


```
pip install dataprep
```

```
Collecting dataprep
  Downloading dataprep-0.4.0-py3-none-any.whl (2.1 MB)
    |████████████████████████████████████████| 2.1 MB 3.6 MB/s
Collecting pydantic<2.0,>=1.6
  Downloading pydantic-1.8.2-cp37-cp37m-manylinux2014_x86_64.whl (10.1 MB)
    |████████████████████████████████████████| 10.1 MB 21.7 MB/s
Collecting regex<2021.0.0,>=2020.10.15
  Downloading regex-2020.11.13-cp37-cp37m-manylinux2014_x86_64.whl (719 kB)
    |████████████████████████████████████████| 719 kB 51.8 MB/s
Collecting metaphone<0.7,>=0.6
  Downloading Metaphone-0.6.tar.gz (14 kB)
Collecting aiohttp<4.0,>=3.6
  Downloading aiohttp-3.8.0-cp37-cp37m-manylinux_2_5_x86_64.manylinux1_x86_64.manyli
    |████████████████████████████████████████| 1.1 MB 46.5 MB/s
Collecting dask[array,dataframe,delayer]<3.0,>=2.25
  Downloading dask-2.30.0-py3-none-any.whl (848 kB)
    |████████████████████████████████████████| 848 kB 37.8 MB/s
Requirement already satisfied: pandas<2.0,>=1.1 in /usr/local/lib/python3.7/dist-pac
Collecting wordcloud<2.0,>=1.8
  Downloading wordcloud-1.8.1-cp37-cp37m-manylinux1_x86_64.whl (366 kB)
    |████████████████████████████████████████| 366 kB 72.7 MB/s
Requirement already satisfied: jinja2<3.0,>=2.11 in /usr/local/lib/python3.7/dist-pa
Collecting levenshtein<0.13.0,>=0.12.0
  Downloading levenshtein-0.12.0-cp37-cp37m-manylinux1_x86_64.whl (158 kB)
    |████████████████████████████████████████| 158 kB 72.7 MB/s
Collecting usaddress<0.6.0,>=0.5.10
  Downloading usaddress-0.5.10-py2.py3-none-any.whl (63 kB)
    |████████████████████████████████████████| 63 kB 2.7 MB/s
Collecting python_stdnum<2.0,>=1.16
  Downloading python_stdnum-1.17-py2.py3-none-any.whl (943 kB)
    |████████████████████████████████████████| 943 kB 53.4 MB/s
Requirement already satisfied: numpy<2,>=1 in /usr/local/lib/python3.7/dist-packages
Requirement already satisfied: tqdm<5.0,>=4.48 in /usr/local/lib/python3.7/dist-pack
Collecting nltk<4.0,>=3.5
  Downloading nltk-3.6.5-py3-none-any.whl (1.5 MB)
    |████████████████████████████████████████| 1.5 MB 58.4 MB/s
Requirement already satisfied: bokeh<3,>=2 in /usr/local/lib/python3.7/dist-packages
Requirement already satisfied: scipy<2,>=1 in /usr/local/lib/python3.7/dist-packages
Requirement already satisfied: ipywidgets<8.0,>=7.5 in /usr/local/lib/python3.7/dist
Requirement already satisfied: bottleneck<2.0,>=1.3 in /usr/local/lib/python3.7/dist
Collecting varname<0.9.0,>=0.8.1
  Downloading varname-0.8.1-py3-none-any.whl (20 kB)
Collecting jsonpath-ng<2.0,>=1.5
  Downloading jsonpath_ng-1.5.3-py3-none-any.whl (29 kB)
Collecting frozenlist>=1.1.1
  Downloading frozenlist-1.2.0-cp37-cp37m-manylinux_2_5_x86_64.manylinux1_x86_64.man
    |████████████████████████████████████████| 192 kB 48.9 MB/s
Requirement already satisfied: attrs>=17.3.0 in /usr/local/lib/python3.7/dist-packag
Requirement already satisfied: charset-normalizer<3.0,>=2.0 in /usr/local/lib/python
Collecting asynctest==0.13.0
  Downloading asynctest-0.13.0-py3-none-any.whl (26 kB)
Collecting multidict<7.0,>=4.5
  Downloading multidict-5.2.0-cp37-cp37m-manylinux_2_5_x86_64.manylinux1_x86_64.many
```

 160 kB 58.7 MB/s
Collecting aiosignal>=1.1.2
Downloading aiosignal-1.2.0-py3-none-any.whl (8.2 kB)
Collecting yarl<2.0,>=1.0

```
from dataprep.datasets import get_dataset_names
get_dataset_names()
```

```
['patient_info',
 'house_prices_test',
 'titanic',
 'wine-quality-red',
 'adult',
 'waste_hauler',
 'covid19',
 'iris',
 'countries',
 'house_prices_train']
```

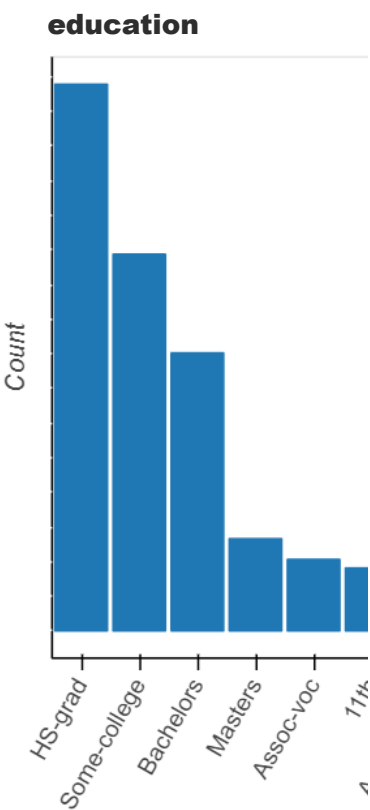
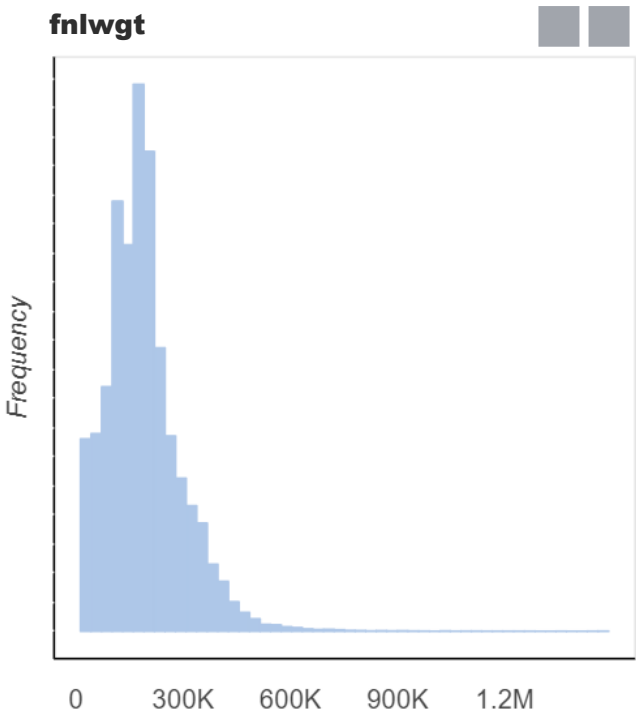
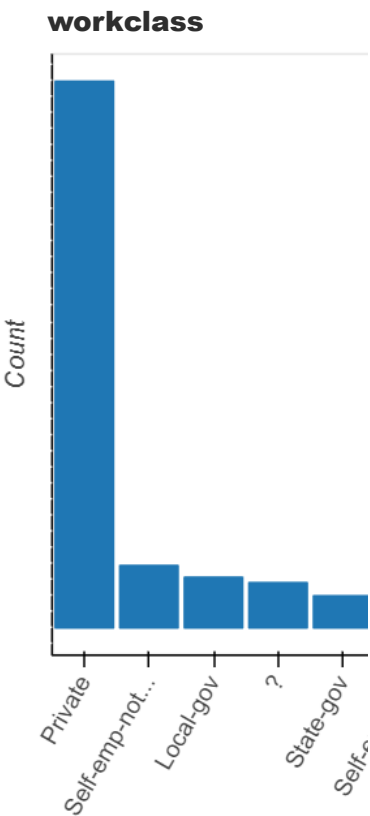
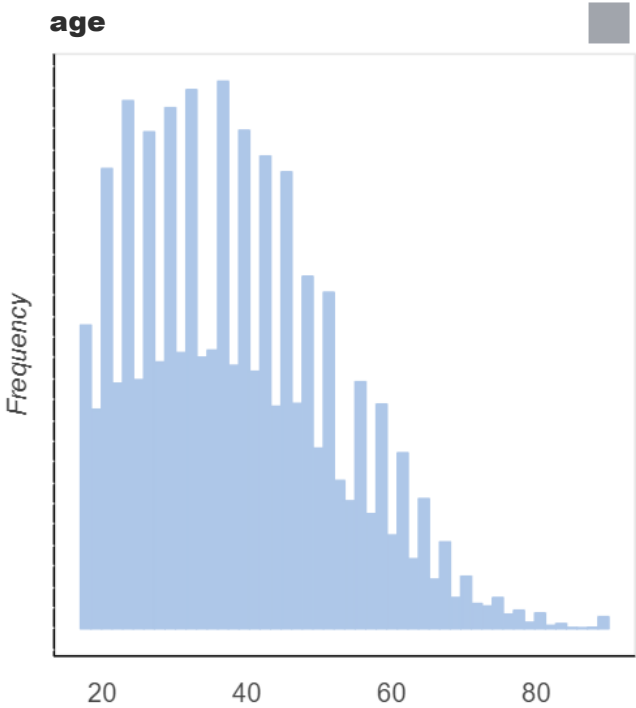
```
import numpy as np
import pandas as pd
```

```
from dataprep.datasets import load_dataset
df = load_dataset('adult')
df.head()
df1 = load_dataset('titanic')
```

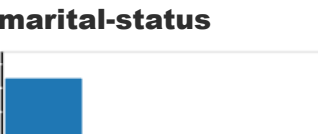
▼ EDA

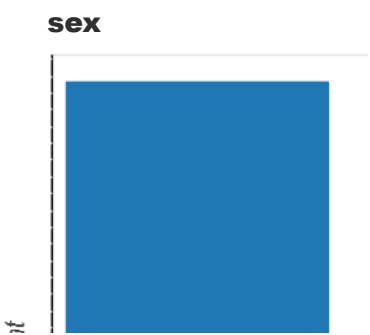
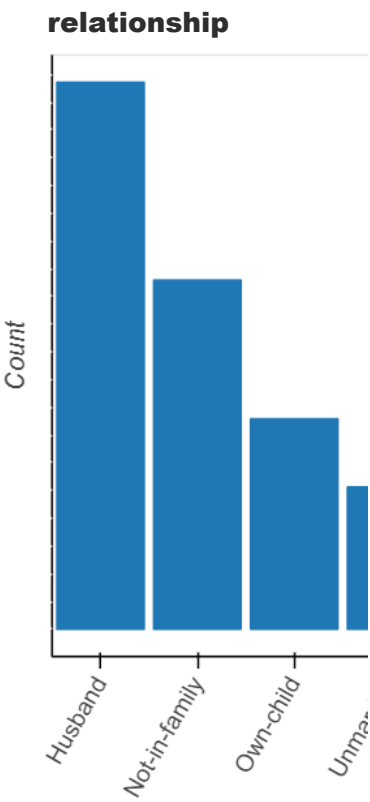
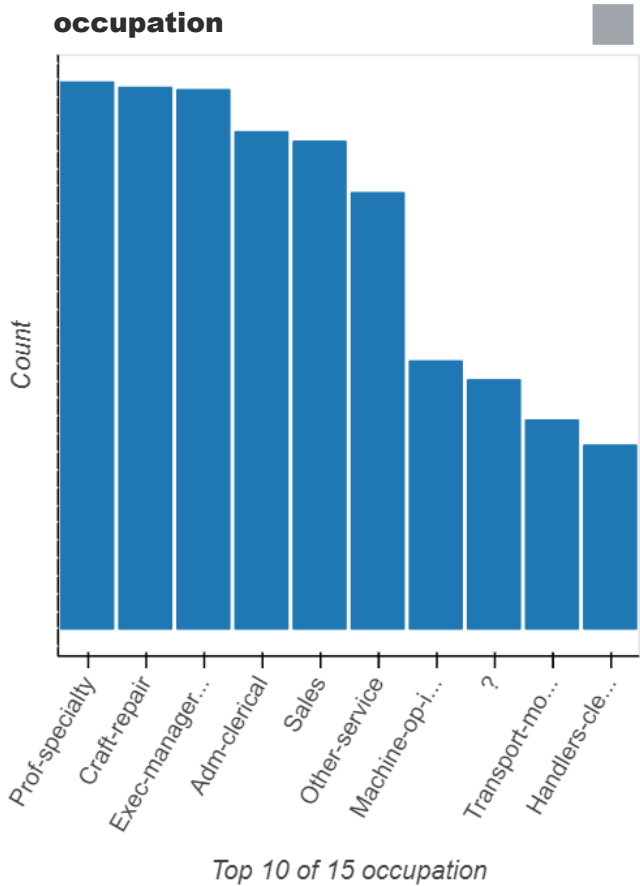
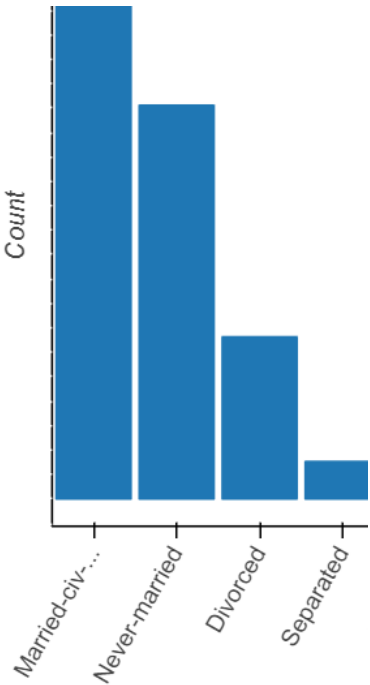
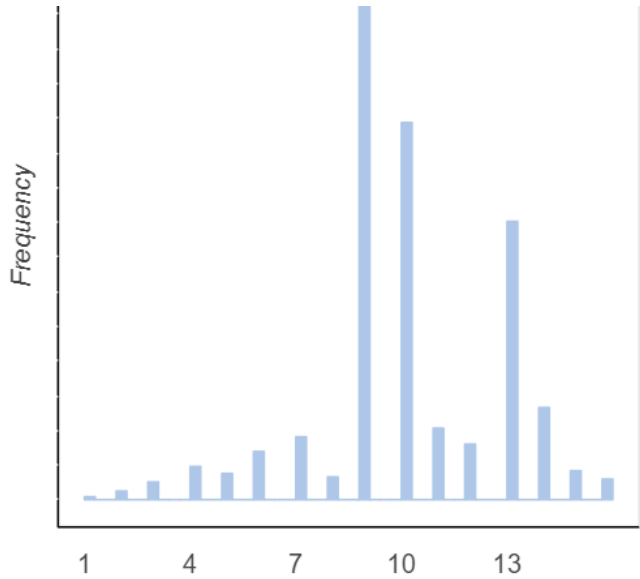
distribution with plot()

```
from dataprep.eda import plot
plot(df)
```



Top 10 of 16 categories





correlation()

```
from dataprep.eda import plot_correlation
plot_correlation(df)
```

Stats	Pearson	Spearman	KendallTau
Pearson Spearman KendallTau			
Highest Positive Correlation	0.144	0.164	0.13
Highest Negative Correlation	-0.077	-0.078	-0.064
Lowest Correlation	0.004	0.001	0.001
Mean Correlation	0.031	0.039	0.033

```
from dataprep.eda import plot_missing
plot_missing(df1)
```

Stats	Bar Chart	Spectrum	Heat Map	Dendrogram
Missing Statistics				
Missing Cells	866			
Missing Cells (%)	8.1%			
Missing Columns	3			
Missing Rows	708			
Avg Missing Cells per Column	72.17			
Avg Missing Cells per Row	0.97			

```
df2 = df.iloc[30000:]
df2.shape
```

(18842, 15)

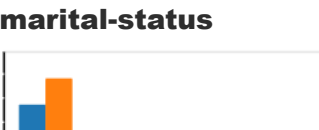
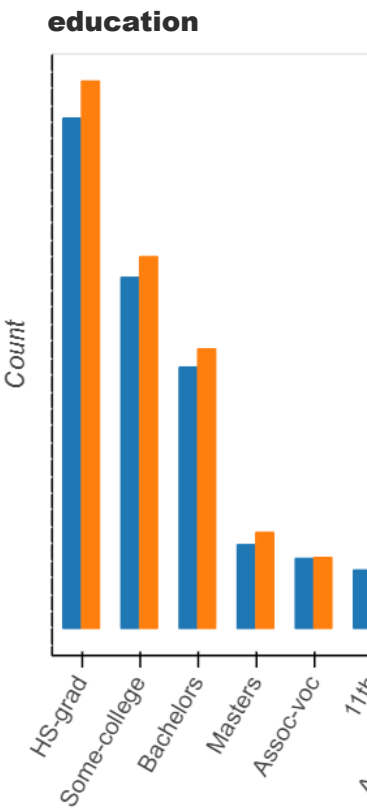
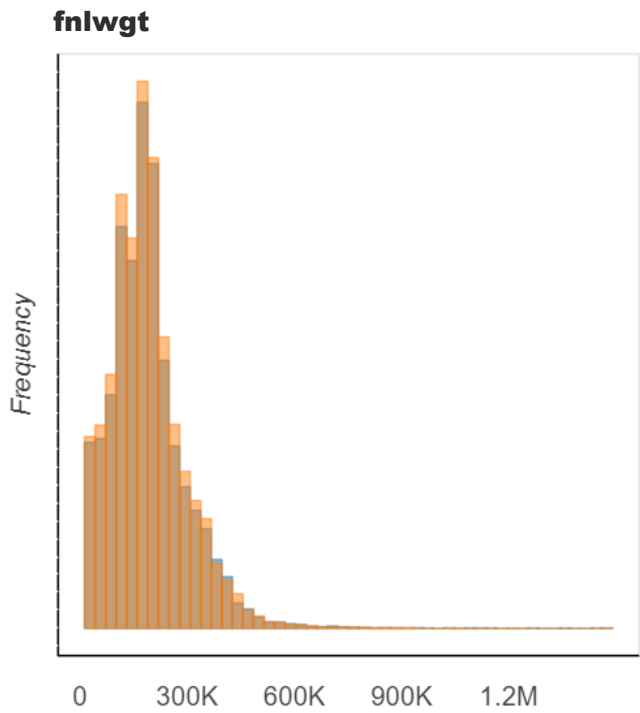
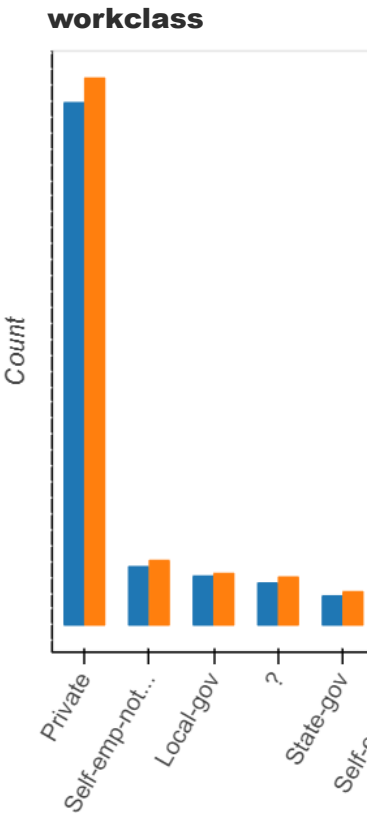
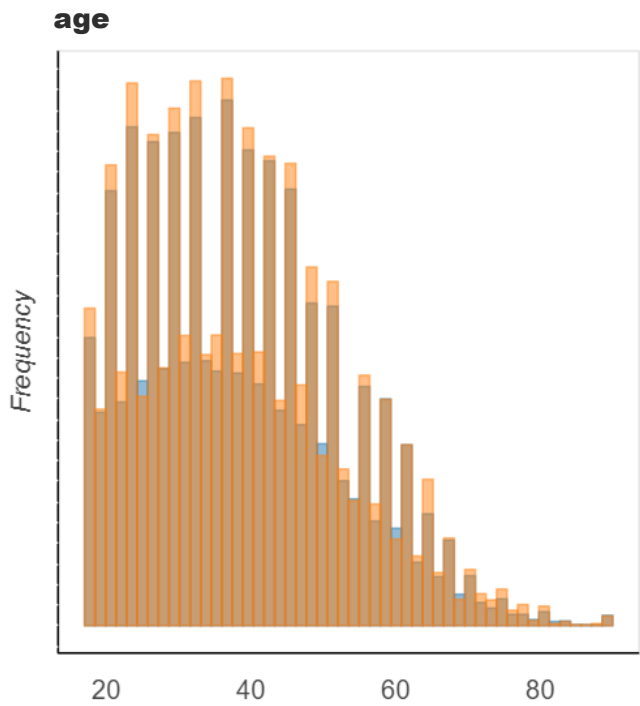
```
df3 = df.iloc[:20000]
df3.shape
```

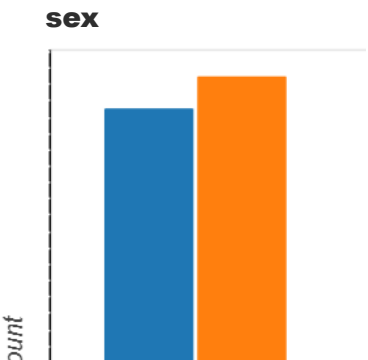
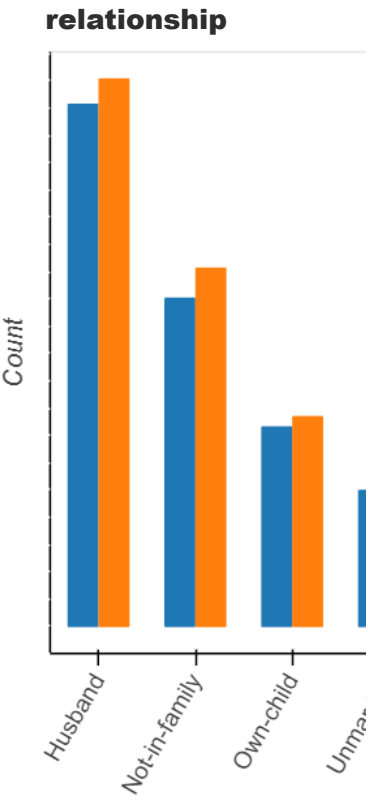
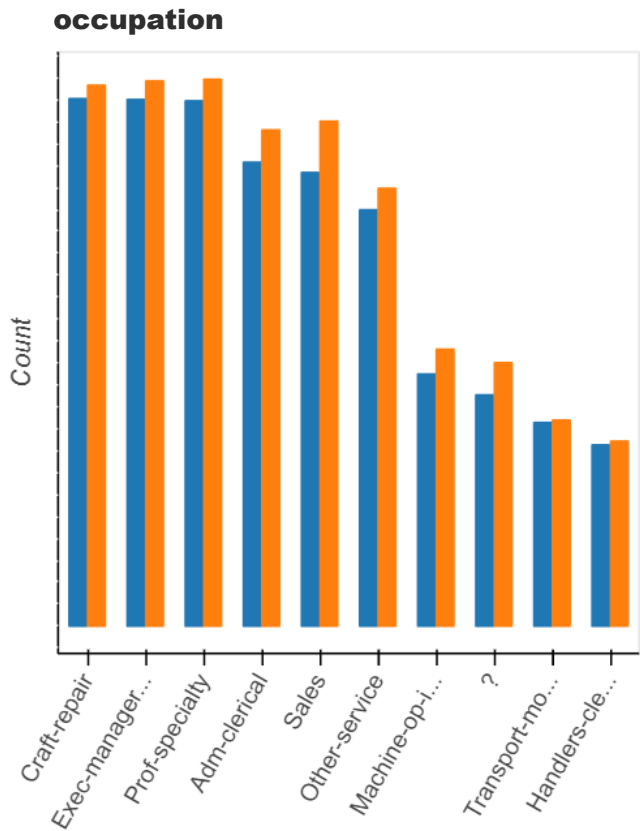
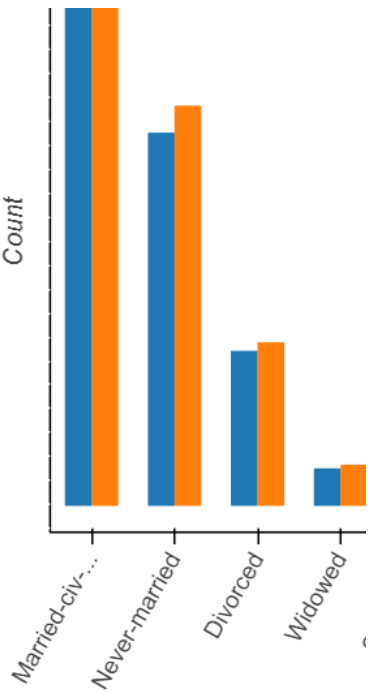
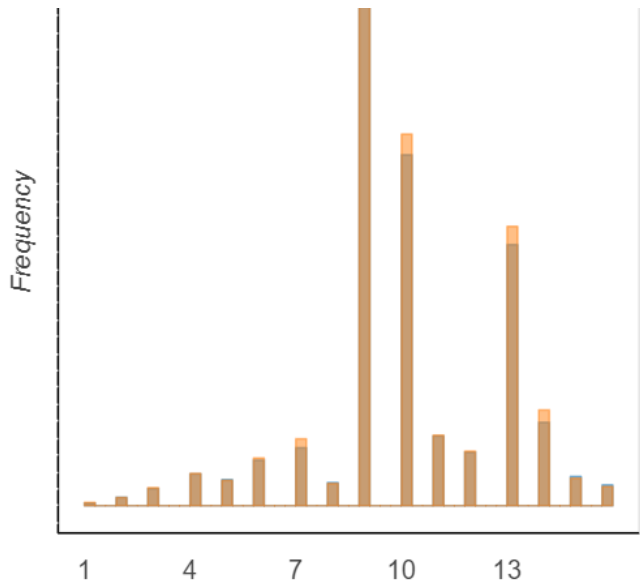
(20000, 15)

```
from dataprep.eda import plot_diff
plot_diff([df2,df3])
```

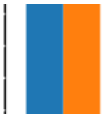
Show Stats

df1 df2





C



C



```
from dataprep.eda import create_report  
report = create_report(df)
```

```
report.show()
```


<div>sex</div> <div>categorical</div> <div>Show Details</div>	Approximate Distinct Count		2	Count	28K		
	Approximate Unique (%)		0.0%		21K		
	Missing		0		14K		
	Missing (%)		0.0%		7000		
	Memory Size		3.3 MB		0		
<div>capital-gain</div> <div>numerical</div> <div>Show Details</div>	Approximate Distinct Count	123	Mean	1079.0676	Frequency	40K	
	Approximate Unique (%)	0.3%	Minimum	0		30K	
	Missing	0	Maximum	99999		20K	
	Missing (%)	0.0%	Zeros	44807		10K	
	Infinite	0	Zeros (%)	91.7%		0	
	Infinite (%)	0.0%	Negatives	0			
	Memory Size	763.2 KB	Negatives (%)	0.0%			
<div>capital-loss</div> <div>numerical</div> <div>Show Details</div>	Approximate Distinct Count	99	Mean	87.5023	Frequency	40K	
	Approximate Unique (%)	0.2%	Minimum	0		30K	
	Missing	0	Maximum	4356		20K	
	Missing (%)	0.0%	Zeros	46560		10K	
	Infinite	0	Zeros (%)	95.3%		0	
	Infinite (%)	0.0%	Negatives	0			
	Memory Size	763.2 KB	Negatives (%)	0.0%			
<div>hours-per-week</div> <div>numerical</div> <div>Show Details</div>	Approximate Distinct Count	96	Mean	40.4224	Frequency	20K	
	Approximate Unique (%)	0.2%	Minimum	1		15K	
	Missing	0	Maximum	99		10K	
	Missing (%)	0.0%	Zeros	0			