EXPLORATORY DATA ANALYSIS IN PYTHON FOR ABSOLUTE BEGINNERS

datalab

1 hidden cell


```
# Import the required packages
import pandas as pd
import plotly.express as px
```

Import the data

```
# Import hotel_bookings_clean_v2.csv
df = pd.read_csv('hotel_bookings_clean_v2.csv')
df
```

	is_...	l..	arrival_date_week_number	arrival_date_day_of_month
0	0	342	27	
1	0	737	27	
2	0	7	27	
3	0	13	27	
4	0	14	27	
5	0	14	27	
6	0	0	27	
7	0	9	27	
8	1	85	27	
9	1	75	27	
10	1	23	27	
11	0	35	27	
12	0	68	27	
13	0	18	27	
14	0	37	27	
15	0	68	27	

Rows: 1,851 ⚠ Truncated from 119,210 rows Expand

Basic exploration

```
# Show dimensions
df.shape
```

```
(119210, 54)
```

```
# Are there missing values?
df.isnull().sum()
```

index
is_canceled
lead_time
arrival_date_week_number
arrival_date_day_of_month
arrival_date_month
stays_in_weekend_nights
stays_in_week_nights
adults
children
babies
is_repeated_guest
previous_cancellations
previous_bookings_not_canceled
required_car_parking_spaces
total_of_special_requests
avg_daily_rate

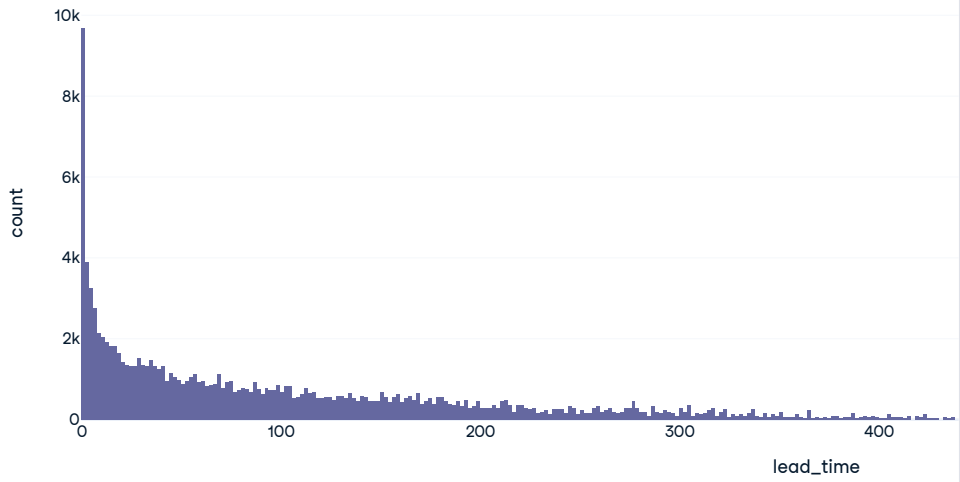
Rows: 54 Expand

```
# Describe with summary statistics
df.describe()
```

	is_ca…	lead_time	arrival_date_week_number	arrival_date_day_of_month
count	119210	119210	119210	
mean	0.3707658753	104.1092274138	27.1633755557	15.79
std	0.4830119026	106.8754499772	13.6011069633	8.78
min	0	0	1	
25%	0	18	16	
50%	0	69	28	
75%	1	161	38	
max	1	737	53	

Rows: 8 Expand

```
# Let's look into lead_time data
px.histogram(df, x='lead_time')
```



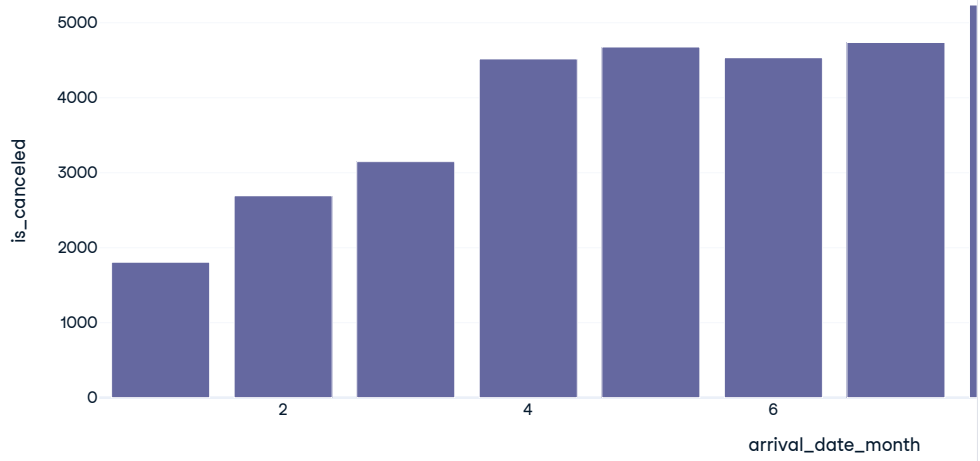

```
# How many bookings were canceled?
# How many bookings were canceled?
n_canceled = df['is_canceled'].sum()
pct_canceled = df['is_canceled'].mean()
print(f'{n_canceled} bookings were canceled, which is {round(pct_canceled * 100, 2)}% of all bookings')
```

```
44199 bookings were canceled, which is 37.08% of all bookings
```

Are the cancellation rates different during different times of the year?

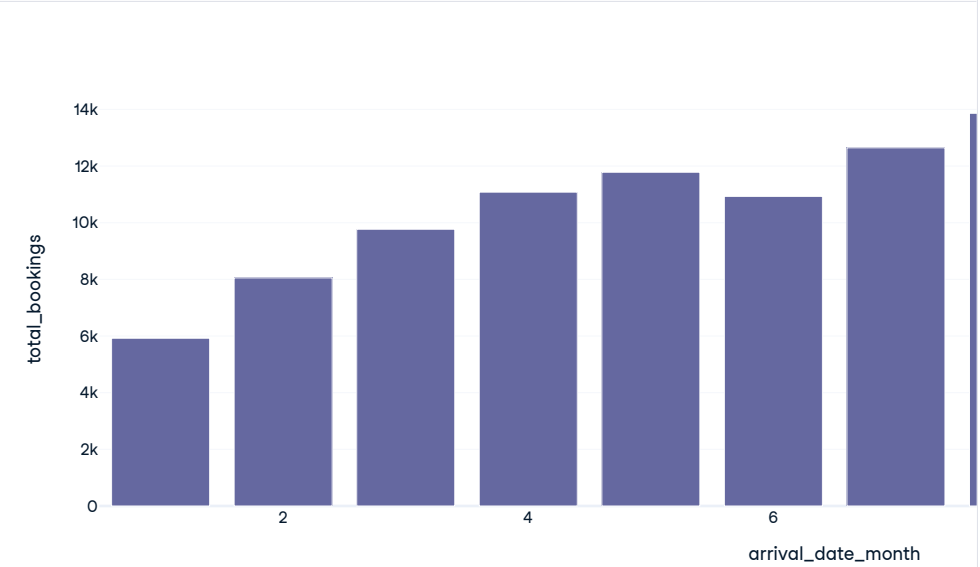
```
# Calculate and plot cancellations every month
cancellations = df\
    .filter(['arrival_date_month', 'is_canceled'])\
    .groupby(by = 'arrival_date_month', as_index=False)\
    .sum()

# Create bar chart of cancellations per month
px.bar(cancellations, x='arrival_date_month', y='is_canceled')
```



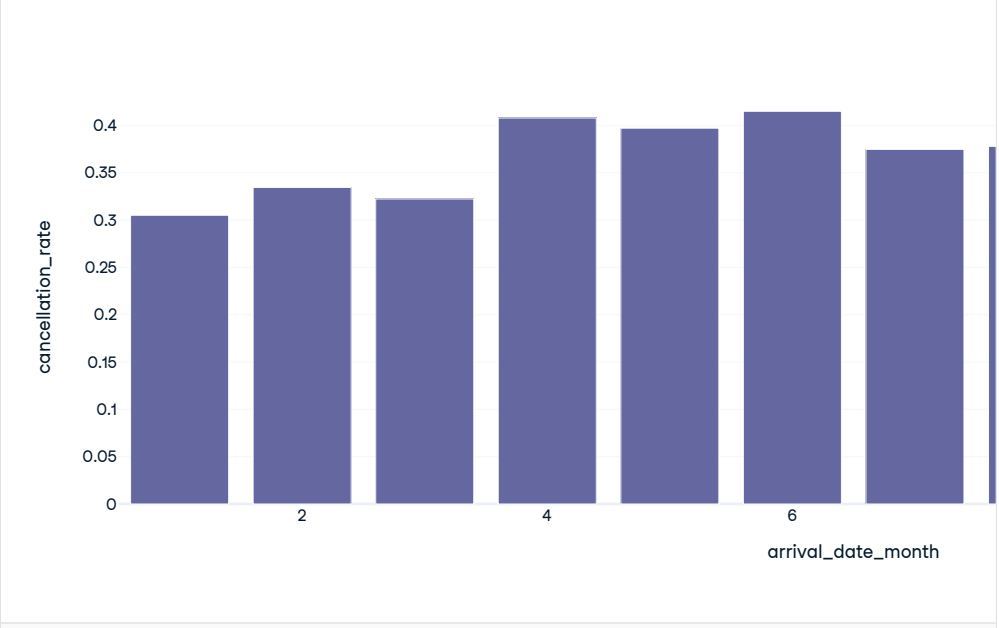

```
# Calculate and plot total bookings every month
total_bookings = df\
    .filter(['arrival_date_month', 'is_canceled'])\
    .groupby(by = 'arrival_date_month', as_index=False)\
    .count()\
    .rename(columns = {'is_canceled': 'total_bookings'})

# Create bar chart of total bookings per month
px.bar(total_bookings, x='arrival_date_month', y='total_bookings')
```



```
# Calculate cancellation rates every month
merged = pd.merge(cancellations, total_bookings, on='arrival_date_month')
merged['cancellation_rate'] = merged['is_canceled'] / merged['total_bookings']
merged

# Create bar chart of cancellation rate every month
px.bar(merged, x='arrival_date_month', y='cancellation_rate')
```



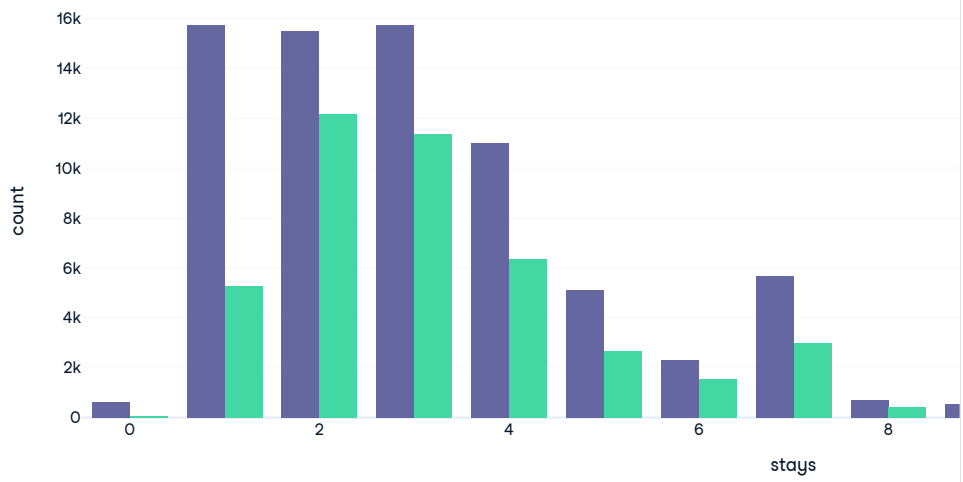
Does the amount of nights influence the cancellation rate?

```
# Prepare the data
df_sel = df\
    .assign(stays = lambda x: x['stays_in_week_nights'] + x['stays_in_weekend_nights'])\
    .query('stays < 15')
df_sel
```

	is_...	l..	arrival_date_week_number	arrival_date_day_of_month
0	0	342	27	
1	0	737	27	
2	0	7	27	
3	0	13	27	
4	0	14	27	
5	0	14	27	
6	0	0	27	
7	0	9	27	
8	1	85	27	
9	1	75	27	
10	1	23	27	
11	0	35	27	
12	0	68	27	
13	0	18	27	
14	0	37	27	
15	0	68	27	

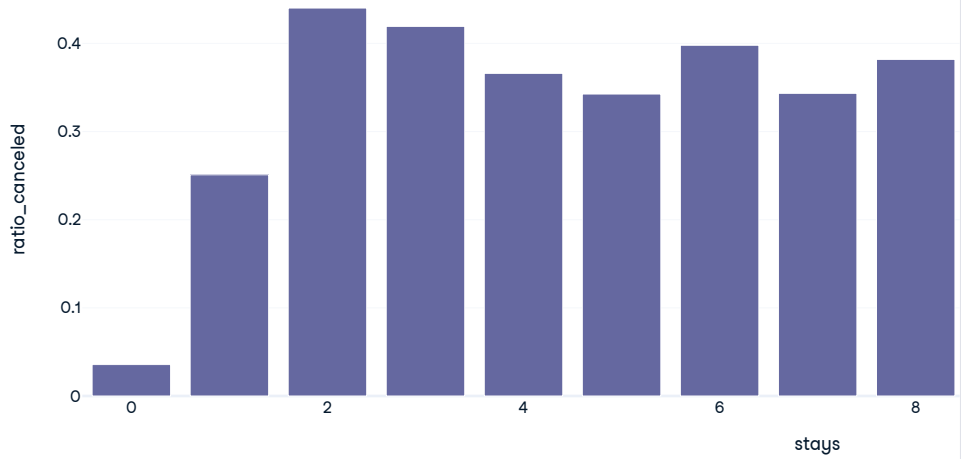
Rows: 1,818 ⚠ Truncated from 118,781 rows Expand

```
# Attempt 1: create a histogram
px.histogram(df_sel, x='stays', color='is_canceled', barmode = 'group')
```

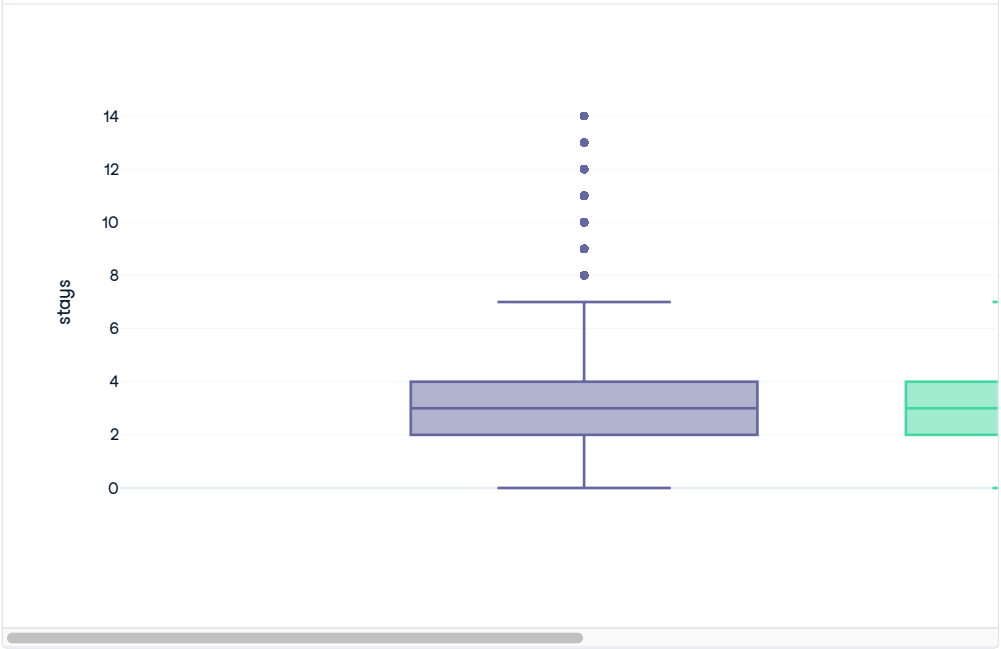



```
# Attempt 2: Calulate cancellation per days of stay
total_bookings = df_sel\
    .filter(['stays', 'is_canceled'])\
    .groupby(by = 'stays', as_index=False)\
    .count()\
    .rename(columns = {'is_canceled': 'total_bookings'})
cancellations = df_sel\
    .filter(['stays', 'is_canceled'])\
    .groupby(by = 'stays', as_index=False)\
    .sum()
merged = pd.merge(total_bookings, cancellations, on='stays')
merged['ratio_canceled'] = merged['is_canceled'] / merged['total_bookings']

# Show on bar chart
px.bar(merged, x='stays', y='ratio_canceled')
```

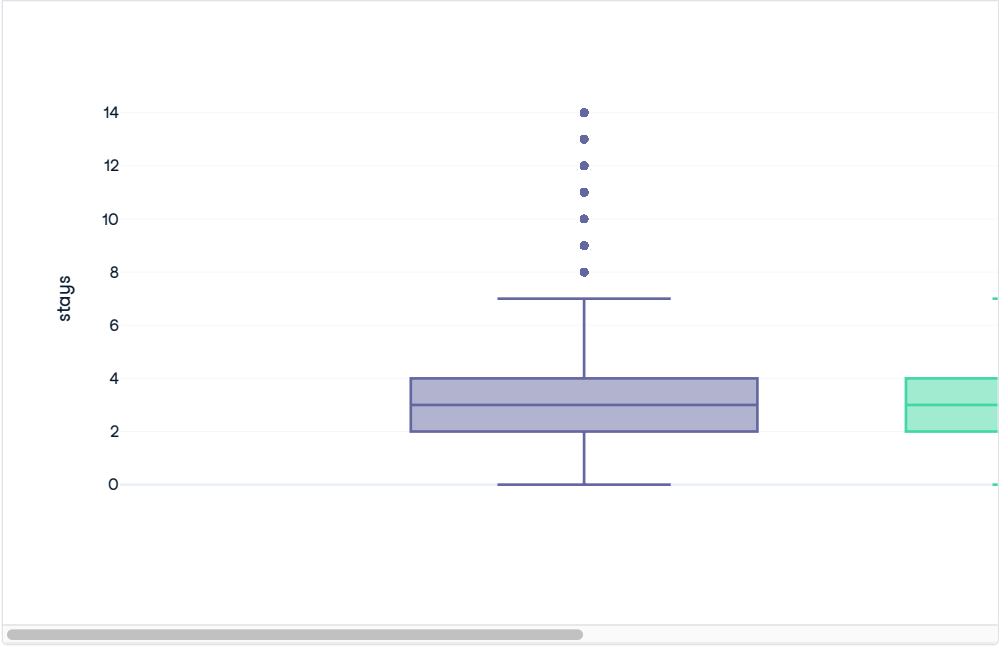


```
## Attempt 3: Boxplot
px.box(df_sel, y='stays', color='is_canceled')
```



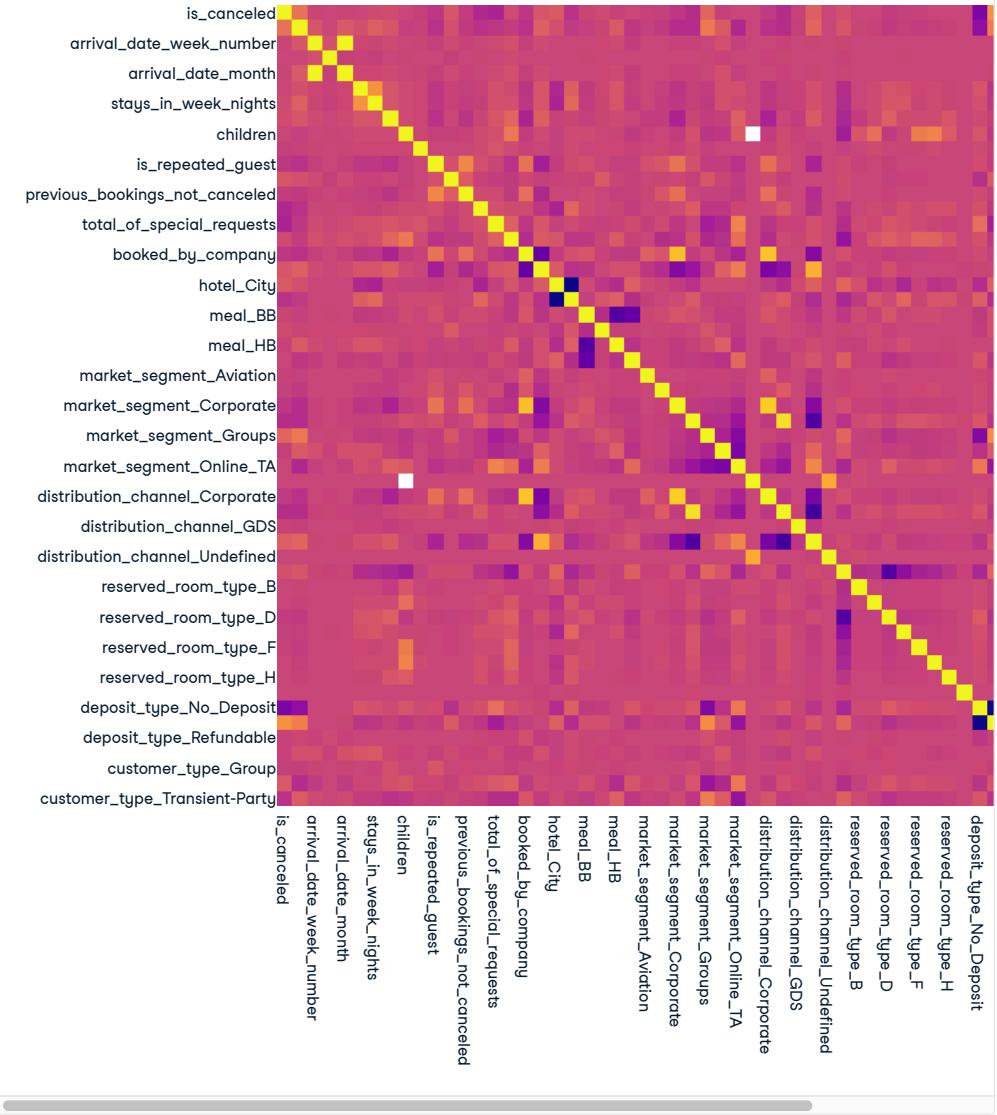
Relationship between daily rate and cancellation

```
# Box plot
df_sel2 = df.query('avg_daily_rate < 1000')
px.box(df_sel, y='stays', color='is_canceled')
```

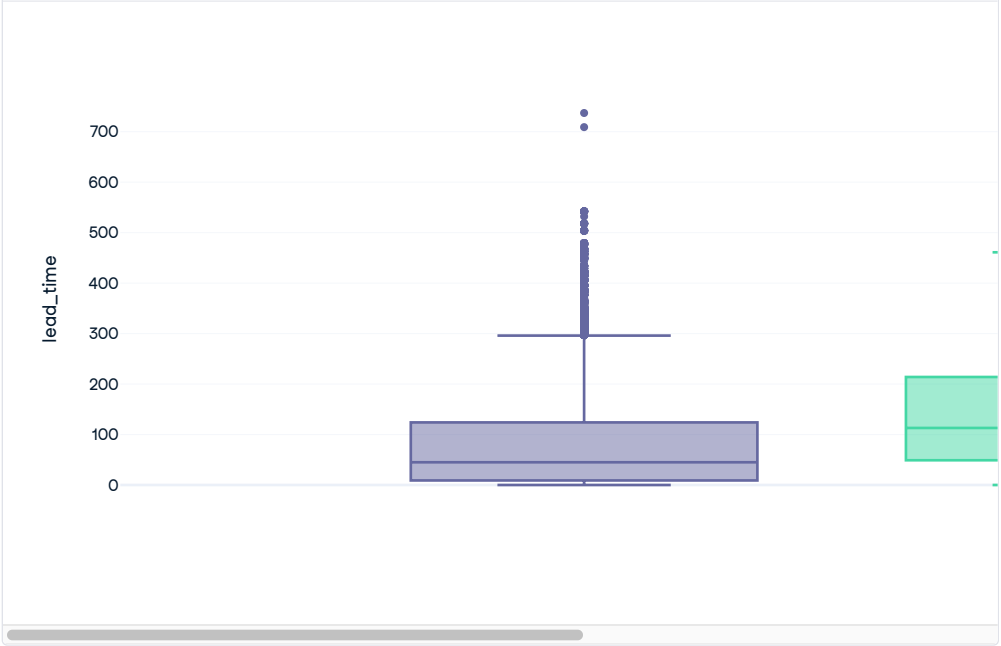


Taking a more systematic approach

```
# Build correlation plot
df.corr()
px.imshow(df.corr(), width=900, height=900)
```



```
# Boxplot of lead time vs cancellations
px.box(df_sel2, y='lead_time', color='is_canceled')
```




```
# Frequency table of cancellation vs deposit_type
freqtable = pd.crosstab(df['is_canceled'], df['deposit_type'], normalize=True)
print(freqtable)
```

```
deposit_type  full deposit made  no deposit made  partial deposit made
is_canceled
```