Validating the *FireEdge* Assessment

Robert W. Szarek

March 10, 2021

1. Overview and Context

This document was created as an R Markdown file from within R-Studio. The benefits of using an R Markdown file for scientific research is the ability to weave live code and theory into a single document, all the while providing the reader a clear look into the analyses conducted to reach the various conclusions outlined in this study. Such reproducible analyses help make scientific studies more transparent and accessible to all stakeholders.

This research study walks the user through a validation study to empirically demonstrate a positive relationship between scores on the *FireEdge* test with ratings of job performance. The benefits of a successful study include:

- Marketing *FireEdge* as a proven screening device to select talent
- Providing legal defensibility for using the *FireEdge*
- Helping minimize bias and adverse impact in observed results

in connection with the *FireEdge* test, which is a cognitive and situation-based judgment test developed for the purposes of hiring entry-level firefighters. The analyses contained in this technical document are post-exam development, where we have a finalized test and overall test score on a number of participants who volunteered to participate in the research study.

The purposes of this research study was to empirically demonstrate (or “validate”) that the *FireEdge* test is able to predict on-the-job performance of firefighters. Three statistical analyses were conducted in this study:

- Analysis of Variance (ANOVA)
- Ordinary Least Squares (OLS) Linear Regression
- Pearson Product Moment Correlation with T-test

To kick the study off, we first load the tidyverse package to ease the manipulation of data that is required to run the analyses outlined above.

```
library(tidyverse)
```

2. Data Import

Once the tidyverse has been loaded, we can proceed to import the cleaned and scored dataset that contains the necessary variables required for conducting this research study. The code below imports the data table into our local R environment.

```
val_data <- read_rds(file = "~/R-lang/FireEdge/data/FireEdge_data.Rds")
```

The original dataset includes item-level data, dimension-level data in conjunction with various other information concerning the criteria. As none of this is required to conduct the validation study, we focus on the overall composite score with the criteria of job performance, and drop all remaining variables from the data table.

```
val_data <-
  val_data %>%
  select(
    Study,
    Gender,
    Race,
    Final_Score,
    Crit_Supv
  ) %>%
  rename(
    "JobPerformance" = Crit_Supv,
    "FireEdge_Score" = Final_Score
  ) %>%
  mutate(Study = as_factor(Study))

str(val_data)
```

```
## tibble [339 × 5] (S3: tbl_df/tbl/data.frame)
##  $ Study         : Factor w/ 3 levels "Maryland","Colorado",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Gender        : Factor w/ 2 levels "Male","Female": 1 2 2 2 1 2 2 1 2 2 ...
##  $ Race          : Factor w/ 7 levels "Black","Asian",..: 6 1 1 1 6 3 6 1 6 1 ...
##  $ FireEdge_Score: num [1:339, 1] 66.6 79.5 82.8 82.1 75.9 ...
##  $ JobPerformance: num [1:339] 45 NA 40 47 44 48 49 48 45 46 ...
```

In the condensed data table, we have the following variables:

- Study is the location of the agency participating in the validation study.
- Gender is the gender identity of the study participant.
- Race is the race of the study participant.
- FireEdge_Score is the final score on the FireEdge test.
- JobPerformance is the supervisory rating of job performance.

3. Descriptive Statistics

Next, we begin to investigate the descriptive statistics concerning the supervisor ratings of job performance. The code below creates a table of information broken down by Study location, such as the average job performance score, the standard deviation of the distribution, and the minimum and maximum scores.

```
val_data %>%
  group_by(Study) %>%
  summarise(
    .groups = "drop",
    RawTotal = n(),
    Missing = sum(is.na(JobPerformance)),
    True_Sample = RawTotal - Missing,
    Mean = mean(JobPerformance, na.rm = TRUE),
    SD = sd(JobPerformance, na.rm = TRUE),
    Min = min(JobPerformance, na.rm = TRUE),
    Max = max(JobPerformance, na.rm = TRUE)
  )
```

```
## # A tibble: 3 x 8
##   Study    RawTotal Missing True_Sample  Mean    SD   Min   Max
##   <fct>       <int>   <int>       <int> <dbl> <dbl> <dbl> <dbl>
## 1 Maryland       65       2          63  44.4  5.85    25    55
## 2 Colorado      205      13         192  45.4  6.30    26    62
## 3 Alabama        69      27          42  43.5  5.24    30    57
```

We see that the agency in Alabama had a total of 27 missing cases of performance ratings, while the Colorado agency had a total of 13 missing cases. The average job performance ratings are all consistent across the three study locations, as are the standard deviations, minimum and maximum scores. Next, we perform the same type of analysis for the *FireEdge* score.

```
val_data %>%
  group_by(Study) %>%
  summarise(
    .groups = "drop",
    RawTotal = n(),
    Missing = sum(is.na(FireEdge_Score)),
    True_Sample = RawTotal - Missing,
    Mean = mean(FireEdge_Score, na.rm = TRUE),
    SD = sd(FireEdge_Score, na.rm = TRUE),
    Min = min(FireEdge_Score, na.rm = TRUE),
    Max = max(FireEdge_Score, na.rm = TRUE)
  )
```

```
## # A tibble: 3 x 8
##   Study    RawTotal Missing True_Sample  Mean    SD   Min   Max
##   <fct>       <int>   <int>       <int> <dbl> <dbl> <dbl> <dbl>
## 1 Maryland       65       0          65  75.2  5.56  59.8  84.1
## 2 Colorado      205       0         205  77.6  5.71  49.4  85.6
## 3 Alabama        69       0          69  72.7  8.95  43.8  85.4
```

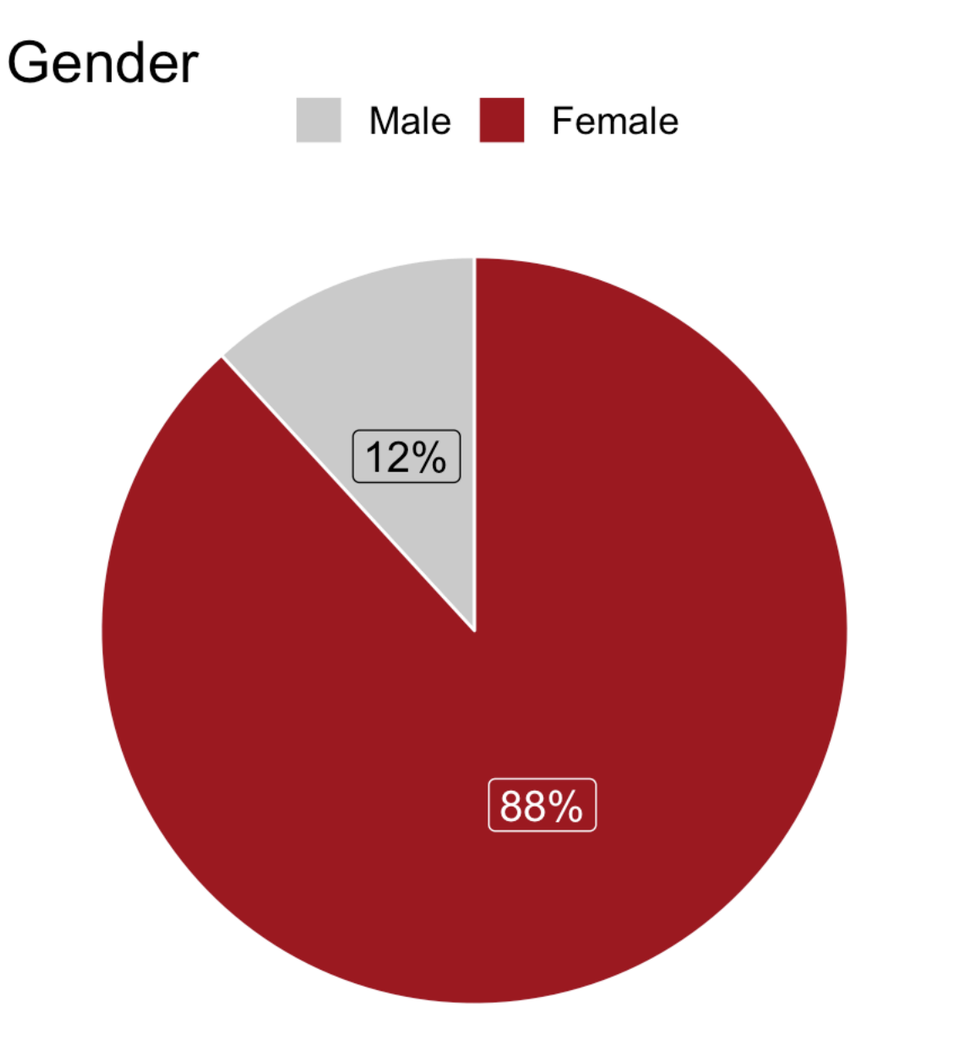
The means all fall in line quite closely, with an average score of around a 75.00 and a standard deviation around 6.5. Minimum and maximum scores were all quite consistent across the three studies. Once we have taken a look at these statistics, we create our finalized dataset by removing the missing data that we observed in the job performance ratings with the following code.

```
val_data <-
val_data %>%
  filter(!is.na(JobPerformance))
```

We remove the missing data points and determine we have a total of 297 rows of data to use for the validation study. Next, we create a series of charts to determine the number of participants based on Race, Gender and Study.

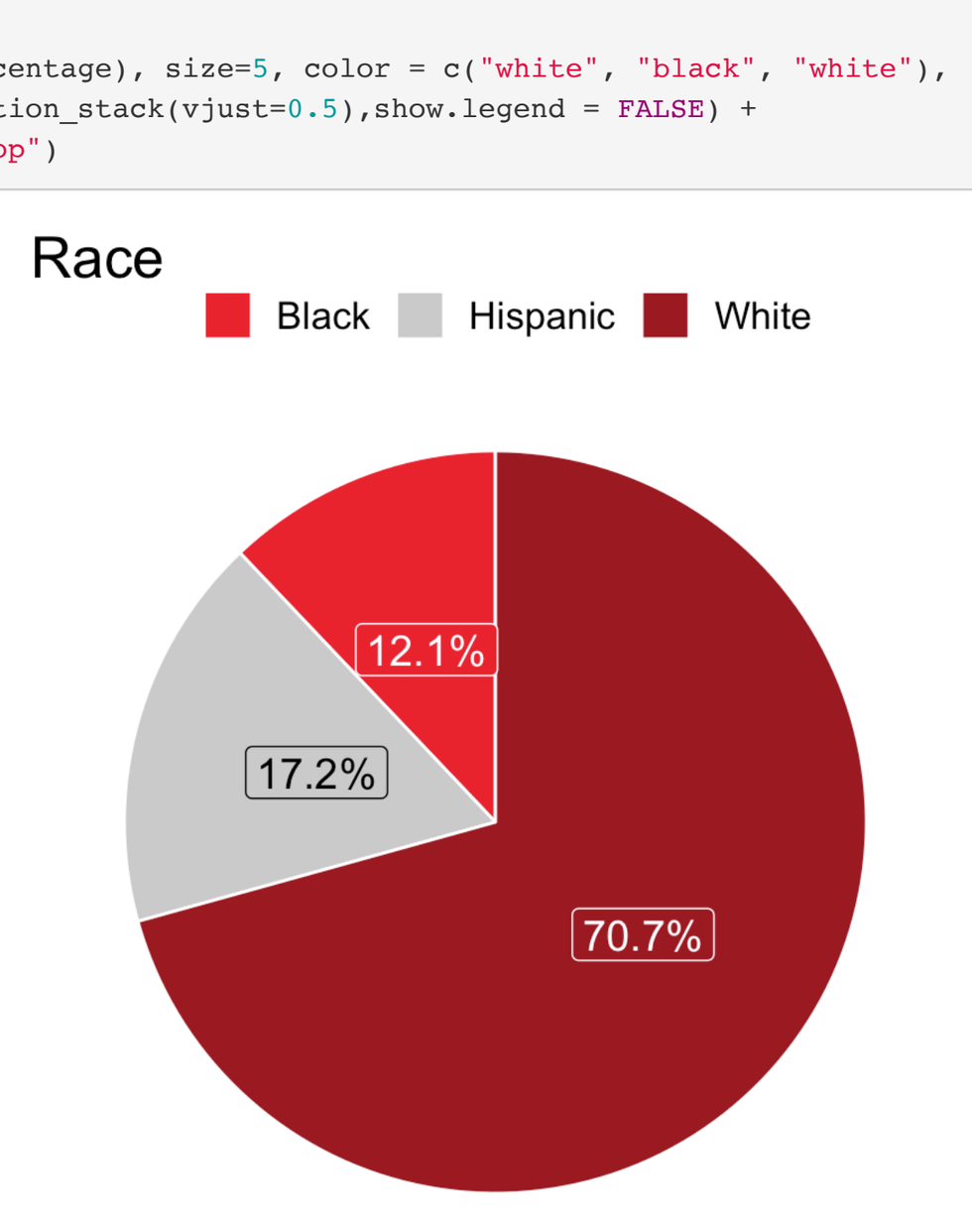
The first pie chart presents counts of Gender represented in the validation study. Unfortunately, a general trend in firefighting is that males usually outweigh females by a considerable portion. This is reflected in the sample size we have obtained where there are significantly more males than females. Such large degrees of difference are therefore not uncommon and care must be taken to ensure the study does attempt to maximize female participation rates to the highest degree possible.

```
val_data %>%
  count(Gender) %>%
  filter(Gender == "Male" | Gender == "Female") %>%
  mutate(Percentage = scales::percent(n / sum(n))) %>%
  ggplot(aes(x = "", y = n, fill = Gender)) +
  geom_bar(stat = "identity",  color = "white", show.legend = TRUE) +
  coord_polar("y") +
  theme_void(base_size = 16) +
  scale_fill_manual(name= "", values = c("#CCCCCC", "#9E1C24")) +
    labs(x = "",
         y = "",
         title = "Gender") +
  geom_label(aes(label = Percentage), size=5, color = c("black", "white"),
             position = position_stack(vjust=0.5),show.legend = FALSE) +
  theme(legend.position = "top")
```



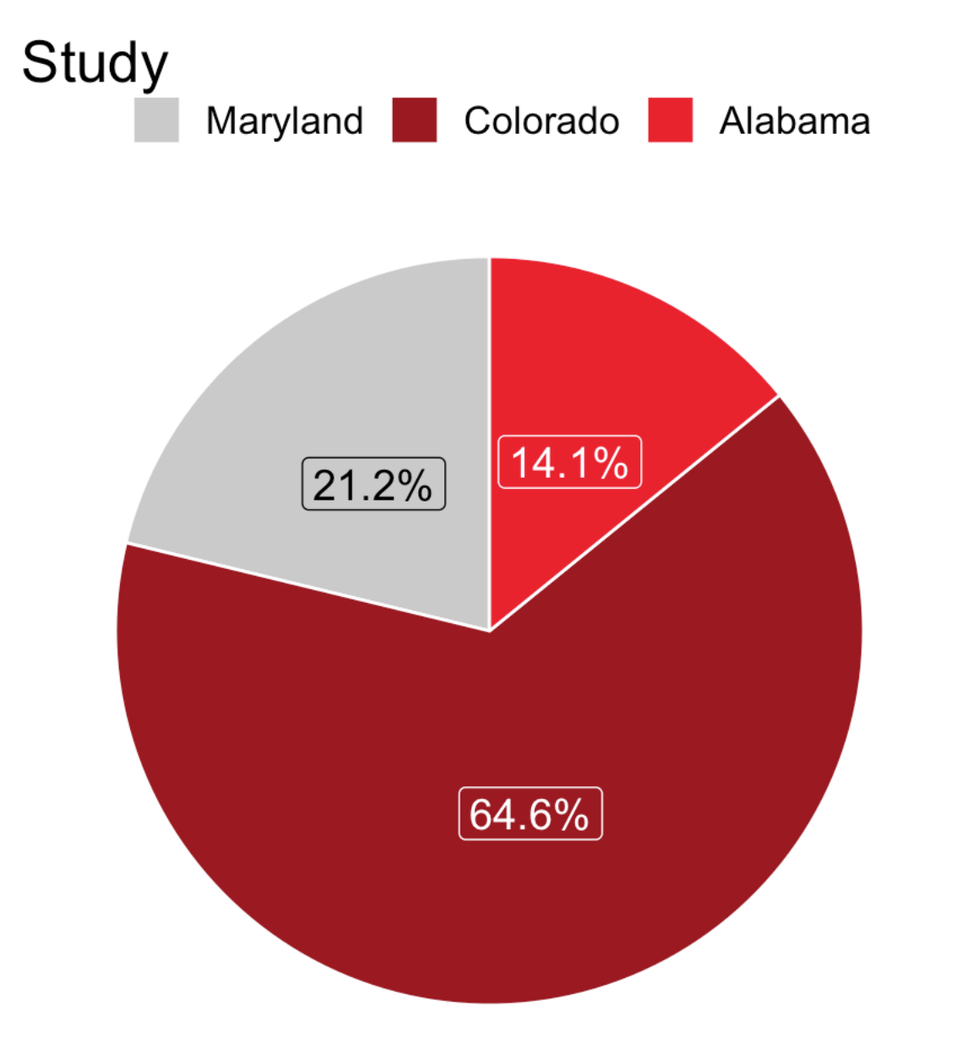
The second pie chart counts the Race breakdown for participants. We can see that minority representation rates are not as high as would be ideal, which would have been in the 20-25% range. However, we still have a robust representation of minority candidates for the study to ensure the sample represents the population of firefighters in general. The original stratified sampling plan sought to ensure minorities are equally represented within the sampling plan.

```
val_data %>%
  count(Race) %>%
  filter(Race == "Hispanic" | Race == "White" | Race == "Black") %>%
  mutate(Percentage = scales::percent(n / sum(n))) %>%
  ggplot(aes(x = "", y = n, fill = Race)) +
  geom_bar(stat = "identity",  color = "white", show.legend = TRUE) +
  coord_polar("y") +
  theme_void(base_size = 16) +
  scale_fill_manual(name= "", values = c( "#EA1D2E", "#CCCCCC", "#9E1C24")) +
    labs(x = "",
         y = "",
         title = "Race") +
  geom_label(aes(label = Percentage), size=5, color = c("white", "black", "white"),
             position = position_stack(vjust=0.5),show.legend = FALSE) +
  theme(legend.position = "top")
```



The final chart outlines the location from which participants were pulled. The majority of participants derived from a firefighting agency located in Colorado, with 64.6% of the total sample. Maryland was next, with 21.2% and Alabama was last, with 14.1%. As we do not expect the firefighting job to differ significantly across the nation, for the purposes of streamlining this validation study we collapsed all three agencies into a single study.

```
val_data %>%
  count(Study) %>%
  mutate(Percentage = scales::percent(n / sum(n))) %>%
  ggplot(aes(x = "", y = n, fill = Study)) +
  geom_bar(stat = "identity",  color = "white", show.legend = TRUE) +
  coord_polar("y") +
  theme_void(base_size = 16) +
  scale_fill_manual(name= "", values = c("#CCCCCC", "#9E1C24", "#EA1D2E")) +
    labs(x = "",
         y = "",
         title = "Study") +
  geom_label(aes(label = Percentage), size=5, color = c("black", "white", "white"),
             position = position_stack(vjust=0.5),show.legend = FALSE) +
  theme(legend.position = "top")
```



In summary, the sample size obtained for the purposes of this validation study were found to be representative of the population in general. We therefore move on to conducting a correlation and linear regression modeling.

4. Pearson Correlation Analysis

A correlation coefficient seeks to establish a relationship between two variables. In this case, we seek to measure the level of association between the *FireEdge* test score and job performance ratings. A positive correlation coefficient would indicate that as one variable increases, so does the other. A negative correlation coefficient would indicate that if one variable increases in value, the other variable decreases. In this particular example, a negative relationship would mean that as test scores go up, job performance goes down, a result we definitely would not like to see! We therefore hypothesize that our correlation will yield a significant, positive association between the two variables.

THe results of the analysis are presented below. Specifically, we obtained a correlation coefficient of .273. We then conducted a Student’s T-test on the value to test whether the obtained correlation is significantly different than zero. The results of the t-test were very significant, $t(295) = 4.884$, $p < .001$. This tells us that the calculated correlation, or relationship between the two variables is significantly different than zero, which helps us gain certainty around the relationship between *FireEdge* test scores and job performance.

```
cor.test(x = val_data$FireEdge_Score,
         y = val_data$JobPerformance,
         exact = TRUE,
         method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  val_data$FireEdge_Score and val_data$JobPerformance
## t = 4.8837, df = 295, p-value = 1.71e-06
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1648158 0.3756179
## sample estimates:
##       cor
## 0.2734977
```

We conclude with the t-test that the observed correlation does empirically demonstrate a significant relationship between test scores and job performance. The linear regression model will further provide a visual means of investigating the degree of this relationship.

5. Linear Regression Model

A linear regression model seeks to summarize the relationship between two variables with a single line of best fit. This line of best fit seeks to minimize the error between an actual observed value and the predicted value. Furthermore, such a relationship could then be used to infer somebody’s job performance rating given their score on the *FireEdge* test. The code below runs the linear regression where the *FireEdge* test score is the predictor, or, independent variable. The dependent variable, or outome variable, is job performance. We seek to predict job performance from test scores.

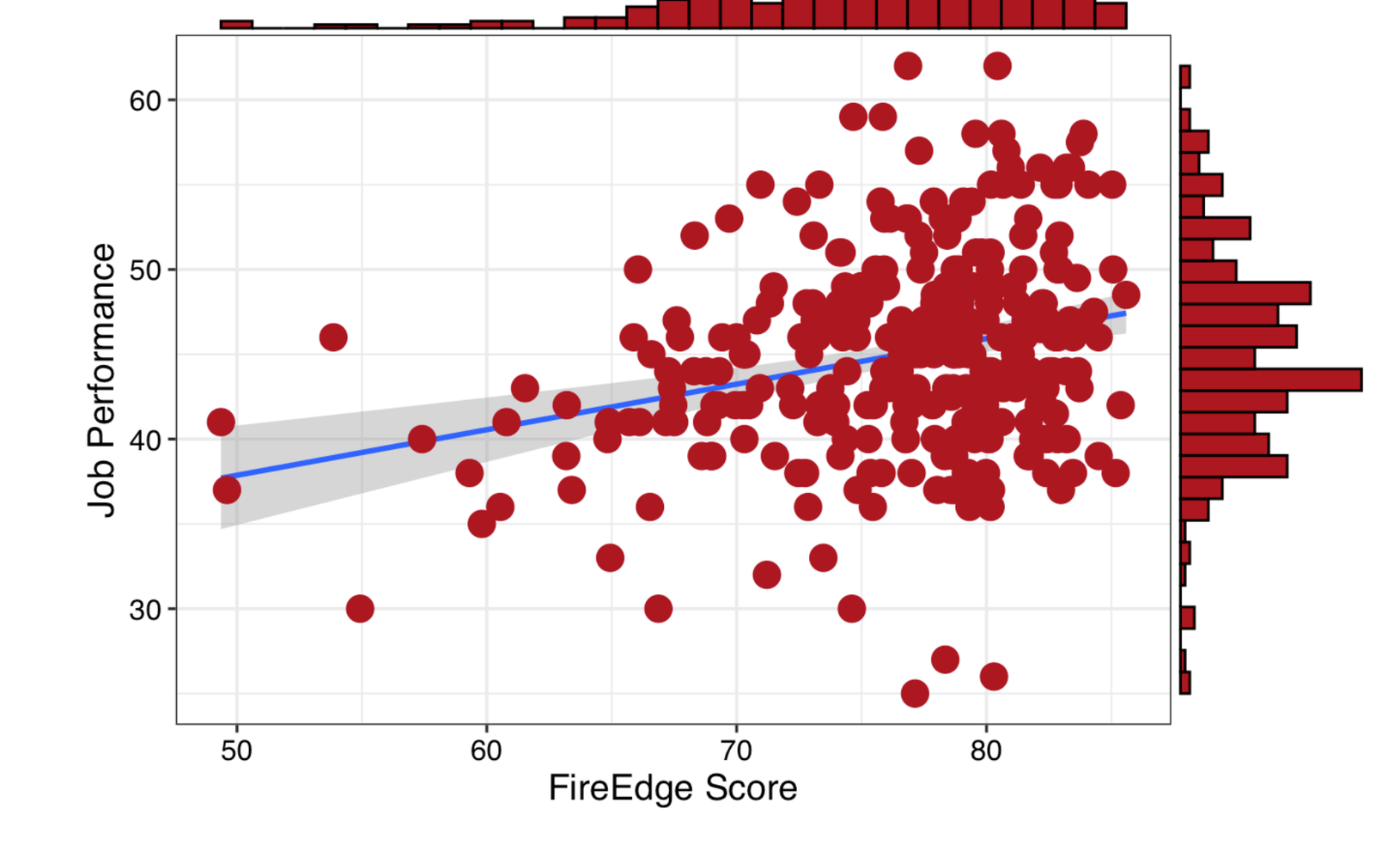
```
summary(
lm(JobPerformance ~ FireEdge_Score,
   data = val_data)
)
```

```
##
## Call:
## lm(formula = JobPerformance ~ FireEdge_Score, data = val_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.1572  -3.4707  -0.2692   3.4955  16.9180
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     24.4378     4.2149   5.798 1.72e-08 ***
## FireEdge_Score   0.2686     0.0550   4.884 1.71e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.865 on 295 degrees of freedom
## Multiple R-squared:  0.0748, Adjusted R-squared:  0.07166
## F-statistic: 23.85 on 1 and 295 DF,  p-value: 1.71e-06
```

The linear regression model looks promising with the results from the model. Including the *FireEdge* score as a predictor had a significant t-test value of $t = 4.884$, $p < .000$. The multiple r-squared was unfortunately not that high, with only .07 of the observed variance being explained by the predictor. The overall F-test was additionally significant, indicating that the model as a whole was significant at $p < .000$. The visualization below shows the extent of the relationship in addition with the relevant histograms for each of the two variables drawn in the margins. We see that the *FireEdge* test scores are negatively skewed, with a significant portion of scores on the higher end of the scale. The job performance ratings followed more of a normal distribution.

```
lin_reg_graph <-
ggplot(val_data, aes(FireEdge_Score, JobPerformance)) +
  geom_point(size = 4, color = "black") +
  geom_smooth(method = "lm", se = TRUE) +
  labs(y = "Job Performance",
       x = "FireEdge Score") +
  geom_point(size = 4.5, color = "#B11822") +
  theme_bw(base_family = "sans", base_size = 12) +
  theme(text = element_text(size = 14),
        axis.text = element_text(colour = "black"))
lin_reg_graph_final <-
  ggExtra::ggMarginal(lin_reg_graph,
                      type = "histogram",
                      fill = "#B11822",
                      size = 5)

lin_reg_graph_final
```



6. Analysis of Variance by Race

Our final analysis seeks to determine if participants who identified across different ethnicities performed significantly better or worse than any other class. Ideally, we’d like to ensure our test functions similarly across Race, Gender, etc. If there are differences in the test by Race, we could potentially have to investigate further in order to determine which component of the examination is driving the differences.

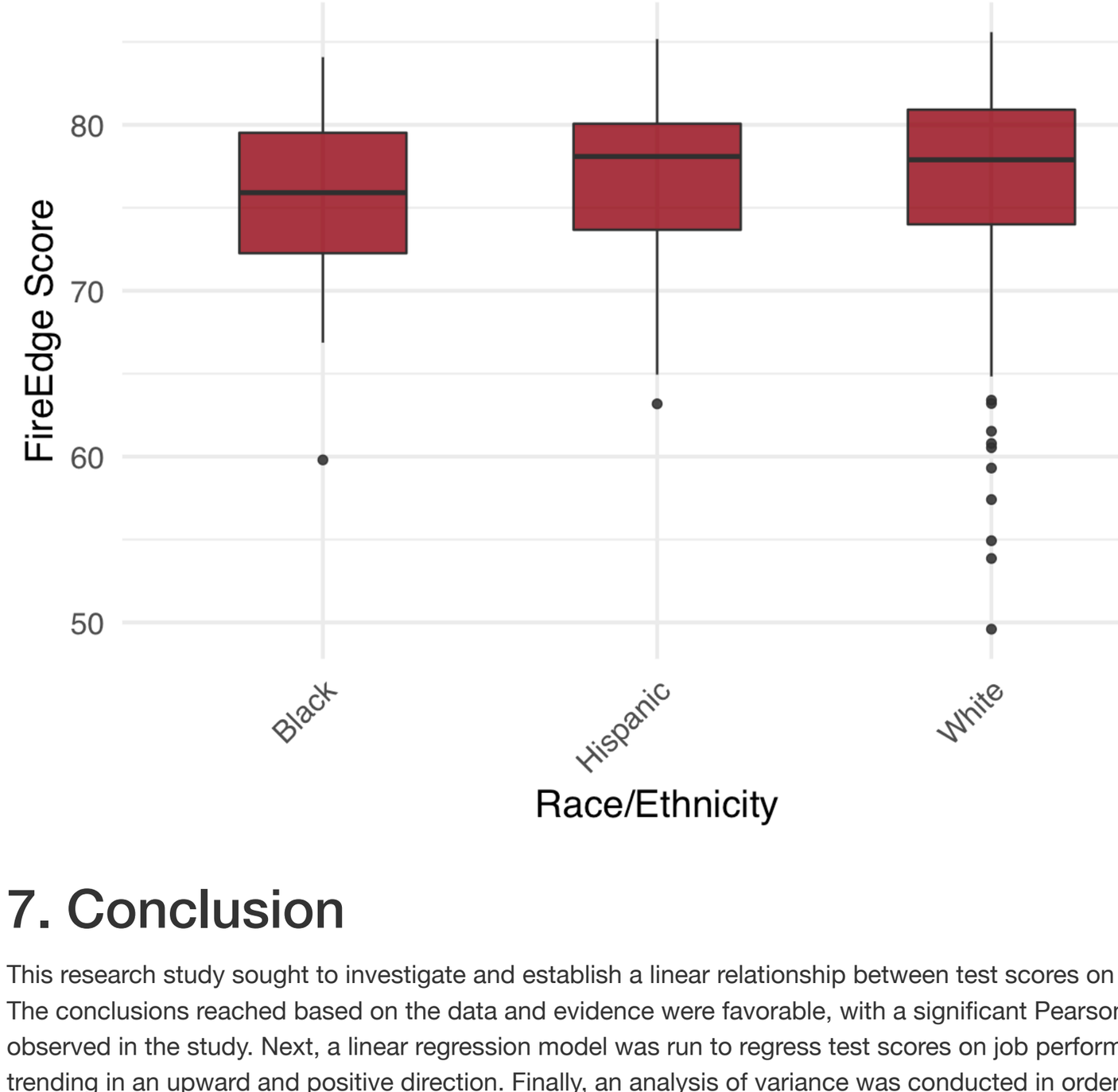
```
summary(
aov(FireEdge_Score ~ Race, data = val_data)
)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Race          4     95   23.80   0.663  0.618
## Residuals   274   9828   35.87
## 18 observations deleted due to missingness
```

The Analysis of Variance, or, ANOVA for short, measures the variance across the different racial groups on *FireEdge* test scores. The resulting test statistic, known as the omnibus F-ratio, was non-significant in this study: $F(4, 274) = 0.663$, $p > .618$. The p value would need to be less than .05 in order for us to reject the null hypothesis and conclude that significant differences exist across Race on *FireEdge* test scores.

Finally, an easy-to-explain way to visually demonstrate the disparity between means on a continuous variable is a boxplot chart. Such a chart places the score value on the y-axis, and the various levels of racial identity on the x-axis. The degree of height difference between the various boxes show the levels of difference.

```
val_data %>%
  filter(Race == "Black" | Race == "White" | Race == "Hispanic") %>%
  ggplot(aes(x=Race, y=FireEdge_Score)) +
  geom_boxplot(fill = "#9E1C24", width= 0.5, alpha = 0.9) +
  theme_minimal(base_size = 16,base_family = "sans") +
    labs(x = "Race/Ethnicity",
         y = "FireEdge Score") +
  theme(legend.position = "none", axis.text.x = element_text(angle = 45, hjust = 1))
```



7. Conclusion

This research study sought to investigate and establish a linear relationship between test scores on the *FireEdge* with ratings of job performance. The conclusions reached based on the data and evidence were favorable, with a significant Pearson product moment correlation ($r = .273$) being observed in the study. Next, a linear regression model was run to regress test scores on job performance, which resulted in the line of best fit trending in an upward and positive direction. Finally, an analysis of variance was conducted in order to ascertain no bias existed in test scores across different Ethnicities.

In summary, the conclusions reached in this study were significant, providing empirical evidence that the *FireEdge* does indeed predict job performance scores. Such information can be used to better inform the clients who use our products.