Data Analysis
Week 3 Task Solutions

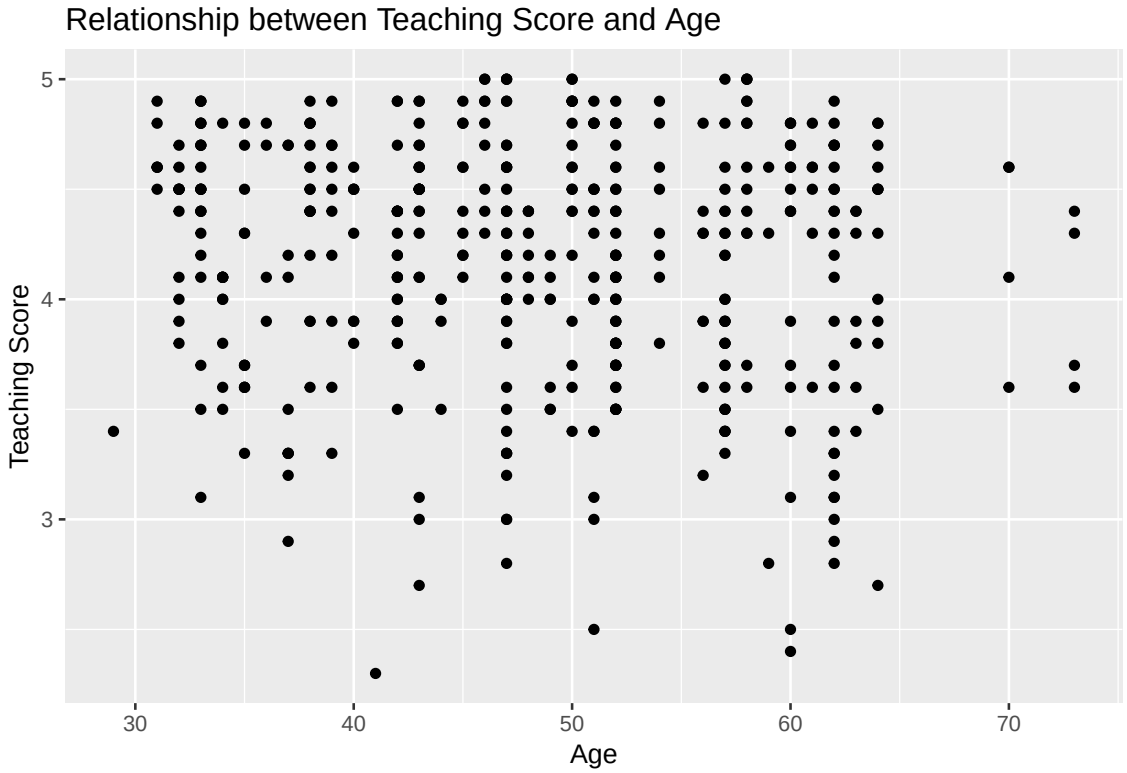
Tasks

1. Examine the relationship between teaching score and age in the `evals` data set. What is the value of the correlation coefficient? How would you interpret this verbally? Finally, produce a scatterplot of teaching score and age.

```
evals.age <- evals %>%
  select(score, age)
evals.age %>%
  get_correlation(formula = score ~ age)
```

```
# A tibble: 1 x 1
  correlation
        <dbl>
1      -0.107
```

```
ggplot(evals.age, aes(x = age, y = score)) +
  geom_point() +
  labs(x = "Age", y = "Teaching Score",
       title = "Relationship between Teaching Score and Age")
```

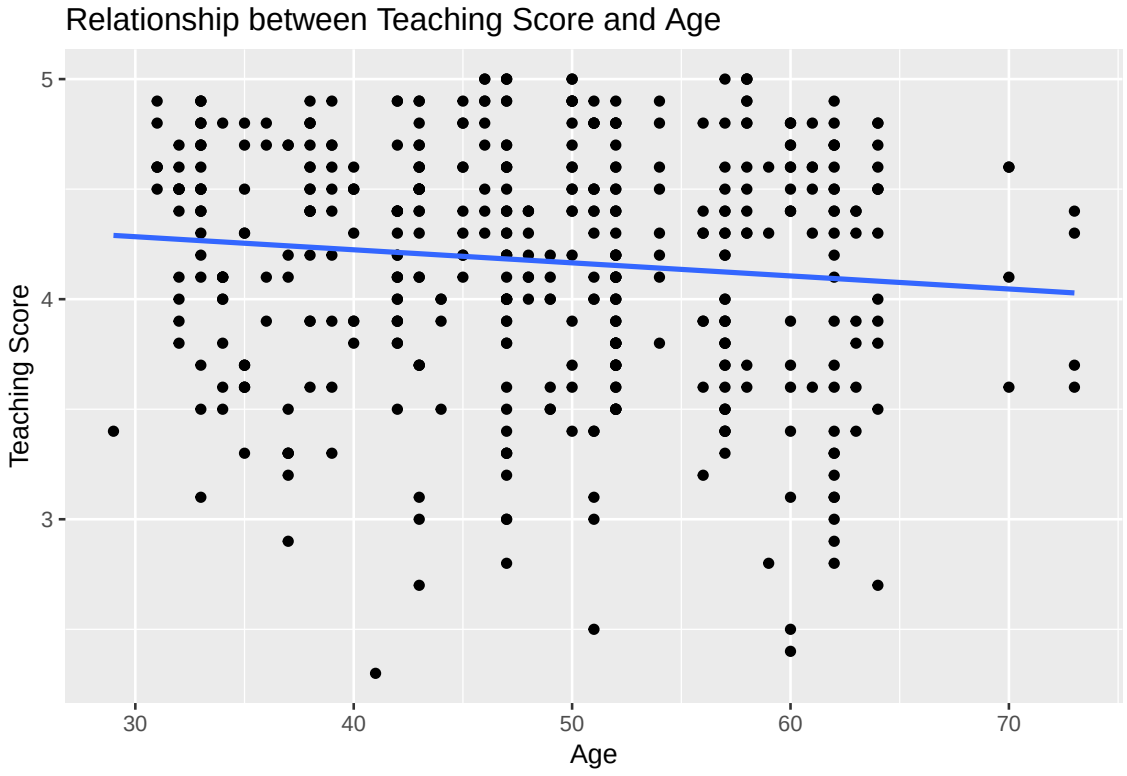


2. Perform a formal analysis of the relationship between teaching score and age by fitting a simple linear regression model. Superimpose your best-fitting line onto your scatterplot from Task 1.

```
evals.age <- evals %>%
  select(score, age)
model <- lm(score ~ age, data = evals.age)
model
```

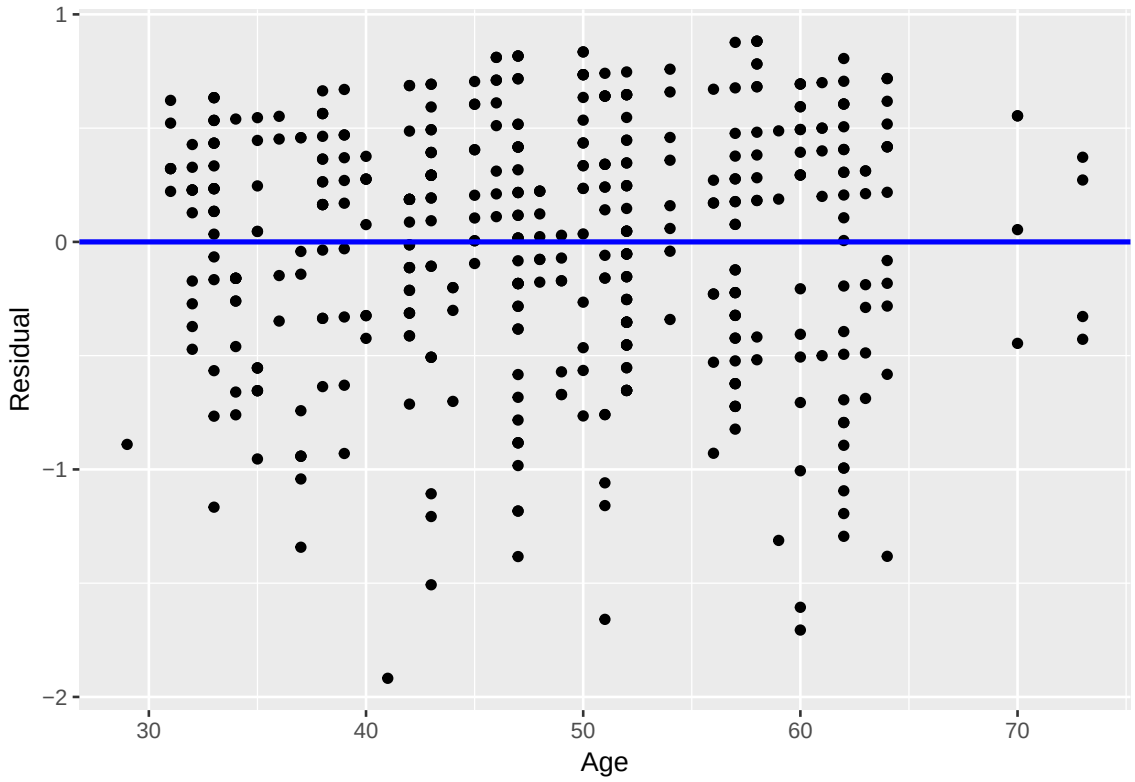
```
Call:
lm(formula = score ~ age, data = evals.age)

Coefficients:
(Intercept)          age
   4.461932    -0.005938
```

```
ggplot(evals.age, aes(x = age, y = score)) +
  geom_point() +
  labs(x = "Age", y = "Teaching Score",
       title = "Relationship between Teaching Score and Age") +
  geom_smooth(method = "lm", se = FALSE)
```

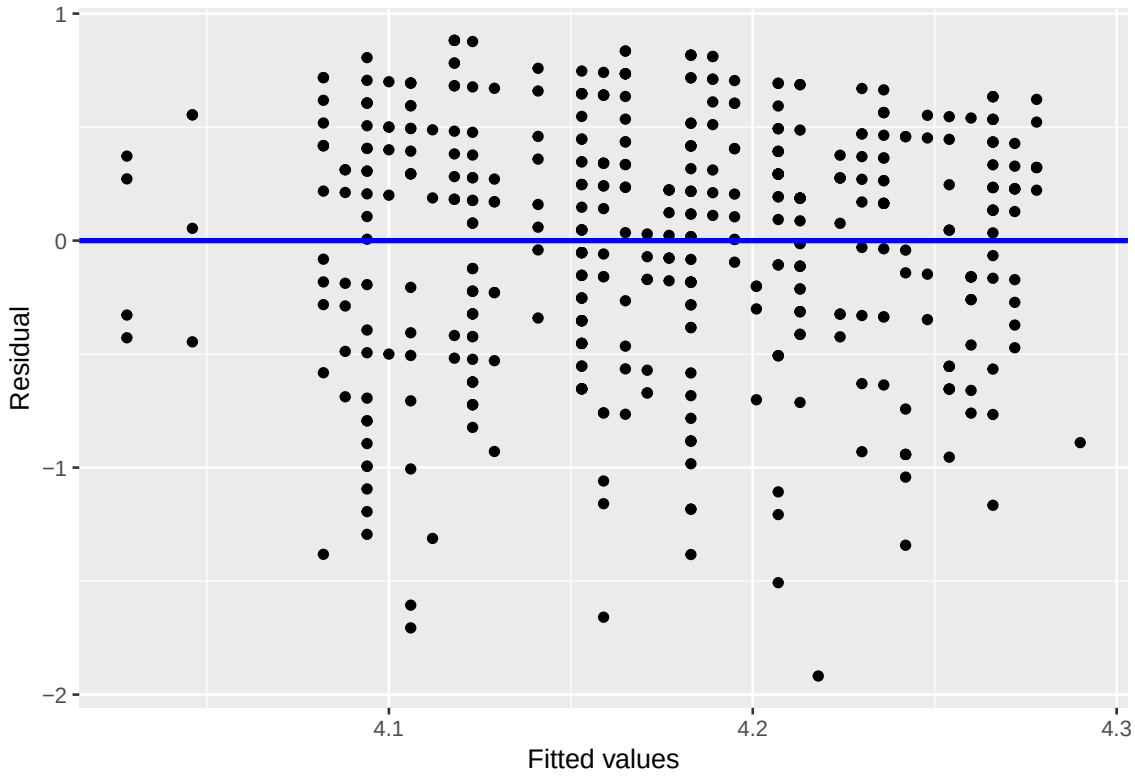


3. Assess the model assumptions from Task 2 by plotting the residuals against the explanatory variable and fitted values, respectively. Also, plot a histogram of the residuals to assess whether they are normally distributed.

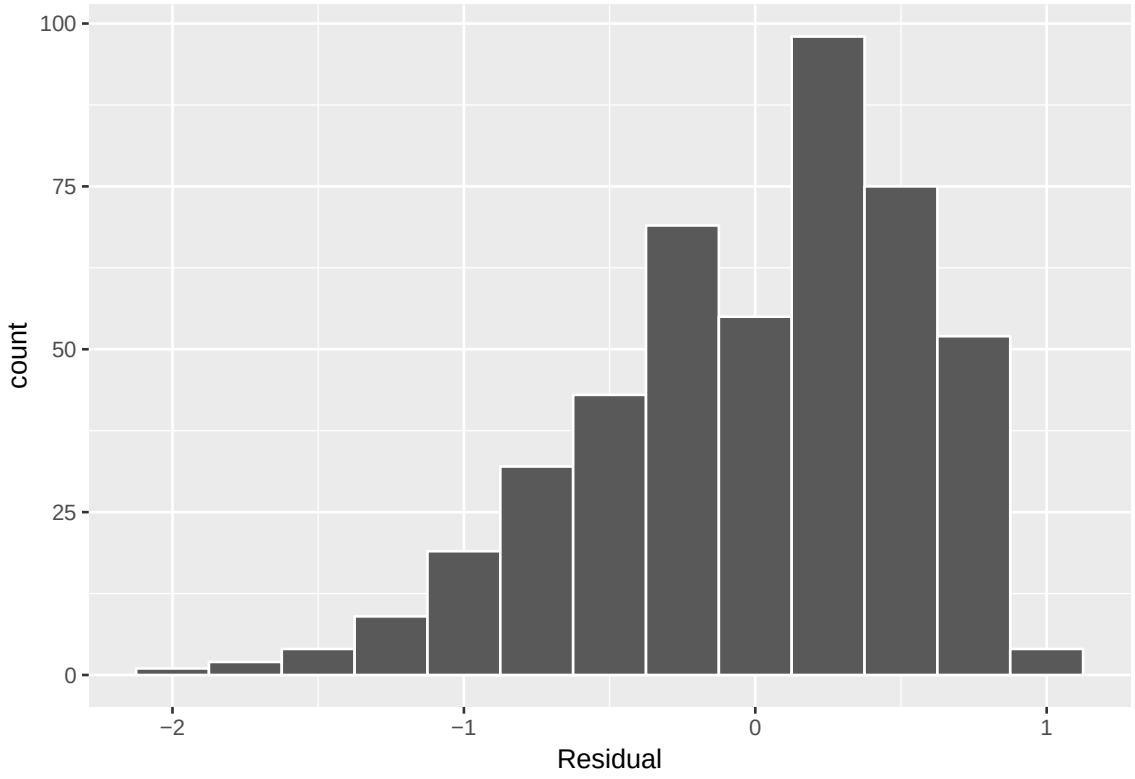
```
evals.age <- evals %>%
  select(score, age)
model <- lm(score ~ age, data = evals.age)
regression.points <- get_regression_points(model)
ggplot(regression.points, aes(x = age, y = residual)) +
  geom_point() +
  labs(x = "Age", y = "Residual") +
  geom_hline(yintercept = 0, col = "blue", size = 1)
```




```
ggplot(regression.points, aes(x = score_hat, y = residual)) +
  geom_point() +
  labs(x = "Fitted values", y = "Residual") +
  geom_hline(yintercept = 0, col = "blue", size = 1)
```




```
ggplot(regression.points, aes(x = residual)) +
  geom_histogram(binwidth = 0.25, color = "white") +
  labs(x = "Residual")
```



4. Perform the same analysis we did on life expectancy from the `gapminder` data set in 2007. However, subset the data for the year 1997. Are there any differences in the results across this 10 year period?

```
gapminder1997 <- gapminder %>%
  filter(year == 1997) %>%
  select(country, continent, lifeExp)

lifeExp.continent <- gapminder1997 %>%
  group_by(continent) %>%
  summarize(median = median(lifeExp), mean = mean(lifeExp))
lifeExp.continent
```

```
# A tibble: 5 x 3
  continent median  mean
  <fct>      <dbl> <dbl>
1 Africa      52.8  53.6
2 Americas    72.1  71.2
3 Asia        70.3  68.0
4 Europe      76.1  75.5
5 Oceania     78.2  78.2
```

```
lifeExp.model <- lm(lifeExp ~ continent, data = gapminder1997)
lifeExp.model
```

```
Call:
lm(formula = lifeExp ~ continent, data = gapminder1997)

Coefficients:
      (Intercept)  continentAmericas      continentAsia
            53.60              17.55              14.42
  continentEurope   continentOceania
            21.91              24.59
```