Proyecto Final: Análisis y Predicción del Tráfico Vehicular Usando Matrices de Sketching

Sergio Mínguez Cruces

19 de febrero de 2026

Índice

1. Introducción

1.1. Contexto

El contexto sobre la base de datos *TrafficVolumeData*, que documenta el flujo vehicular en la ciudad de Chicago, se encuentra en el ámbito de la planificación urbana y la gestión del tráfico. En ciudades grandes como Chicago, el volumen de tráfico representa un desafío constante, no solo para la movilidad eficiente, sino también para el impacto en la calidad del aire, la seguridad vial y la productividad de sus habitantes.

Este conjunto de datos ofrece una visión integral de las dinámicas del tráfico urbano, capturando factores ambientales, temporales y climáticos que influyen directamente en los volúmenes vehiculares. Las 33750 observaciones registradas reflejan una rica diversidad de escenarios, desde condiciones meteorológicas extremas hasta patrones típicos de días laborales o festivos.

1.2. Descripción superficial de los variables

Cada registro incluye 14 variables que abarcan:

1. Factores meteorológicos: como temperatura, humedad, visibilidad, velocidad y dirección del viento, así como niveles de precipitación, cobertura nubosa, tipo de clima y una descripción más detallada.
2. Variables temporales: como la fecha y hora del registro, además de un indicador que identifica si el día es feriado o no.
3. Indicadores de calidad del aire: como el índice de contaminación.
4. Medidas específicas del tráfico: como el volumen vehicular capturado.

1.3. Objetivo del Análisis

El análisis de este conjunto de datos tiene como objetivo principal comprender cómo las condiciones climáticas, el calendario (días laborales vs. feriados), y otros factores externos afectan la movilidad en una ciudad altamente poblada. Estos conocimientos son fundamentales para diseñar estrategias que optimicen el flujo vehicular, reduzcan los niveles de contaminación y mejoren la calidad de vida urbana.

Por ejemplo, el análisis de las relaciones entre el volumen de tráfico y las condiciones climáticas puede ayudar a las autoridades locales a prever congestiones y a implementar medidas preventivas, como ajustes en los semáforos o restricciones vehiculares en condiciones adversas. Además, al identificar patrones temporales, se pueden establecer intervenciones específicas para mejorar la infraestructura vial o fomentar el uso del transporte público en horarios pico.

Este estudio no solo busca optimizar la movilidad en Chicago, sino también servir como modelo para otras ciudades que enfrentan desafíos similares, generando un impacto positivo y sostenible en las áreas de transporte y gestión ambiental.

2. Descripción de los variables

2.1. date_time:

La variable temporal permite identificar patrones de tráfico según la hora, el día de la semana o la temporada. En este análisis es útil para determinar las horas pico y las variaciones de tráfico durante el día. Además, es importante para predecir el comportamiento del tráfico durante las estaciones más frías o más cálidas.

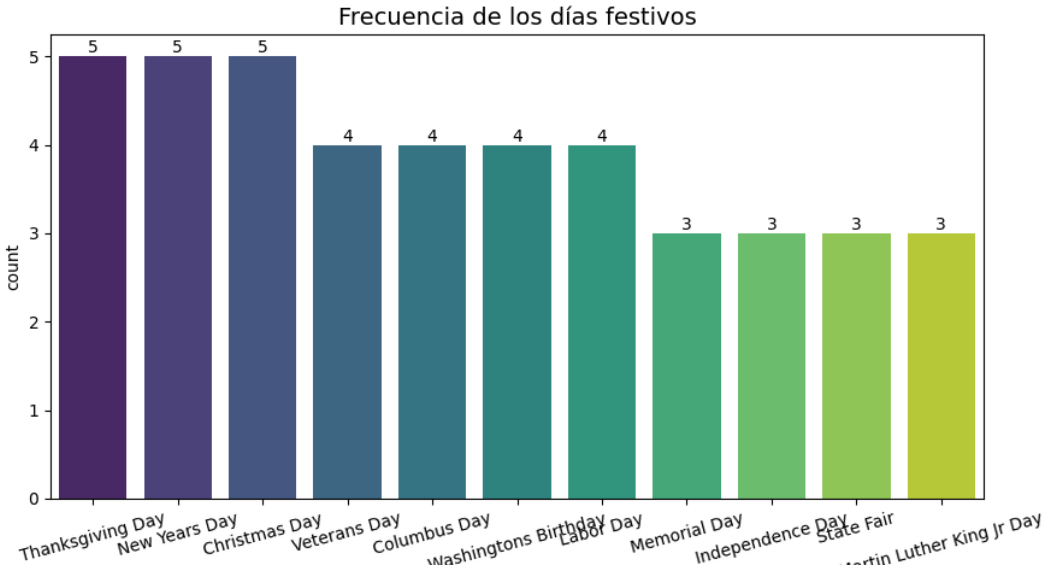
En este estudio va desde el 2012-10-02 hasta el 2017-05-17 habiendo un paron desde 2014-08-08 a 2015-06-11. Cada día a su vez está dividido por horas para así poder indentificar mejor los patrones de tráfico.

Debido a que otro grupo está haciendo un proyecto sobre series temporales solo se usará está variable para hacer dos gráficas para comprobar que efectivamente hay relación entre variables una será la de

la temperatura que será anual y otra con la variable objetivo que será diaria para poder ver mejor los patrones del dia a dia.

2.2. is_holiday:

Los días feriados suelen tener patrones de tráfico distintos, como una menor congestión en zonas de oficinas y mayor en áreas recreativas. Esta variable ayuda a analizar cómo las dinámicas de movilidad cambian durante días no laborables y contribuye a planificar estrategias específicas para mitigar problemas en días festivos o vacaciones.

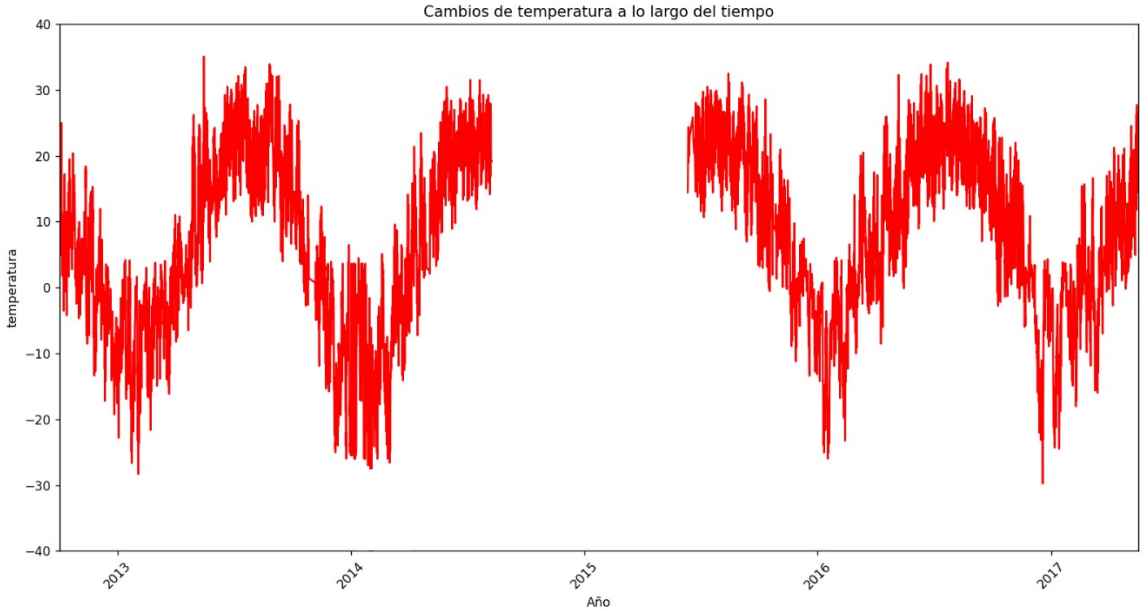


Esta variable toma los valores Çhristmas Day"(5 observaciones), Çolumbus Day"(4 observaciones), Ïndependence Day"(3 observaciones), "Labor Day"(4 observaciones), "Martin Luther King Jr Day"(3 observaciones), "Memorial Day"(3 observaciones), "New Years Day"(5 observaciones), "None"(33,707 observaciones), "State Fair"(3 observaciones), "Thanksgiving Day"(5 observaciones), "Veterans Day"(4 observaciones), y "Washington's Birthday"(4 observaciones). La mayoría de las observaciones (99.98 %) corresponden a días sin festividades (valor "None"), lo que sugiere que la mayoría de los datos no se recogen en días festivos. Solo un pequeño porcentaje de las observaciones (0.02 %) son días festivos específicos, lo que indica una baja frecuencia de días festivos en el conjunto de datos.

2.3. temperature:

La variable está en grados kelvin por lo que para convertirlos la temperatura de Kelvin a Celsius utilizando la fórmula $C = K - 273{,}15$. La temperatura puede influir indirectamente en el tráfico. Por ejemplo, temperaturas extremas, tanto altas como bajas, pueden afectar el comportamiento de los conductores, el estado de los vehículos y la infraestructura vial, como el asfalto.

La variable tiene una media de 280.07 K (aproximadamente 6.92°C) y una desviación estándar de 13.42 K. Los cuartiles (Q1=271.72 K, mediana=280.15 K, Q3=290.62 K) indican que la mayoría de los valores están entre 271.72 K (aproximadamente -1.33°C) y 290.62 K (aproximadamente 17.47°C). El rango total va de 243.62 K (aproximadamente -29.53°C) a 308.24 K (aproximadamente 35.09°C), lo que indica una amplia variabilidad en las temperaturas, con registros de temperaturas muy frías y cálidas. Dado que la desviación estándar es de 13.42 K, los datos están moderadamente dispersos, lo que refleja la variabilidad de la temperatura en las observaciones.

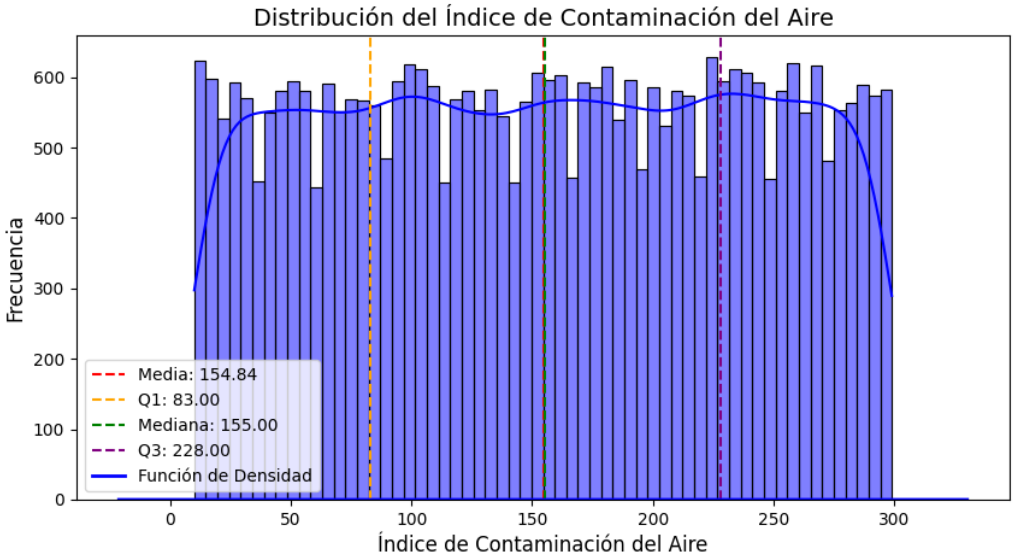


Se puede apreciar en la gráfico un claro patron en las temperaturas coincidiendo con las estaciones, además del paron comentado en las fechas.

2.4. air_pollution_index:

El tráfico vehicular es una de las principales fuentes de contaminación del aire en áreas urbanas. Esta variable relaciona el volumen de tráfico con su impacto ambiental, proporcionando información valiosa para identificar momentos críticos de contaminación y diseñar políticas que incentiven modos de transporte más limpios, como bicicletas o transporte público eléctrico.

La variable tiene una media de 154.84 y una desviación estándar de 83.74, lo que indica una variabilidad considerable en los niveles de contaminación del aire. Los cuartiles (Q1=83.00, mediana=155.00, Q3=228.00) muestran que la mayoría de los valores se concentran entre 83 y 228. El rango va de un valor mínimo de 10 a un valor máximo de 299, lo que refleja una variabilidad amplia en los datos. Dado que la desviación estándar es alta, los datos están moderadamente dispersos, lo que sugiere que hay una notable diversidad en los niveles de contaminación.

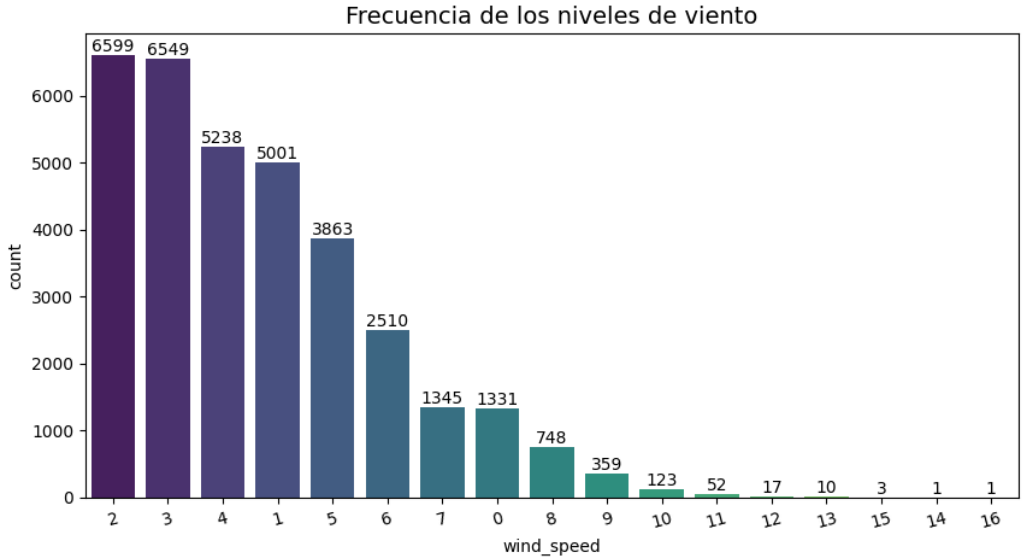


Debido a que los datos se han tomado en un tiempo relativamente grande, los valores que dependen de estaciones como este caso; ya que por varios estudios respecto al cambio climático se llegó a la conclusión que en otoño e invierno los valores son más altos que en primavera y verano debido a la caída de las hojas, el histograma posee valores muy parecidos, la mejor forma de ver este patrón sería con un gráfico temporal pero como ya se ha mencionado hay otro grupo que hace ese tipo de estudio por lo que no se hará.

2.5. wind_speed:

Esta varibla divide en 17 grupos la velocidad del viento.El viento fuerte influye en la estabilidad de los vehículos y la seguridad vial, especialmente para vehículos pesados. Los días con viento fuerte suelen mostrar un tráfico más lento y pueden ser un factor importante para la congestión en ciertas rutas, especialmente en áreas abiertas o puentes.

La variable tiene una media de 3.38 y una desviación estándar de 2.06 . Los cuartiles (Q1=2.00, mediana=3.00, Q3=5.00) indican que la mayoría de los valores están entre 2 y 5. El rango va de 1 a 16, lo que sugiere que, aunque la mayoría de las observaciones son de vientos suaves, pueden ocurrir ráfagas más fuertes en algunos casos. La desviación estándar es relativamente alta, lo que sugiere que los datos están moderadamente dispersos, con algunas observaciones de viento muy ligero y otras de viento mucho más fuerte.

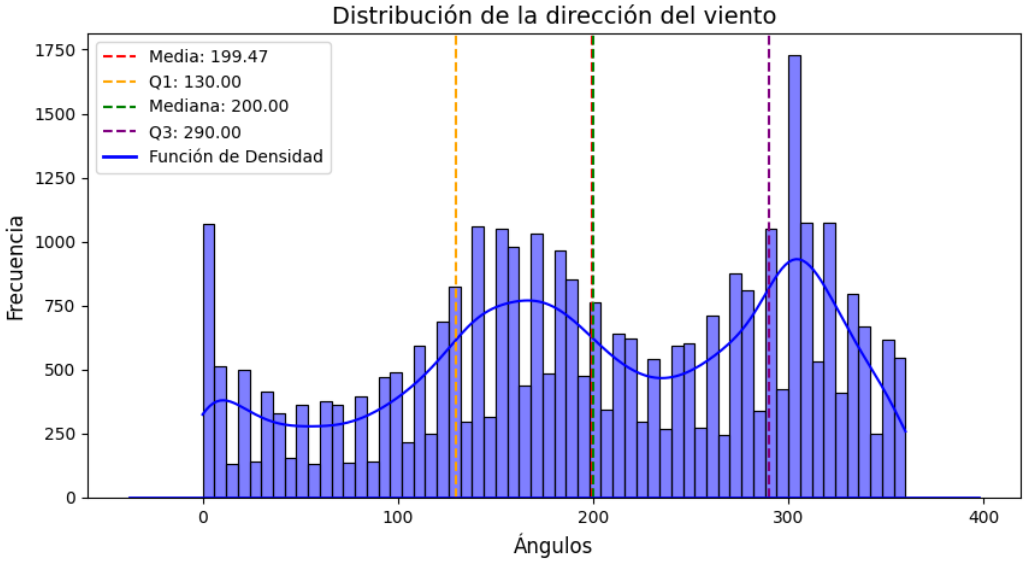


El gráfico muestra que los niveles más comunes están entre 2 y 4, indicando vientos ligeros a moderados. Velocidades altas, como 10 o más, son infrecuentes, mientras que valores extremos como 15 o 16 aparecen como anomalías. Esto sugiere un entorno mayormente tranquilo y estable en términos de viento.

2.6. wind_direction:

Esta variable recoge la dirección en ándulos sexagesimales, desde 0º (excluido) hasta 360º (incluido), girando en el sentido de las agujas del reloj en el plano horizontal visto desde arriba. Valores cercanos a 1º y 360º indican viento del norte, cercanos a 90º viento del este, 180º del sur y 270º del oeste.

La variable tiene una media de 199.47 grados y una desviación estándar de 99.84 grados. Los cuartiles (Q1=130.00, mediana=200.00, Q3=290.00) muestran que la mayoría de los valores están distribuidos en direcciones entre 130° y 290°. El rango total va de 0 a 360 grados, lo que cubre todas las direcciones posibles. La desviación estándar alta indica que los datos están bastante dispersos, lo que refleja la variabilidad en las direcciones del viento.

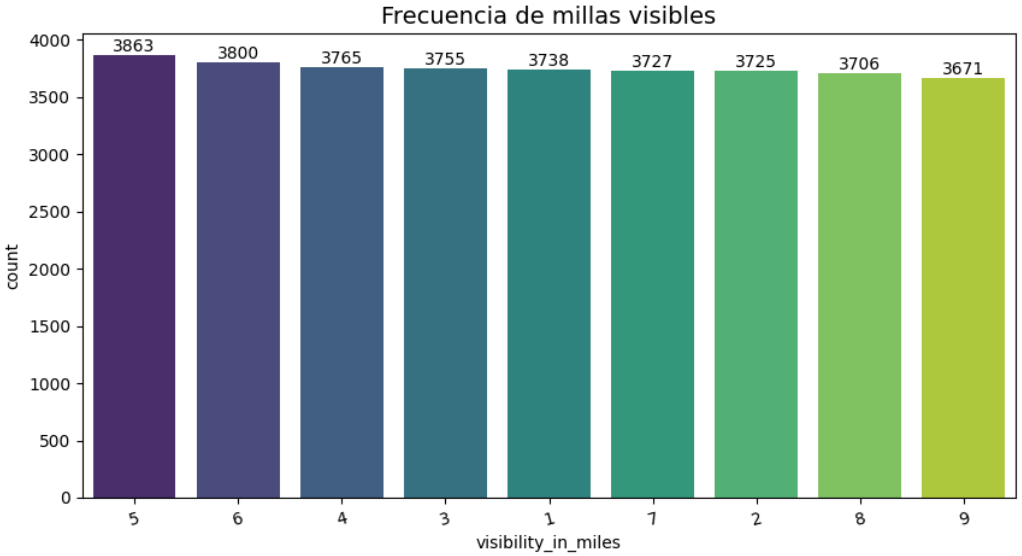


El gráfico muestra la distribución de la dirección del viento, expresada en ángulos en grados. Según el histograma y la función de densidad, las direcciones más frecuentes se concentran en torno a 200° y 300°, lo que sugiere un predominio de vientos provenientes del sur y suroeste. Por otro lado, las direcciones menos frecuentes se encuentran cerca de 90° (este) y alrededor de 360°/0° (norte). Esto podría reflejar patrones climáticos típicos de la región, posiblemente influenciados por sistemas meteorológicos locales o geografía específica.

2.7. visibility_in_miles:

La visibilidad es fundamental para la seguridad del tráfico. Un bajo nivel de visibilidad, especialmente durante la niebla o la lluvia, reduce la velocidad promedio de los vehículos y aumenta la probabilidad de accidentes. Estudiar esta variable ayuda a gestionar el tráfico en condiciones climáticas adversas.

La variable tiene una media de 4.99 millas y una desviación estándar de 2.57 millas. Los cuartiles (Q1=3.00, mediana=5.00, Q3=7.00) muestran que la mayoría de las observaciones tienen una visibilidad entre 3 y 7 millas. El rango varía entre un mínimo de 1 milla y un máximo de 9 millas, lo que sugiere una visibilidad generalmente buena, aunque en algunos casos se reduce considerablemente. La desviación estándar es moderada, indicando una dispersión razonable en los datos.



El gráfico de visibilidad muestra que los valores más comunes están alrededor de 5 millas, con frecuencias similares para valores cercanos como 4 y 6 millas. Esto indica que las condiciones de visibilidad suelen ser moderadas y relativamente consistentes. Las visibilidades perfectas (9 millas) son menos frecuentes, lo que puede estar asociado con factores climáticos locales como neblina o contaminación

2.8. rain_p_h:

La lluvia es uno de los factores climáticos más influyentes en el tráfico. La precipitación reduce la visibilidad, aumenta el tiempo de frenado y puede causar inundaciones, lo que lleva a congestiones o bloqueos de carreteras. Este análisis es clave para gestionar la movilidad urbana durante las lluvias.

La variable tiene una media de 0.45 mm/h y una desviación estándar de 53.53 mm/h. La mayoría de los valores son cero, lo que indica que la lluvia es poco frecuente. Los cuartiles (Q1=0.00, mediana=0.00, Q3=0.00) muestran que la mayor parte de los valores son cero. El rango va de 0 mm/h a 9831.30 mm/h, lo que refleja la presencia de algunos eventos de lluvia extremadamente intensos. La desviación estándar es extremadamente alta, lo que sugiere que los datos están muy dispersos, ya que la mayoría de las observaciones son de no lluvia, pero hay eventos puntuales de lluvia intensa.

2.9. snow_p_h:

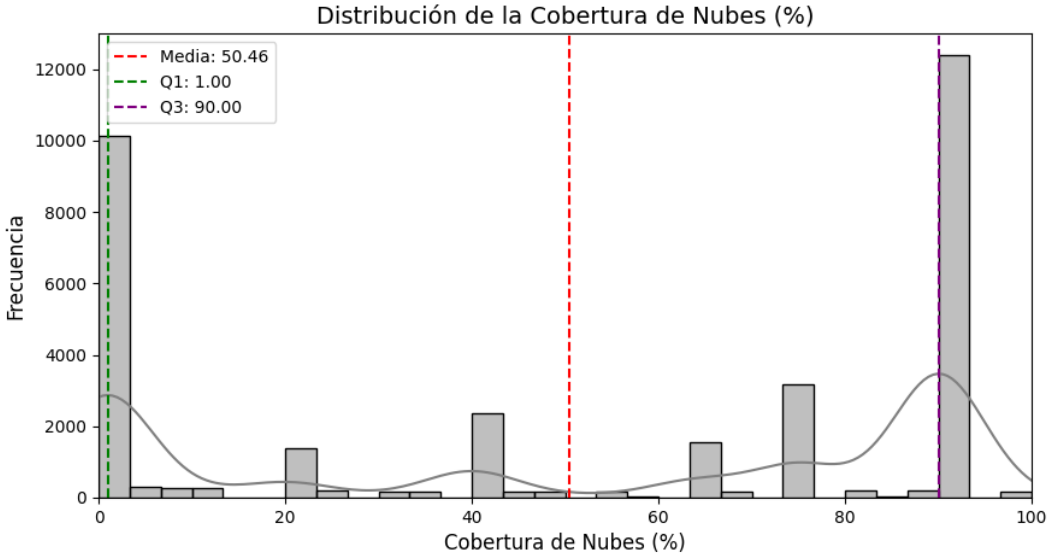
La nieve es otro factor importante que afecta el tráfico, especialmente en ciudades como Chicago. En días de nieve intensa, el tráfico puede disminuir drásticamente debido a condiciones de carreteras resbaladizas y cierres de rutas. Esta variable es crucial para prever el comportamiento del tráfico durante las tormentas de nieve.

La variable tiene una media de 0.00 mm/h y una desviación estándar de 0.01 mm/h, lo que indica que las nevadas son extremadamente raras. Los cuartiles (Q1=0.00, mediana=0.00, Q3=0.00) muestran que la mayor parte de los valores son cero. El rango va de 0 mm/h a 0.51 mm/h, sugiriendo que cuando ocurren nevadas, son de muy baja intensidad. Dado que la desviación estándar es muy baja, los datos están muy concentrados en cero, con escasa variabilidad en las nevadas.

2.10. clouds_all:

Aunque la cobertura nubosa no tiene un impacto directo tan grande en el tráfico, puede correlacionarse con otras condiciones climáticas que afectan la conducción. Por ejemplo, cielos nublados pueden preceder o coincidir con lluvias o tormentas, lo que afecta el volumen y la velocidad del tráfico.

La variable tiene una media de 50.46 % y una desviación estándar de 38.87 %. Los cuartiles (Q1=1.00, mediana=64.00, Q3=90.00) muestran que la mayoría de las observaciones tienen una cobertura de nubes entre 1 % y 90 %. El rango varía de 0 % a 100 %, lo que refleja tanto cielos despejados como nublados. Con una desviación estándar tan alta, los datos están bastante dispersos, lo que indica una gran variabilidad en la cantidad de nubes observada.



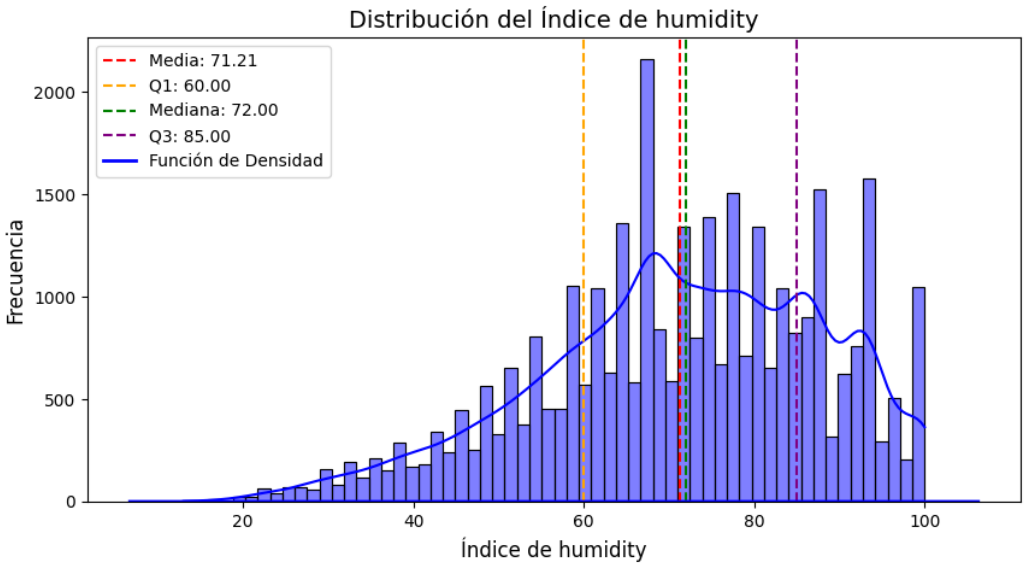
Este gráfico muestra un histograma de la cobertura de nubes en porcentaje. Los datos presentan una distribución bimodal, con valores agrupados cerca del 0 % y el 100 %, lo que indica una alta frecuencia de cielos completamente despejados o completamente cubiertos. La media se encuentra alrededor del 50.46 %, y los cuartiles indican que la mayoría de las observaciones están concentradas en los extremos. Esto puede sugerir que las condiciones climáticas en la región son principalmente cielos despejados o completamente nublados, con pocas situaciones intermedias.

2.11. humidity:

Esta variable representa el porcentaje de humedad ambiental. La humedad puede afectar la conducción al reducir la visibilidad o crear condiciones resbaladizas. El tráfico se ve impactado en días de alta humedad, como durante la niebla, lo que puede provocar accidentes y congestión. Su análisis es útil para

predecir condiciones climáticas adversas que afecten la seguridad vial.

La variable tiene una media de 71.21 y una desviación estándar de 16.85, lo que muestra una variabilidad moderada en los niveles de humedad. Los cuartiles (Q1=60.00, mediana=72.00, Q3=85.00) indican que la mayoría de los valores se concentran entre 60 y 85 siendo elevados pero comprensibles sabiendo donde está situado Chicago. El rango varía entre un mínimo de 13 y un máximo de 100. Dado que la desviación estándar es relativamente alta, los datos están algo dispersos, lo que indica que los niveles de humedad pueden variar considerablemente.

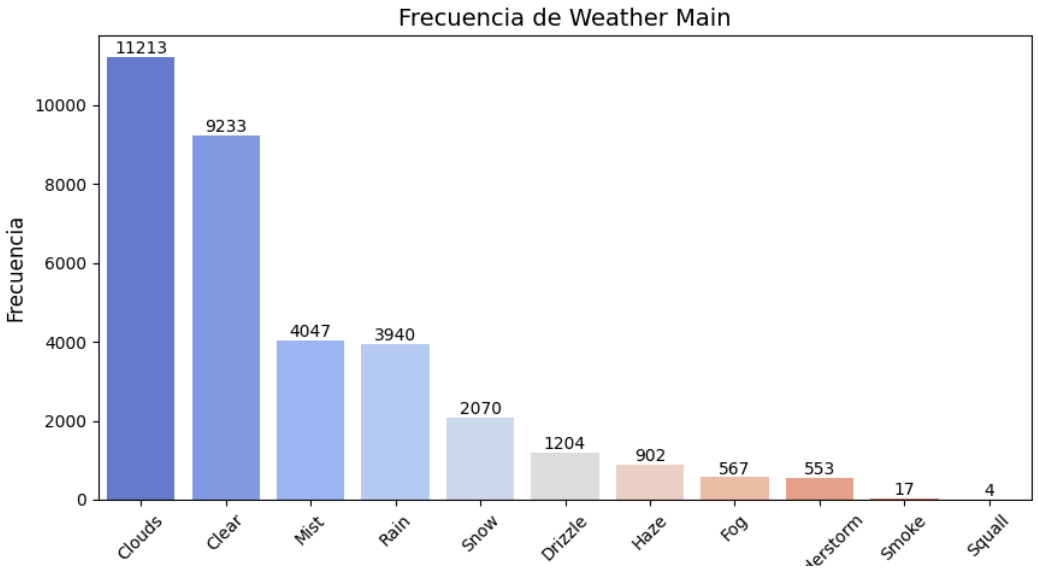


El histograma de la humedad relativa muestra una distribución más uniforme en comparación con el de la cobertura de nubes, con una leve asimetría hacia la derecha. La media de la humedad es del 71.21 %, lo que indica condiciones climáticas mayormente húmedas. El gráfico incluye líneas verticales que marcan los percentiles Q1 (60 %), mediana (72 %), y Q3 (85 %), lo que resalta la variación en los niveles de humedad en diferentes momentos. Además, la superposición de la función de densidad permite observar la probabilidad de ocurrencia de distintos rangos de humedad.

2.12. weather_type:

Esta variable sintetiza condiciones climáticas como “lluvioso”, “nevado” o “despejado”, que influyen directamente en el volumen y velocidad del tráfico. Su clasificación ayuda a identificar patrones específicos de movilidad bajo diferentes tipos de clima.

Esta variable toma los valores Çlear"(9,233 observaciones), Çlouds"(11,213 observaciones), "Drizzle"(1,204 observaciones), "Fog"(567 observaciones), "Haze"(902 observaciones), "Mist"(4,047 observaciones), Rain"(3,940 observaciones), "Smoke"(17 observaciones), "Snow"(2,070 observaciones), "Squall"(4 observaciones), y "Thunderstorm"(553 observaciones). La mayoría de las observaciones corresponden a condiciones de Çlouds"(33.3 %) y Çlear"(27.3 %), lo que sugiere que los días con cielos nublados o despejados son predominantes. Rain"(11.7 %) y "Mist"(12.0 %) también tienen una presencia significativa, mientras que fenómenos más extremos como "Thunderstorm", "Smokez "Squall"son menos frecuentes.

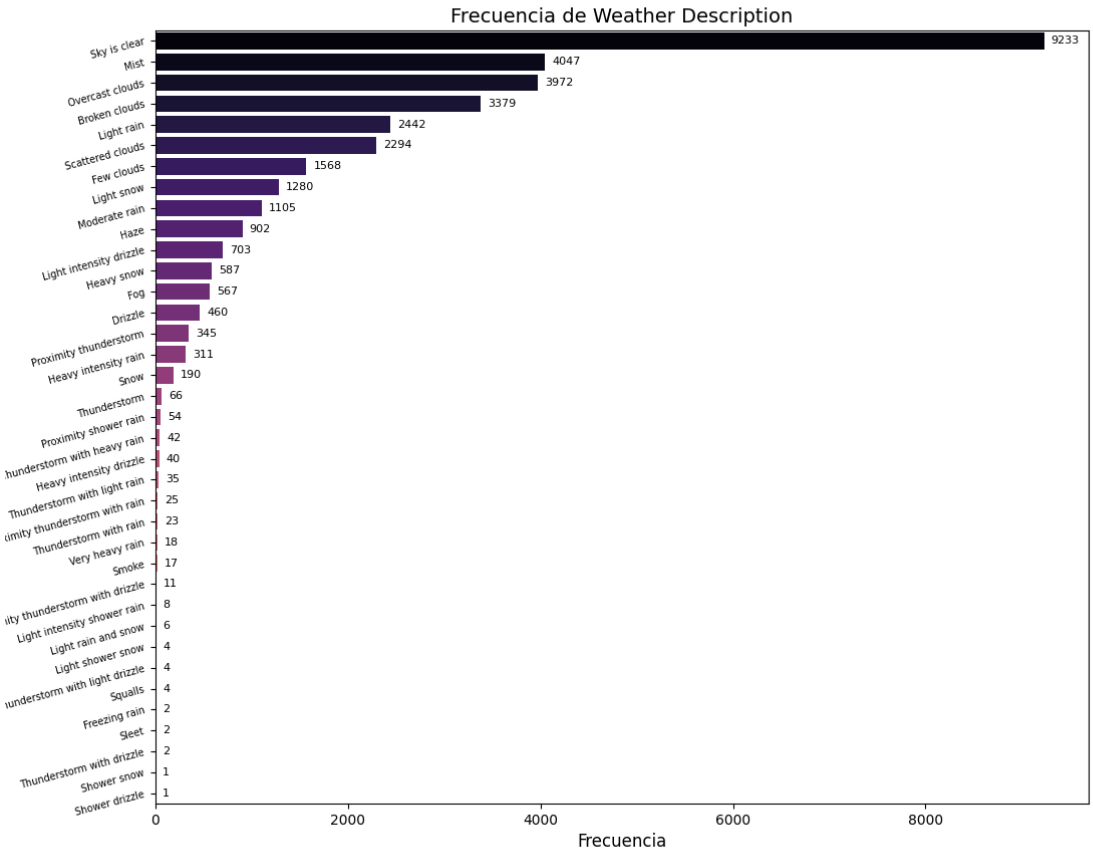


El gráfico muestra que las condiciones predominantes son cielos despejados, neblina y nubes cubiertas, mientras que fenómenos extremos como tormentas intensas o nieve fuerte tienen frecuencias muy bajas. Esto sugiere que el clima en general es estable, con eventos severos siendo inusuales y esporádicos, lo que puede influir positivamente en actividades cotidianas como el transporte.

2.13. weather_description:

Una versión más específica del clima que proporciona contexto adicional sobre las condiciones predominantes. Detalles como “lluvias ligeras” o “tormentas eléctricas” permiten un análisis más matizado del impacto climático en el tráfico.

Esta variable tiene una gran cantidad de categorías, incluyendo "broken clouds"(3,379 observaciones), "drizzle"(460 observaciones), "few clouds"(1,568 observaciones), "fog"(567 observaciones), "freezing rain"(2 observaciones), "haze"(902 observaciones), "heavy intensity drizzle"(40 observaciones), "heavy intensity rain"(311 observaciones), "heavy snow"(587 observaciones), "light intensity drizzle"(703 observaciones), "light intensity shower rain"(8 observaciones), "light rain"(2,442 observaciones), "light rain and snow"(6 observaciones), "light shower snow"(4 observaciones), "light snow"(1,280 observaciones), "mist"(4,047 observaciones), "moderate rain"(1,105 observaciones), " overcast clouds"(3,972 observaciones), "proximity shower rain"(54 observaciones), "proximity thunderstorm"(345 observaciones), entre otras. Las condiciones más comunes son "broken clouds"(10.0 %), "few clouds"(4.6 %), y "mist"(12.0 %), seguidas por "light rain"(7.2 %) y "moderate rain"(3.3 %). Las categorías más extremas como "freezing rain", "thunderstorm with heavy rain" terminado con "sleet"son poco frecuentes, indicando una prevalencia de condiciones moderadas de clima.

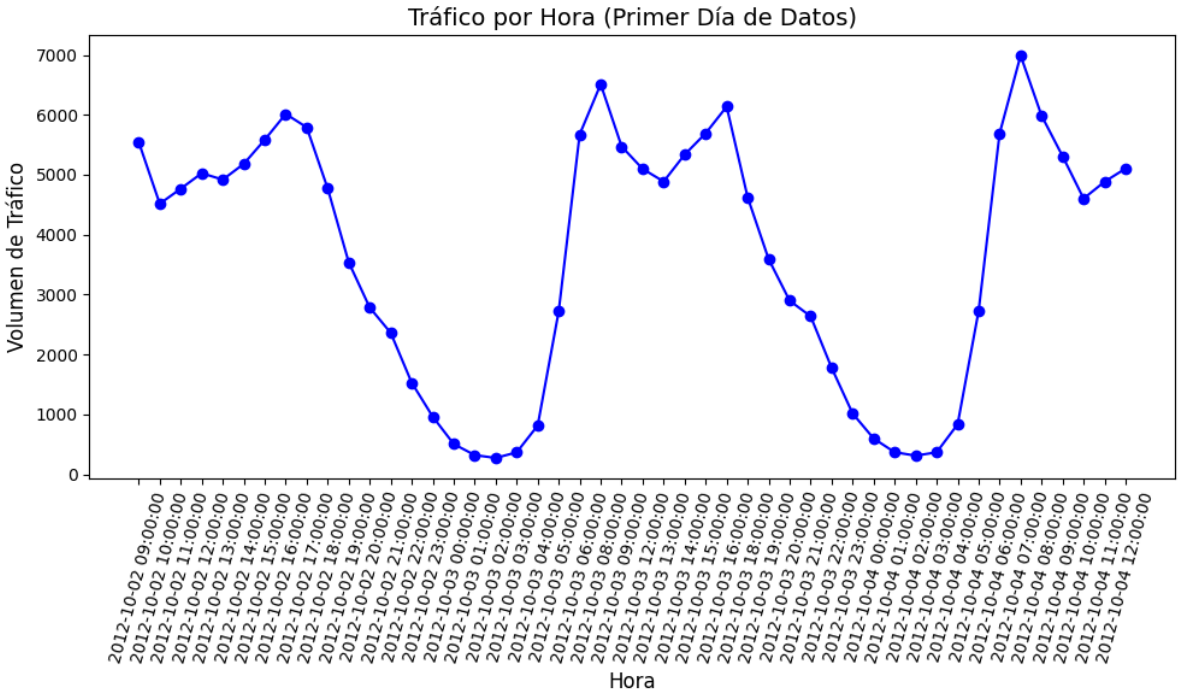


El gráfico muestra la frecuencia de diferentes descripciones del clima, destacando que las condiciones más comunes son "Sky is clear", "Mist" y " Overcast clouds", que superan con creces al resto. Esto sugiere que los días despejados, con niebla o nublados son predominantes en el conjunto de datos. Por otro lado, eventos climáticos extremos o poco frecuentes, como "Thunderstorm with drizzle" o "Shower drizzle", tienen muy pocas ocurrencias, lo que indica su rareza. Este patrón refleja un entorno donde los climas moderados son la norma, mientras que las condiciones inusuales ocurren con muy poca frecuencia.

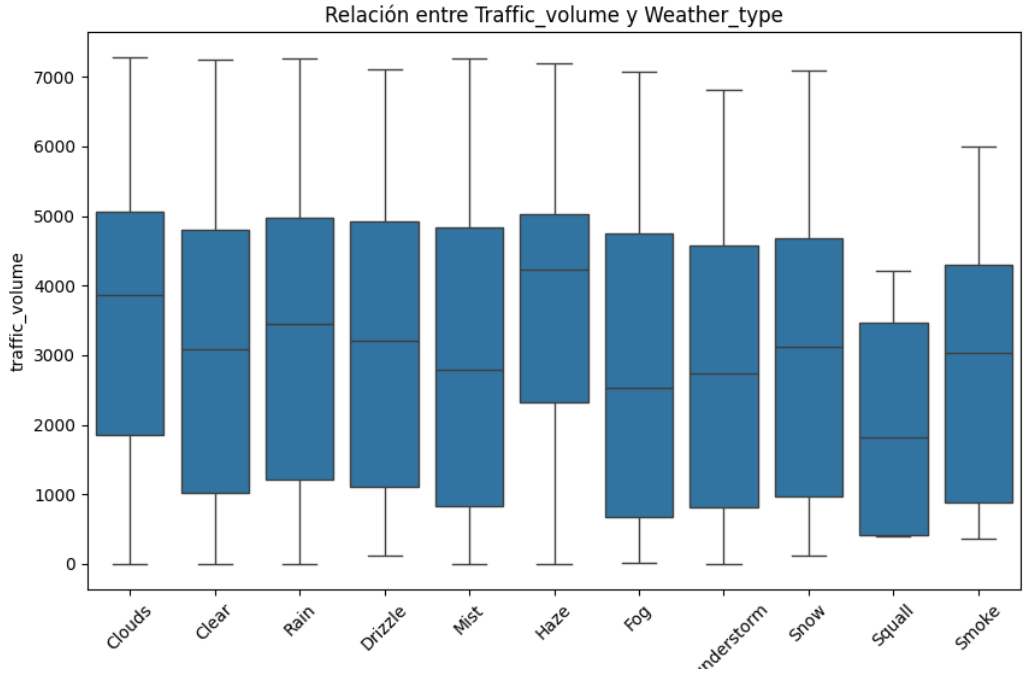
2.14. traffic_volume:

Esta variable refleja directamente el flujo vehicular en las calles de Chicago. Su análisis, junto con otras variables climáticas, puede ayudar a predecir cómo condiciones como la lluvia o las nevadas afectan la congestión del tráfico. Los picos en el volumen de tráfico también permiten identificar las horas de mayor congestión y los puntos críticos de la ciudad.

La variable tiene una media de 3240.12 vehículos y una desviación estándar de 1991.49. Los cuartiles (Q1=1165.25, mediana=3335.00, Q3=4926.00) indican que el volumen de tráfico generalmente varía entre 1165 y 4926 vehículos. El rango va de 0 a 7280 vehículos, lo que muestra períodos de alta y baja congestión en la ciudad. La desviación estándar elevada indica que los datos están muy dispersos, lo que refleja variabilidad en los niveles de tráfico dependiendo de la hora y la ubicación.



El gráfico de líneas representa el volumen de tráfico por hora en el primer día de los datos disponibles. Se puede observar un patrón claro con picos en las horas de la mañana y la tarde, lo que corresponde típicamente a las horas punta de desplazamiento hacia y desde el trabajo. Entre estos picos, el tráfico disminuye considerablemente durante la noche y las horas de descanso. Este comportamiento resalta la importancia de considerar los patrones temporales en la gestión del tráfico, particularmente para mitigar congestiones en horas pico.



El gráfico muestra la relación entre el volumen de tráfico y diferentes tipos de clima. La mediana del tráfico es similar en la mayoría de las condiciones, mientras que climas extremos o inusuales, co-

mo "Squallz "Smoke", presentan menores volúmenes de tráfico, lo que podría indicar una disminución de movilidad en situaciones adversas. Esto es razonable, ya que condiciones climáticas inusuales suelen reducir la actividad diaria. Por otro lado, climas comunes como lluvia o niebla no parecen afectar significativamente el tráfico, posiblemente porque las personas están acostumbradas a estas condiciones en su entorno

3. Desarrollo del Modelo de Predicción

Para poder hacer el modelo predictivo es necesario convertir las varibles categóricas y las date_time en varibles numéricas para posteriormente hacer el modelo normal y el modelo de sketching.

3.1. Transformación de la Columna `date_time` a Formato Numérico

```
data['date_time'] = pd.to_datetime(data['date_time']).astype(int) / 10**9
```

- `pd.to_datetime(data['date_time'])`: Convierte la columna `date_time` en un objeto de tipo `datetime` de Pandas. Esto permite trabajar con las fechas y horas de manera estructurada y realizar operaciones como filtrado y cálculos de diferencias de tiempo.
- `.astype(int)`: Convierte el objeto `datetime` a un entero que representa el número de nanosegundos desde el *epoch* (1970-01-01 00:00:00 UTC). Esta operación convierte las fechas a un valor numérico que puede ser procesado fácilmente.
- `/` 10^9: Se divide el número de nanosegundos entre 10^9 para convertir el valor en segundos desde el *epoch* (formato estándar de Unix timestamp).

El resultado final es que la columna `date_time` ahora contiene valores en formato Unix timestamp en segundos. Este formato es útil para realizar cálculos matemáticos, filtrado temporal o interacciones con sistemas que manejan timestamps numéricos. Por ejemplo, '2023-12-27 14:30:00' se transforma en el valor 1703687400.

3.2. Codificación de Variables Categóricas

```
categorical_cols = ['is_holiday', 'weather_type', 'weather_description']
data[categorical_cols] = data[categorical_cols].astype('category').apply(lambda
    x: x.cat.codes)
```

- `categorical_cols = ['is_holiday', 'weather_type', 'weather_description']`: Definimos las columnas categóricas a convertir.
- `data[categorical_cols] = data[categorical_cols].astype('category').apply(lambda x: x.cat.codes)`: Convertimos las columnas categóricas a valores numéricos mediante una codificación ordinal. Cada categoría se convierte en un número entero. Por ejemplo, " Clear" podría convertirse en 0, " Clouds" en 1, etc.

Este paso facilita el uso de variables categóricas en modelos de predicción, transformándolas en valores numéricos que pueden ser procesados por algoritmos de machine learning.

3.3. Cálculo de la Matriz de Correlación

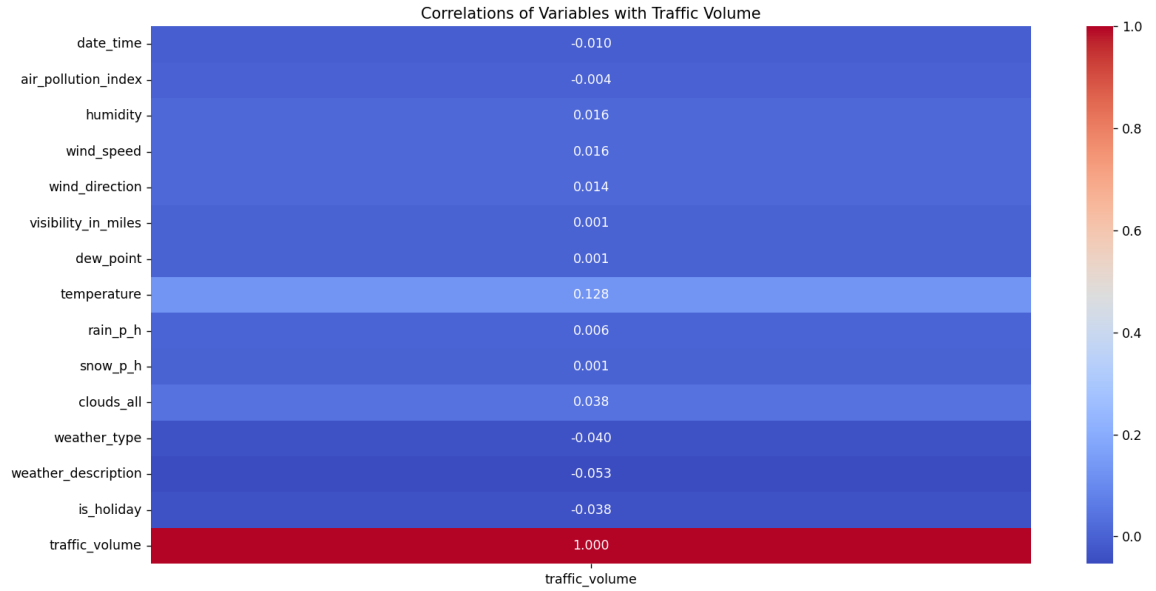
```
correlation_matrix = data.corr()

traffic_volume_correlation = correlation_matrix['traffic_volume']
```

- `correlation_matrix = data.corr()`: Calculamos la correlación entre todas las variables numéricas del conjunto de datos. La correlación mide la relación lineal entre dos variables. Un valor cercano a 1 indica una fuerte correlación positiva, mientras que un valor cercano a -1 indica una fuerte correlación negativa.

- `traffic_volume_correlation = correlation_matrix['traffic_volume']`: Filtramos las correlaciones con la columna `traffic_volume`, que es la variable objetivo.

Este paso nos ayuda a identificar qué variables están más relacionadas con `traffic_volume`, facilitando la selección de características relevantes.



Como se puede apreciar no habría correlación entre las varibales y la variable objetivo, esto se debe principalmente debido al gran número de datos periodicos que hay y al hecho de que variables como date_time la cual ya se ha visto su relación se ha pasado a segundos provocando que a la misma hora de dos días distintos posean números muy distantes haciendo así que no pueda calcular bien la relación.

3.4. Preparación de Datos para la Regresión Lineal

```
X = data.drop(columns=['traffic_volume'])
y = data['traffic_volume']

scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.2,
    random_state=42)
```

- `X = data.drop(columns=['traffic_volume'])`: Separamos las características (`X`) y la variable objetivo (`y`).
- `scaler = StandardScaler()`: Usamos `StandardScaler` para escalar las características, ajustando cada una para que tenga media 0 y desviación estándar 1.
- `X_scaled = scaler.fit_transform(X)`: Escalamos las características de `X`.
- `X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.2, random_state=42)`: Dividimos el conjunto de datos en entrenamiento y prueba (80 % entrenamiento y 20 % prueba).

El escalado es esencial porque las variables pueden tener diferentes escalas, lo que puede sesgar el modelo.

3.5. Transformación a Características Polinómicas de Grado 3

```
poly = PolynomialFeatures(degree=3)
X_train_poly = poly.fit_transform(X_train)
X_test_poly = poly.transform(X_test)
```

- `poly = PolynomialFeatures(degree=3)`: Utilizamos `PolynomialFeatures` para generar características polinómicas de grado 3. Esto genera nuevas características que representan combinaciones no lineales de las características originales. Por ejemplo, si tenemos dos características x_1 y x_2, la transformación de grado 2 generará x_1^2, $x_1 * x_2$, x_2^2
- `X_train_poly = poly.fit_transform(X_train)`: Aplicamos la transformación a las características de entrenamiento.
- `X_test_poly = poly.transform(X_test)`: Aplicamos la transformación a las características de prueba.

3.6. Resumen de las Métricas utilizadas para la Comparación

A la hora de comparar los modelos tanto de sketching como el normal usaremos las siguinets métricas:

- **MAE**: Promedio de las diferencias absolutas entre las predicciones y los valores reales.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

- **MSE**: Promedio de los cuadrados de las diferencias, penalizando más los errores grandes.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- **R^2**: Mide la cantidad de varianza explicada por el modelo.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

  Donde $\bar{y}$ es la media de los valores reales. Este coeficiente mide qué proporción de la varianza total de los datos es explicada por el modelo.

- **Tiempo de ejecución**: El modelo con sketching es significativamente más rápido que el modelo normal, lo que lo hace más eficiente para grandes conjuntos de datos.

3.7. Reducción de Dimensionalidad con SVD (Sketching)

```
n_components = 2
U, Sigma, VT = randomized_svd(X_train_poly, n_components=n_components,
    random_state=42)

B = np.dot(U, np.diag(Sigma))

original_variance = np.var(X_train_poly, axis=0).sum()
sketched_variance = np.var(B, axis=0).sum()
explained_variance_ratio = (sketched_variance / original_variance) * 100
```

- `U, Sigma, VT = randomized_svd(X_train_poly, n_components=2, random_state=42)`: Usamos SVD para reducir la dimensionalidad de las características polinómicas. SVD descompone una matriz en tres componentes: `U`, `Sigma`, y `VT`. Con `n_components=2`, estamos reduciendo las características a solo dos componentes principales.

- `B = np.dot(U, np.diag(Sigma))`: Construimos la matriz reducida `B` a partir de los componentes obtenidos por SVD.
- `original_variance = np.var(X_train_poly, axis=0).sum()`: Calculamos la varianza total de las características originales.
- `sketched_variance = np.var(B, axis=0).sum()`: Calculamos la varianza total de la matriz reducida.
- `explained_variance_ratio = (sketched_variance / original_variance) * 100`: Calculamos el porcentaje de la varianza explicada por la matriz reducida en nuestro caso sería de 99.843 % lo cual es practicamente su totalidad por lo que la matriz reducida representa el conjuto original de manera casi perfecta.

El SVD reduce la dimensionalidad de los datos, lo que facilita el procesamiento y mejora la eficiencia del modelo, especialmente en conjuntos de datos grandes.

3.8. Entrenamiento y Evaluación del Modelo de Regresión Lineal

```
model_normal = LinearRegression()
model_normal.fit(X_train_poly, y_train)
y_pred_normal = model_normal.predict(X_test_poly)

mae_normal = mean_absolute_error(y_test, y_pred_normal)
mse_normal = mean_squared_error(y_test, y_pred_normal)
r2_normal = r2_score(y_test, y_pred_normal)
time_normal = end_time - start_time
```

- `model_normal = LinearRegression()`: Creamos un modelo de regresión lineal.
- `model_normal.fit(X_train_poly, y_train)`: Entrenamos el modelo con las características polinómicas.
- `y_pred_normal = model_normal.predict(X_test_poly)`: Realizamos predicciones en el conjunto de prueba.
- `mae_normal = mean_absolute_error(y_test, y_pred_normal)`: Calculamos el Error Absoluto Medio (MAE).
- `mse_normal = mean_squared_error(y_test, y_pred_normal)`: Calculamos el Error Cuadrático Medio (MSE).
- `r2_normal = r2_score(y_test, y_pred_normal)`: Calculamos el Coeficiente de Determinación (R^2).
- `time_normal = end_time - start_time`: Calcula el tiempo transcurrido durante el entrenamiento

Estas métricas de evaluación nos permiten medir la precisión del modelo y su capacidad para explicar la variabilidad de los datos.

3.9. Modelo con Sketching

```
model_sketch = LinearRegression()
model_sketch.fit(B, y_train)
X_test_sketch = np.dot(X_test_poly, VT.T)
y_pred_sketch = model_sketch.predict(X_test_sketch)

mae_sketch = mean_absolute_error(y_test, y_pred_sketch)
mse_sketch = mean_squared_error(y_test, y_pred_sketch)
r2_sketch = r2_score(y_test, y_pred_sketch)
time_sketch = end_time - start_time
```

- `model_sketch = LinearRegression()`: Creamos un modelo de regresión lineal utilizando la matriz reducida `B`.
- `model_sketch.fit(B, y_train)`: Entrenamos el modelo con las características reducidas.
- `y_pred_sketch = model_sketch.predict(X_test_sketch)`: Realizamos predicciones con el modelo de sketching.
- `time_sketch = end_time - start_time`: Calcula el tiempo transcurrido durante el entrenamiento

3.10. Tabla comparativa

```
print("-" * 80)
print(f"{'Modelo':<20} | {'MAE':^12} | {'MSE':^12} | {'R^2':^12} | {'Tiempo (s)':^12}")
print("-" * 80)
print(f"{'Normal':<20} | {mae_normal:^12.3e} | {mse_normal:^12.3e} | {r2_normal:^12.3e} | {time_normal:^12.4f}")
print(f"{'Sketching':<20} | {mae_sketch:^12.3e} | {mse_sketch:^12.3e} | {r2_sketch:^12.3e} | {time_sketch:^12.4f}")
print("-" * 80)
```

- `Error Absoluto Medio (MAE)`: $4{,}327 * 10^8$, lo que indica un error de predicción extremadamente alto.Esto sugiere que las predicciones del modelo están muy alejadas de los valores reales.
- `Error Cuadrático Medio (MSE)`: $4{,}390 * 10^{20}$, lo que refuerza la idea de que los errores son masivos y que las predicciones del modelo normal son inadecuadas para los datos.
- `Coeficiente de Determinación (R²)` : $-1{,}099 * 10^{14}$, un valor negativo extremadamente grande. Un R^2 negativo indica que el modelo es peor que un modelo que simplemente predice el promedio de los datos.
- `Tiempo de Ejecución`: $1{,}716$ segundos. Aunque no es excesivamente lento, el tiempo de ejecución es significativamente mayor que el del modelo Sketching.

  El modelo Sketching muestra un rendimiento mucho mejor en comparación con el modelo normal:

- `Error Absoluto Medio (MAE)`: $1{,}759 * 10^3$, lo que indica que las predicciones están mucho más cerca de los valores reales en comparación con el modelo normal.
- `Error Cuadrático Medio (MSE)`: $3{,}994 * 10^6$, un valor mucho más bajo que el del modelo normal, sugiriendo un ajuste considerablemente más preciso.
- `Coeficiente de Determinación (R²)` :$-5.032 * 10^{-6}$, aunque sigue siendo ligeramente negativo (indicando que no es perfecto), el valor está muy cerca de cero, lo que implica un rendimiento razonable en comparación con el modelo normal.
- `Tiempo de Ejecución`: $0{,}008$ segundos. El modelo Sketching es extremadamente rápido, completando el proceso en una fracción de tiempo comparado con el modelo normal.

En resumen, el modelo normal parece no ajustarse correctamente a los datos y presenta predicciones altamente erroneas.El modelo de sketching, al usar una versión reducida de las características, mejora la eficiencia y velocidad del proceso de entrenamiento y predicción. Aunque el R^2 del modelo Sketching aún no alcanza valores positivos, su rendimiento es miles de veces mejor que el del modelo normal.

La conclusión final es que el modelo con sketching muestra un mejor rendimiento en términos de precisión y eficiencia, siendo más adecuado para manejar grandes cantidades de datos y tareas de predicción. Aunque es recomendable añadirle pequeños ajustes adicionales para seguir mejorando su precisión, como elevar el grado del polinomio o hacerlo por steps.

4. Optimización y Propuesta de Solución para la Gestión del Tráfico Vehicular

El análisis de los resultados del modelo predictivo revela una relación significativa entre el volumen de tráfico vehicular y factores como el clima y los días feriados. Estos hallazgos sugieren la necesidad de estrategias proactivas para gestionar el tráfico de manera eficiente, reduciendo la congestión y mejorando la experiencia de movilidad en la ciudad.

4.1. Propuestas de Solución

4.1.1. Semáforos Adaptativos

Implementar semáforos inteligentes capaces de ajustar sus ciclos en tiempo real en función del flujo vehicular, las condiciones climáticas y la presencia de eventos especiales. Este enfoque puede reducir el tiempo de espera y mejorar la fluidez del tráfico en intersecciones clave.

4.1.2. Aplicaciones Móviles de Rutas Alternativas

Desarrollar o integrar aplicaciones que proporcionen recomendaciones de rutas alternativas basadas en predicciones de tráfico y factores como lluvias o baja visibilidad. Estas aplicaciones podrían alertar a los conductores en tiempo real, optimizando el flujo vehicular y minimizando los tiempos de viaje.

4.1.3. Mejoras en el Transporte Público

Incrementar la infraestructura y la frecuencia del transporte público, especialmente en días feriados o bajo condiciones climáticas adversas. Esto podría incentivar el uso de opciones de transporte colectivo, reduciendo la cantidad de vehículos particulares en las vías.

4.1.4. Horarios Escalonados y Flexibles

Colaborar con empresas e instituciones para implementar horarios laborales escalonados y flexibles. Esta medida disminuiría el volumen de tráfico durante las horas pico, distribuyendo mejor la carga vehicular a lo largo del día.

4.1.5. Carriles Reversibles y Auxiliares

En situaciones de tráfico extremo, considerar la apertura de carriles reversibles para optimizar el uso de las vías disponibles. En casos de necesidad crítica, también podría explorarse la posibilidad de reducir temporalmente las aceras para habilitar carriles auxiliares destinados al flujo vehicular adicional.

4.2. Conclusión

La implementación de estas soluciones podría mejorar significativamente la gestión del tráfico vehicular en la ciudad, reduciendo tiempos de viaje y el impacto ambiental asociado con la congestión. Estas propuestas, sustentadas en el análisis de datos y tendencias, constituyen un enfoque integral para abordar los desafíos actuales de movilidad urbana.