# Statistics and Data Analysis
# Unit 02 – Lecture 06 Notes
# In-Class Activity: Summarization + Interpretation

Tofik Ali

February 17, 2026

## Purpose of This Activity

This activity is designed to help you practice descriptive statistics end-to-end. You will not only compute numbers, but also interpret what they mean and write a short conclusion.

**Repository.** `https://github.com/tali7c/Statistics-and-Data-Analysis`

## Learning Outcomes

After this activity, you should be able to:

1. compute central tendency (mean, median, mode),
2. compute dispersion (range, IQR, sample variance, sample std),
3. compute Pearson correlation and interpret its sign and magnitude,
4. compare groups using grouped summaries,
5. write clear insights and limitations from statistical summaries.

## 1. Dataset Description

File: `data/activity_student_dataset.csv`

### 1.1 What the columns mean

- `program`: CSE / ECE / AIML (categorical).
- `attendance_pct`: attendance percentage (numeric).
- `study_hours`: study hours (numeric).
- `social_media_hours`: social media hours (numeric).
- `final_score`: final score (numeric).

## 2. Tasks (What You Must Compute)

### Task 1: Central tendency

For `final_score`, compute:

- mean: $\bar{x} = \frac{1}{n}\sum x_i$
- median (middle value after sorting)
- mode (most frequent value)

**Interpretation tip.** Compare mean vs median:

- mean $>$ median often hints right skew,
- mean $<$ median often hints left skew.

This is only a clue; confirm with a plot.

### Task 2: Dispersion

For `final_score`, compute:

- range $=$ max $-$ min
- quartiles $Q_1, Q_3$ and IQR $= Q_3 - Q_1$
- sample variance and sample standard deviation:

$$s^2 = \frac{1}{n-1}\sum (x_i - \bar{x})^2, \quad s = \sqrt{s^2}$$

**Robustness.** IQR is more robust to outliers than standard deviation.

### Task 3: Correlation

Compute Pearson correlation:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2}\sqrt{\sum (y_i - \bar{y})^2}}$$

for:

- (`study_hours`, `final_score`)
- (`social_media_hours`, `final_score`)

**Very important.** Correlation measures **linear association**. It does **not** prove causation.

### Task 4: Grouped summaries

Group by `program` and compute:

- mean and median of `final_score`
- mean of `attendance_pct`

### Task 5: Write-up

Write:

- 3 insights (what the dataset suggests)
- 2 limitations (why your conclusions might not generalize)

## 3. Expected Results (From the Provided Solution)

If you run the provided solution script (next section), you should obtain:

### 3.1 Overall results

- mean(final_score) = 65.50
- median(final_score) = 65.50
- mode(final_score) = 60
- $Q_1 = 60,\ Q_3 = 72,\ IQR = 12$
- sample std(final_score) $\approx 14.11$

### 3.2 Correlations

- corr(study_hours, final_score) $\approx 0.5190$ (moderate positive linear association)
- corr(social_media_hours, final_score) $\approx$ -0.9771 (strong negative linear association)

**Interpretation note.** Even a strong correlation does not prove that one variable causes the other. Other variables (motivation, prior knowledge, teaching quality) could be involved.

### 3.3 By program (mean final score)

- AIML: mean $\approx 75.83$
- CSE: mean $\approx 59.17$
- ECE: mean $\approx 61.50$

## 4. Provided Solution Script (Mini Demo)

After you attempt the activity yourself, run:

```
python demo/activity_solution.py
```

It will save:

- `data/overall_results.csv`
- `data/summary_by_program.csv`
- plots in `images/` (scatter plots, bar chart, histogram)

## 5. Common Mistakes

- **Mixing up correlation and causation.** Correlation is not proof.
- **Using only the mean.** Always look at median and IQR too.
- **Ignoring groups.** A global average can hide group differences.
- **Not writing limitations.** Every conclusion must mention what could be wrong.
- **No plots.** Tables alone can hide skewness and outliers.

## 6. Extension Questions (Optional)

If you finish early, try any two:

1. Identify the lowest and highest scoring students and comment on their study/social media hours.
2. Compare correlation within each program separately (CSE vs ECE vs AIML).
3. Replace mean with median for program comparison and see if ranking changes.
4. Use IQR fences to flag potential outliers in `final_score`.

## References


- Montgomery, D. C., & Runger, G. C. *Applied Statistics and Probability for Engineers*, Wiley, 7th ed., 2020.
- Freedman, D., Pisani, R., & Purves, R. *Statistics*, W. W. Norton, 4th ed., 2007.
- McKinney, W. *Python for Data Analysis*, O'Reilly, 2022.

## Appendix: Slide Deck Content (Reference)

The material below is a reference copy of the slide deck content. Exercise solutions are explained in the main notes where applicable.

### Title Slide

https://github.com/tali7c/Statistics-and-Data-Analysis

## Quick Links

**Activity Tasks Solution Wrap-up**

## Agenda

- Activity Brief
- Tasks and Deliverables
- Solution and Discussion
- Wrap-up

## What We Will Do Today

You will complete a small end-to-end descriptive statistics task:

- compute central tendency (mean/median/mode)
- compute dispersion (range, IQR, variance, std)
- compute correlation for two variable pairs
- create grouped summaries by program
- write 3 insights + 2 limitations

## Time Plan (55 minutes)

- 10 min: attendance + setup
- 25 min: activity work (in pairs)
- 10 min: discussion (compare results and assumptions)
- 5 min: wrap-up + exit question

## Dataset

File: `data/activity_student_dataset.csv`

Columns:

- `program` (CSE/ECE/AIML)
- `attendance_pct`, `study_hours`, `social_media_hours` (numeric)
- `final_score` (numeric)

**Goal:** summarize and interpret what these numbers suggest.

### Task 1: Central Tendency (5 minutes)

Compute for `final_score`:

- mean
- median
- mode

**Checkpoint:** If mean $\neq$ median, what does that hint about skewness?

### Task 2: Dispersion (10 minutes)

Compute for `final_score`:

- range
- $Q_1$, $Q_3$, IQR
- sample variance and sample standard deviation

**Checkpoint:** Which is more robust to outliers: std or IQR?

### Task 3: Correlation (10 minutes)

Compute Pearson correlation:

- corr(`study_hours`, `final_score`)
- corr(`social_media_hours`, `final_score`)

**Checkpoint:** Correlation measures what kind of relationship?

### Task 4: Grouped Summaries (10 minutes)

Group by `program` and compute:

- mean and median of `final_score`
- mean attendance

**Checkpoint:** Which program looks strongest by mean? By median?

### Task 5: Write Insights + Limitations (5 minutes)

Deliver:

- 3 insights (what the numbers suggest)
- 2 limitations (why the conclusion may be weak)

**Example limitation:** small dataset $\Rightarrow$ results may not generalize.

## Final Deliverables (Submit/Show)

- computed values (central tendency, dispersion, correlations)
- 1 grouped summary table by program
- 2 scatter plots OR 1 histogram + 1 bar chart
- short write-up (3 insights + 2 limitations)

## Solution Script (Python)

After attempting yourself, run:

`python demo/activity_solution.py`

Outputs:

- `data/overall_results.csv`
- `data/summary_by_program.csv`
- plots in `images/` (scatter, bar, histogram)

## Expected Key Results (Overall)

| Statistic | Value |
|---|---|
| Mean(final_score) | 65.50 |
| Median(final_score) | 65.50 |
| Mode(final_score) | 60 |
| Range(final_score) | 65 |
| $Q_1$ / $Q_3$ | 60 / 72 |
| IQR | 12 |
| Sample std (final_score) | 14.11 |

## Expected Key Results (Correlation)

| Pair | Pearson $r$ |
|---|---|
| (study_hours, final_score) | 0.5190 |
| (social_media_hours, final_score) | -0.9771 |

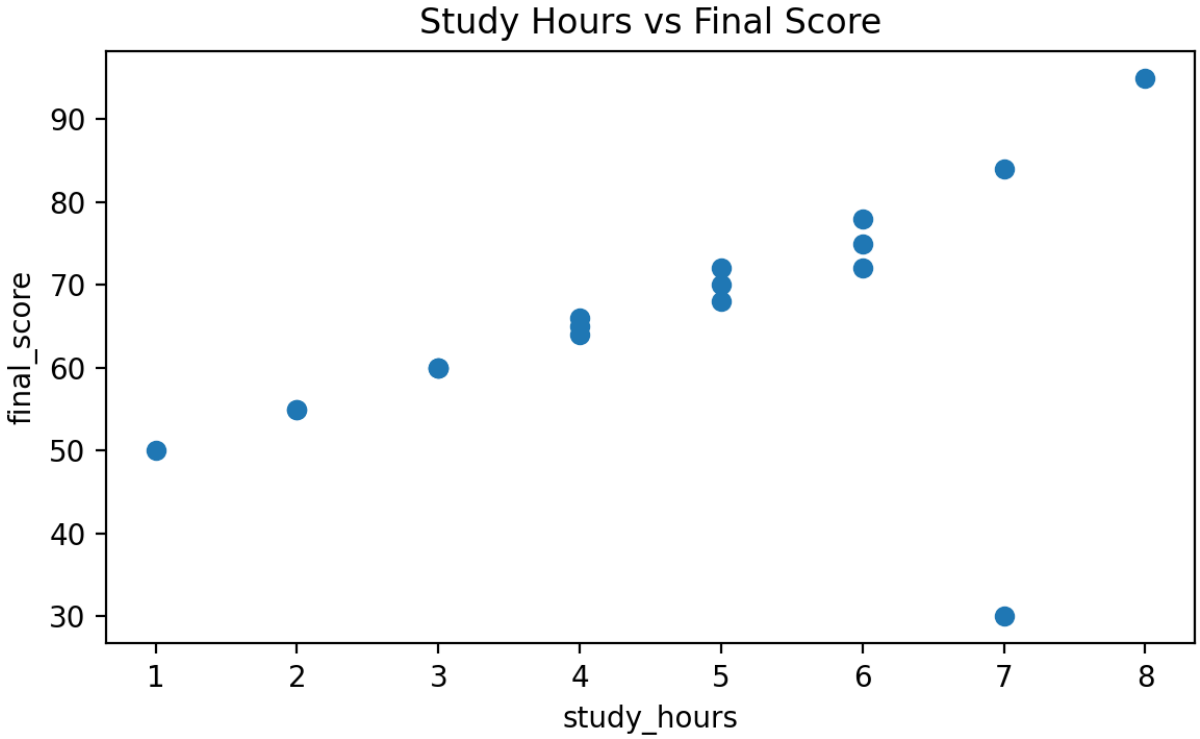
**Question:** Does this prove causation? Why/why not?

## Expected Key Results (By Program)

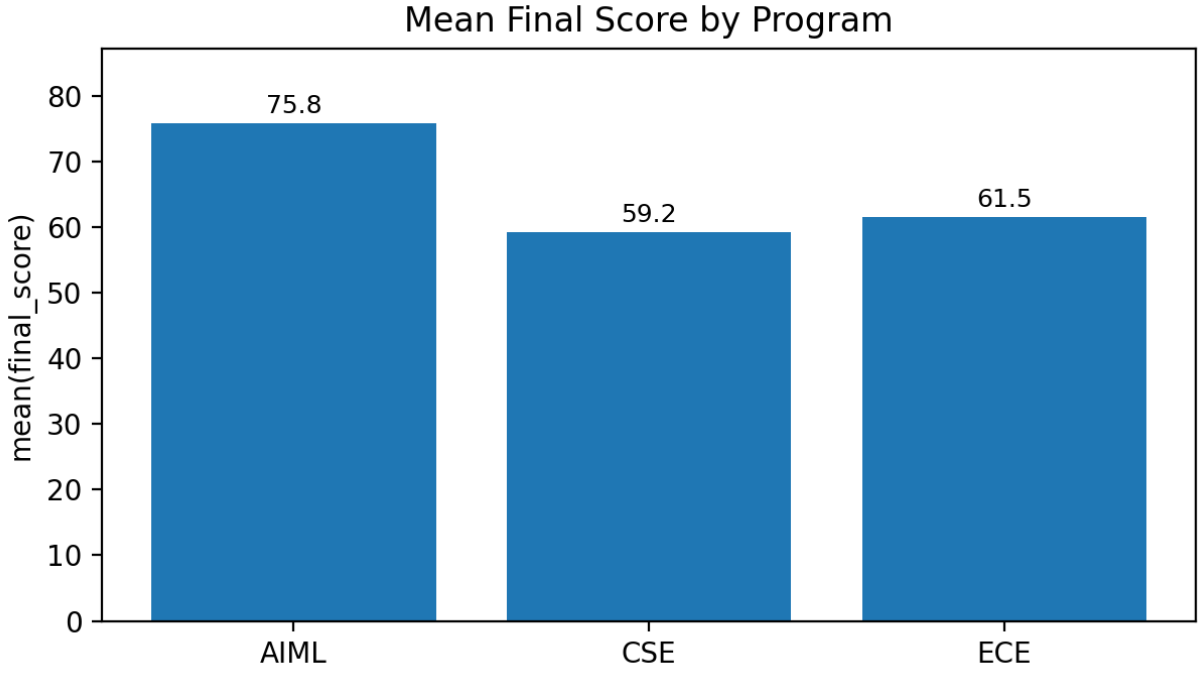
| Program | Mean(final_score) | Median(final_score) |
|---|---|---|
| CSE | 59.17 | 62.50 |
| ECE | 61.50 | 62.00 |
| AIML | 75.83 | 75.00 |

## Example Plots

### Hours vs Final Score




### Mean Final by Program



## Wrap-up

- A good descriptive analysis combines: center + spread + relationships + group comparisons
- Always state assumptions and limitations
- Never confuse correlation with causation

**Exit question:** Which statistic changed your interpretation the most (mean, median, IQR, or correlation)? Why?