

Data Challenge 04

Jake Moore

Objective

In dc_03 you created five base visuals. In this challenge, you will refine them using visualization techniques for clarity and storytelling. You will:

1. add titles, subtitles, axis labels, legend titles, and captions
2. order categories and add direct labels where helpful
3. add reference cues (for example, mean lines or $y = x$) and short on-plot annotations
4. remove legends when redundant and ensure readable axes

Recreate the tidy data objects (Steps 1–6), then create styled versions of the dc_03 visuals (Steps 7–11). Provide a short interpretation note (1–2 sentences) for each.

Data: `us_avg_tuition.xlsx` (same folder as this Rmd).

Special Notes (MUST READ)

- For Steps 1-6, you can replace them all with your Steps 1-6 from your dc_03 if you got 100%, otherwise reference the dc03 “solution” walkthrough video/ contact the instructor to correct any flaws in your submission.
- Be careful pasting any code from your previous Steps 7-11 as the plot object naming is different in dc04. Better to use your previous code as a reference instead.
- Remember you can un-comment code using CMD+shift+c (MacBook) or control+shift+c (Windows)
- This is probably the first week where a few methods appear that we did not explicitly spell out in lecture. That is intentional. Use this challenge to practice two key skills: applying what you know and reading function documentation. I will often name the exact function, but you will choose the arguments. This mirrors real visualization work: there are many valid aesthetic combinations, and you will not memorize them all. Start with what you know, check `?function_name` when stuck, and try small, testable edits. This assignment is the majority of the module time, so pace yourself and iterate. If you hit a wall after an honest try, reach out. I am here to help.

Evaluation

- Steps 1–6 (data prep): 10 points total (for the proper set-up)
- Steps 7–11 (styled visuals): 90 points total
- File naming and knitting each apply a 10% penalty if violated

1) Load the tuition data (2 pts)

```
us_avg_tuition_raw <- read_excel("us_avg_tuition.xlsx")
```

2) Reshape to tidy long format (2 pts)

```
state_tuition_long <- us_avg_tuition_raw %>%  
  rename_with(tolower) %>%  
  pivot_longer(cols = -state, names_to = "school_year", values_to = "avg_cost")
```

3) Factor and missingness checks (2 pts)

```
state_tuition_long <- state_tuition_long %>%  
  mutate(  
    state = factor(state),  
    school_year = as.character(school_year),  
    avg_cost = as.numeric(avg_cost)  
  )
```

4) Build a key: state -> region (2 pts)

```
state_regions <- tibble(  
  state = state.name,  
  region = state.region  
)
```

5) Join tuition with regions (1 pt)

```
tuition_with_region <- state_tuition_long %>%  
  left_join(state_regions, by = "state")
```

6) Regional averages for 2015-16 (1 pt)

```

region_cost_2015 <- tuition_with_region %>%
  filter(school_year == "2015-16") %>%
  group_by(region) %>%
  summarise(avg_cost = mean(avg_cost), .groups = "drop") %>%
  arrange(desc(avg_cost))

```

7) National trend over time — Visual 1 (18 pts)

Builds on dc_03 Visual 1. Using the `national_by_year`, create `p1_national_line_styled` with the following elements:

1. (4 pt) line **and** point markers
2. (4 pt) meaningful title and subtitle
3. (3 pt) meaningful x and y labels with applicable units
4. (3 pt) caption citing the data file as source
5. (2 pt) x labels rotated 45° and horizontal alignment (hjust) of 1 for readability

special note: We convert `school_year` to a factor to preserve the written labels (e.g., “2004-05”) and to fix the axis in chronological order. A factor is categorical, not numeric, so `geom_line()` won’t connect points by default—it treats each x level as its own group. We can add `group = 1` to tell ggplot all points belong to one series, so the line connects across years.

```

national_by_year <- state_tuition_long %>%
  mutate(school_year = factor(school_year, levels = unique(school_year))) %>%
  group_by(school_year) %>%
  summarise(avg_cost = mean(avg_cost), .groups = "drop") %>%
  arrange(school_year)

```

```

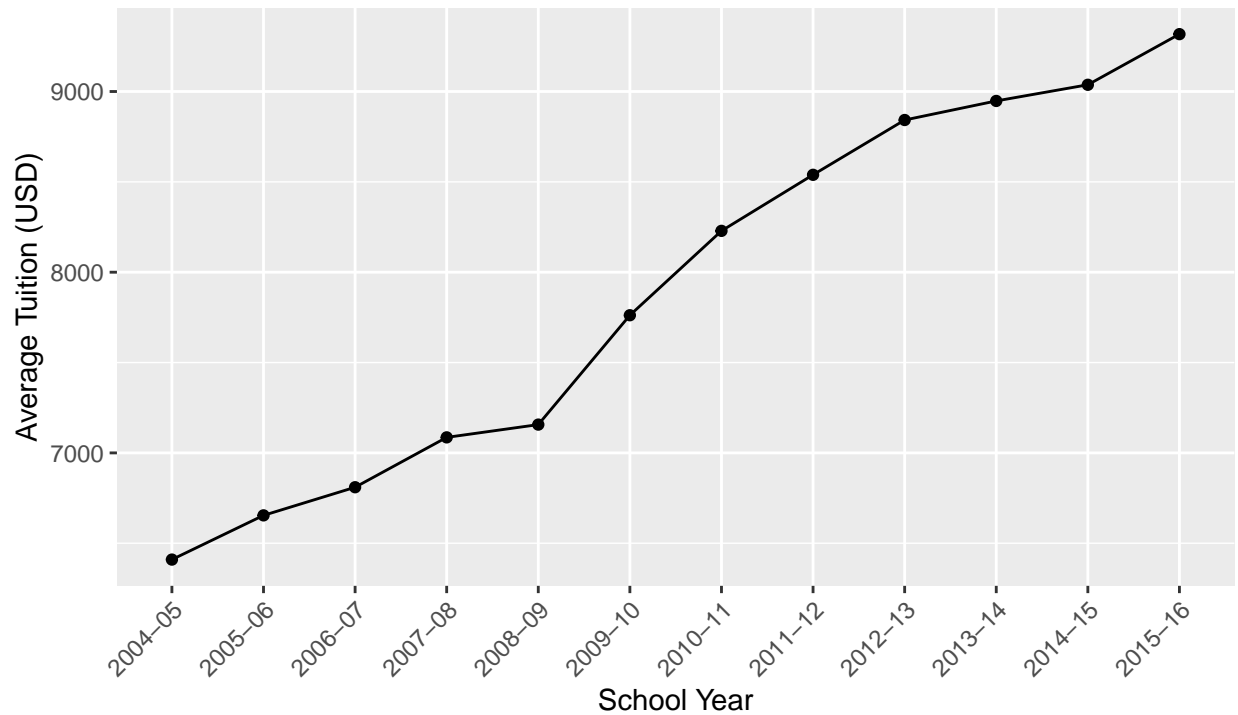
p1_national_line_styled <- ggplot(national_by_year, aes(x = school_year, y = avg_cost , group = 1)) +
  geom_line() +
  geom_point() +
  labs(
    title="Average U.S. Tuition by Year",
    subtitle="There is a positive correlation.",
    x="School Year",
    y="Average Tuition (USD)",
    caption="Source = us_avg_tuition.xlsx"
  ) +
  theme(axis.text.x = element_text(angle=45,hjust=1))

p1_national_line_styled

```

Average U.S. Tuition by Year

There is a positive correlation.



Source = us_avg_tuition.xlsx

Short interpretation (1-2 sentences) → viz1_note_styled. (2 pt)

```
viz1_note_styled <- 'The visualization above proves that average U.S. Tuition increased relatively cons'
```

8) Regional comparison (2015-16) — Visual 2 (18 pts)

Builds on dc_03 Visual 2. Create p2_region_bars_styled from region_cost_2015 with the following elements:

1. ordered bar plot of **region** vs **avg_cost**
2. (4 pt) bars sorted by **avg_cost** ascending
3. (3 pt) bars color filled based on region for enhanced readability
4. (3 pt) direct value labels above each bar (use `round(avg_cost, 0)`, `vjust = -0.3`, `size = 3`)
5. (2 pt) meaningful title and subtitle
6. (1 pt) meaningful x and y labels with applicable units
7. (1 pt) caption citing the data file as source
8. (2 pt) hide redundant legend

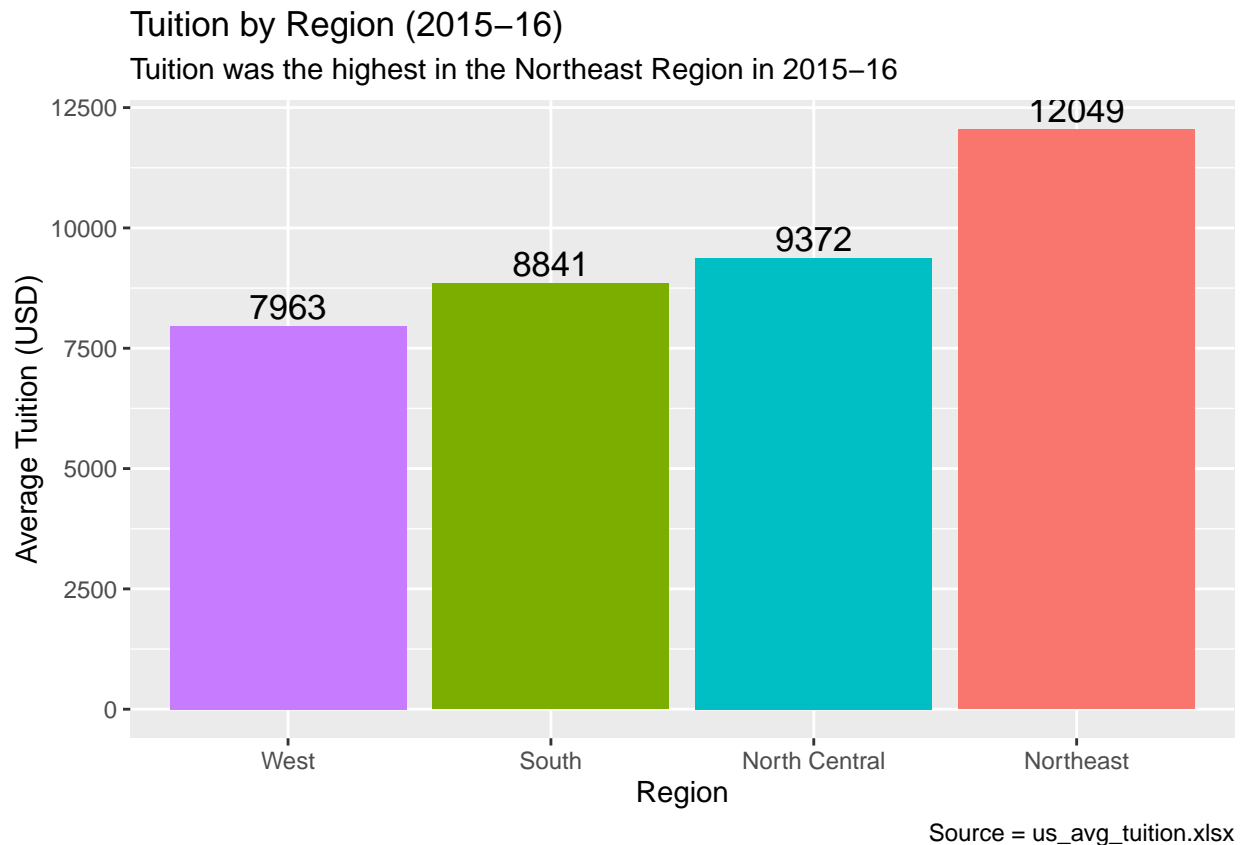
```
p2_region_bars_styled <- ggplot(region_cost_2015, aes(x = fct_reorder(region, avg_cost), y = avg_cost, 
  geom_col() + 
  geom_text(aes(label = round(avg_cost, 0), vjust = -0.3, size = 3)) + 
  labs(
```

```

title="Tuition by Region (2015-16)",
subtitle="Tuition was the highest in the Northeast Region in 2015-16",
x="Region",
y="Average Tuition (USD)",
caption="Source = us_avg_tuition.xlsx"
) +
theme(legend.position= "none")

```

p2_region_bars_styled



Short interpretation (1-2 sentences) → viz2_note_styled. (2 pt)

```
viz2_note_styled <- 'This visualization uses geomcol() to communicate the different regional differences'
```

9) State distribution (2015-16) — Visual 3 (18 pts)

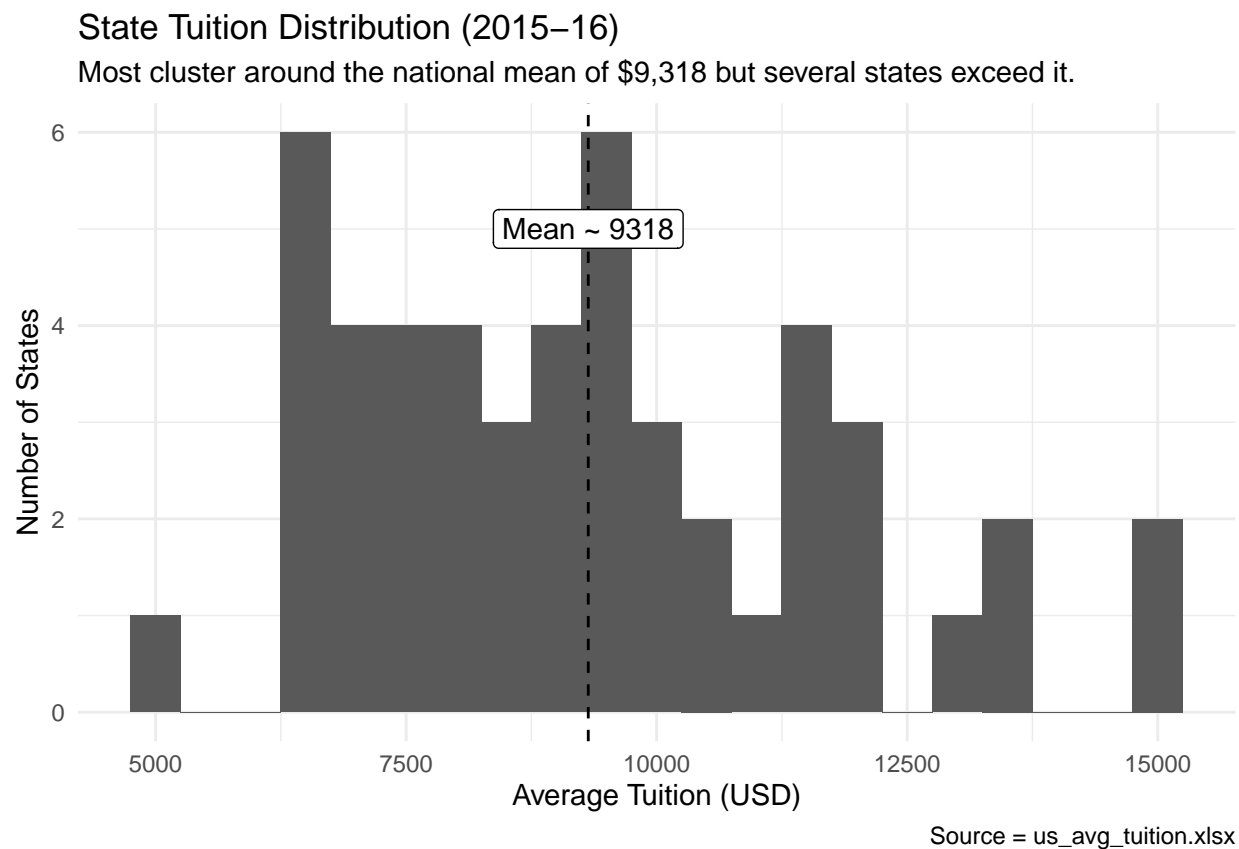
Builds on dc_03 Visual 3. Create p3_state_hist_2015_styled with the following elements:

1. histogram of avg_cost frequencies from 2015-2016
2. (4 pt) a sensible bin width of 500
3. (4 pt) a dashed vertical line at the mean (created before plotting and stored as mean_2015)
4. (3 pt) a small on-plot label annotation at the mean and y=5 "Mean ~"

5. (4 pt) meaningful title, subtitle, axis labels, and caption
6. (1 pt) a minimalist aesthetic using `theme_minimal()`

```
state_2015 <- tuition_with_region %>% filter(school_year == "2015-16")
mean_2015 <- mean(state_2015$avg_cost)
p3_state_hist_2015_styled <- ggplot(state_2015, aes(x = avg_cost)) +
  geom_histogram(binwidth=500) +
  geom_vline(xintercept = mean_2015, linetype = "dashed") +
  annotate("label", x = mean_2015, y = 5, label = paste0("Mean ~ ", round(mean_2015, 0))) +
  labs(
    title="State Tuition Distribution (2015-16)",
    subtitle="Most cluster around the national mean of $9,318 but several states exceed it.",
    x="Average Tuition (USD)",
    y="Number of States",
    caption="Source = us_avg_tuition.xlsx") +
  theme_minimal()

p3_state_hist_2015_styled
```



Short interpretation (1–2 sentences) → viz3_note_styled. (2 pt)

```
viz3_note_styled <- 'The histogram visualized above shows that most states fall around the $9,318 average'
```

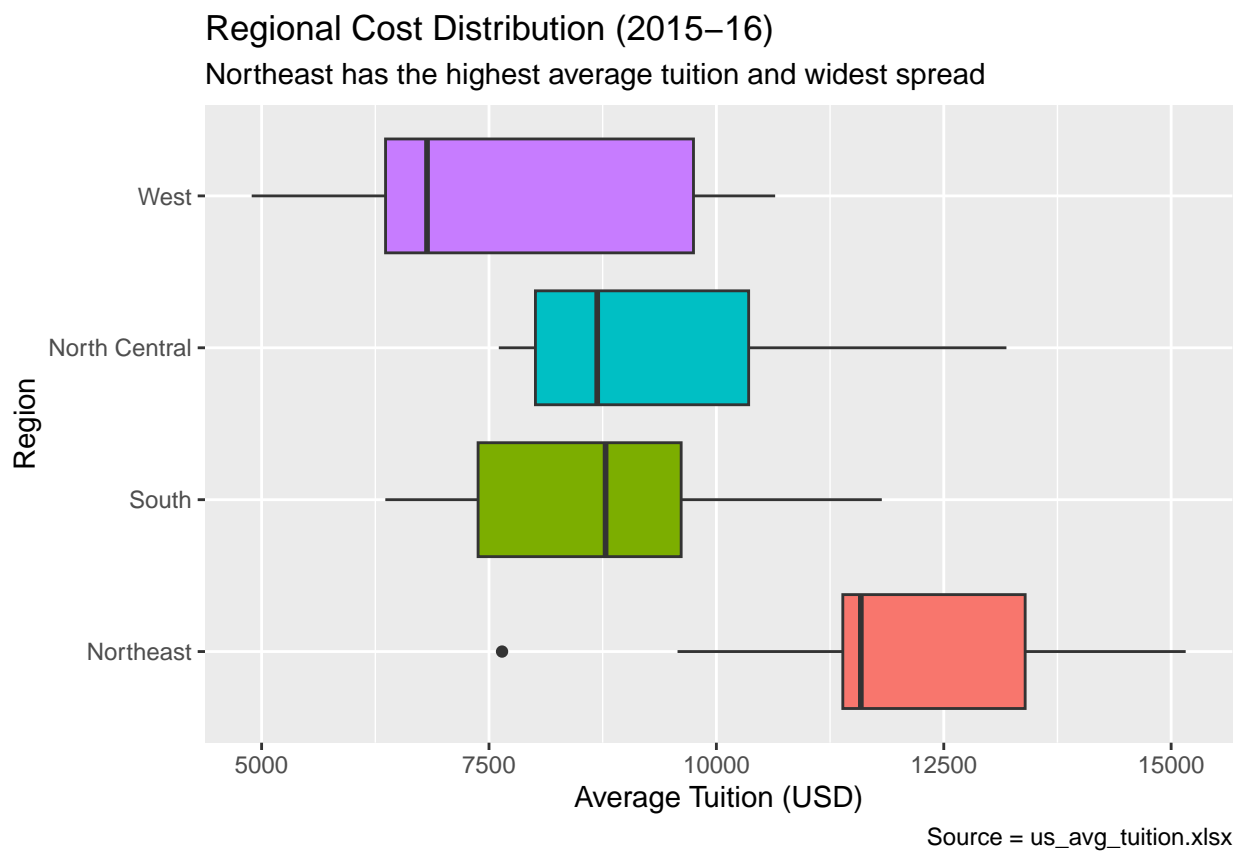
10) Regional spread (2015-16) — Visual 4 (18 pts)

Builds on dc_03 Visual 4. Create `p4_region_box_2015_styled` with the following elements:

- (4 pt) boxplot of `avg_cost` by `region`
- (3 pt) y-axis of `region` for readability
- (3 pt) filled box color based on the `region`
- (4 pt) meaningful title, subtitle, axis labels, and caption
- (2 pt) hidden legend if redundant

```
p4_region_box_2015_styled <- ggplot(state_2015,aes(x=avg_cost,y=region,fill=region)) +  
  geom_boxplot() +  
  labs(  
    title="Regional Cost Distribution (2015-16)",  
    subtitle="Northeast has the highest average tuition and widest spread",  
    x="Average Tuition (USD)",  
    y="Region",  
    caption="Source = us_avg_tuition.xlsx") +  
  theme(legend.position= "none")
```

`p4_region_box_2015_styled`



Short interpretation (1–2 sentences) → `viz4_note_styled`. (2 pt)

```
viz4_note_styled <- 'The boxplot visualized above shows that the Northeast region has the highest average'
```

11) Early vs late comparison — Visual 5 (18 pts)

Builds on dc_03 Visual 5 in three parts:

- A) Recreate `two_years_wide` (2004–05 and 2015–16)
- B) Add a new column `delta = cost_2015 - cost_2004` and find the top 3 delta increases using `slice_max()`
- C) Create `p5_scatter_2004_vs_2015_styled` with the following elements:
 - 1. (3 pt) scatter of `cost_2004` vs `cost_2015` (`alpha = 0.7`)
 - 2. (3 pt) reference line `y = x` (dotted line style)
 - 3. (3 pt) regression line of `y~x` using the `"lm"` method
 - 4. (3 pt) labels for the top 3 increases (based on `delta`) using `geom_label` `size = 3` and `vjust = -0.3`
 - 5. (2 pt) meaningful title, subtitle, axis labels, and caption
 - 6. (2 pt) a minimalist aesthetic using `theme_minimal()`

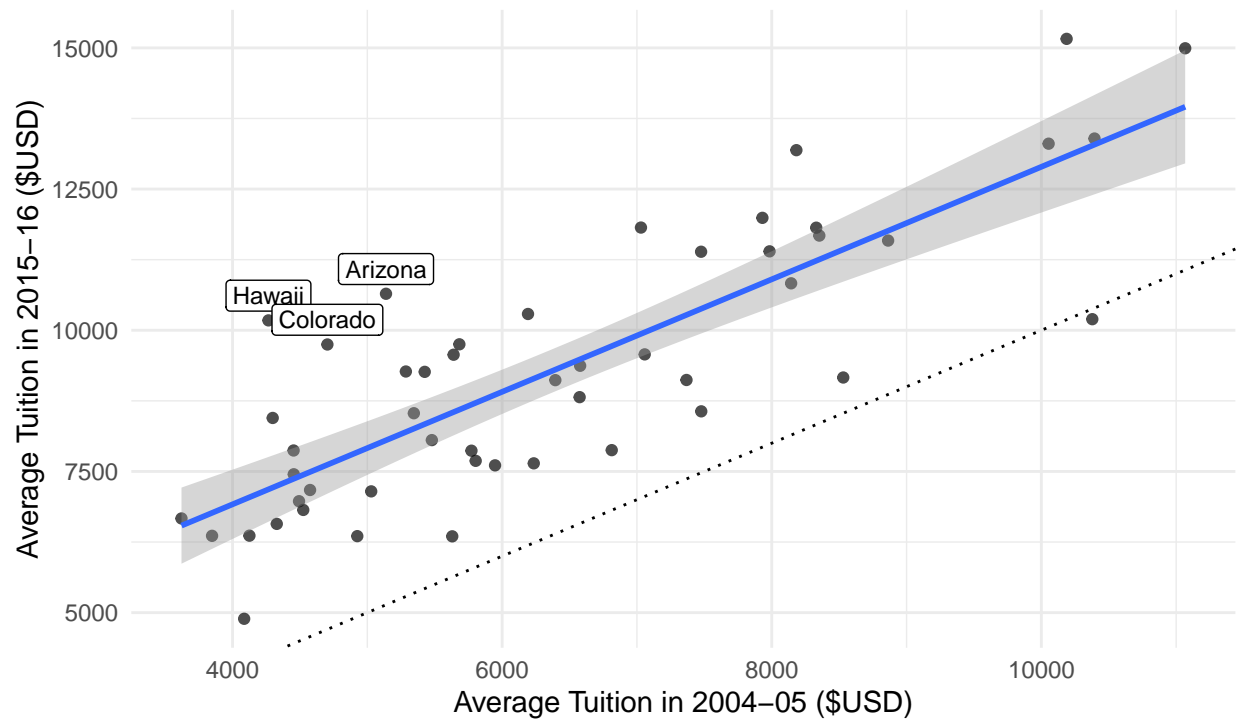
```
two_years_wide <- tuition_with_region %>%  
  filter(school_year %in% c("2004-05", "2015-16")) %>%  
  select(state, school_year, avg_cost) %>%  
  pivot_wider(names_from = school_year, values_from = avg_cost) %>%  
  rename(cost_2004 = `2004-05`,  
         cost_2015 = `2015-16`)
```

```
two_years_wide <- two_years_wide %>% mutate(delta = cost_2015 - cost_2004)  
top3 <- two_years_wide %>% slice_max(delta, n=3)
```

```
p5_scatter_2004_vs_2015_styled <- ggplot(two_years_wide, aes(x = cost_2004, y = cost_2015)) +  
  geom_point(alpha=0.7) +  
  geom_abline(slope=1, intercept=0, linetype = "dotted") +  
  geom_smooth(method="lm") +  
  geom_label(data=top3, aes(label = as.character(state)), size=3, vjust=-0.3) +  
  labs(  
    title="Tuition: 2004-05 vs. 2015-16 ",  
    subtitle="Hawaii, Arizona, and Colorado are top 3 tuition increases.",  
    x="Average Tuition in 2004-05 ($USD)",  
    y="Average Tuition in 2015-16 ($USD)",  
    caption="Source = us_avg_tuition.xlsx") +  
  theme_minimal()  
  
p5_scatter_2004_vs_2015_styled
```

Tuition: 2004–05 vs. 2015–16

Hawaii, Arizona, and Colorado are top 3 tuition increases.



Source = us_avg_tuition.xlsx

Short interpretation (1–2 sentences) → viz5_note_styled. (2 pt)

```
viz5_note_styled <- 'The scatterplot above visualizes a positive relationship between average tuition in
```