CMPE188-F18-HW1 Data Preprocessing

Due Sep 12 at 11:59am **Points** 9 **Questions** 9
Available after Sep 6 at 12pm **Time Limit** None **Allowed Attempts** 5

Instructions

** Please notify me in case you notice mismatch between your answers and the expected correct ones here ***

This assignment needs some work outside the Canvas environment. It consists of three distinct parts: a) short computational questions (to be solved on paper), b) questions based on the weather dataset (some coding needed, to be answered using the provided notebook), and c) questions based on the Titanic tutorial (some coding needed, to be answered using either Kaggle's cloud, or your own laptop).

I recommend that you look at the questions, then take your time to do all that is needed (in Python) and then come back to answer the questions.

Before you begin, do the following:

- Read the description of the weather dataset and go over the data preparation tutorial in your Jupyter notebook.

- Read the description of the Titanic competition. (https://www.kaggle.com/c/titanic) For this assignment we will go through the data preprocessing steps of the introductory data science tutorial (https://www.kaggle.com/helgejo/titanic/an-interactive-data-science-tutorial) using this dataset. You are free to either run online on Kaggle's cloud, or download the notebook and datasets (https://www.kaggle.com/c/titanic/data) (training and test) and run on your laptop.

You are given 5 attempts. Your final score will be the highest score of the 5.

Take the Quiz Again

Attempt History

	Attempt	Time	Score
KEPT	Attempt 2	3 minutes	9 out of 9
LATEST	Attempt 2	3 minutes	9 out of 9
	Attempt 1	78 minutes	7 out of 9

Answers will be shown after your last attempt

Score for this attempt: **9** out of 9

Submitted Sep 11 at 11:16pm

This attempt took 3 minutes.

Question 1

1 / 1 pts

The following table shows the years in position and bonus for the employees of a company.

ID	years in position	bonus
1	4	300
2	-	200
3	6	500
4	3	-
5	3	-
6	2	100

You are asked to impute the missing values using the median of each attribute. What is the imputed value of years in position for employee #2?

○ 3.5

● 3

○ 3.6

○ 4

Question 2

1 / 1 pts

The following table shows the years in position and bonus for the employees of a company.

ID years in position bonus

1	4	300
2	-	200
3	6	500
4	3	-
5	3	-
6	2	100

After imputing the missing values using the median of each attribute, you are asked to normalize the years_in_position and bonus attribute values using min-max normalization.

What are the (imputed, then normalized) values of instance #5?

○ years = 0.5 and bonus = 0.5

○ years = 0.25 and bonus = 0.5

● years = 0.25 and bonus = 0.375

○ years = 0 and bonus = 0.5

Question 3

1 / 1 pts

(for this answer you need to use/update the Data Preparation python script on weather data)

How many samples containing rain accumulation at 9am measurements have missing values?

● 6

○ 3

○ 4

○ 1095

○ 2

○ 1089

Question 4

1 / 1 pts

True or False?

All the attributes of the weather dataset are numerical.

● False

The "number" variable, representing the unique identifier of each tuple, is a nominal (i.e. categorical) value.

○ True

Question 5

1 / 1 pts

(for this answer you need to use/update the Data Preparation python script)

When we remove all the missing values from the dataset, the number of rows is 1064, yet the variable with most missing values has 1089 rows. Why did the number of rows decrease so much?

○ Because rows with missing values as well as rows with 0s are removed

○ Because rows with missing values as well as rows with duplicate values are removed

● Because the missing values in each column are not necessarily in the same row

Question 6

1 / 1 pts

Which of the following is true, looking at the variables' (attributes') descriptions?

○ The youngest passenger was 14 and the oldest was 30 years old.

○ The youngest passenger was an infant and the oldest was 29 years old.

○ We cannot easily derive the age of the youngest and oldest passenger on board.

● The youngest passenger was an infant and the oldest was 80 years old.

○ The youngest passenger was 14 and the oldest was 80 years old.

Question 7

1 / 1 pts

Looking at the attribute correlation heat map, which of the following statements is true?

● The most strong positive correlation to the survival attribute exists between this and the fare.

○ The most strong positive correlation to the survival attribute exists between this and the passenger's class.

○ No answer text provided.

○ The most strong positive correlation to the survival attribute exists between this and the number of spouses/siblings on board.

Question 8

1 / 1 pts

Try to identify indicators for survival, plotting the rate of survival for the following indicators (as instructed in the tutorial). Which of the following is the strongest indicator overall (out of those provided below)?

○ Number of siblings/spouses aboard.

○ Number of parents/children aboard.

● Gender

○ No answer text provided.

Question 9

1 / 1 pts

Turning the categorical variables into numerical, which of the following is a good strategy (take as examples the ones provided in the tutorial)?

○ Create at least two and at most three numerical variables for each categorical one.

● Create as many numerical variables as the possible values of the categorical attribute.

○ There's no need to do this. Most algorithms can work with both numerical and categorical attributes.

○ Only binary attributes can be turned into numerical ones.

Quiz Score: **9** out of 9