*Preprocessing* dan Visualisasi pada Brazilian *Houses to Rent* *Dataset* 2020

Nadya Yuniar Desi Prameswari1), Santi Wulan P2), Kartika Fithriasari3) dan Irhamah4)

Departemen Statistika, Fakultas Sains dan Analitika Data

Institut Teknologi Sepuluh Nopember (ITS)

Jl. Arief Rahman Hakim, Surabaya 60111 Indonesia

*e-mail*: [yuniarnadya4@gmail.com](mailto:yuniarnadya4@gmail.com)1), [santi\_wp@statistika.its.ac.id](mailto:santi_wp@statistika.its.ac.id)2), [kartika\_f@statistika.its.ac.id](mailto:kartika_f@statistika.its.ac.id)3), dan irhamah@statistika.its.ac.id4)

***Abstrak— Persewaan rumah menjadi hal yang menjamur di Brazil. Pemilik properti tidak hanya menyediakan sewa rumah untuk jangka panjang saja, tetapi juga untuk jangka yang pendek, misal pada saat liburan musim panas bagi turis-turis yang berdatangan. Persewaan prpoerti di Brazil dapat dengan mudah ditemukan melalui departemen agen persewaan, hingga mulalui internet. Beberapa website sangat berguna ketika digunakan untuk mencari akomodasi di Brazil, kita dapat terfokus pada area spesifik yang ingin kita tempati. dan juga memungkinkan untuk mencari yang memiliki harga sewa murah, atau berdasarkan tipe dari properti hingga maksimum harga yang ditawarkan. Tersedia dalam berbagai macam kriteria, mulai dengan area properti yang luas, kamar tidur yang banyak hingga jumlah lahan parkir yang tersedia. Harga sewa rumah di Brazil jauh lebih murah dibandingkan dengan harga untuk membeli rumah. Harga untuk sewa rumah juga dapat dinegoisasi dengan pemilik properti. Sehingga pada penelitian kali ini dilakukan untuk mengetahui karakteristik data Brazilian Houses to Rent, mendapatkan insight dari data sehingga diperoleh informasi yang bermanfaat. Pada penelitian ini dilakukan pre-processing data untuk mengetahui karakteristik data, analisis clustering untuk mengetahui ada berapa kelompok dalam data tersebut dan future extraction menggunakan Principal Component Analysis (PCA) untuk menjelaskan minimal 80% varians data menggunakan sotfware Python.***

***Kata Kunci— Data Mining, Preprocessing, Analisis Eksplorasi Data, Clustering, PCA***

1. PENDAHULUAN

K

ebutuhan hierarki yang akan dipenuhi manusia setelah kebutuhan jasmani yang terdiri dari sandang, pangan dan kesehatan ialah rumah. Kebutuhan akan rumah merupakan salah satu motivasi untuk mengembangkan kehidupan yang lebih baik dan layak. dengan memiliki rumah, walaupan tidak besar, pemilik dapat memnguasai ruang yang dapat diatur sesuka hati dan seleranya. Ruang tersebut akan memberikan respon terhadapnya, artinya dapat tercipta suasana timbal balik dan saling menghidpkan. dengan demikian, suasana rumah telah tercipta. bukan lagi hanya untuk menghindari hujan dan panas, tetapi juga memberikan ketenangan, kenyamanan hingga kenangan akan segala peristiwa dalam kehidupannya.

Brazil merupakan negara dengan penduduknya yang padat sehingga kebutuhan akan rumah menjadi hal yang sangat dicari. Penduduk Brazil lebih suka menyewa rumah dibandingkan dengan membeli rumah untuk dirinya. Ketertarikan penduduk Brazil untuk lebih memilih menyewa rumah yang akan ditinggali ini dibuktikan dengan menjamurnya oersewaan-persewaan rumah, utamanya di kota-kota besar seperti Rio de Janeiro, Campinas dan lain sebagainya. Hal ini dikarenakan biaya untuk menyewa rumah jauh lebih murah dibandingkan untuk membeli rumah. Di Brazil, plang “untuk disewakan” dipasang di properti-properti, entah digantung dari jendela apartement atau digantung di gerbang depan rumah. Jika properti tersedia untuk jangka waktu yang pendek, misal liburan musim panas, plang akan bertuliskan “Silahkan kontak pemlik secara langsung”. Hal ini dapat menjadi keuntungan karna harga sewa mungkin bisa dinegosiasi, tetapi tidak dapat pelayanan atau perlindungan yang disediakan oleh agen persewaan.

Berdasarkan latar belakang permasalahan di atas, dan mengacu pada beberapa hasil penelitian yang telah dilakukan maka penelitian ini bertujuan untuk mengetahui karakteristik data *Brazilian Houses to Rent,*  mendapatkan *insight* dari data sehingga diperoleh informasi yang bermanfaat. Pada penelitian ini dilakukan *pre-processing* data untuk mengetahui karakteristik data, analisis *clustering* untuk mengetahui ada berapa kelompok dalam data tersebut dan *future extraction* menggunakan *Principal Component Analysis* (PCA) untuk menjelaskan minimal 80% varians data. Penelitian ini diharapkan dapat memberikan informasi kepada agen rumah. Hasil penelitian ini nantinya juga diharapkan bisa menjadi referensi penelitian lanjutan untuk dapat menganalisis data ini agar menjadi lebih baik.

1. TINJAUAN PUSTAKA
2. *Statistika Deskriptif*

Statistika deskriptif adalah metode-metode yang berkaitan dengan pengumpulan dan penyajian suatu gugus data sehingga memberikan informasi yang berguna [1]. Perlu kiranya dimengerti bahwa statistika deskriptif memberikan informasi hanya mengenai data yang dipunyai dan sama sekali tidak menarik kesimpulan apapun tentang gugus data induknya yang lebih besar.

1. Rata-rata

Rata-rata adalah jumlah semua data yang ada dibagi dengan banyaknya data [1]. Data tunggal memiliki rumus :



(1)

Keterangan :

 : rata-rata

 : sukuk e-n

n : banyaknya data

1. Simpangan Baku

Simpangan baku menunjukkan penarikan akar dari rata-rata kuadrat jarak suatu data terhadap rata-ratanya [1]. Simpangan baku memiliki rumus :



(2)

Keterangan :

S : simpangan baku

n : banyaknya data

 : data ke-i

 : rata-rata

1. Median

Median adalah nilai tengah dari data yang telah disusun berurutan mulai dari data yang terkecil sampai dengan yang terbesar [1]. Secara matematis median dilambangkan dengan *Me* yang dapat dihitung dengan rumus :



(3)

Keterangan :

*Me* : median

n : banyaknya data

x : nilai data

1. Modus

Modus segugus pengamatan adalah nilai yang terjadi paling sering atau yang mempunyai frekuensi paling tinggi. Modus tidak selalu ada. Hal ini terjadi bila semua pengamatan mempunyai frekuensi terjadi yang sama. Untuk data tertentu, mungkin saja terdapat beberapa nilai dengan frekuensi tertinggi, dan dalam hal demikian kita mempunyai lebih dari satu modus [1].

1. Minimum

Nilai minimum merupakan nilai yang paling kecil dalam suatu kelompok pengamatan.

1. Maksimum

Nilai maksimum merupakan nilai yang paling besar dalam suatu kelompok pengamatan.

1. *Preprocessing Data*

*Preprocessing* data merupakan tahapan pertama dan merupakan tahapan yang penting dalam *data mining* atau *data analysis* [2]. Pada umunya data pada dunia nyata yang masih mentah atau biasa disebut data primer memiliki kekurangan diantaranta tidak lengkap, banyak *noise*, dan juga tidak konsisten [3]. Oleh karenanya tahapan ini sangat penting untuk memastikan data sumber diolah sehingga menghasilkan *dataset* yang siap dipakai pada tahapan selanjutnya. Ada tiga tahapan yaitu *data cleaning, data transformation,* dan *feature selection*.

1. *Data cleaning*

*Data cleaning* yaitu tahapan awal pada *preprocessing* data merupakan tahapan yang berusaha memperbaiki kualitas agar menjadi lebih baik seperti mengisi nilai yang hilang *(missing values)*, menghaluskan data yang memiliki *noise*, memperbaiki data yang tidak konsisten hingga mendeteksi adanya *outlier.* *Missing value, noisy* data, dan *outlier* ditemukan pada *dataset* berukuran besar. *Missing value* biasanya dipengaruhi oleh tidak adanya jawaban yang sesuai untuk variabel yang diobservasi. *Missing value* dapat diatasi dengan mengisi data yang hilang tersebut dengan *mean* atau median dari variabel yang mengandung *missing value* tersebut. *Mean* dapat digunakan untuk mengisi *missing value* ketika data berdistribusi normal sedangkan median dapat digunakan untuk mengisi *missing value* ketika data tidak berdistribusi normal [4].

1. *Data transformation*

*Data transformation* dibutuhkan dalam implementasi *data mining* khususnya pada saat *preprocessing* data. Pada tahapan ini data diubah atau dikonsolidasikan sehingga proses penambangan yang dihasilkan dapat lebih efisien dan pola yang ditemukan dapat lebih mudah dipahami.

1. *Feature selection*

*Feature selection* adalah suatu proses yang mencoba untuk menemukan subhimpunan dari himpunan fitur yang tersedia untuk meningkatkan aplikasi dari suatu algoritma pembelajaran [5]. *Feature selection* digunakan dibanyak area aplikasi sebagai alat untuk menghilangkan fitur yang tidak relevan dan atau fitur berlebihan. Sebuah fitur dikatakan tidak relevan jika memberikan sedikit informasi, sedangkan sebuah fitur dikatakan berlebihan jika informasi yang diberikan adalah informasi yang terkandung dalam fitur lain (tidak memberikan informasi baru).

1. Visualisasi Data

Visualisasi data merupakan cara dalam menyajikan data yang detail menjadi informasi yang mudah diterima dengan menggunakan informasi visual yang bisa dipahami secara universal [6]. Bentuk visualisasi data di antaranya yaitu *simple text*, tabel, dan grafik [7]. Teks atau *simple text* digunakan bila informasi yang ingin disampaikan berupa sebuah angka atau dua angka. Tabel digunakan saat pengguna akan melihat beberapa ukuran yang berbeda. Diagram digunakan jika ingin memetakan dimensi menjadi visual. Berikut merupakan beberapa visualisasi data yang sering digunakan.

1. *Pie Chart*

*Pie chart* adalah suatu penyajian data dalam bentuk lingkaran dan biasanya jumlah data dinyatakan dalam persentase. Lingkaran tersebut dibagi menjadi sektor-sektor. Banyaknya sektor tergantung pada banyaknya kelompok data. Besarnya sektor tergantung persentase dari nilai sebuah data terhadap keseluruhan nilai dataanita [7].

1. *Bar Chart*

*Bar Chart* umumnya digunakan untuk menggambarkan perkembangan nilai suatu objek penelitian dalam kurun waktu tertentu. Diagram batang menunjukkan keterangan-keterangan dengan batang-batang tegak atau mendatar dan sama lebar dengan batang-batang terpisah [8].

1. *Correlogram*

*Correlogram* adalah sebuah grafik yang menunjukkan korelasi antara dua variabel.

1. *Boxplot*

*Boxplot* pada dasarnya adalah plot 1 dimensi, kadang kala dikenal sebagai *box-and-whiskers plot*. Pada *boxplot*, dapat diketahui nilai kuartil, median data, *outlier*, nilai ekstrim, persentil [9].

1. *Scatterplot*

*Scatterplot* adalah suatu metode grafis untuk menggambarkan hubungan antara dua variabel dengan cara memplot pasangan nilai *x* dan *y* untuk masing-masing nilai/data pada bentuk 2 dimensi [10].

1. *Clustering*

Analisis *cluster* adalah teknik multivariate yang mempunyai tujuan utama untuk mengelompokkan objek-objek berdasarkan karakteristik yang dimiliki. Analisis *cluster* mengklasifikasi objek sehingga setiap objek yang memiliki sifat mirip (paling dekat kesamaannya) akan mengelompok kedalam satu *cluster* (kelompok) yang sama [11].

1. *Principal Component Analysis (PCA)*

*Principal Compenent Analysis* (analisa komponen utama) adalah salah satu fitur ekstraksi (reduksi) variabel yang banyak digunakan. Bisa dikatakan *principal component analysis* merupakan analisa tertua dan paling terkenal dari teknik statistika multivariate [12].

Perhitungan dari *principal component analysis* didasarkan pada perhitungan nilai eigen dan vektor eigen yang menyatakan penyebaran data dari suatu *dataset.* Tujuan dari analisa PCA adalah untuk mereduksi variabel yang ada menjadi lebih sedikit tanpa harus kehilangan informasi yang termuat dalam data asli. Dengan menggunakan PCA, variabel yang tadinya sebanyak n variabel akan direduksi menjadi k variabel baru *(principal component)* denganjumlah k lebih sedikit dari n dan dengan hanya menggunakan k *principal component* akan menghasilkan nilai yang sama dengan menggunakan n variabel [13].

1. METODOLOGI PENELITIAN
2. *Sumber Data*

Data yang digunakan dalam penelitian ini merupakan data sekunder yang tersedia di *website* https:// www.kaggle.com/datasets dengan menggunakan data *brazilian\_houses\_to\_rent* yaitu data tentang informasi penyewaan rumah mulai dari luas area, jumlah kamar hingga harga sewa di beberapa kota di Brazil.Diakses pada hari Senin, 27 Maret 2020 pukul 18.22 WIB.

1. *Variabel Penelitian*

Variabel penelitian yang digunakan tercantum dalam Tabel 1.

**Tabel 1.** Variabel Penelitian

|  |  |  |
| --- | --- | --- |
| **Variabel** | **Keterangan** | **Jenis Data** |
| *City* | Lokasi area properti | Kategorik |
| *Area* | Luas area properti | Numerik |
| *Rooms* | Jumlah kamar tidur | Numerik |
| *Bathroom* | Jumlah kamar mandi | Numerik |
| *Parking Spaces* | Jumlah lahan parkir | Numerik |
| *Floor* | Lantai | Numerik |
| *Animal* | Apakah menerima hewan atau tidak | Kategorik |
| *Furniture* | Apakah terdapat *furniture* atau tidak | Kategorik |
| *Hoa (R$)* | Pajak pemilik properti | Numerik |
| *Rent Amount (R$)* | Biaya sewa | Numerik |
| *Property Tax (R$)* | Pajak properti | Numerik |
| *Fire Insurance (R$)* | Asuransi kebakaran | Numerik |
| Total (R$) | Total | Numerik |

1. *Struktur Data*

Adapun struktur data yang akan digunakan pada penelitian ini sebagai berikut.

**Tabel 2**. Struktur Data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Data ke-*i*** | ***X1*** | ***X2*** |  | ***X13*** |
| 1 |  |  |  |  |
| 2 |  |  |  |  |
|  |  |  |  |  |
| 10.692 |  |  |  |  |

## Keterangan :

## X1 : data variabel ke-1 *(City)*

## X2 : data variabel ke-2 *(Area)*

*X13*: data variabel ke-13 (*Total (R$))*

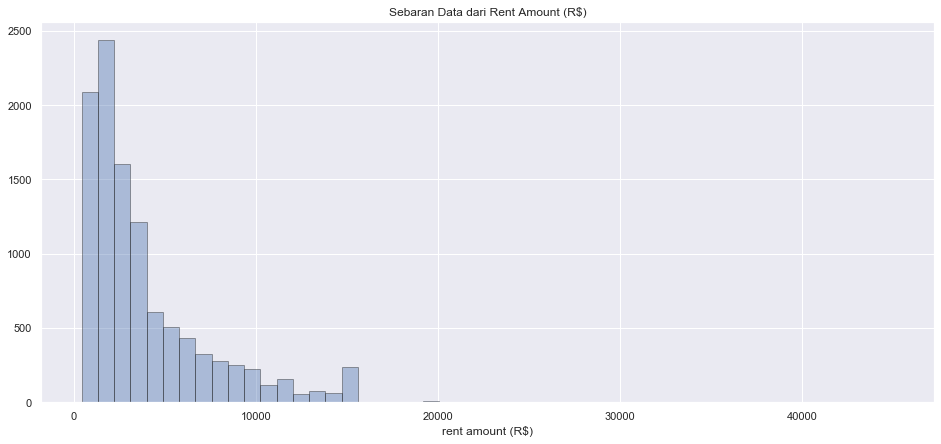
1. *Langkah Analisis*

Langkah-langkah penelitian yang telah dilakukan berdasarkan dengan tujuan penelitian ini adalah sebagai berikut.

1. Mengunduh data melalui *website* <https://www.kaggle.com/datasets>
2. Melakukan *preprocessing* data yaitu mendeteksi serta mengatasi data *outlier*, *noisy*¸ dan *missing value*.
3. Melakukan eksplorasi data menggunakan analisis statistika deskriptif.
4. Melakukan visualisasi data pada tiap variabel dari data *brazilian\_houses\_to\_rent.*
5. Melakukan analisis *clustering* pada data.
6. Melakukan analisis PCA pada data.
7. Melakukan interpretasi.
8. Melakukan penarikan kesimpulan dan saran untuk melakukan metode lanjutan.
9. ANALISIS DAN PEMBAHASAN
10. *Preprocessing Data*

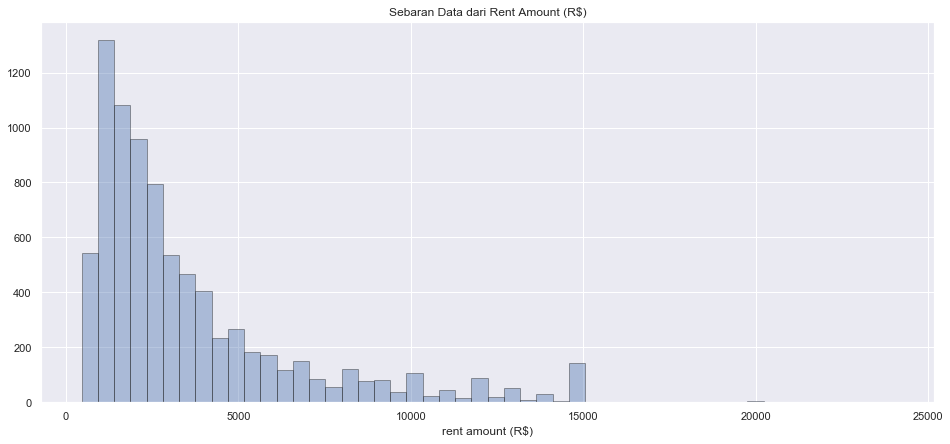
*Preprocessing* yang telah dilakukan menghasilkan beberapa keadaan, diantaranya:

* + - 1. Memeriksa apakah terdapat *missing values* atau tidak. Didapati, ada sebanyak 2461 *missing values* pada variabel *floor.* Selanjutnya menentukan langkah yang digunakan untuk mengatasi *missing values,* yaitu dengan mengetahui data berdistribusi normal atau tidak.



**Gambar 1.** Distribusi dari *Rent Amount (R$)*

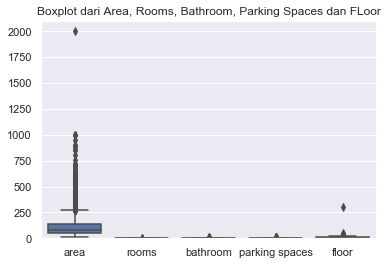
Dari Gambar 1, dapat diketahui dari sebaran data *Rent Amount (R$),* data berdistribusi normal tetapi *skewness* kiri. Maka *missing value* perlu diatasi.



**Gambar 2.** Distribusi dari *Rent Amount (R$)* Setelah Penghapusan Data

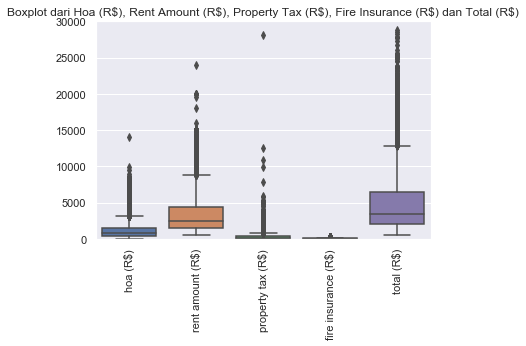
Berdasarkan Gambar 2, setelah data dengan *missing values* dihapus, pada distribusi tidak terdapat perbedaan yang signifikan. Data tetap berdistribusi normal dengan *skewness* kiri, sehingga penghapusan dapat dilakukan.

* + - 1. Mendeteksi apakah terdapat *outlier* pada data atau tidak pada data numerik.



**Gambar 3.** *Boxplot* dari *Area, Rooms, Bathroom, Parking Spaces* dan *Floor*

Dapat dilihat bahwa berdasarkan Gambar 3, terdapat banyak sekali *outlier* dalamtiap variabel, *outlier* paling banyak terdapat pada variabel area.



**Gambar 4.** *Boxplot* dari *Hoa (R$), Rent Amount (R$), Property Tax (R$), Fire Insurance (R$)* dan *Total (R$)*

Sedangkan berdasarkan Gambar 4, terdapat banyak sekali *outlier* dalamtiap variabel, *outlier* paling banyak terdapat pada variabel total (R$). Data *outlier* tidak diatasi karena data *outlier* merupakan data yang cukup penting.

1. *Statistika Deskriptif*

Untuk mengetahui karakteristik data rumah yang disewakan di Brazil, maka dapat dilakukan analisis menggunakan statistika deskriptif yang dapat dilihat pada Tabel 3 berikut.

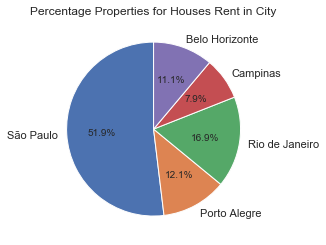
**Tabel 3.** Statistika Deskriptif

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | ***Mean*** | ***Std*** | ***Min*** | ***Med*** | ***Max*** |
| *Area* | 124.58 | 601.03 | 11 | 80 | 46335 |
| *Rooms* | 2.34 | 1.03 | 1 | 2 | 7 |
| *Bathroom* | 2.06 | 1.26 | 1 | 2 | 9 |
| *Parking Spaces* | 1.36 | 1.24 | 0 | 1 | 8 |
| *Floor* | 6.58 | 6.15 | 1 | 3 | 301 |
| *Hoa (R$)*  *Rent Amount (R$)*  *Property Tax (R$)*  *Fire Insurance (R$)* | 1484.77  3582.28  344.74  46.46 | 17755.40  3138.23  3512.01  40.15 | 0  460  0  3 | 750  2500  112  32 | 1117000  24000  313700  305 |
| *Total (R$)* | 5458.47 | 18610.53 | 507 | 3456 | 1120000 |

Pada Tabel 3 menunjukkan nilai rata-rata, simpangan baku, minimum, maksimum, dan nilai tengah dari beberapa variabel kontinu yang digunakan diantaranya luas area properti, jumlah kamar tidur, jumlah kamar mandi, lahan parkir, lantai, pajak pemilik properti, harga sewa, pajak properti, asuransi kebakaran hingga total yang perlu dibayarkan untuk menyewa properti.

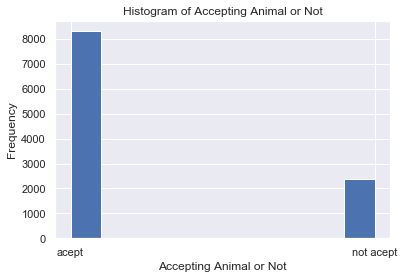
1. *Eksplorasi dan Visualisasi Data*

Analisis eksplorasi data yang dilakukan dalam praktikum kali ini adalah karakteristik data numerik serta kategorik dan juga visualisasi data dalam bentuk grafik.



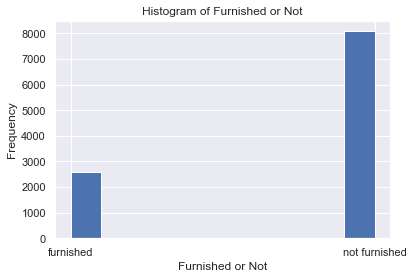
**Gambar 5.** *Pie Chart*  dari *City*

Dari Gambar 5, dapat dilihat bahwa lokasi kota yang terdapat banyak persewaan properti adalah di kota Sao Paulo sebanyak 51.9% disusul dengan Rio de Janeiro dengan presentase 16.9% dan yang paling sedikit berada di kota Campinas sebesar 7.9%.



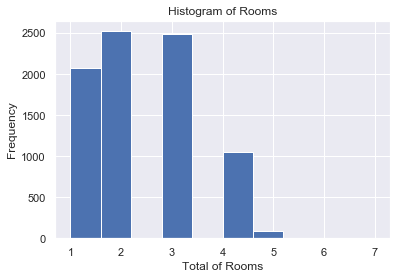
**Gambar 6.** Histogram dari *Accepting Animal or Not*

Dari Gambar 6, dapat dilihat bahwa dari banyaknya persewaan properti memperbolehkan untuk membawa binatang peliharaan dalam area persewaan property ada sebanyak lebih dari 8.000 properti.



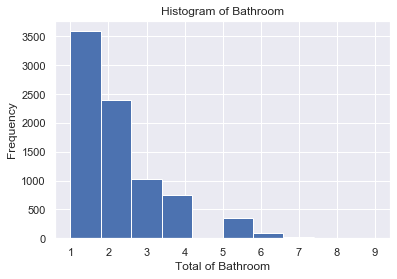
**Gambar 7.** Histogram dari *Furnished or Not*

Dari Gambar 7, dapat dilihat bahwa dari banyaknya persewaan properti belum menyediakan *furniture* dalam area properti yang disewakan ada sebanyak lebih dari 8.000 properti sedangkan yang belum menyediakan *furnish* terdapat kurang dari 3.000 properti.



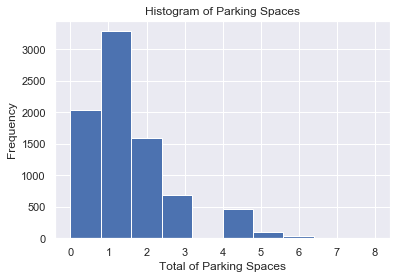
**Gambar 8.** Histogram dari *Rooms*

Dari histogram di atas, properti yang disewakan paling banyak memiliki 2 kamar tidur dengan frekuensi lebih dari 2500 properti, dan 3 kamar tidur sebanyak 2500 properti dan yang paling sedikit ialah properti dengan jumlah kamar tidur sebanyak 5, hanya terdapat kurang dari 500 properti.



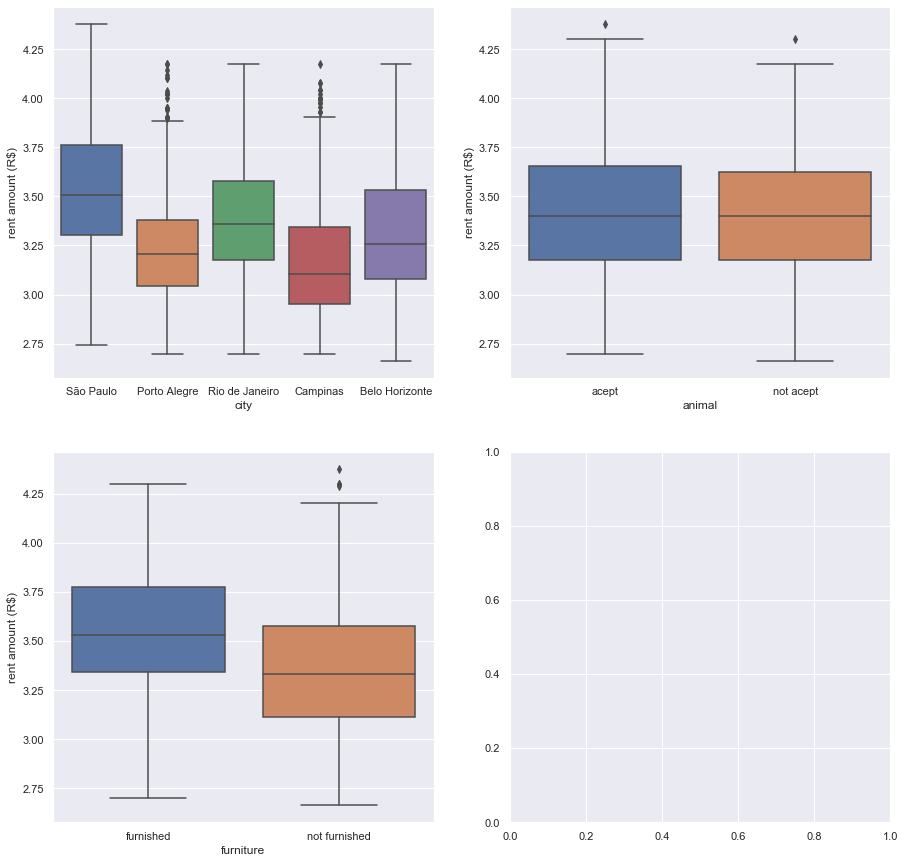
**Gambar 9.** Histogram dari *Bathroom*

Dari histogram pada Gambar 9, properti yang disewakan paling banyak memiliki 1 kamar mandi dengan frekuensi lebih dari 3500 properti, dan 2 kamar mandi sebanyak lebih dari 2000 properti dan yang paling sedikit ialah properti dengan jumlah kamar mandi sebanyak 6, hanya terdapat kurang dari 500 properti.



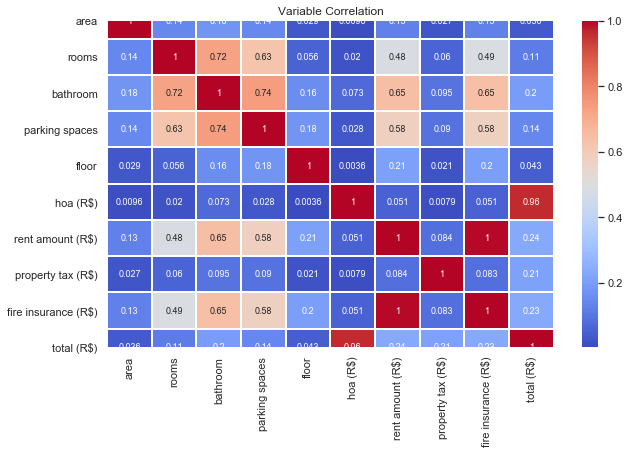
**Gambar 10.** Histogram dari *Parking Spaces*

Dari histogram di atas, properti yang disewakan paling banyak memiliki 1 lahan parkir dengan frekuensi lebih dari 3000 properti, dan tanpa lahan parkir sebanyak lebih dari 2000 properti dan yang paling sedikit ialah properti dengan jumlah lahan parkir sebanyak 6, hanya terdapat kurang dari 500 properti.



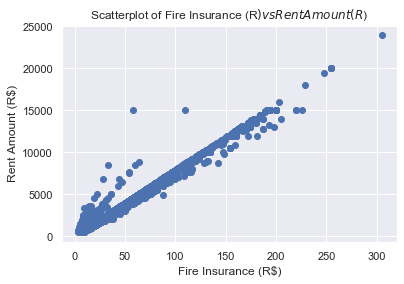
**Gambar 11.** *Boxplot* Variabel Kategorik terhadap Rent Amount (R$)

Pada Gambar 11 di atas, antara variabel *city*  dengan *rent amount (R$)* harga rata-rata sewa rumah tertinggi berada di Kota Sao Paulo dan yang paling murah berada di Kota Campinas. Sedangkan di Kota Porto Alegre dan Campinas juga terdapat beberapa harga persewaan rumah yang berada di luar batas. Pada hubungan variabel *animal* dan *rent amount (R$)* pada masing-masing kategori terdapat 1 *outlier* dan rata-rata harga sewa rumah yang menerima hewan peliharaan dapat masuk ke property memiliki rata-rata harga sewa yang sedikit lebih mahal. dan yang terakhir pada variabel *furniture* dan *rent amount (R$)* rata-rata harga sewa properti antara yang telah di*furnish* dengan yang belum di*furnish* memiliki rata-rata harga sewa yang berbeda, property yang telah di*furnish* memiliki rata-rata harga sewa yang lebih mahal dan pada kategori tidak di*furnish* memiliki 2 *outlier* pada datanya.

**

**Gambar 12.** Korelogram untuk setiap variabel

Berdasarkan grafik korelogram di atas, dapat dilihat bahwa tidak ada variabel yang tidak memiliki hubungan sama sekali, terbukti dari tidak ada nilai korelasi sebesar 0.Namun ada hubungan positif yang renggang antara harga sewa dengan lahan parkir dibuktikkan dari nilai korelasinya sebesar 0,58. Juga ada hubungan positif yang renggang antara variabel *fire insurance* dengan kamar tidur.

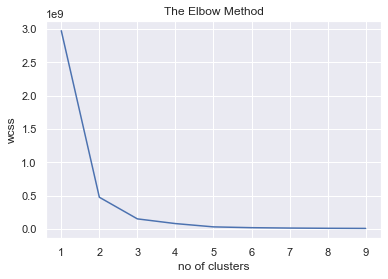


**Gambar 13.** *Scatterplot Fire Insurance (R$)* dan *Rent Amount (R$)*

Berdasarkan visualisasi dari *scatterplot* di atas, dapat diketahui bahwa antara variabel asuransi kebakaran dengan variabel harga sewa memiliki hubungan yang positif, hal ini dapat dibuktikan dari semakin mahal nominal asuransi akibat kebakaran akan dibayarkan, maka nominal harga sewa juga akan semakin meningkat.

1. *Clustering*

Tahap selanjutnya adalah pengelompokan data menggunakan metode *Clustering.* Variabel yang dipilih pada *clustering* ini yaitu variabel *area* dan *floor.* Hasil dari *clustering* menggunakan metode *K-Means* pada observasi ditunjukkan dengan grafik *Elbow Method* di bawah ini.

**

**Gambar 14.** Jumlah *Clusters* Berdasarkan *Elbow Method*

Selisih dari *k* yang mulai konstan atau stabil yaitu mulai dari 4 dan seterusnya. Sehingga, jumlah *k*  yang diambil sebagai *cluster* yaitu sebanyak 3.

* + - 1. *k* = 1, sempit

Luas area yang kecil dan jumlah tingkat lantai yang tidak banyak. Akan mengakibatkan jumlah kamar yang sedikit atau luas kamar yang kecil.

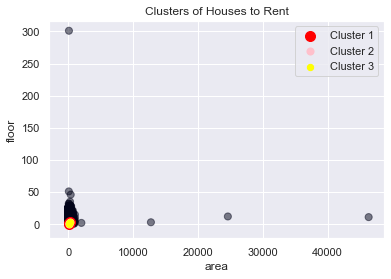
* + - 1. *k* = 2, sedang

Luas area yang cukup besar dan diimbangi dengan tidak hanya 1 lantai. Dapat digunakan untuk membangun kamar dengan luas yang lebih besar, atau dapat membangun kamar dengan ukuran normal tetapi dalam jumlah cukup banyak.

* + - 1. *k* = 3, luas

Luas area yang besar dan diikuti dengan jumlah tingkat lantai yang banyak. Hal ini dapat menyebabkan ukuran kamar yang sangat luas jika hanya dibangun untuk beberapa kamar. Dan menjadi properti dengan jumlah kamar yang sangat banyak dengan ukuran kamar normal.

Penyebaran *cluster* pada tiap variabel dapat dilihat pada gambar berikut.

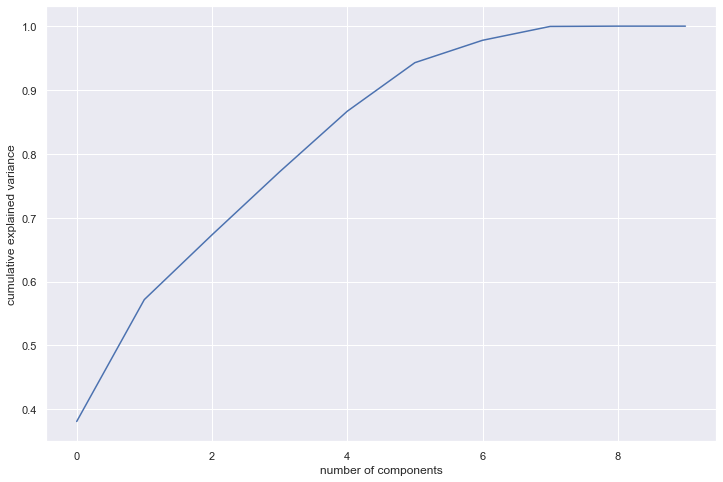


**Gambar 15.** *Cluster* pada Variabel *Area* dan *Floor*

Pada Gambar 15 berikut, menunjukkan bahwa terdapat *cluster* yang tidak menyebar secara rata pada data. Atau dapat dikatakan terdapat *cluster* yang lebih dominan dibanding *cluster* lain. Sehingga dapat disimpulkan bahwa pada setiap *cluster* memiliki jumlah data *cluster* yang berbeda.

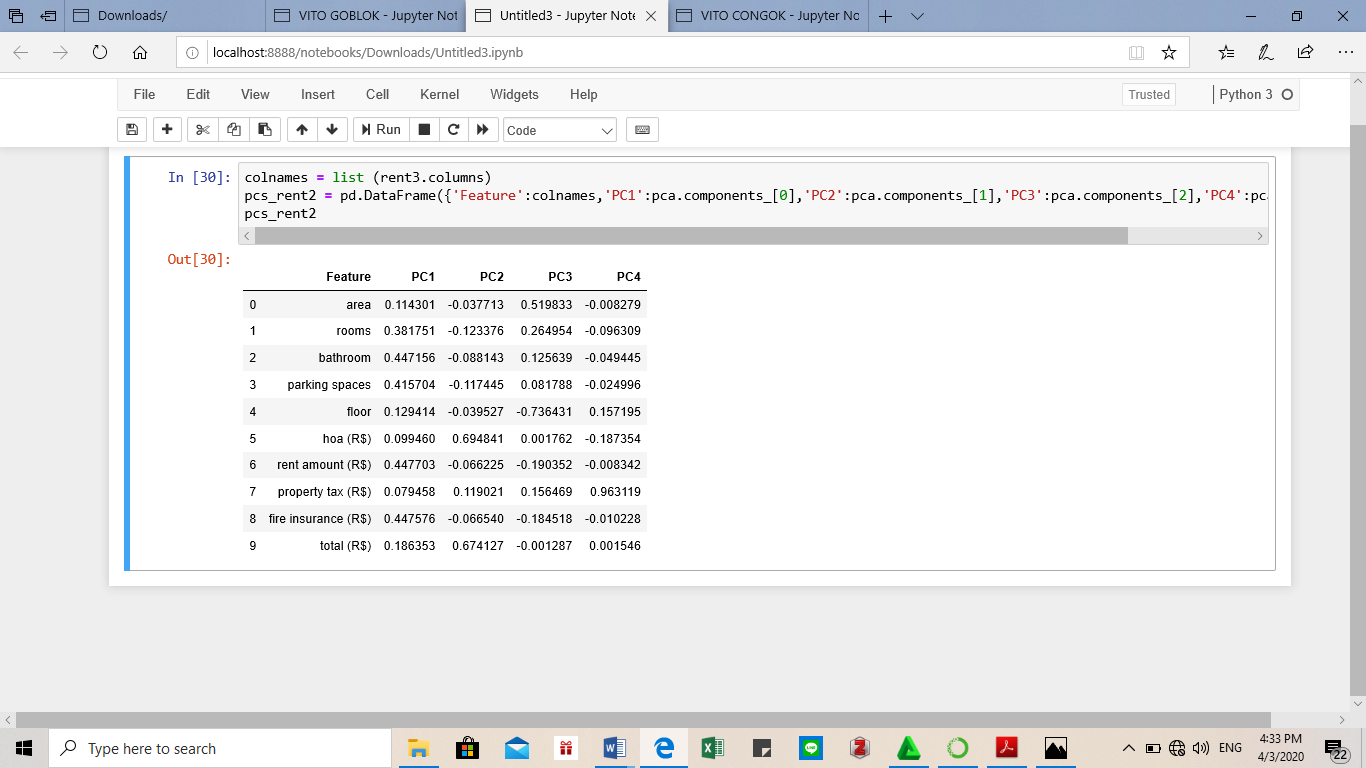
1. *Principal Component Analysis (PCA)*

Analisa dengan menggunakan metodologi PCA dilakukan apabila terdapat korelasi/hubungan antar variabelnya karena tujuan dari analisa dari PCA adalah membuat sejumlah variabel baru yang tidak memiliki korelasi antar variabelnya (korelasi = 0) dan jumlah variabel yang ada lebih sedikit dari jumlah variabel awal *(feature extraction)*.



**Gambar 16.** Banyak PCA

Dari Gambar 16 di atas, nilai *number of components* yang melebihi 80% dari *cumulative explained variance* memiliki nilai 4, sehingga hasil dari PC adalah sebagai berikut.



**Gambar 17.** Persamaan *Principal Component Analysis*

Hasil di atas menunjukkan bahwa faktor-faktor yang dominan pada tiap PC. Pada PC 1 terdapat 5 faktor yang dominan, pada PC 2 terdapat ada 2 faktor yang dominan, 2 faktor dominan pada PC 3 dan pada PC 4 terdapat 1 faktor yang dominan.

1. KESIMPULAN DAN SARAN
2. *Kesimpulan*

Berdasarkan analisis dan pembahasan yang telah dilakukan pada bab IV, maka diperoleh kesimpulan pada data *Brazilian houses to rent*  terdapat *missing value* pada variabel *floor* sebanyak 2461 dan pada tiap variabel terdapat *outlier.* Dari hasil statistika deskriptif juga dapat dilihat bahwa selisih dari dari maksimum dan minimum bernilai cukup besar, sehingga mengidentifikasi adanya *outlier*. Dan pada analisis *clustering* membuat 3 kelompok properti menjadi 3 bagian, yaitu sempit, sedang dan luas. Dan pada *feature extraction* menggunakan *principal component analysis* mereduksi 13 variabel menjadi 4 PC.

1. *Saran*

Berdasarkan kesimpulan yang diperoleh, sebaiknya pemilik properti juga dapat mempertimbangkan tentang harga sewa yang dipasang kepada calon penyewa terhadap fasilitas yang disediakan oleh agen properti. Dan untuk penyewa dapat melakukan survei terlebih dulu terhadap beberapa penyewaan properti yang ada agar bisa mendapatkan rumah yang akan disewa sesuai dengan kriteria yang diinginkan.

Untuk analisis selanjutnya sebaiknya data yang digunakan untuk analisis berdistribusi normal dan tidak memiliki *outlier*, tetapi informasi yang didapat juga tidak terbuang sehingga untuk mencari deteksi *outlier* dan dapat mengatasinya dengan tepat.

DAFTAR PUSTAKA

|  |  |
| --- | --- |
| [1] | R. E. Walpole, Pengantar Statistika, Jakarta: PT. Gramedia Pustaka Utama, 1993. |
| [2] | Q. Y. Fang and L. X. Wei, "A Data Preprocessing Algorith for Classification Model Base On Rough Sets," in *International Conference on Solid State Devices and Material Science*, 2012, pp. 2025-2029. |
| [3] | J. Han, M. Kamber and J. Pei, in *Data Mining Concepts and Techniques*, Morgan Kaufmann, 2012. |
| [4] | M. Djauhary, "Catatan Kuliah Analisis Data," 2001. |
| [5] | M. Dash and L. H., "Feature Selection for Classification, Intelligent Data Analysis," 1997. |
| [6] | B. Fry, "Visualizing Data," Sebastopol, O'Reilly, 2008. |
| [7] | C. N. Knafli, "Storytelling with Data," vol. 2. Wiley, 2015. |
| [8] | A. Adiarsa, ""Jenis Sampel, Grafik dan Tabel.'," 07 03 2012. [Online]. Available: https://blog.ub.ac.id/adiarsa/2012/03/07/jenis-sampel-grafik-dan-tabel-statistik/.. |
| [9] | Tattar, dkk, A Course in Statistics with R, Chennai: John Wiley & Sons, Ltd, 2016. |
| [10] | R. G. Lomax and D. L. Hahs-Vaughn, An Introduction to Statistical Concepts Third Edition, New York: Taylor & Francis Group, 2012. |
| [11] | A. Hidayat, "Penjelasan Lengkap Tentang Analisis Cluster," Statistikian, 26 Maret 2014. [Online]. Available: https://www.statistikian.com/2014/03/analisis-cluster.html#Definisi\_Analisis\_Cluster. [Accessed 2 April 2020]. |
| [12] | I. T. Jollife, Principal Component Analysis Second Edition, New York: Springer-Verlag, 2002. |
| [13] | J. a. Wichern, Applied Multivariate Statistical Analysis Sixth Edition, Pearson Prentice Hall, 2007. |