Assignment 3

Rahul Chowdary Velaga - 120377273

```
import sqlite3
from google.colab import drive
drive.mount('/content/gdrive')
path = '/content/gdrive/MyDrive/lahman_1871-2022.sqlite'
conn = sqlite3.connect(path)

```

```
Drive already mounted at /content/gdrive; to attempt to forcibly remount, call drive.mount("/content/gdrive", force_remount=True).
```

Part 1:

Problem 1:

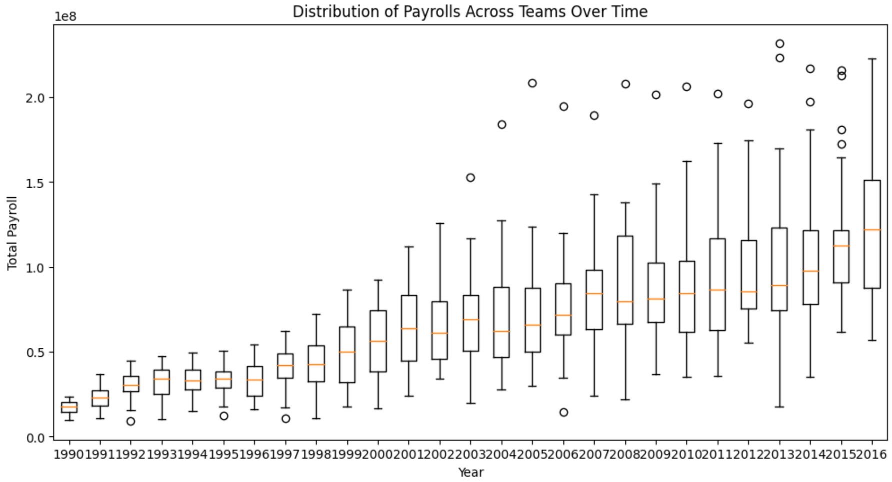
```
# Import the sqlite3 library
import sqlite3
import pandas as pd

# Set the path to your SQLite database
path = '/content/gdrive/MyDrive/lahman_1871-2022.sqlite'

# Connect to the SQLite database
conn = sqlite3.connect(path)

# Write the SQL query
sql_query = '''
SELECT
    t.teamID,
    t.yearID,
    t.franchID,
    t.W,
    t.G,
    s.total_payroll,
    t.W * 100.0 / t.G AS winning_percentage
FROM
    Teams t
JOIN
    (
        SELECT
            teamID,
            yearID,
            SUM(salary) AS total_payroll
        FROM
            Salaries
        GROUP BY
            teamID, yearID
    ) s
ON
    t.teamID = s.teamID AND t.yearID = s.yearID
WHERE
    t.G > 0;
'''
data = pd.read_sql_query(sql_query, conn)
data

```

	teamID	yearID	franchID	W	G	total_payroll	winning_percentage
0	ATL	1985	ATL	66	162	14807000.0	40.740741
1	BAL	1985	BAL	83	161	11560712.0	51.552795
2	BOS	1985	BOS	81	163	10897560.0	49.693252

Part 2:

Problem 2:

```
import matplotlib.pyplot as plt
new_data = data[(data['yearID'] >= 1990) & (data['yearID'] <= 2022)]
plt.figure(figsize=(12, 6))
plt.boxplot([new_data[new_data['yearID'] == year]['total_payroll'] for year in new_data['yearID'].unique()], labels=new_data['yearID'].ur
plt.xlabel('Year')
plt.ylabel('Total Payroll')
plt.title('Distribution of Payrolls Across Teams Over Time')
plt.show()
data
```



	teamID	yearID	franchID	W	G	total_payroll	winning_percentage
0	ATL	1985	ATL	66	162	14807000.0	40.740741
1	BAL	1985	BAL	83	161	11560712.0	51.552795
2	BOS	1985	BOS	81	163	10897560.0	49.693252
3	CAL	1985	ANA	90	162	14427894.0	55.555556
4	CHA	1985	CHW	85	163	9846178.0	52.147239
...	...	...	...	...	...	...	...
913	SLN	2016	STL	86	162	143053500.0	53.086420
914	TBA	2016	TBD	68	162	57097310.0	41.975309
915	TEX	2016	TEX	95	162	176038723.0	58.641975
916	TOR	2016	TOR	89	162	138701700.0	54.938272
917	WAS	2016	WSN	95	162	141652646.0	58.641975

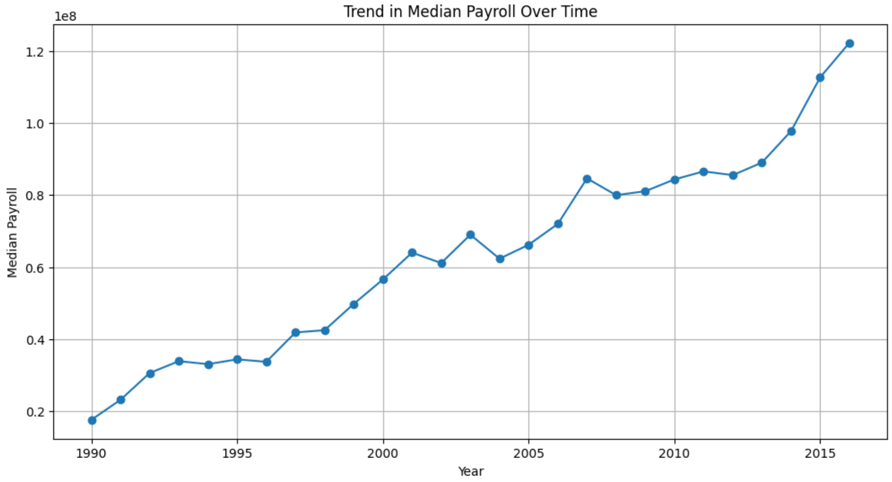
918 rows × 7 columns

Question 1: As the years progressed, the median tends to increase gradually. The change in the width of the inter-quartile range does not increase or decrease linearly and is different for different years.

Also from 2003 we can see outliers above the inter-quartile range which indicates that those players are paid way more than an average player.

Problem 3:

```
payroll_median = new_data.groupby('yearID')['total_payroll'].median().reset_index()
plt.figure(figsize=(12, 6))
plt.plot(payroll_median['yearID'], payroll_median['total_payroll'], marker='o')
plt.xlabel('Year')
plt.ylabel('Median Payroll')
plt.title('Trend in Median Payroll Over Time')
plt.grid(True)
plt.show()
data.head()
```

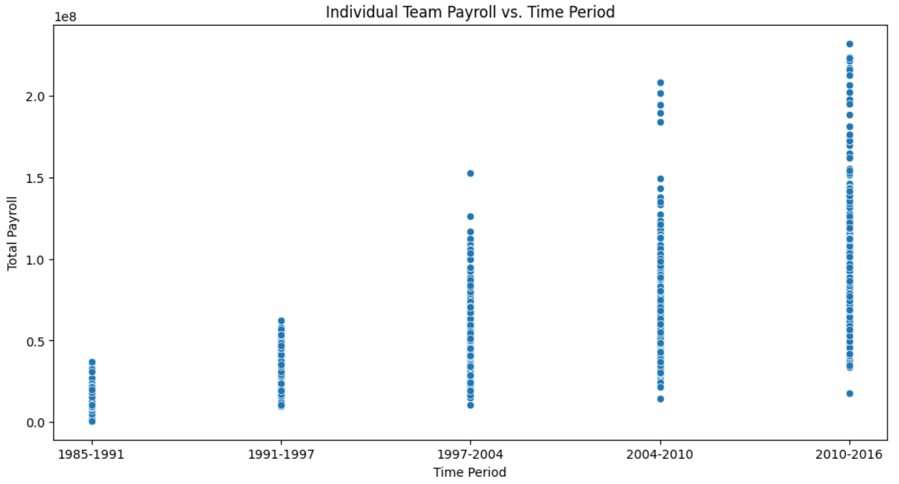


	teamID	yearID	franchID	W	G	total_payroll	winning_percentage
0	ATL	1985	ATL	66	162	14807000.0	40.740741
1	BAL	1985	BAL	83	161	11560712.0	51.552795
2	BOS	1985	BOS	81	163	10897560.0	49.693252
3	CAL	1985	ANA	90	162	14427894.0	55.555556
4	CHA	1985	CHW	85	163	9846178.0	52.147239

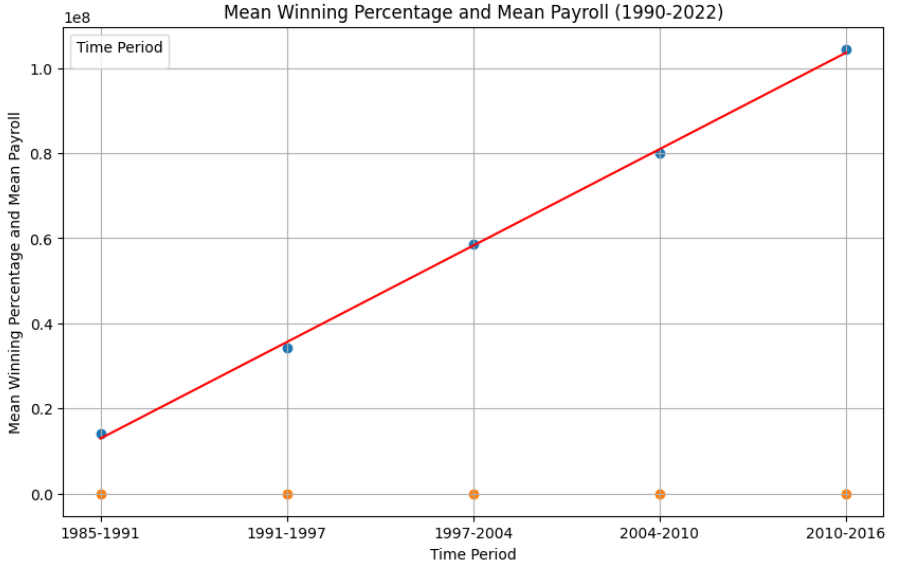
Problem 4:

```
import pandas as pd
import numpy as np
from sklearn.linear_model import LinearRegression
import seaborn as sns

data['time_period'] = pd.cut(data['yearID'], bins=5)
labels = data['time_period'].unique()

xlabels = []

for label in labels:
  xlabels.append('{}-{}'.format(round(label.left), round(label.right)))
data['time_period'] = pd.cut(data['yearID'], bins=5, labels=xlabels)
labels = data['time_period'].unique()

mean_winning_percentage = data.groupby('time_period')['winning_percentage'].mean()
mean_payroll = data.groupby('time_period')['total_payroll'].mean()

```

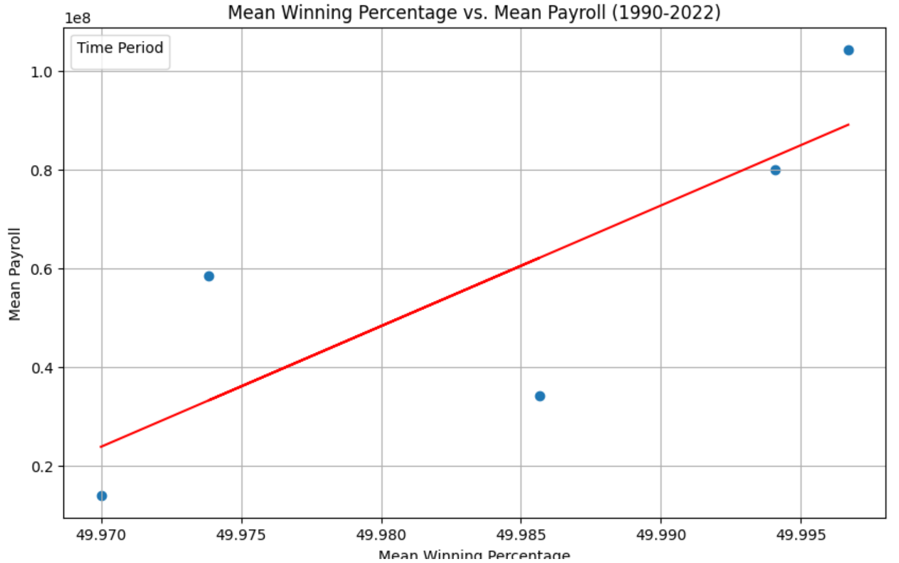
```
# payroll vs time
plt.figure(figsize=(12, 6))
sns.scatterplot(x='time_period', y='total_payroll', data=data)
plt.title('Individual Team Payroll vs. Time Period')
plt.xlabel('Time Period')
plt.ylabel('Total Payroll')
plt.show()


# payroll and win percentage vs time: regression
plt.figure(figsize=(10, 6))
plt.scatter(xlabels, list(mean_payroll))
plt.scatter(xlabels, list(mean_winning_percentage))


plt.title('Mean Winning Percentage and Mean Payroll (1990-2022)')
plt.xlabel('Time Period')
plt.ylabel('Mean Winning Percentage and Mean Payroll')
plt.legend(title='Time Period')
plt.grid(True)

x = np.array(range(len(mean_payroll)))
# x = mean_winning_percentage
y = mean_payroll.values
fit = np.polyfit(x, y, deg=1)
plt.plot(x, fit[0] * x + fit[1], color='red')
plt.show()

# win percentage vs payroll
plt.figure(figsize=(10, 6))
plt.scatter(list(mean_winning_percentage), list(mean_payroll))

plt.title('Mean Winning Percentage vs. Mean Payroll (1990-2022)')
plt.ylabel('Mean Payroll')
plt.xlabel('Mean Winning Percentage')
plt.legend(title='Time Period')
plt.grid(True)

x = mean_winning_percentage
y = mean_payroll.values
fit = np.polyfit(x, y, deg=1)
plt.plot(x, fit[0] * x + fit[1], color='red')
plt.show()
data.head()
```



```
WARNING:matplotlib.legend:No artists with labels found to put in legend.  Note that arti
```

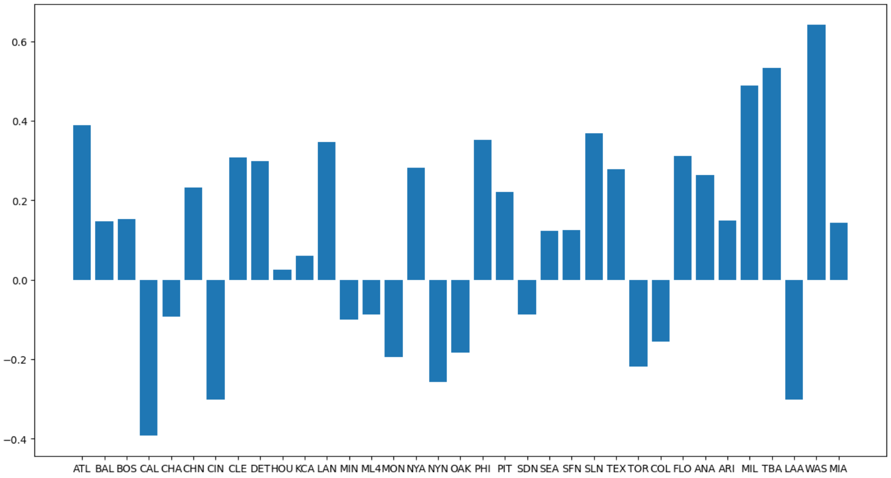


```
WARNING:matplotlib.legend:No artists with labels found to put in legend.  Note that arti
```



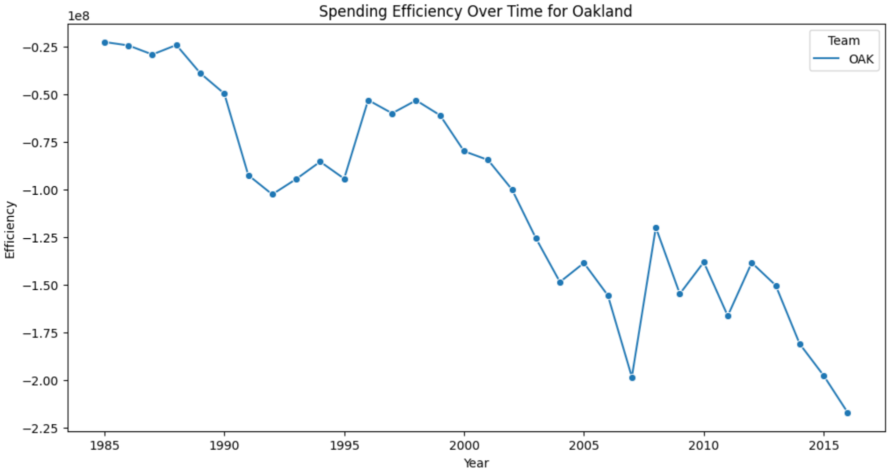
Question 2: The mean payroll of each team has increased over the years but the win percentage has remained almost constant.

```
teams = data['teamID'].unique()
correlation = []
for team_id in teams:
  team_data = data[data['teamID'] == team_id]
  cor = team_data['winning_percentage'].corr(team_data['total_payroll'])
  correlation.append([team_id, cor])

corr = pd.DataFrame(correlation, columns = ['team_id', 'corr'])

plt.figure(figsize = (15, 8))
plt.bar(corr['team_id'], corr['corr'])
plt.show()
```



From the above bar plot, the data on the +ve the y axis indicate that the more money the team spent, the more wins they secured. Similarly the data on -ve y axis indicate that the teams spending more money did not result in more wins. MIA, TBA and MIL were particularly good with paying for their wins.

```
data['expected_win_pct'] = 50 + 2.5 * data['total_payroll']
data['efficiency'] = data['winning_percentage'] - data['expected_win_pct']

selected_team = ['OAK']
selected_team_data = data[data['teamID'].isin(selected_team)]

plt.figure(figsize=(12, 6))
sns.lineplot(x='yearID', y='efficiency', hue='teamID', data=selected_team_data, marker='o')

plt.xlabel('Year')
plt.ylabel('Efficiency')
plt.title('Spending Efficiency Over Time for Oakland')
plt.legend(title='Team')
plt.show()
```



Oakland's spending efficieny decreased non-linearly over the years.

Part 3:

Problem 5:

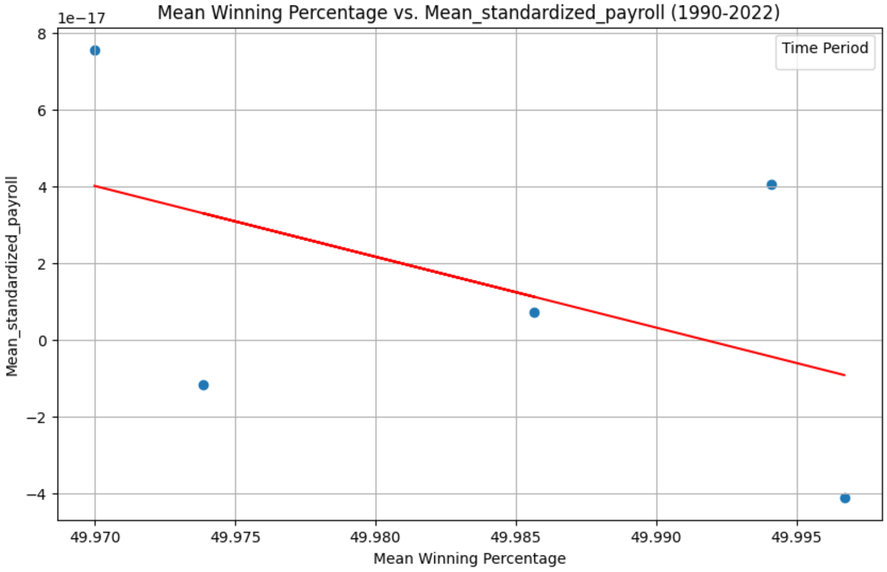
```
data['avg_payroll'] = data.groupby('yearID')['total_payroll'].transform('mean')
data['std_payroll'] = data.groupby('yearID')['total_payroll'].transform('std')
data['standardized_payroll'] = (data['total_payroll'] - data['avg_payroll']) / data['std_payroll']
data.head()

```

	teamID	yearID	franchID	W	G	total_payroll	winning_percentage	time_period	ex
0	ATL	1985	ATL	66	162	14807000.0	40.740741	1985-1991	
1	BAL	1985	BAL	83	161	11560712.0	51.552795	1985-1991	
2	BOS	1985	BOS	81	163	10897560.0	49.693252	1985-1991	
3	CAL	1985	ANA	90	162	14427894.0	55.555556	1985-1991	
4	CHA	1985	CHW	85	163	9846178.0	52.147239	1985-1991	

Problem 6:

```
data['time_period'] = pd.cut(data['yearID'], bins=5)
labels = data['time_period'].unique()

xlabels = []

for label in labels:
  xlabels.append('{}-{}'.format(round(label.left), round(label.right)))
data['time_period'] = pd.cut(data['yearID'], bins=5, labels=xlabels)
labels = data['time_period'].unique()

mean_winning_percentage = data.groupby('time_period')['winning_percentage'].mean()
mean_standardized_payroll = data.groupby('time_period')['standardized_payroll'].mean()



plt.figure(figsize=(10, 6))
plt.scatter(list(mean_winning_percentage), list(mean_standardized_payroll))

plt.title('Mean Winning Percentage vs. Mean_standardized_payroll (1990-2022)')
plt.ylabel('Mean_standardized_payroll')
plt.xlabel('Mean Winning Percentage')
plt.legend(title='Time Period')
plt.grid(True)

# x = np.array(range(len(mean_standardized_payroll)))
x = mean_winning_percentage
y = mean_standardized_payroll.values
fit = np.polyfit(x, y, deg=1)
plt.plot(x, fit[0] * x + fit[1], color='red')
plt.show()
data.head()
```

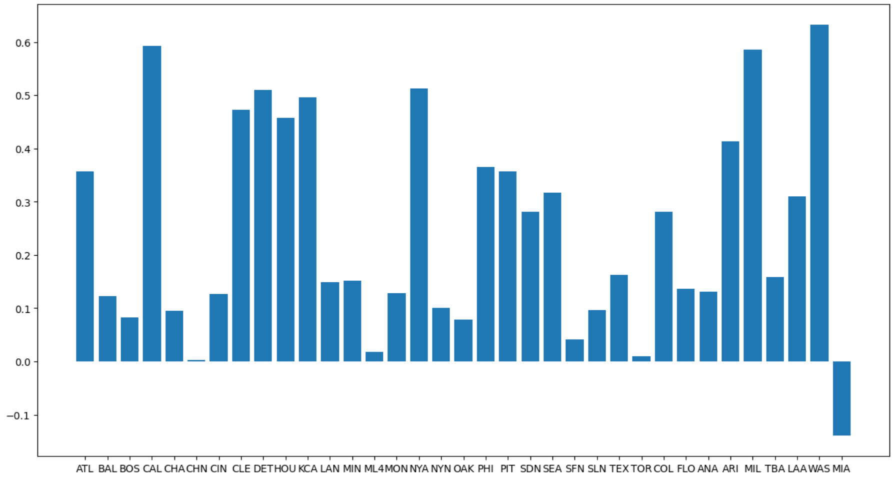
```
WARNING:matplotlib.legend:No artists with labels found to put in legend.  Note that arti
```



	teamID	yearID	franchID	W	G	total_payroll	winning_percentage	time_period	ex
0	ATL	1985	ATL	66	162	14807000.0	40.740741	1985-1991	
1	BAL	1985	BAL	83	161	11560712.0	51.552795	1985-1991	
2	BOS	1985	BOS	81	163	10897560.0	49.693252	1985-1991	
3	CAL	1985	ANA	90	162	14427894.0	55.555556	1985-1991	
4	CHA	1985	CHW	85	163	9846178.0	52.147239	1985-1991	

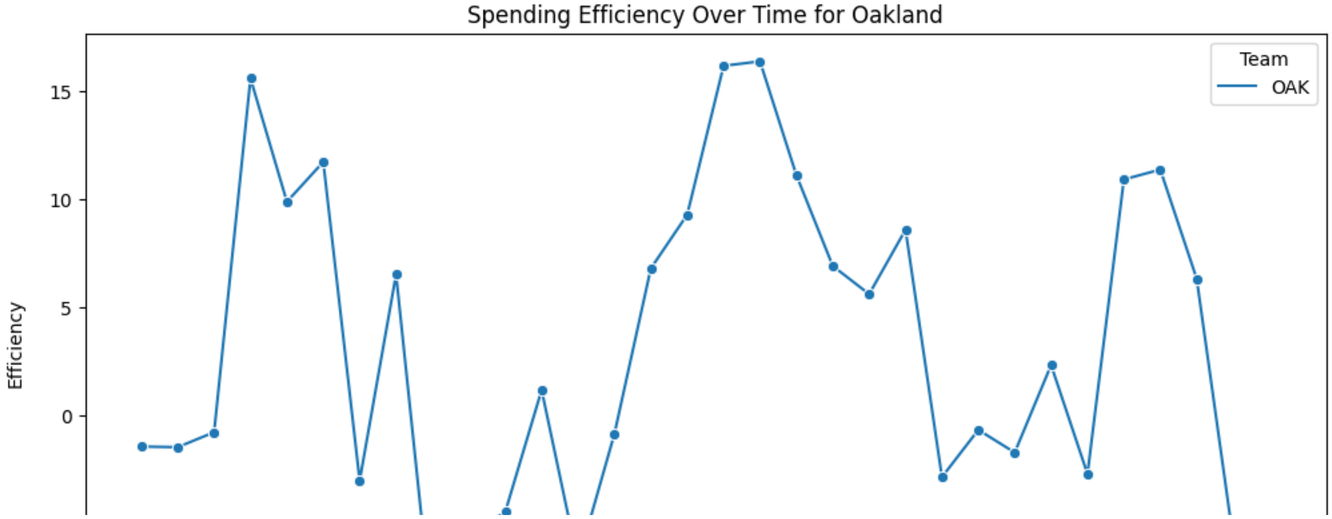
Question 3: Unlike in problem 4, the standardized payroll has decreased over the years but the win percentage has remained constant.

```
teams = data['teamID'].unique()
correlation = []
for team_id in teams:
  team_data = data[data['teamID'] == team_id]
  cor = team_data['winning_percentage'].corr(team_data['standardized_payroll'])
  correlation.append([team_id, cor])

corr = pd.DataFrame(correlation, columns = ['team_id', 'corr'])

plt.figure(figsize = (15, 8))
plt.bar(corr['team_id'], corr['corr'])
plt.show()
```



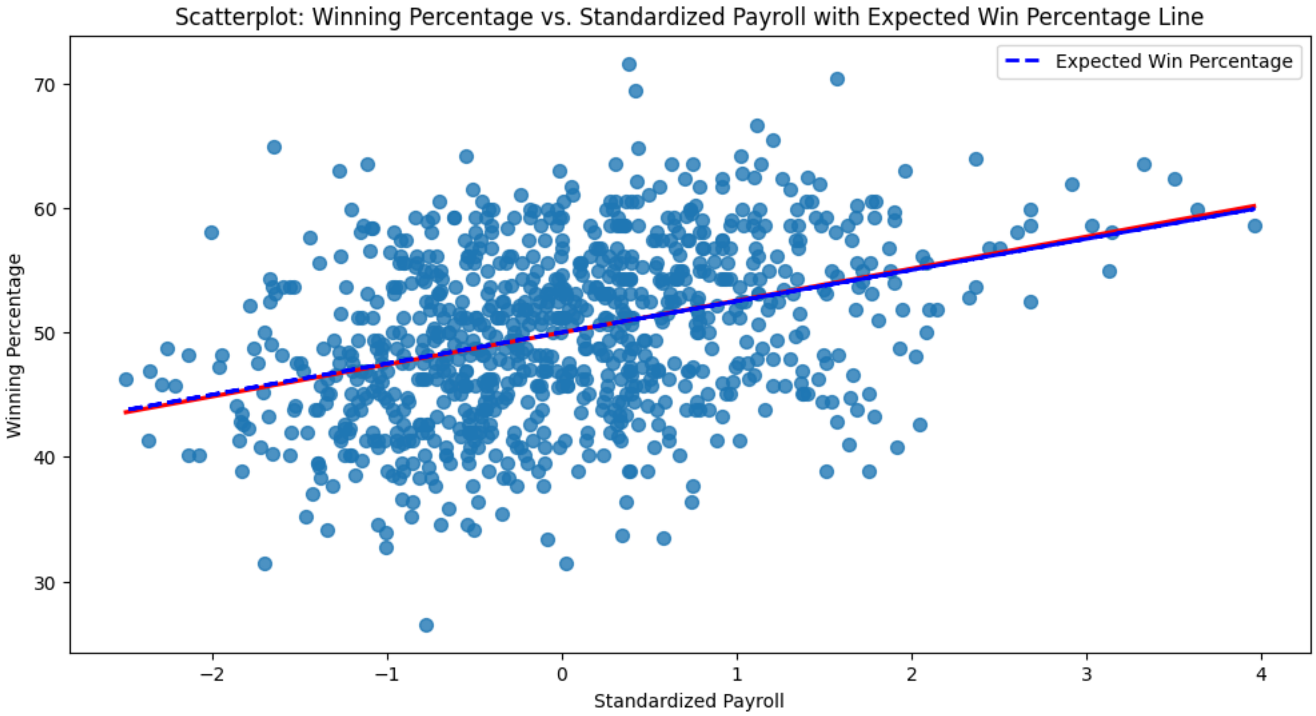
Apart from MIA most of the teams have a +ve correlation between money spent and number of wins secured. The teams that are close to the x axis had no effect on wins even when they spent more money.

```
data['expected_win_pct'] = 50 + 2.5 * data['standardized_payroll']
data['efficiency'] = data['winning_percentage'] - data['expected_win_pct']

selected_team = ['OAK']
selected_team_data = data[data['teamID'].isin(selected_team)]

plt.figure(figsize=(12, 6))
sns.lineplot(x='yearID', y='efficiency', hue='teamID', data=selected_team_data, marker='o')

plt.xlabel('Year')
plt.ylabel('Efficiency')
plt.title('Spending Efficiency Over Time for Oakland')
plt.legend(title='Team')
plt.show()
```



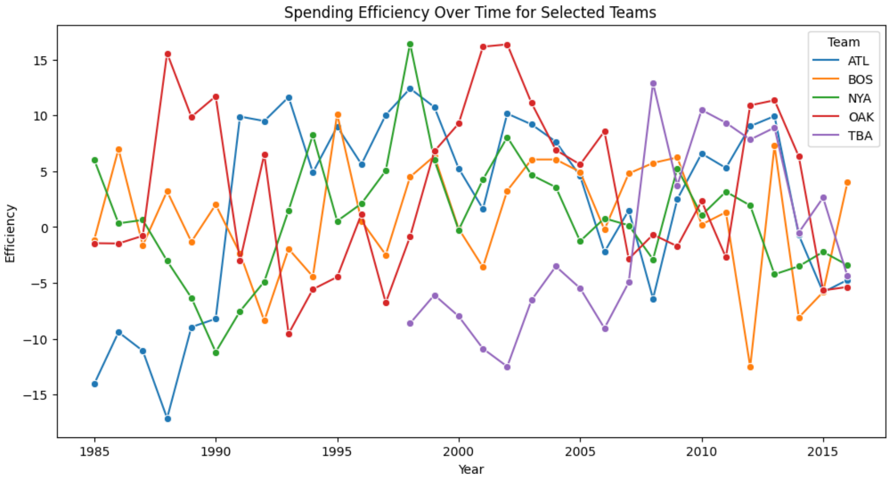
Problem 7:

```
plt.figure(figsize=(12, 6))
sns.regplot(x='standardized_payroll', y='winning_percentage', data=data, scatter_kws={'s': 50}, ci=None, line_kws={'color': 'red'})
expected_win_pct = 50 + 2.5 * data['standardized_payroll']

plt.plot(data['standardized_payroll'], expected_win_pct, color='blue', linestyle='--', linewidth=2, label='Expected Win Percentage')

plt.xlabel('Standardized Payroll')
plt.ylabel('Winning Percentage')
plt.title('Scatterplot: Winning Percentage vs. Standardized Payroll with Expected Win Percentage Line')
plt.legend()
plt.show()

```



Problem 8:

```
data['expected_win_pct'] = 50 + 2.5 * data['standardized_payroll']
data['efficiency'] = data['winning_percentage'] - data['expected_win_pct']

selected_teams = ['OAK', 'NYA', 'BOS', 'ATL', 'TBA']
selected_teams_data = data[data['teamID'].isin(selected_teams)]

plt.figure(figsize=(12, 6))
plt.figure(figsize=(12, 6))
sns.lineplot(x='yearID', y='efficiency', hue='teamID', data=selected_teams_data, marker='o')

plt.xlabel('Year')
plt.ylabel('Efficiency')
plt.title('Spending Efficiency Over Time for Selected Teams')
plt.legend(title='Team')
plt.show()

```


Spending Efficiency Over Time for Selected Teams
Team
ATL
BOS
NYA
OAK
TBA
Efficiency
Year