Enero del 2023

Análisis en la demanda de bebidas

Humberto Flores Páez

Introducción

En el presente documento se expondrá un análisis a un dataset que contiene la información acerca de las ventas diarias de un tipo de bebida para un local comercial situado en el litoral central de Chile. El dataset contiene la información diaria acerca del número de ventas exactas que el local vende por día, además del precio que cobra al día por bebida vendida, así como los ingresos que reporta el local por concepto de bebidas vendidas. Se utilizarán herramientas estadísticas y del análisis de datos para determinar el comportamiento de la demanda a lo largo del tiempo, que permitirán esclarecer que día de la semana las personas consumen más bebidas, si afecta la temporada en las ventas, el efecto de la pandemia y además la creación de un modelo predictivo para determinar el número de ventas diarias para un mes en el futuro.

Análisis descriptivo de los Datos

El análisis de los datos del dataset, se llevará a cabo usando el entorno de Google Colaboratory y Spyder usando Python 3.9, con el fin de poder usar visualizaciones y programación con Python.

Primero que todo se analizarán las diferentes variables que contiene el dataset, las cuales se muestran en la Tabla 1:

Tabla 1: Descripción de las variables que contiene el DataSet.

Ds	Precio_ref	Unidades_total	Monto_total
Es la fecha del dato y contiene la información del año, mes y día.	Es el precio establecido para la venta de la bebida durante cada día.	Es la cantidad total de bebidas vendidas durante un día.	Es el Ingreso total registrado durante un día, por la venta de bebidas.

La información se encuentra dividida en dos datasets, el primero es el conjunto de entrenamiento o “x_train” que contiene la información de las variables antes descritas desde el 1 de enero del 2017 al 31 de enero del 2022 y contiene un total de 1854 filas con 4 columnas en total (ver Ilustración 1). El segunda dataset llamado “y_train”, contiene la data del mes de febrero del 2022, que abarca desde el 1 de febrero al 28 de febrero del 2022, contando con 28 filas en total.

	ds	precio_ref	unidades_total	monto_total
0	2017-01-01	990.0	158	142740
1	2017-01-02	990.0	152	137180
2	2017-01-03	990.0	125	115390
3	2017-01-04	990.0	107	98710
4	2017-01-05	990.0	103	96270
...	...	...	...	...
1849	2022-01-27	1100.0	63	68510
1850	2022-01-28	1100.0	85	93088
1851	2022-01-29	1100.0	95	103920
1852	2022-01-30	1100.0	62	67375
1853	2022-01-31	1100.0	68	74100

1854 rows × 4 columns

Ilustración 1: Visualización de la información del Dataset en Colaboratory

Realizando un análisis preliminar en Python con la librería Pandas, es posible observar que no existen valores nulos, es decir, que cada columna contiene la información completa de cada variable, pero algo llamativo es que dentro del x_train falta la información de 3 días dentro del horizonte temporal que abarca desde el 1 de enero del 2017 al 31 de enero del 2022, es decir existen 3 días en los cuales no existe ninguna información al respecto. Esto último se corrobora con la función .info() de pandas y la función date_range que es capaz de capturar todo el conjunto temporal de la variable (ver Ilustración 2).

```
[188] datos_fechas = pd.date_range(start = "2017-01-01", end = "2022-01-31", freq = "D")

[189] datos_fechas

      DatetimeIndex(['2017-01-01', '2017-01-02', '2017-01-03', '2017-01-04',
                     '2017-01-05', '2017-01-06', '2017-01-07', '2017-01-08',
                     '2017-01-09', '2017-01-10',
                     ...
                     '2022-01-22', '2022-01-23', '2022-01-24', '2022-01-25',
                     '2022-01-26', '2022-01-27', '2022-01-28', '2022-01-29',
                     '2022-01-30', '2022-01-31'],
                    dtype='datetime64[ns]', length=1857, freq='D')
```

Ilustración 2: Creación del data time en pandas para observar el número de días reales.

Los 3 días faltantes en el análisis, representan solo el 0.16% del horizonte temporal que comprende desde el 1 enero del 2017 al 31 de enero del 2022, por lo que, se analizaran cada una de las variables ignorando esta falta en la información, los días faltantes corresponden a 19-04-2017, 15-05-2021 y 16-05-2021. Cabe destacar que el siguiente análisis solo se llevara a cabo con el conjunto de entrenamiento.

La primera variable importante a analizar son las ventas diarias registradas, las cuales podemos observarlas en la siguiente gráfica:

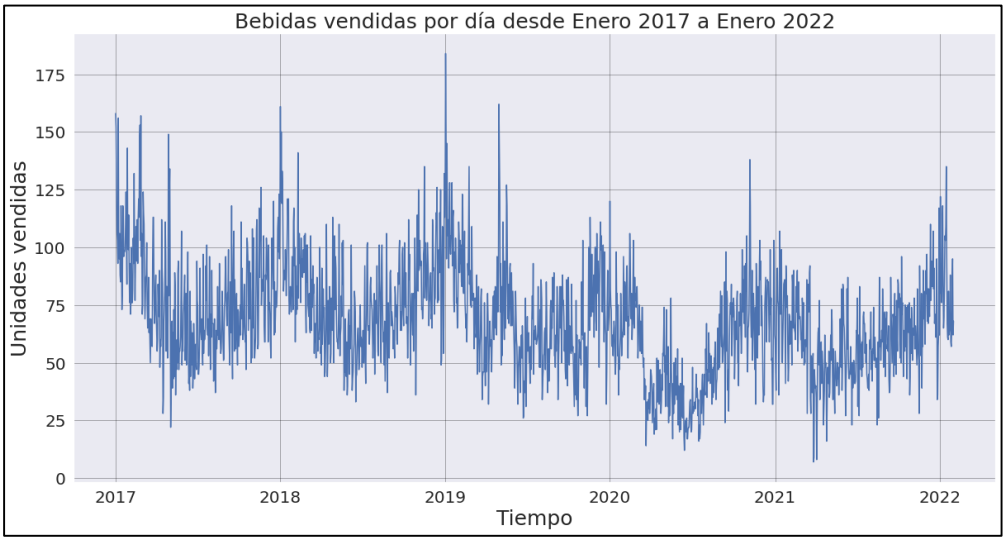

Gráfico 1: Demanda de Bebidas vendidas en el conjunto de entrenamiento.

A partir del Gráfico 1 de ventas diarias a lo largo del tiempo, es posible observar la estacionalidad de las ventas, ya que para el año 2017, 2018 y 2019 se observa claramente una repetición de la señal, es decir el comportamiento de la curva se repite durante estos 3 años, donde la curva alcanza su máxima altura en los de verano y su máximo decrecimiento en el invierno. Esto último es un comportamiento evidente, ya que, al tratarse de un mercado de venta de bebidas en el litoral central, las ventas aumentaran considerablemente durante el verano, debido al aumento de veraneantes, turistas y comercio durante esta época del año, pero durante invierno, se registra el comportamiento decreciente de la curva debido al invierno y con ello a la falta de turismo en la zona, lo que también tiene una relación con las temperaturas registradas durante esa época del año.

También se visualiza el efecto de la pandemia Covid-19 en las ventas, ya que las ventas durante los meses de verano del 2020 son más bajas que en los años anteriores, debido a la incertidumbre social y a la contracción de la economía debido al estallido social registrado desde el 18 de octubre del 2019, pero la situación más crítica es la época de invierno del 2020, ya que aquí se encuentra la zona más baja de toda la curva, por ejemplo, desde marzo hasta septiembre del 2020, el promedio de ventas diario es de 38 bebidas al día, donde el 50% de todas las ventas durante este periodo no superan las 34 bebidas diarias vendidas. También se nota la disminución drástica que sufre desde mediados de marzo, debido al decreto de las cuarentenas y del confinamiento.

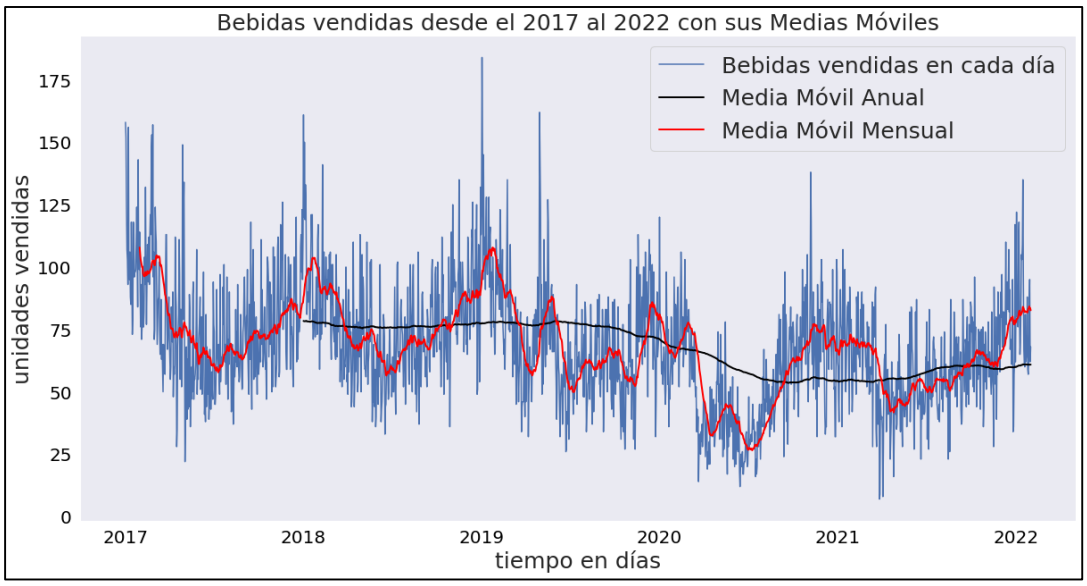

Gráfico 2: Demanda de bebidas vendidas del dataset de entrenamiento con sus medias móviles

Para poder analizar mejor las tendencias en las ventas diarias registradas durante los 5 años, se emplea la media móvil mensual y anual (ver Gráfico 2), donde la primera captura correctamente las variaciones y las tendencias en las ventas diarias, donde es posible mostrar la caída que sufren las ventas en periodos de invierno y las subidas en verano así como también una probable subida en los meses que siguen a enero del 2022, debido a que desde mediados del 2021 hasta el 31 de enero del 2022, la tendencia en las ventas de bebidas es al alza, esta alza en las ventas de bebidas se conectan con la siguiente variable de estudio que es el precio de la bebida, la cual se muestra en el Gráfico 3.

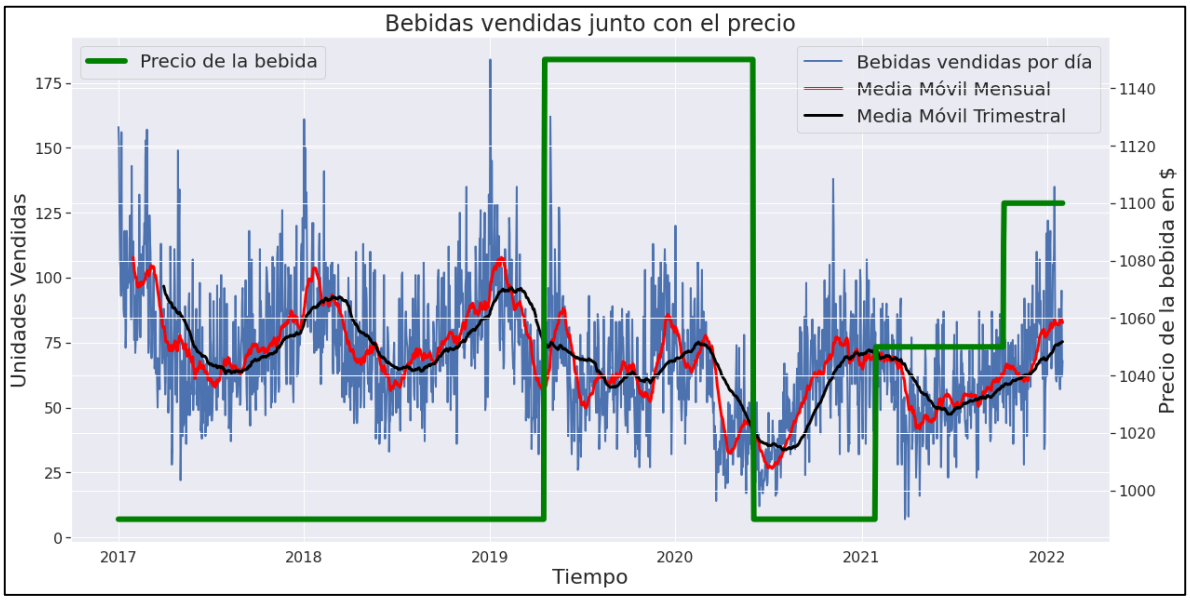

Gráfico 3: Relación entre el precio de la bebida y la demanda con las medias móviles de la demanda.

La línea verde del Gráfico 3 es el precio de la bebida en $, donde existe una relación entre el precio y la demanda en la bebida, ya que el día 18 de abril del 2019 se registra la primera alza de $160 en el precio de la bebida, esto produce que durante los meses que siguen a abril del 2019, específicamente durante los meses de mayo y primera semana de junio, se produce un aumento considerable en la media móvil mensual, la cual baja rápidamente debido a la temporada de invierno y desde noviembre del 2019 se experimenta un incremento en las ventas, pero este incremento es mermado por los contagios en china y el descubrimiento del virus Covid-19, el cual crea especulación en los mercados internacionales e incertidumbre económica, debido a esto, la demanda en las bebidas bajan hasta llegar a su punto más bajo en los meses de invierno del 2020, pero la bajada de $160 en el precio de la bebida, logra aumentar el consumo de bebidas, ya que por ley económica, al disminuir el precio, aumenta la demanda, pero al volver a aumentar el precio en $60 desde el 29 de enero del 2021 hace estancar la media móvil mensual y sufrir una bajada debido a la llegada del invierno. El precio vuelve a subir $50 el día 8 de octubre del 2021 llegando a $1100 por bebida.

La variable precio de la bebida es afectada dos veces a lo largo del horizonte temporal, la primera subida se registra en abril del 2019 y la segunda a fines de enero del 2021, donde se concluye que es más positivo subidas escalonadas y pequeñas, que subidas grandes y de golpe, ya que la primera subida es de golpe, donde la media móvil trimestral sufre bajones importantes en los meses de julio y agosto y no logra alcanzar el nivel de ventas en los meses de verano del 2020, esto se debe al alza en el precio y también a la especulación económica ante el surgimiento del virus covid-19, mientras que la segunda alza en los precios hecha desde comienzos de enero del 2021, hace disminuir y luego aumentar la media móvil trimestral, logrando el suavizamiento en esta curva, lo que demuestra que un aumento controlado y escalonado en los precios logra mejorar la situación económica en el mercado de las bebidas de una mejor forma que una fuerte subida.

Análisis anual de los datos:

En aras de visualizar correctamente la distribución y ubicar el mejor año en términos de ventas, se realiza un boxplot de las ventas diarias (Gráfico 4), segmentándolas por año:

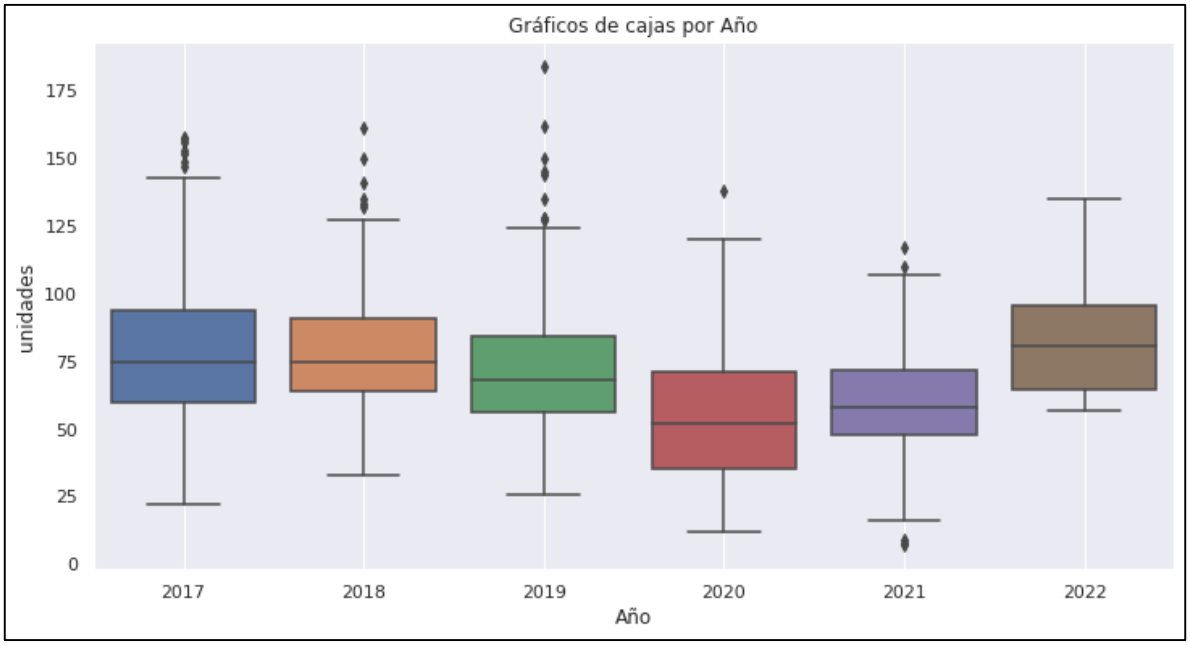

Gráfico 4: Boxplot de las ventas diarias de bebidas segmentadas por año.

Analizando la variable "unidades de bebida vendidas al día" a lo largo del tiempo, es posible ubicar el mejor y el peor año de ventas, donde el mejor año es claramente el 2017, ya que durante ese año el promedio de ventas es de 78 bebidas, donde el 75% de todo ese año se venden a lo máximo 94 bebidas por día (ver Tabla 2), y también es donde existe una distribución más espaciada de la variable en comparación a los demás años. También se aprecia el efecto de la pandemia durante el 2020, evidenciando una caída en las unidades vendidas, pero el aumento del precio y también la implementación de descuentos y promociones hacen que durante el 2021 haya un repunte en las unidades vendidas. Cabe destacar que el año 2022 solo está compuesto por el mes de enero y no es posible comparar o decir que este año es mejor que los demás, ya que aún no se tiene información de todos los demás meses que componen el 2022.

Tabla 2: Resumen de los Estadísticos para la variable "Unidades vendidas de bebidas" por año.

Año	N° de días	Promedio	Desviación Estandar	min	25%	50%	75%	max
2017	364	78.17	24.06	22	60	75	94	158
2018	365	77.42	21.09	33	64	75	91	161
2019	365	71.19	23.59	26	56	68	84	184
2020	366	54.46	22.38	12	35.25	52	71	138
2021	363	59.83	17.89	7	48	58	72	117

También es importante analizar los ingresos totales registrados por el local, el cual se pueden observar en el Gráfico 5:

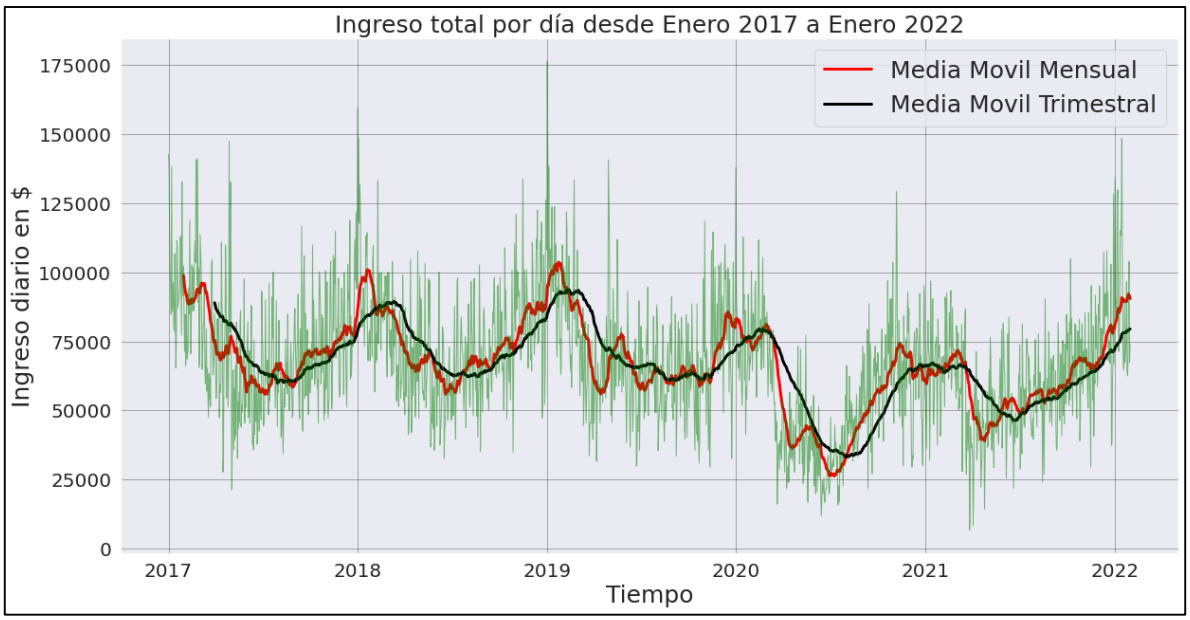

Gráfico 5: Ingreso total real que obtiene el local de venta de bebidas con sus medias móviles durante el conjunto de entrenamiento.

Es importante destacar que el comportamiento del ingreso total es prácticamente idéntico en naturaleza, comportamiento y actividad que el número de bebidas vendidas al día, esto es debido a que el ingreso depende directamente de las bebidas vendidas y el precio es una constante que, si bien sufre variaciones, estas son pequeñas y hacen que la media móvil mensual y trimestral se comporten de igual forma que para las unidades vendidas de bebidas.

Algo importante con respeto al ingreso total es que cuando se multiplica el valor del precio por las unidades vendidas, que es el ingreso estimado, este ingreso estimado es diferente al del ingreso total y generalmente es menor, por lo que se asume que existen muchos días en los cuales el local no cobra a las personas el valor real del precio, sino, que muchas veces aplica descuentos y promociones que hacen que el ingreso real sea menor que el que debiese ser.

Además de las 4 variables descritas anteriormente, es importante mencionar que existen otras variables que también son fundamentales para el estudio, y tienen conexión con la época del año, ya que cada valor en las unidades de bebidas vendidas tiene directa relación con el mes y el día, es decir con la época en la cual están ubicados temporalmente los datos, debido a esto, se agregan 3 nuevas columnas que representaran el año, el mes y el día de la semana de cada dato como lo muestra la Ilustración 3.

```
[231] train["Año"] = train.index.year
      train["Mes"] = train.index.month_name()
      train["Dia"] = train.index.day_name()
      #train["Dia"] = train.index.

[232] print(train)
```

ds	precio	unidades	ingreso_total	Año	Mes	Dia
2017-01-01	990.0	158	142740	2017	January	Sunday
2017-01-02	990.0	152	137180	2017	January	Monday
2017-01-03	990.0	125	115390	2017	January	Tuesday
2017-01-04	990.0	107	98710	2017	January	Wednesday
2017-01-05	990.0	103	96270	2017	January	Thursday
...	...	...	...	...	...	...
2022-01-27	1100.0	63	68510	2022	January	Thursday
2022-01-28	1100.0	85	93088	2022	January	Friday
2022-01-29	1100.0	95	103920	2022	January	Saturday
2022-01-30	1100.0	62	67375	2022	January	Sunday
2022-01-31	1100.0	68	74100	2022	January	Monday

[1854 rows x 6 columns]

Ilustración 3: Adición de tres nuevas variables temporales en el dataset.

Además de este análisis, se decide analizar la distribución en las unidades vendidas por año, por lo que se generan múltiples histogramas según el año para la variable “unidades vendidas de bebidas”, con el fin de identificar si existe o no, una normalidad en la variable, para ello se usara la biblioteca scipy usando el módulo stats para investigar esto en Python.

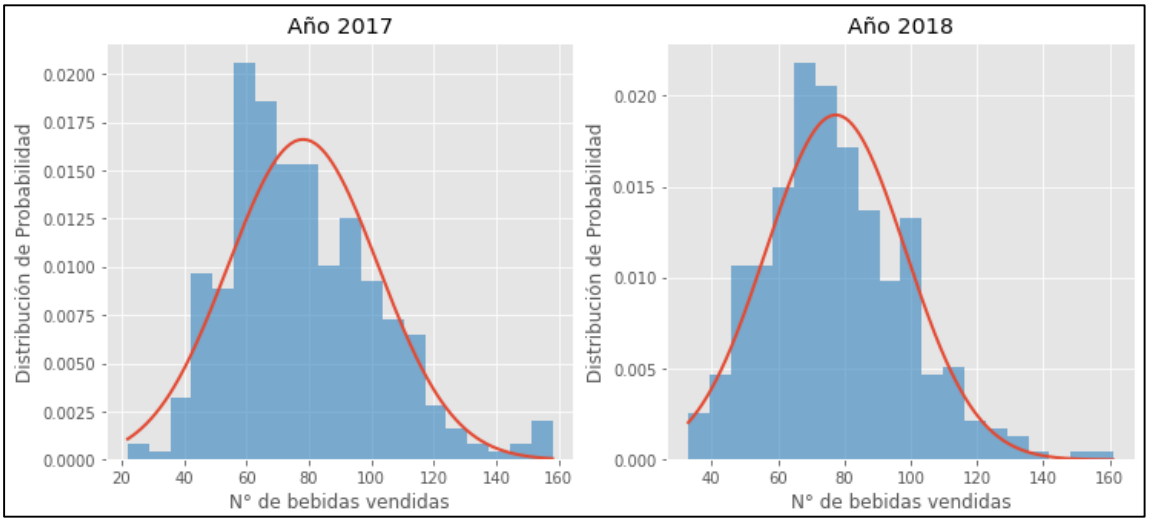

Gráfico 6: Densidad de probabilidad e histograma para los años 2017 y 2018.

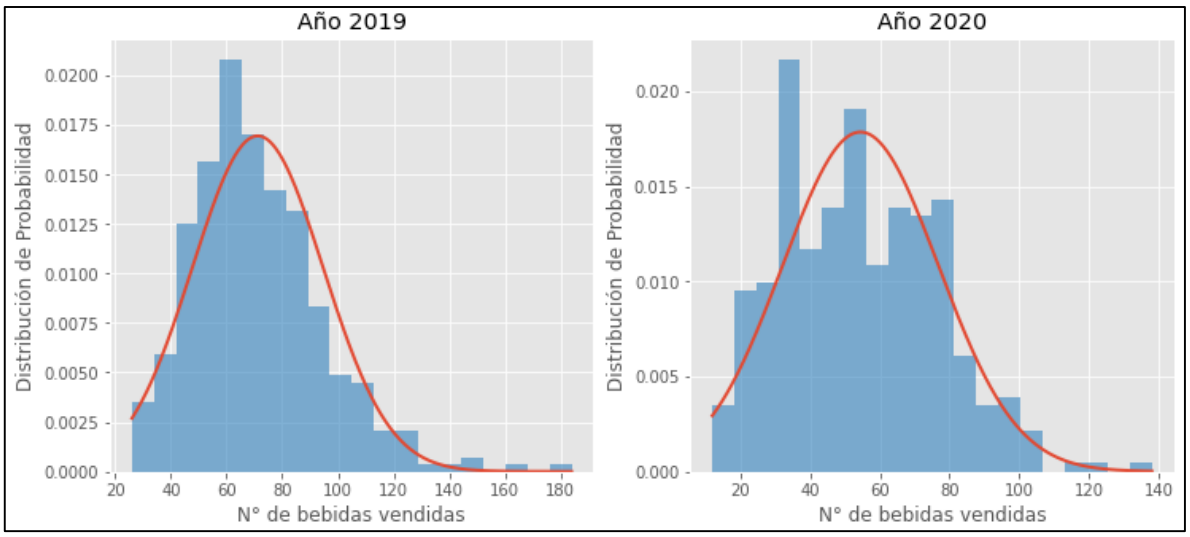

Gráfico 7: Densidad de probabilidad e histograma para los años 2019 y 2020.

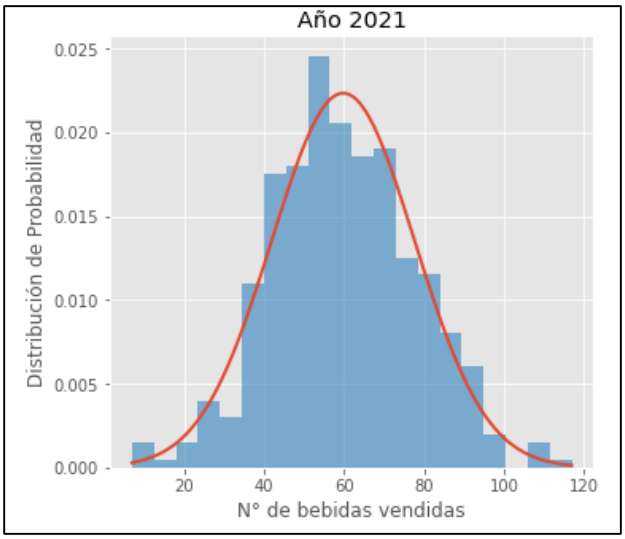

Gráfico 8: Densidad de probabilidad e histograma para el año 2021

A partir de los histogramas generados para cada año en base a la variable “unidades de bebidas vendidas” (Gráfico 6, Gráfico 7 y Gráfico 8), se genera en cada uno de ellos, la distribución normal (curva roja) que es aquella curva que representa la distribución normal usando una media y desviación estándar igual a la que tiene el conjunto de datos por año, por lo que si los rectángulos celestes siguen la forma de la curva roja, se infiere que la distribución de los datos es normal, pero en todos los años se observan diferencias notables entre la forma de los rectángulos y la curva normal, solo para el año 2021, se observa cierta semejanza. Para analizar más en profundidad este tema, se aplica una prueba de contraste de hipótesis a partir de la prueba estadística de Shapiro Wilks, donde usa un estadístico y un p-value, si el p-value es mayor a 0.05 existirán pruebas suficientes de que la variable es normal.

```
Año 2017
ShapiroResult(statistic=0.96815425157547, pvalue=3.838851796444942e-07)
------------------------------------------------------------
Año 2018
ShapiroResult(statistic=0.9795181155204773, pvalue=4.728844214696437e-05)
------------------------------------------------------------
Año 2019
ShapiroResult(statistic=0.9590422511100769, pvalue=1.4611113208218285e-08)
------------------------------------------------------------
Año 2020
ShapiroResult(statistic=0.9767929315567017, pvalue=1.2974011042388156e-05)
------------------------------------------------------------
Año 2021
ShapiroResult(statistic=0.9953119158744812, pvalue=0.3450271189212799)
```

Ilustración 4: Resultados del test Shapiro Wilks en la variable "Unidades vendidas de bebidas" por año para analizar la normalidad de la variable.

El test de Shapiro Wilks (ver Ilustración 4) indica que para los años 2017, 2018, 2019 y 2020, la distribución de la variable “unidades de bebidas vendidas” no sigue una distribución normal ya que el p-value es menor que el nivel de significancia α (0.05) y su estadístico es cercano a 1, además solo para el año 2021 existen pruebas de que en ese año se observa la normalidad en la variable, ya que el p-value es mayor a 0.05.

Efecto del día de la semana en la venta de bebidas

Para analizar el efecto que tiene el día de la semana en las ventas, se construye en Python unos diagramas de cajas, que mostraran la distribución y los cuartiles según cada día de la semana.

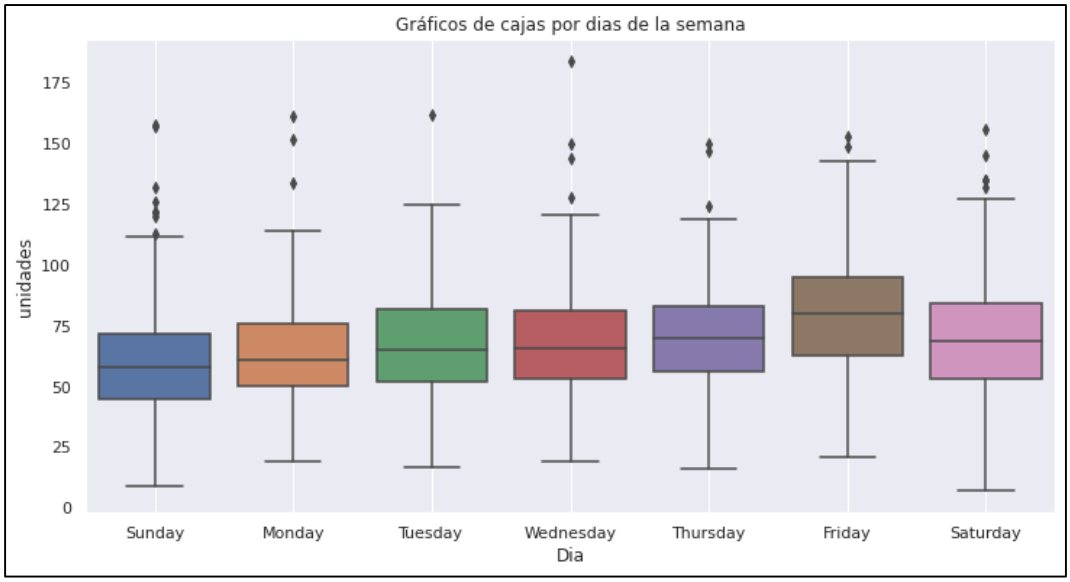

Gráfico 9: Boxplot de las ventas diarias de bebidas segmentadas por el día de la semana.

Tabla 3: Resumen de los Estadísticos para la variable "Unidades vendidas de bebidas" por día de la semana.

Día	N° de días	Promedio	Desviación Estándar	min	25%	50%	75%	máx.
Lunes	266	63.29	21.16	19	50	61	75.75	161
Martes	265	67.26	22.78	17	52	65	82	162
Miércoles	264	68.57	23.85	19	53	66	81.25	184
Jueves	265	70.32	21.36	16	56	70	83	150
Viernes	265	79.37	24.35	21	63	80	95	153
Sabado	264	69.75	25.20	7	53	69	84	156
Domingo	265	60.68	24.04	9	45	58	72	158

A partir del Gráfico 9, es posible evidenciar que el mejor día para las ventas es el día viernes, ya que en este día podemos vender desde 21 hasta casi 153 bebidas diarias, además el valor de la mediana en el día viernes es la más alta, llegando a 80 bebidas y su promedio es de 79 bebidas, por lo que este día también posee una excelente simetría, ya que los datos se encuentran distribuidos de forma normal con respecto al promedio. Además, se puede observar que los días que más se vende bebidas son el viernes y el sábado, debido a que las personas salen de sus trabajos el viernes y por la vida bohemia que tiene el litoral central, las personas desean beber más, aquellos días en los que saben que al siguiente no deberán trabajar, lo que tiene mucho sentido.

El día domingo y lunes son los días en que menos bebidas se vende, pero el domingo es el día de menor venta, debido a que solo se puede vender entre 9 a 158 bebidas, logrando vender 60 bebidas en promedio (ver Tabla 3), donde la desviación estándar en el número de ventas es mayor en comparación a la del día lunes, donde el lunes registra un promedio de 63 bebidas al día.

Es importante destacar que los días viernes y sábados poseen una mayor dispersión o variabilidad de las unidades vendidas de bebidas, en comparación a los días domingos, lunes, martes, miércoles y jueves, esto puede atribuirse al efecto que posee los días en que las personas están recién pagadas, es decir, las personas son más predispuestas a gastar y comprar un mayor número de bebidas cuando tienen más dinero es sus bolsillos, pero esto solo ocurre 1 vez al mes, y además los viernes y sábados se dan solo 1 vez a la semana, por lo que existirá 1 fin de semana al mes, donde las personas compraran más bebidas y los demás fines de semana, las personas decidirán no comprar tantas bebidas como aquel fin de semana donde están recién pagados, pero en los días de semana (domingo, lunes, martes, miércoles y jueves), las personas consumirán de forma normal, ya que independiente si están pagados o no, compraran lo justo, ya que no pueden celebrar ni tampoco disponen de tanto tiempo debido al trabajo.

Efecto de los meses en la demanda de bebidas

Para analizar el efecto que tiene la época del año en las ventas, se decide construir un diagrama de cajas según los meses.

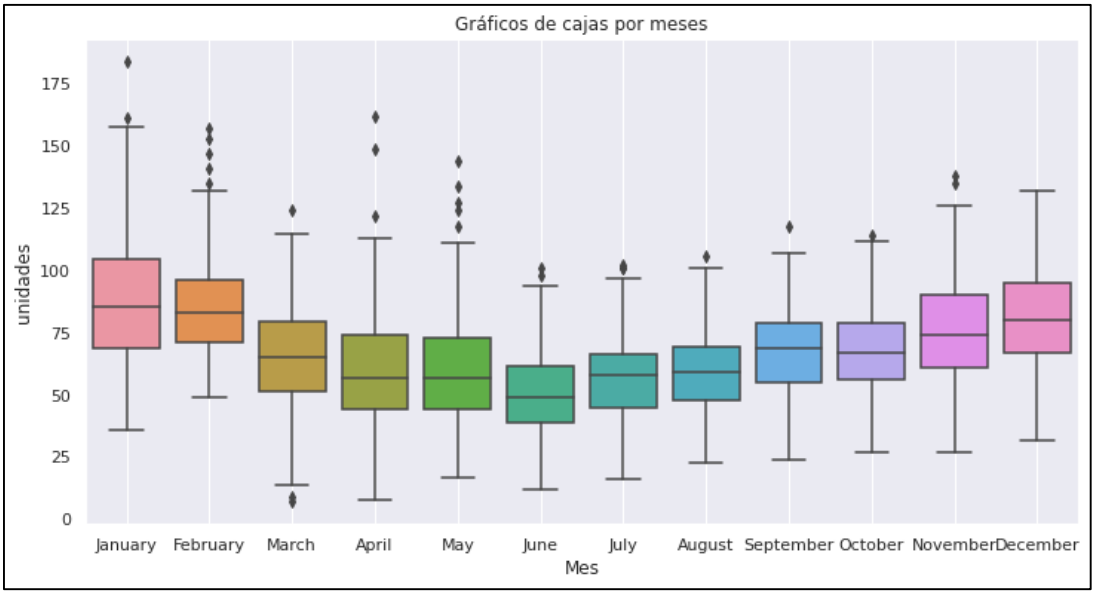

Gráfico 10: Boxplot de las ventas diarias de bebidas segmentadas por meses.

Según el Gráfico 10, se puede observar los mejores meses para la venta de bebidas son enero y febrero, ya que en el mes de enero el 50% de las ventas totales registradas, se venden entre 68 a 104 bebidas al día, con un promedio de 88 bebidas. En febrero, el 50% de todas las ventas comprende desde 71 a 96 bebidas al día, con un promedio de 86 bebidas diarias. Se puede observar la estacionalidad que poseen las ventas de bebidas a lo largo de los meses, ya que la mediana y la cantidad de unidades vendidas ira disminuyendo paulatinamente desde el mes de marzo hasta junio que es el mes donde menos se vende bebidas, y a partir de este mes de junio, las unidades vendidas vuelven a subir de forma constante y creciente hasta llegar a la época de verano, donde tienen su máximo valor.

Es importante mencionar que existen valores que se escapan a los gráficos de cajas o bigotes y el número de datos escapados aumenta notoriamente en el mes de Febrero y Mayo, estos grandes valores de unidades vendidas de bebidas, pueden explicarse debido a las promociones que existen durante estos meses, ya que febrero es el último mes de las vacaciones y es la oportunidad para capturar un mayor número de ventas mensuales, y las promociones son una medida efectiva para lograr un mayor número de ventas puntuales, también en el mes de Mayo y Abril se registran ventas que se encuentran sobre el límite superior del diagrama de caja, existen días en estos meses en los cuales se vende por sobre el máximo del diagrama de cajas, esto es explicable debido a que en Abril existe la semana santa, donde las personas viajan al litoral central y en estos días, existe una gran demanda de bebidas, además de Mayo, donde existen 2 feriados, el día del trabajador y el combate naval de Iquique, días que pueden posibilidad un fin de semana largo y con ello el aumento de ventas por parte de las personas.

Efectos de la pandemia Covid-19 en la demanda de bebidas

Para analizar el efecto de la pandemia en el número de ventas, se grafican el número de ventas a lo largo de tiempo, durante enero del 2017 a enero del 2022, así como también la media móvil mensual (curva roja) y la media móvil anual (curva negra).

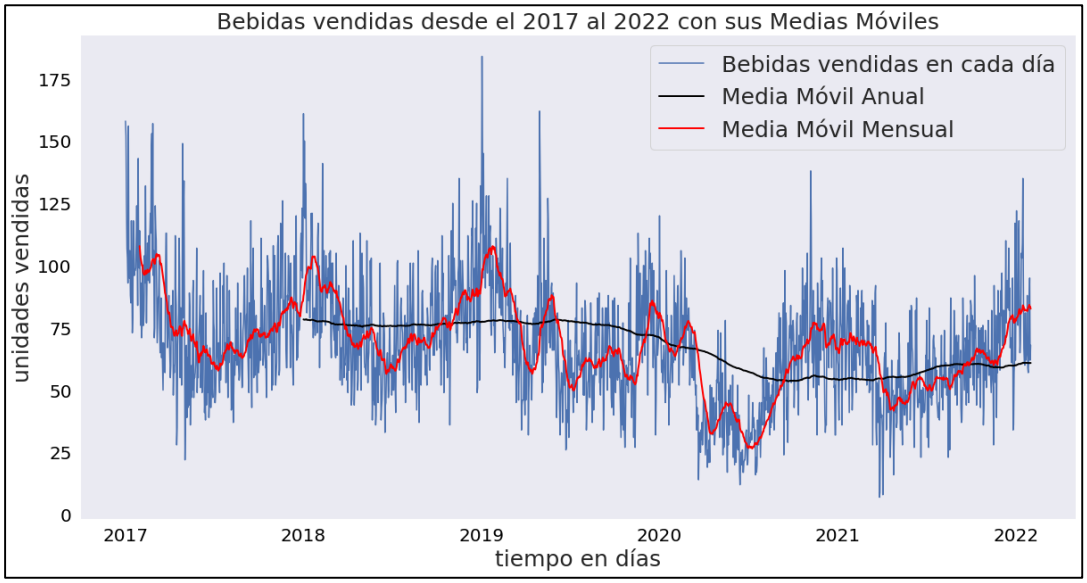

Gráfico 11: Numero de Bebidas vendidas para estudiar el comportamiento de la pandemia.

A partir del Gráfico 11, se observa que la media móvil mensual presenta estacionalidad, es decir se repite de forma cíclica subidas en meses de verano y bajas en los meses de invierno, pero en el año 2020, la media móvil mensual baja más de lo común, experimentando una fuerte caída en las unidades vendidas, esto se debe a la situación de cuarentenas establecidas en el país, que obligan a las personas a no viajar y no salir sus casa, en época de invierno del 2020, ocurre el pick de la pandemia, donde la situación de camas UCI es crítica, y el número de ventas en bebidas es bastante bajo en invierno. Esta situación cambia con la llegada del verano, donde para los meses de enero y febrero del 2021, vuelve a crecer la venta de bebidas, pero de forma paulatina y no alcanza los niveles de ventas registrados en las épocas estivales del 2017 o 2018. Esto último es posible evidenciarlo con los estadísticos para la variable unidades vendidas, segmentados por año, donde count es el número de días de cada año, mean es el promedio en el número de ventas y std es la desviación estándar:

Tabla 4: Resumen de los Estadísticos para la variable "Unidades vendidas de bebidas" segmentado por años.

Año	N° de días	Promedio	Desviación Estandar	min	25%	50%	75%	max
2017	364	78.17	24.06	22	60	75	94	158
2018	365	77.42	21.09	33	64	75	91	161
2019	365	71.19	23.59	26	56	68	84	184
2020	366	54.46	22.38	12	35.25	52	71	138
2021	363	59.83	17.89	7	48	58	72	117
2022	31	83.00	21.10	57	64.5	81	95.5	135

También es posible observar que el promedio de bebidas vendidas durante el año 2020 es el más bajo de los 5 años analizados (54 bebidas por día, ver Tabla 4), y que la pandemia genera una contracción en el número de ventas, lo que solo el pasar del tiempo, la disminución de las restricciones y el fin del confinamiento hace crecer este promedio, pero de forma paulatina y lenta.

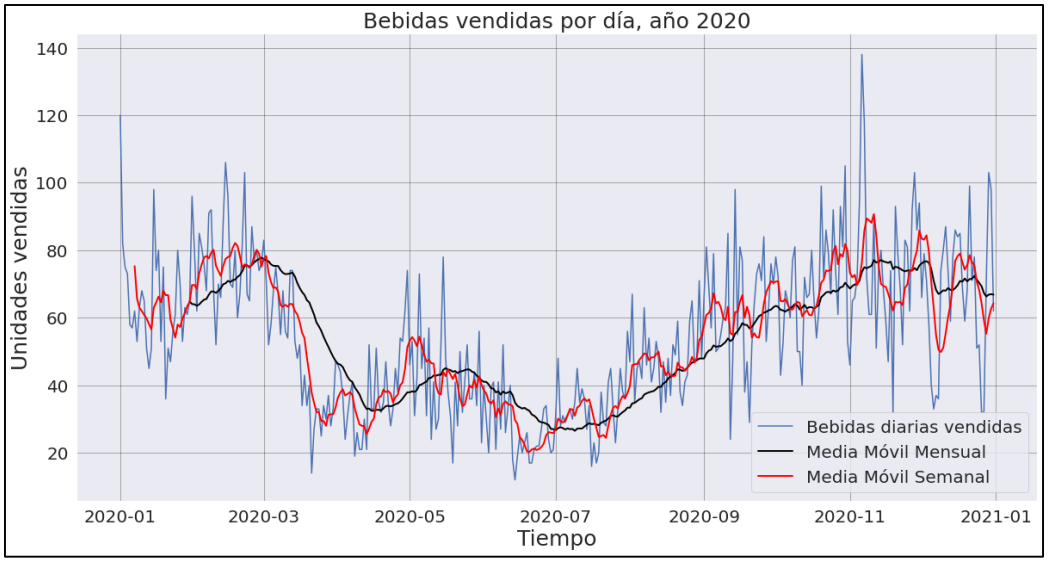

Gráfico 12: Bebidas vendidas por día durante el año 2020 junto con sus medias móviles mensual y trimestral.

Por ejemplo, observando detenidamente solo el año 2020 (ver Gráfico 12), es posible ver que en gran parte de los meses de invierno se venden menos de 40 bebidas al día, lo cual es bastante bajo para un año corriente, donde solo después de agosto del 2020, se vuelve a recuperar las ventas, pero los niveles de ventas en los meses de verano están lejos de ser los habituales.

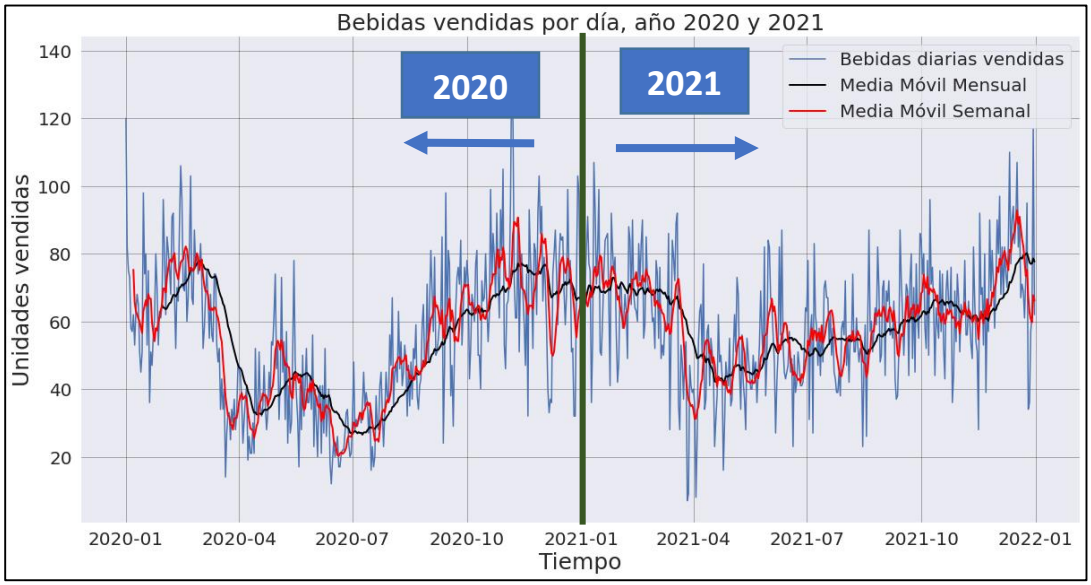

Gráfico 13: Bebidas vendidas durante los años 2020 y 2021 a modo de comparación.

Por ejemplo, analizando la situación de la pandemia, se observa que los meses de verano del 2021 (enero y febrero, ver Gráfico 13) permiten recuperar las ventas y que crezca la media móvil mensual,

pero claramente se aprecia el efecto de la pandemia en la temporada de invierno, donde la pandemia golpea con gran fuerza las ventas en este periodo.

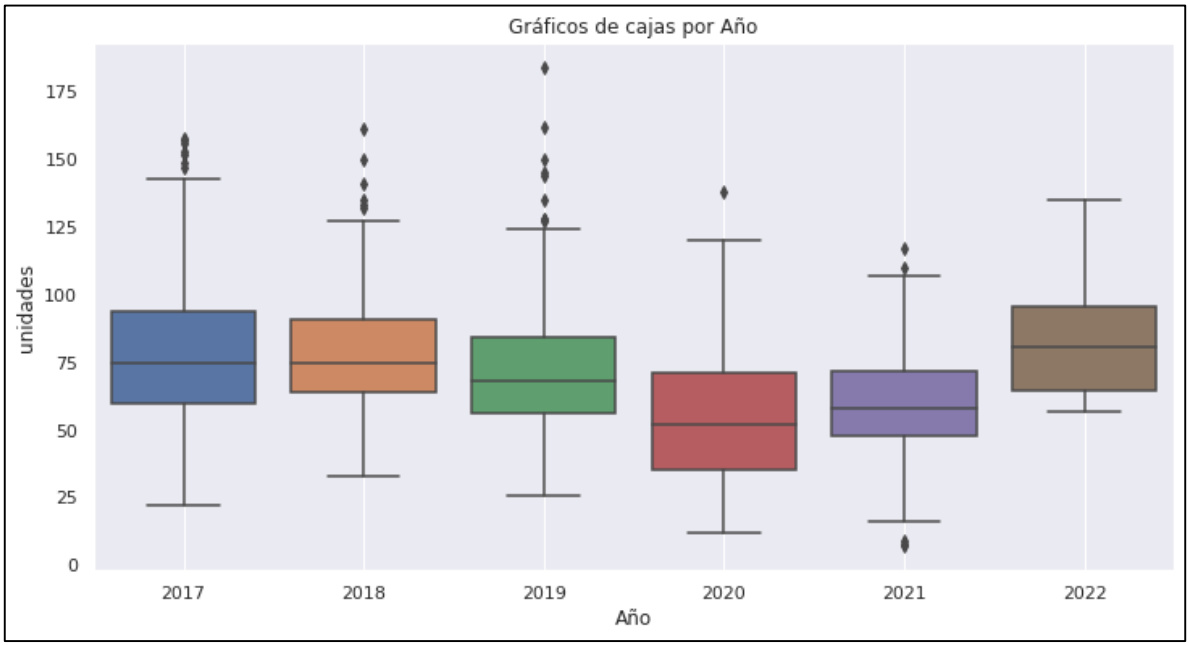

Gráfico 14: Boxplot de las ventas diarias de bebidas segmentadas por años.

Otra forma de visualizar el efecto de la pandemia en las ventas es a través de los boxplot o diagramas de cajas por año, en el Gráfico 14 se observa claramente que el año más crítico es el año 2020, debido al efecto de la pandemia, donde el promedio de ventas diarias es de solo 54 bebidas y donde el 75% de todo el 2020 poseen ventas diarias menores o iguales a 71 bebidas.

De modo adicional la pandemia también genera un efecto notable en la propia distribución de las ventas de bebidas, para ello se construyen múltiples histogramas que reflejan la distribución de probabilidad según el año y como esa distribución se ve modificada para el año 2020, año más golpeado por la pandemia.

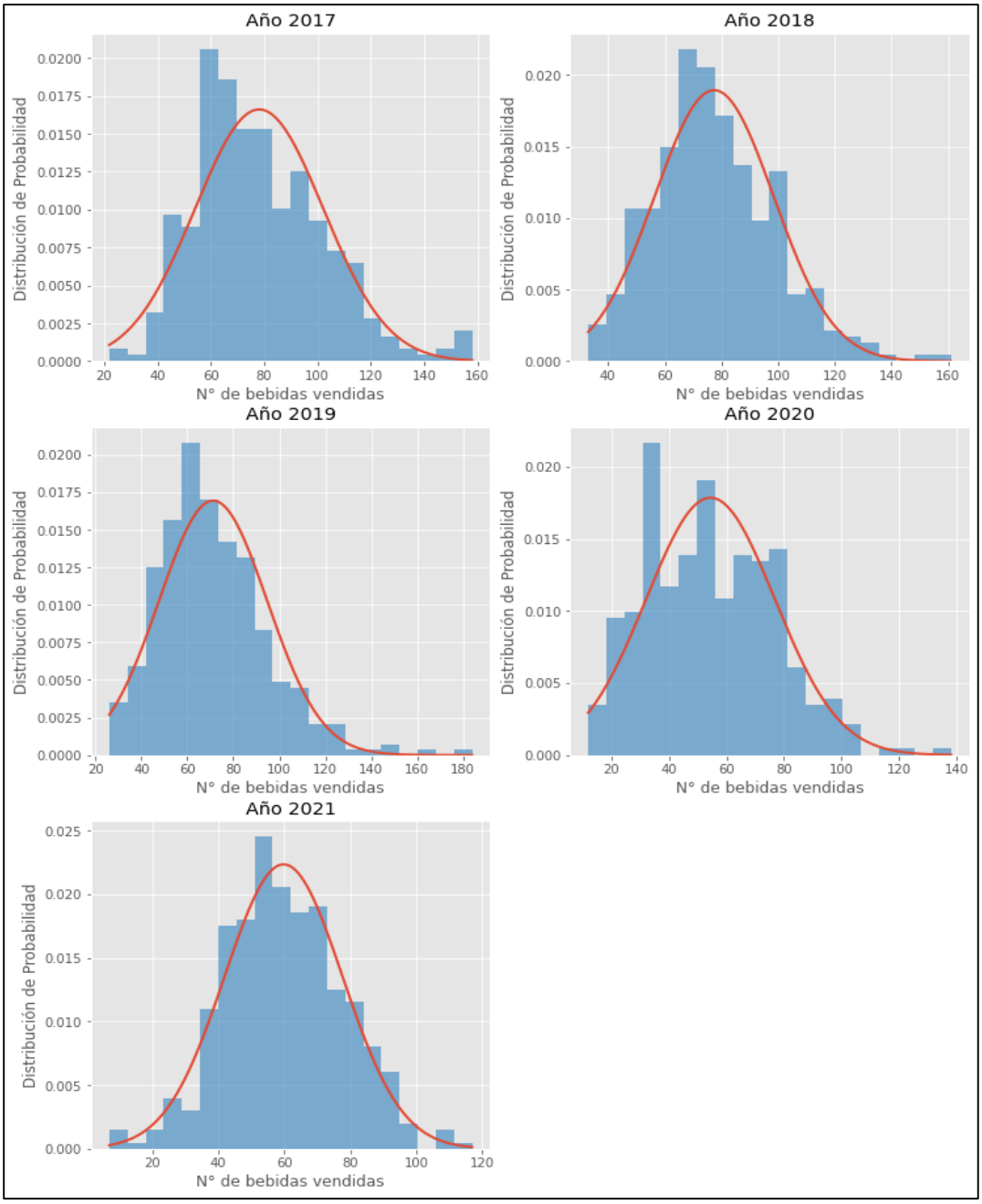

Gráfico 15: Histogramas y distribución de la variable "Unidades vendidas de bebidas" durante todo el horizonte temporal.

A partir del histograma realizado en base al número de bebidas vendidas en el 2020, es posible evidenciar que la distribución en la venta de bebidas durante ese año, se aleja en mayor medida de la distribución normal en comparación a los otros años, es decir, que la pandemia también afecta al comportamiento de compra que posee la demanda sobre la bebida, ya es igual de probable que en un día se vendan 80 bebidas a que se vendan 45 bebidas, es más, vender entre 32 a 37 bebidas al día tiene el doble de probabilidad que vender 60 o 40 bebidas al día, la probabilidad no se distribuye de forma normal, lo que dificulta mucho para predecir el comportamiento en la demanda de bebidas.

Es difícil incluir en el pronóstico o modelo predictivo de bebidas una situación como una pandemia, ya que la pandemia es un fenómeno impredecible y que puede surgir en cualquier momento y en cualquier parte del mundo con una rápida propagación a lo largo del globo, pero lo que sí se puede hacer es analizar ciertos indicadores que demuestren que una infección en un determinado lugar del mundo, puede llegar a escalar y transformarse en una pandemia, es decir, se puede analizar la tasa de contagios por día y la propagación de una nueva enfermedad a penas esta inicia, para determinar de forma anticipada si esta enfermedad puede llegar transformarse en una pandemia global, para poder tomar las acciones necesarias en los mercados económicos, una pandemia que si es capaz de afectar la economía de todo el mundo y obviamente el número de ventas en un local de bebidas en el litoral central.

Efecto del Precio en la demanda de bebidas

Para analizar el efecto que tiene el precio, se decide graficar de forma visual el precio de las bebidas (línea verde), así como también la media móvil del ingreso Real que tiene el local (línea café) y la media móvil del ingreso estimado (línea negra), este ingreso estimado es la multiplicación directa del precio de cada bebida por el número de bebidas, donde se observan diferencias entre cada ingreso, mostrados por la curva de color azul.

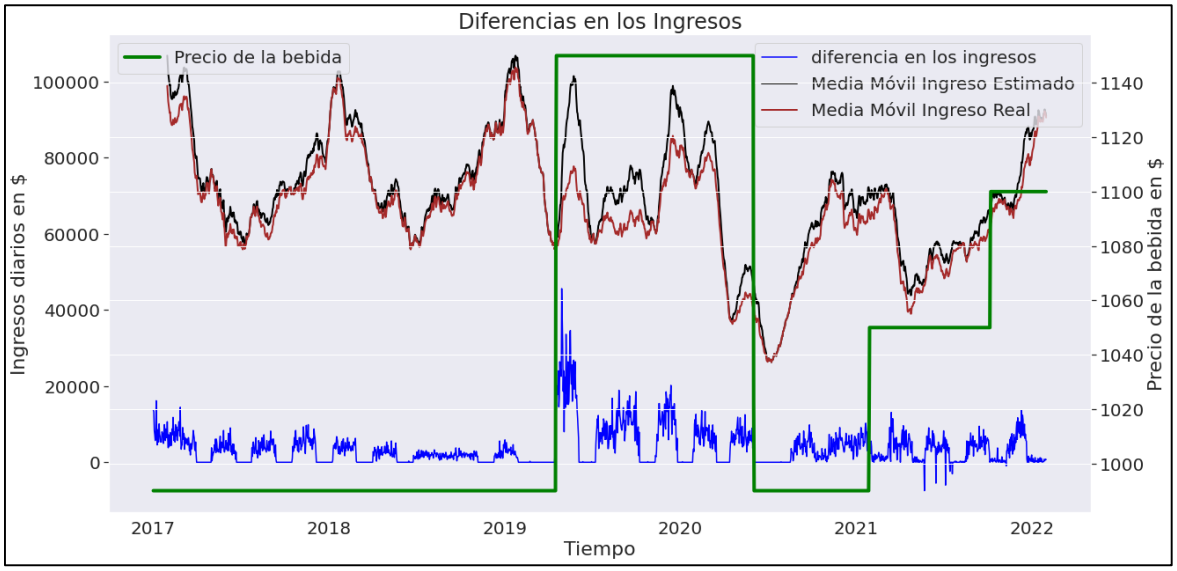

Gráfico 16: Diferencias en los Ingresos, junto con el ingreso total, ingreso estimado y el precio de la bebida.

En un primer acercamiento, notamos que en el Gráfico 16, existe el ingreso total que es lo que realmente gana al día el local de ventas de bebidas, pero en el dataset tenemos la información del precio diario de cada bebida y del número de bebidas vendidas, si multiplicamos el precio (P) de cada bebida por el número de bebidas vendidas (B) obtendríamos el ingreso Estimado que es lo que debería ganar el local y está dado por la siguiente ecuación:

$$Ingreso\ Estimado\ =\ Precio\ \cdot\ Bebidas\ Vendidas.$$

Pero lo que sucede es que, al observar los datos, el Ingreso Total es diferente del Ingreso Estimado diario, esto es observable en la gráfica adjunta, a través de la media móvil mensual entre en Ingreso Total y el Ingreso Estimado, esta diferencia entre los dos ingresos, se debe a la implementación por

parte de la empresa de un sistema de descuentos en el precio que cobra el local por cada bebida vendida, lo que es posible observar en el Gráfico 17, la cual muestra el precio de referencia que impone el local, y el precio real que cobra el local por bebida vendida, donde se aprecia que durante gran parte del horizonte temporal el precio está siendo afectado por descuentos y promociones que alteran la demanda y los ingresos. Los descuentos aplicados son más fuertes durante la primera subida en el precio real de la bebida, esto puede deberse a que, al poco tiempo de aplicar el alza en el precio del producto, la demanda bajo considerablemente y los empleadores se vieron en la necesidad de imponer un descuento más atractivo que hiciese aumentar las ventas diarias, esto refleja lo negativo que puede resultar una subida del 16% en el precio de la bebida.

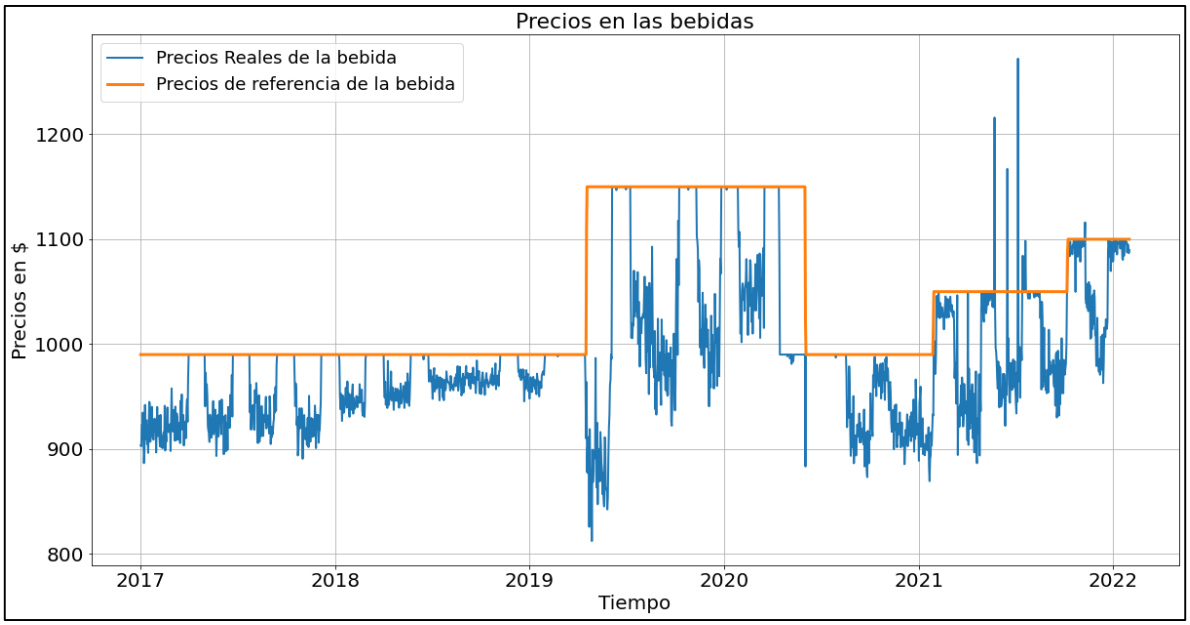

Gráfico 17: Precios Reales y de referencia en el precio de la bebida.

También es posible notar que a partir del 18 de abril del 2019, hasta aproximadamente el 2 de junio del 2020, el precio de las bebidas sube de 990 a 1150, sube 160, lo que produce una mayor diferencia entre el Ingreso Total y el Ingreso Estimado, esto puede deberse, ya que al subir el precio, la demanda disminuye, producto de esto, las personas NO están dispuestas a pagar 160 pesos adicionales por bebida y a que además es época de invierno, donde las personas no consumen tanta bebida, por lo cual, el local se ve obligado a realizar unos descuentos más atractivos para las personas, esto se marca mucho desde la quincena de abril del 2019 a la primera semana de Junio, esto puede deberse a que la subida de precios se realizó durante meses de invierno, durante los cuales la demanda es baja, y por ende al subir el precio, la demanda es más baja aun, y para mantener los Ingresos Totales dentro de los márgenes que espera la compañía, es natural que la empresa impusiera mayores descuentos para recuperar la demanda. Esto último es notable ya que desde el 28 de enero del 2021 también se registra una subida en el precio del producto, pero al ser esta subida de tan solo 60 pesos por bebida, los descuentos que imponen son más bajos que los descuentos impuestos durante la subida de $160, lo cual es positivo para la empresa, esto se debe a que la subida en el precio de la bebida es bajo y paulatino, donde las personas SI están dispuestas a pagar 60 pesos adicionales por bebida, y además el nivel de descuentos es moderado comparado con la situación donde el precio de subió de golpe.

Adicionalmente es posible detectar 4 días durante los meses de invierno del 2021, donde el Ingreso Total es mayor que el Ingreso Estimado, esto puede deberse a alzas puntuales en el precio de la bebida en esos 4 días. las cuales se registran en promedio durante medidos de cada mes y podrían explicarse como una subida puntual en el precio de la bebida, debido a eventos en el litoral central o fiestas realizadas en los alrededores.

Modelo Predictivo para la demanda de bebidas

Se llevará a cabo la creación de un modelo predictivo que sea capaz de predecir el número de bebidas vendidas durante el mes de Febrero del 2022, para ello se creara una red neuronal recurrente, de la librería Keras y Tensorflow, se usa en específico una red neuronal de este tipo debido a que los datos de las ventas son diarias y estas, se encuentran especiadas temporalmente, y el tipo de red neuronal recurrente funciona muy bien para predecir stocks de precios, consumo y volúmenes vendidos a lo largo del tiempo, ya que la variable más fundamental del problema es el tiempo.

Para hacer funcionar esta red, la red necesitara un tensor que es una matriz multidimensional que se genera a partir de la librería Numpy, donde la clave es el tensor temporal que la red necesita, ya que este tensor es una matriz donde cada fila albergara la información de 60 días pasados (timesteps = 60), y estos 60 días se usaran para predecir el día 61 y en la segunda fila, existirán 60 días también, pero comenzará desde el día 2 hasta el día 61, para predecir el 62 y la tercera fila reunirá desde el día 3 al 62, para predecir el 63 y así sucesivamente hasta cubrir todos los días (ver Ilustración 5), es decir, la red estará en constante aprendizaje y por ejemplo para predecir el día 1 de febrero del 2022, necesitara la información del número de ventas realizadas en los meses de diciembre del 2021 y enero del 2022 (60 días) para poder predecir el número de ventas en el día 1 de febrero del 2022.

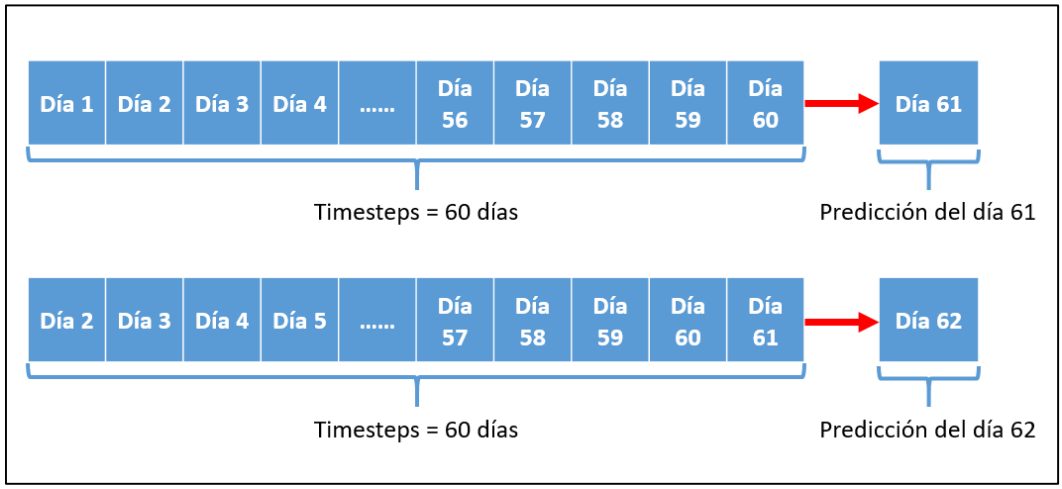

Ilustración 5: Timesteps de 60 días empleado para el aprendizaje de la red neuronal recurrente.

En la Ilustración 6 se observa la matriz de entrenamiento “X_train” a la izquierda y el “y_train” a la derecha, donde los datos de las ventas están estandarizados usando una estandarización normal, de este modo, cada valor numérico en las ventas de bebidas tiene el mismo peso en la red neuronal.

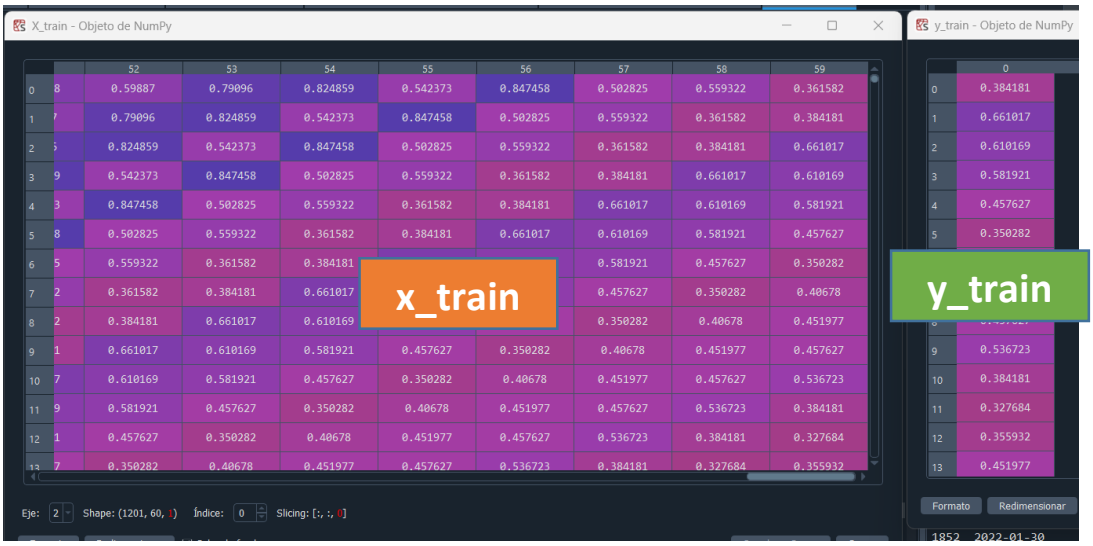

Ilustración 6: Tensor de entrada a la red neuronal recurrente, el cual contiene los datos de las unidades de bebida estandarizados, segmentado en "x_train" e "y_train".

Arquitectura de la Red Neuronal:

El modelo predictivo para las ventas de bebidas se creara a través de una red neuronal recurrente, la cual aprenderá por aprendizaje supervisado los datos de las ventas de bebidas desde enero del 2017 a enero del 2022, para ello la red estará formada por 5 capas de Long Short Term Memory (LSTM), donde cada una de estas capas estará compuesta por 50 neuronas, el número de neuronas de cada capa puede variar, pero se usan 50 neuronas y 5 capas ocultas, debido a que el modelo debe ser capaz de obtener una predicción de un dato especifico (venta diaria), por lo que es un problema de regresión, donde se necesita de un gran número de neuronas para que puedan capturar las tendencias y las variaciones que existen entre cada día para la venta de las bebidas a lo largo del tiempo. También es destacable que después de cada capa oculta se agrega una capa de Dropout, o desactivación neuronal, que básicamente desactiva de forma aleatoria el 10% de las neuronas en cada capa oculta durante en el proceso del aprendizaje hacia adelante para evitar el sobreajuste de la red y que sobreentrene con los datos de entrenamiento. A su vez se usa el optimizador de "RMSprop", debido a que es muy útil en problemas de regresión y fue más eficaz que el optimizador de "Adam".

```
####### modelado de la red neuronal ######

from keras.models import Sequential
from keras.layers import Dense, LSTM, Dropout

regressor = Sequential()
regressor.add(LSTM(units = 50, return_sequences = True,
                   input_shape = (X_train.shape[1], 1)))
regressor.add(Dropout(rate = 0.1))

regressor.add(LSTM(units = 50, return_sequences = True))
regressor.add(Dropout(rate = 0.1))

regressor.add(LSTM(units = 50, return_sequences = True))
regressor.add(Dropout(rate = 0.1))

regressor.add(LSTM(units = 50, return_sequences = True))
regressor.add(Dropout(rate = 0.1))


regressor.add(LSTM(units = 50))
regressor.add(Dropout(rate = 0.1))

regressor.add(Dense(units = 1))


#regressor.compile(optimizer = "adam", loss = "mean_squared_error")
regressor.compile(optimizer = "RMSprop", loss = "mean_squared_error")

regressor.fit(X_train, y_train, epochs = 150, batch_size = 32, verbose = True)
```

Ilustración 7: Modelo de la arquitectura de la Red Neuronal Recurrente en Python compuesta por varias capas apiladas LSTM.

Cabe destacar que el entrenamiento de la red se hace ingresando el x_train y el y_train que corresponden al número de unidades vendidas estandarizadas, donde, por ejemplo, para predecir el día 2 de febrero del 2022, se usaran los 60 días anteriores al 2 de febrero para predecir el 2 de febrero, ya que, en el aprendizaje, la red aprenderá las correlaciones y estacionalidad temporal que presentan los datos, por ende, el número de timesteps o pasos será un hiperparametro importante de la red neuronal, así como también lo es el número de epochs y el batch size, donde el primero representa el número de veces que la red entrena y se ajustan los pesos para cada una de las neuronas y el batch size es el tamaño del lote del entrenamiento.

Predicciones para el mes de febrero del 2022

El Gráfico 18, presenta el resultado de las predicciones de la Red Neuronal Recurrente:

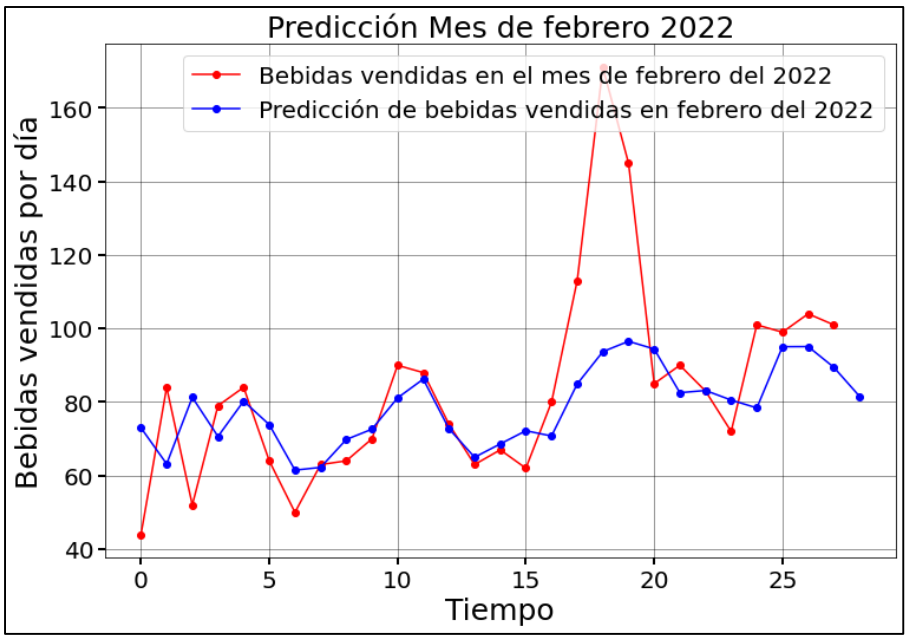

Gráfico 18: Predicción realizada por la Red Neuronal Recurrente para el mes de febrero del 2022.

A partir del gráfico, es posible observar que la red neuronal es capaz de capturar la tendencias en el alza y bajada en el número de bebidas vendidas para el mes de febrero del 2022, no acierta exactamente en el valor exacto, pero es una buena aproximación al número real de bebidas vendidas, la mayor dificultad de la red, fue predecir la brusca subida que existe en el 18 de febrero en el número de bebidas vendidas, esto puede explicarse a que la red neuronal solo está aprendiendo en base a la información de bebidas vendidas y puede existir una variable exógena que esté involucrada y que para este análisis no fue tomada en cuenta.

También se calcula el error porcentual absoluto medio (MAPE) del modelo generado, el cual es 15,81%. Este valor de MAPE refleja que el modelo tiene dificultades para predecir la demanda en el consumo de bebidas con tanta precisión, pero a su vez, es algo normal, ya que se sabe que el local ofrece descuentos y promociones especiales constantemente, además de que la venta de bebidas depende de muchos factores externos, como el día de la semana, si las personas tienen o no dinero, la temperatura del día, si existen eventos o fiestas en el litoral central o si es época de verano o invierno, por lo cual, un valor del 15,81% refleja un modelo que captura de buena manera las tendencias de la demanda en las bebidas a pesar de que no acierte exactamente en su valor.

Posibles mejoras al modelo predictivo

Algunas mejoras importantes que puede recibir el modelo predictivo es la inclusión de nuevas variables que afectan al comportamiento en el número de unidades vendidas de bebidas, como lo es la temperatura del día en el litoral central, esta nueva variable es importante, ya que durante los meses de verano hace más calor y la gente tiende a comprar un mayor número de bebidas, por lo tanto, sería una variable interesante a agregar en el modelo, también es posible incorporar que día de la semana es el que se está prediciendo, ya que el día de la semana influirá notoriamente en el nivel de ventas, así como también el mes en el que se realizan las ventas de bebidas.

A continuación, se expone el código del modelo predictivo realizado en Python:

```
### training set será del 1 de enero del 2017 al 31 de enero del 2022
### test set será el mes de febrero del 2022

### Importacion de las librerias ###
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

dataset_train = pd.read_csv("/content/dataset_ts_ventas_train.csv")
print(dataset_train)

#### definicion del training set compuesto por el número de bebidas vendidas ###
training_set = dataset_train.iloc[:,2:3].values

### escalado de la variable "numero de bebidas vendidas" ###
from sklearn.preprocessing import MinMaxScaler
sc = MinMaxScaler()
training_set_scaler = sc.fit_transform(training_set)

### construccion del tensor X_train e y_train para el posterior proceso de
### entrenamiento de la red neuronal ###
X_train = []
y_train = []
for i in range(60, 1261):
    X_train.append(training_set_scaler[i-60: i, 0])
    y_train.append(training_set_scaler[i])

X_train, y_train = np.array(X_train), np.array(y_train)

X_train = np.reshape(X_train, (X_train.shape[0], X_train.shape[1], 1))
```

Ilustración 8: Código de la Red Neuronal Recurrente creada en Python, parte 1.

```
####### modelado de la red neuronal ######
### Importación de la capa secuencial, densa, LSTM y dropout
from keras.models import Sequential
from keras.layers import Dense, LSTM, Dropout

regressor = Sequential()
### Primera Capa Oculta de la Red Neuronal ###
regressor.add(LSTM(units = 50, return_sequences = True,
                   input_shape = (X_train.shape[1], 1)))
regressor.add(Dropout(rate = 0.1))

### Segunda Capa Oculta de la Red Neuronal ###
regressor.add(LSTM(units = 50, return_sequences = True))
regressor.add(Dropout(rate = 0.1))

### Tercera Capa Oculta de la Red Neuronal ###
regressor.add(LSTM(units = 50, return_sequences = True))
regressor.add(Dropout(rate = 0.1))

### Cuarta Capa Oculta de la Red Neuronal ###
regressor.add(LSTM(units = 50, return_sequences = True))
regressor.add(Dropout(rate = 0.1))

### Quinta Capa Oculta de la Redn Neuronal
regressor.add(LSTM(units = 50))
regressor.add(Dropout(rate = 0.1))

### Capa de Salida de la Red Neuronal
regressor.add(Dense(units = 1))

regressor.compile(optimizer = "RMSprop", loss = "mean_squared_error")

regressor.fit(X_train, y_train, epochs = 150, batch_size = 32, verbose = True)
```

Ilustración 9: Código de la Red Neuronal Recurrente creada en Python, parte 2.

```
######## Etapa de predicción de datos ########
##### unidades de bebidas vendidas en el mes de febrero del 2022 #####
dataset_test = pd.read_csv("/content/dataset_ts_ventas_test.csv")
print(dataset_test)
real_stock_price = dataset_test.iloc[:, 2:3].values
#########################################################


dataset_total = pd.concat((dataset_train["unidades_total"], dataset_test["unidades_total"]), axis = 0)
inputs = dataset_total[len(dataset_total) - len(dataset_test) - 60:].values
inputs = inputs.reshape(-1,1)
inputs = sc.transform(inputs)
X_test = []
## Se crea un nuevo tensor para poder predecir el mes de febrero del 2022
## el cual posee 29 dias por eso va desde el 60 al 89
for i in range(60, 89):
    X_test.append(inputs[i-60:i, 0])
X_test = np.array(X_test)
X_test = np.reshape(X_test, (X_test.shape[0], X_test.shape[1], 1))
### Se crea una instancia para que la red neuronal haga una predicción
predicted_stock_price = regressor.predict(X_test)
### La predicción entrega los resultados normalizados, por lo tanto se debe quitar esta
### normalización para observar el verdadero resultado
predicted_stock_price = sc.inverse_transform(predicted_stock_price)
```

```
fig, ax = plt.subplots(figsize = (12,8))
ax.plot(real_stock_price, color = 'red', label = "bebidas vendidas en el mes de febrero del 2022", marker = "o")
ax.plot(predicted_stock_price, color = 'blue', label = "Prediccion de bebidas vendidas en febrero del 2022", marker = "o")
ax.set_title("Prediction Mes de febrero 2022")
ax.set_xlabel("Tiempo")
ax.set_ylabel("Unidades vendidas de bebidas")
ax.grid()
ax.legend()
```

Ilustración 10: Código de la Red Neuronal Recurrente creada en Python, parte 3.

Conclusiones

La principal variable del dataset estudiado es “el número de bebidas consumidas”, y esta variable presenta estacionalidad, es decir, la forma de su curva a lo largo del tiempo se va repitiendo, con un comportamiento creciente en los meses de verano (enero y febrero) y decrecientes en los meses de invierno (mayo y junio), donde esta regla se repite a lo largo de todo el horizonte temporal estudiado. Es importante que el día en el que más se vende bebidas es el día viernes, donde el 75% de todos los días viernes durante enero del 2017 a enero del 2022, poseen un nivel de ventas no menor a las 95 bebidas diarias, que si lo comparamos con el día más malo de la semana, que es el día domingo, el 75% de todos los días domingos, las ventas diarias no superan las 75 bebidas diarias, siendo su promedio de 61 bebidas al día, en oposición al día viernes que se registran en promedio 79 bebidas diarias vendidas, esto último se debe a que la demanda prefiere consumir más bebida los días en que sabe que al día siguiente tendrá descanso y el día donde termina su jornada laboral.

El efecto de la pandemia también es importante, ya que el covid-19 golpea con fuerza la demanda durante el año 2020, cambiando incluso la distribución de frecuencias en las unidades vendidas, y disminuyendo notoriamente la demanda, alcanzando un promedio anual de solo 54 bebidas al día, donde el 75% de todos los días durante ese año 2020, no superan las 71 ventas diarias, valor bajo si lo comparamos con las 78 bebidas vendidas en promedio en el 2017. Los efectos de la pandemia se agudizan debido a la nueva política que hace aumentar el precio en la bebida, haciendo más fuerte la caída durante el invierno del 2020, debido a las fuertes restricciones de confinamiento que sufre

la población, lo que genera una contracción en la economía nacional. Con relación al precio, es posible destacar que el precio experimenta dos subidas, la primera subida, aumenta el precio en $160 y en la segunda aumenta en $50 pesos y después en $50 pesos más, lo que deja al descubierto que para el negocio es más beneficioso una política que haga aumentar el precio de forma gradual y paulatina en vez de aumentarlo de golpe y de forma brusca, ya que de forma paulatina, hace incrementar la demanda, ya que esta última se adapta bien a cambios lentos en el precio del producto.

Con relación al precio de la bebida, es posible establecer que este sufre constantes modificaciones, ya que el local incorpora durante gran parte del tiempo descuentos y promociones, descuentos que alcanzan los $200 pesos e incluso durante algunos días los $300 pesos, esto se debe al aumento que sufre el precio real de la bebida en un 16%, llegando a $1150 pesos por bebida, valor que hace decrecer la demanda y una medida para recuperarla es disminuir el precio drásticamente con descuentos, lo que hace disminuir la media móvil mensual en las ventas de bebidas, y con ello en los ingresos.

El modelo predictivo creado con una Red Neuronal Recurrente, es capaz de alcanzar un error porcentual medio absoluto del 15,81%, valor aceptable ya que el mercado que trata de predecir es uno muy volátil y que presenta gran variabilidad a lo largo del tiempo, ya que la pandemia, también hace cambiar la distribución de la variable y con ello, también afecta al comportamiento de la demanda, lo que hace más difícil el estudio y predicción del número de unidades vendidas de bebidas, a esto último debemos agregar que durante todo el año, el precio de la bebida si sufre alteraciones, ya que durante muchos días existen descuentos y promociones los que si alteran la demanda y con ello la naturaleza de la curva que debe predecir el modelo.