Data Cleaning

Raw data

	Dress_ID	Style	Price	Rating	Size	Season	NeckLine	SleeveLength	waiseline	Material	FabricType	Decoration	Pattern Type	Recommendation
0	1006032852	Sexy	Low	4.6	M	Summer	o-neck	sleevless	empire	NaN	chiffon	ruffles	animal	1
1	1212192089	Casual	Low	0.0	L	Summer	o-neck	Petal	natural	microfiber	NaN	ruffles	animal	0
2	1190380701	vintage	High	0.0	L	Automn	o-neck	full	natural	polyster	NaN	NaN	print	0
3	966005983	Brief	Average	4.6	L	Spring	o-neck	full	natural	silk	chiffon	embroidary	print	1
4	876339541	cute	Low	4.5	M	Summer	o-neck	butterfly	natural	chiffonfabric	chiffon	bow	dot	0

Data exploration

Dress_ID column not relevant → Remove

Correct typos → "waiseline" to "WaistLine", etc.

Uniform characters → Lowercase categorical columns

NaN values

```
df.isnull().sum(axis=0)
```

```
Style              0
Price              2
Rating             0
Size               0
Season             1
NeckLine           3
SleeveLength       1
waiseline         66
Material         101
FabricType       203
Decoration         0
Pattern Type      81
Recommendation     0
dtype: int64
```

WaistLine, Material, FabricType, PatternType → 66-203 missing → In 500 rows → Test model scores with or without these columns

Price, Season, NeckLine, SleeveLength → 1-3 missing → Fill with most common values

Treating each column

Rating → Ratings can't be zero → Remove

Price → "Average" to "Medium"

SleeveLength →

```
("cap-sleeves", "capsleeves")
("half", "halfsleeve")
("sleveless", "sleeveless")
("sleevless", "sleeveless")
("sleeevless", "sleeveless")
("sleeevless", "sleeveless")
("thressqatar", "threequarter")
("urndowncollor", "turndowncollar")
("turndowncollor", "turndowncollar")
("leapord", "leopard")
```

Reduce Levels

Style →

```
("sexy", "special_occasion")
("party", "special_occasion")
("novelty", "other")
("bohemian", "boho_vintage")
("vintage", "boho_vintage")
("fashion", "other")
("brief", "casual")
("cute", "casual")
```

NeckLine →

```
("bowneck", "bow_and_ruffle")
("ruffled", "bow_and_ruffle")
("slash-neck", "slash_and_boat")
("boat-neck", "slash_and_boat")
("sweetheart", "sweetheart_and_square")
("sqare-collor", "sweetheart_and_square")
("backless", "other")
("peterpan-collor", "other")
("halter", "other")
("open", "other")
("scoop", "other")
```

SleeveLength →

```
("turndowncollar", "full")
("butterfly", "halfsleeve")
("capsleeves", "short")
```

Material →

```
("chiffonfabric", "chiffon")
("shiffon", "chiffon")
("milksilk", "silk")
("model", "modal")
("wool", "wool_and_cashmere")
("cashmere", "wool_and_cashmere")
("knitting", "knit")
("spandex", "knit")
("lycra", "knit")
("modal", "cotton")
```

.....