1 Exercícios da biblioteca Pandas

1.1 Utilização da biblioteca pandas

In [5]:

```
1 import pandas as pd
```

executed in 20ms, finished 21:41:37 2021-08-17

1.2 Definição de objeto - Data Frame

Dados importados do Kagle para critérios didáticos

In [6]:

```
1 HR_df = pd.read_csv('input_Files\WA_Fn-UseC_-HR-Employee-Attrition.csv')
```

executed in 513ms, finished 21:41:37 2021-08-17

1.3 Dimensão do Data Frame

Verificação do número de linhasxcolunas da base de dados

In [7]:

```
1 HR_df.shape
```

executed in 17ms, finished 21:41:37 2021-08-17

Out[7]:

```
(1470, 35)
```

1.4 Verificação do tipo das informações contidas nos dados

In [8]:

```
1 HR_df.dtypes
```

executed in 446ms, finished 21:41:38 2021-08-17

Out[8]:

```
Age                         int64
Attrition                  object
BusinessTravel             object
DailyRate                   int64
Department                 object
DistanceFromHome            int64
Education                   int64
EducationField             object
EmployeeCount               int64
EmployeeNumber              int64
EnvironmentSatisfaction     int64
Gender                     object
HourlyRate                  int64
JobInvolvement              int64
JobLevel                    int64
JobRole                    object
JobSatisfaction             int64
MaritalStatus              object
MonthlyIncome               int64
MonthlyRate                 int64
NumCompaniesWorked          int64
Over18                     object
OverTime                   object
PercentSalaryHike           int64
PerformanceRating           int64
RelationshipSatisfaction    int64
StandardHours               int64
StockOptionLevel            int64
TotalWorkingYears           int64
TrainingTimesLastYear       int64
WorkLifeBalance             int64
YearsAtCompany              int64
YearsInCurrentRole          int64
YearsSinceLastPromotion     int64
YearsWithCurrManager        int64
dtype: object
```

1.5 Visualização dos dados

Visualização das 5 primeiras e 5 últimas colunas

In [9]:

```
1 display(HR_df)
```

executed in 217ms, finished 21:41:38 2021-08-17

	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	E
0	41	Yes	Travel_Rarely	1102	Sales	1	
1	49	No	Travel_Frequently	279	Research & Development	8	
2	37	Yes	Travel_Rarely	1373	Research & Development	2	
3	33	No	Travel_Frequently	1392	Research & Development	3	
4	27	No	Travel_Rarely	591	Research & Development	2	
...	...	...	...	...	...	...	
1465	36	No	Travel_Frequently	884	Research & Development	23	
1466	39	No	Travel_Rarely	613	Research & Development	6	
1467	27	No	Travel_Rarely	155	Research & Development	4	
1468	49	No	Travel_Frequently	1023	Sales	2	
1469	34	No	Travel_Rarely	628	Research & Development	8	

1470 rows × 35 columns

1.6 visualizaão de tipo de dados e qualidade dos dados preliminar

A ausência de null date, dispensa-se preliminarmente a fase de tratamento dos dados

In [10]:

```
1 HR_df.info()
```

executed in 202ms, finished 21:41:38 2021-08-17

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1470 entries, 0 to 1469
Data columns (total 35 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Age                       1470 non-null   int64
 1   Attrition                 1470 non-null   object
 2   BusinessTravel            1470 non-null   object
 3   DailyRate                 1470 non-null   int64
 4   Department                1470 non-null   object
 5   DistanceFromHome          1470 non-null   int64
 6   Education                 1470 non-null   int64
 7   EducationField            1470 non-null   object
 8   EmployeeCount             1470 non-null   int64
 9   EmployeeNumber            1470 non-null   int64
 10  EnvironmentSatisfaction   1470 non-null   int64
 11  Gender                    1470 non-null   object
 12  HourlyRate                1470 non-null   int64
 13  JobInvolvement            1470 non-null   int64
 14  JobLevel                  1470 non-null   int64
 15  JobRole                   1470 non-null   object
 16  JobSatisfaction           1470 non-null   int64
 17  MaritalStatus             1470 non-null   object
 18  MonthlyIncome             1470 non-null   int64
 19  MonthlyRate               1470 non-null   int64
 20  NumCompaniesWorked        1470 non-null   int64
 21  Over18                    1470 non-null   object
 22  OverTime                  1470 non-null   object
 23  PercentSalaryHike         1470 non-null   int64
 24  PerformanceRating         1470 non-null   int64
 25  RelationshipSatisfaction  1470 non-null   int64
 26  StandardHours             1470 non-null   int64
 27  StockOptionLevel          1470 non-null   int64
 28  TotalWorkingYears         1470 non-null   int64
 29  TrainingTimesLastYear     1470 non-null   int64
 30  WorkLifeBalance           1470 non-null   int64
 31  YearsAtCompany            1470 non-null   int64
 32  YearsInCurrentRole        1470 non-null   int64
 33  YearsSinceLastPromotion   1470 non-null   int64
 34  YearsWithCurrManager      1470 non-null   int64
dtypes: int64(26), object(9)
memory usage: 402.1+ KB
```

1.7 Análise descritiva preliminar

In [11]:

```
1 HR_df.describe()
```

executed in 203ms, finished 21:41:39 2021-08-17

Out[11]:

	Age	DailyRate	DistanceFromHome	Education	EmployeeCount	Employe
count	1470.000000	1470.000000	1470.000000	1470.000000	1470.0	
mean	36.923810	802.485714	9.192517	2.912925	1.0	
std	9.135373	403.509100	8.106864	1.024165	0.0	
min	18.000000	102.000000	1.000000	1.000000	1.0	
25%	30.000000	465.000000	2.000000	2.000000	1.0	
50%	36.000000	802.000000	7.000000	3.000000	1.0	
75%	43.000000	1157.000000	14.000000	4.000000	1.0	
max	60.000000	1499.000000	29.000000	5.000000	1.0	

8 rows × 26 columns

1.8 Faixa demonstrativa de valores para cada coluna

In [16]:

```
1 for coluna in HR_df: #dale
2     print(HR_df[coluna].value_counts())
3     print('---------x----------- \n')
```

executed in 124ms, finished 21:43:10 2021-08-17

```
     ..
69    15
53    14
68    14
38    13
34    12
Name: HourlyRate, Length: 71, dtype: int64
---------x-----------

3    868
2    375
4    144
1     83
Name: JobInvolvement, dtype: int64
---------x-----------

1    543
2    534
3    218
4    106
```

1.9 Agrupamento de dados por valor específico de

coluna

```
1 Verificação dos dados daqueles funcionários mais satisfeitos com o trabalho
```

In [19]:

```
1 HR_df.groupby('JobSatisfaction').get_group(4)
```

executed in 86ms, finished 21:51:17 2021-08-17

	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	E
0	41	Yes	Travel_Rarely	1102	Sales	1	
5	32	No	Travel_Frequently	1005	Research & Development	2	
13	34	No	Travel_Rarely	1346	Research & Development	19	
17	22	No	Non-Travel	1123	Research & Development	16	
18	53	No	Travel_Rarely	1219	Sales	2	
...	...	...	...	...	...	...	
1451	38	No	Travel_Rarely	345	Sales	10	
1453	36	No	Travel_Rarely	1120	Sales	11	
1458	35	No	Travel_Rarely	287	Research & Development	1	
1462	39	No	Travel_Rarely	722	Sales	24	
1465	36	No	Travel_Frequently	884	Research & Development	23	

459 rows × 35 columns

1.10 Descrição dos dados agrupados

Análise sob todas as colunas, para insights sobre o que ganrante maior satisfação com o trabalho com apoio da análise descritiva preliminar.

In [25]:

```
with pd.option_context('display.max_rows', None, 'display.max_columns', None):
    display(HR_df.groupby('JobSatisfaction').get_group(4).describe())
```

executed in 145ms, finished 22:07:04 2021-08-17

	MonthlyRate	NumCompaniesWorked	PercentSalaryHike	PerformanceRating	Relationshi
)	459.000000	459.000000	459.000000	459.000000	
5	14103.429194	2.516340	15.440087	3.172113	
3	6942.000178	2.390222	3.775091	0.377891	
)	2094.000000	0.000000	11.000000	3.000000	
)	7790.500000	1.000000	12.500000	3.000000	
)	14075.000000	1.000000	14.000000	3.000000	
)	20140.000000	4.000000	18.000000	3.000000	
)	26968.000000	9.000000	25.000000	4.000000	