Universidad del Valle de Guatemala
Data Science - CC3066 - Sección 20
Catedrático: LUIS ROBERTO FURLAN COLLVER

Integrantes:
JULIO ROBERTO HERRERA SABAN
DIEGO DE JESUS ARREDONDO TURCIOS

Laboratorio 9
Predicción Precios Aguacates en EEUU

Link del Repositorio: https://github.com/arr19422/Lab9-DS

Análisis exploratorio

El *dataset* contiene datos que representan ventas semanales desde 2015 a 2018 a nivel de todos los Estados Unidos, únicamente de la variedad Hass, proviniendo directamente de cajas registradoras. Cabe mencionar que cuando se habla del precio promedio, se refleja el costo por unidad de aguacate, aunque estos se venden en bolsa.

Como parte del análisis exploratorio primero damos un vistazo rápido a los registros que incluye el *dataset*.

	Unnamed: 0	Date	AveragePrice	Total Volume	4046	4225	4770	Total Bags	Small Bags	Large Bags	XLarge Bags	type	year	region
0	0	2015-12-27	1.33	64236.62	1036.74	54454.85	48.16	8696.87	8603.62	93.25	0.0	conventional	2015	Albany
1	1	2015-12-20	1.35	54876.98	674.28	44638.81	58.33	9505.56	9408.07	97.49	0.0	conventional	2015	Albany
2	2	2015-12-13	0.93	118220.22	794.70	109149.67	130.50	8145.35	8042.21	103.14	0.0	conventional	2015	Albany
3	3	2015-12-06	1.08	78992.15	1132.00	71976.41	72.58	5811.16	5677.40	133.76	0.0	conventional	2015	Albany
4	4	2015-11-29	1.28	51039.60	941.48	43838.39	75.78	6183.95	5986.26	197.69	0.0	conventional	2015	Albany

Cuadro 1: Primeros 5 registros del *dataset*.

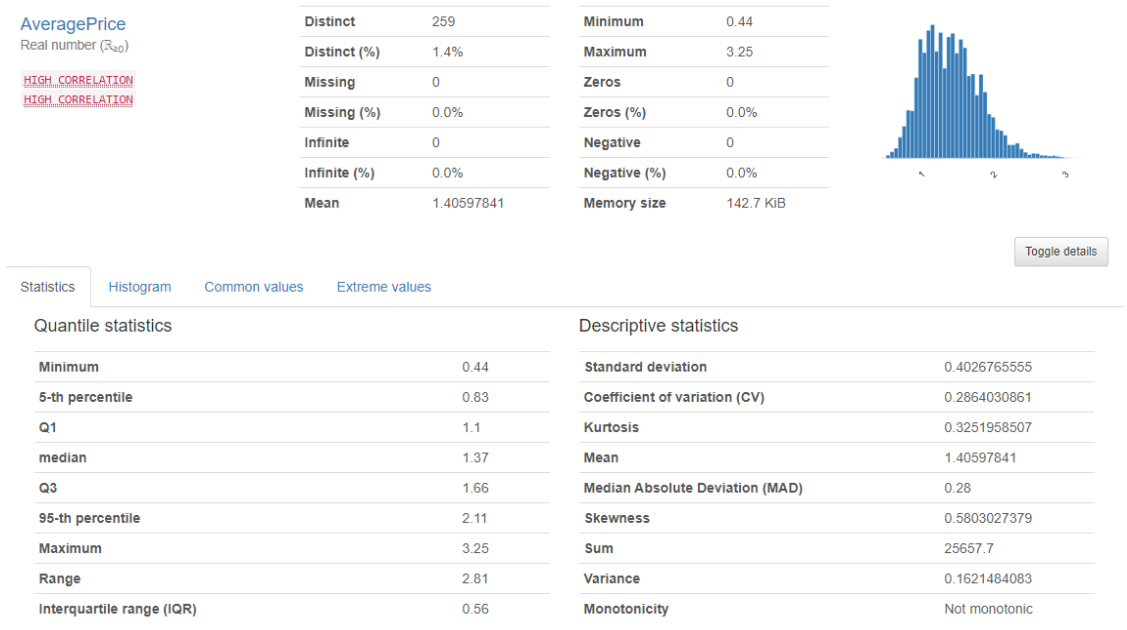
Para comprender cada una de las variables del registro recordemos lo que representa cada una de ellas:

Variable	Descripción	Tipo	Tipo
Date	Fecha de la observación	Cualitativa	Ordinal
AveragePrice	Precio promedio por unidad	Cuantitativa	Continua
type	Convencional u orgánico	Cualitativa	Nominal
year	Año	Cuantitativa	Discreta
Region	La ciudad o región de la observación	Cualitativa	Nominal
Total Volume	Número total de aguacates vendidos	Cuantitativa	Discreta
4046	Número total de aguacates con PLU 4046 vendidos	Cuantitativa	Discreta
4225	Número total de aguacates con PLU 4225 vendidos	Cuantitativa	Discreta
4770	Número total de aguacates con PLU 4770 vendidos	Cuantitativa	Discreta

Cuadro 2: Variables del *dataset*.

Podemos observar que en el *dataset* existen algunas columnas desconocidas, como “Unnamed: 0” la cual representa solo el index de cada registro y es duplicada del index que ya maneja Pandas, Small Bags, Large Bags y XLarge Bags que representan el tamaño de bolsas vendidas y Total Bags que es una suma de estas.

Siguiendo con el análisis exploratorio se realiza un reporte general de cada una de las variables con el fin de observar registros faltantes, porcentaje de valores cero y negativos, distribución de los datos y estadísticas generales como media, mediana y moda.



AveragePrice
Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION
HIGH CORRELATION

Distinct	259	**Minimum**	0.44
Distinct (%)	1.4%	**Maximum**	3.25
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	1.40597841	**Memory size**	142.7 KiB

Toggle details

Statistics | Histogram | Common values | Extreme values

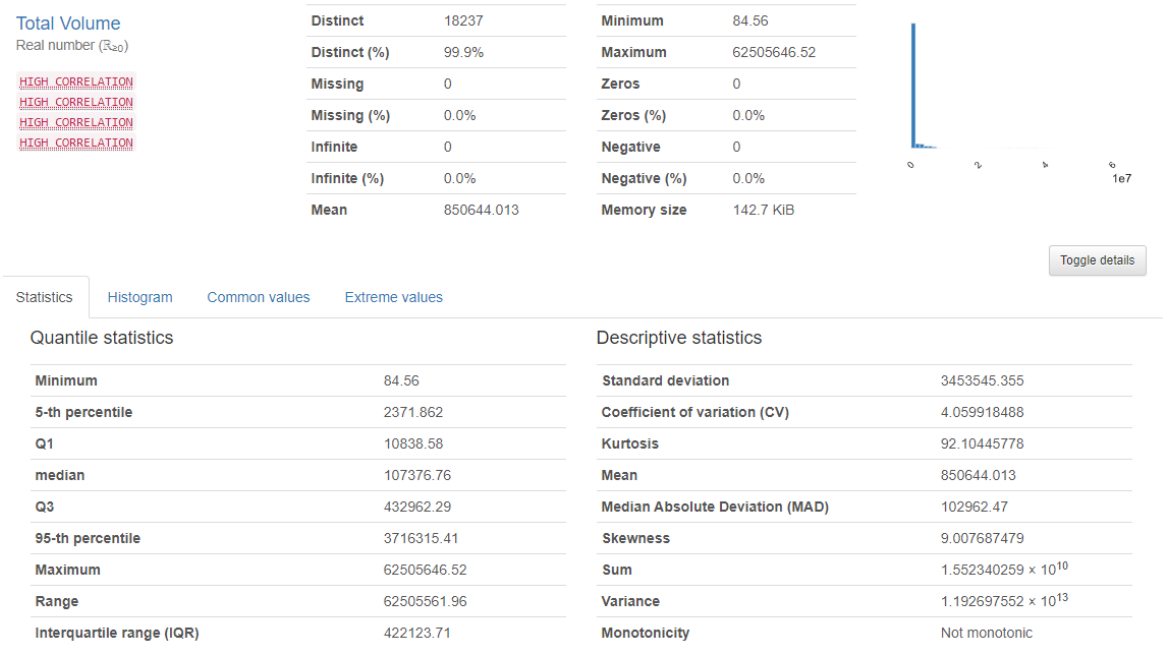
Quantile statistics

Minimum	0.44
5-th percentile	0.83
Q1	1.1
median	1.37
Q3	1.66
95-th percentile	2.11
Maximum	3.25
Range	2.81
Interquartile range (IQR)	0.56

Descriptive statistics

Standard deviation	0.4026765555
Coefficient of variation (CV)	0.2864030861
Kurtosis	0.3251958507
Mean	1.40597841
Median Absolute Deviation (MAD)	0.28
Skewness	0.5803027379
Sum	25657.7
Variance	0.1621484083
Monotonicity	Not monotonic

Figura 1: Reporte de la variable AveragePrice



Total Volume
Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION

Distinct	18237	**Minimum**	84.56
Distinct (%)	99.9%	**Maximum**	62505646.52
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	850644.013	**Memory size**	142.7 KiB

Toggle details

Statistics | Histogram | Common values | Extreme values

Quantile statistics

Minimum	84.56
5-th percentile	2371.862
Q1	10838.58
median	107376.76
Q3	432962.29
95-th percentile	3716315.41
Maximum	62505646.52
Range	62505561.96
Interquartile range (IQR)	422123.71

Descriptive statistics

Standard deviation	3453545.355
Coefficient of variation (CV)	4.059918488
Kurtosis	92.10445778
Mean	850644.013
Median Absolute Deviation (MAD)	102962.47
Skewness	9.007687479
Sum	$1.552340259 \times 10^{10}$
Variance	$1.192697552 \times 10^{13}$
Monotonicity	Not monotonic

Figura 2: Reporte de la variable Total Volume.

4046
Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
ZEROS

Distinct	17702	Minimum	0
Distinct (%)	97.0%	Maximum	22743616.17
Missing	0	Zeros	242
Missing (%)	0.0%	Zeros (%)	1.3%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	293008.4245	Memory size	142.7 KiB

Figura 3: Resumen de reporte de la variable 4046.

4225
Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION

Distinct	18103	Minimum	0
Distinct (%)	99.2%	Maximum	20470572.61
Missing	0	Zeros	61
Missing (%)	0.0%	Zeros (%)	0.3%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	295154.5684	Memory size	142.7 KiB

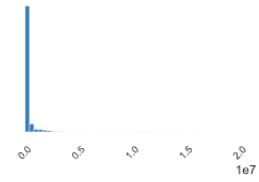


Figura 4: Resumen de reporte de la variable 4225.

4770
Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
ZEROS

Distinct	12071	Minimum	0
Distinct (%)	66.1%	Maximum	2546439.11
Missing	0	Zeros	5497
Missing (%)	0.0%	Zeros (%)	30.1%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	22839.73599	Memory size	142.7 KiB

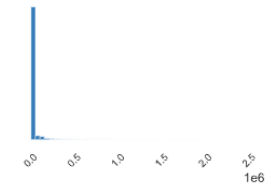


Figura 5: Resumen de reporte de la variable 4770.

Total Bags
Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION

Distinct	18097	Minimum	0
Distinct (%)	99.2%	Maximum	19373134.37
Missing	0	Zeros	15
Missing (%)	0.0%	Zeros (%)	0.1%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	239639.2021	Memory size	142.7 KiB

Figura 6: Resumen de reporte de la variable Total Bags.

Small Bags
Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION

Distinct	17321	Minimum	0
Distinct (%)	94.9%	Maximum	13384586.8
Missing	0	Zeros	159
Missing (%)	0.0%	Zeros (%)	0.9%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	182194.6867	Memory size	142.7 KiB

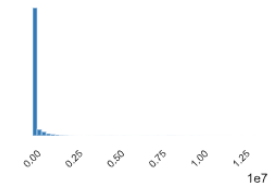


Figura 7: Resumen de reporte de la variable Small Bags.

Large Bags
Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
ZEROS

Distinct	15082	Minimum	0
Distinct (%)	82.6%	Maximum	5719096.61
Missing	0	Zeros	2370
Missing (%)	0.0%	Zeros (%)	13.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	54338.08814	Memory size	142.7 KiB

Figura 8: Resumen de reporte de la variable Large Bags.

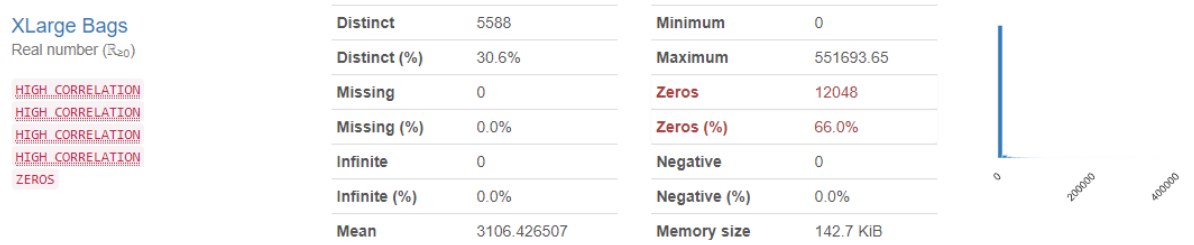


Figura 9: Resumen de reporte de la variable XLarge Bags.

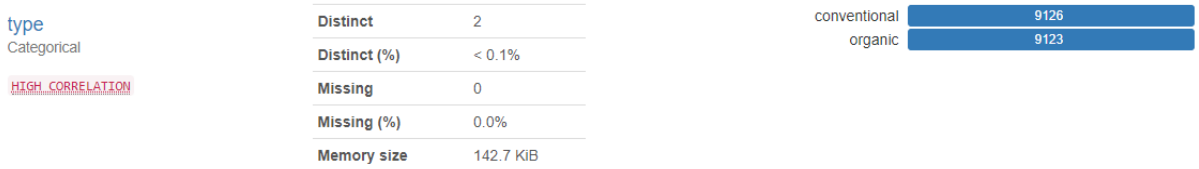


Figura 10: Resumen de reporte de la variable type.

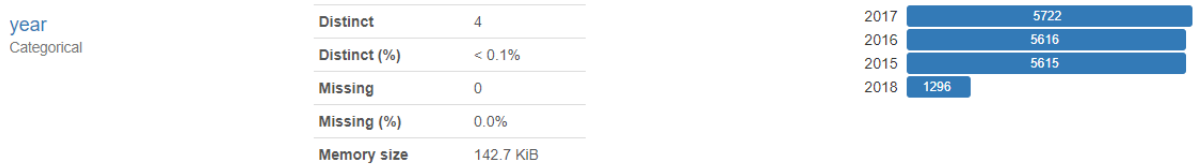


Figura 11: Resumen de reporte de la variable year.

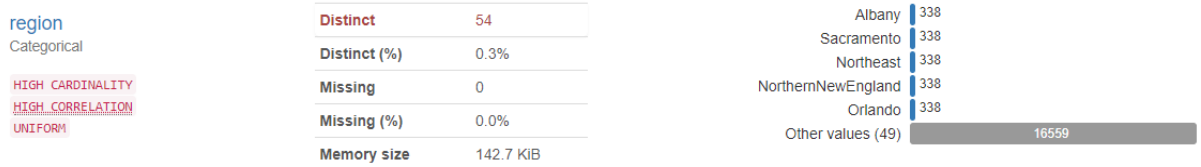


Figura 12: Resumen de reporte de la variable region.

A partir de este reporte no se observa ninguna alerta sobre datos faltantes y en cuanto a las que tienen alto porcentaje de valores cero es aceptable debido a lo que representan. Lo que sí nos indica la distribución de las variables AveragePrice, 4046, 4225, 4770, Small Bags, Large Bags, XLarge Bags y Total Bags es que requiere que se normalicen los datos ya que tienen un rango de valores demasiado alto.

Procesamiento de los datos

Con el fin de asegurarse que los datos estén en orden cronológico, se realiza un cambio en el *dataset* para ordenar a partir de la variable Date y se puede observar como ahora los primeros registros son aquellos con menor fecha y los últimos los que tienen una fecha más reciente.

	Date	AveragePrice	Total Volume	4046	4225	4770	Total Bags	Small Bags	Large Bags	XLarge Bags	type	year	region
51	2015-01-04	1.75	27365.89	9307.34	3844.81	615.28	13598.46	13061.10	537.36	0.00	organic	2015	Southeast
51	2015-01-04	1.49	17723.17	1189.35	15628.27	0.00	905.55	905.55	0.00	0.00	organic	2015	Chicago
51	2015-01-04	1.68	2896.72	161.68	206.96	0.00	2528.08	2528.08	0.00	0.00	organic	2015	HarrisburgScranton
51	2015-01-04	1.52	54956.80	3013.04	35456.88	1561.70	14925.18	11264.80	3660.38	0.00	conventional	2015	Pittsburgh
51	2015-01-04	1.64	1505.12	1.27	1129.50	0.00	374.35	186.67	187.68	0.00	organic	2015	Boise
...	...	...	...	...	...	...	...	...	...	...	...	...	...
0	2018-03-25	1.36	908202.13	142681.06	463136.28	174975.75	127409.04	103579.41	22467.04	1362.59	conventional	2018	Chicago
0	2018-03-25	0.70	9010588.32	3999735.71	966589.50	30130.82	4014132.29	3398569.92	546409.74	69152.63	conventional	2018	SouthCentral
0	2018-03-25	1.42	163496.70	29253.30	5080.04	0.00	129163.36	109052.26	20111.10	0.00	organic	2018	SouthCentral
0	2018-03-25	1.70	190257.38	29644.09	70982.10	0.00	89631.19	89424.11	207.08	0.00	organic	2018	California
0	2018-03-25	1.34	1774776.77	63905.98	908653.71	843.45	801373.63	774634.09	23833.93	2905.61	conventional	2018	NewYork

Cuadro 3: Head y Tail del dataset ordenado cronológicamente.

Visualización de los datos

Para observar cómo se comportan las relaciones entre los datos de algunas de las variables se realizaron algunas gráficas, unas de las más importantes ya que contamos con ellos, es saber cómo se comporta tal variable sobre el tiempo, sobre la fecha, haciendo un cruce de variables sobre x variable por la fecha. Lo primero es el comportamiento de precios promedios sobre la fecha, la tendencia a nivel nacional muestra una fluctuación constante en el precio promedio pero que entre el 2017 y el 2018 se ha acentuado más, por ejemplo teniendo en el mes de junio de 2017 los mayores precios mínimos.

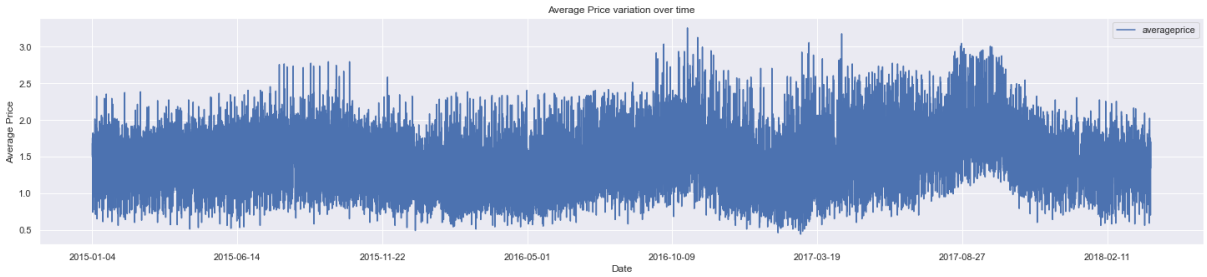


Figura 13: Variación del precio promedio sobre el tiempo.

Como se vió en el análisis exploratorio, existen 54 distintas regiones y para ver cuales son las frecuencias de cada una de ellas, se realizó una gráfica de barras por medio de una agrupación de regiones.

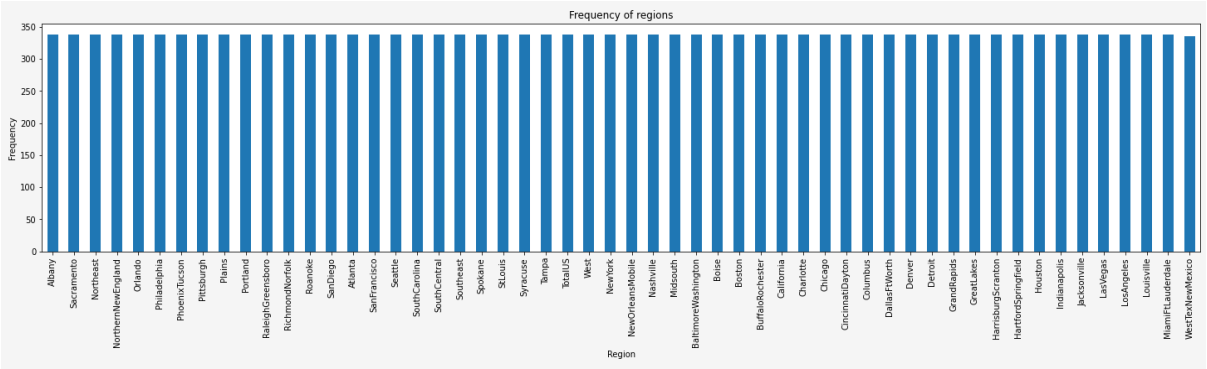


Figura 14: Frecuencia de regiones

Aparte de la variable de fecha, se tiene una variable que indica a qué año pertenece cada registro, así que se realiza una gráfica de barras para saber de qué años se tiene información.

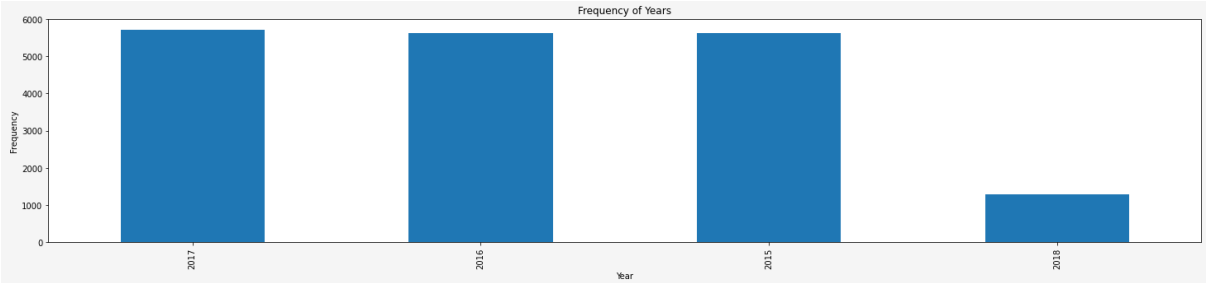


Figura 15: Frecuencia de años.

También graficamos la variación del precio promedio en cada uno de los meses, es decir que se juntó el promedio en cada uno de los meses de todos los años y se agrupó solamente por meses, mostrando así cuáles meses tienen mayor y menor precio promedio, y en esta gráfica se observa como es la variación y la forma que tiene esta gráfica dependiendo de la temporada de cada año.

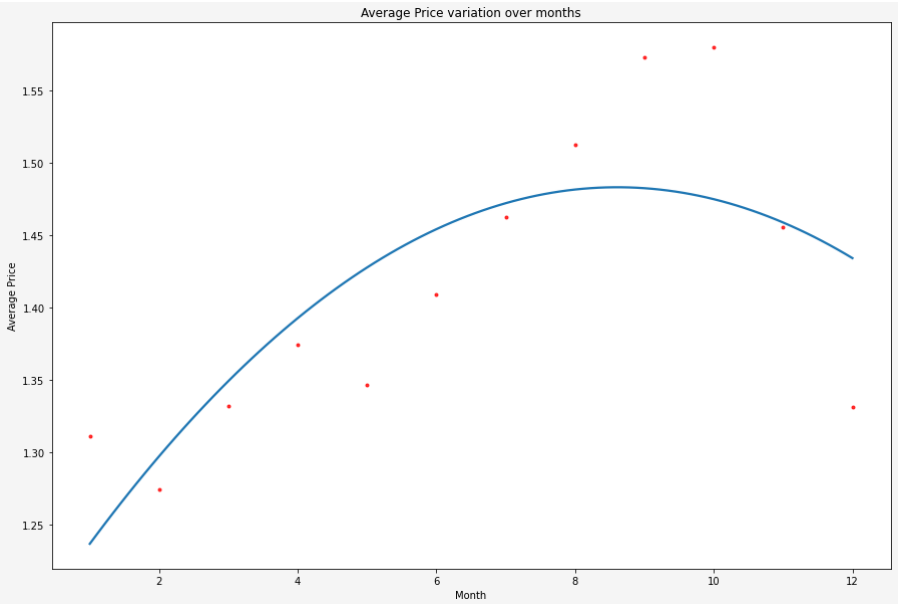


Figura 16: Promedio de AveragePrice por cada mes.

Modelo

Para poder aplicar los datos al modelo se requiere que estos estén normalizados, se utilizó el modelo Tensor de PyTorch el cual permite representar un array multidimensional conteniendo elementos de un solo tipo. Utilizando como evaluación la pérdida de MSE, como optimizador un descenso de gradiente SGD, una secuencia de función de activación Linear, luego una ReLu y luego otra lineal, todo esto con 3 épocas; se obtuvo un accuracy de 0.4353 en MSE con este modelo, lo cual indica que los datos son buenos para realizar predicciones a futuro.

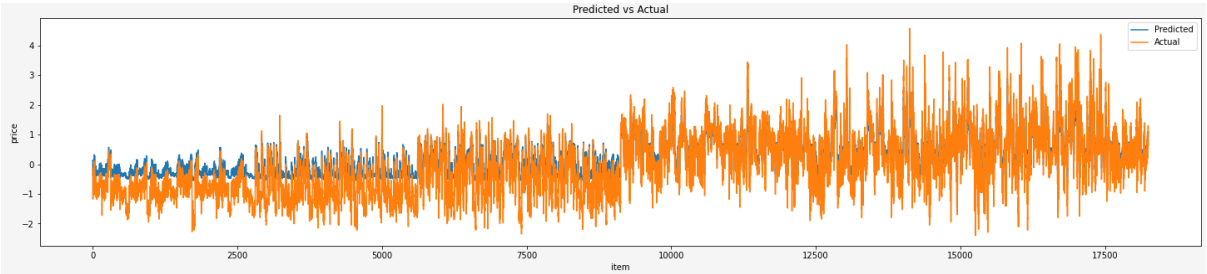


Figura 17: Predicciones vs Reales de AveragePrice.

Haciendo uso del modelo Prophet para predicción de series de tiempo, con un intervalo de incertidumbre del 95% se crearon las predicciones para los siguientes 365 días, dando como resultado la gráfica de la Figura 18, donde se aprecia ese año predicho que va desde los primeros meses del 2018 hasta los primeros meses del 2019.

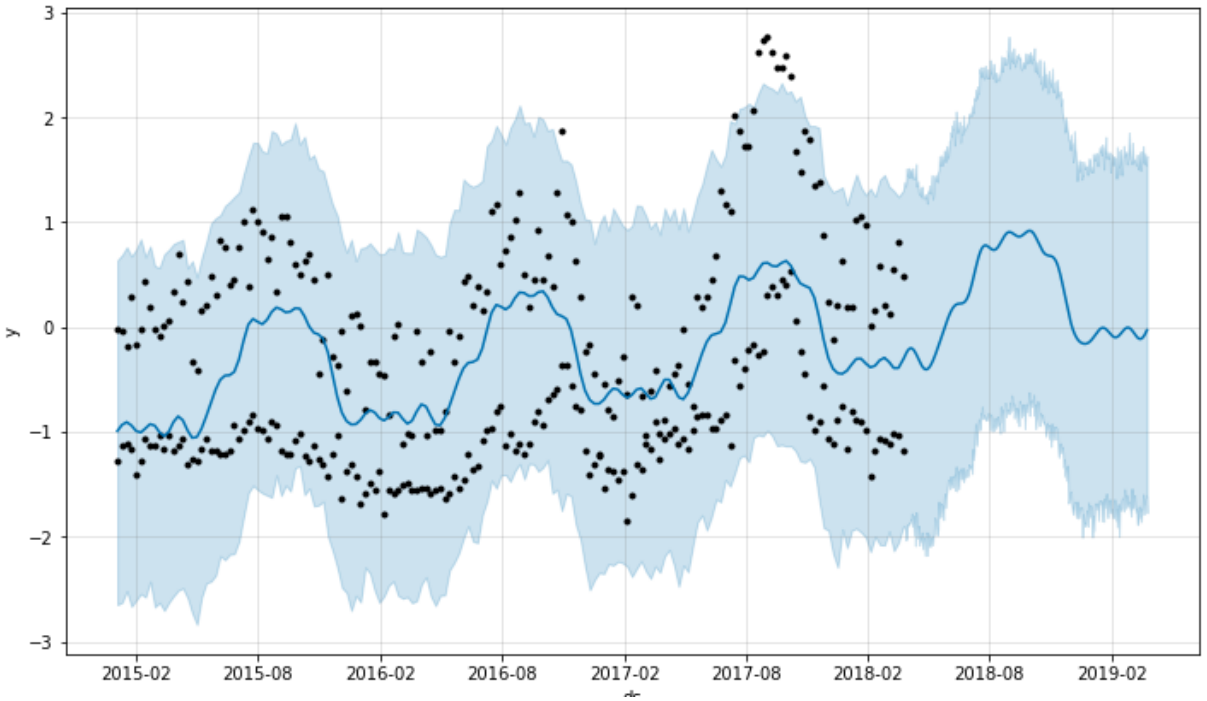


Figura 18: Histórico + Predicción de AveragePrice.

Ahora, haciendo uso de la librería stats model de Python que permite descomponer una serie en sus distintos componentes, trend, seasonal y Residual, se ingresó un dataframe solo con los datos de predicción, es decir, aquellos que van de 2018-03-25 al 2019-3-25. Adicionalmente en el notebook se encuentra la descomposición para los datos históricos, es decir de 2018-03-25 hacia atrás.

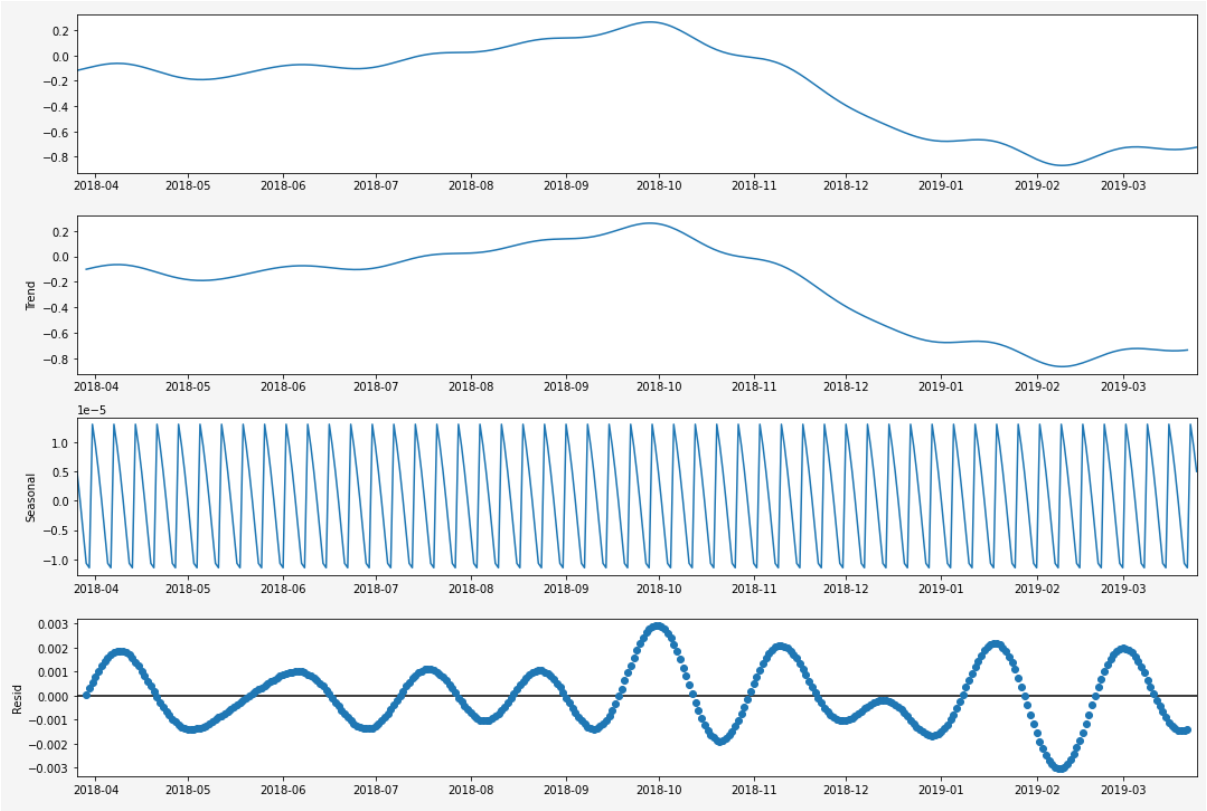


Figura 19: Descomposición de los componentes de predicción.

Parte 2

Para la parte 2 de este laboratorio, se repiten las funciones necesarias de la parte 1 para obtener los resultados requeridos pero ahora solo con los datos que se encuentran en la región Oeste “West” en el *dataset*.

Primero se hace un filtro de los datos, y para volver a observar cómo se comportan, ahora estos datos aislados, se grafican los precios promedios de esta región contra el tiempo. En esta gráfica se puede observar de mejor manera la fluctuación del precio a través de los meses y su repetición durante los años, con la diferencia que a mediados de 2016 y por el tercer mes de 2017 se vió una menor variación en este precio promedio.

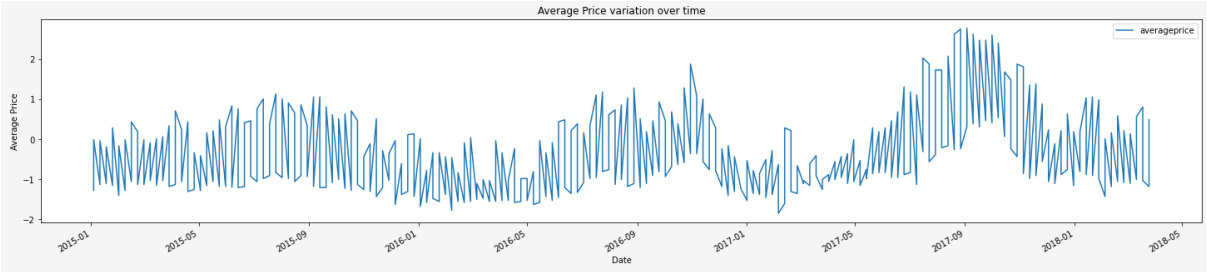


Figura 20: AveragePrice de la región “West” a lo largo del tiempo.

Para predecir los datos a futuro, se creó el modelo Prophet y se especificó una predicción de 365 días futuros, como resultado se obtuvo la siguiente gráfica donde a diferencia de lo

visto en la Figura 18, solo para esta región se predice una disminución más constante que a nivel nacional, donde a finales del 2018 se espera una subida de precios.

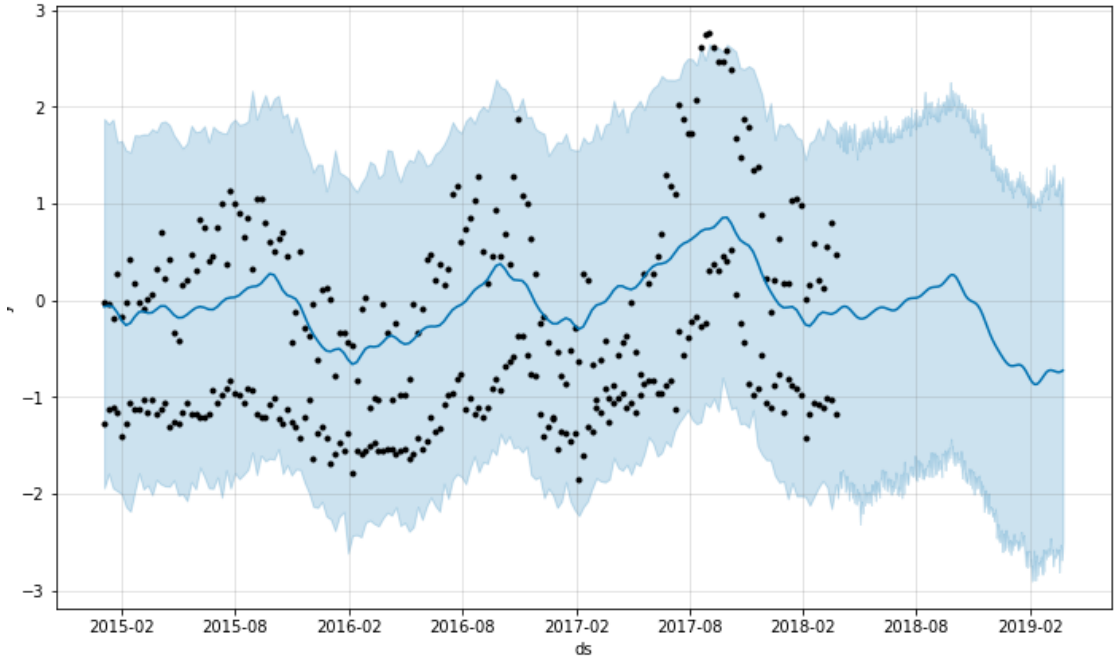


Figura 21: Histórico + Predicción de AveragePrice para la región “West”.

Así mismo se realizó la respectiva descomposición de los componentes Trend, Seasonal y Residual para la predicción de esta región, donde vemos un mayor cambio en la tendencia a comparación de lo visto en la Figura 19 para el nivel nacional.

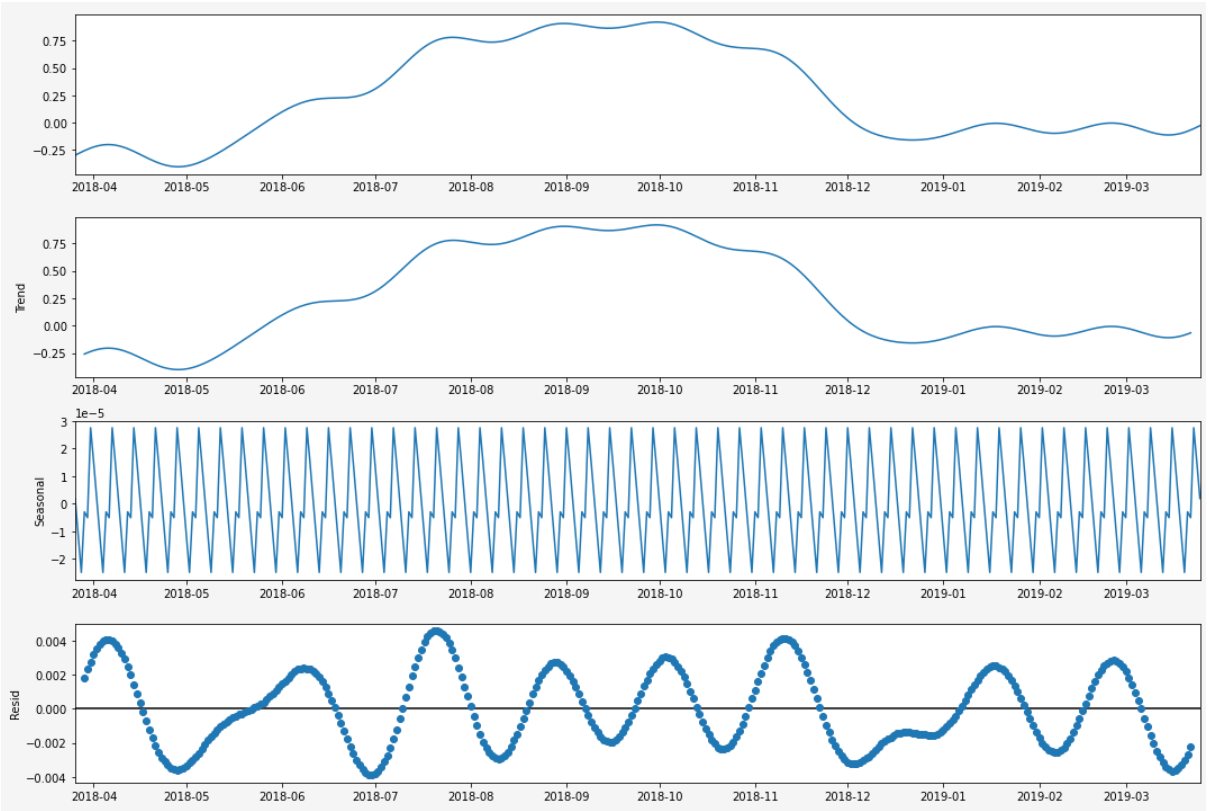


Figura 22: Componentes Trend, Seasonal y Resid para la predicción de la región “West”.