Máster en Data Science

Trabajo Fin de Máster

PREDICCIÓN DE OCUPACIÓN DE PARQUÍMETROS SEGÚN MODELOS PREDICTIVOS ESPACIO-TEMPORALES

Alumnos:

Eva Carbón
Emilio Delgado
Cintia García
Paloma Panadero
Pedro Sánchez

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Análisis Exploratorio de los Datos (EDA)

Este capítulo realiza un análisis exploratorio (EDA) del *dataset* construido en el capítulo anterior (poner referencia). El objetivo es estudiar el *dataset*, para encontrar sus características más relevantes y describir su estructura. El dataset tiene estructura de serie espacio-temporal, por lo que se realizará el análisis en varios niveles:

1. Análisis descriptivo estático, donde se estudian las covariables (variables que no son coordenadas espaciales o temporales y pueden describir o predecir el resultado), estableciendo relaciones entre ellas y extrayendo conclusiones con impacto en capítulos posteriores.

2. Análisis descriptivo dinámico, donde se analiza la estructura temporal y espacial de los datos, describiendo sus principales parámetros y características.

El conjunto de datos bajo análisis consta de 8 columnas y 4113518 observaciones. En la Figura 1.1 se muestra una captura de los 10 primeros registros. El dataset consta de dos columnas con coordenadas espaciales (latitud y longitud) y una columna con coordenadas temporales (timestamp). Además, consta de la variable a predecir (porcentaje de ocupación), y las covariables (temperatura máxima, temperatura mínima y precipitaciones).

1.1. Análisis descriptivo estático

La fuente de datos bajo estudio presenta las siguientes covariables:

- Temperatura máxima
- Temperatura mínima
- Probabilidad de precipitación

	element_key	latitude	longitude	timestamp	occupation_perc	prcp	tmax	tmin
0	35693	47.619158	-122.346457	2016-01-02 00:00:00	28.57	0.0	42	25
1	53549	47.628175	-122.341132	2016-01-02 00:00:00	3.12	0.0	42	25
2	11881	47.619156	-122.333107	2016-01-02 02:00:00	10.00	0.0	42	25
3	9393	47.621441	-122.335970	2016-01-02 03:00:00	20.00	0.0	42	25
4	11133	47.619815	-122.348131	2016-01-02 04:00:00	20.00	0.0	42	25
5	31310	47.619256	-122.339661	2016-01-02 04:00:00	9.09	0.0	42	25
6	13130	47.620816	-122.345711	2016-01-02 04:00:00	22.22	0.0	42	25
7	53126	47.616374	-122.341452	2016-01-02 04:00:00	9.09	0.0	42	25
8	36142	47.617287	-122.338056	2016-01-02 05:00:00	16.67	0.0	42	25
9	76433	47.622804	-122.339860	2016-01-02 05:00:00	9.09	0.0	42	25

Figura 1.1: Extracto de las 10 primeras muestras del dataset

1.1.1. Temperatura máxima

La temperatura máxima registrada durante los días en los que se ha producido una transacción sigue la distribución que se muestra en la Figura 1.2.

A lo largo de los días que recoge el dataset, la temperatura máxima media es de 62.93 ^{o}F, mientras que su desviación típica es 12.68 ^{o}F. En la Tabla 1.1 se presentan los intervalos de confianza para media y varianza de la temperatura máxima de la muestra ($\alpha = 0{,}05$).

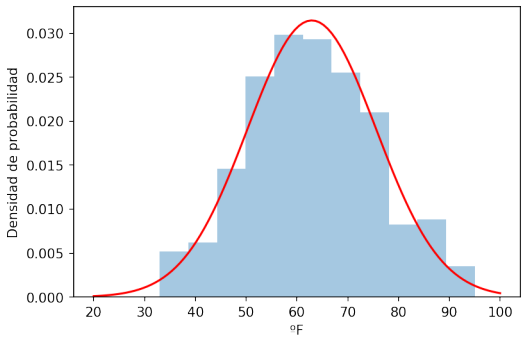

Figura 1.2: Distribución de temperaturas máximas

Como puede observarse, el intervalo de confianza para ambas medidas es muy estrecho, debiéndose al tamaño de la muestra. Esto mismo puede observarse en la Figura 1.2, que es aproximadamente una distribución normal centrada en la media y con desviación típica la de la muestra.

Parámetro	2.5 %	97.5 %
Media (°F)	62.91	62.94
Desv. Est. (°F)	12.67	12.69

Tabla 1.1: Intervalo de confianza para la temperatura máxima

1.1.2. Temperatura mínima

Las mismas conclusiones que se han presentado sobre la temperatura máxima pueden realizarse sobre la temperatura mínima. La temperatura media mínima es 47.76 ^{o}F, mientras que su desviación típica es 8.15 ^{o}F. En la Figura 1.3 se muestra la distribución estadística y en la Tabla 1.2 los intervalos de confianza para cada parámetro, calculados para $\alpha = 0{,}05$.

Parámetro	2.5 %	97.5 %
Media (°F)	47.75	47.77
Desv. Est. (°F)	8.14	8.16

Tabla 1.2: Intervalo de confianza para la temperatura mínima

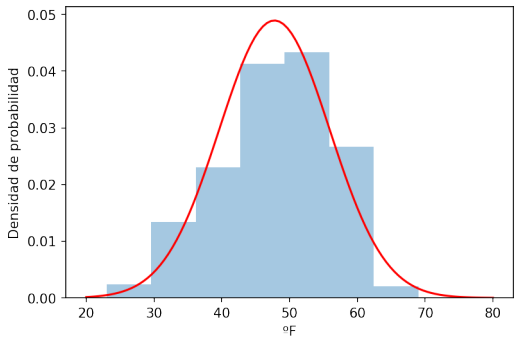

Figura 1.3: Distribución de temperaturas mínimas

De nuevo, debido al tamaño de la muestra, los valores de media y desviación típica calculados son muy precisos. Además, también puede asumirse que la distribución es normal.

La normalidad tanto de temperatura máxima como de temperatura mínima puede ayudar con el desarrollo de los modelos predictivos posteriores, debido a que muchas veces exigen normalidad en las variables que se utilizan para la predicción.

1.1.3. Precipitaciones

La distribución de precipitaciones se muestra en la Figura 1.4, y los intervalos de confianza para media y desviación típica, en 1.3. Presenta una media de 0.13 y una desviación típica muestral de 0.26. Al igual que en secciones anteriores, los intervalos de confianza dan estrechos, como corresponde a una muestra de un gran número de datos. Sin embargo, la distribución en este caso no es gaussiana, principalmente debido a no ser simétrica. Este parámetro tiene mucha dispersión, pues su coeficiente de variación (C_V), calculado como

$$C_V = \frac{\sigma}{\bar{x}} \approx \frac{s}{\bar{x}},$$

da como resultado $C_V = 2$. En porcentaje, el coeficiente de variación es del 200 %, lo que implica que estamos ante una característica con gran variabilidad.

Parámetro	2.5 %	97.5 %
Media	0.126	0.127
Desv. Est.	0.254	0.255

Tabla 1.3: Intervalo de confianza para las precipitaciones

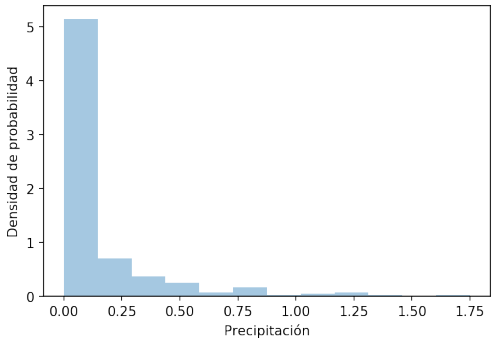

Figura 1.4: Distribución de precipitaciones

1.1.4. Porcentaje de ocupación

Finalmente, se presenta la distribución del porcentaje de ocupación de los parquímetros, tanto gráficamente (Figura 1.5) como con los intervalos de confianza para un 95 % de significación (1.4). La media de ocupación es de 50.81 %, mientras que la desviación típica es 40.08 %.

Parámetro	2.5 %	97.5 %
Media (%)	50.77	50.84
Desv. Est. (ºF)	40.05	40.1

Tabla 1.4: Intervalo de confianza para la temperatura mínima

La distribución no es gaussiana, pues se encuentra acotada inferiormente por cero. Además, está muy polarizada hacia los valores centrales. Del mismo modo que en las variables anteriores, se puede comprobar que los intervalos de confianza son muy estrechos debido al gran número de muestras de que consta el dataset.

Por otro lado, tiene una varianza considerable, especialmente en relación con la media, como consecuencia de la variabilidad de la disponibilidad de plazas de aparcamiento. Sin embargo, la variabilidad de esta característica no es tan grande como en otros casos: su coeficiente de variación es del 80 %.

1.1.5. Análisis de correlaciones entre las covariables y el target

A continuación, se presenta un analisis de correlaciones entre las covariables y el target, que permitirá disponer de información más precisa y detallada sobre el dataset. Además, debido a que

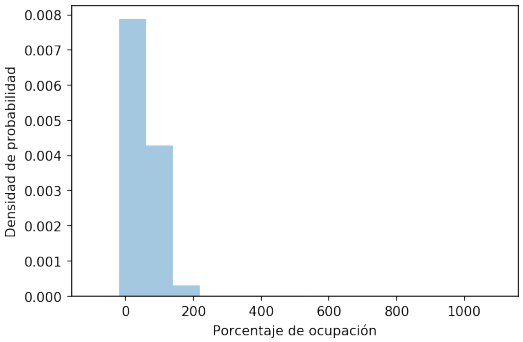

Figura 1.5: Porcentaje de ocupación general de los parquímetros

muchos modelos espacio-temporales constan de una parte regresiva, se podrá analizar con mejor precisión el resultado que se obtenga.

Para el análisis de correlaciones, dado que todas las variables son continuas, se utiliza el coeficiente de correlación de Pearson, definido como

$$r_{X_1X_2} = \frac{E[(X_1 - \bar{x}_1)(X_2 - \bar{x}_2)]}{s_{X_1} s_{X_2}}.$$

La significación estadística de este valor se estudia mediante un test T (poner referencia), que determina si el valor calculado es significativamente distinto de cero.

Para ello, se calcula el estadístico T,

$$T = \frac{r}{\sqrt{1 - r^2}} \cdot \sqrt{N - 2},$$

que se distribuye según una t de Student de $N - 2$ grados de libertad. El p-valor se calcula de forma bilateral, utilizando las tablas de la t de Student, como la probabilidad de obtener un valor más extremo del estadístico T que se ha calculado con la muestra dada. Es decir,

$$p = \text{Prob}(|t| \geq |T|).$$

Establecemos el nivel de significación (α) en el 95 %, por lo que el p-valor deberá ser menor de 0.05 para que sea válido y el resultado tenga significación estadística.

Mantendremos la estructura original del test, tal y como está establecido en (poner referencia), pero debemos tener en cuenta que N es muy grande, y por lo tanto:

- La t de Student se podría aproximar por una distribución normal.
- Los resultados saldrán muy significativos, pues el valor de T será muy elevado, situándose muy a la derecha o muy a la izquierda de la distribución t, quedando muy lejos del valor crítico definido por $\alpha = 0{,}05$.

Los resultados se presentan en la tabla **??**, donde se muestra tanto el coeficiente de correlación de Pearson como el p-valor asociado a cada uno de ellos. Las conclusiones derivadas de esa tabla son extensibles a todas las variables, pues los números son muy parecidos. No hay evidencia de que

haya correlación entre las covariables y el target. Además, esta conclusión estadísticamente es bastante significativa, pues todos los p-valores calculados son menores que el intervalo de significación establecido ($p < 0{,}05$).

Correlación con el porcentaje de ocupación	r	p-valor
Temperatura máxima	0.004	1.59e-15
Temperatura mínima	0.006	4.22e-39
Precipitaciones	-0.006	4.82e-29

Tabla 1.5: Correlaciones entre las covariables y el target

Como se comentó anteriormente, todos los resultados son significativos porque la muestra es muy grande. Por eso, aunque los coeficientes de correlación están próximos a cero, son estadísticamente distintos de cero, lo que es lógico teniendo en cuenta el tamaño de la muestra.

La conclusión, por lo tanto, es que no hay evidencias de un grado de correlación alto entre las covariables y el target. Esto podría dificultar análisis de tipo regresivo, puesto que no hay relaciones lineales directas entre las variables presentadas y el porcentaje de ocupación. Este aspecto se tendrá en cuenta a la hora de realizar el modelo espacio-temporal, pues incluye partes regresivas.

1.1.6. Análisis de correlaciones mutuas entre las covariables

En esta sección, se repite el análisis anterior, pero para estudiar las posibles correlaciones entre cada una de las covariables. De esta forma, se analizará la posible existencia de multicolinealidad, que pudiera influir en la parte regresiva de los modelos espacio-temporales. Por otro lado, se podrá determinar si existen variables que están tan relacionadas que en realidad pertenecen a la misma distribución, con lo que debe tenerse en cuenta para eliminar una de ellas. Los resultados del análisis de correlaciones mutuas entre las covariables se muestran en la Tabla 1.6.

Correlación mutua	r	p-valor
Temperatura máxima - Temperatura mínima	0.878	0
Temperatura máxima - Precipitaciones	-0.275	0
Temperatura mínima - Precipitaciones	-0.108	0

Tabla 1.6: Correlaciones mutuas entre las covariables

De nuevo, debido al tamaño de la muestra, las conclusiones son muy significativas. Se puede observar que hay una correlación muy fuerte entre la temperatura máxima y la temperatura mínima. Esta conclusión es lógica pues, si aumenta la temperatura media, por ejemplo en verano, suben tanto las temperaturas mínimas como las máximas, y además con el mismo signo. Sin embargo, no provienen de la misma distribución estadística, pues su media es claramente diferente, y la muestra es suficientemente grande. Esta afirmación se demuestra mediante la aplicación del test de Kolmogorov-Smirnov (KS), que analiza las diferencias entre las dos funciones de distribución que se están comparando. Se calcula el estadístico D, como

$$D = \text{máx}[F_1(x) - F_2(x)],$$

donde F_1 y F_2 son las funciones de distribución de las variables bajo comparación, en este caso, temperatura máxima y temperatura mínima. El resultado obtenido es que las dos funciones de distribución difieren en $D = 0{,}54$, con una significación estadística altísima, de nuevo debido al tamaño de la muestra (Tabla 1.7).

Variables	D	p-valor
Temperatura máxima - Temperatura mínima	0.54	0

Tabla 1.7: Test de Kolmogorov-Smirnov para las temperaturas

El mismo resultado puede observarse en la Figura 1.6, donde la máxima diferencia entre las dos funciones de distribución es de 0.5 aproximadamente.

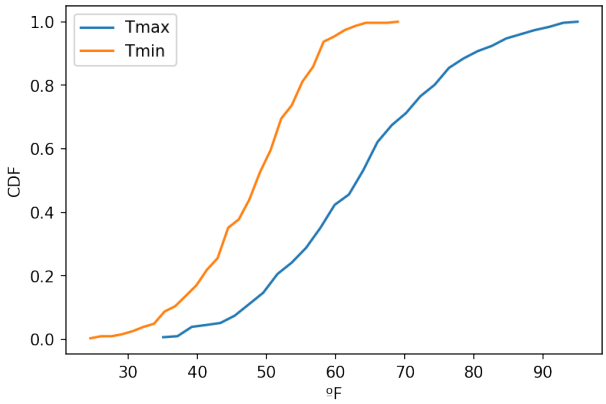

Figura 1.6: Funciones de distribución de las temperaturas máxima y mínima

Este test se presenta únicamente como comprobación formal de lo que se comentó con anterioridad: la temperatura máxima y la mínima están muy relacionadas, pero no provienen de la misma distribución, por lo que tendrán impactos diferentes en el porcentaje de ocupación. Por ejemplo, es posible que las temperaturas mínimas no afecten del mismo modo que las máximas en Seattle, zona de origen de los datos, puesto que Seattle es una zona localizada al norte de EEUU, con lo que las temperaturas máximas no serán tan extremas como las mínimas. Esto podría tener impacto en la cantidad de desplazamientos en coche que hay hacia la zona de los parquímetros.

1.2. Análisis descriptivo dinámico

A continuación se presenta un análisis dinámico del dataset, donde se describen efectos y propiedades del mismo, pero en función del lugar y tiempo en el que se produjeron. Se realiza primero un estudio temporal, donde se relaciona la variable de ocupación (y la distribución de transacciones) con la temporalidad del fenómeno bajo estudio. Después, se analiza de forma geográfica, presentando las distribuciones de ocupación por localización (parquímetro). Por último, se explican cuestiones relativas a la frecuencia de actualización de los parquímetros, que es muy relevante a la hora de decidir qué parquímetros utilizar para realizar la predicción.

1.2.1. Análisis temporal

En primer lugar, en la Figura 1.7, se muestra una representación gráfica en la que aparece la distribución estadística de las transacciones (tickets) en función de la hora en la que se produjeron. Cabe destacar que las horas centrales del día (10h-16h) son las de mayor actividad, e implican una mayor rotación de las plazas de aparcamiento.

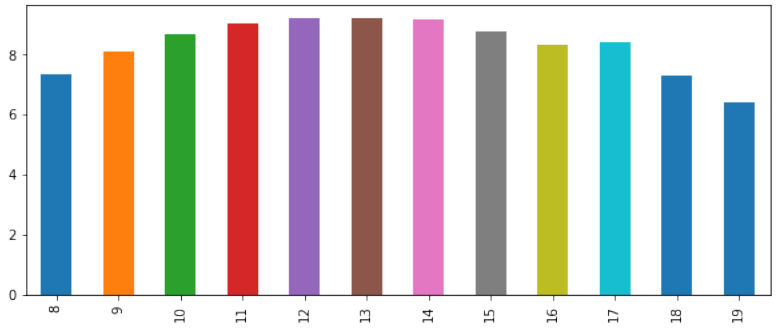

Figura 1.7: Distribución de las transacciones en función de la hora del día

Continuando con el análisis, la Figura 1.8 presenta la distribución del porcentaje de ocupación de los parquímetros en función de la hora del día. Se oberva que la distribución es más uniforme, pero, aún así, tiende a parecerse bastante a la distribución de transacciónes. Esto implica que las horas centrales del día suelen constituir las horas más relevantes para el análisis.

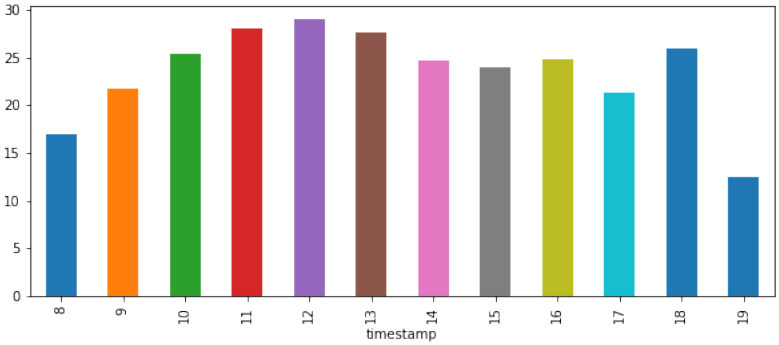

Figura 1.8: Distribución de la ocupación de los parquímetros en función de la hora del día

Dentro de una misma semana, tiende a haber mayor ocupación en los días finales de la semana (Jueves, etiquetado como 3, Viernes, etiquetado como 4, y Sábado, etiquetado como 5), según se representa en la Figura 1.9. Es lógico que se obtenga este resultado, pues los días cercanos al fin de semana suelen llevar aparejados mayores desplazamientos. Nótese que el domingo no aparece representado por no estar activo el sistema de pago por aparcamiento en días festivos.

Si extendemos el análisis a los días dentro de un mes (Figura 1.10), se observa que la distribución es relativamente uniforme: no se aprecia una diferencia significativa entre los días de principio de mes (días 1 a 7, etiquetados como 'begin'), los días de final de mes (días 25 a 31, etiquetados como

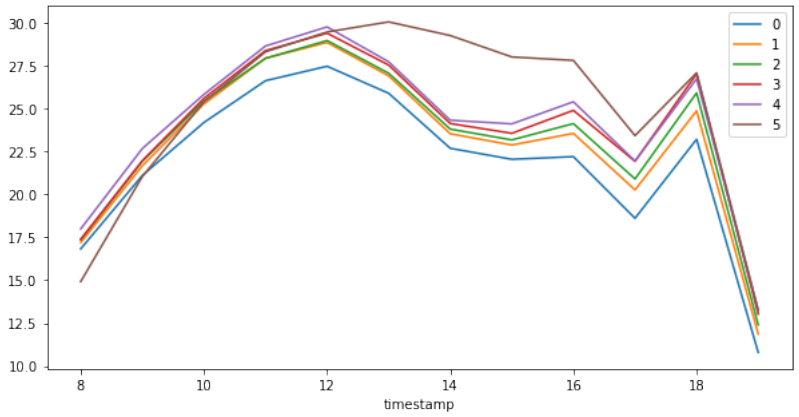

Figura 1.9: Distribución de la ocupación de los parquímetros en función del día de la semana

'end'), y el resto (etiquetados como 'rest'). Además, también se aprecia que la distribución por horas se mantiene tanto a lo largo de una semana como a lo largo de un mes, siempre favoreciéndose ocupaciones mayores en las horas centrales del día.

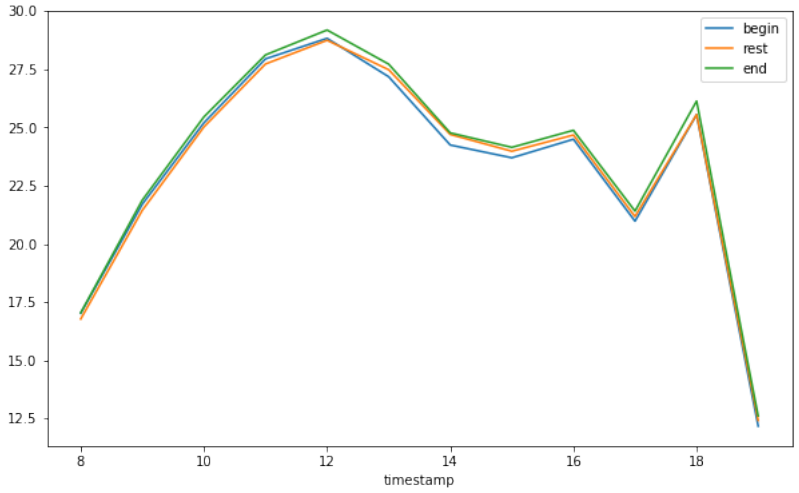

Figura 1.10: Distribución de la ocupación de los parquímetros en función del día del mes

Por último, en un año completo, se aprecia que en los meses de verano hay mayor ocupación que a lo largo del resto del año, posiblemente debido a la influencia de temperaturas más suaves, mientras que en invierno la ocupación es menor (Figura 1.11). Aún así, la diferencia no es muy grande, y se sigue manteniendo la tendencia horaria a mayores ocupaciones entre las 10h y las 14h.

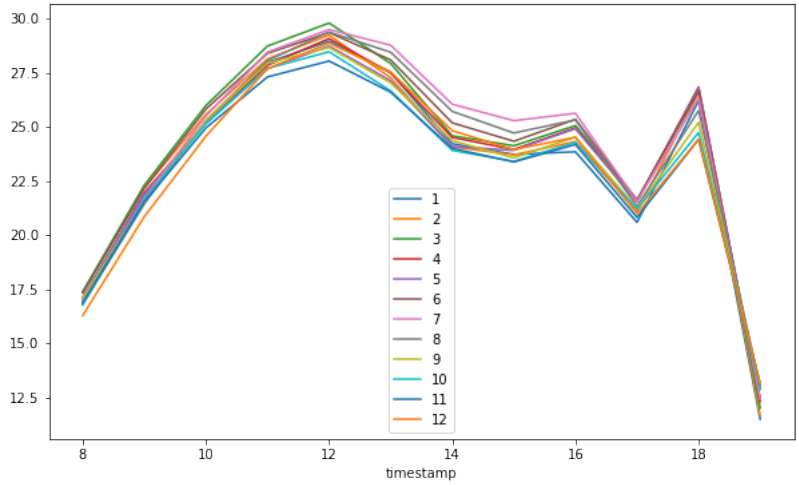

Figura 1.11: Distribución de la ocupación de los parquímetros en función del mes

1.2.2. Análisis espacial

En cuanto a la distribución espacial de la ocupación de los parquímetros, puede observarse en la Figura **??** que es muy heterogénea. Hay parquímetros con una tasa de ocupación elevada durante gran parte del día, posiblemente localizados en el centro de la ciudad, y parquímetros que apenas se llenan durante todo el día. Los parquímetros con mayor ocupación se presentan en la Figura **??**

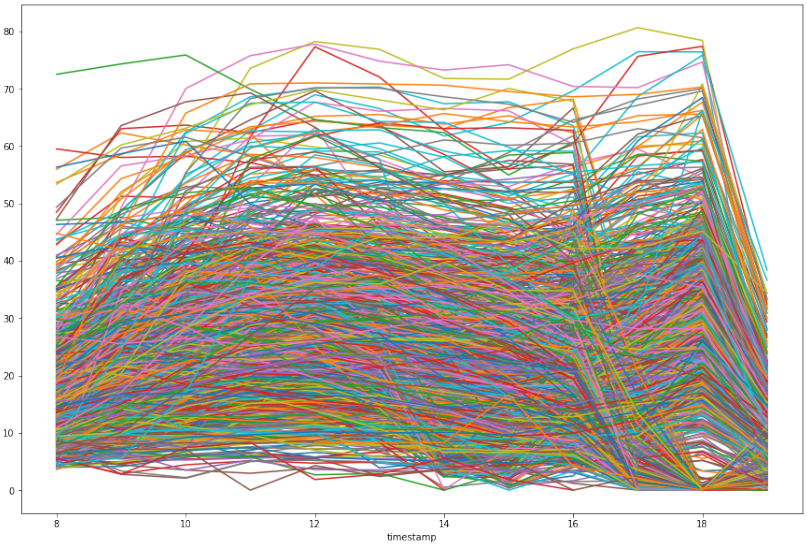

Figura 1.12: Top 100 de parquímetros según su localización

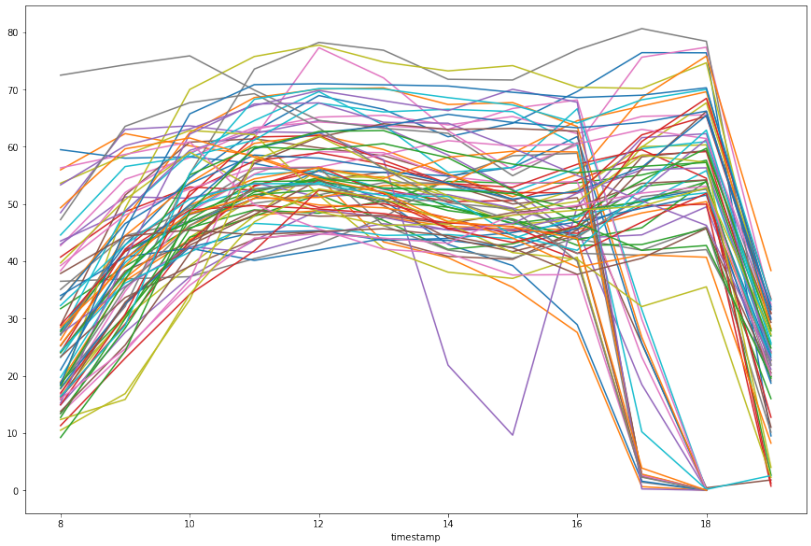

Figura 1.13: Top 100 de parquímetros según su porcentaje de ocupación

1.2.3. Frecuencia de actualización

La frecuencia de actualización de un parquímetro es la tasa a la que se actualiza el estado de sus plazas de aparcamiento. Es una medida de la rotación de vehículos del mismo. Parquímetros con una frecuencia alta de actualización ven circular muchos vehículos por ellos durante el día, mientras que parquímetros con una frecuencia de actualización baja prácticamente no tienen movimiento. Esta variable es importante para realizar la predicción, pues zonas con una tasa de actualización elevada presentan muchas transacciones, es decir, dan mucha información sobre el fenómeno bajo estudio. Por ello, es importante seleccionar parquímetros que permitan realizar buenas generalizaciones. En la Figura **??** se presenta la frecuencia de actualización del parquímetro 12289 durante el mes de enero, en función del tiempo.

Los parquímetros del dataset presentan frecuencias de actualización muy variables (Figura 1.15), siendo algunas de ellas muy bajas. Algunos parquímetros se actualizan solo una o dos veces cada día. Estas series son poco atractivas para utilizarse en un algoritmo predictivo. Los parquímetros que configuran mejores series para utilizarse como referencia en el modelo predictivo son aquellos con una frecuencia alta de actualización (Figura **??**

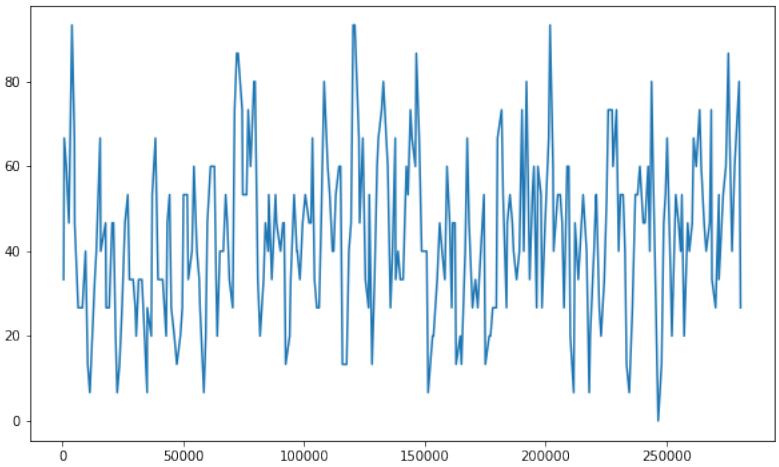

Figura 1.14: Frecuencia de actualización del parquímetro 12289 durante el mes de enero

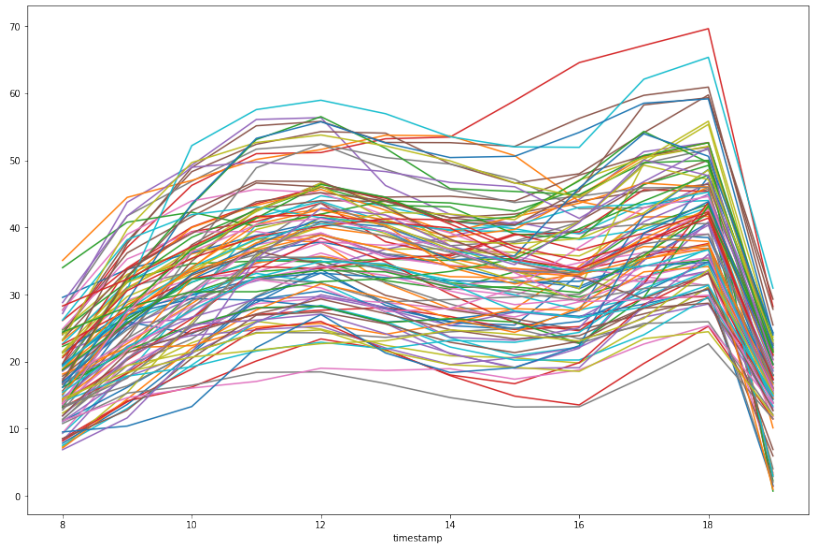

Figura 1.15: Ocupación del Top 100 de parquímetros con mayor frecuencia de actualización