Data Analysis 3, Assignment 1

Tunay Tokmak

2023-01-25

RETAIL SALESPERSONS

The selected occupation is retail salespersons.

```
retail_salespersons<- original_df %>% filter(occ2012 == 4760) %>% select(-occ2012)
```

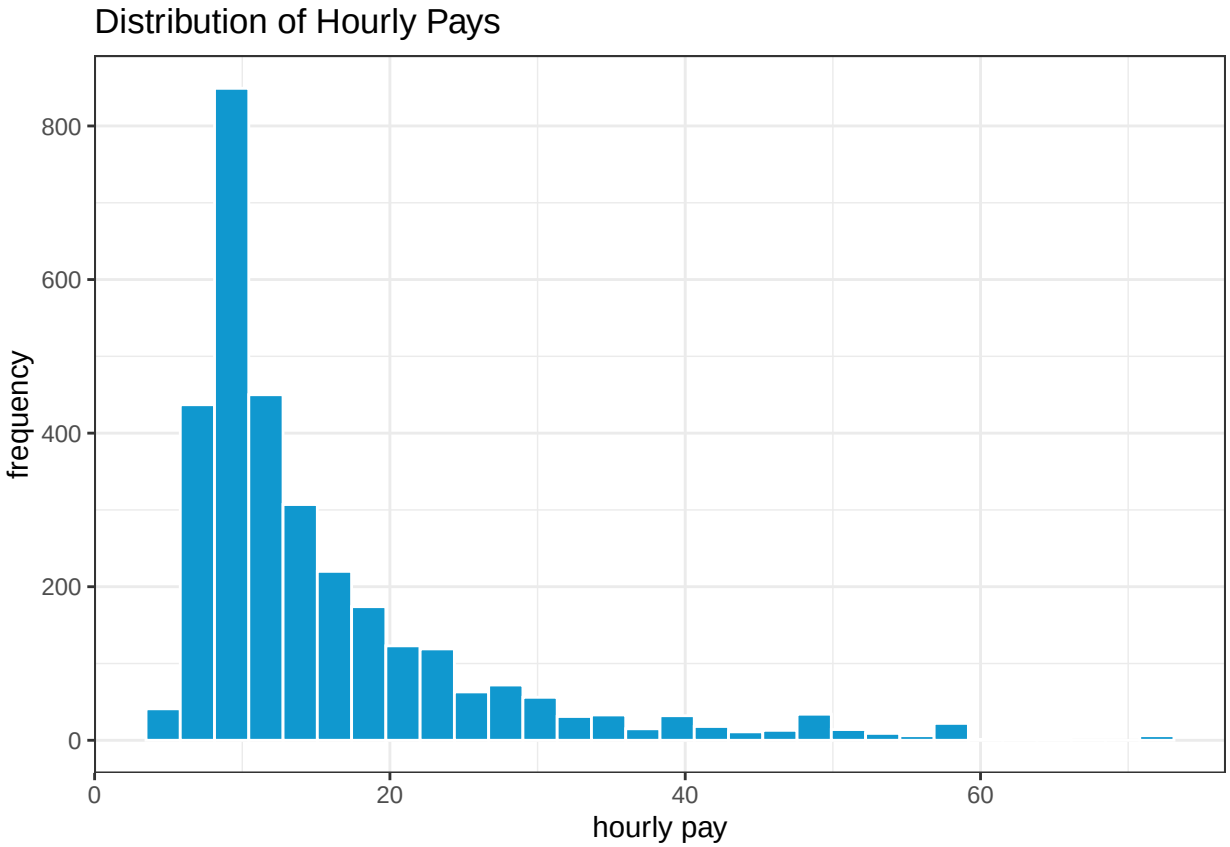
OUTLIER DETECTION

The maximum value is two times higher than the 99th percentile and the minimum value corresponds to 7 cents which is likely an error.

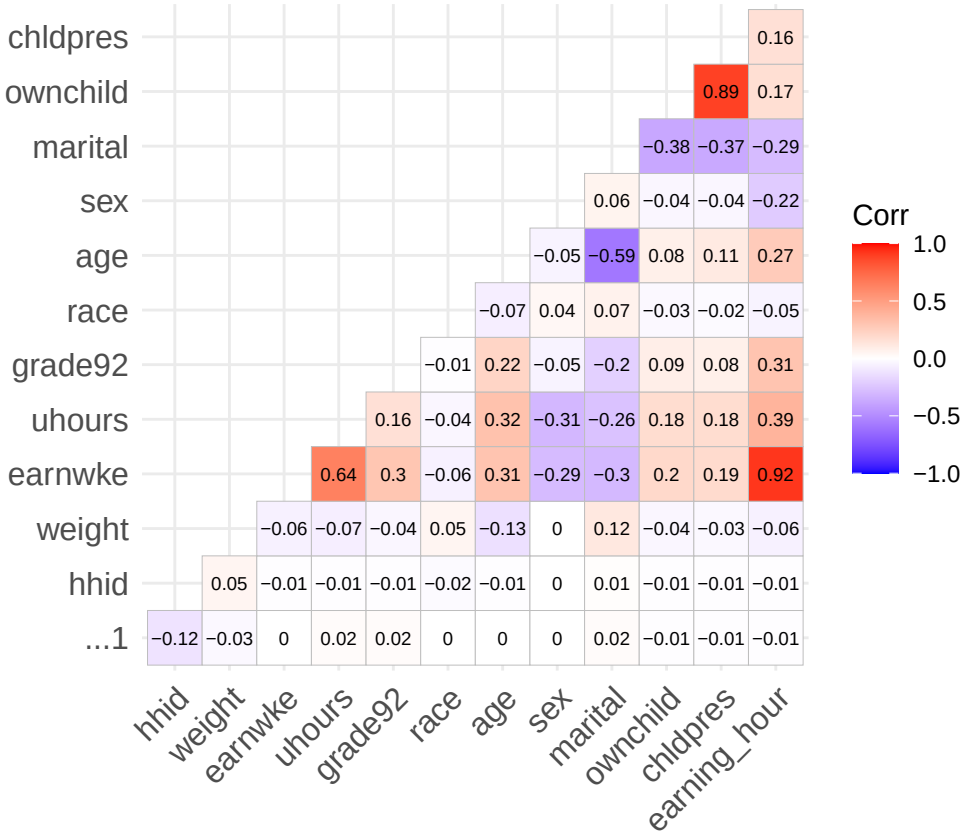
Therefore the lower bound and the upper bound are set as 0.01th and 0.99th percentiles, and the values below or above are considered as outliers.

Let's see how the hourly pay is distributed

	Mean	SD	P0	P01	P25	P50	P75	P99	P100
earning_hour	16.13	12.20	0.07	4.70	9.00	12.00	18.49	72.10	144.23

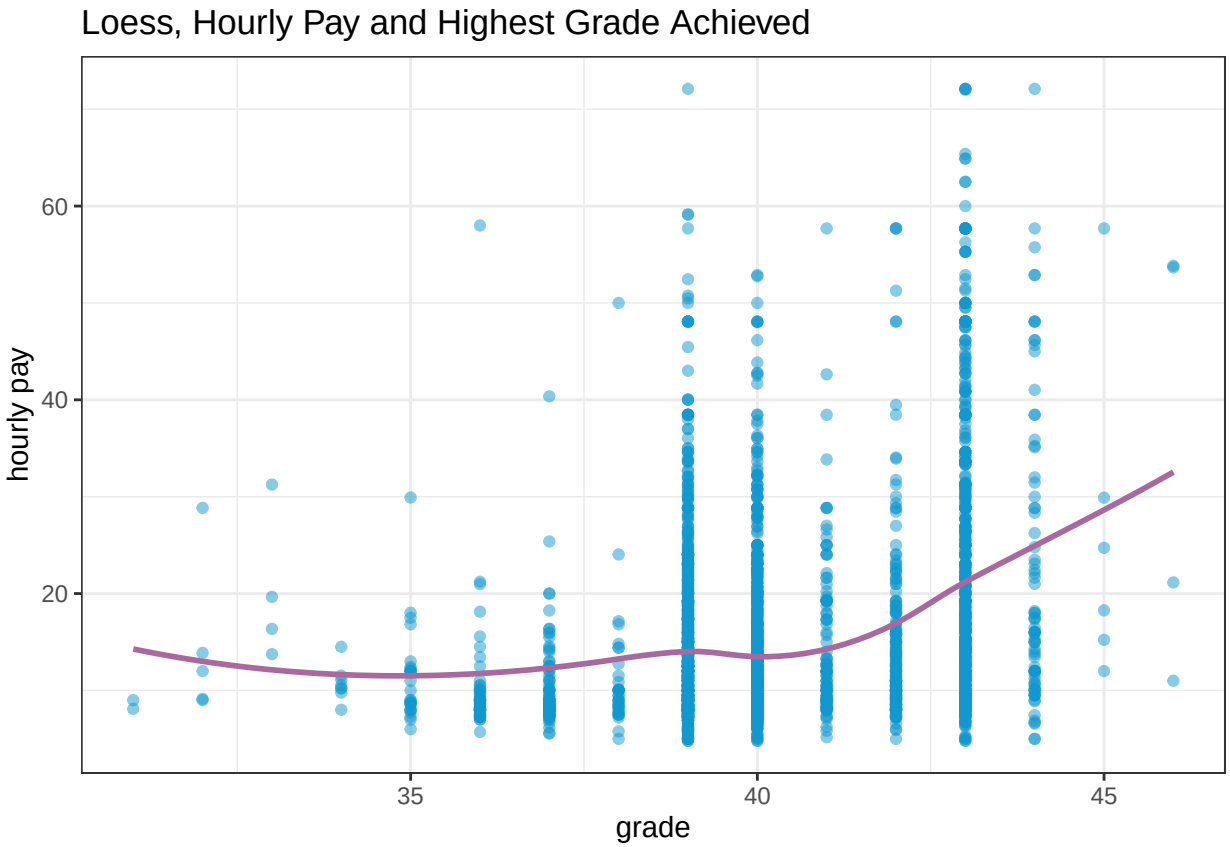

Distribution of Hourly Pays
frequency
800
600
400
200
0
0
20
40
60
hourly pay

PARAMETER SELECTION



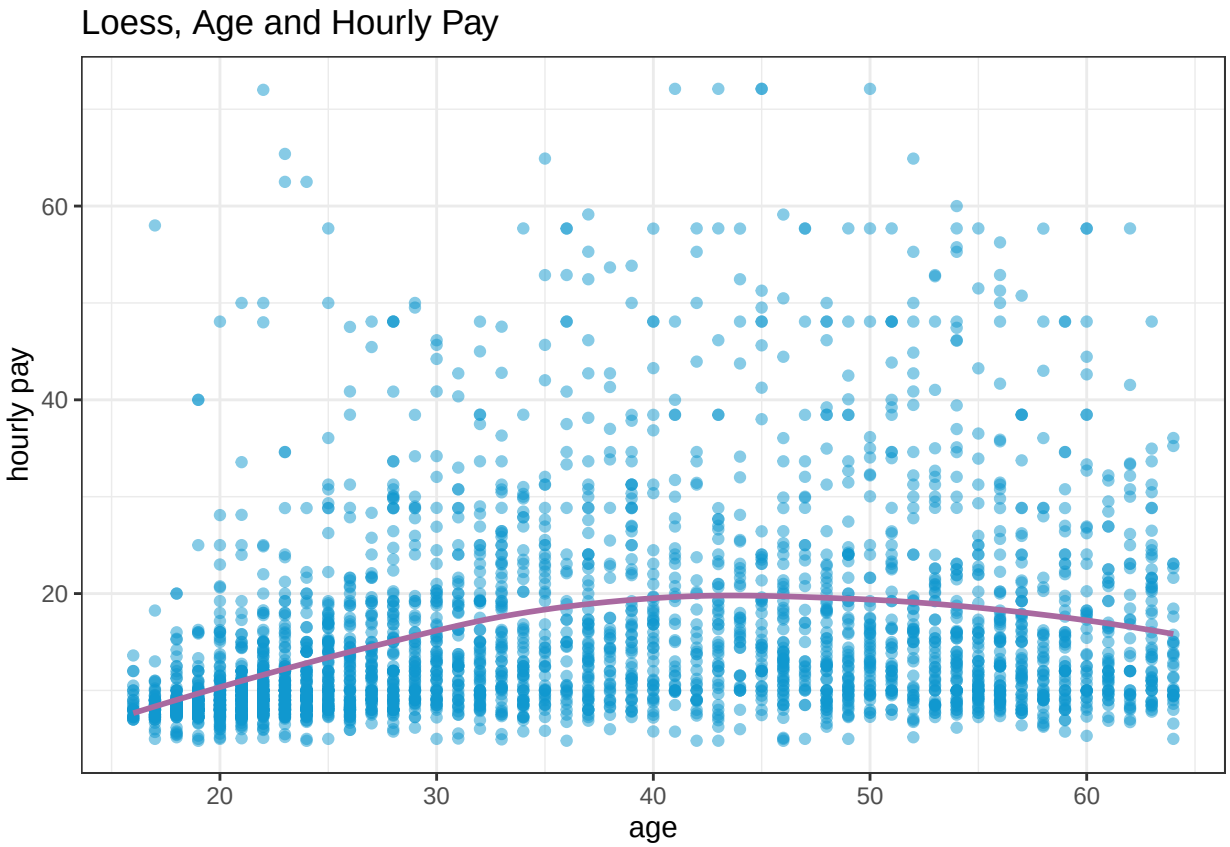
The correlation matrix suggests that hourly pay is positively correlated with the highest grade completed and the age while it is negatively correlated with the marital status and the gender of a retail salesperson.

Highest Grade Achieved and Hourly Pay



It is observed that the hourly pay follows a linear trend after 40. 40 corresponds to 'some college but no degree'. The inference is that retail salespersons who hold a degree tend to earn more in general as their level of education increases. Considering that the highest grade achieved is indeed categorical, I introduce them as factors in the model.

Age and Hourly Pay



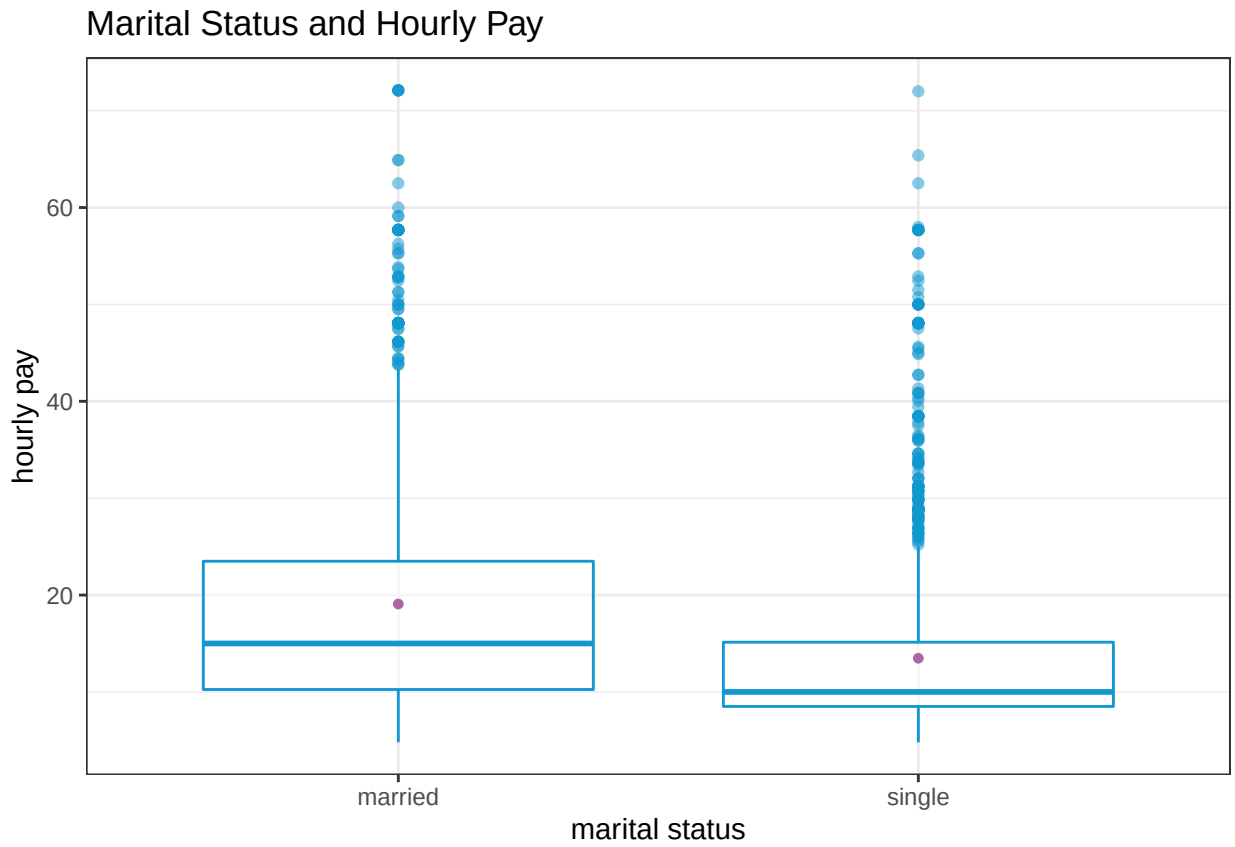
The loess shows that the hourly pay tend to generally increase as the age of a retail sales person increases.

```
retail_salespersons <- retail_salespersons %>% mutate(age_sqr = age ^ 2)
```

Marital Status and Hourly Pay

When the marital attribute is examined, it is seen that the marital status 1 and 2 signifies the presence of a partner when the rest of the values signifies the absence of a partner. Therefore, I create a binary variable named single.

```
retail_salespersons <- retail_salespersons %>% mutate(single = as.numeric(!(marital == 1 |
                                                                             marital == 2)))
```



As it is clearly seen, married retail sales persons have a wider range of hourly pays even though the mean hourly pays are not that different.

Children and Hourly Pay

```
retail_salespersons %>% count(chldpres) %>% arrange(n)
```

```
## # A tibble: 16 x 2
##    chldpres     n
##       <dbl> <int>
##  1       12     1
##  2       15     2
##  3        7     3
##  4       13     4
##  5        9     5
##  6       14     6
##  7       11    17
##  8        6    33
##  9        5    35
## 10        8    58
## 11        2    65
## 12       10    93
## 13        1   104
## 14        4   144
```

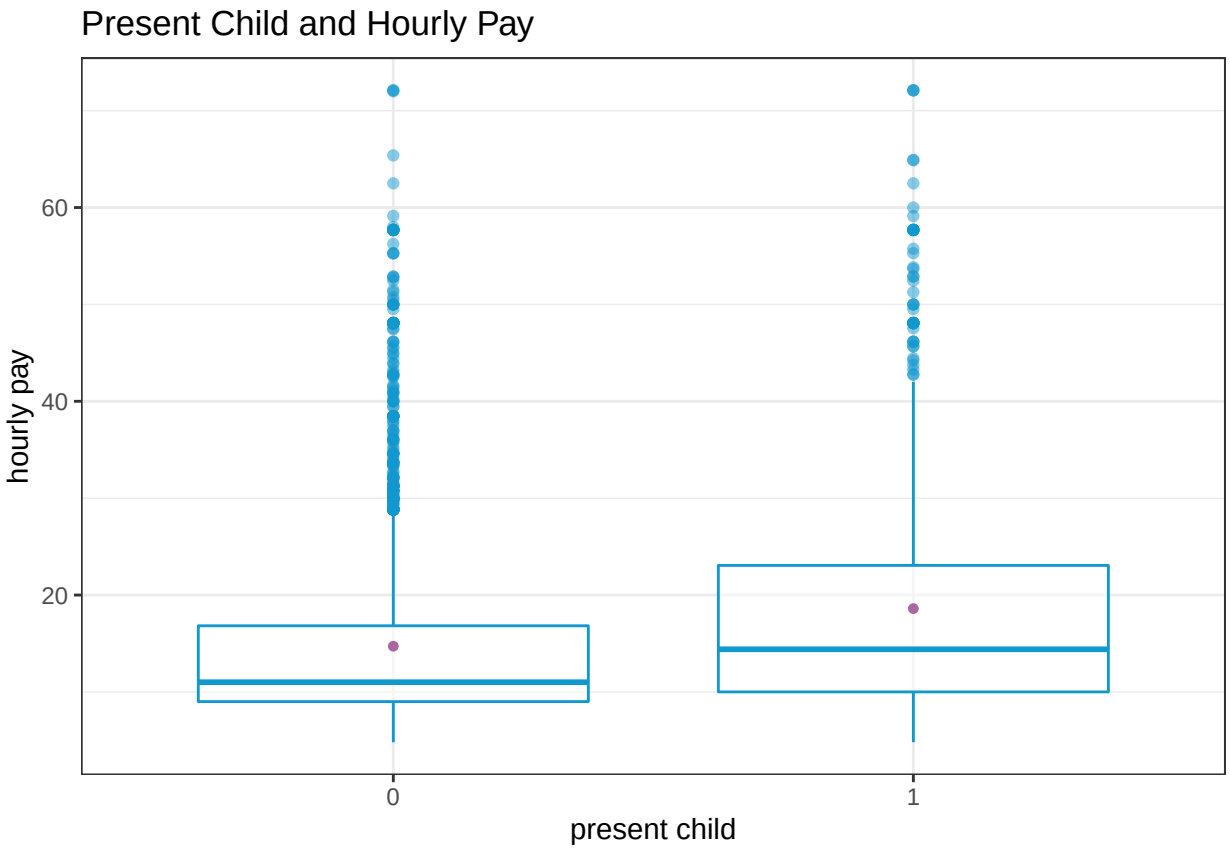
```
## 15         3   190
## 16         0  2401
```

```
# 12,15,7,13,9,14
```

When the present child variable is examined, it is seen that 12,15,7,13,9,15 are observed with a very low frequency. Therefore, they are omitted.

```
retail_salespersons <- retail_salespersons %>% filter( !(chldpres %in% c(12,15,7,13,9,15)))
```

Let's see if the retail salespersons who does not have kids earn less or more.



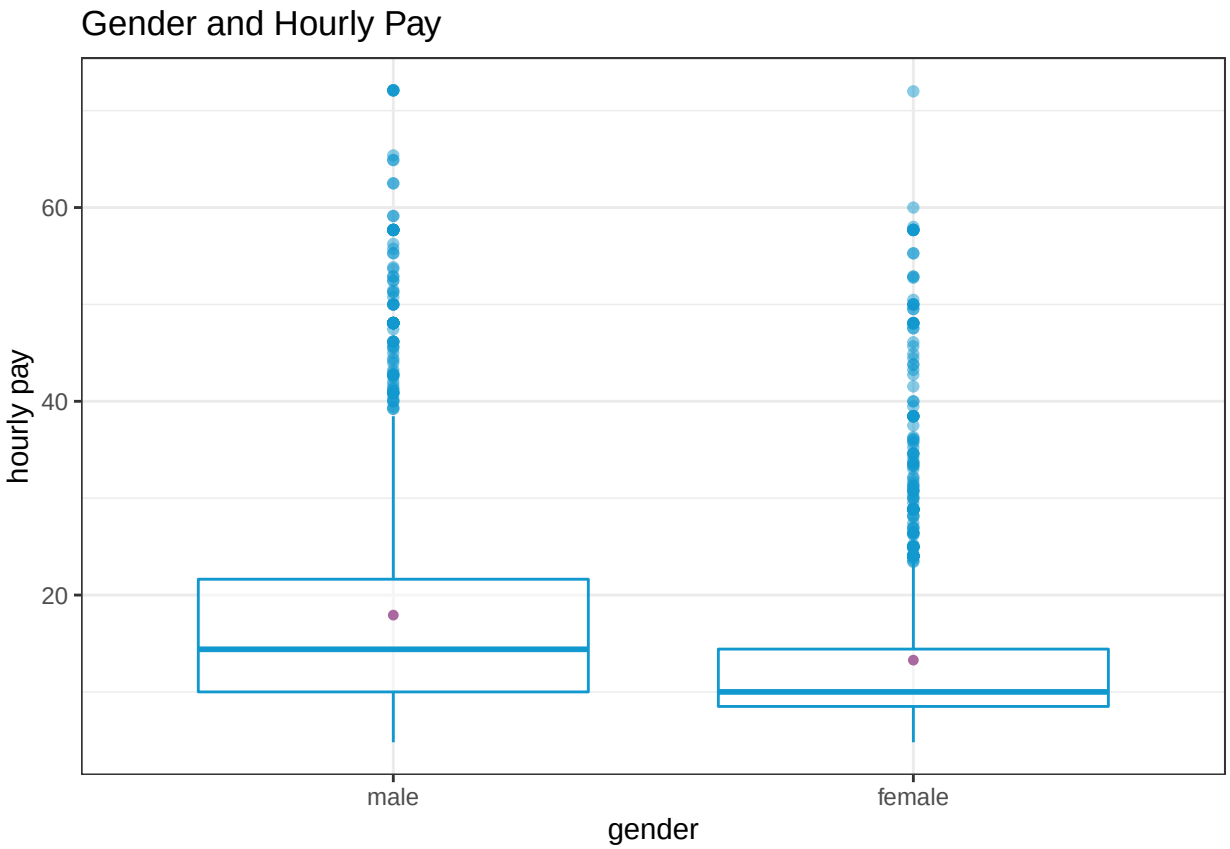
It can be observed that retail salespersons without kids earn less on average with a narrower pay range. Therefore I introduced a binary variable that indicates if a person has a child or not.

```
retail_salespersons <- retail_salespersons %>% mutate(kid=as.numeric(chldpres > 0))
```

Gender and Hourly Pay

I introduce a binary variable to indicate if a retail salesperson is a woman or not.

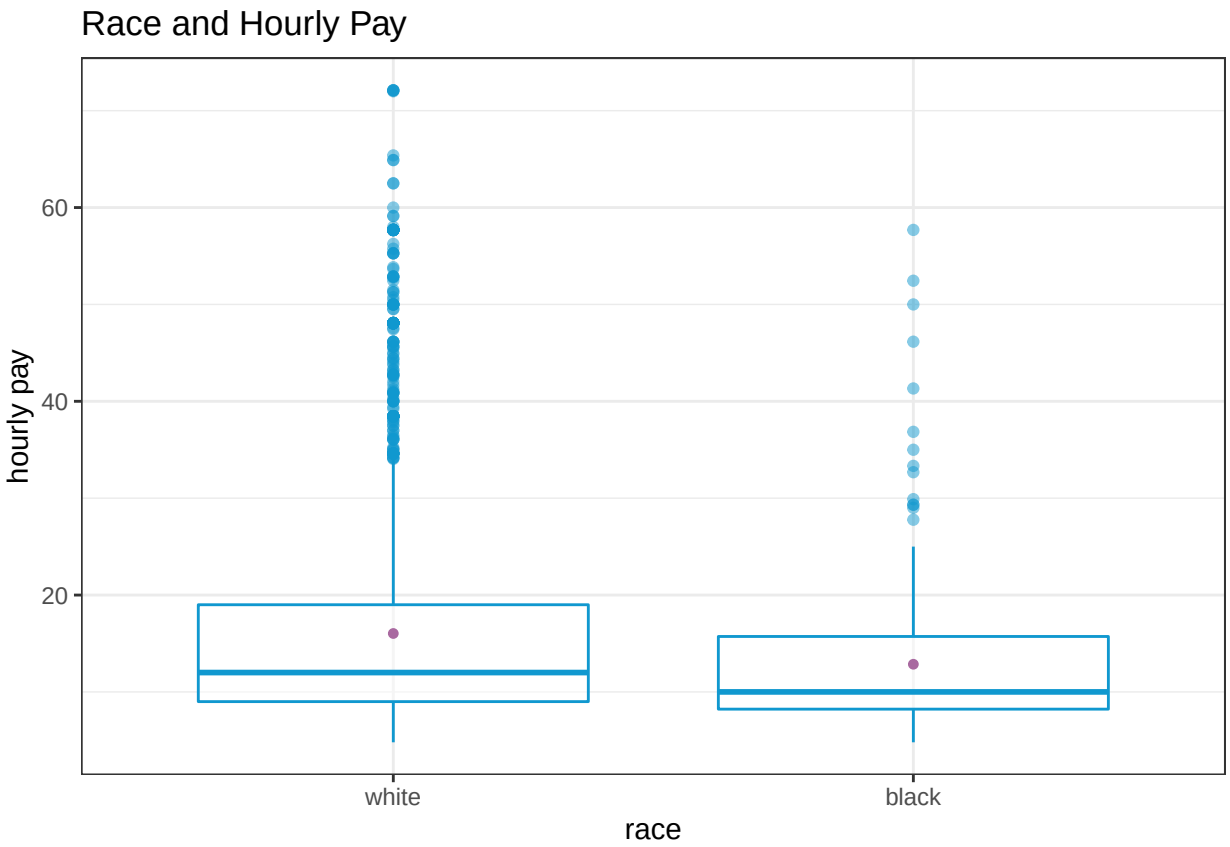
```
retail_salespersons <- retail_salespersons %>% mutate(female = as.numeric(sex == 2))
```



We see that female retail salespersons have a narrower hourly pay range than male retail salespersons.

Race and Hourly Pay

To see if the hourly pay varies with the race of the retail sales person, let's compare white and black people.



The race and hourly pay plot shows that black people have narrower pay range and lower mean hourly pay.Let's create the relevant variables considering asian and hispanic people as well.

```
retail_salespersons <- retail_salespersons %>% mutate(white=as.numeric(race==1), black = as.numeric(rac
                                asian = as.numeric(race==4),hispanic = as.numeric(!is.na(ethnic)),
                                other = as.numeric(white==0 & black==0 & asian==0 & hispanic==0))
```

MODELS

```
model1 <- as.formula(earning_hour ~ age + age_sqr  + as.factor(grade92),
                     retail_salespersons)
model2 <- as.formula(earning_hour ~ age + age_sqr + as.factor(grade92) +
                      female + single + female*single,
                      retail_salespersons)
model3 <- as.formula(earning_hour ~ age + age_sqr + as.factor(grade92) +
                      female + single +
                      kid + female*kid + female*single,
                      retail_salespersons)
model4  <- as.formula(earning_hour ~ age + age_sqr + as.factor(grade92) +
                       female + single +
                       kid + female*kid + female*single +
                       white + black + asian + hispanic + other,
                       retail_salespersons)
```

The parameter choices for models heavily depend on the correlation coefficients. As mentioned above, the age and the highest grade achieved are positively correlated with the hourly payment more than rest of the variables. Apart from this, correlation matrix revealed that the gender and marital status are negatively correlated with the hourly pay. As it is observed in the gender and hourly pay boxplot, female retail salespersons have a narrower pay range even though the average is close to male's mean hourly pay. Therefore, I decided to take gender into account. When it comes to marital status, it makes sense that people who live alone are most likely be satisfied with lower pays. Indeed, when the retail salespersons are categorized as married and single, it is observed clearly from the marital status and hourly pay boxplot that single people earn less on average and have a narrower pay range. Retail salespersons who do not have kids earn less on average. Usually women earn less in case they have children becuase they are caregivers and work part time and have lower salaries. The interaction term of kid,female may represent this as well as the single,female. Lastly, the race and hourly pay boxplot revealed that black people have narrower pay range and lower mean hourly pay. Considering the race factor white, black, asian, hispanic and other variables are created and added as parameters.

Evaluation of Models

Table 1: Model Comparision

Model	N predictors	R-squared	Training RMSE	BIC
(1)	17	0.1982396	9.330455	23132.81
(2)	20	0.2433670	9.064068	22974.72
(3)	22	0.2441243	9.059531	22987.67
(4)	27	0.2488150	9.031377	23008.36

We see that traning RMSE is the lowest for the most complicated model. However the BIC is higher than the second and third models. It is interesting to observe that the BIC score for the least complicated model is the highest as well as the RMSE.

CROSSFOLD VALIDATION

Table 2: 5-fold RMSE results

models	n	rmse_cv
reg1	17	9.316148
reg2	20	9.091817
reg3	22	9.103479
reg4	27	9.096348

The RMSE scores of models do not vary significantly. The score of the second least complicated model is the lowest. Therefore, I would choose the second model with 20 parameters that considers the age, grade, gender and marital status.