변수	통계/시각화	검정	특이사항

숫자 1	min,max,mean,std, median,사분위수
	
	sns.histplot(data=df, x='v', kde=True)
	sns.boxplot(data=df, x='v')
	sns.kdeplot(data=df, x='v')
	sns.displot(data=df, x='v')
		<정규분포(정규성) 검정>
		stats.shapire(df.v) # 5000 미만
		stats.anderson(df.v) # 5000 이상
		stats.kstest(df.v, 'nomal')
		
		귀무가설: 정규분포이다.
			<이상치 판단/처리>
			①사분위수 이용
			ITQ=Q3-Q1
			low=Q1-1.5*ITQ, //high=Q3+1.5ITQ
			df2=df[(df.v <high) & (df.v >low)]
			② zscore
			z=np.abs(df.v -df.v.mean()/df.v.std()
			df2=df[z <2.5]

범주 1	범주별 빈도수(value_counts)/ 범주별 비율 sns.barplot(data=df, x='V') sns.countplot(data=df, x='V')	* 통계, 시각화 추가 p=df.V.value_counts() plt.pie(p.values, label=p.index, autopc='%.2f%%')	
범주(종속) 숫자(독립)	V_1=df[df.V ==0] V_2=df[df.V ==1] sns.kdeplot(x='v', data=V_1) sns.kdeplot(x='v', data=V_2) sns.histplot(x='v', data=df, hue=V)	<로지스틱 회귀모형을 통한 검정> 회귀계수 0인가 import statsmodels.api as sm model=sm.Logit(df.V, df.v) result=model.fit() 귀무: 해당 변수의 회귀계수가 0이다 --독립과 종속이 관련이 없다.	* 로지스틱 회귀모형: 모집단에 대한 가설 아님 그러므로 범주(숫자), 독립(종속) 일때와 같은 검정을 하고 결과를 뒤집어서 적용 : A~B는 !B~!A
숫자2	plt.scatter('v1', 'v2', data=df) sns.regplot(data=df, x='v1', y='v2') sns.relplot(data=df, x='v1', y='v2', hue='V') sns.lmplot(data=df, x='v1', y='v2') g=sns.FacetGrid(df, row=v1, col=v2, size=) g.map(sns.scatterplot,'v3', ' v4 ') sns.pairplot(df)	<상관분석> 상관계수를 통한 분석 stats.pearson(df.x1, df.x2) 종류 spearman( ), kendalltau( ) 귀무: 상관계수는 0이다. 통계량: 상관계수	상관계수: 데이터가 얼마나 직선으로 모여있는지 수치화한 값 (-1~1 : 양의 상관관계/ 음의 상관관계, 0에 가까우면 관련이 없음) # 정규분포 아닐 경우 비모수분석: spearman
숫자 1,2,3...	sns.heatmap(df, vmin=-1, vmax=1, annot=True)	<상관관계> df.corr(numeric_only=True) 판다스의 .corr() 함수	

범주2	교차표
	cr=pd.crosstab(df.V1, df.V2)
	cr=pd.crosstab(df.V1, df.V2, normalize='index')
	: normalize='index' /'columns'/'all'
	
	cr.plot.bar(stacked=True)
	sns.countplot(data=df, x=v, hue='X')
		<독립성 검정 x2(카이제곱)>
		cr=pd.crosstab(df.V1, df.V2) #교차표
		stats.ch2_contingency(cr)
		
		귀무: 두 변수가 독립적이다(연관성 없음)
			카이제곱 검정:
			예측값과 관측값의 차이 측정
			①적합도(선호도) 검정: 변수1
			②독립성 검정: 하나의 모집단에 범주형 2개
			:범주변수들 간 어떤 관계가 있는지
			③동질성 검정: 2개 이상 모집단 2개 이상

범주(독립)
숫자(종속)
	<정규성 검정>
	V_1=df[df.V ==0]['v']
	V_2=df[df.V ==1]['v']
	sns.histplot(V_1, kde=True)
	sns.histplot(V_2, kde=True)
		<단일표본 T검정(one-sample T test)>
		하나의 집단평균이 모평균과 동일한지 검정(신뢰구간)
		stats.ttest_1samp(V_1, df.1.mean())
		stats.ttest_1samp(V_2, df.V.mean()) #귀무: 모평균과 동일
		
		<정규성 검정> --표준정규분포를 갖는지
		stats.shapiro(V_1)
		stats.shapiro(V_2) # 귀무: 정규분포을 가진다
		
		<등분산 검정> --같은 분산을 갖는지 #귀무: 갖는다
		stats.bartett(V_1,V_2) # 표본크기 크고 정규성 가진 데이터
		stats.levene(V_1,V_2) # 정규분포에 영향 안받음(중앙값)
		stats.fligner(V_1,V_2) # 정규분포 안따르는 데이터/중앙값
			<신뢰구간 구하기>
			n=len(v)-1 #자유도
			mn=np.mean(v) #표준편차
			sm=stats.sem(v) #표준오차
			stats.t.interval(0.95, n, mn, sm)
			: 95% 신뢰구간 -1.96~1.96
			
			==>정규성 검정에서 정규성이 아닐때
			<Mann-Whitney U test 검정> 중앙값 차이
			표본수가 적을때 비모수적인 분석
			// 평균비교X, 순위합검정
			stats.mannwhitneyu(V_1,V_2)
			귀무: 두 집단의 중앙값 차이가 없다.

범주2(독립) 숫자1(종속) :종속변수 독립적	sns.barplot(x=V, y=v, data=df) # V 범주들 모두 그래프에 표시됨	<독립표본 T검정> 모집단이 서로 다를때 두 평균의 차이 stats.ttest_ind(V_1,V_2, equal_var=False) 귀무: 유의미한 평균 차이가 없다. t통계량 결과 -2> 또는 >2이면 '차이가 크다'.	비모수검정: 독립적 집단 2, 기간1 <Mann-Whitney U test 검정>
범주2(독립) 숫자1(종속) :종속변수 기간의 변화	(예) happy1 = happy['점수_2019'] - happy['점수_2018'] plt.bar(happy['나라명'], happy1.values) # 증가,감소 보기	<대응표본 T검정> 집단 1, 두 기간에서의 변화 stats.ttest_rel(V_1,V_2, alternative='greater') 귀무:증가한다 ==> alternative='less' 귀무: 감소한다 V_1(before), V_2(after)	<비모수검정> 집단 1, 두 기간에서의 변화 stats.wilcoxon(V_1, V_2, alternative='greater')

범주3 이상 (독립) 숫자1(종속)	V_1=df[df.V ==0]['v'] V_2=df[df.V ==1]['v'] V_3=df[df.V ==2]['v'] sns.barplot(x='V', y='v', data=df)	<F검정(ANOVA)> 신뢰구간 검정 여러 집단의 분산을 통한 집단의 차이가 있는지 stats.f_oneway('V_1','V_2','V_3') 귀무: 차이가 없다 f통계량 결과 2-3이상의 값을 가지면 차이가 있다	<Welchis ANOVA> 3개 집단 등분산 아닐때 pg.welch_anova(data=df,dv=v,between=v) 비모수검정 <kruskal-walis test> : 집단1,두 기간 stats.kruskal(V_1,V_2,V_3) <Friedman test> : 집단1, 3개 이상 기간 stats.friedmanchisquare()