Цель работы

Предобработка данных для дальнейшего применения методов машинного обучения для решения задач.

Датасет – Спрос на бронирование отелей: https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand

Основная информация по датасету:

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_day_of_month	stays_in_weekend_nights
0	Resort Hotel	0	342	2015	July	27	1	0
1	Resort Hotel	0	737	2015	July	27	1	0
2	Resort Hotel	0	7	2015	July	27	1	0
3	Resort Hotel	0	13	2015	July	27	1	0
4	Resort Hotel	0	14	2015	July	27	1	0
...	...	...	...	...	...	...	...	...
119385	City Hotel	0	23	2017	August	35	30	2
119386	City Hotel	0	102	2017	August	35	31	2
119387	City Hotel	0	34	2017	August	35	31	2
119388	City Hotel	0	109	2017	August	35	31	2
119389	City Hotel	0	205	2017	August	35	29	2

119390 rows × 32 columns

```
Data columns (total 32 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
dtypes: float64(4), int64(16), object(12)
memory usage: 29.1+ MB
```

Очистка данных

Были удалены пропуски и дубликатов, произведены минимакс нормализация и нормализация средним (Z-нормализация).

Ввод [136]:
```
df = df.drop_duplicates()
df.dropna(subset=['lead_time', 'arrival_date_year'], inplace=Tru
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 27453 entries, 0 to 119367
Data columns (total 34 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   hotel                           27453 non-null  object
 1   is_canceled                     27453 non-null  int64
 2   lead_time                       27453 non-null  int64
 3   arrival_date_year               27453 non-null  int64
 4   arrival_date_month              27453 non-null  object
 5   arrival_date_week_number        27453 non-null  int64
 6   arrival_date_day_of_month       27453 non-null  int64
 7   stays_in_weekend_nights         27453 non-null  int64
 8   stays_in_week_nights            27453 non-null  int64
 9   adults                          27453 non-null  int64
 10  children                        27449 non-null  float64
 11  babies                          27453 non-null  int64
 12  meal                            27453 non-null  object
 13  country                         27453 non-null  object
 14  market_segment                  27453 non-null  object
 15  distribution_channel            27453 non-null  object
 16  is_repeated_guest               27453 non-null  int64
 17  previous_cancellations          27453 non-null  int64
 18  previous_bookings_not_canceled  27453 non-null  int64
 19  reserved_room_type              27453 non-null  object
 20  assigned_room_type              27453 non-null  object
 21  booking_changes                 27453 non-null  int64
 22  deposit_type                    27453 non-null  object
 23  agent                           19598 non-null  float64
 24  company                         3567 non-null   float64
 25  days_in_waiting_list            27453 non-null  int64
 26  customer_type                   27453 non-null  object
 27  adr                             27453 non-null  float64
 28  required_car_parking_spaces     27453 non-null  int64
 29  total_of_special_requests       27453 non-null  int64
 30  reservation_status              27453 non-null  object
 31  reservation_status_date         27453 non-null  object
 32  stays_in_weeks_nights           27453 non-null  int64
 33  deposit_type_num                27453 non-null  int32
dtypes: float64(4), int32(1), int64(17), object(12)
memory usage: 7.2+ MB
```

Ввод [132]:
```
normalized_df=(df1-df1.min())/(df1.max()-df1.min())
normalized_df
```

Out[132]:

	is_canceled	lead_time	arrival_date_year	stays_in_weeks_nights	is_repeated_guest	previous_cancellations	previous_bookings_not_canceled
0	0.0	0.464043	0.0	0.000000	0.0	0.0	0.0
1	0.0	1.000000	0.0	0.000000	0.0	0.0	0.0
6	0.0	0.000000	0.0	0.035714	0.0	0.0	0.0
7	0.0	0.012212	0.0	0.035714	0.0	0.0	0.0
8	1.0	0.115332	0.0	0.053571	0.0	0.0	0.0
...	...	...	...	...	...	...	...
119317	0.0	0.255088	1.0	0.071429	0.0	0.0	0.0
119340	0.0	0.149254	1.0	0.089286	0.0	0.0	0.0
119357	0.0	0.063772	1.0	0.071429	0.0	0.0	0.0
119366	0.0	0.284939	1.0	0.125000	0.0	0.0	0.0
119367	0.0	0.287653	1.0	0.125000	0.0	0.0	0.0

27453 rows × 7 columns

Ввод [134]:
```
normalized_df=(df1-df1.mean())/df1.std()
normalized_df
```

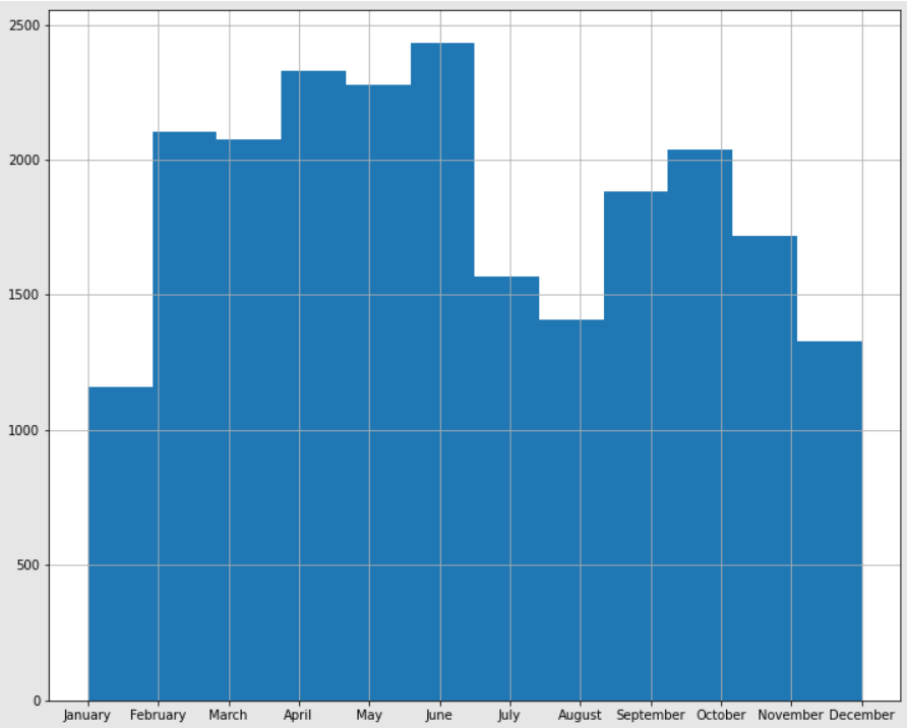
Out[134]:

	is_canceled	lead_time	arrival_date_year	stays_in_weeks_nights	is_repeated_guest	previous_cancellations	previous_bookings_not_canceled
0	-0.744536	3.159445	-1.429093	-1.112042	-0.341876	-0.149038	-0.175089
1	-0.744536	7.666437	-1.429093	-1.112042	-0.341876	-0.149038	-0.175089
6	-0.744536	-0.742812	-1.429093	-0.391894	-0.341876	-0.149038	-0.175089
7	-0.744536	-0.640121	-1.429093	-0.391894	-0.341876	-0.149038	-0.175089
8	1.343070	0.227047	-1.429093	-0.031819	-0.341876	-0.149038	-0.175089
...	...	...	...	...	...	...	...
119317	-0.744536	1.402288	1.332231	0.328255	-0.341876	-0.149038	-0.175089
119340	-0.744536	0.512300	1.332231	0.688329	-0.341876	-0.149038	-0.175089
119357	-0.744536	-0.206537	1.332231	0.328255	-0.341876	-0.149038	-0.175089
119366	-0.744536	1.653311	1.332231	1.408478	-0.341876	-0.149038	-0.175089
119367	-0.744536	1.676131	1.332231	1.408478	-0.341876	-0.149038	-0.175089

27453 rows × 7 columns

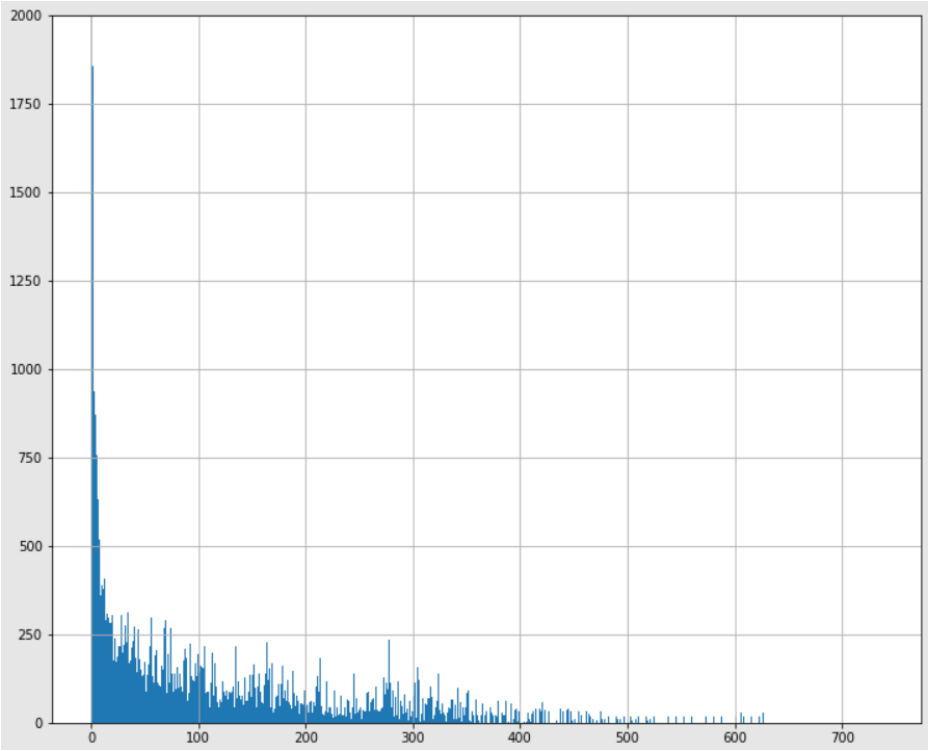
Визуализация значимых признаков

Распределение бронирований по месяцам в 2016 году в Португалии.

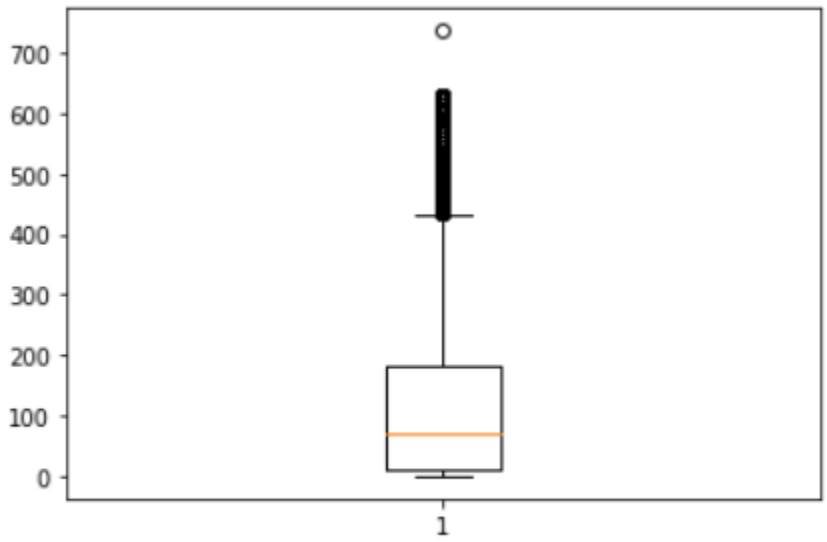


Из диаграммы видно: наиболее популярное время года – весна, осень.

Распределение величины «количество дней, прошедших между датой ввода бронирования и датой прибытия».

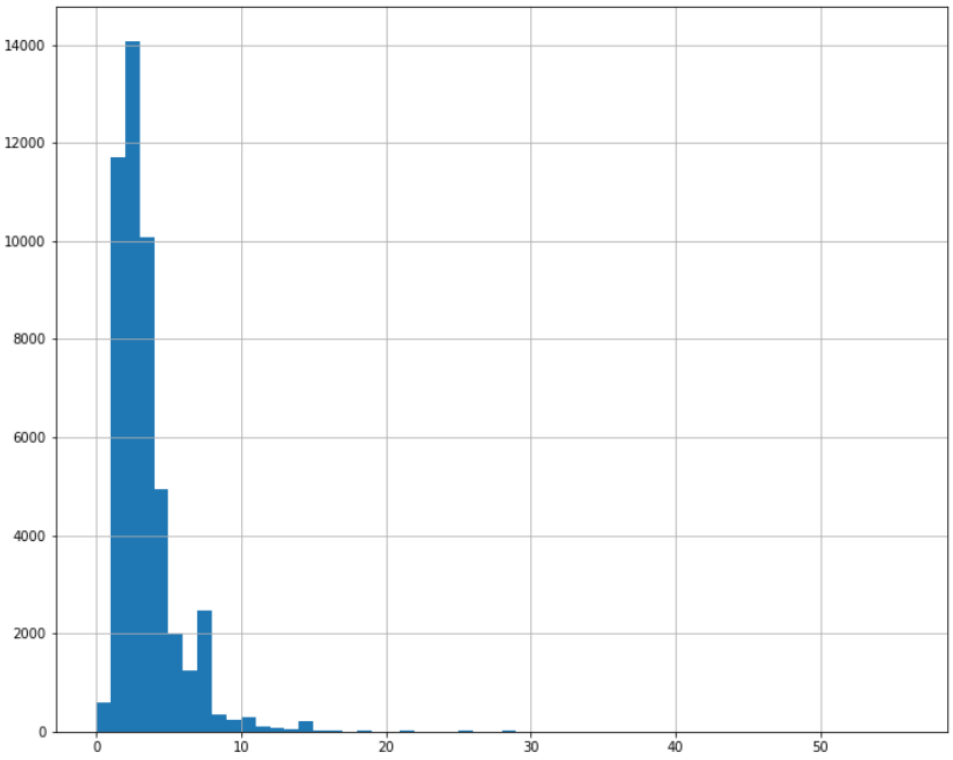


Ящик с усами величины «количество дней, прошедших между датой ввода бронирования и датой прибытия».

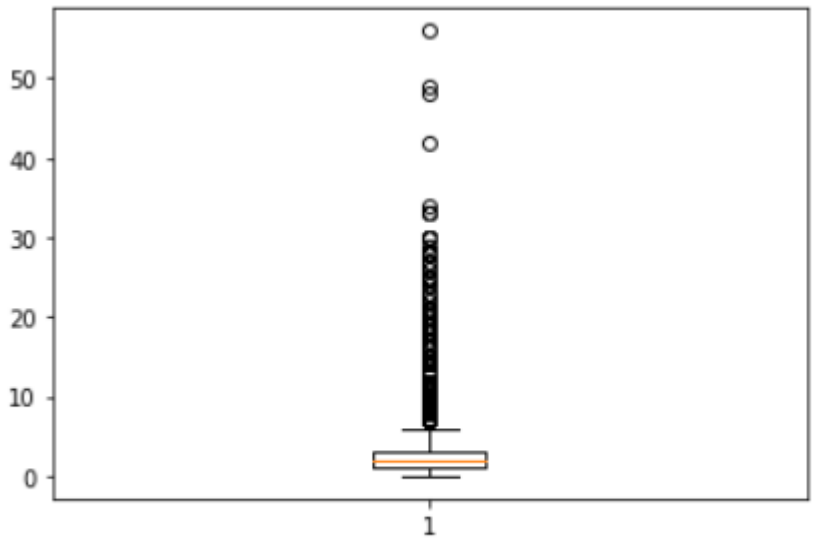



```
Медиана: 72.0, 0,25-квантиль: 12.0, 0,75-квантиль: 181.0
```

Распределение величины «количество ночей, которые гость провел или забронировал для проживания в отеле».

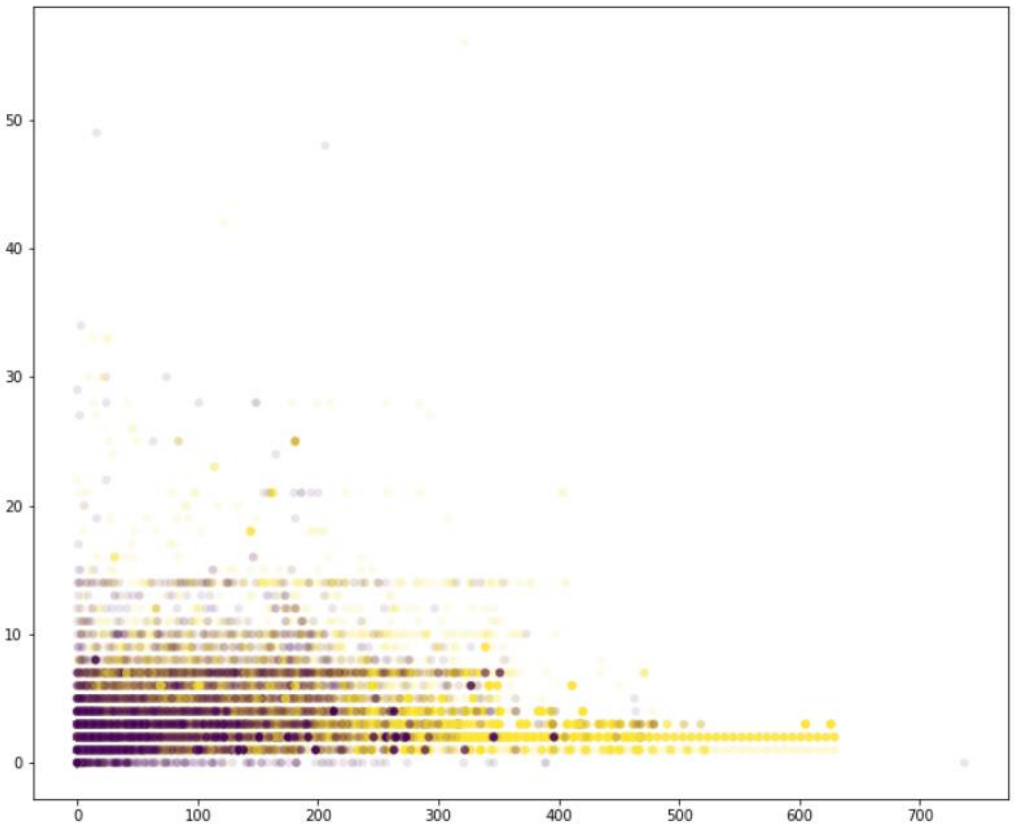


Ящик с усами величины «количество ночей, которые гость провел или забронировал для проживания в отеле».



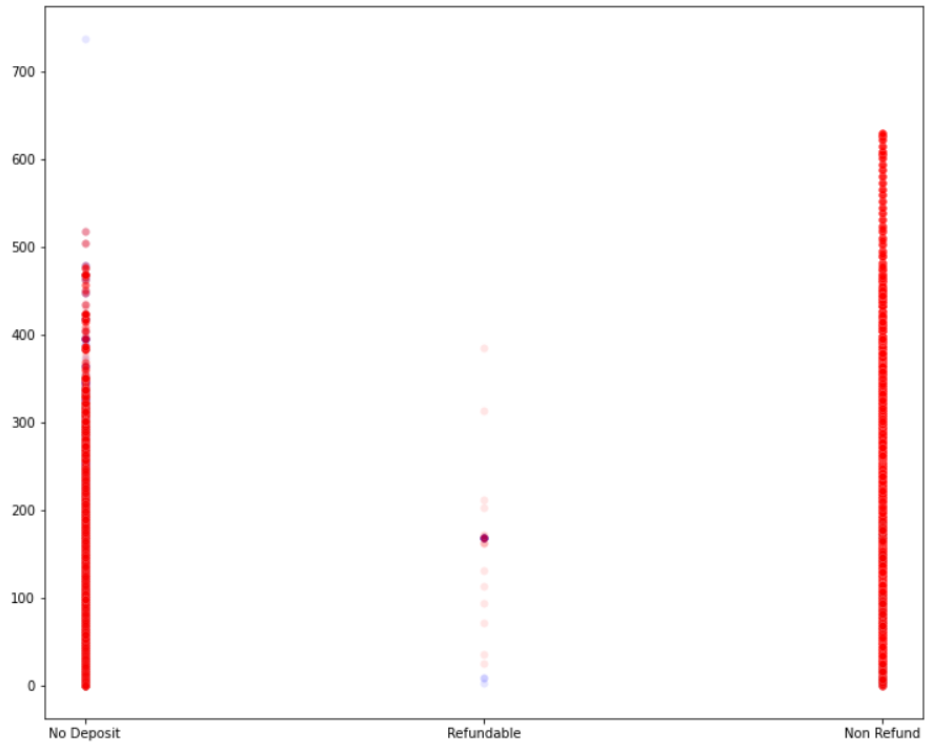
Медиана: 2.0, 0,25-квантиль: 1.0, 0,75-квантиль: 3.0

Диаграмма рассеяния. По оси X - «количество ночей, которые гость провел или забронировал для проживания в отеле», по Y - «количество дней, прошедших между датой ввода бронирования и датой прибытия». Фиолетовый – бронь не отменена, желтый – отменена.



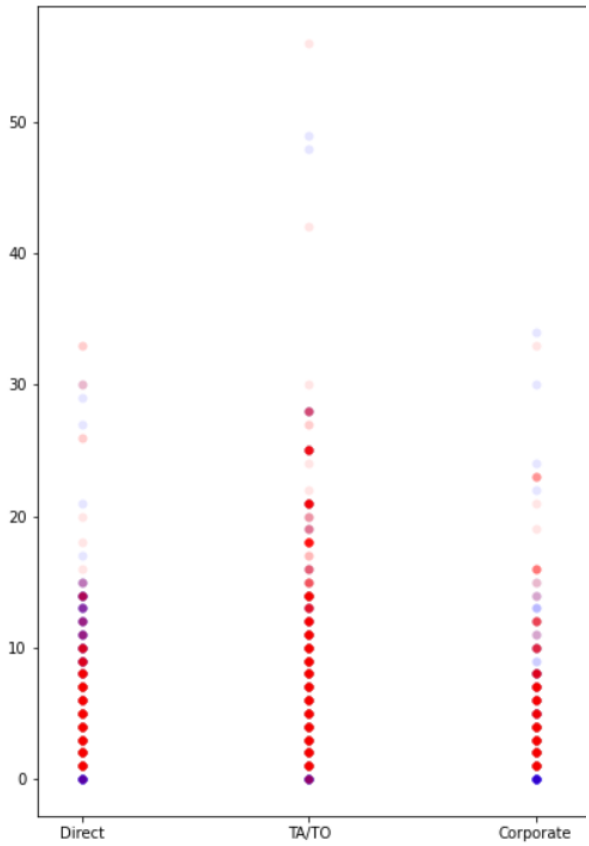
Как видно, что большинство броней, сделанных за очень долгое время до даты прибытия, были отменены. Вероятно, бронирования сделаны по ошибки.

Диаграмма рассеяния отменных броней. По оси X – тип депозита: «Без взноса», «Возвращаемый взнос», «Невозвращаемый депозит», по Y - «количество дней, прошедших между датой ввода бронирования и датой прибытия».



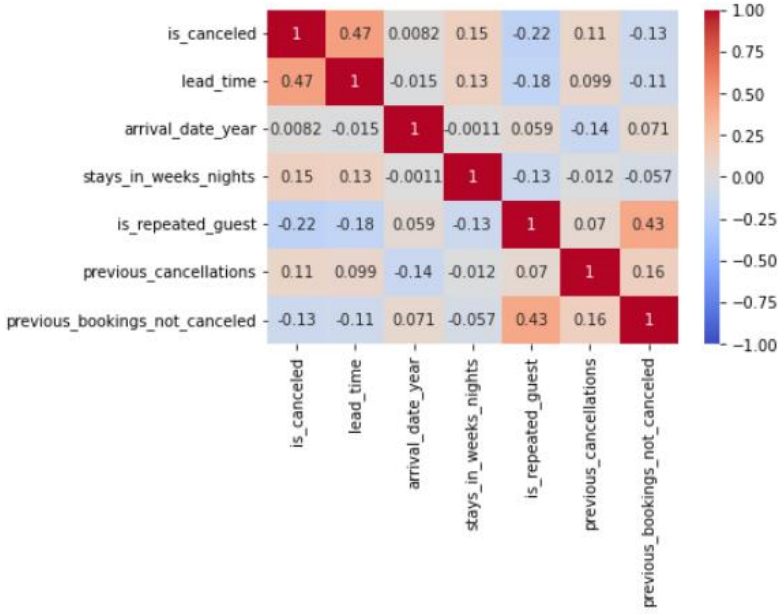
Из диаграммы видно, что очень много отмен с невозвращаемым взносом. Возможно, это связано с тем, что в 2016 в Португалии был большой поток мигрантов. Одно из условий въезда в страну – оплаченное бронирование в отеле.

Диаграмма рассеяния. По оси X - «Прямое бронирование», «Бронирование через туроператора/ турагента», по Y - «количество дней, прошедших между датой ввода бронирования и датой прибытия». Синий – бронь не отменена, красный – отменена.



Корреляция данных

Была построена матрица корреляций.



Наиболее зависимые величины: «статус отмены» и «количество дней, прошедших между датой ввода бронирования и датой прибытия». Брони, сделанные за очень долгое время до даты прибытия, были отменены, такой же вывод был получен из диаграммы рассеяния.

Также наблюдаемость зависимость между величинами: «предыдущее бронирование отменено» и «повторной гость». Те, кто уже делал бронирование в отеле, с меньшей вероятностью отменит повторную бронь в том же отеле.