# DATASCI 347 Quiz 1

Open Book – 10 Minutes

Choose the correct answer(s). There might be more than one right answer in some questions. No calculations are needed.

We use Major League Baseball data for most of the questions. The dataset `baseball.csv` contains payroll and winning for 30 Major League teams for a span of 1998 to 2014.
**Variables:**

- `team`: team name
- `year`:
- `payroll`: team payroll in millions
- `win_num`: number of wins
- `win_pct`: winning percentage

1. Let us first read the data in **Python**:

```
import pandas as pd
baseball = pd.read_csv("baseball.csv")
baseball.shape
# (510, 5)
```

Based on the above *Python* chunk only, choose the correct answer(s):

A. The dataset baseball.csv has 510 rows.

B. The dataset baseball.csv has no missing values

2. We then aggregate the data which only contains the total payroll and average winning percentage for each team. They are `team`, `payroll_total`, and `win_pct_ave`. `payroll_total` is in billions. The data is stored as `data_agg`.

```
# create total and average winning percentage for each team
data_agg = baseball.groupby("team").agg(
    payroll_total = ("payroll", lambda x: x.sum()/1000),
    win_pct_ave   = ("win_pct", "mean")
).reset_index()

# Here are some summaries:
data_agg["win_pct_ave"].mean()  # 0.5
data_agg["win_pct_ave"].std()   # 0.038
```

Based on the information provided above, the sample mean of `win_pct_ave` is 0.5.

A. True

B. False

3. The sample mean of `win_pct_ave` should always be 0.5 because of the nature of variable: one team loses the opponent wins (assume no ties and each pair of team plays one game against each other).

   A. True

   B. False

4. Assume that `win_pct_ave` follows a normal distribution. From the data, Oakland Athletics's `win_pct_ave` = 0.54.

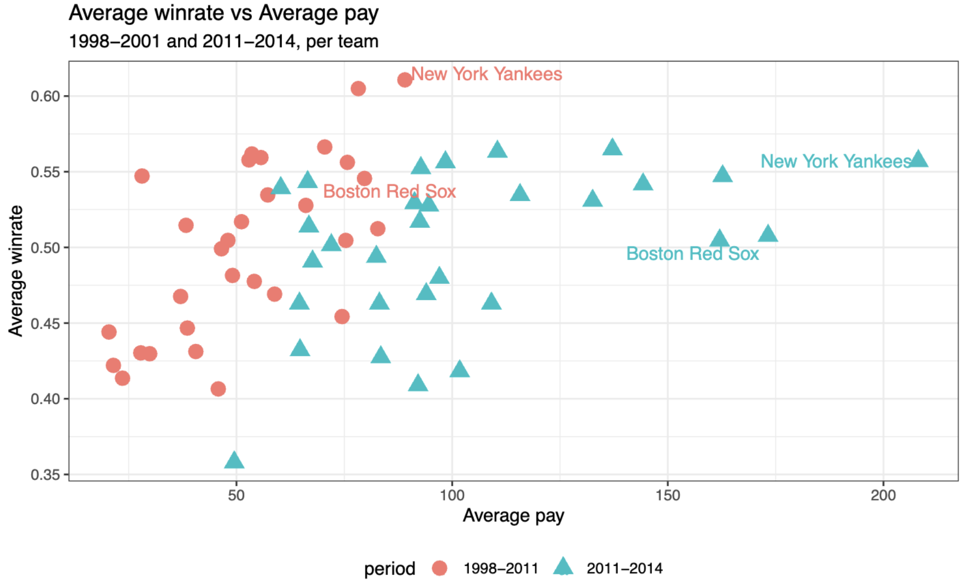
   A. Approximately, 5% of the teams have a higher `win_pct_ave` than that of Oakland Athletics.

   B. Approximately, 2.5% of the teams have a higher `win_pct_ave` than that of Oakland Athletics.

   C. Approximately, 16% of the teams have a higher `win_pct_ave` than that of Oakland Athletics.

   D. Approximately, 16% of the teams have a lower `win_pct_ave` than that of Oakland Athletics.

5. **Question 5 and 6 are based on the following plot:** We use a subset of the Major League Payroll dataset, which only includes average win percentage and average pay for two periods: 1998–2001 (early) and 2011–2014 (late). In the below scatter plot, we plot average win percentage vs average pay for these two periods.



Average per-team spending on payroll increased from the early (red round points) to late periods (blue triangle points).

A. True

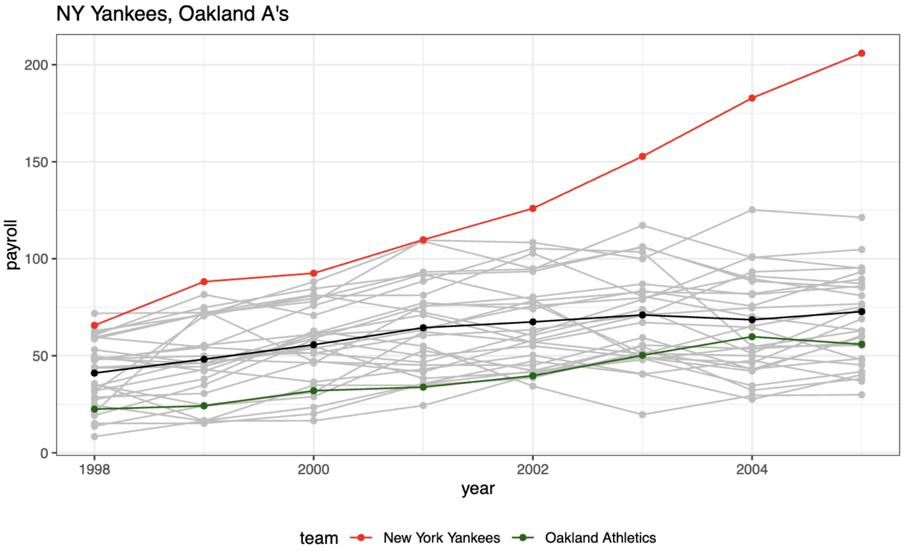
B. False

6. The team that spent the most on players also won the most games, in both periods.

A. True

B. False

7. The following spaghetti plot shows the payroll of each team from 1999 to 2005. The red line is New York Yankees, the green line is Oakland Athletics and the gray lines are the rest of the teams. The black line is the mean payroll of each year.



Choose the correct answer(s):

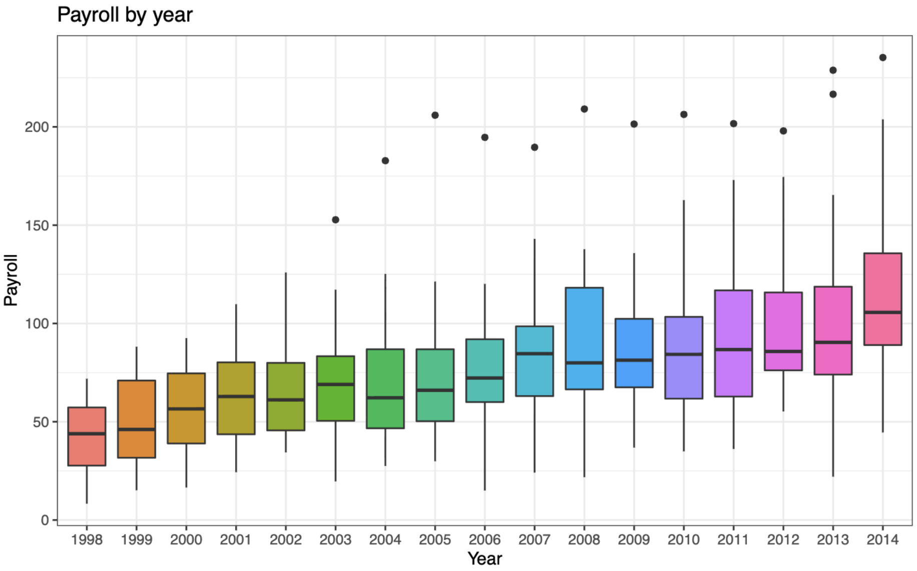
A. New York Yankees is always the highest paid team.

B. The pay of Oakland Athletics is always below average.

C. The increase in the payroll of Oakland Athletics over the period of a year is always below the average raise per year

D. None of the above.

8. The following shows the boxplot of payroll by year from 1998 to 2014.



The median payroll has been increasing every year.

A. True

B. False

9. Define the components of a formal model in the Statistical Learning Framework. Which of the following are *inputs* to the learner?

A. Domain set $X$

B. Label set $Y$

C. Training data $S = ((x_1, y_1), \ldots, (x_m, y_m))$

D. The prediction rule $h : X \to Y$

10. In the Statistical Learning Framework, what is the learner's *output*?

A. A distribution $D$ over $X$

B. A labeling function $f : X \to Y$

C. A prediction rule $h : X \to Y$

D. The training set $S$

11. The error of a classifier $h$ with respect to a distribution $D$ and target function $f$ is defined as:

A. $L_{D,f}(h) = \Pr_{x \sim D}[h(x) \neq f(x)]$

B. $L_{D,f}(h) = \Pr_{x \sim D}[h(x) = f(x)]$

C. $L_{D,f}(h) = \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}[h(x_i) \neq y_i]$

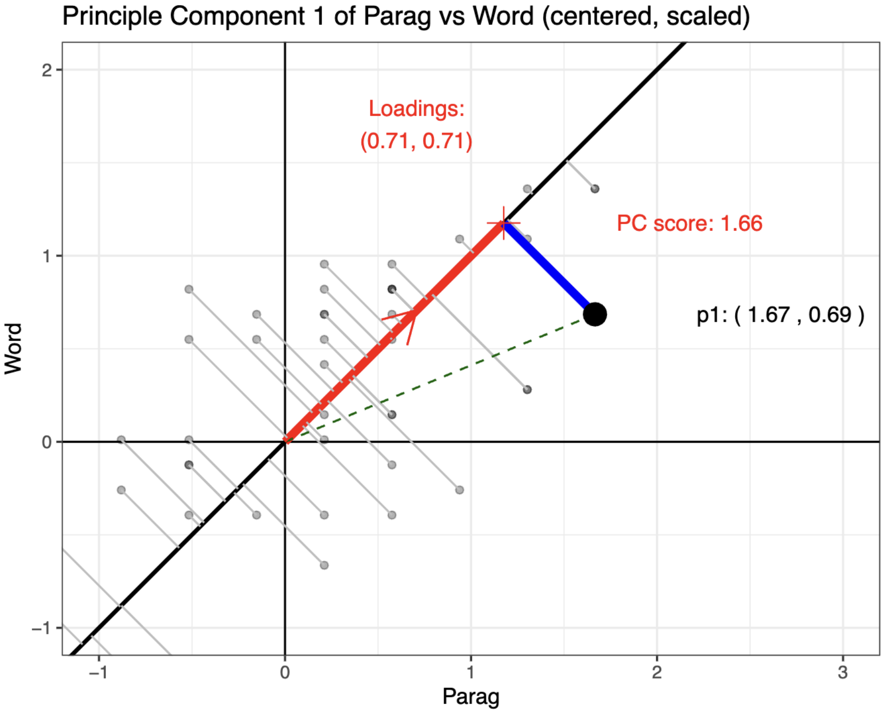
D. None of the above

12. When we perform PCA on SVABS's 10 tests, the leading PC component (or PC1 score)

    A. is a linear combination of 5 tests chosen from SVABS.

    B. is a linear combination of all 10 tests from SVABS.

    C. is the highest score among the 10 tests.



13. Recall the PCA plot of scaled and centered Word and Parag from the AFQT tests from the lecture. The line with $(.71, .71)$ is the PC1 direction. Note $p1$ is the point with $(1.67, 0.69)$ on the graph.

    Choose the correct answer(s):

    A. $p1$ has the largest PC score on PC1.

    B. Loadings of PC1 is $(0.71, 0.71)$

    C. $(-0.71, -0.71)$ is also loadings for PC1