Insights For An HR Manager Using Statistical Graphics Techniques

Eugeniah Arthur

November 19,2022

Contents

Introduction **2**

Background . . . 2

Objective . . . 2

Data Source . . . 2

Exploratory Data Analysis **3**

Data Preprocessing . . . 3

Data Cleaning . . . 3

Univariate plots . . . 3

Bivariate relationships . . . 8

Relationship btween Salary and department . . . 15

Predicting Salary **20**

RandomForest . . . 20

Gradient Boosting . . . 21

Summary and Recommendation for Human Resource Managers. **27**

Conclusion **27**

References **27**

Introduction

Background

In recent years a lot of companies are concerned with diversity and inclusion in the demographics of their associates. A lot of studies have shown that a diverse group in the company helps with productivity and leads to more job satisfaction. There have been calls for inclusion with regards to especially gender and race. Furthermore, fairness in salary is also another aspect in this movement that is critically looked at by both Human resource managers and major stakeholders to ensure the company is upholding such standards. In this analysis, insights with regards to gender, race , salary and other factors in a human resource data set will be found using both statistical graphics and some sophisticated machine learning algorithms.

Objective

1. Unravel insights from the gender, racial, absences and salary distribution of the data set
2. Find the relationship between salary and these other key variables
3. Find the most important factors that influence whether or not an individual's salary will be above the median

Data Source

The data was gotten from kaggle at this website. https://www.kaggle.com/datasets/rhuebner/human-resources-data-set. The data was created by Richard Huebner and Dr. Carla Patalano. It has 36 columns and 311 observations. For this analysis, the following variables will be of key interest: gender, race, married, performance scores,job position,department,absences, salary, recruitment source and employment status.

Exploratory Data Analysis

Data Preprocessing

Data Cleaning

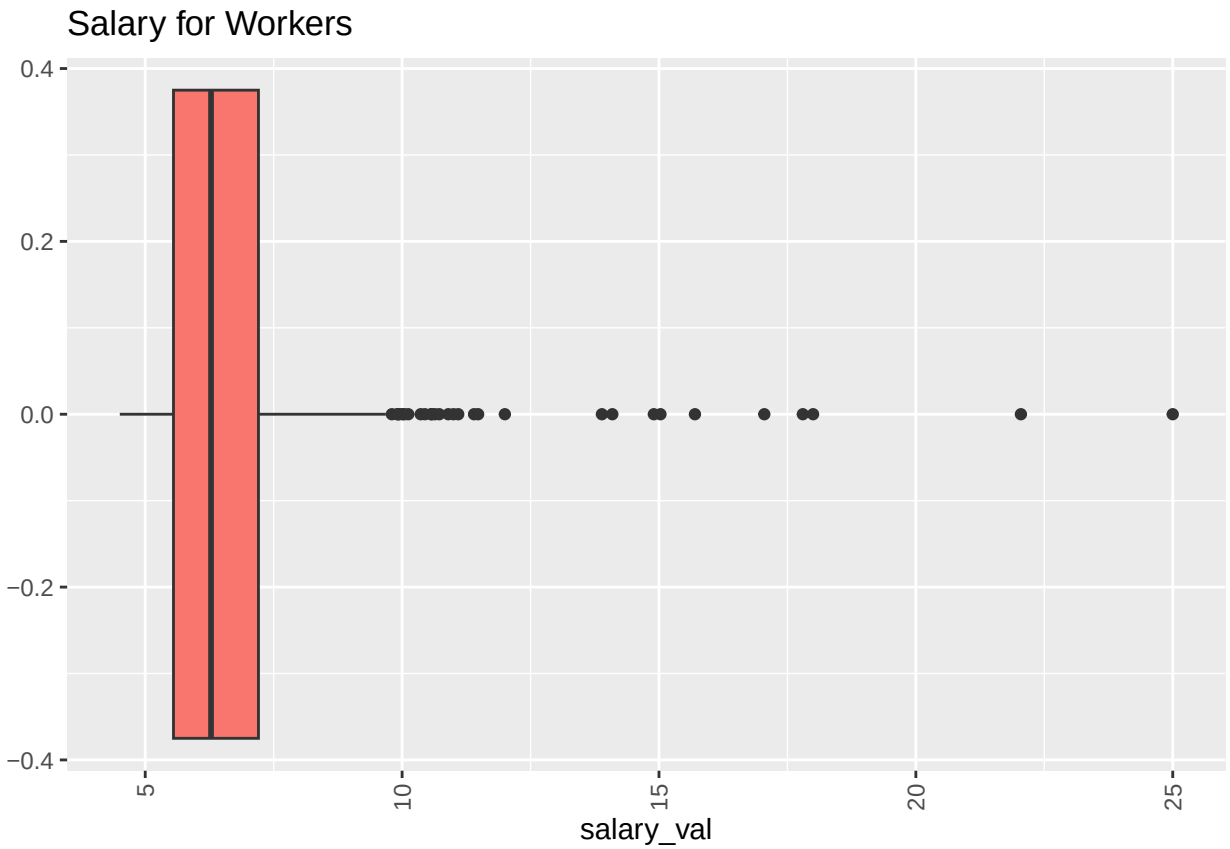
Under data cleaning, the key features were selected. Also, the original job position variable was not well specified so it was cleaned to return the right jobs.

Univariate plots

Distribution of salary

The histogram of the salary data set was plotted using different bin widths. The epanechnikov density was overlayed on it to show the density of the salary data.

The Robust rule of twelve gives a great depiction of the distribution of the salary. The salary variable will be explored to see more trends in the data. The epanchnikov plot does a great job at fitting the kernel of the distribution . Even though it was not able to hit the high frequencies in the highest bin. THe median salary is 62180.




```
## [1] 62810
```

```
## [1] 25
```

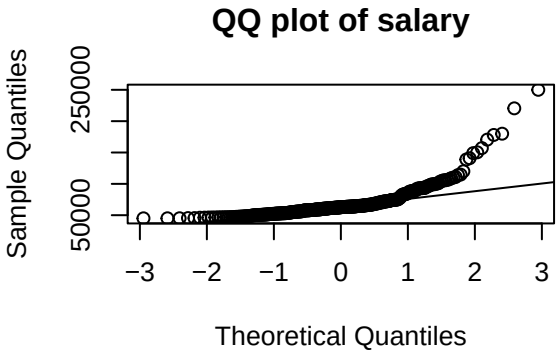
The distribution of the salary of workers is skewed right and this is typical of most salary data set. The data has clusters in them showing different groupings. This may be as a result of the various positions held by the people in the company.

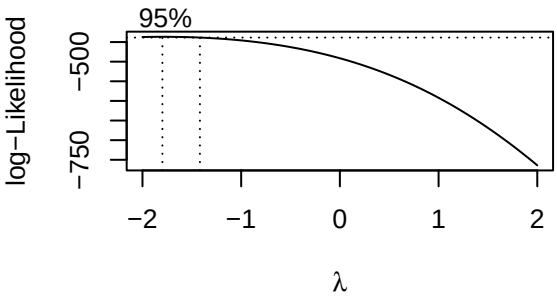
The typical salary is about 63000 per yer with some salary as high as 250000.The outliers are from about 100000. There are about 25 people in this company that have pay more than 100000.

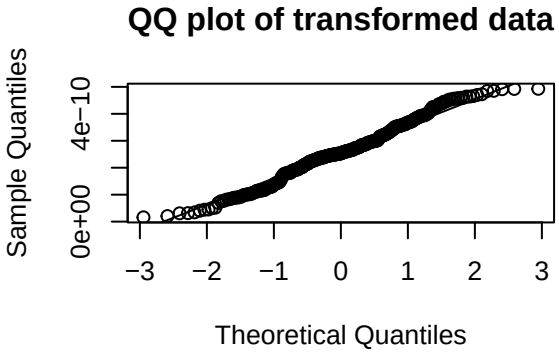
There is very little spread in the data with a lot of outliers at the upper tail. Aside the outliers, the salary seems to be perfectly symmetric .

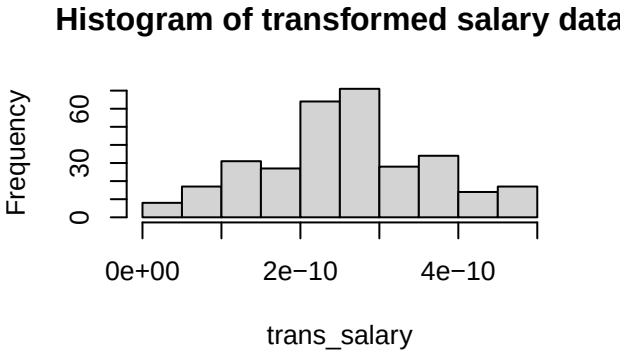
Transforming the salary data

```
## [1] -1.79798
```



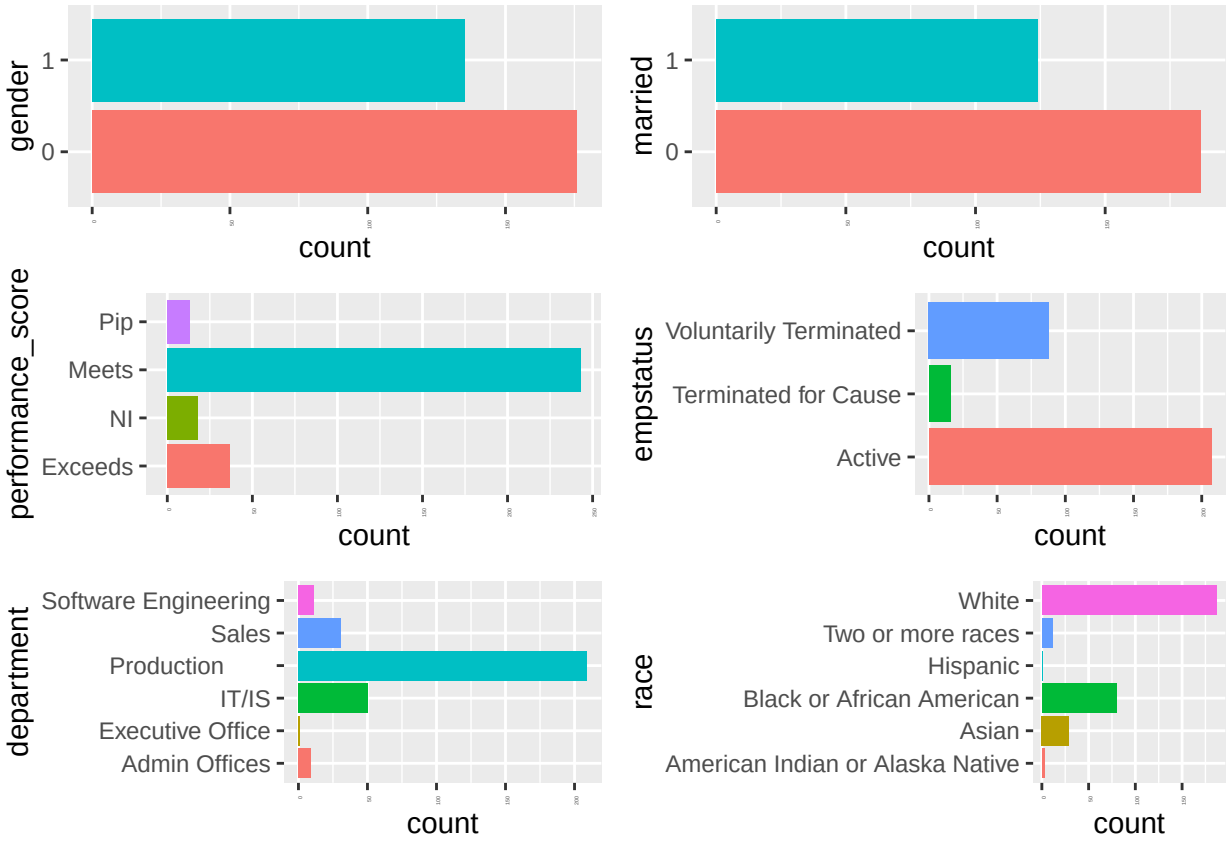


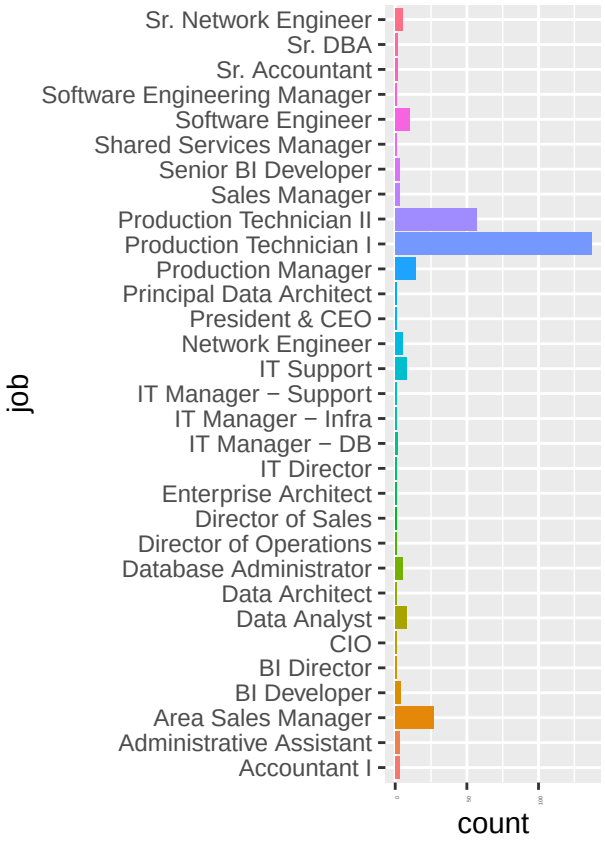


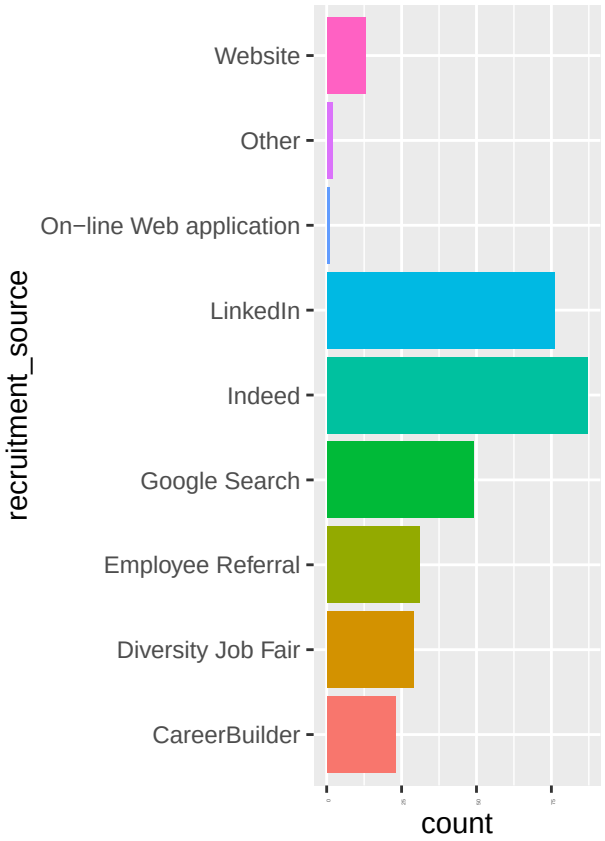


The qq plot of the original salary data shows that it is not normally distributed and hence to normalize this we use the box cox transformation to find the best way to do this, the optimal lambda is -1.8 which is almost -2. Hence, we use the transformation 1/xˆ2. The transformed data looks normally distributed and hence this transformed data can be used for further analysis

Distribution of factor variables





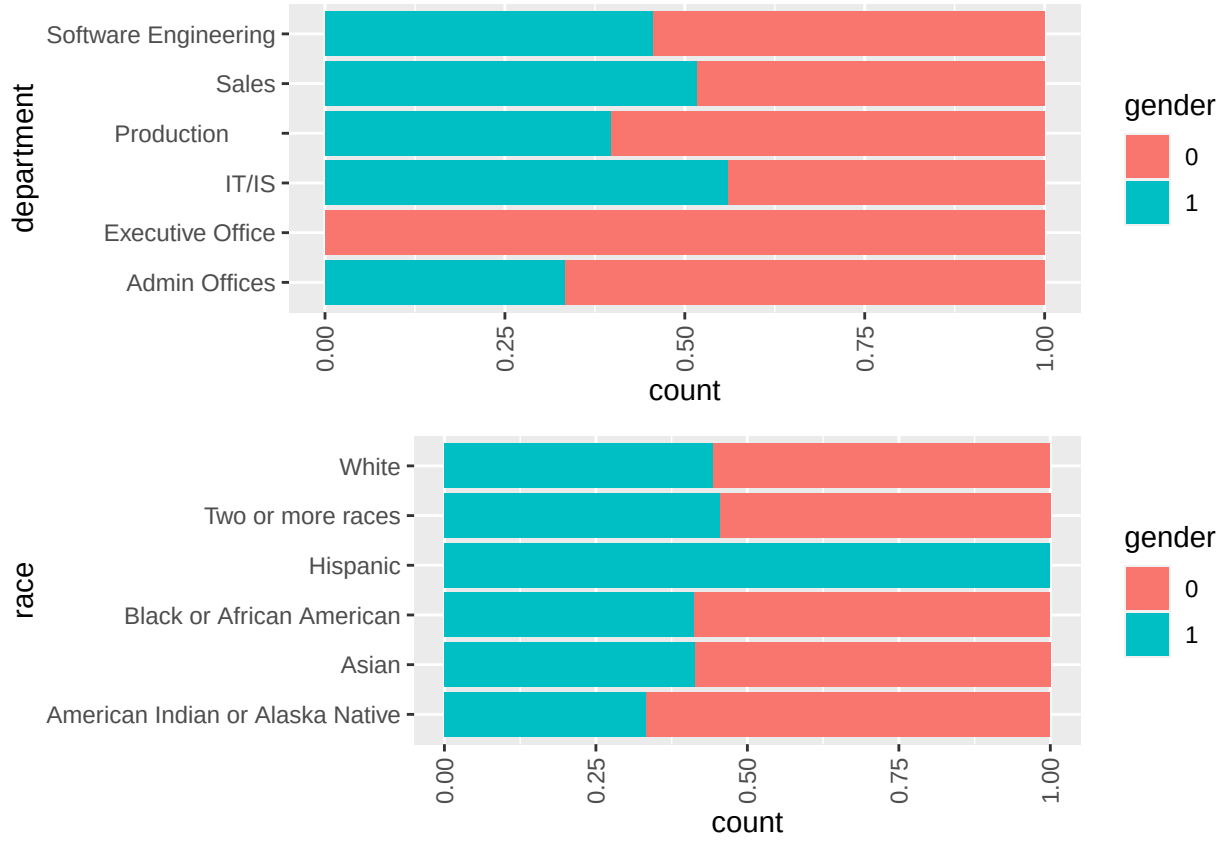


From the plots, the department with the most employees is the production department.There are a lot of production technicians compared to other jobs. Most individuals in this company meets expectation in terms of performance score. Furthermore, a lot of individuals were recruited through Indeed , followed by Linked In. The online web application was not very popular. The company is not very diverse. It has a large proportion of white folks. However, this demographics compared to a lot of other companies in the world looks somewhat better. Almost as many as half of the active individuals voluntarily terminated their job. There are more females than males in the organisation. Also, more than half of the individuals are not married. The dataset has very little diversity. Most of the race are white, with black or african american second almost a half of the number of whites in the group and there is almost an equal distribution of both genders. Hispanic are the fewest people in this organization and they are all male.

This informs me that the company is a production intensive company with lost of female workers.

Bivariate relationships

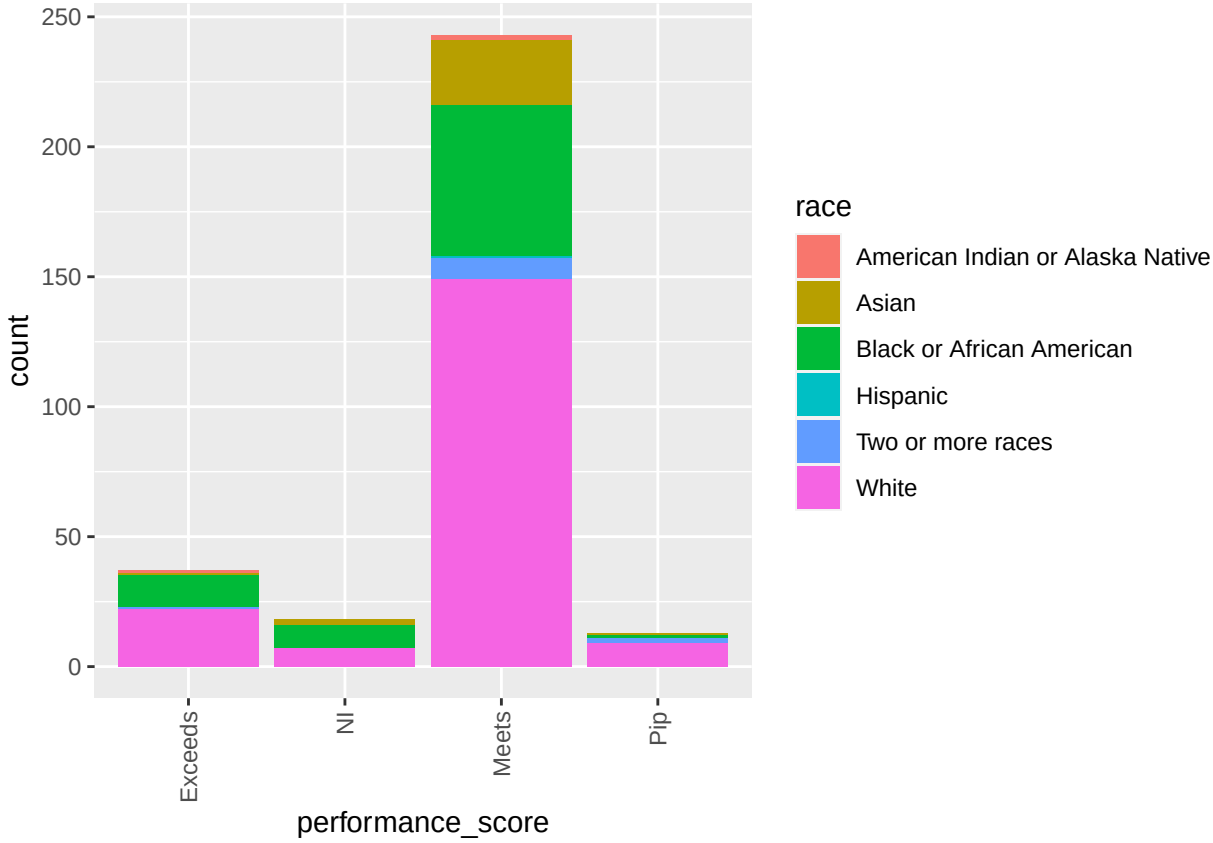
Gender distribution based on race and department



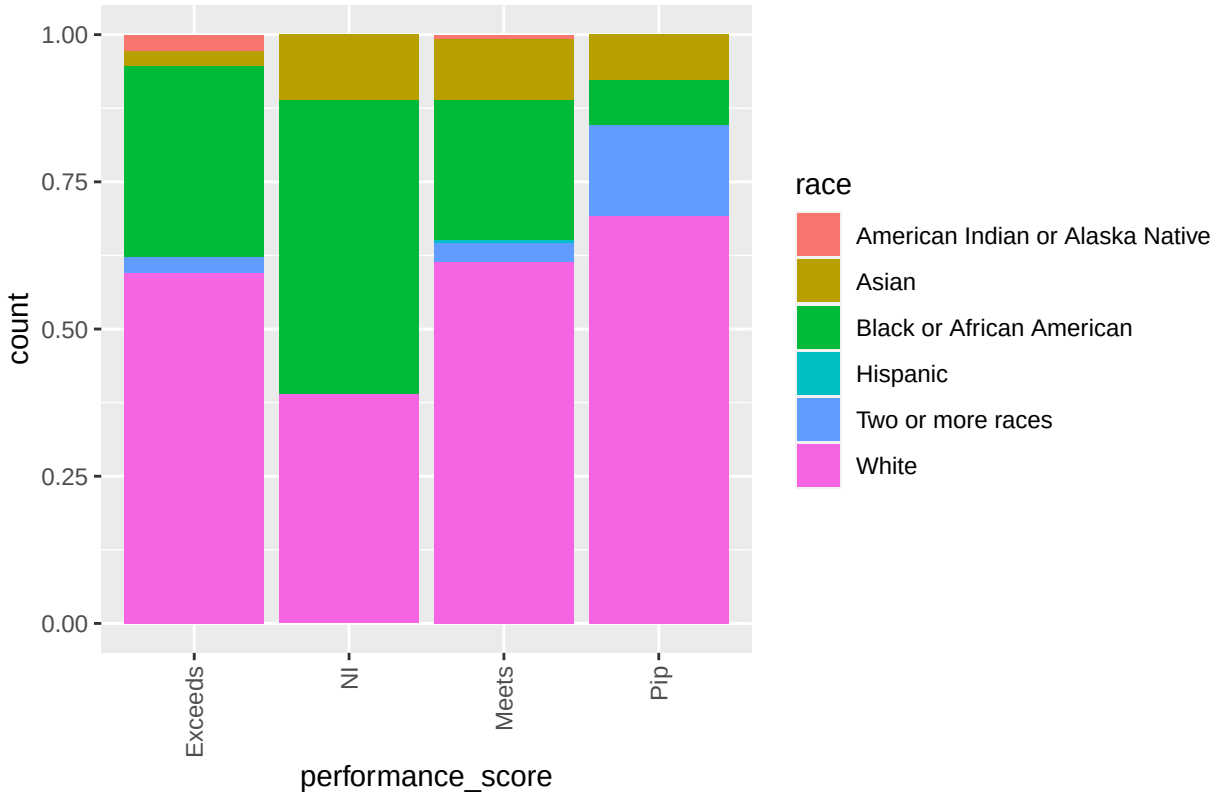
There are more females in the production department. The leader who makes up the executive office is a female.There seems to be a 50- 50 proportion of each gender group in each race except hispanics who are all male. Also, in the IT/IS and sales team, there are more males than females. All the other department have more females. Sales have an equal proportion of gender.

All the Hisanics are male. We have an equal proportion of gender for both the white and two or more races. For all the other races, we have more females.

Relationship between performance score and gender



Conditional Probablity Plot for Performance_score based on race

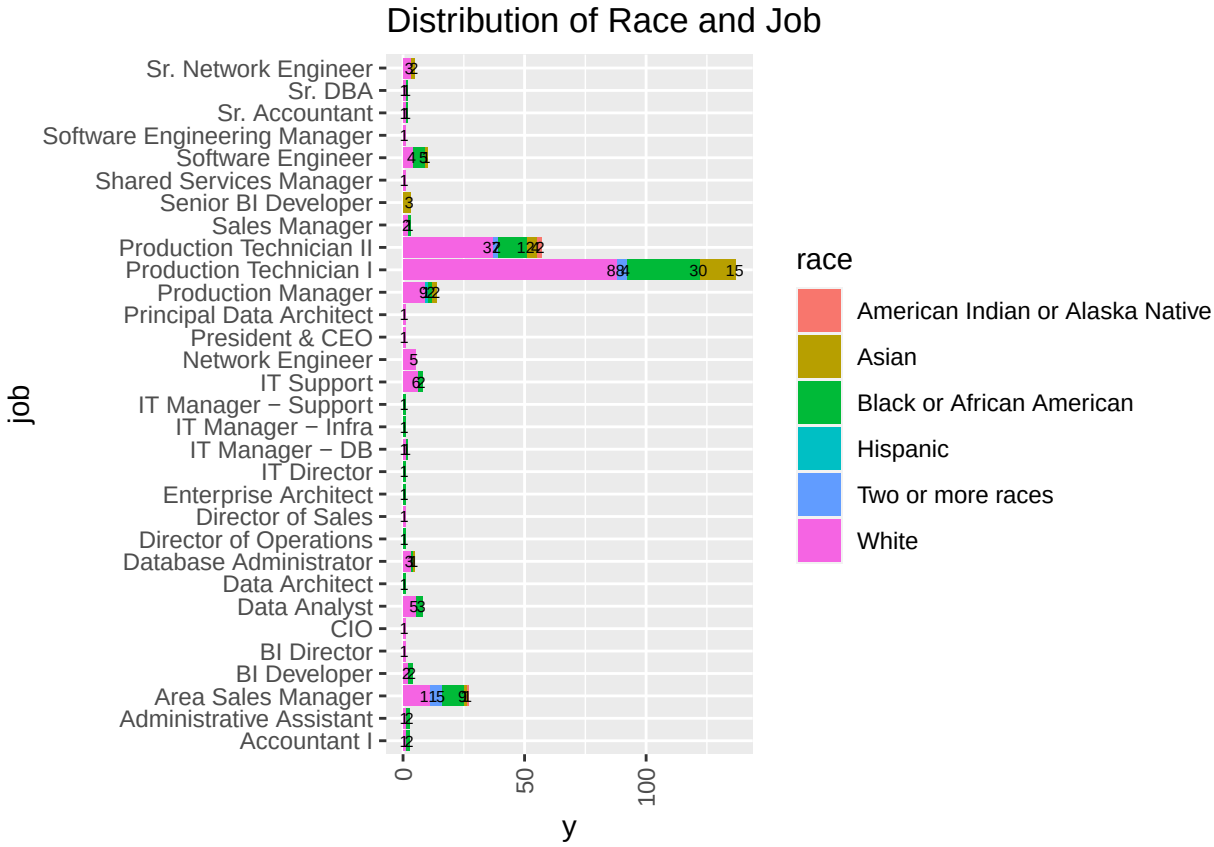


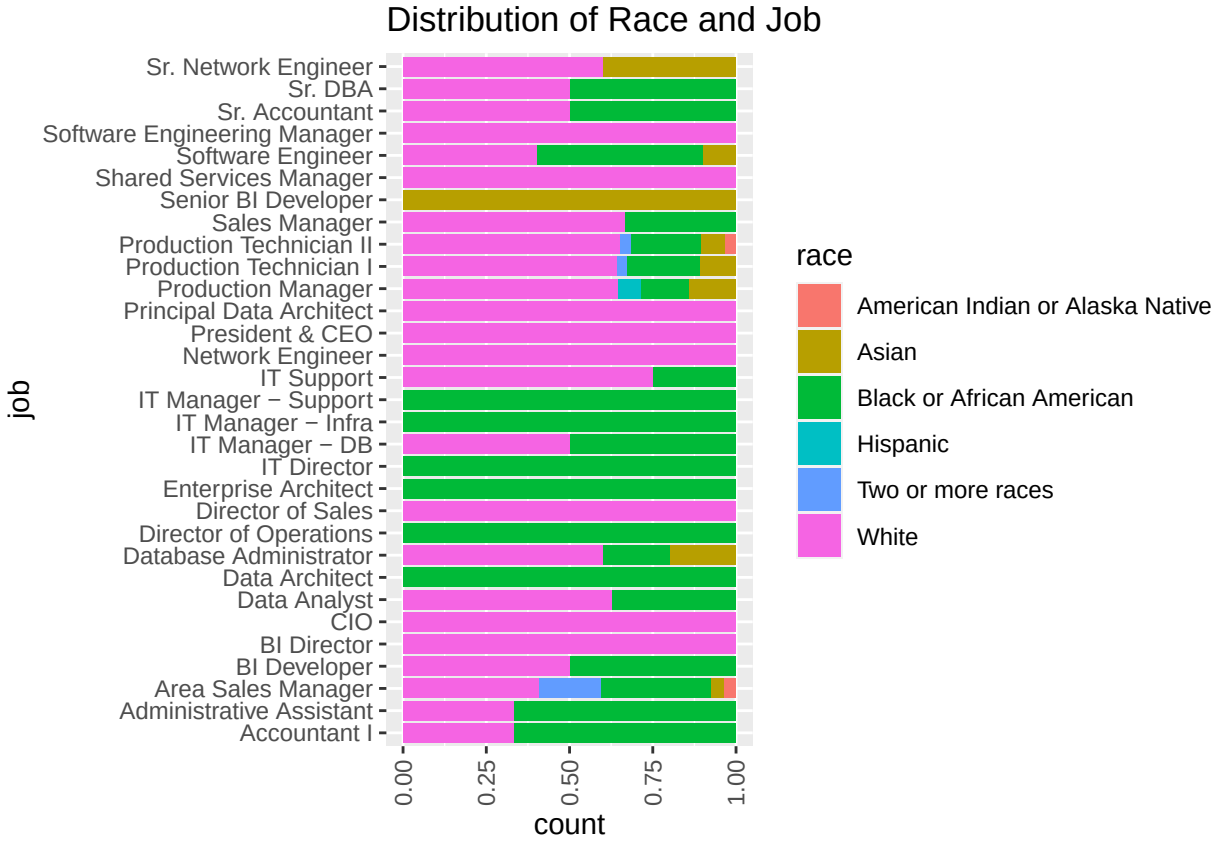
From the dataset, a lot of the individuals fully meets expectation. The proportion of individuals that need improvement and are on the personal improvement program is quite few.

Also, a large proportion of two or more races were in the PIP group relative to all the other performance score group. A large proportion of those that needed improvements were black.

A large proportion of american indians exceeded expectations relative to the other performance score groups and the others exceeded expectation. However, we should be mindful of their small numbers.

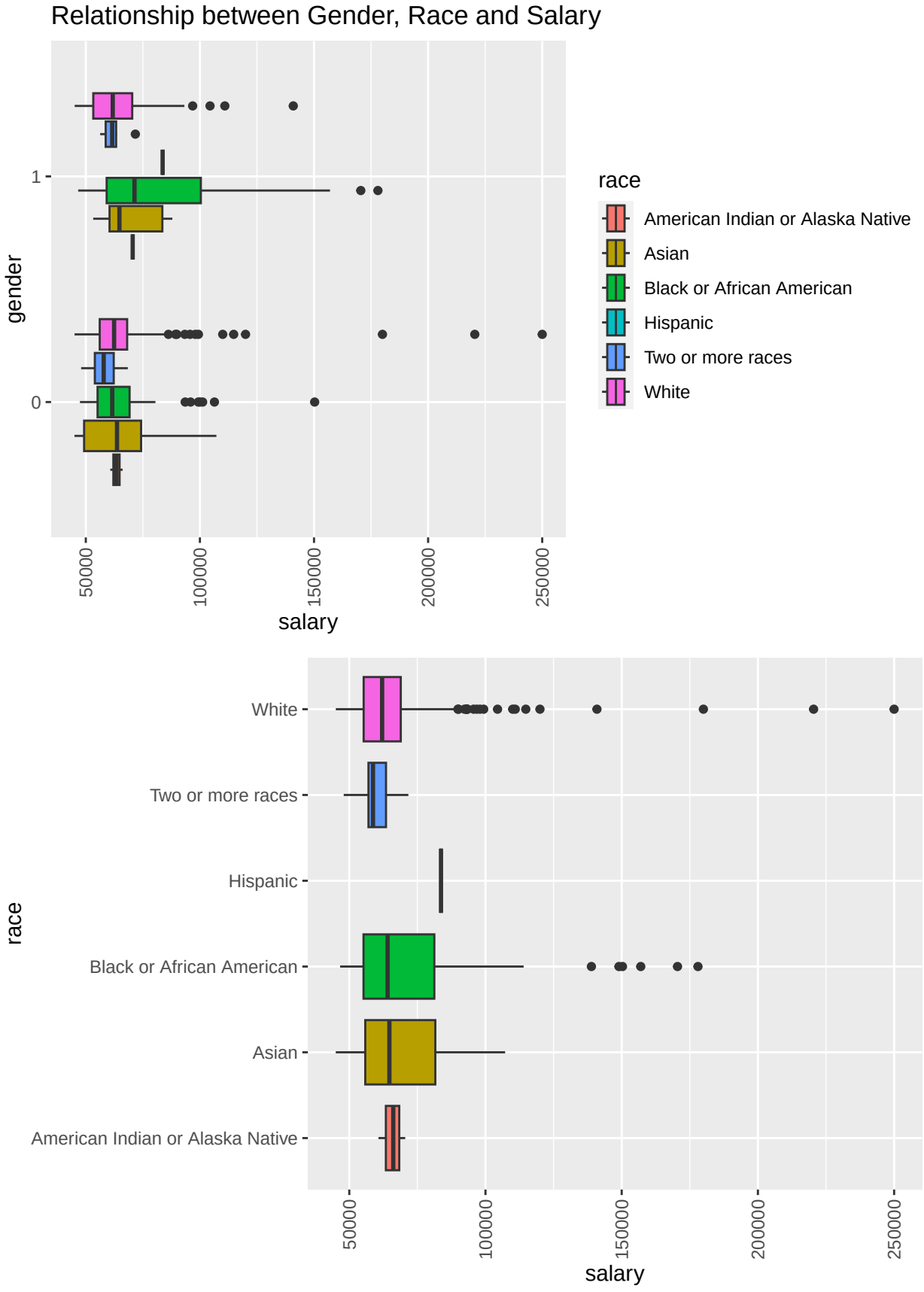
Distribution of race and job descriptions

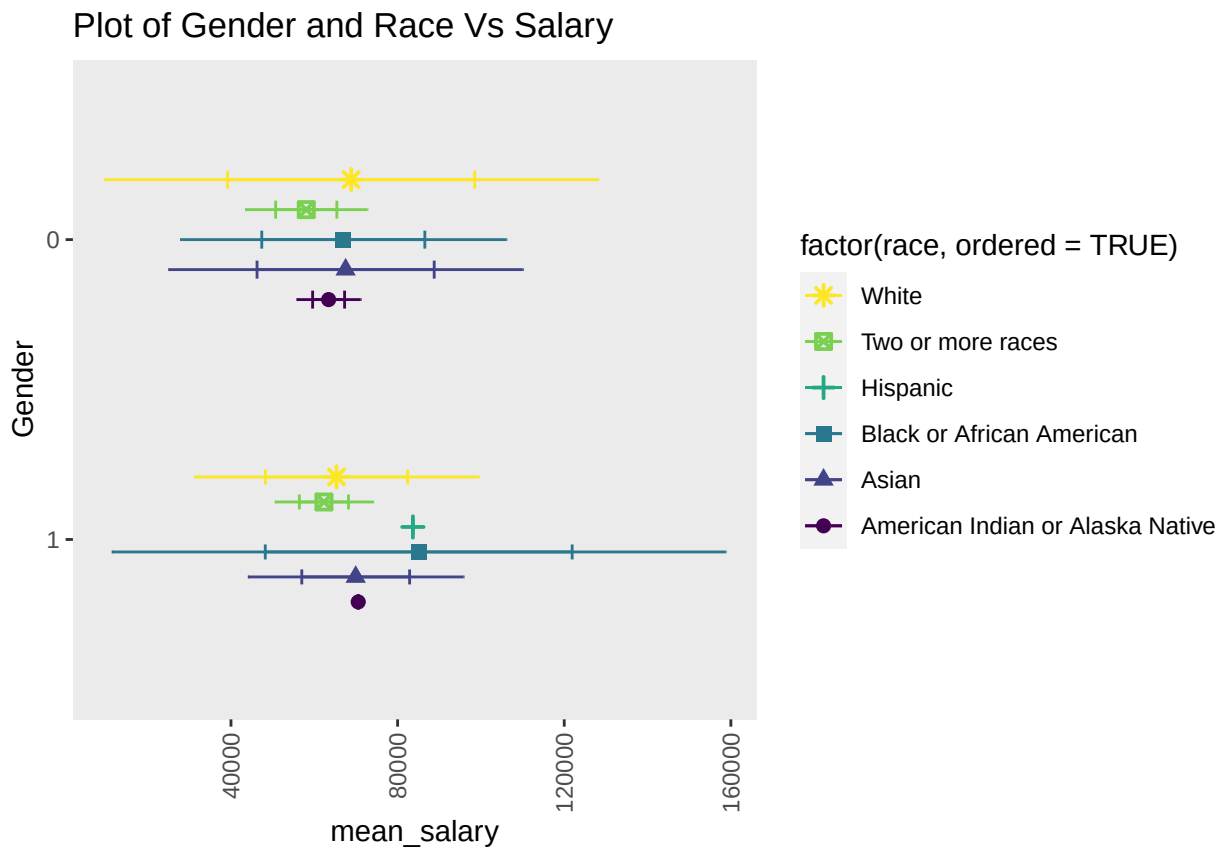




From this plot the president and CEO are seen to be white, there are 3 senior BI developers and they are all Asians, Also, all the 5 network engineers are white. Hence, this job position will be an area of improvement for HR managers when they are filling this role next. Most of the people in this HR data base are Production Technician I and Production Technician II. Production Technician II has quite a mix of individuals from different races. Most of the other roles have just one individual in the position.

Gender Race and Salary





From this plot , The typical salary of all the races is pretty much close together.There are a lot of high salaried individuals in the white and black categories. especially in the white race, this may be as a result of the fact that senior executives among other high paid individuals are white. It is also interesting to note that most of the people with really high salary probably like the CEO are all females.

The spread in the distribution of salary for Asians in the company is quite wide for both females and males. For the white people in the data, aside the very large outliers, their spraed is quite close to each other. The distribution of the salary for the black males is quite spread out than the black females.

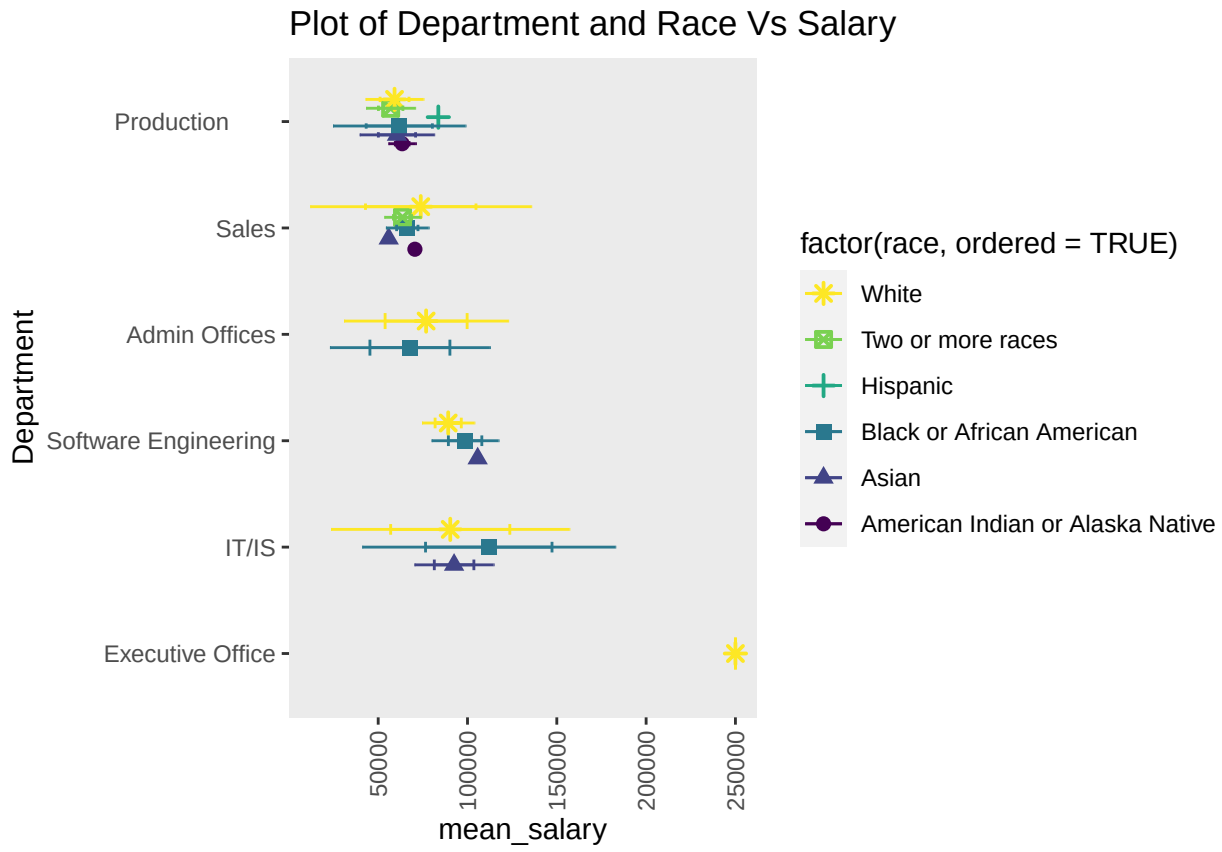
Also, from the dot whisker chat, there is no significant difference in the mean salaries even though average salary for black males is higher than all the other races and gender.

This gives an indication of fair salary in the data.

Relationship btween Salary and department

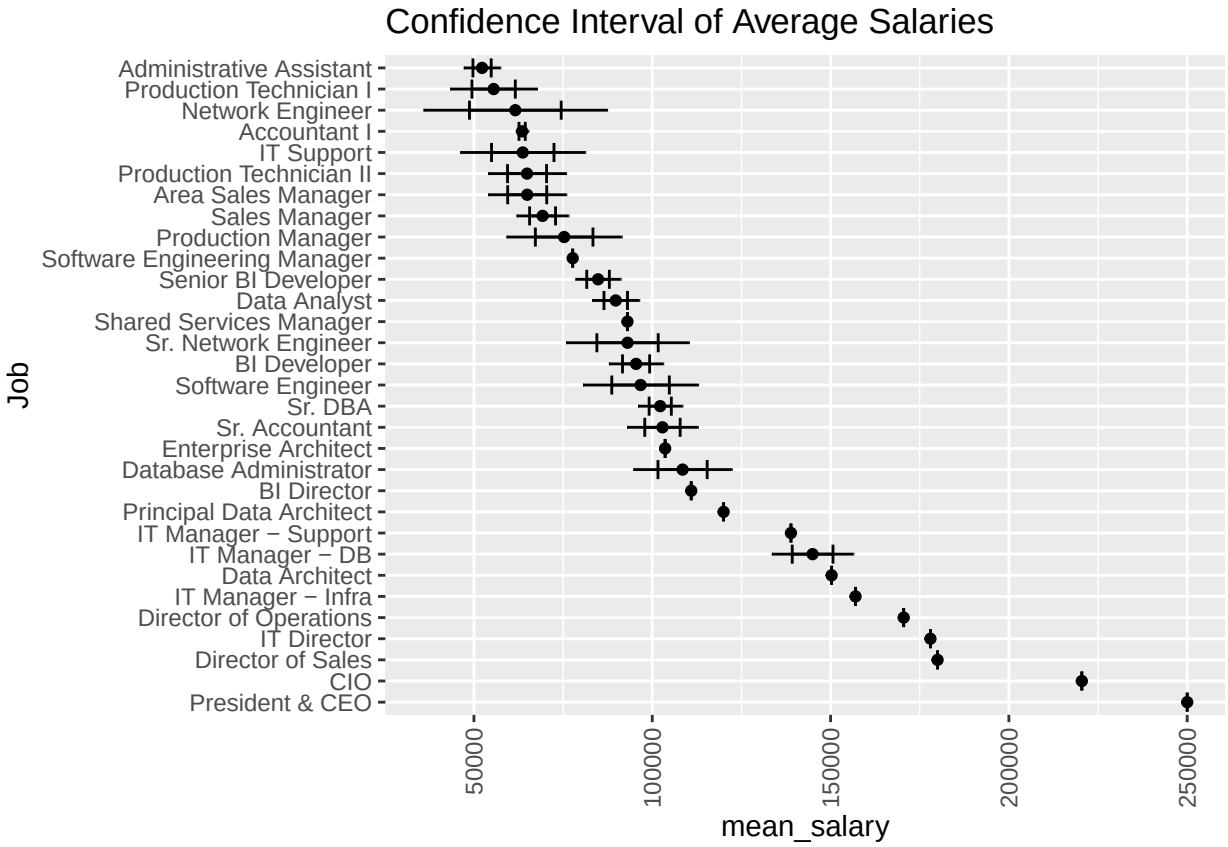
In this company it is noticed that the Chief executive officer is paid around 250000 which is the highest paid salary in the company.This is followed by individuals in the IT/IS department. Its interesting to know that department with the most people is the production department but it has very few variability in their salary with 3 outliers may be the production managers and the head of that department.T The administrative office though had a bit os skewness to it indicating that there may be some salary imbalances to be investigated there.

Thus, we can say that the pay is quite fair in the production team.



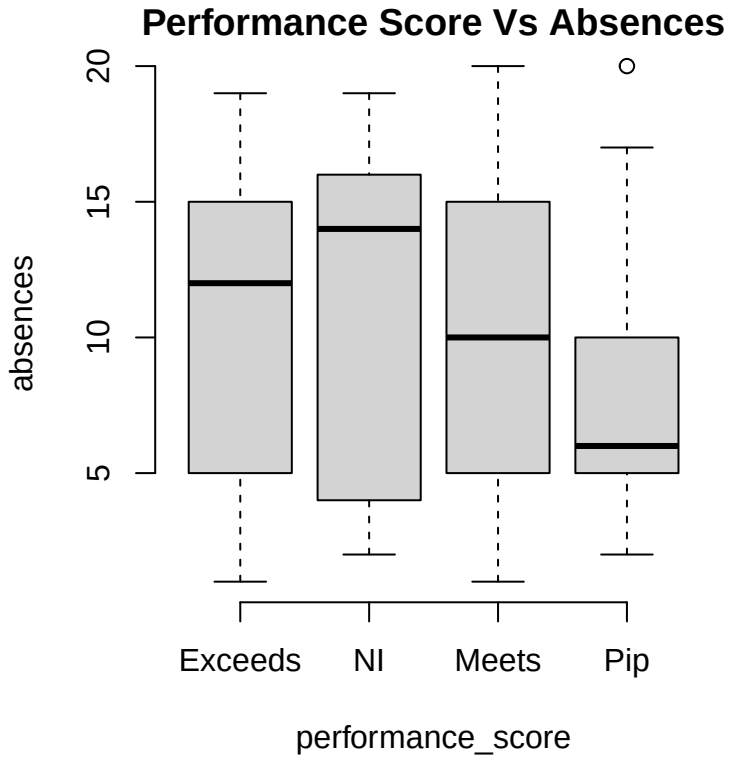
In this figure, we see the confidence interval of the mean salary for each of the departments. Even though for each department, we have different job descriptions, aside the executive office, there does not seem to be a difference in the average salary for different departments. Furthermore, there is no significant difference between the salaries within each department for the races that are present in that department. This shows that there seems to be fair salary distribution even in the administrative office which has only white and black races.

Investigating difference in salary for job position



From this plot, you can see depending on the job title, an associate's salary is significantly different from the other. The least paid workers are the administrative assistants. But their salaries are not significantly higher than Production Technician I and Network Engineer which has a much wider confidence interval than the others. In the plot from IT support to production manager,there no significant difference in the salaries. This plot shows the trends that is usually seen in every cooperation where individual's salary are based on their job positions.. Data Analysts job salaried lies around 80000 compare to Data Architects that is around 130000.

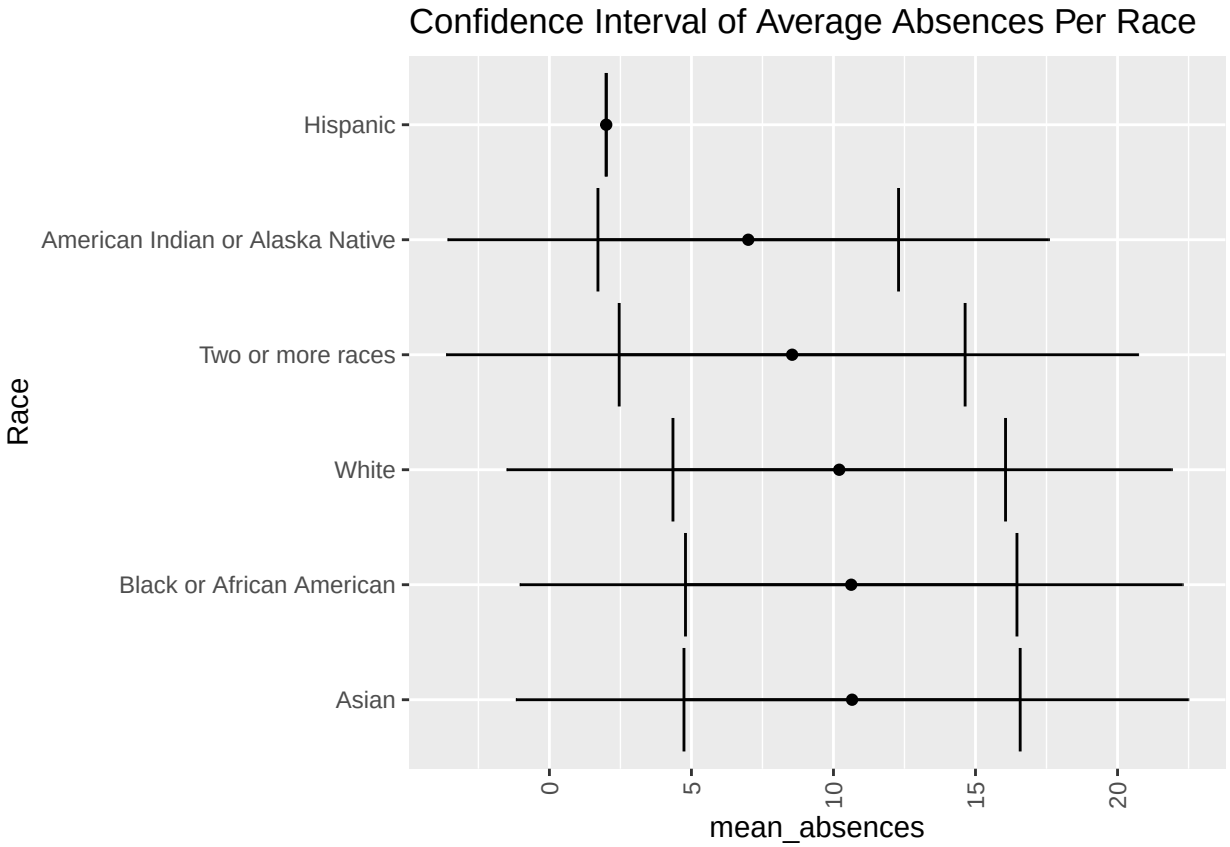
Absences , Race and Performance score.



This plot shows that those on personal improvement have much lower absences than the rest even though we still have one outlier.This is probably because they are on probation and are being monitored carefully.

Those that need improvement have the most absences, these people may be well on the way to being included to the personal improvement program.

Those that exceeds expectation have the second largest average absences .



It is of interest to know if some races are popular for absenteeism at work, from this analysis, there does not seem to a be a significant difference in the average absences of individuals based on their race.

Predicting Salary

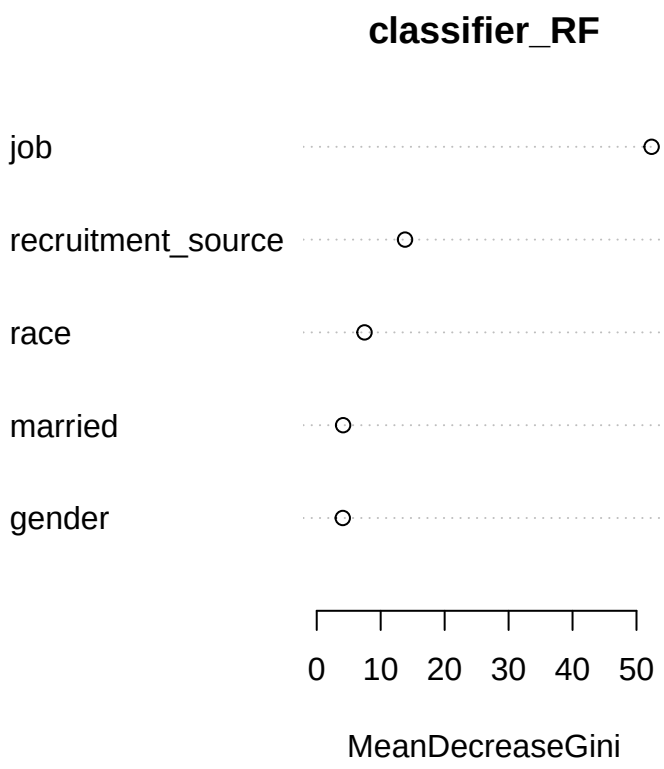
In this part of the analysis, the most important variable that can be used to predict whether or not an individual will have a salary above or below the median is investigated. The random forest and the gradient boosting ensemble method will be used for this analysis. All salaries below the median will be considered as the 0 class and 1 will be for the alternative. Gender, race , job,recruitment source and married will be the variables that will be considered in predicting whether or not an individual's salary will fall below the median. I selected these variables because these are always preknown before a job starts. Also, if there is any disparities in these, it would help the HR manager see where opportunities of fair salaries could be.

The data is split into 70% train and 30% test

```
## Loading required package: lattice
```

RandomForest

```
## Confusion Matrix and Statistics
##
##    y_pred
##      0  1
##   0 37 11
##   1 11 35
##
##                Accuracy : 0.766
##                  95% CI : (0.6674, 0.8471)
##     No Information Rate : 0.5106
##     P-Value [Acc > NIR] : 3.264e-07
##
##                   Kappa : 0.5317
##
##  Mcnemar's Test P-Value : 1
##
##             Sensitivity : 0.7609
##             Specificity : 0.7708
##          Pos Pred Value : 0.7609
##          Neg Pred Value : 0.7708
##              Prevalence : 0.4894
##          Detection Rate : 0.3723
##    Detection Prevalence : 0.4894
##       Balanced Accuracy : 0.7659
##
##        'Positive' Class : 1
##
```



THe randomforest model shows the accuracy at 76.6%. Furthermore, the sensitivity is 76.1% and specificity is 77.0% . This shows that the model does a good job in predicting the various classes.

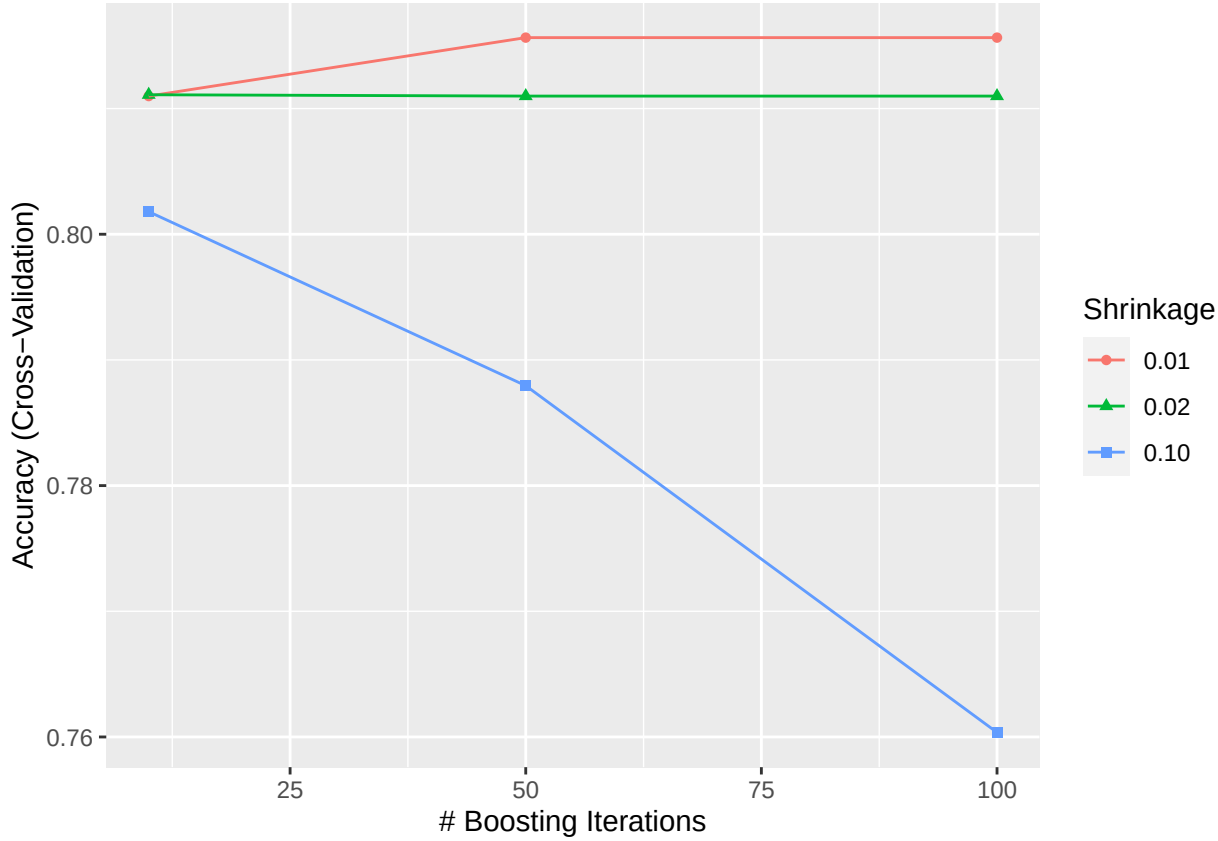
From the variable importance plot, the job, recruitment source and race are the top 3 variables that influenced whether an individual will have a salary above the median.

Gradient Boosting

XGBoost Tuning

```
## eXtreme Gradient Boosting
##
## 217 samples
##   5 predictor
##   2 classes: '0', '1'
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 173, 174, 173, 174, 174
## Resampling results across tuning parameters:
##
##   eta   nrounds  Accuracy   Kappa
##   0.01   10      0.8109937  0.6216086
##   0.01   50      0.8156448  0.6306639
##   0.01  100      0.8156448  0.6306639
##   0.02   10      0.8110994  0.6215730
```

```
##   0.02   50      0.8109937  0.6214273
##   0.02  100      0.8109937  0.6214273
##   0.10   10      0.8017970  0.6030996
##   0.10   50      0.7879493  0.5755753
##   0.10  100      0.7603594  0.5206726
##
## Tuning parameter 'max_depth' was held constant at a value of 7
## Tuning
##
## Tuning parameter 'min_child_weight' was held constant at a value of 1
##
## Tuning parameter 'subsample' was held constant at a value of 0.6
## Accuracy was used to select the optimal model using the largest value.
## The final values used for the model were nrounds = 50, max_depth = 7, eta
##  = 0.01, gamma = 0, colsample_bytree = 0.6, min_child_weight = 1 and
##  subsample = 0.6.
```




```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 36  8
##          1 12 38
##
##                Accuracy : 0.7872
##                  95% CI : (0.6907, 0.8649)
```

```
##     No Information Rate : 0.5106
##     P-Value [Acc > NIR] : 2.894e-08
##
##                   Kappa : 0.575
##
##  Mcnemar's Test P-Value : 0.5023
##
##             Sensitivity : 0.8261
##             Specificity : 0.7500
##          Pos Pred Value : 0.7600
##          Neg Pred Value : 0.8182
##              Prevalence : 0.4894
##          Detection Rate : 0.4043
##    Detection Prevalence : 0.5319
##       Balanced Accuracy : 0.7880
##
##        'Positive' Class : 1
##
```

The output of the cross validation tuning of the parameters shows that the optimal learning rate is 0.02. Also, the optimal number of decision trees in the final model has been selected to be 50.

```
##                         variable mean_abs_shap
## 1:  job.Production.Technician.I    0.48693990
## 2: job.Production.Technician.II    0.09199421
## 3:    recruitment_source.Indeed    0.03887958
## 4:   recruitment_source.Website    0.02384784
## 5:       job.Area.Sales.Manager    0.02108971
```

The shap plot shows that job is the most important variable in predicting whether a salary will fall above the median. It is noticed that for the job position, whether or not the person is a Production Technician I , Technician II or Area Sales Manager has the most influence on whether or not an individual will have a salary below the median salary.

One can see that being in the Production Technician I job position is associated with higher chance of getting a median salary below the median and vice versa. The same is true for Area Sales Manager and Technician II but the impact of not being an Area Sales Manager on predicting whether or not your salary is above the median is much lower than Technician I and II.

Furthermore, if you are recruited through website and linkedIn there is a possibility of your salary being lower than the median salary. This is because most of the people in these category do lower salaried jobs. Interestingly those, that were recruited though Indeed had the most influence and have a high chance of being paid higher.

Being Asian may result in you having a salary higher than the median salary, this may be because more Asians hold high paid salary jobs. the impact of this is very little though.

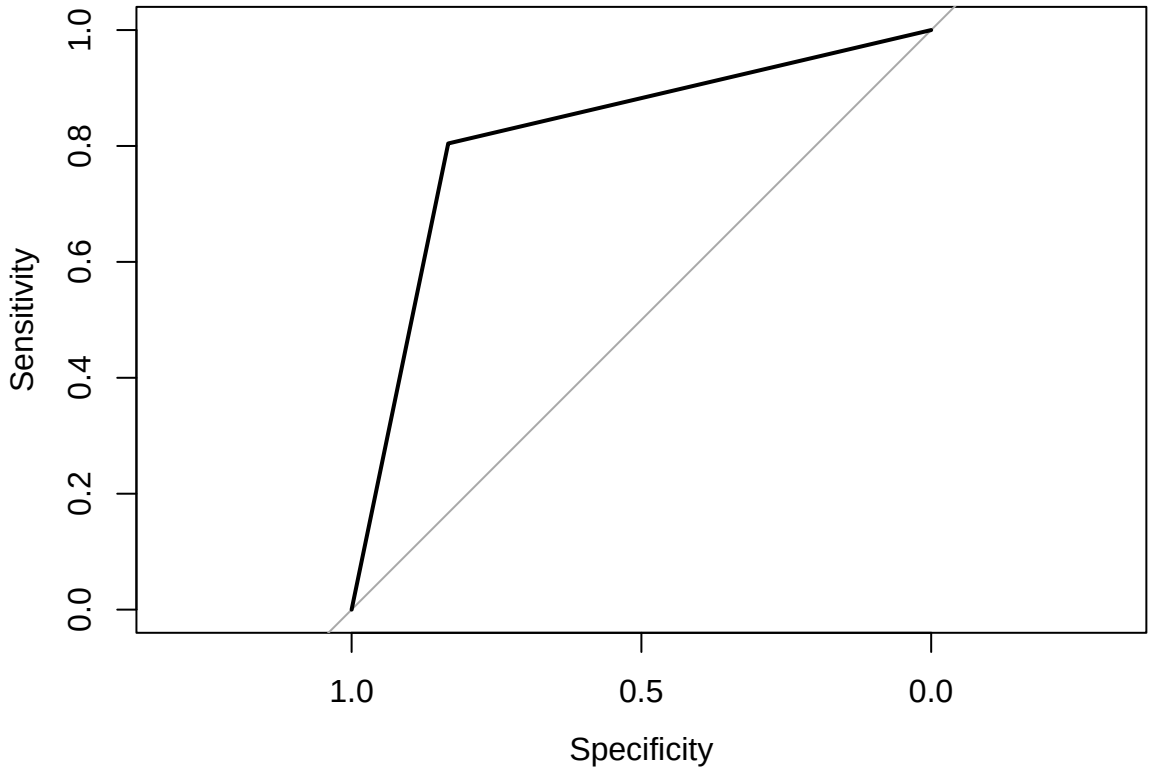
Being female or not married have some impact on whether or not your salary is above the median. This impact is very little and may even be insignificant since it is very close to 0. Even their direction of impact is not directly obvious from the plot.

```
## Confusion Matrix and Statistics
##
##      pred_0_1
## testY  0  1
##     0 40  8
##     1  9 37
##
```

```
##                Accuracy : 0.8191
##                  95% CI : (0.7263, 0.891)
##     No Information Rate : 0.5213
##     P-Value [Acc > NIR] : 1.637e-09
##
##                   Kappa : 0.638
##
##  Mcnemar's Test P-Value : 1
##
##             Sensitivity : 0.8163
##             Specificity : 0.8222
##          Pos Pred Value : 0.8333
##          Neg Pred Value : 0.8043
##              Prevalence : 0.5213
##          Detection Rate : 0.4255
##    Detection Prevalence : 0.5106
##       Balanced Accuracy : 0.8193
##
##        'Positive' Class : 0
##
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```




```
## Area under the curve: 0.8188
```

The prediction accuracy of this model is about 0.7979, and the kappa value is 0.597. The AUC is 0.798 which shows the model does a great job in predicting whether a person's salary falls below the median salary or not. This also indicates that the important variables can be seriously considered

The gradient boosting model seems to have better test prediction metric values than the random forest model. The accuracy, specificity and sensitivity are much higher.

Also, the findings from the teo models are are very consistent.

Summary and Recommendation for Human Resource Managers.

After this analysis, it is very easy to see that the gender distribution of the company is really great even though there is a little bias towards females. Also, there are no non binary gender groups within the data.This may be as a result of it not being reported or the data collection using sex instead of gender.

Furthermore, even though job positions influences salary , it is wonderful to note that within the same department, there is very little diversity in the salary.

Also, there were very few hispanics and American Indians in the data. Human resource managers should do well with targeting such people as well in diversity fairs since currently those help only recruit black people.

In terms of gender and salary distribution, there seems to be fairness.

The recruitment source seems to be influential in decided whether your salary is higher.In terms of recruiting more races aside White and Blacks should be considered since they are few in the company and they are seen to exceed expectations more often. More minority groups needs to be considered during diversity fairs.

For the next senior BI developer role and Network Engineers, other races should be given much higher consideration.

Conclusion

The insights from this analysis shows that job position is the most influential factor to consider to know whether ones salary will fall above the median. The company did not have serious red flag diversity and inclusion issues and the salary distribution seems pretty fair.

Furthermore, techniques from this class has been successful in giving insights about this data set.

References

1. Keen, K. J. (2018), Graphics for Statistics and Data Analysis with R, 2nd Edition, CRC.
2. https://www.kaggle.com/datasets/rhuebner/human-resources-data-set.