* 1. A graph of a number of bars

     Description automatically generated with medium confidenceCreate a histogram of SPEED for each distance.

```{r}

ggplot(racewalking\_df, aes(x = SPEED)) + geom\_histogram(bins = 15, colour = "black", fill = "lightblue") + facet\_wrap(~DISTANCE) + theme\_minimal()

```

* 1. Compare the distribution of SPEED for the 10k races with the distribution for the 20k races.

Both distributions are left skewed, but the 20k is more skewed than the 10k. Additionally, the counts for the 20k are higher than for the 10k.

1. Perform a t.test to determine if there is a significant difference in the mean speed for each distance.

H0: mean 10k speed = mean 20k speed Ha: mean 10k speed mean 20k speed

```{r}

t.test(SPEED~DISTANCE, var.equal=TRUE, data=racewalking\_df)

```

**t =** -0.39936 **p-value =** 0.69; 0.69 > 0.05 **Conclusion:** Do not reject H0;We do not have significant evidence that the mean speed for 10k walkers is significantly different than the mean speed for 20k walkers.

* 1. Create a boxplot comparing the mean speed for each distance and gender.

A graph with blue and black squares

Description automatically generated```{r}

ggplot(racewalking\_df, aes(y = SPEED, x = DISTANCE, fill = DISTANCE)) + geom\_boxplot(show.legend = FALSE) + facet\_wrap(~GENDER) + theme\_minimal() + scale\_fill\_brewer()

```

* 1. Based on the boxplot would you say that there is a significant difference in the mean speed for each distance and for each gender?

There looks to be a considerable difference in the mean speed for each gender and a slight difference in the mean speed for each distance.

1. Use a two-way ANOVA test to determine if there is a significant difference in the mean speed for each distance and for each gender?

Factor A: H0: mean speed 10k = mean speed 20k Ha: mean speed 10k mean speed 20k

Factor B: H0: mean speed W = mean speed M Ha: mean speed W mean speed M

```{r}

amodC <- aov(SPEED~DISTANCE+GENDER,data=racewalking\_df)

summary(amodC)

```

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Source** | **D.f.** | **Sum of Squares** | **Mean Square** | **F-value** | **P-value** |
| **DISTANCE** | 1 | 61 | 61 | 0.325 | 0.569 |
| **GENDER** | 1 | 46054 | 46054 | 245 | <2e-16 |
| **Residual** | 234 | 43985 | 188 |
| **Total** | 236 | 90100 |

Factor A Conclusion: 0.569 > 0.05; Do Not Reject H0; We do not have significant evidence of a difference in the mean speed for each distance.

Factor B Conclusion: <2e-16 < 0.05; Reject H0; We have significant evidence that the mean speed for men is significantly higher than for women.

* 1. Fit the model:

```{r}

mod\_rec <- glm(REC ~ DISTANCE + PODIUM + SPEED + GENDER, family = "binomial", data = racewalking\_df)

mod\_rec |> tidy()

```

* 1. A female athlete walked in the 10k with a podium finish and a speed of 221 meters per minute. Calculate the log(odds), odds, and predicted odds of her making a record.
     1. log(odds):
     2. odds:
     3. :

1. Perform a g-test to assess the overall fit of the multiple logistic regression model:

H0: Ha:

```{r}

record\_mod <- glm(REC ~ DISTANCE + PODIUM + SPEED + GENDER, family = "binomial", data = racewalking\_df)

record\_mod |> summary()

g\_stat <-319.17 - 291.97

# p-value)

1-pchisq(g\_stat,4)

```

**G-stat:** **null deviance – residual deviance** = 319.17-291.97 = 27.2

**P-value:** 1.811123e-05 **Conclusion:** Reject H0; We have significant evidence that DISTANCE, PODIUM, SPEED, and GENDER are effective predictors of whether or not an athlete will make a record.

1. After performing all of these tests on different models, are there any noticeable differences in the different races the athletes competed in? Are there any discernable patterns with records?

It seems that the main difference for walkers’ average speed in events is based on gender whereas distance does not seem to have as big of an impact on their speed. However, with records it does seem that both distance and gender factor in.