Bus 674 Graded HW 1

Email *

will.jiang@gmail.com

First Name

Wei (Will)

Last Name

Jiang

Net ID

Use your OPUS user name! Not your student ID number!

WJIAN63

Q1.1

In the sample code file, what is the R data type of the object AllData? (check all that apply)

- [ ] (a) A vector of integers
- [ ] (b) A vector of doubles
- [x] (c) A list
- [x] (d) A data frame
- [ ] (e) A matrix of integers
- [ ] (f) A matrix of doubles

Q 1.2

In the input csv file, what will be interpreted as a missing value by read_table( ) in the first line of code? (Note: the double quotes below are not a part of the string.)

- ( ) (a) the string "na"
- ( ) (b) a space (i.e. " ")
- ( ) (c) a period (i.e., ".")
- ( ) (d) a blank (i.e., "")
- ( ) (e) the string "NA"
- (•) (f) No input data will be interpreted as a missing value.

Q 1.3

What does line 7 do? Line 7: AllData <- AllData[AllData$Bldg.Type=="1Fam",]

- (a) Sets Bldg.Type to "1Fam"
- (b) Eliminates all single family homes from the data set.
- (c) Keeps only single family homes in the data set
- (d) Keeps only the column (variable) 1Fam
- (e) Changes the number of columns in AllData
- (f)) None of the above

Q 1.4

What does line 8 do? Line 8: AllData <- AllData[RPerm,]

- (a) Reduces the number of rows on AllData
- (b) Reduces the number of columns of AllData
- (c) Creates the training sample
- (d) Reorders the rows of AllData according to the random permutation in RPerm
- (e) Reorders the columns of AllData according to the indices in RPerm
- (f) Creates the validation sample.

Q 1.5a

The code sample correctly creates the training, validation, and test samples as instructed in the problem.

- (●) True
- ( ) False

Q 1.5b

Explain why you answered the way you did in Q15a. Be brief.

The code removes all observations except those that are single family homes first by using logical condition to slice data, then orders the obsevations by using sample() function, and finally splits the training, validation and test samples with correct ratio.

Question 2

Q 2.1

Which variables (in which of the training, validation, and test data sets) have missing values?

	Training Data	Validation Data	Test Data
Lot.Area	☐	☐	☐
Total.Bsmt.SF	☐	☑	☐
Gr.Liv.Area	☐	☐	☐
Full.Bath	☐	☐	☐
Bedroom.AbvGr	☐	☐	☐
Building Age	☐	☐	☐

Q 2.2

Based on the training data, which variables show strong right skewness with upper tail outliers?

- ☑ (a) Lot.Area
- ☑ (b) Total.Bsmt.SF
- ☑ (c) Gr.Liv.Area
- ☐ (d) Full.Bath
- ☐ (e) Bedroom.AbvGr
- ☐ (f) Building Age

Question 3 (Raw, Unstandardized Data)

Q 3.1

What is the root MSE (based on the validation data) when k = 1 (to 2 decimal places).

62833.09

Q3.2

What is the root MSE (based on the validation data) when k = 20 (to 2 decimal places).

50538.81

Question 4 (Raw, Unstandardized Data)

Q 4.1

What is the best k?

12

Q 4.2

What is the root MSE for the test data for the model using the best k (to 2 decimal places).

58539.02

Question 5 (Raw, Standardized Data)

Q 5.1

Did you standardize the SalePrice variable (the Y variable) in the following data sets?

	Yes	No
Training Data	☐	☑
Validation Data	☐	☑
Test Data	☐	☑

Q 5.2

What is the root MSE (based on the validation data) when k = 1 (to 2 decimal places).

50349.44

Q5.3

What is the root MSE (based on the validation data) when k = 20 (to 2 decimal places).

41264.58

Question 6 (Raw, Standardized Data)

Q 6.2

What is the best k?

12

Q 6.3

What is the root MSE for the test data for the model using the best k (to 2 decimal places).

44376.97

Question 8 (Transformed, Unstandardized Data)

Note: No questions on Question 7.

Q 8.1

Did you transform the SalePrice variable (the Y variable) in the following data sets?

	Yes	No
Training Data	☑	☐
Validation Data	☑	☐
Test Data	☑	☐

Q 8.2

What is the best k?

2

Q 8.3

What is the root MSE for the model using the best k (to 2 decimal places) for the validation data.

53278.00

Question 10 (Transformed, Standardized Data)

Note: No questions on Question 9.

Q 10.1

Did you transform and standardize the SalePrice variable (the Y variable) in the following data sets?

	Yes	No
Training Data	☐	☑
Validation Data	☐	☑
Test Data	☐	☑

Q 10.2

What is the best k?

10

Q 10.3

What is the root MSE for the model using the best k (to 2 decimal places) for the validation data.

41776.35

Question 11

Q 11.1

Which model is best overall?

- ( ) (a) The best k model for the raw, unstandardized data
- (•) (b) The best k model for the raw, standardized data
- ( ) (c) The best k model for the transformed, unstandardized data
- ( ) (d) The best k model for the transformed, standardized data
- ( ) (e) There is no best overall model
- ( ) (f) It is not possible to tell if there if there is an overall best model or if there is no overall best model.

Q 11.2

In a paragraph (or perhaps two), explain why you chose the answer you did for Q 11.1

I used validation data for comparing the RMSE of all KNN models thus to decide the best model, and I choosed untransformed, standardized data to build the best model since:
1) it has the lowest root MSE(39809.89) among all the four models.
2) KNN is not a linear model and we even did not build an actual model, so there is no statistical assumption needed and trasnformation is not necessary for KNN.