**Individual Assignment 3**

**Topics: Linear Regression for Prediction; Evaluating Predictive Performance**

**Submission instruction:**

1. Upload a single PDF to Canvas. It should include all the screenshots and answers (a)-(k).
2. Hand in the printed copy of your document in class.

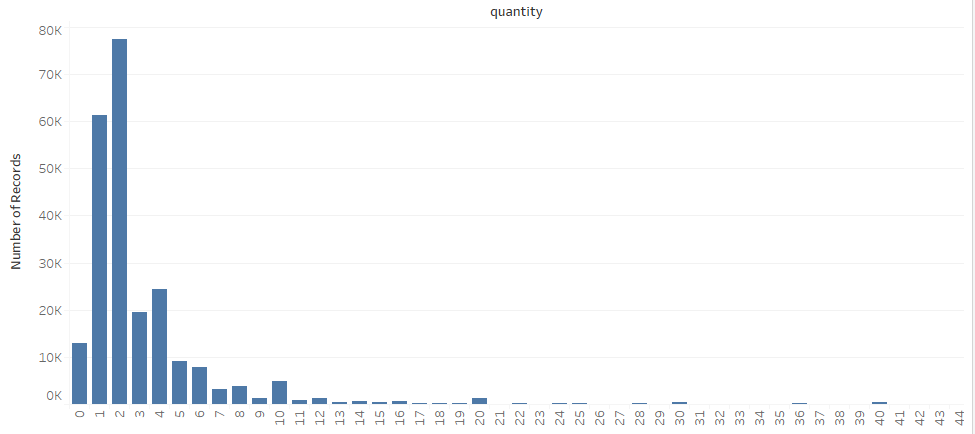
**Data Details and Goal:**

The file “purchase\_item.txt” contains transaction-level information for over 200,000 purchases of restaurant prepaid vouchers and coupons sold on EZTABLE (www.eztable.com). Each transaction includes the purchase of one or more vouchers/coupons. Our goal in this assignment is to build a predictive model for ‘quantity’ (=number of units sold in a single transaction).

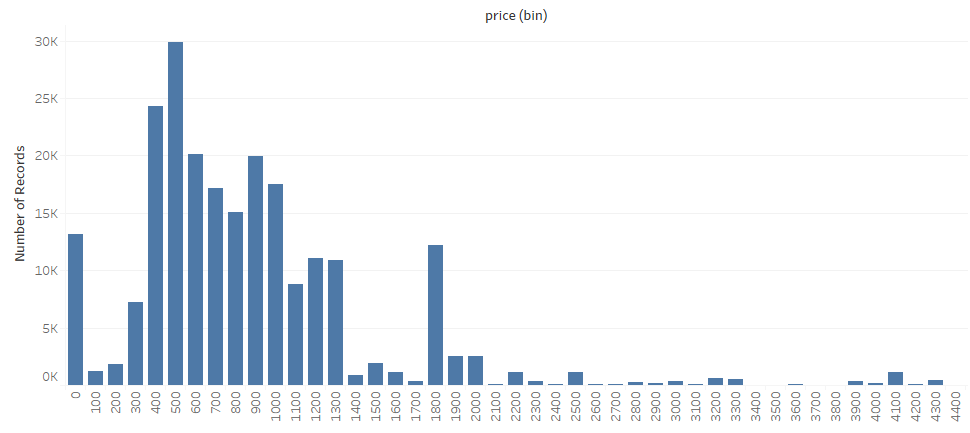
**Data Preparation:**

Step 1: Create a sorted bar chart for each categorical predictor. Create histograms for the numerical columns (in XLMiner: *Explore> Chart Wizard*; or in Tableau: [histogram in Tableau](http://breaking-bi.blogspot.tw/2013/05/creating-histograms-in-tableau.html)).

* Quantity



* Price



1. Fill in the following table:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Column name | Measurement type | If categorical: | | If numerical: |
|  |  | # of categories | % records included in top 2 categories | Is it skewed? |
| id | Categorical | 234,309 | 0.00% |  |
| restaurant\_id | Categorical | 347 | 24.50% |  |
| purchase\_id | Categorical | 216,793 | 0.11% |  |
| type | Categorical | 2 | 100% |  |
| product\_id | Categorical | 1,004 | 7.17% |  |
| name | Categorical | 1,578 | 6.64% |  |
| quantity | Numerical |  |  | Y |
| price | Numerical |  |  | Y |
| cdate | Numerical |  |  |  |

Keep only the records for the “most popular” restaurant. Save into a new file.

[Tableau: On bar chart put mouse over bar, View data… click Underlying tab, Export All]

1. How many records does your new file have? **37,001**

Step 2: Open the new file in Excel. Delete all columns that cannot be used in the model.

Create three derived variables:

* day-of-week of the transaction date (Excel function =WEEKDAY)
* Logarithm (=LN) of Quantity
* Dummy variables for type and for day-of-week (Use Excel’s =IF or in XLMiner: *Transform > Transform Categorical Data > Create Dummies*)

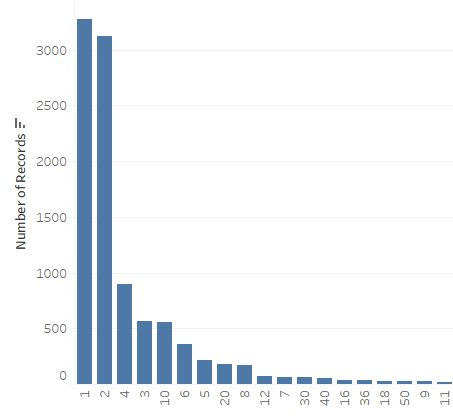
Step 3: Handle missing values for categorical columns by creating a new category NA, and for numerical columns remove rows if there are only a few. Use Excel’s filters and sorting, or in XLMiner: *Transform > Missing Data Handling*

Step 4: Partition the dataset into training, validation, and test sets using XLMiner: *Partition* in Data Mining area. Include only columns price, quantity, LN(quantity), type dummies, and day of week dummies. XLMiner has a limit of 10,000 rows in training and 10,000 in validation.

* Set Training size to 10,000, Validation to 10,000, and the rest goes to Test.
* Use **randomized partition with seed = ddmm** of your birthday**.**

1. Using the information in the training set, what is a *naive prediction* for Quantity in a future purchase? Why?

**1, it has the highest number of records**



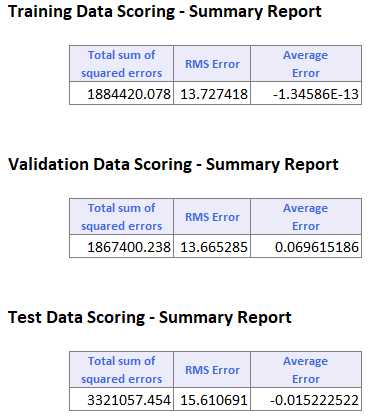
**Model Building and Evaluation:**

**Model 1:** Fit a linear regression model to Quantity as the outcome and all other predictors. XLMiner: *Predict > Multiple Linear Regression*. Make sure not to include LN(Quantity) or too many dummy variables.

* Check the option *Unstandardized Residuals* to get the model residuals.
* Check *Detailed Report* for the test set to get test set predictions.
* Click *Variable Selection* and choose *Stepwise Selection* to obtain variable selection results.

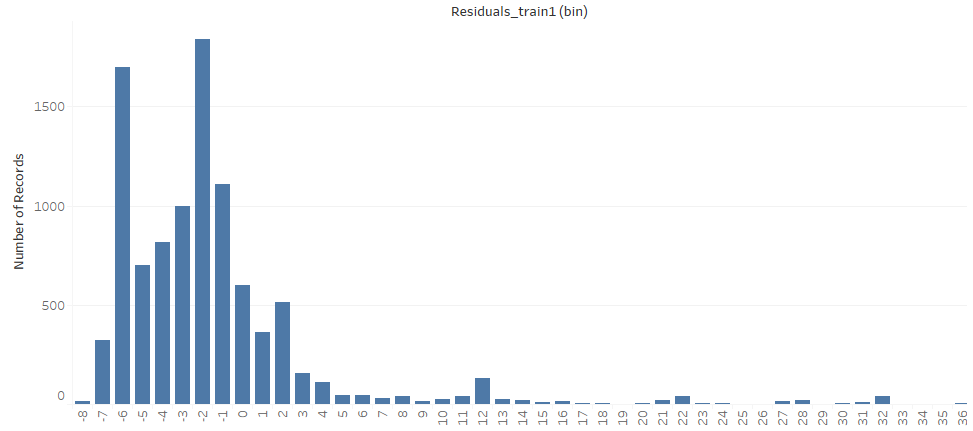
**(d)** From the output, copy the Summary Reports for the training, validation and test sets into your assignment. How does the performance compare across the training, validation and test sets? Is there an indication of over-fitting? Interpret the meaning of the Average Error in the test set.

**Training data scoring is to score how well the model fits the data while validation data scoring is to score the prediction accuracy. Since only the average error is drastically different, there may be an outlier. It is not overfitting**



**(e)** Create a histogram of the model training residuals. Copy the histogram to your assignment. What does this chart tell us about potential prediction errors? How can we improve it? (Recall the Regression for Explanation session).

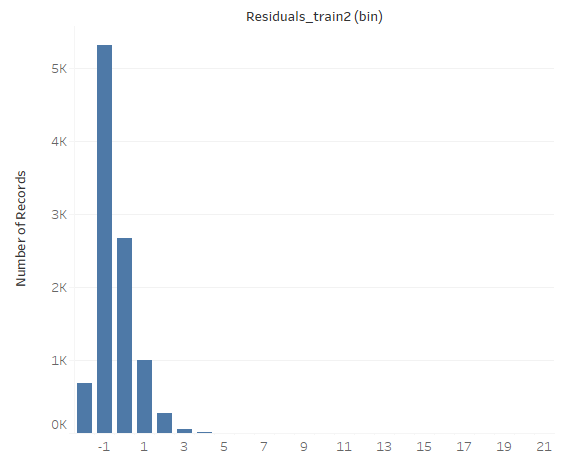
**The histogram of residuals are skewed which violates the assumption of regression. We can do some transformation on the variables such as LN(quantity).**



**Model 2:** Fit another linear regression model, this time with outcome LN(Quantity) and the same predictors from Model 1. Again, check *Unstandardized Residuals* and *Detailed Report* for test.

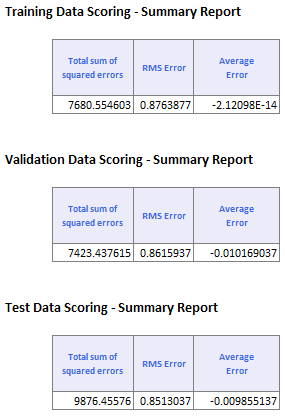
**(f)** Include the histogram of Model 2 training residuals in your report. How does this histogram compare to Model 1 histogram? Which model appears ‘better’?

**The histogram of Model 2 training residuals is not skewed which does not violate the assumption of regression so model 2 appears better.**

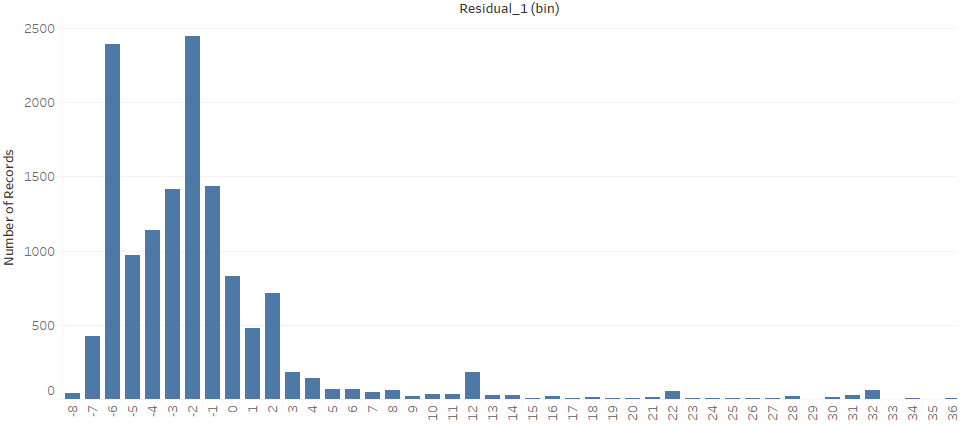


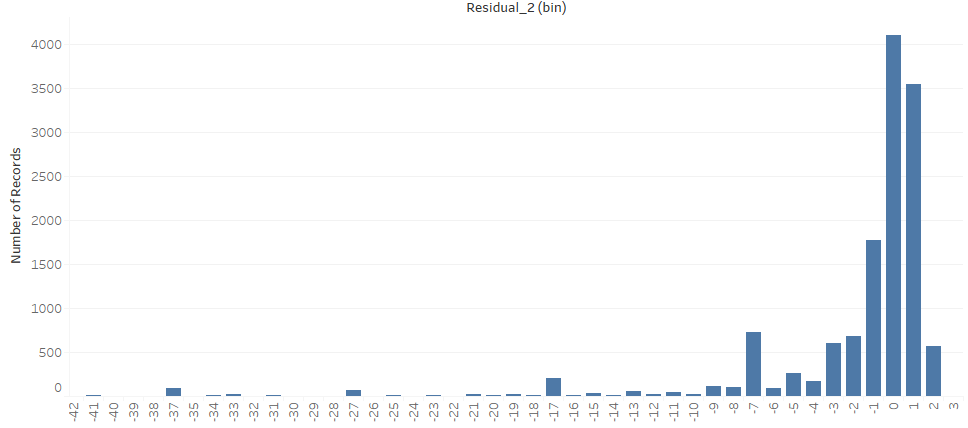
**(g)** From Model 2 output, copy the Summary Reports for the training, validation and test sets. Can these be compared to the Model 1 output? If so, which model has better predictive power?

**Since residuals of validation set in model 2 are smaller than the ones in model 1 so model 2 has better predictive power.**



**(h)** Compare the test set prediction errors of Models 1 and 2 by comparing the two histograms. **Note:** To obtain prediction errors for Model 2, you must first *transform XLMiner’s predictions into the original scale*. Then subtract these numbers from the actual values to obtain prediction errors. Copy the two histograms here.





**(i)** Looking at the histograms from (h), which model is better for predicting Quantity? Why?

**Variable Selection**

Examine the Stepwise results for Model 1.

**(j)** If we use only two predictors, which pair is best in terms of predictive power? If we are not limited to two predictors, which subset of predictors looks most promising in terms of predictive power?

**Price, and if the payment is prepay.**

**Price, if the payment is prepay, and if the purchasing day is Tue. because this subset has the highest adjusted R2**

