Sheet

Miral Mohamed, Nour Badr 900212713, Malak Rakhawy

Assignment 1

Data description:

The data set that we will analyze is called "Hitters", and was found on the website http://jse.amstat.org/jse_data_archive.htm. The data consists of 47 variables, and 6318 observations, and records baseball hitters who were drafted in the major leagues from 1992-2006. We used a subset of these 47 variables to assess them more clearly. We used 12 variables overall. The 5 variables: Total Games Played in Minor League, Total games played in double A and lower, Total Games Played in Triple-A and lower, Seasons Played in Minor League, and Total Plate Appearances in Minor League are continuous. The discrete variables consist of logical, categorical, and character variables. There are 2 logical variables: Throw and If They Appeared in Minor League, and they originally had outputs: Left/Right, and Yes/No, respectively. The 3 variables Position, Minor League Level Started at and Education Type are both categorical, with 6, 7 and 3 levels, respectively. Lastly, the 3 character variables were School, Organization, and Name. We will analyze the frequency of some variables as well as statistical summaries and correlations between them.

```
import pandas as pd
```

```
df=pd.read_csv("data.csv")
```

```
df
```

	Year when drafted	Round drafted	Overall pick	Name	Position	Organization	school	age when drafted	Birth date	Bats	...	Total plate appearances in Short-Season and lower levels	Total games played in Low-A and lower levels	Total plate appearances in Low-A and lower levels	Total games played in High-A and lower levels	Total plate appearances in High-A and lower levels	Tota gam play in Doul A an lowe level
0	1997	16	475	Alex Steele	Outfield	Tigers	SUNY-Cortland	22	12/09/75	Right	...	294	95	780	200	1161	200
1	1996	5	133	Philp Kendall	Catcher	Brewers	Jasper HS (IN)	19	08/22/77	Right	...	77	43	229	61	332	61
2	1999	43	1293	Nathan Rewers	Second Base	Reds	University of Richmond	23	11/30/76	Switch	...	16	3	16	3	16	3
3	2001	9	267	David Mattle	Outfield	Tigers	Kent State University	22	12/21/79	Left	...	159	176	716	279	1472	279
4	1997	1	15	Jason Dellaero	Shortstop	White Sox	University of South Florida	21	12/17/76	Switch	...	16	60	224	235	879	316
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
6312	1999	21	651	Michael Aldridge	Catcher	Yankees	Eastern Michigan University	22	03/17/77	Right	...	7	7	7	7	7	7
6313	2006	13	390	Mikal Garbarino	Outfield	Blue Jays	San Dimas HS, CA	18	04/07/88	Switch	...	1	1	1	1	1	1
6314	1996	31	912	Richard Clark	Outfield	Giants	Countryside HS (Clearwater,FL)	19	11/01/77	Right	...	6	6	6	6	6	6
6315	2001	32	956	Billy Jacobson	Outfield	Astros	Rice University	23	12/16/78	Right	...	25	26	25	26	25	26
6316	1998	40	1199	Michael Baetzel	Shortstop	White Sox	Kishwaukee College	19	10/01/79	Switch	...	13	15	13	15	13	15

6317 rows × 47 columns

```
print(df. columns)
```

```
Index(['Year when drafted', 'Round drafted', 'Overall pick', 'Name',
       'Position', 'Organization', 'school', 'age when drafted ', 'Birth date',
       'Bats', 'Throws ', 'education type', 'Drafted before',
       'Minor-League level started at',
       'if the player appeared in rookie leagues',
       'Number of games played in Rookie Leagues',
       'Number of plate appearances in Rookie Leagues',
       'If the player appeared in Short-Season-A',
       'Number of games played in Short-Season-A',
       'Number of plate appearances in Short-Season-A',
       'If the player appeared in Low-A ', 'Number of games played in Low-A ',
       'Number of plate appearances in Low-A',
       'If the player appeared in High-A', 'Number of games played in High-A',
       'Number of plate appearances in High-A',
       'If the player appeared in Double-A',
       'Number of games played in Double-A',
       'Number of plate appearances in Double-A ',
       'If the player appeared in Triple-A',
       'Number of games played in Triple-A\t',
       'Number of plate appearances in Triple-A ',
       'If the player appeared in the Major Leagues',
       'Number of seasons spent in the minor leagues    ',
       'Total number of games played in the minor leagues',
       'Total number of plate appearances in the minor leagues  ',
       'Total games played in Short-Season and lower levels',
       'Total plate appearances in Short-Season and lower levels',
       'Total games played in Low-A and lower levels',
       'Total plate appearances in Low-A and lower levels',
       'Total games played in High-A and lower levels',
       'Total plate appearances in High-A and lower levels',
       'Total games played in Double-A and lower levels\t',
       'Total plate appearances in Double-A and lower levels',
       'Total games played in Triple-A and lower levels\t',
       'Total plate appearances in Triple-A and lower levels',
       'Average plate appearances per game'],
      dtype='object')
```

Frequency graphs for different variables

Education Type:

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
```

```
df['education type'].value_counts()
```

```
4581/6317
```

```
0.7251860060155136
```

```
1697/6317
```

```
0.26864017729935097
```

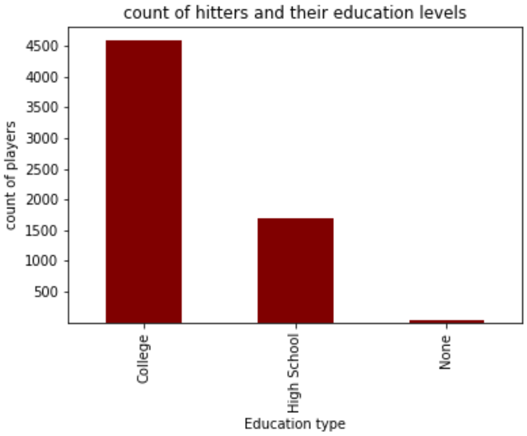
```
39/6317
```

```
0.006173816685135349
```

```
c= [500,1000,1500,2000,2500,3000,3500,4000,4500]
fig= df['education type'].value_counts().plot(kind='bar',color='maroon')
fig.set_ylabel('count of players')
fig.set_xlabel('Education type')
fig.set_yticks(c)

plt.title("count of hitters and their education levels")
```

```
Text(0.5, 1.0, 'count of hitters and their education levels')
```



Explanation for the frequency of the variable "Education type":

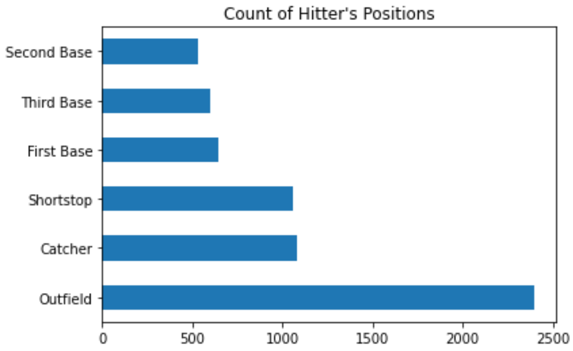
From the above codes, it can be clearly seen that as the education level increases, the frequency of the players in that level increases. For instance, around 72.5% of the players have a college education level, 27% have high school education, and only 0.5% have no education. These frequency results were predicted, because most players are drafted through their schools and most teams wait for players to join college before they get drafted.

Position:

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
```

```
df['Position'].value_counts().plot(kind='barh')
plt.title("Count of Hitter's Positions")
```

```
Text(0.5, 1.0, "Count of Hitter's Positions")
```



Explanation of the frequency of the variable "Position":

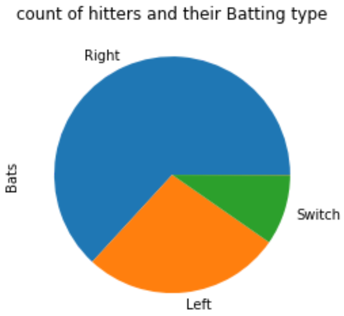
Outfielders have the highest count, which can indicate that they are the ones who are most likely to be drafted for the major leagues. The outfield position also makes up more of the team, which could by why the count is that high. The least common positions are first, second, and third base, which is because they each have one player each.

Throws:

```
import numpy as np
import matplotlib.pyplot as plt
```

```
df['Bats'].value_counts().plot(kind='pie')
plt.title("count of hitters and their Batting type")
```

```
Text(0.5, 1.0, 'count of hitters and their Batting type')
```



This makes sense since most people in the world are right-handed which means that the percentage of people hitting with their right hand is higher than those from their left hand. in addition, some people are both right-handed and left-handed or have the gift to switch between both hands. this explains the fact that there are players who switch between both in their hitting style.

Statistical factors for different variables

Total number of games played in the minor leagues:

```
x=df['Total number of games played in the minor leagues']
```

```
x.mean()
```

```
258.96089916099413
```

```
x.max()-x.min() #range
```

```
1448
```

```
q= np.array([0,0.25,0.5,0.75,1])
q
np.quantile(x,q)  #Min/25%/50%/75%/Max
```

```
x.std() #Standard deviation
```

```
222.6104799830968
```

```
x.var() #Variance
```

```
49555.42579830474
```

```
x.mode()
```

total plate appearances in Triple-A and lower levels

```
y=df['Total plate appearances in Triple-A and lower levels']
```

```
y.mean()
```

```
998.5554851986702
```

```
y.max()-y.min() #range
```

```
5516
```

```
quantile = np.array([0,0.25,0.5,0.75,1.0])
quantile
np.quantile(y,quantile) #Min/25%/50%/75%/Max
```

```
y.std() #standard deviation
```

```
889.7460786740799
```

```
y.var() #variance
```

```
791648.084515902
```

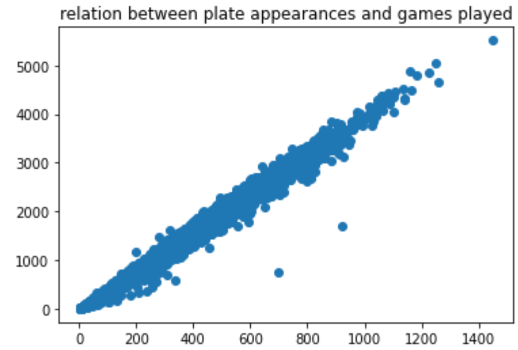
```
y.mode()
```

GRAPHS AND RELATIONSHIPS BETWEEN VARIABLES:

relation between plate appearances and games played:

```
import matplotlib.pyplot as plt
import numpy as np
```

```
y=df['Total plate appearances in Triple-A and lower levels']
x=df['Total number of games played in the minor leagues']
plt.scatter(x,y)
plt.title("relation between plate appearances and games played")
plt.show()
```



```
r= np.corrcoef(x,y)
r
```

As seen in the graph above, there is a strong positive correlation between the total games played and total plate appearances. This is explained since each time you play a game you get more practice. therefore, this makes sense that the players get better when they play more games. This means that their plate appearances get better which explain the strong corrolation between both variables.

Education level vs Total games played in triple A games:

As seen in the data, there are different levels of games in the minor league. The highest and most professional level is the Triple-A level. Therefore, using this variable as an indication of whether the players are successful or not. However, the data does not show the number of games played in this level for each player. Instead, the cumulative number of games played in each level (added to the levels below) is shown.

We must first create the variable "Total triple A games played", by subtracting "Total games played in Triple-A and lower levels\t" and "Total games played in Double-A and lower levels\t"

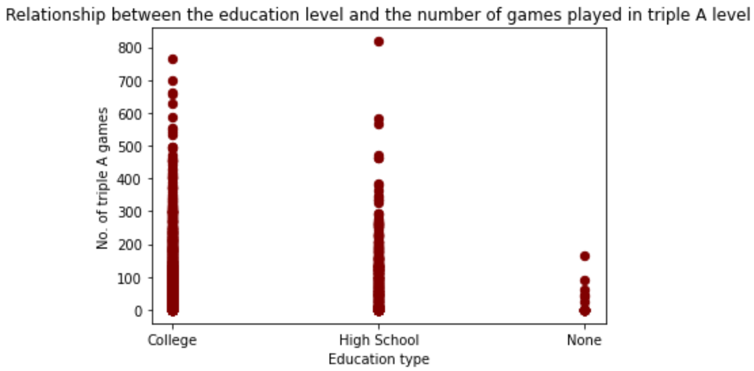
```
df2= df.assign(TotaltripleAgamesplayed = df['Total games played in Triple-A and lower levels\t']-df['Total games played in Double-A an
df2 #we have now created a dataset with a new column
```

	Year when drafted	Round drafted	Overall pick	Name	Position	Organization	school	age when drafted	Birth date	Bats	...	Total games played in Low-A and lower levels	Total plate appearances in Low-A and lower levels	Total games played in High-A and lower levels	Total plate appearances in High-A and lower levels	Total games played in Double-A and lower levels\t	Total plat appearan in Double and lowe levels
0	1997	16	475	Alex Steele	Outfield	Tigers	SUNY-Cortland	22	12/09/75	Right	...	95	780	200	1161	200	1161
1	1996	5	133	Philp Kendall	Catcher	Brewers	Jasper HS (IN)	19	08/22/77	Right	...	43	229	61	332	61	332
2	1999	43	1293	Nathan Rewers	Second Base	Reds	University of Richmond	23	11/30/76	Switch	...	3	16	3	16	3	16
3	2001	9	267	David Mattle	Outfield	Tigers	Kent State University	22	12/21/79	Left	...	176	716	279	1472	279	1472
4	1997	1	15	Jason Dellaero	Shortstop	White Sox	University of South Florida	21	12/17/76	Switch	...	60	224	235	879	316	1623
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
6312	1999	21	651	Michael Aldridge	Catcher	Yankees	Eastern Michigan University	22	03/17/77	Right	...	7	7	7	7	7	7
6313	2006	13	390	Mikal Garbarino	Outfield	Blue Jays	San Dimas HS, CA	18	04/07/88	Switch	...	1	1	1	1	1	1
6314	1996	31	912	Richard Clark	Outfield	Giants	Countryside HS (Clearwater,FL)	19	11/01/77	Right	...	6	6	6	6	6	6
6315	2001	32	956	Billy Jacobson	Outfield	Astros	Rice University	23	12/16/78	Right	...	26	25	26	25	26	25
6316	1998	40	1199	Michael Baetzel	Shortstop	White Sox	Kishwaukee College	19	10/01/79	Switch	...	15	13	15	13	15	13

6317 rows × 48 columns

```
ed= df2['education type']
ta= df2['TotaltripleAgamesplayed']
plt.scatter(ed, ta, color ='maroon',s=40)

plt.xlabel("Education type")
plt.ylabel("No. of triple A games")
plt.title("Relationship between the education level and the number of games played in triple A level")
plt.show()
```



firstly, instead of creating a bar chart, a point/scatter plot was created. This is to have a better representation of the data, as each player is represented by one point in the above chart. In a bar chart, only the player with the maximum number of games played would be shown.

It can be seen that the number of games played by players increases when their education level increases. As stated before, the data consists of mostly college players, some high school players, and a negligible number of players with no education. This means that because college players are more abundant, then there is a higher chance that a player from this category would be more successful and play more games in the triple A levels.

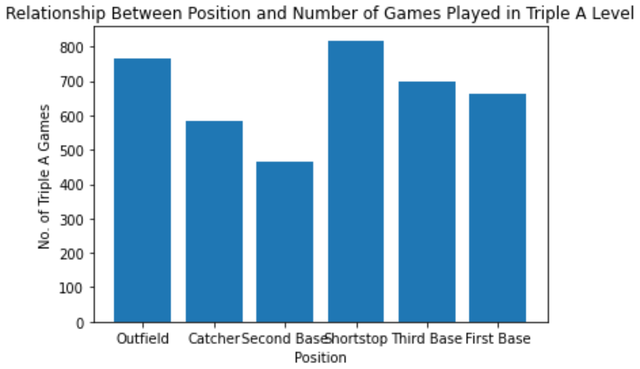
An outlier is observed in the high school education type, which if not noticed would lead to a misinterpretation (That high school players play more games in the triple A level than college players)

we can see that in each category the dots are more condensed at the bottom. This means that it is more common for players to play a smaller number of games in the triple A levels.

Total Triple A Games Played vs Position

```
position= df2['Position']
ta= df2['TotaltripleAgamesplayed']
plt.bar(position, ta)

plt.xlabel("Position")
plt.ylabel("No. of Triple A Games")
plt.title("Relationship Between Position and Number of Games Played in Triple A Level")
plt.show()
```



Here we can see that the most common position in triple A games was shortstop followed by outfield. Triple A games are very competitive and are a measure of the success of a players career. The shortstop position is also considered the hardest position in the team which explains why they make up most of the players in the triple A league.