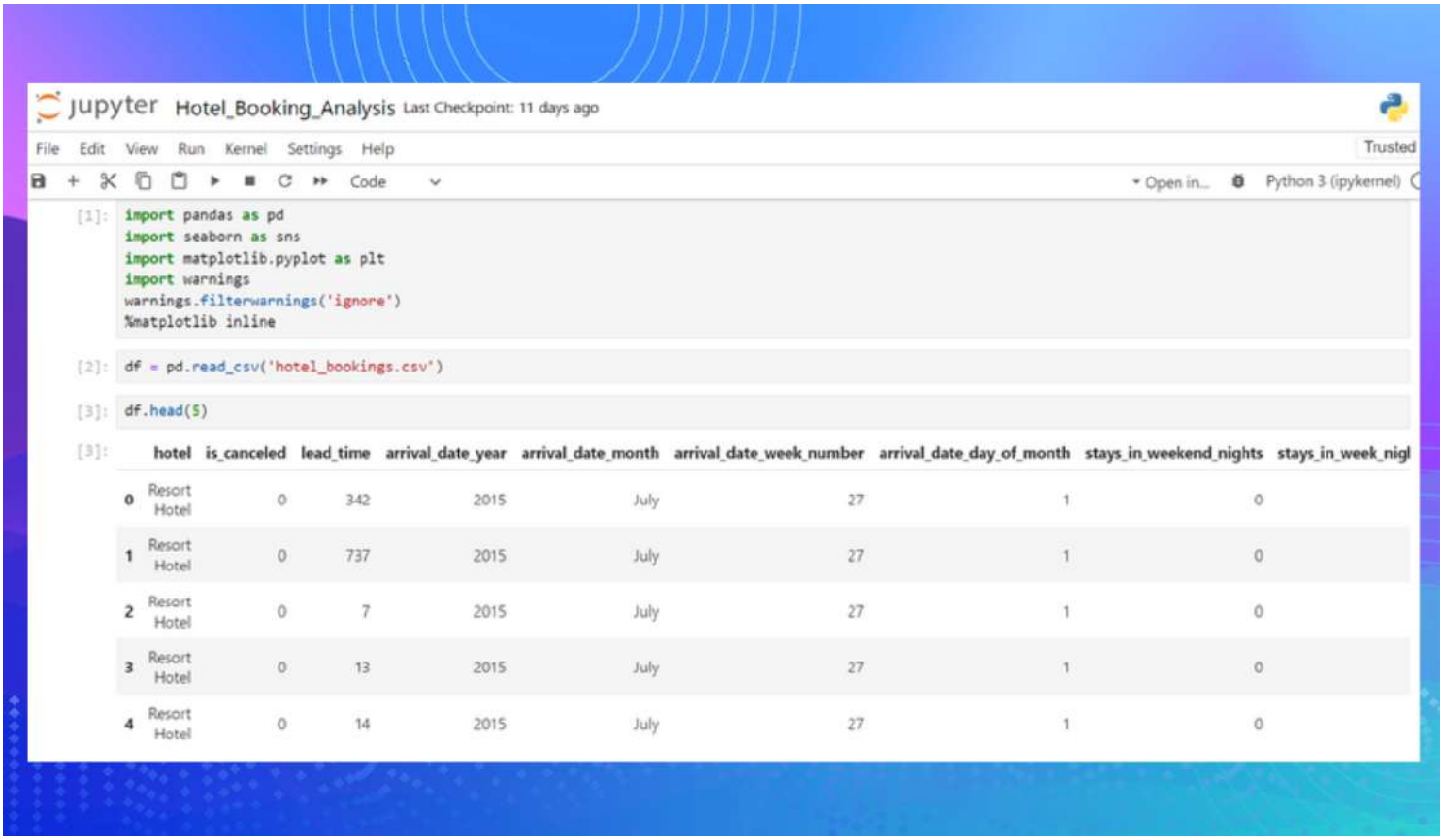
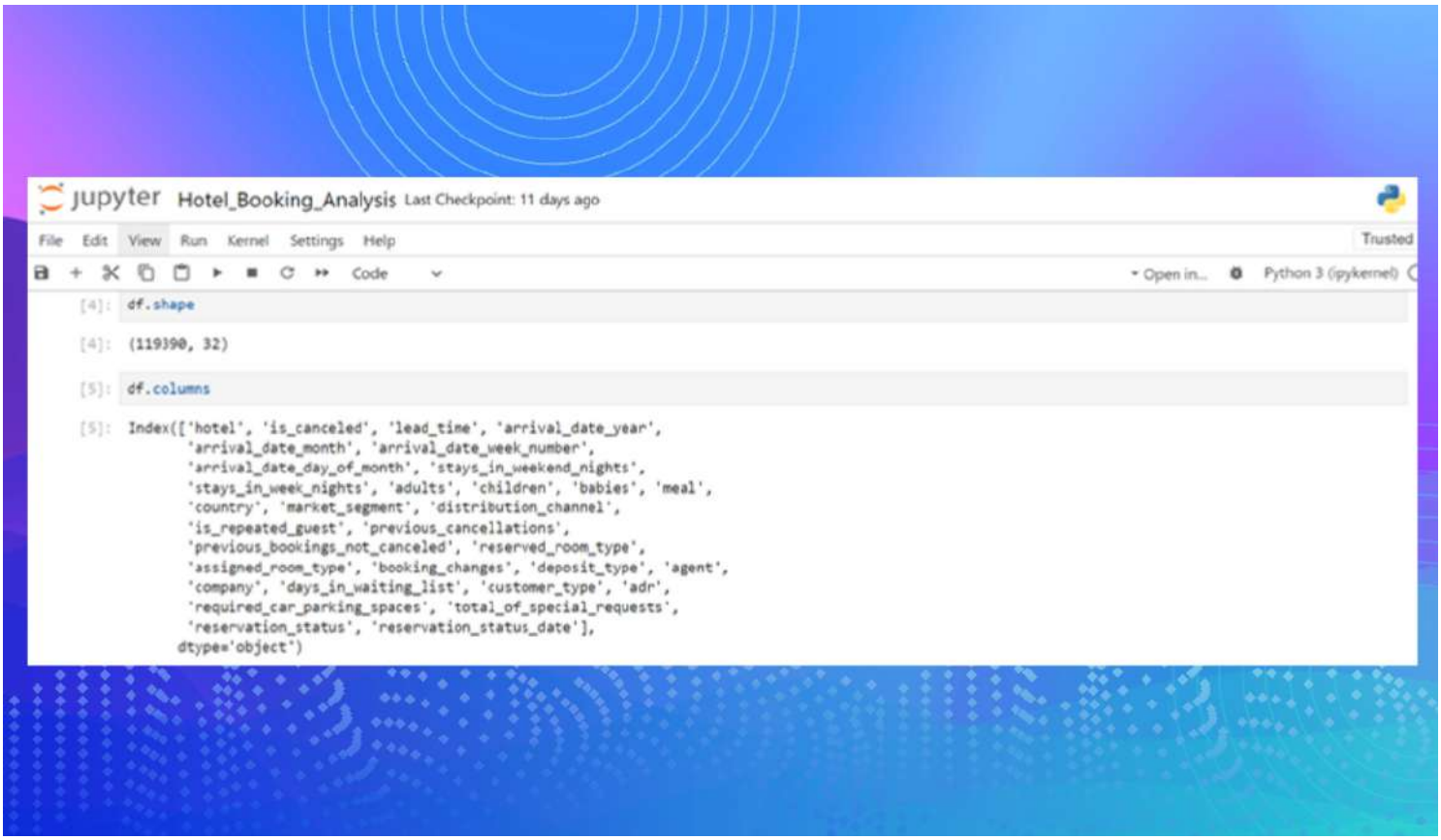
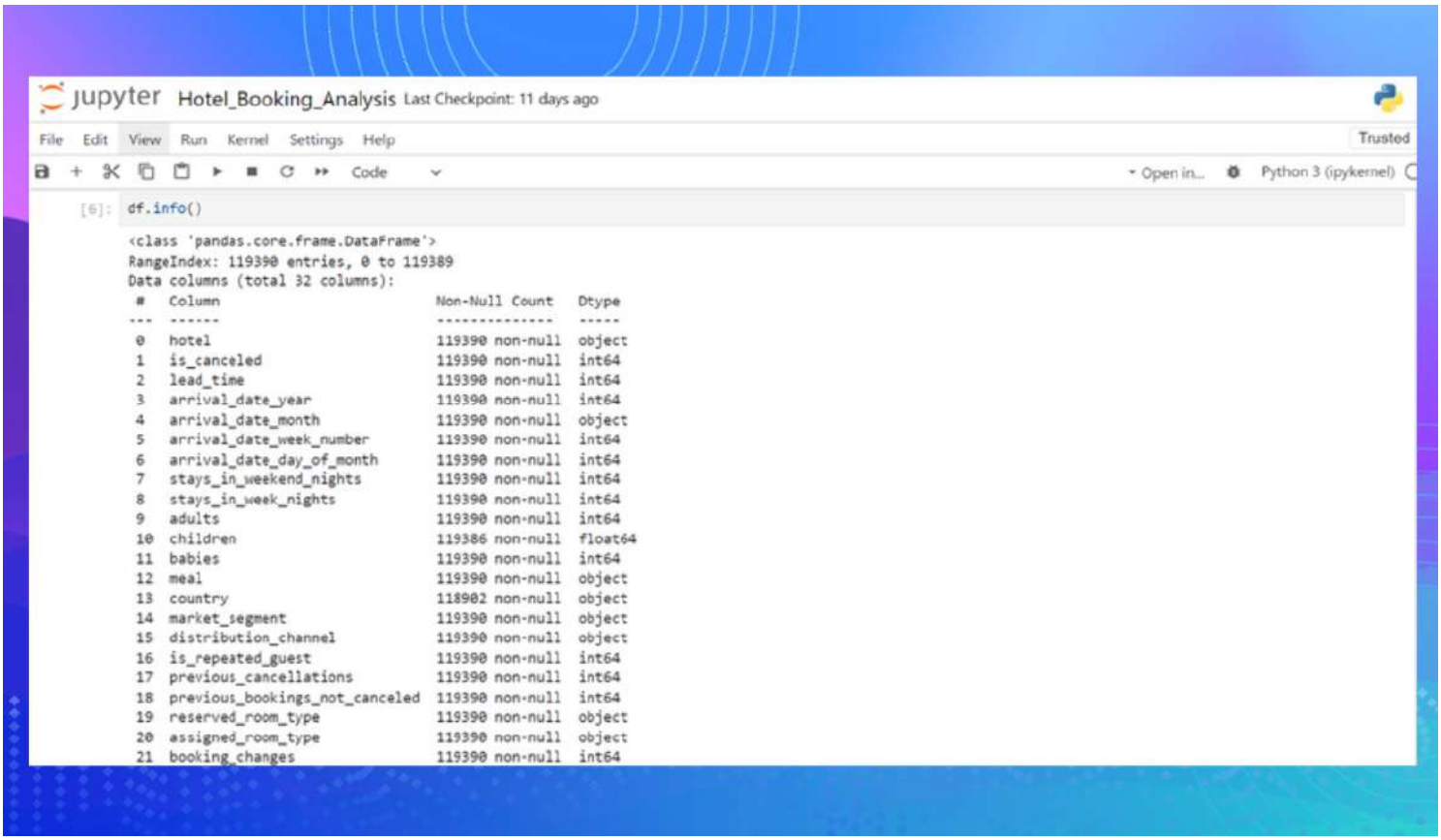
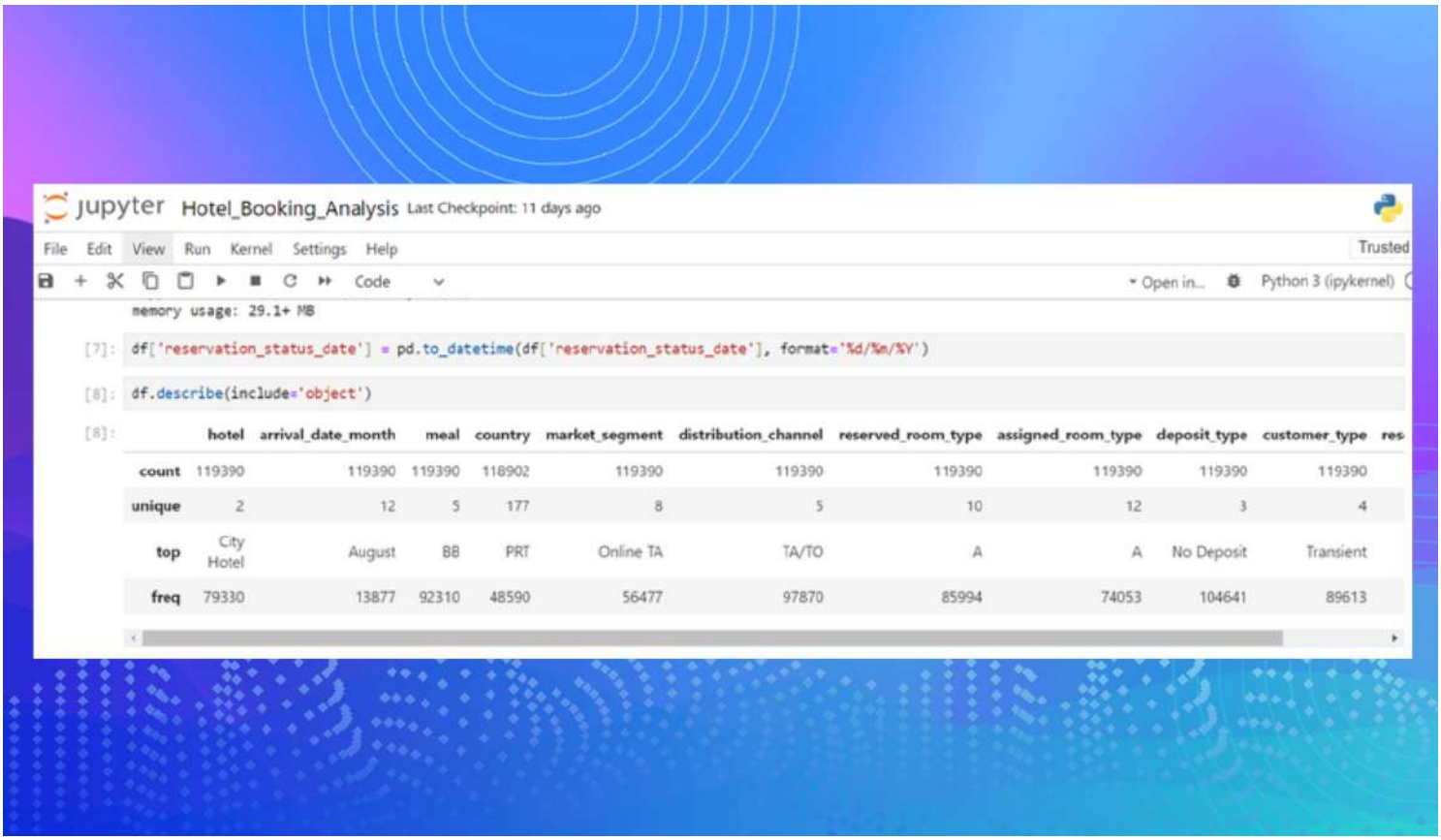
Explore and Clean the Data



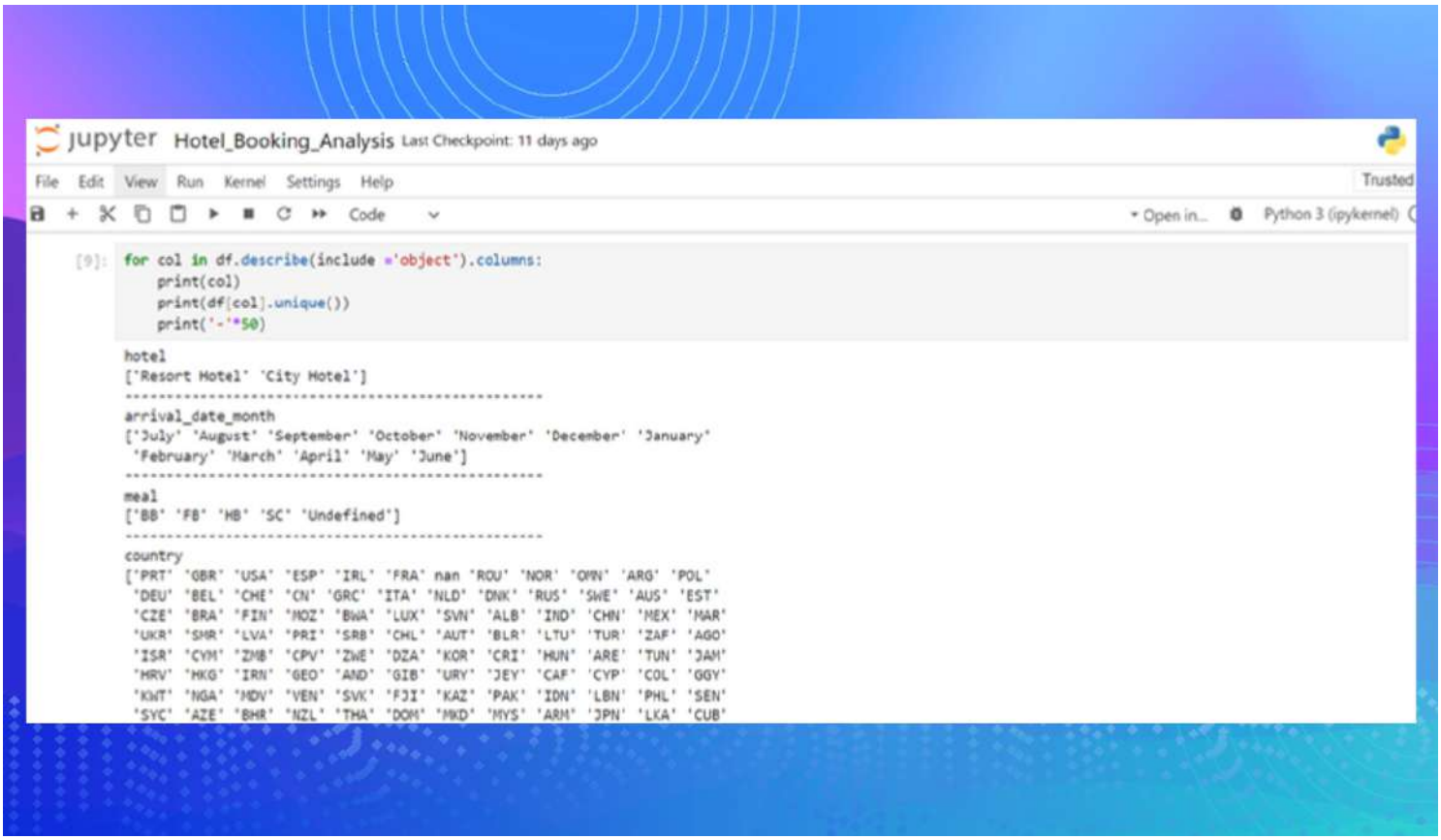
	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_day_of_month	stays_in_weekend_nights	stays_in_week_nigl
0	Resort Hotel	0	342	2015	July	27	1	0	
1	Resort Hotel	0	737	2015	July	27	1	0	
2	Resort Hotel	0	7	2015	July	27	1	0	
3	Resort Hotel	0	13	2015	July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

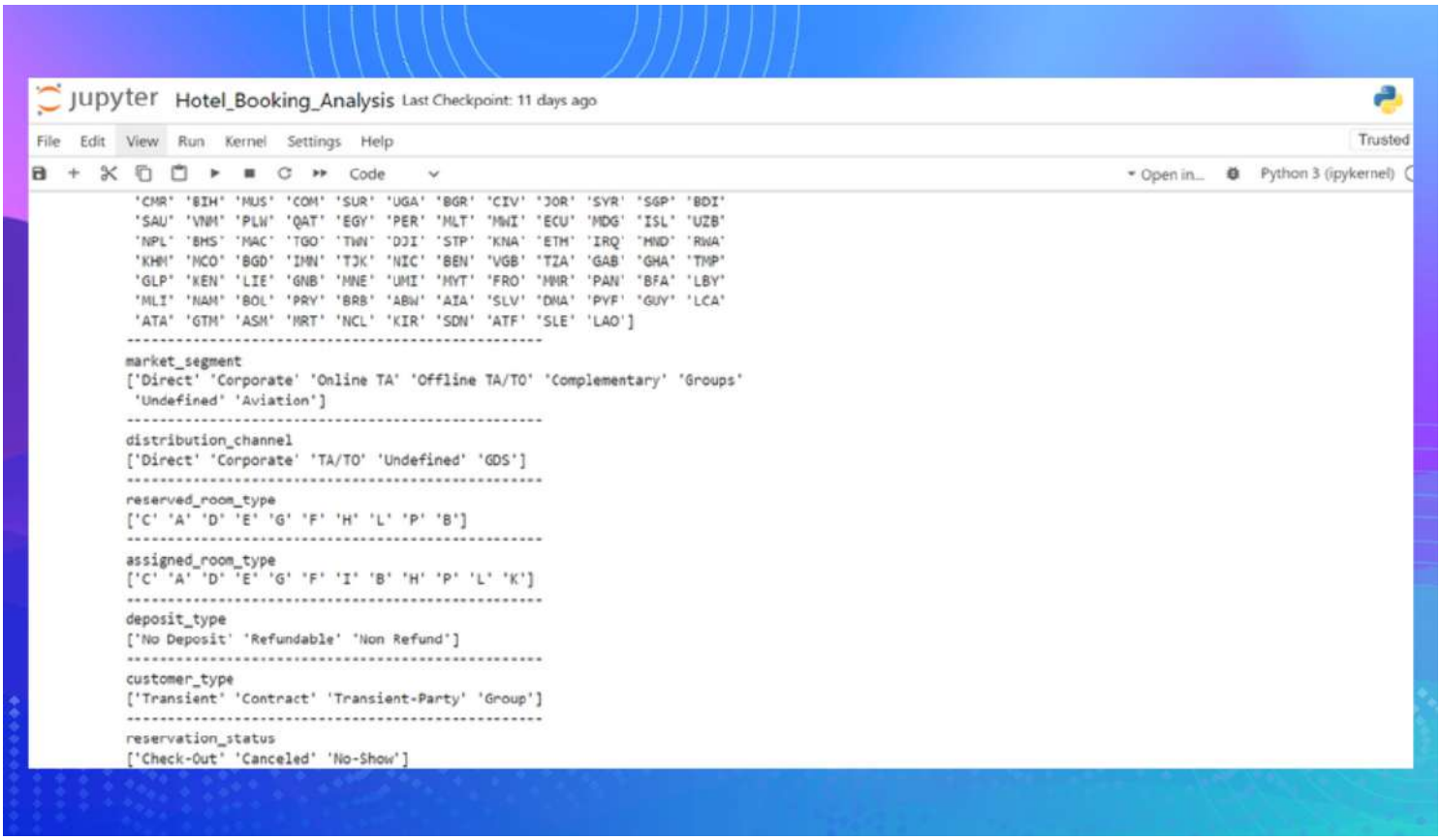

jupyter Hotel_Booking_Analysis Last Checkpoint: 11 days ago
File Edit View Run Kernel Settings Help
Trusted
Code
Open in...
Python 3 (ipykernel)
[4]: df.shape
[4]: (119390, 32)
[5]: df.columns
[5]: Index(['hotel', 'is_canceled', 'lead_time', 'arrival_date_year',
'arrival_date_month', 'arrival_date_week_number',
'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'],
dtype='object')

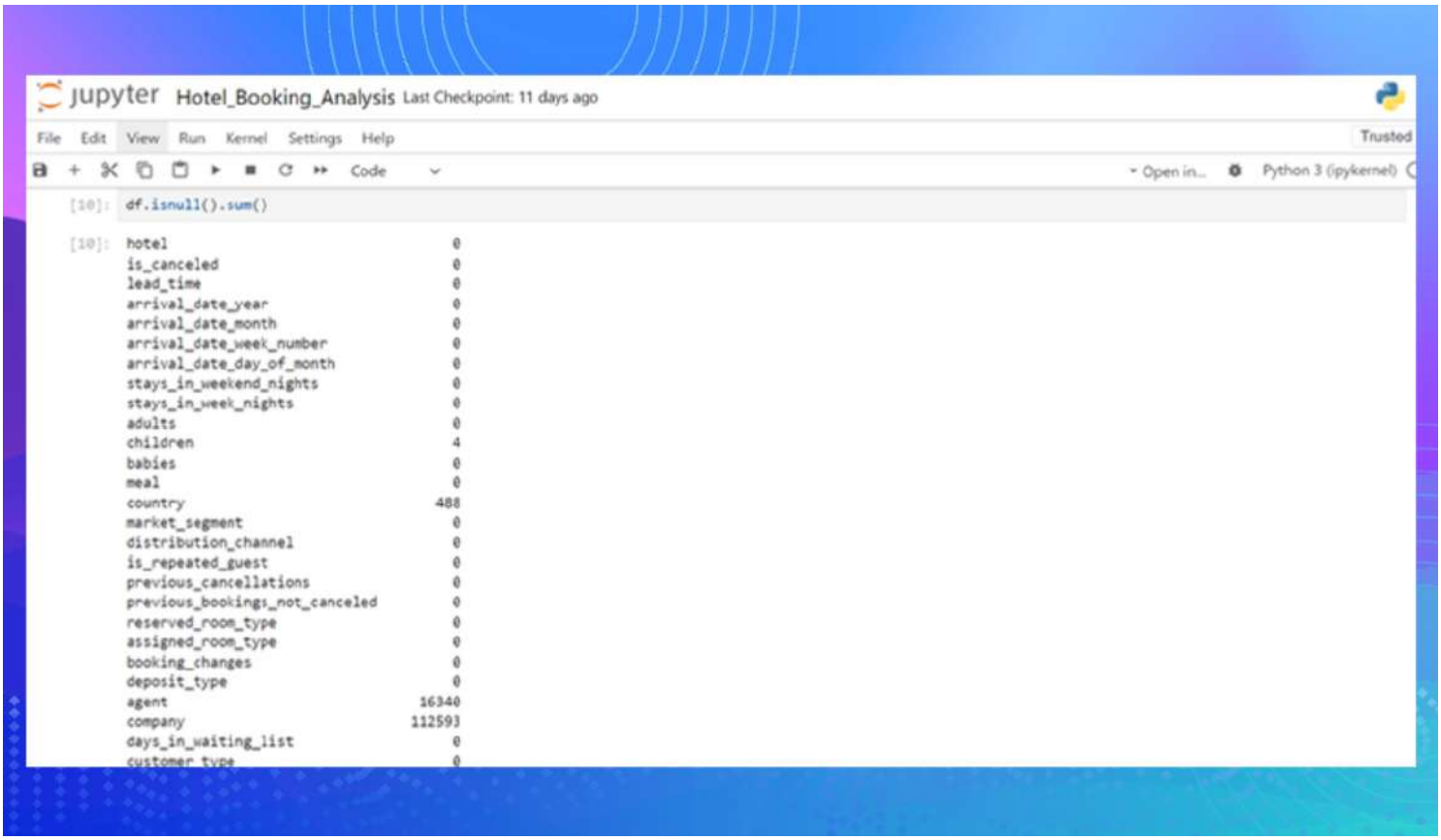

jupyter Hotel_Booking_Analysis Last Checkpoint: 11 days ago
File Edit View Run Kernel Settings Help
Trusted
Code
Open in...
Python 3 (ipykernel)
[6]: df.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 119390 entries, 0 to 119389
Data columns (total 32 columns):
Column Non-Null Count Dtype
--- ------ -------------- -----
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

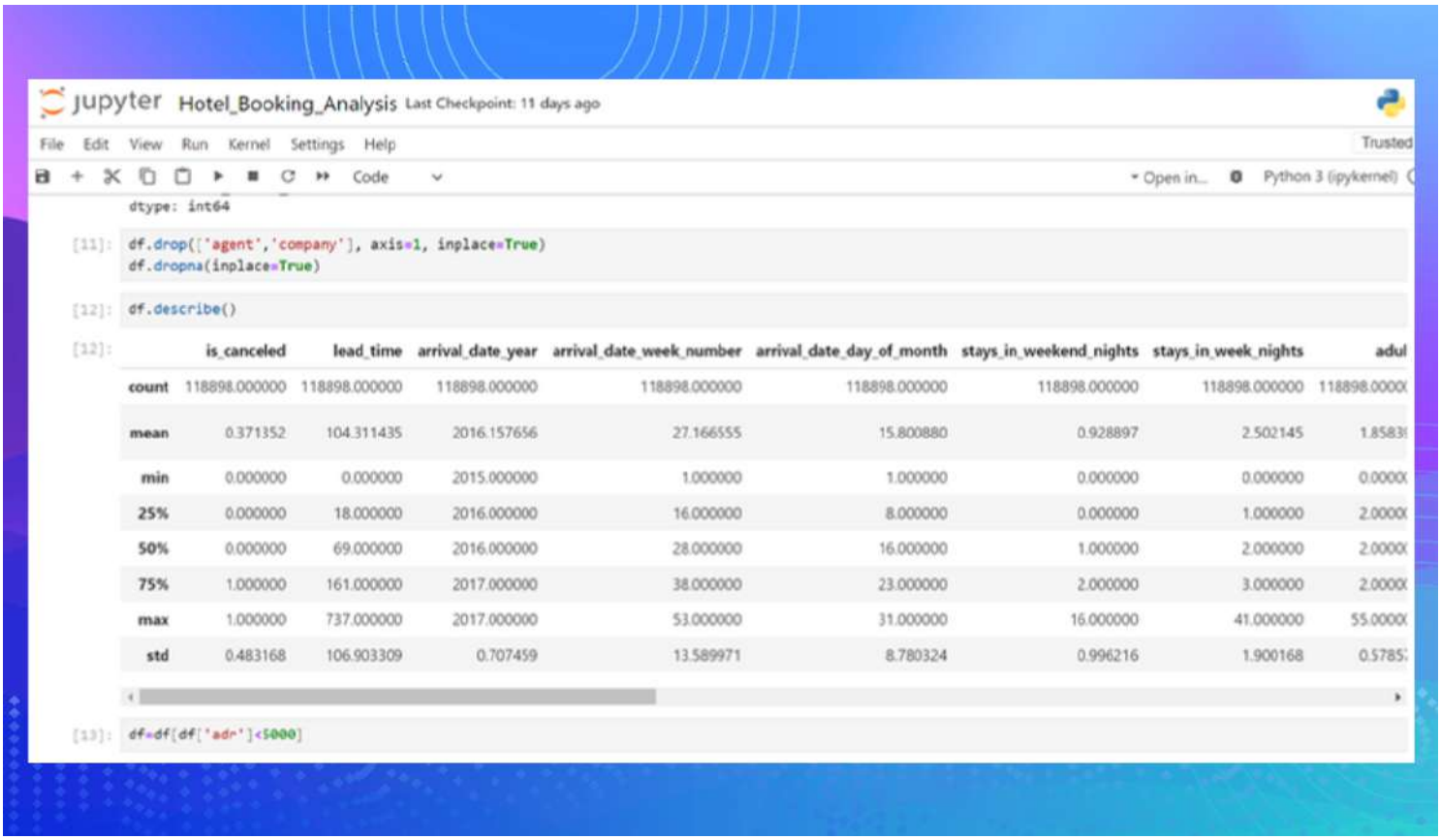


	hotel	arrival_date_month	meal	country	market_segment	distribution_channel	reserved_room_type	assigned_room_type	deposit_type	customer_type	res
count	119390	119390	119390	118902	119390	119390	119390	119390	119390	119390	
unique	2	12	5	177	8	5	10	12	3	4	
top	City Hotel	August	BB	PRT	Online TA	TA/TO	A	A	No Deposit	Transient	
freq	79330	13877	92310	48590	56477	97870	85994	74053	104641	89613	


jupyter Hotel_Booking_Analysis Last Checkpoint: 11 days ago
File Edit View Run Kernel Settings Help
Trusted
Code
Open in...
Python 3 (ipykernel)
[9]:
for col in df.describe(include ='object').columns:
print(col)
print(df[col].unique())
print('-'*50)
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' nan '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'


jupyter Hotel_Booking_Analysis Last Checkpoint: 11 days ago
File Edit View Run Kernel Settings Help
Trusted
Code
Open in...
Python 3 (ipykernel)
'CMR' 'BIH' 'MUS' 'COM' 'SUR' 'UGA' 'BGR' 'CIV' 'JOR' 'SVR' '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'
'Undefined' 'Aviation']
distribution_channel
['Direct' 'Corporate' 'TA/TO' 'Undefined' 'GDS']
reserved_room_type
['C' 'A' 'D' 'E' 'G' 'F' 'H' 'L' 'P' 'B']
assigned_room_type
['C' 'A' 'D' 'E' 'G' 'F' 'I' 'B' 'H' 'P' 'L' 'K']
deposit_type
['No Deposit' 'Refundable' 'Non Refund']
customer_type
['Transient' 'Contract' 'Transient-Party' 'Group']
reservation_status
['Check-Out' 'Canceled' 'No-Show']


jupyter Hotel_Booking_Analysis Last Checkpoint: 11 days ago
File Edit View Run Kernel Settings Help
Trusted
Code
Open in...
Python 3 (ipykernel)
[10]: df.isnull().sum()
[10]: 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


jupyter Hotel_Booking_Analysis Last Checkpoint: 11 days ago
File Edit View Run Kernel Settings Help
Trusted
Code
Open in...
Python 3 (ipykernel)
dtype: int64
[11]: df.drop(['agent','company'], axis=1, inplace=True)
df.dropna(inplace=True)
[12]: df.describe()
[12]:
is_canceled lead_time arrival_date_year arrival_date_week_number arrival_date_day_of_month stays_in_weekend_nights stays_in_week_nights adul
count 118898.000000 118898.000000 118898.000000 118898.000000 118898.000000 118898.000000 118898.000000 118898.0000
mean 0.371352 104.311435 2016.157656 27.166555 15.800880 0.928897 2.502145 1.8583
min 0.000000 0.000000 2015.000000 1.000000 1.000000 0.000000 0.000000 0.0000
25% 0.000000 18.000000 2016.000000 16.000000 8.000000 0.000000 1.000000 2.0000
50% 0.000000 69.000000 2016.000000 28.000000 16.000000 1.000000 2.000000 2.0000
75% 1.000000 161.000000 2017.000000 38.000000 23.000000 2.000000 3.000000 2.0000
max 1.000000 737.000000 2017.000000 53.000000 31.000000 16.000000 41.000000 55.0000
std 0.483168 106.903309 0.707459 13.589971 8.780324 0.996216 1.900168 0.5785
[13]: df=df[df['adr']<5000]

Reservation Status Count

jupyter Hotel_Booking_Analysis Last Checkpoint: 11 days ago

File Edit View Run Kernel Settings Help

Trusted

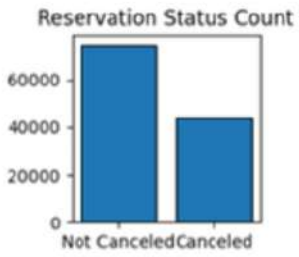
Code

Open in... Python 3 (ipykernel)

[14]:
```
cancelled_percentage =df['is_canceled'].value_counts(normalize=True)
print(cancelled_percentage)
plt.figure(figsize=(2,2))
plt.title('Reservation Status Count')
plt.bar(['Not Canceled','Canceled'],df['is_canceled'].value_counts(), edgecolor='k', width=0.8)
plt.show
```

```
is_canceled
0    0.628653
1    0.371347
Name: proportion, dtype: float64
```

[14]:
```
<function matplotlib.pyplot.show(close=None, block=None)>
```



Reservation status in different hotels

jupyter Hotel_Booking_Analysis Last Checkpoint: 11 days ago

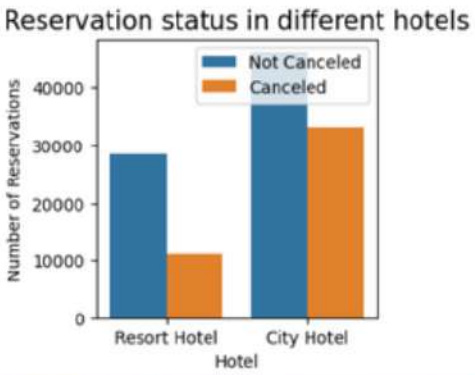
File Edit View Run Kernel Settings Help

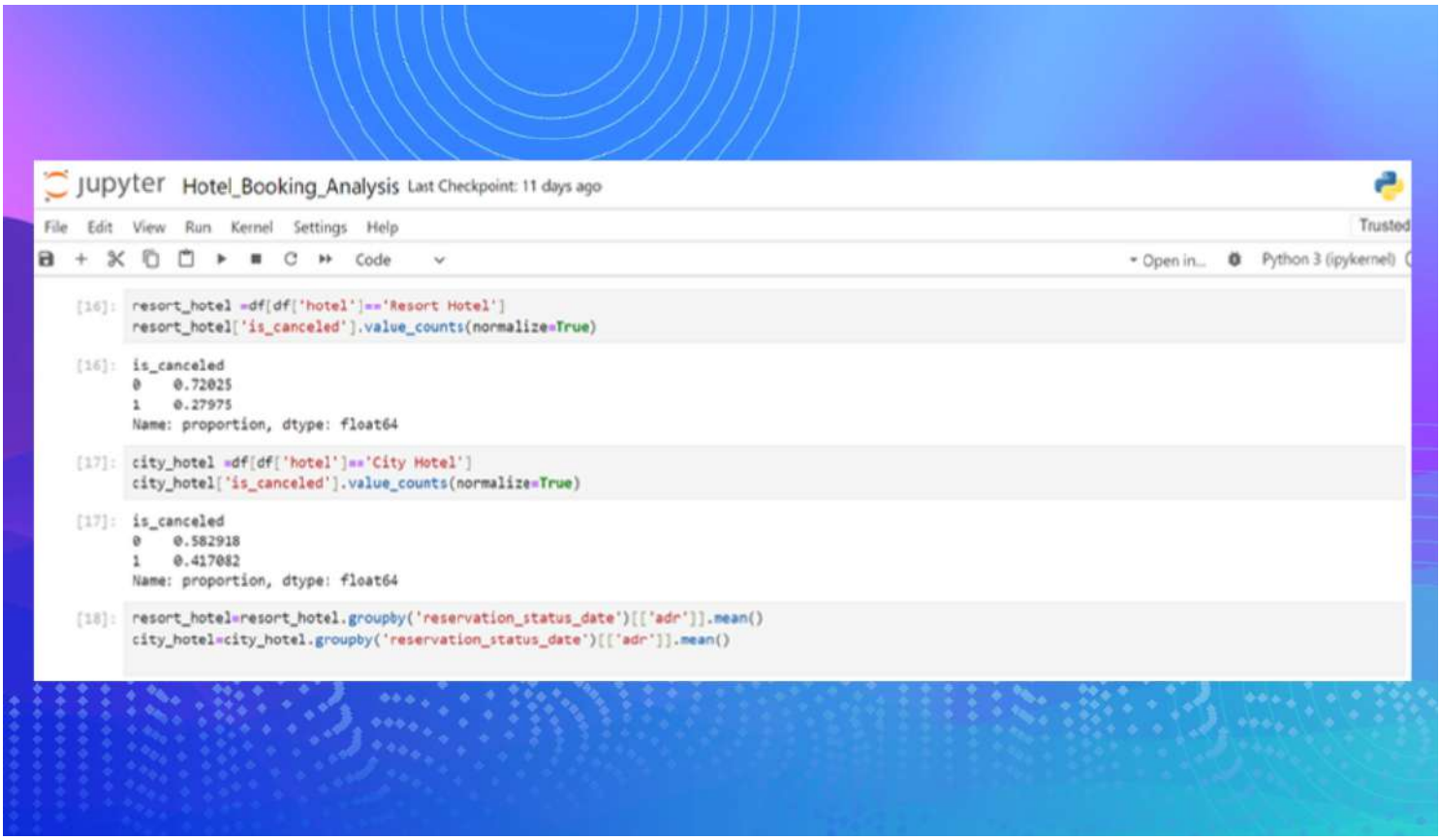
Trusted

Code

Open in... Python 3 (ipykernel)

[15]:
```
plt.figure(figsize=(3,3))
ax1=sns.countplot(x='hotel',hue='is_canceled',data=df,palette='tab10')
legend_labels = ax1.get_legend_handles_labels()
ax1.legend(bbox_to_anchor=(1,1))
plt.title('Reservation status in different hotels', size=15)
plt.xlabel('Hotel')
plt.ylabel('Number of Reservations')
plt.legend(['Not Canceled','Canceled'])
plt.show()
```




jupyter Hotel_Booking_Analysis Last Checkpoint: 11 days ago
File Edit View Run Kernel Settings Help
Trusted
Code
Open in...
Python 3 (ipykernel)
[16]:
resort_hotel =df[df['hotel']=='Resort Hotel']
resort_hotel['is_canceled'].value_counts(normalize=True)
[16]:
is_canceled
0 0.72025
1 0.27975
Name: proportion, dtype: float64
[17]:
city_hotel =df[df['hotel']=='City Hotel']
city_hotel['is_canceled'].value_counts(normalize=True)
[17]:
is_canceled
0 0.582918
1 0.417082
Name: proportion, dtype: float64
[18]:
resort_hotel=resort_hotel.groupby('reservation_status_date')[['adr']].mean()
city_hotel=city_hotel.groupby('reservation_status_date')[['adr']].mean()

Average Daily Rate in City Hotel and Royal Hotel

jupyter Hotel_Booking_Analysis Last Checkpoint: 11 days ago

File Edit View Run Kernel Settings Help

Trusted

Code

Open in... Python 3 (ipykernel)

[31]:
```
plt.figure(figsize=(20,5))
plt.title('Average Daily Rate in City Hotel and Royal Hotel', size=20)
plt.plot(resort_hotel.index, resort_hotel['adr'],label='Resort Hotel')
plt.plot(city_hotel.index, city_hotel['adr'],label='City Hotel')
plt.legend(fontsize=15)
plt.show()
```

Reservation status per month

jupyter Hotel_Booking_Analysis Last Checkpoint: 11 days ago

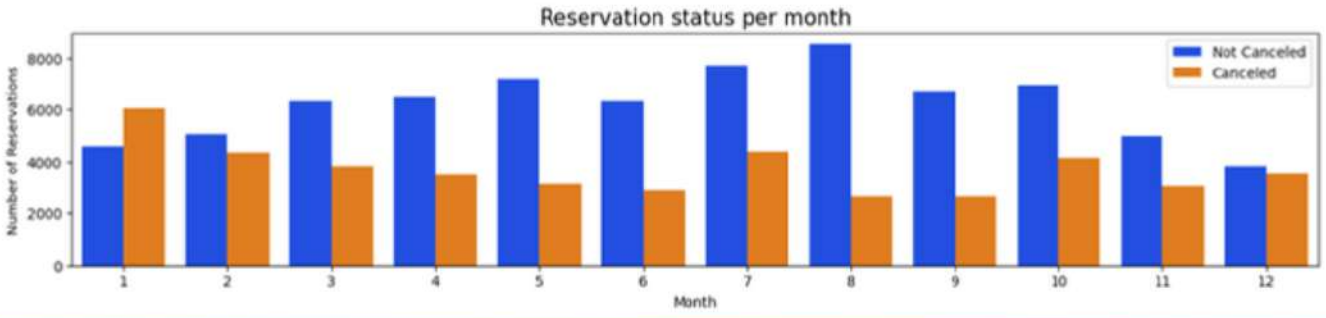
File Edit View Run Kernel Settings Help

Trusted

Code

Open in... Python 3 (ipykernel)

[31]:
```
df['month']=df['reservation_status_date'].dt.month
plt.figure(figsize=(16,3))
ax1=sns.countplot(x='month',hue='is_canceled',data=df,palette='bright')
legend_labels = ax1.get_legend_handles_labels()
ax1.legend(bbox_to_anchor=(1,1))
plt.title('Reservation status per month', size=15)
plt.xlabel('Month')
plt.ylabel('Number of Reservations')
plt.legend(['Not Canceled','Canceled'])
plt.show()
```



Average Daily Rate Per Month

jupyter Hotel_Booking_Analysis Last Checkpoint: 11 days ago

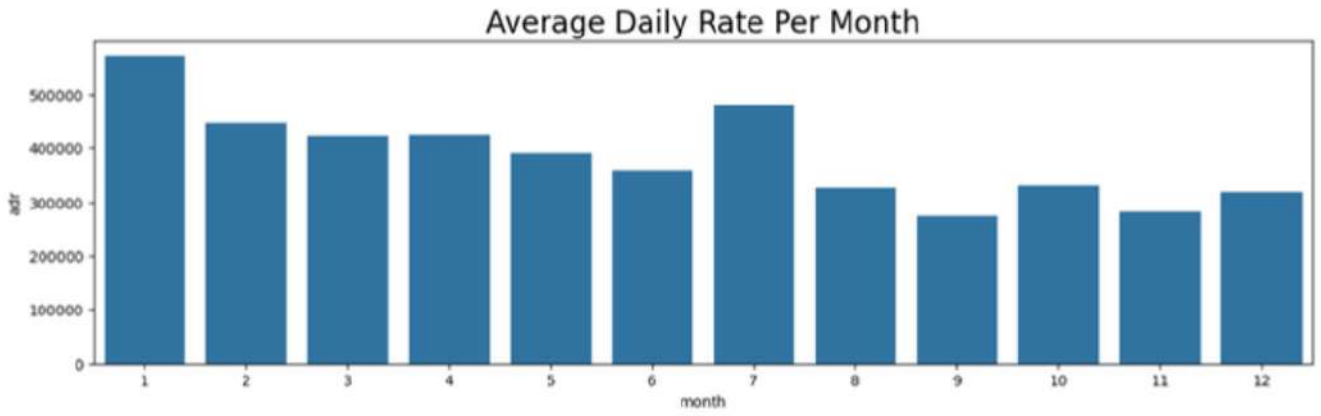
File Edit View Run Kernel Settings Help

Trusted

Code

Open in... Python 3 (ipykernel)

[34]:
```
plt.figure(figsize=(15,4))
plt.title('Average Daily Rate Per Month', fontsize=20)
sns.barplot(x='month', y='adr', data=df[df['is_canceled']==1].groupby('month')[['adr']].sum().reset_index())
plt.show()
```



Top 10 countries with reservations canceled

jupyter Hotel_Booking_Analysis Last Checkpoint: 11 days ago

File Edit View Run Kernel Settings Help

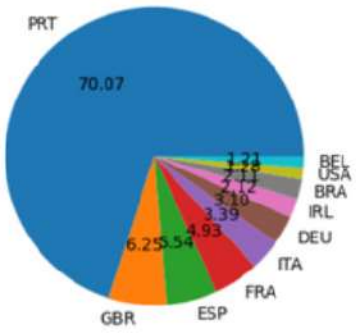
Trusted

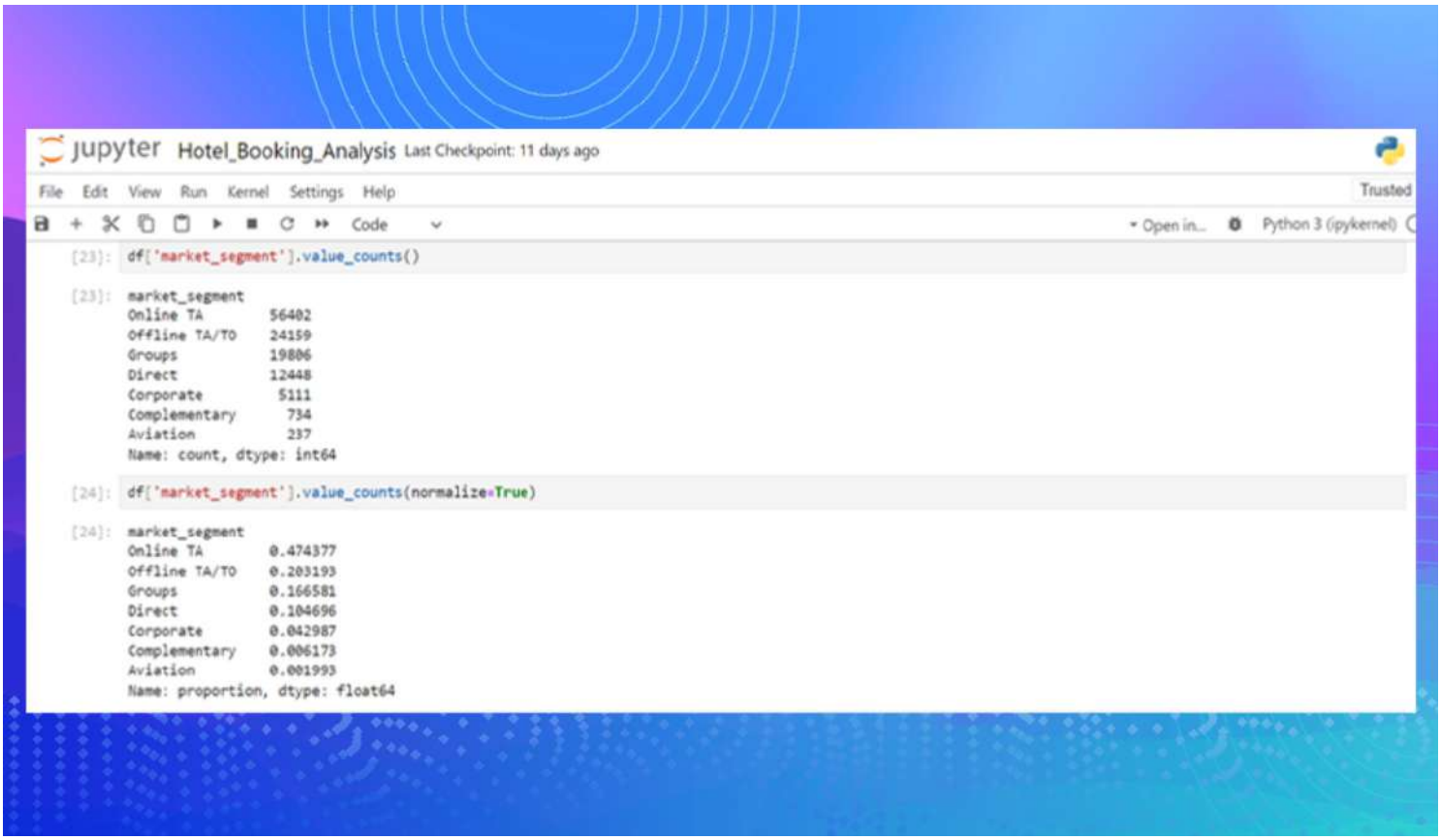
Code

Open in... Python 3 (ipykernel)

[40]:
```
cancelled_data = df[df['is_canceled'] == 1]
top_10_country = cancelled_data['country'].value_counts()[:10]
plt.figure(figsize=(15, 4))
plt.title('Top 10 countries with reservations canceled')
plt.pie(top_10_country, autopct='%.2f', labels=top_10_country.index)
plt.show()
```

Top 10 countries with reservations canceled




jupyter Hotel_Booking_Analysis Last Checkpoint: 11 days ago
File Edit View Run Kernel Settings Help
Trusted
Code
Open in...
Python 3 (ipykernel)
[23]: df['market_segment'].value_counts()
[23]: market_segment
Online TA 56402
Offline TA/TO 24159
Groups 19806
Direct 12448
Corporate 5111
Complementary 734
Aviation 237
Name: count, dtype: int64
[24]: df['market_segment'].value_counts(normalize=True)
[24]: market_segment
Online TA 0.474377
Offline TA/TO 0.203193
Groups 0.166581
Direct 0.104696
Corporate 0.042987
Complementary 0.006173
Aviation 0.001993
Name: proportion, dtype: float64

Average Daily Rate

jupyter Hotel_Booking_Analysis Last Checkpoint: 11 days ago

File Edit View Run Kernel Settings Help

Trusted

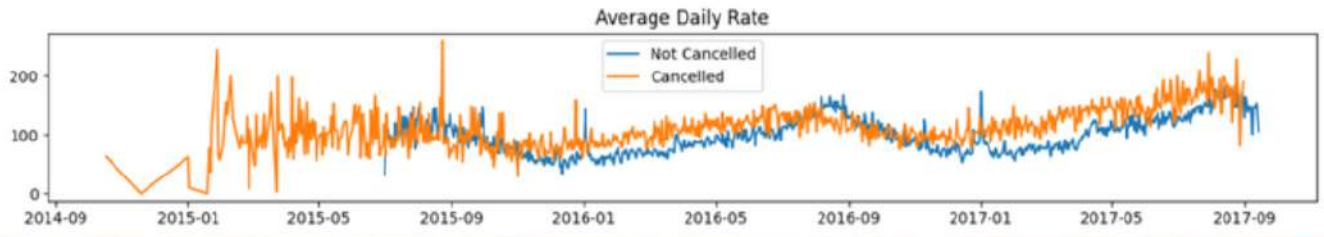
Code

Open in... Python 3 (ipykernel)

```
Name: proportion, dtype: float64
```

[43]:
```
cancelled_df_adr = cancelled_data.groupby('reservation_status_date')[['adr']].mean()
cancelled_df_adr.reset_index(inplace=True)
cancelled_df_adr.sort_values('reservation_status_date', inplace=True)
not_cancelled_data = df[df['is_canceled'] == 0]
not_cancelled_df_adr = not_cancelled_data.groupby('reservation_status_date')[['adr']].mean()
not_cancelled_df_adr.reset_index(inplace=True)
not_cancelled_df_adr.sort_values('reservation_status_date', inplace=True)
plt.figure(figsize=(15, 2))
plt.title('Average Daily Rate')
plt.plot(not_cancelled_df_adr['reservation_status_date'], not_cancelled_df_adr['adr'], label='Not Cancelled')
plt.plot(cancelled_df_adr['reservation_status_date'], cancelled_df_adr['adr'], label='Cancelled')
plt.legend()
```

[43]:
```
<matplotlib.legend.Legend at 0x1c9f7bb5550>
```



jupyter Hotel_Booking_Analysis Last Checkpoint: 11 days ago

File Edit View Run Kernel Settings Help Trusted

Code Open in... Python 3 (ipykernel)

[47]:
```
cancelled_df_adr = cancelled_df_adr[(cancelled_df_adr['reservation_status_date'] > '2016') & (cancelled_df_adr['reservation_status_date'] < '2017')]
not_cancelled_df_adr = not_cancelled_df_adr[(not_cancelled_df_adr['reservation_status_date'] > '2016') & (not_cancelled_df_adr['reservation_status_date']
plt.figure(figsize=(15, 3))
plt.title('Average Daily Rate')
plt.plot(not_cancelled_df_adr['reservation_status_date'], not_cancelled_df_adr['adr'], label='Not Cancelled')
plt.plot(cancelled_df_adr['reservation_status_date'], cancelled_df_adr['adr'], label='Cancelled')
plt.legend(fontsize=10)
plt.show()
```