Pricing Used Rough Terrain Cranes

With Web Scraping and Generalized Linear Regression Models

Diego Beteta

Introducción

Contexto

En un entorno cambiante, con nuevos competidores surgiendo constantemente y en medio de una transformación digital, es importante que las empresas cuenten con criterios adecuados para la fijación de precios (Pricing).

Para que puedan determinar su participación, anticiparse a las tendencias e identificar las características más influyentes en el precio de los productos, las empresas deben disponer de data actual del mercado.

Para resolver esta interrogante es necesario la aplicación de 02 técnicas de Data Science, Web Scraping y Modelos Lineales Generalizados, con el objetivo guiar la toma de decisiones estratégicas que equilibren la rentabilidad de la empresa y la percepción de los clientes.

Producto

Las grúa para terrenos difíciles están montadas en un chasis de sólo 04 ruedas. Tanto las ruedas como la base son más anchas para aumentar la estabilidad al momento de izar cargas.

Vienen con una pluma telescópica, y estabilizadores laterales. Se operan y conducen desde una sóla cabina pequeña, sólo tiene un motor que alimenta tanto la pluma como el motor.

Su tracción en las 04 ruedas permite una fácil maniobrabilidad en espacios reducidos. Sin embargo, no pueden desplazarse a gran velocidad en la vía pública, lo que significa que debe transportarse sobre una camabaja a los lugares de trabajo.

En este mercado, los datos que más influyen en el precio de un equipo usado son: la capacidad de carga (UStons), el año de fabricación y horas trabajadas hasta la fecha (horómetro).

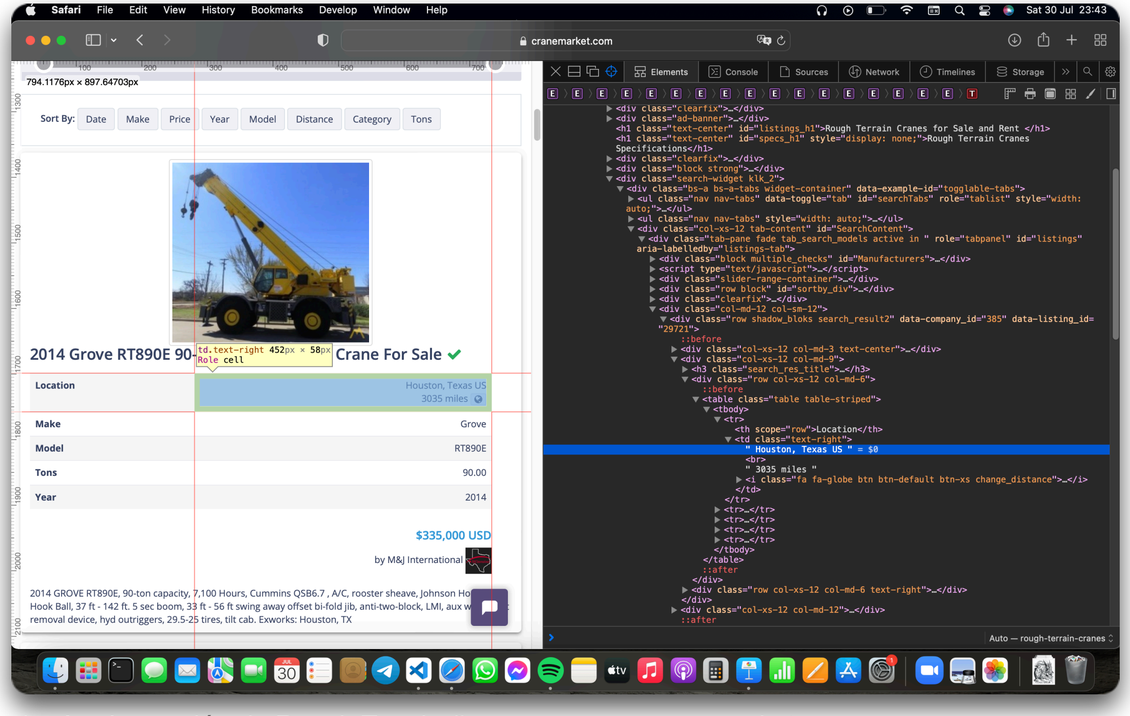
Pricing Used Rough Terrain Cranes

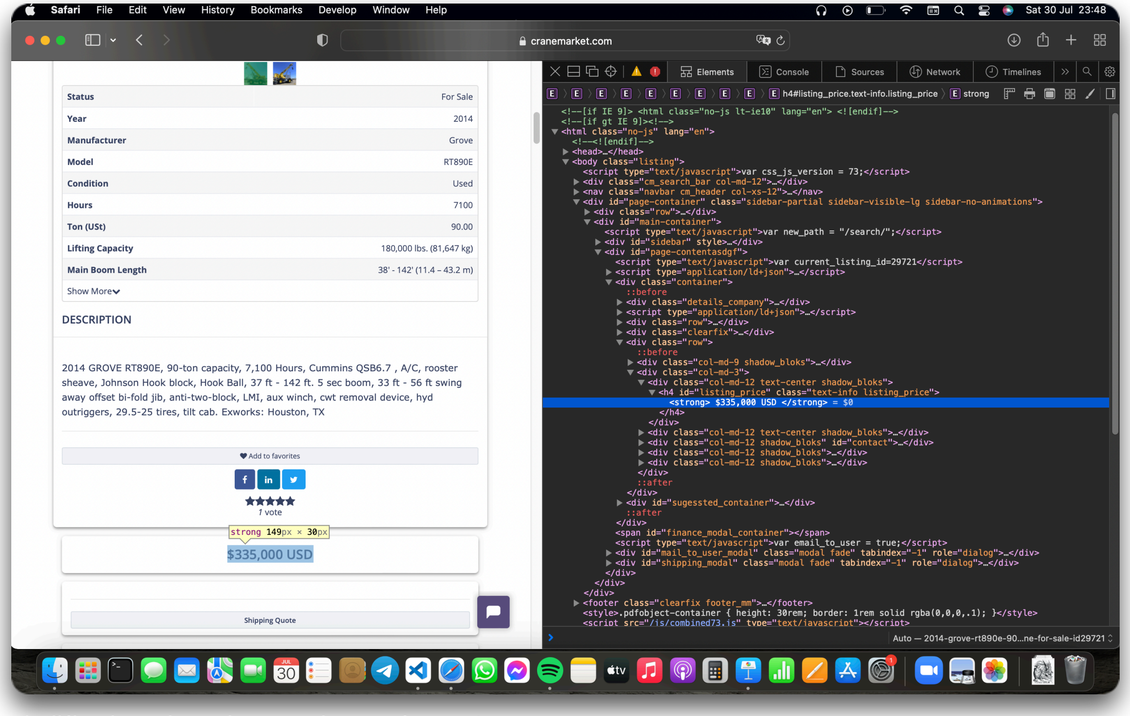
Contenido

1. Extracción de datos con Web Scraping
2. Análisis Exploratorio de Datos (todos los productos)
3. Análisis Exploratorio de Datos (sólo productos con precio disponible)
4. Predicción de precios con Modelo Regresión Lineal Simple
5. Predicción de precios con Modelos Lineales Generalizados
6. Conclusiones

1. Extracción de datos con Web Scraping

Importar, limpiar, filtrar y organizar datos





1. Ir a la sección de Rough Terrain Cranes en www.cranemarket.com
2. Identificar la cantidad de productos yendo al final de las páginas
3. Seleccionar un producto

4. Clic derecho e inspeccionar elementos
5. Analizar la estructura de HTML y etiquetas CSS
6. Ubicar datos relevantes (capacidad, horómetro, horas, precio, etc.)

1. Extracción de datos con Web Scraping

Importar, limpiar, filtrar y organizar datos

```python
# Import scrapy
import scrapy
import numpy as np
import pandas as pd

# turn-off twisted
import sys
if "twisted.internet.reactor" in sys.modules:
    del sys.modules["twisted.internet.reactor"]

class UsedCranesSpider(scrapy.Spider):
    name = "usedcranes"
    allowes_domains = ['https://cranemarket.com']
    url = 'https://cranemarket.com/search/rough-terrain-cranes?page={}'

    # start requests method
    def start_requests( self ):
        for i in range(1, 100):
            yield scrapy.Request(url = self.url.format(i), callback = self.parse_front)

    # First parsing method
    def parse_front( self, response ):
        product_blocks = response.css('h3.search_res_title')
        product_links = product_blocks.xpath('./a/@href')
        links_to_follow = product_links.extract()
        for url in links_to_follow:
            yield response.follow( url = url, callback = self.parse_pages)

    # Second parsing method
    def parse_pages( self, response ):

            table_properties = response.css('table#listing_properties_table > tbody > tr > th::text').getall()
            table_properties_values = response.css('table#listing_properties_table > tbody > tr > td.text-right::text').getall()
            table_dict = dict(zip(table_properties, table_properties_values))

            # Fill empty values that not exist in the table product
            properties = ['Manufacturer', 'Model', 'Ton (USt)', 'Serial Number', 'Year', 'Hours', 'Condition', 'Category']
            for propertie in properties:
                if propertie not in table_dict:
                    table_dict.update({propertie: None})

            yield {
                **dict(list(zip(
                        ['link', 'title', 'location', 'price', 'description'],
                        [response.url,
                        response.css('.col-md-12 > h1#listing_full_name::text').get().strip(),
                        response.css('.col-md-12 > strong::text').get().strip(),
                        response.css('h4#listing_price strong::text').get().strip(),
                        '|'.join(response.css('div#description-content span::text').getall())]))),
                **table_dict
            }
```


```
Rough Terrain Crane For Sale", "location": "Location: Gary, IN, USA", "price": "$599,000 USD", "description": "New 2020 Terex RT100|39' - 154' Main Boom|\nSwing-Away Lattice Jib|\nCummins
For Sale", "location": "Location: 500 World Commerce Pkwy, St. Augustine, Florida 32092, USA", "price": "Price On Request", "description": "S/N: 226299 EQUIPPED WITH 29'-95'4-SECTION BOOM
rane For Sale", "location": "Location: 500 World Commerce Pkwy St. Augustine, Florida 32092", "price": "Price On Request", "description": "ONE 2016 GROVE GRT8100 SN 235616 ROUGH TERRAIN FO
-Ton Rough Terrain Crane For Sale or Rent", "location": "Location: Cropac Equipment Inc, South Service Road West, Oakville, ON, Canada", "price": "Price On Request", "description": "75 US
ation: San Leandro, CA", "price": "$400,000 USD", "description": "", "Status": "For Sale or Rent", "Year": "\n2017 ", "Manufacturer": "\nTadano ", "Model": "\nGR-550XL-3 ", "Condition": "\
ocation: Houston, TX", "price": "$641,025 USD", "description": "", "Status": "For Sale or Rent", "Year": "\n2017 ", "Manufacturer": "\nTadano ", "Model": "\nGR-1000XL-3 ", "Condition": "\n
ation: San Leandro, CA", "price": "$90,000 USD", "description": "The Tadano TR-150XL\u20114 is an excellent hydraulic crane for rough terrains. It has a\u00a0max lift capacity of 15 tons (
ocation: Denver, CO", "price": "$726,495 USD", "description": "", "Status": "For Sale or Rent", "Year": "\n2019 ", "Manufacturer": "\nTadano ", "Model": "\nGR-1000XL-3 ", "Condition": "\nU
ation: San Leandro, CA", "price": "$370,000 USD", "description": "", "Status": "For Sale or Rent", "Year": "\n2016 ", "Manufacturer": "\nTadano ", "Model": "\nGR-550XL-3 ", "Condition": "\
ston, TX", "price": "$140,000 USD", "description": "The Grove RT540E rough terrain crane has a 40-ton capacity and a\u00a0four-section main boom that extends to 102\u00a0feet. With a\u00a0
rrain Crane For Sale or Rent", "location": "Location: Texas, USA", "price": "$706,020 USD", "description": "75-TON RT CRANE; 142' BOOM; 270 HP CUMMINS DIESEL ENGINE; AC & HEATER; MAIN & AU
For Sale", "location": "Location: 500 World Commerce Pkwy, St. Augustine, Florida 32092, USA", "price": "Price On Request", "description": " SN 236662 WITH CUMMINS QSB6.7L TIER 4, VALUE P
RT650E 50-Ton Rough Terrain Crane For Sale or Rent", "location": "Location: Cleveland, OH, USA", "price": "Price On Request", "description": "Cummins\u00a0QSB6.7 TIER 4 Diesel|33'-105' 4
For Sale", "location": "Location: Pittsburgh, PA, USA", "price": "Price On Request", "description": "", "Status": "For Sale", "Year": "\n2008 ", "Manufacturer": "\nGrove ", "Model": "\nRT
For Sale", "location": "Location: Mobile, AL, USA", "price": "Price On Request", "description": "", "Status": "For Sale", "Year": "\n2007 ", "Manufacturer": "\nGrove ", "Model": "\nRT540E
T35/40 40-Ton, 40 Ton, Rough Terrain Crane; CranesList ID: 586 For Sale", "location": "Location: Florida, USA", "price": "$315,000 USD", "description": "with|Five Sheave Quick Reeve Boom H
For Sale", "location": "Location: Toledo, OH, USA", "price": "Price On Request", "description": "", "Status": "For Sale", "Year": "\n2007 ", "Manufacturer": "\nGrove ", "Model": "\nRT650E
on Rough Terrain Crane For Sale", "location": "Location: Columbus, OH, USA", "price": "Price On Request", "description": "Link-Belt RTC80100 II|S/N J737-9786|2007|100 Ton|Rebuilt Detroit D
3 75-Ton Rough Terrain Crane For Sale or Rent", "location": "Location: Chelmsford, MA, USA", "price": "Price On Request", "description": "Available For Sale: 2015 Tadano GR-750XL-3|141.1 F
ion RT60/RT70 70-Ton Rough Terrain Crane; CranesList ID: 594 For Sale", "location": "Location: Florida, USA", "price": "$415,000 USD", "description": "with|Five Sheaves Boom Head and Auxil
For Sale", "location": "Location: 500 World Commerce Pkwy, St. Augustine, Florida 32092, USA", "price": "$345,000 USD", "description": ", SN-234991, 41-128 FT MEGAFORM 4-SECTION FULL POWER
For Sale", "location": "Location: Columbus, OH, USA", "price": "Price On Request", "description": "", "Status": "For Sale", "Year": "\n2006 ", "Manufacturer": "\nGrove ", "Model": "\nRT65
Terrain Crane For Sale or Rent", "location": "Location: DeForest, Wisconsin 53532, USA", "price": "Price On Request", "description": "Please call Darren Reddekopp today for pricing and ad
For Sale", "location": "Location: Atlanta, GA, USA", "price": "Price On Request", "description": "", "Status": "For Sale", "Year": "\n2007 ", "Manufacturer": "\nGrove ", "Model": "\nRT760
For Sale", "location": "Location: Hammond, IN, USA", "price": "Price On Request", "description": "", "Status": "For Sale", "Year": "\n2007 ", "Manufacturer": "\nGrove ", "Model": "\nRT650
ough Terrain Crane For Sale", "location": "Location: Columbus, OH, USA", "price": "Price On Request", "description": "", "Status": "For Sale", "Year": "\n2013 ", "Manufacturer": "\nLink-Be
ough Terrain Crane For Sale", "location": "Location: Nitro, WV, USA", "price": "Price On Request", "description": "", "Status": "For Sale", "Year": "\n2006 ", "Manufacturer": "\nLink-Belt
ough Terrain Crane For Sale or Rent", "location": "Location: Holbrook, MA, USA", "price": "Price On Request", "description": "2015 Tadano GR150XL-1 crane Rough Terrain Crane\u00a0|Includes
rane For Sale", "location": "Location: Cleveland, OH, USA", "price": "Price On Request", "description": "", "Status": "For Sale", "Year": "\n2005 ", "Manufacturer": "\nGrove ", "Model": "\
in Crane For Sale or Rent", "location": "Location: Pflugerville, TX, USA", "price": "$390,000 USD", "description": "33'-57' LATTICE JIB WITH OFFSET|\nAUXILIARY HOIST|\nSTANDARD WIRE ROPE O
Ton Down Cab Rough Terrain Crane For Sale", "location": "Location: Baton Rouge, LA, USA", "price": "Price On Request", "description": "", "Status": "For Sale", "Year": "\n2014 ", "Manufact
ne For Sale", "location": "Location: 9301 E Bloomington Fwy, Minneapolis, Minnesota 55420, USA", "price": "Price On Request", "description": "The Grove GRT8090 90-ton USt Capacity Rough Te
For Sale", "location": "Location: Lima, OH, USA", "price": "Price On Request", "description": "", "Status": "For Sale", "Year": "\n2006 ", "Manufacturer": "\nGrove ", "Model": "\nRT760E "
on Rough Terrain Crane For Sale", "location": "Location: Fort Pierce, FL, USA", "price": "$59,000 USD", "description": "SN \u2013 53I0-1397|60 Ton Capacity|110\u2019 Main Boom|33\u2019 Jib
rain Crane For Sale", "location": "Location: Fort Pierce, FL, USA", "price": "Price On Request", "description": "SN \u2013 58229|60 Ton Capacity|115\u2019 Main Boom|38\u2019 \u2013 60\u201
on Rough Terrain Crane For Sale", "location": "Location: Fort Pierce, FL, USA", "price": "$59,000 USD", "description": "SN \u2013 53I4-1729|60 Ton Capacity|110\u2019 Main Boom|33\u2019 Jib
on Rough Terrain Crane For Sale", "location": "Location: Fort Pierce, FL, USA", "price": "$59,000 USD", "description": "SN \u2013 53I4-0537|50 Ton Capacity|85\u2019 Boom|33\u2019 Off-setta
ough Terrain Crane For Sale", "location": "Location: Pasco, WA, USA", "price": "Price On Request", "description": "15 TON ROUGH TERRAIN CRANE|DETROT ENG|FRONT AND REAR STEER|HYD OUTRIGGERS
For Sale", "location": "Location: Lima, OH, USA", "price": "Price On Request", "description": "Grove RT875E|S/N 224495|2005|75 Ton Capacity,|Cummins Diesel|128'\u00a0Main Boom|56' Jib|Aux
lescopic Crawler Crane For Sale", "location": "Location: Cleveland, OH, USA", "price": "Price On Request", "description": "", "Status": "For Sale", "Year": "\n2004 ", "Manufacturer": "\nLi
rane For Sale", "location": "Location: Fort Wayne, IN, USA", "price": "Price On Request", "description": "Grove RT540E|S/N 233937|2013|40 Ton|Cummins QSB 6.7L\u00a0|102' Main Boom|26'-45'
For Sale", "location": "Location: Knoxville, TN, USA", "price": "Price On Request", "description": "", "Status": "For Sale", "Year": "\n2008 ", "Manufacturer": "\nGrove ", "Model": "\nRT6
Denver, CO", "price": "$165,000 USD", "description": "The Terex RT555-1 Rough Terrain crane has a\u00a0maximum lift capacity of 55 tons and a\u00a0four-section full-power mechanically syn
Terrain Crane For Sale", "location": "Location: South Easton, Easton, MA, USA", "price": "Price On Request", "description": "\n- 39\u2019 to 154\u2019 5-Section Sequenced / Synchronized\u
Crane For Sale", "location": "Location: Nitro, WV, USA", "price": "Price On Request", "description": "", "Status": "For Sale", "Year": "\n2015 ", "Manufacturer": "\nTadano ", "Model": "\n
on Rough Terrain Crane For Sale", "location": "Location: Wilmington, NC, USA", "price": "Price On Request", "description": "Link-Belt RTC-80100 II|S/N J7K0-1842|2010|100 Ton|Rebuilt Detroi
For Sale", "location": "Location: Indianapolis, IN, USA", "price": "Price On Request", "description": "Terex RT665|S/N 16619|2008|65 Ton|Cummins Diesel|110' Main Boom|33'-57' Jib|Aux. Hoi
Rough Terrain Crane For Sale", "location": "Location: 9 Whitmore Avenue, Wayne, New Jersey 07470, USA", "price": "$35,000 USD", "description": "New brakes and a new computer two years ago
ough Terrain Crane For Sale", "location": "Location: Orlando, FL, USA", "price": "Price On Request", "description": "Link-Belt\u00a0|RTC-8065II|S/N J9K2-3186|2012|65 Ton|Cummins Diesel|115
For Sale", "location": "Location: Fort Pierce, FL, USA", "price": "$159,500 USD", "description": "2009 Terex RT780 80 ton FOR SALE|80 ton capacity rough terrain hydraulic crane -\u00a0|CA
or Sale", "location": "Location: Pasco, WA, USA", "price": "Price On Request", "description": "GROVE RT65S\u00a0|ROUGH TERRAIN FOR SALE|2-DRUM|360 SWING|AS IS WHERE IS\u00a0", "Status": "F
For Sale", "location": "Location: 500 World Commerce Pkwy, St. Augustine, Florida 32092, USA", "price": "$577,500 USD", "description": " WITH CUMMINS QSB6.7L TIER 4, 120V ENGINE BLOCK HEA
For Sale", "location": "Location: 10421 Fern Hill Dr Riverview, Florida 33578", "price": "Price On Request", "description": ", 41-134.5 FT MEGAFORM 4-SECTION FULL POWER BOOM, CUMMINS QSB
For Sale", "location": "Location: 10421 Fern Hill Dr Riverview, Florida 33578", "price": "Price On Request", "description": ", 41-134.5 FT MEGAFORM 4-SECTION FULL POWER BOOM, CUMMINS QSB
```


7. Crear un Spider con librería Scrapy en Python
8. Especificar la URL de la página de productos
9. Crear un bucle para extraer los datos de cada producto

10. Importar datos de todos los productos disponibles en formato .JSON
11. Verificar ausencia de errores
12. De esta manera, se logra obtener datos de 932 productos en 13 seg.

1. Extracción de datos con Web Scraping

Importar, limpiar, filtrar y organizar datos

1. Collection

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import json
```

[121] ✓ 0.9s

```
with open("/Users/diegobeteta/Library/Mobile Documents/com~apple~CloudDocs/scrapy/usedcranes/usedcranes.json") as f:
    data = json.load(f)
```

[122] ✓ 0.2s

```
df = pd.DataFrame(columns=[ 'category',
                            'brand',
                            'model',
                            'capacity_ust',
                            'serial_number',
                            'year',
                            'hours',
                            'price',
                            'condition',
                            'location',
                            'description',
                            'link'])

# convert JSON to CSV
for i in range(0, len(data)):
    currentItem = data[i]
    df.loc[i] = [data[i]['Category'],
                 data[i]['Manufacturer'],
                 data[i]['Model'],
                 data[i]['Ton (USt)'],
                 data[i]['Serial Number'],
                 data[i]['Year'],
                 data[i]['Hours'],
                 data[i]['price'],
                 data[i]['Condition'],
                 data[i]['location'],
                 data[i]['description'],
                 data[i]['link']]
df
```

[123] ✓ 0.5s

12. Convertir .JSON en DataFrame

	category	brand	model	capacity_ust	serial_number	year	hours	price	condition	location	description	link
0	\nRough Terrain Cranes	\nTerex	\nRT 100US	100.00	None	\n2020	\n54	$599,000 USD	\nNew	Location: Gary, IN, USA	New 2020 Terex RT100|39' - 154' Main Boom|\nSw...	https://cranemarket.com/new-2020-terex-rt-100-...
1	\nRough Terrain Cranes	\nGrove	\nRT530E-2	30.00	\n226299	\n2007	\n4,578 as of 12/09/22	Price On Request	\nVery Good	Location: 500 World Commerce Pkwy, St. Augusti...	S/N: 226299 EQUIPPED WITH 29'-95'4-SECTION BOO...	https://cranemarket.com/grove-rt530e-30-ton-ro...
2	\nRough Terrain Cranes	\nGrove	\nGRT8100	100.00	\n235616	\n2016	\n4,535	Price On Request	\nVery Good	Location: 500 World Commerce Pkwy St. Augustin...	ONE 2016 GROVE GRT8100 SN 235616 ROUGH TERRAIN...	https://cranemarket.com/grove-grt8100-100-ton-...
3	\nRough Terrain Cranes	\nTadano	\nGR-750XL	75.00	None	\n2013	\n7,927	Price On Request	\nUsed	Location: Cropac Equipment Inc, South Service ...	75 US ton capacity|36.1' - 141.1' five-section...	https://cranemarket.com/2013-tadano-gr-750xl-7...
4	\nRough Terrain Cranes	\nTadano	\nGR-550XL-3	None	None	\n2017	None	$400,000 USD	\nUsed	Location: San Leandro, CA		https://cranemarket.com/tadano-gr-550xl-3-for-...
...	...	...	...	...	...	...	...	...	...	...	...	...
926	\nRough Terrain Cranes	\nGrove	\nRT522	22.00	\n66193	\n1983	\n2672	$29,900 USD	\nAs Is	Location: Sparrow Bush, NY, United States	Detroit 4-53 4 cylinder natural diesel mech |T...	https://cranemarket.com/grove-rt522-22-ton-rou...
927	\nRough Terrain Cranes	\nGrove	\nRT58	14.00	\n30987	\n1975	\n4045	$18,900 USD	\nGood	Location: Sparrow Bush, NY, United States	Transmission: 3 speed automatic |Tires: 17.5-2...	https://cranemarket.com/grove-rt58-14-ton-down...
928	\nRough Terrain Cranes	\nGrove	\nRT755	55.00	None	\n1975	None	$90,000 USD	\nUsed	Location: Mexico	34' - 116' 4-Section Full-Power Boom,|No Jib,|...	https://cranemarket.com/1975-grove-rt755-55-to...
929	\nRough Terrain Cranes	\nGrove	\nRT58	14.00	\n19980	\n1972	\n2511	$18,900 USD	\nGood	Location: Sparrow Bush, NY, United States	Transmission 3-speed automatic |Tires: 17.5-25...	https://cranemarket.com/grove-rt58-14-ton-down...
930	\nRough Terrain Cranes	\nGrove	\nRT58	14.00	\n4005	\n1969	\n 5058	$16,900 USD	\nAs Is	Location: Sparrow Bush, NY, United States	Transmission: 2 speed automatic |Tires: 17.5-2...	https://cranemarket.com/grove-rt58-14-ton-down...

931 rows × 12 columns

13.Limpiar datos nulos, duplicados, caracteres, espacios y signos de número

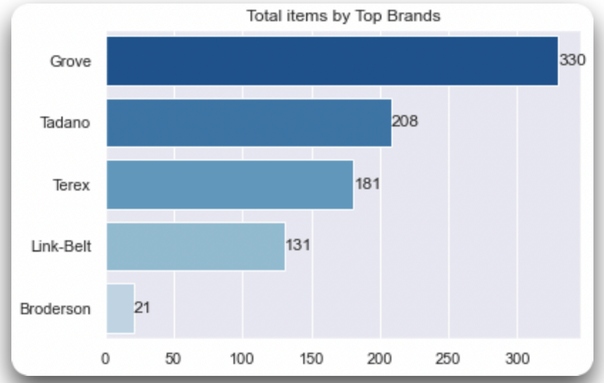
	category	brand	model	capacity_ust	serial_number	year	hours	price	condition	country	description	link
0	Rough Terrain Cranes	Terex	RT 100US	100.00	None	2020	54	599000	New	USA	New 2020 Terex RT100|39' - 154' Main Boom|\nSw...	https://cranemarket.com/new-2020-terex-rt-100-...
1	Rough Terrain Cranes	Grove	RT530E-2	30.00	226299	2007	4578	nan	Very Good	USA	S/N: 226299 EQUIPPED WITH 29'-95'4-SECTION BOO...	https://cranemarket.com/grove-rt530e-30-ton-ro...
2	Rough Terrain Cranes	Grove	GRT8100	100.00	235616	2016	4535	nan	Very Good	USA	ONE 2016 GROVE GRT8100 SN 235616 ROUGH TERRAIN...	https://cranemarket.com/grove-grt8100-100-ton-...
3	Rough Terrain Cranes	Tadano	GR-750XL	75.00	None	2013	7927	nan	Used	Canada	75 US ton capacity|36.1' - 141.1' five-section...	https://cranemarket.com/2013-tadano-gr-750xl-7...
4	Rough Terrain Cranes	Tadano	GR-550XL-3	None	None	2017	None	400000	Used	USA		https://cranemarket.com/tadano-gr-550xl-3-for-...
...	...	...	...	...	...	...	...	...	...	...	...	...
926	Rough Terrain Cranes	Grove	RT522	22.00	66193	1983	2672	29900	As Is	USA	Detroit 4-53 4 cylinder natural diesel mech |T...	https://cranemarket.com/grove-rt522-22-ton-rou...
927	Rough Terrain Cranes	Grove	RT58	14.00	30987	1975	4045	18900	Good	USA	Transmission: 3 speed automatic |Tires: 17.5-2...	https://cranemarket.com/grove-rt58-14-ton-down...
928	Rough Terrain Cranes	Grove	RT755	55.00	None	1975	None	90000	Used	Mexico	34' - 116' 4-Section Full-Power Boom,|No Jib,|...	https://cranemarket.com/1975-grove-rt755-55-to...
929	Rough Terrain Cranes	Grove	RT58	14.00	19980	1972	2511	18900	Good	USA	Transmission 3-speed automatic |Tires: 17.5-25...	https://cranemarket.com/grove-rt58-14-ton-down...
930	Rough Terrain Cranes	Grove	RT58	14.00	4005	1969	5058	16900	As Is	USA	Transmission: 2 speed automatic |Tires: 17.5-2...	https://cranemarket.com/grove-rt58-14-ton-down...

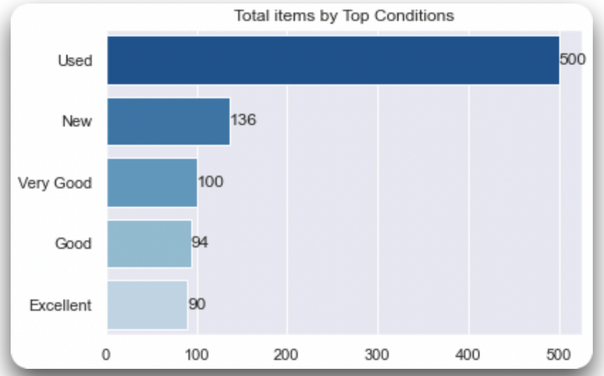
927 rows × 12 columns

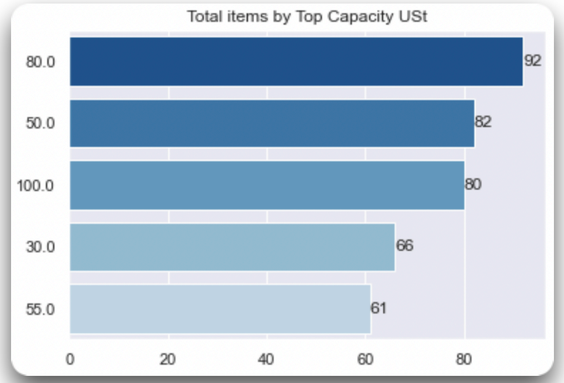
14. Formatear columnas a tipo de datos: category, float and int.

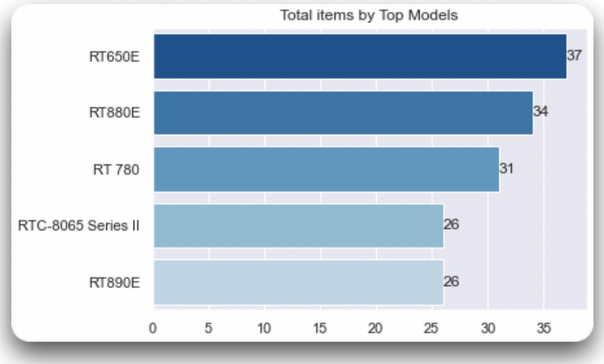
2. Análisis Exploratorio de Datos

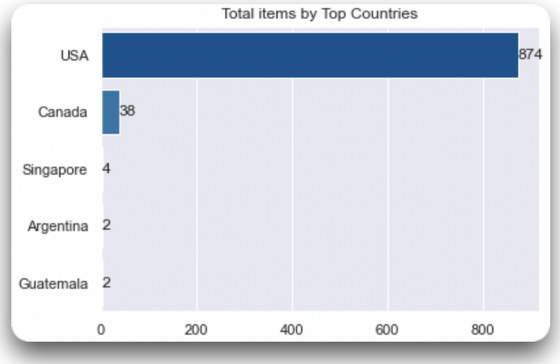
Todos los productos

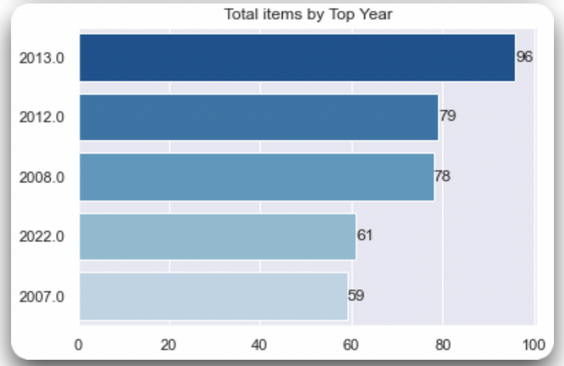






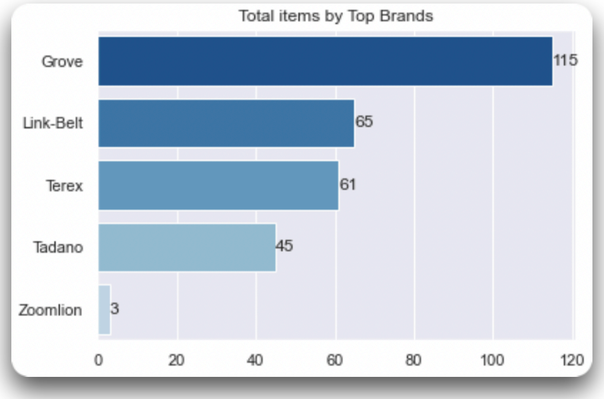


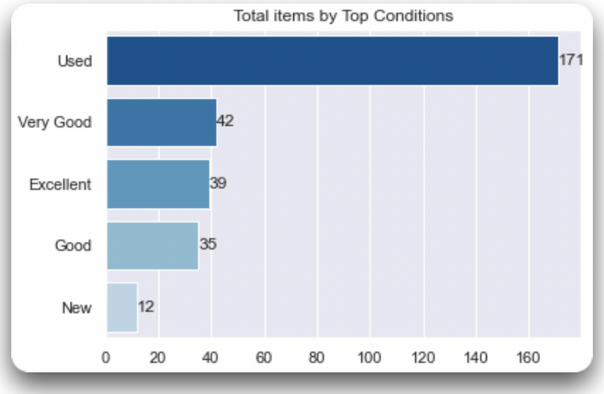


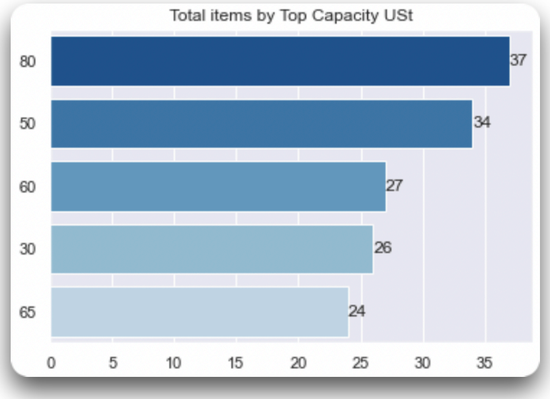


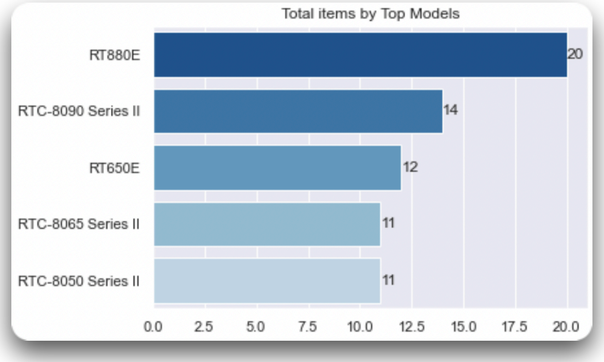
3. Análisis Exploratorio de Datos

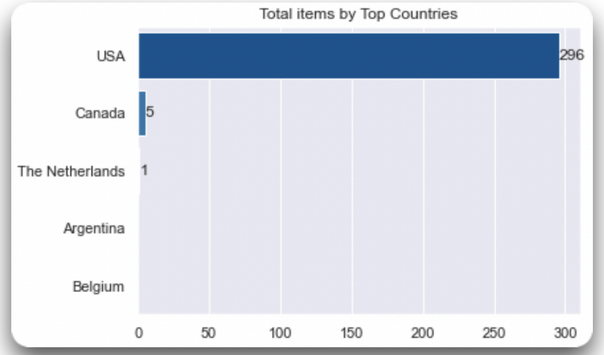
Sólo productos con precio disponible

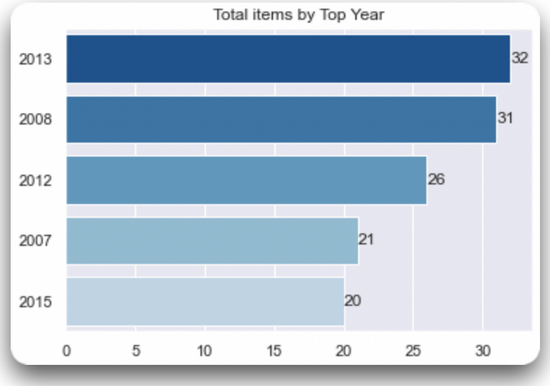






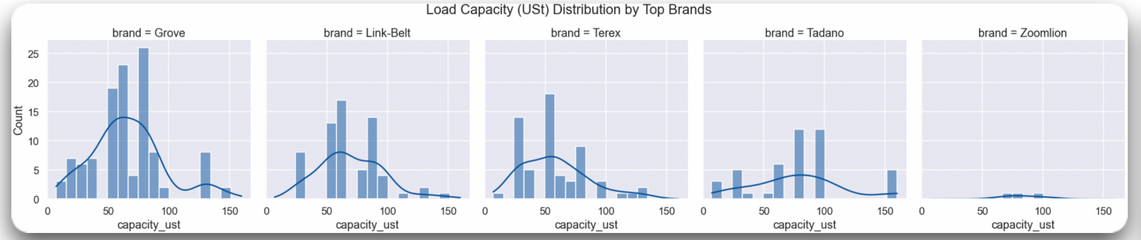


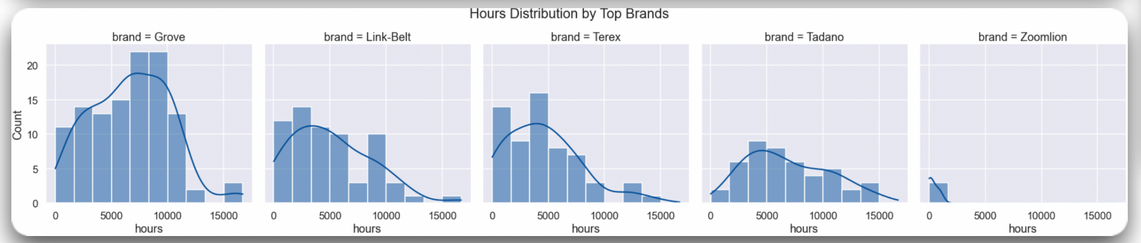


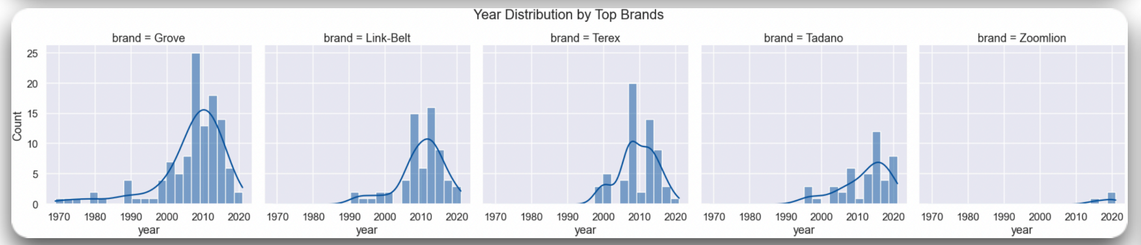


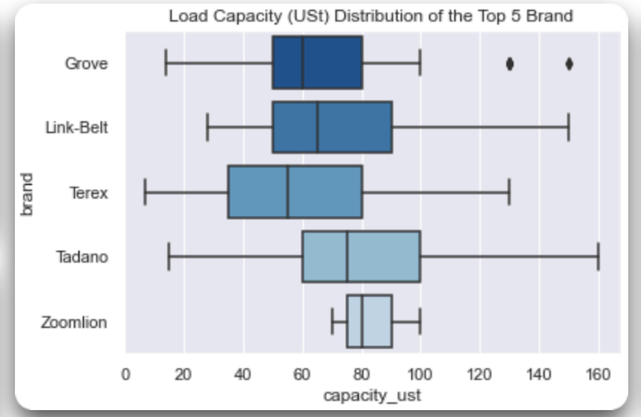
3. Análisis Exploratorio de Datos

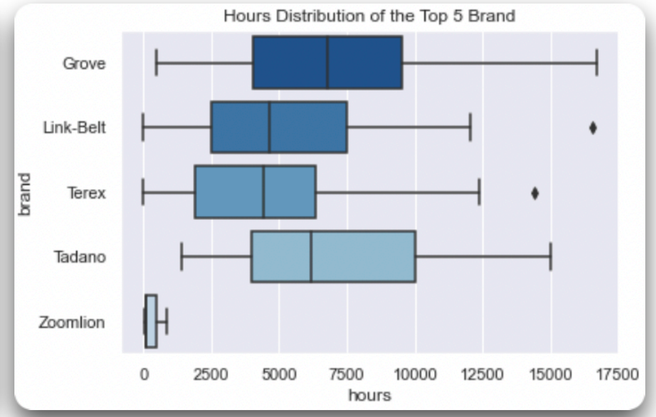
Sólo productos con precio disponible

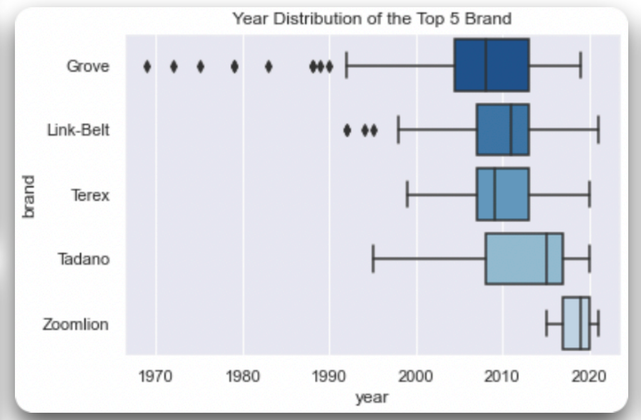






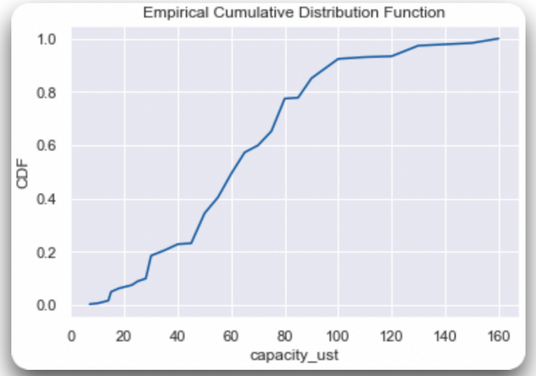






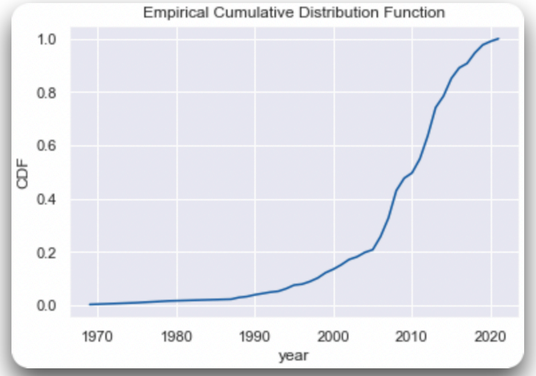
3. Análisis Exploratorio de Datos

Sólo productos con precio disponible



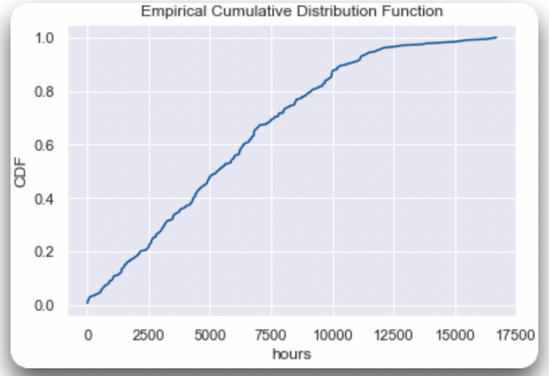
- En CraneMarket.com, la probabilidad de encontrar grúas con capacidad de carga menor a 20 USt es del 6.29%

```
P(x < 20USt):  6.29%
P(x < 40USt):  22.85%
P(x < 60USt):  49.34%
P(x < 80USt):  77.48%
P(x < 100USt):  92.38%
```



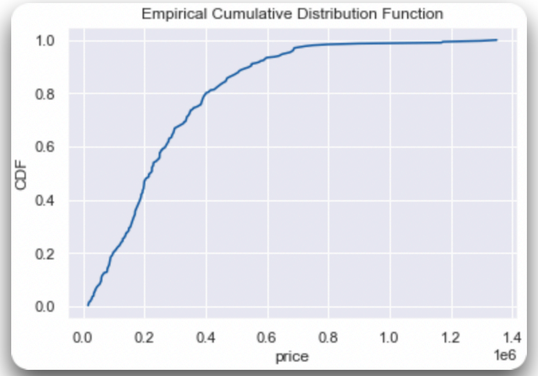
- En CraneMarket.com, la probabilidad de encontrar grúas fabricadas antes del año 2000 es del 13.58%

```
P(x < 2000):  13.58%
P(x < 2005):  20.86%
P(x < 2010):  49.67%
P(x < 2015):  85.10%
P(x < 2020):  99.01%
```



- En CraneMarket.com, la probabilidad de encontrar grúas con horómetro menor a las 1,000 horas es del 8.94%

```
P(x < 1,000horas):  8.94%
P(x < 3,000horas):  27.81%
P(x < 5,000horas):  48.01%
P(x < 7,000horas):  66.89%
P(x < 9,000horas):  79.14%
```

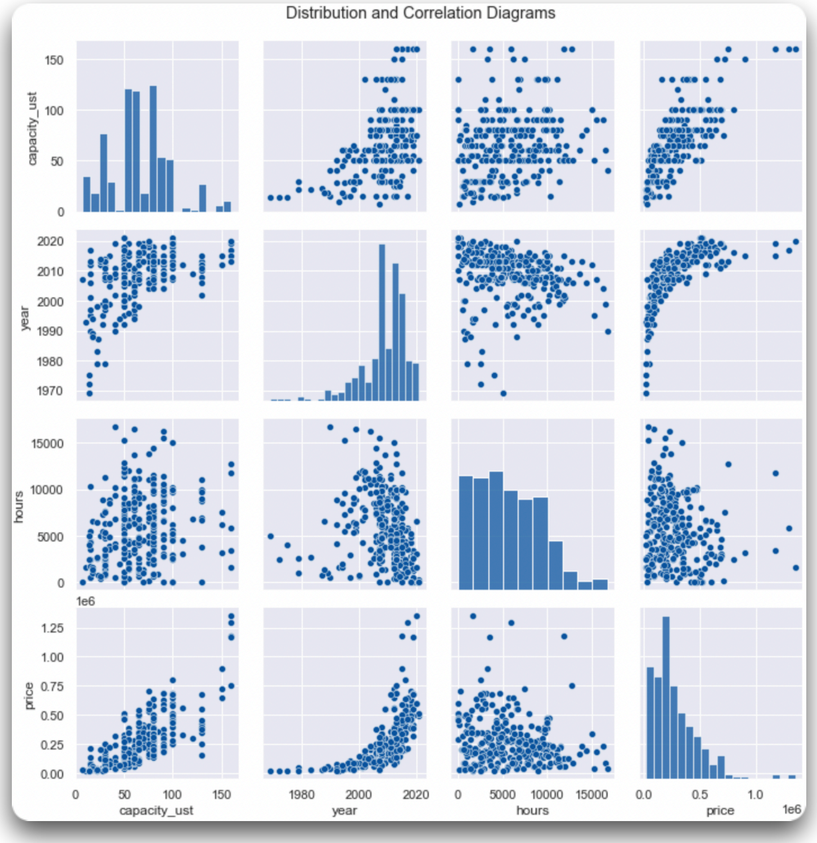


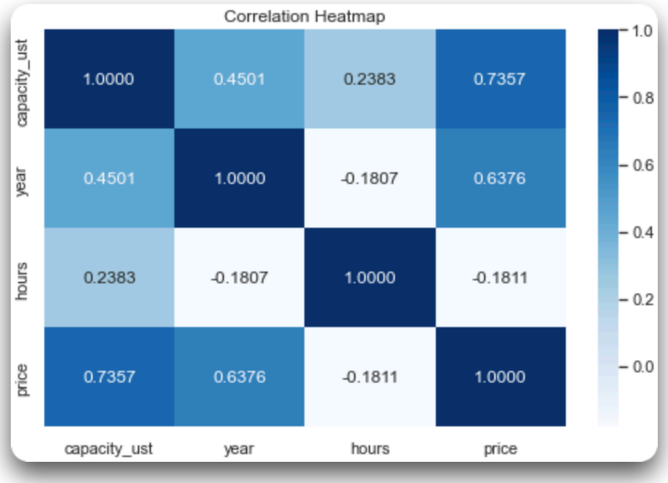
- En CraneMarket.com, la probabilidad de encontrar grúas con precio menor a los $100,000 es del 20.53%

```
P(x < $100,000):  20.53%
P(x < $200,000):  47.02%
P(x < $300,000):  66.89%
P(x < $400,000):  79.80%
P(x < $500,000):  87.42%
```

4. Predicción de precios

Modelo Regresión Lineal Simple





	capacity_ust	year	hours	price
count	302.000000	302.000000	302.000000	3.020000e+02
mean	65.807947	2008.668874	5729.086093	2.736200e+05
std	31.231579	8.222182	3695.562667	2.101391e+05
min	7.000000	1969.000000	5.000000	1.650000e+04
25%	50.000000	2006.000000	2715.000000	1.300000e+05
50%	65.000000	2011.000000	5350.000000	2.217500e+05
75%	80.000000	2014.000000	8459.750000	3.735000e+05
max	160.000000	2021.000000	16700.000000	1.350000e+06

Es muy frecuente pensar que para predecir el precio de un producto basta con aplicar una regresión lineal simple.

Sin embargo, en el mercado de bienes de capital, seguros, retail e inmobiliario los datos no siguen una distribución normal.

En estas situaciones, se hace presente la 'Heterocedasticidad'. Este es un comportamiento en una regresión lineal indicando que la varianza de los errores no es constante en todas las observaciones realizadas.

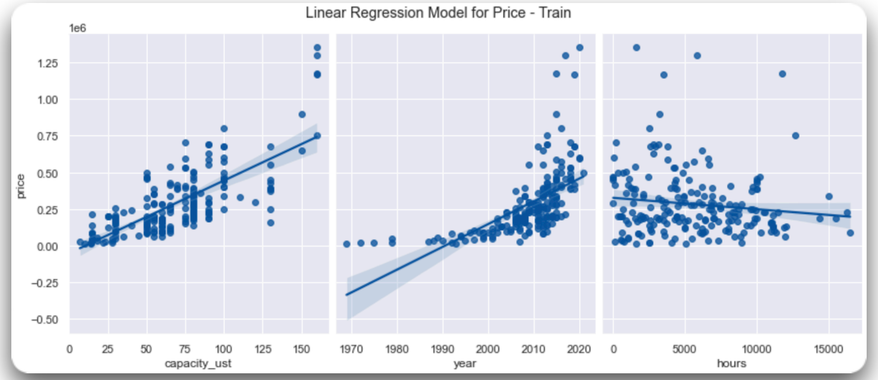
Su presencia significa que aplicar una regresión lineal como modelo predictivo en nuestros datos sería una pésima idea.

En este caso, debemos comprobar estadísticamente si es que la varianza es constante o no a lo largo de las variables explicativas (capacity_ust, year, hours).

De no ser constante, debemos aplicar Modelos Lineales Generalizados.

4. Predicción de precios

Modelo Regresión Lineal Simple



OLS Regression Results

Dep. Variable:	price	R-squared:	0.541
Model:	OLS	Adj. R-squared:	0.539
Method:	Least Squares	F-statistic:	246.6
Date:	Sun, 31 Jul 2022	Prob (F-statistic):	3.23e-37
Time:	00:36:59	Log-Likelihood:	-2811.8
No. Observations:	211	AIC:	5628.
Df Residuals:	209	BIC:	5634.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-5.197e+04	2.37e+04	-2.196	0.029	-9.86e+04	-5316.152
capacity_ust	4957.3634	315.703	15.703	0.000	4334.992	5579.735

OLS Regression Results

Dep. Variable:	price	R-squared:	0.355
Model:	OLS	Adj. R-squared:	0.352
Method:	Least Squares	F-statistic:	115.0
Date:	Sun, 31 Jul 2022	Prob (F-statistic):	1.16e-21
Time:	00:36:59	Log-Likelihood:	-2847.8
No. Observations:	211	AIC:	5700.
Df Residuals:	209	BIC:	5706.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-3.088e+07	2.91e+06	-10.627	0.000	-3.66e+07	-2.52e+07
year	1.551e+04	1446.656	10.725	0.000	1.27e+04	1.84e+04

OLS Regression Results

Dep. Variable:	price	R-squared:	0.017
Model:	OLS	Adj. R-squared:	0.013
Method:	Least Squares	F-statistic:	3.659
Date:	Sun, 31 Jul 2022	Prob (F-statistic):	0.0571
Time:	00:36:59	Log-Likelihood:	-2892.2
No. Observations:	211	AIC:	5788.
Df Residuals:	209	BIC:	5795.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	3.261e+05	2.71e+04	12.021	0.000	2.73e+05	3.8e+05
hours	-7.7825	4.068	-1.913	0.057	-15.803	0.238

		MAE	RMSE	Bias
Linear Regression Model - price ~ capacity_ust (%)				
	Train	39.42	52.43	-0.00
	Test	39.48	50.15	-0.84

		MAE	RMSE	Bias
Linear Regression Model - price ~ year (%)				
	Train	41.21	62.17	-0.00
	Test	39.77	48.96	-8.67

		MAE	RMSE	Bias
Linear Regression Model - price ~ hours (%)				
	Train	54.26	76.74	0.00
	Test	59.41	71.49	-10.45

Un modelo predictivo es adecuado si:

- **R-squared cercano a 100%:** Significa un ajuste lineal perfecto, el modelo explica la variación total de la variable Y.
- **AIC mínimo:** indica que el modelo requiere menos información para predecir, penalizando el sobreajuste.
- **Std-error bajo:** Mide con qué precisión el modelo estima el valor desconocido del coeficiente.
- **P>[t] menor a 0.05:** indica que esa variable explicativa es una adición significativa al modelo porque los cambios en su valor están relacionados con cambios en la variable respuesta.

Sin embargo, al aplicar por separado una regresión lineal en cada variable explicativa no cumplen con las métricas estadísticas mencionadas arriba.

De esta manera, se confirma que una regresión lineal no es un modelo predictivo adecuado para nuestros datos.

4. Predicción de precios

Modelo Regresión Lineal Simple

```python
# capacity_ust vs price
from scipy.stats import bartlett

a = df_clean[(df_clean['capacity_ust'] >= 7) & (df_clean['capacity_ust'] < 37)]['price']
b = df_clean[(df_clean['capacity_ust'] >= 37) & (df_clean['capacity_ust'] < 67)]['price']
c = df_clean[(df_clean['capacity_ust'] >= 67) & (df_clean['capacity_ust'] < 97)]['price']
d = df_clean[(df_clean['capacity_ust'] >= 97) & (df_clean['capacity_ust'] < 127)]['price']
e = df_clean[(df_clean['capacity_ust'] >= 127)]['price']

stat, p = bartlett(a, b, c, d, e)
print([np.var(x, ddof=1) for x in [a, b, c, d, e]])
print('p-value: ',p)

# Since the p-value is less than 0.05, I will not fail to reject the null hypothesis.
# In other words, I have sufficient evidence to say that the groups have different variances.

# By having a statistical difference in the variance, we can be totally sure that
# a linear regression 'capacity_ust-price' will not be an acceptable predictive model.
```

✓ 0.6s

```
[5159522187.995771, 14742554661.742178, 22119434372.384964, 21891684733.333332, 129996041838.48419]
p-value:  2.2578987398717056e-22
```

```python
# year vs price
from scipy.stats import levene

f = df_clean[(df_clean['year'] >= 1969) & (df_clean['year'] < 1979)]['price']
g = df_clean[(df_clean['year'] >= 1979) & (df_clean['year'] < 1989)]['price']
h = df_clean[(df_clean['year'] >= 1989) & (df_clean['year'] < 1999)]['price']
i = df_clean[(df_clean['year'] >= 1999) & (df_clean['year'] < 2009)]['price']
j = df_clean[(df_clean['year'] >= 2009)]['price']

stat, p = levene(f, g, h, i, j)
print([np.var(x, ddof=1) for x in [f, g, h, i, j]])
print('p-value: ',p)

# Since the p-value is less than 0.05, I will not fail to reject the null hypothesis.
# In other words, I have sufficient evidence to say that the groups have different variances.

# By having a statistical difference in the variance, we can be totally sure that
# a linear regression 'year-price' will not be an acceptable predictive model.
```

✓ 0.3s

```
[1333333.3333333333, 71377666.66666666, 547421233.7662338, 6076483664.708308, 46168596943.0238]
p-value:  3.393113922381394e-11
```

```python
# hours vs price
from scipy.stats import bartlett

k = df_clean[(df_clean['hours'] >= 5) & (df_clean['hours'] < 3344)]['price']
l = df_clean[(df_clean['hours'] >= 3344) & (df_clean['hours'] < 6683)]['price']
m = df_clean[(df_clean['hours'] >= 6683) & (df_clean['hours'] < 10022)]['price']
n = df_clean[(df_clean['hours'] >= 10022) & (df_clean['hours'] < 13361)]['price']
o = df_clean[(df_clean['hours'] >= 13361)]['price']

stat, p = bartlett(k, l, m, n, o)
print([np.var(x, ddof=1) for x in [k, l, m, n, o]])
print('p-value: ',p)

# Since the p-value is less than 0.05, I will not fail to reject the null hypothesis.
# In other words, I have sufficient evidence to say that the groups have different variances.

# By having a statistical difference in the variance, we can be totally sure that
# a linear regression 'hours-price' will not be an acceptable predictive model.
```

✓ 0.5s

```
[57218143387.34984, 49660737929.974205, 16381859744.958988, 53762159599.56651, 9643381944.444445]
p-value:  1.4750781304890385e-07
```

Se debe comprobar estadísticamente la presencia de 'Heterocedasticidad' en los diagramas de dispersión:

- Price ~ capacity_ust
- Price ~ year
- Price ~ hours

Para ello, se aplica el test de Barlett y Levene para ver si hay varianzas iguales en 05 grupos (muestras) de igual tamaño a lo largo de los valores de cada variable explicativa.

H0: La varianza entre cada grupo es igual.
H1: Al menos un grupo tiene una varianza que no es igual al resto.

Si el valor p correspondiente de la estadística de prueba es menor que algún nivel de significación (α = 0.5), entonces podemos rechazar la hipótesis nula y concluir que no todos los grupos tienen la misma varianza.

Para muestras de poblaciones aparentemente normales, la prueba de Barlett es la más adecuada.

Para muestras de poblaciones significativamente no normales, la prueba de Levene es más robusta.

En las 03 variables explicativas, se tiene un p valor muy por debajo de 0.05.

Se concluye así que en nuestros datos existe presencia de 'Heterocedasticidad' el cual es un grave problema para los pronósticos.

La solución a este fenómeno es la aplicación de Modelos Lineales Generalizados (GLM).

5. Predicción de precios

Modelos Lineales Generalizados (GLM): Teoría

Distribution	Domain	$\mu = E[Y\|x]$	$v(\mu)$	$\theta(\mu)$	$b(\theta)$	ϕ
Binomial $B(n,p)$	$0,1,\ldots,n$	np	$\mu - \frac{\mu^2}{n}$	$\log \frac{p}{1-p}$	$n\log(1+e^\theta)$	1
Poisson $P(\mu)$	$0,1,\ldots,\infty$	μ	μ	$\log(\mu)$	e^θ	1
Neg. Binom. $NB(\mu,\alpha)$	$0,1,\ldots,\infty$	μ	$\mu+\alpha\mu^2$	$\log(\frac{\alpha\mu}{1+\alpha\mu})$	$-\frac{1}{\alpha}\log(1-\alpha e^\theta)$	1
Gaussian/Normal $N(\mu,\sigma^2)$	$(-\infty,\infty)$	μ	1	μ	$\frac{1}{2}\theta^2$	σ^2
Gamma $N(\mu,\nu)$	$(0,\infty)$	μ	μ^2	$-\frac{1}{\mu}$	$-\log(-\theta)$	$\frac{1}{\nu}$
Inv. Gauss. $IG(\mu,\sigma^2)$	$(0,\infty)$	μ	μ^3	$-\frac{1}{2\mu^2}$	$-\sqrt{-2\theta}$	σ^2
Tweedie $p \geq 1$	depends on p	μ	μ^p	$\frac{\mu^{1-p}}{1-p}$	$\frac{\alpha-1}{\alpha}\left(\frac{\theta}{\alpha-1}\right)^\alpha$	ϕ

https://www.statsmodels.org/dev/glm.html

Los GLM son una gran variedad de modelos de regresión. El supuesto en estos modelos es que la variable respuesta *yi* sigue una distribución dentro de la familia de distribuciones exponenciales con un promedio μi, donde se asume una función μiTβ que frecuentemente no es lineal. Para linealizar μi, es necesario usar un 'link function' para convertir la variable respuesta, *yi.*

Las distribuciones **Gamma** e **Inv. Gauss** son las adecuadas para predecir los valores de la variable respuesta (precio) ya que es **positiva, continua y sesgada.**

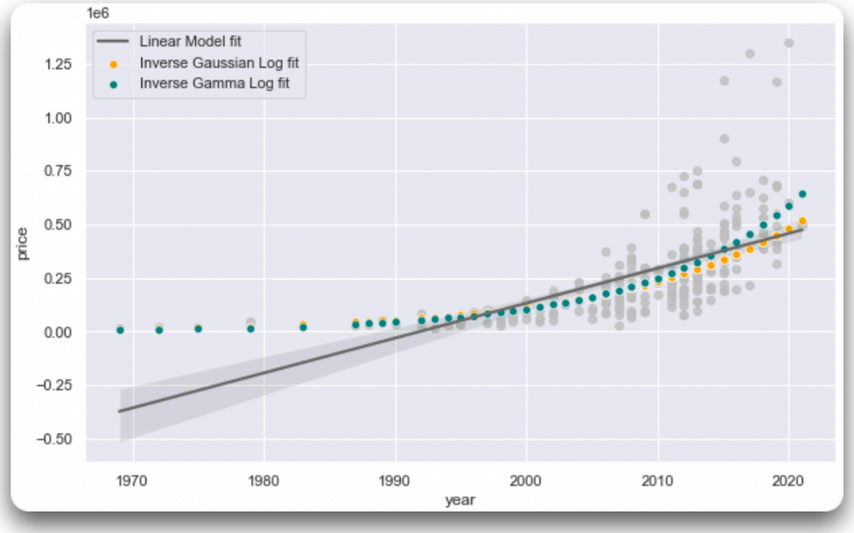
El objetivo de un 'link function' es transformar regresiones no lineales en regresiones lineales. En esta ocasión, emplearé un log link function para que los valores de predicción del precio no sigan una regresión μ, sino log(μ).

```python
# Price ~ year
from statsmodels.formula.api import glm
plt.figure(figsize=(8, 6))

# Inverse Gaussian - log
# Fit
glm_inversegauss_model = glm(formula = 'price ~ year', data=df_clean,
                             family=sm.families.InverseGaussian(link=sm.families.links.log())).fit()

# Gamma - log
# Fit
glm_gamma_model = glm(formula = 'price ~ year', data=df_clean,
                      family=sm.families.Gamma(link=sm.families.links.log())).fit()

df_clean['fit_price'] = glm_inversegauss_model.fittedvalues
df_clean['fit_price2'] = glm_gamma_model.fittedvalues
sns.regplot(x='year', y='price', data=df_clean, fit_reg=True, color='silver',
            line_kws={'color': 'dimgray', 'label': 'Linear Model fit'})
sns.scatterplot(x='year', y='fit_price', data=df_clean, color='orange', label='Inverse Gaussian Log fit')
sns.scatterplot(x='year', y='fit_price2', data=df_clean, color='teal', label='Inverse Gamma Log fit')
plt.show()
```



Para fines demostrativos, se aplicará una regresión lineal simple, un GLM Inverse Gaussian log y GLM Gamma log con la única variable explicativa 'year'. Es decir, predecir 'Price' en función a 'year'.

Esto es porque esta relación tiene una clara distribución exponencial, y demostrar que los GLM pueden reflejar mucho mejor esta distribución.

Como era de esperarse, una distribución Gamma e Inv. Gauss con 'log link function' se ajusta mejor a la variable respuesta (precio) en función a la variable explicativa 'year'.

5. Predicción de precios

Modelos Lineales Generalizados (GLM): Selección

OLS Regression Results

Dep. Variable:	price	R-squared:	0.756
Model:	OLS	Adj. R-squared:	0.753
Method:	Least Squares	F-statistic:	213.9
Date:	Mon, 01 Aug 2022	Prob (F-statistic):	3.77e-63
Time:	20:44:11	Log-Likelihood:	-2750.1
No. Observations:	211	AIC:	5508.
Df Residuals:	207	BIC:	5522.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1.644e+07	2.54e+06	-6.467	0.000	-2.14e+07	-1.14e+07
capacity_ust	4995.9477	289.258	17.272	0.000	4425.679	5566.217
year	8203.9041	1267.696	6.472	0.000	5704.653	1.07e+04
hours	-15.8891	2.250	-7.062	0.000	-20.325	-11.453

Omnibus:	52.705	Durbin-Watson:	2.007
Prob(Omnibus):	0.000	Jarque-Bera (JB):	143.099
Skew:	1.067	Prob(JB):	8.44e-32
Kurtosis:	6.423	Cond. No.	2.38e+06

...

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

[2] The condition number is large, 2.38e+06. This might indicate that there are strong multicollinearity or other numerical problems.

	MAE	RMSE	Bias
Linear Regression Model - all (%)			
Train	28.38	39.66	0.00
Test	32.63	42.34	-6.41

Modelo 1: Regresión Lineal Simple

'Price ~ capacity_ust + year+ hours'

Generalized Linear Model Regression Results

Dep. Variable:	price	No. Observations:	211
Model:	GLM	Df Residuals:	207
Model Family:	InverseGaussian	Df Model:	3
Link Function:	log	Scale:	6.5586e-07
Method:	IRLS	Log-Likelihood:	-2689.0
Date:	Sun, 31 Jul 2022	Deviance:	0.00016819
Time:	22:44:15	Pearson chi2:	0.000136
No. Iterations:	18	Pseudo R-squ. (CS):	0.9901
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-92.9170	4.792	-19.389	0.000	-102.310	-83.524
capacity_ust	0.0176	0.001	16.302	0.000	0.015	0.020
year	0.0519	0.002	21.592	0.000	0.047	0.057
hours	-4.033e-05	5.71e-06	-7.063	0.000	-5.15e-05	-2.91e-05

	MAE	RMSE	Bias
Inverse Gaussian log - all (%)			
Train	26.49	49.69	-5.75
Test	28.60	48.64	-9.24

AIC: 5386.046146025545

Modelo 2: GLM Inverse Gaussian log

'Price ~ capacity_ust + year + hours'

Generalized Linear Model Regression Results

Dep. Variable:	price	No. Observations:	211
Model:	GLM	Df Residuals:	207
Model Family:	Gamma	Df Model:	3
Link Function:	log	Scale:	0.082494
Method:	IRLS	Log-Likelihood:	-2617.0
Date:	Sun, 31 Jul 2022	Deviance:	17.422
Time:	22:44:15	Pearson chi2:	17.1
No. Iterations:	14	Pseudo R-squ. (CS):	0.9987
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-116.6181	6.530	-17.858	0.000	-129.417	-103.819
capacity_ust	0.0142	0.001	19.073	0.000	0.013	0.016
year	0.0638	0.003	19.586	0.000	0.057	0.070
hours	-3.255e-05	5.78e-06	-5.632	0.000	-4.39e-05	-2.12e-05

	MAE	RMSE	Bias
Gamma log - all (%)			
Train	21.09	33.56	-1.40
Test	23.25	37.00	-6.69

AIC: 5242.022103719784

Modelo 3: GLM Gamma log

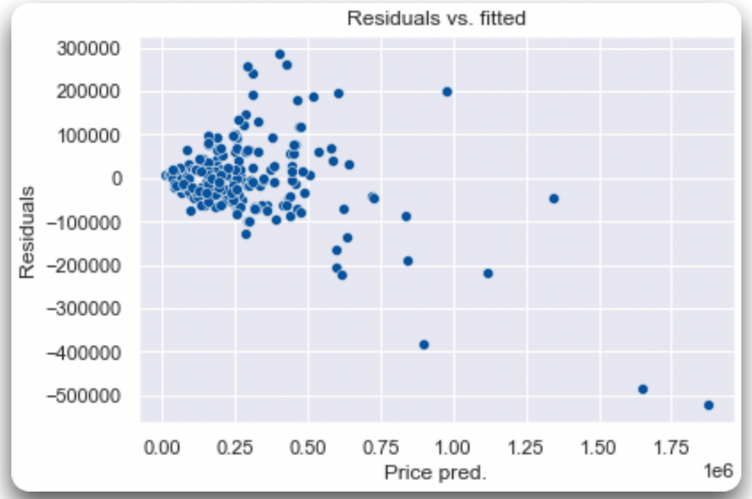
'Price ~ capacity_ust + year + hours'

Se tiene total seguridad en descartar por completo un modelo lineal con varias variables ya que presenta alto niveles de AIC, 'str_errr', % MAE y % RMSE.

De los 02 GLM, Gamma log es el modelo que tiene las mejores métricas estadísticas con el mayor Pseudo R-squ, niveles razonables de 'std_err', menor % MAE, % RMSE, y AIC.

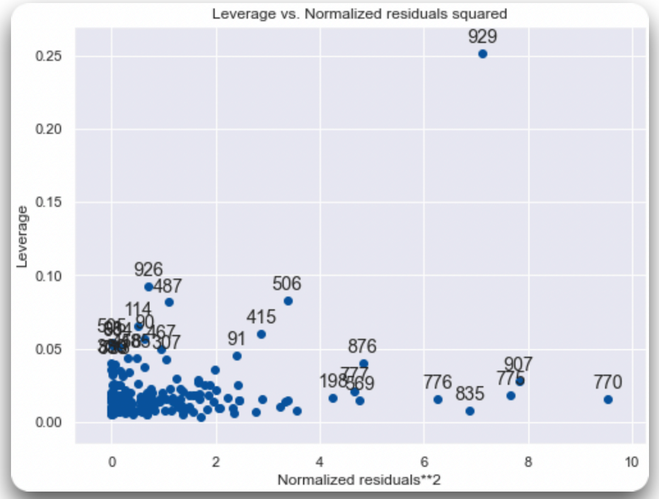
5. Predicción de precios

Modelos Lineales Generalizados (GLM): Diagnóstico



Los residuales no están distribuidos aleatoriamente alrededor de la media cero. La gran mayoría de los residuales no superan los $100,000.

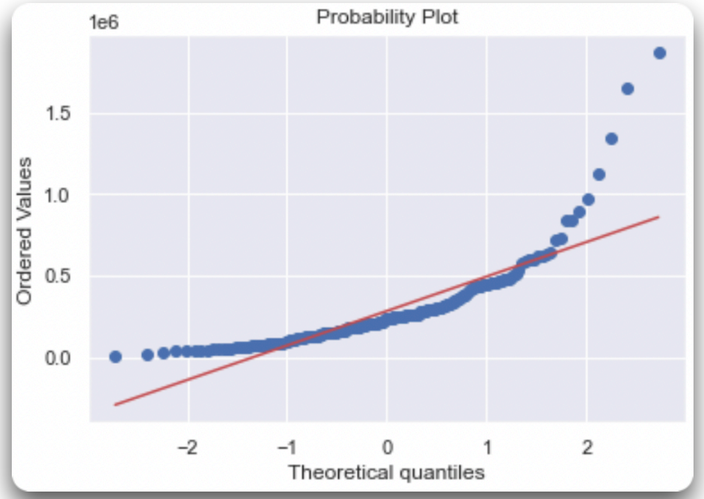
El gráfico se ve demasiado agrupado, sin embargo, puede ser engañoso ya que hay una cierta cantidad de productos con precios extremadamente bajos o altos con respecto al modelo.



Este gráfico nos permite determinar los puntos más influyentes sobre el modelo predictivo aplicado.

En el train dataframe se puede identificar los 20 productos más influyentes sobre el modelo y analizar las causas de ello. Esto puede deberse a la capacidad, año de fabricación o quizás a su configuración técnica (estructuras, componentes o accesorios adicionales).

Quizás, eliminando los datos influyentes en el train dataframe, se puede tener mejores resultados gráficos, sin embargo, no sería justo al momento de comparar con otros GLM y correría el riesgo de sobreajustar el modelo.



Probability Plot nos permite corroborar si es que el GLM aplicado es el adecuado, siempre y cuando los puntos azules siguen la tendencia y están lo más cercano posible a la línea roja.

En el eje X, son los cuantiles de una distribución normal teórica . En el eje Y, son los cuantiles de los residuales.

5. Predicción de precios

Modelos Lineales Generalizados (GLM): Resultados

```python
# Predict new data

# Define your values
my_values = [
            {'capacity_ust': 80, 'year': 2013, 'hours': 5000},
            {'capacity_ust': 80, 'year': 2015, 'hours': 5136},
            {'capacity_ust': 50, 'year': 2012, 'hours': 8740},
            {'capacity_ust': 80, 'year': 2013, 'hours': 10433},
            {'capacity_ust': 130, 'year': 2004, 'hours': 700},
            {'capacity_ust': 130, 'year': 2004, 'hours': 40},
            {'capacity_ust': 30, 'year': 2011, 'hours': 1364},
            {'capacity_ust': 80, 'year': 2012, 'hours': 4618},
            {'capacity_ust': 80, 'year': 2006, 'hours': 617},
            {'capacity_ust': 99, 'year': 1994, 'hours': 24362},
            {'capacity_ust': 99, 'year': 1993, 'hours': 25863},
            {'capacity_ust': 80, 'year': 2006, 'hours': 617},
            {'capacity_ust': 110, 'year': 2000, 'hours': 9729},
            {'capacity_ust': 45, 'year': 2011, 'hours': 4379},
            {'capacity_ust': 130, 'year': 2012, 'hours': 2828},
            {'capacity_ust': 45, 'year': 2011, 'hours': 2036},
            {'capacity_ust': 80, 'year': 2014, 'hours': 1940},
            {'capacity_ust': 130, 'year': 2008, 'hours': 8468},
]
```

1. Construir un diccionario con datos de productos de la empresa con el objetivo de predecir los precios.

```python
newdata = pd.DataFrame(columns=['capacity_ust',
                                'year',
                                'hours'])

for i in range(0, len(my_values)):
    currentItem = my_values[i]
    newdata.loc[i] = [my_values[i]['capacity_ust'],
                      my_values[i]['year'],
                      my_values[i]['hours']]
newdata

# Predict your values with the best model
y_test_pred = glm_gamma_model.predict(newdata)
price_predict = pd.DataFrame(y_test_pred)

# Define +/- 20% of price predict
price_min = price_predict*0.80
price_max = price_predict*1.20
```

2. Convertir el diccionario en un DataFrame,
3. Introducir el DataFrame en el GLM Gamma log
4. Agregar resultados en el DataFrame
5. Definir (arbitrariamente) un +/- 20% para las predicciones

	capacity_ust	year	hours	price_predict	price_min	price_max
0	80	2013	5000	345057.98	276046.38	414069.57
1	80	2015	5136	390276.25	312221.00	468331.50
2	50	2012	8740	187354.03	149883.23	224824.84
3	80	2013	10433	289125.70	231300.56	346950.84
4	130	2004	700	454075.84	363260.67	544891.01
5	130	2004	40	463936.67	371149.33	556724.00
6	30	2011	1364	168318.16	134654.53	201981.79
7	80	2012	4618	327786.46	262229.17	393343.75
8	80	2006	617	254650.46	203720.37	305580.56
9	99	1994	24362	71581.09	57264.88	85897.31
10	99	1993	25863	63955.46	51164.37	76746.55
11	80	2006	617	254650.46	203720.37	305580.56
12	110	2000	9729	197504.81	158003.85	237005.78
13	45	2011	4379	188728.04	150982.43	226473.64
14	130	2012	2828	705758.10	564606.48	846909.72
15	45	2011	2036	203684.98	162947.99	244421.98
16	80	2014	1940	406304.93	325043.95	487565.92
17	130	2008	8468	455112.59	364090.07	546135.11

6. Se exporta el DataFrame en archivo Excel

6. Conclusiones

Puntos claves

- Web Scraping permite extraer datos de miles de productos en cuestión de minutos. Para lograr esto, se debe tener nociones de lenguaje HTML/CSS y un nivel avanzado de Python. Sin embargo, su aplicación será viable siempre y estemos al tanto de cualquier modificación en la estructura HTML/CSS y que las páginas web permitan ejecutar nuestros scrapers.

- Para solucionar esta problemática, los desarrolladores web de estas páginas podrían desarrollar APIs para facilitar la conectividad y simplicidad en la extracción de datos, o que contratemos microservicios que permitan ejecutar nuestros scrapers en estas páginas.

- En este proyecto, se buscó extraer datos de 07 páginas web. Sin embargo, sólo se tuvo acceso a cranemarket.com

- En este tipo de proyectos, es crucial tener productos con precios disponibles para evitar sobreajustes al momento de entrenar nuestros modelos predictivos o generar conclusiones sesgadas del mercado.

- Por temas de un tamaño limitado de datos, interpretación y visualización de efectividad de Modelos Lineales Generalizados sobre Modelo Regresión Simple, sólo se aplicó la paquetería de stats.models y no la de scikit-learn.

- No descarto un posterior estudio sobre la aplicación de modelos de machine learning scikit-learn y la integración con MAPIE para estimar los intervalos de predicción de precios.

6. Conclusiones

Ideas

- Así como se analizó la data del mercado de Rough Terrain Cranes, se puede analizar la data de otro tipo de líneas de grúas, o maquinarias livianas o pesadas en el mercado internacional o nacional.

- Predecir el precio de laptops nuevas en el mercado retail online peruano en función al nivel de procesador, cantidad de memoria RAM, cantidad de SSD GB, tamaño de pantalla, etc.

- Predecir el precio de venta y alquiler de viviendas en la ciudad de Lima en función al m2, cantidad de habitaciones, cantidad de baños, distancia a centros comerciales, etc.

- Predecir el precio de un automóvil usado en el mercado peruano en función al tipo de transmisión, cilindrada, año de fabricación, kilometraje, etc.