# WHO MAKES 100K+ IN TORONTO:EVIDENCE FROM SEPTEMBER 2023 LABOUR FORCE SURVEY

Gaurja Newatia

2024-04-26

### Loading the Data

## 1. Introduction

The world is currently grappling with a global recession that has left no country untouched,including Canada. Despite nominal wage increases, Canada faces the harsh reality that these wage gains have not kept pace with the rising rate of inflation.Moreover,with millions of people immigrating to Canada every year, the labour supply has drastically increased, which has profound implications on wage growth. In response to these challenges, this paper aims to uncover the key factors influencing wages in Canada and seeks to provide policymakers with valuable insights that can inform the development of policies promoting wage equity.

For the purpose of our research, we restrict our analysis to Toronto, one of Canada's largest Census Metropolitan Area (CMA), with population of 6.6 million residents. The problem of interest is to classify the surveyed population as:

- High Earners (Earning an Annual Gross Salary of $100,000 or More)
- Non-High Earners (Earning an Annual Gross Salary of less than $100,000)

## 2. Summary of Previous Work

A lot of research has been conducted in the past and continues to be ongoing, to enable policymakers formulate policies that govern wage determination and empower them address wage differentials.

Mueller, R.E. (2022) analyzed the merged monthly Labour Force Survey data to investigate the gender paygap across four distinct public and private sector definitions, as well as within each of these sectors. They argue that when compared across gender, females tend to exhibit higher wage premiums in the public sector compared to their male counterparts. Moreover, the gender paygap is consistently in favour of males within all sectors, with the disparity slighlty larger in the private sector.

Using 42 years of Current Population Survey Data (1977-2018), McConnell, et al (2023) present a novel finding: marriage has a positive causal effect on the wages of both men and women. They argue that there is a significant shift for married women, as the once-existent wage penalty associated with marriage has transformed into a wage premium. They attribute the premium for men can be predominantly attributed to household specialization, akin to Becker's theory (1985, 1993), where married men tend to focus on market work, resulting in increased market hours and higher earnings compared to their single counterparts. Historically, the wage penalty for women also stemmed from household specialization, with married women working considerably fewer market hours than singles. However, over time, decreased specialization between spouses has enabled married women to nearly match the work hours of singles, eliminating the penalty.

The Quality of Employment Report of Statistics Canada (2021) gives some insights on the state of hourly wages in Canada. It argues that even though average wages in Canada is gradually rising since 1988, certain sections of employees continued to have lower hourly wages than others in 2021. It emphasizes that being a younger employee (aged 15 to 24), having a high school diploma or less, and working part-time were all factors associated with lower hourly wages.

## 3. Description of the Data

The Labour Force Survey (LFS) is a monthly cross-sectional survey conducted by Statistics Canada which are used to produce the unemployment rate as well as other standard labour market indicators such as the employment rate and the participation rate. The monthly dataset used in this study pertains to September 2023 with a 90% response rate.

By applying a filter to the Labour Force Survey (LFS) specific to the Toronto Census Metropolitan Area (CMA), we've narrowed down our sample to encompass 5,645 individuals. This dataset comprises information on 60 variables, collectively addressing a wide range of subjects, including:

- Employment and Unemployment
- Industries
- Labour
- Occupations
- Unionization and industrial relations
- Wages, salaries and other earnings

The 'Earnings' dummy variable, which is our response variables allows us to categorize individuals based on their income level.We classify individuals as 'High Earners' if their hourly wage is greater than or equal to 48.08 [1], and as 'Non High Earners' if their hourly wage is less than 48.08 CAD. This classification is based on the assumption that someone working 40 hours a week at a wage of 48.08 CAD per hour or higher would earn an annual gross salary of $100,000 or more.

## 4. Description of Data Issues

### 4.1 Missing Values

To facilitate our analysis, we organized the dataset into numerical and categorical variables. Upon reviewing the dataset, we noticed the presence of missing values within several numerical variables, even though these missing values hold specific interpretations. For example, in the "paidot" variable, which represents the number of paid overtime hours, blank values signify that an individual does not receive paid overtime. To address this issue and ensure the inclusion of these missing values in our dataset, we made the decision to impute them with the value "999999." This imputation process was carried out using Microsoft Excel.

Similarly for categorical variables, we opted to impute missing values with "Not Applicable." This choice was made to enable the incorporation of these variables in our analysis, even when specific categorical information was absent.

Table 1 below describes the missing values after imputation

Table 1: Percentage of Missing Values

| Variable name | Number of Missing Values | Missing Values(%) |
|---|---|---|

[1] The data is taken from https://ca.talent.com/convert

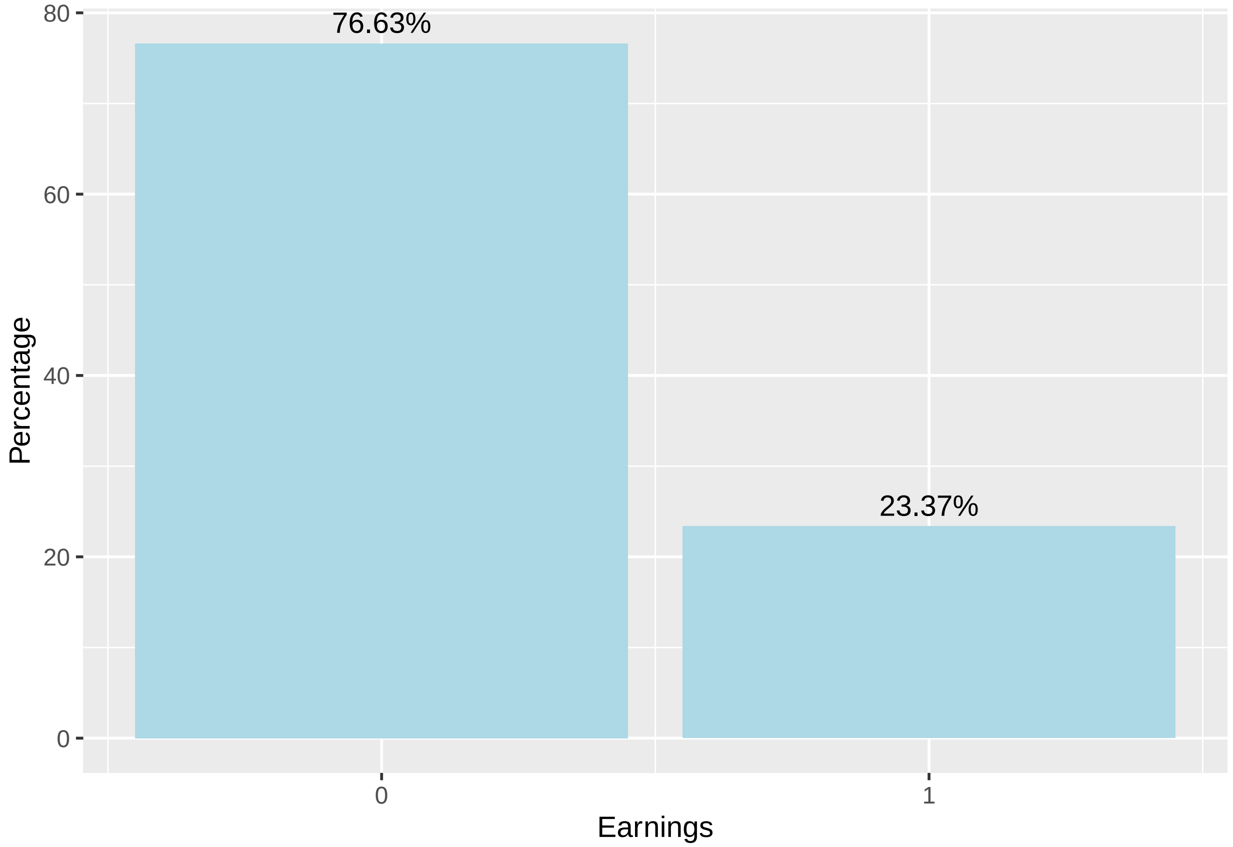
|  |  |  |
|---|---|---|
| everwork | 5645 | 100 |
| ftptlast | 5645 | 100 |
| prevten | 5645 | 100 |
| durunemp | 5645 | 100 |
| flowunem | 5645 | 100 |
| unemftpt | 5645 | 100 |
| whylefto | 5645 | 100 |
| whyleftn | 5645 | 100 |
| durjless | 5645 | 100 |
| availabl | 5645 | 100 |
| lkpubag | 5645 | 100 |
| lkemploy | 5645 | 100 |
| lkrels | 5645 | 100 |
| lkatads | 5645 | 100 |
| lkansads | 5645 | 100 |
| lkothern | 5645 | 100 |
| prioract | 5645 | 100 |
| ynolook | 5645 | 100 |
| tlolook | 5645 | 100 |
| yabsent | 5263 | 93.2 |
| yaway | 5139 | 91.0 |
| whypt | 4816 | 85.3 |
| age_6 | 4342 | 76.9 |
| agyownk | 3511 | 62.2 |
| schooln | 274 | 4.85 |

Table 1 containes categorical variables and those with 100% missing values. Consequently, we made the decision to remove variables that exhibited 100% missing values. Additionally, we opted to exclude "rec_num," "lfsstat," "prov," and "cma" from our dataset. These variables primarily serve as indicators in our filtered dataset and do not contribute meaningful information to our analysis.

## 5. Exploratory Data Analysis

### 5.1 Distribution of High and Non-High Earners

The distribution of high earners and non-high earners is indicated in Figure:1



### 5.2 Differences in Earning Group

While the average hourly wage rates in the Canadian labor market have experienced a gradual increase, it's important to note that certain groups of employees still face challenges, resulting in lower hourly wage rates. This can be observed by the relatively lower proportion of individuals categorized as "High Earners" in the subsequent sections.

5.2.1 Based on Demographic Characteristics

Figure 2 reveals that women clearly have lower proportion of High Earners than men in Toronto. Furthermore, being a younger employee (aged 15-19) or having less than a bachelor's degree significantly impacts the likelihood of falling into the "High Earner" category. These findings are concisely summarized in Figure 3 and Figure 4, respectively.

Figure 5 shows interesting results: Married individuals exhibit a considerably higher representation among high earners compared to other marital categories. Additionally, Figure 6 reveals that immigrants constitute a smaller proportion of high earners when compared to non-immigrants. This phenomenon may be attributed to a substantial portion of immigrants in our dataset who engage in part-time employment and are international students, which leads to reduced hourly earnings.

Figure:2 Percentage of High and Non−High Earners by Gender

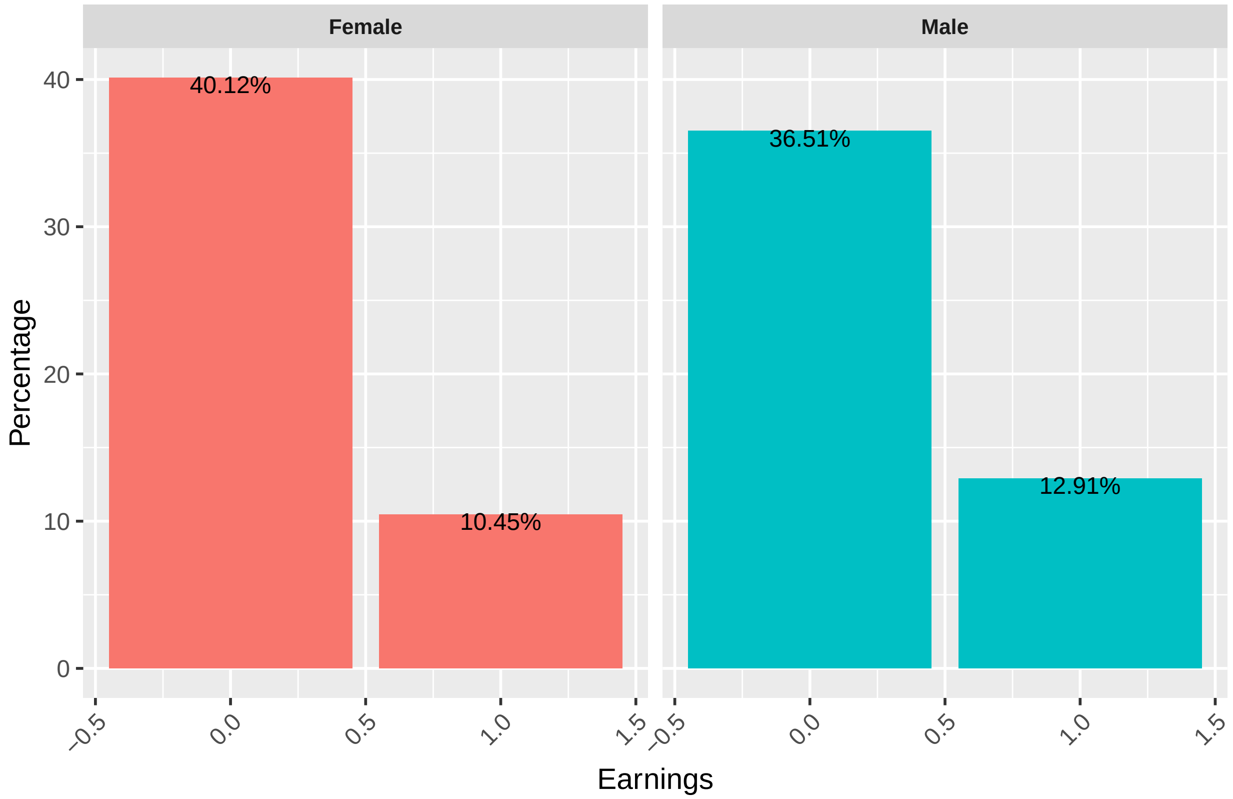


Figure:3 Percentage of High and Non−High Earners by Age Group

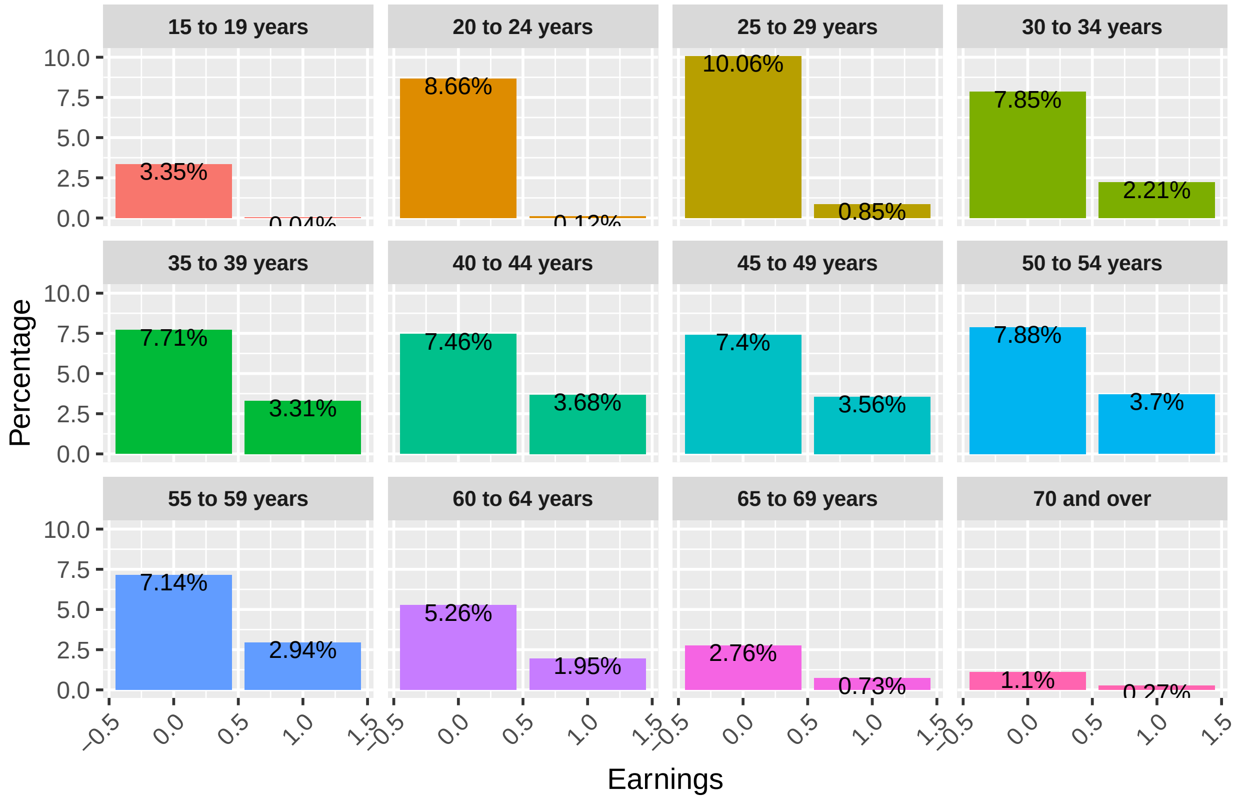


Figure:4 Percentage of High and Non−High Earners by Marital Status

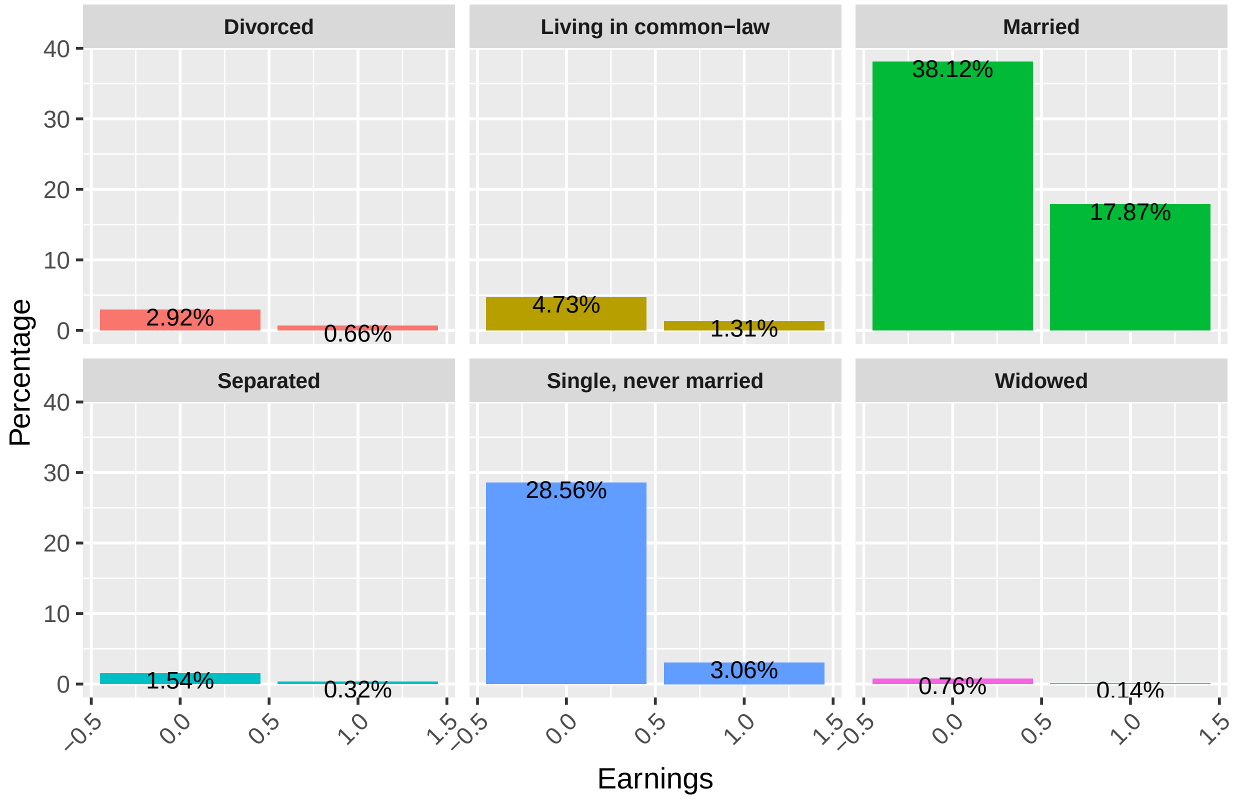
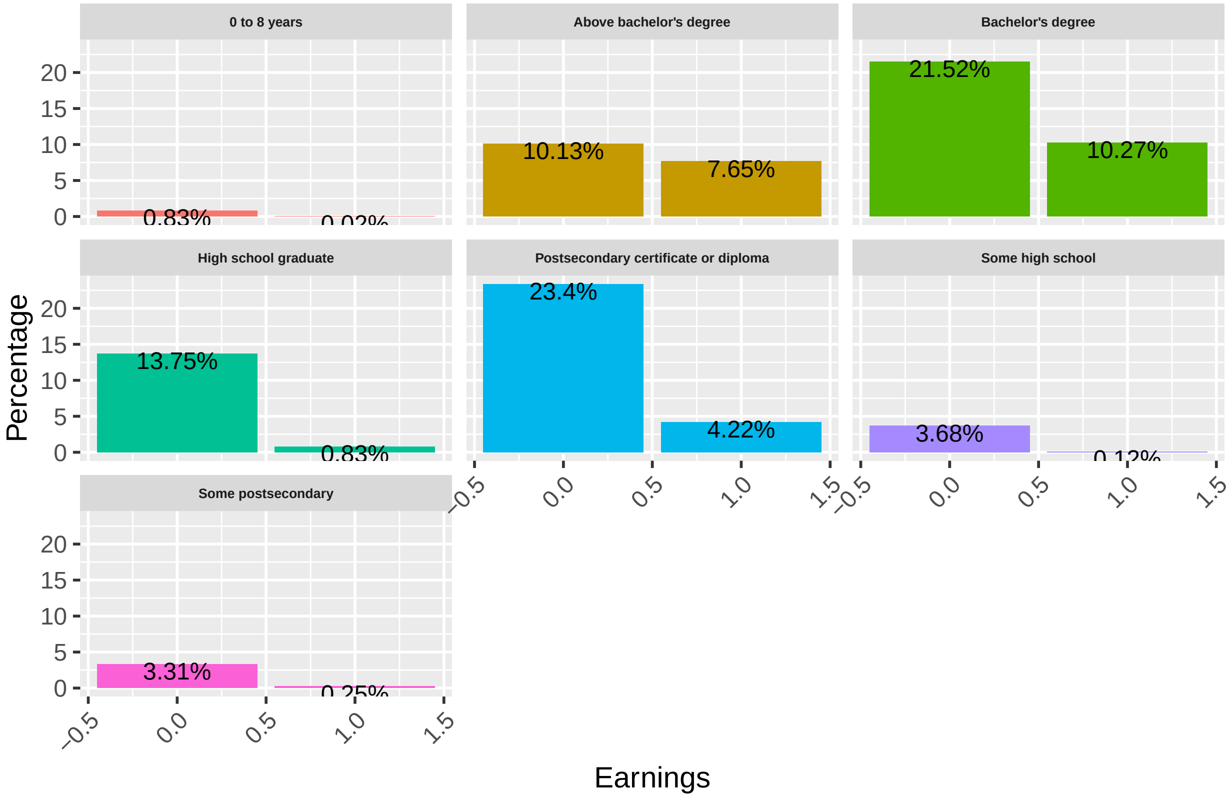


Figure:5 Percentage of High and Non−High Earners by Education



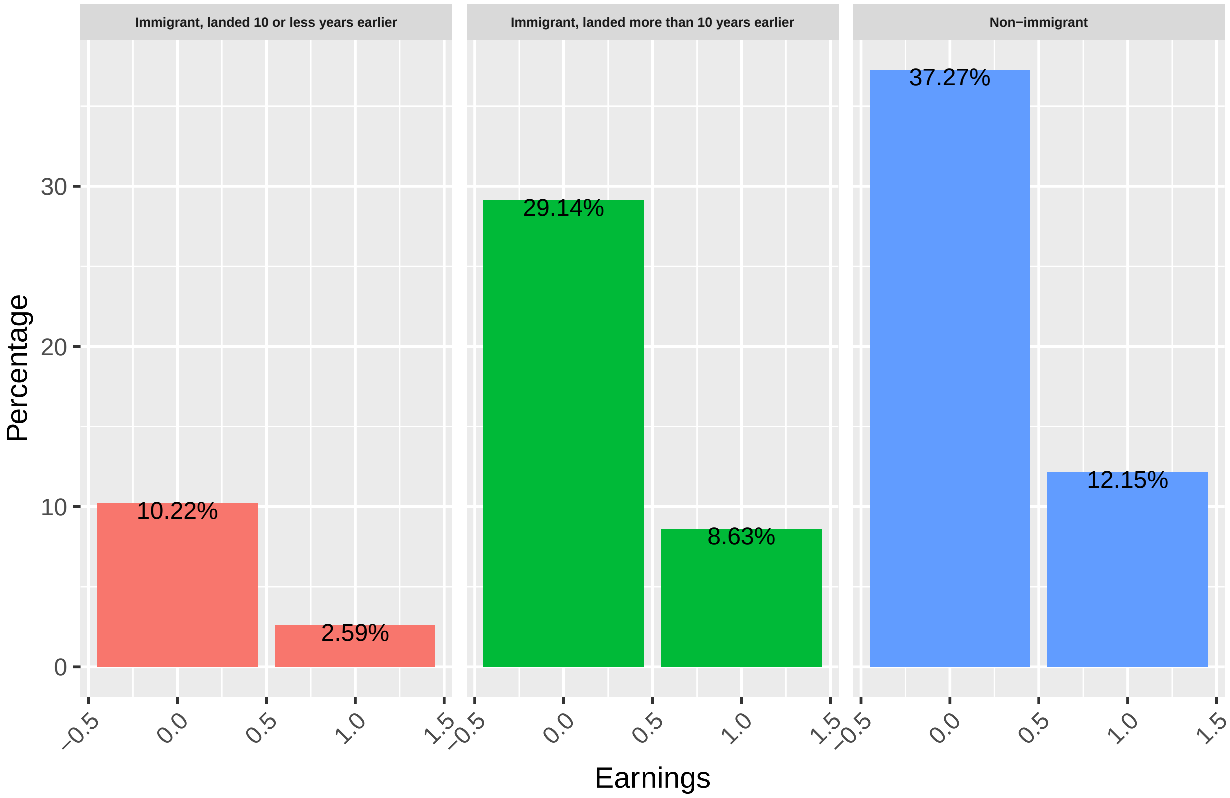


Figure:6 Percentage of High and Non−High Earners by Immigration

5.2.2 Based on Sector

Figure 7 reveals that while private sector has a significantly higher proportion of high earners than the public sector, it also has a higher proportion of “Non-High Earners” perhaps due to higher number of employment opportunities offered by this sector at lower wages.

Figure 7: Percentage of High and Non−High Earners by Sector

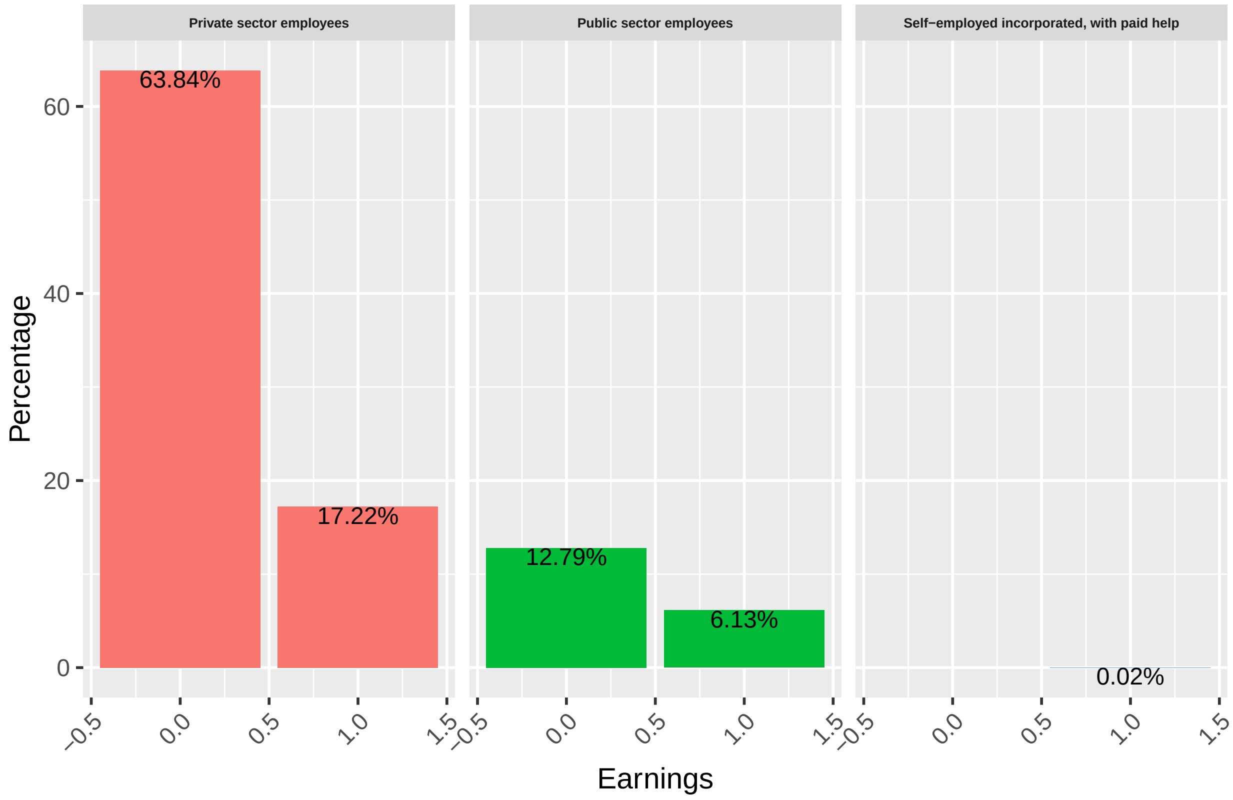
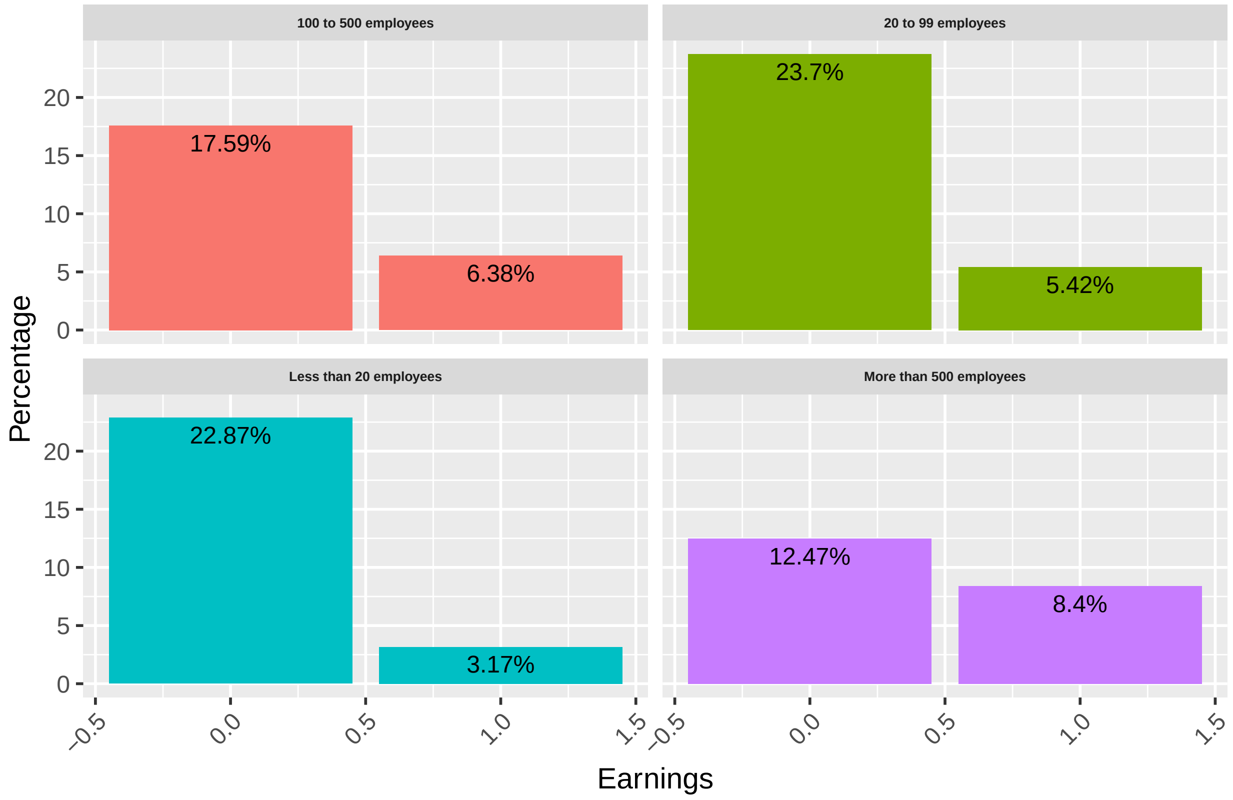


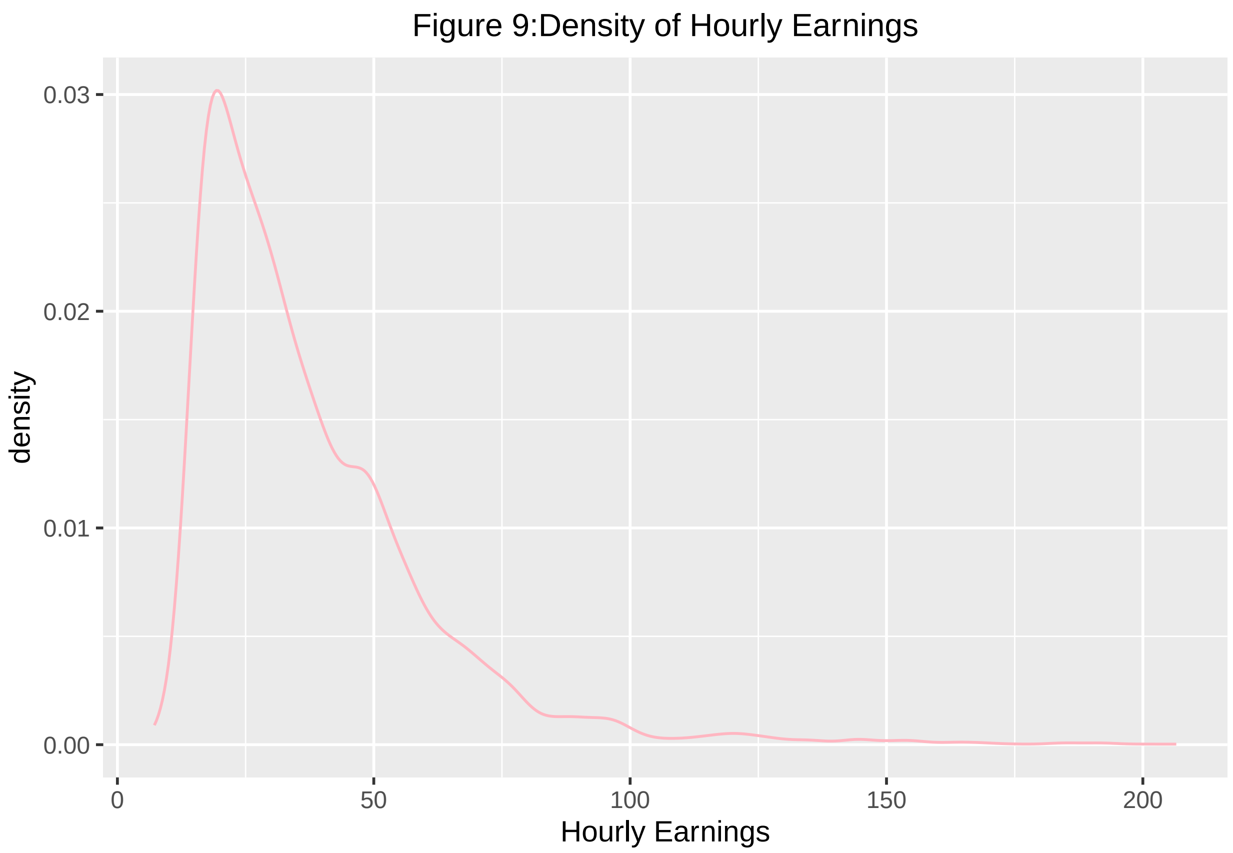
Figure 8 indicates that an establishment with huge number of employees pays its employees well as compared to smaller establishments. Though, a huge proportion of non-high earners are in establishments with 20-99 employees, this could be due to the fact that smaller companies often hire part time workers, particularly students which reduced their wage costs.

Figure 8: Percentage of High and Non−High Earners by Establishment Size



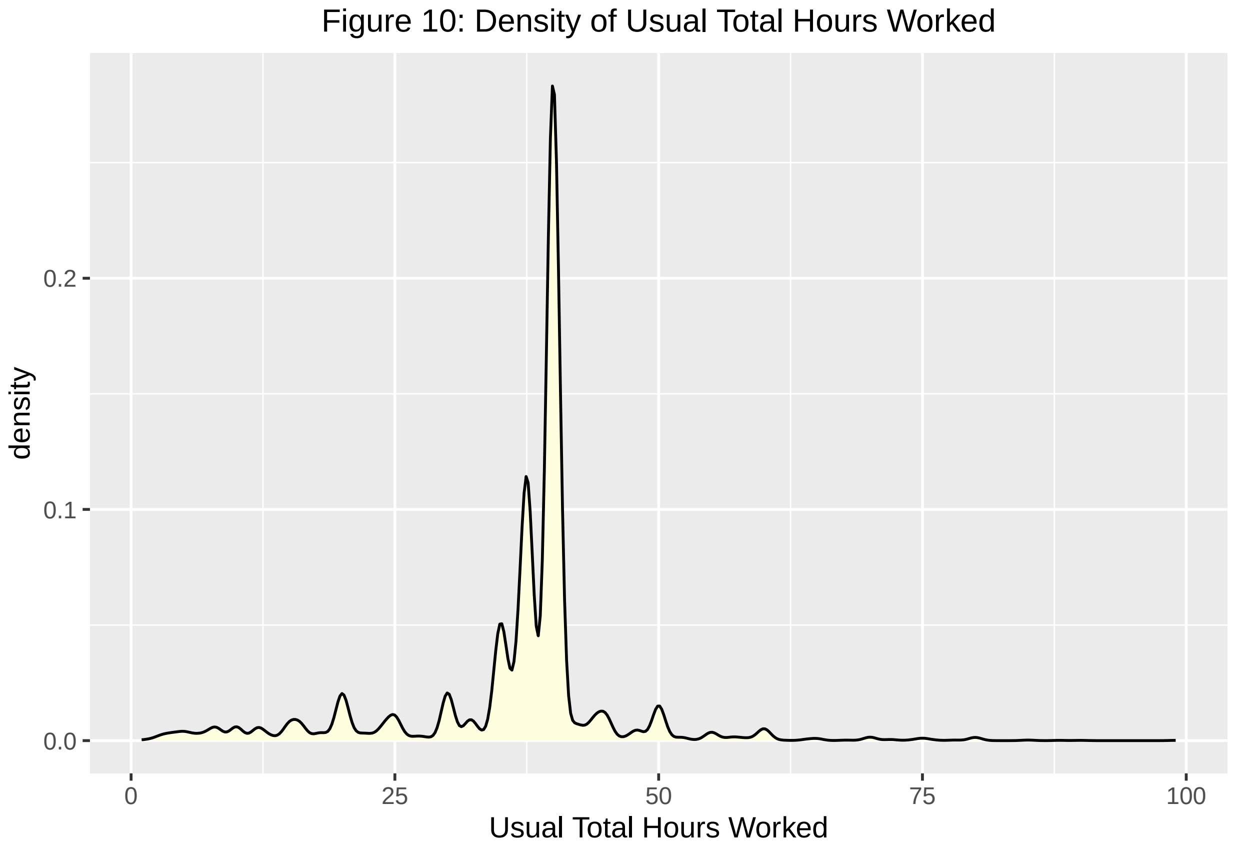
### 5.3 Distribution of Hourly Earnings

The gross hourly earnings (before taxes and other deductions) has a right skewed distribution, i.e. a higher proportion of individuals are in the lower tail, as indicated by Figure:9



### 5.3 Distribution of Total Hours Worked

Figure 10 indicates that most people in Toronto work on an average 40 hour per week, which reveals a good work life balance.



### 5.4 Exploring thought provoking hypothesis

In this section, we seek to validate intriguing findings presented in existing literature through straightforward data analysis and exploration.

#### 5.4.1 Does marriage lead to wage premium?

Inspired by the insights of McConnell, et al (2023), we aim to address this question for Toronto CMA.

Figure 11 and Figure 12 summarize the hourly earnings of females and males respectively based on their marital status. It is interesting to note that married females and males have higher median earnings as compared to their single counterparts, which could possibly be explained by the notion that married people can wait longer for better paying jobs, since they have their partner's income to support themselves during the job search period.

Figure 11: Box Plot showing wage differentials for Females Based on Marital Status

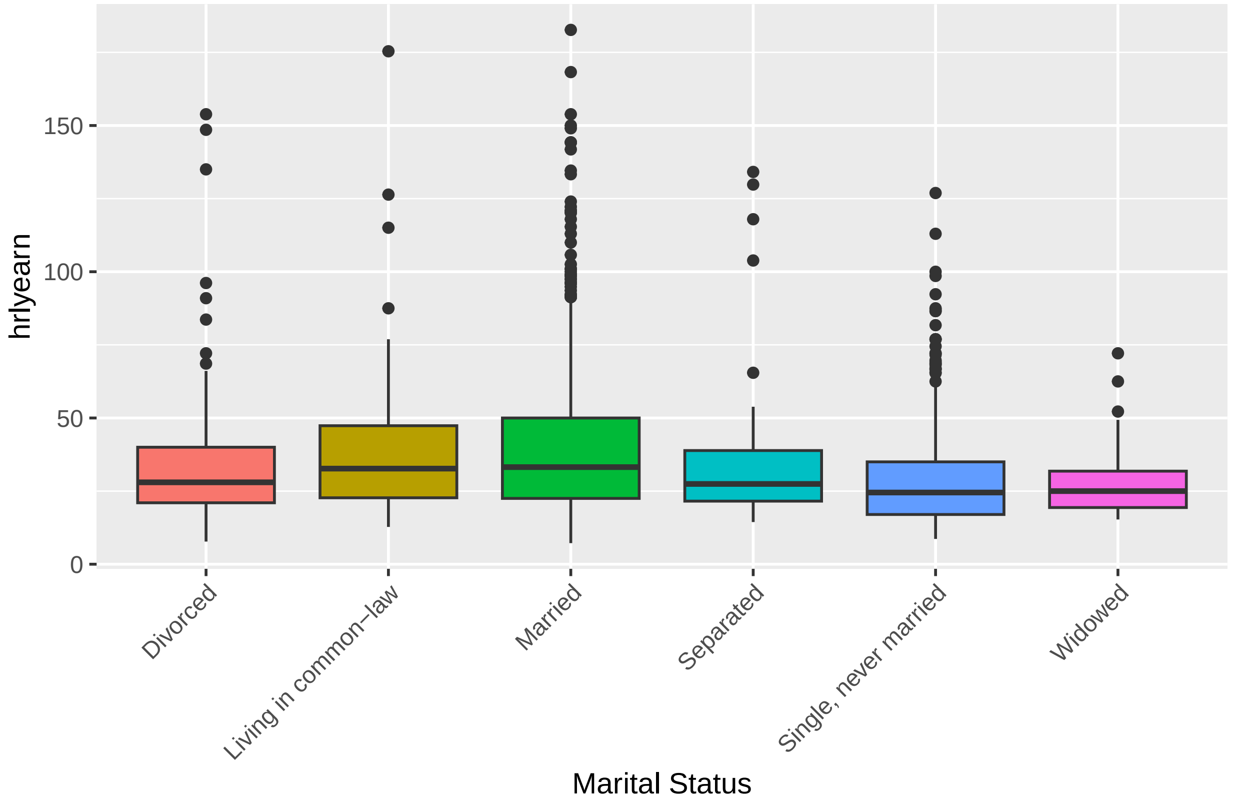
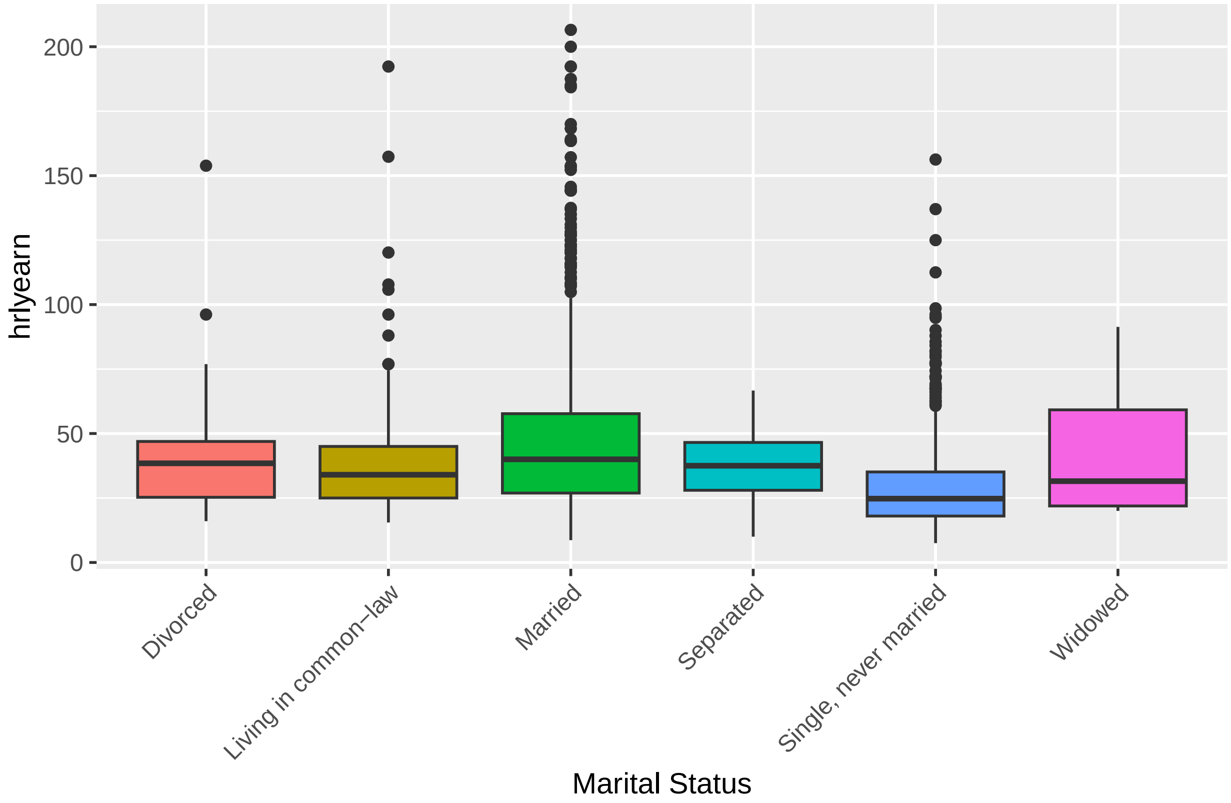


Figure 12: Box Plot showing wage differentials for Males Based on Marital Status



5.4.1 Do females earn better in the public sector?

Inspired by Mueller, R.E. (2022), we aim to address this question for Toronto CMA.

Figure 13 and Figure 14 show a higher proportion of females in the left tail of the distribution in both the private and the public sector. However, in the public sector, the proportion of females earning low wages than males is slightly lower than in the case of private sector.

Figure 13: Density Plot of Earnings of Males and Females in the Public Sector

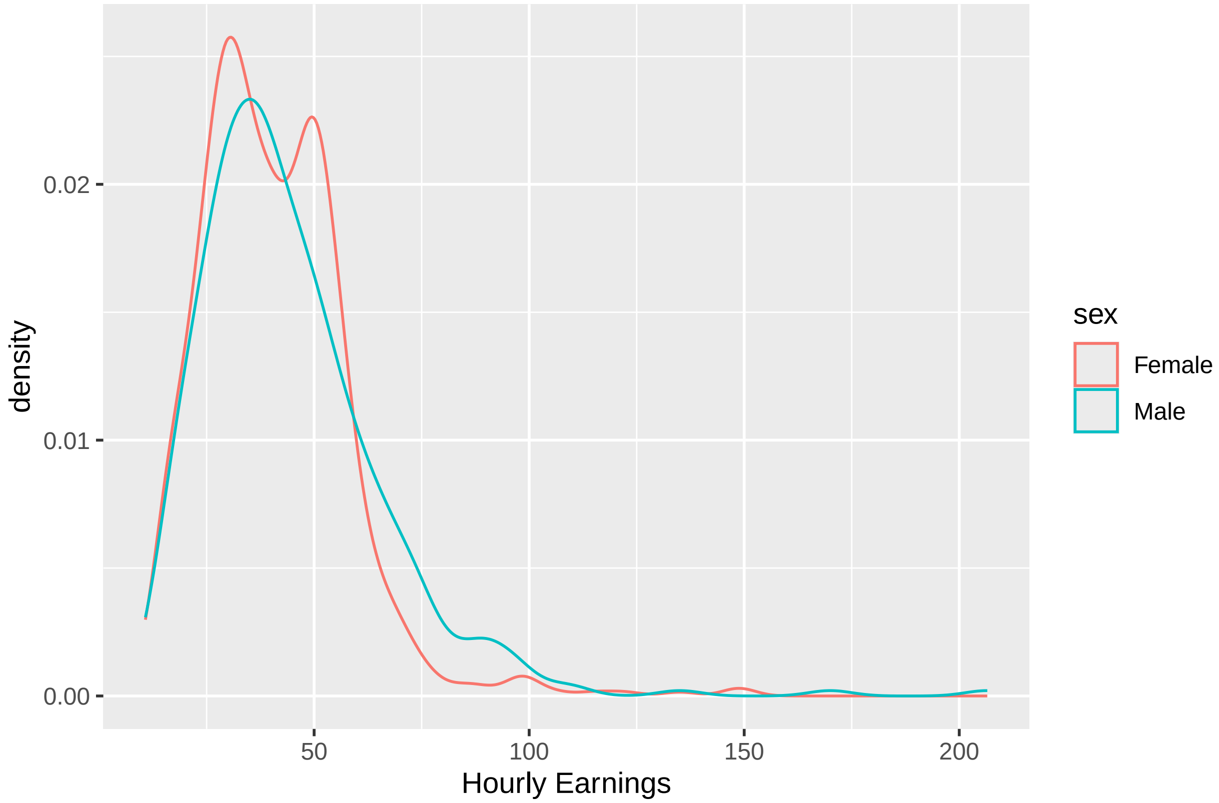
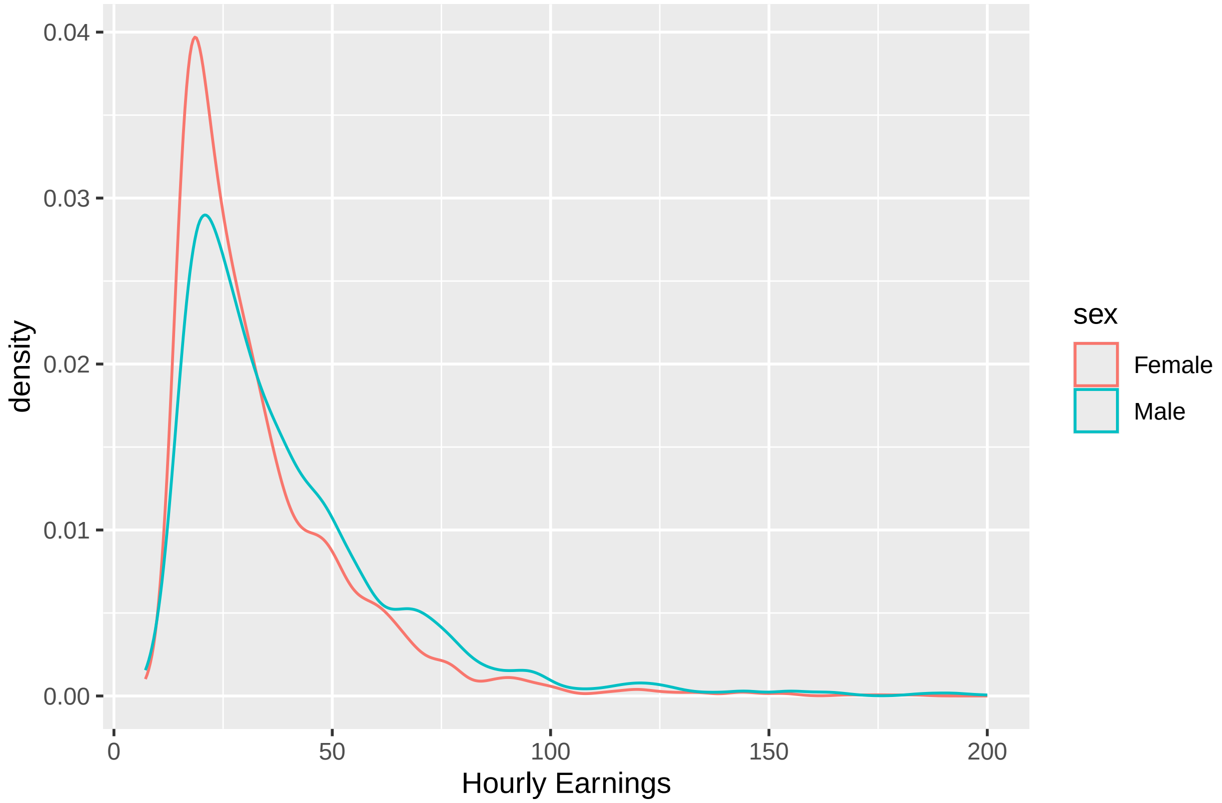


Figure 14: Density Plot of Earnings of Males and Females in the Private Sector



As observed in Figure 15, the proportion of females in the left tail of the distribution is higher in the private sector than in the public sector. The same is true for males, as indicated in Figure 16. This gives us some evidence of lower wage disparity in the public sector.

Figure 15: Density Plot of Earnings of Females in the Private and Public Sector

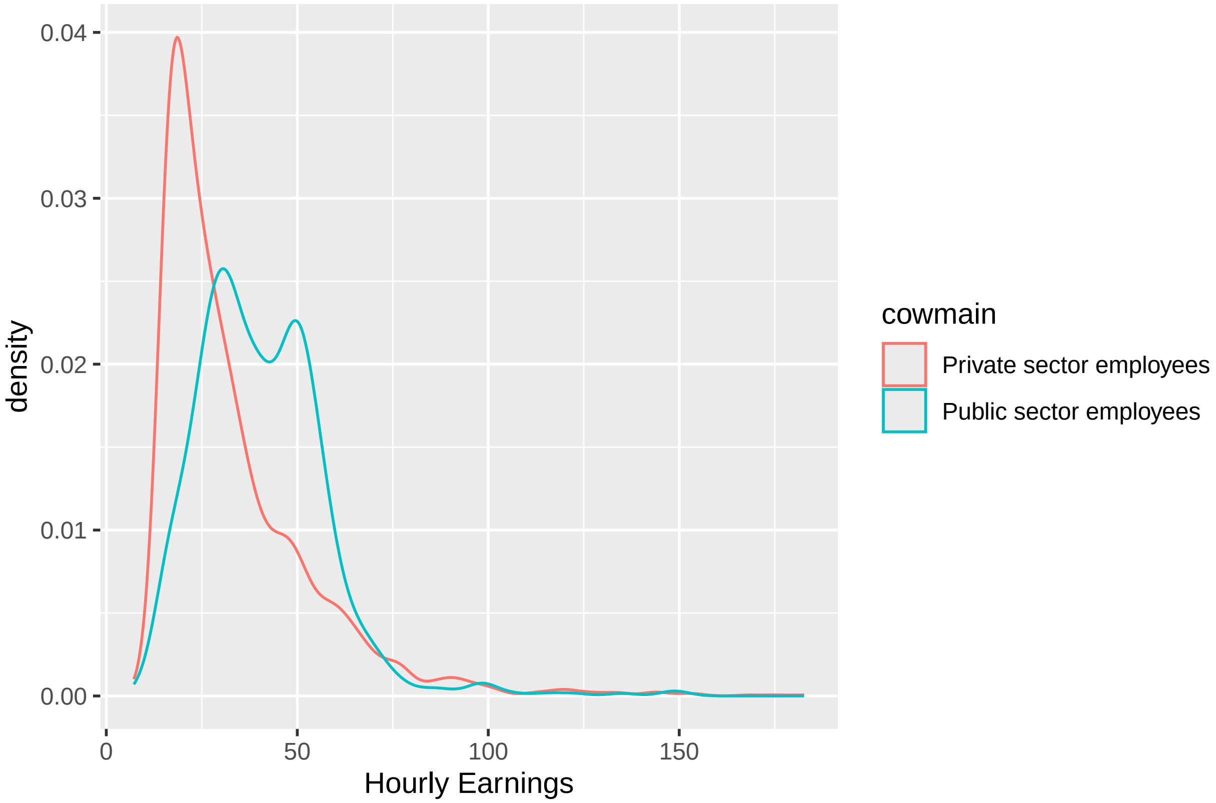
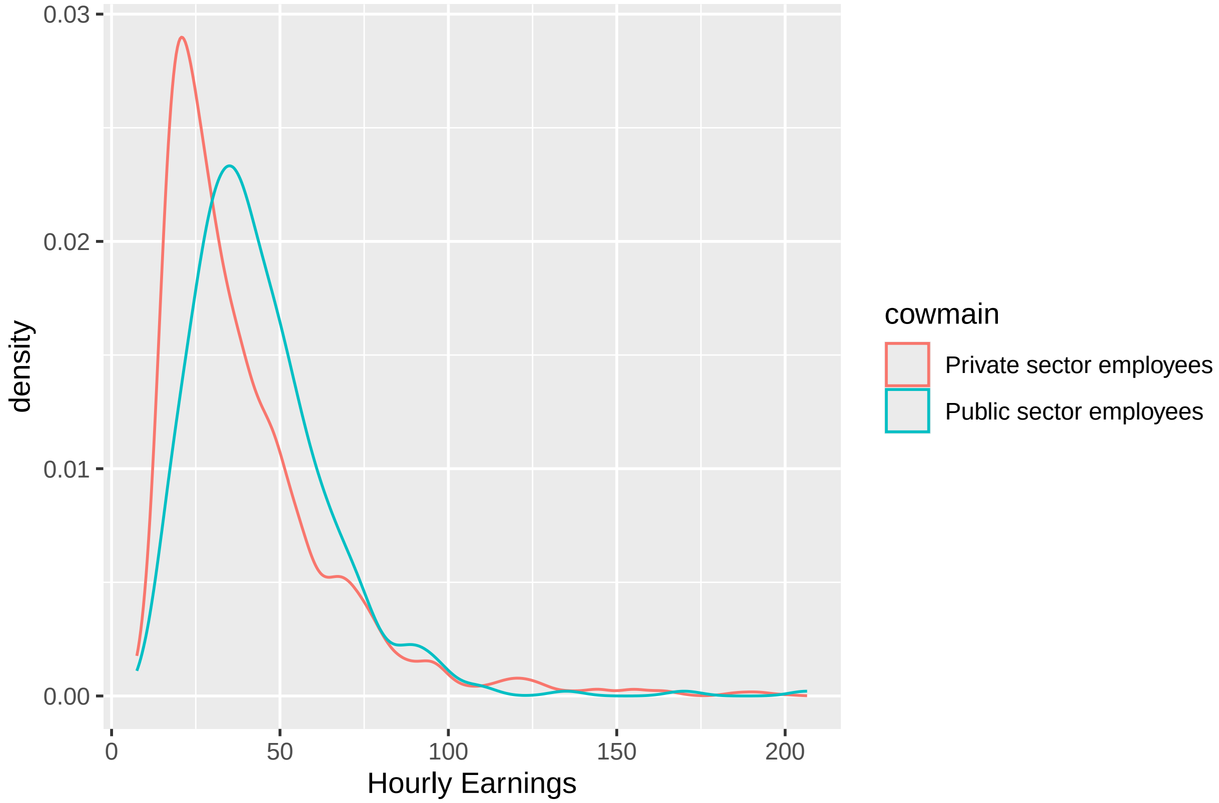


Figure 16: Density Plot of Earnings of Males in the Private and Public Sector



## 6. Variable Selection

The analysis accounts for several individual and job characteristics. Age of the individual is characterized using indicators for five-year age groups. Education is characterized to show the impact of college education on earning capabilities i.e. we divide the individuals as ”College Education” and ”No-College Education” (set equal to ”College” if the person has ”Above bachelor’s degree”,”Postsecondary certificate or diploma” or a ”Bachelor’s degree and ”No-College” otherwise ). An indicator variable for marital status is set equal to one if the individual reports being legally married or living in a commonlaw union and zero otherwise. For union status, we create an indicator for whether a person is covered by a collective agreement (set equal to one if the individual is a member of a union or covered by the collective agreement, zero otherwise). Furthermore, to study the influence of having children, we use the ”ageyonk” variable and divide the individuals as ”No-Children” if variable ageyonk is equal to Not-Applicable and ”Has-Children” otherwise. we also study the impact of overtime eligibility on earnings capability by distinguishing the individuals as overtime-eligible if they get paidot and not-eligible otherwise. A variable for tenure represents the number of months a person has held his or her current job and the variable utothrs indicates the usual total hours that a person works at their main job. Finally, we separate the individuals based on the industry they are working in i.e. we call them working in a ”Top-5” industry if they work in ”Real estate and rental and leasing”,”Finance and insurance”, ”Manufacturing - non-durable goods”,”Manufacturing -durable goods” and ”Health care and social assistance” and ”Not Top-5” otherwise.

## 7. Fitting the Logistic Regression Model

Since the outcome of interest is a binary variable, a binary logistic regression model with a logit link was fit. Variable Selection was done as indicated. Next, we address the issue with having varying ranges amongst our numerical variables by standardizing them so that they have a mean of 0 and standard deviation of 1.This is done to ensure that the scales of the variables does not impact the contribution of the variable to the outcome of interest. we then proceed to build three models:

- Model 1 includes all the covariates
- Model 2 includes all variables from Model 1 in addition to interaction effects between marital status and sex as well as sector where an individual is working and sex. These interaction effects are based on review of literature
- In Model 3, we drop the insignificant variables obtained in Model 2 to create Model

For the three explanatory models, we use the entire dataset, since reducing the sample induces bias but for the predictive models, we divide the dataset into train (80%) and test (20%) as the goal of predictive modelling is to sacrifice bias in order to reduce sampling variance, that is key to improving generalization (model performance on test dataset). The predictive models are furthermore generated via 10 fold repeated cross validation of train data to ensure the robustness and reliability of the model's performance estimates. By repeatedly dividing the data into 10 different subsets and using each subset in turn for validation while training on the remaining data, we were able to mitigate the risk of overfitting and gain a more accurate understanding of how the model is likely to perform on unseen data. Finally,to generate predictions, we use the threshold of 0.50 to get binary outcomes.

Fitting Explanatory Models

Model-1

```
fit_ogl<-glm(Earnings_Dummy~., data = combined_final, family = binomial(link = "logit"))
```

Model-2

```
fit<-glm(Earnings_Dummy ~ `Usual hours` + Tenure + `25 to 29 years` +
    `30 to 34 years` + `35 to 39 years` + `40 to 44 years` +
    `45 to 49 years` + `50 to 54 years` + `55 to 59 years` +
    `60 to 64 years` + `65 to 69 years` + `70 and over` +
    sex_Male + `No College` + `Single jobholder, including job changers` +
    `Public sector employees` + `Self-employed incorporated, ` +
    `Immigrant, landed more than 10 years earlier` + `Non-immigrant` +
    `Unionized Job` + `Temporary, casual or other temporary jobs` +
    `Temporary, seasonal job` + `Temporary, term or contract job` +
    `estsize_100 to 500 employees` + `estsize_Less than 20 employees` +
    `estsize_More than 500 employees` + `No Children` +
    `Overtime Not Eligible` + Industry_Performers + Marital_Status_Single+
    Marital_Status_Single*sex_Male+sex_Male*`Public sector employees`
    ,data = combined_final,family =binomial(link = "logit"))
```

```
fit_reduced<-glm(Earnings_Dummy~`Usual hours` + Tenure + `25 to 29 years` +
 `30 to 34 years` + `35 to 39 years` + `40 to 44 years` + `45 to 49 years` +
   `50 to 54 years` + `55 to 59 years` + `60 to 64 years` + `65 to 69 years` +
   `70 and over` + sex_Male + `No College` + `Public sector employees` +
   `Non-immigrant` + `Unionized Job` + `estsize_100 to 500 employees`
 + `estsize_Less than 20 employees` +  `estsize_More than 500 employees`
 + `No Children` + Marital_Status_Single+
   sex_Male*`Public sector employees`
 ,data =  combined_final,family =binomial(link = "logit"))
```

Fitting Predictive Models

Model-1

```
cv <- trainControl(method = "repeatedcv", number = 10,repeats = 10)
model_cv_ogl<-train(as.factor(Earnings_Dummy)~., data=train_data, method = "glm",
family="binomial",trControl = cv)
```

Model-2

```
cv <- trainControl(method = "repeatedcv", number = 10,repeats = 10)
model_cv_1 <- train( as.factor(Earnings_Dummy)~utothrs + tenure +
      `age_12_25 to 29 years` + `age_12_30 to 34 years` + `age_12_35 to 39 years` +
    `age_12_40 to 44 years` + `age_12_45 to 49 years` +
    `age_12_50 to 54 years` + `age_12_55 to 59 years` +
    `age_12_60 to 64 years` + `age_12_65 to 69 years` +
    `age_12_70 and over` + sex_Male + `College_No College Education` +
    `mjh_Single jobholder, including job changers` + `cowmain_Public sector employees` +
    `cowmain_Self-employed incorporated, with paid help` +
    `immig_Immigrant, landed more than 10 years earlier` +
    `immig_Non-immigrant` + `Union_Status_Unionized Job` +
    `permtemp_Temporary, casual or other temporary jobs` +
    `permtemp_Temporary, seasonal job` + `permtemp_Temporary, term or contract job` +
    `estsize_100 to 500 employees` + `estsize_Less than 20 employees` +
    `estsize_More than 500 employees` + `Children_No Children` +
    `Overtime Eligible_Not Eligible` + Industry_Performers +
    Marital_Status_Single+ Marital_Status_Single*sex_Male+
      sex_Male*`cowmain_Public sector employees`,
    data = train_data, method = "glm", family="binomial",trControl = cv)
```

Model-3

```
cv <- trainControl(method = "repeatedcv", number = 10, repeats = 10)
model_cv_2 <- train(as.factor(Earnings_Dummy)~utothrs + tenure +
    `age_12_25 to 29 years` +
    `age_12_30 to 34 years` + `age_12_35 to 39 years` +
    `age_12_40 to 44 years` + `age_12_45 to 49 years` +
    `age_12_50 to 54 years` + `age_12_55 to 59 years` +
    `age_12_60 to 64 years` + `age_12_65 to 69 years` +
    `age_12_70 and over` + sex_Male + `College_No College Education` +
     `cowmain_Public sector employees` +
    `immig_Non-immigrant` + `Union_Status_Unionized Job` +
    `estsize_100 to 500 employees` + `estsize_Less than 20 employees` +
    `estsize_More than 500 employees` + `Children_No Children` +
    Marital_Status_Single+sex_Male*`cowmain_Public sector employees`,
    data = train_data, method = "glm", family="binomial",trControl = cv)
```

## 8. Model Performance

Model Validation in explanatory modelling involves goodness of fit-tests while in predictive modelling, the focus is on comparing a model's generalization abilities. We therefore use the Likelihood-Ratio Test (LRT) to evaluate model fit and the model test accuracy and the Expected Prediction Error (EPE) to compare models' predictive abilities.

The LRT method, in simple terms compares the log likelihood ratio statistic from the two nested models to a Chi-squared distribution. This determines whether it is beneficial to add more parameters or if the simpler model is preferred.

For predictive power, using 10 fold repeated cross validation, we train Model 1, Model 2 and Model 3 on the train data and evaluate their predictive accuracy on the test data using EPE. The EPE is the average of the difference between predicted values and actual values.

Model Validation for Explanatory Models

```
## Likelihood ratio test
##
## Model 1: Earnings_Dummy ~ ‘Usual hours‘ + Tenure + ‘25 to 29 years‘ +
##     ‘30 to 34 years‘ + ‘35 to 39 years‘ + ‘40 to 44 years‘ +
##     ‘45 to 49 years‘ + ‘50 to 54 years‘ + ‘55 to 59 years‘ +
##     ‘60 to 64 years‘ + ‘65 to 69 years‘ + ‘70 and over‘ + sex_Male +
##     ‘No College‘ + ‘Single jobholder, including job changers‘ +
##     ‘Public sector employees‘ + ‘Self-employed incorporated, ‘ +
##     ‘Immigrant, landed more than 10 years earlier‘ + ‘Non-immigrant‘ +
##     ‘Unionized Job‘ + ‘Temporary, casual or other temporary jobs‘ +
##     ‘Temporary, seasonal job‘ + ‘Temporary, term or contract job‘ +
##     ‘estsize_100 to 500 employees‘ + ‘estsize_Less than 20 employees‘ +
##     ‘estsize_More than 500 employees‘ + ‘No Children‘ + ‘Overtime Not Eligible‘ +
##     Industry_Performers + Marital_Status_Single
## Model 2: Earnings_Dummy ~ ‘Usual hours‘ + Tenure + ‘25 to 29 years‘ +
##     ‘30 to 34 years‘ + ‘35 to 39 years‘ + ‘40 to 44 years‘ +
##     ‘45 to 49 years‘ + ‘50 to 54 years‘ + ‘55 to 59 years‘ +
##     ‘60 to 64 years‘ + ‘65 to 69 years‘ + ‘70 and over‘ + sex_Male +
##     ‘No College‘ + ‘Public sector employees‘ + ‘Non-immigrant‘ +
##     ‘Unionized Job‘ + ‘estsize_100 to 500 employees‘ + ‘estsize_Less than 20 employees‘ +
##     ‘estsize_More than 500 employees‘ + ‘No Children‘ + Marital_Status_Single +
##     sex_Male * ‘Public sector employees‘
## Model 3: Earnings_Dummy ~ ‘Usual hours‘ + Tenure + ‘25 to 29 years‘ +
##     ‘30 to 34 years‘ + ‘35 to 39 years‘ + ‘40 to 44 years‘ +
##     ‘45 to 49 years‘ + ‘50 to 54 years‘ + ‘55 to 59 years‘ +
##     ‘60 to 64 years‘ + ‘65 to 69 years‘ + ‘70 and over‘ + sex_Male +
##     ‘No College‘ + ‘Single jobholder, including job changers‘ +
##     ‘Public sector employees‘ + ‘Self-employed incorporated, ‘ +
##     ‘Immigrant, landed more than 10 years earlier‘ + ‘Non-immigrant‘ +
##     ‘Unionized Job‘ + ‘Temporary, casual or other temporary jobs‘ +
##     ‘Temporary, seasonal job‘ + ‘Temporary, term or contract job‘ +
##     ‘estsize_100 to 500 employees‘ + ‘estsize_Less than 20 employees‘ +
##     ‘estsize_More than 500 employees‘ + ‘No Children‘ + ‘Overtime Not Eligible‘ +
##     Industry_Performers + Marital_Status_Single + Marital_Status_Single *
##     sex_Male + sex_Male * ‘Public sector employees‘
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1  31 -2237.9
## 2  24 -2262.1 -7 48.221  3.224e-08 ***
## 3  33 -2233.9  9 56.336  6.769e-09 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Model with interaction terms and with interaction terms+ insignificant variables shows a better fit.

Model Validation for Predictive Models

Results for 10 fold repeated cross validation

```
predictions_model_cv_ogl<- predict(model_cv_ogl, newdata = test_data,
                                   type = "prob")

predictions_model_cv_1<- predict(model_cv_1, newdata = test_data,
                                 type = "prob")

predictions_model_cv_2<- predict(model_cv_2, newdata = test_data,
                                 type = "prob")
```

```
test_probs_cv_ogl <- predictions_model_cv_ogl[, "1"]
test_probs_cv_1 <- predictions_model_cv_1[, "1"]
test_probs_cv_2 <- predictions_model_cv_2[, "1"]
```

Calculation of Brier Score

```
mean((test_probs_cv_ogl-test_data$Earnings_Dummy)^2)
```

```
## [1] 0.1577106
```

```
mean((test_probs_cv_1-test_data$Earnings_Dummy)^2)
```

```
## [1] 0.1578613
```

```
mean((test_probs_cv_2-test_data$Earnings_Dummy)^2)
```

```
## [1] 0.1601731
```

When evaluating the models via LR test, we find Model 2 and Model 3 have significant Chisq values at 1% significance level, that indicates Model 2 and Model 3 are a better fit when compared to Model 1 and Model 3 performs even better than Model 2 in terms of model fit and in addition, it contains all the significant covariates.

On the contrary, Model 1 has the lowest Brier Score among all the three models. Hence, it would be reasonable to argue that while focusing on explanatory power, since we need to maximize model fit,we should go with Model 3 but when the target is to maximise predictive accuracy, Model 1 would be an appropriate choice.

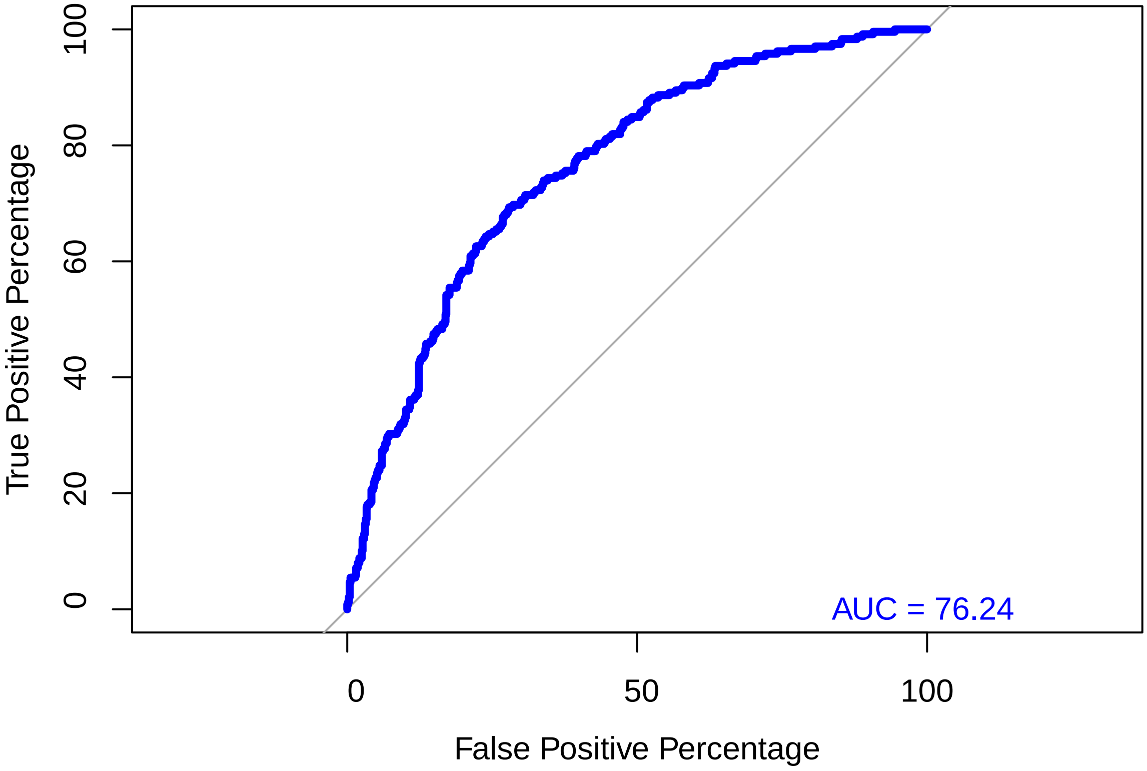
### 9. AUC/ROC Curves

```
roc_object<-roc(test_data$Earnings_Dummy, test_probs_cv_ogl,percent=T)
```

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

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

```r
plot(roc_object,legacy.axes=T,percent=T,xlab="False Positive Percentage",
     ylab="True Positive Percentage",lwd=4,col="blue")
text(0.7, 0.2, paste("AUC =", round(auc(roc_object), 2)), col = "blue")
```

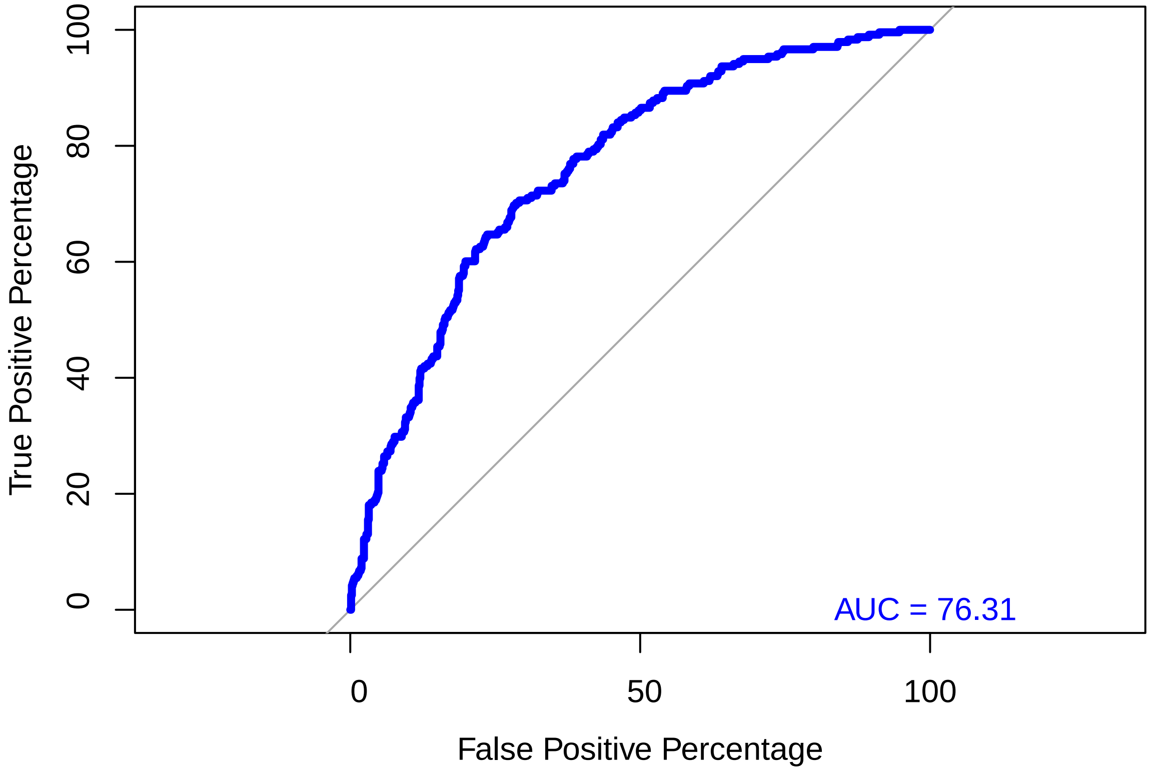


```r
roc_object<-roc(test_data$Earnings_Dummy, test_probs_cv_1,percent=T)
```

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

## Setting direction: controls < cases
```

```r
plot(roc_object,legacy.axes=T,percent=T,xlab="False Positive Percentage",
     ylab="True Positive Percentage",lwd=4,col="blue")
text(0.8, 0.2, paste("AUC =", round(auc(roc_object), 2)), col = "blue")
```

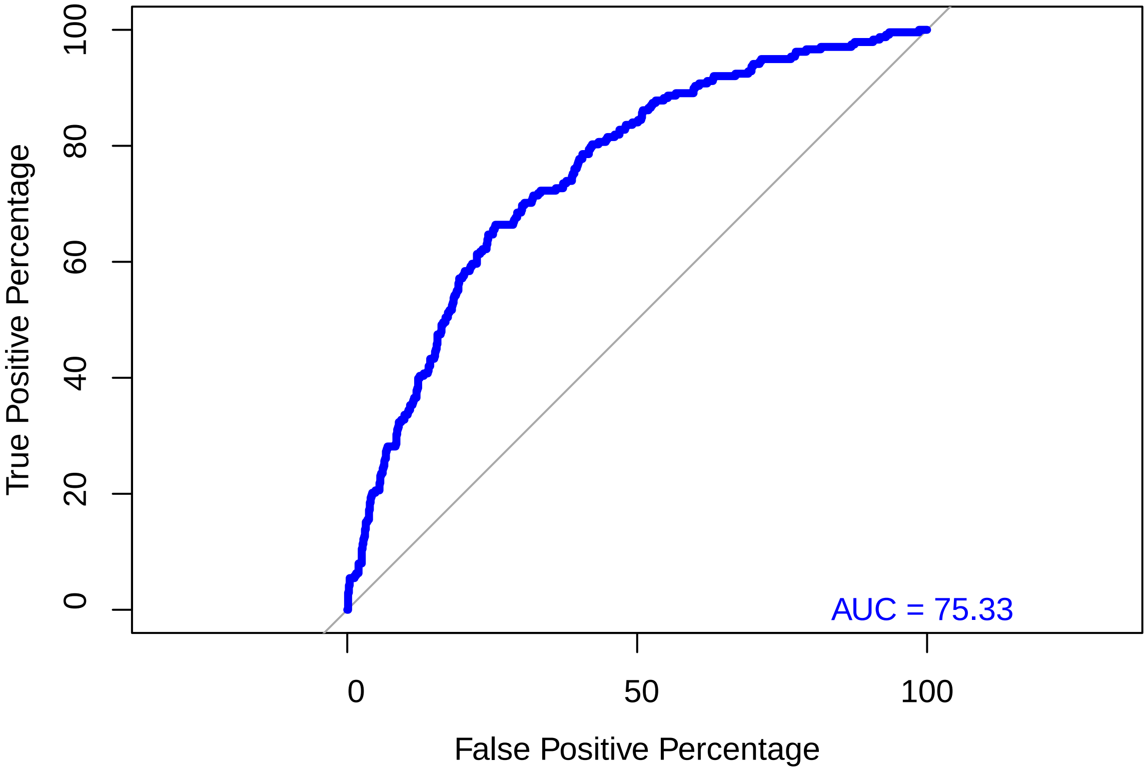


```
roc_object<-roc(test_data$Earnings_Dummy, test_probs_cv_2,percent=T)
```

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

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

```
plot(roc_object,legacy.axes=T,percent=T,xlab="False Positive Percentage",
     ylab="True Positive Percentage",lwd=4,col="blue")
text(0.8, 0.2, paste("AUC =", round(auc(roc_object), 2)), col = "blue")
```



For the test data, Area under the curve is the highest for Model-1 when compared to Model-2 and Model-3.

### 10. Confusion Matrix

```
test_probs_cv_ogl_class<-ifelse(test_probs_cv_ogl>0.50,1,0)
test_probs_cv_1_class<-ifelse(test_probs_cv_1>0.50,1,0)
test_probs_cv_2_class<-ifelse(test_probs_cv_2>0.50,1,0)
```

```
CM1<-confusionMatrix(as.factor(test_data$Earnings_Dummy),
                     as.factor(test_probs_cv_ogl_class))

CM1
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 658  62
##          1 165  73
##
##                Accuracy : 0.763
##                  95% CI : (0.7348, 0.7897)
##     No Information Rate : 0.8591
##     P-Value [Acc > NIR] : 1
```

```
##
##                   Kappa : 0.258
##
##  Mcnemar’s Test P-Value : 1.288e-11
##
##             Sensitivity : 0.7995
##             Specificity : 0.5407
##          Pos Pred Value : 0.9139
##          Neg Pred Value : 0.3067
##              Prevalence : 0.8591
##          Detection Rate : 0.6868
##    Detection Prevalence : 0.7516
##       Balanced Accuracy : 0.6701
##
##        ’Positive’ Class : 0
##
```

```
CM2<-confusionMatrix(as.factor(test_data$Earnings_Dummy),
                     as.factor(test_probs_cv_1_class))
CM2
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 656  64
##          1 165  73
##
##                Accuracy : 0.761
##                  95% CI : (0.7327, 0.7877)
##     No Information Rate : 0.857
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.2539
##
##  Mcnemar’s Test P-Value : 3.891e-11
##
##             Sensitivity : 0.7990
##             Specificity : 0.5328
##          Pos Pred Value : 0.9111
##          Neg Pred Value : 0.3067
##              Prevalence : 0.8570
##          Detection Rate : 0.6848
##    Detection Prevalence : 0.7516
##       Balanced Accuracy : 0.6659
##
##        ’Positive’ Class : 0
##
```

```
CM3<-confusionMatrix(as.factor(test_data$Earnings_Dummy),
                     as.factor(test_probs_cv_2_class))

CM3
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 659  61
##          1 168  70
##
##                Accuracy : 0.761
##                  95% CI : (0.7327, 0.7877)
##     No Information Rate : 0.8633
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.2465
##
##  Mcnemar’s Test P-Value : 2.476e-12
##
##             Sensitivity : 0.7969
##             Specificity : 0.5344
##          Pos Pred Value : 0.9153
##          Neg Pred Value : 0.2941
##              Prevalence : 0.8633
##          Detection Rate : 0.6879
##    Detection Prevalence : 0.7516
##       Balanced Accuracy : 0.6656
##
##        ’Positive’ Class : 0
##
```

Based on the confusion matrices for the three models-

-Sensitivity Model-2>Model-3>Model-1 -Specificity Model-2> Model-3> Model-1

Clearly, Model-2 Outperforms the other two models when it comes to sensitivity (True Positive Rate) and specificity(True Negative Rates).

Thus, while Model-1 gives us highly accurate predictions (lowest Brier Score), it however fails to capture True Positive and True Negatives in the model. This could be due the fact that our dataset is severely imbalanced and has a high proportion of low earners as compared to high earners.