Gapsquare Client 1.5 EDA

Gapsquare

```
client <- dbReadTable(con, "client_ethnic_1_5")
head(client)
```

```
##   hourly_salary full_time_equivalent annual_salary bonus_non_voucher
## 1           294                    0        534843                 0
## 2           235                    0        427864                 0
## 3           220                    1        400000            425000
## 4           192                    1        350004                 0
## 5           175                    1        318609            150000
## 6           151                    1        275002                 0
##   total_bonus bonus_tf gender   location          division
## 1           0    FALSE Female     London Business Services
## 2           0    FALSE   Male     London Business Services
## 3      425000     TRUE   Male Birmingham               CSG
## 4           0    FALSE   Male     London               BAG
## 5      150000     TRUE   Male     London Business Services
## 6           0    FALSE Female     London         Insurance
##                                  post_name        job_level job_group
## 1                                 Director         Director         6
## 2                                 Director         Director         6
## 3 Chief Executive Officer - Claims Solutions         Director         6
## 4                          Salaried Partner Salaried Partner         6
## 5                          Finance Director         Director         6
## 6                          Salaried Partner Salaried Partner         6
##                                ethnic_origin ethnic_grouping white_bame
## 1 British/English/Welsh/Northern Irish/Scottish           White      White
## 2 British/English/Welsh/Northern Irish/Scottish           White      White
## 3                   Any other White background           White      White
## 4                   Any other White background           White      White
## 5 British/English/Welsh/Northern Irish/Scottish           White      White
## 6 British/English/Welsh/Northern Irish/Scottish           White      White
##   total_package   category age years_service length_service employee
## 1      534843.4    Support  73             5     01/09/2016        1
## 2      427863.8    Support  60             2     01/02/2019        2
## 3      824999.6 Fee-Earner  44            14     08/10/2007        3
## 4      350004.2 Fee-Earner  49             2     03/09/2019        4
## 5      468609.2    Support  59             7     14/04/2014        5
## 6      275002.0 Fee-Earner  49             2     01/08/2019        6
```

```
dbClearResult(res)
```

```
## Error in h(simpleError(msg, call)): error in evaluating the argument 'res' in selecting a method for
```

```
dbDisconnect(con)
```

```
summary(client)
```

```
##  hourly_salary    full_time_equivalent annual_salary    bonus_non_voucher
##  Min.   :  7.00   Min.   :0.0000       Min.   : 11994   Min.   :     0.0
##  1st Qu.: 14.00   1st Qu.:1.0000       1st Qu.: 24707   1st Qu.:     0.0
##  Median : 21.00   Median :1.0000       Median : 38557   Median :   499.5
##  Mean   : 27.47   Mean   :0.9862       Mean   : 49994   Mean   :  2988.6
##  3rd Qu.: 34.00   3rd Qu.:1.0000       3rd Qu.: 62754   3rd Qu.:  2857.0
##  Max.   :294.00   Max.   :1.0000       Max.   :534843   Max.   :425000.0
##   total_bonus        bonus_tf          gender            location
##  Min.   :     0.0   Mode :logical   Length:2244        Length:2244
##  1st Qu.:   157.0   FALSE:393       Class :character   Class :character
##  Median :   649.5   TRUE :1851      Mode  :character   Mode  :character
##  Mean   :  3107.9
##  3rd Qu.:  3000.0
##  Max.   :425000.0
##    division          post_name          job_level           job_group
##  Length:2244        Length:2244        Length:2244        Min.   :1.000
##  Class :character   Class :character   Class :character   1st Qu.:1.000
##  Mode  :character   Mode  :character   Mode  :character   Median :2.000
##                                                           Mean   :1.788
##                                                           3rd Qu.:2.000
##                                                           Max.   :6.000
##  ethnic_origin      ethnic_grouping     white_bame         total_package
##  Length:2244        Length:2244        Length:2244        Min.   : 11994
##  Class :character   Class :character   Class :character   1st Qu.: 25652
##  Mode  :character   Mode  :character   Mode  :character   Median : 40161
##                                                           Mean   : 53102
##                                                           3rd Qu.: 65723
##                                                           Max.   :825000
##    category             age        years_service    length_service
##  Length:2244        Min.   :19.00   Min.   : 1.000   Length:2244
##  Class :character   1st Qu.:32.00   1st Qu.: 3.000   Class :character
##  Mode  :character   Median :38.00   Median : 5.000   Mode  :character
##                     Mean   :39.91   Mean   : 8.088
##                     3rd Qu.:48.00   3rd Qu.:11.000
##                     Max.   :79.00   Max.   :56.000
##     employee
##  Min.   :   1.0
##  1st Qu.: 561.8
##  Median :1122.5
##  Mean   :1122.5
##  3rd Qu.:1683.2
##  Max.   :2244.0
```

```
summaryClient<-describe(client, quant=c(.25,.75) ,skew=TRUE)
```

```
## Warning in FUN(newX[, i], ...): no non-missing arguments to min; returning Inf
```

```
## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning -Inf
```

```
summaryClient<-summaryClient[,c(2:5,8,14,15,9)]
kbl(summaryClient)%>%kable_classic(full_width = F, html_font = "Cambria")
```

	n	mean	sd	median	min	Q0.25	Q0.75	max
hourly_salary	2244	2.746613e+01	2.164232e+01	21.00	7.0	14.00	34.00	294.0
full_time_equivalent	2244	9.861854e-01	1.167469e-01	1.00	0.0	1.00	1.00	1.0
annual_salary	2244	4.999367e+04	3.942522e+04	38557.00	11994.0	24706.75	62754.00	534843.0
bonus_non_voucher	2244	2.988571e+03	1.167405e+04	499.50	0.0	0.00	2857.00	425000.0
total_bonus	2244	3.107931e+03	1.166807e+04	649.50	0.0	157.00	3000.00	425000.0
bonus_tf	2244	NaN	NA	NA	Inf	NA	NA	-Inf
gender*	2244	1.351159e+00	4.774388e-01	1.00	1.0	1.00	2.00	2.0
location*	2244	5.128342e+00	2.738819e+00	3.00	1.0	3.00	7.00	11.0
division*	2244	3.266043e+00	1.347406e+00	3.00	1.0	2.00	4.00	6.0
post_name*	2244	1.967540e+02	9.977346e+01	221.00	1.0	113.75	274.00	330.0
job_level*	2244	9.117647e+00	4.002949e+00	9.00	1.0	7.00	11.00	16.0
job_group	2244	1.787879e+00	1.030742e+00	2.00	1.0	1.00	2.00	6.0
ethnic_origin*	2244	1.018627e+01	2.797692e+00	10.00	1.0	10.00	10.00	20.0
ethnic_grouping*	2244	5.323975e+00	1.474326e+00	6.00	1.0	6.00	6.00	6.0
white_bame*	2244	2.668895e+00	6.877661e-01	3.00	1.0	3.00	3.00	3.0
total_package	2244	5.310185e+04	4.559283e+04	40160.96	11993.8	25651.96	65723.31	824999.6
category*	2244	1.332442e+00	4.711935e-01	1.00	1.0	1.00	2.00	2.0
age	2244	3.991355e+01	1.079620e+01	38.00	19.0	32.00	48.00	79.0
years_service	2244	8.088235e+00	7.163977e+00	5.00	1.0	3.00	11.00	56.0
length_service*	2244	4.956016e+02	2.921657e+02	471.00	1.0	243.75	751.25	1023.0
employee	2244	1.122500e+03	6.479313e+02	1122.50	1.0	561.75	1683.25	2244.0

Exploratory Analysis

```
X<-describe(client, quant=c(.25,.75) ,skew=TRUE)
```

```
## Warning in FUN(newX[, i], ...): no non-missing arguments to min; returning Inf
```

```
## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning -Inf
```

```
#colnames(X)
Xprint<-X[,c(2:5,8,14,15,9)]

library(kableExtra)
Xprint %>%
  kbl() %>%
  kable_styling()
```

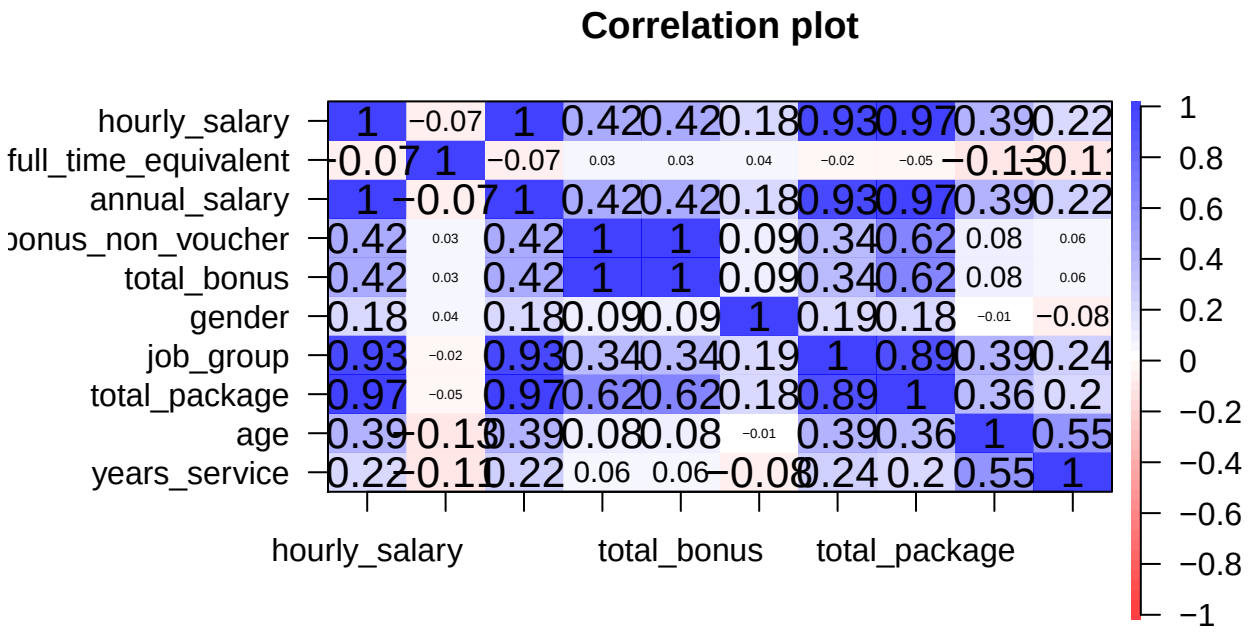
```
completeClient <- client[,c(-21)] #removed bonus_tf and employee, as it was just a count
completeClient$gender <- as.numeric(factor(completeClient$gender))#Men1, Female2

client_numeric <-dplyr::select_if(completeClient, is.numeric)
M<-cor(client_numeric)
corrplot(cor(M),tl.col = 'black',type = "upper")
```

	n	mean	sd	median	min	Q0.25	Q0.75	max
hourly_salary	2244	2.746613e+01	2.164232e+01	21.00	7.0	14.00	34.00	294.0
full_time_equivalent	2244	9.861854e-01	1.167469e-01	1.00	0.0	1.00	1.00	1.0
annual_salary	2244	4.999367e+04	3.942522e+04	38557.00	11994.0	24706.75	62754.00	534843.0
bonus_non_voucher	2244	2.988571e+03	1.167405e+04	499.50	0.0	0.00	2857.00	425000.0
total_bonus	2244	3.107931e+03	1.166807e+04	649.50	0.0	157.00	3000.00	425000.0
bonus_tf	2244	NaN	NA	NA	Inf	NA	NA	-Inf
gender*	2244	1.351159e+00	4.774388e-01	1.00	1.0	1.00	2.00	2.0
location*	2244	5.128342e+00	2.738819e+00	3.00	1.0	3.00	7.00	11.0
division*	2244	3.266043e+00	1.347406e+00	3.00	1.0	2.00	4.00	6.0
post_name*	2244	1.967540e+02	9.977346e+01	221.00	1.0	113.75	274.00	330.0
job_level*	2244	9.117647e+00	4.002949e+00	9.00	1.0	7.00	11.00	16.0
job_group	2244	1.787879e+00	1.030742e+00	2.00	1.0	1.00	2.00	6.0
ethnic_origin*	2244	1.018627e+01	2.797692e+00	10.00	1.0	10.00	10.00	20.0
ethnic_grouping*	2244	5.323975e+00	1.474326e+00	6.00	1.0	6.00	6.00	6.0
white_bame*	2244	2.668895e+00	6.877661e-01	3.00	1.0	3.00	3.00	3.0
total_package	2244	5.310185e+04	4.559283e+04	40160.96	11993.8	25651.96	65723.31	824999.6
category*	2244	1.332442e+00	4.711935e-01	1.00	1.0	1.00	2.00	2.0
age	2244	3.991355e+01	1.079620e+01	38.00	19.0	32.00	48.00	79.0
years_service	2244	8.088235e+00	7.163977e+00	5.00	1.0	3.00	11.00	56.0
length_service*	2244	4.956016e+02	2.921657e+02	471.00	1.0	243.75	751.25	1023.0
employee	2244	1.122500e+03	6.479313e+02	1122.50	1.0	561.75	1683.25	2244.0

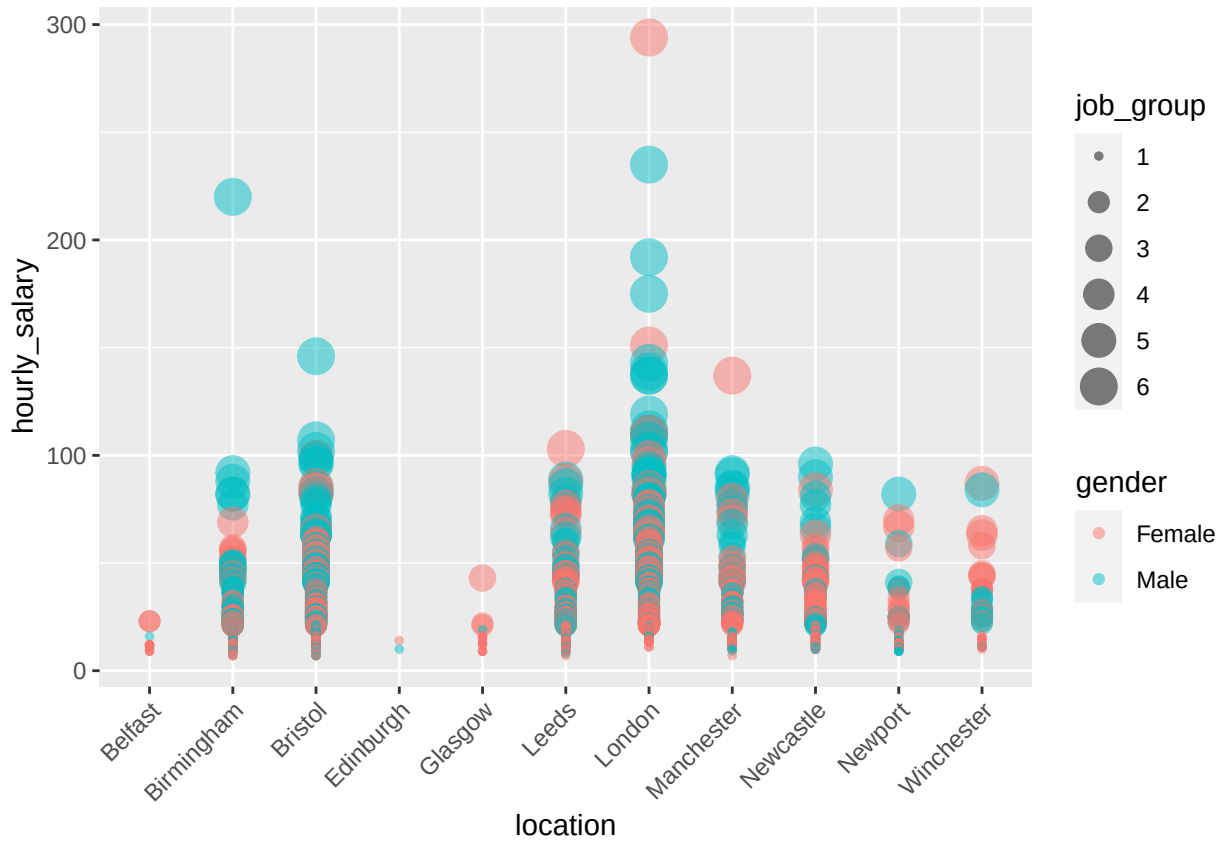
Correlations between all the salary tabs - to be expected, as they are all calculated from one another. Job group and salary are also correlated.

Gender having weak correlations with other factors.
It looks like, there may be more explanation with categorical variables.

```
cor.plot(client_numeric)
```

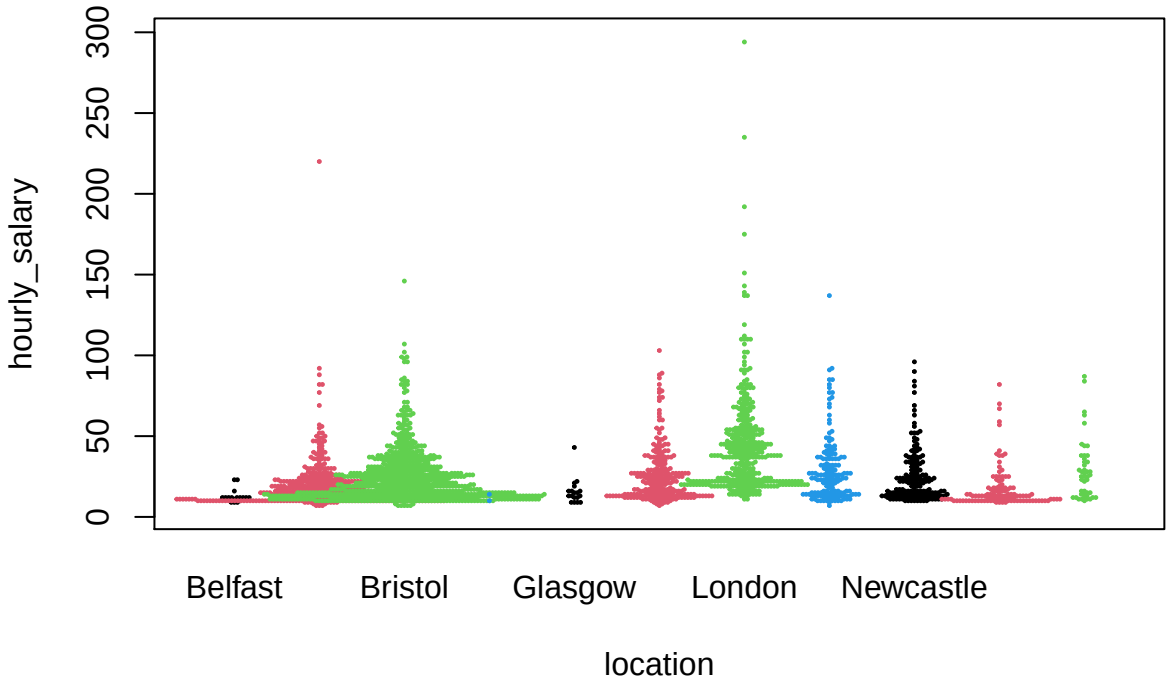



```
client%>%
  ggplot(aes(location, hourly_salary, color = gender, size = job_group )) +
  geom_point(alpha = 0.5)  + theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1))
```

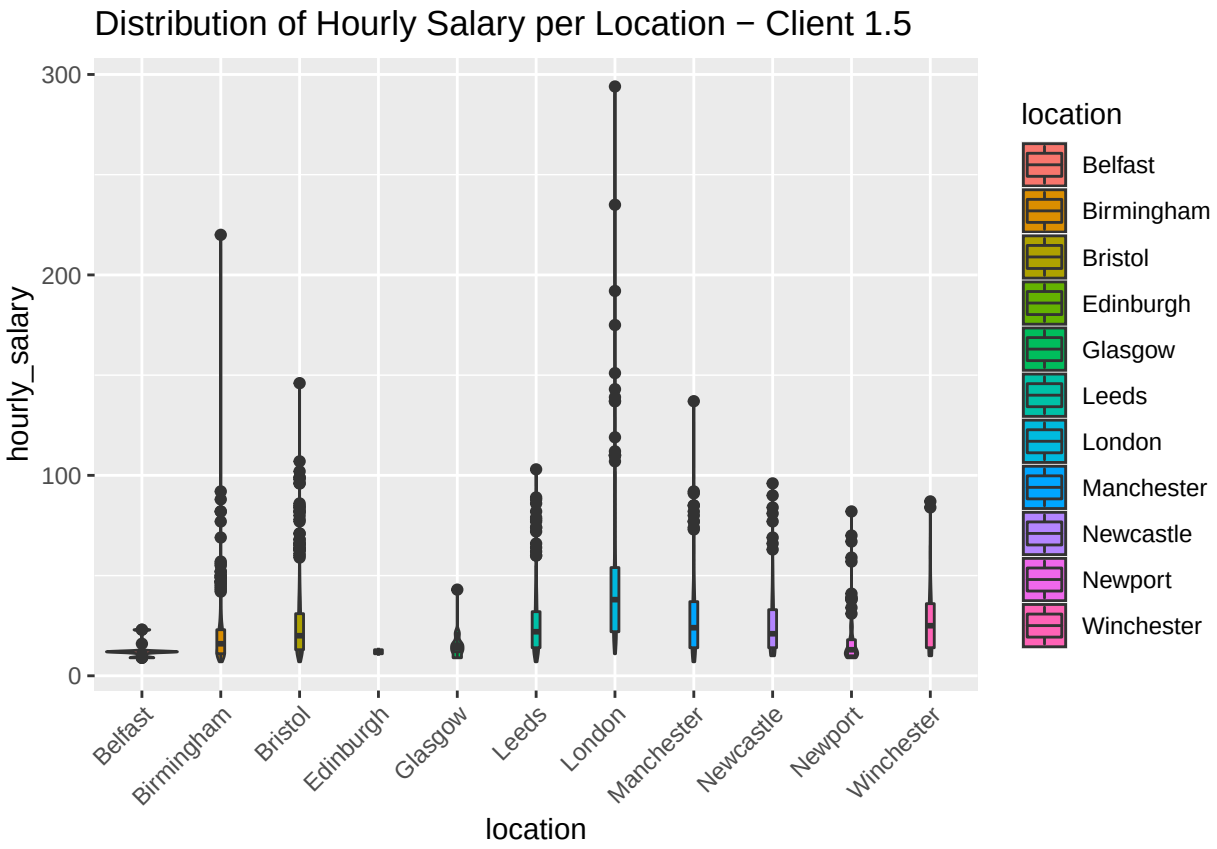


More people working in London / Bristol / Birmingham / Manchester.
This is where the higher salaries seem to be prominent.
Seems to be an outlier in the Birmingham office, probably a director.
Someone with a higher job group seems to be paid a lot less in the London office.

```
beeswarm(hourly_salary ~ location, data=client, col = 1:4, pch=19, method="swarm", cex=0.2)
```




```
D <- ggplot(client, aes(x=location, y=hourly_salary, fill=location)) +
  geom_violin() + theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1))
D + geom_boxplot(width=0.1) + labs(title = "Distribution of Hourly Salary per Location - Client 1.5")
```

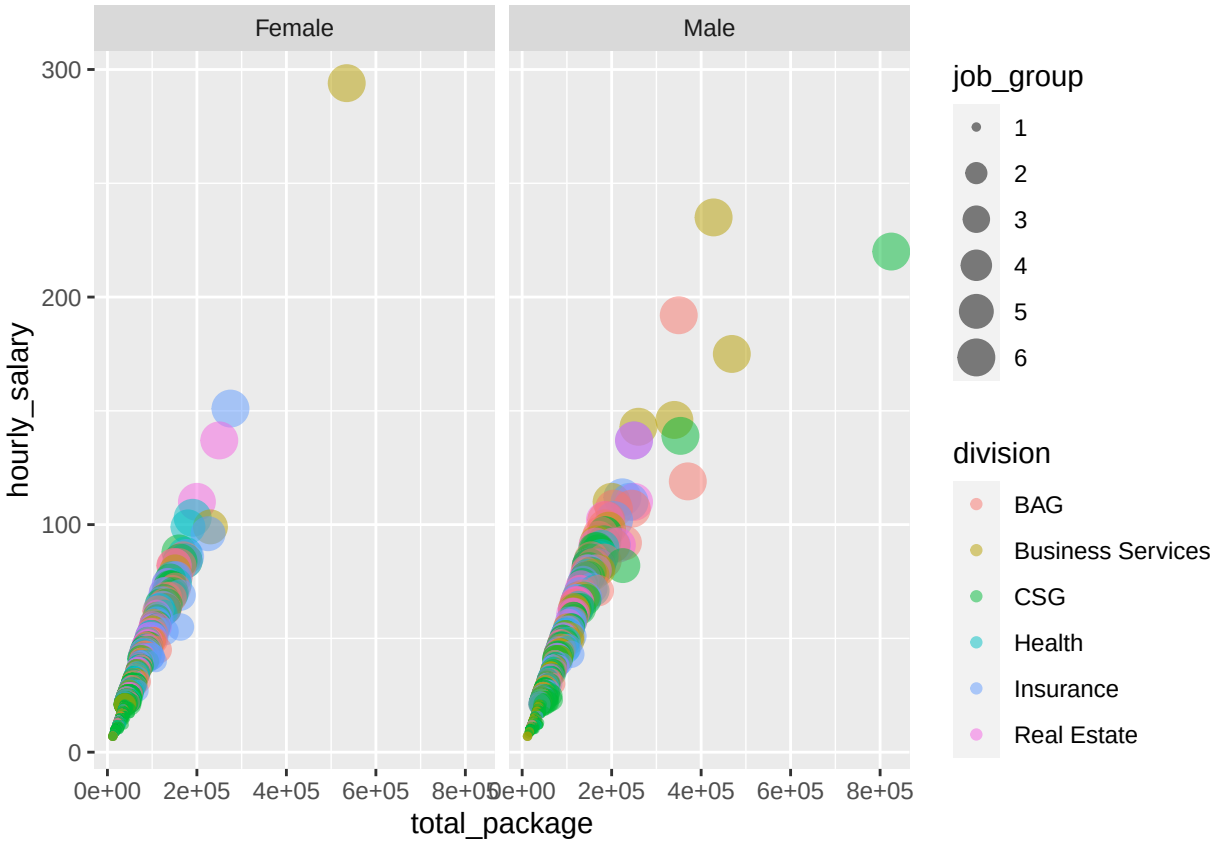


Different version of the graph above. Showing the variation by location, celar to see some of the more extreme results.

```
client%>%
  ggplot(aes(total_package,hourly_salary, color = ethnic_origin, size = job_group)) +
  geom_point(alpha = 0.5) + facet_wrap(~gender)
```

Split by ethnic origin, as expected, strong positive linear correlation with salary and bonus

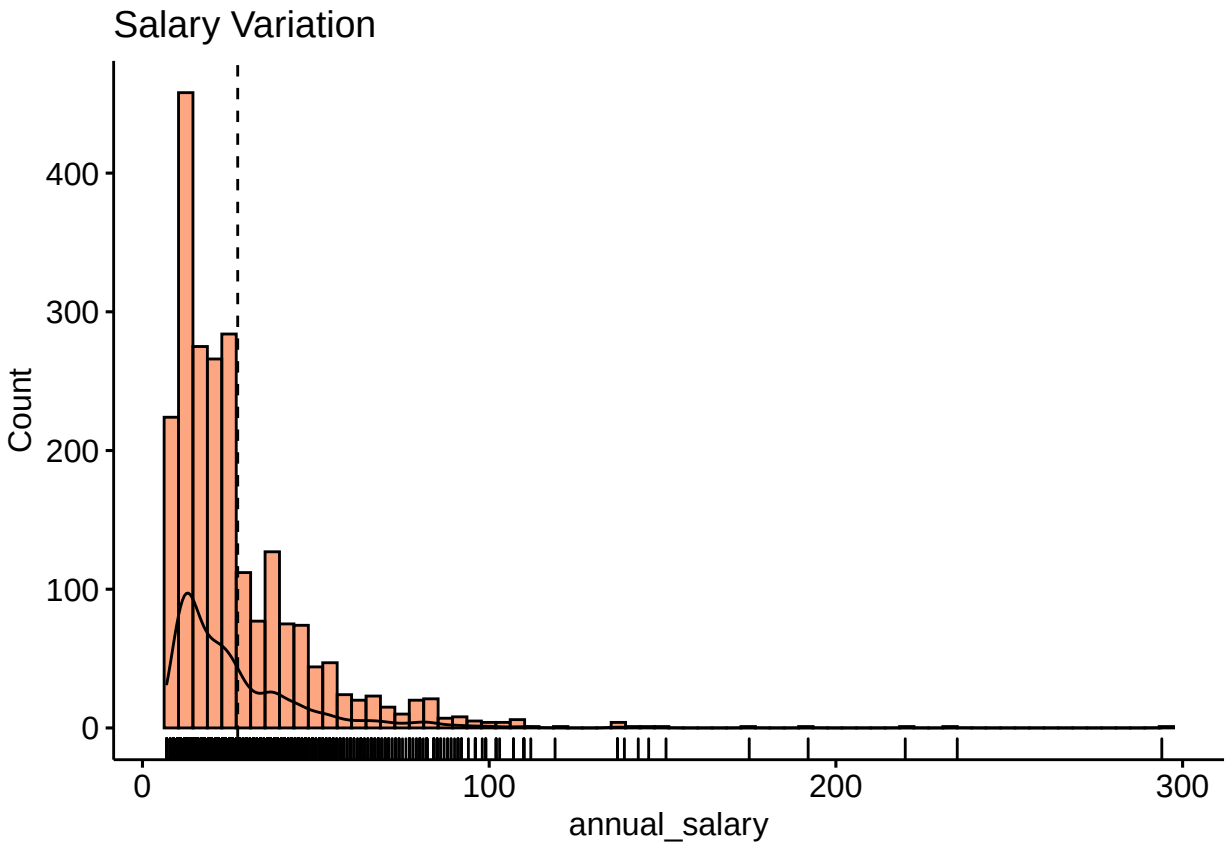
```
client%>%
  ggplot(aes(total_package, hourly_salary, colour = division, size=job_group)) +
  geom_point(alpha = 0.5) + facet_wrap(~gender)
```




```
library(ggpubr)
gghistogram(client, x = "hourly_salary",
            fill = "#FC4E07",
            add = "mean", rug = TRUE,
            bins=70, add_density = TRUE,
            xlab = "annual_salary", ylab = "Count",
            title = "Salary Variation")
```

```
## Warning: geom_vline(): Ignoring ‘mapping‘ because ‘xintercept‘ was provided.
```

```
## Warning: geom_vline(): Ignoring ‘data‘ because ‘xintercept‘ was provided.
```



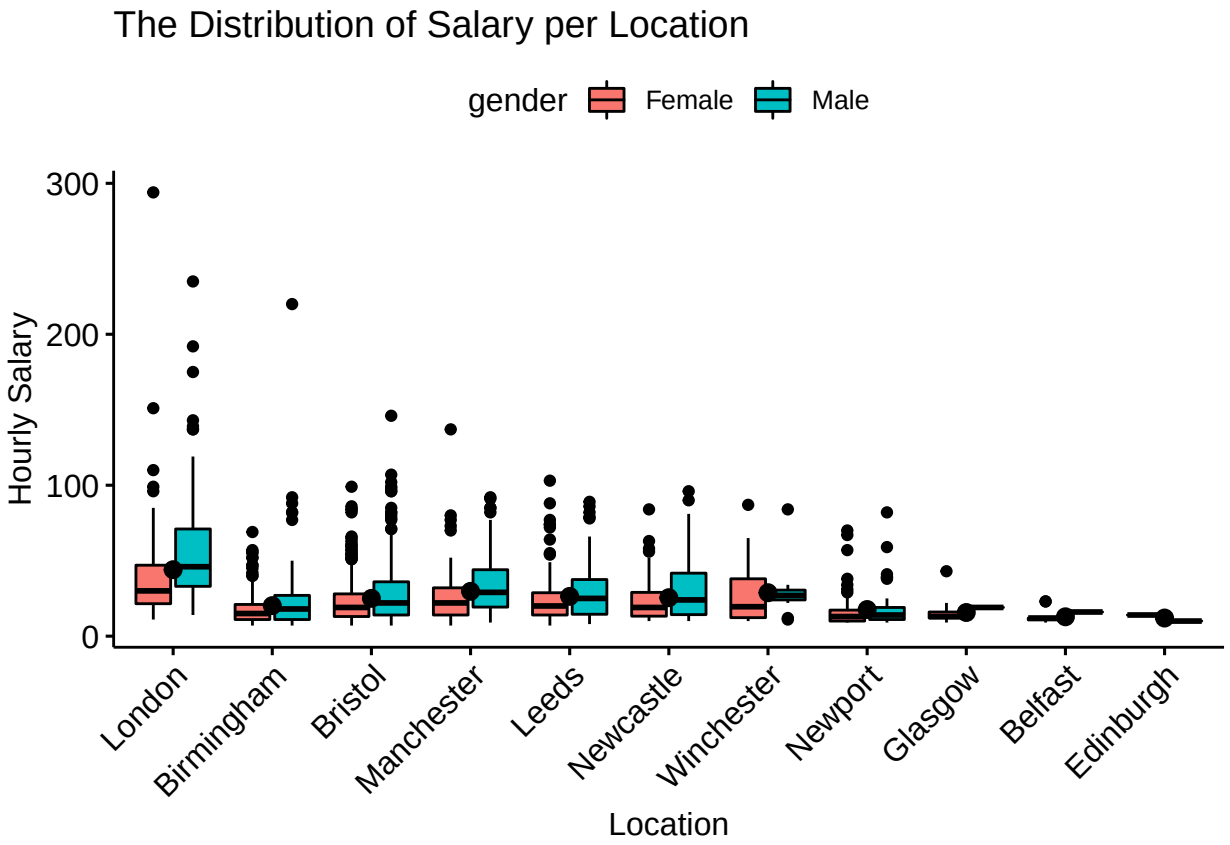

```
ggboxplot(client, x="location", y="hourly_salary",group = "gender",
            fill = "gender",
            add = "mean", rug = TRUE,
            bins=15, add_density = TRUE,
            xlab = "Location", ylab = "Hourly Salary",
            title = "The Distribution of Salary per Location")  + theme(axis.text.x = element_text(angl
```

```
## Warning: Ignoring unknown parameters: bins
```

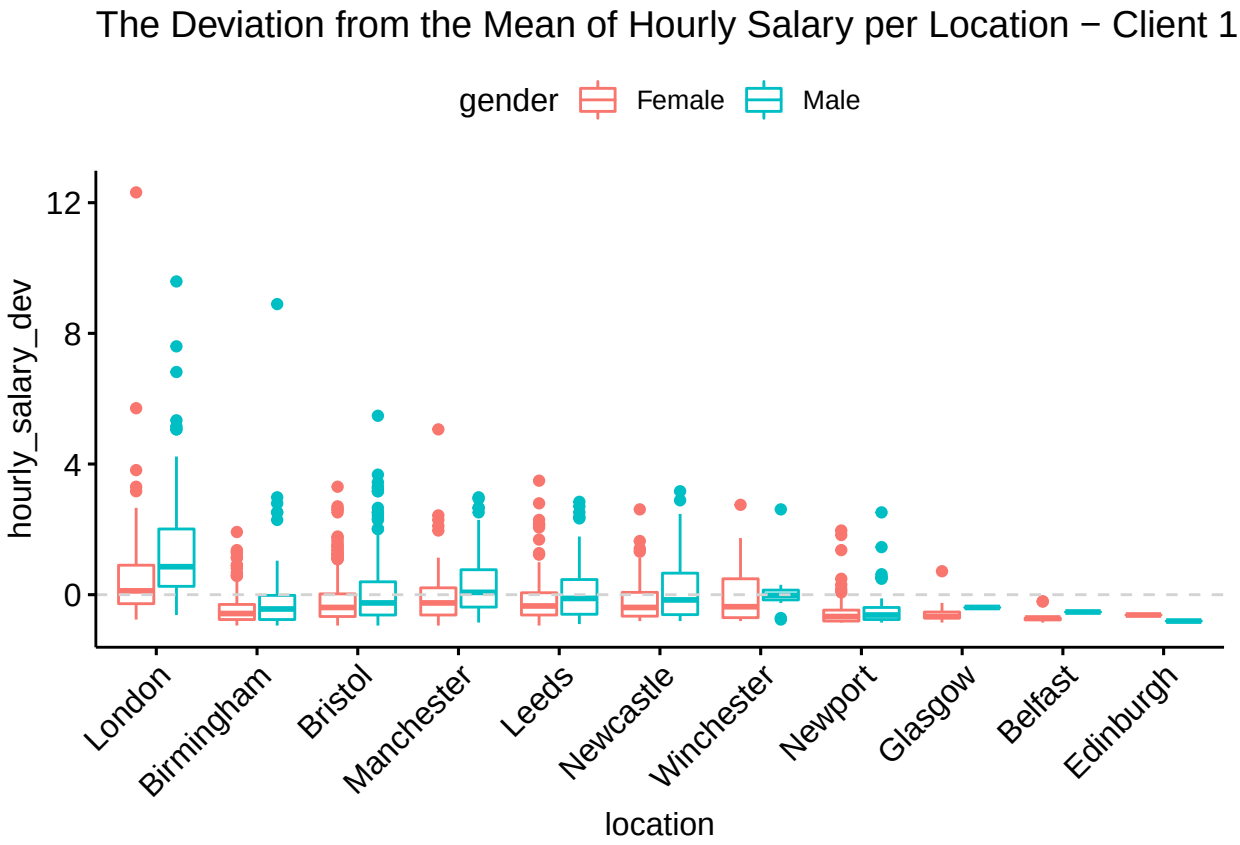
```
## Warning: `fun.y` is deprecated. Use `fun` instead.
```

```
## Warning: `fun.ymin` is deprecated. Use `fun.min` instead.
```

```
## Warning: `fun.ymax` is deprecated. Use `fun.max` instead.
```




```
client$hourly_salary_dev <- (client$hourly_salary - mean(client$hourly_salary)) / sd(client$hourly_sala

ggboxplot(client,
          x = "location",
          y = "hourly_salary_dev",
          title = "The Deviation from the Mean of Hourly Salary per Location - Client 1.5",
          color = "gender",
          sorting = "descending",
          add.params = list(color = "lightgray", size = 1),
          group = "location",
          dot.size = 4) +
 geom_hline(yintercept = 0, linetype = 2, color = "lightgray") + theme(axis.text.x = element_text(angl
```



Deviation from the mean. Seem to be belong average in Glasgow / Belfast / Edinburgh (smaller offices?) Above the average in the larger cities, and male workers more often Birmingham, is generally under the average, apart from the few potential outliers Its only really females in london that are above average

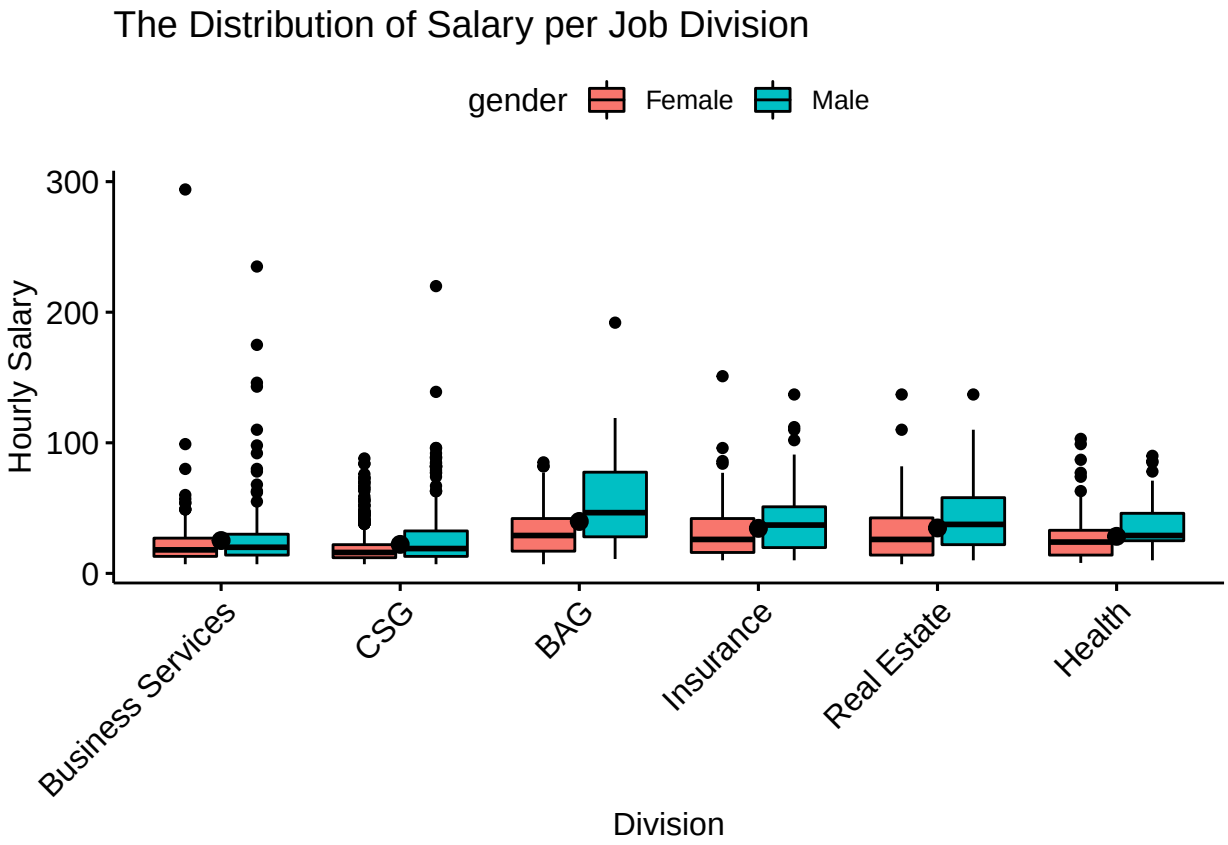
```
ggboxplot(client, x="division", y="hourly_salary",group = "gender",
            fill = "gender",
            add = "mean", rug = TRUE,
            bins=15, add_density = TRUE,
            xlab = "Division", ylab = "Hourly Salary",
            title = "The Distribution of Salary per Job Division")  + theme(axis.text.x = element_text(a
```

```
## Warning: Ignoring unknown parameters: bins
```

```
## Warning: `fun.y` is deprecated. Use `fun` instead.
```

```
## Warning: `fun.ymin` is deprecated. Use `fun.min` instead.
```

```
## Warning: `fun.ymax` is deprecated. Use `fun.max` instead.
```


The Distribution of Salary per Job Division
gender
Female
Male
300
200
100
0
Hourly Salary
Business Services
CSG
BAG
Insurance
Real Estate
Health
Division