HW 2 Manually Graded: Upload PDF here

● Graded

Student

Rey Stone

Total Points

12.5 / 15 pts

Question 1

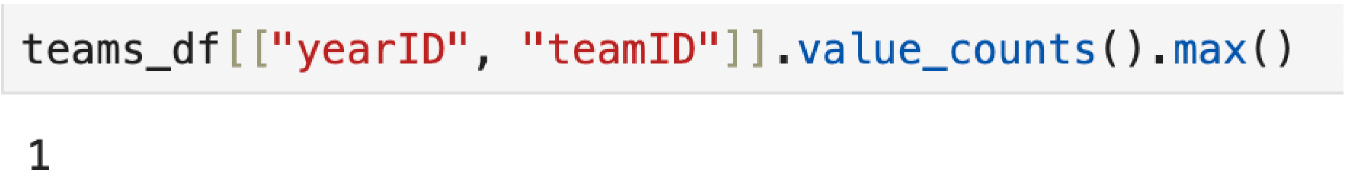
Question 2b

2 / 3 pts

Question 2b

2bi). Represents the number of Games that the Team played in that specific year

2bii). Granularity of the data is a specific team in a specific year.
To prove this notice that the fewest columns that uniquely identify each row in Teams is teamID and yearID:

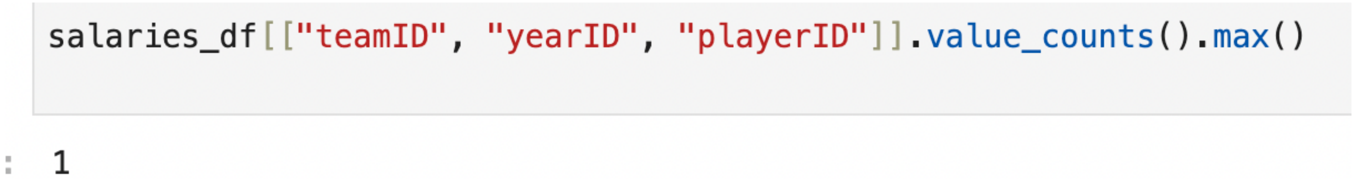


```
teams_df[["yearID", "teamID"]].value_counts().max()
```

```
1
```

Note that while franchID and yearID also uniquely identify each row, when you read the documentation you see that franchID is a primary key in another table, hence why it is included in this table. The data is not aggregated at the franchise level (franchises consists of the MLB team as well as associated minor league teams, and the minor league teams are not aggregated into this data).

2biii). Granularity of the data is a specific player on a specific team in a specific year. To prove this notice that the fewest columns that uniquely identify each row in Salaries is playerID, teamID and yearID:



```
salaries_df[["teamID", "yearID", "playerID"]].value_counts().max()
```

```
:  1
```

✔ **– 1 pt** 2b(iii) incorrect / Missing

Question 2

Question 3a

1.5 / 2 pts

3a (2 pts) !

$$\sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x}$$

$$= \sum_{i=1}^{n} x_i - n \cdot \bar{x}$$

$$= \sum_{i=1}^{n} x_i - n \cdot \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$

$$= \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i = 0$$

✓ **– 0.5 pts** Incorrect or missing substitution

1

Question 3

Question 3b

3 / 4 pts

3b

To find local max/min we start by finding the value(s) of c such that $f'(c) = 0$:

$f(c) = \frac{1}{n}\sum_{i=1}^{n}(x_i - c)^2$

$f'(c) = \frac{d}{dc}\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - c)^2\right)$

$= \frac{1}{n}\sum_{i=1}^{n}\frac{d}{dc}(x_i - c)^2$

$= \frac{1}{n}\sum_{i=1}^{n}2(x_i - c)(-1)$

$= \frac{-2}{n}\sum_{i=1}^{n}(x_i - c)$

Setting this equal to 0 and solving for c:

$f'(c) = 0$

$\implies -\frac{2}{n}\sum_{i=1}^{n}(x_i - c) = 0$

$\implies \sum_{i=1}^{n}(x_i - c) = 0$

$\implies \sum_{i=1}^{n}x_i - \sum_{i=1}^{n}c = 0$

$\implies \sum_{i=1}^{n}x_i - nc = 0$

$\implies \sum_{i=1}^{n}x_i = nc$

$$\implies \boxed{c = \frac{1}{n}\sum_{i=1}^{n}x_i = \bar{x}}$$

To test if this is a max or min, we use the 2nd derivative test:

$f''(c) = \frac{-2}{n}\sum_{i=1}^{n}\frac{d}{dc}(x_i - c)$

$= \frac{-2}{n}\sum_{i=1}^{n}(-1)$

$= \frac{-2}{n}(-1)(n) = 2$

Thus $f''(\bar{x}) = 2 > 0 \implies c = \bar{x}$ is a minimum value

✔ **− 1 pt** Incorrectly solved or did not demonstrate value is a min by calculating $f''(c)$ greater than 0

Question 4

Question 4b

2 / 2 pts

SOLN: `likelihood = (p**4)*(1-p)**6`

(Notice this doesn't have $\binom{10}{4}$ out front because we have been given one specific ordering (we aren't considering all possible orderings with 4 heads))

✔ **– 0 pts** Correct

Question 5

Question 4d

4 / 4 pts

Solution:

$$\log(L(p)) = \log(p^4(1-p)^6) = 4\log(p) + 6\log(1-p).$$

$$\implies \frac{d}{dp}\left(\log(L(p)\right) = \frac{4}{p} - \frac{6}{1-p}$$

Now we solve for where the derivative equals 0:

$$\frac{d}{dp}\left(\log(L(p)\right) = 0$$

$$\implies \frac{4}{p} - \frac{6}{1-p} = 0$$

$$\implies \frac{4}{p} = \frac{6}{1-p}$$

$$\implies 6p = 4(1-p)$$

$$\implies 6p = 4 - 4p$$

$$\implies 10p = 4$$

$$\implies \boxed{\hat{p} = 0.4}$$

✔ **– 0 pts** Correct or follows from likelihood function above.

Question assigned to the following page: 1

0.1 Question 2b (5 pts)

Examine the structure, granularity and faithfulness of the datasets. (Hint: The common utility functions we covered in class will be useful here).

Then answer the following questions:

- i). What does the column `G` represent in the teams dataset? (For a description of the columns, see the documentation in the `data` folder).
- ii). What is the granularity of the `teams.csv` file?
- iii). What is the granularity of the `salary.csv` file?
- iv). How many rows and columns are in the teams dataset? Assign your answer to the variables `team_rows` and `team_col` below.
- v). How many rows and columns are in the salary dataset? Assign your answer to the variables `salary_rows` and `salary_col` below.
- vi). How many entries in the `teams.csv` file are missing Attendance Data? Assign your answer to the variable `missing_attendance` below.

0.1.1 Answer Cell for Questions 2b(i)(ii)(iii)

In this cell, answer questions 2b(i) - (iii) using Markdown (not code).

2b(i) Answer: Games played

2b(ii) Answer: General stats for all MLB teams by year. Columns of data include games played, hits, outs, etc.

2b(iii) Answer: Salary info by player, team, league, and year.

In the code cells below justify your answers to part (ii) and (iii) and then answer parts iv through vi

```
In [563]: team_gran = teams_df.value_counts()

          team_gran

          # Show work in this cell justifying your answer to part 4a(ii) (hint: either use .value_count
```

```
Out[563]: yearID  lgID  teamID  franchID  divID  Rank  G    Ghome  W   L    DivWin  WCWin  LgWin  WSWin
          1995    AL    BAL     BAL       E      3     144  72.0   71  73   N       N      N      N
```

Question assigned to the following page: 1

```
2016    AL    BOS    BOS       E      1      162  81.0    93   69   Y        N       N       N
2013    NL    ARI    ARI       W      2      162  81.0    81   81   N        N       N       N
              ATL    ATL       E      1      162  81.0    96   66   Y        N       N       N
              CHN    CHC       C      5      162  81.0    66   96   N        N       N       N

2004    NL    CIN    CIN       C      4      162  81.0    76   86   N        N       N       N
              COL    COL       W      4      162  81.0    68   94   N        N       N       N
              FLO    FLA       E      3      162  80.0    83   79   N        N       N       N
              HOU    HOU       C      2      162  81.0    92   70   N        Y       N       N
2022    NL    WAS    WSN       E      5      162  81.0    55   107  N        N       N       N
Name: count, Length: 834, dtype: int64
```

```
In [564]: sal_gran = salaries_df.value_counts()

          sal_gran

          # Show work in this cell justifying your answer to part 4a(iii) (hint: either use .value_coun
```

```
Out[564]: yearID  teamID  lgID  playerID   salary
          1985    ATL     NL    barkele01  870000      1
          2006    HOU     NL    quallch01  376000      1
                  KCA     AL    brownem01  1775000     1
                                berroan01  2000000     1
                                bautide01  335500      1
                                                      ..
          1996    KCA     AL    lockhke01  207500      1
                                lennopa01  120000      1
                                jacomja01  150000      1
                                huismri01  118000      1
          2016    WAS     NL    zimmery01  14000000    1
          Name: count, Length: 26428, dtype: int64
```

```
In [565]: # Solution Cell for 2a(iv) and 2a(v)
          #Use code to find the number of rows and columns in the teams data.  Do not enter any values

          team_rows = teams_df.shape[0]

          team_col = teams_df.shape[1]

          salary_rows = salaries_df.shape[0]

          salary_col = salaries_df.shape[1]
```

```
In [566]: # Solution Cell for 2a(vi)
          # Use code to find the number of rows in the Teams data that are missing attendance data

          missing_attendance = sum(teams_df["attendance"].isna())

          missing_attendance
```

```
Out[566]: 279
```

```
In [567]: grader.check("q2b")
```

```
Out[567]: q2b results: All test cases passed!
```

Question assigned to the following page: 2

0.2 Question 3a (2 pts)

We commonly use sigma notation to compactly write the definition of the arithmetic mean (commonly known as the average):

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + ... + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

The ith *deviation from average* is the difference $x_i - \bar{x}$. Prove that the sum of all these deviations is 0 that is, prove that $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$ (write your full solution in the box directly below showing all steps and using LaTeX).

$\sum_{i=1}^{n} x_i - \bar{x}$

By using the summation rules defined above, we can split this summation into to two

$\implies \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x}$

because the summation of $\bar{x}$ is $\bar{x} + \bar{x} + ... + \bar{x}$ to n, we can simplify the second summation

$\implies \sum_{i=1}^{n} x_i - n\bar{x}$

we are given that $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ we can substitute this for the variable $\bar{x}$

$\implies \sum_{i=1}^{n} x_i - n\sum_{i=1}^{n} \frac{1}{n}x_i$

Since $n \cdot \frac{1}{n}$ cancells out we are left with

$\implies \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i = 0$

Question assigned to the following page: 3

0.3 Question 3b (4 pts)

Let $x_1, x_2, \ldots, x_n$ be a list of numbers. You can think of each index i as the label of a household, and the entry x_i as the annual income of Household i.

Consider the function

$$f(c) = \frac{1}{n}\sum_{i=1}^{n}(x_i - c)^2$$

In this scenario, suppose that our data points $x_1, x_2, \ldots, x_n$ are fixed and that c is the only variable.

Using calculus, determine the value of c that minimizes $f(c)$. You must use calculus to justify that this is indeed a minimum, and not a maximum.

$f'(c) = \frac{1}{n}\sum_{i=1}^{n}(x_i - c)^2$

$= \frac{1}{n}\sum_{i=1}^{n} -2(x_i - c)$ by use of the chain rule

$\frac{1}{n}\sum_{i=1}^{n} -2(x_i - c) = 0$ set this equal to zero

$\frac{1}{n}\sum_{i=1}^{n} 2(x_i - c) = 0$ multiply both sides by (-1)

$\frac{1}{n}(2\sum_{i=1}^{n} x_i - 2\sum_{i=1}^{n} c) = 0$ by using the properties of summations.

$\frac{1}{n}(2\sum_{i=1}^{n} x_i - 2c\sum_{i=1}^{n} 1) = 0$ since c is a constant, we can take it out of the summation.

$\frac{2}{n}(\sum_{i=1}^{n} x_i - cn) = 0$ the summation of 1 is $n.$ as well as factoring out the 2.

$\frac{2}{n}\sum_{i=1}^{n} x_i - \frac{2}{n}cn = 0$ expanding the fraction.

$\frac{2\sum_{i=1}^{n} x_i}{n} = 2c$

$\frac{\sum_{i=1}^{n} x_i}{n} = c$ divide by two.

This value is just $c = \bar{x}$ proving that $\bar{x}$ minimizes the function $f(c)$.

Question assigned to the following page: 4

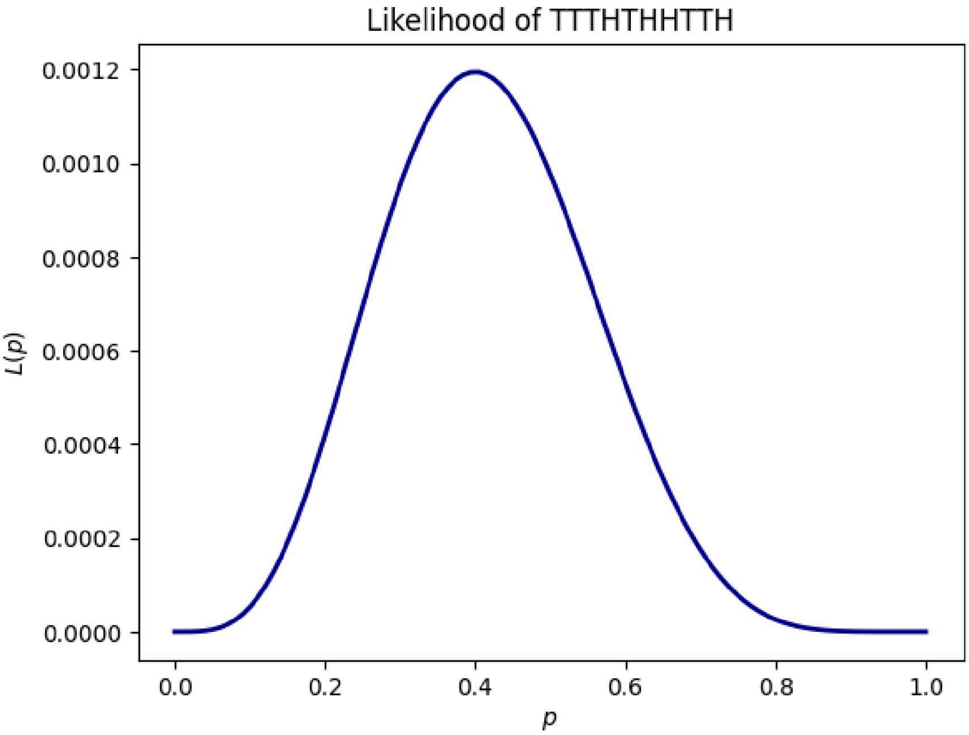
What is $L(p)$ (i.e. the likelihood) for the sequence TTTHTHHTTH?

Enter your answer below by setting the `likelihood` variable equal to the correct function.

(For example `likelihood = sin(p)+2p`, althought that is definitely an incorrect answer!)

Then run the code below to plot the likelihood function.

```
In [597]: #At the top of the notebook we already imported a useful plotting module, matplotlib with ali

          p = np.linspace(0, 1, 100)
          #This creates an array of 100 p-values equally spaced between 0 and 1

          likelihood = pow(1-p, 6) * pow(p, 4)
          #Define the likelihood function above

          plt.plot(p, likelihood, lw=2, color='darkblue')
          #This plots the likelihood function

          plt.xlabel('$p$')
          #This labels the x axis

          plt.ylabel('$L(p)$')
          #This labels the y-axis

          plt.title('Likelihood of TTTHTHHTTH');
          #This titles the plot
```


Likelihood of TTTHTHHTTH
0.0012
0.0010
0.0008
0.0006
0.0004
0.0002
0.0000
$L(p)$
0.0
0.2
0.4
0.6
0.8
1.0
p

Question assigned to the following page: 5

0.4 Question 4d (4 pts)

Notice the value you found graphically for $\hat{p}$ above also intuitively makes sense because it is also the observed proportion of heads in the given sequence TTTHTHHTTH.

Let's prove what you observed graphically above. That is, let's use calculus to find $\hat{p}$.

But **wait before you start trying to find the value p where $L'(p) = 0$ (trust us, the algebra is not pretty...)**

USEFUL TIP: The value $\hat{p}$ at which the function $L(p)$ attains its maximum is the same as the value at which the function $\ln(L(p))$ attains its maximum.

This tip is hugely important in data science because many probabilities are products and the natural log function `ln` function turns products into sums. It's **much simpler to take derivatives of a sum** than a product.

Thus, to find the value p where $L'(p) = 0$: - Take the natural log `ln` of L(p) - Use properties of logs to rewrite products in `ln(L(p))` as sums - Take the derivative of this rewritten version of `ln(L(p))` - Solve $\frac{d}{dp}\left[\ln(L(p))\right] = 0$ for p - You should get the same answer that you found graphically above.

You don't have to check that the value you've found produces a max and not a min – we'll spare you that step.

Show all steps in the cell below using Markdown and LaTeX

$\ln\left((1-p)^6 \cdot p^4\right)$

$\implies \ln\left(1-p\right)^6 + \ln(p)^4$

$\implies 6\ln(1-p) + 4\ln(p)$

$\implies \frac{d}{dp}6\ln(1-p) + 4\ln(p) = 4\frac{1}{p} - 6\frac{1}{1-p}$

$\implies 4\frac{1}{p} - 6\frac{1}{1-p} = 0$

$\implies \frac{4}{p} = \frac{6}{1-p}$

$\implies 4(1-p) = 6p$

$\implies 4 - 4p = 6p$

Question assigned to the following page: 5

$\implies 4 = 10p$

$\implies \frac{4}{10} = p$