Information about the dataset:

Email_hash - Anonymised Personal Identifiable Information (PII)

Company_hash - This represents an anonymized identifier for the company, which is the current employer of the learner.

orgyear - Employment start date

CTC - Current CTC

Job_position - Job profile in the company

CTC_updated_year - Year in which CTC got updated (Yearly increments, Promotions)

In [2]:
```
import re
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as px
import plotly.graph_objects as go
import plotly.figure_factory as ff
import warnings
warnings.filterwarnings("ignore")
from collections import Counter
```

In [3]:
```
import os
os.getcwd()
```

Out[3]: '/Users/phoenix/Documents/Project'

In [4]:
```
os.chdir('/Users/phoenix/Downloads')
```

In [5]:
```
df = pd.read_csv('scaler_clustering.csv')
df.head()
```

Out[5]:

	Unnamed: 0	company_hash	email_hash	orgyear	ctc	job_position	ctc_
0	0	atrgxnnt xzaxv	6de0a4417d18ab14334c3f43397fc13b30c35149d70c05...	2016.0	1100000	Other	
1	1	qtrxvzwt xzegwgbb rxbxnta	b0aaf1ac138b53cb6e039ba2c3d6604a250d02d5145c10...	2018.0	449999	FullStack Engineer	
2	2	ojzwnvwnxw vx	4860c670bcd48fb96c02a4b0ae3608ae6fdd98176112e9...	2015.0	2000000	Backend Engineer	
3	3	ngpgutaxv	effdede7a2e7c2af664c8a31d9346385016128d66bbc58...	2017.0	700000	Backend Engineer	
4	4	qxen sqghu	6ff54e709262f55cb999a1c1db8436cb2055d8f79ab520...	2017.0	1400000	FullStack Engineer	

In [6]:
```
df.shape
```

Out[6]: (205843, 7)

In [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
```

In [8]:
```
df.dtypes
```

Out[8]:
```
Unnamed: 0            int64
company_hash         object
email_hash           object
orgyear             float64
ctc                   int64
job_position         object
ctc_updated_year    float64
dtype: object
```

In [9]:
```
## Checking for null values

df.isnull().sum()
```

Out[9]:
```
Unnamed: 0              0
company_hash           44
email_hash              0
orgyear                86
ctc                     0
job_position        52564
ctc_updated_year        0
dtype: int64
```

In [10]:
```
df["email_hash"].nunique(),df["company_hash"].nunique()
```

Out[10]:
```
(153443, 37299)
```

In [11]:
```
df["email_hash"].value_counts()[:11]
```

Out[11]:
```
email_hash
bbace3cc586400bbc65765bc6a16b77d8913836cfc98b77c05488f02f5714a4b    10
6842660273f70e9aa239026ba33bfe82275d6ab0d20124021b952b5bc3d07e6c     9
298528ce3160cc761e4dc37a07337ee2e0589df251d73645aae209b010210eee     9
3e5e49daa5527a6d5a33599b238bf9bf31e85b9efa9a94f1c88c5e15a6f31378     9
b4d5afa09bec8689017d8b29701b80d664ca37b83cb883376b2e95191320da66     8
faf40195f8c58d5c7edc758cc725a762d51920da996410b80ac4a4d85c803da0     8
4818edfd67ed8563dde5d083306485d91d19f4f1c95d193a1700e79dd245b75c     8
c0eb129061675da412b0deb15871dd06ef0d7cd86eb5f7e8cc6a20b0d1938183     8
d598d6f1fb21b45593c2afc1c2f76ae9f4cb7167156cdf93246d4192a89d8065     8
d15041f58bb01c8ee29f72e33b136e26bc32f3169a40b53d75fe7ae9cbb9a551     8
aacf9473e3cee3e3f7c322e49bb8a6d5346bb05f3ff5bb9e9ac3ae22729ab933     7
Name: count, dtype: int64
```

In [12]:
```
## Removing Special Characters Using Regex

def preprocess_string(string):
    new_string= re.sub('[^A-Za-z ]+', '', string).lower().strip()
    return new_string

mystring='\tAirtel\\\\&&**() X Labs'
preprocess_string(mystring)
```

Out[12]:
```
'airtel x labs'
```

In [13]:
```
df['job_position'] = df['job_position'].apply(lambda x: preprocess_string(str(x)))
df['job_position'].nunique()
```

Out[13]:
```
856
```

In [14]:
```
df['company_hash'] = df['company_hash'].apply(lambda x: preprocess_string(str(x)))
df['company_hash'].nunique()
```

Out[14]:
```
37208
```

In [15]:
```
df.shape
```

Out[15]:
```
(205843, 7)
```

In [16]:
```
## Dropping Unnecessary Columns

df.drop('Unnamed: 0', axis = 1, inplace = True)
```

In [17]:
```
## Dropping Duplicates

df.drop_duplicates(inplace = True)
df.shape
```

Out[17]:
```
(205697, 6)
```

We notice that there are 52564 empty job positions in the data. Here we adopt the following strategy to impute these fields,

1. We first impute the job positions with the most frequently occurring job positions within a company.

2. Then we impute the 'NaN' job positions with those positions whose corresponding 'ctc' is closest to NaN position's ctc.

3. Finally we impute the left over job positions with 'other'.

In [18]:
```
## Concerting 'nan' to 'NaN' value
```

```
df["job_position"] = df["job_position"].replace("nan", np.nan)
df['company_hash'] = df['company_hash'].replace("nan", np.nan)
```

In [19]:
```
df
```

Out[19]:

	company_hash	email_hash	orgyear	ctc	job_position	ctc_upda
0	atrgxnnt xzaxv	6de0a4417d18ab14334c3f43397fc13b30c35149d70c05...	2016.0	1100000	other	
1	qtrxvzwt xzegwgbb rxbxnta	b0aaf1ac138b53cb6e039ba2c3d6604a250d02d5145c10...	2018.0	449999	fullstack engineer	
2	ojzwnvwnxw vx	4860c670bcd48fb96c02a4b0ae3608ae6fdd98176112e9...	2015.0	2000000	backend engineer	
3	ngpgutaxv	effdede7a2e7c2af664c8a31d9346385016128d66bbc58...	2017.0	700000	backend engineer	
4	qxen sqghu	6ff54e709262f55cb999a1c1db8436cb2055d8f79ab520...	2017.0	1400000	fullstack engineer	
...	...	...	...	...	...	
205838	vuurt xzw	70027b728c8ee901fe979533ed94ffda97be08fc23f33b...	2008.0	220000	NaN	
205839	husqvawgb	7f7292ffad724ebbe9ca860f515245368d714c84705b42...	2017.0	500000	NaN	
205840	vwwgrxnt	cb25cc7304e9a24facda7f5567c7922ffc48e3d5d6018c...	2021.0	700000	NaN	
205841	zgn vuurxwvmrt	fb46a1a2752f5f652ce634f6178d0578ef6995ee59f6c8...	2019.0	5100000	NaN	
205842	bgqsvz onvzrtj	0bcfc1d05f2e8dc4147743a1313aa70a119b41b30d4a1f...	2014.0	1240000	NaN	

205697 rows × 6 columns

In [20]:
```
## Imputing NaN Values by grouping on Company hash and CTC

def temp(lst):
    """
    Finds the most frequently occurring element in a list.

    Parameters:
    lst (list): A list of elements.

    Returns:
    any: The element that appears most frequently in the list.
         If multiple elements have the same highest frequency, the first encountered is returned.
    """
    d = Counter(lst)
    return max(d, key=d.get)

df['job_position'] = df['job_position'].fillna(df.groupby(['company_hash', 'ctc'])['job_position'].transfor
```

In [21]:
```
df.isnull().sum()
```

Out[21]:
```
company_hash          44
email_hash             0
orgyear               86
ctc                    0
job_position       37126
ctc_updated_year       0
dtype: int64
```

In [22]:
```
df
```

Out[22]:

	company_hash	email_hash	orgyear	ctc	job_position	ctc_upda
0	atrgxnnt xzaxv	6de0a4417d18ab14334c3f43397fc13b30c35149d70c05...	2016.0	1100000	other	
1	qtrxvzwt xzegwgbb rxbxnta	b0aaf1ac138b53cb6e039ba2c3d6604a250d02d5145c10...	2018.0	449999	fullstack engineer	
2	ojzwnvwnxw vx	4860c670bcd48fb96c02a4b0ae3608ae6fdd98176112e9...	2015.0	2000000	backend engineer	
3	ngpgutaxv	effdede7a2e7c2af664c8a31d9346385016128d66bbc58...	2017.0	700000	backend engineer	
4	qxen sqghu	6ff54e709262f55cb999a1c1db8436cb2055d8f79ab520...	2017.0	1400000	fullstack engineer	
...	...	...	...	...	...	
205838	vuurt xzw	70027b728c8ee901fe979533ed94ffda97be08fc23f33b...	2008.0	220000	NaN	
205839	husqvawgb	7f7292ffad724ebbe9ca860f515245368d714c84705b42...	2017.0	500000	fullstack engineer	
205840	vwwgrxnt	cb25cc7304e9a24facda7f5567c7922ffc48e3d5d6018c...	2021.0	700000	NaN	
205841	zgn vuurxwvmrt	fb46a1a2752f5f652ce634f6178d0578ef6995ee59f6c8...	2019.0	5100000	NaN	
205842	bgqsvz onvzrtj	0bcfc1d05f2e8dc4147743a1313aa70a119b41b30d4a1f...	2014.0	1240000	NaN	

205697 rows × 6 columns

In [23]:
```
## Replacing missing values with Others.

df['company_hash'] = df['company_hash'].replace("",'Others')
```

In [24]:
```
## Grouping on company hash and job position and finally finding mean CTC for each job position

avg_ctc_by_pos = df.groupby(['company_hash','job_position']).agg('mean','ctc').reset_index()
avg_ctc_by_pos
```

Out[24]:

	company_hash	job_position	orgyear	ctc	ctc_updated_year
0	Others	android engineer	2014.0	4.800000e+05	2020.000000
1	Others	backend engineer	2015.5	8.165714e+05	2019.571429
2	Others	data analyst	2019.5	2.100000e+06	2018.500000
3	Others	data scientist	2020.0	2.100000e+06	2019.000000
4	Others	engineering intern	2019.5	7.750000e+05	2019.500000
...	...	...	...	...	...
58451	zyvzwt wgzohrnxzs tzsxzttqo	frontend engineer	2012.0	9.400000e+05	2019.000000
58452	zz	other	2013.0	1.370000e+06	2020.000000
58453	zz wgzztwn mya	other	2009.0	2.380000e+04	2017.000000
58454	zzb ztdnstz vacxogqj ucn rna	fullstack engineer	2017.0	6.000000e+05	2021.000000
58455	zzzbzb	other	1990.0	7.200000e+05	2020.000000

58456 rows × 5 columns

In [25]:
```
## Define function to impute missing job positions

def impute_job_position(row, avg_ctc_by_position):
    """
    Imputes the missing job position based on the closest CTC within the same company.

    Parameters:
    row (pd.Series): A row of the DataFrame containing "job_position", "company_hash", and "ctc".
    avg_ctc_by_position (pd.DataFrame): A DataFrame containing average CTC values by job position and compa

    Returns:
        - The imputed job position if "job_position" is missing.
        - The original job position if it is not missing.
    """
    if pd.isna(row["job_position"]):  # If job_position is NaN
        company_jobs = avg_ctc_by_position[avg_ctc_by_position["company_hash"] == row["company_hash"]]
        if not company_jobs.empty:
            # Find the job with the closest CTC
            closest_job = company_jobs.iloc[(company_jobs["ctc"] - row["ctc"]).abs().argsort()[:1]]["job_pc
            return closest_job
    return row["job_position"]

## Apply the function to fill NaN values
df["job_position"] = df.apply(lambda row: impute_job_position(row, avg_ctc_by_pos), axis=1)
```

```
print(df)
```

```
                  company_hash  \
0                 atrgxnnt xzaxv
1       qtrxvzwt xzegwgbb rxbxnta
2                  ojzwnvwnxw vx
3                      ngpgutaxv
4                     qxen sqghu
...                          ...
205838                 vuurt xzw
205839                 husqvawgb
205840                  vwwgrxnt
205841             zgn vuurxwvmrt
205842             bgqsvz onvzrtj

                                              email_hash  orgyear      ctc  \
0       6de0a4417d18ab14334c3f43397fc13b30c35149d70c05...   2016.0  1100000
1       b0aaf1ac138b53cb6e039ba2c3d6604a250d02d5145c10...   2018.0   449999
2       4860c670bcd48fb96c02a4b0ae3608ae6fdd98176112e9...   2015.0  2000000
3       effdede7a2e7c2af664c8a31d9346385016128d66bbc58...   2017.0   700000
4       6ff54e709262f55cb999a1c1db8436cb2055d8f79ab520...   2017.0  1400000
...                                                   ...      ...      ...
205838  70027b728c8ee901fe979533ed94ffda97be08fc23f33b...   2008.0   220000
205839  7f7292ffad724ebbe9ca860f515245368d714c84705b42...   2017.0   500000
205840  cb25cc7304e9a24facda7f5567c7922ffc48e3d5d6018c...   2021.0   700000
205841  fb46a1a2752f5f652ce634f6178d0578ef6995ee59f6c8...   2019.0  5100000
205842  0bcfc1d05f2e8dc4147743a1313aa70a119b41b30d4a1f...   2014.0  1240000

                      job_position  ctc_updated_year
0                            other            2020.0
1                fullstack engineer            2019.0
2                  backend engineer            2020.0
3                  backend engineer            2019.0
4                fullstack engineer            2019.0
...                             ...               ...
205838                          NaN            2019.0
205839           fullstack engineer            2020.0
205840                 qa engineer             2021.0
205841              frontend engineer            2019.0
205842  associate software developer            2016.0

[205697 rows x 6 columns]
```

In [26]:
```
df.isna().sum()
```

Out[26]:
```
company_hash          44
email_hash             0
orgyear               86
ctc                    0
job_position        4045
ctc_updated_year       0
dtype: int64
```

In [28]:
```
#Filling null values with others

df['job_position'] = df['job_position'].fillna('Others')
df['company_hash'] = df['company_hash'].fillna('Others')
```

In [29]:
```
df.isna().sum()
```

Out[29]:
```
company_hash         0
email_hash           0
orgyear             86
ctc                  0
job_position         0
ctc_updated_year     0
dtype: int64
```

In [30]:
```
df["orgyear"].value_counts()
```

Out[30]:
```
orgyear
2018.0    25248
2019.0    23416
2017.0    23214
2016.0    23012
2015.0    20589
          ...
2107.0        1
1972.0        1
2101.0        1
208.0         1
200.0         1
Name: count, Length: 77, dtype: int64
```

In [31]:
```
## Assigning Company name with count less than 1 with 'Others'

df.loc[df.groupby('company_hash')['ctc'].transform('count') < 5, 'company_hash'] = 'Others'
```

In [32]:
```
df['company_hash'].nunique()
```

Out[32]: 3779

In [33]:
```
df["orgyear"].value_counts()
```

Out[33]:
```
orgyear
2018.0    25248
2019.0    23416
2017.0    23214
2016.0    23012
2015.0    20589
          ...
2107.0        1
1972.0        1
2101.0        1
208.0         1
200.0         1
Name: count, Length: 77, dtype: int64
```

In [164…
```
dictio = dict(df.groupby('company_hash')["orgyear"].median())
```

In [35]:
```
## Filling NaN Values with median values

df.loc[df['orgyear'].isnull(),'orgyear'] = df['company_hash'].map(dictio)
```

In [36]:
```
df.isnull().sum()
```

Out[36]:
```
company_hash        0
email_hash          0
orgyear             0
ctc                 0
job_position        0
ctc_updated_year    0
dtype: int64
```

In [37]:
```
df['ctc_updated_year'] = df[['ctc_updated_year','orgyear']].max(axis = 1)
```

In [38]:
```
## Dropping rows with unusual dates

df = df[~(df['ctc_updated_year'] > 2025)]
```

In [39]:
```
df
```

Out[39]:

	company_hash	email_hash	orgyear	ctc	job_position	ctc_upda
0	atrgxnnt xzaxv	6de0a4417d18ab14334c3f43397fc13b30c35149d70c05...	2016.0	1100000	other	
1	qtrxvzwt xzegwgbb rxbxnta	b0aaf1ac138b53cb6e039ba2c3d6604a250d02d5145c10...	2018.0	449999	fullstack engineer	
2	Others	4860c670bcd48fb96c02a4b0ae3608ae6fdd98176112e9...	2015.0	2000000	backend engineer	
3	ngpgutaxv	effdede7a2e7c2af664c8a31d9346385016128d66bbc58...	2017.0	700000	backend engineer	
4	qxen sqghu	6ff54e709262f55cb999a1c1db8436cb2055d8f79ab520...	2017.0	1400000	fullstack engineer	
...	...	...	...	...	...	
205838	vuurt xzw	70027b728c8ee901fe979533ed94ffda97be08fc23f33b...	2008.0	220000	Others	
205839	husqvawgb	7f7292ffad724ebbe9ca860f515245368d714c84705b42...	2017.0	500000	fullstack engineer	
205840	vwwgrxnt	cb25cc7304e9a24facda7f5567c7922ffc48e3d5d6018c...	2021.0	700000	qa engineer	
205841	zgn vuurxwvmrt	fb46a1a2752f5f652ce634f6178d0578ef6995ee59f6c8...	2019.0	5100000	frontend engineer	
205842	bgqsvz onvzrtj	0bcfc1d05f2e8dc4147743a1313aa70a119b41b30d4a1f...	2014.0	1240000	associate software developer	

205667 rows × 6 columns

Feature Engineering

In [41]:
```
## Creating new feature 'Years of Experience'

df['years_of_experience'] = 2025 - df['orgyear']
```

In [42]:
```
df.head()
```

Out[42]:

	company_hash	email_hash	orgyear	ctc	job_position	ctc_updated_ye
0	atrgxnnt xzaxv	6de0a4417d18ab14334c3f43397fc13b30c35149d70c05...	2016.0	1100000	other	202(
1	qtrxvzwt xzegwgbb rxbxnta	b0aaf1ac138b53cb6e039ba2c3d6604a250d02d5145c10...	2018.0	449999	fullstack engineer	201
2	Others	4860c670bcd48fb96c02a4b0ae3608ae6fdd98176112e9...	2015.0	2000000	backend engineer	202(
3	ngpgutaxv	effdede7a2e7c2af664c8a31d9346385016128d66bbc58...	2017.0	700000	backend engineer	201
4	qxen sqghu	6ff54e709262f55cb999a1c1db8436cb2055d8f79ab520...	2017.0	1400000	fullstack engineer	201

Creating some more new features :

Designation : Learners with a CTC above their company's department average for the same experience get Flag 1, those with an equal CTC get Flag 2, and those with a lower CTC get Flag 1.

Tier : For the company's department, learners with a CTC above the department's average get Flag 1, those with an equal CTC get Flag 2, and those with a lower CTC get Flag 1.

Class : Learners with a CTC above their company average get Flag 1, those with an equal CTC get Flag 2, and those with a lower CTC get Flag 1.

In [44]:
```
## Getting the 5 point summary of CTC (mean, median, max, min, count etc) on the basis of Company, Job Posi

df_new = df.groupby(['company_hash', 'job_position', 'years_of_experience']).agg(
    mean_ctc=('ctc', 'mean'),
    median_ctc=('ctc', 'median'),
    max_ctc=('ctc', 'max'),
    min_ctc=('ctc', 'min'),
    count_ctc=('ctc', 'count')).reset_index()
df_new
```

Out[44]:

	company_hash	job_position	years_of_experience	mean_ctc	median_ctc	max_ctc	min_ctc	count_ctc
0	Others		4.0	4.500000e+05	450000.0	450000	450000	1
1	Others		6.0	4.200000e+05	420000.0	420000	420000	1
2	Others		7.0	3.500000e+05	350000.0	350000	350000	1
3	Others		14.0	7.000000e+05	700000.0	700000	700000	1
4	Others	Others	2.0	2.644000e+06	810000.0	10000000	110000	5
...	...	...	...	...	...	...	...	...
66307	zxztrtvuo	member of technical staff at nineleaps	9.0	1.200000e+06	1200000.0	1200000	1200000	1
66308	zxztrtvuo	other	5.0	4.416667e+05	450000.0	450000	400000	6
66309	zxztrtvuo	other	6.0	4.250000e+05	425000.0	450000	400000	4
66310	zxztrtvuo	software developer intern	9.0	1.200000e+06	1200000.0	1200000	1200000	1
66311	zxztrtvuo	software developer intern	10.0	1.200000e+06	1200000.0	1200000	1200000	1

66312 rows × 8 columns

In [45]:
```
df_merge = df.merge(df_new, on = ['company_hash','job_position','years_of_experience'], how = 'left')
df_merge.head()
```

Out[45]:

	company_hash	email_hash	orgyear	ctc	job_position	ctc_updated_ye
0	atrgxnnt xzaxv	6de0a4417d18ab14334c3f43397fc13b30c35149d70c05...	2016.0	1100000	other	202(
1	qtrxvzwt xzegwgbb rxbxnta	b0aaf1ac138b53cb6e039ba2c3d6604a250d02d5145c10...	2018.0	449999	fullstack engineer	201(
2	Others	4860c670bcd48fb96c02a4b0ae3608ae6fdd98176112e9...	2015.0	2000000	backend engineer	202(
3	ngpgutaxv	effdede7a2e7c2af664c8a31d9346385016128d66bbc58...	2017.0	700000	backend engineer	201(
4	qxen sqghu	6ff54e709262f55cb999a1c1db8436cb2055d8f79ab520...	2017.0	1400000	fullstack engineer	201(

In [46]:
```
## Creating some flags showing learners with CTC greater than the Average of their Company's department hav

def func(ctc, mean_ctc):
    """
    Assigns a flag based on the comparison of CTC with the mean CTC.

    Returns:
    int:
        - 1 if CTC is greater than the mean CTC.
        - 2 if CTC is equal to the mean CTC.
        - 3 if CTC is less than the mean CTC.
    """
    if ctc >  mean_ctc:
        return 1
    elif mean_ctc == ctc:
        return 2
    else:
        return 3
```

In [47]:
```
df_merge['tier'] = df_merge.apply(lambda x: func(x['ctc'], x['mean_ctc']), axis=1)
```

In [48]:
```
df_merge.head()
```

Out[48]:

	company_hash	email_hash	orgyear	ctc	job_position	ctc_updated_ye
0	atrgxnnt xzaxv	6de0a4417d18ab14334c3f43397fc13b30c35149d70c05...	2016.0	1100000	other	202(
1	qtrxvzwt xzegwgbb rxbxnta	b0aaf1ac138b53cb6e039ba2c3d6604a250d02d5145c10...	2018.0	449999	fullstack engineer	201(
2	Others	4860c670bcd48fb96c02a4b0ae3608ae6fdd98176112e9...	2015.0	2000000	backend engineer	202(
3	ngpgutaxv	effdede7a2e7c2af664c8a31d9346385016128d66bbc58...	2017.0	700000	backend engineer	201(
4	qxen sqghu	6ff54e709262f55cb999a1c1db8436cb2055d8f79ab520...	2017.0	1400000	fullstack engineer	201(

In [49]:
```
df_merge.drop(['mean_ctc','median_ctc','max_ctc','min_ctc','count_ctc'],axis = 1,inplace = True)
```

In [50]:
```
df_merge.head()
```

Out[50]:

	company_hash	email_hash	orgyear	ctc	job_position	ctc_updated_ye
0	atrgxnnt xzaxv	6de0a4417d18ab14334c3f43397fc13b30c35149d70c05...	2016.0	1100000	other	202(
1	qtrxvzwt xzegwgbb rxbxnta	b0aaf1ac138b53cb6e039ba2c3d6604a250d02d5145c10...	2018.0	449999	fullstack engineer	201(
2	Others	4860c670bcd48fb96c02a4b0ae3608ae6fdd98176112e9...	2015.0	2000000	backend engineer	202(
3	ngpgutaxv	effdede7a2e7c2af664c8a31d9346385016128d66bbc58...	2017.0	700000	backend engineer	201(
4	qxen sqghu	6ff54e709262f55cb999a1c1db8436cb2055d8f79ab520...	2017.0	1400000	fullstack engineer	201(

In [51]:
```
df_merge.rename({'tier' : 'designation'},axis = 1,inplace = True)
```

In [52]:
```
df_merge.head()
```

Out[52]:

	company_hash	email_hash	orgyear	ctc	job_position	ctc_updated_ye
0	atrgxnnt xzaxv	6de0a4417d18ab14334c3f43397fc13b30c35149d70c05...	2016.0	1100000	other	202(
1	qtrxvzwt xzegwgbb rxbxnta	b0aaf1ac138b53cb6e039ba2c3d6604a250d02d5145c10...	2018.0	449999	fullstack engineer	201(
2	Others	4860c670bcd48fb96c02a4b0ae3608ae6fdd98176112e9...	2015.0	2000000	backend engineer	202(
3	ngpgutaxv	effdede7a2e7c2af664c8a31d9346385016128d66bbc58...	2017.0	700000	backend engineer	201(
4	qxen sqghu	6ff54e709262f55cb999a1c1db8436cb2055d8f79ab520...	2017.0	1400000	fullstack engineer	201(

In [53]:
```
df_new1 = df.groupby(['company_hash', 'job_position']).agg(
    mean_ctc=('ctc', 'mean')
).reset_index()
df_new1
```

Out[53]:

	company_hash	job_position	mean_ctc
0	Others		4.800000e+05
1	Others	Others	3.681324e+06
2	Others	a group chat application	5.000000e+05
3	Others	abap developer	5.000000e+05
4	Others	administrative clerk	5.000000e+05
...	...	...	...
22716	zxztrtvuo	fullstack engineer	8.507273e+05
22717	zxztrtvuo	ios engineer	1.237500e+06
22718	zxztrtvuo	member of technical staff at nineleaps	1.200000e+06
22719	zxztrtvuo	other	4.350000e+05
22720	zxztrtvuo	software developer intern	1.200000e+06

22721 rows × 3 columns

In [54]:
```
## Doing above analysis at Company & Job Position level.flag Class with values [1,2,3]

df_merge1 = df_merge.merge(df_new1, on = ['company_hash','job_position'], how = 'left')
df_merge1.head()
```

Out[54]:

	company_hash	email_hash	orgyear	ctc	job_position	ctc_updated_ye
0	atrgxnnt xzaxv	6de0a4417d18ab14334c3f43397fc13b30c35149d70c05...	2016.0	1100000	other	202(
1	qtrxvzwt xzegwgbb rxbxnta	b0aaf1ac138b53cb6e039ba2c3d6604a250d02d5145c10...	2018.0	449999	fullstack engineer	201(
2	Others	4860c670bcd48fb96c02a4b0ae3608ae6fdd98176112e9...	2015.0	2000000	backend engineer	202(
3	ngpgutaxv	effdede7a2e7c2af664c8a31d9346385016128d66bbc58...	2017.0	700000	backend engineer	201(
4	qxen sqghu	6ff54e709262f55cb999a1c1db8436cb2055d8f79ab520...	2017.0	1400000	fullstack engineer	201(

In [55]:
```
df_merge1['class'] = df_merge1.apply(lambda x: func(x['ctc'], x['mean_ctc']), axis=1)
```

In [56]:
```
df_merge1.drop('mean_ctc',axis = 1,inplace = True)
```

In [57]:
```
df_merge1.head()
```

Out[57]:

	company_hash	email_hash	orgyear	ctc	job_position	ctc_updated_ye
0	atrgxnnt xzaxv	6de0a4417d18ab14334c3f43397fc13b30c35149d70c05...	2016.0	1100000	other	202(
1	qtrxvzwt xzegwgbb rxbxnta	b0aaf1ac138b53cb6e039ba2c3d6604a250d02d5145c10...	2018.0	449999	fullstack engineer	201(
2	Others	4860c670bcd48fb96c02a4b0ae3608ae6fdd98176112e9...	2015.0	2000000	backend engineer	202(
3	ngpgutaxv	effdede7a2e7c2af664c8a31d9346385016128d66bbc58...	2017.0	700000	backend engineer	201(
4	qxen sqghu	6ff54e709262f55cb999a1c1db8436cb2055d8f79ab520...	2017.0	1400000	fullstack engineer	201(

In [58]:
```
df_new2 = df.groupby('company_hash').agg(
    mean_ctc=('ctc', 'mean')
).reset_index()
df_new2
```

Out[58]:

	company_hash	mean_ctc
0	Others	2.599167e+06
1	a ntwyzgrgsxto	1.234688e+06
2	aaqxctz avnv owxtzwto vzvrjnxwo ucn rna	9.850000e+05
3	abwavnv ojontb	7.320000e+05
4	adw ntwyzgrgsj	1.234200e+06
...	...	...
3774	zxxn rna	9.014286e+05
3775	zxyxrtzn	6.887500e+05
3776	zxyxrtzn ntwyzgrgsxto	8.170000e+05
3777	zxzlvwvqn	1.739048e+06
3778	zxztrtvuo	1.172701e+06

3779 rows × 2 columns

In [59]:
```
#Repeating the same analysis at the Company level

df_merge2 = df_merge1.merge(df_new2, on = ['company_hash'], how = 'left')
df_merge2.head()
```

Out[59]:

	company_hash	email_hash	orgyear	ctc	job_position	ctc_updated_ye
0	atrgxnnt xzaxv	6de0a4417d18ab14334c3f43397fc13b30c35149d70c05...	2016.0	1100000	other	202(
1	qtrxvzwt xzegwgbb rxbxnta	b0aaf1ac138b53cb6e039ba2c3d6604a250d02d5145c10...	2018.0	449999	fullstack engineer	201(
2	Others	4860c670bcd48fb96c02a4b0ae3608ae6fdd98176112e9...	2015.0	2000000	backend engineer	202(
3	ngpgutaxv	effdede7a2e7c2af664c8a31d9346385016128d66bbc58...	2017.0	700000	backend engineer	201(
4	qxen sqghu	6ff54e709262f55cb999a1c1db8436cb2055d8f79ab520...	2017.0	1400000	fullstack engineer	201(

In [60]:
```
df_merge2['tier'] = df_merge2.apply(lambda x: func(x['ctc'], x['mean_ctc']), axis=1)
```

In [61]:
```
df_merge2.drop('mean_ctc',axis = 1,inplace = True)
```

In [62]:
```
df_merge2.head()
```

Out[62]:

	company_hash	email_hash	orgyear	ctc	job_position	ctc_updated_ye
0	atrgxnnt xzaxv	6de0a4417d18ab14334c3f43397fc13b30c35149d70c05...	2016.0	1100000	other	202(
1	qtrxvzwt xzegwgbb rxbxnta	b0aaf1ac138b53cb6e039ba2c3d6604a250d02d5145c10...	2018.0	449999	fullstack engineer	201
2	Others	4860c670bcd48fb96c02a4b0ae3608ae6fdd98176112e9...	2015.0	2000000	backend engineer	202(
3	ngpgutaxv	effdede7a2e7c2af664c8a31d9346385016128d66bbc58...	2017.0	700000	backend engineer	201
4	qxen sqghu	6ff54e709262f55cb999a1c1db8436cb2055d8f79ab520...	2017.0	1400000	fullstack engineer	201

In [63]:
```
## Top 10 employees (earning more than most of the employees in the company) - Tier 1

df_merge2[df_merge2['tier'] == 1].sort_values('ctc',ascending = False)[:11]
```

Out[63]:

	company_hash	email_hash	orgyear	ctc	job_position	ctc_up
72805	Others	29a71dd13adf6d2d497571a565bb3096cf66cb46cd1ece...	2015.0	1000150000	Others	
117579	obvqnuqxdwgb	5b4bed51797140db4ed52018a979db1e34cee49e27b488...	2018.0	255555555	fullstack engineer	
3300	Others	06d231f167701592a69cdd7d5c825a0f5b30f0347a4078...	2021.0	250000000	Others	
93783	qvzaonva	fa964c3863d58f39cfa98248869a7aa41fafd4edd106cb...	2012.0	200000000	support engineer	
93784	Others	005f352591238b78024ed320789c30e7815f0eea52ed98...	2020.0	200000000	Others	
8074	zgzt	232e5b21da629a8acb3bdef7b49e57b28fc3bf64ae2687...	2023.0	200000000	engineering leadership	
21289	Others	7e447c2a4390a212cb825a72991d04251b2d943a1daf8d...	2016.0	200000000	Others	
8164	fxuqg rxbxnta	4926be010b57f52aefce6c9e2f01d43c3d4fc9e6860b25...	2017.0	200000000	support engineer	
21283	Others	dbfd099248b99c95261d19ab8f40c0b7b8aacae99d09c3...	2018.0	200000000	fullstack engineer	
101571	ytqt ntwyzgrgsxto	98a90272cbba6e6e9ca94981824f3465f3d34567cb065f...	2015.0	200000000	data analyst	
21240	Others	e8e947453901b8540be1ea26003ddaf5a27e327bdf7c8f...	2018.0	200000000	Others	

In [64]:
```
## Top 10 employees of data science in each company earning more than their peers - Class 1

df_tier1 = df_merge2[df_merge2['tier'] == 1]
df_tier_1_class1 = df_tier1[df_tier1['class'] == 1]
df_ds = df_tier_1_class1[df_tier_1_class1['job_position'] == 'data scientist']

dftop10_ds = (df_ds.groupby('company_hash')
              .apply(lambda x: x.nlargest(10, 'ctc'))
              .reset_index(drop=True))
```

```
dftop10_ds
```

Out[64]:

	company_hash	email_hash	orgyear	ctc	job_position	ctc_updat
0	Others	be8afd4a4e5cedb6bfbabbc88f15e756d69c3f2491d02a...	2021.0	200000000	data scientist	
1	Others	72ed7ced98573f71c8f95bc8b75aac4f0677e8872c6bec...	2019.0	199800000	data scientist	
2	Others	ee8dd42d6ea8365909147d861c7978d19f727a8075ba96...	2020.0	102500000	data scientist	
3	Others	2e1d492bc09bfe0d4cc9757a9c63a296c1527af1c8ecc8...	2021.0	100000000	data scientist	
4	Others	e7722fb701c61e5cad82c39ee8bf3debe160d429b72c64...	2015.0	100000000	data scientist	
...	...	...	...	...	...	
930	zxwt vwnxbxkt	41b39db3bc359c98a43825f654f6b4fa8680dc99207616...	2019.0	3900000	data scientist	
931	zxxn ntwyzgrgsxto	56bc8f0eb4db04e459bf41f51c12b3bbb9c3595d8a172d...	2007.0	5500000	data scientist	
932	zxzlvwvqn	2937acfa6802f83ff11ddbd3de1997b686107dad0c2b5d...	2019.0	1900000	data scientist	
933	zxzlvwvqn	2937acfa6802f83ff11ddbd3de1997b686107dad0c2b5d...	2019.0	1900000	data scientist	
934	zxztrtvuo	1cd0a52ed52dae24d605d9cdc8536499c10ce62bfb070f...	2014.0	2250000	data scientist	

935 rows × 10 columns

In [65]:
```
## Bottom 10 employees of data science in each company earning less than their peers – Class 3

df_tier1 = df_merge2[df_merge2['tier'] == 1]
df_tier_1_class3 = df_tier1[df_tier1['class'] == 3]
df_ds = df_tier_1_class3[df_tier_1_class3['job_position'] == 'data scientist']

dfbottom10_ds = (df_ds.groupby('company_hash')
              .apply(lambda x: x.nsmallest(10, 'ctc'))
              .reset_index(drop=True))

dfbottom10_ds
```

Out[65]:

	company_hash	email_hash	orgyear	ctc	job_position	ctc_updated
0	amo mvzp	fdca92d09d4bb96383be5cd8438d83578e6fa7d09cbee6...	2012.0	2000000	data scientist	
1	amo mvzp	13ed1c77a6961a9683ad363845c62186d052dc0d1f48bc...	2003.0	2100000	data scientist	
2	amo mvzp	cb28bd98e365ca53860af89cd3c96df77d4dc41289899b...	2017.0	2100000	data scientist	
3	amo mvzp	0177202b15c66938ea8a061468928e036a7af0d0c1e882...	2017.0	2100000	data scientist	
4	aqag mvzsvrgqt	dd529581b56bafc7f54cc7236e5f0c6ab793cddcae2618...	2018.0	1300000	data scientist	
...	...	...	...	...	...	
317	zxzlvwvqn	e82cbd0e93c2c0bd0cdc818447ab3dde19b577598d8e9b...	2011.0	1820000	data scientist	
318	zxztrtvuo	f678c67bee8cad9370f6aaf4f4cc22ffd417fd753663c6...	2019.0	1250000	data scientist	
319	zxztrtvuo	f678c67bee8cad9370f6aaf4f4cc22ffd417fd753663c6...	2019.0	1250000	data scientist	
320	zxztrtvuo	3027ca561b65f99da2f65bf3d85c6bb5d5687c67e69e89...	2018.0	1370000	data scientist	
321	zxztrtvuo	10d566c5fca40ffe1d133b79594d071880711ef480da9f...	2017.0	1400000	data scientist	

322 rows × 10 columns

In [66]:
```
## Bottom 10 employees (earning less than most of the employees in the company)– Tier 3

df_merge2[df_merge2['tier'] == 3].sort_values('ctc',ascending = True)[:10]
```

Out[66]:

	company_hash	email_hash	orgyear	ctc	job_position	ctc_updated_
135356	xzntqcxtfmxn	3505b02549ebe2c95840ac6f0a35561a3b4cbe4b79cdb1...	2014.0	2	backend engineer	2
118179	xzntqcxtfmxn	f2b58aeed3c074652de2cfd3c0717a5d21d6fbcf342a78...	2013.0	6	some random title	2
114110	xzntqcxtfmxn	23ad96d6b6f1ecf554a52f6e9b61677c7d73d8a409a143...	2013.0	14	some random title	2
184805	Others	b8a0bb340583936b5a7923947e9aec21add5ebc50cd60b...	2016.0	15	Others	2
183664	Others	75357254a31f133e2d3870057922feddeba82b88056a07...	2019.0	16	Others	2
54808	Others	8786759b95d673466e94f62f1b15e4f8c6bd7de6164074...	2020.0	24	other	2(
91523	Others	512f761579fb116e215cabc9821c7f81153f0763e16018...	2016.0	25	android engineer	2
116891	Others	f7e5e788676100d7c4146740ada9e2f8974defc01f571d...	2022.0	200	other	2(
166280	Others	c411a6917058b50f44d7c62751be9b232155b23211de4c...	2013.0	300	database administrator	2
82002	Others	edcfb902656b736e1f35863298706d9d34ee795b7ed85a...	2018.0	500	cofounder	2

In [67]:
```
## Top 10 companies (based on their CTC)

df_merge2.groupby('company_hash').agg('mean','ctc').sort_values('ctc',ascending = False)[:10].reset_index()
```

Out[67]:

	company_hash	orgyear	ctc	ctc_updated_year	years_of_experience	designation	class	tier
0	xzaxvmhrro	2016.285714	5.374143e+07	2020.142857	8.714286	2.000000	1.857143	2.428571
1	obvqnuqxdwgb	2017.666667	4.395259e+07	2019.166667	7.333333	2.333333	2.500000	2.666667
2	ho tzsxzttqxzs wgbuvzj	2018.714286	4.348714e+07	2021.428571	6.285714	2.000000	2.000000	2.428571
3	ouxwtltn rxbxnta	2012.200000	4.058400e+07	2020.000000	12.800000	2.000000	2.000000	2.600000
4	ovxzn sgmvxz srvoo xzaxv ucnrna	2017.200000	4.056800e+07	2020.200000	7.800000	2.000000	2.000000	2.600000
5	axowgctq agrrvq	2015.400000	4.054400e+07	2018.800000	9.600000	2.000000	2.000000	2.600000
6	exqonoghqwt	2017.000000	4.038480e+07	2020.000000	8.000000	2.000000	2.000000	2.600000
7	uyxrxuo xzaxv	2014.400000	3.770400e+07	2020.200000	10.600000	2.000000	2.200000	2.600000
8	fgqraihvzn rrw	2017.166667	3.495833e+07	2018.833333	7.833333	1.833333	1.833333	2.333333
9	egdwgzz	2017.000000	3.423250e+07	2020.166667	8.000000	2.000000	2.000000	2.666667

In [68]:
```
## Top 2 positions in every company (based on their CTC)
```

In [69]:
```
df_merge2.groupby(['company_hash', 'job_position'])['ctc'].mean().reset_index().sort_values(['company_hash'
```

Out[69]:

	company_hash	job_position	ctc
65	Others	data entry	1.000000e+08
267	Others	telar	1.000000e+08
282	a ntwyzgrgsxto	other	3.716667e+06
279	a ntwyzgrgsxto	backend engineer	9.000000e+05
286	aaqxctz avnv owxtzwto vzvrjnxwo ucn rna	other	3.600000e+06
...	...	...	...
22696	zxyxrtzn ntwyzgrgsxto	backend engineer	9.925000e+05
22708	zxzlvwvqn	product manager	4.900000e+06
22706	zxzlvwvqn	fullstack engineer	2.181250e+06
22710	zxztrtvuo	backend architect	3.750000e+06
22713	zxztrtvuo	devops engineer	2.700000e+06

7435 rows × 3 columns

In [70]:
```
## Top 10 employees in each company - X department - having 5/6/7 years of experience earning more than the

df_filered1 = df_merge2[((df_merge2['years_of_experience'] == 5) | (df_merge2['years_of_experience'] == 6)
df_filered1.groupby('company_hash')['ctc'].mean().reset_index().sort_values(by='ctc', ascending=False).grou
```

Out[70]:

	company_hash	ctc
750	obvqnuqxdwgb	255555555.0
390	ho tzsxzttqxzs wgbuvzj	200000000.0
635	nqvctrnqxvzsrt	200000000.0
976	qtwpgzojo ntwy rvmo ucn rna	200000000.0
876	ovxzn sgmvxz srvoo xzaxv ucnrna	200000000.0
...	...	...
451	jvzatd	100000.0
1167	tuxw	90000.0
804	onvnt evqb	75000.0
308	ftmvrg	70000.0
1683	xzaxvzv hzxctqoxnj mrggbxzsngz	66000.0

1910 rows × 2 columns

In [71]:
```
df_merge2.head()
```

Out[71]:

	company_hash	email_hash	orgyear	ctc	job_position	ctc_updated_ye
0	atrgxnnt xzaxv	6de0a4417d18ab14334c3f43397fc13b30c35149d70c05...	2016.0	1100000	other	202(
1	qtrxvzwt xzegwgbb rxbxnta	b0aaf1ac138b53cb6e039ba2c3d6604a250d02d5145c10...	2018.0	449999	fullstack engineer	201
2	Others	4860c670bcd48fb96c02a4b0ae3608ae6fdd98176112e9...	2015.0	2000000	backend engineer	202(
3	ngpgutaxv	effdede7a2e7c2af664c8a31d9346385016128d66bbc58...	2017.0	700000	backend engineer	201
4	qxen sqghu	6ff54e709262f55cb999a1c1db8436cb2055d8f79ab520...	2017.0	1400000	fullstack engineer	201

In [72]:
```
df_final = df_merge2.copy()
```

In [73]:
```
df_final.head()
```

Out[73]:

	company_hash	email_hash	orgyear	ctc	job_position	ctc_updated_ye
0	atrgxnnt xzaxv	6de0a4417d18ab14334c3f43397fc13b30c35149d70c05...	2016.0	1100000	other	202(
1	qtrxvzwt xzegwgbb rxbxnta	b0aaf1ac138b53cb6e039ba2c3d6604a250d02d5145c10...	2018.0	449999	fullstack engineer	201
2	Others	4860c670bcd48fb96c02a4b0ae3608ae6fdd98176112e9...	2015.0	2000000	backend engineer	202(
3	ngpgutaxv	effdede7a2e7c2af664c8a31d9346385016128d66bbc58...	2017.0	700000	backend engineer	201
4	qxen sqghu	6ff54e709262f55cb999a1c1db8436cb2055d8f79ab520...	2017.0	1400000	fullstack engineer	201

In [74]:
```
df_final['years_of_experience'].max(),df_final['years_of_experience'].min()
```

Out[74]: (2025.0, 0.0)

In [75]:
```
df_exp = df_final['years_of_experience'].value_counts().reset_index()
df_exp
```

Out[75]:

	years_of_experience	count
0	7.0	25257
1	6.0	23419
2	8.0	23225
3	9.0	23025
4	10.0	20626
...	...	...
65	1819.0	1
66	1816.0	1
67	1817.0	1
68	44.0	1
69	1825.0	1

70 rows × 2 columns

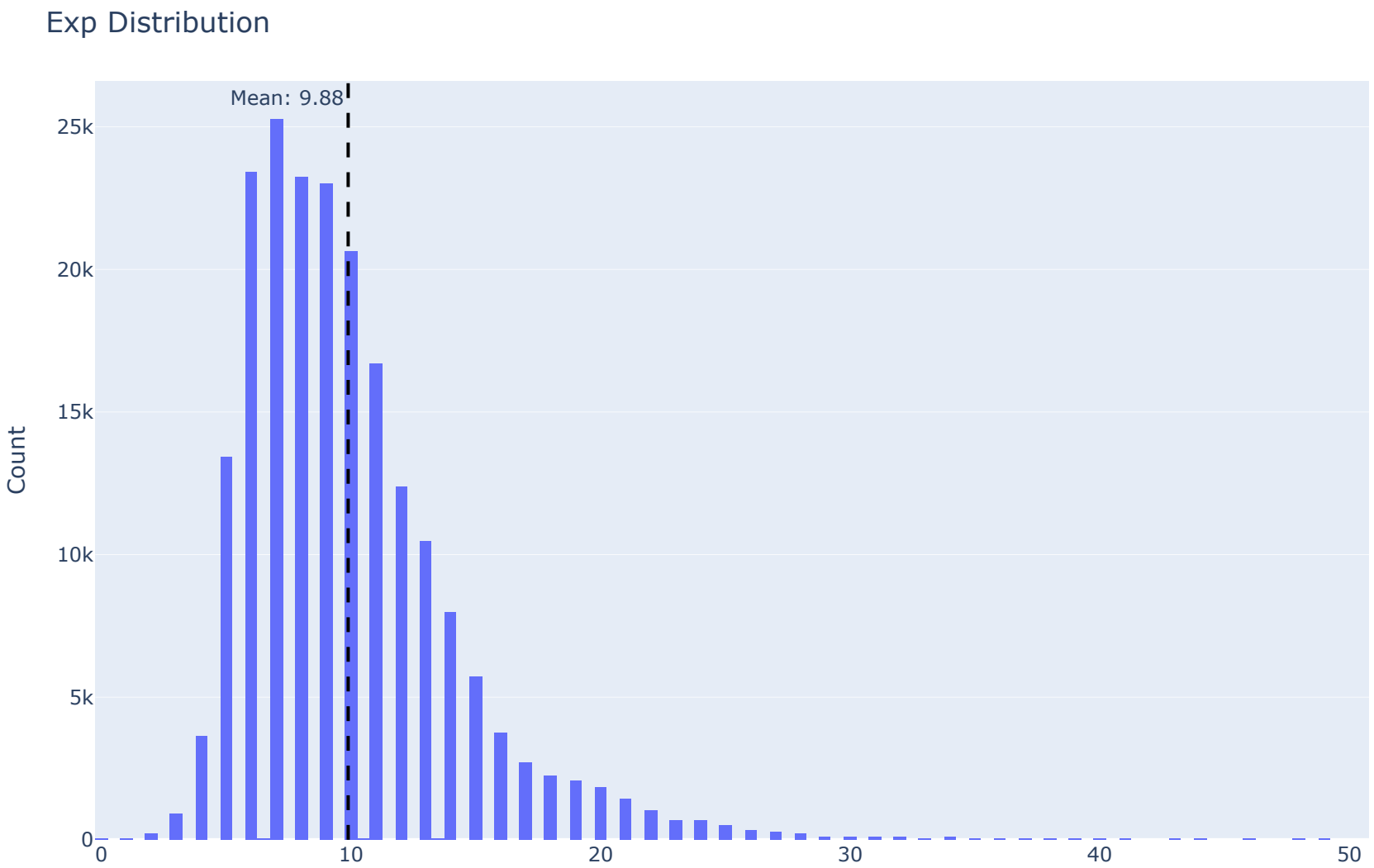
In [76]:
```
## Dropping data points with more than 60 or unusual years of experience

df_final = df_final[~(df_final['years_of_experience'] > 60)]
```

In [77]:
```
df_final.head()
```

Out[77]:

	company_hash	email_hash	orgyear	ctc	job_position	ctc_updated_ye
0	atrgxnnt xzaxv	6de0a4417d18ab14334c3f43397fc13b30c35149d70c05...	2016.0	1100000	other	202(
1	qtrxvzwt xzegwgbb rxbxnta	b0aaf1ac138b53cb6e039ba2c3d6604a250d02d5145c10...	2018.0	449999	fullstack engineer	201
2	Others	4860c670bcd48fb96c02a4b0ae3608ae6fdd98176112e9...	2015.0	2000000	backend engineer	202(
3	ngpgutaxv	effdede7a2e7c2af664c8a31d9346385016128d66bbc58...	2017.0	700000	backend engineer	201
4	qxen sqghu	6ff54e709262f55cb999a1c1db8436cb2055d8f79ab520...	2017.0	1400000	fullstack engineer	201

In [78]:
```
fig_2 = px.histogram(df_final, x = 'years_of_experience')
fig_2.update_layout(
    width=1000,
    height=600,
    title="Exp Distribution",
    xaxis_title="years_of_experience",
    yaxis_title="Count"
)
mean_val = df_final['years_of_experience'].mean()
median_val = df_final['years_of_experience'].median()
fig_2.add_vline(x=mean_val, line_dash="dash", line_color="black", annotation_text=f"Mean: {mean_val:.2f}",
fig_2.show()
```



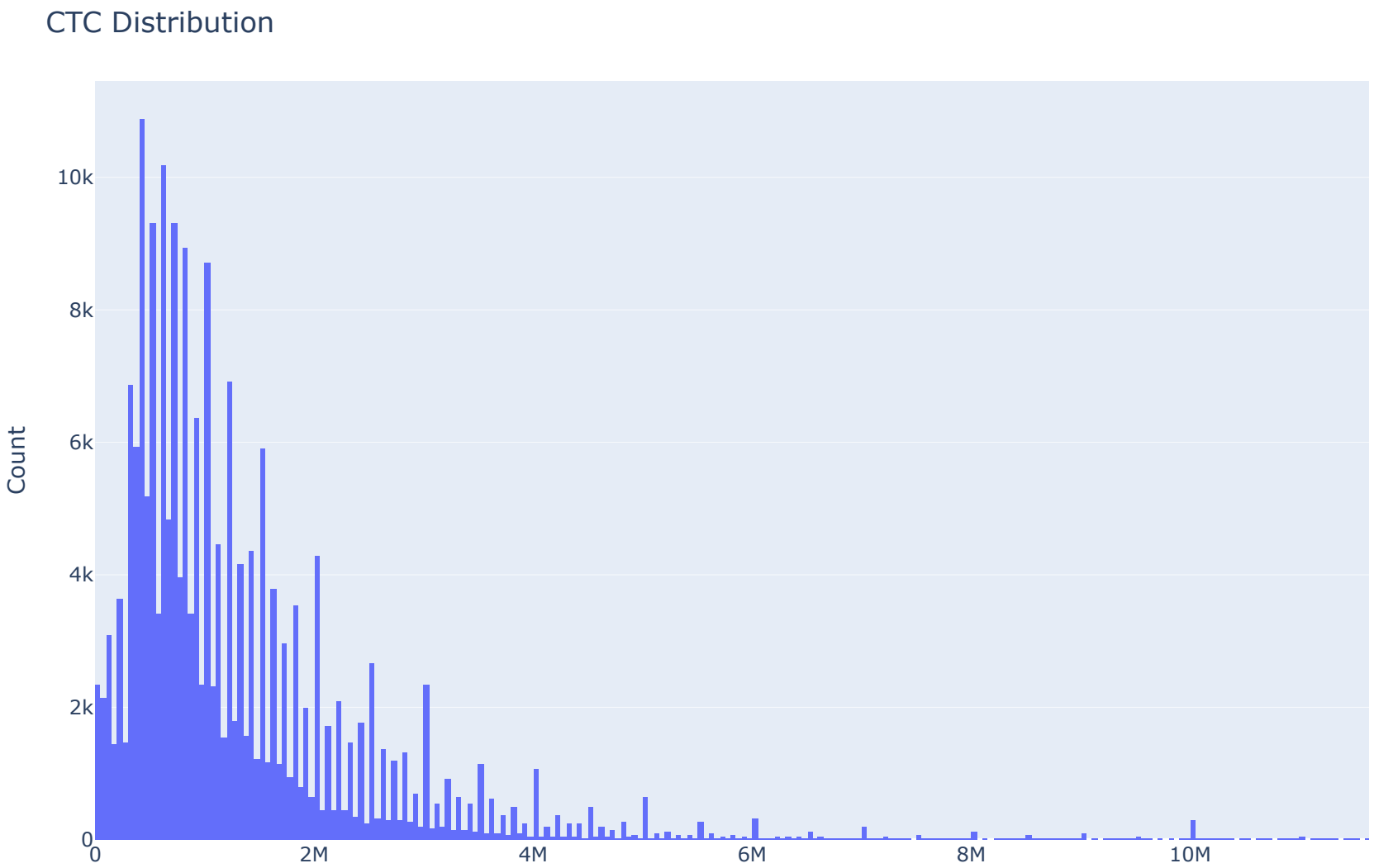
years of experience plot is right skewed with mean value of 9.88.

In [80]:
```
np.percentile(df_final['ctc'],1),np.percentile(df_final['ctc'],99)
```

Out[80]: (37000.0, 12600000.0)

In [81]:
```
df_final['ctc'] = df_final['ctc'].clip(lower = np.percentile(df_final['ctc'],1), upper = np.percentile(df_f
```

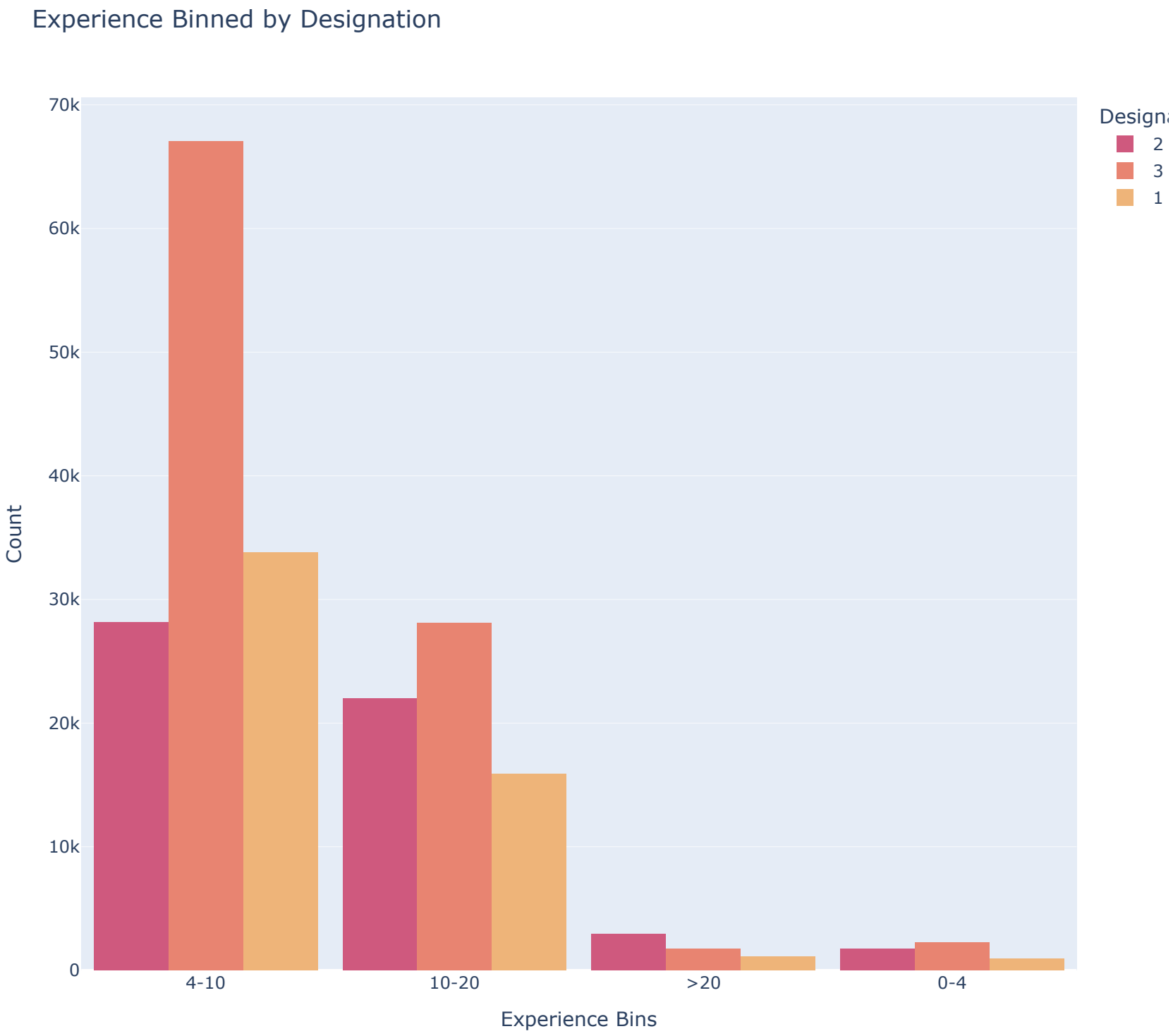
In [82]:
```
fig_1 = px.histogram(df_final, x = 'ctc')

fig_1.update_layout(
    width=1000,
    height=600,
    title="CTC Distribution",
    xaxis_title="CTC",
    yaxis_title="Count"
)

fig_1.show()
```



CTC Distribution plot is heavily right skewed meaning some high values of CTC creating a long tail.

In [84]:
```
df_final['orgyear'] = df_final['orgyear'].clip(lower = np.percentile(df_final['orgyear'],1), upper = np.per
```

In [85]:
```
bins = [0, 4, 10, 20, float("inf")]
labels = ["0-4", "4-10", "10-20", ">20"]
df_final["exp_bin"] = pd.cut(df_final['years_of_experience'],
                      labels = labels,
                      bins = bins)

fig = px.histogram(df_final,
                   x='exp_bin',
                   color='designation',
                   barmode='group',
                   title="Experience Binned by Designation",
                   color_discrete_sequence=px.colors.diverging.Temps_r
                  )

fig.update_layout(
    width=900,
    height=800,
    xaxis_title="Experience Bins",
    yaxis_title="Count",
    legend_title="Designation",
    bargap=0.1
)

fig.show()
```



From the above graph we can clearly see that most employee comes under the 4-10 years of experience range with designation 3

In [87]:
```
df_final.head()
```

Out[87]:

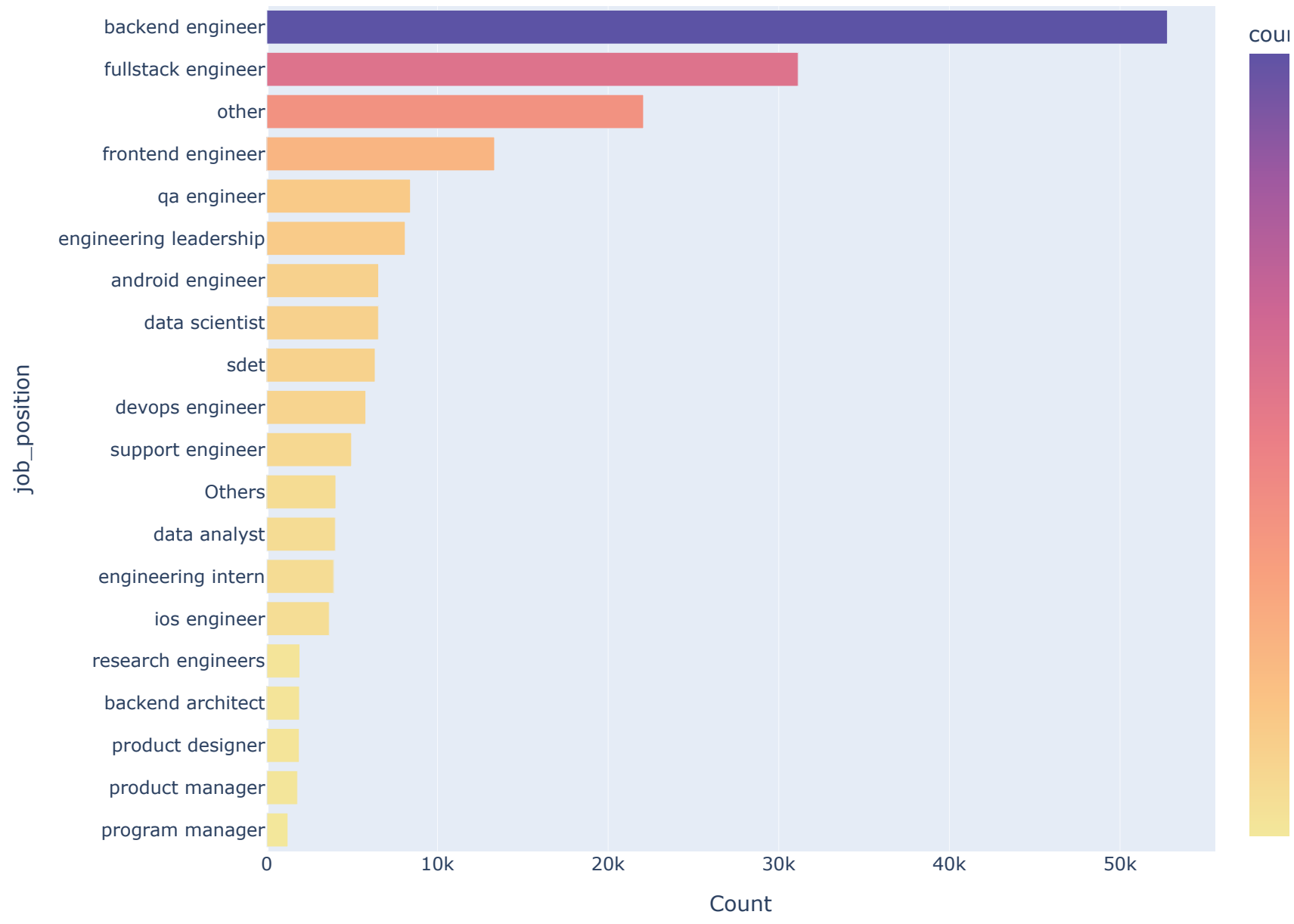
	company_hash	email_hash	orgyear	ctc	job_position	ctc_updated_ye
0	atrgxnnt xzaxv	6de0a4417d18ab14334c3f43397fc13b30c35149d70c05...	2016.0	1100000	other	202(
1	qtrxvzwt xzegwgbb rxbxnta	b0aaf1ac138b53cb6e039ba2c3d6604a250d02d5145c10...	2018.0	449999	fullstack engineer	201
2	Others	4860c670bcd48fb96c02a4b0ae3608ae6fdd98176112e9...	2015.0	2000000	backend engineer	202(
3	ngpgutaxv	effdede7a2e7c2af664c8a31d9346385016128d66bbc58...	2017.0	700000	backend engineer	201
4	qxen sqghu	6ff54e709262f55cb999a1c1db8436cb2055d8f79ab520...	2017.0	1400000	fullstack engineer	201

In [88]:
```
df_final["job_position"].value_counts()
```

Out[88]:
```
job_position
backend engineer                          52777
fullstack engineer                        31132
other                                     22072
frontend engineer                         13343
qa engineer                                8412
                                          ...
research assistant  software engineer         1
trainee decision scientist                    1
data specialist                               1
big data developer                            1
web  ui designer                              1
Name: count, Length: 856, dtype: int64
```

In [89]:
```
job = df_final["job_position"].value_counts().reset_index().sort_values(by ='count', ascending = True).tail
```

In [90]:
```
px1 = px.bar(job, x="count", y="job_position",color='count', color_continuous_scale='sunset')
px1.update_layout(
    width=900,
    height=700,
    xaxis_title="Count",
    yaxis_title="job_position",
)

px1.show()
```

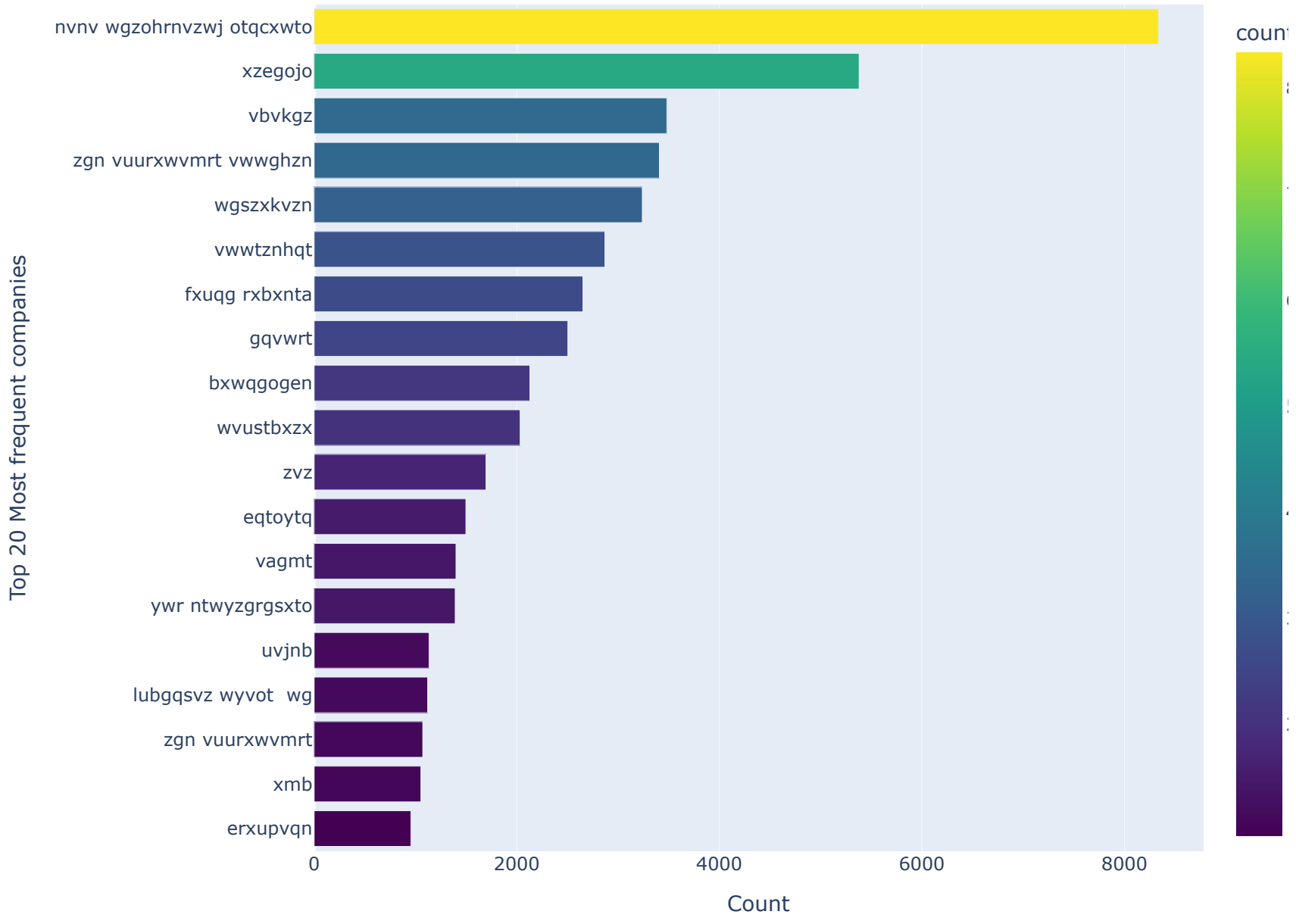


Plotted horizontal bar graph for job positions vs count.

Backend Engineer comes out at top with highest count

In [92]:
```
c_hash = df_final["company_hash"].value_counts().reset_index().sort_values(by ='count', ascending = True).t
```

In [93]:
```
px2 = px.bar(c_hash[~(c_hash['company_hash'] == 'Others')], x="count", y="company_hash",color='count', colo
px2.update_layout(
    width=900,
    height=700,
    xaxis_title="Count",
    yaxis_title="Top 20 Most frequent companies",
)

px2.show()
```



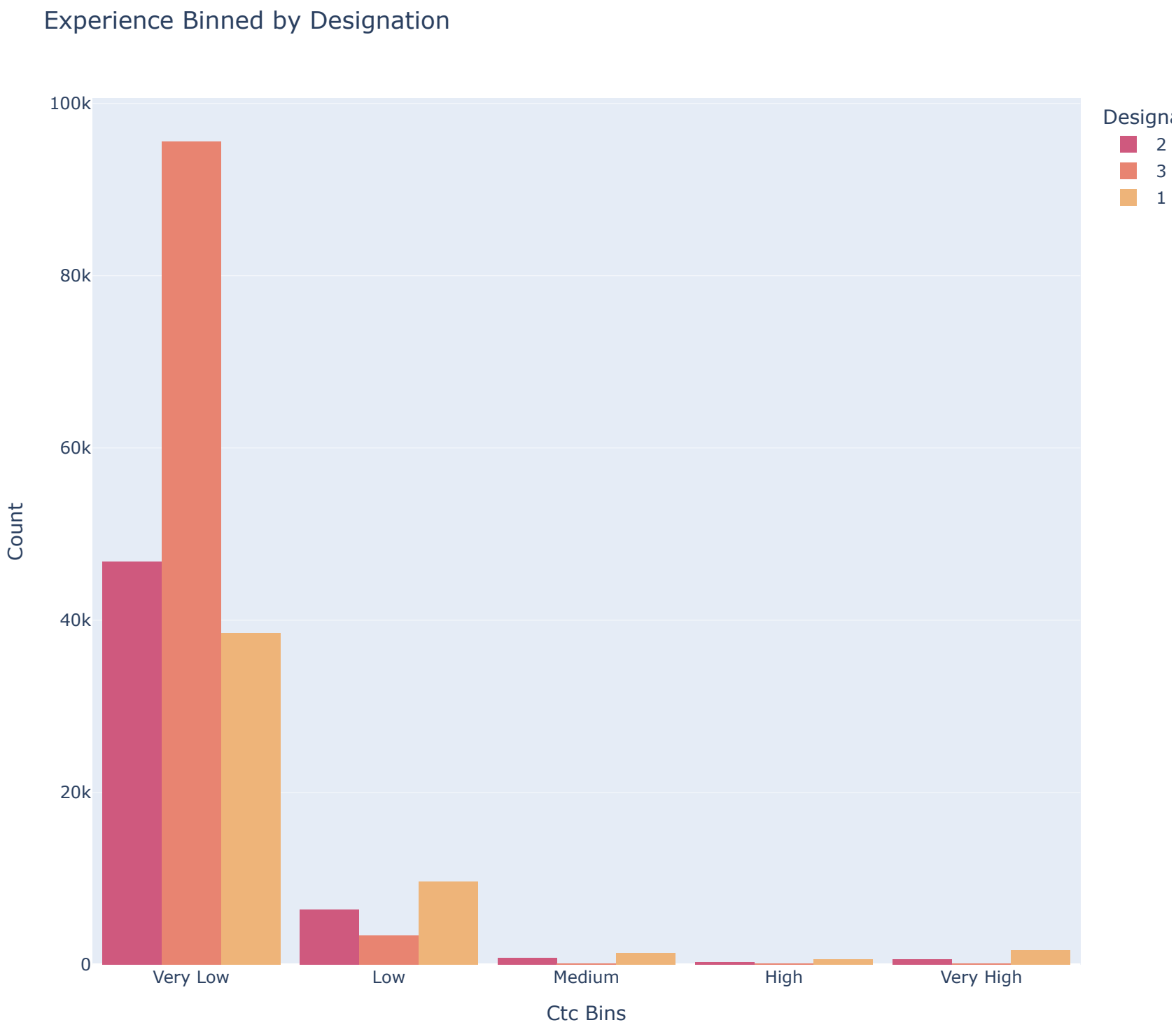
Plotted horizontal bar graph for company vs count.

Company names are hashed

In [95]:
```
df_final['ctc'].max(),df_final['ctc'].min()
```

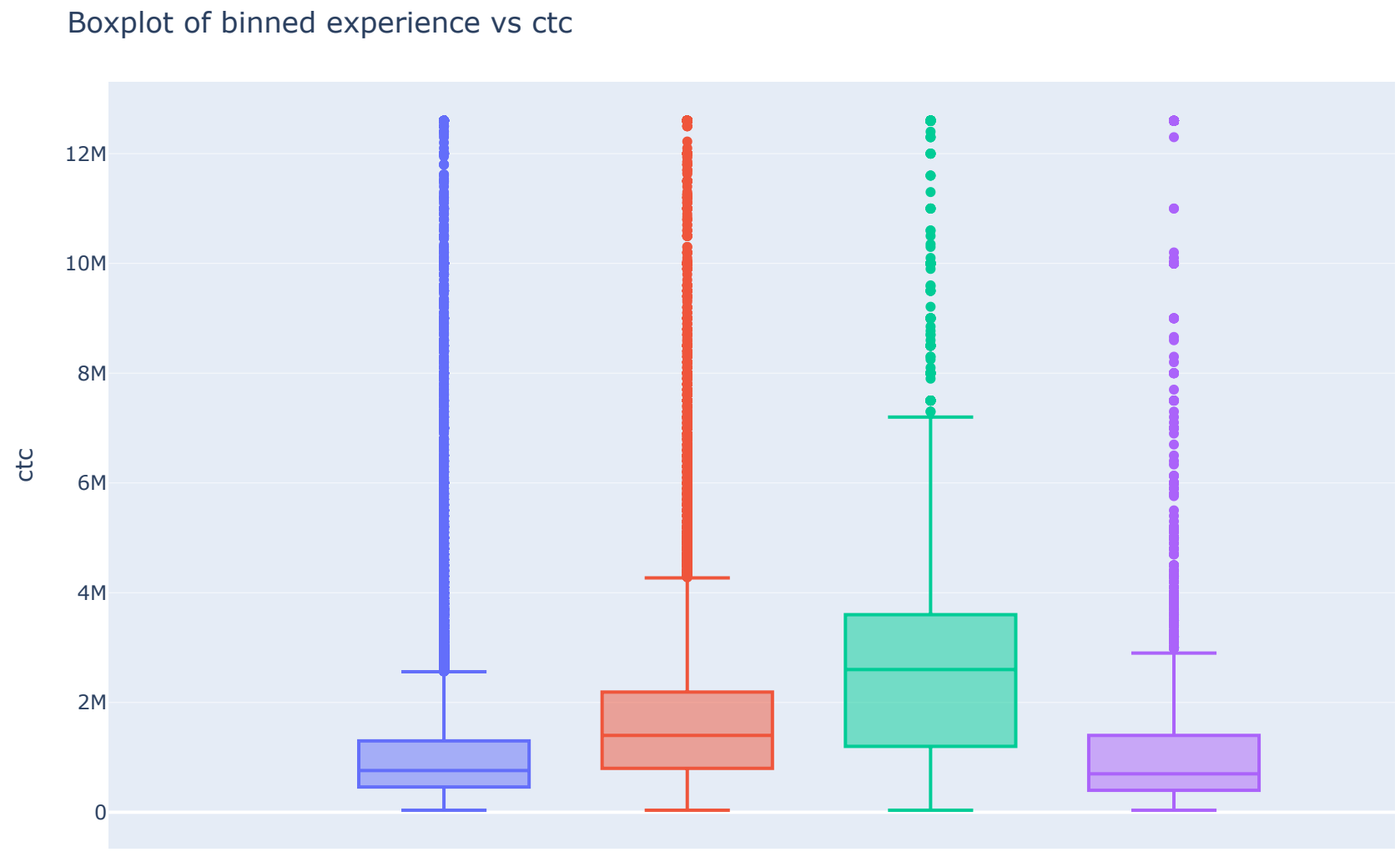
Out[95]:
```
(12600000, 37000)
```

In [96]:
```
min_ctc = 37000
max_ctc = 12600000

bins = pd.cut(df_final['ctc'], bins=5, labels=['Very Low', 'Low', 'Medium', 'High', 'Very High'])

df_final['ctc_category'] = bins

fig3 = px.histogram(df_final,
                    x='ctc_category',
                    color='designation',
                    barmode='group',
                    title="Experience Binned by Designation",
                    color_discrete_sequence=px.colors.diverging.Temps_r
                   )

fig3.update_layout(
    width=900,
    height=800,
    xaxis_title="Ctc Bins",
    yaxis_title="Count",
    legend_title="Designation",
    bargap=0.1
)

fig3.show()
```



We can see a high number of learners with ctc in very low range and the same pattern we have noticed in the above histogram as well

In [98]:
```
fig_4 = px.box(df_final, y="ctc", color = "exp_bin")

fig_4.update_layout(
    width=1100,
    height=600,
    xaxis_title="experience(in years)",
    yaxis_title="ctc",
    legend_title="years_of_experience",
    title = 'Boxplot of binned experience vs ctc'
)

fig_4.show()
```

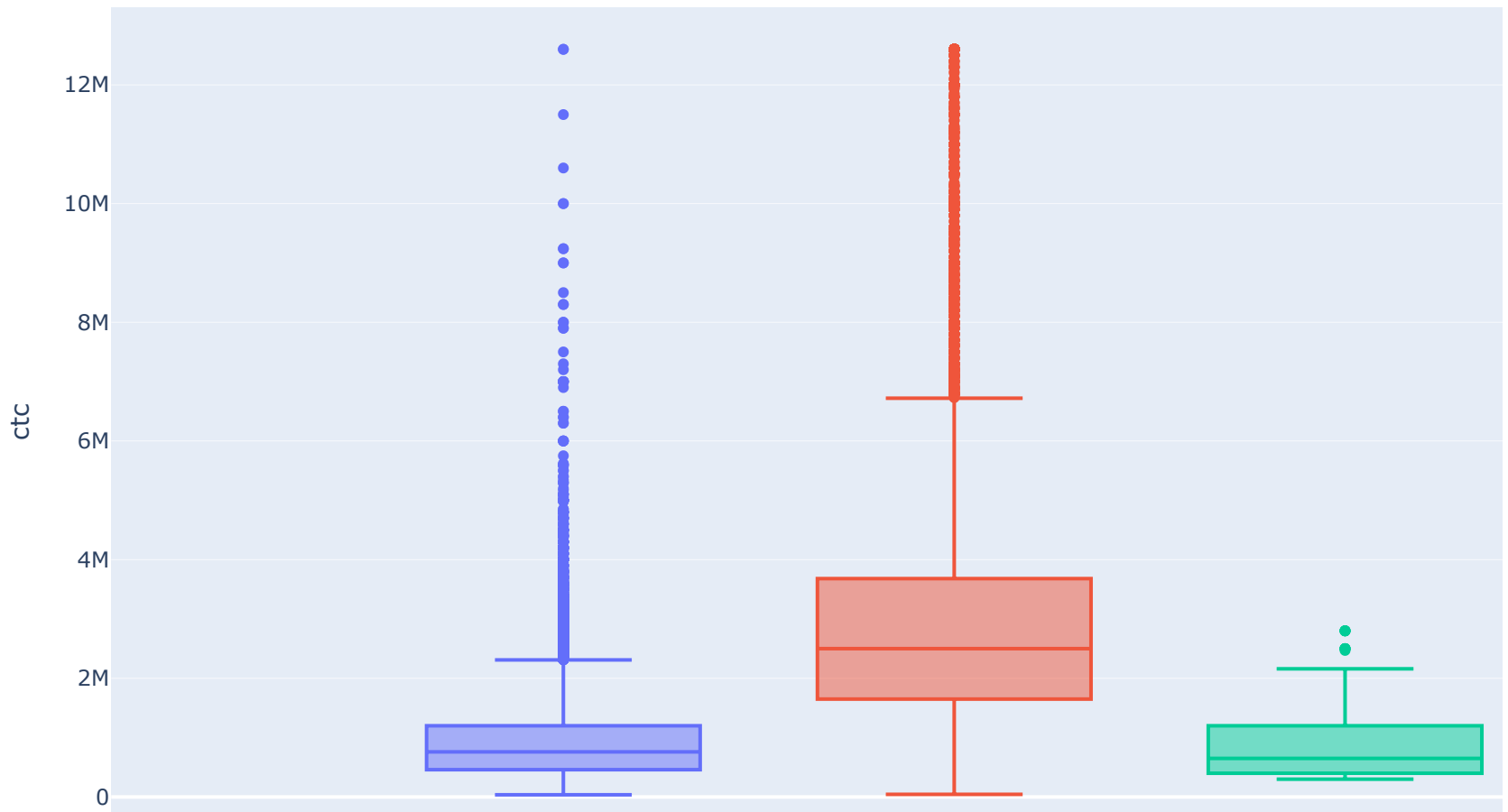


Created box plot for different years of experience bins.

Learners with more than 20 years of experience have highest range.

In [100…

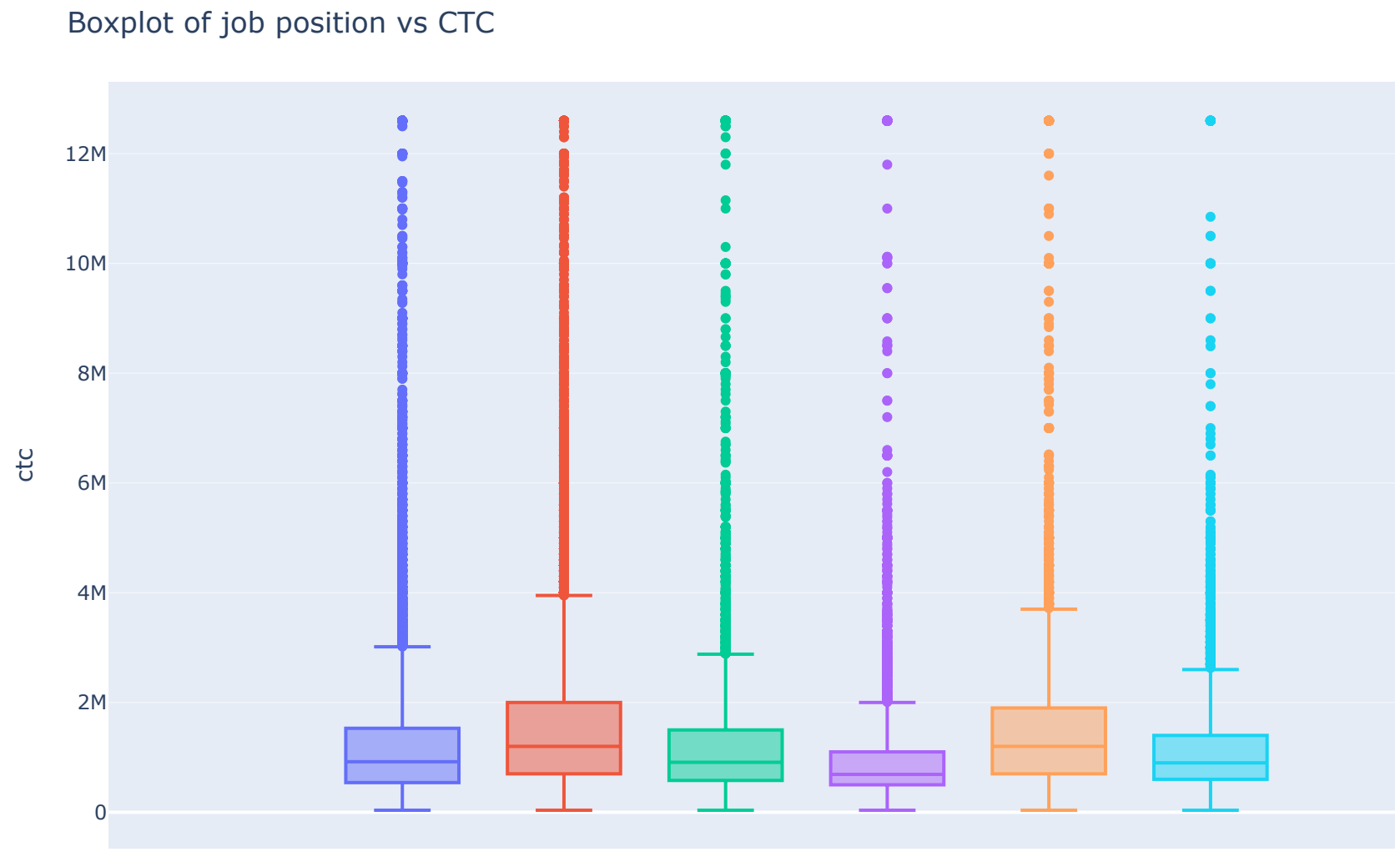
```
fig_5 = px.box(df_final, y="ctc", color = "tier")

fig_5.update_layout(
    width=1100,
    height=600,
    xaxis_title="tier",
    yaxis_title="ctc",
    legend_title="tier",
)

fig_5.show()
```



Box plot for tier 1,2 and 3

In [102…

```
df_job_pos = df_final[(df_final["job_position"] == 'backend engineer') | (df_final["job_position"] == 'full

fig_5 = px.box(df_job_pos, y="ctc", color = "job_position")

fig_5.update_layout(
    width=1100,
    height=600,
    xaxis_title="job_position",
    yaxis_title="ctc",
    legend_title="job_position",
    title='Boxplot of job position vs CTC'
)

fig_5.show()
```



Box Plot for few selected job positions with highest frequency.

In [104…

```
df_job_exp = df_final[((df_final["job_position"] == 'backend engineer') | (df_final["job_position"] == 'ful
df_grouped = df_job_exp.groupby(['job_position','years_of_experience'])['ctc'].mean().reset_index()

fig_6 = px.line(df_grouped,
                x="years_of_experience",
                y="ctc",
                color="job_position",
                title="CTC vs Years of Experience by Job Position")

fig_6.update_layout(
    width=1100,
    height=600,
    xaxis=dict(showgrid=True, gridcolor="green"),
    yaxis=dict(showgrid=True, gridcolor="green"),
)

fig_6.show()
```

Line Graph for the above job positions, comparing mean CTC with years of experience.

The graph exhibits anomalies, likely due to data quality issues.

In [106…
```
## Dropping Uncecessary Columns that do not contribute to clustering

df_final.drop(columns = {'company_hash','email_hash','orgyear','ctc_updated_year','exp_bin','ctc_category'}
```

In [107…
```
df_final.head()
```

Out[107…

	ctc	job_position	years_of_experience	designation	class	tier
0	1100000	other	9.0	2	1	3
1	449999	fullstack engineer	7.0	3	3	3
2	2000000	backend engineer	10.0	1	1	3
3	700000	backend engineer	8.0	3	3	3
4	1400000	fullstack engineer	8.0	2	1	1

In [108…
```
df_final.drop('job_position',axis = 1, inplace = True)
```

In [109…
```
df_final.shape
```

Out[109… (205623, 5)

Scaling

In [111…
```
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
from sklearn.manifold import TSNE
from sklearn.cluster import KMeans
from sklearn.mixture import GaussianMixture
import scipy.cluster.hierarchy as sch
from sklearn.cluster import AgglomerativeClustering
from sklearn.cluster import DBSCAN
```

In [112…
```
scaler = StandardScaler()

df_final_scaled = scaler.fit_transform(df_final)
```

In [113…
```
df_final_scaled
```

Out[113…
```
array([[-0.18800619, -0.20854201, -0.27944008, -1.5087658 ,  0.5153169 ],
       [-0.57816162, -0.68083909,  0.93296065,  0.71210627,  0.5153169 ],
       [ 0.3522082 ,  0.02760654, -1.49184082, -1.5087658 ,  0.5153169 ],
       ...,
       [-0.42810147, -1.38928471, -1.49184082, -1.5087658 ,  0.5153169 ],
       [ 2.21294666, -0.91698763,  0.93296065,  0.71210627,  0.5153169 ],
       [-0.10397284,  0.26375508, -1.49184082, -1.5087658 ,  0.5153169 ]])
```

In [114…
```
# pip install umap-learn
```

In [115…
```
import umap
```

In [116…
```
import umap.umap_ as umap
```

In [117…
```
pca_2d = PCA(n_components = 2)
df_embed = pca_2d.fit_transform(df_final_scaled)
pca_plot = px.scatter(x = df_embed[:,0], y =df_embed[:,1])
pca_plot.update_layout(
    width=1100,
    height=600)
```

Applied Principal Component Analysis (PCA) to the entire dataset, reducing the dimensionality of all features to two principal components.

In [119…
```
tsne_2d = TSNE(n_components = 2, perplexity = 50)
df_embed_tsne = tsne_2d.fit_transform(df_final_scaled)
tsne_plot = px.scatter(x = df_embed_tsne[:,0], y = df_embed_tsne[:,1])
tsne_plot.update_layout(
    width=1100,
    height=600)
```

>

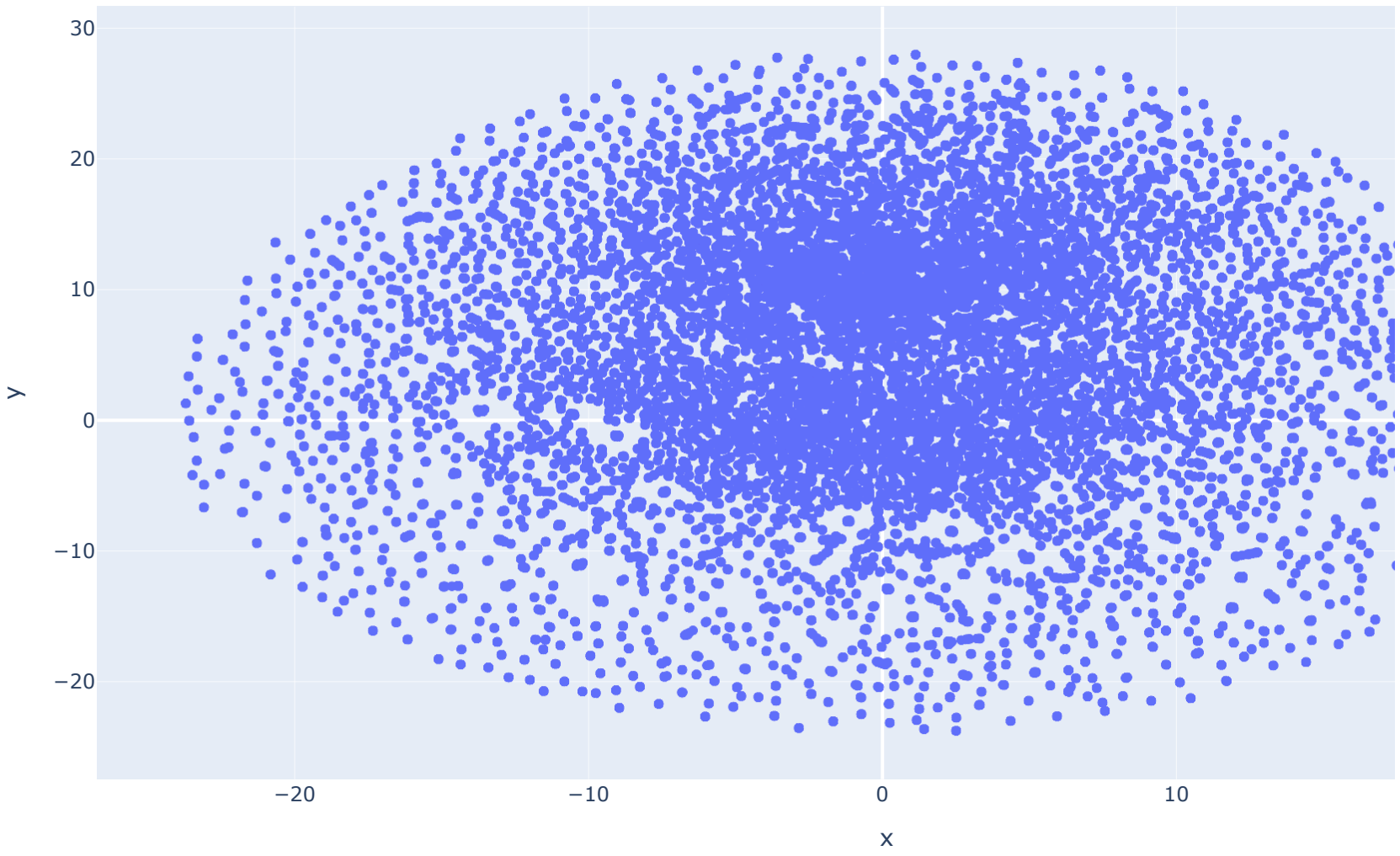
x

Applied t-Distributed Stochastic Neighbor Embedding (t-SNE) to the entire dataset with a perplexity of 50, reducing all features to a 2-dimensional representation for better visualization.

In [121…
```
umap_model = umap.UMAP(n_components=2, random_state=42)
X_umap = umap_model.fit_transform(df_final_scaled)

print(X_umap.shape)
```

```
(205623, 2)
```

In [122…
```
umap_plot = px.scatter(x = X_umap[:,0], y = X_umap[:,1])
umap_plot.update_layout(
    width=1100,
    height=600)
```

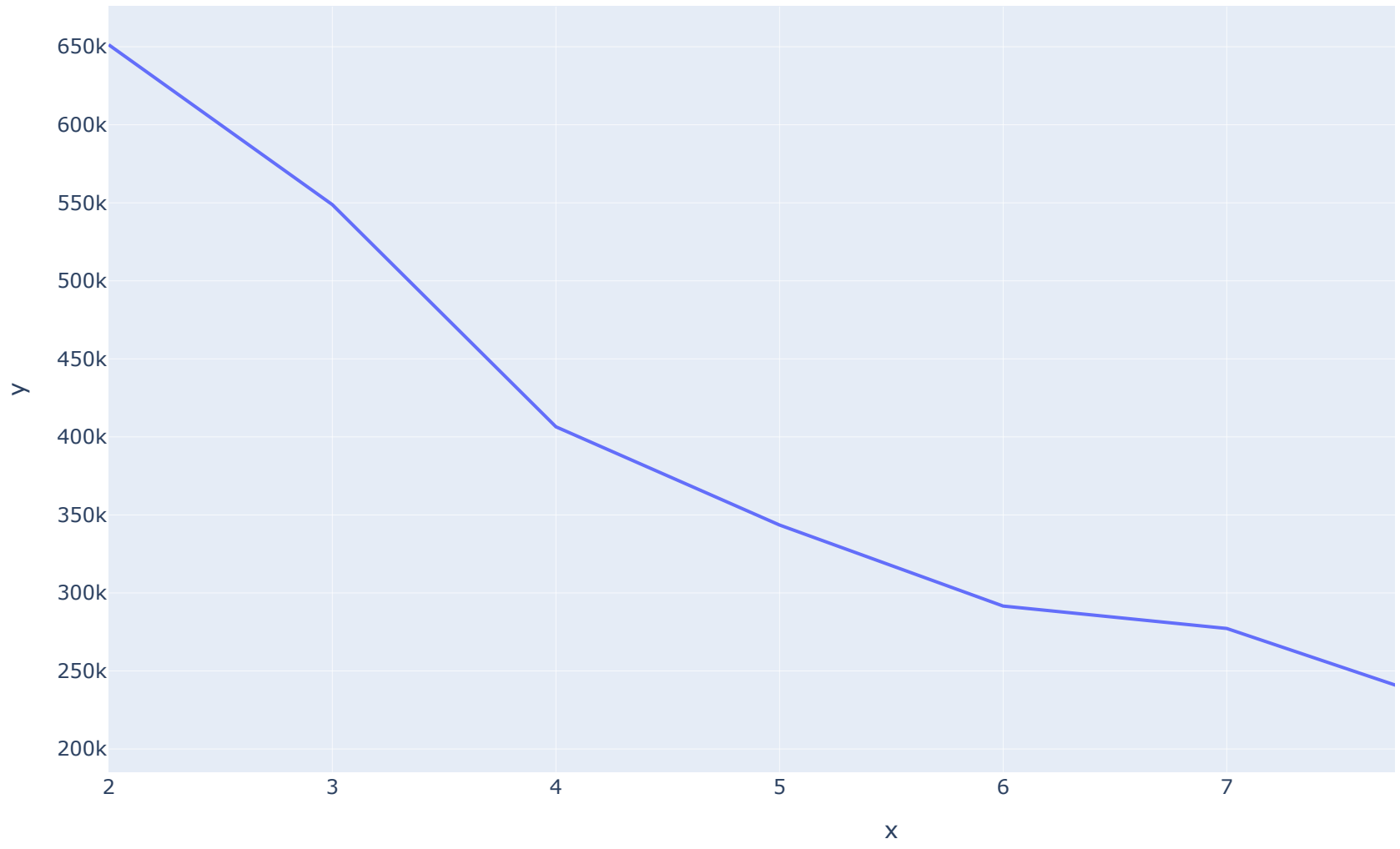


Used UMAP and reducing all features to 2-dimensional representation.

KMeans

In [126…

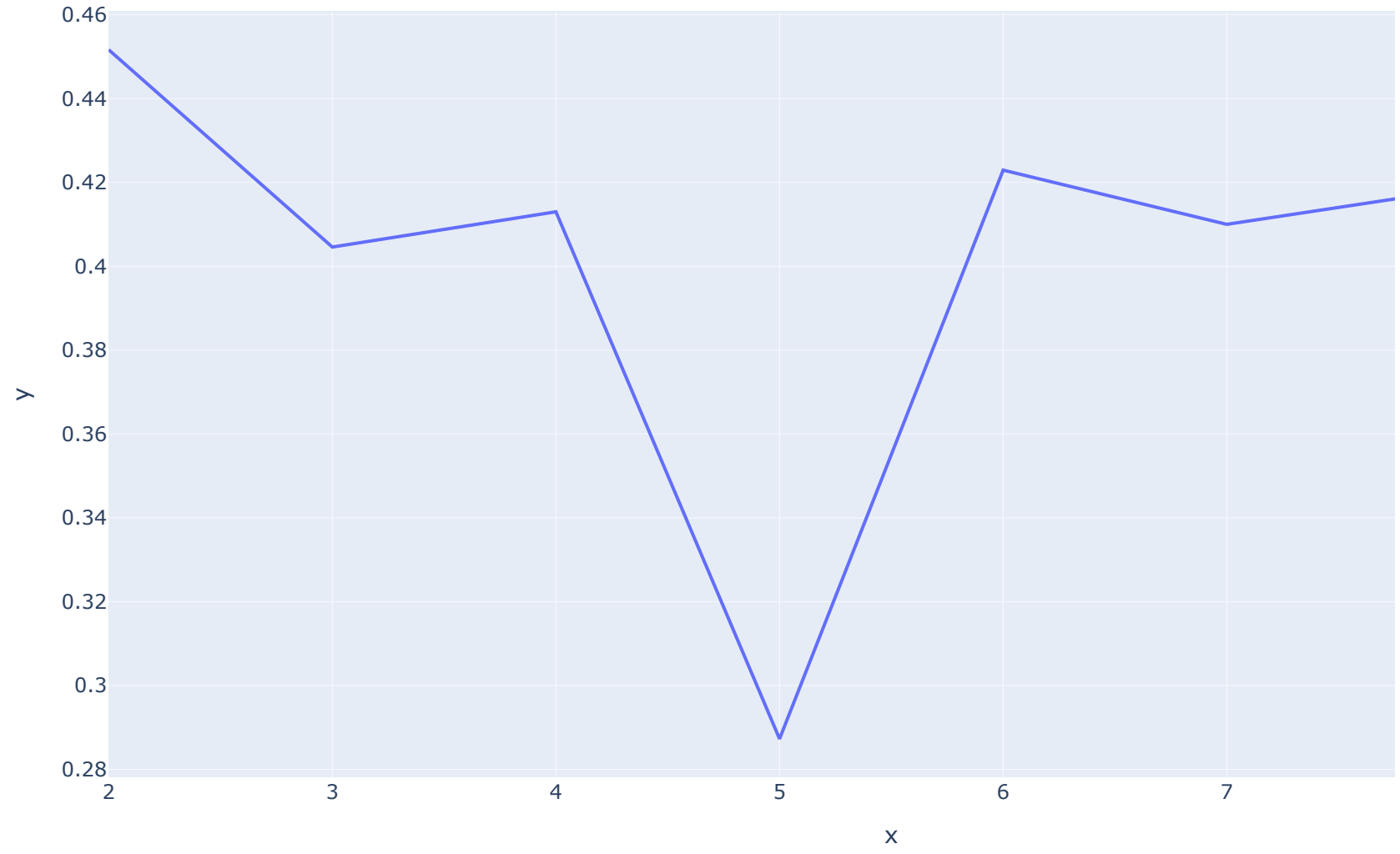
```
elbow = []
for k in range(2,10):
  kmeans = KMeans(n_clusters=k)
  kmeans.fit(df_final_scaled)
  elbow.append(kmeans.inertia_)

elbow_px = px.line(x = range(2,10), y = elbow)
elbow_px.update_layout(
    width=1100,
    height=600)
```



Employed the Elbow Method to determine the optimal value of k. However, the analysis revealed that there was no significant decrease in the Within-Cluster Sum of Squares (WCSS) for any value of k. As a result, I proceeded to explore alternative metrics to identify the best value for k.

In [143…

```
from sklearn.metrics import silhouette_score

sil_score = []
for k in range(2,10):
  kmeans = KMeans(n_clusters=k)
  kmeans.fit(df_final_scaled)
  sil_score.append(silhouette_score(df_final_scaled,kmeans.labels_))

sil_px = px.line(x = range(2,10), y = sil_score)
sil_px.update_layout(
    width=1100,
    height=600)
```

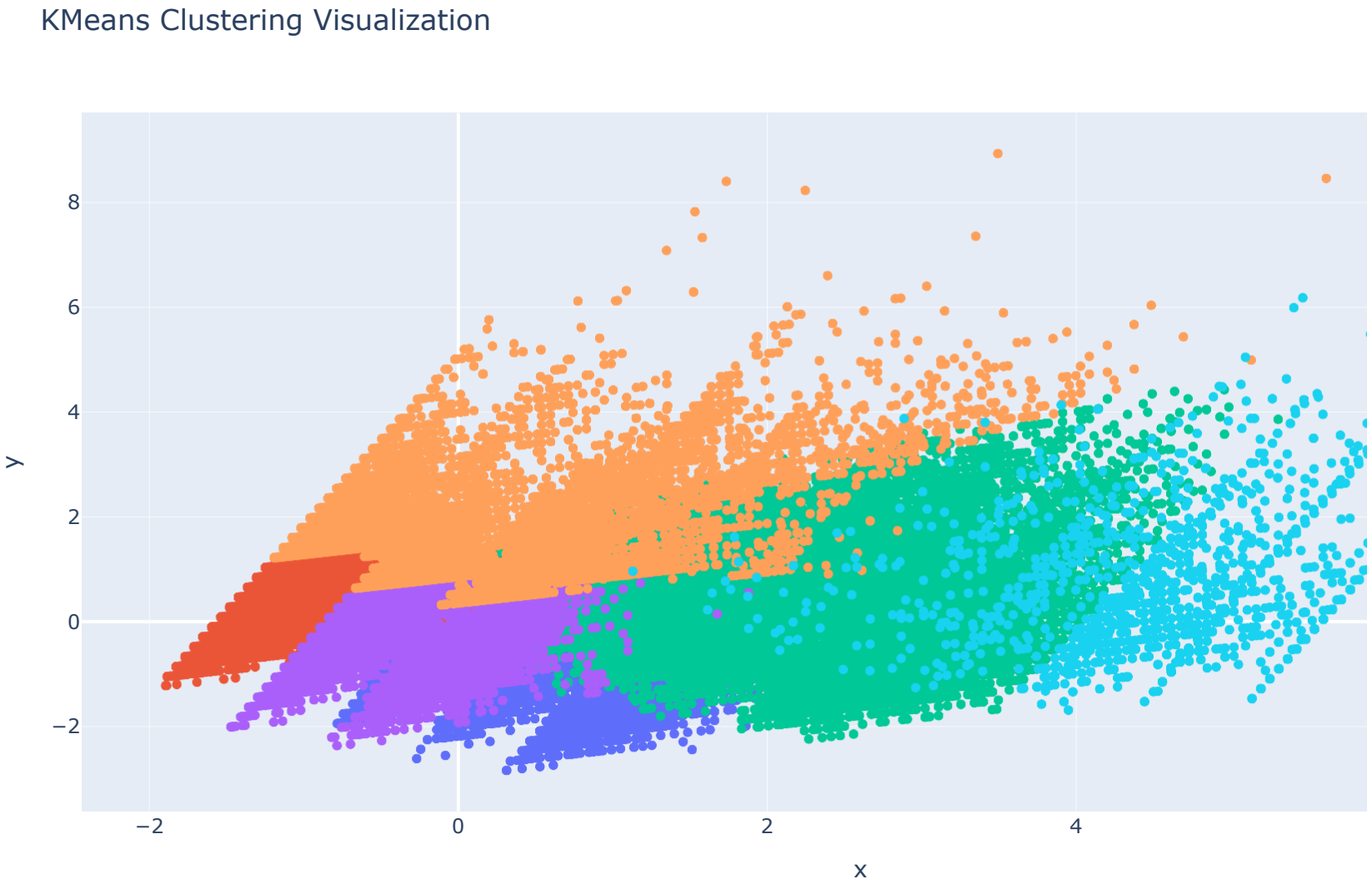


Utilized the silhouette score as an alternative metric to determine the optimal value of k. The highest score was achieved for k = 6, and the Within-Cluster Sum of Squares (WCSS) from the Elbow Method for the same value was also relatively low. Based on this analysis, we can proceed with k = 6 for model development.

In [146…
```
kmeans = KMeans(n_clusters=6)
kmeans.fit(df_final_scaled)
```

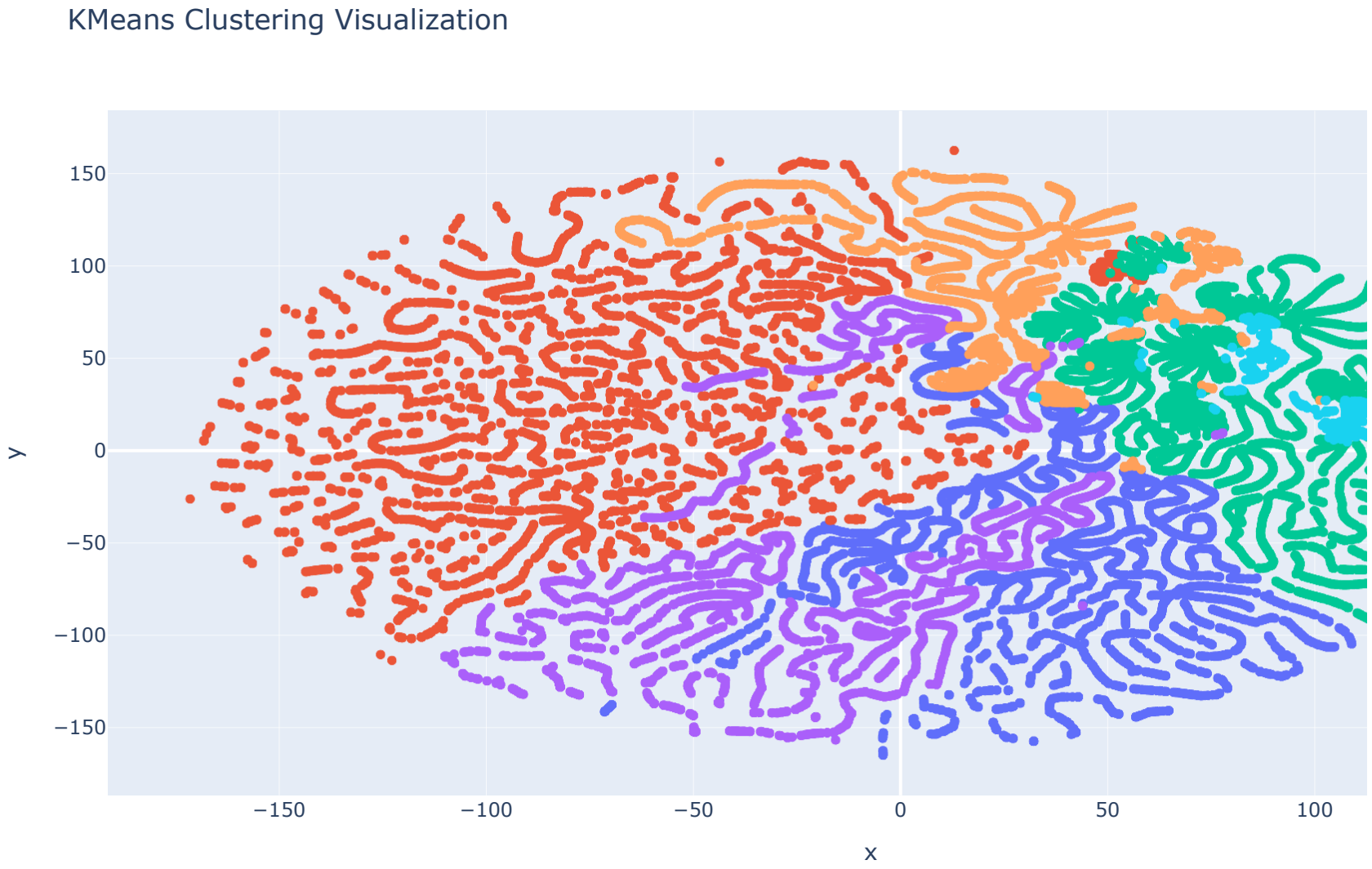
Out[146…
```
KMeans(n_clusters=6)
```

In [148…
```
kmeans_plot = px.scatter(
    x=df_embed[:, 0],
    y=df_embed[:, 1],
    color=kmeans.labels_.astype(str),
    title="KMeans Clustering Visualization"
)

kmeans_plot.update_layout(
    width=1100,
    height=600,
    coloraxis_showscale=True
)

kmeans_plot.show()
```



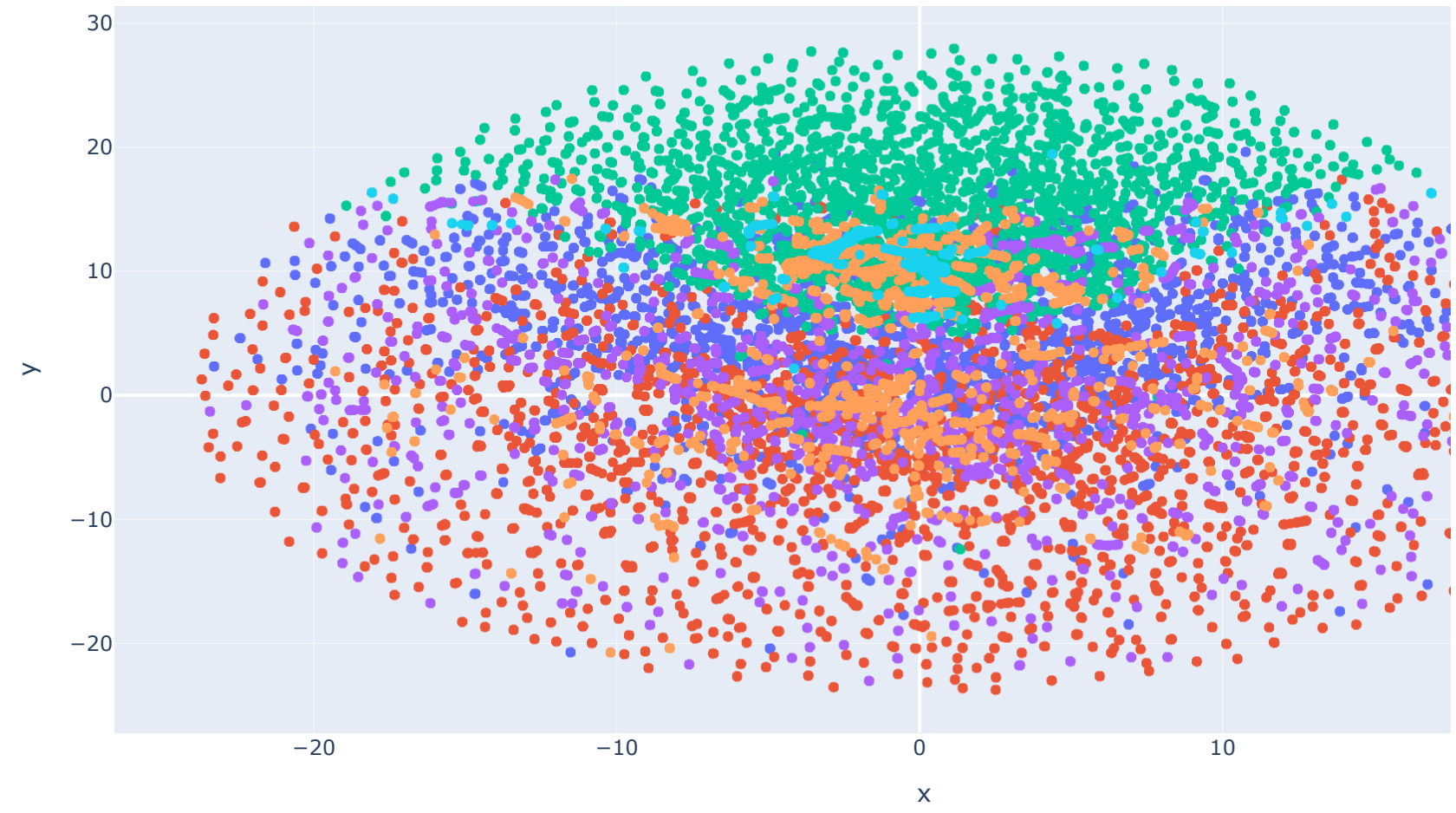
In [150…

```
kmeans_plot_tsne = px.scatter(
    x=df_embed_tsne[:, 0],
    y=df_embed_tsne[:, 1],
    color=kmeans.labels_.astype(str),
    title="KMeans Clustering Visualization"
)

kmeans_plot_tsne.update_layout(
    width=1100,
    height=600,
    coloraxis_showscale=True
)

kmeans_plot_tsne.show()
```



In [152…

```
kmeans_plot_umap = px.scatter(
x = X_umap[:,0], y = X_umap[:,1],
    color=kmeans.labels_.astype(str),
    title="KMeans Clustering Visualization"
)

kmeans_plot_umap.update_layout(
    width=1100,
    height=600,
    coloraxis_showscale=True
)

kmeans_plot_umap.show()
```



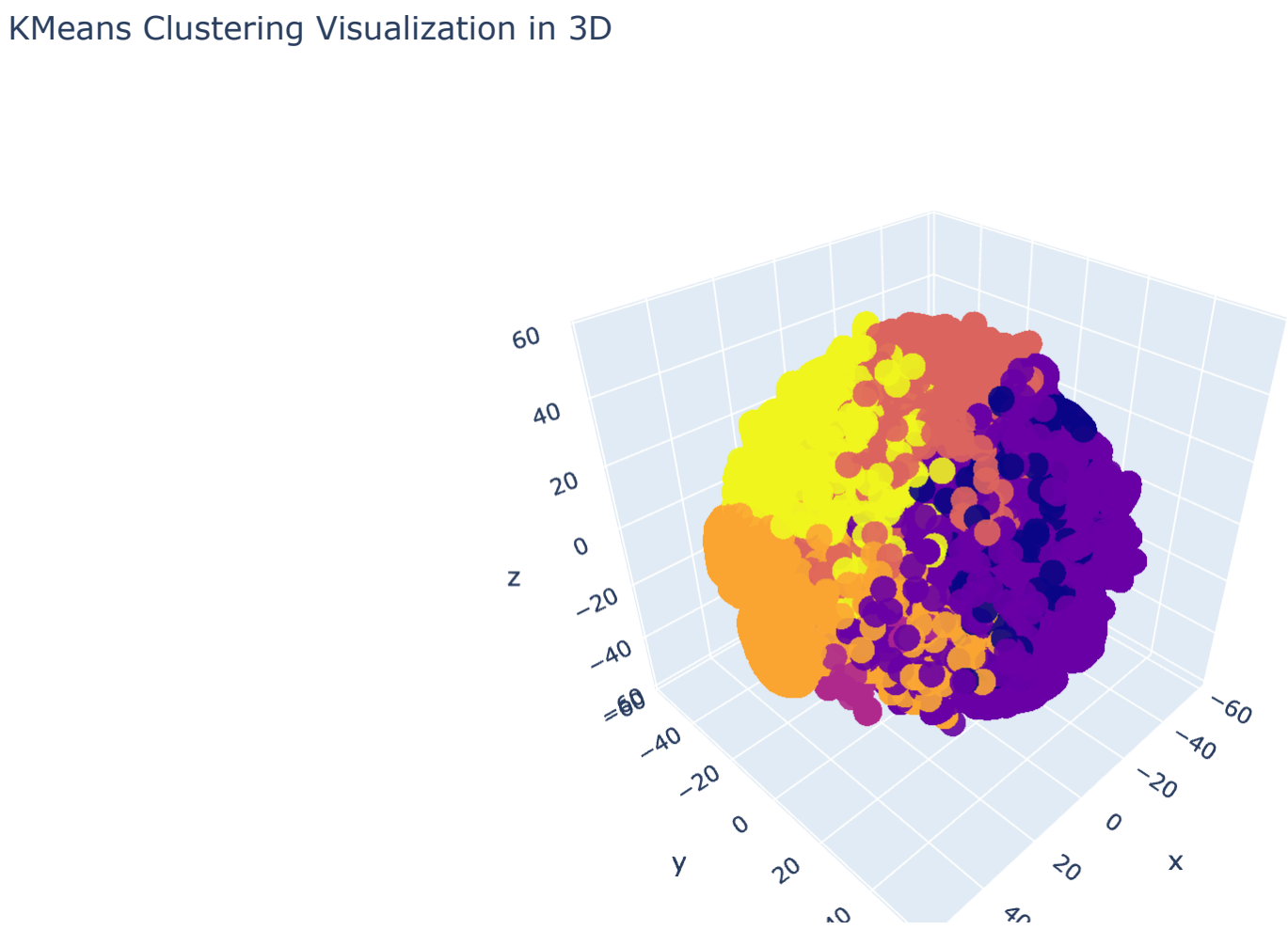
Among PCA, t-SNE, and UMAP, t-SNE provides the most effective visual representation of the clusters.

In [155…
```
tsne_3d = TSNE(n_components = 3, perplexity = 50)
df_embed_tsne_3d = tsne_3d.fit_transform(df_final_scaled)
```

In [157…
```
df_embed_tsne_3d
```

Out[157…
```
array([[ 15.950295 ,  -9.720435 ,  15.822574 ],
       [-21.27045  ,  -3.190481 ,  -3.5663397],
       [ 26.367527 , -20.947083 ,  22.991316 ],
       ...,
       [ 13.31635  , -32.08019  ,   5.652844 ],
       [  5.618409 ,  -3.704048 , -38.70838  ],
       [ 29.790348 ,  -7.976014 ,  42.163017 ]], dtype=float32)
```

In [167…
```
kmeans_plot_tsne_3d = px.scatter_3d(
    x=df_embed_tsne_3d[:, 0],
    y=df_embed_tsne_3d[:, 1],
    z=df_embed_tsne_3d[:, 2],
    color=kmeans.labels_,
    opacity = .9,
    size_max=18,
    title="KMeans Clustering Visualization in 3D"
)

kmeans_plot_tsne_3d.update_layout(
    width=1100,
    height=600
)

kmeans_plot_tsne_3d.show()
```



Applied t-Distributed Stochastic Neighbor Embedding (t-SNE) to the entire dataset, reducing the features to three embedded components for 3D visualization of the data.

Gaussian Mixture Model

In [169…
```
gmm = GaussianMixture(n_components=4, covariance_type='full')
gmm.fit(df_final_scaled)
```

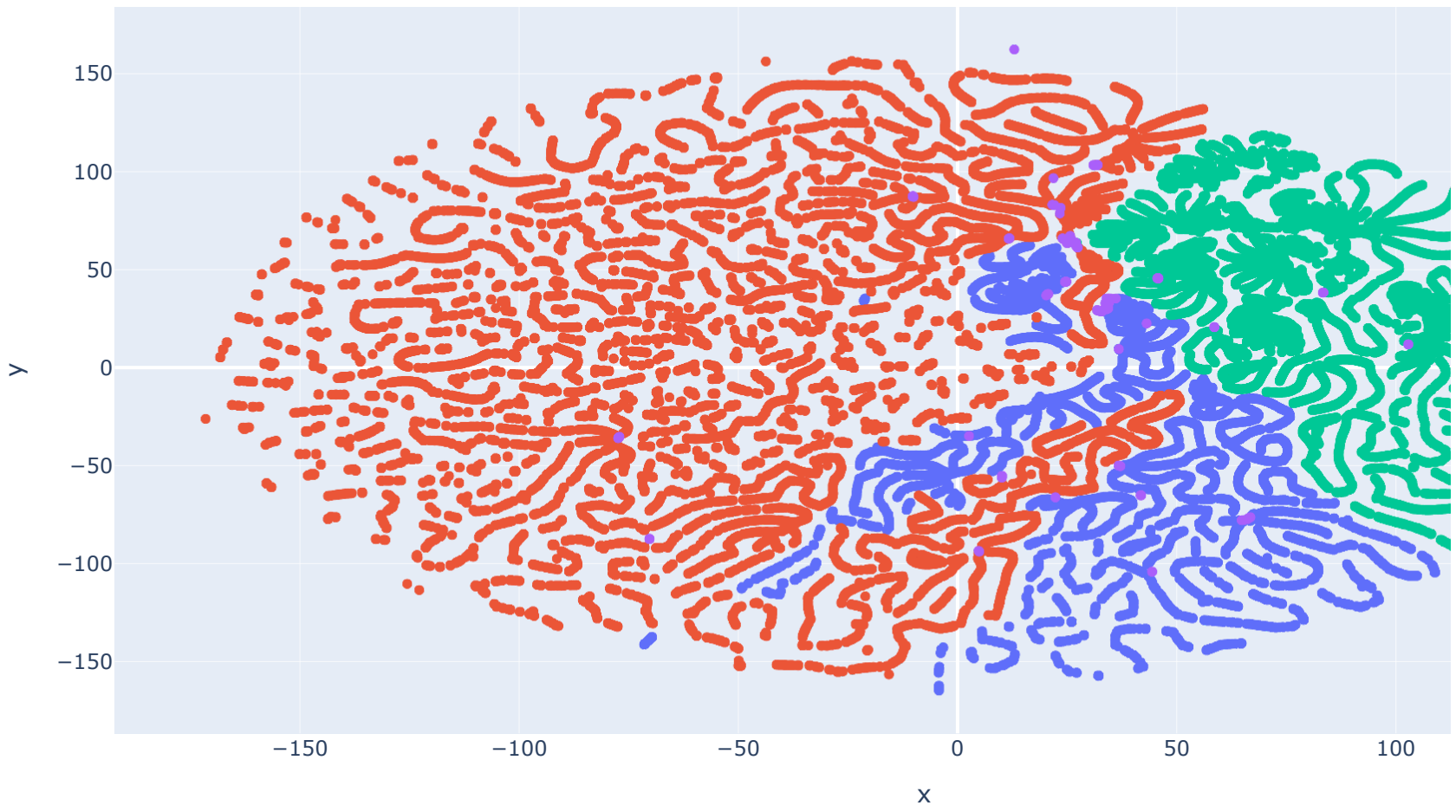
Out[169…
```
GaussianMixture(n_components=4)
```

In [170…
```
gmm_labels = gmm.predict(df_final_scaled)
```

In [171…
```
gmm_labels
```

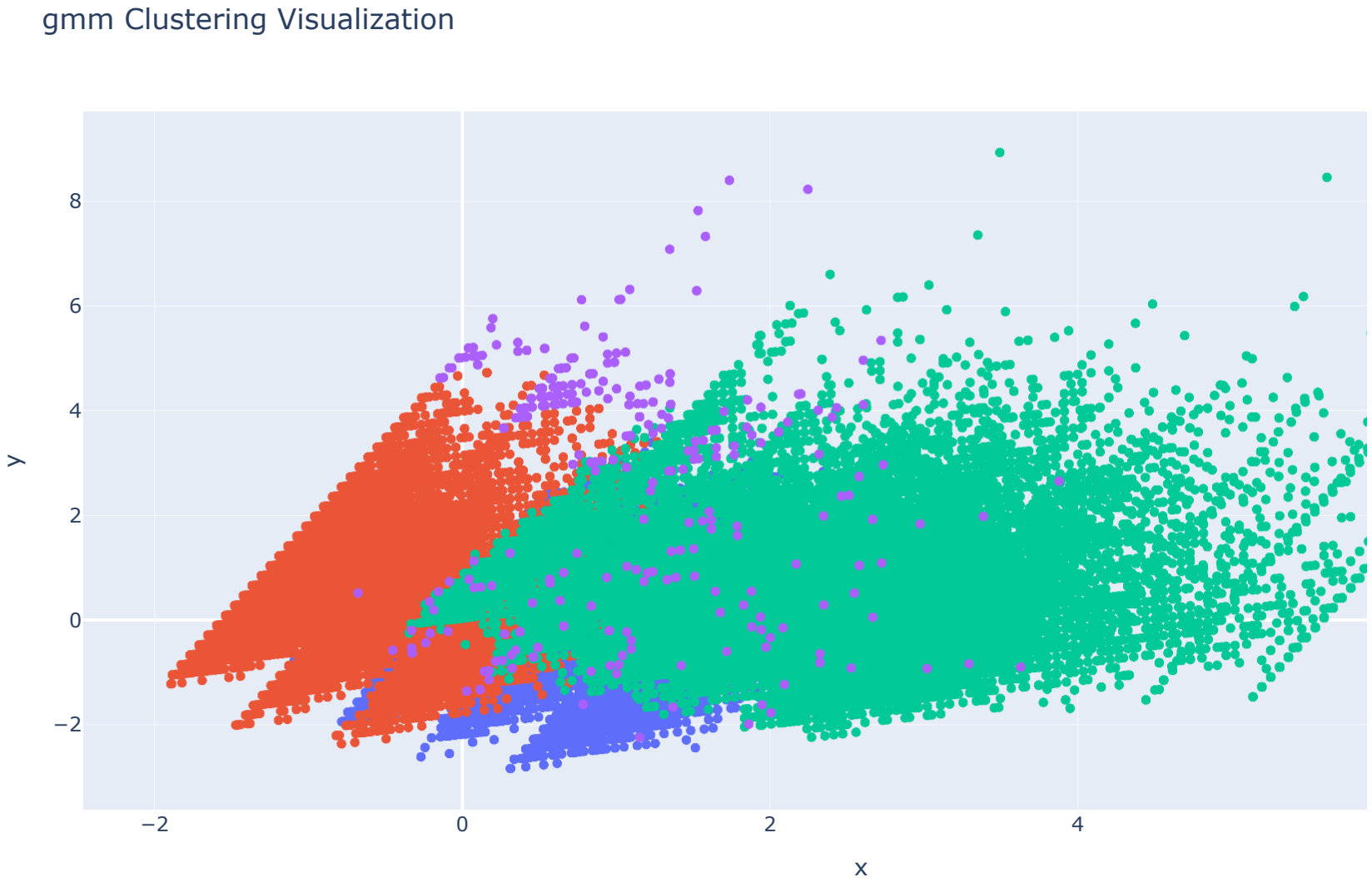
Out[171… array([3, 0, 3, ..., 3, 2, 3])

In [179…
```
gmm_plot_tsne = px.scatter(
    x=df_embed_tsne[:, 0],
    y=df_embed_tsne[:, 1],
    color=gmm_labels.astype(str),
    title="gmm Clustering Visualization"
)

gmm_plot_tsne.update_layout(
    width=1100,
    height=600,
    coloraxis_showscale=True
)

gmm_plot_tsne.show()
```



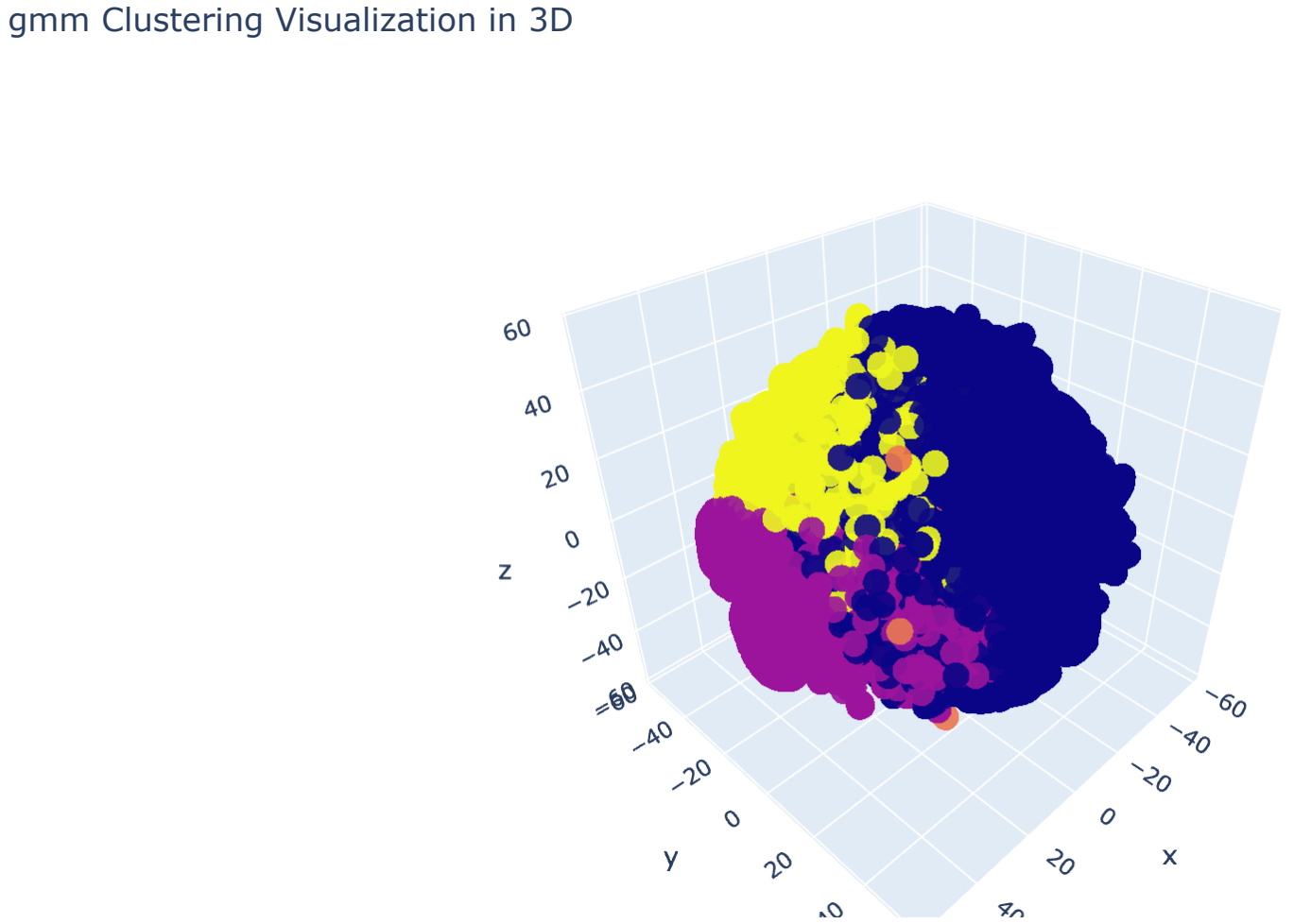
Used tSNE for better visualization of clusters in 2D

In [182…
```
gmm_plot_pca = px.scatter(
    x=df_embed[:, 0],
    y=df_embed[:, 1],
    color=gmm_labels.astype(str),
    title="gmm Clustering Visualization"
)

gmm_plot_pca.update_layout(
    width=1100,
    height=600,
    coloraxis_showscale=True
)

gmm_plot_pca.show()
```



In [188…

```
gmm_plot_tsne_3d = px.scatter_3d(
    x=df_embed_tsne_3d[:, 0],
    y=df_embed_tsne_3d[:, 1],
    z=df_embed_tsne_3d[:, 2],
    color=gmm_labels,
    opacity = .9,
    size_max=18,
    title="gmm Clustering Visualization in 3D"
)

gmm_plot_tsne_3d.update_layout(
    width=1100,
    height=600,
)

gmm_plot_tsne_3d.show()
```



Applied t-Distributed Stochastic Neighbor Embedding (t-SNE) to the entire dataset, reducing the features to three embedded components for 3D visualization of the data.

Hierarchical Clustering

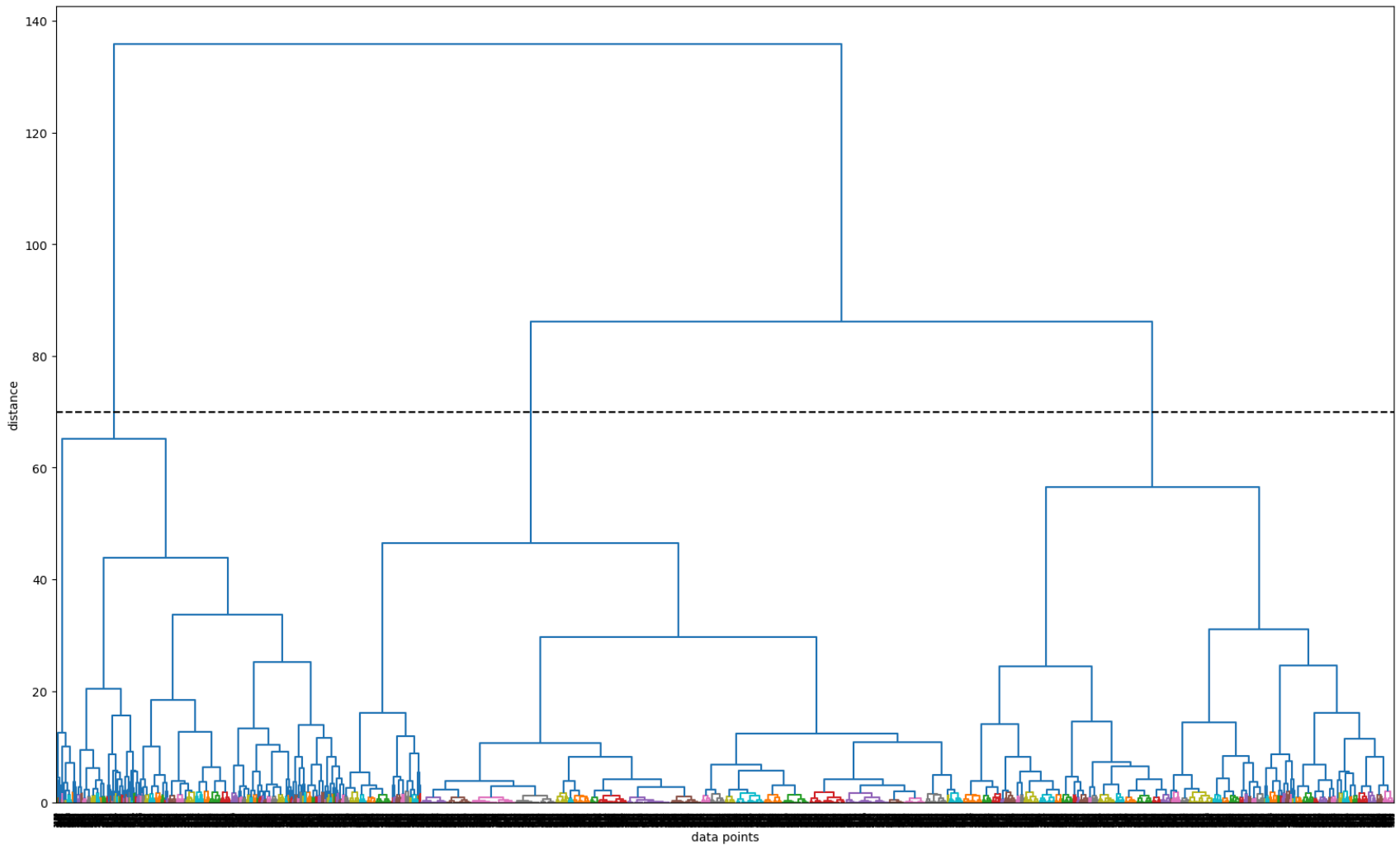
Given the large size of the dataset, we have decided to proceed with a subset of 5000 samples for hierarchical clustering analysis.

In [194…
```
df_clust = pd.DataFrame(df_final_scaled)
```

In [196…
```
h_clust = df_clust.sample(5000)
```

In [197…
```
z = sch.linkage(h_clust, method='ward')
```

In [199…
```
fig, ax = plt.subplots(figsize=(20, 12))

sch.dendrogram(z,ax=ax, color_threshold=2)


plt.xticks(rotation=90)
plt.axhline(y = 70,linestyle = '--', color = 'black')
ax.set_xlabel('data points')
ax.set_ylabel('distance')
plt.show()
```



Plotted Dendogram for the sampled data.

DBSCAN Clustering

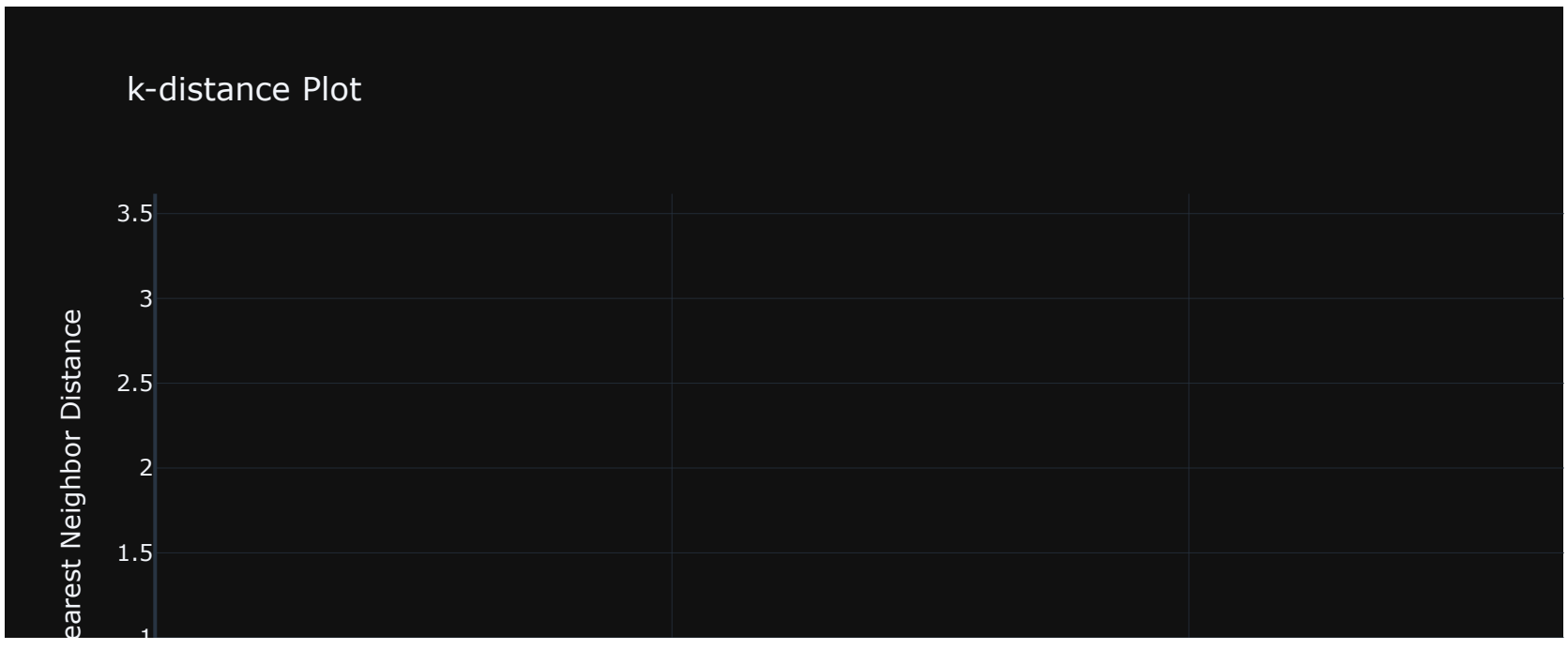
Based on the observations from the visualization, it is clear that the data exhibits varying densities, which may affect the performance of DBSCAN. However, for the purpose of this project, we have still applied DBSCAN for clustering analysis.

In [206…
```
from sklearn.neighbors import NearestNeighbors
```

Hyperparameter tuning - To find best value of epsilon, we plotted k- distance plot and observe for the best value of epsilon where it is sharply increasing or forming an elbow like structure.

In [209…
```
k = 4
neighbors = NearestNeighbors(n_neighbors=k)
neighbors.fit(df_clust)

distances, indices = neighbors.kneighbors(df_clust)

sorted_distances = np.sort(distances[:, k-1])
```

In [211…
```
fig = go.Figure()

fig.add_trace(go.Scatter(
    x=np.arange(len(sorted_distances)),
    y=sorted_distances,
    mode='lines',
    name=f'{k}-th Nearest Neighbor Distance'
))

fig.update_layout(
    title='k-distance Plot',
    xaxis_title='Points sorted by distance',
    yaxis_title=f'{k}-th Nearest Neighbor Distance',
    template='plotly_dark',
    showlegend=False
)

fig.show()
```



Best Value of epsilon comes out to be 0.25

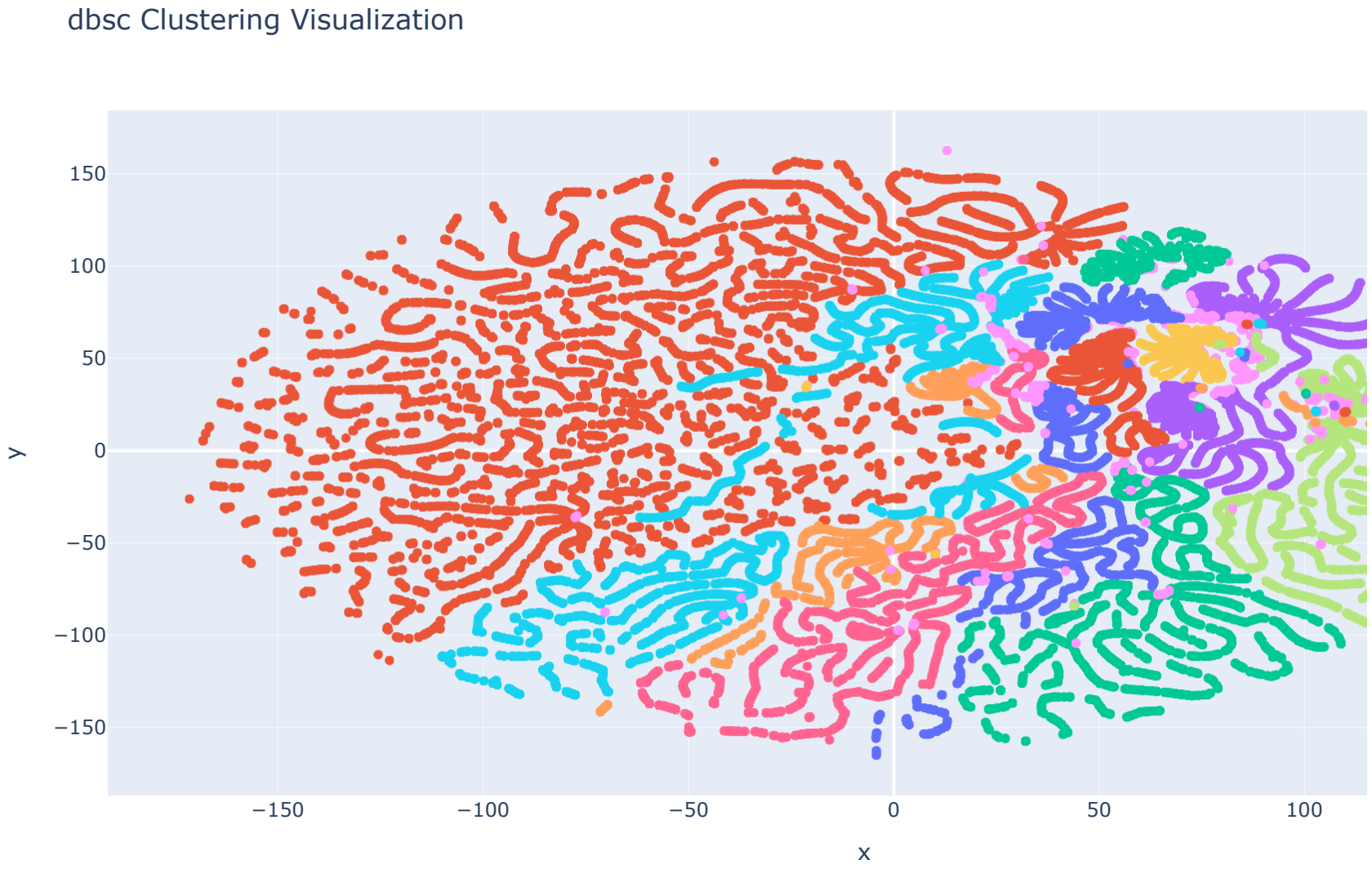
In [214…
```
dbsc = DBSCAN(eps = 0.25, min_samples = 10).fit(df_clust)
```

In [215…
```
dbsc_lbel = dbsc.labels_
```

In [218…
```
dbsc_lbel
```

Out[218… array([0, 1, 2, ..., 2, -1, 2])

In [220…
```
dbsc_plot_tsne = px.scatter(
    x=df_embed_tsne[:, 0],
    y=df_embed_tsne[:, 1],
    color=dbsc_lbel.astype(str),
    title="dbsc Clustering Visualization"
)

dbsc_plot_tsne.update_layout(
    width=1100,
    height=600,
    coloraxis_showscale=True
)

dbsc_plot_tsne.show()
```



Used t-SNE to achieve a clearer 2D visualization of the clusters. Due to the varying density in the dataset, DBSCAN formed numerous clusters that do not appear to provide meaningful insights. However, for the sake of completeness, we have retained the DBSCAN clustering results.

KMeans Clustering Interpretation:

In [225…
```
kmeans.labels_
```

Out[225…
```
array([5, 1, 5, ..., 5, 1, 5], dtype=int32)
```

In [227…
```
df_clust['KMeans_labels'] = kmeans.labels_
```

In [229…
```
df_clust.columns = ['ctc', 'years_of_experience', 'designation', 'class', 'tier', 'KMeans_labels']
df_clust.head()
```

Out[229…

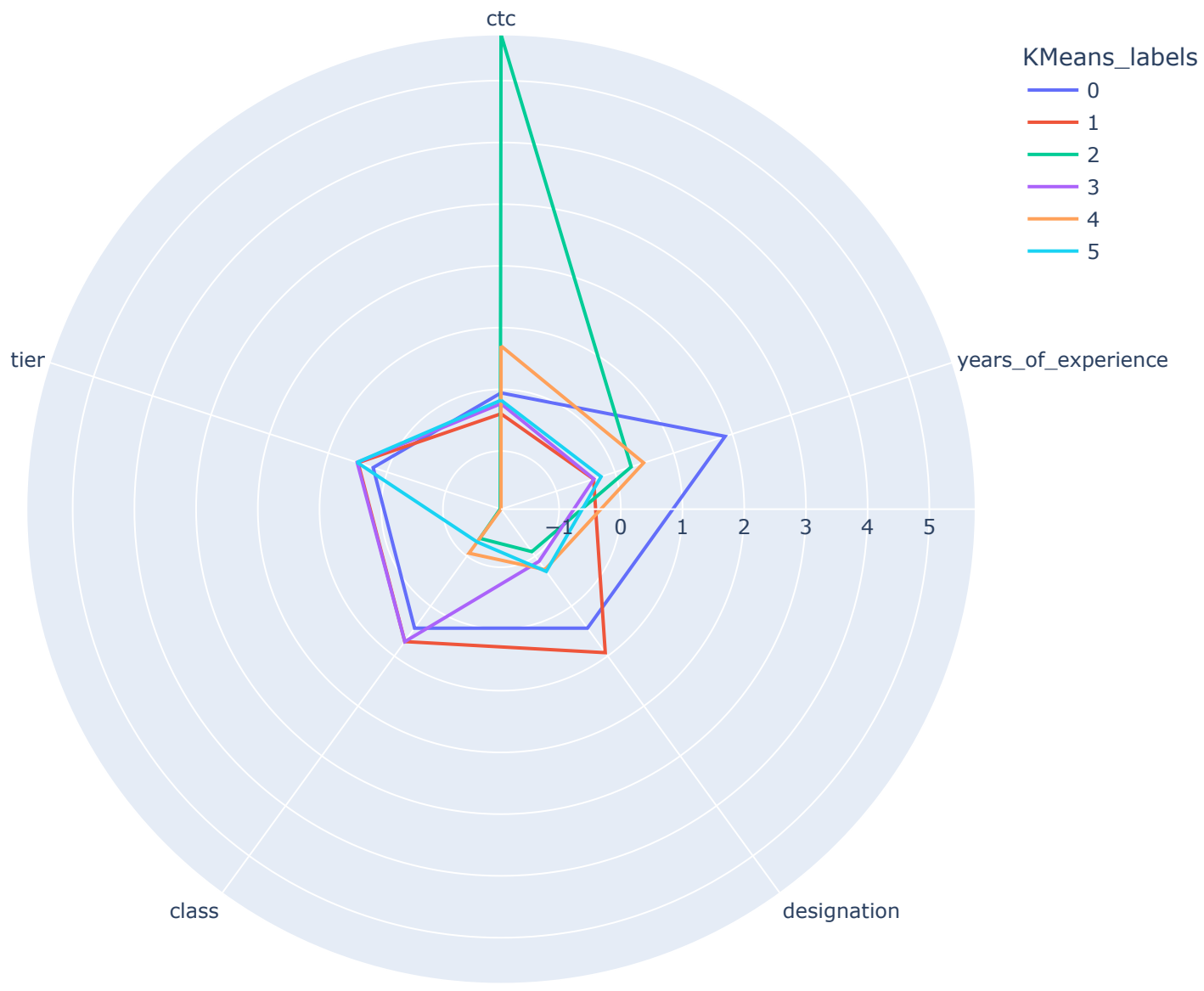
	ctc	years_of_experience	designation	class	tier	KMeans_labels
0	-0.188006	-0.208542	-0.279440	-1.508766	0.515317	5
1	-0.578162	-0.680839	0.932961	0.712106	0.515317	1
2	0.352208	0.027607	-1.491841	-1.508766	0.515317	5
3	-0.428101	-0.444691	0.932961	0.712106	0.515317	1
4	-0.007935	-0.444691	-0.279440	-1.508766	-1.941750	4

In [231…
```
polar = df_clust.groupby("KMeans_labels").mean().reset_index()
polar = pd.melt(polar, id_vars=["KMeans_labels"])
polar.head(4)
```

Out[231…

	KMeans_labels	variable	value
0	0	ctc	-0.057319
1	1	ctc	-0.395467
2	2	ctc	5.748179
3	3	ctc	-0.231929

In [233…
```
fig = px.line_polar(polar, r="value", theta="variable", color="KMeans_labels", line_close=True,height=700,w
fig.show()
```



Label 0 - Cluster of learners who have low tier, class, designation, ctc and experience. Most probably representing a cluster who are early in their careers.

Label 1 - Cluster of learners who have higher designation, class, and experience but similar tier and ctc as compared to above cluster. Probably in roles where growth is slow.

Label 2 - Looks anomolous, non interpretable.

Label 3 - Cluster of learners who have higher designation, class, and tier but lower experience and ctc. Probably belonging to an industry where ctc growth is low but promotions are frequent.

Label 4 - Cannot.

Label 5 - Cluster of learners who have very high experience and ctc as compared to all others, but have lower designation, class, and tier.