

EV Ad Strategy - EDA

José Manuel Rodríguez Vélez

2024-11-11

EDA (Exploratory Data Analysis)

Repository: https://github.com/marthinal/ev_ad_strategy

1. Cargar conjunto de datos.

El primer paso en el Análisis Exploratorio de Datos (EDA) es cargar el dataset para poder comenzar con la exploración y análisis. En este caso, estoy utilizando el archivo CSV que contiene los patrones de carga de vehículos eléctricos. Ver el script en `~/ev_ad_strategy/scripts/load_and_inspect_dataset.R`

2. Exploración Inicial de Datos con `skimr`

Utilicé el paquete `skimr` para llevar a cabo una primera investigación del dataset. `skimr` permite obtener un resumen detallado y compacto, proporcionando información clave como:

- **Tipo de variable:** Identifica si es numérica o categórica.
- **Valores faltantes:** Indica cuántos datos están ausentes en cada columna.
- **Estadísticas descriptivas:** Muestra media, mediana, desviación estándar y percentiles.
- **Distribución:** Proporciona un histograma simplificado para variables numéricas.

Ver `~/ev_ad_strategy/html_report/skim_summary.html`

Análisis de datos faltantes y duplicados

En mi análisis inicial del dataset, identifiqué varios valores faltantes (missing values) y valores duplicados. A continuación, detallo mis observaciones:

1. Valores faltantes

- Existen **66 valores faltantes** en cada una de las siguientes columnas:
 - `Energy.Consumed..kWh.`
 - `Charging.Rate..kW.`
 - `Distance.Driven..since.last.charge...km.`
- Sin embargo, estas variables no serán utilizadas en mi análisis, ya que mi enfoque principal se centrará en otras variables.
- Las variables que voy a utilizar para mi caso de uso específico **no presentan valores faltantes** (`n_missing = 0`). Las variables seleccionadas son:

- **UserType**: para identificar el tipo de usuario (especialmente los de larga distancia).
- **DayOfWeek**: para analizar los días con mayor afluencia.
- **Charging.Station.Location**: para segmentar la audiencia por ciudad.
- **Charging.End.Time**: para identificar los horarios óptimos de carga y, por ende, el mejor momento para mostrar anuncios.
- **Charger.Type**: para filtrar por el tipo de cargador.

2. Valores duplicados en tiempos de carga

- Se detectaron **11 valores duplicados** en **Charging.End.Time**, lo cual no implica automáticamente un error. Dado que los tiempos están redondeados a minutos, es posible que múltiples sesiones de carga hayan terminado al mismo minuto.
- Esto puede ser normal, especialmente en estaciones concurridas donde varios vehículos terminan de cargar simultáneamente. No eliminaré estos duplicados en esta etapa, ya que no afectan a mi análisis principal.

Justificación de las variables seleccionadas para el caso de uso de anuncios de hoteles

El objetivo del análisis es optimizar la colocación de anuncios de hoteles en estaciones de carga para usuarios de larga distancia.

Las variables seleccionadas me permiten cubrir los siguientes aspectos clave:

- **UserType**: Me permite identificar a los usuarios que realizan viajes de larga distancia, quienes son el público objetivo para los anuncios de hoteles.
- **DayOfWeek**: Ayuda a determinar los días de la semana con mayor afluencia de usuarios, para así optimizar el impacto del anuncio.
- **Charging.Station.Location**: Permite segmentar los anuncios por ciudad, mostrando ofertas específicas según la ubicación de la estación de carga.
- **Charging.End.Time**: Indica un tiempo de espera potencial en la estación, y por tanto, el momento ideal para mostrar los anuncios.
- **ChargerType**: Identifica el tipo de cargador utilizado (Level 1, Level 2 o DC Fast Charger). Esto es importante para estimar el tiempo promedio de carga, ya que los cargadores rápidos implican tiempos de espera más cortos, mientras que los cargadores de nivel 1 y 2 permiten una mayor exposición al anuncio debido a tiempos de carga más prolongados.

Justificación para no eliminar valores faltantes y outliers

En mi análisis inicial, identifiqué tanto valores faltantes como outliers en algunas variables del dataset. Sin embargo, considero que para el alcance específico de este proyecto y las preguntas de investigación planteadas, no es necesario realizar una eliminación de estos datos. A continuación, detallo mi justificación:

Valores faltantes

Los valores faltantes identificados corresponden a las siguientes variables:

- **Energy.Consumed..kWh.**
- **Charging.Rate..kW.**
- **Distance.Driven..km.**

Estos valores faltantes no afectan directamente a mi análisis porque las variables seleccionadas para responder a la pregunta del caso de uso (optimización de anuncios de hoteles) no incluyen estas columnas. En otras palabras:

- **Energy.Consumed..kWh.:** La cantidad de energía consumida durante la carga no es relevante para determinar el momento adecuado para mostrar un anuncio, ya que mi análisis se centra en el comportamiento del usuario y no en el consumo energético.
- **Charging.Rate..kW.:** La velocidad de carga tampoco influye en el tiempo de espera del usuario, ya que este tiempo puede variar por otros factores. Además, no afecta a las variables que he seleccionado para mi análisis.
- **Distance.Driven..km.:** Aunque puede haber outliers en esta variable, no afecta mi caso de uso porque no estoy analizando el comportamiento de los usuarios en función de la distancia recorrida, sino en función del tiempo de carga y la localización.

Por estas razones, he decidido no eliminar las filas con valores faltantes. Eliminar estas filas reduciría el número de datos útiles y podría sesgar el análisis, especialmente si los demás datos en la misma fila son válidos y relevantes para las variables que estoy utilizando.

Outliers detectados

En cuanto a los outliers, identifiqué los siguientes casos:

1. **State.of.Charge..Start... y State.of.Charge..End...:** Algunos valores superan el 100%, lo cual no es físicamente posible. Sin embargo, considero que este problema puede ser un error en el sensor del vehículo o en el registro del dato, y no necesariamente implica que toda la fila de datos sea inválida. Para el propósito de mi análisis, que se enfoca en identificar el momento óptimo para mostrar anuncios, estos outliers no tienen un impacto significativo, ya que no estoy utilizando el estado de carga como variable clave.
2. **Distance.Driven..km.:** El valor máximo detectado es de **398.36 km**, que es significativamente superior al percentil 75 (**226.07 km**). Este valor podría ser un outlier, pero como no estoy utilizando esta variable en mi análisis, no afectará mis resultados.

Conclusión

He decidido mantener los valores faltantes y los outliers en el dataset porque:

- Las variables seleccionadas para mi análisis no presentan incidencias de valores faltantes ni outliers.
- El análisis se enfoca en cinco variables clave, las cuales no presentan problemas significativos que justifiquen una limpieza exhaustiva de datos.
- El objetivo de este análisis es demostrar un caso de uso específico para la optimización de anuncios en estaciones de carga. Limitar el alcance a estas variables permite desarrollar una solución concreta y escalable, sin añadir complejidad innecesaria en esta fase inicial.

Considero que este enfoque es adecuado para responder a la pregunta planteada y cumple con los objetivos de la práctica, manteniendo un balance entre la calidad de los datos y la eficiencia del análisis.

Reflexión sobre la escalabilidad del análisis y consideraciones adicionales

Si bien he justificado la decisión de no eliminar los valores faltantes y outliers en este análisis inicial, es importante considerar que este enfoque está alineado con el alcance específico de esta práctica y las preguntas de investigación planteadas. Sin embargo, en un escenario donde se escale el análisis o se añadan nuevas variables, será necesario reevaluar esta decisión.

Escalabilidad y análisis comparativo

En el caso de ampliar el análisis en el futuro con nuevas variables que puedan estar afectadas por outliers o valores faltantes, sería recomendable llevar a cabo un análisis comparativo, tanto con como sin estos valores problemáticos. Este análisis podría ayudar a determinar cómo afectan los outliers y los missing values a las medidas de tendencia central, dispersión y a los resultados generales.

Estrategia para futuras expansiones del análisis

1. **Evaluación de impacto:** Antes de realizar cualquier limpieza de datos, se debe evaluar si los outliers o los valores faltantes afectan de forma significativa las conclusiones del análisis. Esto puede hacerse comparando los resultados con y sin estas incidencias.
2. **Decisiones basadas en evidencia:** Dependiendo del impacto observado, se pueden tomar diferentes acciones:
 - **Eliminar outliers o imputar valores:** Si se identifican como errores claros o datos anómalos que distorsionan los resultados.
 - **Mantener los valores:** Si se determina que los outliers son casos representativos del fenómeno que se está estudiando y no distorsionan el análisis.

Justificación del uso de Charging End Time en lugar de Charging Start Time

He decidido utilizar `Charging.End.Time` en lugar de `Charging.Start.Time` debido a inconsistencias detectadas en los datos de esta última variable.

El principal motivo es que `Charging.Start.Time` sigue un patrón artificial en el que la hora se incrementa de manera constante para todos los usuarios. Por ejemplo:

- El 1 de enero de 2024, un registro comienza a las 12:00, el siguiente a la 1:00, el siguiente a las 2:00, y así sucesivamente, continuando de forma similar al día siguiente.
- Este patrón no tiene sentido en términos reales y sugiere que esta columna podría haber sido generada automáticamente, lo que compromete su validez.

Por otro lado, `Charging.End.Time` no sigue este patrón, y aunque no he profundizado completamente en su análisis, parece tener más sentido. Este razonamiento se basa en que:

1. **Inconsistencia en los tiempos de carga:** Los tiempos de carga observados son inusualmente altos y requieren mayor análisis, especialmente en relación con el tipo de cargador utilizado (`ChargerType`) y los tiempos que cabría esperar para cada uno.
2. **Alcance académico:** Dado que este proyecto es un ejercicio académico, utilizar `Charging.End.Time` es una decisión práctica para avanzar en el análisis, dado que este dato parece más realista que el `Charging.Start.Time`.

Si este fuera un caso real, sería crítico investigar a fondo estas inconsistencias antes de continuar. Sin embargo, para el alcance de este proyecto, considero válido trabajar con `Charging.End.Time`, siempre manteniendo en mente las limitaciones de los datos y el contexto en el que se realiza este análisis.

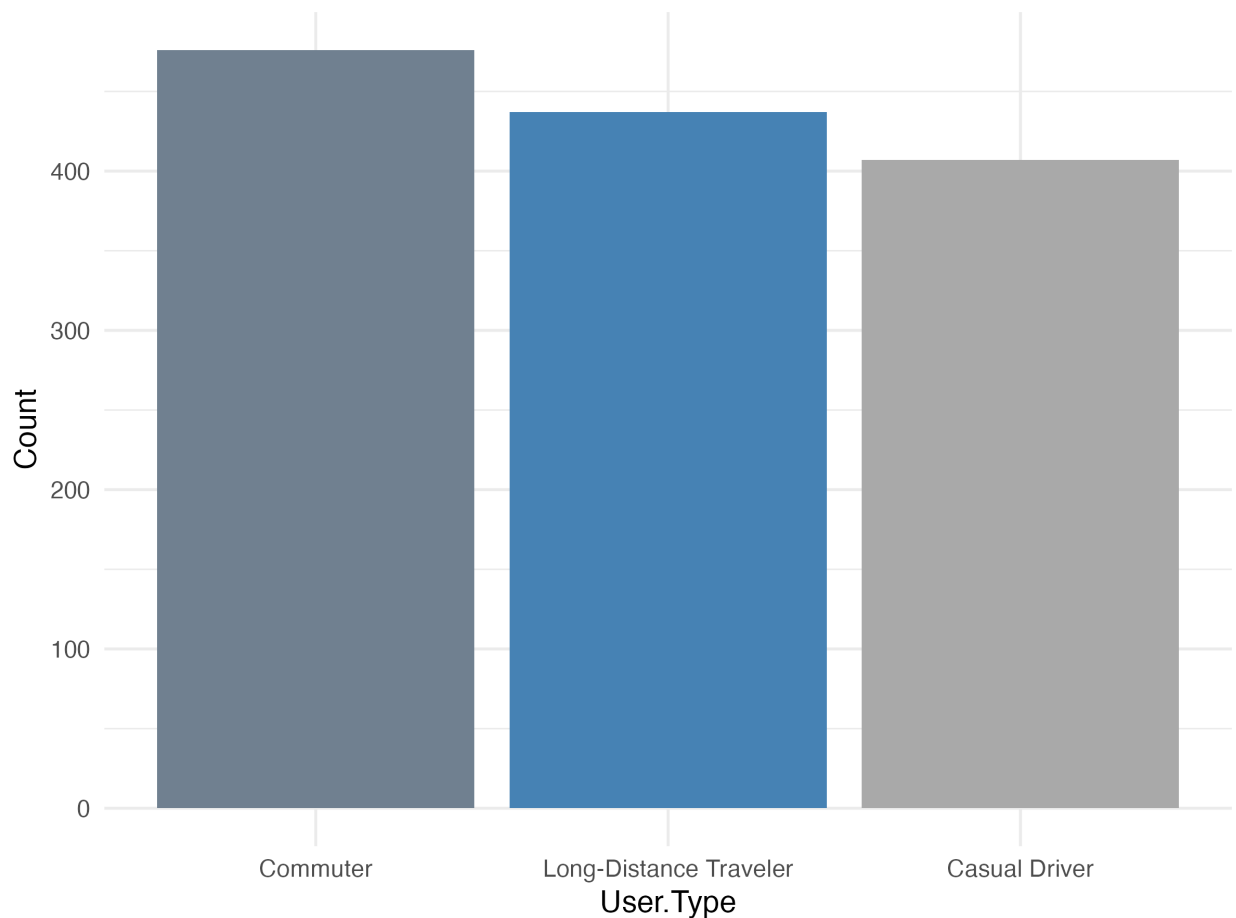
Conclusión

El enfoque adoptado en este proyecto tiene sentido dentro del alcance limitado y específico de esta práctica universitaria, ya que las variables seleccionadas para el análisis no presentan incidencias importantes que requieran una limpieza exhaustiva de datos. Sin embargo, este planteamiento es escalable y podría ajustarse fácilmente en caso de que se amplíe el análisis con nuevas variables. La metodología empleada ha sido diseñada para ser flexible y permitir una revisión crítica de las decisiones tomadas, asegurando así que el análisis se mantenga válido y relevante incluso si se incorporan cambios en etapas futuras.

Análisis Univariado

User Type

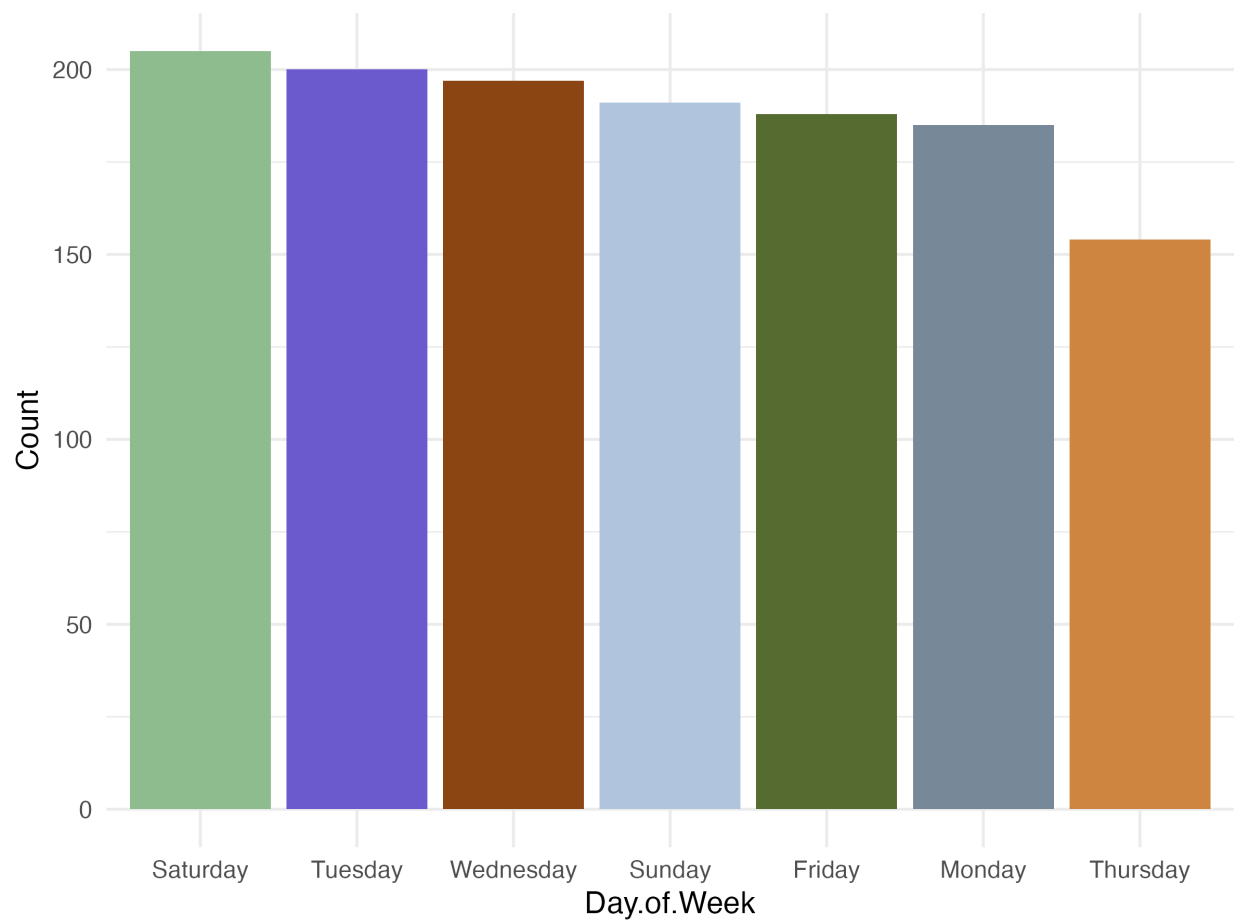
- Commuter 476
- Long-Distance Traveler 437
- Casual Driver 407



- **Commuter:** Representan la mayor proporción en el dataset.
- **Long-Distance Traveler:** Segundo en frecuencia, relevante para análisis de viajes largos.
- **Casual Driver:** La menor frecuencia, pero aún importante para entender patrones generales.

Day of Week

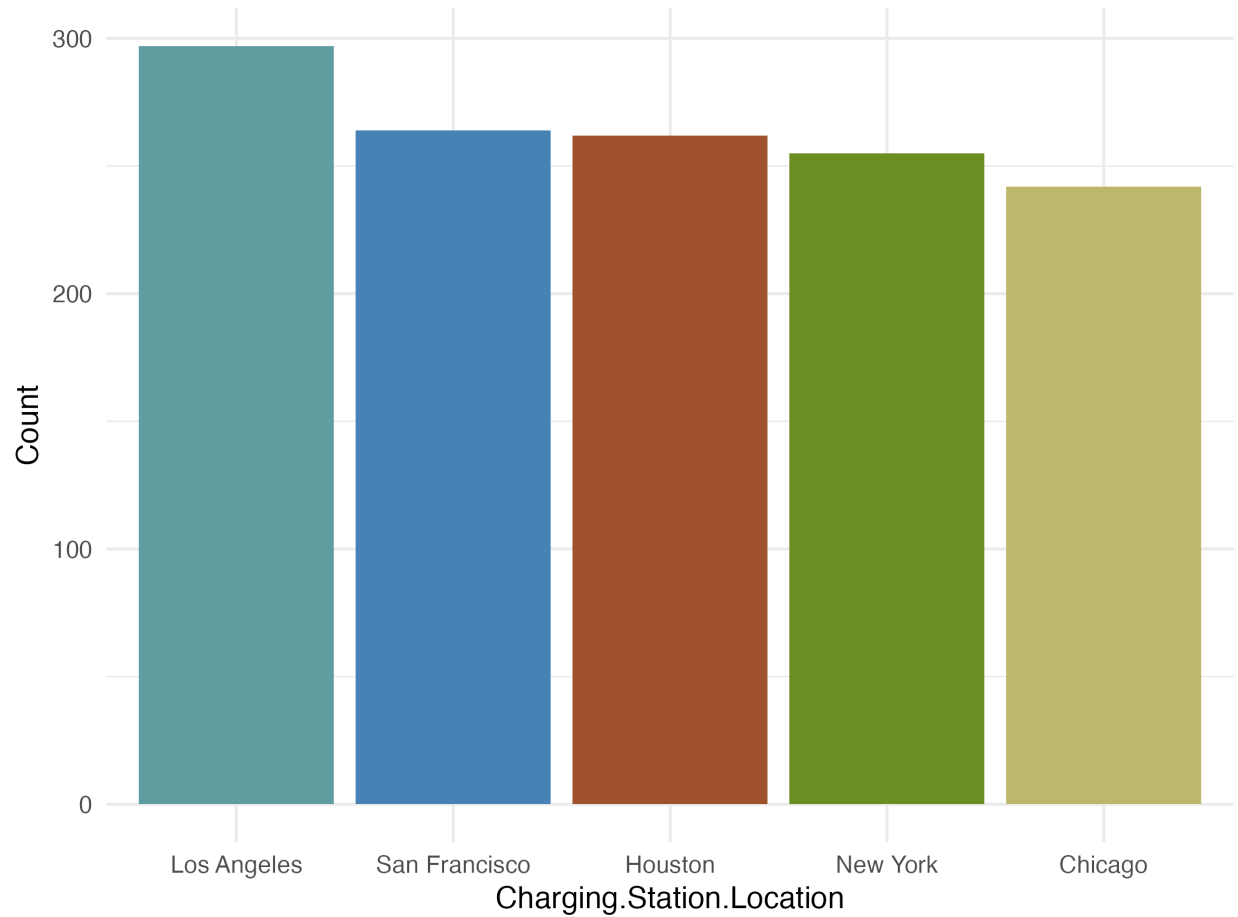
- Saturday 205
- Tuesday 200
- Wednesday 197
- Sunday 191
- Friday 188
- Monday 185
- Thursday 154



- **Saturday y Tuesday** presentan la mayor frecuencia de sesiones de carga, indicando un uso más intenso en estos días.
- **Thursday** tiene la frecuencia más baja, lo que sugiere una menor actividad de carga en comparación con otros días.

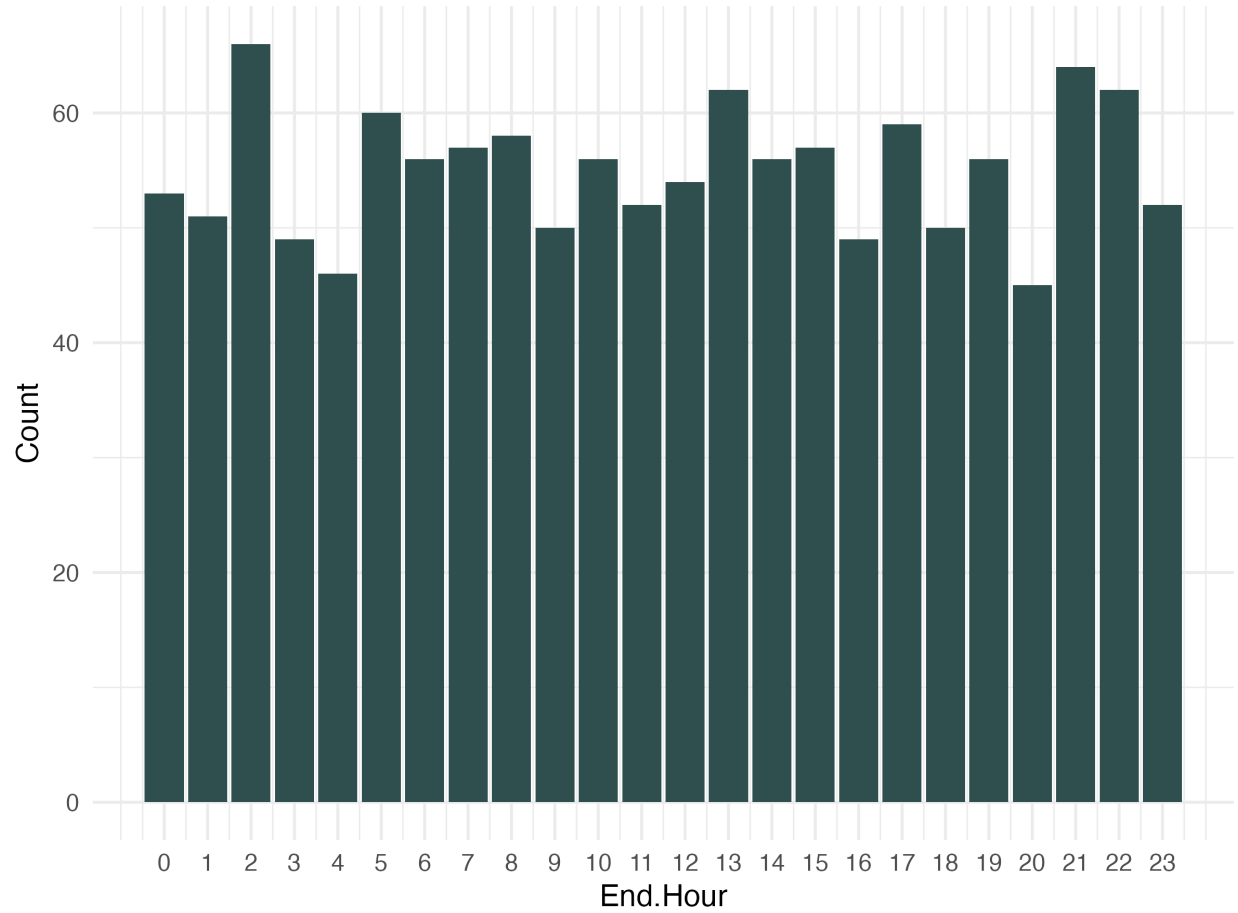
Charging Station Location

- Los Angeles 297
- San Francisco 264
- Houston 262
- New York 255
- Chicago 242

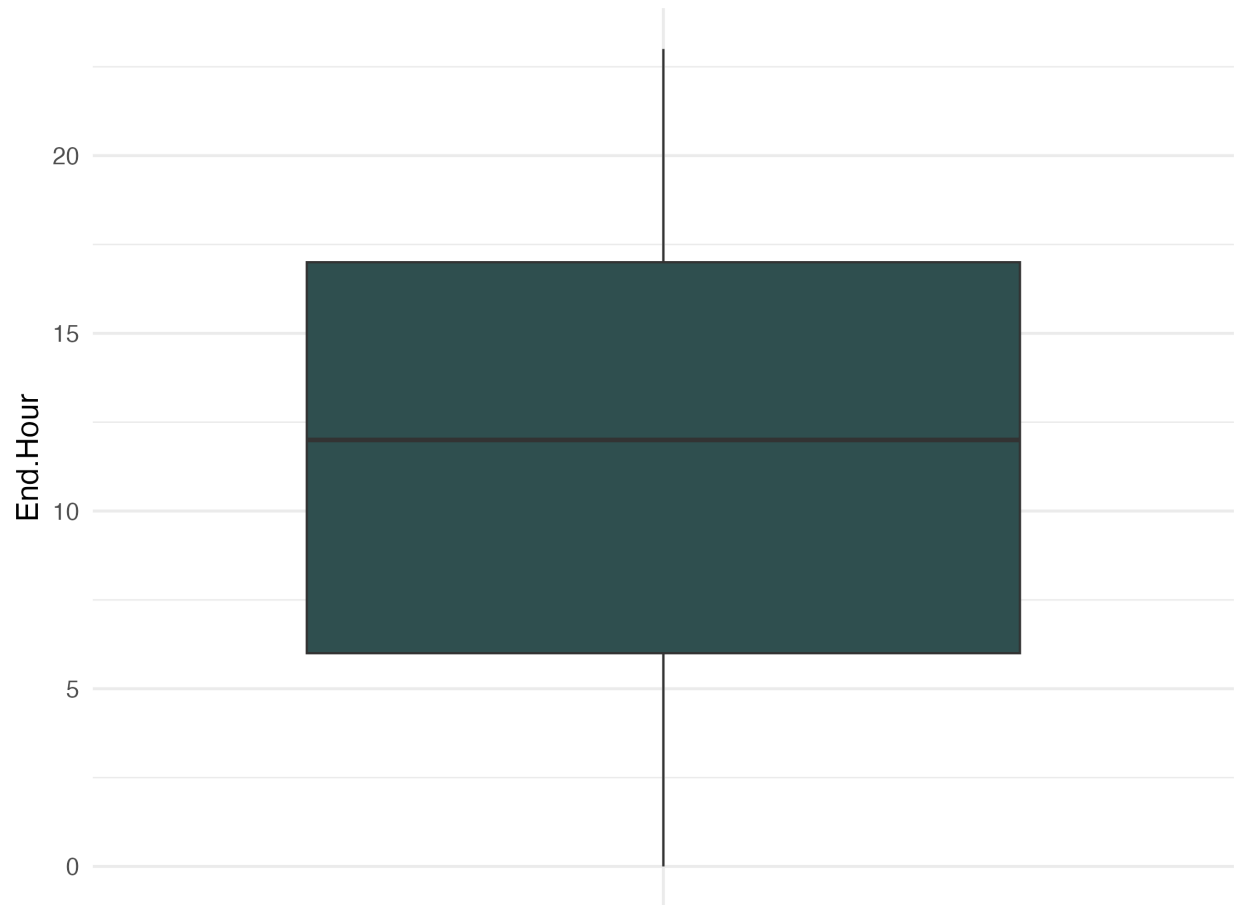


- **Los Ángeles y San Francisco:** Lideran en frecuencia de uso.
- **Chicago:** Presenta menor actividad en comparación con las demás ubicaciones.

End Hour

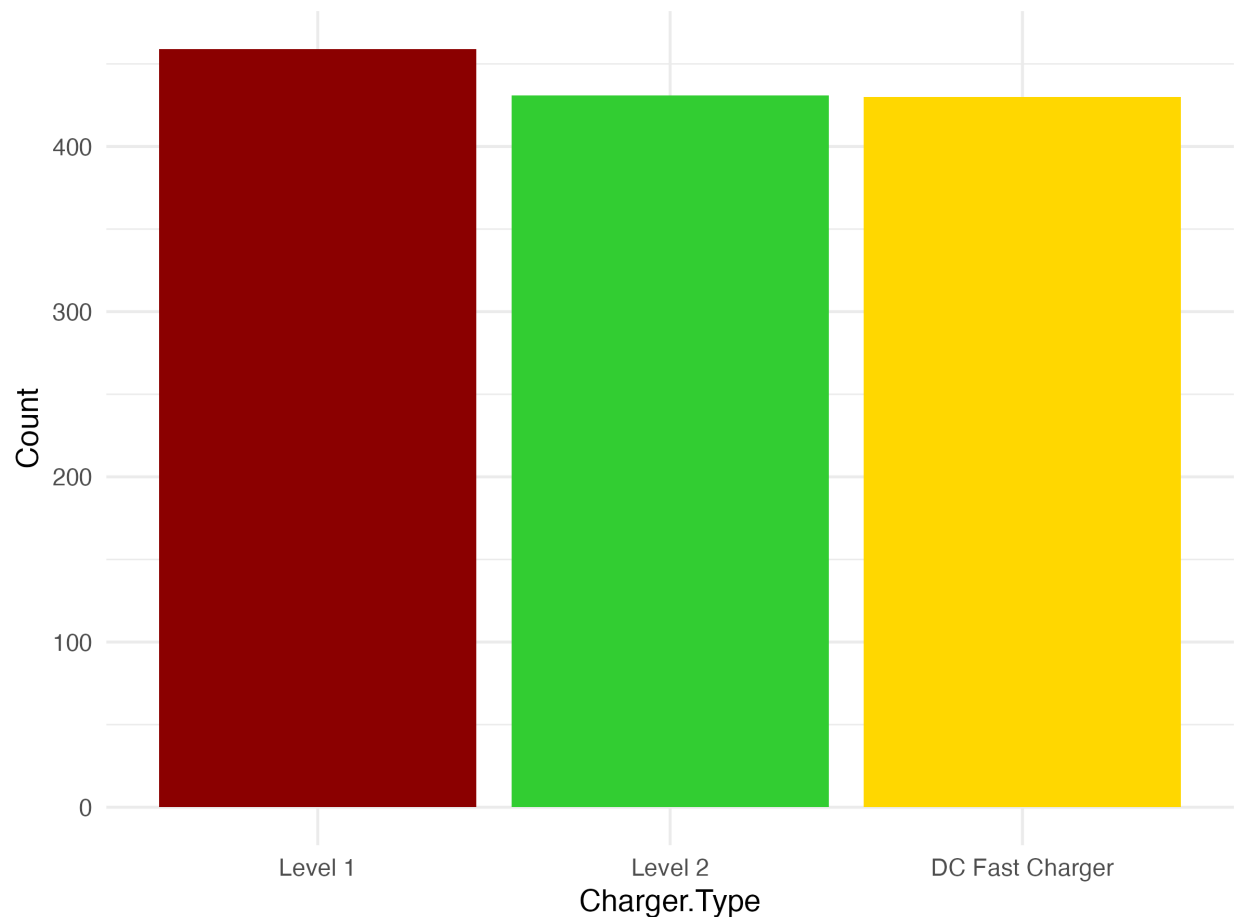


- **2:00 a.m.** destaca como la hora con mayor actividad, representando un pico claro en el comportamiento de los usuarios.
- Las horas de menor actividad se encuentran en la madrugada entre las **3:00 a.m.** y las **5:00 a.m.**
- Durante el resto del día, la actividad es más uniforme, con un aumento hacia las **9:00 p.m.** y **10:00 p.m.**.



El **BoxPlot** de las horas finales de carga muestra que la mediana se encuentra a las 12:00 p.m., con el rango intercuartil (IQR) comprendido entre las 6:00 a.m. (Q1) y las 5:00 p.m. (Q3). Los valores mínimos y máximos están dentro de los bigotes (0:00 a 23:00), lo que indica que **no hay valores atípicos estadísticos**. Aunque el histograma destaca un pico significativo a las 2:00 a.m., este no es considerado un valor atípico en el BoxPlot, sino un comportamiento inusual dentro del rango normal de los datos.

Charger Type



El gráfico de barras muestra la frecuencia de uso de los diferentes tipos de cargadores (**Charger Type**) en las estaciones de carga. Existen tres tipos principales de cargadores representados en el dataset:

- **Level 1:** Este tipo de cargador utiliza una toma de corriente estándar, como la de un hogar, y tiene la velocidad de carga más lenta. Es común para cargas nocturnas o en lugares donde el vehículo estará estacionado por largos períodos.
- **Level 2:** Cargadores más rápidos que utilizan un voltaje más alto, típicamente instalados en estaciones de carga públicas, hogares con instalaciones especiales o en lugares de trabajo.
- **DC Fast Charger:** Este tipo de cargador es el más rápido y eficiente, ya que suministra corriente continua directamente a la batería. Es ideal para viajes largos y estaciones de carga rápida en carretera.

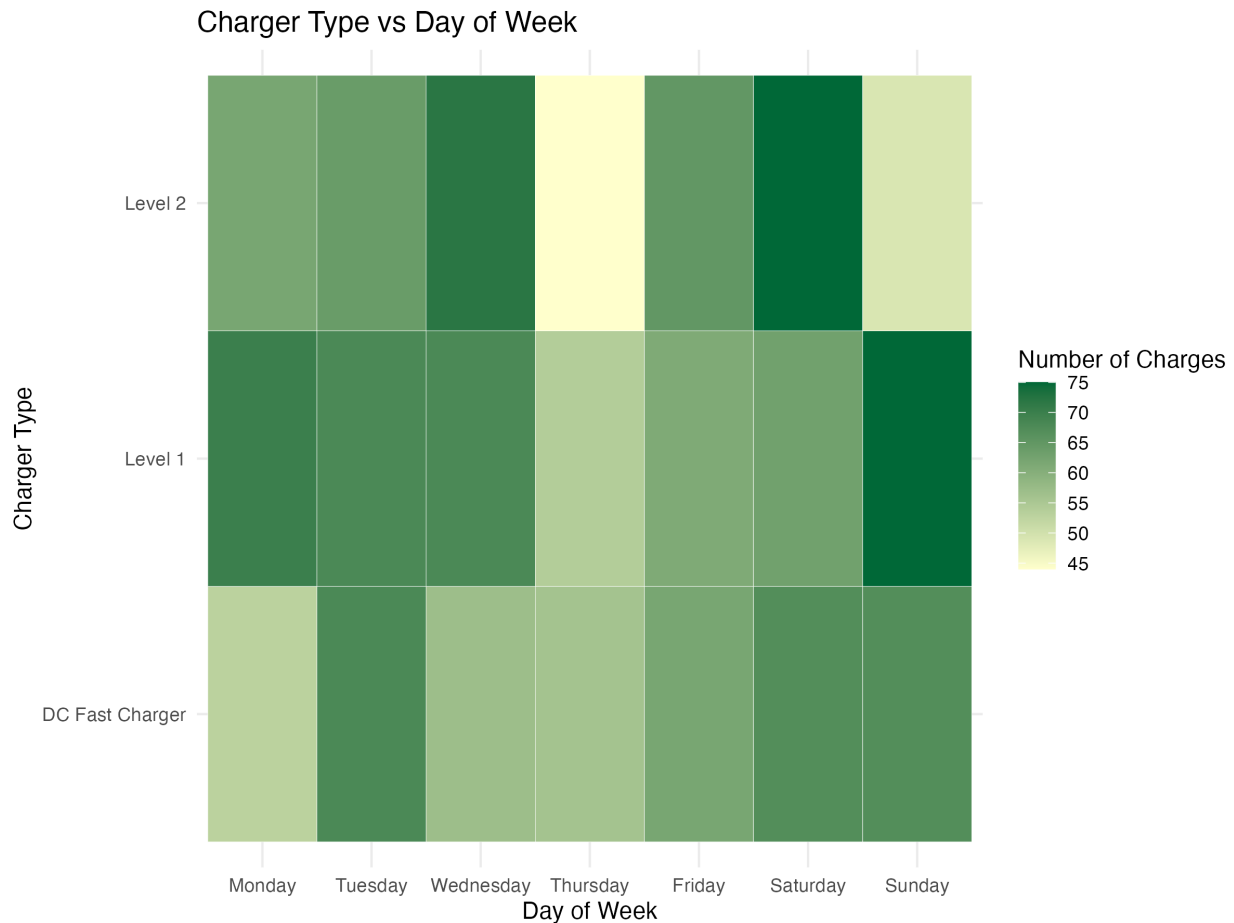
En el gráfico, se observa que:

- **Level 1** es el tipo de cargador más utilizado, posiblemente debido a su disponibilidad en hogares.
- **Level 2** y **DC Fast Charger** tienen frecuencias similares, lo que sugiere un uso balanceado entre velocidades de carga moderadas y rápidas dependiendo de la situación del usuario.

Para explorar los gráficos interactivos y acceder a todos los detalles, puedes visitar el archivo HTML con el reporte completo de análisis univariado en la siguiente ubicación:

Ruta del HTML: `~/ev_ad_strategy/html_report/univariate_analysis.html`

Análisis Bivariado



- **Level 2:**

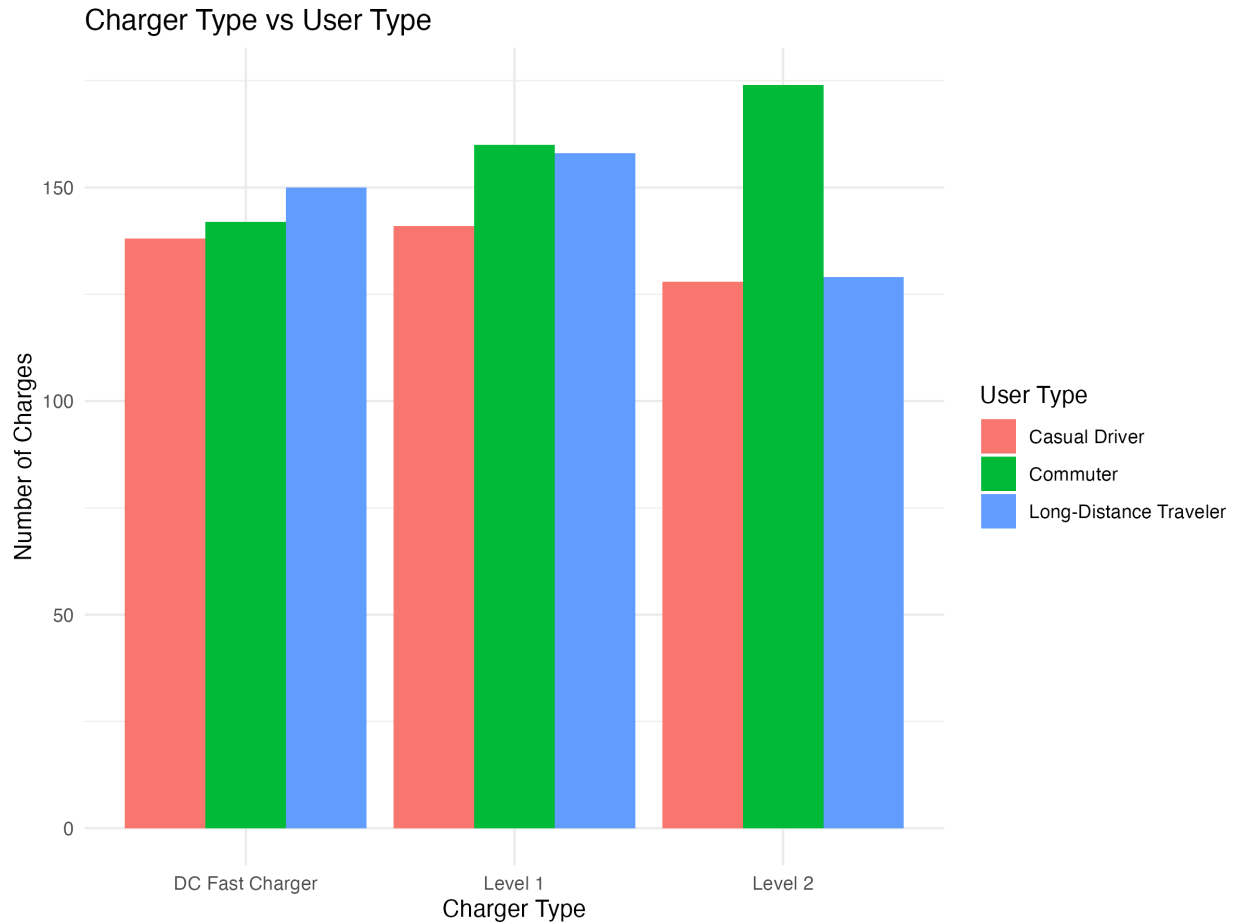
- Los días con mayor número de cargas son el **miércoles** y el **sábado**.
- Los días **lunes** y **martes** presentan un nivel moderado de cargas.
- El **jueves** es el día con la menor cantidad de cargas, seguido del **domingo**, que también tiene una frecuencia relativamente baja.

- **Level 1:**

- El número de cargas es bastante similar los **lunes**, **martes** y **miércoles**, con niveles moderados a altos, destacando ligeramente el **lunes**.
- Hay una disminución el **jueves**, seguida de un leve incremento el **viernes** y el **sábado**.
- El **domingo** es el día con la mayor cantidad de cargas para este tipo de cargador.

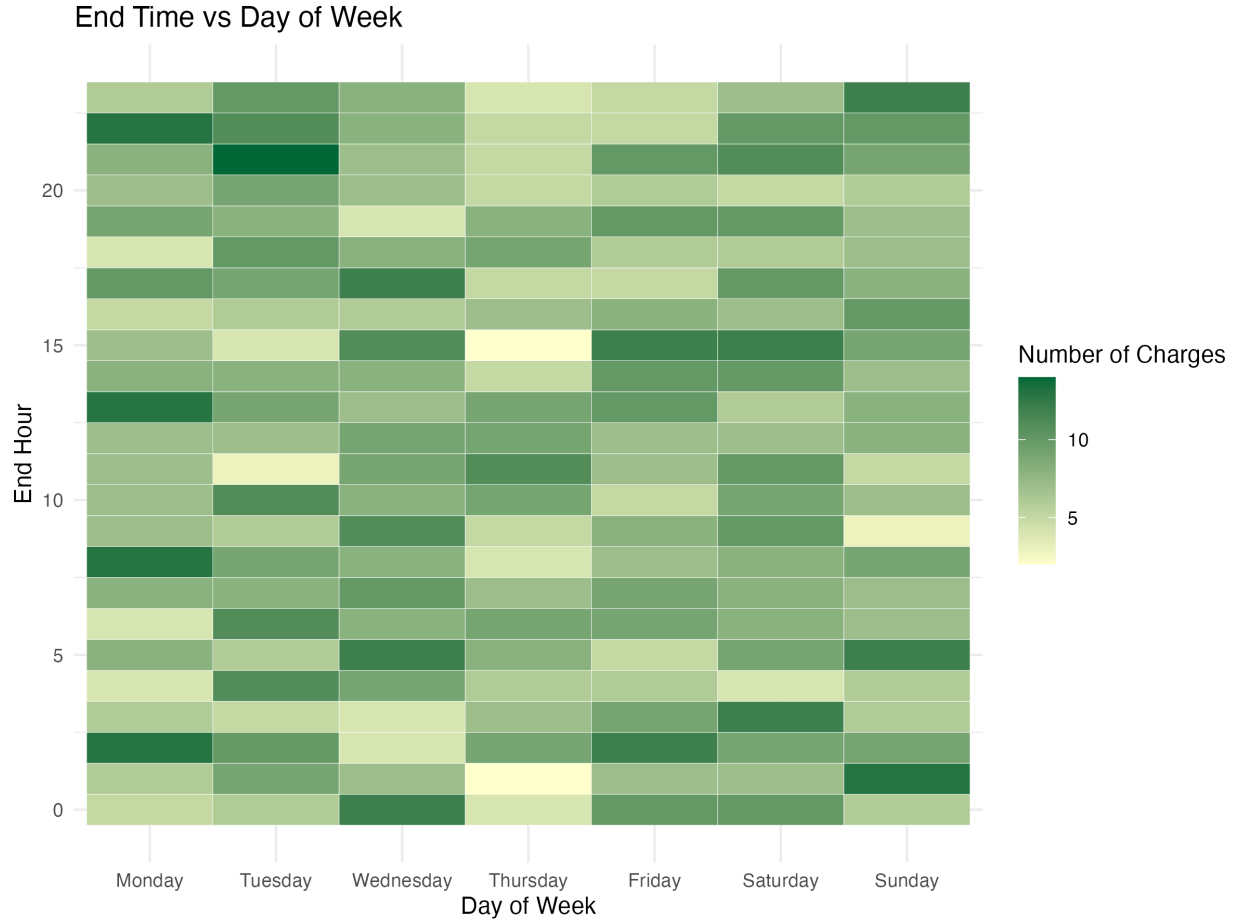
- **DC Fast Charger:**

- El **lunes** tiene un nivel moderado de cargas.
- El **martes** muestra un número de cargas medio-alto.
- El nivel de cargas baja nuevamente a niveles similares al **lunes** el **miércoles**.
- Durante el **viernes** y el fin de semana (**sábado** y **domingo**), se observa una tendencia al alza, aunque muy moderada.



- **DC Fast Charger:**
 - Es utilizado principalmente por los **Long-Distance Traveler**, seguido por los **Commuter** y, finalmente, por los **Casual Driver**.
 - Aunque no hay diferencias extremadamente significativas entre los grupos, los **Long-Distance Traveler** presentan un uso notablemente mayor comparado con los **Casual Driver**.
- **Level 1:**
 - Es el tipo de cargador más utilizado por los **Commuter**, seguido por los **Long-Distance Traveler**.
 - Los **Casual Driver** tienen el menor uso de este tipo de cargador.
- **Level 2:**
 - Es utilizado con mayor frecuencia por los **Commuter**, con una diferencia considerable frente a los otros dos grupos.
 - Los **Casual Driver** y los **Long-Distance Traveler** muestran un uso muy similar, siendo este el cargador menos utilizado por los **Casual Driver**.
- **Patrones generales:**
 - Los **Long-Distance Traveler** tienden a preferir el **Level 1** como su cargador principal, seguido por el **DC Fast Charger**.
 - Los **Casual Driver** muestran una preferencia equilibrada entre el **Level 1** y el **DC Fast Charger**, mientras que el **Level 2** es su menos preferido.

- El **Level 2** destaca como el cargador de mayor preferencia para los **Commuter**, con un uso significativamente superior al de los otros dos tipos.



Lunes: - Picos más altos a las **2:00 a.m.**, **1:00 p.m.**, **8:00 p.m.**, y **10:00 p.m.** - Mínimos observados a las **4:00 a.m.**, **6:00 a.m.**, y **5:00 p.m.**

Martes: - Mayor actividad en la tarde, destacando a partir de las **8:00 p.m.** - El punto más bajo se encuentra a las **11:00 a.m.**

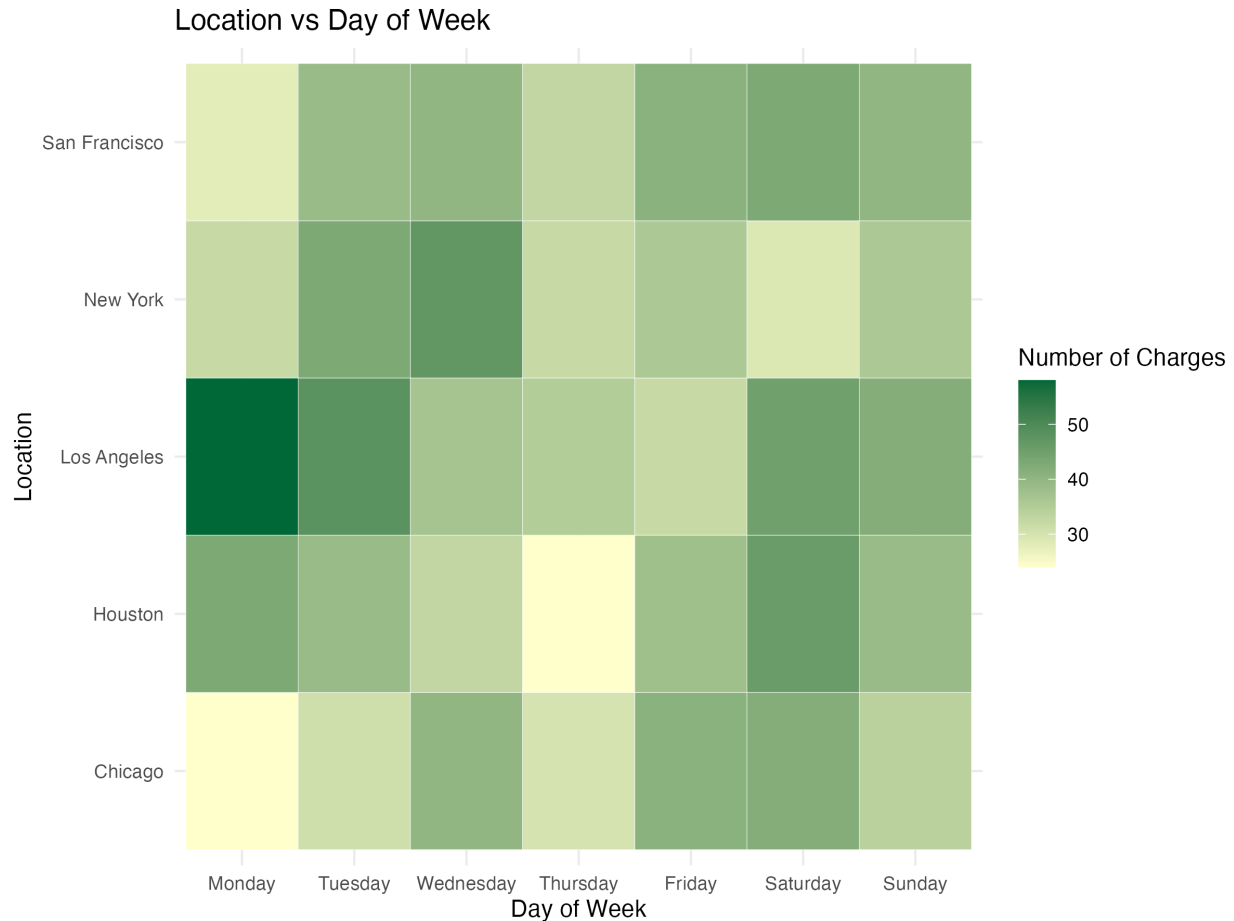
Miércoles: - Pico más alto a la **medianoche (12:00 a.m.)**, con una actividad uniforme el resto del día. - Leve aumento hacia las **5:00 a.m.**

Jueves: - Mínimos entre **12:00 a.m.** y **1:00 a.m.**, con el punto más bajo a la **1:00 a.m.** - Período más activo entre las **10:00 a.m.** y las **2:00 p.m.**, con una caída a las **3:00 p.m.**

Viernes: - Picos a las **2:00 a.m.** y **3:00 p.m.** - Mayor actividad entre la **1:00 p.m.** y las **4:00 p.m.**

Sábado: - Picos a las **3:00 a.m.** y **3:00 p.m.** - Mínimos a las **4:00 a.m.** y **8:00 p.m.**, aunque hay un leve aumento después de esa hora.

Domingo: - Picos principales a la **1:00 a.m.**, **5:00 a.m.**, y **11:00 p.m.** - Actividad más baja a las **9:00 a.m.**



Lunes: - **Los Ángeles** lidera en actividad. - **Houston** tiene una actividad moderada. - **Chicago** registra la menor actividad.

Martes: - **Los Ángeles** sigue siendo la más activa. - **Nueva York** y **San Francisco** tienen actividad moderada. - **Chicago** permanece con la menor actividad.

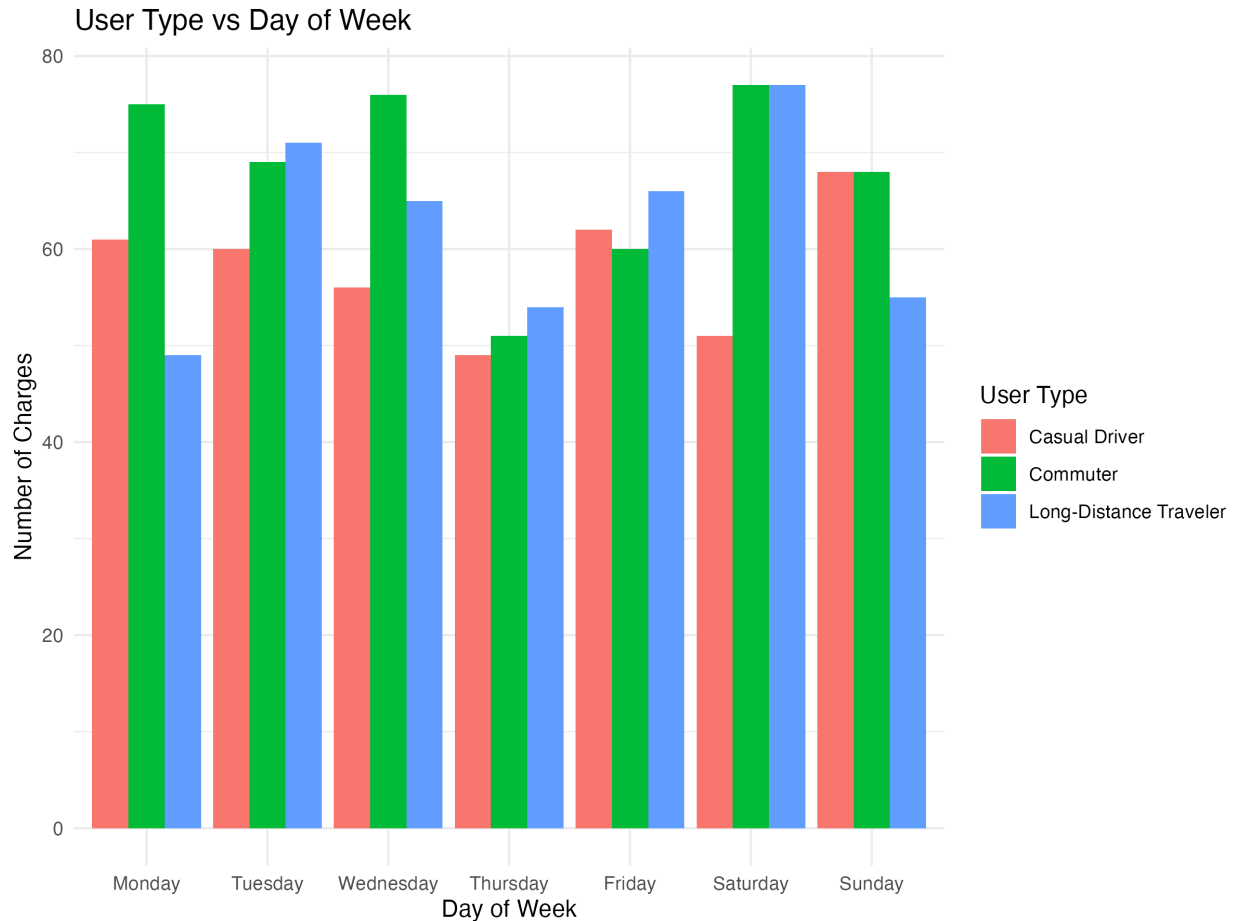
Miércoles: - **Nueva York** presenta la mayor actividad. - **Chicago** es la segunda con mayor actividad, superando a las demás estaciones. - Las otras estaciones tienen niveles moderados.

Jueves: - **Houston** tiene la actividad más baja. - Las demás estaciones muestran actividad baja a moderada.

Viernes: - **Houston**, **Chicago**, y **San Francisco** tienen niveles de actividad moderados. - **Los Ángeles** registra la menor actividad, mientras que **Nueva York** tiene una actividad algo mayor pero aún baja.

Sábado: - **Los Ángeles** y **Houston** destacan como las estaciones más activas. - **Chicago** tiene una actividad moderada. - **Nueva York** registra la menor actividad.

Domingo: - Todas las estaciones muestran niveles moderados de actividad. - **Chicago** tiene la actividad más baja dentro de esta moderación. - **Los Ángeles** y **Houston** destacan levemente como las más activas.



Lunes: - **Commuter** es el más frecuente. - **Long-Distance Traveler** tiene la menor actividad.

Martes: - **Long-Distance Traveler** lidera en actividad. - **Commuter** sigue de cerca. - **Casual Driver** registra la menor frecuencia.

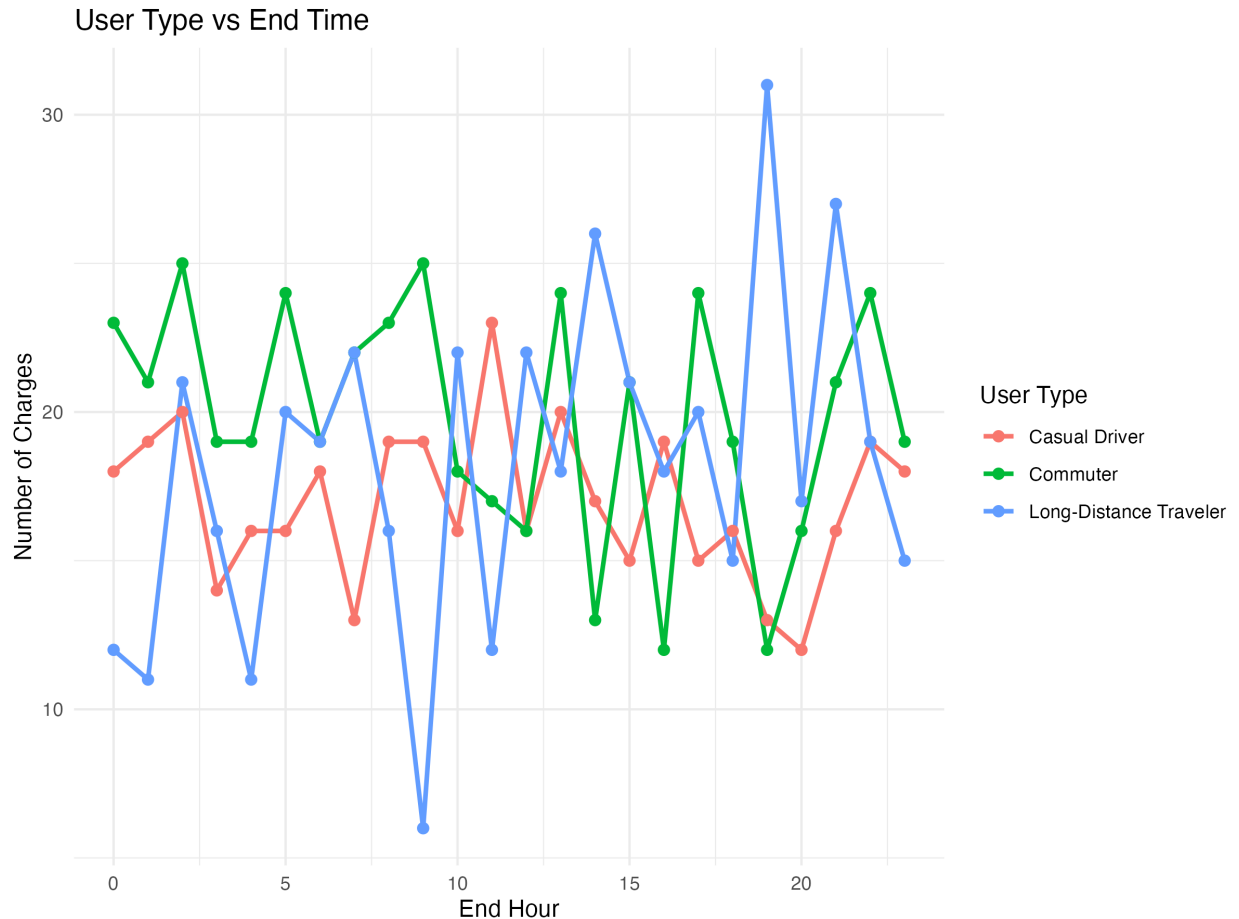
Miércoles: - **Commuter** muestra la mayor actividad, similar al lunes. - **Casual Driver** tiene la frecuencia más baja.

Jueves: - Todos los tipos de usuarios registran una disminución en frecuencia. - **Long-Distance Traveler** lidera, aunque las diferencias entre los tipos son muy pequeñas.

Viernes: - **Long-Distance Traveler** es el más activo. - Las diferencias con los otros tipos de usuarios no son significativas.

Sábado: - **Long-Distance Traveler** y **Commuter** tienen una actividad muy similar y lideran. - **Casual Driver** queda bastante por debajo.

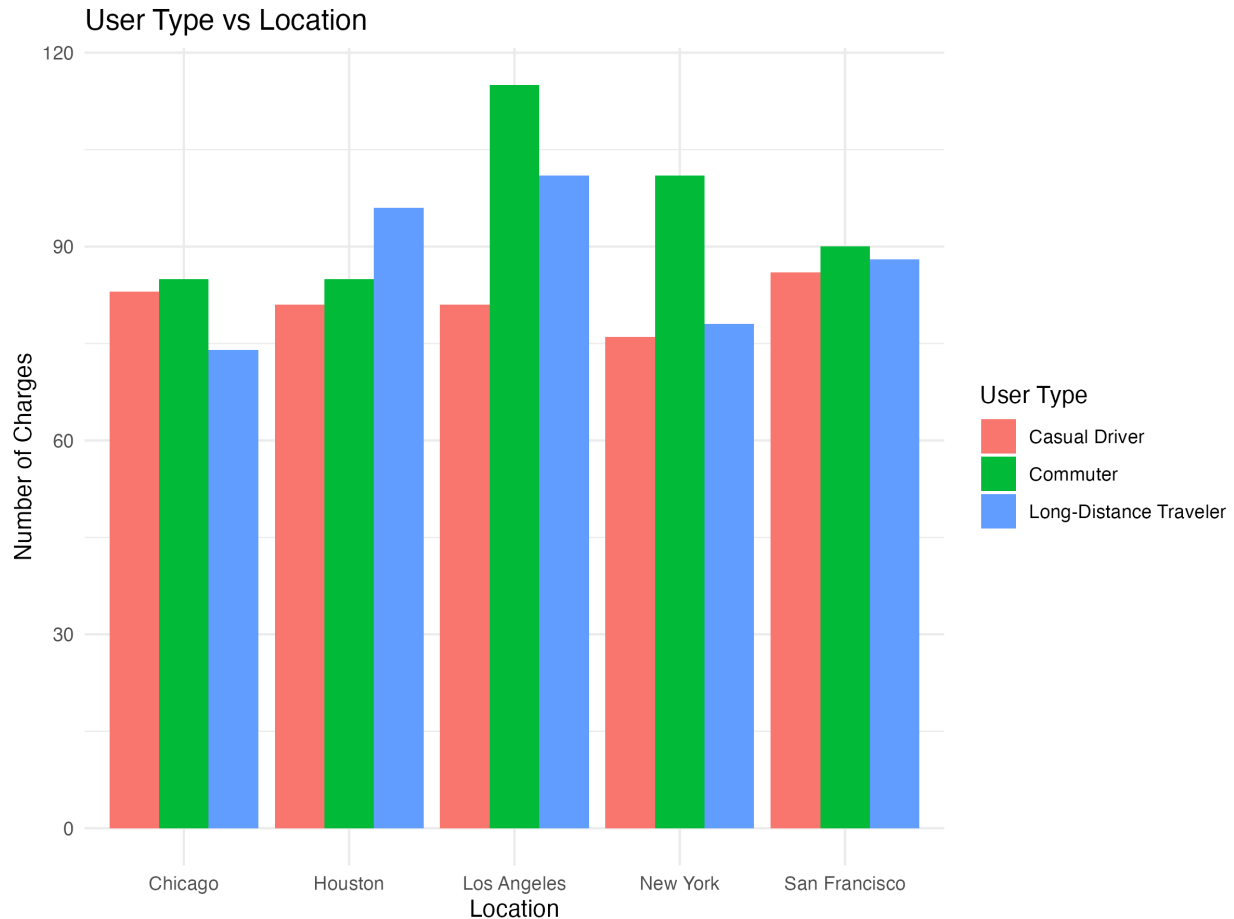
Domingo: - **Commuter** y **Casual Driver** son los más frecuentes. - **Long-Distance Traveler** tiene la menor actividad.



Long-Distance Traveler: - Presenta un patrón de actividad más variable con múltiples picos. - Un pico bajo antes de las **10:00 a.m.** - El pico más alto ocurre cerca de las **7:00 p.m.** - Otro pico destacable aparece antes de las **3:00 p.m.**

Commuter: - Tiene una actividad más estable a lo largo del día. - Más activo en la madrugada, desde las **12:00 a.m.** en adelante. - Muestra mayor frecuencia que el **Casual Driver**, especialmente en horas tempranas de la mañana.

Casual Driver: - Actividad bastante estable en general. - Menor frecuencia en comparación con los otros grupos. - Una disminución notable por la tarde. - Presenta un ligero pico hacia el final de la noche.



Chicago: - **Commuter:** Es el tipo de usuario con mayor frecuencia en esta ciudad. - **Long-Distance Traveler:** Tiene la menor actividad en comparación con los otros tipos de usuario.

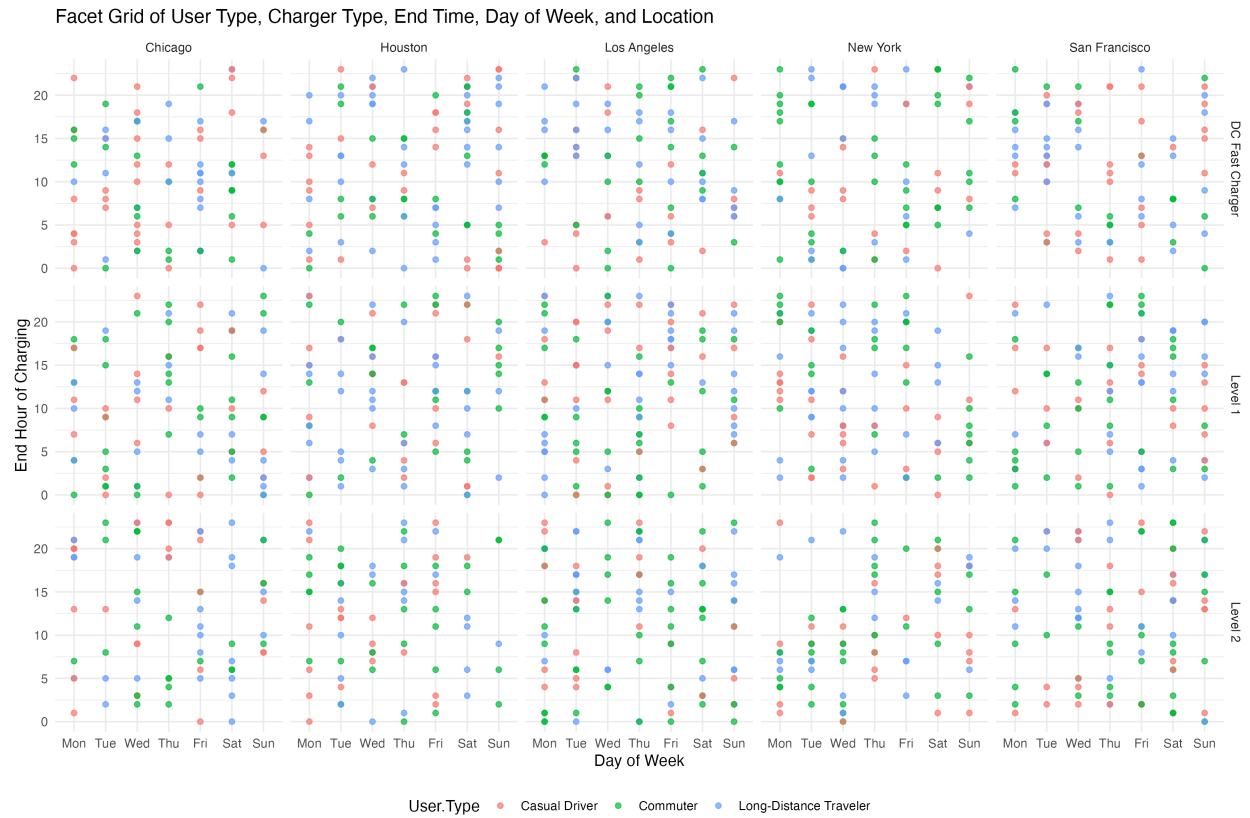
Houston: - **Long-Distance Traveler:** Registra la mayor frecuencia de actividad. - **Casual Driver:** Presenta la menor actividad, con una diferencia considerable frente a los otros grupos.

Los Ángeles: - **Commuter:** Es el más activo de todas las ciudades, con una frecuencia muy alta. - **Long-Distance Traveler:** También tiene una alta frecuencia, siendo la mayor entre este tipo de usuario en todas las ciudades. - **Casual Driver:** Aunque es el grupo menos frecuente, su nivel es comparable con el de otras ciudades.

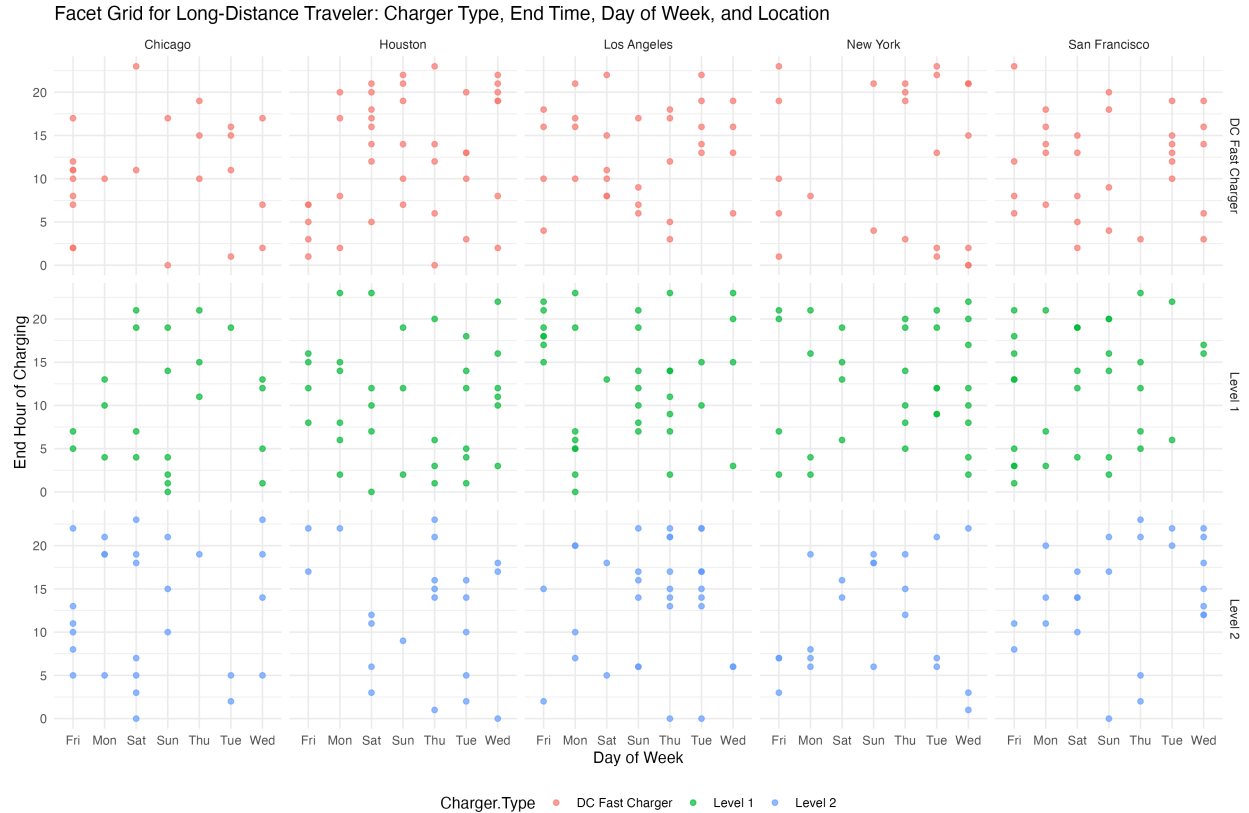
New York: - **Commuter:** Tiene la mayor frecuencia, con una diferencia significativa frente a los demás. - **Long-Distance Traveler y Casual Driver:** Están casi empatados, con niveles de actividad similares.

San Francisco: - **Commuter:** Es el grupo más frecuente, aunque la diferencia con los otros es moderada. - **Long-Distance Traveler y Casual Driver:** Presentan frecuencias equilibradas, sin grandes diferencias.

Análisis Multivariado



Filtrando por long distance traveler:



El análisis multivariado presentado es crucial para entender las relaciones entre las cinco variables analizadas: tipo de usuario (*User Type*), tipo de cargador (*Charger Type*), día de la semana (*Day of Week*), localización (*Location*), y la hora de finalización de la carga (*End Time*).

Este gráfico multivariado, organizado mediante facetas, permite filtrar por los diferentes tipos de usuarios, como el **Long Distance Traveler**, que es el foco principal en este proyecto, especialmente para el caso de uso relacionado con los hoteles. A continuación, se describen algunas observaciones destacadas basadas en las localizaciones y comportamientos detectados:

Observaciones

- **Chicago:**
 - Mayor actividad de los **Long Distance Travelers** durante el fin de semana, especialmente el **viernes** en los cargadores *Level 2* y *Fast Charger*.
 - La actividad se concentra entre las **5:00 a.m. y las 12:00 p.m.**, con un pico notable en el *Fast Charger* desde las **6:00 a.m. hasta las 11:00 a.m.**
 - El sábado por la mañana y el domingo a medianoche también presentan cierta concentración, aunque menos significativa.
- **Houston:**
 - Alta concentración en el *Fast Charger* durante el **sábado por la noche**.
 - El **miércoles a partir de las 8:00 a.m.** muestra una actividad interesante en el *Fast Charger*.
 - También se observa algo de movimiento el **viernes a medianoche**.
- **Los Ángeles:**
 - Mayor actividad el **viernes por la tarde**, a partir de las **3:00 p.m.**
 - El **jueves** también registra cargas finalizadas principalmente a partir de las **3:00 p.m.**

- **Nueva York:**
 - El fin de semana no muestra una alta concentración de actividad.
 - El **miércoles** es el día más destacado, con mayor uso de cargadores *Level 1*, concentrándose principalmente a partir de las **5:00 a.m.**.
- **San Francisco:**
 - La actividad está más distribuida durante el fin de semana.
 - El **miércoles**, los cargadores *Level 2* presentan alta concentración a última hora.
 - El **martes** en los *Fast Chargers* a partir de las **10:00 a.m.** muestra un pico interesante.

Conclusión

El gráfico multivariado permite observar patrones claros de comportamiento en relación con las localizaciones, días de la semana, y tipos de cargadores. Por ejemplo, es evidente que ciertos días y horarios presentan concentraciones significativas dependiendo del tipo de cargador y usuario. Este análisis proporciona insights clave para identificar tendencias en el uso de cargadores en diferentes contextos y horarios, facilitando una mejor comprensión del comportamiento de los usuarios. A pesar de que podrían extraerse aún más detalles, la visualización ya permite interpretar de manera eficiente cuándo y dónde terminan las cargas.

Generación de Gráficos Interactivos

Todos los gráficos generados han sido exportados como reportes interactivos en formato HTML, utilizando facetas para proporcionar una experiencia de análisis dinámica y filtrable. Estos gráficos permiten explorar las relaciones entre las variables seleccionadas de forma intuitiva y detallada, utilizando filtros para ajustar la visualización según las necesidades del análisis.

Los reportes interactivos se encuentran organizados en la carpeta `html_report`, listos para ser utilizados en análisis adicionales o presentaciones.

Ruta del directorio de reportes HTML: `~/ev_ad_strategy/html_report`

Experiencia durante el desarrollo de la tarea

A lo largo del desarrollo de esta tarea, he aprendido aspectos clave sobre el análisis de datos y la importancia de comprender el dataset desde las primeras etapas del proyecto. Uno de los puntos más relevantes ha sido reconocer que, al comenzar a analizar los datos, es cuando realmente se empieza a entender mejor cómo abordar la resolución del problema planteado.

En este caso, he tenido que adaptarme a las características específicas del dataset, incluyendo la detección de datos erróneos. Como hemos discutido en clase, cualquier decisión sobre modificar un outlier o eliminar datos debe realizarse de forma controlada y con un sentido claro, ya que puede afectar significativamente los resultados del análisis. Por esta razón, he decidido mantener los datos tal y como están en esta fase inicial, sin realizar modificaciones. Sin embargo, reconozco que en futuras iteraciones del proyecto podría ser necesario revisar y ajustar los datos si surgen nuevas incidencias o detecto patrones adicionales que requieran análisis más profundo.

Este enfoque refleja la importancia de analizar los datos con cuidado, aceptando que el proceso puede requerir varias iteraciones para alcanzar un entendimiento más completo. Este aprendizaje ha sido fundamental y probablemente se convierta en un hábito útil para futuros proyectos de análisis de datos.