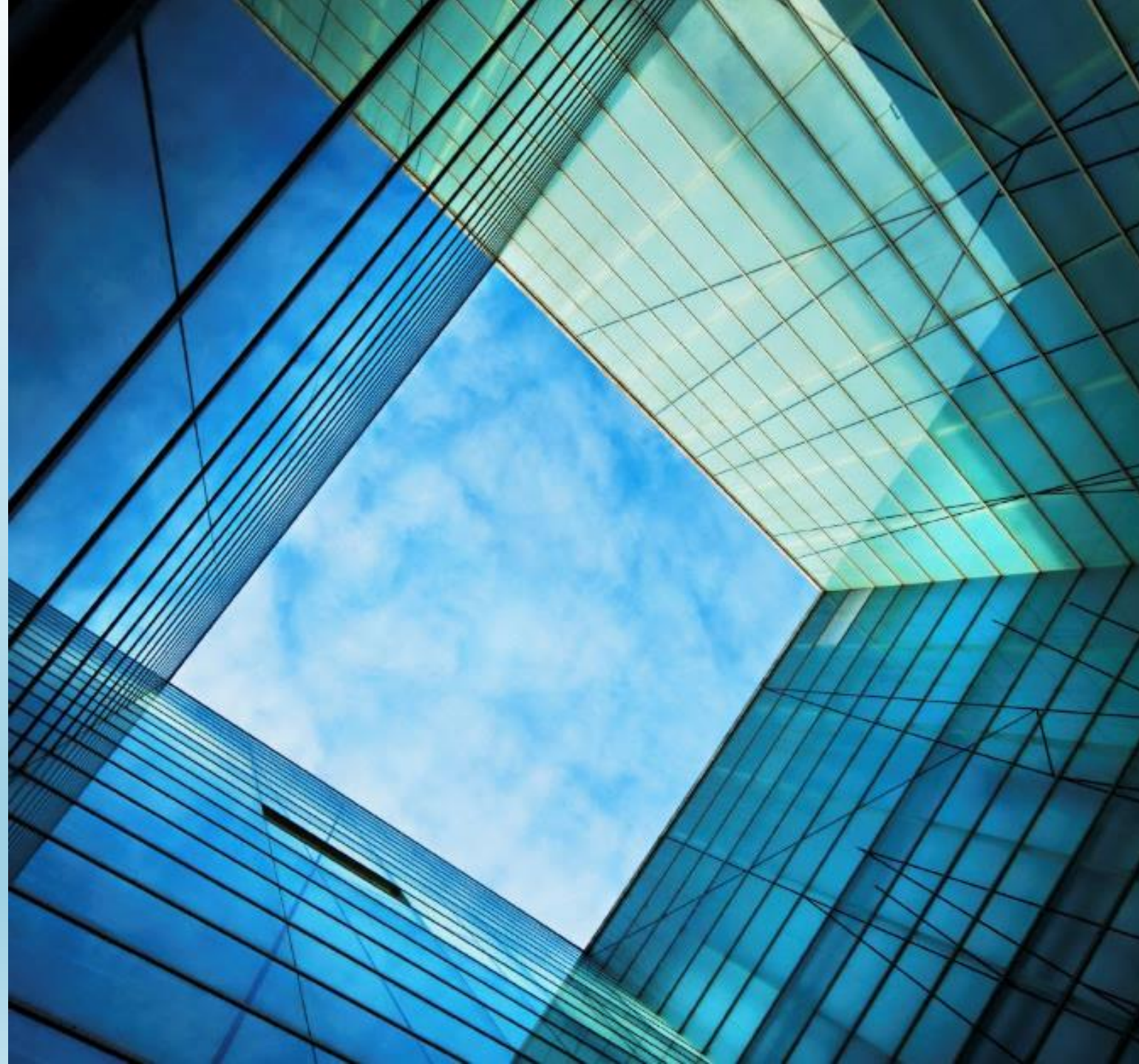
Activity #4

Linear Regression



TOPICS

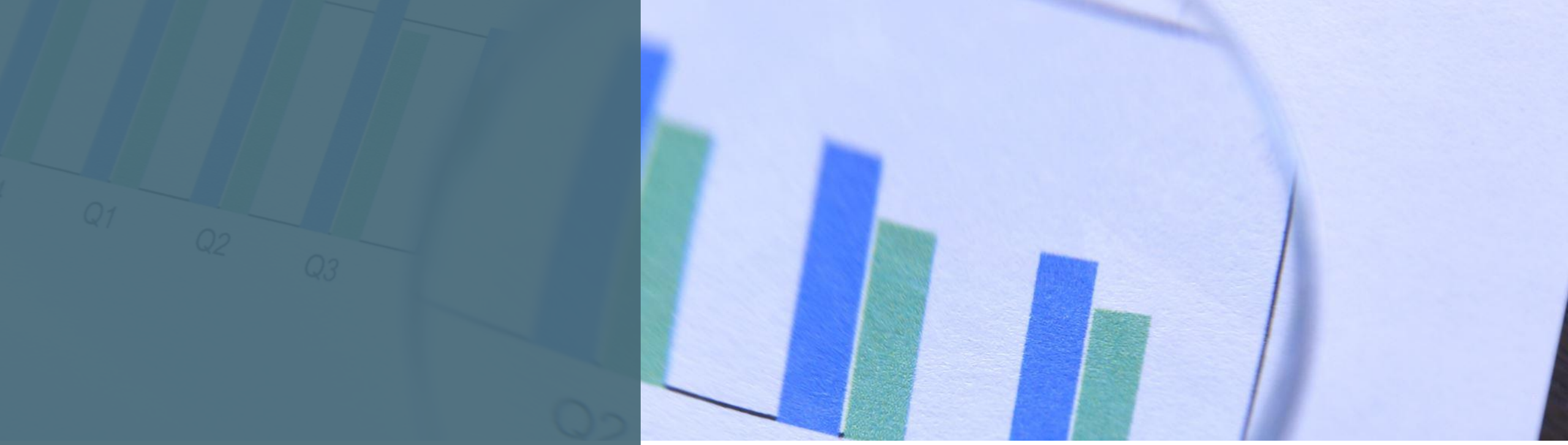
4.1 Data Exploration / Transform / Feature Selection

4.2 PCA: Feature Dimensional Reduction

4.3 Linear Regression

LIBRARIES

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn import preprocessing
from sklearn import preprocessing
from sklearn.decomposition import PCA
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score
```



4.1

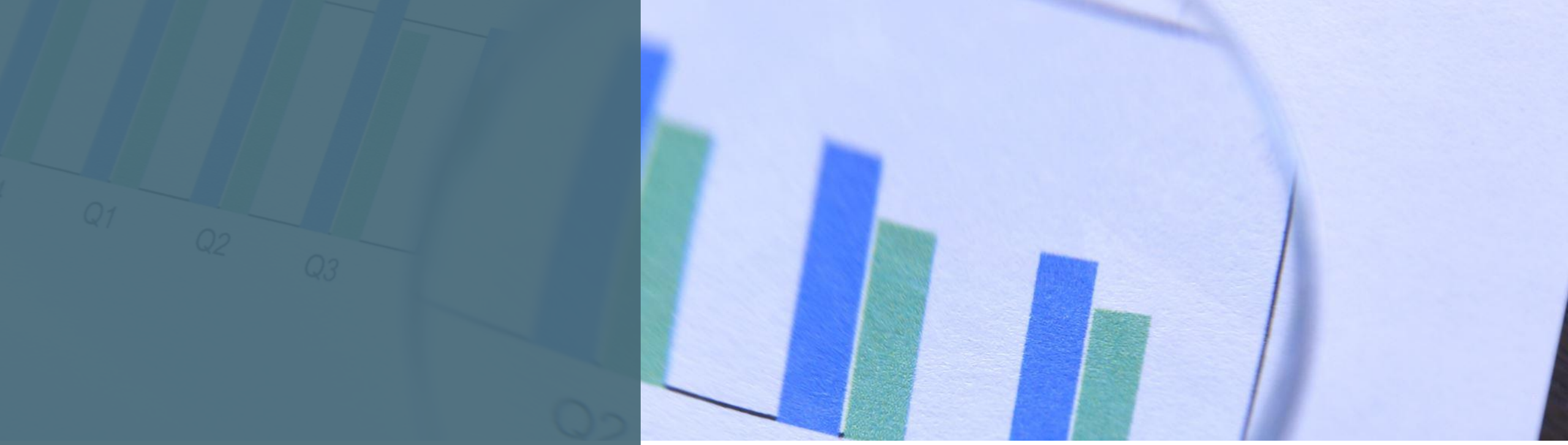
Data Exploration / Transform / Feature Selection

4.1 (a) Data exploration

1. Read .csv file
   - read_csv("CarPrice.csv")
2. View Data Array Shape
   - # Variables
   - # Samples
   - # Statistics -> describe()
3. Remove
   - 'car_ID','CarName'
4. View Variable info
   - Info()
     - Data Type / # non null
5. Fill NA
   - fillna()

4.1 (b) Data Transform and Feature Selection

1. • Standardized Data for continuous data columns for only continuous data columns
2. • Calculate correlation between variables for only continuous data columns
   • corr()
3. • Reduce Corr() to Lower Matrix
4. • Drop columns if correlation value > 0.86
5. • OneHotEncode for categorical columns (try from Pandas)
   • pd.get_dummies(data, columns = categorical_features, drop_first=True)



4.2

PCA: Feature Dimensional Reduction

4.2 PCA Dimensional Reduction

1

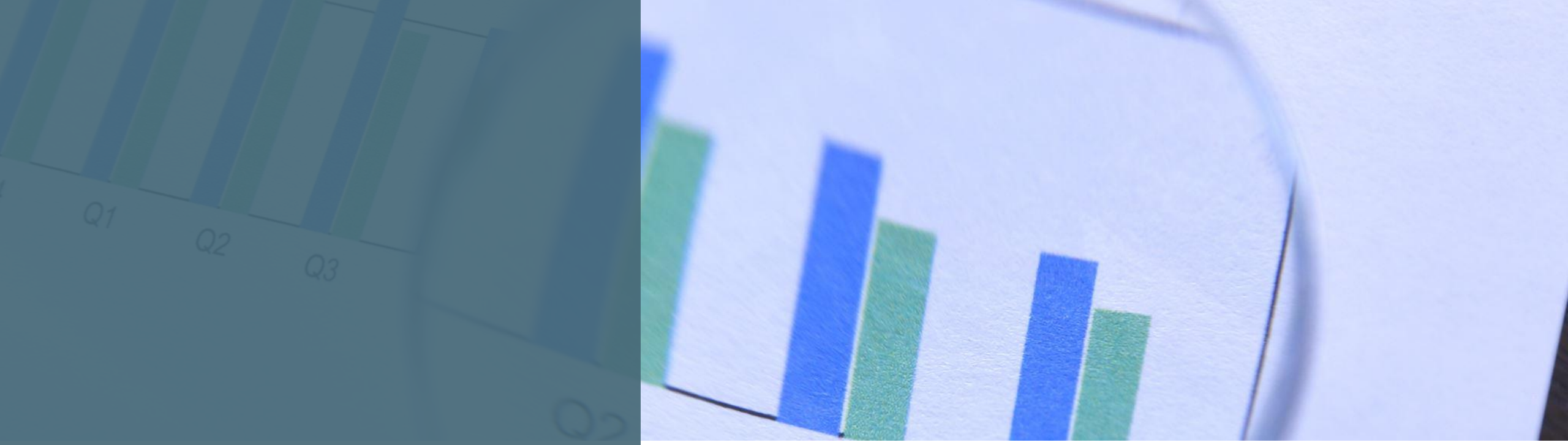
- # PCA all variables
  - pca = PCA()
  - X_pca = pca.fit_transform(X_standard)

2

- Visualize Explained Variance Ratio (% eigenvalues)
  - plt.bar() ค่าของ pca.explained_variance_ratio_

3

- # PCA n_components (ทดลองเปลี่ยนค่า n_components อย่างน้อย 3 ค่า เพื่อเลือกค่าดีที่สุด)
  - pca2 = PCA(n_components)
  - X_pca_2 = pca2.fit_transform(X_standard)



4.3

Linear Regression

4.3 Linear Regression

1

- # Shuffle Split (Train / Test Split)
  - Rseed กำหนด ค่าใดก็ได้
  - x_train_set, x_test, y_train_set, y_test = train_test_split(X, Y, test_size = 0.3, random_state = Rseed)

2

- # Shuffle Split (Train / Validation Split)
  - x_train, x_validate, y_train, y_validate = train_test_split(x_train_set, y_train_set, test_size = 0.3, random_state = Rseed)

3

- # Perform Linear Regression -> All variables
  - lr = LinearRegression()
  - # Train
  - lr.fit(x_train, y_train)
  - # Validate
  - y_pred_lr = lr.predict(x_validate)
  - # Test
  - y_test_pred_lr = lr.predict(x_test)

4

- # Measure Accuracy Validation and Test
  - r2_score(y_pred_lr, y_validate)
  - r2_score(y_test_pred_lr, y_test)
  - lr.score(x_validate, y_validate)
  - lr.score(x_test, y_test)

เปรียบเทียบประสิทธิภาพ Linear Regression