Project 3: Finding Relationships In Baseball

Brigham Eaquinto

Project Summary

This report analyses statistics from baseball. Basic findings such are batting averages are explored. Students who attended BYU-Idaho who now play in the major leagues will be shown too! The final question will show who has won the most World Series between the two socks teams.

Grand Question 1:

Write an SQL query to create a new dataframe about baseball players who attended BYU-Idaho. The new table should contain five columns: playerID, schoolID, salary, and the yearID/teamID associated with each salary. Order the table by salary (highest to lowest) and print out the table in your report.

	PlayerID	SchoolID	Salary	Year	TeamID
0	lindsma01	idbyuid	4000000	2014	CHA
1	lindsma01	idbyuid	3600000	2012	BAL
2	lindsma01	idbyuid	2800000	2011	COL
3	lindsma01	idbyuid	2300000	2013	CHA
4	lindsma01	idbyuid	1625000	2010	HOU
5	stephga01	idbyuid	1025000	2001	SLN
6	stephga01	idbyuid	900000	2002	SLN
7	stephga01	idbyuid	800000	2003	SLN
8	stephga01	idbyuid	550000	2000	SLN
9	lindsma01	idbyuid	410000	2009	FLO
10	lindsma01	idbyuid	395000	2008	FLO

	PlayerID	SchoolID	Salary	Year	TeamID
11	lindsma01	idbyuid	380000	2007	FLO
12	stephga01	idbyuid	215000	1999	SLN
13	stephga01	idbyuid	185000	1998	PHI
14	stephga01	idbyuid	150000	1997	PHI

Grand Question 2

This three-part question requires you to calculate batting average (number of hits divided by the number of at-bats):

a) Write an SQL query that provides playerID, yearID, and batting average for players with at least 1 at bat that year. Sort the table from highest batting average to lowest, and then by playerid alphabetically. Show the top 5 results in your report.

	playerID	yearID	batting_avg
0	aberal01	1957	1
1	abernte02	1960	1
2	abramge01	1923	1
3	acklefr01	1964	1
4	alanirj01	2019	1

b) Use the same query as above, but only include players with at least 10 at bats that year. Print the top 5 results.

	playerID	yearID	batting_avg
0	nymanny01	1974	0.642857
1	carsoma01	2013	0.636364
2	altizda01	1910	0.6
3	johnsde01	1975	0.6
4	silvech01	1948	0.571429

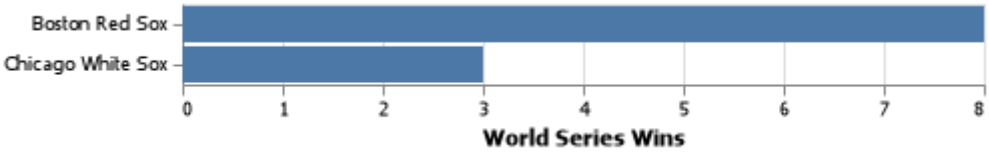
c) Now calculate the batting average for players over their entire careers (all years combined). Only include players with at least 100 at bats, and print the top 5 results.

	playerid	batting_avg
0	hazlebo01	0.402985
1	daviscu01	0.380952
2	fishesh01	0.374016
3	woltery01	0.369565
4	cobbty01	0.366299

Grand Question 3

Pick any two baseball teams and compare them using a metric of your choice (average salary, home runs, number of wins, etc). Write an SQL query to get the data you need, then make a graph in Altair to visualize the comparison.

Below we see that the Boston Red Sox team has won 5 more World Series's than the Chicago White Sox. Both are very prominant teams due to the famousness of their respective cities. It is common that the more money the team has, the better players they receive, and we can *loosely* infer that the Red Sox has more funding towards their baseball team.



Appendix A

```
#%%
import datadotworld as dw
import altair as alt

# %%

# GQ 1) Write an SQL query to create a new dataframe about baseball players who attended BYU-Ida
# using aliases in this query. s for salaries and cp for collegeplaying. They qualify the columr

gq1 = dw.query('byuidss/cse-250-baseball-database',
    """
    SELECT DISTINCT s.playerID AS PlayerID,
                    schoolID AS SchoolID,
                    s.salary AS Salary,
                    s.yearID AS Year,
                    s.teamid AS TeamID
    FROM salaries AS s
    JOIN collegeplaying AS cp
    ON s.playerID = cp.playerID
    WHERE cp.schoolid = "idbyuid" #filtering the schools to BYU-I
    ORDER BY s.salary DESC
    """)
print(gq1.dataframe.to_markdown())


# %%
# Grand Question 2: This three-part question requires you to calculate batting average (number c

# %%
# 2.1) Write an SQL query that provides playerID, yearID, and batting average for players with a

gq2_1 = dw.query('byuidss/cse-250-baseball-database',
    """
    SELECT  playerID
          , yearID
          , (h / ab) AS batting_avg
    FROM batting
    WHERE (h / ab) > 0.5
    ORDER BY batting_avg DESC, playerID
    LIMIT 5
    """)
print(gq2_1.dataframe.to_markdown())


#%%
# 2.2) Use the same query as above, but only include players with at least 10 at bats that year.

gq2_2 = dw.query('byuidss/cse-250-baseball-database',
    """
    SELECT  playerID
```

```
          , yearID
          , (h / ab) as batting_avg
    FROM batting
    WHERE ab >= 10
    ORDER BY batting_avg DESC, playerID
    LIMIT 5
    """)
print(gq2_2.dataframe.to_markdown())

# %%
# 2.3) Now calculate the batting average for players over their entire careers (all years combir

gq2_3 = dw.query('byuidss/cse-250-baseball-database',
    """
    SELECT playerid
         , (sum(h) / sum(ab)) as batting_avg
    FROM batting
    WHERE ab > 100
    GROUP BY playerid
    ORDER BY batting_avg DESC, playerid
    LIMIT 5
    """)
print(gq2_3.dataframe.to_markdown())

# %%
# Pick any two baseball teams and compare them using a metric of your choice (average salary, hc

gq3 = dw.query('byuidss/cse-250-baseball-database',
    """
    SELECT name AS team
         , COUNT_IF(wswin = "Y") AS world_series_wins
    FROM teams
    WHERE name = "Boston Red Sox" OR name = "Chicago White Sox"
    GROUP BY teamid
    """)
# gq3.dataframe
gq3_dat = gq3.dataframe

chart_3 = (
alt.Chart(gq3_dat)
.mark_bar()
.encode(x = alt.X('world_series_wins:Q', title = "World Series Wins"),
        y = alt.Y('team:O', title = ""))

)
chart_3.save("gq_chart3.png")
```