Project Perfume

Chris Chen, Peter Liu, Yueqi Xu

Research Questions

- Does the prestige of the **brands** contribute to the most of the perfume’s retail **price**?
- Does the **customer rating**, **seller**, and **seller rating** contribute to the perfume’s retail **price**?
- Does the perfume's **department** (female, male, or unisex), **scents** (woody, floral, fruity, etc.), and **notes** potentially leads to difference in retail prices?

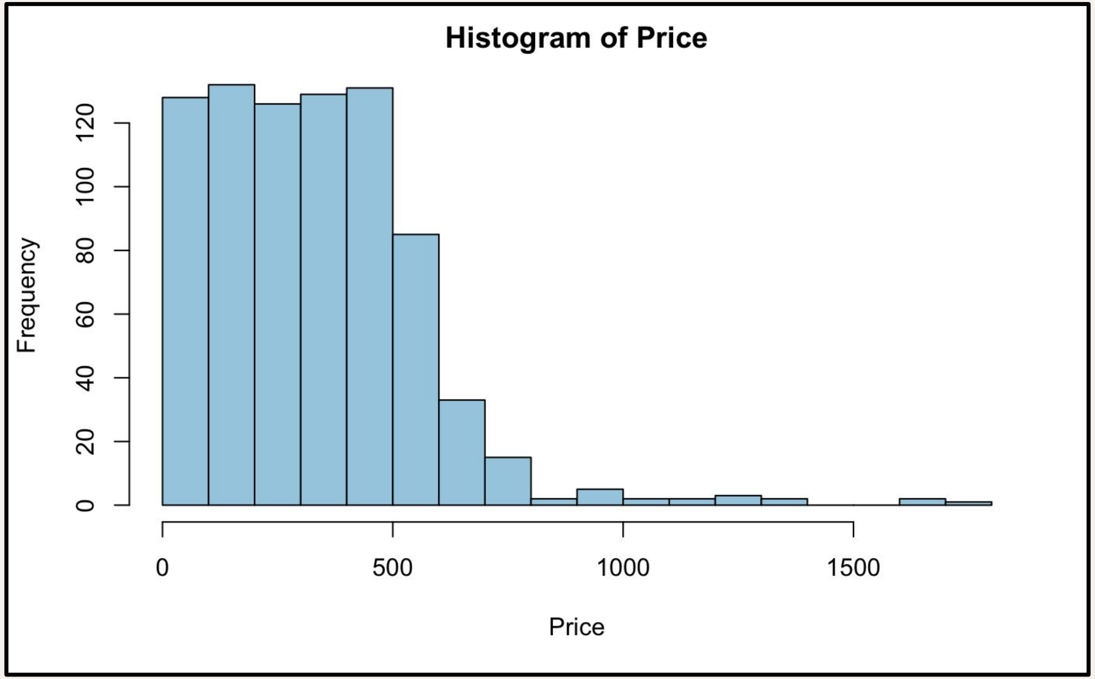
Response and Predictors

Response

- Price

Predictors

- Brand
- Volume
- Concentration
- Department
- Scent
- Base note
- Middle note
- Item rating
- Seller
- Seller rating
- # of seller rating

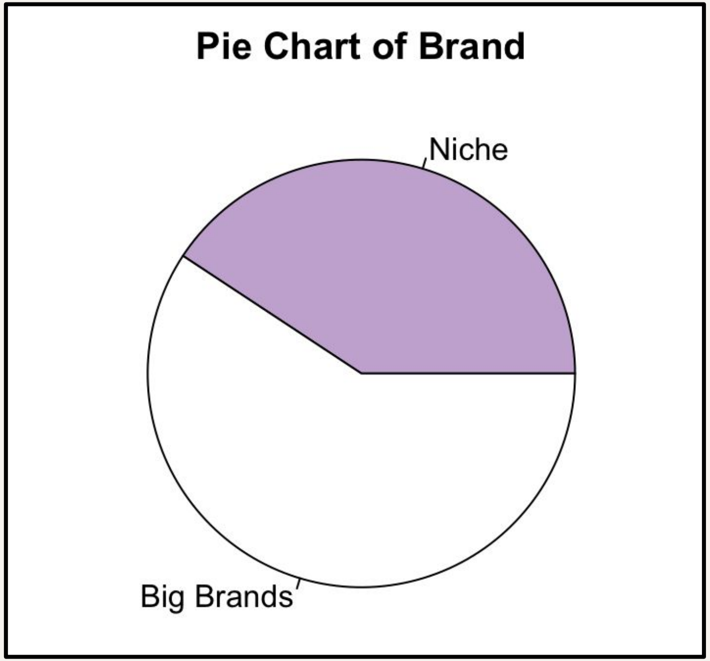


Data cleaning: Remove Nonsense

- The data contains no NA, which is good!
- Lots of typos
- Characters in other languages (Arabian, Latin, etc.)
- Nonsense induced by web page crawling. For instance, considering *Yves Saint Laurent* (A famous cosmetic brand, often abbreviated into YSL) into three different brands.
- 889 observations after removing nonsenses

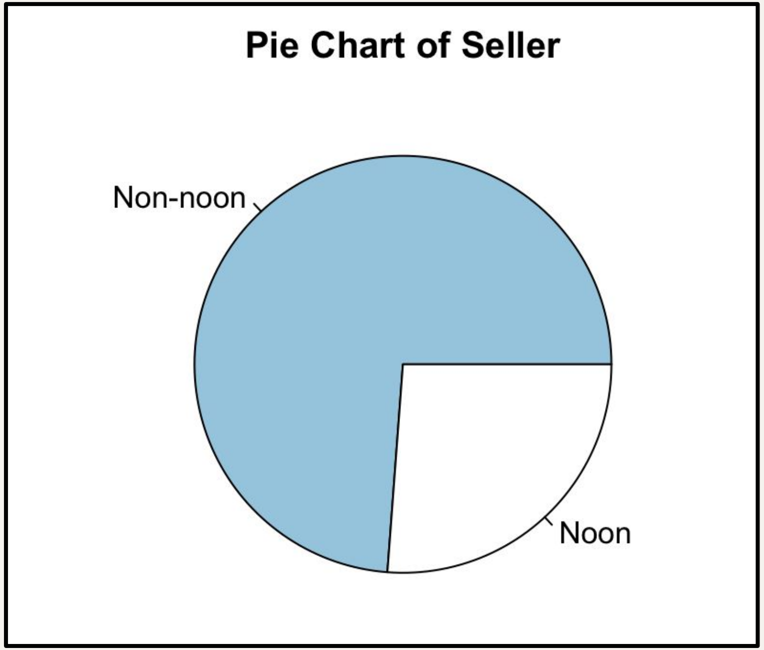
Data Cleaning: Merging Categorical Covariates

- Covariate **Brand** originally contain 148 levels.
- We converted it into a categorical variable called **big_brand** with 2 levels:
- **1 (Big Brand)**: brand that contains more than 10 listed individual perfumes;
- **0 (Niche brand)**: otherwise.



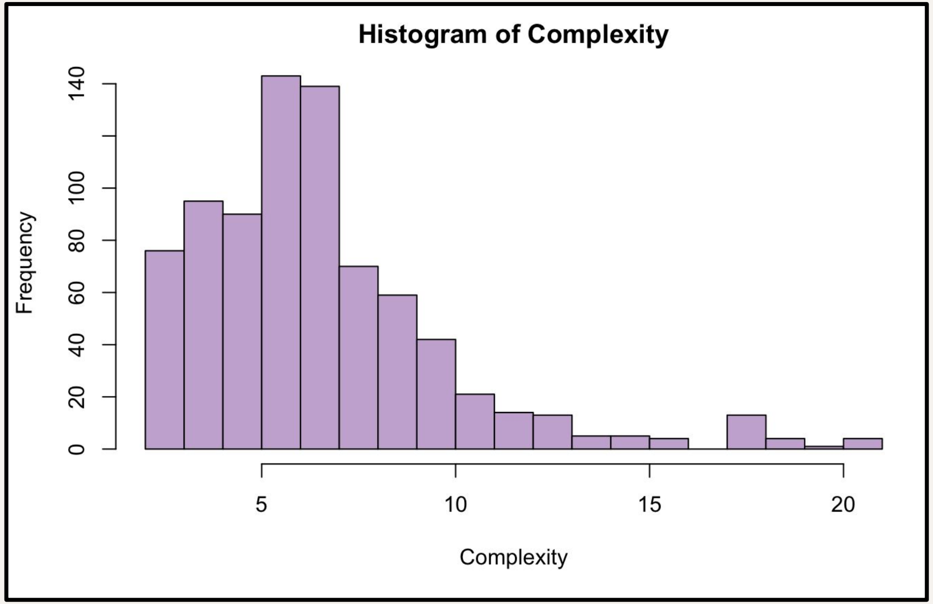
Data Cleaning: Merging Categorical Covariates

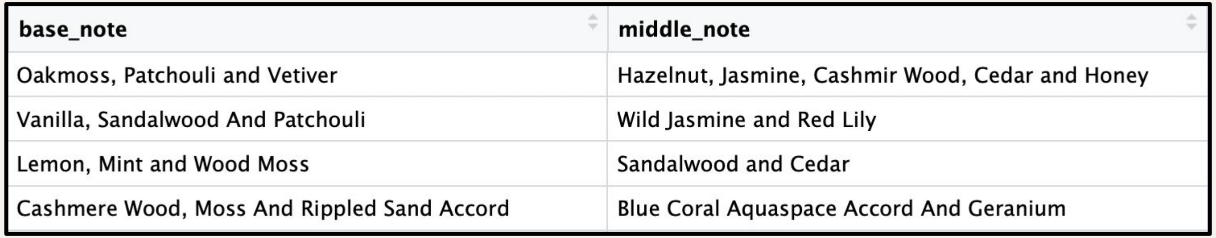
- Covariate **Seller** originally contains 115 levels.
- We converted it into a categorical variable called **is_noon** with two levels:
- **1 (Noon)**: the perfume is sold by noon official
- **0 (Non-noon)**: the perfume is sold by individual sellers



Data Cleaning

Base Notes & **Middle Notes**: We counted the total number of distinct notes in base notes and middle notes, and stored this information in a new variable called **Complexity**.

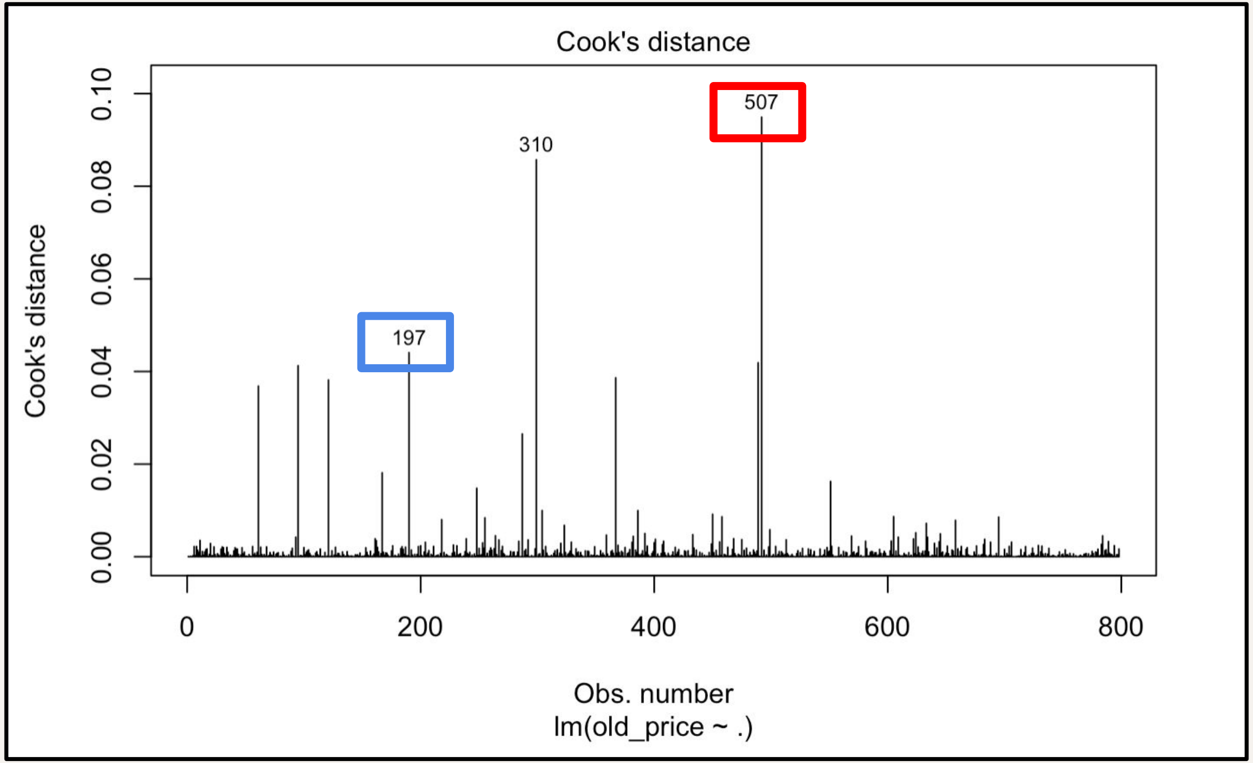


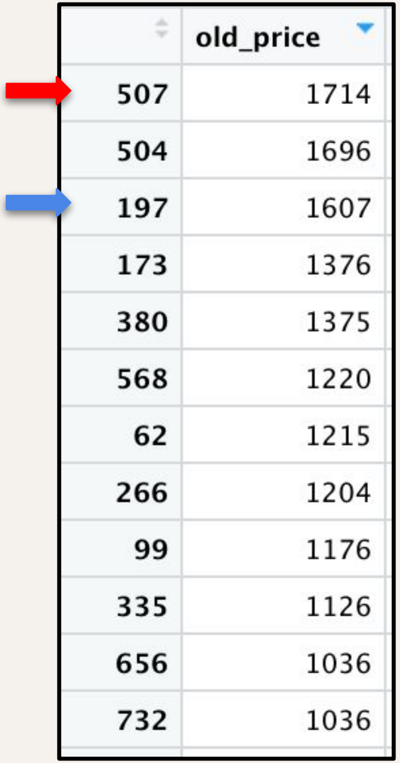


base_note	middle_note
Oakmoss, Patchouli and Vetiver	Hazelnut, Jasmine, Cashmir Wood, Cedar and Honey
Vanilla, Sandalwood And Patchouli	Wild Jasmine and Red Lily
Lemon, Mint and Wood Moss	Sandalwood and Cedar
Cashmere Wood, Moss And Rippled Sand Accord	Blue Coral Aquaspace Accord And Geranium

Data Modeling: Outliers

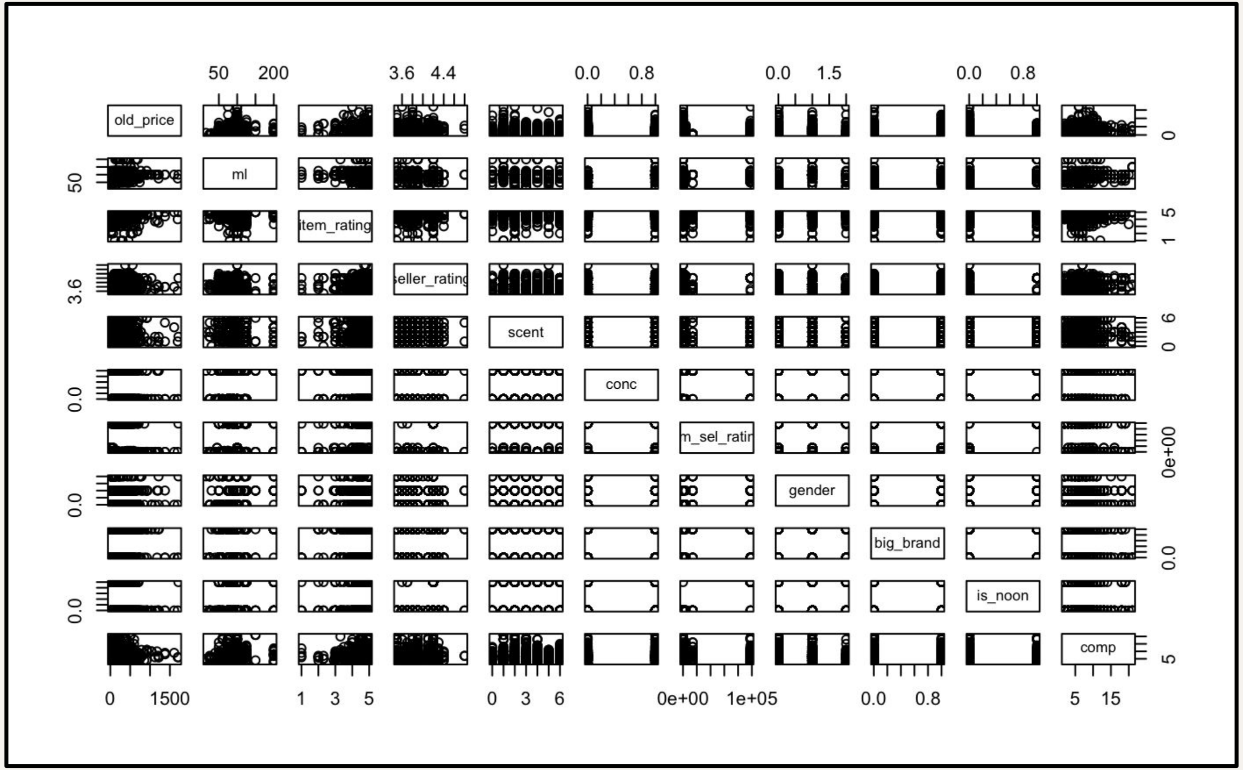
lm(old_price ~ . , data = perfume)



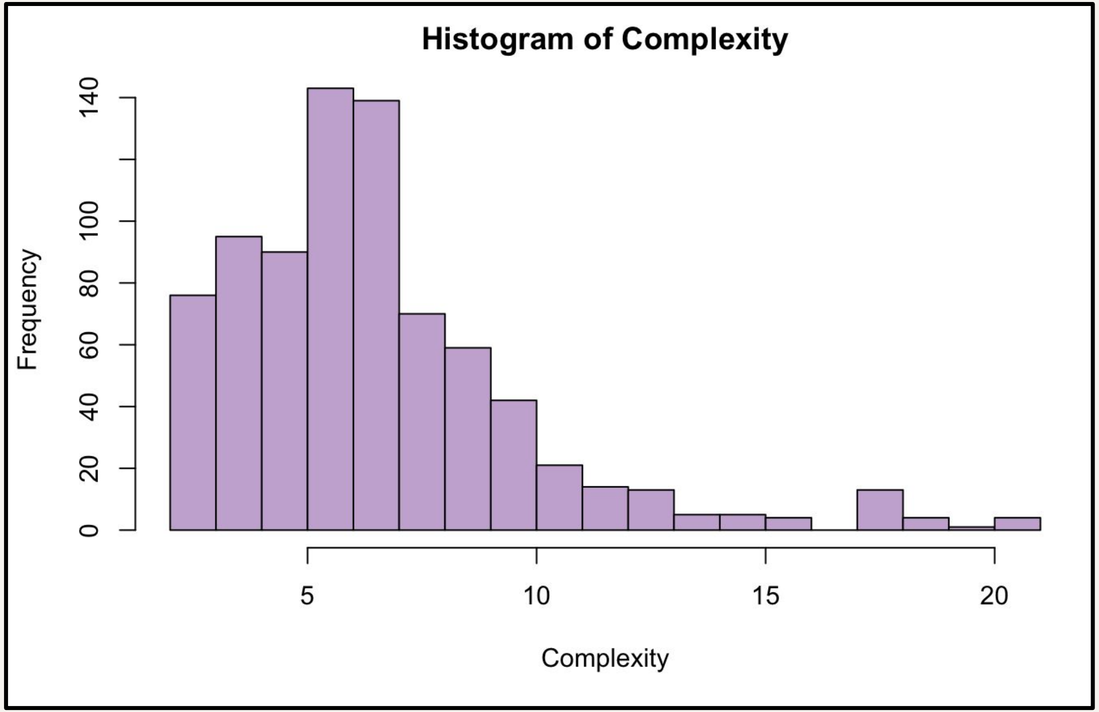


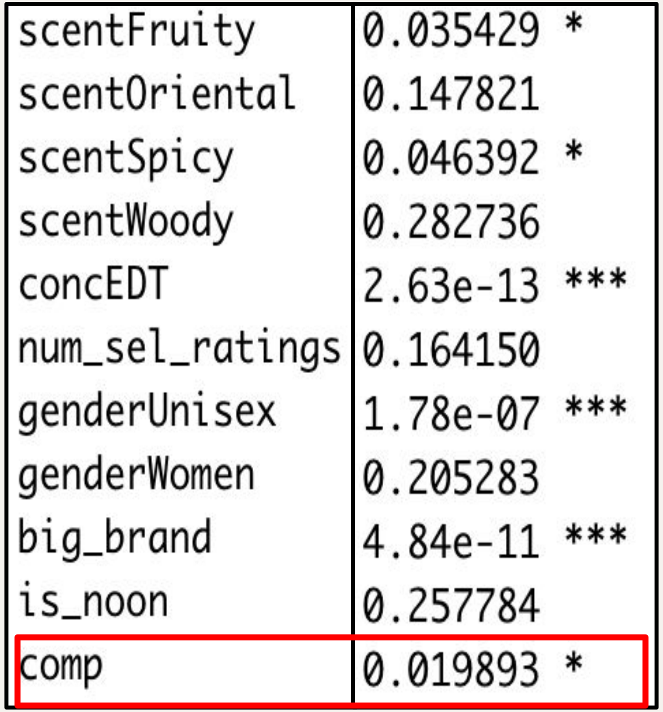
	old_price
507	1714
504	1696
197	1607
173	1376
380	1375
568	1220
62	1215
266	1204
99	1176
335	1126
656	1036
732	1036

Data Modeling: Interaction



Data Modeling: Interaction

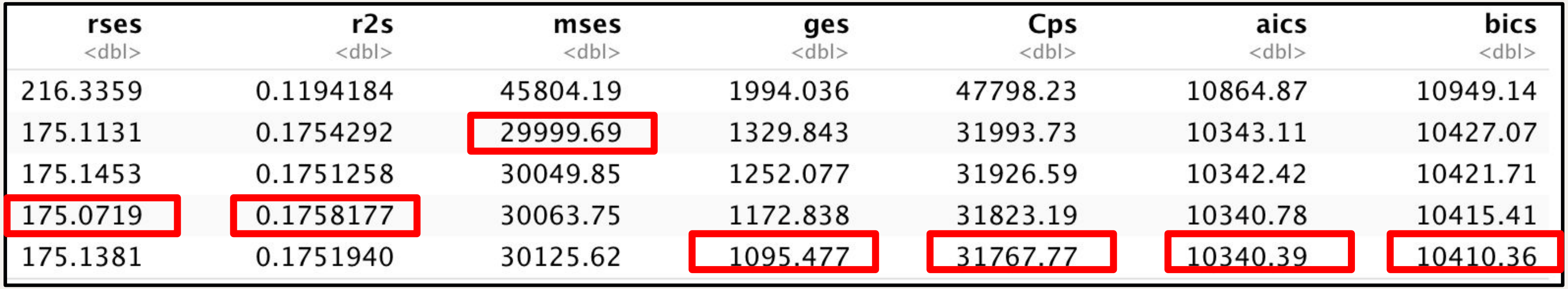




```
scentFruity       0.035429 *
scentOriental     0.147821
scentSpicy        0.046392 *
scentWoody        0.282736
concEDT           2.63e-13 ***
num_sel_ratings   0.164150
genderUnisex      1.78e-07 ***
genderWomen       0.205283
big_brand         4.84e-11 ***
is_noon           0.257784
comp              0.019893 *
```

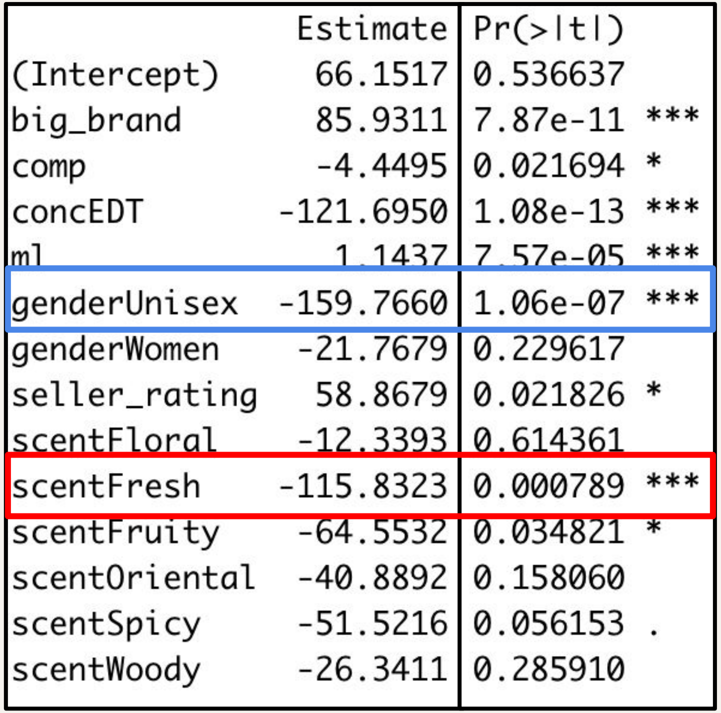
Model Selection: First Round

lm.1 = lm(old_price ~ ., data = perfume_original)
lm.2 = lm(old_price ~ ., data = perfume)
lm.3 = lm(old_price ~ . - is_noon, data = perfume)
lm.4 = lm(old_price ~ . - is_noon - item_rating, data = perfume)
lm.5 = lm(old_price ~ . - is_noon - item_rating - num_sel_ratings, data = perfume)

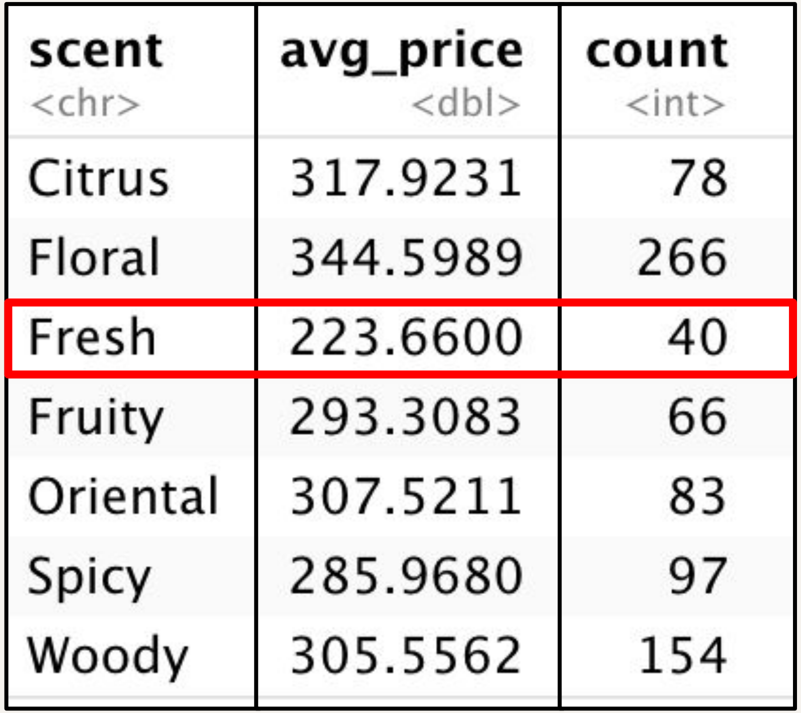


rses <dbl>	r2s <dbl>	mses <dbl>	ges <dbl>	Cps <dbl>	aics <dbl>	bics <dbl>
216.3359	0.1194184	45804.19	1994.036	47798.23	10864.87	10949.14
175.1131	0.1754292	29999.69	1329.843	31993.73	10343.11	10427.07
175.1453	0.1751258	30049.85	1252.077	31926.59	10342.42	10421.71
175.0719	0.1758177	30063.75	1172.838	31823.19	10340.78	10415.41
175.1381	0.1751940	30125.62	1095.477	31767.77	10340.39	10410.36

Model Selection: Merging Scent/Gender

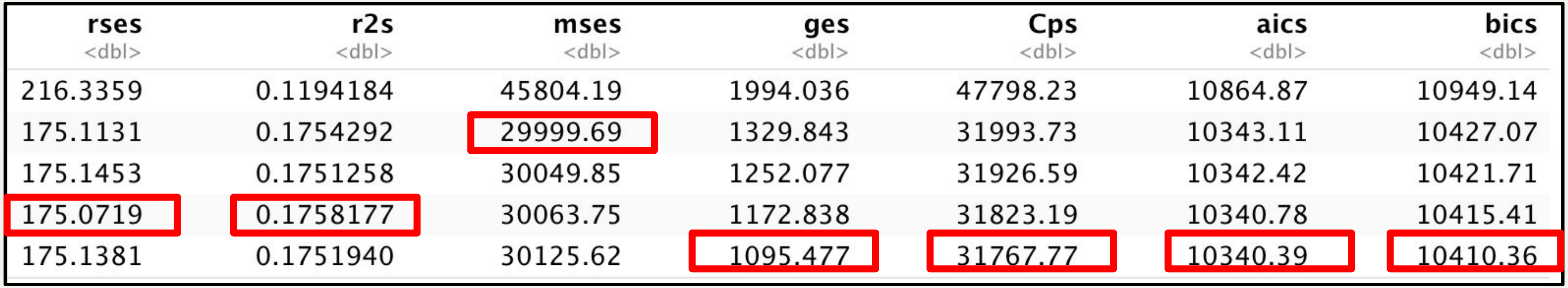


```
                Estimate  Pr(>|t|)
(Intercept)      66.1517  0.536637
big_brand        85.9311  7.87e-11 ***
comp             -4.4495  0.021694 *
concEDT        -121.6950  1.08e-13 ***
ml                1.1437  7.57e-05 ***
genderUnisex   -159.7660  1.06e-07 ***
genderWomen     -21.7679  0.229617
seller_rating    58.8679  0.021826 *
scentFloral     -12.3393  0.614361
scentFresh     -115.8323  0.000789 ***
scentFruity     -64.5532  0.034821 *
scentOriental   -40.8892  0.158060
scentSpicy      -51.5216  0.056153 .
scentWoody      -26.3411  0.285910
```

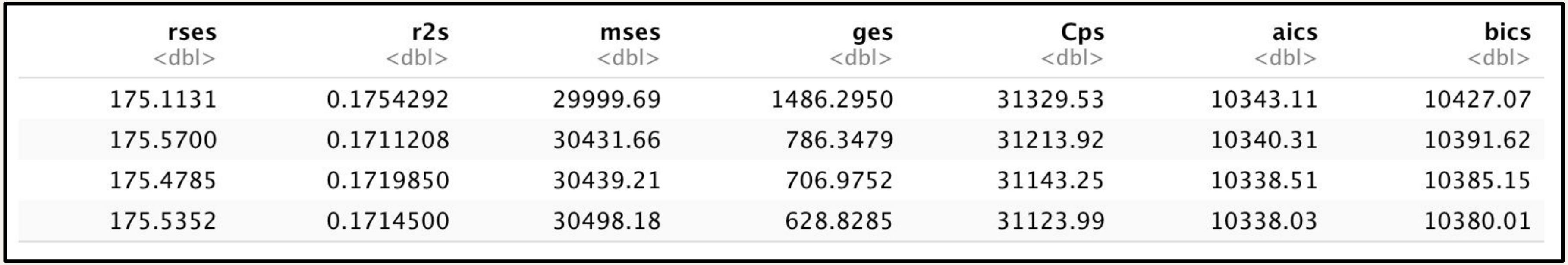


scent <chr>	avg_price <dbl>	count <int>
Citrus	317.9231	78
Floral	344.5989	266
Fresh	223.6600	40
Fruity	293.3083	66
Oriental	307.5211	83
Spicy	285.9680	97
Woody	305.5562	154

Model Selection: Second Round



rses <dbl>	r2s <dbl>	mses <dbl>	ges <dbl>	Cps <dbl>	aics <dbl>	bics <dbl>
216.3359	0.1194184	45804.19	1994.036	47798.23	10864.87	10949.14
175.1131	0.1754292	29999.69	1329.843	31993.73	10343.11	10427.07
175.1453	0.1751258	30049.85	1252.077	31926.59	10342.42	10421.71
175.0719	0.1758177	30063.75	1172.838	31823.19	10340.78	10415.41
175.1381	0.1751940	30125.62	1095.477	31767.77	10340.39	10410.36



rses <dbl>	r2s <dbl>	mses <dbl>	ges <dbl>	Cps <dbl>	aics <dbl>	bics <dbl>
175.1131	0.1754292	29999.69	1486.2950	31329.53	10343.11	10427.07
175.5700	0.1711208	30431.66	786.3479	31213.92	10340.31	10391.62
175.4785	0.1719850	30439.21	706.9752	31143.25	10338.51	10385.15
175.5352	0.1714500	30498.18	628.8285	31123.99	10338.03	10380.01

Final Model

Price = 23.34

+ 85.56·big_brand

- 111.02·I(concentration = EDT)

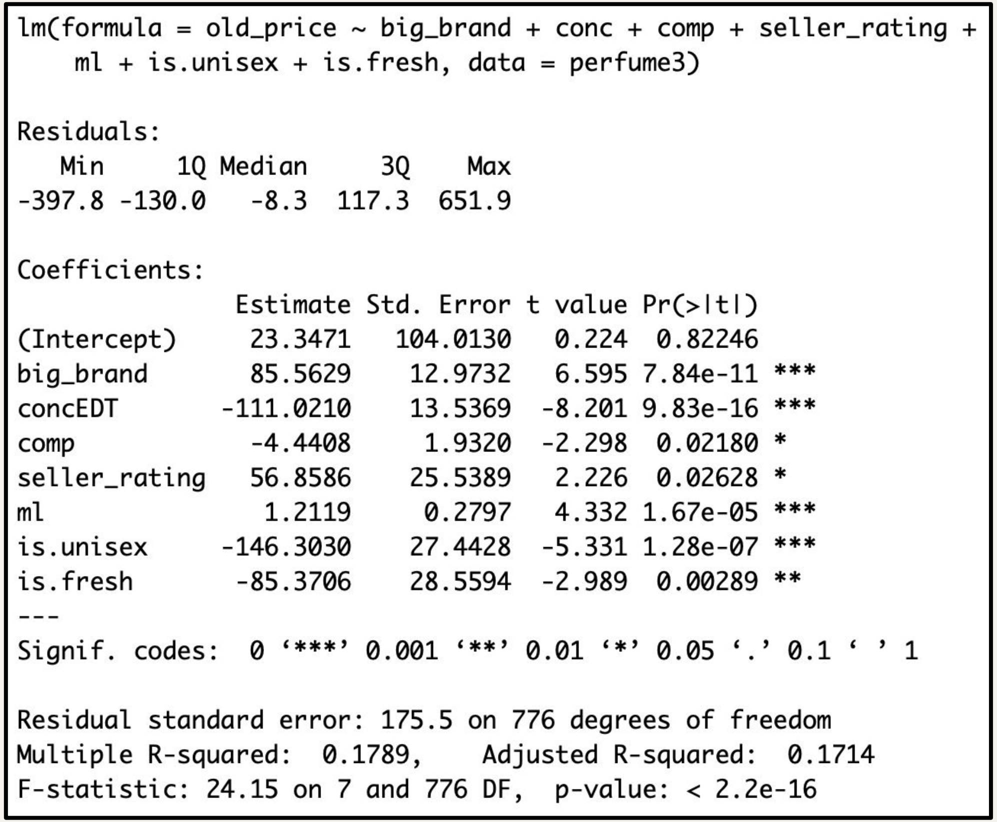
- 4.44·comp

+ 56.85·seller_rating

+ 1.21·volume

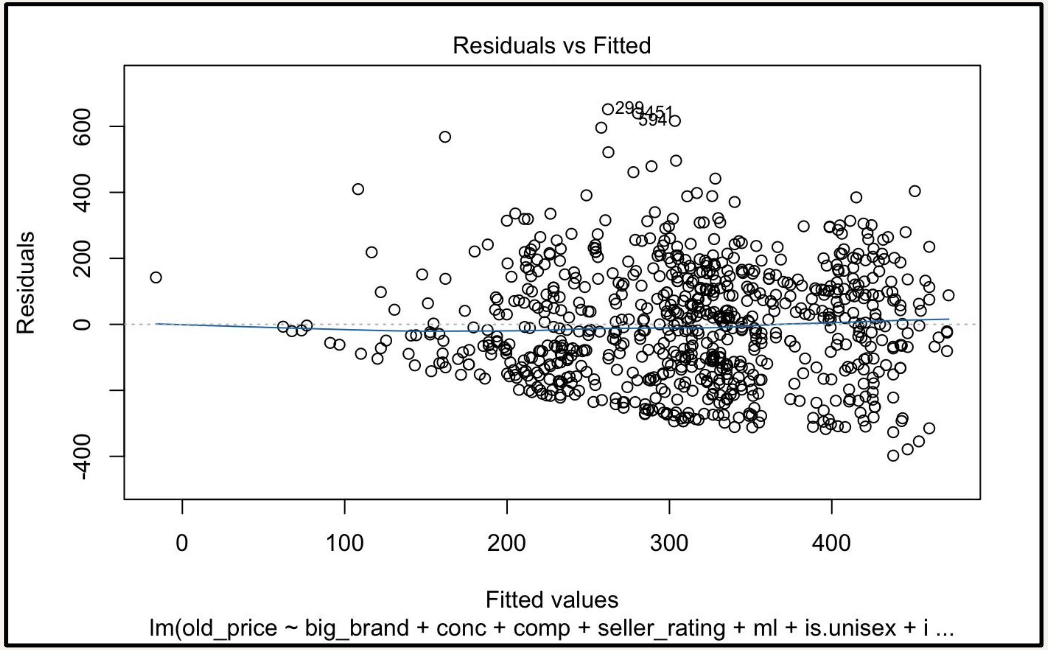
- 146.3·I(gender = Unisex)

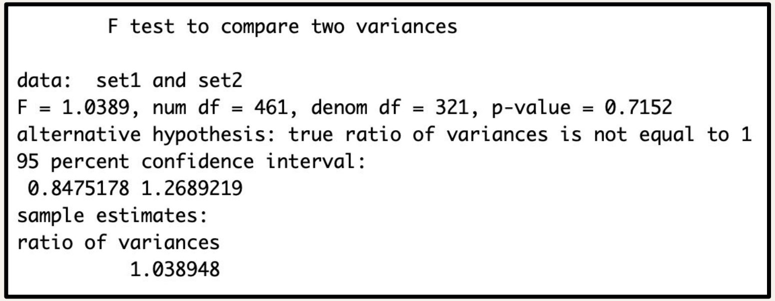
- 85.37·I(scent = fresh)



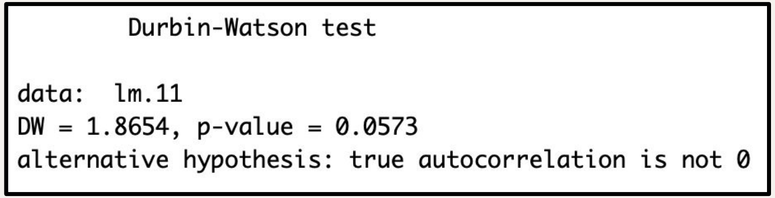
```
lm(formula = old_price ~ big_brand + conc + comp + seller_rating +
    ml + is.unisex + is.fresh, data = perfume3)

Residuals:
   Min     1Q Median     3Q    Max 
-397.8 -130.0   -8.3  117.3  651.9 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    23.3471   104.0130   0.224  0.82246    
big_brand      85.5629    12.9732   6.595 7.84e-11 ***
concEDT      -111.0210    13.5369  -8.201 9.83e-16 ***
comp           -4.4408     1.9320  -2.298  0.02180 *  
seller_rating  56.8586    25.5389   2.226  0.02628 *  
ml              1.2119     0.2797   4.332 1.67e-05 ***
is.unisex    -146.3030    27.4428  -5.331 1.28e-07 ***
is.fresh      -85.3706    28.5594  -2.989  0.00289 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 175.5 on 776 degrees of freedom
Multiple R-squared:  0.1789,    Adjusted R-squared:  0.1714 
F-statistic: 24.15 on 7 and 776 DF,  p-value: < 2.2e-16
```

Final Model



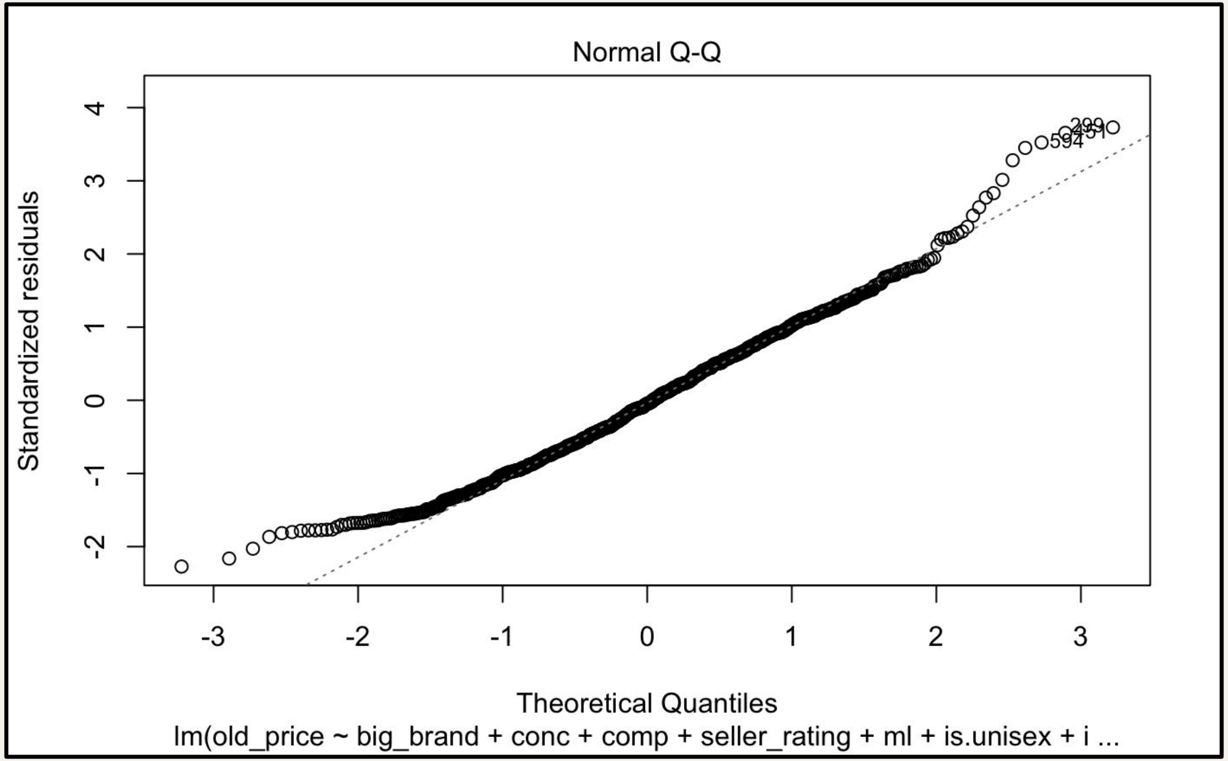


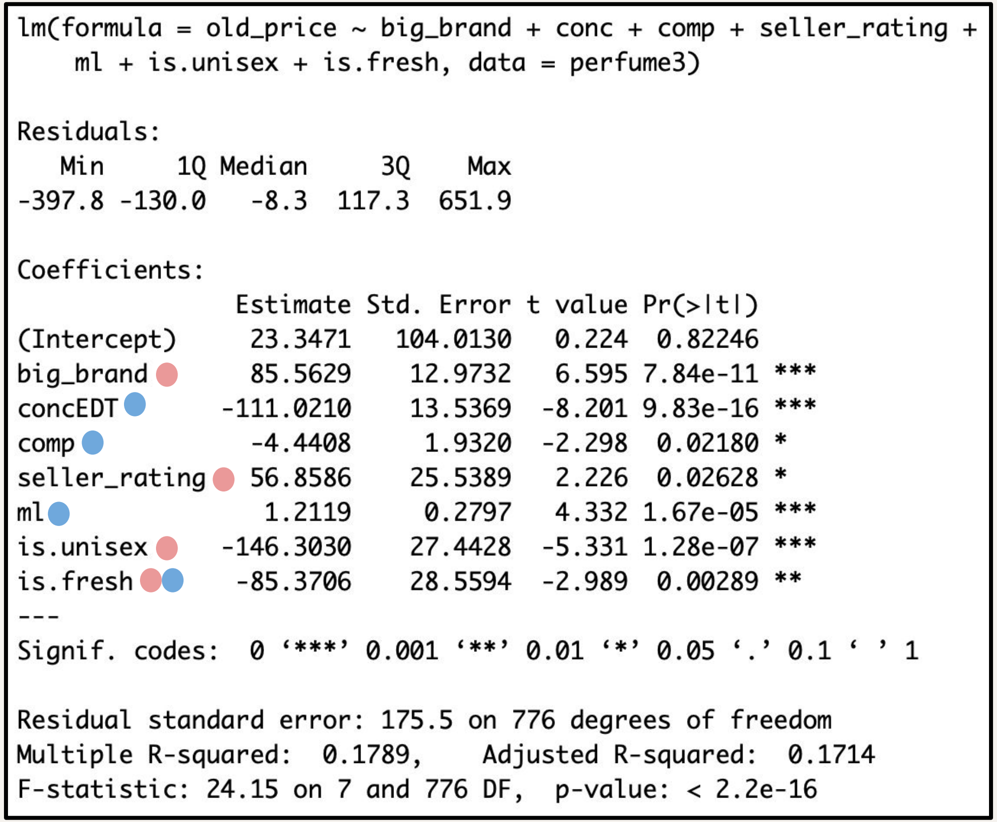
```
        F test to compare two variances

data:  set1 and set2
F = 1.0389, num df = 461, denom df = 321, p-value = 0.7152
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.8475178 1.2689219
sample estimates:
ratio of variances
          1.038948
```



```
        Durbin-Watson test

data:  lm.11
DW = 1.8654, p-value = 0.0573
alternative hypothesis: true autocorrelation is not 0
```

Final Model


Normal Q-Q
Standardized residuals
4
3
2
1
0
-1
-2
299
451
594
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(old_price ~ big_brand + conc + comp + seller_rating + ml + is.unisex + i ...



```
lm(formula = old_price ~ big_brand + conc + comp + seller_rating +
    ml + is.unisex + is.fresh, data = perfume3)

Residuals:
   Min     1Q Median     3Q    Max
-397.8 -130.0   -8.3  117.3  651.9

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     23.3471   104.0130   0.224  0.82246
big_brand ●     85.5629    12.9732   6.595 7.84e-11 ***
concEDT ●     -111.0210    13.5369  -8.201 9.83e-16 ***
comp ●          -4.4408     1.9320  -2.298  0.02180 *
seller_rating ● 56.8586    25.5389   2.226  0.02628 *
ml ●             1.2119     0.2797   4.332 1.67e-05 ***
is.unisex ●   -146.3030    27.4428  -5.331 1.28e-07 ***
is.fresh ●●    -85.3706    28.5594  -2.989  0.00289 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 175.5 on 776 degrees of freedom
Multiple R-squared:  0.1789,	Adjusted R-squared:  0.1714
F-statistic: 24.15 on 7 and 776 DF,  p-value: < 2.2e-16
```

Conclusion

- The prestige of the **brands** contribute the most to the perfume’s retail **price**.
- The **customer rating** and **seller** do not contribute to the perfume’s retail **price**. However, **seller rating does**.
- The perfume's **department** (female/male/unisex), **scents** (woody/floral/fruity/etc…), and **notes** leads to difference in retail prices and customer ratings.

Thanks

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**,and infographics & images by **Freepik**