Quiz 1

Modern Data Mining

February 9, 2021

Name: ____

Section (571, 701): ____

This is an open book, 10-minute quiz. 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 team for a span of 1998 to 2014.

`team`: team name

`year`:

`payroll`: team payroll in millions

`win_num`: number of wins

`win_pct`: winning percentage

Let us first read the data:

```r
baseball <- read.csv("baseball.csv", header = TRUE, stringsAsFactors = F)
dim(baseball)
```

```
## [1] 510   5
```

1. Based on the above r-chunk only, choose the correct answer(s):

(A) The dataset `baseball.csv` has 510 rows.

(B) The dataset `baseball.csv` has no missing values

Answer is (A). One would need more information to know if there are any missing values.

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`.

```r
# create total and average winning percentage for each team
data_agg <-baseball %>%
  group_by(team) %>%
  summarise(
    payroll_total = sum(payroll)/1000,
    win_pct_ave = mean(win_pct))
```

Here are some summaries:

```
mean(data_agg$win_pct_ave)
```

```
## [1] 0.5
```

```
sd(data_agg$win_pct_ave)
```

```
## [1] 0.038
```

Based on the information provided above,

2. The sample mean of `win_pct_ave` is 0.5.

(A) True

(B) False

Answer is (A).

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

Answer is (A).

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`.

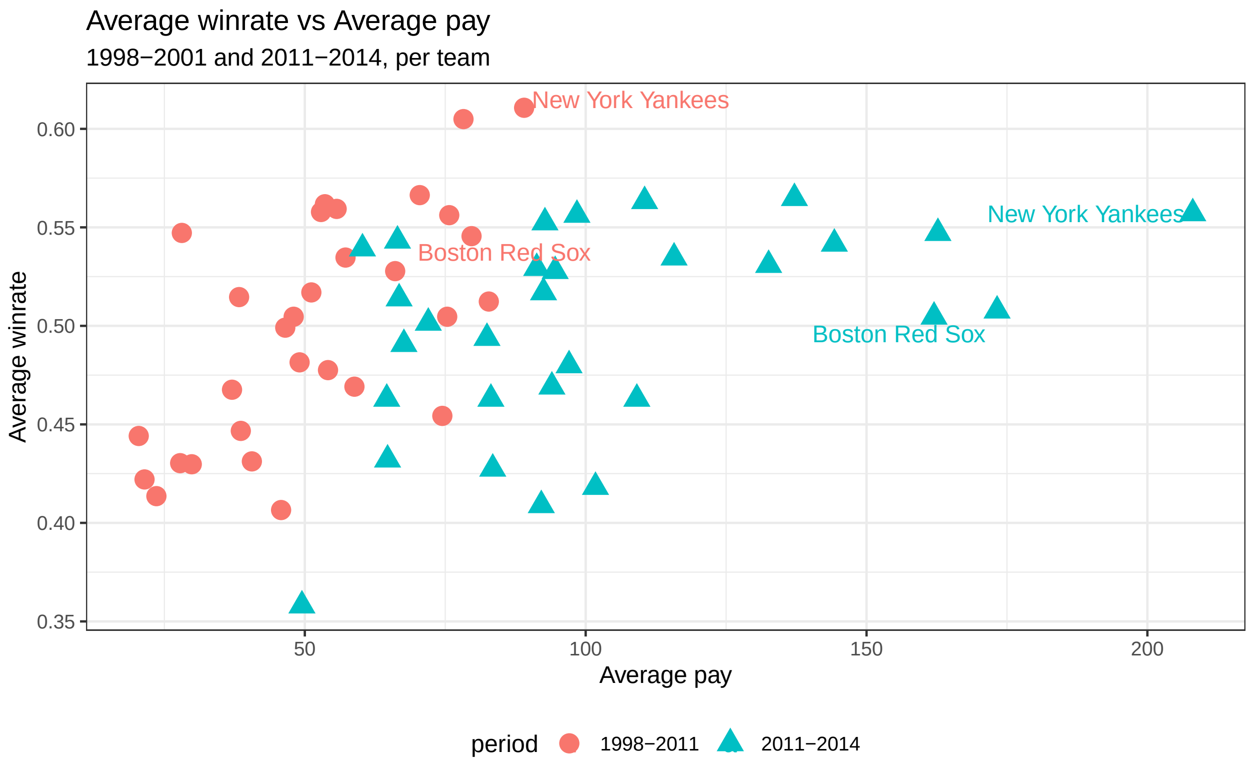
(C) Approximately, 16% of the teams have a higher `win_pct_ave` than that of `Oakland Athletics`.

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

Answer is (C). By normality we know that approximately 68% of the observations fall within one standard deviation from the mean. This is equivalent to say that 16% of the observations fall above one standard deviation above the mean.

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.

```
## `summarise()` has grouped output by 'team'. You can override using the `.groups` argument.
```



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

(A) True

(B) False

Answer is (A).

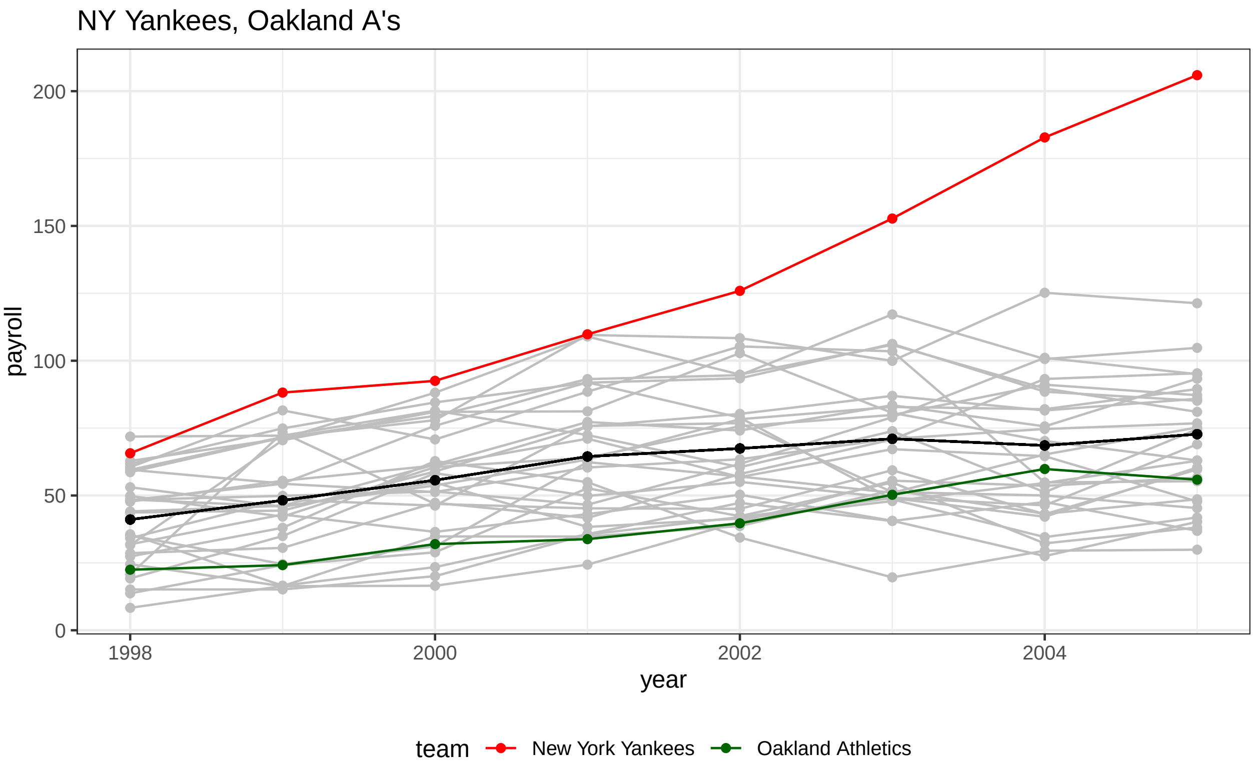
6 The team that spent the most on players also won the most games, in both periods.

(A) True

(B) False

Answer is (B).

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.

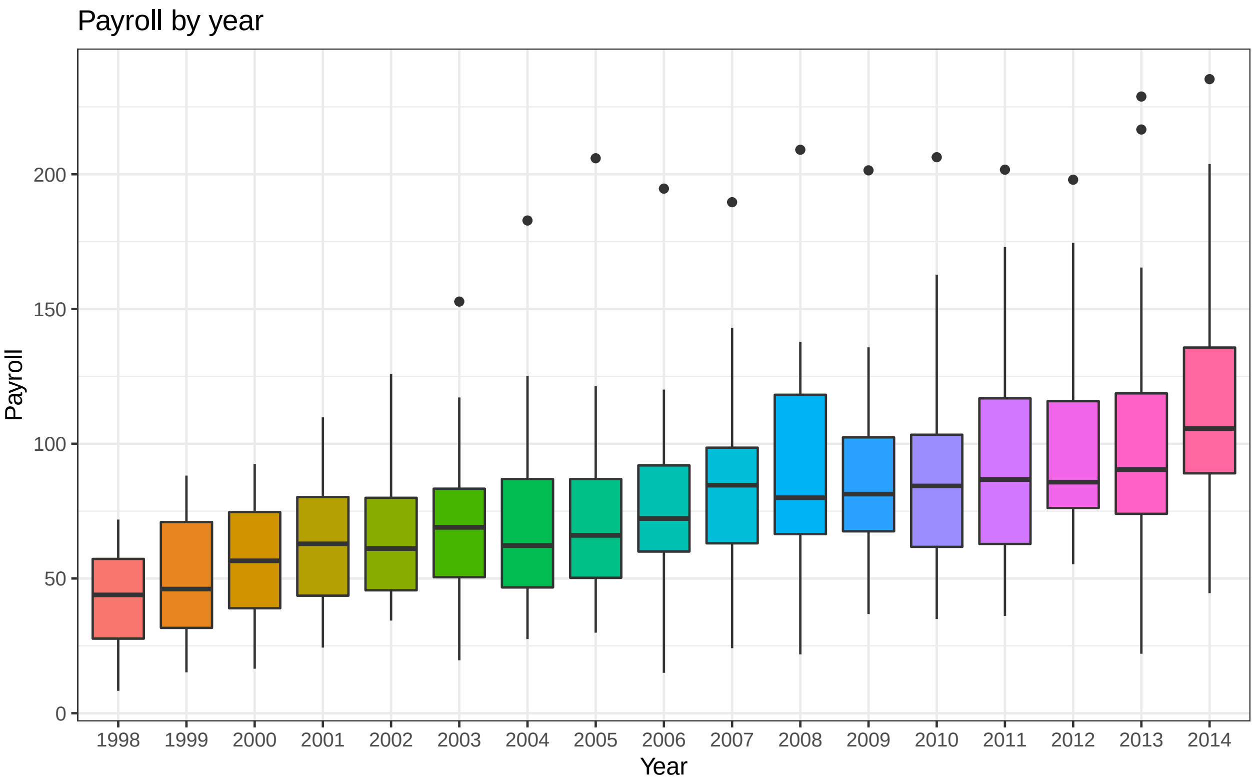
(B) The pay of Oakland Athletics is always below average.

(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.

Answer is (B). (C) is wrong, by noting that there is a larger increase from 02 to 03 for Oakland A's.

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

Answer is (B). The middle bars show medians for a given year. It is clear they are not strictly increasing.

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

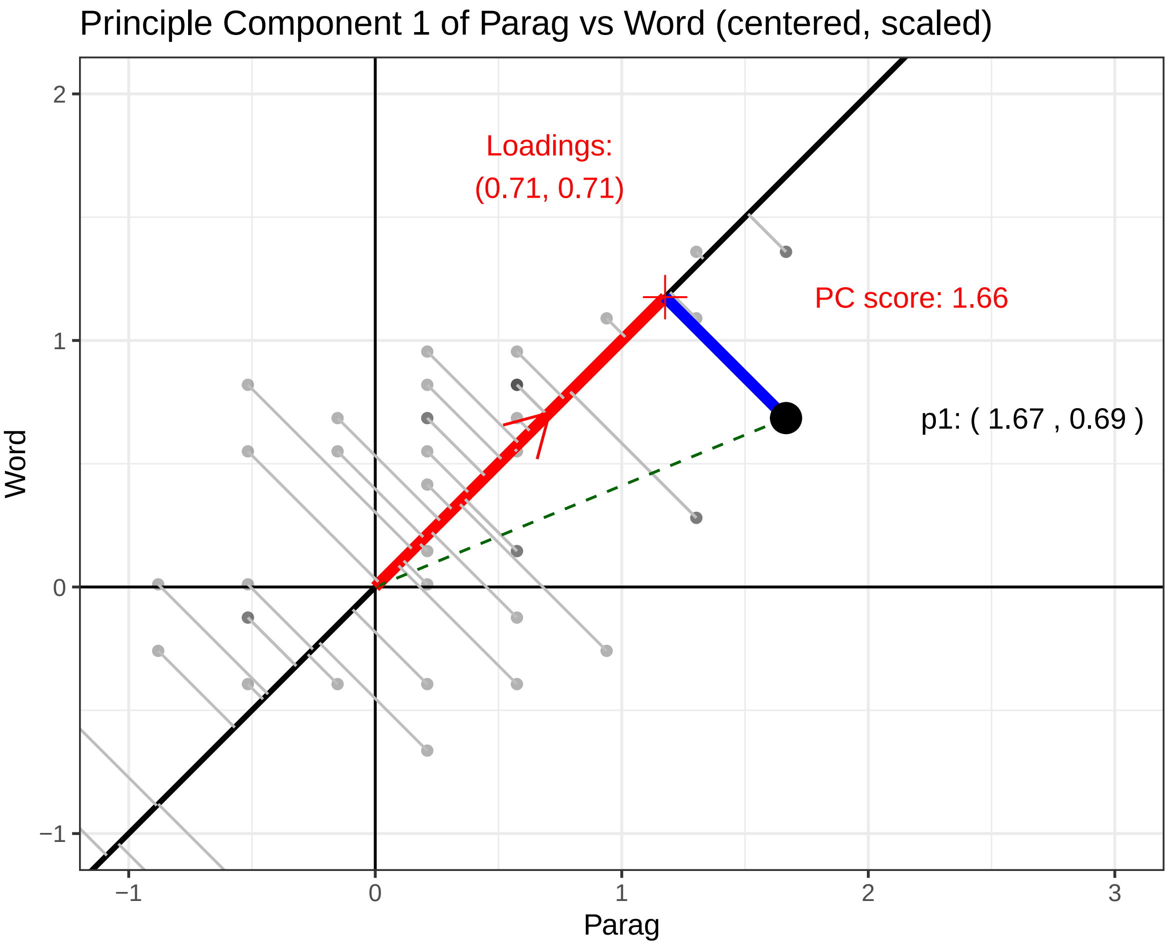
(A) is a linear combination of 5 tests chosen from SVABS.

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

(C) is the highest score among the 10 tests

Answer is (B).

10 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

Answers are (B) and (C). Loadings are unique to the sign. We know $(0.71, 0.71)$ is PC1 loadings so $-(0.71, 0.71) = (-.71, -.71)$ is another set of loadings as well.