What patterns can be seen between students preferred programming language?

Quinlan Wilson

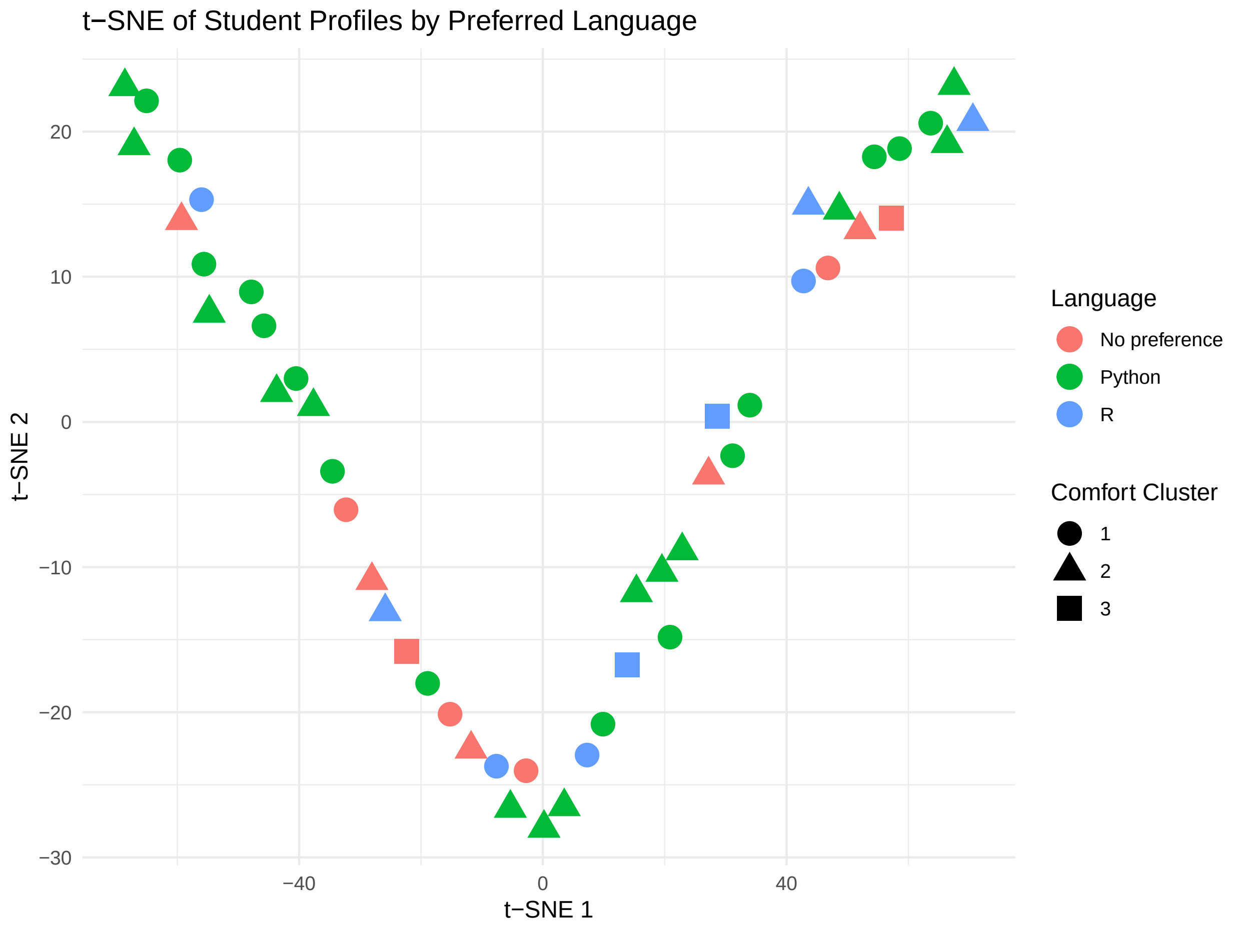
Dimensionality Reduction and Visualization with t-SNE

To explore patterns in students preference between using the programming languages R and Python, I applied t-distributed Stochastic Neighbor Embedding (t-SNE). Which is a reduction technique that works well for visualizing high-dimensional data like the survey data which had many columns. t-SNE projects these features into a 2D graph while preserving local structure, so that students with similar profiles appear closer together in the plot. This visualization helps uncover hidden groupings and relationships that may not be evident through summary statistics or traditional plots.

```r
background <- read_csv(here('data/background-clean.csv'))
interest <- read_csv(here('data/interest-clean.csv'))

df <- inner_join(background, interest, by = "response_id")
```

```r
df$lang[df$response_id == 29] <- "Python"
num_df <- df %>%
  select(where(is.numeric))
```

```r
set.seed(1999)
bad_cols <- sapply(as.data.frame(scale(num_df)), function(x) any(!is.finite(x)))

num_df_clean <- df %>% select(where(is.numeric))
num_df_clean <- num_df_clean[, sapply(num_df_clean, function(x) all(is.finite(x)))]

tsne <- Rtsne(num_df_clean, perplexity = 5)

tsne_df <- data.frame(
  x1 = tsne$Y[,1],
  x2 = tsne$Y[,2],
  lang = df$lang[as.numeric(rownames(num_df_clean))]
)

comf_df <- df %>% select(prog.comf, math.comf, stat.comf)
comf_cluster <- kmeans(comf_df, centers = 3)

tsne_df$comf_cluster <- as.factor(comf_cluster$cluster)

ggplot(tsne_df, aes(x1, x2, color = lang, shape = comf_cluster)) +
  geom_point(size = 5) +
  labs(
    title = "t-SNE of Student Profiles by Preferred Language",
    x = "t-SNE 1",
    y = "t-SNE 2",
    color = "Language",
    shape = "Comfort Cluster"
  ) +
  theme_minimal()
```



Interpretation

The graph shows student profiles in two dimensions (t-SNE 1 and t-SNE 2) which are abstract coordinates that reflect how similar each student profile is to others based on all the numeric features the were fed into t-SNE. Students who are close together in the plot have similar profiles. Points are color coded by preferred programming language and shaped by comfort cluster coming from k-means clustering on programming, math, and statistics comfort.

From the plot we can observe that students who prefer Python tend to cluster together, especially in higher-comfort groups. Suggesting a link between programming confidence and language preference. Students preferring R are more evenly distributed and those who have no preference appear more scattered and often fall into lower comfort clusters. Implying that uncertainty in language choice may be linked to lower confidence.