Proiect Pachete Software

Tănăsescu Alexandru-Gabriel

Grupa 1096

Organizația analizată: British Airways

Setul de date utilizat: Customer booking virtual experience program

Link către setul de date: https://www.kaggle.com/code/shtrausslearning/british-airways-data-science-programme/input

Setul de date conține următoarele coloane:

- num_passengers: numărul de pasageri pe bilet;
- sales_channel: modalitatea de booking (internet/mobil);
- trip_type: tipul de cursă (fără/cu întoarcere);
- purchase_lead: numărul de zile de la achiziția biletului până la data de îmbarcare;
- length_of_stay: durata sejurului (în zile);
- flight_hour: ora de plecare (în formatul de 24h);
- flight_day: ziua plecării (în abreviere);
- route: ruta avionului (cod ICAO);
- booking_origin: țara de plecare;
- wants_extra_baggage: dacă s-a optat pentru un bagaj adițional (valoare booleană);
- wants_preffered_seat: dacă s-a optat pentru un loc specific (valoare booleană);
- wants_in_flight_meals: dacă s-a optat pentru a lua masa (valoare booleană);
- flight_duration: durata zborului (în ore);
- booking_complete: dacă s-a finalizat booking-ul cu succes (valoare booleană).

Pe acest set de date se vor analiza primele 200 de intrări.

---

1. Analiza datelor utilizând programul SAS

1.1 Importarea datelor dintr-un CSV (setul de date aferent) în SAS

Algoritm:

```
libname home '/home/u63848710/';
proc import datafile="/home/u63848710/customer_booking.csv" out=home.bookings
dbms=csv replace;
  getnames=yes;
run;
```

Rezultat:

Total rows: 200 Total columns: 14 Rows 1-100

	num_passengers	sales_channel	trip_type	purchase_lead	length
1	2	Internet	RoundTrip	262	
2	1	Internet	RoundTrip	112	

1.2 Crearea de formate pentru canalul de booking, prin afișarea Website în loc de Internet

Interpretare: Ajută la exemplificarea mai concisă a platformei de accesare

Algoritm:

```
proc format;
    value $sales_channel_format 'Internet' = 'Website';
run;

data booking_formatat;
    set home.bookings;
    format sales_channel $sales_channel_format.;
run;
```

Rezultat:

	num_passengers	sales_channel
1	2	Website
2	1	Website
3	2	Website
4	1	Website
5	2	Website
6	1	Website
7	3	Website
8	2	Website
9	1	Website
10	1	Mobile
11	2	Website

1.3 Afișarea rândurilor unde pasagerul a vrut un loc specific

Interpretare: Determină preferințele pasagerilor pentru acest serviciu

Algoritm:

```
proc print data=home.bookings;
  where wants_preferred_seat = 1;
run;
```

Rezultat:

Obs	num_passengers	sales_channel	trip_type	purchase_lead	length_of_stay	flight_hour	flight_day	route	booking_origin	wants_extra_baggage	wants_preferred_seat
3	2	Internet	RoundTrip	243	22	17	Wed	AKLDEL	India	1	1
11	2	Internet	RoundTrip	185	25	14	Tue	AKLDEL	United Kingdom	1	1
12	1	Internet	RoundTrip	8	43	2	Sat	AKLDEL	New Zealand	1	1
15	1	Internet	RoundTrip	245	34	4	Tue	AKLDEL	New Zealand	1	1
31	1	Internet	RoundTrip	16	35	23	Wed	AKLICN	New Zealand	1	1
35	1	Internet	RoundTrip	21	22	7	Fri	AKLICN	New Zealand	1	1
38	6	Internet	RoundTrip	20	22	3	Fri	AKLICN	South Korea	1	1
42	1	Internet	RoundTrip	71	90	7	Thu	AKLICN	South Korea	1	1
49	1	Internet	RoundTrip	107	24	14	Fri	AKLICN	South Korea	0	1

1.4 Crearea unui subset de date pentru pasagerii cu plecare din India

Interpretare: Identifică numărul de plecări din India

Algoritm:

```
proc print data=home.bookings;
  where wants_preferred_seat = 1;
run;
```

Rezultat:

	num_passengers	sales_channel	trip_type	purchase_lead	length_of_stay	flight_hour	flight_day	r
1	2	Internet	RoundTrip	243	22	17	Wed	A
2	2	Internet	RoundTrip	68	22	15	Wed	A
3	2	Internet	RoundTrip	238	19	14	Mon	A
4	1	Mobile	RoundTrip	378	30	12	Sun	A
5	1	Internet	RoundTrip	185	17	14	Fri	A
6	1	Internet	RoundTrip	192	18	14	Thu	A
7	1	Internet	RoundTrip	259	37	6	Sun	A

1.5 Calcularea duratei medii a zborurilor:

Interpretare: Oferă informații despre durata medie a zborurilor în vederea eficientizării ulterioare

Algoritm:

```
proc means data=home.bookings;
  title 'Durata medie a zborurilor';
  var flight_duration;
run;
```

Rezultat:

Durata medie a zborurilor

The MEANS Procedure

Analysis Variable : flight_duration				
N	Mean	Std Dev	Minimum	Maximum
200	7.7623500	1.3579670	4.7500000	8.8300000

1.6 Afișarea prețurilor pentru fiecare rând din lista de booking, folosind un set de date de prețuri

Interpretare: Corelarea setului de date cu alte informații elocvente, precum prețul biletelor pentru o eventuală analiză a achizițiilor

Algoritm:

```
data pret_tichete;
  input route $ price;
  datalines;
AKLDEL 133
AKLHGH 150
AKLHND 167
AKLICN 189
AKLKIX 201
AKLKTM 211
AKLKUL 230
;
```

```
run;

proc sql;
  create table booking_cu_preturi as
  select b.*, p.price
  from home.bookings as b
  left join pret_tichete as p
  on b.route = p.route;
quit;
```

Rezultat:

el	trip_type	purchase_lead	length_of_stay	flight_hour	flight_day	route	booking_origin	price
	RoundTrip	67	155	8	Sun	AKLDEL	New Zealand	133
	RoundTrip	192	18	14	Thu	AKLDEL	India	133
	RoundTrip	351	17	3	Sun	AKLHGH	China	150
	CircleTri	228	29	23	Wed	AKLHNL	New Zealand	167
	RoundTrip	250	23	9	Sun	AKLICN	South Korea	189

1.7 Stocarea tuturor rutelor distincte existente în setul de date într-un masiv ordonat

Interpretarea: Identificare tuturor rutelor distincte pentru analize ulterioare (precum cea de sus de corelarea cu prețul biletelor pe ruta respectivă)

Algoritm:

```
proc sort data=home.bookings out=home.sorted nodupkey;
  by route;
run;

data _null_;
  set home.sorted end=last;
  array routes[*] $20 _TEMPORARY_;

  if _n_ = 1 then do;
    do i to dim(routes);
            routes[i] = '';
    end;
  end;

  routes[_n_] = route;

  if last then do;
    call sortn(of routes[*]);
    put 'Routes';
    do i = 1 to dim(routes) while (routes[i] ne '');
            put routes[i];
    end;
  end;
run;
```

Rezultat:

	route
1	AKLDEL
2	AKLHGH
3	AKLHND
4	AKLICN
5	AKLKIX
6	AKLKTM
7	AKLKUL

1.8 Crearea unui raport detaliat al pasagerilor

Interpretare: Identificarea unor tendințe generale în vederea optimizărilor diferitelor strategii

Algoritm:

```
proc report data=home.bookings;
  column num_passengers sales_channel trip_type booking_origin;
run;
```

Rezultat:

num_passengers	sales_channel	trip_type	booking_origin
2	Internet	RoundTrip	New Zealand
1	Internet	RoundTrip	New Zealand
2	Internet	RoundTrip	India
1	Internet	RoundTrip	New Zealand
2	Internet	RoundTrip	India
1	Internet	RoundTrip	New Zealand
3	Internet	RoundTrip	New Zealand
2	Internet	RoundTrip	India
1	Internet	RoundTrip	New Zealand
1	Mobile	RoundTrip	India
2	Internet	RoundTrip	United Kingdom

1.9 Analiza frecvenței zborurilor în funcție de ziua săptămânii

Interpretare: Analiză pentru optimizarea programului de zboruri și diferitelor procese de marketing

Algoritm:

```
proc freq data=home.bookings;
  tables flight_day;
run;
```

Rezultat:

Analiza frecventei dupa ziua saptamanii

The FREQ Procedure

flight_day	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Fri	22	11.00	22	11.00
Mon	36	18.00	58	29.00
Sat	21	10.50	79	39.50
Sun	27	13.50	106	53.00
Thu	27	13.50	133	66.50
Tue	40	20.00	173	86.50
Wed	27	13.50	200	100.00

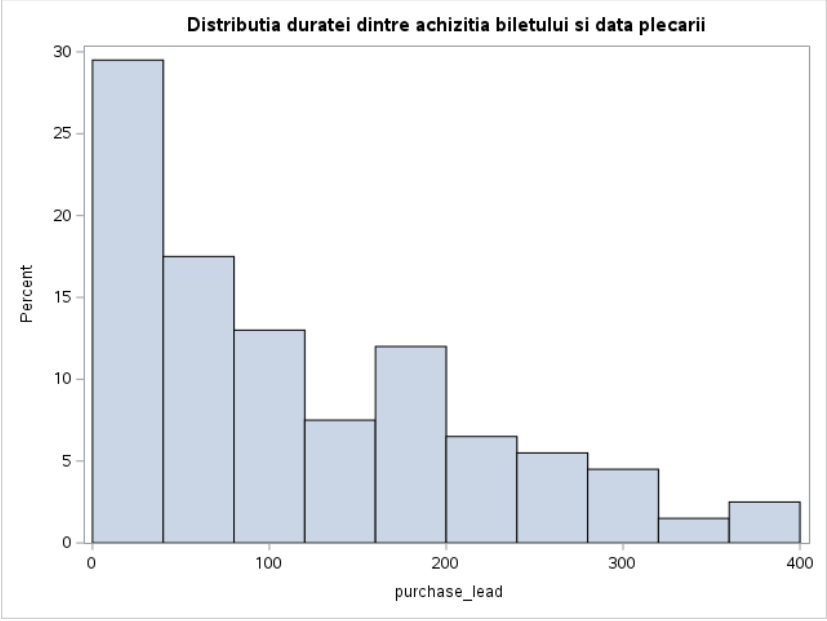
1.10 Vizualizarea distribuției dintre data achiziționării biletului și data plecării

Interpretare: Vizualizarea pentru identificarea tendințelor clienților în materie de achiziție a biletelor (ca durată de timp de dinaintea zborului)

Algoritm:

```
proc sgplot data=home.bookings;
  histogram purchase_lead;
  title 'Distributia duratei dintre achizitia biletului si data plecarii';
run;
```

Rezultat:



2. Analiza datelor utilizând limbajul de programare Python

2.1 Încărcarea setului de date aferent în Python

Algoritm:

```
import pandas as pd
df = pd.read_csv('./customer_booking.csv')
df
```

Rezultat:

	num_passengers	sales_channel	trip_type	purchase_lead	length_of_stay	flight_hour	flight_day	route	booking_origin	wants_extra_baggage	wants_p
0	2	Internet	RoundTrip	262	19	7	Sat	AKLDEL	New Zealand	1	
1	1	Internet	RoundTrip	112	20	3	Sat	AKLDEL	New Zealand	0	
2	2	Internet	RoundTrip	243	22	17	Wed	AKLDEL	India	1	
3	1	Internet	RoundTrip	96	31	4	Sat	AKLDEL	New Zealand	0	
4	2	Internet	RoundTrip	68	22	15	Wed	AKLDEL	India	1	
...	...	...	...	...	...	...	...	...	...	...	
195	1	Internet	RoundTrip	380	21	21	Thu	AKLKUL	New Zealand	1	
196	5	Internet	RoundTrip	206	17	13	Sun	AKLKUL	Malaysia	1	
197	1	Internet	RoundTrip	13	27	1	Wed	AKLKUL	New Zealand	1	
198	1	Internet	RoundTrip	229	31	2	Fri	AKLKUL	New Zealand	0	
199	1	Internet	RoundTrip	132	20	9	Mon	AKLKUL	Slovakia	1	

200 rows × 14 columns

2.2 Afișarea informațiilor pasagerilor care au numărul de zile de la booking până la zbor mai mare de 350 de zile

Interpretare: Vizualizarea datelor unde pasagerii au realizat booking-ul cu un număr mare de zile în avans pentru identificarea criteriilor de alegere al acestora

Algoritm:

```
rows = []
for index, row in df.iterrows():
    if row['purchase_lead'] > 350:
        rows.append(row)
pd.DataFrame(rows)
```

Rezultat:

	num_passengers	sales_channel	trip_type	purchase_lead	length_of_stay	flight_hour	flight_day	route	booking_origin	wants_extra_baggage	wants_
9	1	Mobile	RoundTrip	378	30	12	Sun	AKLDEL	India	0	
19	1	Website	RoundTrip	351	17	3	Sun	AKLHGH	China	0	
135	4	Website	RoundTrip	366	17	16	Sun	AKLKUL	Malaysia	1	
162	1	Website	RoundTrip	384	18	5	Thu	AKLKUL	Malaysia	0	
190	1	Website	RoundTrip	396	23	9	Mon	AKLKUL	New Zealand	1	
195	1	Website	RoundTrip	380	21	21	Thu	AKLKUL	New Zealand	1	

2.3 Determinarea numărului de pasageri care au ales opțiunea de bagaje extra

Interpretare: Identificarea tendinței pasagerilor de a alege această opțiune

Algoritm:

```
pasageri_bagaje_extra = df[df['wants_extra_baggage'] == 1]
print(f'Nr. de pasageri care vor bagaje extra: {pasageri_bagaje_extra.shape[0]} din nr. total de pasageri {df.shape[0]}')
```

Rezultat:

```
pasageri_bagaje_extra = df[df['wants_extra_baggage'] == 1]
print(f'Nr. de pasageri care vor bagaje extra: {pasageri_bagaje_extra.shape[0]} din nr.
```

```
Nr. de pasageri care vor bagaje extra: 147 din nr. total de pasageri 200
```

2.4 Afișarea primelor cinci înregistrări și a primelor nouă coloane

Interpretare: Identificarea datelor și a tipului acestora

Algoritm:

```
df.iloc[:5, :9]
```

Rezultat:

	num_passengers	sales_channel	trip_type	purchase_lead	length_of_stay	flight_hour	flight_day	route	booking_origin
0	2	Website	RoundTrip	262	19	7	Sat	AKLDEL	New Zealand
1	1	Website	RoundTrip	112	20	3	Sat	AKLDEL	New Zealand
2	2	Website	RoundTrip	243	22	17	Wed	AKLDEL	India
3	1	Website	RoundTrip	96	31	4	Sat	AKLDEL	New Zealand
4	2	Website	RoundTrip	68	22	15	Wed	AKLDEL	India

2.5 Calcularea duratei medii a zborurilor

Interpretare: Oferirea de informații utile pentru managementul timpilor de zbor

Algoritm:

```
print(f'Durata medie a zborurilor: {df["flight_duration"].mean()}')
```

Rezultat:

```
print(f'Durata medie a zborurilor: {df["flight_duration"].mean()}')
```

```
Durata medie a zborurilor: 7.762349999999999
```

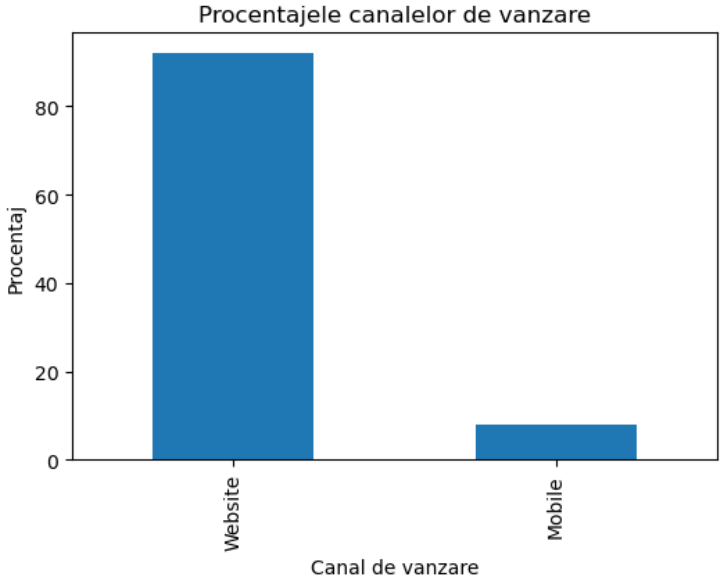
2.6 Afișarea grafică a procentajului canelelor de vânzare utilizate, cu redenumirea prealabilă a valorilor „Internet” în „Website”

Interpretare: Oferirea de informații concrete despre platforma utilizată pentru booking

Algoritm:

```
import matplotlib.pyplot as plt
df['sales_channel'] = df['sales_channel'].replace('Internet', 'Website')
canale_vanzari = df['sales_channel'].value_counts()
procente = (canale_vanzari / canale_vanzari.sum()) * 100
plt.figure(figsize=(6,4))
procente.plot(kind='bar')
plt.title('Procentajele canalelor de vanzare')
plt.xlabel('Canal de vanzare')
plt.ylabel('Procentaj')
plt.show()
```

Rezultat:



2.7 Gruparea datelor după ziua zborului și afișarea mediilor despre numărul de pasageri, durata ședерii și numărul de zile de la booking până la zbor

Interpretare: Identificarea tendințelor generale pe zile ale săptămânii despre media numărului de pasageri, media zilelor de la booking în avans și media zilelor duratei ședерii

Algoritm:

```
grupare = df.groupby('flight_day').agg({
    'num_passengers':'mean',
    'purchase_lead':'mean',
    'length_of_stay':'mean'
}).reset_index()
grupare
```

Rezultat:

	flight_day	num_passengers	purchase_lead	length_of_stay
0	Fri	1.772727	95.136364	28.681818
1	Mon	1.666667	130.416667	36.833333
2	Sat	1.619048	116.142857	37.761905
3	Sun	1.592593	165.481481	40.074074
4	Thu	1.740741	99.666667	52.074074
5	Tue	1.175000	111.375000	48.875000
6	Wed	1.185185	97.111111	36.814815

2.8 Analiza frecvenței zborurilor pe ore

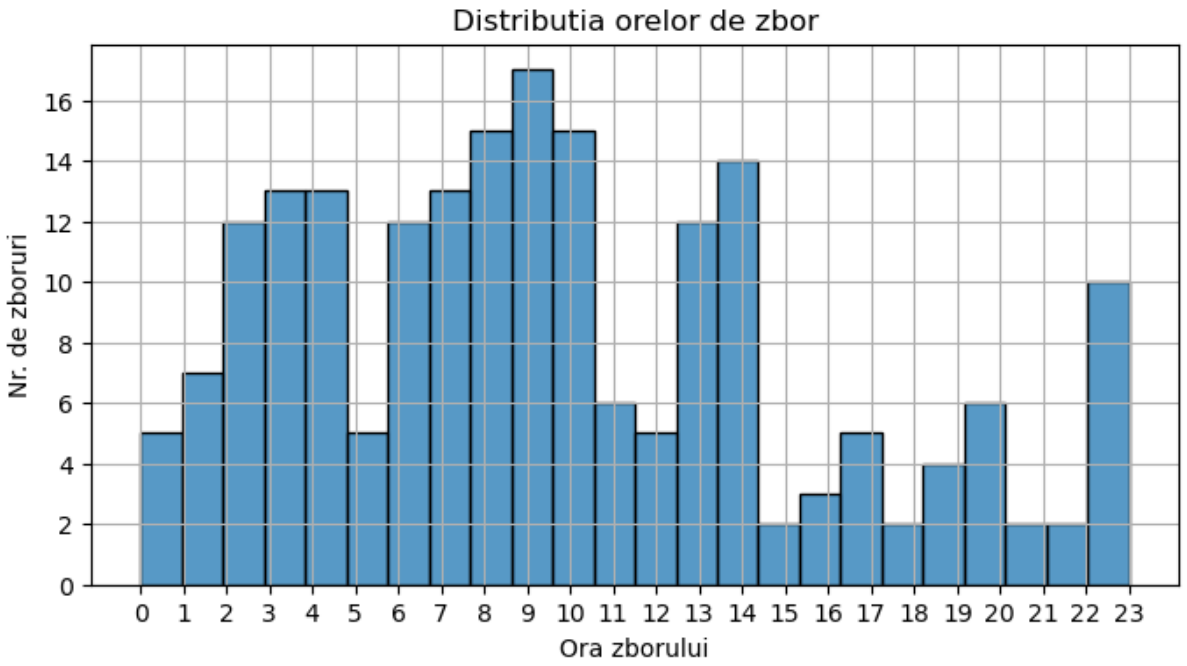
Interpretare: Optimizarea programului de zbor pe ore (fluxului de pasageri)

Algoritm:

```
import seaborn as sns
plt.figure(figsize=(8, 4))
sns.histplot(df['flight_hour'], bins=24, kde=False)
plt.title('Distributia orelor de zbor')
plt.xlabel('Ora zborului')
```

```
plt.ylabel('Nr. de zboruri')
plt.xticks(range(0,24))
plt.grid(True)
plt.show()
```

Rezultat:



2.9 Clusterizarea pasagerilor în funcție de preferințe

Interpretare: Segmentarea pasagerilor după preferințe pentru a le oferii oferte personalizate sau pentru alte criterii

Algoritm:

```
from sklearn.cluster import KMeans
features = df[['wants_extra_baggage', 'wants_preferred_seat', 'wants_in_flight_meals']]
kmeans = KMeans(n_clusters=3)
kmeans.fit(features)
df['cluster'] = kmeans.labels_
df
```

Rezultat:

y	flight_hour	flight_day	route	booking_origin	wants_extra_baggage	wants_preferred_seat	wants_in_flight_meals	flight_duration	booking_complete	cluster
9	7	Sat	AKLDEL	New Zealand	1	0	0	5.52	0	2
0	3	Sat	AKLDEL	New Zealand	0	0	0	5.52	0	0
2	17	Wed	AKLDEL	India	1	1	0	5.52	0	2
1	4	Sat	AKLDEL	New Zealand	0	0	1	5.52	0	0
2	15	Wed	AKLDEL	India	1	0	1	5.52	0	1
..	...	...	...	...	...	...	...	...	...	...
1	21	Thu	AKLKUL	New Zealand	1	1	0	8.83	0	2
7	13	Sun	AKLKUL	Malaysia	1	0	0	8.83	0	2
7	1	Wed	AKLKUL	New Zealand	1	0	1	8.83	0	1
1	2	Fri	AKLKUL	New Zealand	0	0	1	8.83	0	0
0	9	Mon	AKLKUL	Slovakia	1	0	0	8.83	0	2

2.10 Analiza influenței diferitelor variabile asupra orei zborurilor

Interpretare: Identificarea factorilor ce influențează alegerea orei zborului

Algoritm:

```
import statsmodels.api as sm
X = df[['num_passengers', 'purchase_lead', 'length_of_stay', 'flight_duration']].to_numpy()
y = df['flight_hour'].to_numpy()
X = sm.add_constant(X)
model = sm.OLS(y, X).fit()
model.summary()
```

Rezultat:

OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.043
Model:	OLS	**Adj. R-squared:**	0.023
Method:	Least Squares	**F-statistic:**	2.181
Date:	Fri, 24 May 2024	**Prob (F-statistic):**	0.0726
Time:	12:09:39	**Log-Likelihood:**	-641.22
No. Observations:	200	**AIC:**	1292.
Df Residuals:	195	**BIC:**	1309.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	5.5581	2.594	2.142	0.033	0.442	10.674
x1	-0.4677	0.446	-1.048	0.296	-1.348	0.413
x2	0.0061	0.004	1.365	0.174	-0.003	0.015
x3	-0.0179	0.011	-1.643	0.102	-0.039	0.004
x4	0.6151	0.317	1.938	0.054	-0.011	1.241

Omnibus:	9.109	**Durbin-Watson:**	2.260
Prob(Omnibus):	0.011	**Jarque-Bera (JB):**	8.844
Skew:	0.466	**Prob(JB):**	0.0120
Kurtosis:	2.563	**Cond. No.**	957.

Concluzii: Această analiză a datelor permite identificarea preferințelor și tendințelor pasagerilor, prin care se vor optimiza ofertele și serviciile companiei, adică îmbunătățirea eficienței operaționale, ca de exemplu identificarea factorilor decisivi în alegerea rutei, oferirea de bonusuri, strategii de marketing mai bine dezvoltate după preferințe și opțiuni adiț.