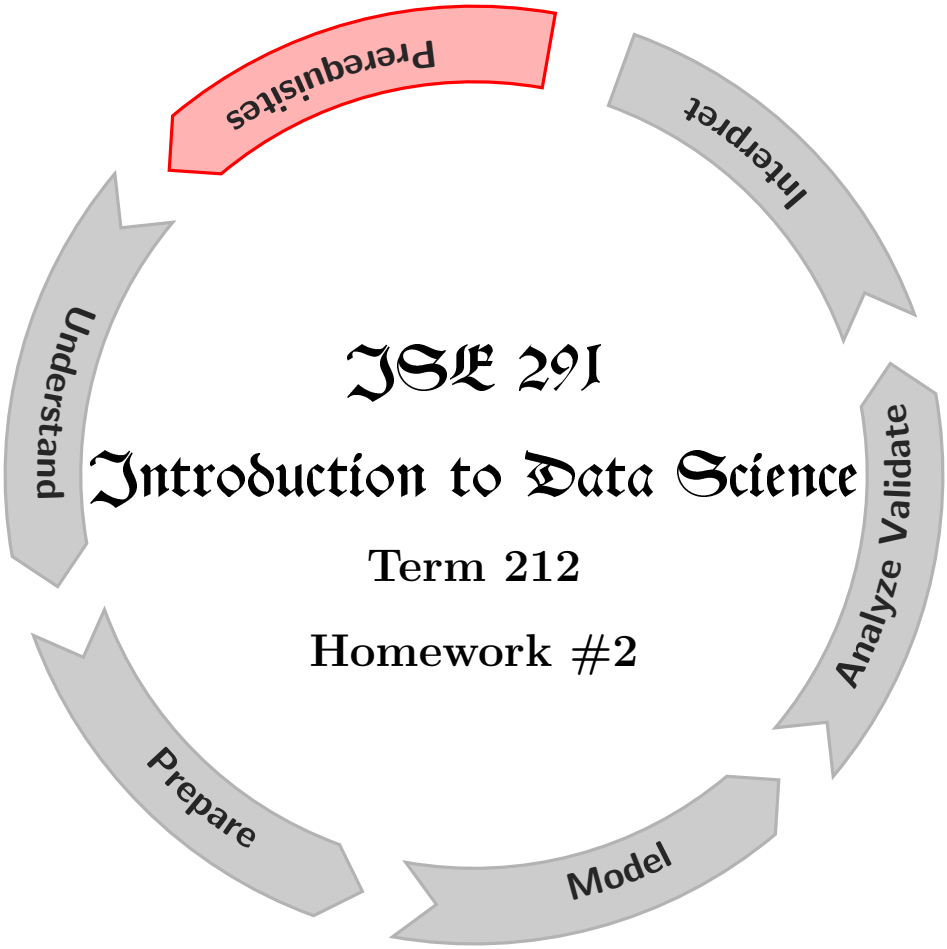
[The HW must be submitted as one .ipynb file. Write name & ID in the provided template.]



JSE 291

Introduction to Data Science

Term 212

Homework #2

Homework Guidelines

To receive full credit, you should make sure you adhere to the following guidelines. For any questions/-comments contact your section instructor.

Homework Presentation & Submission:

- You should submit the solutions for the **FIRST TWO** problems only.
- Every sub-problem (part) should be answered on a DIFFERENT CELL as given in the template.
- 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.
- Any text should be written as comment in the code cell. Do NOT modify code cell into markdown cell.
- Submit entire HW as ONE single .ipynb document and ONE additional pdf file containing figures.
- **Do NOT add/delete** any cell in the given template.

Problem # A **50 marks**

☞ *Note: Solve **Problem # A** by hand. You can use calculator and mathematical set (or geometry box). Do NOT use laptop/computer/mobile.*

Consider data given in Tables 1 & 2 obtained from a company selling gift products to customers living in the United Kingdom, Australia, Norway, Belgium, and Germany.

Table 1: Gift Products Data

Invoice No	Description	Qty	Unit Price	Cust ID	Country	Cust Pr.
536365	WHITE HANGING HEART T-LIGHT HOLDER	6	2.55	17850	United Kingdom	High
536367	ASSORTED COLOUR BIRD ORNAMENT	32	1.69	13047	United Kingdom	Low
536371	PAPER CHAIN KIT 50'S CHRISTMAS	80	2.55	13748	United Kingdom	Medium
536372	HAND WARMER RED POLKA DOT	6	1.85	17850	United Kingdom	Low
536374	VICTORIAN SEWING BOX LARGE	32	10.95	15100	United Kingdom	Medium
536375	WHITE HANGING HEART T-LIGHT HOLDER	6	2.55	17850	United Kingdom	High
536389	CHRISTMAS LIGHTS 10 REINDEER	6	8.5	12431	Australia	Low
536390	CHRISTMAS LIGHTS 10 REINDEER	2	8.5	17511	United Kingdom	High
581476	PANTRY MAGNETIC SHOPPING LIST	48	1.25	12433	Norway	Medium
581493	RETRO PLASTIC POLKA TRAY	15	0.42	12423	Belgium	Medium
581494	RABBIT NIGHT LIGHT	24	1.79	12518	Germany	Low
581495	WALL ART BICYCLE SAFETY	12	5.95	14051	United Kingdom	High
581570	CHRISTMAS CRAFT TREE TOP ANGEL	6	2.1	12662	Germany	Low
581571	HANGING MINI COLOURED BOTTLES	6	0.63	15311	United Kingdom	Low
581574	ROUND SNACK BOXES SET OF4 WOODLAND	6	2.95	12526	Germany	Medium
581579	JUMBO BAG PINK POLKADOT	10	1.79	17581	United Kingdom	High

Table 2: Data Description

Invoice No.	Invoice number. A 6-digit number uniquely assigned to each transaction.
Description	Product (item) name.
Qty.	Quantity: The quantities of each product (item) per transaction.
Unit Price	Product price per unit
Cust ID.	Customer number: A 5-digit integral number uniquely assigned to each customer
Country	Country name. the name of the country where each customer resides
Cust Pr.	Customer Priority

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

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

A-2: Identify the variable types in Table 1.

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

A-4: Draw the histogram for **Quantity** with three bins. *Submit all the drawings as one pdf file.*

A-5: Draw the pie chart for **Country**. *Submit all the drawings as one pdf file.*

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

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

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

A-9: Find the 25^{th} percentile (Q1) and 75^{th} percentile (Q3) for **Unit Price**. 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.

A-10: Draw the box-plot for **Unit Price**. *Submit all the drawings as one pdf file.*

Problem #B **50 marks**

☞ *Note: Solve all **Problem #B** questions using Python (not by hand).*

Consider the following data related to the vehicle mileage in city (miles per gallon) of the five different sport utility vehicle (SUV) car models (i.e., Model A to Model E):

Model A	3.13	5.67	4.73	4.7	6.38	6.04	7.88	4.1	6.17	4.72	5.12	4.34	4.83	4.43	6.87
	5.55	4.4	5.32	2.41	4.36	4.28	5.39	6.15	4.26	5.95	3.64	5.15	4.11	4.97	6.76
Model B	5.08	6.41	5.8	4.87	6.94	5.29	5.02	6.34	5.72	5.47	4.48	5.94	5.28	5.82	4.64
	5.06	5.57	5.36	7.1	6.3	6.08	4.52	5.25	4.91	4.47	4.4	5.09	5.72	5.64	4.78
Model C	4.88	4.15	5.45	5.05	3.96	6.45	5.82	8.5	3.83	6.25	7.18	6.9	5.82	5.07	3.82
	5.87	6.49	7.1	6.07	5.41	7.63	6.66	5.08	5.81	6.82	5.22	6.02	5.67	5.64	5.87
Model D	8.3	5.75	5.45	5.53	5.77	7.13	9.2	8.12	5.75	5.45	5.53	4.92	5.22	5.56	6.02
	5.1	5.21	6.58	6.14	9.2	6.93	9.2	5.36	6.39	3.1	5.85	6.75	4.75	9.2	6.41
Model E	6.43	5.18	5.29	5.31	4.04	6.43	3.95	5.17	4.74	5.45	4.91	6.69	5.29	8.4	5.17
	8.4	3.68	5.26	7.52	4.77	8.4	6.43	6.43	5.29	5.29	6.88	6.83	5.09	4.18	6.88

Answer the following questions:

B-1: To draw the frequency distribution of data, draw the histogram for each of the above model's data. Fix the figure width as 3 units and height as 5 units, and use 5 bins for each histogram.

B-2: To show the basic statistical summary, calculate the measures of central tendency (i.e., mean, median and mode) and measures of dispersion (population variance, sample standard deviation and interquartile range) for each model's data.

B-3: To find the outliers in data, draw the box-plot for each of the above model's data in one figure. Label x-axis with the name of the model. Indicate as comment which model have outlier(s).

B-4: To assess the normality, run the shapiro wilk test for each model's data by assuming 0.05 level of significance. State your findings.

B-5: To check the equality of means for all possible pairs of Model A to Model E, implement the appropriate test (i.e., either student's t-test or Mann-Whitney U test but not both) by assuming 0.05 level of significance. Set kwarg `alternative='two-sided'` and state your findings.

To read the data in python and repeat for each section, you can use the following code:

In [1]:
```
Data = {
   "Model_A": [3.13,5.67,4.73,4.70,6.38,6.04,7.88,4.10,6.17,4.72,5.12,4.34,4.83,4.43,6.87,5.55,
   4.40,5.32,2.41,4.36,4.28,5.39,6.15,4.26,5.95,3.64,5.15,4.11,4.97,6.76],
   "Model_B": [5.08,6.41,5.80,4.87,6.94,5.29,5.02,6.34,5.72,5.47,4.48,5.94,5.28,5.82,4.64,5.06,
   5.57,5.36,7.10,6.30,6.08,4.52,5.25,4.91,4.47,4.40,5.09,5.72,5.64,4.78],
   "Model_C": [4.88,4.15,5.45,5.05,3.96,6.45,5.82,8.50,3.83,6.25,17.18,6.90,5.82,5.07,3.82,5.87,
   6.49,7.10,6.07,5.41,7.63,6.66,5.08,5.81,6.82,5.22,6.02,5.67,5.64,5.87],
   "Model_D":[8.30,5.75,5.45,5.53,5.77,7.13,9.20,8.12,5.75,5.45,5.53,14.92,5.22,5.56,6.02,5.10,
   5.21,6.58,6.14,9.20,6.93,9.20,5.36,6.39,3.10,5.85,6.75,4.75,9.20,6.41],
   "Model_E": [6.43,5.18,5.29,5.31,4.04,6.43,3.95,5.17,4.74,5.45,4.91,6.69,5.29,8.40,5.17,8.40,
   3.68,5.26,7.52,4.77,8.40,6.43,6.43,5.29,5.29,6.88,6.83,5.09,4.18,6.88]
}

for section in Data: #dictionary's default iterator is key
   series=Data[section].copy() # list is mutable, so we copy to avoid chages in actual data
   # your code for each section
   #
   #
   #
   # 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: `plt.boxplot()`

C-6: `plt.title()`

C-7: `np.mean()`

C-8: `np.median()`

C-9: `scipy.stats.mode()`

C-10: `np.std()`

C-11: `np.var()`

C-12: `np.percentile()`

C-13: `scipy.stats.shapiro()`

C-14: `scipy.stats.ttest_ind()`

C-15: `scipy.stats.mannwhitneyu()`

C-16: `ax = fig.add_subplot()`

C-17: `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]:
```
plt.figure()
plt.hist(array,bins=[10,21,31,41,51])
plt.show()
```

Code-2:

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

Code-3:

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

Code-4:

In [4]:
```
_, A = shapiro(array)
print(A)
```

Code-5:

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

Code-6:

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

Code-7:

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

Code-8:

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