Week 3 Tasks

Task 1

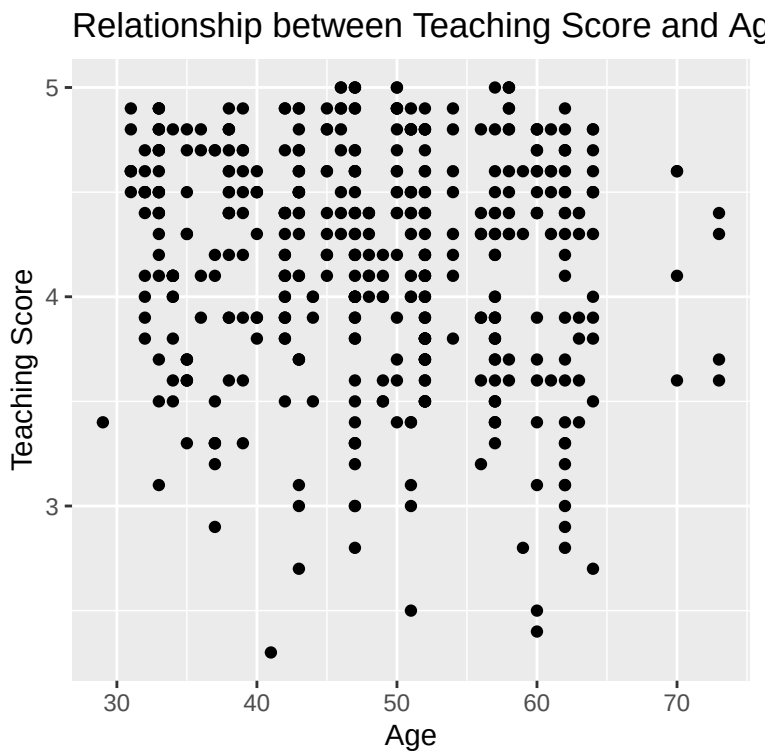
Examine the relationship between teaching score and age in the `evals` data set. What is the value of the correlation coefficient? How would you interpret this verbally? Finally, produce a scatterplot of teaching score and age.
Click here to see the solution

```
evals.age <- evals %>%
  dplyr::select(score, age)

cor(evals.age$score, evals.age$age)
```

```
[1] -0.107032
```

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



Task 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 the previous Task .
Click here to see the solution

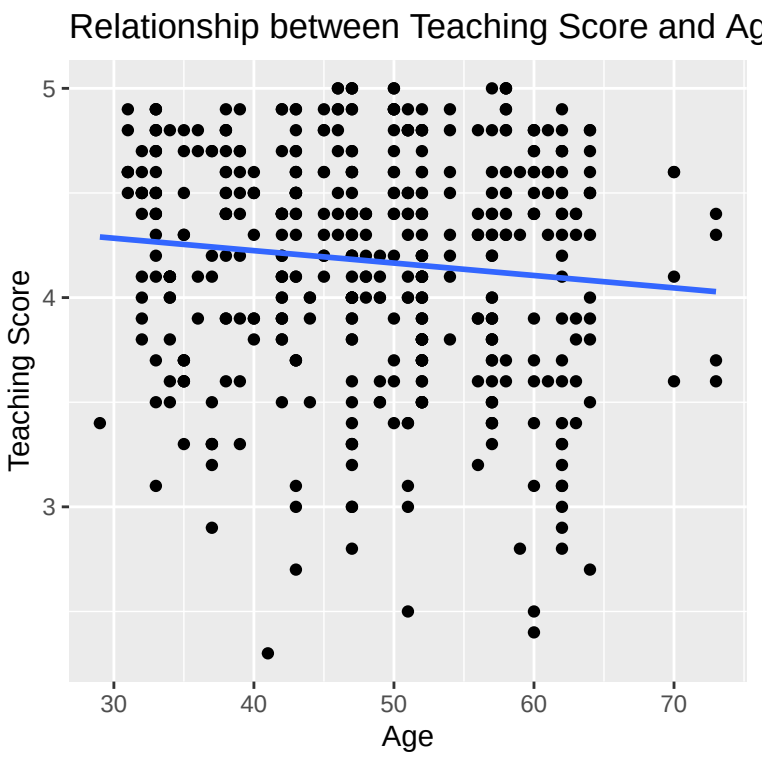
```
model <- lm(score ~ age, data = evals.age)
model
```

```
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)
```

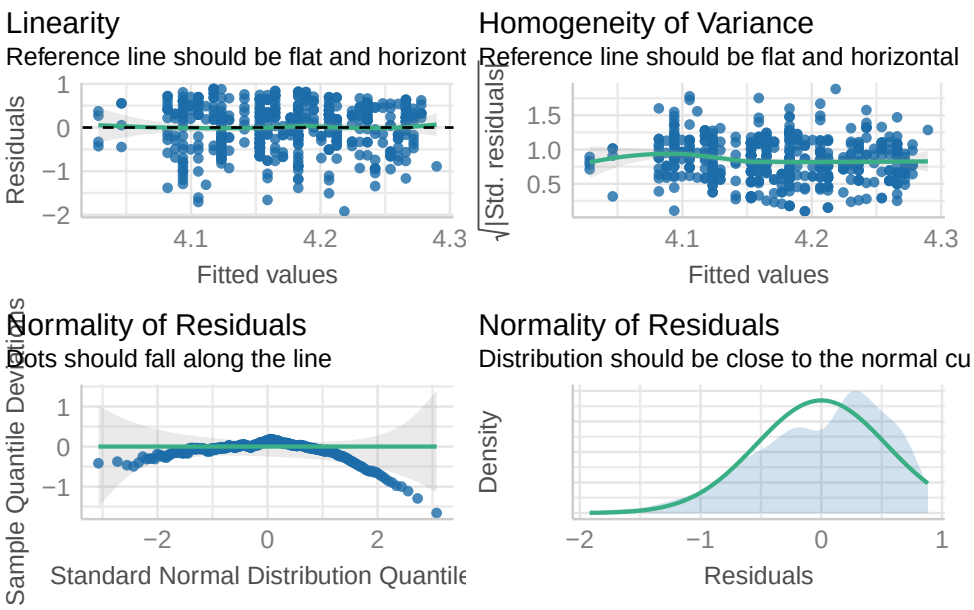
```
`geom_smooth()` using formula = 'y ~ x'
```



Task 3

Assess the model assumptions from Task 2 by plotting the residuals diagnostic plots.
Click here to see the solution

```
check_model(model,check=c("linearity","homogeneity","qq","normality"))
```



Task 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?
Click here to see the solution

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

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

```
# A tibble: 5 x 3
  continent median  mean
  <fct>      <dbl> <dbl>
1 Asia        70.3  68.0
2 Europe      76.1  75.5
3 Africa      52.8  53.6
4 Americas    72.1  71.2
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    continentEurope
            53.60              17.55              14.42              21.91
```

```
continentOceania
           24.59
```

Task 5

Return to the `Credit` data set and fit a multiple regression model with `Balance` as the outcome variable, and `Income` and `Age` as the explanatory variables, respectively. Assess the assumptions of the multiple regression model.
Click here to see the solution

```
# Select variables of interest
Cred <- Credit %>%
  select(Balance, Income, Age)
# Explore the data
Cred %>%
  skim()
```

Table 1: Data summary

Name	Piped data
Number of rows	400
Number of columns	3
Column type frequency:	
numeric	3
Group variables	None

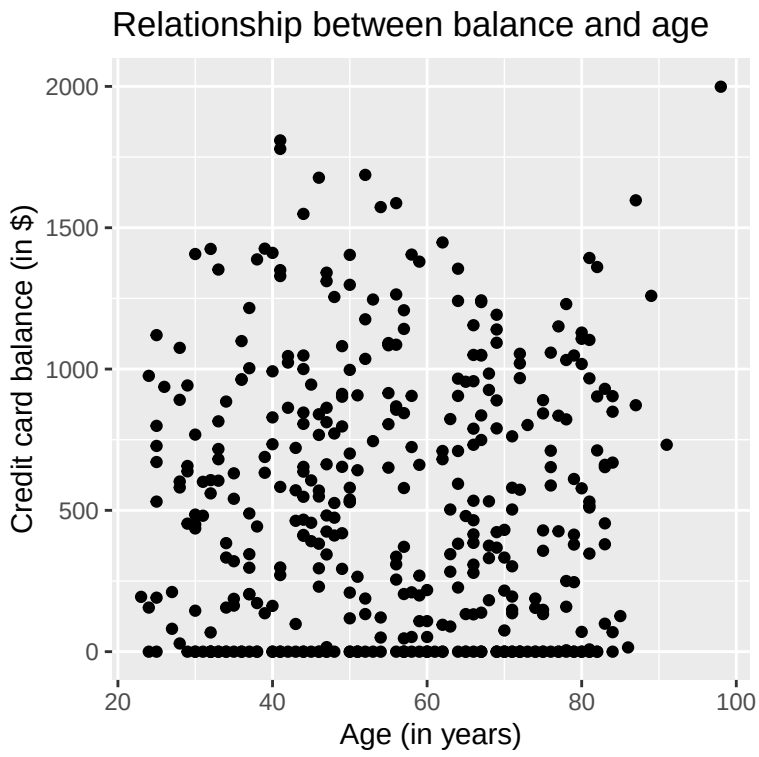
Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Balance	0	1	520.02	459.76	0.00	68.75	459.50	863.00	1999.00	▇▃▂▂▁
Income	0	1	45.22	35.24	10.35	21.01	33.12	57.47	186.63	▇▃▁▁▁
Age	0	1	55.67	17.25	23.00	41.75	56.00	70.00	98.00	▃▆▇▆▂

```r
# Correlation between covariates
Cred %>%
  cor()
```

```
            Balance    Income         Age
Balance 1.000000000 0.4636565 0.001835119
Income  0.463656457 1.0000000 0.175338403
Age     0.001835119 0.1753384 1.000000000
```

```r
# Scatterplot
ggplot(Cred, aes(x = Age, y = Balance)) +
  geom_point() +
  labs(x = "Age (in years)", y = "Credit card balance (in $)",
       title = "Relationship between balance and age")
```




```r
# Fit the model
Balance.model <- lm(Balance ~ Age + Income, data = Cred)
# Model output
tab_model(Balance.model,show.ci = F)
```

	Balance	
Predictors	Estimates	p
(Intercept)	359.67	**<0.001**
Age	-2.19	0.069
Income	6.24	**<0.001**
Observations	400	
R^2 / R^2 adjusted	0.221 / 0.218	

```
# Check assumptions
check_model(Balance.model,check=c("linearity","homogeneity","qq","normality"))
```