In [1]:

```
#import pandas
import pandas as pd
```

In [7]:

```
#Steps to make the data messy.
#The original dataset is pretty clean, to practice the data cleaning steps, we can make

hotel_bookings_clean = pd.read_csv('hotel_bookings.csv')

hotel_bookings_messy = hotel_bookings_clean.copy()

#Change some of the int dtypes to floats and some to strings
hotel_bookings_messy['lead_time'] = hotel_bookings_messy['lead_time'].astype('str')

#Combine arrival date and month and drop separate columns
hotel_bookings_messy['arrival_date'] = hotel_bookings_messy['arrival_date_month'] + "-"
hotel_bookings_messy.drop(columns = ['arrival_date_month', 'arrival_date_year'], inplace

column_to_move = hotel_bookings_messy.pop('arrival_date')
hotel_bookings_messy.insert(3, 'arrival_date', column_to_move)

#Add random special characters
indexes = hotel_bookings_messy['hotel'].sample(10000).index.tolist()
character_to_add = '^'
hotel_bookings_messy.loc[indexes, 'hotel'] = character_to_add + hotel_bookings_messy['ho


indexes = hotel_bookings_messy['hotel'].sample(10000).index.tolist()
character_to_add = '**'
hotel_bookings_messy.loc[indexes, 'hotel'] = hotel_bookings_messy['hotel'] + character_t

indexes = hotel_bookings_messy['hotel'].sample(10000).index.tolist()
character_to_add = '\n'
hotel_bookings_messy.loc[indexes, 'hotel'] = hotel_bookings_messy['hotel'] + character_t

#Save to csv to read in the next step, comment out once you do this
#hotel_bookings_messy.to_csv('hotel_bookings_messy.csv')
```

In [9]:

```
#Read in the dataset
hotel_bookings = pd.read_csv('hotel_bookings_messy.csv')
```

In [11]:

```
#First look at the dataset
hotel_bookings
```

Out[11]:

| | Unnamed: 0 | hotel | is_canceled | lead_time | arrival_date | arrival_date_week_number | arrival_date_ |
|---|---|---|---|---|---|---|---|
| **0** | 0 | Resort Hotel | 0 | 342 | July-2015 | 27 | |
| **1** | 1 | Resort Hotel | 0 | 737 | July-2015 | 27 | |
| **2** | 2 | Resort Hotel** | 0 | 7 | July-2015 | 27 | |

| | Unnamed: 0 | hotel | is_canceled | lead_time | arrival_date | arrival_date_week_number | arrival_date_ |
|---|---|---|---|---|---|---|---|
| **3** | 3 | Resort Hotel | 0 | 13 | July-2015 | 27 | |
| **4** | 4 | Resort Hotel | 0 | 14 | July-2015 | 27 | |
| **...** | ... | ... | ... | ... | ... | ... | |
| **119385** | 119385 | City Hotel | 0 | 23 | August-2017 | 35 | |
| **119386** | 119386 | City Hotel | 0 | 102 | August-2017 | 35 | |
| **119387** | 119387 | City Hotel | 0 | 34 | August-2017 | 35 | |
| **119388** | 119388 | City Hotel** | 0 | 109 | August-2017 | 35 | |
| **119389** | 119389 | City Hotel | 0 | 205 | August-2017 | 35 | |

119390 rows × 32 columns

# Drop / Rename Columns

In [15]:

```
hotel_bookings.columns
```

Out[15]:

```
Index(['Unnamed: 0', 'hotel', 'is_canceled', 'lead_time', 'arrival_date',
       '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 [28]:

```
hotel_bookings.rename({'adults': 'num_adults', 'children': 'num_children', 'babies': 'nu
hotel_bookings.head(10)
```

Out[28]:

| | hotel | is_canceled | lead_time | arrival_date | arrival_date_week_number | arrival_date_day_of_month | s |
|---|---|---|---|---|---|---|---|
| **0** | Resort Hotel | 0 | 342 | July-2015 | 27 | 1 | |
| **1** | Resort Hotel | 0 | 737 | July-2015 | 27 | 1 | |
| **2** | Resort Hotel** | 0 | 7 | July-2015 | 27 | 1 | |

| | hotel | is_canceled | lead_time | arrival_date | arrival_date_week_number | arrival_date_day_of_month | s |
|---|---|---|---|---|---|---|---|
| **3** | Resort Hotel | 0 | 13 | July-2015 | 27 | 1 | |
| **4** | Resort Hotel | 0 | 14 | July-2015 | 27 | 1 | |
| **5** | Resort Hotel | 0 | 14 | July-2015 | 27 | 1 | |
| **6** | Resort Hotel | 0 | 0 | July-2015 | 27 | 1 | |
| **7** | Resort Hotel | 0 | 9 | July-2015 | 27 | 1 | |
| **8** | Resort Hotel | 1 | 85 | July-2015 | 27 | 1 | |
| **9** | Resort Hotel | 1 | 75 | July-2015 | 27 | 1 | |

10 rows × 31 columns

# NaNs

In [72]:

```
hotel_bookings.isnull().sum()
```

Out[72]:

```
hotel                           0
is_canceled                     0
lead_time                       0
arrival_date                    0
arrival_date_week_number        0
arrival_date_day_of_month       0
stays_in_weekend_nights         0
stays_in_week_nights            0
num_adults                      0
num_children                    0
num_babies                      0
meal                            0
country                         0
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                           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
dtype: int64
```

In [70]:

```
hotel_bookings['agent'].unique()
hotel_bookings[hotel_bookings['agent'] == 5]

hotel_bookings.fillna({ 'agent' : -1}, inplace=True)

hotel_bookings[hotel_bookings['num_children'].isna()]

hotel_bookings[hotel_bookings['country'].isna()][['hotel','is_canceled','country']]

hotel_bookings.fillna({'country': 'Unknown'}, inplace=True)

hotel_bookings.dropna(subset=['num_children'], inplace=True)

hotel_bookings.drop(columns=['company'], inplace=True)
```

# Check column data types

In [75]:

```
hotel_bookings.dtypes
```

Out[75]:

```
hotel                            object
is_canceled                       int64
lead_time                         int64
arrival_date                     object
arrival_date_week_number          int64
arrival_date_day_of_month         int64
stays_in_weekend_nights           int64
stays_in_week_nights              int64
num_adults                        int64
num_children                    float64
num_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
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 [77]:

```
hotel_bookings = hotel_bookings.astype({'is_canceled' : 'boolean','is_repeated_guest' :
hotel_bookings.dtypes
```

Out[77]:

```
hotel                            object
is_canceled                     boolean
lead_time                         int64
arrival_date                     object
arrival_date_week_number          int64
arrival_date_day_of_month         int64
stays_in_weekend_nights           int64
stays_in_week_nights              int64
num_adults                        int64
num_children                      int32
num_babies                        int64
meal                             object
country                          object
market_segment                   object
distribution_channel             object
is_repeated_guest               boolean
previous_cancellations            int64
previous_bookings_not_canceled    int64
reserved_room_type               object
assigned_room_type               object
booking_changes                   int64
deposit_type                     object
agent                             int32
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
```

# Bin Columns

In [82]:

```
hotel_bookings['lead_time'].unique()
hotel_bookings['lead_time'].describe()
```

Out[82]:

```
count    119386.000000
mean        104.014801
std         106.863286
min           0.000000
25%          18.000000
50%          69.000000
75%         160.000000
max         737.000000
Name: lead_time, dtype: float64
```

In [86]:

```
bins = [0,100,200,300,400,500,600,700,800]
labels = ['0-100','101-200','201-300','301-400','401-500','501-600','601-700','701-800']

hotel_bookings['lead_time_binned'] = pd.cut(hotel_bookings['lead_time'], bins=bins, labe
```

```
hotel_bookings[['lead_time','lead_time_binned']]
```

Out[86]:

| | lead_time | lead_time_binned |
|---|---|---|
| **0** | 342 | 301-400 |
| **1** | 737 | 701-800 |
| **2** | 7 | 0-100 |
| **3** | 13 | 0-100 |
| **4** | 14 | 0-100 |
| **...** | ... | ... |
| **119385** | 23 | 0-100 |
| **119386** | 102 | 101-200 |
| **119387** | 34 | 0-100 |
| **119388** | 109 | 101-200 |
| **119389** | 205 | 201-300 |

119386 rows × 2 columns

# Seperate columns

In [91]:

```
hotel_bookings['arrival_date_month'] = hotel_bookings['arrival_date'].str.split('-', exp
hotel_bookings['arrival_date_year'] = hotel_bookings['arrival_date'].str.split('-', expa
hotel_bookings.head(5)
```

Out[91]:

| | hotel | is_canceled | lead_time | arrival_date | arrival_date_week_number | arrival_date_day_of_month | s |
|---|---|---|---|---|---|---|---|
| **0** | Resort Hotel | False | 342 | July-2015 | 27 | 1 | |
| **1** | Resort Hotel | False | 737 | July-2015 | 27 | 1 | |
| **2** | Resort Hotel** | False | 7 | July-2015 | 27 | 1 | |
| **3** | Resort Hotel | False | 13 | July-2015 | 27 | 1 | |
| **4** | Resort Hotel | False | 14 | July-2015 | 27 | 1 | |

5 rows × 33 columns

In [97]:

```
column_to_move = hotel_bookings.pop('arrival_date_month')
hotel_bookings.insert(4, 'arrival_date_month', column_to_move)
hotel_bookings.head()
```

```
column_to_move = hotel_bookings.pop('arrival_date_year')
hotel_bookings.insert(5, 'arrival_date_year', column_to_move)
hotel_bookings.head()
```

Out[97]:

| | hotel | is_canceled | lead_time | arrival_date | arrival_date_month | arrival_date_year | arrival_date_week |
|---|---|---|---|---|---|---|---|
| 0 | Resort Hotel | False | 342 | July-2015 | July | 2015 | |
| 1 | Resort Hotel | False | 737 | July-2015 | July | 2015 | |
| 2 | Resort Hotel** | False | 7 | July-2015 | July | 2015 | |
| 3 | Resort Hotel | False | 13 | July-2015 | July | 2015 | |
| 4 | Resort Hotel | False | 14 | July-2015 | July | 2015 | |

5 rows × 33 columns

In [99]:

```
hotel_bookings['hotel'].unique()
```

Out[99]:

```
array(['Resort Hotel', 'Resort Hotel**', '^Resort Hotel',
       '^Resort Hotel**', 'Resort Hotel\n', '^Resort Hotel\n',
       '^Resort Hotel**\n', 'Resort Hotel**\n', '^City Hotel',
       'City Hotel', 'City Hotel**', 'City Hotel\n', '^City Hotel**',
       'City Hotel**\n', '^City Hotel\n', '^City Hotel**\n'], dtype=object)
```

# String cleaning

In [102]:

```
hotel_bookings['hotel'] = hotel_bookings['hotel'].replace(r"[\*\n\^]",'',regex=True)
hotel_bookings['hotel'].unique()
```

Out[102]:

```
array(['Resort Hotel', 'City Hotel'], dtype=object)
```

# Remove duplicates

In [114]:

```
hotel_bookings.loc[hotel_bookings.duplicated(keep=False)]

hotel_bookings.drop_duplicates(keep='first', inplace=True)
```

In [116]:

```
hotel_bookings.loc[hotel_bookings.duplicated(keep=False)]
```

Out[116]:

| hotel | is_canceled | lead_time | arrival_date | arrival_date_month | arrival_date_year | arrival_date_week_nu |
|---|---|---|---|---|---|---|

0 rows × 33 columns

In [ ]:

```
```

In [ ]:

```
```