DNSC6279 Data Mining
Quiz 01 Name:

1. **(2 pts.)** According to Tan, name four measurement levels for attributes (e.g. columns) in a data set.
Nominal; ordinal; interval; ratio – ½ pt. each

2. **(2 pts.)** Data Set 1 below suffers from the common problem known as being *untidy*. Write out a *tidy* representation of this data set.

Location	Income: < $50K	Income: $51K – $100K	Income: > $101K
On-site	409	104	14
Remote	57	0	0

Data Set 1

Location	Income	Frequency
Onsite	<$50K	409
Onsite	$51K – $100K	104
Onsite	> $101K	14
Remote	<$50K	57

Row values must match, but column and row order don't matter.

½ pt. for each correct row

3. (**1 pt.)** Describe the fundamental difference between data storage in Base SAS and Pandas.
Base SAS stores data on disk. Pandas holds data in RAM.

4. **(1 pt.)** Given Data Set 2 below, write out the results of selecting the rows where X1 ≥ 5.

ID	X1	X2
1111	3	D
1112	4	C
1113	5	A
1114	6	A

Data Set 2

ID	X1	X2
1113	5	A
1114	6	A

Row values must match, but column and row order don't matter.

½ pt. for each correct row

5. **(1 pt.)** Given Data Set 3 below, write out the results of grouping Data Set 3 by ID and summing X3.

ID	X3
1111	11
1112	1
1112	2
1111	13

Data Set 3

Row values must match, but column and row order don't matter.

½ pt. for each correct row

ID	X3
1111	24
1112	3

6. **(2 pts.)** Given Data Sets 2 and 3 above, write out the results of ***left*** joining Data Set 3 onto Data Set 2 by the key variable shared between the two data sets. (Treat Data Set 2 as the left table.)

ID	X1	X2	X3
1111	3	D	11
1111	3	D	13
1112	4	C	1
1112	4	C	2
1113	5	A	
1114	6	A	

Row values must match, but column and row order don't matter.

1/3 pt. for each correct row

7. **(1 pt.)** Name the author and prolific R contributor who wrote "Tidy Data" and who is the primary developer of the R packages `ggplot` and `dplyr`. Hadley Wickham