Data Challenge 03

Jake Moore

Objective

Tidy (clean and transform) tuition data by creating a clean key table (state → region), and joining by key (combine tables by a key). Create five basic visuals and write short interpretation notes (1–2 sentences each). Remember to un-comment any code you want graded and fill in all blanks.

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

---

1) Load the tuition data

Goal: create new dataframe object `us_avg_tuition_raw` by loading in `us_avg_tuition.xlsx`. Success check: `nrow(us_avg_tuition_raw) == 50`.

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

---

2) Reshape to tidy long format

Goal: First, ensure all column names are lowercase. Second, create a new dataframe called `state_tuition_long` that has columns `state`, `school_year`, `avg_cost`. Hint: rename `State` → `state`, pivot year columns to long. Success check: `nrow(state_tuition_long) == 600`.

```
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

Goal: Ensure then confirm `state` is factor and no NA in `avg_cost`. You may store helper objects for checks. Success check: `length(levels(state_tuition_long$state)) == 50` and `sum(is.na(state_tuition_long$avg_cost)) == 0`.

```r
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

Goal: create dataframe `state_regions` with `state`, `region` (Northeast, South, Midwest, West) using built-in vectors `state.name` and `state.region` to Hint: take a look at what `state.name` and `state.region` look like in your console. Success check: `nrow(state_regions) == 50` and `n_distinct(state_regions$region) == 4`.

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

---

5) Join tuition with regions

Goal: create `tuition_with_region` by getting `state_tuition_long` left join `state_regions`. Success check: `nrow(tuition_with_region) == 600` and no NA in `tuition_with_region$region`.

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

---

6) Regional averages for 2015–16

Goal: create `region_cost_2015` with mean `avg_cost` by `region` for `school_year == "2015-16"`. Success check: `nrow(region_cost_2015) == 4`.

```r
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))
```

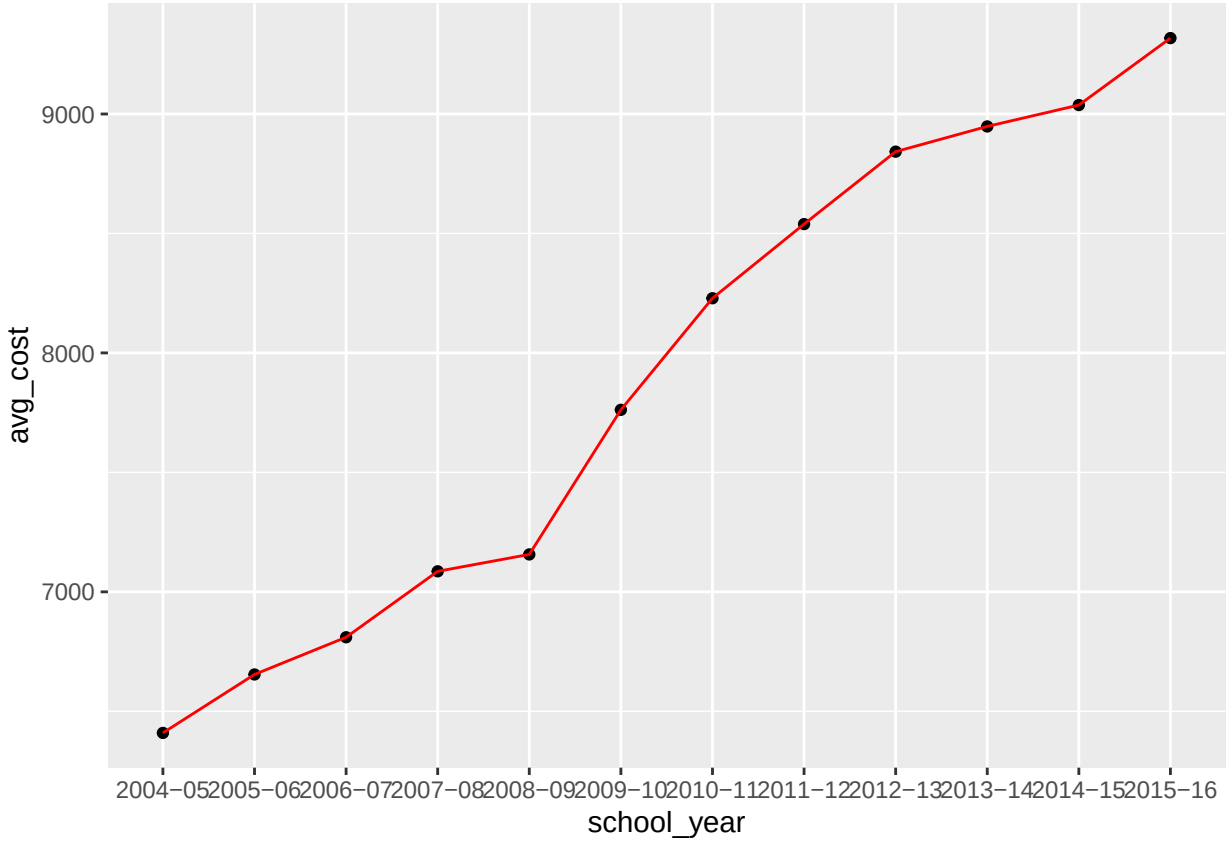
---

7) National trend over time — Visual 1

Goal: Using `national_by_year` (provided for you), create plot `p1_national_line` that maps `x = school_year` to `y = avg_cost`. Success check: `nrow(national_by_year) == 12`.

```
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 <- national_by_year %>%
  ggplot(aes(x = school_year, y = avg_cost, group = 1))+
  geom_point(color = "Black")+
  geom_line(color = "Red")
p1_national_line
```



Describe what plot this is with a short interpretation (1–2 sentences) → store in `viz1_note`.

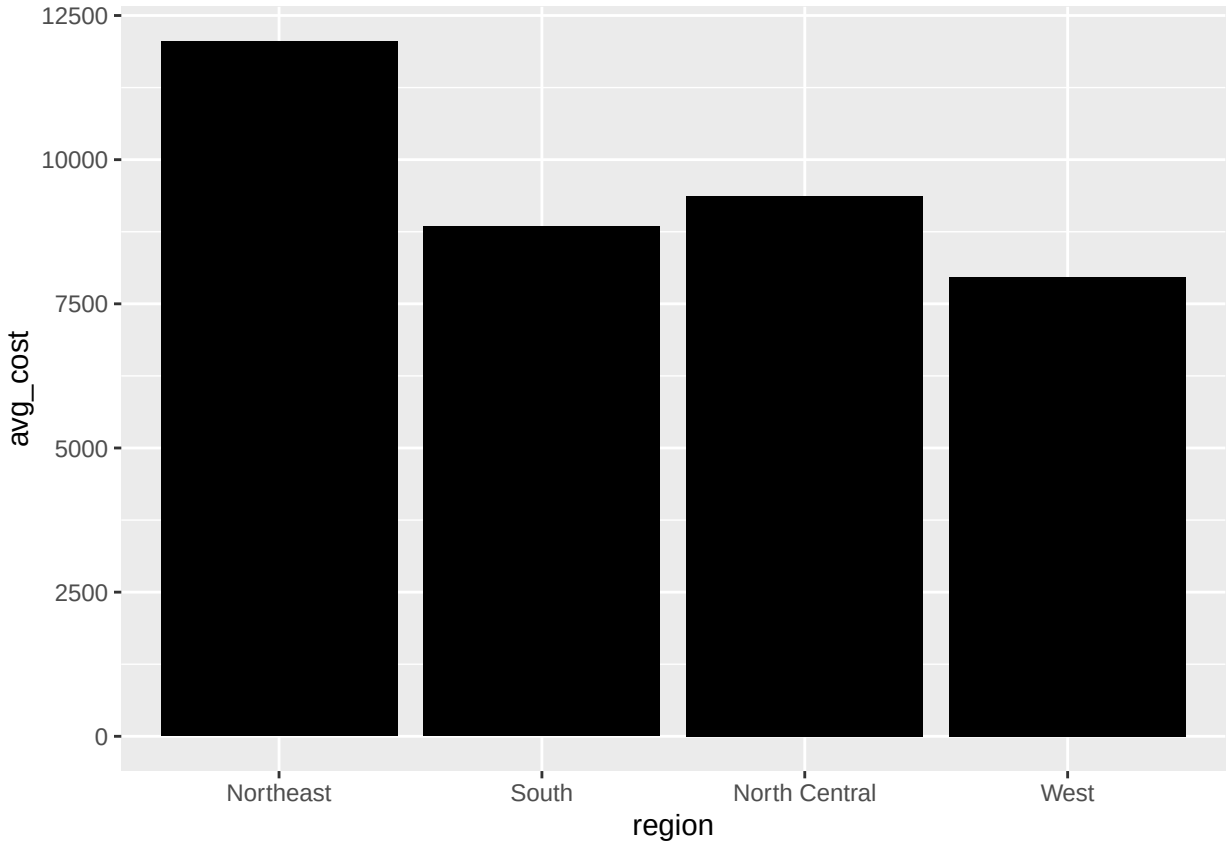
```
viz1_note <- 'This ggplot shows the national average cost of tuition by school year. The plot demonstra
```

---

8) Regional comparison (2015–16) — Visual 2

Goal: create `p2_region_bars` (chart of `region_cost_2015`). Plot must map `x = region`, `y = avg_cost`.

```r
p2_region_bars <- region_cost_2015 %>%
  ggplot(aes(x = region, y = avg_cost)) +
  geom_col(fill = "Black")
p2_region_bars
```



Describe what plot this is with a short interpretation (1–2 sentences) → `viz2_note`.

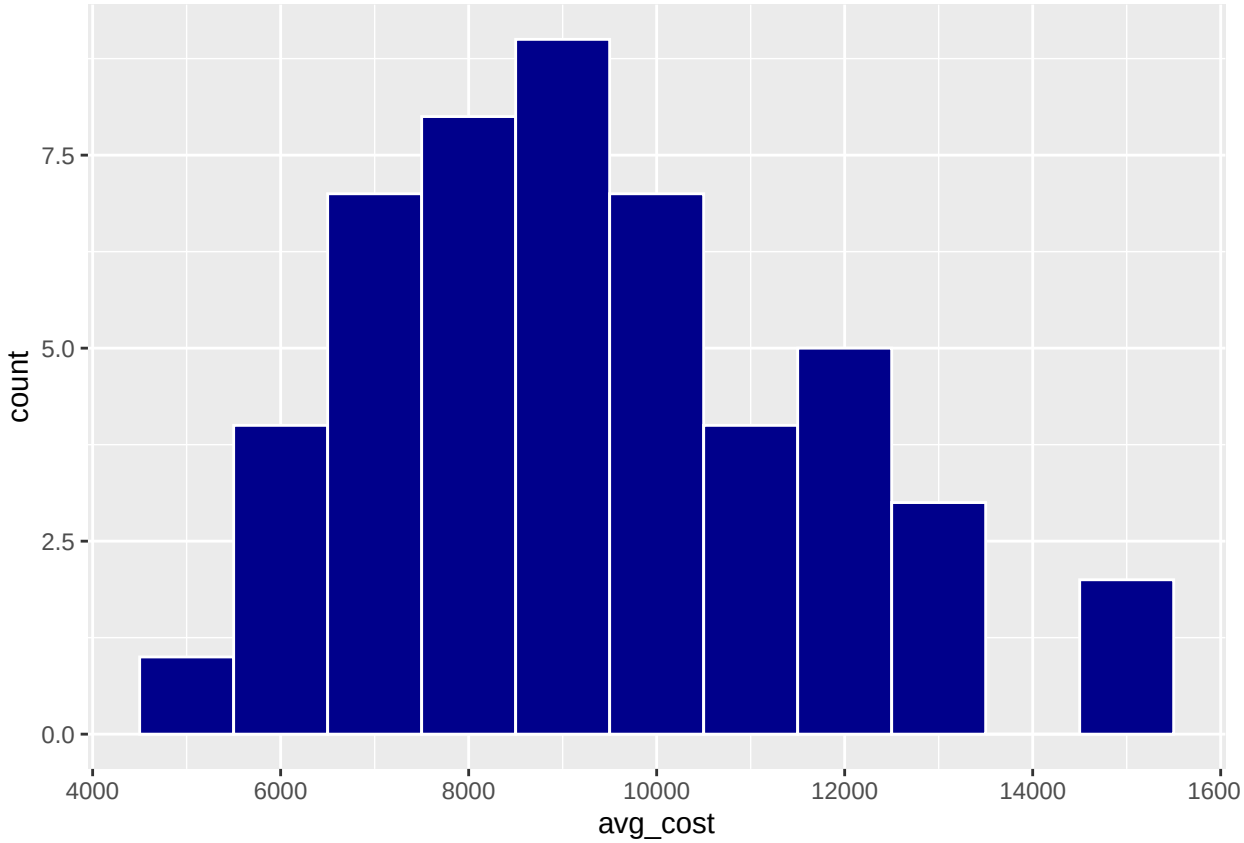
```r
viz2_note <- 'This ggplot visualizes the average tuition cost by region for the 2015-2016 school year. I
```

---

9) State distribution (2015–16) — Visual 3

Goal: plot of state `avg_cost` in 2015–16 → `p3_state_hist_2015`. Plot must map `x = avg_cost`.

```r
p3_state_hist_2015 <- tuition_with_region %>%
  filter(school_year == "2015-16") %>%
  ggplot(aes(x = avg_cost)) +
  geom_histogram(binwidth = 1000, fill = "darkblue", color = "white")
p3_state_hist_2015
```



Describe what plot this is with a short interpretation (1–2 sentences) → `viz3_note`.

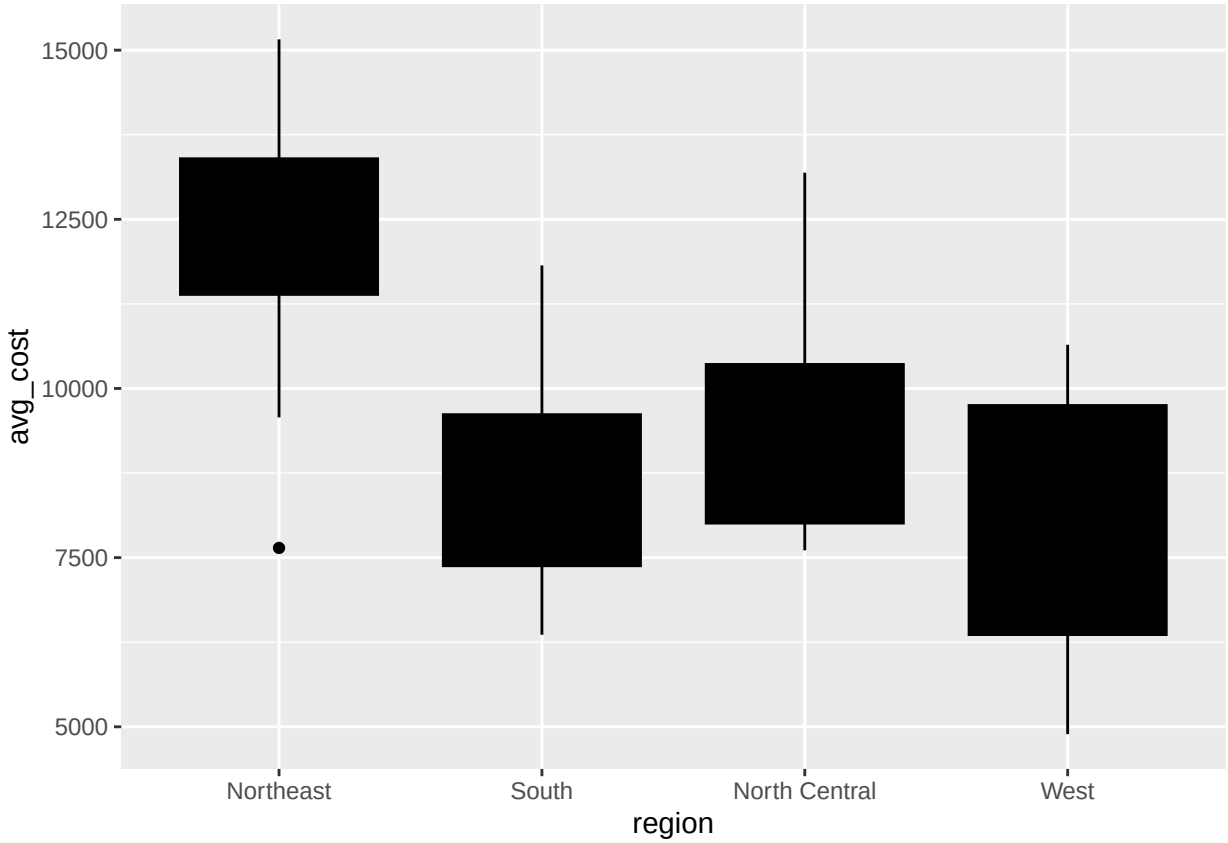
```
viz3_note <- 'This histogram visualizes the distribution of average tuition cost across states in 2015-
The histogram shows that most states fall within the range of 7,000 to 10,000 for average tuition cost
```

---

10) Regional spread (2015–16) — Visual 4

Goal: plot of `avg_cost` by `region` in 2015–16 → `p4_region_box_2015`. Plot must map `x = region`, `y = avg_cost`.

```
p4_region_box_2015 <- tuition_with_region %>%
  filter(school_year == "2015-16") %>%
  ggplot(aes(x = region, y = avg_cost)) +
  geom_boxplot(fill = "black", color = "black")
p4_region_box_2015
```



Describe what plot this is with a short interpretation (1–2 sentences) → `viz4_note`.

```
viz4_note <- 'This boxplot compares the distrbibution of average tuition across different state regions
```

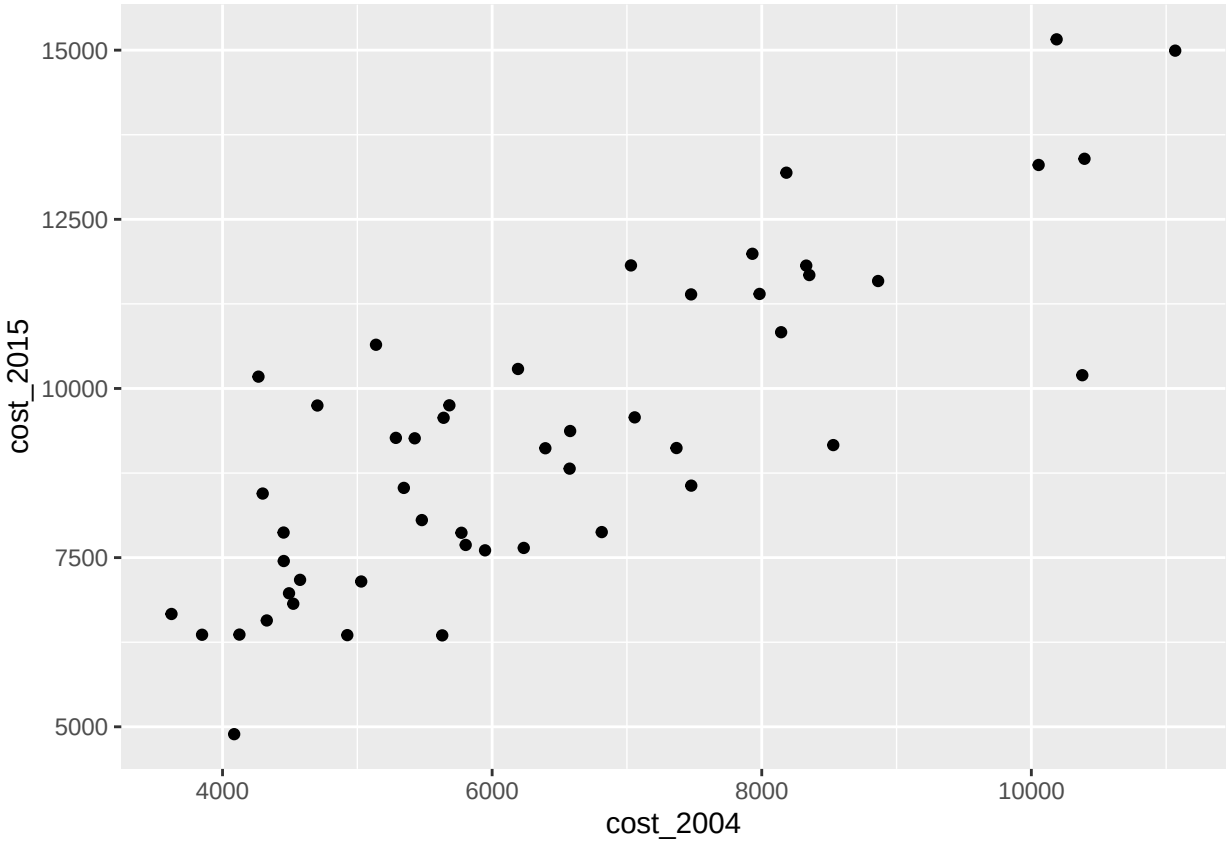
---

11) Early vs late comparison — Visual 5 (,004–05 vs 2015–16)

Goal: build a two-column table for 2004–05 and 2015–16 named `two_years_wide`, then create plot `p5_scatter_2004_vs_2015`. Plot must map `x = cost_2004` to `y = cost_2015`.

```
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`)
```

```
p5_scatter_2004_vs_2015 <-two_years_wide %>%
  ggplot(aes(x = cost_2004, y = cost_2015))+
  geom_point(color = "black")
p5_scatter_2004_vs_2015
```



Describe what plot this is with a short interpretation (1–2 sentences) → `viz5_note`.

```
viz5_note <- 'This scatterplot compares average state tuition costs in 2004-5 versus 2015-6. The graph s
```