## **Data Understanding**

Langkah pertama dalam *data understanding* adalah dengan melakukan pemuatan data ke dalam bentuk DataFrame menggunakan *library* Python Pandas. *Library* Pandas menerima *directory* dari data yang akan dimuat kemudian memasukkan dalam variabel “df”.

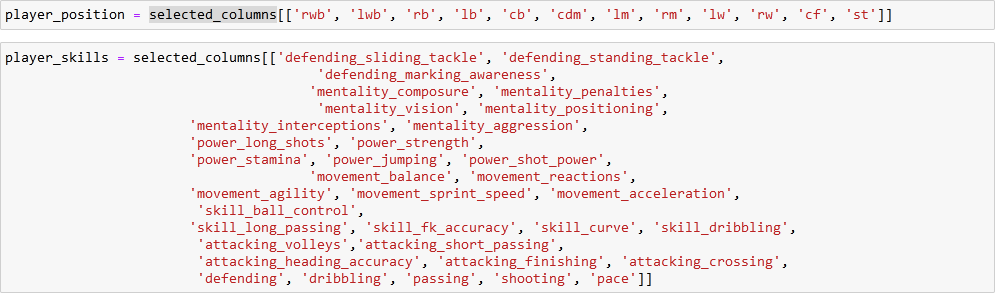
Langkah kedua adalah memilih data yang diinginkan, yaitu data terbaru dan variabel yang diinginkan. Data terbaru dapat didapatkan dengan memfitler kolom “fifa\_version”, sedangkan variabel yang diinginkan dapat dilakukan dengan memasukkan nama kolom yang diingkan ke dalam sebuah list dan kemudian menjadikannya filter sebuah DataFrame.

Langkah ketiga adalah menghilangkan data yang bersifut NaN. Penghilangan data tersebut dilakukan dengan fungsi *library* Pandas dropna(). Fungsi dropna() akan menghilangkan semua baris yang terdapat data yang tidak lengkap. Ditambahkan argumen “inplace=True” pada fungsi tersebut untuk melakukan *overwrite* terhadap DataFrame yang lama dengan yang baru dimana baris data tidak memiliki nilai yang bersifat NaN.



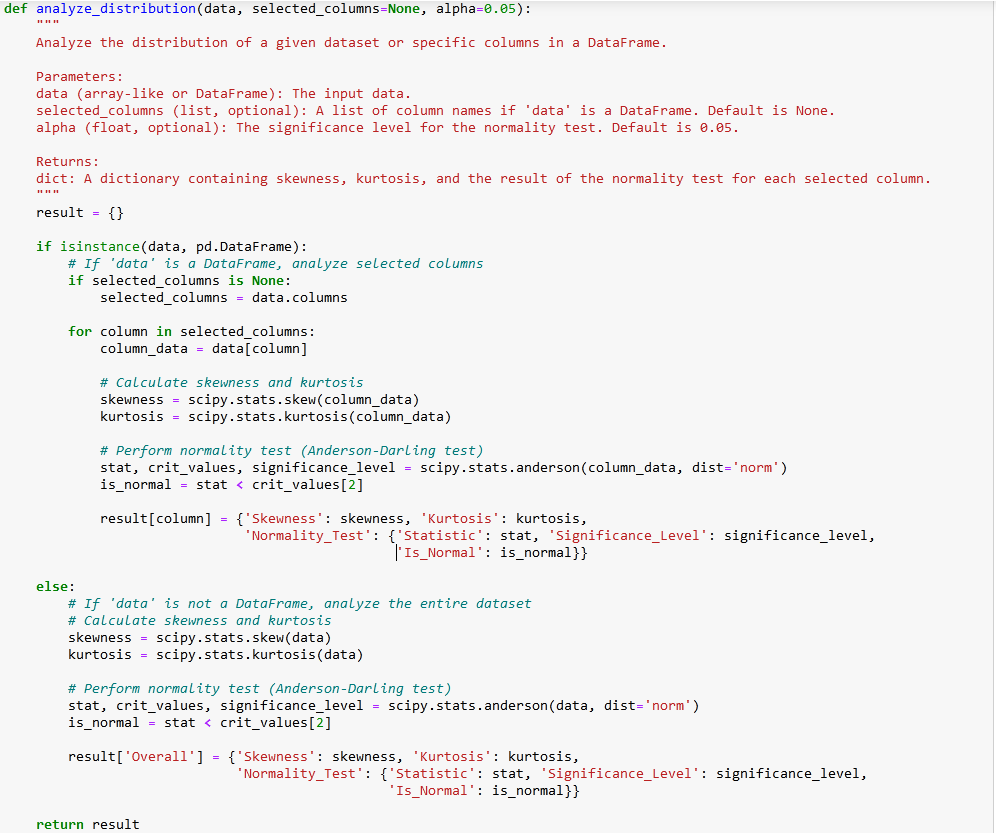
Gambar 1. 1 Implementasi Langkah Pertama hingga Ketiga.

Langkah keempat adalah dengan memisah variabel posisi dan keahlian pemain sepakbola. Pemisahan variabel ini dilakukan untuk kemudahan proses *data mining*. Variabel posisi dimasukkan ke dalam “player\_position” dan variabel keahlian dimasukkan ke dalam “player\_skills.” Semua variasi variabel posisi dan keahlian pemain sepakbola dipilih untuk analisis data.



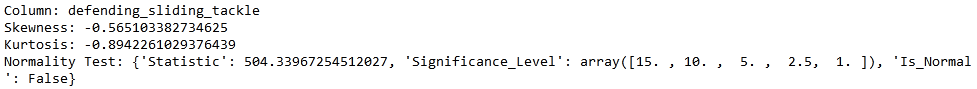
Gambar 1. 2 Implementasi Langkah Keempat

Langkah kelima adalah melakukan uji normalitas data menggunakan Anderson-Darling Test dan melihat *skewness* dan kurtosis dari distribusi. Anderson-Darling test adalah uji statistik yang digunakan untuk menentukan apakah sampel data berasal dari distribusi tertentu, seperti distribusi normal. Semakin besar nilai statistik uji Anderson-Darling, semakin banyak bukti yang menunjukkan bahwa sampel tidak berasal dari distribusi normal. *Skewness* atau kemiringan adalah ukuran statistik yang digunakan untuk mengukur sejauh mana distribusi data cenderung miring ke kiri atau ke kanan dari nilai tengahnya. Nilai *skewness* positif menunjukkan bahwa data condong ke kanan, sedangkan nilai negatif menuunjukkan data condong ke kiri. Data berdistribusi normal akan memiliki nilai *skewness* 0. Kurtosis adalah ukuran statistik yang menggambarkan seberapa curam atau landai bentuk distribusi data. Kurtosis positif menunjukkan bahwa distribusi memiliki ekor yang lebih tebal dan lebih banyak nilai ekstrem di kedua sisi nilai tengah, sedangkan kurtosis negatif menunjukkan bahwa distribusi memiliki ekor yang lebih tipis dan nilai-nilai data lebih terkumpul di sekitar nilai tengah. Uji normalitas ini dilakukan untuk menentukan jenis model dan uji hipotesis hubungan antar variabel nantinya karena beberapa model dan uji hipotesis memiliki asumsi distribusi normal.



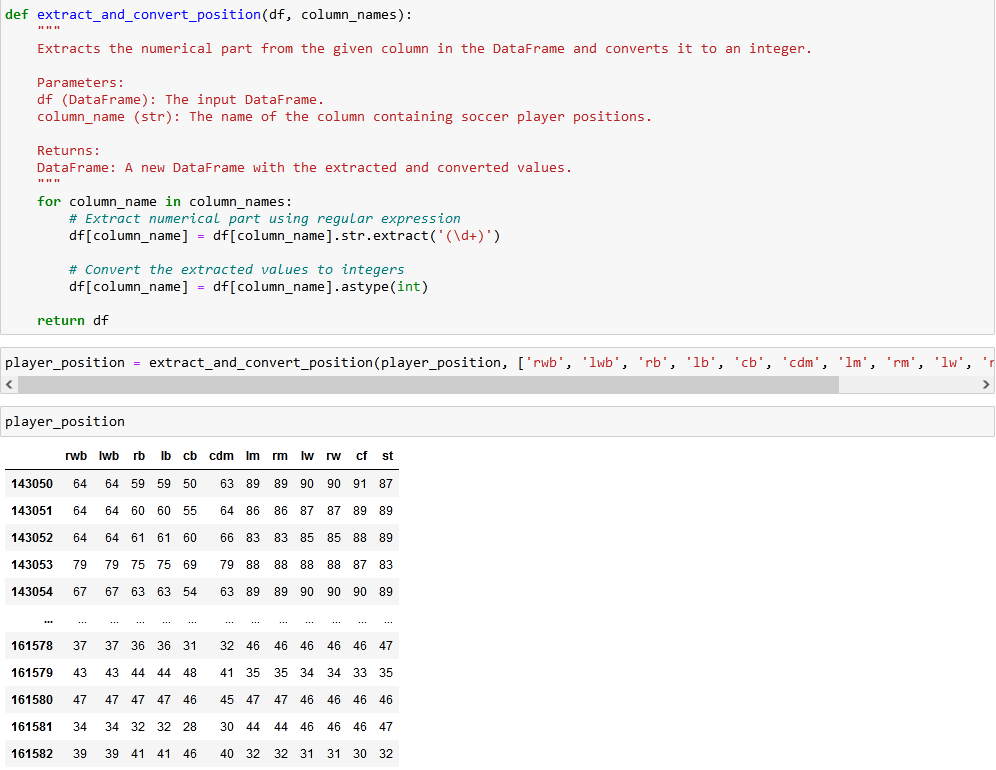
Gambar 1. 3 Implementasi Langkah Kelima

Contoh hasil dari implementasi langkah kelima terlihat pada Gambar 1. 4. Pada gambar tersebut terlihat bahwa *skewness* bernilai negatif, sehingga data condong ke kiri. Nilai kurtosis negatif, sehingga menunjukkan bahwa distribusi memiliki ekor yang lebih tipis dan nilai-nilai data lebih terkumpul di sekitar nilai tengah. Dari hasil juga terlihat bahwa data bukan merupakan data berdistribusi normal.



Gambar 1. 4 Contoh Hasil Langkah Kelima

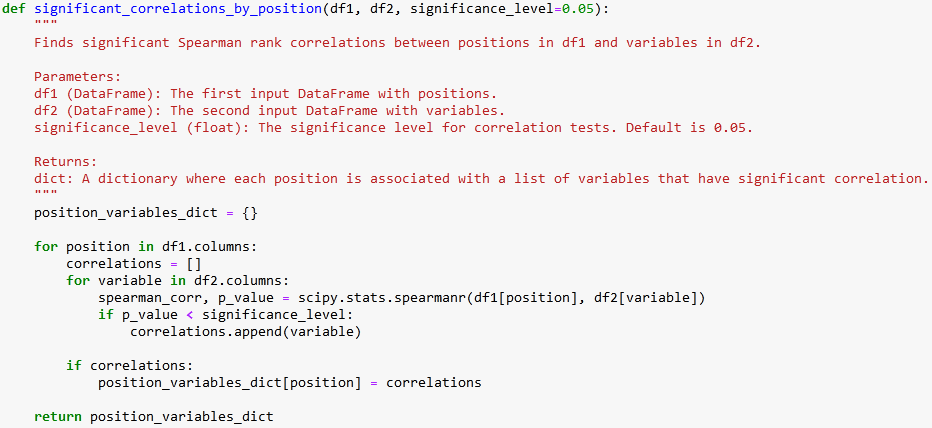
Langkah selanjutnya adalah dengan mengubah variabel posisi pemain sepakbola dari tipe Object menjadi tipe Integer. Pengubahan dari tipe Object menjadi tipe Integer dilakukan dengan implementasi fungsi “extract\_and\_convert\_position” seperti pada Gambar 1.5.



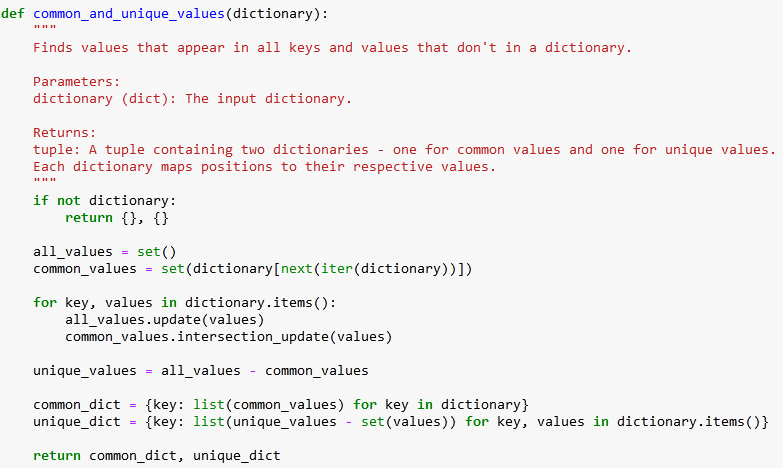
Gambar 1. 5 Implementasi Langkah Keenam

Langkah keenam adalah dengan mengecek signifikansi variabel terhadap masing-masing posisi pemain sepakbola. Pengecekan signifikansi variabel ini dilakukan untuk memangkas fitur yang digunakan demi mengurangi dimensi dan agar tidak menggunakan variabel yang tidak perlu. Pengecekan signifikansi variabel digunakan menggunakan uji korelasi peringkat Spearman. Uji korelasi peringkat Spearman adalah metode statistik yang digunakan untuk menentukan apakah ada hubungan monotonic antara dua variabel. Korelasi peringkat Spearman tidak memerlukan asumsi bahwa data harus terdistribusi normal, sehingga dapat digunakan pada data yang tidak terdistribusi normal.

Pengecekan signifikansi variabel ini dilakukan dengan 2 tahap. Tahapan pertama adalah dengan mengecek signifikansi dan menyimpan variabel yang signifikan terhadap masing-masing posisi pemain, kemudian selanjutnya dilakukan pengecekan terhadap variabel yang tidak signifikan untuk tidak dipilih. Implementasi 2 tahapan ini terlihat pada Gambar 1. 6 dan Gambar 1. 7 secara berurutan.

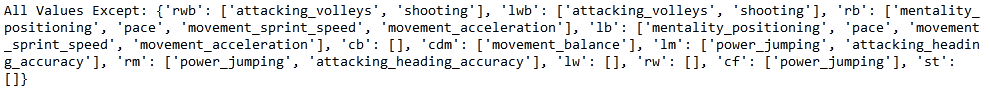


Gambar 1. 6 Tahap Pertama pada Langkah Keenam



Gambar 1. 7 Tahap Kedua pada Langkah Keenam

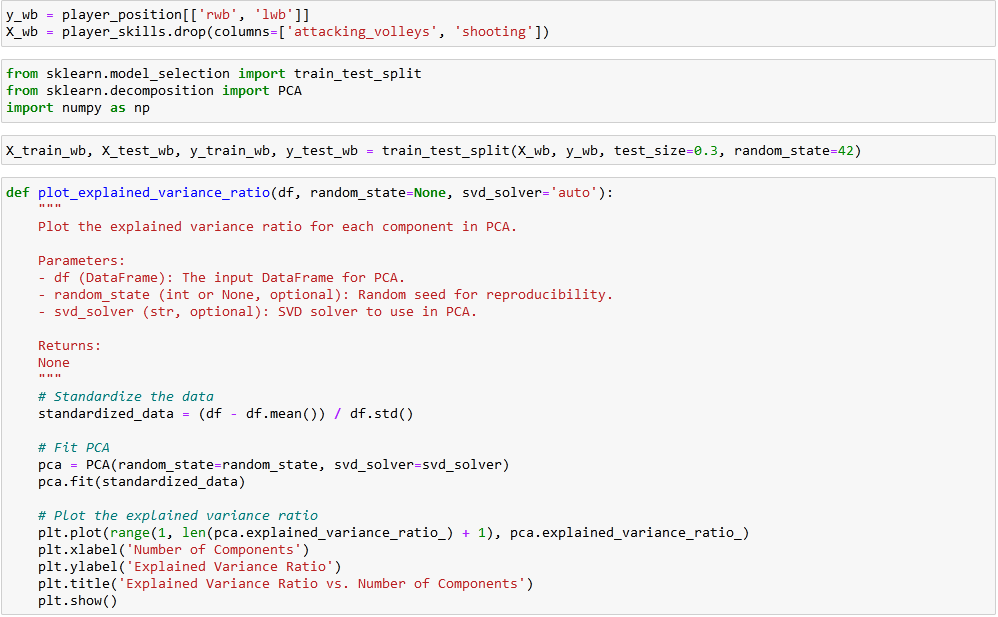
Hasil dari pengecekan signifikansi ditunjukkan pada Gambar 1. 8. Gambar 1. 8 menunjukkan bahwa pada *Right Wing Back* (RWB) dan *Left Wing Back* (LWB) terdapat variabel “attacking\_volleys” dan “shooting” yang tidak signifikan, pada *Right Back* (RB) dan *Left Back* (LB) terdapat variabel 'mentality\_positioning', 'pace', 'movement\_sprint\_speed', dan 'movement\_acceleration' yang tidak signifikan, pada *Center Defensive Midfielder* (CDM) terdapat variabel “movement\_balance” yang tidak signifikan, pada *Left Midfielder* (LM)dan *Right Midfielder* (RM) terdapat variabel 'power\_jumping' dan 'attacking\_heading\_accuracy' yang tidak signifikan, dan terakhir pada variabel *Center Forward* (CF) terdapat variabel 'power\_jumping' yang tidak signifikan. Berdasarkan hal tersebut akan dikecualikan variabel-variabel yang tidak memiliki signifikansi pada tahap *data preparation*.



Gambar 1. 8 Hasil dari Langkah Keenam

1. **Data Preparation**

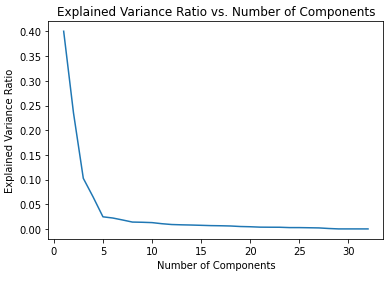
*Data preparation* dilakukan dengan cara mengeculaikan variabel independen yang tidak signifikan, mengimplementasikan *Principal Component Analysis* (PCA), dan juga memilah data *training* dan *testing* dengan perbandingan 7:3. Implementasi *data preparation* ditunjukkan pada Gambar 1. 9.



Gambar 1. 9 Implementasi Data Preparation.

Analisis Komponen Utama (Principal Component Analysis atau PCA) adalah teknik statistik yang digunakan untuk menyederhanakan struktur multidimensi dari dataset, dengan tujuan untuk mengidentifikasi pola dan hubungan antarvariabel. PCA akan mereduksi struktur multidimensi dari dataset menjadi n-komponen yang dapat kita tentukan. Penentuan n-komponen yang ideal dapat menggunakan metode *elbow*. Metode Elbow (Elbow Method) dalam Analisis Komponen Utama (PCA) adalah teknik yang digunakan untuk menentukan jumlah komponen utama yang paling sesuai untuk menggambarkan variasi dalam dataset. Metode ini memungkinkan kita untuk memilih jumlah komponen utama yang memaksimalkan penjelasan variabilitas data.

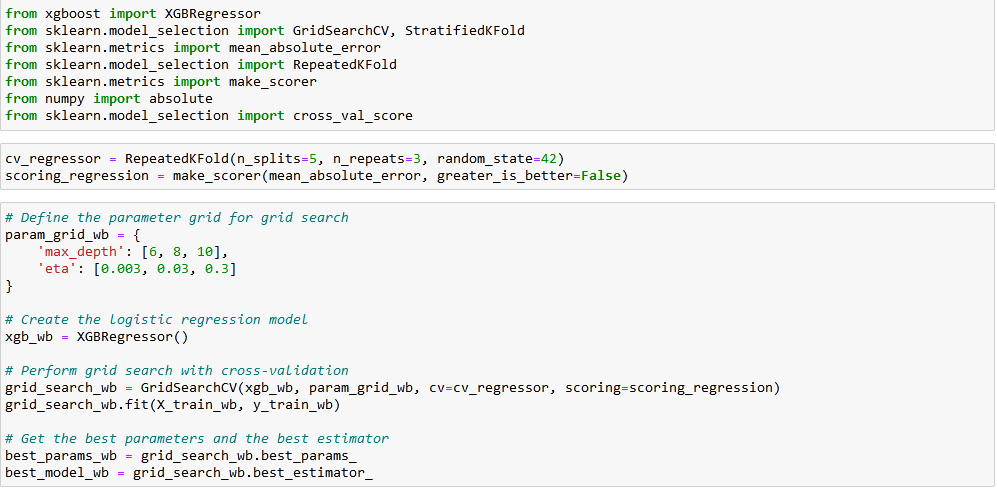
Hasil metode *elbow* dalam PCA dapat dilihat pada Gambar 1. 10. Gambar 1. 10 memperlihatkan visualisasi metode *elbow* pada variabel yang signifikan pada posisi *Wingback*. Terlihat bahwa dalam grafik sumbu x menunjukkan jumlah n-komponen dan sumbu y menunjukkan *explained variance* dari jumlah n-komponen tersebut. Variansi yang dijelaskan (*Explained Variance*) dalam Analisis Komponen Utama (*Principal Component Analysis* atau PCA) merujuk pada sejauh mana setiap komponen utama mampu menjelaskan variabilitas total dalam dataset. Ketika kita melakukan PCA, komponen-komponen utama dibentuk sedemikian rupa sehingga mereka mampu merangkum informasi yang paling penting dari dataset. Berdasarkan hal tersebut dipilih jumlah komponen paling sedikit namun juga paling mewakili informasi dari dataset yang asli. Terlihat pada Gambar 1. 10 bahwa setelah reduksi dimensi menjadi komponen ke-5 menambah jumlah n-komponen tidak memberikan tambahan *explained variance* yang signifikan, sehingga cukup dipilih jumlah reduksi dimensi menjadi 5 komponen.



Gambar 1. 10 Grafik Explained Variance dari Metode Elbow dalam PCA.

1. **Modelling**

Tahapan *modelling* dilakukan dengan cara melakukan *hyperparameter tuning* model *XGBRegressor*. Tuning hiperparameter (*hyperparameter tuning*) adalah proses mencari nilai yang optimal untuk hiperparameter dalam suatu model pembelajaran mesin dengan menggunakan teknik atau metode tertentu. Hiperparameter adalah parameter yang nilainya tidak diajarkan oleh model, tetapi harus diatur sebelum proses pelatihan dimulai. Tujuannya adalah untuk meningkatkan kinerja dan generalisasi model pada dataset yang tidak terlihat selama pelatihan. XGBoost adalah algoritma penguatan (*boosting*) yang menggabungkan beberapa model pohon keputusan untuk meningkatkan kinerja prediktif. Algoritma ini bekerja dengan cara membangun pohon secara berurutan, di mana setiap pohon berusaha untuk memperbaiki kesalahan prediksi pohon sebelumnya. Tahapan *modelling* dilakukan dengan kode Python pada Gambar 1. 11.

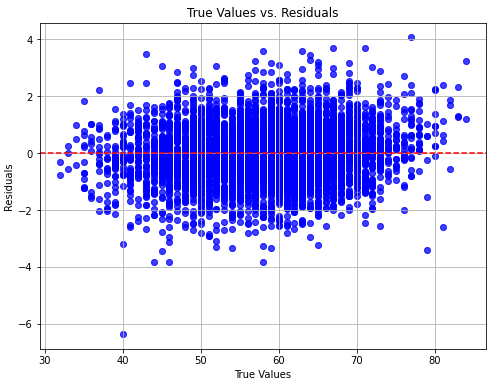


Gambar 1. 11 Kode Implementasi Tahapan Modelling

1. **Evaluasi**

Dipilih *metrics* *Mean Absolute Error* (MAE) dan metode analisis residual untuk melakukan evaluasi terhadap model. Mean Absolute Error (MAE) atau Kesalahan Rata-rata Absolut adalah suatu metrik evaluasi yang digunakan dalam statistika dan ilmu data untuk mengukur sejauh mana selisih antara nilai prediksi dan nilai sebenarnya pada suatu model regresi. MAE dihitung dengan cara mengambil nilai absolut dari selisih antara setiap prediksi dan nilai sebenarnya, kemudian menghitung rata-rata dari nilai-nilai absolut tersebut. Analisis residu (residual analysis) adalah suatu pendekatan dalam statistika yang digunakan untuk mengevaluasi seberapa baik model statistik atau model prediksi menyesuaikan data yang diamati. Residu adalah selisih antara nilai yang diamati dan nilai yang diprediksi oleh model. Analisis residu membantu untuk memeriksa apakah model telah menangkap dengan baik pola-pola dalam data ataukah ada pola yang masih tertinggal dalam residu.

Contoh hasil evaluasi adalah pada Gambar 1. 12 dimana menunjukkan analisis residual dari model untuk regresi posisi *Wingback*. Pada Gambar 1. 12 menunjukkan bahwa data bersifat homoskedastis atau residu memiliki variabilitas yang konstan menunjukkan bahwa model bekerja dengan baik, namun terlihat bahwa masih terdapat beberapa *outlier* yang perlu dihilangkan.



Gambar 1. 12 Visualisasi Analisis Residual dari Regresi Posisi Wingback