data-discovery

The purpose of this notebook will be to answer several important questions about our data set, which must be answered before we start build any models.

Those questions include:

- What assumptions can we make about the data?
- What is the quality of the data, and if the quality is low, what can we do about it?
- What are the data types?
- What does the data look and feel like?
- Should any features be removed before they are ingested into our models into our models?

The data used in this notebook has been altered slightly from the raw .json.gz's that we are using as our absolute source of truth. There are no actual transformations to the data, just an unzipping and repacking of the data to make it more accessible. The transformation is done within `_00_unzip_and_repackage.py` .

install dependancies

In [16]:
```
import pandas as pd
import polars as pl
import seaborn as sns
import matplotlib.pyplot as plt
from IPython.display import display
```

read in the data

Let's read in our data from the creditcard.parquet file in the data folder. The original .csv has been reformatted to specify data types and reduce the size of the file. The code that was used to do this is included in the scripts folder.

In [31]:
```
# read in data from parquet
df = pl.read_parquet('../data/00_products_unzipped.parquet')

# display with styling
display(df.to_pandas().head(3).style.set_properties(
    **{'max-width': '75px', 'text-overflow': 'ellipsis'})
)
```

	Category	Price	Title	Features	Description	Details	SKU	Manufacturer
0	Automotive	nan	Bosal 813-767 Exhaust Pipe	['Package Dimensions: 5.0 H x 44.0 L x 8.0 W (inches)' 'Easy Installation' 'Package Weight: 30.0 pounds' 'Country of Origin : United States']	['One piece, direct fit exhaust system component that saves installation time by eliminating the need for labor intensive bending and welding required with split system exhaust components.']	{'Manufacturer': 'Bosal', 'Brand': 'Bosal', 'Item Weight': '6.85 pounds', 'Product Dimensions': '44"L x 8"W', 'Item model number': '813-767', 'Manufacturer Part Number': '813-767', 'Best Sellers Rank': {'Automotive': 2432472.0, 'Steering Wheel Covers': None, 'Steering Wheel Accessories': None, 'Automotive Exhaust Systems & Parts': 4897.0, 'Automotive Replacement Exhaust Pipes': 5406.0, 'Automotive Replacement Distributors': None, 'Audio & Video Connectors & Adapters': None, 'Sports & Outdoors': None, 'Yoga Starter Sets': None, 'Camera Cases': None, 'Patio, Lawn & Garden': None, 'Lawn Mower	813-767	Bosal

Category	Price	Title	Features	Description	Details	SKU	Manufacturer
					Belts': None, 'Tools & Home Improvement': None, 'Extension Cords': None, 'Powersports Clutches': None, 'Arts, Crafts & Sewing': None, 'Sewing Products': None, 'Powersports Batteries': None, 'External Hard Drives': None, 'Clothing, Shoes & Jewelry': None, 'Luggage Scales': None, 'Automotive Replacement Chassis Steering Knuckles': None, 'Lamp Shades': None, 'Home & Kitchen': None, 'Statues': None, 'Office Products': None, 'PBX Phones & Systems': None, 'Exterior Car Care Microfiber Cloths': None, "Kids' Bed Blankets": None, 'TV Wall & Ceiling Mounts':		

Category	Price	Title	Features	Description	Details	SKU	Manufacturer
					None, 'Electronics Mounts': None, 'Toys & Games': None, "Kids' Stickers": None, 'Kitchen & Dining': None, 'Kitchen Storage Accessories': None, 'Replacement Under-Sink Water Filters': None, 'Automotive Replacement Shock Lift Supports': None, 'Bumper Covers': None, 'Kitchen Rugs': None, 'Cell Phones & Accessories': None, 'Cell Phone Screen Protectors': None, "Women's Cold Weather Mittens": None, 'Climate Pledge Friendly': None, 'Computer Monitors': None, 'Climate Pledge Friendly: Computers': None, 'Decorative Signs & Plaques': None, 'Cell		

Category	Price	Title	Features	Description	Details	SKU	Manufacturer
					Phone Basic Cases': None, 'Fishing Equipment': None, 'Powersports Plows': None, 'Snowshoes': None, 'Sliding Door Hardware': None, 'Automotive Replacement Oxygen Sensors': None, 'Mouse Pads': None, 'Chopsticks': None, 'Patio Furniture Pillows': None, 'Chip & Dip Sets': None, 'Automotive Replacement Transmission Oil Pressure Sensors': None, 'Automotive Replacement Transmissions & Parts': None, 'Automotive Replacement Electrical Accessories': None, 'Automotive Replacement Engine Oil Drain Plugs': None, 'Throw Pillow Covers': None, 'Pool Rafts & Inflatable Ride-ons': None, 'Outdoor Décor':		

Category	Price	Title	Features	Description	Details	SKU	Manufacturer
					None, 'Internal USB Port Cards': None, "Kids' Paint With Water Kits": None}, 'Date First Available': 'December 15, 2007', 'Shape': 'Round', 'Installation Type': 'Bolt-On', 'Package Dimensions': None, 'Material Leather': None, '': None, 'Country of Origin': None, 'Color Name': None, 'Size': None, 'Color': None, 'Item Dimensions LxWxH': None, 'Included Components': None, 'Compatible Devices': None, 'Connector Gender': None, 'Unit Count': None, 'Indoor/ Outdoor Usage': None, 'Is Discontinued By Manufacturer': None, 'UPC': None, 'Item Package Dimensions L x W x H': None, 'Package		

Category	Price	Title	Features	Description	Details	SKU	Manufacturer
					Weight': None, 'Item Dimensions LxWxH': None, 'Brand Name': None, 'Material': None, 'Part Number': None, 'Model Year': None, 'Theme': None, 'Age Range (Description)': None, 'Package Type': None, 'Style': None, 'Closure Type': None, 'Standing screen display size': None, 'Form Factor': None, 'Shell Type': None, 'Other display features': None, 'Compatible Phone Models': None, 'Pattern': None, 'Model': None, 'Exterior': None, 'Number of Pieces': None, 'Mounting Type': None, 'Vehicle Service Type Scooter': None, 'Battery Cell Composition Lead Calcium': None, 'Hard Drive': None,		

Category	Price	Title	Features	Description	Details	SKU	Manufacturer
					'Hardware Platform': None, 'Flash Memory Size': None, 'Hard Drive Interface': None, 'Digital Storage Capacity': None, 'Hard Disk Interface': None, 'Connectivity Technology': None, 'Special Feature': None, 'Hard Disk Form Factor': None, 'Hard Disk Description': None, 'Item Package Quantity': None, 'Batteries Required?': None, 'Warranty Description': None, 'Domestic Shipping': None, 'International Shipping': None, 'Exterior Finish': None, 'Weight Limit': None, 'Department': None, 'Batteries': None, 'Number of Items': None, 'Top Width': None, 'Base Width': None, 'Special		

Category	Price	Title	Features	Description	Details	SKU	Manufacturer
					Features': None, 'Shade Color': None, 'Shade Material': None, 'Batteries Included?': None, 'Product Care Instructions': None, 'Collection Name': None, 'Assembly required': None, 'Number of pieces': None, 'Batteries required': None, 'Telephone Type': None, 'Answering System Type': None, 'Multiline Operation': None, 'Caller Identification': None, 'Fabric Type': None, 'Seasons': None, 'Fabric Warmth Description': None, 'Sport': None, 'Movement Type': None, 'TV Size': None, 'Minimum Compatible Size': None, 'Maximum Tilt Angle': None, 'Room Type': None, 'Cartoon Character': None, 'Reusability':		

Category	Price	Title	Features	Description	Details	SKU	Manufacturer
					None, 'Material Vinyl': None, 'Shape Irregular': None, 'Hair Type': None, 'Scent': None, 'Item Form': None, 'Occasion': None, 'Number of Sets': None, 'Corner Style': None, 'Weight': None, 'Duration': None, 'External Testing Certification': None, 'Product Benefits': None, 'Position': None, 'Vehicle Service Type': None, 'OEM Part Number': None, 'ABPA Partslink Number': None, 'Pile Height': None, 'Construction Type': None, 'Back Material Type': None, 'Is Stain Resistant': None, 'Rug Form Type': None, 'Screen Size': None, 'Screen Surface Description': None, 'Connector Type': None, 'Input		

Category	Price	Title	Features	Description	Details	SKU	Manufacturer
					Voltage': None, 'Wattage': None, 'Output Current': None, 'Specification Met': None, 'Output Voltage': None, 'Power Source': None, 'Suggested Users': None, 'Sport Type': None, 'Hand Orientation': None, 'Import': None, 'Frame Type': None, 'Voltage': None, 'Plug Format': None, 'Handle Material': None, 'Closure': None, 'Material Type': None, 'Refresh Rate': None, 'Aspect Ratio': None, 'Language': None, 'Care instructions': None, 'Manufacturer recommended age': None, 'Release date': None, 'Handle Type': None, 'Finish Type': None, 'Metal Type': None, 'Finish': None, 'Dishwasher compatible': None,		

Category	Price	Title	Features	Description	Details	SKU	Manufacturer
					'Target Species': None, 'Weave Type': None, 'Maximum Weight Recommendation': None, 'Frame Material': None, 'Number of Handles': None, 'Style Classic': None, 'Finish Type Powder Coated': None, 'Auto Part Position': None, 'Fill Material': None, 'Pillow Type': None, 'Item Firmness Description': None, 'Model Name': None, 'Thread Count': None, 'Item Display Dimensions': None, 'Number Of Pieces': None, 'Bowl Material': None, 'Is Dishwasher Safe': None, 'Is Microwaveable': None, 'Model number': None, 'Recommended Uses For Product': None, 'Battery Cell Composition': None,		

	Category	Price	Title	Features	Description	Details	SKU	Manufacturer
						'Wireless Type': None, 'Series': None, 'Operating System': None, 'Hardware Interface': None, 'Special features': None}		
1	Automotive	nan	NICEASY Bling Diamond Steering Wheel Cover for Women Girls,Fashionable,Elegant,Universal Fit 15 Inch,Leather Steering Wheel Accessory	['【Fashionable Design】 Fashion charming design of the diamond steering wheel cover, bling under the light, sparkling through windshield and shining your car.Simple Elegant Design style,Excellent Gift for Man Woman Girl Mom Wife' '【Eco-friendly & Healthy】 Our bling steering wheel covers are totally eco-friendly and healthy. PU leather with rhinestone bling design add fashion to your car interior, make your car personalized. Top diamond craftwork, more stable, will not hurt	['【Fashionable Design】 Fashion charming design of the diamond steering wheel cover, bling under the light, sparkling through windshield and shining your car.' '【Eco-friendly & Healthy】 Our bling steering wheel covers are totally eco-friendly and healthy. PU leather with rhinestone bling design add fashion to your car interior, make your car personalized. Top diamond craftwork, more stable, will not hurt your hands.' '【Universal Fit Most Steering Wheel 】 It will fit to any size of the	{'Manufacturer': 'Hi-Well', 'Brand': 'NICEASY', 'Item Weight': '1.5 pounds', 'Product Dimensions': None, 'Item model number': None, 'Manufacturer Part Number': 'NEYSWC1', 'Best Sellers Rank': {'Automotive': 1032865.0, 'Steering Wheel Covers': 3212.0, 'Steering Wheel Accessories': 4369.0, 'Automotive Exhaust Systems & Parts': None, 'Automotive Replacement Exhaust Pipes': None, 'Automotive Replacement Distributors': None, 'Audio & Video Connectors & Adapters': None, 'Sports &	NEYSWC1	Hi-Well

Category	Price	Title	Features	Description	Details	SKU	Manufacturer
			your hands.' '【Universal Fit Most Steering Wheel 】 It will fit to any size of the car steering wheel. Universal Fit 14.5-15 inch (37-38 cm) Steering Wheel' "【Easy Use】Don't worry about the installation even, because it is easy to operate, Any adult can easy install this steering wheel cover, no tools required." '【Absolutely 100% Customer Satisfaction Service 】If you are not satisfied with your purchase of Steering wheel cover , please feel free to contact us anytime and anywhere.']	car steering wheel. Universal Fit 14.5-15 inch (37-38 cm) Steering Wheel' "【Easy Use】Don't worry about the installation even, because it is easy to operate, Any adult can easy install this steering wheel cover, no tools required." '【Absolutely 100% Customer Satisfaction Service 】If you are not satisfied with your purchase of Steering wheel cover , please feel free to contact us anytime and anywhere.']	Outdoors': None, 'Yoga Starter Sets': None, 'Camera Cases': None, 'Patio, Lawn & Garden': None, 'Lawn Mower Belts': None, 'Tools & Home Improvement': None, 'Extension Cords': None, 'Powersports Clutches': None, 'Arts, Crafts & Sewing': None, 'Sewing Products': None, 'Powersports Batteries': None, 'External Hard Drives': None, 'Clothing, Shoes & Jewelry': None, 'Luggage Scales': None, 'Automotive Replacement Chassis Steering Knuckles': None, 'Lamp Shades': None, 'Home & Kitchen': None, 'Statues': None, 'Office Products': None, 'PBX Phones & Systems': None,		

Category	Price	Title	Features	Description	Details	SKU	Manufacturer
					'Exterior Car Care Microfiber Cloths': None, "Kids' Bed Blankets": None, 'TV Wall & Ceiling Mounts': None, 'Electronics Mounts': None, 'Toys & Games': None, "Kids' Stickers": None, 'Kitchen & Dining': None, 'Kitchen Storage Accessories': None, 'Replacement Under-Sink Water Filters': None, 'Automotive Replacement Shock Lift Supports': None, 'Bumper Covers': None, 'Kitchen Rugs': None, 'Cell Phones & Accessories': None, 'Cell Phone Screen Protectors': None, "Women's Cold Weather Mittens": None, 'Climate Pledge Friendly': None, 'Computer		

Category	Price	Title	Features	Description	Details	SKU	Manufacturer
					Monitors': None, 'Climate Pledge Friendly: Computers': None, 'Decorative Signs & Plaques': None, 'Cell Phone Basic Cases': None, 'Fishing Equipment': None, 'Powersports Plows': None, 'Snowshoes': None, 'Sliding Door Hardware': None, 'Automotive Replacement Oxygen Sensors': None, 'Mouse Pads': None, 'Chopsticks': None, 'Patio Furniture Pillows': None, 'Chip & Dip Sets': None, 'Automotive Replacement Transmission Oil Pressure Sensors': None, 'Automotive Replacement Transmissions & Parts': None, 'Automotive Replacement Electrical Accessories': None, 'Automotive Replacement Engine Oil Drain Plugs':		

Category	Price	Title	Features	Description	Details	SKU	Manufacturer
					None, 'Throw Pillow Covers': None, 'Pool Rafts & Inflatable Ride-ons': None, 'Outdoor Décor': None, 'Internal USB Port Cards': None, "Kids' Paint With Water Kits": None}, 'Date First Available': 'June 27, 2020', 'Shape': None, 'Installation Type': None, 'Package Dimensions': '15.8 x 15.8 x 1.8 inches', 'Material Leather': '', '': 'Vehicle Service Type Car', 'Country of Origin': None, 'Color Name': None, 'Size': None, 'Color': None, 'Item Dimensions LxWxH': None, 'Included Components': None, 'Compatible Devices': None, 'Connector Gender': None, 'Unit Count': None, 'Indoor/ Outdoor		

Category	Price	Title	Features	Description	Details	SKU	Manufacturer
					Usage': None, 'Is Discontinued By Manufacturer': None, 'UPC': None, 'Item Package Dimensions L x W x H': None, 'Package Weight': None, 'Item Dimensions LxWxH': None, 'Brand Name': None, 'Material': None, 'Part Number': None, 'Model Year': None, 'Theme': None, 'Age Range (Description)': None, 'Package Type': None, 'Style': None, 'Closure Type': None, 'Standing screen display size': None, 'Form Factor': None, 'Shell Type': None, 'Other display features': None, 'Compatible Phone Models': None, 'Pattern': None, 'Model': None, 'Exterior': None, 'Number of Pieces': None,		

Category	Price	Title	Features	Description	Details	SKU	Manufacturer
					'Mounting Type': None, 'Vehicle Service Type Scooter': None, 'Battery Cell Composition Lead Calcium': None, 'Hard Drive': None, 'Hardware Platform': None, 'Flash Memory Size': None, 'Hard Drive Interface': None, 'Digital Storage Capacity': None, 'Hard Disk Interface': None, 'Connectivity Technology': None, 'Special Feature': None, 'Hard Disk Form Factor': None, 'Hard Disk Description': None, 'Item Package Quantity': None, 'Batteries Required?': None, 'Warranty Description': None, 'Domestic Shipping': None, 'International Shipping': None, 'Exterior Finish': None, 'Weight Limit': None,		

Category	Price	Title	Features	Description	Details	SKU	Manufacturer
					'Department': None, 'Batteries': None, 'Number of Items': None, 'Top Width': None, 'Base Width': None, 'Special Features': None, 'Shade Color': None, 'Shade Material': None, 'Batteries Included?': None, 'Product Care Instructions': None, 'Collection Name': None, 'Assembly required': None, 'Number of pieces': None, 'Batteries required': None, 'Telephone Type': None, 'Answering System Type': None, 'Multiline Operation': None, 'Caller Identification': None, 'Fabric Type': None, 'Seasons': None, 'Fabric Warmth Description': None, 'Sport': None, 'Movement Type': None, 'TV Size':		

Category	Price	Title	Features	Description	Details	SKU	Manufacturer
					None, 'Minimum Compatible Size': None, 'Maximum Tilt Angle': None, 'Room Type': None, 'Cartoon Character': None, 'Reusability': None, 'Material Vinyl': None, 'Shape Irregular': None, 'Hair Type': None, 'Scent': None, 'Item Form': None, 'Occasion': None, 'Number of Sets': None, 'Corner Style': None, 'Weight': None, 'Duration': None, 'External Testing Certification': None, 'Product Benefits': None, 'Position': None, 'Vehicle Service Type': None, 'OEM Part Number': None, 'ABPA Partslink Number': None, 'Pile Height': None, 'Construction Type': None, 'Back Material Type': None, 'Is Stain Resistant':		

Category	Price	Title	Features	Description	Details	SKU	Manufacturer
					None, 'Rug Form Type': None, 'Screen Size': None, 'Screen Surface Description': None, 'Connector Type': None, 'Input Voltage': None, 'Wattage': None, 'Output Current': None, 'Specification Met': None, 'Output Voltage': None, 'Power Source': None, 'Suggested Users': None, 'Sport Type': None, 'Hand Orientation': None, 'Import': None, 'Frame Type': None, 'Voltage': None, 'Plug Format': None, 'Handle Material': None, 'Closure': None, 'Material Type': None, 'Refresh Rate': None, 'Aspect Ratio': None, 'Language': None, 'Care instructions': None, 'Manufacturer recommended age': None,		

Category	Price	Title	Features	Description	Details	SKU	Manufacturer
					'Release date': None, 'Handle Type': None, 'Finish Type': None, 'Metal Type': None, 'Finish': None, 'Dishwasher compatible': None, 'Target Species': None, 'Weave Type': None, 'Maximum Weight Recommendation': None, 'Frame Material': None, 'Number of Handles': None, 'Style Classic': None, 'Finish Type Powder Coated': None, 'Auto Part Position': None, 'Fill Material': None, 'Pillow Type': None, 'Item Firmness Description': None, 'Model Name': None, 'Thread Count': None, 'Item Display Dimensions': None, 'Number Of Pieces': None, 'Bowl Material': None, 'Is Dishwasher Safe': None, 'Is		

	Category	Price	Title	Features	Description	Details	SKU	Manufacturer
						Microwaveable': None, 'Model number': None, 'Recommended Uses For Product': None, 'Battery Cell Composition': None, 'Wireless Type': None, 'Series': None, 'Operating System': None, 'Hardware Interface': None, 'Special features': None}		
2	Automotive	78.910000	WAI World Power Systems World Power Systems DST1835 Distributor	['World Power Systems DST1835 Distributor' 'The package length is 10.78 inches' 'The package height is 10.25 inches' 'The package width is 6.38 inches']	['Wai Distributors/ Wiper Motors Dst1835 New Ignition Distributor']	{'Manufacturer': 'WAI World Power Systems', 'Brand': 'WAI World Power Systems', 'Item Weight': '2.98 pounds', 'Product Dimensions': '16 x 8 x 6 inches', 'Item model number': 'DST1835', 'Manufacturer Part Number': 'DST1835', 'Best Sellers Rank': {'Automotive': 3594961.0, 'Steering Wheel Covers': None, 'Steering Wheel Accessories': None,	DST1835	WAI World Power Systems

Category	Price	Title	Features	Description	Details	SKU	Manufacturer
					'Automotive Exhaust Systems & Parts': None, 'Automotive Replacement Exhaust Pipes': None, 'Automotive Replacement Distributors': 2961.0, 'Audio & Video Connectors & Adapters': None, 'Sports & Outdoors': None, 'Yoga Starter Sets': None, 'Camera Cases': None, 'Patio, Lawn & Garden': None, 'Lawn Mower Belts': None, 'Tools & Home Improvement': None, 'Extension Cords': None, 'Powersports Clutches': None, 'Arts, Crafts & Sewing': None, 'Sewing Products': None, 'Powersports Batteries': None, 'External Hard Drives': None, 'Clothing, Shoes & Jewelry': None, 'Luggage Scales': None,		

Category	Price	Title	Features	Description	Details	SKU	Manufacturer
					'Automotive Replacement Chassis Steering Knuckles': None, 'Lamp Shades': None, 'Home & Kitchen': None, 'Statues': None, 'Office Products': None, 'PBX Phones & Systems': None, 'Exterior Car Care Microfiber Cloths': None, "Kids' Bed Blankets": None, 'TV Wall & Ceiling Mounts': None, 'Electronics Mounts': None, 'Toys & Games': None, "Kids' Stickers": None, 'Kitchen & Dining': None, 'Kitchen Storage Accessories': None, 'Replacement Under-Sink Water Filters': None, 'Automotive Replacement Shock Lift Supports': None, 'Bumper Covers': None, 'Kitchen Rugs': None,		

Category	Price	Title	Features	Description	Details	SKU	Manufacturer
					'Cell Phones & Accessories': None, 'Cell Phone Screen Protectors': None, "Women's Cold Weather Mittens": None, 'Climate Pledge Friendly': None, 'Computer Monitors': None, 'Climate Pledge Friendly: Computers': None, 'Decorative Signs & Plaques': None, 'Cell Phone Basic Cases': None, 'Fishing Equipment': None, 'Powersports Plows': None, 'Snowshoes': None, 'Sliding Door Hardware': None, 'Automotive Replacement Oxygen Sensors': None, 'Mouse Pads': None, 'Chopsticks': None, 'Patio Furniture Pillows': None, 'Chip & Dip Sets': None, 'Automotive Replacement		

Category	Price	Title	Features	Description	Details	SKU	Manufacturer
					Transmission Oil Pressure Sensors': None, 'Automotive Replacement Transmissions & Parts': None, 'Automotive Replacement Electrical Accessories': None, 'Automotive Replacement Engine Oil Drain Plugs': None, 'Throw Pillow Covers': None, 'Pool Rafts & Inflatable Ride-ons': None, 'Outdoor Décor': None, 'Internal USB Port Cards': None, "Kids' Paint With Water Kits": None}, 'Date First Available': 'September 6, 2008', 'Shape': None, 'Installation Type': None, 'Package Dimensions': None, 'Material Leather': None, '': None, 'Country of Origin': 'USA', 'Color Name': None, 'Size': None, 'Color': None, 'Item		

Category	Price	Title	Features	Description	Details	SKU	Manufacturer
					Dimensions LxWxH': None, 'Included Components': None, 'Compatible Devices': None, 'Connector Gender': None, 'Unit Count': None, 'Indoor/ Outdoor Usage': None, 'Is Discontinued By Manufacturer': None, 'UPC': None, 'Item Package Dimensions L x W x H': None, 'Package Weight': None, 'Item Dimensions LxWxH': None, 'Brand Name': None, 'Material': None, 'Part Number': None, 'Model Year': None, 'Theme': None, 'Age Range (Description)': None, 'Package Type': None, 'Style': None, 'Closure Type': None, 'Standing screen display size': None, 'Form Factor': None, 'Shell Type': None, 'Other		

Category	Price	Title	Features	Description	Details	SKU	Manufacturer
					display features': None, 'Compatible Phone Models': None, 'Pattern': None, 'Model': None, 'Exterior': None, 'Number of Pieces': None, 'Mounting Type': None, 'Vehicle Service Type Scooter': None, 'Battery Cell Composition Lead Calcium': None, 'Hard Drive': None, 'Hardware Platform': None, 'Flash Memory Size': None, 'Hard Drive Interface': None, 'Digital Storage Capacity': None, 'Hard Disk Interface': None, 'Connectivity Technology': None, 'Special Feature': None, 'Hard Disk Form Factor': None, 'Hard Disk Description': None, 'Item Package Quantity': None, 'Batteries		

Category	Price	Title	Features	Description	Details	SKU	Manufacturer
					Required?': None, 'Warranty Description': None, 'Domestic Shipping': None, 'International Shipping': None, 'Exterior Finish': None, 'Weight Limit': None, 'Department': None, 'Batteries': None, 'Number of Items': None, 'Top Width': None, 'Base Width': None, 'Special Features': None, 'Shade Color': None, 'Shade Material': None, 'Batteries Included?': None, 'Product Care Instructions': None, 'Collection Name': None, 'Assembly required': None, 'Number of pieces': None, 'Batteries required': None, 'Telephone Type': None, 'Answering System Type': None,		

Category	Price	Title	Features	Description	Details	SKU	Manufacturer
					'Multiline Operation': None, 'Caller Identification': None, 'Fabric Type': None, 'Seasons': None, 'Fabric Warmth Description': None, 'Sport': None, 'Movement Type': None, 'TV Size': None, 'Minimum Compatible Size': None, 'Maximum Tilt Angle': None, 'Room Type': None, 'Cartoon Character': None, 'Reusability': None, 'Material Vinyl': None, 'Shape Irregular': None, 'Hair Type': None, 'Scent': None, 'Item Form': None, 'Occasion': None, 'Number of Sets': None, 'Corner Style': None, 'Weight': None, 'Duration': None, 'External Testing Certification': None, 'Product Benefits': None, 'Position': None, 'Vehicle Service		

Category	Price	Title	Features	Description	Details	SKU	Manufacturer
					Type': None, 'OEM Part Number': None, 'ABPA Partslink Number': None, 'Pile Height': None, 'Construction Type': None, 'Back Material Type': None, 'Is Stain Resistant': None, 'Rug Form Type': None, 'Screen Size': None, 'Screen Surface Description': None, 'Connector Type': None, 'Input Voltage': None, 'Wattage': None, 'Output Current': None, 'Specification Met': None, 'Output Voltage': None, 'Power Source': None, 'Suggested Users': None, 'Sport Type': None, 'Hand Orientation': None, 'Import': None, 'Frame Type': None, 'Voltage': None, 'Plug Format': None, 'Handle Material':		

Category	Price	Title	Features	Description	Details	SKU	Manufacturer
					None, 'Closure': None, 'Material Type': None, 'Refresh Rate': None, 'Aspect Ratio': None, 'Language': None, 'Care instructions': None, 'Manufacturer recommended age': None, 'Release date': None, 'Handle Type': None, 'Finish Type': None, 'Metal Type': None, 'Finish': None, 'Dishwasher compatible': None, 'Target Species': None, 'Weave Type': None, 'Maximum Weight Recommendation': None, 'Frame Material': None, 'Number of Handles': None, 'Style Classic': None, 'Finish Type Powder Coated': None, 'Auto Part Position': None, 'Fill Material': None, 'Pillow Type': None, 'Item Firmness Description': None, 'Model		

Category	Price	Title	Features	Description	Details	SKU	Manufacturer
					Name': None, 'Thread Count': None, 'Item Display Dimensions': None, 'Number Of Pieces': None, 'Bowl Material': None, 'Is Dishwasher Safe': None, 'Is Microwaveable': None, 'Model number': None, 'Recommended Uses For Product': None, 'Battery Cell Composition': None, 'Wireless Type': None, 'Series': None, 'Operating System': None, 'Hardware Interface': None, 'Special features': None}		

data at a glance

Just from looking at a few elements in this dataset I can tell this is going to be a feature engineering problem. Sometimes you get lucky and you get a high quality, homogenous data set that contains several valuable numeric fields that have reasonable distrubitions. Not here. No sir. Without going 1 step further I can tell that at least 80% of the work getting value out of this model is going to be beating and bending these elements in something useful.

Here are some other early observations:

- mixed data types (string, float, lists, nested structures)
- high missingness rate in some fields (particularly Price)
- inconsistent elements between rows (varying Detail field structures)

- extremely sparse data in the Details field (many null values)
- heterogeneous text data formats (titles, features as lists, descriptions as lists)
- variable length text fields (particularly in Features and Description)
- nested/hierarchical data structures that will require flattening or specialized processing
- truncated values that may be missing information
- diverse product attributes that vary

I can also tell that what is now the details column is essentially it's own nested datframe that contains many more rows that the root level columns, so let's go ahead and break that off before we start digging deeper.

In [27]:
```
# seperate columns at root level from details
root_df = df.drop('Details')

# break otu details into it's own dataframe
details_df =  df.select(pl.col("Details").struct.field('*'))
```

null values, data types, and distinct values

There are three things I normally want to now first about any data set I'm analyzing.

- data types
  - specifically, I want to know if the data is categorical or numeric
  - this will have a large impact on the model chosen, especially with regard to the dependant variable
- null values
  - how many elements in a column are null, NaN, empty
- distinct values
  - how many distinct values are there for each field

data types

In [28]:
```
# Create a dataframe from dtypes
root_dtype_list = [(col, str(dtype)) for col, dtype in zip(root_df.columns, root_df.dtypes)]
root_dtype = pd.DataFrame(root_dtype_list, columns=['Column', 'Data Type'])

# Display the result
print("root dataframe dtypes")
display(root_dtype)

# repeat for details
details_dtype_list = [(col, str(dtype)) for col, dtype in zip(details_df.columns, details_df.d
details_dtype = pd.DataFrame(details_dtype_list, columns=['Column', 'Data Type'])

# Reorganize into groups of 25
num_cols = len(details_dtype)
group_size = 58
num_groups = (num_cols + group_size - 1) // group_size

# Create compact display dataframe
compact_df = pd.DataFrame()

for i in range(num_groups):
```

```
    start_idx = i * group_size
    end_idx = min(start_idx + group_size, num_cols)

    group_data = details_dtype.iloc[start_idx:end_idx].reset_index(drop=True)

    # Add to result dataframe
    if i == 0:
        compact_df = group_data
    else:
        # Rename columns for additional groups
        group_data.columns = [f'Column_{i+1}', f'Data Type_{i+1}']
        compact_df = pd.concat([compact_df, group_data], axis=1)

# Display result
print("details dataframe dtypes")
display(compact_df)
```

```
root dataframe dtypes
```

	Column	Data Type
0	Category	String
1	Price	Float64
2	Title	String
3	Features	List(String)
4	Description	List(String)
5	SKU	String
6	Manufacturer	String

```
details dataframe dtypes
```

	Column	Data Type	Column_2	Data Type_2	Column_3	Data Type_3
0	Manufacturer	String	Hard Disk Description	String	Screen Size	String
1	Brand	String	Item Package Quantity	String	Screen Surface Description	String
2	Item Weight	String	Batteries Required?	String	Connector Type	String
3	Product Dimensions	String	Warranty Description	String	Input Voltage	String
4	Item model number	String	Domestic Shipping	String	Wattage	String
5	Manufacturer Part Number	String	International Shipping	String	Output Current	String
6	Best Sellers Rank	Struct({'Automotive': Int64, 'Steering Wheel C...	Exterior Finish	String	Specification Met	String
7	Date First Available	String	Weight Limit	String	Output Voltage	String
8	Shape	String	Department	String	Power Source	String
9	Installation Type	String	Batteries	String	Suggested Users	String
10	Package Dimensions	String	Number of Items	String	Sport Type	String
11	Material Leather	String	Top Width	String	Hand Orientation	String
12		String	Base Width	String	Import	String
13	Country of Origin	String	Special Features	String	Frame Type	String
14	Color Name	String	Shade Color	String	Voltage	String
15	Size	String	Shade Material	String	Plug Format	String
16	Color	String	Batteries Included?	String	Handle Material	String
17	Item Dimensions LxWxH	String	Product Care Instructions	String	Closure	String
18	Included Components	String	Collection Name	String	Material Type	String
19	Compatible Devices	String	Assembly required	String	Refresh Rate	String
20	Connector Gender	String	Number of pieces	String	Aspect Ratio	String
21	Unit Count	String	Batteries required	String	Language	String
22	Indoor/Outdoor Usage	String	Telephone Type	String	Care instructions	String

	Column	Data Type	Column_2	Data Type_2	Column_3	Data Type_3
23	Is Discontinued By Manufacturer	String	Answering System Type	String	Manufacturer recommended age	String
24	UPC	String	Multiline Operation	String	Release date	String
25	Item Package Dimensions L x W x H	String	Caller Identification	String	Handle Type	String
26	Package Weight	String	Fabric Type	String	Finish Type	String
27	Item Dimensions LxWxH	String	Seasons	String	Metal Type	String
28	Brand Name	String	Fabric Warmth Description	String	Finish	String
29	Material	String	Sport	String	Dishwasher compatible	String
30	Part Number	String	Movement Type	String	Target Species	String
31	Model Year	String	TV Size	String	Weave Type	String
32	Theme	String	Minimum Compatible Size	String	Maximum Weight Recommendation	String
33	Age Range (Description)	String	Maximum Tilt Angle	String	Frame Material	String
34	Package Type	String	Room Type	String	Number of Handles	String
35	Style	String	Cartoon Character	String	Style Classic	String
36	Closure Type	String	Reusability	String	Finish Type Powder Coated	String
37	Standing screen display size	String	Material Vinyl	String	Auto Part Position	String
38	Form Factor	String	Shape Irregular	String	Fill Material	String
39	Shell Type	String	Hair Type	String	Pillow Type	String
40	Other display features	String	Scent	String	Item Firmness Description	String
41	Compatible Phone Models	String	Item Form	String	Model Name	String
42	Pattern	String	Occasion	String	Thread Count	String
43	Model	String	Number of Sets	String	Item Display Dimensions	String
44	Exterior	String	Corner Style	String	Number Of Pieces	String
45	Number of Pieces	String	Weight	String	Bowl Material	String
46	Mounting Type	String	Duration	String	Is Dishwasher Safe	String

	Column	Data Type	Column_2	Data Type_2	Column_3	Data Type_3
47	Vehicle Service Type Scooter	String	External Testing Certification	String	Is Microwaveable	String
48	Battery Cell Composition Lead Calcium	String	Product Benefits	String	Model number	String
49	Hard Drive	String	Position	String	Recommended Uses For Product	String
50	Hardware Platform	String	Vehicle Service Type	String	Battery Cell Composition	String
51	Flash Memory Size	String	OEM Part Number	String	Wireless Type	String
52	Hard Drive Interface	String	ABPA Partslink Number	String	Series	String
53	Digital Storage Capacity	String	Pile Height	String	Operating System	String
54	Hard Disk Interface	String	Construction Type	String	Hardware Interface	String
55	Connectivity Technology	String	Back Material Type	String	Special features	String
56	Special Feature	String	Is Stain Resistant	String	NaN	NaN
57	Hard Disk Form Factor	String	Rug Form Type	String	NaN	NaN

Main dataframe structure:

- Core fields use consistent types (strings, float64 for Price)
- Text fields use string or list-of-string types
- All identifier fields are strings

Details dataframe structure:

- Contains 172 attributes as separate columns
- Almost entirely string-typed despite containing numeric data
- Contains NaN values in bottom rows
- Includes struct field with nested information
- Contains redundant columns (multiple dimension fields)
- Includes dates stored as strings
- Contains boolean concepts as strings ("Is Dishwasher Safe")
- Multiple columns represent the same concept (dimensions, weight)

Data characteristics:

- String fields contain numeric values
- Temporal data stored as text

- Categorical information spread across many columns
- Inconsistent representation of similar data points

nulls

In [20]:
```
# calculate nulls and null perctangates
nulls = root_df.null_count().unpivot(
    variable_name='column',
    value_name='null_count'
).with_columns(
    null_percentage = (pl.col('null_count')/df.height)*100
).sort(by='null_percentage', descending=False)

# stylized output
styled_nulls = (nulls.to_pandas()
    .style
    .background_gradient(subset=['null_percentage'], cmap='RdYlGn_r')
    .bar(subset=['null_percentage'], color='#4a90e2', align='mid')
    .format({'null_percentage': '{:,.2f}%'})
    .set_caption('root nulls')
)

display(styled_nulls)
```

root nulls

	column	null_count	null_percentage
0	Category	0	0.00%
1	Title	0	0.00%
2	Features	0	0.00%
3	Description	0	0.00%
4	SKU	0	0.00%
5	Manufacturer	0	0.00%
6	Price	24857	58.58%

In [21]:
```
# details null percentages
# calculate nulls and null perctangates
nulls = details_df.null_count().unpivot(
    variable_name='column',
    value_name='null_count'
).with_columns(
    null_percentage = (pl.col('null_count')/df.height)*100
).filter(
    pl.col('null_percentage') <= 90
).sort(by='null_percentage', descending=False)

# stylized output
styled_nulls = (nulls.to_pandas()
    .style
    .background_gradient(subset=['null_percentage'], cmap='RdYlGn_r')
    .bar(subset=['null_percentage'], color='#4a90e2', align='mid')
    .format({'null_percentage': '{:,.2f}%'})
    .set_caption('details nulls w/null_percentage <= %90')
)
```

```
display(styled_nulls)
```

details nulls w/null_percentage <= %90

	column	null_count	null_percentage
0	Manufacturer	4285	10.10%
1	Date First Available	8586	20.24%
2	Item model number	9395	22.14%
3	Brand	15528	36.60%
4	Item Weight	16232	38.26%
5	Product Dimensions	20036	47.22%
6	Best Sellers Rank	25501	60.10%
7	Is Discontinued By Manufacturer	25878	60.99%
8	Color	26461	62.37%
9	Material	29162	68.73%
10	Package Dimensions	29173	68.76%
11	Department	35007	82.51%
12	Manufacturer Part Number	35659	84.04%
13	Part Number	36146	85.19%
14	Style	36627	86.33%
15	Size	36781	86.69%
16	Special Feature	36828	86.80%
17	Country of Origin	37610	88.64%
18	Number of Items	37647	88.73%
19	Item Dimensions LxWxH	37951	89.45%

- In the main dataframe, Price is the only field with missing values (58.58%)
- All other primary fields (Category, Title, Features, Description, SKU, Manufacturer) are complete
- In the details dataframe, null rates are extremely high across all fields
- Even the most populated detail fields have significant missingness:
  - Manufacturer (10.10%)
  - Date First Available (20.24%)
  - Item model number (22.14%)
- Most detail fields have very high null rates, exceeding 80%
- The data shows a clear pattern of sparsity, particularly in the Details table
- Certain fields like Item Dimensions LxWxH (89.45%) are present in only about 10% of records
- The Details fields appear to vary substantially by product type/category

distincts

In [22]:
```
# calculate unique values and unique value percentages for root_df
```

```
distincts = root_df.select([
    pl.all().n_unique()
]).transpose(
    include_header=True,
    column_names=["distinct_count"]
).with_columns(
    distinct_percentage=(pl.col('distinct_count')/df.height)*100
).sort(by='distinct_percentage', descending=True)

# stylize
styled_distincts = (distincts.to_pandas()
    .style
    .background_gradient(subset=['distinct_percentage'], cmap='RdYlGn_r')
    .bar(subset=['distinct_percentage'], color='#4a90e2', align='mid')
    .format({'distinct_percentage': '{:,.2f}%'})
    .set_caption('root distincts')
)

display(styled_distincts)
```

root distincts

	column	distinct_count	distinct_percentage
0	Title	42362	99.84%
1	SKU	41484	97.77%
2	Features	35322	83.25%
3	Description	29145	68.69%
4	Manufacturer	28041	66.09%
5	Price	5288	12.46%
6	Category	28	0.07%

In [23]:
```
# calculate unique values and unique value percentages for root_df
distincts = details_df.select([
    pl.all().n_unique()
]).transpose(
    include_header=True,
    column_names=["distinct_count"]
).with_columns(
    distinct_percentage=(pl.col('distinct_count')/df.height)*100
).sort(by='distinct_percentage', descending=True)

# stylize
styled_distincts = (distincts.to_pandas()
    .style
    .background_gradient(subset=['distinct_percentage'], cmap='RdYlGn_r')
    .bar(subset=['distinct_percentage'], color='#4a90e2', align='mid')
    .format({'distinct_percentage': '{:,.2f}%'})
    .set_caption('details distincts')
)

display(styled_distincts)
```

details distincts

	column	distinct_count	distinct_percentage
0	Item model number	32258	76.03%
1	Manufacturer	25164	59.31%
2	Product Dimensions	19622	46.25%
3	Brand	18424	43.42%
4	Best Sellers Rank	13789	32.50%
5	Package Dimensions	12692	29.91%
6	Manufacturer Part Number	6691	15.77%
7	Part Number	6190	14.59%
8	Date First Available	6167	14.53%
9	Color	5216	12.29%
10	Item Dimensions LxWxH	4026	9.49%
11	UPC	3571	8.42%
12	Size	3322	7.83%
13	Item Weight	3222	7.59%
14	Item Package Dimensions L x W x H	2507	5.91%
15	Brand Name	2428	5.72%
16	Included Components	2405	5.67%
17	Material	2397	5.65%
18	Special Feature	2196	5.18%
19	Style	1936	4.56%
20	Model Name	1795	4.23%
21	Item Dimensions LxWxH	1650	3.89%
22	Model	1340	3.16%
23	OEM Part Number	1294	3.05%
24	Special Features	878	2.07%
25	Theme	866	2.04%
26	Compatible Phone Models	780	1.84%
27	Package Weight	670	1.58%
28	Compatible Devices	667	1.57%
29	Recommended Uses For Product	626	1.48%
30	Release date	613	1.44%
31	Warranty Description	557	1.31%
32	Unit Count	530	1.25%

	column	distinct_count	distinct_percentage
33	Special features	523	1.23%
34		506	1.19%
35	Material Type	484	1.14%
36	Finish Type	427	1.01%
37	Occasion	391	0.92%
38	Shape	368	0.87%
39	Fabric Type	348	0.82%
40	Mounting Type	323	0.76%
41	Scent	322	0.76%
42	Room Type	317	0.75%
43	Department	315	0.74%
44	Product Benefits	312	0.74%
45	Item Form	280	0.66%
46	Pattern	274	0.65%
47	Series	269	0.63%
48	Wattage	266	0.63%
49	Finish	257	0.61%
50	Vehicle Service Type	252	0.59%
51	Cartoon Character	225	0.53%
52	Manufacturer recommended age	196	0.46%
53	ABPA Partslink Number	191	0.45%
54	Connector Type	187	0.44%
55	Sport Type	184	0.43%
56	Product Care Instructions	181	0.43%
57	Handle Material	164	0.39%
58	Age Range (Description)	159	0.37%
59	Color Name	150	0.35%
60	Number of Pieces	148	0.35%
61	Weight	148	0.35%
62	Installation Type	147	0.35%
63	Connectivity Technology	146	0.34%
64	Screen Size	145	0.34%
65	Exterior	143	0.34%
66	Closure Type	142	0.33%

	column	distinct_count	distinct_percentage
67	Power Source	142	0.33%
68	Voltage	142	0.33%
69	Suggested Users	140	0.33%
70	Form Factor	135	0.32%
71	Operating System	134	0.32%
72	Sport	133	0.31%
73	Batteries	122	0.29%
74	Hair Type	118	0.28%
75	Exterior Finish	116	0.27%
76	Specification Met	113	0.27%
77	Shade Material	109	0.26%
78	Standing screen display size	107	0.25%
79	Hard Drive	96	0.23%
80	Target Species	94	0.22%
81	Frame Material	81	0.19%
82	Maximum Weight Recommendation	79	0.19%
83	Back Material Type	72	0.17%
84	Model number	72	0.17%
85	Number of Items	68	0.16%
86	Care instructions	68	0.16%
87	Hardware Platform	66	0.16%
88	Country of Origin	61	0.14%
89	Weight Limit	60	0.14%
90	Hardware Interface	57	0.13%
91	Closure	53	0.12%
92	Item Display Dimensions	50	0.12%
93	Collection Name	47	0.11%
94	Shade Color	46	0.11%
95	Metal Type	46	0.11%
96	Wireless Type	46	0.11%
97	Seasons	44	0.10%
98	Auto Part Position	38	0.09%
99	Flash Memory Size	37	0.09%
100	Fill Material	35	0.08%

	column	distinct_count	distinct_percentage
101	Number Of Pieces	34	0.08%
102	Position	32	0.08%
103	Input Voltage	32	0.08%
104	Item Package Quantity	30	0.07%
105	Digital Storage Capacity	29	0.07%
106	Number of pieces	26	0.06%
107	Output Voltage	26	0.06%
108	Weave Type	25	0.06%
109	Hand Orientation	24	0.06%
110	Language	23	0.05%
111	Hard Drive Interface	22	0.05%
112	TV Size	22	0.05%
113	Handle Type	22	0.05%
114	Item Firmness Description	22	0.05%
115	Model Year	21	0.05%
116	Construction Type	20	0.05%
117	Thread Count	20	0.05%
118	Package Type	18	0.04%
119	Hard Disk Description	18	0.04%
120	Duration	18	0.04%
121	Import	18	0.04%
122	Frame Type	18	0.04%
123	Minimum Compatible Size	17	0.04%
124	Plug Format	17	0.04%
125	Connector Gender	16	0.04%
126	Other display features	16	0.04%
127	Number of Sets	16	0.04%
128	Hard Disk Interface	15	0.04%
129	Battery Cell Composition	15	0.04%
130	Output Current	13	0.03%
131	Bowl Material	13	0.03%
132	Shell Type	12	0.03%
133	Base Width	11	0.03%
134	Screen Surface Description	10	0.02%

	column	distinct_count	distinct_percentage
135	Refresh Rate	10	0.02%
136	Top Width	9	0.02%
137	Movement Type	9	0.02%
138	External Testing Certification	8	0.02%
139	Hard Disk Form Factor	7	0.02%
140	Maximum Tilt Angle	7	0.02%
141	Reusability	7	0.02%
142	Domestic Shipping	6	0.01%
143	Fabric Warmth Description	6	0.01%
144	Pile Height	6	0.01%
145	Rug Form Type	6	0.01%
146	Aspect Ratio	6	0.01%
147	Number of Handles	6	0.01%
148	Pillow Type	6	0.01%
149	Is Discontinued By Manufacturer	5	0.01%
150	Batteries Required?	5	0.01%
151	Batteries Included?	5	0.01%
152	Telephone Type	5	0.01%
153	Indoor/Outdoor Usage	4	0.01%
154	Batteries required	4	0.01%
155	Corner Style	4	0.01%
156	International Shipping	3	0.01%
157	Assembly required	3	0.01%
158	Answering System Type	3	0.01%
159	Multiline Operation	3	0.01%
160	Caller Identification	3	0.01%
161	Is Stain Resistant	3	0.01%
162	Dishwasher compatible	3	0.01%
163	Is Dishwasher Safe	3	0.01%
164	Is Microwaveable	3	0.01%
165	Material Leather	2	0.00%
166	Vehicle Service Type Scooter	2	0.00%
167	Battery Cell Composition Lead Calcium	2	0.00%
168	Material Vinyl	2	0.00%

	column	distinct_count	distinct_percentage
169	Shape Irregular	2	0.00%
170	Style Classic	2	0.00%
171	Finish Type Powder Coated	2	0.00%

Main dataframe:

- Title (99.84%) and SKU (97.77%) are nearly unique for each product
- Features (83.25%) and Description (68.69%) have high uniqueness, suggesting custom text per product
- Price has only 5,288 distinct values (12.46%), indicating price point clustering
- Category has exactly 28 distinct values - this is our target variable

Details dataframe: -I tem model number (76.03%) is the most unique field in Details

- Manufacturer (59.31%) and Product Dimensions (46.23%) show moderate uniqueness
- Significant duplication in many attributes, with most fields under 10% uniqueness
- Multiple fields for similar concepts (e.g., Item Dimensions appears 3 times)
- Some fields like Special Feature, Special Features, and Special features have different uniqueness levels despite similar names

analysis of individual fields

category

Let's get to know our target variable a little bit better.

In [24]:
```
# display information about the catgory column
categories = root_df.group_by(
    'Category'
).agg(pl.len().alias('count')
).with_columns(
    total_percentage = (pl.col('count')/df.height)*100
).sort(by='total_percentage', descending=True)

# stylize
styled_categories = (categories.to_pandas()
    .style
    .background_gradient(subset=['total_percentage'], cmap='RdYlGn_r')
    .bar(subset=['total_percentage'], color='#4a90e2', align='mid')
    .format({'total_percentage': '{:,.2f}%'})
    .set_caption('categories')
)

display(styled_categories)
```

categories

	Category	count	total_percentage
0	Fashion	6420	15.13%
1	Home	5060	11.93%
2	Automotive	4651	10.96%
3	Tools & Home Improvement	3245	7.65%
4	Sports & Outdoors	2810	6.62%
5	All Beauty	2147	5.06%
6	Office Products	2101	4.95%
7	Toys & Games	2012	4.74%
8	Cell Phones & Accessories	1890	4.45%
9	Industrial & Scientific	1591	3.75%
10	Grocery	1543	3.64%
11	All Electronics	1463	3.45%
12	Computers	1275	3.01%
13	Arts, Crafts & Sewing	1082	2.55%
14	Health & Personal Care	929	2.19%
15	Pet Supplies	787	1.85%
16	Camera & Photo	679	1.60%
17	Digital Music	492	1.16%
18	Musical Instruments	467	1.10%
19	Home Audio & Theater	433	1.02%
20	Baby	332	0.78%
21	Appliances	249	0.59%
22	Video Games	211	0.50%
23	Premium Beauty	200	0.47%
24	Car Electronics	146	0.34%
25	GPS & Navigation	73	0.17%
26	Portable Audio & Accessories	73	0.17%
27	Software	68	0.16%

- The dataset contains 28 distinct product categories
- Distribution is imbalanced, with Fashion (15.13%), Home (11.93%), and Automotive (10.96%) making up over a third of all products
- The smallest categories (Software, GPS & Navigation, Portable Audio) each represent less than 0.2% of the data

- There's a significant drop-off after the top 3 categories
- Related categories exist in the dataset (e.g., "All Beauty" and "Premium Beauty")

numeric fields

Let's generate some descriptive stats and visualizations to understand the look and feel of the two un-anonymized input fields.

In [25]:
```
# summary stats for time and amount
numerics = root_df.select(pl.selectors.numeric())

display(numerics.describe())
```

shape: (9, 2)

statistic	Price
str	f64
"count"	17572.0
"null_count"	24857.0
"mean"	60.408664
"std"	210.016231
"min"	0.01
"25%"	12.88
"50%"	21.99
"75%"	49.09
"max"	18999.0

My absolute favorite plot, is the violin plot. It displays your mean, your quartiles, your outliers, and in my opinion there is not better plot to gain a quick understand of the general nature of a field. Unfortunately for the amount field the plot is not very interesting, other than telling us we have some extreme outliers on the upper end.

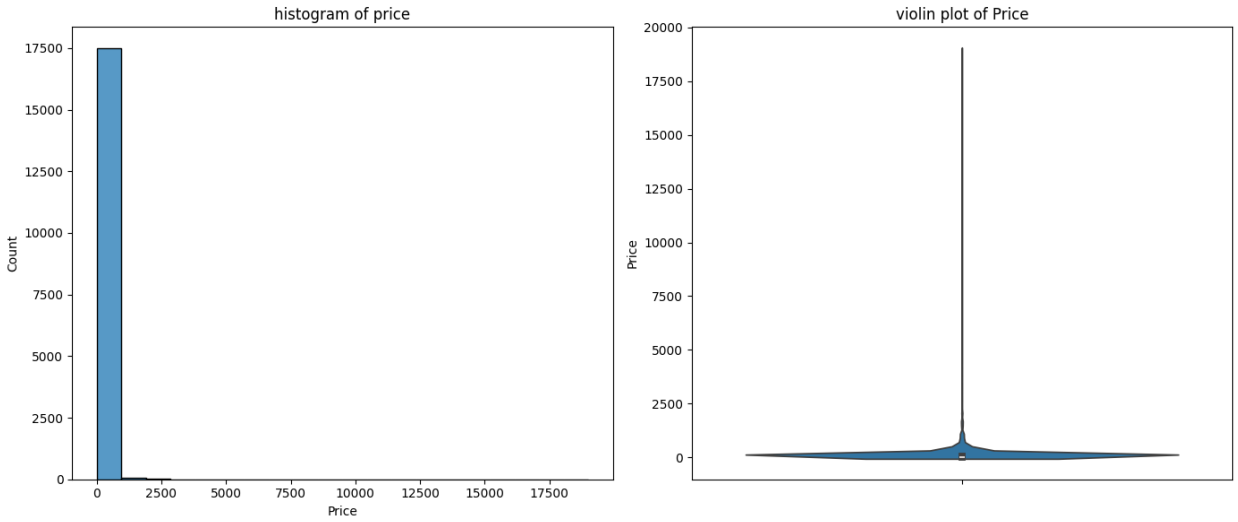
In [26]:
```
# visualize numeric fields
numerics_pdf = numerics.to_pandas()

# Create a figure with two subplots side by side
fig, axes = plt.subplots(1, 2, figsize=(14, 6))

# Plot time on histogram
sns.histplot(numerics_pdf['Price'], bins=20, ax=axes[0])
axes[0].set_title('histogram of price')

# Plot time on violin plot
sns.violinplot(y=df['Price'], ax=axes[1])
axes[1].set_title('violin plot of Price')

# Display the plots
plt.tight_layout()
plt.show()
```



- High missing value rate (58.58% nulls - 24,857 out of 42,429 products)
- Extreme range from $0.01 to 18,999, with mean at $60.40
- Highly right-skewed distribution (histogram shows most values clustered near zero)
- Large standard deviation ($210.02) relative to median ($21.99)
- The violin plot reveals outliers extending to nearly $20,000, while most prices concentrate below $50

conclusions

Features to Included

- Transformed Price: Log-transformed and normalized the price field despite 58.58% missing values, preserving nulls as meaningful signals
- Transformed Text Fields: Generated embeddings for Title, Features, and Description using DistilBERT, then applied dimensionality reduction (TruncatedSVD) to create 50 components for each field
- Details Fields: Created binary indicators (1/0) for presence/absence of each attribute in the Details structure, then applied SparsePCA to reduce from 172 sparse columns to 50 components

Features to Excluded

- SKU: Removed due to extremely high uniqueness (97.77%), which would lead to overfitting
- Manufacturer: Excluded as a standalone feature since the semantic information is captured in the text embeddings
- Raw Text Fields: Removed original text fields after embedding generation to reduce dimensionality

Planned Feature Transformations

- Price Field: Despite high missingness (58.58%), I'll apply log transformation and normalization while preserving nulls as potentially meaningful signals
- Text Fields: Will convert Title, Features, and Description into usable features, likely using embeddings or vectorization to capture their semantic content
- Details Structure: Plan to handle the nested structure by creating indicators for field presence/absence, then applying dimensionality reduction since there are too many sparse columns

Implementation Considerations

- Need to standardize inconsistent field naming across the dataset
- Must address redundant fields (like multiple dimension formats) through consolidation
- Will require dimensionality reduction techniques to handle the large feature space
- Text processing approach needs to capture semantic relationships while remaining computationally efficient