An Anova Approach

THE ROLE OF RISK FACTORS IN INSURANCE CLAIM COSTS:

OUR MEMBER

odi ketchiel

vito mungiel

INSURANCE DATASET

- **age** → umur peserta asuransi.
- **sex** → jenis kelamin (male / female).
- **bmi** → Body Mass Index, ukuran berat badan terhadap tinggi.
- **children** → jumlah tanggungan anak.
- **smoker** → status perokok (yes / no).
- **region** → wilayah tempat tinggal (southeast, southwest, dll).
- **charges** → total biaya asuransi kesehatan yang dibebankan.

	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

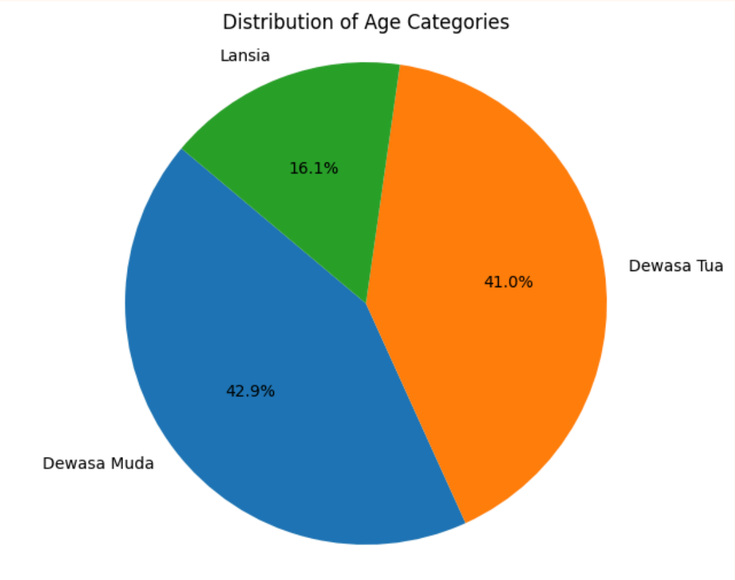
FEATURING ENGINEERING

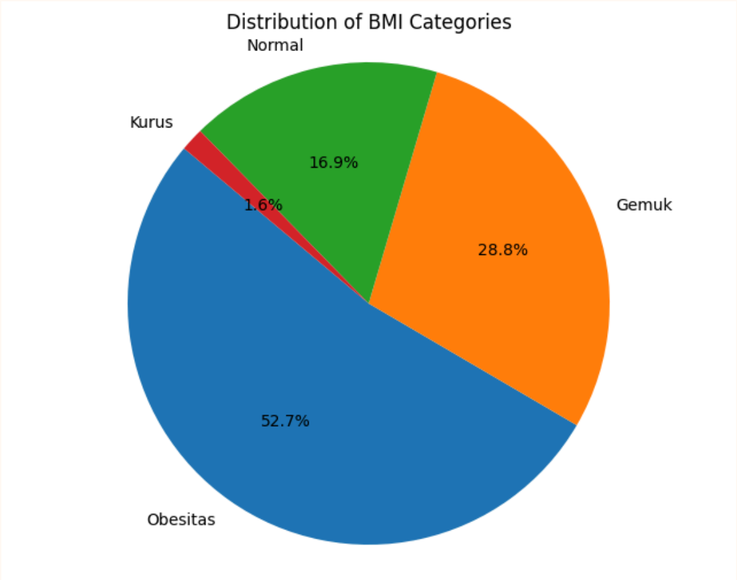
AGE_CATEGORY, BMI_CATEGORY, CHILDREN_CATEGORY

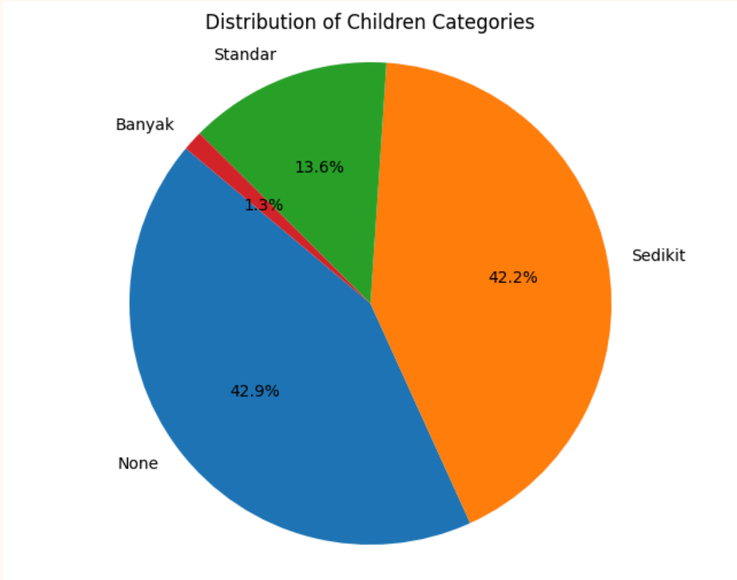
	age	sex	bmi	children	smoker	region	charges	age_category	bmi_category	children_category
0	19	female	27.900	0	1	southwest	16884.92400	Dewasa Muda	Gemuk	None
1	18	male	33.770	1	0	southeast	1725.55230	Dewasa Muda	Obesitas	Sedikit
2	28	male	33.000	3	0	southeast	4449.46200	Dewasa Muda	Obesitas	Standar
3	33	male	22.705	0	0	northwest	21984.47061	Dewasa Muda	Normal	None
4	32	male	28.880	0	0	northwest	3866.85520	Dewasa Muda	Gemuk	None
...	...	...	...	...	...	...	...	...	...	...
1333	50	male	30.970	3	0	northwest	10600.54830	Dewasa Tua	Obesitas	Standar
1334	18	female	31.920	0	0	northeast	2205.98080	Dewasa Muda	Obesitas	None
1335	18	female	36.850	0	0	southeast	1629.83350	Dewasa Muda	Obesitas	None
1336	21	female	25.800	0	0	southwest	2007.94500	Dewasa Muda	Gemuk	None
1337	61	female	29.070	0	1	northwest	29141.36030	Lansia	Gemuk	None

1. **Kolom** `age` → `age_category`
   - Interval usia:
     - `0 – 35` → **Dewasa Muda**
     - `36 – 55` → **Dewasa Tua**
     - `> 55` → **Lansia**
2. **Kolom** `bmi` → `bmi_category`
   - Interval BMI:
     - `0 – 18.5` → **Kurus**
     - `18.5 – 25` → **Normal**
     - `25 – 30` → **Gemuk**
     - `> 30` → **Obesitas**
3. **Kolom** `children` → `children_category`
   - Jumlah anak:
     - `0` → **None**
     - `1 – 2` → **Sedikit**
     - `3 – 4` → **Standar**
     - `≥ 5` → **Banyak**

VISUALIZATION





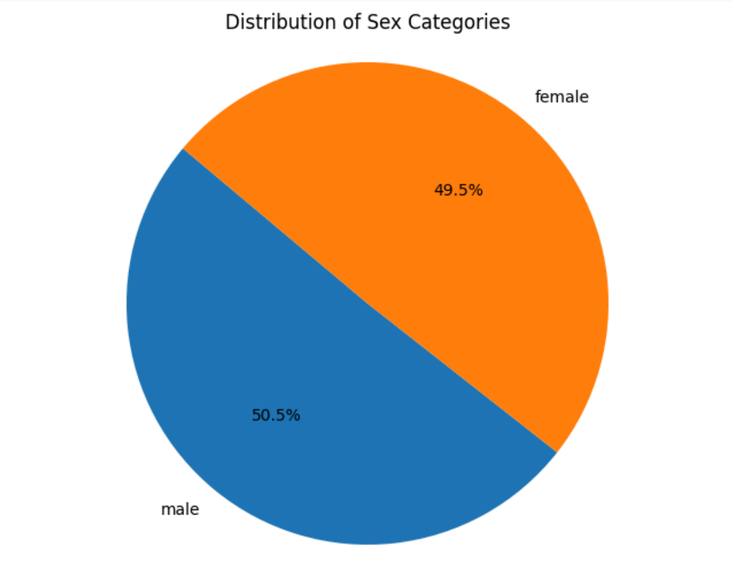


Age_Category

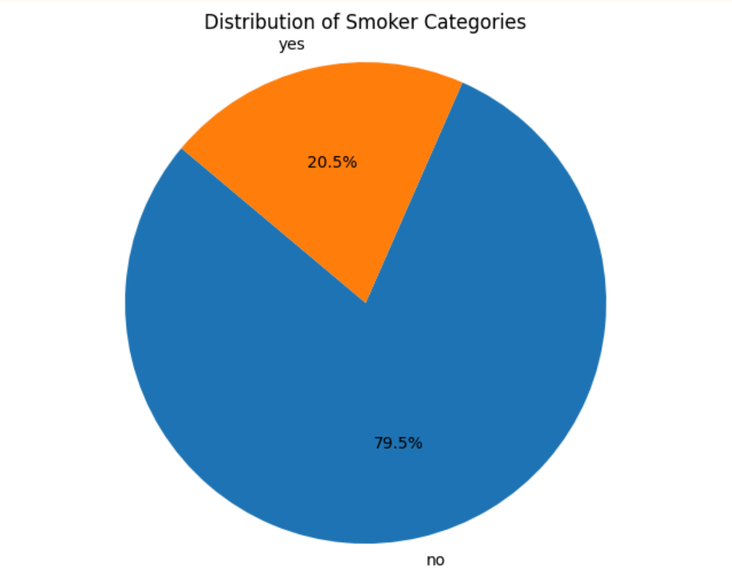
BMI_Category

Children_Category

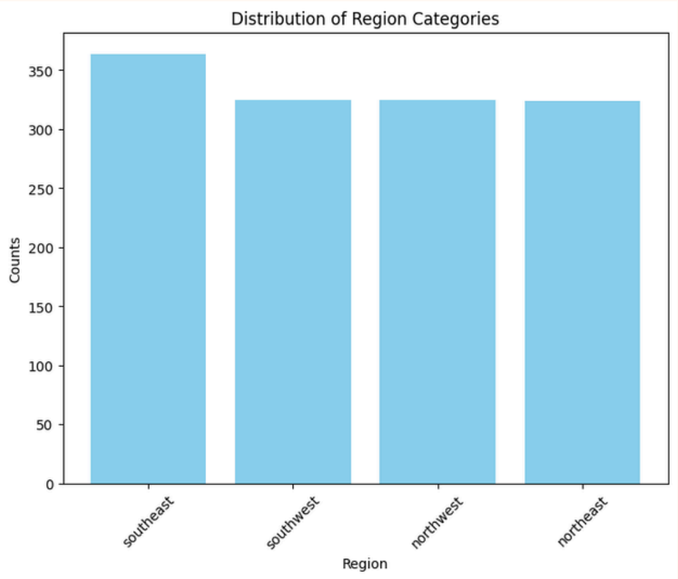
VISUALIZATION



Sex



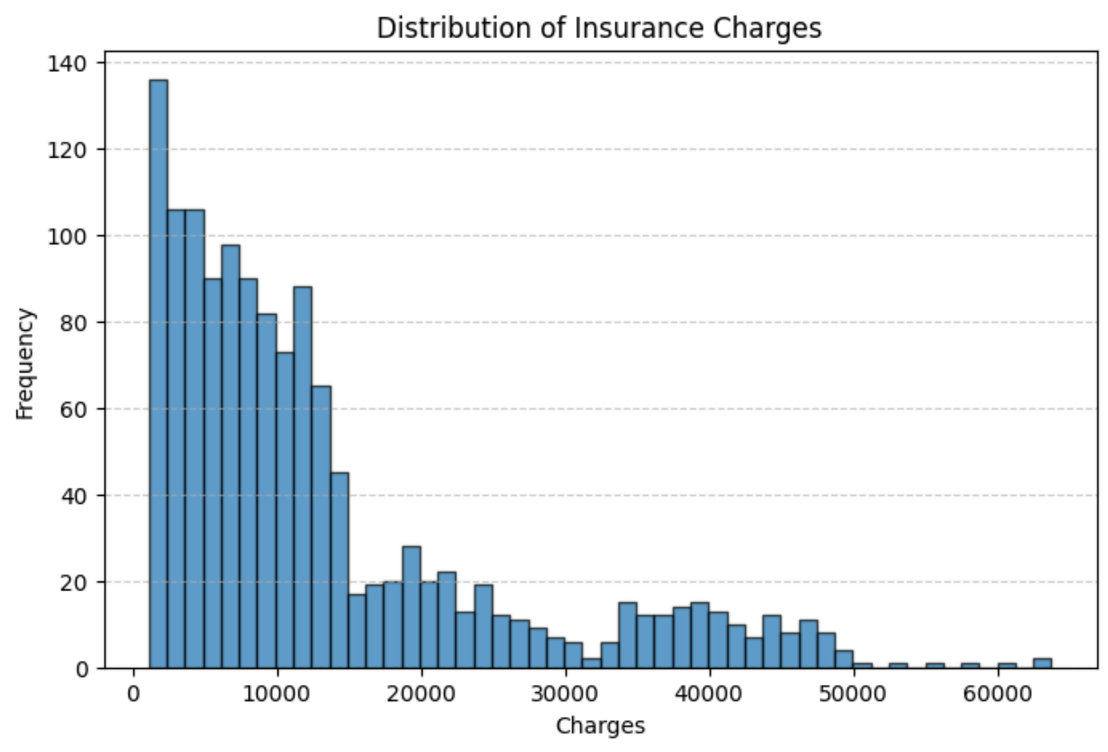
Smoker

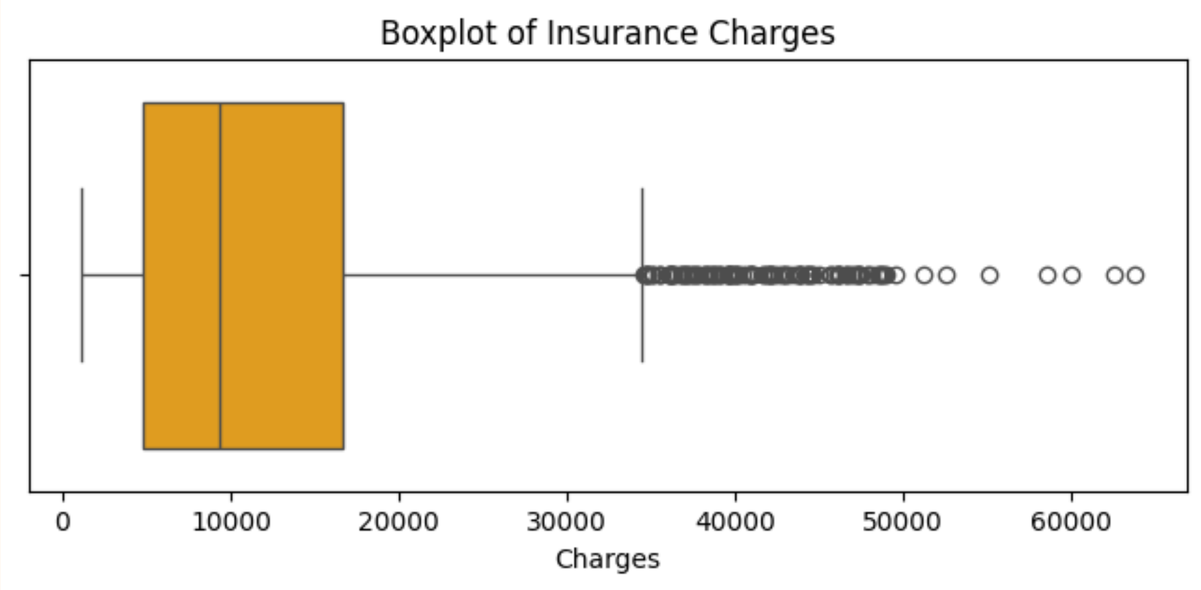


Region

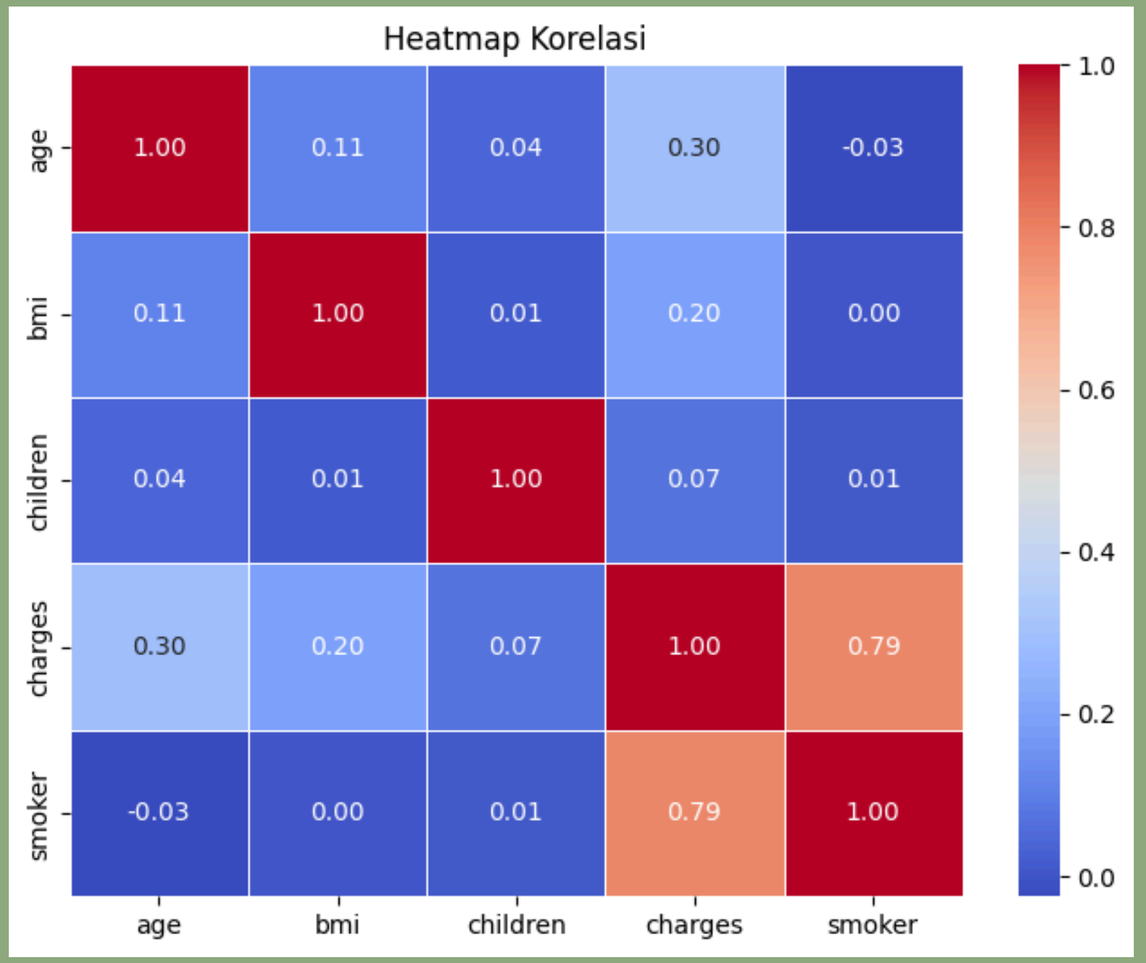
VISUALIZATION

Charges





CORRELATION TEST



ONE-WAY ANOVA

ONE-WAY ANOVA

Bentuk Hipotesis Statistik :

$H_0 : \mu_1 = \mu_2 = ..., = \mu_i (Semua\ perlakuan\ memberikan\ respon\ yang\ sama)$

$H_1 : paling\ sedikit\ satu\ \tan da = tidak\ berlaku$

Statistik Uji

Struktur Tabel ANOVA

Sumber variasi	db	JK	KT	Fhitung
Perlakuan / Antar kelompok *(between)*	t-1	JKP	KTP	KTP/KTG
Galat / Error / Dalam kelompok *(within)*	t(r-1)	JKG	KTG	
Total	tr-1	JKT	-	-

Kriteria Uji

Tolak H_0 jika $F_{hitung} \geq F_{tabel}$ atau $p - value \leq \alpha$

Gunakan tabel distribusi F dengan nilai peluang (1-α) dan derajat kebebasan $db = (\upsilon_1, \upsilon_2)$ dimana υ_1 pembilang dan υ_2 penyebut..

Dimana :

$$FK = {Y_{..}^2}/{tr}$$

$$JKT = \sum_{i=1}^{t} \sum_{j=1}^{r_i} Y_{ij}^2 - FK$$

$$JKP = \sum_{i=1}^{t} \frac{Y_{i.}^2}{r_i} - FK$$

$$JKG = JKT - JKP$$
$$KTP = JKP/(t-1)$$
$$KTG = JKG/(N-k)$$

AGE_CATEGORY - CHARGES

Hipotesis Uji:

H0: Rata-rata biaya asuransi antar kategori usia tidak berbeda signifikan.
H1: Rata-rata biaya asuransi antar kategori usia berbeda signifikan.

Statistik Uji:

```
One-Way ANOVA with age_category as a Factor:
                     sum_sq      df          F        PR(>F)
C(age_category)  1.498926e+10     2.0  55.252143  8.826862e-24
Residual         1.810850e+11  1335.0        NaN           NaN

Hasil ANOVA:
Nilai F-hitung  : 55.2521
Nilai F-kritis  : 3.0025 (alpha=0.05, df1=2, df2=1335)
p-value         : 0.0000
Keputusan       : Tolak H0
Kesimpulan      : Ada perbedaan signifikan dalam biaya asuransi antara kategori usia.
```

SEX - CHARGES

Hipotesis Uji:

H0 = Rata-rata biaya asuransi antara kategori jenis kelamin tidak berbeda signifikan
H1 = Rata-rata biaya asuransi antara kategori jenis kelamin berbeda signifikan .

Statistik Uji:

```
One-Way ANOVA with sex as a Factor:
                sum_sq      df         F    PR(>F)
C(sex)    6.435902e+08     1.0  4.399702  0.036133
Residual  1.954306e+11  1336.0       NaN       NaN

Hasil ANOVA:
Nilai F-hitung  : 4.3997
Nilai F-kritis  : 3.8484 (alpha=0.05, df1=1, df2=1336)
p-value         : 0.0361
Keputusan       : Tolak H0
Kesimpulan      : Ada perbedaan signifikan dalam biaya asuransi antara kategori jenis kelamin.
```

BMI_CATEGORY - CHARGES

Hipotesis Uji:

H0: Rata-rata biaya asuransi antar kategori BMI tidak berbeda signifikan.
H1: Rata-rata biaya asuransi antar kategori BMI berbeda signifikan.

Statistik Uji:

```
One-Way ANOVA with bmi_category as a Factor:
                       sum_sq      df          F        PR(>F)
C(bmi_category)  7.955555e+09     3.0  18.804992  5.997613e-12
Residual         1.881187e+11  1334.0        NaN           NaN

Hasil ANOVA:
Nilai F-hitung  : 18.8050
Nilai F-kritis  : 2.6116 (alpha=0.05, df1=3, df2=1334)
p-value         : 0.0000
Keputusan       : Tolak H0
Kesimpulan      : Ada perbedaan signifikan dalam biaya asuransi antara kategori BMI.
```

CHILDREN_CATEGORY - CHARGES

Hipotesis Uji:

H0: Rata-rata biaya asuransi antar kategori jumlah anak tidak berbeda signifikan.
H1: Rata-rata biaya asuransi antar kategori jumlah anak berbeda signifikan.

Statistik Uji:

```
One-Way ANOVA with children_category as a Factor:
                          sum_sq      df         F    PR(>F)
C(children_category)  1.591613e+09     3.0  3.639078  0.012411
Residual              1.944826e+11  1334.0       NaN       NaN

Hasil ANOVA:
Nilai F-hitung  : 3.6391
Nilai F-kritis  : 2.6116 (alpha=0.05, df1=3, df2=1334)
p-value         : 0.0124
Keputusan       : Tolak H0
Kesimpulan      : Ada perbedaan signifikan dalam biaya asuransi antara kategori anak.
```

SMOKER- CHARGES

Hipotesis Uji:

H0: Rata-rata biaya asuransi antar kategori perokok tidak berbeda signifikan.
H1: Rata-rata biaya asuransi antar kategori perokok berbeda signifikan.

Statistik Uji:

```
One-Way ANOVA with smoker as a Factor:
                sum_sq      df            F         PR(>F)
C(smoker)  1.215199e+11     1.0  2177.614868  8.271436e-283
Residual   7.455432e+10  1336.0          NaN            NaN

Hasil ANOVA:
Nilai F-hitung  : 2177.6149
Nilai F-kritis  : 3.8484 (alpha=0.05, df1=1, df2=1336)
p-value         : 0.0000
Keputusan       : Tolak H0
Kesimpulan      : Ada perbedaan signifikan dalam biaya asuransi antara kategori perokok.
```

REGION - CHARGES

Hipotesis Uji:

H0: Rata-rata biaya asuransi antar kategori region tidak berbeda signifikan.
H1: Rata-rata biaya asuransi antar kategori region berbeda signifikan.

Statistik Uji:

```
One-Way ANOVA with region as a Factor:
                sum_sq      df         F    PR(>F)
C(region)  1.300760e+09     3.0  2.969627  0.030893
Residual   1.947735e+11  1334.0       NaN       NaN

Hasil ANOVA:
Nilai F-hitung  : 2.9696
Nilai F-kritis  : 2.6116 (alpha=0.05, df1=3, df2=1334)
p-value         : 0.0309
Keputusan       : Tolak H0
Kesimpulan      : Ada perbedaan signifikan dalam biaya asuransi antara kategori region.
```

TWO-WAYS ANOVA

TWO-WAYS ANOVA

Statistik Uji :

Struktur Tabel Anova

Sumber Variasi	db	JK	KT	F hitung
A B AB Galat/error dlm kelompok	(a-1) (b-1) (a-1)(b-1) ab(c-1)	JKA JKB JKAB JKG	KTA KTB KTAB KTG	KTA/KTG KTB/KTG KTAB/KTG -
Total	abc-1	JKT	-	-

Kriteria Uji

Tolak H_0 jika $F_{hitung} \geq F_{tabel}$ atau $p - value \leq \alpha$

Gunakan tabel distribusi F dengan nilai peluang (1-α) dan derajat kebebasan $db = (\upsilon_1, \upsilon_2)$ dimana υ_1 db pembilang dan υ_2 db penyebut..

dimana

$$FK = \frac{Y_{...}^{\ 2}}{abr}$$

$$JKT = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r}\left(Y_{ijk} - \bar{Y}_{...}\right)^2 = \sum\sum\sum Y_{ijk}^{\ 2} - FK$$

$$JKA = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r}\left(\bar{Y}_{i..} - \bar{Y}_{...}\right)^2 = \sum \frac{Y_{i..}^{\ 2}}{br} - FK \qquad JKB = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r}\left(\bar{Y}_{.j.} - \bar{Y}_{...}\right)^2 = \sum \frac{Y_{.j.}^{\ 2}}{ar} - FK$$

$$JKAB = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r}\left(\bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...}\right)^2 = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r}\left(\bar{Y}_{ij.} - \bar{Y}_{...}\right)^2 - JKA - JKB$$

$$JKAB = JKP - JKA - JKB$$

$$JKP = \sum\sum\sum\left(\bar{Y}_{ij.} - \bar{Y}_{...}\right)^2 = \sum\sum \frac{Y_{ij.}^{\ 2}}{r} - FK$$

$$JKG = JKT - JKP$$

AGE_CATEGORY, SMOKER - CHARGES

Hipotesis Uji:

H0: Rata-rata biaya asuransi tidak berbeda signifikan menurut kategori usia dan perokok.
H1: Rata-rata biaya asuransi berbeda signifikan menurut kategori usia dan perokok.

Statistik Uji:

Two-Way ANOVA with age_category and smoker as Factors:

	sum_sq	df	F	PR(>F)
C(age_category)	1.667833e+10	2.0	192.305426	4.249708e-74
C(smoker)	1.232090e+11	1.0	2841.262866	0.000000e+00
C(age_category):C(smoker)	1.149365e+08	2.0	1.325247	2.660874e-01
Residual	5.776106e+10	1332.0	NaN	NaN

```
=== Uji untuk C(age_category):C(smoker) ===
F-hitung : 1.3252
F-kritis : 3.0025 (alpha=0.05, df1=2, df2=1332)
p-value  : 0.266087
Keputusan: Gagal Tolak H0
```

SMOKER, BMI_CATEGORY- CHARGES

Hipotesis Uji:

H0: Rata-rata biaya asuransi tidak berbeda signifikan menurut kategori perokok dan BMI.
H1: Rata-rata biaya asuransi berbeda signifikan menurut kategori perokok dan BMI.

Statistik Uji:

Two-Way ANOVA with bmi_category and smoker as Factors:

	sum_sq	df	F	PR(>F)
C(bmi_category)	8.283097e+09	3.0	80.234289	1.040883e-47
C(smoker)	1.218474e+11	1.0	3540.829000	0.000000e+00
C(bmi_category):C(smoker)	2.050310e+10	3.0	198.603422	2.011173e-106
Residual	4.576812e+10	1330.0	NaN	NaN

```
=== Uji untuk C(bmi_category):C(smoker) ===
F-hitung : 198.6034
F-kritis : 2.6116 (alpha=0.05, df1=3, df2=1330)
p-value  : 0.000000
Keputusan: Tolak H0
```

AGE_CATEGORY, BMI_CATEGORY- CHARGES

Hipotesis Uji:

H0: Rata-rata biaya asuransi tidak berbeda signifikan menurut kategori usia dan BMI.
H1: Rata-rata biaya asuransi berbeda signifikan menurut kategori usia dan BMI.

Statistik Uji:

Two-Way ANOVA with age_category and smoker as Factors:

	sum_sq	df	F	PR(>F)
C(age_category)	1.667833e+10	2.0	192.305426	4.249708e-74
C(smoker)	1.232090e+11	1.0	2841.262866	0.000000e+00
C(age_category):C(smoker)	1.149365e+08	2.0	1.325247	2.660874e-01
Residual	5.776106e+10	1332.0	NaN	NaN

```
=== Uji untuk C(age_category):C(smoker) ===
F-hitung : 1.3252
F-kritis : 3.0025 (alpha=0.05, df1=2, df2=1332)
p-value  : 0.266087
Keputusan: Gagal Tolak H0
```

MULTI-WAYS ANOVA

AGE_CATEGORY, SMOKER, BMI_CATEGORY - CHARGES

Hipotesis Uji:

H0: Rata-rata biaya asuransi tidak berbeda signifikan menurut kategori usia, status perokok, dan BMI.
H1: Rata-rata biaya asuransi berbeda signifikan menurut kategori usia, status perokok, dan BMI.

Statistik Uji:

Multi-Way ANOVA (age_category, bmi_category, smoker):

	sum_sq	df	F	PR(>F)
age_category	1.529061e+10	2.0	330.749699	4.381382e-117
bmi_category	6.516567e+09	3.0	93.972852	4.186626e-55
smoker	1.234517e+11	1.0	5340.743699	0.000000e+00
age_category:bmi_category	5.294714e+07	6.0	0.381765	8.909735e-01
age_category:smoker	2.560787e+06	2.0	0.055392	9.461163e-01
bmi_category:smoker	2.073358e+10	3.0	298.990843	5.849836e-148
age_category:bmi_category:smoker	2.633549e+07	6.0	0.189887	9.797189e-01
Residual	3.039632e+10	1315.0	NaN	NaN

```
=== Uji untuk age_category:bmi_category:smoker ===
F-hitung : 0.1899
F-kritis : 2.1055 (alpha=0.05, df1=6, df2=1315)
p-value  : 0.979719
Keputusan: Gagal Tolak H0
```

SUMMARY

1. Berdasarkan hasil statistik uji untuk One-Way ANOVA, variabel charges memiliki perbedaan rata-rata secara signifikan terhadap semua variabel kategori (menolak H0).

2. Berdasarkan hasil statistik uji untuk Two-Ways ANOVA, variabel kategori smoker dan bmi memiliki perbedaan rata-rata secara signifikan terhadap charges (menolak H0)

3. Berdasarkan hasil statistik uji untuk Multi-Way ANOVA, variabel kategori age, smoker, dan bmi tidak terdapat perbedaan rata-rata secara signifikan terhadap charges (menerima H0)

THANK YOU

“If life is full of differences, don’t worry.
ANOVA will tell you if they’re truly significant.” - ChatGPT