Data Cleaning and EDA

Umi Yamaguchi, Benjamin Rodovinski

```
# Load the dataset
data <- read.csv("data.csv", header=T)

# head(data)
# str(data)
# summary(data)
```

```
# Install
install.packages("ggplot2")
```

```
## Installing package into '/usr/local/lib/R/site-library'
## (as 'lib' is unspecified)
```

```
library("ggplot2")
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

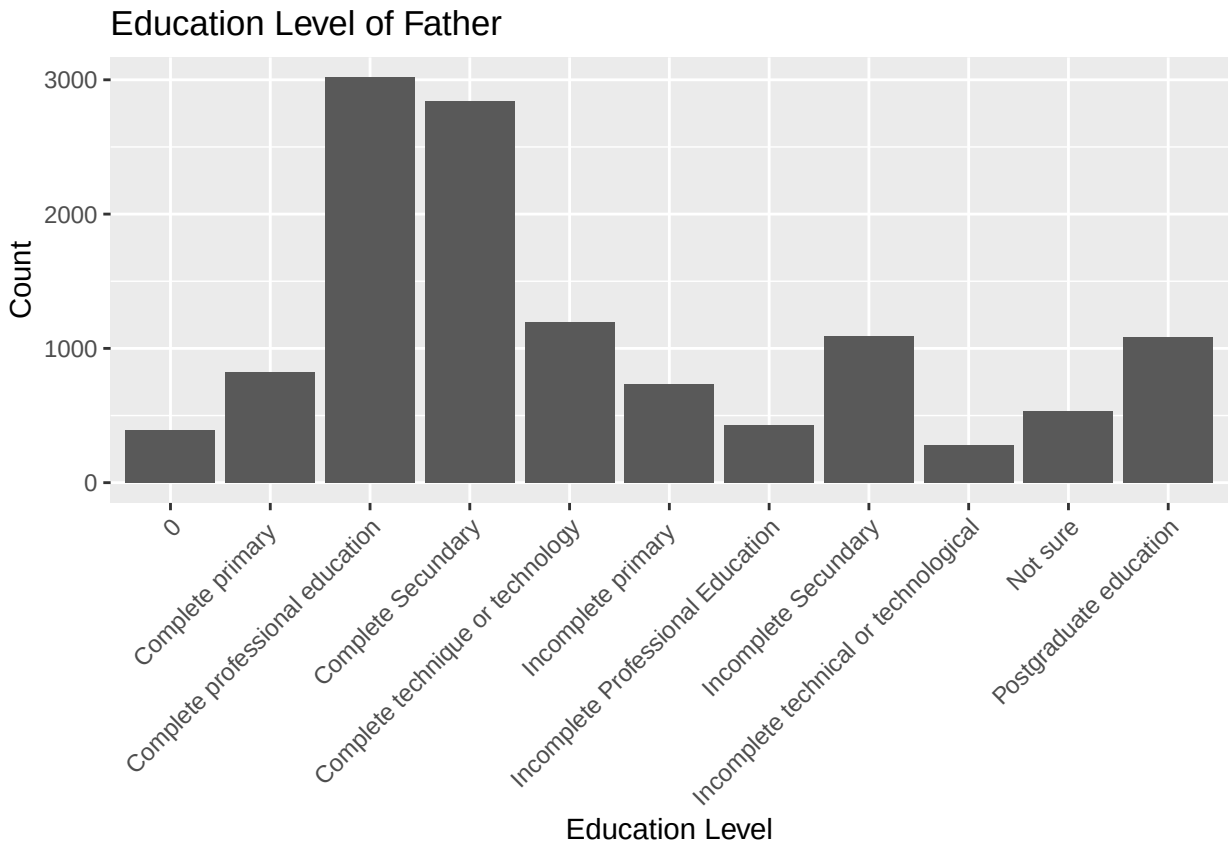
```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

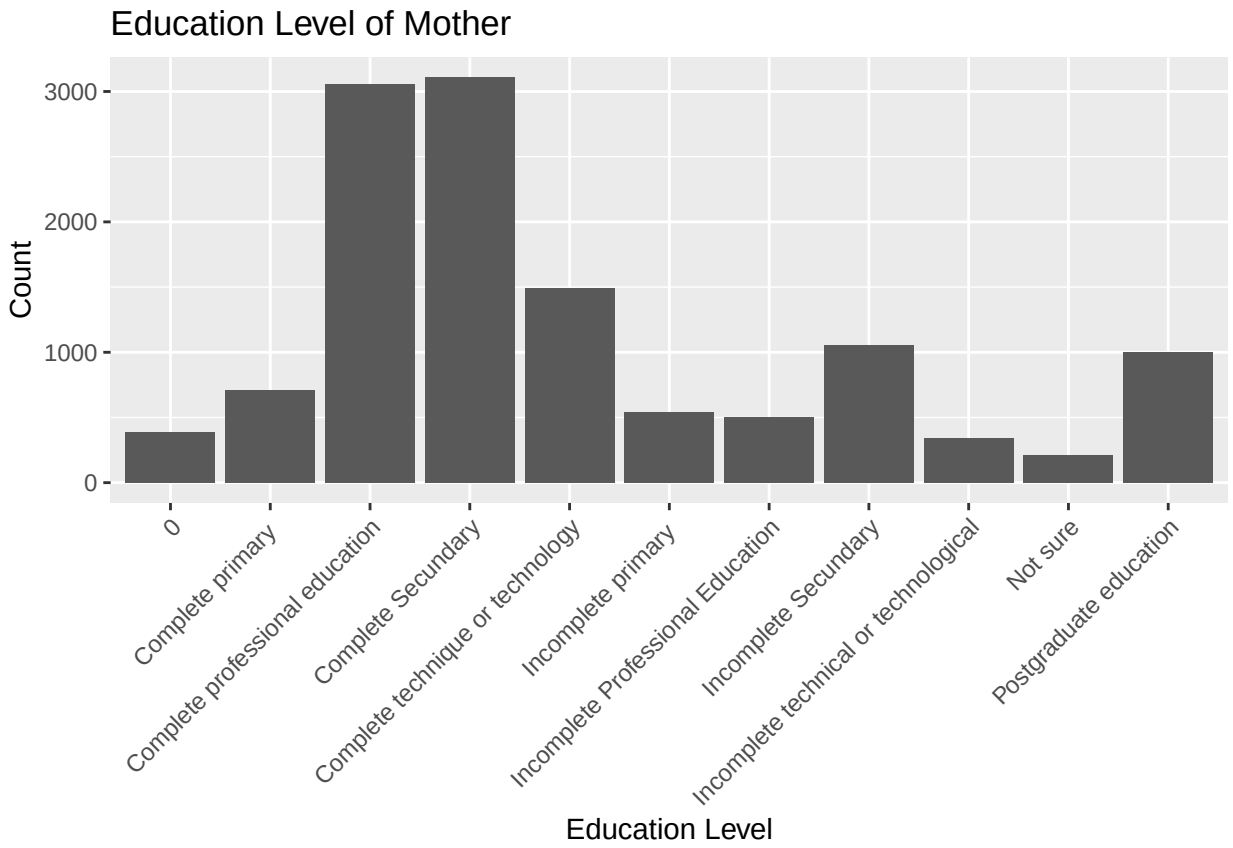
Encode Categorical Variables

1. Education of Father and Mother

```
ggplot(data, aes(x = EDU_FATHER)) +
  geom_bar() +
  labs(title = "Education Level of Father", x = "Education Level", y = "Count") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```r
ggplot(data, aes(x = EDU_MOTHER)) +
  geom_bar() +
  labs(title = "Education Level of Mother", x = "Education Level", y = "Count") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Clean up the education levels.

We will keep the Complete and Incomplete Educations data for analyzing difference between Complete and Incomplete going to affect on Grading Mark

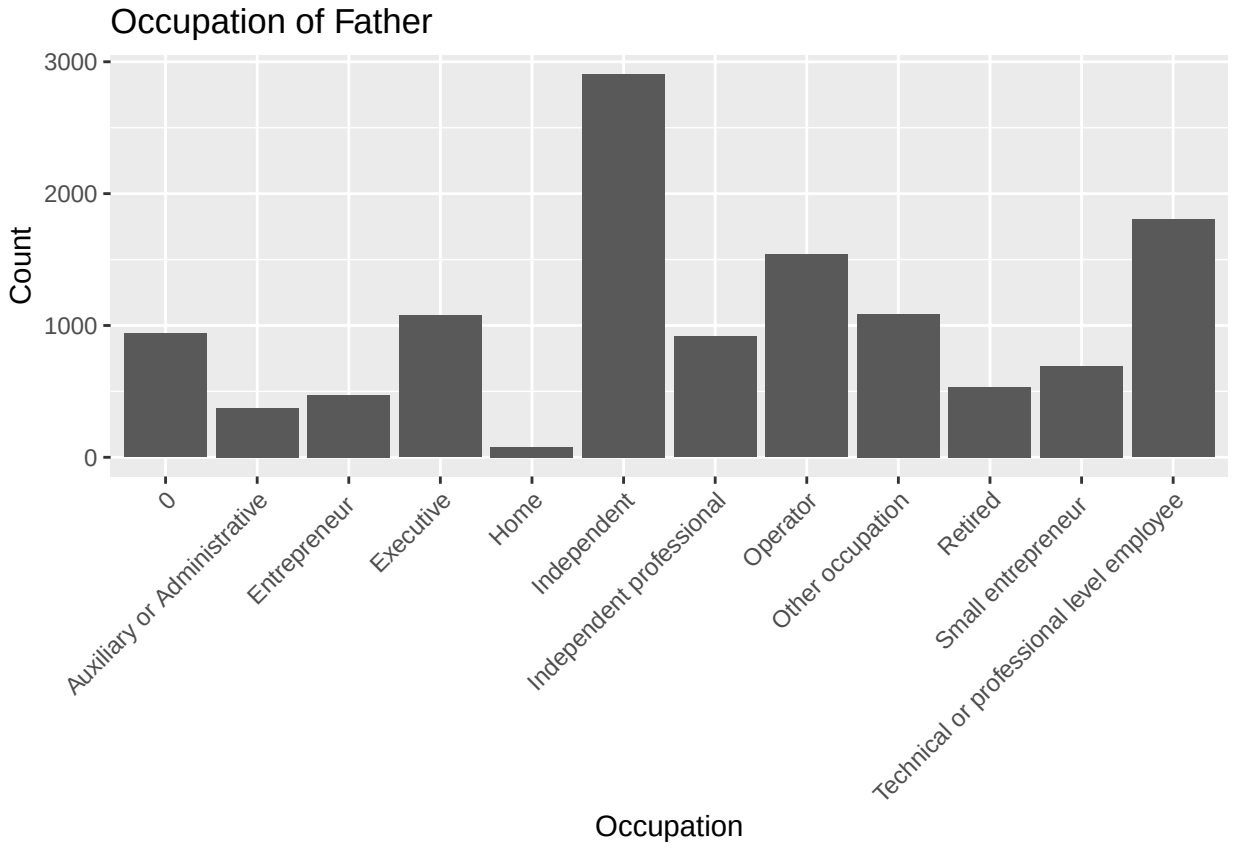
```
# Assuming 'EDU_Father' contains the education levels
data$EDU_FATHER <- dplyr::case_when(
  data$EDU_FATHER == "0" ~ "Unknown",
  data$EDU_FATHER == "Complete primary " ~ "Complete Primary Education",
  data$EDU_FATHER == "Incomplete primary " ~ "Incomplete Primary Education",
  data$EDU_FATHER == "Complete Secundary" ~ "Complete Secondary Education",
  data$EDU_FATHER == "Incomplete Secundary" ~ "Incomplete Secondary Education",
  data$EDU_FATHER == "Complete technique or technology" ~ "Complete Technical Education",
  data$EDU_FATHER == "Incomplete technical or technological" ~ "Incomplete Technical Education",
  data$EDU_FATHER == "Complete professional education" ~ "Complete Professional Education",
  data$EDU_FATHER == "Incomplete Professional Education" ~ "Incomplete Professional Education",
  data$EDU_FATHER == "Not sure" ~ "Unknown",
  data$EDU_FATHER == "Postgraduate education" ~ "Postgraduate Education"
)

# Assuming 'EDU_Father' contains the education levels
data$EDU_MOTHER <- dplyr::case_when(
  data$EDU_MOTHER == "0" ~ "Unknown",
  data$EDU_MOTHER == "Complete primary " ~ "Complete Primary Education",
  data$EDU_MOTHER == "Incomplete primary " ~ "Incomplete Primary Education",
  data$EDU_MOTHER == "Complete Secundary" ~ "Complete Secondary Education",
  data$EDU_MOTHER == "Incomplete Secundary" ~ "Incomplete Secondary Education",
  data$EDU_MOTHER == "Complete technique or technology" ~ "Complete Technical Education",
```

```
  data$EDU_MOTHER == "Incomplete technical or technological" ~ "Incomplete Technical Education",
  data$EDU_MOTHER == "Complete professional education" ~ "Complete Professional Education",
  data$EDU_MOTHER == "Incomplete Professional Education" ~ "Incomplete Professional Education",
  data$EDU_MOTHER == "Not sure" ~ "Unknown",
  data$EDU_MOTHER == "Postgraduate education" ~ "Postgraduate Education"
)
```

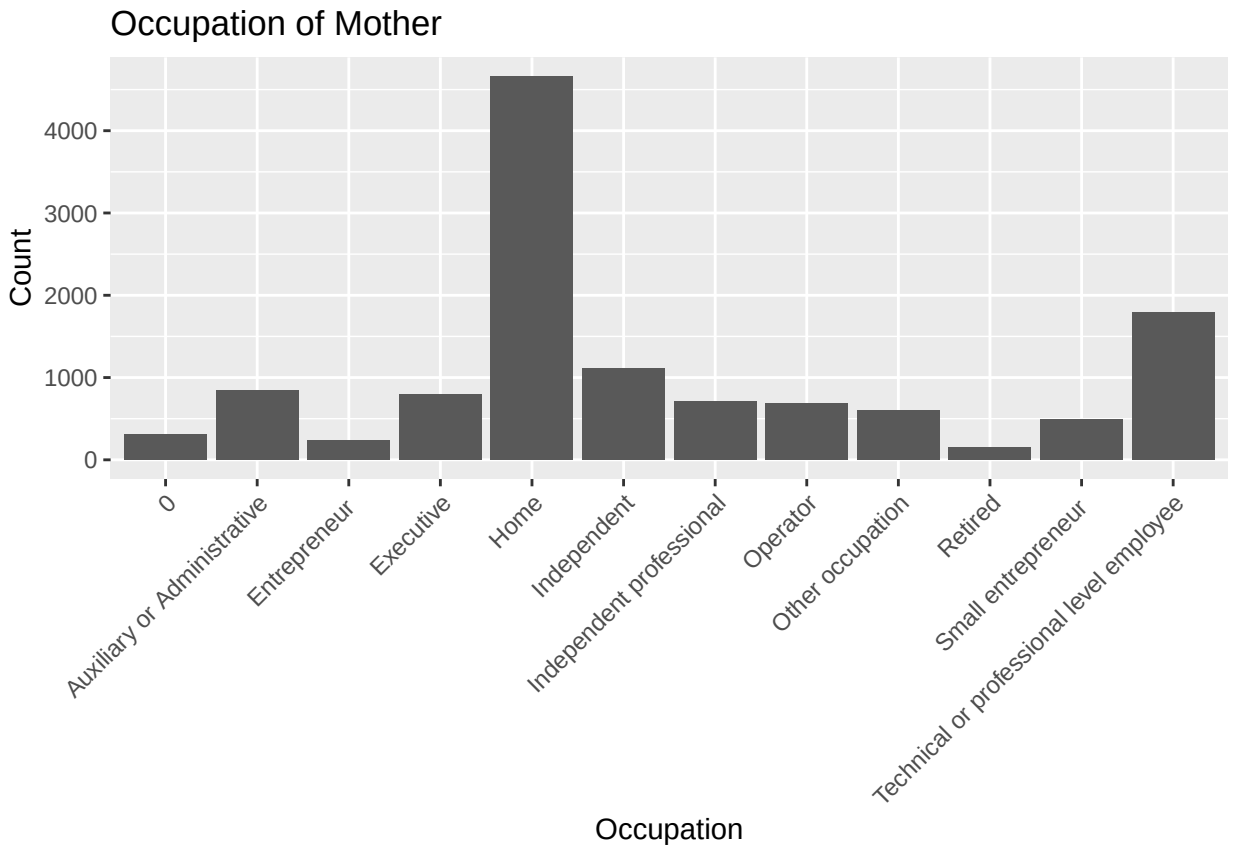
2. Occupation of Father and Mother

We will keep this as is

```
ggplot(data, aes(x = OCC_FATHER)) +
  geom_bar() +
  labs(title = "Occupation of Father", x = "Occupation", y = "Count") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```
ggplot(data, aes(x = OCC_MOTHER)) +
  geom_bar() +
  labs(title = "Occupation of Mother", x = "Occupation", y = "Count") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



2.

University as Tier list

```
#unique(data$UNIVERSITY)
```

Using University Ranking System to identify correlation between University level and Overall Score

```
data <- data %>%
  mutate(UNIVERSITY = case_when(
    UNIVERSITY == "UNIVERSIDAD DE SANTANDER - UDES-BUCARAMANGA"  ~ 4,
    UNIVERSITY == "UNIVERSIDAD DE LOS ANDES-BOGOTÁ D.C." ~ 5,
    UNIVERSITY == "UNIVERSIDAD NACIONAL ABIERTA Y A DISTANCIA UNAD-BOGOTÁ D.C." ~ 3,
    UNIVERSITY == "UNIVERSIDAD CATOLICA DE PEREIRA-PEREIRA" ~ 2,
    UNIVERSITY == "UNIVERSIDAD INDUSTRIAL DE SANTANDER-BUCARAMANGA" ~ 4,
    UNIVERSITY == "UNIVERSIDAD TECNOLOGICA DE PEREIRA - ITP-PEREIRA" ~ 4,
    UNIVERSITY == "UNIVERSIDAD ANTONIO NARIÑO-BOGOTÁ D.C." ~ 2,
    UNIVERSITY == "UNIVERSIDAD COOPERATIVA DE COLOMBIA-BOGOTÁ D.C." ~ 3,
    UNIVERSITY == "UNIVERSIDAD DEL NORTE-BARRANQUILLA" ~ 4,
    UNIVERSITY == "UNIVERSIDAD EAFIT--MEDELLIN" ~ 4,
    UNIVERSITY == "POLITECNICO GRANCOLOMBIANO-BOGOTÁ D.C." ~ 3,
    UNIVERSITY == "INSTITUCION UNIVERSITARIA CENTRO DE ESTUDIOS SUPERIORES MARIA GORETTI-PASTO" ~ 1,
    UNIVERSITY == "UNIVERSIDAD DE NARIÑO-PASTO" ~ 2,
    UNIVERSITY == "CORPORACION UNIVERSITARIA AUTONOMA DE NARIÑO -AUNAR-PASTO" ~ 2,
    UNIVERSITY == "UNIVERSIDAD NACIONAL DE COLOMBIA-MANIZALES" ~ 3,
    UNIVERSITY == "UNIVERSIDAD MARIANA-PASTO" ~ 2,
    UNIVERSITY == "UNIVERSIDAD AUTONOMA DE OCCIDENTE-CALI" ~ 2,
    UNIVERSITY == "PONTIFICIA UNIVERSIDAD JAVERIANA-CALI" ~ 5,
    UNIVERSITY == "UNIVERSIDAD DEL VALLE-CALI" ~ 4,
    UNIVERSITY == "UNIVERSIDAD NACIONAL DE COLOMBIA-BOGOTÁ D.C." ~ 5,
    UNIVERSITY == "UNIVERSIDAD DE SAN BUENAVENTURA-CALI" ~ 2,
```

```
    UNIVERSITY == "UNIVERSIDAD ICESI-CALI" ~ 4,
    UNIVERSITY == "UNIVERSIDAD DE IBAGUE-IBAGUE" ~ 2,
    UNIVERSITY == "UNIVERSIDAD SANTIAGO DE CALI-CALI" ~ 2,
    UNIVERSITY == "UNIVERSIDAD AUTONOMA DEL CARIBE-BARRANQUILLA" ~ 2,
    UNIVERSITY == "CORPORACION UNIVERSIDAD DE LA COSTA, CUC-BARRANQUILLA" ~ 3,
    UNIVERSITY == "UNIVERSIDAD DEL ATLANTICO-BARRANQUILLA" ~ 3,
    UNIVERSITY == "UNIVERSIDAD MILITAR\"NUEVA GRANADA\"-BOGOTÁ D.C." ~ 3,
    UNIVERSITY == "CORPORACION UNIVERSITARIA MINUTO DE DIOS -UNIMINUTO-BOGOTÁ D.C." ~ 3,
    UNIVERSITY == "FUNDACION UNIVERSIDAD DE AMERICA-BOGOTÁ D.C." ~ 2,
    UNIVERSITY == "UNIVERSIDAD SERGIO ARBOLEDA-BOGOTÁ D.C." ~ 3,
    UNIVERSITY == "PONTIFICIA UNIVERSIDAD JAVERIANA-BOGOTÁ D.C." ~ 4,
    UNIVERSITY == "UNIVERSIDAD CATOLICA DE COLOMBIA-BOGOTÁ D.C." ~ 4,
    UNIVERSITY == "FUNDACION UNIVERSITARIA LOS LIBERTADORES-BOGOTÁ D.C." ~ 2,
    UNIVERSITY == "CORPORACION TECNOLOGICA INDUSTRIAL COLOMBIANA - TEINCO-BOGOTÁ D.C." ~ 1,
    UNIVERSITY == "UNIVERSITARIA AGUSTINIANA- UNIAGUSTINIANA-BOGOTÁ D.C." ~ 2,
    UNIVERSITY == "UNIVERSIDAD DEL CAUCA-POPAYAN" ~ 4,
    UNIVERSITY == "UNIVERSIDAD DE ANTIOQUIA-MEDELLIN" ~ 4,
    UNIVERSITY == "UNIVERSIDAD DE LA SABANA-CHIA" ~ 4,
    UNIVERSITY == "ESCUELA COLOMBIANA DE INGENIERIA\"JULIO GARAVITO\"-BOGOTÁ D.C." ~ 2,
    UNIVERSITY == "UNIVERSIDAD SANTO TOMAS-BOGOTÁ D.C." ~ 3,
    UNIVERSITY == "FUNDACION UNIVERSITARIA INTERNACIONAL DEL TROPICO AMERICANO-YOPAL" ~ 2,
    UNIVERSITY == "FUNDACION UNIVERSIDAD DE BOGOTA\"JORGE TADEO LOZANO\"-BOGOTÁ D.C." ~ 4,
    UNIVERSITY == "UNIVERSIDAD SANTO TOMAS-TUNJA" ~ 3,
    UNIVERSITY == "UNIVERSIDAD LIBRE-PEREIRA" ~ 3,
    UNIVERSITY == "UNIVERSIDAD PONTIFICIA BOLIVARIANA-BUCARAMANGA" ~ 4,
    UNIVERSITY == "CORPORACION UNIVERSITARIA DE INVESTIGACION Y DESARROLLO -\"UDI\"-BUCARAMANGA" ~ 2,
    UNIVERSITY == "UNIVERSIDAD MANUELA BELTRAN-UMB--BOGOTÁ D.C." ~ 2,
    UNIVERSITY == "UNIVERSIDAD AUTONOMA DE BUCARAMANGA-UNAB-BUCARAMANGA" ~ 3,
    UNIVERSITY == "UNIVERSIDAD LIBRE-BARRANQUILLA" ~ 2,
    UNIVERSITY == "UNIVERSIDAD SANTO TOMAS-BUCARAMANGA" ~ 3,
    UNIVERSITY == "UNIVERSIDAD MANUELA BELTRAN-UMB--BUCARAMANGA" ~ 2,
    UNIVERSITY == "UNIVERSIDAD DE BOYACA - UNIBOYACA-TUNJA" ~ 2,
    UNIVERSITY == "UNIVERSIDAD PEDAGOGICA Y TECNOLOGICA DE COLOMBIA-TUNJA" ~ 4,
    UNIVERSITY == "UNIVERSIDAD DE PAMPLONA-PAMPLONA" ~ 3,
    UNIVERSITY == "UNIVERSIDAD DISTRITAL\"FRANCISCO JOSE DE CALDAS\"-BOGOTÁ D.C." ~ 3,
    UNIVERSITY == "FUNDACION UNIVERSITARIA DE SAN GIL - UNISANGIL -SAN GIL" ~ 2,
    UNIVERSITY == "UNIVERSIDAD EL BOSQUE-BOGOTÁ D.C." ~ 4,
    UNIVERSITY == "UNIDADES TECNOLOGICAS DE SANTANDER-BUCARAMANGA" ~ 4,
    UNIVERSITY == "CORPORACION UNIFICADA NACIONAL DE EDUCACION SUPERIOR-CUN-BOGOTÁ D.C." ~ 2,
    UNIVERSITY == "CORPORACION UNIVERSITARIA DEL CARIBE - CECAR-SINCELEJO" ~ 3,
    UNIVERSITY == "UNIVERSIDAD DEL SINÚ 'Elías Bechara Zainúm' - UNISINÚ-MONTERIA" ~ 2,
    UNIVERSITY == "UNIVERSIDAD DE SUCRE-SINCELEJO" ~ 2,
    UNIVERSITY == "UNIVERSIDAD NACIONAL DE COLOMBIA-MEDELLIN" ~ 5,
    UNIVERSITY == "UNIVERSIDAD DEL MAGDALENA - UNIMAGDALENA-SANTA MARTA" ~ 4,
    UNIVERSITY == "UNIVERSIDAD SIMON BOLIVAR-BARRANQUILLA" ~ 2,
    UNIVERSITY == "CORPORACION UNIVERSITARIA ANTONIO JOSE DE SUCRE - CORPOSUCRE-SINCELEJO" ~ 1,
    UNIVERSITY == "UNIVERSIDAD DE CORDOBA-MONTERIA" ~ 4,
    UNIVERSITY == "FUNDACION UNIVERSITARIA TECNOLOGICO COMFENALCO - CARTAGENA      -CARTAGENA" ~ 2,
    UNIVERSITY == "UNIVERSIDAD DE LA GUAJIRA-RIOHACHA" ~ 2,
    UNIVERSITY == "UNIVERSIDAD AUTONOMA LATINOAMERICANA-UNAULA-MEDELLIN" ~ 2,
    UNIVERSITY == "UNIVERSIDAD TECNOLOGICA DE BOLIVAR-CARTAGENA" ~ 4,
    UNIVERSITY == "UNIVERSIDAD AUTONOMA DE MANIZALES-MANIZALES" ~ 2,
    UNIVERSITY == "CORPORACION UNIVERSIDAD PILOTO DE COLOMBIA-GIRARDOT" ~ 2,
```

```
    UNIVERSITY == "UNIVERSIDAD DEL QUINDIO-ARMENIA" ~ 3,
    UNIVERSITY == "ESCUELA DE ADMINISTRACION Y MERCADOTECNIA DEL QUINDIO-ARMENIA" ~ 1,
    UNIVERSITY == "UNIVERSIDAD LA GRAN COLOMBIA-BOGOTÁ D.C." ~ 2,
    UNIVERSITY == "UNIVERSIDAD CENTRAL-BOGOTÁ D.C." ~ 2,
    UNIVERSITY == "FUNDACION UNIVERSITARIA AGRARIA DE COLOMBIA -UNIAGRARIA-BOGOTÁ D.C." ~ 2,
    UNIVERSITY == "UNIVERSIDAD ECCI-BOGOTÁ D.C." ~ 2,
    UNIVERSITY == "UNIVERSIDAD PEDAGOGICA Y TECNOLOGICA DE COLOMBIA-SOGAMOSO" ~ 3,
    UNIVERSITY == "UNIVERSIDAD TECNOLOGICA DEL CHOCO\"DIEGO LUIS CORDOBA\"-QUIBDO" ~ 2,
    UNIVERSITY == "UNIVERSIDAD DE MEDELLIN-MEDELLIN" ~ 4,
    UNIVERSITY == "UNIVERSIDAD COOPERATIVA DE COLOMBIA-MEDELLIN" ~ 3,
    UNIVERSITY == "CORPORACION UNIVERSITARIA DEL HUILA-CORHUILA-NEIVA" ~ 1,
    UNIVERSITY == "UNIVERSIDAD SURCOLOMBIANA-NEIVA" ~ 2,
    UNIVERSITY == "CORPORACION UNIVERSITARIA AUTONOMA DEL CAUCA-POPAYAN" ~ 2,
    UNIVERSITY == "UNIVERSIDAD PONTIFICIA BOLIVARIANA-MEDELLIN" ~ 4,
    UNIVERSITY == "UNIVERSIDAD FRANCISCO DE PAULA SANTANDER-CUCUTA" ~ 2,
    UNIVERSITY == "UNIVERSIDAD POPULAR DEL CESAR-VALLEDUPAR" ~ 2,
    UNIVERSITY == "CORPORACION UNIVERSITARIA REPUBLICANA-BOGOTÁ D.C." ~ 2,
    UNIVERSITY == "FUNDACION UNIVERSITARIA COLOMBO INTERNACIONAL - UNICOLOMBO-CARTAGENA" ~ 1,
    UNIVERSITY == "CORPORACION UNIVERSITARIA AMERICANA-BARRANQUILLA" ~ 2,
    UNIVERSITY == "CORPORACION UNIVERSITARIA DEL META-VILLAVICENCIO" ~ 1,
    UNIVERSITY == "UNIVERSIDAD DE LOS LLANOS-VILLAVICENCIO" ~ 2,
    UNIVERSITY == "UNIVERSIDAD DE LA SALLE-BOGOTÁ D.C." ~ 4,
    UNIVERSITY == "UNIVERSIDAD INCCA DE COLOMBIA-BOGOTÁ D.C." ~ 2,
    UNIVERSITY == "INSTITUCION UNIVERSITARIA ITSA-SOLEDAD" ~ 2,
    UNIVERSITY == "UNIVERSIDAD EIA-MEDELLIN" ~ 2,
    UNIVERSITY == "UNIVERSIDAD DE SAN BUENAVENTURA-MEDELLIN" ~ 2,
    UNIVERSITY == "POLITECNICO COLOMBIANO\"JAIME ISAZA CADAVID\"-MEDELLIN" ~ 2,
    UNIVERSITY == "INSTITUCION  UNIVERSITARIA PASCUAL BRAVO-MEDELLIN" ~ 2,
    UNIVERSITY == "COLEGIO MAYOR DE ANTIOQUIA-MEDELLIN" ~ 2,
    UNIVERSITY == "UNIVERSIDAD PONTIFICIA BOLIVARIANA-MONTERIA" ~ 4,
    UNIVERSITY == "INSTITUTO TECNOLOGICO METROPOLITANO-MEDELLIN" ~ 4,
    UNIVERSITY == "CORPORACION UNIVERSITARIA LASALLISTA-CALDAS" ~ 2,
    UNIVERSITY == "INSTITUCION UNIVERSITARIA DE ENVIGADO-ENVIGADO" ~ 2,
    UNIVERSITY == "CORPORACION UNIVERSITARIA  UNITEC-BOGOTÁ D.C." ~ 2,
    UNIVERSITY == "UNIVERSIDAD LIBRE-BOGOTÁ D.C." ~ 2,
    UNIVERSITY == "INSTITUTO TOLIMENSE DE FORMACION TECNICA PROFESIONAL -ESPINAL" ~ 1,
    UNIVERSITY == "FUNDACION UNIVERSITARIA DE POPAYAN-POPAYAN" ~ 2,
    UNIVERSITY == "UNIVERSIDAD LIBRE-CALI" ~ 3,
    UNIVERSITY == "UNIVERSIDAD LIBRE-CUCUTA" ~ 3,
    UNIVERSITY == "UNIVERSIDAD FRANCISCO DE PAULA SANTANDER-OCAÑA" ~ 2,
    UNIVERSITY == "CORPORACION UNIVERSITARIA EMPRESARIAL ALEXANDER VON HUMBOLDT - C.U.E.-ARMENIA" ~ 2,
    UNIVERSITY == "INSTITUCION UNIVERSITARIA ANTONIO JOSE CAMACHO - UNIAJC-CALI" ~ 2,
    UNIVERSITY == "UNIDAD CENTRAL DEL VALLE DEL CAUCA-TULUA" ~ 2,
    UNIVERSITY == "CORPORACION UNIVERSIDAD PILOTO DE COLOMBIA-BOGOTÁ D.C." ~ 2,
    UNIVERSITY == "FUNDACION UNIVERSITARIA AUTONOMA DE COLOMBIA -FUAC--BOGOTÁ D.C." ~ 3,
    UNIVERSITY == "FUNDACION UNIVERSITARIA KONRAD LORENZ-BOGOTÁ D.C." ~ 3,
    UNIVERSITY == "UNIVERSIDAD DE CARTAGENA-CARTAGENA" ~ 4,
    UNIVERSITY == "CORPORACION POLITECNICO DE LA COSTA ATLANTICA-BARRANQUILLA" ~ 1,
    UNIVERSITY == "FUNDACION UNIVERSITARIA CAFAM-BOGOTÁ D.C." ~ 3,
    UNIVERSITY == "INSTITUCION UNIVERSITARIA DE COLOMBIA - UNIVERSITARIA DE COLOMBIA-BOGOTÁ D.C." ~ 5,
    UNIVERSITY == "UNIVERSIDAD EAN-BOGOTÁ D.C." ~ 2,
    UNIVERSITY == "UNIVERSIDAD DE CUNDINAMARCA-UDEC-FUSAGASUGA" ~ 2,
    UNIVERSITY == "ESCUELA MILITAR DE AVIACION\"MARCO FIDEL SUAREZ\"-CALI" ~ 1,
```

```
    UNIVERSITY == "ESCUELA TECNOLOGICA INSTITUTO TECNICO CENTRAL -BOGOTÁ D.C." ~ 2,
    UNIVERSITY == "UNIVERSIDAD DEL SINÚ 'Elías Bechara Zainúm' - UNISINÚ-CARTAGENA" ~ 2,
    UNIVERSITY == "CORPORACION UNIVERSITARIA COMFACAUCA - UNICOMFACAUCA-POPAYAN" ~ 2,
    UNIVERSITY == "FUNDACION UNIVERSITARIA EMPRESARIAL DE LA CAMARA DE COMERCIO DE Bogotá -BOGOTÁ D.C."
    UNIVERSITY == "ESCUELA DE INGENIEROS MILITARES-BOGOTÁ D.C." ~ 1,
    UNIVERSITY == "FUNDACION UNIVERSITARIA CLARETIANA - UNICLARETIANA-QUIBDO" ~ 2,
    UNIVERSITY == "FUNDACION UNIVERSITARIA NAVARRA - UNINAVARRA-NEIVA" ~ 1,
    TRUE ~ 0  # Assign NA to any university not listed
  ))
```

```
data$UNIVERSITY <- as.factor(data$UNIVERSITY)
data$STRATUM <- as.factor(data$STRATUM)
data$PEOPLE_HOUSE <- as.factor(data$PEOPLE_HOUSE)
```

```
model1 <- lm(OVERALL_SCORE ~ GENDER + EDU_FATHER + EDU_MOTHER + OCC_FATHER + OCC_MOTHER + STRATUM + PEO
#summary(model1)
```

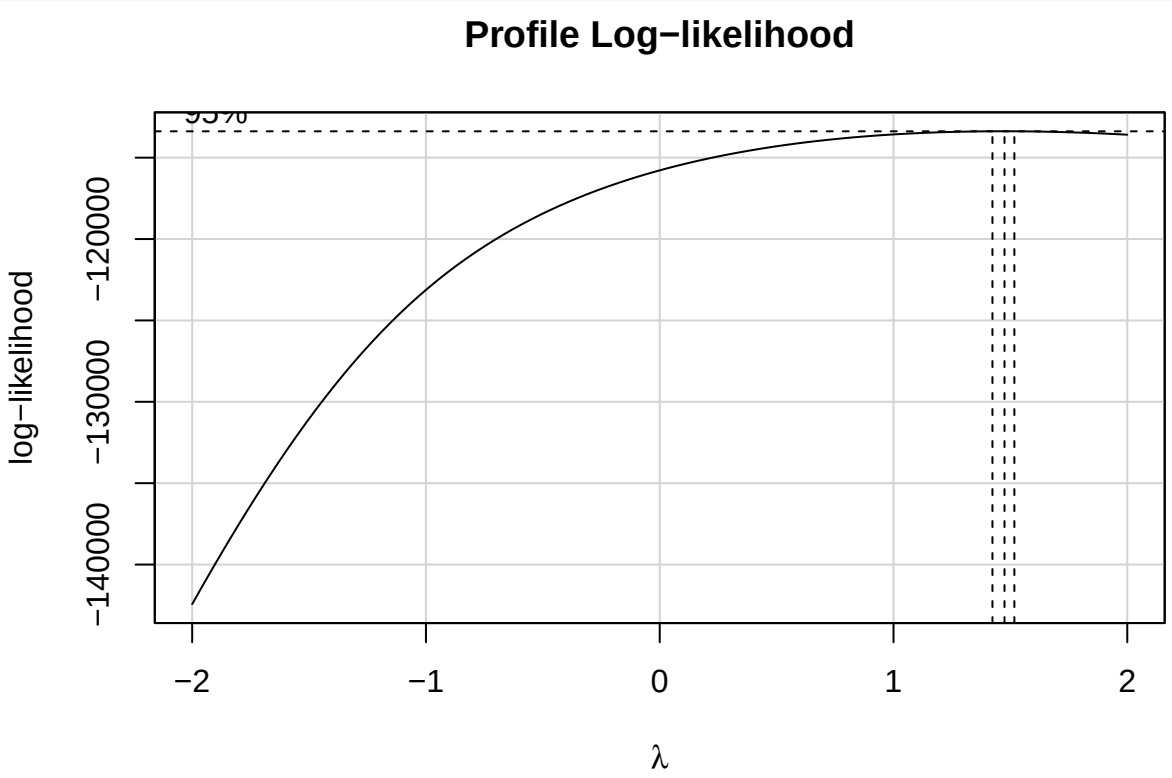
```
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
boxCox(model1)
```




```
p1 <- powerTransform(model1)
summary(p1)
```

```
## bcPower Transformation to Normality
##    Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## Y1    1.4685        1.47       1.4199       1.5171
##
## Likelihood ratio test that transformation parameter is equal to 0
##  (log transformation)
##                            LRT df       pval
## LR test, lambda = (0) 4801.293  1 < 2.22e-16
##
## Likelihood ratio test that no transformation is needed
##                            LRT df       pval
## LR test, lambda = (1) 388.5269  1 < 2.22e-16
```

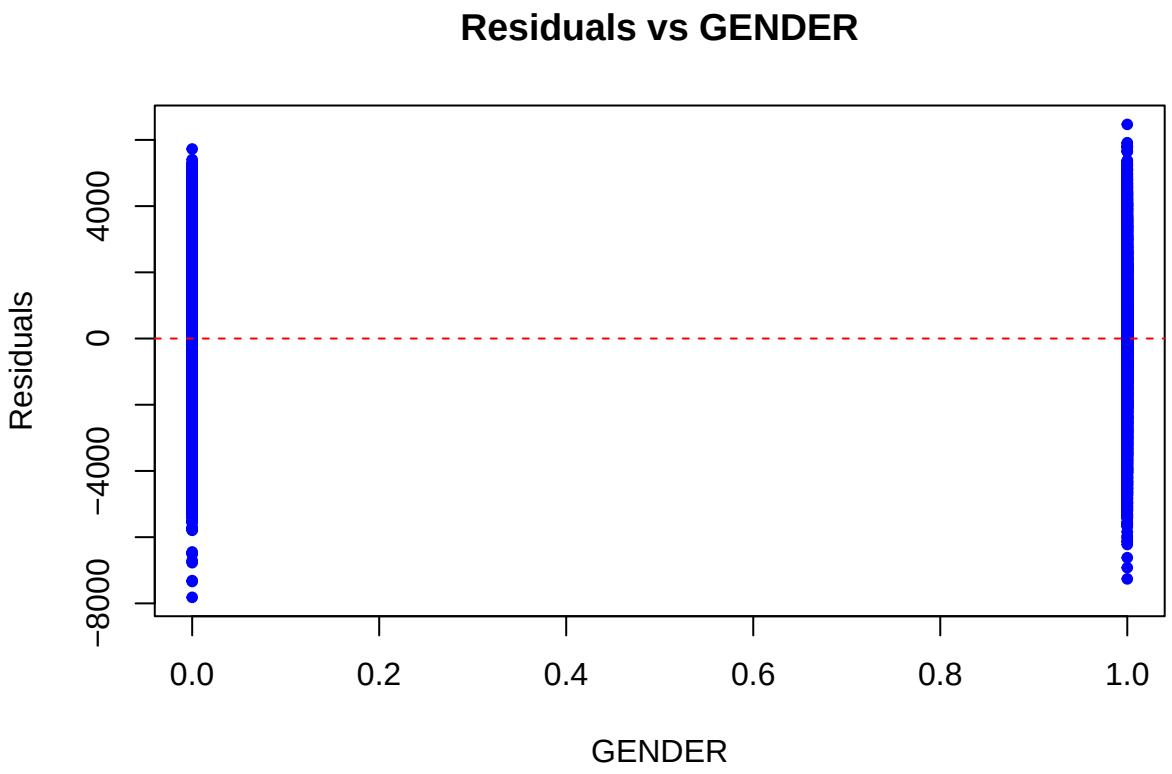
In the graph of Response variable "Overall_Score" has a left skew so by using powerTransform and boxCox function, we found that the optimal lambda is 1.46. We will round it to 1.5 for simplicity.

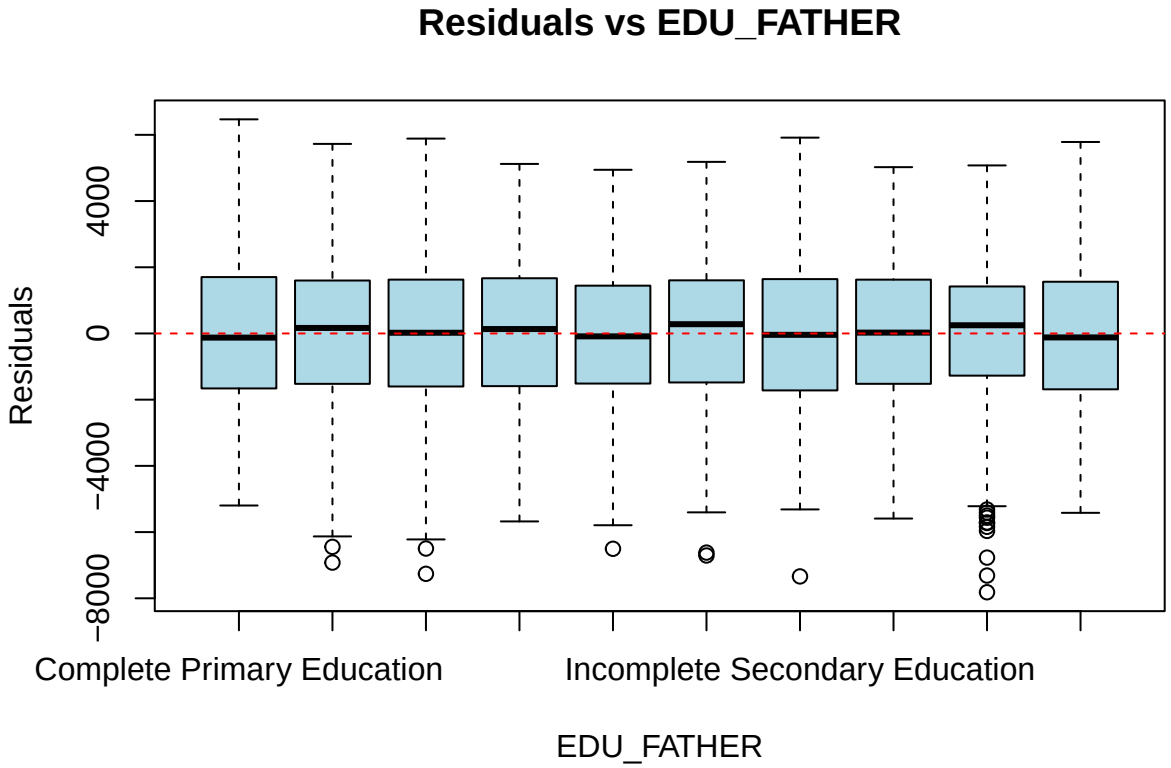
```
powerSCORE <- data$OVERALL_SCORE^1.5

model1 <- lm(powerSCORE ~ GENDER + EDU_FATHER + EDU_MOTHER + OCC_FATHER + OCC_MOTHER + STRATUM + PEOPLE_
# summary(model1)
# hist(powerSCORE)
# residuals <- residuals(model1)
# qqnorm(residuals, main = "Q-Q Plot of Residuals")
# qqline(residuals, col = "red", lwd = 2)

# y_value <- resid(model1)
# x_value <- fitted(model1)

# ggplot(data, aes(x = x_value, y = y_value)) +
#   geom_point(alpha = 0.5) +                    # Scatter plot of residuals
#   geom_smooth(method = "loess", color = "blue") + # LOESS smooth line to see trend
#   geom_hline(yintercept = 0, color = "red", linetype = "dashed") + # Line at zero
#   labs(title = "Residuals vs. Fitted",
#        x = "Fitted Values",
#        y = "Residuals")
```

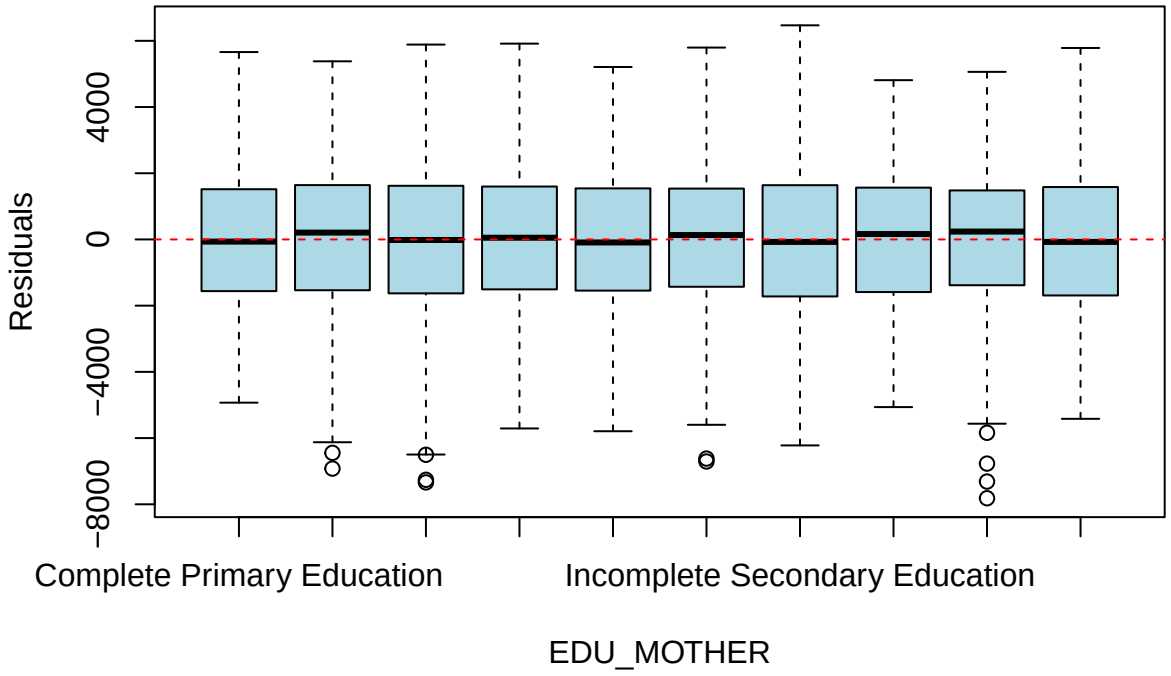
Based on above histogram, we can see a clean normal distribution and less tails for Q-Q plot.

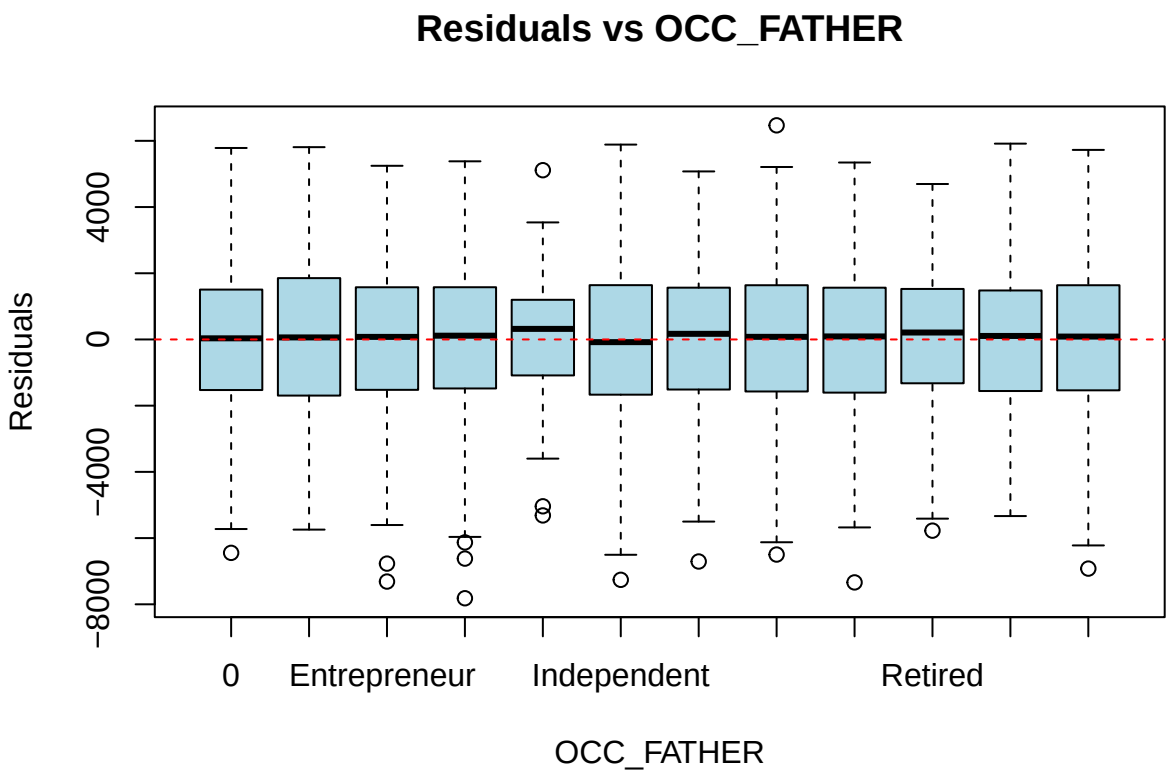
```
residuals <- residuals(model1)

# Extract the data used in the model (this excludes observations with missing values)
model_data <- model1$model

# Add residuals to the model data
model_data$residuals <- residuals

# Identify the response variable (first column in model data)
response_var <- names(model_data)[1]

# Get the list of predictors from the model data
predictors <- setdiff(names(model_data), c(response_var, "residuals"))

# Set up the plotting area
#par(mfrow = c(3, 4), mar = c(4, 4, 2, 1))  # Adjust rows and columns as needed

# Loop over each predictor and create a residual plot
```

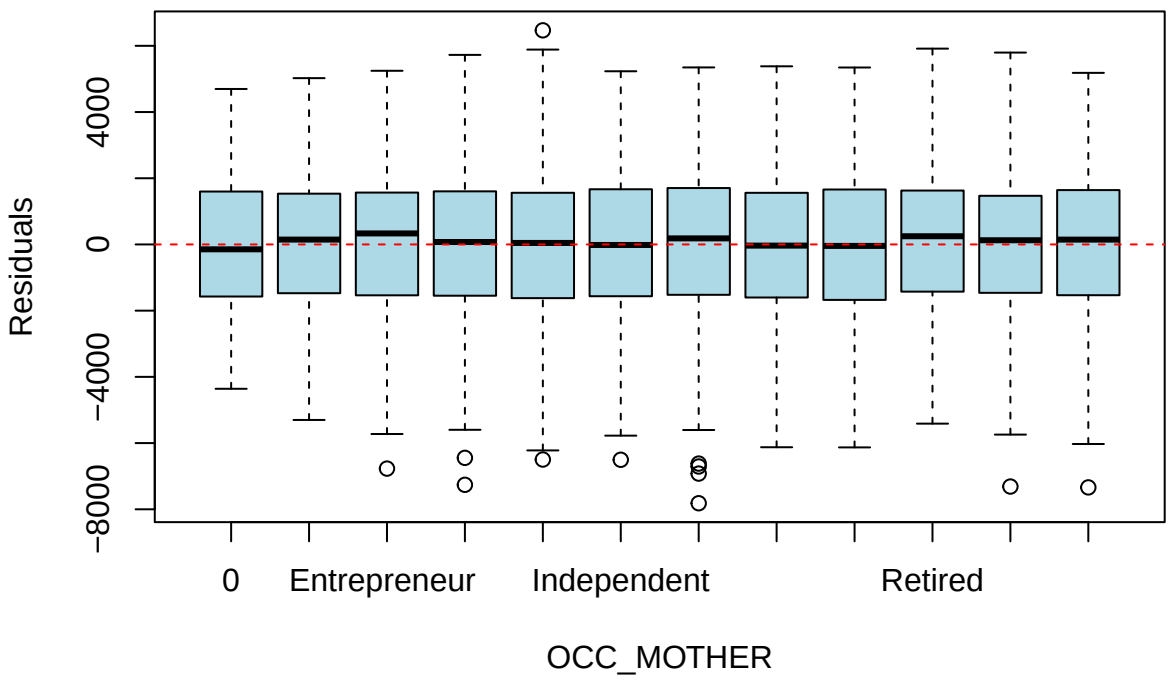
```r
for (pred in predictors) {
  predictor_data <- model_data[[pred]]

  # Check if predictor is categorical or numerical
  if (is.factor(predictor_data) || is.character(predictor_data)) {
    # Create a boxplot for categorical predictors
    boxplot(residuals ~ predictor_data,
            data = model_data,
            xlab = pred,
            ylab = "Residuals",
            main = paste("Residuals vs", pred),
            col = "lightblue")
    abline(h = 0, col = "red", lty = 2)
  } else {
    # Create a scatter plot for numerical predictors
    plot(predictor_data, model_data$residuals,
         xlab = pred,
         ylab = "Residuals",
         main = paste("Residuals vs", pred),
         pch = 20,
         col = "blue")
    abline(h = 0, col = "red", lty = 2)
  }
}
```

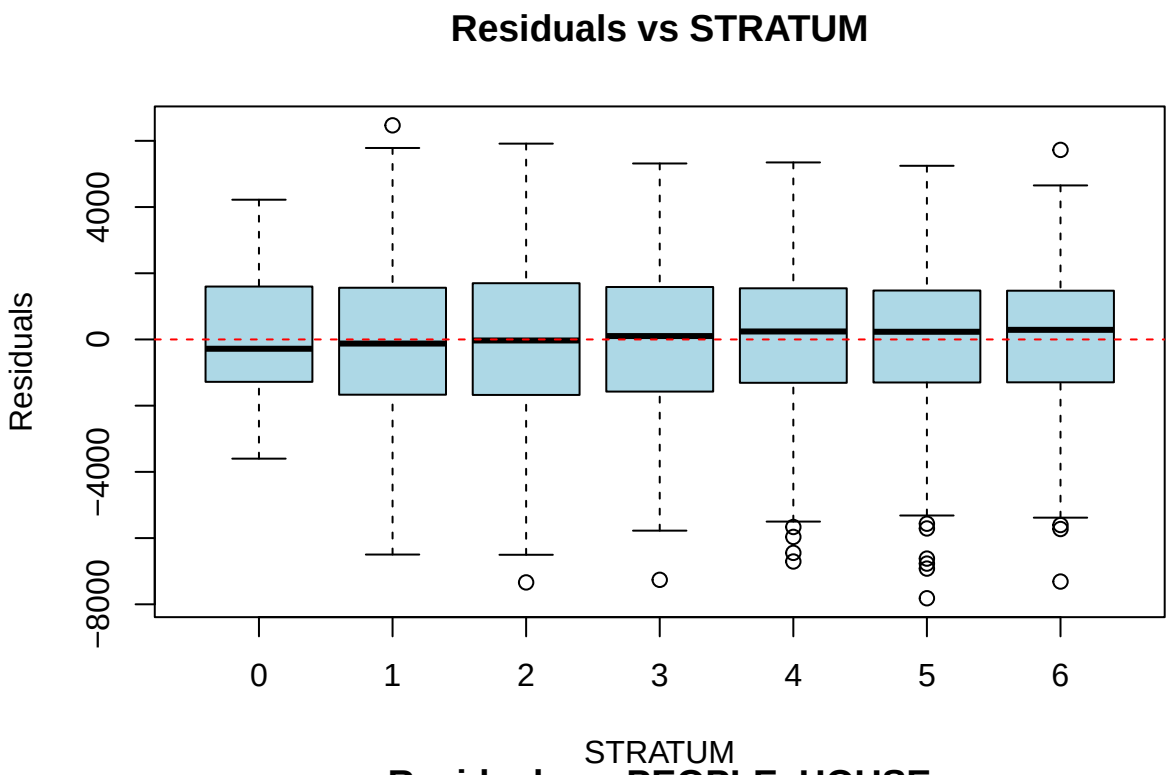


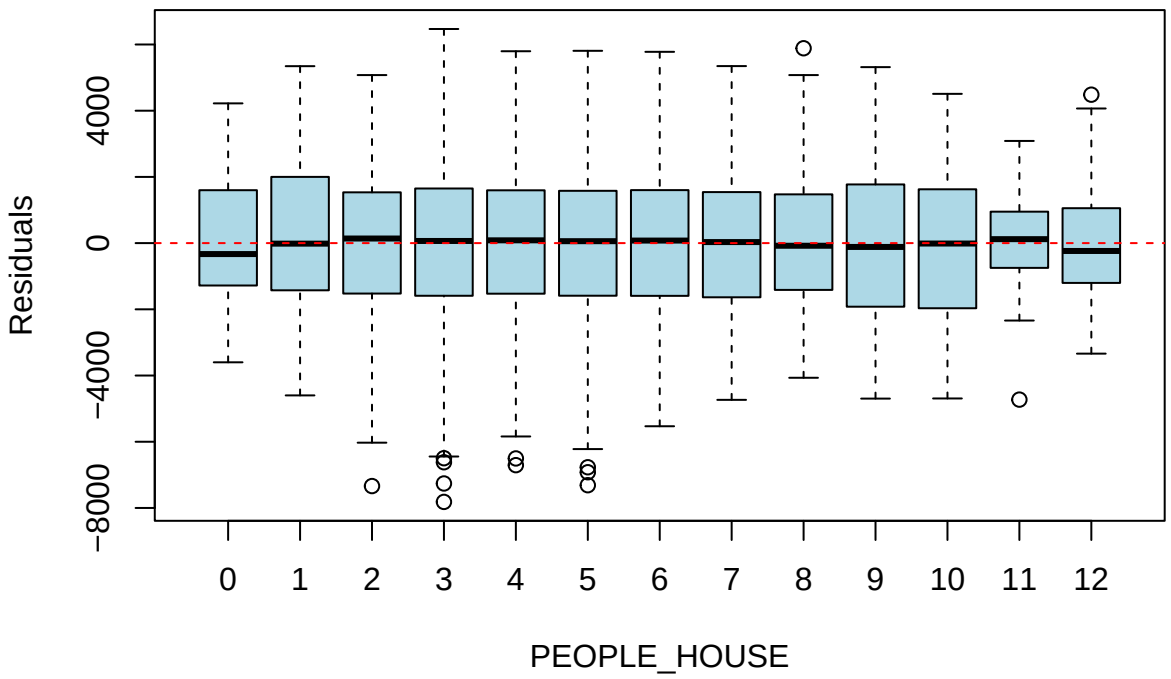

Residuals vs EDU_FATHER
Residuals
4000
0
−4000
−8000
Complete Primary Education
Incomplete Secondary Education
EDU_FATHER

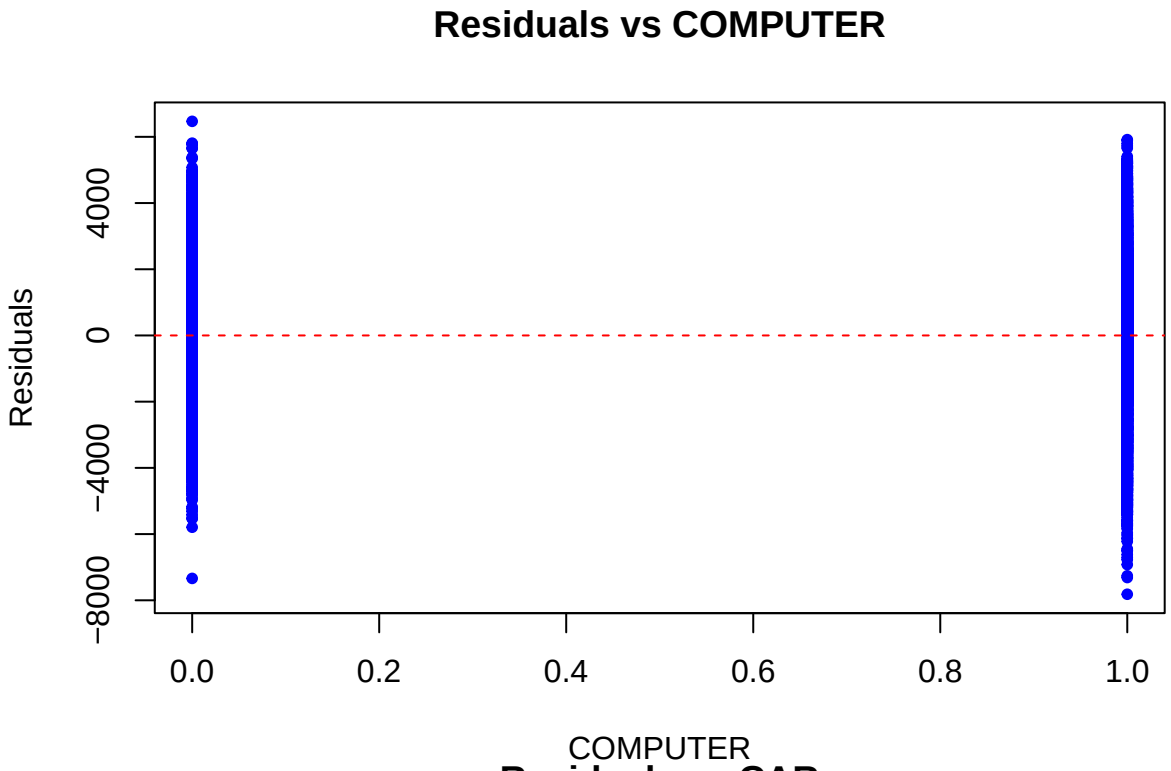

Residuals vs EDU_MOTHER
Residuals
4000
0
−4000
−8000
Complete Primary Education
Incomplete Secondary Education
EDU_MOTHER

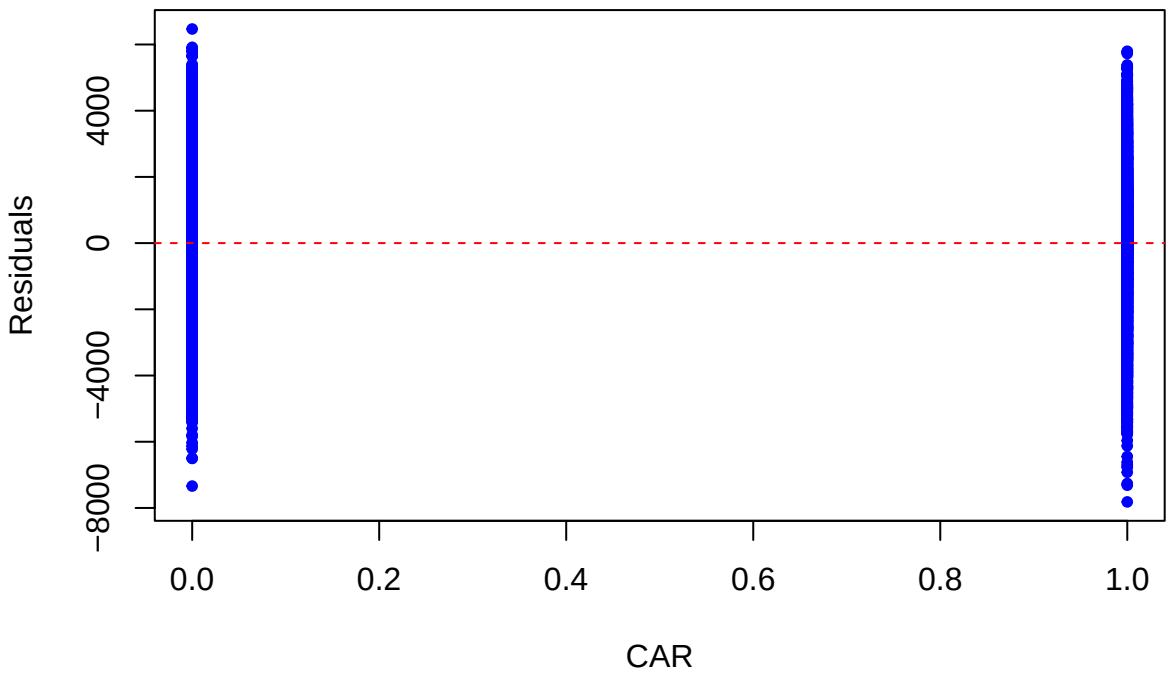

Residuals vs OCC_FATHER
Residuals
4000
0
−4000
−8000
0
Entrepreneur
Independent
Retired
OCC_FATHER

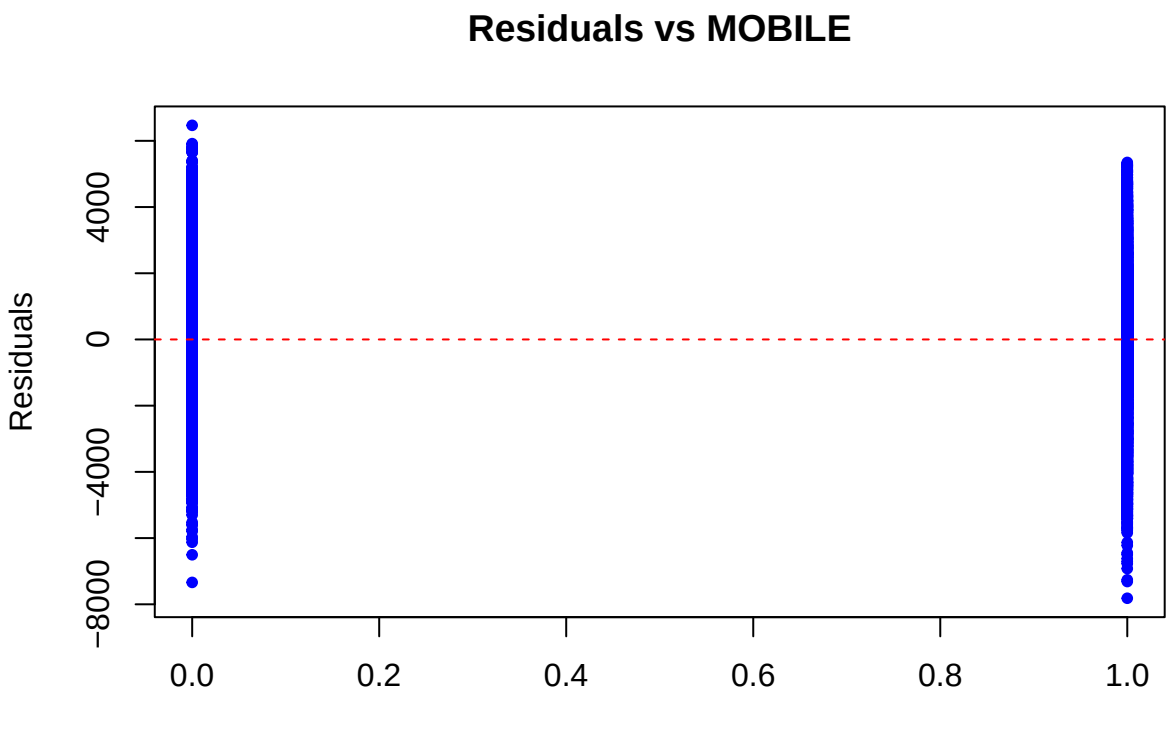

Residuals vs OCC_MOTHER
Residuals
4000
0
−4000
−8000
0
Entrepreneur
Independent
Retired
OCC_MOTHER

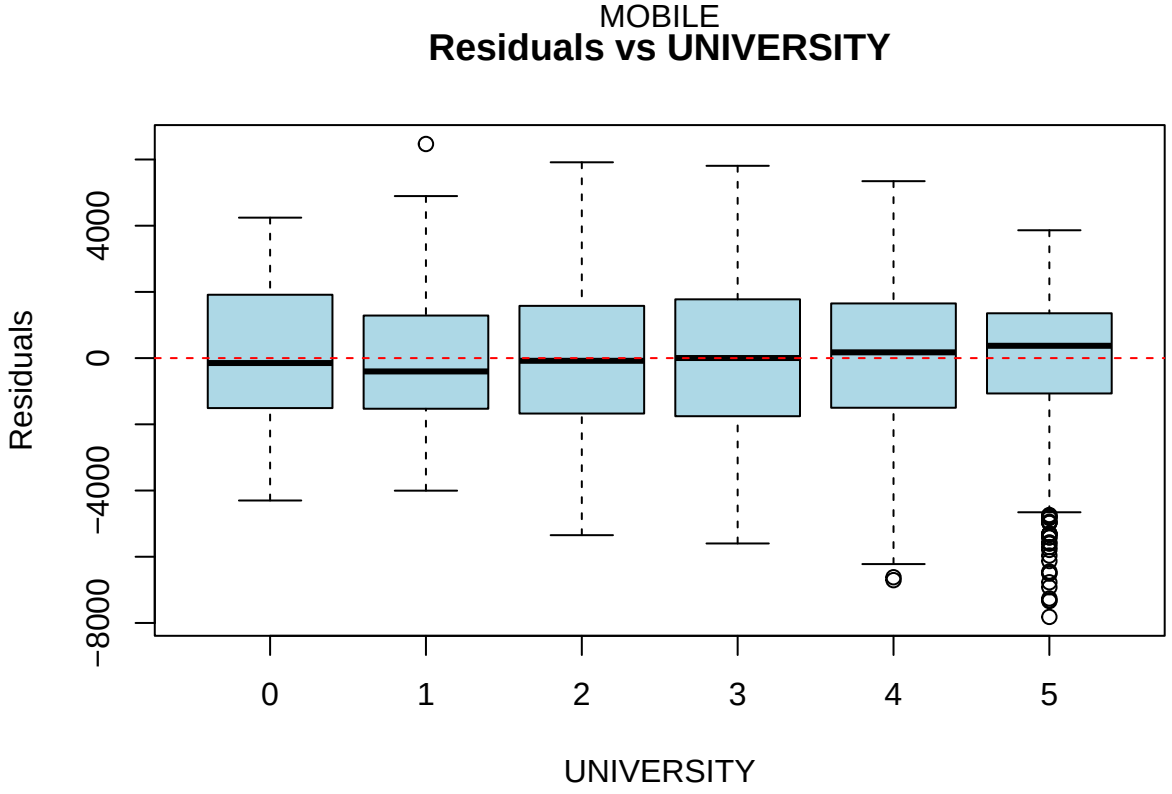

Residuals vs STRATUM
Residuals
4000
0
−4000
−8000
0
1
2
3
4
5
6
STRATUM


Residuals vs PEOPLE_HOUSE
Residuals
4000
0
−4000
−8000
0
1
2
3
4
5
6
7
8
9
10
11
12
PEOPLE_HOUSE


Residuals vs COMPUTER
Residuals
4000
0
−4000
−8000
0.0
0.2
0.4
0.6
0.8
1.0
COMPUTER


Residuals vs CAR
Residuals
4000
0
−4000
−8000
0.0
0.2
0.4
0.6
0.8
1.0
CAR





Conduct ANOVA Test

```
f_test <- qf(0.95, 67, 12343)
```

By Conducting ANOVA Test, We test with 95% F-statistics with 67 and 12343 DFs. We found that f_test

is 1.30, therefore 63.66 > 1.30. We rejct null and conclude significant linear relationship exists for at least one predictor.

Analyze T-tests

Estimate Std. Error t value Pr(>|t|)

(Intercept) 4098.833 699.631 5.859 4.79e-09 ***GENDER 144.373 39.978 3.611 0.000306*** EDU_FATHERComplete Professional Education 148.354 103.467 1.434 0.151646
EDU_FATHERComplete Secondary Education 9.219 92.477 0.100 0.920596
EDU_FATHERComplete Technical Education 265.841 110.336 2.409 0.015995 *
EDU_FATHERIncomplete Primary Education 4.443 114.941 0.039 0.969167
EDU_FATHERIncomplete Professional Education 431.772 139.023 3.106 0.001902 ** EDU_FATHERIncomplete Secondary Education -58.386 104.140 -0.561 0.575048
EDU_FATHERIncomplete Technical Education 273.722 159.527 1.716 0.086217 .
EDU_FATHERPostgraduate Education 553.392 123.824 4.469 7.92e-06 ***EDU_FATHERUnknown -56.710 125.693 -0.451 0.651868***
EDU_MOTHERComplete Professional Education 347.647 110.163 3.156 0.001605
EDU_MOTHERComplete Secondary Education 20.454 97.686 0.209 0.834147
EDU_MOTHERComplete Technical Education 194.642 111.768 1.741 0.081624 .
EDU_MOTHERIncomplete Primary Education -114.569 129.518 -0.885 0.376400
EDU_MOTHERIncomplete Professional Education 356.371 138.219 2.578 0.009940 ***EDU_MOTHERIncomplete Secondary Education -76.486 109.513 -0.698 0.484928***
EDU_MOTHERIncomplete Technical Education 167.576 153.004 1.095 0.273431
EDU_MOTHERPostgraduate Education 615.088 131.133 4.691 2.75e-06 EDU_MOTHERUnknown -120.955 149.191 -0.811 0.417533
OCC_FATHERAuxiliary or Administrative 83.296 149.507 0.557 0.577444
OCC_FATHEREntrepreneur -65.845 140.912 -0.467 0.640308
OCC_FATHERExecutive 171.707 115.854 1.482 0.138341
OCC_FATHERHome 19.141 266.464 0.072 0.942737
OCC_FATHERIndependent 110.334 107.228 1.029 0.303516
OCC_FATHERIndependent professional 159.800 119.281 1.340 0.180371
OCC_FATHEROperator 135.958 115.253 1.180 0.238162
OCC_FATHEROther occupation 225.674 118.393 1.906 0.056654 .
OCC_FATHERRetired 323.250 134.939 2.396 0.016611 *
OCC_FATHERSmall entrepreneur -66.323 130.454 -0.508 0.611182
OCC_FATHERTechnical or professional level employee 96.850 110.189 0.879 0.379448
OCC_MOTHERAuxiliary or Administrative 32.721 180.202 0.182 0.855916
OCC_MOTHEREntrepreneur -198.946 220.379 -0.903 0.366678
OCC_MOTHERExecutive -143.428 182.542 -0.786 0.432041
OCC_MOTHERHome 10.455 160.864 0.065 0.948180
OCC_MOTHERIndependent -128.531 174.650 -0.736 0.461784
OCC_MOTHERIndependent professional -74.995 183.304 -0.409 0.682454
OCC_MOTHEROperator 71.632 183.357 0.391 0.696049
OCC_MOTHEROther occupation -226.100 185.527 -1.219 0.222985
OCC_MOTHERRetired 337.552 238.983 1.412 0.157843
OCC_MOTHERSmall entrepreneur 28.873 192.776 0.150 0.880944
OCC_MOTHERTechnical or professional level employee -35.702 174.300 -0.205 0.837710
STRATUM1 77.636 760.267 0.102 0.918666
STRATUM2 458.423 759.046 0.604 0.545890
STRATUM3 706.136 758.809 0.931 0.352087
STRATUM4 993.144 760.569 1.306 0.191648
STRATUM5 1011.501 764.057 1.324 0.185576
STRATUM6 1211.438 768.141 1.577 0.114797

PEOPLE_HOUSE1 -907.447 876.598 -1.035 0.300600
PEOPLE_HOUSE2 -577.771 640.559 -0.902 0.367085
PEOPLE_HOUSE3 -827.336 636.858 -1.299 0.193937
PEOPLE_HOUSE4 -815.136 636.336 -1.281 0.200223
PEOPLE_HOUSE5 -770.964 636.882 -1.211 0.226100
PEOPLE_HOUSE6 -848.956 638.846 -1.329 0.183909
PEOPLE_HOUSE7 -869.699 646.421 -1.345 0.178519
PEOPLE_HOUSE8 -1293.297 657.895 -1.966 0.049343 *
PEOPLE_HOUSE9 -981.047 685.073 -1.432 0.152160
PEOPLE_HOUSE10 -768.875 703.353 -1.093 0.274346
PEOPLE_HOUSE11 -451.045 808.321 -0.558 0.576853
PEOPLE_HOUSE12 -891.910 744.536 -1.198 0.230963
COMPUTER 75.185 62.674 1.200 0.230304
CAR -75.387 47.139 -1.599 0.109793
MOBILE 362.839 49.353 7.352 2.08e-13 ***UNIVERSITY1 -484.925 353.468 -1.372 0.170118***
UNIVERSITY2 475.028 334.347 1.421 0.155411
UNIVERSITY3 950.475 335.251 2.835 0.004588 *UNIVERSITY4 1870.271 334.741 5.587 2.36e-08*
UNIVERSITY5 3303.381 339.151 9.740 < 2e-16 *

#Analysis

Based on the summary of the model, we can see that OCC_FATHER and OCC_MOTHER were neither significance with OVERALL_SCORE. Let's compare using partial-F Test.

OCC_FATHER and OCC_MOTHER

```
model_no_OCC <- lm(powerSCORE ~ GENDER + EDU_FATHER + EDU_MOTHER + STRATUM + PEOPLE_HOUSE + COMPUTER + C
#summary(model_no_OCC)

#anova(model_no_OCC, model1)
```

```
model1_df <- 12343
model_no_OCC_df <- 12365
occ_test <- qf(0.95, model_no_OCC_df - model1_df, model1_df)
```

in this test we left null hypothesis that OCC_FATHER and OCC_MOTHER = 0 and alternative hypothesis that at least one not equal to 0.

By conduct partial F test, we got F = 1.721. By comparing with F-test, we found that 1.721 > 1.543. We reject the null hypothesis. Which means there's some significant linear relationship exists between OVERALL_SCORE and either OCC_FATHER and OCC_MOTHER so we will keep the full model.

STRATUM

```
model_no_STRATUM <- lm(powerSCORE ~ GENDER + EDU_FATHER + EDU_MOTHER + OCC_FATHER + OCC_MOTHER + PEOPLE_
#summary(model_no_STRATUM)

#anova(model_no_STRATUM, model1)
```

```
model1_df <- 12343
model_no_STRATUM_df <- 12349
STRATUM_test <- qf(0.95, model_no_STRATUM_df - model1_df, model1_df)
```

in this test we left null hypothesis that STRATUM = 0 and alternative hypothesis that STRATUM is not equal to 0.

By conduct partial F test, we got F = 21.153. By comparing with F-test, we found that 21.153 > 2.0993. We reject the null hypothesis. Which means there's significant linear relationship exists between OVERALL_SCORE and STRATUM so we will keep the full model.

PEOPLE_HOUSE

```
model_no_PEOPLE <- lm(powerSCORE ~ GENDER + EDU_FATHER + EDU_MOTHER + OCC_FATHER + OCC_MOTHER + STRATUM
#summary(model_no_PEOPLE)

#anova(model_no_PEOPLE, model1)
```

```
model1_df <- 12343
model_no_PEOPLE_df <- 12355
PEOPLE_test <- qf(0.95, model_no_PEOPLE_df - model1_df, model1_df)
```

in this test we left null hypothesis that PEOPLE_HOUSE = 0 and alternative hypothesis that PEOPLE_HOUSE is not equal to 0.

By conduct partial F test, we got F = 1.5755. By comparing with F-test, we found that 1.5755 < 1.7529. We fail to reject the null hypothesis. Which means there's no significant linear relationship exists between PEOPLE_HOUSE so we will keep the model_no_PEOPLE.

COMPUTER and CAR

```
model_no_COM_CAR <- lm(powerSCORE ~ GENDER + EDU_FATHER + EDU_MOTHER + OCC_FATHER + OCC_MOTHER + STRATU
#summary(model_no_COM_CAR)

#anova(model_no_COM_CAR, model1)
```

```
model1_df <- 12343
model_no_COM_CAR_df <- 12357
COM_CAR_test <- qf(0.95, model_no_COM_CAR_df - model1_df, model1_df)
```

in this test we left null hypothesis that COMPUTER and CAR = 0 and alternative hypothesis that COMPUTER or CAR is not equal to 0.

By conduct partial F test, we got F = 1.6365. By comparing with F-test, we found that 1.6365 < 1.6926 We fail to reject the null hypothesis. Which means there's no significant linear relationship exists between COMPUTER and CAR so we will keep the model_no_COM_CAR.

Further analysis for OCC_MOTHER

```
model_no_MOTHER <- lm(powerSCORE ~ GENDER + EDU_FATHER + EDU_MOTHER + OCC_FATHER + STRATUM + MOBILE + U
#summary(model_no_MOTHER)

#anova(model_no_MOTHER, model1)
```

```
model1_df <- 12343
model_no_MOTHER_df <- 12368
MOTHER_test <- qf(0.95, model_no_MOTHER_df - model1_df, model1_df)
```

in this test we left null hypothesis that OCC_MOTHER = 0 and alternative hypothesis that is not equal to 0.

By conduct partial F test, we got F = 1.6827. By comparing with F-test, we found that 1.6827 > 1.5070. We reject the null hypothesis. Which means there's some significant linear relationship exists between

OVERALL_SCORE and OCC_MOTHER so we will keep the model_no_COM_CAR.

Final Check

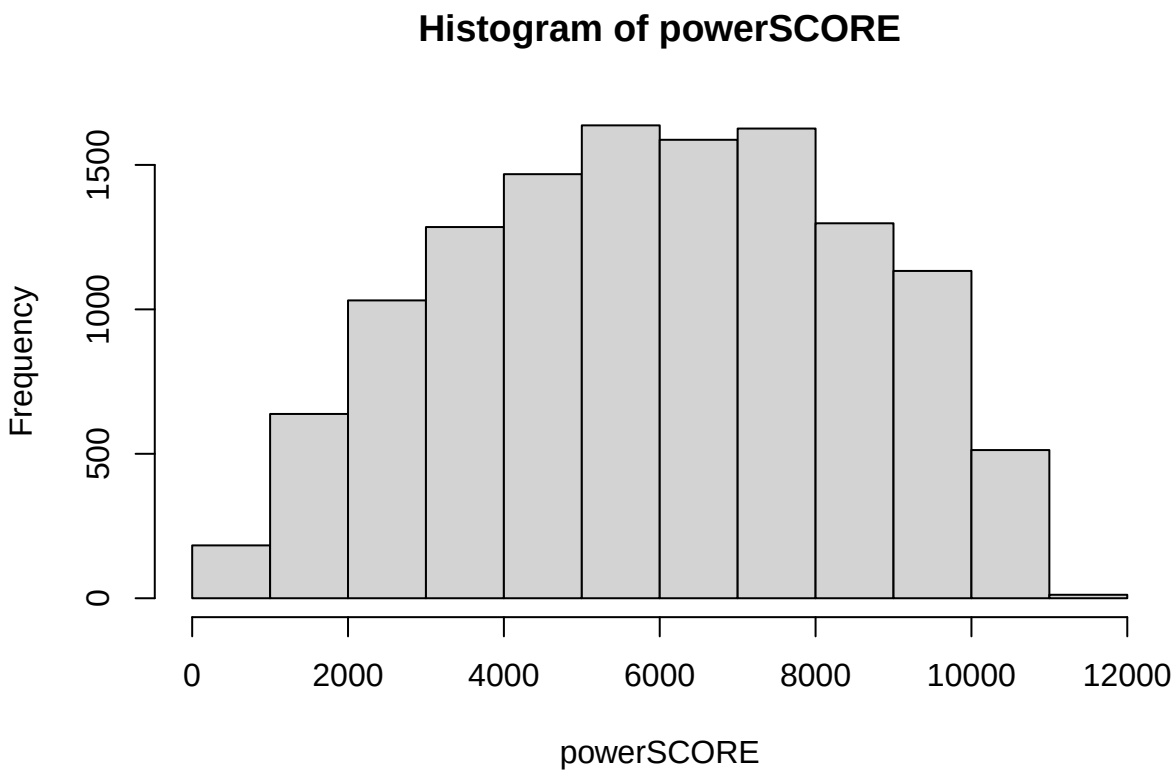
```
powerSCORE <- data$OVERALL_SCORE^1.5

model1 <- lm(powerSCORE ~ GENDER + EDU_FATHER + EDU_MOTHER + OCC_FATHER + OCC_MOTHER + STRATUM + MOBILE
summary(model1)
```

```
##
## Call:
## lm(formula = powerSCORE ~ GENDER + EDU_FATHER + EDU_MOTHER +
##     OCC_FATHER + OCC_MOTHER + STRATUM + MOBILE + UNIVERSITY,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7847.0 -1564.5    63.2  1597.1  6433.7
##
## Coefficients:
##                                                  Estimate Std. Error t value
## (Intercept)                                      3919.113    685.681   5.716
## GENDER                                            144.748     39.974   3.621
## EDU_FATHERComplete Professional Education         156.213    103.304   1.512
## EDU_FATHERComplete Secondary Education             17.011     92.352   0.184
## EDU_FATHERComplete Technical Education            271.250    110.121   2.463
## EDU_FATHERIncomplete Primary Education              3.420    114.884   0.030
## EDU_FATHERIncomplete Professional Education       438.104    138.947   3.153
## EDU_FATHERIncomplete Secondary Education          -51.023    104.045  -0.490
## EDU_FATHERIncomplete Technical Education          286.504    159.376   1.798
## EDU_FATHERPostgraduate Education                  561.578    123.647   4.542
## EDU_FATHERUnknown                                 -40.622    125.621  -0.323
## EDU_MOTHERComplete Professional Education         356.720    109.914   3.245
## EDU_MOTHERComplete Secondary Education             33.903     97.551   0.348
## EDU_MOTHERComplete Technical Education            208.489    111.627   1.868
## EDU_MOTHERIncomplete Primary Education           -119.405    129.311  -0.923
## EDU_MOTHERIncomplete Professional Education       367.304    138.090   2.660
## EDU_MOTHERIncomplete Secondary Education          -72.108    109.458  -0.659
## EDU_MOTHERIncomplete Technical Education          186.343    152.814   1.219
## EDU_MOTHERPostgraduate Education                  625.376    130.744   4.783
## EDU_MOTHERUnknown                                -123.522    148.454  -0.832
## OCC_FATHERAuxiliary or Administrative             108.590    149.062   0.728
## OCC_FATHEREntrepreneur                            -48.189    140.368  -0.343
## OCC_FATHERExecutive                               185.125    115.219   1.607
## OCC_FATHERHome                                     45.551    266.123   0.171
## OCC_FATHERIndependent                             128.152    106.620   1.202
## OCC_FATHERIndependent professional                169.317    118.836   1.425
## OCC_FATHEROperator                                161.063    114.578   1.406
## OCC_FATHEROther occupation                        255.109    117.731   2.167
## OCC_FATHERRetired                                 349.290    134.277   2.601
## OCC_FATHERSmall entrepreneur                      -50.275    129.706  -0.388
## OCC_FATHERTechnical or professional level employee 115.346    109.484   1.054
## OCC_MOTHERAuxiliary or Administrative              15.573    175.900   0.089
```

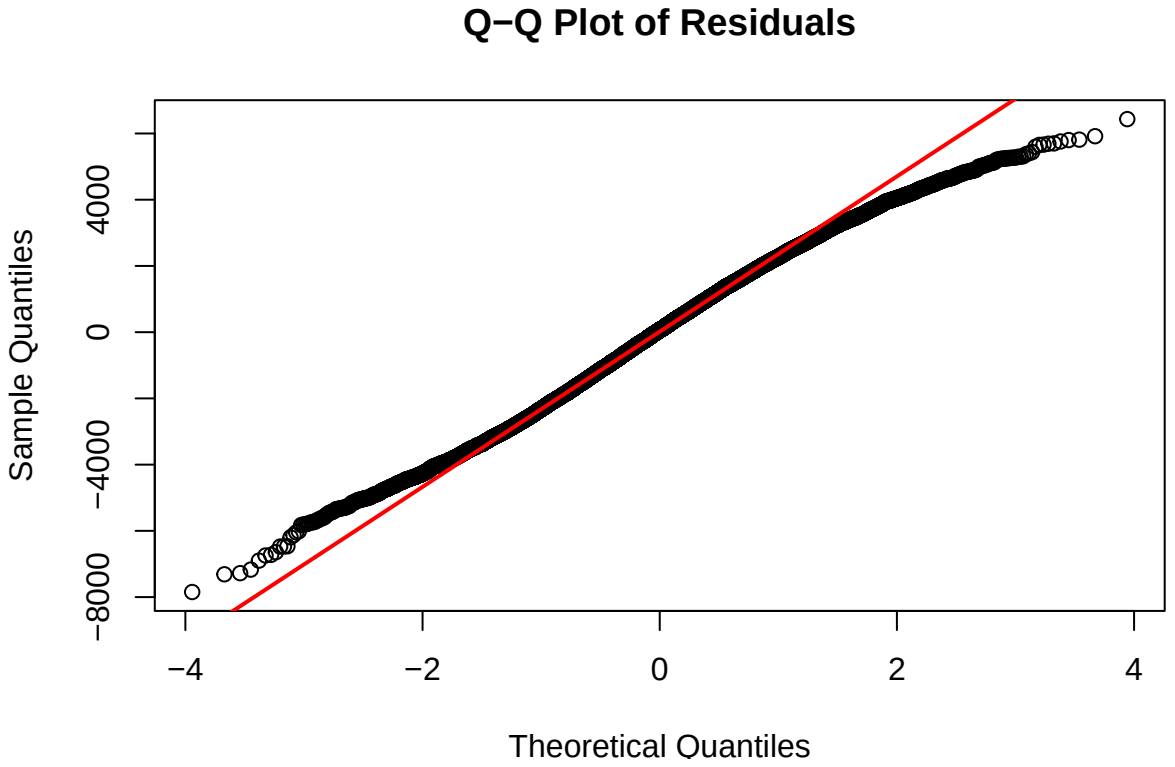
```
## OCC_MOTHEREntrepreneur                                -225.020    217.053  -1.037
## OCC_MOTHERExecutive                                   -170.663    178.419  -0.957
## OCC_MOTHERHome                                         -16.735    156.804  -0.107
## OCC_MOTHERIndependent                                 -150.078    170.468  -0.880
## OCC_MOTHERIndependent professional                     -96.321    179.269  -0.537
## OCC_MOTHEROperator                                      56.859    179.602   0.317
## OCC_MOTHEROther occupation                            -253.509    181.608  -1.396
## OCC_MOTHERRetired                                      326.669    235.675   1.386
## OCC_MOTHERSmall entrepreneur                            -2.105    189.062  -0.011
## OCC_MOTHERTechnical or professional level employee     -52.032    170.079  -0.306
## STRATUM1                                              -536.320    593.819  -0.903
## STRATUM2                                              -148.902    592.455  -0.251
## STRATUM3                                                96.982    592.826   0.164
## STRATUM4                                               373.898    595.311   0.628
## STRATUM5                                               382.766    600.137   0.638
## STRATUM6                                               579.205    604.994   0.957
## MOBILE                                                 357.063     48.454   7.369
## UNIVERSITY1                                           -487.566    353.330  -1.380
## UNIVERSITY2                                            480.545    334.266   1.438
## UNIVERSITY3                                            954.432    335.161   2.848
## UNIVERSITY4                                           1872.076    334.639   5.594
## UNIVERSITY5                                           3309.881    339.073   9.762
##                                                      Pr(>|t|)
## (Intercept)                                          1.12e-08 ***
## GENDER                                               0.000295 ***
## EDU_FATHERComplete Professional Education            0.130515
## EDU_FATHERComplete Secondary Education               0.853863
## EDU_FATHERComplete Technical Education               0.013784 *
## EDU_FATHERIncomplete Primary Education               0.976252
## EDU_FATHERIncomplete Professional Education          0.001620 **
## EDU_FATHERIncomplete Secondary Education             0.623868
## EDU_FATHERIncomplete Technical Education             0.072255 .
## EDU_FATHERPostgraduate Education                     5.63e-06 ***
## EDU_FATHERUnknown                                    0.746419
## EDU_MOTHERComplete Professional Education            0.001176 **
## EDU_MOTHERComplete Secondary Education               0.728192
## EDU_MOTHERComplete Technical Education               0.061823 .
## EDU_MOTHERIncomplete Primary Education               0.355820
## EDU_MOTHERIncomplete Professional Education          0.007827 **
## EDU_MOTHERIncomplete Secondary Education             0.510056
## EDU_MOTHERIncomplete Technical Education             0.222711
## EDU_MOTHERPostgraduate Education                     1.74e-06 ***
## EDU_MOTHERUnknown                                    0.405393
## OCC_FATHERAuxiliary or Administrative                0.466329
## OCC_FATHEREntrepreneur                               0.731376
## OCC_FATHERExecutive                                  0.108141
## OCC_FATHERHome                                       0.864097
## OCC_FATHERIndependent                                0.229404
## OCC_FATHERIndependent professional                   0.154241
## OCC_FATHEROperator                                   0.159836
## OCC_FATHEROther occupation                           0.030263 *
## OCC_FATHERRetired                                    0.009299 **
## OCC_FATHERSmall entrepreneur                         0.698314
## OCC_FATHERTechnical or professional level employee   0.292110
```

```
## OCC_MOTHERAuxiliary or Administrative              0.929455
## OCC_MOTHEREntrepreneur                             0.299894
## OCC_MOTHERExecutive                                0.338826
## OCC_MOTHERHome                                     0.915009
## OCC_MOTHERIndependent                              0.378667
## OCC_MOTHERIndependent professional                 0.591071
## OCC_MOTHEROperator                                 0.751564
## OCC_MOTHEROther occupation                         0.162765
## OCC_MOTHERRetired                                  0.165741
## OCC_MOTHERSmall entrepreneur                       0.991117
## OCC_MOTHERTechnical or professional level employee 0.759666
## STRATUM1                                           0.366453
## STRATUM2                                           0.801563
## STRATUM3                                           0.870054
## STRATUM4                                           0.529969
## STRATUM5                                           0.523617
## STRATUM6                                           0.338398
## MOBILE                                             1.83e-13 ***
## UNIVERSITY1                                        0.167638
## UNIVERSITY2                                        0.150569
## UNIVERSITY3                                        0.004411 **
## UNIVERSITY4                                        2.26e-08 ***
## UNIVERSITY5                                         < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2176 on 12357 degrees of freedom
## Multiple R-squared:  0.2554, Adjusted R-squared:  0.2522
## F-statistic: 79.99 on 53 and 12357 DF,  p-value: < 2.2e-16
```

```
hist(powerSCORE)
```



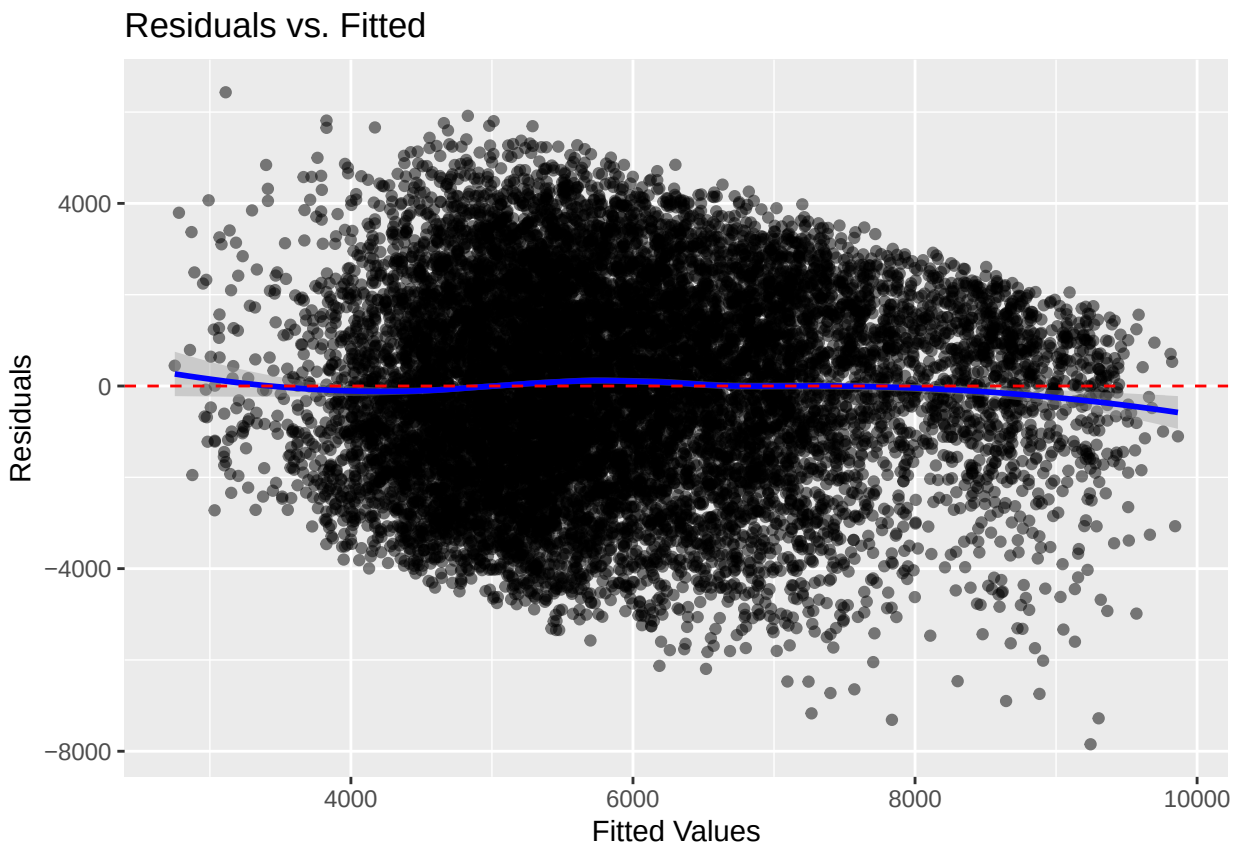

```r
residuals <- residuals(model1)
qqnorm(residuals, main = "Q-Q Plot of Residuals")
qqline(residuals, col = "red", lwd = 2)
```




```r
y_value <- resid(model1)
```

```r
x_value <- fitted(model1)

ggplot(data, aes(x = x_value, y = y_value)) +
  geom_point(alpha = 0.5) +                    # Scatter plot of residuals
  geom_smooth(method = "loess", color = "blue") + # LOESS smooth line to see trend
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") + # Line at zero
  labs(title = "Residuals vs. Fitted",
       x = "Fitted Values",
       y = "Residuals")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Verifying Multicollinearity and Goodness

```r
#cor(data[, c(1,9:11)])
vif1 <- 1/(1 - 0.2554)
print(vif1)
```

```
## [1] 1.343003
```

```r
vif(model1)
```

```
##                GVIF Df GVIF^(1/(2*Df))
## GENDER     1.010637  1        1.005304
## EDU_FATHER 7.629199  9        1.119506
## EDU_MOTHER 8.655672  9        1.127385
## OCC_FATHER 5.473231 11        1.080330
```

```
## OCC_MOTHER 4.913772 11        1.075048
## STRATUM    2.112027  6        1.064286
## MOBILE     1.260082  1        1.122534
## UNIVERSITY 1.226190  5        1.020601
```

Based on the VIF, we can see that all predictors have VIF < 2 which means there's not much big concern in multicollinearity.

Let's check for the Goodness

Outliers and Leverage Observations

```
# Calculate hat values
hat_values <- hatvalues(model1)
threshold_leverage <- 2 * (length(coef(model1)) / nrow(data))

lev_val <- which(hat_values > threshold_leverage)
```

This shows that there's lot of leverage observation that exceeds the threshold

```
# Calculate standardized residuals
standardized_residuals <- rstandard(model1)

# Identify outliers (standardized residuals > |2|)
outliers <- which(abs(standardized_residuals) > 2)
```

There's lot of outliers. To check that if there's any influence to the significance, we going to look at cooks distance.

```
di <- cooks.distance(model1)

p <- 53  # Number of predictors
n <- 12411  # Total number of observations

cutoff_di <- qf(0.5, p+1, n-p-1)
influencial_points <- which(di > cutoff_di)
```

Seems like it won't affect on the significance

Let's move on if there's any influence on own fitted values

```
dffits <- dffits(model1)
cutoff_dffits <-  2/sqrt(   (length(coef(model1)) / nrow(data))   )
which(abs(dffits) > cutoff_dffits)
```

```
## named integer(0)
```

It seems like there's no influence on own fitted values.

Let's check influential on coefficients.

```
dfbetas <- dfbetas(model1)
dim(dfbetas)
```

```
## [1] 12411    54
```

```
cutoff_dfbetas <- 2/sqrt(nrow(data))
```

```r
dfbeta_v_val_left <- which(abs(dfbetas[,1])>cutoff_dfbetas)
dfbeta_v_val_right <- which(abs(dfbetas[,2])>cutoff_dfbetas)
```

There's lot of disproportionate influence on the regression coefficient.

By inspecting the outliers, I found that there's high leverage outliers and high residuals outliers. To avoid deleting the high leverage outliers we're going to discard that is actually affecting our data to high noise.

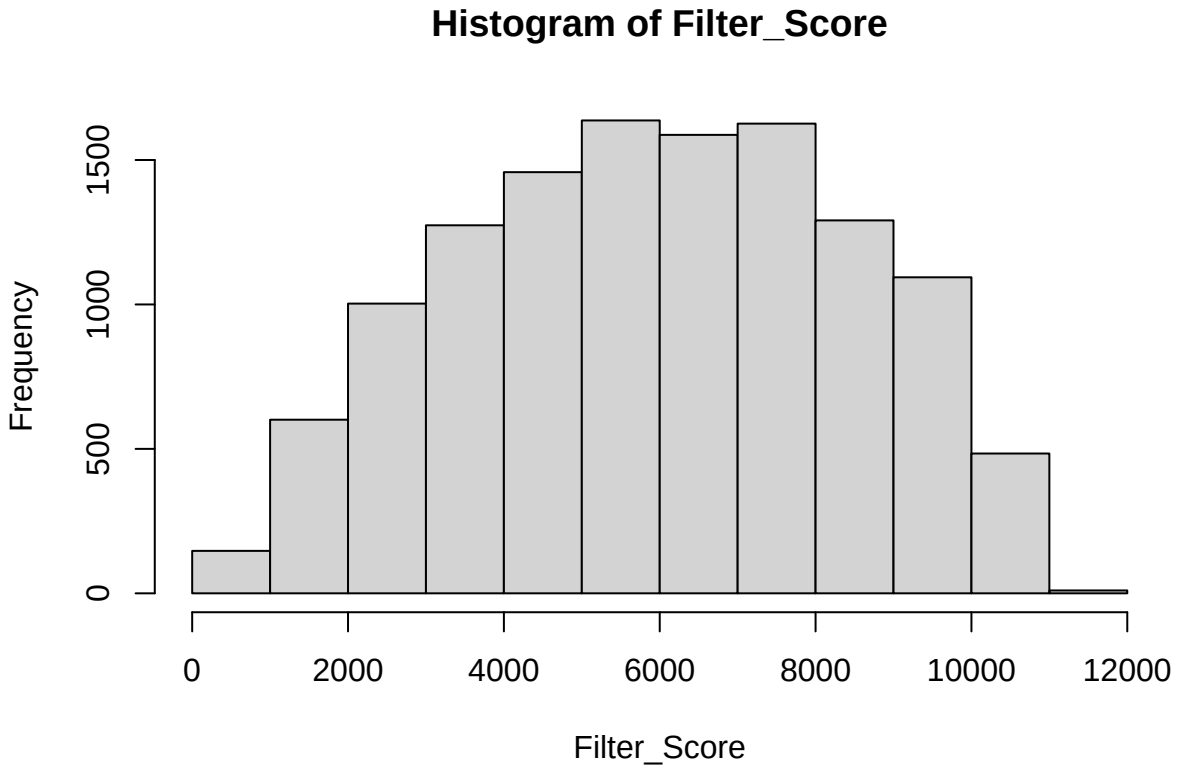
```r
filtered_large_residuals <- setdiff(
  outliers,
  intersect(outliers, c(lev_val, dfbeta_v_val_left, dfbeta_v_val_right))
)
```

From what we found, let's start discarding the any leverage observation, outliers and disproportionate values.

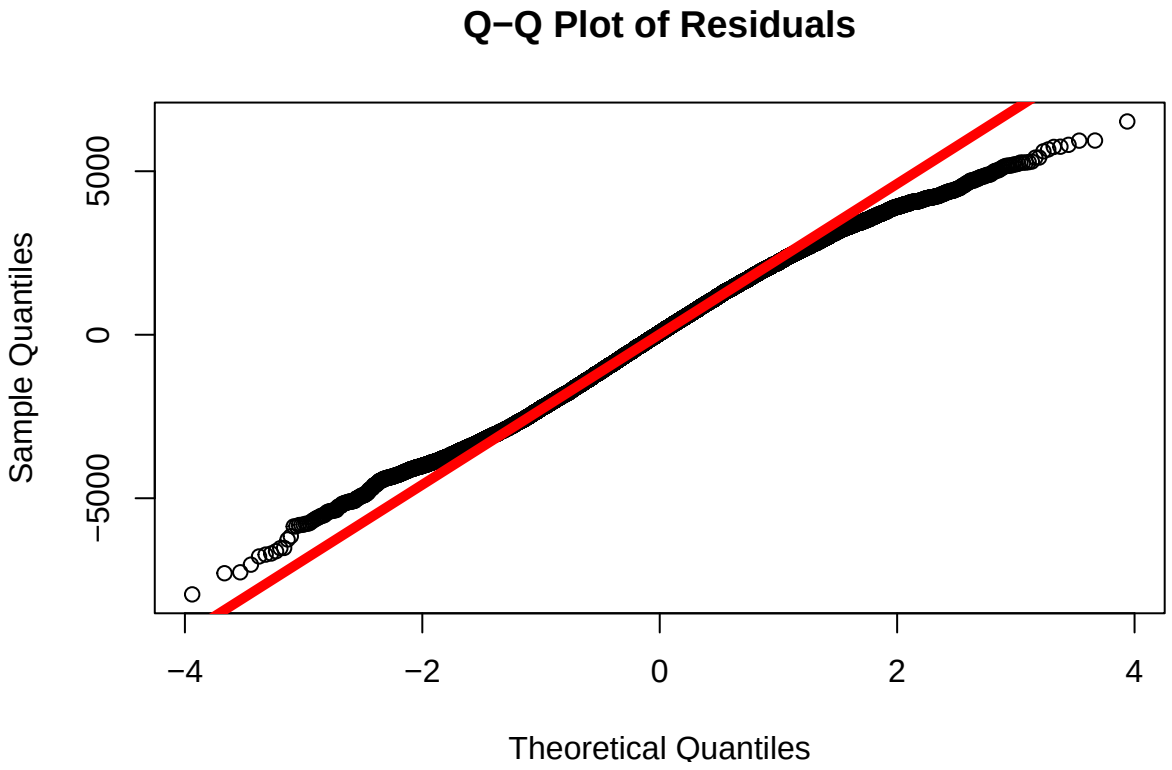
```r
# filer out all the outliers, leverage obs, and disproportionate values. Then, sort it out.     outlier
drop_data <- sort(unique(rbind(filtered_large_residuals)))

new_data <- data[-c(drop_data), ]

Filter_Score <- new_data$OVERALL_SCORE^1.5

model2 <- lm(Filter_Score ~ GENDER + EDU_FATHER + EDU_MOTHER + OCC_FATHER + OCC_MOTHER + STRATUM + MOBI
```

Verify

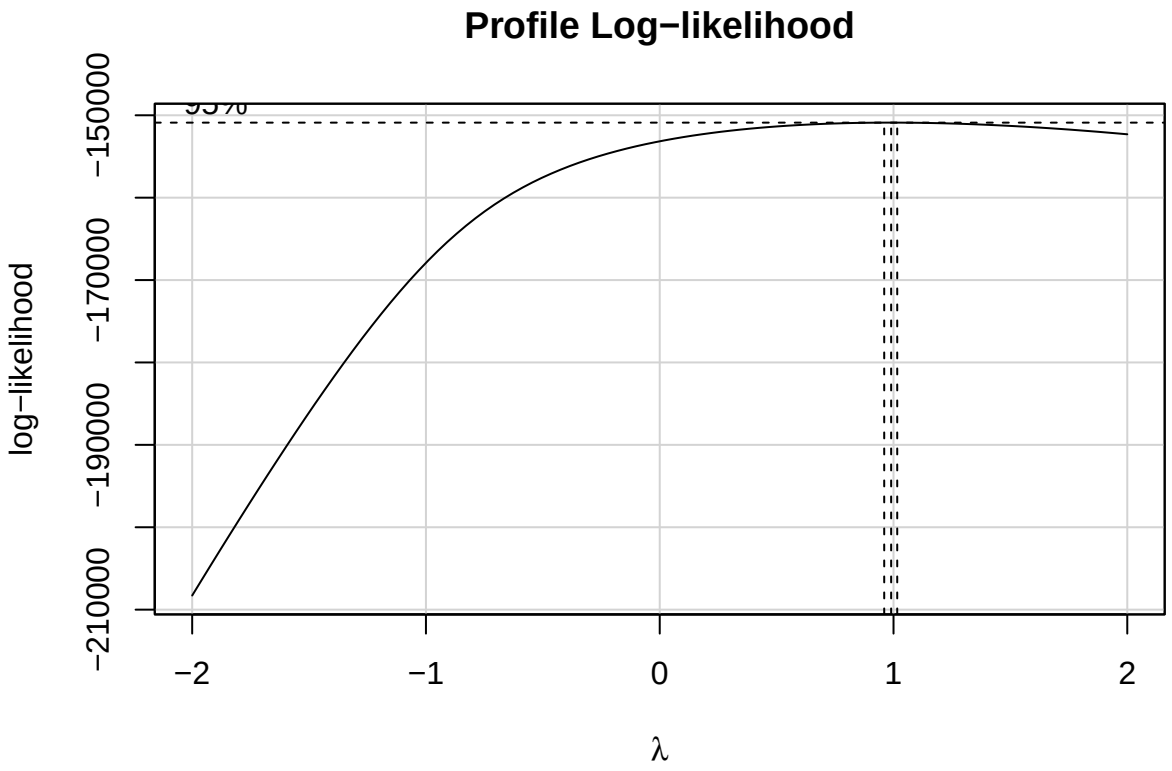
```r
hist(Filter_Score)
```



```r
residuals <- residuals(model2)
qqnorm(residuals, main = "Q-Q Plot of Residuals")
qqline(residuals, col = "red", lwd = 5)
```




```r
boxCox(model2)
```

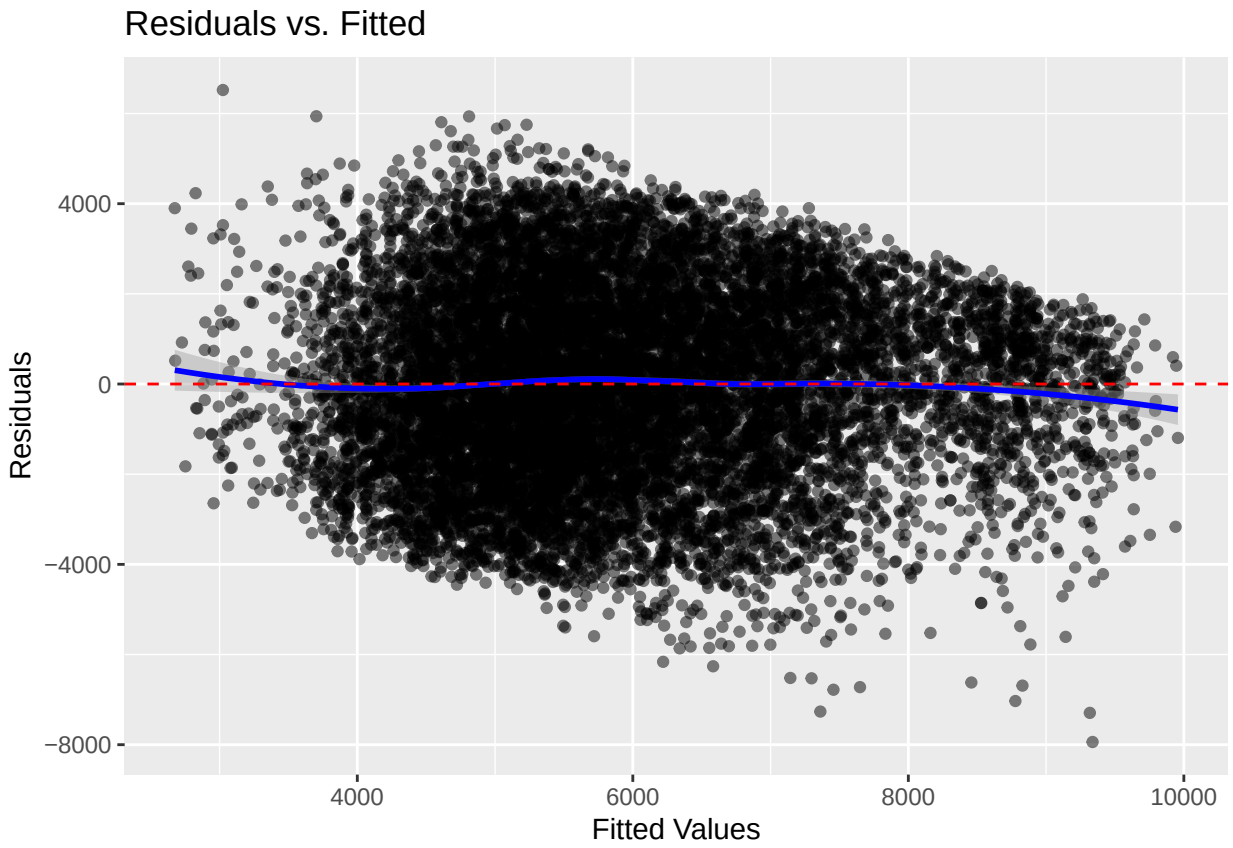


```r
p1 <- powerTransform(model2)
summary(p1)
```

```
## bcPower Transformation to Normality
##    Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## Y1    0.9884           1       0.9549       1.0218
##
## Likelihood ratio test that transformation parameter is equal to 0
##  (log transformation)
##                            LRT df       pval
## LR test, lambda = (0) 4533.512  1 < 2.22e-16
##
## Likelihood ratio test that no transformation is needed
##                             LRT df    pval
## LR test, lambda = (1) 0.4624622  1 0.49648
```

```r
y_value <- resid(model2)
x_value <- fitted(model2)

ggplot(data[-c(drop_data), ], aes(x = x_value, y = y_value)) +
  geom_point(alpha = 0.5) +                     # Scatter plot of residuals
  geom_smooth(method = "loess", color = "blue") + # LOESS smooth line to see trend
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") + # Line at zero
  labs(title = "Residuals vs. Fitted",
       x = "Fitted Values",
       y = "Residuals")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
summary(model2)
```

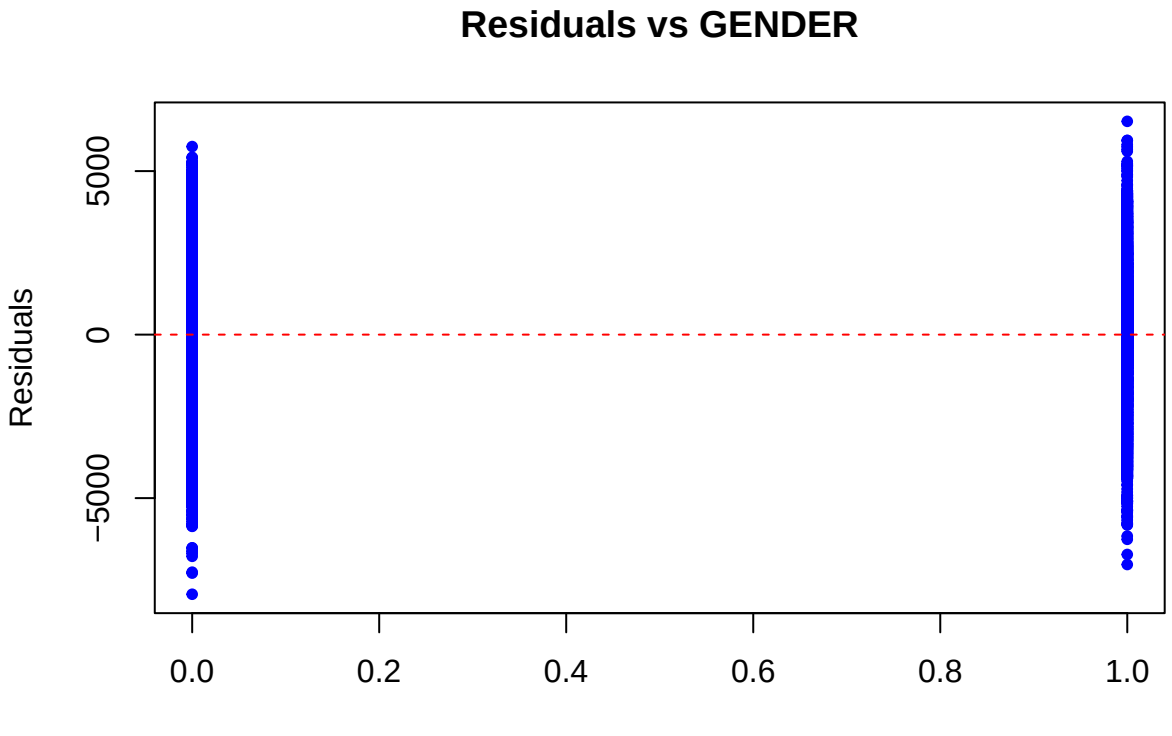
```
##
## Call:
## lm(formula = Filter_Score ~ GENDER + EDU_FATHER + EDU_MOTHER +
##     OCC_FATHER + OCC_MOTHER + STRATUM + MOBILE + UNIVERSITY,
##     data = new_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7940.0 -1537.7    51.2  1565.2  6521.8
##
## Coefficients:
##                                                  Estimate Std. Error t value
## (Intercept)                                      3943.265    664.211   5.937
## GENDER                                            169.960     38.913   4.368
## EDU_FATHERComplete Professional Education         143.487    100.947   1.421
## EDU_FATHERComplete Secondary Education              1.068     90.136   0.012
## EDU_FATHERComplete Technical Education            239.944    107.304   2.236
## EDU_FATHERIncomplete Primary Education             -5.418    111.909  -0.048
## EDU_FATHERIncomplete Professional Education       443.343    135.813   3.264
## EDU_FATHERIncomplete Secondary Education          -63.525    101.416  -0.626
## EDU_FATHERIncomplete Technical Education          287.382    155.561   1.847
## EDU_FATHERPostgraduate Education                  553.113    120.678   4.583
## EDU_FATHERUnknown                                -122.774    122.554  -1.002
## EDU_MOTHERComplete Professional Education         370.463    107.311   3.452
## EDU_MOTHERComplete Secondary Education             25.591     95.176   0.269
```

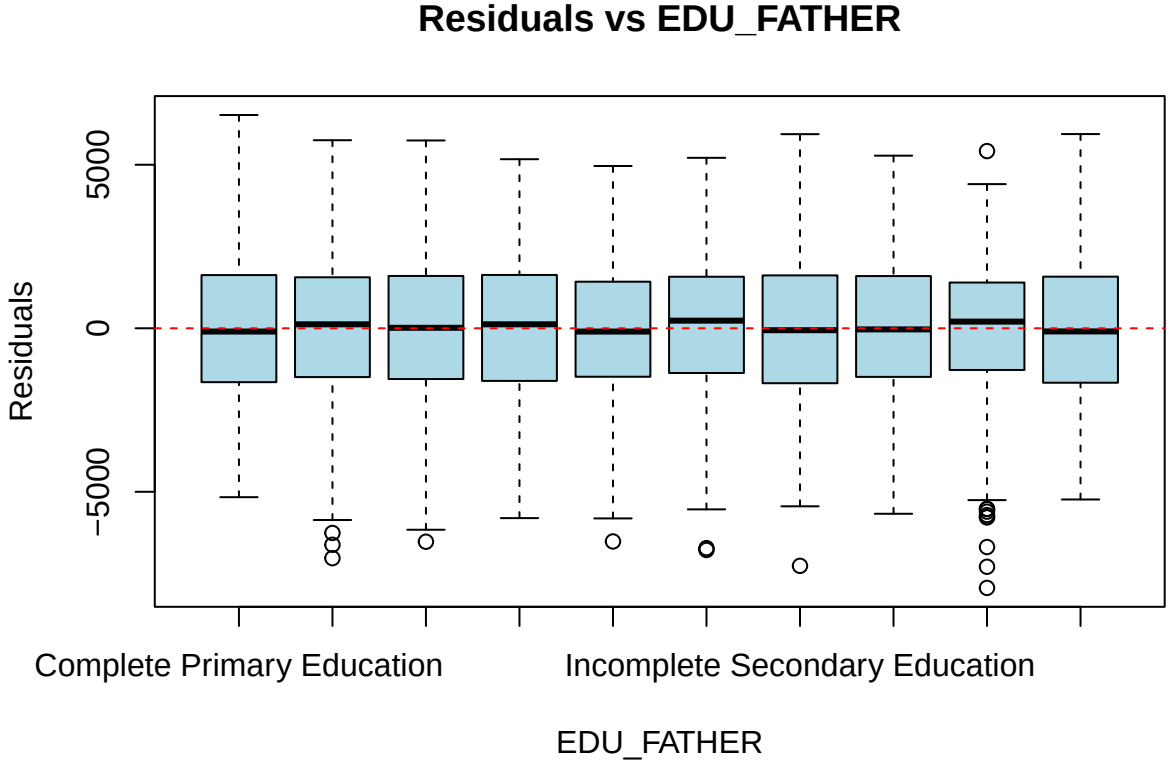
```
## EDU_MOTHERComplete Technical Education              202.014    108.769   1.857
## EDU_MOTHERIncomplete Primary Education             -113.834    125.792  -0.905
## EDU_MOTHERIncomplete Professional Education         336.471    134.866   2.495
## EDU_MOTHERIncomplete Secondary Education           -102.195    106.894  -0.956
## EDU_MOTHERIncomplete Technical Education            176.295    148.557   1.187
## EDU_MOTHERPostgraduate Education                    589.099    127.526   4.619
## EDU_MOTHERUnknown                                   -97.034    144.523  -0.671
## OCC_FATHERAuxiliary or Administrative               145.565    145.339   1.002
## OCC_FATHEREntrepreneur                              -83.857    137.042  -0.612
## OCC_FATHERExecutive                                 220.067    112.560   1.955
## OCC_FATHERHome                                       39.866    258.002   0.155
## OCC_FATHERIndependent                               116.105    104.143   1.115
## OCC_FATHERIndependent professional                  151.546    116.134   1.305
## OCC_FATHEROperator                                  179.640    111.930   1.605
## OCC_FATHEROther occupation                          285.784    114.982   2.485
## OCC_FATHERRetired                                   368.702    131.554   2.803
## OCC_FATHERSmall entrepreneur                        -61.049    126.452  -0.483
## OCC_FATHERTechnical or professional level employee  118.062    106.906   1.104
## OCC_MOTHERAuxiliary or Administrative                13.162    170.885   0.077
## OCC_MOTHEREntrepreneur                            -270.159    211.016  -1.280
## OCC_MOTHERExecutive                               -142.382    173.584  -0.820
## OCC_MOTHERHome                                     -25.295    152.266  -0.166
## OCC_MOTHERIndependent                             -156.495    165.558  -0.945
## OCC_MOTHERIndependent professional                 -65.081    174.286  -0.373
## OCC_MOTHEROperator                                  51.395    174.736   0.294
## OCC_MOTHEROther occupation                        -293.507    176.861  -1.660
## OCC_MOTHERRetired                                  348.309    229.354   1.519
## OCC_MOTHERSmall entrepreneur                        12.781    183.701   0.070
## OCC_MOTHERTechnical or professional level employee -50.316    165.258  -0.304
## STRATUM1                                          -632.055    575.129  -1.099
## STRATUM2                                          -194.221    573.769  -0.339
## STRATUM3                                            49.046    574.139   0.085
## STRATUM4                                           337.693    576.583   0.586
## STRATUM5                                           291.900    581.349   0.502
## STRATUM6                                           498.099    586.138   0.850
## MOBILE                                             390.445     47.261   8.261
## UNIVERSITY1                                       -509.201    342.359  -1.487
## UNIVERSITY2                                        466.436    323.743   1.441
## UNIVERSITY3                                       1004.316    324.622   3.094
## UNIVERSITY4                                       1918.866    324.105   5.921
## UNIVERSITY5                                       3414.695    328.486  10.395
##                                                  Pr(>|t|)
## (Intercept)                                      2.99e-09 ***
## GENDER                                           1.27e-05 ***
## EDU_FATHERComplete Professional Education        0.155221
## EDU_FATHERComplete Secondary Education           0.990544
## EDU_FATHERComplete Technical Education           0.025363 *
## EDU_FATHERIncomplete Primary Education           0.961386
## EDU_FATHERIncomplete Professional Education      0.001100 **
## EDU_FATHERIncomplete Secondary Education         0.531074
## EDU_FATHERIncomplete Technical Education         0.064714 .
## EDU_FATHERPostgraduate Education                 4.62e-06 ***
## EDU_FATHERUnknown                                0.316462
## EDU_MOTHERComplete Professional Education        0.000558 ***
```

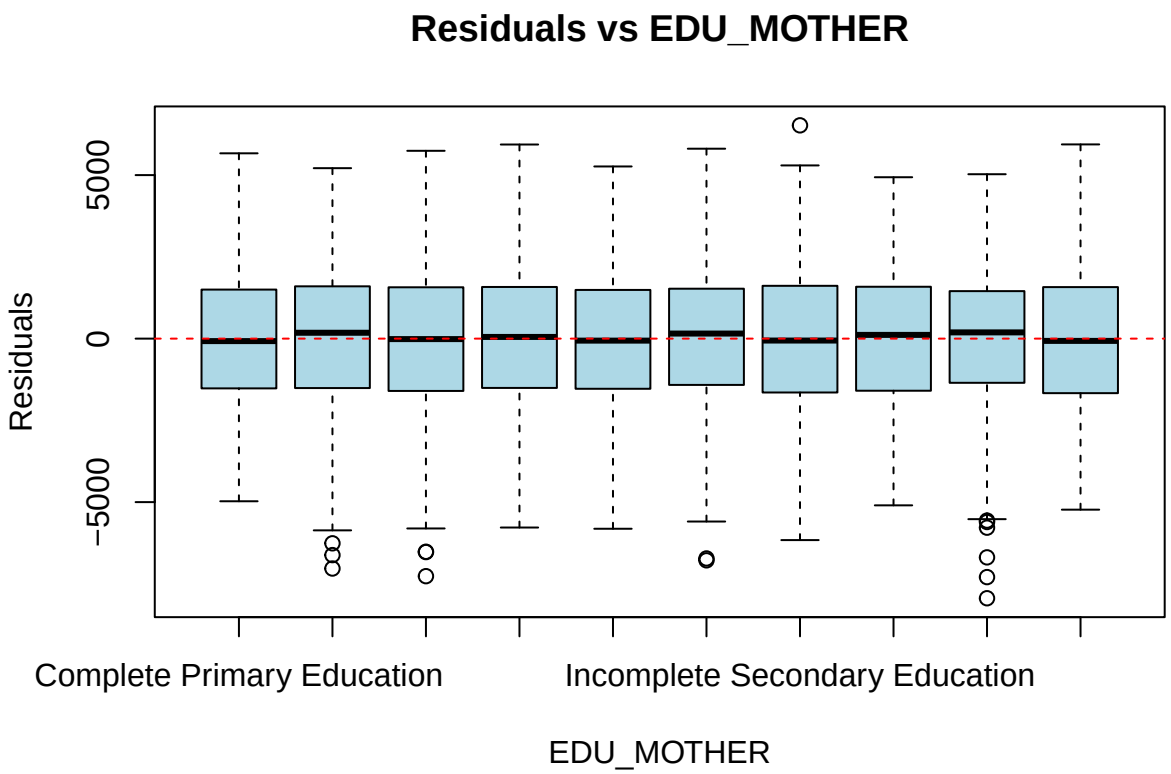
```
## EDU_MOTHERComplete Secondary Education             0.788026
## EDU_MOTHERComplete Technical Education             0.063296 .
## EDU_MOTHERIncomplete Primary Education             0.365517
## EDU_MOTHERIncomplete Professional Education        0.012614 *
## EDU_MOTHERIncomplete Secondary Education           0.339071
## EDU_MOTHERIncomplete Technical Education           0.235363
## EDU_MOTHERPostgraduate Education                   3.89e-06 ***
## EDU_MOTHERUnknown                                  0.501971
## OCC_FATHERAuxiliary or Administrative              0.316579
## OCC_FATHEREntrepreneur                             0.540609
## OCC_FATHERExecutive                                0.050593 .
## OCC_FATHERHome                                     0.877205
## OCC_FATHERIndependent                              0.264933
## OCC_FATHERIndependent professional                 0.191944
## OCC_FATHEROperator                                 0.108533
## OCC_FATHEROther occupation                         0.012952 *
## OCC_FATHERRetired                                  0.005076 **
## OCC_FATHERSmall entrepreneur                       0.629256
## OCC_FATHERTechnical or professional level employee 0.269460
## OCC_MOTHERAuxiliary or Administrative              0.938606
## OCC_MOTHEREntrepreneur                             0.200472
## OCC_MOTHERExecutive                                0.412090
## OCC_MOTHERHome                                     0.868062
## OCC_MOTHERIndependent                              0.344548
## OCC_MOTHERIndependent professional                 0.708845
## OCC_MOTHEROperator                                 0.768663
## OCC_MOTHEROther occupation                         0.097033 .
## OCC_MOTHERRetired                                  0.128877
## OCC_MOTHERSmall entrepreneur                       0.944531
## OCC_MOTHERTechnical or professional level employee 0.760776
## STRATUM1                                           0.271799
## STRATUM2                                           0.734992
## STRATUM3                                           0.931925
## STRATUM4                                           0.558102
## STRATUM5                                           0.615601
## STRATUM6                                           0.395454
## MOBILE                                              < 2e-16 ***
## UNIVERSITY1                                        0.136954
## UNIVERSITY2                                        0.149678
## UNIVERSITY3                                        0.001981 **
## UNIVERSITY4                                        3.30e-09 ***
## UNIVERSITY5                                         < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2107 on 12158 degrees of freedom
## Multiple R-squared:  0.2819, Adjusted R-squared:  0.2788
## F-statistic: 90.07 on 53 and 12158 DF,  p-value: < 2.2e-16
```

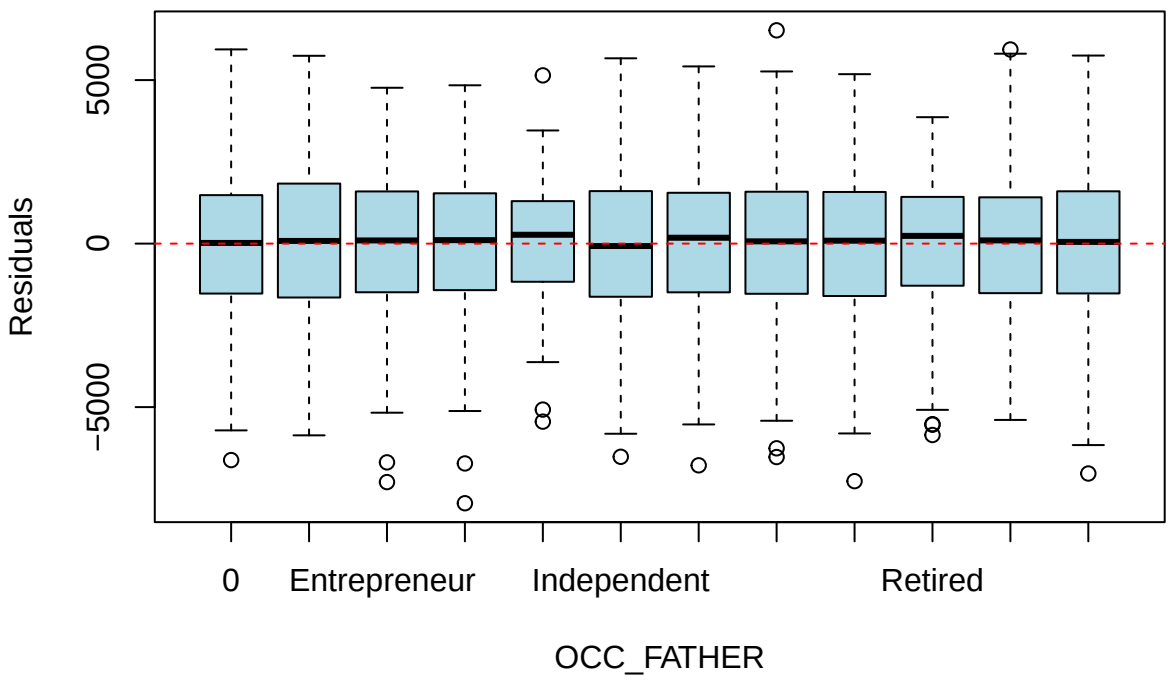
```r
residuals <- residuals(model2)

# Extract the data used in the model (this excludes observations with missing values)
model_data <- model2$model
```

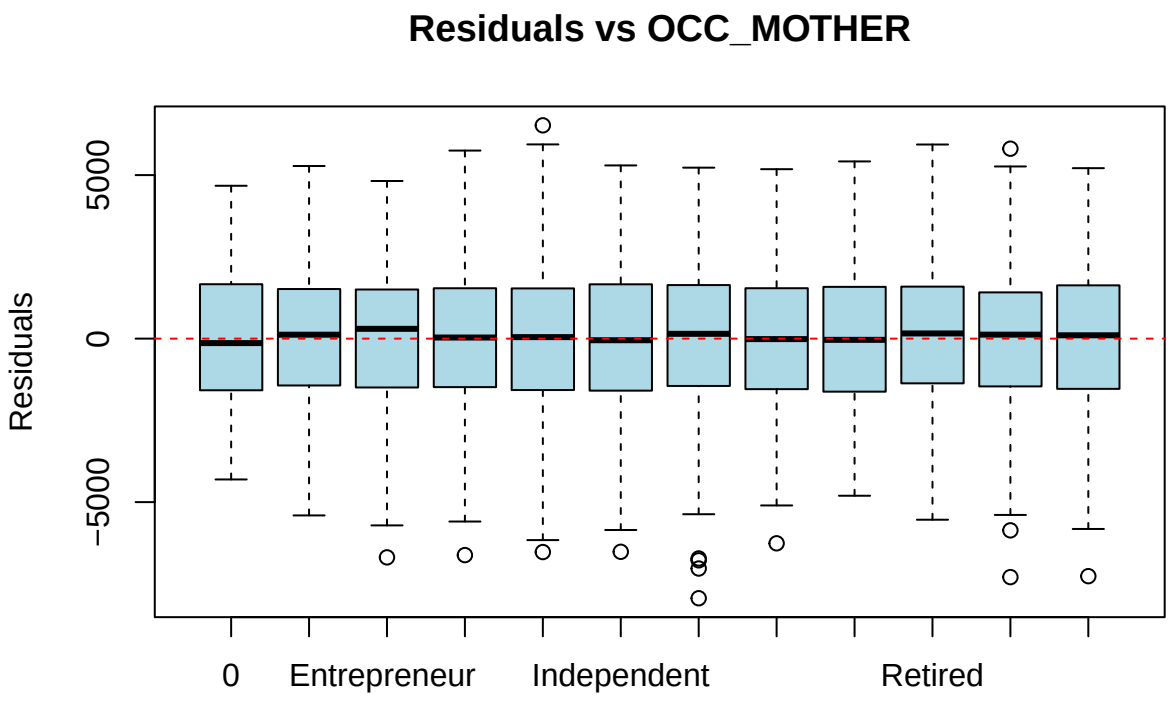
```r
# Add residuals to the model data
model_data$residuals <- residuals

# Identify the response variable (first column in model data)
response_var <- names(model_data)[1]

# Get the list of predictors from the model data
predictors <- setdiff(names(model_data), c(response_var, "residuals"))

# Set up the plotting area
#par(mfrow = c(3, 4), mar = c(4, 4, 2, 1))  # Adjust rows and columns as needed

# Loop over each predictor and create a residual plot
for (pred in predictors) {
  predictor_data <- model_data[[pred]]

  # Check if predictor is categorical or numerical
  if (is.factor(predictor_data) || is.character(predictor_data)) {
    # Create a boxplot for categorical predictors
    boxplot(residuals ~ predictor_data,
            data = model_data,
            xlab = pred,
            ylab = "Residuals",
            main = paste("Residuals vs", pred),
            col = "lightblue")
    abline(h = 0, col = "red", lty = 2)
  } else {
    # Create a scatter plot for numerical predictors
    plot(predictor_data, model_data$residuals,
         xlab = pred,
         ylab = "Residuals",
         main = paste("Residuals vs", pred),
         pch = 20,
         col = "blue")
    abline(h = 0, col = "red", lty = 2)
  }
}
```

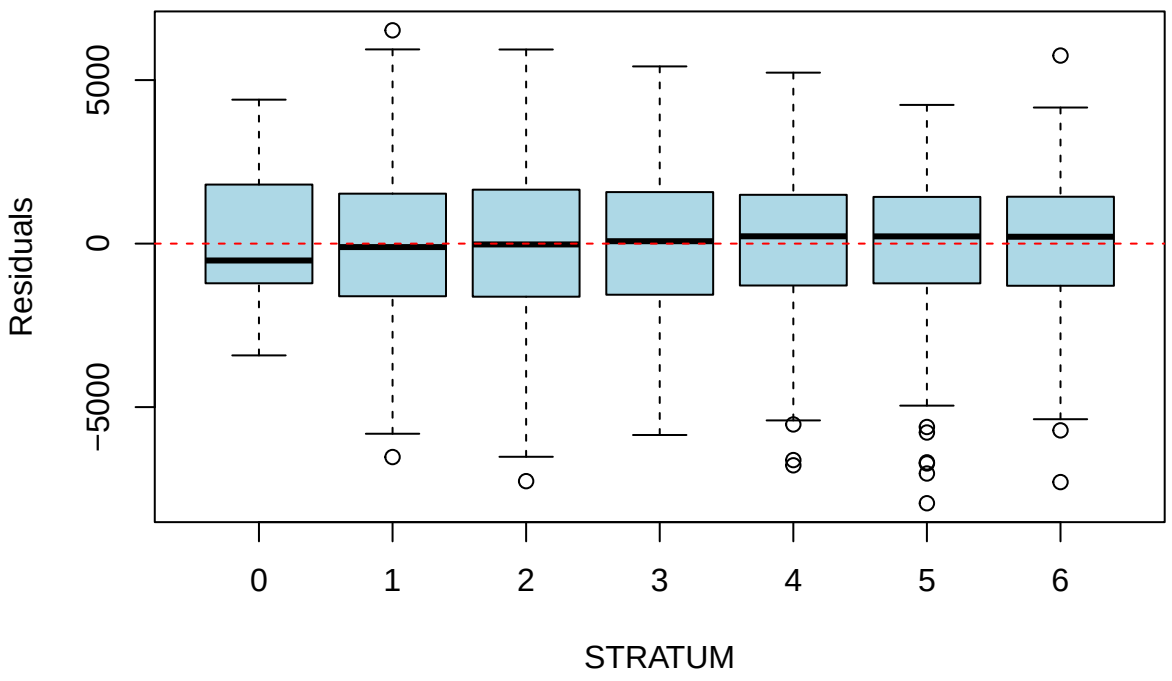

Residuals vs GENDER
Residuals
5000
0
−5000
0.0
0.2
0.4
0.6
0.8
1.0
GENDER

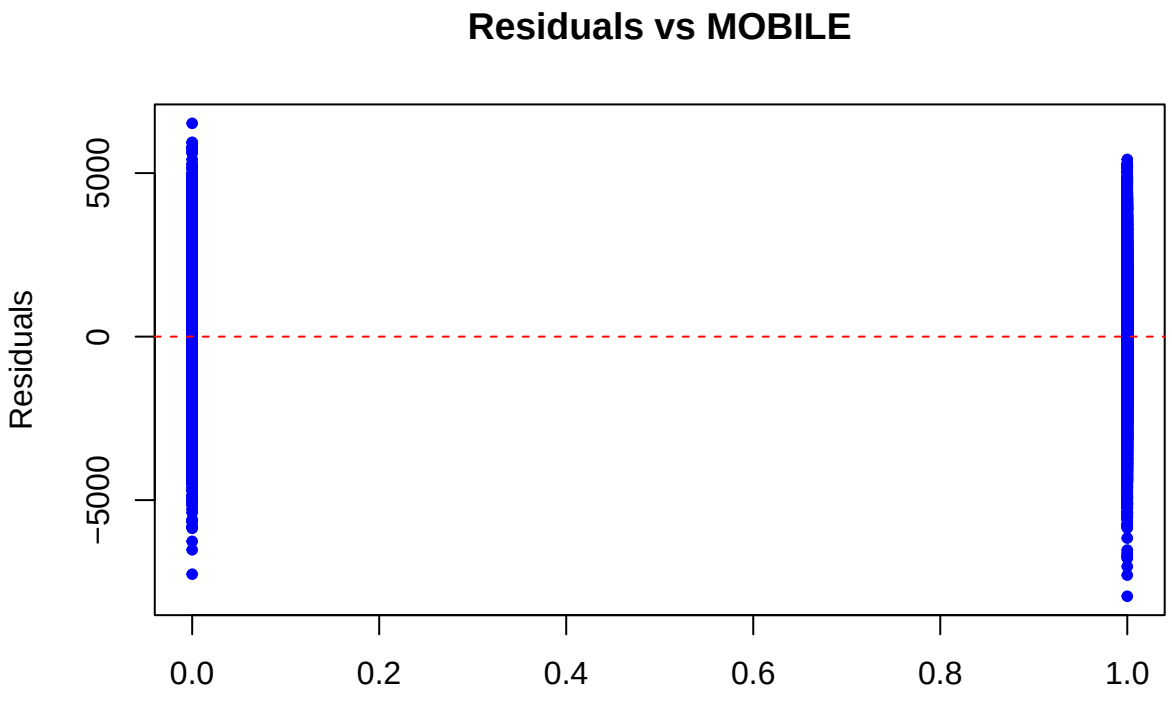

Residuals vs EDU_FATHER
Residuals
5000
0
−5000
Complete Primary Education
Incomplete Secondary Education
EDU_FATHER

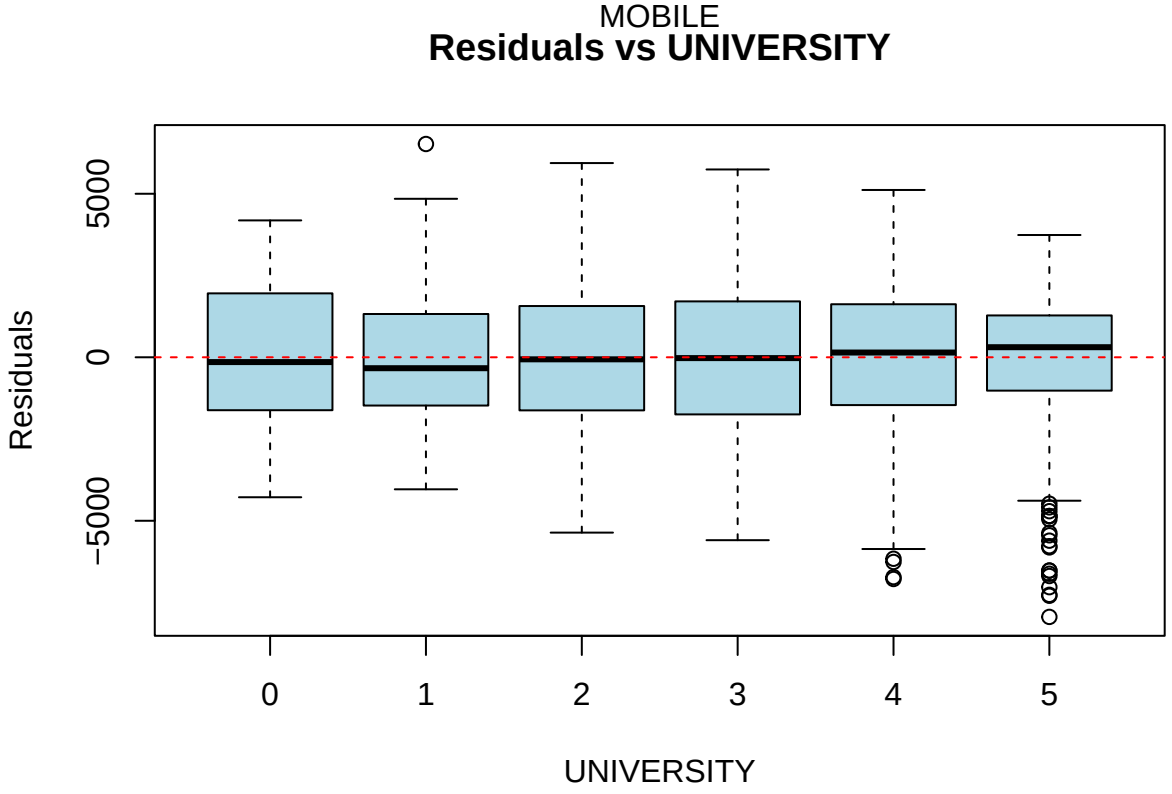

Residuals vs EDU_MOTHER
Residuals
5000
0
−5000
Complete Primary Education
Incomplete Secondary Education
EDU_MOTHER


Residuals vs OCC_FATHER
Residuals
5000
0
−5000
0
Entrepreneur
Independent
Retired
OCC_FATHER


Residuals vs OCC_MOTHER
Residuals
5000
0
−5000
0
Entrepreneur
Independent
Retired
OCC_MOTHER


Residuals vs STRATUM
Residuals
5000
0
−5000
0
1
2
3
4
5
6
STRATUM





Lastly use Automated Selection for final check

```
library(MASS)
```

```
##
```

```
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select
```

```
stepAIC(lm(OVERALL_SCORE ~ ., data=new_data[,-1]), direction = "both", k=2)
```

```
## Start:  AIC=107671.5
## OVERALL_SCORE ~ GENDER + EDU_FATHER + EDU_MOTHER + OCC_FATHER +
##     OCC_MOTHER + STRATUM + PEOPLE_HOUSE + COMPUTER + CAR + MOBILE +
##     UNIVERSITY
##
##                 Df Sum of Sq      RSS    AIC
## - PEOPLE_HOUSE 12    136980 81619616 107668
## - COMPUTER      1      5777 81488413 107670
## <none>                      81482636 107672
## - CAR           1     17584 81500220 107672
## - OCC_MOTHER   11    154793 81637429 107673
## - OCC_FATHER   11    159575 81642211 107673
## - GENDER        1    144095 81626731 107691
## - EDU_MOTHER    9    349128 81831764 107706
## - EDU_FATHER    9    371231 81853867 107709
## - MOBILE        1    474924 81957560 107740
## - STRATUM       6   1011384 82494020 107810
## - UNIVERSITY    5  13685851 95168487 109558
##
## Step:  AIC=107668
## OVERALL_SCORE ~ GENDER + EDU_FATHER + EDU_MOTHER + OCC_FATHER +
##     OCC_MOTHER + STRATUM + COMPUTER + CAR + MOBILE + UNIVERSITY
##
##                 Df Sum of Sq      RSS    AIC
## - COMPUTER      1      6072 81625688 107667
## <none>                      81619616 107668
## - CAR           1     19842 81639458 107669
## - OCC_FATHER   11    161133 81780749 107670
## - OCC_MOTHER   11    162170 81781786 107670
## + PEOPLE_HOUSE 12    136980 81482636 107672
## - GENDER        1    144817 81764433 107688
## - EDU_MOTHER    9    360316 81979932 107704
## - EDU_FATHER    9    372003 81991619 107706
## - MOBILE        1    458990 82078606 107735
## - STRATUM       6   1031958 82651574 107809
## - UNIVERSITY    5  13702736 95322352 109553
##
## Step:  AIC=107666.9
## OVERALL_SCORE ~ GENDER + EDU_FATHER + EDU_MOTHER + OCC_FATHER +
##     OCC_MOTHER + STRATUM + CAR + MOBILE + UNIVERSITY
##
##                 Df Sum of Sq      RSS    AIC
## <none>                      81625688 107667
## - CAR           1     17096 81642783 107668
## + COMPUTER      1      6072 81619616 107668
## - OCC_MOTHER   11    160060 81785748 107669
## - OCC_FATHER   11    162998 81788686 107669
```

```
## + PEOPLE_HOUSE 12    137275 81488413 107670
## - GENDER        1    145321 81771009 107687
## - EDU_MOTHER    9    365400 81991088 107703
## - EDU_FATHER    9    374430 82000118 107705
## - MOBILE        1    486426 82112114 107738
## - STRATUM       6   1075930 82701618 107815
## - UNIVERSITY    5  13697783 95323470 109551
##
## Call:
## lm(formula = OVERALL_SCORE ~ GENDER + EDU_FATHER + EDU_MOTHER +
##     OCC_FATHER + OCC_MOTHER + STRATUM + CAR + MOBILE + UNIVERSITY,
##     data = new_data[, -1])
##
## Coefficients:
##                                         (Intercept)
##                                           241.29108
##                                              GENDER
##                                             7.04066
##          EDU_FATHERComplete Professional Education
##                                             6.08168
##             EDU_FATHERComplete Secondary Education
##                                             0.18473
##             EDU_FATHERComplete Technical Education
##                                             9.77249
##             EDU_FATHERIncomplete Primary Education
##                                             0.69293
##        EDU_FATHERIncomplete Professional Education
##                                            17.50108
##           EDU_FATHERIncomplete Secondary Education
##                                            -2.31437
##           EDU_FATHERIncomplete Technical Education
##                                            11.09802
##                  EDU_FATHERPostgraduate Education
##                                            21.39616
##                                EDU_FATHERUnknown
##                                            -4.30520
##          EDU_MOTHERComplete Professional Education
##                                            14.65846
##             EDU_MOTHERComplete Secondary Education
##                                             1.40386
##             EDU_MOTHERComplete Technical Education
##                                             8.37271
##             EDU_MOTHERIncomplete Primary Education
##                                            -4.96804
##        EDU_MOTHERIncomplete Professional Education
##                                            13.66763
##           EDU_MOTHERIncomplete Secondary Education
##                                            -4.10135
##           EDU_MOTHERIncomplete Technical Education
##                                             7.83392
##                  EDU_MOTHERPostgraduate Education
##                                            22.47396
##                                EDU_MOTHERUnknown
```

```
##                                         -4.74419
##               OCC_FATHERAuxiliary or Administrative
##                                          4.42012
##                              OCC_FATHEREntrepreneur
##                                         -3.59037
##                                 OCC_FATHERExecutive
##                                          7.60585
##                                     OCC_FATHERHome
##                                          1.04591
##                                OCC_FATHERIndependent
##                                          3.82146
##                   OCC_FATHERIndependent professional
##                                          5.01714
##                                 OCC_FATHEROperator
##                                          6.17007
##                           OCC_FATHEROther occupation
##                                         10.76984
##                                  OCC_FATHERRetired
##                                         13.92468
##                         OCC_FATHERSmall entrepreneur
##                                         -2.58385
## OCC_FATHERTechnical or professional level employee
##                                          3.40629
##               OCC_MOTHERAuxiliary or Administrative
##                                         -1.01303
##                              OCC_MOTHEREntrepreneur
##                                        -12.45485
##                                 OCC_MOTHERExecutive
##                                         -6.80829
##                                     OCC_MOTHERHome
##                                         -2.34624
##                                OCC_MOTHERIndependent
##                                         -7.92624
##                   OCC_MOTHERIndependent professional
##                                         -4.10604
##                                 OCC_MOTHEROperator
##                                          1.01717
##                           OCC_MOTHEROther occupation
##                                        -13.05450
##                                  OCC_MOTHERRetired
##                                         11.22013
##                         OCC_MOTHERSmall entrepreneur
##                                         -0.06209
## OCC_MOTHERTechnical or professional level employee
##                                         -3.60932
##                                          STRATUM1
##                                        -23.89599
##                                          STRATUM2
##                                         -5.62191
##                                          STRATUM3
##                                          4.76286
##                                          STRATUM4
##                                          15.72953
##                                          STRATUM5
```

```
##                                          13.83152
##                                          STRATUM6
##                                          20.38974
##                                               CAR
##                                          -2.80424
##                                            MOBILE
##                                          15.70626
##                                        UNIVERSITY1
##                                          -19.07932
##                                        UNIVERSITY2
##                                          20.89761
##                                        UNIVERSITY3
##                                          41.70545
##                                        UNIVERSITY4
##                                          77.00710
##                                        UNIVERSITY5
##                                         129.05208
```

We have got very similar selection that we got in model2. We will keep PEOPLE_HOUSE and OCC_MOTHER for answering research questions

Answering to research questions

Final Linear Regression Eqaution

```
model_Fin <- lm(Filter_Score ~ GENDER + EDU_FATHER + EDU_MOTHER + OCC_FATHER + OCC_MOTHER + STRATUM + M

summary(model_Fin)
```

```
##
## Call:
## lm(formula = Filter_Score ~ GENDER + EDU_FATHER + EDU_MOTHER +
##     OCC_FATHER + OCC_MOTHER + STRATUM + MOBILE + UNIVERSITY +
##     PEOPLE_HOUSE, data = new_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7901.7 -1542.4    51.8  1553.5  6531.6
##
## Coefficients:
##                                                  Estimate Std. Error
## (Intercept)                                    4158.3714   677.5739
## GENDER                                          169.6681    38.9119
## EDU_FATHERComplete Professional Education       139.4791   100.9799
## EDU_FATHERComplete Secondary Education           -1.2216    90.1544
## EDU_FATHERComplete Technical Education          240.4916   107.3314
## EDU_FATHERIncomplete Primary Education           -3.0093   111.9621
## EDU_FATHERIncomplete Professional Education     442.8764   135.8104
## EDU_FATHERIncomplete Secondary Education        -63.5822   101.4340
## EDU_FATHERIncomplete Technical Education        281.1494   155.6635
## EDU_FATHERPostgraduate Education                546.4697   120.7328
## EDU_FATHERUnknown                              -134.3326   122.5972
## EDU_MOTHERComplete Professional Education       355.5117   107.3900
## EDU_MOTHERComplete Secondary Education           11.3420    95.2466
```

```
## EDU_MOTHERComplete Technical Education              187.9145   108.8332
## EDU_MOTHERIncomplete Primary Education             -119.1032   125.8419
## EDU_MOTHERIncomplete Professional Education         323.3621   134.8811
## EDU_MOTHERIncomplete Secondary Education           -105.4806   106.9012
## EDU_MOTHERIncomplete Technical Education            154.5358   148.6675
## EDU_MOTHERPostgraduate Education                    569.9602   127.6536
## EDU_MOTHERUnknown                                  -111.5103   144.7871
## OCC_FATHERAuxiliary or Administrative               137.0549   145.3790
## OCC_FATHEREntrepreneur                              -98.2779   137.0685
## OCC_FATHERExecutive                                 214.1163   112.5767
## OCC_FATHERHome                                       34.5237   258.0558
## OCC_FATHERIndependent                               112.4914   104.1784
## OCC_FATHERIndependent professional                  149.0674   116.1398
## OCC_FATHEROperator                                  174.8562   111.9511
## OCC_FATHEROther occupation                          276.3549   115.0170
## OCC_FATHERRetired                                   358.1862   131.5990
## OCC_FATHERSmall entrepreneur                        -66.3075   126.4704
## OCC_FATHERTechnical or professional level employee  113.7518   106.9413
## OCC_MOTHERAuxiliary or Administrative                66.1474   172.5716
## OCC_MOTHEREntrepreneur                            -217.3104   212.3664
## OCC_MOTHERExecutive                                 -85.6381   175.3017
## OCC_MOTHERHome                                       29.1225   154.2943
## OCC_MOTHERIndependent                              -102.9761   167.2697
## OCC_MOTHERIndependent professional                  -14.0233   175.9297
## OCC_MOTHEROperator                                  100.4776   176.3293
## OCC_MOTHEROther occupation                         -234.2423   178.5812
## OCC_MOTHERRetired                                   391.8157   230.6988
## OCC_MOTHERSmall entrepreneur                         69.0465   185.2686
## OCC_MOTHERTechnical or professional level employee   -0.1351   167.0302
## STRATUM1                                             21.8253   736.2199
## STRATUM2                                            454.0756   735.0752
## STRATUM3                                            695.3044   734.8162
## STRATUM4                                            981.7169   736.4472
## STRATUM5                                            938.9346   739.8209
## STRATUM6                                           1149.4393   743.8606
## MOBILE                                              399.5137    47.4073
## UNIVERSITY1                                        -523.3513   342.4416
## UNIVERSITY2                                         450.5230   323.8069
## UNIVERSITY3                                         990.2758   324.6914
## UNIVERSITY4                                        1906.2938   324.1781
## UNIVERSITY5                                        3396.5162   328.5283
## PEOPLE_HOUSE1                                     -1451.9208   865.6322
## PEOPLE_HOUSE2                                      -674.5260   620.2624
## PEOPLE_HOUSE3                                      -910.6468   616.6269
## PEOPLE_HOUSE4                                      -899.1426   616.0433
## PEOPLE_HOUSE5                                      -837.9881   616.5760
## PEOPLE_HOUSE6                                      -917.1072   618.5083
## PEOPLE_HOUSE7                                      -960.5737   625.9571
## PEOPLE_HOUSE8                                     -1385.4384   637.5296
## PEOPLE_HOUSE9                                      -926.6062   664.5998
## PEOPLE_HOUSE10                                     -849.9841   683.5705
## PEOPLE_HOUSE11                                     -512.4834   782.5731
## PEOPLE_HOUSE12                                    -1116.3771   724.0555
##                                                   t value Pr(>|t|)
```

```
## (Intercept)                                          6.137 8.66e-10 ***
## GENDER                                               4.360 1.31e-05 ***
## EDU_FATHERComplete Professional Education            1.381 0.167226
## EDU_FATHERComplete Secondary Education              -0.014 0.989189
## EDU_FATHERComplete Technical Education               2.241 0.025067 *
## EDU_FATHERIncomplete Primary Education              -0.027 0.978557
## EDU_FATHERIncomplete Professional Education          3.261 0.001113 **
## EDU_FATHERIncomplete Secondary Education            -0.627 0.530780
## EDU_FATHERIncomplete Technical Education             1.806 0.070922 .
## EDU_FATHERPostgraduate Education                     4.526 6.06e-06 ***
## EDU_FATHERUnknown                                   -1.096 0.273222
## EDU_MOTHERComplete Professional Education            3.310 0.000934 ***
## EDU_MOTHERComplete Secondary Education               0.119 0.905214
## EDU_MOTHERComplete Technical Education               1.727 0.084260 .
## EDU_MOTHERIncomplete Primary Education              -0.946 0.343937
## EDU_MOTHERIncomplete Professional Education          2.397 0.016527 *
## EDU_MOTHERIncomplete Secondary Education            -0.987 0.323804
## EDU_MOTHERIncomplete Technical Education             1.039 0.298605
## EDU_MOTHERPostgraduate Education                     4.465 8.08e-06 ***
## EDU_MOTHERUnknown                                   -0.770 0.441215
## OCC_FATHERAuxiliary or Administrative                0.943 0.345832
## OCC_FATHEREntrepreneur                              -0.717 0.473389
## OCC_FATHERExecutive                                  1.902 0.057200 .
## OCC_FATHERHome                                       0.134 0.893576
## OCC_FATHERIndependent                                1.080 0.280255
## OCC_FATHERIndependent professional                   1.284 0.199336
## OCC_FATHEROperator                                   1.562 0.118338
## OCC_FATHEROther occupation                           2.403 0.016288 *
## OCC_FATHERRetired                                    2.722 0.006502 **
## OCC_FATHERSmall entrepreneur                        -0.524 0.600085
## OCC_FATHERTechnical or professional level employee   1.064 0.287493
## OCC_MOTHERAuxiliary or Administrative                0.383 0.701501
## OCC_MOTHEREntrepreneur                              -1.023 0.306196
## OCC_MOTHERExecutive                                 -0.489 0.625192
## OCC_MOTHERHome                                       0.189 0.850295
## OCC_MOTHERIndependent                               -0.616 0.538151
## OCC_MOTHERIndependent professional                  -0.080 0.936470
## OCC_MOTHEROperator                                   0.570 0.568804
## OCC_MOTHEROther occupation                          -1.312 0.189651
## OCC_MOTHERRetired                                    1.698 0.089460 .
## OCC_MOTHERSmall entrepreneur                         0.373 0.709391
## OCC_MOTHERTechnical or professional level employee  -0.001 0.999355
## STRATUM1                                             0.030 0.976351
## STRATUM2                                             0.618 0.536767
## STRATUM3                                             0.946 0.344051
## STRATUM4                                             1.333 0.182542
## STRATUM5                                             1.269 0.204416
## STRATUM6                                             1.545 0.122316
## MOBILE                                               8.427  < 2e-16 ***
## UNIVERSITY1                                         -1.528 0.126466
## UNIVERSITY2                                          1.391 0.164150
## UNIVERSITY3                                          3.050 0.002294 **
## UNIVERSITY4                                          5.880 4.20e-09 ***
## UNIVERSITY5                                         10.339  < 2e-16 ***
```

```
## PEOPLE_HOUSE1                                        -1.677 0.093510 .
## PEOPLE_HOUSE2                                        -1.087 0.276844
## PEOPLE_HOUSE3                                        -1.477 0.139750
## PEOPLE_HOUSE4                                        -1.460 0.144441
## PEOPLE_HOUSE5                                        -1.359 0.174140
## PEOPLE_HOUSE6                                        -1.483 0.138161
## PEOPLE_HOUSE7                                        -1.535 0.124916
## PEOPLE_HOUSE8                                        -2.173 0.029789 *
## PEOPLE_HOUSE9                                        -1.394 0.163273
## PEOPLE_HOUSE10                                       -1.243 0.213727
## PEOPLE_HOUSE11                                       -0.655 0.512564
## PEOPLE_HOUSE12                                       -1.542 0.123139
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2106 on 12146 degrees of freedom
## Multiple R-squared:  0.2832, Adjusted R-squared:  0.2793
## F-statistic: 73.82 on 65 and 12146 DF,  p-value: < 2.2e-16
```

Correlation

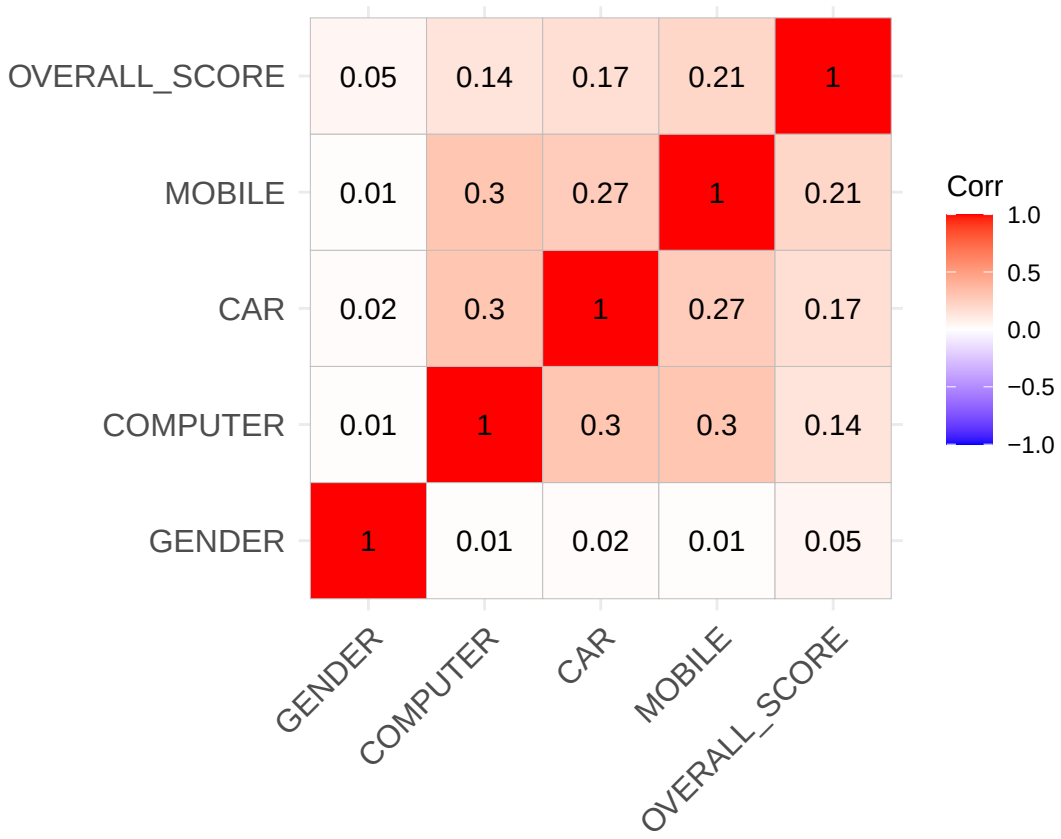
```
new_data <- new_data[, !names(new_data) %in% "X"]

new_corr <- cor(new_data[, sapply(new_data, is.numeric)])

new_corr
```

```
##                    GENDER    COMPUTER        CAR      MOBILE OVERALL_SCORE
## GENDER        1.000000000 0.008992342 0.02036606 0.007173569    0.05125525
## COMPUTER      0.008992342 1.000000000 0.29946210 0.296572331    0.14261851
## CAR           0.020366058 0.299462095 1.00000000 0.268420558    0.17482929
## MOBILE        0.007173569 0.296572331 0.26842056 1.000000000    0.21407682
## OVERALL_SCORE 0.051255248 0.142618513 0.17482929 0.214076820    1.00000000
```

```
install.packages("ggcorrplot")
```

```
## Installing package into '/usr/local/lib/R/site-library'
## (as 'lib' is unspecified)
```

```
library(ggcorrplot)
ggcorrplot(new_corr, lab = TRUE)
```



Correlation

for every categorical values

Gender

```
lm_Gender <- lm(Filter_Score ~ GENDER, data = new_data)

summary(lm_Gender)
```

```
##
## Call:
## lm(formula = Filter_Score ~ GENDER, data = new_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -5993  -1931     53   1949   5382
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5798.73      34.90  166.15  < 2e-16 ***
## GENDER        252.31      45.55    5.54 3.09e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2478 on 12210 degrees of freedom
## Multiple R-squared:  0.002507,   Adjusted R-squared:  0.002425
## F-statistic: 30.69 on 1 and 12210 DF,  p-value: 3.093e-08
```

EDU_FATHER

```
lm_edu_father <- lm(Filter_Score ~ EDU_FATHER, data = new_data)

summary(lm_edu_father)
```

```
##
## Call:
## lm(formula = Filter_Score ~ EDU_FATHER, data = new_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6771.3 -1807.7    83.4  1845.5  5795.7
##
## Coefficients:
##                                           Estimate Std. Error t value
## (Intercept)                                5137.21      83.46  61.555
## EDU_FATHERComplete Professional Education  1380.95      94.26  14.650
## EDU_FATHERComplete Secondary Education      366.43      94.84   3.863
## EDU_FATHERComplete Technical Education      959.50     108.35   8.856
## EDU_FATHERIncomplete Primary Education       13.44     121.38   0.111
## EDU_FATHERIncomplete Professional Education 1366.70     143.33   9.535
## EDU_FATHERIncomplete Secondary Education    122.37     110.56   1.107
## EDU_FATHERIncomplete Technical Education    789.70     166.46   4.744
## EDU_FATHERPostgraduate Education           2363.06     110.71  21.344
## EDU_FATHERUnknown                           219.87     114.87   1.914
##                                             Pr(>|t|)
## (Intercept)                                  < 2e-16 ***
## EDU_FATHERComplete Professional Education    < 2e-16 ***
## EDU_FATHERComplete Secondary Education      0.000112 ***
## EDU_FATHERComplete Technical Education       < 2e-16 ***
## EDU_FATHERIncomplete Primary Education      0.911822
## EDU_FATHERIncomplete Professional Education  < 2e-16 ***
## EDU_FATHERIncomplete Secondary Education    0.268395
## EDU_FATHERIncomplete Technical Education    2.12e-06 ***
## EDU_FATHERPostgraduate Education             < 2e-16 ***
## EDU_FATHERUnknown                           0.055628 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2380 on 12202 degrees of freedom
## Multiple R-squared:  0.08074,    Adjusted R-squared:  0.08006
## F-statistic: 119.1 on 9 and 12202 DF,  p-value: < 2.2e-16
```

EDU_MOTHER

```
lm_edu_mother <- lm(Filter_Score ~ EDU_MOTHER, data = new_data)

summary(lm_edu_mother)
```

```
##
## Call:
## lm(formula = Filter_Score ~ EDU_MOTHER, data = new_data)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6507.5 -1805.0    86.5  1842.5  5869.5
##
## Coefficients:
##                                           Estimate Std. Error t value
## (Intercept)                                5131.11      89.82  57.128
## EDU_MOTHERComplete Professional Education  1508.95      99.80  15.119
## EDU_MOTHERComplete Secondary Education      345.60      99.65   3.468
## EDU_MOTHERComplete Technical Education      886.47     109.05   8.129
## EDU_MOTHERIncomplete Primary Education      -54.20     136.34  -0.398
## EDU_MOTHERIncomplete Professional Education 1267.80     139.89   9.063
## EDU_MOTHERIncomplete Secondary Education      16.07     116.44   0.138
## EDU_MOTHERIncomplete Technical Education     756.30     157.31   4.808
## EDU_MOTHERPostgraduate Education            2195.24     117.55  18.674
## EDU_MOTHERUnknown                             70.86     132.60   0.534
##                                             Pr(>|t|)
## (Intercept)                                  < 2e-16 ***
## EDU_MOTHERComplete Professional Education    < 2e-16 ***
## EDU_MOTHERComplete Secondary Education      0.000526 ***
## EDU_MOTHERComplete Technical Education      4.74e-16 ***
## EDU_MOTHERIncomplete Primary Education      0.690973
## EDU_MOTHERIncomplete Professional Education  < 2e-16 ***
## EDU_MOTHERIncomplete Secondary Education    0.890235
## EDU_MOTHERIncomplete Technical Education    1.55e-06 ***
## EDU_MOTHERPostgraduate Education             < 2e-16 ***
## EDU_MOTHERUnknown                           0.593075
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2381 on 12202 degrees of freedom
## Multiple R-squared:  0.07936,    Adjusted R-squared:  0.07868
## F-statistic: 116.9 on 9 and 12202 DF,  p-value: < 2.2e-16
```

OCC_FATHER

```
lm_occ_father <- lm(Filter_Score ~ OCC_FATHER, data = new_data)

summary(lm_occ_father)
```

```
##
## Call:
## lm(formula = Filter_Score ~ OCC_FATHER, data = new_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6260.1 -1877.5    77.1  1919.1  5481.4
##
## Coefficients:
##                                               Estimate Std. Error t value
## (Intercept)                                    6075.06      80.00  75.941
## OCC_FATHERAuxiliary or Administrative          -229.66     150.18  -1.529
## OCC_FATHEREntrepreneur                          494.06     138.71   3.562
## OCC_FATHERExecutive                             733.48     109.70   6.686
```

```
## OCC_FATHERHome                                        -273.71     289.15  -0.947
## OCC_FATHERIndependent                                 -598.93      92.09  -6.504
## OCC_FATHERIndependent professional                     424.20     114.14   3.717
## OCC_FATHEROperator                                    -610.11     101.61  -6.005
## OCC_FATHEROther occupation                            -427.82     109.32  -3.914
## OCC_FATHERRetired                                      193.15     133.87   1.443
## OCC_FATHERSmall entrepreneur                          -339.97     122.74  -2.770
## OCC_FATHERTechnical or professional level employee     216.72      98.72   2.195
##                                                      Pr(>|t|)
## (Intercept)                                           < 2e-16 ***
## OCC_FATHERAuxiliary or Administrative                0.126233
## OCC_FATHEREntrepreneur                               0.000370 ***
## OCC_FATHERExecutive                                  2.39e-11 ***
## OCC_FATHERHome                                       0.343862
## OCC_FATHERIndependent                                8.12e-11 ***
## OCC_FATHERIndependent professional                   0.000203 ***
## OCC_FATHEROperator                                   1.97e-09 ***
## OCC_FATHEROther occupation                           9.15e-05 ***
## OCC_FATHERRetired                                    0.149100
## OCC_FATHERSmall entrepreneur                         0.005618 **
## OCC_FATHERTechnical or professional level employee 0.028158 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2438 on 12200 degrees of freedom
## Multiple R-squared:  0.03504,    Adjusted R-squared:  0.03417
## F-statistic: 40.27 on 11 and 12200 DF,  p-value: < 2.2e-16
```

OCC_MOTHER

```
lm_occ_mother <- lm(Filter_Score ~ OCC_MOTHER, data = new_data)

summary(lm_occ_mother)
```

```
##
## Call:
## lm(formula = Filter_Score ~ OCC_MOTHER, data = new_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -6593  -1866     69   1909   5489
##
## Coefficients:
##                                                    Estimate Std. Error t value
## (Intercept)                                         5389.94     137.91  39.083
## OCC_MOTHERAuxiliary or Administrative                730.93     161.60   4.523
## OCC_MOTHEREntrepreneur                              1176.50     209.59   5.613
## OCC_MOTHERExecutive                                 1328.04     163.40   8.127
## OCC_MOTHERHome                                       234.39     142.56   1.644
## OCC_MOTHERIndependent                                140.76     156.35   0.900
## OCC_MOTHERIndependent professional                  1392.78     165.79   8.401
## OCC_MOTHEROperator                                   291.34     167.05   1.744
## OCC_MOTHEROther occupation                            99.39     170.51   0.583
## OCC_MOTHERRetired                                   1407.85     239.12   5.888
```

```
## OCC_MOTHERSmall entrepreneur                            751.00     176.69   4.250
## OCC_MOTHERTechnical or professional level employee  1034.65     149.59   6.916
##                                                    Pr(>|t|)
## (Intercept)                                         < 2e-16 ***
## OCC_MOTHERAuxiliary or Administrative              6.15e-06 ***
## OCC_MOTHEREntrepreneur                             2.03e-08 ***
## OCC_MOTHERExecutive                                4.81e-16 ***
## OCC_MOTHERHome                                       0.1002
## OCC_MOTHERIndependent                                0.3680
## OCC_MOTHERIndependent professional                  < 2e-16 ***
## OCC_MOTHEROperator                                   0.0812 .
## OCC_MOTHEROther occupation                           0.5600
## OCC_MOTHERRetired                                  4.02e-09 ***
## OCC_MOTHERSmall entrepreneur                       2.15e-05 ***
## OCC_MOTHERTechnical or professional level employee 4.87e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2440 on 12200 degrees of freedom
## Multiple R-squared:  0.03377,    Adjusted R-squared:  0.0329
## F-statistic: 38.76 on 11 and 12200 DF,  p-value: < 2.2e-16
```

STRATUM

```
lm_stratum <- lm(Filter_Score ~ STRATUM, data = new_data)

summary(lm_stratum)
```

```
##
## Call:
## lm(formula = Filter_Score ~ STRATUM, data = new_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6975.4 -1758.0    55.4  1806.2  6046.6
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4860.1      625.3   7.772 8.32e-15 ***
## STRATUM1      -126.4      627.9  -0.201 0.840495
## STRATUM2       595.5      626.4   0.951 0.341775
## STRATUM3      1234.7      626.4   1.971 0.048732 *
## STRATUM4      2156.5      628.1   3.433 0.000598 ***
## STRATUM5      2653.1      632.3   4.196 2.74e-05 ***
## STRATUM6      3055.9      636.3   4.803 1.59e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2340 on 12205 degrees of freedom
## Multiple R-squared:  0.1111, Adjusted R-squared:  0.1107
## F-statistic: 254.3 on 6 and 12205 DF,  p-value: < 2.2e-16
```

MOBILE

```
lm_mobile <- lm(Filter_Score ~ MOBILE, data = new_data)

summary(lm_mobile)
```

```
##
## Call:
## lm(formula = Filter_Score ~ MOBILE, data = new_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6224.8 -1850.4    56.6  1897.8  5932.9
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5113.57      40.92  124.96   <2e-16 ***
## MOBILE       1169.30      48.47   24.12   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2424 on 12210 degrees of freedom
## Multiple R-squared:  0.04549,    Adjusted R-squared:  0.04541
## F-statistic: 581.9 on 1 and 12210 DF,  p-value: < 2.2e-16
```

UNIVERSITY

```
lm_university <- lm(Filter_Score ~ UNIVERSITY, data = new_data)

summary(lm_university)
```

```
##
## Call:
## lm(formula = Filter_Score ~ UNIVERSITY, data = new_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8400.2 -1622.7    81.6  1638.1  6029.7
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4567.0      334.2  13.664  < 2e-16 ***
## UNIVERSITY1   -755.1      355.3  -2.125  0.03357 *
## UNIVERSITY2    383.0      336.0   1.140  0.25435
## UNIVERSITY3    922.7      336.9   2.739  0.00618 **
## UNIVERSITY4   2096.7      336.1   6.238 4.57e-10 ***
## UNIVERSITY5   3929.5      339.8  11.566  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2192 on 12206 degrees of freedom
## Multiple R-squared:   0.22,  Adjusted R-squared:  0.2196
## F-statistic: 688.4 on 5 and 12206 DF,  p-value: < 2.2e-16
```

PEOPLE_HOUSE

```
lm_household <- lm(Filter_Score ~ PEOPLE_HOUSE, data = new_data)

summary(lm_household)
```

```
##
## Call:
## lm(formula = Filter_Score ~ PEOPLE_HOUSE, data = new_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6122.6 -1906.4    56.1  1952.0  5317.0
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     5439.31     540.03  10.072   <2e-16 ***
## PEOPLE_HOUSE1   -141.24     895.54  -0.158    0.875
## PEOPLE_HOUSE2    779.56     549.61   1.418    0.156
## PEOPLE_HOUSE3    563.05     542.49   1.038    0.299
## PEOPLE_HOUSE4    608.69     541.24   1.125    0.261
## PEOPLE_HOUSE5    493.79     542.04   0.911    0.362
## PEOPLE_HOUSE6    218.55     545.28   0.401    0.689
## PEOPLE_HOUSE7     60.71     555.31   0.109    0.913
## PEOPLE_HOUSE8   -406.52     574.38  -0.708    0.479
## PEOPLE_HOUSE9    -63.42     613.76  -0.103    0.918
## PEOPLE_HOUSE10   226.28     643.52   0.352    0.725
## PEOPLE_HOUSE11   263.29     783.56   0.336    0.737
## PEOPLE_HOUSE12  -272.79     699.43  -0.390    0.697
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2475 on 12199 degrees of freedom
## Multiple R-squared:  0.006033,   Adjusted R-squared:  0.005056
## F-statistic: 6.171 on 12 and 12199 DF,  p-value: 6.048e-11
```

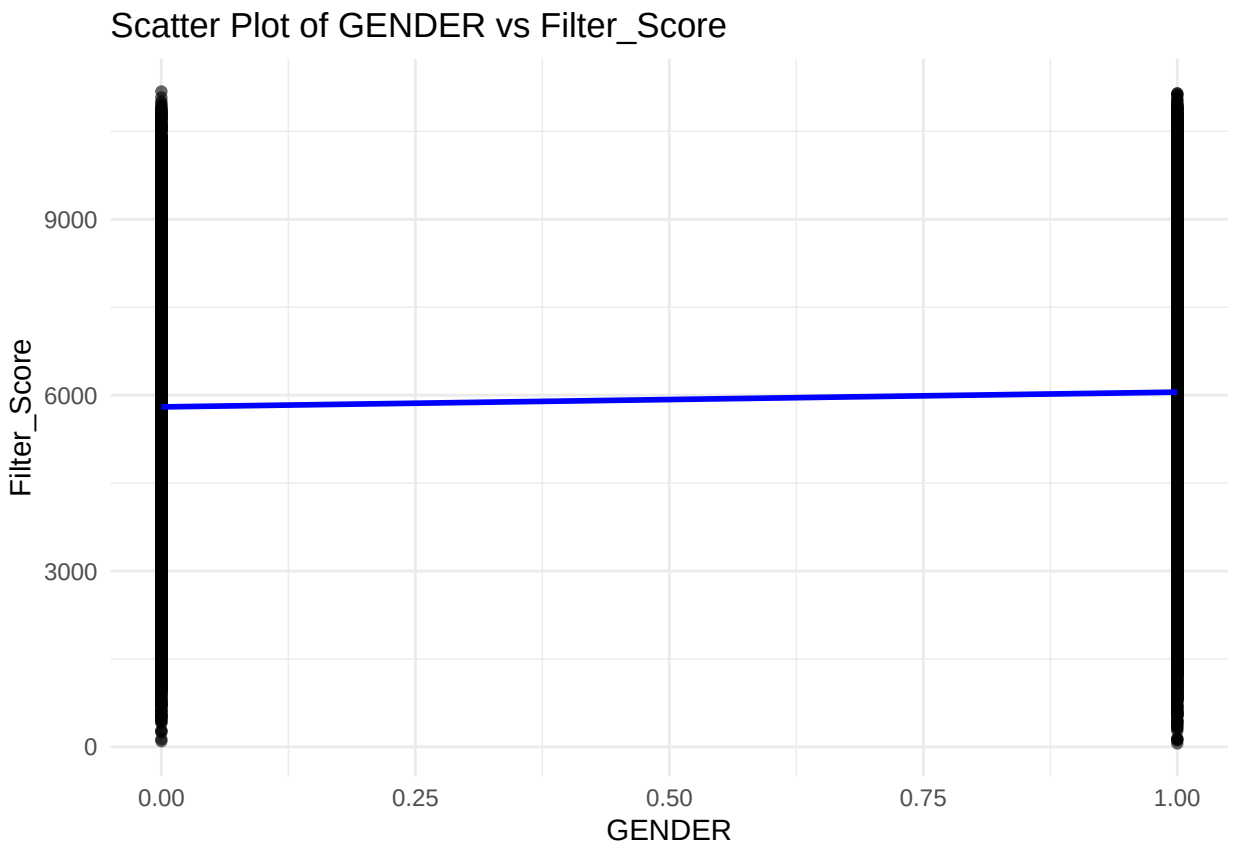
Graphs for each predictors

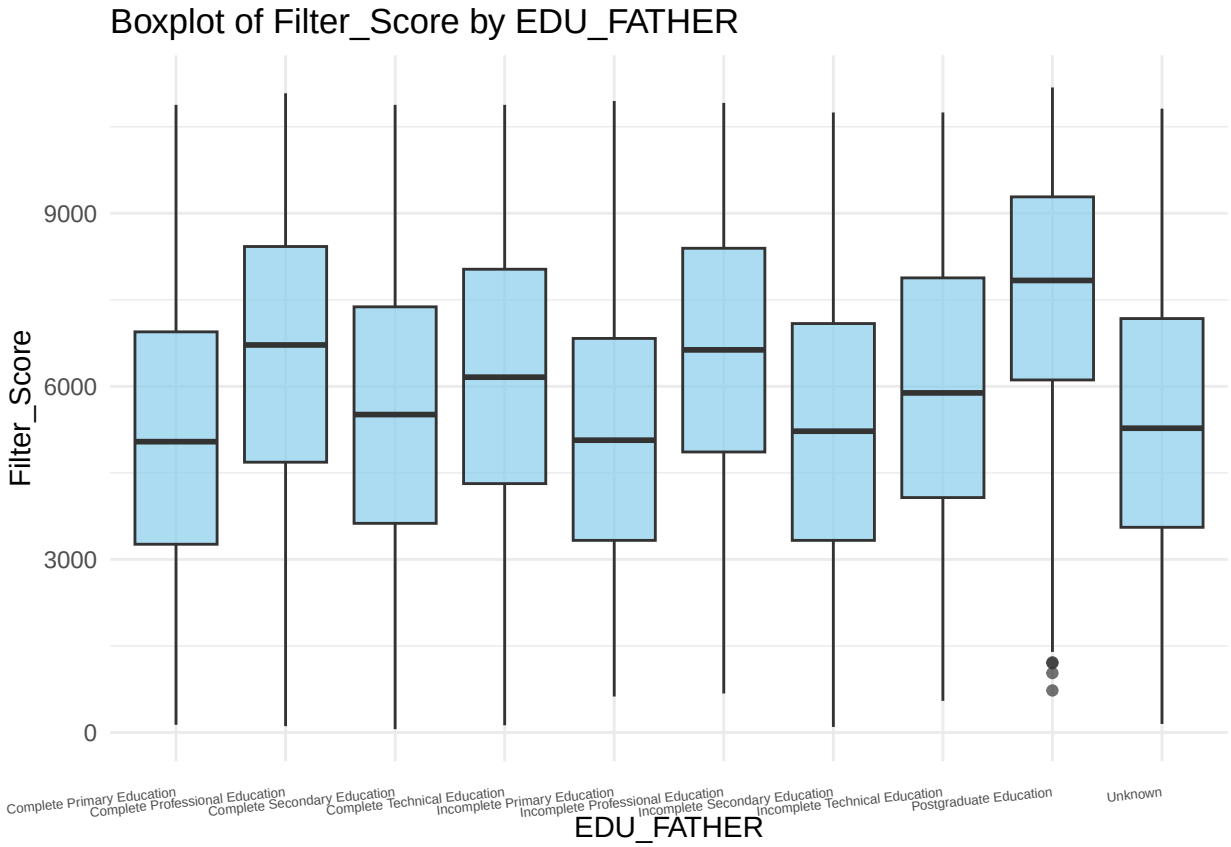
```
# Define the outcome variable and predictors
outcome_var <- "Filter_Score"
predictors <- c("GENDER", "EDU_FATHER", "EDU_MOTHER", "OCC_FATHER",
                "OCC_MOTHER", "STRATUM", "MOBILE", "UNIVERSITY", "PEOPLE_HOUSE")


# Loop through predictors to generate plots
for (pred in predictors) {
  if (is.numeric(new_data[[pred]])) {
    # For numerical predictors, use a scatter plot with a regression line
    p <- ggplot(new_data, aes_string(x = pred, y = outcome_var)) +
      geom_point(alpha = 0.6) +
      geom_smooth(method = "lm", color = "blue", se = FALSE) +
      labs(title = paste("Scatter Plot of", pred, "vs", outcome_var),
           x = pred, y = outcome_var) +
      theme_minimal()
```

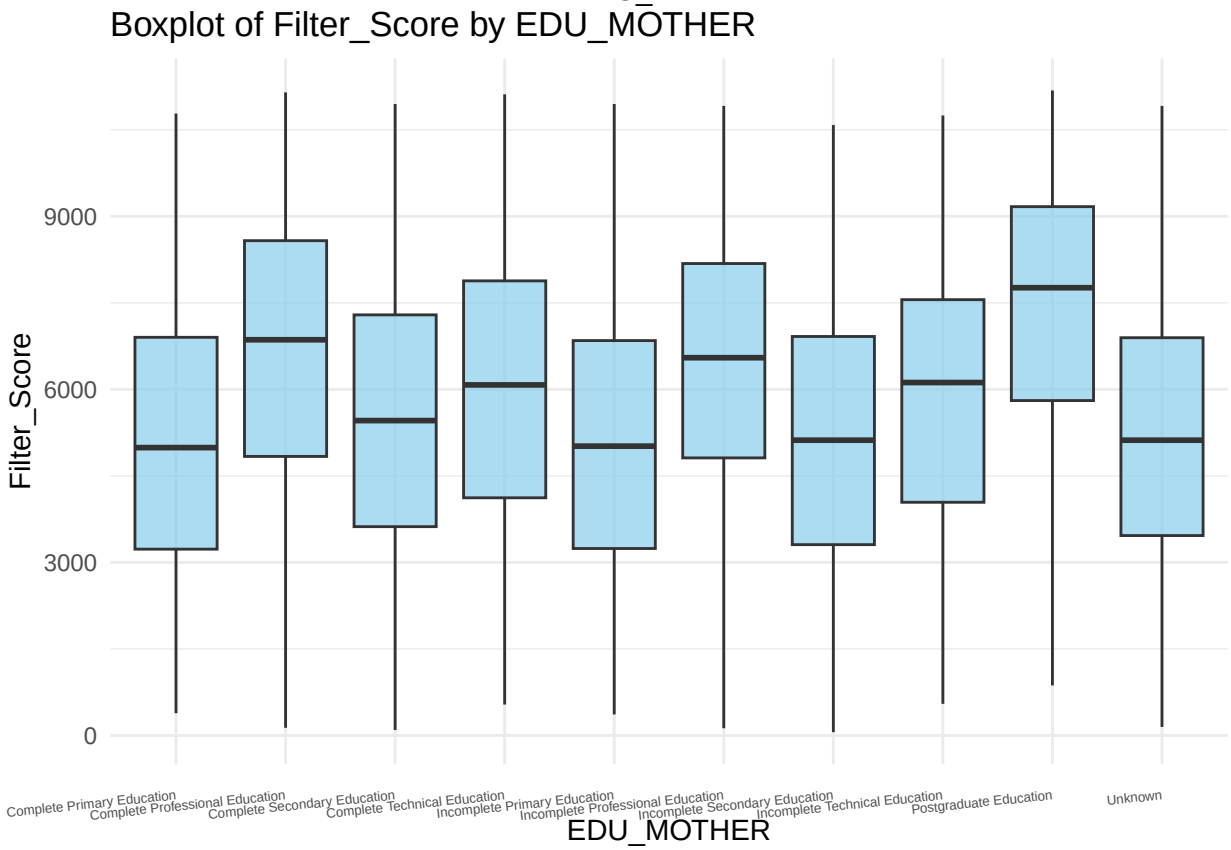
```
  } else {
    # For categorical predictors, use a boxplot
    p <- ggplot(new_data, aes_string(x = pred, y = outcome_var)) +
      geom_boxplot(fill = "skyblue", alpha = 0.7) +
      labs(title = paste("Boxplot of", outcome_var, "by", pred),
           x = pred, y = outcome_var) +
      theme_minimal() + theme(axis.text.x = element_text(angle = 6, vjust = 0.5, hjust=1, size = 5))
  }

  # Add plot to the list
  print(p)
}
```

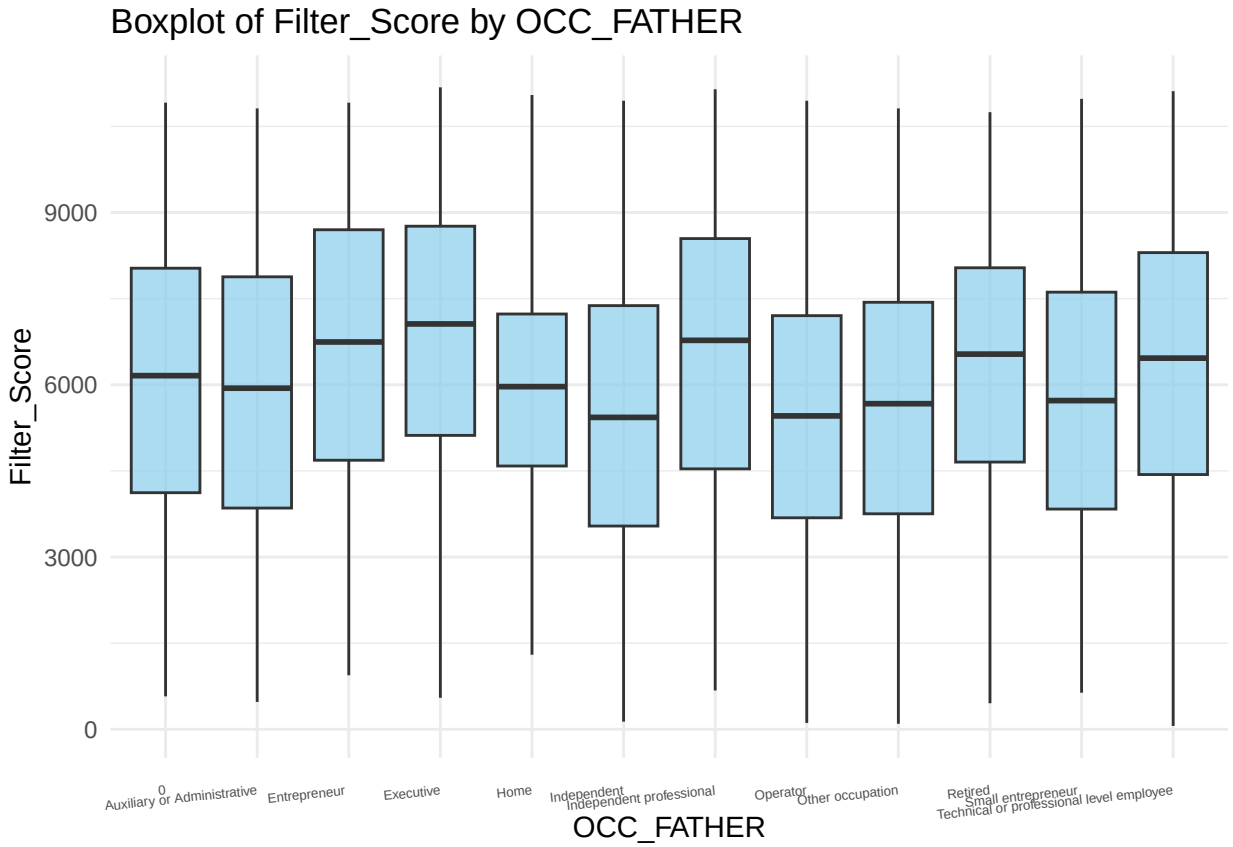
```
## Warning: `aes_string()` was deprecated in ggplot2 3.0.0.
## i Please use tidy evaluation idioms with `aes()`.
## i See also `vignette("ggplot2-in-packages")` for more information.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

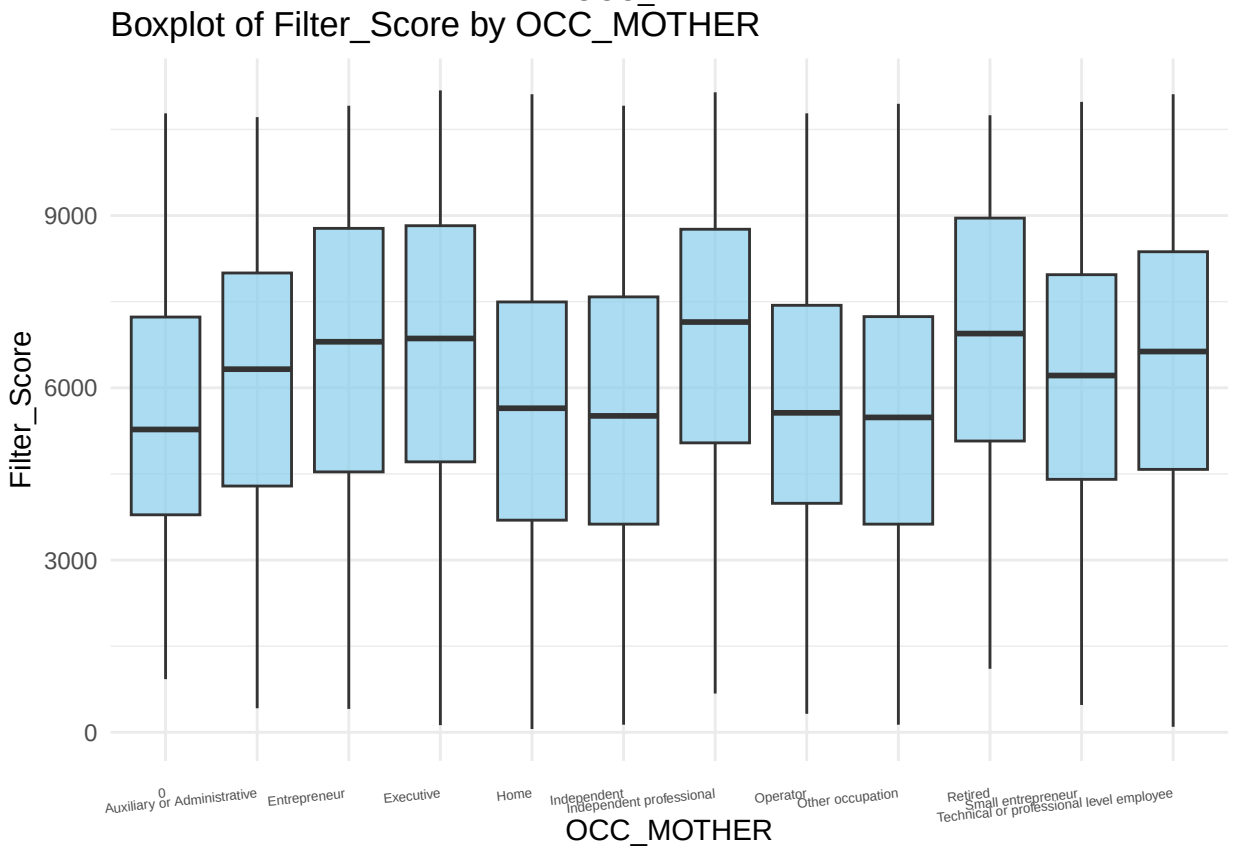
```
## `geom_smooth()` using formula = 'y ~ x'
```

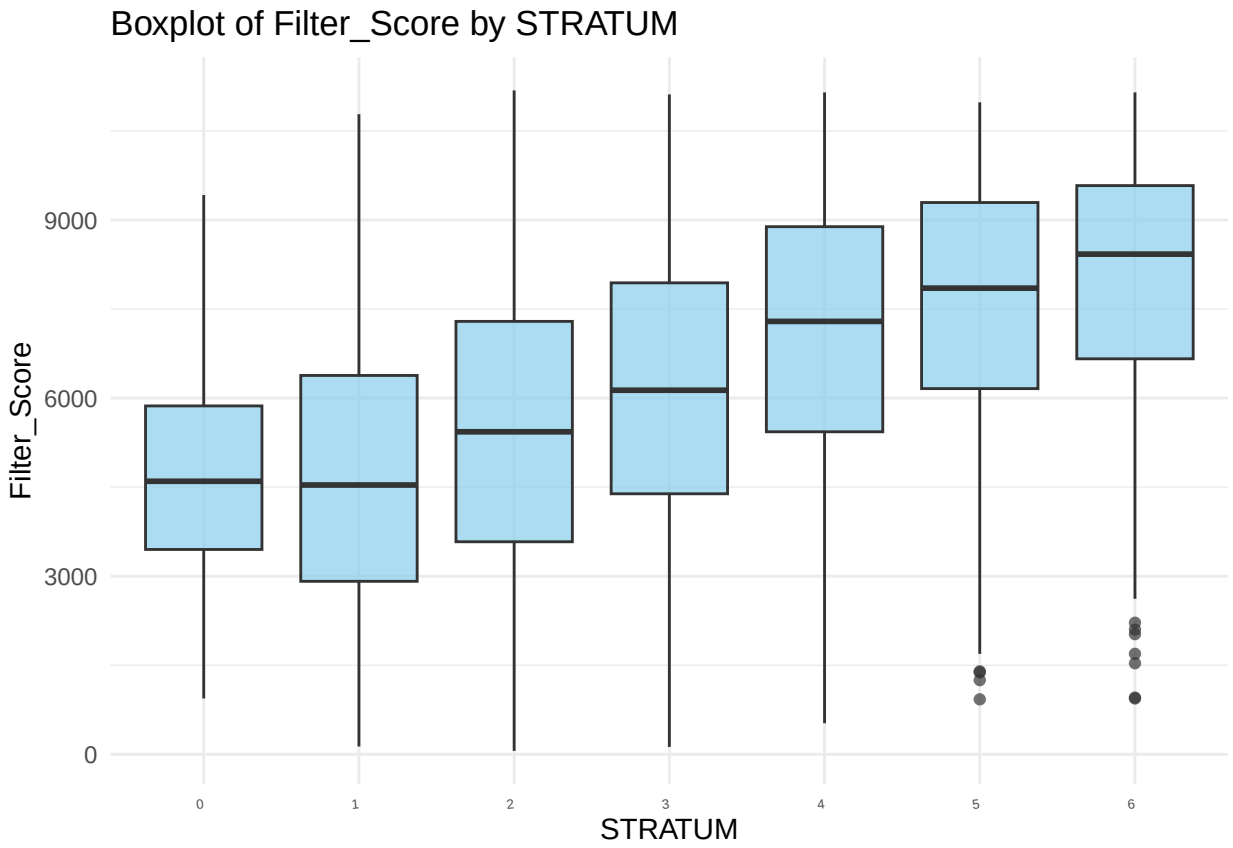



Boxplot of Filter_Score by EDU_FATHER
Filter_Score
9000
6000
3000
0
Complete Primary Education
Complete Professional Education
Complete Secondary Education
Complete Technical Education
Incomplete Primary Education
Incomplete Professional Education
Incomplete Secondary Education
Incomplete Technical Education
Postgraduate Education
Unknown
EDU_FATHER

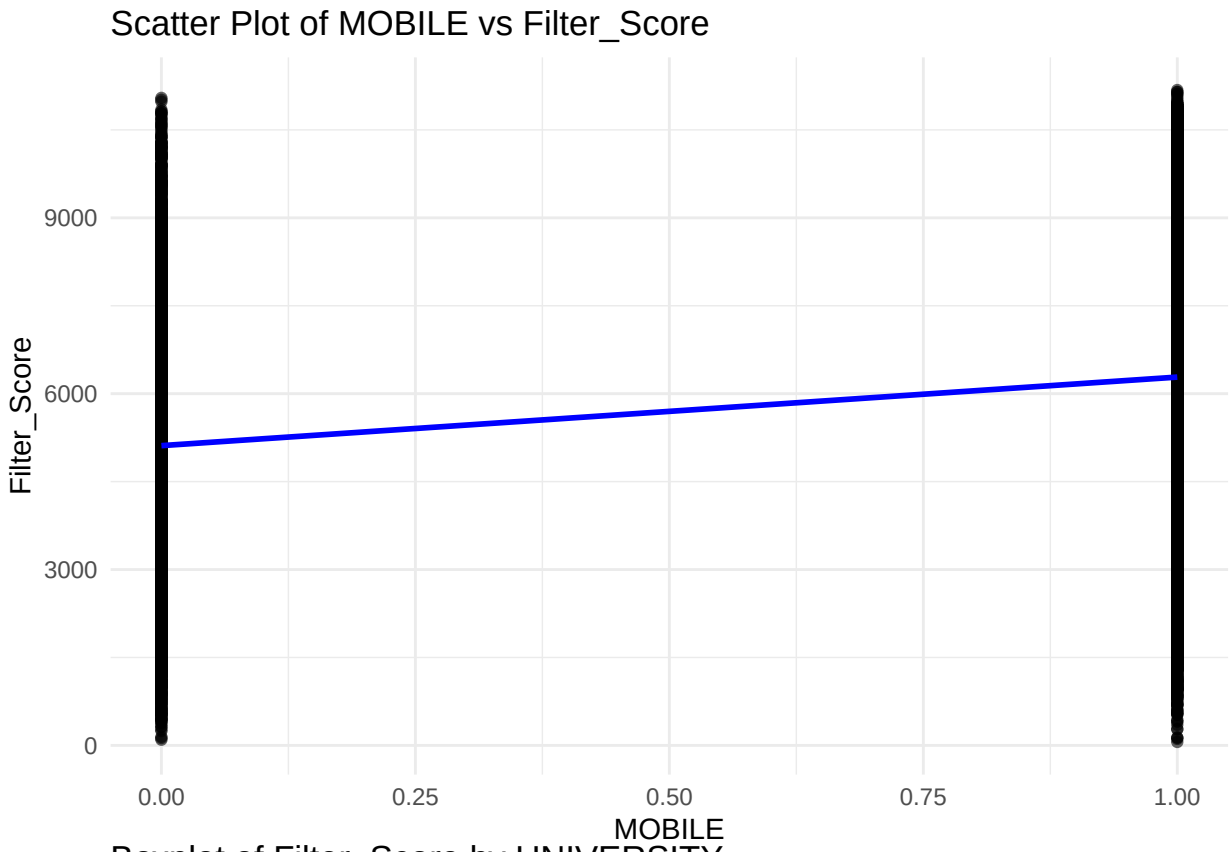

Boxplot of Filter_Score by EDU_MOTHER
Filter_Score
9000
6000
3000
0
Complete Primary Education
Complete Professional Education
Complete Secondary Education
Complete Technical Education
Incomplete Primary Education
Incomplete Professional Education
Incomplete Secondary Education
Incomplete Technical Education
Postgraduate Education
Unknown
EDU_MOTHER

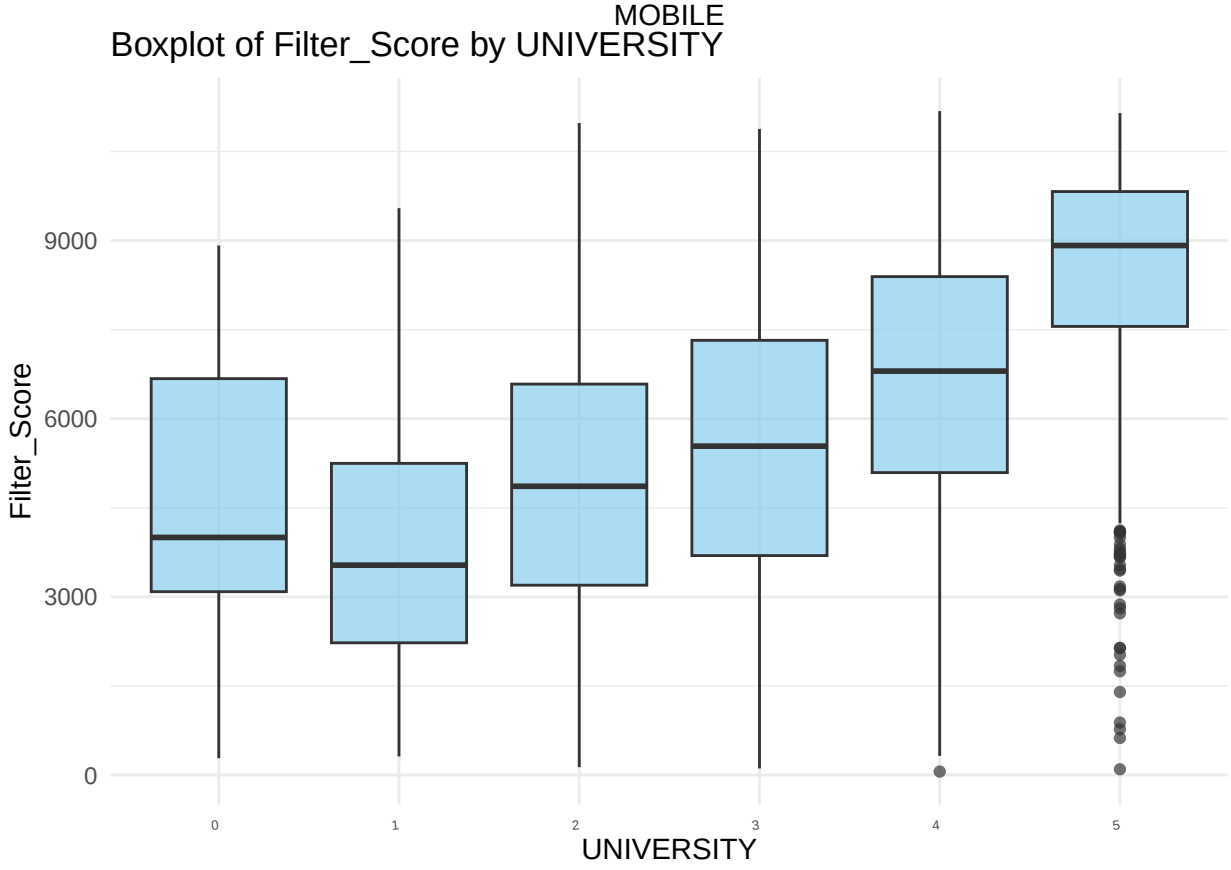

Boxplot of Filter_Score by OCC_FATHER
Filter_Score
9000
6000
3000
0
0
Auxiliary or Administrative
Entrepreneur
Executive
Home
Independent
Independent professional
Operator
Other occupation
Retired
Small entrepreneur
Technical or professional level employee
OCC_FATHER

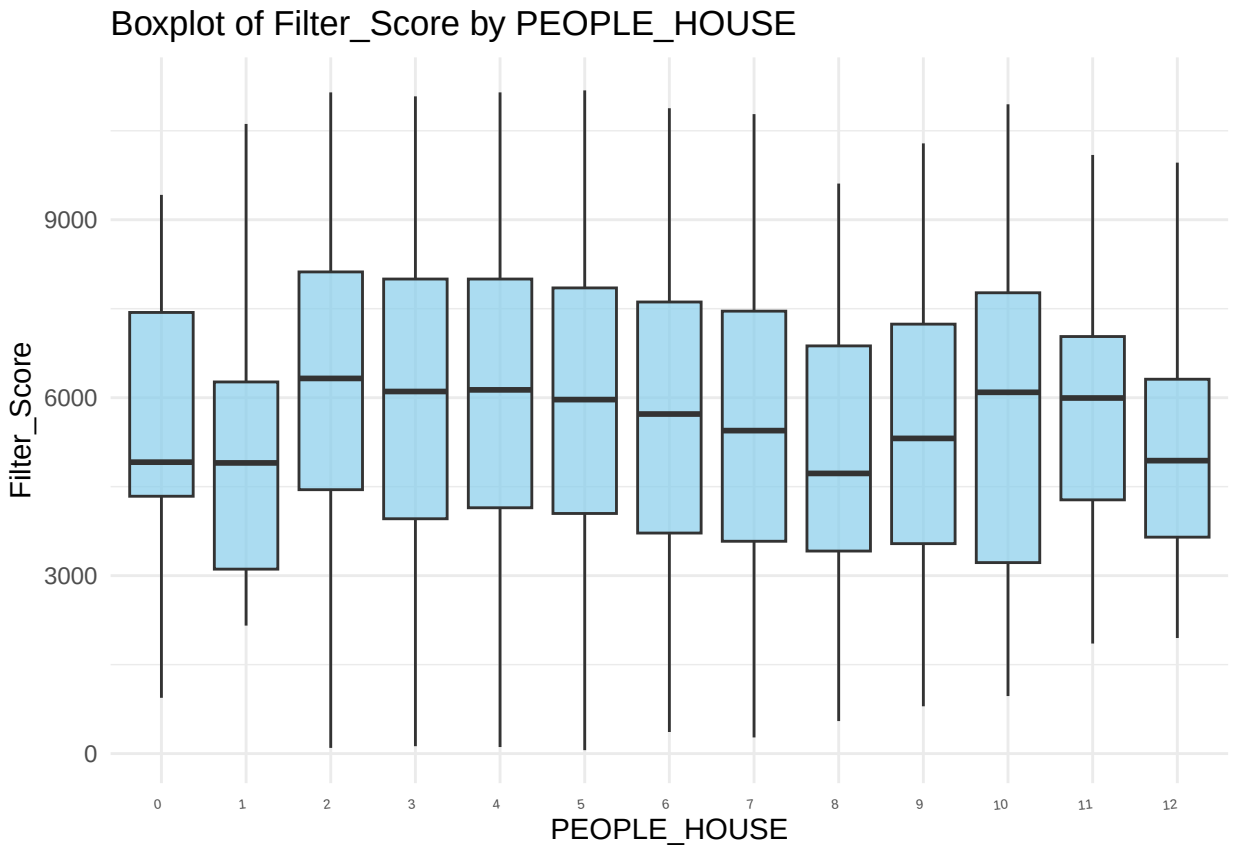

Boxplot of Filter_Score by OCC_MOTHER
Filter_Score
9000
6000
3000
0
0
Auxiliary or Administrative
Entrepreneur
Executive
Home
Independent
Independent professional
Operator
Other occupation
Retired
Small entrepreneur
Technical or professional level employee
OCC_MOTHER




```
## `geom_smooth()` using formula = 'y ~ x'
```


Scatter Plot of MOBILE vs Filter_Score
Filter_Score
9000
6000
3000
0
0.00
0.25
0.50
0.75
1.00
MOBILE


Boxplot of Filter_Score by UNIVERSITY
Filter_Score
9000
6000
3000
0
0
1
2
3
4
5
UNIVERSITY


Boxplot of Filter_Score by PEOPLE_HOUSE
Filter_Score
9000
6000
3000
0
0
1
2
3
4
5
6
7
8
9
10
11
12
PEOPLE_HOUSE