Investigate Business Hotel using Data Visualization

Created by:
Andi Eka Nugraha
an.ekanugraha@gmail.com
linkedin.com/in/andi-eka-nugraha

Bachelor in Physics with major expertise in Instrumentation & Robotics and has attended a Datascience bootcamp for 4 months. Experienced in programming microcontrollers and machine learning to process data or images, as well as creating robotic systems that can support human work. Able to understand business, especially for data analysis, studying statistics and machine learning, as well as the ability to create regression models, classification, and clustering. Skills in identifying and analyzing patterns in data and presenting analytical results well.

Supported by:
Rakamin Academy
Career Acceleration School
www.rakamin.com

Overview

“Sangat penting bagi suatu perusahaan untuk selalu menganalisa performa bisnisnya. Pada kesempatan kali ini, kita akan lebih mendalami bisnis dalam bidang perhotelan. Fokus yang kita tuju adalah untuk mengetahui bagaimana perilaku pelanggan kita dalam melakukan pemesanan hotel, dan hubungannya terhadap tingkat pembatalan pemesanan hotel. Hasil dari insight yang kita temukan akan kita sajikan dalam bentuk data visualisasi agar lebih mudah dipahami dan bersifat lebih persuasif. ”

Data Preprocessing

```
Data columns (total 30 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  int64
 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_weekdays_nights        119390 non-null  int64
 9   meal                            119390 non-null  object
 10  city                            119390 non-null  object
 11  market_segment                  119390 non-null  object
 12  distribution_channel            119390 non-null  object
 13  is_repeated_guest               119390 non-null  int64
 14  previous_cancellations          119390 non-null  int64
 15  previous_bookings_not_canceled  119390 non-null  int64
 16  booking_changes                 119390 non-null  int64
 17  deposit_type                    119390 non-null  object
 18  agent                           119390 non-null  float64
 19  company                         119390 non-null  object
 20  days_in_waiting_list            119390 non-null  int64
 21  customer_type                   119390 non-null  object
 22  adr                             119390 non-null  float64
 23  required_car_parking_spaces     119390 non-null  int64
 24  total_of_special_requests       119390 non-null  int64
 25  reservation_status              119390 non-null  object
 26  total_customers                 119390 non-null  int64
 27  stay_duration                   119390 non-null  int64
 28  duration                        119390 non-null  object
 29  category_lead_time              119390 non-null  object
```

- Terdapat 29 kolom pada Dataset, dengan total baris sebanyak 119390.
- Terdiri dari beberapa jenis tipe data diantaranya string, integer, dan float.
- Terdapat beberapa nilai kosong pada beberapa kolom

Untuk selengkapnya, dapat melihat jupyter notebook disini

Data Preprocessing

Rakamin Academy

Data preprocessing adalah langkah-langkah yang dilakukan untuk membersihkan dan merapikan data sebelum dilakukan analisis lebih lanjut. Berikut adalah hasil dari data preprocessing yang telah disebutkan

- **Handling Missing Values “city”:** Jika terdapat data kosong pada kolom ini, langkahnya adalah mengisi data kosong tersebut dengan kota yang paling sering muncul dalam dataset. Hal ini dilakukan untuk mempertahankan jumlah data yang ada dan menghindari kehilangan informasi yang berharga.
- **Handling Missing Values "agent", "company", and "children":** Jika terdapat data kosong pada kolom-kolom ini, langkahnya adalah mengisinya dengan nilai 0. Ini dilakukan karena diasumsikan bahwa data yang kosong tersebut tidak melibatkan agen (agent), tidak atas nama perusahaan (company), atau tidak ada anak (children) saat menginap.
- **Handling Invalid Values:** Pada kolom "meal" terdapat nilai "undefined" dalam kolom ini, langkahnya adalah menggantinya dengan nilai "No Meal". Hal ini dilakukan untuk memberikan konsistensi dan memastikan bahwa tidak ada nilai yang ambigu atau tidak valid dalam dataset.
- **Handling Unnecessary Data:** Untuk menghitung total jumlah pelanggan, kita dapat menggabungkan jumlah orang dewasa (adult), bayi (babies), dan anak-anak (children) untuk mendapatkan total pelanggan dalam satu entitas. Dalam hasil preprocessing ini, kolom-kolom ini akan digantikan dengan satu kolom baru yang menunjukkan jumlah total pelanggan.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Monthly Hotel Booking Analysis Based on Hotel Type

```
result = df.groupby(['hotel','arrival_date_year', 'arrival_date_month'])['total_customers'].count().reset_index()
result['arrival_date_year'] = result['arrival_date_year'].replace(2017,2019)
result = result.groupby(['hotel','arrival_date_year', 'arrival_date_month'])['total_customers'].sum().reset_index()
result
```

- Menggunakan metode **groupby** pada Dataset untuk mengelompokkan data berdasarkan kolom **'hotel'**, **'arrival_date_year'**, dan **'arrival_date_month'**. Memilih kolom **'total_customers'** menggunakan indexing **['total_customers']**. Kemudian menghitung jumlah entri pada setiap kelompok menggunakan metode **count()**. Menggunakan metode **reset_index()** untuk mereset indeks dan mengubah hasil pengelompokan menjadi DataFrame baru.

- Mengganti nilai **2017** dalam kolom **'arrival_date_year'** dengan nilai **2019** menggunakan metode **replace()**, hal tersebut dilakukan karena ada kesalahan input data bagian tahun sehingga dilakukan penyesuaian data. Kemudian mengupdate kolom **'arrival_date_year'** dalam Dataset **result** dengan nilai yang telah diganti.

- Kembali menggunakan metode groupby pada Dataset result setelah perubahan pada kolom **'arrival_date_year',** lalu mengelompokkan data berdasarkan kolom **'hotel'**, **'arrival_date_year'**, dan **'arrival_date_month'**. Memilih kolom **'total_customers'** menggunakan **indexing ['total_customers']** dan menghitung total nilai **'total_customers'** pada setiap kelompok menggunakan metode **sum()**. Menggunakan metode **reset_index()** untuk mereset indeks dan mengubah hasil pengelompokan menjadi Dataset baru.

Untuk selengkapnya, dapat melihat jupyter notebook disini

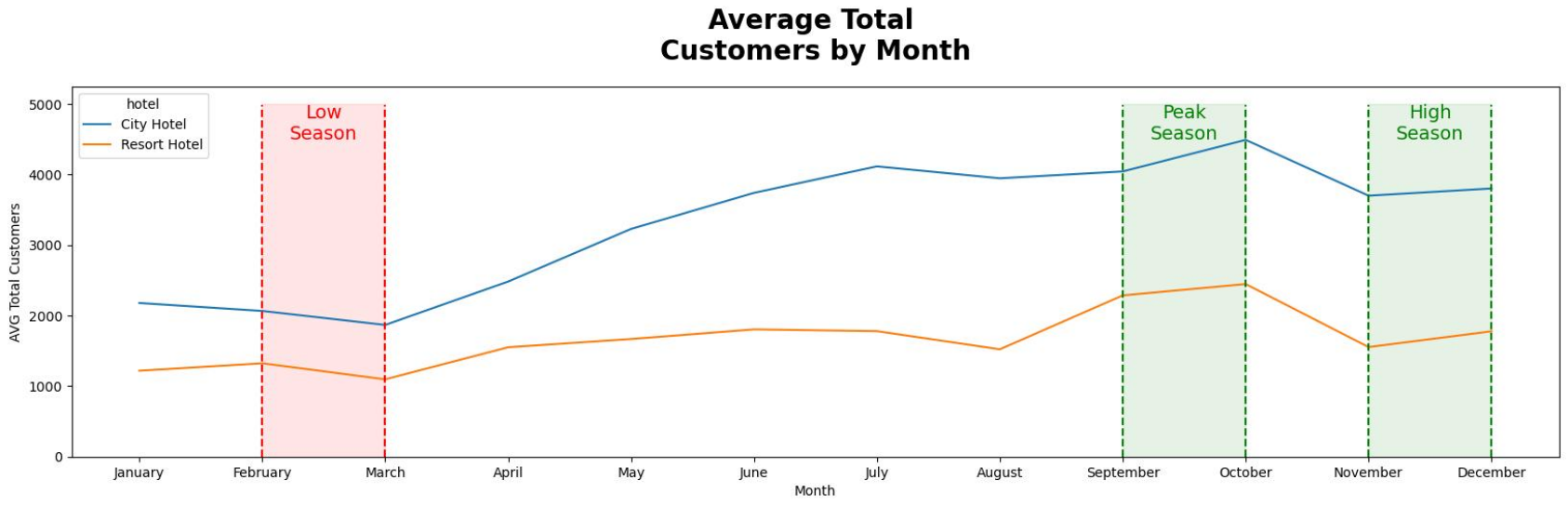
Monthly Hotel Booking Analysis Based on Hotel Type

```
x = result.groupby(['hotel', 'arrival_date_month'])['total_customers'].mean().reset_index()
```

- Menggunakan metode **groupby** pada Dataset **result** untuk mengelompokkan data berdasarkan kolom **'hotel'** dan **'arrival_date_month'**. Memilih kolom **'total_customers'** menggunakan **indexing ['total_customers']**.
- Menghitung rata-rata nilai **'total_customers'** pada setiap kelompok menggunakan metode **mean()** untuk mengetahui berapa rata rata pengunjung pada bulan tertentu setiap tahunnya. Menggunakan metode **reset_index()** untuk mereset indeks dan mengubah hasil pengelompokan menjadi Dataset baru.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Monthly Hotel Booking Analysis Based on Hotel Type



- **Peak Season:** Untuk City Hotel, nilai tertinggi terjadi pada bulan Oktober dengan rata-rata 4491.5 pelanggan. Untuk Resort Hotel, nilai tertinggi terjadi pada bulan Oktober dengan rata-rata 2447.0 pelanggan. Oleh karena itu, Oktober adalah bulan dengan jumlah pelanggan terbanyak untuk kedua jenis hotel.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Monthly Hotel Booking Analysis Based on Hotel Type

- **Low Season:** Untuk City Hotel, rata-rata jumlah pelanggan pada bulan Februari adalah 2066, sedangkan pada bulan Maret adalah 1868. Untuk Resort Hotel, rata-rata jumlah pelanggan pada bulan Februari adalah 1324, sedangkan pada bulan Maret adalah 1096,5. Pada kedua jenis hotel, **terjadi penurunan** jumlah pelanggan pada bulan Februari dan Maret dibandingkan dengan bulan-bulan lainnya. Dapat disimpulkan bahwa bulan **Februari dan Maret adalah periode low season** atau periode dengan jumlah pelanggan yang relatif lebih rendah.

- **High Season:** Untuk City Hotel, rata-rata jumlah pelanggan pada bulan November adalah 3700, sedangkan pada bulan Desember adalah 3802,5. Untuk Resort Hotel, rata-rata jumlah pelanggan pada bulan November adalah 1554, sedangkan pada bulan Desember adalah 1777.5. Pada kedua jenis hotel, terjadi **peningkatan jumlah pelanggan** pada bulan November dan Desember dibandingkan dengan bulan-bulan sebelumnya. Dapat disimpulkan bahwa bulan **November dan Desember adalah periode high season** atau periode dengan jumlah pelanggan yang relatif lebih tinggi.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Impact Analysis of Stay Duration on Hotel Bookings Cancellation Rates

Rakamin Academy

```
df['stay_duration'] = df['stays_in_weekdays_nights'] + df['stays_in_weekend_nights']
```

- Membuat kolom **‘stay_duration’,** Kolom ini berisi jumlah total durasi menginap (dalam hari) yang dihitung berdasarkan kolom **‘stays_in_weekdays_nights’** (jumlah menginap pada hari kerja) dan **‘stays_in_weekend_nights’** (jumlah menginap pada akhir pekan).

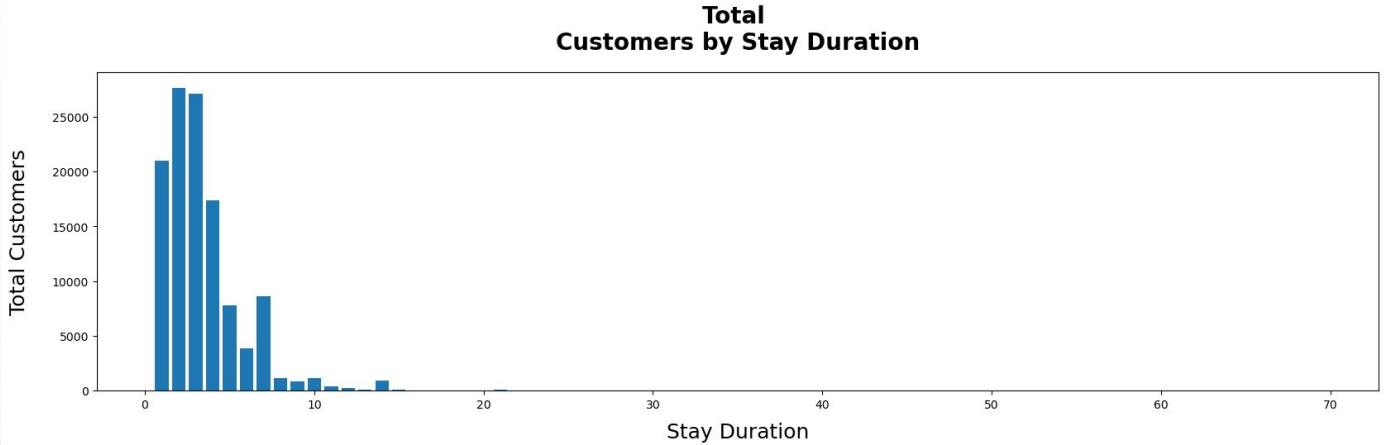
```
x = df.groupby(['stay_duration'])['total_customers'].count().reset_index()
x = x[x['stay_duration'] != 0]
```

  - Menggunakan metode **groupby** pada Dataset untuk mengelompokkan data berdasarkan kolom **'stay_duration'**. Memilih kolom **'total_customers'** menggunakan **indexing ['total_customers']**. Menghitung jumlah entri pada setiap kelompok menggunakan metode **count()**. Menggunakan metode **reset_index()** untuk mereset indeks dan mengubah hasil pengelompokan menjadi Dataset baru. Maka akan terbentuk Dataset kolom pengelompokan jumlah customer berdasarkan **‘stay_duration’.**

  - Melakukan filtering pada Dataset dengan menggunakan kondisi **x['stay_duration'] != 0**. Menghilangkan baris yang memiliki nilai **'stay_duration'** sama dengan **0** dari Dataset. Hal ini bertujuan untuk menghilangkan customer yang tidak menginap di hotel.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Impact Analysis of Stay Duration on Hotel Bookings Cancellation Rates



- Durasi menginap paling umum adalah 2 dan 3 hari, dengan masing-masing memiliki jumlah tamu sebanyak 27.643 dan 27.076. Durasi menginap yang lebih lama cenderung memiliki jumlah tamu yang lebih sedikit, seperti durasi 4 hari (17.383 tamu), 5 hari (7.784 tamu), dan seterusnya. Terdapat sedikit tamu yang menginap dalam durasi yang sangat lama, seperti 69 hari (1 tamu) dan 60 hari (1 tamu).

- Sebagian besar tamu cenderung menginap dalam durasi yang relatif singkat dengan puncak terjadi pada durasi 2 dan 3 hari, terdapat penurunan jumlah tamu secara signifikan seiring dengan bertambahnya durasi menginap. Durasi menginap yang lebih lama mungkin mencerminkan pengunjung yang memiliki alasan khusus atau keperluan bisnis, seperti konferensi, pertemuan, atau kegiatan yang membutuhkan waktu lebih lama.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Impact Analysis of Stay Duration on Hotel Bookings Cancellation Rates

Rakamin Academy

```
df['duration'] = df['stay_duration'].apply(lambda value: 'A few days' if value < 7 else ('A few weeks' if value <= 31 else ('Several months')))
df['duration'].value_counts()
```

- Menggunakan metode apply pada kolom **"stay_duration"** untuk menerapkan suatu fungsi pada setiap nilai dalam kolom tersebut. Fungsi yang diterapkan menggunakan ekspresi lambda untuk menentukan kategori durasi berdasarkan nilai **"stay_duration"**. Jika nilai **"stay_duration"** kurang dari 7, maka kategori durasi diatur sebagai **"A few days"**. Jika nilai "stay_duration" antara 7 dan 31 (inklusif), maka kategori durasi diatur sebagai "A few weeks". Jika nilai **"stay_duration"** lebih dari 31, maka kategori durasi diatur sebagai **"Several months"**. Segmentasi ini dilakukan untuk mempermudah dalam melakukan anlisis dan menarik insight dari Dataset.

```
x = df[df['is_canceled'] == 0].groupby(['hotel','duration'])['is_canceled'].count().reset_index()
y = df[df['is_canceled'] == 1].groupby(['hotel','duration'])['is_canceled'].count().reset_index()
x['canceled'] = y['is_canceled']
x.columns = ['hotel', 'duration','not canceled','canceled']
x['rasio'] = x['canceled'] / (x['not canceled']+x['canceled']) * 100
x
```

- Pada baris program pertama, Melakukan filtering pada DataFrame df dengan kondisi **df['is_canceled'] == 0**, yang mengambil hanya data yang tidak dibatalkan (tidak dibatalkan = 0). Menggunakan metode **groupby** pada Dataset hasil filtering untuk mengelompokkan data berdasarkan kolom **'hotel'** dan **'duration'**.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Impact Analysis of Stay Duration on Hotel Bookings Cancellation Rates

- Kemudian menghitung jumlah entri (yang tidak dibatalkan) pada setiap kelompok menggunakan metode **count(), m**enggunakan metode **reset_index()** untuk mereset indeks dan mengubah hasil pengelompokan menjadi Dataset baru dan menyimpan hasil pengelompokan dan penghitungan tersebut dalam variabel x.

- Pada program baris ke dua, melakukan filtering pada DataFrame df dengan kondisi **df['is_canceled'] == 1**, yang mengambil hanya data yang dibatalkan (dibatalkan = 1). Menggunakan metode **groupby** pada Dataset hasil filtering untuk mengelompokkan data berdasarkan kolom **'hotel'** dan **'duration'**. Memilih kolom **'is_canceled'** menggunakan **indexing ['is_canceled']** lalu Menghitung jumlah entri (yang dibatalkan) pada setiap kelompok menggunakan metode **count()**. Menggunakan metode **reset_index()** untuk mereset indeks dan mengubah hasil pengelompokan menjadi Dataset baru dan menyimpan hasil pengelompokan dan penghitungan tersebut dalam variabel y.

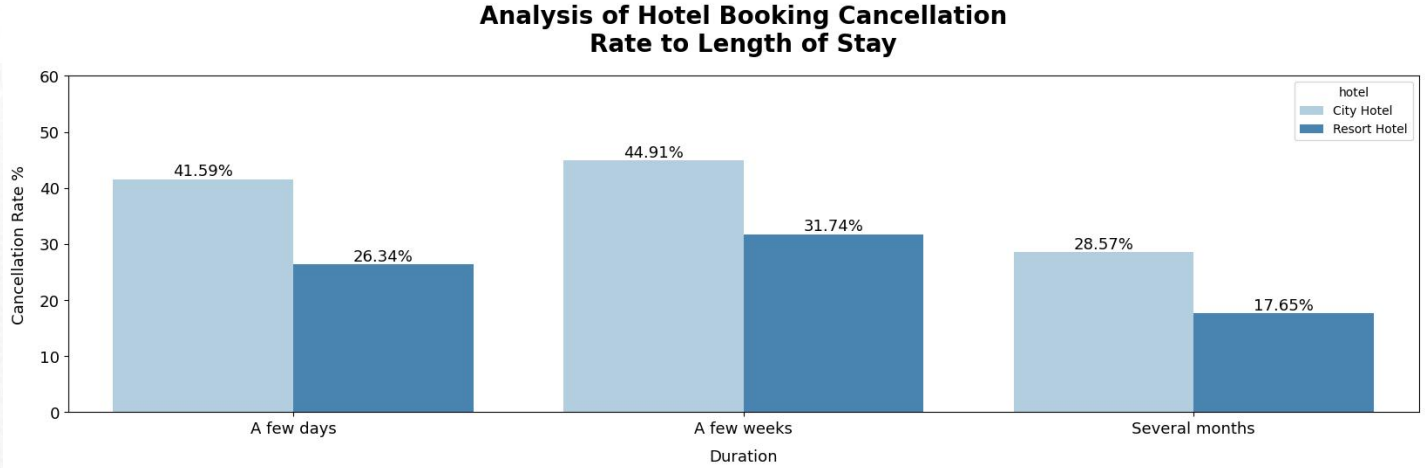
- Pada program baris ke tiga, menambahkan kolom **'canceled'** pada Dataset x yang berisi nilai dari kolom **'is_canceled'** pada Dataset y lalu menggabungkan informasi jumlah pembatalan dari Dataset y ke Dataset x berdasarkan **hotel** dan **durasi**.

- Pada baris ke empat, mengubah nama kolom dalam Dataset x menjadi **'hotel'**, **'duration'**, **'not canceled'**, dan **'canceled'**. **'not canceled'** merujuk pada jumlah entri yang tidak dibatalkan (hasil dari baris1).**'canceled'** merujuk pada jumlah entri yang dibatalkan (hasil dari baris 3).

- Pada baris ke lima, menambahkan kolom **'rasio'** pada Dataset x yang menghitung **rasio pembatalan** dalam persentase. Rasio pembatalan dihitung dengan membagi jumlah pembatalan (kolom 'canceled') dengan jumlah total (jumlah pembatalan + jumlah yang tidak dibatalkan) dan dikalikan dengan 100%.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Impact Analysis of Stay Duration on Hotel Bookings Cancellation Rates



- Durasi menginap yang lebih lama (kategori **"A few weeks"** dan **"Several months"**) cenderung memiliki rasio pembatalan yang **lebih tinggi** daripada durasi menginap yang **lebih singkat** (kategori **"A few days"**). Pembatalan lebih umum terjadi pada durasi menginap yang lebih panjang, yang mungkin dikaitkan dengan perubahan rencana atau faktor lain yang mempengaruhi keputusan tamu untuk membatalkan. Hotel City Hotel memiliki r**asio pembatalan yang lebih tinggi** secara keseluruhan dibandingkan dengan **Resort Hotel**, terlepas dari durasi menginapnya. Manajemen hotel dapat menggunakan informasi ini untuk memahami pola pembatalan dan mengembangkan strategi untuk mengurangi jumlah pembatalan, seperti meningkatkan kepuasan tamu atau menawarkan fleksibilitas dalam kebijakan pembatalan.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Impact Analysis of Lead Time on Hotel Bookings Cancellation Rate

```
df['category_lead_time'] = df['lead_time'].apply(lambda value: 'A day' if value == 0 else ('A few days' if value <= 7 else
                                                                ('A few weeks' if value <= 31 else ('A few months'))))
```

- Menggunakan metode apply pada kolom **"lead_time"** untuk menerapkan suatu fungsi pada setiap nilai dalam kolom tersebut. Fungsi yang diterapkan menggunakan **ekspresi lambda** untuk menentukan kategori waktu pemesanan berdasarkan nilai **"lead_time"**. Jika nilai **"lead_time"** adalah **0**, maka kategori waktu pemesanan diatur sebagai **"A day"**. Jika nilai **"lead_time"** kurang dari atau sama dengan **7**, maka kategori waktu pemesanan diatur sebagai **"A few days"**. Jika nilai **"lead_time"** kurang dari atau sama dengan **31**, maka kategori waktu pemesanan diatur sebagai **"A few weeks"**. Jika nilai **"lead_time"** lebih dari **31**, maka kategori waktu pemesanan diatur sebagai **"A few months"**.
- Dataset df akan memiliki kolom baru **"category_lead_time"** yang membagi waktu pemesanan menjadi beberapa segementasi berdasarkan nilai **"lead_time"**. Segmentasi ini dapat membantu dalam analisis dan pemahaman pola pemesanan berdasarkan durasi waktu antara pemesanan dan tanggal kedatangan.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Impact Analysis of Lead Time on Hotel Bookings Cancellation Rate

```
x = df[df['is_canceled'] == 0].groupby(['hotel','category_lead_time'])['is_canceled'].count().reset_index()
y = df[df['is_canceled'] == 1].groupby(['hotel','category_lead_time'])['is_canceled'].count().reset_index()
x['canceled'] = y['is_canceled']
x = x.rename(columns={'is_canceled': 'not_canceled'})
x['Cancellation_rate'] = x['canceled']/(x['not_canceled'] +x['canceled']) * 100
```

- Pada baris pertama, melakukan filtering pada Dataset df dengan kondisi **df['is_canceled'] == 0**, yang mengambil hanya data yang tidak dibatalkan (tidak dibatalkan = 0). Menggunakan metode **groupby** pada Dataset hasil filtering untuk mengelompokkan data berdasarkan kolom **'hotel'** dan **'category_lead_time'**. Memilih kolom **'is_canceled'** menggunakan **indexing ['is_canceled']**. Menghitung jumlah entri (yang tidak dibatalkan) pada setiap kelompok menggunakan metode **count()**. Menggunakan metode **reset_index()** untuk mereset indeks dan mengubah hasil pengelompokan menjadi Dataset baru. Menyimpan hasil pengelompokan dan penghitungan tersebut dalam variabel x.

- Pada baris ke dua, melakukan filtering pada Dataset df dengan kondisi **df['is_canceled'] == 1**, yang mengambil hanya data yang dibatalkan (dibatalkan = 1).Menggunakan metode **groupby** pada Dataset hasil filtering untuk mengelompokkan data berdasarkan kolom **'hotel'** dan **'category_lead_time'**. kemudian memilih **kolom 'is_canceled'** menggunakan **indexing ['is_canceled']**. Menghitung jumlah entri (yang dibatalkan) pada setiap kelompok menggunakan metode **count()**. Menggunakan metode **reset_index()** untuk mereset indeks dan mengubah hasil pengelompokan menjadi Dataset baru. Menyimpan hasil pengelompokan dan penghitungan tersebut dalam variabel y.

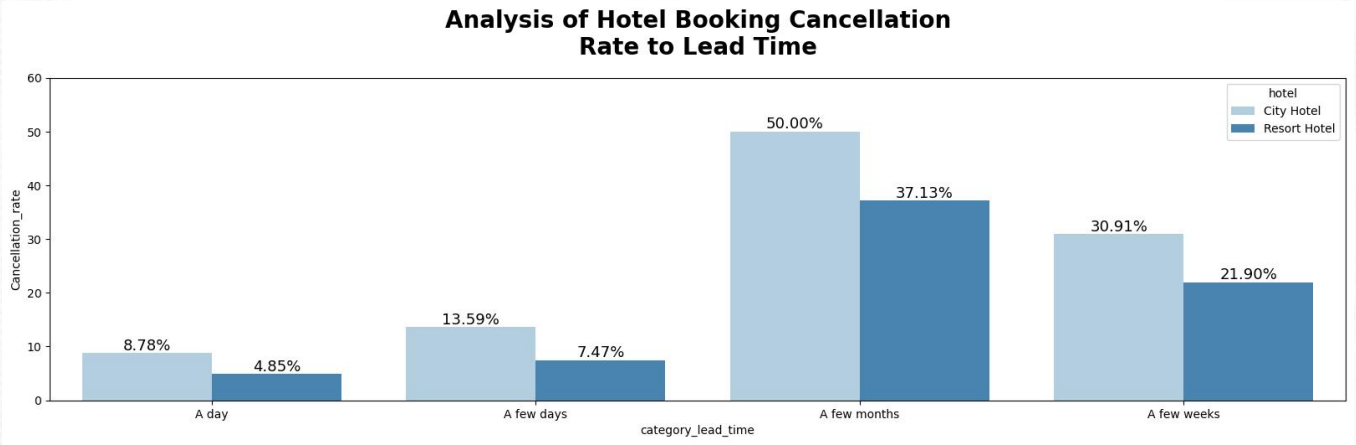
Untuk selengkapnya, dapat melihat jupyter notebook disini

Impact Analysis of Lead Time on Hotel Bookings Cancellation Rate

- Pada baris ketiga, menambahkan kolom **'canceled'** pada Dataset x yang berisi nilai dari kolom **'is_canceled'** pada Dataset y. Menggabungkan informasi jumlah pembatalan dari Dataset y ke Dataset x berdasarkan hotel dan kategori waktu pemesanan.
- Pada baris ke empat, mengubah nama kolom **'is_canceled'** dalam Dataset x menjadi **'not_canceled'**. Ini dilakukan untuk memberikan pemahaman yang lebih jelas bahwa kolom tersebut berisi jumlah tidak dibatalkan.
- Pada baris ke lima, menambahkan kolom **'Cancellation_rate'** pada Dataset x yang menghitung tingkat pembatalan dalam persentase. Rasio tingkat pembatalan dihitung dengan **membagi jumlah pembatalan (kolom 'canceled') dengan jumlah total (jumlah pembatalan + jumlah yang tidak dibatalkan)** dan dikalikan dengan **100%.**

Untuk selengkapnya, dapat melihat jupyter notebook disini

Impact Analysis of Lead Time on Hotel Bookings Cancellation Rate



- Hotel dengan kategori waktu pemesanan yang lebih lama (beberapa bulan) memiliki tingkat pembatalan yang **lebih tinggi**, sedangkan hotel dengan kategori waktu pemesanan yang **lebih pendek** (beberapa hari atau beberapa minggu) memiliki tingkat pembatalan yang lebih rendah.

- Kategori waktu pemesanan **"A few months"** adalah yang paling **sering mengalami pembatalan** di kedua jenis hotel. Hal ini mungkin berkaitan dengan alasan pembatalan yang terkait dengan perubahan jadwal liburan atau rencana perjalanan dalam jangka waktu yang lebih panjang.

- Kategori waktu pemesanan **"A day"** adalah yang paling jarang mengalami pembatalan di kedua jenis hotel. Ini mungkin disebabkan oleh keputusan yang diambil dengan cepat atau reservasi yang dibuat untuk situasi mendesak atau keperluan mendesak.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Impact Analysis of Lead Time on Hotel Bookings Cancellation Rate

- Resort Hotel cenderung memiliki tingkat **pembatalan yang lebih rendah** dibandingkan dengan City Hotel. Hal ini mungkin berkaitan dengan karakteristik dan kebijakan pembatalan yang berbeda antara kedua jenis hotel tersebut dan jumlah pelanggan yang memesan lebih sedikit dibandingkan dengan City Hotel.

Untuk selengkapnya, dapat melihat jupyter notebook disini