Importing Necessory libraris and packages

In [1]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import matplotlib_inline
import plotly.express as px
```

In [2]:
```
pd.set_option('display.max_columns',50)
pd.set_option('display.max_rows',500)
```

Reading & Understanding Data

In [3]:
```
df = pd.read_csv("hotel_booking.csv")
df.head()
```

Out[3]:

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_day_of_month	s
0	Resort Hotel	0	342	2015	July	27	1	
1	Resort Hotel	0	737	2015	July	27	1	
2	Resort Hotel	0	7	2015	July	27	1	
3	Resort Hotel	0	13	2015	July	27	1	
4	Resort Hotel	0	14	2015	July	27	1	

In [4]:
```
# Shape of the dataset
df.shape
```

Out[4]: (119390, 36)

In [5]:

```python
# Columns/features of the dataset
df.columns
```

Out[5]:

```
Index(['hotel', 'is_canceled', 'lead_time', 'arrival_date_year',
       'arrival_date_month', 'arrival_date_week_number',
       'arrival_date_day_of_month', 'stays_in_weekend_nights',
       'stays_in_week_nights', 'adults', 'children', 'babies', 'meal',
       'country', 'market_segment', 'distribution_channel',
       'is_repeated_guest', 'previous_cancellations',
       'previous_bookings_not_canceled', 'reserved_room_type',
       'assigned_room_type', 'booking_changes', 'deposit_type', 'agent',
       'company', 'days_in_waiting_list', 'customer_type', 'adr',
       'required_car_parking_spaces', 'total_of_special_requests',
       'reservation_status', 'reservation_status_date', 'name', 'email',
       'phone-number', 'credit_card'],
      dtype='object')
```

In [6]:

```
# Information about the dataset
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 119390 entries, 0 to 119389
Data columns (total 36 columns):
 #   Column                          Non-Null Count   Dtype
---  ------                          --------------   -----
 0   hotel                           119390 non-null  object
 1   is_canceled                     119390 non-null  int64
 2   lead_time                       119390 non-null  int64
 3   arrival_date_year               119390 non-null  int64
 4   arrival_date_month              119390 non-null  object
 5   arrival_date_week_number        119390 non-null  int64
 6   arrival_date_day_of_month       119390 non-null  int64
 7   stays_in_weekend_nights         119390 non-null  int64
 8   stays_in_week_nights            119390 non-null  int64
 9   adults                          119390 non-null  int64
 10  children                        119386 non-null  float64
 11  babies                          119390 non-null  int64
 12  meal                            119390 non-null  object
 13  country                         118902 non-null  object
 14  market_segment                  119390 non-null  object
 15  distribution_channel            119390 non-null  object
 16  is_repeated_guest               119390 non-null  int64
 17  previous_cancellations          119390 non-null  int64
 18  previous_bookings_not_canceled  119390 non-null  int64
 19  reserved_room_type              119390 non-null  object
 20  assigned_room_type              119390 non-null  object
 21  booking_changes                 119390 non-null  int64
 22  deposit_type                    119390 non-null  object
 23  agent                           103050 non-null  float64
 24  company                         6797 non-null    float64
 25  days_in_waiting_list            119390 non-null  int64
 26  customer_type                   119390 non-null  object
 27  adr                             119390 non-null  float64
 28  required_car_parking_spaces     119390 non-null  int64
 29  total_of_special_requests       119390 non-null  int64
 30  reservation_status              119390 non-null  object
 31  reservation_status_date         119390 non-null  object
 32  name                            119390 non-null  object
```

```
 33  email                           119390 non-null  object
 34  phone-number                    119390 non-null  object
 35  credit_card                     119390 non-null  object
dtypes: float64(4), int64(16), object(16)
memory usage: 32.8+ MB
```

The datatype for `reservation_status_date` is object, so we need to change it to datetime.

In [7]:

```python
# Change the datatype
df['reservation_status_date'] = pd.to_datetime(df['reservation_status_date'])
df['reservation_status_date']
```

Out[7]:

```
0        2015-07-01
1        2015-07-01
2        2015-07-02
3        2015-07-02
4        2015-07-03
            ...
119385   2017-09-06
119386   2017-09-07
119387   2017-09-07
119388   2017-09-07
119389   2017-09-07
Name: reservation_status_date, Length: 119390, dtype: datetime64[ns]
```

In [8]:

```
# Check for null values
df.isnull().sum()
```

Out[8]:

```
hotel                                0
is_canceled                          0
lead_time                            0
arrival_date_year                    0
arrival_date_month                   0
arrival_date_week_number             0
arrival_date_day_of_month            0
stays_in_weekend_nights              0
stays_in_week_nights                 0
adults                               0
children                             4
babies                               0
meal                                 0
country                            488
market_segment                       0
distribution_channel                 0
is_repeated_guest                    0
previous_cancellations               0
previous_bookings_not_canceled       0
reserved_room_type                   0
assigned_room_type                   0
booking_changes                      0
deposit_type                         0
agent                            16340
company                         112593
days_in_waiting_list                 0
customer_type                        0
adr                                  0
required_car_parking_spaces          0
total_of_special_requests            0
reservation_status                   0
reservation_status_date              0
name                                 0
email                                0
phone-number                         0
credit_card                          0
dtype: int64
```

Exploratory Data Analysis

Data Cleaning

In [9]:

```
# Lets check percentage missing values
df.isnull().sum()*100/len(df)
```

Out[9]:

hotel	0.000000
is_canceled	0.000000
lead_time	0.000000
arrival_date_year	0.000000
arrival_date_month	0.000000
arrival_date_week_number	0.000000
arrival_date_day_of_month	0.000000
stays_in_weekend_nights	0.000000
stays_in_week_nights	0.000000
adults	0.000000
children	0.003350
babies	0.000000
meal	0.000000
country	0.408744
market_segment	0.000000
distribution_channel	0.000000
is_repeated_guest	0.000000
previous_cancellations	0.000000
previous_bookings_not_canceled	0.000000
reserved_room_type	0.000000
assigned_room_type	0.000000
booking_changes	0.000000
deposit_type	0.000000
agent	13.686238
company	94.306893
days_in_waiting_list	0.000000
customer_type	0.000000
adr	0.000000
required_car_parking_spaces	0.000000
total_of_special_requests	0.000000
reservation_status	0.000000
reservation_status_date	0.000000
name	0.000000
email	0.000000
phone-number	0.000000
credit_card	0.000000

dtype: float64

In [10]:

```
# Dropping `company`, `agent`
df.drop(['company','agent'], axis=1, inplace=True)
```

In [11]:

```
# checking null again if any
df.isnull().sum()
```

Out[11]:

```
hotel                             0
is_canceled                       0
lead_time                         0
arrival_date_year                 0
arrival_date_month                0
arrival_date_week_number          0
arrival_date_day_of_month         0
stays_in_weekend_nights           0
stays_in_week_nights              0
adults                            0
children                          4
babies                            0
meal                              0
country                         488
market_segment                    0
distribution_channel              0
is_repeated_guest                 0
previous_cancellations            0
previous_bookings_not_canceled    0
reserved_room_type                0
assigned_room_type                0
booking_changes                   0
deposit_type                      0
days_in_waiting_list              0
customer_type                     0
adr                               0
required_car_parking_spaces       0
total_of_special_requests         0
reservation_status                0
reservation_status_date           0
name                              0
email                             0
phone-number                      0
credit_card                       0
dtype: int64
```

In [12]:

```python
# Dropping all the null values
# Check shape before the dataset
print('Before Dropping null: ', df.shape)
df.dropna(inplace=True)
print('After Dropping null : ', df.shape)
```

```
Before Dropping null:  (119390, 34)
After Dropping null :  (118898, 34)
```

In [13]:

```
# Statistical summary for the numerical analysis
df.describe().T
```

Out[13]:

	count	mean	std	min	25%	50%	75%	max
is_canceled	118898.0	0.371352	0.483168	0.00	0.0	0.0	1.0	1.0
lead_time	118898.0	104.311435	106.903309	0.00	18.0	69.0	161.0	737.0
arrival_date_year	118898.0	2016.157656	0.707459	2015.00	2016.0	2016.0	2017.0	2017.0
arrival_date_week_number	118898.0	27.166555	13.589971	1.00	16.0	28.0	38.0	53.0
arrival_date_day_of_month	118898.0	15.800880	8.780324	1.00	8.0	16.0	23.0	31.0
stays_in_weekend_nights	118898.0	0.928897	0.996216	0.00	0.0	1.0	2.0	16.0
stays_in_week_nights	118898.0	2.502145	1.900168	0.00	1.0	2.0	3.0	41.0
adults	118898.0	1.858391	0.578576	0.00	2.0	2.0	2.0	55.0
children	118898.0	0.104207	0.399172	0.00	0.0	0.0	0.0	10.0
babies	118898.0	0.007948	0.097380	0.00	0.0	0.0	0.0	10.0
is_repeated_guest	118898.0	0.032011	0.176029	0.00	0.0	0.0	0.0	1.0
previous_cancellations	118898.0	0.087142	0.845869	0.00	0.0	0.0	0.0	26.0
previous_bookings_not_canceled	118898.0	0.131634	1.484672	0.00	0.0	0.0	0.0	72.0
booking_changes	118898.0	0.221181	0.652785	0.00	0.0	0.0	0.0	21.0
days_in_waiting_list	118898.0	2.330754	17.630452	0.00	0.0	0.0	0.0	391.0
adr	118898.0	102.003243	50.485862	-6.38	70.0	95.0	126.0	5400.0
required_car_parking_spaces	118898.0	0.061885	0.244172	0.00	0.0	0.0	0.0	8.0
total_of_special_requests	118898.0	0.571683	0.792678	0.00	0.0	0.0	1.0	5.0

Univariate Analysis

In [14]:

```
# Lets separate the categorical and numerical columns
cat_list = df.select_dtypes(include="object").columns.to_list()
num_list = df.select_dtypes(include=['number', 'datetime']).columns.to_list()
```

In [15]:

```
# Categorical columns
cat_list
```

Out[15]:

```
['hotel',
 'arrival_date_month',
 'meal',
 'country',
 'market_segment',
 'distribution_channel',
 'reserved_room_type',
 'assigned_room_type',
 'deposit_type',
 'customer_type',
 'reservation_status',
 'name',
 'email',
 'phone-number',
 'credit_card']
```

In [16]:

```python
# Numerical columns
num_list
```

Out[16]:

```
['is_canceled',
 'lead_time',
 'arrival_date_year',
 'arrival_date_week_number',
 'arrival_date_day_of_month',
 'stays_in_weekend_nights',
 'stays_in_week_nights',
 'adults',
 'children',
 'babies',
 'is_repeated_guest',
 'previous_cancellations',
 'previous_bookings_not_canceled',
 'booking_changes',
 'days_in_waiting_list',
 'adr',
 'required_car_parking_spaces',
 'total_of_special_requests',
 'reservation_status_date']
```

In [17]:

```python
# Verify if all the features are selected
len(cat_list) + len(num_list) == df.shape[1]
```

Out[17]:

```
True
```

In [18]:

```python
# Univariate Categorical Analysis - Unique categories
for col in cat_list:
    print(col)
    print(df[col].unique())
    print("="*80)
```

```
hotel
['Resort Hotel' 'City Hotel']
================================================================================
arrival_date_month
['July' 'August' 'September' 'October' 'November' 'December' 'January'
 'February' 'March' 'April' 'May' 'June']
================================================================================
meal
['BB' 'FB' 'HB' 'SC' 'Undefined']
================================================================================
country
['PRT' 'GBR' 'USA' 'ESP' 'IRL' 'FRA' 'ROU' 'NOR' 'OMN' 'ARG' 'POL' 'DEU'
 'BEL' 'CHE' 'CN' 'GRC' 'ITA' 'NLD' 'DNK' 'RUS' 'SWE' 'AUS' 'EST' 'CZE'
 'BRA' 'FIN' 'MOZ' 'BWA' 'LUX' 'SVN' 'ALB' 'IND' 'CHN' 'MEX' 'MAR' 'UKR'
 'SMR' 'LVA' 'PRI' 'SRB' 'CHL' 'AUT' 'BLR' 'LTU' 'TUR' 'ZAF' 'AGO' 'ISR'
 'CYM' 'ZMB' 'CPV' 'ZWE' 'DZA' 'KOR' 'CRI' 'HUN' 'ARE' 'TUN' 'JAM' 'HRV'
 'HKG' 'IRN' 'GEO' 'AND' 'GIB' 'URY' 'JEY' 'CAF' 'CYP' 'COL' 'GGY' 'KWT'
 'NGA' 'MDV' 'VEN' 'SVK' 'FJI' 'KAZ' 'PAK' 'IDN' 'LBN' 'PHL' 'SEN' 'SYC'
 'AZE' 'BHR' 'NZL' 'THA' 'DOM' 'MKD' 'MYS' 'ARM' 'JPN' 'LKA' 'CUB' 'CMR'
 'BIH' 'MUS' 'COM' 'SUR' 'UGA' 'BGR' 'CIV' 'JOR' 'SYR' 'SGP' 'BDI' 'SAU'
 'VNM' 'PLW' 'QAT' 'EGY' 'PER' 'MLT' 'MWI' 'ECU' 'MDG' 'ISL' 'UZB' 'NPL'
 'BHS' 'MAC' 'TGO' 'TWN' 'DJI' 'STP' 'KNA' 'ETH' 'IRQ' 'HND' 'RWA' 'KHM'
 'MCO' 'BGD' 'IMN' 'TJK' 'NIC' 'BEN' 'VGB' 'TZA' 'GAB' 'GHA' 'TMP' 'GLP'
 'KEN' 'LIE' 'GNB' 'MNE' 'UMI' 'MYT' 'FRO' 'MMR' 'PAN' 'BFA' 'LBY' 'MLI'
 'NAM' 'BOL' 'PRY' 'BRB' 'ABW' 'AIA' 'SLV' 'DMA' 'PYF' 'GUY' 'LCA' 'ATA'
 'GTM' 'ASM' 'MRT' 'NCL' 'KIR' 'SDN' 'ATF' 'SLE' 'LAO']
================================================================================
market_segment
['Direct' 'Corporate' 'Online TA' 'Offline TA/TO' 'Complementary' 'Groups'
 'Aviation']
================================================================================
distribution_channel
['Direct' 'Corporate' 'TA/TO' 'Undefined' 'GDS']
================================================================================
reserved_room_type
['C' 'A' 'D' 'E' 'G' 'F' 'H' 'L' 'B' 'P']
================================================================================
assigned_room_type
```

```
['C' 'A' 'D' 'E' 'G' 'F' 'I' 'B' 'H' 'L' 'K' 'P']
================================================================================
deposit_type
['No Deposit' 'Refundable' 'Non Refund']
================================================================================
customer_type
['Transient' 'Contract' 'Transient-Party' 'Group']
================================================================================
reservation_status
['Check-Out' 'Canceled' 'No-Show']
================================================================================
name
['Ernest Barnes' 'Andrea Baker' 'Rebecca Parker' ... 'Wesley Aguilar'
 'Caroline Conley MD' 'Ariana Michael']
================================================================================
email
['Ernest.Barnes31@outlook.com' 'Andrea_Baker94@aol.com'
 'Rebecca_Parker@comcast.net' ... 'Mary_Morales@hotmail.com'
 'MD_Caroline@comcast.net' 'Ariana_M@xfinity.com']
================================================================================
phone-number
['669-792-1661' '858-637-6955' '652-885-2745' ... '395-518-4100'
 '531-528-1017' '422-804-6403']
================================================================================
credit_card
['************4322' '************9157' '************3734' ...
 '************9170' '************6349' '************7959']
================================================================================
```

In [19]:

```python
# Looking at the above result, we can drop `name`, `email`, `phone-number`, `credit_card`
drop_list = ["name","email",'phone-number','credit_card']
df.drop(drop_list, axis=1, inplace=True)
```

In [20]:
```
# We are left with
df.shape
```

Out[20]:
```
(118898, 30)
```

In [21]:
```
# Removing the drop_list from cat_list
cat_list = [item for item in cat_list if item not in drop_list ]
cat_list
```

Out[21]:
```
['hotel',
 'arrival_date_month',
 'meal',
 'country',
 'market_segment',
 'distribution_channel',
 'reserved_room_type',
 'assigned_room_type',
 'deposit_type',
 'customer_type',
 'reservation_status']
```

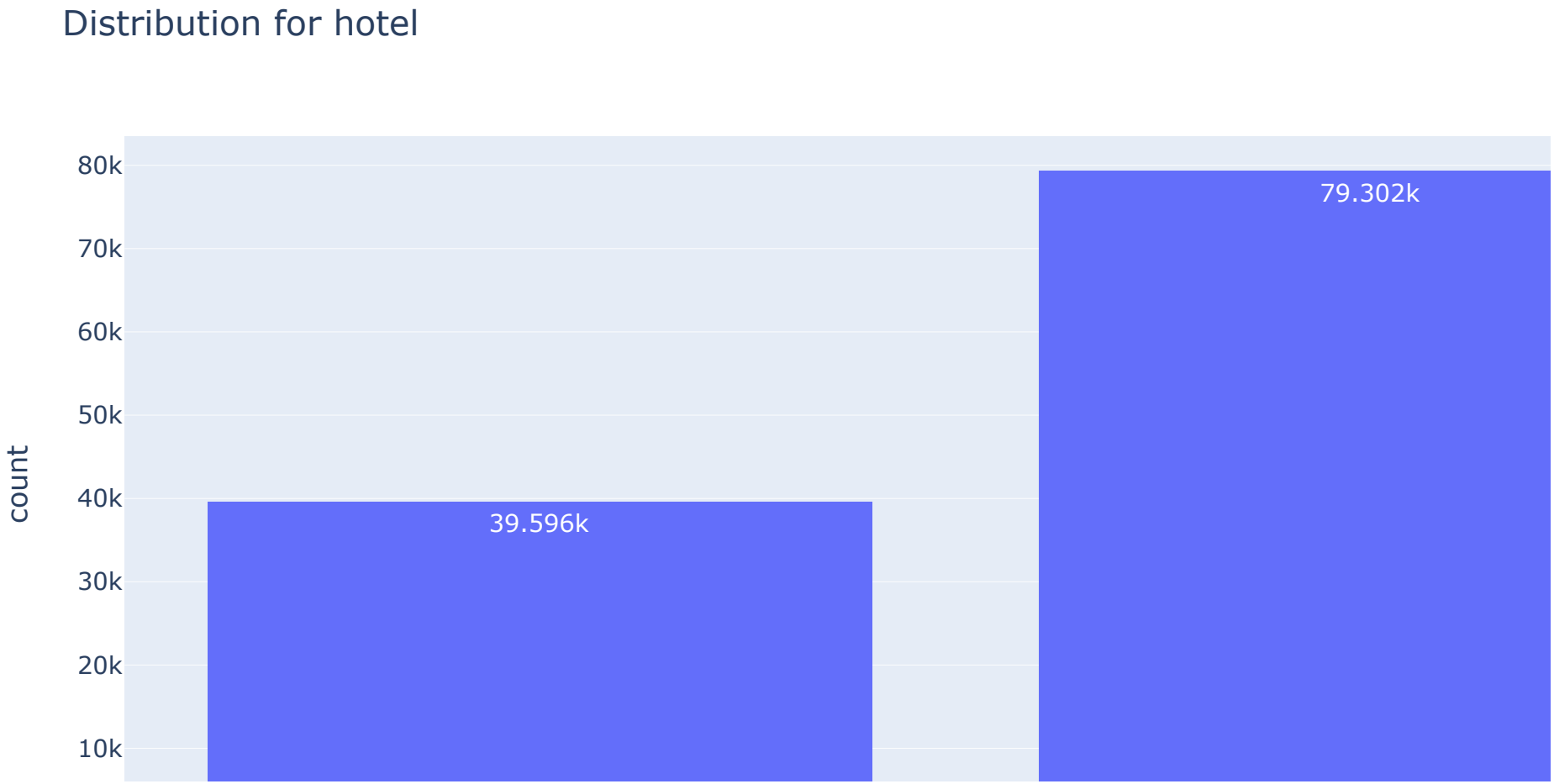
In [29]:
```
list(df['hotel'].value_counts().values)
```

Out[29]:
```
[79302, 39596]
```

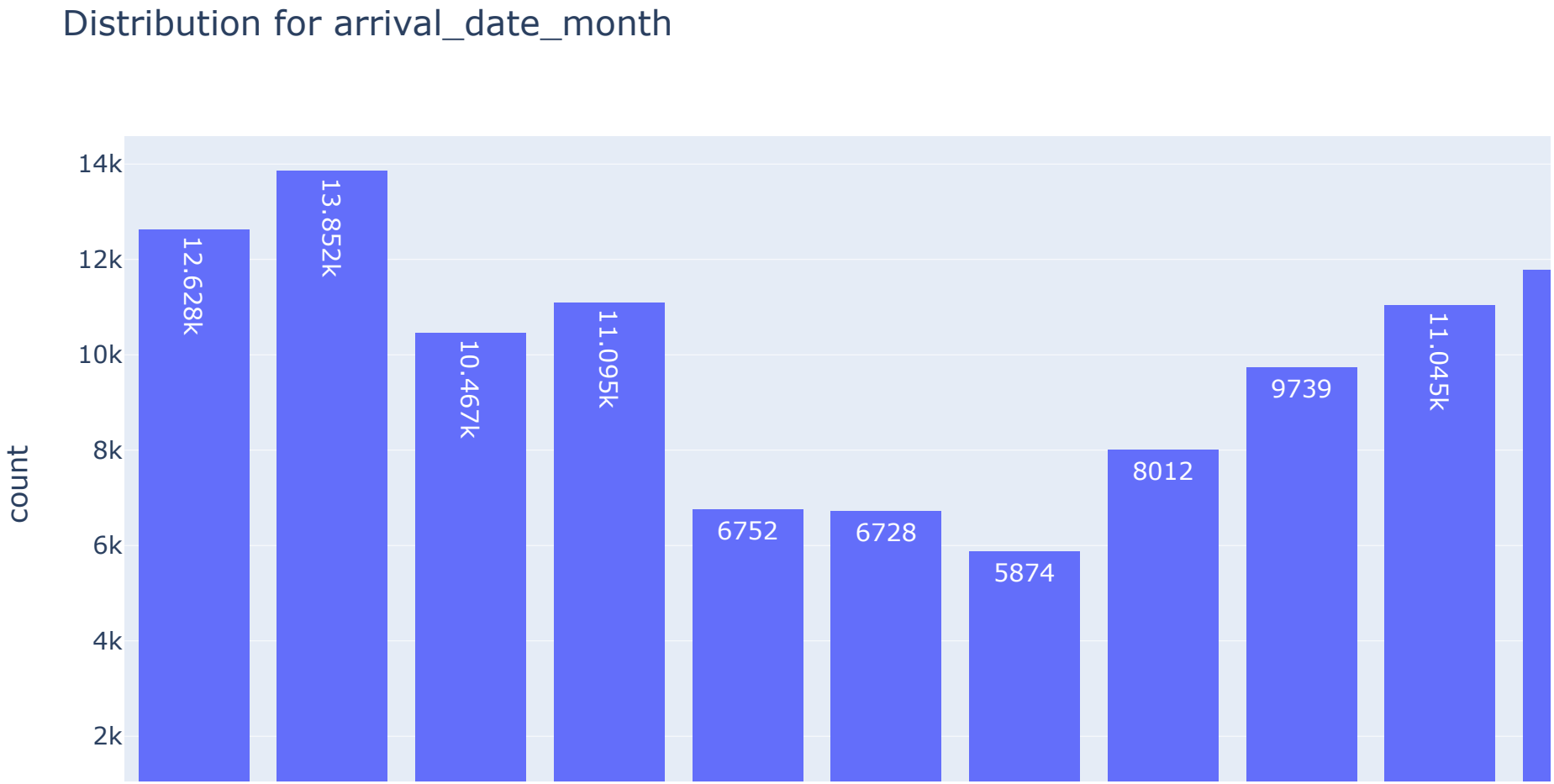
In [38]:

```
# Countplot for categorical variable
for col in cat_list:
    print(df[col].value_counts(normalize=True))
    fig = px.histogram(data_frame=df,
                       x = col,
                       text_auto=True,
                       title="Distribution for {}".format(col))
    fig.show()
    print('='*80)
```

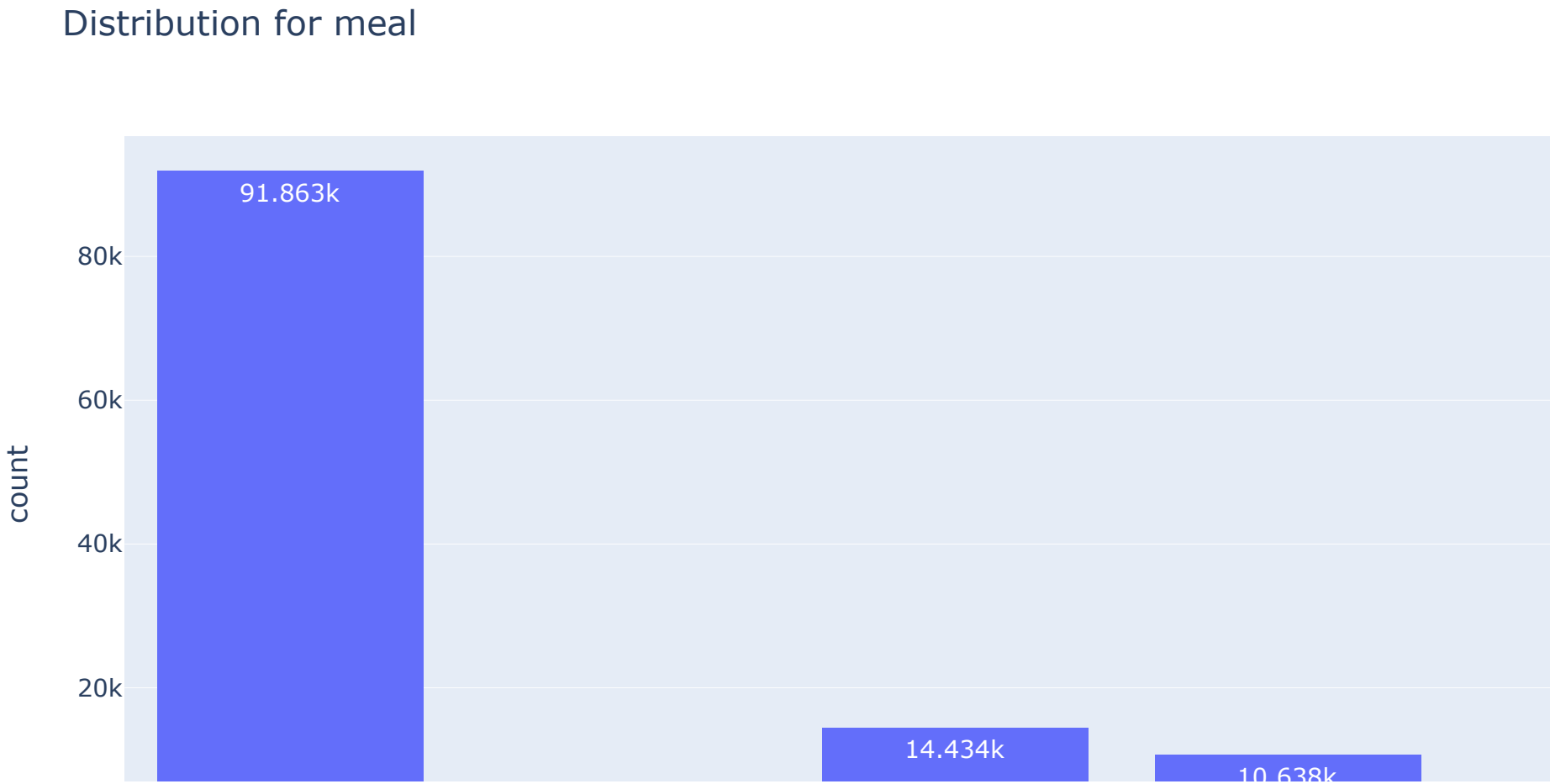
```
City Hotel      0.666975
Resort Hotel    0.333025
Name: hotel, dtype: float64
```


Distribution for hotel
count
80k
70k
60k
50k
40k
30k
20k
10k
39.596k
79.302k

```
================================================================================
August       0.116503
July         0.106209
May          0.099068
October      0.093315
April        0.092895
June         0.091902
September    0.088033
March        0.081911
February     0.067385
November     0.056788
December     0.056586
January      0.049404
Name: arrival_date_month, dtype: float64
```


Distribution for arrival_date_month
count
14k
12k
10k
8k
6k
4k
2k
12.628k
13.852k
10.467k
11.095k
6752
6728
5874
8012
9739
11.045k

```
================================================================================
BB           0.772620
HB           0.121398
SC           0.089472
Undefined    0.009798
FB           0.006712
Name: meal, dtype: float64
```


Distribution for meal
count
80k
60k
40k
20k
91.863k
14.434k
10.638k

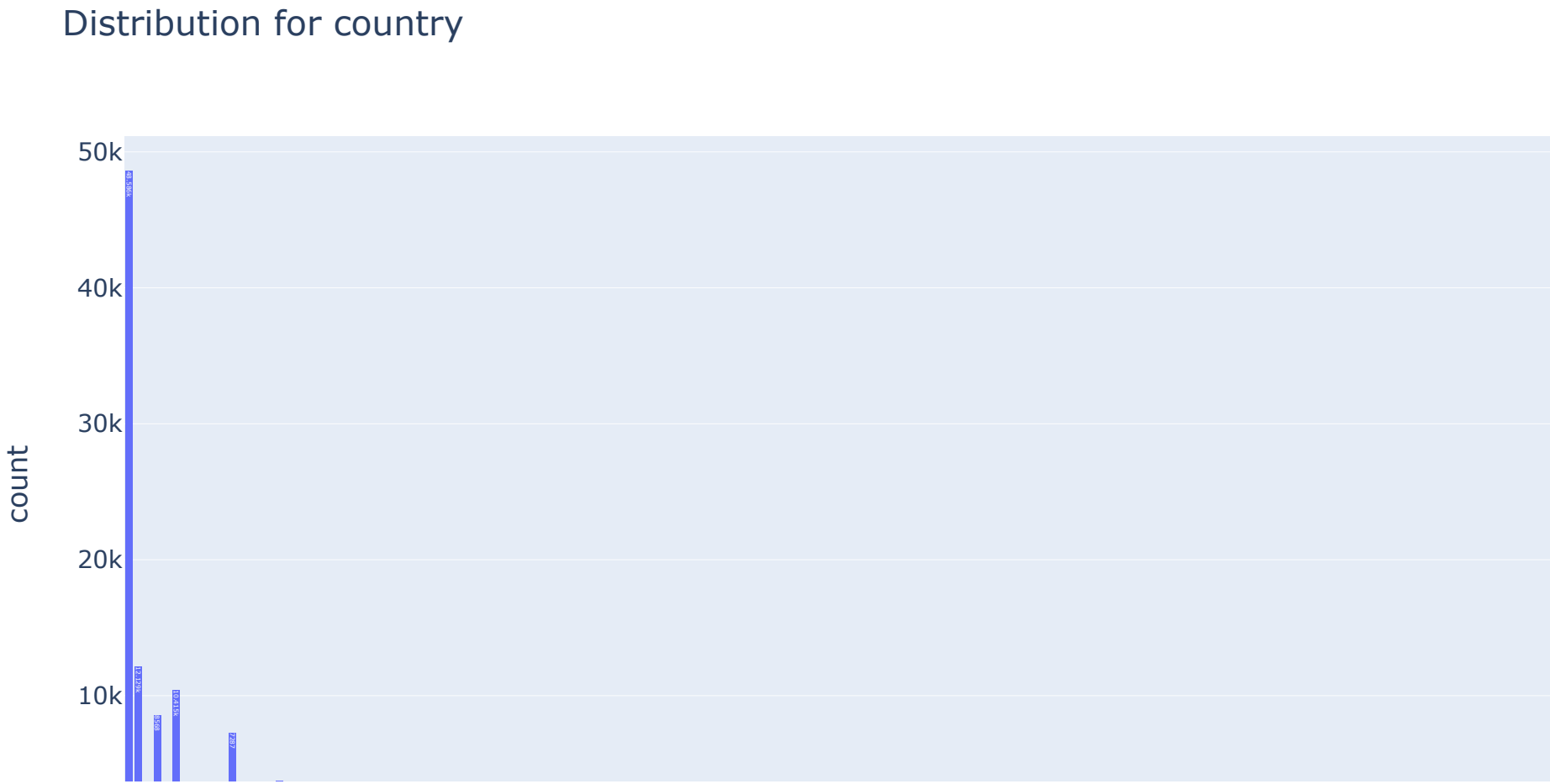
```
==============================================================================
PRT    0.408636
GBR    0.102012
FRA    0.087596
ESP    0.072062
DEU    0.061288
ITA    0.031674
IRL    0.028386
BEL    0.019698
BRA    0.018705
NLD    0.017696
USA    0.017637
CHE    0.014550
CN     0.010757
AUT    0.010623
SWE    0.008612
CHN    0.008402
POL    0.007729
ISR    0.005627
RUS    0.005315
NOR    0.005105
ROU    0.004205
FIN    0.003760
DNK    0.003659
AUS    0.003583
AGO    0.003045
LUX    0.002414
MAR    0.002178
TUR    0.002086
HUN    0.001934
ARG    0.001800
JPN    0.001657
CZE    0.001438
IND    0.001278
KOR    0.001119
GRC    0.001077
DZA    0.000866
SRB    0.000849
```

HRV	0.000841
MEX	0.000715
EST	0.000698
IRN	0.000698
LTU	0.000681
ZAF	0.000673
BGR	0.000631
NZL	0.000622
COL	0.000597
UKR	0.000572
MOZ	0.000564
CHL	0.000547
SVK	0.000547
THA	0.000496
SVN	0.000479
ISL	0.000479
LVA	0.000463
ARE	0.000429
CYP	0.000429
TWN	0.000429
SAU	0.000404
PHL	0.000336
TUN	0.000328
SGP	0.000328
IDN	0.000294
NGA	0.000286
EGY	0.000269
URY	0.000269
LBN	0.000261
PER	0.000244
HKG	0.000244
MYS	0.000235
ECU	0.000227
VEN	0.000219
BLR	0.000219
CPV	0.000202
GEO	0.000185
JOR	0.000177

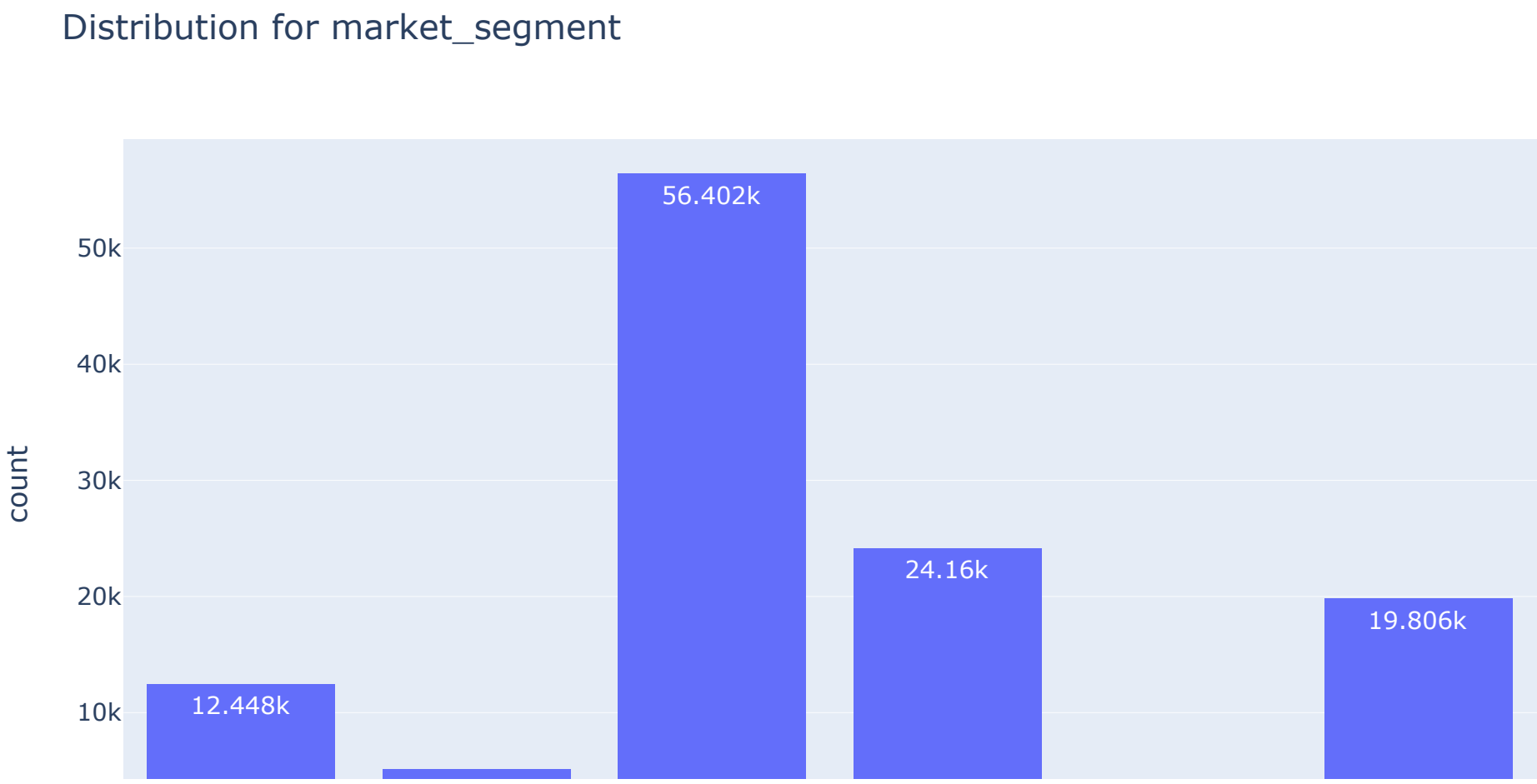
KAZ	0.000160
CRI	0.000160
GIB	0.000151
MLT	0.000151
OMN	0.000151
AZE	0.000143
KWT	0.000135
MAC	0.000135
QAT	0.000126
IRQ	0.000118
DOM	0.000118
PAK	0.000118
BIH	0.000109
MDV	0.000101
BGD	0.000101
ALB	0.000101
PRI	0.000101
SEN	0.000093
CMR	0.000084
MKD	0.000084
BOL	0.000084
PAN	0.000076
GNB	0.000076
TJK	0.000076
VNM	0.000067
CUB	0.000067
ARM	0.000067
JEY	0.000067
LBY	0.000067
AND	0.000059
MUS	0.000059
LKA	0.000059
CIV	0.000050
JAM	0.000050
KEN	0.000050
FRO	0.000042
MNE	0.000042
TZA	0.000042

BHR	0.000042
CAF	0.000042
SUR	0.000042
PRY	0.000034
BRB	0.000034
GTM	0.000034
UZB	0.000034
MCO	0.000034
GAB	0.000034
GHA	0.000034
ZWE	0.000034
ETH	0.000025
TMP	0.000025
LIE	0.000025
GGY	0.000025
SYR	0.000025
BEN	0.000025
GLP	0.000017
SLV	0.000017
ATA	0.000017
MYT	0.000017
ABW	0.000017
KHM	0.000017
LAO	0.000017
STP	0.000017
ZMB	0.000017
MWI	0.000017
IMN	0.000017
COM	0.000017
TGO	0.000017
UGA	0.000017
KNA	0.000017
RWA	0.000017
SYC	0.000017
KIR	0.000008
SDN	0.000008
NCL	0.000008
AIA	0.000008

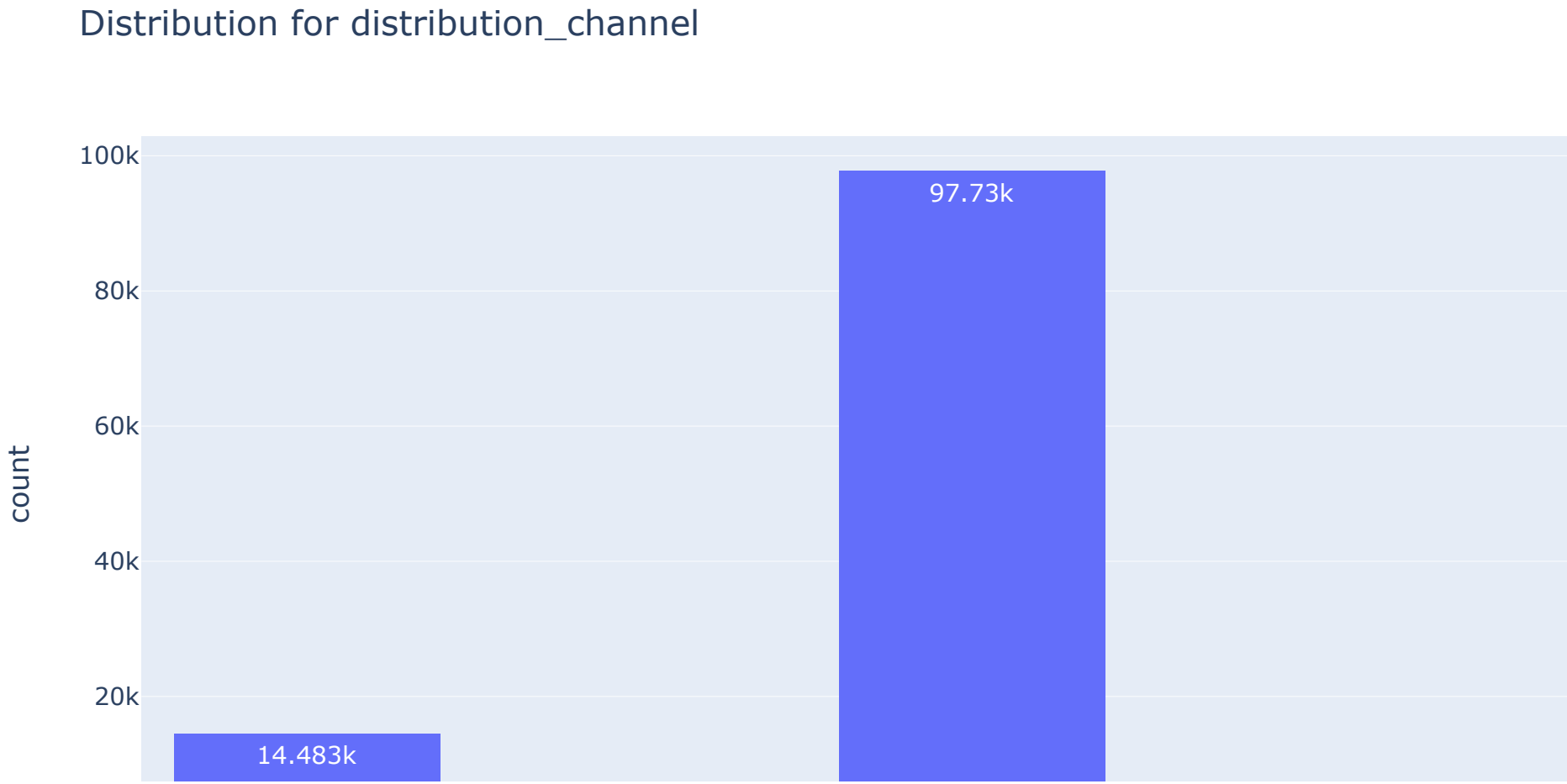
```
ASM    0.000008
FJI    0.000008
ATF    0.000008
LCA    0.000008
GUY    0.000008
PYF    0.000008
DMA    0.000008
SLE    0.000008
MRT    0.000008
NIC    0.000008
BDI    0.000008
PLW    0.000008
MLI    0.000008
CYM    0.000008
BFA    0.000008
MDG    0.000008
MMR    0.000008
NPL    0.000008
BHS    0.000008
UMI    0.000008
SMR    0.000008
DJI    0.000008
BWA    0.000008
HND    0.000008
VGB    0.000008
NAM    0.000008
Name: country, dtype: float64
```


Distribution for country
count
50k
40k
30k
20k
10k

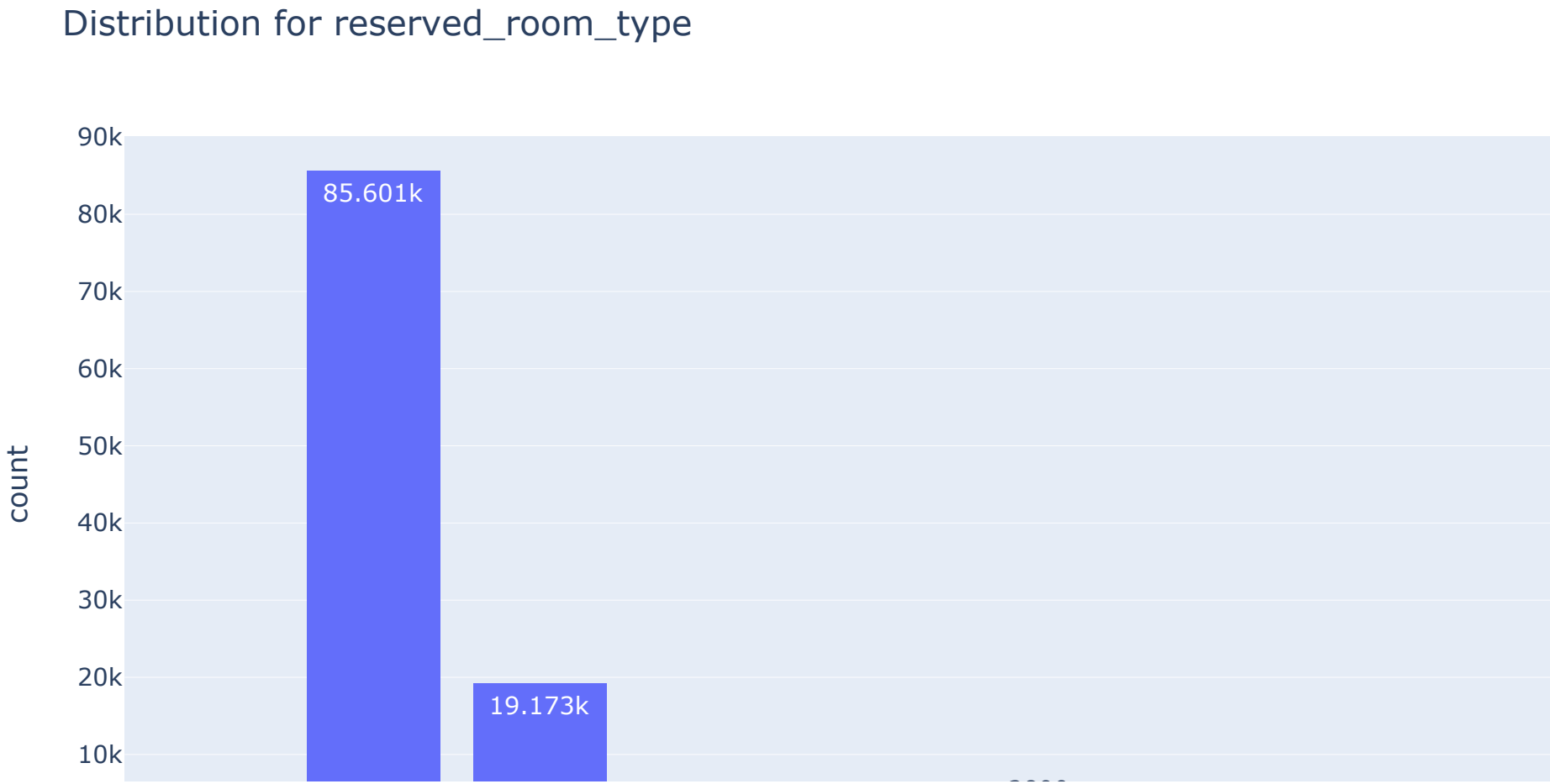
```
================================================================================
Online TA        0.474373
Offline TA/TO    0.203199
Groups           0.166580
Direct           0.104695
Corporate        0.042986
Complementary    0.006173
Aviation         0.001993
Name: market_segment, dtype: float64
```


Distribution for market_segment
count
50k
40k
30k
20k
10k
56.402k
24.16k
19.806k
12.448k

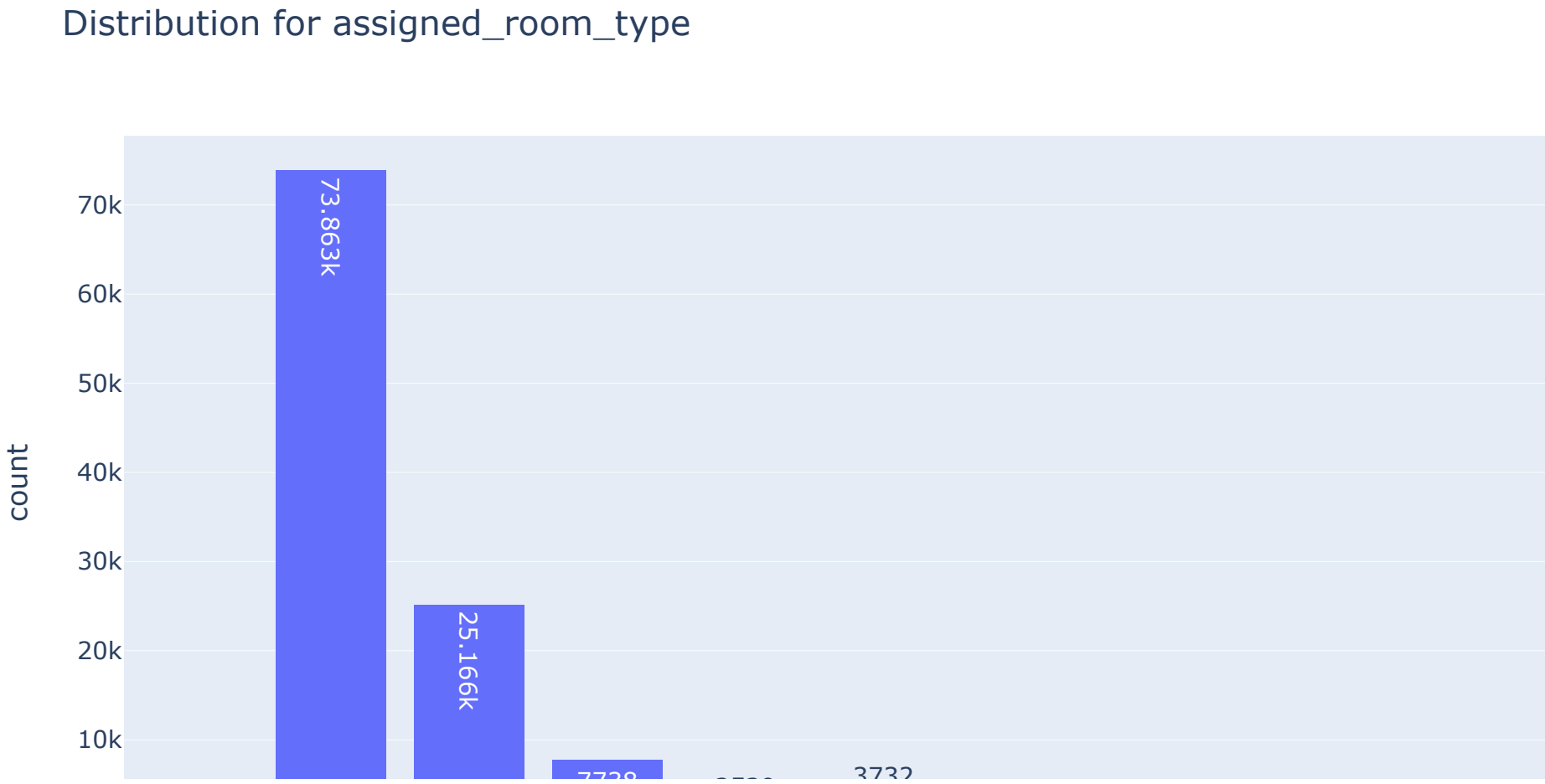
```
================================================================================
TA/TO        0.821965
Direct       0.121810
Corporate    0.054593
GDS          0.001623
Undefined    0.000008
Name: distribution_channel, dtype: float64
```


Distribution for distribution_channel
count
100k
80k
60k
40k
20k
14.483k
97.73k

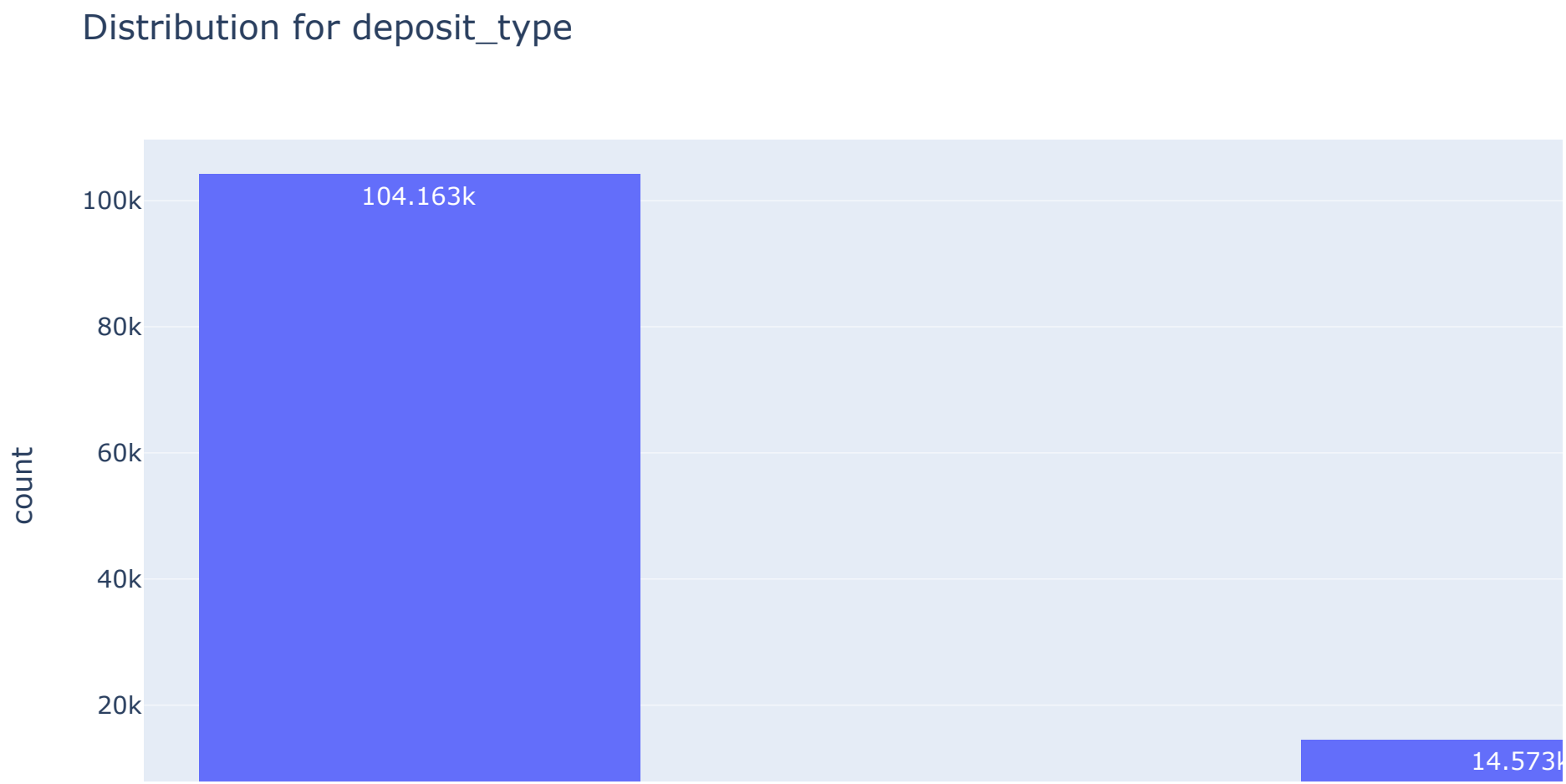
```
================================================================================
A    0.719953
D    0.161256
E    0.054643
F    0.024307
G    0.017519
B    0.009369
C    0.007830
H    0.005055
L    0.000050
P    0.000017
Name: reserved_room_type, dtype: float64
```


Distribution for reserved_room_type
count
90k
80k
70k
60k
50k
40k
30k
20k
10k
85.601k
19.173k

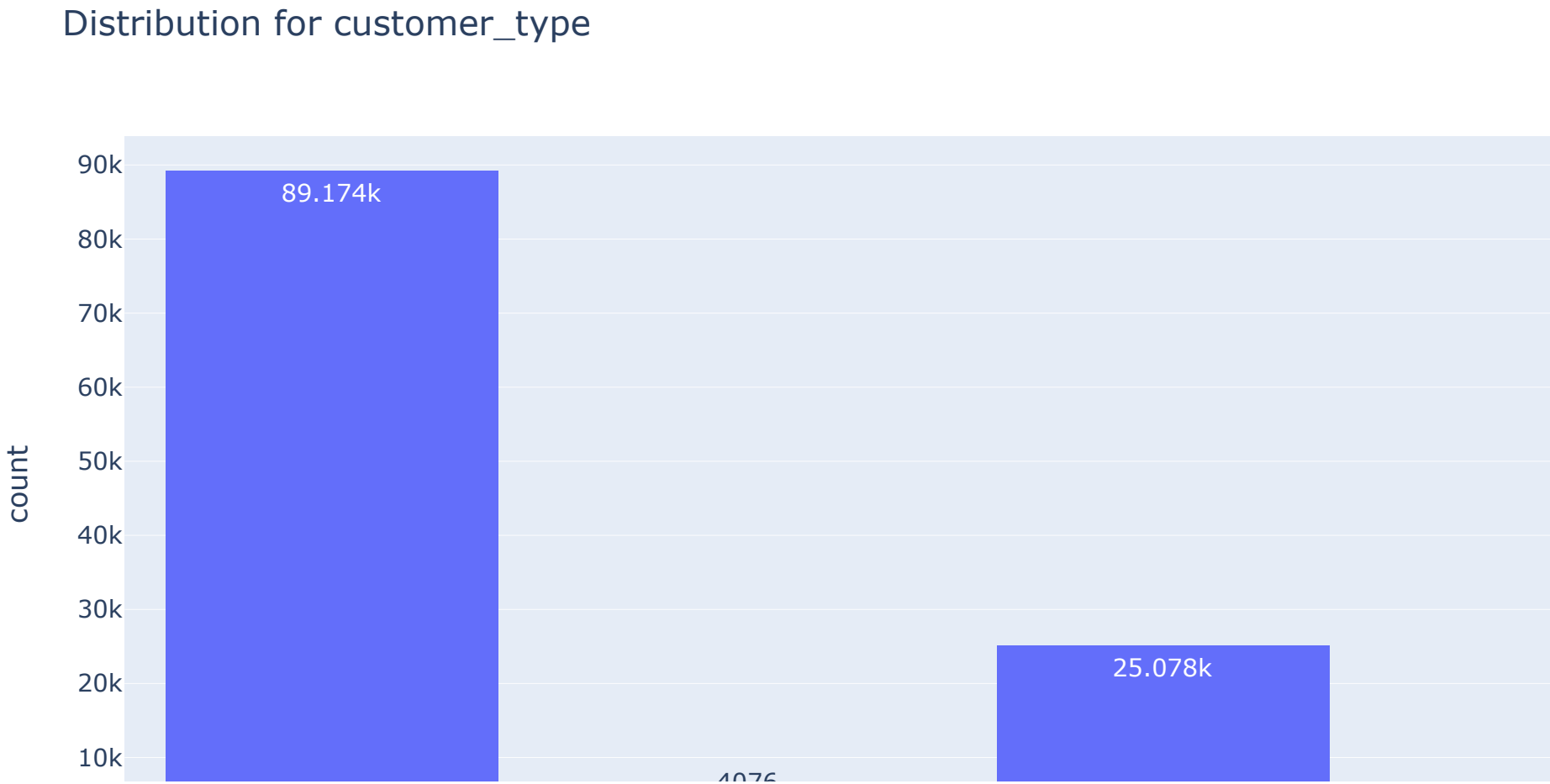
```
================================================================================
A    0.621230
D    0.211660
E    0.065081
F    0.031388
G    0.021354
C    0.019798
B    0.018158
H    0.005955
I    0.003003
K    0.002347
P    0.000017
L    0.000008
Name: assigned_room_type, dtype: float64
```



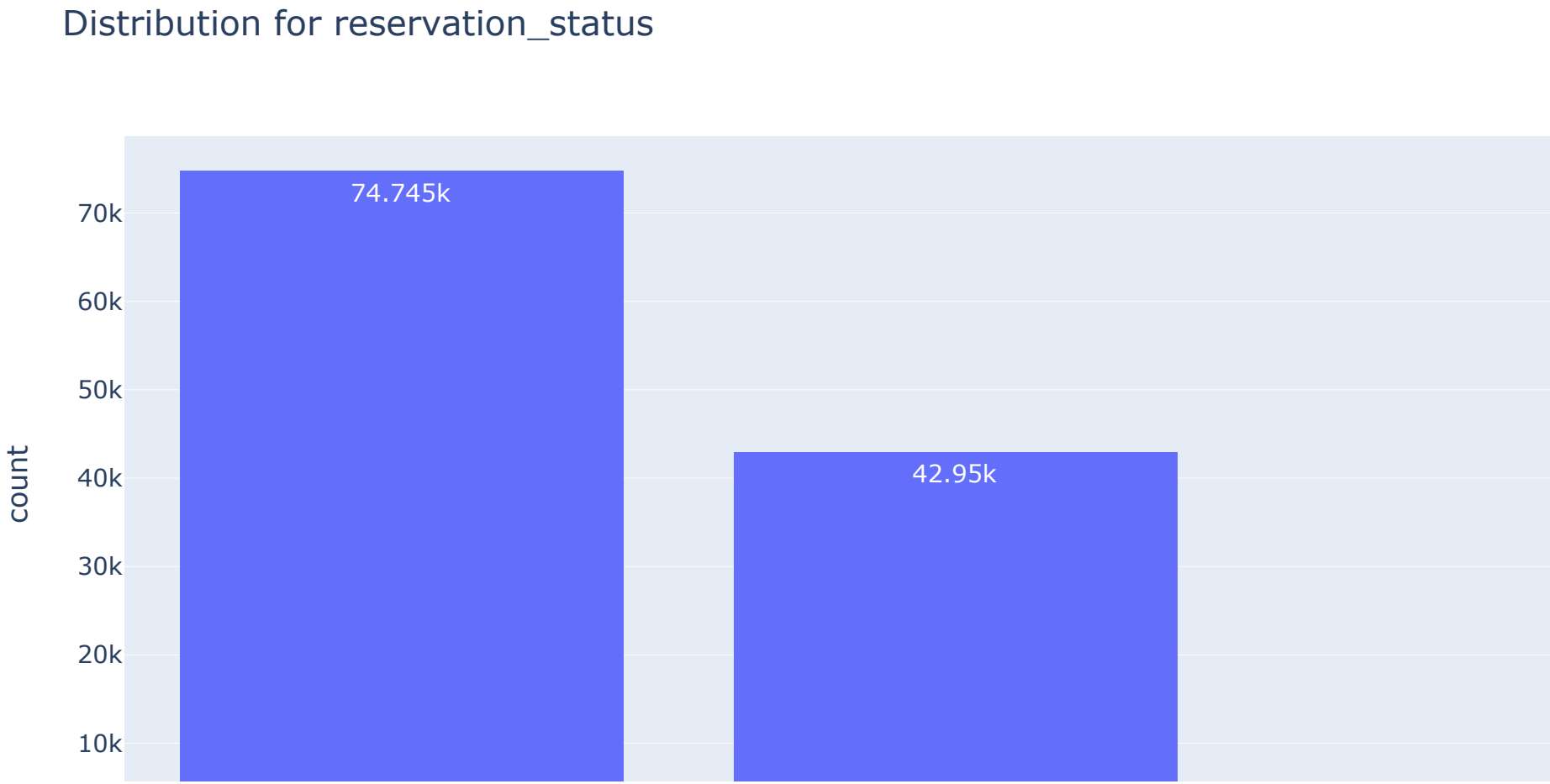

```
==============================================================================
No Deposit    0.876070
Non Refund    0.122567
Refundable    0.001363
Name: deposit_type, dtype: float64
```




```
==============================================================================
Transient          0.750004
Transient-Party    0.210920
Contract           0.034281
Group              0.004794
Name: customer_type, dtype: float64
```




```
==============================================================================
Check-Out    0.628648
Canceled     0.361234
No-Show      0.010118
Name: reservation_status, dtype: float64
```


Distribution for reservation_status
count
70k
60k
50k
40k
30k
20k
10k
74.745k
42.95k

==========

In [ ]:

```
1
```