Data Integration

Setup

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

```
df = pd.read_csv('/content/Data Science Jobs Salaries.csv', skiprows = 2)
df.head()
```

	work_year	experience_level	employment_type	job_title	salary	salary_curr
0	2021e	EN	FT	Data Science Consultant	54000	
1	2020	SE	FT	Data Scientist	60000	
2	2021e	EX	FT	Head of Data Science	85000	
3	2021e	EX	FT	Head of Data	230000	
4	2021e	EN	FT	Machine Learning Engineer	125000	

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 245 entries, 0 to 244
Data columns (total 11 columns):
 #   Column              Non-Null Count  Dtype 
---  ------              --------------  ----- 
 0   work_year           245 non-null    object
 1   experience_level    245 non-null    object
 2   employment_type     245 non-null    object
 3   job_title           245 non-null    object
 4   salary              245 non-null    int64 
 5   salary_currency     245 non-null    object
 6   salary_in_usd       245 non-null    int64 
 7   employee_residence  245 non-null    object
 8   remote_ratio        245 non-null    int64 
 9   company_location    245 non-null    object
 10  company_size        245 non-null    object
dtypes: int64(3), object(8)
memory usage: 21.2+ KB
```

Concatenation

> Documentation: https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.concat.html

```
df_1 = df[['company_location','job_title','experience_level','salary_in_usd']].sam
df_2 = df[['company_location','job_title','experience_level','salary_in_usd']].sam
df_3 = df[['company_location','job_title','experience_level','salary_in_usd']].sam
```

```
df_1
```

	company_location	job_title	experience_level	salary_in_usd
16	US	Data Engineer	MI	90000
125	IN	Data Scientist	MI	16949
25	PL	Director of Data Science	EX	154963
22	US	ML Engineer	MI	270000
41	US	Head of Data	EX	235000

```
df_2
```

	company_location	job_title	experience_level	salary_in_usd
216	ES	Data Scientist	MI	38776
73	US	Data Analyst	MI	93000
137	US	Data Scientist	MI	147000
28	GB	Research Scientist	EN	83000
45	DE	Data Science Consultant	EN	77481

```
df_3
```

	company_location	job_title	experience_level	salary_in_usd
92	AE	Lead Data Scientist	MI	115000
70	US	Data Scientist	MI	105000
242	US	Data Scientist	EN	105000
130	CA	Data Analyst	SE	71968
84	GB	Data Engineer	MI	72625

```
df_cat1 = pd.concat([df_1,df_2,df_3], axis=0)
df_cat1
```

	company_location	job_title	experience_level	salary_in_usd
16	US	Data Engineer	MI	90000
125	IN	Data Scientist	MI	16949
25	PL	Director of Data Science	EX	154963
22	US	ML Engineer	MI	270000
41	US	Head of Data	EX	235000
216	ES	Data Scientist	MI	38776
73	US	Data Analyst	MI	93000
137	US	Data Scientist	MI	147000
28	GB	Research Scientist	EN	83000
45	DE	Data Science Consultant	EN	77481
92	AE	Lead Data Scientist	MI	115000
70	US	Data Scientist	MI	105000
242	US	Data Scientist	EN	105000
130	CA	Data Analyst	SE	71968
84	GB	Data Engineer	MI	72625

```
df_cat1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 15 entries, 16 to 84
Data columns (total 4 columns):
 #   Column            Non-Null Count  Dtype 
---  ------            --------------  ----- 
 0   company_location  15 non-null     object
 1   job_title         15 non-null     object
 2   experience_level  15 non-null     object
 3   salary_in_usd     15 non-null     int64 
dtypes: int64(1), object(3)
memory usage: 600.0+ bytes
```

```
df_cat2 = pd.concat([df_1,df_2,df_3], axis=1)
df_cat2
```

	company_location	job_title	experience_level	salary_in_usd	company_loc
16	US	Data Engineer	MI	90000.0	
125	IN	Data Scientist	MI	16949.0	
25	PL	Director of Data Science	EX	154963.0	
22	US	ML Engineer	MI	270000.0	
41	US	Head of Data	EX	235000.0	
216	NaN	NaN	NaN	NaN	
73	NaN	NaN	NaN	NaN	
137	NaN	NaN	NaN	NaN	
28	NaN	NaN	NaN	NaN	
45	NaN	NaN	NaN	NaN	
92	NaN	NaN	NaN	NaN	
70	NaN	NaN	NaN	NaN	
242	NaN	NaN	NaN	NaN	
130	NaN	NaN	NaN	NaN	
84	NaN	NaN	NaN	NaN	

```
df_cat2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 15 entries, 16 to 84
Data columns (total 12 columns):
 #   Column            Non-Null Count  Dtype  
---  ------            --------------  -----  
 0   company_location  5 non-null      object 
 1   job_title         5 non-null      object 
 2   experience_level  5 non-null      object 
 3   salary_in_usd     5 non-null      float64
 4   company_location  5 non-null      object 
 5   job_title         5 non-null      object 
 6   experience_level  5 non-null      object 
 7   salary_in_usd     5 non-null      float64
 8   company_location  5 non-null      object 
 9   job_title         5 non-null      object 
 10  experience_level  5 non-null      object 
 11  salary_in_usd     5 non-null      float64
dtypes: float64(3), object(9)
memory usage: 1.5+ KB
```

Merging

> Documentation:
> https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.merge.html

```
df_1=df[['company_location','experience_level','salary_in_usd']][0:5]
df_1
```

	company_location	experience_level	salary_in_usd
0	DE	EN	64369
1	US	SE	68428
2	RU	EX	85000
3	RU	EX	230000
4	US	EN	125000

```
df_2=df[['company_location','job_title','salary_in_usd']][0:5]
df_2
```

	company_location	job_title	salary_in_usd
0	DE	Data Science Consultant	64369
1	US	Data Scientist	68428
2	RU	Head of Data Science	85000
3	RU	Head of Data	230000
4	US	Machine Learning Engineer	125000

```
pd.merge(df_1,df_2,on='company_location',how='inner')
```

	company_location	experience_level	salary_in_usd_x	job_title	salary_in_u
0	DE	EN	64369	Data Science Consultant	6
1	US	SE	68428	Data Scientist	6
2	US	SE	68428	Machine Learning Engineer	12
3	US	EN	125000	Data Scientist	6
4	US	EN	125000	Machine Learning Engineer	12
5	RU	EX	85000	Head of Data Science	8
6	RU	EX	85000	Head of Data	23
7	RU	EX	230000	Head of	

```
pd.merge(df_1,df_2,on='company_location',how='inner').drop_duplicates()
```

	company_location	experience_level	salary_in_usd_x	job_title	salary_in_u
0	DE	EN	64369	Data Science Consultant	6
1	US	SE	68428	Data Scientist	6
2	US	SE	68428	Machine Learning Engineer	12
3	US	EN	125000	Data Scientist	6
4	US	EN	125000	Machine Learning Engineer	12
5	RU	EX	85000	Head of Data Science	8
6	RU	EX	85000	Head of Data	23
7	RU	EX	230000	Head of Data	8

```
df_3=df[['company_location','job_title','experience_level',]][2:6]
df_3
```

	company_location	job_title	experience_level
2	RU	Head of Data Science	EX
3	RU	Head of Data	EX
4	US	Machine Learning Engineer	EN
5	US	Data Analytics Manager	SE

```
pd.merge(df_1,df_3,on='company_location',how='inner').drop_duplicates()
```

	company_location	experience_level_x	salary_in_usd	job_title	experience_
0	US	SE	68428	Machine Learning Engineer	
1	US	SE	68428	Data Analytics Manager	
2	US	EN	125000	Machine Learning Engineer	
3	US	EN	125000	Data Analytics Manager	
4	RU	EX	85000	Head of Data Science	
5	RU	EX	85000	Head of Data	
6	RU	EX	230000	Head of Data Science	
7	RU	EX	230000	Head of Data	

```
pd.merge(df_1,df_3,on='company_location',how='outer').drop_duplicates()
```

	company_location	experience_level_x	salary_in_usd	job_title	experience_
0	DE	EN	64369	NaN	
1	US	SE	68428	Machine Learning Engineer	
2	US	SE	68428	Data Analytics Manager	
3	US	EN	125000	Machine Learning Engineer	
4	US	EN	125000	Data Analytics Manager	
5	RU	EX	85000	Head of Data Science	
6	RU	EX	85000	Head of Data	
7	RU	EX	230000	Head of Data Science	
8	RU	EX	230000	Head of Data	

Joining

> documentation:
> https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.join.html

```
df_1=df[['experience_level']][0:5]
df_1
```

	experience_level
0	EN
1	SE
2	EX
3	EX
4	EN

```
df_2=df[['job_title']][2:7]
df_2
```

	job_title
2	Head of Data Science
3	Head of Data
4	Machine Learning Engineer
5	Data Analytics Manager
6	Research Scientist

```
df_1.join(df_2,how='left').drop_duplicates()
```

	experience_level	job_title
0	EN	NaN
1	SE	NaN
2	EX	Head of Data Science
3	EX	Head of Data
4	EN	Machine Learning Engineer

```
df_1.join(df_2,how='right').drop_duplicates()
```

	experience_level	job_title
2	EX	Head of Data Science
3	EX	Head of Data
4	EN	Machine Learning Engineer
5	NaN	Data Analytics Manager
6	NaN	Research Scientist

```
df_1.join(df_2,how='inner').drop_duplicates()
```

	experience_level	job_title
2	EX	Head of Data Science
3	EX	Head of Data
4	EN	Machine Learning Engineer

```
df_1.join(df_2,how='outer').drop_duplicates()
```

	experience_level	job_title
0	EN	NaN
1	SE	NaN
2	EX	Head of Data Science
3	EX	Head of Data
4	EN	Machine Learning Engineer
5	NaN	Data Analytics Manager
6	NaN	Research Scientist

Colab paid products - Cancel contracts here