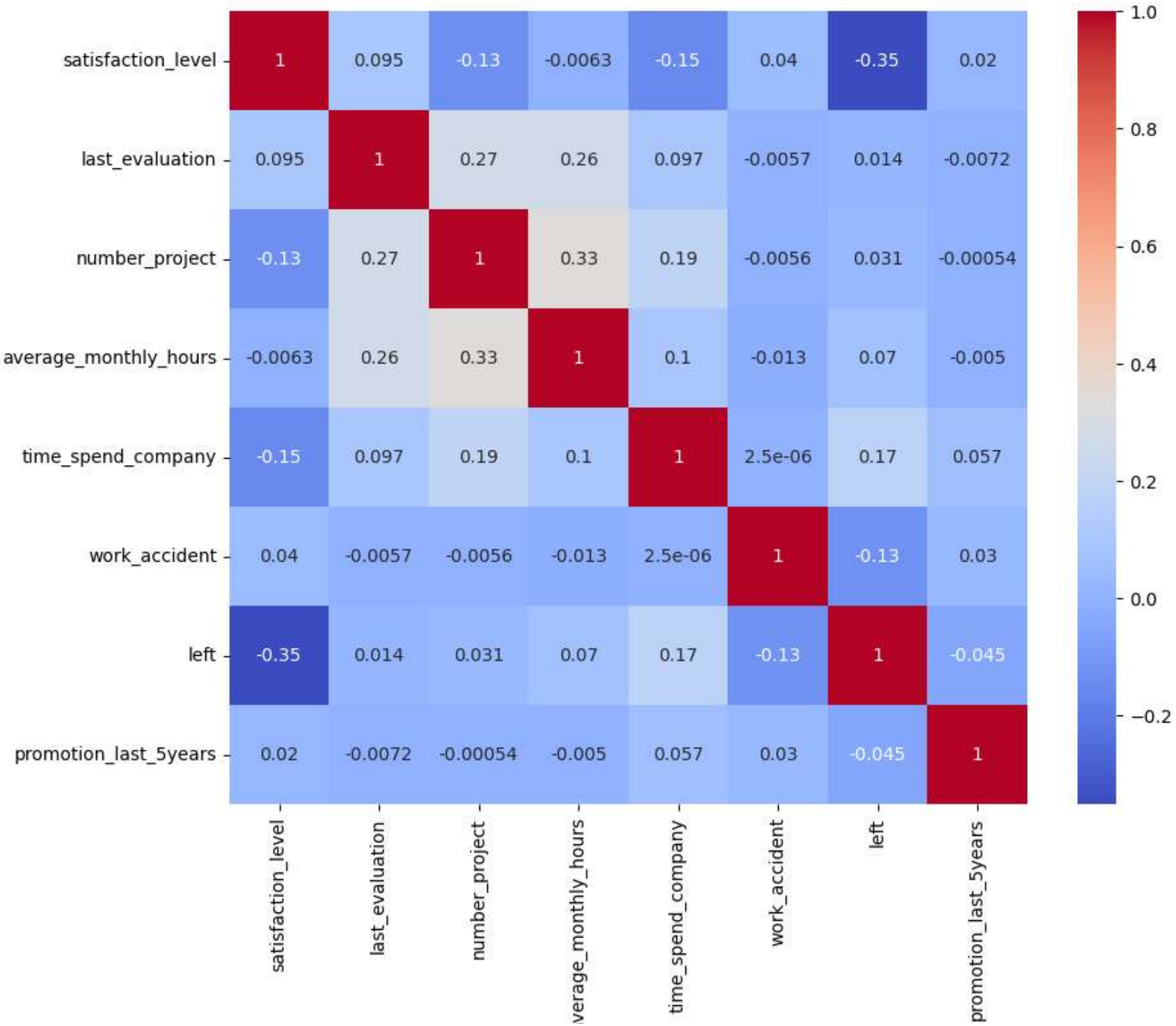
Data visualizations

Our dataset consists of numerical data, and we aim to gain a general understanding of the patterns that exist within it, particularly with regards to employee turnover. Therefore, I will be using a correlation matrix as a starting point to identify any potential correlations between the various variables and employee turnover.

In [92]:
```
# assuming the data is already loaded into a pandas dataframe named df1
corr_matrix = df1.corr(numeric_only=True)

# print the correlation matrix (This is for me personally i like looking at numbers w/o nay )
#print(corr_matrix)

# Plot the correlation matrix as a heatmap
fig, ax = plt.subplots(figsize=(10, 8))
cm = sns.heatmap(corr_matrix, annot=True, cmap='coolwarm')
plt.show()
```


satisfaction_level
last_evaluation
number_project
average_monthly_hours
time_spend_company
work_accident
left
promotion_last_5years
1 0.095 -0.13 -0.0063 -0.15 0.04 -0.35 0.02
0.095 1 0.27 0.26 0.097 -0.0057 0.014 -0.0072
-0.13 0.27 1 0.33 0.19 -0.0056 0.031 -0.00054
-0.0063 0.26 0.33 1 0.1 -0.013 0.07 -0.005
-0.15 0.097 0.19 0.1 1 2.5e-06 0.17 0.057
0.04 -0.0057 -0.0056 -0.013 2.5e-06 1 -0.13 0.03
-0.35 0.014 0.031 0.07 0.17 -0.13 1 -0.045
0.02 -0.0072 -0.00054 -0.005 0.057 0.03 -0.045 1
satisfaction_level
last_evaluation
number_project
verage_monthly_hours
time_spend_company
work_accident
left
promotion_last_5years
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

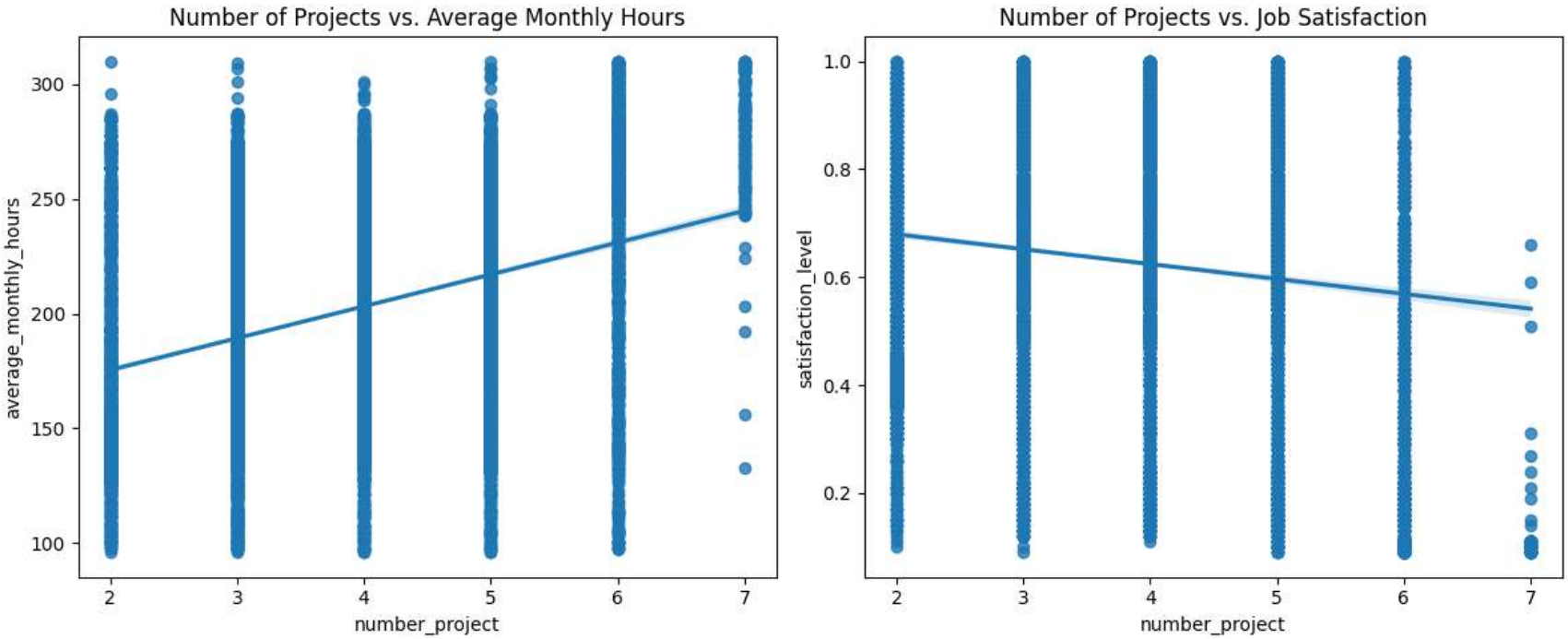
The correlation matrix shows the relationships between the different variables in the dataset. Based on the matrix, we can draw some insights about how different variables are related to each other.

- Employees who have more projects tend to work longer hours, receive higher evaluations, and be slightly more satisfied with their jobs, but also tend to be less satisfied overall and more likely to leave the company.
- Employees who have been with the company longer tend to work on more projects, be slightly more satisfied with their jobs, and be slightly more likely to have been promoted.
- Work accidents have a weak negative correlation with employee turnover.

From the correlation matrix that has been provided, it appears that job satisfaction plays a crucial role in retaining employees. The correlation between job satisfaction and employee retention rate is quite strong, with higher job satisfaction being linked to lower employee turnover. In other words, if employees are more satisfied with their jobs, they are less likely to leave the company. Therefore, improving job satisfaction could be a crucial factor in enhancing employee retention.

To showcase the correlation between the number of projects and the average monthly hours worked, as well as the connection between the number of projects and job satisfaction, we can utilize a scatterplot with a regression line.

In [46]:
```
# Scatterplot with regression line
fig, axs = plt.subplots(ncols=2, figsize=(12, 5))
sns.regplot(x='number_project', y='average_monthly_hours', data=df1, ax=axs[0])
axs[0].set_title('Number of Projects vs. Average Monthly Hours')
sns.regplot(x='number_project', y='satisfaction_level', data=df1, ax=axs[1])
axs[1].set_title('Number of Projects vs. Job Satisfaction')
plt.tight_layout()
plt.show()
```



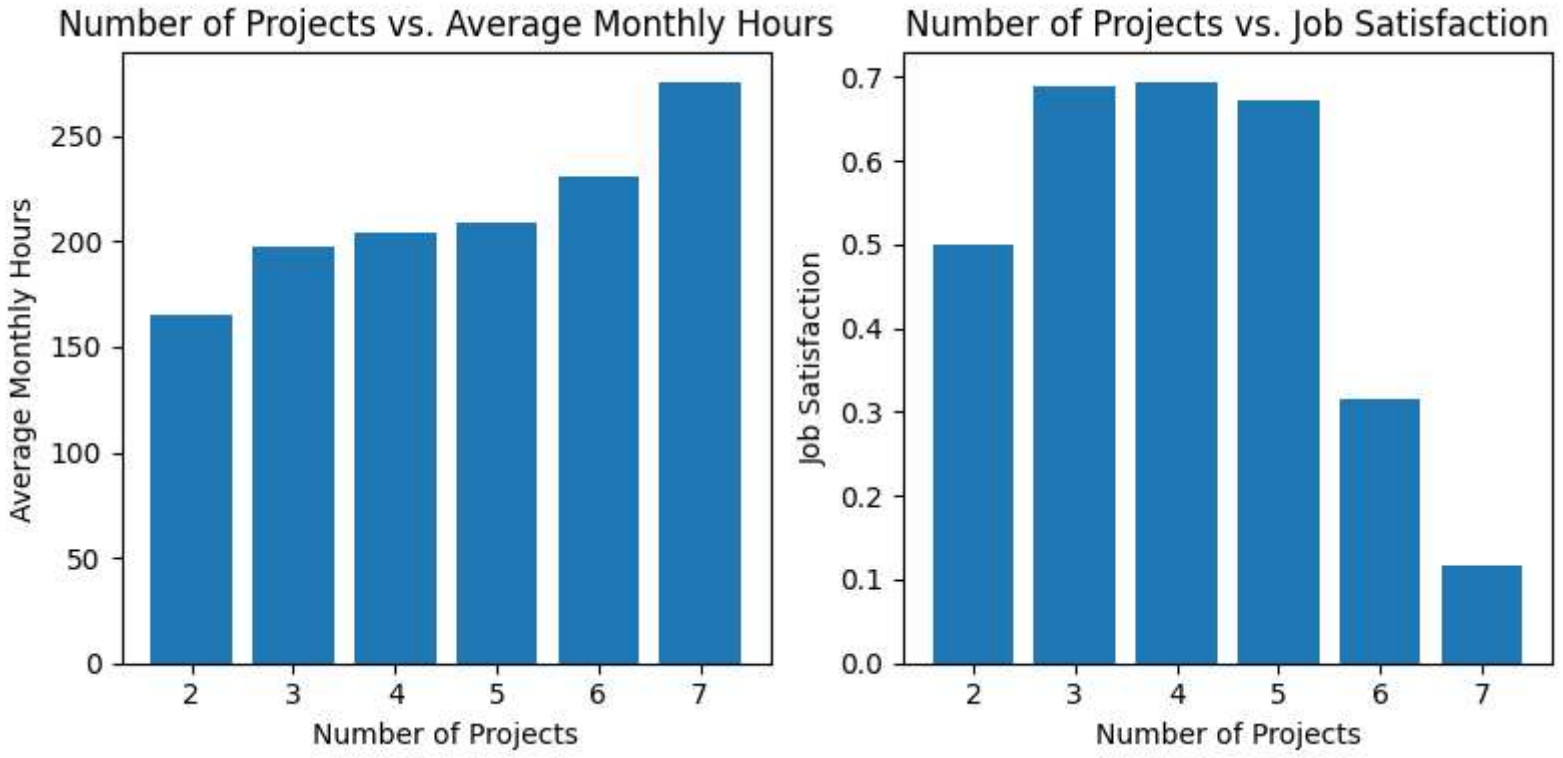
The regression line shows the trends:

- That more projects an employee has the more houres they work.
- More projects they have the less satsfied they are with there job

Let's delve into the average amount of time that employees typically spend working based on the number of projects they are currently managing.

In [90]:
```
project_groups = df1.groupby('number_project')
means = project_groups.mean()

fig, axs = plt.subplots(nrows=1, ncols=2, figsize=(8, 4))

axs[0].bar(np.arange(2, 8), means['average_monthly_hours'])
axs[0].set_xlabel('Number of Projects')
axs[0].set_ylabel('Average Monthly Hours')
axs[0].set_xticks(np.arange(2, 8))
axs[0].set_title('Number of Projects vs. Average Monthly Hours')

axs[1].bar(np.arange(2, 8), means['satisfaction_level'])
```

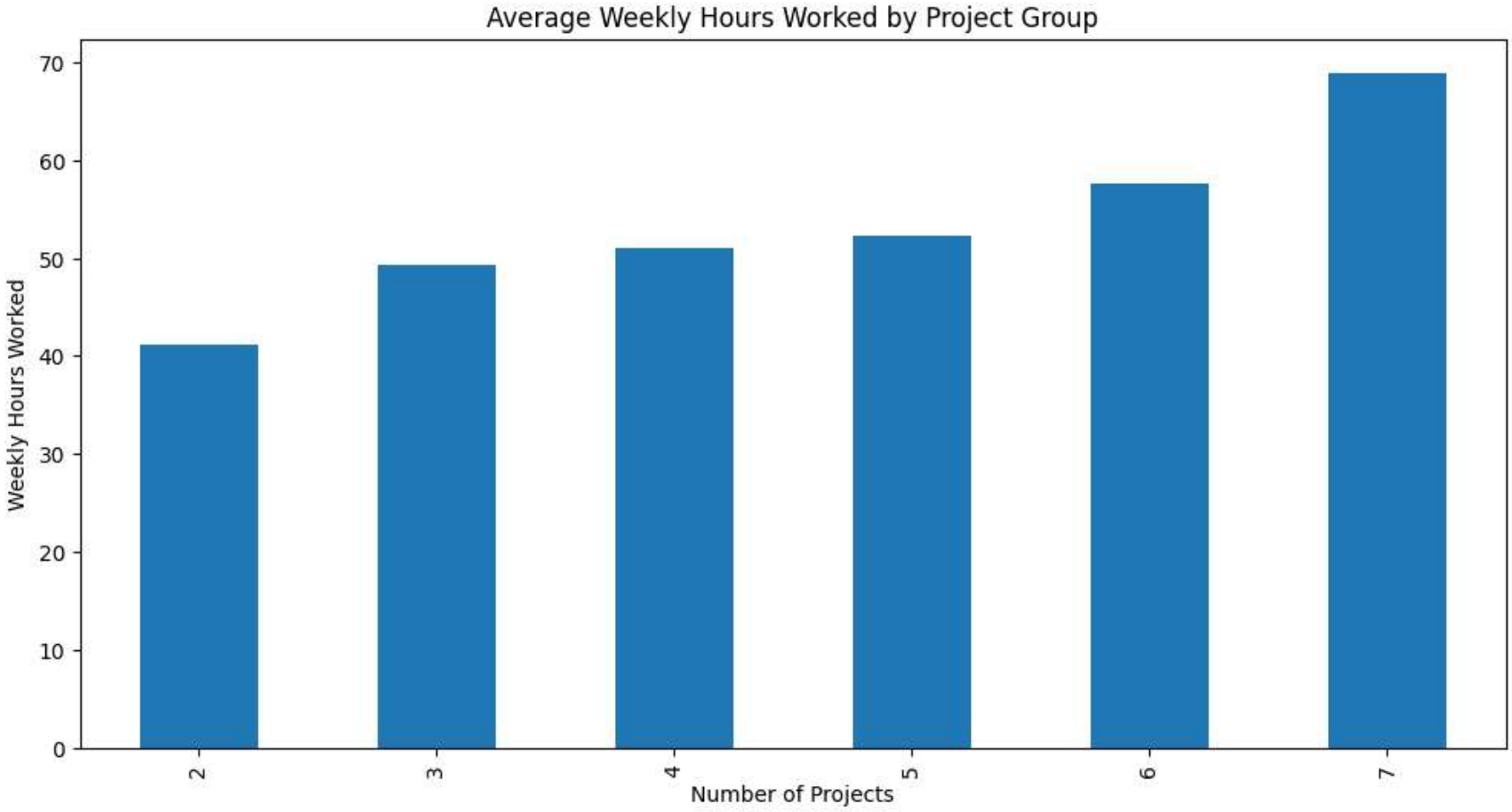
```
axs[1].set_xlabel('Number of Projects')
axs[1].set_ylabel('Job Satisfaction')
axs[1].set_xticks(np.arange(2, 8))
axs[1].set_title('Number of Projects vs. Job Satisfaction')

plt.tight_layout()
plt.show()
```



Upon initial observation, it appears that the employees are experiencing an excessive workload. Let's explore this matter further to gain a better understanding.

In [87]:
```
# Group the data by number of projects and calculate the mean of average monthly hours
grouped = df1.groupby('number_project')['average_monthly_hours'].mean()

# Convert the mean of average monthly hours to weekly hours
grouped_weekly = grouped / 4

# Plot a bar chart of the grouped and converted data
grouped_weekly.plot(kind='bar')
```

```
# Set axis labels and title
plt.xlabel('Number of Projects')
plt.ylabel('Weekly Hours Worked')
plt.title('Average Weekly Hours Worked by Project Group')

# Show the plot
plt.show()
```

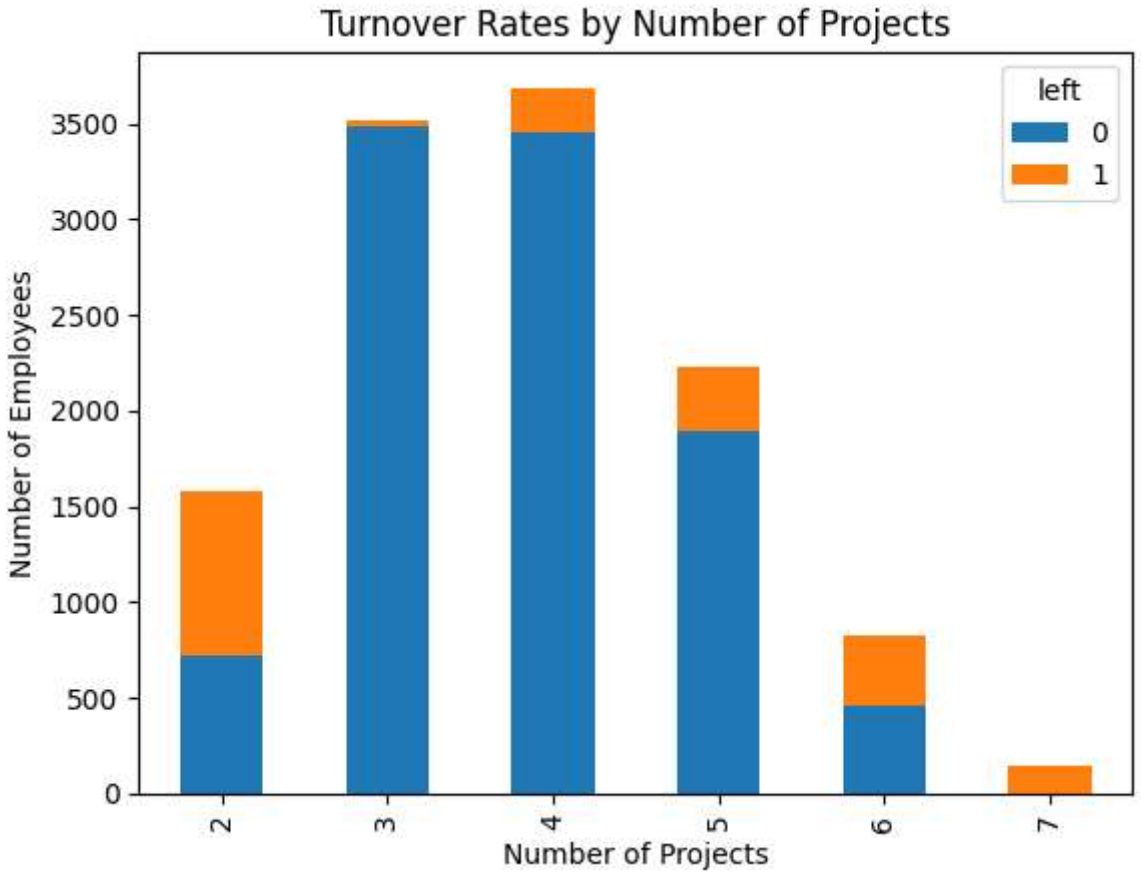


It appears that employees tend to work over 50 hours a week.

The 2 projects managed group seems to be the only one that is working close to 40hrs per week.

let's review the turnover per project to ensure that we haven't overlooked anything.

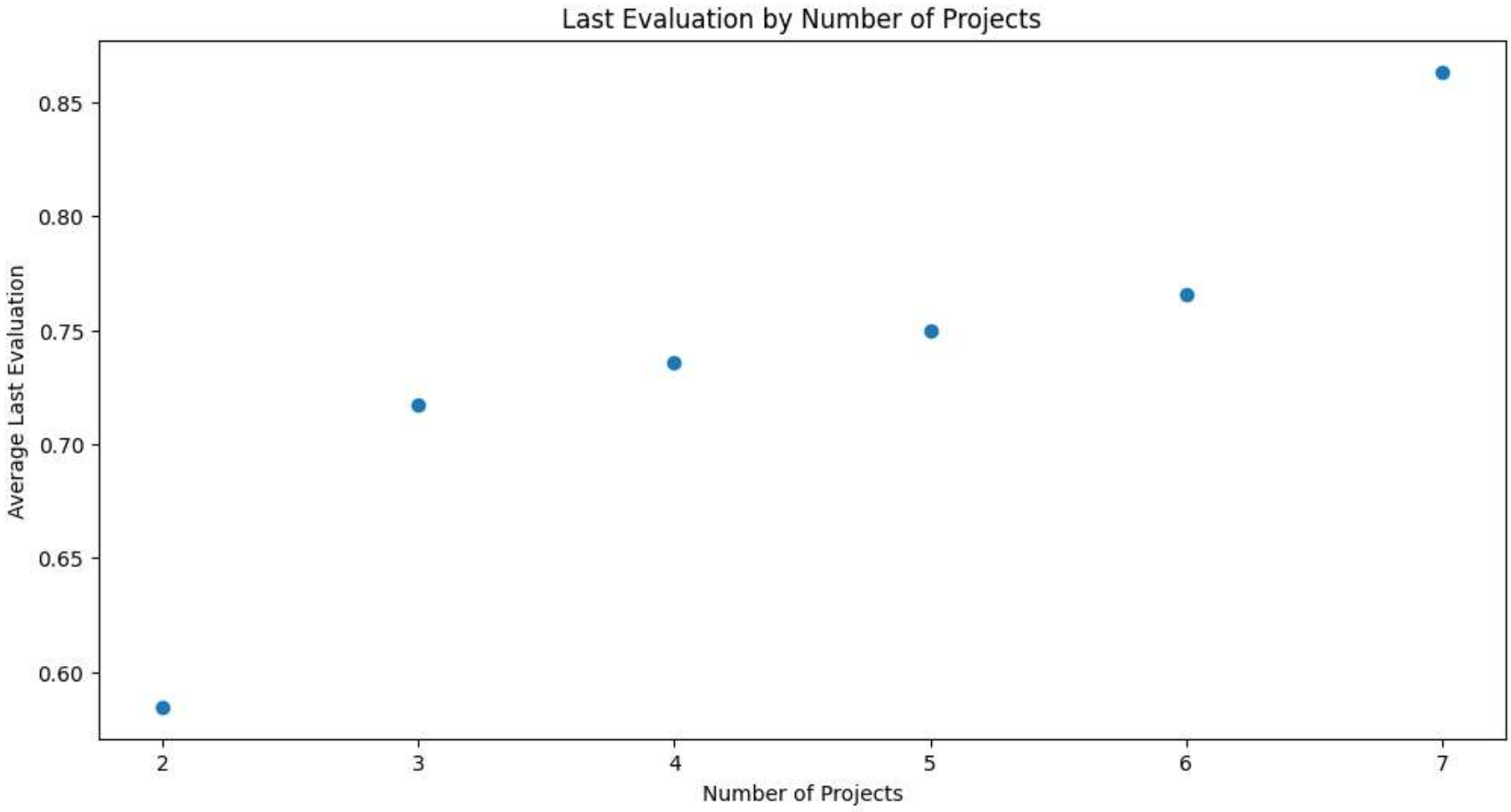
In [36]:
```
left_counts = df1.groupby(['number_project', 'left']).size().unstack()
left_counts.plot(kind='bar', stacked=True)
plt.title('Turnover Rates by Number of Projects')
plt.xlabel('Number of Projects')
plt.ylabel('Number of Employees')
plt.show()
```



This show the proportion of employees who left the company for each number of projects, and to compare this to the proportion who stayed.

We've come across some unexpected data that suggests when employees are assigned only 2 projects, more of them tend to leave compared to those who are assigned 3 projects. One possible explanation for this could be that the employees with 2 projects may not be performing as well, which could be why they end up being let go.

In [85]:
```
# Group by number of projects and calculate mean of last evaluation
df_eval_project = df1.groupby('number_project')['last_evaluation'].mean().reset_index()

# Create a scatter plot
plt.scatter(df_eval_project['number_project'], df_eval_project['last_evaluation'])
plt.xlabel('Number of Projects')
plt.ylabel('Average Last Evaluation')
plt.title('Last Evaluation by Number of Projects')
plt.show()
```



From the graph, it seems clear that individuals who are only handling two projects are not meeting expectations, and there's a good chance that they may be dismissed from their position.