Московский государственный технический университет им. Н.Э. Баумана
Факультет «Информатика и системы управления»
Кафедра «Автоматизированные системы обработки информации и управления»

Отчет

Лабораторная работа № 1

По курсу «Технологии машинного обучения»

«Разведочный анализ данных. Исследование и визуализация данных»

ИСПОЛНИТЕЛЬ:

Тарасов Владислвав
Группа ИУ5-64

_____

"__"_____2020 г.

ПРЕПОДАВАТЕЛЬ:

Гапанюк Ю.Е.

_____

"__"_____2020 г.

Москва 2020

Цель работы

Изучить различные методы визуализации данных

Задание

- Выбрать набор данных
- Создать ноутбук, который содержить следующие разделы:
  - Текстовое описание выбранного наборы данных
  - Основные характеристики датасета
  - Визуальное исследование датасета
  - Информация о корелляции признаков
- Сформировать отчет и разместить его на своем репозитории GitHub

Ход выполнения лабораторной работы

1. Набор данных

Этот набор данных содержит информацию о бронировании для городской гостиницы и курортного отеля и включает в себя такую информацию, как, например, время бронирования, продолжительность пребывания, количество взрослых, детей и / или детей и количество доступных парковочных мест.

- hotel - Отель (H1 = Курортный отель или H2 = Городской отель)
- is_canceled
- lead_time - время выполнения заказа
- 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 - Тип еды забронирован. Категории представлены в стандартных пакетах питания для гостей: Undefined / SC - без питания; BB - кровать и завтрак; HB - полупансион (завтрак и еще один прием пищи - обычно ужин); FB - полный пансион (завтрак, обед и ужин)
- country
- market_segment - Обозначение сегмента рынка. В категориях термин «TA» означает «Туристические агенты», а «TO» означает «Туроператоры».
- distribution_channel - Канал распределения бронирования. Термин «TA» означает «Туристические агенты», а «TO» означает «Туроператоры».
- is_repeated_guest
- previous_cancellations - Количество предыдущих заказов, которые были отменены клиентом до текущего бронирования
- previous_bookings_not_canceled
- reserved_room_type
- assigned_room_type
- booking_changes
- deposit_type
- agent
- company
- days_in_waiting_list
- customer_type
- adr Средняя дневная ставка, определенная путем деления суммы всех транзакций на проживание на общее количество ночей проживания.
- required_car_parking_spaces
- total_of_special_requests
- reservation_status
- reservation_status_date

In [10]:
```
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import pandas
%matplotlib inline
sns.set(style="ticks")

data = pandas.read_csv('../data/hotel_bookings.csv')
```

2. Основные харектеристики датасета

In [12]:
```
data.head(10)
```

Out[12]:

	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
5	Resort Hotel	0	14	2015	July	27	1	0
6	Resort Hotel	0	0	2015	July	27	1	0
7	Resort Hotel	0	9	2015	July	27	1	0
8	Resort Hotel	1	85	2015	July	27	1	0
9	Resort Hotel	1	75	2015	July	27	1	0

10 rows × 32 columns

In [13]:
```
data.shape
```

Out[13]:
```
(119390, 32)
```

In [14]:
```
data.columns
```

Out[14]:
```
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'],
      dtype='object')
```

In [15]:
```
data.dtypes
```

Out[15]:
```
hotel                            object
is_canceled                       int64
lead_time                         int64
arrival_date_year                 int64
arrival_date_month               object
arrival_date_week_number          int64
arrival_date_day_of_month         int64
stays_in_weekend_nights           int64
stays_in_week_nights              int64
adults                            int64
children                        float64
babies                            int64
meal                             object
country                          object
market_segment                   object
distribution_channel             object
is_repeated_guest                 int64
previous_cancellations            int64
previous_bookings_not_canceled    int64
reserved_room_type               object
assigned_room_type               object
booking_changes                   int64
deposit_type                     object
agent                           float64
company                         float64
days_in_waiting_list              int64
customer_type                    object
adr                             float64
required_car_parking_spaces       int64
total_of_special_requests         int64
reservation_status               object
reservation_status_date          object
dtype: object
```

In [16]:
```
data.describe()
```

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_
count	119390.000000	119390.000000	119390.000000	119390.000000	119390.000000	119390.000000	119390.(
mean	0.370416	104.011416	2016.156554	27.165173	15.798241	0.927599	2.5
std	0.482918	106.863097	0.707476	13.605138	8.780829	0.998613	1.5
min	0.000000	0.000000	2015.000000	1.000000	1.000000	0.000000	0.(
25%	0.000000	18.000000	2016.000000	16.000000	8.000000	0.000000	1.(
50%	0.000000	69.000000	2016.000000	28.000000	16.000000	1.000000	2.(
75%	1.000000	160.000000	2017.000000	38.000000	23.000000	2.000000	3.(
max	1.000000	737.000000	2017.000000	53.000000	31.000000	19.000000	50.(

In [17]:
```
data['is_repeated_guest'].unique()
```

Out[17]:
```
array([0, 1])
```

In [18]:
```
# Проверим наличие пустых значений
# Цикл по колонкам датасета
for col in data.columns:
    # Количество пустых значений - все значения заполнены
    temp_null_count = data[data[col].isnull()].shape[0]
    print('{} - {}'.format(col, temp_null_count))
```

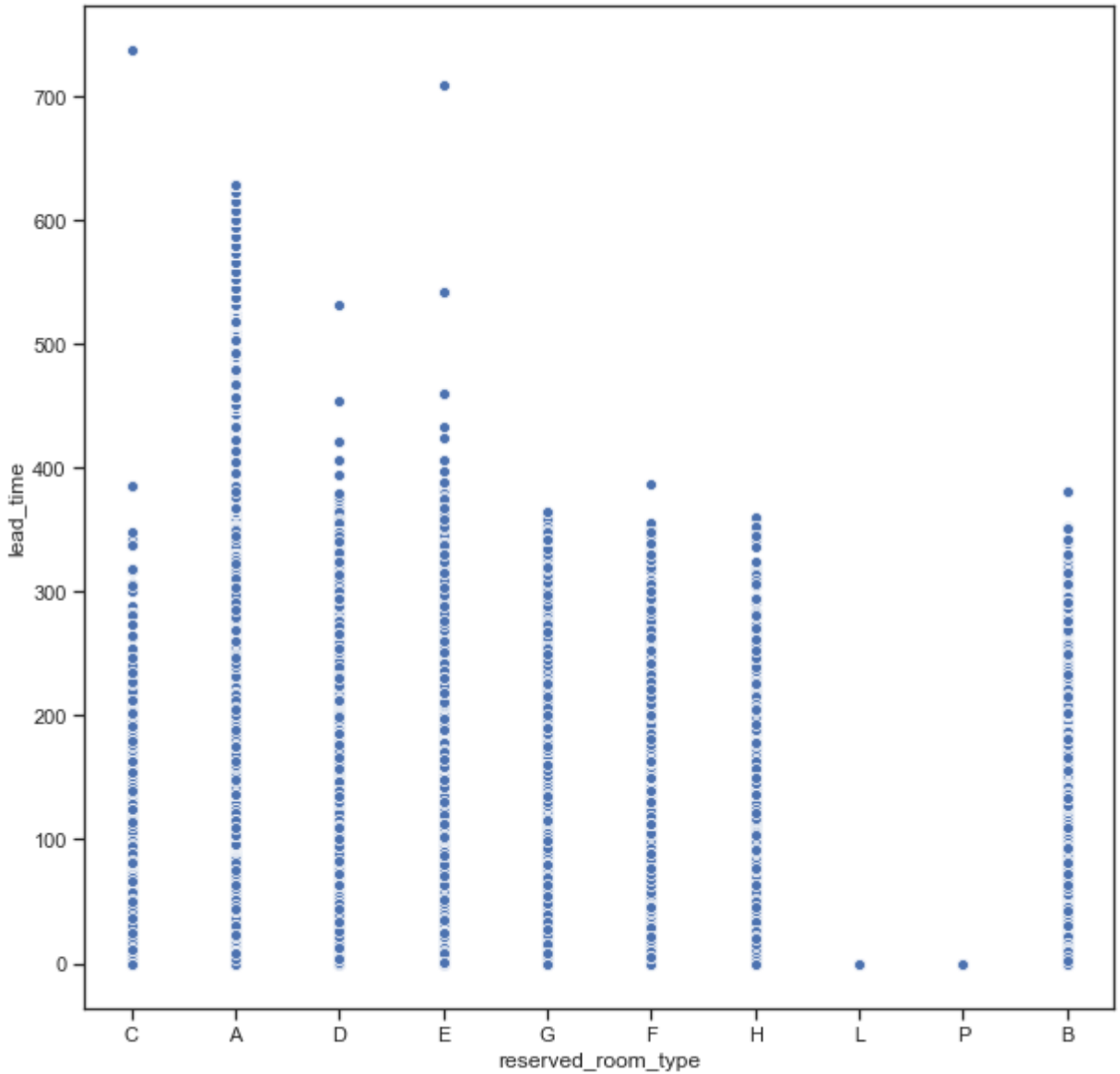
```
hotel - 0
is_canceled - 0
lead_time - 0
arrival_date_year - 0
arrival_date_month - 0
arrival_date_week_number - 0
arrival_date_day_of_month - 0
stays_in_weekend_nights - 0
stays_in_week_nights - 0
adults - 0
children - 4
babies - 0
meal - 0
country - 488
market_segment - 0
distribution_channel - 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
```

3. Визуальное исследования датасета

Диаграмма расеяния

In [30]:
```
fig, ax = plt.subplots(figsize=(10,10))
sns.scatterplot(ax=ax, x='reserved_room_type', y='lead_time', data=data)
```

Out[30]: <matplotlib.axes._subplots.AxesSubplot at 0x127dcba10>



По диаграмме расеяния можно понять, что в среденем люди бронируют раньше тип A. Но если исследовать более глубоко, то в комнату B

в среднем бронируют за более ранний срок.

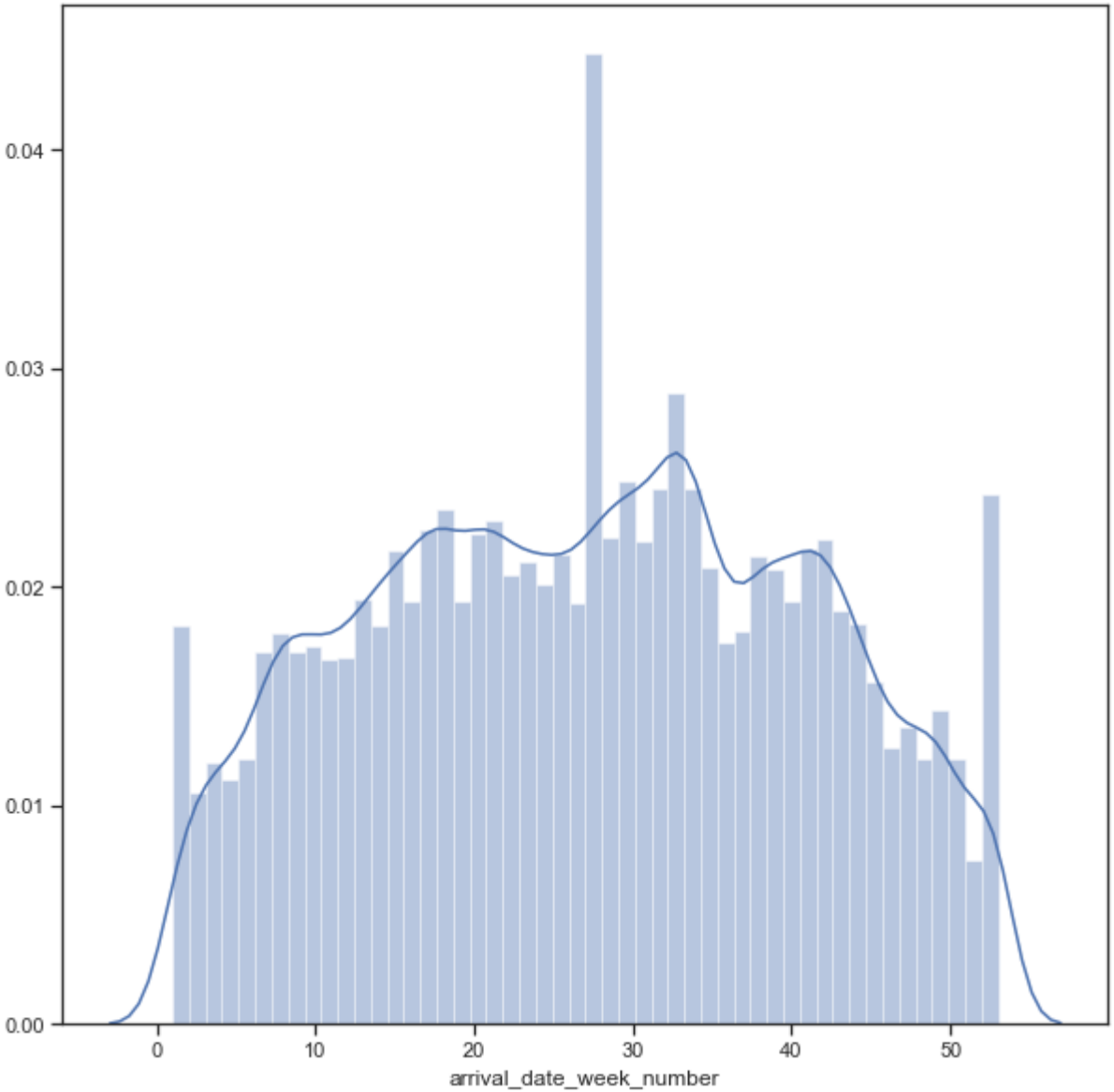
```
df[df['reserved_room_type’] == ‘A’][‘lead_time’].mean() === 110

df[df[‘reserved_room_type’] == ‘b’][‘lead_time’].mean() === 113
```

Гистограмма

In [41]:
```
fig, ax = plt.subplots(figsize=(10,10))
sns.distplot(data['arrival_date_week_number'])
```

Out[41]: <matplotlib.axes._subplots.AxesSubplot at 0x128c7cf10>

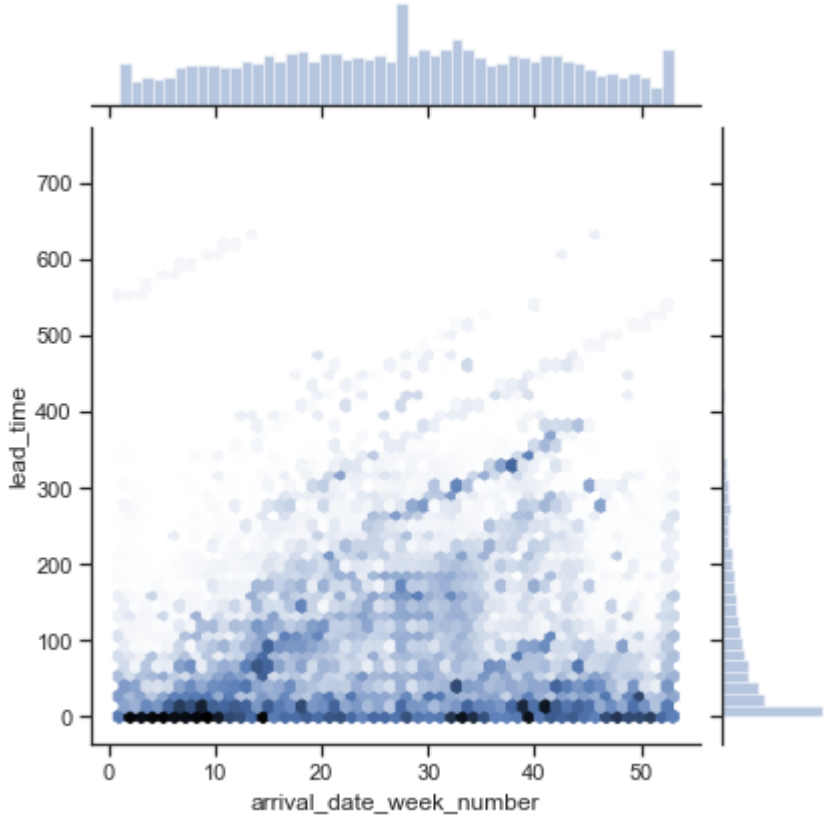


По гистограмме видно, что в 26 неделю года больше всего людей бронируют отели.

Jointplot

In [44]:
```
sns.jointplot(x='arrival_date_week_number', y='lead_time', data=data, kind="hex")
```

Out[44]:
```
<seaborn.axisgrid.JointGrid at 0x12e1373d0>
```

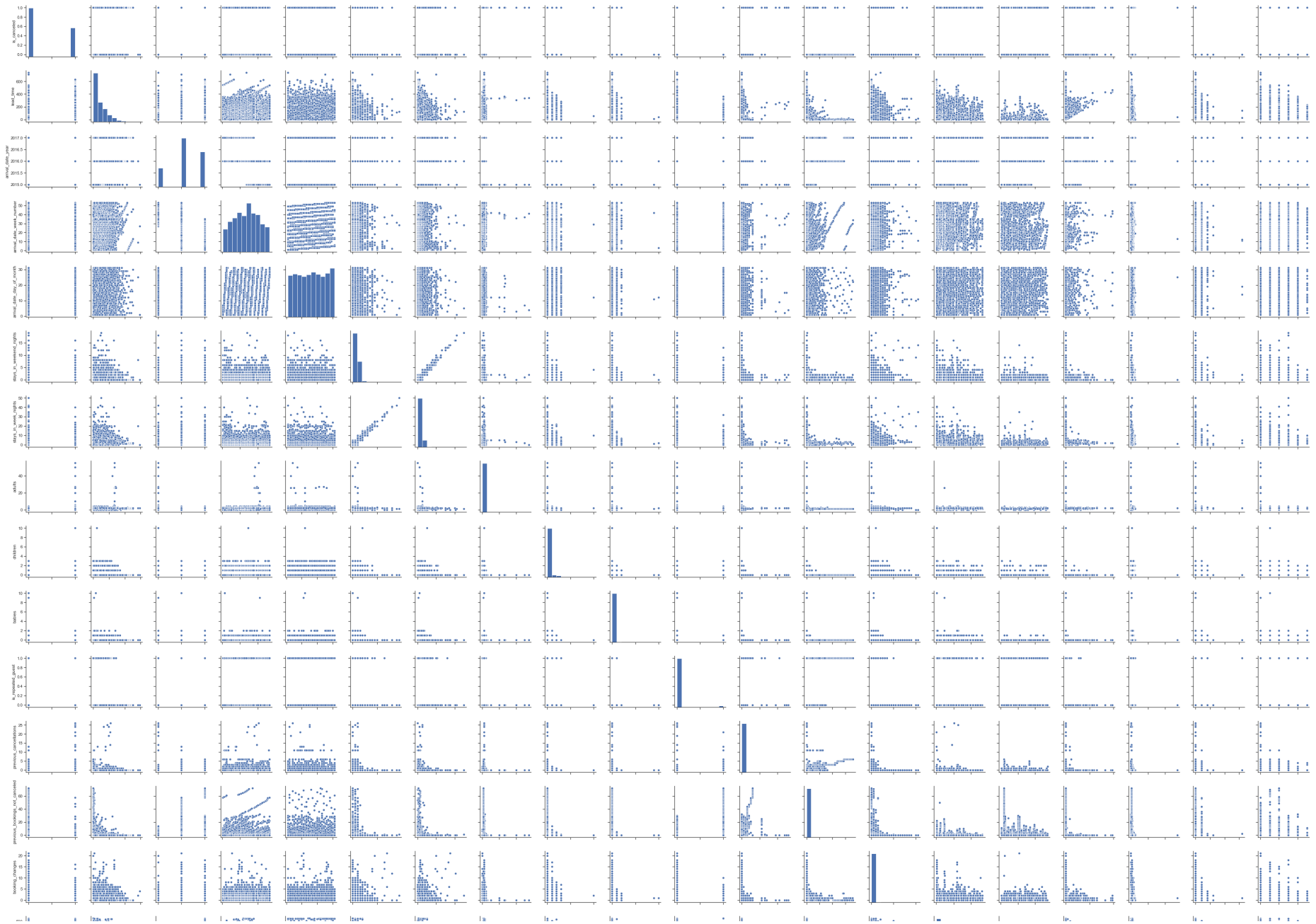


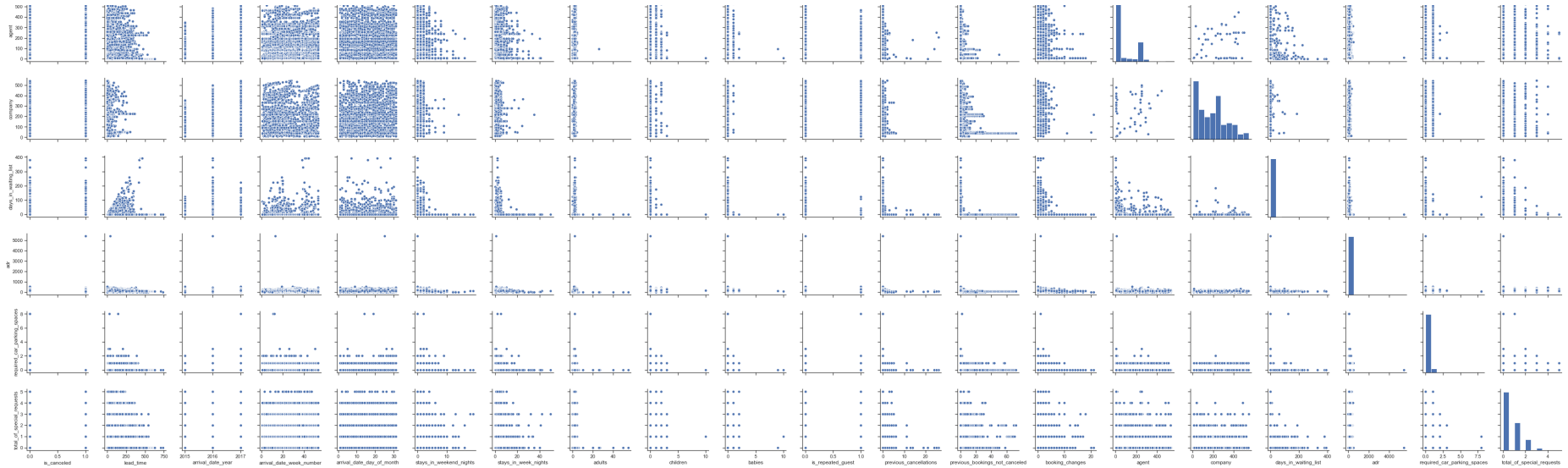
Можно сделать вывод что чем позже в году прибытие, тем дольше время прибывания.

Парная диаграмма

In [48]:
```
sns.pairplot(data)
```

Out[48]: <seaborn.axisgrid.PairGrid at 0x1eb58b210>

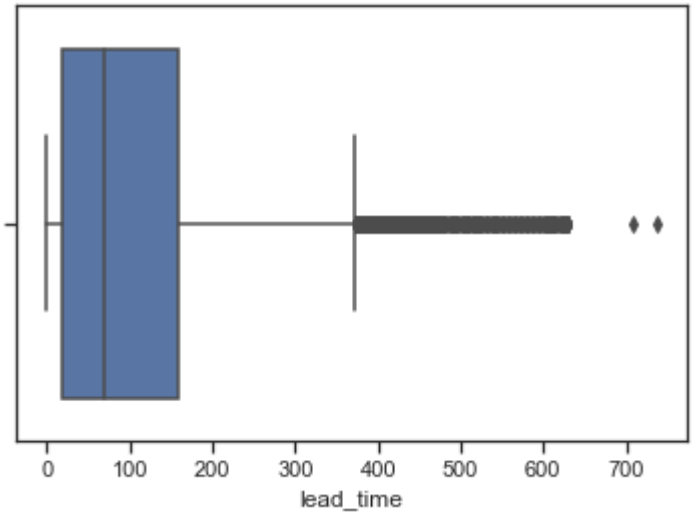




Ящик с усами

In [50]:
```
sns.boxplot(x=data['lead_time'])
```

Out[50]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x1eca81f10>
```



Информация о корреляции признаков

Построим корреляционную матрицу по всему набору данных Проверка корреляции признаков позволяет решить две задачи:

- Понять какие признаки (колонки датасета) наиболее сильно коррелируют с целевым признаком (в нашем примере это колонка "lead_team").Именно эти признаки будут наиболее информативными для моделей машинного обучения. Признаки, которые слабо коррелируют с целевым признаком, можно попробовать исключить из построения модели, иногда это повышает качество модели. Нужно отметить, что некоторые алгоритмы машинного обучения автоматически определяют ценность того или иного признака для построения модели.
- Понять какие нецелевые признаки линейно зависимы между собой. Линейно зависимые признаки, как правило, очень плохо влияют на качество моделей. Поэтому если несколько признаков линейно зависимы, то для построения модели из них выбирают какой-то один признак.

In [51]:
```
data.corr()
```

Out[51]:

	is_canceled	lead_time	arrival_date_year	arrival_date_week_number	arrival_date_day_of_month	stays_in_weekend_night
is_canceled	1.000000	0.293123	0.016660	0.008148	-0.006130	-0.0017
lead_time	0.293123	1.000000	0.040142	0.126871	0.002268	0.0856
arrival_date_year	0.016660	0.040142	1.000000	-0.540561	-0.000221	0.0214
arrival_date_week_number	0.008148	0.126871	-0.540561	1.000000	0.066809	0.0182
arrival_date_day_of_month	-0.006130	0.002268	-0.000221	0.066809	1.000000	-0.0163
stays_in_weekend_nights	-0.001791	0.085671	0.021497	0.018208	-0.016354	1.0000
stays_in_week_nights	0.024765	0.165799	0.030883	0.015558	-0.028174	0.4989
adults	0.060017	0.119519	0.029635	0.025909	-0.001566	0.0918
children	0.005048	-0.037622	0.054624	0.005518	0.014544	0.0457
babies	-0.032491	-0.020915	-0.013192	0.010395	-0.000230	0.0184
is_repeated_guest	-0.084793	-0.124410	0.010341	-0.030131	-0.006145	-0.0872

Видим, что: lead_time кореллирует с is_canceled (0.3). От того что отменена бронь или нет, сильно зависит время с покупки до вьезда

In [54]:
```
sns.heatmap(data.corr())
```

Out[54]: <matplotlib.axes._subplots.AxesSubplot at 0x229bf04d0>