Predictive Analysis

Riya Amin

1/27/2020

```
# exploring dataset

str(ANZ)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    12043 obs. of  23 variables:
##  $ status           : chr  "authorized" "authorized" "authorized" "authorized" ...
##  $ card_present_flag: num  1 0 1 1 1 NA 1 1 1 NA ...
##  $ bpay_biller_code : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ account          : chr  "ACC-1598451071" "ACC-1598451071" "ACC-1222300524" "ACC-103705056
4" ...
##  $ currency         : chr  "AUD" "AUD" "AUD" "AUD" ...
##  $ long_lat         : chr  "153.41 -27.95" "153.41 -27.95" "151.23 -33.94" "153.10 -27.66"
...
##  $ txn_description  : chr  "POS" "SALES-POS" "POS" "SALES-POS" ...
##  $ merchant_id      : chr  "81c48296-73be-44a7-befa-d053f48ce7cd" "830a451c-316e-4a6a-bf25-e3
7caedca49e" "835c231d-8cdf-4e96-859d-e9d571760cf0" "48514682-c78a-4a88-b0da-2d6302e64673" ...
##  $ merchant_code    : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ first_name       : chr  "Diana" "Diana" "Michael" "Rhonda" ...
##  $ balance          : num  35.39 21.2 5.71 2117.22 17.95 ...
##  $ date             : POSIXct, format: "2018-08-01" "2018-08-01" ...
##  $ gender           : chr  "F" "F" "M" "F" ...
##  $ age              : num  26 26 38 40 26 20 43 43 27 40 ...
##  $ merchant_suburb  : chr  "Ashmore" "Sydney" "Sydney" "Buderim" ...
##  $ merchant_state   : chr  "QLD" "NSW" "NSW" "QLD" ...
##  $ extraction       : chr  "2018-08-01T01:01:15.000+0000" "2018-08-01T01:13:45.000+0000" "201
8-08-01T01:26:15.000+0000" "2018-08-01T01:38:45.000+0000" ...
##  $ amount           : num  16.25 14.19 6.42 40.9 3.25 ...
##  $ transaction_id   : chr  "a623070bfead4541a6b0fff8a09e706c" "13270a2a902145da9db4c951e04b51
b9" "feb79e7ecd7048a5a36ec889d1a94270" "2698170da3704fd981b15e64a006079e" ...
##  $ country          : chr  "Australia" "Australia" "Australia" "Australia" ...
##  $ customer_id      : chr  "CUS-2487424745" "CUS-2487424745" "CUS-2142601169" "CUS-161422687
2" ...
##  $ merchant_long_lat: chr  "153.38 -27.99" "151.21 -33.87" "151.21 -33.87" "153.05 -26.68"
...
##  $ movement         : chr  "debit" "debit" "debit" "debit" ...
```

```
summary(ANZ)
```

```
##    status          card_present_flag bpay_biller_code   account
##  Length:12043       Min.   :0.000     Min.   :0        Length:12043
##  Class :character   1st Qu.:1.000     1st Qu.:0        Class :character
##  Mode  :character   Median :1.000     Median :0        Mode  :character
##                     Mean   :0.803     Mean   :0
##                     3rd Qu.:1.000     3rd Qu.:0
##                     Max.   :1.000     Max.   :0
##                     NA's   :4326      NA's   :11160
##    currency           long_lat         txn_description    merchant_id
##  Length:12043       Length:12043       Length:12043       Length:12043
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  merchant_code    first_name           balance
##  Min.   :0       Length:12043       Min.   :     0.24
##  1st Qu.:0       Class :character   1st Qu.:  3158.59
##  Median :0       Mode  :character   Median :  6432.01
##  Mean   :0                          Mean   : 14704.20
##  3rd Qu.:0                          3rd Qu.: 12465.94
##  Max.   :0                          Max.   :267128.52
##  NA's   :11160
##       date                          gender               age
##  Min.   :2018-08-01 00:00:00   Length:12043       Min.   :18.00
##  1st Qu.:2018-08-24 00:00:00   Class :character   1st Qu.:22.00
##  Median :2018-09-16 00:00:00   Mode  :character   Median :28.00
##  Mean   :2018-09-15 21:27:39                      Mean   :30.58
##  3rd Qu.:2018-10-09 00:00:00                      3rd Qu.:38.00
##  Max.   :2018-10-31 00:00:00                      Max.   :78.00
##
##  merchant_suburb    merchant_state      extraction            amount
##  Length:12043       Length:12043       Length:12043       Min.   :   0.10
##  Class :character   Class :character   Class :character   1st Qu.:  16.00
##  Mode  :character   Mode  :character   Mode  :character   Median :  29.00
##                                                           Mean   : 187.93
##                                                           3rd Qu.:  53.66
##                                                           Max.   :8835.98
##
##  transaction_id       country           customer_id        merchant_long_lat
##  Length:12043       Length:12043       Length:12043       Length:12043
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    movement
##  Length:12043
##  Class :character
##  Mode  :character
##
```

```
##
##
##
```

Predictive Analysis

Indentifying the annual salary for each cutomer.

Annual salary for each cutomer can be estimated by filtering on the txn_description column and looking only at those rows marked ‘PAY/SALARY’. Multiplying the amount paid by the frequency of payment gives the tri-monthly payment. Multipyling by 4 gives the year salary (12/3).

Filtering on the txn description column and looking only at those rows marked ’ PAY/SALARY ’ can estimate the annual salary for each customer. The 3 months payment is given by multiplying the amount paid by the frequency of payment. The multiplication by 4 gives salary for the year (12/3).

```
salary <- ANZ %>% filter(txn_description == 'PAY/SALARY')
salary
```

status <chr>	card_present_flag <dbl>	bpay_biller_code <dbl>	account <chr>	currency <chr>	long_lat <chr>	txn <ch
posted	*NA*	0	ACC-588564840	AUD	151.27 -33.76	PA
posted	*NA*	0	ACC-1650504218	AUD	145.01 -37.93	PA
posted	*NA*	0	ACC-3326339947	AUD	151.18 -33.80	PA
posted	*NA*	0	ACC-3541460373	AUD	145.00 -37.83	PA
posted	*NA*	0	ACC-2776252858	AUD	144.95 -37.76	PA
posted	*NA*	0	ACC-1598451071	AUD	153.41 -27.95	PA
posted	*NA*	0	ACC-3485804958	AUD	138.52 -35.01	PA
posted	*NA*	0	ACC-1973887809	AUD	115.78 -31.90	PA
posted	*NA*	0	ACC-4059612845	AUD	130.98 -12.49	PA
posted	*NA*	0	ACC-819621312	AUD	145.04 -37.85	PA

1-10 of 883 rows | 1-7 of 23 columns Previous **1** 2 3 4 5 6 … 89 Next

```
Annual_salary <- salary %>% group_by(account)%>% summarize(age= mean(age), gender = unique(gende
r), annual_salary = sum(amount*4))

Annual_salary
```

account <chr>	age <dbl>	gender <chr>	annual_salary <dbl>

account <chr>	age <dbl>	gender <chr>	annual_salary <dbl>
ACC-1037050564	40	F	46388.68
ACC-1056639002	22	M	76680.24
ACC-1199531521	52	M	106001.84
ACC-1217063613	27	F	38908.96
ACC-1222300524	38	M	52110.76
ACC-1243371644	42	M	40357.92
ACC-1279356312	44	M	69296.16
ACC-1334819143	33	M	65244.24
ACC-1344825761	46	F	59290.80
ACC-1349834573	39	M	85991.92

1-10 of 100 rows Previous **1** 2 3 4 5 6 … 10 Next

This task explores correlations between annual salary and various customer attributes (eg: age, gender).

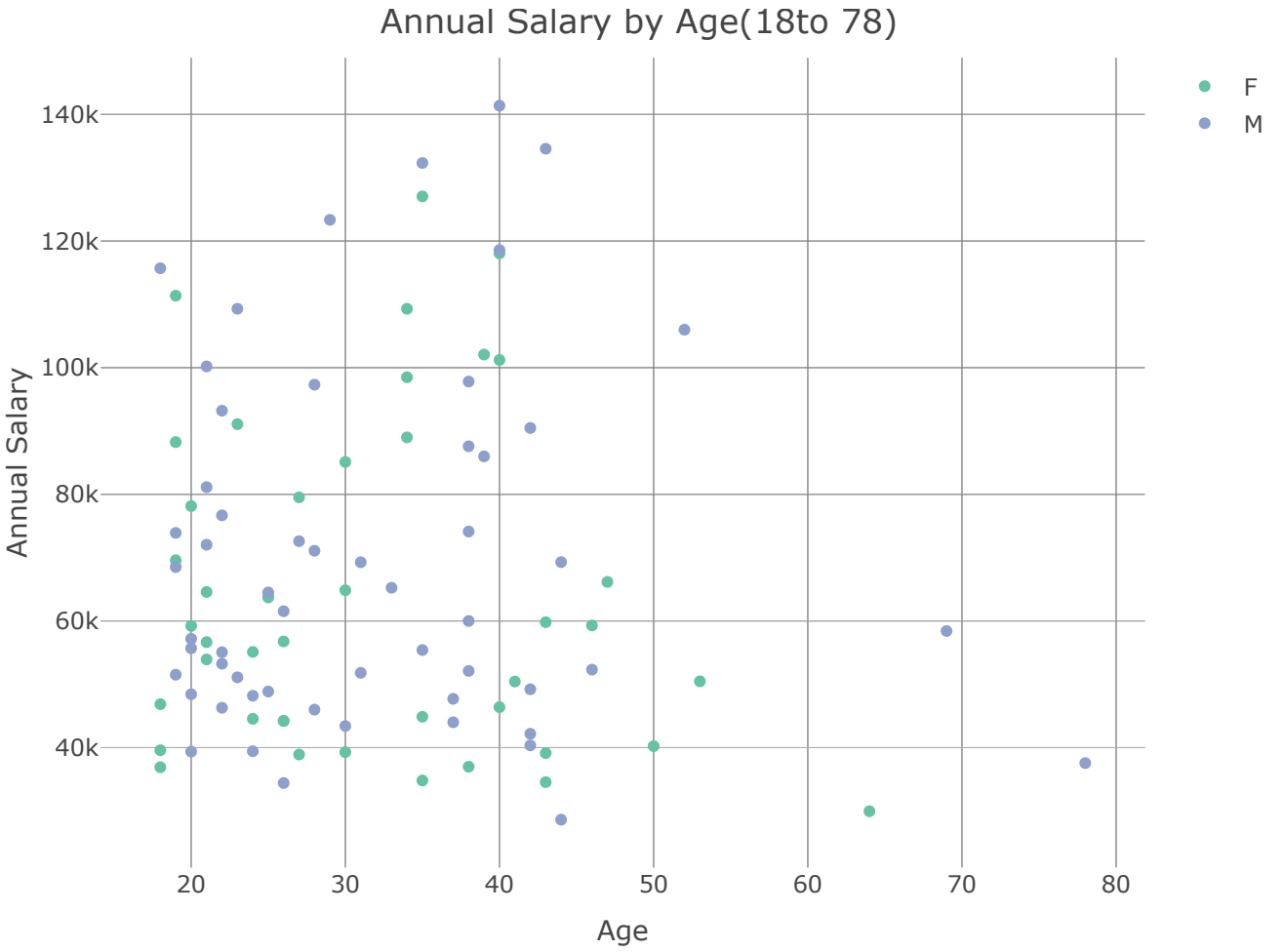
AGE

As the age varies from 18 to 78 in this dataset-that is, from people only beginning their careers to pensioners-there is a very non-linear relationship between the age and the annual salary of then given the entire age range. One reason is that customers above the age of 60 all have incomes at the bottom end of the spectrum.

```
plot_ly(data = Annual_salary,
        x=~age, y=~annual_salary, color =~gender , type = "scatter",
        text =~ paste("AGE:", age))%>%
  layout(title = "Annual Salary by Age(18to 78)",
         xaxis = list(title = "Age"),
         yaxis = list(title = "Annual Salary"))
```

```
## No scatter mode specifed:
##   Setting the mode to markers
##   Read more about this attribute -> https://plot.ly/r/reference/#scatter-mode
```

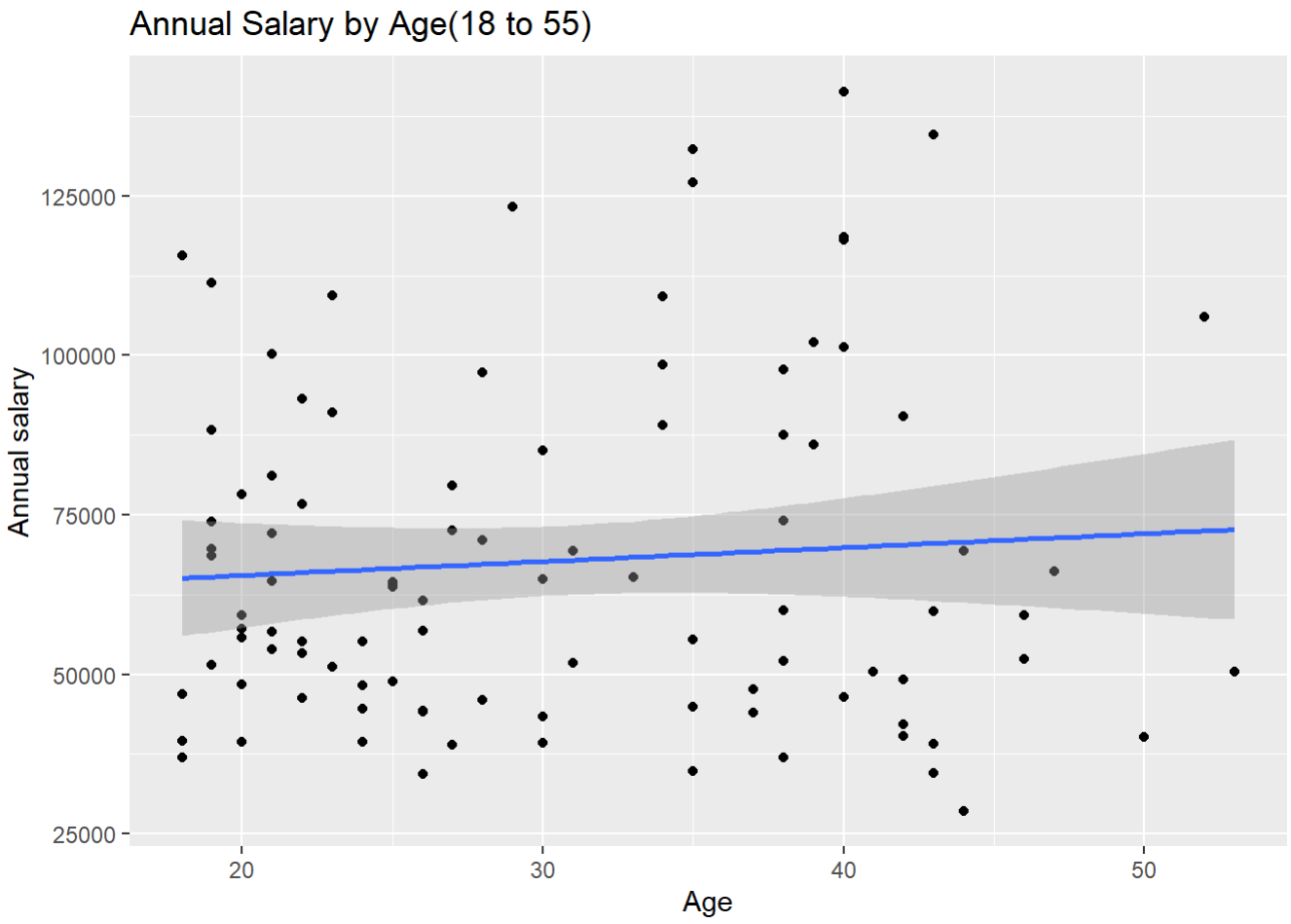
```
## Warning in RColorBrewer::brewer.pal(N, "Set2"): minimal value for n is 3, returning requested
palette with 3 different levels

## Warning in RColorBrewer::brewer.pal(N, "Set2"): minimal value for n is 3, returning requested
palette with 3 different levels
```




```
cor(Annual_salary$annual_salary,Annual_salary$age)
```

```
## [1] -0.0365039
```

```
Annual_salary_18to55 <- Annual_salary%>% filter(Annual_salary$age <= 55)

Annual_salary_18to55 %>% ggplot(aes(x= age, y= annual_salary))+geom_point()+
  geom_smooth(method = lm)+
  theme(legend.position = "none")+
  labs(y = "Annual salary",
       x = "Age",
       title = 'Annual Salary by Age(18 to 55)')
```



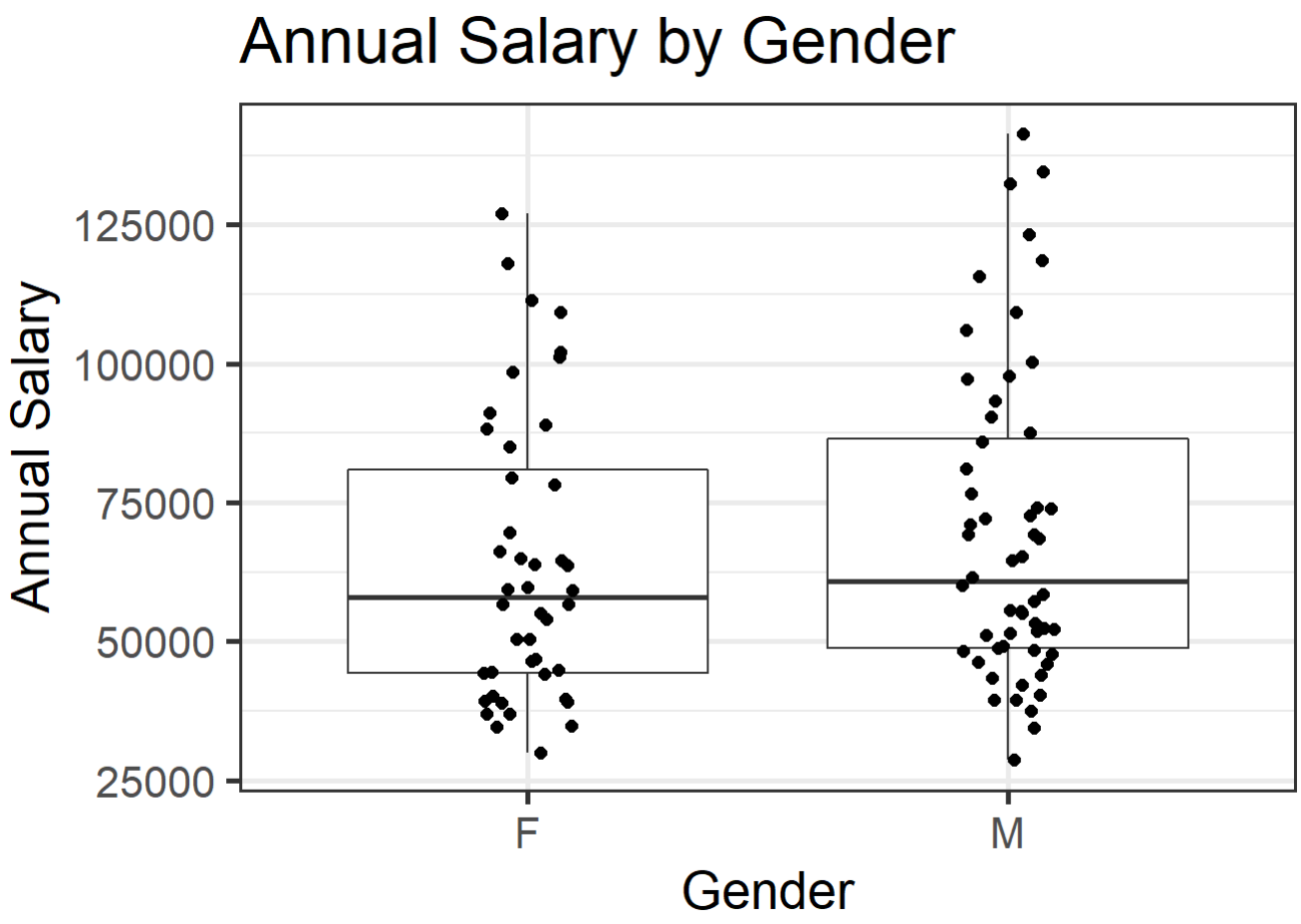

```
cor(Annual_salary_18to55$annual_salary,Annual_salary_18to55$age)
```

```
## [1] 0.07595699
```

GENDER

It seems from the estimation of the annual salary by class that men will earn more than women, but this problem can be formalized by a hypothesis check. We would like to use a two-sample test, but first we need to verify the presumption that the usually distributed data is being followed.

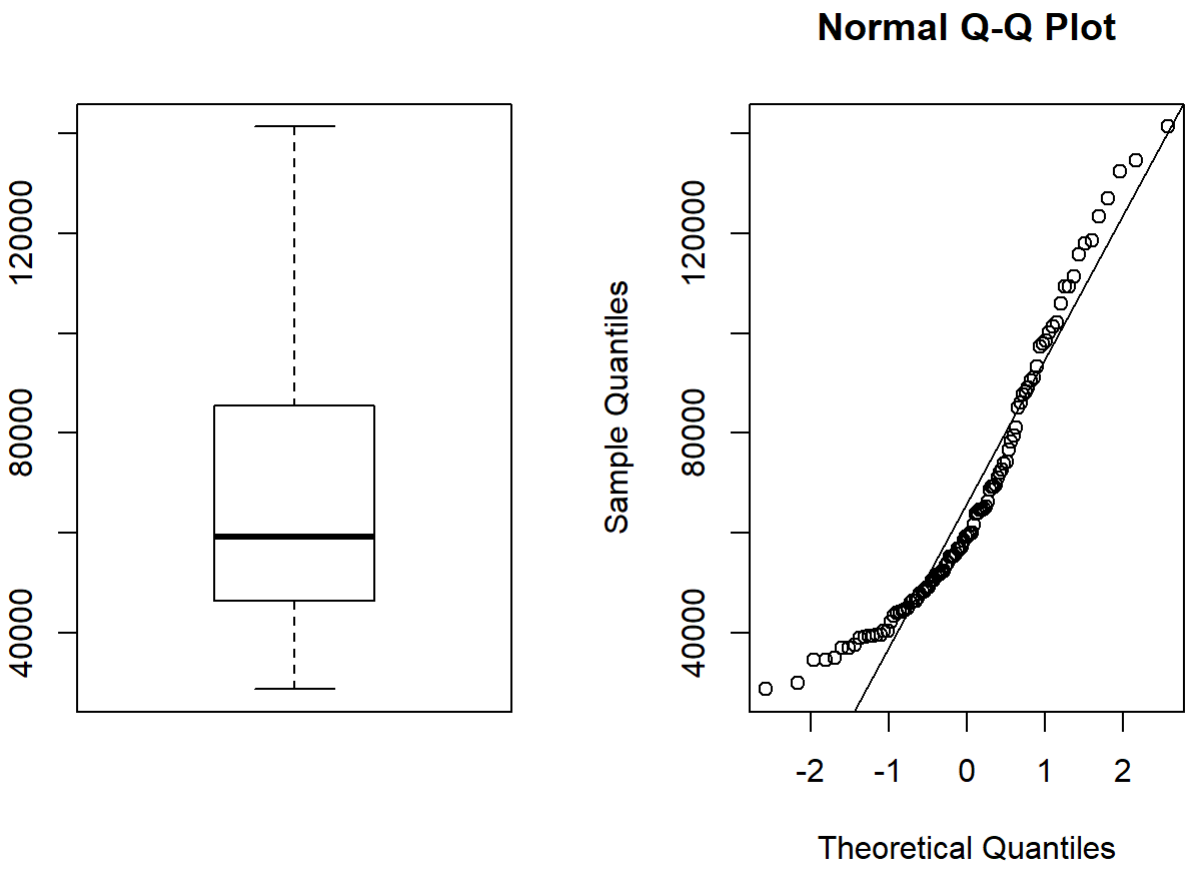
```
Annual_salary %>% ggplot(aes(y= annual_salary, x= gender))+
  geom_boxplot(coef=10)+
                geom_jitter(width = 0.1, size = 2)+
                theme_bw(base_size = 20)+
                theme(legend.position = "none")+
                labs(y = "Annual Salary",
                     x = "Gender",
                     title = 'Annual Salary by Gender')
```



Testing Normality Assumptions:

Plotting the annual salary data box plot and the Q-Q plot to check the normality, we see that the box plot does not appear quite symmetrical suggesting that the mean is in the lower part of the spectrum. The Q-Q diagram looks fairly continuous except for the lower statistical quantiles. From this we can infer that the data tends to be fairly commonly distributed, and the number of observations(100) helps us to depend on the central limit theorem for the normal distribution of the sample means. Therefore, we will assume that the hypotheses needed to perform the t-test are fulfilled.

```
par(mfrow =c(1,2))
boxplot(Annual_salary$annual_salary)
qqnorm(Annual_salary$annual_salary)
qqline(Annual_salary$annual_salary)
```




```
library(purrr)
salary_men <- as_vector(Annual_salary%>% filter(gender == 'M') %>% select(annual_salary))
salary_women <- as_vector(Annual_salary %>% filter(gender == 'F')%>% select(annual_salary))

t.test(salary_men, salary_women, alternative = 'greater')
```

```
##
##  Welch Two Sample t-test
##
## data:  salary_men and salary_women
## t = 1.0279, df = 95.592, p-value = 0.1533
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -3403.147       Inf
## sample estimates:
## mean of x mean of y
##  69494.33  63968.75
```

With the test figure of 1.0279 providing a p-value of 0.1533 at a 5% significance level, there is evidence to suggest that we should not dreject the null hypothesis that the average annual salary of men is no different from the average annual salary of women. In view of this result, it is questionable whether gender would be a good predictor of annual salary in the regression model.

Spending transactions

We will consider total yearly spending for each cutomer as a metric summarising their habits. Frequency of debit transactions does not matter as much as the total amount spent each year as it is possible to engage in very frequent small transactions yet not spend much overall, or to engagein fewer but larger transactions and spend much more. We will consider yearly spending to keep the time scale consistent with yearly salary.

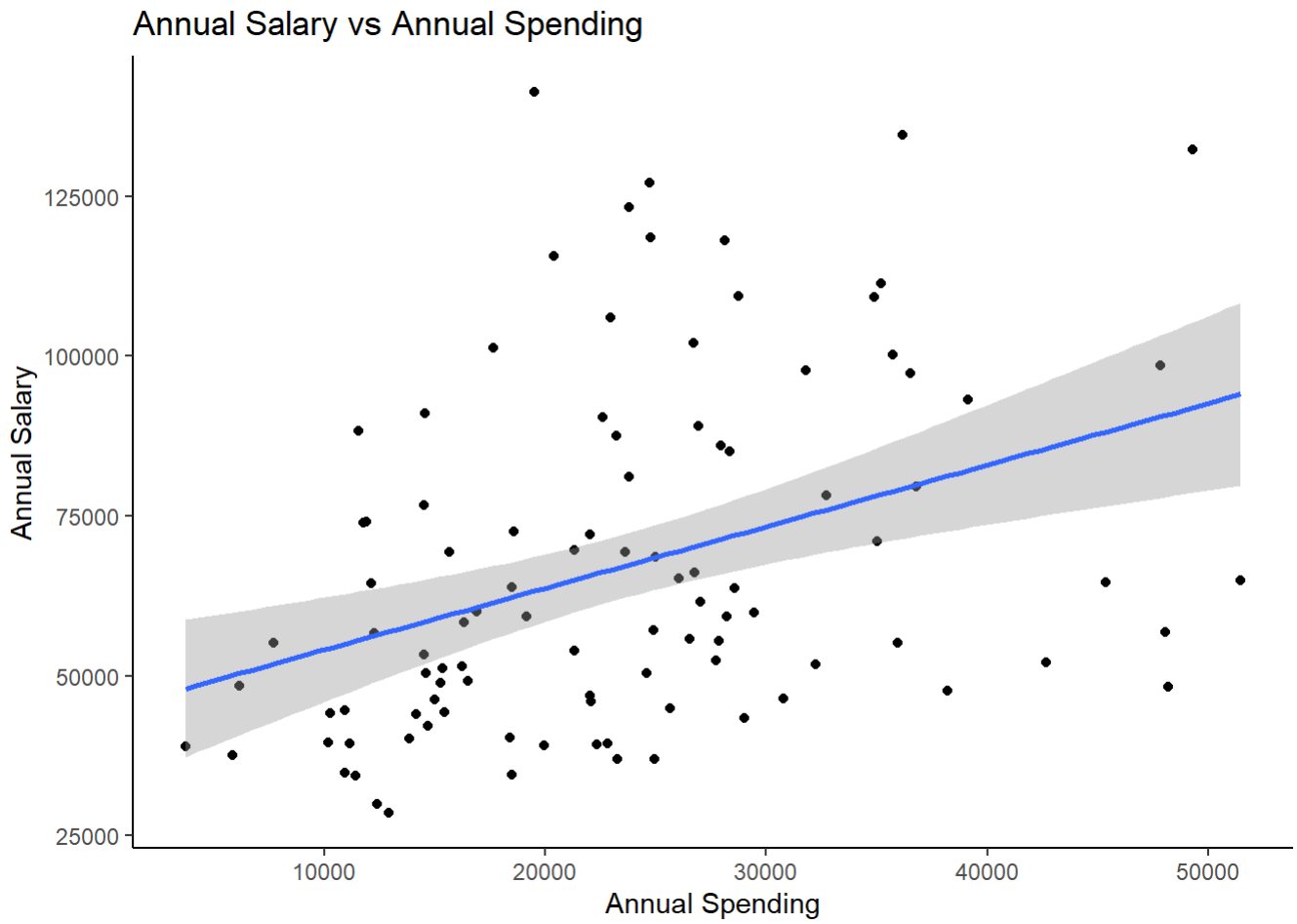
For each customer, we will consider total annual expenditure as a metric summarizing their habits. The frequency of debit transactions does not matter as much as the total amount spent each year as it is possible to engage in very frequent small transactions but not spend a lot of money overall, or to engage in less but larger transactions and spend a lot more. we recommend investing on an annual basis to maintain the time scale aligned with the annual salary.

```
spending <- ANZ %>% filter(movement == 'debit') %>% group_by(account)%>% summarize(annual_spendi
ng = sum(amount*4), monthly_spending = sum(amount/3))

salary_factors <- merge(Annual_salary,spending)

head(salary_factors)
```

account <chr>	…	gender <chr>	annual_salary <dbl>	annual_spending <dbl>	monthly_spending <dbl>
1 ACC-1037050564	40	F	46388.68	30757.08	2563.0900
2 ACC-1056639002	22	M	76680.24	14526.12	1210.5100
3 ACC-1199531521	52	M	106001.84	22970.84	1914.2367
4 ACC-1217063613	27	F	38908.96	3701.92	308.4933
5 ACC-1222300524	38	M	52110.76	42675.04	3556.2533
6 ACC-1243371644	42	M	40357.92	18411.04	1534.2533

6 rows

```
salary_factors %>% ggplot(aes(x= annual_spending, y= annual_salary)) +geom_point()+ geom_smooth
(method = lm)+ theme_classic() +
  labs(y = "Annual Salary",
       x = "Annual Spending",
       title = 'Annual Salary vs Annual Spending')
```




```
cor(salary_factors$annual_salary, salary_factors$annual_spending)
```

```
## [1] 0.3734772
```

Through comparing the annual salary against the annual spending for each client and changing the regression axis, we can see that there is a rather strong positive correlation between the two, with a correlation coefficient of 0.37.

Expenditure on an annual basis is likely to be the best indicator of the annual salary we have found.

Multiple Regression Model

We will fit a multiple regression model using the three matrixes discussed-age, gender and expenditure-and then assess the effectiveness of the first gender to be recorded as a binary numeric variable. We're going to use 1 for males and 0 for females.

```
salary_factors$gender <- ifelse(salary_factors$gender == "M",1,0)
```

Currently, we use 60 percent of the dataset as training data to fit the layout. This allows us to test the predictive accuracy of the model using the remaining 40%.

```
set.seed(101)

sample_data <- sample.split(names(salary_factors), SplitRatio = 0.6)
train <- subset(salary_factors, sample_data == TRUE)
test <- subset(salary_factors, sample_data == FALSE)

lin_mod <- lm(annual_salary ~ age + gender + annual_spending, data = train)

summary(lin_mod)
```

```
##
## Call:
## lm(formula = annual_salary ~ age + gender + annual_spending,
##     data = train)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -41604 -18062  -5348  16165  56980
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     54845.1419 13651.4164   4.018 0.000221 ***
## age               -70.9596   326.2929  -0.217 0.828823
## gender           2137.2293  7792.9080   0.274 0.785145
## annual_spending     0.7161     0.3442   2.080 0.043216 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25770 on 45 degrees of freedom
## Multiple R-squared:  0.0888, Adjusted R-squared:  0.02806
## F-statistic: 1.462 on 3 and 45 DF,  p-value: 0.2376
```

As indicated in the initial tests for the relationship between the three predictors and the annual salary, age and gender are not important predictors of large p-value annual salaries of 0.828823 and.785145, respectively. Nonetheless, total spending is a very strong indicator of annual salaries with a lower p-value of 0.043216. It indicates that it would be useful to reduce the age and gender of the sample by merely utilizing annual investment to estimate annual salaries.

Model Accuracy

Two measures will be used: 1. Root Mean Squared Error (RMSE) and 2. Mean Absolute Error (MAE).

We will test the accuracy predciting both in sample, using the original training set, and out of sample using the test set. We must check the accuracy forecasting both in the study, using the initial training package, and out of the field using the test collection.

In sample Accuracy :

```
predictions_in <- predict(lin_mod, newdata = train, interval = "prediction", level = 0.90)

rmse_in <- mean((train$annual_salary - predictions_in)^2)
sqrt(rmse_in)
```

```
## [1] 44284.96
```

```
mae_in <- mean(abs(train$annual_salary - predictions_in))
mae_in
```

```
## [1] 37213.59
```

Out of sample Accuracy:

```
predictions_out <- predict(lin_mod, newdata = test, interval = "prediction", level = 0.90)

rmse_out <- mean((test$annual_salary- predictions_out)^2)

sqrt(rmse_out)
```

```
## [1] 44835.12
```

```
mae_out <- mean(abs(test$annual_salary- predictions_out))

mae_out
```

```
## [1] 37751.16
```

In the sample and out of the sample are fairly similar, as is shown in and out of the MAE sample, suggesting that the model generalizes fairly well new data. Nonetheless, both of these parameters tend to be quite high, indicating that there is generally a fairly significant prediction bias using this model. On review, though, both are in the region of 35000 to 45000. As income is usually segmented into roughly $50,000 brackets, this is perhaps an acceptable level of error when trying to predict a customer's revenue bracket. There is room for improvement, though, perhaps by identifying some low association predictors and integrating them into the model for better accuracy.

Decision Tree Based Model

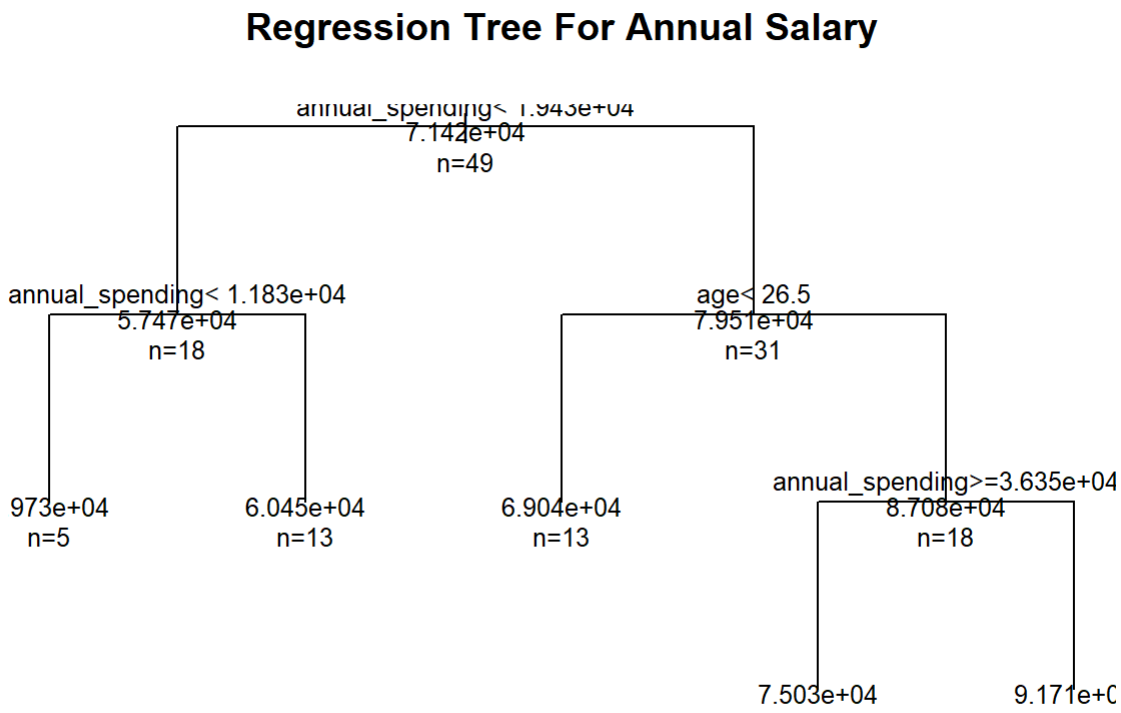
In order to fit the regression tree, we use the same method as for multiple regression, estimating annual salaries based on age, class and annual spending. We use the same training package that we set up to fit the pattern.

```
D_tree <- rpart(annual_salary ~ age+ gender + annual_spending, data = train, method = 'anova', c
ontrol = rpart.control(minsplit = 14))


printcp(D_tree)
```

```
##
## Regression tree:
## rpart(formula = annual_salary ~ age + gender + annual_spending,
##     data = train, method = "anova", control = rpart.control(minsplit = 14))
##
## Variables actually used in tree construction:
## [1] age             annual_spending
##
## Root node error: 3.2807e+10/49 = 669537108
##
## n= 49
##
##         CP nsplit rel error xerror    xstd
## 1 0.168616      0   1.00000 1.0773 0.18542
## 2 0.074873      1   0.83138 1.1789 0.23077
## 3 0.030631      2   0.75651 1.3224 0.23361
## 4 0.012654      3   0.72588 1.3720 0.24647
## 5 0.010000      4   0.71323 1.3502 0.22616
```

Specifying that the minimum number of measurements in the node is 14 results in a tree that divides the annual salary into six values, which seems fair, since it is similar to the number of brackets in which the ATO splits the profits. This also seems fair, as it means that we are not over-adjusting the model to the results.

```
plot(D_tree, uniform = TRUE,
     main = "Regression Tree For Annual Salary")
text(D_tree, use.n = TRUE, all = TRUE, cex= .8)
```

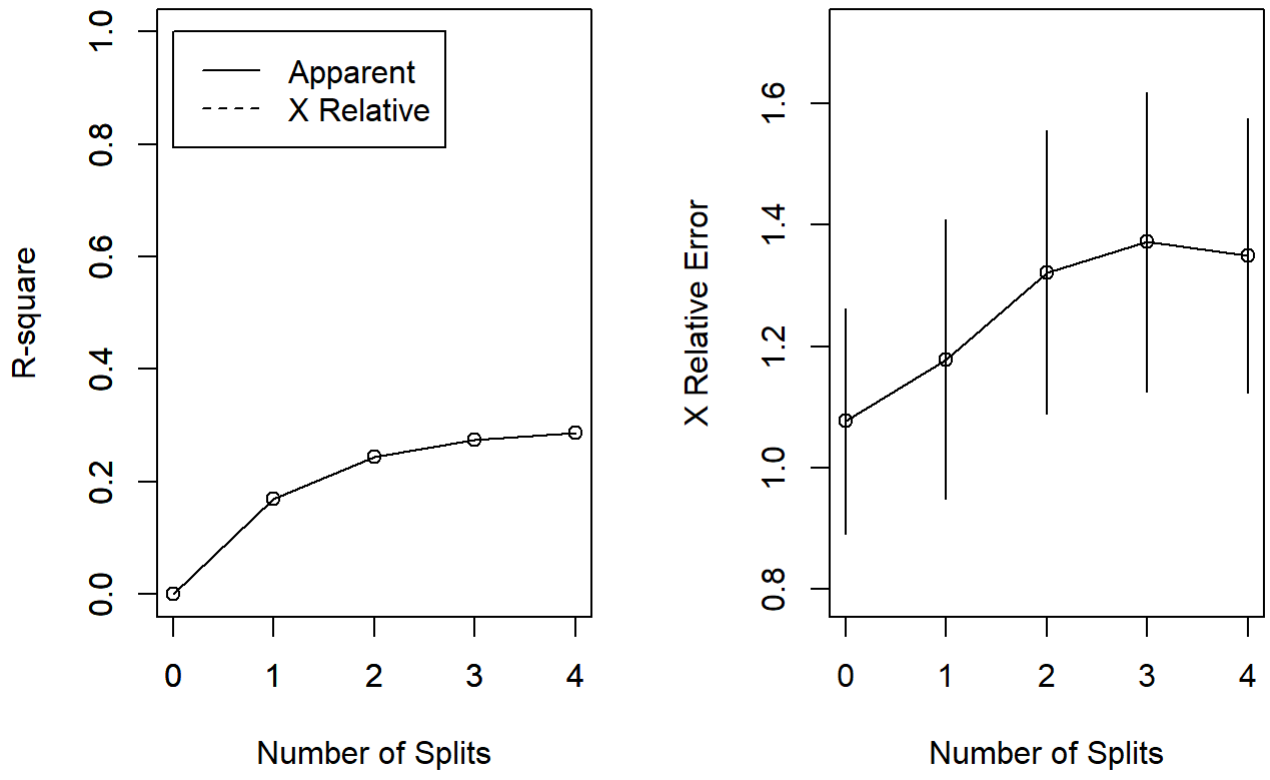


Accuracy of Model

The plot of the average R-square and relative error reveals that R-square (total variance factor described by the model) decreases with the number of splits as predicted, and the error tends to be constant at 1.5 with 4 splits.

```
par(mfrow = c(1,2))
rsq.rpart(D_tree)
```

```
##
## Regression tree:
## rpart(formula = annual_salary ~ age + gender + annual_spending,
##     data = train, method = "anova", control = rpart.control(minsplit = 14))
##
## Variables actually used in tree construction:
## [1] age             annual_spending
##
## Root node error: 3.2807e+10/49 = 669537108
##
## n= 49
##
##         CP nsplit rel error xerror    xstd
## 1 0.168616      0   1.00000 1.0773 0.18542
## 2 0.074873      1   0.83138 1.1789 0.23077
## 3 0.030631      2   0.75651 1.3224 0.23361
## 4 0.012654      3   0.72588 1.3720 0.24647
## 5 0.010000      4   0.71323 1.3502 0.22616
```



As before, we calculate the consistency of the sample by predicting the annual salary from the training set and the quality of the sample by predicting the annual salary from the study set. We use RMSE and MAE to calculate predictive precision.

in sample accuracy:

```
tree_pred_in <- predict(D_tree, newdata = train, interval = "prediction", level = 0.90)

tree_rmse_in <- mean((train$annual_salary - tree_pred_in)^2)

sqrt(tree_rmse_in)
```

```
## [1] 21852.49
```

```
tree_mae_in <- mean(abs(train$annual_salary- tree_pred_in))
tree_mae_in
```

```
## [1] 17347.06
```

Out of Sample Accuracy :

```
tree_pred_out <- predict(D_tree, newdata = test, interval = "prediction", levels = 0.90)
tree_rmse_out <- mean((test$annual_salary - tree_pred_out)^2)
sqrt(tree_rmse_out)
```

```
## [1] 28180.3
```

```
tree_mae_out <- mean(abs(test$annual_salary - tree_pred_out))
tree_mae_out
```

```
## [1] 23143.74
```

All RMSE and MAE are fairly similar in and out of the study and, however, both are significantly lower for this model than for the linear model, in the region of 15000 and 30000. This suggests that as suspected, the relationships of annual salary with some of the predictors being used (such as age) were not actually linear.

Conclusion

The tree-based model has better statistical precision, making it the preferred model for ANZ to classify consumers in sales brackets for reporting purposes.

The tree based model has greter predictive accuracy, so it is the recommended model for ANZ to segment customers into income brackets for reporting purposes. More increase in precision may be feasible by strategies such as pruning, cross-validation, experimenting with parameters (e.g. amount of measurements required for splitting, different value of cost-complexity factors) and a consensus tree-based model such as random forest.