In [1]:

```
# Task 1 Load and Study the data
# Number of employees, number of features and the type of features
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

In [2]:

```
# Read data
emp = pd.read_csv(r'C:\Desktop\DataAnalytics\UnifiedMentor\Employee Datase
t.csv')
```

In [3]:

```
# Look at the data
emp.head()
```

Out[3]:

	id	groups	age	healthy_eating	active_lifestyle	salary
0	0	A	36	5	5	2297
1	1	A	26	3	5	1134
2	2	A	61	8	1	4969
3	3	O	24	3	6	902
4	4	O	39	6	2	3574

In [4]:

```
# Dimensions of data
emp.shape
```

Out[4]:

```
(50, 6)
```

In [5]:

```
# Rows of data
emp.index
```

Out[5]:

```
RangeIndex(start=0, stop=50, step=1)
```

In [6]:

```
# Columns of data
emp.columns
```

Out[6]:

```
Index(['id', 'groups', 'age', 'healthy_eating', 'active_lifestyle', 'salar
y'], dtype='object')
```

In [7]:

```
# Basic Information
emp.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 50 entries, 0 to 49
Data columns (total 6 columns):
 #   Column            Non-Null Count  Dtype 
---  ------            --------------  ----- 
 0   id                50 non-null     int64 
 1   groups            50 non-null     object
 2   age               50 non-null     int64 
 3   healthy_eating    50 non-null     int64 
 4   active_lifestyle  50 non-null     int64 
 5   salary            50 non-null     int64 
dtypes: int64(5), object(1)
memory usage: 2.5+ KB
```

In [9]:

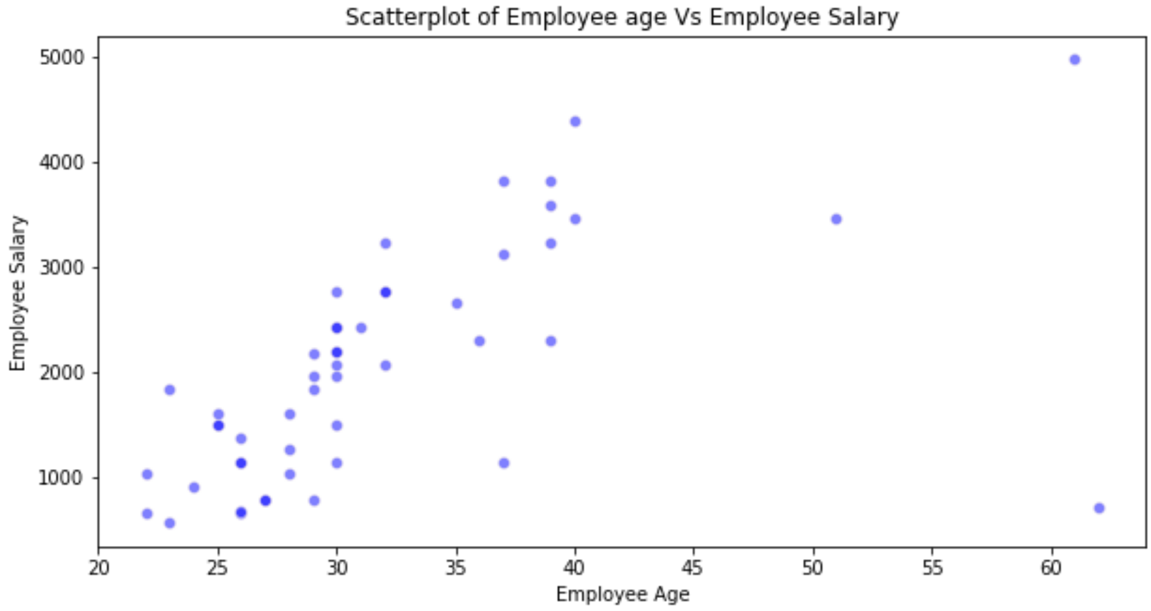
```
# Observation of task 1
# There are 50 rows and 6 columns.Each row contains details of employees
```

In [11]:

```
# Task 2
# Visualise the distributions of ratings and compensations
```

In [12]:

```
# Scatter plot for age and Emp salary
plt.figure(figsize = (10,5))
sns.scatterplot(data = emp, x = 'age', y = 'salary', color = 'blue', edgeco
lor = 'linen', alpha = 0.5)
plt.title("Scatterplot of Employee age Vs Employee Salary")
plt.xlabel('Employee Age')
plt.ylabel('Employee Salary')
plt.show()
```

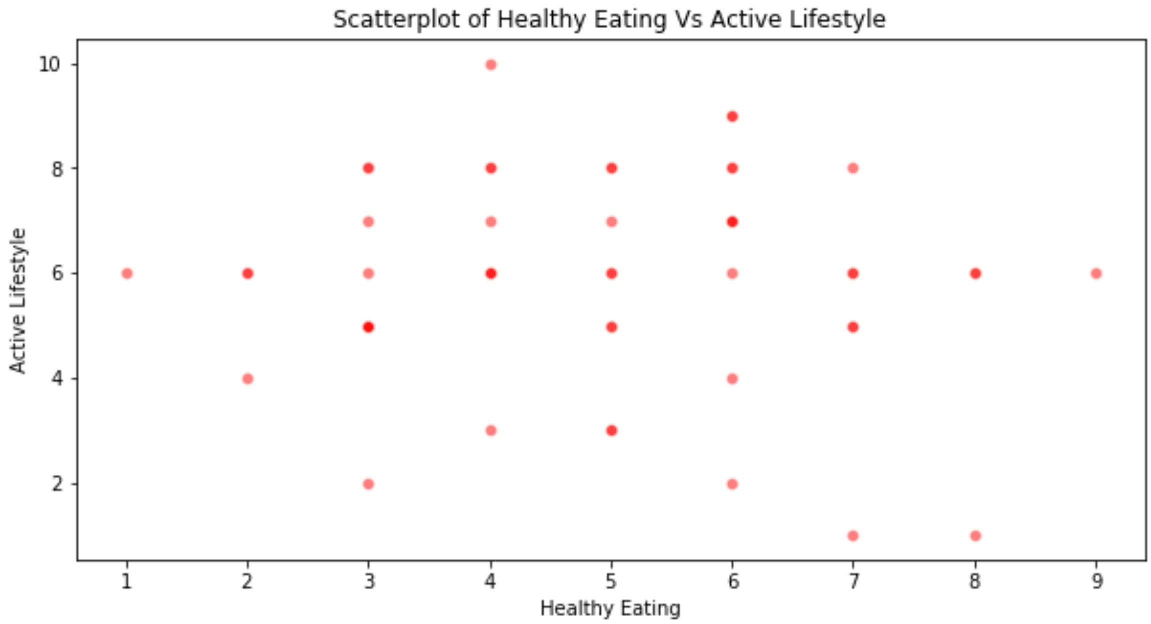


In [13]:

```
# Observation. As age increases salary also increases.
# we can see more employees between 25 to 35 age
```

In [14]:

```
# Scatter plot for healthy eating and active life style
plt.figure(figsize = (10,5))
sns.scatterplot(data = emp, x = 'healthy_eating', y = 'active_lifestyle', c
olor = 'red', edgecolor = 'linen', alpha = 0.5)
plt.title("Scatterplot of Healthy Eating Vs Active Lifestyle")
plt.xlabel('Healthy Eating')
plt.ylabel('Active Lifestyle')
plt.show()
```

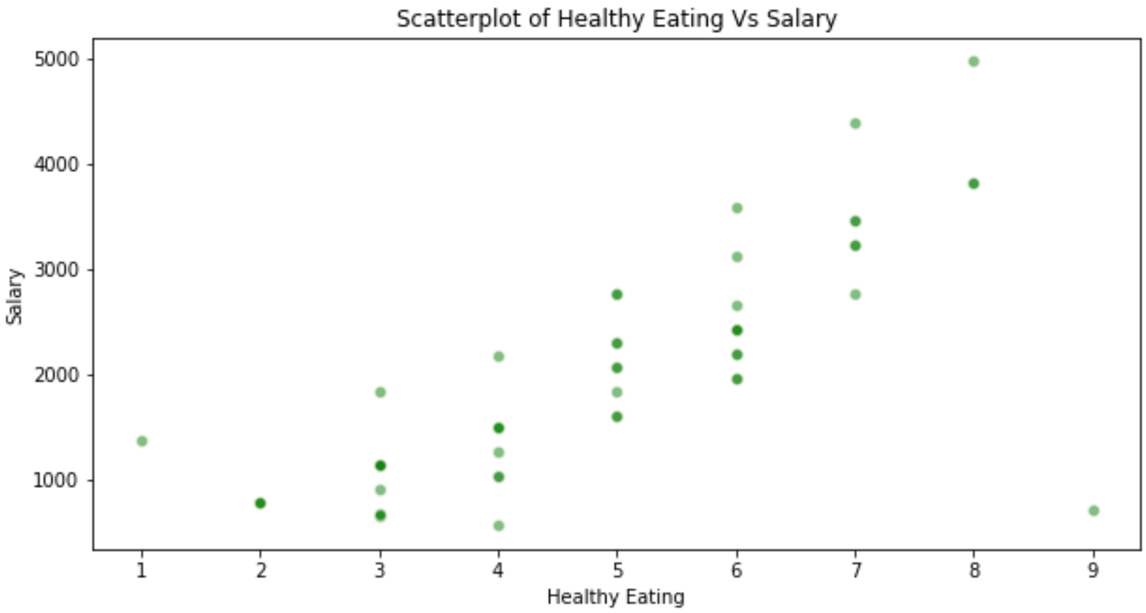


In [15]:

```
# Observations
# As Healthy eating increases active lifestyle increases
```

In [16]:

```
# Scatter plot of healthy eating and Salary
plt.figure(figsize = (10,5))
sns.scatterplot(data = emp, x = 'healthy_eating', y = 'salary', color = 'green', edgecolor = 'linen', alpha = 0.5)
plt.title("Scatterplot of Healthy Eating Vs Salary")
plt.xlabel('Healthy Eating')
plt.ylabel('Salary')
plt.show()
```

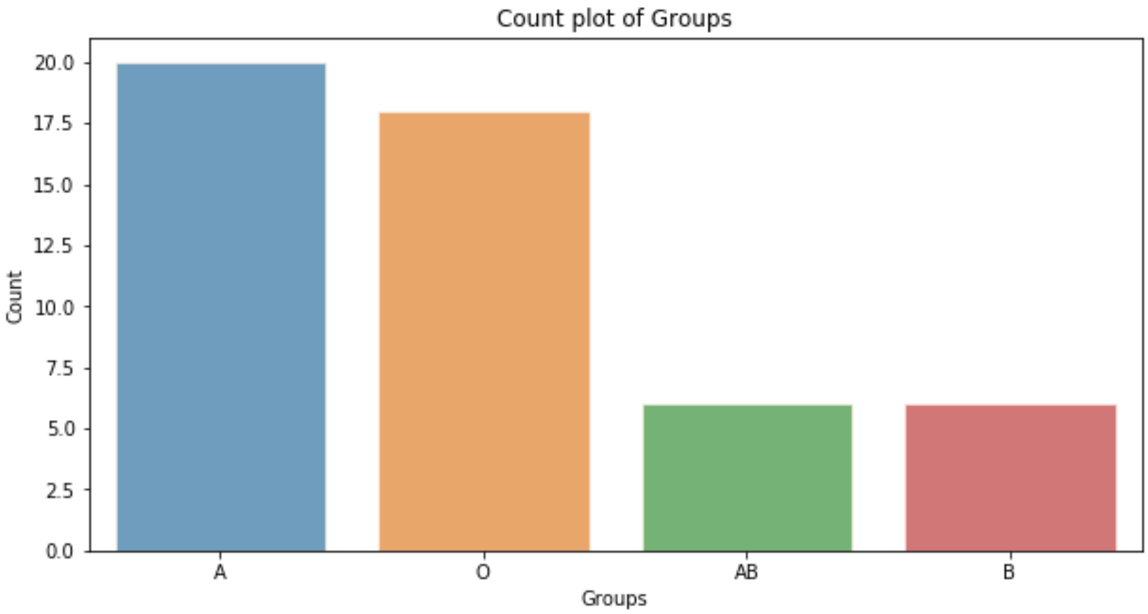


In [18]:

```
# Observations
# As healthy eating increases salary increases.
# Because they take less offs. their productive hours increases
```

In [19]:

```
# Create a count plot of groups
plt.figure(figsize = (10,5))
sns.countplot(data = emp, x = 'groups', edgecolor = 'linen', alpha = 0.7 )
plt.title("Count plot of Groups")
plt.xlabel('Groups')
plt.ylabel('Count')
plt.show()
```

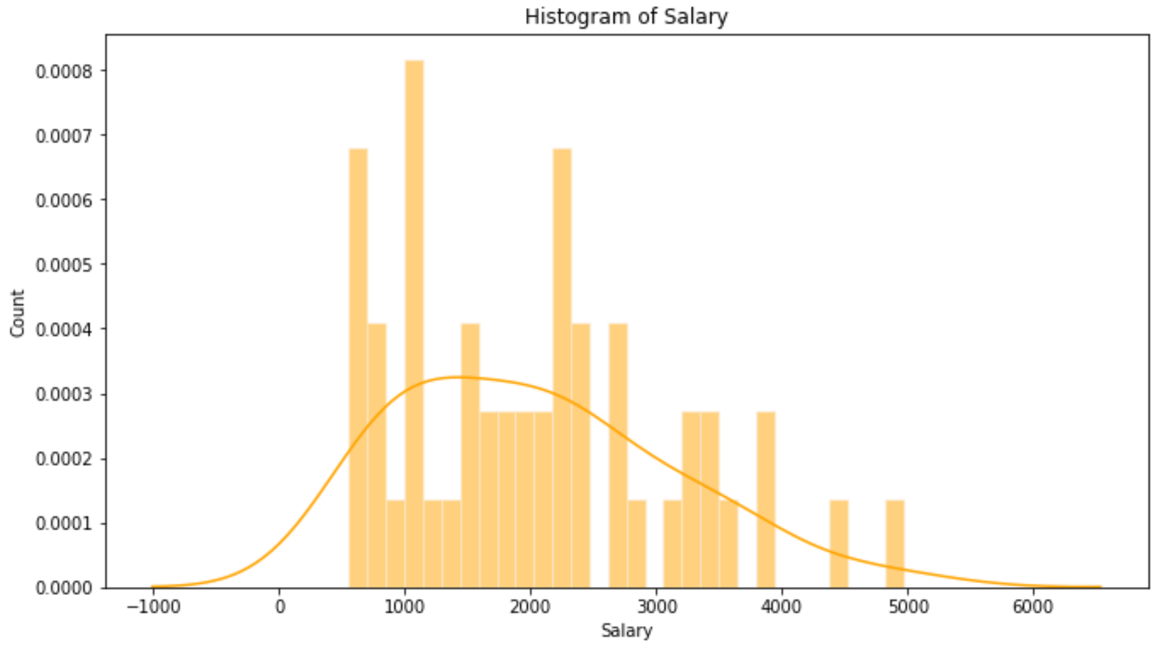


In [20]:

```
# Observation
# We see that most employees either belong to blood group A or O, with grou
p A having maximum frequency
```

In [21]:

```
# Create histogram of Salary
plt.figure(figsize = (11,6))
sns.distplot(emp['salary'], color='orange', hist_kws={'edgecolor': 'linen',
'alpha': 0.5}, bins=30)
plt.title("Histogram of Salary")
plt.xlabel('Salary')
plt.ylabel('Count')
plt.show()
```



In [23]:

```
#Observation
```

In [24]:

```
# Observation Task 2
```

In [25]:

```
# Task 3
# Subset of data based on thresholds.
# We will now subset the original data frame based on the following conditi
ons
# Employee with healthyeating  > 8
# Employee with Salary < 1000
# Employees with healthly > 8 & Salary < 1000
```

In [26]:

```
# Healthy > 8
sub1 = emp[emp['healthy_eating']>8]
sub1
```

Out[26]:

	id	groups	age	healthy_eating	active_lifestyle	salary
26	26	A	62	9	6	700

In [27]:

```
# Salary < 1000
sub2 = emp[emp['salary'] < 1000]
sub2
```

Out[27]:

	id	groups	age	healthy_eating	active_lifestyle	salary
3	3	O	24	3	6	902
15	15	B	26	3	8	662
18	18	A	27	2	6	779
26	26	A	62	9	6	700
32	32	A	22	3	8	662
35	35	O	27	2	4	785
38	38	AB	26	3	7	670
39	39	B	29	2	6	779
43	43	O	23	4	10	556

In [28]:

```
# Both Health> 8 and Salary < 1000
sub = emp[(emp['healthy_eating']>8) & (emp['salary'] < 1000)]
sub
```

Out[28]:

	id	groups	age	healthy_eating	active_lifestyle	salary
26	26	A	62	9	6	700

In [29]:

```
# Observation
 # the only employee seemingly facing a discrepency in salary as compared
#to healthy eating is employee with emp id = 26 and salary 700
```

In [30]:

```
# Final Conclusion
# From the given data, we can use simple visualisations to get a sense of h
ow data are distributed.
# we  can conduct preliminary analyses simply by subseting data sets using
well thought out thresholds and conditions
```