Thomas Koutsidis

Assignment 3

In [1]:
```
import pandas as pd
```

In [2]:
```
csv_df = pd.read_csv("employees.csv")
```

In [3]:
```
csv_df
```

Out[3]:

	First Name	Gender	Start Date	Last Login Time	Salary	Bonus %	Senior Management	Team
0	Douglas	Male	8/6/1993	12:42 PM	97308	6.945	True	Marketing
1	Thomas	Male	3/31/1996	6:53 AM	61933	4.170	True	NaN
2	Maria	Female	4/23/1993	11:17 AM	130590	11.858	False	Finance
3	Jerry	Male	3/4/2005	1:00 PM	138705	9.340	True	Finance
4	Larry	Male	1/24/1998	4:47 PM	101004	1.389	True	Client Services
...	...	...	...	...	...	...	...	...
995	Henry	NaN	11/23/2014	6:09 AM	132483	16.655	False	Distribution
996	Phillip	Male	1/31/1984	6:30 AM	42392	19.675	False	Finance
997	Russell	Male	5/20/2013	12:39 PM	96914	1.421	False	Product
998	Larry	Male	4/20/2013	4:45 PM	60500	11.985	False	Business Development
999	Albert	Male	5/15/2012	6:24 PM	129949	10.169	True	Sales

1000 rows × 8 columns

In [4]:
```
noNaN_df = csv_df.dropna()
```

In [5]:
```
noNaN_df
```

Out[5]:

	First Name	Gender	Start Date	Last Login Time	Salary	Bonus %	Senior Management	Team
0	Douglas	Male	8/6/1993	12:42 PM	97308	6.945	True	Marketing
2	Maria	Female	4/23/1993	11:17 AM	130590	11.858	False	Finance
3	Jerry	Male	3/4/2005	1:00 PM	138705	9.340	True	Finance
4	Larry	Male	1/24/1998	4:47 PM	101004	1.389	True	Client Services
5	Dennis	Male	4/18/1987	1:35 AM	115163	10.125	False	Legal
...	...	...	...	...	...	...	...	...
994	George	Male	6/21/2013	5:47 PM	98874	4.479	True	Marketing
996	Phillip	Male	1/31/1984	6:30 AM	42392	19.675	False	Finance
997	Russell	Male	5/20/2013	12:39 PM	96914	1.421	False	Product
998	Larry	Male	4/20/2013	4:45 PM	60500	11.985	False	Business Development
999	Albert	Male	5/15/2012	6:24 PM	129949	10.169	True	Sales

764 rows × 8 columns

In [6]:
```
duplicated_df = noNaN_df[noNaN_df.duplicated()]
```

In [7]:
```
duplicated_df
```

Out[7]:

First Name	Gender	Start Date	Last Login Time	Salary	Bonus %	Senior Management	Team

In [8]:
```
dummy_matrix = pd.get_dummies(csv_df["Team"].tail(20))
```

In [9]:
```
dummy_matrix
```

Out[9]:

	Business Development	Client Services	Distribution	Engineering	Finance	Human Resources	Legal	Marketing	Product	Sales
980	False	False	False	True	False	False	False	False	False	False
981	False	False	False	False	False	False	True	False	False	False
982	False	False	False	False	False	True	False	False	False	False
983	False	False	False	True	False	False	False	False	False	False
984	False	False	False	True	False	False	False	False	False	False
985	False	False	False	False	False	False	True	False	False	False
986	False	False	False	False	False	False	False	True	False	False
987	False	False	False	False	True	False	False	False	False	False
988	False	False	False	False	False	True	False	False	False	False
989	False	False	False	False	False	False	True	False	False	False
990	False	True	False	False	False	False	False	False	False	False
991	False	False	False	False	False	False	False	True	False	False
992	False	False	False	False	True	False	False	False	False	False
993	False	False	False	True	False	False	False	False	False	False
994	False	False	False	False	False	False	False	True	False	False
995	False	False	True	False	False	False	False	False	False	False
996	False	False	False	False	True	False	False	False	False	False
997	False	False	False	False	False	False	False	False	True	False
998	True	False	False	False	False	False	False	False	False	False
999	False	False	False	False	False	False	False	False	False	True

In [10]:
```
unique_values = csv_df['Team'].unique()
```

In [11]:
```
unique_values
```

Out[11]:
```
array(['Marketing', nan, 'Finance', 'Client Services', 'Legal', 'Product',
       'Engineering', 'Business Development', 'Human Resources', 'Sales',
       'Distribution'], dtype=object)
```

In [12]:
```
mapping = {
    'Marketing': 'Building1',
    'Client Services': 'Building1',
    'Product': 'Building1',
    'Sales': 'Building1',
    'Finance': 'Building2',
    'Legal': 'Building2',
    'Human Resources': 'Building2',
    'Engineering': 'Building3',
    'Business Development': 'Building3',
    'Distribution': 'Building4',
}
```

In [13]:
```
csv_df["location"] = csv_df["Team"].map(mapping).fillna("Administration")
```

In [14]:
```
csv_df
```

Out[14]:

	First Name	Gender	Start Date	Last Login Time	Salary	Bonus %	Senior Management	Team	location
0	Douglas	Male	8/6/1993	12:42 PM	97308	6.945	True	Marketing	Building1
1	Thomas	Male	3/31/1996	6:53 AM	61933	4.170	True	NaN	Administration
2	Maria	Female	4/23/1993	11:17 AM	130590	11.858	False	Finance	Building2
3	Jerry	Male	3/4/2005	1:00 PM	138705	9.340	True	Finance	Building2
4	Larry	Male	1/24/1998	4:47 PM	101004	1.389	True	Client Services	Building1
...	...	...	...	...	...	...	...	...	...
995	Henry	NaN	11/23/2014	6:09 AM	132483	16.655	False	Distribution	Building4
996	Phillip	Male	1/31/1984	6:30 AM	42392	19.675	False	Finance	Building2
997	Russell	Male	5/20/2013	12:39 PM	96914	1.421	False	Product	Building1
998	Larry	Male	4/20/2013	4:45 PM	60500	11.985	False	Business Development	Building3
999	Albert	Male	5/15/2012	6:24 PM	129949	10.169	True	Sales	Building1

1000 rows × 9 columns

In [15]:
```
print("min. salary =", csv_df["Salary"].min())
print("max. salary =", csv_df["Salary"].max())
```

```
min. salary = 35013
max. salary = 149908
```

In [16]:
```
bins = [35013, 63736, 92460, 121184, 149908]
groups = pd.cut(csv_df["Salary"], bins)
```

In [17]:
```
groups
```

Out[17]:
```
0       (92460, 121184]
1       (35013, 63736]
2      (121184, 149908]
3      (121184, 149908]
4       (92460, 121184]
             ...
995    (121184, 149908]
996     (35013, 63736]
997     (92460, 121184]
998     (35013, 63736]
999    (121184, 149908]
Name: Salary, Length: 1000, dtype: category
Categories (4, interval[int64, right]): [(35013, 63736] < (63736, 92460] < (92460, 121184] < (121184, 149908]]
```

In [18]:
```
groups.value_counts()
```

Out[18]:
```
Salary
(35013, 63736]      261
(63736, 92460]      261
(92460, 121184]     245
(121184, 149908]    232
Name: count, dtype: int64
```

In [19]:
```
bonus_outliers = csv_df[csv_df['Bonus %'] > 9.3]
print(bonus_outliers)
```

```
     First Name  Gender  Start Date Last Login Time  Salary  Bonus %  \
2         Maria  Female   4/23/1993       11:17 AM  130590   11.858
3         Jerry    Male    3/4/2005        1:00 PM  138705    9.340
5       Dennis    Male   4/18/1987        1:35 AM  115163   10.125
6          Ruby  Female   8/17/1987        4:20 PM   65476   10.012
7           NaN  Female   7/20/2015       10:43 AM   45906   11.598
..          ...     ...         ...            ...     ...      ...
993        Tina  Female   5/15/1997        3:53 PM   56450   19.040
995       Henry     NaN  11/23/2014        6:09 AM  132483   16.655
996     Phillip    Male   1/31/1984        6:30 AM   42392   19.675
998       Larry    Male   4/20/2013        4:45 PM   60500   11.985
999      Albert    Male   5/15/2012        6:24 PM  129949   10.169

    Senior Management                Team   location
2               False             Finance  Building2
3                True             Finance  Building2
5               False               Legal  Building2
6                True             Product  Building1
7                 NaN             Finance  Building2
..                ...                 ...        ...
993              True         Engineering  Building3
995             False        Distribution  Building4
996             False             Finance  Building2
998             False  Business Development  Building3
999              True               Sales  Building1

[536 rows x 9 columns]
```

In [20]:
```
boolean_indexing = csv_df['First Name'].fillna('').str.startswith('S')

firstname_df = csv_df[boolean_indexing]
```

In [21]:
```
firstname_df
```

Out[21]:

	First Name	Gender	Start Date	Last Login Time	Salary	Bonus %	Senior Management	Team	location
17	Shawn	Male	12/7/1986	7:45 PM	111737	6.414	False	Product	Building1
27	Scott	NaN	7/11/1991	6:58 PM	122367	5.218	False	Legal	Building2
38	Stephanie	Female	9/13/1986	1:52 AM	36844	5.574	True	Business Development	Building3
54	Sara	Female	8/15/2007	9:23 AM	83677	8.999	False	Engineering	Building3
65	Steve	Male	11/11/2009	11:44 PM	61310	12.428	True	Distribution	Building4
...	...	...	...	...	...	...	...	...	...
960	Stephen	Male	10/29/1989	11:34 PM	93997	18.093	True	Business Development	Building3
975	Susan	Female	4/7/1995	10:05 PM	92436	12.467	False	Sales	Building1
977	Sarah	Female	12/4/1995	9:16 AM	124566	5.949	False	Product	Building1
978	Sean	Male	1/17/1983	2:23 PM	66146	11.178	False	Human Resources	Building2
985	Stephen	NaN	7/10/1983	8:10 PM	85668	1.909	False	Legal	Building2

80 rows × 9 columns

In [22]:
```
data = csv_df.iloc[:10, :3]
hierarchical_indexing = pd.MultiIndex.from_product([range(len(data)), data.columns])
hierarchical_series = pd.Series(data.values.flatten(), index = hierarchical_indexing)
print(hierarchical_series)
```

```
0  First Name         Douglas
   Gender                Male
   Start Date       8/6/1993
1  First Name          Thomas
   Gender                Male
   Start Date      3/31/1996
2  First Name           Maria
   Gender              Female
   Start Date      4/23/1993
3  First Name           Jerry
   Gender                Male
   Start Date       3/4/2005
4  First Name           Larry
   Gender                Male
   Start Date      1/24/1998
5  First Name          Dennis
   Gender                Male
   Start Date      4/18/1987
6  First Name            Ruby
   Gender              Female
   Start Date      8/17/1987
7  First Name             NaN
   Gender              Female
   Start Date      7/20/2015
8  First Name          Angela
   Gender              Female
   Start Date     11/22/2005
9  First Name          Frances
   Gender              Female
   Start Date       8/8/2002
dtype: object
```

In [23]:
```
averages = csv_df.groupby('Team')[['Salary', 'Bonus %']].mean()
print(averages)
```

```
                            Salary    Bonus %
Team
Business Development  91866.316832  10.572376
Client Services       88224.424528  10.495104
Distribution          88500.466667   9.615644
Engineering           94269.195652  10.462989
Finance               92218.480392  10.186873
Human Resources       90944.527473   9.993879
Legal                 89303.613636  10.322830
Marketing             90435.591837  10.353449
Product               88665.505263   9.791484
Sales                 92173.436170  10.116915
```

In [24]:
```
def top(dataframe):
    return dataframe.nlargest(5, 'Salary')

paid = top(csv_df)
print("Five Highest Paid:\n", paid)
```

```
Five Highest Paid:
     First Name  Gender Start Date Last Login Time  Salary  Bonus %  \
644  Katherine  Female  8/13/1996       12:21 AM  149908   10.912
429       Rose  Female  5/28/2015        8:40 AM  149903    5.630
828   Cynthia  Female  7/12/2006        8:55 AM  149684    7.864
186       NaN  Female  2/23/2005        9:50 PM  149654    1.825
160     Kathy  Female  3/18/2000        7:26 PM  149563   16.991

    Senior Management            Team   location
644             False         Finance  Building2
429             False   Human Resources  Building2
828             False         Product  Building1
186               NaN           Sales  Building1
160              True         Finance  Building2
```

In [25]:
```
genders = csv_df.groupby('Gender')

for i, j in genders:
    print(f"Five Highest Paid {i}:")
    genders_paid = top(j)
    print(genders_paid, "\n")
```

```
Five Highest Paid Female:
     First Name  Gender Start Date Last Login Time  Salary  Bonus %  \
644  Katherine  Female  8/13/1996       12:21 AM  149908   10.912
429       Rose  Female  5/28/2015        8:40 AM  149903    5.630
828   Cynthia  Female  7/12/2006        8:55 AM  149684    7.864
186       NaN  Female  2/23/2005        9:50 PM  149654    1.825
160     Kathy  Female  3/18/2000        7:26 PM  149563   16.991

    Senior Management            Team   location
644             False         Finance  Building2
429             False   Human Resources  Building2
828             False         Product  Building1
186               NaN           Sales  Building1
160              True         Finance  Building2

Five Highest Paid Male:
     First Name Gender Start Date Last Login Time  Salary  Bonus %  \
981      James   Male  1/15/1993        5:19 PM  149885   19.288
800   Clarence   Male   8/5/1989        6:11 PM  148941   11.517
858    Charles   Male   9/3/1997       10:04 AM  148281    6.002
318        Roy   Male   8/6/2006        7:52 AM  148225    1.841
83       Shawn   Male   9/23/2005        2:55 AM  148115    6.539

    Senior Management      Team   location
981             False   Legal  Building2
800              True  Product  Building1
858             False      NaN  Administration
318              True  Finance  Building2
83               True  Finance  Building2
```