Business Context

A problem about classification products in some categories was found and to solve this problem was decided to create a model for products' classification.

Data Description

The dataset have about 38507 rows and 15 columns to start with and was need to drop one column and 60 rows in reason of NA values in these place, letting about 38447 rows and 14 columns on the dataset and these statistical description below:

	attributes	mean	median	std	min	max	range	skew	kurtosis
0	search_page	1.49	1.00	0.98	1.00	5.00	4.00	2.12	3.69
1	position	16.89	16.00	11.59	0.00	38.00	38.00	0.21	-1.21
2	price	84.12	28.53	211.95	0.07	11509.38	11509.31	17.46	523.48
3	weight	361.75	9.00	1820.75	0.00	65009.00	65009.00	16.77	411.34
4	express_delivery	0.78	1.00	0.41	0.00	1.00	1.00	-1.36	-0.16
5	minimum_quantity	14.60	7.00	43.80	0.00	3000.00	3000.00	29.63	1355.95
6	view_counts	545.99	243.00	1417.47	1.00	45010.00	45009.00	14.74	321.12

Feature Engineering

For create better models was did a feature engineering and derived these feature below:

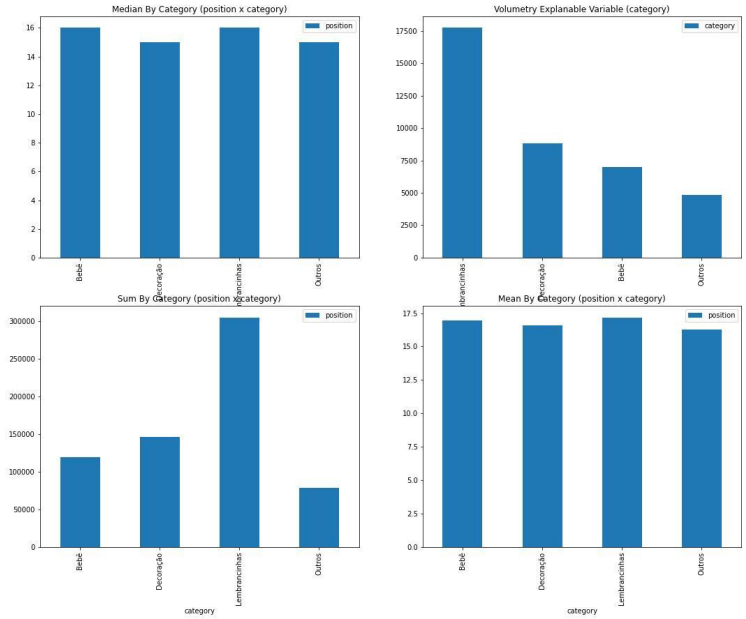
- **page_position**: indicate the position on the page of the product.
- **price_per_weight**: indicate the price per weight of this product.
- **day_of_week**: indicate the day of week of creation of this product.
- **week_of_year**: indicate the week of year of creation of this product.
- **month**: indicate the month of creation of this product.
- **year**: indicate the year of creation of this product.

In addition, it was decided to concat the categories of “Papel e Cia” and “Bijuterias e Jóias” with “Outros” because they don’t have a significant quantity in front of others’ categories.

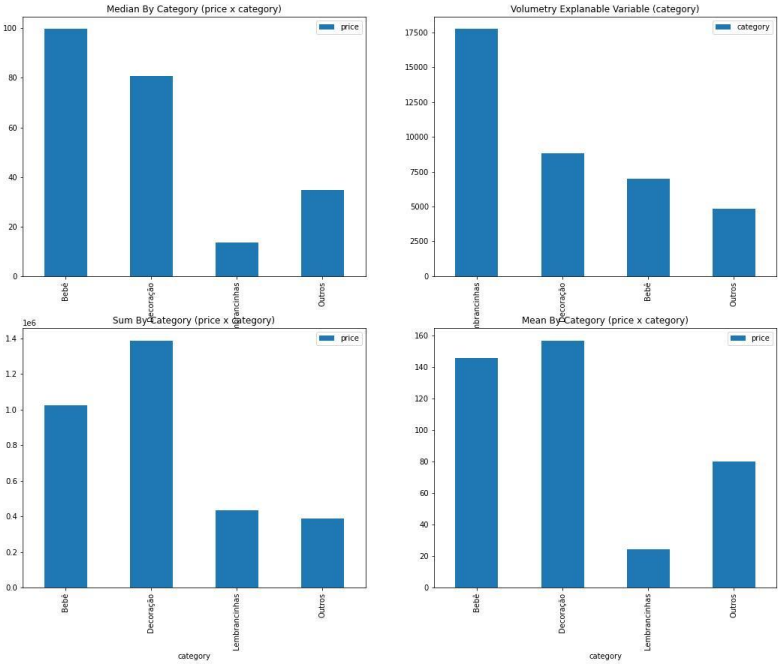
Data Analysis

Was did some data analysis for select the best feature and selected nine feature of all the dataset for modeling, the features are:

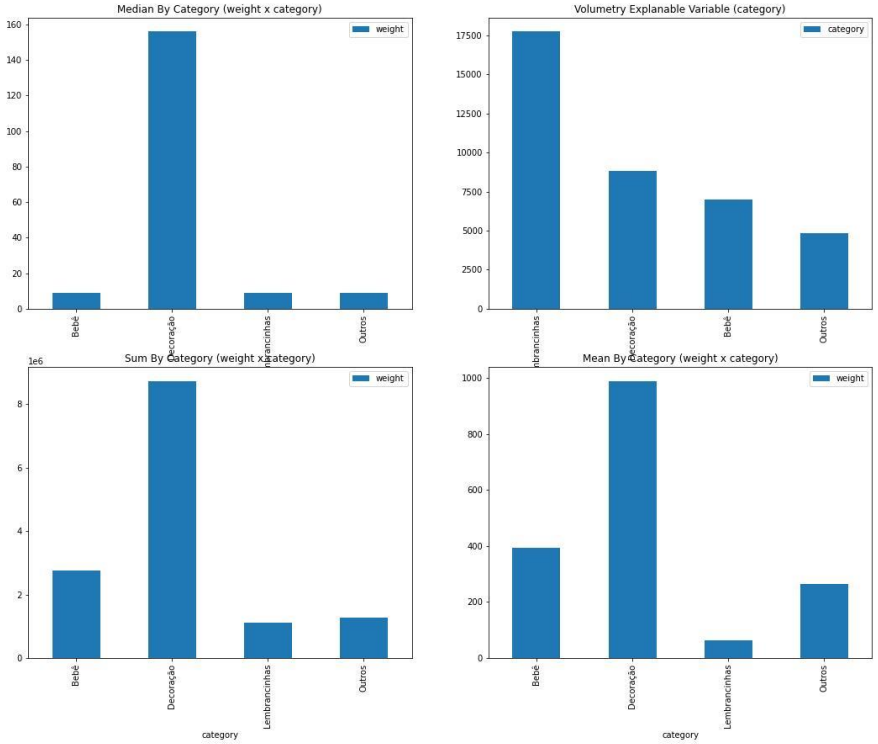
- **position**: a little impact, but have some, has a low quantity of features, was decided to follow with it.



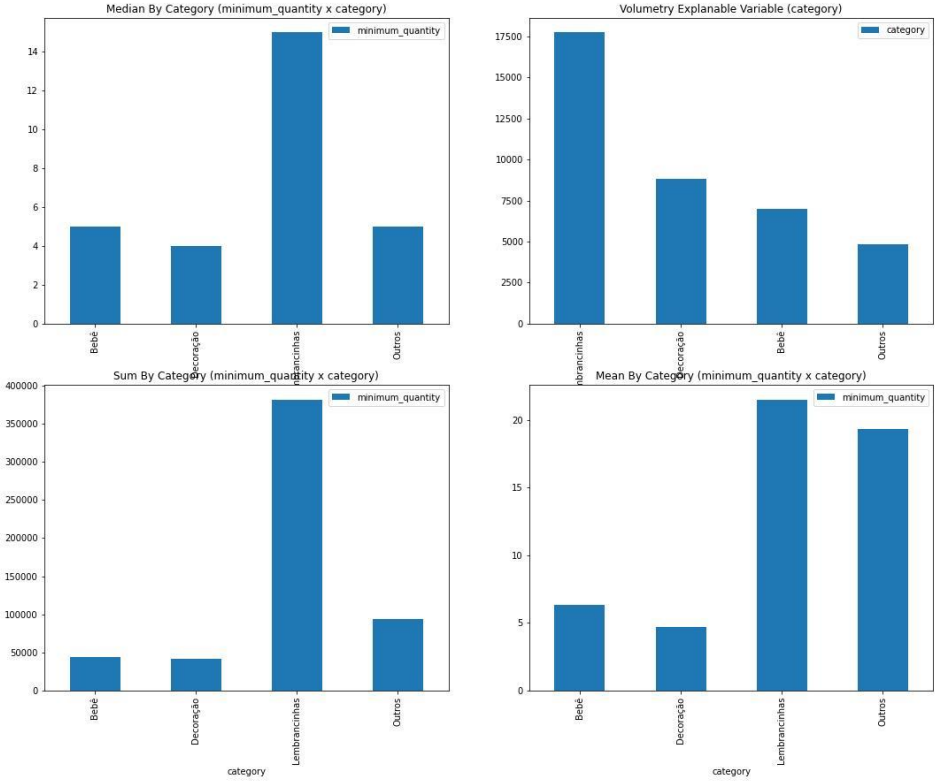
- **price**: a big difference can be noticed in this feature, decided to follow with her beucase this.



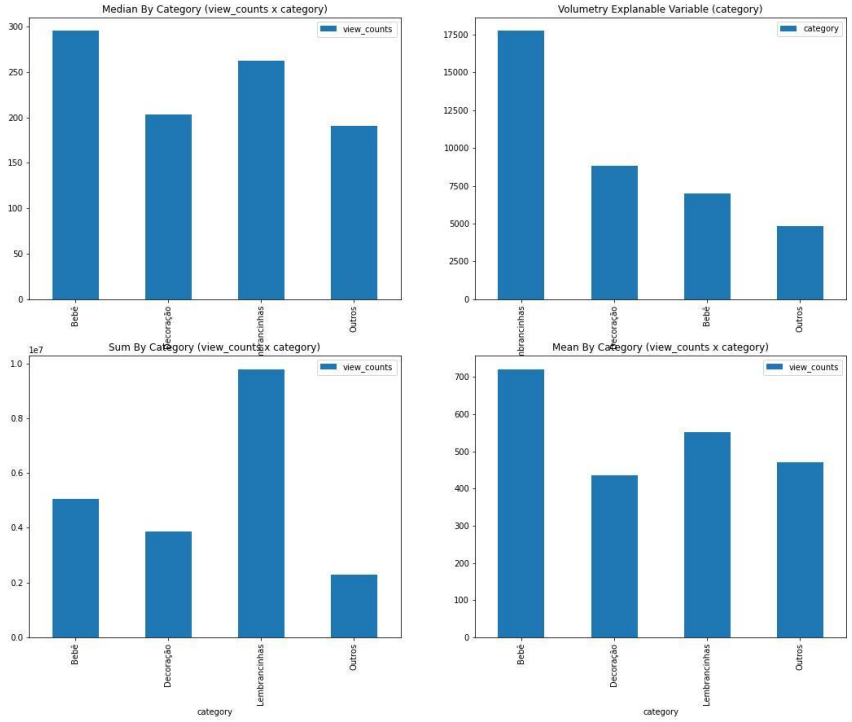
- **weight**: help to differentiate the “Decoração” category of others.



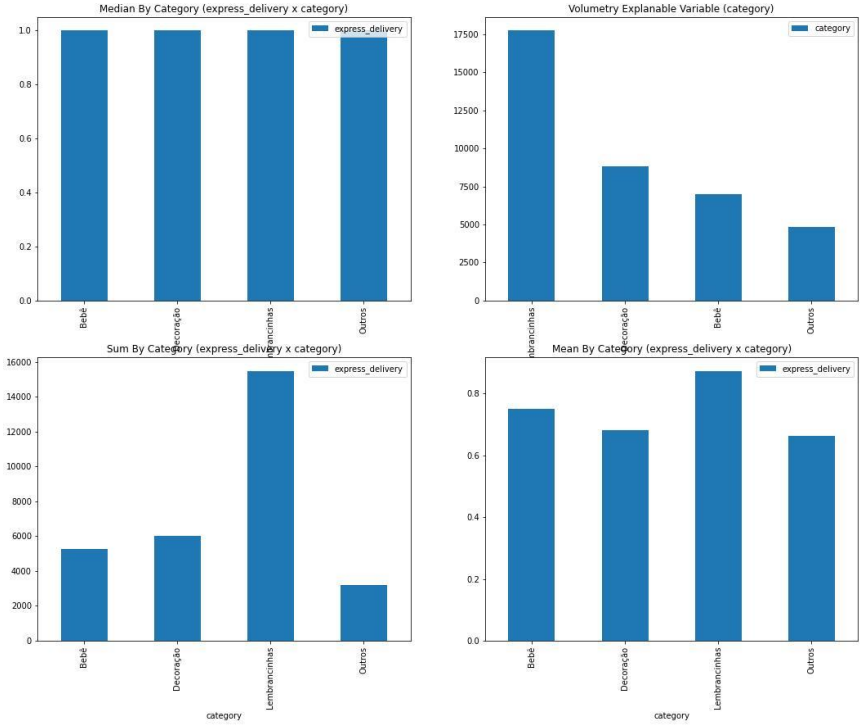
- **minimum_quantity**: help to differentiate “Lembrancinhas” of the others.



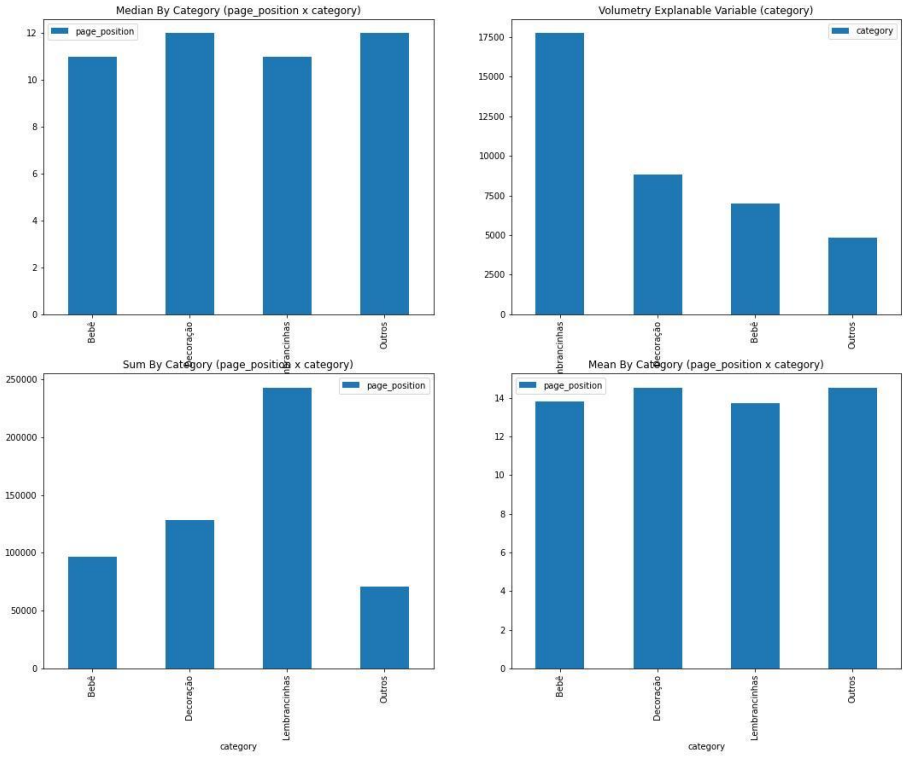
- **view_counts**: selected because it's a good feature with balanced differences between the categories.



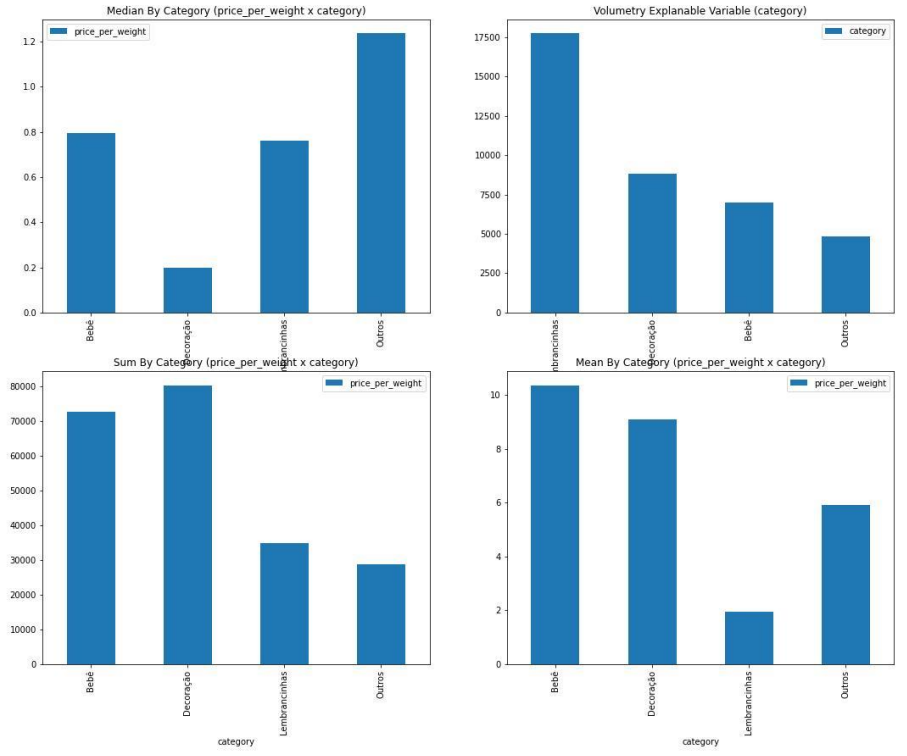
- **express_delivery**: the worst feature selected, but has some changes in mean, was decided to follow with it because of this.



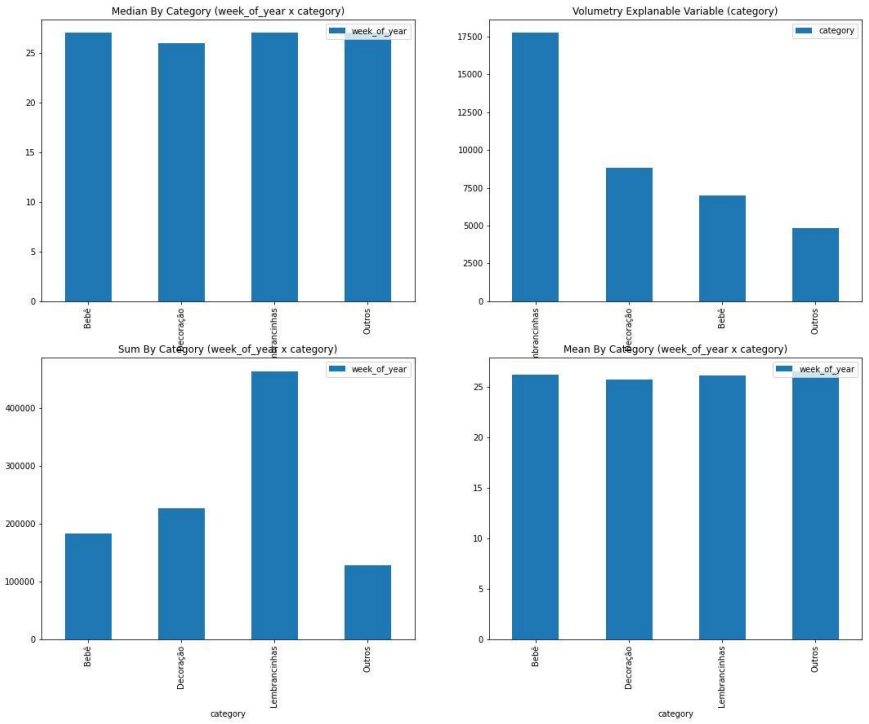
- **page_position**: low impact, but has some, decided to follow with it because of this.



- **price_per_weight**: a very good feature, going to be used because it differentiates capacity on median values.



- **week_of_year**: decided to use it to catch some seasonal effect invisible on the dataset.



Data Transformation

On data transformation step was decided to follow two main steps:

- Rescale all features for the same scale, Min-Max Scale, for testing some models who depend on calculations between features.
- Decided to apply the “Label Encoder '' technique from “skelarn.preprocessing” for encoding the response variable.

Model' Performance

For select a model was tested three algorithms with the performance below (K-Folds Cross Validation):

	Model Name	Accuracy CV	Kappa CV	Precision CV	Recall CV	F1-Score CV
0	Extreme Gradient Boosting Classifier	0.64	0.45	0.55	0.50	0.50
0	Random Forest Classifier	0.61	0.40	0.51	0.46	0.43
0	K-Nearest Neighbour Classifier	0.49	0.13	0.37	0.31	0.28

The best performance was from XGBoost, then I decided to follow with it in the deployment step.
Below has the feature importance of the features for this model selected: