

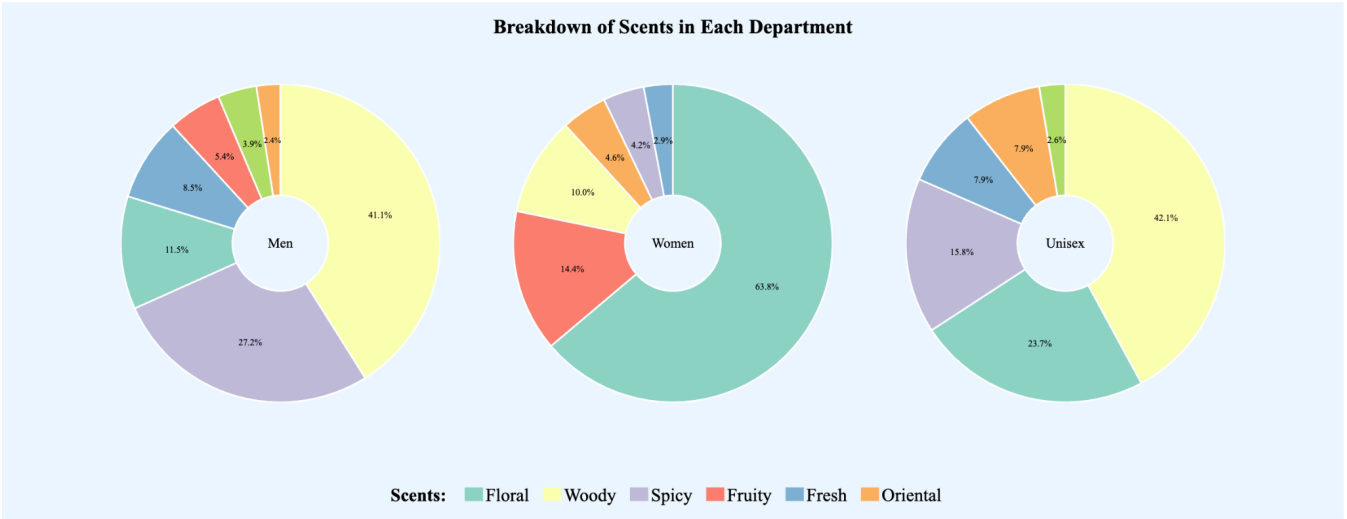
INFO 3300 Project 1 Final Report

Amanda Chen (ac2534), Vanessa Tam (vt259), Monica Nguyenduy (mn473), and Jenny Yu (jy865)

Chart Visualizations

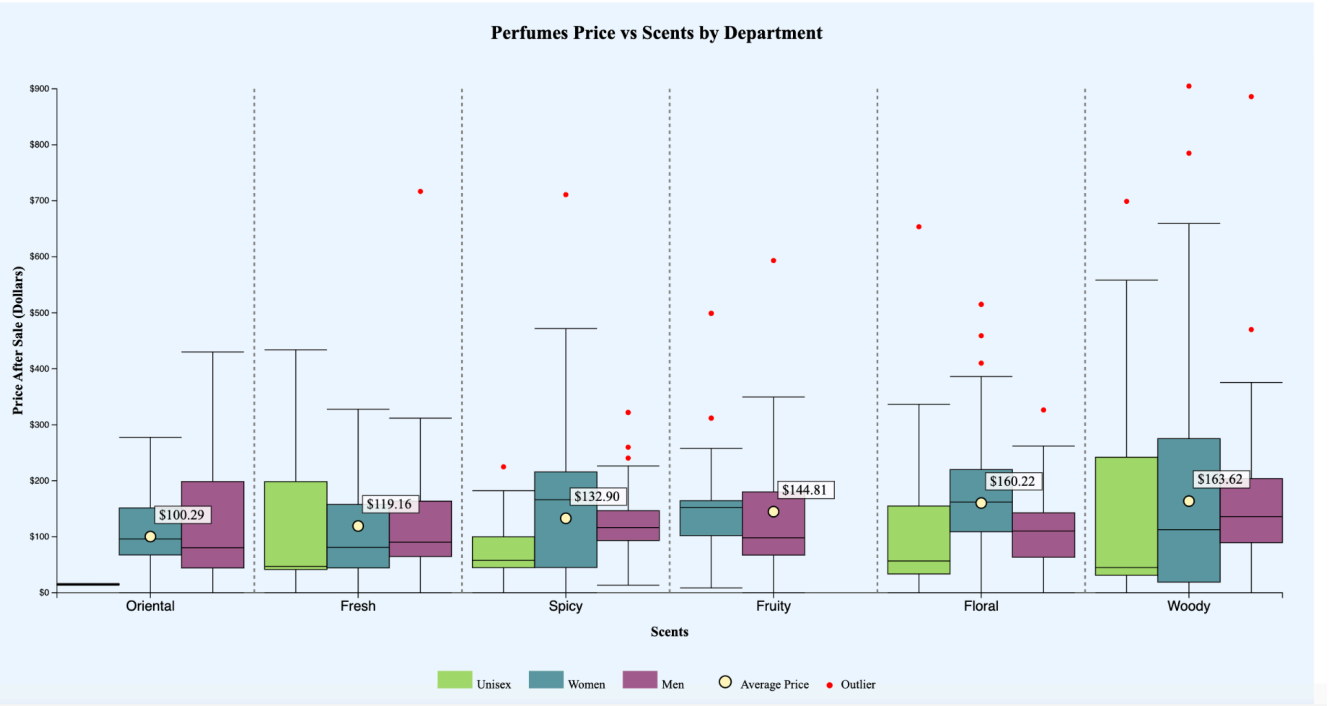
Scents Across Different Departments: Men, Women, & Unisex

This chart aims to explore the proportion of scents of the perfumes sold in each of the three different departments during retail to answer the following question: Which types of scents are targeted to be more for men or more for women? This may be a peek into which smells and notes are traditionally considered masculine, feminine, or androgynous. If there's an overwhelming majority for one specific scent in any of the departments, it could be a sign to shop in that department specifically if they specialize and have a diversity of options in that scent that you prefer.



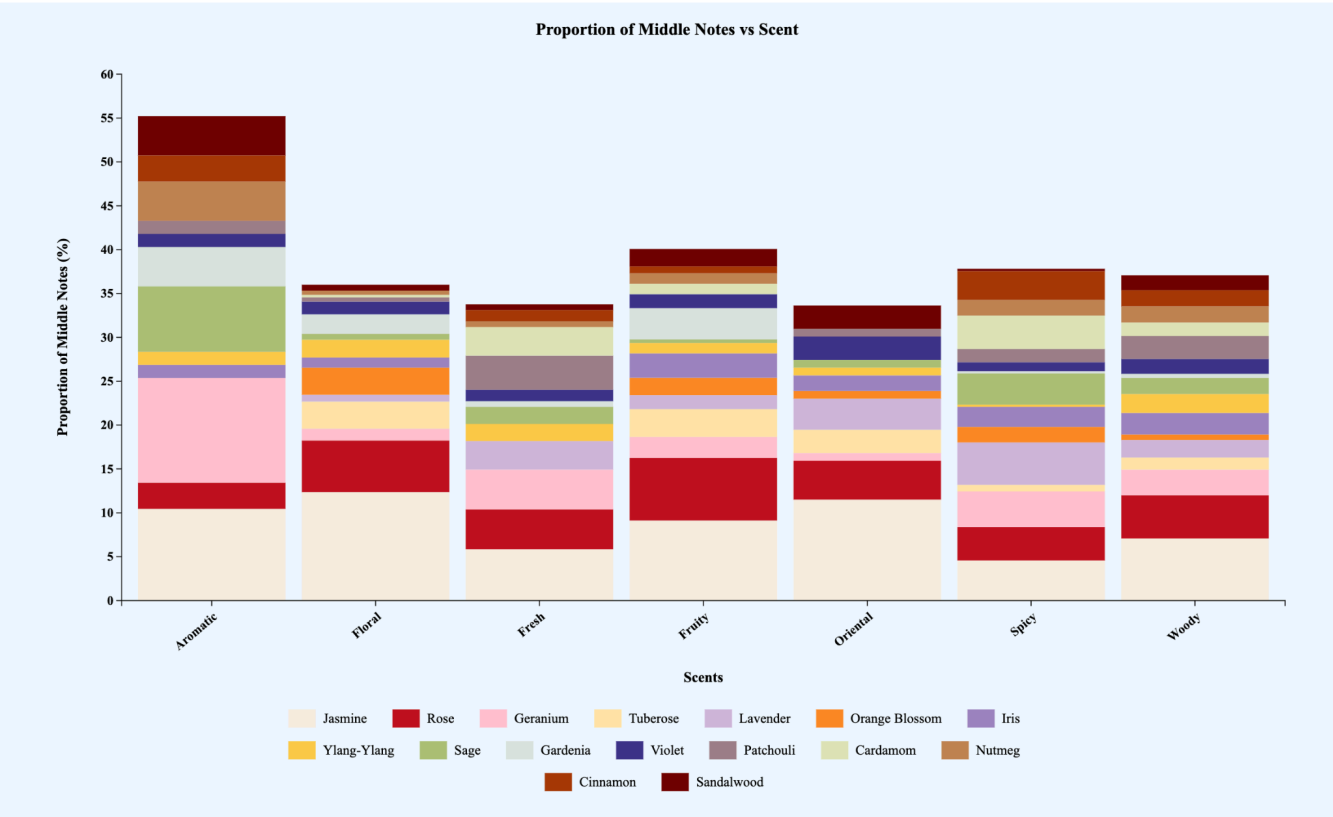
Correlation between Perfume Price, Scent, and Department

How do scent categories relate to pricing across men's, women's, and unisex departments? Are some scent types more consistently priced across departments, or do we see significant differences? This chart explores price variations among scent groups within each department. Differences in central or distributions of prices could suggest targeted price structuring based on higher demand for certain scents or reflect branding and marketing strategies.



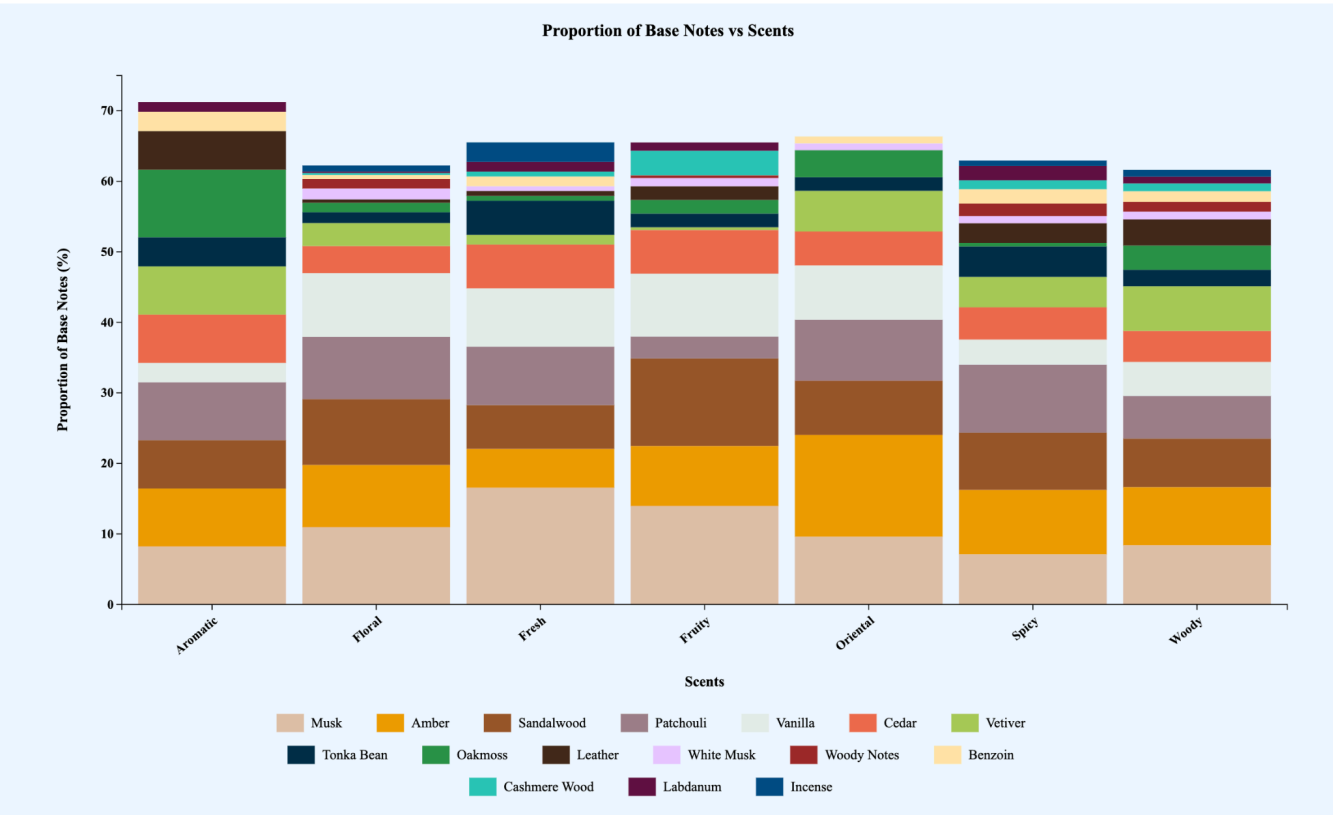
Middle Notes in Scent Families

We pulled the top 15 most frequently occurring middle notes out of all the perfumes, and calculated how much they appear in each of 7 scent groups to answer the following question: How much do the 15 most common middle notes take up in different perfume scents? The greater the total proportion of the scent group, the more that the top 15 middle notes are used, as opposed to less popular notes. Additionally, we can see just how many perfumes have each of the 15 middle notes out of each scent group!



Base Notes in Scent Families

Following the same idea as before, we pulled the top 15 most frequently occurring base notes out of all the perfumes and calculated how much they appear in each of 7 scent groups. How much variation in note use is there between each scent group? Which scent group's perfumes have a smaller share of popular notes, thus leaving more variation in less common notes? Lastly, how much do the 15 most common base notes take up in different perfume scents?



Data Description

For our project, we sourced our dataset from [Kaggle](#), which focused on the perfume section of the leading lifestyle shopping destination, noon, in Saudi Arabia.

The variables in our dataset that were used in our visualization are listed as follows:

- **new_price**: The discounted price of the perfume.
- **department**: The target gender for each perfume (e.g., feminine for women, masculine for men, unisex, and kids).
- **scents**: The general scent profile of each perfume.
- **base_note**: The base notes that provide depth and stability to the fragrance.
- **middle_note**: The main thematic scent of the perfume.

First, we filtered the scents. We started with 22 unique scent groups and created a new dataset with the top 10 so we could focus our analysis on the most commonly referenced scents. After retrieving the new dataset, we manually removed the following scents: Arabian and Vanilla since these are notes that can be combined with larger scent categories (oriental scents). We removed “Aromatic” scents when visualizing for the pie chart and box plots as it is technically a subcategory of fresh scents, but kept it for the stacked bar charts since there was an interesting disparity between the aromatic scent and the other scent groups.

For both the pie chart and boxed plot, we cleaned the dataset by removing Kids from Department since there was only 1 record and then filtered the data based on the scents in the following list: Men, Women, Unisex. Since in the pie chart we would create new arrays for counts of the scents for each department, we trimmed the ends of the scents and department of possible white spaces.

For our stacked bar chart, we encountered some challenges with messy data, particularly with the middle and base notes. For example, some entries used commas to separate notes while others used "and" or other delimiters. We addressed the misspellings and inconsistencies by filtering out entries with missing or incomplete values in those respective fields. Then, we found the frequency of the distinct notes, which totaled 656, and from this, we found the top 15 most common notes create a more concise and interpretable visual representation.

Design Rationale

Pie Chart

We choose to use a pie chart to explore the distribution of perfume scents across key departments: Men, Women, and Unisex. This visualization allowed for us to highlight which scents were predominant in each department, showing how traditionally masculine, feminine or androgynous scents are represented. By splitting the departments into three different pie charts, it allowed for easy comparison and identification of the predominant scents in each department.

The visual channels used in the pie charts were positioned or tilted for the angular position of the pie slice, size and shape of the slice shape, and the color hues for the scents. The slice size is proportional to the percentage of perfumes featuring the particular scent in each department, intuitively indicating large slices with more prevalent scents and smaller slices with less prevalent scents. As for the color scheme, we used categorical color schemes to

distinct scents. We used text marks to indicate the three department's pie charts and for the percentage of the pie slice.

One of the trade-offs of the pie chart is that it provides a simple overview of the scent proportions in each department and does not display the exact count of the scents in the dataset. However, we had intentionally chosen to display the general trend in each department's scent proportion to maintain visual clarity with the absence of extra text displaying count. With the decision to display the percentage breakdown, this introduced another trade-off. Without the text displaying the percentage of each slice, it reduced the clarity of the pie chart and eliminated the ability to see the exact breakdown of the proportions. However with adding the text for the percentage, a trade-off would be possible crowding especially the areas surrounded by small slices. We found a compromise to show detailed information visualization by adding the text representing the percentages and decreasing the font size for the smaller slices.

Box Plot

We chose a box plot to display the distribution of perfume prices, including max and min values, quartiles, and outliers, across different scent groups and departments. This was intended to help us understand the center and spread of prices between departments, across different perfume scents. The scent groups are represented along the X-axis, and prices are displayed on the Y-axis. The boxes and whiskers for each scent group represent the price ranges, with color-coded boxes used to differentiate the departments: Unisex, Women, and Men. We used analogous colors to indicate related departments while keeping them distinct enough for comparison. Yellow circles indicate the average price for each group and are accompanied by translucent labels to ensure the price values remain visible without obscuring the box plots behind them.

The tradeoff in using this box plot is that it provides a summary of the data, e.g. key metrics like quartiles and outliers with a consistent structure that makes it easier to compare the pricing trends across departments within the same scent group; however, it doesn't show individual data points, such as the number of perfumes sold in each category. This could be important for more clearly analyzing how popularity (number sold) may impact price. Another trade-off involves the placement of the average price markers, which may seem connected to specific scent groups due to their location, but we mitigated this by using a distinctive yellow color for the averages to differentiate them from the department-based color coding.

The visual channels we used are vertically aligned length, horizontally and vertically aligned positions, and color hue. The vertically aligned length is shown as the length of the whiskers extending from the boxes visually encodes the range of perfume prices, indicating the minimum and maximum prices within each scent group and department. The horizontally aligned positions shown are on the x-axis of the different scent groups. The vertically aligned positions shown are on the y-axis of the price after sale in dollars, with the vertical positioning of boxes and whiskers showing the price distribution within each scent group. The color hue represents the different departments (Unisex, Women, and Men), with intuitive colors representing each category. The marks we used are circles and boxes. Circles are used for both the outliers (red circles) and the average price (yellow circles) to distinguish these key points from the rest of the plot. Boxes are used to represent quartiles/min max values.

Stacked Bar

We chose stacked bar plots to represent the proportion of middle and base notes used in different perfume scent groups, effectively showing both the distribution of popular notes and the variation across scent categories. We decided to use stacked bars because they allow us to visualize both categorical variables at once: the total use of

the top 15 notes and the breakdown of individual note usage. This design helps us compare the proportion of each note within a given scent group, such as how much of the Floral scent group is composed of jasmine, while also providing an overview of how the most common notes are used across all scent groups. The stacked format also enables us to display both the overall trends and the specific contribution of individual notes to each scent category.

The visual channels we used are horizontally aligned positions, vertically aligned length, and color hue. The horizontally aligned positions are shown on the x-axis, representing the different scent groups. The vertically aligned lengths are on the y-axis; the length of each bar shows the proportion of notes within each scent group. The color hue represents different notes with their natural associations that viewers may have (ie. rose with red, cedar with brown, jasmine with off-white, etc.) to enhance the intuitive understanding of the chart, and the consistency across the scent groups allows users to easily compare the usage of notes across the groups, supported by a legend that corresponds the colors to their respective notes.

One tradeoff is that it is more difficult to compare specific notes across categories when their positions within the stack vary since the baseline for comparison isn't consistent across groups. Additionally, while the stacked bar format provides an overall picture, it doesn't show the absolute number of notes and perfumes each scent group contains, which could be useful for further analysis. Lastly, assigning colors for each note required careful selection to avoid confusion, but with many colors involved, the chart could become overwhelming for some viewers, especially with subtle color differences like the multiple shades of greens. We also considered creating a segment for each bar representing "other notes", notes that are not in the top 15 notes included in our visualization, to represent what share of 100% the less common, niche notes (out of 600+ total notes in the dataset) take up, but increasing the y-axis scale to 100 would have decreased visibility of each of the individual 15 notes, so we decided against this.

The Story

Pie Chart

For the pie chart, we wanted to investigate which scents are targeted more towards men, women, and unisex by analyzing the distribution of the scents sold in each department. Surprisingly, we found that both men and unisex departments have a dominance in woody scents, with a percentage of 41.1% and 42.1% respectively. However in the women's department, the scent that is most popularly targeted to women is floral scents with 63.8%. This is not surprising since floral scents are a popular choice for women's perfumes. The second and third most prominent scents in the men department are spicy with 27.2% and floral with 11.5%. Interestingly for unisex, the second and third scents are about the same proportion to mens but flipped, with floral being 23.7% and spicy being 15.8%. Once again, the scents in the womens department differ since the following two prominent scents are fruity being 14.4% and woody being 10%.

Furthermore, the findings show that the men's department is the most diverse out of the three since it includes all of the 7 scents: floral, woody, spicy, fruity, fresh, oriental, and aromatic. However, the women's department does not sell aromatic scents and the unisex department does not sell fruity scents. Although the primary scents sold in mens and unisex align, the finding overall shows the unique different distribution of scents sold in each department.

Box Plot

From the box plot, we wanted to explore whether there were differences in how perfumes are priced across departments. One surprising finding is that perfume prices seem to correlate with the popularity (based on the previous visualization) of certain scents. This suggests that the more popular a scent is, the higher its central prices tend to be, both within and across departments. For example, woody and floral scents are among the top 3 most popular scents across all 3 departments, and have the highest average prices within their scent categories in the box plot. The median price of spicy perfumes for men is higher than for unisex, and spicy scents are proportionally more popular in the men's department than women's or unisex. Similarly, fruity perfumes have a higher median price for women than men and are proportionally more popular in women's perfumes than in men's. Similarly, the price ranges appear to be smaller between departments the more popular the scent is within the department, which suggests some possible degree of targeted pricing taking place by popularity. For example, woody and spicy scents are comparatively less popular for women than either men or unisex departments, and the price terms of both the middle 50% and the range is comparatively greater. Fresh and floral scents have similar popularities and similar differences in upper/lower ranges of the central spreads for price.

Another surprising finding is that women's perfumes across each scent category generally have higher median prices (4 out of 7 scent categories) compared to men's and unisex perfumes in the same scent groups. In contrast, unisex perfumes tend to have lower median prices (in all 7 categories) compared to both men's and women's perfumes. Both in a couple cases confounded with our earlier notion that popularity may contribute to the prices of the perfumes, as spicy scents were both less popular and had lower percentages in women's than in both men's or unisex departments, but had a higher median price than both. For woody, fresh, and oriental scents, unisex perfumes within the category were priced lower than women's despite having higher popularity and percentages than women's. This suggests that there may be some departmental correlation with perfume pricing as well and departmental differences in perceived value of the perfumes.

Stacked Bar Chart

For our first stacked bar chart, which displays the proportion of the 15 most common middle notes in different perfume scent groups (excluding aromatic, which can be considered a subcategory of fresh scents), it highlights the differences in middle note composition across the top 7 scent groups: aromatic, floral, fresh, fruity, oriental, spicy, and woody. One insight is that the floral group leans heavily on rose and jasmine, which is not surprising given the association between floral scents and flower-based notes. However, it was surprising to see that the Woody group showed the most even distribution of notes compared to the rest, indicating that a broader mix of middle notes is used to create complexity within woody perfumes.

One of the focuses of the stacked bar charts is to examine how much of perfumes of any given scent group are taken up by the 15 most common notes, which arguably shows a lack of diversity from the most popular scents. Considering that many of these perfumes are high on the mass-produced, commercially available level (e.g., Chanel, YSL, Dior, Burberry, Gucci, Jo Malone, De Marly, etc.), a lot of these notes could potentially share a lot of similarities in their layered fragrances. If we had chosen a dataset of smaller, independent perfumers and brands, it's possible that there would be more variation in the notes used outside of a dominance of a majority falling within 15 notes, so many of them being floral notes.

Another useful insight is seeing how individual notes can be categorized as certain scents (eg., sandalwood and cedar are woody, citrus is fruity, jasmine and rose are floral), but any of these notes could alternatively take up a larger share of perfumes outside of the scent group they actually belong to. We see this with cinnamon and

nutmeg, gourmand notes that are considered oriental, actually taking up the largest share of *aromatic* perfumes over any other scent. This is somewhat expected, considering middle notes traditionally consist of strong floral scents, and stronger scents in general than notes that would be used in the easily fading, short-term top notes. Rose and jasmine are some of these stronger, longer lasting notes.

For our second stacked bar chart, which displays the proportion of the 15 most common base notes in different perfume scent groups, it highlights the differences in base note compositions across the top 6 scent groups.

Interestingly, the top 15 notes account for about 35–40% of the middle notes and 60% of the base notes used in each scent group, indicating that while these are the dominant notes, the remaining percentage of the middle and base notes comes from a larger variety of less common notes. This suggests that while perfumes rely on a core set of popular notes, there is still significant variety and customization when it comes to selecting the remaining notes in different scent groups. All of the scent groups, with the exception of aromatic, fall between 60-65% of using the same top 15 notes, and aromatic has over 70% of its perfumes containing these 15 notes. So from this perspective, aromatic perfumes have somewhat less diversity in note use.

Overall, for both the middle note and base note charts, it's interesting quantifying the exact distribution of exactly how common or uncommon the most ready-on-the-market perfumes share. Many perfumes must rely on notes with established chemistry and formulas for mass production and will include the same notes across different scents even if the perfume itself may not be marketed with that note being its central component. For example, Jo Malone's Orange Blossom has a strong orange blossom note but is supported by notes of vetiver, lilac, and iris in the middle and base notes. At a glance, these two charts may express a depressing review of lack of diversity in the popular perfume market, but one factor that the chart does not take into account is exactly how strong any single note is in the perfume, or more importantly, how strongly that note may be perceived by its users. Some people like Jo Malone's Orange Blossom because of its olfactory similarity to a jasmine fragrance, but others may deem it to be too citrusy and will prefer Diptyque's Do Son for its muskier take on jasmine. This goes to show just how much diversity and combinations of the same notes can result in completely different fragrances and experiences. The same notes on paper usually never translate to the same layers and complexities of smell when used in real life.

Team Contributions

As a team, we worked towards finalizing our visualizations in a cohesive way, ensuring that the data represent our accuracy and flow together. Individually, for creating the visualization, we did the following:

- Amanda: Created stacked bar charts (4 hours), descriptions in index file, wrote stacked bar charts in report, polished chart color schemes and other visual chart elements (2 hours), wrote captions and descriptions in data visualization preview (1 hour).
- Vanessa: Created stacked bar charts (6 hours), made general styling edits to code, created report layout, wrote data description and stacked bar chart in report (1 hour)
- Monica: Created pie charts (3 hrs), formatted and styled index.html (1.5 hr), wrote pie chart in report (1 hr), small code revisions (1.5 hr)
- Jenny: Created box plot (5 hrs), contributed to data description in report and wrote box plot in report (2 hrs)