In [1]:

```
import pandas as pd
```

In [3]:

```
df=pd.read_csv("chicago.csv")
df["Department"]=df["Department"].astype("category")#optimisation
df.head()
```

In [6]:

```
df.head()
```

Out[6]:

	Name	Position Title	Department	Employee Annual Salary
0	AARON, ELVIA J	WATER RATE TAKER	WATER MGMNT	$90744.00
1	AARON, JEFFERY M	POLICE OFFICER	POLICE	$84450.00
2	AARON, KARINA	POLICE OFFICER	POLICE	$84450.00
3	AARON, KIMBERLEI R	CHIEF CONTRACT EXPEDITER	GENERAL SERVICES	$89880.00
4	ABAD JR, VICENTE M	CIVIL ENGINEER IV	WATER MGMNT	$106836.00

In [12]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32063 entries, 0 to 32062
Data columns (total 4 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   Name                   32062 non-null  object
 1   Position Title         32062 non-null  object
 2   Department             32062 non-null  object
 3   Employee Annual Salary  32062 non-null  object
dtypes: object(4)
memory usage: 1002.1+ KB
```

In [15]:

```
len(df["Department"].unique())
```

Out[15]:

```
36
```

In [19]:

```
df["Department"].nunique()
```

Out[19]:

```
35
```

In [20]:

```
df["Department"].nunique(dropna=False)
```

Out[20]:

```
36
```

In [22]:

```
df["Department"]=df["Department"].astype("category")#optimisation
```

In [23]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32063 entries, 0 to 32062
Data columns (total 4 columns):
 #   Column                  Non-Null Count  Dtype   
---  ------                  --------------  -----   
 0   Name                    32062 non-null  object  
 1   Position Title          32062 non-null  object  
 2   Department              32062 non-null  category
 3   Employee Annual Salary  32062 non-null  object  
dtypes: category(1), object(3)
memory usage: 784.2+ KB
```

In [24]:

```
df.nunique()
```

Out[24]:

```
Name                     31776
Position Title            1093
Department                  35
Employee Annual Salary    1156
dtype: int64
```

Common string methods loweer, upper, title, len

In [25]:

```
"HELLO WORLD".lower()
```

Out[25]:

```
'hello world'
```

In [26]:

```
"hello world".upper()
```

Out[26]:

```
'HELLO WORLD'
```

In [27]:

```
"hello world".title()
```

Out[27]:

```
'Hello World'
```

In [28]:

```
len("hello world")
```

Out[28]:

```
11
```

In [30]:

```
(df["Name"].str.lower()).head()
```

Out[30]:

```
0        aaron,  elvia j
1      aaron,  jeffery m
2         aaron,  karina
3    aaron,  kimberlei r
4    abad jr,  vicente m
Name: Name, dtype: object
```

In [32]:

```
(df["Name"].str.title()).head()
```

Out[32]:

```
0        Aaron,  Elvia J
1      Aaron,  Jeffery M
2         Aaron,  Karina
3    Aaron,  Kimberlei R
4    Abad Jr,  Vicente M
Name: Name, dtype: object
```

In [34]:

```
(df["Department"].str.len()).head()
```

Out[34]:

```
0    11.0
1     6.0
2     6.0
3    16.0
4    11.0
Name: Department, dtype: float64
```

.srt.replace()

In [56]:

```
df=pd.read_csv("chicago.csv")
df["Department"]=df["Department"].astype("category")#optimisation
df.head()
```

Out[56]:

	Name	Position Title	Department	Employee Annual Salary
0	AARON, ELVIA J	WATER RATE TAKER	WATER MGMNT	$90744.00
1	AARON, JEFFERY M	POLICE OFFICER	POLICE	$84450.00
2	AARON, KARINA	POLICE OFFICER	POLICE	$84450.00
3	AARON, KIMBERLEI R	CHIEF CONTRACT EXPEDITER	GENERAL SERVICES	$89880.00
4	ABAD JR, VICENTE M	CIVIL ENGINEER IV	WATER MGMNT	$106836.00

In [57]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32063 entries, 0 to 32062
Data columns (total 4 columns):
 #   Column                 Non-Null Count  Dtype   
---  ------                 --------------  -----   
 0   Name                   32062 non-null  object  
 1   Position Title         32062 non-null  object  
 2   Department             32062 non-null  category
 3   Employee Annual Salary  32062 non-null  object  
dtypes: category(1), object(3)
memory usage: 784.2+ KB
```

In [60]:

```
df["Department"]=df["Department"].str.replace("MGMNT","MANAGEMENT")#replace department
df.head()
```

Out[60]:

	Name	Position Title	Department	Employee Annual Salary
0	AARON, ELVIA J	WATER RATE TAKER	WATER MANAGEMENT	$90744.00
1	AARON, JEFFERY M	POLICE OFFICER	POLICE	$84450.00
2	AARON, KARINA	POLICE OFFICER	POLICE	$84450.00
3	AARON, KIMBERLEI R	CHIEF CONTRACT EXPEDITER	GENERAL SERVICES	$89880.00
4	ABAD JR, VICENTE M	CIVIL ENGINEER IV	WATER MANAGEMENT	$106836.00

In [61]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32063 entries, 0 to 32062
Data columns (total 4 columns):
 #   Column                 Non-Null Count  Dtype 
---  ------                 --------------  ----- 
 0   Name                   32062 non-null  object
 1   Position Title         32062 non-null  object
 2   Department             32062 non-null  object
 3   Employee Annual Salary  32062 non-null  object
dtypes: object(4)
memory usage: 1002.1+ KB
```

In [67]:

```
df["Employee Annual Salary"]=df["Employee Annual Salary"].str.replace("$","")#remove dollar
```

```
<ipython-input-67-8c74eee03de5>:1: FutureWarning: The default value of regex
will change from True to False in a future version. In addition, single char
acter regular expressions will *not* be treated as literal strings when rege
x=True.
  df["Employee Annual Salary"]=df["Employee Annual Salary"].str.replace
("$","")
```

In [68]:

```
df.head()
```

Out[68]:

	Name	Position Title	Department	Employee Annual Salary
0	AARON, ELVIA J	WATER RATE TAKER	WATER MANAGEMENT	90744.00
1	AARON, JEFFERY M	POLICE OFFICER	POLICE	84450.00
2	AARON, KARINA	POLICE OFFICER	POLICE	84450.00
3	AARON, KIMBERLEI R	CHIEF CONTRACT EXPEDITER	GENERAL SERVICES	89880.00
4	ABAD JR, VICENTE M	CIVIL ENGINEER IV	WATER MANAGEMENT	106836.00

In [69]:

```
df["Employee Annual Salary"]=df["Employee Annual Salary"].astype(float)
```

In [70]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32063 entries, 0 to 32062
Data columns (total 4 columns):
 #   Column                 Non-Null Count  Dtype  
---  ------                 --------------  -----  
 0   Name                   32062 non-null  object 
 1   Position Title         32062 non-null  object 
 2   Department             32062 non-null  object 
 3   Employee Annual Salary  32062 non-null  float64
dtypes: float64(1), object(3)
memory usage: 1002.1+ KB
```

In [71]:

```
df["Employee Annual Salary"].nlargest(3)#only particular column
```

Out[71]:

```
8184     300000.0
7954     216210.0
25532    202728.0
Name: Employee Annual Salary, dtype: float64
```

In [73]:

```
df.nlargest(3,"Employee Annual Salary")#entire DF
```

Out[73]:

	Name	Position Title	Department	Employee Annual Salary
8184	EVANS, GINGER S	COMMISSIONER OF AVIATION	AVIATION	300000.0
7954	EMANUEL, RAHM	MAYOR	MAYOR'S OFFICE	216210.0
25532	SANTIAGO, JOSE A	FIRE COMMISSIONER	FIRE	202728.0

filter using string methods

In [74]:

```
df=pd.read_csv("chicago.csv").dropna(how="all")
df["Department"]=df["Department"].astype("category")
df.tail()
```

Out[74]:

	Name	Position Title	Department	Employee Annual Salary
32057	ZYGADLO, MICHAEL J	FRM OF MACHINISTS - AUTOMOTIVE	GENERAL SERVICES	$99528.00
32058	ZYGOWICZ, PETER J	POLICE OFFICER	POLICE	$87384.00
32059	ZYMANTAS, MARK E	POLICE OFFICER	POLICE	$84450.00
32060	ZYRKOWSKI, CARLO E	POLICE OFFICER	POLICE	$87384.00
32061	ZYSKOWSKI, DARIUSZ	CHIEF DATA BASE ANALYST	DoIT	$113664.00

In [76]:

```
df["Position Title"].head()
```

Out[76]:

```
0            WATER RATE TAKER
1              POLICE OFFICER
2              POLICE OFFICER
3    CHIEF CONTRACT EXPEDITER
4           CIVIL ENGINEER IV
Name: Position Title, dtype: object
```

In [86]:

```
mask=df["Department"].str.lower().str.contains("water")#contains
df[mask].head()
```

Out[86]:

	Name	Position Title	Department	Employee Annual Salary
0	AARON, ELVIA J	WATER RATE TAKER	WATER MGMNT	$90744.00
4	ABAD JR, VICENTE M	CIVIL ENGINEER IV	WATER MGMNT	$106836.00
20	ABDUL-KARIM, MUHAMMAD A	ENGINEERING TECHNICIAN VI	WATER MGMNT	$108228.00
25	ABDULSATTAR, MUDHAR	CIVIL ENGINEER II	WATER MGMNT	$58536.00
34	ABRAHAM, GIRLEY T	CIVIL ENGINEER IV	WATER MGMNT	$106836.00

In [89]:

```
(df[df["Position Title"].str.lower().str.startswith("laborer")]).head()#startswith
```

Out[89]:

	Name	Position Title	Department	Employee Annual Salary
168	ADEWOLE, KAREEM A	LABORER - APPRENTICE	WATER MGMNT	$73382.40
250	AHMAD, FAROOQ	LABORER	AVIATION	$69825.60
367	ALEXANDER, CLEMMIE	LABORER	TRANSPORTN	$81536.00
411	ALICEA JR, FELIX	LABORER	AVIATION	$69825.60
819	ANKUM, CHARMAIN D	LABORER - APPRENTICE	WATER MGMNT	$73382.40

In [88]:

```
(df[df["Position Title"].str.lower().str.endswith("ist")]).head()#startswith
```

Out[88]:

	Name	Position Title	Department	Employee Annual Salary
184	AFROZ, NAYYAR	PSYCHIATRIST	HEALTH	$99840.00
308	ALARCON, LUIS J	LOAN PROCESSING SPECIALIST	COMMUNITY DEVELOPMENT	$81948.00
422	ALLAIN, CAROLYN	SENIOR TELECOMMUNICATIONS SPECIALIST	DoIT	$89880.00
472	ALLEN, ROBERT	MACHINIST	WATER MGMNT	$94328.00
705	ANDERSON, EDWARD M	SR PROCUREMENT SPECIALIST	PROCUREMENT	$91476.00

lstrip, rstrip, strip methods

In [90]:

```
df=pd.read_csv("chicago.csv").dropna(how="all")
df["Department"]=df["Department"].astype("category")
df.tail()
```

Out[90]:

	Name	Position Title	Department	Employee Annual Salary
32057	ZYGADLO, MICHAEL J	FRM OF MACHINISTS - AUTOMOTIVE	GENERAL SERVICES	$99528.00
32058	ZYGOWICZ, PETER J	POLICE OFFICER	POLICE	$87384.00
32059	ZYMANTAS, MARK E	POLICE OFFICER	POLICE	$84450.00
32060	ZYRKOWSKI, CARLO E	POLICE OFFICER	POLICE	$87384.00
32061	ZYSKOWSKI, DARIUSZ	CHIEF DATA BASE ANALYST	DoIT	$113664.00

In [91]:

```
word="    hello  "
```

In [95]:

```
word.strip()
```

Out[95]:

```
'hello'
```

In [98]:

```
df["Name"]=df["Name"].str.lstrip().str.rstrip()
```

In [99]:

```
df["Position Title"]=df["Position Title"].str.strip()
```

Srting methods on index and columns

In [133]:

```
df=pd.read_csv("chicago.csv",index_col="Name").dropna(how="all")
df["Department"]=df["Department"].astype("category")
df.tail()
```

Out[133]:

Name	Position Title	Department	Employee Annual Salary
ZYGADLO, MICHAEL J	FRM OF MACHINISTS - AUTOMOTIVE	GENERAL SERVICES	$99528.00
ZYGOWICZ, PETER J	POLICE OFFICER	POLICE	$87384.00
ZYMANTAS, MARK E	POLICE OFFICER	POLICE	$84450.00
ZYRKOWSKI, CARLO E	POLICE OFFICER	POLICE	$87384.00
ZYSKOWSKI, DARIUSZ	CHIEF DATA BASE ANALYST	DoIT	$113664.00

In [137]:

```
df.index
head()
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
<ipython-input-137-17dd73253cb4> in <module>
      1 df.index
----> 2 head()

NameError: name 'head' is not defined
```

In [104]:

```
df.index=df.index.str.strip().str.title()
```

In [105]:

```
df.head()
```

Out[105]:

Name	Position Title	Department	Employee Annual Salary
Aaron, Elvia J	WATER RATE TAKER	WATER MGMNT	$90744.00
Aaron, Jeffery M	POLICE OFFICER	POLICE	$84450.00
Aaron, Karina	POLICE OFFICER	POLICE	$84450.00
Aaron, Kimberlei R	CHIEF CONTRACT EXPEDITER	GENERAL SERVICES	$89880.00
Abad Jr, Vicente M	CIVIL ENGINEER IV	WATER MGMNT	$106836.00

In [106]:

```
df.columns
```

Out[106]:

```
Index(['Position Title', 'Department', 'Employee Annual Salary'], dtype='obj
ect')
```

In [108]:

```
df.columns=df.columns.str.upper()
```

In [109]:

```
df.head()
```

Out[109]:

Name	POSITION TITLE	DEPARTMENT	EMPLOYEE ANNUAL SALARY
Aaron, Elvia J	WATER RATE TAKER	WATER MGMNT	$90744.00
Aaron, Jeffery M	POLICE OFFICER	POLICE	$84450.00
Aaron, Karina	POLICE OFFICER	POLICE	$84450.00
Aaron, Kimberlei R	CHIEF CONTRACT EXPEDITER	GENERAL SERVICES	$89880.00
Abad Jr, Vicente M	CIVIL ENGINEER IV	WATER MGMNT	$106836.00

split method

In [110]:

```
df=pd.read_csv("chicago.csv").dropna(how="all")
df["Department"]=df["Department"].astype("category")
df.tail()
```

Out[110]:

	Name	Position Title	Department	Employee Annual Salary
32057	ZYGADLO, MICHAEL J	FRM OF MACHINISTS - AUTOMOTIVE	GENERAL SERVICES	$99528.00
32058	ZYGOWICZ, PETER J	POLICE OFFICER	POLICE	$87384.00
32059	ZYMANTAS, MARK E	POLICE OFFICER	POLICE	$84450.00
32060	ZYRKOWSKI, CARLO E	POLICE OFFICER	POLICE	$87384.00
32061	ZYSKOWSKI, DARIUSZ	CHIEF DATA BASE ANALYST	DoIT	$113664.00

In [111]:

```
"Hello how are you".split(" ")
```

Out[111]:

```
['Hello', 'how', 'are', 'you']
```

In [114]:

```
(df["Name"]).head()
```

Out[114]:

```
0        AARON,  ELVIA J
1      AARON,  JEFFERY M
2         AARON,  KARINA
3    AARON,  KIMBERLEI R
4    ABAD JR,  VICENTE M
Name: Name, dtype: object
```

In [130]:

```
df["Name"].str.split(",").str.get(0).str.title().value_counts().head()
```

Out[130]:

```
Williams    293
Johnson     244
Smith       241
Brown       185
Jones       183
Name: Name, dtype: int64
```

In [129]:

```
df["Position Title"].str.split(" ").str.get(0).value_counts().head()
```

Out[129]:

```
POLICE             10856
FIREFIGHTER-EMT     1509
SERGEANT            1186
POOL                 918
FIREFIGHTER          810
Name: Position Title, dtype: int64
```

More practice on split methods

In [138]:

```
df=pd.read_csv("chicago.csv").dropna(how="all")
df["Department"]=df["Department"].astype("category")
df.head()
```

Out[138]:

	Name	Position Title	Department	Employee Annual Salary
0	AARON, ELVIA J	WATER RATE TAKER	WATER MGMNT	$90744.00
1	AARON, JEFFERY M	POLICE OFFICER	POLICE	$84450.00
2	AARON, KARINA	POLICE OFFICER	POLICE	$84450.00
3	AARON, KIMBERLEI R	CHIEF CONTRACT EXPEDITER	GENERAL SERVICES	$89880.00
4	ABAD JR, VICENTE M	CIVIL ENGINEER IV	WATER MGMNT	$106836.00

In [141]:

```
df["Name"].str.split(",").str.get(0).value_counts().head()
```

Out[141]:

```
WILLIAMS    293
JOHNSON     244
SMITH       241
BROWN       185
JONES       183
Name: Name, dtype: int64
```

In [154]:

```
df["Name"].str.split(",").str.get(1).str.strip(" ").str.split(" ").str.get(0).head()#first
```

Out[154]:

```
0        ELVIA
1      JEFFERY
2       KARINA
3    KIMBERLEI
4      VICENTE
Name: Name, dtype: object
```

In [155]:

```
df["Name"].str.split(",").str.get(1).str.strip(" ").str.split(" ").str.get(0).value_counts(
```

Out[155]:

```
MICHAEL    1153
JOHN        899
JAMES       676
ROBERT      622
JOSEPH      537
Name: Name, dtype: int64
```

Advanced split -expand param, n splits

In [156]:

```
df=pd.read_csv("chicago.csv").dropna(how="all")
df["Department"]=df["Department"].astype("category")
df.head()
```

Out[156]:

	Name	Position Title	Department	Employee Annual Salary
0	AARON, ELVIA J	WATER RATE TAKER	WATER MGMNT	$90744.00
1	AARON, JEFFERY M	POLICE OFFICER	POLICE	$84450.00
2	AARON, KARINA	POLICE OFFICER	POLICE	$84450.00
3	AARON, KIMBERLEI R	CHIEF CONTRACT EXPEDITER	GENERAL SERVICES	$89880.00
4	ABAD JR, VICENTE M	CIVIL ENGINEER IV	WATER MGMNT	$106836.00

In [162]:

```
df["Name"].str.split(",",expand=True).head()##Data frame first and last name in two clms
```

Out[162]:

	0	1
0	AARON	ELVIA J
1	AARON	JEFFERY M
2	AARON	KARINA
3	AARON	KIMBERLEI R
4	ABAD JR	VICENTE M

In [163]:

```
df[["First Name","Last Name"]]=df["Name"].str.split(",",expand=True)##Data frame first and
```

In [164]:

```
df.head()
```

Out[164]:

	Name	Position Title	Department	Employee Annual Salary	First Name	Last Name
0	AARON, ELVIA J	WATER RATE TAKER	WATER MGMNT	$90744.00	AARON	ELVIA J
1	AARON, JEFFERY M	POLICE OFFICER	POLICE	$84450.00	AARON	JEFFERY M
2	AARON, KARINA	POLICE OFFICER	POLICE	$84450.00	AARON	KARINA
3	AARON, KIMBERLEI R	CHIEF CONTRACT EXPEDITER	GENERAL SERVICES	$89880.00	AARON	KIMBERLEI R
4	ABAD JR, VICENTE M	CIVIL ENGINEER IV	WATER MGMNT	$106836.00	ABAD JR	VICENTE M

In [172]:

```
df[["First title word","remaining words"]]=df["Position Title"].str.strip().str.split(" ",e
```

In [173]:

```
df.head()
```

Out[173]:

	Name	Position Title	Department	Employee Annual Salary	First Name	Last Name	First title word	remaining words
0	AARON, ELVIA J	WATER RATE TAKER	WATER MGMNT	$90744.00	AARON	ELVIA J	WATER	RATE TAKER
1	AARON, JEFFERY M	POLICE OFFICER	POLICE	$84450.00	AARON	JEFFERY M	POLICE	OFFICER
2	AARON, KARINA	POLICE OFFICER	POLICE	$84450.00	AARON	KARINA	POLICE	OFFICER
3	AARON, KIMBERLEI R	CHIEF CONTRACT EXPEDITER	GENERAL SERVICES	$89880.00	AARON	KIMBERLEI R	CHIEF	CONTRACT EXPEDITER
4	ABAD JR, VICENTE M	CIVIL ENGINEER IV	WATER MGMNT	$106836.00	ABAD JR	VICENTE M	CIVIL	ENGINEER IV