# Assignment1-R

January 20, 2020

## 1 Assignment 1 - Data preprocessing and manual introspection

Load the data file Wage.rds or Wage.csv:

[2]:
```r
load(file = "../ISLR/data/Wage.rda")
```

Display the number of features and their names:

[3]:
```r
dim(Wage)[2]
names(Wage)
```

11

1. ’year’ 2. ’age’ 3. ’maritl’ 4. ’race’ 5. ’education’ 6. ’region’ 7. ’jobclass’ 8. ’health’ 9. ’health_ins’ 10. ’logwage’ 11. ’wage’

Delete the feature ‘logwage’ and display the number of features and their names again:

[11]:
```r
drops <- c("logwage")
Wage <- Wage[ , !(names(Wage) %in% drops)]
dim(Wage)[2]
names(Wage)
```

10

1. ’year’ 2. ’age’ 3. ’maritl’ 4. ’race’ 5. ’education’ 6. ’region’ 7. ’jobclass’ 8. ’health’ 9. ’health_ins’ 10. ’wage’

Display the number of data points:

[12]:
```r
dim(Wage)[1]
```

3000

Display the data in a table (subset of rows is sufficient):

[13]:
```r
Wage
```

A data.frame: 3000 × 10

| | year | age | maritl | race | education | region |
|---|---|---|---|---|---|---|
| | <int> | <int> | <fct> | <fct> | <fct> | <fct> |
| 231655 | 2006 | 18 | 1. Never Married | 1. White | 1. < HS Grad | 2. Mid |
| 86582 | 2004 | 24 | 1. Never Married | 1. White | 4. College Grad | 2. Mid |
| 161300 | 2003 | 45 | 2. Married | 1. White | 3. Some College | 2. Mid |
| 155159 | 2003 | 43 | 2. Married | 3. Asian | 4. College Grad | 2. Mid |
| 11443 | 2005 | 50 | 4. Divorced | 1. White | 2. HS Grad | 2. Mid |
| 376662 | 2008 | 54 | 2. Married | 1. White | 4. College Grad | 2. Mid |
| 450601 | 2009 | 44 | 2. Married | 4. Other | 3. Some College | 2. Mid |
| 377954 | 2008 | 30 | 1. Never Married | 3. Asian | 3. Some College | 2. Mid |
| 228963 | 2006 | 41 | 1. Never Married | 2. Black | 3. Some College | 2. Mid |
| 81404 | 2004 | 52 | 2. Married | 1. White | 2. HS Grad | 2. Mid |
| 302778 | 2007 | 45 | 4. Divorced | 1. White | 3. Some College | 2. Mid |
| 305706 | 2007 | 34 | 2. Married | 1. White | 2. HS Grad | 2. Mid |
| 8690 | 2005 | 35 | 1. Never Married | 1. White | 2. HS Grad | 2. Mid |
| 153561 | 2003 | 39 | 2. Married | 1. White | 4. College Grad | 2. Mid |
| 449654 | 2009 | 54 | 2. Married | 1. White | 2. HS Grad | 2. Mid |
| 447660 | 2009 | 51 | 2. Married | 1. White | 3. Some College | 2. Mid |
| 160191 | 2003 | 37 | 1. Never Married | 3. Asian | 4. College Grad | 2. Mid |
| 230312 | 2006 | 50 | 2. Married | 1. White | 5. Advanced Degree | 2. Mid |
| 301585 | 2007 | 56 | 2. Married | 1. White | 4. College Grad | 2. Mid |
| 153682 | 2003 | 37 | 1. Never Married | 1. White | 3. Some College | 2. Mid |
| 158226 | 2003 | 38 | 2. Married | 3. Asian | 4. College Grad | 2. Mid |
| 11141 | 2005 | 40 | 4. Divorced | 1. White | 2. HS Grad | 2. Mid |
| 448410 | 2009 | 75 | 2. Married | 1. White | 4. College Grad | 2. Mid |
| 305116 | 2007 | 40 | 2. Married | 1. White | 4. College Grad | 2. Mid |
| 233002 | 2006 | 38 | 1. Never Married | 1. White | 2. HS Grad | 2. Mid |
| 8684 | 2005 | 49 | 2. Married | 1. White | 5. Advanced Degree | 2. Mid |
| 229379 | 2006 | 43 | 2. Married | 1. White | 2. HS Grad | 2. Mid |
| 86064 | 2004 | 34 | 2. Married | 4. Other | 2. HS Grad | 2. Mid |
| 378472 | 2008 | 57 | 2. Married | 1. White | 2. HS Grad | 2. Mid |
| 157244 | 2003 | 18 | 1. Never Married | 2. Black | 2. HS Grad | 2. Mid |
| 304184 | 2007 | 59 | 2. Married | 3. Asian | 2. HS Grad | 2. Mid |
| 154351 | 2003 | 29 | 1. Never Married | 4. Other | 3. Some College | 2. Mid |
| 447182 | 2009 | 22 | 1. Never Married | 1. White | 2. HS Grad | 2. Mid |
| 13962 | 2005 | 54 | 2. Married | 1. White | 2. HS Grad | 2. Mid |
| 154728 | 2003 | 46 | 2. Married | 2. Black | 2. HS Grad | 2. Mid |
| 380298 | 2008 | 51 | 2. Married | 1. White | 2. HS Grad | 2. Mid |
| 230171 | 2006 | 35 | 1. Never Married | 1. White | 3. Some College | 2. Mid |
| 307415 | 2007 | 49 | 2. Married | 1. White | 2. HS Grad | 2. Mid |
| 161305 | 2003 | 53 | 2. Married | 1. White | 2. HS Grad | 2. Mid |
| 451605 | 2009 | 61 | 2. Married | 1. White | 3. Some College | 2. Mid |
| 301838 | 2007 | 40 | 2. Married | 2. Black | 1. < HS Grad | 2. Mid |
| 154752 | 2003 | 52 | 2. Married | 1. White | 1. < HS Grad | 2. Mid |
| 8804 | 2005 | 40 | 2. Married | 1. White | 4. College Grad | 2. Mid |
| 158531 | 2003 | 56 | 2. Married | 1. White | 1. < HS Grad | 2. Mid |
| 379706 | 2008 | 39 | 2. Married | 1. White | 2. HS Grad | 2. Mid |
| 306214 | 2007 | 30 | 2. Married | 1. White | 2. HS Grad | 2. Mid |
| 158084 | 2003 | 58 | 2. Married | 1. White | 3. Some College | 2. Mid |
| 305029 | 2007 | 33 | 2. Married | 3. Asian | 5. Advanced Degree | 2. Mid |
| 307412 | 2007 | 51 | 2. Married | 1. White | 5. Advanced Degree | 2. Mid |
| 377739 | 2008 | 32 | 1. Never Married | 1. White | 4. College Grad | 2. Mid |

Print a statistic summary of the features (year, age, maritl, race, education, region, jobclass, health, health_ins) and the label (wage):

[14]:
```
print(summary(Wage))
```

```
     year           age                     maritl           race
Min.   :2003   Min.   :18.00   1. Never Married: 648   1. White:2480
1st Qu.:2004   1st Qu.:33.75   2. Married      :2074   2. Black: 293
Median :2006   Median :42.00   3. Widowed      :  19   3. Asian: 190
Mean   :2006   Mean   :42.41   4. Divorced     : 204   4. Other:  37
3rd Qu.:2008   3rd Qu.:51.00   5. Separated    :  55
Max.   :2009   Max.   :80.00

             education                     region                jobclass
1. < HS Grad       :268   2. Middle Atlantic   :3000   1. Industrial :1544
2. HS Grad         :971   1. New England       :   0   2. Information:1456
3. Some College    :650   3. East North Central:   0
4. College Grad    :685   4. West North Central:   0
5. Advanced Degree:426   5. South Atlantic    :   0
                          6. East South Central:   0
                          (Other)              :   0
           health      health_ins        wage
1. <=Good     : 858   1. Yes:2083   Min.   : 20.09
2. >=Very Good:2142   2. No : 917   1st Qu.: 85.38
                                    Median :104.92
                                    Mean   :111.70
                                    3rd Qu.:128.68
                                    Max.   :318.34
```

**For the numerical features**, check the correlation, i.e., the relation of feature to lable variations. Therefore, for **each** such feature perform the following steps:

1. Plot the feature against the lable values
2. Test the normality of the feature and lable values
3. Test their correlation using an appropriate test
4. Interprete the results

In R, we need to download and install some libraries.

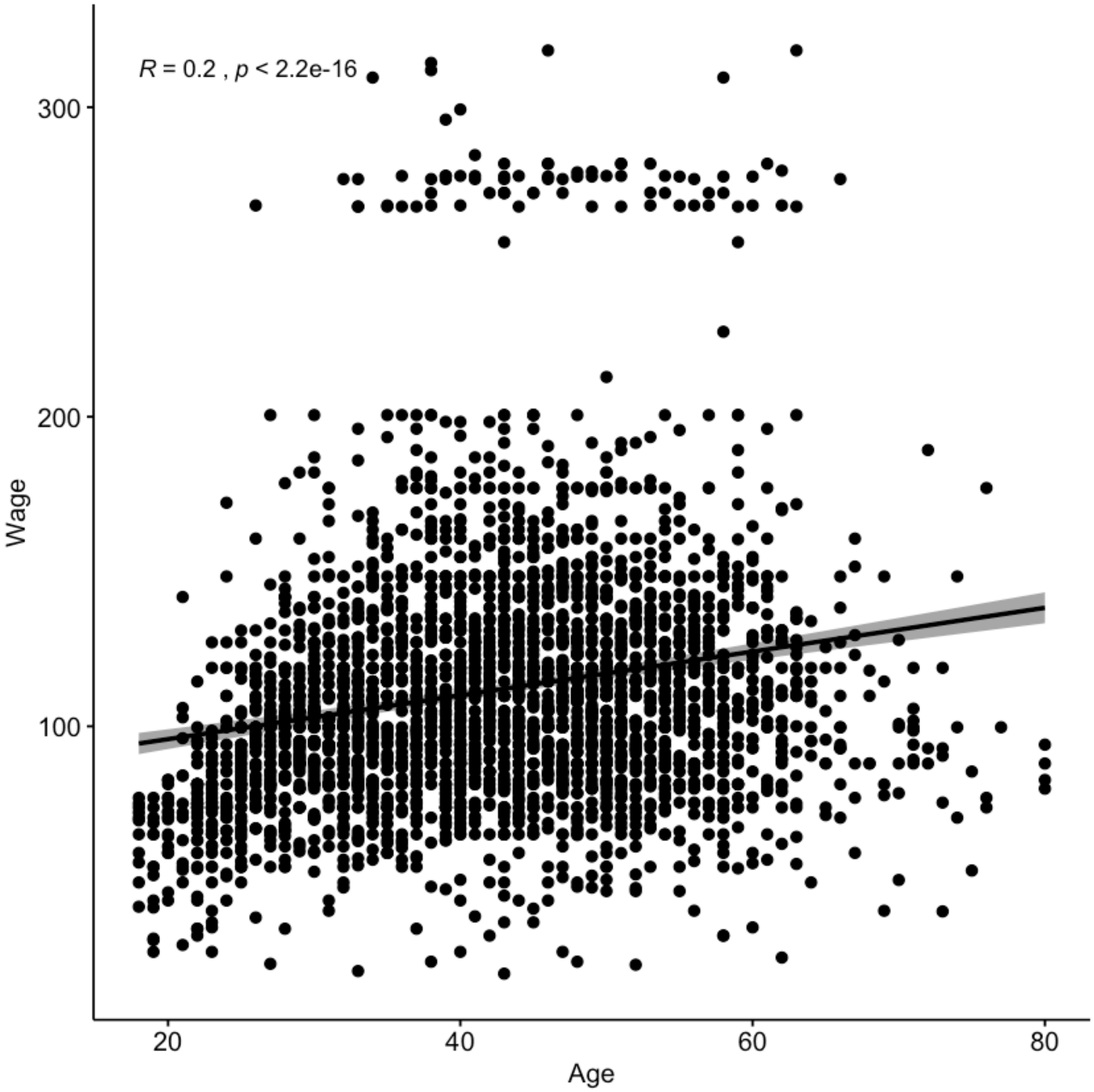
[15]:
```
install.packages("ggpubr")
library("ggpubr")
```

```
The downloaded binary packages are in
/var/folders/ct/4pcck8t94sdfc73rhymq4t140000gp/T//Rtmp57Ghca/downloaded_packages

Loading required package: ggplot2


Loading required package: magrittr
```

Step 1: Plot the feature against the lable values:

[16]:

```
ggscatter(Wage, x = "age", y = "wage",
          add = "reg.line", conf.int = TRUE,
          cor.coef = TRUE, cor.method = "pearson",
          xlab = "Age", ylab = "Wage")
```



Step 2: Test the normality of the feature and lable values:

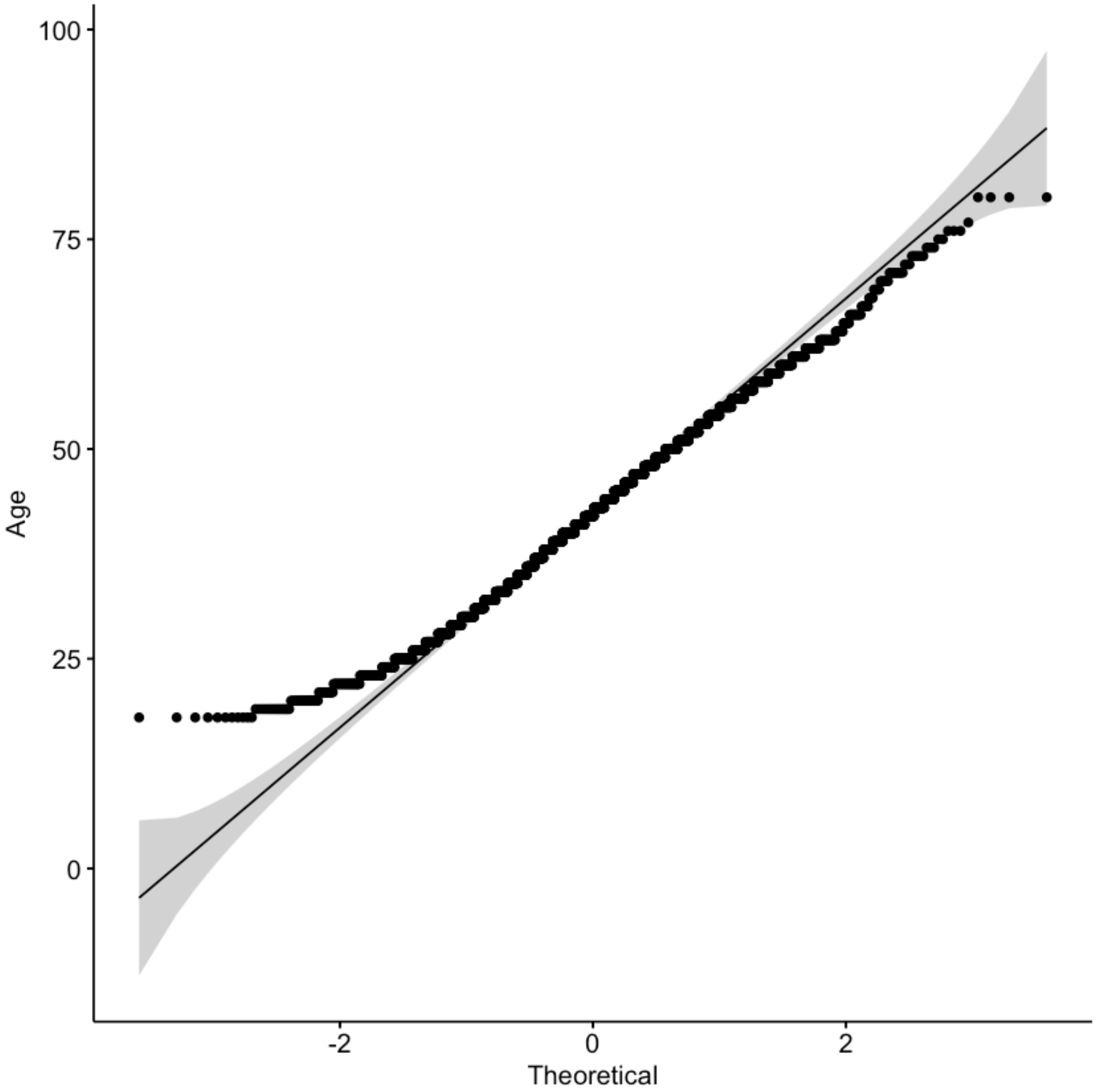
[17]:

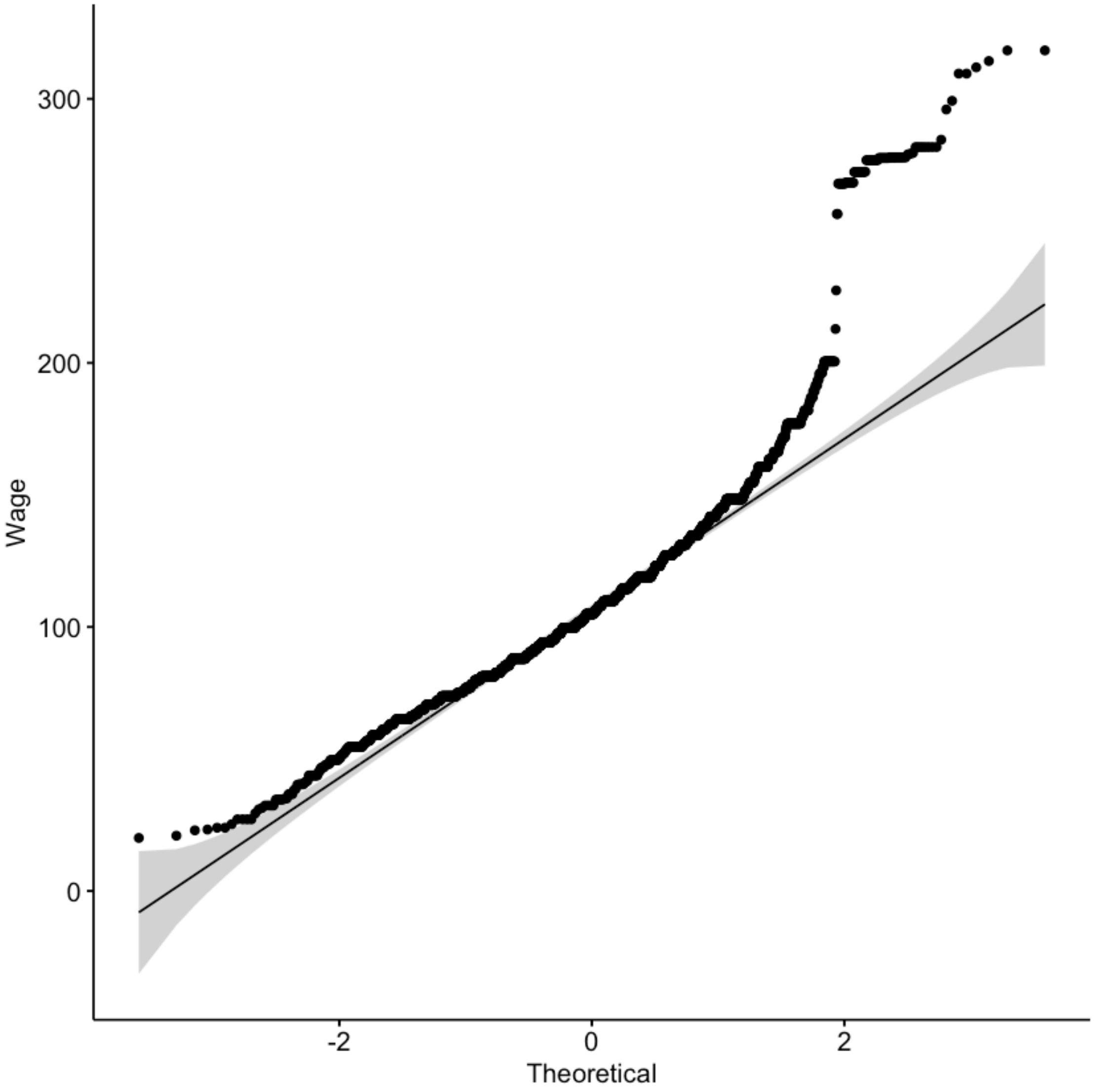
```
shapiro.test(Wage$age)
ggqqplot(Wage$age, ylab = "Age")
```

```
shapiro.test(Wage$wage)
ggqqplot(Wage$wage, ylab = "Wage")
```

```
Shapiro-Wilk normality test

data:  Wage$age
W = 0.99106, p-value = 9.416e-13
```

```
Shapiro-Wilk normality test

data:  Wage$wage
W = 0.87957, p-value < 2.2e-16
```





Step 3: Perform the Pearson correlation test:

[18]:

```
res <- cor.test(Wage$age, Wage$wage, method = "pearson")
res
```

```
Pearson's product-moment correlation

data:  Wage$age and Wage$wage
t = 10.923, df = 2998, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.1609777 0.2298147
sample estimates:
      cor
0.1956372
```

Step 4: *Your interpretation of results goes here!*

**For the non-numerical features**, analyse the variance (ANOVA), to study differences between the means of the label values for groups of data points with the same feature value. Therefore, for **each** such feature perform the following steps:

1. List the possible feature values
2. Plot (box plot) the label values for each group of of data points with the same feature value.
3. Perform the one way ANOVA test
4. Interprete the results

Step 1: List the possible feature values:

[19]:

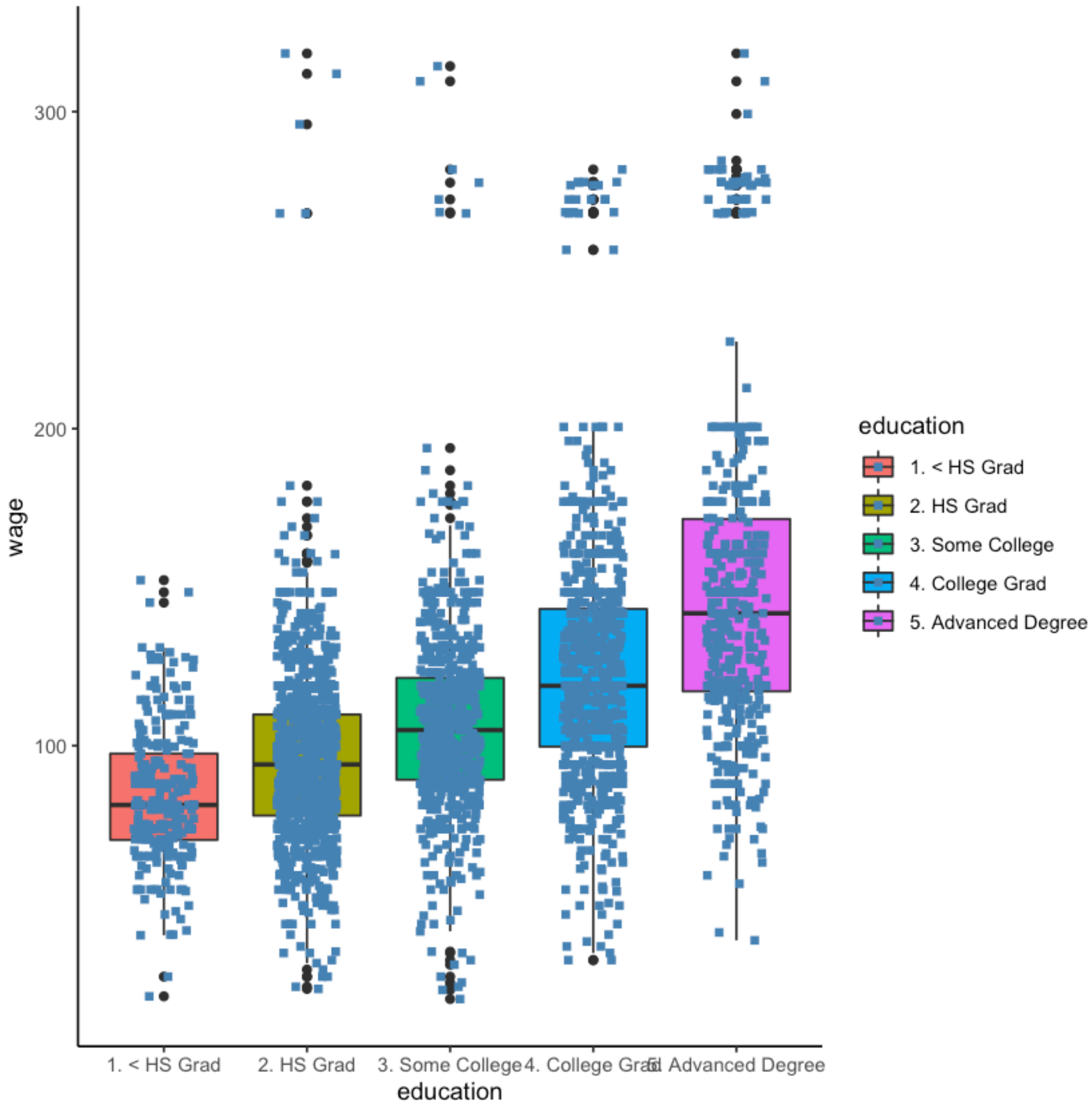
```r
levels(Wage$education)
```

1. ’1. < HS Grad’ 2. ’2. HS Grad’ 3. ’3. Some College’ 4. ’4. College Grad’ 5. ’5. Advanced Degree’

Step 2: Plot (box plot) the label values for each group of of data points with the same feature value:

[20]:

```r
ggplot(Wage, aes(x = education, y = wage, fill = education)) +
    geom_boxplot() +
    geom_jitter(shape = 15,
        color = "steelblue",
        position = position_jitter(0.21)) +
    theme_classic()
```



Step 3. Perform the one-way ANOVA test:

```
[150]: anova_one_way <- aov(wage~education, data = Wage)
       summary(anova_one_way)
```

```
                 Df  Sum Sq Mean Sq F value Pr(>F)    
education         4 1226364  306591   229.8 <2e-16 ***
Residuals      2995 3995721    1334                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Step 4: *Your interpretation of results goes here!*