Class Survey Code

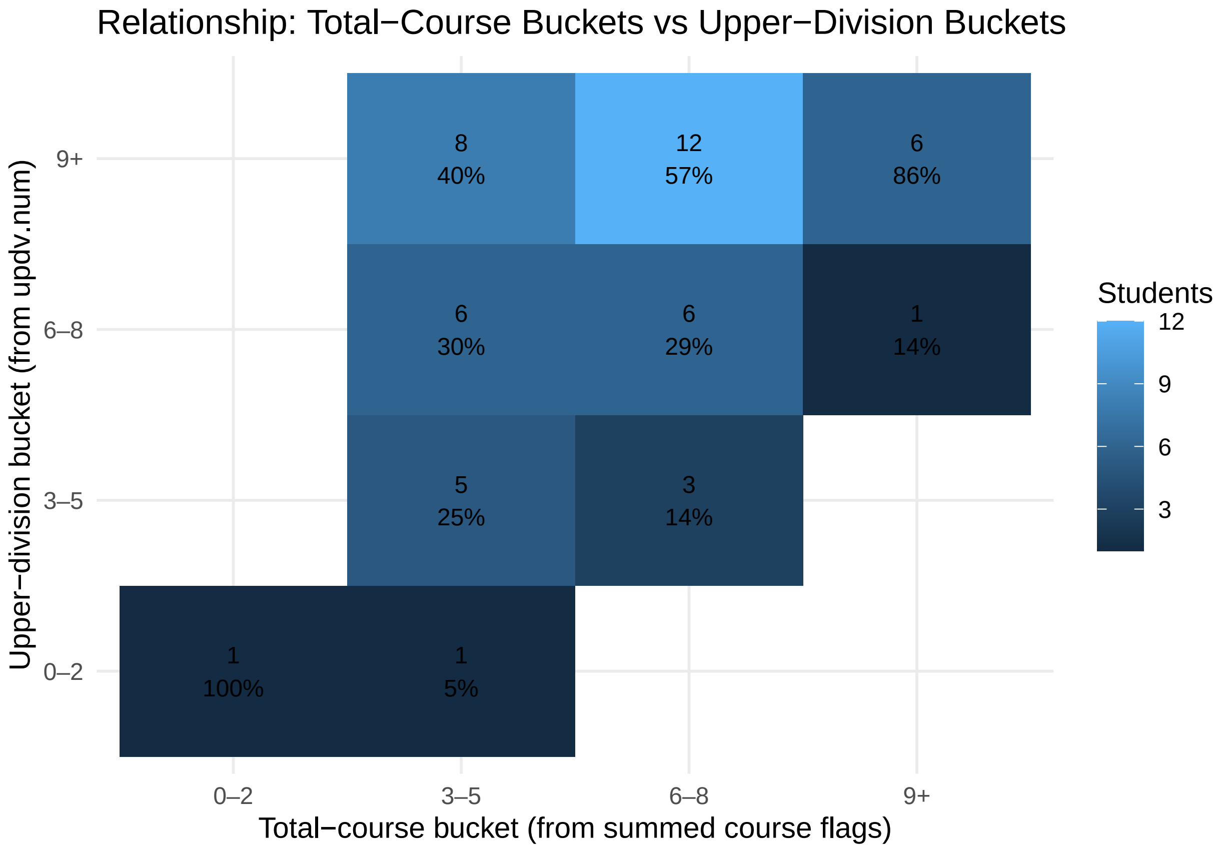
Aidan Frazier

2025-10-16

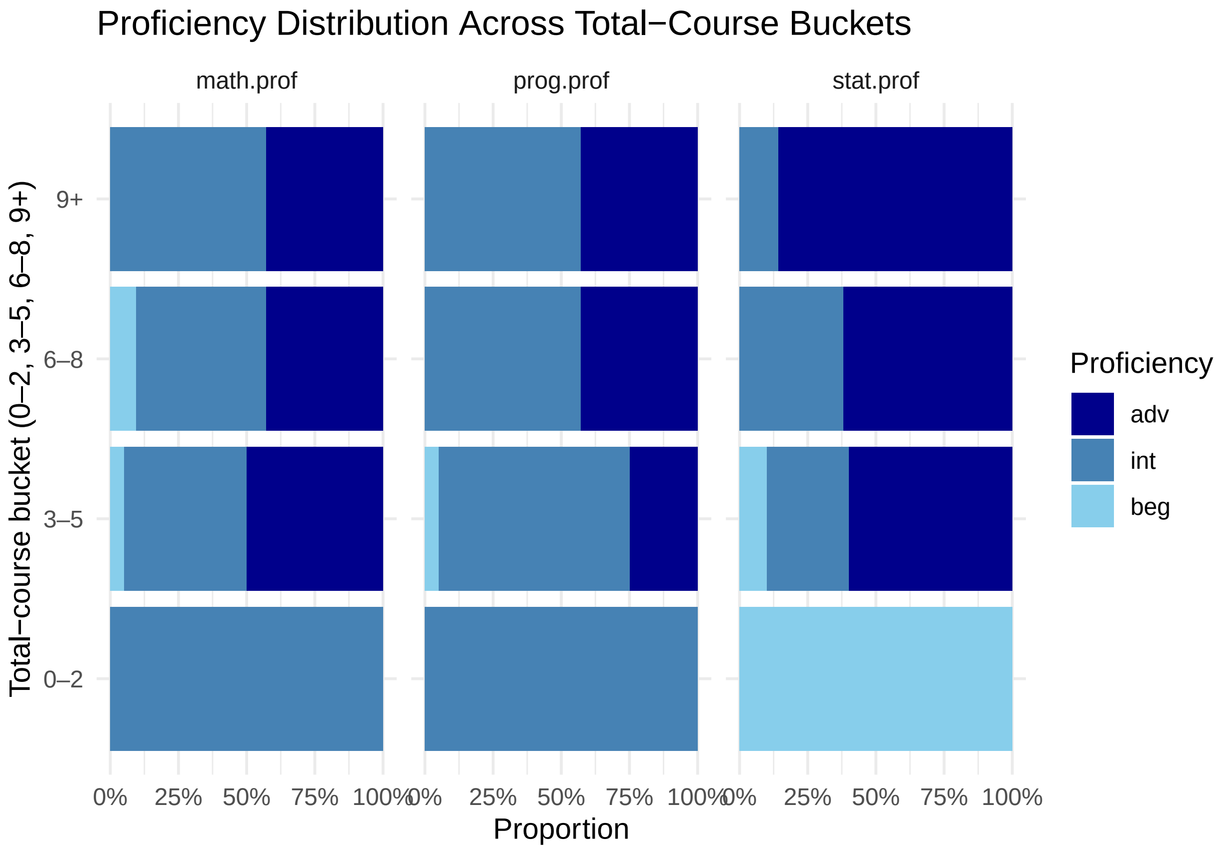
Question: Analysis of courses taken within a subject vs Stat/Prog/Math for both number rating and factors

```r
# Extracting the bucketed values for upper divisions from the merged data set
merged <- merged %>%
  mutate(
    updv_bucket = case_when(
      str_detect(updv.num, "0\\s*-\\s*2") ~ "0-2",
      str_detect(updv.num, "3\\s*-\\s*5") ~ "3-5",
      str_detect(updv.num, "6\\s*-\\s*8") ~ "6-8",
      str_detect(updv.num, "9\\s*\\+") ~ "9+",
      TRUE ~ updv.num
    ),
    updv_bucket = factor(updv_bucket, levels = c("0-2","3-5","6-8","9+"))
  )

course_cols <- names(merged) %>% stringr::str_subset("^(PSTAT|CS|ECON|LING)\\d+$")

merged <- merged %>%
  mutate(
    num_courses = rowSums(across(all_of(course_cols), ~ replace_na(.x, 0)), na.rm = TRUE),
    pstat_courses = rowSums(across(starts_with("PSTAT"), ~ replace_na(.x, 0)), na.rm = TRUE),
    cs_courses = rowSums(across(starts_with("CS"), ~ replace_na(.x, 0)), na.rm = TRUE),
    econ_courses=  rowSums(across(starts_with("ECON"), ~ replace_na(.x, 0)), na.rm = TRUE),
    ling_courses = rowSums(across(starts_with("LING"), ~ replace_na(.x, 0)), na.rm = TRUE)
  )

# Totals of courses which are individualized into the same intervals as upper-division buckets
# sorted by 0-2, 3-5, 6-8, 9+
merged <- merged %>%
  mutate(
    total_bucket = cut(
      num_courses,
      breaks = c(-Inf, 2, 5, 8, Inf),
      labels = c("0-2","3-5","6-8","9+"),
      include.lowest = TRUE, right = TRUE),
    total_bucket = factor(total_bucket, levels = c("0-2","3-5","6-8","9+"))
  )
```

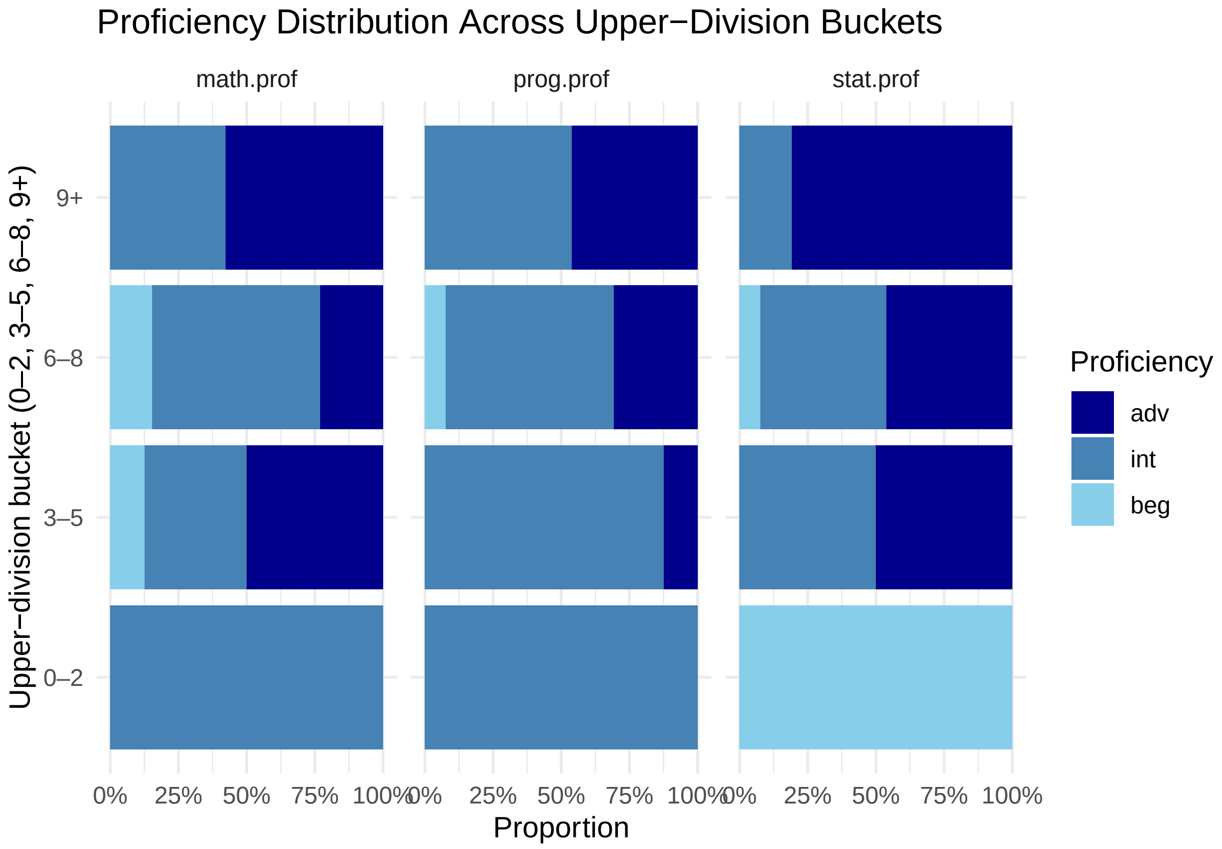
```r
# Sort prof by (beg -> int -> adv
merged <- merged %>%
  mutate(
    stat.prof = factor(stat.prof, levels = c("beg","int","adv")),
    prog.prof = factor(prog.prof, levels = c("beg","int","adv")),
    math.prof = factor(math.prof, levels = c("beg","int","adv"))
  )


# Heatmap of counts and percentages across the two created bucketings
rel_df <- merged %>%
  count(total_bucket, updv_bucket, name = "n") %>%
  group_by(total_bucket) %>%
  mutate(pct_within_total = n / sum(n)) %>%
  ungroup()

ggplot(rel_df, aes(x = total_bucket, y = updv_bucket, fill = n)) +
  geom_tile() +
  geom_text(aes(label = paste0(n, "\n", scales::percent(pct_within_total, accuracy = 1))), size = 3) +
  scale_fill_continuous(name = "Students") +
  labs(title = "Relationship: Total-Course Buckets vs Upper-Division Buckets",
       x = "Total-course bucket (from summed course flags)",
       y = "Upper-division bucket (from updv.num)") +
  theme_minimal()
```



```r
# Proficiency by our summed total courses buckets (beg -> adv)
prof_long <- merged %>%
  pivot_longer(c(stat.prof, prog.prof, math.prof),
               names_to = "Field", values_to = "Proficiency") %>%
  mutate(Proficiency = factor(Proficiency, levels = c("adv","int","beg")))

ggplot(prof_long, aes(x = total_bucket, fill = Proficiency)) +
  geom_bar(position = "fill") +
  facet_wrap(~ Field) +
  scale_fill_manual(values = c("beg" = "skyblue", "int" = "steelblue", "adv" = "darkblue")) +
  scale_y_continuous(labels = scales::percent) +
  coord_flip() +
  labs(title = "Proficiency Distribution Across Total-Course Buckets",
       x = "Total-course bucket (0-2, 3-5, 6-8, 9+)", y = "Proportion") +
  theme_minimal()
```



```r
# Proficiency by UPPER-DIVISION buckets (beg -> adv)
prof_long_ud <- merged %>%
  pivot_longer(c(stat.prof, prog.prof, math.prof),
               names_to = "Field", values_to = "Proficiency") %>%
  mutate(Proficiency = factor(Proficiency, levels = c("adv","int","beg")))  # beg at bottom

ggplot(prof_long_ud, aes(x = updv_bucket, fill = Proficiency)) +
  geom_bar(position = "fill") +
  facet_wrap(~ Field) +
  scale_fill_manual(values = c("beg" = "skyblue", "int" = "steelblue", "adv" = "darkblue")) +
  scale_y_continuous(labels = scales::percent) +
  coord_flip() +
  labs(title = "Proficiency Distribution Across Upper-Division Buckets",
       x = "Upper-division bucket (0-2, 3-5, 6-8, 9+)",
       y = "Proportion") +
  theme_minimal()
```



## NOTE: total courses is a summation of the specific courses which one may have checked off on the intake form.

Plot 1

In this plot rows are the upper-division (UD) buckets from updv.num (0–2, 3–5, 6–8, 9+), and the columns are created buckets of total courses checked off (0–2, 3–5, 6–8, 9+). What we can observe from this plot is the mass is strongly diagonal, the higher total buckets pair well with higher UD buckets, with large within-column shares in the matching. So we can conclude that the total and UD measures are consistent both very consistent measures. Thus, in the plots and analysis to come we see that we can use either bucket, and it suggests UD captures depth rather than just quantity, which can also be concluded from the data collection where updv.num can account for more math and stats and CS classes than the totals which were limited to only what we were able to check off.

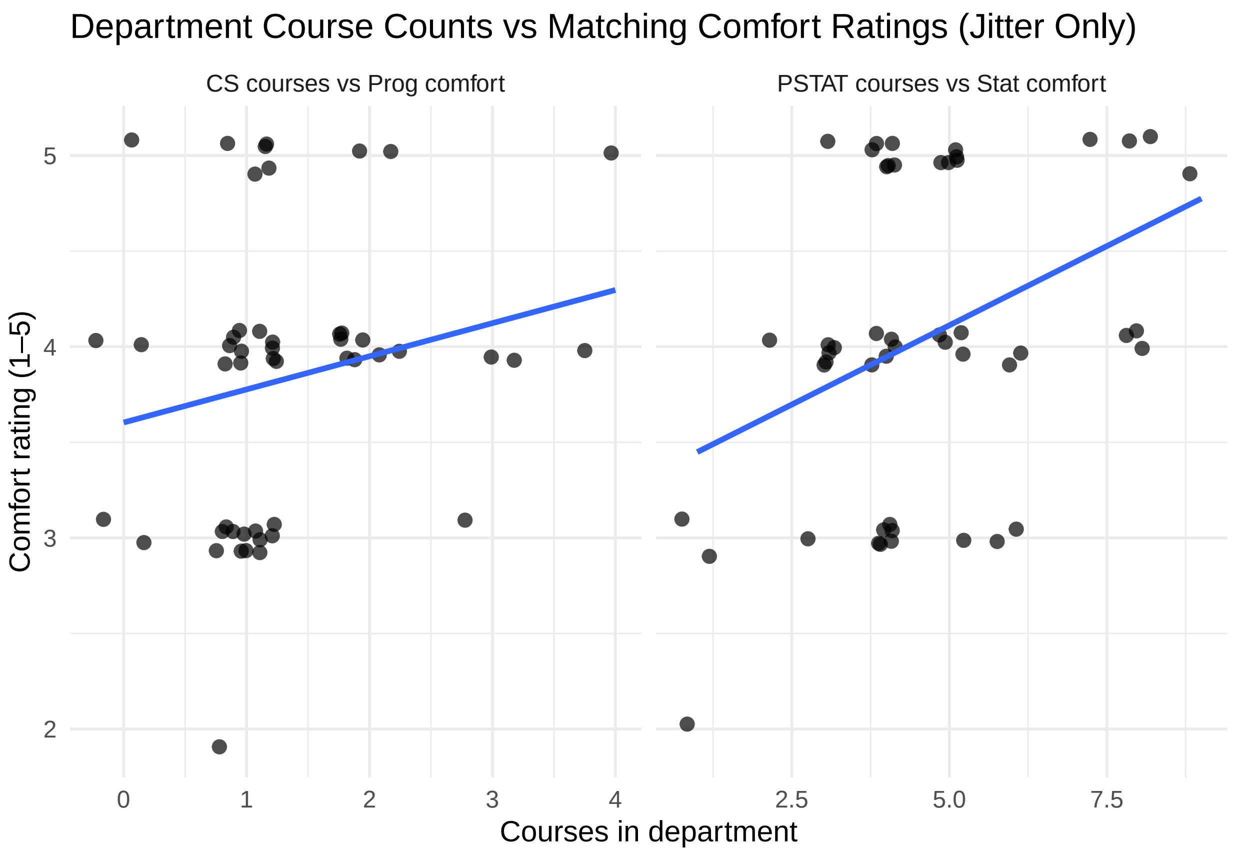
Plot 2

In plot 2 for the Stat/Prog/Math, each total-course bucket is split into beg/int/adv in relation to proficiency which were marked on the intake forms. In the plot we can see that as total courses increase, beg shrinks and adv grows in all three fields, this gradient is most apparent in Statistics, moderate in Math, and weaker in Programming (where many students are already “adv/int” at the low totals). This plot validates the idea that the more coursework which is done the more likely someone is to believe that they have a higher proficiency in the respective fields.

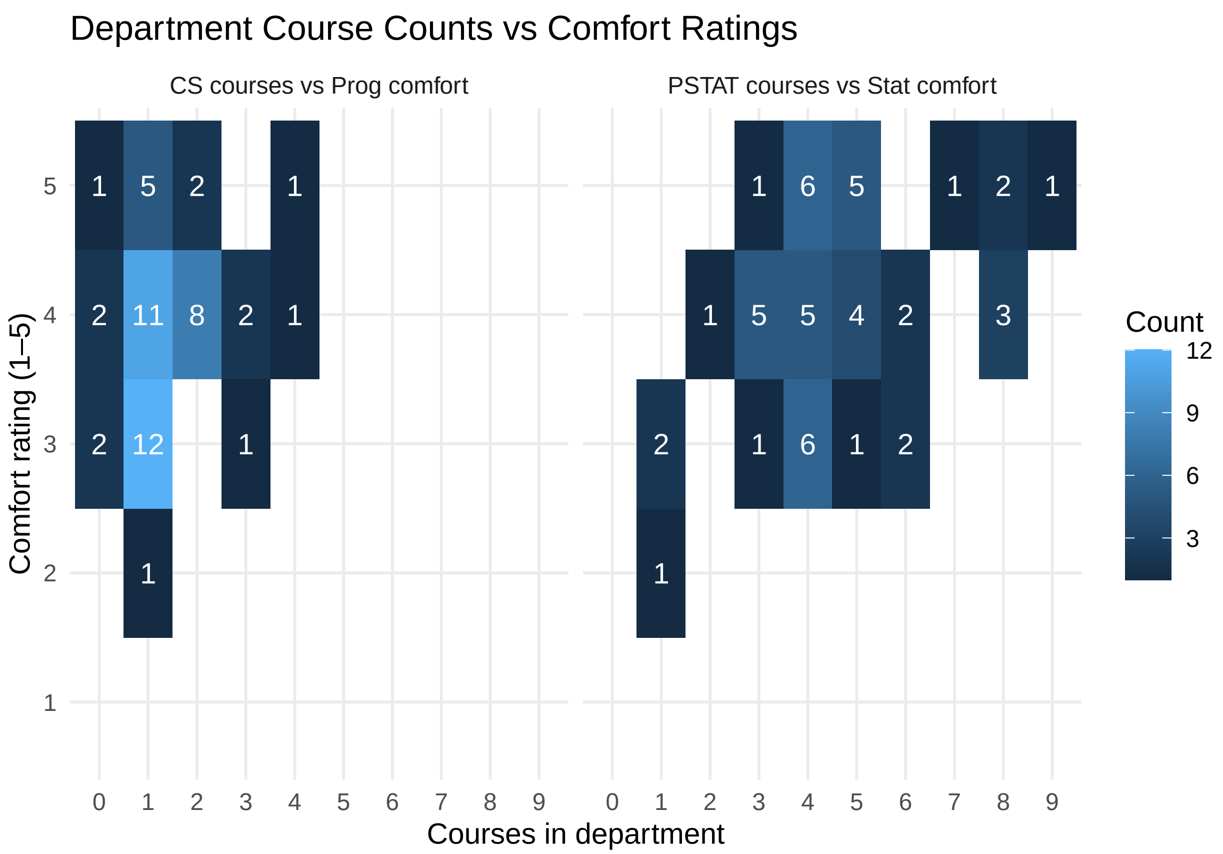
Plot 3

Plot 3 follows the exact same ideas and principles as plot 2; however, this is using the number of upper divisions one has taken instead of the total courses which could be checked. There is a different result found though. While we observe similar results to plot 2 they are far less dramatic and this could be for numerous reasons, but it is more likely that with the removal of the inclusion of specific class which could help in proficiency that one may put a lower score. Many factors could cause for this difference. But, we can conclude that upper division counts show better indications of depth in the data set.

```
# System to flag the courses we want
course_cols <- names(merged) %>%
  stringr::str_subset("^(PSTAT|CS)\\d+$")

# Convert the values to 0/1
course_mat_bin <- merged %>%
  transmute(across(all_of(course_cols), ~ {
    v <- suppressWarnings(as.numeric(.x))
    v <- dplyr::coalesce(v, 0)
    as.integer(v > 0)
  }))

# Summing the totals based on the course department title
course_cols_pstat <- course_cols %>%
  stringr::str_subset("^PSTAT\\d+$")
course_cols_cs    <- course_cols %>%
  stringr::str_subset("^CS\\d+$")


dept_plots_df <- dplyr::bind_rows(
  merged %>%
    transmute(x = pstat_courses, y = `stat.comf`, lab = "PSTAT courses vs Stat comfort"),
  merged %>%
    transmute(x = cs_courses, y = `prog.comf`, lab = "CS courses vs Prog comfort")
)

merged <- merged %>%
  mutate(
    num_courses = rowSums(course_mat_bin[course_cols], na.rm = TRUE),
    pstat_courses = if (length(course_cols_pstat)) rowSums(course_mat_bin[course_cols_pstat], na.rm = T
    cs_courses = if (length(course_cols_cs)) rowSums(course_mat_bin[course_cols_cs], na.rm = TRUE) else


dept_plots_df <- dplyr::bind_rows(
  merged %>% transmute(x = pstat_courses, y = `stat.comf`, lab = "PSTAT courses vs Stat comfort"),
  merged %>% transmute(x = cs_courses, y = `prog.comf`, lab = "CS courses vs Prog comfort")
)


ggplot(dept_plots_df, aes(x = x, y = y)) +
  geom_jitter(width = 0.25, height = 0.1, size = 2, alpha = 0.7) +
  facet_wrap(~ lab, scales = "free_x") +
  labs(title = "Department Course Counts vs Matching Comfort Ratings (Jitter Only)",
       x = "Courses in department", y = "Comfort rating (1-5)") +
  geom_smooth(method = "lm", formula = y ~ x, se = FALSE, size = 1) +
  theme_minimal()
```



```
dept_counts <- dept_plots_df %>%
  transmute(
    lab = as.character(lab),
    x   = as.integer(round(x)),
    y   = factor(as.integer(as.character(y)),
                 levels = 1:5)
  ) %>%
  filter(!is.na(x), !is.na(y)) %>%
  count(lab, x, y, name = "n")

# Heatmap with count labels version
ggplot(dept_counts, aes(x = factor(x), y = y, fill = n)) +
  geom_tile() +
  geom_text(aes(label = n), color = "white") +
  facet_wrap(~ lab, scales = "free_x") +
  scale_x_discrete(drop = FALSE) +
  scale_y_discrete(drop = FALSE) +
  labs(
    title = "Department Course Counts vs Comfort Ratings",
    x = "Courses in department",
    y = "Comfort rating (1-5)",
    fill = "Count"
  ) +
  theme_minimal()
```



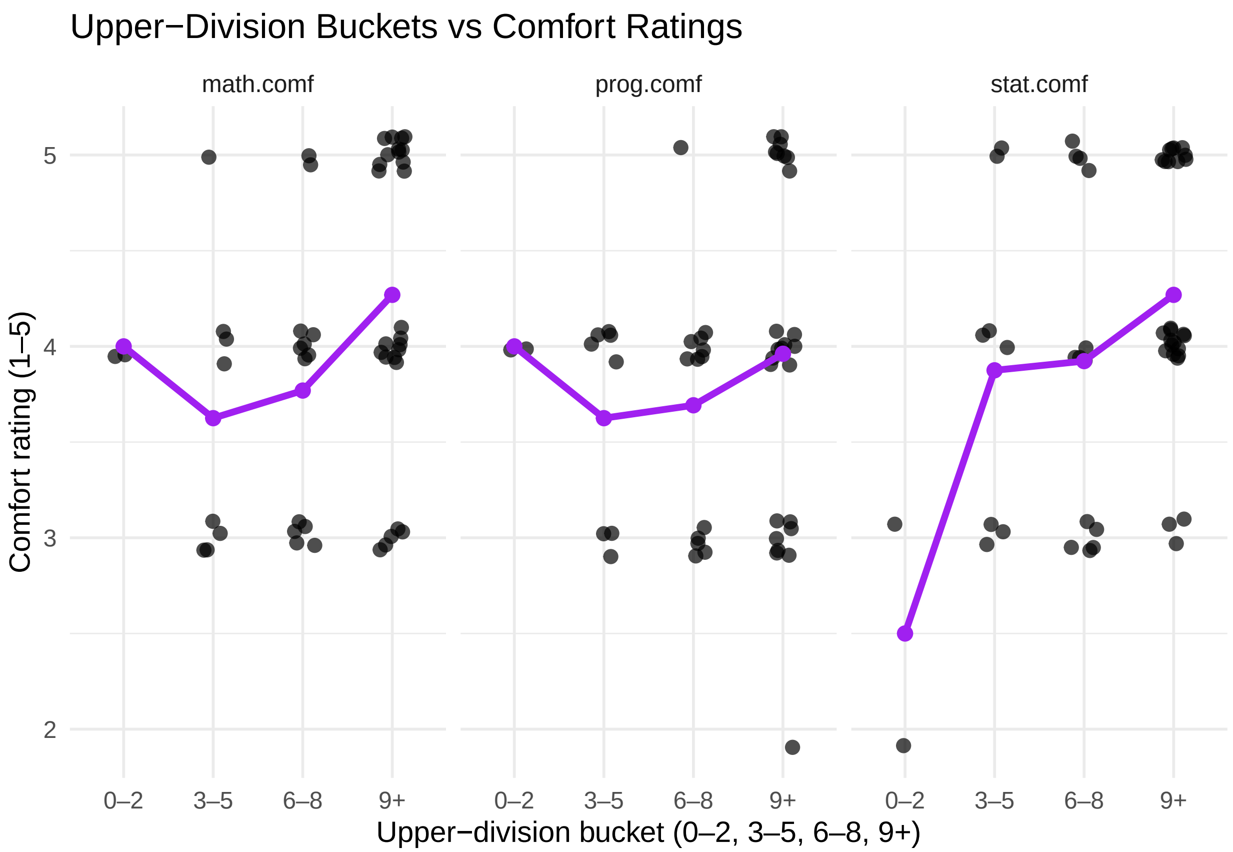
## Plots 4 and 5

In the above plots 4 and 5 we are measuring the same thing within both, however they just provided different visuals for us to analyze. Now my group has deduced that comfort rating is a better indicator of ability within respective fields. In these plots we additionally have not measured math because there is no inclusion of math department courses so we can not run this analysis on math comfort rating.

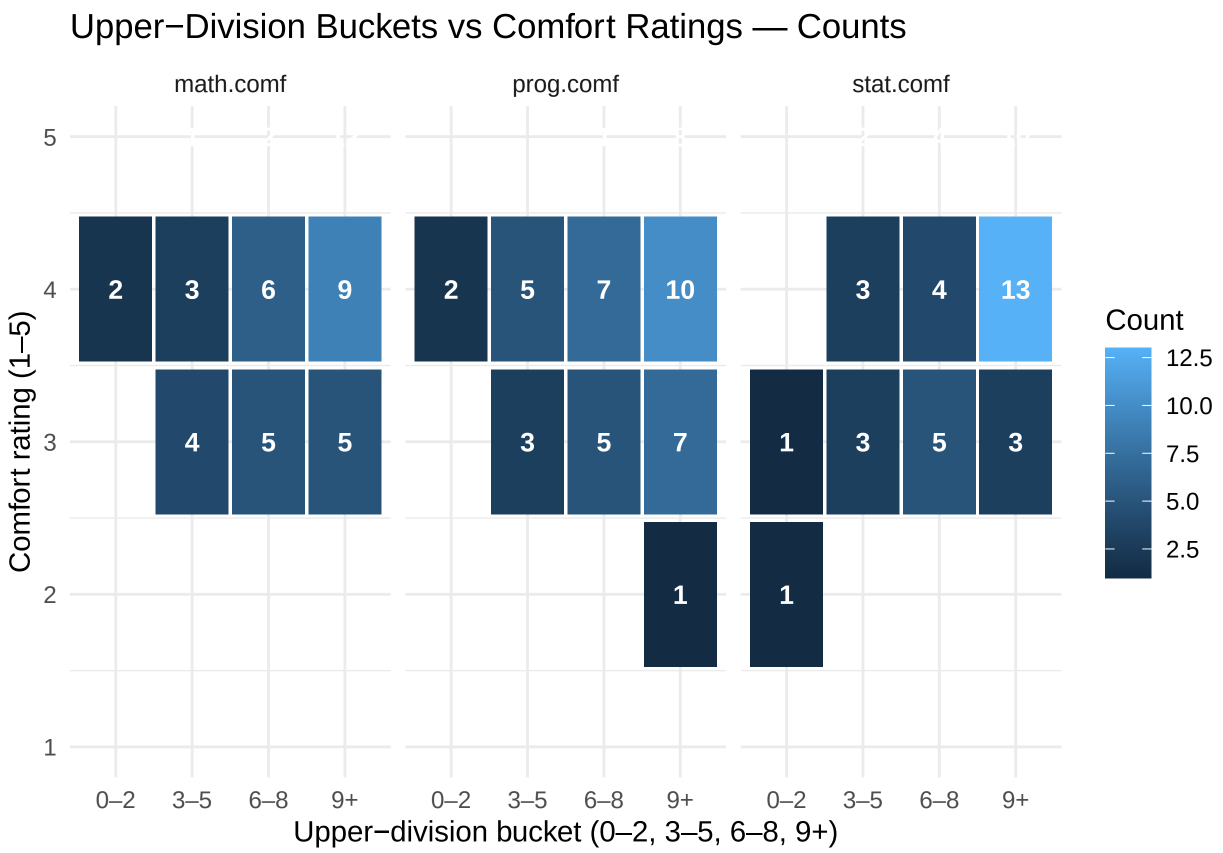
In these plots of courses which have been taken in in CS and PSTAT that were able to be checked off in the intake survey and plotted them against ones comfort rating in programming and statistics respectively. We observe that the plots model: student department-specific counts (x) vs matching comfort (y), faceted for CS→Prog and PSTAT→Stat. Additionally we observe that both slopes are positive. The PSTAT→Stat slope is steeper: more PSTAT courses is closely associated with higher Stat comfort. The CS→Prog slope is positive but flatter. Now from these plots we can conclude that more within-subject courses implies higher comfort is clearly true for Statistics and present but weaker for Programming, likely because programming comfort comes from PSTAT/R work, prior experience, and self-study—not just CS courses.

Specifically within plot 5 we can see in the comparison of CS courses with programming that there is a heavy concentration around x=1 with y of about 3–4. While in PSTAT courses relation with statistical comfort, the mass sits at x of about 4–6 with y=4–5, plus a thin tail at higher x with y=5.

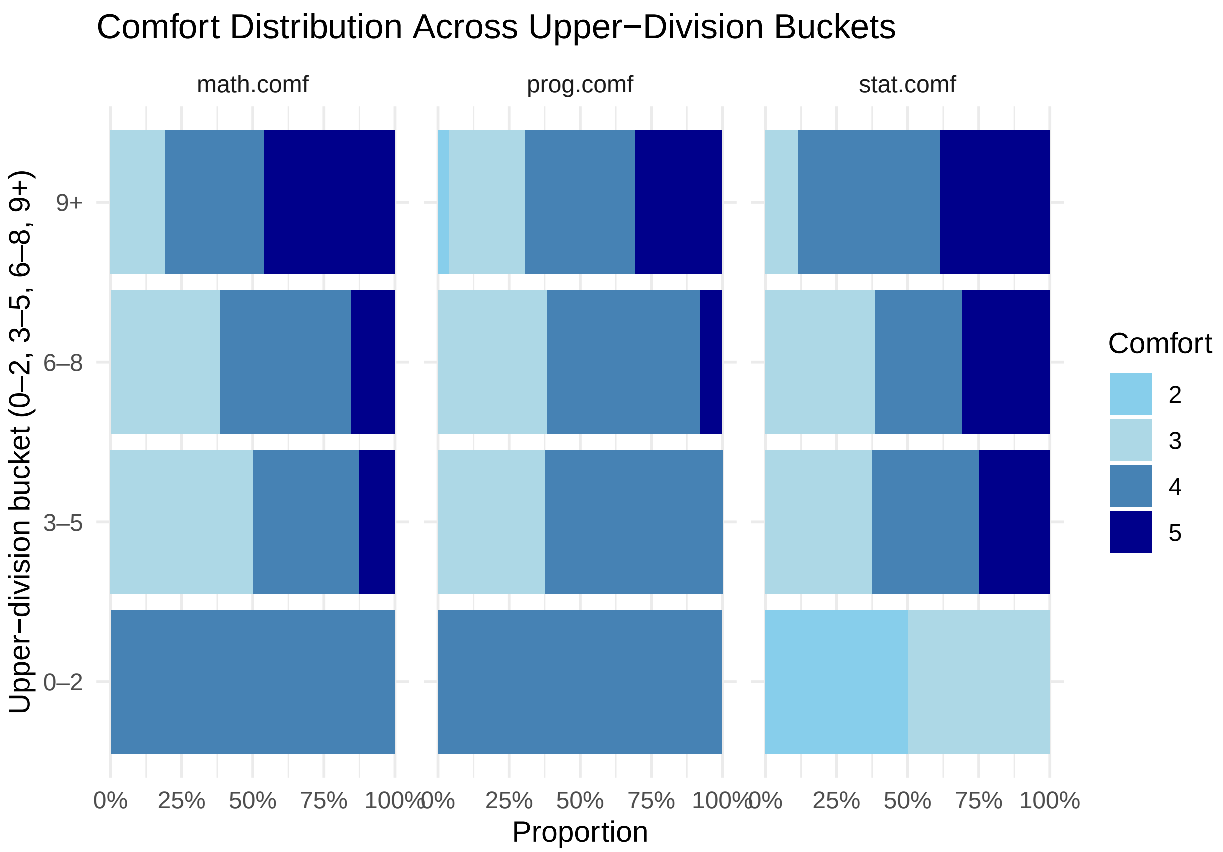
```r
merged <- merged %>%
  mutate(
    updv_bucket = case_when(
      str_detect(updv.num, "0\\s*-\\s*2") ~ "0-2",
      str_detect(updv.num, "3\\s*-\\s*5") ~ "3-5",
      str_detect(updv.num, "6\\s*-\\s*8") ~ "6-8",
      str_detect(updv.num, "9\\s*\\+")   ~ "9+",
      TRUE ~ NA_character_
    ),
    updv_bucket = factor(updv_bucket, levels = c("0-2","3-5","6-8","9+"))
  )

# Restructuring of comfort ratings
ratings_updv <- merged %>%
  pivot_longer(c(stat.comf, prog.comf, math.comf),
               names_to = "Field", values_to = "Rating")

# Plot of upper divisions bucket counts vs comfort ratings
ggplot(ratings_updv, aes(x = updv_bucket, y = Rating)) +
  geom_jitter(width = 0.15, height = 0.1, alpha = 0.7, size = 2) +
  stat_summary(fun = mean, geom = "line", group = 1, size = 1.2, color = "purple") +
  stat_summary(fun = mean, geom = "point", size = 2.2, color = "purple") +
  facet_wrap(~ Field, nrow = 1) +
  labs(title = "Upper-Division Buckets vs Comfort Ratings",
       x = "Upper-division bucket (0-2, 3-5, 6-8, 9+)", y = "Comfort rating (1-5)") +
  theme_minimal()
```



```r
hm_counts <- ratings_updv %>%
  count(Field, updv_bucket, Rating, name = "n")

ggplot(hm_counts, aes(x = updv_bucket, y = Rating, fill = n)) +
  geom_tile(width = 0.95, height = 0.95) +
  geom_text(aes(label = n), color = "white", fontface = "bold", size = 3.6) +
  facet_wrap(~ Field, nrow = 1) +
  scale_y_continuous(breaks = 1:5, limits = c(1, 5)) +
  labs(title = "Upper-Division Buckets vs Comfort Ratings - Counts",
    x = "Upper-division bucket (0-2, 3-5, 6-8, 9+)",
    y = "Comfort rating (1-5)", fill = "Count") +
  theme_minimal()
```



```
# Long format
ratings_ud <- merged %>%
  pivot_longer(c(stat.comf, prog.comf, math.comf),
               names_to = "Field", values_to = "Comfort") %>%
  mutate(
    Comfort = parse_number(as.character(Comfort)),
    Comfort = factor(Comfort, levels = 1:5, ordered = TRUE)
  )

# Stacked proportional bars of comfort
ggplot(ratings_ud, aes(x = updv_bucket, fill = Comfort)) +
  geom_bar(position = position_fill(reverse = TRUE)) +
  facet_wrap(~ Field) +
  scale_fill_manual(
    values = c("5" = "darkblue", "4" = "steelblue", "3" = "lightblue", "2" = "skyblue", "1" = "blue"),
    name = "Comfort",
    breaks = c("1","2","3","4","5")
  ) +
  scale_y_continuous(labels = scales::percent) +
  coord_flip() +
  labs(
    title = "Comfort Distribution Across Upper-Division Buckets",
    x = "Upper-division bucket (0-2, 3-5, 6-8, 9+)",
    y = "Proportion"
  ) +
  theme_minimal()
```



## Plots 6 and 7
Much like plots 4 and 5 these two plots are of the same information however are different visuals. In plot 6 we see UD bucket on x, comfort on y, with a per-bucket mean line. We observe from this plot that the mean line rises with UD buckets for all fields. Largest for stats, moderate for math, and lowest for programming. These findings do not suprise me at all because this means having learned from more classes leads to an increase of comfort in such field. Thus we can conclude that comfort increases with depth of coursework, reinforcing the department-specific results above.

Now for plot 7 we see the counts at each UD bucket × comfort rating cell, per field on y. We see for Statistics, high counts concentrate at (6–8, 4–5) and (9+, 4–5). Programming shows many 4's (and some 5's) even at 0–2/3–5. Math shows a slow increase toward higher comfort in higher UD buckets. Thus, we reinforce our findings from plot 6.

Plot 8

In plot 8 we see that for each field, UD buckets are structured into comfort levels 1–5. We observe a shared shift upward with UD buckets i.e. lower comfort shrinks, higher comfort grows. The shift is most pronounced for Stat, then Math and lastly for Programming it still trends upward but maintains a large high-comfort share even at low UD. Thus, we have found that the distributional view supports the same conclusions we have made so far that the depth of UD course buckets tracks comfort best.

Results

A re proposed question. How does coursework relate to comfort in Statistics/Programming/Math?

We find quite trivially that using the upper division buckets for this comparison is far more insightful than the use of summation of classes which could be checked off on the intake form. Additionally, our group has also found that comfort rating is a better indicator of ones abilities in a field, which is explained to be because proficiency measures individuals against the entire school including non-STEM majors, whereas, we find that comfort rating is one's own assessment of themselves based on what they have done so far. Thus, the question above is able to be answered.

Now what I have found is that the within-subject exposure does indeed affect someones answers. For Statistics, more PSTAT courses and higher upper-division exposure show a strong correlation with higher comfort. Programming has a correlation between CS courses and programming comfort which is positive but weaker because many students report high comfort even with few CS courses—consistent with programming practice gained PSTAT courses in R, prior experience, or self-study. We find that math shows a much more moderate gradient. Comfort increases with both total courses and UD, but less than Statistics. However, we must note this is far more challenging to measure because no one had the ability to check off math courses on the intake form. The heat maps and stacks show that UD buckets align with higher comfort as well as, and typically better than, total courses.