Proyecto Análisis de Datos

Parte 1: Auditoría de datos

Preparación, auditoria y análisis de variables.

José Manuel Bustos Muñoz
Santiago Martinez De la Riva

1. **Descripción de las variables y valores estadísticos (mínimo, máximo, media, desviación, mediana, etc.). Estudia qué valores estadísticos son los convenientes según el tipo de variable y procede en consecuencia.**

Lo primero sería la carga de datos. Se realiza la carga con los datos que se va a trabajar en el proyecto, que son los datos del dataset de "Bike-Sharing-Dataset".
Este dataset tiene como objetivo el predecir el uso diario y horario de un sistema de alquiler de bicicletas basándose en datos climatológicos, día de la semana, temporada, etc. El dataset incluye dos años de uso de bicicletas del sistema público de Washington DC.

Se han cargado los dos dataset proporcionados, uno con los datos por día, y otro por hora de cada día. Ambos datasets tienen las mismas columnas, excepto la columna "hr" que sólo tiene el dataset con los datos por hora en lugar de diarios.

Los tamaños de cada uno de los dataset son los siguientes:
- Dataset diario: num_rows: 731, Columnas: 16.
- Dataset horario: num_rows: 17379, Columnas: 17.

Las columnas del dataset horario:
- instant: record index
- dteday: date
- season: season (1:springer, 2:summer, 3:fall, 4:winter)
- yr: year (0: 2011, 1:2012)
- mnth: month (1 to 12)
- hr: hour (0 to 23)
- holiday: weather day is holiday or not
- weekday: day of the week
- workingday: if day is neither weekend nor holiday is 1, otherwise is 0.
- weathersit:
    1: Clear, Few clouds, Partly cloudy, Partly cloudy
    2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
    3: Light Snow, Light Rain + Thunderstorm + Scattered clouds
    Light Rain + Scattered clouds
    4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
- temp: Normalized temperature in Celsius. The values are divided to 41 (max)
- atemp: Normalized feeling temperature in Celsius. The values are divided to 50 (max)
- hum: Normalized humidity. The values are divided to 100 (max)
- windspeed: Normalized wind speed. The values are divided to 67 (max)
- casual: count of casual users
- registered: count of registered users
- cnt: count of total rental bikes including both casual and registered

A partir de este punto nos quedamos con el dataset con los datos horarios, que tienen las mismas variables que el otro dataset más el dato de la hora, además que tiene bastantes más registros para analizar. Con este dataset se puede hacer por ejemplo un análisis por la hora de uso, además de por el mes o la estación del año, así que nos proporciona mayor información y posibilidades.

Para continuar con el análisis del dataset no tendremos en cuenta la primera columna del mismo “instant” que actúa de índice y por tanto no contiene información relevante.
df_bike_hours = df_bike_hour.iloc[:,1:]

Usamos *describe* para ver los primeros datos estadísticos sobre cada atributo del dataset:

	count	mean	std	min	25%	50%	75%	max
season	17379.0	2.501640	1.106918	1.00	2.0000	3.0000	3.0000	4.0000
yr	17379.0	0.502561	0.500008	0.00	0.0000	1.0000	1.0000	1.0000
mnth	17379.0	6.537775	3.438776	1.00	4.0000	7.0000	10.0000	12.0000
hr	17379.0	11.546752	6.914405	0.00	6.0000	12.0000	18.0000	23.0000
holiday	17379.0	0.028770	0.167165	0.00	0.0000	0.0000	0.0000	1.0000
weekday	17379.0	3.003683	2.005771	0.00	1.0000	3.0000	5.0000	6.0000
workingday	17379.0	0.682721	0.465431	0.00	0.0000	1.0000	1.0000	1.0000
weathersit	17379.0	1.425283	0.639357	1.00	1.0000	1.0000	2.0000	4.0000
temp	17379.0	0.496987	0.192556	0.02	0.3400	0.5000	0.6600	1.0000
atemp	17379.0	0.475775	0.171850	0.00	0.3333	0.4848	0.6212	1.0000
hum	17379.0	0.627229	0.192930	0.00	0.4800	0.6300	0.7800	1.0000
windspeed	17379.0	0.190098	0.122340	0.00	0.1045	0.1940	0.2537	0.8507
casual	17379.0	35.676218	49.305030	0.00	4.0000	17.0000	48.0000	367.0000
registered	17379.0	153.786869	151.357286	0.00	34.0000	115.0000	220.0000	886.0000
cnt	17379.0	189.463088	181.387599	1.00	40.0000	142.0000	281.0000	977.0000

Usamos la función Describe para llevar a cabo un análisis de los valores de cada atributo, obteniendo información sobre la distribución de los datos.
Sabemos que, cuando la media prácticamente coincide con la mitad del valor mínimo y el máximo, si además el percentil 50% también es prácticamente similar a la media, indica que la dispersión de los datos es pequeña y la mitad de los datos están en la parte izquierda del conjunto y la otra mitad en la parte derecha.
Siguiendo esto, podemos concluir que la variable "cnt", que será la variable elegida como target, no está uniformemente distribuida en el rango de valores que toma, ya que hay más valores en la parte baja del rango que hace que su media y percentil 50% sean más bajos de lo que podría ser. Además el valor máximo que toma está muy alejado de estos valores, lo que sugiere que puede haber valores que podríamos considerar como outliers.

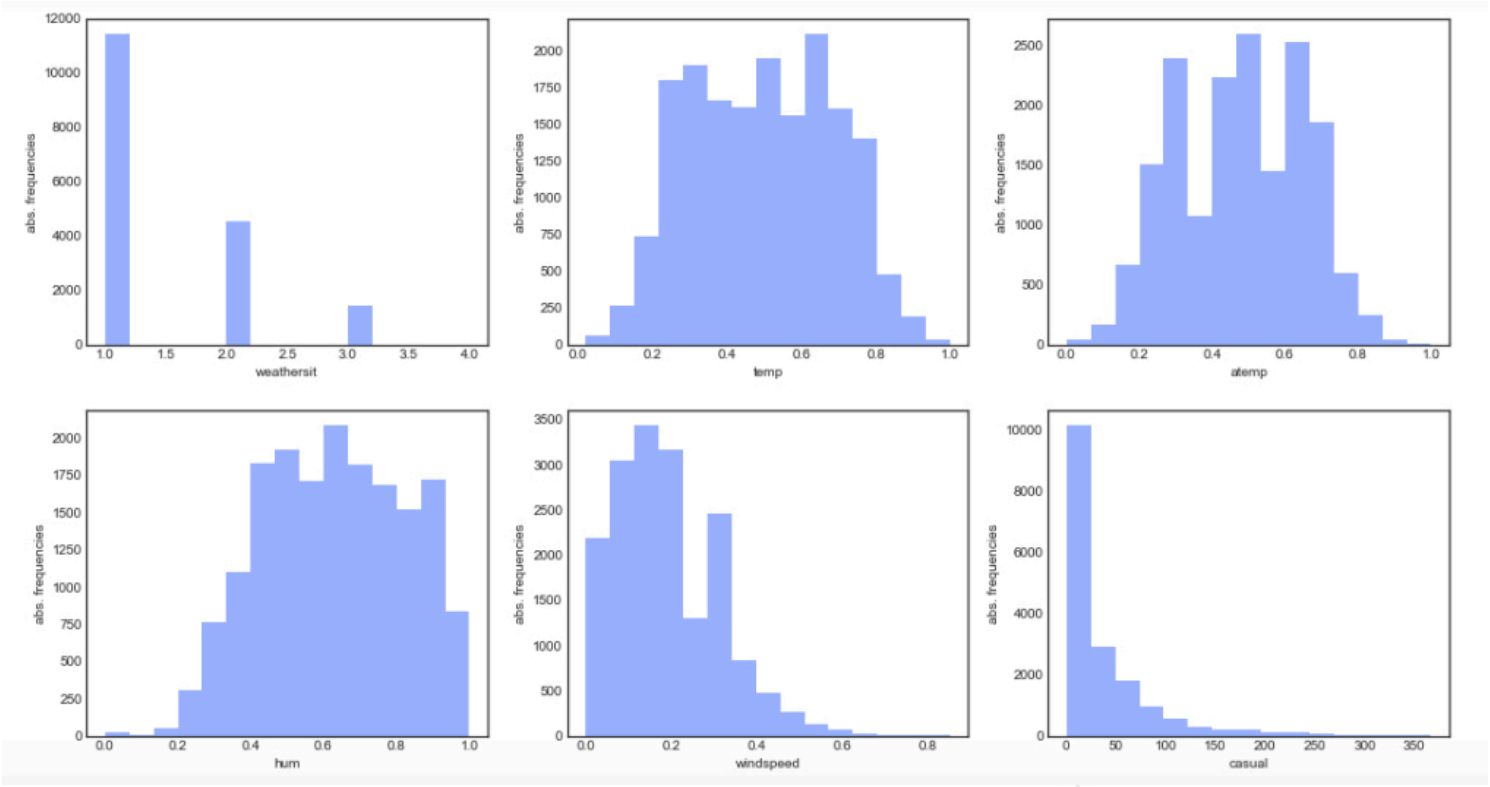
Con este primer análisis introductorio no podemos inferir mucha información útil de las variables categóricas a traves de la información que conseguimos con la función Describe, pero por ejemplo si se puede apreciar en la variable "weathersit", que indica el tipo de día en base al tiempo que realizaba ese día, que a lo largo del año la mayor parte de días fueron de tipo 1, que indica que son días con buen tiempo, ya que la mediana que deja por debajo al 50% de los datos del conjunto es 1.

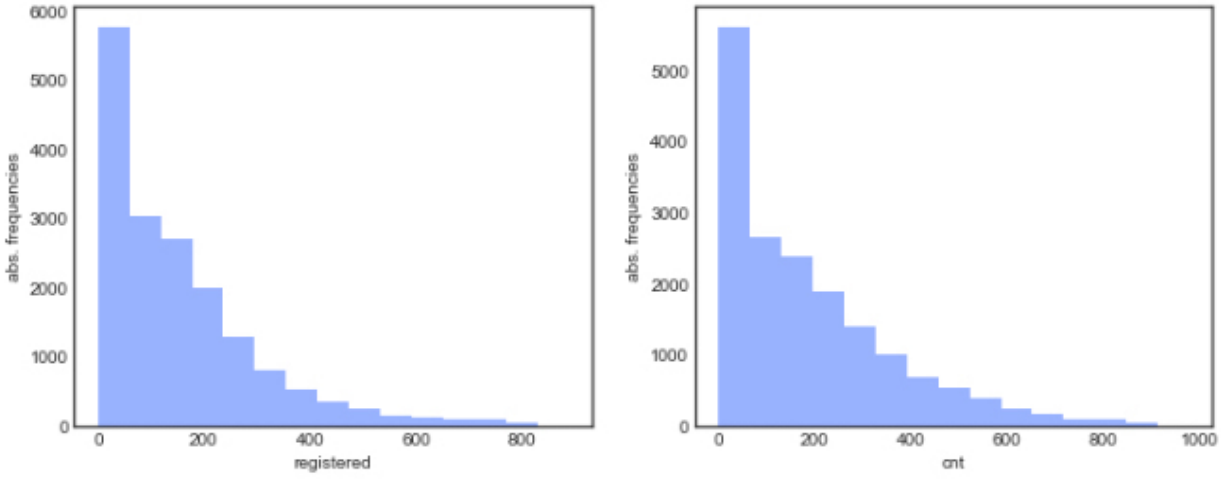
Por otro lado, de las variables numéricas si se pueden sacar más conclusiones, gracias a la información obtenido con la función Describe. Por ejemplo podemos observar como las variables "temp" y "atemp" que representan la temperatura y temperatura ambiental, parecen estar muy relacionadas entre sí y ya nos puede indicar una correlación antes de estudiarlo en otro apartado.

Para finalizar también podemos ver como la variable "cnt" tiene un valor mínimo de 1, lo que indicaría que todos los días del año al menos se hizo una utilización del servicio.

Número de clases: La variable que pensamos que nos va a aportar más información como clase para separar los datos y estudiarlos en base a cada una, es la variable "year" que indica los datos del año 1 a estudiar, y los datos del año 2. Así pueden enfrentarse los datos y comparar la evolución de un año a otro, lo que podemos utilizar a la hora de hacer un estudio del negocio y la posterior mejora del servicio año tras año.

Hemos graficado los histogramas de algunas variables del dataset, aquellas en las que este tipo de gráfica podría dar cierta información, como las variables numéricas.





A través de los histogramas podemos apreciar la distribución de un conjunto de datos. Podemos analizar los picos, la simetría y asimetría del conjunto, o incluso datos atípicos u outliers.

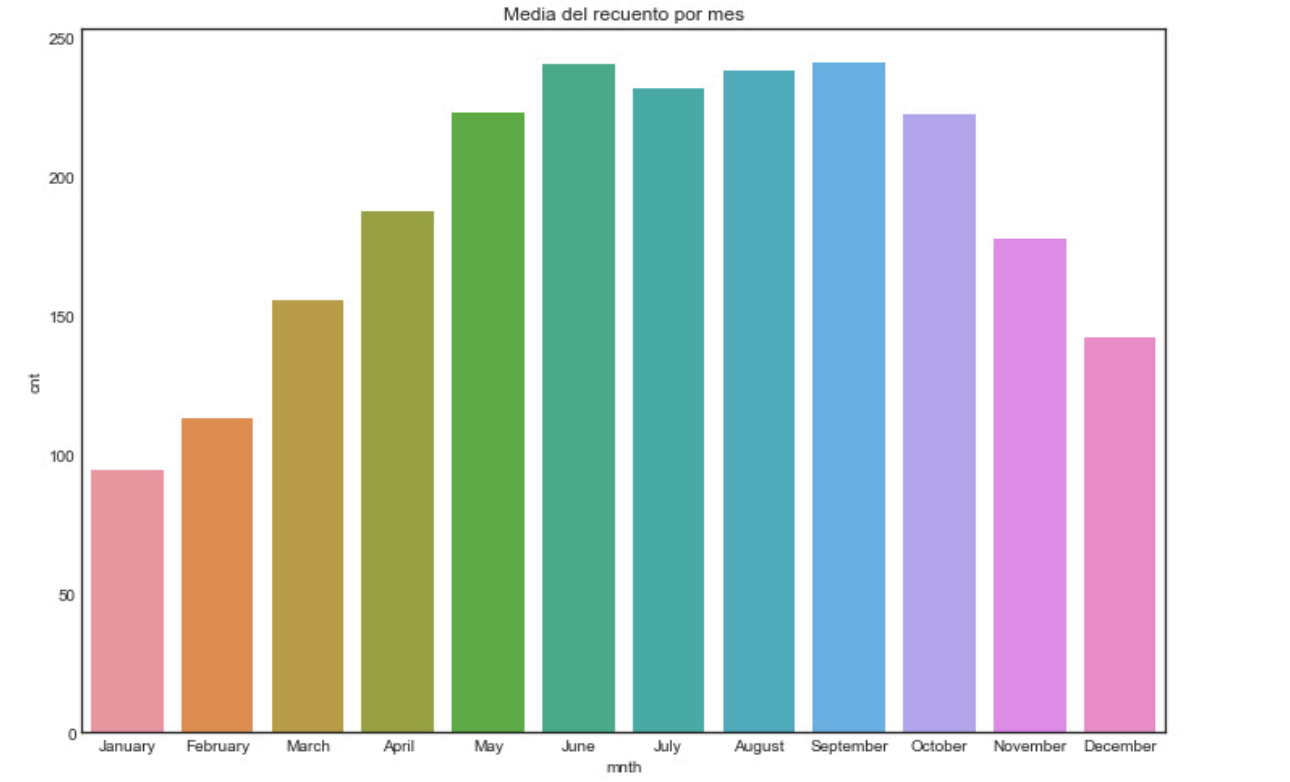
Hemos pintado los histogramas de las variables numéricas, y ahora podemos analizar mejor la distribución de sus datos respecto a lo visto con el primer análisis con Describe, incluso en alguno de los gráficos ya podemos ver la presencia de valores outliers, aunque estos valores los podremos analizar mejor posteriormente mediante los boxplots.

De nuevo vemos como las variables de "temp" y "atemp" parecen estar relacionadas, ya que sus histogramas siguen un mismo patrón o distribución de datos.
En la variable "hum", que indica la humedad, vemos algún valor que podría ser un valor atípico, y que corresponde cuando la humedad es prácticamente nula, lo que indicaría días totalmente secos en el ambiente, lo que según el histograma no sería algo habitual.
La variable que indica la velocidad del viento, "windspeed", tiene una distribución donde lo habitual son días con poco viento, y a medida que el viento crece decrece la frecuencia.
El histograma de la variable "weathersit" nos indica lo mismo que vimos con Describe. Los días más habituales a lo largo del año son aquellos claros y con buen tiempo, y decrece la frecuencia a medida que aumenta el grado de mal tiempo.

Los histogramas de las variables que recogen el recuento de usos diarios, siguen una distribución donde los picos y la mayor concentración de datos están en los valores pequeños, por tanto los días con mayor registro de usos son los menos habituales, al contrario que los días donde hubo un uso limitado del servicio.

Lo que nos interesa ahora, es analizar las distintas variables en relación al recuento final, para poder estudiar: en que época del año hay mayor uso, en que día de la semana, en que hora del día, bajo que condiciones climatológicas y de esta forma tratar de inferir cuando son los tramos de más uso de bicicletas en la zona de Washington DC.

Por ejemplo obtener la media del uso mensual del servicio, y poder ver los meses con mayor uso.



También sería valioso obtener las horas de mayor uso del servicio: analizando el uso mensual y diario del servicio, podemos observar como en el histograma de "Uso de servicio por meses" son los meses de Verano donde más se utiliza el servicio de renting, algo que puede tener lógica debido a que son los meses donde suele hacer mejor tiempo.

En el caso del Histograma de "Uso del servicio por horas", la acumulación de uso la apreciamos a primera hora, entre las 7 y las 9, que es el horario de entra de trabajo y a la tarde sobre las 17 y 18 lo que podría corresponder a la salida de trabajo de los usuarios.

2. Describe y realiza modificaciones en la base datos si lo consideras necesario. Por ejemplo, qué harías con valores nominales, si los hubiera.

Revisando el dataset vemos que las variables "season", "yr", "mnth" o "weathersit" son variables categóricas. Estas variables deberían tener valores nominales, pero ya vienen transformadas a valores númericos, donde cada valor está asociado a cada una de las categorías definidas para cada variable.
Por ejemplo, para la variable season los valores 1,2,3 y 4 han sustituido a los valores nominales springer, summer, fall, winter. (1:springer, 2:summer, 3:fall, 4:winter)

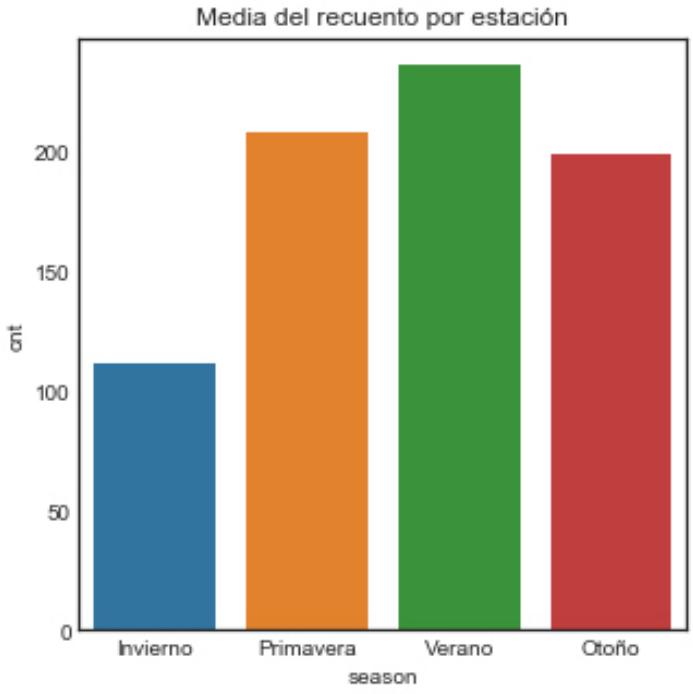
En caso de encontraros valores nominales en el dataset, realizaríamos el mimos proceso de transformación de asignarles un rango de valores numéricos asociados a sus correspondientes valores categóricos.

Con los valores nominales en caso de que los hubiera haríamos esto mismo, asignarles un rango de valores numéricos y asociarlo a los distintos valores origen.

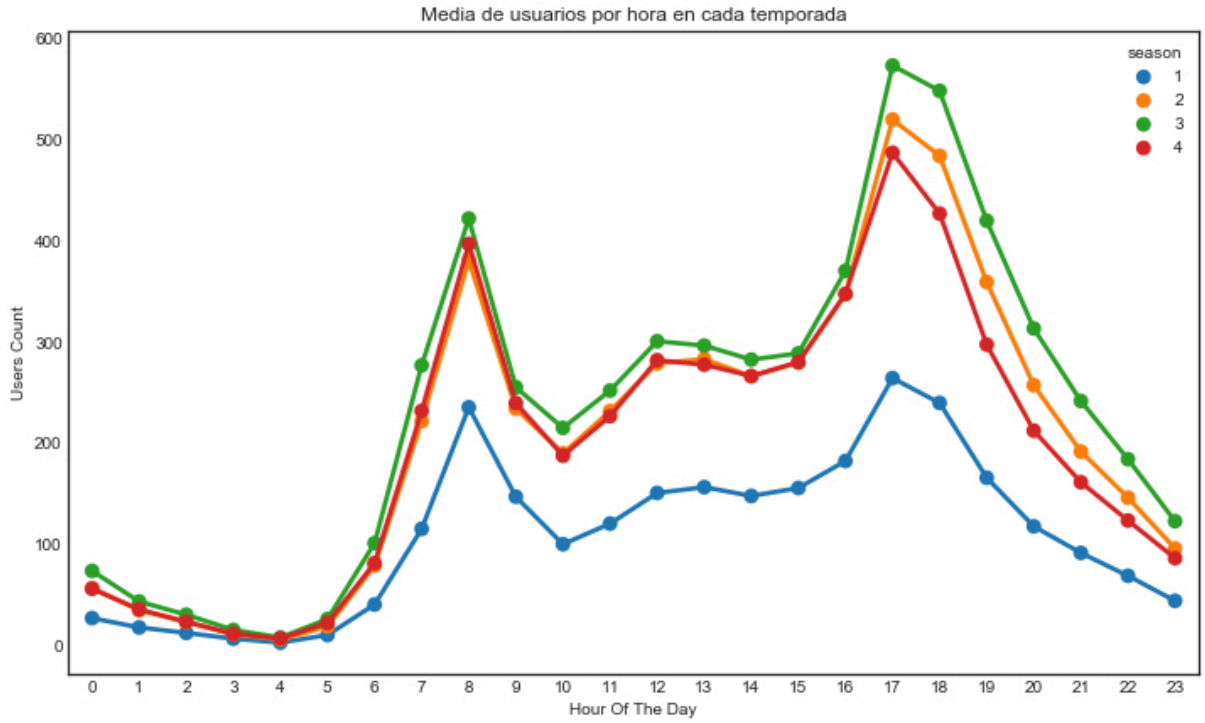
Hemos visto que en la definición de las variables del dataset se indica que los valores de la variable "season" son 1's, y nos hemos dado cuenta en un primer análisis exploratorio que el valor de esta variable no está correctamente relacionado con la fecha del mismo registro, por lo que estos valores están mal. Para solucionar esto podríamos optar o por eliminar dicha variable ya que aportaría información incorrecta o actualizarla con el valor de la season extraído de la fecha de cada uno de los registros, con la siguiente asignación en el orden correcto: (1: winter, 2: springer, 3: summer, 4: fall).

Ejemplo de modificación de los valores del atributo season, si quisiéramos cambiar los valores y adecuarlos a lo que dice la definición del atributo en la documentación del dataset: df_bike_hours.replace({'season': {1: '4', 2: '1', 3: '3', 4: '2'}}, inplace = True)

Una vez modificada correctamente la variable season, o simplemente teniendo en cuenta que el valor que viene no coincide con el que dice la definición del dataset, se pueden realizar análisis en base a ella, ya que parece interesante estudiar por temporadas el uso del servicio.



También podría hacerse el estudio de las horas donde mayor uso del servicio existe, contando también por temporada del año, y se obtendría el siguiente gráfico:



Hemos decidido eliminar del dataset, la primera variable que actúa como índice ya que consideramos que no nos aporta información de valor.

Otro caso de variables relacionadas son las variables "season" y "month". Depende del tipo de análisis que queramos realizar de la información, que podría interesarnos estudiar los datos por estación, lo que nos daría también información por extensión de los meses, o estudiar el uso del servicio mes a mes.

3. Estudia si es necesario normalizar los datos y cómo lo harías. Procede a modificar la base de datos (normalizar) si lo consideras necesario.

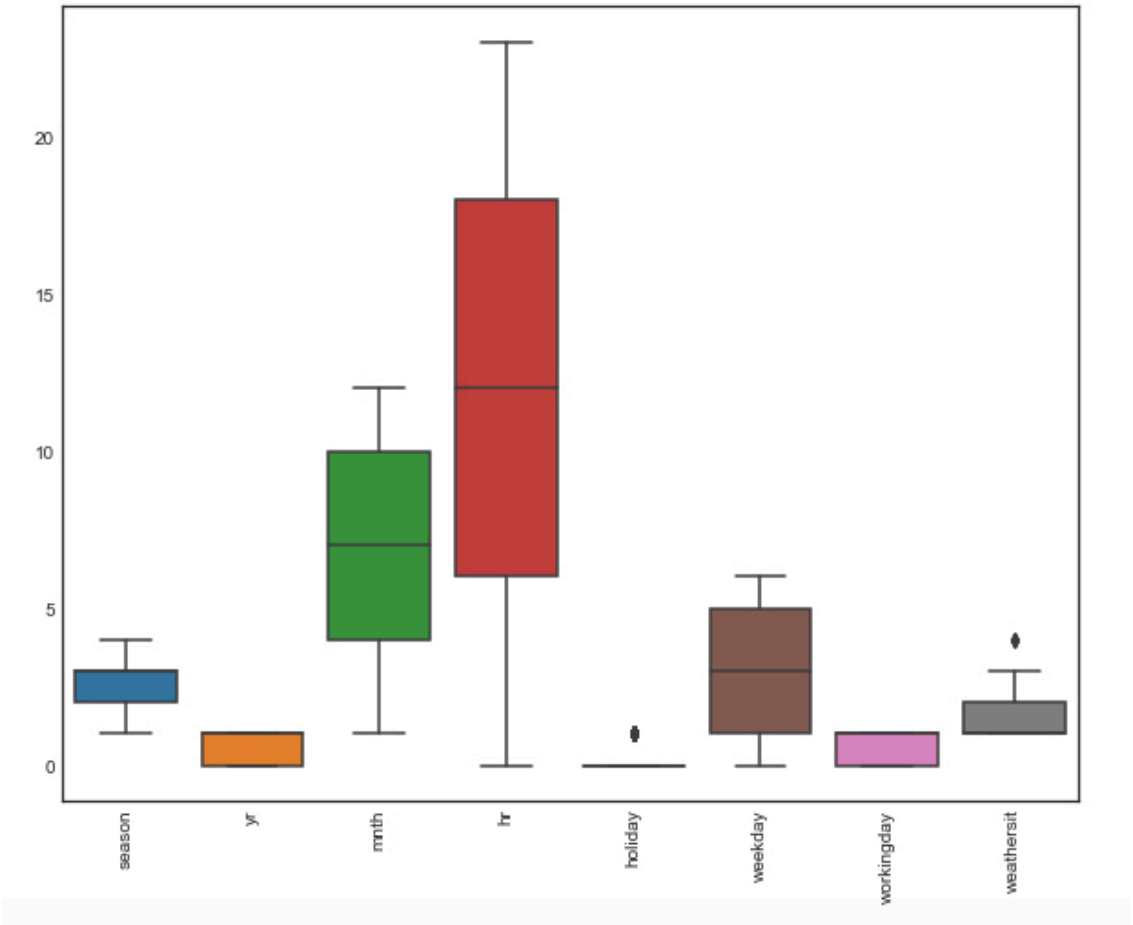
Consideramos que no es necesario realizar ningún proceso de normalización sobre las variables. Las variables que miden fenómenos ambientales ya vienen normalizadas y tienen una escala similar, con lo que pueden estudiarse en conjunto.
Por otro lado tenemos las variables que hacen el recuento de bicicletas por hora, según sean registradas o no, las cuales también están en la misma escala.
El resto de variables serían categóricas, guardan la misma escala entre ellas y pueden relacionarse de algún otro modo si fuera necesario.
Así que podemos concluir que las variables del dataset no parece que necesiten una normalización extra, pero si fuera necesario llevar a cabo un futuro estudio predictivo mediante modelización y se necesitará la normalización de alguna/s variable/s para que no haya variables que pudieran tomar un peso superior a otras en el modelo, se llevaría a cabo.

Ejemplo de normalización en base a la media de las variables y la desviación que es una de las formas más habituales de normalizar un dataset:
df_normalizado = (df - df.mean()) / df.std()

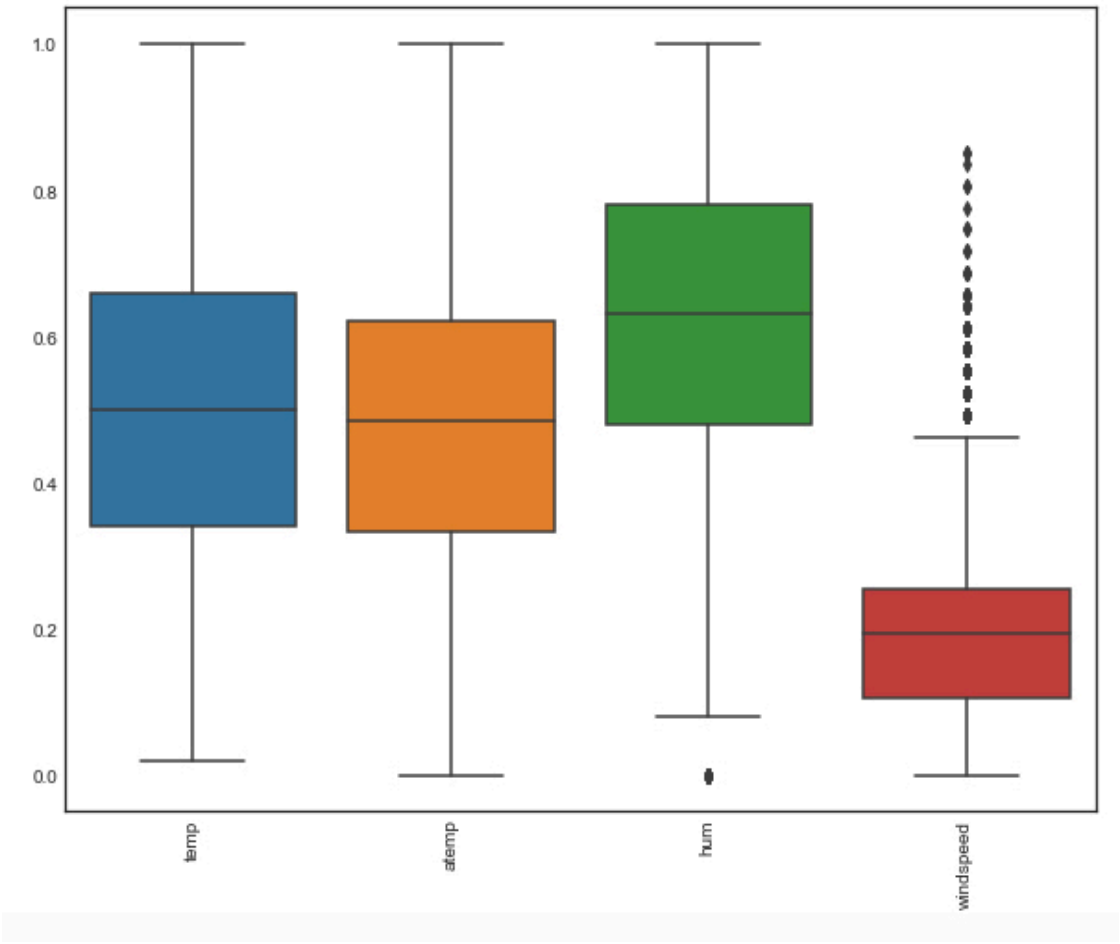
4. Detección de valores extremos (outliers) y descripción de qué harías en cada caso.

Mostramos una gráfica con los boxplots de cada variable del dataframe horario. Este es un tipo de gráfica donde puede apreciarse fácilmente la dispersión de los datos o los valores outliers. Dividimos los gráficos boxplots por variables categóricas, cuantitativas y las de recuento.

Boxplots de las variables categóricas:



Boxplot de las variables numéricas:



Los boxplots son gráficos donde se aprecia fácilmente la distribución de los datos y los outliers.
La caja que representa el boxplot, está delimitada por los percentiles 25 y 75, que dejan al 25% y 75% de los valores por debajo, por tanto, todos los datos que entren en los límites de la caja representarían el 50% de los datos y la línea interior de la caja representa el percentil 50 de la distribución o mediana.
Otra característica del boxplot son los bigotes inferior y superior. Estos bigotes representan la distribución de los datos restantes hasta llegar a los límites del gráfico tanto inferior como superior. Los datos que estén fuera de los límites son los considerados outliers.

En los boxplots de las variables de recuento ("casual", "registered" y "cnt") se aprecia como la mayor parte de valores están en el rango más bajo de cada atributo, y los valores más altos son considerados outliers. De aquí podemos extraer que lo menos habitual son días donde el uso de bicicletas sea muy elevado.
En el caso del boxplot de la variable que mide el viento, "windspeed", tiene muchos valores outliers ya que es poco habitual que haya días con mucho viento.
Los boxplots de las variables de temperatura vuelven a indicarnos su alta correlación, son prácticamente iguales, y son muy centrados y con una distribución uniforme, lo que sugiere que lo más habitual a lo largo del año son valores medios de temperatura.

Las variables categóricas arrojan boxplots más singulares debido a la naturaleza de estas variables. Algunas no tiene ni siquiera sentido analizarlas con este tipo de gráficos, por ejemplo la variable "yr" se representa con 0 o 1, según sean del primer año o del segundo. Como son un 50% de valores 0 y otro 50% un 1, el boxplot que sale no tiene ni bigotes.

Por último el boxplot de los días festivos. Esta variable también toma valores 0 o 1, y el boxplot, nos indica que todos los valores 1 son outliers, ya que son muy pocos debido a que la mayoría de días del año son no festivos.

Los valores outliers son datos que están lo suficientemente alejados de la mayoría de valores, de la media y mediana del conjunto, como para ser considerados valores extraños, ya sea porque son muy poco habituales o porque son datos erróneos.
Dependiendo de lo que se esté representando y de lo que se quiera analizar, estos valores deberán o no tenerse en cuenta para el análisis de esos atributos.

Un ejemplo de como eliminar outliers sería la que mostramos a continuación, eliminando del dataframe con los datos aquellos valores más alejados en base a la utilización de la media y la desviación:

```
# Una forma de eliminar outliers sería por ejemplo la siguiente, que eliminaría los datos que estén muy alejados
df_WithoutOutliers = df_bike_hours[np.abs(df_bike_hours["cnt"]-df_bike_hours["cnt"].mean()) \
                                   <=(3*df_bike_hours["cnt"].std())]

print("Before outer analysis:{}".format(df_bike_hours.shape))
print("After outer analysis:{}".format(df_WithoutOutliers.shape))
```

```
Before outer analysis:(17379, 16)
After outer analysis:(17135, 16)
```

5. Detección de valores perdidos (missing values) y descripción de cómo actuarías para solventar el problema.

Para ver si hay datos vacíos o nulos con el uso de las siguientes funciones podemos ver como el dataframe no tiene valores nulos o NAN, todas las variables tienen datos no nulos para cada registro:
df_bike_hours.notnull().all()

La no presencia de Missing Values o NaN, nos hace pensar que tenemos un juego de datos de calidad, es decir en buen estado.
Los valores faltantes o missing pueden tener una influencia significativa en nuestro análisis y futuro modelo predictivo, por lo que siempre es una decisión importante determinar la forma en que los vamos a manejar.

Las alternativas que tenemos para manejarlos son:

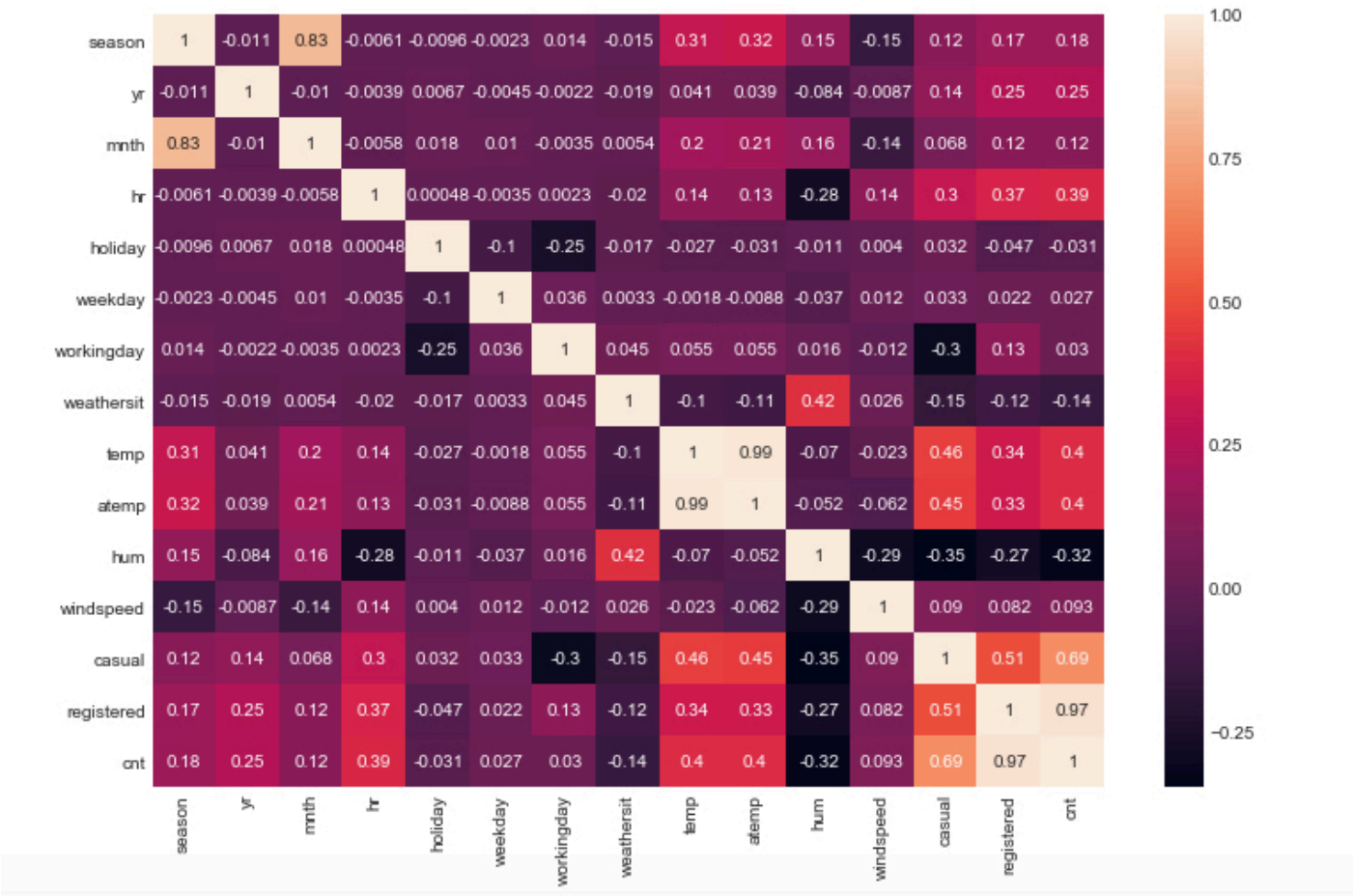
1. Dejarlos como están, lo que a la larga nos va a traer bastantes dolores de cabeza ya que en general los algoritmos no los suelen procesar correctamente y provocan errores.
2. Eliminarlos, lo que es una alternativa viable, aunque dependiendo de la cantidad de valores nulos, puede afectar significativamente el resultado final de nuestro modelo predictivo.
3. Inferir su valor. En este caso, lo que podemos hacer es tratar de inferir el valor faltante y reemplazarlo por el valor inferido. Esta suele ser generalmente la mejor alternativa a seguir.

Nosotros utilizaríamos la última alternativa, inferiremos los valores faltantes, usando la media aritmética para los datos cuantitativos, y la moda para los datos categóricos.

6. Buscar correlaciones entre:

a. **las variables predictoras, lo que permitirá ver si hay variables redundantes.**
b. **variables predictoras y la clase (target).**

Graficamos la matriz de correlación con el valor en cada celda correspondiente a la correlación entre cada par de variables, y un mapa de calor con un gradiente de color en consecuencia al valor pintado.



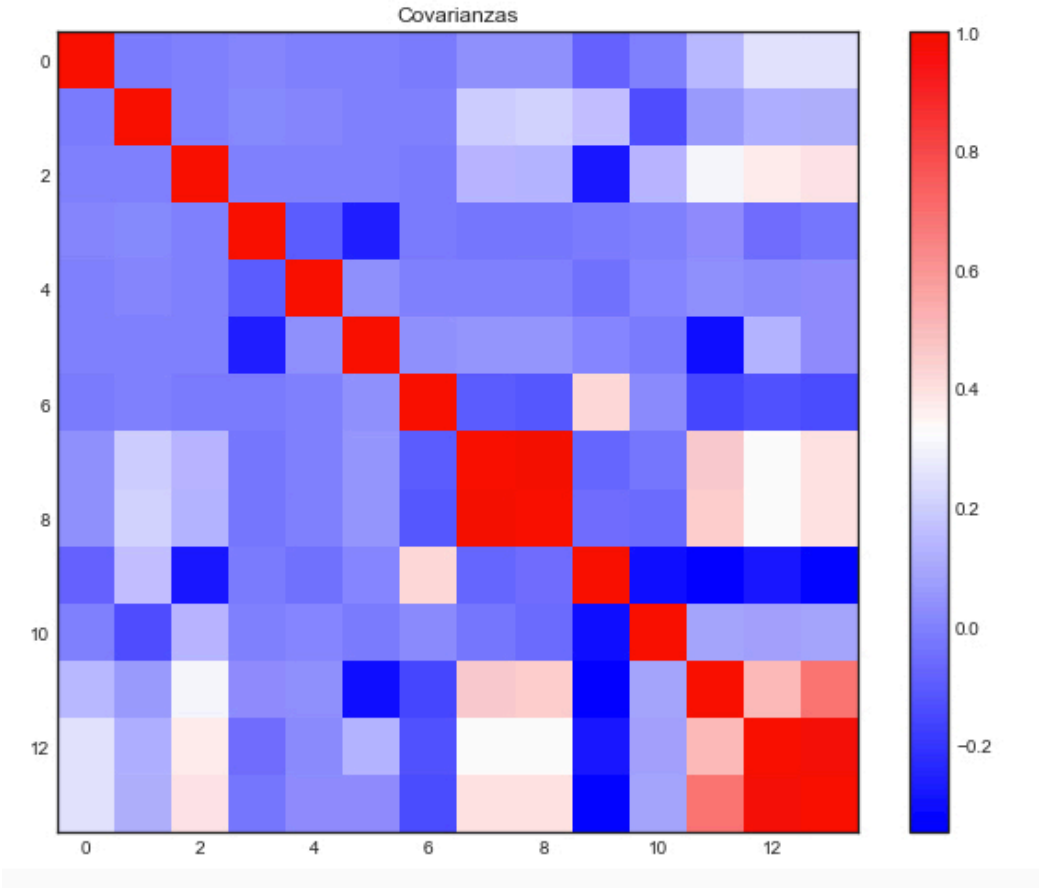
El gráfico de correlación representa la matriz de correlación entre cada par de variables del conjunto de datos. Cada cuadrado en el eje "x" y eje "y" representa a una de las variables del conjunto de datos.
La diagonal de la gráfica será la correlación total, valor 1, ya que relaciona a cada variable consigo misma.
Al lado del gráfico está la leyenda de colores, lo que nos permite según el color de cada cuadrado ver el grado de relación entre cada par de variables.
Dos variables están correlacionadas cuando el valor es cercano a 1 o a -1, correlación positiva o negativa.
Se aprecia alta relación positiva entre las variables "temp" y "atemp", y las variables "registered" y "cnt".

- La variable "temp" indica la temperatura, y "atemp" la sensación de temperatura, y es por esa razón por la cual son muy similares y están correlacionadas.
- Por otro lado, la variable "registered" indica el recuento de bicicletas registradas, y la variable "cnt" el recuento de bicicletas registradas y casuales. Como las casuales suelen ser un número bastante menor que las registradas, al final hay poca diferencia en el recuento total (variable "cnt") y el recuento de registradas.

Por tanto, para el análisis se podría prescindir de una variable de cada uno de estos dos pares de variables altamente correlacionadas, ya que tienen información muy similar.

Ahora mostramos la gráfica de covarianza de los datos. No cogemos la primera variable que es la fecha, ya que daría error al realizar las operaciones.

Para apreciar bien en la gráfica de covarianza los valores entre las variables, normalizamos todo el conjunto de datos ya que sino no se podría apreciar visualmente de forma satisfactoria.
El gráfico de covarianzas representa la matriz de covarianza entre cada par de variables del conjunto de datos. La diagonal de la gráfica representa la varianza de cada propiedad que al normalizar es igual a 1.

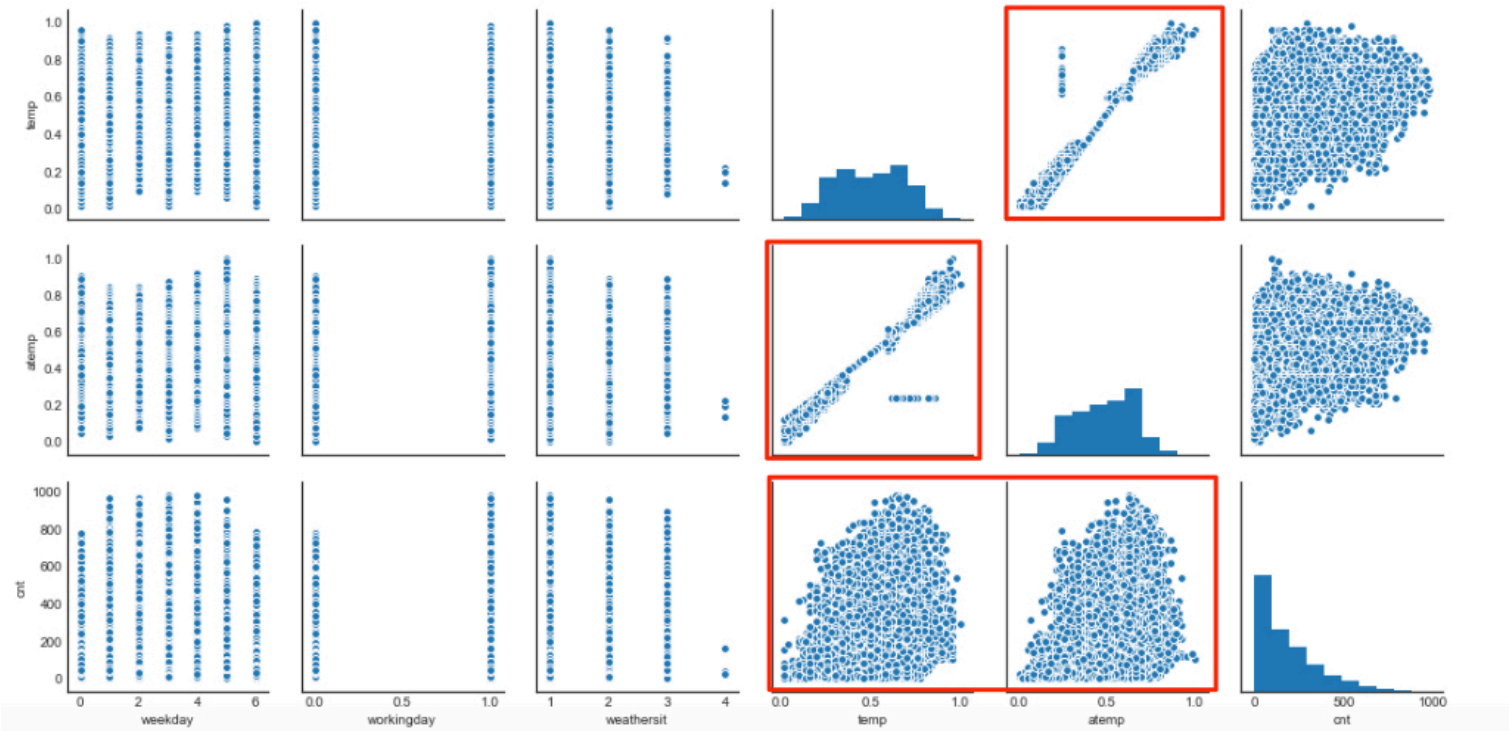


La covarianza positiva entre dos variables indica una relación directa, y que cuando una variable crece la otra variable también lo hace. Y una covarianza negativa indica una relación inversa, que cuando una variable crece la otra variable decrece.
Entonces se puede ver de nuevo como los pares de variables que guardaban mayor relación anteriormente con las correlaciones vuelven a hacerlo ahora con las covarianzas. Estos pares de variables relacionados tienen valor positivo que indica relación directa.
Podemos ver como hay una alta relación positiva entre las variables "temp" y "atemp", y las variables "registered" y "cnt".

También podemos utilizar las gráficas de matriz de dispersión para estudiar la relación entre las variables del dataset.
Seleccionamos la variable target y el resto de variables predictoras para enfrentarlas y ver posibles relaciones con matrices de dispersión. La variable target que será "cnt", y sería la variable que se quisiera poder predecir, y las otras variables quitando la primera que sería la fecha serían las variables predictoras:

```
vars_bike_hour   = df_bike_hours.columns[1:15]
target_bike_hour = df_bike_hours.columns[-1]
```

Se pinta la matriz de dispersión entre las variables del dataframe, en la diagonal los histogramas de cada variable:
Se señalan las relaciones altas que se han visto en las gráficas, que serían entre las variables de temp y atemp, y de registered y cnt.



Con las matrices de dispersión también podemos ver la distribución de los datos y la relación entre cada par de variables. En la diagonal de la matriz se pueden poner distintos tipos de gráficas, en este caso hemos dibujado los histogramas de cada variable. Los histogramas son más similares para los pares de variables que están correlacionadas como hemos visto anteriormente.
Donde mejor se ve la relación entre variables, es en la parte de la matriz de dispersión que las relaciona, donde las variables que tienen alta relación forman prácticamente una recta. Si la recta es creciente significa que cuando aumenta el valor de la variable "x", también aumenta el valor de la variable "y". Al contrario si la recta es decreciente.
En los diagramas de dispersión vuelve a apreciarse la relación lineal entre los pares de variables que se dijo en secciones precias. Vemos como los datos enfrentados de cada variable acaban formando prácticamente una linea recta. En las variables "atemp" y "temp" se pueden ver algunos datos que se separan de la gran mayoría, que forma una linea recta, y que posiblemente sean datos que corresponden a días donde hubo una diferencia no habitual entre la sensación de temperatura y la real.

7. Detecta, si hubiera, falsos predictores.

Puede haber variables predictoras que podrían llevar a análisis erróneos, o al menos sesgados, ya sea porque tienen directamente información falsa o porque no tienen en cuenta algún punto.

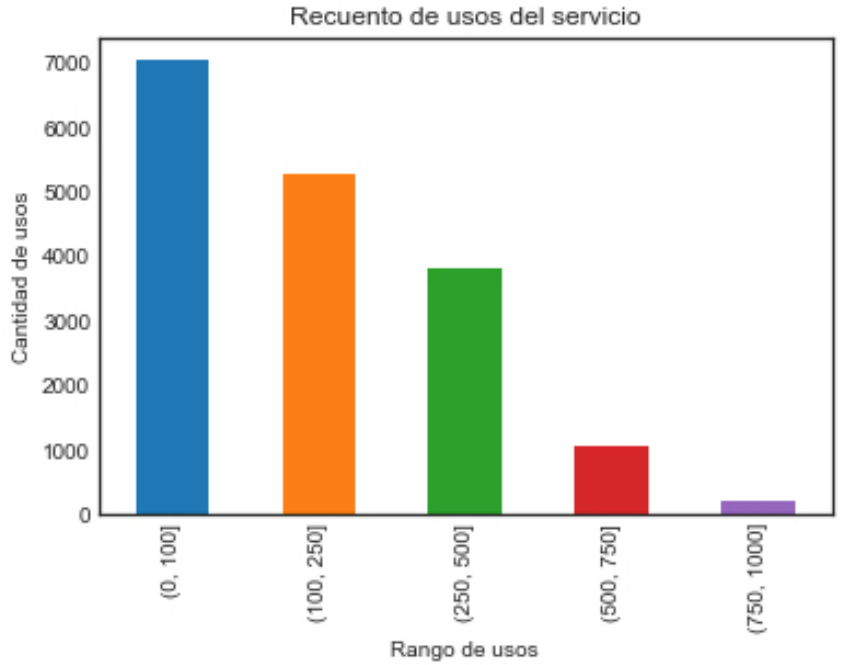
Pensamos por ejemplo, que estudiar y sacar las conclusiones sólo y exclusivamente a través de la variable "season", la cual indica la temporada del año de cada registro, podría ser un falso predictor, ya que no todos los días de cada estación hace el mismo tiempo, aunque si es más habitual que sigan un mismo patrón, pero podría no tenerse en cuenta los días con condiciones climatológicas distintas.

Por otro lado podríamos crear una nueva variable, en base a las variables de "season" y "weathersit", uniendo así la información de la estación del año y las condiciones climatológicas de cada registro, lo que nos aportaría una información interesante para el resultado del estudio.

8. Estudia si fuera conveniente segmentar alguna de las variables.

Una variable que podríamos segmentar sería la variable target, cnt, para agrupar en rangos el recuento de usos del servicio y estudiar los datos en base a estos rangos que podrían definirse según el estudio de negocio que quiera realizarse para mejorar y aumentar el servicio.

Podríamos listar por cada rango el número de recuento, y además mostrarlo con una gráfica de barras.



9. Estudia si fuera conveniente crear nuevas variables sintéticas basadas en las variables originales.

Como comentamos en el apartado de los falsos predictores, una opción de una nueva variable a utilizar para analizar los datos, sería unir los datos de las variables "season" y "weathersit", para así de esta forma poder estudiar los datos por grupos de estación-condicion_climatológica, y ser capaces de ver si los días con buenas o malas condiciones se comportan igual en diferentes estaciones del año.

Otra variable que vemos que podría ser necesaria, es la variable día. Imaginemos que queremos hacer un estudio a nivel de día del funcionamiento del servicio de rentíng en los meses de agosto de ambos años, para saber su comportamiento en este periodo vacacional concreto.
Para poder hacer esto simplemente tenemos que extraer el día de la fecha, para así poder usarlo en estudios posteriores, o por ejemplo ver si hay alguna relación entre los últimos días del mes en los cuales la gente cobra, y el uso del servicio al tener más dinero puede bajar al ser más propensos a usar otros medios de locomoción más caros.

	dteday	season	yr	mnth	hr	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt	season_weather	day
0	2011-01-01	1	0	1	0	0	6	0	1	0.24	0.2879	0.81	0.0000	3	13	16	1	1
1	2011-01-01	1	0	1	1	0	6	0	1	0.22	0.2727	0.80	0.0000	8	32	40	1	1
2	2011-01-01	1	0	1	2	0	6	0	1	0.22	0.2727	0.80	0.0000	5	27	32	1	1
3	2011-01-01	1	0	1	3	0	6	0	1	0.24	0.2879	0.75	0.0000	3	10	13	1	1
4	2011-01-01	1	0	1	4	0	6	0	1	0.24	0.2879	0.75	0.0000	0	1	1	1	1
5	2011-01-01	1	0	1	5	0	6	0	2	0.24	0.2576	0.75	0.0896	0	1	1	2	1
6	2011-01-01	1	0	1	6	0	6	0	1	0.22	0.2727	0.80	0.0000	2	0	2	1	1
7	2011-01-01	1	0	1	7	0	6	0	1	0.20	0.2576	0.86	0.0000	1	2	3	1	1
8	2011-01-01	1	0	1	8	0	6	0	1	0.24	0.2879	0.75	0.0000	1	7	8	1	1
9	2011-01-01	1	0	1	9	0	6	0	1	0.32	0.3485	0.76	0.0000	8	6	14	1	1

10. Conclusiones finales.

Terminar eliminando las variables que puedan quedar y hayamos visto que queremos quitar porque no nos aportan valor al análisis que queremos realizar. Y después exportar el dataframe resultante a un fichero que sería el de partida para el posterior análisis predictivo.

Eliminamos la variable "dtday", pues tenemos esa misma información en las variables "yr", "mnth" y “day", y de esta forma tenemos más granuralidad pudiendo hacer análisis más profundos. La variable "atemp" por tener una fuerte correlación con “temp", y "registered" por su alta correlación con "cnt".

Un vistazo a los primeros registros del dataset resultante:

	season	yr	mnth	hr	holiday	weekday	workingday	weathersit	temp	hum	windspeed	casual	cnt	season_weather	day
0	1	0	1	0	0	6	0	1	0.24	0.81	0.0000	3	16	1	1
1	1	0	1	1	0	6	0	1	0.22	0.80	0.0000	8	40	1	1
2	1	0	1	2	0	6	0	1	0.22	0.80	0.0000	5	32	1	1
3	1	0	1	3	0	6	0	1	0.24	0.75	0.0000	3	13	1	1
4	1	0	1	4	0	6	0	1	0.24	0.75	0.0000	0	1	1	1
5	1	0	1	5	0	6	0	2	0.24	0.75	0.0896	0	1	2	1
6	1	0	1	6	0	6	0	1	0.22	0.80	0.0000	2	2	1	1
7	1	0	1	7	0	6	0	1	0.20	0.86	0.0000	1	3	1	1
8	1	0	1	8	0	6	0	1	0.24	0.75	0.0000	1	8	1	1
9	1	0	1	9	0	6	0	1	0.32	0.76	0.0000	8	14	1	1