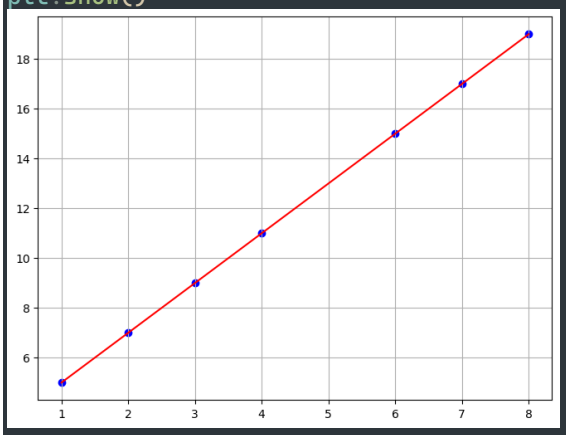
Viikko 39 -tehtävät

Tehtävä 1

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sb
from sklearn.linear_model import LinearRegression

x = [1, 2, 3, 4, 6, 7, 8]
y = [2*x + 3 for x in x]
df = pd.DataFrame({'X': x, 'Y': y})

plt.figure(figsize=(8, 6))
plt.scatter(df['X'], df['Y'], color='blue')
plt.plot(df['X'], df['Y'], color='red')
plt.grid(True)
plt.show()
```

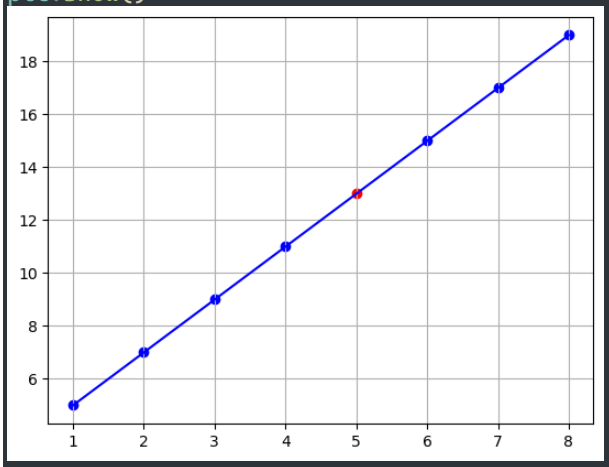


Tehtävä 2

```
model = LinearRegression()
model.fit(df[['X']], df['Y'])

coef = model.coef_
inter = model.intercept_

print(f'Suoran yhtälö: y = {coef[0]}x + {inter}')
```

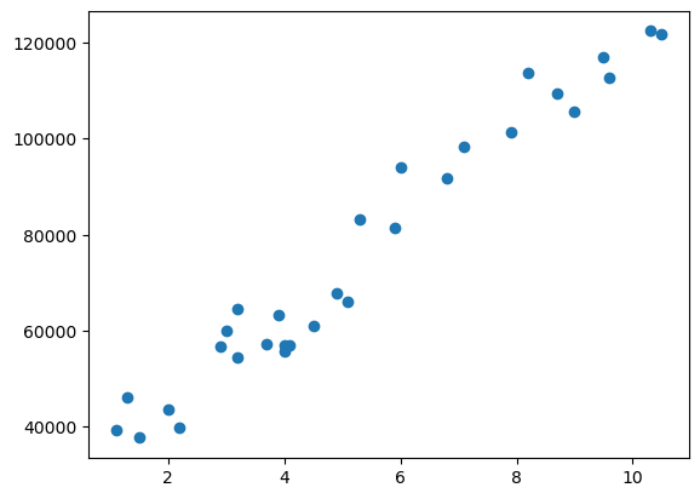
```
Suoran yhtälö: y = 2.0000000000000004x + 2.9999999999999982
```

```
plt.scatter(df['X'], df['Y'], color='blue')
plt.plot(df['X'], df['Y'], color='blue')
plt.scatter(x=5, y=model.predict([[5]]), color='red')
plt.grid(True)
plt.show()
```

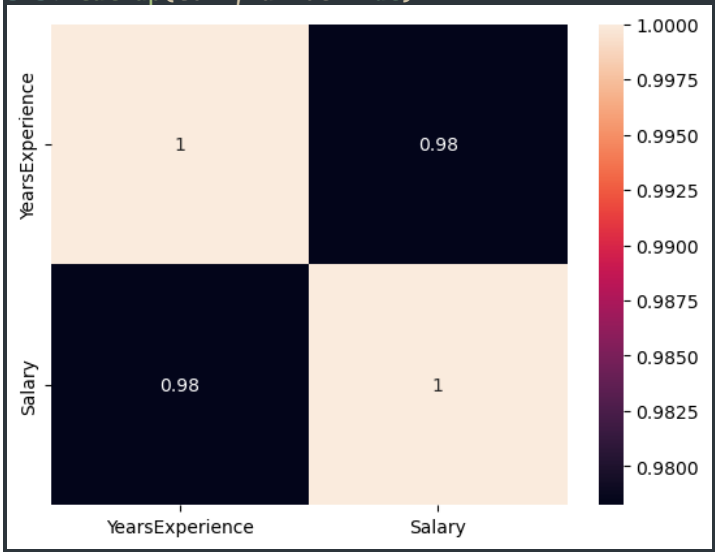


Tehtävä 3

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.stats import pearsonr
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import r2_score, mean_absolute_error, mean_squared_error

df = pd.read_csv('./work/viikko5/datasets/salary.csv')

plt.scatter(df['YearsExperience'], df['Salary'])
plt.show()
```




```
corr = df.corr()
sns.heatmap(corr, annot=True)
```




```
pearsonr(df['YearsExperience'], df['Salary'])
```

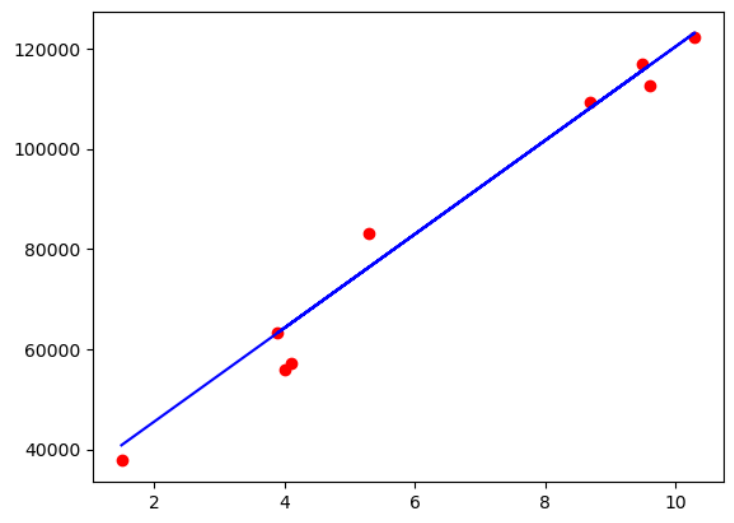
```
PearsonRResult(statistic=0.9782416184887599, pvalue=1.1430681092271567e-20)
```

```
X = df.iloc[:, [0]]
y = df.iloc[:, [1]]

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3,
random_state=0)

model = LinearRegression()
model.fit(X_train, y_train)
```

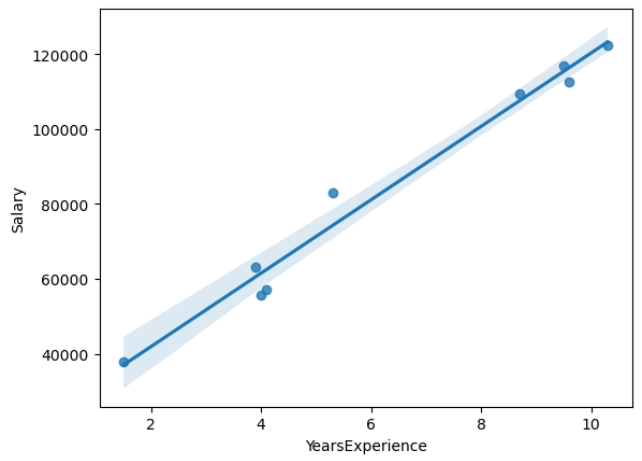
```
coef = model.coef_
inter = model.intercept_

print(f'Suoran yhtälö: Salary = {coef[0]} * YearsOfExperience + {inter}')
```

```
Suoran yhtälö: Salary = [9360.26128619] * YearsOfExperience + [26777.3913412]
```

```
y_pred = model.predict(X_test)

plt.scatter(X_test, y_test, color='red')
plt.plot(X_test, y_pred, color='blue')
plt.show()
```




```
sns.regplot(x=X_test, y=y_test)
plt.show()
```




```
r2 = r2_score(y_test, y_pred)
mae = mean_absolute_error(y_test, y_pred)
```

```
rmse = mean_squared_error(y_test, y_pred, squared=False)


print(f'r2: {r2}')
print(f'mae: {mae}')
print(f'rmse: {rmse}')
```

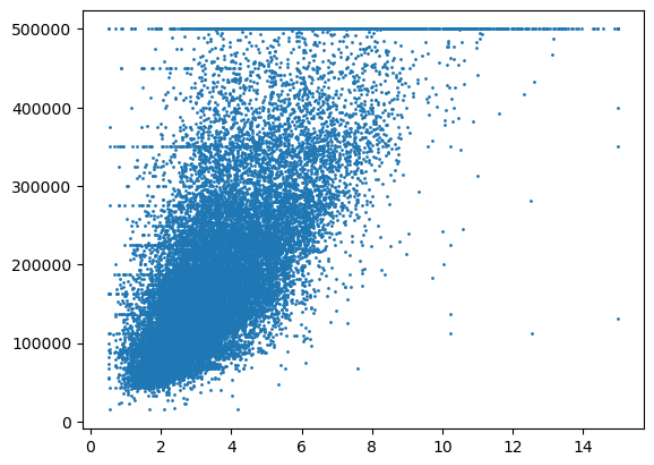
```
r2: 0.9740993407213511
mae: 3737.417861878896
rmse: 4834.260936361728
```

```
print(f'Palkka 7 vuoden kokemuksella: {model.predict([[7]])}')
```

```
Palkka 7 vuoden kokemuksella: [[92299.22034455]]
```

Tehtävä 4

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.stats import pearsonr
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import r2_score, mean_absolute_error, mean_squared_error

df = pd.read_csv('./work/viikko5/datasets/housing.csv')

plt.scatter(x=df['median_income'], y=df['median_house_value'], s=1)
plt.show()
```

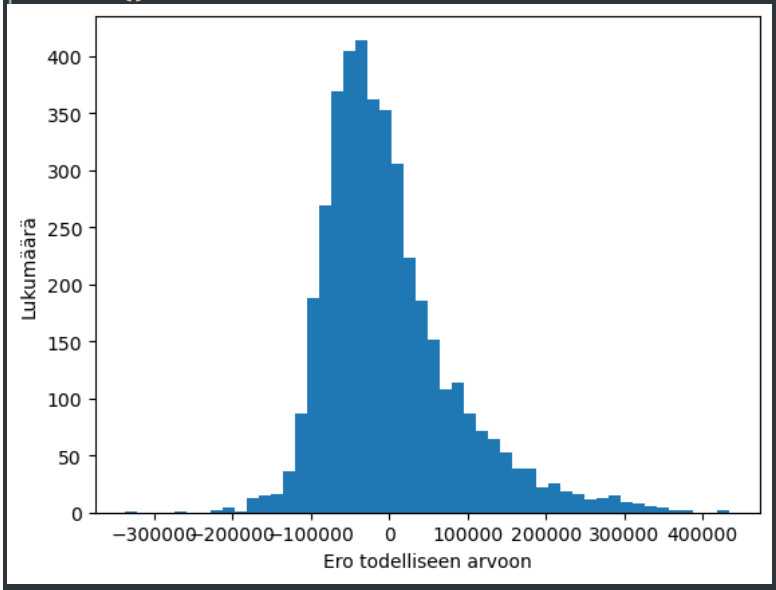



```
X = df[['median_income']]
```

```
y = df[['median_house_value']]

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=0)



model = LinearRegression()
model.fit(X_train, y_train)

coef = model.coef_
inter = model.intercept_

print(f'Suoran yhtälö: MedianHouseValue = {coef[0]} * MedianIncome + {inter}')
```

```
Suoran yhtälö: MedianHouseValue = [42032.17769894] * MeidanIncome + [44320.63522766]
```

```
y_pred = model.predict(X_test)

plt.hist(y_test - y_pred, bins=50)
plt.xlabel('Ero todelliseen arvoon')
plt.ylabel('Lukumäärä')
plt.show()
```




```
r2 = r2_score(y_test, y_pred)
mae = mean_absolute_error(y_test, y_pred)
rmse = mean_squared_error(y_test, y_pred, squared=False)
```

```
print(f'r2: {r2}')
print(f'mae: {mae}')
print(f'rmse: {rmse}')
```

```
r2: 0.44668468048959444
mae: 63521.30348040669
rmse: 84941.05152406936
```

```
print(f'Ennustettu talon arvo 30 000$ vuosituloilla: {model.predict([[30]])}')
```

```
Ennustettu talon arvo 30 000$ vuosituloilla: [[1305285.96619586]]
```