MSA 6702

Analytics Project

San Francisco Employee Salaries



Karishma Garewal

Instructor: Prof. John Chen

Table of Content:

1. Objective
2. Meet the data
3. Exploratory Data Analysis
4. Gender Gap
5. Regression Model
6. Decision Tree
7. Clustering
8. Conclusion

Objective:

The objective here is to find out how benefits vary for administrative analyst title based on base pay, overtime pay, other pay and total pay. Then showing how base pay varies over a course of 4 years and correlation between base pay, overtime pay, benefits and year. So, with the help of base pay we can predict what will be the benefit.

Meet the Data:

This dataset was taken from kaggle having multiple job titles and salary of employees between year 2011 and 2012.

Variables:

ID

EmployeeName

JobTitle

Basepay

OvertimePay

Otherpay

Benefits

TotalPay Benefits

Gender

Sample Data:

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Id** | **Employee Name** | **JobTitle** | **BasePay** | **OvertimePay** | **OtherPay** | **Benefits** | **TotalPay** | **TotalPayBenefits** | **Year** | **Agency** | **gender** |
| 11056 | maria mckee | administrative analyst iii | 92550.03 | 0 | 0 | 0 | 92550.03 | 92550.03 | 2011 | San Francisco | f |
| 14517 | simone jacques | administrative analyst | 75306.3 | 0 | 4448 | 0 | 79754.03 | 79754.03 | 2011 | San Francisco | f |
| 15477 | sherry tan | administrative analyst | 75005.63 | 0 | 1000 | 0 | 76005.63 | 76005.63 | 2011 | San Francisco | f |
| 15488 | shirley li | administrative analyst | 75005.63 | 0 | 960 | 0 | 75965.63 | 75965.63 | 2011 | San Francisco | f |
| 15613 | megan filly | administrative analyst ii | 75585.82 | 0 | 0 | 0 | 75585.82 | 75585.82 | 2011 | San Francisco | f |

Exploratory Data Analysis:

Here I have taken subset of data, that is data which is related to Job title **“Administrative Analyst”.**

The main aim here is to predict the benefits how it will vary based on the basepay and try to find correlation between some variables.

R Code:

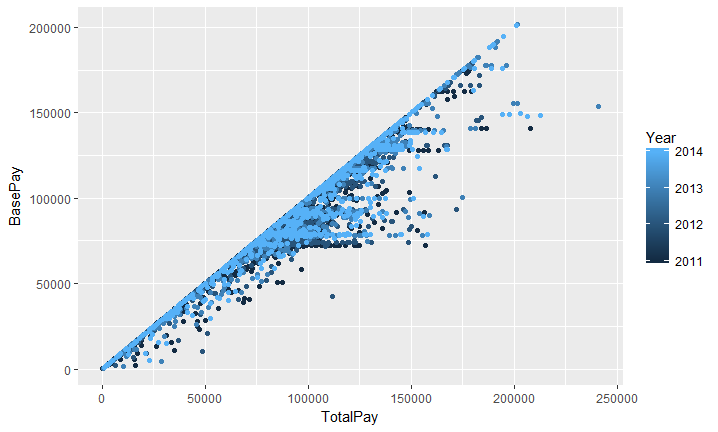
ggplot(subset(Analytics\_Data, Year %in% c(2011, 2012,2013,2014)),

       aes(x=TotalPay,

           y=BasePay,

           color=Year))+

  geom\_point()



From this plot, we can see that over the time as the TotalPay is increasing BasePay is also increasing.

R Code:

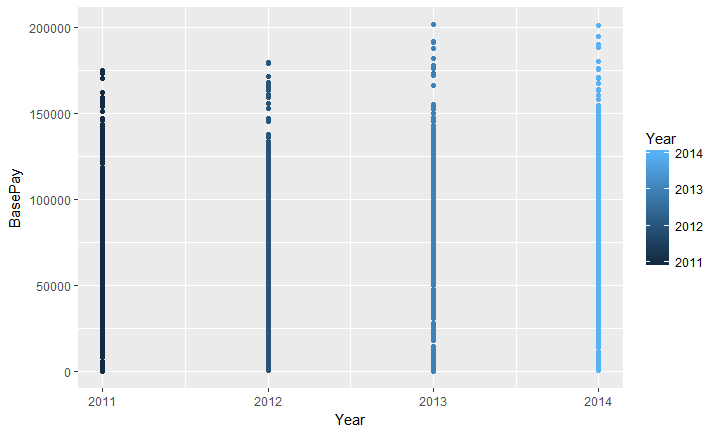
ggplot(subset(Analytics\_Data, Year %in% c(2011, 2012,2013,2014)),

       aes(x=Year,

           y=BasePay,

           color= BasePay))+

  geom\_point()



Here we can see that how BasePay is changing for each year. One interesting fact here is the Basepay for 2012 is lower then what it is for 2011 and then in 2013 again it increased, but in case of 2014 and 2014 there is not much difference.

R Code:

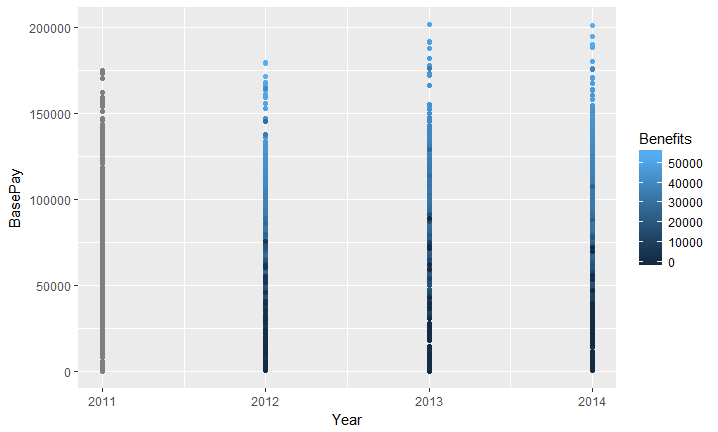
ggplot(subset(Analytics\_Data, Year %in% c(2011, 2012,2013,2014)),

       aes(x=Year,

           y=BasePay,

           color=Benefits))+

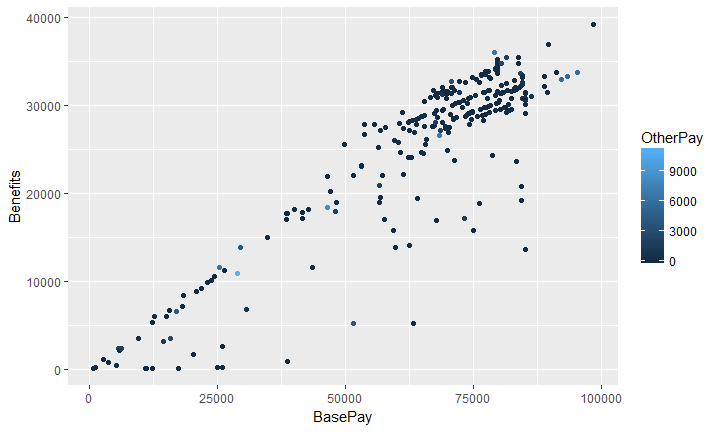
  geom\_point()



R Code:

library(ggplot2)

ggplot(final\_data, aes(BasePay, Benefits, color = OtherPay)) + geom\_point()



From here we can see that as the BasePay is increasing Benefits are also increasing. But there is one interesting fact as the overtime pay is increasing Benefits are going down.

Gender Gap:

R Code:

attach(male)

attach(female)

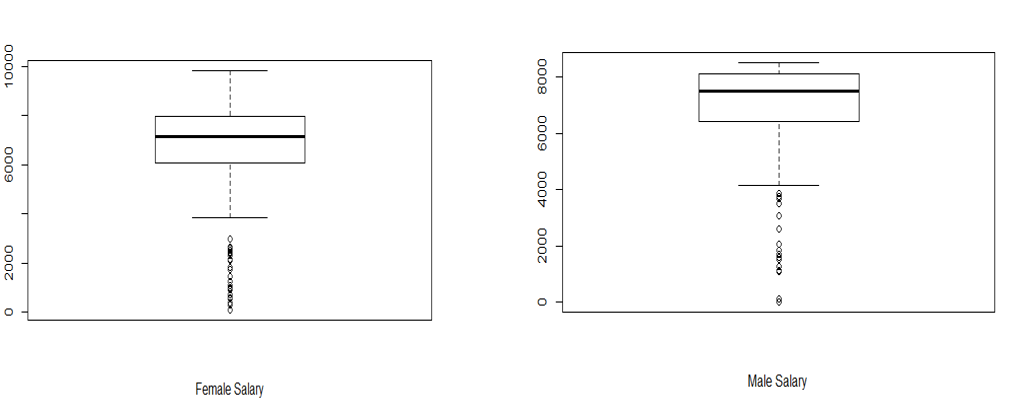
salaryfe=female$BasePay/10

boxplot(salaryfe,xlab = "Female Salary",color="orange")

salaryma=male$BasePay/10

boxplot(salaryma,xlab = "Male Salary ",color="orange")

Here I have tried to show if there is any biasness between BasePay of Males and Females.



What I found here was that the Mean for Males was higher than Mean for females. However in case of females the data seems to be same I the upper and lower whiskers however in case of Males more data lies in lower whisker.

Bootstrap Distribution:

R Code:

times.Basic <- male$BasePay

times.Ext   <- female$BasePay

n.Basic <- length(times.Basic)

n.Ext <- length(times.Ext)

B <- 10^4

times.diff.mean <- numeric(B)

for (i in 1:B)

{

  Basic.boot <- sample(times.Basic, n.Basic, replace=TRUE)

  Ext.boot <- sample(times.Ext, n.Basic, replace=TRUE)

  times.diff.mean[i] <- mean(Basic.boot)-mean(Ext.boot)

}

hist(times.diff.mean, main="Bootstrap distribution of difference in means",xlab="Means")

abline(v = mean(times.diff.mean), col = "red")

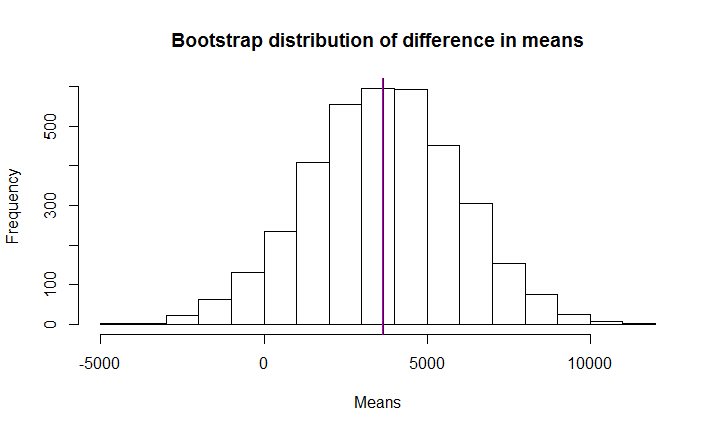
abline(v = mean(times.Basic) - mean(times.Ext), col = "blue")

an = times.diff.mean[![is.na](http://is.na/)(times.diff.mean)]

times.Basi

times.diff.mec = times.Basic[![is.na](http://is.na/)(times.Basic)]

times.Ext = times.Ext[![is.na](http://is.na/)(times.Ext)]



From here we can see that there is difference in mean salaries of males and females, which further confirms our earlier assumption about difference I men.

Regression Model:

R Code:

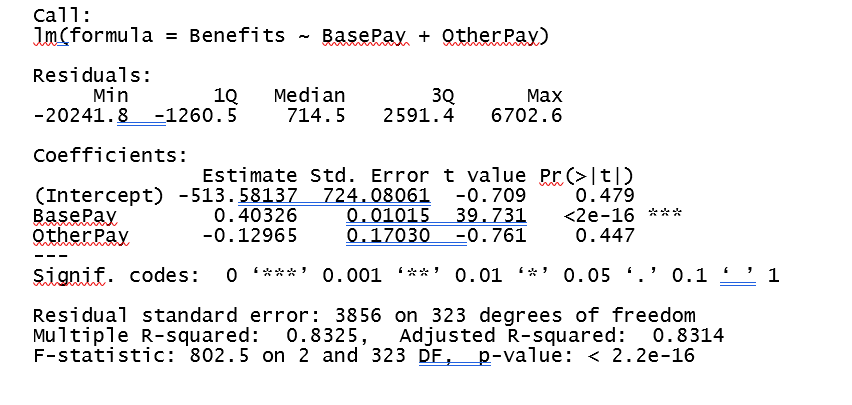
modelbenefits=lm(Benefits~BasePay+OtherPay+TotalPay)

summary(modelbenefits)

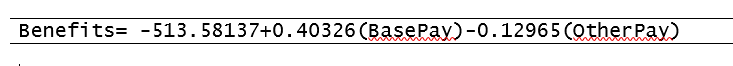
modelbenefits2=lm(Benefits~ BasePay +OtherPay+)

summary(modelbenefits2)

First I tried to build full model which was not having significant variables. So, I have built reduced model.



Regression Equation:



So, we can say that for each unit increase in Basepay the Benefits increase by 0.40326.

Decision Tree:

R Code:

library(rpart)

fittree=rpart(Benefits~BasePay+OtherPay+TotalPay, data=final\_data, method="class")

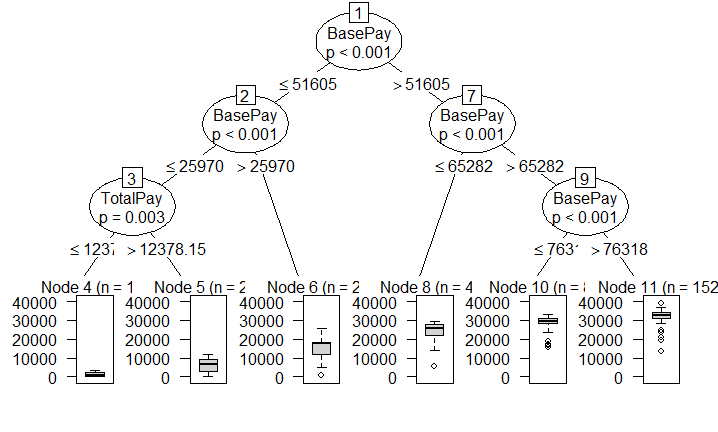
plot(fittree,uniform="TRUE")

library(party)

str(Fin\_Data)

decisiontree=ctree(Fin\_Data$Benefits~Fin\_Data$BasePay+Fin\_Data$TotalPay+Fin\_Data$OtherPay)

plot(decisiontree)



Here from the leaf node only we can see that as the base pay is increasing the benefits are increasing.

So, there is a gradual increase in mean in the boxplot.

The tree is first divided into two parts Base pay <= 51605 is on the left side and greater than 51605 is on the right side, and then again there are further slits.

So, for example if the base pay is 40,000 and then a person may get benefits around 9000.And consider a case where a person is having Basepay of 20,000 and total pay of 10000 then the chances of benefits are around 500.

Clustering:

R Code:

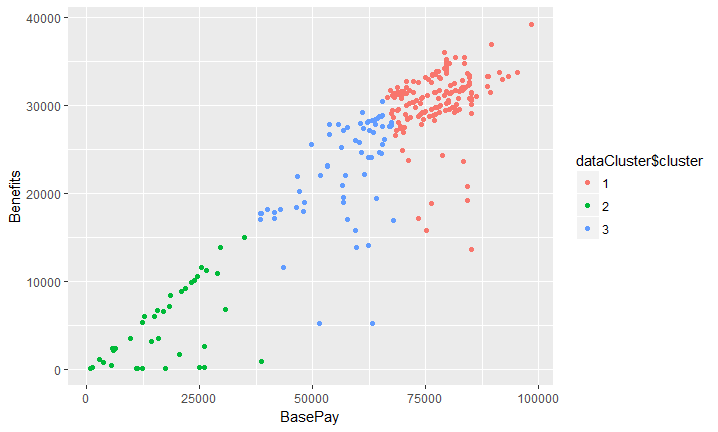
set.seed(20)

dataCluster <- kmeans(final\_data[, c(1:3)], 3, nstart = 20)

dataCluster$cluster <- as.factor(dataCluster$cluster)

ggplot(final\_data, aes(BasePay, Benefits, color = dataCluster$cluster)) + geom\_point()

Here I have used k-means clustering to find out the distribution of benefits based on the range of basepay.



So, from here we can see that if the basepay is in the range of 0 to 25,000 then the benefits will lie in the range of 0 to 15000.And again as the base pay is increasing the benefits are increasing.

Conclusion:

After building regression model, plots, decision tree and clustering finally we can conclude that as the base pay increases benefits also increase. And if we get experience, education level in the data we can predict the base pay, but also the data should be of a particular company as the pay varies from one company to another.