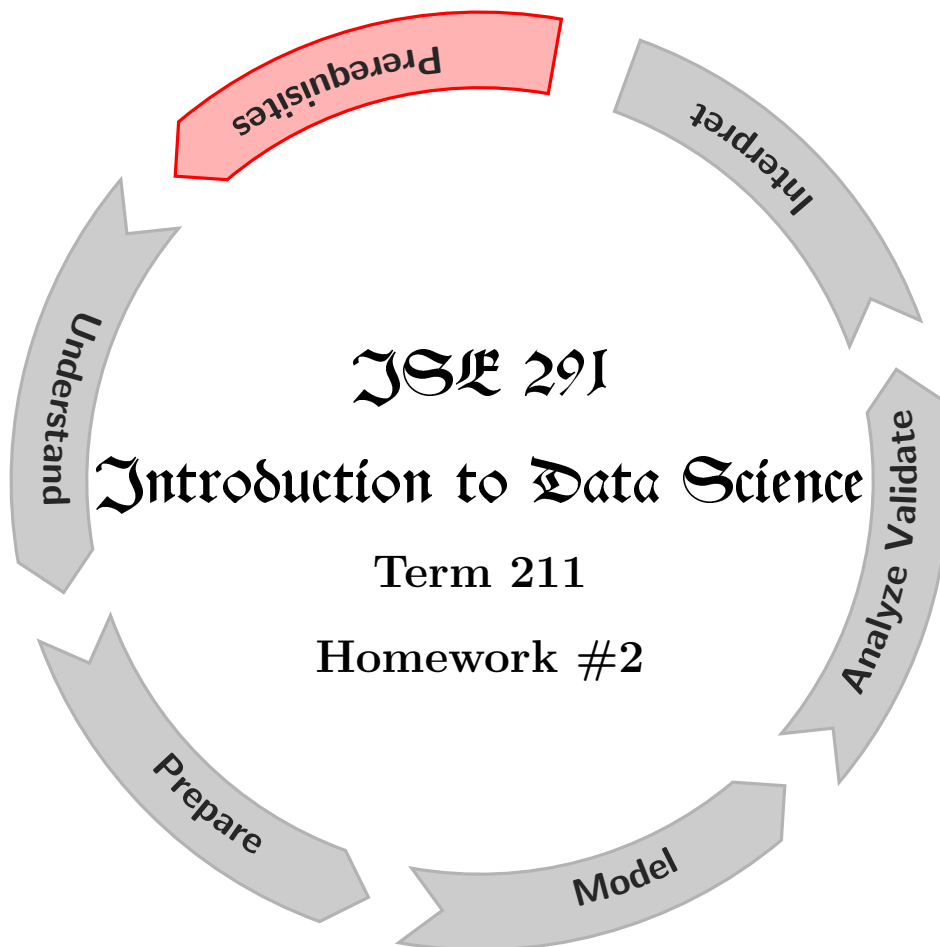


[The HW must be submitted as one .pdf file for Problem-A, and one .ipynb file for Problem-B. Write names & IDs of all the group members on both files.]



Homework Guidelines

To receive full credit, you should make sure you adhere to the following guidelines. You should submit the solutions for the FIRST TWO problems only. ONE HW per group should be submitted. For any questions/comments contact your section instructor.

Homework Presentation: For Problem-A

- The problem and part number should be clearly written.
- All solutions of your homework should be in CHRONOLOGICAL order.
- Your NAMES, IDs, and the homework number should be clearly indicated in the FIRST page of the pdf.

Homework Presentation: For Problem-B

- Every sub-problem (part) should be answered on a DIFFERENT CELL.
- EVERY CELL should have problem and part number clearly written in the first line.
- All cells of your homework should be in CHRONOLOGICAL order. One cell per sub-problem.
- Your NAMES, IDs, and the homework number should be clearly indicated in the FIRST CELL of the notebook.

Problem # A**50 marks**

Consider data given in Tables 1 & 2 obtained from GameStation company located at Burj AlAswasd in AlKhobar.

Table 1: Popular Games Data

S. No	Game	Platform	Price	Rating	Multiplayer	Sold
1	MineKraft	PS4	250	E	1	232
2	Hassasin Creed	Xbox	320	M	1	211
3	Cities & Skyes	PC	200	E	0	115
4	Age of Emirates	PC	289	E	0	540
5	Grave Raider	PS4	170	T	1	279
6	FourthKnight	Xbox	220	T	0	345
7	Plastic Gear Solid	PS4	289	M	1	198
8	Pocket League	PS4	350	E	1	571
9	Project Bikes	Xbox	197	E	1	309
10	Hall of Duty	PS4	266	M	1	295

Table 2: Data Description

S. No	Serial number (unique number) assigned to the games.
Game	Name of the game.
Platform	The console that is to be used for the game.
Price	The selling price for one game in SAR.
Rating	The ESRB rating, E = everyone, T= teen, M = Mautre. The ratings can be sorted as $E < T < M$.
Multiplayer	Yes = 1, No = 0.
Sold	Total copies sold last month. The data provider believes that the name of the game, platform, price, rating, and number of players can be used to predict the copies that can be sold in the future months.

Answer the following questions using data given in Tables 1 & 2:

A-1: List all the variables shown in Table 1. What is the use of Table 2's data.

A-2: Classify the variables in Table 1 into field types.

A-3: Classify the variables in Table 1 into the following terminology: Independent and dependent variables.

A-4: Draw the histogram for **Rating**.

A-5: Draw the pie chart for **Platform**.


A-6: From **Rating's** histogram, can you conclude that the distribution is symmetrical?

A-7: Find the mean, median and mode for the **Price**.

A-8: Find the variance and standard deviation for the **Price**.

A-9: Find the upper and lower quartile for **Price**.

A-10: Draw the box-plot for **Price**.

 *Note: Solve all the above questions by hand. You can use calculator and mathematical set (or geometry box). Do NOT use laptop/computer/mobile. To get the quartiles (Q1, Q2, Q3) do the following: First, sort the data. Now, Q1 is the median of the first half of the data, Q2 is the median of the full data, and Q3 is the median of the second half of the data.*

Problem #B

50 marks

Consider the following data related to the final scores of the four sections of a data science course:

Section	Scores
Clytherin:	95, 95, 65, 70, 65, 65, 75, 75, 80, 65, 95, 75, 65, 75, 80, 65, 95, 75, 65, 95, 75, 65, 80, 65, 70, 70, 65
Jriffinodor:	81, 65, 73, 77, 70, 71, 74, 63, 85, 79, 86, 68, 75, 66, 62, 88, 83, 87, 64, 72, 69, 84, 78, 80, 76, 82, 67
Hufflebuff:	84, 77, 77, 61, 77, 84, 75, 89, 66, 75, 61, 66, 80, 61, 70, 66, 73, 89, 73, 70, 73, 70, 80, 84, 89, 80, 75
Ravenklaw:	75, 75

Answer the following questions:

B-1: Draw the histograms of scores for each of the above sections. Take bin size as 10 units, starting from score 40. Draw each histogram in one figure. Write the title on each histogram.

B-2: Find the mean, median and mode for each of the above sections.

B-3: Find the population variance and population standard deviation for each of the above sections.

B-4: Find the upper and lower quartile for each of the above sections.

B-5: Draw the box-plot for each of the above sections in one figure. Label x-axis with the name of the section.

B-6: Do hypothesis testing for each section to check if the data follows normal distribution. Assume p-value less than 0.05 or 5% as very small.

B-7: Do pair-wise student's t-test to compare the means of two distribution. Set kwarg `alternative='two-sided'`. Assume p-value less than 0.05 or 5% as very small.

B-8: Do pair-wise Mann-Whitney U test to compare the underlying distributions. Set kwarg `alternative='two-sided'`. Assume p-value less than 0.05 or 5% as very small.

📌 *Note: Solve all the above questions using Python (not by hand). To read the data in python and repeat for each section, you can use the following code:*

```
In [1]: 1 Data = {
2     "Jriffindor": [81, 65, 73, 77, 70, 71, 74, 63, 85, 79, 86, 68, 75, 66, 62,
3         88, 83, 87, 64, 72, 69, 84, 78, 80, 76, 82, 67],
4     "Clytherin": [95, 95, 65, 70, 65, 65, 75, 75, 75, 80, 65, 95, 75, 65, 75, 80, 65, 95, 75, 65,
5         95, 75, 65, 80, 65, 70, 70, 65],
6     "Hufflebuff": [84, 77, 77, 61, 77, 84, 75, 89, 66, 75, 61, 66, 80, 61, 70,
7         66, 73, 89, 73, 70, 73, 70, 80, 84, 89, 80, 75],
8     "Ravenklaw": [75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75, 75,
9         75, 75, 75, 75, 75, 75]
10 }
11
12 for section in Data: #dictionary's default iterator is key
13     series=Data[section].copy() # list is mutable, so we copy to avoid chages in actual data
14     # your code for each section
15     #
16     #
17     #
18     # end of the code for each section
```

Problem #C (Practice only. No submission required.)

Consider the following python methods, available in naive python, or numpy library:

C-1: `plt.figure()`
C-2: `plt.show()`
C-3: `plt.hist()`
C-4: `plt.pie()`
C-5: `np.mean()`
C-6: `np.median()`
C-7: `scipy.stats.mode()`
C-8: `np.std()`
C-9: `np.var()`
C-10: `np.percentile()`
C-11: `plt.boxplot()`
C-12: `plt.title()`
C-13: `scipy.stats.ttest_1samp()`
C-14: `scipy.stats.ttest_ind()`
C-15: `ax = fig.add_subplot()`
C-16: `ax.set_xticklabels()`

Answer the following questions for each of the above methods:

- State the purpose/usage of the method.
- List all the argument of the method.
- Classify the arguments as positional or keyword arguments.
- Write the default values for each of the keyword arguments.

☞ Note: You must use ***help()*** function from python to answer all the above questions.

Problem #D (Practice only. No submission required.)

Consider the economics data given in Tables 3 & 4 obtained from <http://research.stlouisfed.org/fred2>. The data contains 20 rows and 6 variables, as described below:

Table 3: Economics Data

DATE	PSR	PCE	MWU	TP	UL
7/1/1967	12.6	506.7	4.5	198712	2944
8/1/1967	12.6	509.8	4.7	198911	2945
9/1/1967	11.9	515.6	4.6	199113	2958
10/1/1967	12.9	512.2	4.9	199311	3143
11/1/1967	12.8	517.4	4.7	199498	3066
12/1/1967	11.8	525.1	4.8	199657	3018
1/1/1968	11.7	530.9	5.1	199808	2878
2/1/1968	12.3	533.6	4.5	199920	3001
3/1/1968	11.7	544.3	4.1	200056	2877
4/1/1968	12.3	544	4.6	200208	2709
5/1/1968	12	549.8	4.4	200361	2740
6/1/1968	11.7	556.3	4.4	200536	2938
7/1/1968	10.7	563.2	4.5	200706	2883
8/1/1968	10.5	567	4.2	200898	2768
9/1/1968	10.6	568.2	4.6	201095	2686
10/1/1968	10.8	571.6	4.8	201290	2689
11/1/1968	10.6	576.7	4.4	201466	2715
12/1/1968	11.1	576.5	4.4	201621	2685
1/1/1969	10.3	583.5	4.4	201760	2718
2/1/1969	9.7	588.7	4.9	201881	2692

Table 4: Data Description

DATE	month of data collection.
PSR	personal savings rate.
PCE	personal consumption expenditures, in billions of dollars.
MWU	median duration of unemployment, in week.
TP	total population, in thousands.
UL	number of unemployed in thousands.

Answer the following questions using data given in Tables 3 & 4:

D-1: List all the variables shown in Table 3. What is the use of Table 4's data.

D-2: Classify the variables in Table 3 into the following categories: Numerical, Categorical, Nominal, Ordinal.

D-3: Draw the histogram for MWU with 14 bins

D-4: From MWU's histogram, can you conclude that the distribution is symmetrical?

D-5: Find the mean, median and mode for the UL.

D-6: Find the variance and standard deviation for the UL.

D-7: Find the upper and lower quartile for UL.

D-8: Draw the box-plot for UL.

☞ *Note: Solve all the above questions using Python (not by hand).*

Problem #E (Practice only. No submission required.)

Explain the following **Python** codes. In the following codes, **array** stands for **numpy** array, **np** stands for **numpy** library, and **plt** stands for **matplotlib.pyplot** library:

Code-1: _____

```
In [1]: 1 plt.figure()
        2 plt.hist(array,bins=[10,21,31,41,51])
        3 plt.show()
```

Code-2: _____

```
In [2]: 1 subj_labels, subj_counts = np.unique(array, return_counts = True)
        2 plt.figure()
        3 plt.pie(subj_counts, subj_labels = unique,autopct='%.2f%%')
        4 plt.show()
```

Code-3: _____

```
In [3]: 1 A, B, C = np.percentile(array, list(range(25,100,25)))
        2 print(A, B, C)
```

Code-4: _____

```
In [4]: 1 _, A = ttest_1samp(array, 20)
        2 print(A)
```

Code-5: _____

```
In [5]: 1 _,B = ttest_ind(array,np.random.permutation(array))
        2 print(B)
```

Code-6: _____

```
In [6]: 1 print("Reject") if np.random.rand() <0.5 else print("Accept")
```

Code-7: _____

```
In [7]: 1 fig,ax = plt.subplots()
        2 plt.boxplot([array,np.random.permutation(array)-10])
        3 plt.title('Box Plot')
        4 ax.set_xticklabels(['Group1', 'Group2'])
        5 plt.show()
```

Code-8: _____

```
In [8]: 1 fig = plt.figure()
        2 ax = fig.add_subplot(121)
        3 plt.boxplot(array)
        4 plt.title('Plot 1')
        5 ax = fig.add_subplot(122)
        6 plt.hist(array,bins=5)
        7 plt.title('Plot 2')
        8 plt.show()
```