**Analítica Computacional para La Toma de Decisiones**

**Proyecto 1 – Bicicletas compartidas**

**URL del Repositorio:** [**https://github.com/sagomezg1**](https://github.com/sagomezg1)

Sara Alejandra Gómez – 202022625

Andres Felipe Garcia - 202021148

**Roles:**

Sara: Ingeniería de datos, Análisis de datos, Despliegue

Andrés: Ciencia de datos, Análisis de negocio, Tablero de datos

**Cliente Objetivo:**

El diseño de producto se realizó con enfoque en un cliente del área de finanzas del operador del sistema, que debe asegurar que los ingresos recibidos sean suficientes para cubrir todos los costos (operación, mantenimiento, capital) y alcance la rentabilidad esperada.

**Preguntas de negocio y plan de acción**

Preguntas por responder a través de visualizaciones de los datos y análisis descriptivo:

1. ¿Cómo varía la cantidad de bicicletas alquiladas a lo largo del día? ¿Cuáles son las horas pico de mayor demanda? Gráfico de líneas o de barras que muestre el número de bicicletas alquiladas por hora del día.
2. ¿Cómo cambia la demanda de bicicletas según la estación del año (invierno, primavera, verano, otoño)? Gráfico de barras o boxplot que compare la cantidad de alquileres entre estaciones.
3. ¿Qué efecto tienen las condiciones climáticas (temperatura, lluvia, nieve) en la demanda de bicicletas? Gráficos de dispersión o diagramas de caja (boxplots) para comparar la demanda bajo diferentes condiciones climáticas
4. ¿Cuál es la distribución de ingresos diarios y cómo se comparan con los costos operativos? ¿Qué tan frecuente es que los ingresos superen los costos? Histograma o gráfico de densidad para la distribución de ingresos, junto con una comparación frente a los costos en un gráfico de líneas o barras.
5. ¿Hay patrones en la demanda de bicicletas durante los diferentes días de la semana? ¿Qué días tienen mayor o menor demanda? Gráfico de barras que muestre el promedio de bicicletas alquiladas por día de la semana.
6. ¿Qué variables están más correlacionadas con la demanda de bicicletas? ¿Cómo se relacionan entre sí las variables climáticas y de operación? Matriz de correlación o heatmap para visualizar las relaciones entre diferentes variables.

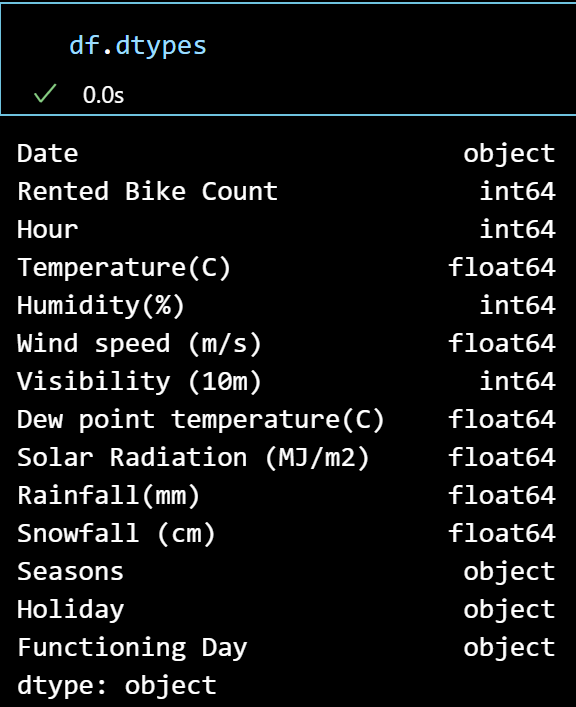
**Preguntas por responder a través del modelo predictivo de regresión lineal múltiple:**

1. ¿Cómo podemos predecir los ingresos diarios basados en variables como la cantidad de bicicletas alquiladas, la temporada, el clima (temperatura, humedad, velocidad del viento), y si es un día festivo?
2. ¿Qué combinación de factores (e.g., hora del día, condiciones climáticas, días festivos) tiene el mayor impacto en el número de bicicletas alquiladas, y por lo tanto en los ingresos?
3. ¿Cómo podemos modelar los costos operativos (e.g., mantenimiento, personal) en función de la demanda diaria (número de bicicletas alquiladas) y las condiciones climáticas?
4. ¿Qué tan bien los ingresos proyectados cubrirán los costos fijos y variables en diferentes condiciones?
5. ¿Qué variables deben modificarse para maximizar la rentabilidad, considerando los costos operativos y las variables externas?
6. ¿Qué días del año o en qué condiciones específicas (clima, temporadas, días festivos) es más probable alcanzar una rentabilidad más alta, y cómo se puede ajustar la operación para capitalizar esos días?
7. ¿Podemos predecir caídas en la demanda durante ciertos periodos y ajustar los costos de operación en consecuencia?

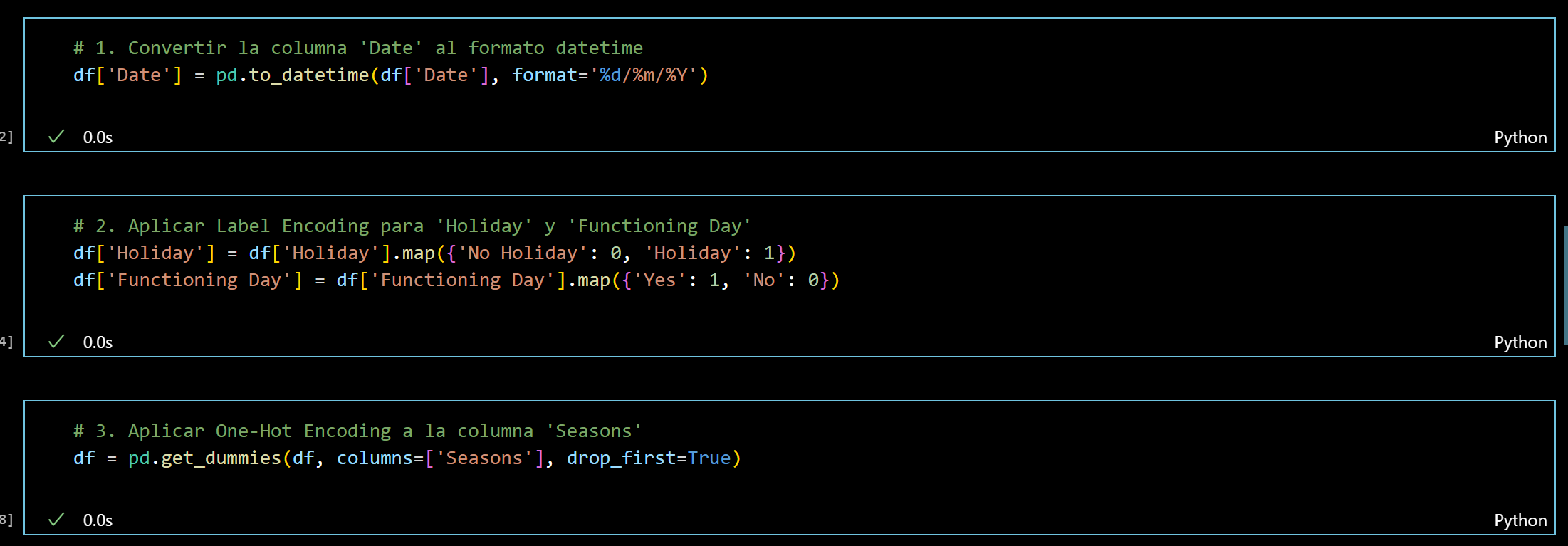
**Datos**

**Tarea 2 - Limpieza y alistamiento de datos**

Para esta tarea, en primer lugar, realizamos un análisis del tipo de variables mediante un comando que se llama dtypes, que nos muestra para cada variable qué tipo de dato tenemos:



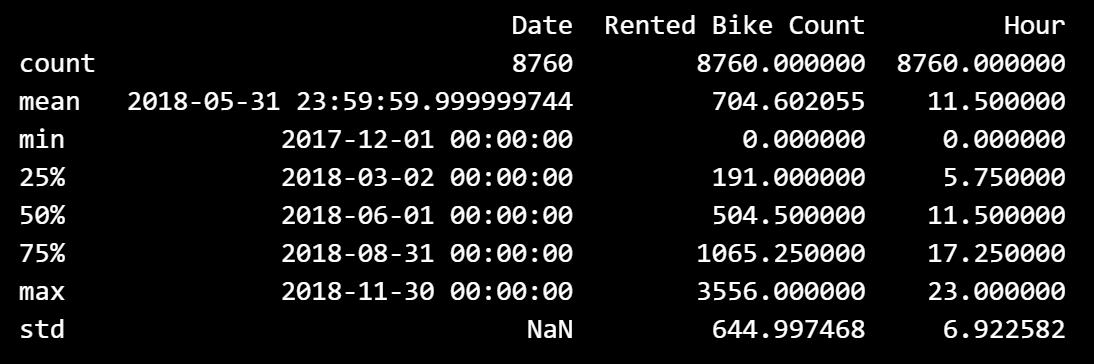
Para la limpieza de datos nos vamos a enfocar en aquellos que tengan tipo “object” porque nos interesa hacer una transformación de datos categóricos a numéricos, para mayor facilidad en el análisis estadístico y descriptivo.

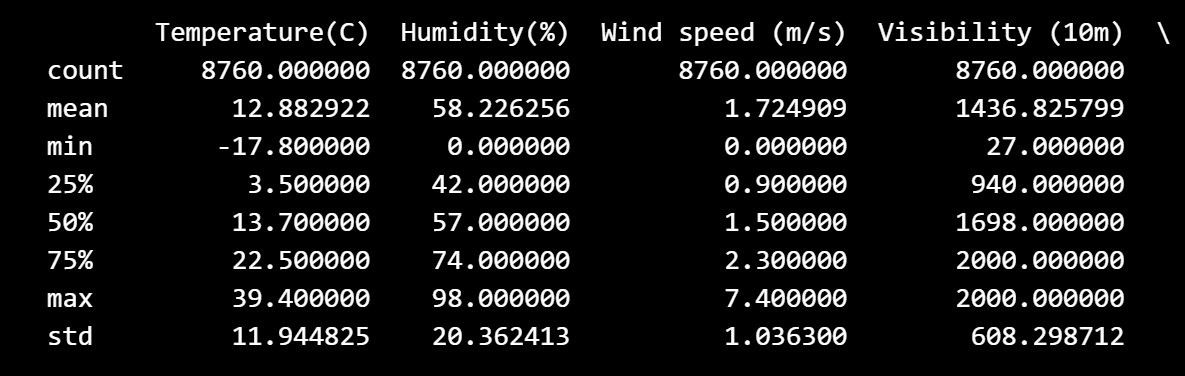


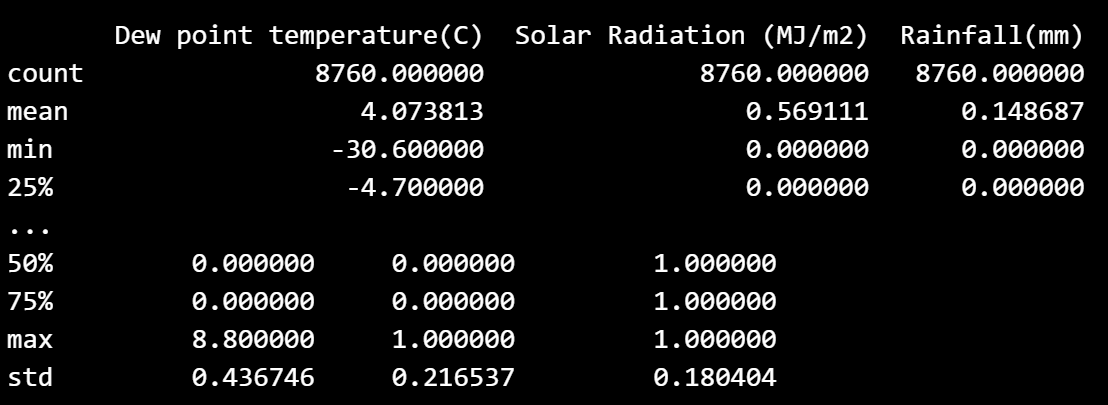
En este caso, realizamos tres líneas de código diferentes, para el primer caso, quisimos convertir los datos de las fechas en un formato más agradable, que sería de alguna manera el default que trae Python. Seguidamente, seguimos los lineamientos del taller 1 para la transformación de variables categóricas y nuevamente usamos la función map que transforma variables categóricas en números. Teniendo en cuenta que “Holiday” y “Functioning day” solo contienen dos valores posibles, lo que hicimos fue convertirlo en una variable binaria con posibles valores [0,1]. Para el caso de la variable “Seasons”, el tema no se puede manejar igual que para las otras variables porque hay mas de dos categorías, entonces en este caso, lo que hicimos fue utilizar el “One-Hot encoding”, mediante el cual se separan por columnas todos los valores posibles, en este caso son las 4 estaciones y para cada una de las 4 columnas nuevas se asignan variables binarias [0,1]. Resultando en el siguiente df:

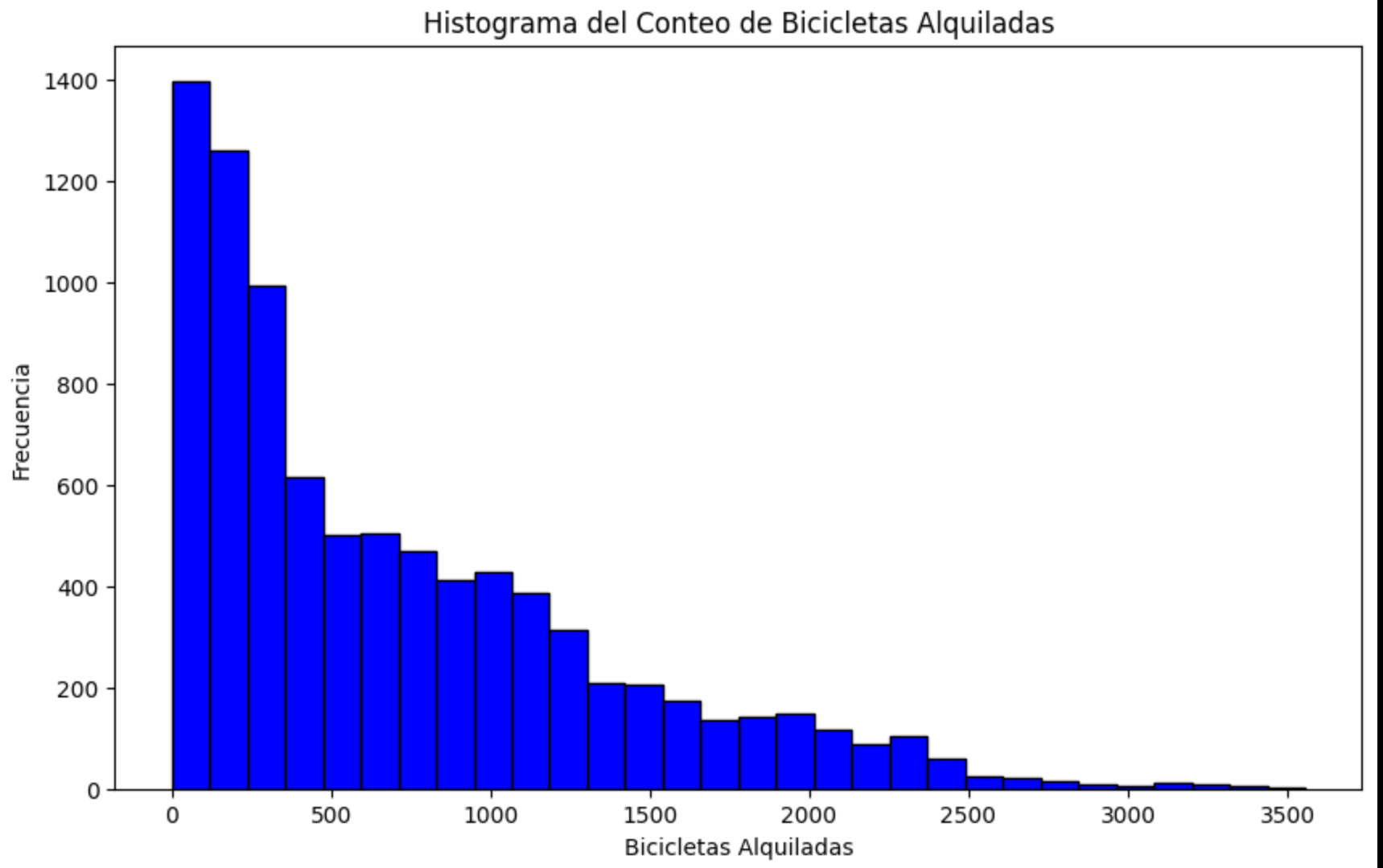


**Tarea 3 - Exploración de datos**

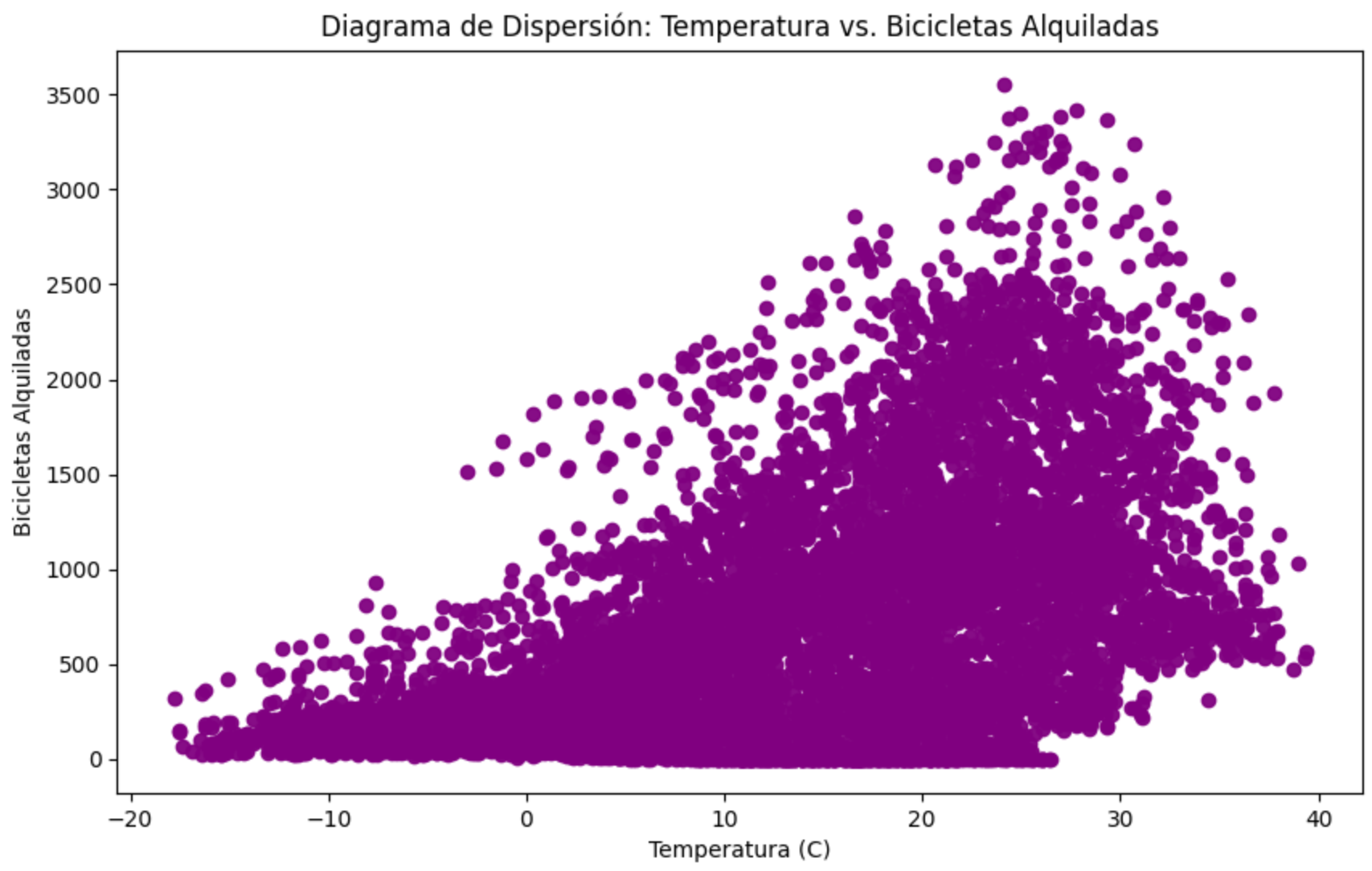




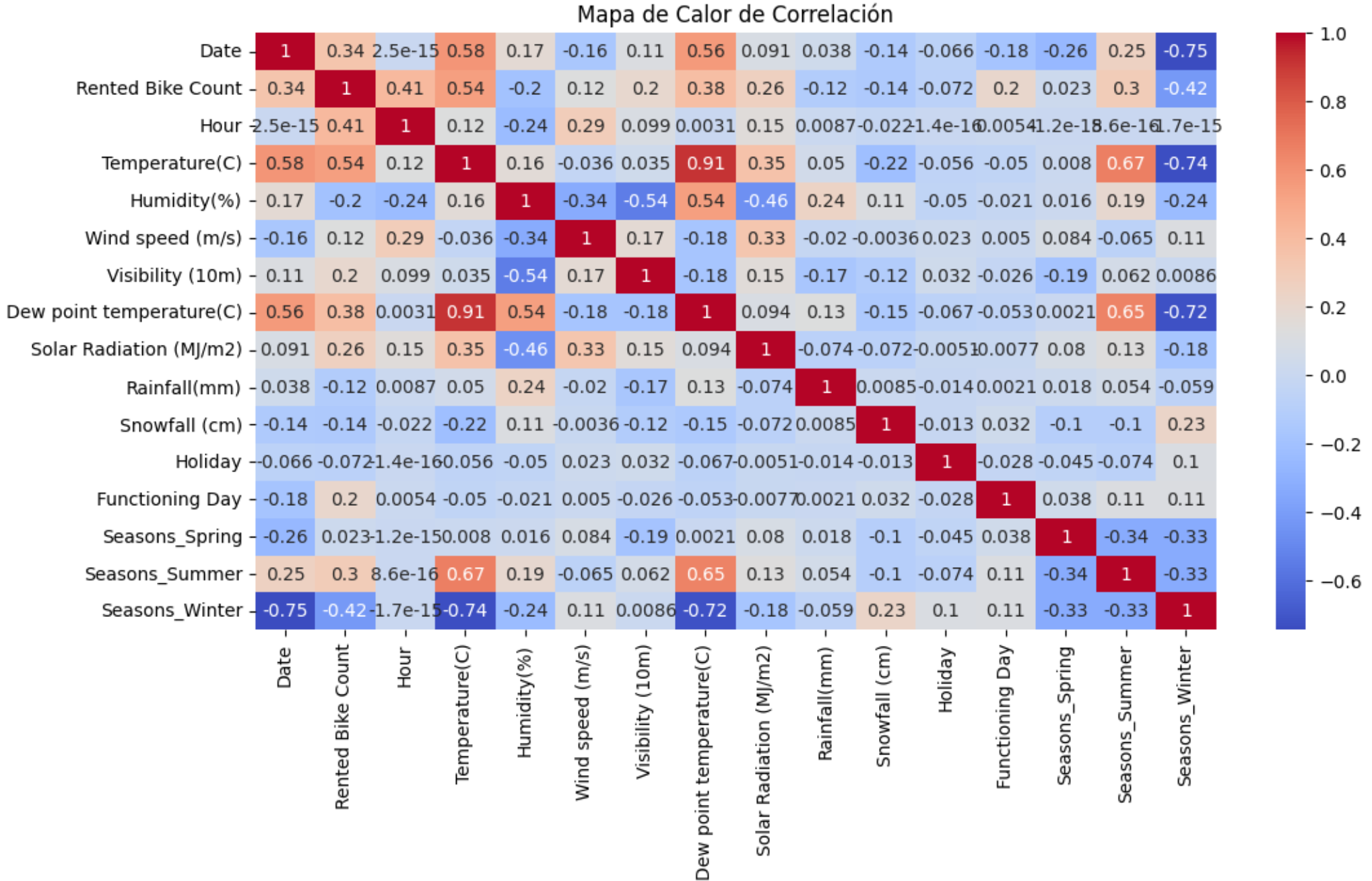




En cuanto al histograma, es posible analizar que diariamente se pueden llegar a alquilar principalmente de 0 a 500 bicicletas. Los valores restantes que serian para rentas mayores a 500, representan un valor por temporada, lo que quiere decir que no son las rentas que se tendrían en temporadas normales. Además, por las estadísticas descriptivas podemos analizar que el máximo valor de renta de bicicletas de 3556 bicicletas. Por lo que se podría decir que la empresa, sin ningún tipo de festivo o temporalidad especifica puede rentar de 0 a 500 bicicletas.

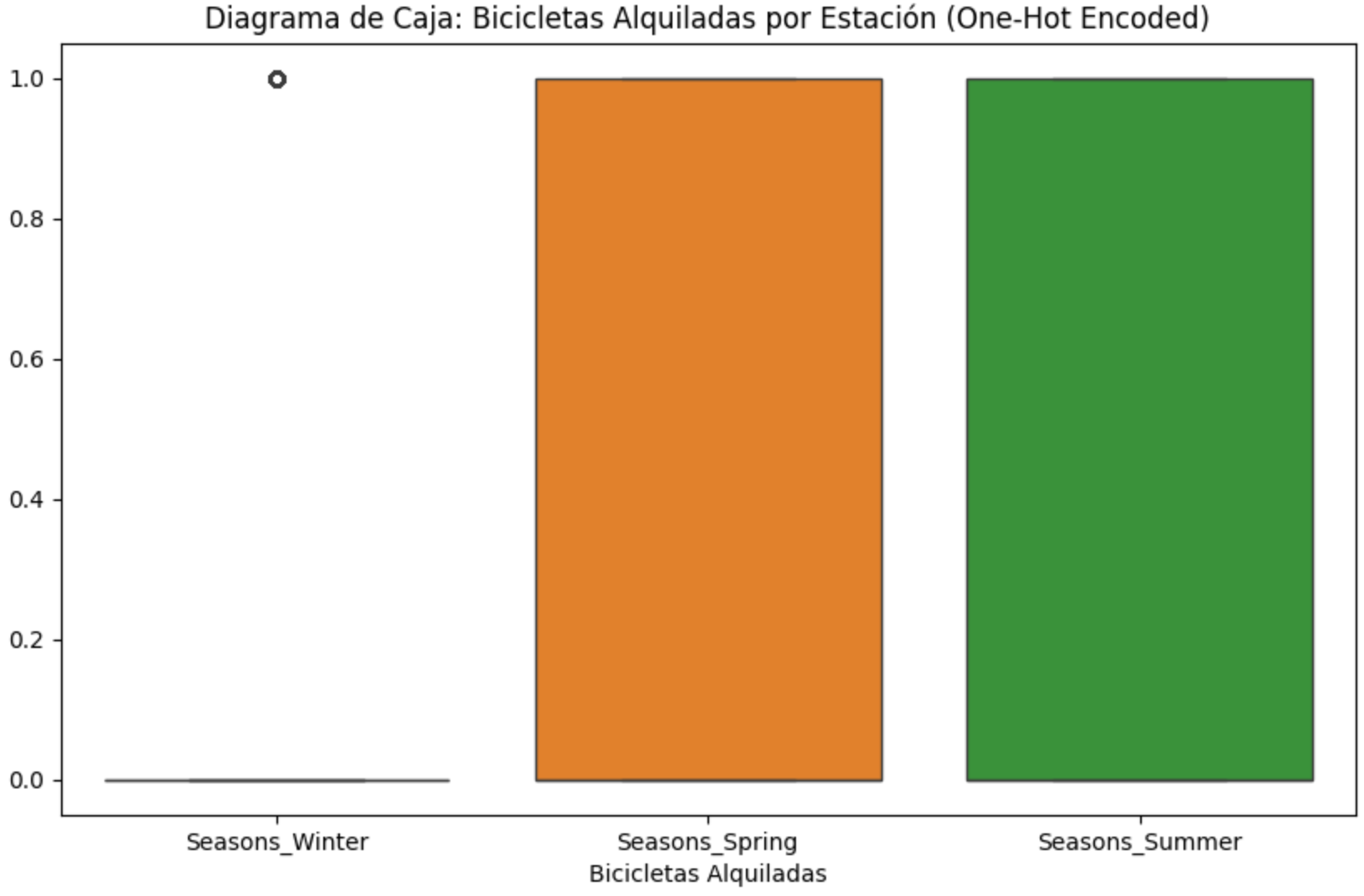


Respecto a la temperatura ideal para rentar bicicletas para los clientes seria entre 15 y 30 grados centígrados, debido a que estos son los valores en los que se ha rentado la mayor cantidad de bicicletas para la empresa. Por otro lado, podemos ver otro factor favorable y es que a pesar de que haya temporadas de temperaturas de -10 grados igual se están rentando hasta mas de 500 bicicletas, por lo que podemos ver que se esta volviendo en un servicio casi esencial para los clientes. Además, se recomienda para la empresa realizar especiales inversiones a publicidad, nuevas innovaciones y otras tecnologías para las temporadas donde haya temperaturas entre 15 y 35 grados centígrados porque representa el cuantil 75% y 100% de las rentas de bicicleta lo que representan ingresos superiores para la empresa.

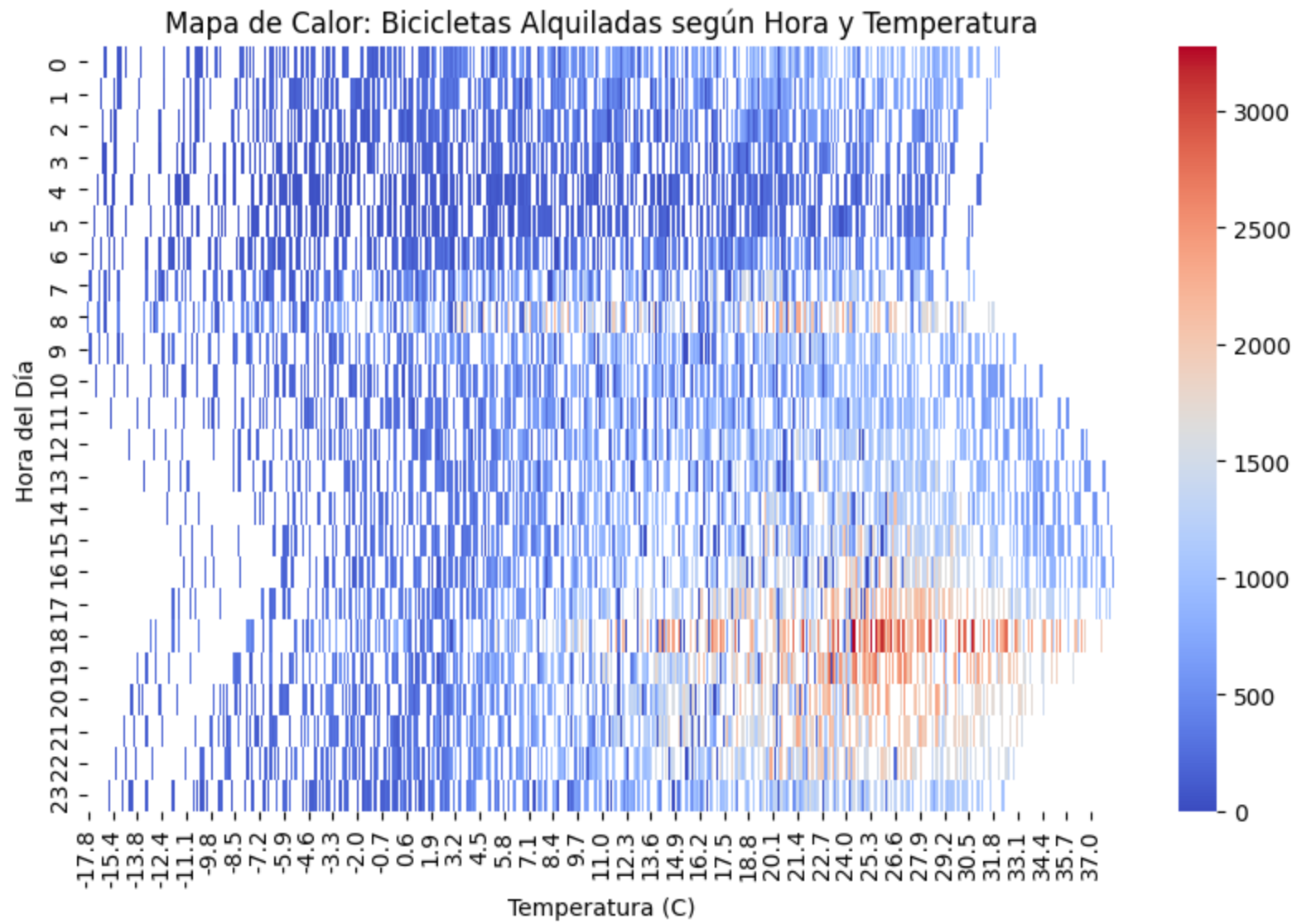


El gráfico muestra la correlación entre diferentes variables relacionadas con la renta de bicicletas y las condiciones climáticas. Se observa que el conteo de bicicletas rentadas tiene una fuerte correlación positiva con la temperatura (0.54) y la radiación solar (0.35), lo que sugiere que en días cálidos y soleados se rentan más bicicletas. También existe una correlación negativa con las estaciones de invierno (-0.42) y la humedad (-0.24), indicando que en climas fríos y húmedos la demanda de bicicletas disminuye. La hora del día también influye, con una correlación de 0.41, lo que muestra que ciertos momentos del día pueden afectar el número de bicicletas alquiladas.

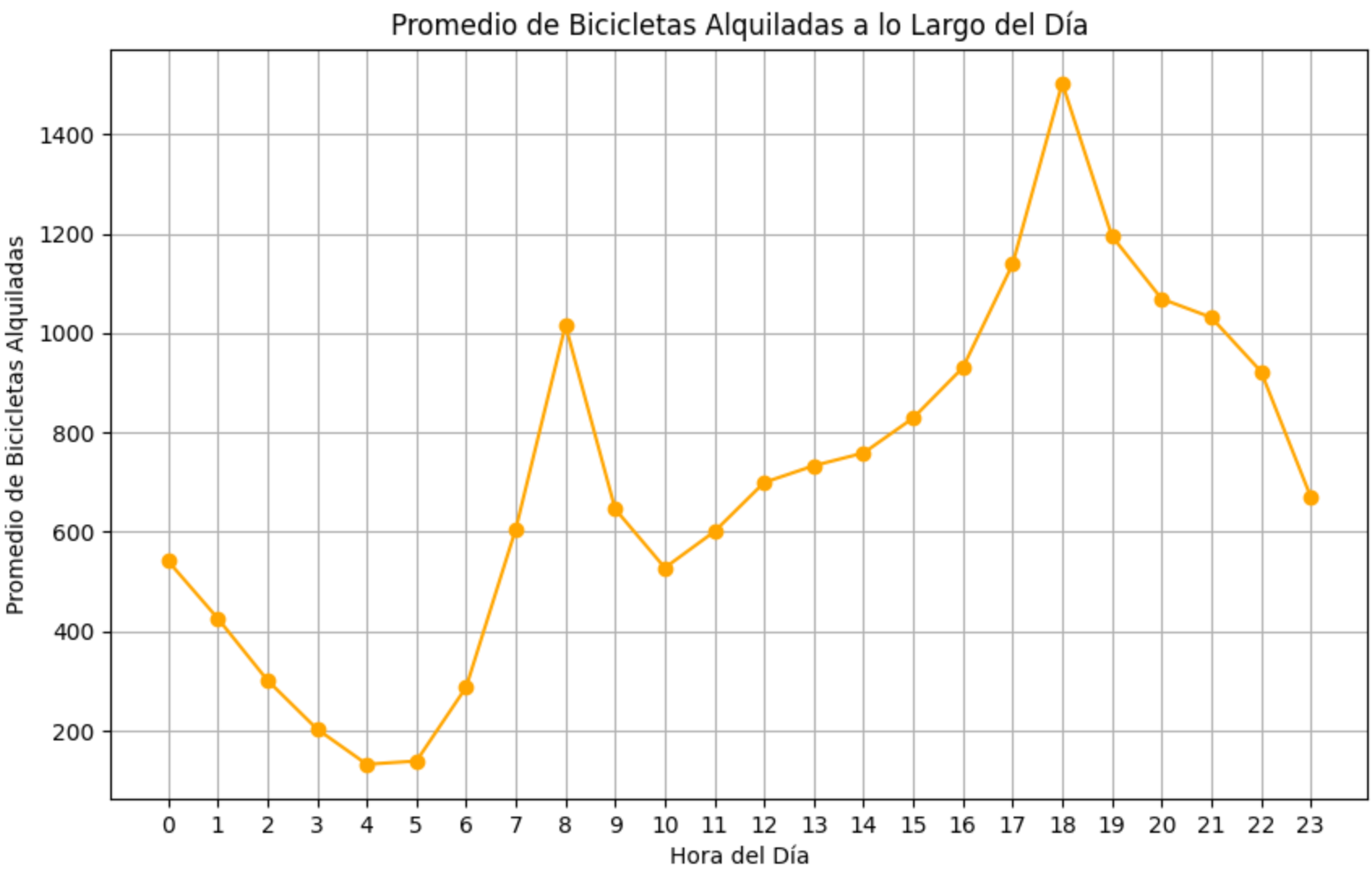
Además, las variables climáticas presentan relaciones claras entre sí, como la temperatura y el punto de rocío (0.91), que están fuertemente ligadas. Las estaciones del año tienen un impacto notable en el clima, donde invierno tiene una fuerte correlación negativa con la temperatura (-0.74) y la radiación solar (-0.72), mientras que verano muestra una correlación positiva con la temperatura (0.65). Las variables relacionadas con días festivos o laborales, como Holiday o Functioning Day, no parecen tener un impacto significativo en las demás variables del gráfico.



Para el análisis de las estaciones mas favorables para el alquiler de bicicletas, podríamos analizar que la primavera y verano, son las más rentables para la empresa porque representan el promedio mas alto de bicicletas alquiladas. Sin embargo, el invierno no representa altas utilidades para la empresa, aunque hay un outlier que podrían ser los clientes que lo utilizan como método de transporte indispensable, esto no es tan representativo en cuanto a lo que sería el público objetivo y sobre todo la estación en la que mayor inversión debe haber.

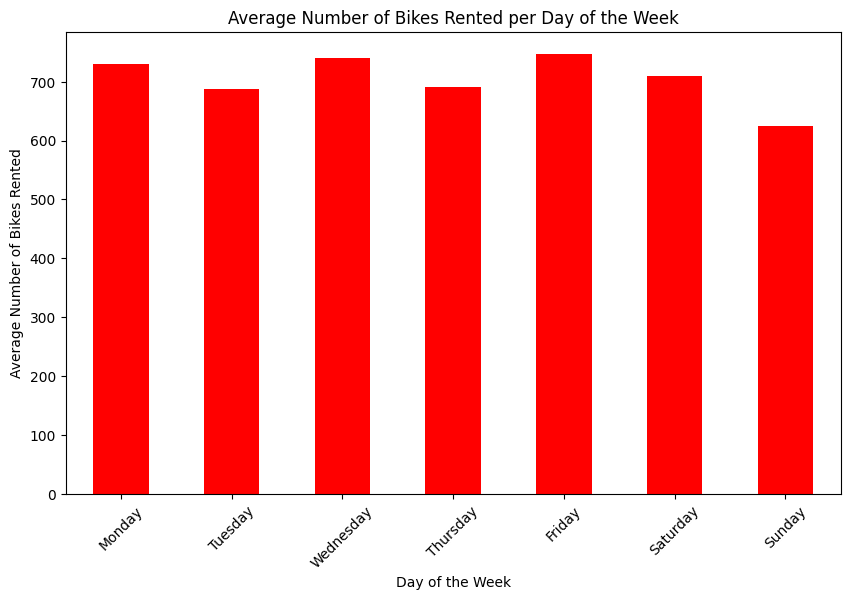


Con este mapa de calor, podríamos crear algunas relaciones, que son combinaciones optimas para entender la mejor hora y temperatura que se debe tener en cuenta. En este caso podemos analizar un primer rango que son las 8am con temperaturas entre 4.5 y 30.5 grados centígrados. En cuanto al bloque de la tarde, el rango de 5pm a 10pm tienen un color naranja a rojo lo que indica que es donde mayor numero de bicicletas se rentan. Además, con temperaturas de 12.3 y 37 grados centígrados es donde habría mayor renta de bicicletas en el bloque de tarde y nocturno. Por último, haciendo un análisis de cual seria el mejor bloque teniendo en cuenta los colores, este seria a las 6pm en temperaturas de 25.4 y 27.9 grados centígrados.



Retomando, el análisis anterior para añadir más soportes al tema de la hora mas concurrida de alquiler de bicicletas. Tomamos en cuenta los puntos de peak en la gráfica, estos se encuentran a las 8 am y a las 6pm, sobresaltando el bloque de la tarde podemos ver que hay una diferencia promedio de 400 bicicletas más arrendadas a esa hora. Además, podemos ver que la gráfica incrementa en mayor manera cunado se sobrepasa el medio día, por lo que el bloque de la tarde es el de mayor productividad para la empresa.

Finalmente evaluamos el numero promedio de bicicletas alquiladas a lo largo de la semana esto nos da una idea de cómo se comporta la variable en cualquier dia disponible.



Según el grafico podemos concluir que no existe una diferencia aparente entre los días de la semana y el numero promedio de bicicletas rentadas, en este orden de ideas se aconsejaría verificar otro tipo de variables y su relación para así poder ver su influencia.

**Modelos**

**Producto**

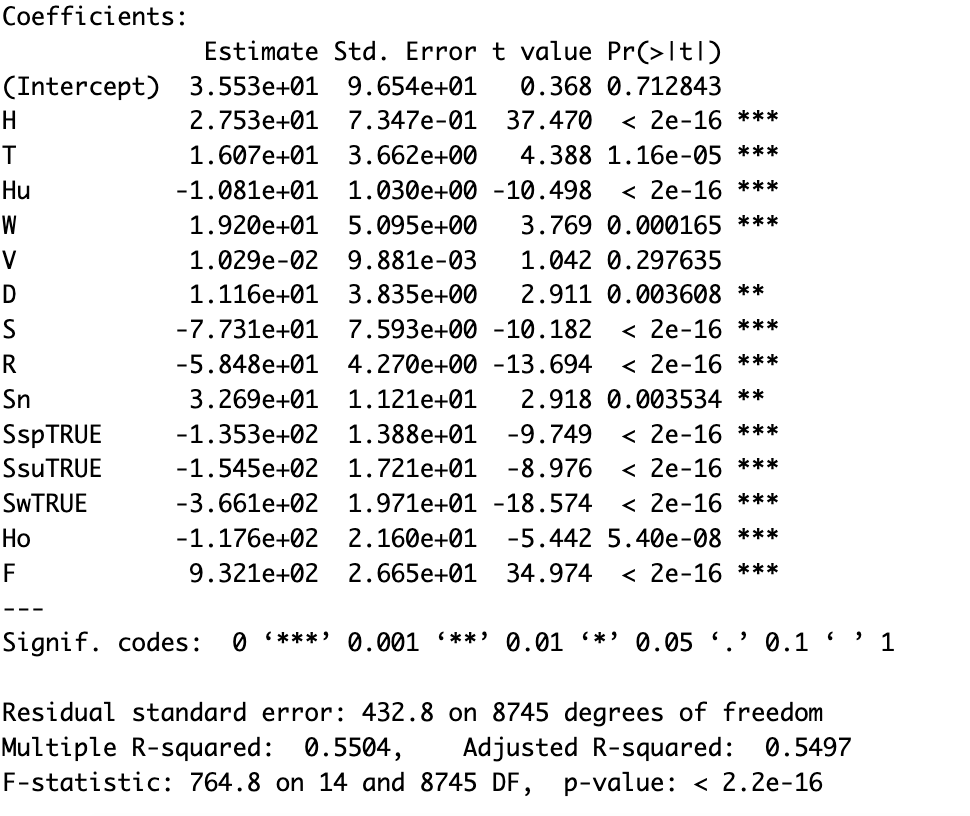
Debido a que la información con la que se tiene no incluye los precios, nuestro análisis financiero va a tener ciertos supuestos, como los son la unidad de renta o venta (que tanto se cobra por el alquiler de una bicicleta). Adicionalmente, tendríamos supuestos de costos operativos (mantenimiento y personal) y finalmente la proyección de estos ingresos. Los valores de supuesto que utilizamos son los siguientes: Precio de arrendamiento de una bicicleta estándar (**Por día:** $10.000 -$15.000COP. 4 viajes de 60 minutos es decir $10.000/4=$2.500 y $15.000/4=$3.7500, es decir $2.500-$3.750 por hora) según la secretaría de movilidad de Bogotá (2024). Finalmente, tendríamos los costos operativos (**Por bicicleta:** $50.000-$60.000) según Megasbyke (2024)., adicionalmente el arreglo de una biclicleta se hace cada 200 horas de uso.

Con el objetivo de realizar un modelo de regresión lineal múltiple que permita realizar predicciones con base a los datos obtenidos, y teniendo en cuenta lo mencionado anteriormente sobre los supuestos presentes en el modelo. Se decidió realizar 2 modelos, en los que cada uno de ellos se valoran ciertas características para llegar a una conclusión (prediccion de demanada y factores que impactan la rentabilidad, inicialmente en cada modelo se evaluará supuestos de regresión lineal como multicolinealidad, homocedasticidad, especificación y demás. La idea de evaluar los supuestos de regresión para cada modelo es poder corregirlos y decidir el mejor modelo Anova para la predicción de los ingresos y costos.

### **Modelo para predecir la demanda (cantidad de bicicletas alquiladas):**

* Variable dependiente (Y): Rented Bike Count (Número de bicicletas alquiladas). Variables independientes (X): **Hour:** Hora del día, **Temperature(C):** Temperatura, **Humidity(%):** Humedad, **Wind speed (m/s):** Velocidad del viento, **Visibility (10m):** Visibilidad, **Dew point temperature(C):** Punto de rocío, **Solar Radiation (MJ/m2):** Radiación solar, **Rainfall(mm):** Lluvia., **Snowfall (cm):** Nieve, **Seasons:** Estación del año, **Holiday:** Si es día festivo y finalmente **Functioning Day:** Si es un día hábil.

La idea de este modelo es evaluar todas las variables para ver cómo afectan la cantidad de bicicletas rentadas. Para este modelo inicialmente generamos un resumen de la regresion para ver la significancia global del modelo y ver si sirve el modelo generado. Estos fueron los resultados:

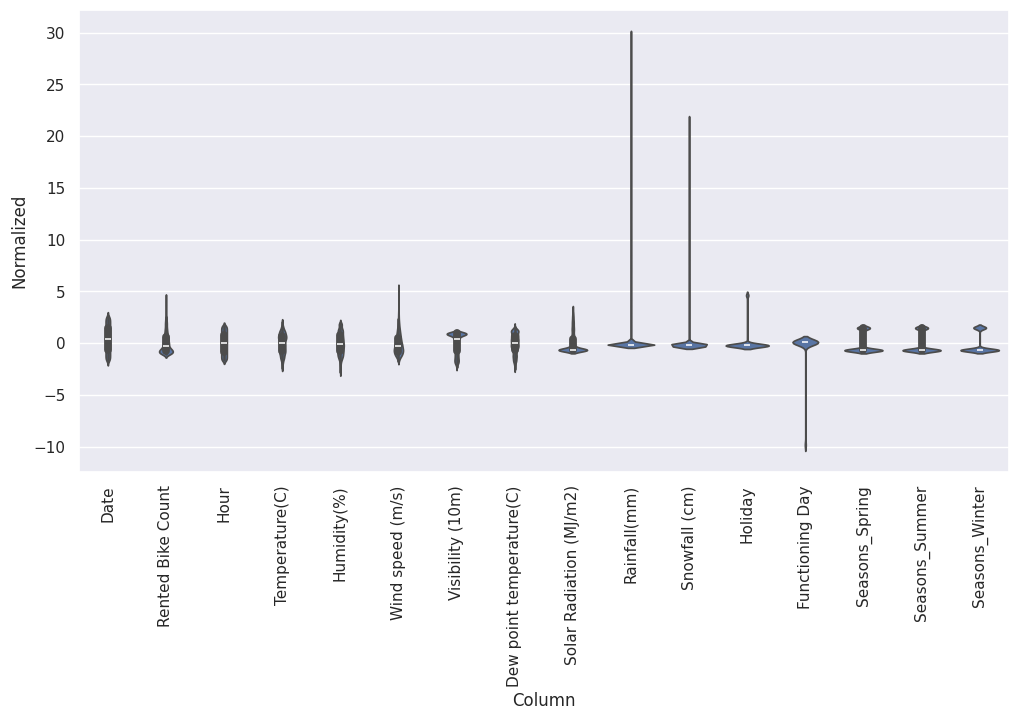


Como podemos observar en la imagen el p-value del modelo es menor al nivel de significancia por lo tanto tenemos suficiente prueba Estadística para decir que los coeficientes de las variables auxiliares son diferentes de cero. Posteriormente, se realizó un Anova de la regresión para identificar las variables significativas, las variables importantes fueron: “Hour”, “Temperature”, “Humidity”, “Wind speed”, “Solar Radiation”, “*Rainfall”, “Las tres Seasons” , “Holiday” y “Functional Day”.* Una vez analizado las variables continuamos a evaluar supuestos de Regresión Lineal como Autocorrelación, Heterocedasticidad y Multicolinealidad, la idea de evaluar estos supuestos es poder afirmar que el modelo generado se cumpla en cualquier circunstancia. Al evaluar Autocorrelación hicimos la prueba de Durbin Watson dando un resultado de 0.50948 con un p-value menor a 5%, esto indicaría que hay problemas de Autocorrelación en las variables. Para solucionarlo hicimos la transformación de Cochrane-Orcutt y volvimos nuevamente a realizar Durbin Watson, dando ahora un resultado de 1.5741 estando en el rango donde no se rechaza corrigiendo así la Autocorrelación.

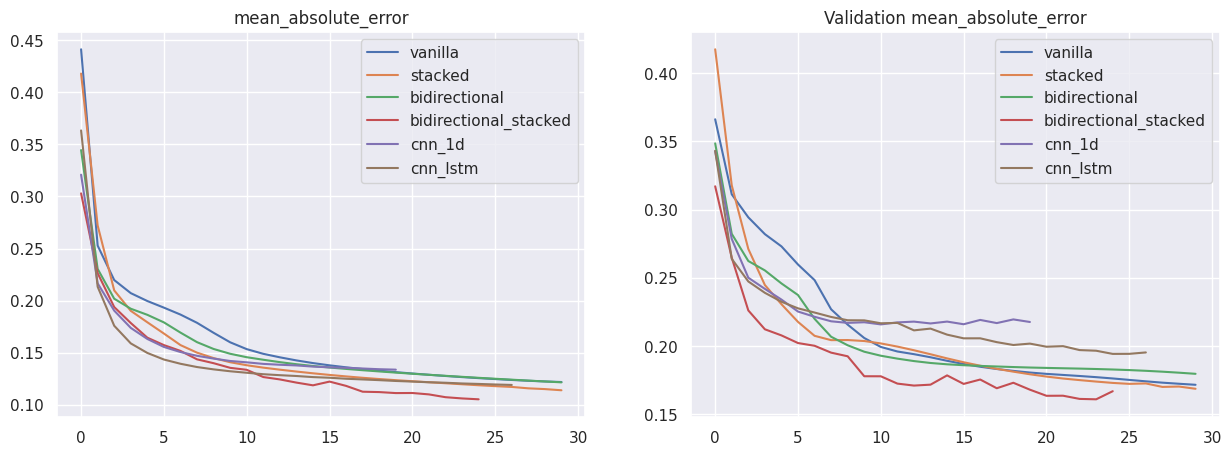
Posteriormente al Modelo corregido evaluamos Multicolinealidad, sacando el VIF de cada variable, a lo cual dieron dos variables “Temperatura” y “Dew point Temperature” mayor a 10 por ende con problemas de Multicolinealidad. Como una de estas variables en nuestro analisis Anova dio significativo (Temperatura) no es ideal quitar directamente la variable, por lo que se solucionó promediar las variables directamente y sacar nuevamente el VIF, dando ninguna variable mayor a 10.

Finalmente, se evaluó Heterocedasticidad a través de Brush Pagan del modelo corregido de Autocorrelación y Multicolinealidad. El estadístico de prueba de Busch Pagan dio 474.65 con un p-value menor a 5% dando evidencia estadística a problemas de Heterocedasticidad. Para corregir, se decidió calcular los residuos y pesos y pesos mediante la inversa del cuadrado de los residuos y así volver a generar una nueva regresión con el nuevo modelo teniendo un BP= 0.103636. Dando, así como mejor modelo el modelo denominado WLS.

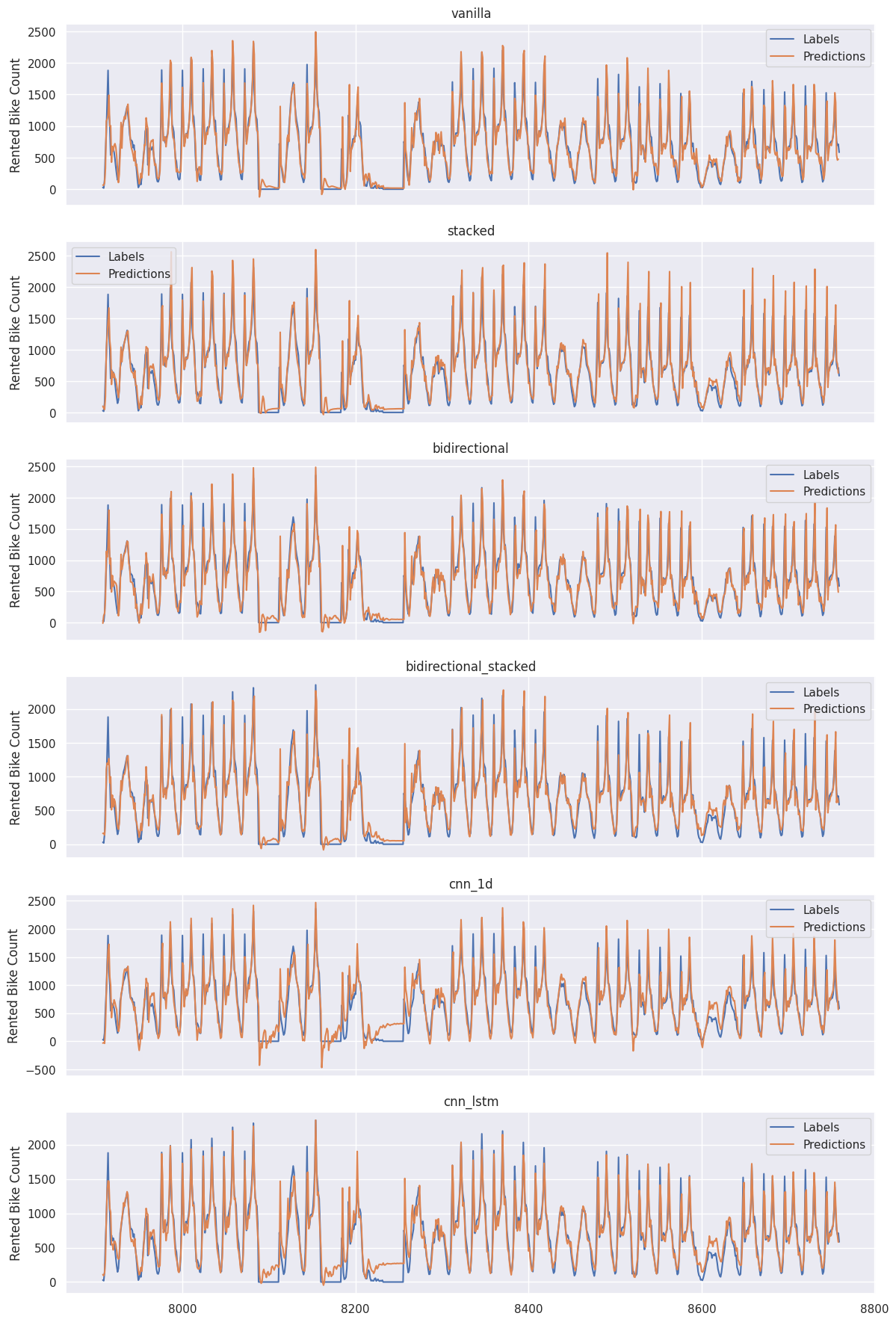
Para esta instancia se decidió hacer un modelo de predicción de redes neuronales, la idea es poder a futuro predecir los valores de la demanda y multiplicar por los supuestos de ingresos por hora de la variable “Rented Bike” y asi generar los ingresos. El primer paso, fue convertir todos los datos del archivo de datos filtrados a tipo float para evitar distinciones y posteriormente separar los datos de prueba y validación. Una vez separados los datos los normalizamos para poder evaluarlos en el modelo predictivo.



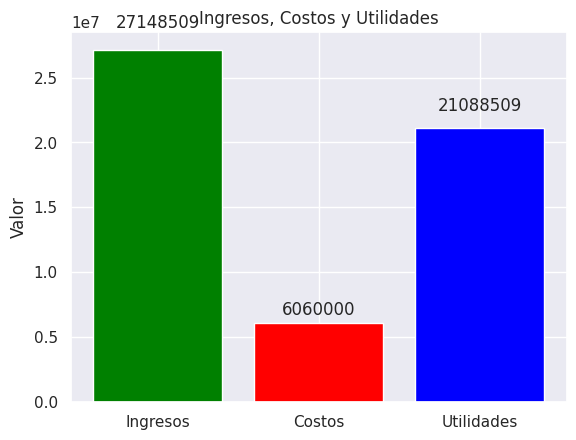
Luego creamos las series de tiempo con los datos proporcionados, para la creación de un modelo de Paso único con 6 diferentes variaciones. Creamos ahora el modelo con redes neuronales con 30 épocas y como valor de minimización el Mean Absolute Error, lo corremos y d los siguientes resultados:



Graficamos para verificar las predicciones de nuestro modelo y nos damos cuenta de que el valor de “prediction” es igualito al valor de “Labels”, es decir que nuestras predicciones son muy similares a los valores reales por ende hicimos bien nuestro modelo.



Finalmente, para realizar el análisis financiero tenemos que sacar los costos y los ingresos para identificar la utilidad. Lo que realizamos fue multiplicar los valores de Prediction de nuestro modelo con el valor $3,125 que según los supuestos que hicimos es el promedio del costo de arrendamiento de una hora por bicicleta, de esta manera sacamos los ingresos. Hacemos exactamente lo mismo para sacar los costos, pero esta vez se realizó un contador, ya que en promedio se da mantenimiento a una bicicleta cada 100 hora de uso, por ende, se coje la variable “Hour” y se va sumando, una vez llegue a 100 se suman los $60.000 del costo promedio de mantenimiento. Y finalmente las utilidades es Ingresos – costos. La simulación que hicimos de las utilidades va para el primer año de operación es decir 12 meses, dando como valor de ganancia $21088509.96501064. Graficamos:

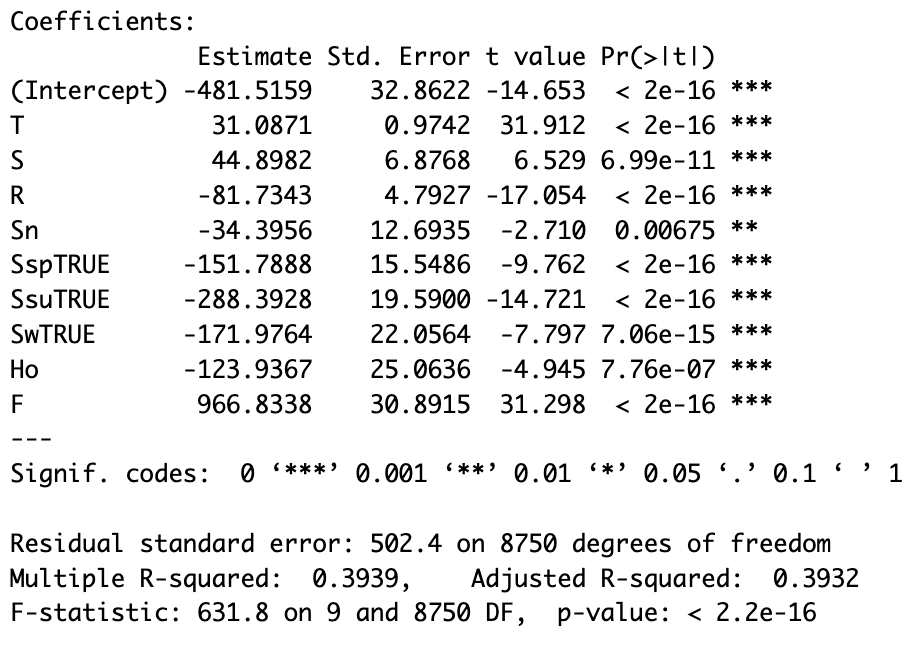


* **Segundo modelo de regresión lineal:**

**Modelo para determinar los factores que más impactan en la rentabilidad:**

* Variable dependiente (Y): Rented Bike Count (Número de bicicletas alquiladas). Variables independientes (X): **Temperature(C):** Temperatura, **Solar Radiation (MJ/m2):** Radiación solar, **Rainfall(mm):** Lluvia., **Snowfall (cm):** Nieve, **Seasons:** Estación del año, **Holiday:** Si es día festivo y finalmente **Functioning Day:** Si es un día hábil.

La idea de este modelo es evaluar ciertas variables que inicialmente pueden tener un efecto e la rentabilidad. El valor Y es nuevamente las bicicletas rentadas ya que con ese comportamiento podemos sacar la rentabilidad. De esta manera, ver cómo afectan la cantidad de bicicletas rentadas. Para este modelo inicialmente generamos un resumen de la regresión para ver la significancia global del modelo y ver si sirve el modelo generado. Estos fueron los resultados:



Como podemos observar en la imagen el p-value del modelo es menor al nivel de significancia por lo tanto tenemos suficiente prueba Estadística para decir que los coeficientes de las variables auxiliares son diferentes de cero (modelo significativo). Posteriormente, se realizó un Anova de la regresión para identificar las variables significativas, las variables importantes fueron: “Temperature”, “Solar Radiation”, “*Rainfall”, “Seasons\_Winter”, “Season\_Summer”, “Holiday” y “Functional Day”.* Una vez analizado las variables continuamos a evaluar supuestos de Regresión Lineal como Autocorrelación, Heterocedasticidad y Multicolinealidad, la idea de evaluar estos supuestos es poder afirmar que el modelo generado se cumpla en cualquier circunstancia. Al evaluar Autocorrelación hicimos la prueba de Durbin Watson dando un resultado de 0.35915 con un p-value menor a 5%, esto indicaría que hay problemas de Autocorrelación en las variables. Para solucionarlo hicimos la transformación de Cochrane-Orcutt y volvimos nuevamente a realizar Durbin Watson, dando ahora un resultado de 1.4532 estando en el rango donde no se rechaza corrigiendo así la Autocorrelación.

Posteriormente al Modelo corregido evaluamos Multicolinealidad, sacando el VIF de cada variable, ninguna variable dio mayor a 10 por lo que no se presentan problemas de Multicolinealidad.

Finalmente, se evaluó Heterocedasticidad a través de Brush Pagan del modelo corregido de Autocorrelación y Multicolinealidad. El estadístico de prueba de Busch Pagan dio 437.64 con un p-value menor a 5% dando evidencia estadística a problemas de Heterocedasticidad. Para corregir, se decidió calcular los residuos y pesos, estos pesos mediante la inversa del cuadrado de los residuos y así volver a generar una nueva regresión con el nuevo modelo teniendo un BP= 0.05697. Dando, así como mejor modelo el modelo denominado modelorWLS.

Inicialmente este modelo solo nos indica las variables que afectan la rentabilidad, mas no nos dice de qué manera y que tan eficiente seria cambiar estos datos. Para idéntificar esto se realizó una autocorrelación cruzada de las variables explicativas con lag –40 a 40. En conclusión, si se quisiera general un impacto en la rentabilidad positivo en el corto plazo se recomendaria centrarse en las variable de temperatura, Season\_Spring, Season\_Summer y Functionind Day.

