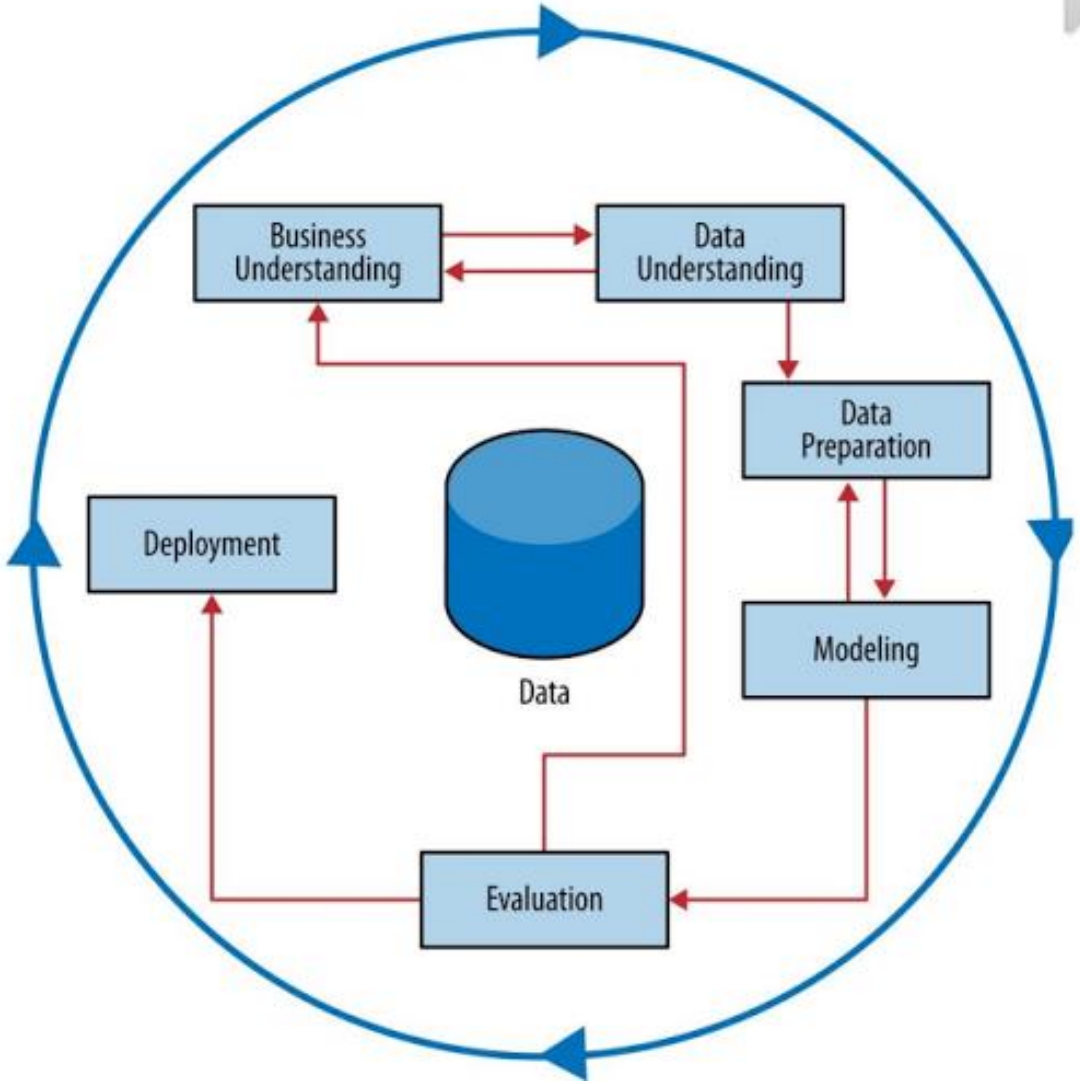
Ciclo CRISP-DM



Problema de negócio

O contexto não foi definido. Entretanto, para contextualizar o projeto **estarei adotando a seguinte abordagem:**

Contexto: Uma empresa chamada QualityShoes opera no ramo de e-commerce dentro da área de artigos esportivos. Como a empresa ainda não tem um time de dados estruturado, ela fez um contrato com uma empresa de consultoria especializada no ramo de transformação digital. Com esse contrato, fui chamado para atuar como cientista de dados. Comecei a marcar algumas reuniões com as áreas de negócio da empresa cliente, e após uma primeira reunião com o CFO da quality e o nosso time de dados, constatamos que o primeiro projeto a ser realizado deveria ser a projeção de receita bruta/líquida. O portfólio da quality é bem amplo, com mais de 100 mil clientes atacadistas e varejistas. A ideia inicial do projeto é entender as principais variáveis que estão correlacionadas à receita e retirar insights importantes para o time de negócio.

Objetivo de negócio: Projeção de receita para os próximos 6 meses

Problema de negócio

Entendimento do negócio:

1. Qual a motivação:

    * A projeção de receita nos próximos 6 meses surgiu a partir da necessidade de desenhar o melhor budget para investimentos nas áreas da empresa, quanto mais seguro é a minha predição receita, menos riscos a empresa estará correndo com investimentos mais arriscados.

2. Qual a causa raíz do problema:
    - Dificuldade em determinar o melhor budget para investimentos internos.

3. Quem é o dono do problema:
    - Diretor financeiro (CFO) da QualityShoes

4. Qual é o formato da solução?

    Granularidade: Previsão de receita diária nos próximos 183 dias (6 meses)

    Tipo de problema: Previsão de receita (Regressão)

    Potenciais métodos: Séries temporais e regressão com algumas modificações

    Formato de entrega:

    * O valor total da receita líquida no final dos 6 meses.
    * A entrega será pelo app do streamlit
    * Checagem trimestral

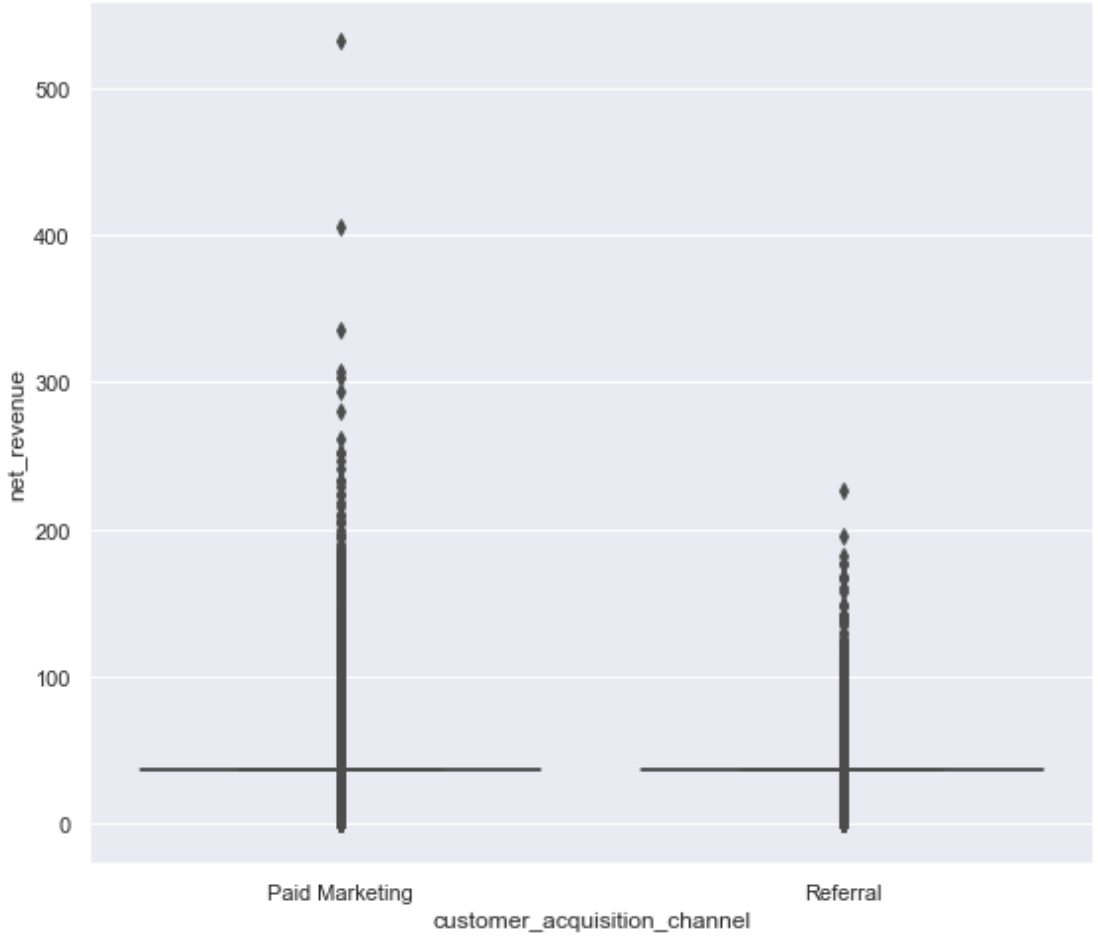
Conhecimento sobre os dados

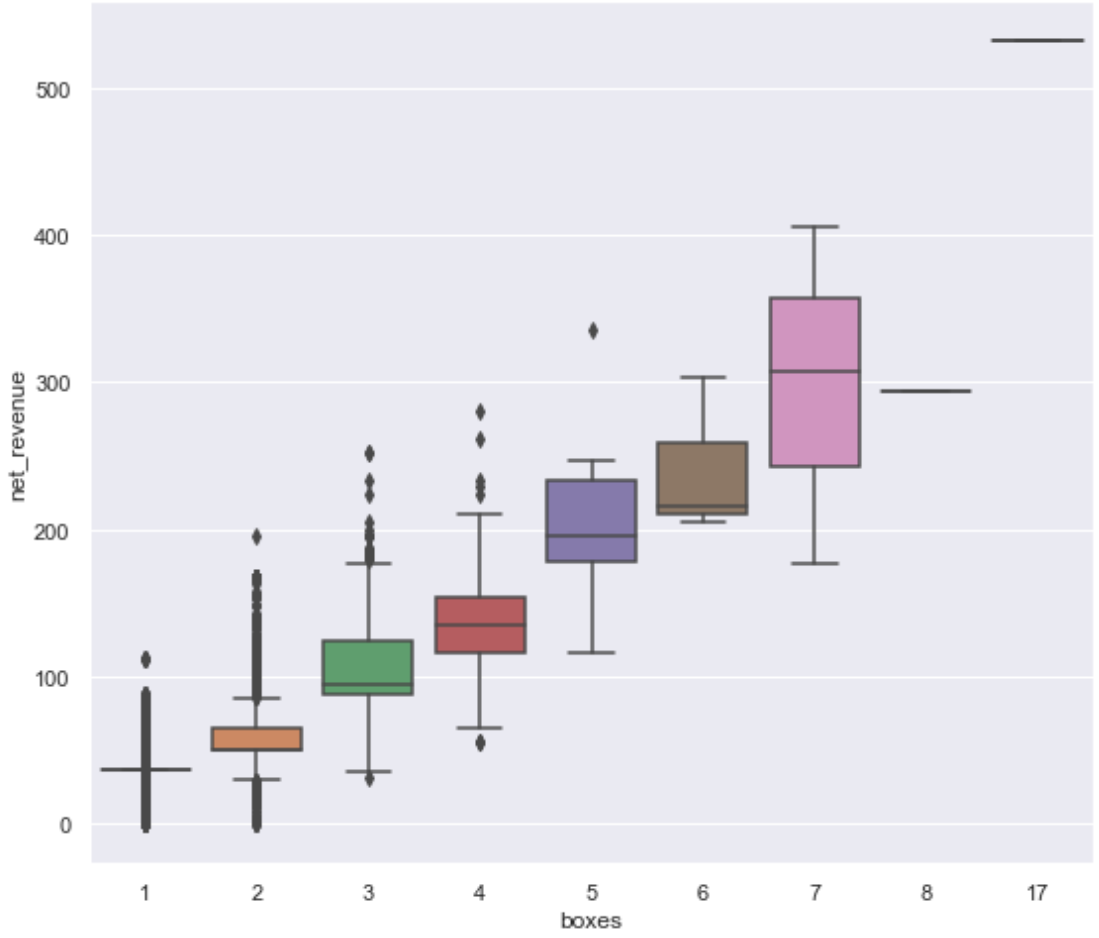
Atributos numéricos

Attributes	count	mean	std	min	25%	50%	75%	max	range	skew	kurtosis
customer_id	715875.0	323664.862498	186136.720066	103.0	159325.0	289541.0	476431.0	746721.0	746618.0	0.403586	-1.048270
year	715875.0	2014.196999	0.754568	2013.0	2014.0	2014.0	2015.0	2015.0	2.0	-0.342490	-1.179912
net_revenue	715875.0	36.415437	13.646022	0.0	37.0	37.0	37.0	532.0	532.0	0.526808	10.403764
gross_revenue	715875.0	41.430585	9.752502	14.0	37.0	37.0	40.0	532.0	518.0	3.435058	36.807084
boxes	715875.0	1.034697	0.190199	1.0	1.0	1.0	1.0	17.0	16.0	6.621214	112.525487
weekofyear_num	715875.0	27.491899	15.105624	1.0	14.0	28.0	41.0	53.0	52.0	-0.061525	-1.238845

Conhecimento sobre os dados

Atributos categóricos

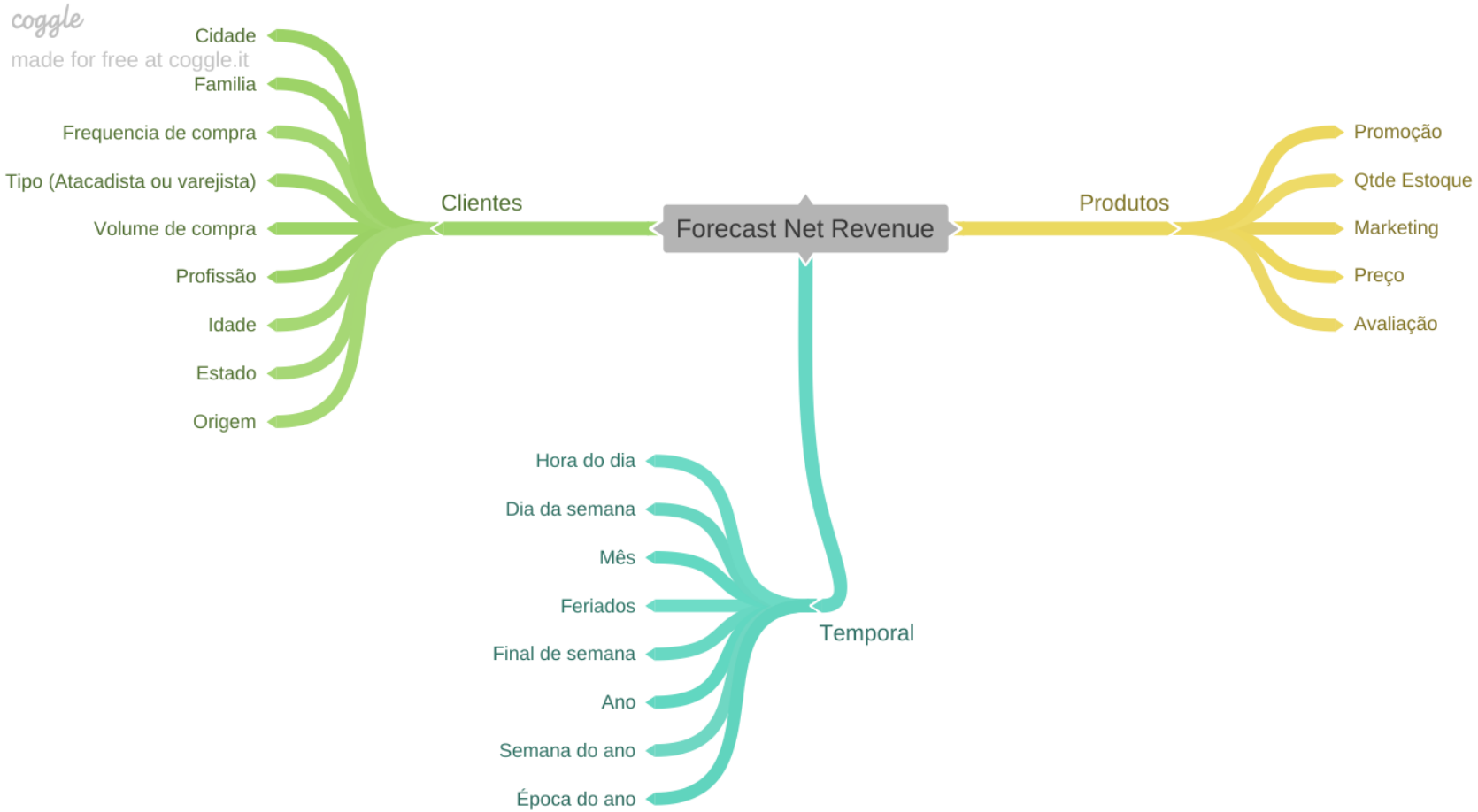




Preparação dos dados

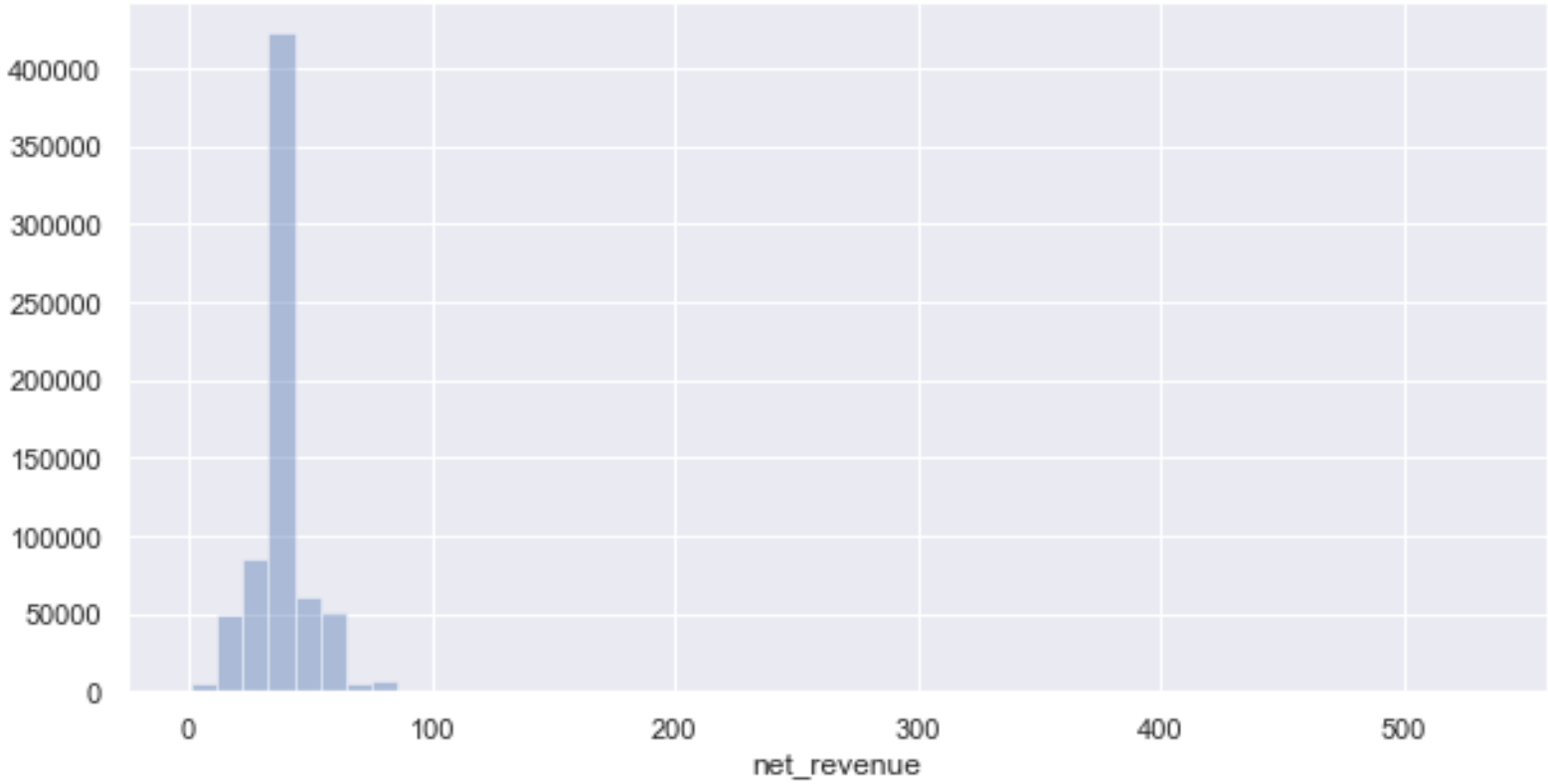
Possíveis features

- Quais dados complementares você julga que seriam importantes para incrementar suas análises em relação a resultados de vendas e análises sobre os clientes?



EDA – Exploratory Data Analysis

Univariada - Variável target



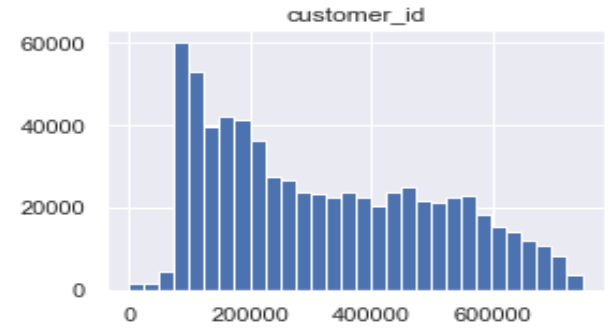
Obs: A maioria dos algoritmos de ML foram performados com algumas condições, e uma delas, normalmente, é a curva de distribuição da variável target, de forma que quanto mais se aproxima de uma variável normal, melhor é o seu resultado.

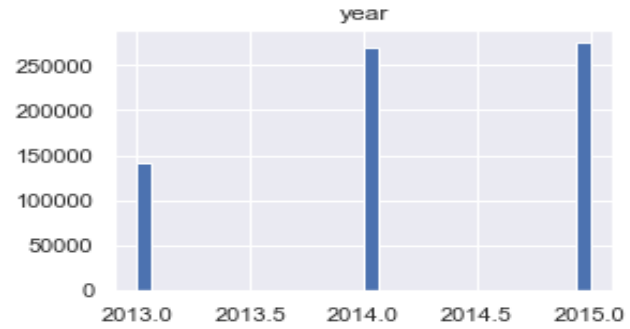
Por isso, precisamos transformar a variável target de forma que corresponda a uma curva normal ou pelo menos tenha uma aparência semelhante.

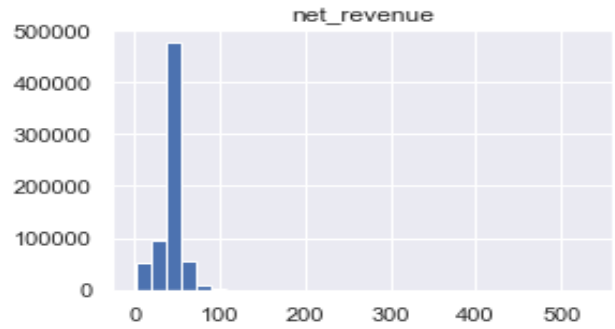
Existem várias técnicas de transformação, como por exemplo o uso do log.

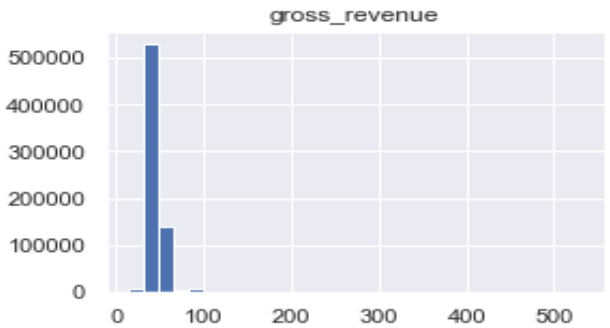
EDA – Exploratory Data Analysis

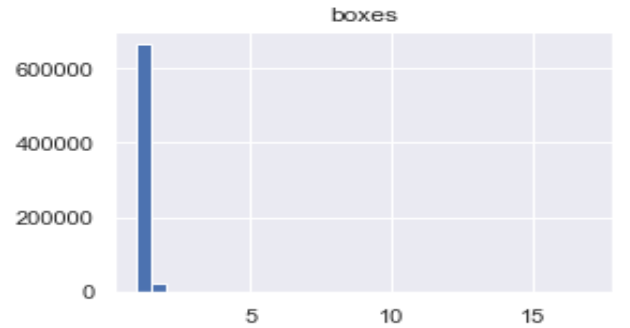
Univariada - Variáveis numéricas

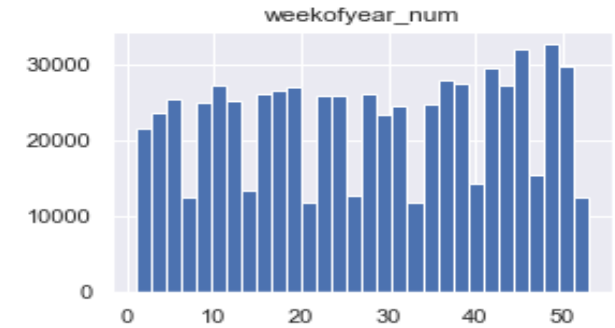


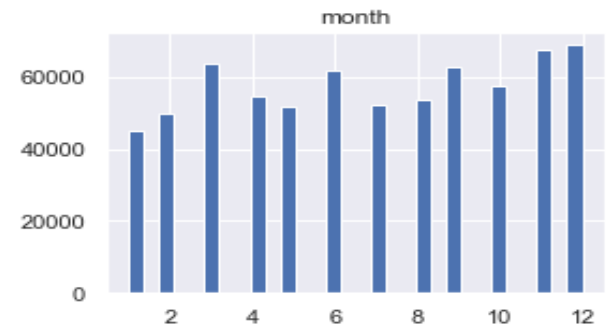


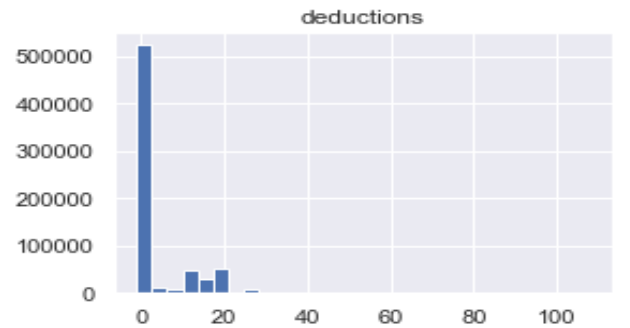






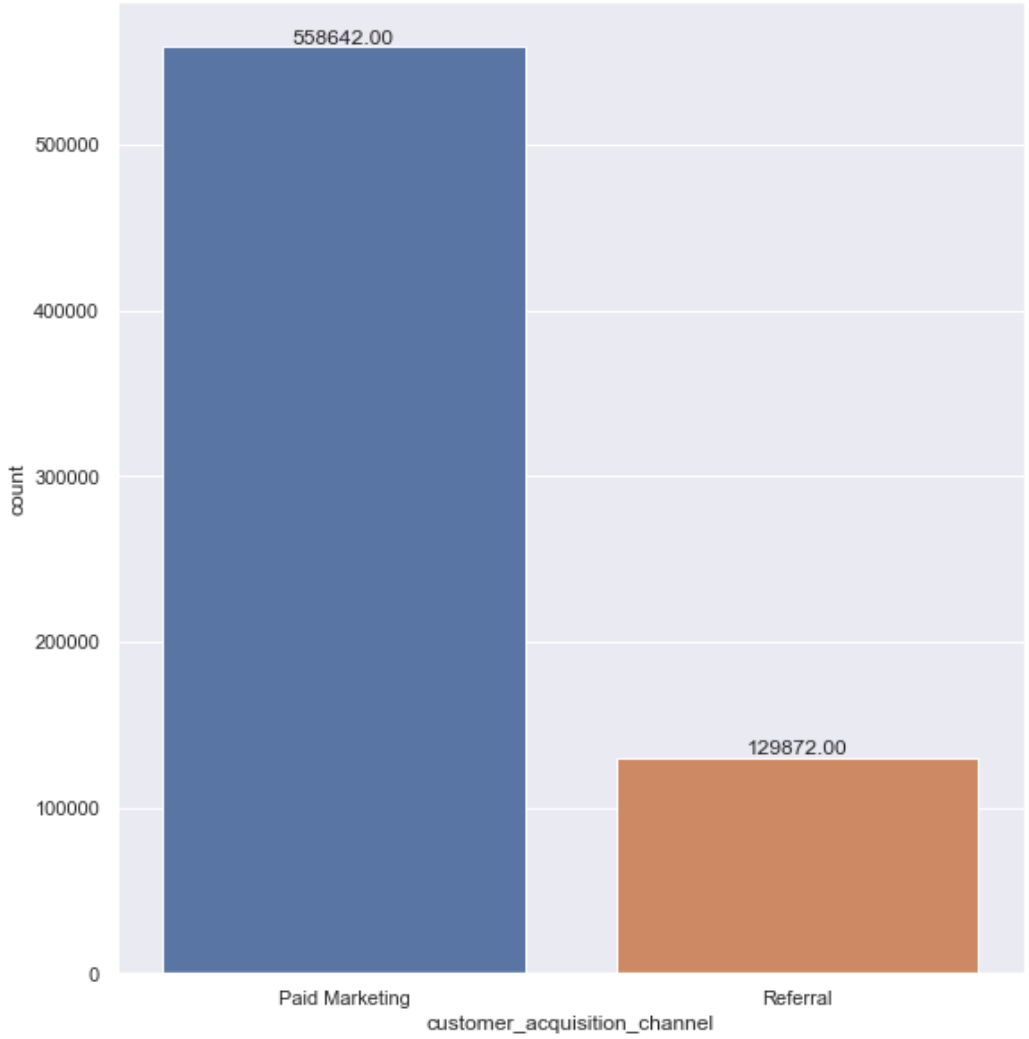


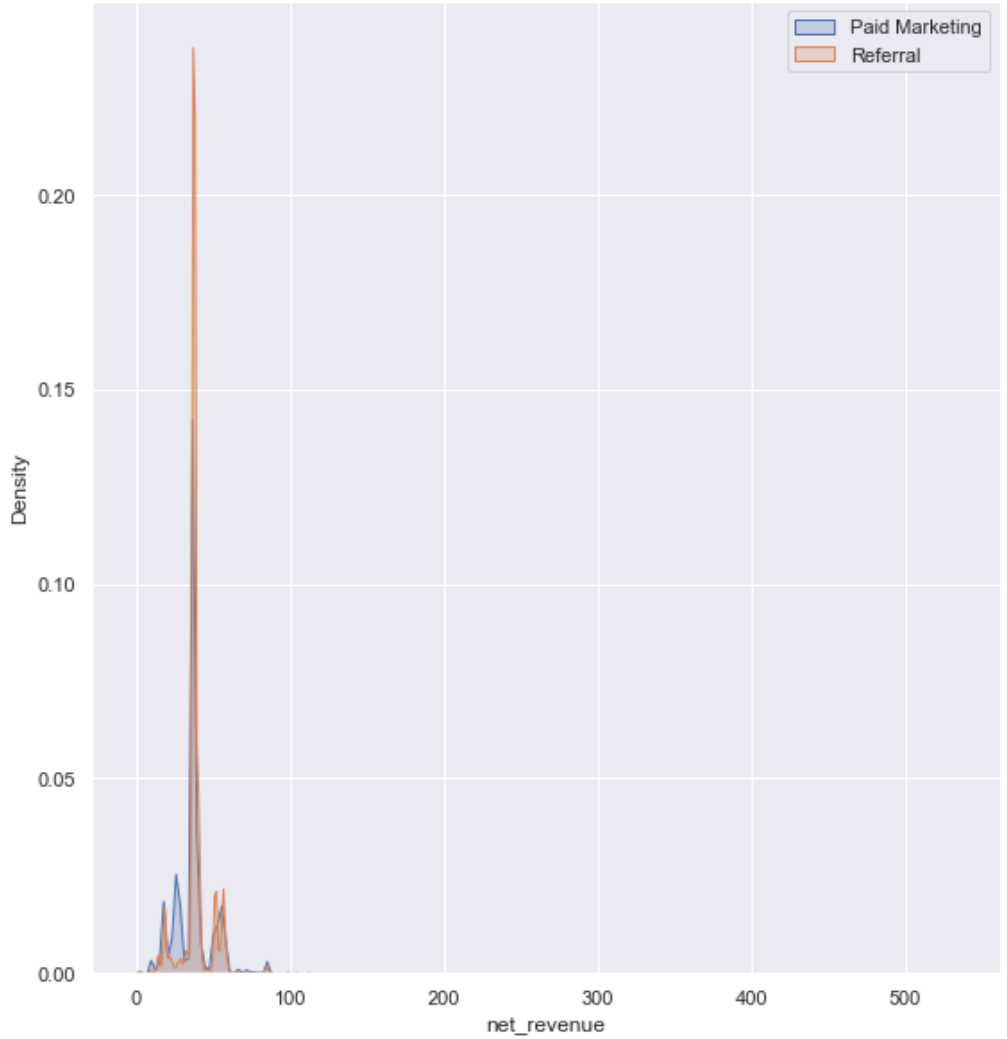




EDA – Exploratory Data Analysis

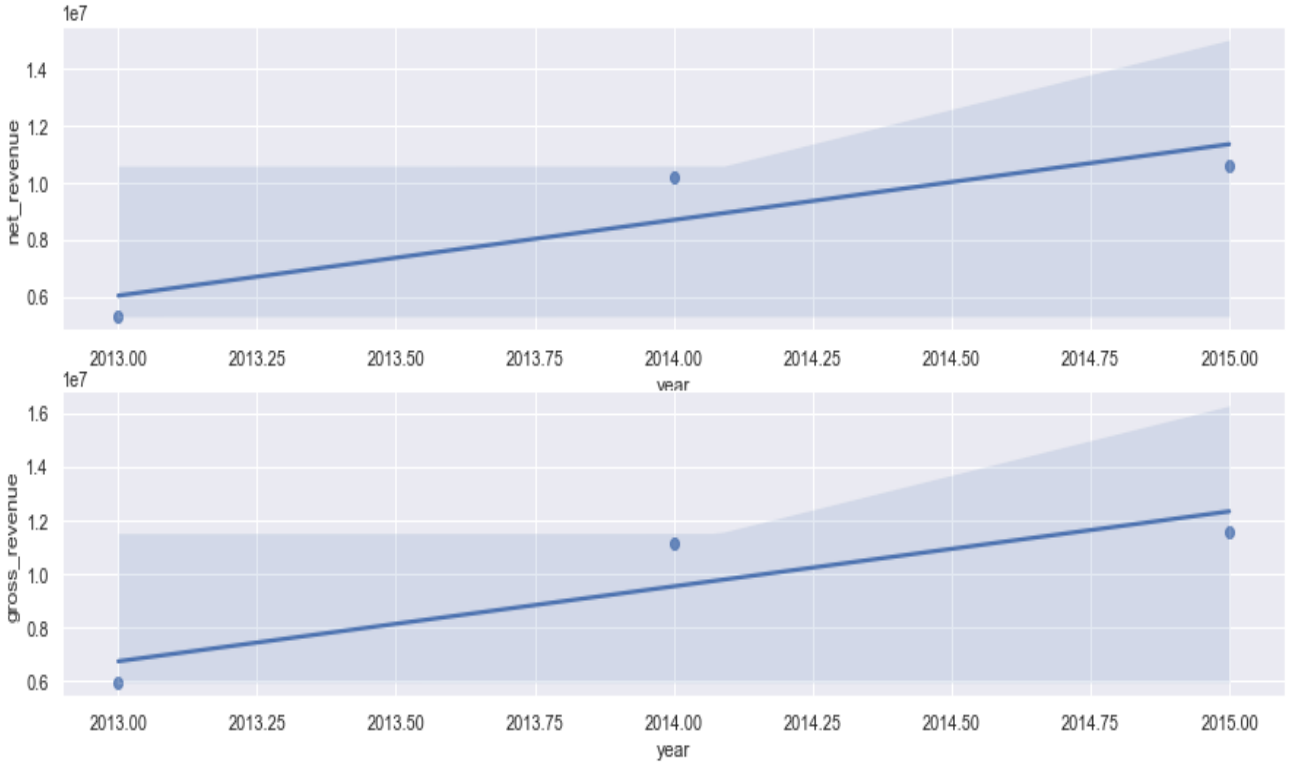
univariada - Variáveis categóricas

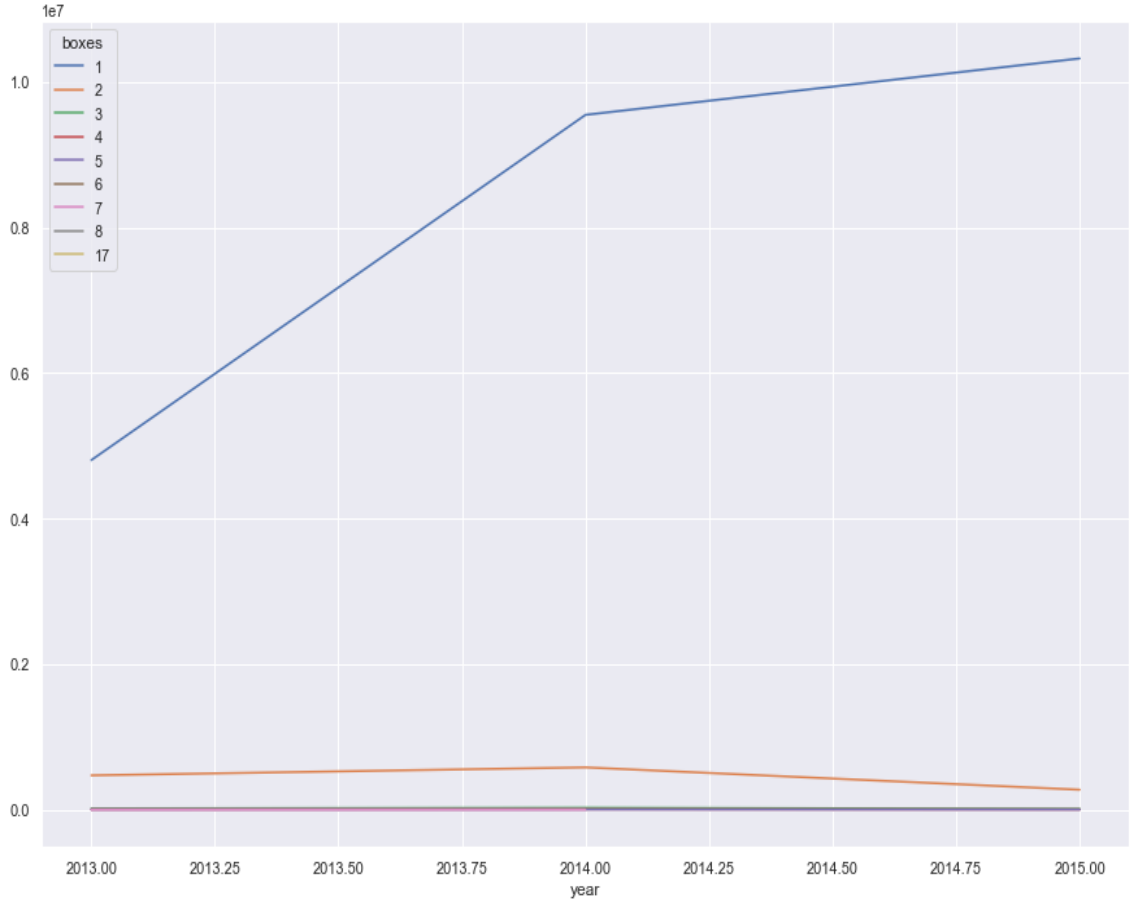




EDA – Exploratory Data Analysis

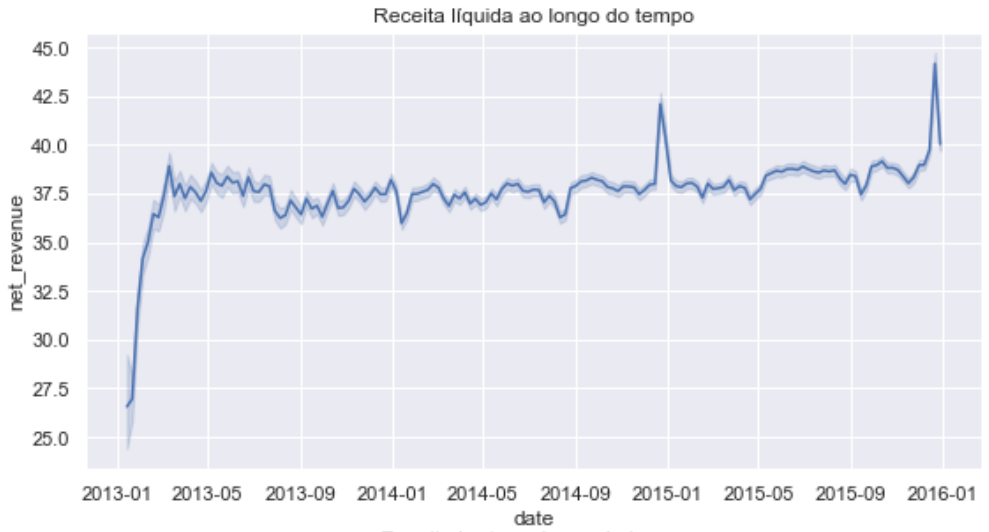
Análise bivariada

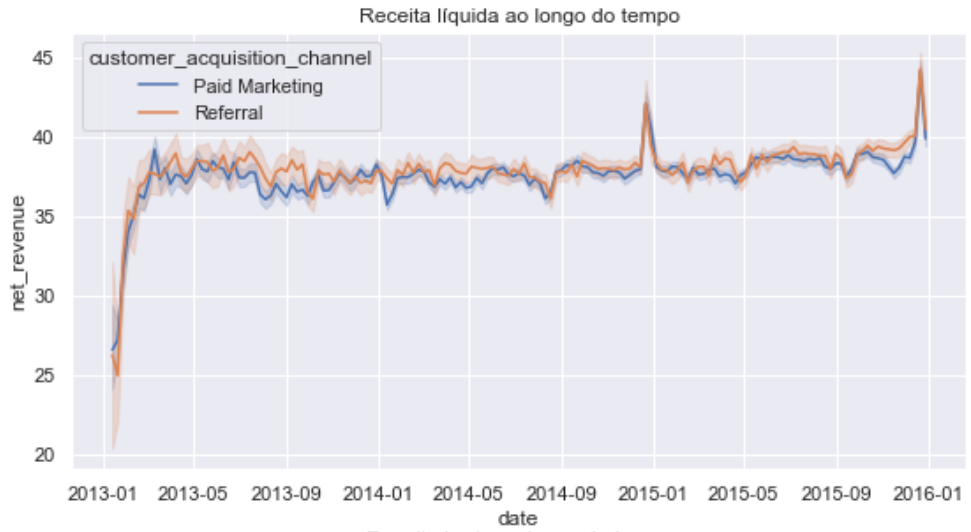


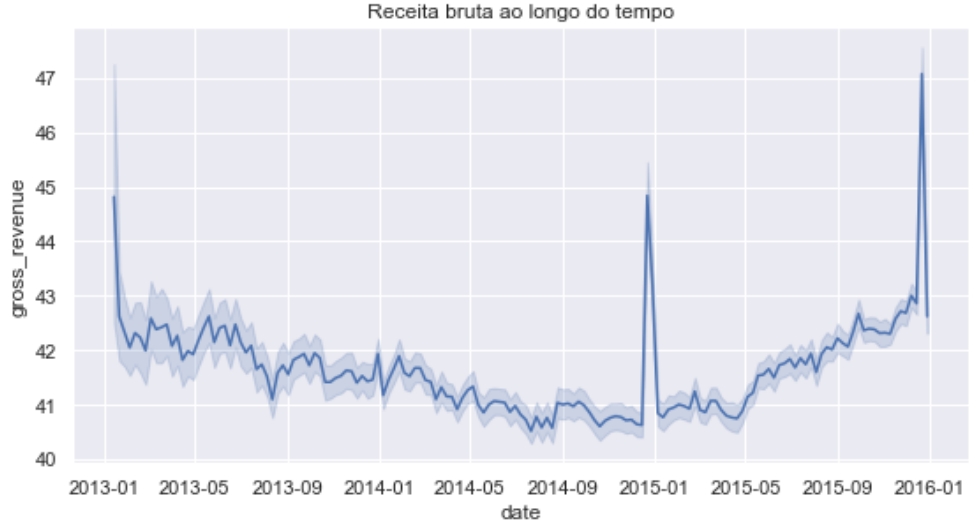


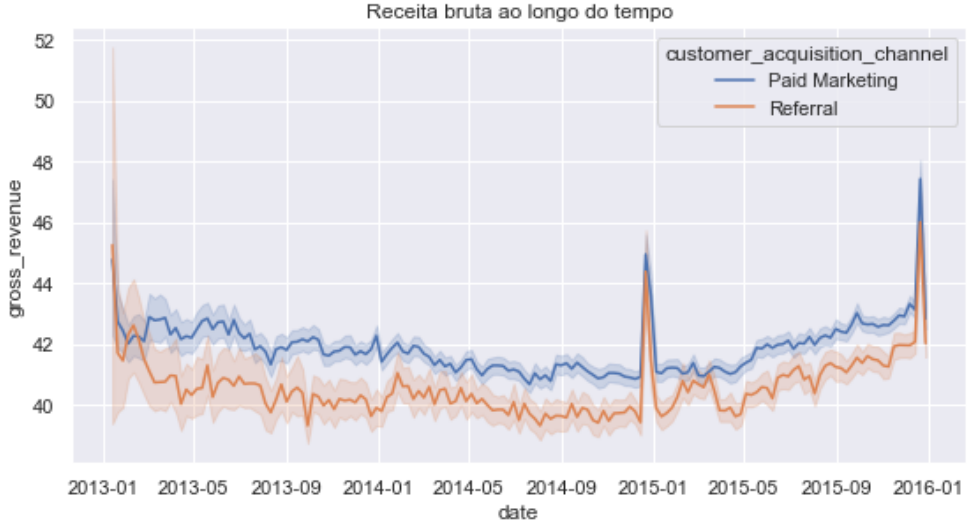
EDA – Exploratory Data Analysis

Análise bivariada



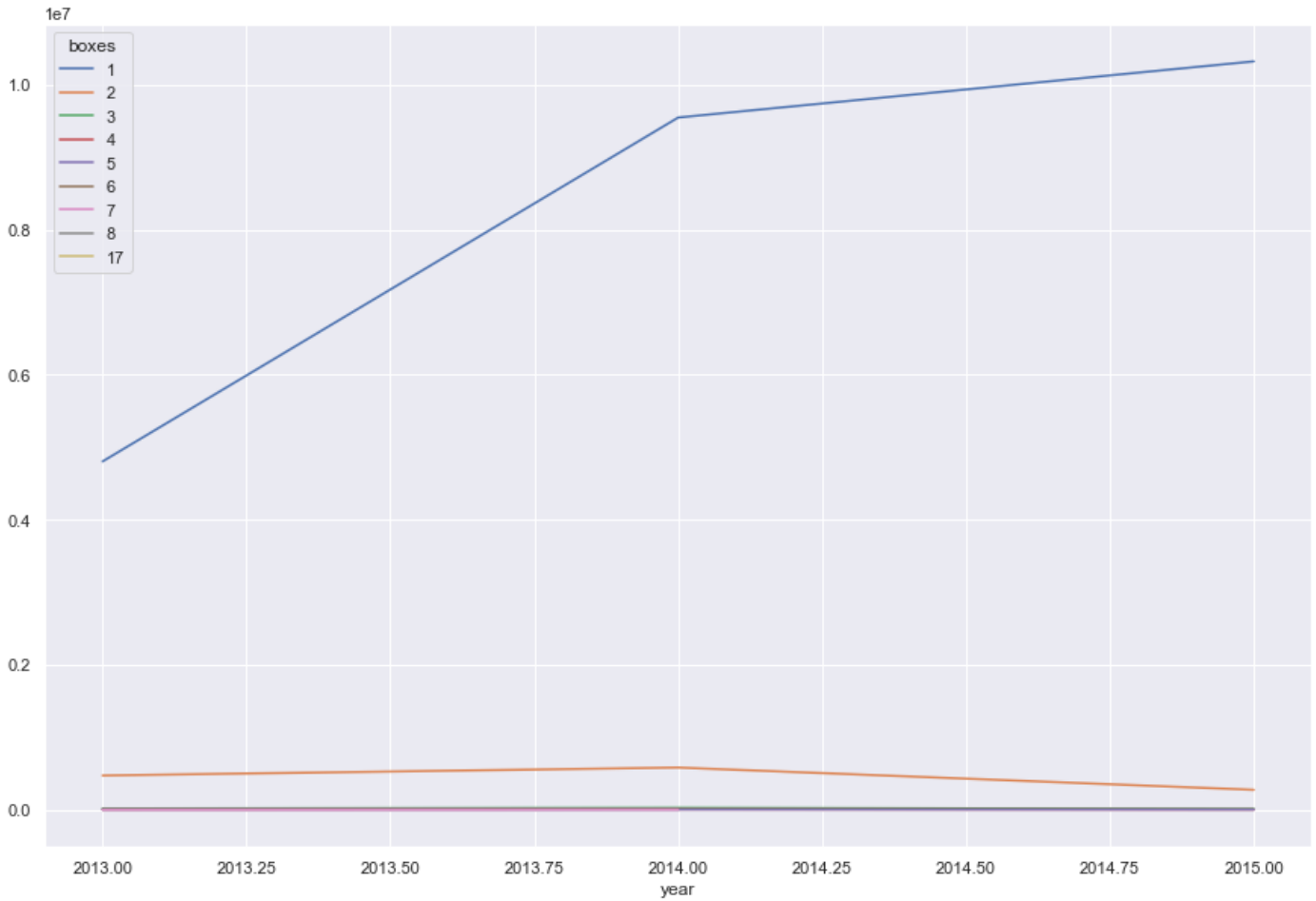






EDA – Exploratory Data Analysis

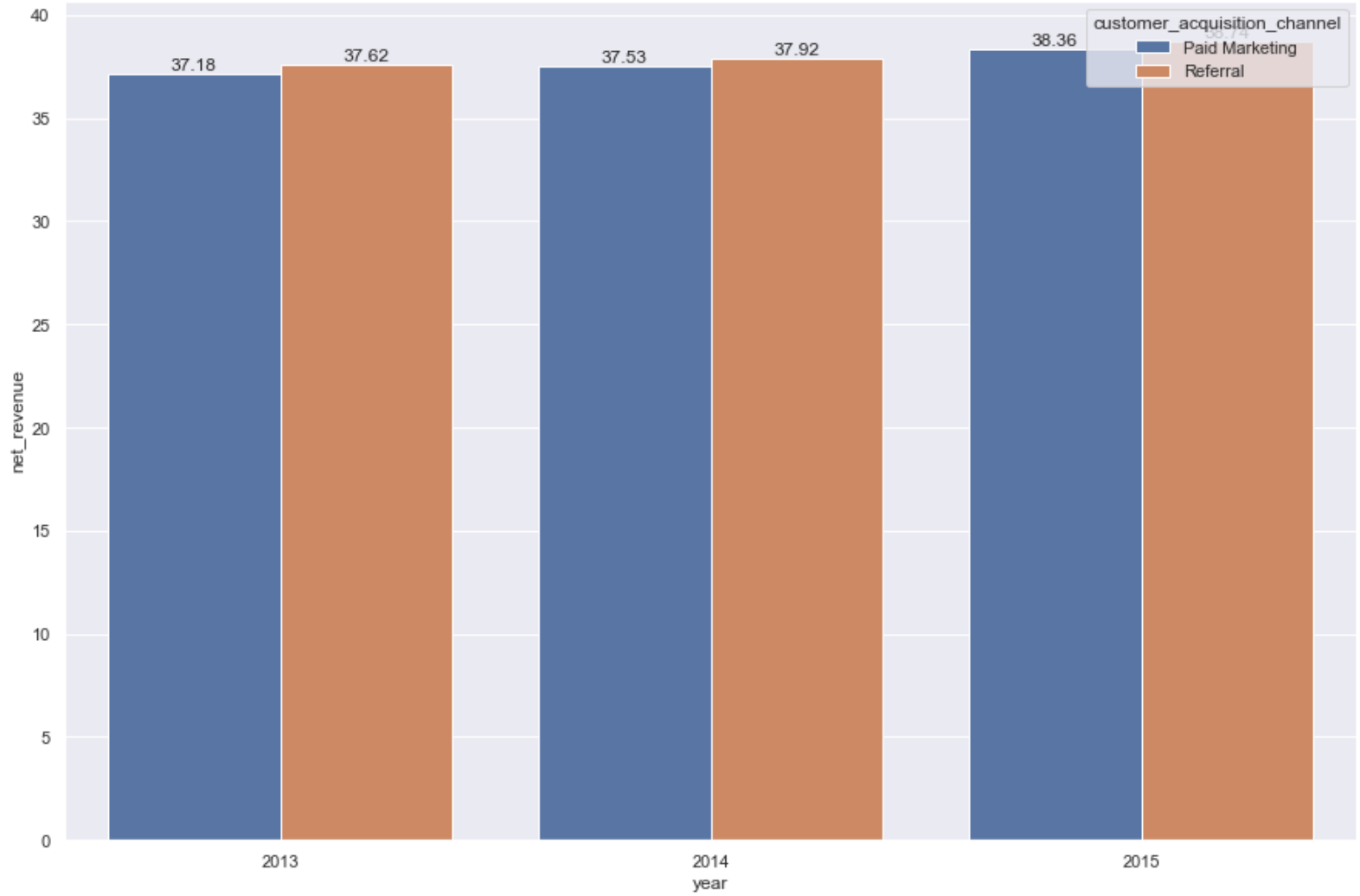
- Aponte resultados de vendas por ano: Gross Revenue, Net Revenue, Boxes (em Gráfico e Tabela)



year	boxes	net_revenue	gross_revenue
2013	1	4803558.0	5401141.0
2013	2	470482.0	523677.0
2013	3	17038.0	17995.0
2013	4	2217.0	2351.0
2013	5	1115.0	1170.0
2013	6	303.0	322.0
2013	7	583.0	602.0
2013	17	532.0	532.0
2014	1	9549856.0	10443559.0
2014	2	580355.0	640311.0
2014	3	25075.0	26519.0
2014	4	4626.0	4835.0
2014	5	946.0	946.0
2014	7	308.0	308.0
2014	8	294.0	294.0
2015	1	10325228.0	11234870.0
2015	2	273228.0	290262.0
2015	3	10688.0	10960.0
2015	4	1822.0	1976.0
2015	5	226.0	263.0
2015	6	421.0	432.0

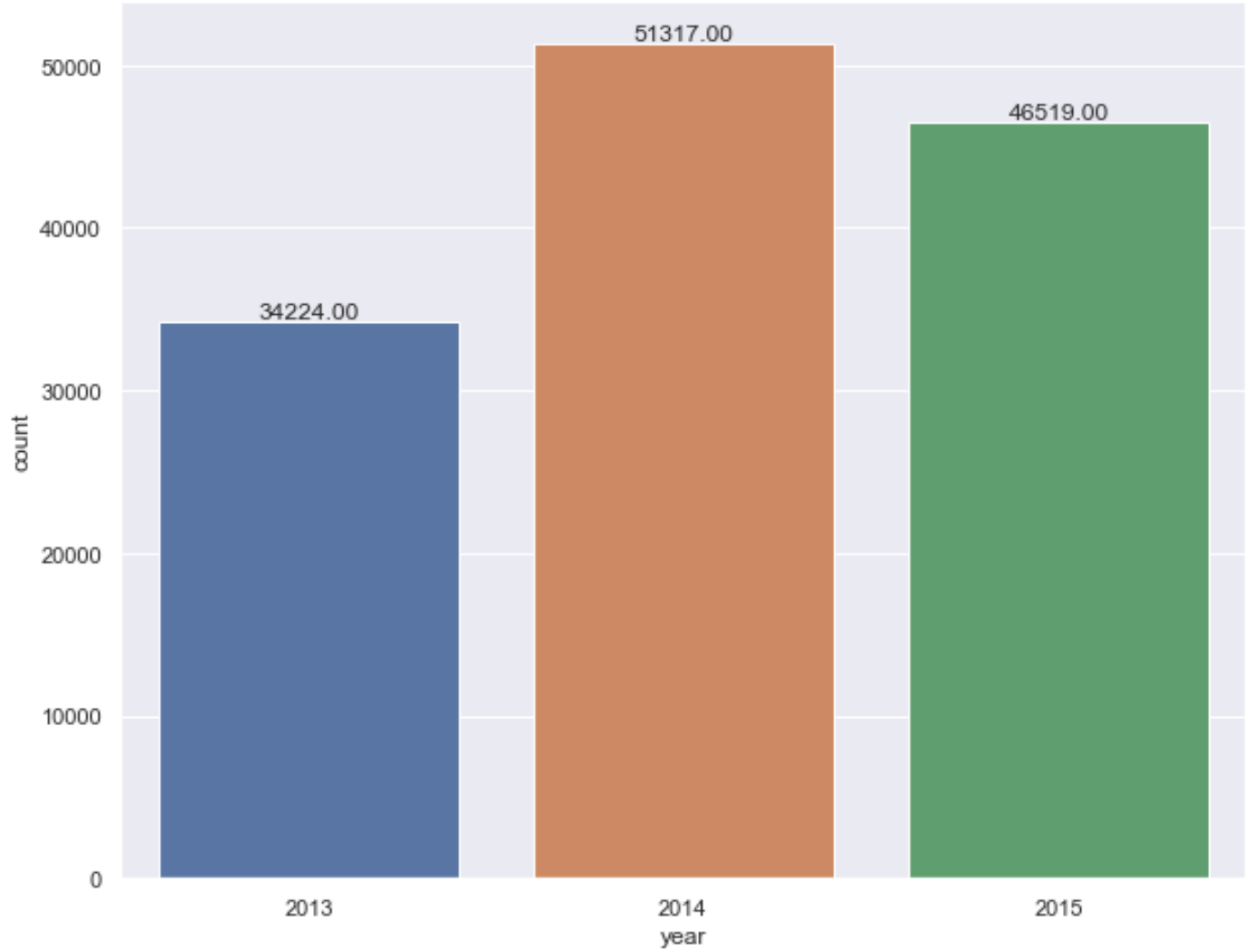
EDA – Exploratory Data Analysis

- Qual customer_acquision_channel teve maior Ticket Médio em 2013 e em 2015?



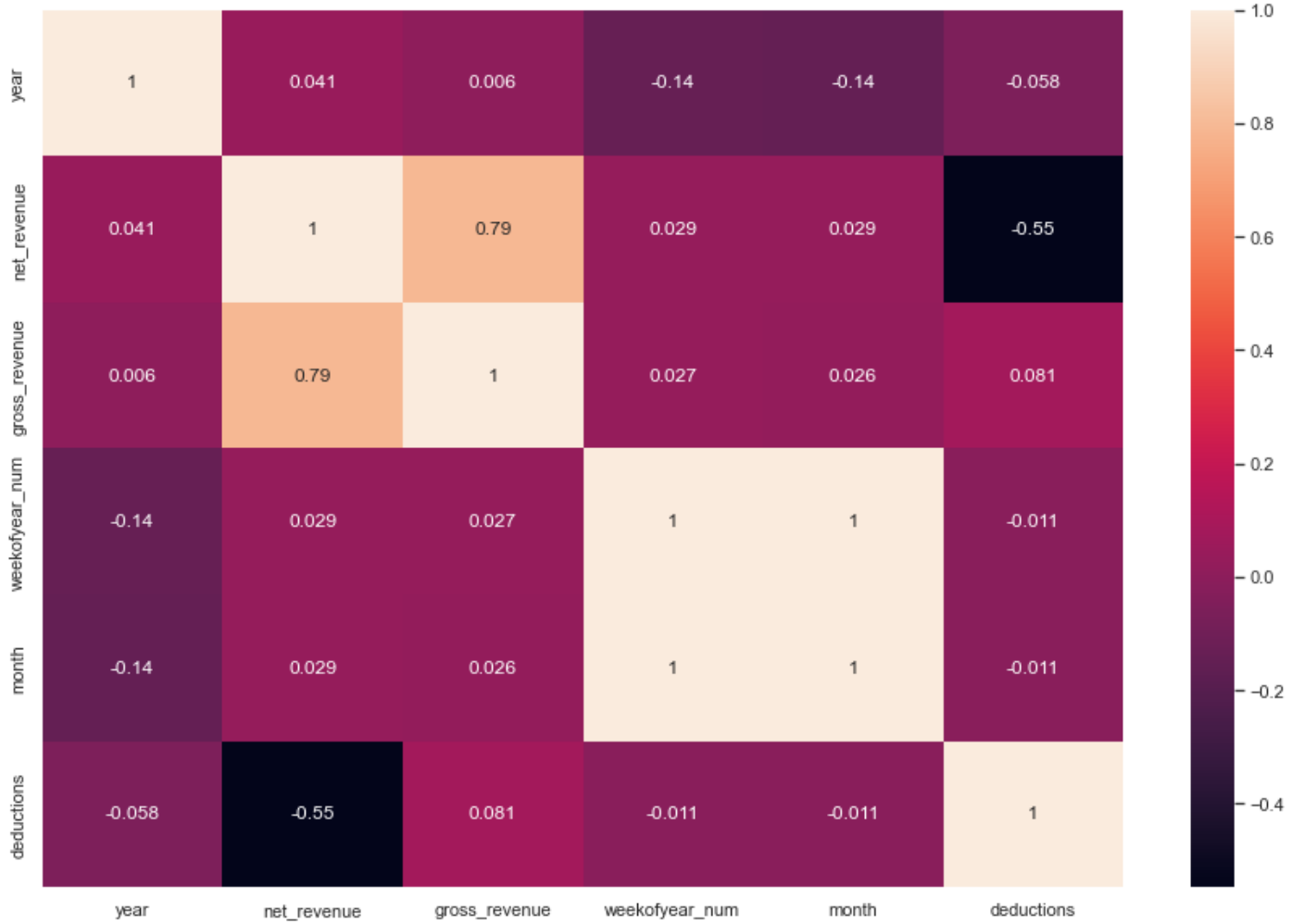
EDA – Exploratory Data Analysis

- Número de clientes únicos por Ano e comparativo desse resultado 2013x2015 (em gráfico e tabela)



EDA – Exploratory Data Analysis

Análise multivariada



Performance Model

Model Name	MAE	MAPE	RMSE
XGBoost Regressor	6.556673	0.249037	10.508128
Random Forest	6.553688	0.248369	10.508171
Linear Regression	6.630673	0.253155	10.635224
Average Model	6.950695	0.262839	11.140697
ARIMA Model	23611.044409	0.121334	29900.633820

scenarios	Values
predictions	R$ 5,317,009.00
worst_scenario	R$ 5,086,986.54
best_scenario	R$ 5,547,031.31

Performance Model

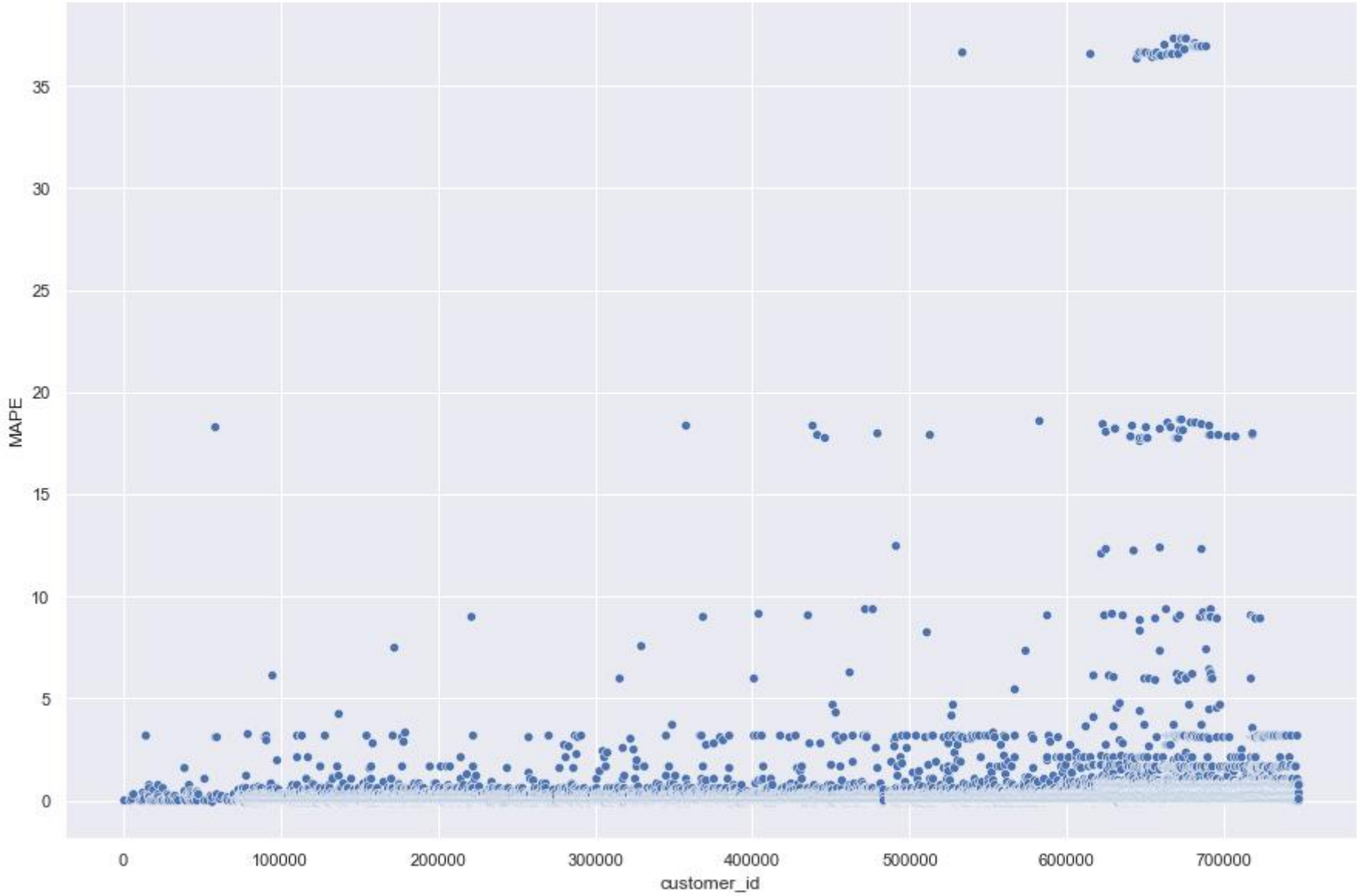
Top 10 piores clientes para predição

customer_id	predictions	worst_scenario	best_scenario	MAE	MAPE
667620	38.321270	1.0	75.642540	37.321270	37.321270
672403	38.321270	1.0	75.642540	37.321270	37.321270
674792	38.321270	1.0	75.642540	37.321270	37.321270
674322	38.321270	1.0	75.642540	37.321270	37.321270
671981	38.321270	1.0	75.642540	37.321270	37.321270
671963	38.321270	1.0	75.642540	37.321270	37.321270
680180	38.109562	1.0	75.219124	37.109562	37.109562
680881	38.109562	1.0	75.219124	37.109562	37.109562
680032	38.109562	1.0	75.219124	37.109562	37.109562
661100	38.028324	1.0	75.056648	37.028324	37.028324

Top 10 melhores clientes para predição

customer_id	predictions	worst_scenario	best_scenario	MAE	MAPE
325361	36.988667	36.977333	37.0	0.011333	0.000306
234492	36.988667	36.977333	37.0	0.011333	0.000306
391661	36.988667	36.977333	37.0	0.011333	0.000306
123313	36.988667	36.977333	37.0	0.011333	0.000306
157872	36.988667	36.977333	37.0	0.011333	0.000306
569232	36.988667	36.977333	37.0	0.011333	0.000306
77222	36.988667	36.977333	37.0	0.011333	0.000306
77000	36.988667	36.977333	37.0	0.011333	0.000306
309920	36.988667	36.977333	37.0	0.011333	0.000306
452002	36.988667	36.977333	37.0	0.011333	0.000306

Performance Model



Performance Model