502 Project EDA

Madeline Chang

```r
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- 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 become errors
```

```r
library(dplyr)
library(forcats)
library(ggplot2)
library(rpart)
library(caret)
```

```
## Loading required package: lattice
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:purrr':
##
##     lift
```

```r
library(rpart.plot)
library(C50)
library(Metrics)
```

```
##
## Attaching package: 'Metrics'
##
## The following objects are masked from 'package:caret':
##
##     precision, recall
```

```r
library(e1071)
library(glmnet)
```

```
## Loading required package: Matrix
##
## Attaching package: 'Matrix'
##
## The following objects are masked from 'package:tidyr':
```

```
##
##     expand, pack, unpack
##
## Loaded glmnet 4.1-8
```

Reading in data

```r
recruitment<- read.csv('/Users/mtc/ADS/ADS 502/Project/recruitment_data.csv')

recruitment<- recruitment %>%
  mutate(Gender = as.factor(Gender),
         EducationLevel = as.factor(EducationLevel),
         RecruitmentStrategy = as.factor(RecruitmentStrategy),
         HiringDecision = as.factor(HiringDecision),
         Gender_name = as.factor(ifelse(Gender == 0, "Male", "Female")),
         Hiring_name = as.factor(ifelse(HiringDecision == 0, "Not Hired", "Hired")),
         Recruitment_name = fct_collapse(RecruitmentStrategy,
                                        Aggressive = 1,
                                        Moderate = 2,
                                        Conservative = 3),
         Education_name = fct_collapse(EducationLevel,
                                       Bachelor_1 = 1,
                                       Bachelor_2 = 2,
                                       Masters = 3,
                                       PhD = 4))
```

Data Quality

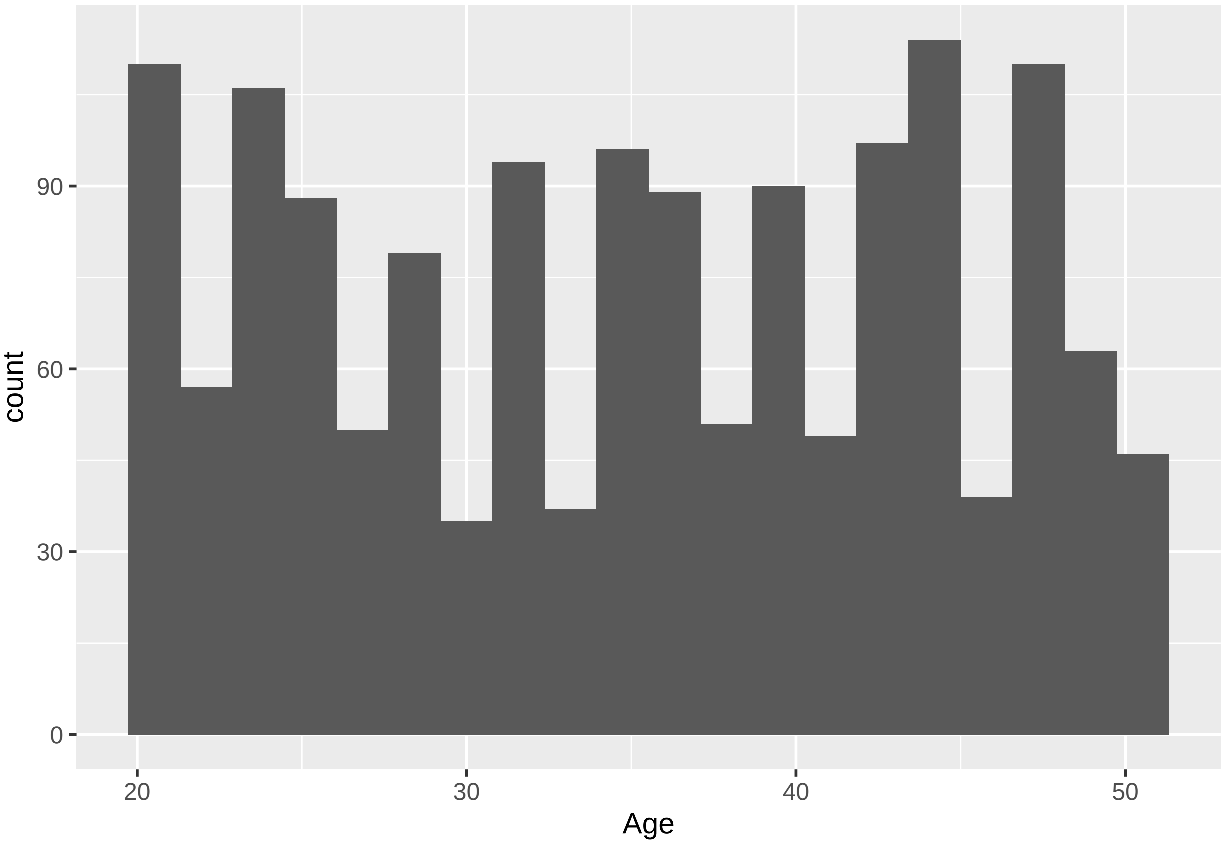
```r
colSums(is.na(recruitment))
```

```
##                 Age              Gender      EducationLevel     ExperienceYears
##                   0                   0                   0                   0
##   PreviousCompanies DistanceFromCompany      InterviewScore          SkillScore
##                   0                   0                   0                   0
##    PersonalityScore RecruitmentStrategy      HiringDecision         Gender_name
##                   0                   0                   0                   0
##         Hiring_name    Recruitment_name      Education_name
##                   0                   0                   0
```

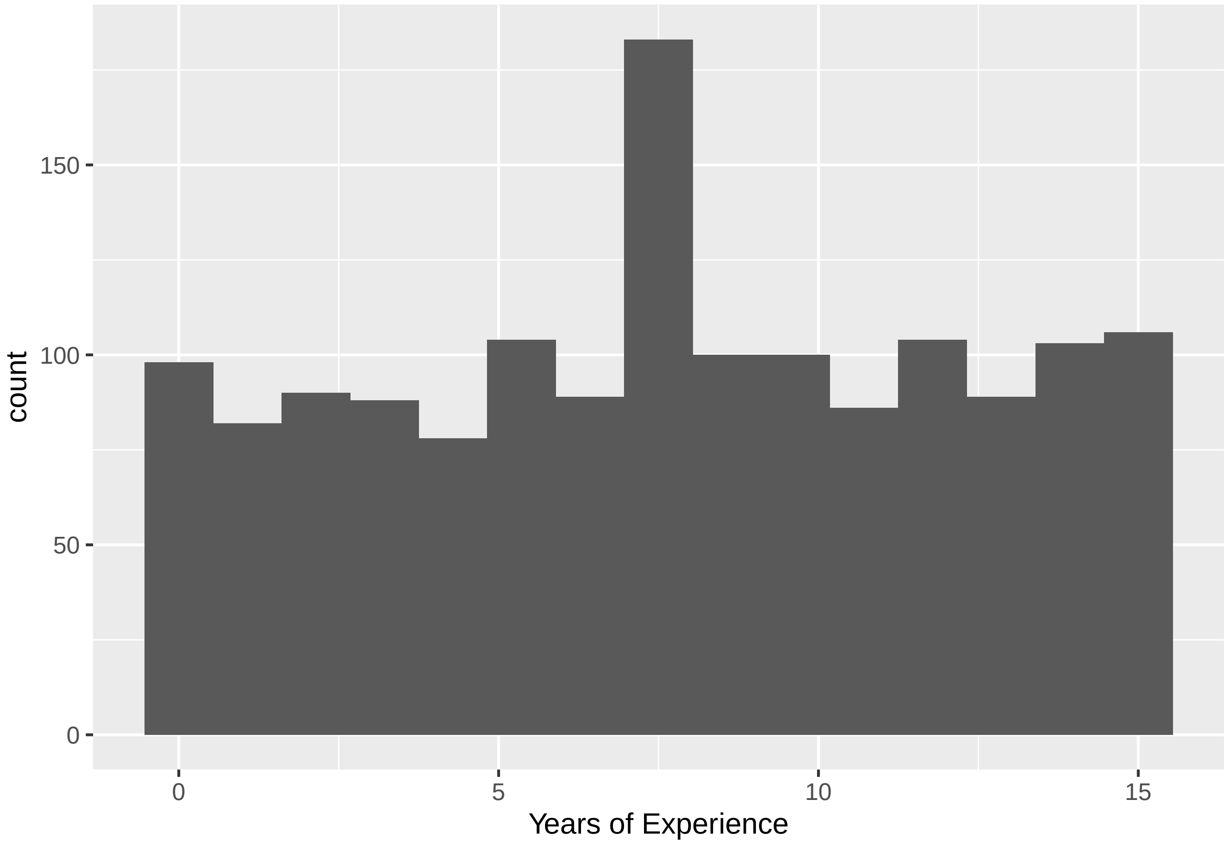
```r
near_zero<- nearZeroVar(recruitment) # no columns with zero or near-zero variance

corr<- cor(recruitment[,c(1, 4, 5, 6, 7, 8, 9)])
high_corr <- findCorrelation(corr, cutoff = 0.75) # no numeric columns with high correlation with each
```

Distribution of Variables

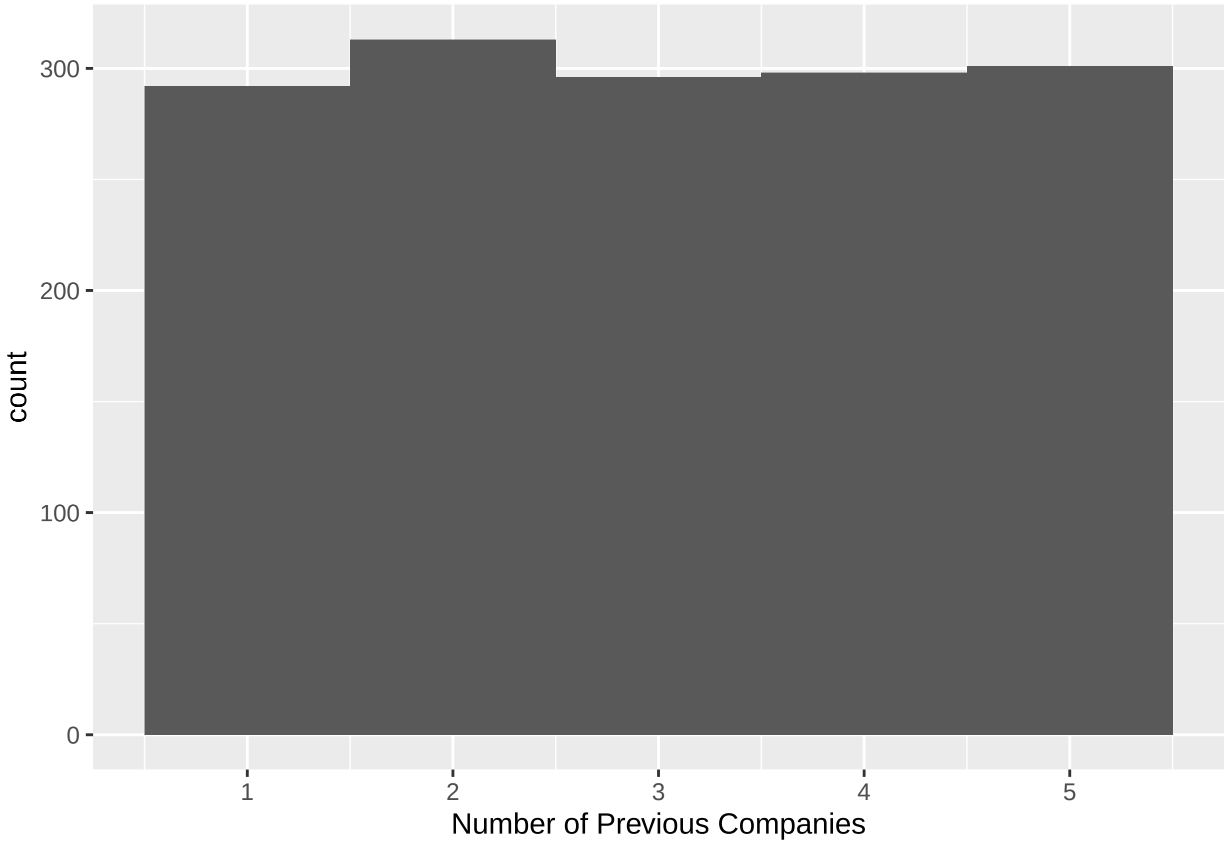
```r
hist<- function(col, bin_num){
  ggplot(data = recruitment) +
    geom_histogram(aes(x = .data[[col]]), bins = bin_num) +
    xlab(col)
}


hist("Age", 20)
```



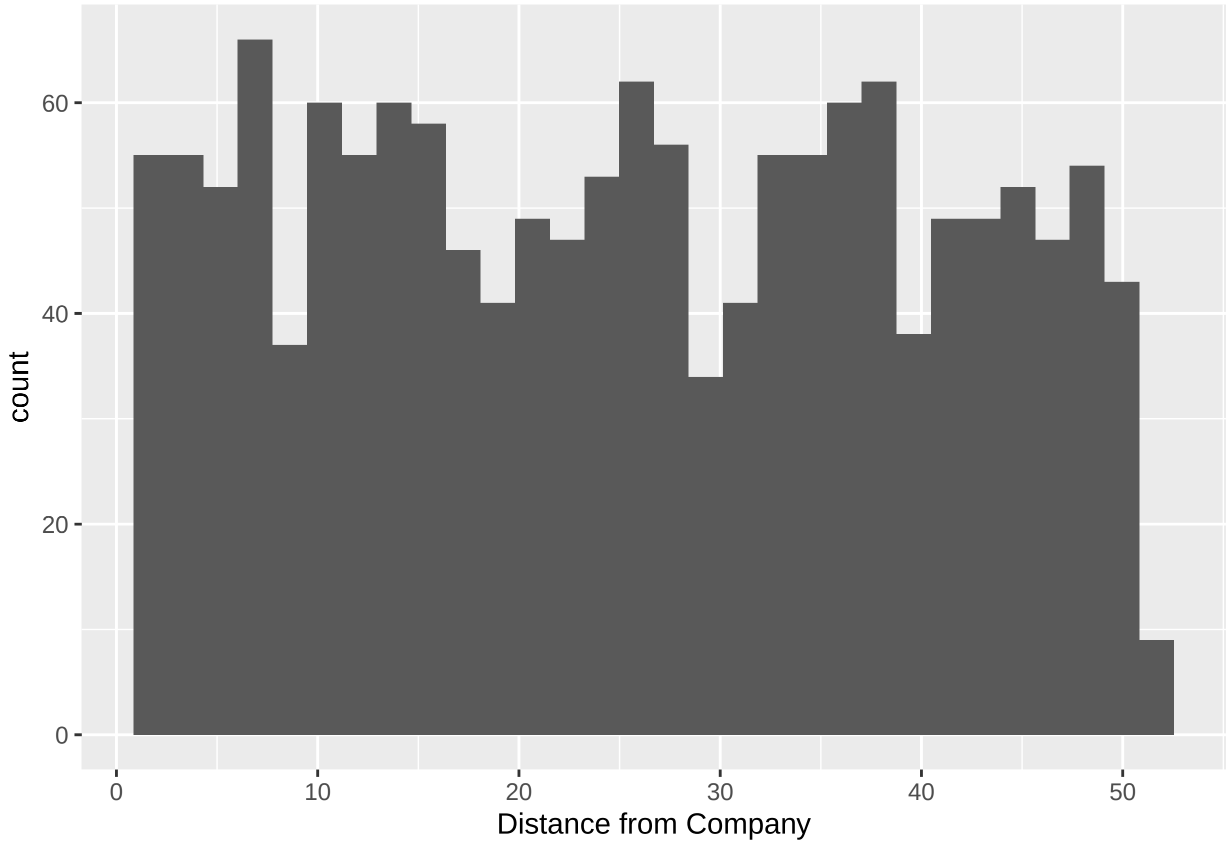
```
hist("ExperienceYears", 15) +
  xlab("Years of Experience")
```



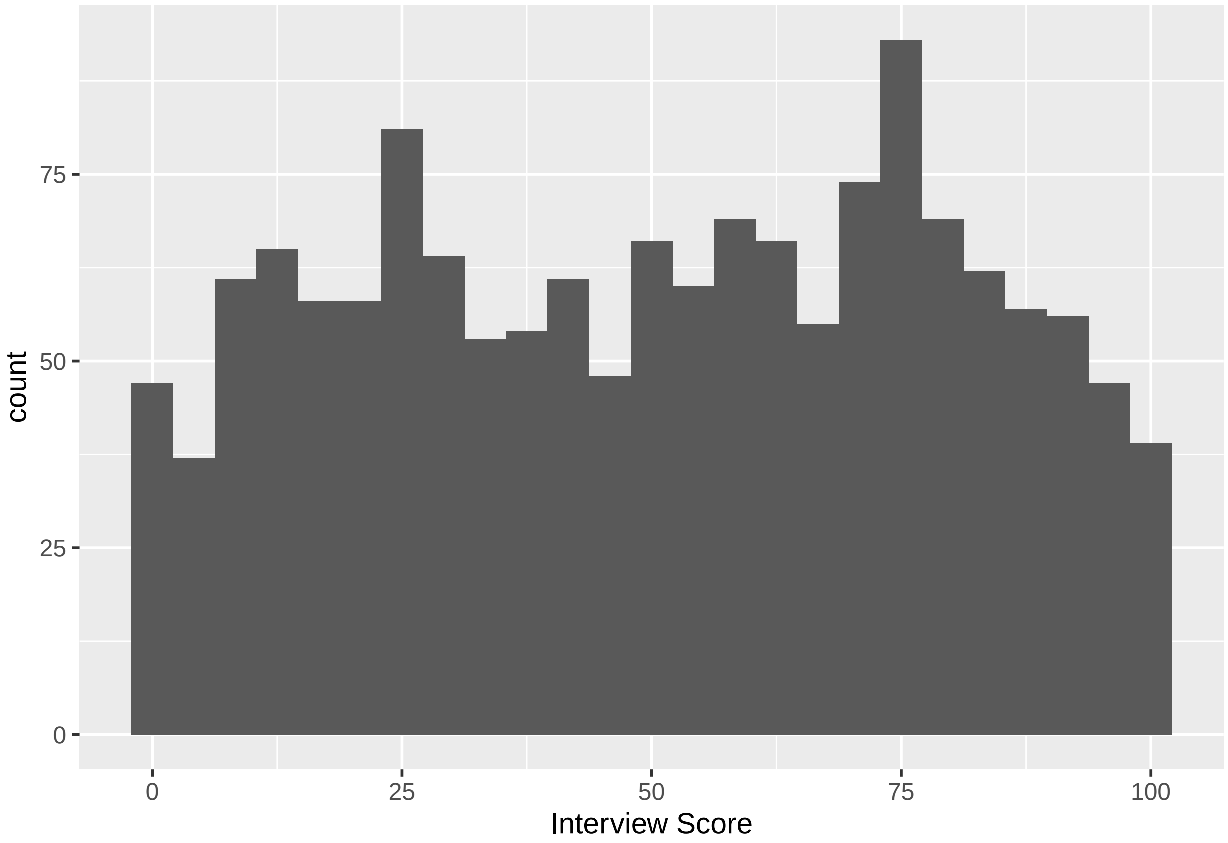
```
hist("PreviousCompanies", 5) +
  xlab("Number of Previous Companies")
```



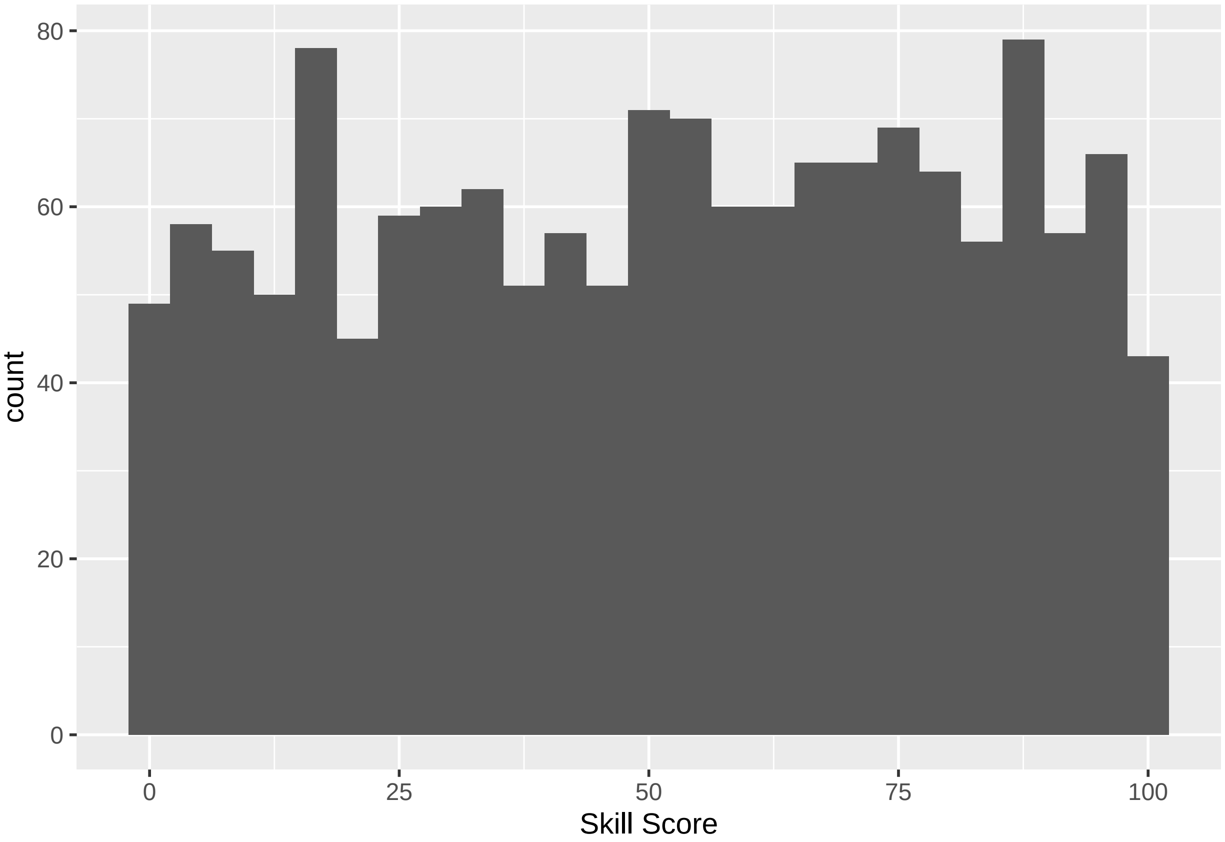
```r
hist("DistanceFromCompany", 30) +
  xlab("Distance from Company")
```



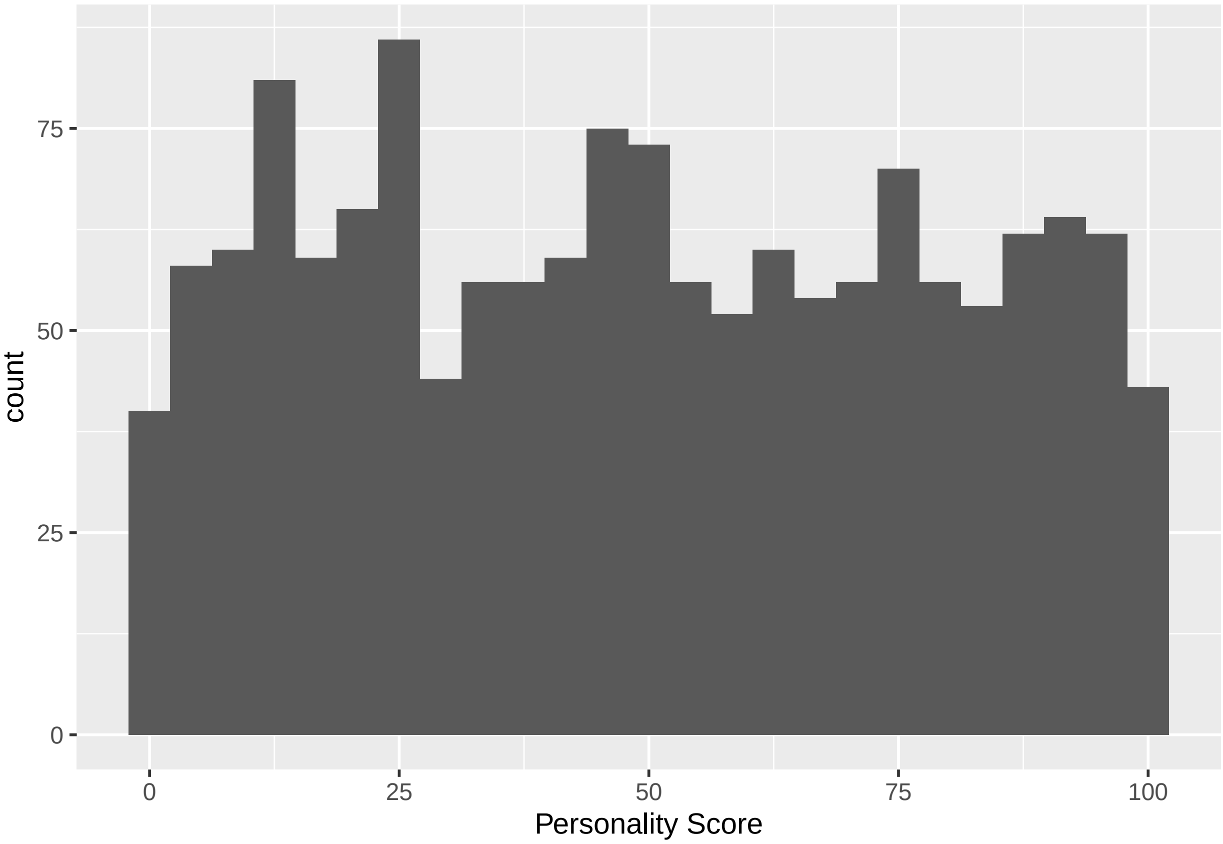
```r
hist("InterviewScore", 25) +
  xlab("Interview Score")
```



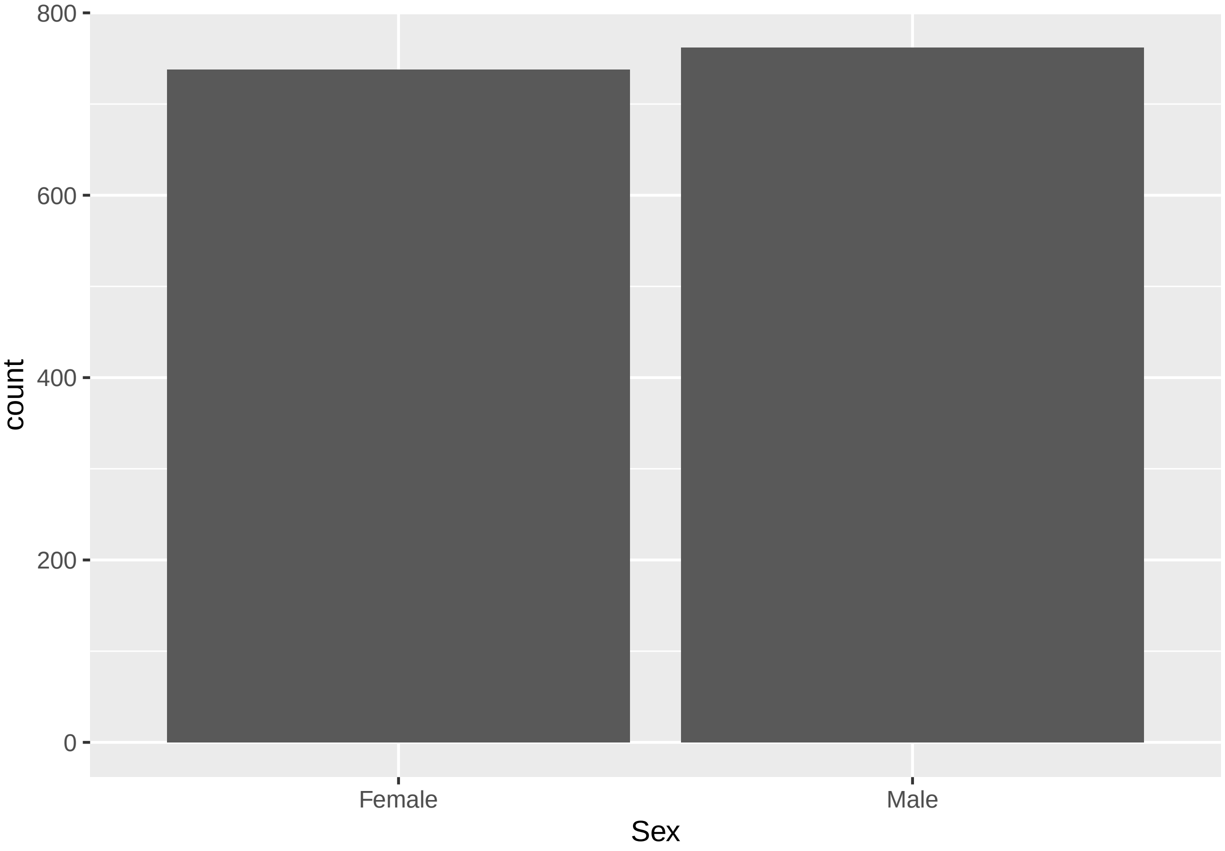
```
hist("SkillScore", 25)+
  xlab("Skill Score")
```



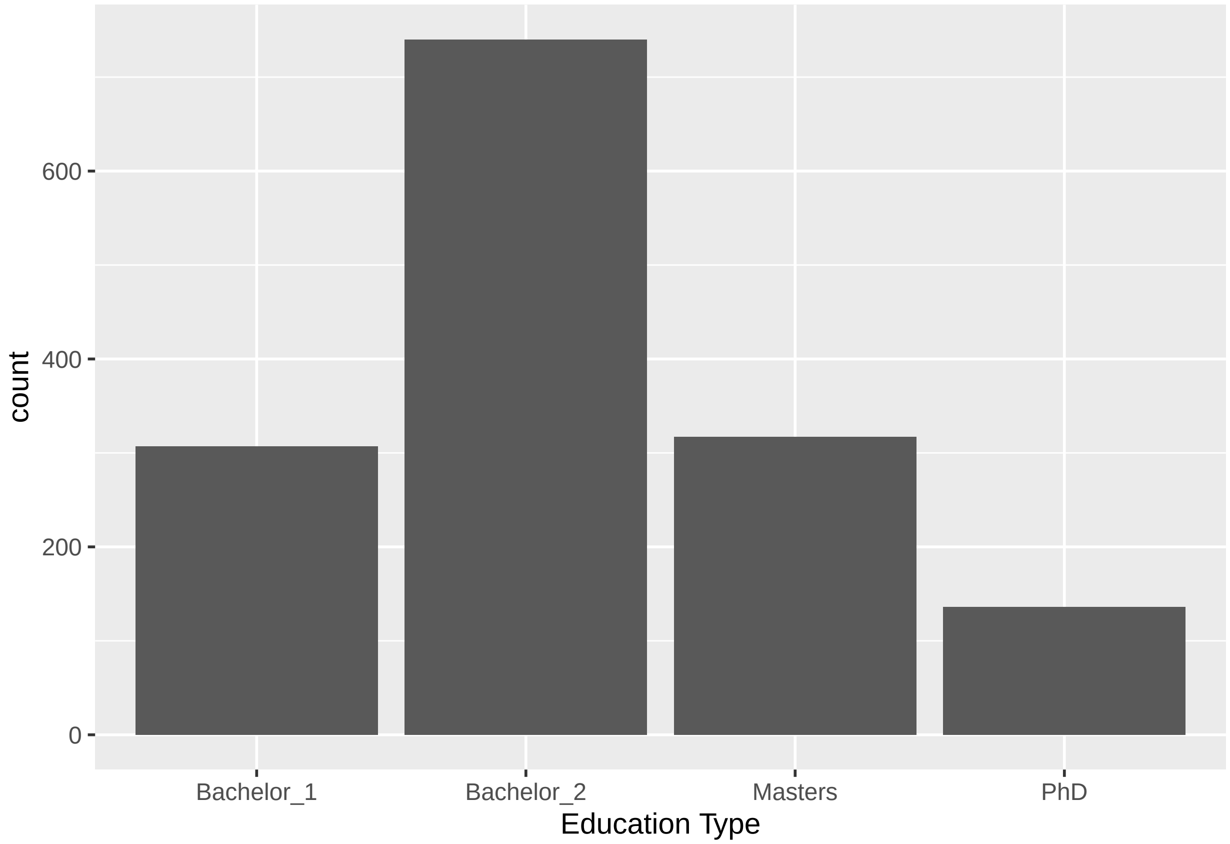
```r
hist("PersonalityScore", 25) +
  xlab("Personality Score")
```



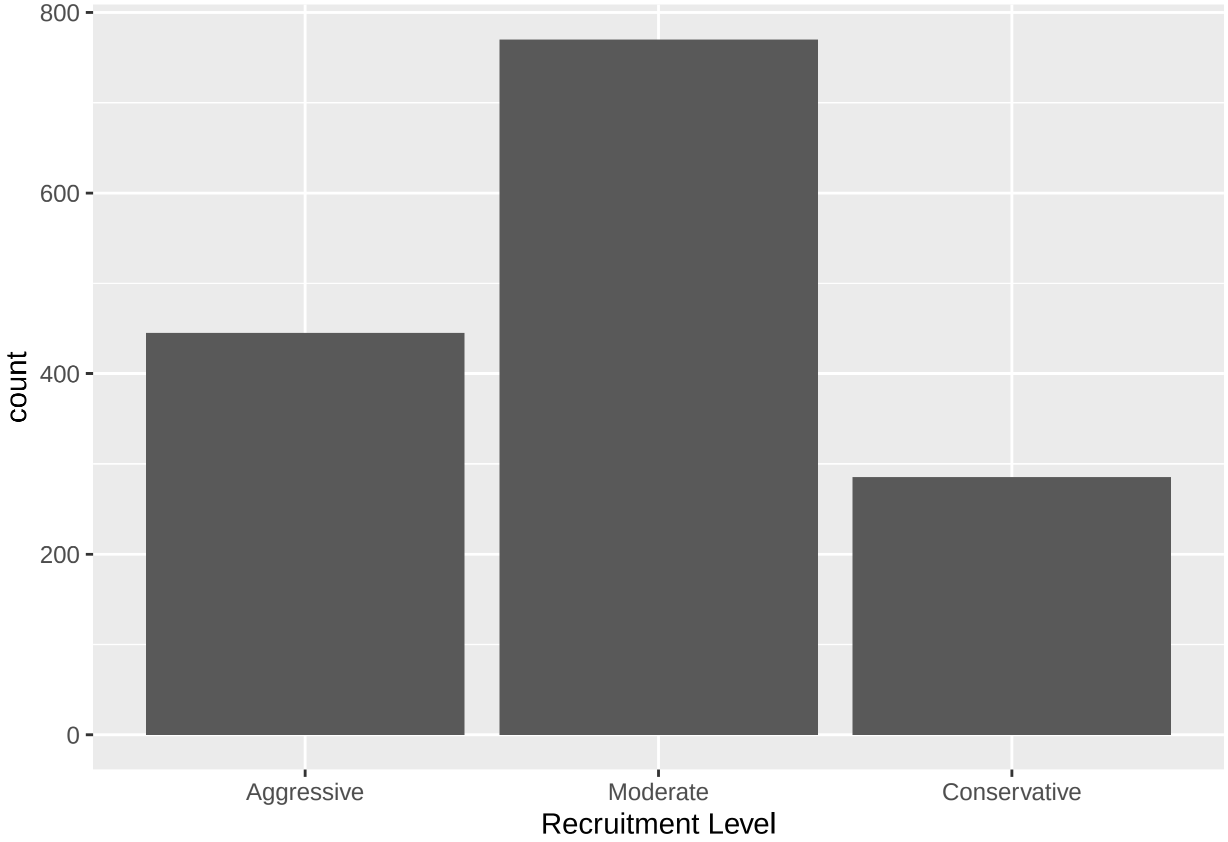
```
bar<- function(col){
  ggplot(data = recruitment) +
    geom_bar(aes(x = .data[[col]])) +
    xlab(col)
}

bar("Gender_name") +
  xlab("Sex")
```



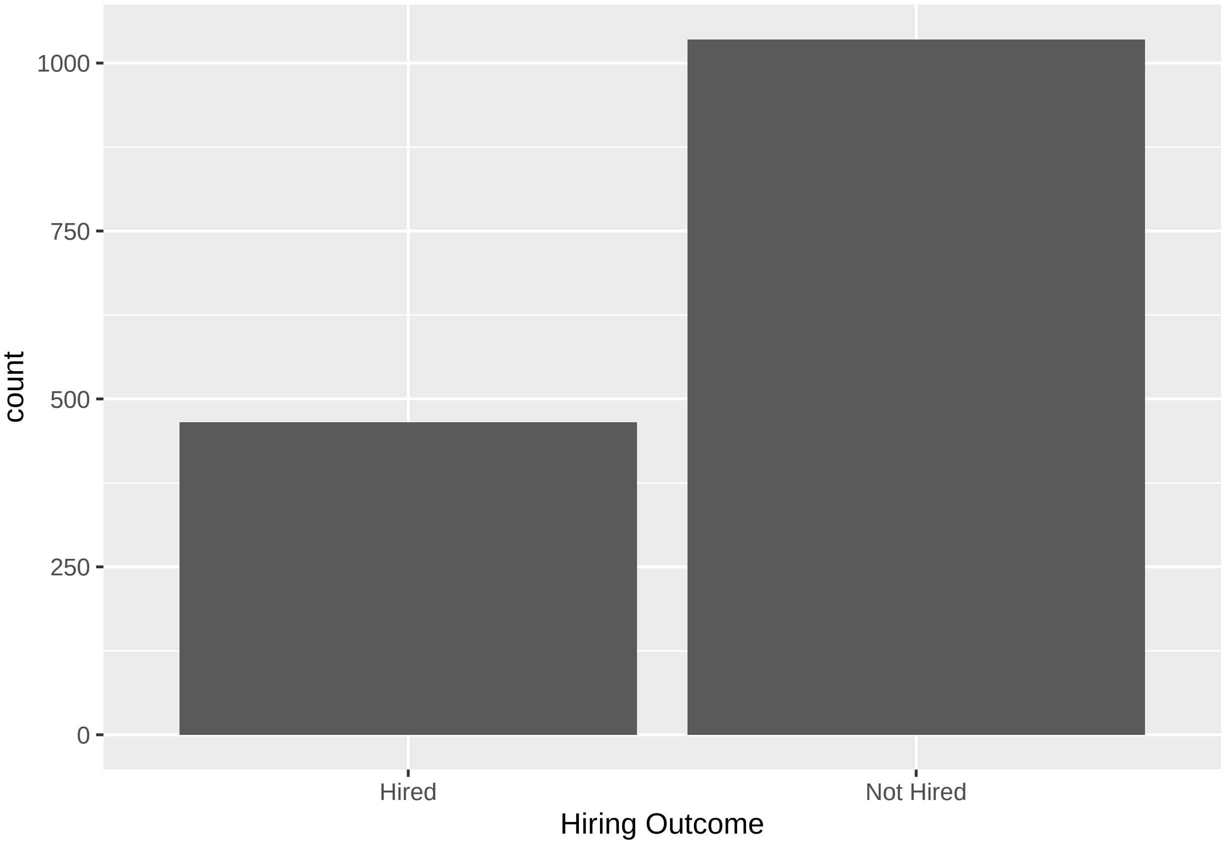
```r
bar("Education_name") +
  xlab("Education Type")
```



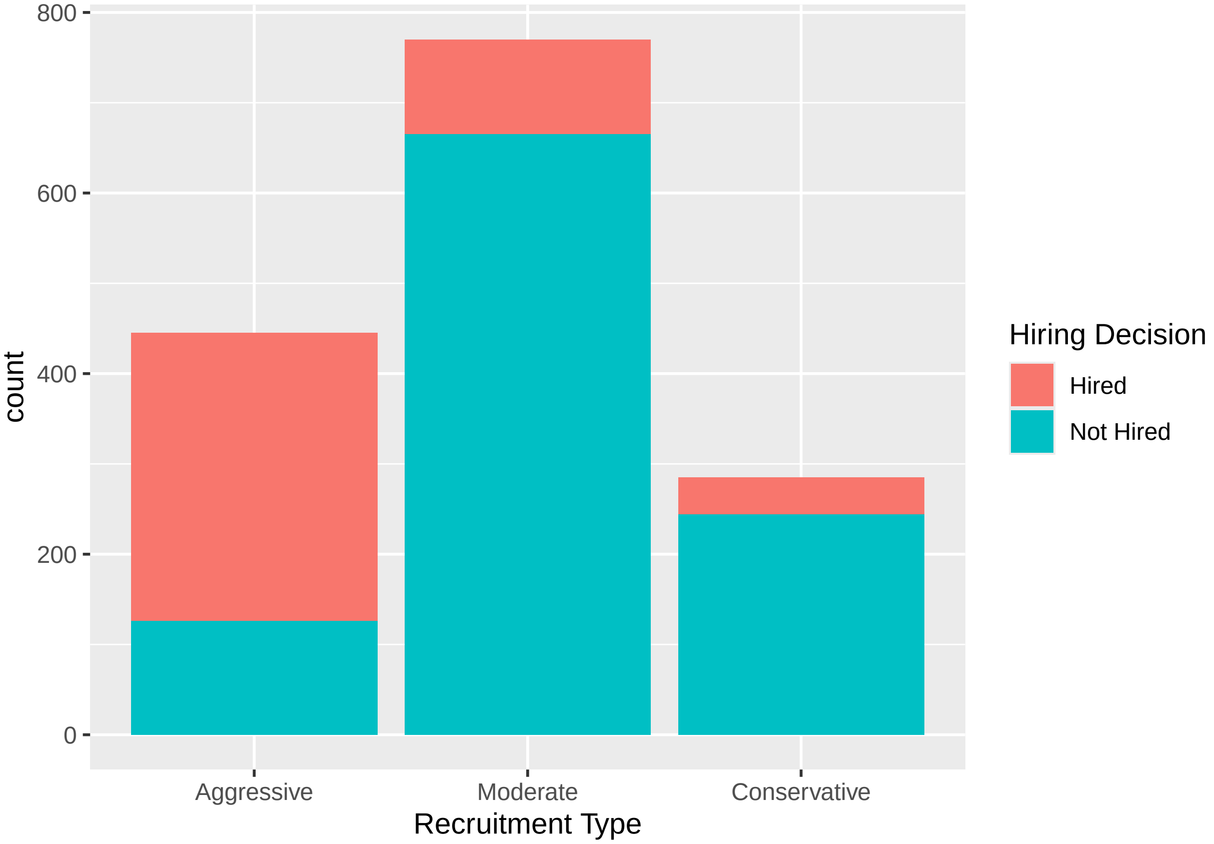
```r
bar("Recruitment_name") +
  xlab("Recruitment Level")
```



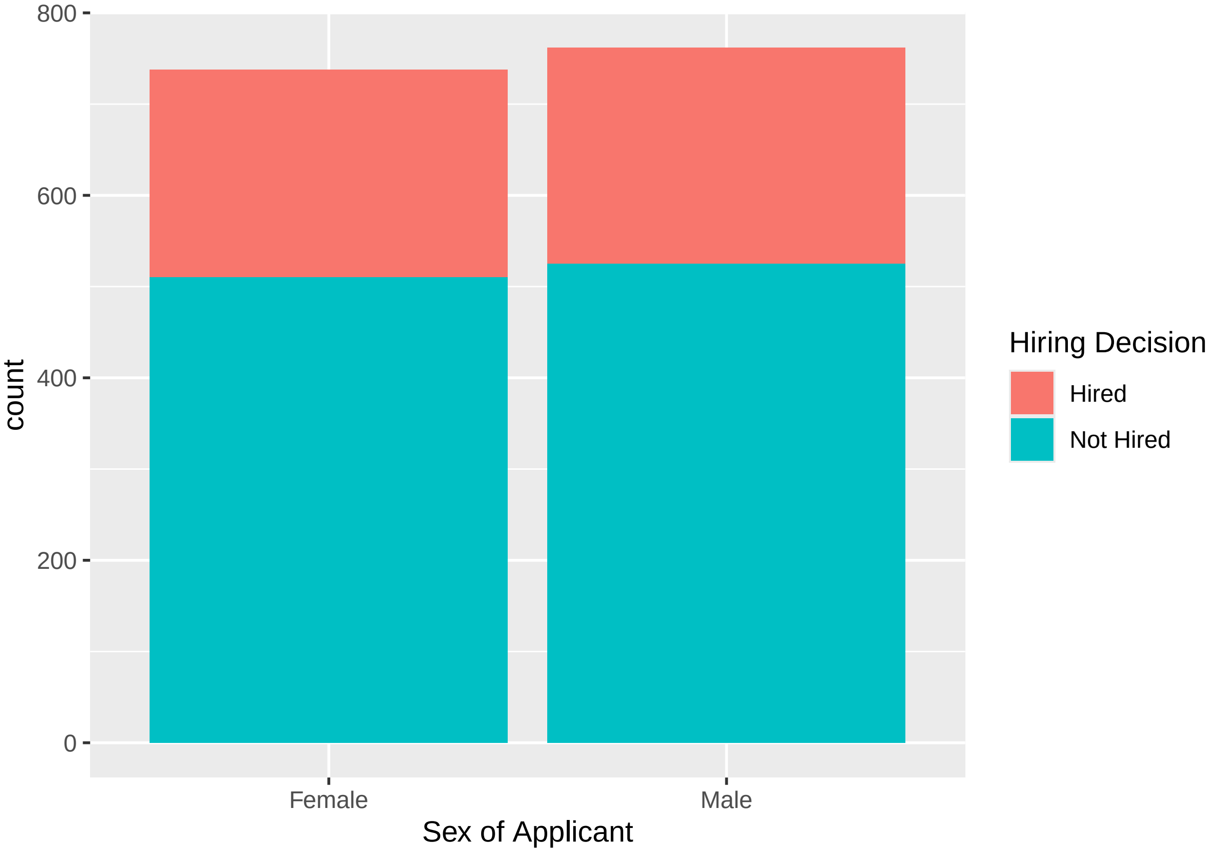
```
bar("Hiring_name") +
  xlab("Hiring Outcome")
```



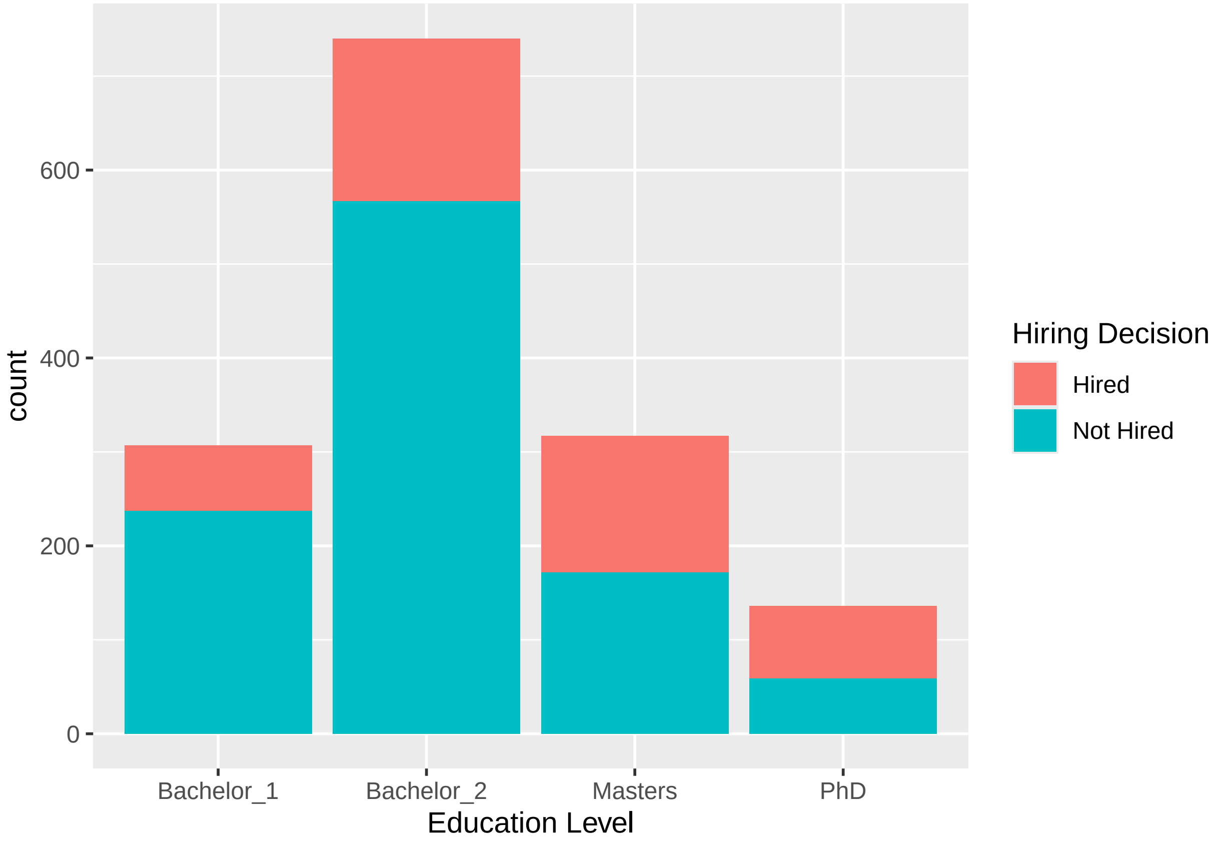
```
ggplot(data = recruitment) +
  geom_bar(aes(x = Recruitment_name, fill = Hiring_name)) +
  xlab("Recruitment Type") +
  guides(fill = guide_legend(title = "Hiring Decision"))
```



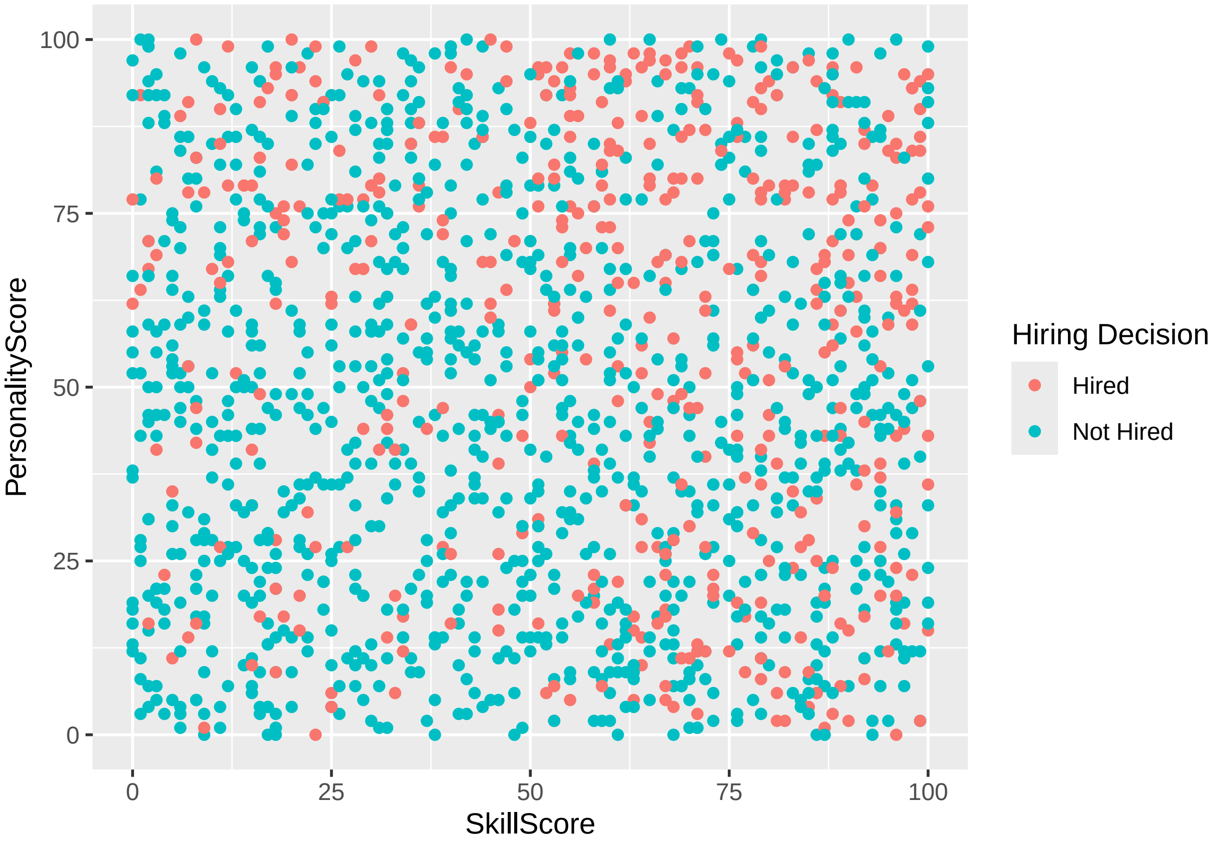
```r
ggplot(data = recruitment) +
  geom_bar(aes(x = Gender_name, fill = Hiring_name)) +
  xlab("Sex of Applicant") +
  guides(fill = guide_legend(title = "Hiring Decision"))
```



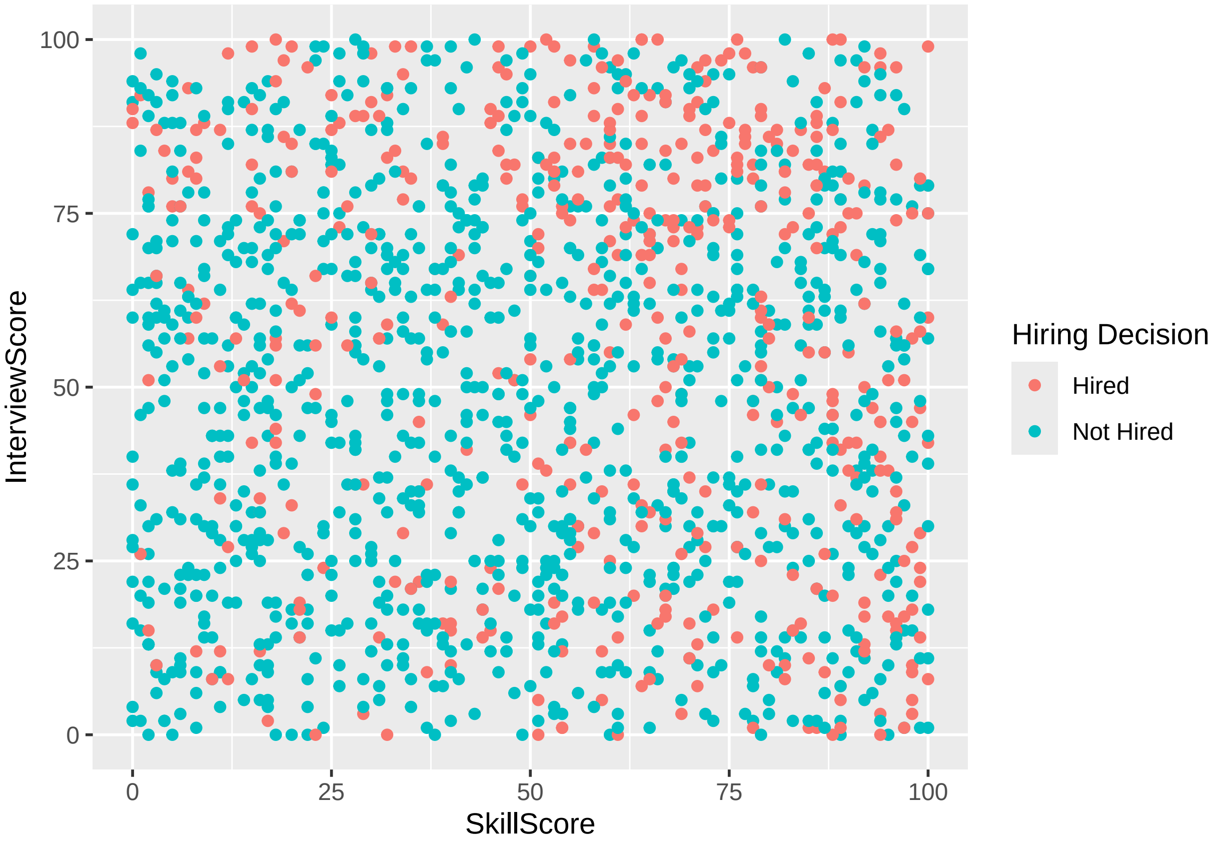
```
ggplot(data = recruitment) +
  geom_bar(aes(x = Education_name, fill = Hiring_name)) +
  xlab("Education Level") +
  guides(fill = guide_legend(title = "Hiring Decision"))
```



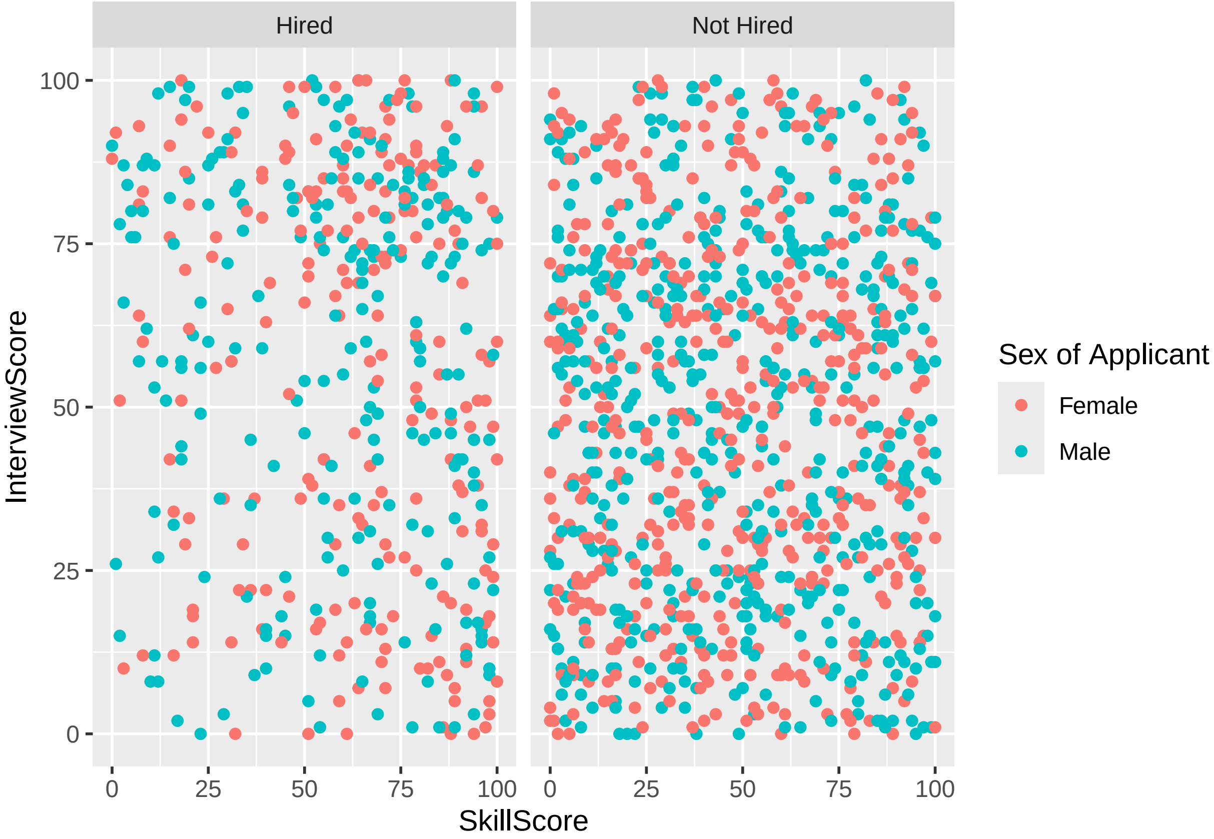
```
hiring_point<- function(x, y){
  ggplot(data = recruitment) +
    geom_point(aes(x = .data[[x]], y = .data[[y]], color = Hiring_name))
}

hiring_point("SkillScore", "PersonalityScore") +
  guides(color = guide_legend(title = "Hiring Decision"))
```



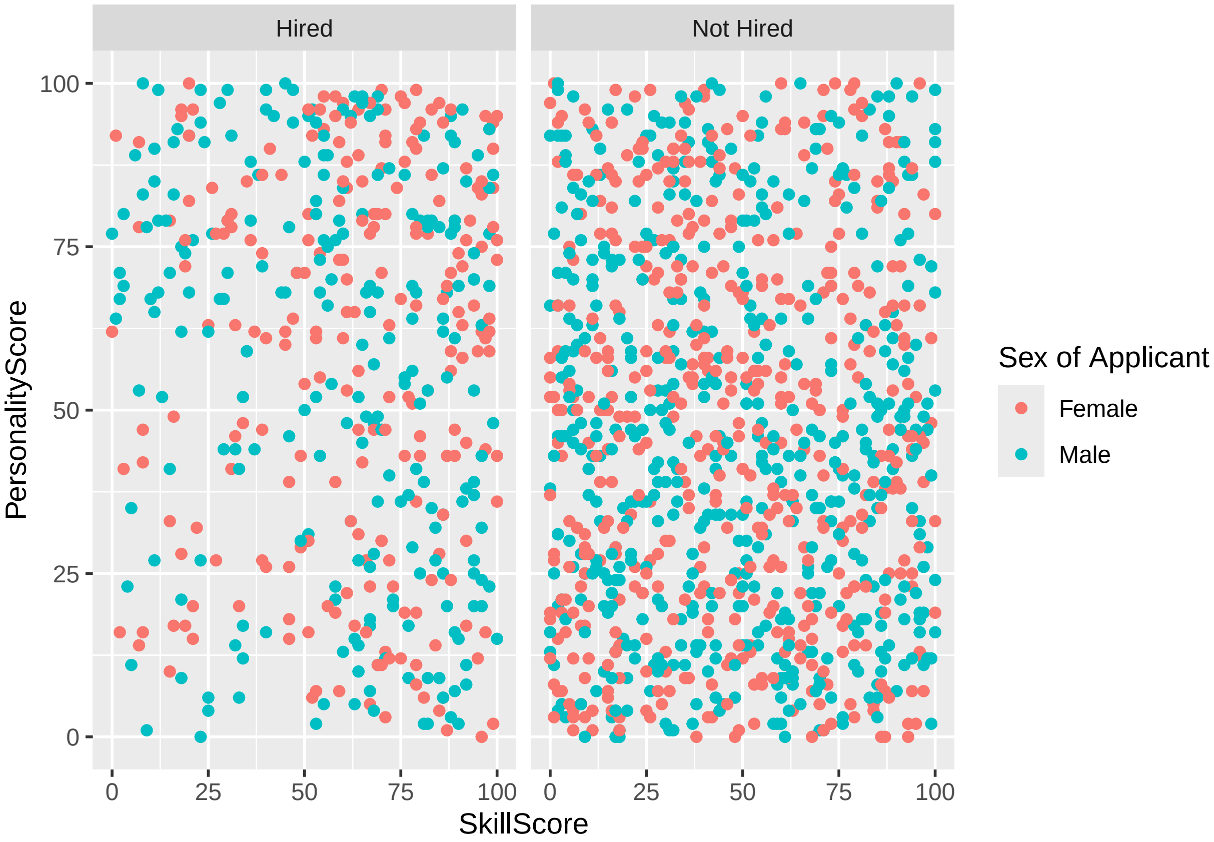
```
hiring_point("SkillScore", "InterviewScore") +
  guides(color = guide_legend(title = "Hiring Decision"))
```



```r
ggplot(data = recruitment) +
  geom_point(aes(x = SkillScore, y = InterviewScore, color = Gender_name))+
  facet_grid(~Hiring_name) +
  guides(color = guide_legend(title = "Sex of Applicant"))
```



```
ggplot(data = recruitment) +
  geom_point(aes(x = SkillScore, y = PersonalityScore, color = Gender_name)) +
  facet_grid(~Hiring_name) +
  guides(color = guide_legend(title = "Sex of Applicant"))
```



```
set.seed(720)

trainingRows <- createDataPartition(recruitment$Hiring_name, p = .75, list = FALSE)

recruit_train <- recruitment[trainingRows, ]
recruit_test <- recruitment[-trainingRows, ]
```

```
log_reg<- glm(HiringDecision ~ Age + Gender + EducationLevel + ExperienceYears + PreviousCompanies + Di

summary(log_reg)
```

```
##
## Call:
## glm(formula = HiringDecision ~ Age + Gender + EducationLevel +
##     ExperienceYears + PreviousCompanies + DistanceFromCompany +
##     InterviewScore + SkillScore + PersonalityScore + RecruitmentStrategy,
##     family = "binomial", data = recruit_train)
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -4.929373   0.677568  -7.275 3.46e-13 ***
## Age                  -0.007693   0.010976  -0.701    0.483
## Gender1               0.003851   0.196709   0.020    0.984
## EducationLevel2       0.248679   0.275474   0.903    0.367
## EducationLevel3       2.284181   0.323342   7.064 1.61e-12 ***
## EducationLevel4       2.605979   0.385030   6.768 1.30e-11 ***
## ExperienceYears       0.150380   0.023296   6.455 1.08e-10 ***
## PreviousCompanies     0.098294   0.069243   1.420    0.156
## DistanceFromCompany   0.001221   0.006804   0.180    0.858
## InterviewScore        0.028174   0.003870   7.279 3.35e-13 ***
## SkillScore            0.032259   0.003700   8.718  < 2e-16 ***
## PersonalityScore      0.025023   0.003598   6.955 3.53e-12 ***
## RecruitmentStrategy2 -4.274524   0.276639 -15.452  < 2e-16 ***
## RecruitmentStrategy3 -4.300213   0.340508 -12.629  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1394.12  on 1125  degrees of freedom
## Residual deviance:  692.71  on 1112  degrees of freedom
## AIC: 720.71
##
## Number of Fisher Scoring iterations: 6
```

```
exp(log_reg$coefficients)
```

```
##          (Intercept)                  Age              Gender1
##          0.007231034          0.992336997          1.003858292
##      EducationLevel2      EducationLevel3      EducationLevel4
##          1.282330490          9.817639576         13.544481127
##      ExperienceYears    PreviousCompanies  DistanceFromCompany
##          1.162275999          1.103287643          1.001222146
##       InterviewScore           SkillScore     PersonalityScore
##          1.028574795          1.032784980          1.025338967
## RecruitmentStrategy2 RecruitmentStrategy3
##          0.013918670          0.013565667
```

Combine both bachelors degrees into one level- not statistically significant No need for age or gender (seen in EDA and logistic model), previous companies and distance are not statistically significant.

What seems to be most important are aggressive recruiting strategy and education at masters level or above. (See exponentiated coefficients from logistic regression model)