Data Integration

Setup

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

```
df = pd.read_csv('/content/sample_data/california_housing_test.csv')
df.head()
```

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	popula
0	-122.05	37.37	27.0	3885.0	661.0	1
1	-118.30	34.26	43.0	1510.0	310.0	
2	-117.81	33.78	27.0	3589.0	507.0	1
3	-118.36	33.82	28.0	67.0	15.0	
4	-119.67	36.33	19.0	1241.0	244.0	

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3000 entries, 0 to 2999
Data columns (total 9 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   longitude           3000 non-null   float64
 1   latitude            3000 non-null   float64
 2   housing_median_age  3000 non-null   float64
 3   total_rooms         3000 non-null   float64
 4   total_bedrooms      3000 non-null   float64
 5   population          3000 non-null   float64
 6   households          3000 non-null   float64
 7   median_income       3000 non-null   float64
 8   median_house_value  3000 non-null   float64
dtypes: float64(9)
memory usage: 211.1 KB
```

Concatenation

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

```
df_1 = df[['longitude','latitude','median_income']].sample(n=5)
df_2 = df[['longitude','latitude','median_income']].sample(n=5)
df_3 = df[['longitude','latitude','median_income']].sample(n=5)
```

```
df_1
```

	longitude	latitude	median_income
362	-117.19	32.77	3.8571
2425	-121.32	38.62	3.0864
1863	-118.36	33.82	3.3565
1059	-119.75	36.78	2.3333
1751	-121.96	37.34	5.7910

```
df_2
```

	longitude	latitude	median_income
2286	-122.20	37.47	4.2083
1933	-118.27	33.93	2.6458
1214	-121.00	37.60	2.6899
2372	-122.04	37.97	2.3152
483	-115.90	32.69	1.5417

```
df_3
```

	longitude	latitude	median_income
2731	-117.69	34.04	4.0096
1902	-117.90	36.95	1.7292
2683	-118.05	34.14	8.9728
937	-121.27	38.14	2.2883
1671	-117.98	33.76	4.4545

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

	longitude	latitude	median_income
362	-117.19	32.77	3.8571
2425	-121.32	38.62	3.0864
1863	-118.36	33.82	3.3565
1059	-119.75	36.78	2.3333
1751	-121.96	37.34	5.7910
2286	-122.20	37.47	4.2083
1933	-118.27	33.93	2.6458
1214	-121.00	37.60	2.6899
2372	-122.04	37.97	2.3152
483	-115.90	32.69	1.5417
2731	-117.69	34.04	4.0096
1902	-117.90	36.95	1.7292
2683	-118.05	34.14	8.9728
937	-121.27	38.14	2.2883
1671	-117.98	33.76	4.4545

```
df_cat1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 15 entries, 362 to 1671
Data columns (total 3 columns):
 #   Column         Non-Null Count  Dtype  
---  ------         --------------  -----  
 0   longitude      15 non-null     float64
 1   latitude       15 non-null     float64
 2   median_income  15 non-null     float64
dtypes: float64(3)
memory usage: 480.0 bytes
```

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

	longitude	latitude	median_income	longitude	latitude	median_income	lo
362	-117.19	32.77	3.8571	NaN	NaN	NaN	
2425	-121.32	38.62	3.0864	NaN	NaN	NaN	
1863	-118.36	33.82	3.3565	NaN	NaN	NaN	
1059	-119.75	36.78	2.3333	NaN	NaN	NaN	
1751	-121.96	37.34	5.7910	NaN	NaN	NaN	
2286	NaN	NaN	NaN	-122.20	37.47	4.2083	
1933	NaN	NaN	NaN	-118.27	33.93	2.6458	
1214	NaN	NaN	NaN	-121.00	37.60	2.6899	
2372	NaN	NaN	NaN	-122.04	37.97	2.3152	
483	NaN	NaN	NaN	-115.90	32.69	1.5417	
2731	NaN	NaN	NaN	NaN	NaN	NaN	
1902	NaN	NaN	NaN	NaN	NaN	NaN	
2683	NaN	NaN	NaN	NaN	NaN	NaN	
937	NaN	NaN	NaN	NaN	NaN	NaN	
1671	NaN	NaN	NaN	NaN	NaN	NaN	

```
df_cat2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 15 entries, 362 to 1671
Data columns (total 9 columns):
 #   Column         Non-Null Count  Dtype  
---  ------         --------------  -----  
 0   longitude      5 non-null      float64
 1   latitude       5 non-null      float64
 2   median_income  5 non-null      float64
 3   longitude      5 non-null      float64
 4   latitude       5 non-null      float64
 5   median_income  5 non-null      float64
 6   longitude      5 non-null      float64
 7   latitude       5 non-null      float64
 8   median_income  5 non-null      float64
dtypes: float64(9)
memory usage: 1.2 KB
```

```
df_1 = df[['longitude']][:5]
df_2 = df[['latitude']][:5]
df_3 = df[['median_income']][:5]
df_1, df_2, df_3
```

```
(   longitude
 0    -122.05
 1    -118.30
 2    -117.81
 3    -118.36
 4    -119.67,
    latitude
 0     37.37
 1     34.26
 2     33.78
 3     33.82
 4     36.33,
    median_income
 0         6.6085
 1         3.5990
 2         5.7934
 3         6.1359
 4         2.9375)
```

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

	longitude	latitude	median_income
0	-122.05	37.37	6.6085
1	-118.30	34.26	3.5990
2	-117.81	33.78	5.7934
3	-118.36	33.82	6.1359
4	-119.67	36.33	2.9375

Merging

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

```
df_1=df[['longitude','median_income']][0:5]
df_1
```

	longitude	median_income
0	-122.05	6.6085
1	-118.30	3.5990
2	-117.81	5.7934
3	-118.36	6.1359
4	-119.67	2.9375

```
df_2=df[['longitude','median_house_value']][0:5]
df_2
```

	longitude	median_house_value
0	-122.05	344700.0
1	-118.30	176500.0
2	-117.81	270500.0
3	-118.36	330000.0
4	-119.67	81700.0

```
pd.merge(df_1,df_2,on=['longitude'],how='inner')
```

	longitude	median_income	median_house_value
0	-122.05	6.6085	344700.0
1	-118.30	3.5990	176500.0
2	-117.81	5.7934	270500.0
3	-118.36	6.1359	330000.0
4	-119.67	2.9375	81700.0

```
df_3=df[['longitude','population',]][2:7]
df_3
```

	longitude	population
2	-117.81	1484.0
3	-118.36	49.0
4	-119.67	850.0
5	-119.56	663.0
6	-121.43	604.0

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

	longitude	median_income	population
0	-117.81	5.7934	1484.0
1	-118.36	6.1359	49.0
2	-119.67	2.9375	850.0

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

	longitude	median_income	population
0	-122.05	6.6085	NaN
1	-118.30	3.5990	NaN
2	-117.81	5.7934	1484.0
3	-118.36	6.1359	49.0
4	-119.67	2.9375	850.0
5	-119.56	NaN	663.0
6	-121.43	NaN	604.0

Joining

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

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

	longitude
0	-122.05
1	-118.30
2	-117.81
3	-118.36
4	-119.67

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

	latitude
2	33.78
3	33.82
4	36.33
5	36.51
6	38.63

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

	longitude	latitude
0	-122.05	NaN
1	-118.30	NaN
2	-117.81	33.78
3	-118.36	33.82
4	-119.67	36.33

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

	longitude	latitude
2	-117.81	33.78
3	-118.36	33.82
4	-119.67	36.33
5	NaN	36.51
6	NaN	38.63

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

	longitude	latitude
2	-117.81	33.78
3	-118.36	33.82
4	-119.67	36.33

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

	longitude	latitude
0	-122.05	NaN
1	-118.30	NaN
2	-117.81	33.78
3	-118.36	33.82
4	-119.67	36.33
5	NaN	36.51
6	NaN	38.63

Colab paid products - Cancel contracts here