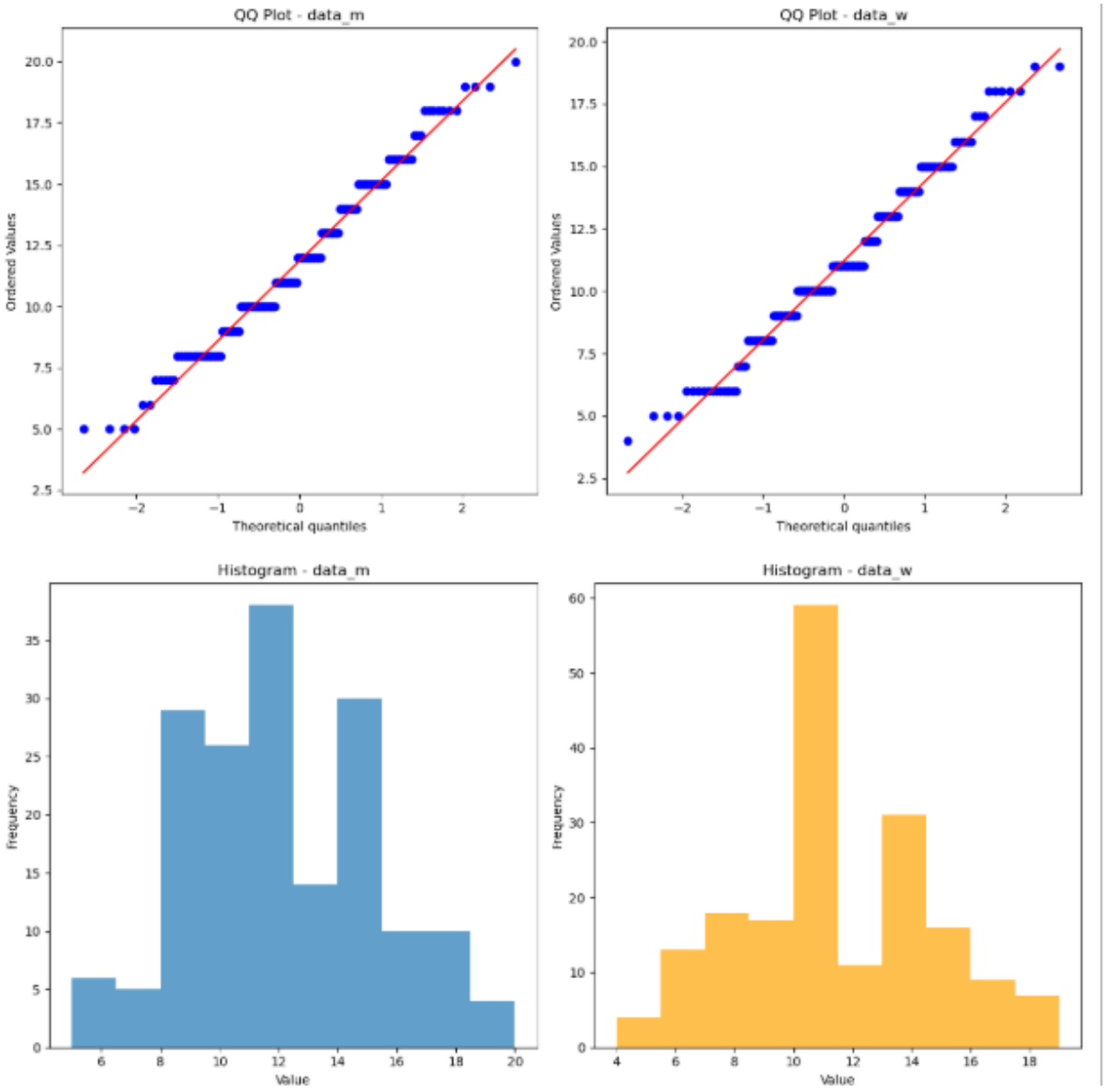
• 9_1 student grade example에서 남,녀 의 수학 기말고사 점수가 차이가 나는지 test를 하였다. 하지만 t-test를 하기 위한 조건을 체크 하지 않고 진행을 하였다. 이를 통계적 가정에 맞춰서 다시 분석을 진행하세요.

t-test를 하기 위해 내가 생각한 조건 -각 데이터가 정규분포를 따르는지

이유 - t-test는 두 집단의 데이터가 정규분포를 따르고 있다고 가정 , t-test가 정규분포의 특성을 활용하여 표본 평균의 차이를 평가, 따라서 왜곡될 가능성이 있기에 확인해야 함



Qqplot과 히스토그램을 그려봤는데 시각적으로 정규분포를 따른다고 판단하는 것은 명확한 수치도 아니고 주관적인 판단일 수 있으므로 테스트를 통해 정규성 검정을 하려한

다. – 주관적 판단으로는 정규분포를 따르는 것 같지는 않다.

```
print(len(data_m))
print(len(data_w))
```

executed in 20ms, finished 22

```
172
185
```

샘플의 개수를 확인

이 정도면 샤피로 테스트를 적용했을 때 올바른 결론이 나올 확률이 높다. 샘플갯수가 너무크면 민감한 샤피로 테스트는 pvalue가 낮게 나올 가능성이 높아지기 때문이다.

9R에서는 5000개 이하에서만 수행가능)

```
data_m.describe()
```

executed in 140ms, finished 14:06:15 2024-12-08

```
count    172.000000
mean      11.866279
std        3.258748
min        5.000000
25%       10.000000
50%       12.000000
75%       14.000000
max       20.000000
Name: G3, dtype: float64
```

```
data_w.describe()
```

executed in 30ms, finished 14:06:16 2024-12-08

```
count    185.000000
mean      11.205405
std        3.174452
min        4.000000
25%        9.000000
50%       11.000000
75%       13.000000
max       19.000000
Name: G3, dtype: float64
```

통계치 간단하게 확인해주고

```
print("data_m의 정규성",stats.shapiro(data_m))
print("data_w의 정규성",stats.shapiro(data_w))
```

executed in 29ms, finished 14:07:02 2024-12-08

```
data_m의 정규성 ShapiroResult(statistic=0.9788146615028381, pvalue=0.009935521520674229)
data_w의 정규성 ShapiroResult(statistic=0.9785570502281189, pvalue=0.006041768938302994)
```

샤피로테스트를 해봤더니 둘다 정규분포를 따른다는 귀무가설을 기각해야 된다고 나왔다. 지금 데이터의 샘플 개수를 고려하면 샤피로가 가장 유효한 검정이라고 판단되는데 다른 검정들도 해보겠다.

```
mean_m, std_m = np.mean(data_m), np.std(data_m)
mean_w, std_w = np.mean(data_w), np.std(data_w)

print("ks_test & data_m의 결과 ",stats.kstest(data_m, 'norm', args=(mean_m, std_m)))
print("ks_test & data_w의 결과" ,stats.kstest(data_w, 'norm', args=(mean_w, std_w)))
```

executed in 136ms, finished 12:58:54 2024-12-08

```
ks_test & data_m의 결과  KstestResult(statistic=0.10086228450774237, pvalue=0.05628834890664047)
ks_test & data_w의 결과 KstestResult(statistic=0.12586577693158563, pvalue=0.005161903291796414)
```

ks test에서는 남자들의 성적분포가 정규분포를 따를 수 있다는 귀무가설 유지

```
from scipy.stats import jarque_bera
stat, p_value = jarque_bera(data_w)
print("Test Statistic:", stat)
print("p-value:", p_value)
```

executed in 34ms, finished 14:09:20 2024-12-08

```
Test Statistic: 2.2043559063770712
p-value: 0.3321468949708064
```

```
from scipy.stats import jarque_bera
stat, p_value = jarque_bera(data_m)
print("Test Statistic:", stat)
print("p-value:", p_value)
```

executed in 16ms, finished 14:21:37 2024-12-08

```
Test Statistic: 3.0406254237317034
p-value: 0.2186435038426241
```

```
from scipy.stats import normaltest
stat, p_value = normaltest(data_w)
print("Test Statistic:", stat)
print("p-value:", p_value)
```

executed in 22ms, finished 14:09:21 2024-12-08

```
Test Statistic: 2.323816262090783
p-value: 0.31288857850171053
```

```
from scipy.stats import normaltest
stat, p_value = normaltest(data_m)
print("Test Statistic:", stat)
print("p-value:", p_value)
```

executed in 34ms, finished 14:21:37 2024-12-08

```
Test Statistic: 3.7174901419749595
p-value: 0.15586811109986803
```

왜도와 첨도를 기준으로 정규성 검증을 해보니 둘다 정규 분포를 따른다는 귀무가설 유지 결과가 나왔다. 사용한 테스트 – 자크베라,normaltest

정규성에 대한 결론 - 샤피로 테스트를 근거로 정규성이 없다는 대립가설을 채택

첨도외 왜도를 기준으로는 정규분포와 유사하다 볼 수 있지만 분포의 작은 비대칭성이나 이상치는 무시될 있음음 샤피로-윌크 검정은 이러한 미세한 특징을 잡아내는 데 적합

#샘플의 갯수가 100~200 이므로 샤피로 테스트가 민감하게 반응하는 것이 오히려 정확할 것이라고 판단

#샘플의 갯수가 너무 많으면 너무 민감한게 오히려 독이 될 수 있겠지만 이 샘플에서는 유효한 검정이라고 생각한다.

```
from scipy.stats import levene
from scipy.stats import fligner
print("levene테스트 결과:",stats.levene(data_m, data_w))
print("fligner테스트 결과:",stats.fligner(data_m,data_w))
print("결론: 분산이 같다는 귀무가설을 기각할 수 없다")
```

executed in 22ms, finished 13:16:20 2024-12-08

```
levene테스트 결과: LeveneResult(statistic=0.6144552033049334, pvalue=0.4336379490863752)
fligner테스트 결과: FlignerResult(statistic=0.4455866504689745, pvalue=0.5044382692544688)
결론: 분산이 같다는 귀무가설을 기각할 수 없다
```

정규분포를 따르지 않는다는 결론으로 등분산성 검정을 해야되는데 levene , filgner 테스트가 적합 – 둘의 결론으로 분산이 같다는 귀무가설을 기각할 수 없다.

```
from scipy.stats import ranksums

w_stat, p_ranksum = ranksums(data_m, data_w)
print(f"Wilcoxon Rank-Sum Test p-value: ",ranksums(data_m, data_w))
```

executed in 23ms, finished 13:56:36 2024-12-08

```
Wilcoxon Rank-Sum Test p-value:  RanksumsResult(statistic=1.764823191570922, pvalue=0.077593485676993)
```

따라서 비모수적 방법을 사용해 데이터의 분포에 대한 가정을 하지 않으므로, 정규성을 따를 필요가 없습니다Ranksum 테스트 사용

이유 - 데이터의 분포에 대한 가정을 하지 않으므로, 정규성을 따를 필요가 없다.

두 독립적인 그룹 간 중앙값의 차이를 비교

결론 – 남녀 학생의 성적 차이가 없다 귀무가설을 기각할 수 없다.

추가적으로

표본의 독립성 검증

data_m은 남성 그룹에서, data_w는 여성 그룹에서 수집된 데이터라면 독립적일 가능성 높다

+ 서로가 서로의 성적에 영향을 주지 않는다 판단 – 관측값들간의 독립성

하지만 무수히 많은 컬럼이 있고 성별의 따른 성적으로 차이를 보기에는 성적에 영향을 줄 컬럼들이 많은 것으로 판단하였다.(예시 주당 공부시간 등) 추가적으로 (밑에 글)

1. **school** - 학생이 다니는 학교 (이진 값: 'GP' - Gabriel Pereira 또는 'MS' - Mousinho da Silveira)
2. **sex** - 학생의 성별 (이진 값: 'F' - 여자, 'M' - 남자)
3. **age** - 학생의 나이 (숫자형: 15세부터 22세까지)
4. **address** - 학생의 집 주소 유형 (이진 값: 'U' - 도시, 'R' - 농촌)
5. **famsize** - 가족 규모 (이진 값: 'LE3' - 3명 이하, 'GT3' - 3명 초과)
6. **Pstatus** - 부모의 동거 상태 (이진 값: 'T' - 함께 거주, 'A' - 따로 거주)
7. **Medu** - 어머니의 학력 (숫자형: 0 - 없음, 1 - 초등학교 (4학년), 2 - 5~9학년, 3 - 고등학교, 4 - 대학교 이상)
8. **Fedu** - 아버지의 학력 (숫자형: 0 - 없음, 1 - 초등학교 (4학년), 2 - 5~9학년, 3 - 고등학교, 4 - 대학교 이상)
9. **Mjob** - 어머니의 직업 (명목형: 'teacher' - 교사, 'health' - 의료 관련, 'services' - 행정 또는 경찰, 'at_home' - 가정주부, 'other' - 기타)
10. **Fjob** - 아버지의 직업 (명목형: 'teacher' - 교사, 'health' - 의료 관련, 'services' - 행정 또는 경찰, 'at_home' - 가정주부, 'other' - 기타)
11. **reason** - 이 학교를 선택한 이유 (명목형: 'home' - 집에서 가까움, 'reputation' - 학교 평판, 'course' - 학업 선호도, 'other' - 기타)
12. **guardian** - 학생의 보호자 (명목형: 'mother' - 어머니, 'father' - 아버지, 'other' - 기타)
13. **traveltime** - 집에서 학교까지 통학 시간 (숫자형: 1 - 15분 미만, 2 - 15~~30분, 3 - 30분~~1시간, 4 - 1시간 초과)
14. **studytime** - 주당 공부 시간 (숫자형: 1 - 2시간 미만, 2 - 2~~5시간, 3 - 5~~10시간, 4 - 10시간 초과)
15. **failures** - 과거 수업 낙제 횟수 (숫자형: nnn - 1 이상 3 미만, 또는 4 이상)
16. **schoolsup** - 추가 학습 지원 여부 (이진 값: 'yes' - 있음, 'no' - 없음)
17. **famsup** - 가족 학습 지원 여부 (이진 값: 'yes' - 있음, 'no' - 없음)
18. **paid** - 수업 과목(수학 또는 포르투갈어) 내 추가 수업 여부 (이진 값: 'yes' - 있음, 'no' - 없음)
19. **activities** - 방과 후 활동 여부 (이진 값: 'yes' - 있음, 'no' - 없음)
20. **nursery** - 유치원을 다녔는지 여부 (이진 값: 'yes' - 다님, 'no' - 다니지 않음)
21. **higher** - 고등 교육(대학교 등)을 원하는지 여부 (이진 값: 'yes' - 원함, 'no' - 원하지 않음)
22. **internet** - 집에서 인터넷 사용 가능 여부 (이진 값: 'yes' - 가능, 'no' - 불가능)
23. **romantic** - 연애 중인지 여부 (이진 값: 'yes' - 연애 중, 'no' - 연애 중 아님)
24. **famrel** - 가족 관계의 질 (숫자형: 1 - 매우 나쁨, 5 - 매우 좋음)
25. **freetime** - 방과 후 자유 시간 (숫자형: 1 - 매우 적음, 5 - 매우 많음)
26. **goout** - 친구와 외출 빈도 (숫자형: 1 - 매우 적음, 5 - 매우 많음)
27. **Dalc** - 평일 알코올 소비량 (숫자형: 1 - 매우 적음, 5 - 매우 많음)
28. **Walc** - 주말 알코올 소비량 (숫자형: 1 - 매우 적음, 5 - 매우 많음)
29. **health** - 현재 건강 상태 (숫자형: 1 - 매우 나쁨, 5 - 매우 좋음)
30. **absences** - 학교 결석 횟수 (숫자형: 0부터 93까지)

학생들이 서로 다른 개인으로 간주 즉,

한 학생의 데이터가 다른 학생의 데이터에 영향을 주지 않는 경우.

하지만, 같은 학교나 가족 구조 또는 친구 관계에서 발생할 수 있는 의존성이 독립성을 위반할 가능성이 존재해 고려필요

```
print(len(data_m))
print(len(data_w))
```

executed in 10ms, finished

```
258
376
```

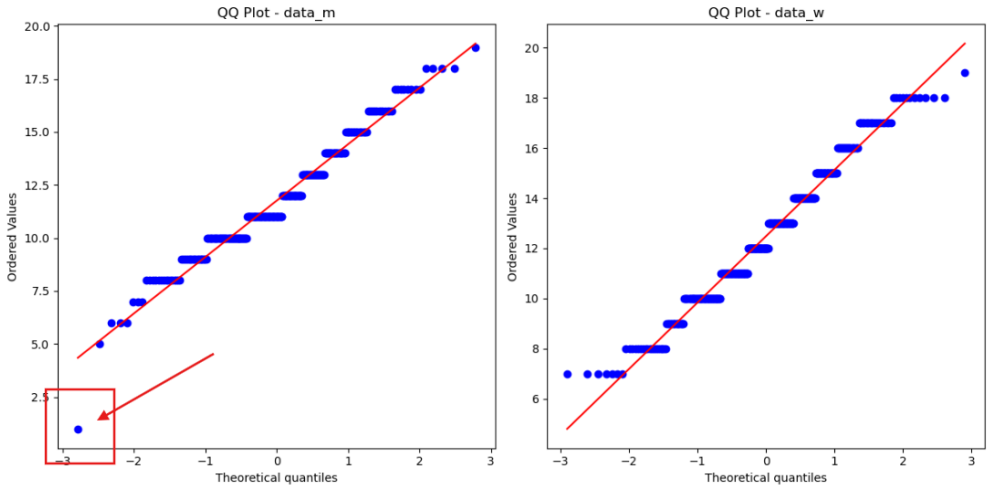
데이터의 개수가 조금 차이남

```
print(data_m.describe())
print(data_w.describe())
```

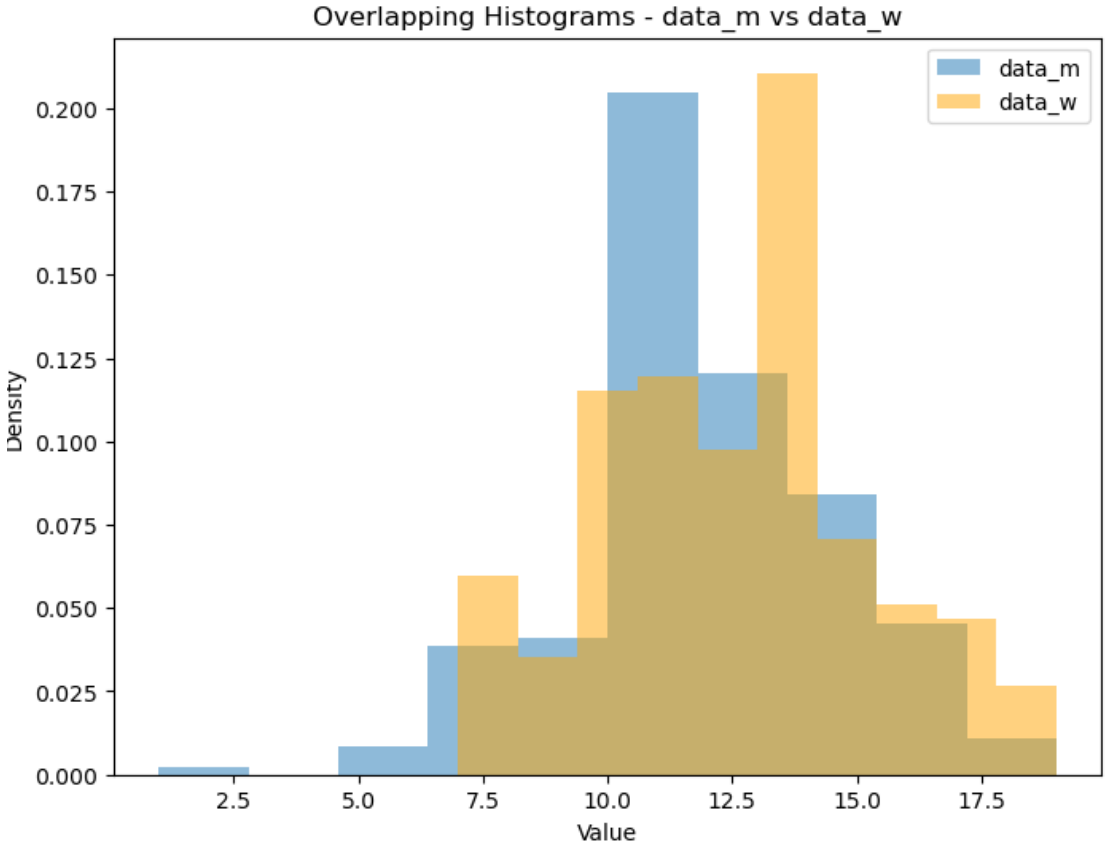
executed in 30ms, finished 15:10:00 2024-12-0

```
count    258.000000
mean      11.759690
std        2.682339
min        1.000000
25%       10.000000
50%       11.000000
75%       13.750000
max       19.000000
Name: G3, dtype: float64
count    376.000000
mean      12.481383
std        2.662265
min        7.000000
25%       10.000000
50%       12.000000
75%       14.000000
max       19.000000
Name: G3, dtype: float64
```

평균도 조금 차이남 – 과연 유의미한 차이일까?



남자 데이터중 점수가 1인 사람이 있음 – outlier긴 한데 1~20의 바운더리에 있는 성적 데이터고 시험을 보긴 봤으므로 제거하기 어려움



겹쳐서 히스토그램을 보니 전체적으로 여자가 성적을 잘 본 것처럼 나왔다

정말 그런지 테스트를 통해 확인해봤다 (정규성 검정, 등분산성 검정)

```
print("data_m의 정규성",stats.shapiro(data_m))
print("data_w의 정규성",stats.shapiro(data_w))
```

executed in 104ms, finished 15:16:00 2024-12-08

```
data_m의 정규성 ShapiroResult(statistic=0.9713147282600403, pvalue=4.664006701204926e-05)
data_w의 정규성 ShapiroResult(statistic=0.9747232794761658, pvalue=3.8834627957839984e-06)
```

둘 다 귀무가설을 기각할 수 없게나옴 (정규분포를 따른다고 할 수 없다.)

```
mean_m, std_m = np.mean(data_m), np.std(data_m)
mean_w, std_w = np.mean(data_w), np.std(data_w)

print("ks_test & data_m의 결과 ",stats.kstest(data_m, 'norm', args=(mean_m, std_m)))
print("ks_test & data_w의 결과" ,stats.kstest(data_w, 'norm', args=(mean_w, std_w)))
```

executed in 158ms, finished 15:16:11 2024-12-08

```
ks_test & data_m의 결과  KstestResult(statistic=0.1427142905103481, pvalue=4.767702005030989e-05)
ks_test & data_w의 결과 KstestResult(statistic=0.10757389252617444, pvalue=0.0003040247723250106)
```

KS 테스트도 마찬가지로 정규분포를 따른다고 할 수 없게 나온다.

```
from scipy.stats import jarque_bera
stat, p_value = jarque_bera(data_m)
print("Test Statistic:", stat)
print("p-value:", p_value)
```

executed in 36ms, finished 15:16:18 2024-12-08

```
Test Statistic: 3.088497365471416
p-value: 0.21347219633441383
```

```
from scipy.stats import jarque_ber
stat, p_value = jarque_bera(data_w)
print("Test Statistic:", stat)
print("p-value:", p_value)
```

executed in 11ms, finished 15:26:40 2024-12-08

```
Test Statistic: 7.555644810811773
p-value: 0.022872444145530 58
```

```
from scipy.stats import normaltest
stat, p_value = normaltest(data_m)
print("Test Statistic:", stat)
print("p-value:", p_value)
```

executed in 62ms, finished 15:16:25 2024-12-08

```
Test Statistic: 2.9847340734493 58
p-value: 0.2248398210552 1774
```

```
from scipy.stats import normaltest
stat, p_value = normaltest(data_w)
print("Test Statistic:", stat)
print("p-value:", p_value)
```

executed in 67ms, finished 15:26:39 2024-12-08

```
Test Statistic: 13.478901038958876
p-value: 0.001183297176761 1806
```

왜도와 첨도를 기준으로 정규성 검증을 해보니 둘다 정규 분포를 따른다는 귀무가설 유지 결과가 나왔다. 사용한 테스트 – 자크베라,normaltest

```
from scipy.stats import levene
from scipy.stats import fligner
print("levene테스트 결과:",stats.levene(data_m, data_w))
print("fligner테스트 결과:",stats.fligner(data_m,data_w))
print("결론: 분산이 같다는 귀무가설을 기각할 수 없다")
```

executed in 55ms, finished 15:16:32 2024-12-08

```
levene테스트 결과: LeveneResult(statistic=0.5699053541647334, pvalue=0.45057774863577 17)
fligner테스트 결과: FlignerResult(statistic=0.37879627131036775, pvalue=0.538248156579413)
결론: 분산이 같다는 귀무가설을 기각할 수 없다
```

등분산성 검정 – 분산이 같다는 귀무가설 채택

```
from scipy.stats import ranksums

w_stat, p_ranksum = ranksums(data_m, data_w)
print(f"Wilcoxon Rank-Sum Test p-value: ",ranksums(data_m, data_w))
```

executed in 30ms, finished 15:17:35 2024-12-08

```
Wilcoxon Rank-Sum Test p-value:  RanksumsResult(statistic=-3.2557446208016954, pvalue=0.001130954259649573)
```

Ranksums 테스트 결과 둘의 평균 차이가 유의미 하다라는 대립가설 채택

이유 -pvalue가 0.05보다 낮으므로 정말 차이가 존재한다.

추가적으로

```
: from scipy.stats import mannwhitneyu


u_stat, p_mwu = mannwhitneyu(data_m, data_w, alternative='less')
print(f"Mann-Whitney U Test p-value: {p_mwu}")
```

executed in 28ms, finished 15:32:13 2024-12-08

```
Mann-Whitney U Test p-value: 0.0005226337837144511
```

Mannwhitneyu test 코드에 alternative="less" 추가해 돌려보니 p-value가 작아서 data_m이 data_w보다 작다는 대립가설 채택

즉, 여자의 포루투칼어 성적이 유의미하게 높다고 말할 수 있다.

추가적 고려

하지만 첫번째 데이터와 동일하듯 학교마다 포루투칼어 수업이 있는 학교의 여학생이 더 많을 수도 있고 다른 무수한 컬럼들이 포루투칼어의 성적에 관련되기 때문에 꼭 성별에 의한 차이가 아닐 가능성도 배제할 수는 없다.