Analisis Perbedaan Resort Hotel dan City Hotel

Hasri Wiji Aqsari, Santi Wulan Purnami, Kartika Fithriasari dan Irhamah
Departemen Statistika, Fakultas Sains dan Analitika Data,
Institut Teknologi Sepuluh Nopember (ITS)
Jl. Arief Rahman Hakim, Surabaya 60111 Indonesia
e-mail : santiwulan08@gmail.com

Abstrak—Hotel adalah salah satu hal terpenting di dunia pariwisata. Perusahaan hotel berjumlah sangat banyak. Dan masing-masing dari perusahaan tersebut berlomba-lomba untuk mendapatkan pelanggan terbanyak. Pada penelitian ini akan dilihat perbedaan antara resort hotel dan city hotel. Data diperoleh melalui *website Kaggle*. Paca proses *preprocessing* data diketahui bahwa terdapat 3 variabel yang mengandung *missing value* diantaranya variabel *company, agent* dan *country*. Menangani *missing value* dilakukan dengan cara menghapusnya, karena variabel yang terdapat *missing value* tersebut adaah variabel kategorik, dan kedepannya tidak dilibatkan dalam visualisasi maupun analisis selanjutnya. Setelah dilakukan visualisasi data, dapat diketahui bahwa pelanggan lebih banyak memilih resort hotel daripada city hotel, diperjelas dengan waktu tunggu atau *waiting list* pada resort hotel lebih lama daripada city hotel. Analisis *cluster* menggunakan metode *K-Means* tidak menghasilkan kesimpulan yang baik. Analisis menggunakan PCA didapatkan hasil 32 variabel asal direduksi menjadi 10 PC.

Kata Kunci: City Hotel, *Missing value,* Resort Hotel, *Preprocessing*, Visualisasi,

I. PENDAHULUAN

Pariwisata merupakan salah satu industry yang memiliki peran penting dalam mengingkatkan devisa negara. Hal ini dibuktikan dengan pemberitaan di media cetak, online maupun TV yang selalu mengekspos informasi mengenai dunia pariwisata. Oleh karena itu, ilmuilmu yang berkaitan dengan pariwisata mempengaruhi berbagai aspek diantaranya budaya, ekonomi, hukum/politik, pemasaran, psikologi pelayanan, agro dan pelayanan masyarakat.

Pengertian hotel menurut Surat Keputusan Mentri Pariwisata, Pos dan Telekomunikasi No. KN/34/HK/103/MPPT tahun 2987 sebagai berikut, hotel adalah satu jenis akomodasi yang menggunakan sebagaian maupun keseluruhan bangunan untuk menyediakan jasa pelaanan penginapan, makanan dan minuman serta jasa-jasa lainnya bagi umum yang dikelola secaara komersial serta memenuhi persyaratan yang telah ditetapakan oleh pemerintah.

Suatu hotel pasti membutuhkan kerjasama dari berbagai pihak seperti travel agent, dinas pariwisata, dinas perijinan dan kepolisian. Dewasa ini, industry perhotelan semakin berkembang oesat. Mulai dari pembangunan hotel yang meningkat sampai penambahan jumlah kamar dan fasilitas yang dibutuhkan untuk memenuhi kepuasan tamu. Oleh karena itu, penulis menyadari bahwa industry perhotelan sangat dibutuhkan oleh wisatawan sebagai penunjang akomodasi bagi kegiatan wisata.

Hotel resort dan hotel city adalah salah satu hotel yang berada di New York. Akan dianalisis karakteristik kedua hotel tersebut melalui visualisasi data, *clustering festure extraction* dan PCA.

II. TINJAUAN PUSTAKA

A. Preprocessing Data

Preprocessing data adalah salah satu teknik dari data mining yang termasuk persiapan dan transformasi data ke bentuk yang cocok untuk dilakukan prosedur mining. Data preprocessing bertujuan untuk mengurangi dimensi data, mencari hubungan antar data, menormalisasi data, mengeluarkan data pencilan, dan mengekstrak fitur untuk data. Beberapa teknik yang termasuk dalam proses data preprocessing adalah pembersihan data, integrasi, tranformasi, dan reduksi. Berikut beberapa istilah dalam *preprocessing* data :

1. *Missing values.* Bila ada data yang nilainya tidak tercatat, maka nilai yang hilang ini dapat diisi dengan menggunakan nilai rata-rata.
2. *Binning*. Metode ini digunakan untuk menyimpan data berdasarkan "tetangga", yaitu nilai di sekitarnya. Nilai yang telah diurutkan dibagi ke beberapa kelompok. Metode ini bergantung pada data disekitarnya, sehingga dilakukan local smoothing.
3. Reduksi dimensi. Metode ini menggunakan mekanisme pengkodean untuk mengurangi dataset. Reduksi dimensi dapat dilakukan dengan tanpa ada informasi yang hilang. Ada dua tipe reduksi dimensi, yaitu feature selection dan feature extraction.
4. *Feature selection.* Digunakan untuk menyaring fitur atau variabel yang tidak berhubungan atau mubazir dari dataset. Beberapa metode *feature selection* yaitu :
   a. *Variance Thresholds. Variance thresholds* menghilangkan fitur yang nilainya tidak berubah banyak dari observasi ke observasi.
   b. *Correlation Thresholds. Correlation Thre-sholds* menghilakan fitur yang berkorelasi atau berhubungan tinggi dengan fitur lain.

B. Missing value

Missing value adalah informasi yang tidak tersedia untuk sebuah objek (kasus). *Missing value* terjadi karena informasi untuk sesuatu tentang objek tidak diberikan, sulit dicari, atau memang informasi tersebut tidak ada. *Missing value* pada dasarnya tidak bermasalah bagi keseluruhan data, aplagi jika jumlahnya hanya sedikit, misahl hanya 1% dari seluruh data. Namun jika persentase data yang hilang tersebut cukup besar, maka perlu dilakukan pengujian apakah data yang mengandung banyak *missing* tersebut masih layak diproses lebih lanjut atau tidak [3].

C. Outlier

Outlier adalah kasus atau data yang memiliki karakteristik unik yng terlihat sangat berbeda jauh dari observasi-observasi lainnya dan muncul dalam bentuk nilai ekstrim baik untuk sebuah variabel tunggal atau kombinasi. Empat penyebab munculnya *outlier* yaitu

kesalahan dalam memasukkan data, gagal menspesifikasi adanya *missing value*, *outlier* bukan merupakan anggota populasi yang diambil sebagai sampel, dan tidak berdistribusi secara normal [4].

D. Statistika Deskriptif

Statistika deskriptif merupakan metode-metode yang berkaitan dengan pengumpulan dan penyajian suatu gugus data sehingga memberikan informasi yang berguna. Statistika deskriptif dalam praktikum ini terdiri dari *mean*, simpangan baku, median,nilai maksimum dan nilai minimum.

Mean atau rata-rata adalah jumlah semua data yang ada dibagi dengan banyaknya data. Data tunggal memiliki rumus :

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n} \quad (1)$$

Keterangan :

$\bar{x}$: rata-rata

x_n : sukuk e-n

n : banyaknya data

Simpangan baku menunjukkan penarikan akar dari rata-rata kuadrat jarak suatu data terhadap rata-ratanya. Simpangan baku memiliki rumus :

$$S = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}} \quad (2)$$

Keterangan :

S : simpangan baku

n : banyaknya data

x_i : data ke-i

$\bar{x}$: rata-rata

Median adalah nilai tengah dari data yang telah disusun berurutan mulai dari data yang terkecil sampai dengan yang terbesar. Secara matematis median dilambangkan dengan *Me* yang dapat dihitung dengan rumus :

$$Me = x_{\frac{n+1}{2}} \quad (3)$$

Keterangan :

Me : median

n : banyaknya data

x : nilai data

Statistik lima serangkai. Nilai minimum, maksimum, dan kuartil memberikan informasi mengenai nilai tengah dan variasi dari sebuah variabel. Ditulis dengan urutan dari yang terkecil hingga terbesar, maka hal ini dikatakan *five-number summary* dari sebuah variabel. Statistik lima serangkai memberikan informasi tentang lokasi dan sebaran suatu data.

Min, Q1, Q2, Q3, Max (4)

Keterangan :

Min : nilai minimum

Q1 : kuartil bawah

Q2 : median

Q3 : kuartil atas

Max : nilai maksimum

E. Visualisasi Data

Bidang visualisasi difokuskan pada penciptaan gambar yang menyampaikan infornasi penting tentang data yang mendasari [1]. Visualisasi data dilihat oleh banyak bidang ilmu sebagai komunikasi visual modern. Visualisasi data tidak berada di bawah bidang manapun, melainkan interpretasi di antara banyak bidang misalnya, terkadang dilihat sebagai cabang modern dari statistik deskriptif oleh beberapa orang, tetapi juga sebagai dasar alat pengembangan oleh yang lain. Visualisasi data mengikutkan pembuatan dan kajian dari representasi visual dari data, artinya informasi yang telah diabstraksikan dalam bentuk skematis, termasuk atribut atau variabel dari unit informasi.

1. Histogram

Data yang telah disusun dalam bentuk tabel distribusi frekuensi dapat disajikan dalam bentuk diagram yang disebut histogram, berikut adalah beberapa pengertian histogram. Histogram yaitu merupakan grafik dari distribusi frekuensi suatu variabel. Tampilan histogram berupa petak-petak empat persegi panjang. Sebagai sumbu horizontal (absis, sumbu x) boleh memakai tepi-tepi kelas, batas-batas kelas atau nilainilai variabel yang diobservasi, sedang sumbu vertical (ordinat, sumbu y) menunjukan frekuensi. Untuk distribusi bergolong/ kelompok yang menjadi absis adalah nilai tengah dari masing-masing kelas [2].

Histogram adalah grafik yang menggambarkan suatu distribusi frekuensi dengan bentuk beberapa segi empat [3]. Histogram merupakan grafik batang dari distribusi frekuensi [4]. Histogram adalah suatu bentuk grafik yang menggambarkan sebaran (distribusi) frekuensi suatu perangkat data dalam bentuk batang [5]. Histogram digunakan untuk menggambarkan secara visual frekuensi data yang bersifat kontinu.

Jadi histogram adalah diagram kotak yang lebarnya menunjukkan interval kelas, sedangkan batas-batas tepi kotak merupakan tepi bawah dan tepi atas kelas, dan tingginya menunjukkan frekuensi pada kelas tersebut.

2. *Scatter Diagram*

Scatter diagram adalah gambaran yang menunjukkan kemungkinan hubungan antara keeratan hubungan antara dua variabel tersebut yang sering diwujudkan sebagai koefisien korelasi.

3. *Density Plot*

Density plot memvisualisasikan distribusi data melalui interval atau periode waktu yang berkesinambungan. Plot ini adalah variasi dari histogram yang menggunakan kernel smoothing untuk memplot nilai, memungkinkan distribusi yang lebih halus dengan menghaluskan kebisingan. Puncak dari density plot membantu menampilkan di mana nilai terkonsentrasi selama interval [6].

4. *Correlogram*

Correlogram adalah sebuah grafik yang menunjukkan korelasi antara dua variabel.

5. *Line Plot*

Line plot atau diagram garis dinilai merupakan bentuk yang sangat tepat untuk menggambarkan data yang berhubungan dengan runtutan waktu (*time series data*). Dengan kata lain, corak diagram ini paling sesuai digunakan untuk memvisualisasikan perkembagan dinamika keadaan dari satu kurun waktu ke rentang

waktu berikutnya [8]

F. Feature Extraction menggunakan *Principal Component Analysis*

Ekstraksi fitur merupakan proses yang bertujuan untuk menentukan ciri-ciri. Adapun tahapan ekstraksi fitur menggunakan PCA sebagai berikut [11].

1. Menghitung rata-rata keseluruhan sampel data diperoleh dengan menggunakan persamaan sebagai berikut.

$$\bar{x}_j = \frac{\sum_{ij=1}^{n} x_{ij}}{n} \quad (5)$$

2. *Adjusted* data (data yang telah disesuaikan adalah hasil pengurangan dari setiap data dengan rata-rata setiap data yang diperoleh degan persamaan sebagai berikut.

$$Adjusted\ data = x_{ij} - \bar{x}_j$$

$$X^{'} = Adjusted\ data$$

3. Menghitung matriks kovarian dihitung dengan menggunakan persamaan sebagai berikut.

$$c = \frac{1}{M} X^{'} X^{'T}$$

4. Menghitung nilai eigen dan vector eigen dari matriks kovarian dihitung dengan menggunakan persamaan karakteristik berikut ini. (7)

$$c - \lambda I = 0$$

$$(c - \lambda I)v = 0$$

5. Hitung nilai eigen yang terbesar yang berkorespondensi terhadap nilai vector eigen yang (8) dipilih menjadi *principal component*. Vektor eigen yyang disusun dari yang terbesar ke yang terkecil dipilih menjadi vector fitur.

$$v = (eig_1, eig_2, eig_3, ..., eig_n)$$

6. Mencari PC sebagai rata-rata

$$PC = X^{'} \times v$$

7. Langkah berikutnya untuk melakukan trans (9) data untuk menghasilkan PCA.

$$PCA = PC^T \times X^{'T} \quad (10)$$

III. METODE PENELITIAN (10)

A. SumberData

Data yang digunakan dalam penelitian ini merupakan data sekunder yang tersedia di website https:// www.kaggle.com/datasets dengan menggunakan data *hotel booking demand* yaitu data tentang perusahaan hotel resort dan hotel city. Diakses pada hari Senin, 29 Maret 2020 pukul 15.00 WIB.

B. Variabel Penelitian

Variabel penelitian yang digunakan dalam penelitian ini adalah sebagai berikut.

Tabel 1 Nama Variabel, Deskripsi dan Tipe Data

Variabel	Deskripsi	Tipe Data
Hotel	Nama hotel (resort dan city)	Kategorik
is_canceled	*Booking* tercancel atau tidak. Kode 1 : *cancel* Kode 0 : tidak *cancel*	Kategorik
lead_time	Selisih data dari *booking* sampai pelanggan *check in*	Numerik
arrival_date_year	Tahun kedatangan pelanggan	Numerik
arrival_date_month	Bulan kedatangan pelanggan	Numerik
arrival_date_week_ number	Minggu kedatanagn pelang-gan akumulasi 1 tahun	Numerik
arrival_date_day_of_ month	Tanggal kedatangan pelanggan	Numerik
stays_in_weekend_ nights	Jumlah weekend (sabtu/ minggu) pelanggan berada di hotel	Numerik
stays_in_week_ nights	Jumlah hari biasa (senin sampai jumat) pelanggan berada di hotel	Numerik
adults	Jumlah pelanggan dewasa	Numerik
children	Jumlah pelanggan anak-anak	Numerik
babies	Jumlah pelanggan bayi	Numerik
meal	kategori fasilitas yang dipilih	Kategorik
country	asal negara pelanggan	Kategorik
market_segment	asal kedatanagn pelanggan (*Travel agent*, atau *Tour Operators*)	Kategorik
distribution_channel	sumber *booking* hoteln. Melalui *travel agent* atau tour operators.	Kategorik
is_repeated_guest	kode untuk pengunjung yang pernah datang atau tidak. Kode 1 : pengunjung pernah datang Kode 0 : pengunjung belum pernah datang.	Kategorik
previous_cancellation	jumlah *booking* yang di *cancel*	Numerik
reserved_room_type	Jumlah *booking* yang tidak di *cancel*	Numerik
reserved_room_ typ	kode ruangan yang dipesan	Kategorik
assigned_room_type	kode ruangan yang bisa beda/hampir sama dengan *served room* type.	Kategorik
booking_changes	Jumlah perubahan selama *booking*	Numerik
deposit_type	deposit pelanggan (*no deposite, non refund, refunable*)	Kategorik
agent	ID travel	Kategorik
company	ID *company*	Kategorik
days_in_waiting_list	jumlah hari selama berada di *waiting list* (belum dikomfirmasi)	Numerik
customer_type	tipe elanggan (*contract, group, transient, transient-party*)	Kategorik
adr	jumlah semua transaksi penginapan dibagi dengan total menginap	Numerik
required_car_parking_ spaces	jumlah ruang parkir yang dibutuhkan pelanggan	Numerik
total_of_special_ requests	jumlah *special request*	Numerik
reservation_status	status pemesanan *canceled, check-out, no-show*)	Kategorik
reservation_status_ date	tanggal status terakhihr	Kategorik

C. Langkah Analisis

Langkah analisis yang dilakukan pada penelitian ini adalah sebagai berikut.

1. Mengunduh data dari website https://www.kaggle.com/datasets
2. Melakukan *preprocessing* data
3. Melakukan visualisasi data
4. Memberikan deskripsi hasil *preprocessing* dan visualisasi data

5. Melakukan *cluster* dan PCA
6. Memberikan kesimpulan dan saran

IV. HASIL DAN PEMBAHASAN

A. Hasil Preprocessing Data

Preprocessing yang telah dilakukan menghasilkan beberapa keadaan, diantaranya :

1. Deteksi *missing value.* Pada saat dideteksi *missing value* diketahui bahwa pada variabel *agent, country* dan *company* terdapar *missing value.*488 pada varibel *country*, 16340 pada variabel *agent*, dan 112593 pada variabel *company*. Langkah yang dilakukan dalam mengatasi *missing value* ini adalah mendrop atau menghapus data yang *missing*. Alasan dilakukan penghapusan yaitu karena ketiga variabel kedepannya tidak dilibatkan ke dalam visualisasi, sehingga dihapus juga tidak mempengaruhi hasil analisis yang dilakukan.
2. Memeriksa tipe data pada masing-masing variabel. Terdapat 16 variabel dengan tipe data int, 12 variabel dengan tipe *object* dan sisanya 5 variabel dengan tipe *float*.
3. Deteksi *outlier*. Digunakan *boxplot* untuk mendeteksi adanya *outlier* pada data. Variabel yang dicek diantaranya *lead_time, stays_in_week_night, agent, children, babies, booking change, adult*. Dari varibel-varibel yang dicoba, hanya variabel *adult* yang tidak terdapat outlier. Tidak dilakukan *treatment* apapun untuk mengatasi outlier pada data tersebut.

B. Statistika Deskriptif

Statistika deskriptif yang dihasilkan dari data tersebut ditampilkan pada tabel berikut :

Tabel 1 Statistika Deskriptif

Variable	*Mean*	Max	Min	Q1	Q2	Q3
Lead_time	40,52	364	0	12	27	36
Stay_in_weekend	1.57	9	0	0	2	2
Stay_in_week_night	4.63	21	0	2	4	6
Adult	1.41	3	1	1	1	2
Children	0,04	2	0	0	0	0
Baby	0	0	0	0	0	0

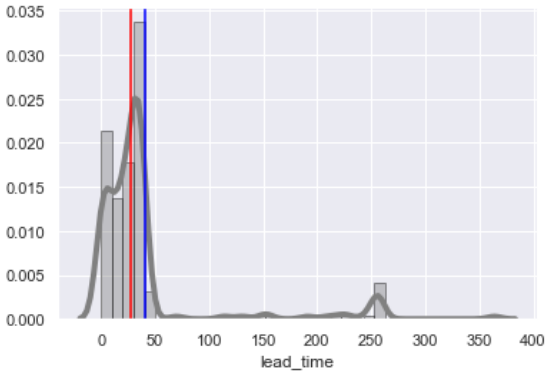
Berdasarkan tabel tersebut dapat diketahui bahwa rata-rata pelanggan yang memesan kedua hotel tersebut 41 hari sebelum kedatangannya. Tetapi ada juga beberapa pelanggan yang memesan dan langsung ditempati pada hari itu juga. Sebaliknya ada beberapa pelanggan yang memesan bahkan 1 tahun sebelum hari kedatangannya.

Pelanggan lebih banyak menginap pada hari-hari kerja, yaitu rata-rata sekitar 5 malam pada hari kerja, dna rata-rata 2 malam pada *weekend*.

Pelanggan yang menginap di hotel ini rata-rata hanya 2 orang, dengan minimal 1 orang dewasa dan maksimal 3 orang dewasa. Sangat jarang penggan membawa anak-anak, dan tidak ada sama sekali pelanggan yang membawa bayi.

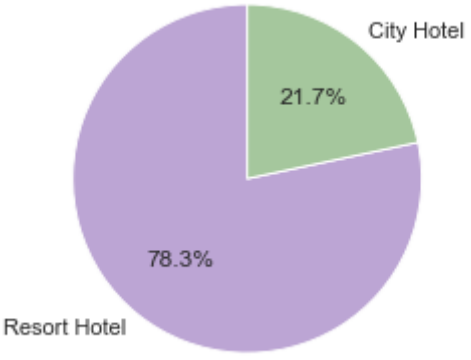
C. Visualisasi Data

Visualisasi data digunakan untuk mengetahui beberapa informasi yang dapat disampaikan dari data yang ada. Berikut adalah visualisasi yang dapat dileh pada data hotel tersebut.



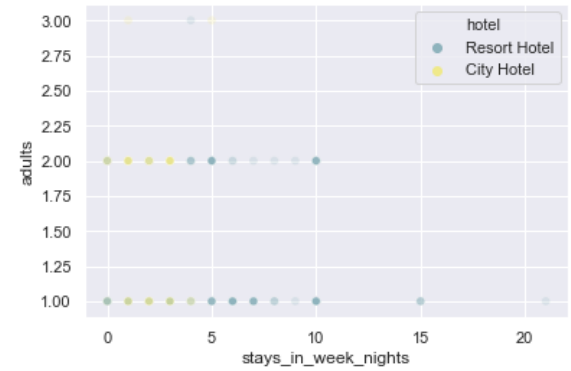
Gambar 1. Uji normalitas variabel lead_time

Gambar tersebut menunjukkan diagram dengan density plot dari variabel *lead_time*. Variabel *lead_time* adalah variabel yang menjelaskan tentang selisih waktu dari *booking* hingga kedatangan pelanggan. Informasi yang didapatkan dari gambar tersebut adalah varibel *lead_time* tidak berdistribusi normal, artinya selisih waktu dari *booking* hingga kedatangan pelanggan memiliki nilai yang sangat besar dan banyak juga yang memiliki nilai yang sangat kecil. Dari gambar dapat diketahui bahwa pelanggan paling banyak memiliki *lead_time* kurang dari 50 hari. Diatas dari 50 hari ada, tetapi hanya sedikit sedikit.



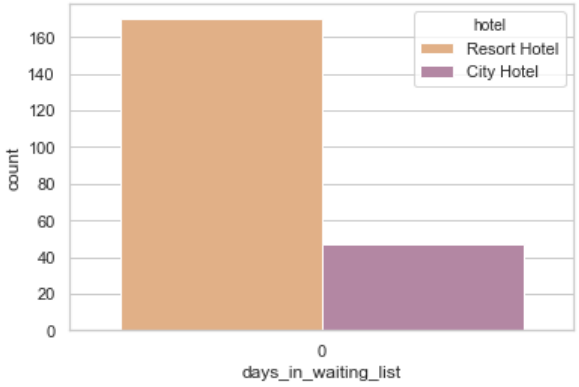
Gambar 2. Perbandingan pengguna hotel

Gambar tersebut menjelaskan proporsi banyak pelanggan untuk resort hotel dan city hotel. Resort hotel memiliki lebih banyak pelanggan daripada city hotel. Yitu dengan presentasi masing masing 78,3% untuk resort hotel dan 21,7% untuk city hotel.



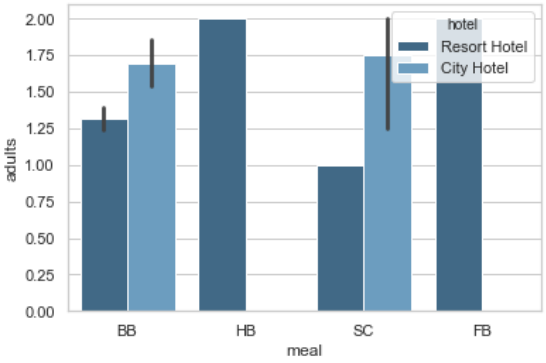
Gambar 3. Perbandingan lama orang menginap di 2 hotel

Gambar tersebut menjelaskan bahwa resort hotel dan city hotel memiliki lama pelanggan menginapnya hampir sama. Untuk pelanggan dewasa 1 orang tidak terdapat perbedaan untuk resort hotel dan city hotel, yaitu berkisar antara 0 sampai 15 malam. Selanjutnya untuk pelanggan dewasa 2 orang juga tidak terdapat perbedaan untuk resort hotel dan city hotel, yaitu berkisar dari 0 sampai 10 malam.



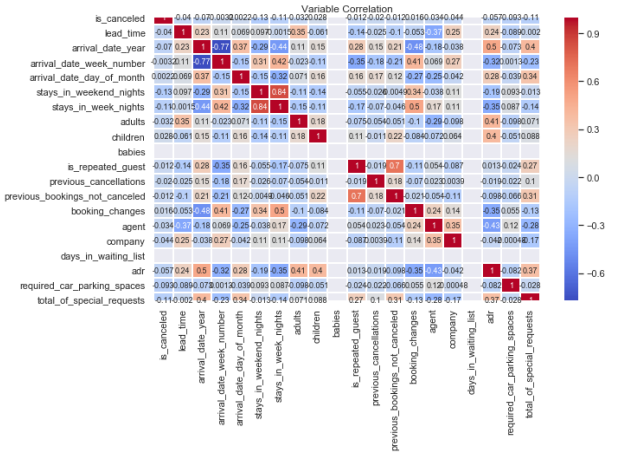
Gambar 4. Perbandingan lama waiting list untuk kedua hotel

Gambar tersebut menjelaskan perbedaan jumlah waiting list untuk resort hotel dan city hotel. Diketahui bahwa jumlah waiting list resort hotel jauh lebih banyak daripada city hotel. Sehingga dapat diduga jika resort hotel memiliki kamar sering full sehingga pelanggan masing harus berada di posisi *waiting_list*. Hal tersebut cocok dengan visualisasi sebelumnya tentang jumlah pengguna resort hotel dan city hotel. Sebelumnya dikatakan bahwa jumlah pengguna resort hotel sebesar 78.3% jauh lebih banyak daripada penggunak city hotel. Sehingga *match* dengan jumlah hari pada *waiting_list*.



Gambar 5. Perbandingan pemilihan meal pada 2 hotel

Diulas kembali variabel *meal* adalah pilihan fasilitas yang dipilih oleh pelanggan dari kedua hotel tersebut. Pada city hotel tidak terdapat pelanggan yang memilih fasilitas HB dan FB. Sebelumnya diketahui bahwa HB adalah *Half Board*, fasilitas sarapan dan satu macam makan lain, biasanya makan malam. Sedangkan FB adalah *Full Board*, fasilitas sarapan, makan siang dan makan malam. Jika di city hotel tidak terdapat seorangpun pelanggan yang memilih HB dan FB, sebaliknya di resort hotel, pelanggan paling banyak memilih 2 fasilitas tersebut, HB dan FB.

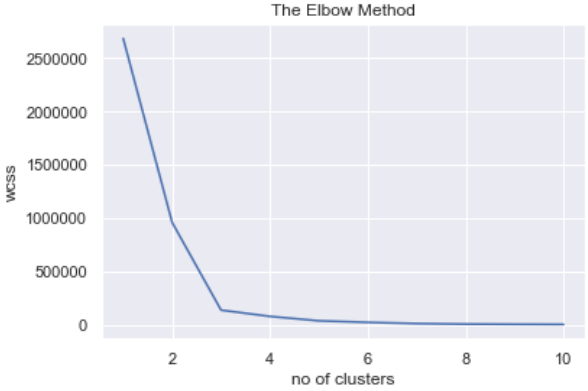


Gambar 6. Korelasi tiap variabel

Gambar tersebut menjelaskan korelais untuk tiap variabel dengan variabel lain. Korelasi tertinggi ada pada variabel *previous_booking_not_canceled* dengan variabel *not_repeated* yaitu koefisien korelasi sebesar 0.7. Artinya status pelanggan, pernah atau tidak pernah menginap, mempengaruhi pelanggan tersebut untuk mencancel atau tidak pesanannya.

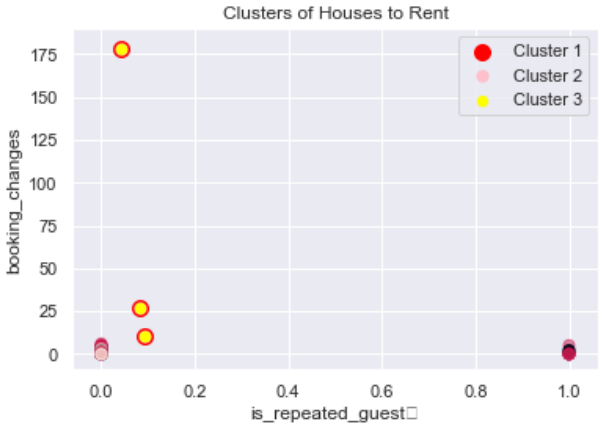
D. Clustering

Clustering dilakukan menggunakan metode *K-Means.* Variabel yang digunakan adalah variabel *is_repeated_guest* dan *booking_canceled*. Berikut ditampilkan grafik dari metode *elbow* untuk menentukan jumlah k yang digunakan.



Gambar 7. Grafik *Elbow*

Berdasarkan gambar tersebut dapat diketahui bahwa pada saat k=3 sudah menunjukkan hasil yang konvergen. Sehingga nilai k yang diambil adalah k=3. Berkut adalah hasil *cluster* untuk 2 variabel tersebut.



Gambar 8. Hasil *clustering*

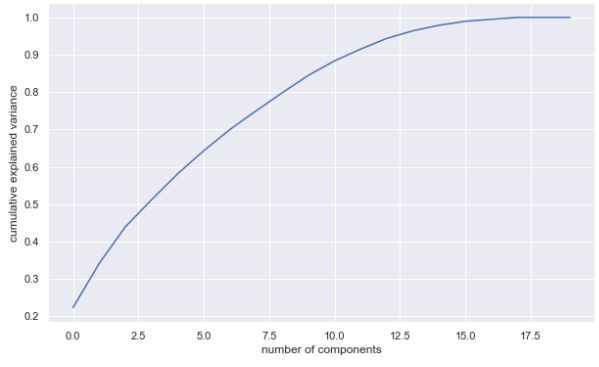
Sebelumnya dijelaskan bahwa variabel *is_repeated_guest* adalah variabel kategorik dengan kode 1 adalah pelanggan yang pernah datang ke hotel dan kode 0 adalah pelanggan yang belum pernah datang ke hotel. Sedangkan variabel *booking_changes* adalah varibel yang menyatakan jumlah pelanggan mengganti booking nya.

Berdasarkan gambar tersebut dapat diketahui bahwa *cluster* dibagi menjadi 3 kelompok. Untuk variabel *is_repeated*_order kode 0 dan kode 1 tersebar pada *cluster* 1, 2 dan 3. Sehingga dapat diduga bahwa variaabel tersebut tidak menjadi penentu pembagian *cluster*. Kelompok pertama dengan warna merah berisi pelanggan dengan perubahan jadwal sebanyak 15 sampai 25 kali, tetapi ada juga pelanggan dengan jumlah perubahan 175 kali. Kelompok kedua berwarna pink berisi pelanggan yang tidak pernah mengganti jadwal *booking* nya. Terakhir kelompok ketiga berwarna kuning yang berisi pelanggan dengan perubahan jadwal sebanyak 15 sampai 25 kali.

Evaluasi dari metode *cluster* mengguunakan *K-Means* yaitu tidak bisa membagi data *booking_changes* dengan baik. Karena pada beberapa kriteria masih menjadi 2 bagian kelompok yang berbeda.

F. Principal Componen Analysis

PCA digunakan untuk menyederhanakan variabel dengan cara mereduksi jumlah variabelnya. Berikut adalah gambar yang menjelaskan tentang pemilihan jumlah komponen untuk PCA.



Gambar 9. Pemilihan jumlah komponen untuk PCA

Berdasarkan gambar tersebut dapat diketahui bahwa pada komponen=10 sudah melihatkan grafik yang konvergen. Sehingga komponen yang dipilih adalah 10. Dapat disimpulkan bahwa PCA menggunakan 10 variabel. Artinya dari 32 varibel yang ada sebelumnya direduksi hanya menjadi 10 variabel saja.

Untuk memilih variabel yang masuk ke dalam 10 PC tersebut disajikan pada cuplikan tabel yang berbentuk gambar berikut ini.

	Feature	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
0	is_canceled	2.537634e-03	8.628512e-03	-0.211940	2.516353e-01	0.347331	3.728845e-01	9.492210e-02	-2.075983e-01	7.132304e-01
1	lead_time	6.421335e-02	-4.251889e-01	0.223004	-2.477354e-01	-0.017309	2.917924e-01	-3.605941e-01	-1.343655e-01	8.527255e-02
2	arrival_date_year	4.017773e-01	3.755854e-02	0.039186	-2.549617e-01	-0.204444	1.405653e-02	-2.060080e-01	6.917473e-02	2.277040e-01
3	arrival_date_week_number	-3.481058e-01	-2.072657e-01	0.096637	3.482595e-02	0.220979	1.087191e-01	2.206891e-01	-2.968901e-01	-3.274521e-01
4	arrival_date_day_of_month	2.586195e-01	-1.534375e-02	0.115552	-8.483333e-02	-0.066313	2.061509e-01	4.508042e-01	-4.527336e-01	-1.475130e-01
5	stays_in_weekend_nights	-2.660736e-01	-2.461210e-02	0.517617	-1.185914e-02	-0.180837	-6.061219e-03	-1.187435e-02	6.877103e-02	2.713318e-01
6	stays_in_week_nights	-3.585927e-01	1.166760e-02	0.408013	-5.230638e-03	-0.080274	-6.341085e-02	8.091337e-03	1.743212e-01	2.307683e-01
7	adults	1.475149e-01	-3.781816e-01	0.121090	8.831790e-02	0.232818	-1.200788e-01	-1.035120e-01	2.393954e-01	-1.366052e-01
8	children	1.594283e-01	-2.091923e-02	0.102555	-1.603461e-01	0.526868	-3.656163e-01	3.910603e-01	2.115925e-01	1.558358e-01
9	babies	-6.776264e-21	6.938894e-18	0.000000	-1.665335e-16	0.000000	-2.775558e-17	-1.387779e-17	-0.000000e+00	-1.387779e-17
10	is_repeated_guest	1.861498e-01	4.460884e-01	0.229621	-1.798587e-02	0.229544	-1.327557e-02	-3.033637e-01	-1.742568e-01	-4.628229e-02
11	previous_cancellations	8.155440e-02	1.365052e-01	0.036091	-1.867002e-01	-0.268673	4.496248e-01	4.666838e-01	4.739698e-01	2.608045e-02

Gambar 9. Pemilihan variabel yang masuk PC

Berdasarkan gambar tersebut dapat diketahui bahwa PC 1 sampai PC 10 memiliki anggota sebagai berikut :

Tabel 2 Anggota tiap PC

PC	Variabel
1	*arrival_date_year*, *previous_cancellations*
2	*is_canceled, required_car_parking_spaces*
3	*company*
4	*arrival_date_day_of_month*
5	*total_of_special_requests*
6	*days_in_waiting_list*
7	*stays_in_week_nights, children*
8	*stays_in_weekend_nights, adr*
9	*lead_time, is_repeated_guest, previous_bookings_not_canceled*
10	*arrival_date_week_number, adults, babies, booking_changes, agent*

V. KESIMPULAN DAN SARAN

A. Kesimpulan

Berdasarkan hasil visualisasi data pada bab sebelumnya didapatkan kesimpulan sebagai berikut :

1. Ada 3 variabel yang terdapat *missing value* didalamnya. *Missing value* diatasi dengan cara menghapus datanya, karena variabel yang terdapat *missing value* tersebut adalah variabel kategorik dan kedepannya variabel tersebut tidak digunakan di dalam analisis
2. Resort hotel lebih banyak digemari daripada city hotel. Selanjutnya diperjelas dengan lama *waiting list* pada resort hotel lebih lama daripada city hotel. Pada city hotel tidak ada pelanggan yang memilih fasilitas HB dan FB. Sebaliknya di resort hotel justru HB dan FB adalah fasilitas yang paling banyak dipilih.
3. Hasil dari *cluster* dengan metode *K-Means* tidak begitu baik karena ada beberapa kriteria terdapat pada 2 kelompok yang berbeda. Sehingga metode *K-Means* dinilai tidak bisa membagi kelompok dengan baik pada variabel tersebut
4. Hasil dari analisis PCA yaitu mereduksi variabel menjadi 10 PC, sedangkan sebelumnya terdapat 32 variabel.

B. Saran

Saran yang diberikan penulis untuk penelitian selanjutnya yaitu, sebaiknya tidak menggunakan *K-Means* jika akan melakukan pengelompokan variabel.

DAFTAR PUSTAKA

[1] Hansen, C. D. & Johnson, C. R. 2005. The Visualization handbook. USA: Elsevier Inc.

[2] Somantri, Ating dan Sambas Ali Muhidin. 2006. Aplikasi statistika dalam Penelitian. Bandung : Pustaka Ceria

[3] Riduwan . 2010. Dasar-dasar Statistika. Bandung : Alfabeta.

[4] Hasan, M. Iqbal. 2011. Pokok – Pokok Materi Statistika (Statistik Deskriptif). Jakarta : PT Bumi Aksara.

[5] Furqon. 1999. Statistika Terapan Untuk Penelitian. AFABETA: Bandung.

[6] Website : http://www.visualcomplexity.com/ diakses pada tanggal 02 Maret 2020

[7] Website : : http://datavisualization.ch/showcases/ diakses pada tanggal 02 Maret 2020

[8] Santosa, P. B., & Hamdani, M. (2007). *Statistika* Edisi Kedua. Jakarta: Remaja Rosdakarya