What Defines Gender in Perfumes?

De Xuan Tan, Final Assignment
Advanced Quantitative Methods, Spring '25

Introduction

Gender has always shaped how perfumes are marketed and consumed. For years, perfumes have been categorized into "men" and "women" based on their scent profiles and its respective social narratives. Floral and powdery notes are often associated with femininity, while woody and spicy notes are associated with masculinity. Today, these social conventions continue to persist, despite being more of a cultural norms than having any scientific proof. (Doe 2020).

One of the key reasons gendered perfumes have remained so dominant is due to deeply embedded stereotypes within the fragrance industry and consumer psychology. (Smith 2019). The rising trend of unisex and gender-neutral fragrances, however, challenges these long-standing norms and invites a re-evaluation of how scent and identity interact (Lopez 2021).

The analysis begins by identifying the most influential variables contributing to gender differentiation—such as specific top, heart, and base notes—before building a classification model that predicts gender category using these features (Sissel and Collins 2020). Finally, all these can be further visualized by reducing the high-dimensional data into a visual "scent space" using clustering techniques like PCA. This will allow for an exploration of overlaps and divergences in how perfumes are gendered, ultimately revealing biases, trends, and outliers in gender-based fragrance marketing (Morris 2021).

Literature Review

The gendering of perfumes has deep historical and cultural roots. Scent has long been used as a tool to express identity, with ancient civilizations using aromatic oils for both men and women without strict gender boundaries (Classen, Howes, and Synnott 1994). However, in the modern perfume industry, particularly post-20th century, fragrances became sharply divided along gender lines, largely influenced by Western marketing practices and evolving gender roles (Roden 2015).

One might argue that a common trend in perfume marketing suggests that fragrances targeted towards women often exhibit a higher price point compared than those marketed towards men, possibly due to differences in branding, and packaging (Morrison 2019). Also, Perfumes with more notes may provide the model with richer compositional features to differentiate gendered trends (Fragrantica 2023).

The relationship between perfume ingredients and scent composition significantly affects gender classification in fragrance marketing. The goal of this analysis is to identify which combinations of scents, ingredients, and concentration types most strongly influence whether a perfume is marketed as masculine or feminine (Morrison 2019; Scentbird 2022). These data points will be visualized as clusters to reveal how unrelated batches of perfumes can share characteristics based on their features (Fragrantica 2023).

Research Questions

1. How does genders classification in perfumes affect pricing?

2. Which scent(s) determines whether a perfume is classified as masculine or feminine?

3. How does the diversity of fragrance notes affect gender classification in perfumes?

4. Which features contribute most to the accuracy of the model?

Methodology

Logistic regression is highly suitable in analyzing binary outcomes. This study revolves around building a logistic regression model which can identify the factors that are significant to the survivability of the passengers of Titanic. The process is as follows:

Data Processing The dataset will undergo cleaning, graceful handling of missing values. Some cells under the variable Age are missing value, so we will replace these NaN values with an average age instead of removing the respective entries to maintain the number of observations within the dataset. In addition, categorical variable, such as gender is encoded into binary format, where female = 1, male = 0.

Model Building The cleaned dataset is split into training (80%) and testing (20%) sets. A logistic regression model will be trained using data that is included in the

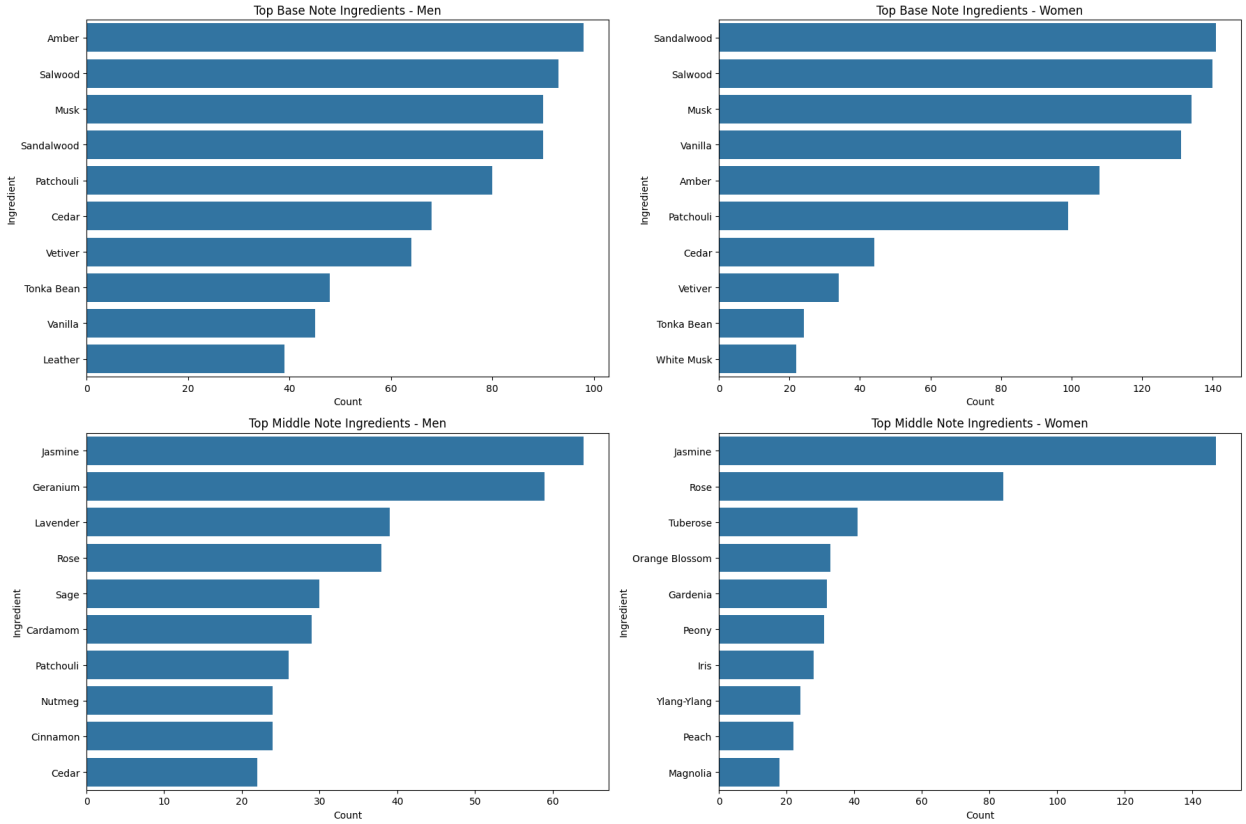
preprocessing steps for model fitting. All predictors are standardized or encoded during preprocessing to ensure compatibility with the logistic regression model.

Model Training The logistic regression model is fitted using the training dataset. Each predictor contributes to the model by adjusting the odds of a perfume being classified as male or female.

Model Testing The model is evaluated on the "unseen" 20% testing dataset, to predict the gender of perfumes based on the predictors. Performance metrics such as prediction accuracy, and ROC-AUC score are calculated to assess how well the model generalizes to new data.

Final Evaluation The final evaluation involves interpreting the model's coefficients and assessing its predictions, such as: A positive coefficient for "Notes Diversity" would suggest perfumes with more diverse notes are more likely to be classified as Male. A PCA analysis will be done to identify the most influential but indiscernible factors in gender classification and understand which note or scent determines the gender of the perfume.

Exploratory Data Analysis

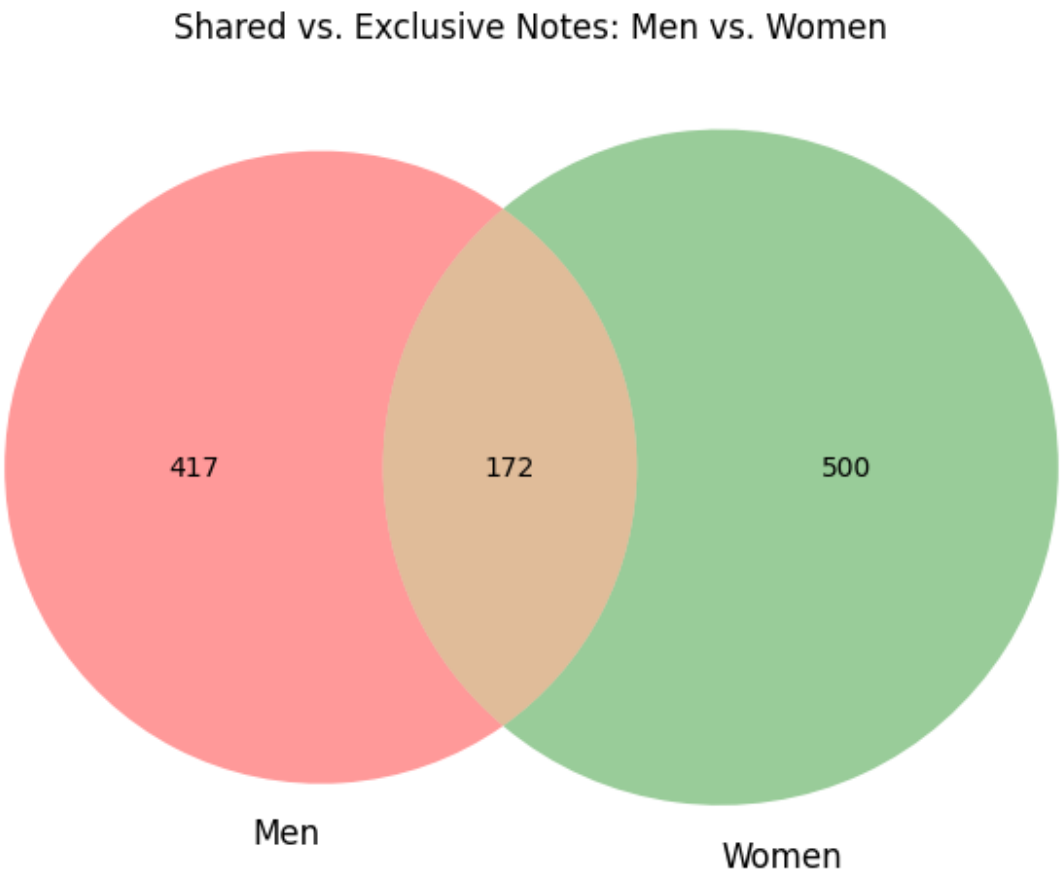


Top Base Notes – Men Amber, Salwood, and Musk are the most common base notes in men's fragrances, with usage counts approaching 100.

Top Base Notes – Women Sandalwood, Salwood, and Musk dominate the base notes in women’s fragrances, each with over 120 counts.

Top Middle Notes – Men Jasmine and Geranium are the most frequent middle notes in men's perfumes, significantly ahead of others like Lavender and Sage.

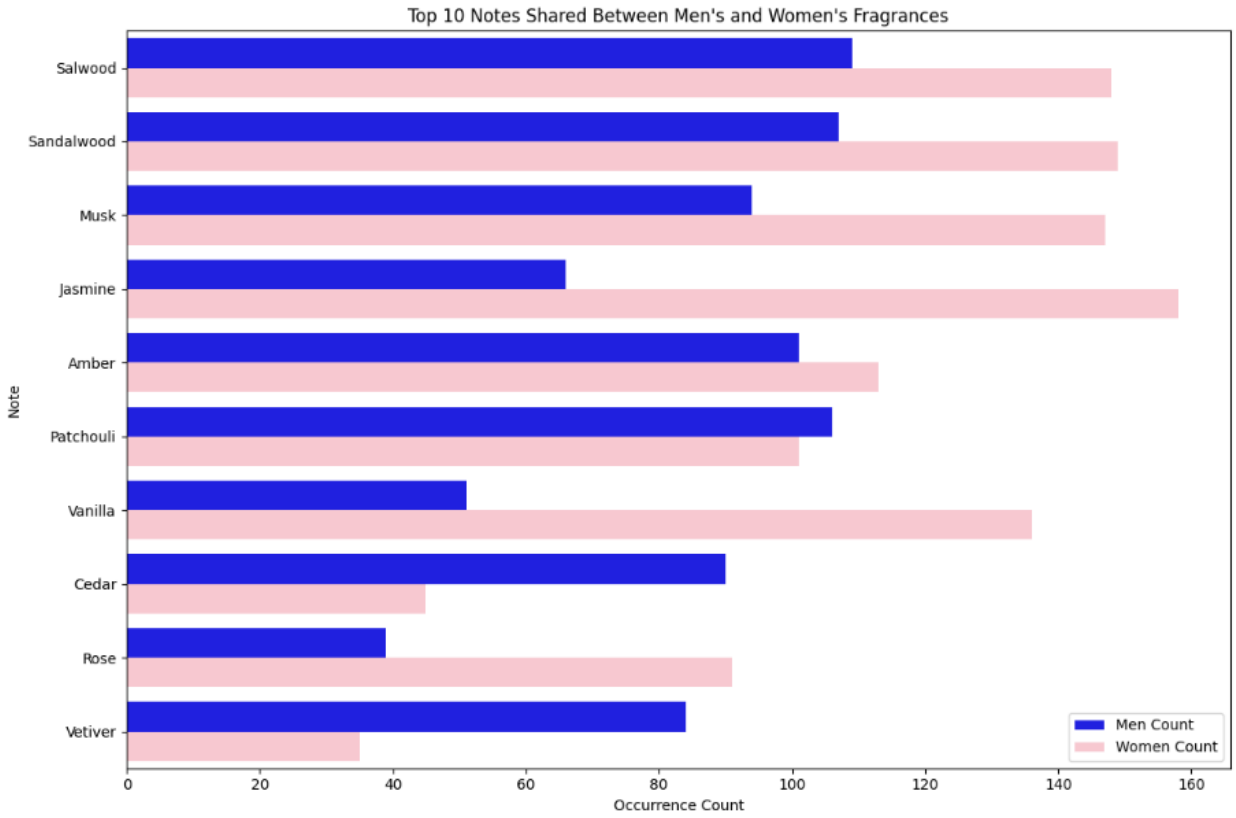
Top Middle Notes – Women Jasmine stands out as the top middle note by a wide margin for women, followed by Rose and Tuberose in popularity.



Gender Exclusive Notes

Exclusive Notes - Men	Occurrences	Exclusive Notes - Women	Occurrences
Clary Sage	14	Peony	31
Fir	10	Jasmine Sambac	13
Brazilian Rosewood	7	Wild Jasmine	9
Olibanum	7	Lilac	8
Geranium	7	Watermelon	5
Black Basil	6	Plum	5
Tonka Bean	6	Red Lily	5
Coumarin	6	Ylang Ylang	4
Birch	6	Cotton Candy	4
Pimento	4	Blackberry	4

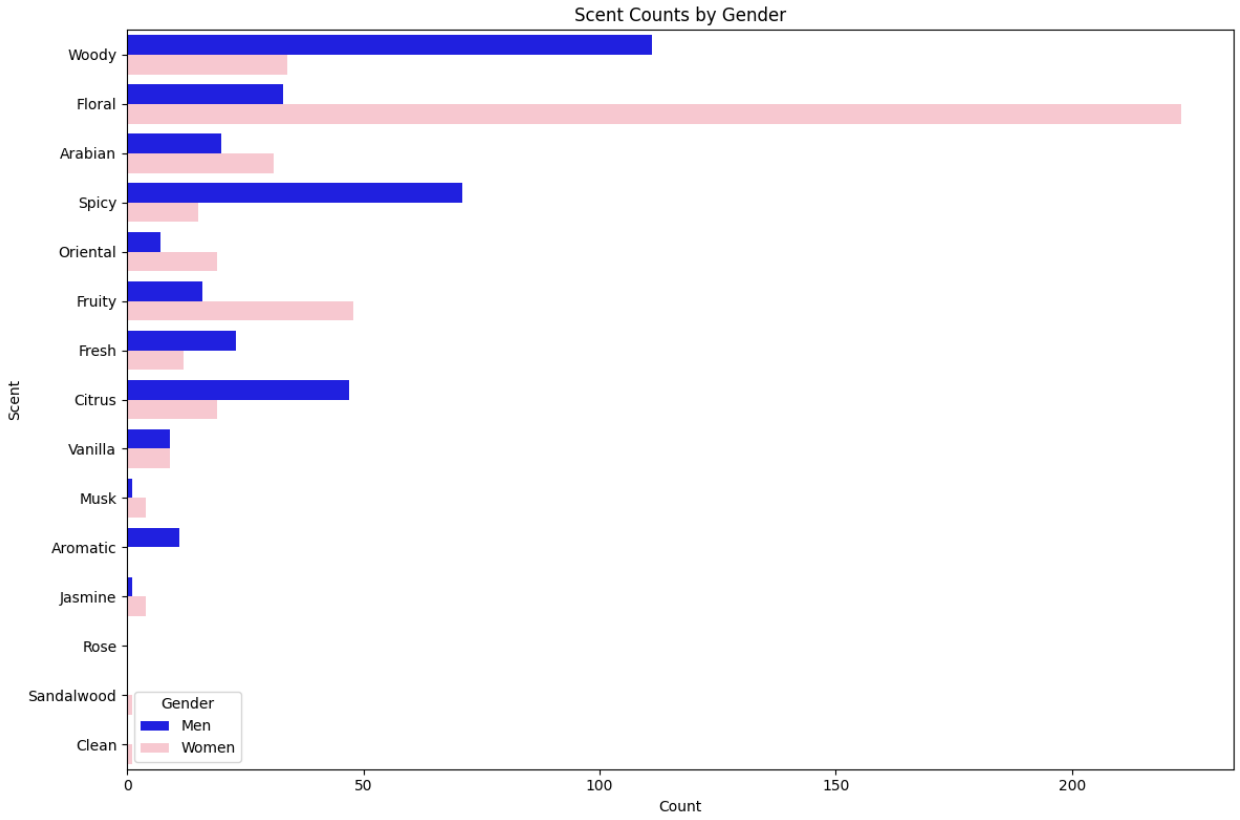
There are more fragrance notes exclusively used in women's perfumes (500) than in men's (417), with 172 notes shared across both. Peony and Jasmine Sambac are top exclusives for women, while Clary Sage and Fir dominate the men's exclusives.



Most Common Notes between Men's and Women's Fragrances:

Notes	Men Count	Women Count	Men %	Women %	Absolute Difference %
Patchouli	106	101	51.2	48.8	2.4
Amber	101	113	47.2	52.8	5.6
Salwood	109	148	42.4	57.6	15.2
Sandalwood	107	149	41.8	58.2	16.4
Musk	94	147	39.0	61.0	22.0
~~Oakmoss~~	~~35~~	~~20~~	~~63.6~~	~~36.4~~	~~27.2~~
~~Cedar~~	~~90~~	~~45~~	~~66.7~~	~~33.3~~	~~33.4~~
~~Tonka Bean~~	~~53~~	~~25~~	~~67.9~~	~~32.1~~	~~35.8~~
~~Rose~~	~~39~~	~~91~~	~~30.0~~	~~70.0~~	~~40.0~~
~~Jasmine~~	~~66~~	~~158~~	~~29.5~~	~~70.5~~	~~41.0~~

Patchouli and Amber show the smallest gender preference gaps (2.4% and 5.6% respectively), making them reliable indicators of unisex perfumes, followed by Salwood, Sandalwood, and Musk. This information can be ideal for identifying unisex fragrances by targeting notes with minimal gender distribution.

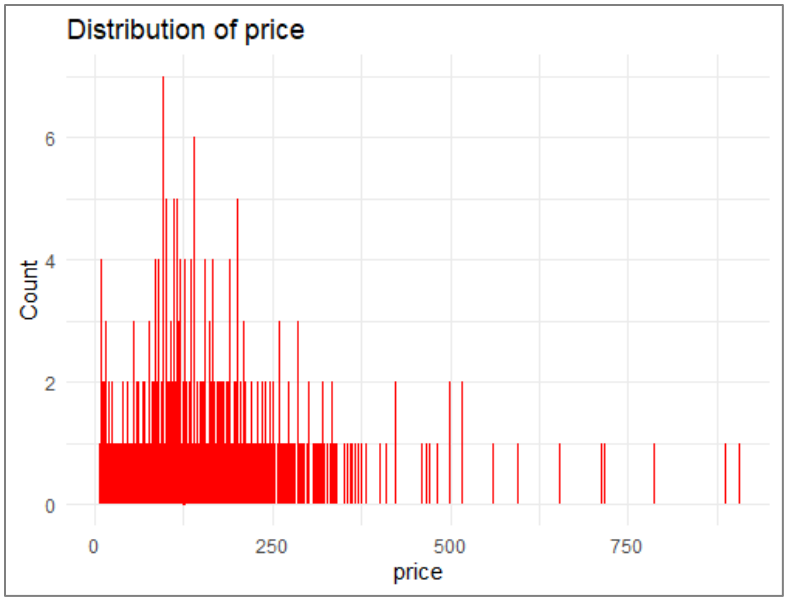


Genders and their Respective Scents and Counts

	Woody	Floral	Arabian	Spicy	Oriental	Fruity	Fresh	Citrus
Men	111	33	20	71	7	16	23	47
Women	34	223	31	15	19	48	12	19
Total	145	256	51	86	26	64	35	66

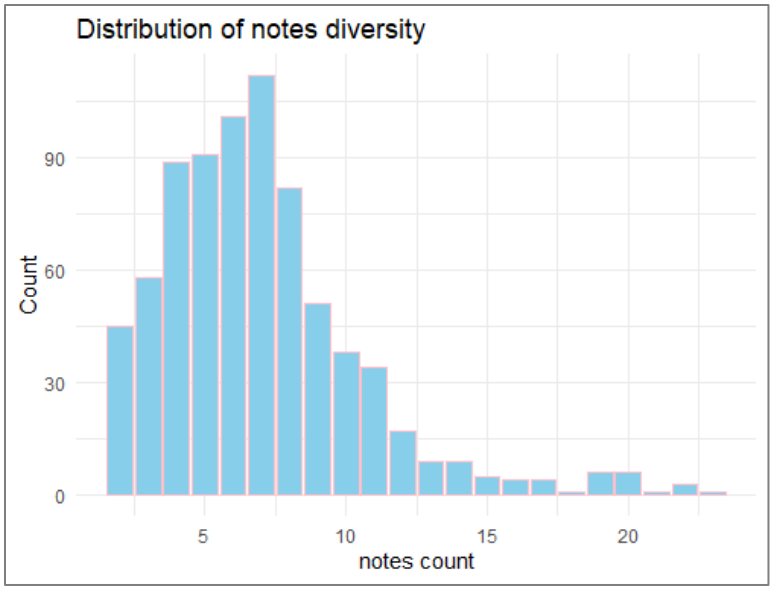
	Vanilla	Musk	Aromatic	Jasmine	Rose	Sandalwood	Clean
Men	9	1	11	1	0	0	0
Women	9	4	0	4	0	1	1
Total	18	5	11	5	0	1	1

Gender Scent Counts Several scents shared almost equally between men and women, such as Arabian and Vanilla. Distinct scents emerge with Floral, Oriental, Fruity, Musk, Jasmine leaning towards feminine tastes, while Woody, Spicy, Fresh, Citrus and Aromatic scents are more favored by men. These scents may be beneficial in determining perfume gender.

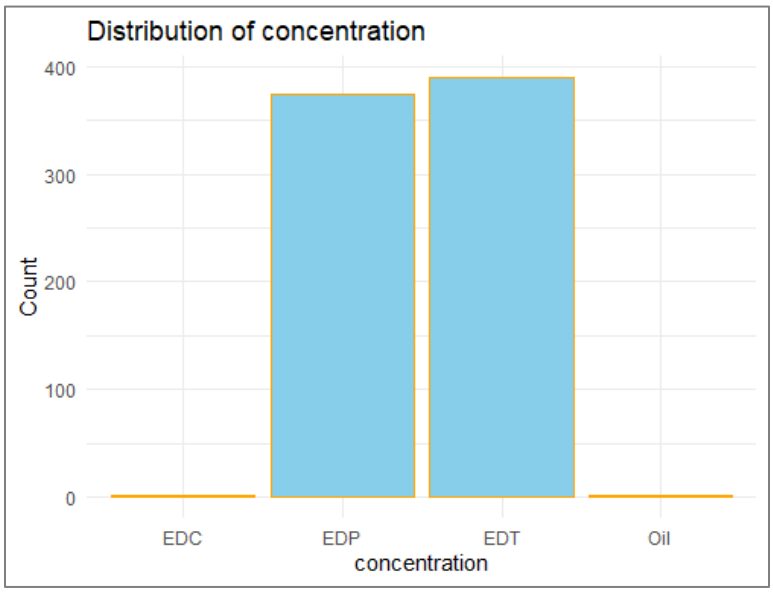


Perfume Prices (US$) In the dataset, majority of the perfumes are priced lower than US$350. There are some outliers of perfumes of higher prices, at around US$720 to even US$1000, which may result in bias to perfumes of lower costs.

Gender Distribution The chart indicates that despite there are more women than men in the dataset, distribution is still fairly even and the gap is not huge hence, the chance of gender bias is greatly reduced.



Notes Diversity The distribution shows a peak in the number of perfumes with around 7 notes, showing that perfumes with 6 notes are the most common in the dataset. With the number of notes increasing from this peak, the number of perfumes declines significantly, showing that most perfumes in dataset are on a low note count.



Perfume Concentrations Majority of the perfumes in the dataset are EDP and EDT, with the EDC and Oil Extrait on significantly lower counts. Hence, the results may introduce bias towards EDPs and EDTs.

Feature Selection

Referring to the literature review, relevant variables are identified and included as predictors in the model.

The variables that are present in both literature review and the dataset are as follow:

- Price
- Scent Strength

In addition, the following variables in the dataset will also be added due to suspected correlations:

- Notes Diversity
- Top Common Notes
- Top Scents

The model was built with the predicters added in the following order:

1. Price
2. Scent Strength
3. Notes Diversity
4. Top Common Notes
5. Top Scents

Other noteworthy variables that are missing which may be helpful in determining perfume gender include:

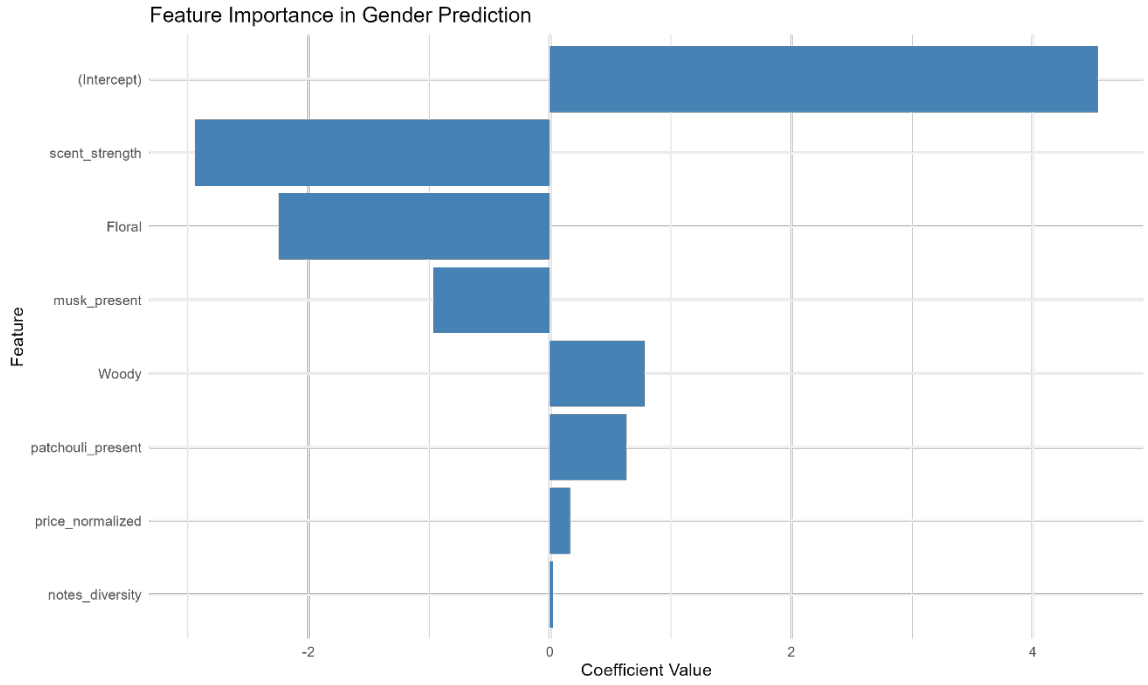
-Seasonal or Temporal Trends Some notes are more popular in certain seasons (e.g., citrus in summer, musk in winter). Gendered usage could shift with seasonal offerings, but this is not captured.

-Age Group Fragrance preferences can vary significantly across age groups. Understanding whether certain notes appeal more to younger or older users could provide insight into gendered patterns not captured by raw usage counts.

-Cultural or Regional Influence Certain notes may be gendered differently across cultures or regions. Including geographic data or market segmentation could help contextualize gender associations.

Results

Evaluating Model Performance A trial and error is conducted to find the most suitable combination of predictors within the model while ensuring that the predicters remain statistically significant and yet retain a high accuracy of prediction on the test data.



The final logistic regression model indicates that the gender of a perfume can be accurately predicted by its Price, Scent Strength (Concentration), Notes Diversity, Top Common Notes (Patchouli and Musk), Top Scents (Woody and Floral).

Positive coefficients increase the odds of the perfume being masculine, whereas negative coefficients decrease the odds of the perfume being feminine.

The greater the swing of the bar to the left, the stronger the negative influence of that feature, making the perfume more likely to be classified as the feminine. In this model, Scent Strength and Floral scent are the major discriminators for female perfumes, followed by a same, smaller influence by Musk note. Woody scent and Patchouli note are the major discriminators for male perfumes. The intercept suggests a baseline tendency towards masculine perfumes.

The following insights are derived from these diagnostics:

-A perfume is 93.5% more likely to be labelled Masculine than Feminine when all other predictors are not considered.

-For every unit increase in the Price, the perfume is 6% less likely for it to be classified as a "Male" perfume. However, due to its high p-value (> 0.05), this shows that this predictor is not statistically significant. It is, however included, due to suspected correlation that woman's perfumes tend to cost more.

-For every unit increase in Scent Strength, the perfume is 5.5% more likely to be classified as a "Male" perfume. This is highly statistically significant as it's p-value < 0.05.

-Every every unit increase in Notes Diversity, the perfume is 2% more likely to be classified as a "Male" perfume. The high p-value > 0.05 indicates it's not statistically significant. It is, however included, due to suspected correlation that perfumes with more notes might give the model more features to interpret its gender.

-Every every unit increase in presence of Patchouli, the perfume is TWICE as likely to be classified as a "Male" perfume. This is highly statistically significant as it's p-value < 0.05.

-For every unit increase in presence of Musk, the perfume is 62% more likely to be classified as a "Female" perfume. This is highly statistically significant as it's p-value < 0.05.

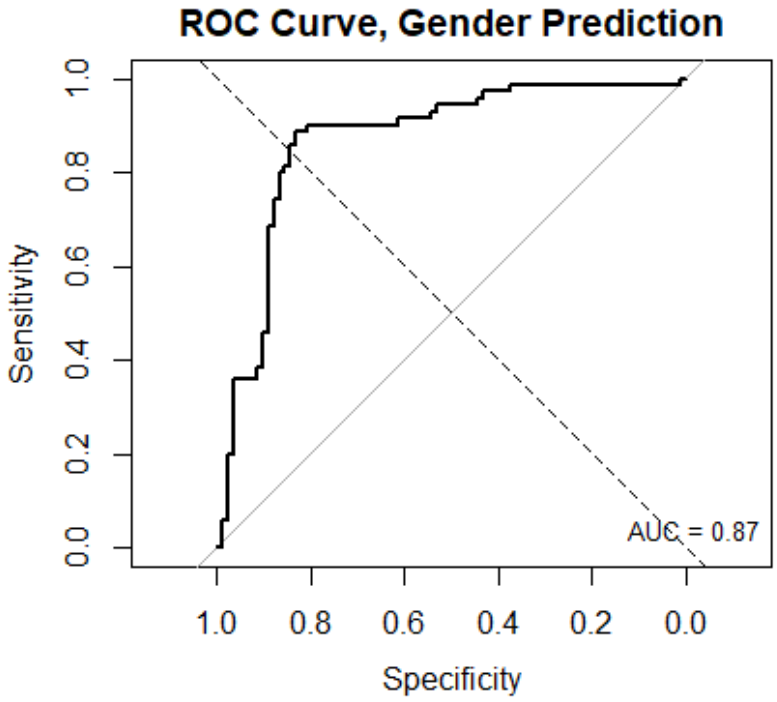
-For every unit increase in presence of Woody scent, the perfume is 76% more likely to be classified as a "Male" perfume. This is highly statistically significant as it's p-value < 0.05.

-For every unit increase in presence of Floral scent, the perfume is 89% more likely to be classified as a "Female" perfume. This is highly statistically significant as it's p-value < 0.05.

-Nagelkerke Value The value of 0.584 suggests that the model explains almost 60% of the variance in perfume gender, with the remaining factors outside of the predictors in the model.

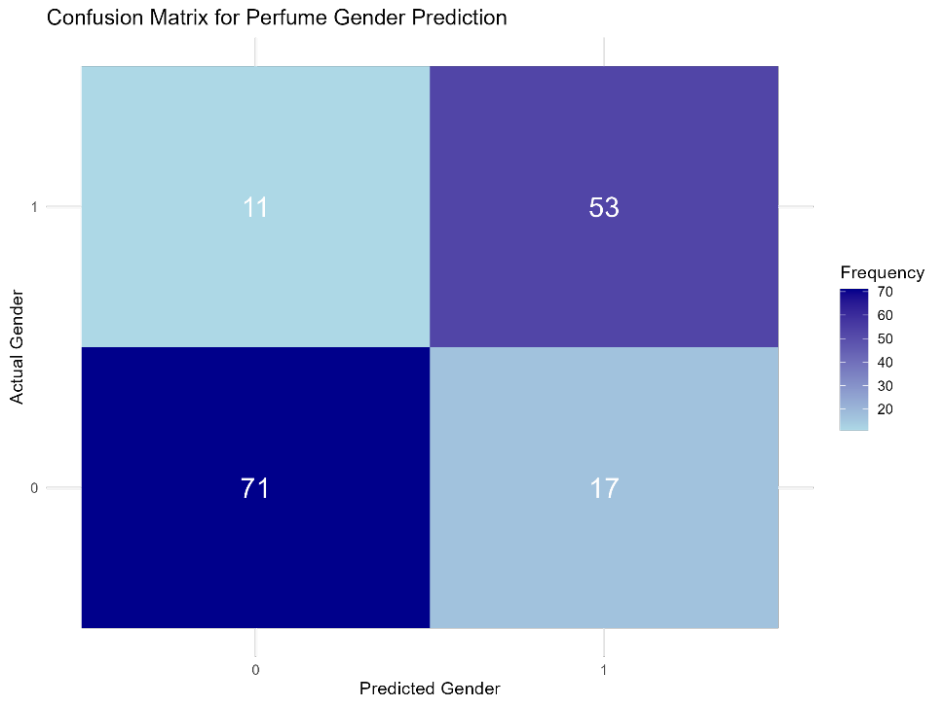
Evaluating Training Performance

The final logistic regression model is also tested on the training data, and it is revealed to have an **accuracy score of 83.52%**, suggesting that out of 100 perfumes, 83 of them have their gender correctly predicted.



In addition, the final logistic regression model has been tested on **"the area under the curve" on the ROC curve, which stands at 0.87.** Where the closer the value is to 1, the better the model. The remaining gap might be due to omitted variables bias as factors key in determining perfume gender are not available in the dataset, and hence outside of our control. As the sample of 760 observations is considered moderate, the potential bias associated with a small sample size is likely minimal.

Evaluating Test Performance



The final logistic regression model is further tested on a set of test data to evaluate its error rate on datasets not "seen" by the model. The test data has details of 152 perfumes with the same number of variables as the training dataset. It is able to correctly predict genders of the 124 out of the 152 perfumes, computing an **accuracy score of 81.58%.**

Where Women = 0, Men = 1

These peaks for each gender at opposite ends demonstrate a good overall discriminatory power. However, the overlap in the middle highlights the model's uncertainty for some perfumes and contributes to the fact that the AUC was 0.87 (not a perfect 1.0). Since we would want the pink distribution to be as near to 0 and the blue distributed as near to 1 as possible, this overlap suggest there's room to improve the model, usually by including more features.

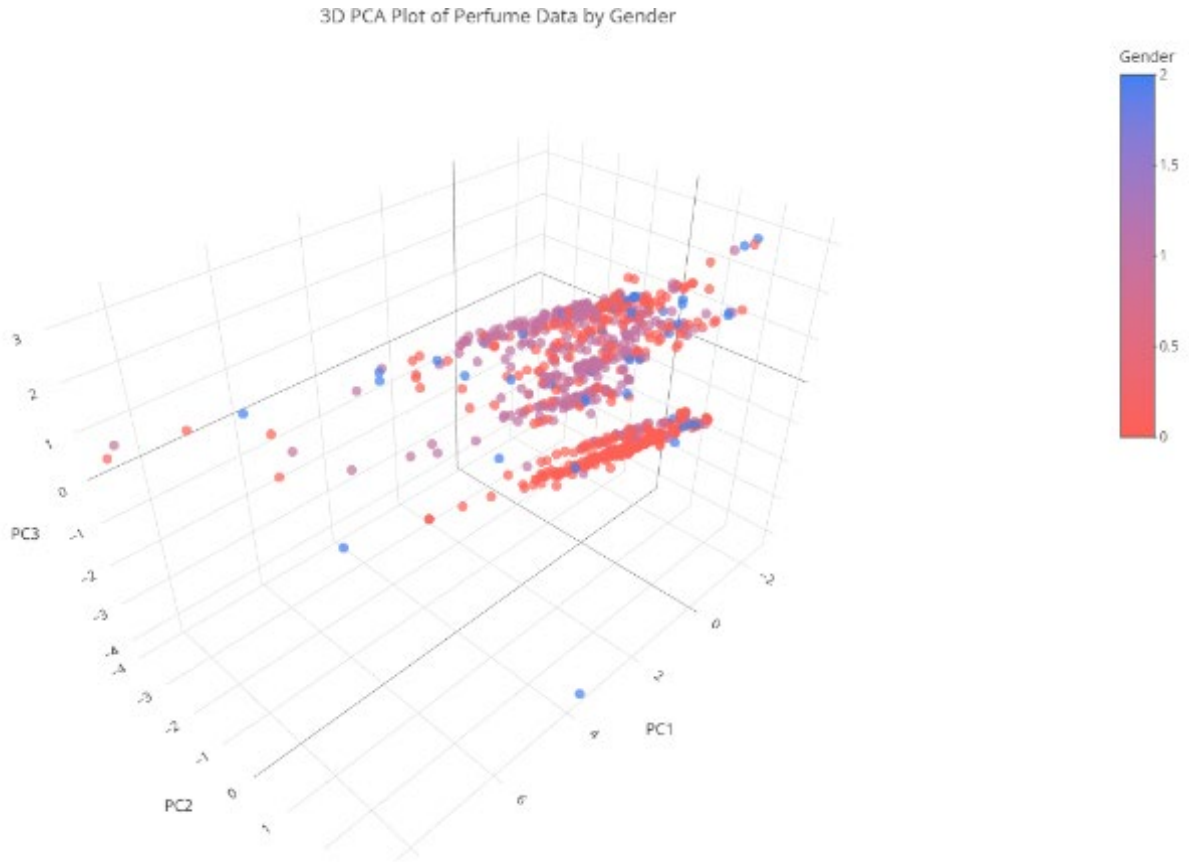
Pink Peak; Women There is also a smaller peak at 0.85 probability, hinting that some female perfumes were incorrectly predicted or due to their features, are seen to be masculine.

Teal Peak; Men Similar to the Pink Peak, there is smaller peak at 0.15 probability, suggesting some male perfumes were incorrectly predicted or due to their features, are seen to be feminine.

Cluster Features into Genders

In the features driving our perfume gender prediction model, we see that scent strength and floral notes tend to push the prediction towards the feminine category. To further explore the underlying structure and groupings within our perfume data based on these same predictive features, we can consider c**reating clusters from predictors through** Principal Component Analysis (PCA)**, which** potentially reveal patterns that might not be immediately obvious or indiscernible through direct gender prediction alone.

PCA simplifies the number of feature within a "space" above 3 dimensions by reducing number of features, while not sacrificing variance and cluster these perfume as points, which reveal these "clusters", groups where the data points are related to each other, allowing us to see perfumes naturally group together based on combinations of these principal components, providing a different lens to understand the intricate mechanics of perfume gender categorization.



Where 0 is female, 1 is male, 2 is unisex. Github Pages

A 3D Principal Component Analysis (PCA) of the perfume data is plotted, colour coded by gender. The three axes, PC1, PC2, and PC3, represent the first three principal components, which are linear combinations of the original perfume features that capture the most variance in the data.

Each point in the 3D space represents a perfume, and the colour indicates its gender category according to the legend: Red for female (0), Purple for Male (1), Blue for Unisex (2). A color gradient interpolated on these points to represent a mix or where the classification does not fall strictly as Male or Female.

From this plot, a few insights can be derived:

Distinction of Perfume Gender The plot suggests the extent to which perfumes determine which genders as they cluster in distinct regions of the reduced 3D space. Upon rotating the space, clusters can be spotted between red and purple points, indicating that the use of PCA has effectively captured differences in perfume characteristics that define masculine and feminine scents.

Overlap or Outliers as Unisex Perfume Areas where red and purple points intersect might suggest perfumes with characteristics appealing to both genders, hinting the possibility of marketing them as unisex fragrance – something most probably previously unthought of. The presence and distribution of intermediate colours further highlight these overlaps. These can be put into contrast with the blue points, illustrate perfumes intentionally marketed as unisex.

Dominant Factors Determining Gender Due to the use of PCA, the plot does not directly label the original variables, as the principal components themselves are derived from these variables (like scents, concentration, etc.). The spread and orientation of the clusters along the PC1, PC2, and PC3 axes can indirectly suggest which underlying factors contribute most to the differentiation between perfume gender.

	PC1	*PC2*	*PC3*
Price	0.6328	-0.2348	-0.0482
Scent Strength	0.3313	0.4477	0.2737
Notes Diversity	-0.0692	-0.085	-0.1804
Woody	0.0202	-0.4286	0.3164
Floral	0.1898	0.565	-0.4592
Arabian	0.0164	0.0314	0.2826

Spicy	-0.132	-0.2145	-0.1674
Oriental	-0.043	0.1753	0.1817
Fruity	-0.0124	0.0567	0.4
Fresh	-0.089	0.0009	0.0968
Citrus	-0.1055	-0.2483	-0.3856
Vanilla	0.0412	-0.0577	0.0557
Musk	-0.0281	0.1104	0.2338
Aromatic	-0.018	-0.1131	-0.1861
Jasmine	0.0499	-0.0132	0.0675
Rose	0.0583	0.0667	-0.1213
Sandalwood	-0.0296	0.0454	0.0464
Clean	-0.015	0.0398	0.066

For example, if the clusters are widely separated along PC1, then the features that have a high loading, seen here is Price and Scent Strength, on PC1 are likely strong discriminators between male and female perfumes.

Limits

1. Missing Values and Bias The dataset had some missing values or misspelling that are diffcult to anticipate. Although they were removed, this can mean a lost opportunity to work on uncharted data and also affect overall data distribution, which can introduce unnecessary bias into the model.

2. Missing Key Variables The dataset does not have other potentially significant variables, such as target age group or seasonal scents/ notes. Including these variables could improve the model's predictive accuracy.

3. Generalizability While the model performed equally well on both training and test data, the predictive accuracy on the test data (at 81.58%) suggests a rather high level of predictive capability, but further evaluation on truly novel data would be necessary to fully assess its generalizability, since it has only been tested on 20% of the initial dataset.

4. Majority Blindspot The training dataset has an issue of data imbalanced such as, having more EDPs than Extrait Oil perfumes; more low priced perfumes than perfumes of high price points. This could lead to bias toward the majority (regularly-priced perfumes), affecting its ability to accurately predict genders for the minority (luxury perfumes).

Conclusion

In summary, the final model fits the data well, as evident in the AUC score of 0.87. It identified scents strength and ironically, gender common notes as the most significant predictors of gender in perfumes. While the model achieved good predictive accuracy, there is additional room for improvement by addressing data imbalance in the training dataset, such as the number of notes, prices of the perfumes and concentration of perfumes can be made more equally distributed in the dataset. The findings reinforce the factors highlighted in the literature review, stating that gender stereotypes and their respective scents greatly influenced which gender the perfume is marketed.

References

Doe, Jane. 2020. *The Scent of Gender: A Cultural History of Perfume*. New York: Fragrance Press.

Lopez, Maria. 2021. "Unisex Fragrance Trends and Gender Fluidity." *Perfume & Society* 8 (1): 27–29.

Smith, John. 2019. "Packaging Identity: Marketing and Gender in Consumer Products." *Journal of Advertising Research* 55 (3): 112–125.

Classen, Constance, David Howes, and Anthony Synnott. 1994. *Aroma: The Cultural History of Smell*. London: Routledge.

Milotic, Daniel. 2003. "The Impact of Visual Design on the Perception of Product Flavor." *Journal of Consumer Behaviour* 3 (2): 179–193.

Mintel. 2021. *Fragrance and Gender: Consumer Attitudes Toward Gender-Neutral Perfumes in the US*. London: Mintel Group Ltd.

Roden, Barbara. 2015. "Fragrance and Femininity: The Gender Divide in the Perfume Industry." *Cultural Studies Review* 21 (1): 45–58.

Fragrantica. 2023. *Fragrance Classification by Notes and Gender Bias*. Accessed March 20, 2025. https://www.fragrantica.com.

Morrison, Karen. 2019. *The Economics of Perfume Marketing: Gender and Pricing Trends*. New York: Fashion and Fragrance Press.

Morris, Lila. 2021. "Gendered Marketing in the Fragrance Industry: A Scented Spectrum." *Journal of Consumer Culture* 21 (3): 415–32.

Scentbird. 2022. "Most Popular Perfume Notes by Gender: Patchouli, Musk, and More." *Scentbird Magazine*, July 18, 2022. https://www.scentbird.com/blog/popular-perfume-notes-by-gender.

Sissel, Morgan, and Ryan Collins. 2020. *Olfactory Data Science: Machine Learning in the Perfume Industry*. Cambridge: Data Press.

Technical Appendix – Steps

Step 1: Literature review & Record expectations

Based on literature review, the following variables will be considered to be included in the logistic regression model.

1. Passenger Class (Pclass): Socioeconomic status was a major determinant of survival. Historical records suggest that first-class passengers had preferential access to lifeboats, while third-class passengers faced higher mortality rates (Lord 1955). Including this variable helps capture class-based survival disparities.

2. Sex: One of the strongest predictors of survival, female passengers had a significantly higher survival rate due to the "women and children first" evacuation protocol (Gleicher 1996). Gender-based survival bias is evident in multiple statistical analyses of the dataset (Dawson 1995).

Step 2: Reconcile with available variables. Add relevant variables from dataset which were not part of step 1

The variables that are present in both literature review and the dataset are as follow:

- Price
- Scent Strength

In addition, the following variables in the dataset will also be added due to suspected correlations:

- Notes Diversity
- Top Common Notes
- Top Scents

Step 3: Make note of missing variables (omitted variables bias)

Some missing variables that maybe helpful in determine survival rates include:

- Seasonal or Temporal Trends Some notes are more popular in certain seasons (e.g., citrus in summer, musk in winter). Gendered usage could shift with seasonal offerings, but this is not captured.

- Age Group Fragrance preferences can vary significantly across age groups. Understanding whether certain notes appeal more to younger or older users could provide insight into gendered patterns not captured by raw usage counts.

- Cultural or Regional Influence Certain notes may be gendered differently across cultures or regions. Including geographic data or market segmentation could help contextualize gender associations.

Step 4: Rank order the candidate variables

1. Price
2. Scent Strength

3. Notes Diversity
4. Top Common Notes
5. Top Scents

Step 5: Check descriptive statistics

Missing and Nan values are removed.

```
> describe(profumo)
                vars   n   mean     sd median trimmed    mad min    max  range  skew kurtosis   se
brand              1 767  68.75  42.77  61.00   66.53  50.41 1.0 149.00 148.00  0.39    -1.09 1.54
name               2 767 274.90 155.97 274.00  275.14 195.70 1.0 548.00 547.00 -0.01    -1.18 5.63
price              3 767 154.87 108.80 139.95  143.82  83.99 5.6 904.85 899.25  2.10     9.02 3.93
concentration      4 767   2.51   0.51   3.00    2.51   0.00 1.0   4.00   3.00 -0.04    -1.84 0.02
gender             5 767   1.54   0.50   2.00    1.55   0.00 1.0   2.00   1.00 -0.18    -1.97 0.02
scents             6 767   8.67   5.46   7.00    8.53   5.93 1.0  17.00  16.00  0.43    -1.33 0.20
base_note          7 767 347.44 198.98 349.00  348.10 255.01 1.0 690.00 689.00 -0.02    -1.21 7.18
middle_note        8 767 342.29 197.15 340.00  342.72 252.04 1.0 679.00 678.00  0.00    -1.19 7.12
Gender_encoded     9 767   0.46   0.50   0.00    0.45   0.00 0.0   1.00   1.00  0.18    -1.97 0.02
scent_strength    10 765   1.49   0.50   1.00    1.49   0.00 1.0   2.00   1.00  0.04    -2.00 0.02
price_normalized  11 767   0.17   0.12   0.15    0.15   0.09 0.0   1.00   1.00  2.10     9.02 0.00
notes_diversity   12 767   6.98   3.59   6.00    6.57   2.97 2.0  23.00  21.00  1.42     3.04 0.13
base_note_0       13 767 105.79  58.64 117.00  108.28  72.65 1.0 201.00 200.00 -0.37    -1.20 2.12
base_note_1       14 767  91.86  55.75 106.00   94.00  57.82 1.0 180.00 179.00 -0.40    -1.25 2.01
base_note_2       15 767  58.26  50.48  60.00   55.91  72.65 1.0 151.00 150.00  0.11    -1.59 1.82
base_note_3       16 767  31.01  37.82   1.00   26.56   0.00 1.0 108.00 107.00  0.68    -1.28 1.37
base_note_4       17 767  10.78  18.53   1.00    6.70   0.00 1.0  67.00  66.00  1.59     0.87 0.67
base_note_5       18 767   4.66  10.03   1.00    1.55   0.00 1.0  45.00  44.00  2.67     5.68 0.36
base_note_6       19 767   2.20   4.95   1.00    1.00   0.00 1.0  34.00  33.00  4.51    20.22 0.18
base_note_7       20 767   1.53   2.70   1.00    1.00   0.00 1.0  22.00  21.00  5.60    31.86 0.10
base_note_8       21 767   1.13   0.96   1.00    1.00   0.00 1.0  10.00   9.00  7.72    61.54 0.03
base_note_9       22 767   1.07   0.62   1.00    1.00   0.00 1.0   8.00   7.00  9.84    99.57 0.02
base_note_10      23 767   1.02   0.28   1.00    1.00   0.00 1.0   6.00   5.00 14.11   213.93 0.01
base_note_11      24 767   1.01   0.17   1.00    1.00   0.00 1.0   4.00   3.00 15.69   253.96 0.01
middle_note_0     25 767 144.49  80.28 130.00  143.57  94.89 1.0 295.00 294.00  0.13    -1.03 2.90
middle_note_1     26 767 118.42  84.63 107.00  116.35 117.13 1.0 272.00 271.00  0.10    -1.27 3.06
middle_note_2     27 767  67.04  68.40  49.00   59.92  71.16 1.0 214.00 213.00  0.55    -1.11 2.47
middle_note_3     28 767  25.53  39.15   1.00   17.72   0.00 1.0 129.00 128.00  1.33     0.29 1.41
middle_note_4     29 767   8.81  19.19   1.00    3.21   0.00 1.0  84.00  83.00  2.45     4.74 0.69
middle_note_5     30 767   3.32   7.92   1.00    1.01   0.00 1.0  47.00  46.00  3.67    12.92 0.29
middle_note_6     31 767   2.17   5.15   1.00    1.00   0.00 1.0  37.00  36.00  4.73    22.18 0.19
middle_note_7     32 767   1.57   2.98   1.00    1.00   0.00 1.0  24.00  23.00  5.51    30.33 0.11
middle_note_8     33 767   1.23   1.50   1.00    1.00   0.00 1.0  16.00  15.00  7.10    52.41 0.05
middle_note_9     34 767   1.14   1.01   1.00    1.00   0.00 1.0  13.00  12.00  8.61    78.31 0.04
middle_note_10    35 767   1.02   0.30   1.00    1.00   0.00 1.0   6.00   5.00 13.24   182.56 0.01
middle_note_11    36 767   1.00   0.08   1.00    1.00   0.00 1.0   3.00   2.00 22.18   515.34 0.00
Woody             37 767   0.19   0.39   0.00    0.11   0.00 0.0   1.00   1.00  1.59     0.51 0.01
Floral            38 767   0.33   0.47   0.00    0.29   0.00 0.0   1.00   1.00  0.70    -1.51 0.02
Arabian           39 767   0.07   0.25   0.00    0.00   0.00 0.0   1.00   1.00  3.47    10.08 0.01
Spicy             40 767   0.11   0.32   0.00    0.02   0.00 0.0   1.00   1.00  2.45     4.03 0.01
Oriental          41 767   0.03   0.18   0.00    0.00   0.00 0.0   1.00   1.00  5.14    24.46 0.01
Fruity            42 767   0.08   0.28   0.00    0.00   0.00 0.0   1.00   1.00  3.01     7.05 0.01
Fresh             43 767   0.05   0.21   0.00    0.00   0.00 0.0   1.00   1.00  4.35    16.91 0.01
Citrus            44 767   0.09   0.28   0.00    0.00   0.00 0.0   1.00   1.00  2.95     6.69 0.01
Vanilla           45 767   0.02   0.15   0.00    0.00   0.00 0.0   1.00   1.00  6.28    37.53 0.01
Musk              46 767   0.01   0.08   0.00    0.00   0.00 0.0   1.00   1.00 12.24   148.01 0.00
Aromatic          47 767   0.01   0.12   0.00    0.00   0.00 0.0   1.00   1.00  8.15    64.57 0.00
Jasmine           48 767   0.01   0.08   0.00    0.00   0.00 0.0   1.00   1.00 12.24   148.01 0.00
Rose              49 767   0.00   0.00   0.00    0.00   0.00 0.0   0.00   0.00   NaN      NaN 0.00
Sandalwood        50 767   0.00   0.04   0.00    0.00   0.00 0.0   1.00   1.00 27.59   760.01 0.00
Clean             51 767   0.00   0.04   0.00    0.00   0.00 0.0   1.00   1.00 27.59   760.01 0.00
```

```
> summary(profumo)
    brand               name               price         concentration        gender             scents
 Length:767         Length:767         Min.   :  5.60   Length:767         Length:767         Length:767
 Class :character   Class :character   1st Qu.: 86.88   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Median :139.95   Mode  :character   Mode  :character   Mode  :character
                                       Mean   :154.88
                                       3rd Qu.:200.28
                                       Max.   :904.85

  base_note          middle_note        Gender_encoded   scent_strength price_normalized    notes_diversity
 Length:767         Length:767         Min.   :0.0000   Min.   :1.00   Min.   :0.0002779   Min.   : 2.000
 Class :character   Class :character   1st Qu.:0.0000   1st Qu.:1.00   1st Qu.:0.0906337   1st Qu.: 4.500
 Mode  :character   Mode  :character   Median :0.0000   Median :1.00   Median :0.1496387   Median : 6.000
                                       Mean   :0.4563   Mean   :1.49   Mean   :0.1662312   Mean   : 6.975
                                       3rd Qu.:1.0000   3rd Qu.:2.00   3rd Qu.:0.2167037   3rd Qu.: 8.000
                                       Max.   :1.0000   Max.   :2.00   Max.   :1.0000000   Max.   :23.000
                                                        NA's   :2
 base_note_0        base_note_1        base_note_2        base_note_3        base_note_4        base_note_5
 Length:767         Length:767         Length:767         Length:767         Length:767         Length:767
 Class :character   Class :character   Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character

 base_note_6        base_note_7        base_note_8        base_note_9        base_note_10       base_note_11
 Length:767         Length:767         Length:767         Length:767         Length:767         Length:767
 Class :character   Class :character   Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character

middle_note_0      middle_note_1      middle_note_2      middle_note_3      middle_note_4      middle_note_5
Length:767         Length:767         Length:767         Length:767         Length:767         Length:767
Class :character   Class :character   Class :character   Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character

middle_note_6      middle_note_7      middle_note_8      middle_note_9      middle_note_10     middle_note_11
Length:767         Length:767         Length:767         Length:767         Length:767         Length:767
Class :character   Class :character   Class :character   Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character

     Woody            Floral           Arabian            Spicy           Oriental          Fruity
Min.   :0.000   Min.   :0.0000   Min.   :0.00000   Min.   :0.0000   Min.   :0.0000   Min.   :0.00000
1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000
Median :0.000   Median :0.0000   Median :0.00000   Median :0.0000   Median :0.0000   Median :0.00000
Mean   :0.189   Mean   :0.3338   Mean   :0.06649   Mean   :0.1121   Mean   :0.0339   Mean   :0.08344
3rd Qu.:0.000   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.00000
Max.   :1.000   Max.   :1.0000   Max.   :1.00000   Max.   :1.0000   Max.   :1.0000   Max.   :1.00000

     Fresh             Citrus           Vanilla             Musk             Aromatic          Jasmine
Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.000000   Min.   :0.00000   Min.   :0.000000
1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.:0.00000   1st Qu.:0.000000
Median :0.00000   Median :0.00000   Median :0.00000   Median :0.000000   Median :0.00000   Median :0.000000
Mean   :0.04563   Mean   :0.08605   Mean   :0.02347   Mean   :0.006519   Mean   :0.01434   Mean   :0.006519
3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.000000   3rd Qu.:0.00000   3rd Qu.:0.000000
Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.000000   Max.   :1.00000   Max.   :1.000000

     Rose      Sandalwood          Clean
Min.   :0   Min.   :0.000000   Min.   :0.000000
1st Qu.:0   1st Qu.:0.000000   1st Qu.:0.000000
Median :0   Median :0.000000   Median :0.000000
Mean   :0   Mean   :0.001304   Mean   :0.001304
3rd Qu.:0   3rd Qu.:0.000000   3rd Qu.:0.000000
Max.   :0   Max.   :1.000000   Max.   :1.000000
```

1. Price
2. Gender
3. Concentration
4. Note Diversity

```
mean_price <- mean(profumo$price, na.rm = TRUE)
median_price <- median(profumo$price, na.rm = TRUE)
sd_price <- sd(profumo$price, na.rm = TRUE)
cat("Mean Price:", mean_price, "\nMedian Price:", median_price, "\nSD Price:", sd_price, "\n\n")
```

```
> cat("Mean Price:", mean_price, "\nMedian Price:", median_price, "\nSD Price:", sd_price, "\n\n")
Mean Price: 154.875
Median Price: 139.95
SD Price: 108.8005
```

```
mean_sex <- mean(profumo$Gender_encoded, na.rm = TRUE)
median_sex <- median(profumo$Gender_encoded, na.rm = TRUE)
sd_sex <- sd(profumo$Gender_encoded, na.rm = TRUE)
cat("Mean Gender:", mean_sex, "\nMedian Gender:", median_sex, "\nSD Gender:", sd_sex, "\n\n")
```

```
> mean_sex <- mean(profumo$Gender_encoded, na.rm = TRUE)
> median_sex <- median(profumo$Gender_encoded, na.rm = TRUE)
> sd_sex <- sd(profumo$Gender_encoded, na.rm = TRUE)
> cat("Mean Gender:", mean_sex, "\nMedian Gender:", median_sex, "\nSD Gender:", sd_sex, "\n\n")
Mean Gender: 0.4563233
Median Gender: 0
SD Gender: 0.4984137
```

```
mean_cont <- mean(profumo$scent_strength, na.rm = TRUE)
median_cont <- median(profumo$scent_strength, na.rm = TRUE)
sd_cont <- sd(profumo$scent_strength, na.rm = TRUE)
cat("Mean Concentration:", mean_cont, "\nMedian Concentration:", median_cont, "\nSD Concentration:", sd_cont,
"\n\n")
```

```
> mean_cont <- mean(profumo$scent_strength, na.rm = TRUE)
> median_cont <- median(profumo$scent_strength, na.rm = TRUE)
> sd_cont <- sd(profumo$scent_strength, na.rm = TRUE)
> cat("Mean Concentration:", mean_cont, "\nMedian Concentration:", median_cont, "\nSD Concentration:", sd_cont, "\n
\n")
Mean Concentration: 1.490196
Median Concentration: 1
SD Concentration: 0.5002309
```

```
mean_div <- mean(profumo$notes_diversity, na.rm = TRUE)
median_div <- median(profumo$notes_diversity, na.rm = TRUE)
sd_div <- sd(profumo$notes_diversity, na.rm = TRUE)
cat("Mean Notes Diversity:", mean_div, "\nMedian Notes Diversity:", median_div, "\nSD Notes Diversity:", sd_div,
"\n\n")
```

```
> mean_div <- mean(profumo$notes_diversity, na.rm = TRUE)
> median_div <- median(profumo$notes_diversity, na.rm = TRUE)
> sd_div <- sd(profumo$notes_diversity, na.rm = TRUE)
> cat("Mean Notes Diversity:", mean_div, "\nMedian Notes Diversity:", median_div, "\nSD Notes Diversity:", sd_div, "\n
\n")
Mean Notes Diversity: 6.975228
Median Notes Diversity: 6
SD Notes Diversity: 3.586042
```

Step 6: Cross tabs. Check all quantitative variables in a correlation/crosstabs analysis. Think about what type of data you are working with and what type of analysis is necessary. Also think about potential multicollinearity.

```
> CrossTable(profumo$Gender_encoded, profumo$scent_strength, expected = TRUE, format = "SPSS")

   Cell Contents
|-------------------------|
|                   Count |
|         Expected Values |
| Chi-square contribution |
|             Row Percent |
|          Column Percent |
|           Total Percent |
|-------------------------|

Total Observations in Table:  817

                     | profumo$scent_strength
profumo$Gender_encoded |         1 |         2 | Row Total |
-----------------------|-----------|-----------|-----------|
                     0 |       100 |       316 |       416 |
                       |   202.654 |   213.346 |           |
                       |    51.999 |    49.393 |           |
                       |    24.038% |    75.962% |    50.918% |
                       |    25.126% |    75.418% |           |
                       |    12.240% |    38.678% |           |
-----------------------|-----------|-----------|-----------|
                     1 |       290 |        59 |       349 |
                       |   170.015 |   178.985 |           |
                       |    84.678 |    80.434 |           |
                       |    83.095% |    16.905% |    42.717% |
                       |    72.864% |    14.081% |           |
                       |    35.496% |     7.222% |           |
-----------------------|-----------|-----------|-----------|
                     2 |         8 |        44 |        52 |
                       |    25.332 |    26.668 |           |
                       |    11.858 |    11.264 |           |
                       |    15.385% |    84.615% |     6.365% |
                       |     2.010% |    10.501% |           |
                       |     0.979% |     5.386% |           |
-----------------------|-----------|-----------|-----------|
          Column Total |       398 |       419 |       817 |
                       |    48.715% |    51.285% |           |
-----------------------|-----------|-----------|-----------|


Statistics for All Table Factors


Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  289.6253     d.f. =  2     p =  1.284269e-63


      Minimum expected frequency: 25.3317
```

The CrossTable (gender vs scent strength), it's Chi-square test indicates a very strong association between these 2 variables, with a p value of 1.284269e-63.

Dummy Variables are introduced on the following variables:
Continuous variables (price) are normalized. Categorical variables (concentration, gender and scent) encoding of the– gender (2 categories), concentration as scent strength (4 categories) and scent (15 categories) as dummy variables.

Step 7: Build Model

Model 1 – Price + Scent Strength

```
> perfume_1 <- glm(
+     Gender_encoded ~ price_normalized + scent_strength,
+     family = binomial,
+     data = train_data)
> summary(perfume_1)

Call:
glm(formula = Gender_encoded ~ price_normalized + scent_strength,
    family = binomial, data = train_data)

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)        3.6795     0.3033  12.130   <2e-16 ***
price_normalized   1.2178     0.8180   1.489    0.137
scent_strength    -2.7863     0.2163 -12.880   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 844.95  on 611  degrees of freedom
Residual deviance: 622.34  on 609  degrees of freedom
  (2 observations deleted due to missingness)
AIC: 628.34

Number of Fisher Scoring iterations: 4
```

Model 1 – Diagnostics

```
> exp(coef(perfume_1))
     (Intercept) price_normalized   scent_strength
     39.62714743       3.37974594       0.06165154
> log.pseudo.r2(perfume_1)
Pseudo R^2 for Logistic Regression
Cohen R^2                0.263
Cox and Snell R^2        0.305
Nagelkerke R^2           0.407
```

Model 2 – Price + Scent Strength + Notes Diversity

```
> perfume_2 <- glm(
+     Gender_encoded ~ price_normalized + scent_strength + notes_diversity,
+     family = binomial,
+     data = train_data)
> summary(perfume_2)

Call:
glm(formula = Gender_encoded ~ price_normalized + scent_strength +
    notes_diversity, family = binomial, data = train_data)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       3.626681   0.369799   9.807   <2e-16 ***
price_normalized  1.226484   0.818222   1.499    0.134
scent_strength   -2.783890   0.216517 -12.858   <2e-16 ***
notes_diversity   0.006751   0.027282   0.247    0.805
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 844.95  on 611  degrees of freedom
Residual deviance: 622.28  on 608  degrees of freedom
  (2 observations deleted due to missingness)
AIC: 630.28

Number of Fisher Scoring iterations: 4
```

Model 2 – Diagnostics

```
> exp(coef(perfume_2))
     (Intercept) price_normalized   scent_strength  notes_diversity
     37.58783837       3.40922229       0.06179763       1.00677389
> log.pseudo.r2(perfume_2)
Pseudo R^2 for Logistic Regression
Cohen R^2                 0.264
Cox and Snell R^2         0.305
Nagelkerke R^2            0.407
```

Model 3 – Price + Scent Strength + Notes Diversity + Scent – Woody + Scent – Floral

```
> perfume_3 <- glm(
+     Gender_encoded ~ price_normalized + scent_strength + notes_diversity + Woody + Floral,
+     family = binomial,
+     data = train_data)
> summary(perfume_3)

Call:
glm(formula = Gender_encoded ~ price_normalized + scent_strength +
    notes_diversity + Woody + Floral, family = binomial, data = train_data)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       4.1670757  0.4369226   9.537  < 2e-16 ***
price_normalized  0.9876645  0.8514396   1.160   0.2461
scent_strength   -2.7582750  0.2393813 -11.523  < 2e-16 ***
notes_diversity  -0.0001092  0.0302930  -0.004   0.9971
Woody             0.6103181  0.2929469   2.083   0.0372 *
Floral           -2.2045377  0.2773274  -7.949 1.88e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 844.95  on 611  degrees of freedom
Residual deviance: 520.42  on 606  degrees of freedom
  (2 observations deleted due to missingness)
AIC: 532.42

Number of Fisher Scoring iterations: 5
```

Model 3 – Diagnostics

```
> exp(coef(perfume_3))
     (Intercept) price_normalized   scent_strength  notes_diversity           Woody
     64.52648222       2.68495636       0.06340104       0.99989076       1.84101700
          Floral
      0.11030151
> log.pseudo.r2(perfume_3)
Pseudo R^2 for Logistic Regression
Cohen R^2                 0.384
Cox and Snell R^2         0.412
Nagelkerke R^2            0.55
```

Model 4 – Price + Scent Strength + Notes Diversity + 4x Gender Exclusive Notes

```
> perfume_4 <- glm(
+     Gender_encoded ~ price_normalized + scent_strength + notes_diversity + clary_sage_present +
+         fir_present + peony_present + jasmine_sambac_present,
+     family = binomial,
+     data = train_data
+ )
> summary(perfume_4)

Call:
glm(formula = Gender_encoded ~ price_normalized + scent_strength +
    notes_diversity + clary_sage_present + fir_present + peony_present +
    jasmine_sambac_present, family = binomial, data = train_data)

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)             3.738e+00  3.950e-01   9.464   <2e-16 ***
price_normalized        1.135e+00  8.383e-01   1.353    0.176
scent_strength         -2.759e+00  2.284e-01 -12.080   <2e-16 ***
notes_diversity        -7.287e-03  3.022e-02  -0.241    0.809
clary_sage_present      1.877e+01  1.465e+03   0.013    0.990
fir_present             1.723e+01  1.628e+03   0.011    0.992
peony_present          -1.811e+01  9.531e+02  -0.019    0.985
jasmine_sambac_present -1.643e+01  1.789e+03  -0.009    0.993
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 844.95  on 611  degrees of freedom
Residual deviance: 559.66  on 604  degrees of freedom
  (2 observations deleted due to missingness)
AIC: 575.66

Number of Fisher Scoring iterations: 17
```

Model 4 – Diagnostics

```
> exp(coef(perfume_4))
          (Intercept)       price_normalized         scent_strength        notes_diversity
         4.202351e+01           3.109926e+00           6.332540e-02           9.927400e-01
    clary_sage_present            fir_present          peony_present jasmine_sambac_present
         1.412199e+08           3.051236e+07           1.362183e-08           7.297262e-08
> log.pseudo.r2(perfume_4)
Pseudo R^2 for Logistic Regression
Cohen R^2               0.338
Cox and Snell R^2       0.373
Nagelkerke R^2          0.498
```

Model 5 – Price + Scent Strength + Notes Diversity + 5x Gender Common Notes

```
> perfume_5 <- glm(
+     Gender_encoded ~ price_normalized + scent_strength + notes_diversity +
+         patchouli_present + amber_present + salwood_present +
+         sandalwood_present + musk_present,
+     family = binomial,
+     data = train_data
+ )
>
> summary(perfume_5)

Call:
glm(formula = Gender_encoded ~ price_normalized + scent_strength +
    notes_diversity + patchouli_present + amber_present + salwood_present +
    sandalwood_present + musk_present, family = binomial, data = train_data)

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)        3.9919801  0.4050890   9.855  < 2e-16 ***
price_normalized   0.6764967  0.8668271   0.780   0.4351
scent_strength    -2.8932443  0.2300639 -12.576  < 2e-16 ***
notes_diversity    0.0198476  0.0369293   0.537   0.5910
patchouli_present  0.5808284  0.2280876   2.547   0.0109 *
amber_present      0.2823665  0.2238857   1.261   0.2072
salwood_present   -0.0741853  0.4625411  -0.160   0.8726
sandalwood_present 0.0005664  0.4564817   0.001   0.9990
musk_present      -1.0065848  0.2221005  -4.532 5.84e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 844.95  on 611  degrees of freedom
Residual deviance: 589.77  on 603  degrees of freedom
  (2 observations deleted due to missingness)
AIC: 607.77

Number of Fisher Scoring iterations: 4
```

Model 5 – Diagnostics

```
> exp(coef(perfume_5))
       (Intercept)   price_normalized     scent_strength    notes_diversity  patchouli_present
         54.1620301          1.9669746          0.0553962          1.0200458          1.7875186
     amber_present    salwood_present sandalwood_present       musk_present
         1.3262647          0.9284996          1.0005665          0.3654650
> log.pseudo.r2(perfume_5)
Pseudo R^2 for Logistic Regression
Cohen R^2                0.302
Cox and Snell R^2        0.341
Nagelkerke R^2           0.455
```

Model 6 – Scent Strength + Significant Gender Common Notes + Significant Scents

```
> perfume_6 <- glm(
+     Gender_encoded ~ scent_strength + patchouli_present +
+         musk_present + Woody + Floral,
+     family = binomial,
+     data = train_data)
> summary(perfume_6)

Call:
glm(formula = Gender_encoded ~ scent_strength + patchouli_present +
    musk_present + Woody + Floral, family = binomial, data = train_data)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)         4.6792     0.4288  10.911  < 2e-16 ***
scent_strength     -2.8416     0.2437 -11.659  < 2e-16 ***
patchouli_present   0.7422     0.2448   3.032  0.00243 **
musk_present       -0.9580     0.2334  -4.104 4.06e-05 ***
Woody               0.5925     0.2922   2.028  0.04260 *
Floral             -2.2454     0.2887  -7.777 7.46e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 844.95  on 611  degrees of freedom
Residual deviance: 493.67  on 606  degrees of freedom
  (2 observations deleted due to missingness)
AIC: 505.67

Number of Fisher Scoring iterations: 5
```

Model 6 – Diagnostics

```
> exp(coef(perfume_6))
      (Intercept)    scent_strength patchouli_present      musk_present             Woody
     107.68433650         0.05833364         2.10050877         0.38365343         1.80845204
           Floral
        0.10588949
> log.pseudo.r2(perfume_6)
Pseudo R^2 for Logistic Regression
Cohen R^2                0.416
Cox and Snell R^2        0.437
Nagelkerke R^2           0.583
```

Modelling Process

1. Price + Scent Strength
    - Price is not significant with p-value at > 0.05
    - Scent strenghh is highly significant with p-value at < 0.05
    - Nagelkerke at 0.407, explaining 40.7% of the variance in perfume gender.

2. Price + Scent Strength + Notes Diversity
    a. Adding notes diversity with p = 0.805, shows that it is not statistically significant.
    b. Nagelkerke remained at 0.407

3. Price + Scent Strength + Notes Diversity + Scent – Woody + Scent – Floral
    a. Top 2 scents are added. Nagelkerke increased to 0.55
    b. This changed scent strength to become statistically significant, along with floral and to a lesser extent, woody scents

4. Price + Scent Strength + Notes Diversity + 4x Gender Exclusive Notes
    a. Top 2 scent exchanged with 4x gender exclusive notes, none are significant
    b. Nagelkerke decreased to 0.498

5. Price + Scent Strength + Notes Diversity + 5x Gender Common Notes
    a. Replaced 4x gender exclusive notes with 5x Gender Common Notes, 2 common notes are significant
    b. Nagelkerke decreased further to 0.455

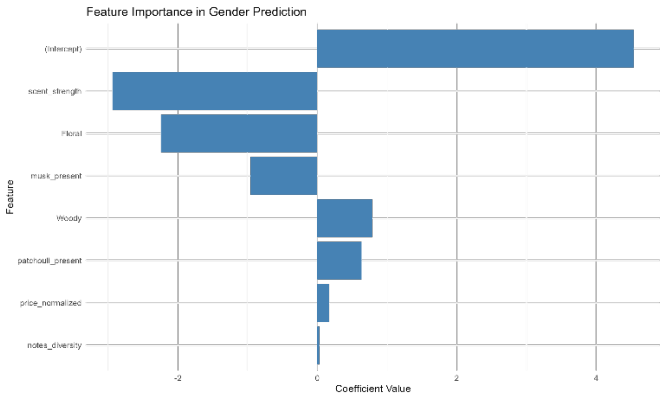
6. Scent Strength + Significant Gender Common Notes + Significant Scents
    a. Removed price and notres diversity and experimented solely with significant common notes – notes above 100 counts and significant scents – scents that appear the most in the datatset
    b. All predictors become significant
    c. Nagelkerke increased to 0.583

Final Model – Price + Scent Strength + Notes Diversity + 5x Gender Common Notes

a. Most predictor are significant, except for price and notes diversity, they are inserted due to literature review
b. Nagelkerke increased to 0.584

```
> perfume_final <- glm(
+     Gender_encoded ~ price_normalized + scent_strength + notes_diversity +
patchouli_present + musk_present + Woody + Floral,
+     family = binomial,
+     data = train_data)
> summary(perfume_final)

Call:
glm(formula = Gender_encoded ~ price_normalized + scent_strength +
    notes_diversity + patchouli_present + musk_present + Woody +
    Floral, family = binomial, data = train_data)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        4.53787    0.47090   9.637  < 2e-16 ***
price_normalized   0.66169    0.87817   0.753  0.45116
scent_strength    -2.89285    0.25539 -11.327  < 2e-16 ***
notes_diversity    0.01825    0.03319   0.550  0.58243
patchouli_present  0.71376    0.25155   2.838  0.00455 **
musk_present      -0.98009    0.24556  -3.991 6.57e-05 ***
Woody              0.56780    0.29390   1.932  0.05337 .
Floral            -2.25238    0.29001  -7.767 8.06e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 844.95  on 611  degrees of freedom
Residual deviance: 492.81  on 604  degrees of freedom
  (2 observations deleted due to missingness)
AIC: 508.81

Number of Fisher Scoring iterations: 5
```

Final Model – Diagnostics




```
> exp(coef(perfume_final))
      (Intercept)  price_normalized    scent_strength   notes_diversity patch
ouli_present
      93.49116285        1.93805713        0.05541828        1.01841471
2.04165477
     musk_present             Woody             Floral
       0.37527896        1.76438742        0.10514873
> log.pseudo.r2(perfume_final)
Pseudo R^2 for Logistic Regression
Cohen R^2                 0.417
Cox and Snell R^2         0.438
Nagelkerke R^2            0.584
```

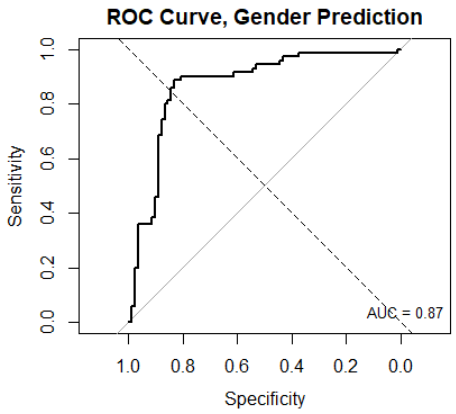
Final Model – Train Data Diagnostics

```
> train_data$prob_final <- predict(perfume_final, newdata = train_data, type = "response")
> quantile(train_data$prob_final)
       0%        25%        50%        75%       100%
0.01129352 0.10207279 0.37955711 0.84609203 0.96991485
```

```
> table(train_data$pred_final)

  0   1
337 276
```

```
> accur_final <- 1 - mean(train_data$pred_final != train_data$Gender_encoded)
> accur_final
[1] 0.8352365
```

The accuracy of the model on training data stands at 83.52%



The area under the curve stands at 0.87. Where the close the value is to 1, the better the model. The remaining gap might be use omitted variables bias. As the sample of 760 observations is considered moderate, the potential bias of having a small sample size can be ruled out.

Step 8: Test Model

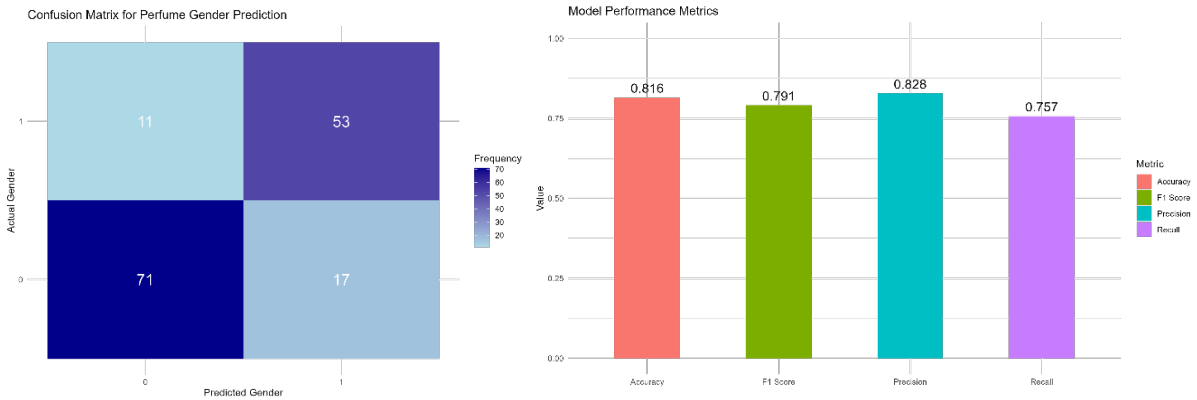
Final Model – Test Data Diagnostics

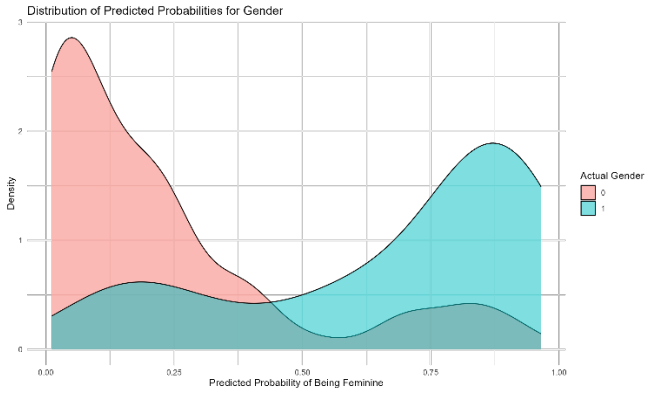
```
> quantile(test_data$prob_final)
        0%        25%        50%        75%       100%
0.01129569 0.10845765 0.29965279 0.82427478 0.96521733
```

```
> table(test_data$pred_final)

 0  1
88 64
```

```
> accur_final <- 1 - mean(test_data$pred_final != test_data$Gender_encoded)
>
> accur_final
[1] 0.8157895
```

The accuracy of the model on test data stands at 81.58%






```
> cat("Precision:", precision, "\n")
Precision: 0.8955224
> cat("Recall:", recall, "\n")
Recall: 0.8450704
> cat("F1 Score:", f1_score, "\n")
F1 Score: 0.8695652
```

PCA


```
> print("Loadings Matrix:")
[1] "Loadings Matrix:"
> print(loadings)
                      PC1          PC2         PC3
price           0.63277315 -0.234805033 -0.04817566
scent_strength  0.33125666  0.447707566  0.27366577
price_normalized 0.63277315 -0.234805033 -0.04817566
notes_diversity -0.06692069 -0.085046612 -0.18042235
Woody           0.02015684 -0.428551918  0.31635602
Floral          0.18981913  0.564999858 -0.45923082
Arabian         0.01638684  0.031433525  0.28256112
Spicy          -0.13197040 -0.214549350 -0.16739828
Oriental       -0.04299850  0.175252772  0.18170648
Fruity         -0.01241147  0.056722909  0.39975962
Fresh          -0.08903387  0.000945395  0.09682102
Citrus         -0.10554499 -0.248265069 -0.38559489
Vanilla         0.04118764 -0.057689391  0.05573523
Musk           -0.02812265  0.110394687  0.23381527
Aromatic       -0.01799333 -0.113127415 -0.18611369
Jasmine         0.04989190 -0.013190841  0.06753055
Rose            0.05825794  0.066707124 -0.12126575
Sandalwood     -0.02960905  0.045447892  0.04644281
Clean          -0.01496576  0.039842813  0.06600201
```