None

HW4 Solutions

Note:

1. Only one possible right answer is shown. All possible right answers will be given full credit.
2. Only the final solution is shown, and the details of actual code is not shown.
3. Some of the solutions have extra customization that is not specifically mentioned in the HW problem statement. Those customizations are for the purpose of illustration only. No credits will be deducted for extra customization.
4. You may come to the office hours or the help sessions to discuss the HW solutions.
5. If you find any typos or issues, kindly contact your section instructor.

Table of Contents

Problem-A

```
--------------------------------------------------Problem #A--------------------------------------------------
A-1:
```

	Channel	City_Town	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
406	NaN	NaN	$29,729.00	479	$7,326.00	$6,130.00	361.0	1083
247	Export	Dhahran	$6,338.00	226	$1,668.00	$1,492.00	311.0	686
233	NaN	NaN	$1,206.00	362	$2,857.00	$1,945.00	353.0	967
31	NaN	NaN	$43,088.00	210	$2,609.00	$1,200.00	1107.0	823
71	Hotels	Dammam	$39,228.00	143	$764.00	$4,510.00	93.0	2346
13	NaN	NaN	$9,385.00	153	$1,422.00	$3,019.00	227.0	684
415	Hotels	Khobar	$40,254.00	64	$3,600.00	$1,042.00	436.0	18
43	Hotels	Tabuk	$444.00	88	$2,060.00	$264.00	290.0	259
376	Hotels	Qaseem	$47,493.00	257	$3,779.00	$5,243.00	828.0	2253
143	Hotels	Dhahran	$8,861.00	378	$2,223.00	$633.00	1580.0	1521
292	Hotels	Khobar	$2,126.00	329	$3,281.00	$1,535.00	235.0	4365
67	Hotels	Dhahran	$14,100.00	213	$3,445.00	$1,336.00	1491.0	548
268	Retail	Khobar	$85.00	2096	$45,828.00	$36.00	24231.0	1423
428	Hotels	Dammam	$5,809.00	74	$803.00	$1,393.00	79.0	429
284	Hotels	Dammam	$2,615.00	87	$1,524.00	$1,103.00	514.0	468
166	NaN	NaN	$3,157.00	489	$2,500.00	$4,477.00	273.0	2165
29	Hotels	Dammam	$11,173.00	252	$3,355.00	$1,517.00	310.0	222
373	Hotels	Hail	$53,205.00	496	$7,336.00	$3,012.00	967.0	818
381	Hotels	Khobar	$25,767.00	361	$2,013.00	$10,303.00	314.0	1384
93	Hotels	Dhahran	$11,442.00	103	$582.00	$5,390.00	74.0	247

```
The fields for each column is presented in A-2.

The number of rows in the data is 451, and the number of columns in the data is 8.

The non-null values in each column are:
Channel             417
City_Town           417
Fresh               451
Milk                451
Grocery             451
Frozen              451
Detergents_Paper    446
Delicassen          451
dtype: int64
```

```
----------------------------------------------------------------------------------------------------
A-2:
-------------------------------------------------------------------------
|Field           |Actual Data Type  |Python Data Type |Consistency Check |
-------------------------------------------------------------------------
|Channel         |Categorical       |object           |                  |
|City_Town       |Categorical       |object           |                  |
|Fresh           |Numerical         |object           |inconsistent      |
|Milk            |Numerical         |int64            |                  |
|Grocery         |Numerical         |object           |inconsistent      |
|Frozen          |Numerical         |object           |inconsistent      |
|Detergents_Paper|Numerical         |float64          |                  |
|Delicassen      |Numerical         |int64            |                  |
-------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------
A-3:
The values in Channel column are: ['Retail', 'Hotels', 'Export', nan].
Removing the rows containing Export in Channel.
```

	Channel	City_Town	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
0	Retail	Dhahran	$11,170.00	1077	$8,814.00	$2,194.00	1976.0	143
1	Retail	Dammam	$21,217.00	621	$14,982.00	$3,095.00	6707.0	602
2	Hotels	Khobar	$5,963.00	365	$6,192.00	$425.00	1716.0	750
3	Retail	Khobar	$9,413.00	826	$5,126.00	$666.00	1795.0	1451
4	Hotels	Dhahran	$5,969.00	199	$3,417.00	$5,679.00	1135.0	290
...	...	...	...	...	...	...	...	...
446	Hotels	Dammam	$717.00	359	$6,532.00	$7,530.00	529.0	894
447	Hotels	Dhahran	$22,335.00	120	$2,406.00	$2,046.00	101.0	558
448	NaN	NaN	$45,640.00	696	$6,536.00	$7,368.00	1532.0	230
449	NaN	NaN	$23,632.00	673	$3,842.00	$8,620.00	385.0	819
450	Retail	Madinah	$918.00	2066	$13,567.00	$1,465.00	6846.0	806

440 rows × 8 columns

```
----------------------------------------------------------------------------------------------------
A-4:
```

	Channel	City_Town	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
0	Retail	Dhahran	11170.0	10555.0	8814.0	2194.0	1976.0	143
1	Retail	Dammam	21217.0	6086.0	14982.0	3095.0	6707.0	602
2	Hotels	Khobar	5963.0	3577.0	6192.0	425.0	1716.0	750
3	Retail	Khobar	9413.0	8095.0	5126.0	666.0	1795.0	1451
4	Hotels	Dhahran	5969.0	1950.0	3417.0	5679.0	1135.0	290
...	...	...	...	...	...	...	...	...
446	Hotels	Dammam	717.0	3518.0	6532.0	7530.0	529.0	894
447	Hotels	Dhahran	22335.0	1176.0	2406.0	2046.0	101.0	558
448	NaN	NaN	45640.0	6821.0	6536.0	7368.0	1532.0	230
449	NaN	NaN	23632.0	6595.0	3842.0	8620.0	385.0	819
450	Retail	Madinah	918.0	20247.0	13567.0	1465.0	6846.0	806

440 rows × 8 columns

```
----------------------------------------------------------------------------------------------------
A-5:

Initial number of NaN values:
Channel             34
City_Town           34
Fresh                0
Milk                 0
Grocery              0
Frozen               0
Detergents_Paper     5
Delicassen           0
dtype: int64

Number of NaN values after transfromation:
Channel             0
City_Town           0
Fresh               0
Milk                0
Grocery             0
Frozen              0
Detergents_Paper    0
Delicassen          0
dtype: int64
----------------------------------------------------------------------------------------------------
A-6:

Before transformation:
```

	Channel	City_Town	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
0	Retail	Dhahran	11170.0	10555.0	8814.0	2194.0	1976.0	143
1	Retail	Dammam	21217.0	6086.0	14982.0	3095.0	6707.0	602
2	Hotels	Khobar	5963.0	3577.0	6192.0	425.0	1716.0	750
3	Retail	Khobar	9413.0	8095.0	5126.0	666.0	1795.0	1451
4	Hotels	Dhahran	5969.0	1950.0	3417.0	5679.0	1135.0	290

```
After transformation:
```

	Channel	City_Town	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
0	0	Dhahran	11170.0	10555.0	8814.0	2194.0	1976.0	143
1	0	Dammam	21217.0	6086.0	14982.0	3095.0	6707.0	602
2	1	Khobar	5963.0	3577.0	6192.0	425.0	1716.0	750
3	0	Khobar	9413.0	8095.0	5126.0	666.0	1795.0	1451
4	1	Dhahran	5969.0	1950.0	3417.0	5679.0	1135.0	290

```
----------------------------------------------------------------------------------------------------
A-7:

After transformation:
```

	Channel	City_Town	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen	Region
153	1	Khobar	6269.0	1078.0	1980.0	3860.0	609.0	2162	Eastern
8	1	Qaseem	9193.0	4792.0	2157.0	327.0	780.0	548	Central
6	1	Dammam	16260.0	578.0	1296.0	848.0	445.0	258	Eastern
360	0	Dammam	3103.0	13789.0	21955.0	1668.0	6792.0	1452	Eastern
146	0	Tabuk	8090.0	3136.0	6986.0	1455.0	3712.0	531	Western
315	1	Dhahran	6300.0	1264.0	2591.0	1170.0	199.0	326	Eastern
336	0	Dammam	1406.0	16395.0	28986.0	673.0	836.0	3	Eastern
373	1	Hail	53205.0	4861.0	7336.0	3012.0	967.0	818	Central

```
----------------------------------------------------------------------------------------------------
A-8:

Since we are asked to not to delete the oringinal column. We can duplicate the column before doing one hot enc
oding.

We can use 'df.rename' to rename the columns later. Look for help(df.rename).

After transformation, we deleted the first encoded column. The final df looks like:
```

	Channel	City_Town	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen	Region	Region_Eastern	Region_Western
278	1	Dammam	42312.0	911.0	1510.0	1718.0	410.0	1819	Eastern	1	0
163	0	Khobar	31714.0	12074.0	11757.0	287.0	3881.0	2931	Eastern	1	0
301	1	Dhahran	5841.0	1421.0	1162.0	597.0	476.0	70	Eastern	1	0
148	1	Dammam	13779.0	1931.0	1648.0	596.0	227.0	436	Eastern	1	0
107	1	Dammam	16225.0	1793.0	1765.0	853.0	170.0	1067	Eastern	1	0
188	1	Hail	3317.0	6468.0	6861.0	1329.0	3961.0	1215	Central	0	0
4	1	Dhahran	5969.0	1950.0	3417.0	5679.0	1135.0	290	Eastern	1	0
208	1	Dammam	10405.0	1568.0	1096.0	8425.0	399.0	318	Eastern	1	0

```
----------------------------------------------------------------------------------------------------
A-9:

After transformation:
```

	Channel	City_Town	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen	Region	Region_Eastern	Region_Western
359	1	Khobar	-0.296233	-0.468946	-0.611468	0.169941	-0.549294	-0.315904	Eastern	1	0
26	1	Dammam	-0.384019	-0.481126	-0.706808	-0.299626	-0.347716	-0.495889	Eastern	1	0
280	0	Makkah	0.352072	0.060091	-0.167638	-0.463572	-0.046188	0.170446	Western	0	1
254	1	Dhahran	1.286849	-0.658856	-0.611574	0.166847	-0.479161	-0.445479	Eastern	1	0
52	1	Qaseem	-0.779493	-0.634356	-0.133505	0.258203	0.000007	-0.395424	Central	0	0
39	1	Khobar	-0.303199	0.088606	-0.717659	-0.183729	-0.592969	-0.453644	Eastern	1	0
83	1	Dhahran	-0.874325	-0.043031	-0.149096	-0.564827	0.000007	-0.290344	Eastern	1	0
3	0	Khobar	-0.204806	0.334161	-0.297637	-0.496155	-0.228239	-0.026224	Eastern	1	0

```
--------------------------------------------------------------------------------------------------------
A-10:

Clean & Prepared Data:
```

	Channel	City_Town	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen	Region	Region_Eastern	Region_Western
200	1	Qaseem	-0.264728	-0.413301	-0.574175	0.778088	-0.459423	-0.193429	Central	0	0
411	0	Dhahran	0.794644	0.350356	2.827614	-0.624837	2.038675	1.031322	Eastern	1	0
202	1	Makkah	0.307189	-0.690001	-0.488422	0.130140	-0.296481	-0.460034	Western	0	1
432	1	Dammam	0.557724	-0.410533	-0.577546	1.192181	-0.503729	0.420011	Eastern	1	0
90	1	Qaseem	-0.372779	-0.649305	-0.620423	-0.367267	-0.554753	0.088796	Central	0	0
212	1	Tabuk	0.155918	-0.581480	-0.697538	-0.499455	-0.574491	-0.265139	Western	0	1
243	1	Khobar	-0.929735	-0.004965	-0.213675	-0.561527	0.359698	-0.069889	Eastern	1	0
39	1	Khobar	-0.303199	0.088606	-0.717659	-0.183729	-0.592969	-0.453644	Eastern	1	0

```
Summary Statistics Numerical Columns:
```

	count	mean	std	min	25%	50%	75%	max
Channel	440.0	6.772727e-01	0.468052	0.000000	0.000000	1.000000	1.000000	1.000000
Fresh	440.0	-3.885781e-17	1.001138	-0.949683	-0.702334	-0.276760	0.390523	7.927738
Milk	440.0	8.074349e-18	1.001138	-0.778173	-0.578850	-0.294607	0.188613	9.183973
Grocery	440.0	-7.292147e-17	1.001138	-0.837334	-0.610836	-0.336668	0.284911	8.936528
Frozen	440.0	1.892426e-18	1.001138	-0.628343	-0.480431	-0.318804	0.099464	11.919002
Detergents_Paper	440.0	4.743680e-17	1.001138	-0.604518	-0.551236	-0.433701	0.218383	7.967591
Delicassen	440.0	-5.942216e-17	1.001138	-0.540264	-0.396401	-0.198577	0.104860	16.478447
Region_Eastern	440.0	7.181818e-01	0.450397	0.000000	0.000000	1.000000	1.000000	1.000000
Region_Western	440.0	1.068182e-01	0.309234	0.000000	0.000000	0.000000	0.000000	1.000000

```
Summary Statistics Categorical Columns:
```

	count	unique	top	freq
City_Town	440	9	Khobar	125
Region	440	3	Eastern	316

```
-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~
```

Problem-B

```
--------------------------------------------------Problem #B--------------------------------------------------
B-1:
Display the data:
```

	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score	gk score
0	female	group B	bachelor's degree	standard	none	21.6	72.00	29.6	71.0
1	female	group C	some college	standard	completed	20.7	22.50	35.2	NaN
2	female	group B	master's degree	standard	none	27.0	23.75	37.2	93.0
3	male	group A	associate's degree	free/reduced	none	14.1	NaN	17.6	56.0
4	male	group C	some college	standard	none	76.0	19.50	30.0	77.0
...	...	...	...	...	...	...	...	...	...
995	female	group E	master's degree	standard	completed	26.4	24.75	38.0	93.0
996	male	group C	high school	free/reduced	none	18.6	13.75	55.0	57.0
997	female	group C	high school	free/reduced	completed	17.7	71.00	26.0	63.0
998	female	group D	some college	standard	completed	20.4	19.50	30.8	75.0
999	female	group D	some college	free/reduced	none	77.0	21.50	-34.4	85.0

1000 rows × 9 columns

```
The number of rows in the data is 1000, and the number of columns in the data is 9.

The missing values in each column are:
gender                          0
race/ethnicity                  0
parental level of education     0
lunch                           9
test preparation course         0
math score                     36
reading score                  78
writing score                  39
gk score                      113
dtype: int64

Summary Statistics Numerical Columns:
```

	count	mean	std	min	25%	50%	75%	max
math score	964.0	19.727905	14.076062	-29.1	16.2	19.8	23.40	98.0
reading score	922.0	18.649675	15.468953	-25.0	14.5	17.5	20.25	100.0
writing score	961.0	26.394589	15.261831	-36.8	22.4	27.6	32.00	100.0
gk score	887.0	59.310034	35.740312	-100.0	55.0	66.0	77.00	100.0

```
Summary Statistics Categorical Columns:
```

	count	unique	top	freq
gender	1000	2	female	518
race/ethnicity	1000	5	group C	319
parental level of education	1000	6	some college	226
lunch	991	2	standard	639
test preparation course	1000	2	none	642

```
----------------------------------------------------------------------------------------------------
B-2:
-----------------------------------------------------------------------------------------
|Field                             |Actual Data Type  |Python Data Type  |Consistency Check |
-----------------------------------------------------------------------------------------
|gender                            |Categorical       |object            |                  |
|race/ethnicity                    |Categorical       |object            |                  |
|parental level of education       |Ordinal           |object            |                  |
|lunch                             |Categorical       |object            |                  |
|test preparation course           |Categorical       |object            |                  |
|math score                        |Numerical         |float64           |                  |
|reading score                     |Numerical         |float64           |                  |
|writing score                     |Numerical         |float64           |                  |
-----------------------------------------------------------------------------------------
```

```
----------------------------------------------------------------------------------------------------
B-3:

After transformation:
```

	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score	gk score
0	female	group B	bachelor's degree	standard	none	21.6	18.00	29.6	71.0
1	female	group C	some college	standard	completed	20.7	22.50	35.2	NaN
2	female	group B	master's degree	standard	none	27.0	23.75	37.2	93.0
3	male	group A	associate's degree	free/reduced	none	14.1	NaN	17.6	56.0
4	male	group C	some college	standard	none	22.8	19.50	30.0	77.0
...	...	...	...	...	...	...	...	...	...
995	female	group E	master's degree	standard	completed	26.4	24.75	38.0	93.0
996	male	group C	high school	free/reduced	none	18.6	13.75	22.0	57.0
997	female	group C	high school	free/reduced	completed	17.7	17.75	26.0	63.0
998	female	group D	some college	standard	completed	20.4	19.50	30.8	75.0
999	female	group D	some college	free/reduced	none	23.1	21.50	34.4	85.0

1000 rows × 9 columns

```
----------------------------------------------------------------------------------------------------
B-4:

After transformation:
```

	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score	gk score
0	female	group B	bachelor's degree	standard	none	21.6	18.00	29.6	71.000000
1	female	group C	some college	standard	completed	20.7	22.50	35.2	82.333333
2	female	group B	master's degree	standard	none	27.0	23.75	37.2	93.000000
3	male	group A	associate's degree	free/reduced	none	14.1	12.25	17.6	56.000000
4	male	group C	some college	standard	none	22.8	19.50	30.0	77.000000

```
----------------------------------------------------------------------------------------------------
B-5:

Identify non-numeric columns with NaN values:
gender                         False
race/ethnicity                 False
parental level of education    False
lunch                           True
test preparation course        False
dtype: bool
Value occurrences in lunch column before transformation:
standard        639
free/reduced    352
Name: lunch, dtype: int64
When 'race/ethnicity'=group A, then the mode of corresponding 'lunch' rows is = standard
When 'race/ethnicity'=group B, then the mode of corresponding 'lunch' rows is = standard
When 'race/ethnicity'=group C, then the mode of corresponding 'lunch' rows is = standard
When 'race/ethnicity'=group D, then the mode of corresponding 'lunch' rows is = standard
When 'race/ethnicity'=group E, then the mode of corresponding 'lunch' rows is = standard

Value occurrences in lunch column after transformation:
standard        648
free/reduced    352
Name: lunch, dtype: int64
```

```
----------------------------------------------------------------------------------------------------
B-6:

Before transformation:
```

	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score	gk score
0	female	group B	bachelor's degree	standard	none	21.6	18.00	29.6	71.000000
1	female	group C	some college	standard	completed	20.7	22.50	35.2	82.333333
2	female	group B	master's degree	standard	none	27.0	23.75	37.2	93.000000
3	male	group A	associate's degree	free/reduced	none	14.1	12.25	17.6	56.000000
4	male	group C	some college	standard	none	22.8	19.50	30.0	77.000000

```
After transformation:
```

	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score	gk score
0	female	group B	4	1	none	21.6	18.00	29.6	71.000000
1	female	group C	2	1	completed	20.7	22.50	35.2	82.333333
2	female	group B	5	1	none	27.0	23.75	37.2	93.000000
3	male	group A	3	0	none	14.1	12.25	17.6	56.000000
4	male	group C	2	1	none	22.8	19.50	30.0	77.000000

```
----------------------------------------------------------------------------------------------------
B-7:

Before transformation:
```

	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score	gk score
0	female	group B	4	1	none	21.6	18.00	29.6	71.000000
1	female	group C	2	1	completed	20.7	22.50	35.2	82.333333
2	female	group B	5	1	none	27.0	23.75	37.2	93.000000
3	male	group A	3	0	none	14.1	12.25	17.6	56.000000
4	male	group C	2	1	none	22.8	19.50	30.0	77.000000

```
After transformation:
```

	gender	race/ethnicity	parental level of education	lunch	math score	reading score	writing score	gk score	test preparation course_none
0	female	group B	4	1	21.6	18.00	29.6	71.000000	1
1	female	group C	2	1	20.7	22.50	35.2	82.333333	0
2	female	group B	5	1	27.0	23.75	37.2	93.000000	1
3	male	group A	3	0	14.1	12.25	17.6	56.000000	1
4	male	group C	2	1	22.8	19.50	30.0	77.000000	1

```
----------------------------------------------------------------------------------------------------
B-8:

Before transformation:
```

	gender	race/ethnicity	parental level of education	lunch	math score	reading score	writing score	gk score	test preparation course_none
0	female	group B	4	1	21.6	18.00	29.6	71.000000	1
1	female	group C	2	1	20.7	22.50	35.2	82.333333	0
2	female	group B	5	1	27.0	23.75	37.2	93.000000	1
3	male	group A	3	0	14.1	12.25	17.6	56.000000	1
4	male	group C	2	1	22.8	19.50	30.0	77.000000	1

```
After transformation:
```

	gender	race/ethnicity	parental level of education	lunch	math score	reading score	writing score	gk score	test preparation course_none
0	female	group B	4	1	0.72	0.662651	0.711111	0.681319	1
1	female	group C	2	1	0.69	0.879518	0.866667	0.805861	0
2	female	group B	5	1	0.90	0.939759	0.922222	0.923077	1
3	male	group A	3	0	0.47	0.385542	0.377778	0.516484	1
4	male	group C	2	1	0.76	0.734940	0.722222	0.747253	1

```
----------------------------------------------------------------------------------------------------
B-9:

Clean & Prepared Data:
```

	gender	race/ethnicity	parental level of education	lunch	math score	reading score	writing score	gk score	test preparation course_none
986	female	group C	3	1	0.400000	0.506024	0.455556	0.450549	1
402	female	group A	2	0	0.490000	0.578313	0.500000	0.560440	1
413	male	group B	0	1	0.630000	0.602410	0.633333	0.604396	0
97	female	group E	2	1	0.630000	0.662651	0.666667	0.659341	0
71	male	group D	2	1	0.593333	0.457831	0.588889	0.560440	0
545	male	group E	0	0	0.780000	0.795181	0.777778	0.791209	0
785	female	group B	0	1	0.320000	0.409639	0.377778	0.461538	0
95	male	group C	3	0	0.780000	0.751004	0.800000	0.758242	0

```
Summary Statistics Numerical Columns:
```

	count	mean	std	min	25%	50%	75%	max
parental level of education	1000.0	2.081000	1.460333	0.0	1.000000	2.000000	3.000000	5.0
lunch	1000.0	0.648000	0.477833	0.0	0.000000	1.000000	1.000000	1.0
math score	1000.0	0.660910	0.151296	0.0	0.570000	0.660000	0.770000	1.0
reading score	1000.0	0.626747	0.175521	0.0	0.506024	0.638554	0.746988	1.0
writing score	1000.0	0.645844	0.168194	0.0	0.533333	0.655556	0.766667	1.0
gk score	1000.0	0.645696	0.159264	0.0	0.538462	0.648352	0.758242	1.0
test preparation course_none	1000.0	0.642000	0.479652	0.0	0.000000	1.000000	1.000000	1.0

```
Summary Statistics Categorical Columns:
```

	count	unique	top	freq
gender	1000	2	female	518
race/ethnicity	1000	5	group C	319

```
~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-~-
```