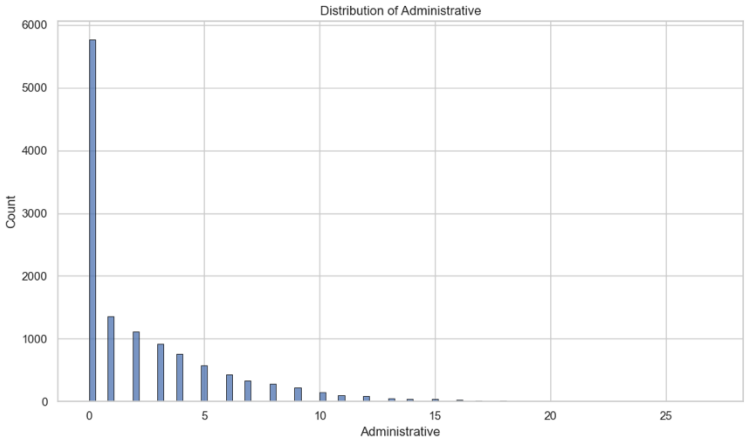
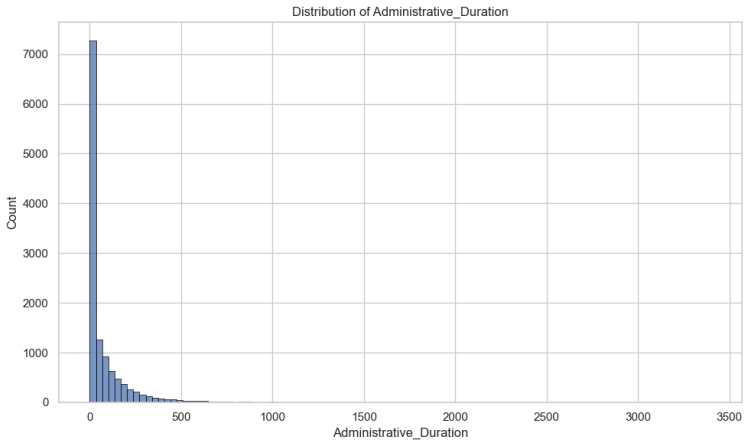
Tema 2 IA

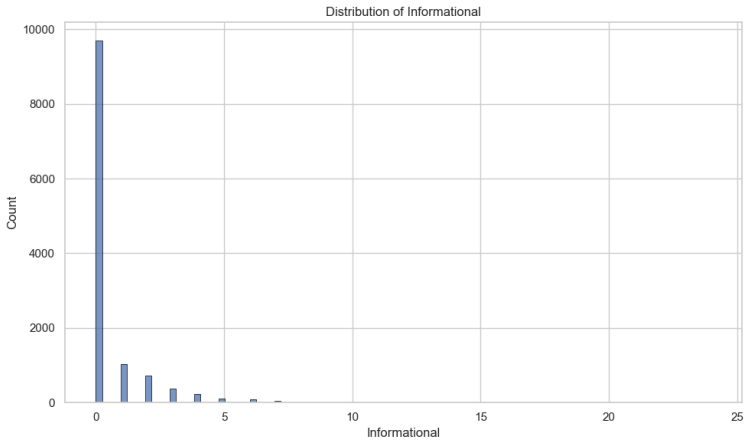
When does Santa’s shopping bring revenue?

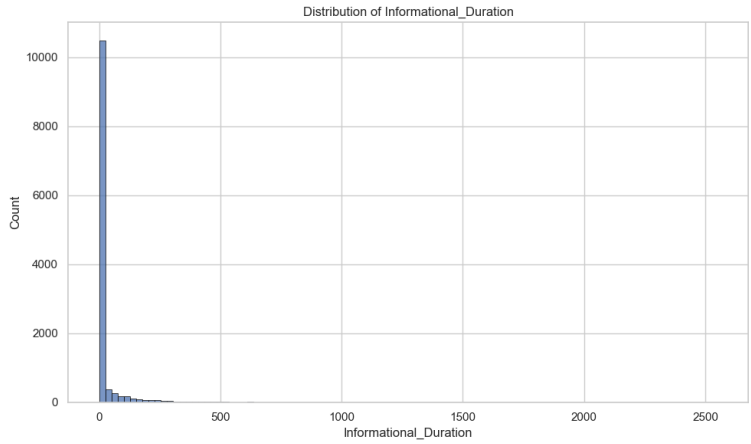
3.1. Explorarea datelor

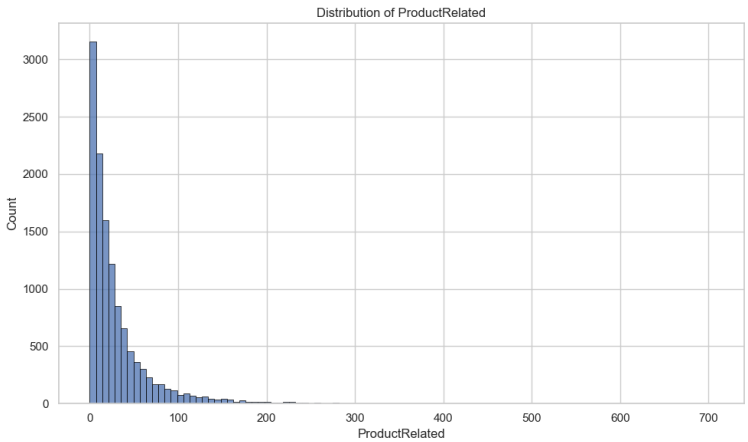
1. Analiza echilibrului de clase

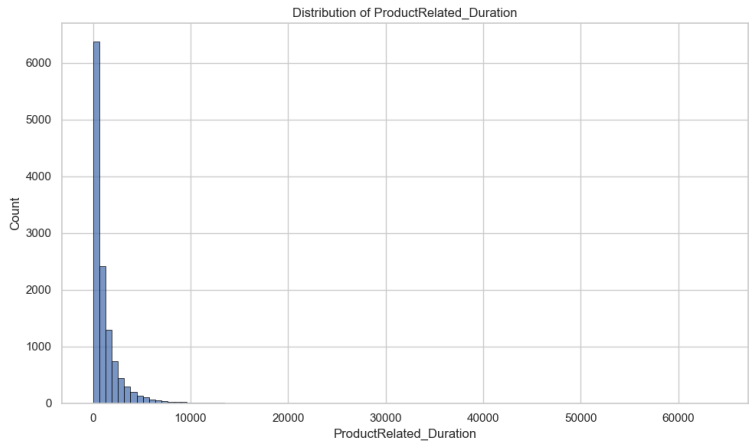


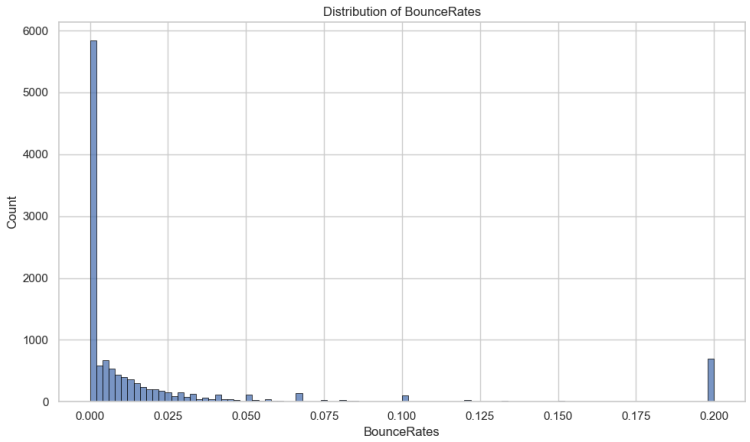


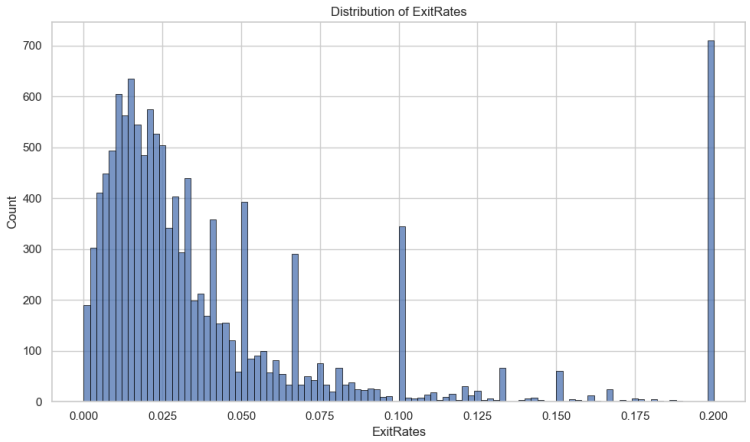


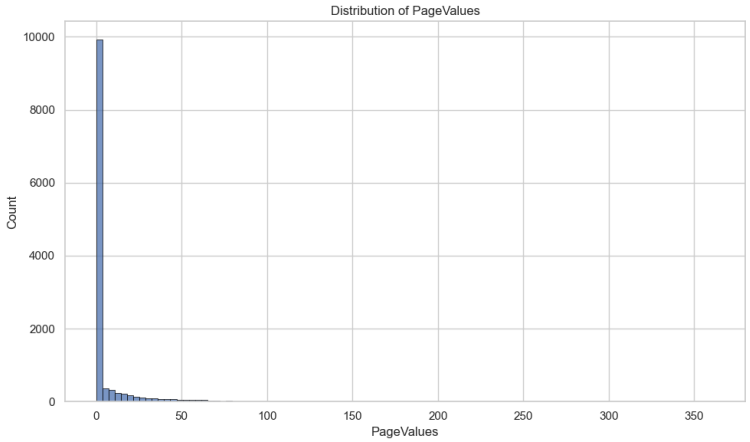


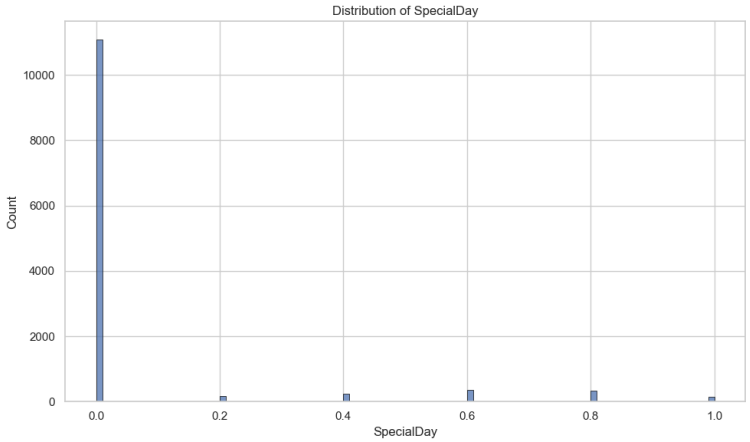


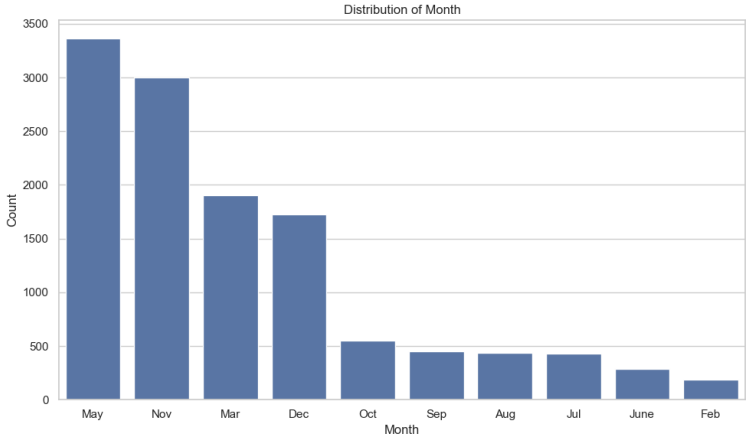


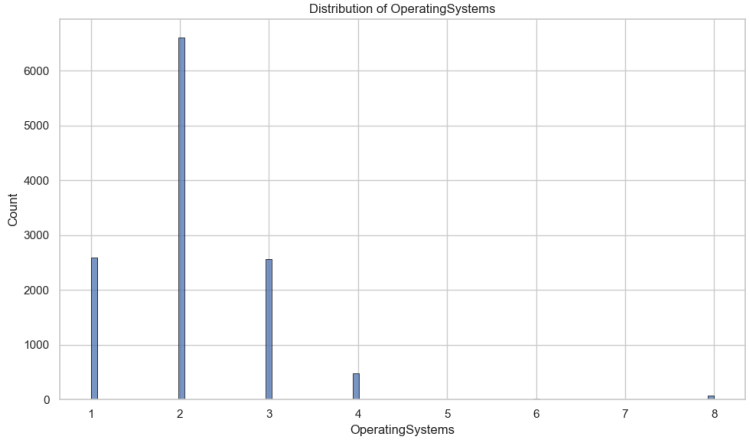

Distribution of BounceRates
Count
BounceRates

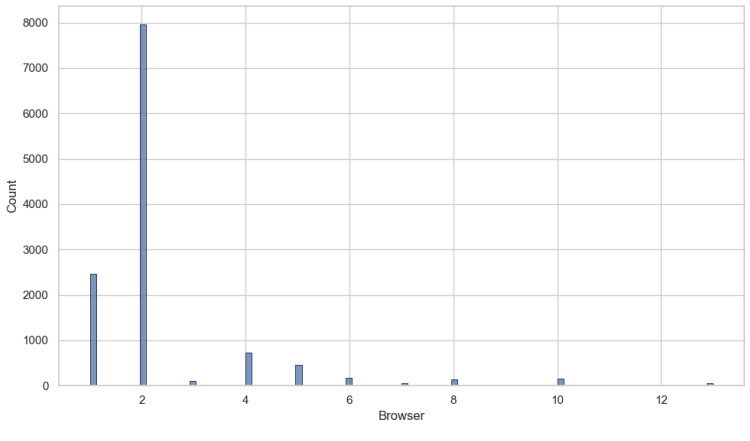

Distribution of ExitRates
Count
ExitRates

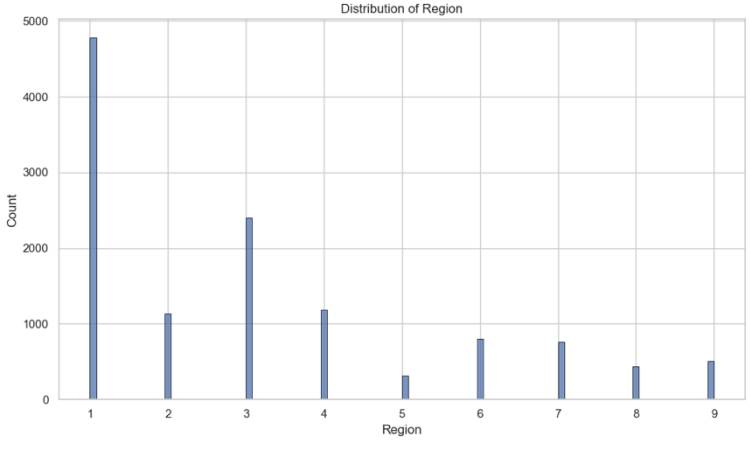

Distribution of PageValues
Count
PageValues

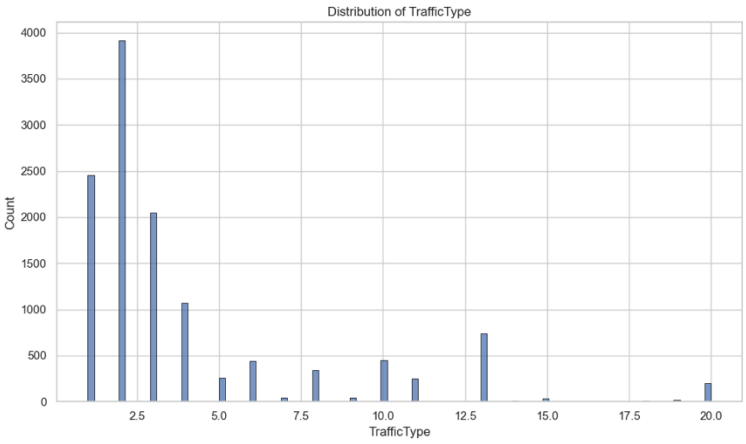

Distribution of SpecialDay
Count
SpecialDay

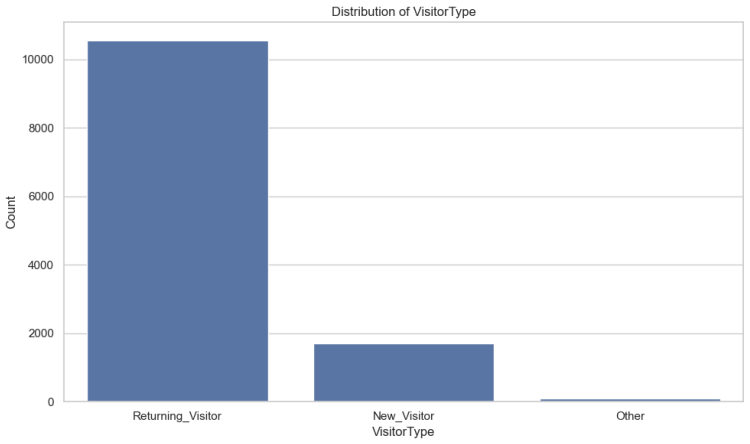

Distribution of Month
Count
May
Nov
Mar
Dec
Oct
Sep
Aug
Jul
June
Feb
Month

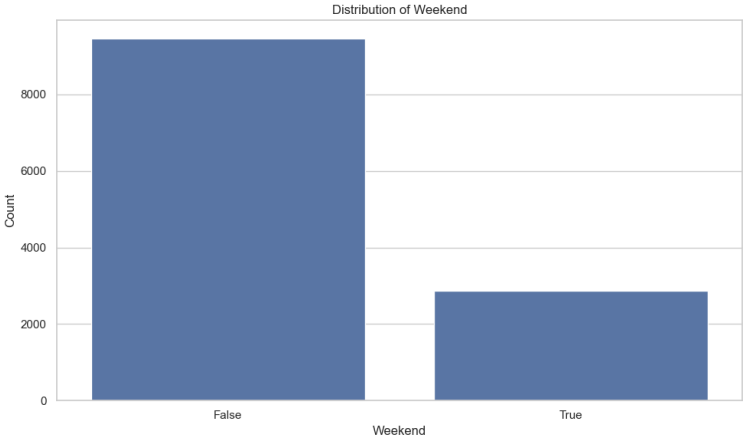

Distribution of OperatingSystems
Count
OperatingSystems

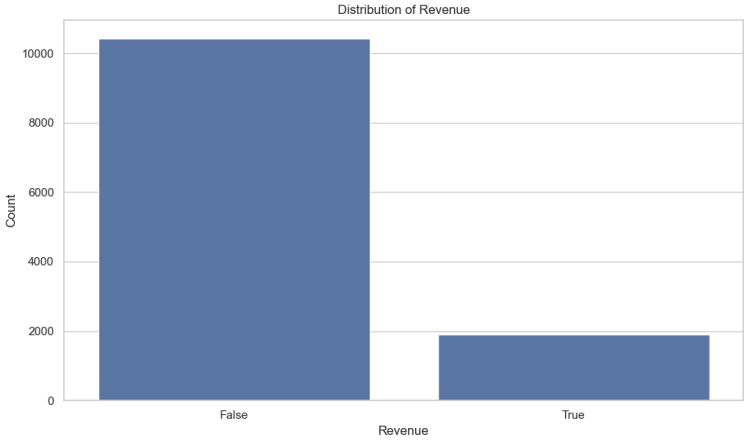

Distribution of Browser
Count
Browser


Distribution of Region
Count
Region

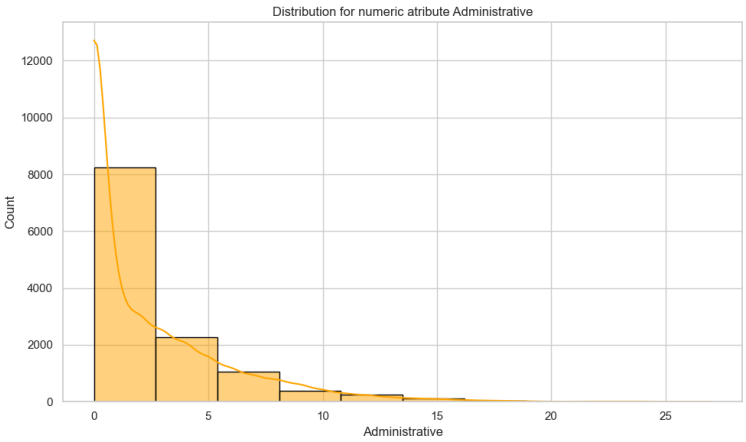


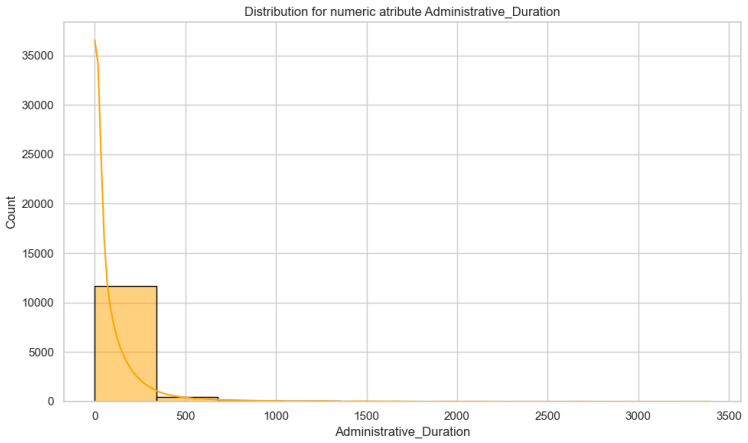


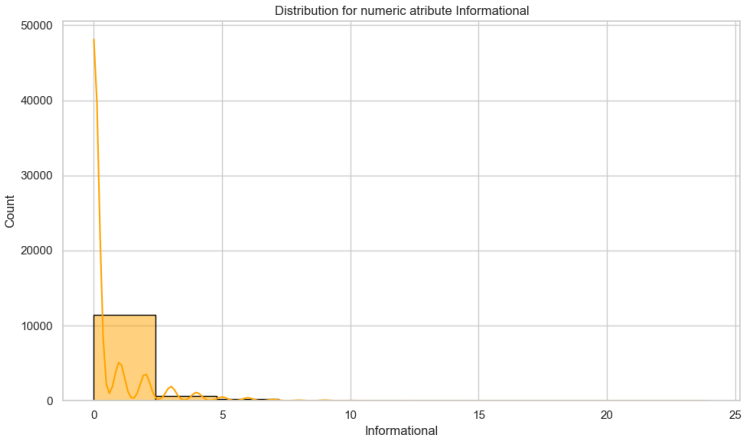


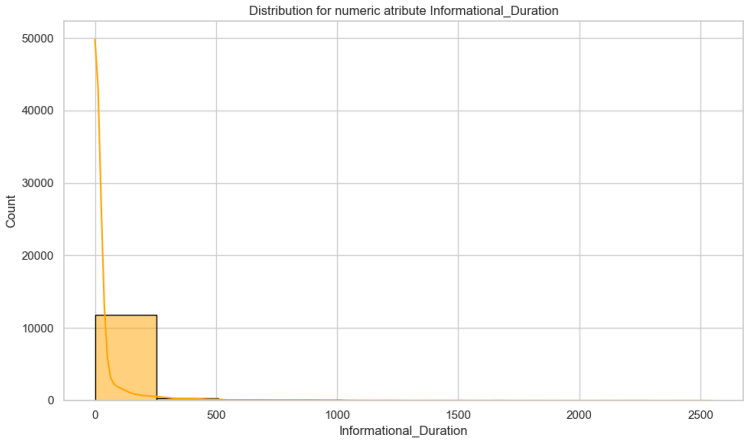


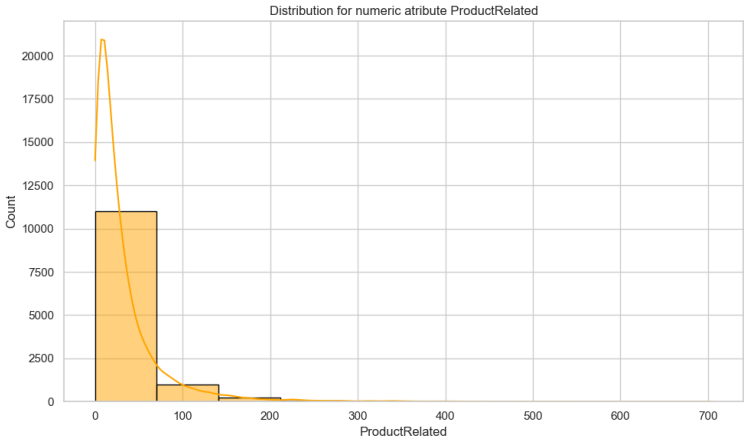
2. A. 1. Vizualizarea atributelor numerice

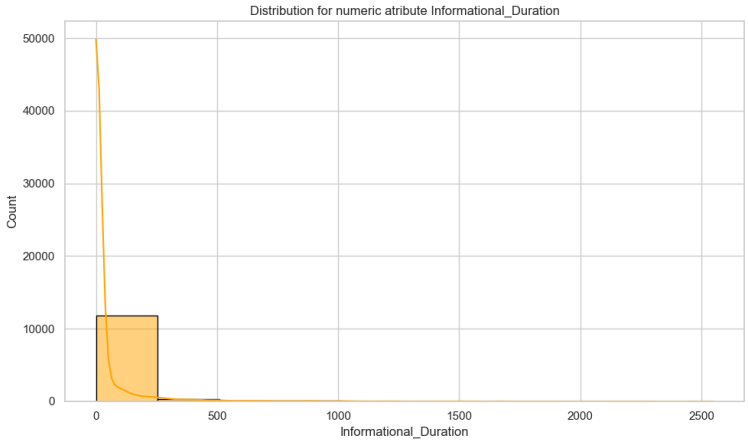


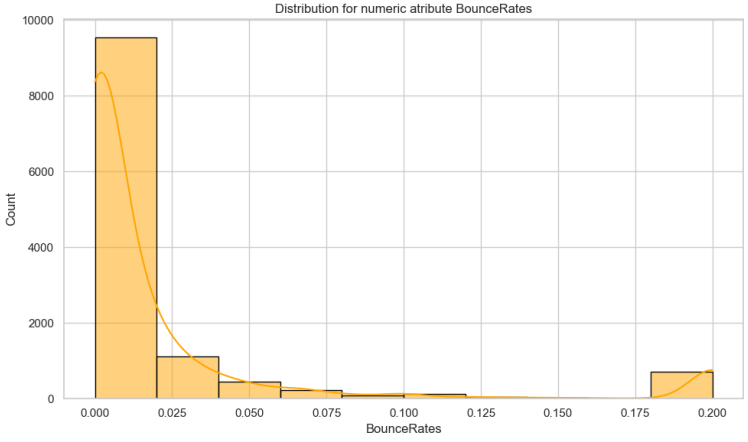


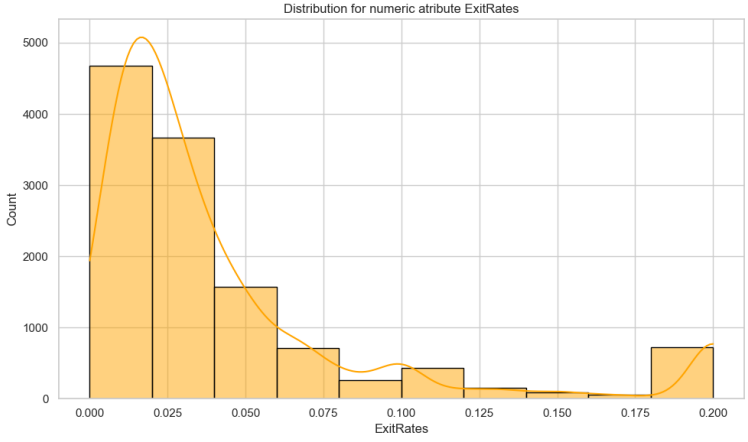


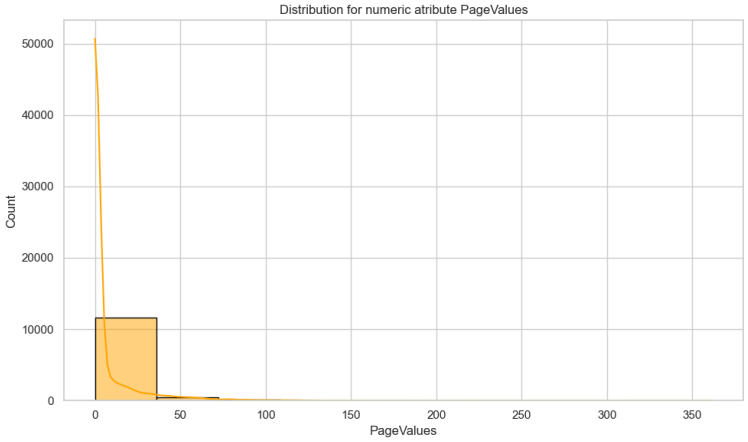


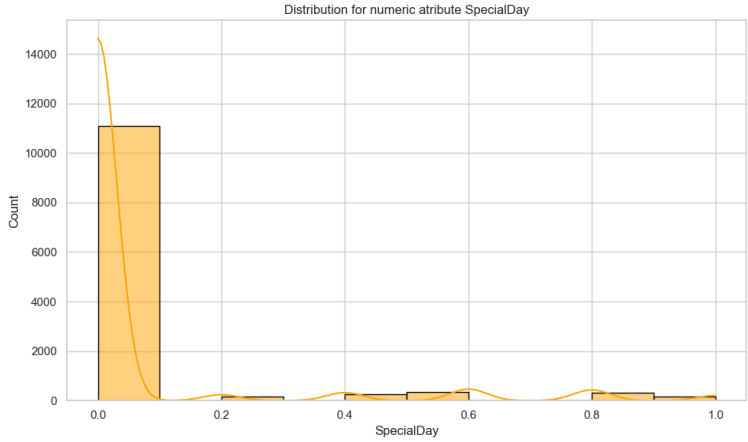

Distribution for numeric atribute ProductRelated
Count
ProductRelated

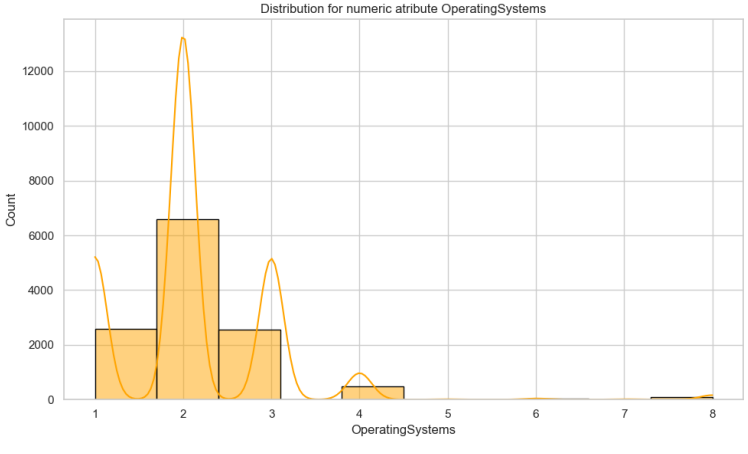

Distribution for numeric atribute Informational_Duration
Count
Informational_Duration

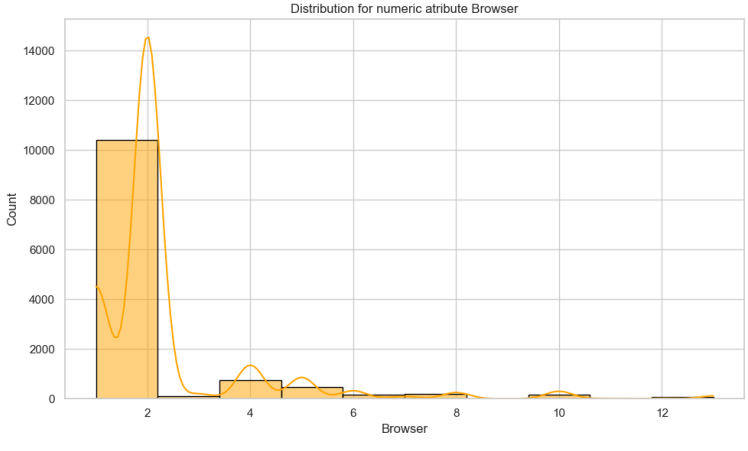

Distribution for numeric atribute BounceRates
Count
BounceRates

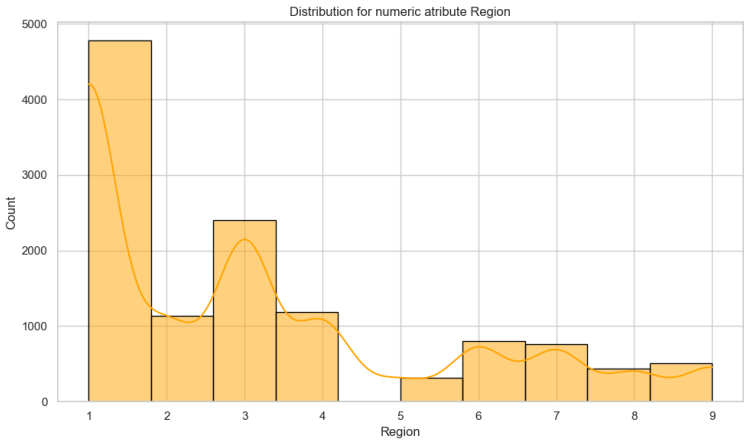

Distribution for numeric atribute ExitRates
Count
ExitRates

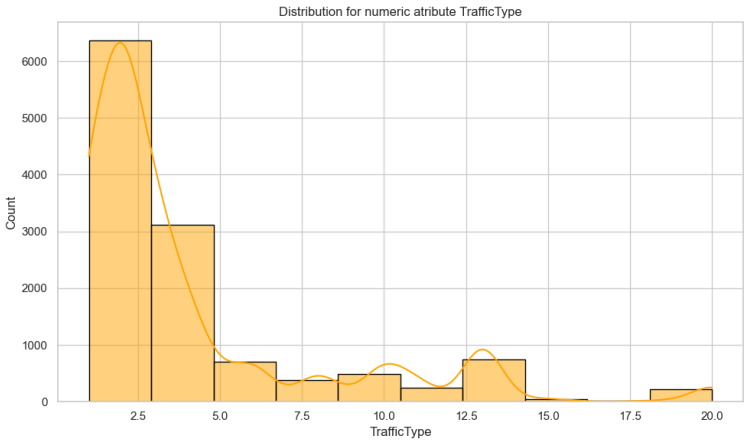

Distribution for numeric atribute PageValues
Count
PageValues


Distribution for numeric atribute SpecialDay
Count
SpecialDay

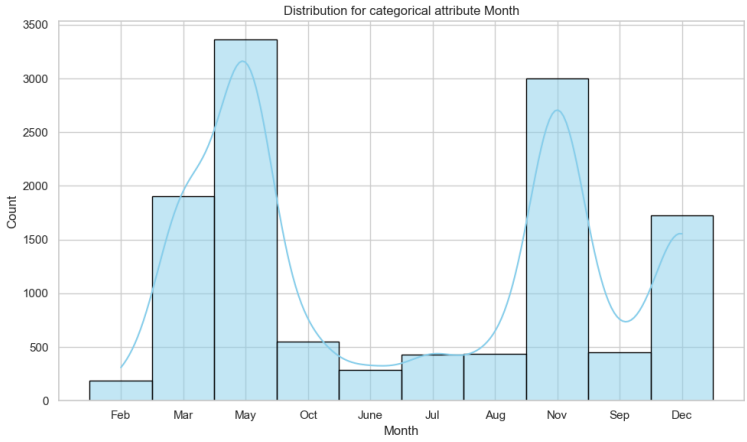

Distribution for numeric atribute OperatingSystems
Count
OperatingSystems

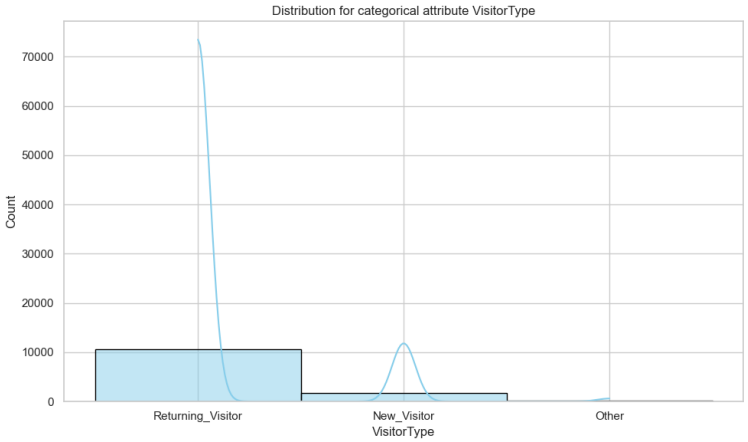

Distribution for numeric atribute Browser
Count
Browser

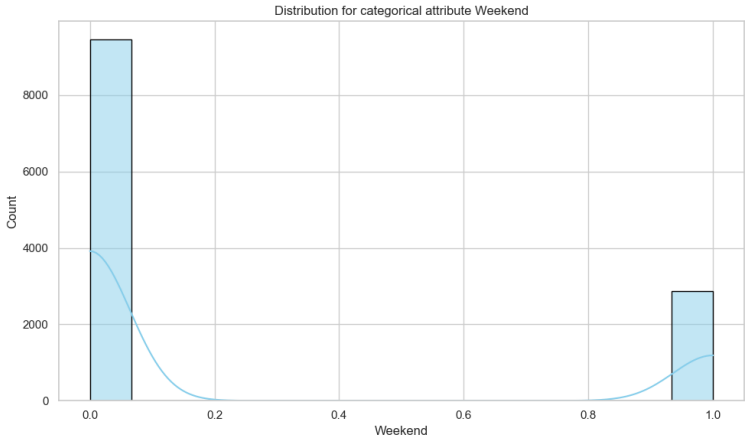


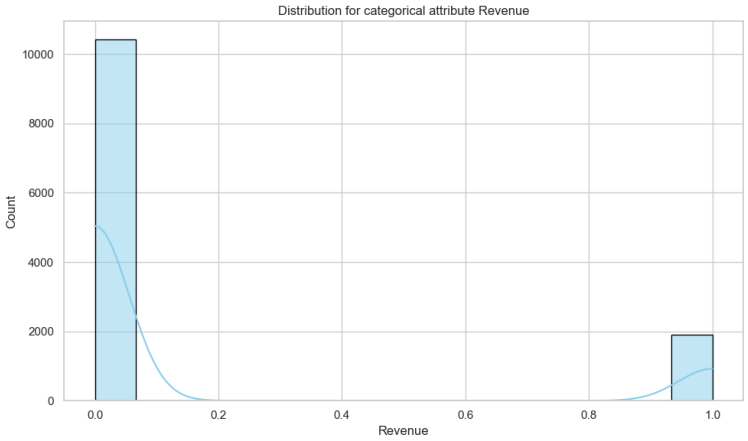


2. A. 2. Vizualizarea atributelor categorice



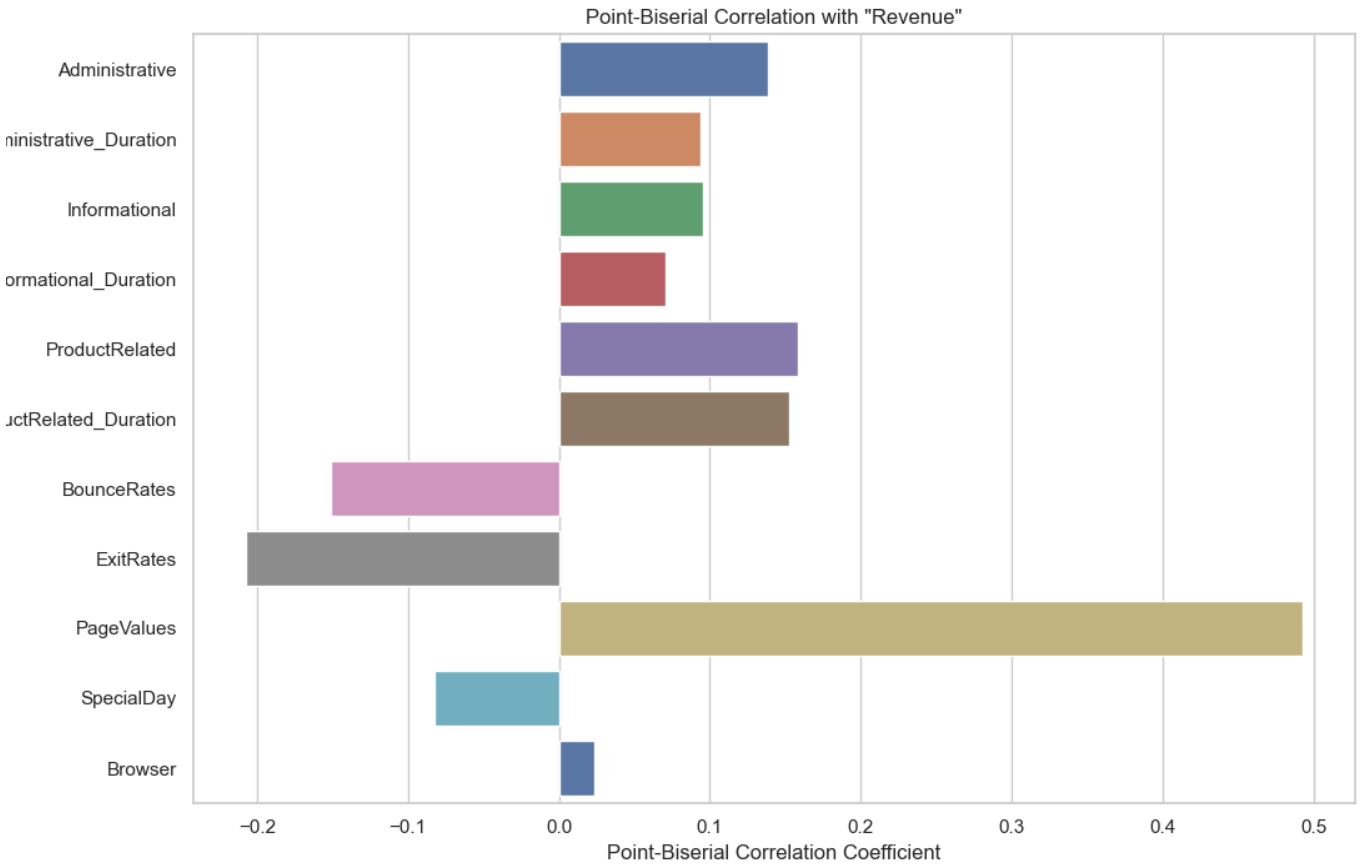






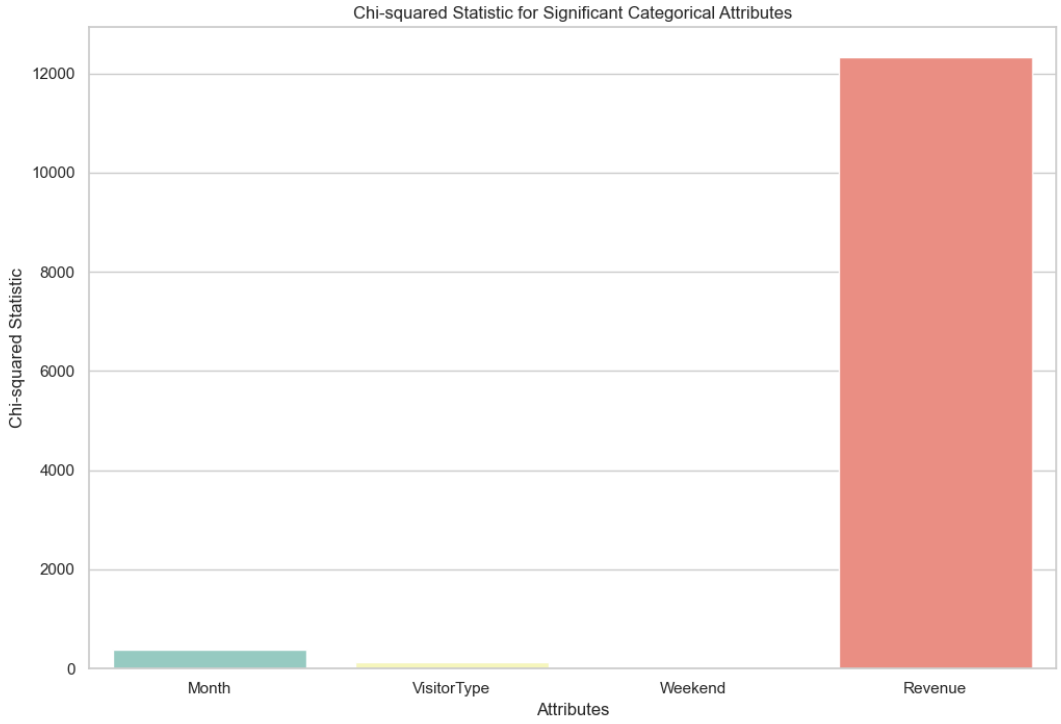
2. B. 1. Point-Biserial Corelation

Attribute	Point-Biserial Correlation	p-value
Administrative	0.138917	3.52E-54
Administrative_Duration	0.093587	2.15E-25
Informational	0.0952	3.17E-26
Informational_Duration	0.070345	5.28E-15
ProductRelated	0.158538	3.24E-70
ProductRelated_Duration	0.152373	6.12E-65
BounceRates	-0.150673	1.59E-63
ExitRates	-0.207071	1.66E-119
PageValues	0.492569	0.00E+00
SpecialDay	-0.082305	5.50E-20
OperatingSystems	-0.014668	1.03E-01
Browser	0.023984	7.74E-03
Region	-0.011595	1.98E-01
TrafficType	-0.005113	5.70E-01



2. B. 2. Pearson Chi-Squared

Attribute	Chi-squared	p-value
Month	384.934762	2.24E-77
VisitorType	135.251923	4.27E-30
Weekend	10.390978	1.27E-03
Revenue	12322.35585	0.00E+00



Concluzii atribute numerice:

- PageValues: coeficientul de 0.492569 si p-value 0.00E+00 indica o corelatie semnificativa. Acest atribut are o influenta foarte semnificativa asupra variabilei tinta “Revenue”.
- ProductRelated si ProductRelated_Duration ofera de asemenea informatii utile despre predictie.
- BounceRates, ExitRates si SpecialDay au coeficienti negative, indicat o asociere inversa cu varaibila tinta.

- OperatingSystems, Browser, Region si TrafficType au P-value mari (>0.05), astfel indicand lipsa unei corelatii semnificative cu vatiabila tinta.

Concluzii atribute categorice:

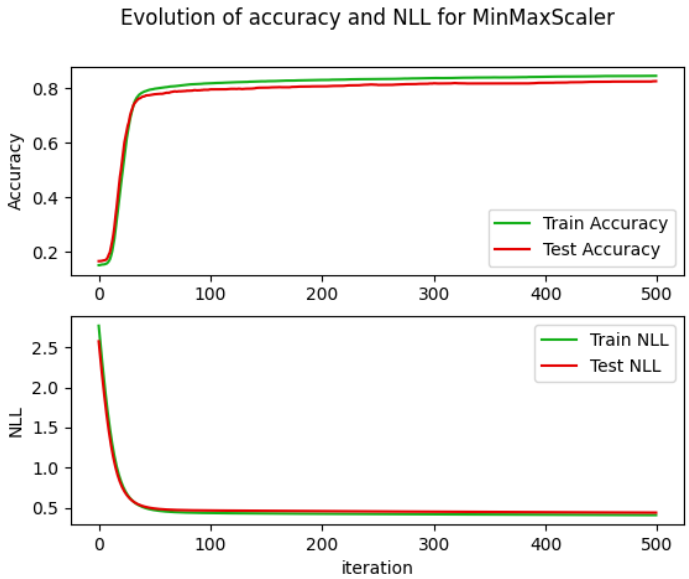
- Month, VisitorType si Weekend ofera o corelatie semnificativa.
- Revenue este variabila tinta.

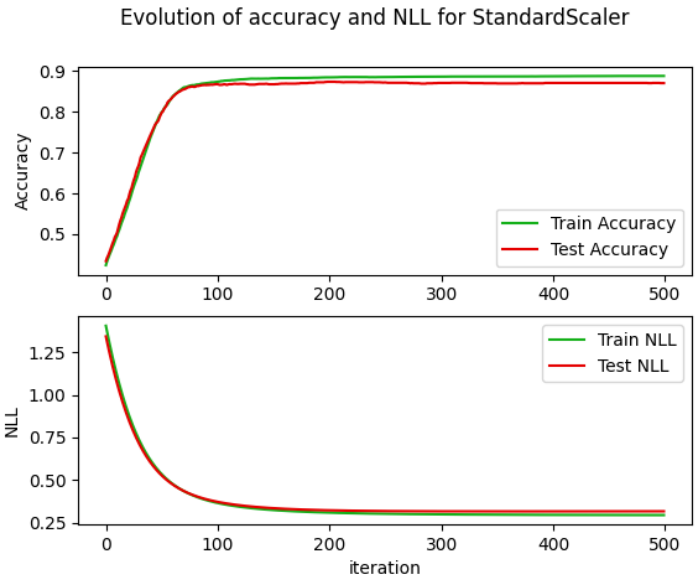
Concluzii generale:

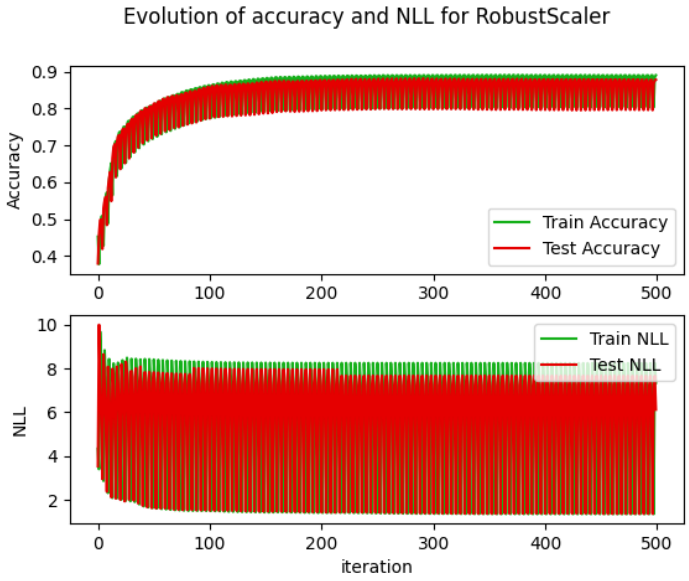
- Atributele PageValues, Month, VisitorType, Weekend, ProductRelated si ProductRelated_Duration sunt cele mai relevante pentru predicita variabilei Revenue.
- Atributele cu coeficienti mari si p-value-uri mici au o influenta importanta asupra variabilei tinta.
- Astfel putem folosi preponderant variabilele relevante pentru a crea un model predictibil mai eficient si performant.

3. 2. 1. Regresie Logistica

- Implementare manuala







3. 2. 3. Evaluarea Comparativa a Rezultatelor

- Regresie Logistica – implementare manuala:

MinMaxScaler:

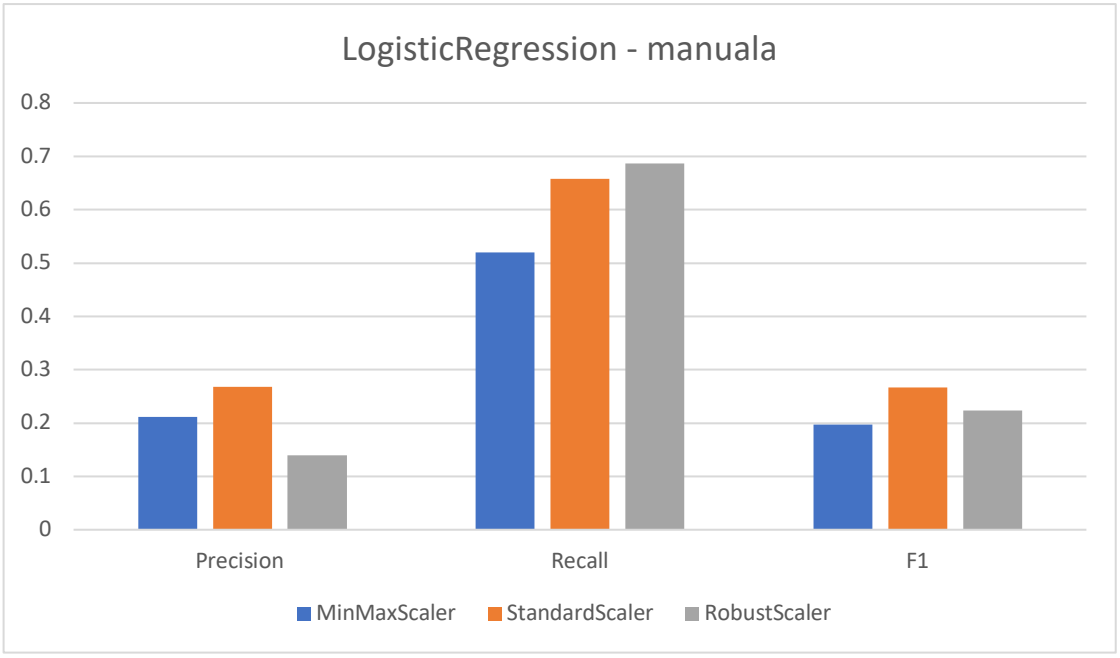
Precision: 0.2110
Recall: 0.5195
F1 Score: 0.1967

StandarScaler:

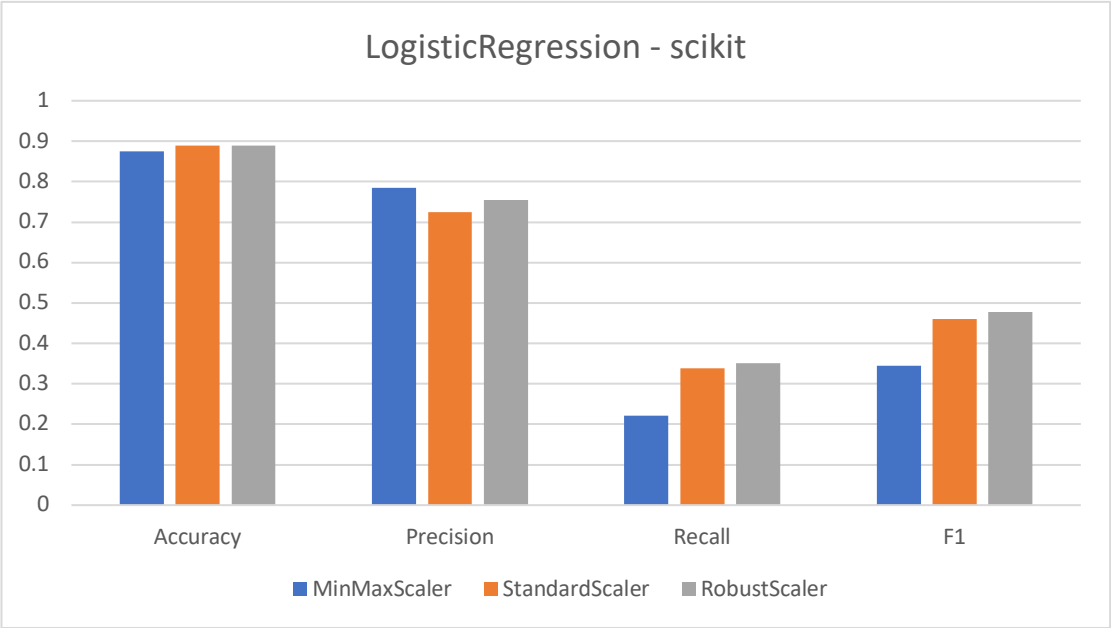
Precision: 0.2674
Recall: 0.6572
F1 Score: 0.2672

RobustScaler:

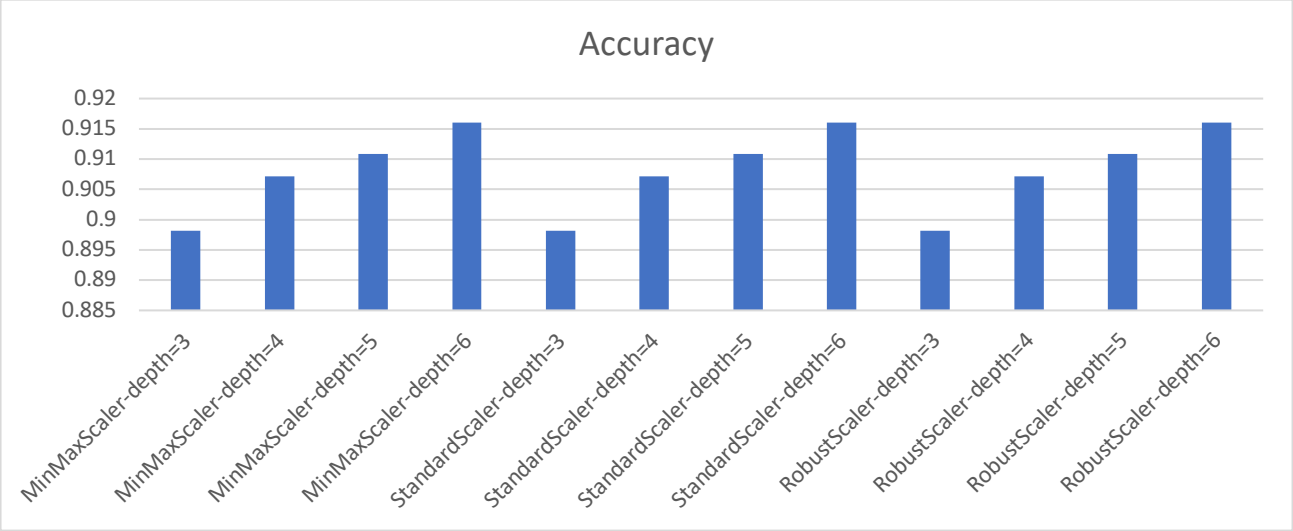
Precision: 0.1397
Recall: 0.6861
F1 Score: 0.2239

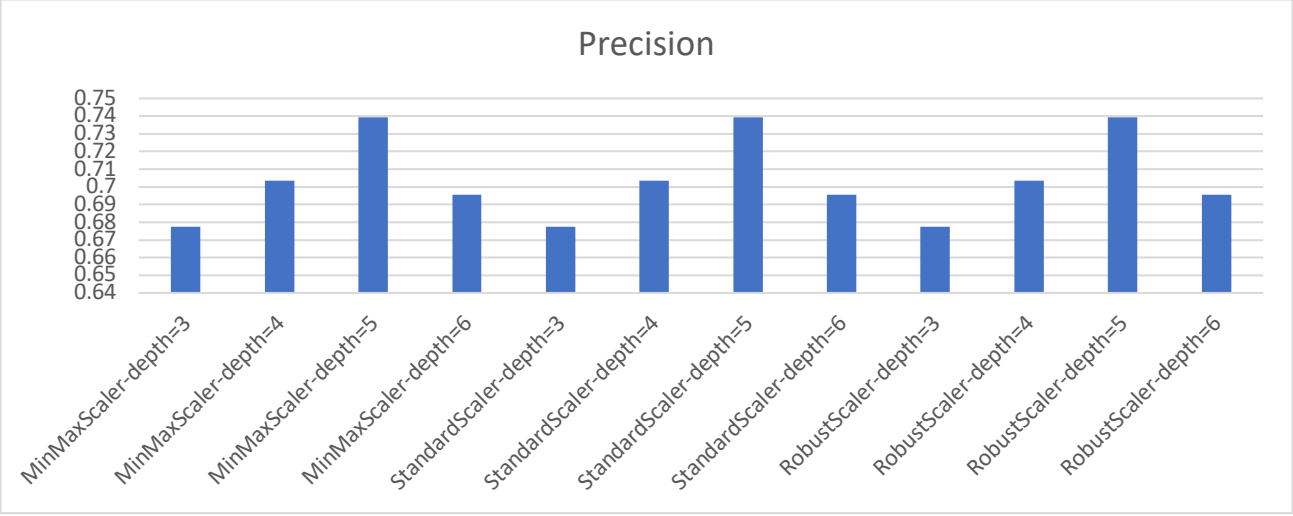


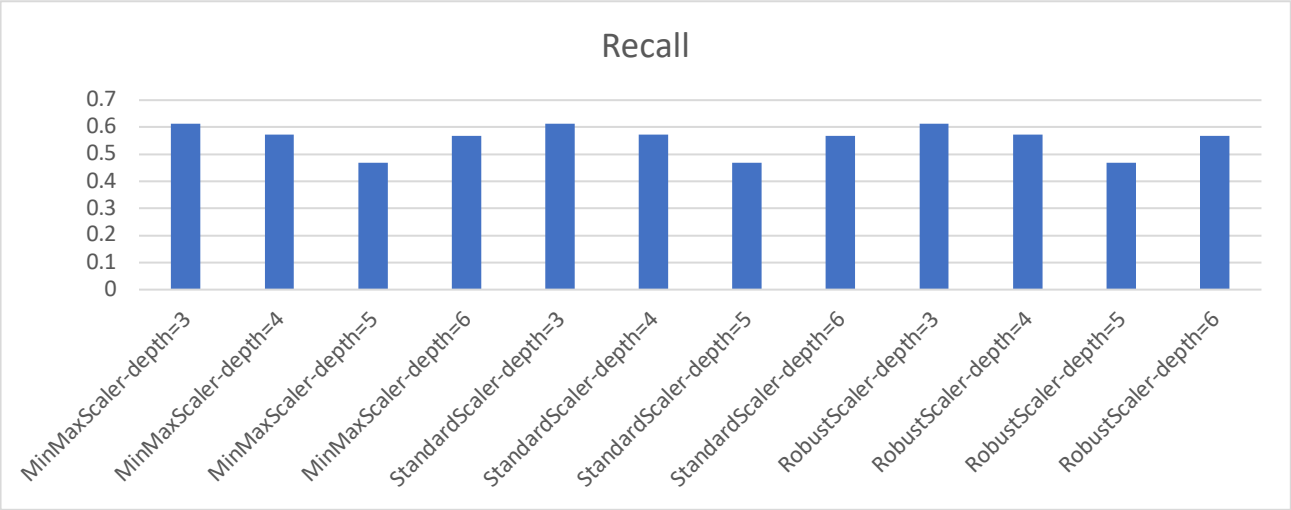
- Regresie Logistica – implementare folosind biblioteca scikit-learn:

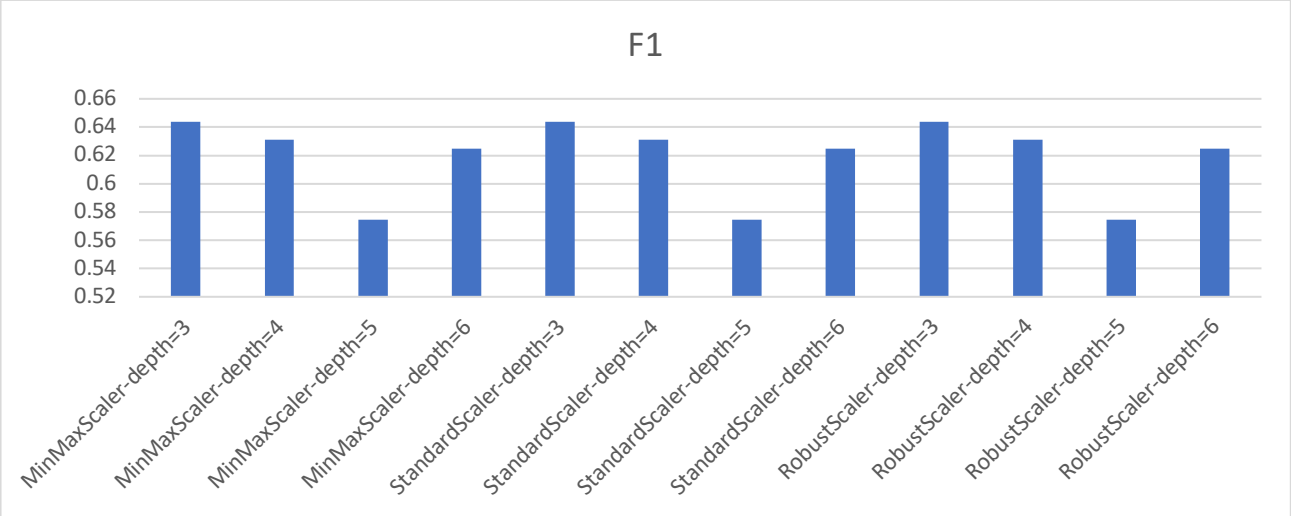


- Arbori de decizie – implementare folosind biblioteca scikit-learn:









Tabel general:

	Precision - mean	Precision - var	Recall - mean	Recall - var	F1 - mean	F1 - var
manual-logistic - MinMaxScaler	0.211	0.0059	0.5195	0.0112	0.1967	0.0059
manual-logistic - StandardScaler	**0.2674**	**0.0014**	**0.6572**	**0.0137**	**0.2672**	**0.0033**
manual-logistic - RobustScaler	0.1397	0.0059	0.6861	0.0167	0.2239	0.0072
scikit-logistic - MinMaxScaler	0.7845	0	0.2214	0	0.3454	0
scikit-logistic - StandardScaler	0.724	0	0.3382	0	0.461	0
scikit-logistic - RobustScaler	**0.7539**	**0**	**0.3504**	**0**	**0.4784**	**0**
scikit-tree - MinMaxScaler - depth=3	0.6774	0	0.6131	0	0.6437	0
scikit-tree - MinMaxScaler - depth=4	0.7036	0	0.5718	0	0.6309	0
scikit-tree - MinMaxScaler - depth=5	0.7395	0	0.4696	0	0.5744	0
scikit-tree - MinMaxScaler - depth=6	0.6955	0	0.5669	0	0.6247	0
scikit-tree - StandardScaler - depth=3	0.6774	0	0.6131	0	0.6437	0

scikit-tree - StandardScaler - depth=4	0.7036	0	0.5718	0	0.6309	0
scikit-tree - StandardScaler - depth=5	0.7395	0	0.4696	0	0.5744	0
scikit-tree - StandardScaler - depth=6	0.6955	0	0.5669	0	0.6247	0
scikit-tree - RobustScaler - depth=3	**0.6774**	**0**	**0.6131**	**0**	**0.6437**	**0**
scikit-tree - RobustScaler - depth=4	0.7036	0	0.5718	0	0.6309	0
scikit-tree - RobustScaler - depth=5	0.7395	0	0.4696	0	0.5744	0
scikit-tree - RobustScaler - depth=6	0.6955	0	0.5669	0	0.6247	0

Concluzii generale:

- Se observa ca biblioteca scikit ofera performante mult mai bune decat implementarea manuala pentru regresia logisitica
- Biblioteca scikit ofera rezultate constante (deviatia standard =1e-20), neschimbandu-se intre rulari consecutive
- Pentru implementarile folosind scikit RobustScaler tinde sa obtina cele mai bune valori si pentru regresia logistica si pentru arborii de decizie
- Arborii de decizie obtin rezultate mai bune comparativ cu regresia logistica, iar depth nu conteaza foarte mult (se poate observa ca indiferent de valoarea adancimii rezultatele obtinute sunt destul de constante)
- Pentru implementara manuala folosind regresia logistica Standard Scaler se potriveste cel mai bine.