# Proyecto Final Curso Data Science – CODERHOUSE

## ANÁLISIS DE VENTAS EN AMAZON PARA UNA EMPRESA QUE OPERA EN EL MERCADO CANADIENSE

**FELIPE BORDAGARAY LANCIERI**

**MAYO DE 2024**

**Índice**

## 1. ABSTRACT

El objetivo de este análisis será el de determinar si existen patrones estacionales en las ventas de los productos de la compañía, así como también determinar si existe correlación y/o causalidad entre las ventas y ciertos fatores externos como la temperatura máxima promedio y las horas de luz.

El público objetivo del análisis es la gerencia de la empresa que está interesada en conocer aspectos ocultos de los patrones de consumo de sus productos.

Las conclusiones de este trabajo le deberían permitir tomar decisiones de alto impacto en la compañía como lo son:

-La determinación de niveles de inventarios óptimos,

-La decisión de discontinuar ciertas marcas o productos

-La planificación de la producción ante variables externas que determinen cierto grado de causalidad sobre las ventas

## 2. Objetivo

El objetivo principal del trabajo es poder predecir la estacionalidad de la venta de los productos de la empresa, tanto durante las distintas estaciones del año como cuando suceden fenómenos externos como condiciones climáticas y horas de luz del día.

También nos interesa saber el impacto real en las ventas de las visitas totales a las publicaciones para poder maximizar los presupuestos de publicidad y pay-per-click.

## 3. Contexto y Audiencia

Esta base fue obtenida de la empresa en la que trabajo actualmente y refiere a 4 años de venta diaria de una variedad de productos.

Es una empresa que se dedica a la reventa online de diversos productos de diversas industrias, con la característica de tener exclusividad en el canal para el país en cuestión.

La empresa se encuentra constantemente reevaluando proveedores para maximizar sus ganancias y a la vez tener un mix de productos contra estacionales entre sí para asegurar la rentabilidad durante todo el año.

La base de datos combina Ventas totales en unidades y $, ventas para cada una de las 3 marcas (HB, RYE y MED) en unidades y $, page views (equivalente a clicks en la pagina de compra del producto), y datos climatológicos promedio de la zona, todo agrupado por día.

## 4. Hipótesis y preguntas de interés

El dataset elegido corresponde a las ventas por día en la plataforma AMAZON de una empresa que cuenta con una línea de productos. Dentro del dataset hay información sobre las unidades vendidas, las visitas a las publicaciones. En principio establecemos las siguientes hipótesis:

1) Las unidades vendidas son directamente proporcionales a la cantidad de visitas.

2) Los productos tienen una estacionalidad positiva en el verano (de Junio a Septiembre)

3) Las horas de luz del día, pueden afectar el comportamiento de los consumidores

4) El inicio de la pandemia (Marzo 2020) y el inicio de la guerra de Ucrania (FEB-MAR 2022) afectaron a las ventas

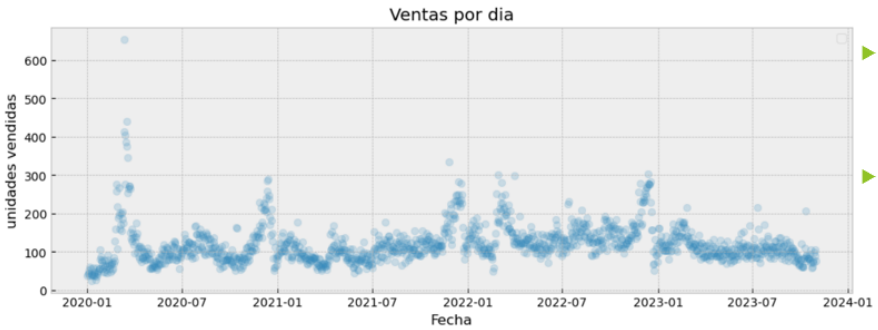
## 5. MetaData

Tipos de datos a utilizar en el presente trabajo:

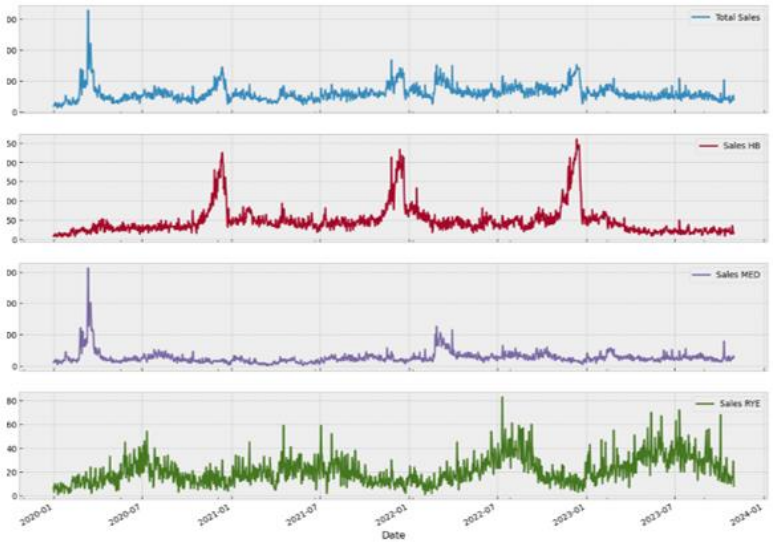
```
AmzProd.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1399 entries, 0 to 1398
Data columns (total 27 columns):
 #   Column                                      Non-Null Count  Dtype
---  ------                                      --------------  -----
 0   Date                                        1399 non-null   object
 1   Sales ($) Heartbeat Hot Sauce               1399 non-null   float64
 2   Sales (units) Heartbeat Hot Sauce           1399 non-null   int64
 3   Avg item price ($/unit) Heartbeat Hot Sauce  1399 non-null   float64
 4   Sessions Heartbeat Hot Sauce                1399 non-null   int64
 5   Page Views Heartbeat Hot Sauce              1399 non-null   int64
 6   Sales ($) Medentech                         1399 non-null   float64
 7   Sales (units) Medentech                     1399 non-null   int64
 8   Avg item price ($/unit) Medentech           1399 non-null   float64
 9   Sessions Medentech                          1399 non-null   int64
 10  Page Views Medentech                        1399 non-null   int64
 11  Sales ($) Ryeka                             1398 non-null   float64
 12  Sales (units) Ryeka                         1398 non-null   float64
 13  Avg item price ($/unit) Ryeka               1398 non-null   float64
 14  Sessions Ryeka                              1398 non-null   float64
 15  Page Views Ryeka                            1398 non-null   float64
 16  Total Sales ($)                             1399 non-null   float64
 17  Total Sales (units)                         1399 non-null   int64
 18  Avg item price ($/unit)                     1399 non-null   float64
 19  Total Sessions                              1399 non-null   int64
 20  Total Page Views                            1399 non-null   int64
 21  max_temperature                             1399 non-null   float64
 22  avg_temperature                             1399 non-null   float64
 23  min_temperature                             1399 non-null   float64
 24  rain                                        1399 non-null   float64
 25  snow                                        1399 non-null   float64
 26  daylight                                    1399 non-null   float64
dtypes: float64(17), int64(9), object(1)
memory usage: 295.2+ KB
```

Los datos son en su totalidad numéricos, por lo que los análisis serán orientados al tipo cuantitativos.




- Tenemos una base de datos de casi 4 años completes de ventas diarias en las plataformas identificadas
- Vamos a analizar el comportamiento por cada una de las 3 lineas de productos asignadas a el presente análisis




- Las 3 lineas de productos son
- HB: una linea de salsas de mesa
- MED: una linea de productos para la purificación de agua
- RYE: una linea de productos de suplementos alimenticios para entrenamiento profesional

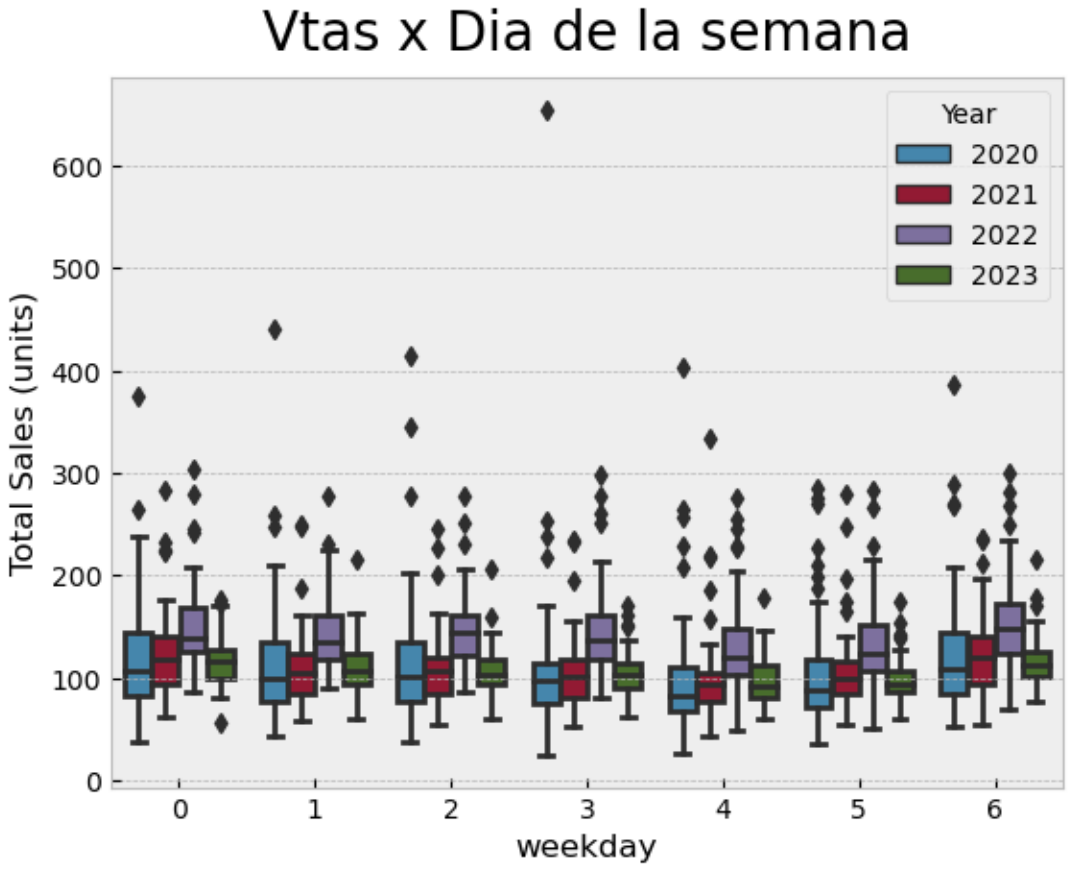
## 6. EDA: Análisis exploratorio de Datos

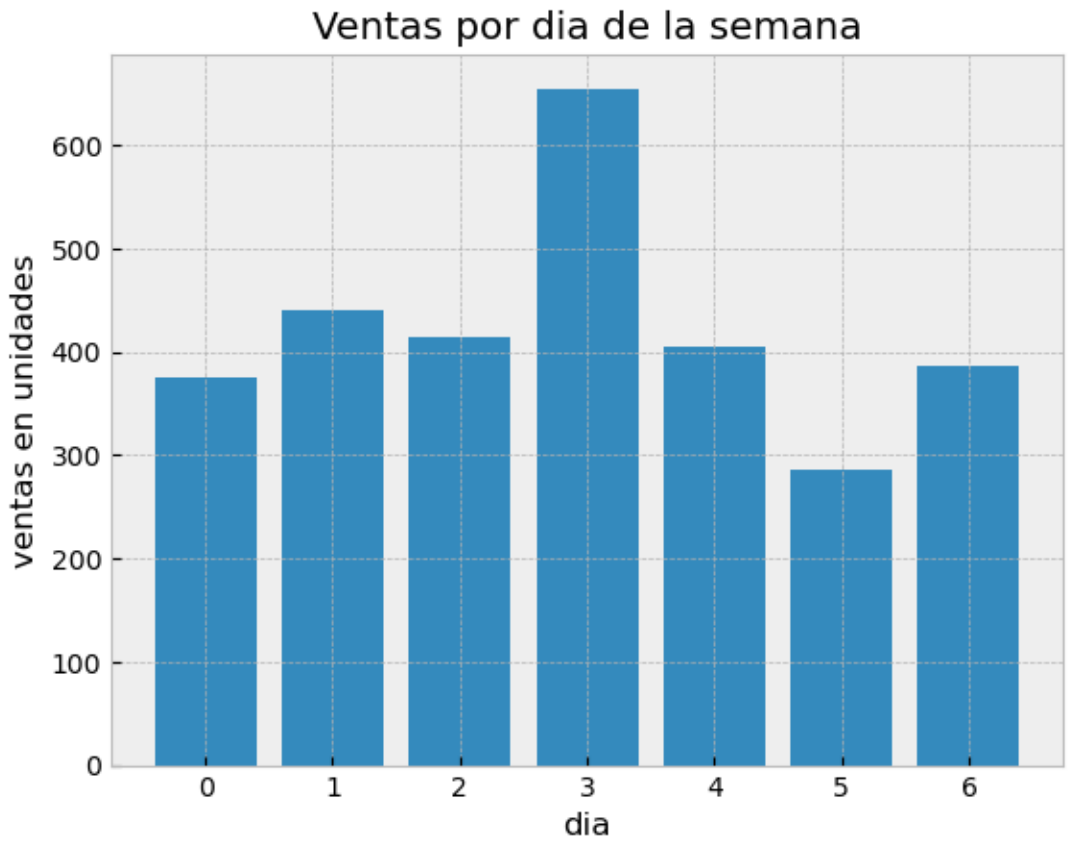
A continuación, se realizará un análisis exploratorio de los datos para comenzar a encontrar patrones, características y correlaciones en los datos.

Tanto el análisis EDA como el subsiguiente análisis bivariado y multivariado se realizarán utilizando las bases de datos de las 3 líneas de productos involucradas.

Luego a partir de las conclusiones preliminares, nos enfocaremos en una de las marcas para implementar modelos de MACHINE LEARNING.

Las tres marcas se analizan según su evolución de ventas en los distintos años.



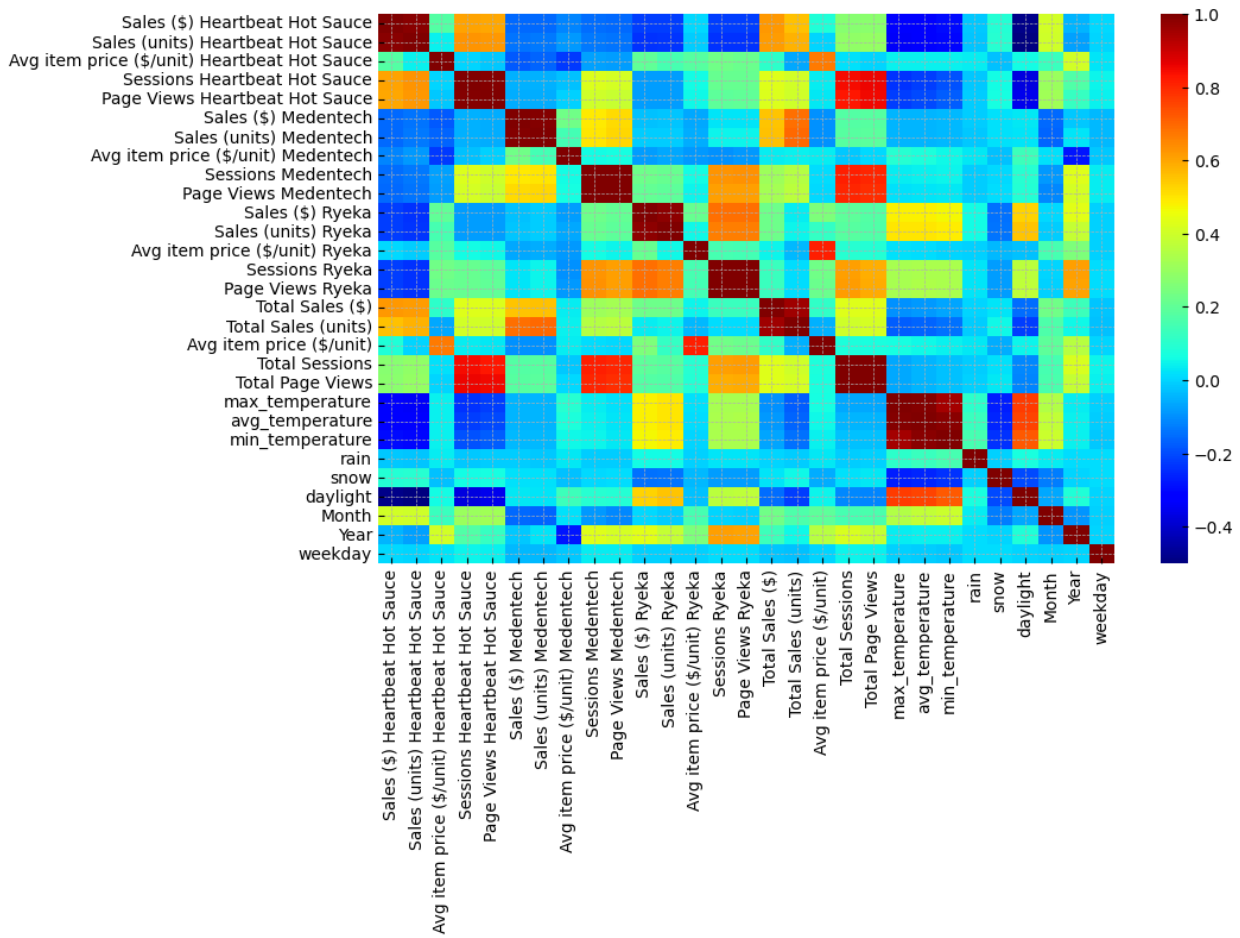



En este grafico se observa que la mayoría de las correlaciones fuertes son de variables que expresan cosas similares u obvias, como por ejemplo las horas de lux y las temperaturas, o las sesiones con las page views de los usuarios.

Pero podemos identificar algunas correlaciones mas leves que podrían ayudar a encarar el análisis:

-correlación negativa entre Daylight con Ventas de la marca HB.

-correlación positiva entre la temperatura y las ventas de la marca RYE.



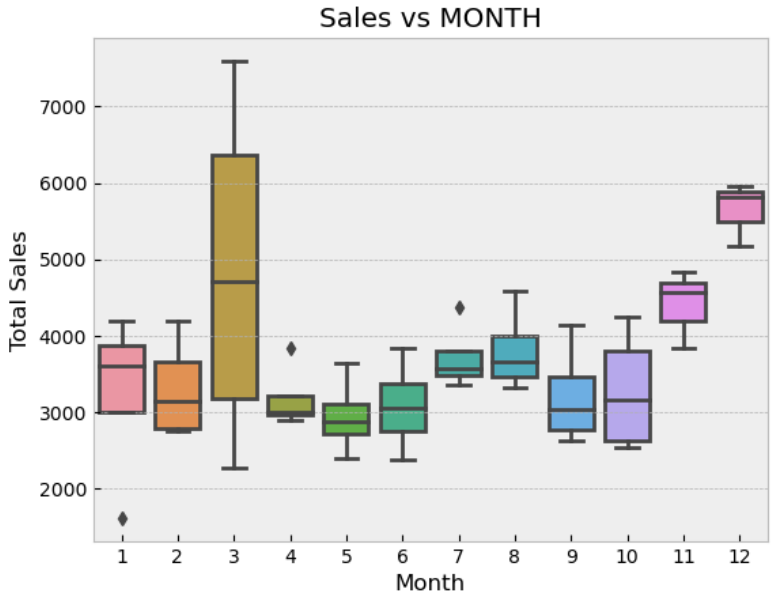

En este grafico se observa que la mayoría de las correlaciones fuertes son de variables que expresan cosas similares u obvias, como por ejemplo las horas de lux y las temperaturas, o las sesiones con las page views de los usuarios.

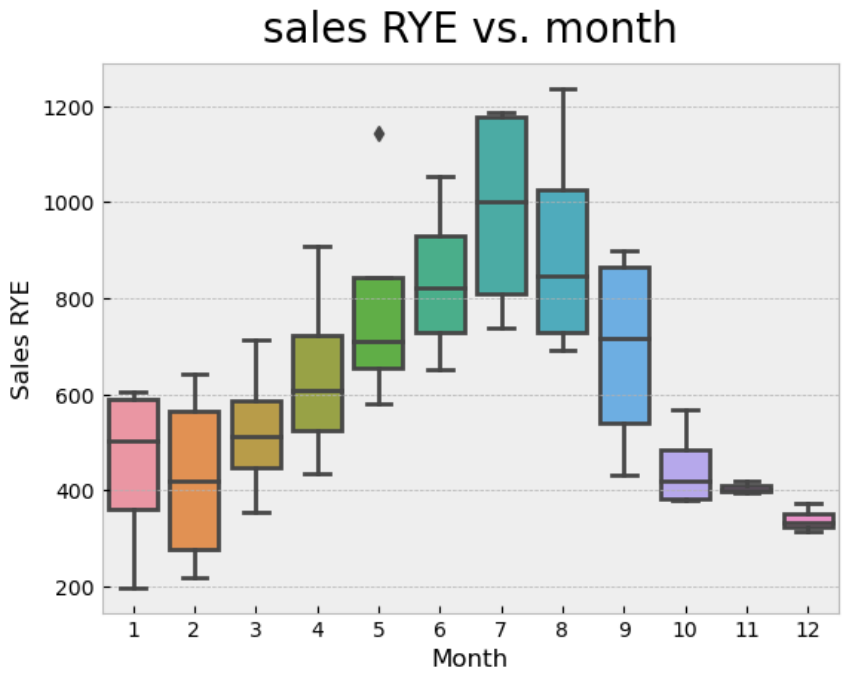
Pero podemos identificar algunas correlaciones mas leves que podrían ayudar a encarar el análisis:

-correlación negativa entre Daylight con Ventas de la marca HB.

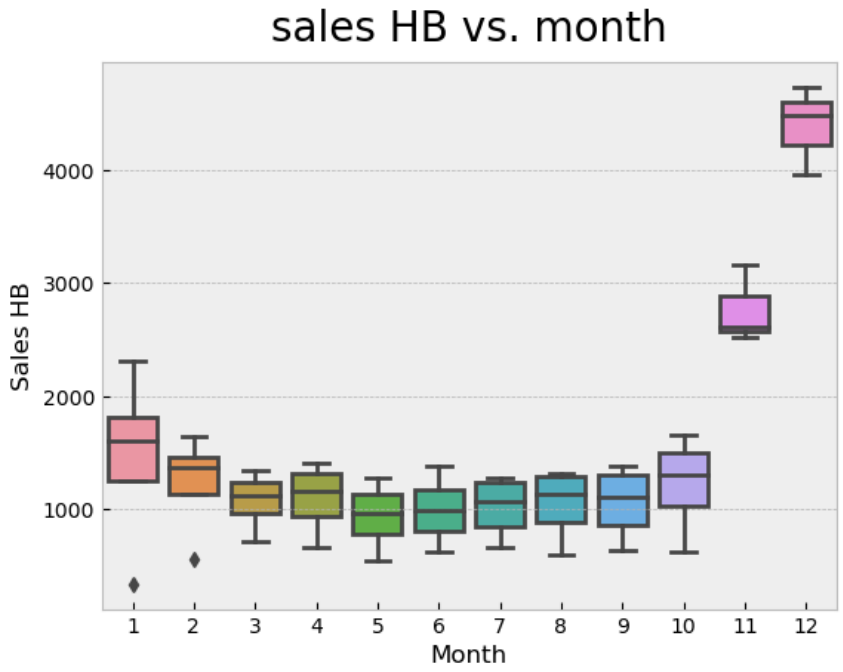
-correlación positiva entre la temperatura y las ventas de la marca RYE.

BOXPLOTS de ventas conjuntas y por separado según cada marca por mes del año.

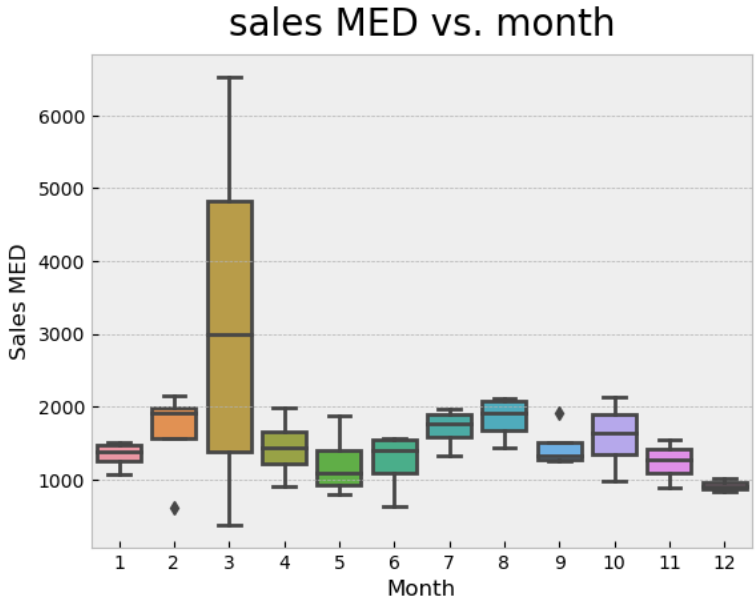



ESTA MARCA TIENE ESTACIONALIDAD DE VENTAS COMPATIBLE CON EL VERANO EN EL HEMISFERIO NORTE




ESTA MARCA TIENE UNA FUERTE ASIMETRIA DE VENTAS EN EL PERIODO DE LAS FIESTA DE FIN DE AÑO




AQUI SE PUEDE OBSERVAR CLARAMENTE LO PLANTEADO ANTERIORMENTE RESPECTO A LA ANOMALIA DE LAS VENTAS OBSERVABLE EN EL MES DE MARZO PARA ESTA MARCA RELACIONADA CON LA PURIFICACION DE AGUA PARA CONSUMO HUMANO

Como conclusión de esta evaluación de correlación, sumado al análisis de OLS de las diferentes variables, podemos afirmar que si bien las ventas totales de la combinación de las 3 líneas de productos no se pueden correlacionar con el clima o las horas de luz.

OLS PARA HORAS DIARIAS DE LUZ VS VENTAS RYE

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | Q("Daylight hs") | R-squared: | 0.536 |
| Model: | OLS | Adj. R-squared: | 0.526 |
| Method: | Least Squares | F-statistic: | 50.91 |
| Date: | Fri, 03 May 2024 | Prob (F-statistic): | 7.24e-09 |
| Time: | 19:03:01 | Log-Likelihood: | -83.299 |
| No. Observations: | 46 | AIC: | 170.6 |
| Df Residuals: | 44 | BIC: | 174.3 |
| Df Model: | 1 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Intercept | 8.5176 | 0.580 | 14.691 | 0.000 | 7.349 | 9.686 |
| Q("Sales RYE") | 0.0061 | 0.001 | 7.135 | 0.000 | 0.004 | 0.008 |

| | | | |
|---|---|---|---|
| Omnibus: | 3.802 | Durbin-Watson: | 0.437 |
| Prob(Omnibus): | 0.149 | Jarque-Bera (JB): | 1.823 |
| Skew: | 0.133 | Prob(JB): | 0.402 |
| Kurtosis: | 2.062 | Cond. No. | 1.78e+03 |

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.78e+03. This might indicate that there are strong multicollinearity or other numerical problems.

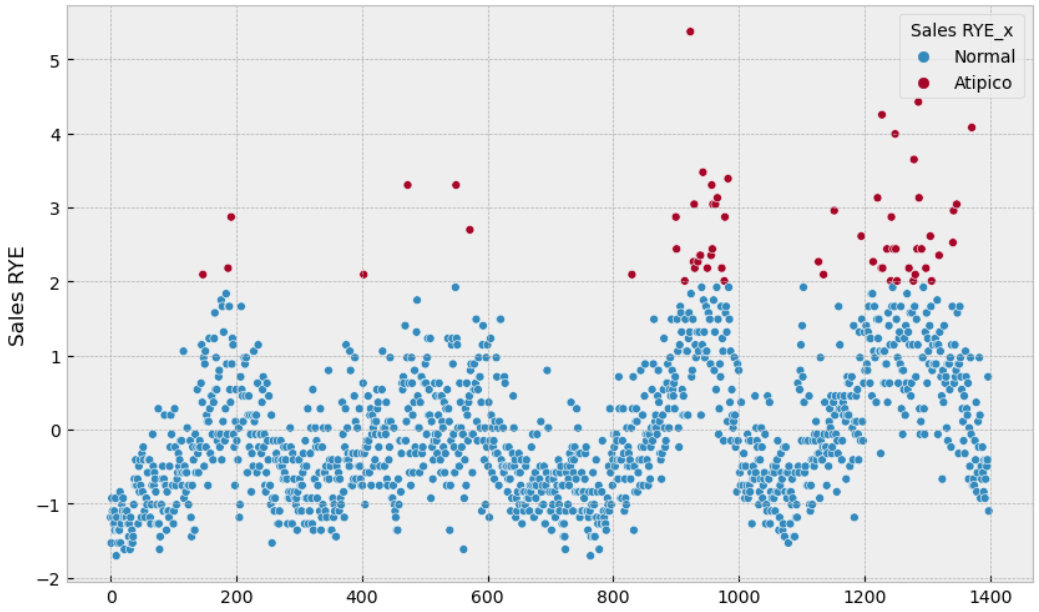
Cuando analizamos cada marca en particular, la línea RYE tiene una fuerte correlación positiva con la temperatura, siendo su F-estadístico de 47.6%  al igual que con las horas de sol, que tiene un F-estadístico del 50.9%

## 7. Data wrangling e implementación de modelo de ML

Habiendo realizado el análisis EDA y las correlaciones de las 3 líneas de producto, se decide enfocar los esfuerzos en una de las líneas para enfocar el trabajo y definir el modelo de ML mas adecuado.

Continuamos el análisis solo con la línea de productos RYE

Filtramos Valores atípicos




Luego de la estructuración de los datos se definen las variables independientes y la variable objetivo:

## Genero la variable target a predecir, Sales RYE y dejo la base x con las variables independientes a utilizar

In [200]:
```
X = datosML.drop(['Sales RYE', 'Year'],axis=1)
y = datosML['Sales RYE']
X.head()
```

Out[200]:

| | Month | weekday | Page Views RYE | Daylight | Max Temp |
|---|---|---|---|---|---|
| 0 | 1 | 2 | 27.0 | 8.98 | -0.2 |
| 1 | 1 | 3 | 20.0 | 8.98 | 6.4 |
| 2 | 1 | 4 | 23.0 | 9.00 | 6.8 |
| 3 | 1 | 5 | 28.0 | 9.02 | 2.5 |
| 4 | 1 | 6 | 32.0 | 9.05 | 1.3 |

In [201]:
```
y.head()
```

Out[201]:
```
0     7.0
1     3.0
2    10.0
3     8.0
4     7.0
Name: Sales RYE, dtype: float64
```

#Convierto las variables categoricas en columnas para hacerlas numericas

In [202]:
```
numerical = X.drop(["Month","weekday"], axis=1)
numerical.head()
```

Out[202]:

| | Page Views RYE | Daylight | Max Temp |
|---|---|---|---|
| 0 | 27.0 | 8.98 | -0.2 |
| 1 | 20.0 | 8.98 | 6.4 |
| 2 | 23.0 | 9.00 | 6.8 |

**Modelo de Machine LEARNING INICIAL RANDOM FOREST**

```
mean absolut error:  4.506079368485591
mean squared error:  36.60637197549331
Root mean squared error:  6.050319989512398
```

## 8. Crossvalidation

Se utilizaron 3 modelos de ML, Regresion Lineal, Random Forest y Red neuronal con los siguienes resultados en la crossvalidation

| Modelo: Regresión Lineal | Modelo: Random Forest | Modelo: Red Neuronal |
|---|---|---|
| MSE MARCA RYE: 78.51 | MSE MARCA RYE: 72.21 | MSE MARCA RYE: 77.30 |
| MSE MARCA RYE: 136.05 | MSE MARCA RYE: 60.62 | MSE MARCA RYE: 52.48 |
| MSE MARCA RYE: 34.83 | MSE MARCA RYE: 39.29 | MSE MARCA RYE: 72.00 |
| MSE MARCA RYE: 46.27 | MSE MARCA RYE: 59.05 | MSE MARCA RYE: 42.83 |
| MSE MARCA RYE: 54.90 | MSE MARCA RYE: 53.72 | MSE MARCA RYE: 58.10 |

De los 3 modelos analizados el que tiene resultados mas uniformes es el de Random Forest.

Conclusiones preliminares del Modelo de RANDOM FOREST

EL MODELO ASI IMPLEMENTADO No tiene muy buenos resultados para poder predecir las ventas con un error promedio aceptable, a continuación vamos a implementar una optimización para encontrar la mejor combinación de hiperparámetros.

Varianza en el rendimiento:

Se observa una variación en los valores del MSE en cada uno de los 5 splits de la validación cruzada. Esto sugiere que el rendimiento del modelo puede ser sensible a la partición específica de los datos en conjuntos de entrenamiento y prueba. Esta variación puede ser normal, pero se debe tener en cuenta que podría afectar la estabilidad del modelo.

Tendencia general:

Aunque hay variación, se puede observar una tendencia general en los valores del MSE. Por ejemplo, el tercer split tiene el MSE más bajo (39.29), lo que indica un mejor rendimiento del modelo en comparación con los otros splits. Por otro lado, el primer split tiene el MSE más alto (72.21), lo que indica un peor rendimiento en comparación con los otros splits.

## 9. Optimización

Se utiliza el método de GRIDSEARCH para buscar la optimización del Modelo de Random Forest,

Para la implementación del modelo de RANDOM FOREST los mejores hiperparámetros encontrados son los siguientes:

{'max_depth': 20, 'min_samples_leaf': 4, 'min_samples_split': 5, 'n_estimators': 50}

Obteniéndose los siguientes resultados

```
mean absolut error:  4.554172654015214
mean squared error:  37.607378379749086
Root mean squared error:  6.132485497720242
```

Esta optimización sigue dando resultados que no serían los óptimos para que el modelo sea exitoso. So confiabilidad es muy baja porque el desvío absoluto es alto comparado con las ventas promedio diarias en unidades.

## 10.Conclusiones y Recomendaciones

Como Conclusiones sugerimos analizar otras variables no contenidas en el presente trabajo para encontrar algún tipo de correlación más fuerte que permita predecir de manera más óptima las ventas de la marca RYE.

La base de datos de ventas es muy completa en cuanto a la cantidad de datos obtenidos, pero no es suficiente para poder generar un modelo que prediga de manera efectiva lo que necesitamos