중간 발표

사이버 물리 시스템
박수창 교수님

2016038005 김홍길

CONTENTS.

01_문제 정의

02_EDA

01

문제 정의

전복 나이 예측 경진대회

01. 문제

주어진 자료를 받아, 전복의 나이를 예측하기.

주어지는 자료 (train 2)

	id	Gender	Lenght	Diameter	Height	Whole We	Shucked V	Viscra Wei	Shell Weig	Target
1154	1155	M	0.57	0.48	0.18	0.9395	0.399	0.2	0.295	14
830	831	M	0.56	0.425	0.135	0.9415	0.509	0.2015	0.1975	9
1162	1163	F	0.655	0.51	0.15	1.043	0.4795	0.223	0.305	9
599	600	I	0.31	0.225	0.05	0.1445	0.0675	0.0385	0.045	6
1058	1059	M	0.31	0.225	0.075	0.1295	0.0455	0.0335	0.044	9
220	221	F	0.565	0.4	0.13	0.6975	0.3075	0.1665	0.18	8
923	924	I	0.525	0.4	0.145	0.6095	0.248	0.159	0.175	9
790	791	M	0.655	0.515	0.2	1.373	0.443	0.3375	0.49	16
485	486	F	0.645	0.51	0.18	1.6195	0.7815	0.322	0.4675	12

02. input, output

id, gender, Lenght, Diameter, Height, Whole Weight, Shucked Weight, Viscra Weight, Target

전복 나이 (Target)

01_문제 정의

02_EDA

03_디자인

04_평가 및 결론

02

EDA

EDA

01 데이터 column(feature) 분석

02 high level perspective

03 VC

데이터 타입

```
In [ ]: df.dtypes
```

```
Out[ ]: NaN                int64
        id                 int64
        Gender            object
        Lenght           float64
        Diameter         float64
        Height           float64
        Whole Weight     float64
        Shucked Weight   float64
        Viscra Weight    float64
        Shell Weight     float64
        Target             int64
```

iint, float, object 값 들어있음
target 은 int 값.

데이터 개수

```
In [ ]: df.shape
```

```
Out[ ]: (1127, 11)
```

데이터 개수는, 1127개임을 확인.

EDA

01 데이터 column(feature) 분석

02 상관관계 분석

결측치.

값이 표기되지 않은 값. 학습과정에서 좋지 못한 영향

```
# 결측치 분석 결측치 없음으로 확인.
df.isnull().sum()
```

```
NaN               0
Gender            0
Lenght            0
Diameter          0
Height            0
Whole Weight      0
Shucked Weight    0
Viscra Weight     0
Shell Weight      0
Target            0
dtype: int64
```

EDA

01 데이터 column(feature) 분석

02 상관관계 분석

데이터 통계

```
df.describe()
```

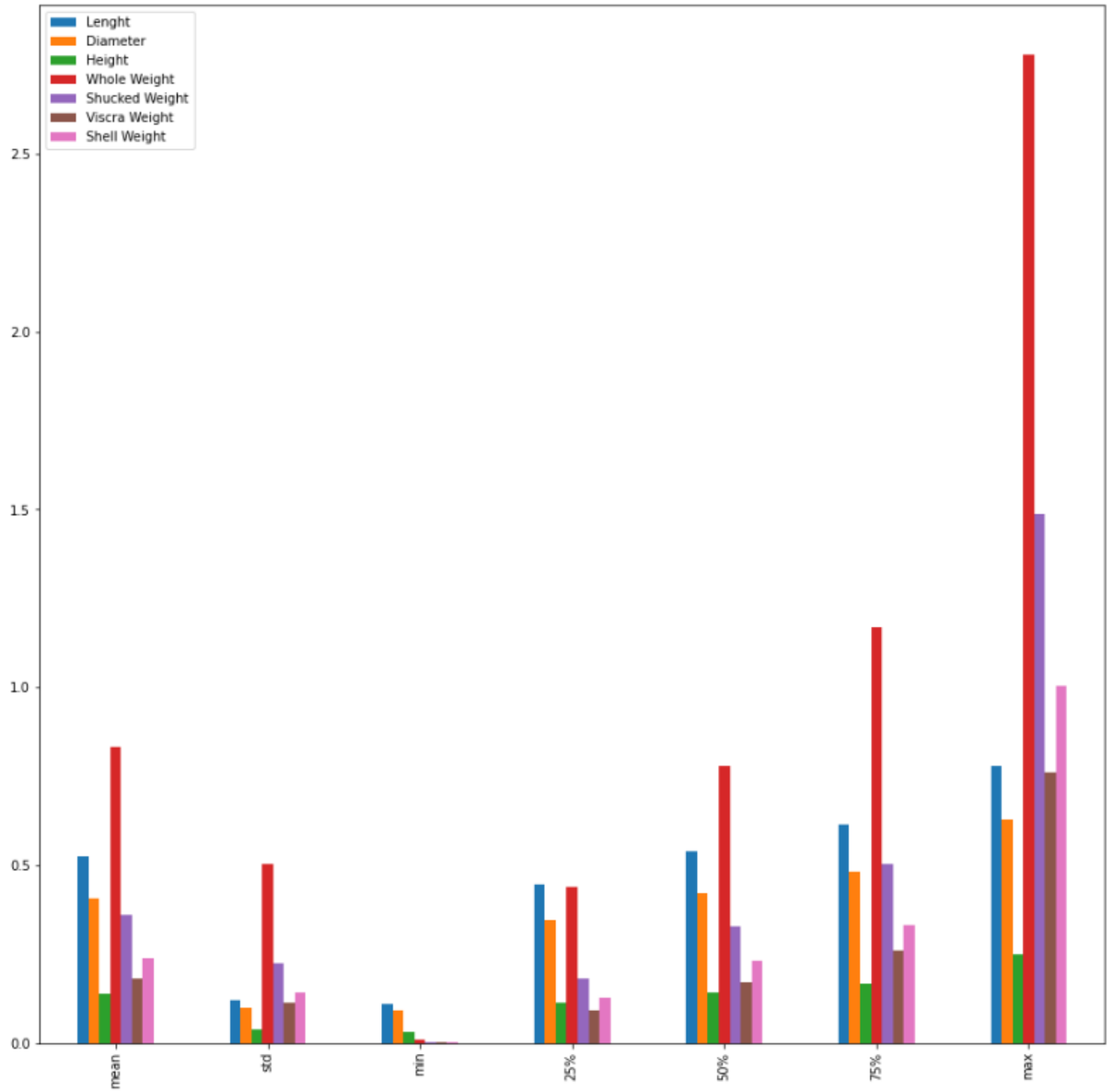
	NaN	Lenght	Diameter	Height	Whole Weight	Shucked Weight	Viscra Weight	Shell Weight	Target
count	1127.000000	1127.000000	1127.000000	1127.000000	1127.000000	1127.000000	1127.000000	1127.000000	1127.000000
mean	625.349601	0.522924	0.407036	0.139476	0.831996	0.358900	0.181458	0.239849	9.921029
std	363.756770	0.121090	0.100372	0.039082	0.502113	0.225445	0.112655	0.142285	3.236664
min	0.000000	0.110000	0.090000	0.030000	0.008000	0.002500	0.002000	0.003000	3.000000
25%	312.500000	0.445000	0.345000	0.112500	0.440250	0.180750	0.092500	0.127500	8.000000
50%	623.000000	0.540000	0.420000	0.140000	0.777500	0.326500	0.168500	0.230500	10.000000
75%	939.500000	0.615000	0.480000	0.165000	1.167000	0.503500	0.259000	0.330000	11.000000
max	1252.000000	0.780000	0.630000	0.250000	2.779500	1.488000	0.760000	1.005000	29.000000

EDA

01 데이터 column(feature) 분석

02 상관관계 분석

데이터 통계

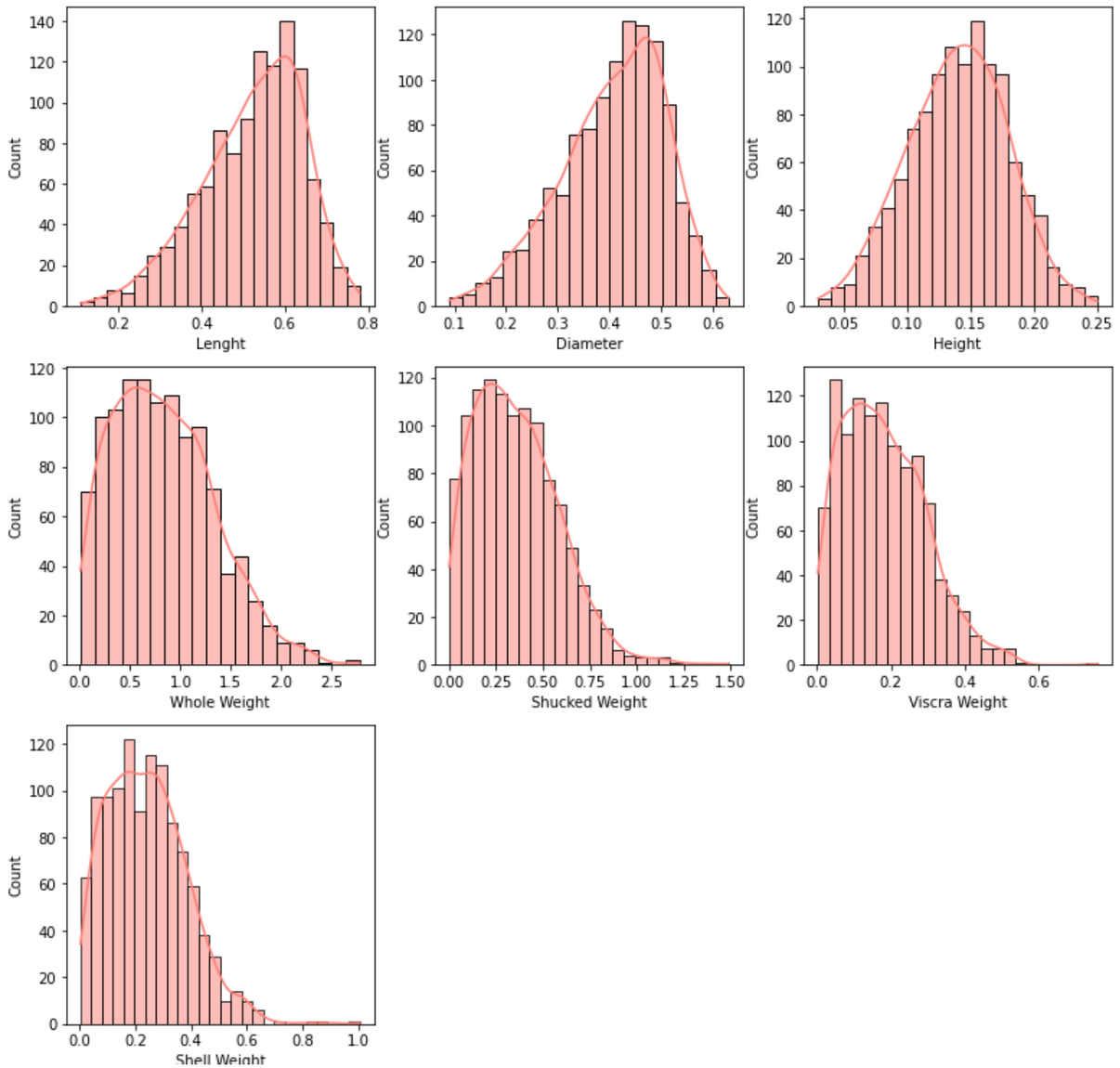


EDA

01 데이터 column(feature) 분석

02 상관관계 분석

데이터 통계 - 값의 분포



EDA

01 데이터 column(feature) 분석

02 상관관계 분석

주어진 자료들은, 성별별로 다를수가 있음

성별별 통계

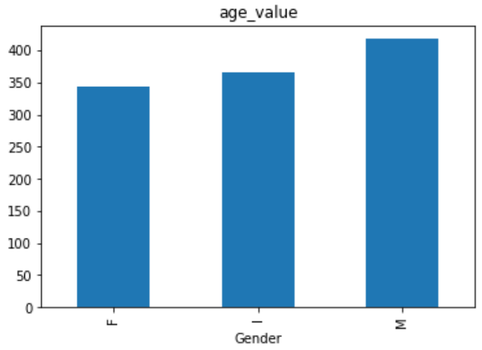
```
dfg
```

Gender	count	mean	std	min	25%	50%	75%	max
F	344.0	0.578605	0.087516	0.290	0.525	0.590	0.64000	0.780
I	365.0	0.431082	0.109195	0.110	0.360	0.435	0.52000	0.725
M	418.0	0.557297	0.108445	0.155	0.495	0.580	0.63375	0.775

성별당 개수

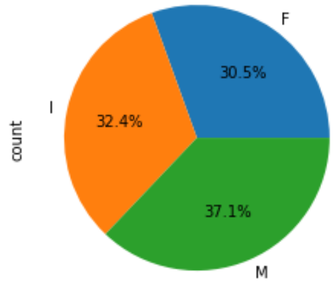
```
dfg['count'].plot.bar(title='age_value')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f120ff59710>
```




```
dfg['count'].plot.pie(autopct='%.1f%%')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f120e42b
```



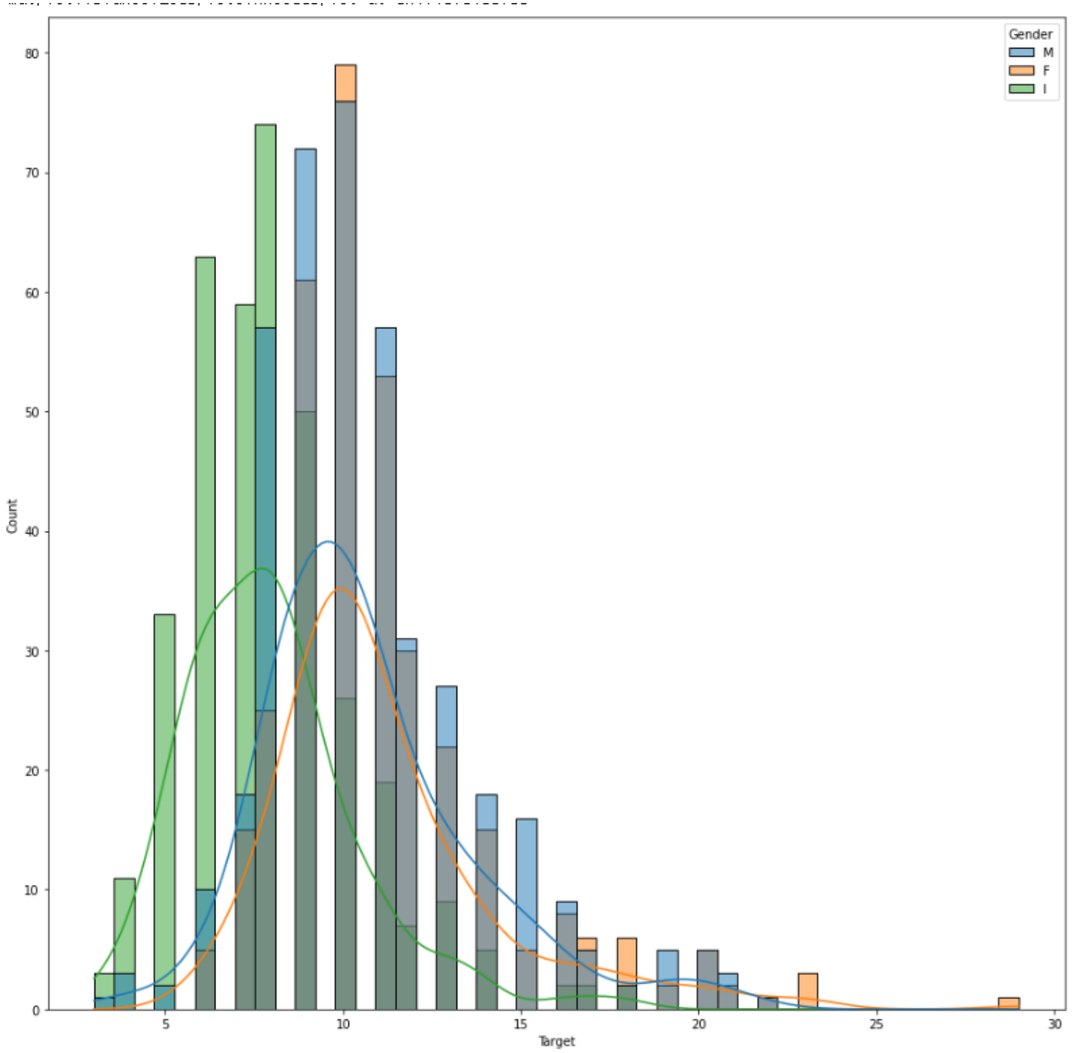
EDA

01 데이터 column(feature) 분석

02 상관관계 분석

성별당 나이

Target인 나이는 중요하기에, 성별당 나이를 따로 표시 -> I가 나이가 대체로 적음



EDA

01 구현의 배경 및 관련 작업

02 상관관계 분석

상관관계란?

데이터들간의 관계를 표시하는 수치

```
df_co
```

	Lenght	Diameter	Height	Whole Weight	Shucked Weight	Viscra Weight	Shell Weight	Target
Lenght	1.000000	0.987336	0.905323	0.926032	0.895539	0.902617	0.897628	0.544288
Diameter	0.987336	1.000000	0.912596	0.928018	0.891400	0.901182	0.907122	0.569380
Height	0.905323	0.912596	1.000000	0.897565	0.840838	0.868035	0.902457	0.606440
Whole Weight	0.926032	0.928018	0.897565	1.000000	0.967998	0.966080	0.952435	0.536748
Shucked Weight	0.895539	0.891400	0.840838	0.967998	1.000000	0.931633	0.872618	0.418847
Viscra Weight	0.902617	0.901182	0.868035	0.966080	0.931633	1.000000	0.903189	0.494249
Shell Weight	0.897628	0.907122	0.902457	0.952435	0.872618	0.903189	1.000000	0.624020
Target	0.544288	0.569380	0.606440	0.536748	0.418847	0.494249	0.624020	1.000000

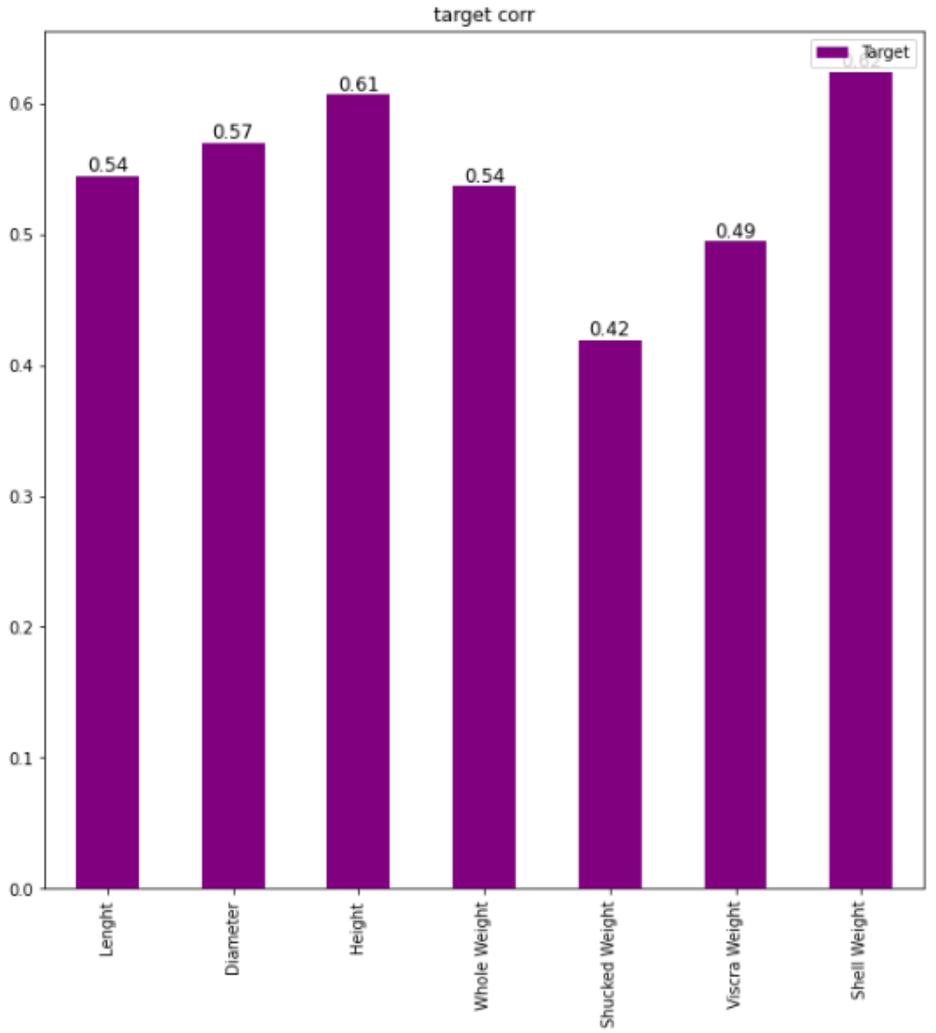
EDA

01 구현의 배경 및 관련 작업

02 상관관계 분석

Target 과의 상관관계

값들만 따로 빼서 막대그래프로 표현



비슷

EDA

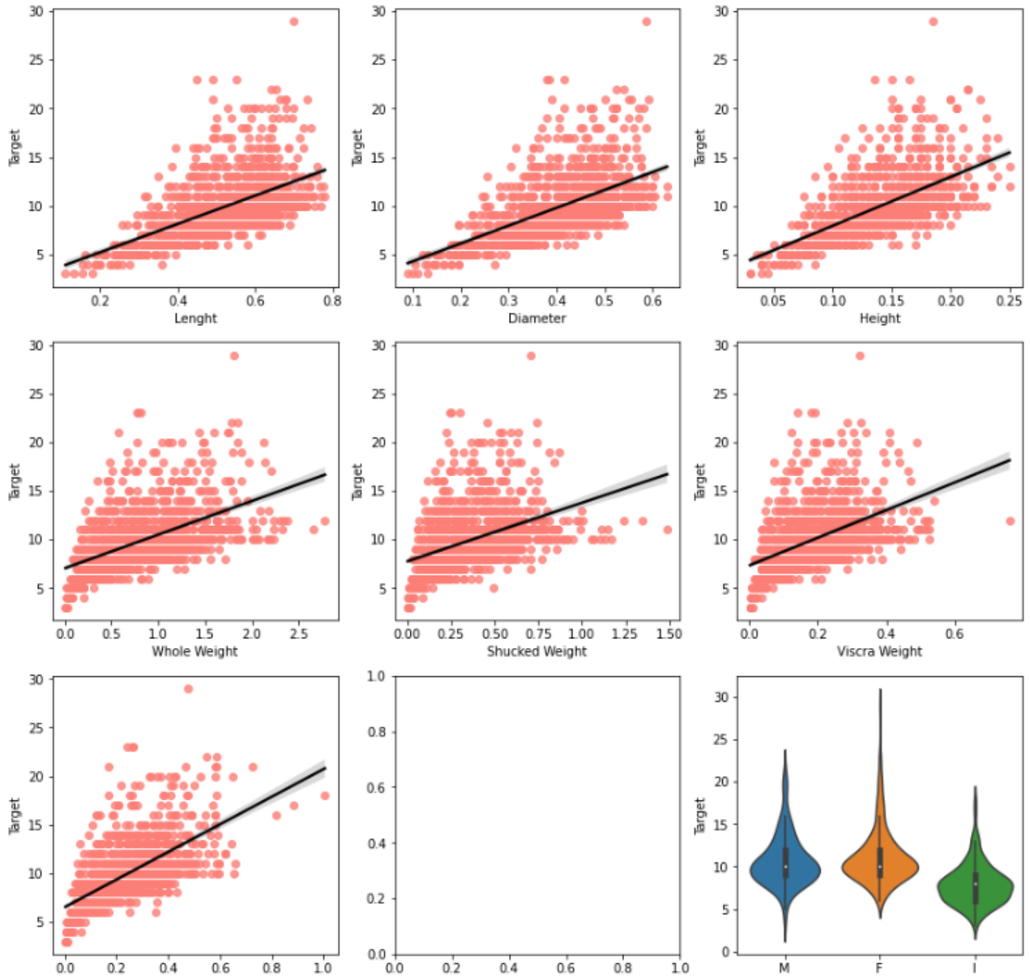
01 구현의 배경 및 관련 작업

02 상관관계 분석

Target 과의 상관관계

Target과의 상관관계 표현

corr_target



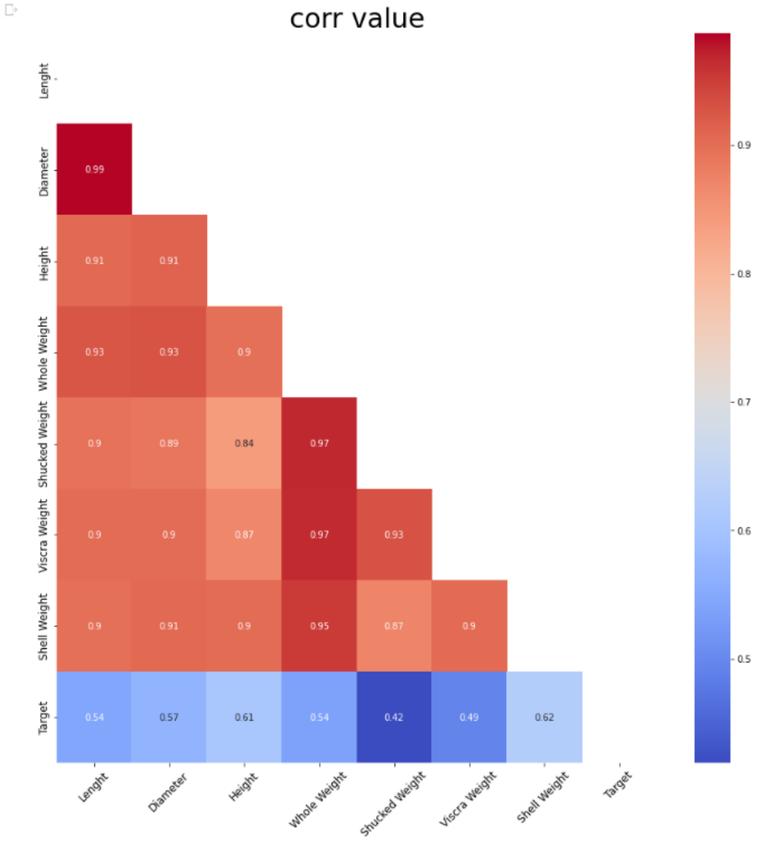
EDA

01 구현의 배경 및 관련 작업

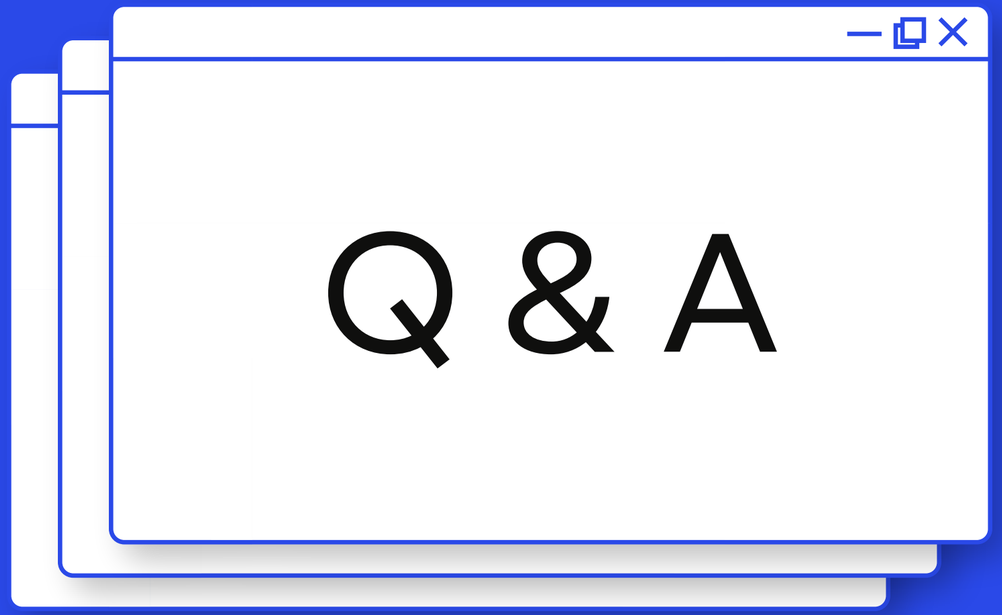
02 상관관계 분석

서로의 상관관계 그래프

다른 상관관계들도 중요하다고 생각해서, 시각화 - heatmap



서로간의 상관관계가 강하다고 나타남



CONTACT

miri77@miridih.com
010.1234.5678
instagram.com/miri_77