STAT 408 Final Project

Shreya Bhagi

1. Data Import

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.1     v stringr   1.5.2
v ggplot2   4.0.0     v tibble    3.3.0
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.1.0
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

```
library(MASS)
```

```
Attaching package: 'MASS'

The following object is masked from 'package:dplyr':

    select
```

```
library(lmtest)
```

```
Loading required package: zoo

Attaching package: 'zoo'
```

```
The following objects are masked from 'package:base':

    as.Date, as.Date.numeric
```

```
prostate <- read.csv("/Users/shreyabhagi/Desktop/StatFinal/synthetic_prostate_cancer_risk.csv

str(prostate)
```

```
'data.frame':   1000 obs. of  13 variables:
 $ id                     : int  1 2 3 4 5 6 7 8 9 10 ...
 $ age                    : int  39 43 72 60 51 68 39 67 58 78 ...
 $ bmi                    : num  24.7 25.6 22.4 25.6 26.6 29.4 24.4 25.1 20.6 26.9 ...
 $ smoker                 : chr  "No" "Yes" "No" "Yes" ...
 $ alcohol_consumption    : chr  "None" "None" "Moderate" "None" ...
 $ diet_type              : chr  "Fatty" "Mixed" "Mixed" "Fatty" ...
 $ physical_activity_level: chr  "Moderate" "High" "Moderate" "Moderate" ...
 $ family_history         : chr  "No" "No" "No" "No" ...
 $ mental_stress_level    : chr  "High" "High" "High" "High" ...
 $ sleep_hours            : num  5.6 6.9 7.8 5.6 5.9 8.3 6 6.4 8.3 8.3 ...
 $ regular_health_checkup : chr  "No" "Yes" "Yes" "Yes" ...
 $ prostate_exam_done     : chr  "No" "No" "No" "No" ...
 $ risk_level             : chr  "Medium" "Low" "Low" "Medium" ...
```

```
head(prostate)
```

```
  id age  bmi smoker alcohol_consumption diet_type physical_activity_level
1  1  39 24.7     No                None     Fatty                 Moderate
2  2  43 25.6    Yes                None     Mixed                     High
3  3  72 22.4     No            Moderate     Mixed                 Moderate
4  4  60 25.6    Yes                None     Fatty                 Moderate
5  5  51 26.6    Yes                None     Mixed                      Low
6  6  68 29.4    Yes            Moderate     Mixed                 Moderate
  family_history mental_stress_level sleep_hours regular_health_checkup
1             No                High         5.6                     No
2             No                High         6.9                    Yes
3             No                High         7.8                    Yes
4             No                High         5.6                    Yes
5             No              Medium         5.9                     No
6            Yes              Medium         8.3                     No
  prostate_exam_done risk_level
```

```
1                  No     Medium
2                  No        Low
3                  No        Low
4                  No     Medium
5                  No     Medium
6                  No     Medium
```

```
summary(prostate)
```

```
       id              age              bmi            smoker
 Min.   :   1.0   Min.   :30.00   Min.   :17.70   Length:1000
 1st Qu.: 250.8   1st Qu.:43.00   1st Qu.:23.90   Class :character
 Median : 500.5   Median :55.00   Median :26.00   Mode  :character
 Mean   : 500.5   Mean   :55.16   Mean   :26.00
 3rd Qu.: 750.2   3rd Qu.:68.00   3rd Qu.:28.02
 Max.   :1000.0   Max.   :80.00   Max.   :36.80
 alcohol_consumption  diet_type         physical_activity_level
 Length:1000         Length:1000        Length:1000
 Class :character    Class :character   Class :character
 Mode  :character    Mode  :character   Mode  :character


 family_history     mental_stress_level  sleep_hours    regular_health_checkup
 Length:1000        Length:1000         Min.   :4.500   Length:1000
 Class :character   Class :character    1st Qu.:5.600   Class :character
 Mode  :character   Mode  :character    Median :6.700   Mode  :character
                                        Mean   :6.726
                                        3rd Qu.:7.900
                                        Max.   :9.000
 prostate_exam_done  risk_level
 Length:1000        Length:1000
 Class :character   Class :character
 Mode  :character   Mode  :character
```

2. Data Exploration & Preparation

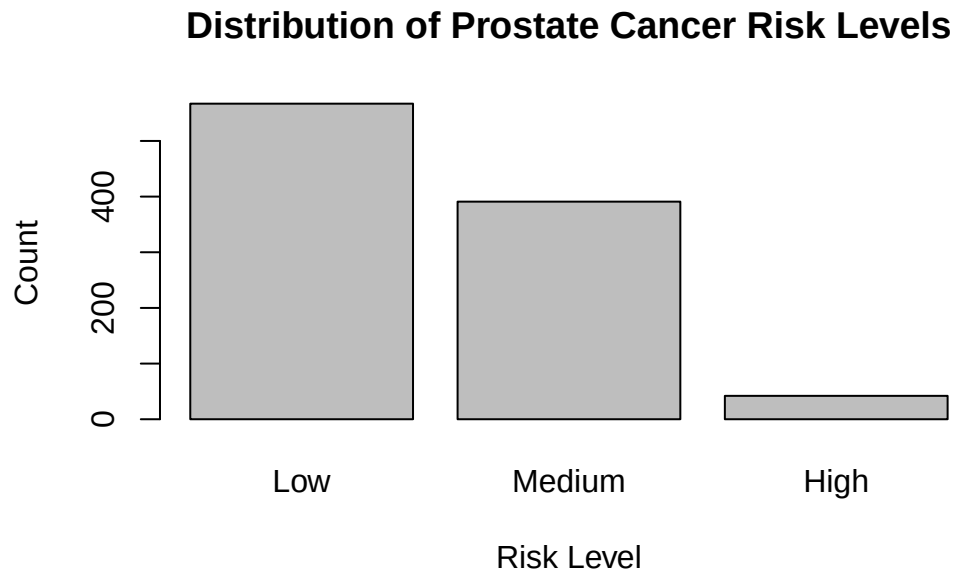
```r
colSums(is.na(prostate))
```

```
                     id                     age                     bmi 
                      0                       0                       0 
                 smoker     alcohol_consumption               diet_type 
                      0                       0                       0 
physical_activity_level          family_history     mental_stress_level 
                      0                       0                       0 
            sleep_hours  regular_health_checkup      prostate_exam_done 
                      0                       0                       0 
             risk_level 
                      0 
```

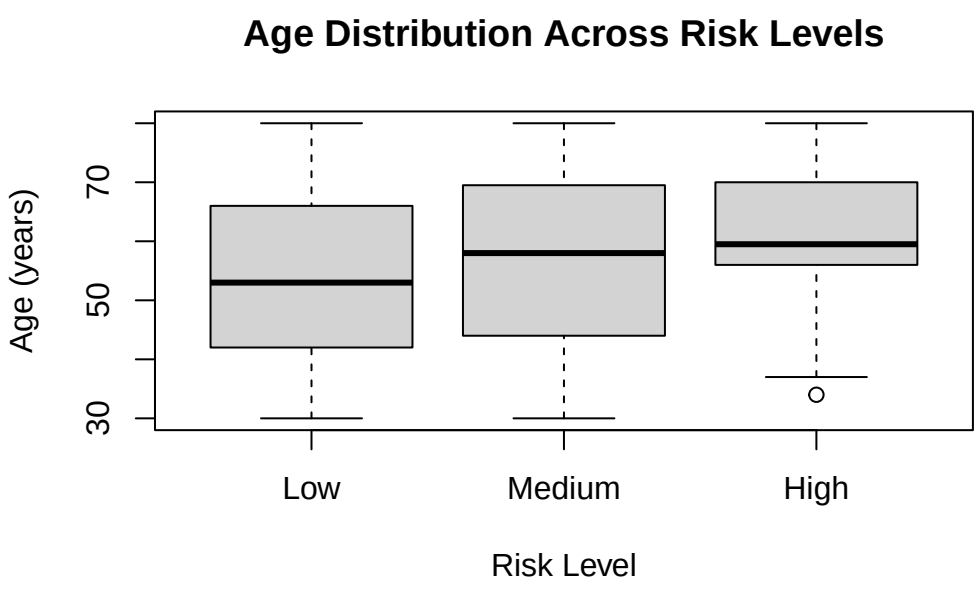
```r
prostate <- prostate %>%
  mutate(
    risk_level = factor(
      risk_level,
      levels  = c("Low","Medium","High"),
      ordered = TRUE
    ),
    risk_numeric = as.numeric(risk_level),

    smoker                  = factor(smoker),
    alcohol_consumption     = factor(alcohol_consumption,
                                     levels  = c("None","Moderate","High"),
                                     ordered = TRUE),
    diet_type               = factor(diet_type,
                                     levels  = c("Fatty","Mixed","Healthy"),
                                     ordered = TRUE),
    physical_activity_level = factor(physical_activity_level,
                                     levels  = c("Low","Moderate","High"),
                                     ordered = TRUE),
    family_history          = factor(family_history),
    mental_stress_level     = factor(mental_stress_level,
                                     levels  = c("Low","Medium","High"),
                                     ordered = TRUE),
    regular_health_checkup  = factor(regular_health_checkup),
    prostate_exam_done      = factor(prostate_exam_done)
  )

summary(prostate$risk_level)
```

```
   Low Medium   High
   567    391     42
```

```r
barplot(
  table(prostate$risk_level),
  main = "Distribution of Prostate Cancer Risk Levels",
  xlab = "Risk Level",
  ylab = "Count"
)
```




```r
boxplot(
  age ~ risk_level,
  data = prostate,
  main = "Age Distribution Across Risk Levels",
  xlab = "Risk Level",
  ylab = "Age (years)"
)
```




```
cor(prostate[, c("age","bmi","sleep_hours")])
```

```
                    age         bmi sleep_hours
age          1.00000000 -0.01187699 -0.04033382
bmi         -0.01187699  1.00000000 -0.03424433
sleep_hours -0.04033382 -0.03424433  1.00000000
```

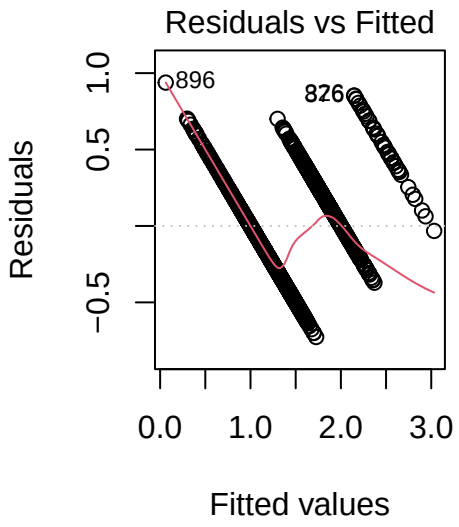
3. Baseline Multiple Linear Regression

```
lm_full <- lm(
  risk_numeric ~ age + bmi + smoker + alcohol_consumption +
    diet_type + physical_activity_level + family_history +
    mental_stress_level + sleep_hours +
    regular_health_checkup + prostate_exam_done,
  data = prostate
)

summary(lm_full)
```

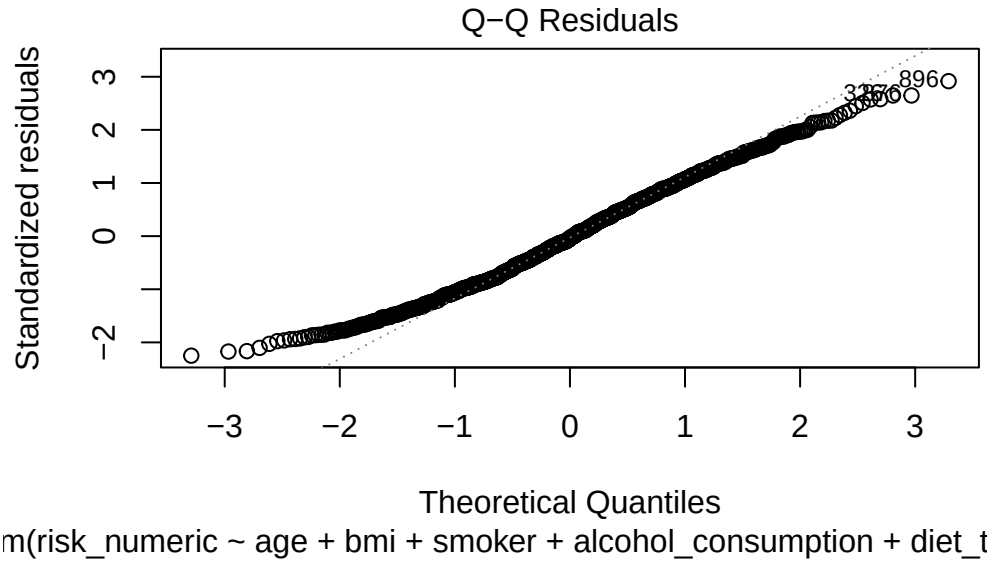
```
Call:
```

```
lm(formula = risk_numeric ~ age + bmi + smoker + alcohol_consumption +
    diet_type + physical_activity_level + family_history + mental_stress_level +
    sleep_hours + regular_health_checkup + prostate_exam_done,
    data = prostate)

Residuals:
     Min       1Q   Median       3Q      Max
-0.72639 -0.25720 -0.01183  0.23785  0.93801

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 0.8181990  0.1156454   7.075 2.83e-12 ***
age                         0.0117948  0.0008115  14.534  < 2e-16 ***
bmi                         0.0214887  0.0034295   6.266 5.54e-10 ***
smokerYes                   0.3811546  0.0207313  18.385  < 2e-16 ***
alcohol_consumption.L       0.2571458  0.0202957  12.670  < 2e-16 ***
alcohol_consumption.Q       0.2002582  0.0175621  11.403  < 2e-16 ***
diet_type.L                -0.2598049  0.0186421 -13.936  < 2e-16 ***
diet_type.Q                 0.1848742  0.0174975  10.566  < 2e-16 ***
physical_activity_level.L -0.2940465  0.0184664 -15.923  < 2e-16 ***
physical_activity_level.Q  0.1825625  0.0174165  10.482  < 2e-16 ***
family_historyYes           0.4131165  0.0265514  15.559  < 2e-16 ***
mental_stress_level.L       0.1355132  0.0190670   7.107 2.27e-12 ***
mental_stress_level.Q       0.0650641  0.0171377   3.797 0.000156 ***
sleep_hours                -0.0615599  0.0078271  -7.865 9.67e-15 ***
regular_health_checkupYes -0.4126357  0.0207222 -19.913  < 2e-16 ***
prostate_exam_doneYes      -0.3406904  0.0250172 -13.618  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3248 on 984 degrees of freedom
Multiple R-squared:  0.6886,    Adjusted R-squared:  0.6839
F-statistic: 145.1 on 15 and 984 DF,  p-value: < 2.2e-16
```

```r
par(mfrow = c(1,2))
plot(lm_full, which = 1)
```




```
plot(lm_full, which = 2)
```




```
m(risk_numeric ~ age + bmi + smoker + alcohol_consumption + diet_type +
```

```r
par(mfrow = c(1,1))
```

```r
bptest(lm_full)
```

```
    studentized Breusch-Pagan test

data:  lm_full
BP = 25.079, df = 15, p-value = 0.04889
```

```r
ks.test(rstandard(lm_full), "pnorm")
```

```
    Asymptotic one-sample Kolmogorov-Smirnov test

data:  rstandard(lm_full)
D = 0.044184, p-value = 0.04031
alternative hypothesis: two-sided
```

4. Ordinal Logistic Regression

```r
ord_main <- polr(
  risk_level ~ age + bmi + smoker + alcohol_consumption +
    diet_type + physical_activity_level + family_history +
    mental_stress_level + sleep_hours +
    regular_health_checkup + prostate_exam_done,
  data = prostate,
  Hess = TRUE
)

summary(ord_main)
```

```
Call:
polr(formula = risk_level ~ age + bmi + smoker + alcohol_consumption +
    diet_type + physical_activity_level + family_history + mental_stress_level +
    sleep_hours + regular_health_checkup + prostate_exam_done,
    data = prostate, Hess = TRUE)

Coefficients:
```

```
                            Value Std. Error t value
age                        0.1902    0.01724  11.029
bmi                        0.3711    0.05378   6.900
smokerYes                  7.1713    0.61334  11.692
alcohol_consumption.L      4.2175    0.40891  10.314
alcohol_consumption.Q      3.1775    0.32956   9.642
diet_type.L               -4.5420    0.41648 -10.906
diet_type.Q                2.8674    0.31477   9.110
physical_activity_level.L -4.7589    0.42336 -11.241
physical_activity_level.Q  3.2702    0.34640   9.441
family_historyYes          6.5150    0.57455  11.339
mental_stress_level.L      2.1947    0.30082   7.296
mental_stress_level.Q      1.2414    0.24685   5.029
sleep_hours               -1.0363    0.12775  -8.112
regular_health_checkupYes -7.0598    0.58766 -12.013
prostate_exam_doneYes     -6.0001    0.56287 -10.660

Intercepts:
            Value    Std. Error t value
Low|Medium   12.2499   1.7722     6.9121
Medium|High  25.4321   2.4548    10.3601

Residual Deviance: 363.7767
AIC: 397.7767
```

```r
ord_int_full <- polr(
  risk_level ~ age + bmi + smoker + alcohol_consumption +
    diet_type + physical_activity_level + family_history +
    mental_stress_level + sleep_hours +
    regular_health_checkup + prostate_exam_done +
    age:physical_activity_level + age:family_history,
  data = prostate,
  Hess = TRUE
)

summary(ord_int_full)
```

```
Call:
polr(formula = risk_level ~ age + bmi + smoker + alcohol_consumption +
    diet_type + physical_activity_level + family_history + mental_stress_level +
    sleep_hours + regular_health_checkup + prostate_exam_done +
    age:physical_activity_level + age:family_history, data = prostate,
```

```
    Hess = TRUE)

Coefficients:
                                   Value Std. Error  t value
age                             0.193555    0.01835  10.5494
bmi                             0.371910    0.05395   6.8930
smokerYes                       7.210857    0.62087  11.6141
alcohol_consumption.L           4.235913    0.41223  10.2756
alcohol_consumption.Q           3.189033    0.33070   9.6434
diet_type.L                    -4.571123    0.42174 -10.8387
diet_type.Q                     2.884131    0.31715   9.0939
physical_activity_level.L      -5.331725    1.10418  -4.8287
physical_activity_level.Q       3.048176    0.95950   3.1768
family_historyYes               6.943194    1.43553   4.8367
mental_stress_level.L           2.211480    0.30338   7.2895
mental_stress_level.Q           1.249115    0.24749   5.0471
sleep_hours                    -1.036344    0.12803  -8.0943
regular_health_checkupYes      -7.087598    0.59226 -11.9671
prostate_exam_doneYes          -6.011217    0.56647 -10.6118
age:physical_activity_level.L   0.009980    0.01777   0.5617
age:physical_activity_level.Q   0.004368    0.01653   0.2642
age:family_historyYes          -0.007361    0.02386  -0.3084

Intercepts:
             Value    Std. Error t value
Low|Medium   12.4667   1.8243     6.8335
Medium|High  25.6817   2.5060    10.2480

Residual Deviance: 363.3498
AIC: 403.3498
```

```r
ord_final <- step(ord_int_full,
                  direction = "backward",
                  trace = 0)

summary(ord_final)
```

```
Call:
polr(formula = risk_level ~ age + bmi + smoker + alcohol_consumption +
    diet_type + physical_activity_level + family_history + mental_stress_level +
    sleep_hours + regular_health_checkup + prostate_exam_done,
    data = prostate, Hess = TRUE)
```

```
Coefficients:
                            Value Std. Error t value
age                        0.1902    0.01724  11.029
bmi                        0.3711    0.05378   6.900
smokerYes                  7.1713    0.61334  11.692
alcohol_consumption.L      4.2175    0.40891  10.314
alcohol_consumption.Q      3.1775    0.32956   9.642
diet_type.L               -4.5420    0.41648 -10.906
diet_type.Q                2.8674    0.31477   9.110
physical_activity_level.L -4.7589    0.42336 -11.241
physical_activity_level.Q  3.2702    0.34640   9.441
family_historyYes          6.5150    0.57455  11.339
mental_stress_level.L      2.1947    0.30082   7.296
mental_stress_level.Q      1.2414    0.24685   5.029
sleep_hours               -1.0363    0.12775  -8.112
regular_health_checkupYes -7.0598    0.58766 -12.013
prostate_exam_doneYes     -6.0001    0.56287 -10.660

Intercepts:
            Value    Std. Error t value
Low|Medium   12.2499   1.7722     6.9121
Medium|High  25.4321   2.4548    10.3601

Residual Deviance: 363.7767
AIC: 397.7767
```

```r
ctable_final <- coef(summary(ord_final))
p_vals_final <- 2 * pnorm(abs(ctable_final[, "t value"]), lower.tail = FALSE)
results_final <- cbind(ctable_final, "p value" = p_vals_final)
results_final
```

```
                              Value Std. Error    t value      p value
age                       0.1901857 0.01724467  11.028668 2.779512e-28
bmi                       0.3711001 0.05378292   6.899962 5.201633e-12
smokerYes                 7.1713207 0.61334320  11.692183 1.397483e-31
alcohol_consumption.L     4.2174919 0.40891443  10.313874 6.099208e-25
alcohol_consumption.Q     3.1775121 0.32956386   9.641567 5.336693e-22
diet_type.L              -4.5420258 0.41648063 -10.905731 1.082196e-27
diet_type.Q               2.8673697 0.31476570   9.109537 8.273446e-20
physical_activity_level.L -4.7588507 0.42335971 -11.240679 2.573948e-29
physical_activity_level.Q  3.2702164 0.34640269   9.440505 3.709864e-21
```

```
family_historyYes           6.5149663 0.57454695  11.339310 8.380213e-30
mental_stress_level.L       2.1946875 0.30082186   7.295638 2.972462e-13
mental_stress_level.Q       1.2413532 0.24684770   5.028822 4.935020e-07
sleep_hours                -1.0363034 0.12775394  -8.111714 4.991074e-16
regular_health_checkupYes  -7.0597591 0.58766311 -12.013276 3.026152e-33
prostate_exam_doneYes      -6.0001293 0.56286884 -10.659907 1.567547e-26
Low|Medium                 12.2498630 1.77223668   6.912092 4.775580e-12
Medium|High                25.4321223 2.45481779  10.360086 3.766250e-25
```

```r
AIC(ord_main, ord_int_full, ord_final)
```

```
             df      AIC
ord_main     17 397.7767
ord_int_full 20 403.3498
ord_final    17 397.7767
```

5. Odd Ratios for Final Model

```r
ORs <- exp(coef(ord_final))
final_OR_table <- data.frame(
  Predictor  = names(ORs),
  Odds_Ratio = as.numeric(ORs)
)
final_OR_table
```

```
                   Predictor   Odds_Ratio
1                        age 1.209474e+00
2                        bmi 1.449328e+00
3                  smokerYes 1.301562e+03
4      alcohol_consumption.L 6.786306e+01
5      alcohol_consumption.Q 2.398700e+01
6                diet_type.L 1.065181e-02
7                diet_type.Q 1.759069e+01
8  physical_activity_level.L 8.575460e-03
9  physical_activity_level.Q 2.631703e+01
10         family_historyYes 6.751712e+02
11     mental_stress_level.L 8.977195e+00
12     mental_stress_level.Q 3.460293e+00
13               sleep_hours 3.547637e-01
14 regular_health_checkupYes 8.589850e-04
15     prostate_exam_doneYes 2.478432e-03
```

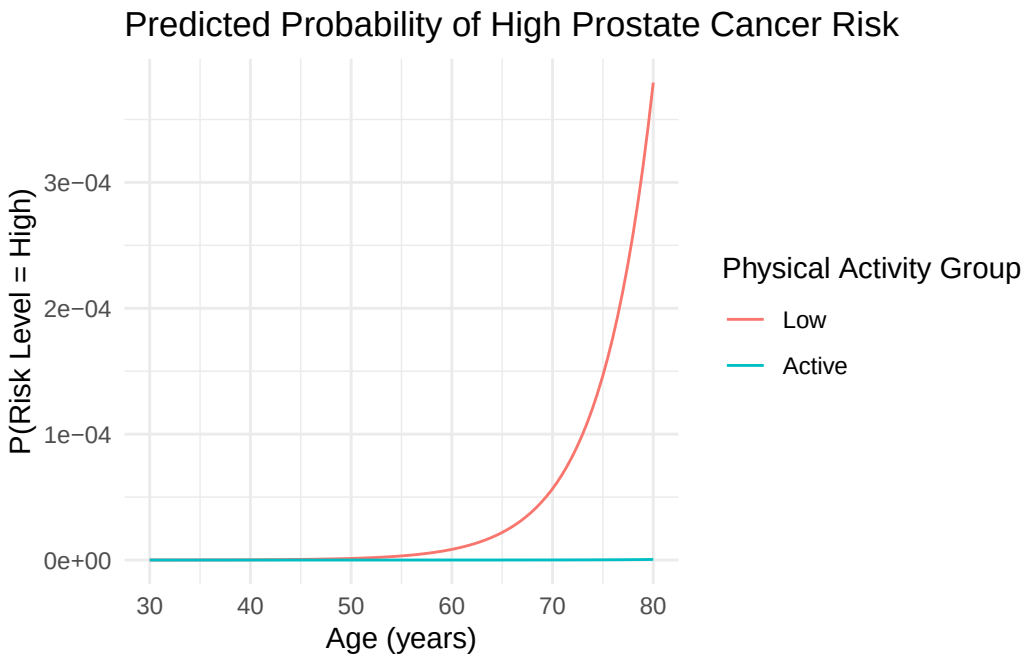
6. Predicted Probabilities Plot

```
newdat <- expand.grid(
  age        = seq(min(prostate$age), max(prostate$age), length.out = 100),
  act_group  = c("Low","Active")
)

newdat <- newdat %>%
  mutate(
    physical_activity_level = ifelse(act_group == "Low", "Low", "High"),
    physical_activity_level = factor(
      physical_activity_level,
      levels  = levels(prostate$physical_activity_level),
      ordered = is.ordered(prostate$physical_activity_level)
    ),

    bmi                   = mean(prostate$bmi),
    smoker                = "No",
    alcohol_consumption   = "None",
    diet_type             = "Mixed",
    family_history        = "No",
    mental_stress_level   = "Medium",
    sleep_hours           = mean(prostate$sleep_hours),
    regular_health_checkup = "No",
    prostate_exam_done     = "No"
  )

newdat <- newdat %>%
  mutate(
    smoker                = factor(smoker, levels = levels(prostate$smoker)),
    alcohol_consumption   = factor(alcohol_consumption,
                                   levels  = levels(prostate$alcohol_consumption),
                                   ordered = is.ordered(prostate$alcohol_consumption)),
    diet_type             = factor(diet_type,
                                   levels  = levels(prostate$diet_type),
                                   ordered = is.ordered(prostate$diet_type)),
    family_history        = factor(family_history,
                                   levels = levels(prostate$family_history)),
    mental_stress_level   = factor(mental_stress_level,
                                   levels  = levels(prostate$mental_stress_level),
                                   ordered = is.ordered(prostate$mental_stress_level)),
    regular_health_checkup = factor(regular_health_checkup,
                                    levels = levels(prostate$regular_health_checkup)),
```

```
    prostate_exam_done     = factor(prostate_exam_done,
                                    levels = levels(prostate$prostate_exam_done))
  )
```

```
pred_probs <- predict(ord_final, newdat, type = "probs")
pred_df <- cbind(newdat, as.data.frame(pred_probs))
```

```
ggplot(pred_df,
       aes(x = age, y = High, color = act_group)) +
  geom_line() +
  labs(
    title = "Predicted Probability of High Prostate Cancer Risk",
    x = "Age (years)",
    y = "P(Risk Level = High)",
    color = "Physical Activity Group"
  ) +
  theme_minimal()
```




```
prostate %>%
  group_by(risk_level) %>%
  summarize(
    n           = n(),
```

```
    mean_age    = mean(age),
    mean_bmi    = mean(bmi),
    prop_smoker = mean(smoker == "Yes"),
    prop_famhx  = mean(family_history == "Yes")
  )
```

```
# A tibble: 3 x 6
  risk_level     n mean_age mean_bmi prop_smoker prop_famhx
  <ord>      <int>    <dbl>    <dbl>       <dbl>      <dbl>
1 Low          567     53.8     25.7       0.384      0.106
2 Medium       391     56.5     26.4       0.660      0.279
3 High          42     60.2     27.0       0.881      0.405
```