MEDICAL INSURANCE COST

GAMMA DISTRIBUTION

CALVIN JANITRA

NATHANAEL JUNICO ODI PERDANA

Agenda

01	Konsep Distribusi Gamma
02	Dataset Overview
03	Penentuan Jenis Distribusi
04	Pemilihan Model
05	Hasil & Evaluasi

Konsep Distribusi Gamma

Pengertian

Distribusi Gamma adalah distribusi probabilitas kontinu yang sangat fleksibel untuk memodelkan variabel acak positif yang skewed (miring).

Sifat-Sifat Distribusi Gamma

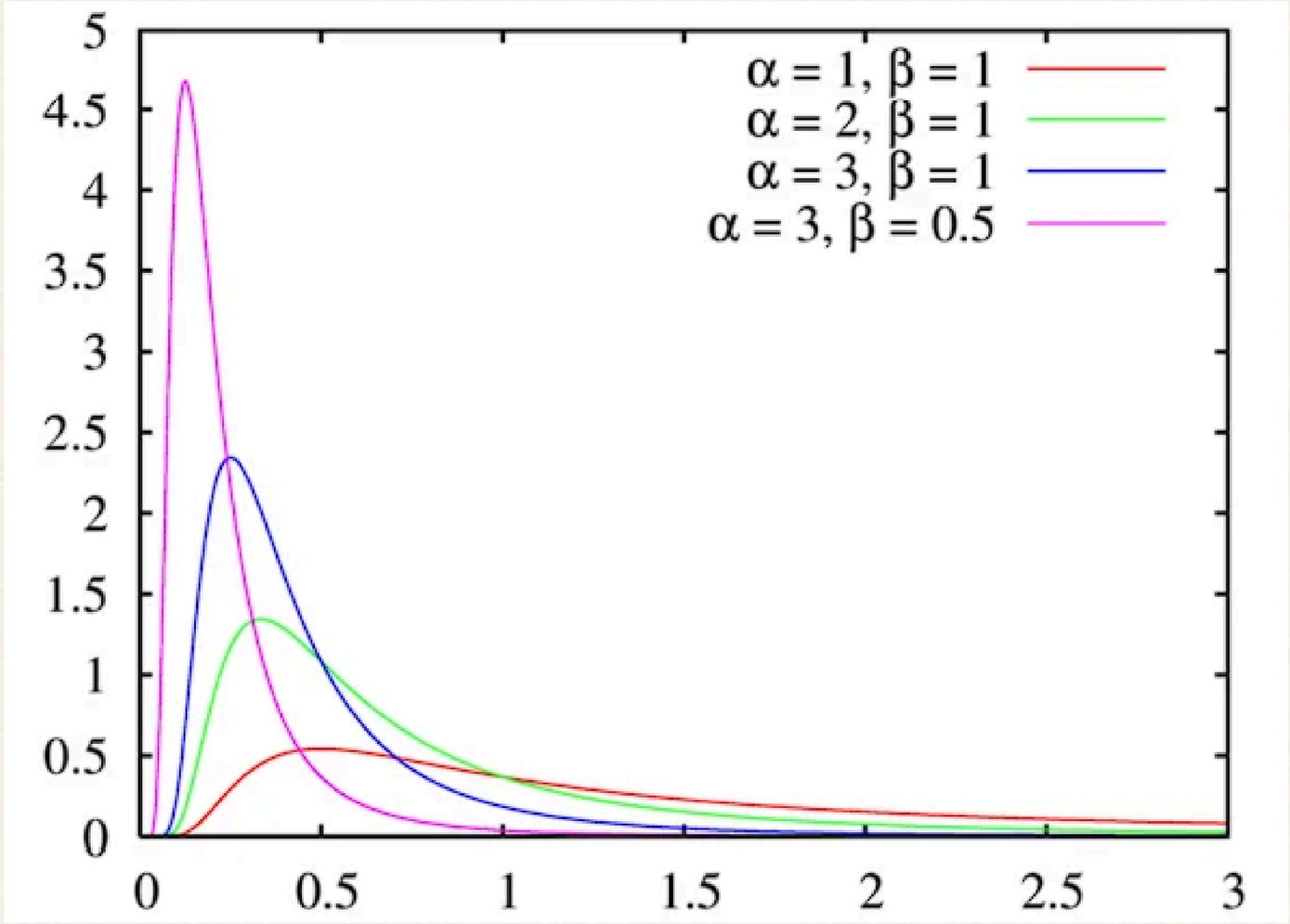
1. Hanya bernilai kontinu positif ($x > 0$) : Data tidak boleh bernilai negatif.
2. Bersifat right-skewed (miring ke kanan/ekor panjang ke kanan).
3. Varians meningkat seiring bertambahnya rata-rata (ada heteroskedastisitas).

Kapan menggunakan Distribusi Gamma?

1. Biaya (medical cost, insurance claims)
2. Waktu tunggu (waiting time) untuk terjadinya kejadian tertentu.
3. Curah hujan (rainfall).
4. Lifetime data (waktu sampai kegagalan).

Konsep Distribusi Gamma

Kurva Distribusi Gamma



Konsep Distribusi Gamma

Persamaan Rumus PDF (Probability Density Function):

$$f(x;\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$$

- **f(x;α,β)**: fungsi kepadatan probabilitas dari distribusi Gamma. Nilai ini memberikan densitas probabilitas pada titik x untuk parameter α dan β.
- **x**: variabel acak kontinu yang dimodelkan oleh distribusi Gamma di mana nilai x harus positif (x>0)
- ***α* (alpha)**: parameter bentuk (shape parameter).
- ***β* (beta)**: parameter laju (rate parameter).
- **Γ(*α*)**: fungsi Gamma yang merupakan generalisasi dari faktorial untuk bilangan real dan kompleks.
- ***e***: bilangan Euler atau bilangan natural (2,71828...)

Konsep Distribusi Gamma

Persamaan Rumus CDF (Cumulative Distribution Function):

$$F(x;k,\theta) = \frac{1}{\Gamma(k)}\gamma\left(k, \frac{x}{\theta}\right)$$

- **f(x;*k*,θ)**: fungsi distribusi kumulatif (CDF).
- **x**: nilai acak yang diamati.
- ***k***: parameter bentuk (shape parameter).
- ***θ* (theta)**: skala penyebaran (scale parameter).
- **γ**: integral dari fungsi gamma, tapi hanya dari 0 sampai x/θ, bukan sampai ∞.
- **Γ(*α*)**: fungsi gamma yang merupakan generalisasi dari faktorial untuk bilangan real dan kompleks.

Konsep Distribusi Gamma

Perbedaan PDF dan CDF:

Aspek	PDF	CDF
Apa?	Kepadatan probabilitas di sekitar x	Probabilitas kumulatif sampai x
Range	Bisa > 1 (bukan probabilitas langsung)	Selalu antara 0 – 1
Interpretasi	Seberapa sering nilai tertentu relatif muncul	Seberapa banyak data di bawah nilai tertentu
Visual	Kurva berbentuk lonceng/skewed	Kurva naik monoton dari 0 → 1

Dataset Overview

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520
...	...	...	...	...	...	...	...
1333	50	male	30.970	3	no	northwest	10600.54830
1334	18	female	31.920	0	no	northeast	2205.98080
1335	18	female	36.850	0	no	southeast	1629.83350
1336	21	female	25.800	0	no	southwest	2007.94500
1337	61	female	29.070	0	yes	northwest	29141.36030

1338 rows × 7 columns

01 Age : Umur Pelanggan

02 Sex : Jenis Kelamin

03 BMI : Body Mass Index

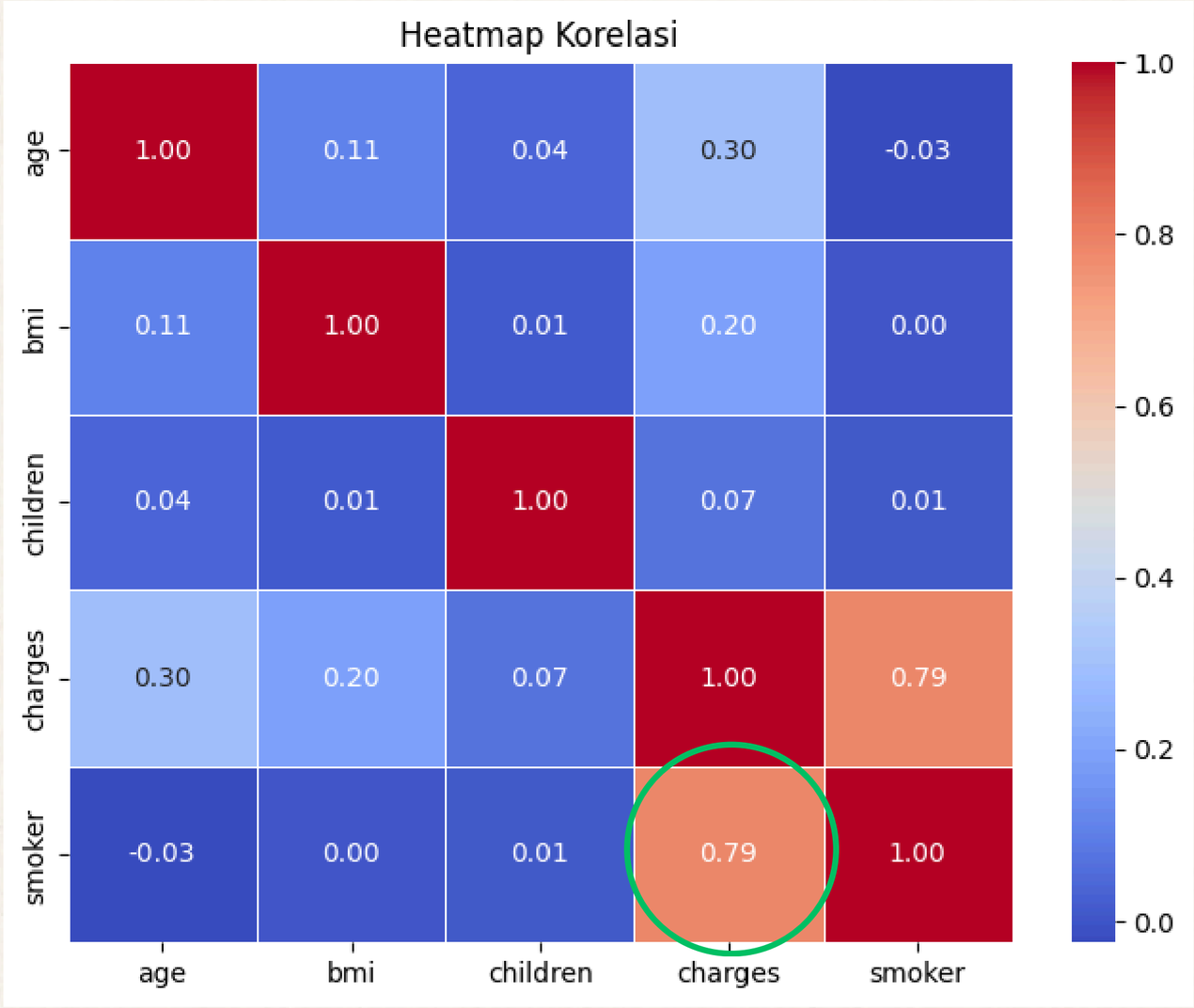
04 Children : Jumlah tanggungan anak.

05 Smoker : Status perokok (yes / no).

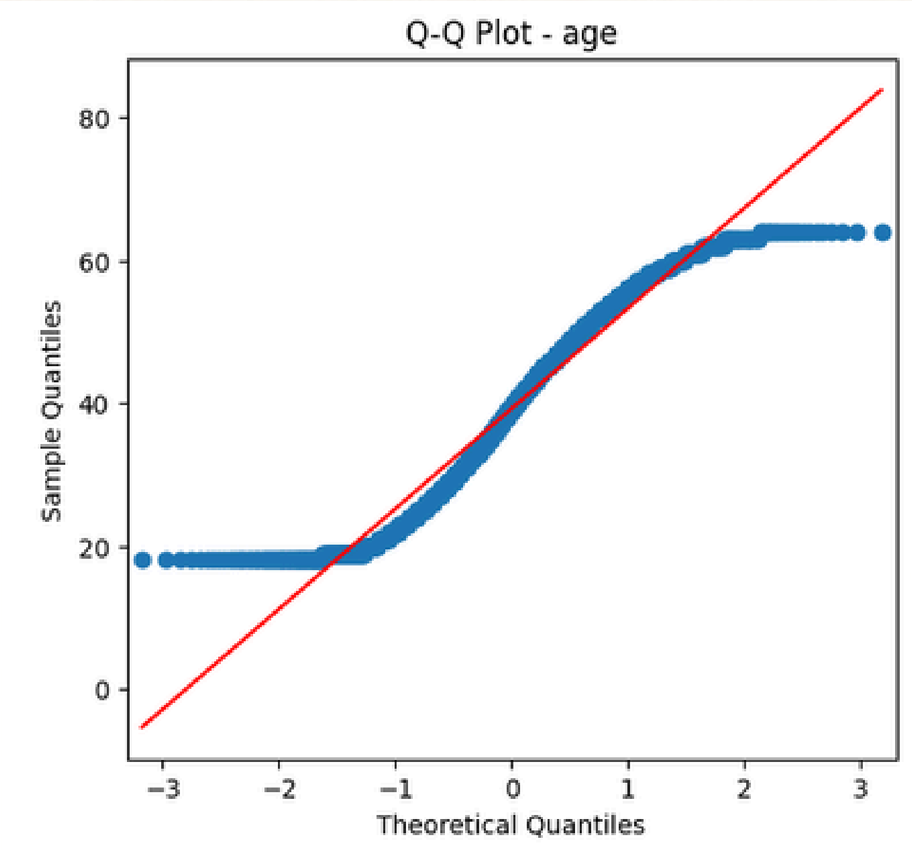
06 Region : wilayah tempat tinggal (southeast, southwest, dll).

07 Charges : Total biaya asuransi kesehatan yang dibebankan.

Uji Korelasi



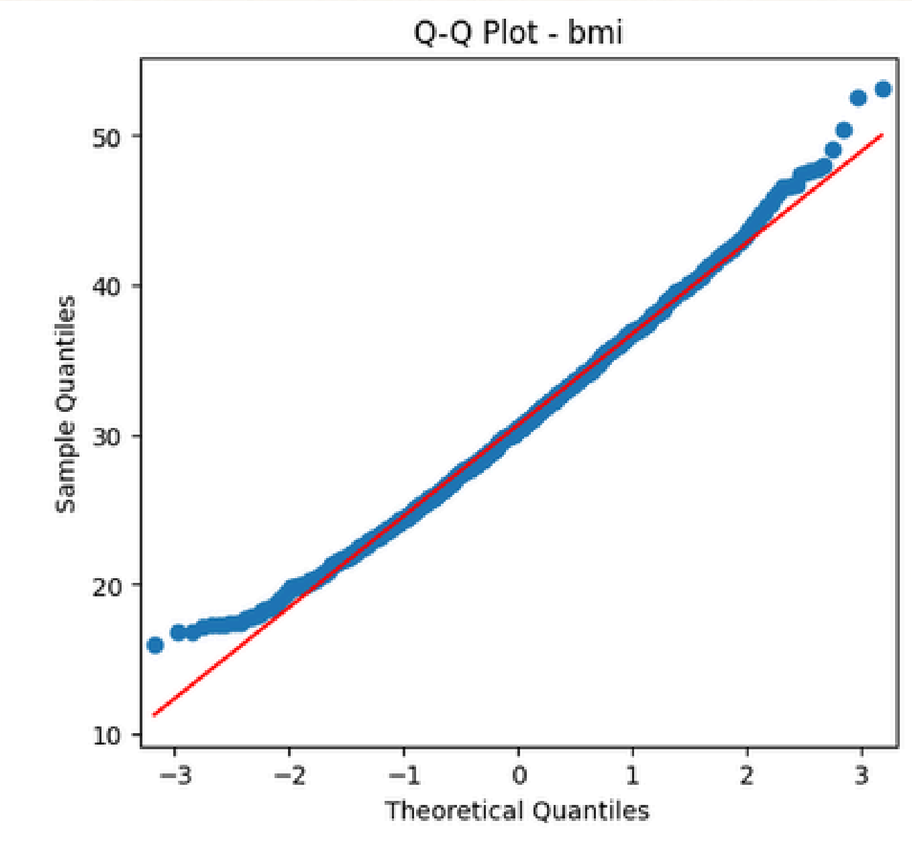
Uji Normalitas



AGE

```
Kolom: age
KS Statistic = 0.0790, p-value = 0.0000
➡ Data tidak berdistribusi normal (tolak H0).
```

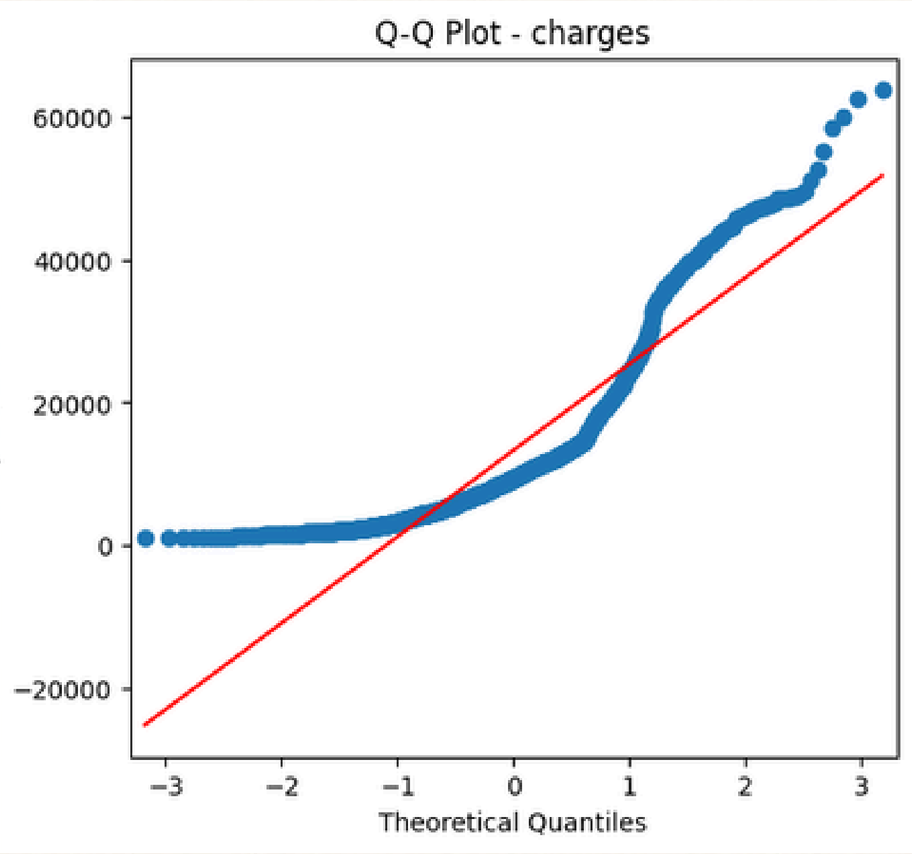
Uji Normalitas



BMI

```
Kolom: bmi
KS Statistic = 0.0261, p-value = 0.3145
➡ Data berdistribusi normal (gagal tolak H0).
```

Uji Normalitas

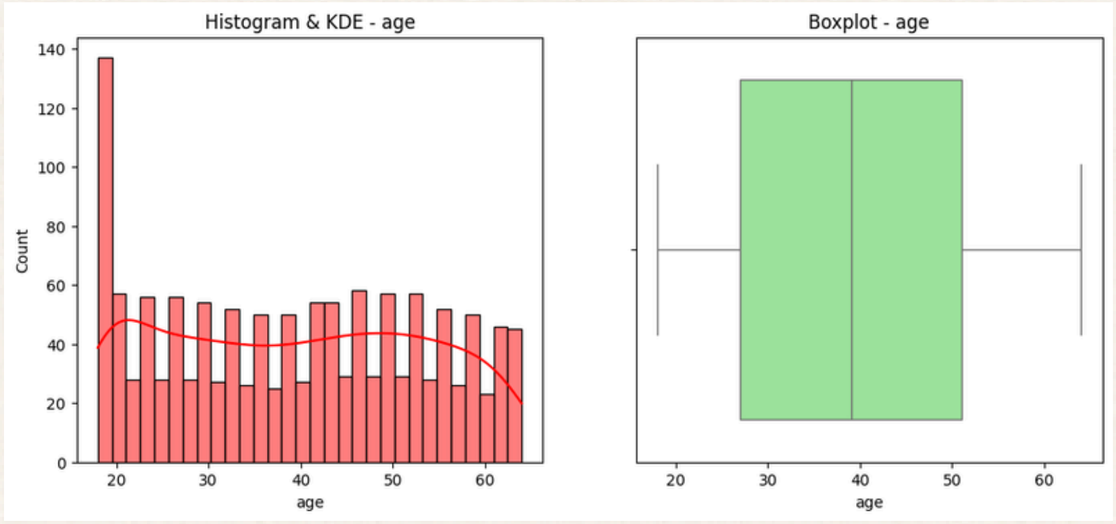


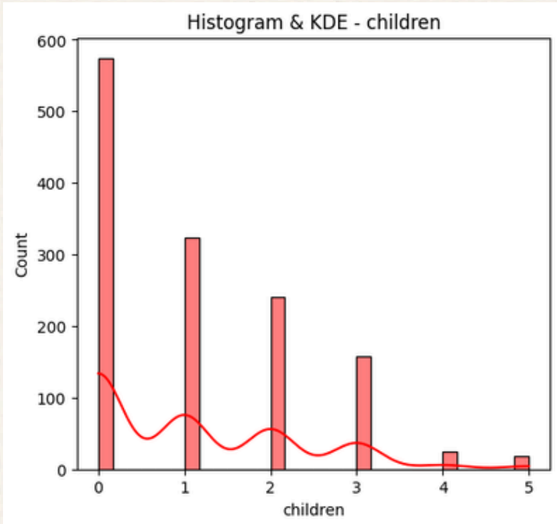
CHARGES

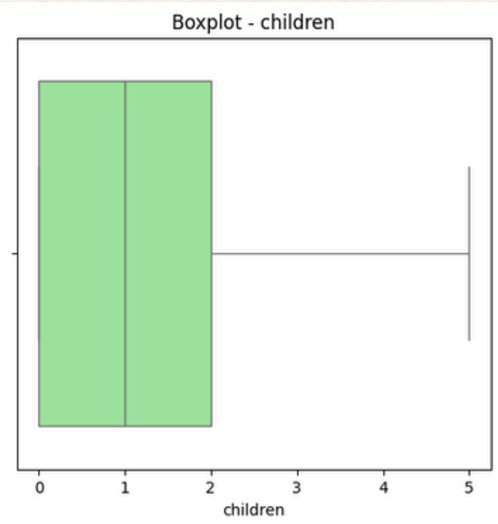
```
Kolom: charges
KS Statistic = 0.1885, p-value = 0.0000
➡ Data tidak berdistribusi normal (tolak H0).
```

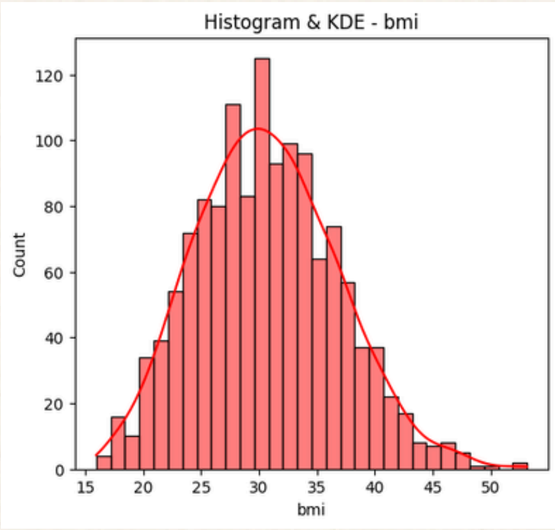
Looking For Gamma Distribution

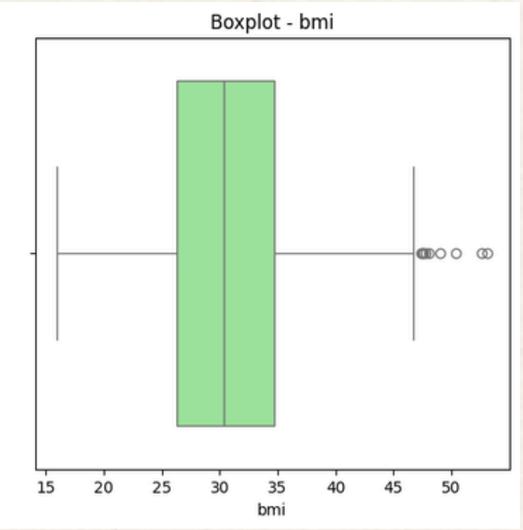
Kolom Numerikal :

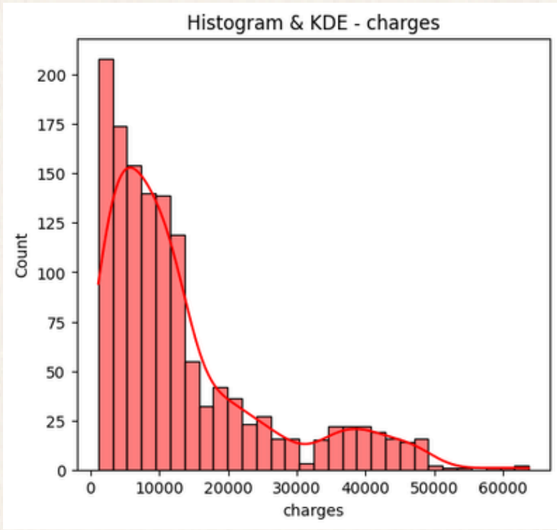


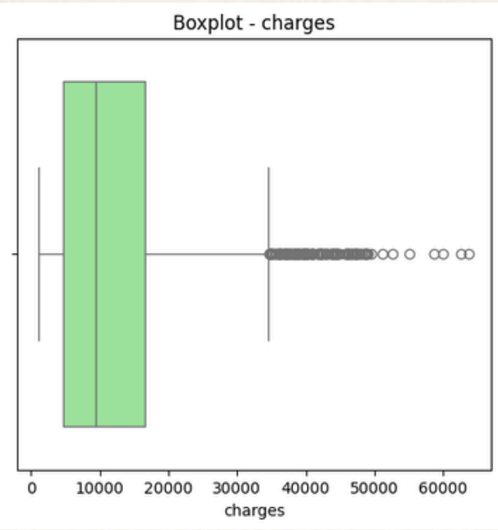






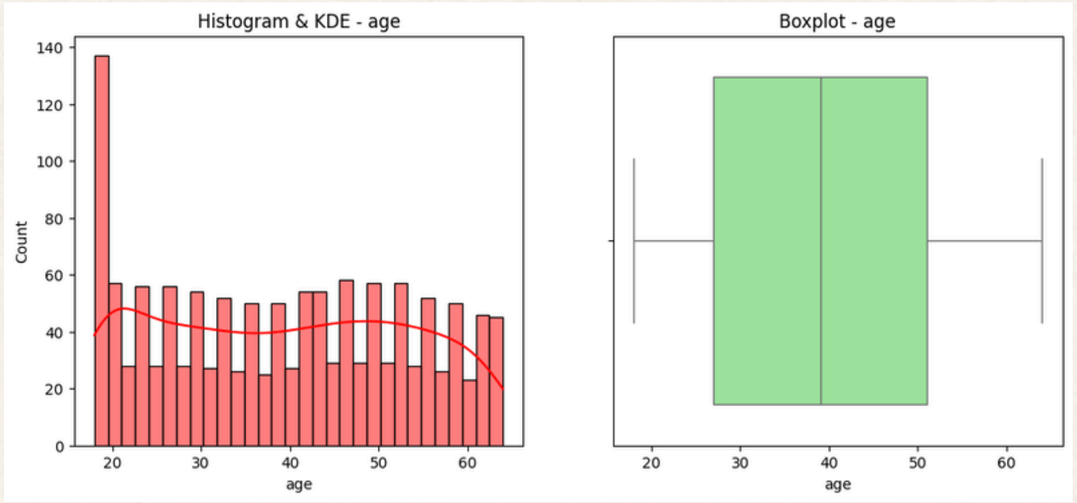


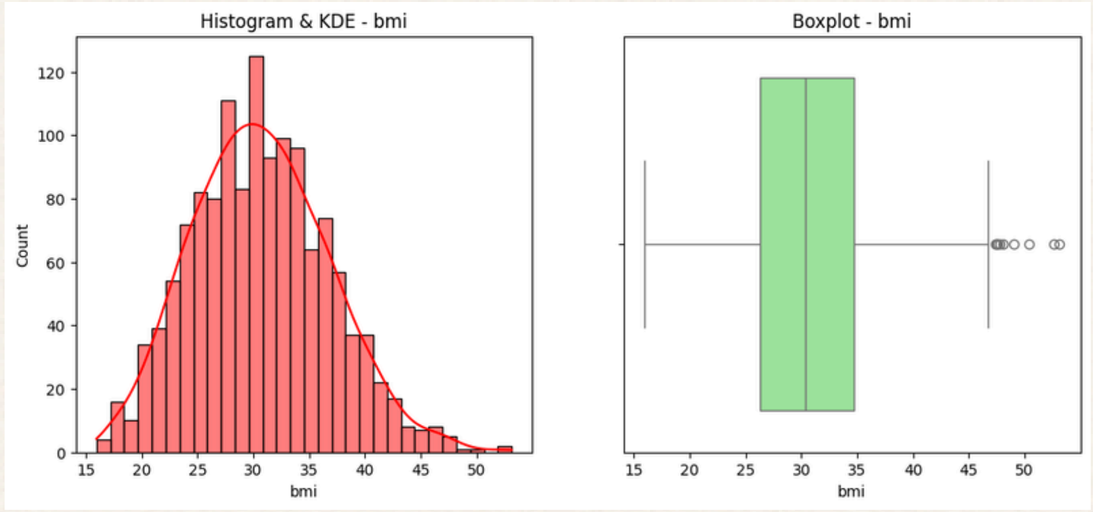




Looking For Gamma Distribution

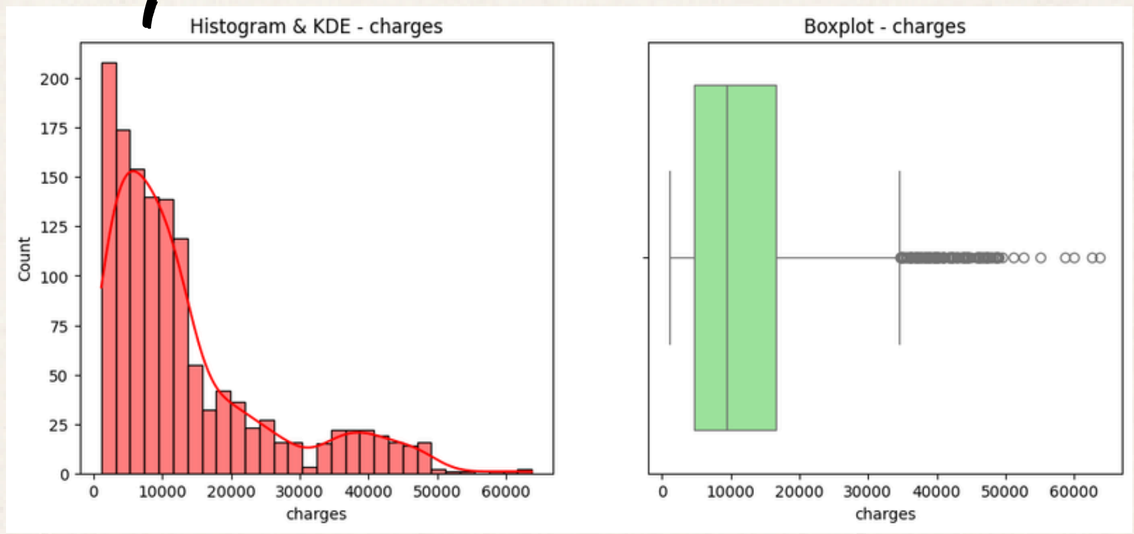
Kolom Numerikal Kontinu :





Sesuai dengan sifat distribusi gamma :

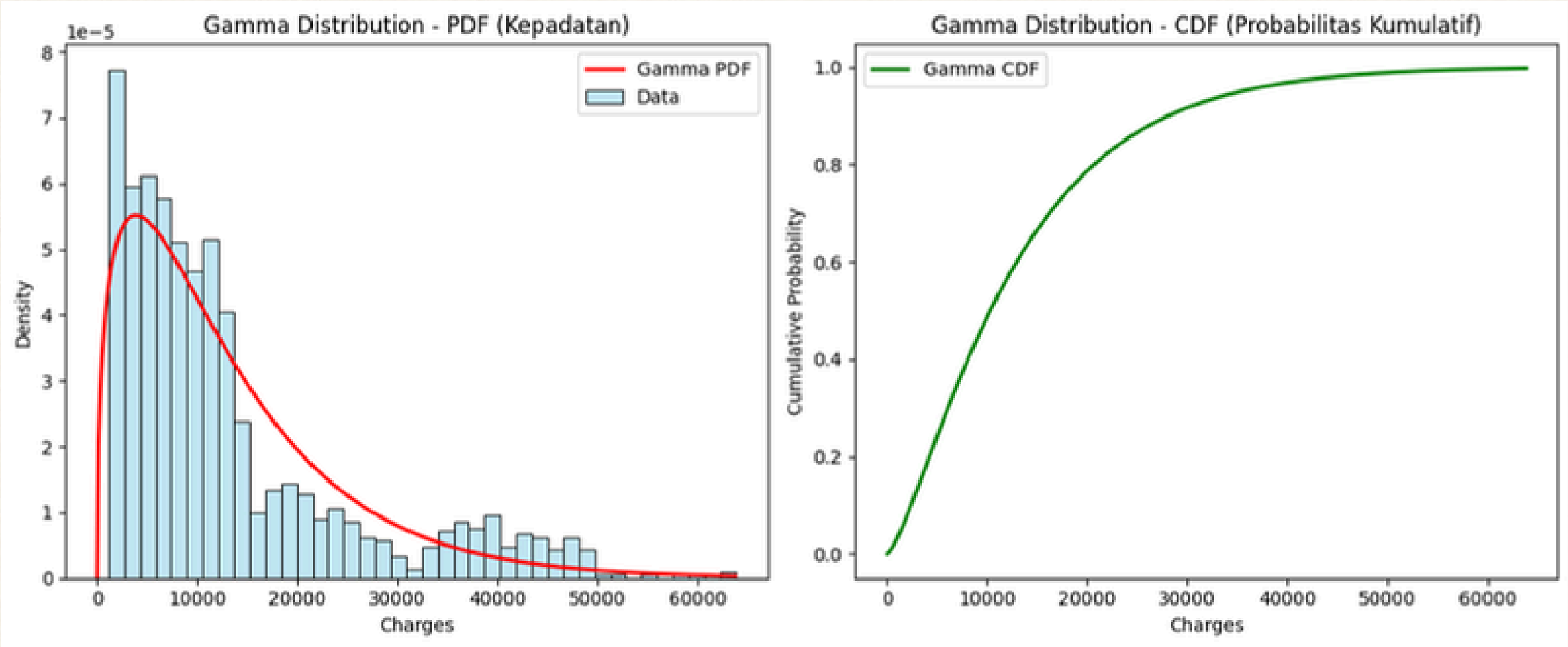
1. Data bernilai kontinu positif > 0
2. Bersifat right skewed



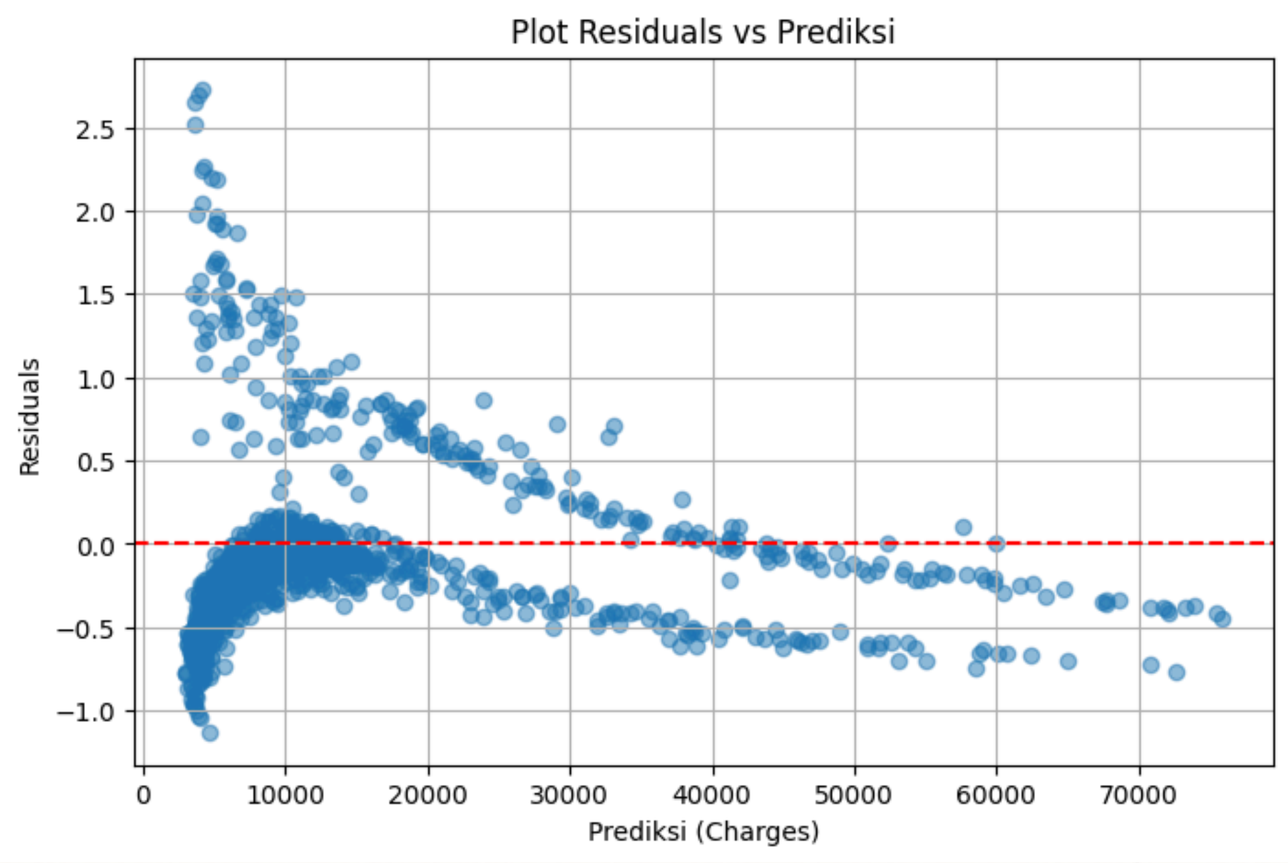
Uji Skewness

age Skewness: 0.055610083072599126
bmi Skewness: 0.28372857291709386
charges Skewness: 1.5141797118745743

Pengujian Distribusi Gamma

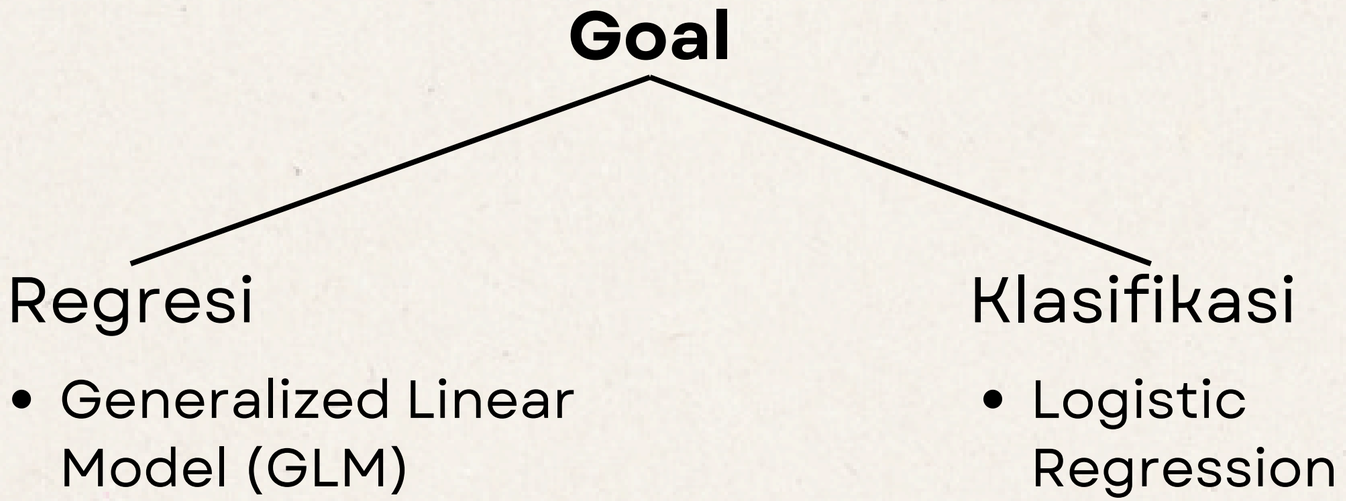


Uji Heteroskedastisitas




```
Uji Breusch-Pagan:
LM Statistic: 104.4443
LM-Test p-value: 0.0000
F Statistic: 14.0657
F-Test p-value: 0.0000
Kesimpulan: Terdapat bukti heteroskedastisitas (p-value < 0.05).
```

Pemilihan Model untuk Distribusi Gamma



Pemilihan Model untuk Distribusi Gamma

Regresi

Train Test Split

Dilakukan split data untuk train 70% dan test 30%.

Library Model

```
import statsmodels.formula.api as smf

model = smf.glm(
  formula="charges ~ age + bmi + smoker + children + sex + region",
  data=train_df,
  family=sm.families.Gamma(sm.families.links.log())
).fit()
```

Hasil Model

Generalized Linear Model Regression Results

Dep. Variable:	charges	No. Observations:	936
Model:	GLM	Df Residuals:	927
Model Family:	Gamma	Df Model:	8
Link Function:	log	Scale:	0.46874
Method:	IRLS	Log-Likelihood:	-9325.8
Date:	Fri, 26 Sep 2025	Deviance:	241.83
Time:	09:57:06	Pearson chi2:	435.
No. Iterations:	19	Pseudo R-squ. (CS):	0.6748
Covariance Type:	nonrobust		

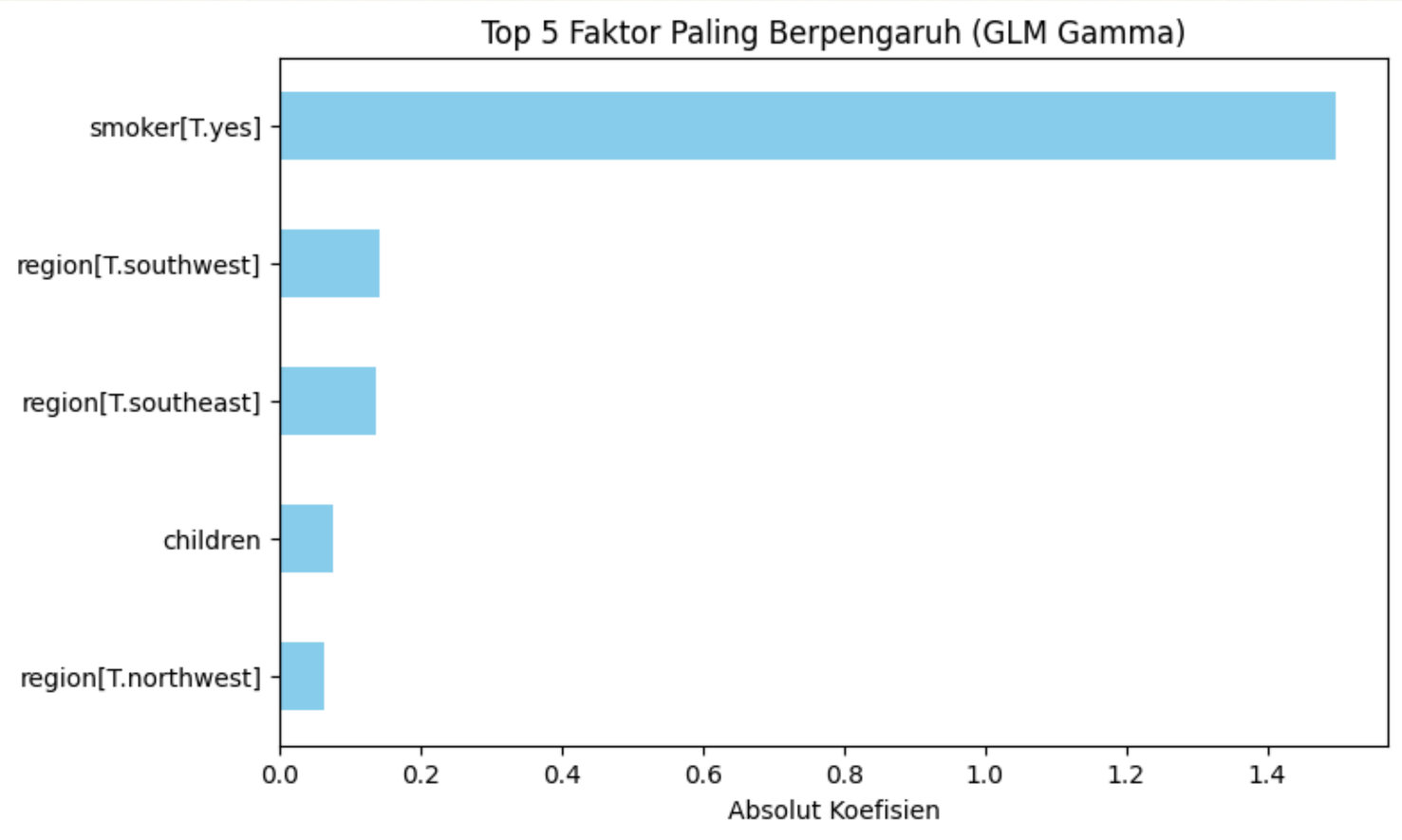
	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	7.4034	0.134	55.135	0.000	7.140	7.667
smoker[T.yes]	1.4963	0.055	27.054	0.000	1.388	1.605
sex[T.male]	-0.0587	0.045	-1.307	0.191	-0.147	0.029
region[T.northwest]	-0.0634	0.063	-1.002	0.316	-0.187	0.061
region[T.southeast]	-0.1358	0.064	-2.112	0.035	-0.262	-0.010
region[T.southwest]	-0.1415	0.064	-2.205	0.027	-0.267	-0.016
age	0.0285	0.002	17.823	0.000	0.025	0.032
bmi	0.0142	0.004	3.639	0.000	0.007	0.022
children	0.0758	0.019	4.091	0.000	0.039	0.112

Evaluasi pada Test Set:
MAE: 4020.38
RMSE: 7104.76
R^2: 0.66

Persamaan Model (Log-Link)

log(E[charges]) = 7.4034 + 1.4963*smoker[T.yes] + -0.1358*region[T.southeast] + -0.1415*region[T.southwest] + 0.0285*age + 0.0142*bmi + 0.0758*children

Features Importance



Pemilihan Model untuk Distribusi Gamma

Klasifikasi

Train Test Split

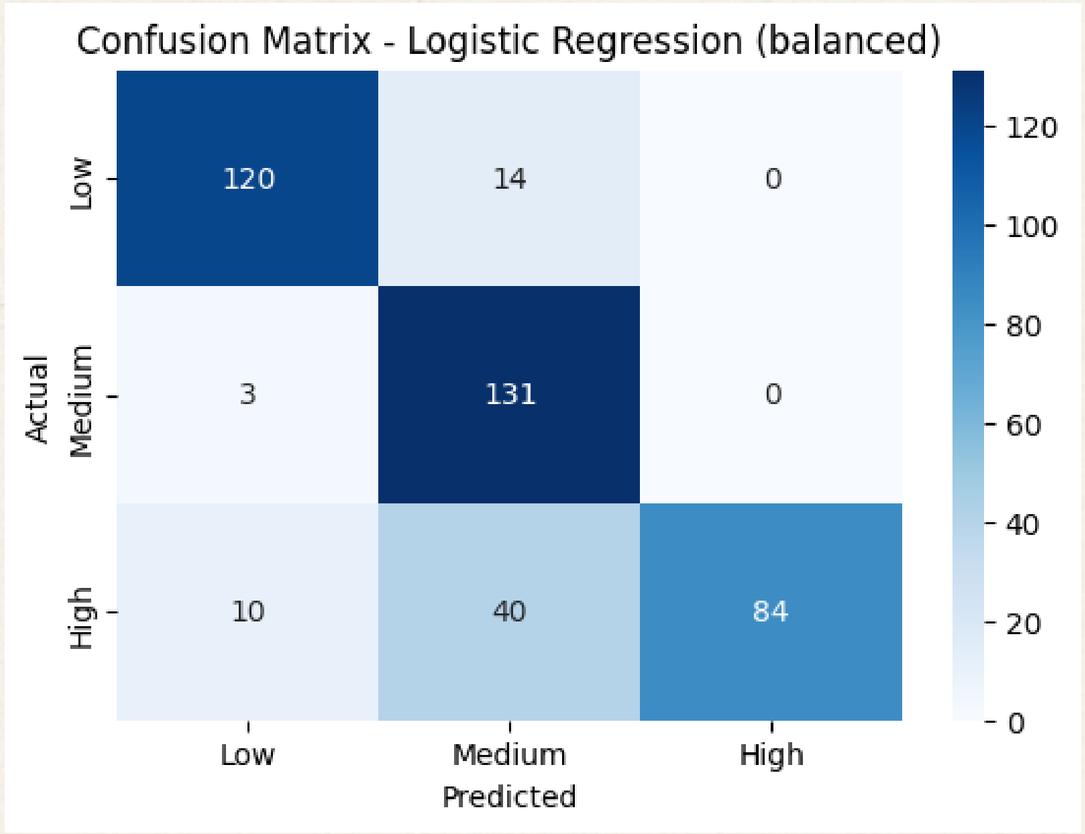
Dilakukan split data untuk train 70% dan test 30%.

Library Model

```
lr_pipeline = Pipeline(steps=[
  ('pre', preprocessor),
  ('clf', LogisticRegression(max_iter=2000,
class_weight='balanced', random_state=42))
])
```

Hasil Model

Confusion Matrix



Evaluation

```
=== Logistic Regression (class_weight='balanced') ===
              precision    recall  f1-score   support

        High     1.0000    0.6269    0.7706       134
         Low     0.9023    0.8955    0.8989       134
      Medium     0.7081    0.9776    0.8213       134

    accuracy                         0.8333       402
   macro avg     0.8701    0.8333    0.8303       402
weighted avg     0.8701    0.8333    0.8303       402
```

Thank you

"Gamma tells the story of time: sometimes short, sometimes long, but always positive." - ChatGPT