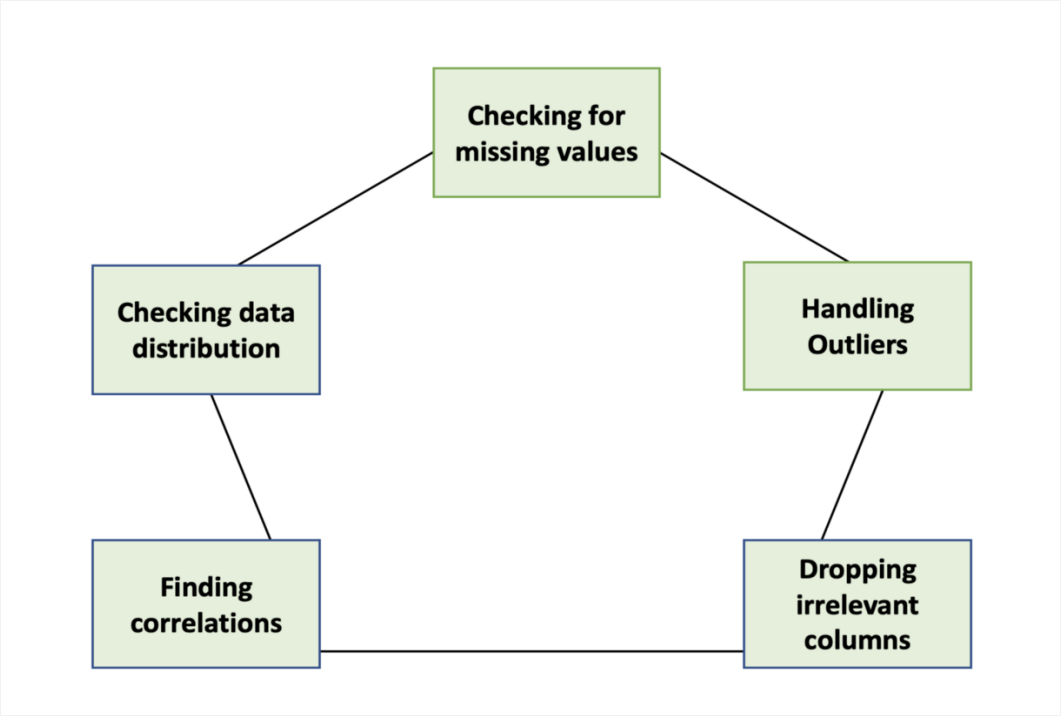
Kaggle Notebook

Automatic EDA Libraries Comparisson

일반적인 EDA 절차



데이터 분석 시 EDA과정은 필수적
시각화를 지원하는 Auto DEA 라이브러리 비교

main idea

- EDA과정에서 소비되는 plot을 그리는 시간과 결과를 비교하는 시간을 단축시켜 가장 best work를 진행 할 수 있도록 함

List

- Dataprep
- AutoViz
- Pandas Profiling
- SweetViz
- Lux

https://eda-ai-lab.tistory.com/484
https://www.kaggle.com/andreshg/automatic-eda-libraries-comparisson/comments

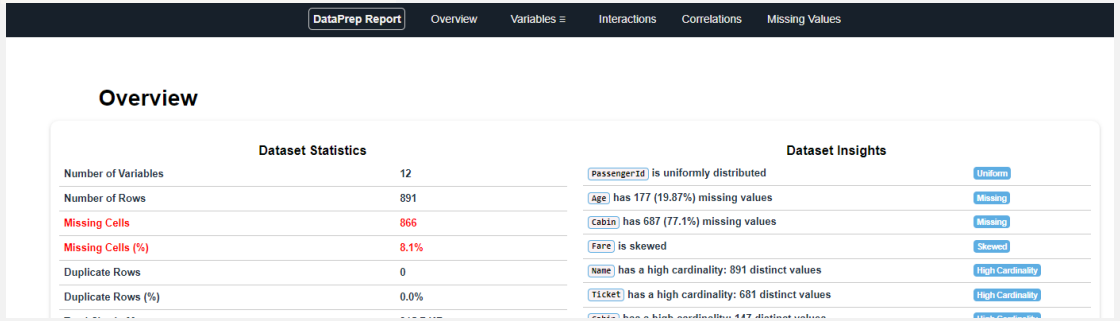
Dataprep

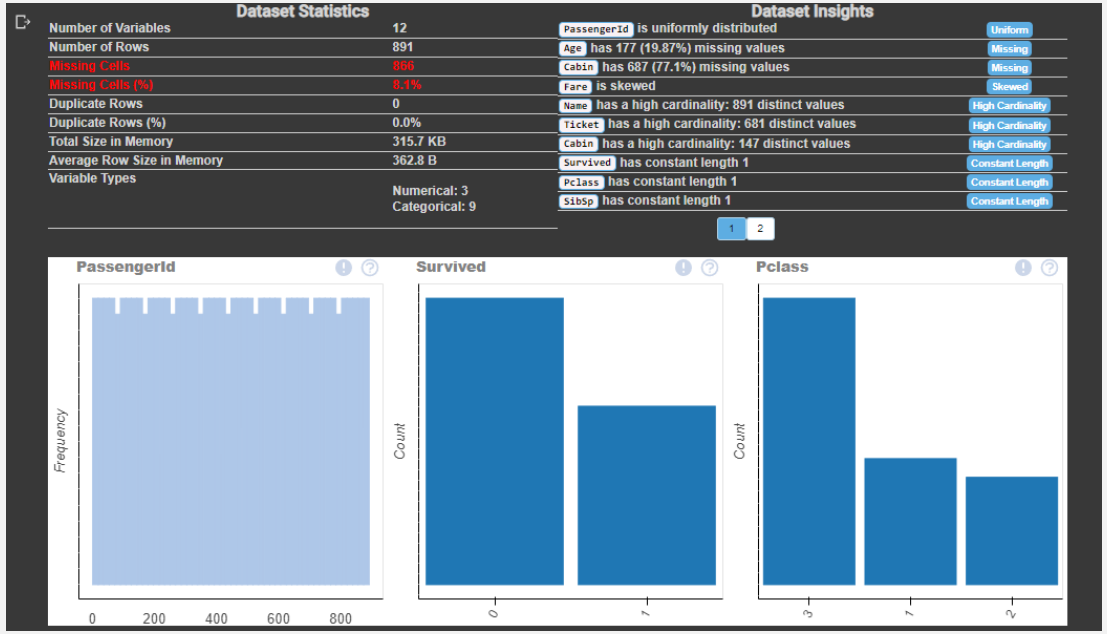
(공식 link: https://docs.dataprep.ai/user_guide/eda/introduction.html)

- dataprep.eda 패키지에는 4개의 하위 모듈이 있음
  - dataprep.eda.basic package : API plot 제공; bar chart, 각 변수에 대한 분포 등 제공
  - dataprep.eda.correlation package : 변수 간의 상관관계 분석을 위한 API plot_correlation 제공
  - dataprep.eda.create_report: 데이터 세트의 통계량, 결측값, 간단한 plot 등을 레포트 형식으로 제공
  - dataprep.eda.missing package: 결측값의 패턴과 영향을 분석을 위한 API plot_missing 제공

```
!pip install dataprep
from dataprep.eda import plot, plot_correlation, create_report, plot_missing

df = pd.read_csv('train_titanic.csv')
plot(df)
```

```
create_report(df)
```



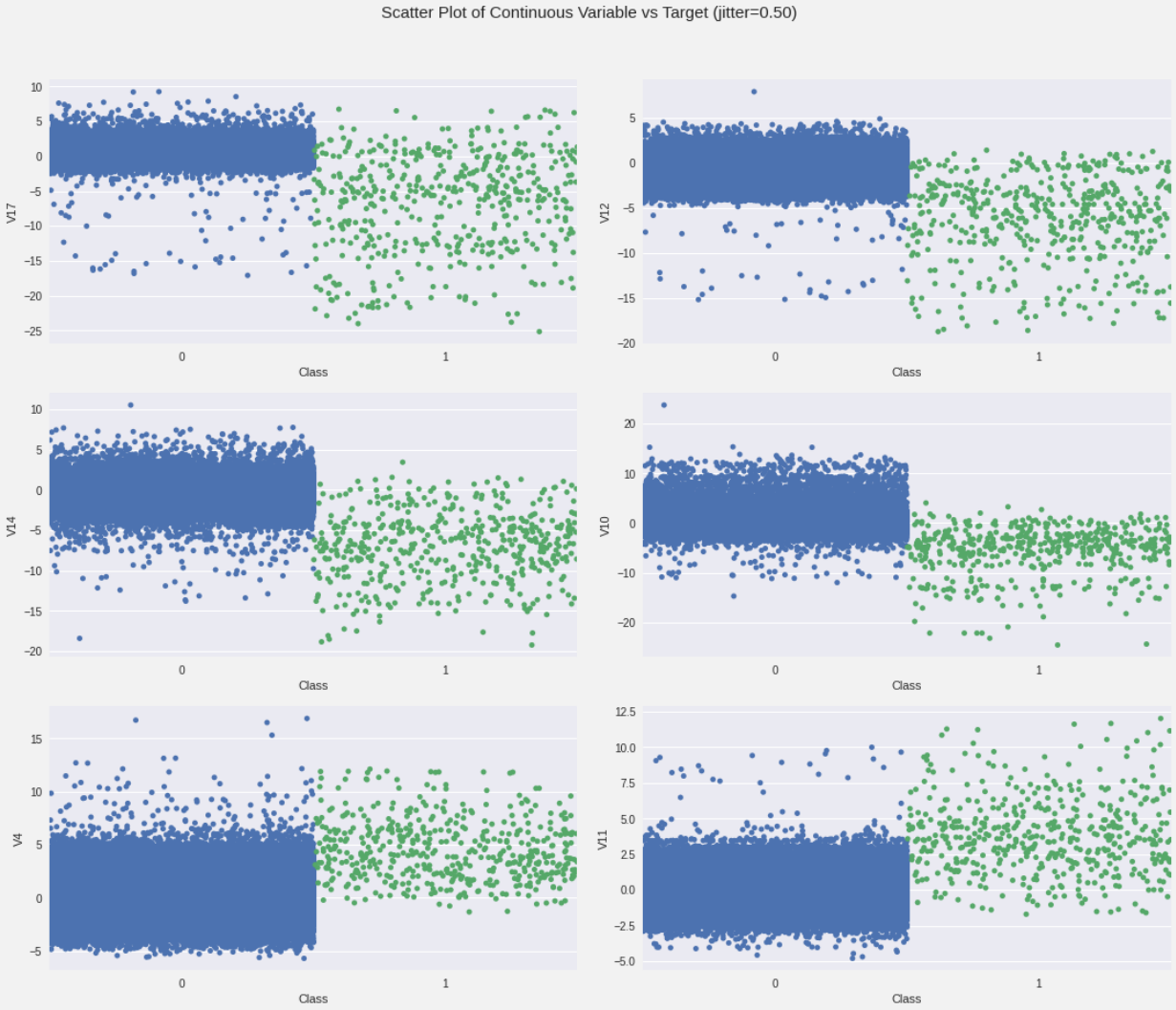


AutoViz

(공식 link: https://github.com/AutoViML/AutoViz)

- 그래프 보고서 생성
  - 데이터 세트 개요
  - 연속 변수 별 산점도
  - 범주형 변수의 분포
  - 연속형 변수의 히트 맵
  - 각 범주 형 변수 별 평균 수치 변수

```
!pip install git+git://github.com/AutoViML/AutoViz.git
!pip install xlrd

from autoviz.AutoViz_Class import AutoViz_Class

df = pd.read_csv('creditcard.csv')

AV = AutoViz_Class()
dftc = AV.AutoViz(
    filename='',
    sep='' ,
    depVar='Class',
    dfte=df,
    header=0,
    verbose=1,
    lowess=False,
    chart_format='png',
    max_rows_analyzed=300000,
    max_cols_analyzed=30
)
```

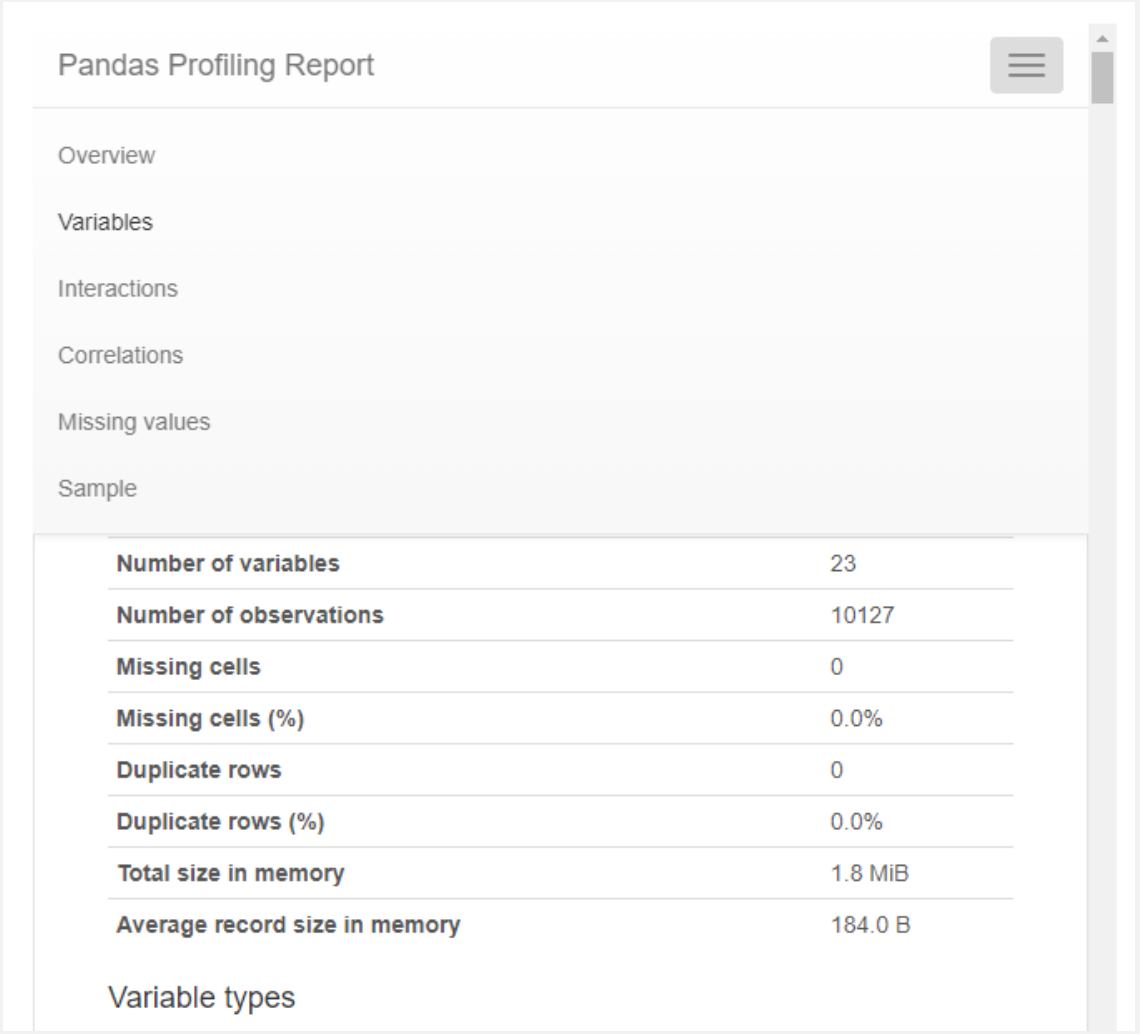


Pandas Profiling

(참고 link: https://wikidocs.net/47193)

- 보고서 생성
  - Overview: 전체적인 개요로 데이터의 크기, 변수의 수, 결측값(missing value) 비율, 데이터의 종류 제공
  - Variables: 모든 특성 변수들에 대한 결측값, 중복을 제외한 유일한 값(unique values)의 개수 등의 통계치

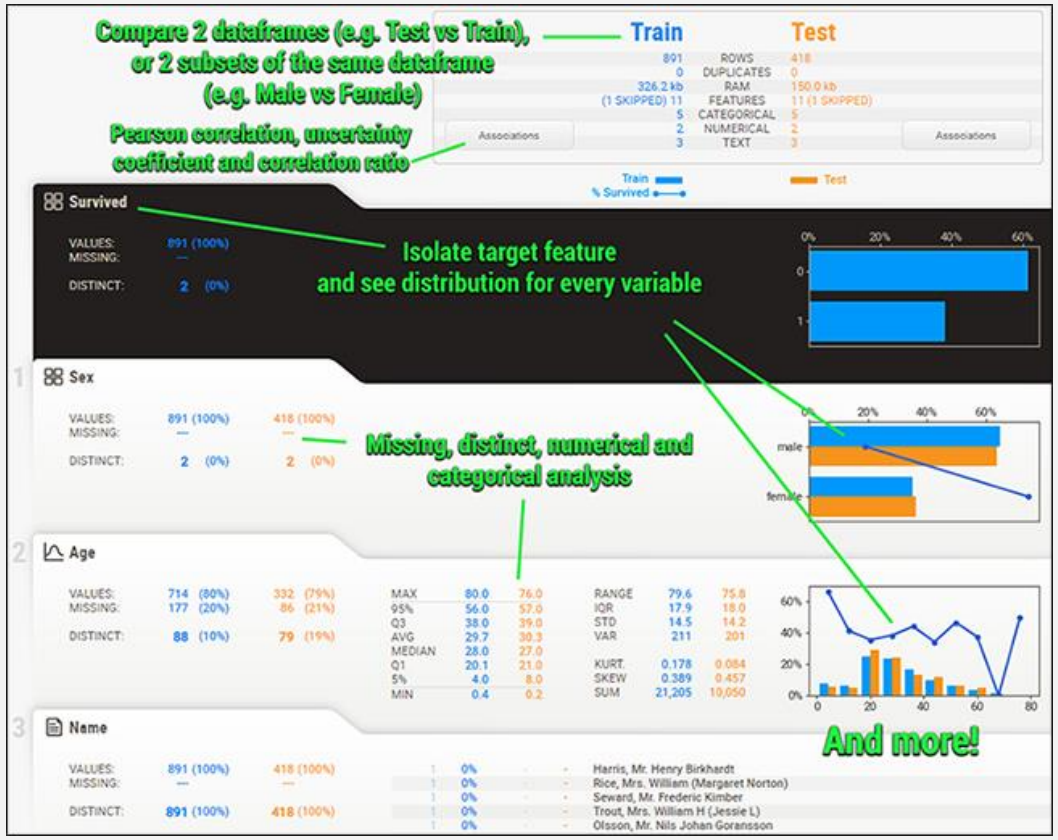
```
from pandas_profiling import ProfileReport

df = pd.read_csv('BankChurners.csv')

report = ProfileReport(df)
report
```



SweetViz

(참고 link: https://pypi.org/project/sweetviz/)

- 보고서 생성

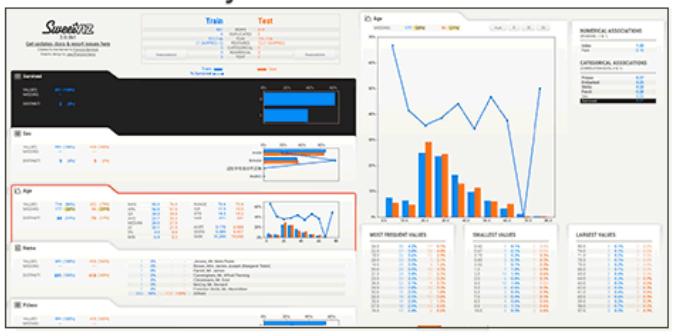



```
!pip install sweetviz
import sweetviz as sv

df = pd.read_csv('BankChurners.csv').head(2000)

advert_report = sv.analyze([df, 'Data'])
advert_report.show_html()
```

SweetViz 2.0 LAYOUT PARAMETERS

```
show_[html/notebook](...)
    layout='widescreen'
    layout='vertical'
    scale=1.0

show_notebook(...)
    w=500
    w="100%"
    h="full"
```

layout='widescreen'



Lux

(참고 link: https://github.com/lux-org/lux)

- 추세 파악 및 패턴 파악 가능
- 특정 시각화 내보내기 기능
- 관계 구분을 위한 컬러 지정 기능 등

```
!pip install lux-api
import lux

df = pd.read_csv('train_titanic.csv')
df
```

Toggle Pandas/…

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarke
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
				Cumings,