Salary Predictions of Data Professions

By:Pratiksha Saheb

Problem Statement

Salaries in the field of data professions vary widely based on factors such as experience, job role, and performance. Accurately predicting salaries for data professionals is essential for both job seekers and employers

Methodology

Exploratry Data Analysis

Data Preprocessing

Feature Engineering

Machine Learning Model

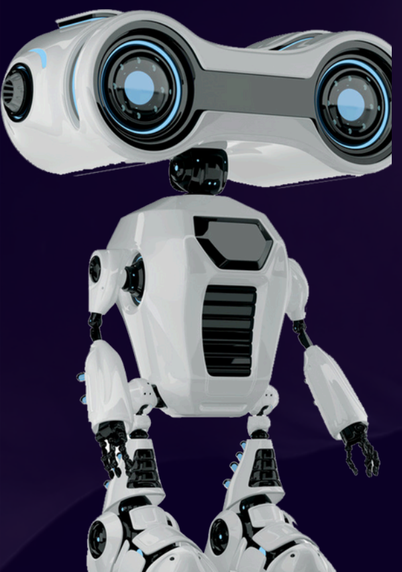
Model Evaluation

Exploratry Data Analysis

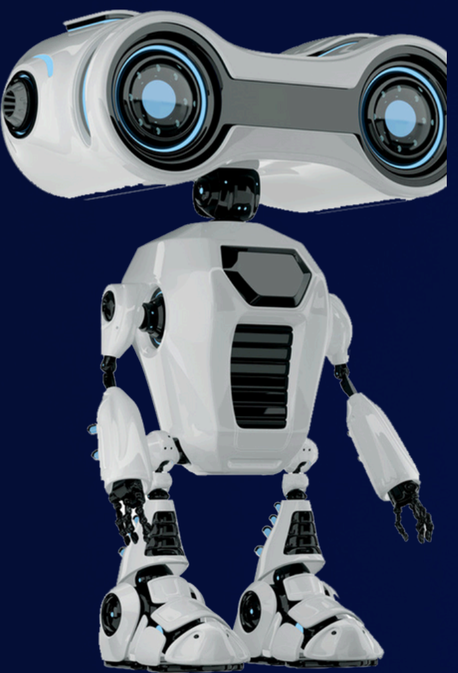
```
1 df = pd.read_csv('Salary Prediction of Data Professions.csv')
```

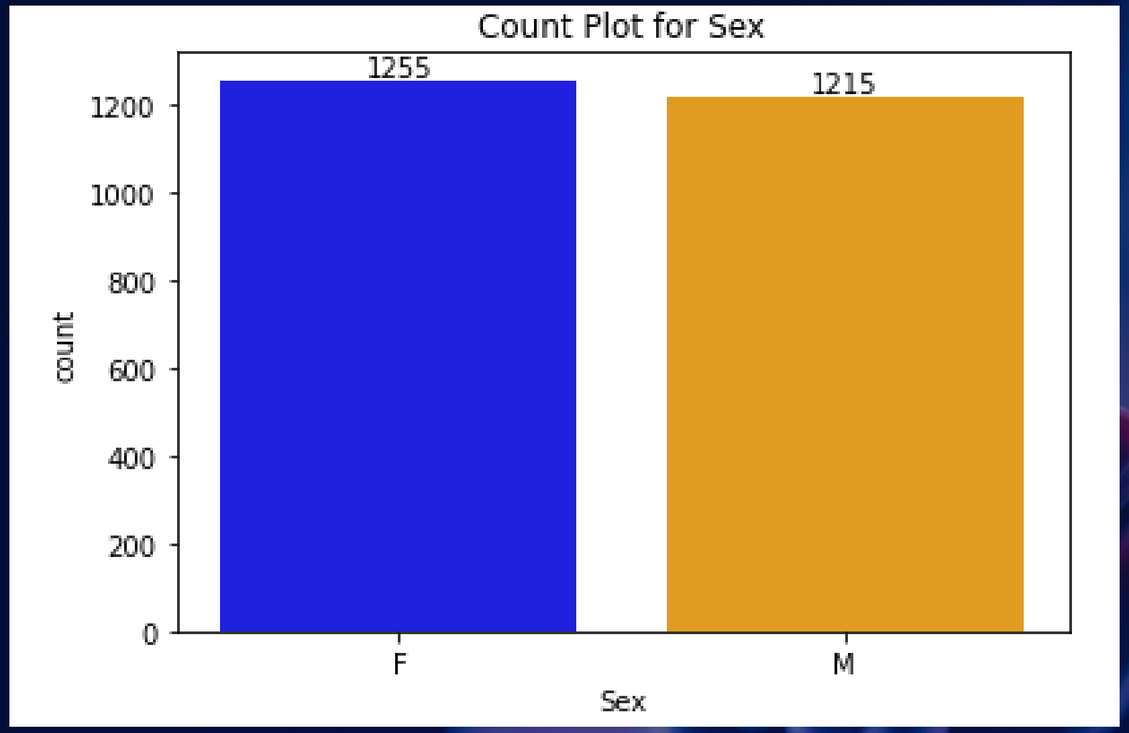
```
1 df.head()
```

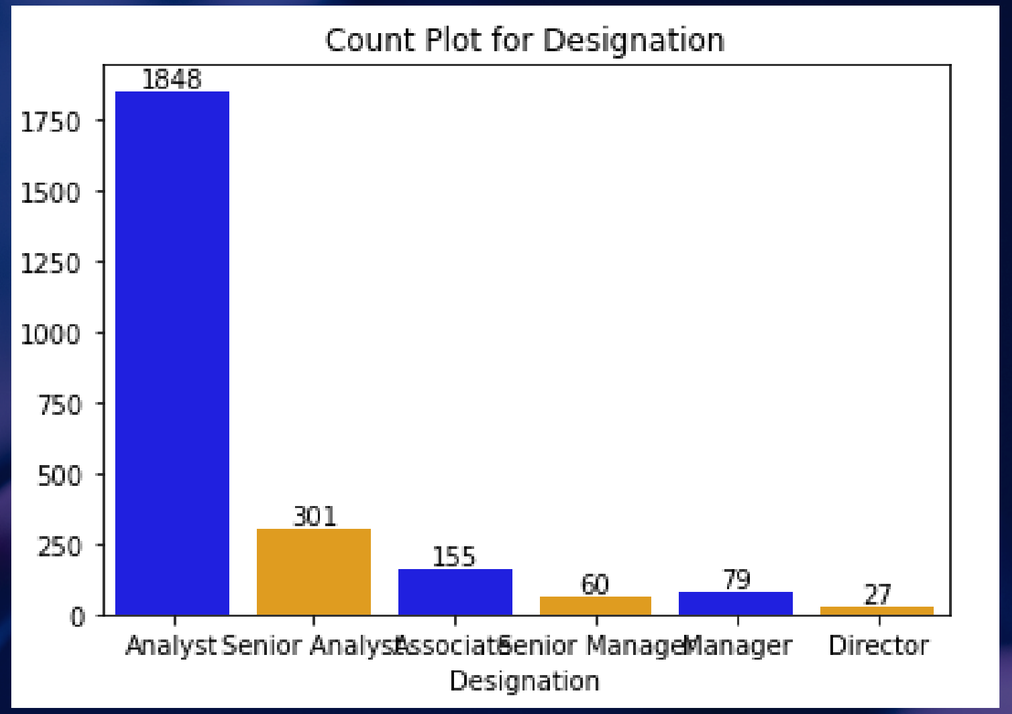
	FIRST NAME	LAST NAME	SEX	DOJ	CURRENT DATE	DESIGNATION	AGE	SALARY	UNIT	LEAVES USED	LEAVES REMAINING	RATINGS	PAST EXP
0	TOMASA	ARMEN	F	5-18-2014	01-07-2016	Analyst	21.0	44570	Finance	24.0	6.0	2.0	0
1	ANNIE	NaN	F	NaN	01-07-2016	Associate	NaN	89207	Web	NaN	13.0	NaN	7
2	OLIVE	ANCY	F	7-28-2014	01-07-2016	Analyst	21.0	40955	Finance	23.0	7.0	3.0	0
3	CHERRY	AQUILAR	F	04-03-2013	01-07-2016	Analyst	22.0	45550	IT	22.0	8.0	3.0	0
4	LEON	ABOULAHOUD	M	11-20-2014	01-07-2016	Analyst	NaN	43161	Operations	27.0	3.0	NaN	3

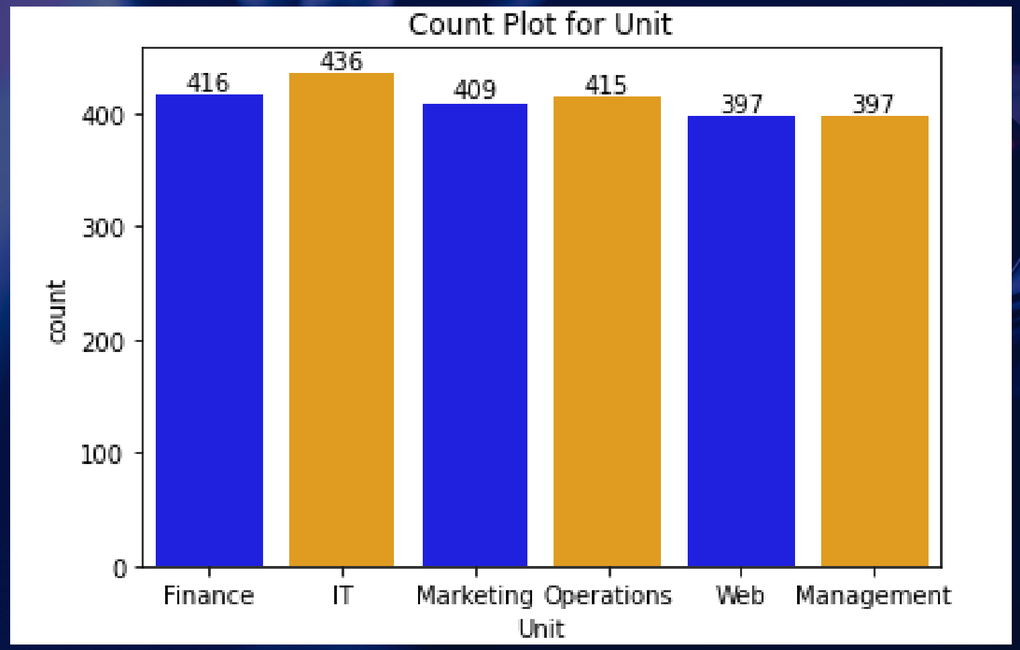


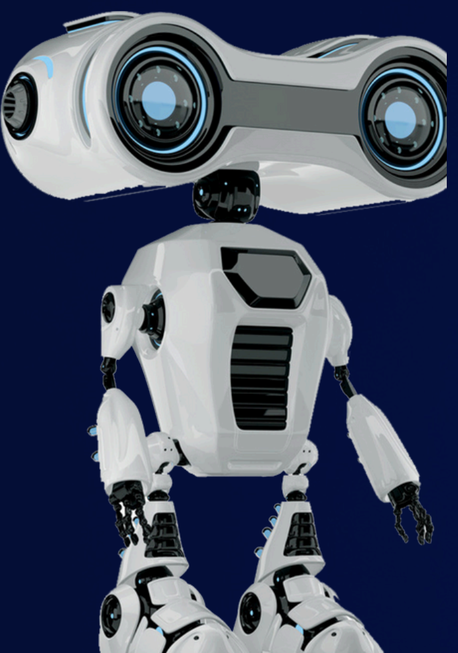
count	2636.000000	2639.000000	2636.000000	2637.000000	2637.000000	2639.000000
mean	24.756449	58136.678287	22.501517	7.503223	3.486159	1.566881
std	3.908228	36876.956944	4.604469	4.603193	1.114933	2.728416
min	21.000000	40001.000000	15.000000	0.000000	2.000000	0.000000
25%	22.000000	43418.000000	19.000000	4.000000	2.000000	0.000000
50%	24.000000	46781.000000	22.000000	8.000000	3.000000	1.000000
75%	25.000000	51401.500000	26.000000	11.000000	4.000000	2.000000
max	45.000000	388112.000000	30.000000	15.000000	5.000000	23.000000

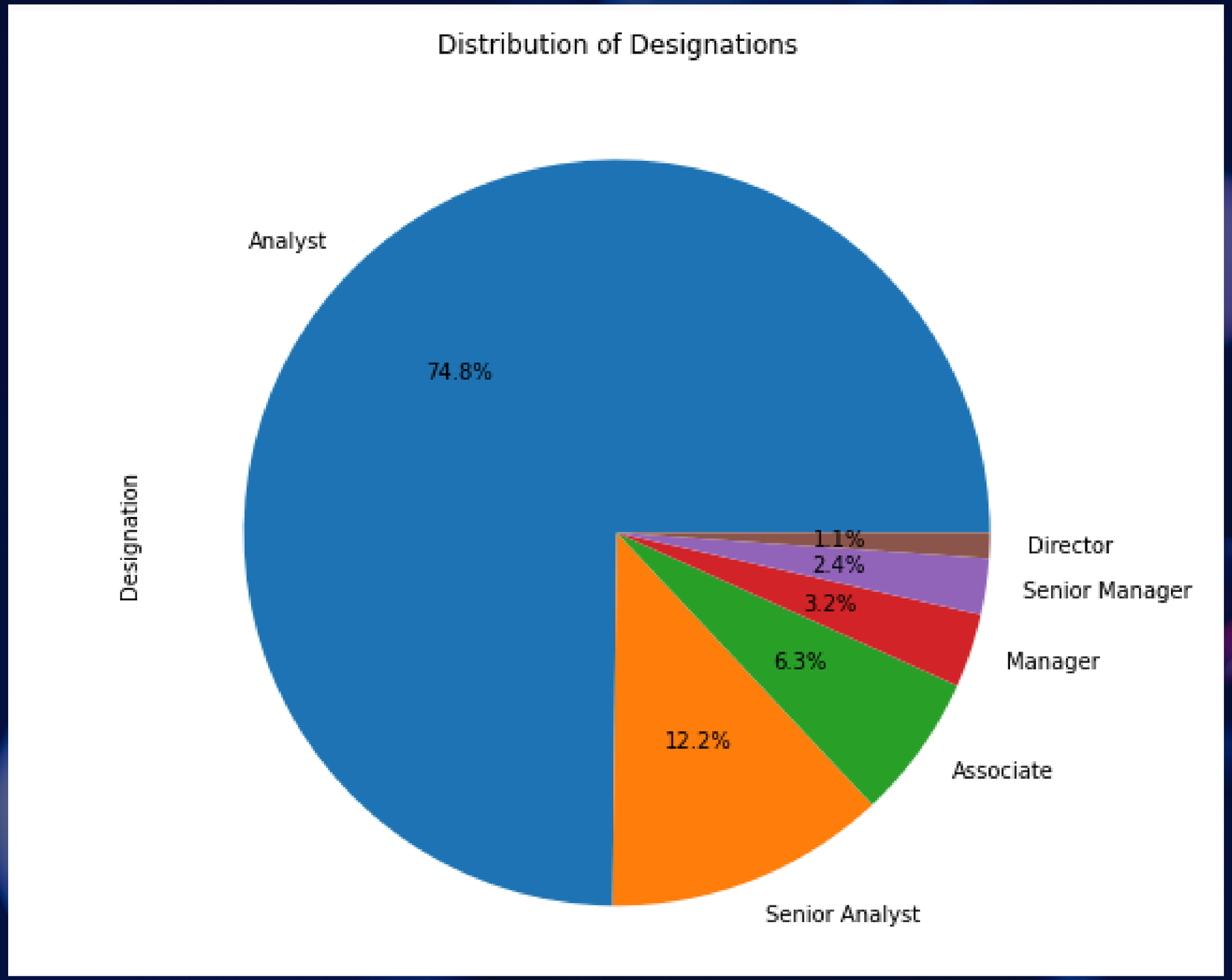


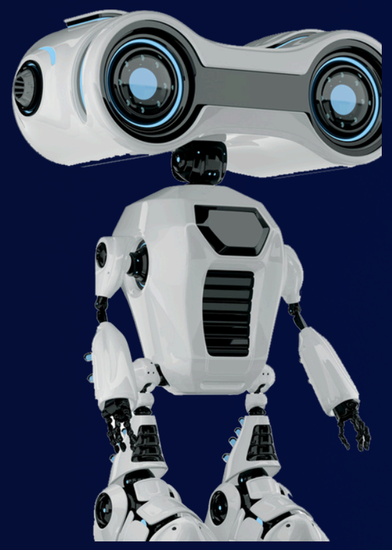

Count Plot for Sex
1255
1215
1200
1000
800
600
400
200
0
count
F
M
Sex

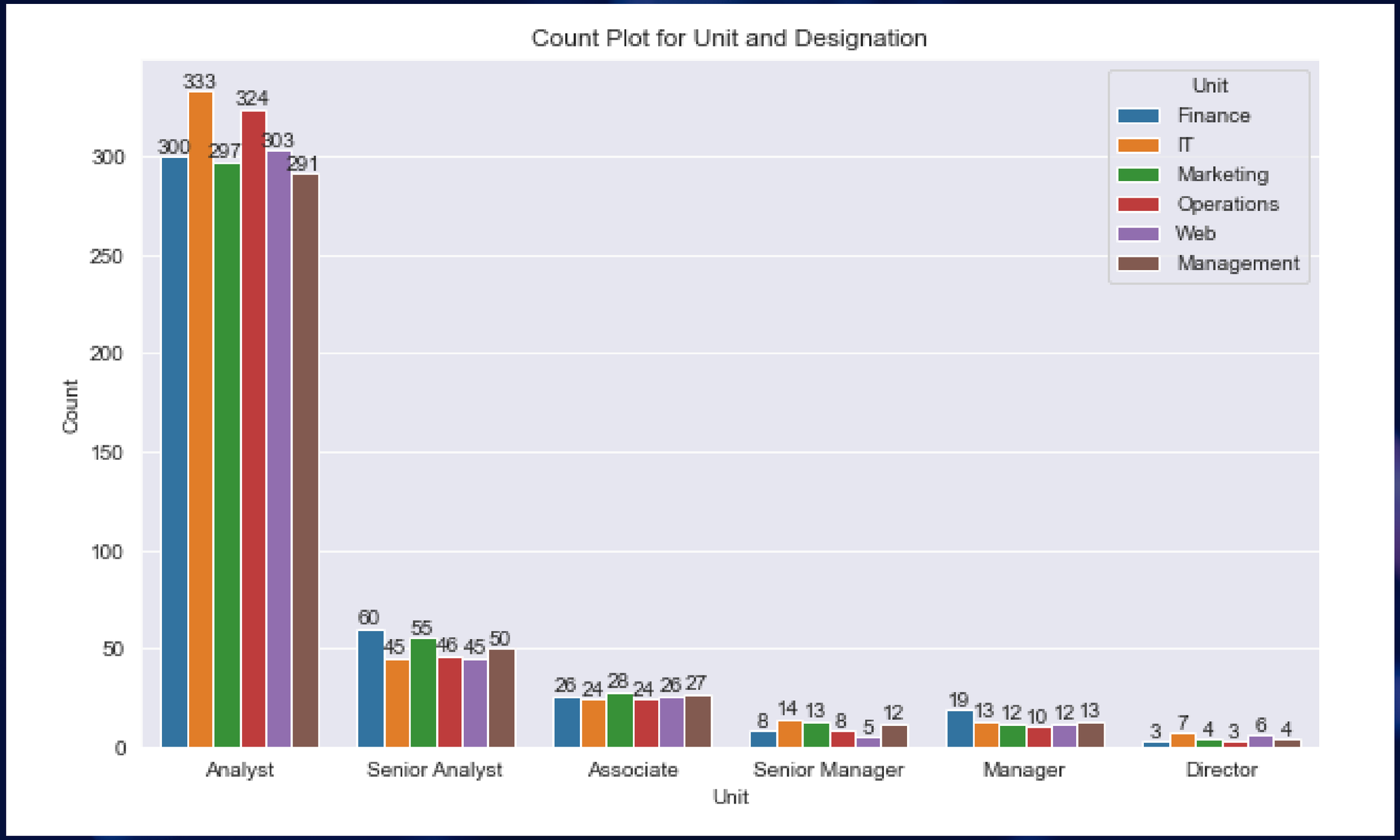

Count Plot for Designation
1848
1750
1500
1250
1000
750
500
250
0
301
155
60
79
27
Analyst
Senior Analyst
Associate
Senior Manager
Manager
Director
Designation

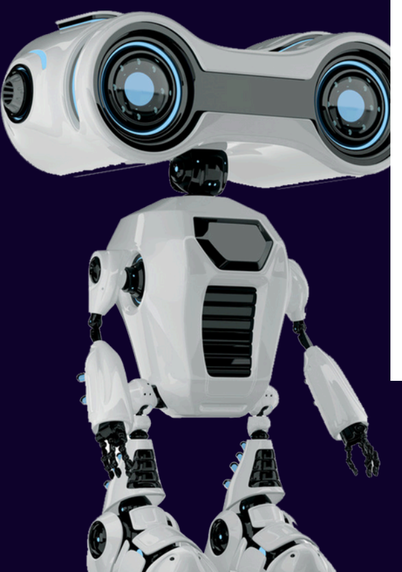

Count Plot for Unit
416
436
409
415
397
397
400
300
200
100
0
count
Finance
IT
Marketing
Operations
Web
Management
Unit

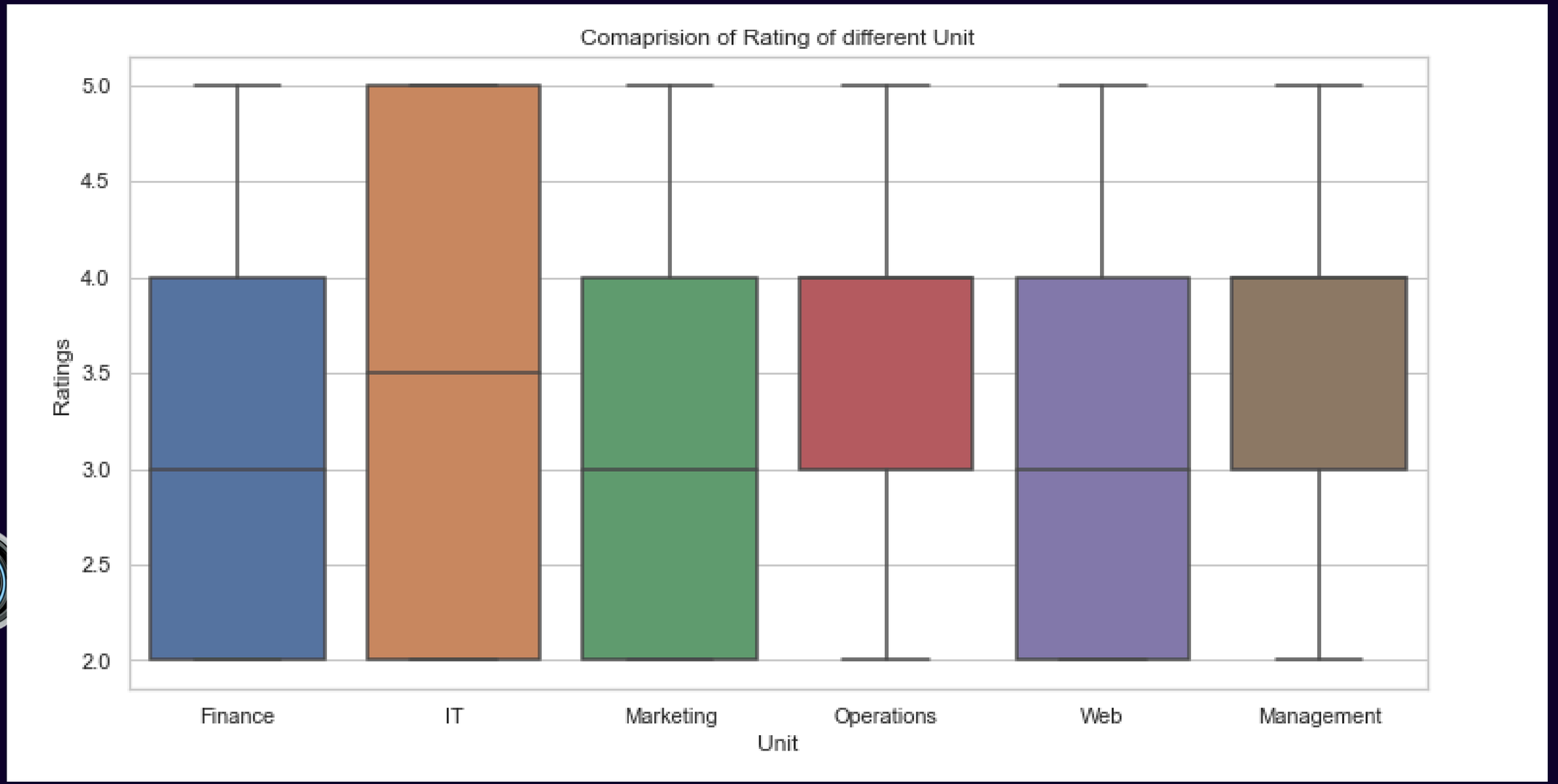



Distribution of Designations
Designation
Analyst
74.8%
Director
1.1%
Senior Manager
2.4%
Manager
3.2%
Associate
6.3%
Senior Analyst
12.2%

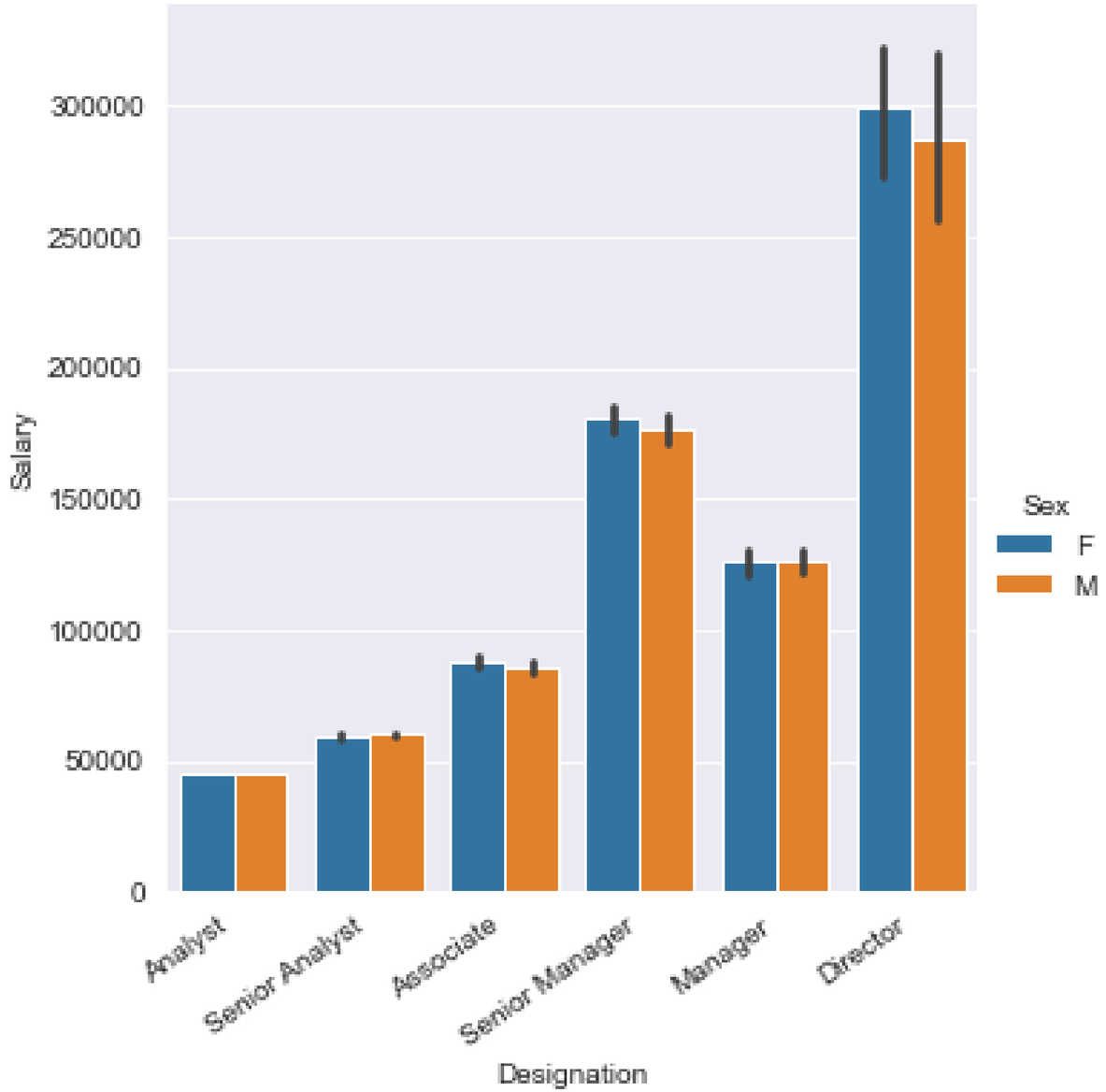


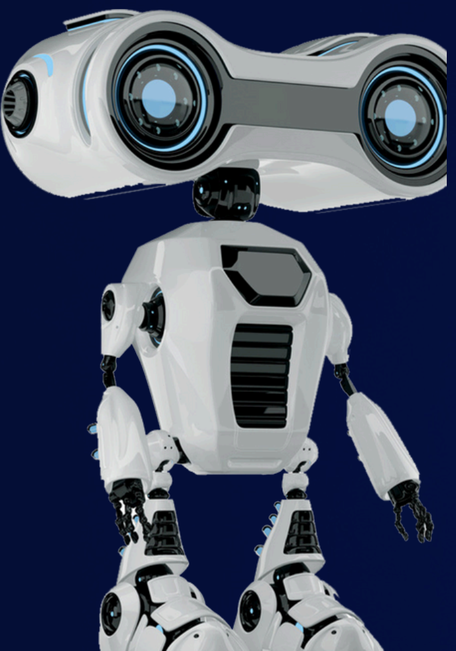

Count Plot for Unit and Designation
Unit
Finance
IT
Marketing
Operations
Web
Management
Count
0
50
100
150
200
250
300
300
333
297
324
303
291
60
45
55
46
45
50
26
24
28
24
26
27
8
14
13
8
5
12
19
13
12
10
12
13
3
7
4
3
6
4
Analyst
Senior Analyst
Associate
Senior Manager
Manager
Director
Unit

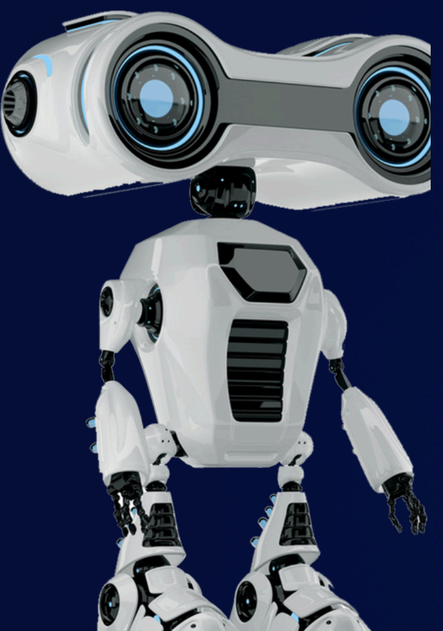


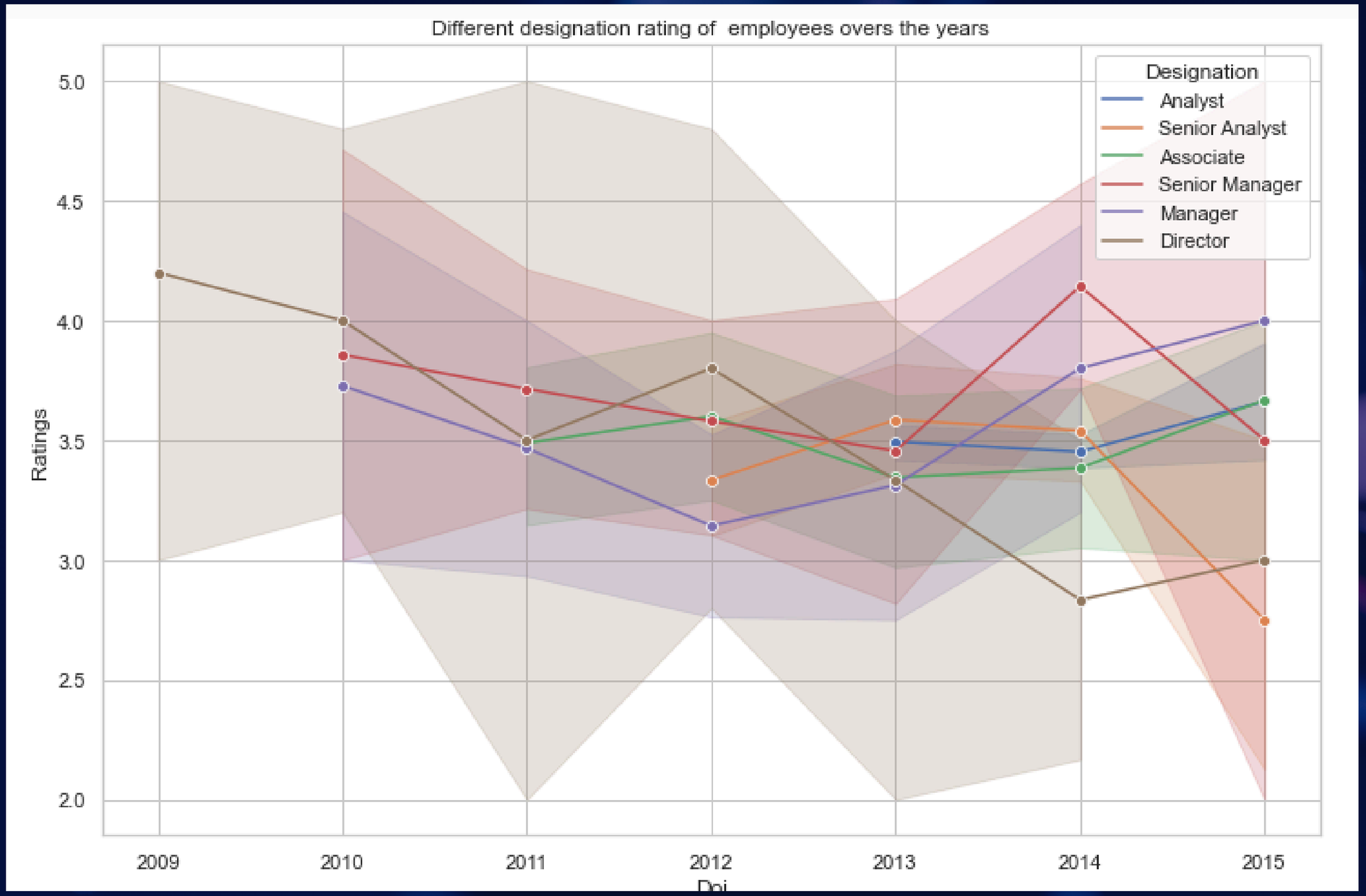

Comaprision of Rating of different Unit
Ratings
5.0
4.5
4.0
3.5
3.0
2.5
2.0
Finance
IT
Marketing
Operations
Web
Management
Unit

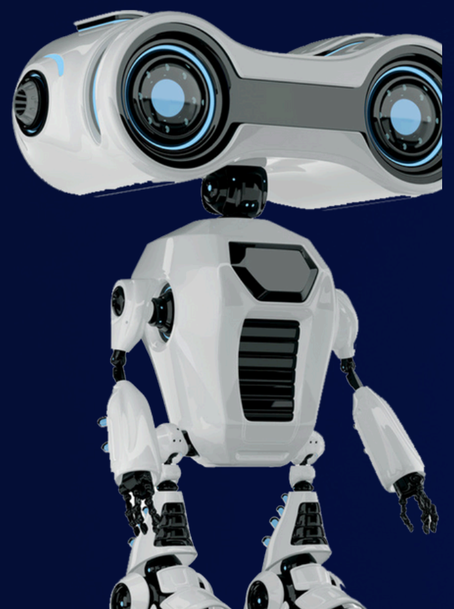
<Figure size 864x432 with 0 Axes>

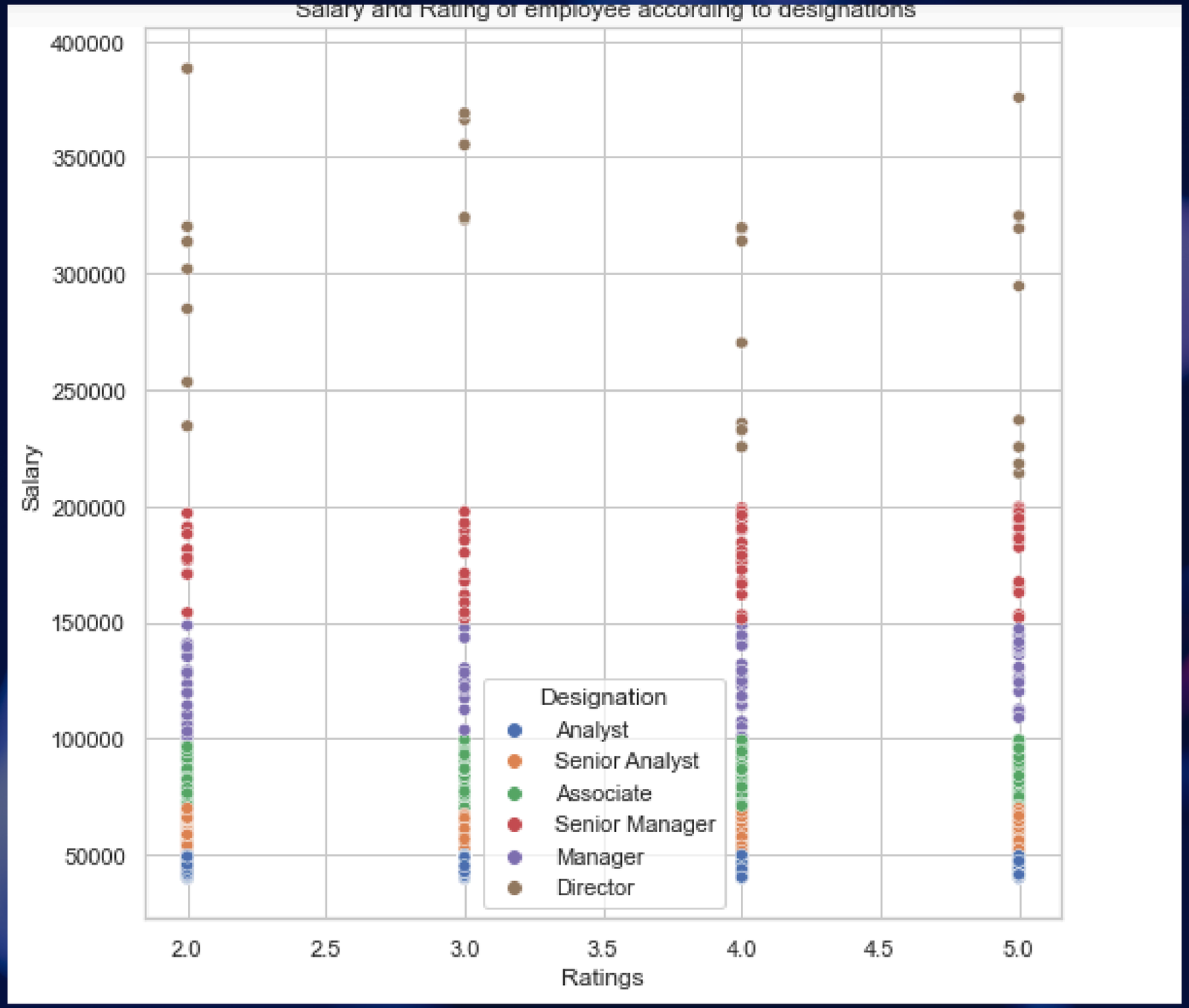







Different designation rating of employees overs the years
Designation
Analyst
Senior Analyst
Associate
Senior Manager
Manager
Director
Ratings
5.0
4.5
4.0
3.5
3.0
2.5
2.0
2009
2010
2011
2012
2013
2014
2015
Dpi




Salary and Rating of employee according to designations
400000
350000
300000
250000
200000
150000
100000
50000
Salary
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Ratings
Designation
Analyst
Senior Analyst
Associate
Senior Manager
Manager
Director

Model Development

ML Model Development

```
x = df.drop(columns = ['index','FirstName','LastName','CurrentDate','LeavesRemaining','Doj','Salary'])
y = df['Salary'].values
```

```
from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
```

```
x['Sex'] = le.fit_transform(x['Sex'])
x['Designation'] = le.fit_transform(x['Designation'])
x['Unit'] = le.fit_transform(x['Unit'])
x = x.values
```

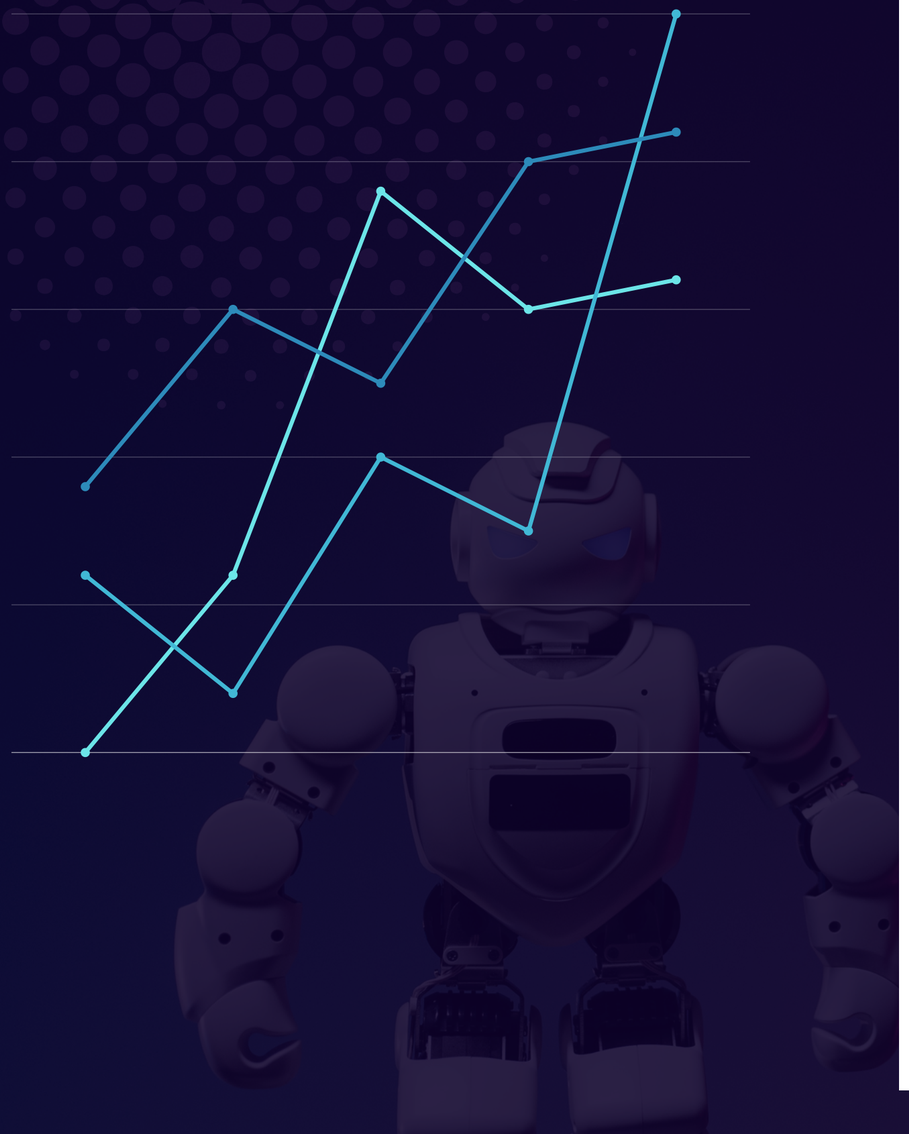
```
x
```

```
array([[ 0.,  0., 21., ...,  2.,  0.,  2.],
       [ 0.,  0., 21., ...,  3.,  0.,  1.],
       [ 0.,  0., 22., ...,  3.,  0.,  3.],
       ...,
       [ 0.,  0., 21., ...,  5.,  0.,  2.],
       [ 0.,  0., 24., ...,  3.,  1.,  2.],
       [ 1.,  0., 24., ...,  2.,  2.,  1.]])
```

```python
models = {'Random Forest' : RandomForestRegressor(max_depth = 10 ,min_samples_leaf = 4,n_estimators = 10),
          'Linear Regression' : LinearRegression(),
          'Decision Tree': DecisionTreeRegressor(max_depth = 10 ,min_samples_leaf = 2,min_samples_split = 10),
          'GradeientBoost': GradientBoostingRegressor(min_samples_split = 10,n_estimators= 50)}
```

```python
for name,model in models.items():
    model.fit(X_train,y_train)
    y_pred = model.predict(X_test)
    mae = mean_absolute_error(y_test,y_pred)
    mse = mean_squared_error(y_test,y_pred)
    rmse = mean_squared_error(y_test,y_pred, squared = False)
    r2 =r2_score(y_test,y_pred)
    print(f' Evaluating {name} model')
    print("Mean absolute Error (MAE) :", mae)
    print("Mean squared Error (MSE) :" ,mse)
    print(" Root Mean squared Error (RMSE) :",rmse)
    print("R Squared error (R2) :",r2)
    print('\n')
```

Evaluation Metrics



```
 Evaluating Random Forest model
Mean absolute Error (MAE) : 3602.4431497168825
Mean squared Error (MSE) : 31803874.223804925
 Root Mean squared Error (RMSE) : 5639.492372882946
R Squared error (R2) : 0.9669763222597569


 Evaluating Linear Regression model
Mean absolute Error (MAE) : 10601.420801059045
Mean squared Error (MSE) : 191750434.92097548
 Root Mean squared Error (RMSE) : 13847.39812820356
R Squared error (R2) : 0.8008951826176555


 Evaluating Decision Tree model
Mean absolute Error (MAE) : 4116.304480419694
Mean squared Error (MSE) : 42925278.649209544
 Root Mean squared Error (RMSE) : 6551.7385974418685
R Squared error (R2) : 0.955428368284748


 Evaluating GradeientBoost model
Mean absolute Error (MAE) : 3653.907678576465
Mean squared Error (MSE) : 31995626.0611464
 Root Mean squared Error (RMSE) : 5656.467631052652
R Squared error (R2) : 0.9667772159861658
```


Thank You!