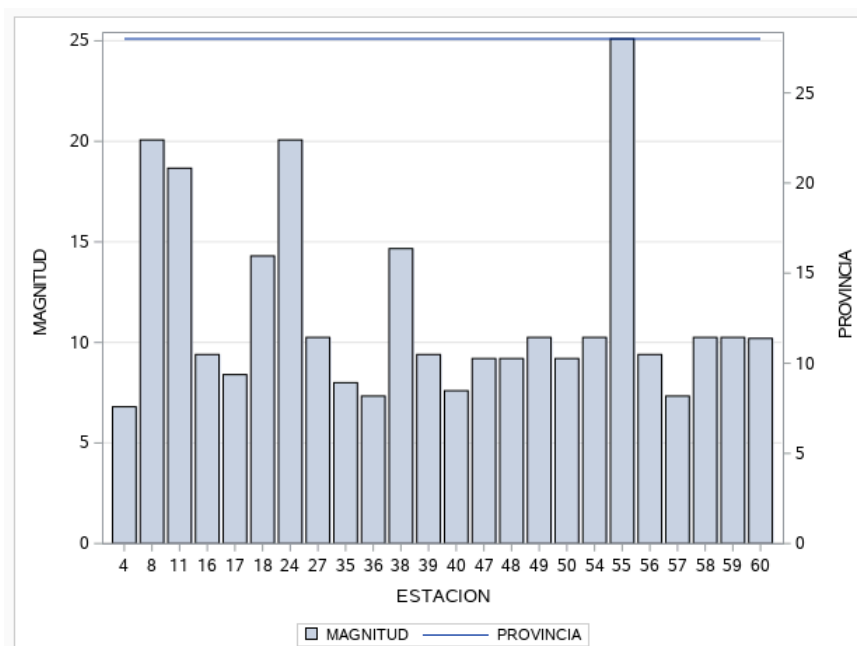
En la siguiente tabla se muestra el resultado de la consulta anterior.

	Meses	Media
1	Enero	15,81193
2	Febrero	14,91193
3	Marzo	15,91193
4	Abril	13,91293
5	Mayo	16,81193
6	Junio	14,91194
7	Julio	15,91194
8	Agosto	13,91294
9	Septiembre	17,81193
10	Octubre	14,91195
11	Noviembre	15,91195
12	Diciembre	13,91295

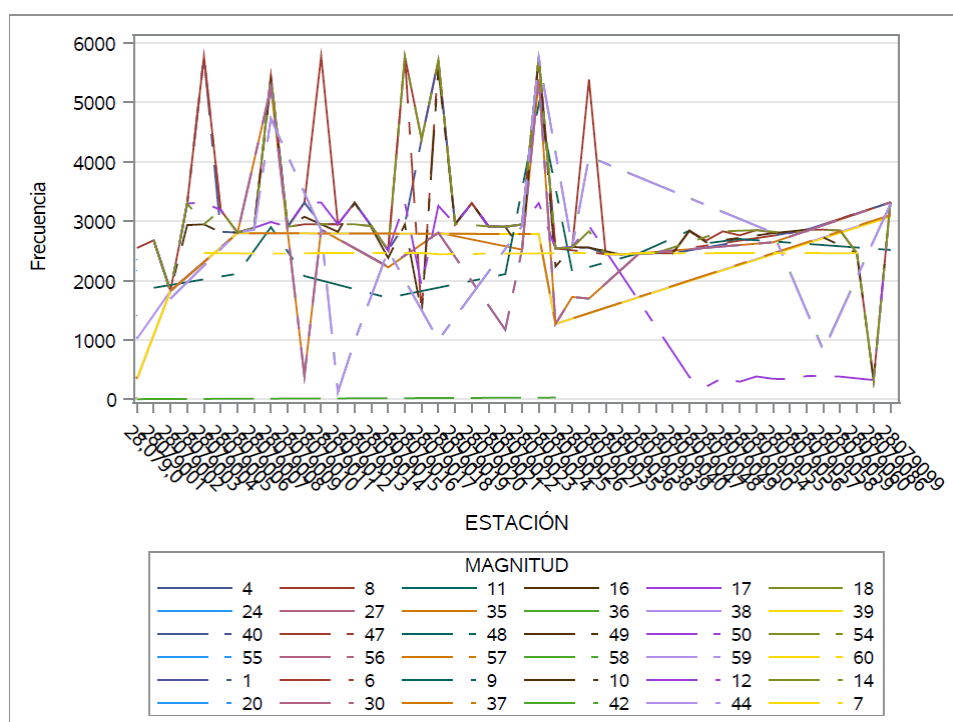
En el siguiente gráfico se muestra la frecuencia de datos de CO medidos en las estaciones de control.



En este gráfico se muestran las magnitudes medidas por estación de control en la provincia de Madrid.



En este gráfico se muestra la frecuencia de magnitudes medidas por cada estación de control.



El trafico por zona en Madrid lo podemos consultar con el siguiente código:

```
PROC SQL;
CREATE TABLE Mistabla.traficozona AS
SELECT identif ,fecha,intensidad,tipo,ocupacion FROM MISTABLA.TRAFICO2013;
RUN;
QUIT;

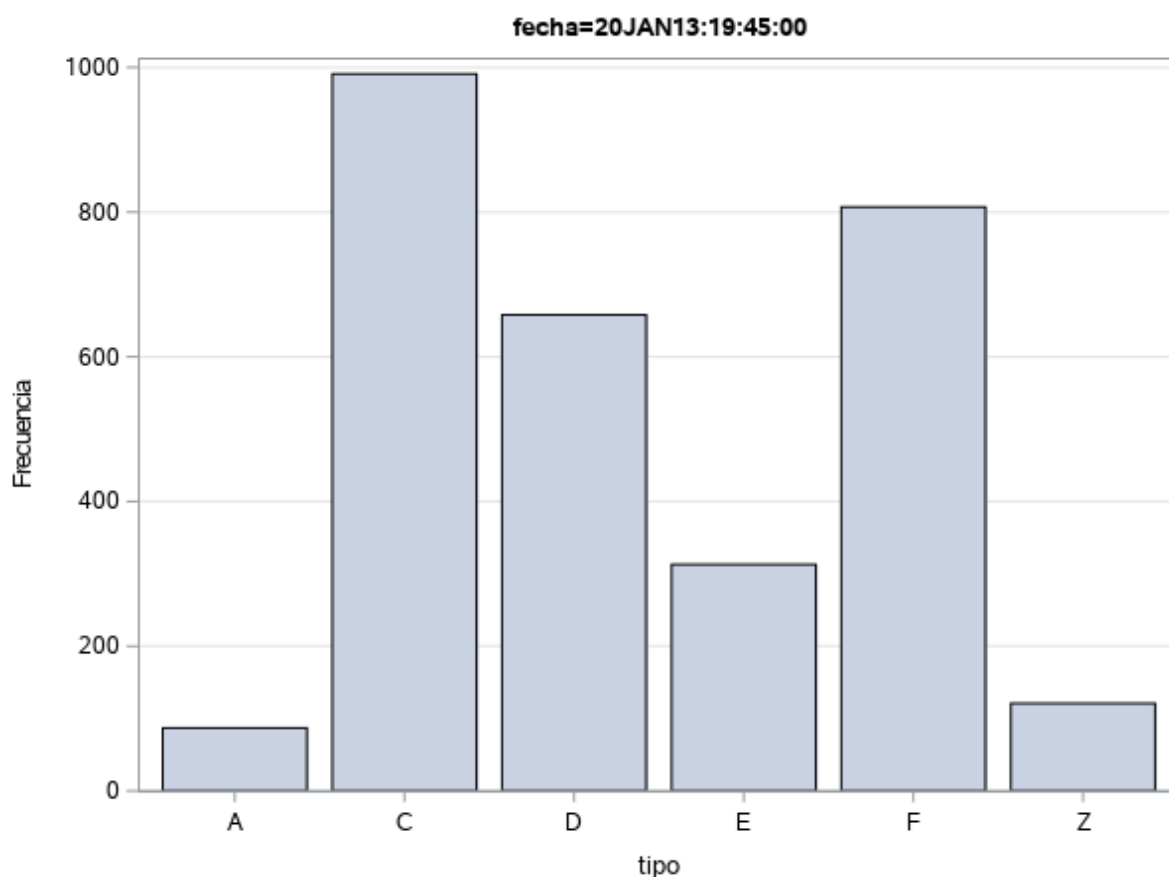
PROC DATASETS NOLIST NODETAILS;
CONTENTS DATA=WORK.query OUT=WORK.details;
RUN;

PROC PRINT DATA=WORK.details;
RUN;
```

Mostrándonos los datos en la tabla siguiente.

identif	fecha	intensidad	tipo
61081	20JAN13:20:15:00	336	E
61082	20JAN13:20:15:00	320	E
61083	20JAN13:20:15:00	704	E
61085	20JAN13:20:15:00	328	E
61086	20JAN13:20:15:00	236	E
61087	20JAN13:20:15:00	268	E
61088	20JAN13:20:15:00	264	E
61089	20JAN13:20:15:00	212	E
61090	20JAN13:20:15:00	288	E
61091	20JAN13:20:15:00	180	E
62085	20JAN13:20:15:00	484	E
62100	20JAN13:20:15:00	800	E
62501	20JAN13:20:15:00	240	E

Dándonos la opción de crear graficas agrupando por fecha para saber el tipo de vehículo:



CONCLUSIONES SOBRE EL TRABAJO Y AUTOEVALUACIÓN

Sobre SAS, la herramienta utiliza un formato amigable en el uso de los datos y en las distintas herramientas que proporciona para editar además de para la propia representación de los datos mediante distintos tipos de gráficas. Es fácil de utilizar y la interfaz gráfica permite la mayoría de funciones, además de que hay una gran comunidad en Internet para obtener documentación y tutoriales para el aprendizaje en la utilización del software.

Como puntos negativos podríamos mencionar que, al estar el software integrado en una máquina virtual y tener que utilizar el protocolo HTTP para las peticiones entre las máquinas host y guest, si las tablas son muy grandes (como en nuestro caso, en el que algún archivo sobrepasaba el límite de filas que soportaba Microsoft Excel) acabarían surgiendo problemas a la hora de representar los datos, ya que la petición quedaba desechada al haber un gran número de datos. La solución a este problema consistió en filtrar los datos y volcarlos en otras tablas, como explicamos anteriormente en la sección [“Descripción de los procesos de Extracción, Transformación y Carga”](#), para así poder representarlos.

Además, durante el transcurso de las prácticas nos surgió un error con la carpeta *boot* de SAS que nos obligó a tener que empezar el proyecto de nuevo. Por desgracia no hay documentación disponible para estos tipos de errores.

En conclusión, el desarrollo de la práctica en general nos ha gustado. Nos pareció interesante ver cómo se hace el tratamiento de los datos desde la visión de una empresa o para llevar a cabo un estudio, como en nuestro caso, no sólo en el ámbito de un usuario normal, sino en el ámbito de un administrador de un almacén de datos. Durante el transcurso de la práctica hemos aprendido la importancia que tiene realizar un buen estudio inicial de los conjuntos de datos y las herramientas de trabajo para poder elegir los elementos idóneos para el trabajo a desempeñar, y a su vez poder observar y entender las propias limitaciones del software y los problemas que pueden llegar a surgir durante su uso. Hemos aprendido una forma nueva de tratamiento de datos, así que estamos contentos.

BIBLIOGRAFÍA

-Conjuntos de datos:

- [Portal de datos abiertos del Ayuntamiento de Madrid](#)
- [Kaggle](#)

-Documentación/tutoriales:

- [SAS \(How To Tutorials\)](#)
- [Manual de Introducción a SAS \(UAB\)](#)
- [Curso de introducción a la programación SAS v8 \(UCM\)](#)
- [tutorialspoint](#)
- [edureka!](#)
- [SAS Support Communities](#)

-Recursos interesantes:

- [Mapa interactivo de estaciones de control](#)
- [Tabla resumen de legislación de calidad del aire](#)