In [1]:
```
import pandas as pd
import numpy as np
```

In [2]:
```
df=pd.read_csv("/Users/akanksha/Documents/study/Reva/Worksheet in 2_HR analytics.csv", sep='|')
```

In [3]:
```
df
```

Out[3]:

	satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spent_company	Work_accident	If employee has left	promotion_last_5years	department	salary bracket
0	0.38	0.53	2	157	3	0	1	0	sales	low
1	0.80	0.86	5	262	6	0	1	0	sales	medium
2	0.11	0.88	7	272	4	0	1	0	sales	medium
3	0.72	0.87	5	223	5	0	1	0	sales	low
4	0.37	0.52	2	159	3	0	1	0	sales	low
5	0.41	0.50	2	153	3	0	1	0	sales	low
6	0.10	0.77	6	247	4	0	1	0	sales	low
7	0.92	0.85	5	259	5	0	1	0	sales	low
8	0.89	1.00	5	224	5	0	1	0	sales	low
9	0.42	0.53	2	142	3	0	1	0	sales	low
10	0.45	0.54	2	135	3	0	1	0	sales	low
11	0.11	0.81	6	305	4	0	1	0	sales	low
12	0.84	0.92	4	234	5	0	1	0	sales	low
13	0.41	0.55	2	148	3	0	1	0	sales	low
14	0.36	0.56	2	137	3	0	1	0	sales	low
15	0.38	0.54	2	143	3	0	1	0	sales	low
16	0.45	0.47	2	160	3	0	1	0	sales	low
17	0.78	0.99	4	255	6	0	1	0	sales	low
18	0.45	0.51	2	160	3	1	1	1	sales	low
19	0.76	0.89	5	262	5	0	1	0	sales	low
20	0.11	0.83	6	282	4	0	1	0	sales	low
21	0.38	0.55	2	147	3	0	1	0	sales	low
22	0.09	0.95	6	304	4	0	1	0	sales	low
23	0.46	0.57	2	139	3	0	1	0	sales	low
24	0.40	0.53	2	158	3	0	1	0	sales	low
25	0.89	0.92	5	242	5	0	1	0	sales	low
26	0.82	0.87	4	239	5	0	1	0	sales	low
27	0.40	0.49	2	135	3	0	1	0	sales	low
28	0.41	0.46	2	128	3	0	1	0	accounting	low
29	0.38	0.50	2	132	3	0	1	0	accounting	low
...	...	...	...	...	...	...	...	...	...	...
14969	0.43	0.46	2	157	3	0	1	0	sales	medium
14970	0.78	0.93	4	225	5	0	1	0	sales	medium
14971	0.39	0.45	2	140	3	0	1	0	sales	medium
14972	0.11	0.97	6	310	4	0	1	0	accounting	medium
14973	0.36	0.52	2	143	3	0	1	0	accounting	medium
14974	0.36	0.54	2	153	3	0	1	0	accounting	medium
14975	0.10	0.79	7	310	4	0	1	0	hr	medium
14976	0.40	0.47	2	136	3	0	1	0	hr	medium
14977	0.81	0.85	4	251	6	0	1	0	hr	medium
14978	0.40	0.47	2	144	3	0	1	0	hr	medium
14979	0.09	0.93	6	296	4	0	1	0	technical	medium
14980	0.76	0.89	5	238	5	0	1	0	technical	high
14981	0.73	0.93	5	162	4	0	1	0	technical	low
14982	0.38	0.49	2	137	3	0	1	0	technical	medium
14983	0.72	0.84	5	257	5	0	1	0	technical	medium
14984	0.40	0.56	2	148	3	0	1	0	technical	medium
14985	0.91	0.99	5	254	5	0	1	0	technical	medium
14986	0.85	0.85	4	247	6	0	1	0	technical	low

	satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spent_company	Work_accident	If employee has left	promotion_last_5years	department	salary bracket
14987	0.90	0.70	5	206	4	0	1	0	technical	low
14988	0.46	0.55	2	145	3	0	1	0	technical	low
14989	0.43	0.57	2	159	3	1	1	0	technical	low
14990	0.89	0.88	5	228	5	1	1	0	support	low
14991	0.09	0.81	6	257	4	0	1	0	support	low
14992	0.40	0.48	2	155	3	0	1	0	support	low
14993	0.76	0.83	6	293	6	0	1	0	support	low
14994	0.40	0.57	2	151	3	0	1	0	support	low
14995	0.37	0.48	2	160	3	0	1	0	support	low
14996	0.37	0.53	2	143	3	0	1	0	support	low
14997	0.11	0.96	6	280	4	0	1	0	support	low
14998	0.37	0.52	2	158	3	0	1	0	support	low

14999 rows × 10 columns

In [4]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14999 entries, 0 to 14998
Data columns (total 10 columns):
satisfaction_level       14999 non-null float64
last_evaluation          14999 non-null float64
number_project           14999 non-null int64
average_montly_hours     14999 non-null int64
time_spent_company       14999 non-null int64
Work_accident            14999 non-null int64
If employee has left     14999 non-null int64
promotion_last_5years    14999 non-null int64
department               14999 non-null object
salary bracket           14999 non-null object
dtypes: float64(2), int64(6), object(2)
memory usage: 1.1+ MB
```

In [5]:
```
df.corr()
```

Out[5]:

	satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spent_company	Work_accident	If employee has left	promotion_last_5years
satisfaction_level	1.000000	0.105021	-0.142970	-0.020048	-0.100866	0.058697	-0.388375	0.025605
last_evaluation	0.105021	1.000000	0.349333	0.339742	0.131591	-0.007104	0.006567	-0.008684
number_project	-0.142970	0.349333	1.000000	0.417211	0.196786	-0.004741	0.023787	-0.006064
average_montly_hours	-0.020048	0.339742	0.417211	1.000000	0.127755	-0.010143	0.071287	-0.003544
time_spent_company	-0.100866	0.131591	0.196786	0.127755	1.000000	0.002120	0.144822	0.067433
Work_accident	0.058697	-0.007104	-0.004741	-0.010143	0.002120	1.000000	-0.154622	0.039245
If employee has left	-0.388375	0.006567	0.023787	0.071287	0.144822	-0.154622	1.000000	-0.061788
promotion_last_5years	0.025605	-0.008684	-0.006064	-0.003544	0.067433	0.039245	-0.061788	1.000000

In [6]:
```
import matplotlib.pyplot as plt
import seaborn as sns
```

In [7]:
```
#finding total employee count
number_of_employees =  df.shape[0]+1
print('Number of employees = ', number_of_employees)
```

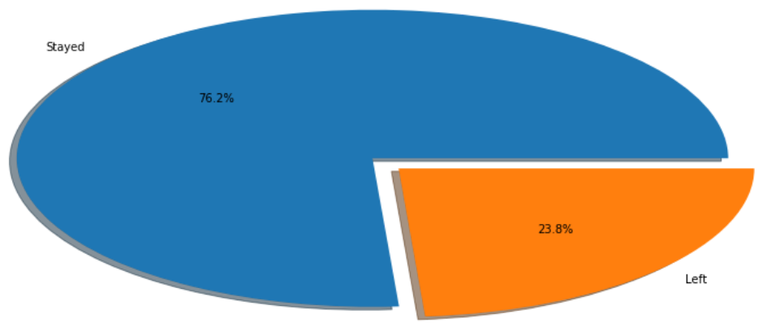
```
Number of employees =  15000
```

In [8]:
```
#finding % of employees who left vs stayed back
employees_who_left = len(df[df['If employee has left'] == 1])
print('Number of employees who have left = ', employees_who_left)
print('Percentage of employees who have left = %0.2f' % (employees_who_left/15000 * 100))
```

```
Number of employees who have left =  3571
Percentage of employees who have left = 23.81
```

In [9]:
```
fig1 = plt.figure(figsize=(14,6))
labels = ['Stayed', 'Left']


plt.pie(df['If employee has left'].value_counts(), explode=[0,0.1], labels=labels,autopct='%1.1f%%', shadow=True)
```

Out[9]:
```
([<matplotlib.patches.Wedge at 0x1a1bd15630>,
  <matplotlib.patches.Wedge at 0x1a1bd15fd0>],
 [Text(-0.806387,0.748158,'Stayed'), Text(0.879695,-0.816172,'Left')],
 [Text(-0.439847,0.408086,'76.2%'), Text(0.513155,-0.4761,'23.8%')])
```

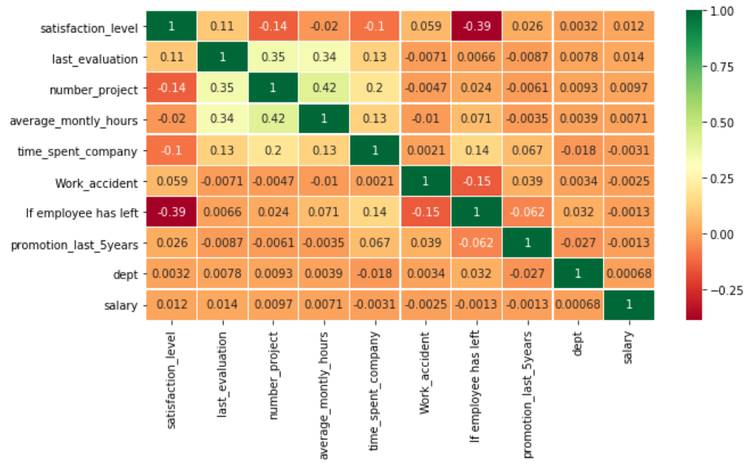


In [10]:
```
#the no.of employees left belonged to any specific salary bracket or department.

from sklearn import preprocessing

le_dept = preprocessing.LabelEncoder()
le_dept.fit(df['department'])
df['dept'] = le_dept.transform(df['department'])

le_salary = preprocessing.LabelEncoder()
le_salary.fit(df['salary bracket'])
df['salary'] = le_salary.transform(df['salary bracket'])

df_x = df.drop(columns=['department', 'salary bracket'])
df_x.head()
```

Out[10]:

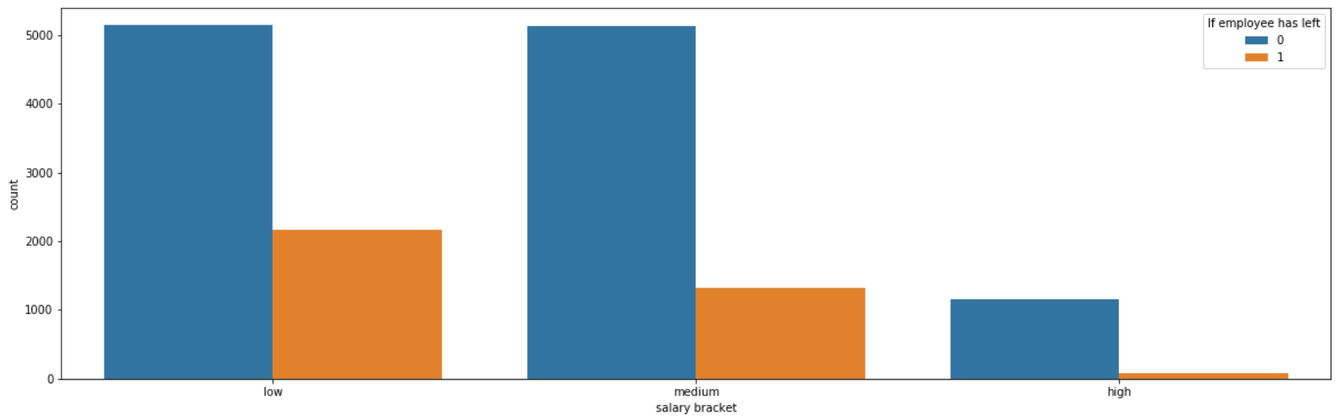
	satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spent_company	Work_accident	If employee has left	promotion_last_5years	dept	salary
0	0.38	0.53	2	157	3	0	1	0	7	1
1	0.80	0.86	5	262	6	0	1	0	7	2
2	0.11	0.88	7	272	4	0	1	0	7	2
3	0.72	0.87	5	223	5	0	1	0	7	1
4	0.37	0.52	2	159	3	0	1	0	7	1

In [11]:
```
fig2 = plt.figure()
sns.heatmap(df.corr(),annot=True,cmap='RdYlGn',linewidths=0.2)
fig2=plt.gcf()
fig2.set_size_inches(10,5)
```



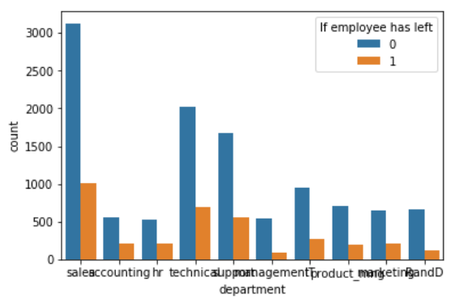
In [12]:
```
from matplotlib import gridspec

fig3 = plt.figure(figsize=(20,6))
gs = gridspec.GridSpec(1, 2, width_ratios=[2, 3])

sns.countplot('salary bracket', hue='If employee has left', data=df)
```

Out[12]: <matplotlib.axes._subplots.AxesSubplot at 0x1a1f64f160>



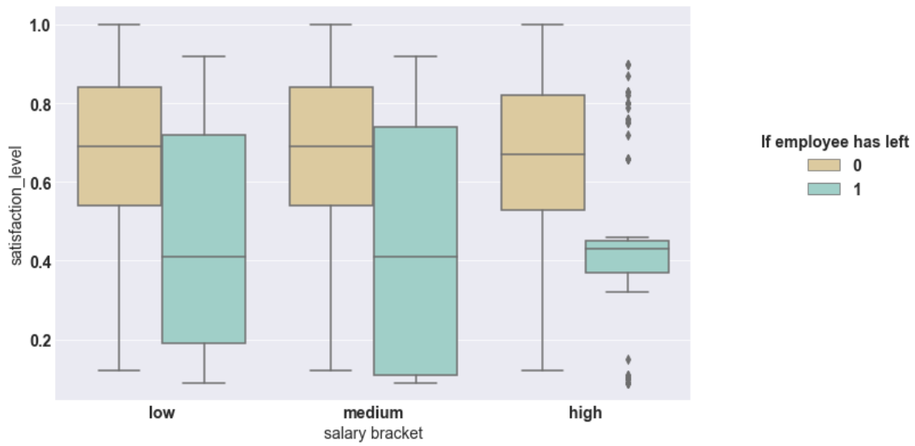
In [13]:
```
sns.countplot('department', hue='If employee has left', data=df)
```

Out[13]: <matplotlib.axes._subplots.AxesSubplot at 0x1a1f8d7710>



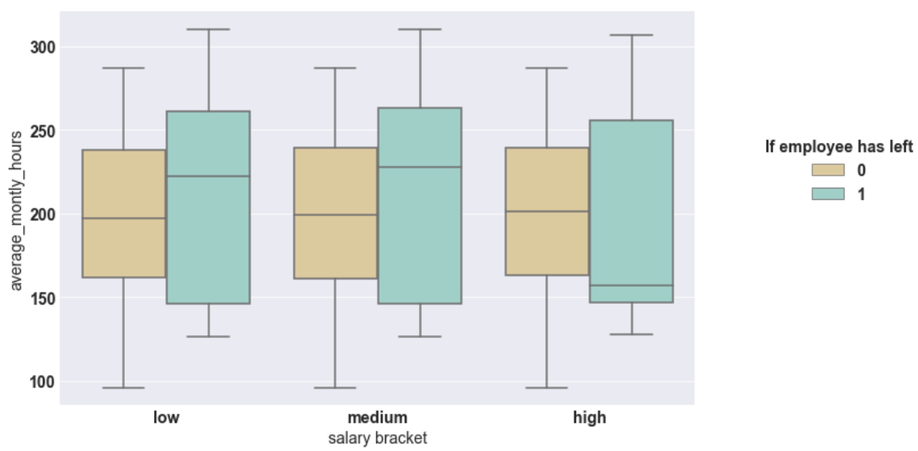
In [14]:
```
#low satisfaction level maybe?

import matplotlib
font = {'family' : 'normal',
        'weight' : 'bold',
        'size'   : 14}
matplotlib.rc('font', **font)
sns.set_style('darkgrid')
fig4 = plt.figure(figsize=(10,14))
ax1 = fig4.add_subplot(2,1,1)
ax1 = sns.boxplot(x='salary bracket', y="satisfaction_level", hue="If employee has left", data=df, palette="BrBG")
ax1.legend(loc=(1.1, 0.5), title='If employee has left')
```

Out[14]: <matplotlib.legend.Legend at 0x1a1fa019b0>



In [15]:
```
#long working hours maybe?

import matplotlib
font = {'family' : 'normal',
        'weight' : 'bold',
        'size'   : 14}
matplotlib.rc('font', **font)
sns.set_style('darkgrid')
fig4 = plt.figure(figsize=(10,14))
ax1 = fig4.add_subplot(2,1,1)
ax1 = sns.boxplot(x='salary bracket', y="average_montly_hours", hue="If employee has left", data=df, palette="BrBG")
ax1.legend(loc=(1.1, 0.5), title='If employee has left')
```

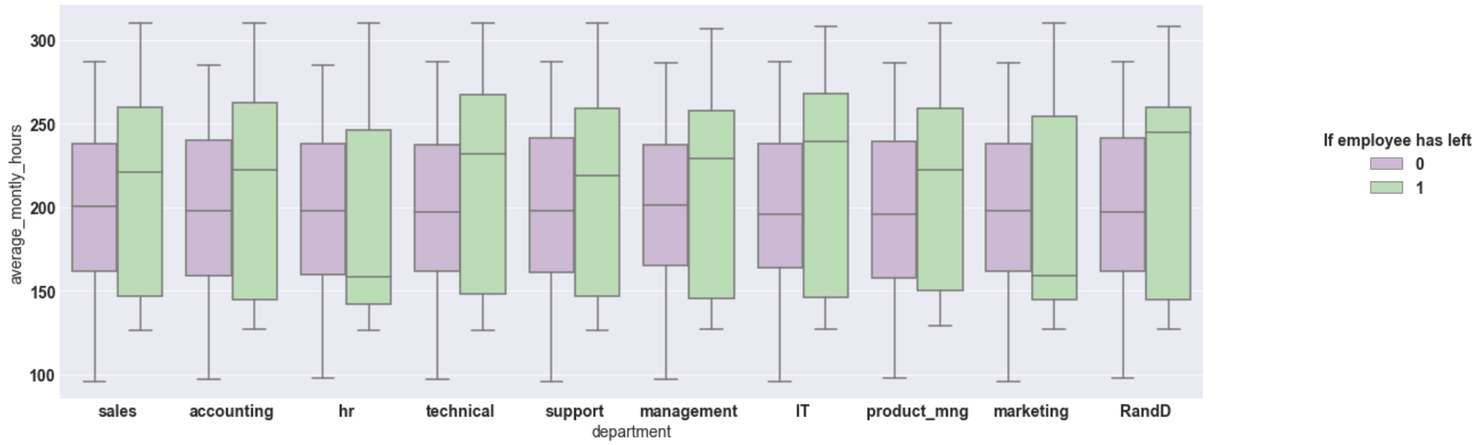
Out[15]: <matplotlib.legend.Legend at 0x1a20330898>



Employees belonging to the low and medium salary brackets who left were working longer hours.

In [17]:
```
fig8 = plt.figure(figsize=(18,14))
ax2 = fig8.add_subplot(2,1,2)
ax2 = sns.boxplot(x='department', y="average_montly_hours", hue="If employee has left", data=df, palette="PRGn")
ax2.legend(loc=(1.1, 0.5), title='If employee has left')
```

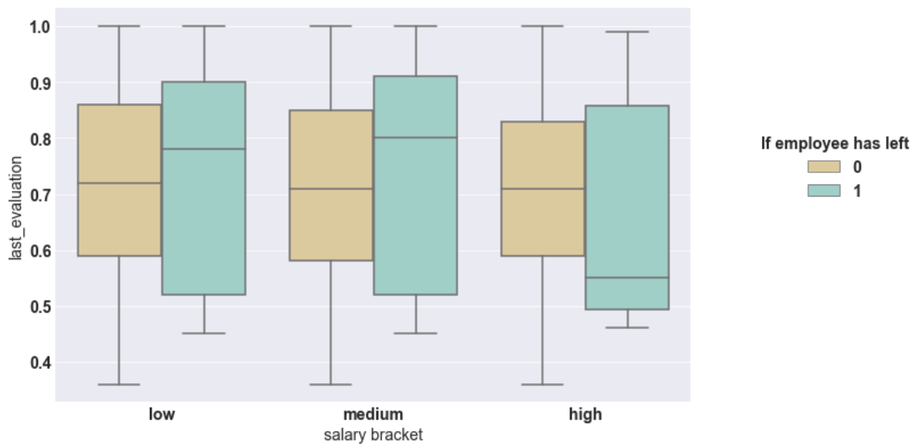
Out[17]: <matplotlib.legend.Legend at 0x1a20d5ec18>



In [ ]:
```
# Employees who left were working longer hours in most departments.
```

In [18]:
```
font = {'family' : 'normal',
        'weight' : 'bold',
        'size'   : 14}
matplotlib.rc('font', **font)
sns.set_style('darkgrid')
fig9 = plt.figure(figsize=(10,14))
ax1 = fig9.add_subplot(2,1,1)
ax1 = sns.boxplot(x='salary bracket', y="last_evaluation", hue="If employee has left", data=df, palette="BrBG")
ax1.legend(loc=(1.1, 0.5), title='If employee has left')
```

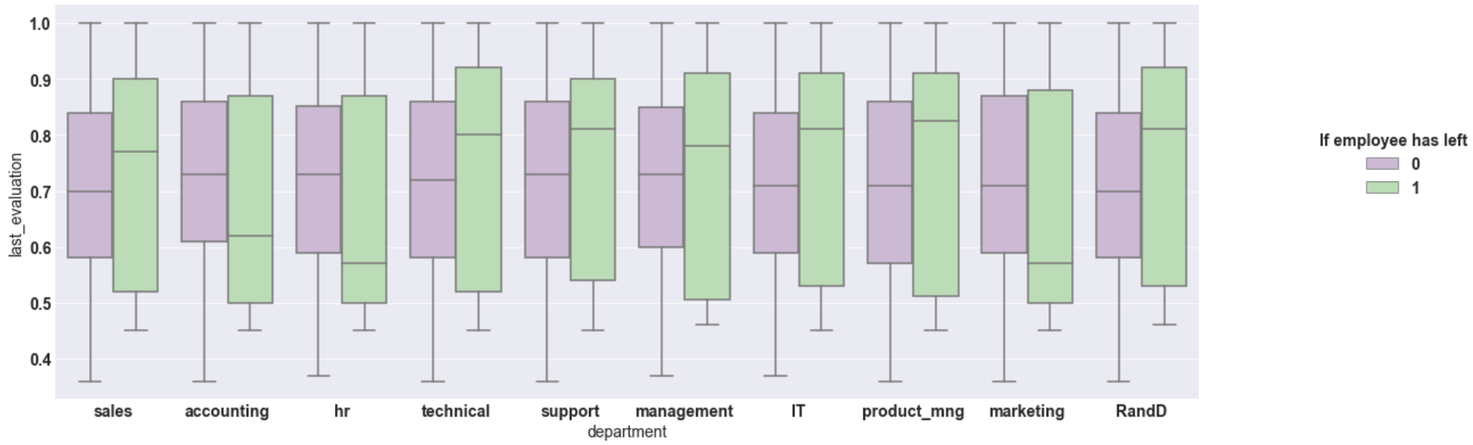
Out[18]: <matplotlib.legend.Legend at 0x1a20f25a90>



In [ ]:
```
# employees belonging to the low and medium salary brackets who left had higher evaluation scores.
```

In [19]:
```
fig10 = plt.figure(figsize=(18,14))
ax2 = fig10.add_subplot(2,1,2)
ax2 = sns.boxplot(x='department', y="last_evaluation", hue="If employee has left", data=df, palette="PRGn")
ax2.legend(loc=(1.1, 0.5), title='If employee has left')
```

Out[19]: <matplotlib.legend.Legend at 0x1a2131a6a0>



In [20]:
```
#Employees who left most departments had higher evaluation scores.
```

In [22]:
```
# inference 1: Employees who have left the company were good at their work, worked longer hours but were highly unsatisfied.
```

In [23]:
```
#inference 2: Employees in the HR and Marketing department were highly demotivated as they didn't work hard
```