NBA SALARIES /STATISTICAL MODELS

Casavecchia Dario

CONTENTS:

/01

Data Preparation

Transformation of the 'Salary' Variable

/02

Descriptive Analysis

Normality Test and Confidence Interval

/03

Hypothesis Testing

Two-sided test on the mean:

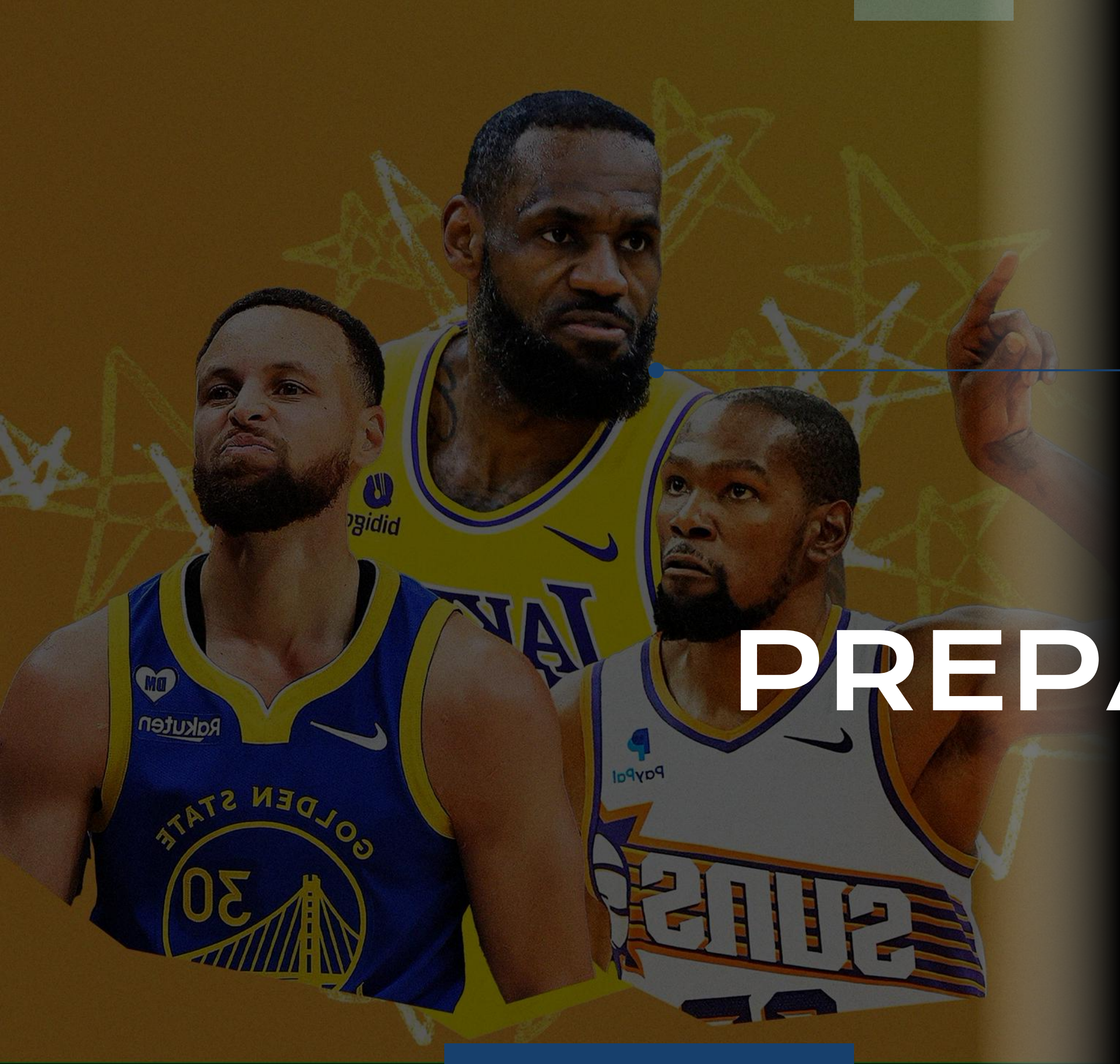
- Equal to $5m
- Greater than $4m

/04

Linear Regression

Two-sided test on the mean:

- Equal to $5m
- Greater than $4m

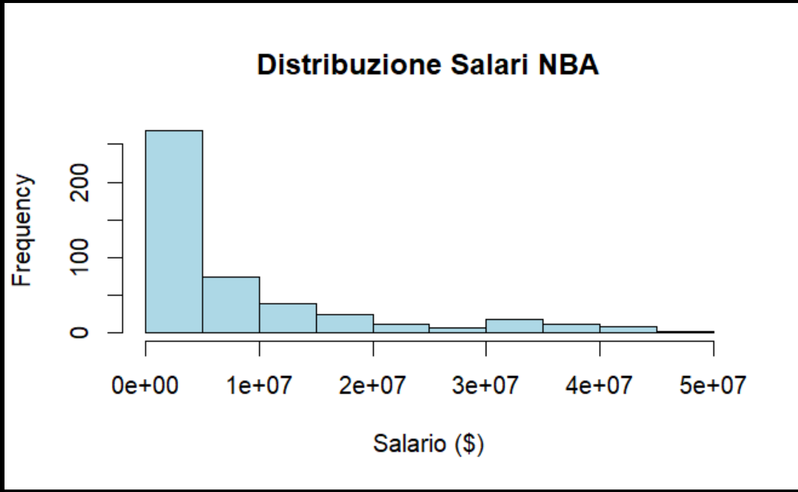


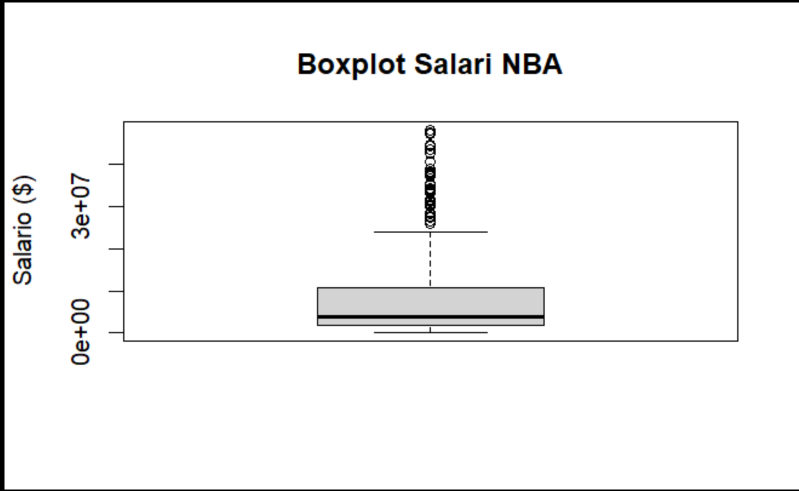
/01

DATA PREPARATION

WITHOUT USING LOG10()

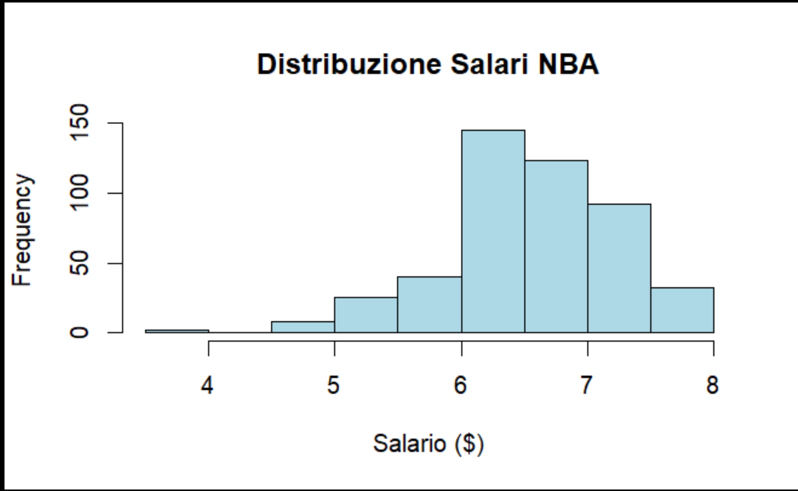
- NBA salaries have a very asymmetrical distribution (few earn a lot and many earn little).
- Less readable graphs.
- Mean influenced by outliers.

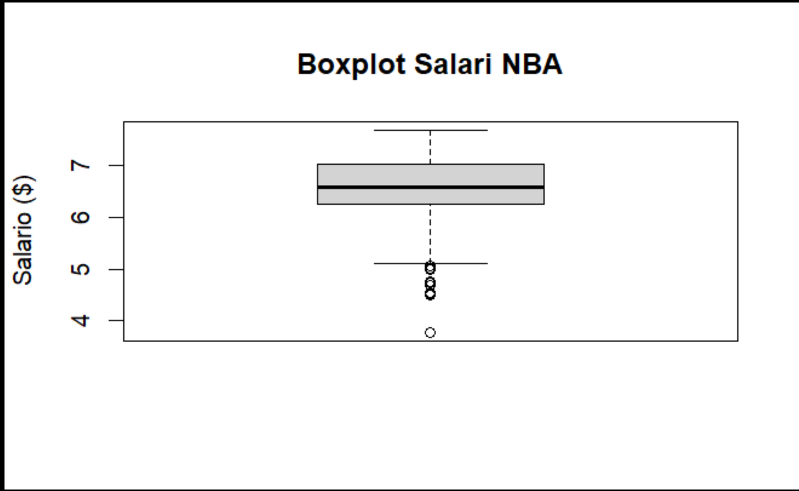




USING LOG10()

- More symmetric and linear distribution.
- Outliers are more contained.
- Data more suitable for building regression models.





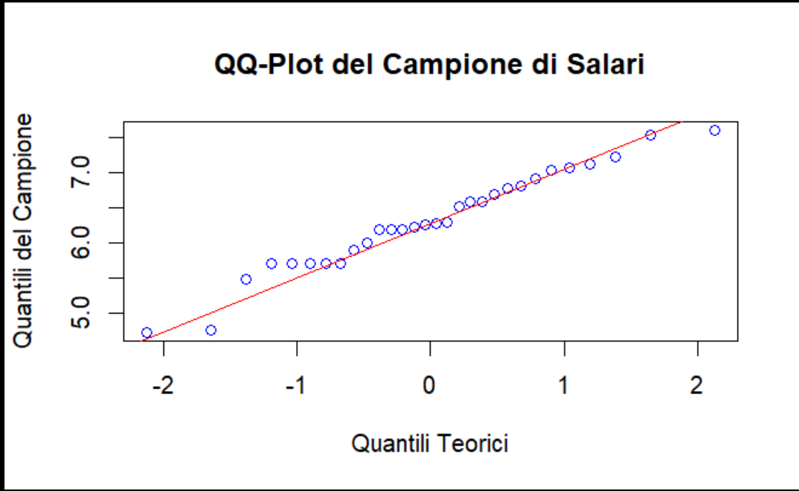
/02

DESCRIPTIVE ANALYSIS

SHAPIRO-WILK TEST

```
        Shapiro-Wilk normality test

data:  sample_salary
W = 0.97083, p-value = 0.5623
```

- **Hypothesis H0 (null):** The sample follows a normal distribution.
- **H1 (alternative):** It does not follow a linear distribution.
- **P-value > 0.05:** Do not reject H0.
- **QQ-Plot:** The data follows the "perfect normality" line quite well but with slight deviations at the extremes. (Graph consistent with the Test).



CONFIDENCE INTERVAL

The mean salary is between 6.279 and 6.823 on a logarithmic scale.
With monetary conversion, we are talking about an interval between approximately $2m and $6.5m.
High salaries but with high variability.

```
mean_salary <- mean(sample_salary)
sd_salary <- sd(sample_salary)
n <- length(sample_salary)
error_std <- qt(0.975, df = n-1) * sd_salary / sqrt(n)

conf_interval <- c(mean_salary - error_log, mean_salary + error_log)
conf_interval
```

/03

HYPOTHESIS TESTING

Is the Mean Salary different from $5m? /t-test

```
        One Sample t-test

data:  sample_salary
t = -2.9426, df = 29, p-value = 0.006342
alternative hypothesis: true mean is not equal to 6.69897
95 percent confidence interval:
 6.052179 6.582601
sample estimates:
mean of x
  6.31739
```

H0: The mean **is equal** to 5,000,000.

H1: The mean is different from 5,000,000.

P-value = 0.006: Very small -> reject H0 (the sample mean of salaries is significantly different from $5m)

The confidence interval does not include 6.69897, confirming the result.

Is the Mean Salary greater than $4m? /t-test

```
        One Sample t-test

data:  sample_salary
t = -2.1953, df = 29, p-value = 0.9819
alternative hypothesis: true mean is greater than 6.60206
95 percent confidence interval:
 6.097059      Inf
sample estimates:
mean of x
  6.31739
```

H0: The mean is less than or equal to 4,000,000.

H1: The mean is greater than 4,000,000.

P-value = 0.9819: Much larger than 0.05 -> do not reject H0 (we do not have sufficient evidence to say that the mean salary is greater than $4m).

/03

LINEAR REGRESSION TEST

DOES TS% INFLUENCE SALARY?

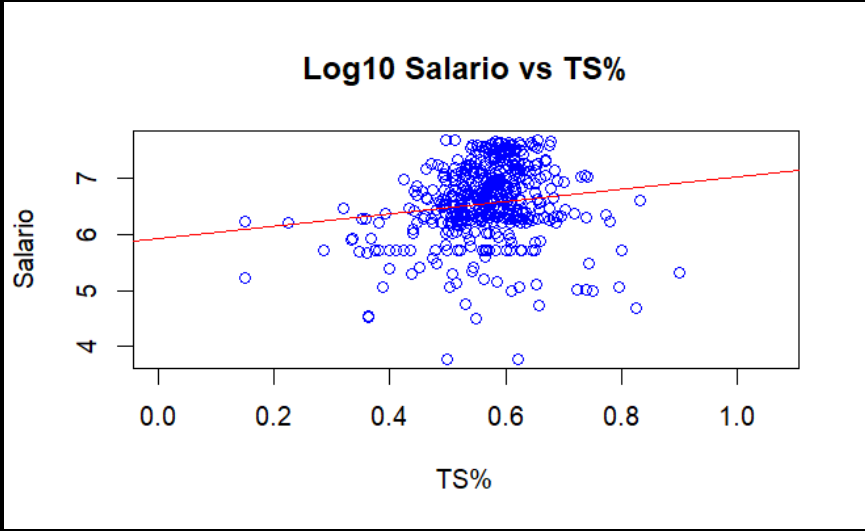
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.9320     0.1742  34.062  < 2e-16 ***
`TS%`         1.0947     0.3045   3.595 0.000359 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6615 on 464 degrees of freedom
  (1 osservazione eliminata a causa di un valore mancante)
Multiple R-squared:  0.0271,	Adjusted R-squared:  0.02501
F-statistic: 12.93 on 1 and 464 DF,  p-value: 0.0003588
```

Very low P-value: High significance.
Low R^2: Salary is influenced by TS% but only to a small extent.

If TS% increased by 10%?

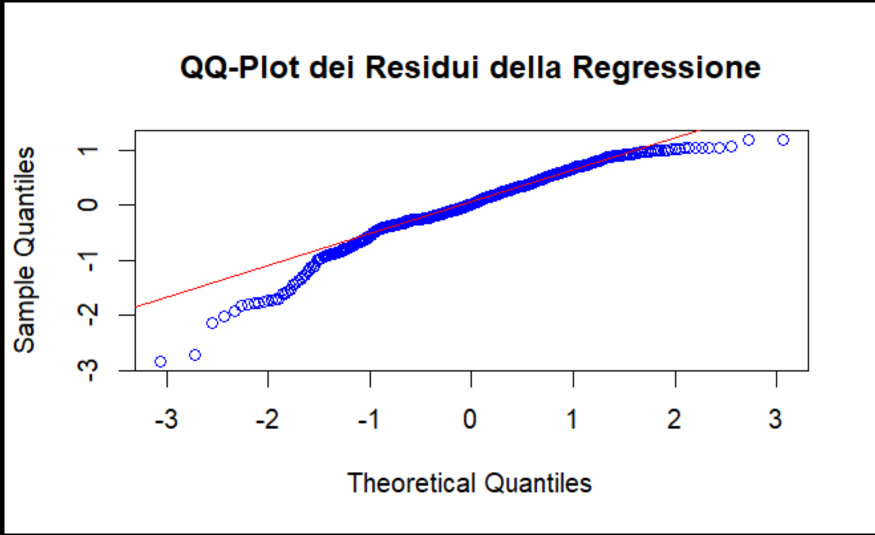
1.0947 * 0.01 = 0.10947 -> 10^0.10947 -> 1.29
The salary would therefore increase by 29%.

Assuming a player earns 5 million, with a 10% increase in TS%, the player will earn 6.45 million dollars (+29%).



REGRESSION RESIDUALS: Salary vs TS%

- The residuals are distributed along the theoretical line in the central part, indicating a good approximation to normality.
- Some slight deviations at the extremes are present, but not such that they compromise the reliability of the model.



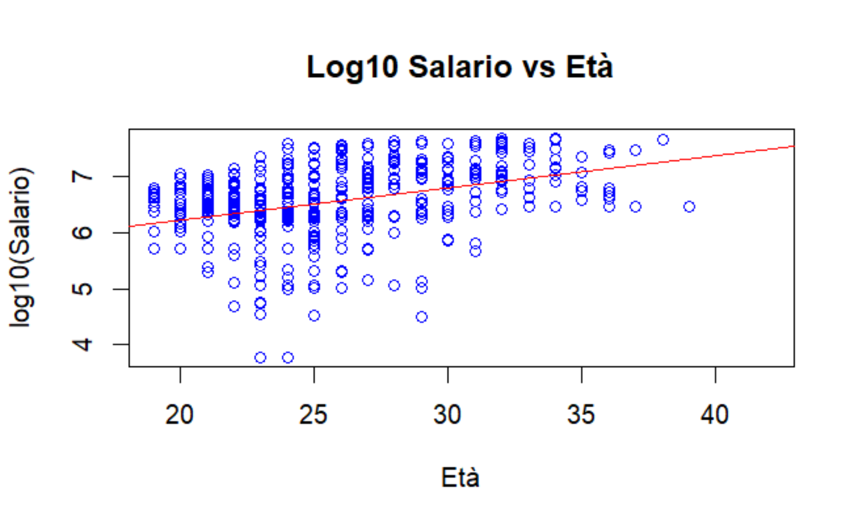
DOES AGE INFLUENCE SALARY?

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.066438   0.177423  28.556  < 2e-16 ***
Age         0.057300   0.006779   8.452 3.71e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6256 on 465 degrees of freedom
Multiple R-squared:  0.1332,	Adjusted R-squared:  0.1313
F-statistic: 71.44 on 1 and 465 DF,  p-value: 3.712e-16
```

Very low P-value: High significance.
Low R^2: Salary is influenced by age but only to a small extent.

Every year of age brings a salary increase of 0.057.
10^0,057 = 1,14 → 14%

If a player earns $5m, the following year (if still in the league), they will earn $5.7m (+14%)

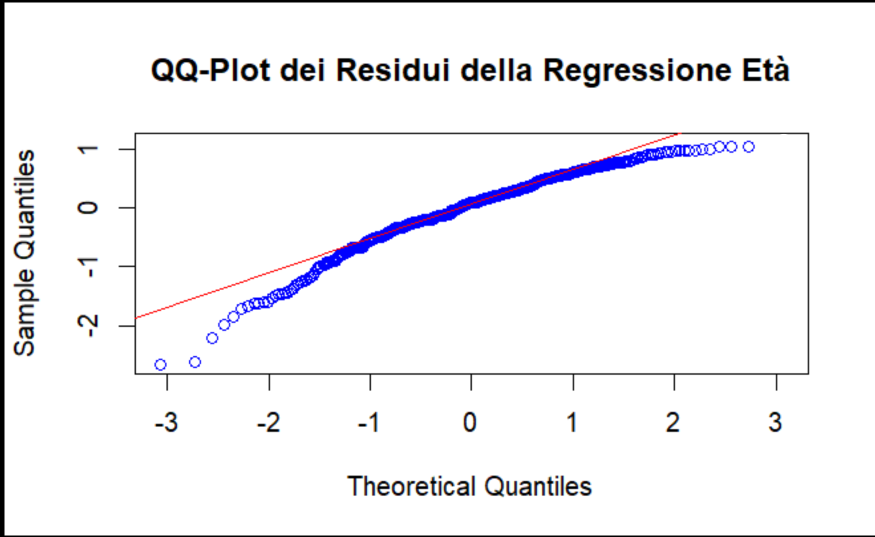


REGRESSION RESIDUALS: Salary vs Age

The QQ-Plot of residuals on the Age regression is very similar to the one seen for TS%.

The residuals follow the line quite well in the center.

At the extremes (especially for very high or very low values) there is some deviation.



THANKS FOR YOUR ATTENTION