Premise Data Science Problem

The following is a 245k sample of Premise food staples observations from Ghana (21mb gzipped csv), collected since October 2014.

http://archaeopteryx.premise.com/data_sample/premise_GH_food-staples_2015-07-20.csv.gz

Data Analysis and Statistical Modeling Questions

1. i. For the data set, identify high price outliers. Explain how you identified these outliers.
2. i. For the data set, describe some variables that could be sources of sampling bias when estimating price trends in Ghana. Explain why each of these variables could cause sampling bias.
   ii. Pick one bias (not u_uuid). Write some code that attempts to estimate the potential amount of sample bias caused by this variable, and describe your methodology.

Modeling Question

1. Create a model that predicts price from various metadata.
   i. Explain how your model works, and why you chose it.
   ii. Why did you use the metadata you used?
   iii. How can you be sure that you're not over-fitting the model?

Submit the output and the code that generated the output (should be runnable if the environment has the requisite packages).

Dataset details: An observation represents the price of an item (p_item_uuid) at a particular location (l_place_uuid) and time (t_time). Each observation contains a set of metadata that provides more information about that capture.

Metadata fields

Field	Type	Example	Description
t_time	datetime	"2014-06-12T 13.21.34.000Z"	Time of observation submission
p_item_manufacturer_lc	string	"Nabisco"	Manufacturer of observation (if applicable)
p_item_brand_lc	string	"Chips Ahoy"	Brand of observation (if applicable)
p_item_sub_brand_lc	string	"Double Chocolate"	Sub-brand of observation (if applicable)
p_item_product_lc	string	"Chocolate Chip Cookies"	Observed product
p_item_description_lc	string		Observation description
p_item_uuid	string	"495115e1cf193baadb0504b7a87c49d450eb1"	Unique identifier for observed product
packaging	string	"1.0 x 12.4oz"	Quantity in packaging x unit size
p_quantity	string	"1"	Quantity in packaging
p_size	string	"12.4"	Unit size
p_units_lc	string	"oz"	Unit
p_price	double	3.49	Sale price of observed product in indicated packaging
normalized_price	double	0.00992792158	Price normalized according to quantity and size
normalized_size_units	string	"g"	Units of the product's size after normalization
p_currency	string	"BRL"	Currency of sale price
g_language	string	PT	Language of locale (ISO-639-2)
city	string	"Belo Horizonte"	City of observation
g_country	string	"BR"	Two-letter code of country of observation (ISO-3166-alpha2)
l_place_name	string	"Carrefour"	Name of store where product was observed
l_place_uuid	string	"780b3709-fab3-48bf-9a4e-1e81db02b33a"	Unique identifier for store where product was observed
g_lat	double	-22.9246044159	Latitude where product was observed
g_lon	double	-43.23856354	Longitude where product was observed
g_loc_accuracy	double	20.3999996185	Android location accuracy.
t_created	timestamp	2014-11-21 02:40:46.397000	ISO-8601 timestamp of item creation.
t_modified	timestamp	2014-11-21 02:40:46.397000	ISO-8601 timestamp of last modification.
t_uploaded	timestamp	2014-11-21 02:40:46.397000	ISO-8601 timestamp of item upload.
g_source	string	offline	(Unused) data source
thumbnail_0x0	string	https://img.premise.com/300x300/cd038d42ecd0d4151e3abbee9cc2a1ee12572d2f	Full size image
thumbnail_300x300	string	https://img.premise.com/300x300/cd038d42ecd0d4151e3abbee9cc2a1ee12572d2f	300x300 thumbnail of image
u_uuid	string	cd038d42ecd0d4151e3abbee9cc2a1ee12572d2f	User unique identifier.