Multivariate Data Analysis Assignment #1

Multiple Linear Regression (MLR)

산업경영공학부 2020170856 이우진

[Q1] 본인이 스스로 Multivariate Linear Regression 모델을 적용하는데 적합한 데이터셋을 선정하고 선정이유를 설명하시오. 데이터셋 탐색은 아래에서 제시된 Data Repository를 포함하여 여러 Repository를 검색해서 결정하시오. 보고서에는 데이터를 다운로드할 수 있는 링크를 반드시 제공하시오.

Dataset : **Predict Purity and Price of Honey**

다운로드 링크 : https://www.kaggle.com/datasets/stealthtechnologies/predict-purity-and-price-of-honey

선정 이유 : 해당 데이터셋의 종속 변수가 연속형 변수이며, Instance의 개수가 247903개로 많은 데이터를 가지고 있다. 종속 변수로 사용할 Price를 제외하고, 총 10개의 설명 변수들을 가지고 있으며, 해당 데이터가 Numerical Data 뿐 아니라 Categorical Data도 포함하고 있기 때문에 수업 때 배운 1-of-C coding 변환도 활용할 수 있다고 생각하였다. 또한, 꿀이 어떤 꽃을 통해 만들어졌는지, 꿀의 순도, 밀도, 수분 함량 등의 변수들이 종속 변수인 가격에 영향을 미칠 것이라 생각하였고, 따라서 해당 데이터셋에 MLR 모델을 적용하는데 적합하다 판단하여 선정하였다.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 247903 entries, 0 to 247902
Data columns (total 11 columns):
 #   Column           Non-Null Count   Dtype
---  ------           --------------   -----
 0   CS               247903 non-null  float64
 1   Density          247903 non-null  float64
 2   WC               247903 non-null  float64
 3   pH               247903 non-null  float64
 4   EC               247903 non-null  float64
 5   F                247903 non-null  float64
 6   G                247903 non-null  float64
 7   Pollen_analysis  247903 non-null  object
 8   Viscosity        247903 non-null  float64
 9   Purity           247903 non-null  float64
 10  Price            247903 non-null  float64
dtypes: float64(10), object(1)
memory usage: 20.8+ MB
```

[Q2] 해당 데이터셋의 종속변수와 설명변수는 어떤 것들이 있는가? 분석 전에 아래 세 가지 질문에 대해서 스스로 생각해보고 답변을 하시오.

해당 데이터 셋에는 1개의 종속 변수와 10개의 설명 변수가 있다.

<종속 변수>

- Price : 꿀의 계산된 가격

<설명 변수>

- CS (Color Score) : 꿀 샘플의 색상 점수, 1.0~10.0 범위로 값이 높을수록 어두운 색.
- Density : 꿀 샘플의 밀도, 1.21 to 1.86 범위
- WC (Water Content) : 꿀 샘플의 수분 함량, 12.0% ~ 25.0% 범위
- pH : 꿀 샘플의 pH 수준, 2.50~7.50 범위
- EC (Electrical Conductivity) : 꿀 샘플의 전기 전도도
- F (Fructose Level) : 꿀 샘플의 과당 수준, 20~50 범위
- G (Glucose Level) : 꿀 샘플의 포도당 수준, 20~45 범위
- Pollen_analysis : 꿀 샘플의 꽃가루 출처
- Viscosity : 꿀 샘플의 점도, 1500~10000 범위
- Purity : 꿀 샘플의 순도, 0.01~1.00 범위

2-1. 이 데이터는 종속변수와 설명변수들 사이에 실제로 "선형 관계"가 있다고 가정할 수 있겠는가? 가정할 수 있음/없음 판단에 대한 본인의 생각을 서술하시오.

분석에 앞서 명목형 변수인 'Pollen_analysis'를 살펴보기 위해 dataframe의 unique 함수를 사용하면 아래와 같은 결과를 얻을 수 있다.

```
['Blueberry' 'Alfalfa' 'Chestnut' 'Borage' 'Sunflower' 'Orange Blossom'
 'Acacia' 'Tupelo' 'Clover' 'Wildflower' 'Thyme' 'Sage' 'Avocado'
 'Lavender' 'Eucalyptus' 'Buckwheat' 'Rosemary' 'Heather' 'Manuka']
```

'Pollen_analysis'에는 총 19개의 다양한 종류의 꽃가루 출처들이 있고, 꿀의 꽃가루 출처에 따라 꿀의 특징 및 사람들의 선호도가 달라지기 때문에, 종속 변수인 Price에 큰 영향을 미칠 것이라 생각하고 의미있는 상관관계를 가질 것으로 보인다. 또한, 꿀을 구매하는데 있어, 순수한 꿀의 선

호도가 매우 높기 때문에 꿀의 순도 또한 종속 변수와 선형 관계를 가질 것으로 보인다.

그러나 꿀 샘플의 Electrical Conductivity, pH, Viscosity 등은 사람들이 꿀을 구매할 때, 신경 쓰는 요소가 아니기 때문에, 가격에 크게 영향을 미치지 않는다고 생각하고, 이와 같은 변수들은 선형 관계가 없을 것으로 보인다.

독립 변수마다 차이가 있겠지만, 꿀의 가격에 영향을 줄 수 있는 설명 변수들이 존재한다고 생각하기 때문에 이 데이터는 종속 변수와 설명 변수들 사이에 '선형 관계'가 있다고 가정할 수 있다.

2-2. 이 데이터에서 제공된 설명변수들 중에서 높은 상관관계가 있을 것으로 예상되는 변수들은 어떤 것들이 있는가? 왜 그렇게 생각하는가?

종속변수 Price와 높은 상관관계가 있을 것으로 예상되는 변수에는 'Pollen_analysis', 'Purity'가 있다. 2-1에서 답한 것처럼 Price에 직접적인 영향을 미치는 요소가 '사람들의 선호' 라고 생각하는데, 아카시아 꿀, 밤 꿀처럼, 특정 꿀의 꽃가루 출처가 사람들의 선호도에 큰 영향을 주기 때문에 높은 상관관계가 있을 것이라고 생각한다. 동일한 이유로 사람들이 꿀을 구매할 때, 불순물이 첨가되지 않은 순도 높은 꿀을 원하기 때문에 'Purity'도 높은 상관관계를 가질 것으로 예상된다.

이것과 관련하여 WC(water Content), Density 등의 변수는 'Purity'와 서로 영향을 줄 수 있는 변수라고 생각하여, 이러한 변수들도 상관관계가 있을 가능성이 존재한다고 생각한다.

2-3. 제공된 설명변수들 중에서 종속변수를 예측하는데 필요하지 않을 것으로 예상되는 변수들은 어떤 것들이 있는가? 왜 그렇게 생각하는가?

종속변수 Price를 예측하는데 필요하지 않을 것으로 예상되는 변수에는 CS (Color Score), pH, EC (Electrical Conductivity) 등이 있다. 이들은 평소 사람들이 꿀을 구매할 때, 고려하는 요소가 아니기 때문에 가격을 예측하는 데 필요하지 않을 것이라고 생각한다. 또한, 'Pollen_analysis'의 차이 즉, 근본적인 꿀의 출처 차이로 인하여 pH, EC와 같은 꿀의 특징들도 차이를 보일 것으로 예상하지만, 해당 변수들이 Price와 선형관계를 띄고 있다고 보기엔 어렵다고 생각한다.

[Q3] 개별 입력 변수들에 대하여 각각 다음과 같은 단변량 통계량을 계산하고 Box plot을 도시하시오: Mean, Standard deviation, Skewness, Kurtosis. 전체 변수 중에서 정규분포를 따른다고 할 수 있는 변수들은 몇 개인가? 정규분포를 따른다고 가정한 근거는 무엇인가?

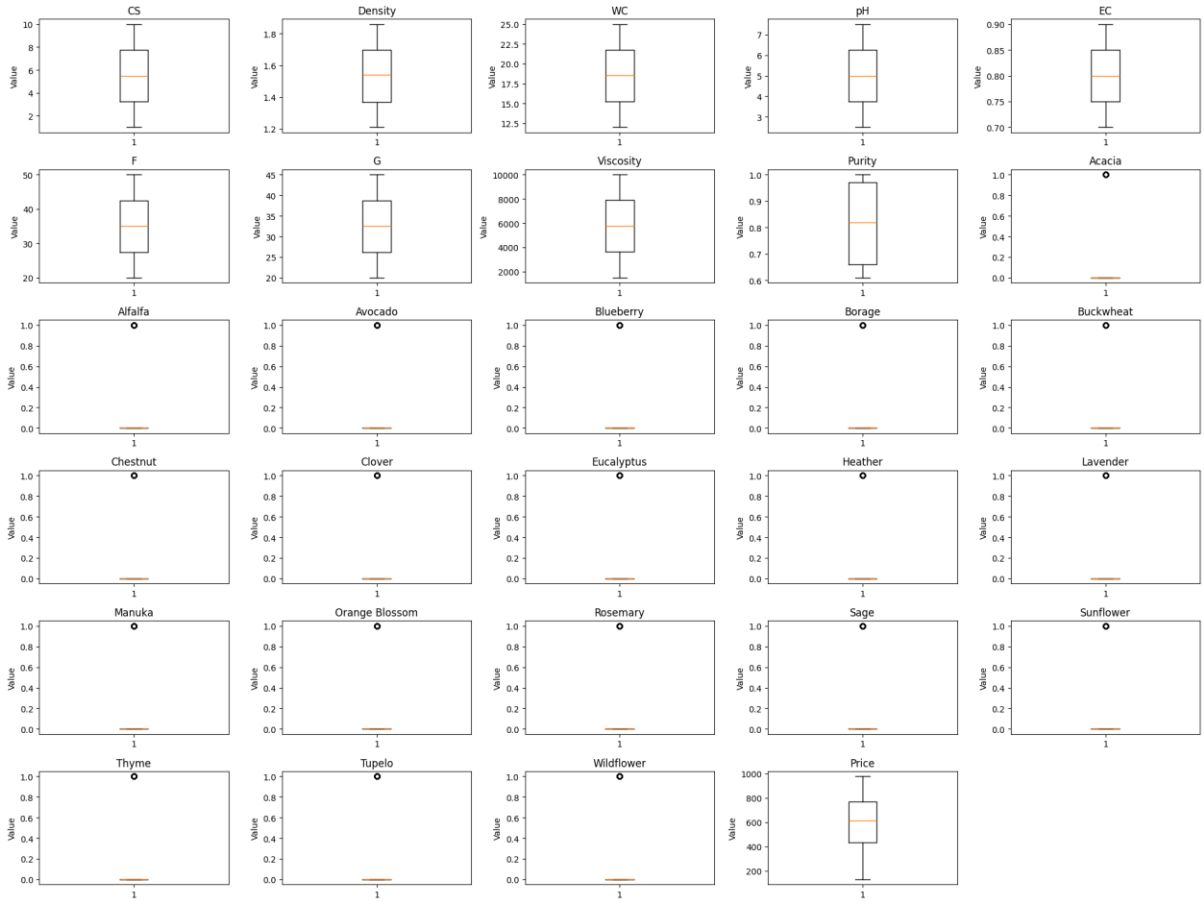
개별 입력 변수들에 대한 단변량 통계량을 계산하기 전에, 명목형 변수인 'Pollen_analysis'에 대해서 1-of-C coding 변환을 진행하였다.

	Acacia	Alfalfa	Avocado	Blueberry	Borage	Buckwheat	Chestnut	Clover	Eucalyptus	Heather	Lavender	Manuka	Orange Blossom	Rosemary	Sage	Sunflower	Thyme	Tupelo	Wildflower
0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
247898	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
247899	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
247900	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
247901	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
247902	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0

그 후 개별 입력 변수들에 대한 단변량 통계량 계산을 진행하였고 결과는 아래와 같다.

	Mean	STD	Skewness	Kurtosis
CS	5.500259	2.593947	0.000062	-1.196323
Density	1.535523	0.187824	-0.000805	-1.200521
WC	18.502625	3.748635	-0.002823	-1.199574
pH	4.996047	1.444060	0.002843	-1.201069
EC	0.799974	0.057911	0.000804	-1.187623
F	34.970573	8.655898	0.000789	-1.199137
G	32.501006	7.226290	-0.000315	-1.202248
Viscosity	5752.893888	2455.739903	-0.003851	-1.201837
Purity	0.824471	0.139417	-0.071791	-1.521310
Acacia	0.052343	0.222718	4.019946	14.159966
Alfalfa	0.052638	0.223309	4.006671	14.053416
Avocado	0.053210	0.224453	3.981149	13.849545
Blueberry	0.052867	0.223769	3.996379	13.971048
Borage	0.052440	0.222913	4.015570	14.124804
Buckwheat	0.052585	0.223204	4.009028	14.072303
Chestnut	0.052521	0.223075	4.011932	14.095601
Clover	0.052077	0.222182	4.032039	14.257339
Eucalyptus	0.053222	0.224477	3.980614	13.845285
Heather	0.053194	0.224421	3.981862	13.855227
Lavender	0.052787	0.223608	3.999983	13.999867
Manuka	0.052452	0.222937	4.015024	14.120417
Orange Blossom	0.052408	0.222848	4.017027	14.136510
Rosemary	0.052157	0.222345	4.028365	14.227727
Sage	0.052912	0.223858	3.994400	13.955235
Sunflower	0.053037	0.224107	3.988836	13.910814
Thyme	0.053069	0.224172	3.987403	13.899385
Tupelo	0.051577	0.221171	4.054998	14.443010
Wildflower	0.052504	0.223042	4.012659	14.101434
Price	594.807644	233.627972	-0.244700	-0.876182

각 변수들에 대한 box plot은 아래와 같다.



먼저, Skewness와 Kurtosis를 기준으로 정규분포 가정을 확인하려면, 두 값이 0에 가까워야한다. 그러나 이 통계만으로 해당 변수가 정규분포를 따르는지 완벽하게 판단하기는 어렵기 때문에 boxplot의 형태도 같이 확인하였다.

'절댓값 2 이하의 Skewness'와 '절댓값 3 이하의 Kurtosis'를 동시에 만족하는 변수를 찾고, boxplot의 형태도 종합적으로 고려해보았을 때, CS, Density, WC, pH, EC, F, G, Viscosity, Purity, Price 이렇게 총 10개의 변수가 정규분포를 따른다고 가정할 수 있다.

이때, 정규분포를 따른다고 가정한 변수들 모두 Kurtosis 값이 음수인 것을 확인할 수 있는데, 이는 해당 변수들이 정규분포보다 더 완만한 분포를 나타낸다고 해석할 수 있다.

▶ 정규성 검정

추가로 각 변수에 대한 정규성 검정을 위해 Shapiro 함수를 이용해 Shapiro-Wilks test를 진행하였다. Shapiro-Wilks test의 결과는 아래와 같다.

	test_statistic	p_value
CS	0.955292	3.085967e-108
Density	0.954694	1.342649e-108
WC	0.955062	2.237951e-108
pH	0.954827	1.614045e-108
EC	0.953865	4.306984e-109
F	0.955063	2.242424e-108
G	0.954663	1.286668e-108
Viscosity	0.954691	1.338300e-108
Purity	0.862613	6.415777e-141
Acacia	0.230652	1.421543e-199
Alfalfa	0.231485	1.559207e-199
Avocado	0.233100	1.865441e-199
Blueberry	0.232134	1.675706e-199
Borage	0.230926	1.465433e-199
Buckwheat	0.231337	1.533770e-199
Chestnut	0.231154	1.503021e-199
Clover	0.229896	1.307380e-199
Eucalyptus	0.233134	1.872509e-199
Heather	0.233054	1.856058e-199
Lavender	0.231907	1.633889e-199
Manuka	0.230960	1.471012e-199
Orange Blossom	0.230835	1.450658e-199
Rosemary	0.230125	1.340991e-199
Sage	0.232259	1.699151e-199
Sunflower	0.232612	1.766965e-199
Thyme	0.232703	1.784892e-199
Tupelo	0.228471	1.116678e-199
Wildflower	0.231109	1.495429e-199
Price	0.961219	2.077458e-104

결과를 확인해보면 모든 변수의 p-value 값이 유의수준(0.05)보다 작은 것을 확인할 수 있고, 따라서 모든 변수가 정규분포를 따르지 않는다고 할 수 있다.

[Q4] [Q3]의 Box plot을 근거로 각 변수들에 대한 이상치(너무 값이 크거나 작은 객체) 조건을 정의하고, 해당하는 객체들을 데이터셋에서 제거해 보시오.

명목형 변수인 'Pollen_analysis'에 대해 1-of-C coding 변환을 진행하였는데, 이들의 boxplot을 확인해보면 1의 값이 모두 이상치라고 표현된 것을 확인할 수 있다. 그 이유를 확인하기 위해

value_counts 함수를 통해 각 변수의 개수를 확인해보면 아래와 같다.

```
Pollen_analysis
Eucalyptus        13194
Avocado           13191
Heather           13187
Thyme             13156
Sunflower         13148
Sage              13117
Blueberry         13106
Lavender          13086
Alfalfa           13049
Buckwheat         13036
Chestnut          13020
Wildflower        13016
Manuka            13003
Borage            13000
Orange Blossom    12992
Acacia            12976
Rosemary          12930
Clover            12910
Tupelo            12786
Name: count, dtype: int64
```

즉, 247903개의 데이터 중 각각의 범주에 해당하는 개수가 13000개 정도로, 각 column에서 1의 값의 비율이 약 5퍼센트 정도 되기 때문에 이상치로 표시된 것을 알 수 있다. 그렇지만 이들은 1-of-C coding 변환을 통한 정상적인 데이터이므로 이상치 제거 고려대상에서 제외한다.

Q1을 1사분위수, Q3를 3사분위수라고 할 때, Q1, Q3에서 IQR(Q3-Q1)의 1.5배 떨어진 거리를 최소, 최대로 정의하고 이 밖의 값들은 이상치(outlier)라고 간주한다. 즉, boxplot에서 구간 (Q1 – 1.5 * IQR , Q3 +1.5 * IQR)에 포함되지 않는 데이터를 이상치로 정의하였고, 데이터셋에서 이상치를 제거하였다. 이상치를 제거한 결과는 아래와 같다.

```
Before removing outliers : (247903, 29)
After removing outliers : (247903, 29)
```

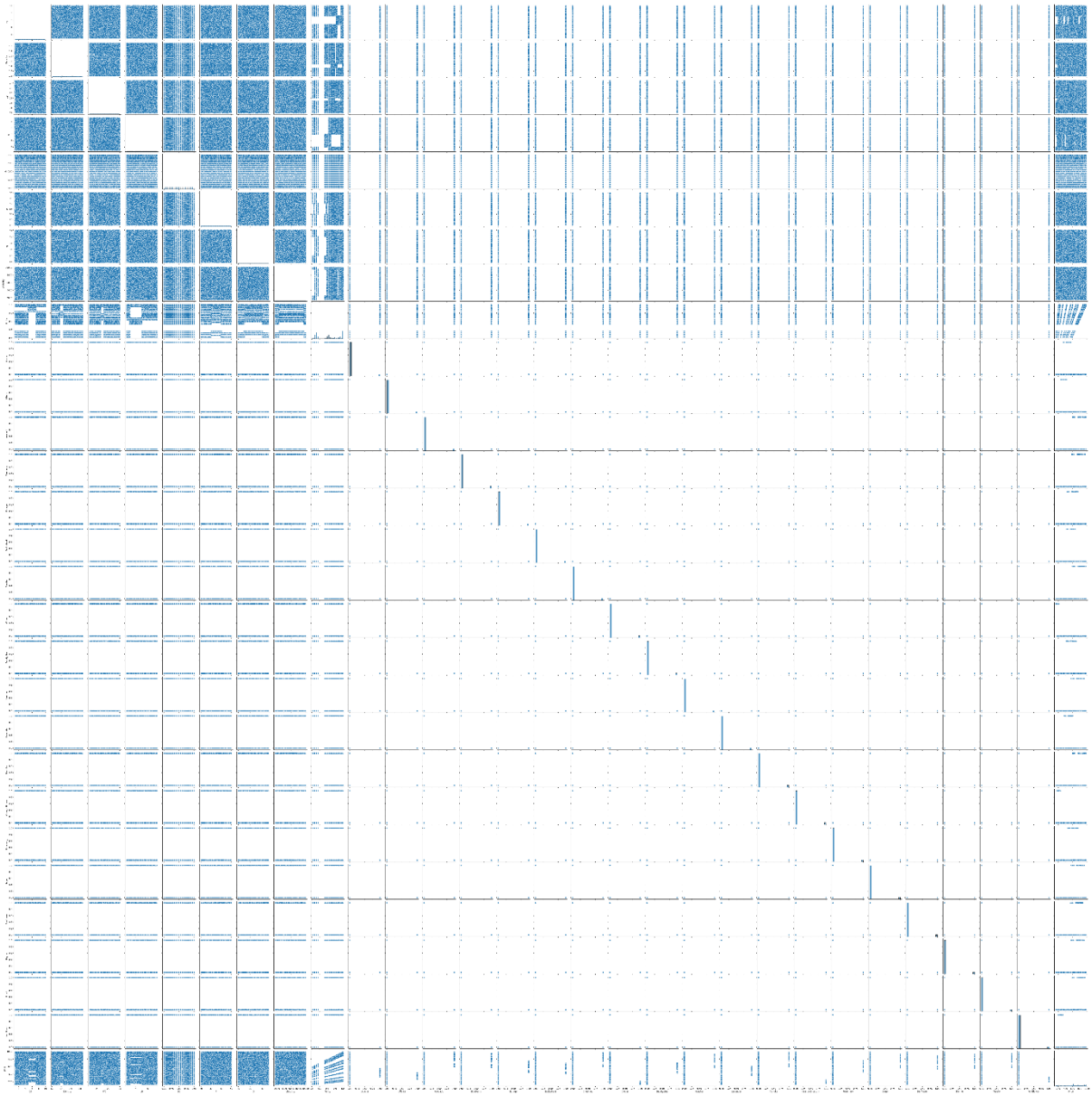
이상치 제거 전, 후의 데이터가 동일한 것을 알 수 있는데 이는 boxplot에서도 확인해볼 수 있다. boxplot을 보면, 1-of-C coding 변환을 한 'Pollen_analysis'의 column들을 제외하고, 모든 column에서 데이터들이 전부 구간 (Q1 – 1.5 * IQR , Q3 +1.5 * IQR)에 포함되어 있는 것을 확인할 수 있다.

다음 각 물음에 대해서는 [Q4]에서 제거된 객체들을 제외시킨 데이터프레임을 사용하여 답하시오.

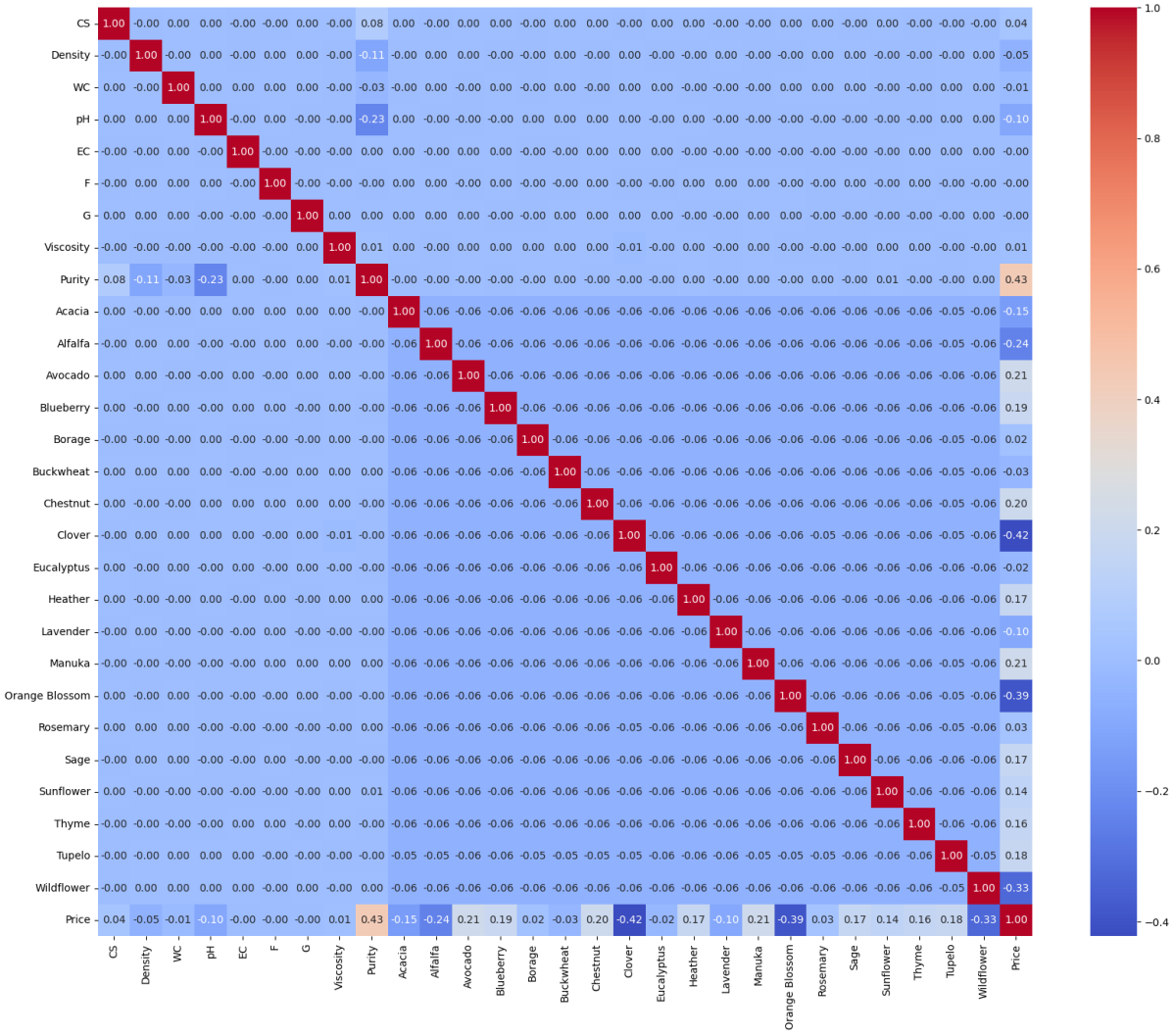
[Q5] 가능한 모든 두 쌍의 입력 변수 조합에 대한 산점도(scatter plot) 등을 도시하여 입력변수 간 상관성에 대한 분석을 수행해 보시오. 어떤 두 조합의 변수들이 서로 강한 상관관계가 있다고 할 수 있는가? 이렇게 강한 상관관계가 발생한 변수들은 상식적으로도 상관관계가 높은 변수들이라고 할 수 있는가?

변수 간 상관관계를 Scatter plot과 Heatmap으로 시각화 해보았고, 그 결과는 아래와 같다.

▶ Scatter plot



▶ Heatmap



상관계수는 -1부터 1까지의 값을 가지며, 0에서 멀리 떨어질수록 강한 선형관계를 가진다고 말할 수 있다. 이때, 상관계수가 양수인 경우 양의 상관관계, 음수인 경우 음의 상관관계를 보인다고 할 수 있다.

상관관계를 나타내는 변수들을 살펴보면, [Price, Purity]의 조합이 가장 높은 양의 상관관계를 보이는 것을 확인할 수 있는데, 사람들이 순도 높은 꿀을 선호한다는 점에서 가격과 양의 상관관계를 나타내는 것은 상식적으로 납득할 수 있다.

또한, [Price, Alfalfa], [Price, Clover], [Price, Orange Blossom], [Price, Wildflower]조합은 음의 상관관계를 보이는데 반면, [Price, Avocado], [Price, Chestnut], [Price, Manuka] 조합은 양의 상관관계를 보이는 것을 확인할 수 있다. 이들은 모두 'Pollen_analysis' column을 1-of-C coding 변환을 한 것으로 'Pollen_analysis'의 변수들 중 일부가 'Price'와 상관관계를 보인다고 할 수 있고, 사람들이 꿀을 구매할 때, 꿀의 종류에 따른 선호도가 다르기 때문에 해당 변수들이 가격과 상관관계를 보이는 것은 상식적으로 납득할 수 있다.

상관관계 분석 결과를 아래의 표로 정리하였다.

상관관계	변수
양의 상관관계가 존재한다. (상관계수 0.2 이상)	[Price, Purity] [Price, Avocado] [Price, Chestnut] [Price, Manuka]
음의 상관관계가 존재한다. (상관계수 -0.2 이하)	[Purity, pH] [Price, Alfalfa] [Price, Clover] [Price, Orange Blossom] [Price, Windflower]

특히, [Price, Purity], [Price, Clover] 조합은 다른 조합에 비해 상관관계가 높다.

[Q6] 전체 데이터셋을 70%의 학습 데이터와 30%의 테스트 데이터로 분할한 후 모든 변수를 사용하여 MLR 모델을 학습한 뒤, Adjusted R2값을 통해 데이터의 선형성(linearity)을 판단해 보시오. Residual plot과 Q-Q Plot을 도시하고 Ordinary Least Square 방식의 Solution이 만족해야 하는 가정들이 만족될만한 수준인지 정성적으로 판단해 보시오.

Categorical 변수의 원핫인코딩을 진행하고, sklearn의 train_test_split을 통해 70%의 학습 데이터와 30%의 테스트 데이터로 분할한 후, MLR 모델을 학습하였다.

```
from sklearn.model_selection import train_test_split
seed = 12345
test_size = 0.3
honey_trn_data, honey_valid_data = train_test_split(new_honey_data,test_size=test_size,random_state=seed)
```

```
import statsmodels.api as sm

feature_names = list(honey_input.columns)
x_with_const = sm.add_constant(x_trn)

# Fit the linear regression model
model = sm.OLS(y_trn, x_with_const)
results = model.fit()

summary_table = results.summary()
# Replace generic labels with variable names in the summary table
summary_table = results.summary(xname=['const'] + feature_names)

print(summary_table)
```

아래는 MLR 모델의 결과이다.

OLS Regression Results

Dep. Variable:	y	R-squared:	0.977
Model:	OLS	Adj. R-squared:	0.977
Method:	Least Squares	F-statistic:	2.742e+05
Date:	Sun, 31 Mar 2024	Prob (F-statistic):	0.00
Time:	15:41:17	Log-Likelihood:	-8.6481e+05
No. Observations:	173532	AIC:	1.730e+06
Df Residuals:	173504	BIC:	1.730e+06
Df Model:	27		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-4.2583	1.618	-2.631	0.009	-7.430	-1.086
CS	0.0082	0.033	0.251	0.802	-0.056	0.073
Density	-0.0304	0.454	-0.067	0.947	-0.920	0.859
WC	-0.0486	0.023	-2.147	0.032	-0.093	-0.004
pH	0.0827	0.060	1.369	0.171	-0.036	0.201
EC	2.7755	1.465	1.894	0.058	-0.097	5.648
F	-0.0053	0.010	-0.544	0.587	-0.025	0.014
G	0.0050	0.012	0.429	0.668	-0.018	0.028
Viscosity	2.601e-05	3.45e-05	0.753	0.451	-4.17e-05	9.37e-05
Purity	723.9310	0.632	1145.995	0.000	722.693	725.169
Acacia	-146.2727	0.370	-395.671	0.000	-146.997	-145.548
Alfalfa	-237.9838	0.371	-641.864	0.000	-238.710	-237.257
Avocado	210.3363	0.368	572.072	0.000	209.616	211.057
Blueberry	185.3064	0.367	504.392	0.000	184.586	186.026
Borage	19.4084	0.370	52.499	0.000	18.684	20.133
Buckwheat	-30.3981	0.369	-82.353	0.000	-31.122	-29.675
Chestnut	201.5615	0.370	544.154	0.000	200.836	202.288
Clover	-419.6284	0.371	-1131.747	0.000	-420.355	-418.902
Eucalyptus	-22.1285	0.367	-60.249	0.000	-22.848	-21.409
Heather	169.0417	0.368	459.048	0.000	168.320	169.763
Lavender	-96.7261	0.369	-261.848	0.000	-97.450	-96.002
Manuka	210.5395	0.371	567.181	0.000	209.812	211.267
Orange Blossom	-387.6060	0.371	-1044.407	0.000	-388.333	-386.879
Rosemary	27.7007	0.372	74.440	0.000	26.971	28.430
Sage	168.9473	0.369	457.799	0.000	168.224	169.671
Sunflower	136.0120	0.369	368.979	0.000	135.290	136.735
Thyme	160.4800	0.368	436.412	0.000	159.759	161.201
Tupelo	176.9617	0.374	473.513	0.000	176.229	177.694
Wildflower	-329.8104	0.372	-886.507	0.000	-330.540	-329.081

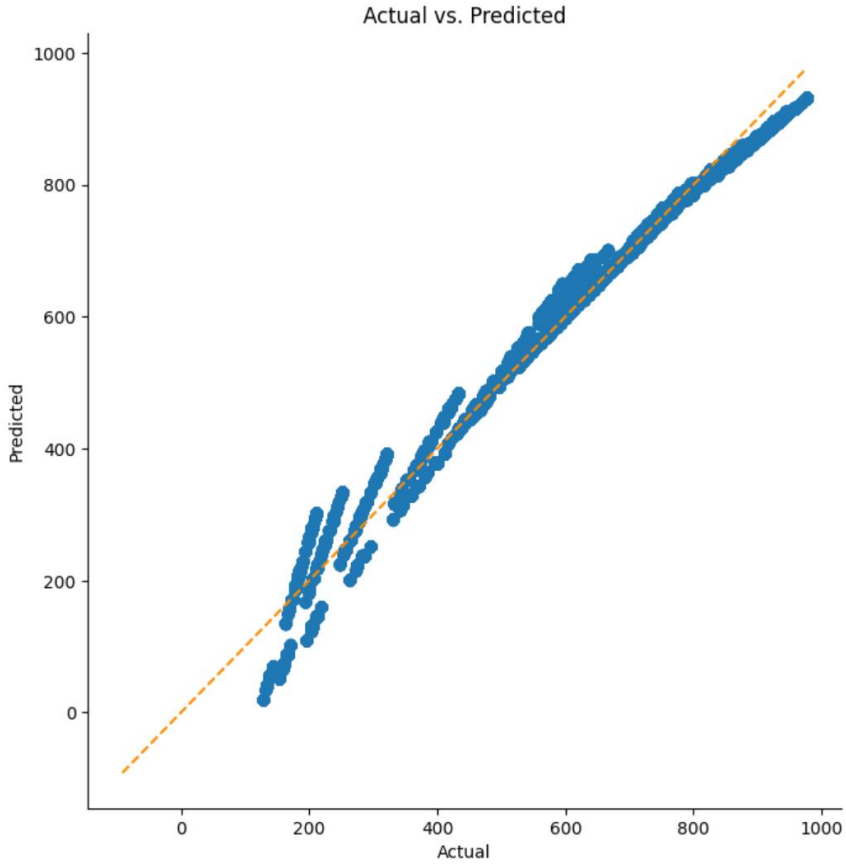
Omnibus:	608.564	Durbin-Watson:	2.006
Prob(Omnibus):	0.000	Jarque-Bera (JB):	783.772
Skew:	0.051	Prob(JB):	6.40e-171
Kurtosis:	3.313	Cond. No.	1.13e+19

MLR 모델 학습 결과, Adjusted R-squared 값이 0.977로, 종속변수 Price에 대한 모델의 설명력이 97.7%인 것을 확인할 수 있다. Adjusted R-squared 값이 1에 가까울수록 종속변수와 설명변수 간에 강한 선형 관계가 있음을 뜻하므로, 데이터의 입력변수들과 출력변수 사이에 강한 선형 관계가 있음을 파악할 수 있다.

이때, 회귀 모형 즉, Ordinary Least Square 방식의 solution은 다음의 가정을 만족하여야 한다.

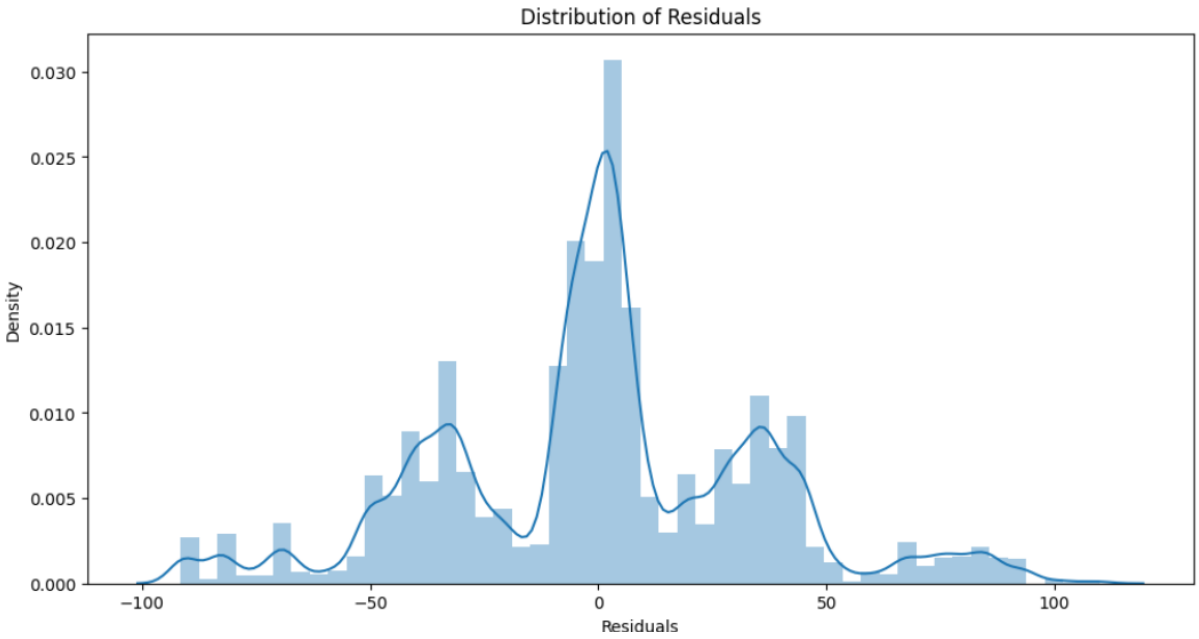
1. 설명변수와 종속변수 사이에 선형관계가 성립한다.
2. 오차항들이 정규분포를 따른다.
3. 각 관측치들은 서로 독립이다.
4. 종속변수 Y에 대한 오차항은 설명변수 값의 범위에 관계없이 일정하다. (homoskedasticity)

① 설명변수와 종속변수 사이의 선형관계



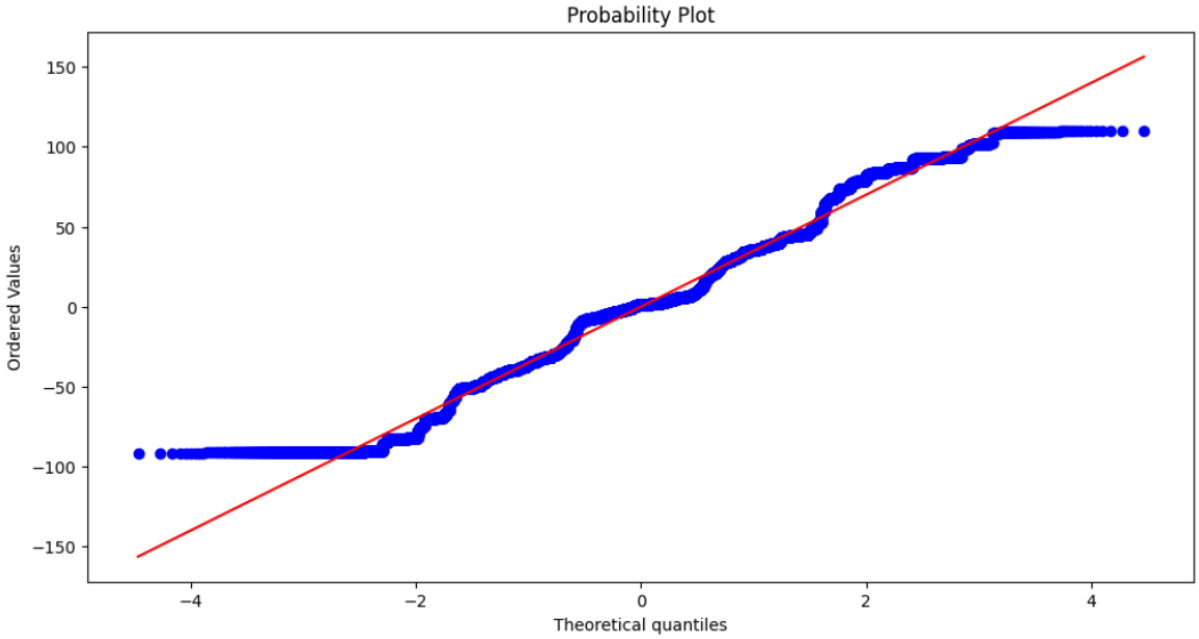
해당 plot을 통해 설명변수와 종속변수간 선형관계를 보인다고 해석할 수 있다.

② Residual Distribution



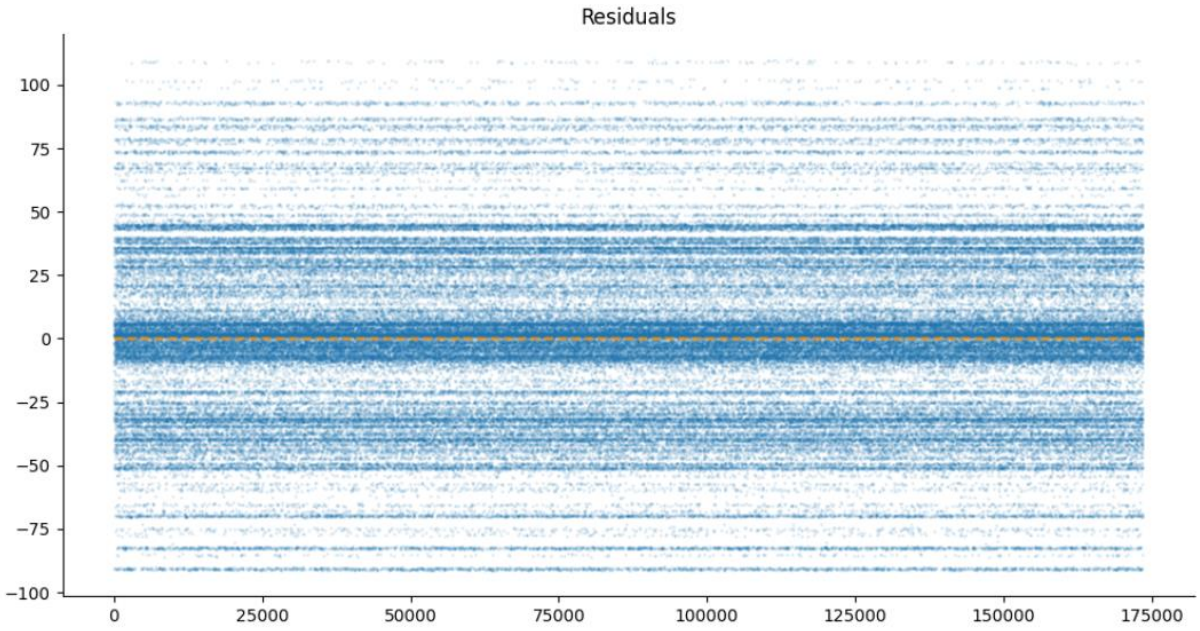
Residual Distribution을 나타낸 그래프를 보았을 때, 오차항이 정규분포를 띄지 않는 것을 확인할 수 있다.

③ QQ plot



QQ plot을 보았을 때, 양 끝이 직선에서 크게 벗어나는 모습을 관찰할 수 있고, 따라서 잔차들이 정규성을 띄지 않는다고 볼 수 있다.

④ Residual Plot



Residual plot의 결과, 잔차들이 무작위하게 분포해있다고 볼 수 있고 homoscedasticity에 위배되지 않는다고 할 수 있다.

결론적으로 Residual Distribution, QQ plot 해석 결과, 잔차들이 정규성을 띈다고 보기 어렵고, 따라서 잔차가 정규분포를 따라야한다는 가정은 만족하지 않는다. 즉, Ordinary Least Square 방식의 Solution이 만족해야 하는 가정이 만족되지 않는다고 할 수 있다.

[Q7] 유의수준 0.01에서 모형 구축에 통계적으로 유의미한 변수들은 어떤 것들이 있는가? 해당 변수들은 종속변수와 양/음 중에서 어떤 상관관계를 갖고 있는가?

	coef	std err	t	P>|t|	[0.025	0.975]
const	-4.2583	1.618	-2.631	0.009	-7.430	-1.086
CS	0.0082	0.033	0.251	0.802	-0.056	0.073
Density	-0.0304	0.454	-0.067	0.947	-0.920	0.859
WC	-0.0486	0.023	-2.147	0.032	-0.093	-0.004
pH	0.0827	0.060	1.369	0.171	-0.036	0.201
EC	2.7755	1.465	1.894	0.058	-0.097	5.648
F	-0.0053	0.010	-0.544	0.587	-0.025	0.014
G	0.0050	0.012	0.429	0.668	-0.018	0.028
Viscosity	2.601e-05	3.45e-05	0.753	0.451	-4.17e-05	9.37e-05
Purity	723.9310	0.632	1145.995	0.000	722.693	725.169
Acacia	-146.2727	0.370	-395.671	0.000	-146.997	-145.548
Alfalfa	-237.9838	0.371	-641.864	0.000	-238.710	-237.257
Avocado	210.3363	0.368	572.072	0.000	209.616	211.057
Blueberry	185.3064	0.367	504.392	0.000	184.586	186.026
Borage	19.4084	0.370	52.499	0.000	18.684	20.133
Buckwheat	-30.3981	0.369	-82.353	0.000	-31.122	-29.675
Chestnut	201.5615	0.370	544.154	0.000	200.836	202.288
Clover	-419.6284	0.371	-1131.747	0.000	-420.355	-418.902
Eucalyptus	-22.1285	0.367	-60.249	0.000	-22.848	-21.409
Heather	169.0417	0.368	459.048	0.000	168.320	169.763
Lavender	-96.7261	0.369	-261.848	0.000	-97.450	-96.002
Manuka	210.5395	0.371	567.181	0.000	209.812	211.267
Orange Blossom	-387.6060	0.371	-1044.407	0.000	-388.333	-386.879
Rosemary	27.7007	0.372	74.440	0.000	26.971	28.430
Sage	168.9473	0.369	457.799	0.000	168.224	169.671
Sunflower	136.0120	0.369	368.979	0.000	135.290	136.735
Thyme	160.4800	0.368	436.412	0.000	159.759	161.201
Tupelo	176.9617	0.374	473.513	0.000	176.229	177.694
Wildflower	-329.8104	0.372	-886.507	0.000	-330.540	-329.081

유의수준 0.01에서 MLR 모델 구축에 통계적으로 유의미한 변수들을 찾기 위해, 위의 summary에서 p-value 값을 참조하였다.

다음은 p-value 값이 0.01 이내인 변수들이며, 통계적으로 유의미한 변수라고 할 수 있다.

▶ Purity, Acacia, Alfalfa, Avocado, Blueberry, Borage, Buckwheat, Chestnut, Clover, Eucalyptus, Heather, Lavender, Manuka, Orange Blossom, Rosemary, Sage, Sunflower, Thyme, Tupelo, Wildflower

이때, 'Purity'를 제외한 모든 변수는 'Pollen_analysis'를 원핫인코딩한 변수들이다.

▶ 즉, 'Purity'와 'Pollen_analysis'가 통계적으로 유의미한 변수임을 확인할 수 있다.

유의미한 변수들 중, coefficient가 양수인 것은 종속변수와 양의 상관관계를 가진다고 볼 수 있고, 음수인 것은 종속변수와 음의 상관관계를 가진다고 할 수 있다. 이를 정리하면 아래와 같다.

- 양의 상관관계

  Purity, Avocado, Blueberry, Borage, Chestnut, Heather, Manuka, Rosemary, Sage, Sunflower, Thyme, Tupelo

- 음의 상관관계

  Acacia, Alfalfa, Buckwheat, Clover, Eucalyptus, Lavender, Orange Blossom, Wildflower

[Q8] Test 데이터셋에 대하여 MAE, MAPE, RMSE를 계산하고 그에 대한 해석을 해 보시오.

Test 데이터셋에 대한 성능을 측정한 결과는 아래와 같다.

	RMSE	MAE	MAPE
Honey	35.32162	26.26264	0.07341

각 평가지표는 유효숫자 5자리로 통일하여 출력하였다.

- RMSE는 실제 값과 예측 값 사이 제곱 오차의 평균의 제곱근으로, 해당 값은 35.32162 이다.

- MAE는 실제 값과 예측 값 사이의 절대적인 오차의 평균으로, 해당 값은 26.26264 이다.

- MAPE는 실제 값 대비 예측 값이 얼마나 차이가 있는지를 비율로 측정한 값으로, 해당 값은 0.07341 이다. 즉, MAPE 값에 의하면 MLR 모델이 약 7.341%의 오차가 있음을 알 수 있다.

전체적인 성능 수치를 고려해 보았을 때, 특히 MAPE 값을 보았을 때, MLR 모델이 약 7.3%의 오차를 보이고 있다는 점에서, MLR 모델이 정상적으로 잘 적용된 것으로 해석할 수 있다. 또한, 해당 데이터셋에 Multiple Linear Regression을 적용하는 것이 적합하다고 해석할 수 있다.

[Q9] 만약 원래 변수 수의 절반 이하로 입력 변수를 사용하여 모델을 구축해야 할 경우 어떤 변수들을 선택하겠는가? [Q5]와 [Q7]의 답변을 바탕으로 본인이 선택한 변수들에 대한 근거를 제시하시오.

Q5의 답변에 의하면 이 데이터셋에서 다른 조합들에 비해 강한 상관관계가 있는 변수들의 조합은 [Price, Purity], [Price, Clover] 이었다. 이때, Clover 뿐 아니라 'Pollen_analysis'의 여러 다른 변수들도 Price와 상관관계를 보이는 경우가 많았다.

Q7의 답변에 의하면 통계적으로 유의미한 변수는 Purity와 원핫인코딩을 진행한 Pollen_analysis의 변수들이었다. 이때, WC 변수의 p-value는 0.032로 유의수준(0.01)보다 높았지만, 유의수준을 0.05로 한다면 통계적으로 유의미한 변수라고 할 수 있는 여지가 있다.

추가로 전체 데이터 셋을 이용하여 VIF를 체크해 보았을 때, WC 변수의 VIF 값이 1.00101276로, Multicollinearity도 없다고 할 수 있다.

따라서, 위의 내용들을 바탕으로 최종적으로 선택한 변수의 목록은 다음과 같다.

['Purity', 'WC', 'Pollen_analysis']

이때, 'Pollen_analysis'는 1-of-C coding 변환을 진행한다.

[Q10] [Q9]에서 선택한 변수들만을 사용하여 MLR 모형을 다시 학습하고 Adjusted R2, Test 데이터셋에 대한 MAE, MAPE, RMSE를 산출한 뒤, 두 모형(모든 변수 사용 vs. 선택된 변수만 사용)을 비교해 보시오.

3개의 설명변수들['Purity', 'WC', 'Pollen_analysis']을 선택하여 honey_data_2에 저장한 후 학습을 진행하였다.

```
from sklearn.model_selection import train_test_split
seed = 12345
test_size = 0.3
honey_trn_data_2, honey_valid_data_2 = train_test_split(honey_data_2,test_size=test_size,random_state=seed)
```

전과 동일하게 전체 데이터셋의 70%를 학습 데이터로, 30%를 테스트 데이터로 분할하여 MLR 모델 학습을 진행하였고, 그 결과는 아래와 같다.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.977
Model:                            OLS   Adj. R-squared:                  0.977
Method:                 Least Squares   F-statistic:                 3.702e+05
Date:                Tue, 02 Apr 2024   Prob (F-statistic):               0.00
Time:                        01:59:28   Log-Likelihood:            -8.6482e+05
No. Observations:              173532   AIC:                         1.730e+06
Df Residuals:                  173511   BIC:                         1.730e+06
Df Model:                          20
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          -1.4987      0.635     -2.362      0.018      -2.742      -0.255
Purity        723.7569      0.609   1188.927      0.000     722.564     724.950
WC             -0.0487      0.023     -2.154      0.031      -0.093      -0.004
Acacia       -146.1237      0.361   -404.607      0.000    -146.832    -145.416
Alfalfa      -237.8415      0.362   -656.302      0.000    -238.552    -237.131
Avocado       210.4814      0.359    586.025      0.000     209.777     211.185
Blueberry     185.4530      0.359    516.594      0.000     184.749     186.157
Borage         19.5538      0.361     54.104      0.000      18.845      20.262
Buckwheat     -30.2525      0.361    -83.881      0.000     -30.959     -29.546
Chestnut      201.7099      0.362    557.513      0.000     201.001     202.419
Clover       -419.4801      0.362  -1158.572      0.000    -420.190    -418.770
Eucalyptus    -21.9826      0.359    -61.255      0.000     -22.686     -21.279
Heather       169.1838      0.360    469.830      0.000     168.478     169.890
Lavender      -96.5875      0.361   -267.396      0.000     -97.295     -95.880
Manuka        210.6815      0.363    580.233      0.000     209.970     211.393
Orange Blossom -387.4601    0.363  -1068.264      0.000    -388.171    -386.749
Rosemary       27.8427      0.364     76.513      0.000      27.129      28.556
Sage          169.0903      0.361    468.587      0.000     168.383     169.798
Sunflower     136.1558      0.360    378.027      0.000     135.450     136.862
Thyme         160.6245      0.359    446.873      0.000     159.920     161.329
Tupelo        177.1130      0.365    485.068      0.000     176.397     177.829
Wildflower   -329.6604      0.363   -907.610      0.000    -330.372    -328.948
==============================================================================
Omnibus:                      614.923   Durbin-Watson:                   2.006
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              791.536
Skew:                           0.052   Prob(JB):                     1.32e-172
Kurtosis:                       3.314   Cond. No.                      1.93e+16
==============================================================================
```

새로운 MLR 모델의 Adjusted R-squared 값은 0.977로 설명변수들이 종속변수에 대해 97.7%의 설명력을 가짐을 알 수 있다. 새로운 모델의 Adjusted R-squared 값이 기존 모델의 값과 차이가 없음을 알 수 있고, 아래의 오른쪽 표는 새로운 모델의 RMSE, MAE, MAPE를 계산한 결과이다.

	RMSE	MAE	MAPE
Honey	35.32162	26.26264	0.07341

	RMSE	MAE	MAPE
Honey	35.32195	26.26040	0.07340

RMSE, MAE, MAPE 값 모두 기존 모델 대비 큰 차이가 없는 것을 확인할 수 있다. 이를 통해 선택되지 않은 설명변수들은 모델의 성능에 거의 영향을 끼치지 않는다고 해석할 수 있다.

즉, ['Purity', 'WC', 'Pollen_analysis'] 이렇게 3개의 설명변수만으로도 종속변수에 대한 설명력을 가진다는 것이고, 변수 재선정 때 선택되지 않은 변수들은 전체적으로 봤을 때, 통계적으로 유의미하지 않고, 선택하지 않아도 크게 문제되지 않는다는 결론을 내릴 수 있다.

[Extra Question] 이 외 해당 데이터셋을 통해 MLR 관점에서 가능한 추가적인 분석을 웹에서 검색해서 수행하고 그 결과를 해석해 보시오.

선형회귀의 방법 중 Lasso Regression과 Ridge Regression은 규제화 기법을 통해 모델의 복잡성을 줄이는데 유용하다. 즉, 손실함수에 penalty항을 추가함으로써, 과최적화를 방지할 수 있다.

Lasso 회귀와 Ridge 회귀의 차이점은 다음과 같다.

- Lasso 회귀는 L1 규제화를 사용하는데 이는 회귀 계수의 절댓값을 합한 것을 기반으로 한다. 즉, 규제로 인하여 중요하지 않은 일부 파라메터의 계수를 0으로 만들 수 있고, 이를 통해 '변수 선택'이 가능하게 해준다.

- Ridge 회귀는 L2 규제화를 사용하는데 이는 회귀 계수의 제곱을 합한 것을 기반으로 한다. 즉, 규제로 인하여 중요하지 않은 파라메터의 계수를 0에 가깝게 만들 수 있지만 0이 되지는 않는다.

현재 데이터셋의 전체 변수의 개수는 11개이지만, 지금까지의 분석에 의하면, 이들 중 통계적으로 유의미하지 않은 변수들이 있는 것을 확인할 수 있다. 또한, Categorical 변수('Pollen_analysis')에 대해 1-of-C coding 변환을 진행하면 19개 column이 새롭게 만들어는데, 전체 데이터셋을 이용하여 이들의 VIF 값을 확인해보면 아래와 같이 모두 20이 넘어가는 것을 확인할 수 있다.

```
Acacia: 22.198490598618715
Alfalfa: 22.06653729954508
Avocado: 22.441478057414223
Blueberry: 22.478495520757345
Borage: 22.161086373392976
Buckwheat: 22.300125982580575
Chestnut: 22.123103276973065
Clover: 22.115057995069773
Eucalyptus: 22.494182126238673
Heather: 22.2989532319776
Lavender: 22.15429784517692
Manuka: 21.941763466463996
Orange Blossom: 22.032638327433283
Rosemary: 21.863202888513555
Sage: 22.228132801385673
Sunflower: 22.344144196580537
Thyme: 22.405054477869196
Tupelo: 21.7564588443912
Wildflower: 22.01257752788145
```

따라서 특정 파라메터들의 계수를 0으로 만들 수 있어 해당 특성을 모델 구축에서 배제할 가능성이 있는 Lasso 회귀 분석이 현재 데이터셋에 적합할 것으로 판단하였다.

Lasso 회귀 분석에는 alpha라는 규제의 강도를 결정하는 하이퍼파라메터가 있는데, 이 값을 조절함으로써 모델의 복잡도를 조절할 수 있다. 이 값이 너무 크면 underfitting, 너무 작으면 overfitting의 문제가 발생한다.

alpha의 값을 0.1로 하여 Lasso 회귀분석을 진행한 후, 각 변수의 계수들을 확인하면 아래와 같다.

```
from sklearn.linear_model import Lasso

feature_names = list(new_honey_data.columns)[:-1]

# alpha는 규제의 강도
lasso = Lasso(alpha=0.1)
lasso.fit(x_trn, y_trn)

# Create a DataFrame to display the coefficients
coeff_df = pd.DataFrame(lasso.coef_, feature_names, columns=['Coefficient'])
coeff_df.loc['Intercept'] = lasso.intercept_

print(coeff_df)
```

	Coefficient
CS	0.017057
Density	-0.000000
WC	-0.048470
pH	-0.000000
EC	0.000000
F	-0.004445
G	0.001433
Viscosity	0.000031
Purity	718.593166
Acacia	-172.071599
Alfalfa	-263.762061
Avocado	180.767287
Blueberry	155.739678
Borage	-6.380471
Buckwheat	-56.199339
Chestnut	171.958438
Clover	-445.413766
Eucalyptus	-47.949711
Heather	139.467331
Lavender	-122.525398
Manuka	180.931633
Orange Blossom	-413.377185
Rosemary	0.000000
Sage	139.364567
Sunflower	106.447564
Thyme	130.906709
Tupelo	147.331341
Wildflower	-355.567373
Intercept	30.423389

같은 데이터로 학습한 기존 모델의 계수와 비교해보면, 'Density', 'pH', 'EC', 'Rosemary' 변수의 계수가 원래는 0이 아니었지만, Lasso 회귀 분석 결과에서는 0이 된 것을 확인할 수 있고, L1 규제화로 인한 변수선택이 잘 일어난 것을 확인할 수 있다.

	coef
const	-4.2583
CS	0.0082
Density	-0.0304
WC	-0.0486
pH	0.0827
EC	2.7755
F	-0.0053
G	0.0050
Viscosity	2.601e-05
Purity	723.9310
Acacia	-146.2727
Alfalfa	-237.9838
Avocado	210.3363
Blueberry	185.3064
Borage	19.4084
Buckwheat	-30.3981
Chestnut	201.5615
Clover	-419.6284
Eucalyptus	-22.1285
Heather	169.0417
Lavender	-96.7261
Manuka	210.5395
Orange Blossom	-387.6060
Rosemary	27.7007
Sage	168.9473
Sunflower	136.0120
Thyme	160.4800
Tupelo	176.9617
Wildflower	-329.8104

	Coefficient
CS	0.017057
Density	-0.000000
WC	-0.048470
pH	-0.000000
EC	0.000000
F	-0.004445
G	0.001433
Viscosity	0.000031
Purity	718.593166
Acacia	-172.071599
Alfalfa	-263.762061
Avocado	180.767287
Blueberry	155.739678
Borage	-6.380471
Buckwheat	-56.199339
Chestnut	171.958438
Clover	-445.413766
Eucalyptus	-47.949711
Heather	139.467331
Lavender	-122.525398
Manuka	180.931633
Orange Blossom	-413.377185
Rosemary	0.000000
Sage	139.364567
Sunflower	106.447564
Thyme	130.906709
Tupelo	147.331341
Wildflower	-355.567373
Intercept	30.423389

아래의 오른쪽 표는 새로운 모델의 RMSE, MAE, MAPE를 계산한 결과이다.

	RMSE	MAE	MAPE
Honey	35.32162	26.26264	0.07341

	RMSE	MAE	MAPE
Honey	35.38467	26.46893	0.07264

평가지표에서는 기존 모델 대비 큰 차이가 없는 것을 확인할 수 있고, 따라서 L1 규제화로 인한 변수 선택이 적절하게 되었다고 판단할 수 있다.