여러 변수의 상관분석
#01. 작업준비
패키지 참조
데이터 가져오기
#02. 데이터 전처리
분석 대상 컬럼만 추출
#03. 산점도 행렬
1. 기본 사용
2. 파라미터 설정
대각선 기준 다르게 표현하기
#04. 인터렉티브 그래프
상관계수 행렬
산점도 행렬 시각화
분석결과 검정

여러 변수의 상관분석

#01. 작업준비

패키지 참조

```
import numpy as np
from pandas import read_excel
from scipy import stats

import sys
import seaborn as sb
from matplotlib import pyplot as plt
```

데이터 가져오기

```
df = read_excel("https://data.hossam.kr/E03/mtcars.xlsx", index_col='nam
df
```

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	ca
name											
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4

여러 변수의 상관분석

#01. 작업준비

패키지 참조

데이터 가져오기

#02. 데이터 전처리

분석 대상 컬럼만 추출

#03. 산점도 행렬

1. 기본 사용

2. 파라미터 설정

대각선 기준 다르게 표현하기

#04. 인터렉티브 그래프

상관계수 행렬

산점도 행렬 시각화

분석결과 검정

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	ca
name											
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.440	18.90	1	0	4	4
Merc 450SE	16.4	8	275.8	180	3.07	4.070	17.40	0	0	3	3
Merc 450SL	17.3	8	275.8	180	3.07	3.730	17.60	0	0	3	3
Merc 450SLC	15.2	8	275.8	180	3.07	3.780	18.00	0	0	3	3

여러 변수의 상관분석

#01. 작업준비

패키지 참조

데이터 가져오기

#02. 데이터 전처리

분석 대상 컬럼만 추출

#03. 산점도 행렬

1. 기본 사용

2. 파라미터 설정

대각선 기준 다르게 표현하기

#04. 인터렉티브 그래프

상관계수 행렬

산점도 행렬 시각화

분석결과 검정

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	ca
name											
Cadillac Fleetwood	10.4	8	472.0	205	2.93	5.250	17.98	0	0	3	4
Lincoln Continental	10.4	8	460.0	215	3.00	5.424	17.82	0	0	3	4
Chrysler Imperial	14.7	8	440.0	230	3.23	5.345	17.42	0	0	3	4
Fiat 128	32.4	4	78.7	66	4.08	2.200	19.47	1	1	4	1
Honda Civic	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2
Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.90	1	1	4	1
Toyota Corona	21.5	4	120.1	97	3.70	2.465	20.01	1	0	3	1
Dodge Challenger	15.5	8	318.0	150	2.76	3.520	16.87	0	0	3	2
AMC Javelin	15.2	8	304.0	150	3.15	3.435	17.30	0	0	3	2
Camaro Z28	13.3	8	350.0	245	3.73	3.840	15.41	0	0	3	4
Pontiac	19.2	8	400.0	175	3.08	3.845	17.05	0	0	3	2

여러 변수의 상관분석

#01. 작업준비

패키지 참조

데이터 가져오기

#02. 데이터 전처리

분석 대상 컬럼만 추출

#03. 산점도 행렬

1. 기본 사용

2. 파라미터 설정

대각선 기준 다르게 표현하기

#04. 인터렉티브 그래프

상관계수 행렬

산점도 행렬 시각화

분석결과 검정

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	ca
name											
Firebird											
Fiat X1-9	27.3	4	79.0	66	4.08	1.935	18.90	1	1	4	1
Porsche 914-2	26.0	4	120.3	91	4.43	2.140	16.70	0	1	5	2
Lotus Europa	30.4	4	95.1	113	3.77	1.513	16.90	1	1	5	2
Ford Pantera L	15.8	8	351.0	264	4.22	3.170	14.50	0	1	5	4
Ferrari Dino	19.7	6	145.0	175	3.62	2.770	15.50	0	1	5	6
Maserati Bora	15.0	8	301.0	335	3.54	3.570	14.60	0	1	5	8
Volvo 142E	21.4	4	121.0	109	4.11	2.780	18.60	1	1	4	2

#02. 데이터 전처리

분석 대상 컬럼만 추출

```
df2 = df.filter(['mpg', 'disp', 'hp', 'drat', 'wt', 'qsec', 'vs'])
df2['vs'] = df2['vs'].astype('category')
```

여러 변수의 상관분석

#01. 작업준비

패키지 참조

데이터 가져오기

#02. 데이터 전처리

분석 대상 컬럼만 추출

#03. 산점도 행렬

1. 기본 사용

2. 파라미터 설정

대각선 기준 다르게 표현하기

#04. 인터렉티브 그래프

상관계수 행렬

산점도 행렬 시각화

분석결과 검정

```
df2
```

	mpg	disp	hp	drat	wt	qsec	vs
name							
Mazda RX4	21.0	160.0	110	3.90	2.620	16.46	0
Mazda RX4 Wag	21.0	160.0	110	3.90	2.875	17.02	0
Datsun 710	22.8	108.0	93	3.85	2.320	18.61	1
Hornet 4 Drive	21.4	258.0	110	3.08	3.215	19.44	1
Hornet Sportabout	18.7	360.0	175	3.15	3.440	17.02	0
Valiant	18.1	225.0	105	2.76	3.460	20.22	1
Duster 360	14.3	360.0	245	3.21	3.570	15.84	0
Merc 240D	24.4	146.7	62	3.69	3.190	20.00	1
Merc 230	22.8	140.8	95	3.92	3.150	22.90	1
Merc 280	19.2	167.6	123	3.92	3.440	18.30	1
Merc 280C	17.8	167.6	123	3.92	3.440	18.90	1
Merc 450SE	16.4	275.8	180	3.07	4.070	17.40	0
Merc 450SL	17.3	275.8	180	3.07	3.730	17.60	0
Merc 450SLC	15.2	275.8	180	3.07	3.780	18.00	0
Cadillac Fleetwood	10.4	472.0	205	2.93	5.250	17.98	0
Lincoln Continental	10.4	460.0	215	3.00	5.424	17.82	0

여러 변수의 상관분석

#01. 작업준비

패키지 참조

데이터 가져오기

#02. 데이터 전처리

분석 대상 컬럼만 추출

#03. 산점도 행렬

1. 기본 사용

2. 파라미터 설정

대각선 기준 다르게 표현하기

#04. 인터렉티브 그래프

상관계수 행렬

산점도 행렬 시각화

분석결과 검정

	mpg	disp	hp	drat	wt	qsec	vs
name							
Chrysler Imperial	14.7	440.0	230	3.23	5.345	17.42	0
Fiat 128	32.4	78.7	66	4.08	2.200	19.47	1
Honda Civic	30.4	75.7	52	4.93	1.615	18.52	1
Toyota Corolla	33.9	71.1	65	4.22	1.835	19.90	1
Toyota Corona	21.5	120.1	97	3.70	2.465	20.01	1
Dodge Challenger	15.5	318.0	150	2.76	3.520	16.87	0
AMC Javelin	15.2	304.0	150	3.15	3.435	17.30	0
Camaro Z28	13.3	350.0	245	3.73	3.840	15.41	0
Pontiac Firebird	19.2	400.0	175	3.08	3.845	17.05	0
Fiat X1-9	27.3	79.0	66	4.08	1.935	18.90	1
Porsche 914-2	26.0	120.3	91	4.43	2.140	16.70	0
Lotus Europa	30.4	95.1	113	3.77	1.513	16.90	1
Ford Pantera L	15.8	351.0	264	4.22	3.170	14.50	0
Ferrari Dino	19.7	145.0	175	3.62	2.770	15.50	0
Maserati Bora	15.0	301.0	335	3.54	3.570	14.60	0
Volvo 142E	21.4	121.0	109	4.11	2.780	18.60	1

여러 변수의 상관분석
#01. 작업준비
패키지 참조
데이터 가져오기
#02. 데이터 전처리
분석 대상 컬럼만 추출
#03. 산점도 행렬
1. 기본 사용
2. 파라미터 설정
대각선 기준 다르게 표현하기
#04. 인터렉티브 그래프
상관계수 행렬
산점도 행렬 시각화
분석결과 검정

#03. 산점도 행렬

각 변수들을 교차로 표현한 산점도 그래프의 묶음

```
plt.rcParams["font.family"] = 'AppleGothic' if sys.platform == 'darwin'
plt.rcParams["font.size"] = 12
plt.rcParams["figure.figsize"] = (20, 20)
plt.rcParams["axes.unicode_minus"] = False
```

1. 기본 사용

대각 원소자리에 히스토그램이 표현된다.

`pairplot()` 메서드에 `diag_kind='hist'`를 적용 (기본값이므로 생략시 자동 적용)

카테고리 타입은 자동으로 제외됨

```
sb.pairplot(df2)
plt.show()
plt.close()
```

여러 변수의 상관분석

#01. 작업준비

패키지 참조

데이터 가져오기

#02. 데이터 전처리

분석 대상 컬럼만 추출

#03. 산점도 행렬

1. 기본 사용

2. 파라미터 설정

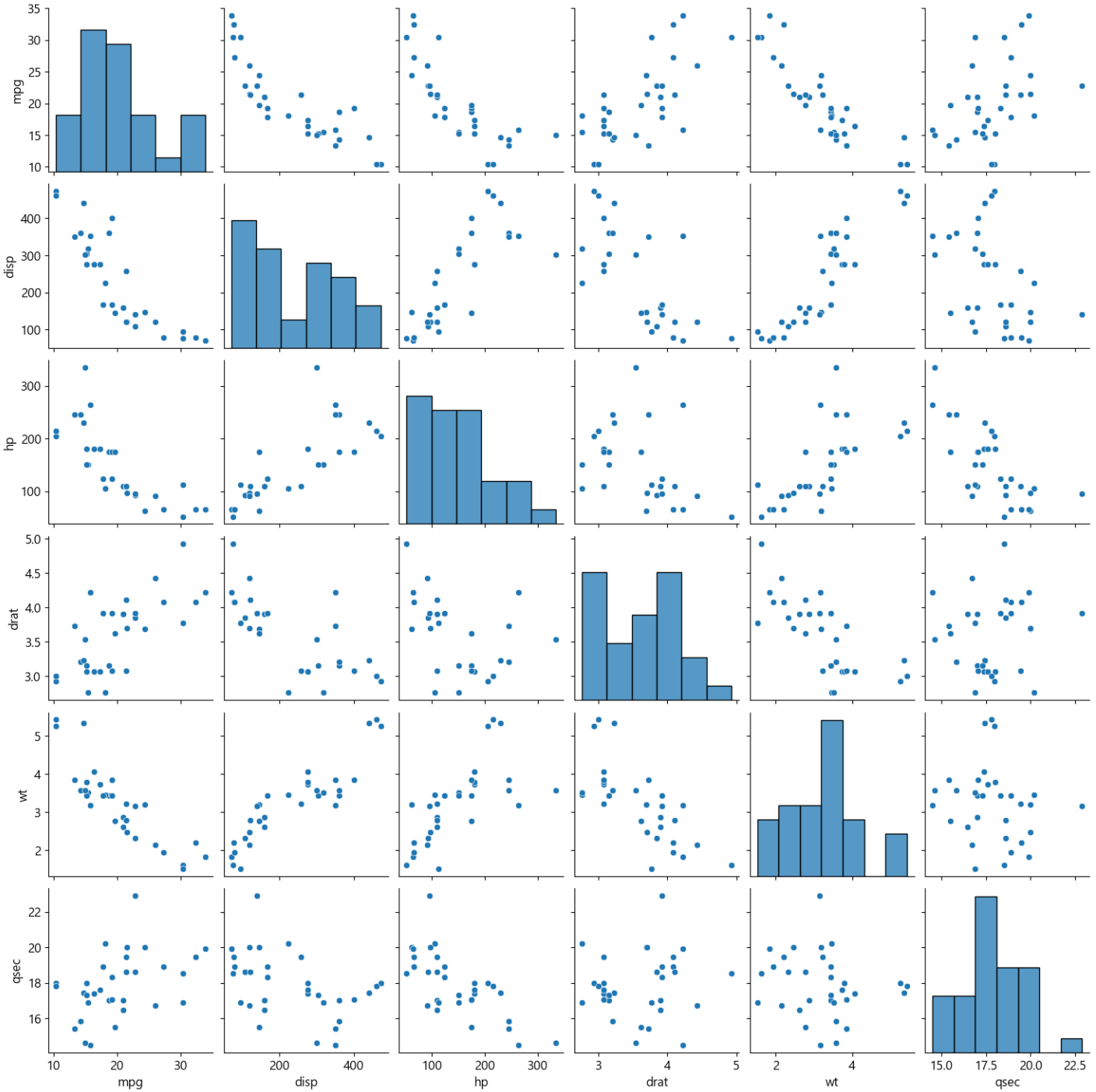
대각선 기준 다르게 표현하기

#04. 인터렉티브 그래프

상관계수 행렬

산점도 행렬 시각화

분석결과 검정



여러 변수의 상관분석

#01. 작업준비

패키지 참조

데이터 가져오기

#02. 데이터 전처리

분석 대상 컬럼만 추출

#03. 산점도 행렬

1. 기본 사용

2. 파라미터 설정

대각선 기준 다르게 표현하기

#04. 인터렉티브 그래프

상관계수 행렬

산점도 행렬 시각화

분석결과 검정

2. 파라미터 설정

```
sb.pairplot(df2,
            diag_kind='kde',  # 대각선에 커널밀도분포 표시
            hue='vs',  # 범주별 색상 구분
            palette='pastel'  # pastel, bright, deep, muted, colorblir
)
plt.show()
plt.close()
```

여러 변수의 상관분석

#01. 작업준비

패키지 참조

데이터 가져오기

#02. 데이터 전처리

분석 대상 컬럼만 추출

#03. 산점도 행렬

1. 기본 사용

2. 파라미터 설정

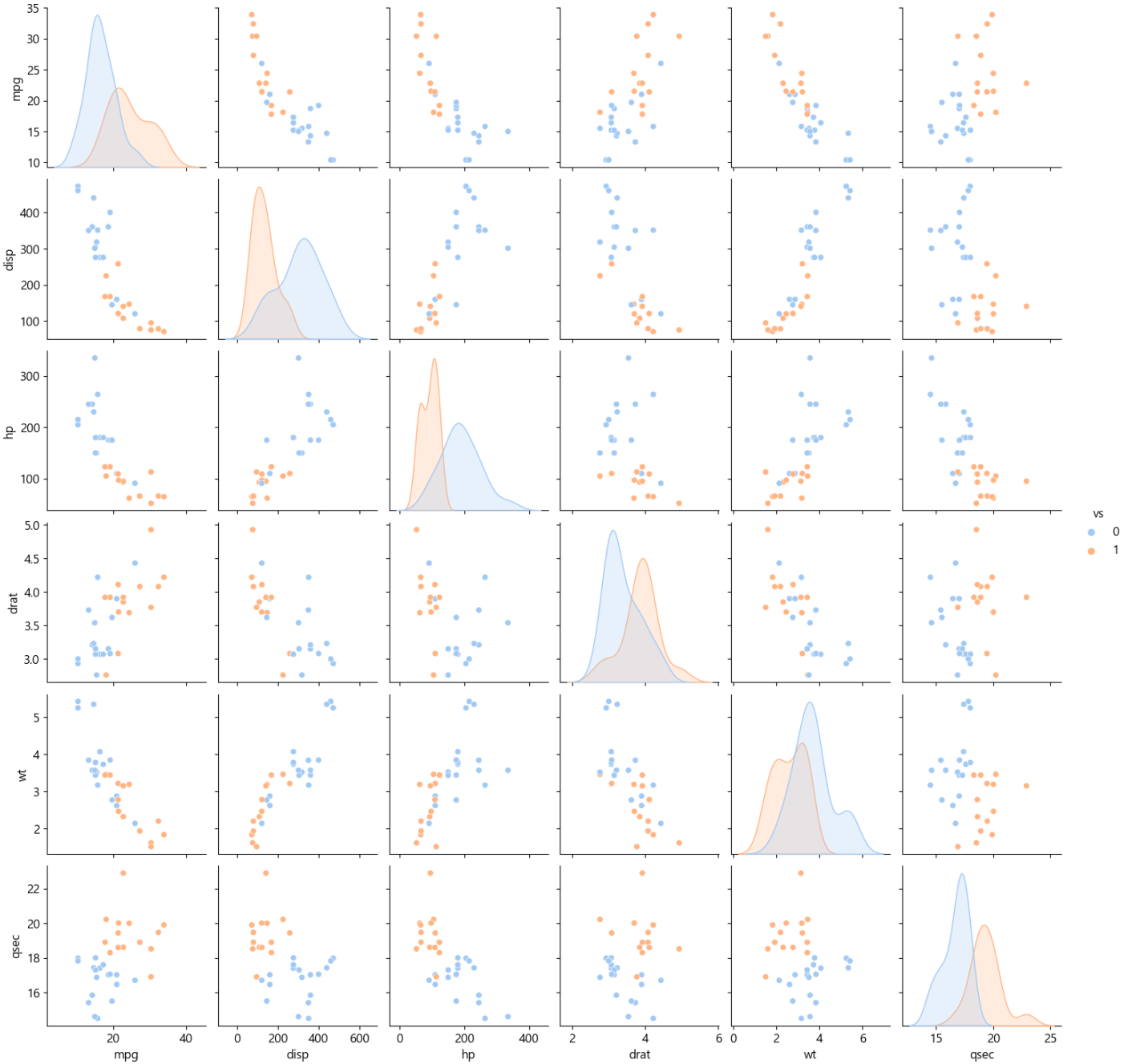
대각선 기준 다르게 표현하기

#04. 인터렉티브 그래프

상관계수 행렬

산점도 행렬 시각화

분석결과 검정



여러 변수의 상관분석

#01. 작업준비

패키지 참조

데이터 가져오기

#02. 데이터 전처리

분석 대상 컬럼만 추출

#03. 산점도 행렬

1. 기본 사용

2. 파라미터 설정

대각선 기준 다르게 표현하기

#04. 인터렉티브 그래프

상관계수 행렬

산점도 행렬 시각화

분석결과 검정

대각선 기준 다르게 표현하기

- 대각선 위 : 2차원 밀도함수 + 추세선
- 대각선 아래 : 2차원 밀도함수 + 산포도

```
g = sb.pairplot(df2,
                diag_kind='kde',  # 대각선에 커널밀도분포 표시
                hue='vs',  # 범주별 색상 구분
                palette='pastel'  # pastel, bright, deep, muted, colorblir
)

g.map_upper(sb.kdeplot, alpha=0.3)
g.map_lower(sb.regplot, scatter=False, truncate=False, ci=False)

plt.show()
plt.close()
```

여러 변수의 상관분석

#01. 작업준비

패키지 참조

데이터 가져오기

#02. 데이터 전처리

분석 대상 컬럼만 추출

#03. 산점도 행렬

1. 기본 사용

2. 파라미터 설정

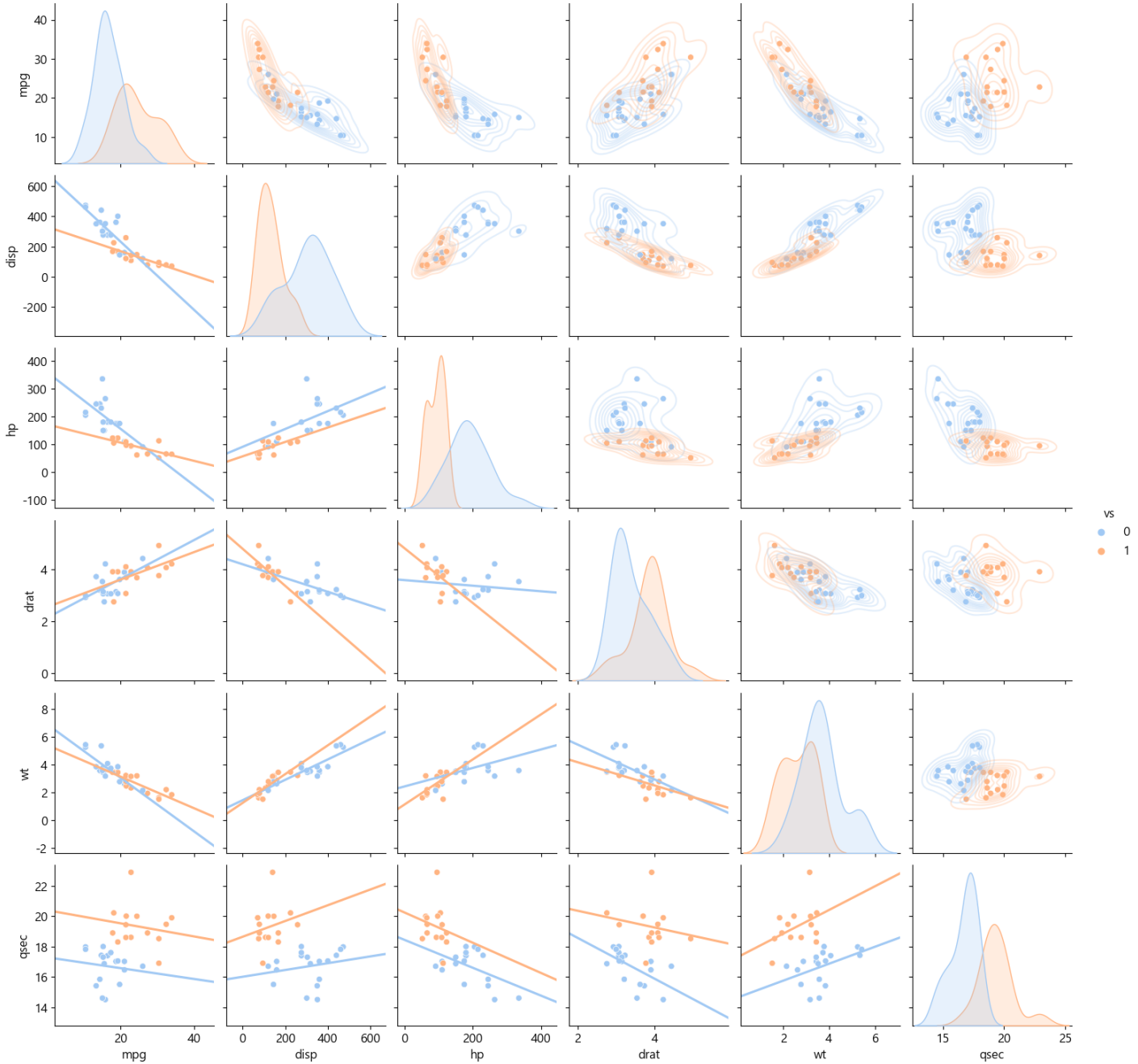
대각선 기준 다르게 표현하기

#04. 인터렉티브 그래프

상관계수 행렬

산점도 행렬 시각화

분석결과 검정



여러 변수의 상관분석

#01. 작업준비

패키지 참조

데이터 가져오기

#02. 데이터 전처리

분석 대상 컬럼만 추출

#03. 산점도 행렬

1. 기본 사용

2. 파라미터 설정

대각선 기준 다르게 표현하기

#04. 인터렉티브 그래프

상관계수 행렬

산점도 행렬 시각화

분석결과 검정

#04. 인터렉티브 그래프

`plotly`, `cufflinks` 패키지 설치 필요

```
import plotly.figure_factory as ff

# Cufflinks wrapper on plotly
import cufflinks as cf

# plotly + cufflinks in offline mode
from plotly.offline import iplot
cf.go_offline()

# set the global theme
cf.set_config_file(world_readable=True, theme='pearl', offline=True)
```

```
fig = ff.create_scatterplotmatrix(
    df2,
    height=1200,
    width=1200,
    diag='histogram')  # scatter, histogram, box

iplot(fig)
```

여러 변수의 상관분석

#01. 작업준비

패키지 참조

데이터 가져오기

#02. 데이터 전처리

분석 대상 컬럼만 추출

#03. 산점도 행렬

1. 기본 사용

2. 파라미터 설정

대각선 기준 다르게 표현하기

#04. 인터렉티브 그래프

상관계수 행렬

산점도 행렬 시각화

분석결과 검정

여러 변수의 상관분석

#01. 작업준비

패키지 참조

데이터 가져오기

#02. 데이터 전처리

분석 대상 컬럼만 추출

#03. 산점도 행렬

1. 기본 사용

2. 파라미터 설정

대각선 기준 다르게 표현하기

#04. 인터렉티브 그래프

상관계수 행렬

산점도 행렬 시각화

분석결과 검정

상관계수 행렬

여러 개의 변수를 갖는 데이터프레임에 대해서도 사용 가능

```
df3 = df2.drop('vs', axis=1)
corr = df3.corr(method='pearson')
corr
```

	mpg	disp	hp	drat	wt	qsec
mpg	1.000000	-0.847551	-0.776168	0.681172	-0.867659	0.418684

여러 변수의 상관분석

#01. 작업준비

패키지 참조

데이터 가져오기

#02. 데이터 전처리

분석 대상 컬럼만 추출

#03. 산점도 행렬

1. 기본 사용

2. 파라미터 설정

대각선 기준 다르게 표현하기

#04. 인터렉티브 그래프

상관계수 행렬

산점도 행렬 시각화

분석결과 검정

	mpg	disp	hp	drat	wt	qsec
disp	-0.847551	1.000000	0.790949	-0.710214	0.887980	-0.433698
hp	-0.776168	0.790949	1.000000	-0.448759	0.658748	-0.708223
drat	0.681172	-0.710214	-0.448759	1.000000	-0.712441	0.091205
wt	-0.867659	0.887980	0.658748	-0.712441	1.000000	-0.174716
qsec	0.418684	-0.433698	-0.708223	0.091205	-0.174716	1.000000

산점도 행렬 시각화

```
plt.rcParams["figure.figsize"] = (10,8)
sb.heatmap(df3.corr(method='pearson'), annot = True, cmap = 'Greens', vm
plt.show()
plt.close()
```

여러 변수의 상관분석

#01. 작업준비

패키지 참조

데이터 가져오기

#02. 데이터 전처리

분석 대상 컬럼만 추출

#03. 산점도 행렬

1. 기본 사용

2. 파라미터 설정

대각선 기준 다르게 표현하기

#04. 인터렉티브 그래프

상관계수 행렬

산점도 행렬 시각화

분석결과 검정

분석결과 검정

여러 변수의 상관분석

#01. 작업준비

패키지 참조

데이터 가져오기

#02. 데이터 전처리

분석 대상 컬럼만 추출

#03. 산점도 행렬

1. 기본 사용

2. 파라미터 설정

대각선 기준 다르게 표현하기

#04. 인터렉티브 그래프

상관계수 행렬

산점도 행렬 시각화

분석결과 검정

```
print(stats.pearsonr(df3['mpg'], df3['disp']))
print(stats.pearsonr(df3['disp'], df3['hp']))
print(stats.pearsonr(df3['hp'], df3['drat']))
print(stats.pearsonr(df3['drat'], df3['wt']))
print(stats.pearsonr(df3['wt'], df3['qsec']))
print(stats.pearsonr(df3['qsec'], df3['mpg']))
```

```
PearsonRResult(statistic=-0.8475513792624787, pvalue=9.380326537381379e-
PearsonRResult(statistic=0.7909485863698065, pvalue=7.14267865573725e-08
PearsonRResult(statistic=-0.4487591168729195, pvalue=0.00998877189352624
PearsonRResult(statistic=-0.7124406466973718, pvalue=4.7842600661325326e
PearsonRResult(statistic=-0.17471587871340488, pvalue=0.3388682841349162
PearsonRResult(statistic=0.41868403392177833, pvalue=0.01708198849651956
```