Rameez Raja

Linear Regression - Predicting Airfares on New Routes

1. **Perform Typecasting. Which variables are categorical?**

Categorical Variables:

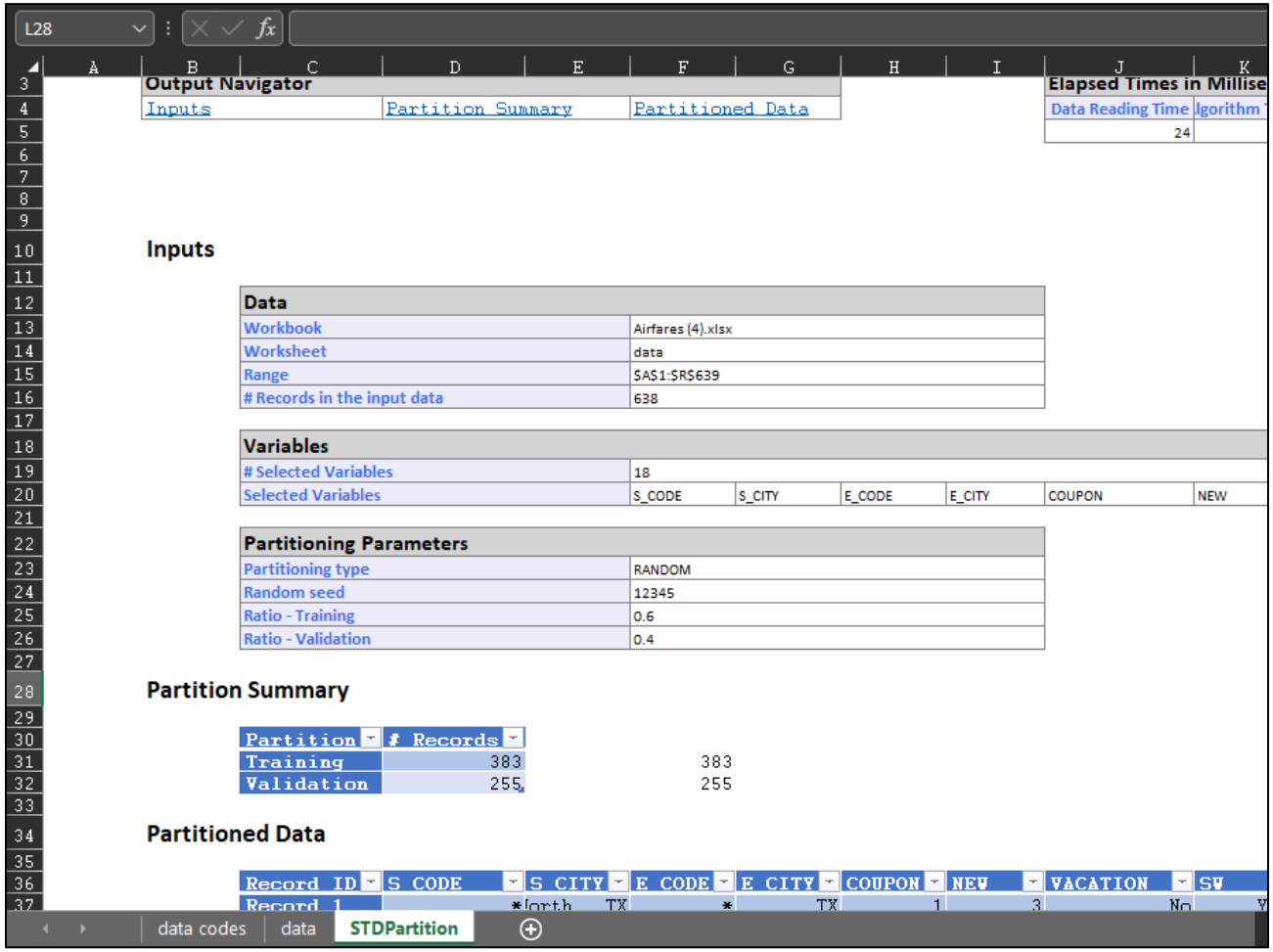
- S_CODE: starting airport's code
- S_CITY: starting city
- E_CODE: ending airport's code
- E_CITY: ending city
- VACATION: whether a vacation route (Yes) or not (No)
- SW: whether Southwest Airlines serves that route (Yes) or not (No)
- SLOT: whether either endpoint airport is slot controlled or not; this is a measure of airport congestion
- GATE: whether either endpoint airport has gate constraints or not; this is another measure of airport congestion

Numerical Variables:

- COUPON: average number of coupons
- NEW: number of new carriers
- HI: Herfindel Index – measure of market concentration
- S_INCOME: starting city's average personal income
- E_INCOME: ending city's average personal income
- S_POP: starting city's population
- E_POP: ending city's population
- DISTANCE: distance between two endpoint airports in miles
- PAX: number of passengers on that route during the period of data collection
- FARE: average fare on that route

2. **Partition the data into training (60%) and validation sets (40%). The model will be fit to the training data and evaluated on the validation set.**

When we partition the data using standard partition and the 60/40 split, we get the following:



Output Navigator

Inputs | Partition Summary | Partitioned Data

Elapsed Times in Millise

Data Reading Time	lgorithm
24	

Inputs

Data	
Workbook	Airfares (4).xlsx
Worksheet	data
Range	$A$1:$R$639
# Records in the input data	638

Variables						
# Selected Variables	18					
Selected Variables	S_CODE	S_CITY	E_CODE	E_CITY	COUPON	NEW

Partitioning Parameters	
Partitioning type	RANDOM
Random seed	12345
Ratio - Training	0.6
Ratio - Validation	0.4

Partition Summary

Partition	# Records	
Training	383	383
Validation	255	255

Partitioned Data

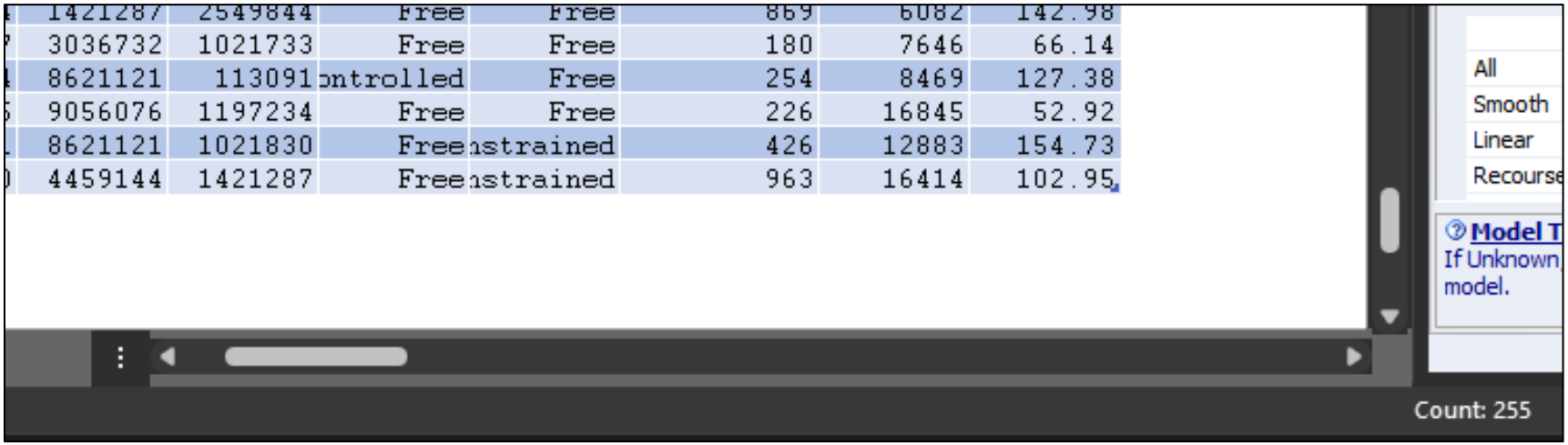
Record ID	S_CODE	S_CITY	E_CODE	E_CITY	COUPON	NEW	VACATION	SU
Record 1		*lorth	TX	*	TX	1	3	No

data codes | data | STDPartition

3. **What is the Fare value for the last three records included in validation set?**

Unfortunately, when the partition is run it does not order the records in order, so as per the example in class the first 383 records (60%) belong to the training and the rest of the 255 (40%) records belong to the validation set. Therefore, the fare value for the last three records included in the validation set are:

- 52.92
- 154.73
- 102.95



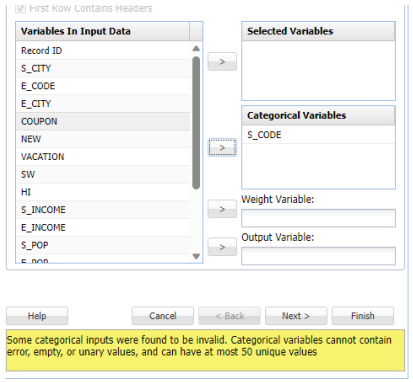
4. **Try to add all categorical variables to create a linear regression model. Which variable cannot be added (not allowed by the software) and why is that?**

Analytical solver we have two types:

- Classify: target variable is a **categorical variable**
- Predict: price, numerical variable

Steps: We click on data mining and then click on classify.

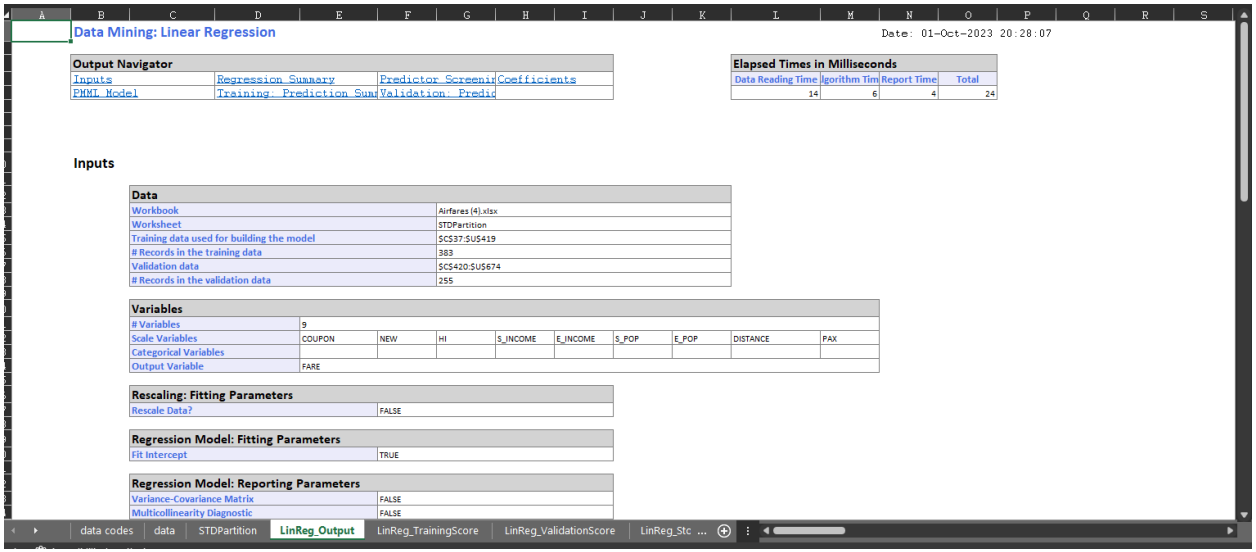
We then select linear regression and proceed to select the categorical values. However, when S_CITY is selected it doesn't work - we get the below error, because there are blanks in the column for these.



The following categorical variables are not added:

- S_CITY
- E_CITY

5. **Create a linear regression model only based numerical variables to predict the Fare variable. For scoring Training and Validation data, have detailed and summary reports created. What is RMSE for the validation?**



Data Mining: Linear Regression

Date: 01-Oct-2023 20:28:07

Output Navigator

Inputs	Regression_Summary	Predictor_Screeni	Coefficients
PMML Model	Training: Prediction_Sum	Validation: Predic	

Elapsed Times in Milliseconds

Data Reading Time	lgorithm Tim	Report Time	Total
14	6	4	24

Inputs

Data	
Workbook	Airfares (4).xlsx
Worksheet	STDPartition
Training data used for building the model	$C$37:$U$419
# Records in the training data	383
Validation data	$C$420:$U$674
# Records in the validation data	255

Variables									
# Variables	9								
Scale Variables	COUPON	NEW	HI	S_INCOME	E_INCOME	S_POP	E_POP	DISTANCE	PAX
Categorical Variables									
Output Variable	FARE								

Rescaling: Fitting Parameters	
Rescale Data?	FALSE

Regression Model: Fitting Parameters	
Fit Intercept	TRUE

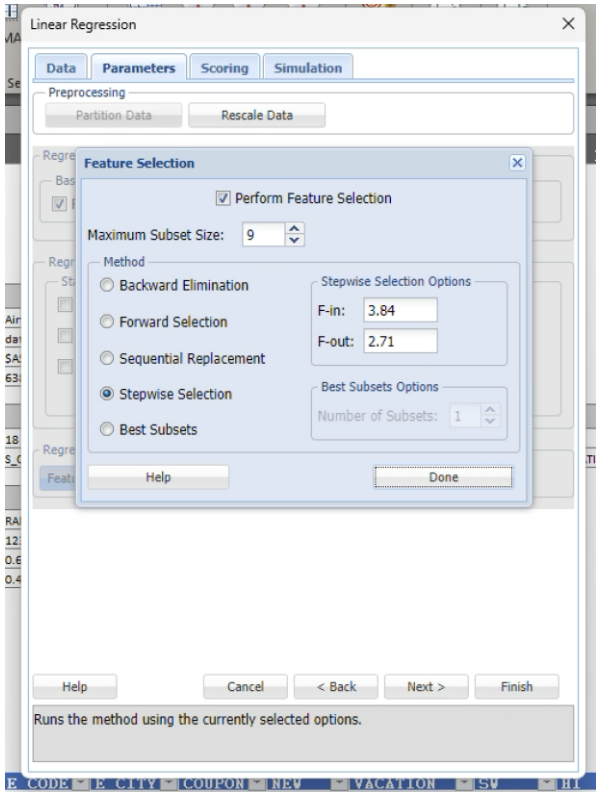
Regression Model: Reporting Parameters	
Variance-Covariance Matrix	FALSE
Multicollinearity Diagnostic	FALSE

data codes | data | STDPartition | LinReg_Output | LinReg_TrainingScore | LinReg_ValidationScore | LinReg_Stc ...

Validation: Prediction Summary

Metric	Value
SSE	487875.6
MSE	1913.238
RMSE	43.74057
MAD	35.29559
R2	0.665976

6. **Redo step 5 and this time use stepwise variable selection feature to reduce the number of predictors. How many subsets are created, and which variable(s) are omitted at the end?**



See excel spreadsheet attached to assignment.