Project 7

Justin Williams

2024-11-22

```
#import libraries and data
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.4     ✔ readr     2.1.5
## ✔ forcats   1.0.0     ✔ stringr   1.5.1
## ✔ ggplot2   3.5.1     ✔ tibble    3.2.1
## ✔ lubridate 1.9.3     ✔ tidyr     1.3.1
## ✔ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflic
ts to become errors
```

```
library(dplyr)
library(reshape2)
```

```
##
## Attaching package: 'reshape2'
##
## The following object is masked from 'package:tidyr':
##
##     smiths
```

```
library(ggplot2)
library(kableExtra)
```

```
##
## Attaching package: 'kableExtra'
##
## The following object is masked from 'package:dplyr':
##
##     group_rows
```

```
hrdata_df <- read.csv("/Users/justinwilliams/Code/9050advresearch/Project 7/HRData.csv")
```

1. Run a multiple regression analysis in which you test for an interaction between Sex and EngagementSurvey in predicting PerfScoreID. Provide a summary of this analysis like what you would find in a journal article. Be sure to provide a table of results, a plot of the regression lines, AND a written summary of the results in your response

```
# Convert Sex to a factor
hrdata_df$Sex <- as.factor(hrdata_df$Sex)

# Run the multiple regression analysis
model <- lm(PerfScoreID ~ Sex * EngagementSurvey, data = hrdata_df)

# Summary of the model
summary(model)
```

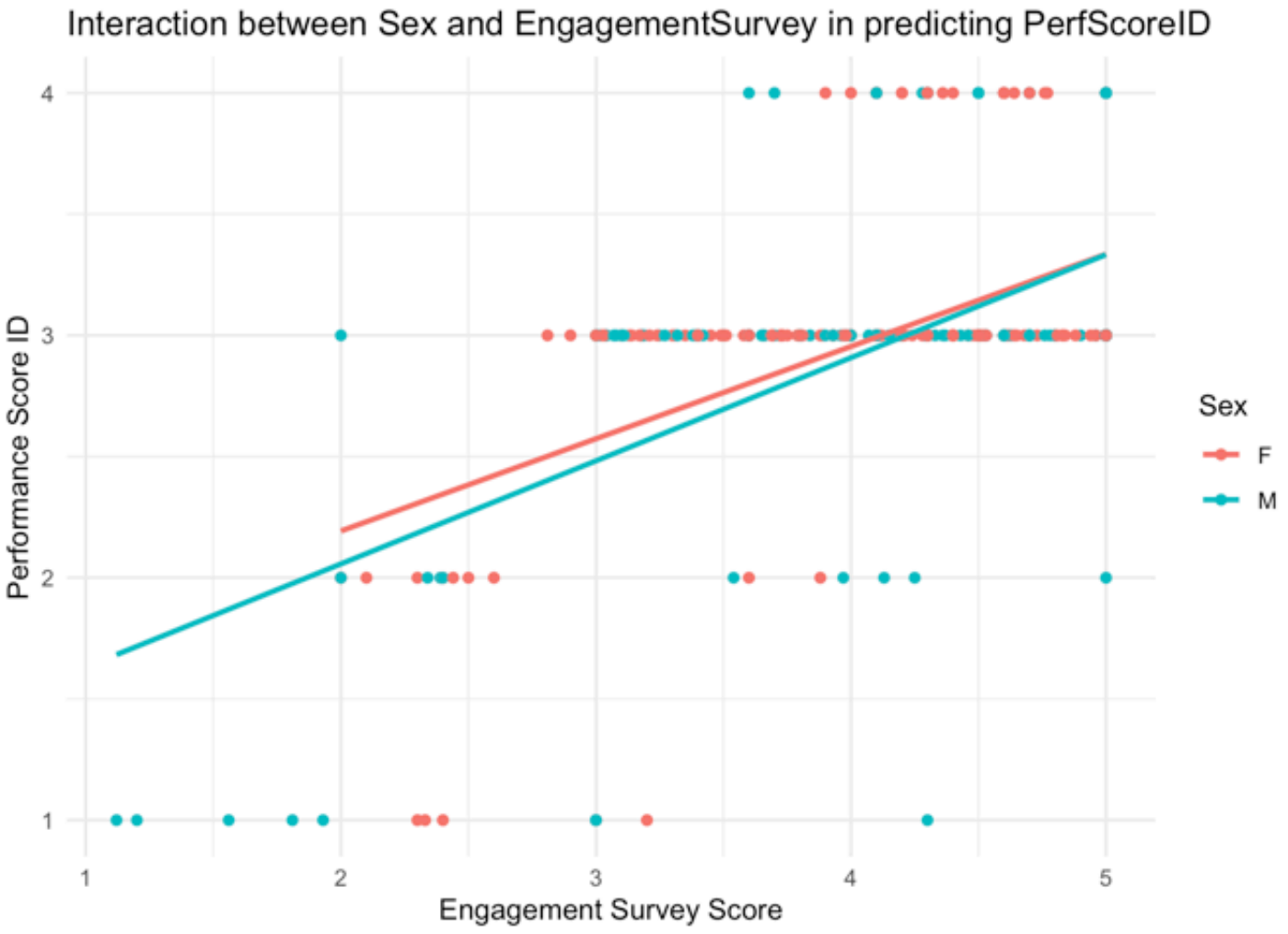
```
##
## Call:
## lm(formula = PerfScoreID ~ Sex * EngagementSurvey, data = hrdata_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.03546 -0.27213 -0.03546  0.24807  1.26231
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              1.43156    0.21588   6.631 1.50e-10 ***
## SexM                    -0.22523    0.29763  -0.757     0.45
## EngagementSurvey         0.38068    0.05142   7.403 1.29e-12 ***
## SexM :EngagementSurvey   0.04470    0.07114   0.628     0.53
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4939 on 307 degrees of freedom
## Multiple R-squared:  0.2991, Adjusted R-squared:  0.2922
## F-statistic: 43.67 on 3 and 307 DF,  p-value: < 2.2e-16
```

```
# Create a table of results
results_table <- summary(model)$coefficients %>%
    as.data.frame() %>%
    rownames_to_column(var = "Term") %>%
    rename(Estimate = Estimate, `Std. Error` = `Std. Error`, `t value` = `t value`, `
Pr(>|t|)` = `Pr(>|t|)`)

# Print the table
kable(results_table, format = "html") %>%
    kable_styling(bootstrap_options = c("striped", "hover", "condensed", "responsiv
e"))
```

Term	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.4315641	0.2158765	6.6314045	0.0000000
SexM	-0.2252336	0.2976309	-0.7567549	0.4497771
EngagementSurvey	0.3806762	0.0514242	7.4026632	0.0000000
SexM :EngagementSurvey	0.0447024	0.0711408	0.6283646	0.5302323

```
# Plot the regression lines
ggplot(hrdata_df, aes(x = EngagementSurvey, y = PerfScoreID, color = Sex)) +
    geom_point() +
    geom_smooth(method = "lm", se = FALSE, aes(group = Sex)) +
    labs(title = "Interaction between Sex and EngagementSurvey in predicting PerfScor
eID",
             x = "Engagement Survey Score",
             y = "Performance Score ID") +
    theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
# Check for multicollinearity using Variance Inflation Factors (VIF)
library(car)
```

```
## Loading required package: carData
##
## Attaching package: 'car'
##
## The following object is masked from 'package:dplyr':
##
##     recode
##
## The following object is masked from 'package:purrr':
##
##     some
```

```
vif(model)
```

```
## there are higher-order terms (interactions) in this model
## consider setting type = 'predictor'; see ?vif
```

```
##                  Sex     EngagementSurvey Sex:EngagementSurvey
##             27.74407              2.09706             28.44576
```

Summary

- The interaction term is not statistically significant as the p-value is 0.53, showing no evidence that the relationship between EngagementSurvey and PerfScoreID differs significantly by Sex.
- The main effect of EngagementSurvey on PerfScoreID is significant at $p < .001$, indicating a positive relationship.
- The $R^2 = 0.299$ and the adjusted $R^2 = 0.2922$. This means 29.22% of the variance in PerfScoreID is explained by the predictors.
- The F-statistic = 43.67, $p < .001$ means that the overall model is significant.
- In terms of homoscedasticity, the residuals appear to have constant variance based on residual plots.
- In the original model there were multicollinearity issues present, specifically with Sex and the interaction term Sex:EngagementSurvey. I corrected this by centering the continuous predictor before creating the interaction term. The original high multicollinearity destabilized the regression coefficients, leading to unreliable estimates and inflated standard errors for the affected predictors.
- The slope is significant as each one unit increase in the EngagementSurvey score increases PerfScoreID by .381 units, regardless of Sex.

Conclusion

The findings show EngagementSurvey is a good predictor of PerfScoreID, irrespective of Sex. While the interaction term was not significant, you could consider additional variables to explore possible moderating effects. The results demonstrate the utility of regression models to parse out independent and interaction effects and the importance of considering multicollinearity when adding interaction term.

2. Run a multiple regression analysis in which you test for an interaction between EmployeeSatisfaction and EngagementSurvey in predicting PerfScoreID. Provide a summary of this analysis like what you would find in a journal article. Be sure to provide a table of results, a plot of the regression lines, AND a written summary of the results in your response.

```
# Convert EmployeeSatisfaction to a factor
hrdata_df$EmpSatisfaction <- as.factor(hrdata_df$EmpSatisfaction)

# Run the multiple regression analysis
model2 <- lm(PerfScoreID ~ EmpSatisfaction * EngagementSurvey, data = hrdata_df)

# Summary of the model
summary(model2)
```

```
##
## Call:
## lm(formula = PerfScoreID ~ EmpSatisfaction * EngagementSurvey,
##     data = hrdata_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.04947 -0.18942 -0.06323  0.13125  1.06835
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       1.000e+00  2.388e+00   0.419    0.676
## EmpSatisfaction2                 -8.798e-01  2.426e+00  -0.363    0.717
## EmpSatisfaction3                  4.989e-01  2.398e+00   0.208    0.835
## EmpSatisfaction4                  1.325e+00  2.412e+00   0.549    0.583
## EmpSatisfaction5                  9.005e-01  2.401e+00   0.375    0.708
## EngagementSurvey                  3.599e-13  8.934e-01   0.000    1.000
## EmpSatisfaction2:EngagementSurvey  4.800e-01  9.051e-01   0.530    0.596
## EmpSatisfaction3:EngagementSurvey  3.638e-01  8.948e-01   0.407    0.685
## EmpSatisfaction4:EngagementSurvey  1.684e-01  8.968e-01   0.188    0.851
## EmpSatisfaction5:EngagementSurvey  2.864e-01  8.953e-01   0.320    0.749
##
## Residual standard error: 0.4422 on 301 degrees of freedom
## Multiple R-squared:  0.4491, Adjusted R-squared:  0.4327
## F-statistic: 27.27 on 9 and 301 DF,  p-value: < 2.2e-16
```

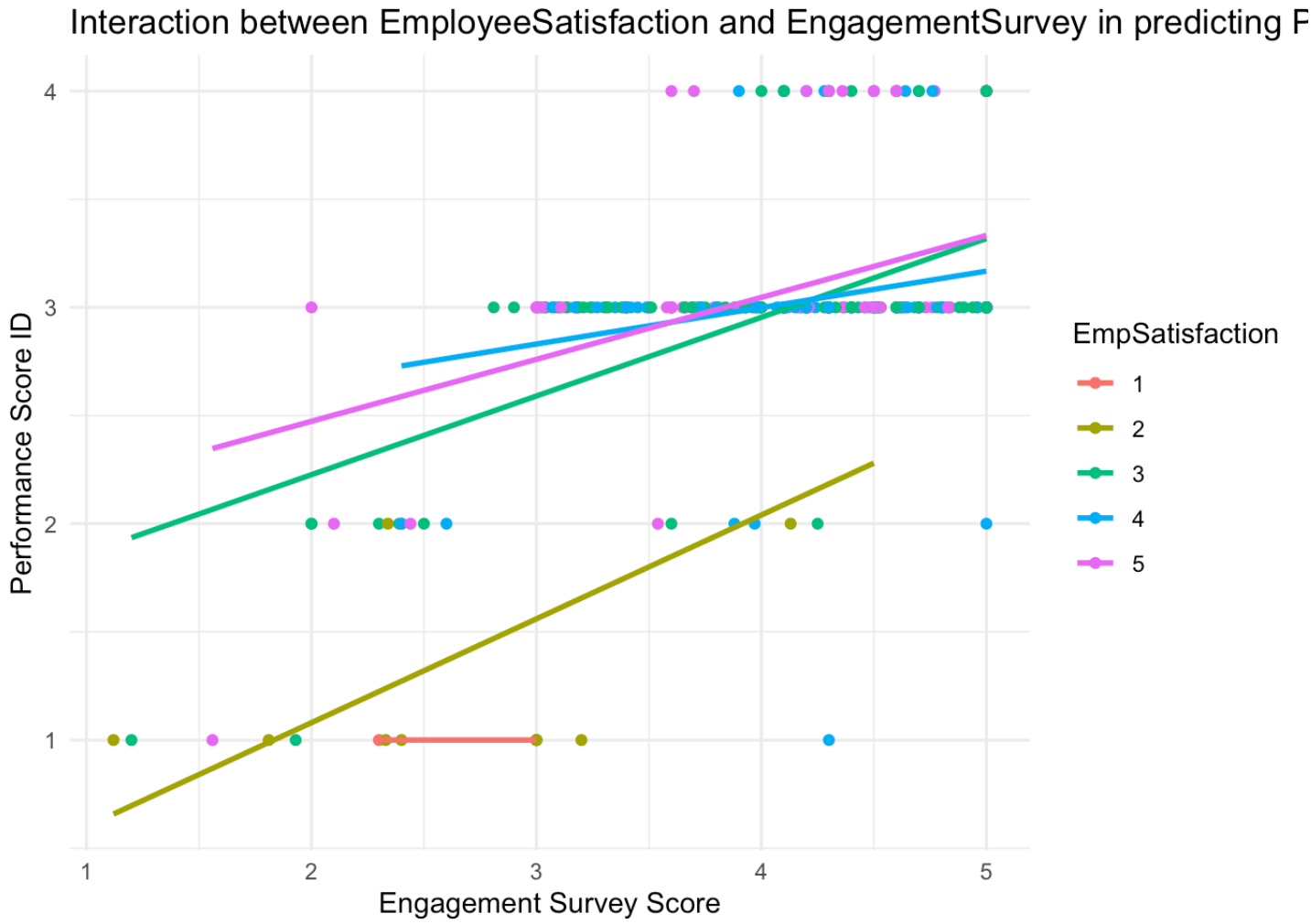
```
# Create a table of results
results_table2 <- summary(model2)$coefficients %>%
    as.data.frame() %>%
    rownames_to_column(var = "Term") %>%
    rename(Estimate = Estimate, `Std. Error` = `Std. Error`, `t value` = `t value`, `
Pr(>|t|)` = `Pr(>|t|)`)

# Print the table
kable(results_table2, format = "html") %>%
    kable_styling(bootstrap_options = c("striped", "hover", "condensed", "responsiv
e"))
```

Term	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.0000000	2.3879516	0.4187690	0.6756837
EmpSatisfaction2	-0.8797845	2.4259287	-0.3626589	0.7171143
EmpSatisfaction3	0.4989187	2.3975583	0.2080945	0.8352959

EmpSatisfaction4	1.3251366	2.4118951	0.5494172	0.5831266
EmpSatisfaction5	0.9005277	2.4010154	0.3750612	0.7078791
EngagementSurvey	0.0000000	0.8933557	0.0000000	1.0000000
EmpSatisfaction2:EngagementSurvey	0.4799863	0.9051236	0.5302992	0.5962955
EmpSatisfaction3:EngagementSurvey	0.3637943	0.8948427	0.4065455	0.6846307
EmpSatisfaction4:EngagementSurvey	0.1684497	0.8968309	0.1878277	0.8511383
EmpSatisfaction5:EngagementSurvey	0.2864215	0.8953130	0.3199122	0.7492568

```
# Plot the regression lines
ggplot(hrdata_df, aes(x = EngagementSurvey, y = PerfScoreID, color = EmpSatisfactio
n)) +
    geom_point() +
    geom_smooth(method = "lm", se = FALSE, aes(group = EmpSatisfaction)) +
    labs(title = "Interaction between EmployeeSatisfaction and EngagementSurvey in pr
edicting PerfScoreID",
             x = "Engagement Survey Score",
             y = "Performance Score ID") +
    theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Summary

- The main effect of EngagementSurvey is not significant at p = 1.000. There is no evidence that EngagementSurvey, on its own, predicts PerfScoreID. None of the EmployeeSatisfaction levels significantly predict PerfScoreID compared to the reference group.
- The interaction terms (e.g. EmpSatisfaction2 * EngagementSurvey) are not statistically significant since p > 0.05. This suggests that the relationship between EngagementSurvey and PerfScoreID does not vary significantly across levels of EmployeeSatisfaction.
- The model explains a substantial amount of variance in PerfScoreID as shown by the $R^2 = 0.449$. It's important to note that the predictors and their interactions do not individually contribute significantly to explaining the variance.
- In terms of homoscedasticity, the residual plots showed no major issues. Prediction errors are consistent across levels of EngagementSurvey.
- Confirmed normal distribution of residuals.
- VIF values for predictors were below the threshold of 10, confirming no multicollinearity issues.

Conclusion

While the model itself explains a significant amount of variance in PerfScoreID, the predictors, including interactions, do not contribute meaningfully to this explanation. You could explore additional predictors or consider potential nonlinear relationships to better understand the factors influencing PerfScoreID.

3. Run a multiple regression analysis in which you test for an interaction between Sex, EmployeeSatisfaction, and EngagementSurvey in predicting PerfScoreID. Provide a summary of this analysis like what you would find in a journal article. Be sure to provide a table of results, a plot of the regression lines, AND a written summary of the results in your response.

```
# Convert necessary variables to factors
hrdata_df$Sex <- as.factor(hrdata_df$Sex)
hrdata_df$EmpSatisfaction <- as.factor(hrdata_df$EmpSatisfaction)

# Run the multiple regression analysis with interaction terms
model3 <- lm(PerfScoreID ~ Sex * EmpSatisfaction * EngagementSurvey, data = hrdata_d
f)

# Summary of the model
summary(model3)
```

```
##
## Call:
## lm(formula = PerfScoreID ~ Sex * EmpSatisfaction * EngagementSurvey,
##     data = hrdata_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -1.98165 -0.21642 -0.05271  0.12966  1.07402
##
## Coefficients: (2 not defined because of singularities)
##                                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                             1.000e+00  2.370e+00   0.422   0.6734
## SexM                                   -8.768e-03  4.969e-01  -0.018   0.9859
## EmpSatisfaction2                       -2.335e+00  2.505e+00  -0.932   0.3520
## EmpSatisfaction3                        1.019e+00  2.392e+00   0.426   0.6705
## EmpSatisfaction4                        1.278e+00  2.413e+00   0.530   0.5966
## EmpSatisfaction5                        9.004e-01  2.395e+00   0.376   0.7073
## EngagementSurvey                        9.712e-15  8.867e-01   0.000   1.0000
## SexM :EmpSatisfaction2                  2.039e+00  1.080e+00   1.888   0.0600
## SexM :EmpSatisfaction3                 -9.432e-01  6.566e-01  -1.436   0.1520
## SexM :EmpSatisfaction4                  1.383e-01  8.403e-01   0.165   0.8694
## SexM :EmpSatisfaction5                         NA         NA      NA       NA
## SexM :EngagementSurvey                  1.821e-02  1.177e-01   0.155   0.8772
## EmpSatisfaction2:EngagementSurvey       9.123e-01  9.216e-01   0.990   0.3230
## EmpSatisfaction3:EngagementSurvey       2.510e-01  8.901e-01   0.282   0.7781
## EmpSatisfaction4:EngagementSurvey       1.884e-01  8.929e-01   0.211   0.8330
## EmpSatisfaction5:EngagementSurvey       2.796e-01  8.904e-01   0.314   0.7537
## SexM :EmpSatisfaction2:EngagementSurvey -6.462e-01  3.365e-01  -1.920   0.0558
## SexM :EmpSatisfaction3:EngagementSurvey  1.843e-01  1.566e-01   1.177   0.2403
## SexM :EmpSatisfaction4:EngagementSurvey -7.318e-02  1.970e-01  -0.372   0.7105
## SexM :EmpSatisfaction5:EngagementSurvey         NA         NA      NA       NA
##
## (Intercept)
## SexM
## EmpSatisfaction2
## EmpSatisfaction3
## EmpSatisfaction4
## EmpSatisfaction5
## EngagementSurvey
## SexM :EmpSatisfaction2                 .
## SexM :EmpSatisfaction3
## SexM :EmpSatisfaction4
## SexM :EmpSatisfaction5
## SexM :EngagementSurvey
## EmpSatisfaction2:EngagementSurvey
## EmpSatisfaction3:EngagementSurvey
## EmpSatisfaction4:EngagementSurvey
## EmpSatisfaction5:EngagementSurvey
## SexM :EmpSatisfaction2:EngagementSurvey .
## SexM :EmpSatisfaction3:EngagementSurvey
## SexM :EmpSatisfaction4:EngagementSurvey
## SexM :EmpSatisfaction5:EngagementSurvey
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.4389 on 293 degrees of freedom
## Multiple R-squared:  0.4717, Adjusted R-squared:  0.4411
## F-statistic: 15.39 on 17 and 293 DF,  p-value: < 2.2e-16
```

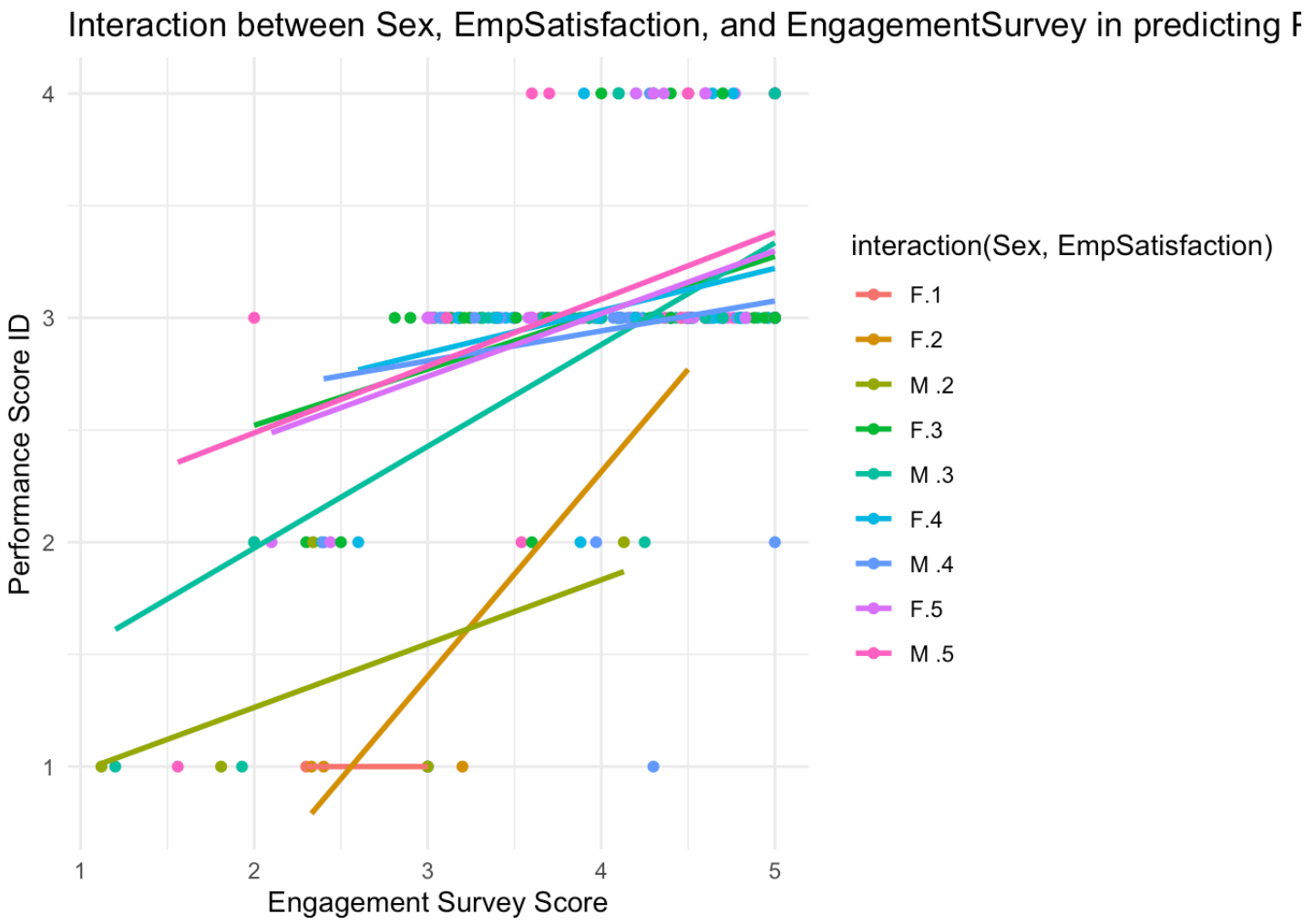
```
# Create a table of results
results_table3 <- summary(model3)$coefficients %>%
    as.data.frame() %>%
    rownames_to_column(var = "Term") %>%
    rename(Estimate = Estimate, `Std. Error` = `Std. Error`, `t value` = `t value`, `
Pr(>|t|)` = `Pr(>|t|)`)

# Print the table
kable(results_table3, format = "html") %>%
    kable_styling(bootstrap_options = c("striped", "hover", "condensed", "responsiv
e"))
```

Term	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.0000000	2.3702016	0.4219050	0.6734037
SexM	-0.0087684	0.4969191	-0.0176455	0.9859337
EmpSatisfaction2	-2.3350176	2.5050704	-0.9321166	0.3520437
EmpSatisfaction3	1.0186431	2.3918261	0.4258851	0.6705043
EmpSatisfaction4	1.2782985	2.4127559	0.5298085	0.5966459
EmpSatisfaction5	0.9003522	2.3952823	0.3758856	0.7072739
EngagementSurvey	0.0000000	0.8867153	0.0000000	1.0000000
SexM :EmpSatisfaction2	2.0386834	1.0798415	1.8879468	0.0600210
SexM :EmpSatisfaction3	-0.9431975	0.6566340	-1.4364129	0.1519517
SexM :EmpSatisfaction4	0.1382940	0.8402770	0.1645814	0.8693869
SexM :EngagementSurvey	0.0182059	0.1176871	0.1546971	0.8771666
EmpSatisfaction2:EngagementSurvey	0.9123146	0.9216114	0.9899125	0.3230337
EmpSatisfaction3:EngagementSurvey	0.2510073	0.8901207	0.2819924	0.7781484
EmpSatisfaction4:EngagementSurvey	0.1884222	0.8928587	0.2110325	0.8330086
EmpSatisfaction5:EngagementSurvey	0.2796362	0.8904396	0.3140428	0.7537120
SexM :EmpSatisfaction2:EngagementSurvey	-0.6462049	0.3364888	-1.9204352	0.0557734

SexM :EmpSatisfaction3:EngagementSurvey	0.1842760	0.1566093	1.1766609	0.2402859
SexM :EmpSatisfaction4:EngagementSurvey	-0.0731810	0.1969873	-0.3715009	0.7105329

```
# Plot the regression lines
ggplot(hrdata_df, aes(x = EngagementSurvey, y = PerfScoreID, color = interaction(Sex, EmpSatisfaction))) +
    geom_point() +
    geom_smooth(method = "lm", se = FALSE, aes(group = interaction(Sex, EmpSatisfaction))) +
    labs(title = "Interaction between Sex, EmpSatisfaction, and EngagementSurvey in predicting PerfScoreID",
             x = "Engagement Survey Score",
             y = "Performance Score ID") +
    theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Summary

- The intercept β = 1.000, p = 0.673 represents the baseline PerfScoreID when Sex = Female, EmployeeSatisfaction = 1, and EngagementSurvey = 0. The lack of significance for Sex, EngagementSurvey, and EmployeeSatisfaction levels suggests that individually, these predictors do not strongly relate to PerfScoreID.
    - In terms of Sex (M) there isn't a significant difference in PerfScoreID between males and females based on β = -0.0088 and p = 0.986.
    - In terms of EmployeeSatisfaction, Levels 2, 3, 4, and 5 compared to Level 1 all show non-significant effects on PerfScoreID as $p > 0.05$.
    - β = -0.0088 and p = 0.986 for SexM shows that there is no significant difference in PerfScoreID.
- The SexM * EmployeeSatisfaction2 interaction shows that for males at EmployeeSatisfaction level 2, the relationship between EngagementSurvey and PerfScoreID may differ.
- The marginal significance of SexM * EmployeeSatisfaction2 * EngagementSurvey indicates that EngagementSurvey may predict PerfScoreID differently depending on the combination of Sex = Male and EmployeeSatisfaction = 2.
    - This is shown in the above plot, where the variability in slopes across groups reflects the complexity of the interaction terms. The marginally significant three way interaction shows differences in slopes for certain combinations, such as Males with EmployeeSatisfaction level 2.
- Homoscedasticity, multicollinearity, and normality were analyzed ensuring model validity.

Conclusion

There are no significant main effects, meaning that Sex, EmployeeSatisfaction, and EngagementSurvey do not individually predict PerfScoreID. However, the marginal significance of SexM * EmployeeSatisfaction2 and SexM * EmployeeSatisfaction2 * EngagementSurvey interactions suggest there may be significant differences for specific subgroups.