**TUGAS AKHIR DATA SCIENCE PADA DOMAIN SPESIFIK**

**STUDI KASUS PADA DATASET HOTEL**

****

**KELOMPOK: 01**

**DISUSUN OLEH:**

**Samuel Edward Winoto / 6182101040**

**Jenson Mark Lowell / 6182001001**

**Steffi Widjaya / 6182101054**

**KELAS: A**

**Dosen Kelas: Lydia Mutiara Dewi, MA**

**UNIVERSITAS KATOLIK PARAHYANGAN**

**PROGRAM STUDI INFORMATIKA**

**FAKULTAS TEKNOLOGI INFORMASI DAN SAINS**

**BANDUNG**

**2024**

1. **LATAR BELAKANG**

**Definisi Pariwisata**

Pariwisata adalah kegiatan perjalanan yang dilakukan oleh individu atau kelompok ke suatu tempat di luar lingkungan sehari-hari untuk tujuan rekreasi, bisnis, atau keperluan lainnya dalam jangka waktu sementara. Aktivitas ini mencakup perjalanan domestik maupun internasional, dan sering kali melibatkan penginapan, makanan, serta hiburan selama masa tinggal di destinasi yang dikunjungi.

**Peran Pariwisata Secara Umum**

Pariwisata memiliki peran yang sangat penting dalam perekonomian banyak negara. Secara umum, peran pariwisata mencakup:

1. **Sumber Pendapatan:** Pariwisata menjadi sumber pendapatan utama bagi banyak negara dan daerah. Pengeluaran wisatawan untuk akomodasi, makanan, transportasi, dan aktivitas rekreasi berkontribusi signifikan terhadap pendapatan nasional dan lokal.
2. **Penciptaan Lapangan Kerja:** Industri pariwisata menyediakan berbagai jenis pekerjaan, mulai dari sektor perhotelan, restoran, transportasi, hingga layanan pemandu wisata. Hal ini membantu mengurangi angka pengangguran dan meningkatkan kesejahteraan masyarakat.
3. **Pengembangan Infrastruktur:** Permintaan wisatawan sering kali mendorong pemerintah dan pihak swasta untuk mengembangkan infrastruktur seperti bandara, jalan raya, dan fasilitas umum lainnya, yang juga bermanfaat bagi penduduk setempat.
4. **Pertukaran Budaya:** Pariwisata membuka kesempatan bagi pertukaran budaya antara wisatawan dan penduduk lokal, yang dapat meningkatkan pemahaman dan toleransi antarbudaya.

**Karakteristik Pariwisata Secara Umum**

Pariwisata memiliki beberapa karakteristik umum, yaitu:

1. **Temporality:** Aktivitas pariwisata bersifat sementara dan tidak menetap. Wisatawan biasanya hanya tinggal di destinasi untuk jangka waktu tertentu sebelum kembali ke tempat asal mereka.
2. **Heterogen dan Kompleks:** Pariwisata melibatkan berbagai jenis aktivitas seperti wisata alam, wisata budaya, wisata kuliner, dan lainnya. Keanekaragaman ini mencerminkan minat dan preferensi wisatawan yang berbeda-beda.
3. **Perishable :** Pariwisata merupakan jasa yang produknya tidak dapat disimpan untuk dikonsumsi di kemudian hari.
4. **Intangible:** Pengalaman pariwisata tidak berwujud dan subjektif, bergantung pada persepsi dan kepuasan individu terhadap layanan dan fasilitas yang diterima.

**Definisi Hotel**

**Hotel** adalahusaha penyediaan akomodasi berupa kamar-kamar di dalam suatu bangunan, yang dapat dilengkapi dengan jasa pelayanan makan dan minum, kegiatan hiburan dan/atau fasilitas lainnya secara harian dengan tujuan memperoleh keuntungan

**Klasifikasi Hotel**

Hotel dapat diklasifikasikan berdasarkan beberapa kriteria, yaitu:

1. **Penentuan standar hotel menurut ukuran (size) hotel**
   1. **Hotel besar** adalah hotel yang memiliki 300 kamar lebih
   2. **Hotel menengah** adalah hotel yang memiliki 100-299 kamar
   3. **Hotel kecil** adalah hotel yang memiliki 25-99 kamar
2. **Penentuan hotel menurut operasinya**
   1. **Around the year operation hotel** adalah hotel yang beroperasi sepanjang tahun.
   2. **Seasonal hotel** adalah hotel yang beroperasi pada musim-musim tertentu
3. **Penentuan hotel menurut lokasinya**
   1. **City Hotel** adalah hotel yang berlokasi di daerah perkotaan
   2. **Residential Hotel** adalah hotel yang berlokasi di pinggiran yang cukup jauh dari keramaian
   3. **Resort Hotel** adalah hotel yang berlokasi di daerah wisata seperti pegunungan (mountain hotel) atau di tepi pantai (beach hotel)
   4. **Motel** adalah hotel yang berlokasi di pinggir atau di sepanjang jalan raya yang menghubungkan satu kota dengan kota lainnya

Berikut merupakan Tabel Klasifikasi Standard Hotel Bintang :

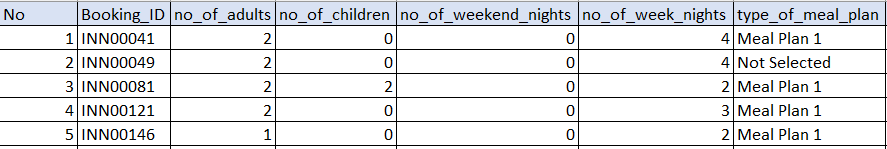
| **Klasifikasi**  **Hotel Bintang** | **Persyaratan** | |
| --- | --- | --- |
| **Jumlah Unit Kamar** | **Luasan Standar Minimum**  **Kamar** |
| **Bintang 1** | Unit standard minimum 15 kamar  Kamar mandi di dalam | Luas unit standard minimum 20m2 |
| **Bintang 2** | Unit standard minimum 20 kamar  Unit suite minimum 2 kamar  Kamar mandi di dalam | Luas unit standard minimum 22m2  Luas unit suite minimum 44m2 |
| **Bintang 3** | Unit standard minimum 30 kamar  Unit suite minimum 2 kamar  Kamar mandi di dalam | Luas unit standard minimum 24m2  Luas unit suite minimum 48m2 |
| **Bintang 4** | Unit standard minimum 50 kamar  Unit suite minimum 3 kamar  Kamar mandi di dalam | Luas unit standard minimum 24m2  Luas unit suite minimum 48m2 |
| **Bintang 5** | Unit standard minimum 100 kamar  Unit suite minimum 4 kamar  Kamar mandi di dalam | Luas standard minimum 26m2  Luas minimum suite 52m2 |

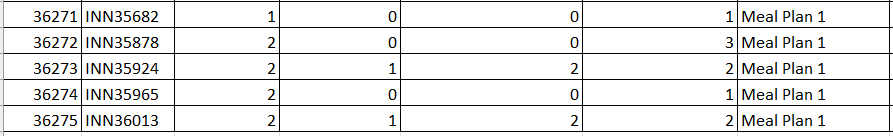
1. **DATASET Hotel Reservation Dataset**

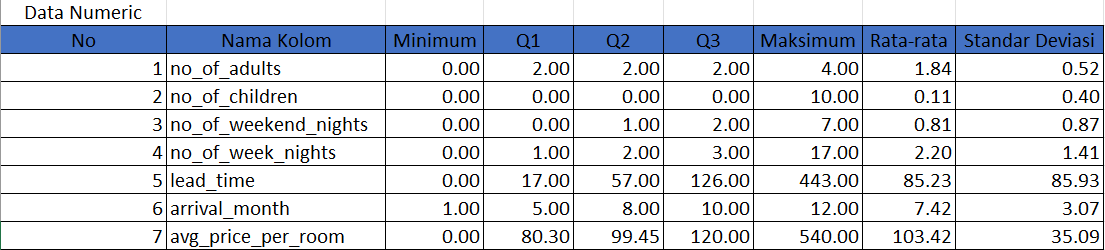
Eksplorasi umum dataset (lihat file tugas 1 yang sudah diperbaiki):

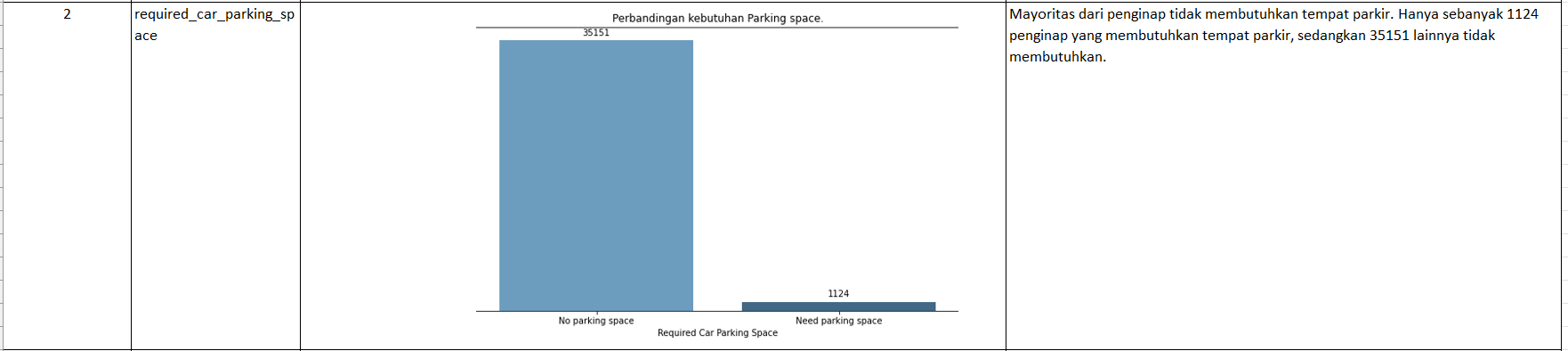
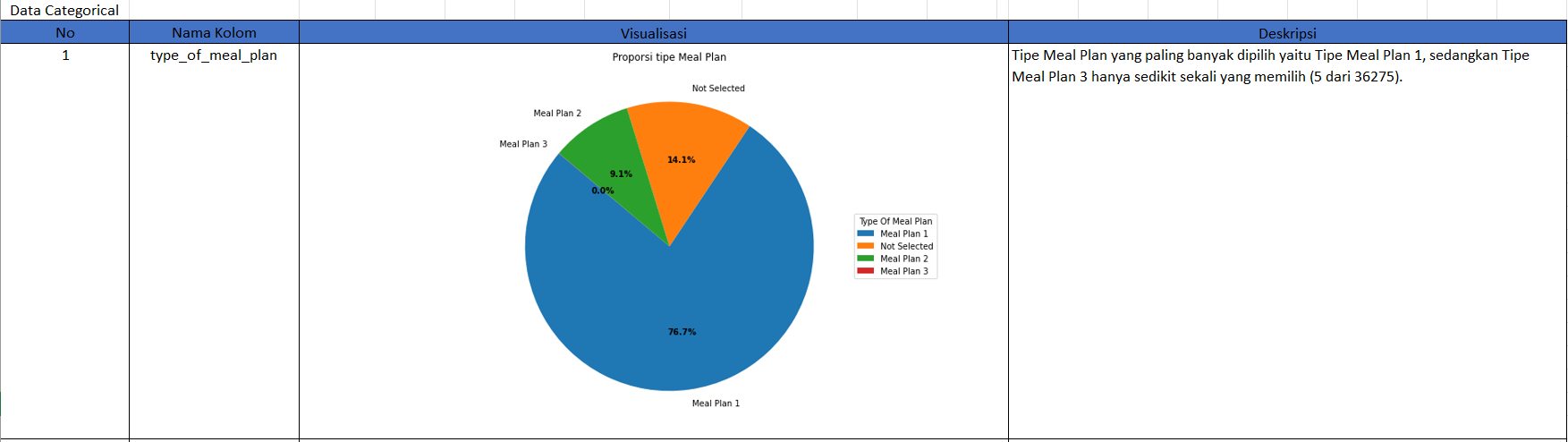
* 1. Narasi singkat mengenai dataset   
     Saluran reservasi hotel online telah mengubah cara pemesanan dilakukan dan mempengaruhi perilaku pelanggan. Kini, pelanggan dapat memesan dan membatalkan reservasi dengan mudah, yang mengakibatkan peningkatan pembatalan dan ketidakhadiran (no-show). Pembatalan biasanya terjadi karena perubahan rencana atau konflik jadwal, dan sering kali didorong oleh kebijakan pembatalan gratis atau berbiaya rendah. Sehingga untuk mengatasi masalah ini, hotel mulai menerapkan kebijakan pembatalan yang lebih ketat dan meningkatkan program loyalitas untuk memastikan reservasi yang lebih pasti dan juga perlu untuk memprediksi pelanggan mana saja yang akan membatalkan atau tidak membatalkan reservasi
  2. Link sumber dataset: [Hotel Reservation Dataset](https://www.kaggle.com/datasets/ahsan81/hotel-reservations-classification-dataset?authuser=0)
  3. Jumlah kolom: 20 Kolom
  4. Jumlah baris: 36275 baris
  5. Identifikasi null values dan Outliers

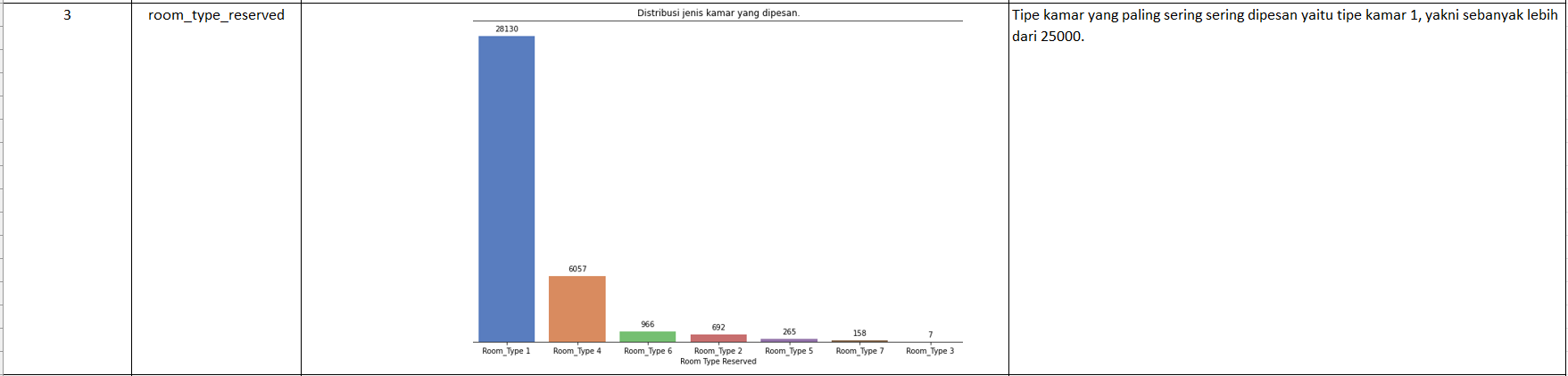
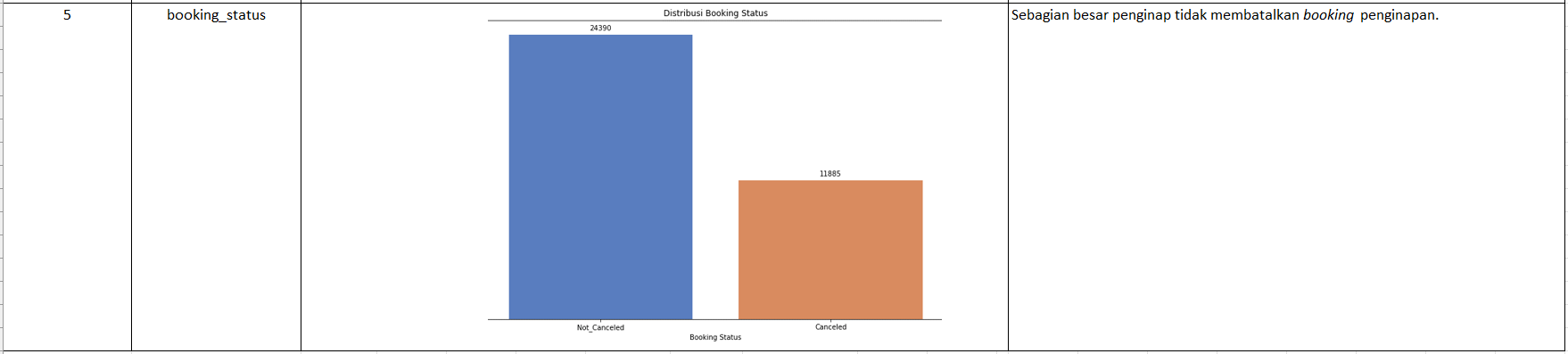
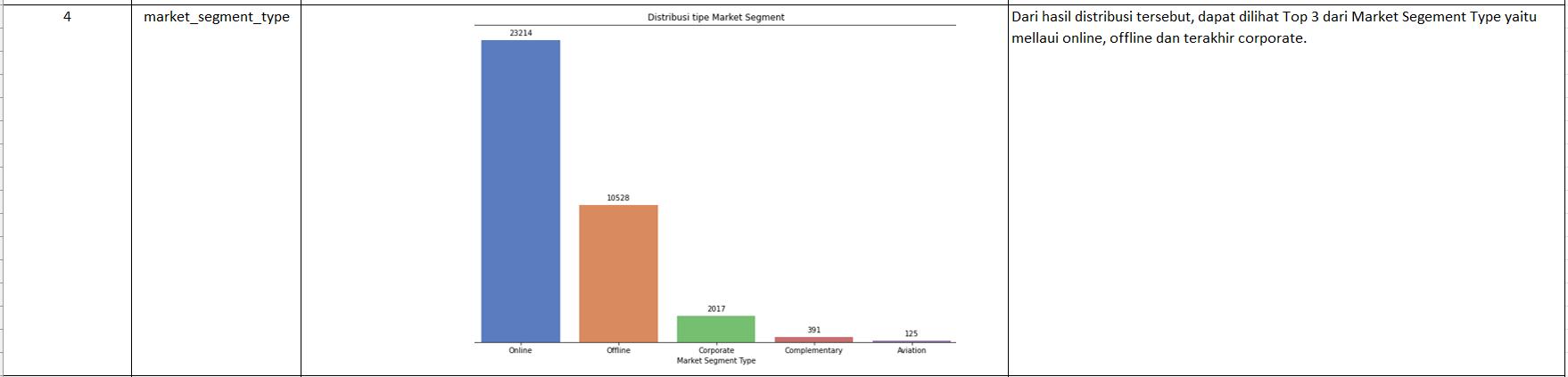


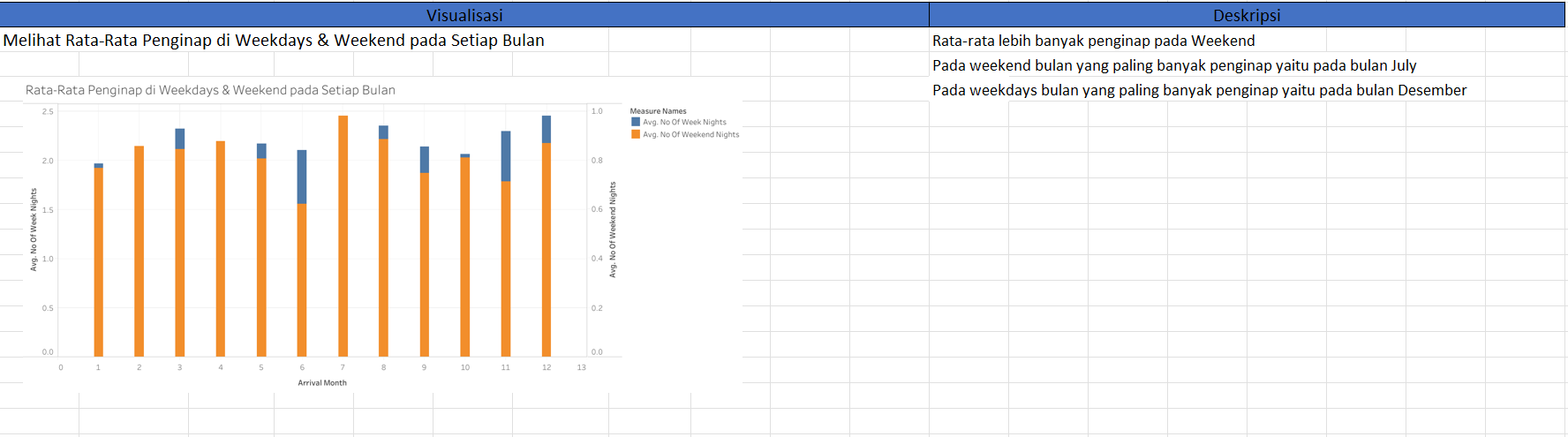
* 1. Sample 5 baris data pertama dan 5 baris data terakhir untuk 6 kolom pertama saja dari Hotel Reservation Dataset  
     

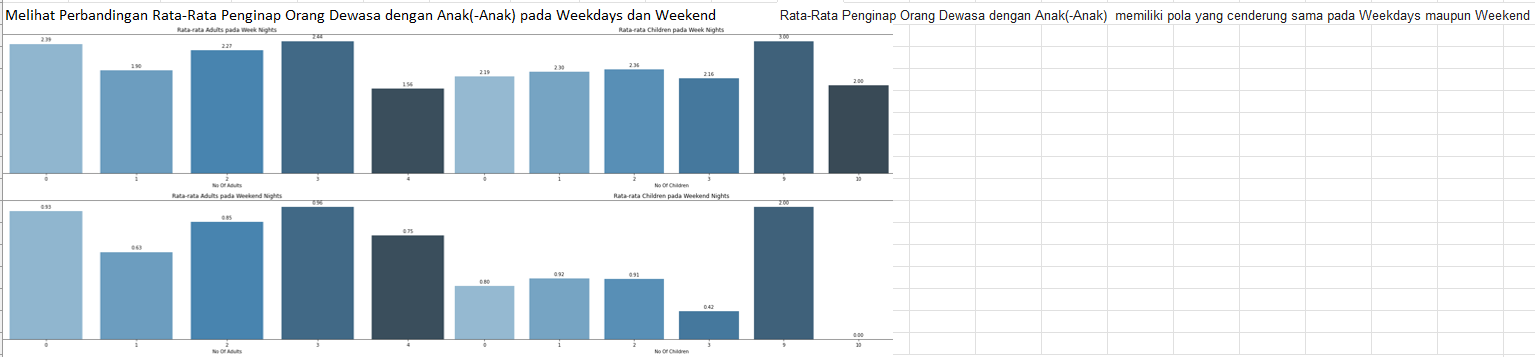


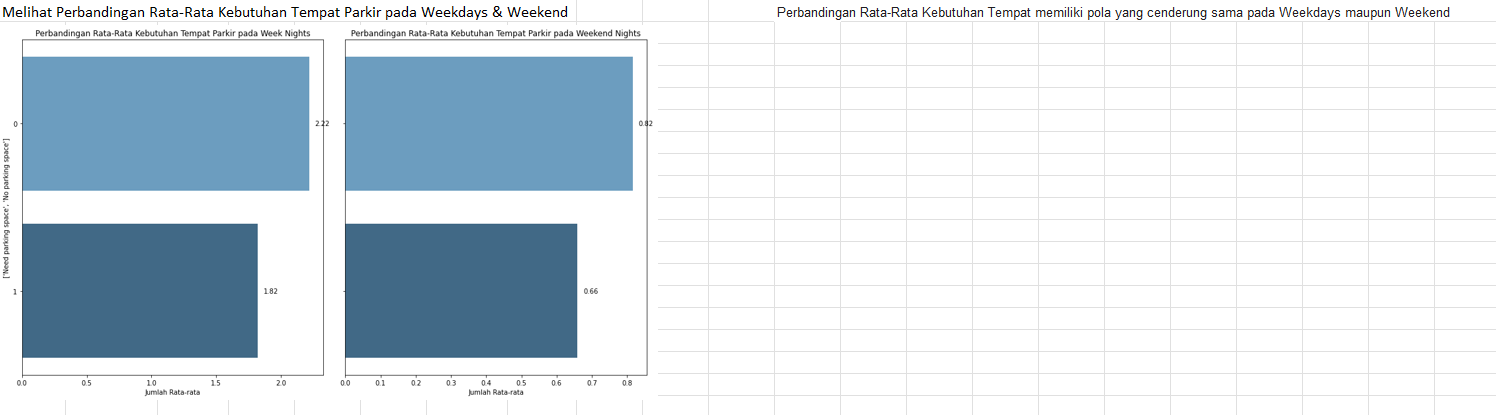
* 1. Descriptive Analysis

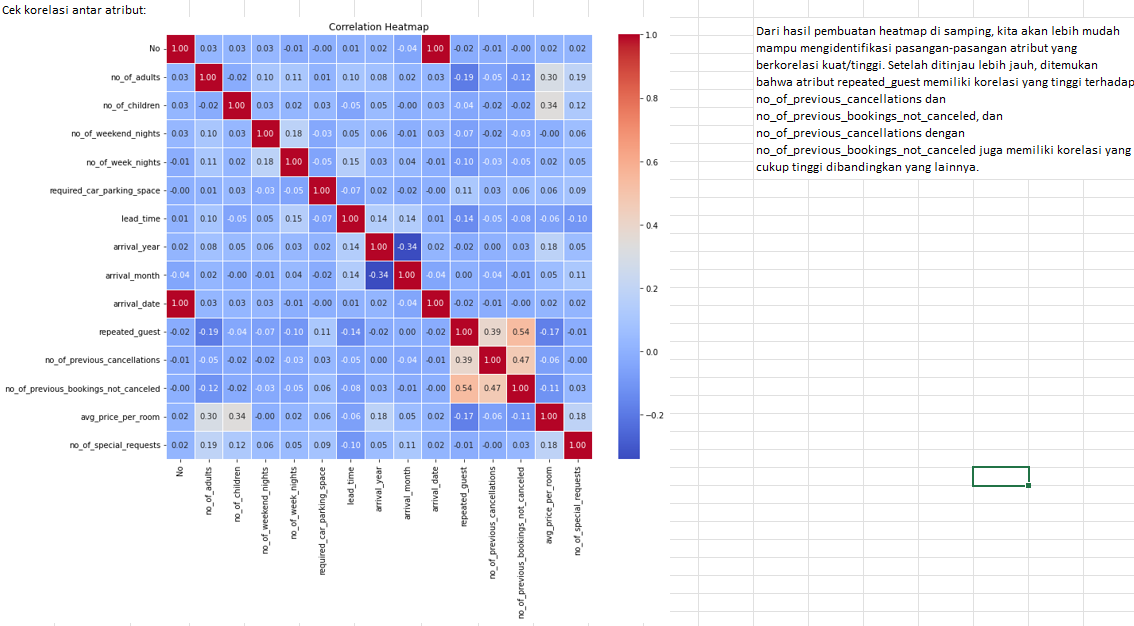




* 1. Pola Data

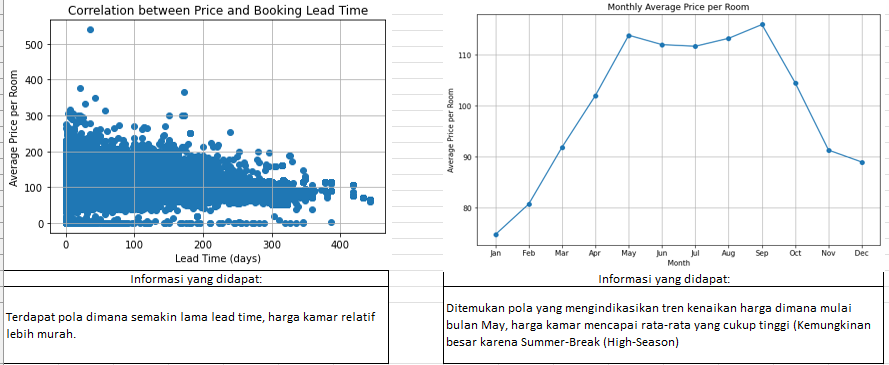
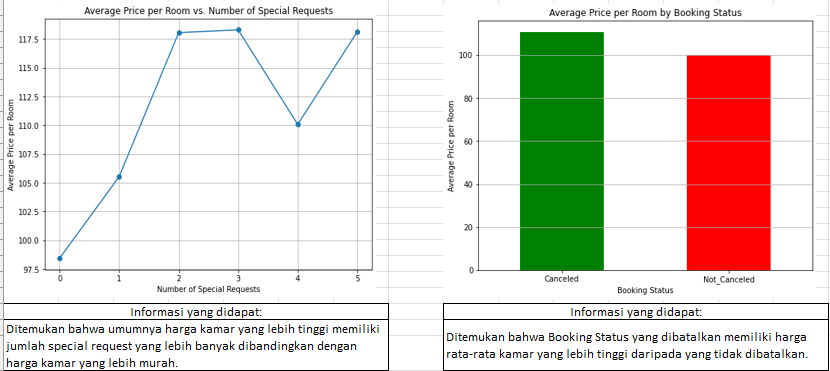
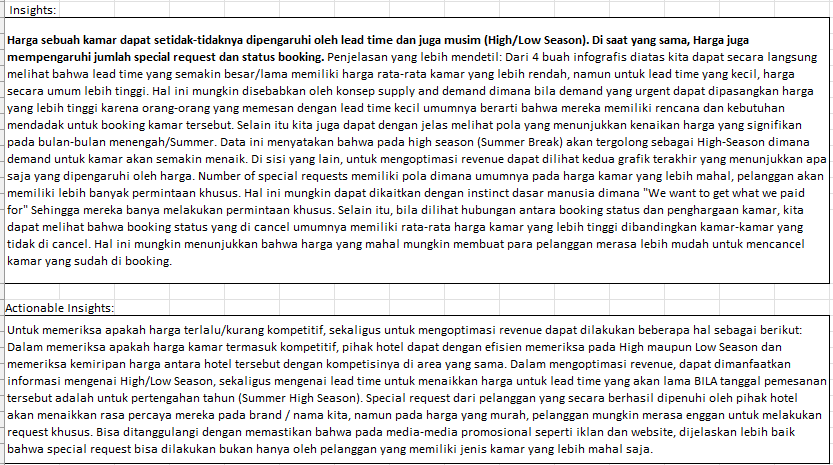


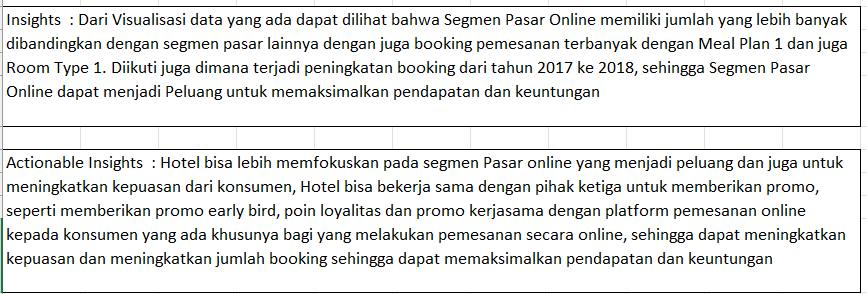
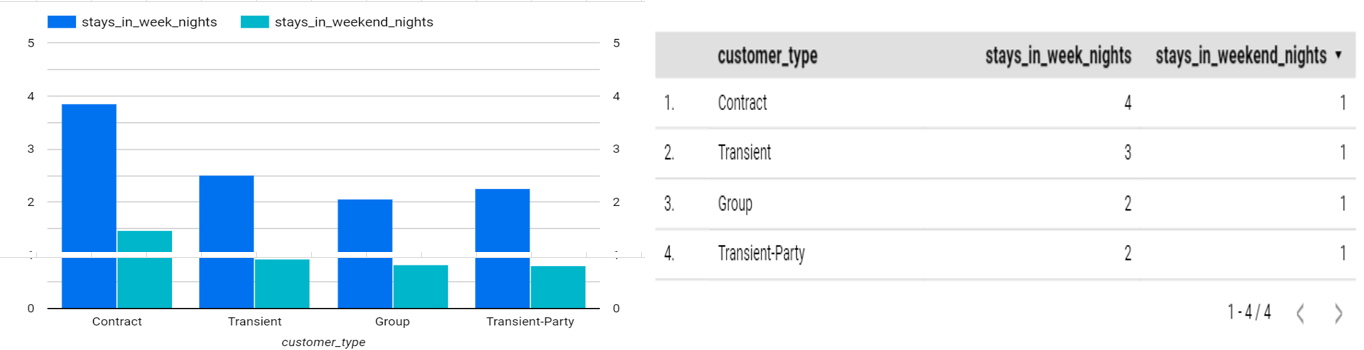
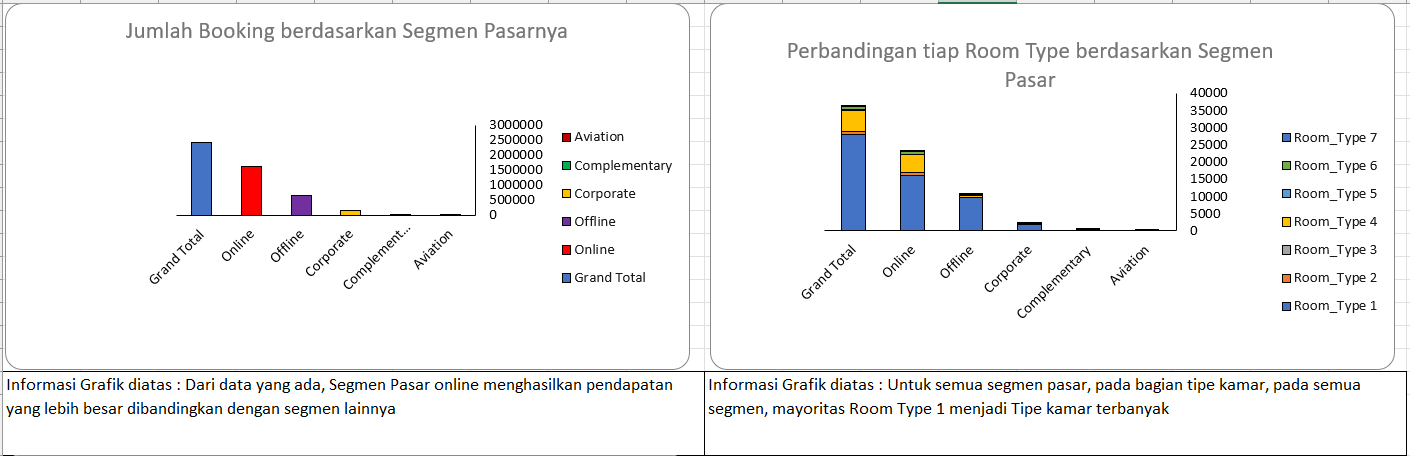
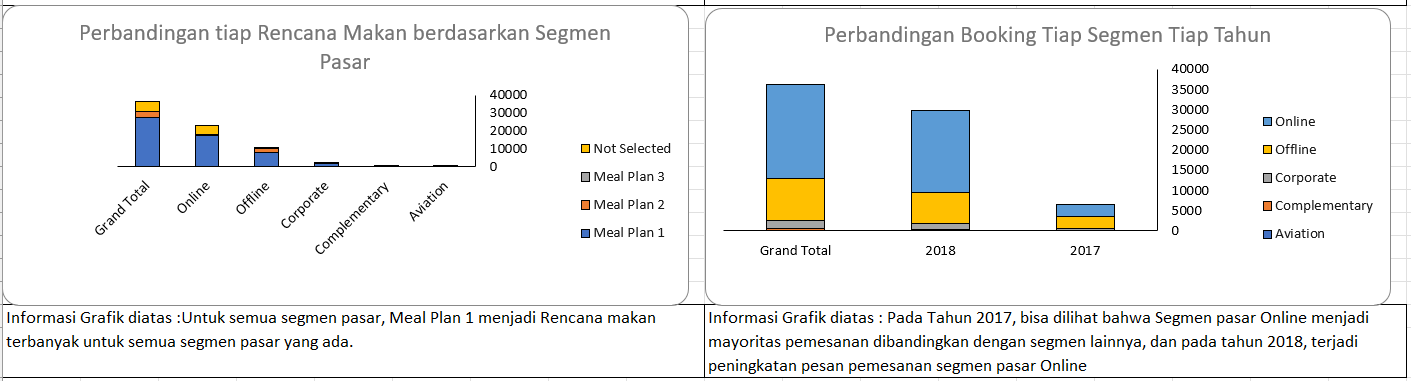
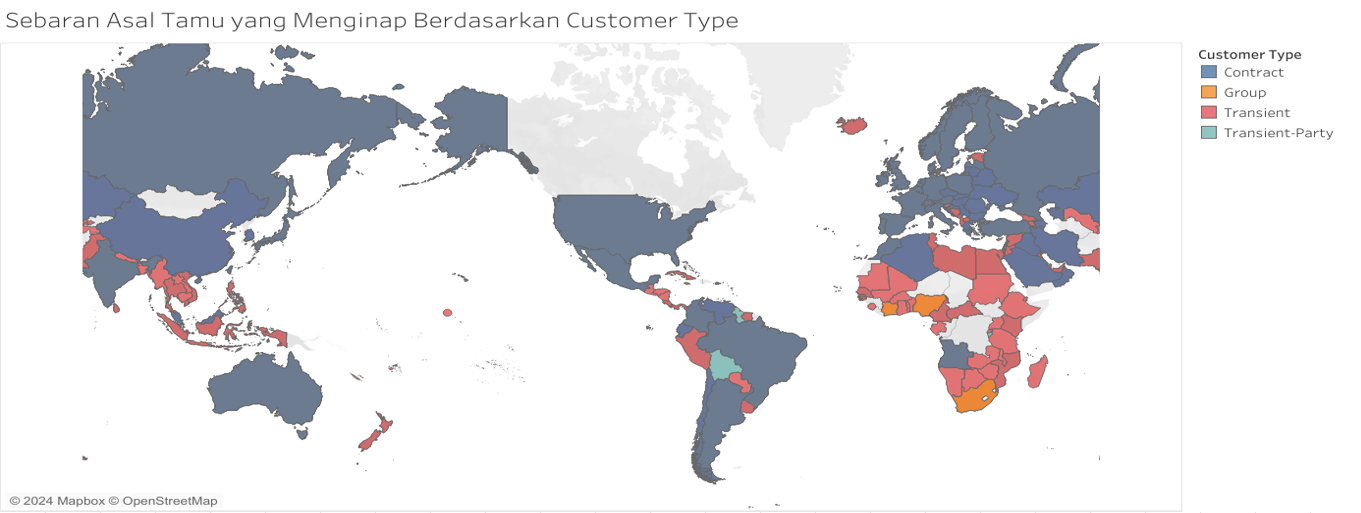
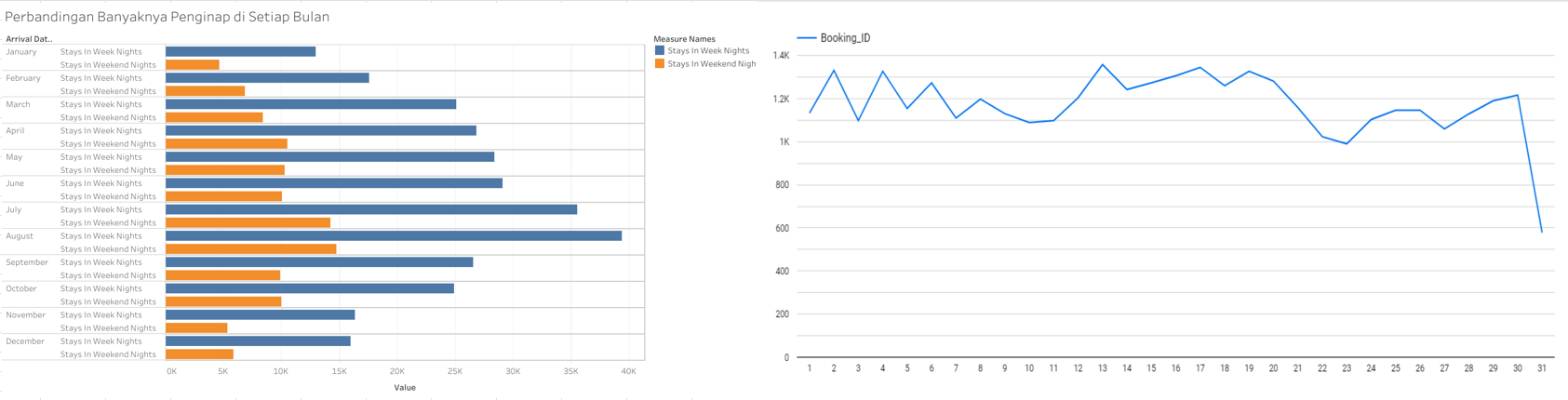
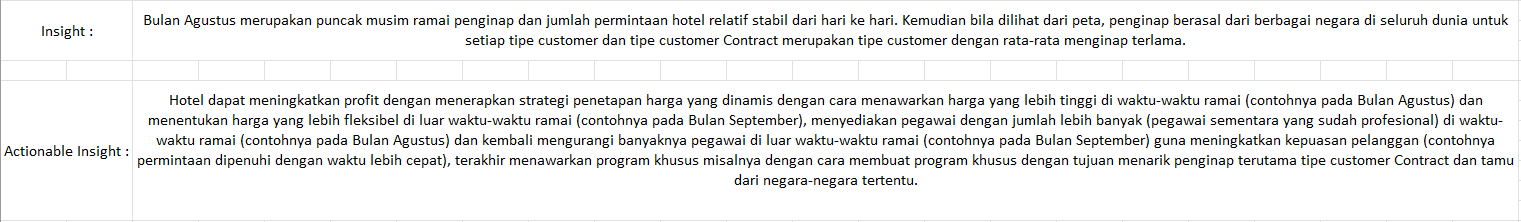


****

1. **Dashboard Hotel Reservation Dataset** :

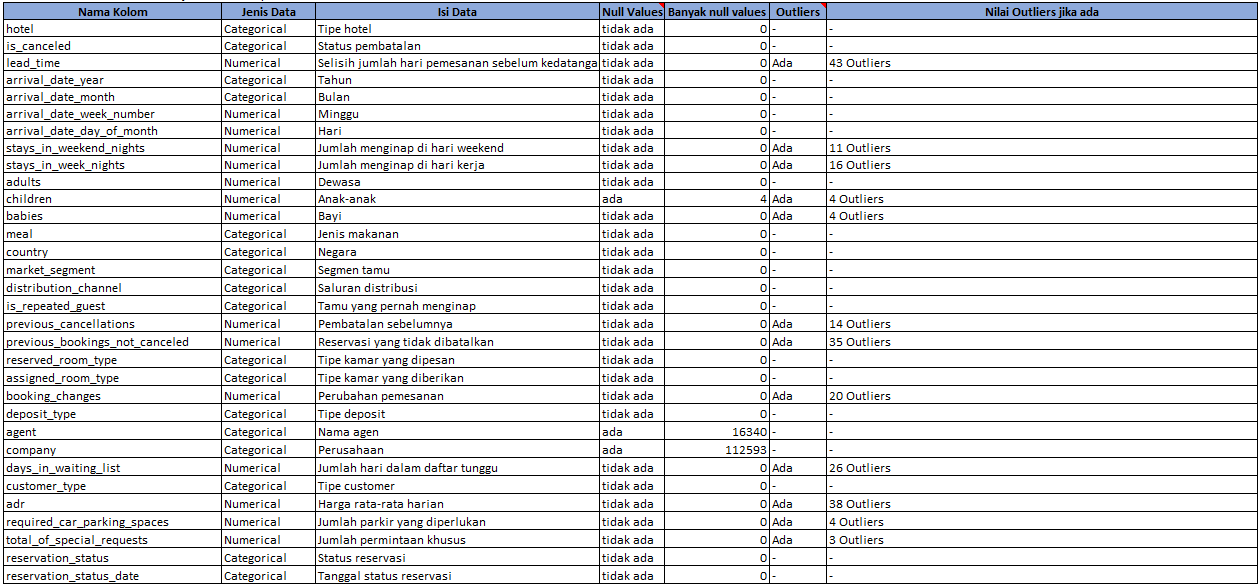
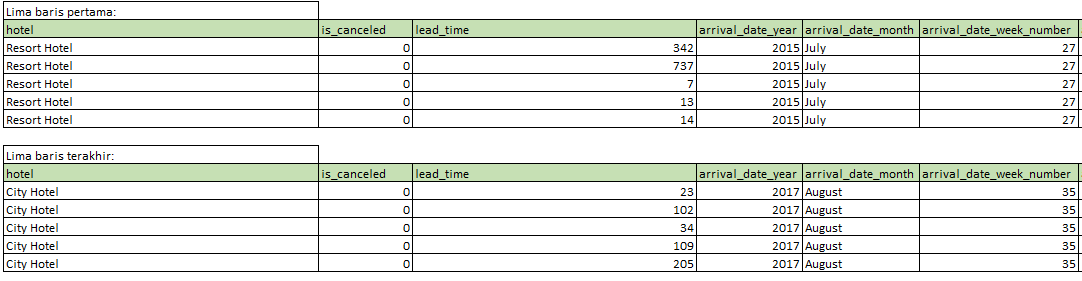
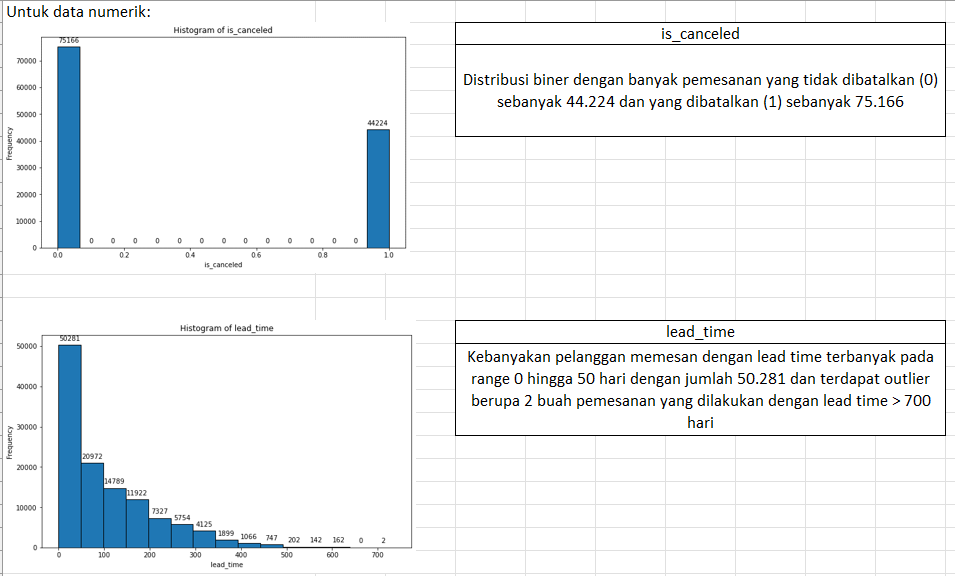
**Dashboard 1 :**

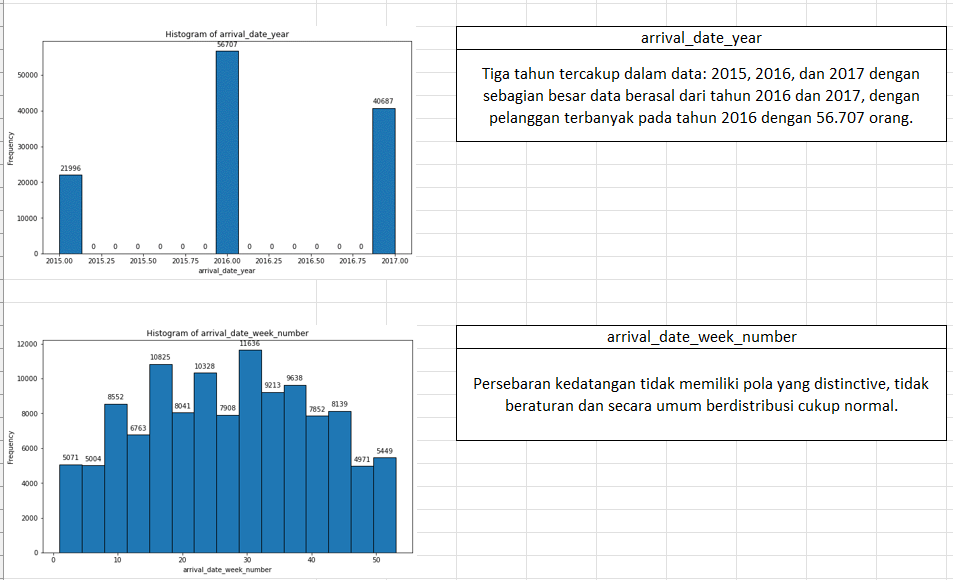
**  
**

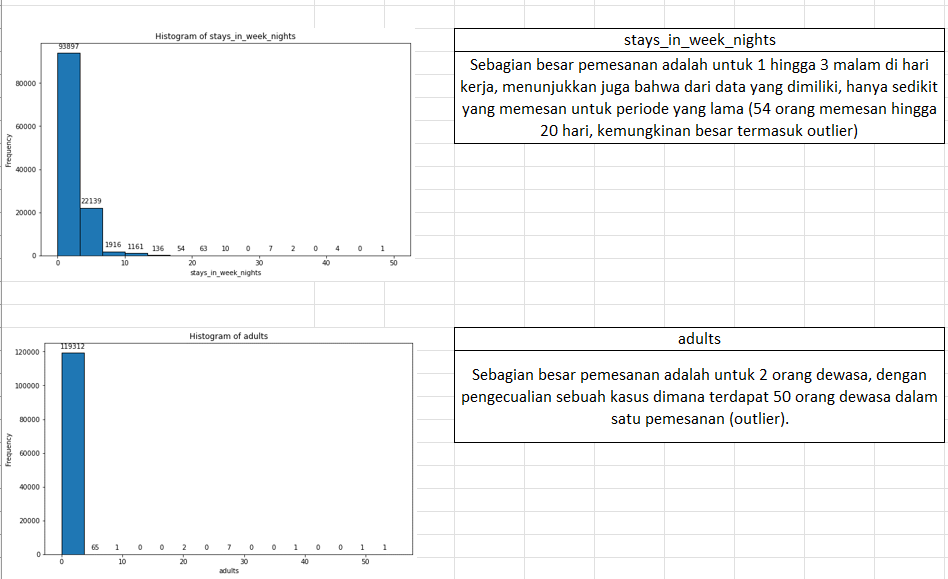
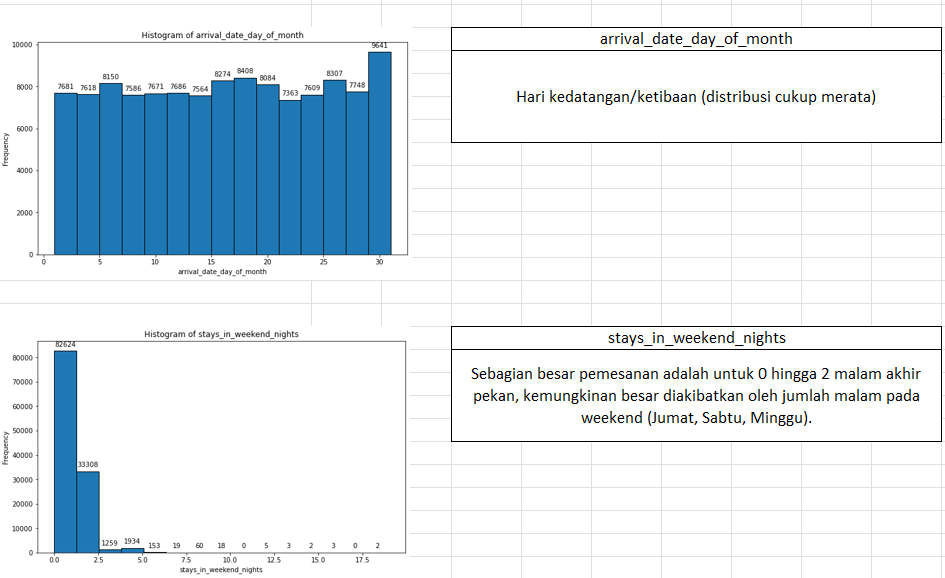
**Dashboard 2: Strategi Meningkatkan Profit berdasarkan Perilaku Penginap  
Dashboard 3 :   
Dashboard 3 :**

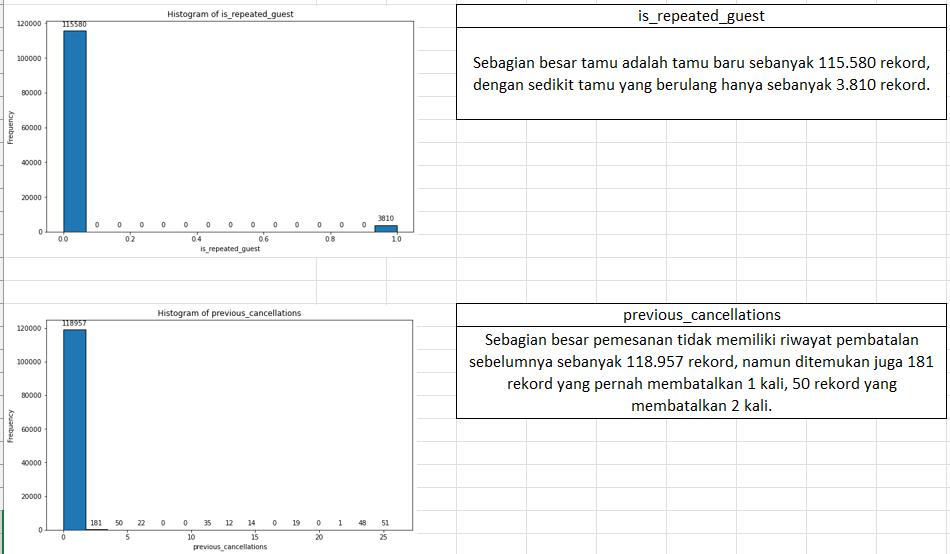
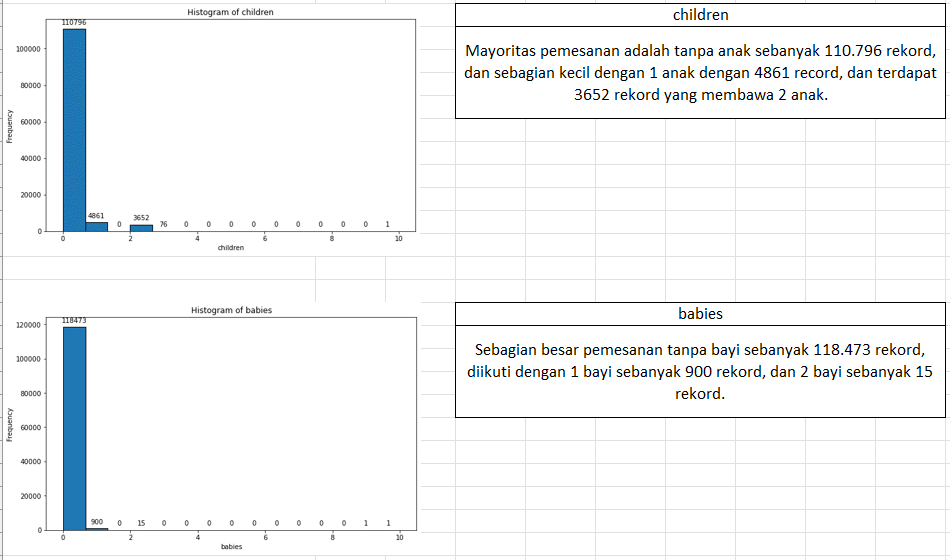
1. **DATASET Hotel Booking Demand**

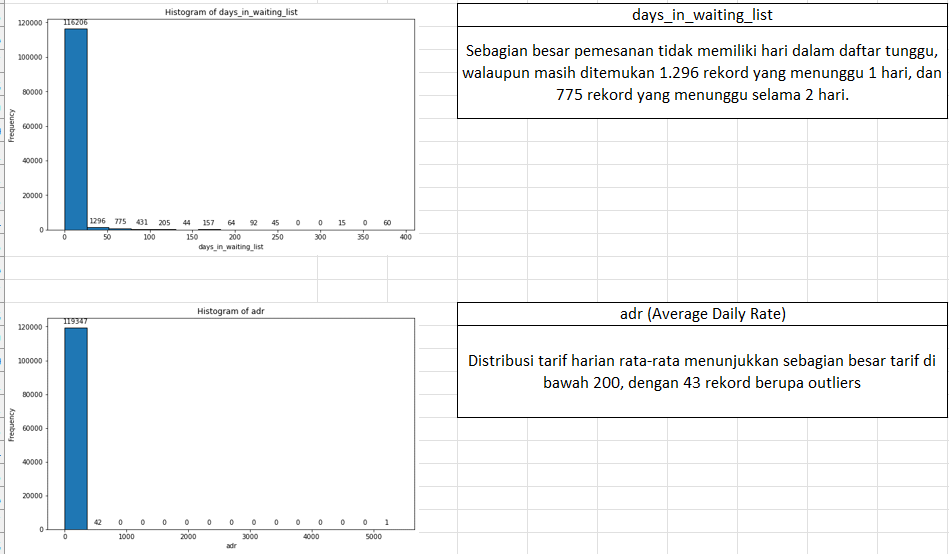
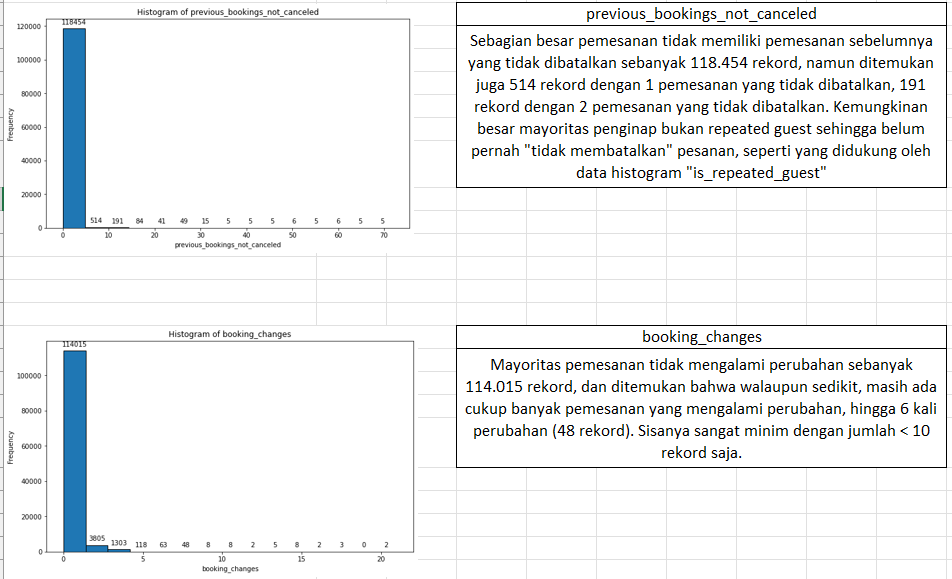
Eksplorasi umum dataset ():

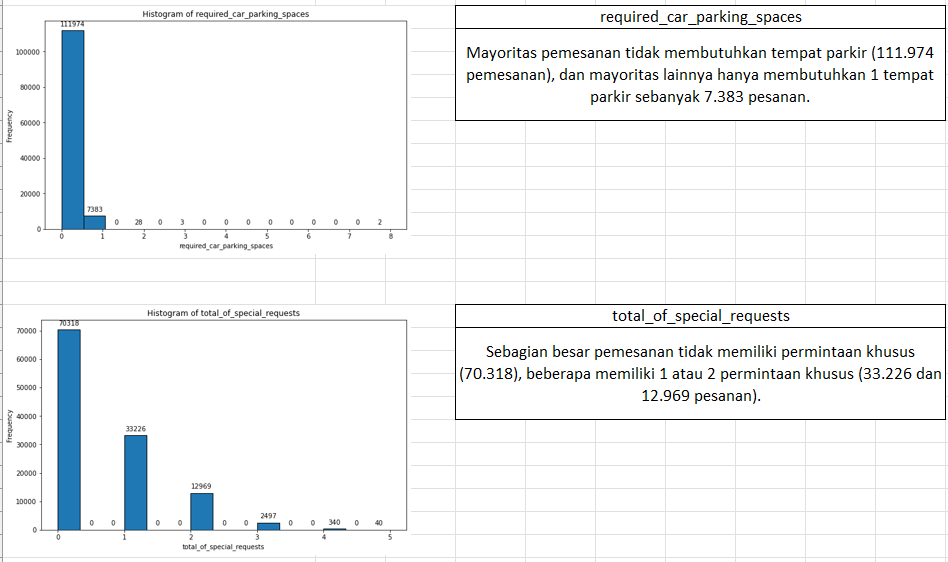
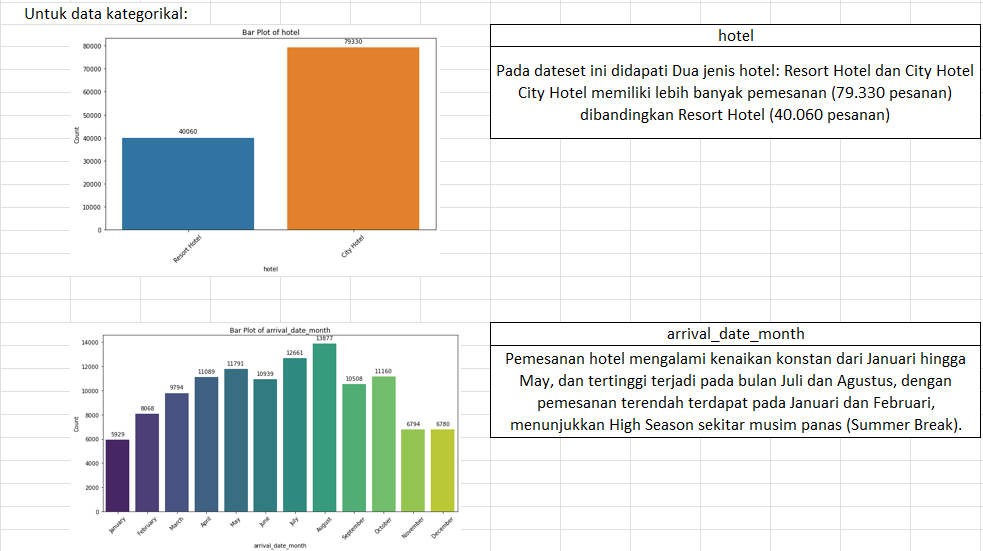
* + Narasi singkat mengenai dataset:  
    Dataset pemesanan hotel ini dapat membantu untuk menjawab pertanyaan kapan waktu yang terbaik dalam setahun untuk memesan kamar hotel, berapa lama waktu menginap yang optimal untuk mendapatkan tarif harian terbaik, dan Bagaimana memprediksi apakah sebuah hotel kemungkinan akan menerima jumlah permintaan khusus yang secara tidak proporsional tinggi
  + Link sumber dataset: [Hotel Booking Dataset](https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand)
  + Jumlah baris: 119391 Baris
  + Jumlah kolom: 32 Kolom
  + Identifikasi null values dan Outliers 
  + Sample 5 baris data pertama dan 5 baris data terakhir untuk 6 kolom pertama saja dari Hotel Booking Demand Dataset
  + Descriptive Analysis

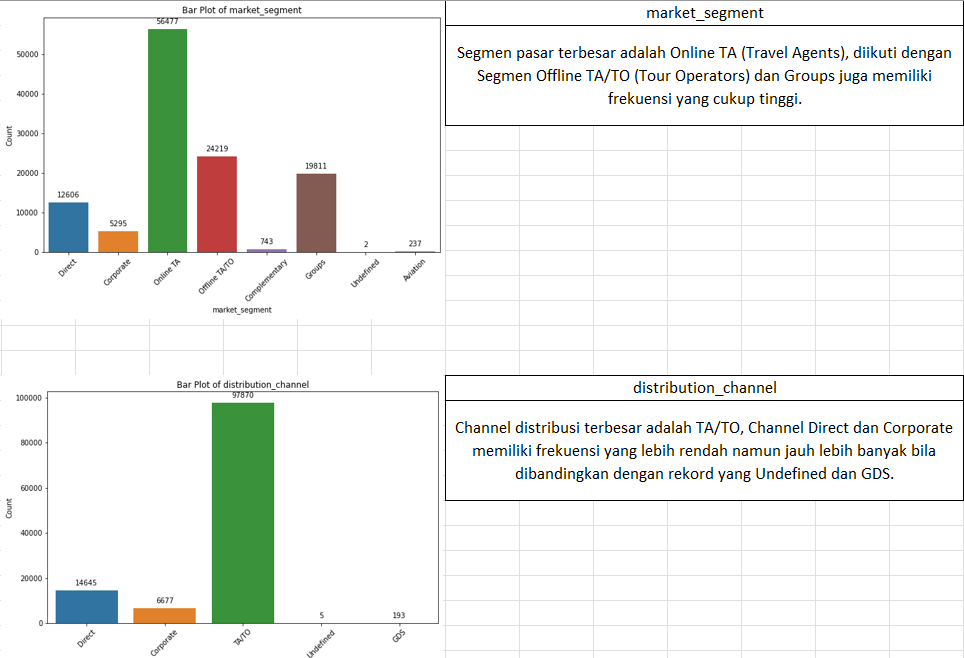
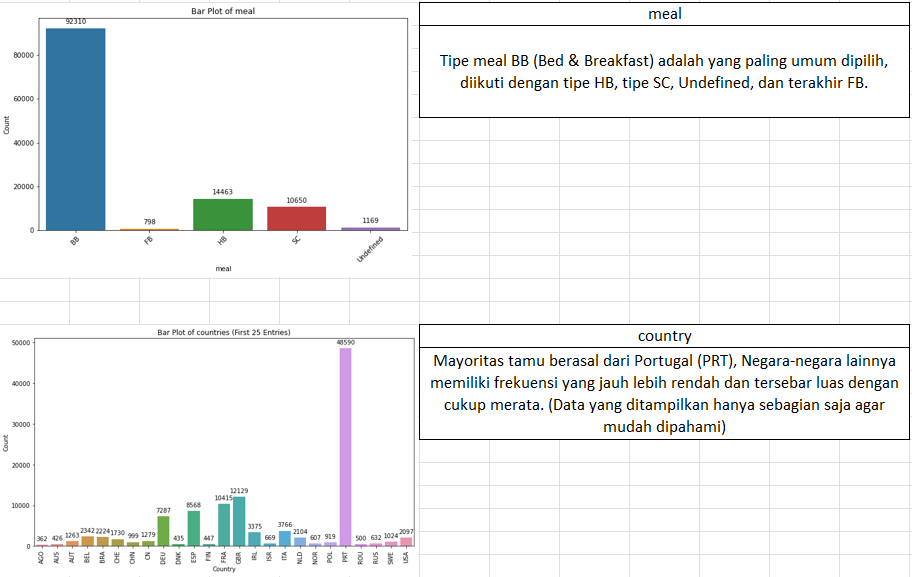


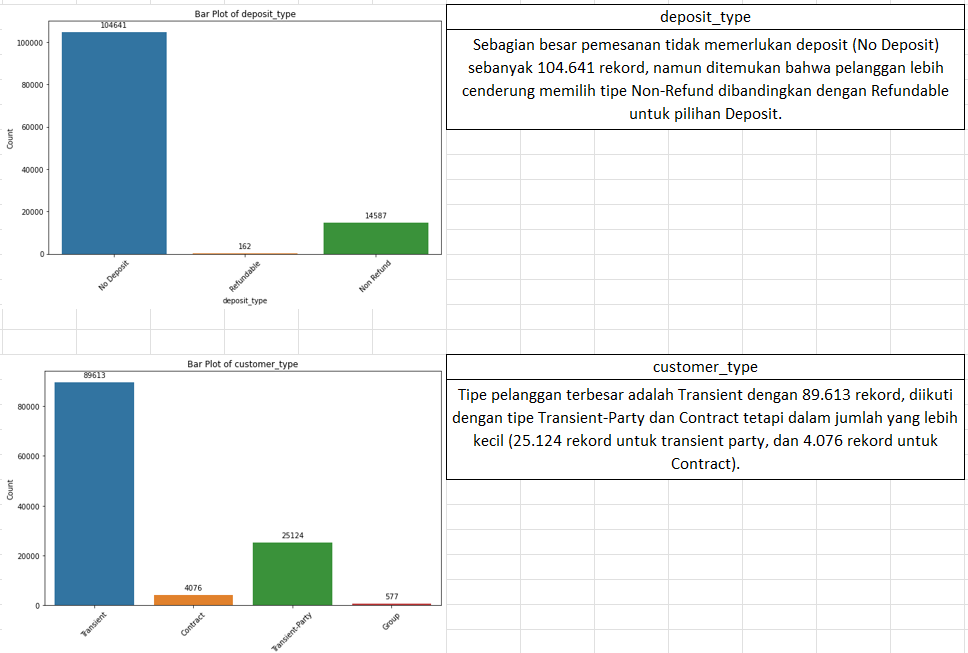


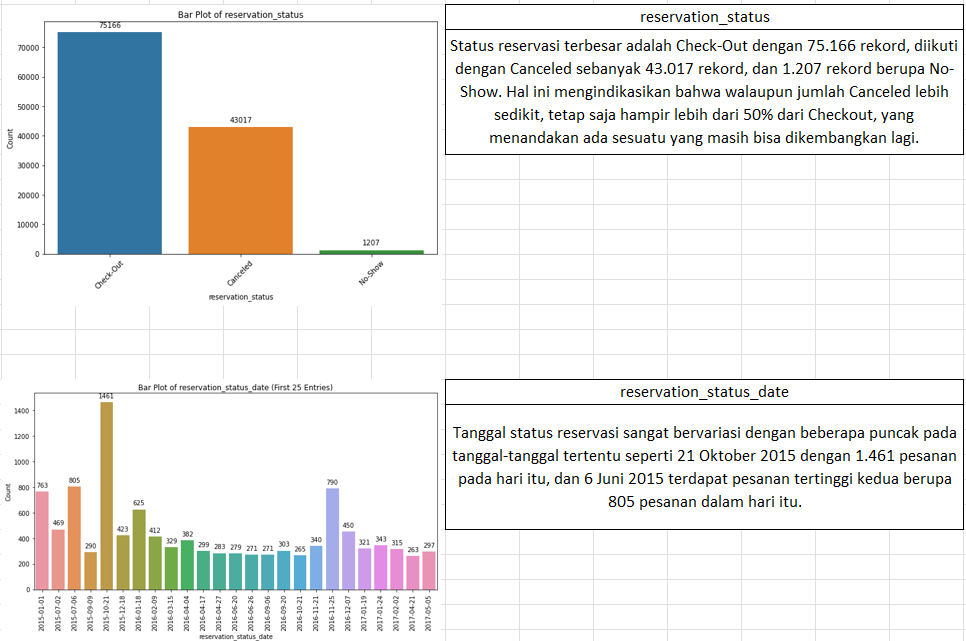


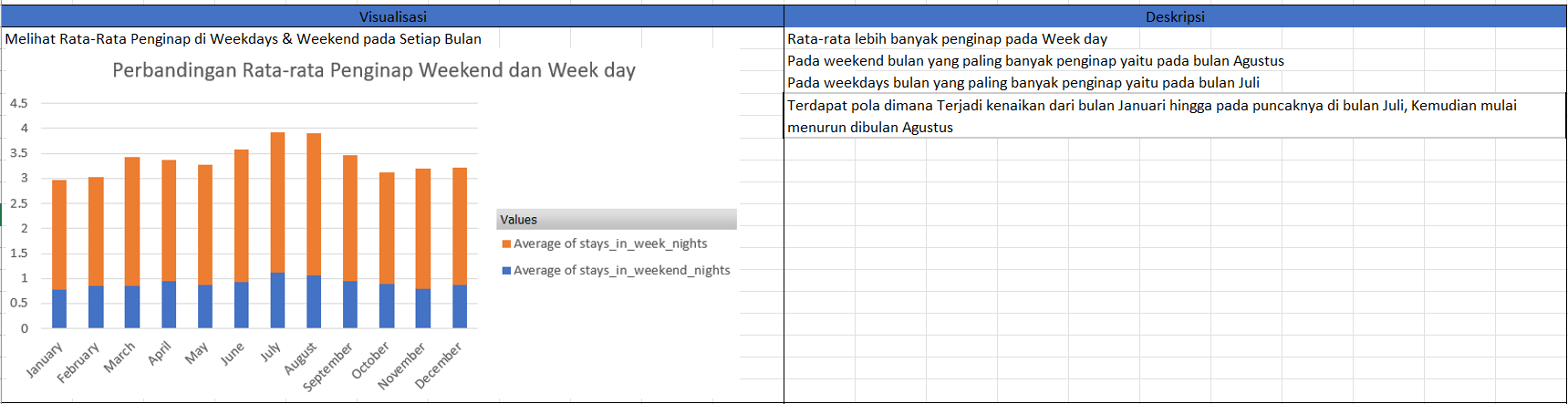










* + Pola Data

Catatan: Eksplorasi detail dapat dilihat pada file Tugas 3 versi final di folder UASDSDS\_3

* Enam model yang sudah dibuat dari dataset Hotel Booking (lihat dari tugas 4 yang sudah diperbaiki, di tabel no 6 sampai dengan kolom Alasan Model Dipilih):

| **No** | **Nama File** | **Pengukuran** | **Informasi yang diperoleh dari model** |
| --- | --- | --- | --- |
| 1 | M1\_KlasifikasiDTree\_01.ows | Accuracy: 0.585  Precision: 0.783  Recall: 0.673 | Dari Hasil model yang ada, dapat dilihat bila dilihat dari pengukuran, model cukup dapat mendeteksi atau memisahkan pemesanan yang dibatalkan berdasarkan Previous\_Cancellation dan Previous\_booking\_not\_canceled. Bila divisualisasikan juga bahwa semakin tinggi nilai previous\_booking\_not\_canceled, maka kemungkinan sebuah pemesanan kamar tidak dibatalkan semakin tinggi, tapi bila nilai Previous\_Cancellation tinggi, maka kemungkinan pemesanan kamar dibatalkan cukup tinggi juga. Ada kemungkinan kecil dimana jika jumlah previous\_cancellations nya lebih dari 2, maka booking akan dibatalkan dan juga jika perbandingan antara previous\_cancellations dan previous\_booking\_not\_canceled nya cukup jauh berbeda. Sehingga dapat disimpulkan bahwa Previous\_Cancellation dan Previous\_booking\_not\_canceled cukup berpengaruh untuk melihat pemesanan dibatalkan atau tidak |
| 2 | M2\_RegresiLR\_01.ows | MSE : 1285.376  RMSE : 35.852  MAE : 25.751  R² : 0.438 | Dari Hasil model yang ada, dapat dilihat bila dilihat bahwa nilai adr dipengaruhi oleh beberapa faktor seperti total\_of\_special\_requests , assigned\_room\_type ,hotel , adults, lead\_time, meal, dan arrival\_date\_month.Akan tetapi melihat pengukuran model yang ada model regresi linear ini tidak bisa dianggap sangat bagus. ² yang rendah dan MAPE yang sangat besar menunjukkan bahwa model ini memiliki performa yang kurang baik dalam memprediksi data. |
| 3 | M3\_KlasifikasiKNN\_01.py | Accuracy: 0.7556884167060949  Precision: 0.6701030927835051  Recall: 0.6408124480342083 | Informasi yang kami peroleh dari model ini :  Pembatalan pemesanan tampaknya cukup umum, mengingat recall model sebesar 64.08%. Hal ini menunjukkan bahwa ada sejumlah pemesanan yang dibatalkan di hotel tersebut, dan model mampu mendeteksi sebagian besar dari pembatalan tersebut. Selain itu, model memiliki akurasi yang cukup tinggi sebesar 75.56%, presisi sebesar 67.01% mengindikasikan bahwa sebagian besar prediksi model benar, precision yang lebih tinggi mengurangi kesalahan ketika model memprediksi pembatalan. |
| 4 | M4\_ClusteringKmeans\_01.py | - | Informasi yang kami peroleh dari model ini :  **Cluster 1 (Short Stay Short Lead Time)** cenderung memiliki risiko pembatalan yang moderat. Hal ini mungkin terjadi karena tamu dengan lead time yang singkat (pemesan kamar dalam waktu dekat dengan tanggal kedatangan) cenderung membuat keputusan pemesanan yang lebih cepat dan impulsif yang mana berpotensi lebih mudah membatalkan jika ada perubahan rencana mendadak dan juga mungkin mereka memiliki jadwal yang lebih fleksibel, sehingga lebih mungkin untuk mencari alternatif lain jika ada kesempatan yang lebih baik atau lebih murah.  **Cluster 1 (Long Stay Long Lead Time)** cenderung memiliki risiko pembatalan yang lebih rendah. Tamu dalam kelompok ini mungkin lebih yakin dengan rencana perjalanan mereka karena mereka memesan jauh-jauh hari dan menginap lebih lama. |
| 5 | M5\_KlasifikasiDTree\_01.py | Accuracy: 1.0  Precision: 1.0  Recall: 1.0  F1 Score: 1.0  Cross-Validation F1 Scores: [1. 1. 1. 1. 1.]  Mean F1 Score: 1.0  Standard Deviation of F1 Scores: 0.0 | Hasil yang didapatkan:  Model yang dibuat berhasil memprediksi dengan tepat sempurna seperti yang dapat dilihat dari skor yang didapatkan. Dari model ini kita dapat memahami fitur yang tergolong penting dan sangat berpengaruh dalam memprediksi pembatalan (Cancellation). |
| 6 | M6\_ClusteringK-Means\_01.py | Silhouette Score untuk k=4 setelah t-SNE: 0.5432828664779663 | Cluster 0 berisi customer yang cenderung melakukan booking dari jauh hari yang diindikasikan dengan median dan Inter Quartil Range yang lebih lebar. Cluster 1 dan 2 menunjukkan lead time yang sangat pendek, berarti booking yang terburu-buru (Last Minute). Segmen ini dapat ditarget dengan promo-promo last minute maupun policy yang lebih fleksibel. Cluster 3 memiliki lead time yang umum, mengindikasikan kestabilan waktu membooking (tidak terlalu cepat maupun terlalu lama). |

* Empat dari enam model terbaik yang diusulkan beserta alasannya (lihat dari tugas 4 yang sudah diperbaiki):

| **No** | **Nama File** | **Pengukuran** | **Informasi yang diperoleh dari model** |
| --- | --- | --- | --- |
| 1 | M3\_KlasifikasiKNN\_01.py | Accuracy: 0.7556884167060949  Precision: 0.6701030927835051  Recall: 0.6408124480342083 | Informasi yang kami peroleh dari model ini :  Pembatalan pemesanan tampaknya cukup umum, mengingat recall model sebesar 64.08%. Hal ini menunjukkan bahwa ada sejumlah pemesanan yang dibatalkan di hotel tersebut, dan model mampu mendeteksi sebagian besar dari pembatalan tersebut. Selain itu, model memiliki akurasi yang cukup tinggi sebesar 75.56%, presisi sebesar 67.01% mengindikasikan bahwa sebagian besar prediksi model benar, precision yang lebih tinggi mengurangi kesalahan ketika model memprediksi pembatalan. |
| 2 | M1\_KlasifikasiDTree\_01.ows | Accuracy: 0.585  F1-Score:0.582  Precision: 0.783  Recall: 0.673 | Dari Hasil model yang ada, dapat dilihat bila dilihat dari pengukuran, model cukup dapat mendeteksi atau memisahkan pemesanan yang dibatalkan berdasarkan Previous\_Cancellation dan Previous\_booking\_not\_canceled. Bila divisualisasikan juga bahwa semakin tinggi nilai previous\_booking\_not\_canceled, maka kemungkinan sebuah pemesanan kamar tidak dibatalkan semakin tinggi, tapi bila nilai Previous\_Cancellation tinggi, maka kemungkinan pemesanan kamar dibatalkan cukup tinggi juga. Ada kemungkinan kecil dimana jika jumlah previous\_cancellations nya lebih dari 2, maka booking akan dibatalkan dan juga jika perbandingan antara previous\_cancellations dan previous\_booking\_not\_canceled nya cukup jauh berbeda. Sehingga dapat disimpulkan bahwa Previous\_Cancellation dan Previous\_booking\_not\_canceled cukup berpengaruh untuk melihat pemesanan dibatalkan atau tidak |
| 3 | M4\_ClusteringKmeans\_01.py | - | Informasi yang kami peroleh dari model ini :  **Cluster 1 (Short Stay Short Lead Time)** cenderung memiliki risiko pembatalan yang moderat. Hal ini mungkin terjadi karena tamu dengan lead time yang singkat (pemesan kamar dalam waktu dekat dengan tanggal kedatangan) cenderung membuat keputusan pemesanan yang lebih cepat dan impulsif yang mana berpotensi lebih mudah membatalkan jika ada perubahan rencana mendadak dan juga mungkin mereka memiliki jadwal yang lebih fleksibel, sehingga lebih mungkin untuk mencari alternatif lain jika ada kesempatan yang lebih baik atau lebih murah.  **Cluster 1 (Long Stay Long Lead Time)** cenderung memiliki risiko pembatalan yang lebih rendah. Tamu dalam kelompok ini mungkin lebih yakin dengan rencana perjalanan mereka karena mereka memesan jauh-jauh hari dan menginap lebih lama. |
| 4 | M6\_ClusteringK-Means\_01.py | Silhouette Score untuk k=4 setelah t-SNE: 0.5432828664779663 | Cluster 0 berisi customer yang cenderung melakukan booking dari jauh hari yang diindikasikan dengan median dan Inter Quartil Range yang lebih lebar. Cluster 1 dan 2 menunjukkan lead time yang sangat pendek, berarti booking yang terburu-buru (Last Minute). Segmen ini dapat ditarget dengan promo-promo last minute maupun policy yang lebih fleksibel. Cluster 3 memiliki lead time yang umum, mengindikasikan kestabilan waktu membooking (tidak terlalu cepat maupun terlalu lama). |

Catatan: keseluruhan model yang sudah dibuat dapat dilihat pada folder tugas 4 versi final di folder UASDSDS\_NoKel (ganti dengan No kelompok anda)

* Insights yang diperoleh dari 4 Model di atas beserta actionable insightsnya (lihat file T4\_DB\_Nokel.xlsx)

| Insight Model 1:  Cluster 1: Memiliki lead time yang lebih pendek (di bawah 10 hari), stays\_in\_weekend\_nights yang sedikit lebih rendah, dan stays\_in\_week\_nights yang lebih rendah dibandingkan dengan Cluster 2. Hal ini menunjukkan bahwa Cluster 1 mungkin terdiri dari tamu yang memesan dalam waktu yang lebih singkat dan tinggal untuk waktu yang lebih singkat.  Cluster 2: Memiliki lead time yang jauh lebih panjang (di atas 10 hari), stays\_in\_weekend\_nights dan stays\_in\_week\_nights yang lebih tinggi. Hal ini menunjukkan bahwa Cluster 2 mungkin terdiri dari tamu yang memesan jauh-jauh hari sebelumnya dan tinggal untuk waktu yang lebih lama.  Actionable insight Model 1:  Insight Model 2:  Cluster 1 (sama dengan Cluster 0): 'Short Lead Time, Short Stay' (kelompok paling berpotensi melakukan pembatalan)  Cluster 2 (sama dengan Cluster 1): 'Long Lead Time, Long Stay'  Actionable insight Model 2:  Insight Model 3:  Label 0 merupakan pemesanan yang tidak dibatalkan Label 1 merupakan pemesanan yang dibatalkan  Dari Visualisasi diatas dapat dilihat, jika jumlah previous\_cancellations-nya lebih dari 2, maka pemesanan cenderung tidak dibatalkan, dan juga jika perbandingan antara previous\_cancellations dan previous\_booking\_not\_canceled cukup jauh berbeda.  Sehingga dapat disimpulkan bahwa Previous\_Cancellation dan Previous\_booking\_not\_canceled cukup berpengaruh untuk melihat pemesanan dibatalkan atau tidak.  Actionable insight Model 3:  Insight Model 4:  Hasil dari clustering menunjukkan 2 cluster yang berisi bookings yang memiliki kemiripan dalam berbagai karakteristik dilihat dari gabungan fitur yang digunakan dalam komponen t-SNE.  Cluster 0 berisi booking yang dekat dengan hari penginapan (Lead time kecil) karena kebanyakan booking di cluster ini memiliki lead time yang rendah, dengan beberapa pengecualian berupa outlier dengan lead time yang tinggi.  Cluster 1 berisi booking yang dilakukan dari jauh hari (memiliki lead time yang cenderung lebih besar) yang diindikasikan dengan inter-quartil range yang lebih lebar dan besar.  Actionable insight Model 4: |
| --- |

1. **Lampiran**

* Tabel kontribusi tugas akhir (urutkan berdasarkan NPM secara ascending):

| **No** | **NPM** | **Nama** | **Kontribusi Pekerjaan (%)** | **Tanda Tangan** | **Kelas** |
| --- | --- | --- | --- | --- | --- |
| 1 | 6182001001 | Jenson Mark Lowell | 100% |  | A |
| 2 | 6182101040 | Samuel Edward Winoto | 100% |  |
| 3 | 6182101054 | Steffi Widjaya | 100% |  |

* Lampiran script coding (jika ada) terlampir di dalam folder folder UASDSDS\_01

1. Referensi

UNWTO. (2019). Tourism definitions.  
<https://webunwto.s3.eu-west-1.amazonaws.com/s3fs-public/2019-11/unwto-tourism-definitions.pdf>

Ibrahim, N. (2024, March 5). Technology insights in the hotel industry.