MVA Project - Data Cleaning, EDA, tests

Chun-Jung Chen & Akshay Arora

Data Preparation

Descriptive Statistics

```
summary(d3)
```

```
##  school   sex          age        address famsize   Pstatus      Medu
##  GP:331   F:195   Min.   :15.00   R: 81   GT3:266   A: 38   Min.   :0.0
##  MS: 39   M:175   1st Qu.:16.00   U:289   LE3:104   T:332   1st Qu.:2.0
##                   Median :17.00                             Median :3.0
##                   Mean   :16.58                             Mean   :2.8
##                   3rd Qu.:17.00                             3rd Qu.:4.0
##                   Max.   :22.00                             Max.   :4.0
##       Fedu             Mjob           Fjob     internet    guardian
##  Min.   :0.000   at_home : 53   at_home : 16   no : 57   father: 88
##  1st Qu.:2.000   health  : 33   health  : 17   yes:313   mother:266
##  Median :3.000   other   :134   other   :205             other : 16
##  Mean   :2.557   services: 93   services:103
##  3rd Qu.:3.750   teacher : 57   teacher : 29
##  Max.   :4.000
##    traveltime    famsup    romantic      famrel         freetime
##  Min.   : 1.00   no :139   no :251   Min.   :1.000   Min.   :1.000
##  1st Qu.: 5.00   yes:231   yes:119   1st Qu.:4.000   1st Qu.:3.000
##  Median :11.00                       Median :4.000   Median :3.000
##  Mean   :15.62                       Mean   :3.935   Mean   :3.224
##  3rd Qu.:23.00                       3rd Qu.:5.000   3rd Qu.:4.000
##  Max.   :89.00                       Max.   :5.000   Max.   :5.000
##    FirstMath       SecondMath      FinalMath       FirstPort
##  Min.   : 3.00   Min.   : 0.00   Min.   : 0.00   Min.   : 0.00
##  1st Qu.: 8.00   1st Qu.: 9.00   1st Qu.: 8.00   1st Qu.:10.00
##  Median :11.00   Median :11.00   Median :11.00   Median :12.00
##  Mean   :10.89   Mean   :10.75   Mean   :10.46   Mean   :12.14
##  3rd Qu.:13.00   3rd Qu.:13.00   3rd Qu.:14.00   3rd Qu.:14.00
##  Max.   :19.00   Max.   :19.00   Max.   :20.00   Max.   :19.00
##    SecondPort      FinalPort       FinalAvg
##  Min.   : 5.00   Min.   : 0.00   Min.   : 0.00
##  1st Qu.:11.00   1st Qu.:11.00   1st Qu.: 9.50
##  Median :12.00   Median :13.00   Median :11.50
##  Mean   :12.27   Mean   :12.55   Mean   :11.51
##  3rd Qu.:14.00   3rd Qu.:14.00   3rd Qu.:13.50
##  Max.   :19.00   Max.   :19.00   Max.   :18.50
```

Calculating Mode

- The names of the levels present below and the mode of each factor also illustrates below.

```
table(str(d3))
```

```
## 'data.frame':    370 obs. of  24 variables:
##  $ school    : Factor w/ 2 levels "GP","MS": 1 1 1 1 1 1 1 1 1 1 ...
##  $ sex       : Factor w/ 2 levels "F","M": 1 1 1 1 1 1 1 1 1 1 ...
##  $ age       : int  15 15 15 15 15 15 15 15 15 15 ...
##  $ address   : Factor w/ 2 levels "R","U": 1 1 1 1 1 1 1 1 1 2 ...
##  $ famsize   : Factor w/ 2 levels "GT3","LE3": 1 1 1 1 1 1 1 2 2 1 ...
##  $ Pstatus   : Factor w/ 2 levels "A","T": 2 2 2 2 2 2 2 2 2 1 ...
##  $ Medu      : int  1 1 2 2 3 3 3 2 3 3 ...
##  $ Fedu      : int  1 1 2 4 3 4 4 2 1 3 ...
##  $ Mjob      : Factor w/ 5 levels "at_home","health",..: 1 3 1 4 4 4 4 2 3 3 ...
##  $ Fjob      : Factor w/ 5 levels "at_home","health",..: 3 3 3 2 4 2 5 4 3 2 ...
##  $ internet  : Factor w/ 2 levels "no","yes": 2 2 1 2 2 2 2 2 2 1 ...
##  $ guardian  : Factor w/ 3 levels "father","mother",..: 2 2 2 2 3 2 1 2 1 1 ...
##  $ traveltime: num  23 11 6 1 19 3 23 23 29 8 ...
##  $ famsup    : Factor w/ 2 levels "no","yes": 2 2 2 2 2 2 2 2 2 1 ...
##  $ romantic  : Factor w/ 2 levels "no","yes": 1 2 1 1 2 1 2 1 1 1 ...
##  $ famrel    : int  3 3 4 4 4 4 4 4 4 4 ...
##  $ freetime  : int  1 3 3 3 2 3 2 1 4 3 ...
##  $ FirstMath : int  7 8 14 10 10 12 12 8 16 10 ...
##  $ SecondMath: int  10 6 13 9 10 12 0 9 16 11 ...
##  $ FinalMath : int  10 5 13 8 10 11 0 8 16 11 ...
##  $ FirstPort : int  13 13 14 10 13 11 10 11 15 10 ...
##  $ SecondPort: int  13 11 13 11 13 12 11 10 15 10 ...
##  $ FinalPort : int  13 11 12 10 13 12 12 11 15 10 ...
##  $ FinalAvg  : num  11.5 8 12.5 9 11.5 11.5 6 9.5 15.5 10.5 ...
```

```
## < table of extent 0 >
```

- variables' Mode

school	sex	age	address	famsize	Pstatus
"GP"	"F"	"16"	"U"	"GT3"	"T"
Medu	**Fedu**	**Mjob**	**Fjob**	**internet**	**guardian**
"4"	"2"	"other"	"other"	"yes"	"mother"
traveltime	**famsup**	**romantic**	**famrel**	**freetime**	**FirstMath**
" 3"	"yes"	"no"	"4"	"3"	"10"
SecondMath	**FinalMath**	**FirstPort**	**SecondPort**	**FinalPort**	**FinalAvg**
" 9"	"10"	"12"	"12"	"13"	"12.0"

Question 1: Deos Family conditions affect students' final grade in Math?

Type	Variable	Descriptions
DV	FinalMath	Final Grade in Math
ID	famsize	Family Size
ID	Pstatus	Parents' Cohabitation
ID	famrel	Quality of family relationships
ID	famsup	Family's educational support

Descriptive Statistics

```
##    FinalMath     famsize   Pstatus     famrel      famsup
##  Min.   : 0.00   GT3:266   A: 38   Min.   :1.000   no :139
##  1st Qu.: 8.00   LE3:104   T:332   1st Qu.:4.000   yes:231
##  Median :11.00                     Median :4.000
##  Mean   :10.46                     Mean   :3.935
##  3rd Qu.:14.00                     3rd Qu.:5.000
##  Max.   :20.00                     Max.   :5.000
```

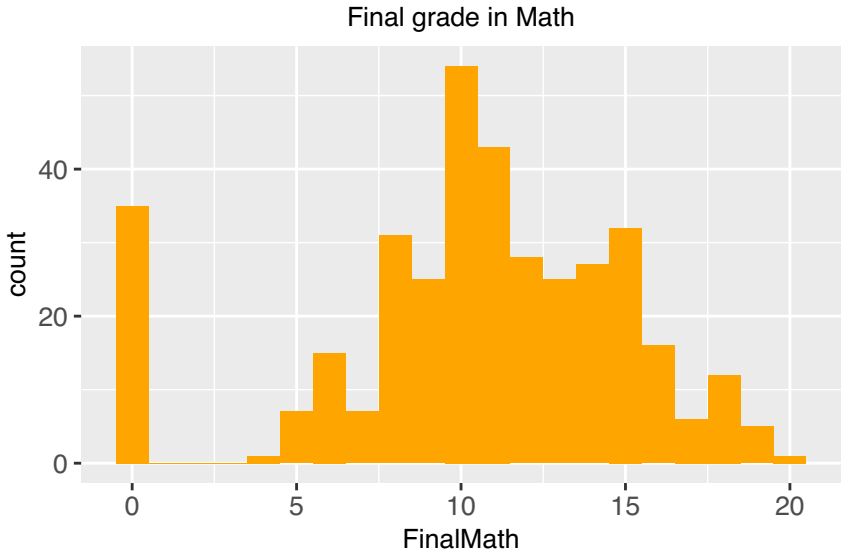
Final Math

- Variance of Final Math Score

```
var(d3.q1$FinalMath)
```

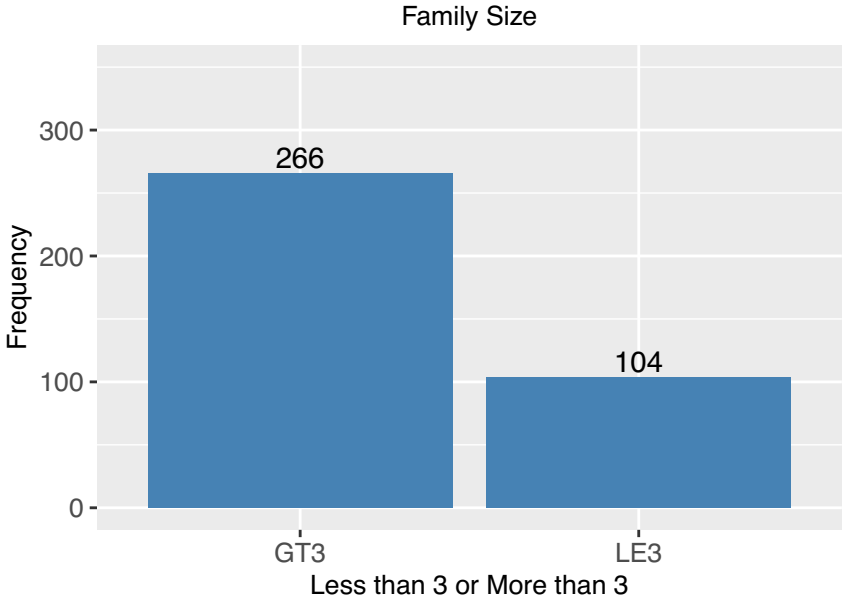
```
## [1] 21.24131
```

- Histogram of Final grade in Math:
  - The histogram presents bell shape so it is likely to follow normal distribution.



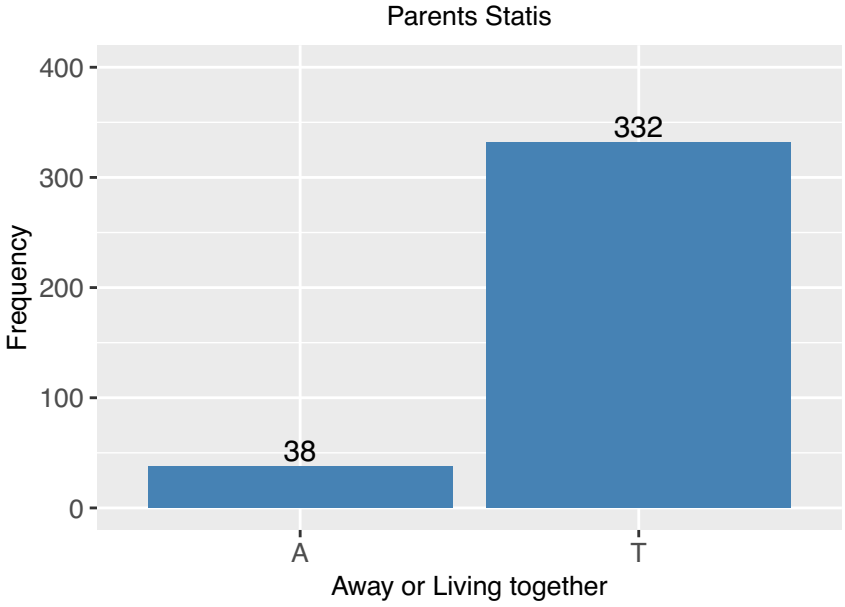
Family Size

- Most of the family size are greater than 3 family members



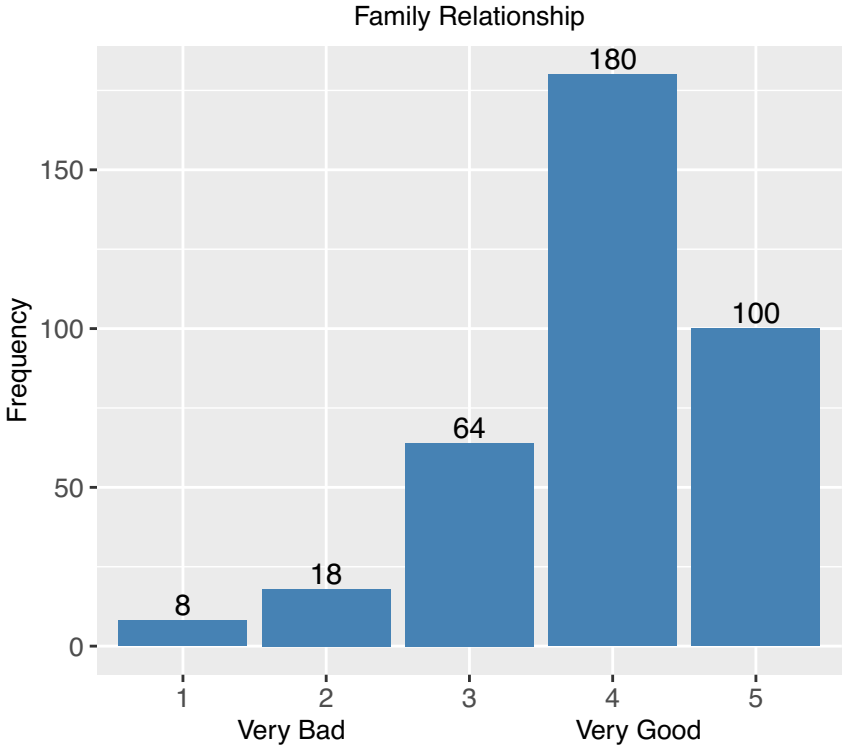
Parents Status

- Most of the family are living together



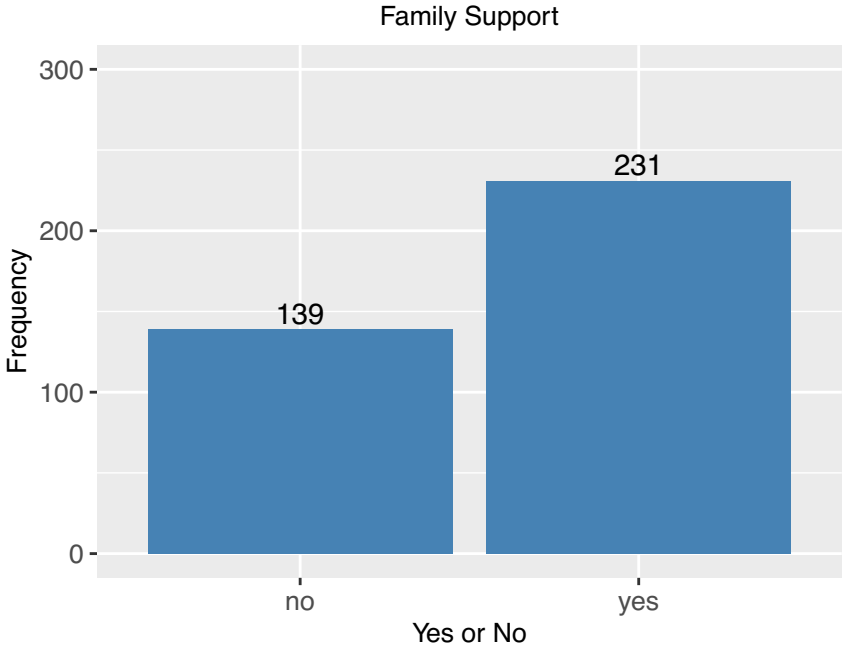
Family Relationship

- Most of the students rated their family relationship for 4 out of 5. it can be considered as a good family relationship.



Family Support

- Over 60% of students received family support



Question 2: Does parents' jobs and education level influence students' first period of grade in Math?

Type	Variable	Descriptions
DV	FirstMath	First Grade in Math
ID	Medu	Mother's Education Level
ID	Fedu	Father's Education Level
ID	Mjob	Mother's Job
ID	Fjob	Father's Job

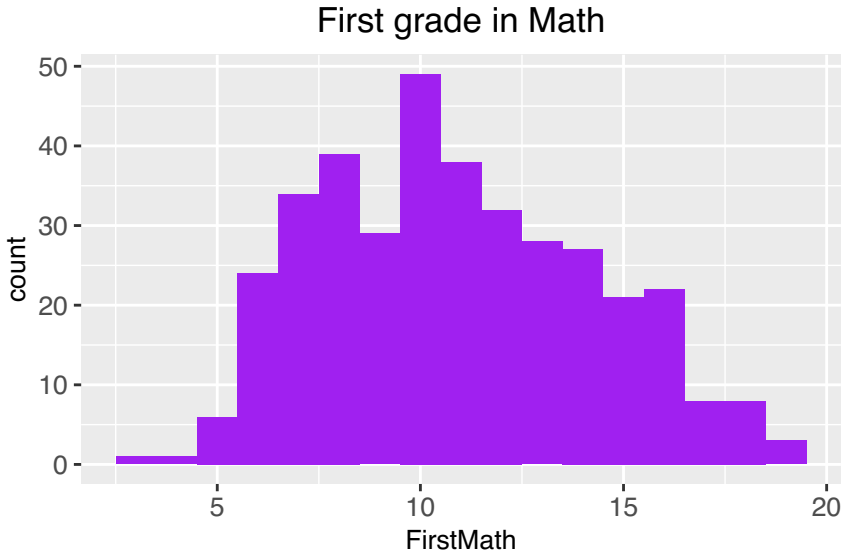
Descriptive Statistics

```
##    FirstMath          Medu          Fedu             Mjob
##  Min.   : 3.00   Min.   :0.0   Min.   :0.000   at_home : 53
##  1st Qu.: 8.00   1st Qu.:2.0   1st Qu.:2.000   health  : 33
##  Median :11.00   Median :3.0   Median :3.000   other   :134
##  Mean   :10.89   Mean   :2.8   Mean   :2.557   services: 93
##  3rd Qu.:13.00   3rd Qu.:4.0   3rd Qu.:3.750   teacher : 57
##  Max.   :19.00   Max.   :4.0   Max.   :4.000
```

```
##        Fjob
##  at_home : 16
##  health  : 17
##  other   :205
##  services:103
##  teacher : 29
##
```

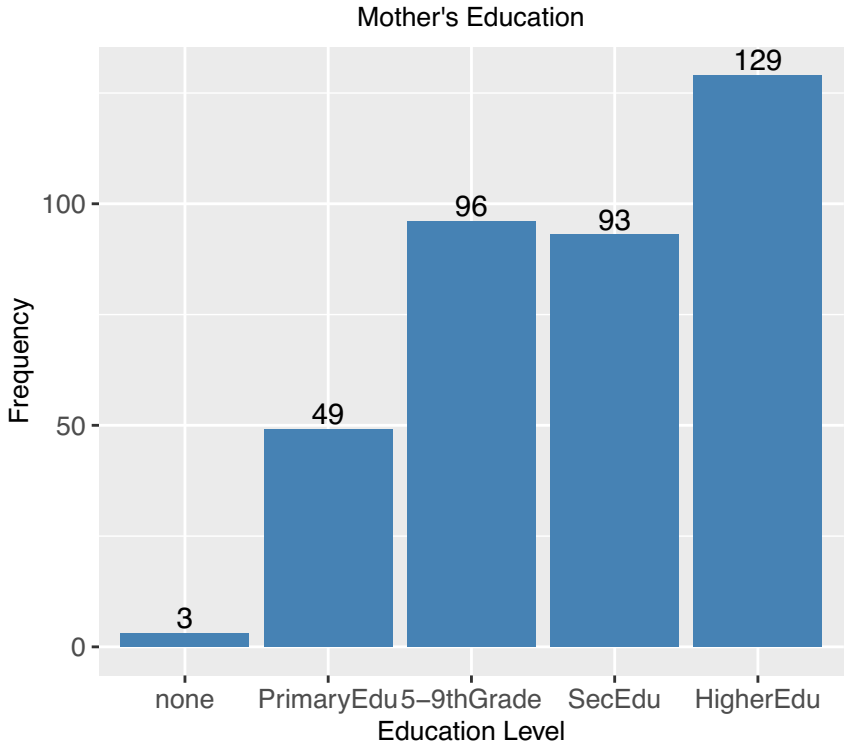
First grade in Math

- The data pattern is normally distributed.

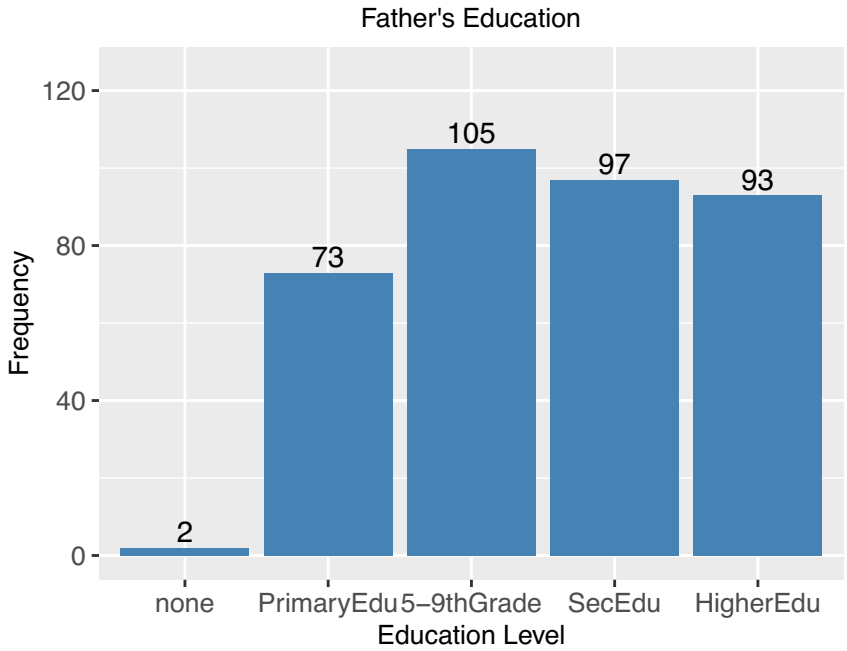


Parents Education

1. Mother's education
   * Majority of Mother received higher education



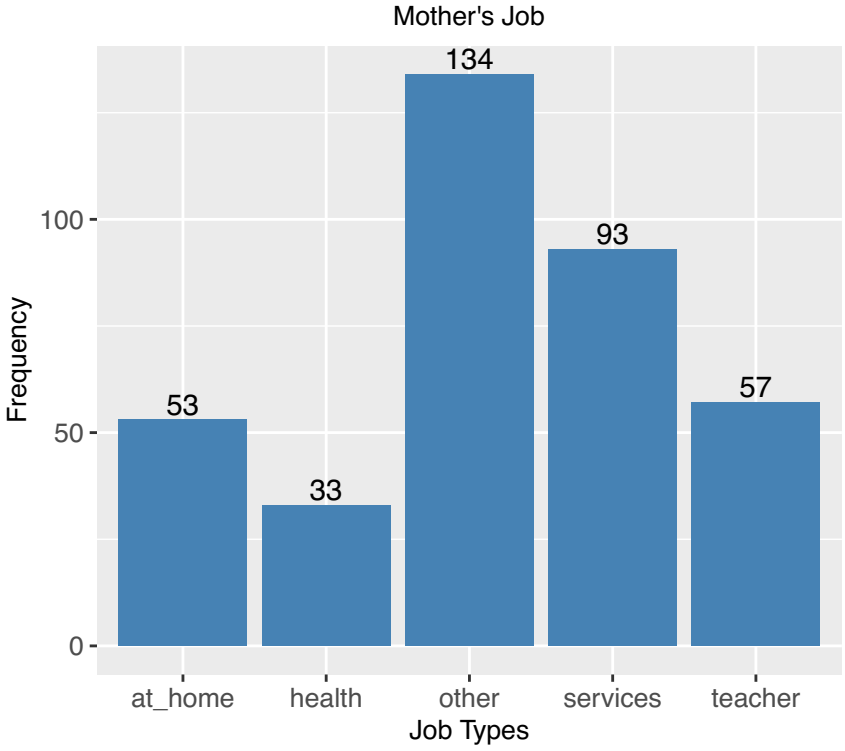
2. Father's education
   * Comparing to mother's education level, it is surprised that most father only received 5-9 grade.



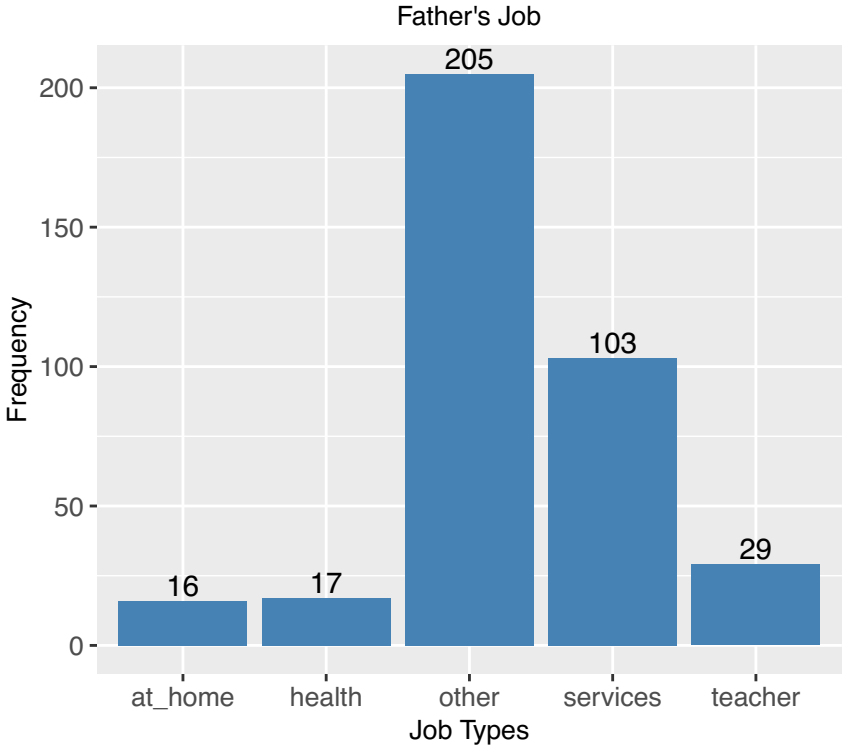
Parents' Job

- Besides other types of job, most of the parents are working in services industry

```
1. Mother's Job
```




```
2. Father's Job
```



Question 3: Does student's learning conditions really impact students' final grade math score and Portuguese scores in average?

Type	Variable	Descriptions
DV	FinalAvg	Average of final grade math score and Portuguese scores
ID	internet	Internet access
ID	traveltime	Home to school travel time
ID	romantic	Romantic relationship
ID	freetime	Free time after school

Descriptive Statistics

```
##     FinalAvg     internet    traveltime    romantic     freetime
##  Min.   : 0.00   no : 57   Min.   : 1.00   no :251   Min.   :1.000
##  1st Qu.: 9.50   yes:313   1st Qu.: 5.00   yes:119   1st Qu.:3.000
##  Median :11.50             Median :11.00             Median :3.000
##  Mean   :11.51             Mean   :15.62             Mean   :3.224
##  3rd Qu.:13.50             3rd Qu.:23.00             3rd Qu.:4.000
##  Max.   :18.50             Max.   :89.00             Max.   :5.000
```

```
cov(d3.q3$FinalAvg, d3.q3$traveltime)
```

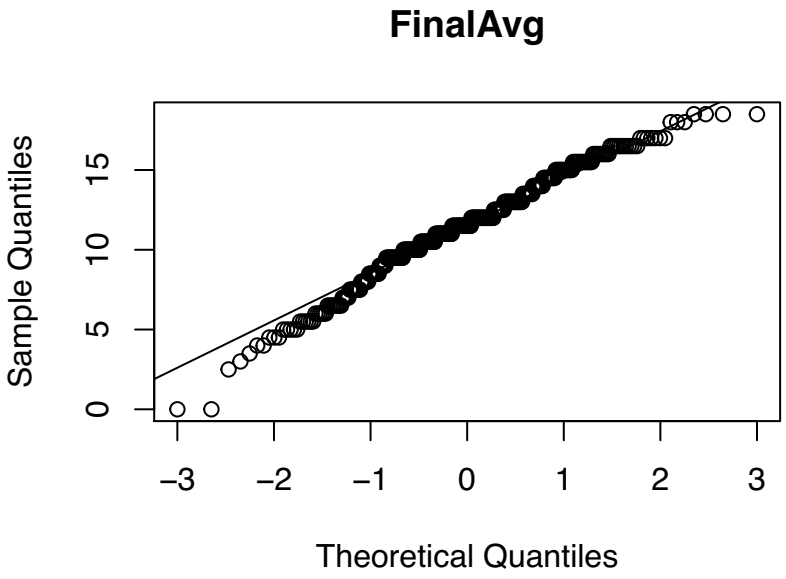
```
## [1] -2.98252
```

FinalAverage

```
var(d3.q3$FinalAvg)
```

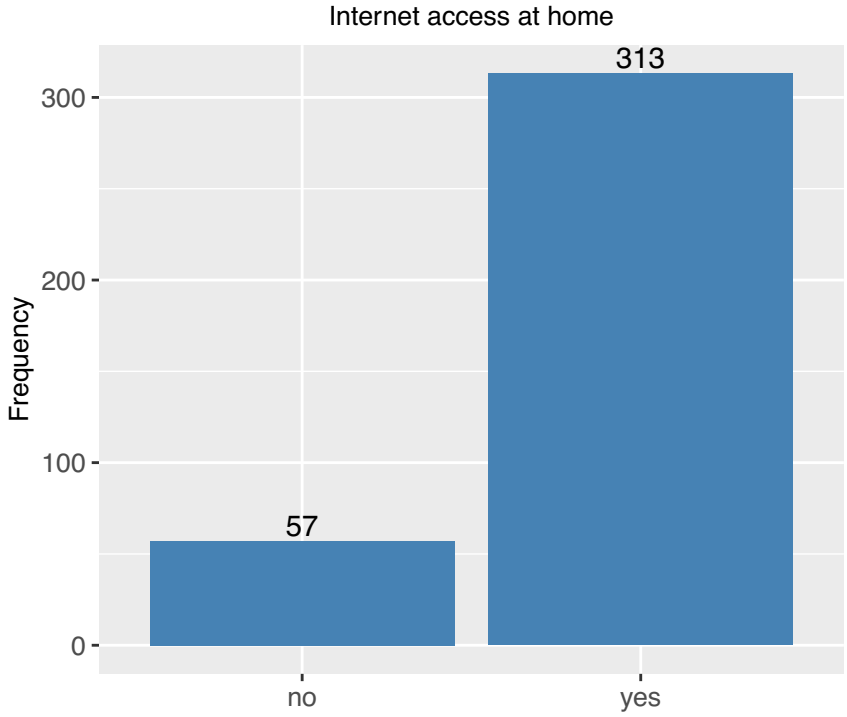
```
## [1] 10.82173
```

- According to the qqplot, most of the points lay on the line, means that there does not exist extreme outliers and data follows normal distribution.



Internet

- The family with internet access is 5 times more than the family without it.

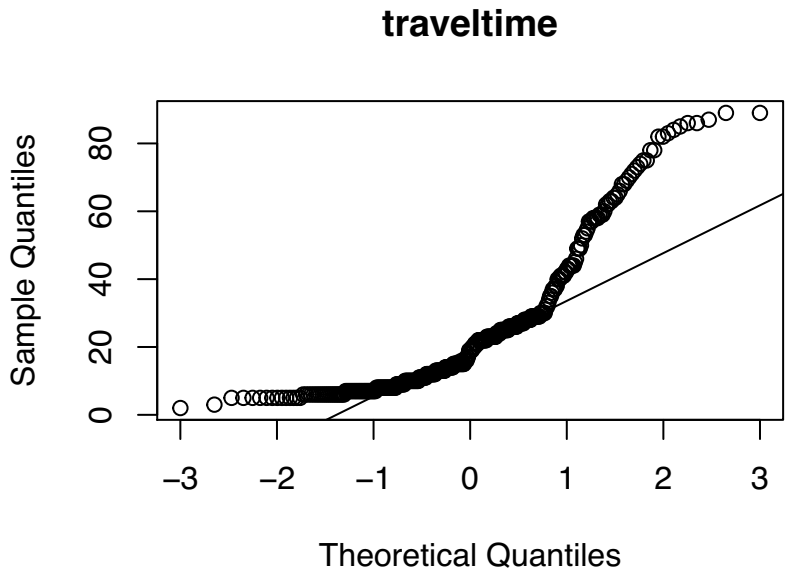


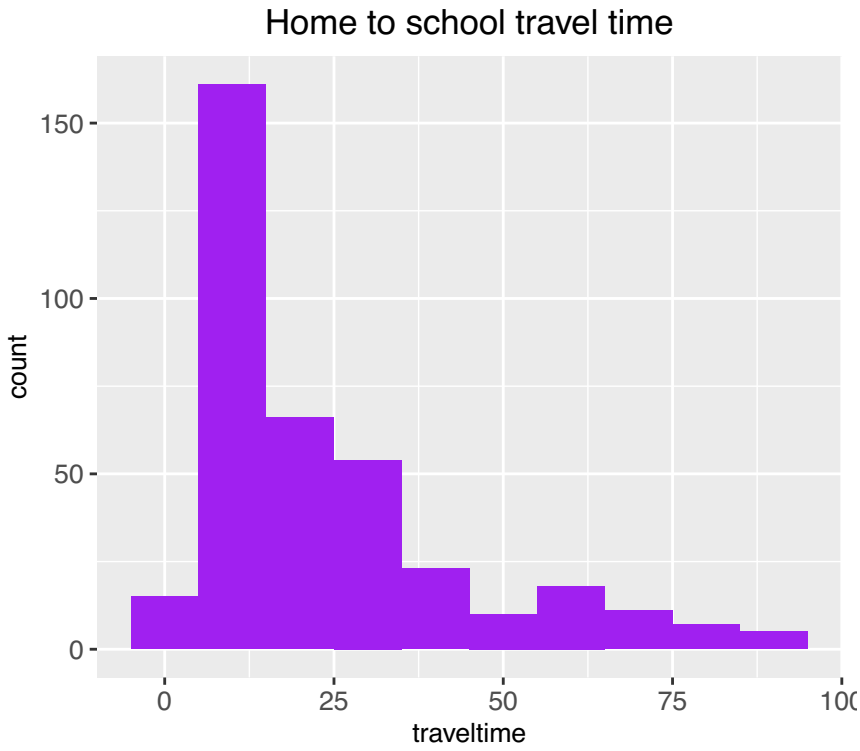
Travel time

- According to the qq plot, the data is extremely **right skewed**, the normal transformation is necessary and presented below by box-cox transforamtion.

```
var(d3.q3$traveltime)
```

```
## [1] 396.59
```



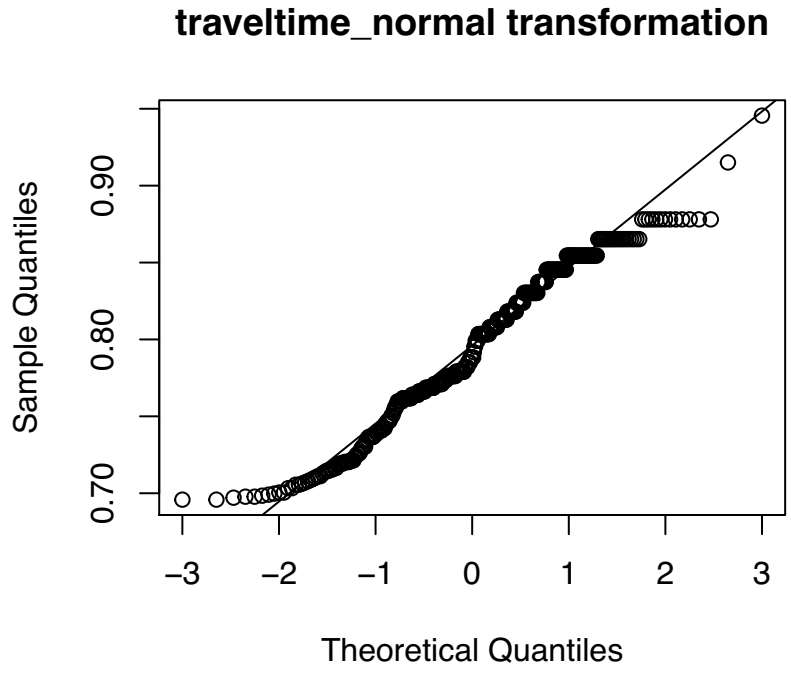


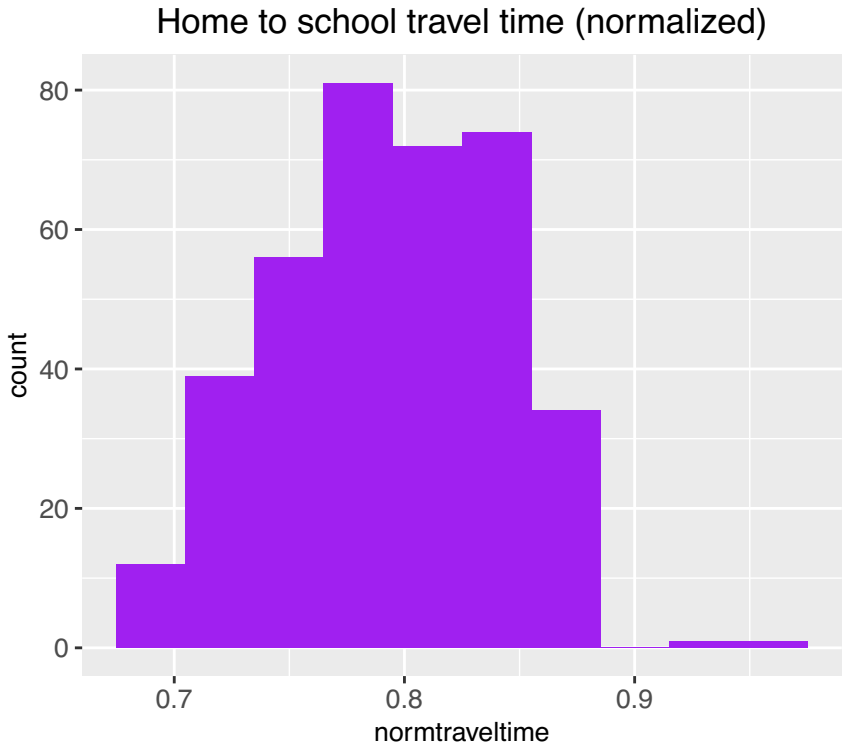
- Shapiro test for normality
  - The test result proves that travel time data is not normally distributed.

```
shapiro.test(d3.q3$traveltime)  #P-value << 0.05. not significantly normal distributed
```

```
##
##  Shapiro-Wilk normality test
##
## data:  d3.q3$traveltime
## W = 0.82559, p-value < 2.2e-16
```

- Power transformation
  - After power transformation, the qqplot shows that data is close to normal distribution and can be used for statistical inference.





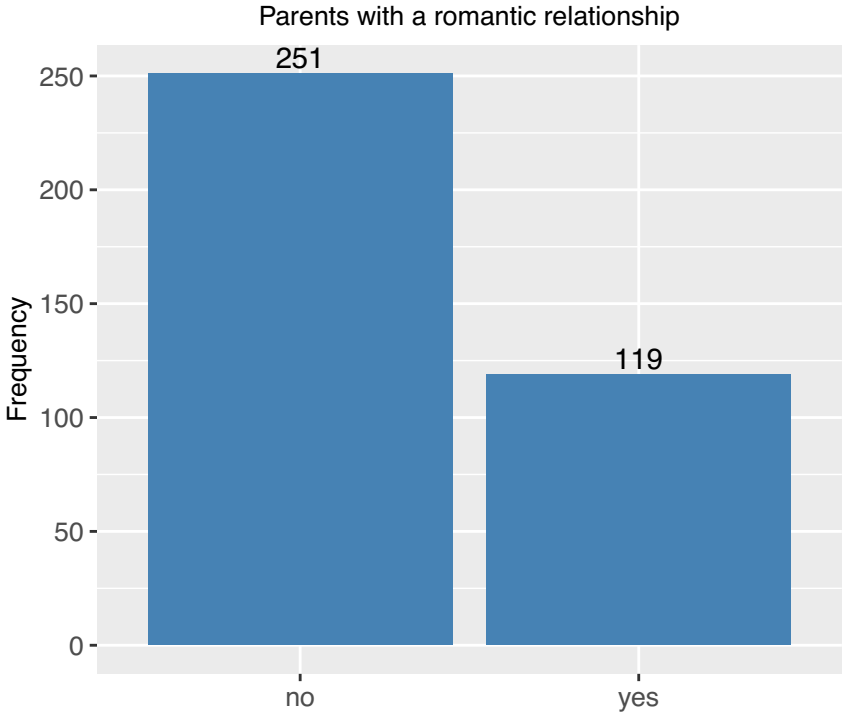
- T test on travel time and internet assess
  - P-value of the t test shows that there is not difference of travel time between family with or without internet access.

```
set.seed(20201001)
with(data = d3.q3, t.test(normtraveltime[internet == "yes"],
    normtraveltime[internet == "no"], var.equal = TRUE))
```

```
##
##  Two Sample t-test
##
## data:  normtraveltime[internet == "yes"] and normtraveltime[internet == "no"]
## t = -0.075648, df = 368, p-value = 0.9397
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.01471347  0.01362336
## sample estimates:
## mean of x mean of y
## 0.7934071 0.7939522
```

Romantic

- twice of the parents are not in the romantic relationship than other parents in the romantic relationship



Free time (free time after school)

- Majority of students score their level of free time as 3 out of 5, means that they found normal for their spare time.