Statistics and Data Analysis

Unit 02 – Lecture 06: In-Class Activity (Summarization + Interpretation)

Tofik Ali

School of Computer Science, UPES Dehradun

February 14, 2026

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

Quick Links

Activity Tasks Solution Wrap-up

Agenda

What We Will Do Today

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

- compute central tendency (mean/median/mode)

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)

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

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

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

Time Plan (55 minutes)

- 10 min: attendance + setup
- 25 min: activity work (in pairs)

Time Plan (55 minutes)

- 10 min: attendance + setup
- 25 min: activity work (in pairs)
- 10 min: discussion (compare results and assumptions)

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

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

Task 1: Central Tendency (5 minutes)

Compute for `final_score`:

- mean
- median

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

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

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

Task 2: Dispersion (10 minutes)

Compute for `final_score`:

- range
- Q_1, Q_3, IQR

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

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`)

Checkpoint: Correlation measures what kind of relationship?

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`

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

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)

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

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)

Final Deliverables (Submit/Show)

- computed values (central tendency, dispersion, correlations)
- 1 grouped summary table by program

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

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

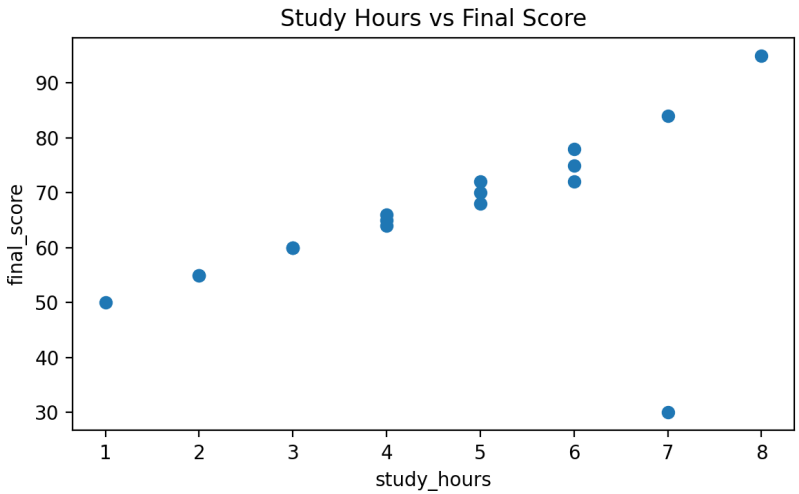
Question: Does this prove causation? Why/why not?

Expected Key Results (By Program)

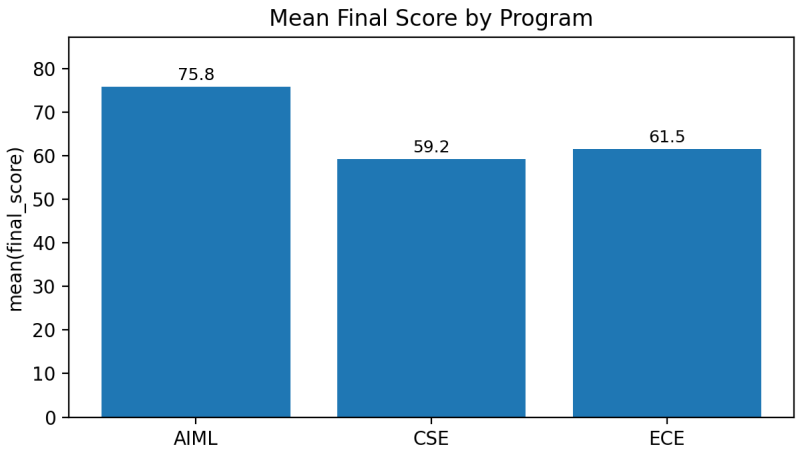
Program	Mean(final_score)	Median(final_score)
CSE	59.17	62.50
ECE	61.50	62.00
AIML	75.83	75.00

Example Plots

Hours vs Final Score



Mean Final by Program



Wrap-up

- A good descriptive analysis combines: center + spread + relationships + group comparisons

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

Wrap-up

- A good descriptive analysis combines: center + spread + relationships + group comparisons
- Always state assumptions and limitations

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

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?