Study Guide for 10-Minute In-Class Quiz 1

Rescheduled: Thursday in class (not Tuesday)

This guide will help you succeed on the upcoming quiz. You will have **10 minutes in class**, and the quiz is **open-book**. I will be walking around during the quiz to make sure you are **not using generative AI tools**. If you prepare the material below, the quiz will feel straightforward and you will be set up to earn an **A**.

—

1. The Data

The dataset is called `baseball.csv` and covers Major League Baseball teams from 1998–2014.

It includes:

- `team`: Team name
- `year`: Season year
- `payroll`: Team payroll in millions of dollars
- `win_num`: Number of wins
- `win_pct`: Winning percentage (0 to 1)

There are **510 rows** and **5 columns** — one row for each team-season.

—

2. Python Skills You Need

Be ready to:

- Load the dataset with `pandas`.
- Check its size with `.shape` and check for missing values with `.isnull().sum()`.
- Group by team and calculate totals and averages using `groupby().agg()`.
- Compute mean and standard deviation with `.mean()` and `.std()`.
- Create simple plots (scatterplots, line plots, boxplots) using `matplotlib` or `seaborn`.

—

3. Statistics Concepts

Make sure you understand:

- What the sample mean and standard deviation tell you.
- Why the league-wide average winning percentage is about 0.5.
- The **68–95–99.7 rule** of the normal distribution and how to interpret data relative to the mean.
- How to interpret changes and comparisons across teams and years.

—

4. Visualization Interpretation

Be ready to explain what you see in:

- **Scatterplots**: payroll vs. win percentage, across different time periods.
- **Line plots**: payroll trends over time for specific teams (e.g., high-spending vs. low-spending teams).
- **Boxplots**: payroll distributions by year (and what the median shows).

—

5. Statistical Learning Framework

Know the basics:

- Inputs: domain X, labels Y, training data S.
- Output: prediction rule $h : X \to Y$.
- Error: $L_{D,f}(h) = \Pr_{x \sim D}[h(x) \neq f(x)]$.
- Difference between **true error** and **empirical error**.

—

6. Principal Component Analysis (PCA)

You should understand:

- PCA creates new variables (principal components) as linear combinations of all features.
- Loadings define the direction of components and are unique up to a sign change.
- How to read PCA plots: PC1 vs. PC2, scores of points, and meaning of loadings.

—

7. Quiz Logistics

- The quiz will be on **Thursday**, in class.
- Time limit: **10 minutes**.
- It is open-book: you may use notes, textbook, or laptop.
- No generative AI tools are allowed — I will be checking.

—

How to Earn an A

- Practice loading the dataset and computing summaries in Python.
- Review statistical basics: mean, standard deviation, normal distribution.
- Be confident interpreting scatterplots, line plots, and boxplots.
- Refresh your understanding of the Statistical Learning Framework and PCA.

If you prepare these areas, this will be an **easy 10-minute quiz**.