Data Trio Proyecto 1

Preguntas de negocio y plan de acción

Área de negocio: El área de operaciones del operador del sistema

Tarea 1 – Determine la pregunta (o preguntas) de negocio que quiere resolver para su cliente seleccionado - **Catalina**

Tomando como usuario final al área de operaciones, junto con la Secretaría de Movilidad de la ciudad, se plantearon las siguientes preguntas de negocio, con el fin de desarrollar un plan para satisfacer las necesidades del cliente:

- ¿Cómo varía la demanda de bicicletas a lo largo del día?
- ¿Cómo afectan las condiciones meteorológicas el uso del sistema de bicicletas compartidas?
- ¿Cuál de las condiciones meteorológicas tiene un mayor impacto en la demanda?
- ¿Cuál sería un valor esperado para la demanda de bicicletas para enero del 2019?

Por lo tanto, con el fin de visualizar los datos y obtener insights clave, se realizarán las siguientes gráficas:

- Gráfico de barras que relacione la demanda de bicicletas arrendadas con las diferentes estaciones del año.
- Gráfico de series temporales que muestre la demanda de bicicletas en función de la hora para diferentes días.
- Gráfico de dispersión que relacione la demanda de las bicicletas con variables climáticas como la temperatura, humedad, velocidad del viento y radiación solar.

Estas visualizaciones ayudarán a identificar patrones de uso y cómo las distintas condiciones afectan la demanda de bicicletas compartidas, complementando el modelo predictivo para una toma de decisiones más informada.

De igual forma, se realizará un modelo de regresión múltiple para predecir el número de bicicletas arrendadas. Inicialmente, las variables independientes utilizadas en este modelo serán: fecha, hora del día, estación, temperatura, humedad, velocidad del viento, visibilidad, radiación solar, temperatura del punto de rocío, centímetros de lluvia y nieve, si es día laboral o no, y si es festivo o no.

El propósito de este modelo es anticipar la demanda, permitiendo así optimizar la asignación de recursos.

Datos

Tarea 2 - Limpieza y alistamiento de datos - María Viviana

En primer lugar, se buscaron valores nulos o perdidos y entradas duplicadas, pero no había en la base de datos y por tanto no se tuvieron que eliminar registros.

Así como en la *Tarea 3 – Exploración de datos*, se graficaron mapas de calor las correlaciones entre variables y se observó una alta correlación entre *dew point temperature* y *temperature*, por lo cual se debía eliminar una de las dos para no afectar el modelo. Tras ver esto, se graficó en un mapa de calor las correlaciones entre las variables características y la variable de respuesta y se eliminó *dew point temperature* pues *temperature* tenía mayor correlación (0.56), lo que sugiere que es un predictor más fuerte de la variable de respuesta (Ver Tarea 3).

Por otro lado, se calcularon tasas para *Snowfall*, *Rainfall* y *Solar Radiation* para analizar estadísticamente los valores atípicos en estas (ceros). A partir de esto, se eliminaron las

variables *Snowfall y Rainfall* pues los ceros representaban el 94.9% y el 94.0% de los datos respectivamente. Luego se calculó una tasa para los No en *Holiday* y los No en *Functioning Day*, y representaban el 95.1% y el 3.4% de los datos respectivamente. En cuanto *Functioning Day,* primero se filtró el DataFrame para solo dejar "Yes" (lo que borraron 295 registros) y luego, se quitaron ambas variables pues no eran muy dicientes para la evaluación posterior del modelo.

En cuanto al tratamiento de datos, se transformaron las fechas (*Date)* con Datetime con el formato "%d/%m/%Y" y se convirtió la columna de *Seasons* con Label Encoder para tratarlos posteriormente numéricamente y quedaron así:

```
{'Autumn': 0, 'Spring': 1, 'Summer': 2, 'Winter': 3}
```

Finalmente, en la última celda del cuaderno de "Limpieza de datos" se generó una nueva base de datos "limpios" en la carpeta de data para el Modelo y el Producto.

Tarea 3 - Exploración de datos – Juan

Los datos de bicicletas compartidas cuentan con las siguientes características generales, cuenta con 13 variables explicativas: Date, Hour, Temperature(c), Humidity (%) ,Wind speed (m/s), Visibility (10m), Dew point temperature(c), Solar radiation (mj/m2), Rainfall(mm), Snowfall (cm), Seasons, Holiday, Functioning day

La variable de respuesta es: Rented bike count, Conteo de bicicletas alquiladas en cada observación

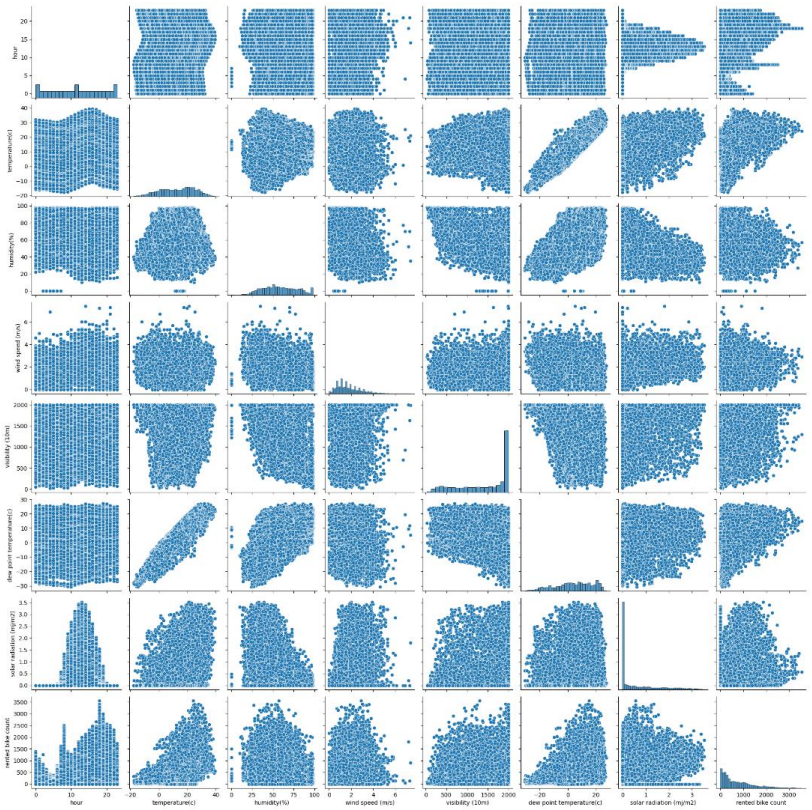
Estadísticas Descriptivas:

Al hacer una revisión preliminar de las estadísticas descriptivas de las variables disponibles observamos un comportamiento especial en las variables *rainfall* y *snowfall*, ya que sus cuartiles son iguales a 0.

	date	rented bike count	hour	temperature(c)	humidity(%)	wind speed (m/s)	visibility (10m)	dew point temperature(c)	solar radiation (mj/m2)	rainfall(mm)	snowfall (cm)
count	8760	8760.000000	8760.000000	8760.000000	8760.000000	8760.000000	8760.000000	8760.000000	8760.000000	8760.000000	8760.000000
mean	2018-05-31 23:59:59.999999744	704.602055	11.500000	12.882922	58.226256	1.724909	1436.825799	4.073813	0.569111	0.148687	0.075068
min	2017-12-01 00:00:00	0.000000	0.000000	-17.800000	0.000000	0.000000	27.000000	-30.600000	0.000000	0.000000	0.000000
25%	2018-03-02 00:00:00	191.000000	5.750000	3.500000	42.000000	0.900000	940.000000	-4.700000	0.000000	0.000000	0.000000
50%	2018-06-01 00:00:00	504.500000	11.500000	13.700000	57.000000	1.500000	1698.000000	5.100000	0.010000	0.000000	0.000000
75%	2018-08-31 00:00:00	1065.250000	17.250000	22.500000	74.000000	2.300000	2000.000000	14.800000	0.930000	0.000000	0.000000
max	2018-11-30 00:00:00	3556.000000	23.000000	39.400000	98.000000	7.400000	2000.000000	27.200000	3.520000	35.000000	8.800000
std	NaN	644.997468	6.922582	11.944825	20.362413	1.036300	608.298712	13.060369	0.868746	1.128193	0.436746

Revisando más cerca los datos encontramos que las variables *rainfall* y *snowfall* son cero en el ~94 y ~95%, respectivamente, por lo tanto, nuevamente se consideran que no son significativas para abordar el problema. A partir de ahora se retirarán de los análisis y modelos.

Histogramas y Correlación entre variables:

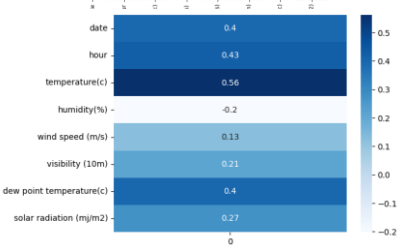


En general los histogramas muestran comportamiento normal o uniforme para la mayoría de las variables presentes, lo cual nos permite asumir que, si tienen un comportamiento aleatorio, sin embargo, *solar radiation* y *visibility* están un poco sesgados, sin embargo, es valor de moda es razonable con la naturaleza de las métricas.

Correlación entre variables explicativas:

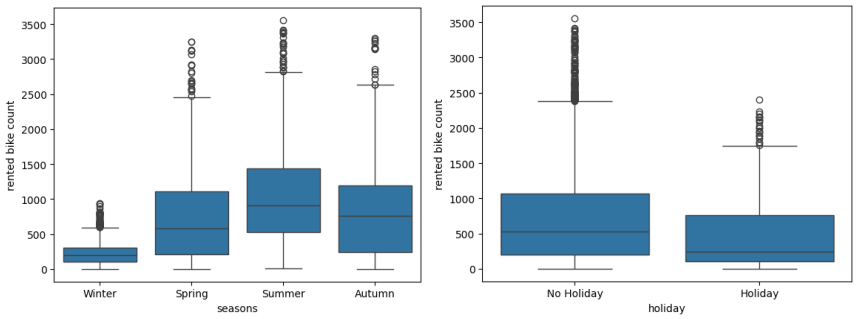
Revisando la matriz de correlaciones se observa un comportamiento normal entre las variables actuales, sin embargo, se encuentra una alta correlación, entre *dew point temperature* y *temperature*, por lo cual se recomienda eliminar una de las dos para no afectar el modelo. Por claridad se remueve *dew point temperature* ya que la otra variable tiene más sentido de negocio.

Correlación con variable de respuesta:



Revisando la correlación entre las variables explicativas y las de respuesta se observa algún nivel de correlación entre ellas, lo cual puede sugerir un buen nivel de explicación en futuros modelos. Igualmente, estos valores de correlación confirman la selección de *temperature* sobre *dew point temperature*, ya que se observa que *temperature* puede tener un mayor nivel explicativo que *dew point temperature* al tener mayor correlación con la variable de respuesta.

Análisis de variables categóricas:

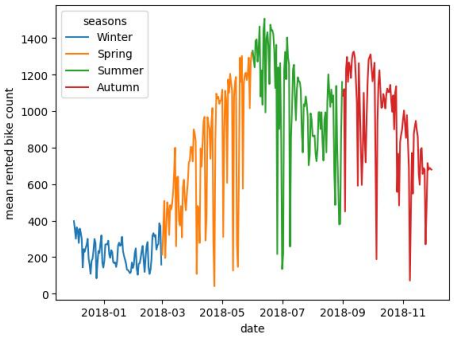


Con ayuda de los diagramas de caja se puede deducir una influencia entre la temporada climática el número de bicicletas rentadas, especialmente en la temporada de invierno. Sin embargo, no se sospecha tanta influencia entre si es un festivo o no sobre el número de bicicletas rentadas ya que los dos primeros cuartiles se ven similares. Igualmente, se observa una posible presencia de datos atípicos algunas categorías, especialmente los puntos que se encuentran en la parte superior de *Autumn*. Sin embargo, no se considera que el tamaño de estos datos justifique un tratamiento especial o que vayan a cambiar drásticamente los resultados.

Revisando más de cerca la variable *holiday*, observamos lo siguiente:

holiday	Number of dates
Holiday	17
No Holiday	336

Con ayuda del diagrama de cajas, y observando el número de días que son festivos dentro de los datos sumando a que no se observa una gran influencia sobre la variable de respuesta, tomamos la decisión de remover la variable *holiday* de las variables explicativas que puedan ser usadas en el modelo.



Al revisar la correlación entre las estaciones y la fecha de arrendamiento observamos, que en cierta medida el efecto de estas variables sobre la variable de respuesta es similar, por lo que se decide utilizar únicamente la variable *seasons* para la regresión lineal, la cual nos permitirá revisar más interacciones con las otras variables que necesariamente no sean tan evidentes.

Finalmente, se concluye que de las 13 variables disponibles para modelamiento del problema se excluirán 6 variables explicativas del problema, ya que o no apoyan a explicar la variable de respuesta o su información ya se encuentra contenida en otras variables que si se incluirán en el problema. El data set final después de este punto cuenta con 9 variables, tanto explicativas como de respuesta y con 8465 registros para modelamiento y el desarrollo del tablero.

Modelos

Tarea 4 – Modelamiento - María Viviana

Se generaron dos tipos de modelos complementarios: regresión lineal y de pronóstico y posteriormente se calcularon métricas como R^2, RMSE, MSE y MAE para la comparación y evaluación de los mismos.

Tras la exploración y limpieza de los datos, se generó un primer modelo de regresión lineal con 7 variables características, sin tener en cuenta las fechas. Para ello, se dividieron los datos en conjuntos de entrenamiento y prueba (80% para entrenar y 20% para probar) con función train_test_split de scikit-learn. Se llevó a cabo una regresión lineal utilizando el método de Mínimos Cuadrados Ordinarios (OLS), y el modelo se hizo con los datos de entrenamiento.

$$y = \beta_0 + \beta_1 \times \text{Hour} + \beta_2 \times \text{Temperature} + \beta_3 \times \text{Humidity} + \beta_4 \times \text{Wind speed} + \beta_5 \times \text{Visibility} + \beta_6 \times \text{Solar Radiation} + \beta_7 \times \text{Seasons}$$

A partir del resumen del anterior modelo, se concluyó que la variable de *Wind Speed* no era significativa para el modelo. Por tanto, se generó un segundo modelo de regresión lineal de la misma manera, sin Wind Speed y este arrojó que todas las variables eran significativas.

$$y = \beta_0 + \beta_1 \times \text{Hour} + \beta_2 \times \text{Temperature} + \beta_3 \times \text{Humidity} + \beta_4 \times \text{Visibility} + \beta_5 \times \text{Solar Radiation} + \beta_6 \times \text{Seasons}$$

De manera complementaria, se generó un modelo de Forecasting: Arima. Este es un modelo estadístico reconocido y utilizado para la predicción de series temporales (forecasting). “Este modelo consta de tres componentes. El elemento autorregresivo (AR) relaciona el valor actual con valores pasados (lags). El elemento de media móvil (MA) asume que el error de predicción es una combinación lineal de los errores de predicción pasados. Por último, el componente integrado (I) indica que los valores de la serie original han sido reemplazados por la diferencia entre valores consecutivos (y este proceso de diferencia puede haberse realizado más de una vez)” (Amat, J y Ortiz, J., septiembre 2023).

Para empezar, se convirtieron las fechas en formato de serie temporal, se ordenaron los datos en función de la fecha y la hora y se estableció *Date* como índice. Se graficó la variable de respuesta *Rented Bike Count* para confirmar la estacionalidad de la serie. Para verificar la que la serie es estacionaria (es decir, si las propiedades estadísticas, como la media y la varianza, son constantes a lo largo del tiempo), se utilizó la Prueba de Dickey-Fuller Aumentada (ADF). Dado que el p-value es menor a 0.05, la serie es estacionaria y el parámetro *d* para el modelo es 0. Para determinar el mejor modelo, con auto_arima se buscan los parámetros (p, d, q) que minimicen el error. Luego si se ajusta el modelo ARIMA y los valores ajustados se almacenan en una nueva columna “predicted”.

Desempeño de los modelos

Modelo	R2	MAE	MSE	RMSE
Regresión Lineal 1	0.5308	332.7635	198807.7325	445.8786
Regresión Lineal 2	0.5303	332.7544	199038.7356	446.1376
ARIMA	0.8320	165.8052	69303.0570	263.2547

Los resultados demuestran que el modelo ARIMA es mejor que los modelos de Regresión Lineal en términos de capacidad predictiva y ajuste a los datos históricos. Esto es coherente dado que ARIMA es un modelo especializado en series temporales y captura mejor las tendencias y patrones estacionales presentes en los datos. Es importante destacar que estos modelos son complementarios en el análisis de datos ya que los modelos de regresión ayudan a identificar y cuantificar el efecto de variables en la demanda, útil para intervenciones estratégicas, mientras que los modelos de pronóstico ofrecen estimaciones más exactas de la demanda futura, esenciales para la gestión de inventario y la optimización de recursos.

Producto

Tarea 5 - Diseño y desarrollo del tablero ─ Catalina

Gráfico de barras que relacione la demanda de bicicletas arrendadas con las diferentes estaciones del año

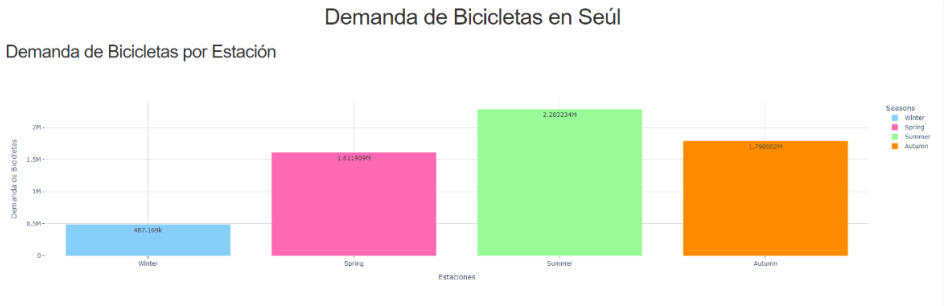


Gráfico de línea que muestre la demanda de bicicletas en función de la hora para la fecha seleccionada por el usuario

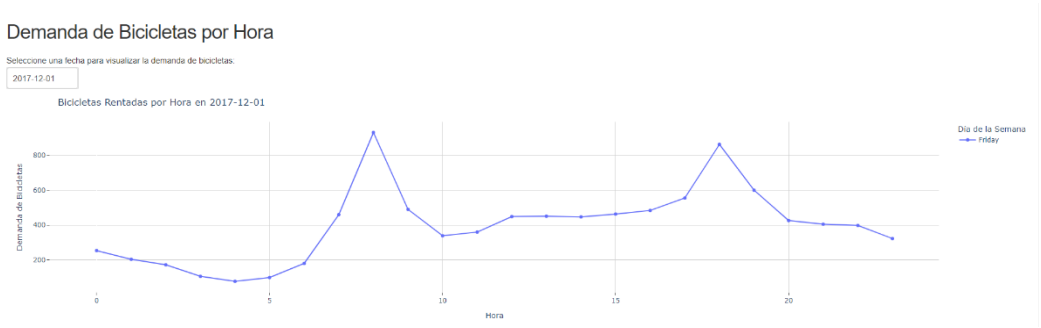
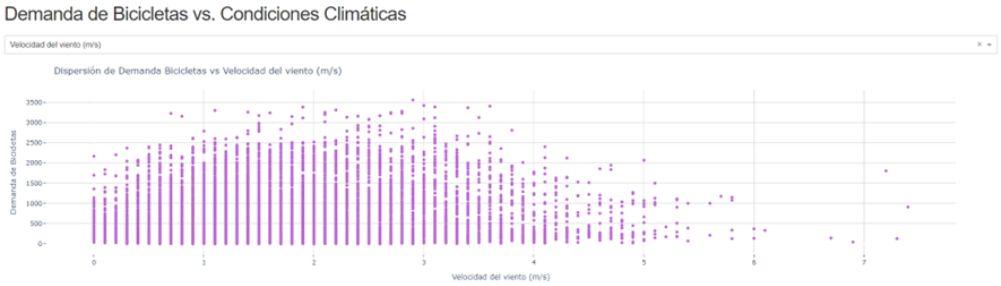
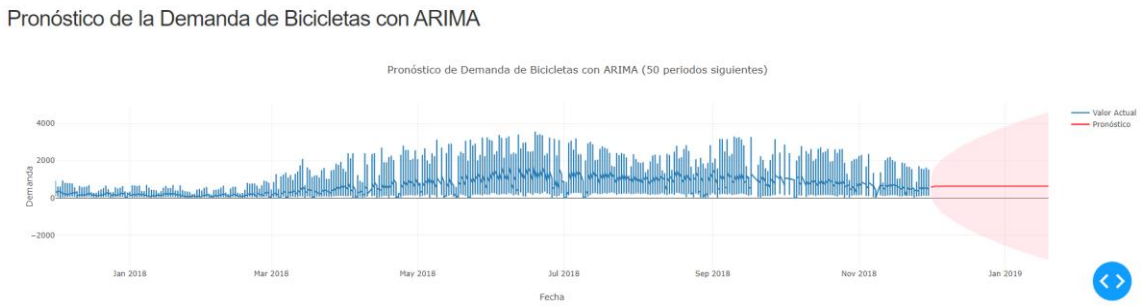


Gráfico de dispersión que relacione la demanda de las bicicletas con variables climáticas como la temperatura, humedad, velocidad del viento y radiación solar

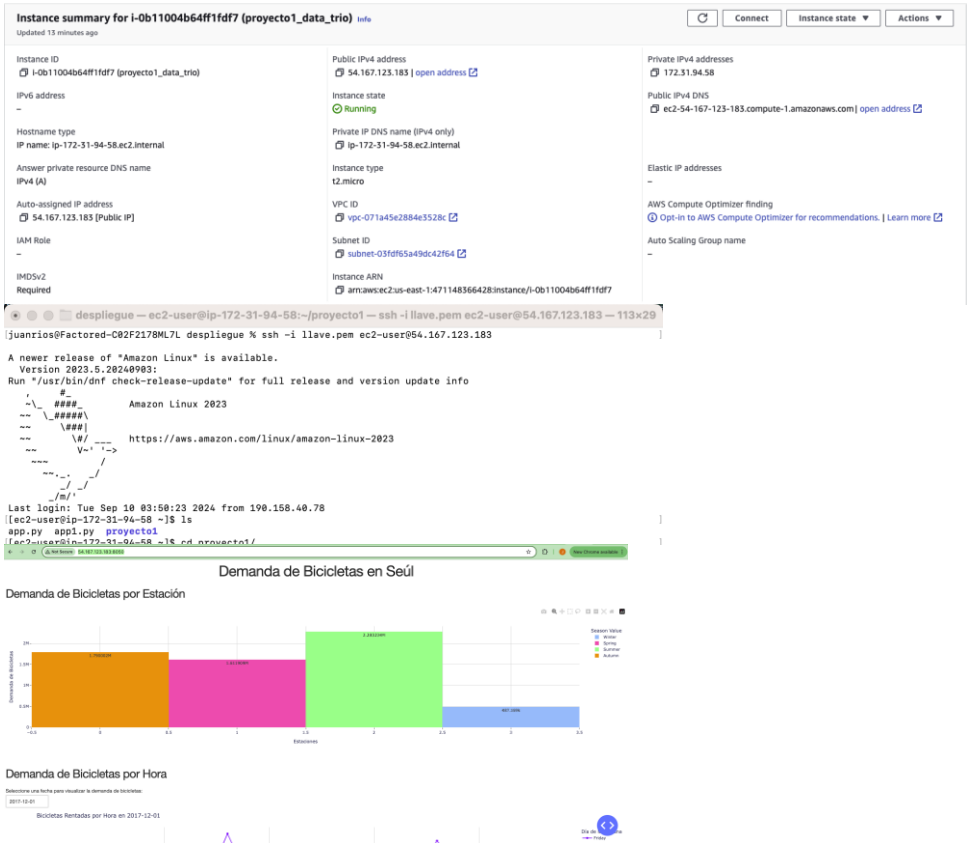


La aplicación contiene un menú de cascada donde el usuario puede escoger las variables a relacionar con la demanda de bicicletas.

Pronóstico de la demanda de bicicletas para los siguiente 50 periodos a partir de la última fecha



Tarea 6 – Despliegue – Juan



URL Tablero en ejecución: http://54.167.123.183:8050/

Referencias:

Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and Practice. OTexts. Recuperado de: https://otexts.com/fpp2/

Rodrigo, J & Ortiz, J. (septiembre 2023). Modelos ARIMA y SARIMAX con Python. Recuperado de: https://cienciadedatos.net/documentos/py51-modelos-arima-sarimax-python