1. Create a histogram of *Wins* and describe the distribution.

A graph of a number of different numbers

Description automatically generated with medium confidence

Right skewed, values clustered at the lower end of the x-axis.

```{r}

hist(badminton\_df$Wins)

```

1. Obtain the summary statistics of *Wins* and fill them in below.

**Mean:** 10.76 **Standard Deviation:** 15.84 **Minimum:** 0 **Maximum:** 107 **Median:** 4

```{r}

mean(badminton\_df$Wins)

sd(badminton\_df$Wins)

min(badminton\_df$Wins)

max(badminton\_df$Wins)

median(badminton\_df$Wins)

```

1. Calculate the range for *Wins* and for *pts\_for*.

*Wins* = 107 - 0 = **107** *pts\_for* = 6042 - 62 = **5980**

```{r}

max(badminton\_df$pts\_for)

min(badminton\_df$pts\_for)

```

1. The ranges for *Wins* and *pts\_for* are both quite large. What do you think is a possible reason for this flaw in the data?

Some players have played for longer than others, for example, there could be a player who only played in the 2023 World Tour whereas there could also be a player who's played in all the World Tours since 2018.

1. Create a side-by-side box plot of *Wins* for each *Category*. Describe the distribution.

The range for Doubles appears bigger than that for Singles, both seem to have a lot of outliers.

```{r}

boxplot(Wins~Category, badminton\_df)

```

1. Perform a t.test to determine if there is a significant difference in the mean number of *Wins* for *Doubles* vs. *Singles*.

**t-value:** 0.17

**p-value:** 0.87

**Conclusion:** Do Not Reject H0; 0.87 > 0.05

We do not have significant evidence that the mean *Wins* for doubles is higher than the mean *Wins* for Singles.

```{r}

t.test(Wins~Category, data = badminton\_df)

```

1. Fit a simple linear model of *pts\_for* predicting *Wins*.

```{r}

lm(Wins~ pts\_for, data = badminton\_df)

```

* 1. Interpret the slope of the model you just fit.

For every additional point a player scores, their wins increase by 0.017.

* 1. If a player has 216 points calculate their predicted *Wins*.

They would have about 3 wins.

* 1. Assess the overall fit of the model.

**t-value:** 124.73

**p-value:** <2e-16

**Conclusion:** Reject H0; <2e-16 < 0.05

We have significant evidence of a positive slope when using *pts\_for* to predict *Wins*.

```{r}

mod <- lm(Wins~pts\_for, data = badminton\_df)

summary(mod)

```

* 1. Find the 90% confidence interval of the slope of the model.

**Lower:** 0.01679255 **Upper:** 0.01724367

We are 90% confident that the slope of the model of *pts\_for* predicting *Wins* is between 0.0168 and 0.0172.

```{r}

mod <- lm(Wins~pts\_for, data = badminton\_df)

confint(mod, level=0.90)

```

1. Do you think that *pts\_for* is really the most effective predictor of *Wins*? Why or why not? What could be a better predictor?

pts\_for is an obvious predictor of Wins as the more points a player has, the more wins they will have typically. However, this could make it ineffective as a predictor of Wins as they are so closely related to one another. A better predictor could be shot\_pct or pts\_agst.