```
!gdown "https://d2beiqkhq929f0.cloudfront.net/public_assets/assets/000/002/856/original/scaler_clustering.csv"
```

```
Downloading...
From: https://d2beiqkhq929f0.cloudfront.net/public_assets/assets/000/002/856/original/scaler_clustering.csv
To: /content/scaler_clustering.csv
100% 24.7M/24.7M [00:12<00:00, 1.91MB/s]
```

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

```
df = pd.read_csv('scaler_clustering.csv')
```

```
df.ndim
# The data is 2 dimensional
```

```
2
```

```
df.shape
# The data has 205843 rows with 7 columns (features)
```

```
(205843, 7)
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 205843 entries, 0 to 205842
Data columns (total 7 columns):
 #   Column            Non-Null Count   Dtype  
---  ------            --------------   -----  
 0   Unnamed: 0        205843 non-null  int64  
 1   company_hash      205799 non-null  object 
 2   email_hash        205843 non-null  object 
 3   orgyear           205757 non-null  float64
 4   ctc               205843 non-null  int64  
 5   job_position      153279 non-null  object 
 6   ctc_updated_year  205843 non-null  float64
dtypes: float64(2), int64(2), object(3)
memory usage: 11.0+ MB
```

```
df.describe()
```

	Unnamed: 0	orgyear	ctc	ctc_updated_year
count	205843.000000	205757.000000	2.058430e+05	205843.000000
mean	103273.941786	2014.882750	2.271685e+06	2019.628231
std	59741.306484	63.571115	1.180091e+07	1.325104
min	0.000000	0.000000	2.000000e+00	2015.000000
25%	51518.500000	2013.000000	5.300000e+05	2019.000000
50%	103151.000000	2016.000000	9.500000e+05	2020.000000
75%	154992.500000	2018.000000	1.700000e+06	2021.000000
max	206922.000000	20165.000000	1.000150e+09	2021.000000

```
df.isna().sum()
# we have null values in 3 columns
# There are many null values in the column job_position. 25% (52564/205843 = 25.5%)of this column has null values
```

```
Unnamed: 0              0
company_hash           44
email_hash              0
orgyear                86
ctc                     0
job_position        52564
ctc_updated_year        0
dtype: int64
```

```
len(df[df.duplicated()])
# There are no duplicated rows in the data
```

```
0
```

```
df.columns
```

```
Index(['Unnamed: 0', 'company_hash', 'email_hash', 'orgyear', 'ctc',
       'job_position', 'ctc_updated_year'],
      dtype='object')
```

```
for i in df.columns:
  print(i," - ",df[i].value_counts().values[0])
# we can see the number of unique values for each column
```

```
Unnamed: 0  -  1
company_hash  -  8337
email_hash  -  10
```

```
orgyear  -  25256
ctc  -  7832
job_position  -  43554
ctc_updated_year  -  68688
```

```
df['email_hash'].value_counts()
```

```
email_hash
bbace3cc586400bbc65765bc6a16b77d8913836cfc98b77c05488f02f5714a4b    10
6842660273f70e9aa239026ba33bfe82275d6ab0d20124021b952b5bc3d07e6c     9
298528ce3160cc761e4dc37a07337ee2e0589df251d73645aae209b010210eee     9
3e5e49daa5527a6d5a33599b238bf9bf31e85b9efa9a94f1c88c5e15a6f31378     9
b4d5afa09bec8689017d8b29701b80d664ca37b83cb883376b2e95191320da66     8
                                                                    ..
bb2fe5e655ada7f7b7ac4a614db0b9c560e796bdfcaa4e5367e69eedfea93876     1
d6cdef97e759dbf1b7522babccbbbd5f164a75d1b4139e02c945958720f1ed79     1
700d1190c17aaa3f2dd9070e47a4c042ecd9205333545dbfaee0f85644d00306     1
c2a1c9e4b9f4e1ed7d889ee4560102c1e2235b2c1a0e59cea95a6fe55c658407     1
0bcfc1d05f2e8dc4147743a1313aa70a119b41b30d4a1f7e738a6a87d3712c31     1
Name: count, Length: 153443, dtype: int64
```

```
df[df['email_hash'] == 'bbace3cc586400bbc65765bc6a16b77d8913836cfc98b77c05488f02f5714a4b']
# For same email_hash and company_hash there exists multiple rows
```

	Unnamed: 0	company_hash	email_hash	orgyear	ctc	job_position	ctc_updated_year
24109	24129	oxej ntwyzgrgsxto rxbxnta	bbace3cc586400bbc65765bc6a16b77d8913836cfc98b7...	2018.0	720000	NaN	2020.0
45984	46038	oxej ntwyzgrgsxto rxbxnta	bbace3cc586400bbc65765bc6a16b77d8913836cfc98b7...	2018.0	720000	Support Engineer	2020.0
72315	72415	oxej ntwyzgrgsxto rxbxnta	bbace3cc586400bbc65765bc6a16b77d8913836cfc98b7...	2018.0	720000	Other	2020.0
102915	103145	oxej ntwyzgrgsxto rxbxnta	bbace3cc586400bbc65765bc6a16b77d8913836cfc98b7...	2018.0	720000	FullStack Engineer	2020.0
117764	118076	oxej ntwyzgrgsxto rxbxnta	bbace3cc586400bbc65765bc6a16b77d8913836cfc98b7...	2018.0	720000	Data Analyst	2020.0
121483	121825	oxej ntwyzgrgsxto rxbxnta	bbace3cc586400bbc65765bc6a16b77d8913836cfc98b7...	2018.0	660000	Other	2019.0
124476	124840	oxej ntwyzgrgsxto rxbxnta	bbace3cc586400bbc65765bc6a16b77d8913836cfc98b7...	2018.0	660000	Support Engineer	2019.0
144479	145021	oxej ntwyzgrgsxto rxbxnta	bbace3cc586400bbc65765bc6a16b77d8913836cfc98b7...	2018.0	660000	FullStack Engineer	2019.0
152801	153402	oxej ntwyzgrgsxto rxbxnta	bbace3cc586400bbc65765bc6a16b77d8913836cfc98b7...	2018.0	660000	Devops Engineer	2019.0
159835	160472	oxej ntwyzgrgsxto rxbxnta	bbace3cc586400bbc65765bc6a16b77d8913836cfc98b7...	2018.0	660000	NaN	2019.0

```
df = df.groupby('email_hash').first().reset_index()
```

```
df['YoE'] = df['ctc_updated_year'] - df['orgyear']
```

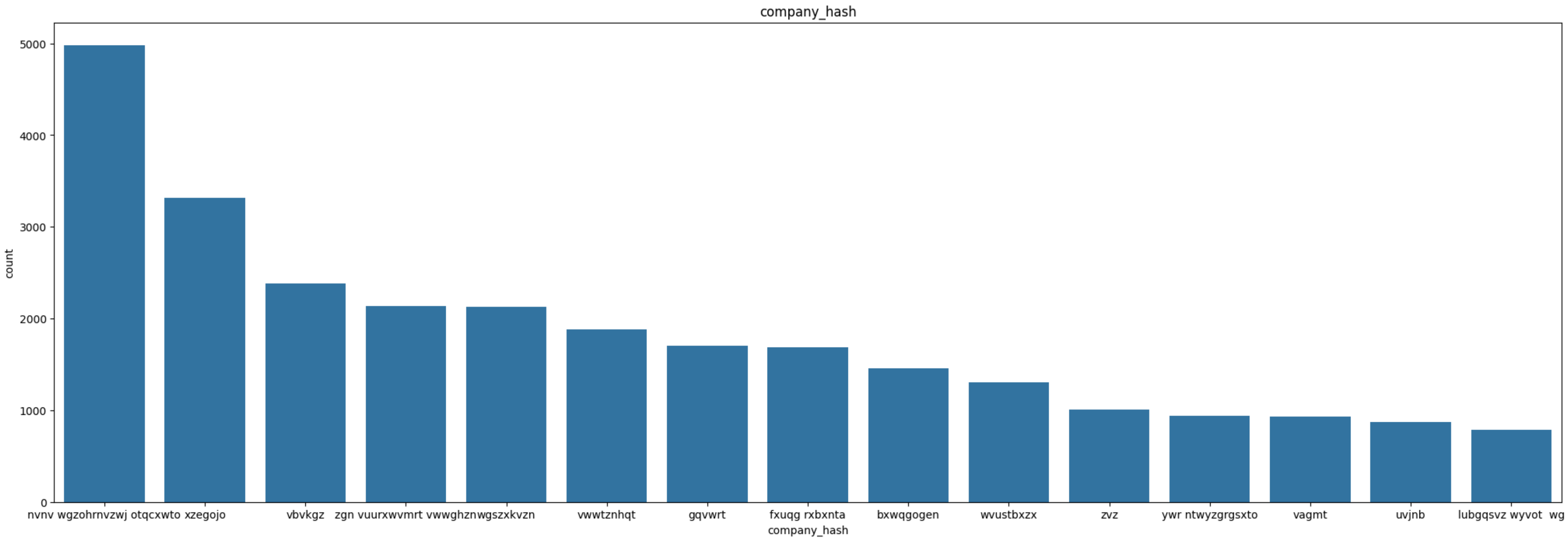
```
feat = 'company_hash'
df[feat] = df[feat].fillna('na')
enc_nom = (df.groupby(feat).size()) / len(df)
df[feat+'_encode'] = df[feat].apply(lambda x : enc_nom[x])

feat = 'job_position'
df[feat] = df[feat].fillna('na')
enc_nom = (df.groupby(feat).size()) / len(df)*10000
df[feat+'_encode'] = df[feat].apply(lambda x : enc_nom[x])
```

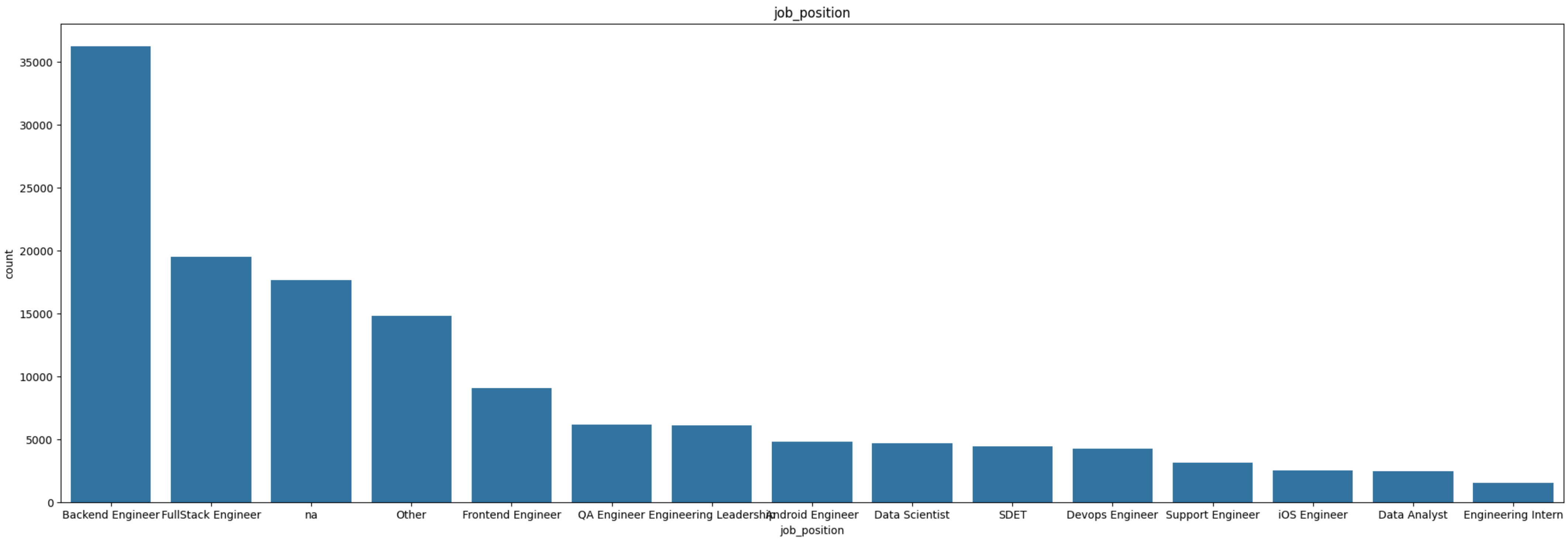
```
data = df[~df['orgyear'].isin( sorted(df['orgyear'].fillna(0).astype(int).unique()) )]
# removing outliers from orgyear column
```

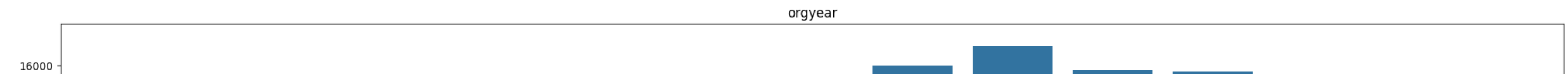
```
df = df[~(df['YoE']<0)]
```

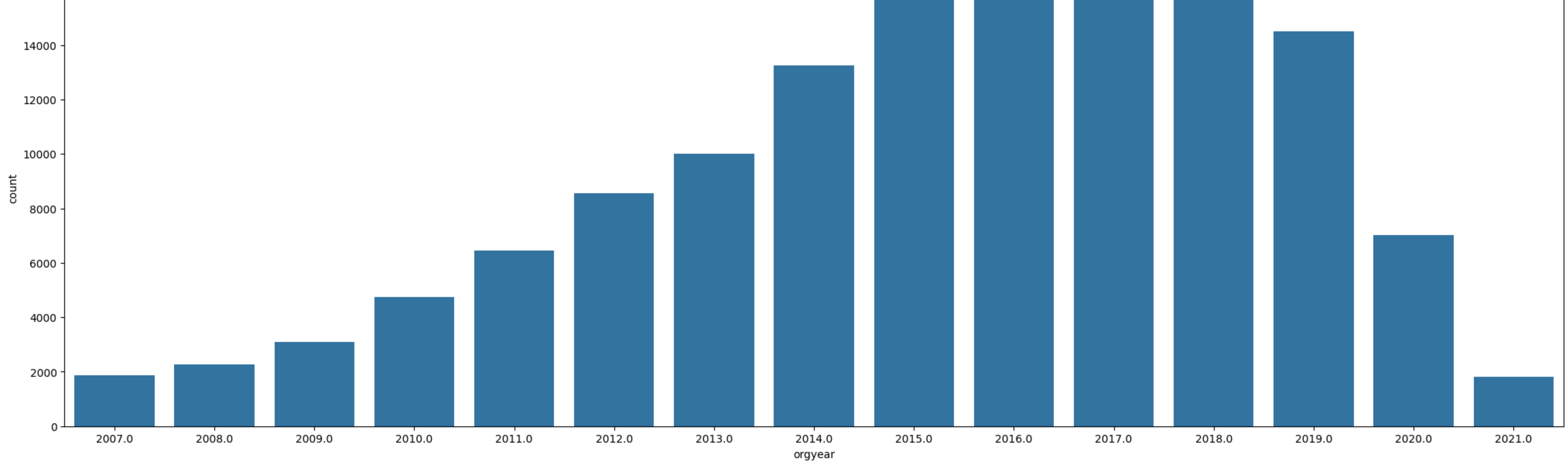
```
categroical_columns = [ 'company_hash','job_position','orgyear','ctc_updated_year']
```

```
import matplotlib.pyplot as plt
import seaborn as sns
for i in categroical_columns:
    tmp = df.copy()
    tmp['count'] = 1
    tmp = tmp.groupby(i).sum()['count'].reset_index().sort_values('count',ascending=False).head(15)
    plt.figure(figsize=(25,8))
    sns.barplot(data=tmp,y='count',x=i).set(title=i)

    plt.show()
```

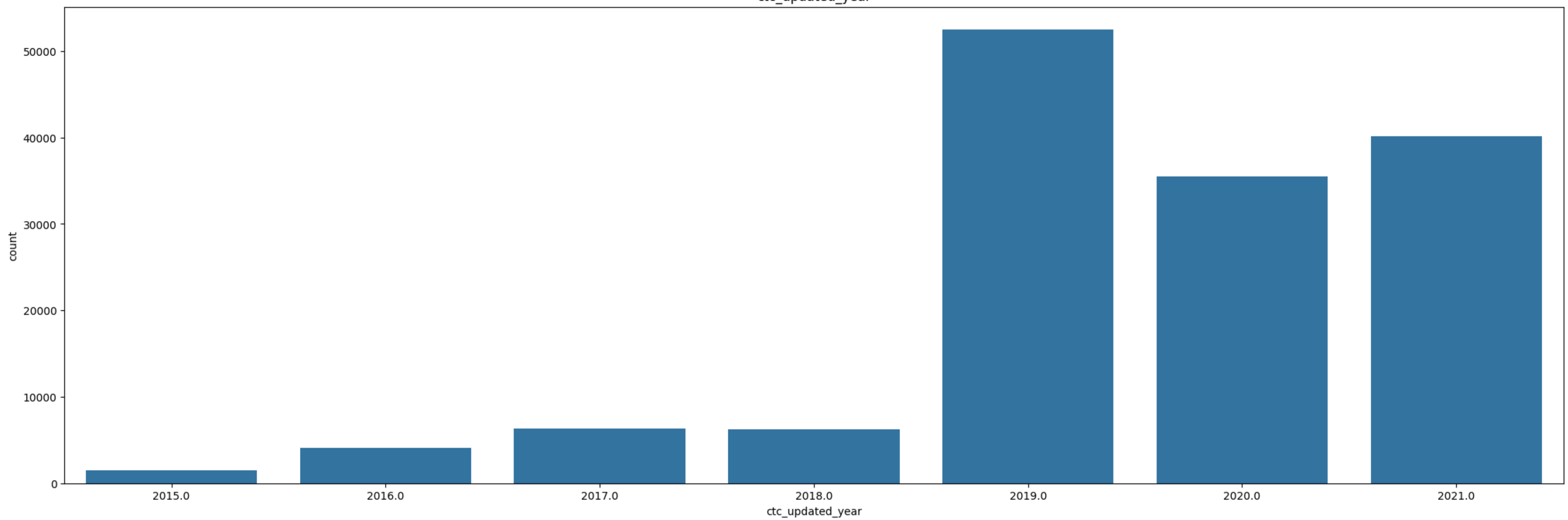

company_hash
count
0
1000
2000
3000
4000
5000
nvnv wgzohrnvzwj otqcxwto
xzegojo
vbvkgz
zgn vuurxwvmrt
vwwghzn
wgszxkvzn
vwwtznhqt
gqvwrt
fxuqg rxbxnta
bxwqgogen
wvustbxzx
zvz
ywr ntwyzgrgsxto
vagmt
uvjnb
lubgqsvz wyvot wg
company_hash


job_position
count
0
5000
10000
15000
20000
25000
30000
35000
Backend Engineer
FullStack Engineer
na
Other
Frontend Engineer
QA Engineer
Engineering Leadership
Android Engineer
Data Scientist
SDET
Devops Engineer
Support Engineer
iOS Engineer
Data Analyst
Engineering Intern
job_position


orgyear
16000

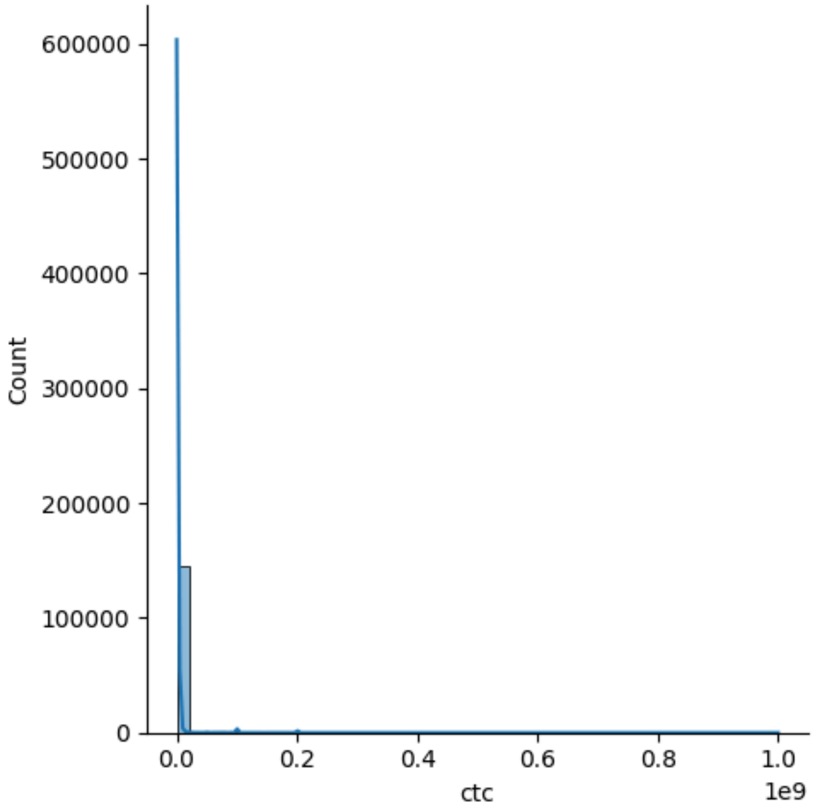


count
0
2000
4000
6000
8000
10000
12000
14000
2007.0
2008.0
2009.0
2010.0
2011.0
2012.0
2013.0
2014.0
2015.0
2016.0
2017.0
2018.0
2019.0
2020.0
2021.0
orgyear

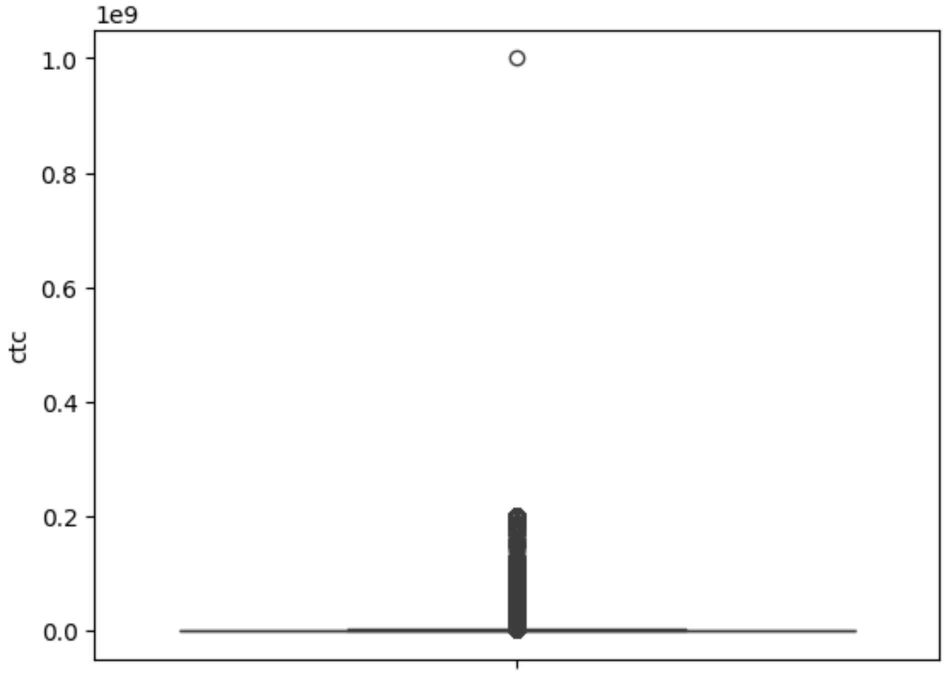


ctc_updated_year
count
0
10000
20000
30000
40000
50000
2015.0
2016.0
2017.0
2018.0
2019.0
2020.0
2021.0
ctc_updated_year

```
sns.displot(df['ctc'],kde=True,bins=50)
plt.show()
```




```
v = df['ctc']
sns.boxplot(v)
plt.show()
```




```
df.sort_values(['ctc']).iloc[1000:1020,:]
```

	email_hash	Unnamed: 0	company_hash	orgyear	ctc	job_position	ctc_updated_year	orgyear_na	ctc_updated_ye
18660	1ed6102745820ab25394b6c68ae9ce00d16da570cd5269...	66178	gqvwrt	2017.0	20000	Backend Engineer	2019.0	False	
50742	54ef23798f69b74b097501d53ba650082e7fe88d9ef087...	121968	ctqxex	2019.0	20000	Product Designer	2019.0	False	
48629	516cce2379bb216da1a1facf09db479d4453a126517989...	135854	wgbgag	2014.0	20000	Backend Engineer	2018.0	False	
5376	08bac5026bf379045813ce0e99e8df5601a56d913424fe...	84968	uqtowqxmtq360 ogrhnxgzo	2013.0	20000	FullStack Engineer	2019.0	False	
90663	975e224e718de0d75c2d33d2bf24e75c4b7559664763b1...	125784	wgcvrtzot ntwyzgrgsxto	2016.0	20000	Backend Engineer	2019.0	False	
120930	c9f0c1b5a2a71b0b754abcf68bc68c55188763ed0c8837...	99951	wtrxsg xzw	2009.0	20000	Backend Engineer	2019.0	False	
18584	1eb23fa16469ad4ec4bfd9c02a4878bee17986453eb87a...	81272	gqvzst	2018.0	20000	FullStack Engineer	2019.0	False	
57264	5f897bb2d379c6725963ce768e321023df76b61d89685b...	186777	egdxn ogenfvqt xzw	2015.0	20000	na	2016.0	False	
133452	decd2e0b07fec24774a6f9b99fcfe7243736eff34558a8...	190232	rtzaergf	2019.0	20000	Backend Engineer	2020.0	False	
48884	51d95ac4c20482bc4a898308d5b6ccb1454c2020ad3f69...	172981	taqvvp	2017.0	20000	FullStack Engineer	2019.0	False	
26393	2bcb5c9ed20f93a9e3e212dd76adf520057ff1764c9e72...	41221	onhatzn	2017.0	20000	Frontend Engineer	2021.0	False	
150206	fa7635744336651e986400fe9afd74f35ef6675e36c246...	130387	ertdxouytqt ogenfvqt otqcxwto uqxcvnt rxbxnta	2016.0	20000	Frontend Engineer	2019.0	False	
60598	6516b25bd2b7ad378caafe1f7d9eab7b3dbd33d713f986...	194344	hwyxqh	2016.0	20000	FullStack Engineer	2019.0	False	
74637	7ccf11cb54a14b0d7384a7edaf6b7bfd5ea760e55aff60...	47609	btzngq sqvuyxwo	2017.0	20000	Backend Engineer	2019.0	False	
62321	68071ee5df5210fe9264fbad4609a751ad30dbe6fc05fc...	91159	xmtd	2016.0	20000	FullStack Engineer	2021.0	False	
36725	3d37626eb7c103fb80700a1f223ba36951da26e789d87d...	188345	ntog	2015.0	20000	Engineering Leadership	2019.0	False	
23996	27d261a44415f8c6596501a33c8fdd9fe658e1a40db0ce...	52338	otqcxej	2014.0	20000	na	2019.0	False	
30954	3394eeca520d9029ce6bd56e83faa5d4d82c396f453e2a...	97003	mrvmmtg	2011.0	20000	Android E...	2019.0	False	

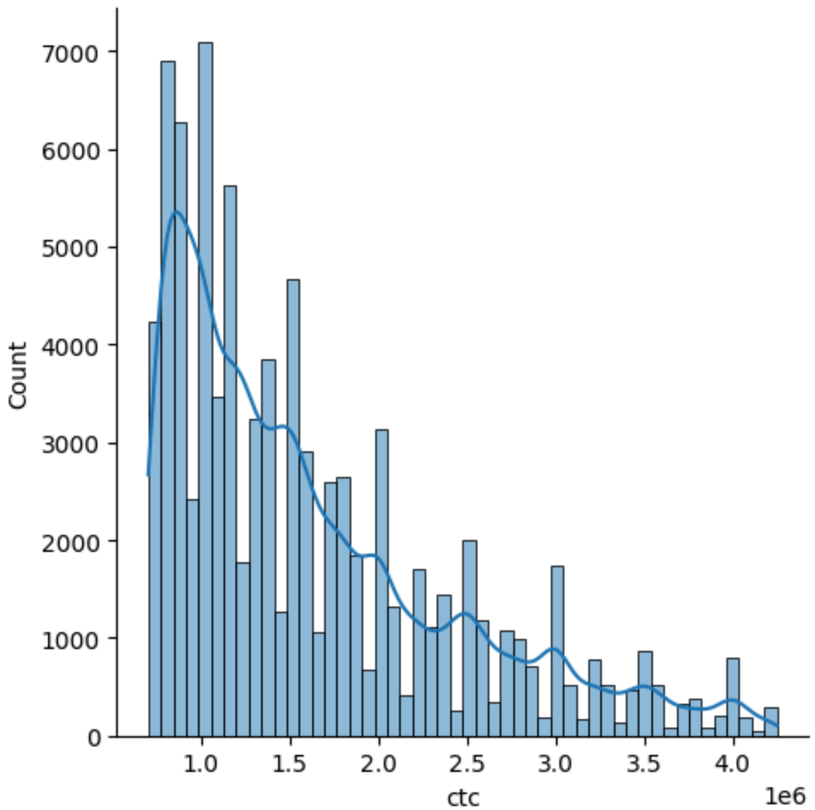
							Engineer		
80218	86085043c7ed1ffee48ce667750708e099656959d943a3...	168082	jo xzegogen ucn rna	2017.0	20000	Android Engineer		2019.0	False
61728	66faca4dac89b8a8aa598bbd666f279787b27a01f85efc...	193417	rtvz zgat	2018.0	20000	Frontend Engineer		2020.0	False

```
df = df[df['ctc'] >702475]
```

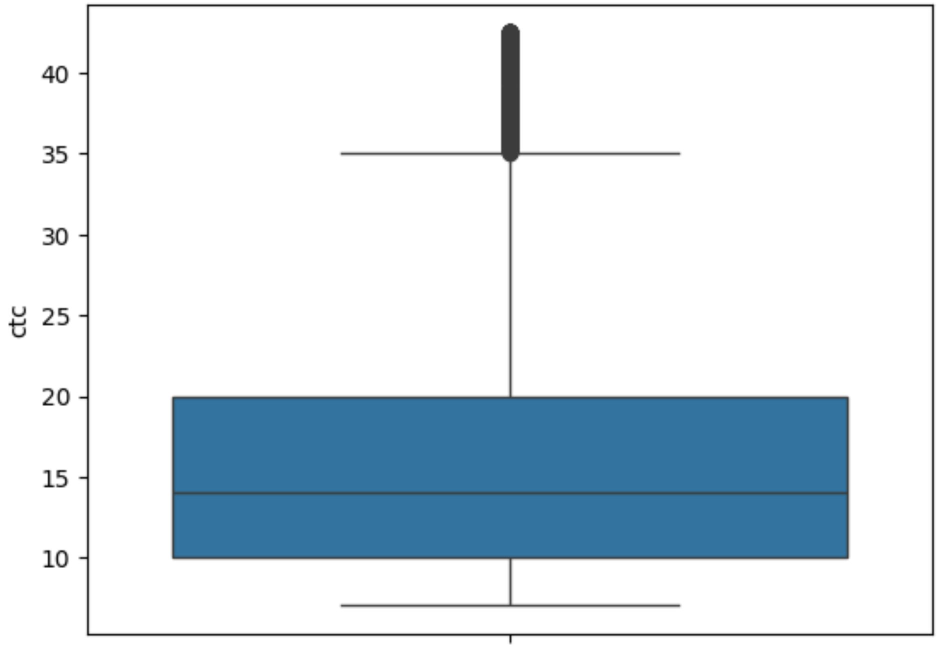
```
# Outlier removal using IQR
dftmp = df.copy()
print(dftmp.shape)
cols = ['ctc'] # one or more

Q1 = dftmp[cols].quantile(0.25)
Q3 = dftmp[cols].quantile(0.75)
IQR = Q3 - Q1

dftmp = dftmp[~((dftmp[cols] < (Q1 - 1.5 * IQR)) |(dftmp[cols] > (Q3 + 1.5 * IQR))).any(axis=1)]
print(dftmp.shape)
```

```
(92586, 20)
(86491, 20)
```

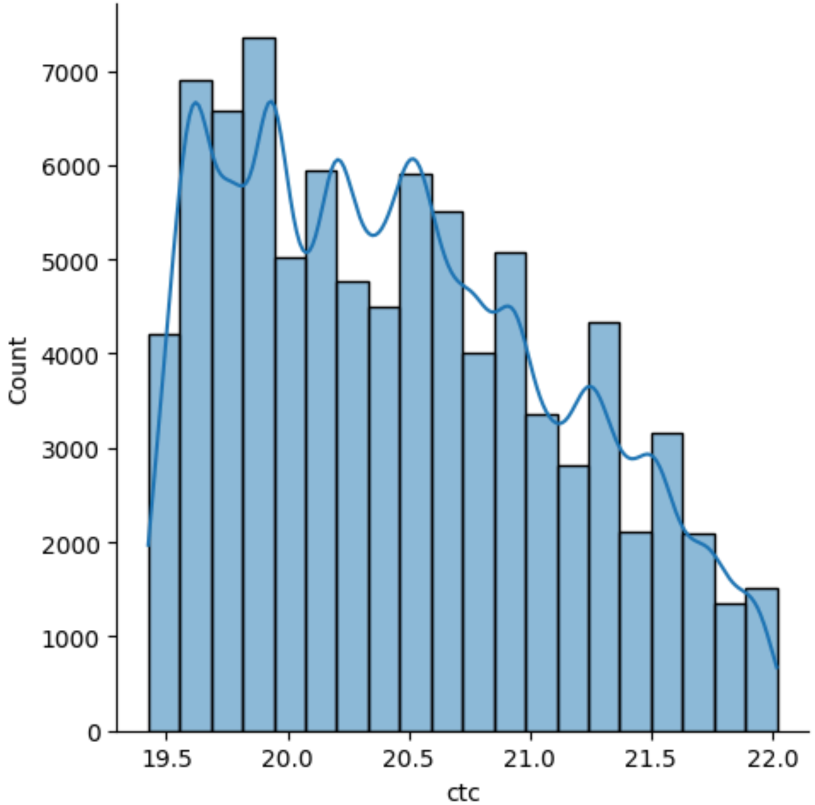
```
v = dftmp['ctc']
sns.displot(v,kde=True,bins=50)
plt.show()
```



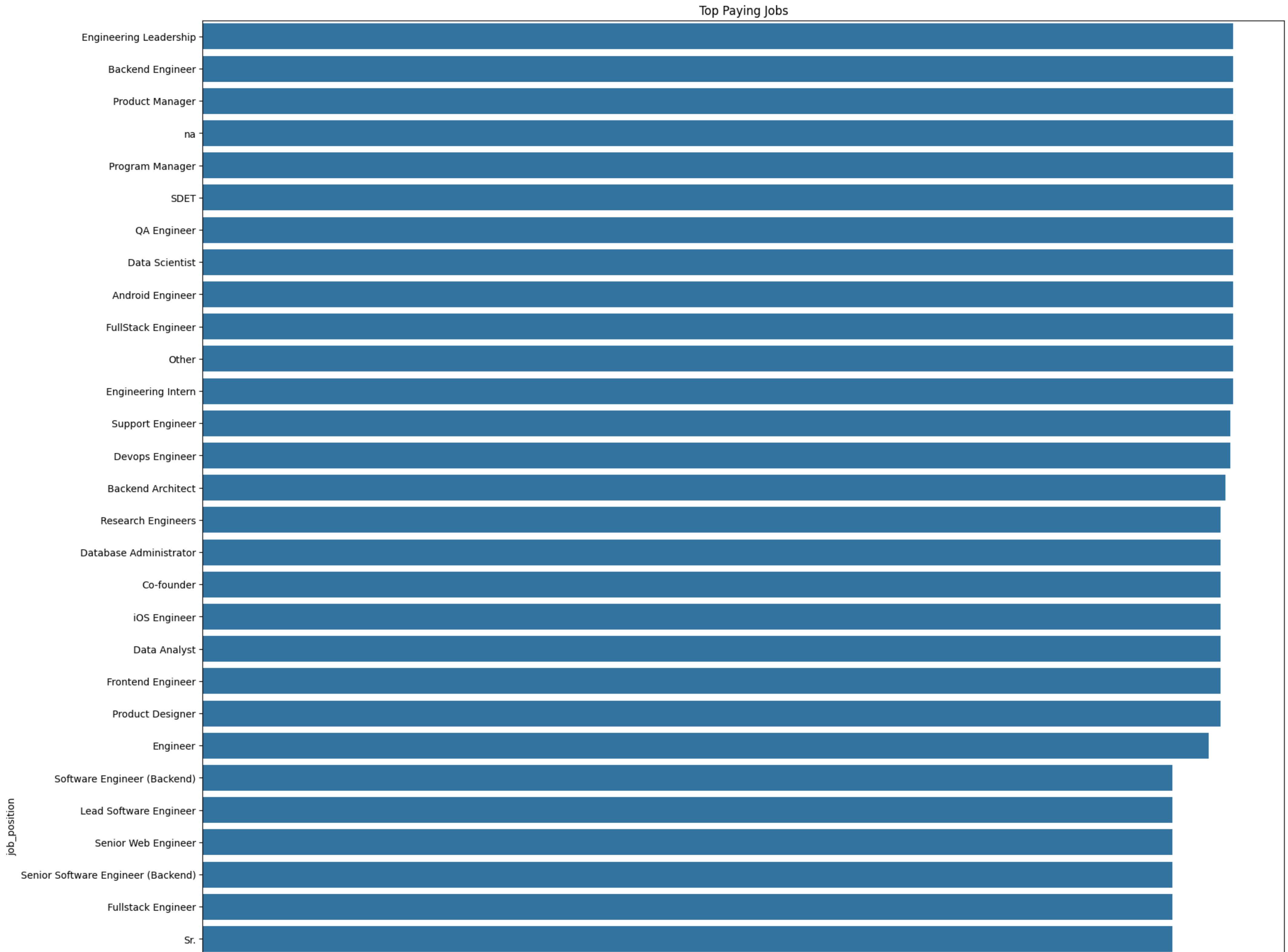

```
v = dftmp['ctc']/100000
sns.boxplot(v)
plt.show()
```



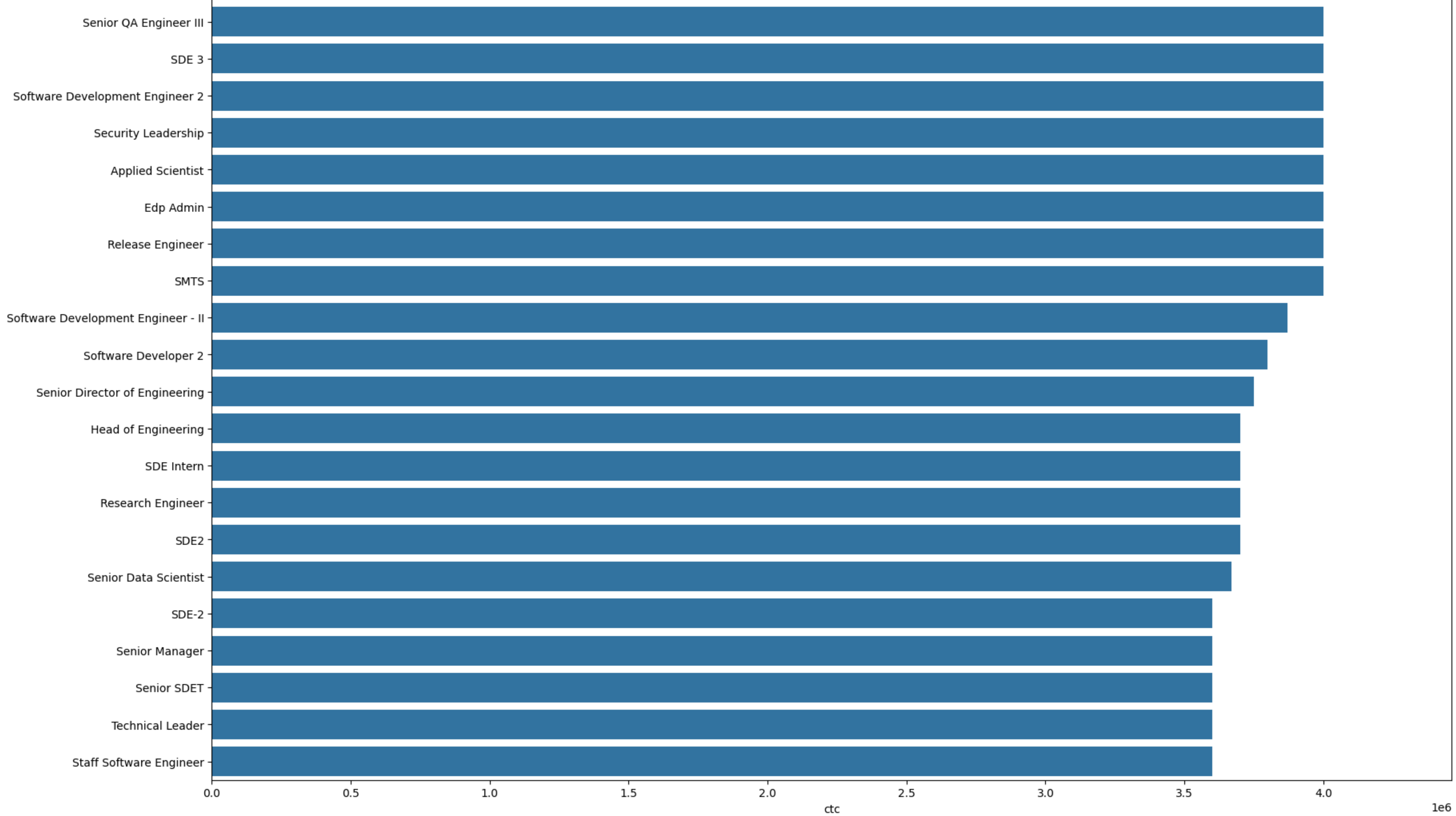

```
v = np.log2(dftmp['ctc'])
sns.displot(v,kde=True,bins=20)
plt.show()
```




```
tmp = dftmp.copy()
tmp = tmp.groupby(['job_position']).max()['ctc'].reset_index().sort_values('ctc',ascending=False).head(50)
plt.figure(figsize=(20,30))
sns.barplot(data=tmp,x='ctc',y='job_position').set(title="Top Paying Jobs")
plt.show()
list(tmp['job_position'])
```



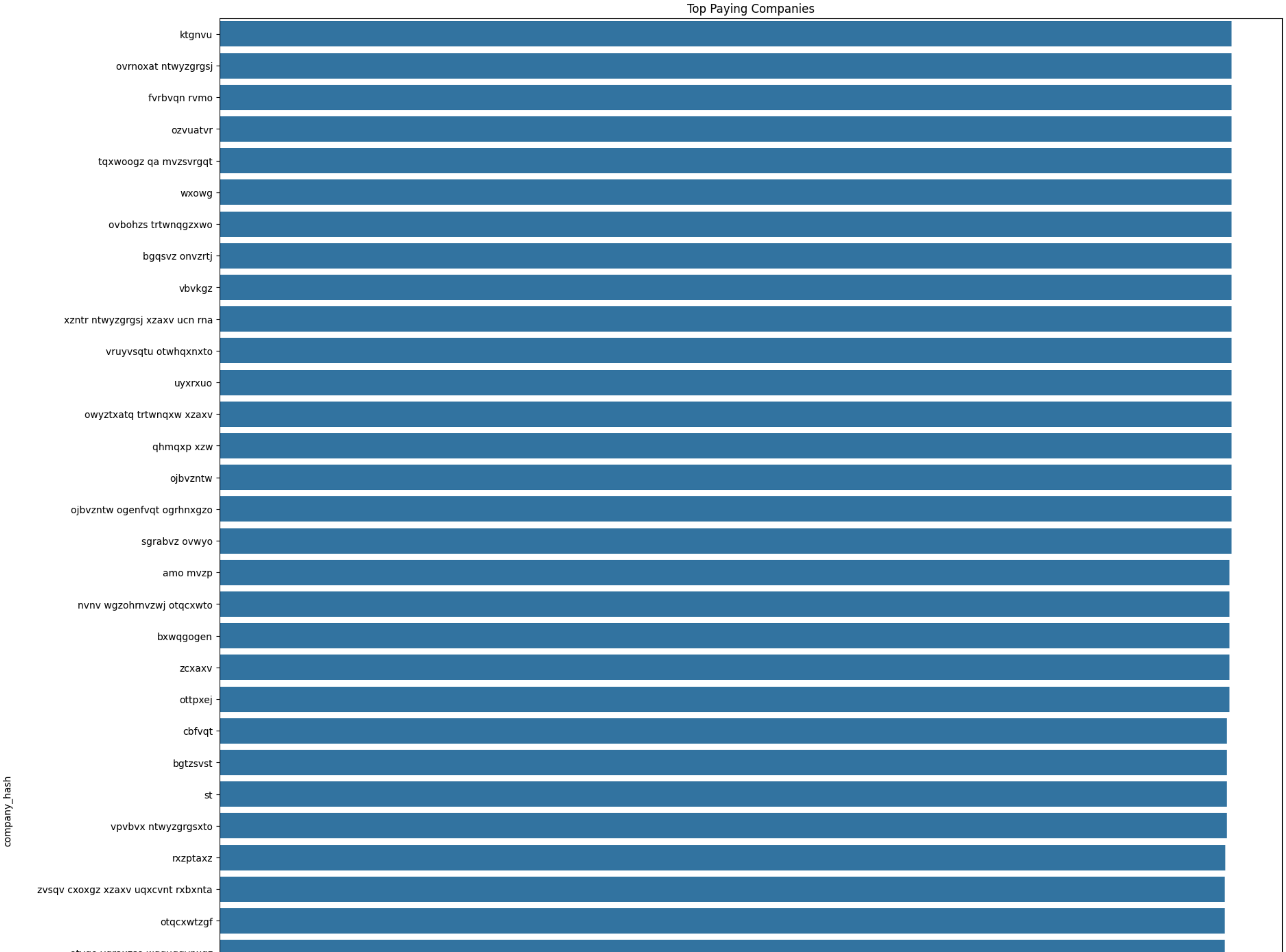
Top Paying Jobs
Engineering Leadership
Backend Engineer
Product Manager
na
Program Manager
SDET
QA Engineer
Data Scientist
Android Engineer
FullStack Engineer
Other
Engineering Intern
Support Engineer
Devops Engineer
Backend Architect
Research Engineers
Database Administrator
Co-founder
iOS Engineer
Data Analyst
Frontend Engineer
Product Designer
Engineer
Software Engineer (Backend)
Lead Software Engineer
Senior Web Engineer
Senior Software Engineer (Backend)
Fullstack Engineer
Sr.
job_position

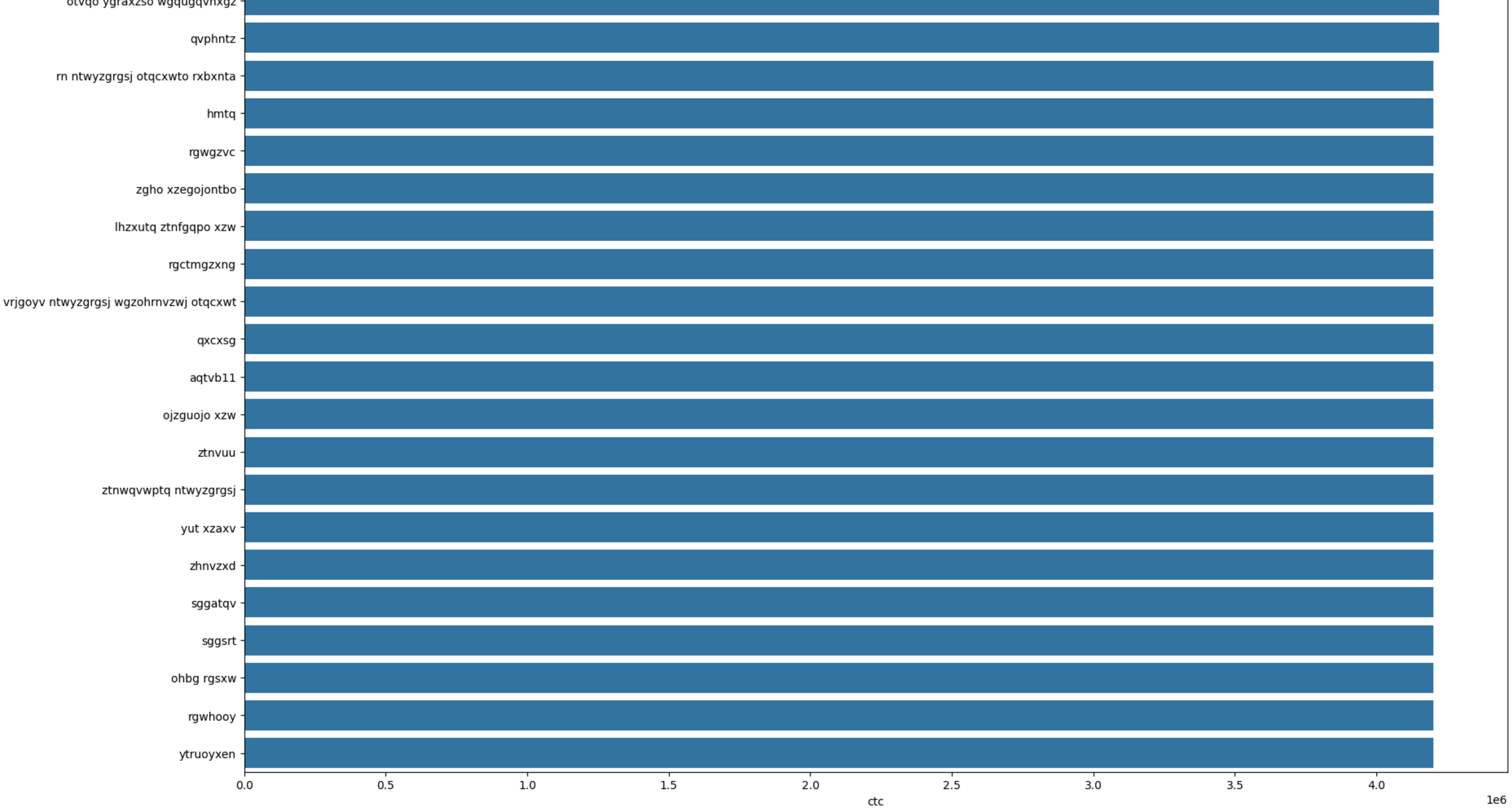



```
['Engineering Leadership',
 'Backend Engineer',
 'Product Manager',
 'na',
 'Program Manager',
 'SDET',
 'QA Engineer',
 'Data Scientist',
 'Android Engineer',
 'FullStack Engineer',
 'Other',
 'Engineering Intern',
 'Support Engineer',
 'Devops Engineer',
```

```
 'Devops Engineer',
 'Backend Architect',
 'Research Engineers',
 'Database Administrator',
 'Co-founder',
 'iOS Engineer',
 'Data Analyst',
 'Frontend Engineer',
 'Product Designer',
 'Engineer',
 'Software Engineer (Backend)',
 'Lead Software Engineer',
 'Senior Web Engineer',
 'Senior Software Engineer (Backend)',
 'Fullstack Engineer',
 'Sr.',
 'Senior QA Engineer III',
 'SDE 3',
 'Software Development Engineer 2',
 'Security Leadership',
 'Applied Scientist',
 'Edp Admin',
 'Release Engineer',
 'SMTS',
 'Software Development Engineer - II',
 'Software Developer 2',
 'Senior Director of Engineering',
 'Head of Engineering',
 'SDE Intern',
 'Research Engineer',
 'SDE2',
 'Senior Data Scientist',
 'SDE-2',
 'Senior Manager',
 'Senior SDET',
 'Technical Leader',
 'Staff Software Engineer']
```

```
tmp = dftmp.copy()
tmp = tmp.groupby(['company_hash']).max()['ctc'].reset_index().sort_values('ctc',ascending=False).head(50)
plt.figure(figsize=(20,30))
sns.barplot(data=tmp,x='ctc',y='company_hash').set(title="Top Paying Companies")
plt.show()

list(tmp['company_hash'])
```


Top Paying Companies
company_hash
ktgnvu
ovrnoxat ntwyzgrgsj
fvrbvqn rvmo
ozvuatvr
tqxwoogz qa mvzsvrgqt
wxowg
ovbohzs trtwnqgzxwo
bgqsvz onvzrtj
vbvkgz
xzntr ntwyzgrgsj xzaxv ucn rna
vruyvsqtu otwhqxnxto
uyxrxuo
owyztxatq trtwnqxw xzaxv
qhmqxp xzw
ojbvzntw
ojbvzntw ogenfvqt ogrhnxgzo
sgrabvz ovwyo
amo mvzp
nvnv wgzohrnvzwj otqcxwto
bxwqgogen
zcxaxv
ottpxej
cbfvqt
bgtzsvst
st
vpvbvx ntwyzgrgsxto
rxzptaxz
zvsqv cxoxgz xzaxv uqxcvnt rxbxnta
otqcxwtzgf




```
['ktgnvu',
 'ovrnoxat ntwyzgrgsj',
 'fvrbvqn rvmo',
 'ozvuatvr',
 'tqxwoogz qa mvzsvrgqt',
 'wxowg',
 'ovbohzs trtwnqgzxwo',
 'bgqsvz onvzrtj',
 'vbvkgz',
 'xzntr ntwyzgrgsj xzaxv ucn rna',
 'vruyvsqtu otwhqxnxto',
 'uyxrxuo',
 'owyztxatq trtwnqxw xzaxv',
 'qhmqxp xzw',
 'ojbvzntw',
```

```
 'ojbvzntw ogenfvqt ogrnnxgzo',
 'sgrabvz ovwyo',
 'amo mvzp',
 'nvnv wgzohrnvzwj otqcxwto',
 'bxwqgogen',
 'zcxaxv',
 'ottpxej',
 'cbfvqt',
 'bgtzsvst',
 'st',
 'vpvbvx ntwyzgrgsxto',
 'rxzptaxz',
 'zvsqv cxoxgz xzaxv uqxcvnt rxbxnta',
 'otqcxwtzgf',
 'otvqo ygraxzso wgqugqvnxgz',
 'qvphntz',
 'rn ntwyzgrgsj otqcxwto rxbxnta',
 'hmtq',
 'rgwgzvc',
 'zgho xzegojontbo',
 'lhzxutq ztnfgqpo xzw',
 'rgctmgzxng',
 'vrjgoyv ntwyzgrgsj wgzohrnvzwj otqcxwt',
 'qxcxsg',
 'aqtvb11',
 'ojzguojo xzw',
 'ztnvuu',
 'ztnwqvwptq ntwyzgrgsj',
 'yut xzaxv',
 'zhnvzxd',
 'sggatqv',
 'sggsrt',
 'ohbg rgsxw',
 'rgwhooy',
 'ytruoyxen']
```

```
# Manual Clustering
dateda = dftmp.copy()
```

```
grp = ['company_hash','job_position','YoE']
data_tmp1 = dateda.groupby(grp).agg({'ctc':['mean','median','min','max','count']}).reset_index()
data_tmp1.columns  = ["{} {}".format(b_, a_) if a_ not in grp else "{}".format(a_) for a_, b_ in zip(data_tmp1.columns.droplevel(1), data_tmp1.columns.d
data_tmp1.head(100).tail(50)

datatmp = dateda.merge(data_tmp1[['company_hash', 'job_position', 'YoE', 'mean ctc']],on=['company_hash', 'job_position', 'YoE'],how='left')



col1 = 'ctc'
col2 = 'mean ctc'
conditions  = [ datatmp[col1] > datatmp[col2], datatmp[col1] == datatmp[col2], datatmp[col1] < datatmp[col2] ]
choices     = [ 1, 2, 3 ]

datatmp['Designation'] = np.select(conditions, choices, default=np.nan)
```

```
grp = ['company_hash','job_position']
data_tmp1 = datatmp.groupby(grp).agg({'ctc':[('mean2','mean'),'median','min','max','count']}).reset_index()
data_tmp1.columns  = ["{} {}".format(b_, a_) if a_ not in grp else "{}".format(a_) for a_, b_ in zip(data_tmp1.columns.droplevel(1), data_tmp1.columns.d
data_tmp1.head(100).tail(50)


datatmp = datatmp.merge(data_tmp1[grp + ['mean2 ctc']],on=grp,how='left')


col1 = 'ctc'
col2 = 'mean2 ctc'
conditions  = [ datatmp[col1] > datatmp[col2], datatmp[col1] == datatmp[col2], datatmp[col1] < datatmp[col2] ]
choices     = [ 1, 2, 3 ]

datatmp['Class'] = np.select(conditions, choices, default=np.nan)
```

```
grp = ['company_hash']
data_tmp1 = datatmp.groupby(grp).agg({'ctc':[('mean3','mean'),'median','min','max','count']}).reset_index()
data_tmp1.columns  = ["{} {}".format(b_, a_) if a_ not in grp else "{}".format(a_) for a_, b_ in zip(data_tmp1.columns.droplevel(1), data_tmp1.columns.d
data_tmp1.head(100).tail(50)


datatmp = datatmp.merge(data_tmp1[grp + ['mean3 ctc']],on=grp,how='left')


col1 = 'ctc'
col2 = 'mean3 ctc'
conditions  = [ datatmp[col1] > datatmp[col2], datatmp[col1] == datatmp[col2], datatmp[col1] < datatmp[col2] ]
choices     = [ 1, 2, 3 ]

datatmp['Tier'] = np.select(conditions, choices, default=np.nan)
```

```
datatmp['diff_desig'] = datatmp['ctc'] - datatmp['mean ctc']
datatmp['diff_class'] = datatmp['ctc'] - datatmp['mean2 ctc']
datatmp['diff_tier'] = datatmp['ctc'] - datatmp['mean3 ctc']
```

```
# Top 10 employees (earning more than most of the employees in the company) - Tier 1
datatmp[datatmp['Tier'] == 1].sort_values('diff_tier',ascending=False).head(10)[['email_hash','ctc','mean3 ctc']]
```

	email_hash	ctc	mean3 ctc
76206	e15abfd41c005995728191e49ef001e83e813cd3ed5104...	4240000	1.051315e+06
49043	90d5114ca752c55babef2c517ac8b17aaee3d9ff5740de...	4200000	1.051315e+06
59592	b022b84623593cc38a3c1d39d4545b368a7b5f286be1c7...	4200000	1.051315e+06
54775	a1c1c8919e2918b24241a40271e02381daf199c61d7a3b...	4200000	1.143837e+06
70692	d13d7376e9ced16b4e250d0643f9139f8b36a62847f71b...	4200000	1.147773e+06
31657	5d872e52cb535a71fc75a5a97e779bb4c1554d0baa920d...	4200000	1.158025e+06
14811	2b10e1d996c6ab5e175eea35ca25ea7afbaacd1237ab64...	4200000	1.158025e+06
47739	8d0ed00904247626f5557f5983feeb5a0567d7726eea39...	4200000	1.176534e+06
31834	5dff6a65d548553262b6a289f014b2b72a5d47ff6dfa5c...	4170000	1.165011e+06
45639	86b90dd64ddb663ea35be98422947a01ba9ab837fb76df...	4000000	1.051315e+06

```
# Top 10 employees of data science in Amazon / TCS etc earning more than their peers - Class 1
datatmp[(datatmp['Tier'] == 1)&(datatmp['Class'] == 1)&(datatmp['job_position'].isin(['Data Science Analyst','Data Scientist','Data Scientist II','Assoc
```

	email_hash	ctc	mean2 ctc
81316	f04a0228e5af6f8f6ecc33e089892e80d85b3c749b3244...	4000000	1.533750e+06
56247	a63f3f44de7586430615a8a9bd13d41e7b0d541ca0f690...	4200000	1.862000e+06
16849	31616edfc502824631b11793313d35d5bb2288319dcb25...	3800000	1.513842e+06
21448	3efbb8c4d67b4a4c6ba4c639cd84e9ff98b85d5f57d82f...	3979999	1.716000e+06
33521	62f705ba342cb9e51117446a5522c2e42c14db27b9b20e...	4250000	2.025000e+06
83423	f67ae342b7431f7ab05eca998d904647b02711538aa839...	3750000	1.565556e+06
83551	f6e8c41a40ec308c996d498e22729359d2b564cae037a0...	3500000	1.410000e+06
191	009ded427ebcb5c2fb1970017a683693a7abef0fa96f5e...	3900000	1.834333e+06
79556	eb35a5d34977c6135372e46d6cc4f85332f1a4f9578bd5...	4080000	2.020000e+06
36095	6aa8cfeb5b98da66158e0af4ca8869362174abdba84a02...	3200000	1.233235e+06

```
# Bottom 10 employees of data science in Amazon / TCS etc earning less than their peers - Class 3
datatmp[(datatmp['Tier'] == 1)&(datatmp['Class'] == 3)&(datatmp['job_position'].isin(['Data Science Analyst','Data Scientist','Data Scientist II','Assoc
```

	email_hash	ctc	mean2 ctc
14517	2a3136f6e2d03a3dbfa3f683e4ae1b744b4815a8e0177c...	1700000	3125000.0
55809	a4f1770283497277f8cd3b7cb04e9b5c3135815eebb4cf...	2300000	3292500.0
48883	9069f6772b1e7959734a115bf49b2168a888608496af50...	1900000	2850000.0
82797	f49bd18e7fe914929f6cc23bb4e7979d58290119f2adcf...	1600000	2500000.0
51661	987a063524741381c302a096e4b019f46088e519f59f4a...	2000000	2750000.0
65969	c371eff30d6983ab69401441f359fed64397f7699c7aff...	1630000	2350000.0
79601	eb5552cf683e3072a7e2e2c6e63ebb46183a716b2bd2a1...	1780000	2496000.0
2813	080c3b2cc8fe9e7743520a3771a3b4db72e49ef2542ebf...	1400000	1986000.0
26915	4fcbc73fbd3da62f8750d69c13846ada4d1302f4817865...	1700000	2250000.0
61650	b63f00fbd2f8774eccde057bbf3f99ae1742adf496b2cc...	1600000	2102500.0

```
# Bottom 10 employees (earning less than most of the employees in the company)- Tier 3
datatmp[datatmp['Tier'] == 3].sort_values('diff_tier',ascending=True).head(10)[['email_hash','ctc','mean3 ctc']]
```

	email_hash	ctc	mean3 ctc
12124	2322345290a1926df62347d45f06b68932e219cb010bf8...	850000	3.262923e+06
64087	bda6e0f742115289a27f304078935331a5563d90c91461...	750000	2.929000e+06
15911	2e7e946b56a245338d8da1daf60ef851031c9964cffd25...	950000	2.950000e+06
4336	0c535bb44414d62cab133425339bd7e156ec79823899ae...	810000	2.770000e+06
73501	d96a6540ff59456abe30f51f68e954388b1f6922c4bb0c...	900000	2.840543e+06
49917	935480e039d80833292d858a553a4bc0f628b9b97ce9ec...	900000	2.840543e+06
19351	38d71a484d7663f7c14df8432620bbbab718933173a295...	1368000	3.262923e+06
70317	d034e386dbce817ee1ea099b161379d3341af0a16573d8...	800000	2.683125e+06
36015	6a6d1a4452505b678e264700fd0c28f247c4522d27f112...	770000	2.637273e+06
2613	077fd3f95d8dbf89c112a8eca6601db3729f51b53b57a0...	720000	2.577054e+06

```
# Top 10 employees in Amazon- X department - having 5/6/7 years of experience earning more than their peers - Tier X
datatmp[(datatmp['YoE'].isin([5,6,7]))&(datatmp['company_hash'].isin(['Amazon']))].sort_values('diff_desig',ascending=False).head(10)[['email_hash','ctc
```

email_hash	ctc	mean ctc

```
datatmp.groupby('company_hash').mean('ctc').reset_index().sort_values('ctc',ascending=False).head(10)[['company_hash','ctc']]
```

	company_hash	ctc
12501	tqxwoogz qa mvzsvrgqt	4250000.0
9064	ovrnoxat ntwyzgrgsj	4250000.0
19700	zvsqv cxoxgz xzaxv uqxcvnt rxbxnta	4220000.0
11150	rvzabvqp sqghu mvzsvrgqt xzaxv	4200000.0
6979	ntrtwgb bvzvsta otqcxwto gqsvzxovnxgz	4200000.0
14976	vsvqfvrwgzohrnxzs	4200000.0
7736	obvqnmxnuxdtr ntwy ucn rna	4200000.0
14621	vqgfzv wgzohrnxzs rrw	4200000.0
5546	lxoq yq	4200000.0
3679	fttduvz wgzohrnvzn	4200000.0

```python
import pandas as pd

# Top 2 positions in every company (based on their CTC)
tmp = datatmp[datatmp['job_position'].notna()]

tmp = tmp[pd.to_numeric(tmp['ctc'], errors='coerce').notna()]

tmp['ctc'] = pd.to_numeric(tmp['ctc'])

tmp = tmp.groupby(['company_hash', 'job_position']).mean('ctc').sort_values(['company_hash', 'ctc']).reset_index()

tmp = tmp.groupby('company_hash').head(2)[['company_hash', 'job_position']]
tmp
```

	company_hash	job_position
0	01 ojztqsj	Frontend Engineer
1	05mz exzytvrny uqxcvnt rxbxnta	Backend Engineer
2	1 jtvq	Backend Engineer
3	10 axsxnvr ahmvx rgzagz	Android Engineer
4	1001 vuuo	Frontend Engineer
...	...	...
35729	zyuw rxbxnta	Frontend Engineer
35730	zyvzwt fgga qtztfvr eqvzwyxogq yi	na
35731	zyvzwt wgzohrnxzs tzsxzttqo	Frontend Engineer
35732	zz	Other
35733	zzzbzb	Other

24644 rows × 2 columns

Next steps: Generate code with tmp | View recommended plots

```python
# Preparing the model for training
data = dateda.copy()
data
```

	email_hash	Unnamed: 0	company_hash	orgyear	ctc	job_position	ctc_updated_year	orgyear_na	ctc_updated_y
0	00003288036a44374976948c327f246fdbdf0778546904...	84782	bxwqgogen	2012.0	3500000	Backend Engineer	2019.0	False	
3	000120d0c8aa304fcf12ab4b85e21feb80a342cfea03d4...	53905	bxwqgotbx wgqugqvnxgz	2004.0	2000000	FullStack Engineer	2021.0	False	
4	00014d71a389170e668ba96ae8e1f9d991591acc899025...	138707	fvrbvqn rvmo	2009.0	3400000	na	2018.0	False	
6	00022dc29c7f77032275182b883d4f273ea1007aefc437...	7782	xzeqvwrgha ntwyzgrgsxto	2016.0	750000	Frontend Engineer	2019.0	False	
7	00036c2c5212d88d07acdc5bda7eef5653f8b09bbe30b7...	30543	ocu xnivz gbvz	2011.0	2300000	Other	2021.0	False	
...	...	...	...	...	...	...	...	...	
153432	fffa3a7b849802580a1972f11d192b43ff1c871bb43002...	79890	nvnv wgzohrnvzwj otqcxwto	2014.0	1800000	Backend Engineer	2019.0	False	
153438	fffc254e627e4bd1bc0ed7f01f9aebbba7c3cc56ac914e...	39683	tqxwoogz ogenfvqt wvbuho	2004.0	3529999	QA Engineer	2019.0	False	
153439	fffcf97db1e9c13898f4eb4cd1c2fe862358480e104535...	186656	trnqvcg	2015.0	1600000	na	2018.0	False	
153440	fffe7552892f8ca5fb8647d49ca805b72ea0e9538b6b01...	148878	znn avnv srgmvr atrxctqj otqcxwto	2014.0	900000	Devops Engineer	2019.0	False	
153442	ffffa3eb3575f43b86d986911463dce7bcadcea227e5a4...	117170	sgrabvz ovwyo	2018.0	1500000	FullStack Engineer	2021.0	False	

86491 rows × 20 columns

Next steps: Generate code with data View recommended plots

```
# Transforming ctc feature using log function
data['ctc_log'] = np.log2(data['ctc'])
```

```
# Columns that are non numeric are temporarily being removed so that we can perform imputation
drop_cols = ['job_position','email_hash','Unnamed: 0','company_hash']
for i in drop_cols:
    try:
        data.drop([i],axis=1,inplace=True)
    except:
        print('no')
```

```
no
no
no
```

```
data.isna().sum()
```

```
orgyear                  40
ctc                       0
ctc_updated_year          0
orgyear_na                0
ctc_updated_year_na       0
company_hash_na           0
email_hash_na             0
Unnamed: 0_na             0
ctc_na                    0
job_position_na           0
orgyear_na_na             0
ctc_updated_year_na_na    0
company_hash_na_na        0
YoE                      40
company_hash_encode       0
job_position_encode       0
ctc_log                   0
dtype: int64
```

```
from sklearn.impute import KNNImputer
from sklearn.pipeline import Pipeline
from sklearn.preprocessing import StandardScaler
from sklearn.impute import SimpleImputer
from sklearn.cluster import MiniBatchKMeans, KMeans
from sklearn.metrics import silhouette_score
from sklearn.cluster import KMeans
from sklearn.decomposition import PCA
```

```
# Kmeans Clustering
# Training the model with unscaled features
pipe_knn = Pipeline([('scaler', StandardScaler()), ('knn_imputer',  KNNImputer(n_neighbors=2, weights="uniform"))])
pipe_knn_5 = Pipeline([('scaler', StandardScaler()), ('knn_imputer',  KNNImputer(n_neighbors=5, weights="uniform"))])
pipe = Pipeline([('scaler', StandardScaler()), ('simple_imputer',  SimpleImputer(missing_values=np.nan, strategy='mean'))])
pipe_knn_pca = Pipeline([('scaler', StandardScaler()), ('knn_imputer',  KNNImputer(n_neighbors=2, weights="uniform")),('pca',PCA(n_components=8))])
pipe_unscaled = Pipeline([('knn_imputer',  KNNImputer(n_neighbors=5, weights="uniform"))])
```

```
data.describe()
```

	orgyear	ctc	ctc_updated_year	YoE	company_hash_encode	job_position_encode	ctc_log
count	86451.000000	8.649100e+04	86491.000000	86451.000000	86491.000000	86491.000000	86491.000000
mean	2013.207644	1.626626e+06	2019.441399	6.233855	0.002635	1209.606859	20.474986
std	34.638584	8.080777e+05	1.283691	34.620973	0.005557	878.498484	0.662663
min	0.000000	7.040000e+05	2015.000000	0.000000	0.000007	0.065171	19.425216
25%	2012.000000	1.000000e+06	2019.000000	3.000000	0.000033	319.141310	19.931569
50%	2015.000000	1.400000e+06	2019.000000	5.000000	0.000371	1318.079026	20.416995
75%	2017.000000	2.000000e+06	2020.000000	8.000000	0.002170	2431.717315	20.931569
max	2021.000000	4.250000e+06	2021.000000	2021.000000	0.034221	2431.717315	22.019031

```
data.info()
```

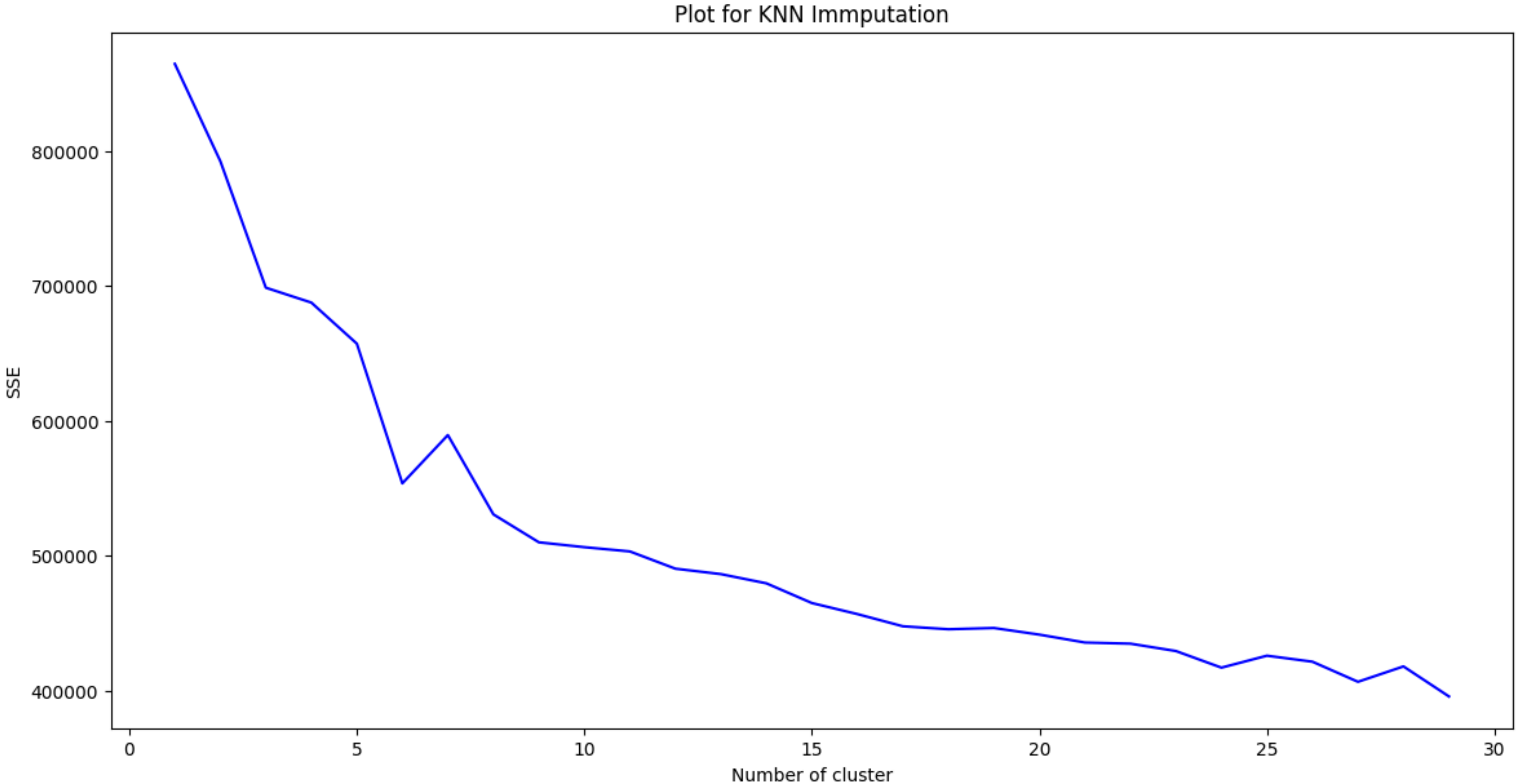
```
<class 'pandas.core.frame.DataFrame'>
Index: 86491 entries, 0 to 153442
Data columns (total 17 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   orgyear               86451 non-null  float64
 1   ctc                   86491 non-null  int64
 2   ctc_updated_year      86491 non-null  float64
 3   orgyear_na            86491 non-null  bool
 4   ctc_updated_year_na   86491 non-null  bool
 5   company_hash_na       86491 non-null  bool
 6   email_hash_na         86491 non-null  bool
 7   Unnamed: 0_na         86491 non-null  bool
 8   ctc_na                86491 non-null  bool
 9   job_position_na       86491 non-null  bool
 10  orgyear_na_na         86491 non-null  bool
 11  ctc_updated_year_na_na 86491 non-null  bool
 12  company_hash_na_na    86491 non-null  bool
 13  YoE                   86451 non-null  float64
 14  company_hash_encode   86491 non-null  float64
 15  job_position_encode   86491 non-null  float64
```

```
 16  ctc_log                 86491 non-null  float64
dtypes: bool(10), float64(6), int64(1)
memory usage: 6.1 MB
```

```
# Finding optimal num of clusters using Elbow method
for name,pipeline in [('KNN Immputation',pipe_knn),('KNN Imputation with (default) 5 neighbours',pipe_knn_5),('Mean Imputation ',pipe),
                      ('KNN Immputation + PCA', pipe_knn_pca),('KNN Imputation Unscaled data',pipe_unscaled )]:

    X = pipeline.fit_transform(data)
    X = pd.DataFrame(X)
    if "PCA" not in name :
        X.columns= data.columns

    sse = {}
    #sil_score = {}
    print("Running for ",name)
    for k in range(1, 30):
        #print('K :',k)
        kmeans = MiniBatchKMeans(init="k-means++",n_clusters=k,
                                 random_state=0).fit(X)
        label = kmeans.labels_
        data["clusters"] = label
        #print(data["clusters"])
        sse[k] = kmeans.inertia_

        #sil_score[k] = silhouette_score(X, label, metric='euclidean')

    plt.figure(figsize=(14,7))
    plt.plot(list(sse.keys()), list(sse.values()),'b-',label='Sum of squared error')
    plt.xlabel("Number of cluster")
    plt.ylabel("SSE")
    plt.title("Plot for "+name)
    plt.show()
```

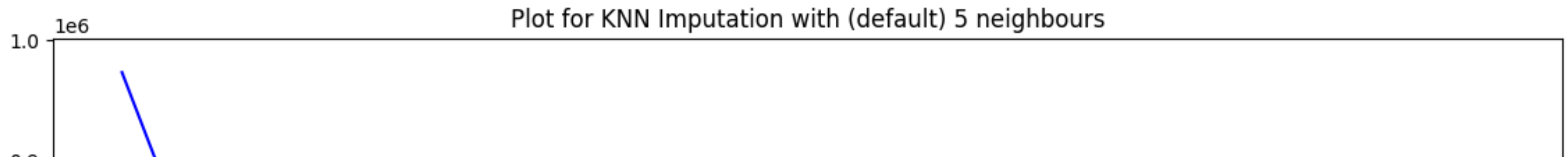
```
Running for  KNN Immputation
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
```

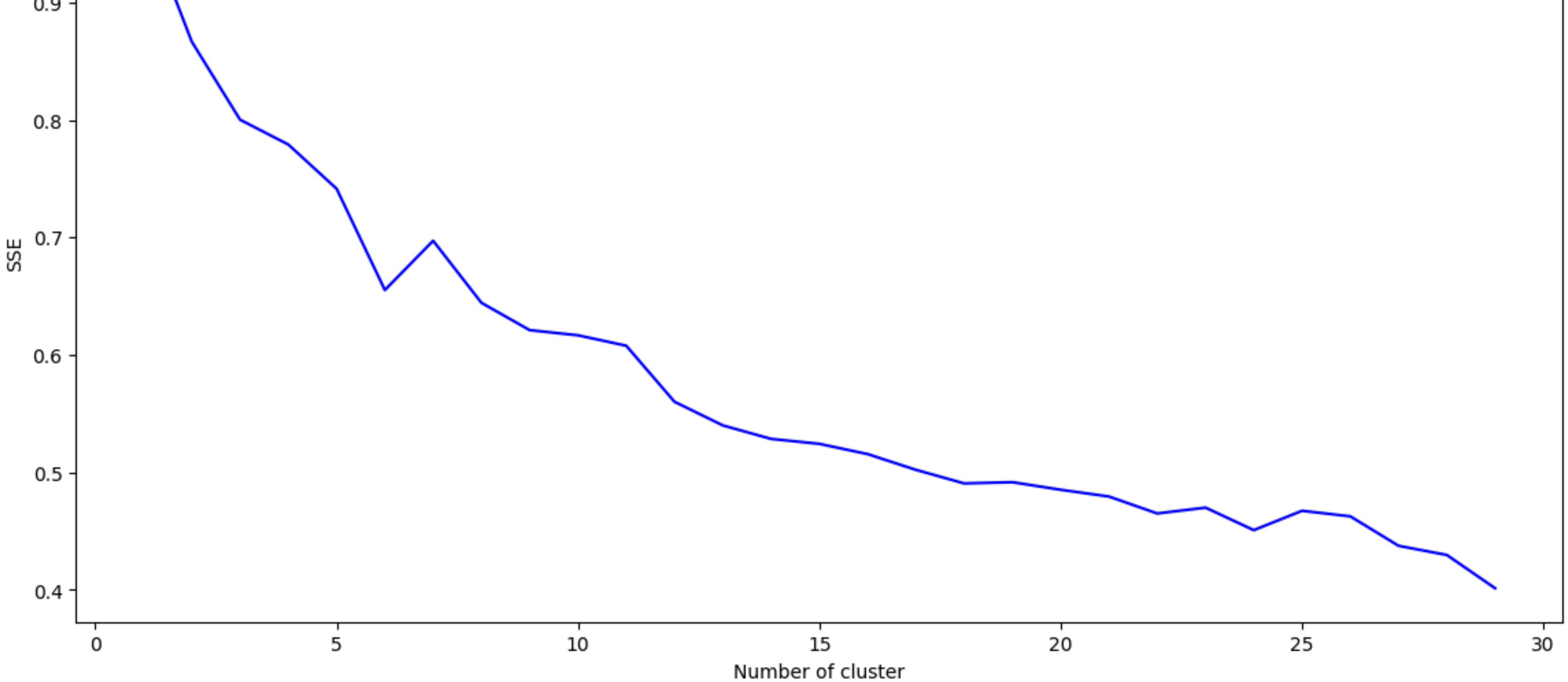
```
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
```




```
Running for  KNN Imputation with (default) 5 neighbours
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
```

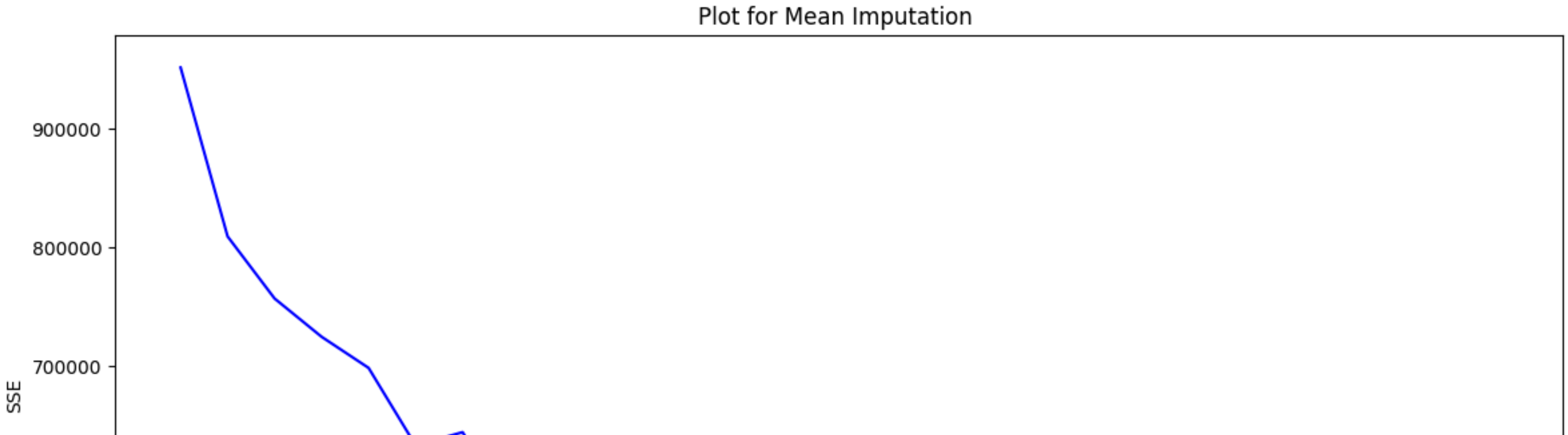
```
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
```

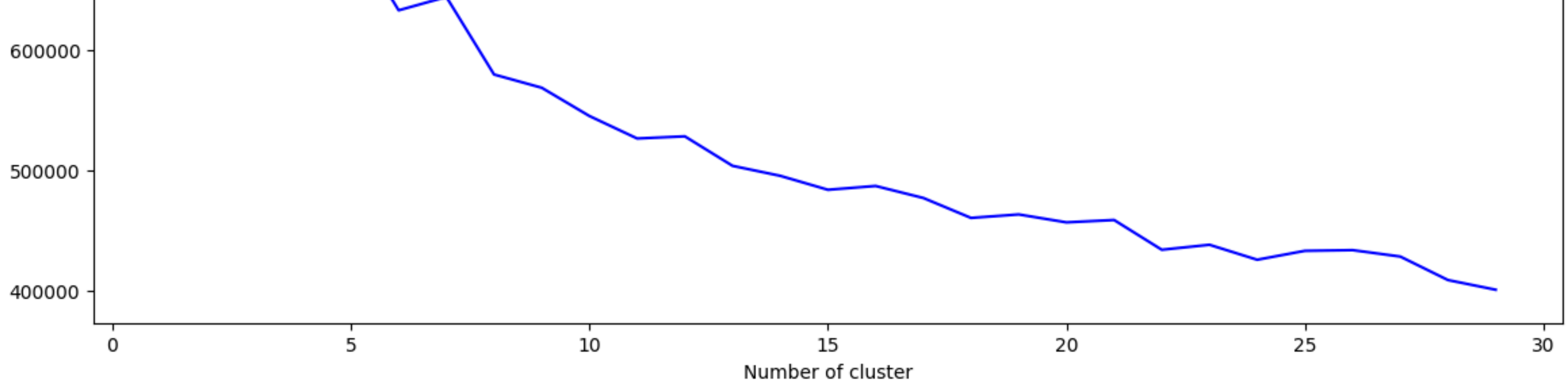





```
Running for  Mean Imputation
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
```

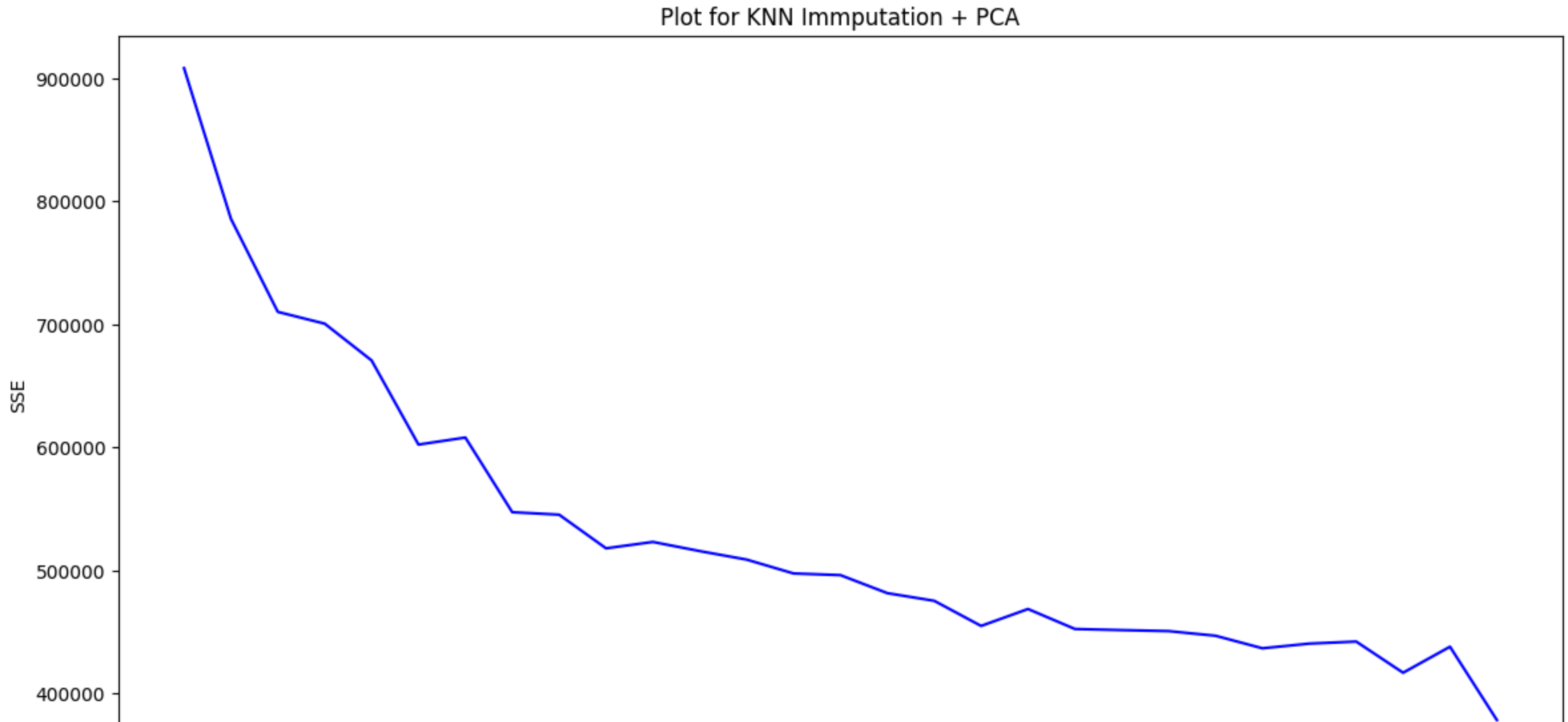
```
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
```

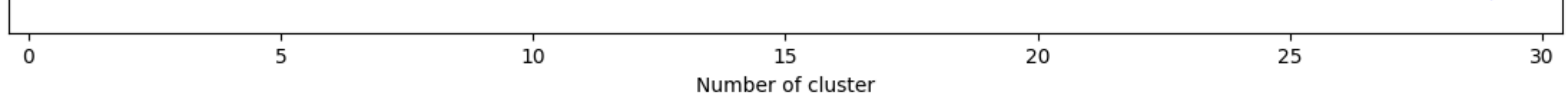





```
Running for  KNN Immputation + PCA
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
```

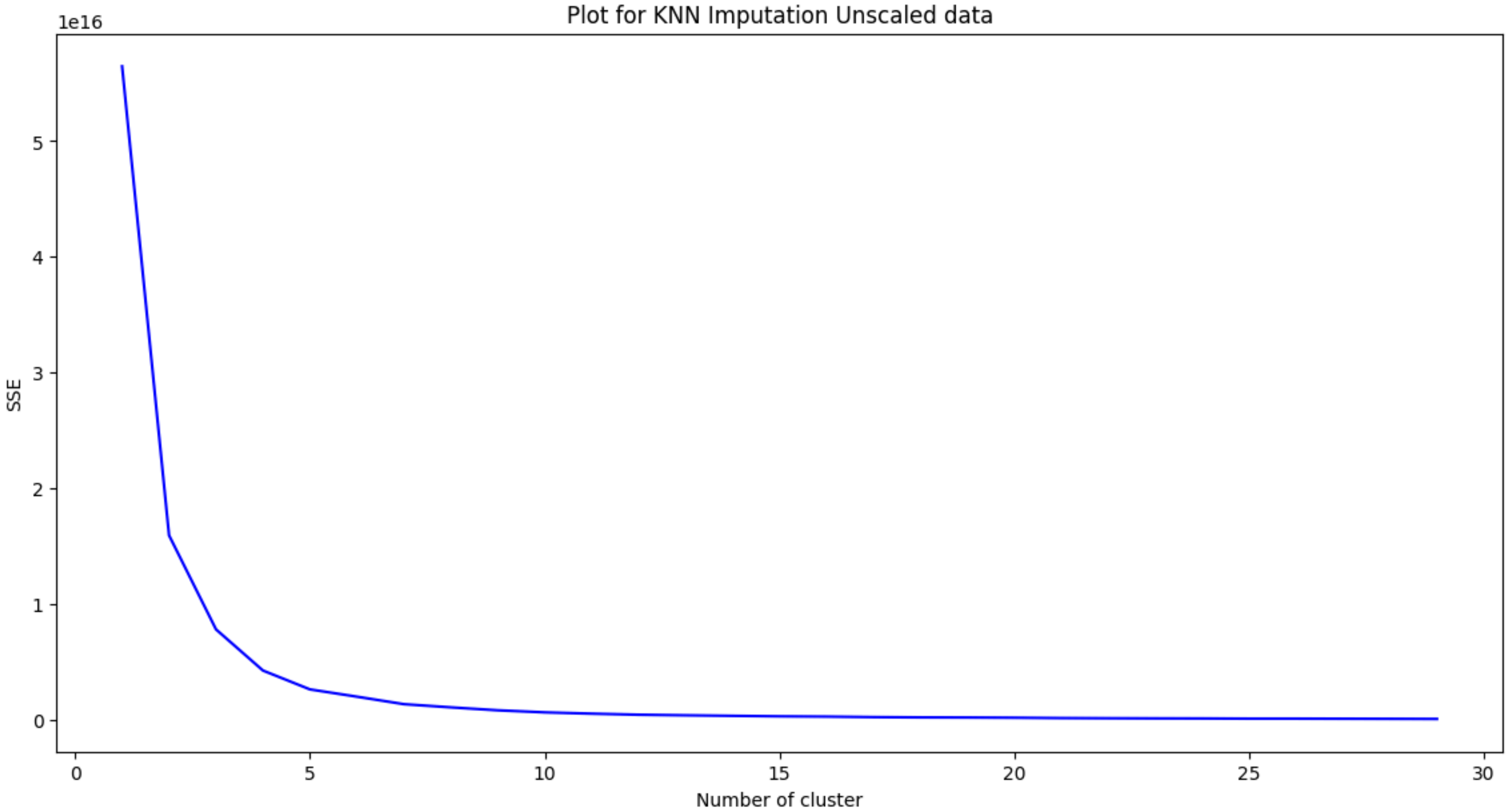
```
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
```






```
Running for  KNN Imputation Unscaled data
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
```

```
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 3 to 'auto' in
  warnings.warn(
```



```
# OBSERVATION
# Number of clusters is around 16-20 for scaled data,
while around 5 for unscaled data
# Number of clusters around 2 seems optimal in most
cases, while in last plot(with single linkage) number of
clusters around 16 is optimal
# Top Paying job titles include 'Engineering Leadership',
'Backend Engineer', 'Product Manager', 'Program Manager',
'SDET', 'QA Engineer', 'Data Scientist', 'Android
Engineer' and 'FullStack Engineer'.
# AMong Top paying companies mean salary for these
company is increasing every year
# Avg CTC seems to be decreasing with year.
```

```
# RECOMMENDATIONS
# Freshers who want to work on technical side should look
for roles related to Backend Engineer, SDET, QA engineer,
Dataa Scientist,
# Android Engineer,Full stack engineer to get good
salaries as experience increases.
# Other Observations and Recommendations are given at the
cell level itself.
```