EDA on San Francisco Salaries

Aouidane Imed eddine

```
library(readr)
library(tidyverse)
```

```
## Warning: le package 'tidyverse' a été compilé avec la version R 4.3.2
```

```
## Warning: le package 'ggplot2' a été compilé avec la version R 4.3.3
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v purrr     1.0.2
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.5.0     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## -- 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
```

About the dataset :

- this data contains the names, job title, and compensation for San Francisco city employees on an annual basis from 2011 to 2014.

```
salaries <- read_csv("Salaries.csv")
dim(salaries)
```

```
## [1] 148654     13
```

```
head(salaries)
```

```
## # A tibble: 6 x 13
##      Id EmployeeName     JobTitle BasePay OvertimePay OtherPay Benefits TotalPay
##   <dbl> <chr>            <chr>      <dbl>       <dbl>    <dbl>    <dbl>    <dbl>
## 1     1 NATHANIEL FORD   GENERAL~ 167411.          0   400184.       NA  567595.
## 2     2 GARY JIMENEZ     CAPTAIN~ 155966.     245132.  137811.       NA  538909.
## 3     3 ALBERT PARDINI   CAPTAIN~ 212739.     106088.   16453.       NA  335280.
## 4     4 CHRISTOPHER CHO~ WIRE RO~  77916       56121.  198307.       NA  332344.
## 5     5 PATRICK GARDNER  DEPUTY ~ 134402.       9737   182235.       NA  326373.
## 6     6 DAVID SULLIVAN   ASSISTA~ 118602        8601   189083.       NA  316286.
## # i 5 more variables: TotalPayBenefits <dbl>, Year <dbl>, Notes <lgl>,
## #   Agency <chr>, Status <chr>
```

- The dataset contains 13 columns and 148654 rows. The columns are:
    1. Id
    2. EmployeeName
    3. JobTitle
    4. BasePay
    5. OvertimePay
    6. OtherPay
    7. Benefits
    8. TotalPay
    9. TotalPayBenefits
    10. Year
    11. Notes
    12. Agency
    13. Status

Removing unnecessary columns:

```
salaries <- salaries %>%
  select(-c(Notes, Agency,Id,EmployeeName,Year))
```

Exploratory Data Analysis:

```
summary(salaries)
```

```
##    JobTitle            BasePay        OvertimePay          OtherPay
##  Length:148654      Min.   :  -166   Min.   :    -0.01   Min.   : -7058.6
##  Class :character   1st Qu.: 33588   1st Qu.:     0.00   1st Qu.:     0.0
##  Mode  :character   Median : 65007   Median :     0.00   Median :   811.3
##                     Mean   : 66325   Mean   :  5066.06   Mean   :  3648.8
##                     3rd Qu.: 94691   3rd Qu.:  4658.18   3rd Qu.:  4236.1
##                     Max.   :319275   Max.   :245131.88   Max.   :400184.2
##                     NA's   :609      NA's   :4           NA's   :4
##     Benefits           TotalPay       TotalPayBenefits      Status
##  Min.   :  -33.89   Min.   :  -618.1   Min.   :  -618.1   Length:148654
##  1st Qu.:11535.40   1st Qu.: 36169.0   1st Qu.: 44065.7   Class :character
##  Median :28628.62   Median : 71426.6   Median : 92404.1   Mode  :character
##  Mean   :25007.89   Mean   : 74768.3   Mean   : 93692.6
##  3rd Qu.:35566.86   3rd Qu.:105839.1   3rd Qu.:132876.5
##  Max.   :96570.66   Max.   :567595.4   Max.   :567595.4
##  NA's   :36163
```

- We can observe that the minimum salary is negative which is not possible. So we need to clean the data. # Data Cleaning:

```
salaries <- salaries %>% filter(BasePay > 0,
                                OvertimePay > 0,
                                OtherPay > 0,
```

```
                              Benefits > 0,
                              TotalPay > 0,
                              TotalPayBenefits >0)
summary(salaries)
```

```
##    JobTitle            BasePay          OvertimePay          OtherPay
##  Length:48401       Min.   :    37.6   Min.   :     0.02   Min.   :     0.59
##  Class :character   1st Qu.: 55972.0   1st Qu.:  2080.86   1st Qu.:  1189.03
##  Mode  :character   Median : 71368.0   Median :  6056.18   Median :  3794.08
##                     Mean   : 78229.7   Mean   : 11797.29   Mean   :  6432.75
##                     3rd Qu.:105506.7   3rd Qu.: 15377.36   3rd Qu.:  8652.42
##                     Max.   :318835.5   Max.   :220909.48   Max.   :203735.92
##     Benefits           TotalPay        TotalPayBenefits      Status
##  Min.   :    4.92   Min.   :   312.2   Min.   :   402.1   Length:48401
##  1st Qu.:26335.27   1st Qu.: 63216.4   1st Qu.: 89636.5   Class :character
##  Median :32194.95   Median : 88235.6   Median :119868.7   Mode  :character
##  Mean   :30337.34   Mean   : 96459.8   Mean   :126797.1
##  3rd Qu.:37049.01   3rd Qu.:128221.9   3rd Qu.:164789.8
##  Max.   :89540.23   Max.   :390112.0   Max.   :479652.2
```

- Now we have cleaned the data and removed the negative values. We can explore the data further. ## Full time jobs vs Part time jobs:

```
# Full time jobs
summary(subset(salaries,Status == "FT")$TotalPay)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   29107   74337   99481  108208  136497  390112
```

```
# Part time jobs
summary(subset(salaries,Status == "PT")$TotalPay)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1083   30116   52055   60831   82839  277582
```

- The average salary for full-time jobs is 107,000 and for part-time jobs is 31,000. This shows that full-time jobs have higher salaries compared to part-time jobs.

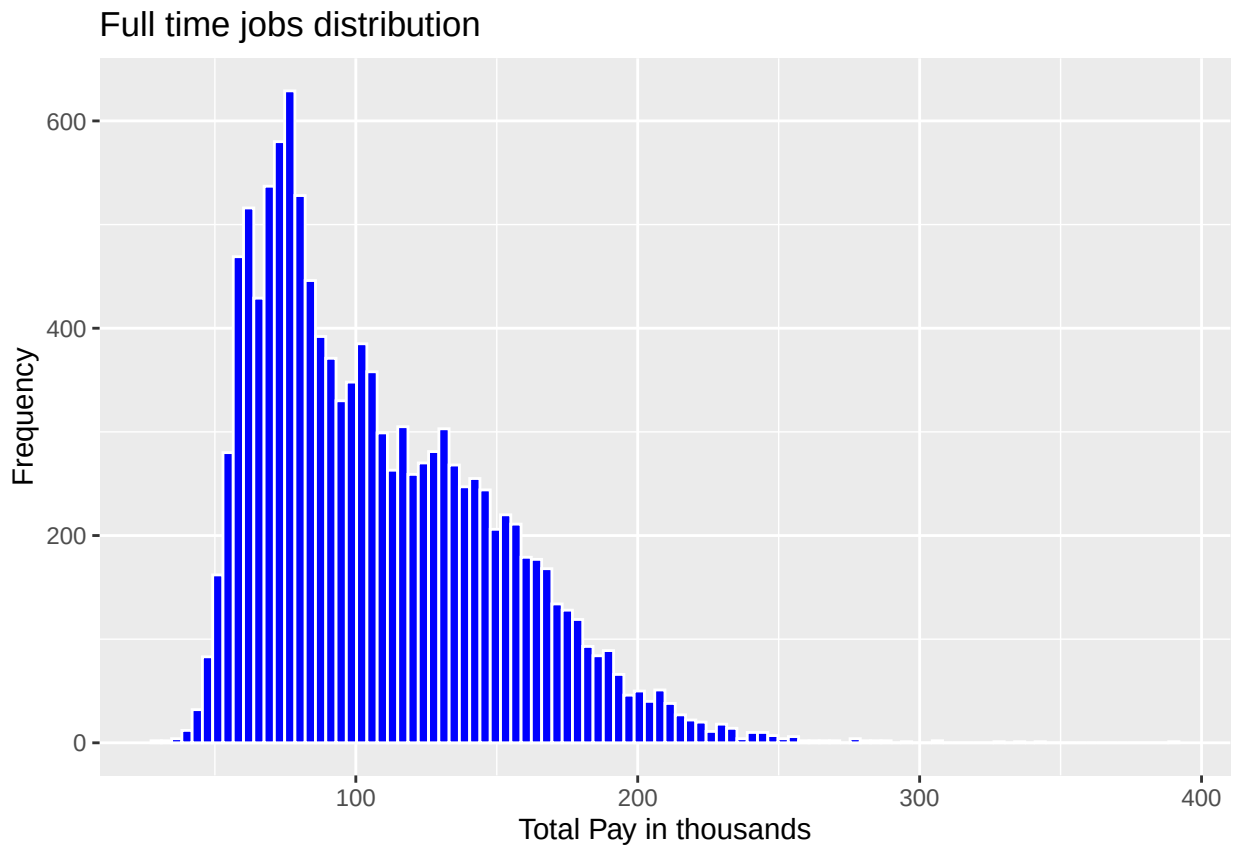
Visualizing the data:

```
salaries_ft <- salaries[which(salaries$Status == "FT"),]
salaries_pt <- salaries[which(salaries$Status == "PT"),]
cbind(dim(salaries_ft),dim(salaries_pt))
```

```
##       [,1] [,2]
## [1,] 12166 4340
## [2,]     8    8
```

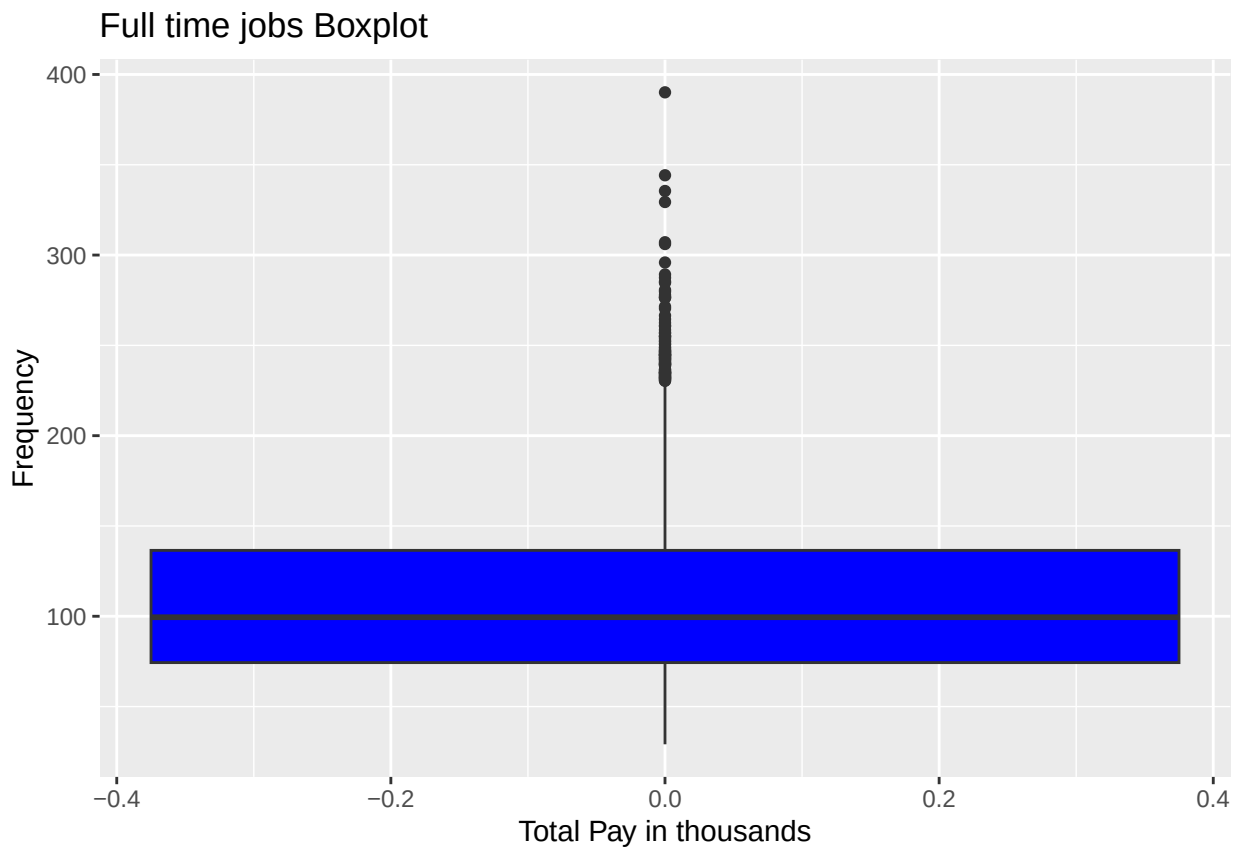
Full time jobs distribution:

```
salaries_ft %>% ggplot(.,aes(x = TotalPay/1000))+
  geom_histogram(fill = "blue",bins = 100,,color = "white")+
  labs(title = "Full time jobs distribution",
       x = "Total Pay in thousands",
       y = "Frequency")
```



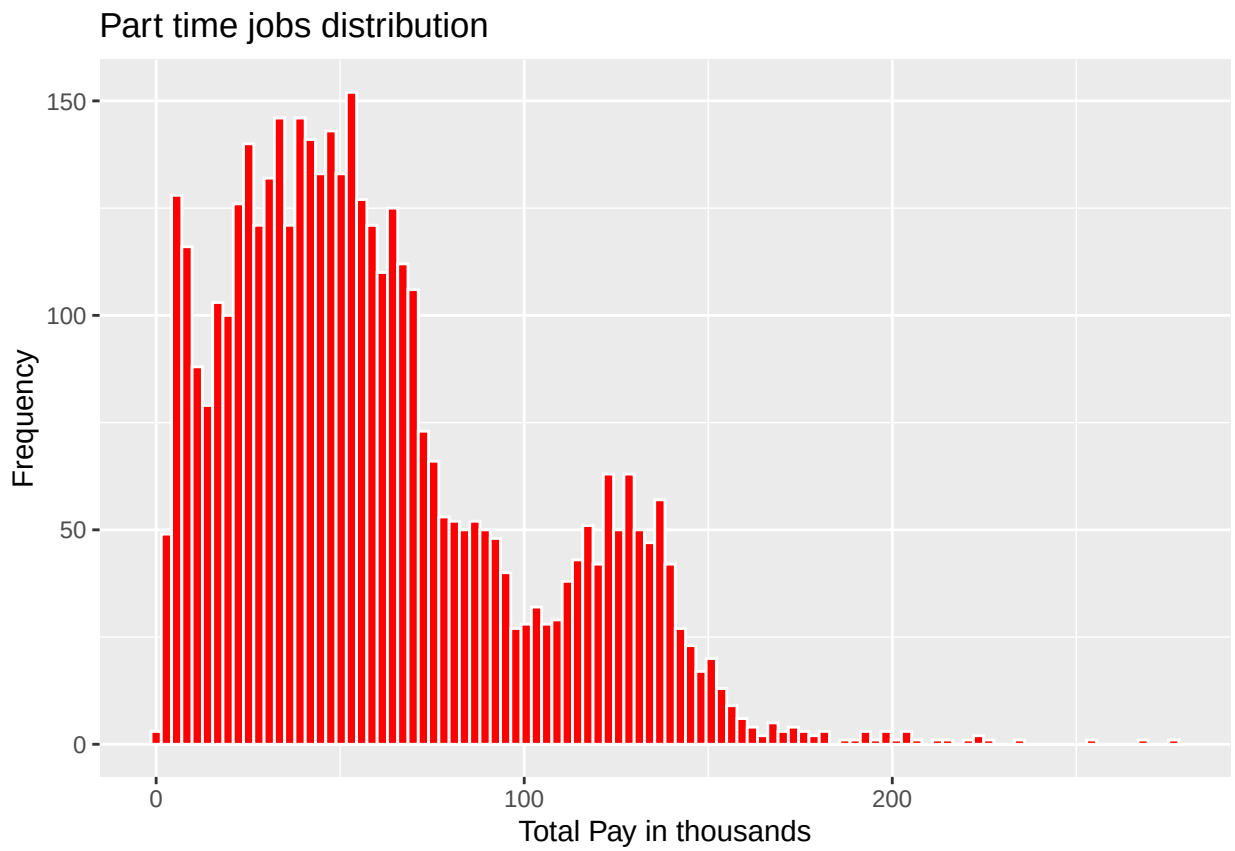
- The distribution of full-time jobs is right-skewed. Most of the employees earn between 0 to 200,000.

```
salaries_ft %>% ggplot(.,aes(y = TotalPay/1000))+
  geom_boxplot(fill = "blue")+
  labs(title = "Full time jobs Boxplot",
       x = "Total Pay in thousands",
       y = "Frequency")
```



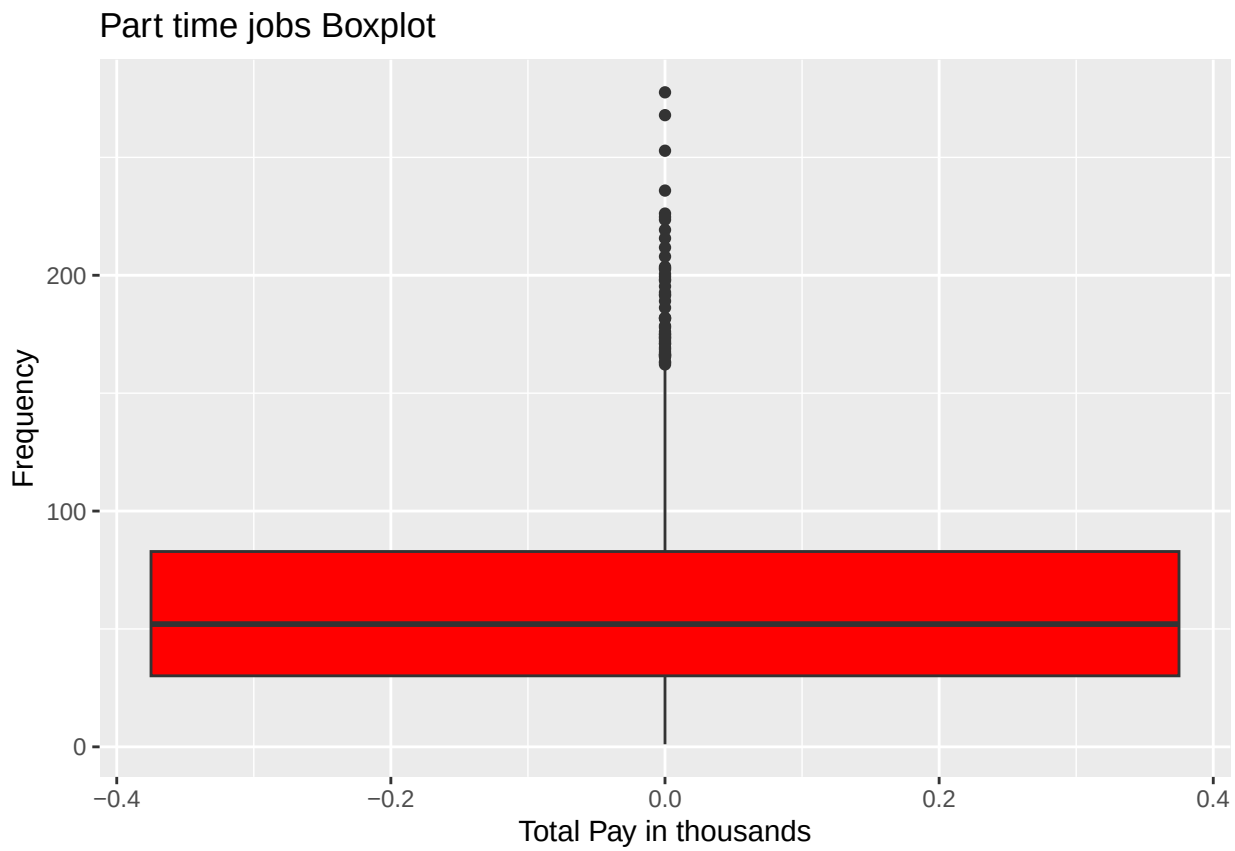
Part time jobs distribution:

```
salaries_pt %>% ggplot(.,aes(x = TotalPay/1000))+
  geom_histogram(fill = "red",bins = 100,color = "white")+
  labs(title = "Part time jobs distribution",
       x = "Total Pay in thousands",
       y = "Frequency")
```



- The distribution of part-time jobs is right-skewed. Most of the employees earn between 0 to 100,000.

```
salaries_pt %>% ggplot(.,aes(y = TotalPay/1000))+
  geom_boxplot(fill = "red")+
  labs(title = "Part time jobs Boxplot",
       x = "Total Pay in thousands",
       y = "Frequency")
```



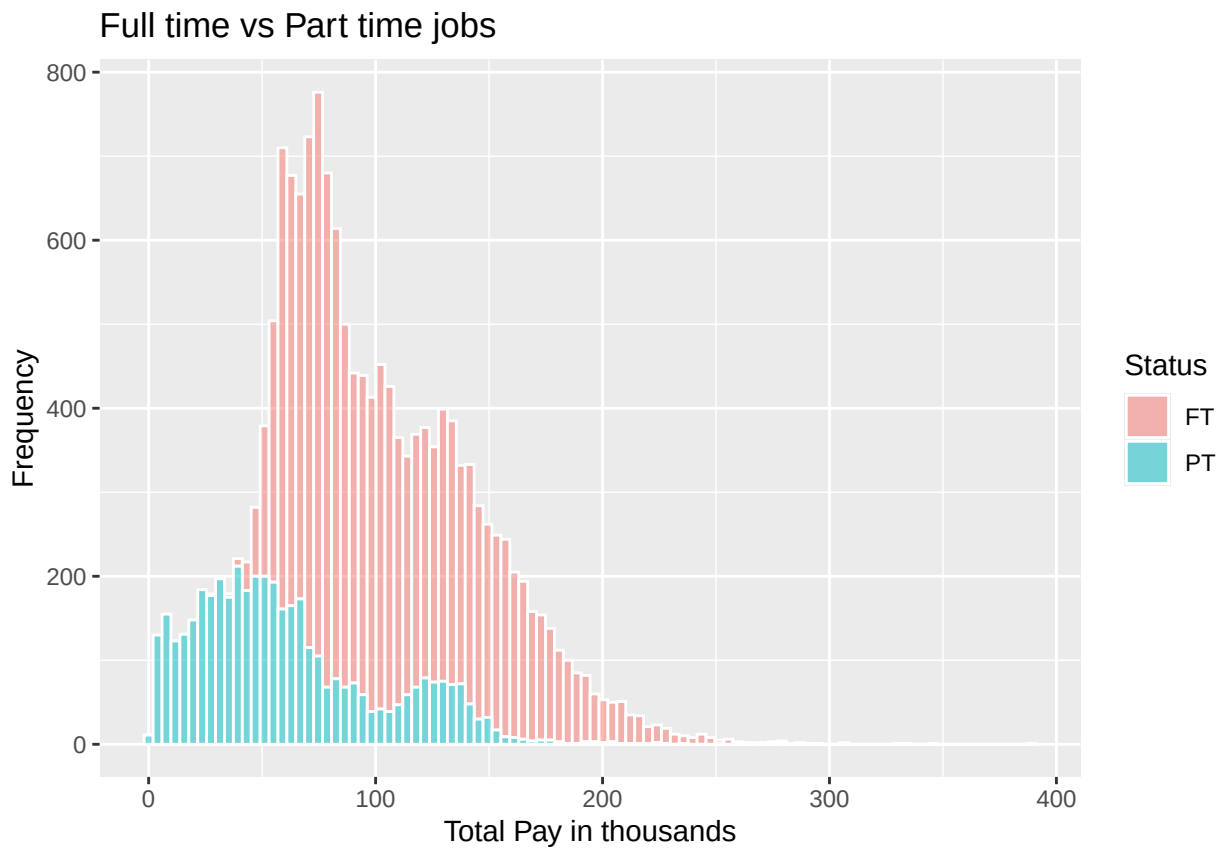
Handeling outliers:

```
salaries_ft <- salaries_ft %>% filter(TotalPay < mean(TotalPay) + 3*sd(TotalPay),
                                      TotalPay > mean(TotalPay) - 3*sd(TotalPay))

salaries_pt <- salaries_pt %>% filter(TotalPay < mean(TotalPay) + 3*sd(TotalPay),
                                      TotalPay > mean(TotalPay) - 3*sd(TotalPay))
```

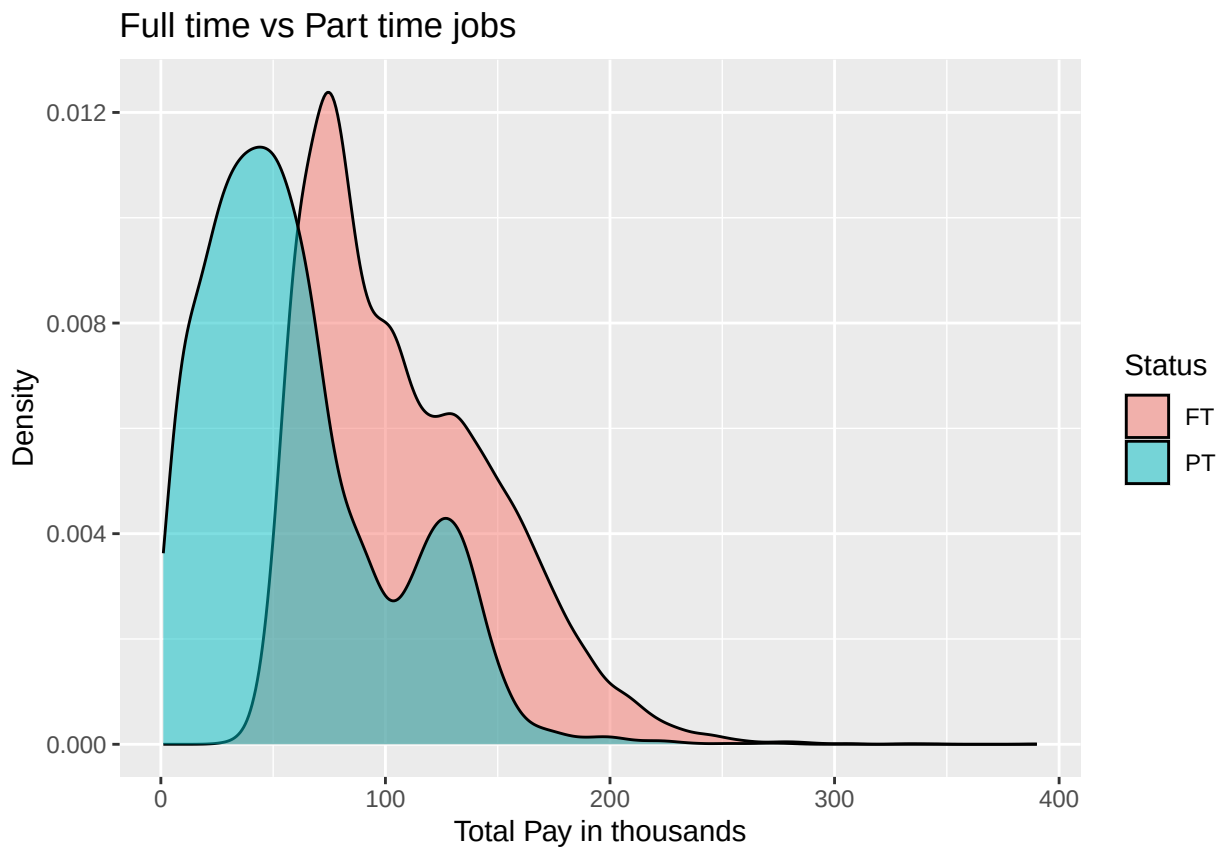
Comparing full-time and part-time jobs:

```
salaries <- salaries %>% filter(!is.na(Status))
salaries %>%
  ggplot(.,aes(x = TotalPay/1000,fill = Status))+
  geom_histogram(bins = 100,color = "white",alpha = 0.5)+
  labs(title = "Full time vs Part time jobs",
       x = "Total Pay in thousands",
       y = "Frequency")
```



Density plot:

```
salaries %>%
  ggplot(.,aes(x = TotalPay/1000,fill = Status))+
  geom_density(alpha = 0.5)+
  labs(title = "Full time vs Part time jobs",
       x = "Total Pay in thousands",
       y = "Density")
```



Conclusion:

- The average salary for full-time jobs is 107,000 and for part-time jobs is 31,000. This shows that full-time jobs have higher salaries compared to part-time jobs.