Programa de Maestría en Ciencia de los datos y analítica

Modelado e Implementación para la predicción de ventas de motos en Colombia

Proyecto integrador

Semestre I

Ciencia de los datos y analítica

PRESENTA:

Cristian Castro Arias

Javier Patiño Serna

Ssndra Marcela Díaz Cordero

Simón Madrid Álvarez

Tutores Principales:

Henrry Laniado Rodas

Edwin Nelson Montoya Murillo

José Antonio Solano Atehortua

Colombia,Medellin. (Diciembre) 2022

Contents

1 Marco De Referencia

El marco de referencia del problema tratado en este proyecto es el mercado de motos nuevas en Colombia:

1.1 Problema general

El mercado de motos nuevas en Colombia vende alrededor de 750 mil unidades al año; en este participan diferentes marcas (Auteco, Yamaha, Akt, Hero, Honda, Suzuki, Uma) que compiten por ganar share (porcentaje de las ventas).

Estos competidores tienen varias referencias de motos que se ubican en diferentes segmentos.

Estos segmentos fueron acordados por el mercado y buscan segmentar el uso que cada consumidor le da a la motocicleta:

Figure 1: Segmentos principales del mercado de motos nuevas en Colombia

El problema que se quiere resolver con este proyecto integrador es: Predecir las unidades vendidas de las referencias de moto marca AUTECO (variable dependiente) como resultado del comportamiento de sus competidores (variables explicativas).

1.2 Impacto de solución

Con esta soluciónn, la empresa podrá hacer un presupuesto dinámico (de ser posible diario) de las ventas; Incluso se podrá hacer una intervención temprana de precios si así se requiere (promociones para atraer más demanda).

Adicionalmente, a través de la matriz de coeficientes, la empresa tendrá la posibilidad de identificar cuáles son los competidores más directos

2 Desarrollo Tecnológico

2.1 Arquitectura de almacenamiento en AWS (Batch)

Antes de plantear la arquitectura completa del proyecto, se identificaron las zonas del mismo:

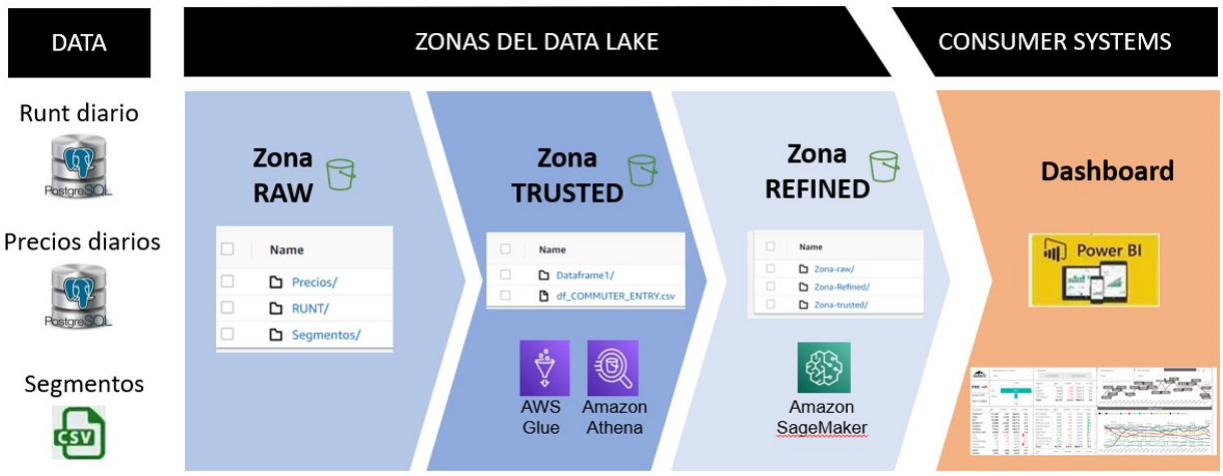


Figure 2: Zonas del proyecto

Cabe aclarar que el datalake está compuesto de datos estructurados y por tanto pueden ser consultados a través del lenguaje SQL.

Se definieron tres zonas, las cuales se alojaron en el bucket S3 del proyecto.

En la Zona Raw se encuentran los datos crudos, sin procesar y sin ningún tratamiento (tal y cual se leen de la fuente original)

En la Zona trusted se encuentra la información catalogada y procesada con Glue y Athena. En este bucket se encuentra el dataframe listo para poder evaluar el modelo aplicable.

Por último, en la Zona refined se encuentra el resultado del modelo, es decir los betas con los cuales se predecirán las ventas.

La arquitectura dieñada está en batch, ya que no se necesita tener información en línea (incluso para la predicción se utiliza una variable de rezago). Aicional, el proveedor de la API expone el servicio de consulta de las unidades vendidas con la información del día anterior.

Esta arquitectura fue elegida para una mejor optimización de los recursos en el procesamiento de la información.

Esta arquitectura lleva la entrada, procesamiento y salida de los datos a una base de datos que luego será gestionada en un visor que se ha elegido para la visualización por el usuario.

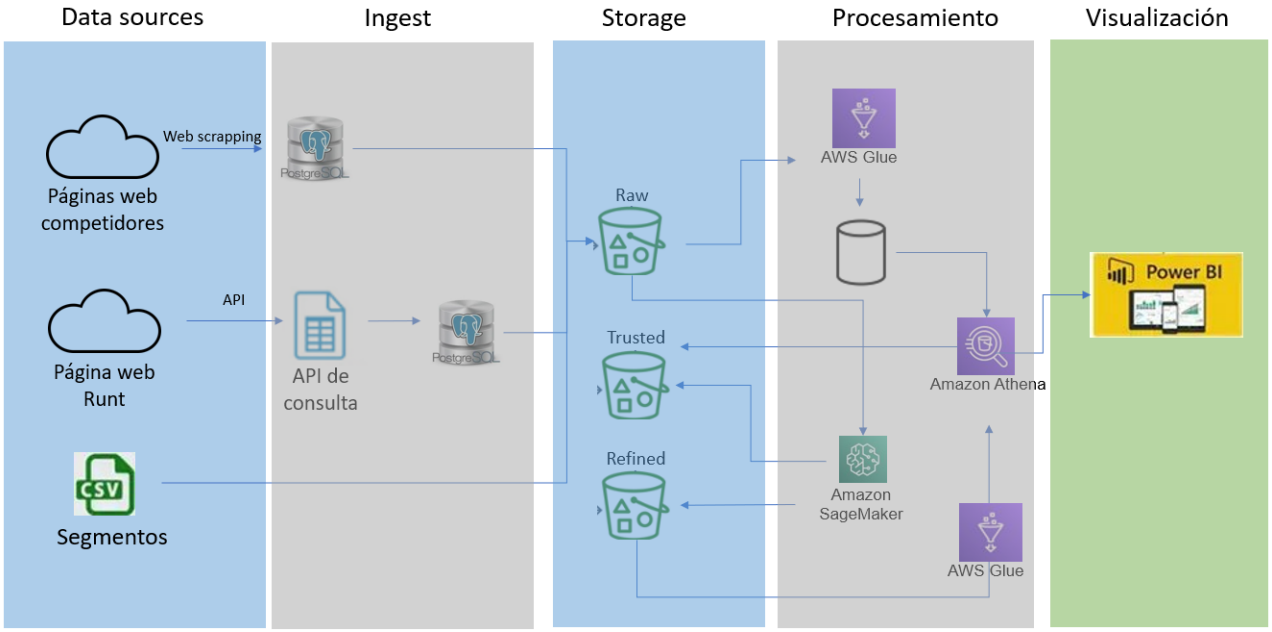


Figure 3: Arquitectura batch del proyecto

2.1.1 Data Source

La fuente de datos correspondiente a los **precios** es recolectada con ayuda de un web scraping que consulta la información en las páginas web de los competidores. Esta información se consulta con periodicidad diaria y se hace la ingesta automática a una base de datos que aloja el consolidado de los precios consultados ese día.

Las páginas consultadas son:

- grupouma: GrupoUma®, 2022
- incolmotos-yamaha: Yamaha®, 2022
- aktmotos: motos®, 2022
- suzuki: suzuki®, 2022
- auteco: auteco®, 2022
- heromotos:Heromotos®, 2022

En el bucket S3 del proyecto, se encuentra como un archivo csv llamado precios. Esta información es pública (no confidencial) y proviene de las páginas web de cada una de las ensambladoras que pertenecen al mercado de motos en Colombia. Esta base contiene los precios diarios de las referencias de motos que se venden en Colombia y está discriminada por cada ensambladora.

La fuente de datos correspondiente a **RUNT motocicletas** se recolecta a través de una API con acceso pago, expuesta por el proveedor del Runt en Colombia. Esta información es pública (no confidencial), pero con acceso limitado a través de un pago mensual.

La base RUNT de motocicletas contiene las ventas diarias de motos nuevas en Colombia discriminadas por referencias, marca y empresa ensambladora responsable. La ingesta del consumo de la API se hace de manera automática a una base de datos.

En el bucket S3 del proyecto, se encuentra como un archivo csv llamado Runt.

La fuente de datos correspondiente al **Listado de referencias de motos por segmento** es un archivo compartido de construcción interna, donde los jefes comerciales actualizan la información de segmentos ante cambios en el mercado.

Esta información contiene los segmentos a los que pertenecen cada una de las referencias de motos del mercado (incluyendo las motos propias y de los competidores). Estos segmentos son motos de trabajo, deportivas, deluxe, scooter y moped.

Como la información es manual y de construcción interna, se propone una ingesta manual desde el archivo csv al bucket S3.

2.1.2 Storage (S3)

Para este proyecto se crearon tres S3 (Servicio simple de almacenamiento de AWS): zona-raw, zona-refined, zona-truested.

El buccket del proyecto se llama "prr-integrador-grupo3". Se configuró para ser público dados los objetivos académicos, sin embargo, en el contexto de empresa este bucket se configurará privado con accesos restringidos para los ingenieros de datos de la compañía.

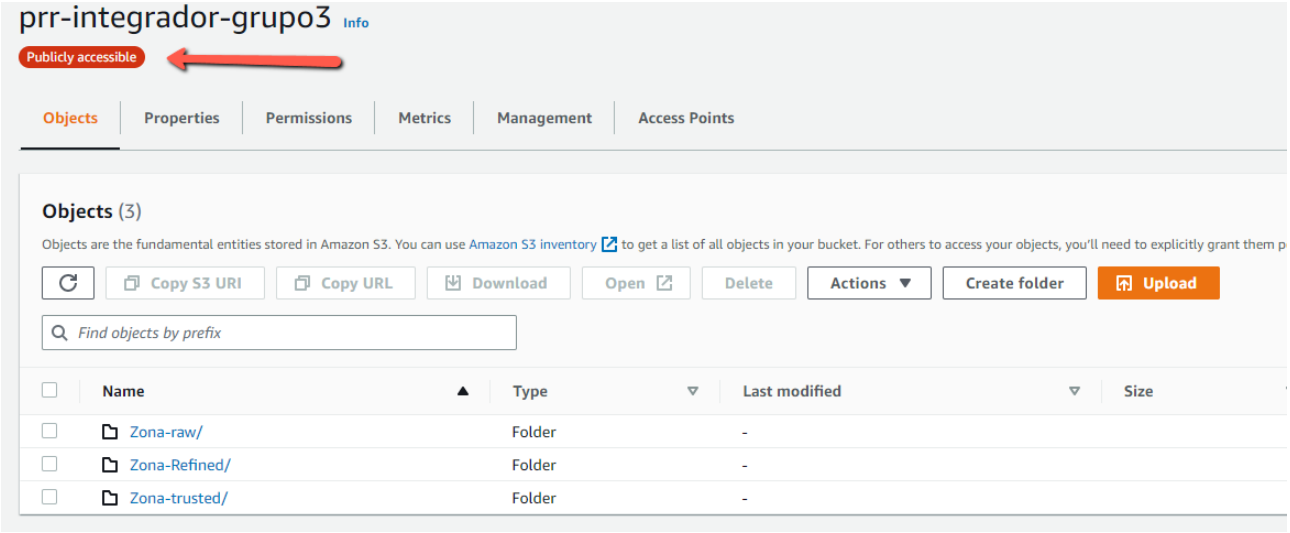


Figure 4: Bucket del proyecto

La URL para acceder al bucket público del proyecto es: https://s3.console.aws.amazon.com/s3/buckets/prr-integrador-grupo3?region=us-east-1tab=objects

Y en este se encuentran cada una de la zonas definidas en la figura anterior:

- **Zona-Raw**: Contienen los datasets necesarios para resolver el porblema planeato: runt diario, precios diarios, segmentos.

  En la **zona-raw** se tienen los siguientes puntos de datos: Precios/, RUNT/, Segmentos/

  - **Precios** contiene un archivo tipo csv con precios por modelos.
  - **RUNT** contiene un archivo .txt que describre las ventas diarias de motos nuevas en Colombia discriminadas por referencias.
  - **Segmentos** contienen un archivo tipo .csv; Contiene los segmentos a los que pertenecen cada una de las referencias de motos del mercado

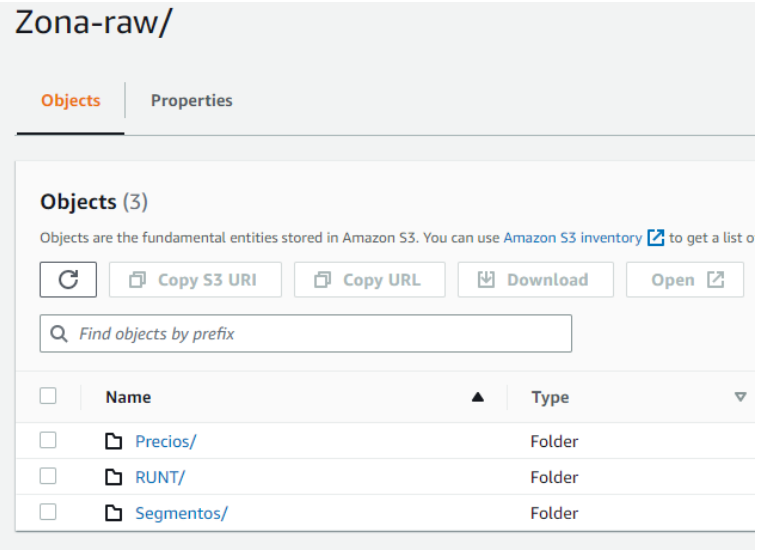


Figure 5: Bucket del proyecto: Zona Raw

- **Zona-Trusted**: Contiene la informacion procesada en datasets consolidados y organizados con estructura de fácil lectura. En esta zona se alojaron 3 tipos de archivo con diferentes versiones del dataset requerido:

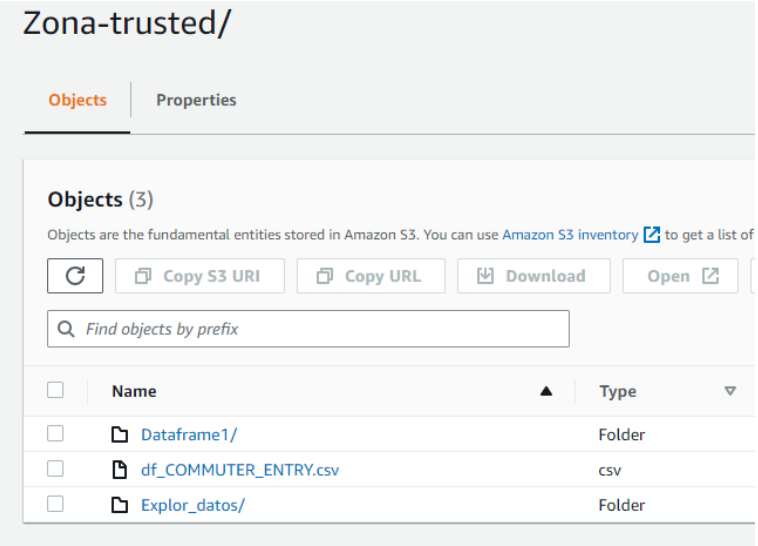


Figure 6: Bucket del proyecto: Zona trusted

- **Zona-Refined**: Contiene la informacion con la que se usa el modelo de datos y se guarda el refinamiento de estos para utilizarlo en la presentación. Este aloja los betas generados en el modelo de regresión y con los cuales se predecirán la unidades.

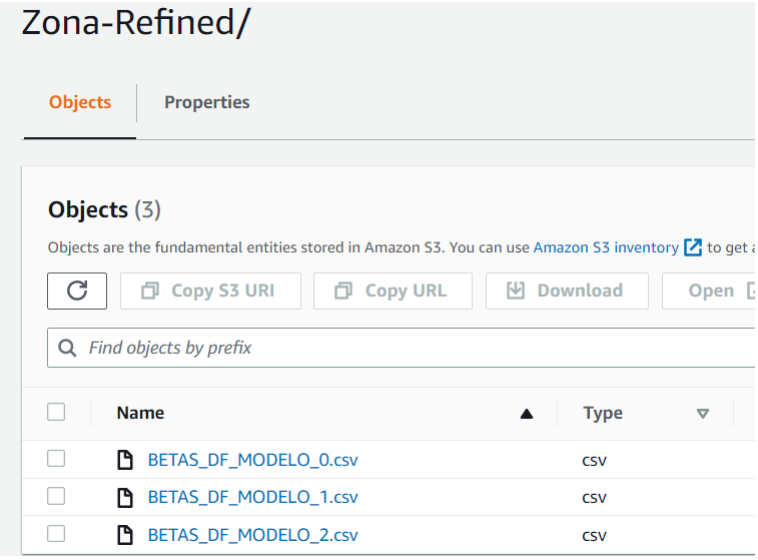


Figure 7: Bucket del proyecto: Zona refnined

2.1.3 Procesamiento

El procesamiento se realiza en dos fases. La primera fase corresponde al procesamiento de los datos alamcenados en la Zona Raw (histórico de precios, históricos de unidades vendidades, segmentación del mercado por referencias). Aquí se cataloga la información a través de Amazon Glue, y por medio del crawler designado se crean las tablas correspondientes:

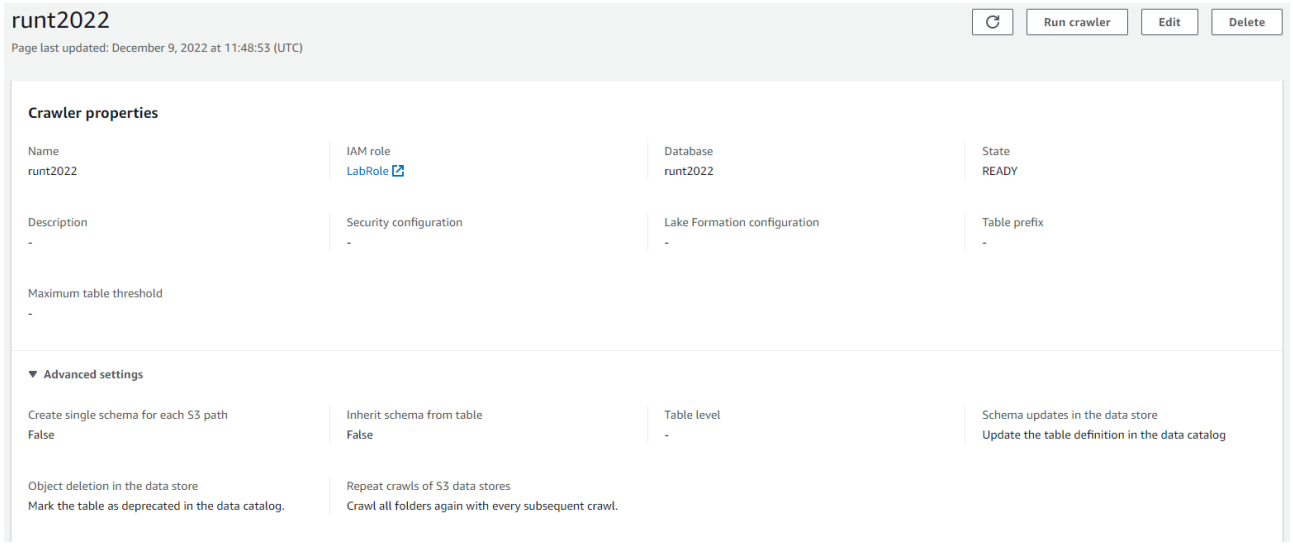
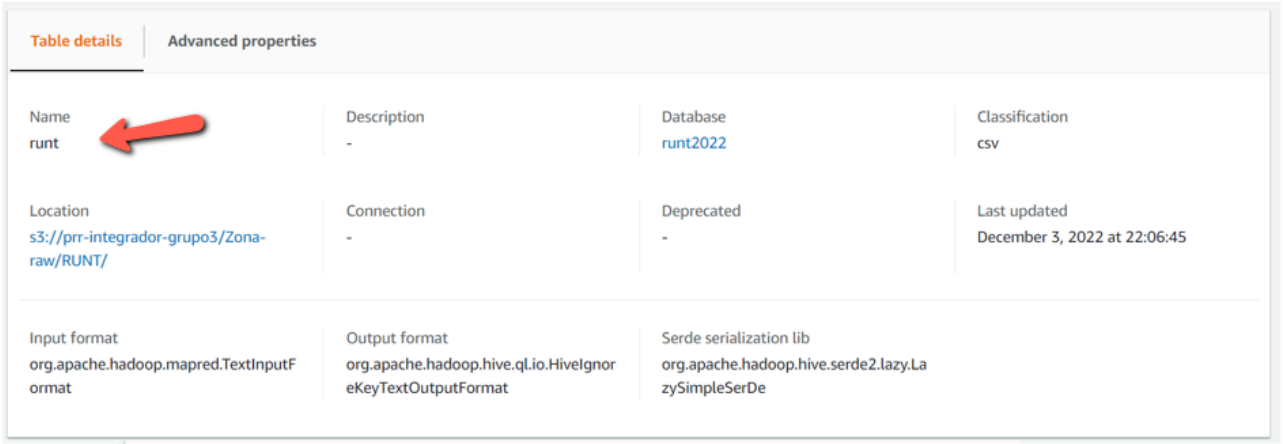


Figure 8: Crawler utilizado para catalogar los archivos de la zona Raw

Se crea la base de datos runt2022 y en ella se alojan las tablas y los esquemas de cada uno de los archivos, como se muestra en las siguientes figuras:

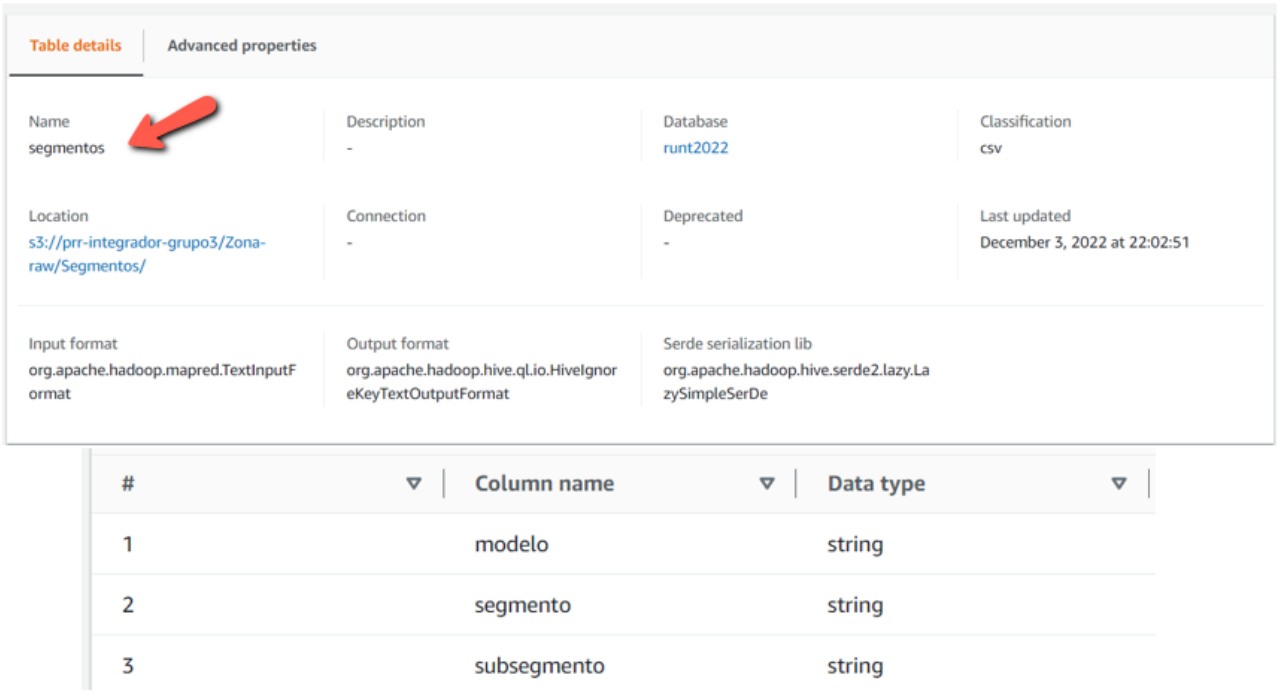


#	Column name	Data type
1	categoria	string
2	combustible	string
3	empresa	string
4	fecha	string
5	marca	string
6	modelo	string
7	cantidad	bigint

Figure 9: Tabla con las cantidades diarias vendidas

Para la tabla que se observa en la figura anterior, el crawler catalogó correctamente las columnas y el tipo de dato, excepto para la columna fecha; sin embargo esta columna que quedó mal catalogada preferimos transformarla directamente en Athena para utilizarla en el query.

La segunda tabla catalagoda fue la de segmentación:



#	Column name	Data type
1	modelo	string
2	segmento	string
3	subsegmento	string

Figure 10: Tabla con los segmentos del mercado por referencia de moto

La catalogación hecha por el crawler en esta tabla omitió los encabezados, por lo cual se debió editar las propiedades de la tabla de modo que reconociera la primera fila como el encabezado de la tabla de datos. Así mismo, se cambiaron los nombres de algunas columnas de modo que no hubieran titulos o caracteres especiales.

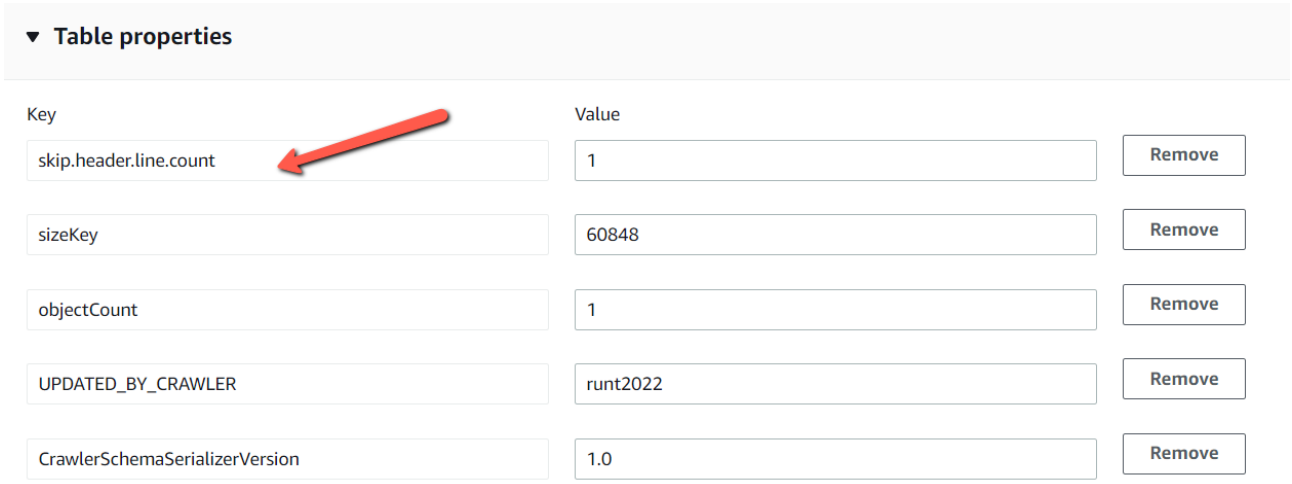
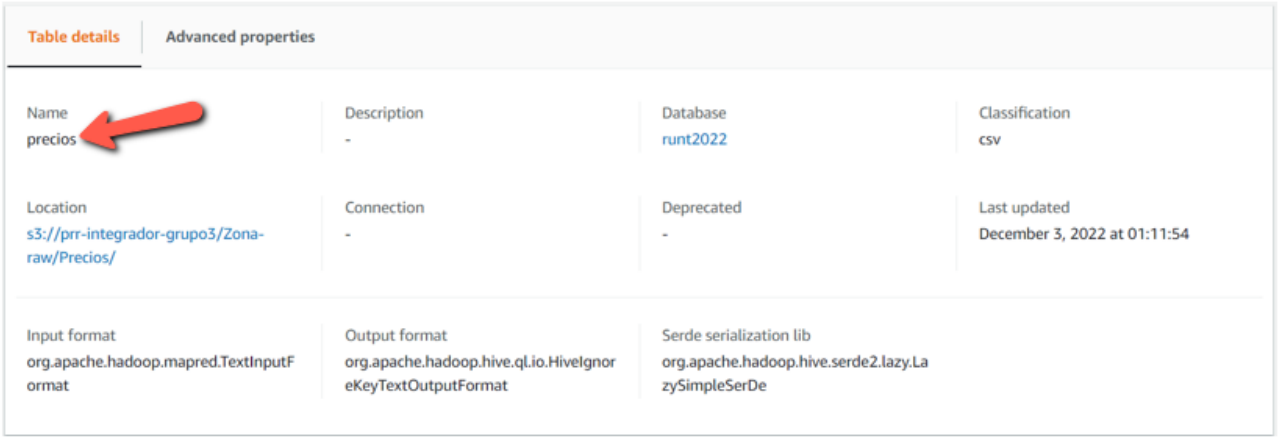


Figure 11: Modificación propiedades de la table Segmento

La última tabla catalogada fue la tabla de precios:



#	Column name	Data type
1	modelo	string
2	precio	bigint
3	fecha	date
4	variacion_precio	double

Figure 12: Tabla con los precios diarios de las motos

El crawler catalogó correctamente las columnas (incluso la de fecha). Para esta tabla sólo se cambiaron algunos nombres de modo que fuera mas sencillo construir un query desde Athena.

Después de tener las tablas catalogadas de manera correcta, se hizo uso de Athena para la primera exploración de los datos y la construcción de un dataframe consolidado que pudiera ser manipulable. Un ejemplo de la primera exploración de datos se observa en la siguiente figura:

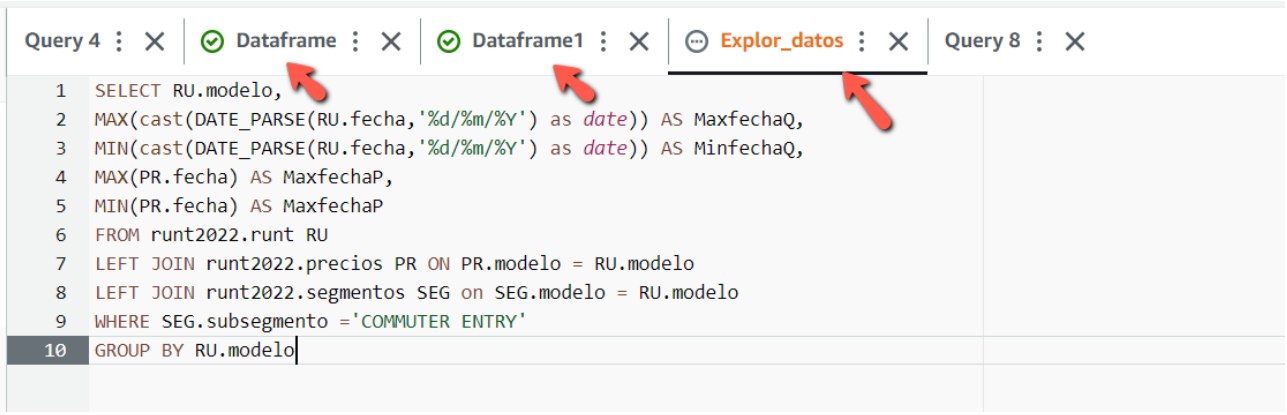


Figure 13: Ejemplo query: Exploración de datos con Athena

Con esta información se identificó que no todas las fechas existían para todas las referencias de motos, y que por tanto era necesario construir una extrapolación de datos en los peridodos donde fuera necesario.

Adicional, se construyó un query que fusionaba las 3 tablas anteriores y que permitía tener un primer dataframe en la zona trusted. Este dataframe se ejecutó y alamcenó desde Athema hasta el bucket de S3, editando las propiedades de la base de datos de modo que le apuntará a la ruta de la Zona trusted:

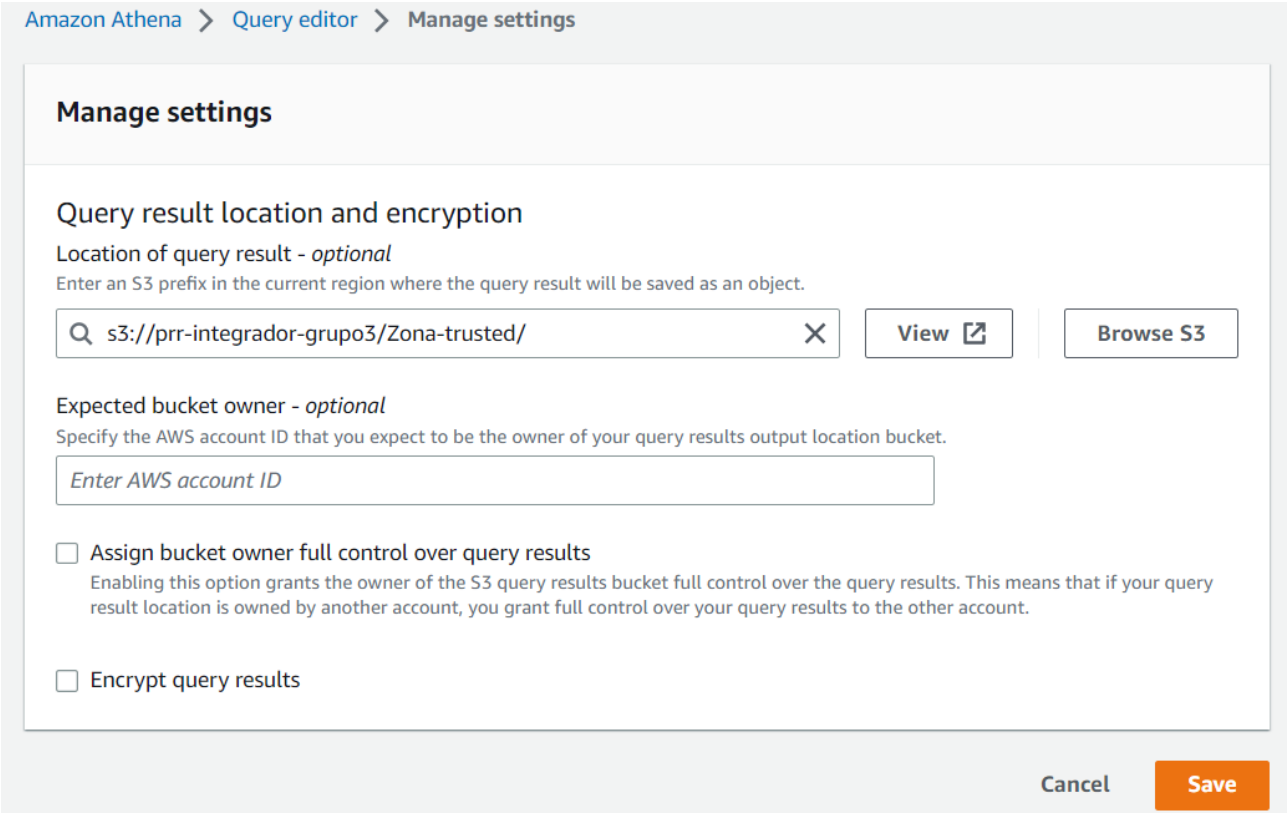


Figure 14: Conexión Athena con el bucket de S3

Los queries ejecutados se encuentran en los anexos a este documento.

Además de Athena, también se usó leguaje python para explorar los datos y contruir otro dataframe consolidado y depurado. Para este propósito, se hizo uso de ***Amazon Sagemaker studio***

Se utiliza ***sagemaker studio*** para implementar el modelo y las trasformación de la data, accesando la información desde los S3 anteriormente mencionados:

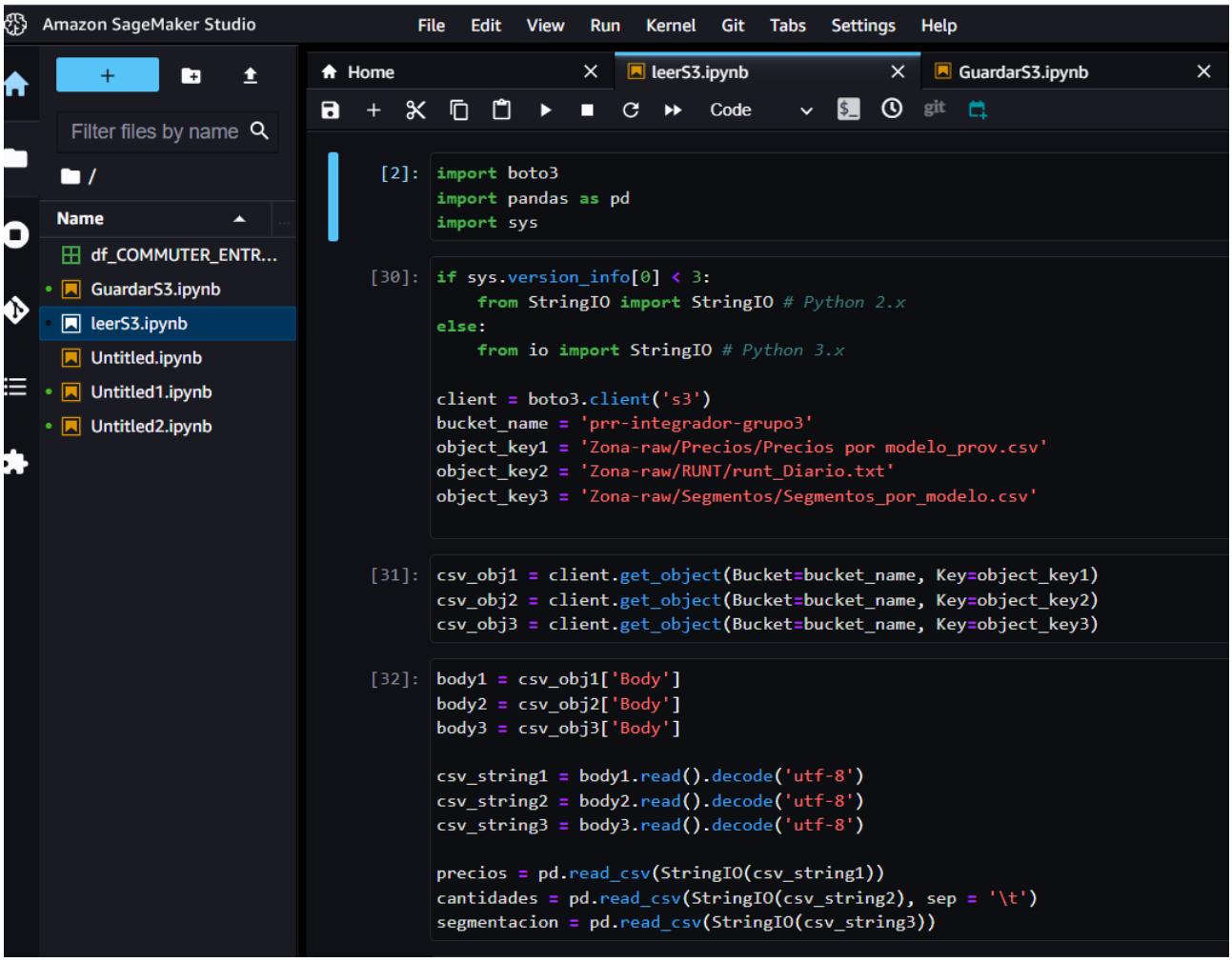


Figure 15: Ejemplo del uso de sagemaker: Lectura de los datos directamente a S3

El almacén de características de Amazon SageMaker es un repositorio completamente administrado y creado exclusivamente para almacenar, actualizar, recuperar y compartir características de aprendizaje automático (ML). Las características son entradas para los modelos de ML que se usan durante el entrenamiento y la inferencia.

Esta herramienta permitió gestionar de manera eficiente y ágil toda la programación, almacenamiento y uso de los datos, así como la implementación del modelo previamente configurado con las métricas y nuestro propio trabajo estadístico en lenguaje Python.

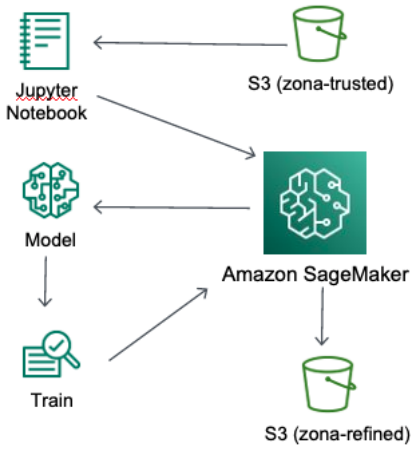
La limpieza del dataset hecha hasta aquí consta de :

- Identificación de outliers.
- Consolidación de los datos diarios en semanas.
- Selección de la mejor manera de agrupar los datos semanales (mínimo precio de la semana) y sumatoria de las unidades vendidas en a semana
- Eliminar las referencias de motos que no tienen un segmento definido.
- Filtrar por el subsegmento que nos interesa predecir (Commuter Entry).
- Eliminar los datos del periodo de pandemia 2020.
- Renombrar columnas para identificar por cada moto los precios y las cantidades.

Una vez se tiene el dataset limpio con la estructura requerida, se procede a entrenar un modelo que responda la pregunta del problema ¿Cuántas unidades va a vender Auteco?

La creación y entrenamiento del modelo también se hace a través de Sagemaker studio (notebook de python) leyendo los datos directamente del bucket del proyecto en la Zona trusted.

Se utiliza un notebook con la herramienta Jupyter Notebook para entrenar el modelo con nuestro propio código y métricas calculadas.



Teniendo el modelo entrenado, se procede a crear un dataframe que contenga los betas del modelo y con los cuales se hará la predicción de unidades a vender. Esta información se almacenará en el bucket del proyecto en la Zona refined.

2.1.4 Presentacion de datos

La presentacion del datos se realiza en modo de visualización con la herramienta Power BI. Usando el conector de power BI de amazon athena: (AmazonAWS®, 2022)

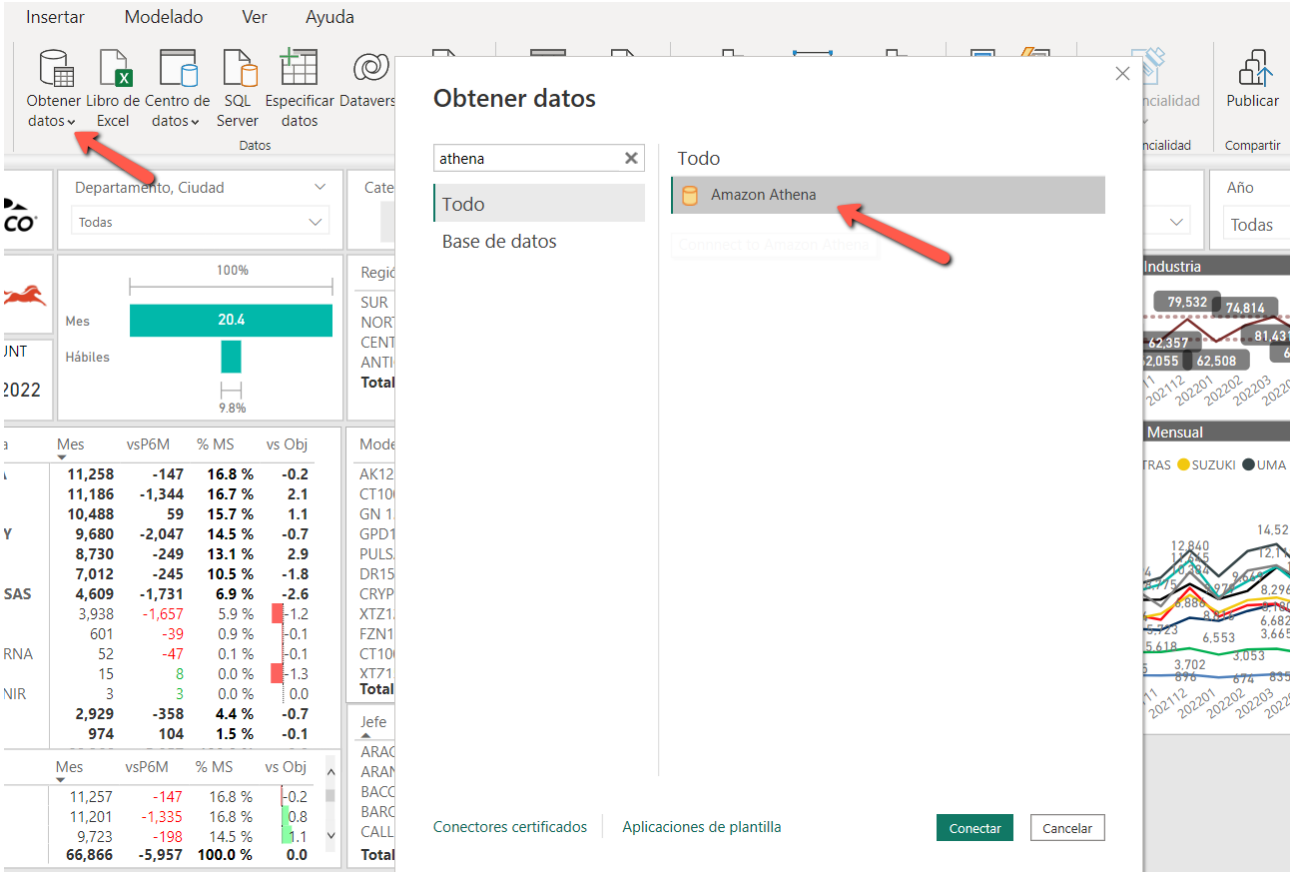


Figure 16: Conexión de Athena y Power BI

Se conectará la base de datos que aloja las ventas diarias, los segmentos y los betas del modelo para predecir las unidades que se venderán la próxima semana.

Con esta información disponible, se hará uso del lenguaje dax para calcular la predicción de unidades (aplicar los betas a las unidades vendidas por cada una de las referencias en el periodo anterior).

De esta manera, se procede a construir un dashboard que pueda ser leído por el equipo comercial de la compañía y que le permita tomar decisiones:

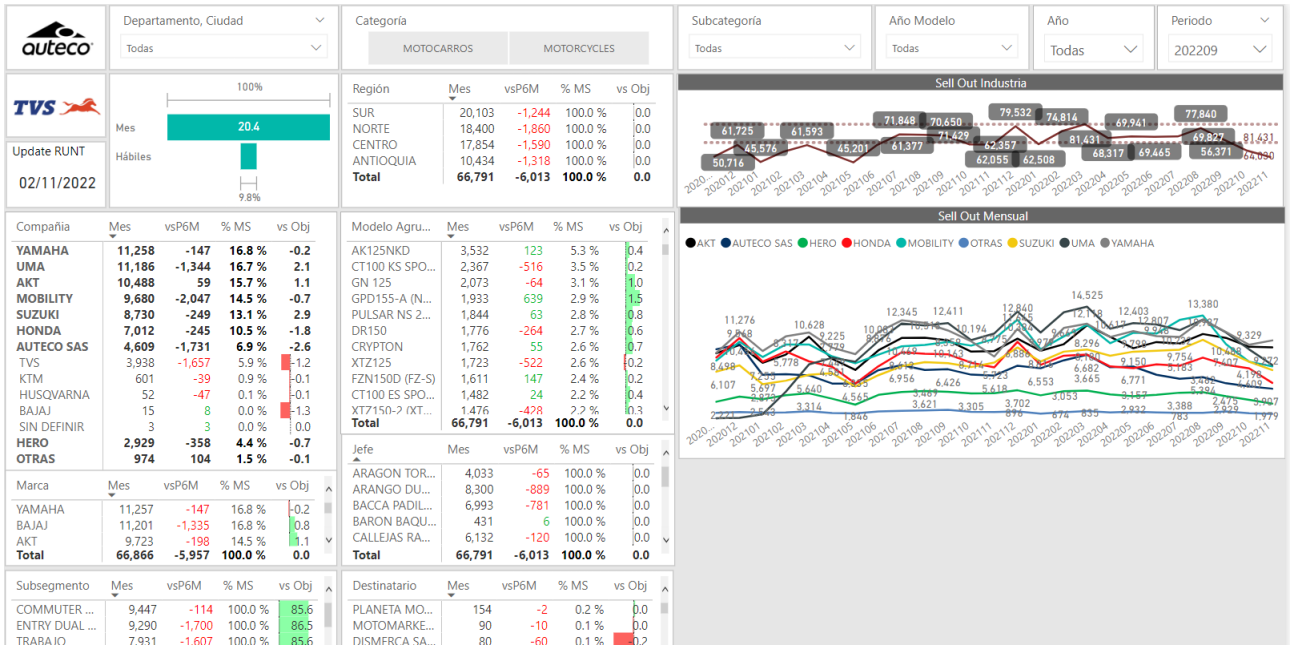


Figure 17: Dashboard en Power BI

3 Desarrollo Metodológico

3.1 Análisis Exploratorio de Datos

Se utilizaron datos diarios del runt para contabilizar la cantidad de motos vendidas para cada referencia de cada marca. Se eliminaron registros nulos, vacíos e inválidos. Se escogieron referencias del segmento COMMUTER ENTRY debido a que se tenía mayor disponibilidad de datos para este segmento desde el 2021. Finalmente, se escogieron las referencias más vendidas de cada una de las marcas. En la Tabla 1 se muestran las referencias con el número total de unidades vendidas.

Marca	Referencia	Cantidad
AKT	AK125 EIII	38962
HERO	ECO DELUXE CW	30986
SUZUKI	GN 125	52894
UMA	BOXER S	8179
YAMAHA	YC 110D (YC-Z)	16204
AUTECO	SPORT 100 ELS	26205

Table 1: Tabla referencias seleccionadas

En este caso nuestra variable a predecir es la referencia de Auteco SPORT 100 ELS. Como parte del pretratamiento de los datos, se removieron registros vacíos o nulos del dataframe. Adicionalmente, se organizaron los datos en semanas desde la primera semana del 2021 hasta la semana 44 del 2022.

Una vez reducido el dataframe a las columnas de interés se analizó cual de las variables tenía mayor variabilidad. En las Figuras 18 y 19 se muestra el cambio semanal en el precio y la cantidad de unidades vendidas respectivamente.

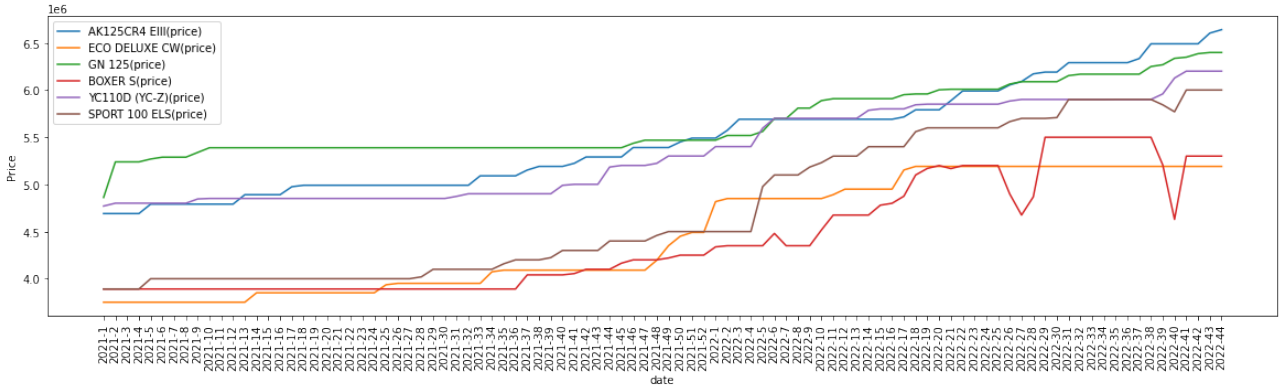


Figure 18: Variación semanal en los precios

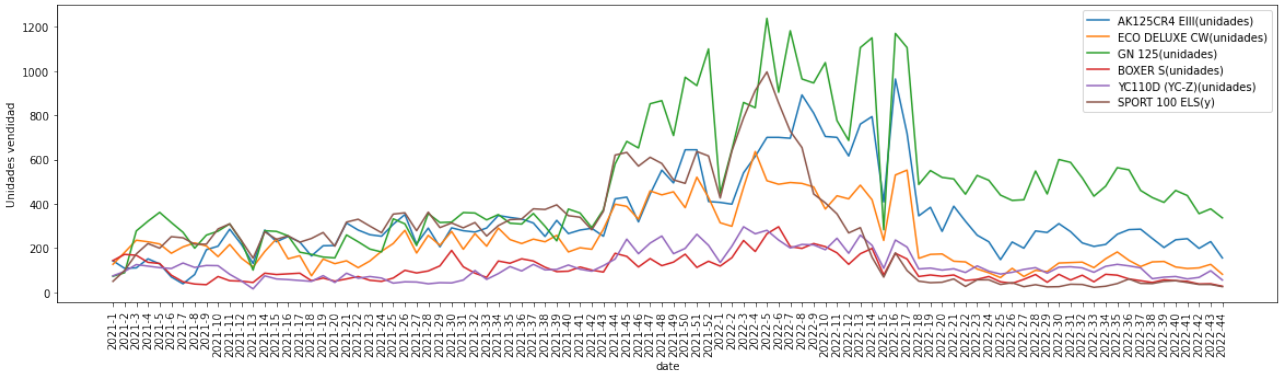


Figure 19: Variación semanal en las unidades vendidas

En las Figuras se evidencia que la variación en las unidades es más notoria, mientras que la variación en los precios tiende a mantenerse constante en largos periodos de tiempo. Por esta razón se decidió utilizar las unidades vendidas semanalmente para predecir la demanda de la referencia Sport 100 ELS en la semana siguiente. Además, se estableció un delay de una semana en la demanda de motos Sport 100 ELS y se utilizó como variable adicional.

Posteriormente se calculó la matriz de correlación del dataframe. En la Tabla 2 se muestran las variables que tienen mayor correlación con la variable objetivo (Sport 100 ELS delay)

Marca	**Referencia**	**Factor de Correlación**
AKT	AK125 EIII	0.41
HERO	ECO DELUXE CW	0.68
SUZUKI	GN 125	0.34
UMA	BOXER S	0.63
YAMAHA	YC 110D (YC-Z)	0.55
AUTECO	SPORT 100 ELS	0.95

Table 2: Tabla de correlación

Podemos observar que la cantidad de unidades vendidas de la referencia SPORT 100 ELS tiene una gran correlación con la cantidad de unidades vendidas en la siguiente semana.

Una vez obtenemos el factor de corrleación para la variable objetivo, continuamos calculando los coeficientes de correlación múltiple de cada variable con el resto. Este valor es una medida de que tanto es explicada una variable en términos de las otras. Los resultados se muestran en la Tabla 3

Marca	Referencia	Coeficiente de Correlación Múltiple
AKT	AK125 EIII	0.84
HERO	ECO DELUXE CW	0.88
SUZUKI	GN 125	0.87
UMA	BOXER S	0.68
YAMAHA	YC 110D (YC-Z)	0.83
AUTECO	SPORT 100 ELS	0.93
AUTECO	SPORT 100 ELS (delay)	0.92

Table 3: Tabla de correlación múltiple

Observamos que la variable delay tiene un factor de correlación múltiple de 0.92. Esto indica que es una variable que es explicada bien por las variables restantes.

De esta manera concluimos el Análisis Exploratorio de Datos, escogimos nuestra variable objetivo y las variables explicativas que van a ser la entrada para los diferentes modelos que serán descritos en la próxima sección.

3.2 Selección de Modelos

3.2.1 Regresión Lineal

El primer modelo seleccionado utilizado es una Regresión lineal, en el cual la Y son las unidades de SPORT 100 ESL de la siguiente semana, a partir de ahora se llamará y_{delay} a esta variable. Las variables explicativas X son las unidades vendidas de las marcas AKT, HERO, SUZUKI, UMA, YAMAHA y las unidades de SPORT 100 ESL vendidas esa semana. En la Ecuación 1 se observa la fórmula general de la regresión lineal (Peña, 2002/1).

$$Y = \beta_0 + X_1\beta_1 + ... + X_n\beta_n \tag{1}$$

Para aplicar el modelo se toman 10 semanas aleatorias del dataframe como grupo de testeo, y las 85 restantes utilizan como grupo de entrenamiento. En la Tabla 4 se muestran los betas calculados utilizando la regresión lineal con los datos del grupo de entrenamiento.

Betas	Valor
β_0	3.33×10
β_1	-9.01×10^{-02}
β_2	$-1,57 \times 10^{-02}$
β_3	-1.41×10^{-01}
β_4	1.85×10^{-01}
β_5	4.08×10^{-01}
β_6	9.81×10^{-01}

Table 4: Betas regresión lineal

Para evaluar que tan buenas son las predicciones de la regresión se utilizaron dos métricas: i) Mean Absolute Eror (MAE) y ii) Mean Absolute Percentage Error (MAPE). En las Ecuaciones 2 y 3 se definen el MAE y el MAPE respectivametente. En donde N es el número de puntos ajustados, x_i son las observaciones actuales y $\hat{x}_i$ son las observaciones estimadas (Peña, 2002/1).

$$MAE = \frac{\sum_{i=1}^{N}|x_i - \hat{x}_i|}{N} \tag{2}$$

$$MAPE = \frac{100\%}{N}\sum_{i=1}^{N}\frac{|x_i - \hat{x}_i|}{x_i} \tag{3}$$

En la Tabla 5 se muestran los valores calculados de MAE y MAPE para los conjuntos de entrenamiento y testeo.

	Entrenamiento	Testeo
MAE	42.10	60.20
MAPE	0.22	0.20

Table 5: MAE y MAPE de Regresión Lineal

Observamos que el valor de MAE para testeo es bastante alto, este valor de 60 se puede interpretar como el número de unidades en promedio en las que está fallando la predicción. Es decir, en promedio la predicción erra al valor verdadero por 60 unidades.

3.2.2 Regresión Ridge

Se seleccionó el modelo Ridge para realizar una segunda regresión y comparar los valores del MAE y el MAPE obtenidos con la regresión lineal. En la Tabla 6 se muestran los valores de MAE y MAPE obtenidos para la regresión Ridge con un parámetro alpha de 0.5.

	Entrenamiento	Testeo
MAE	42.10	60.30
MAPE	0.22	0.20

Table 6: MAE y MAPE de Regresión Ridge

Al comparar los resultados de MAE y MAPE de la regresión lineal y la regresión Ridge no se observa una mejora sustancial.

3.2.3 Eliminación de Outliers

Para eliminar datos anormales del dataframe, se utilizó como medida de distancia la distancia de Mahalanobis. En la Ecuación 4 se muestra la definición de la distancia de mahalanobis, en donde $\vec{x}$ y $\vec{y}$ son los vectore entre los cuales se calcula la distancia, y $\sum^{-1}$ es la matriz de covarianzas de los vectores (Peña, 2002/1).

$$d_m(\vec{x}, \vec{y}) = \sqrt{(\vec{x} - \vec{y})^T \sum^{-1} (\vec{x} - \vec{y})} \tag{4}$$

De esta manera, se calculó la distancia de mahalanobis entre los datos de entrenamiento y la media de los mismos, y se eliminó del conjunto de entranmiento al 20% más alejados. De esta manera se redujo el dataframe de entrenamiento de 85 semanas a 68 semanas.

Con el nuevo conjunto de entrenamiento se realizó nuevamente una regresión lineal. En la Tabla 7 se muestran los betas recalculados para la regresión.

Betas	Valor
β_0	4.41×10
β_1	-3.83×10^{-02}
β_2	$1,28 \times 10^{-01}$
β_3	-1.48×10^{-01}
β_4	2.96×10^{-01}
β_5	8.82×10^{-02}
β_6	8.61×10^{-01}

Table 7: Betas recalculados regresión lineal

Con los nuevos betas calculamos el MAE y el MAPE nuevamente. Los resultados se muestran en la Tabla 8

	Entrenamiento	Testeo
MAE	30.82	45.70
MAPE	0.22	0.15

Table 8: MAE y MAPE Regresión Lineal sin outliers

Se observa una mejora considerable en el MAE de entrenamiento, en el MAE y MAPE de testeo. En este punto el modelo la diferencia entre la demanda estimada y la real es de 46 unidades en promedio.

3.2.4 Análisis de Componentes Principales (PCA)

Para intentar mejorar las predicciones, se decidió reducir la dimensionalidad de las variables explicativas. Para cumplir este objetivo se utilizó un Análisis de Componentes Principales (PCA). El primer paso para llevar a cabo el PCA fue escalar y transformar los vectores X y Y de entrenamiento. Una vez finalizado este proceso se aplicó el algoritmo PCA inicialmente con 2 componentes. Se calcula la varianza explicada de las dos primeras componentes obteniendo un valor de 0.89.

Se calculan las proyecciones del vector X escalado sobre las dos primeras componentes. Como resultado se obtuvieron los vectores $Z1$ y $Z2$. En la Figura 20 se observa la gráfica de los vectores $Z1$ y $Z2$.

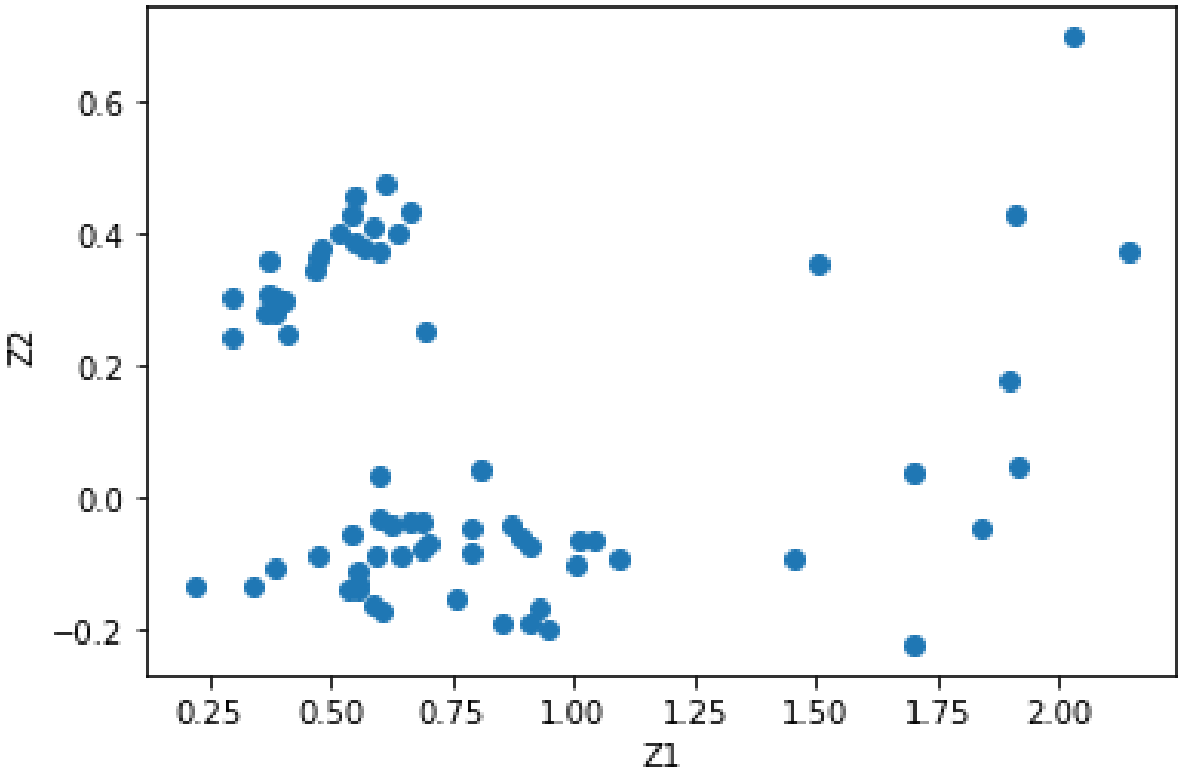


Figure 20: Proyección del vector X sobre las dos primeras componentes

Con los nuevos vectores $Z1$, $Z2$ y el vector Y escalado se realiza nuevamente la regresión lineal. En la Tabla 9 se muestran los nuevos valores de los betas.

Betas	Valor
β_0	0.11
β_1	0.39
β_2	−0.75

Table 9: Betas recalculados PCA

Las predicciones obtenidas con la regresión tienen aplicado el escalamiento inicial que se le realizó al vector Y. Por esta razón se aplica una transformación escalar inversa para que las predicciones sean comparables con el vector original. Finalmente en la Tabla 12 se muestra el MAE y el MAPE de entrenamiento después de haber aplicado PCA.

	Entrenamiento
MAE	33.89
MAPE	0.22

Table 10: MAE y MAPE PCA entrenamiento

3.2.5 Algoritmo K-means

Con el objetivo de seguir mejorando las predicciones se decidió modificar la gráfica de $Z1$ y $Z2$ añadiendole un indicador de color relacionado con los valores del vector Y que es el número de unidades vendidas que queremos predecir. El resultado se observa en la Figura 21

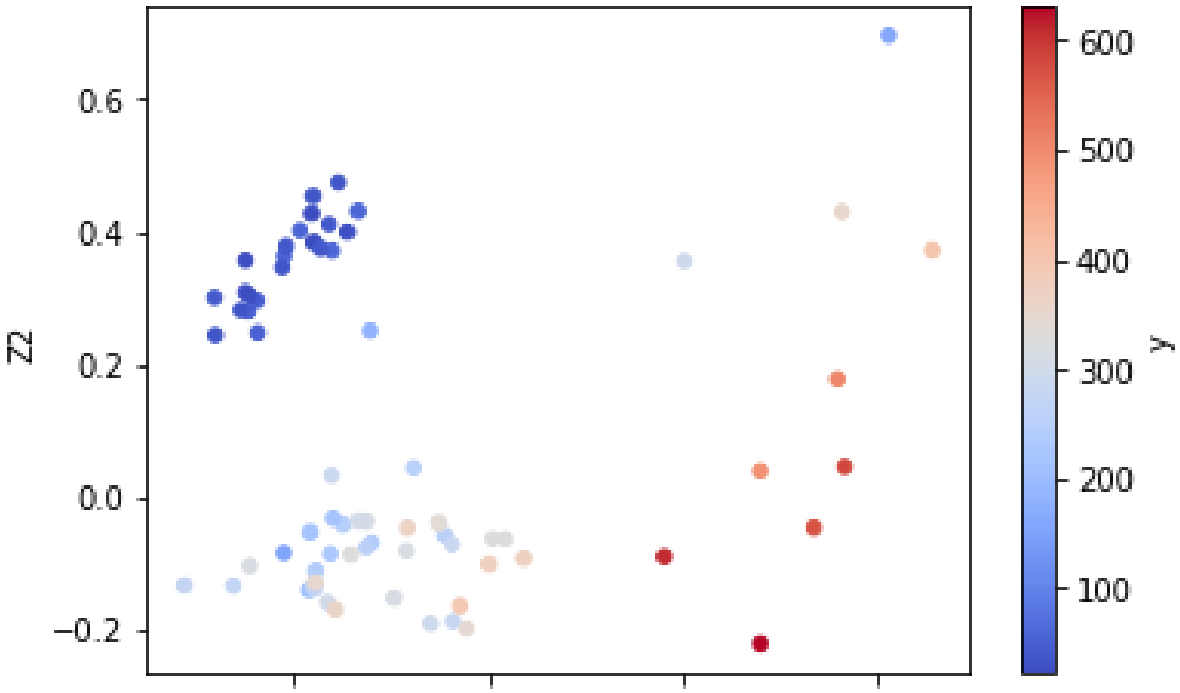


Figure 21: Mapa de color Z1 y Z2

Podemos observar que se forma un cluster bastante sólido de puntos azules en la parte superior izquierda, adicionalmente hay un cluster un poco más heterogéneo en la parte inferior izquierda y un último grupo un poco más disperso que abarca la parte derecha.

Con los resultados de la figura anterior, se decidió aplicar sobre los vectores $Z1$ y $Z2$ el algoritmo de K-means utilizando la distancia de mahalanobis y definiendo 3 grupos, con el objetivo de clusterizar los datos. El algoritmo aplicado fue programado por el equipo de trabajo y se utilizó la distancia de mahalanobis en lugar de la distancia euclídea. En la Figura 22 se muestran los clusters formados por el algoritmo.

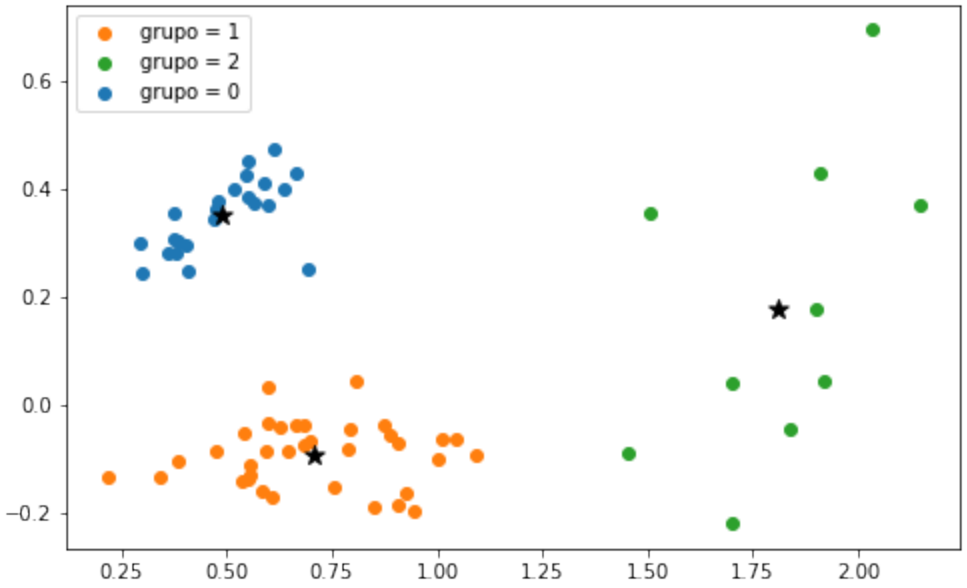


Figure 22: Resultados clusterización K-means

Podemos observar la formación clara de tres clusters, los centroides de cada uno se marcan en la gráfica. Una vez clusterizados los datos, se dividió el data frame de entrenamiento en tres data frames para cada grupo y se realizó la regresión lineal. En la Tabla 11 se muestran los betas calculados.

	Grupo 0	**Grupo 1**	**Grupo 2**
β_0	23.80	$1.63 \times 10^{+02}$	$2.70 \times 10^{+02}$
β_1	-0.04	1.77×10^{-01}	9.71×10^{-02}
β_2	0.39	-1.02×10^{-01}	-1.46×10^{-01}
β_3	-0.20	-2.25×10^{-01}	-7.49×10^{-02}
β_4	0.32	7.20×10^{-01}	-4.29×10^{-01}
β_5	0.13	3.06×10^{-01}	-9.78×10^{-01}
β_6	0.52	2.84×10^{-01}	1.07

Table 11: Betas recalculados clusterización

Con los nuevos betas de las regresiones para cada grupo se calculó el MAE y el MAPE de entrenamiento. Los resultados se muestran en la Tabla.

	Grupo 0	**Grupo 1**	**Grupo 2**
MAE	11.86	33.30	16.52
MAPE	0.33	0.12	0.04

Table 12: MAE y MAPE K-means entrenamiento

Observamos que el MAE para los grupos 0 y 2 mejora considerablemente, esto es congruente con lo observado en la Figura 21 en donde estos grupos tienen colores más homogéneos.

Una vez calculados las métricas para el conjunto de entrenamiento, el siguiente paso fue revisar el conjunto de testeo. Inicialmente, se tomó el data frame de testeo y se escalaron sus valores. Posteriormente se proyectaron los datos sobre los vectores $Z1$ y $Z2$ y se graficaron. En la Figura 23 se muestra la proyección del conjunto de entrenamiento sobre los vectores de las componentes principales.

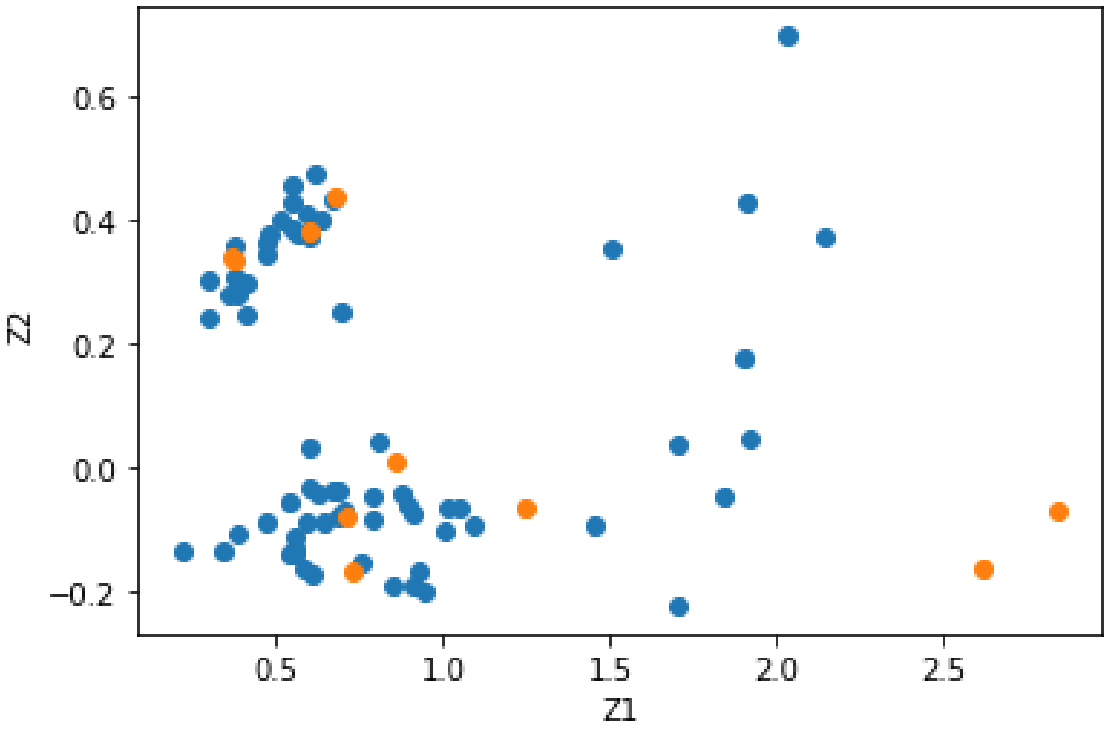


Figure 23: Proyección conjunto de testeo

Los puntos anaranjados son la proyección de los puntos de testeo que se tomaron. Posteriormente se aplicó el algoritmo de k-means y se asignó cada punto a un determinado grupo. Utilizando los betas calculados con la regresión se obtuvieron los resultados estimados y se calculó el MAE y el MAPE para el conjunto de testeo. Los resultados se muestran en la Tabla 13

	Grupo 0	**Grupo 1**	**Grupo 2**
MAE	12.14	124.46	17.42
MAPE	0.30	0.32	0.02

Table 13: MAE y MAPE K-means testeo

Finalmente, se calculó el MAE y el MAPE general, sin distinción por grupos para el conjunto de testeo. En la Tabla 14 se muestran los resultados

	Testeo
MAE	13.90
MAPE	0.20

Table 14: MAE y MAPE general

Con este resultado, observamos una mejora considerable desde la primera regresión. En este punto la diferencia entre la demanda estimada y la real es de 14 unidades en promedio. Inicialmente se encontraba en 60 unidades en promedio.

3.3 Análisis y Conclusiones

3.3.1 Análisis de variabilidad de los betas

Se analizó la varianza de cada uno los betas tomando muestras aleatorias del conjunto de test (10 semanas), por tanto, el conjunto de train se modificaba con cada muestreo, se realizaron 1000 muestreos diferentes para los cuales se obtuvieron los siguientes graficos (Figura 24).

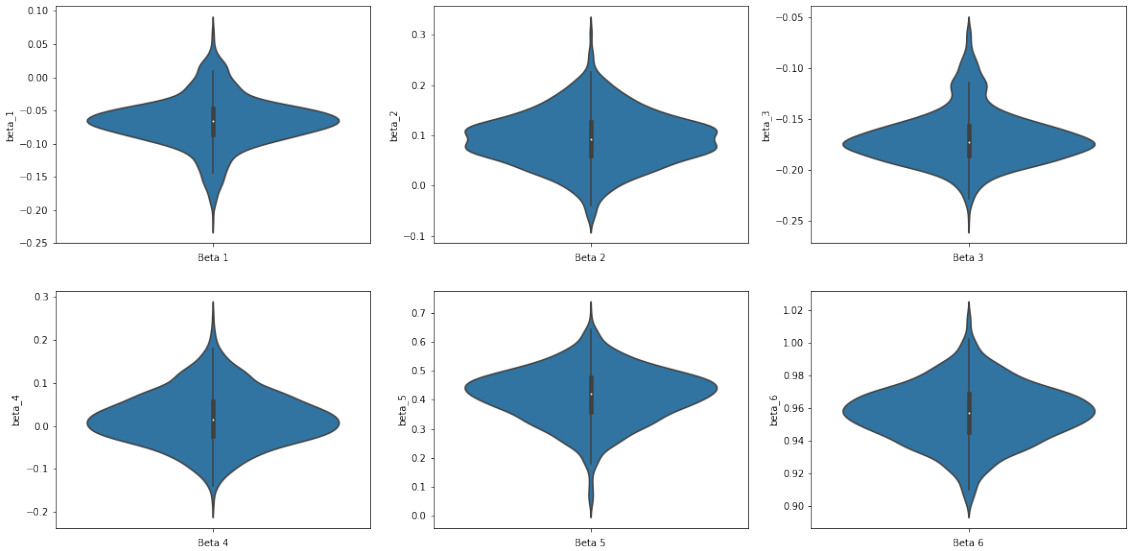


Figure 24: Variación de los betas

Se observa que el valor de los betas no varia mucho y se concentra alrededor de la media independientemente del conjunto de train y test.

Adicionalmente, se analizó la varianza del intercepto b. El resultado se muestra en la Figura 25

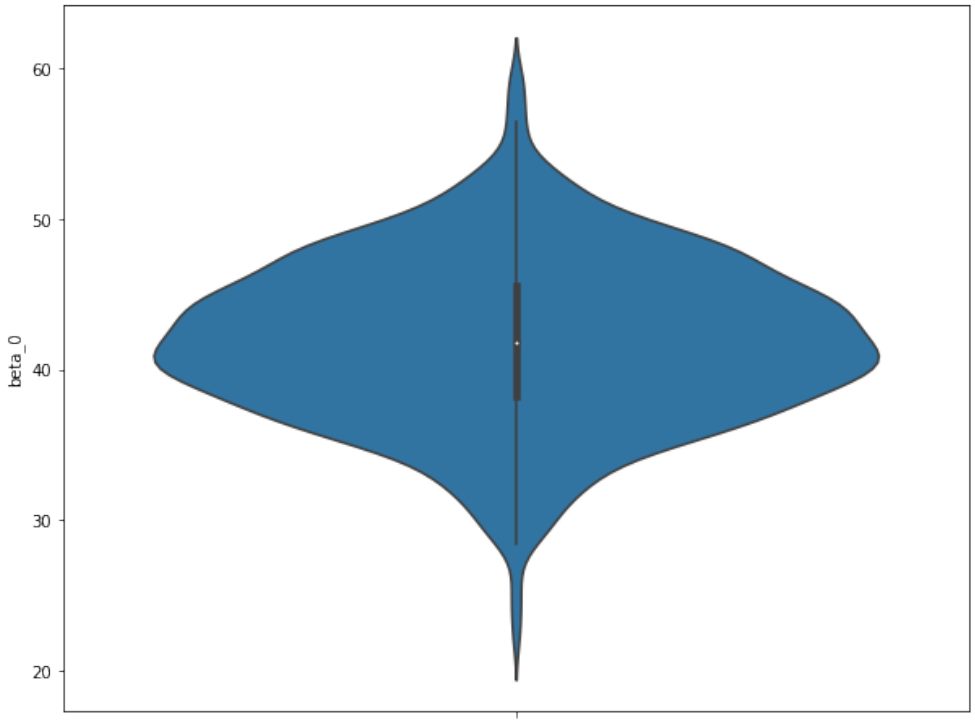


Figure 25: Variación intercepto b

4 Conclusiones

- Usando las técnicas vistas en los diferentes cursos, se logró implementar un modelo que combina la regresión lineal, PCA y clusterización capaz de predecir el número de motos vendidas por AUTECO cada semana, llegando a tener un nivel de error de tan solo 12 unidades por semana (MAE 12).

- Este modelo es una buena proyección para el futuro, cuando se tengan más datos se puede generar un entrenamiento más robusto; por el momento sólo se tienen 96 semanas lo cual es una muestra muy pequeña de la población, ya que los datos durante la pandemia no son aportantes para el modelo.
- El modelo no necesita ser entrenado como una serie de tiempo, es decir, no importan los registros pasados, lo único que importa es conocer las unidades vendidas de la marca y su competencia en cada segmento la semana inmediatamente anterior. Esto permite que los datos de train y test puedan ser escogidos aleatoriamente en cada entrenamiento y garantizar que el modelo converge a los mismos resultados, ya que los betas tienen poca varianza ante cambios en el set de entrenamiento.
- La naturaleza de los datos utilizados para este modelo son estructurados, lo cual permite que sean consultados através de lenguaje SQL
- La arquitectura se diseñó de tipo Batch, si bien los datos se podrían obtener en línea, el modelo necesita un rezago de una semana para ser calculado y el equipo comercial toma decisiones con las ventas semanales por lo que no se ve la necesidad de incrementar los costos con una arquitectura de ingesta en streaming.
- Se propone un despliegue del modelo en AWS combinado con Power BI. Esto debido a que AWS nos da los criterios de seguridad y rapidez que requiere el proyecto. Power BI se selecciona por el costo de la herramienta y por la facilidad de manejo (los jefes comerciales están acostumbrados a su interfaz visual).
- EAFIT$_e studiantesMCDA$®, 2022

References

AmazonAWS®. (2022). Conector power bi con aws [Accedido en octubre de 2022].
auteco®. (2022). Auteco [Accedido en octubre de 2022].
EAFIT$_e studiantesMCDA$®. (2022). Proyecto integrador ciencias de dato y analitica [Accedido en diciembre de 2022].
GrupoUma®. (2022). Grupouma [Accedido en octubre de 2022].
Heromotos®. (2022). Heromotos [Accedido en octubre de 2022].
motos®, A. (2022). Akt motos [Accedido en octubre de 2022].
Peña, D. (2002/1). *Análisis de datos multivariantes* (e*, Vol. v*) [Notas y observaciones*]. McGraw-hill.
suzuki®. (2022). Suzuki [Accedido en octubre de 2022].
Yamaha®, I. (2022). Bibtex [Accedido en octubre de 2022].