Objective: A

1. Model Building Process

Data Preprocessing

1.1Data Understanding

- Read the Excel sheet called “data codes” to understand the dataset.
- Make assumption which columns are not needed, and which columns are relevant with each other. We find S_CODE, S_CITY, E_CODE, E_CITY are not needed to build model, DISTANCE and FARE are strongly related, and it looks like a linear relationship.

1.2 Data Cleaning:

- Find outliers in the dataset using 3-sigma rule. Create a col fill formula with first cell.
- Outlier Formula: IF(OR(mean - 3*stdev; mean+3*stdev),"outlier", "ok") .
- The value of mean stands for average value of selected column, and it is calculated by function average(). The value of stdev stands for standard deviation, and it is calculated by function stdev.P().
- Replace values of outliers with mean value.
- Find missing data using Excel Conditional Formatting, in the Excel click Home-Conditional Formatting-Highlight Cell Rules-Equals. Change value to Blanks. But there is no empty value in this dataset.

1.3 Data Type Conversion

- In the dataset, there are 4 columns that need to be binarized including VACATION, SW, SLOT, GATE.
  a) In terms of VACATION, values of “NO” are binarized to 0, value of “Yes” to 1.
  b) In terms of SW, values of “NO” are binarized to 0, value of “Yes” to 1.
  c) In terms of Slot, values of “Controlled” are binarized to 0, value of “Free” to 1.
  d) In terms of Slot, values of Free are “Constrained” are binarized to 0, value of “Free” to 1.

1.4 Feature Selection

- At this stage, we can make sure that some features are not necessary for us to build model, and these features should be removed, such as S_CODE, S_CITY, E_CODE, E_CITY.

1.5 Scale Data

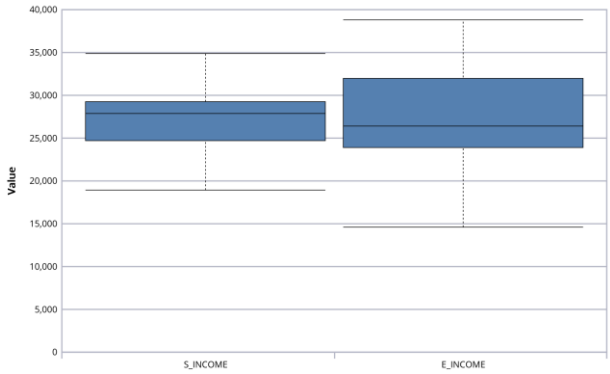

Figure 1. Income Box Plot

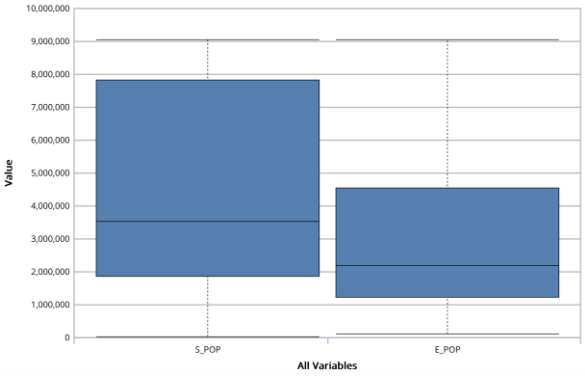

Figure 2. Population Box Plot

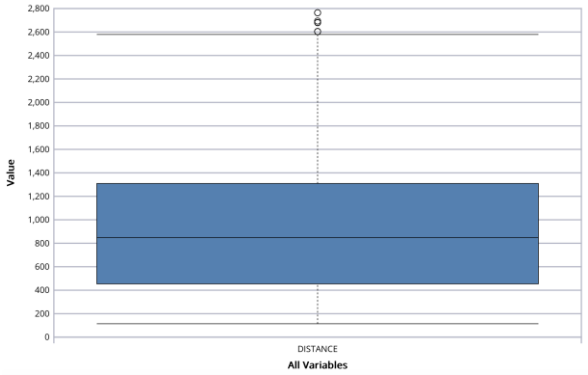

Figure 3. Distance Box Plot

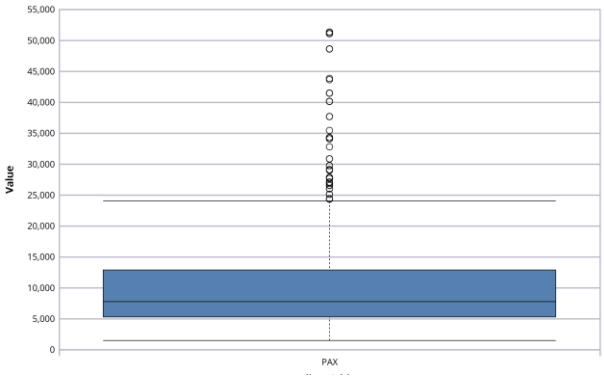

Figure 4. Distance Box Plot

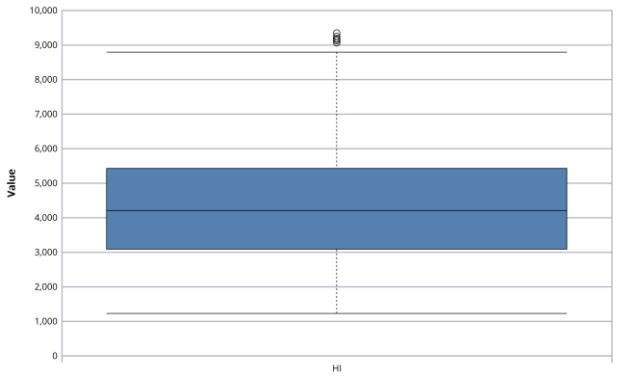

Figure 5. Distance Box Plot

By above box plot chart, the number value of income, population, distance, pax and hi are very huge and the range is big. To improve performance and accuracy, these values should be normalization.

By using transform function of XLMiner, clicking Data Science Tab -Transform-Rescale Continuous Data, and then select HI, S_INCOME, E_INCOME, S_POP, E_POP, DISTANCE, PAX, Change to Parameters Tab and choose normalization radio button.

After scale some feature, we can get some new value rang from 0 to 1. The numbers below are some of normalized values

HI	S_INCOME	E_INCOME	S_POP	E_POP	DISTANCE	PAX
0.50020767	0.60851557	0.26894662	0.33312815	0.01050565	0.07472019	0.127572661
0.5158696	0.50542421	0.62933124	0.38807076	0.78643691	0.17434208	0.146748647
0.97969704	0.70176183	0.62933124	0.63785765	0.78643691	0.09434268	0.09924997
0.17573121	0.64758243	0.62933124	0.86420212	0.78643691	0.18792688	0.474184629
0.17573121	0.64758243	0.62933124	0.86420212	0.78643691	0.18792688	0.474184629
0.26819241	0.44604008	0.62933124	0.24385763	0.78643691	0.07358812	0.238336046
0.68032453	0.60851557	0.62933124	0.33312815	0.78643691	0.41735911	0.062602976

Figure 6. Normalized Values

1.6 Partition Data

- Using partition function of XLMiner to split whole dataset into training set (60%) and test set (40%). Based on the 1.4 feature selection section, these features (S_CODE, S_CITY, E_CODE, E_CITY.) are not need. Additionally, we create new columns (VACATION_DUMMY_CODE, SW_DUMMY_CODE, SLOT_DUMMY_CODE, GATE_DUMMY_CODE) for the category features (VACATION, SW, SLOT, GATE). So in the partition process, these features (S_CODE, S_CITY, E_CODE, E_CITY, VACATION, SW, SLOT, GATE) are not selected for partition.

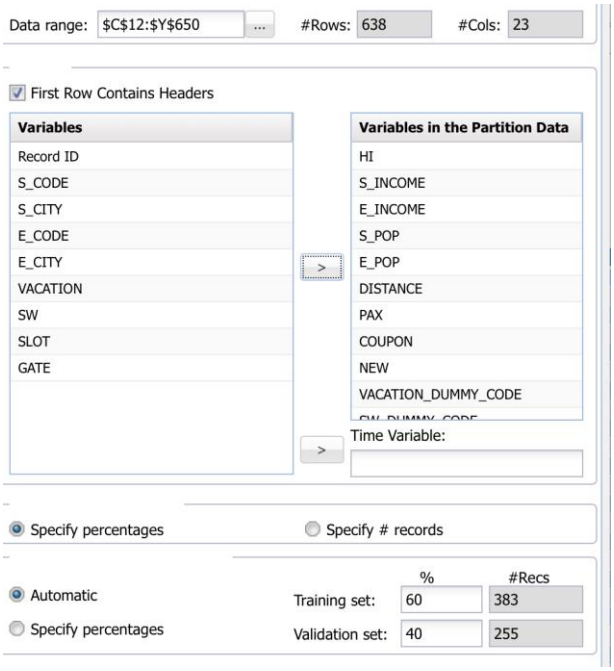

Figure 7. Partition Values

1.7 Build Model

We choose to build a Linear Regression model to predict the fare, click Data Science – Predict- Linear Regression and then select FARE as output variable, others go into selected variables.

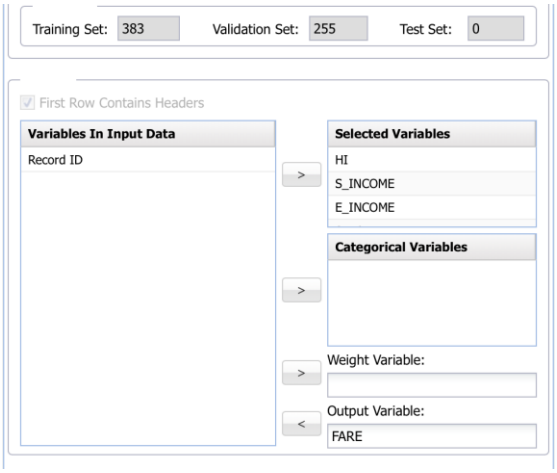

Figure 8. Select Variables

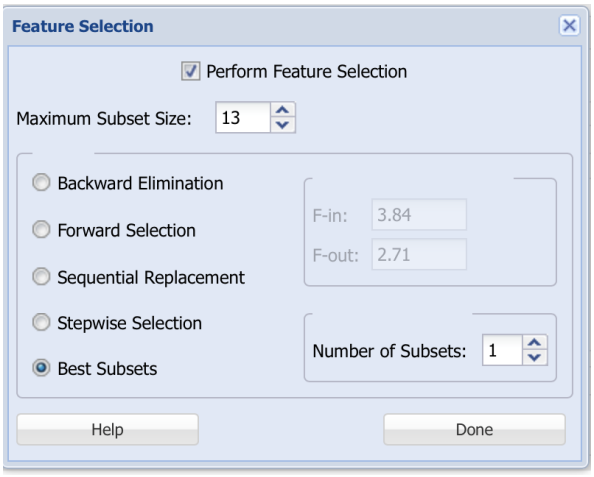

Figure 9. Feature Selection

After select variables, change to Parameters Tab. We already have scaled data, so we don't need to use Rescale Data function. The only thing that we must do is to use Feature Selection function and choose Best Subsets radio button. Lastly, change to Scoring Tab and click Detailed Report and Summary Report.

The first model contains all selected features, and we don't make sure it is the best model.

Metric	Value
SSE	322123.518
MSE	1263.22948
RMSE	35.5419398
MAD	28.6428091
R2	0.77945823

Figure 10. Model with all selected variables (Subset 14)

So, we build new model using Subsets and compare reports with each report. We find that the model built by subset 12 has better metric value. By comparing the key metrics, it has the lowest SSE and RMSE values and the R^2 value is also slightly higher than others. In our Excel file the model built by subset 12 is stored in LinReg_Stored1.

Subset 12

Metric	Value
SSE	320954.63
MSE	1258.6456
RMSE	35.477396
MAD	27.893329
R2	0.7802585

Subset 9

Metric	Value
SSE	334489.45
MSE	1311.7233
RMSE	36.217721
MAD	28.497948
R2	0.7709919

Subset 8

Metric	Value
SSE	344973.79
MSE	1352.8384
RMSE	36.780951
MAD	29.288507
R2	0.7638138

Subset 6

Metric	Value
SSE	360174.0036
MSE	1412.447073
RMSE	37.58253681
MAD	29.66366936
R2	0.75340698

Figure 11. Summary Reports

Variables selected by Subset 12:

# Variables	11										
Scale Variables	HI	E_INCOME	S_POP	E_POP	DISTANCE	PAX	COUPON	VACATION_DUMMY_CODE	SW_DUMMY_CODE	SLOT_DUMMY_CODE	GATE_DUMMY_CODE

Variables selected by Subset 9:

# Variables	8							
Scale Variables	HI	S_POP	E_POP	DISTANCE	PAX	VACATION_DUMMY_CODE	SW_DUMMY_CODE	SLOT_DUMMY_CODE

Variables selected by Subset 8:

# Variables	7						
Scale Variables	HI	S_POP	E_POP	DISTANCE	COUPON	VACATION_DUMMY_CODE	SW_DUMMY_CODE

Variables selected by Subset 6:

# Variables	5				
Scale Variables	HI	DISTANCE	VACATION_DUMMY_CODE	SW_DUMMY_CODE	SLOT_DUMMY_CODE

To summarize, although the model in Subset 12 contains several variables, the necessary core variables are retained, but too many variables are screened to avoid the problem of overfitting the model conclusions due to too many variables, which leads to high model complexity and increases the computational burden of the model. The models in Subset 12 achieve the best balance of model performance, simplicity and practicality of application.

Validation against overfitting

The model of Subset 12 is constructed by separating the training set and test set, and by comparing the performance metrics of different subsets through cross-validation, regularization techniques, feature selection, control of model complexity, and residual analysis, we validate that the model has a stable performance without obvious overfitting problem. Meanwhile, the lower MSE and RMSE also indicate that the model does not have obvious overfitting problem, which makes the subset 12 model the best choice for fare prediction of new routes.

Practicality and Explanation

Compared with the other subset models, the Subset 12 model has better performance and retains the key explanatory variables, which can provide valuable support for the actual route fare setting. The Subset 12 model can be better used to forecast fares in real time, helping airlines to adjust prices in response to changes in market demand.

2. Deploy and Score Linear Regression Model

2.1 Test data item process

Predict for required data item:

COUPON	NEW	VACATION_D	SW_DUMMY	HI	S_INCOME	E_INCOME	S_POP	E_POP	SLOT_DUMM	GATE_DUMM	DISTANCE	PAX
1.202	3	0	0	4442.141	28760	27664	4557004	3195503	1	1	1976	12782

Figure 12. Test Data Item

Since in the dataset these features (HI, S_INCOME, E_INCOME, S_POP, E_POP, DISTANCE, PAX) have been normalized, so the test data item also needs to be normalized from 0 to 1 using the formular below. The Max and Min values are found by excel formular function MAX() and MIN().Finally, we get the normalized data and use the best model built by subset 12 to predict fare.

$$x_i^{new} = \frac{(x_i - x_{min})}{(x_{max} - x_{min})}$$

Figure 13. Normalization Formular

COUPON	NEW	VACATION_D	SW_DUMMY	HI	S_INCOME	E_INCOME	S_POP	E_POP	SLOT_DUMM	GATE_DUMM	DISTANCE	PAX
1.202	3	0	0	4442.141	28760	27664	4557004	3195503	1	1	1976	12782
1.202	3	0	0	0.395583867	0.616228758	0.539544873	0.501556241	0.344772348	1	1	0.702641509	0.226220564
				Max								
				9350	34880	38813	9056076	9056076			2764	51358
				Min								
				1230	18933	14600	29838	111745			114	1504

Figure 14. Test Data Item after Normalized

2.2 Predict

The value of perdition fare is 237.07.

Record ID	Prediction: FARE
Record 1	237.0785936

Figure 15. Predict Result

2.3 Predict the reduction in average fare on the above route if Southwest Airlines decides to cover this route.

If Southwest Airlines decides to cover this route, the value of SW must change to 1.

COUPON	NEW	VACATION_D	SW_DUMMY	HI	S_INCOME	E_INCOME	S_POP	E_POP	SLOT_DUMM	GATE_DUMM	DISTANCE	PAX
1.202	3	0	1	0.39558387	0.61622876	0.53954487	0.50155624	0.34477235	1	1	0.70264151	0.22622056

The value of perdition fare is 202.09. So, the reduction in the average fare can be 34.98.

Record ID	Prediction: FARE
Record 1	202.0933788

Figure 16. Predict Result with SW =1

3.Comparation between Liner Regression, KNN and Regression Tree Model

3.1 Data Understanding

By reading the assignment documentation uploaded by Canvas, we understood the content of the data represented by each heading. At the same time we began to hypothesise which of the data given would be useful for modelling and which would not be very useful for modelling, based on the data available in real life and comparing it to the documented data we concluded that data such as DISTANCE, VACATION etc. are positively correlated with ticket prices.

3.2 Data Cleaning

Using 3-sigma for querying the outliers present in the data is different from part 1, in this part instead of replacing the mean of the selected columns with the value of MEAN, we directly remove it and proceed to the next steps. Replacement is done using Transform and Partition is used to split the data in different allocation ratios.

3.3 Data Type Conversion

Use Categorical Data Create Dummies to replace 'Yes/No' with '0/1' in the VACATION and SW columns, and 'Free' in the SLOT and GATE columns. 'Free' in the SLOT and GATE columns are replaced with "1" and vice versa.

For 'Liner Regression', 'K-Nearest Neighbours' and 'Regression Tree', the data used in the columns of SLOT and GATE are replaced with '1', and vice versa with '0'. The data used in 'Liner Regression', 'K-Nearest Neighbours' and 'Regression Tree' are split according to the ratio of 60/40; 70/30 to find the optimal result.

3.4 Feature Selection

After the assumptions were made, we decided to delete S_CODE and E_CODE because there are too many invalid values '*' and they represent the same meaning as CITY. The rest remained in the document.

3.5 Scaling Data

Instead of doing a STATISTICAL RULE OF THUMB directly with prices, we chose to do a STATISTICAL RULE OF THUMB on price/mile, we calculate that first and in that way we reckon preserved more reasonable data.

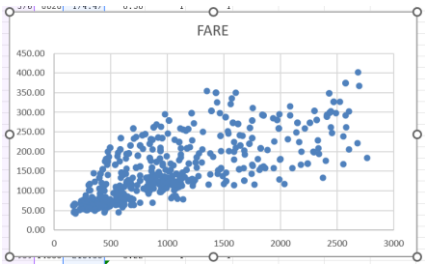

Figure 17. Scatterplot (linear)

Variables																			
# Selected Variables	19																		
Selected Variables	Record ID	COUPON	NEW	HI	S_INCOME	E_INCOME	S_POP	E_POP	DISTANCE	PAX	FARE	VACATIO	VACATION	SW_No	SW_Yes	SLOT_Controlled	SLOT_Free	GATE_Co	GATE_Free

Figure 18. Selected Variables

3.6 Data differentiation

Using XLMiner, the entire dataset was differentiated between the training and test sets by selecting Specify percentages, the first time we tried 60% (training set) and 40% (test set).

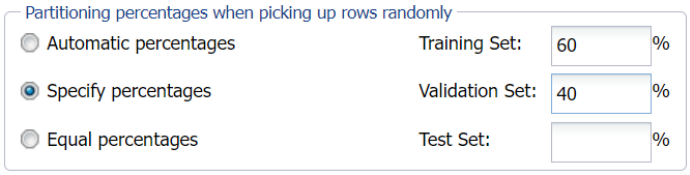

Figure 19.. 70% (training set) and 30% (test set)

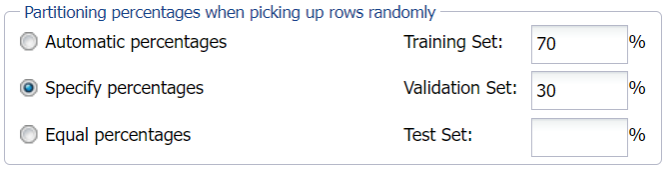

Figure 20. 70% (training set) and 30% (test set)

After trying 60% (training set) and 40% (test set) we proceeded to try 70% (training set) and 30% (test set). And compare which allocation gives the best results.

3.7 Constructing the model

We first choose to use 70% (training set) and 30% (test set) of the test method.

The Regression Tree model is used to predict the fare. The specific process is as follows 'Data Science→Predict→K-Nearest Neighbours', then select 'Fare' as the output variable, except for the' Record ID' column, the rest of the partition to the right side of the "Selected Variables", click "Next", in the Decision Model, select Prune. Select Prune and choose Fully Grown in Tree for Scoring (because it is the most general choice). Click 'Show Feather Importance' and select 'Fully Grown, Best Pruned and Minimum Specified' in 'Tree to Display'. Minimum Specified', complete the above steps and click "Next", select the "Detailed Report" for the training set and test set respectively, and select "Finish". And then Finish.

Metric	Value
SSE	253358
MSE	1473.012
RMSE	38.37983
MAD	27.31667
R2	0.761708

Figure 21. Regression Tree

The data shown in the report matched our hypothesis, but in order to analyse the data more comprehensively, we carried out a second construction of a different model to find the optimal solution.

We decided to use 'KNN' for the second construction, and similar to the Regression Tree construction, we got the data generated by KNN and found that the data generated by the KNN model was far from the optimal solution, so we gave up and chose 'Liner Regression'. Liner Regression'.

Metric	Value
SSE	620814.4
MSE	3609.386
RMSE	60.07817
MAD	36.09777
R2	0.416103

Figure 22. KNN

Metric	Value
SSE	322123.518
MSE	1263.22948
RMSE	35.5419398
MAD	28.6428091
R2	0.77945823

Figure 23. Liner Regression

Next, we used 60% (training set) and 40% (test set) testing method.

In 60% (training set) and 40% (test set) test method we found that in this ratio Regression Tree is the best model. But after comparing SEE, MSE, RMSE, MAD and R^2 we decided to select Liner Regression as the best model.

Metric	Value
SSE	187789.3
MSE	545.8991
RMSE	23.36448
MAD	15.98919
R2	0.912629

Figure 24. Regression Tree

Metric	Value
SSE	834797
MSE	3629.552
RMSE	60.24576
MAD	37.7771
R2	0.379682

Figure 25. KNN

Objective: B

1.In reality, which of the factors (predictor variables) will not be available for predicting the average fare from a new airport?

Not Available variables:

PAX: Before the operation of the new route, there are no passengers.

New, **SW, HI, COUPON**: This is a new airport, there are no carries before.

SLOT, GATE: Information on slot control or gate constraints may not be readily available for new airports until the airport gets operational assessments.

Assumptions

Positive correlation between DISTANCE and FARES

The longer the distance, the higher the fares usually are. This is because longer routes incur higher operating costs, such as fuel costs, crew member's salaries, and so on. Airlines will adjust their fares appropriately according to the distance in order to cover costs and maintain profits. Therefore, it can be surmised that there is a positive correlation between route distance and fare.

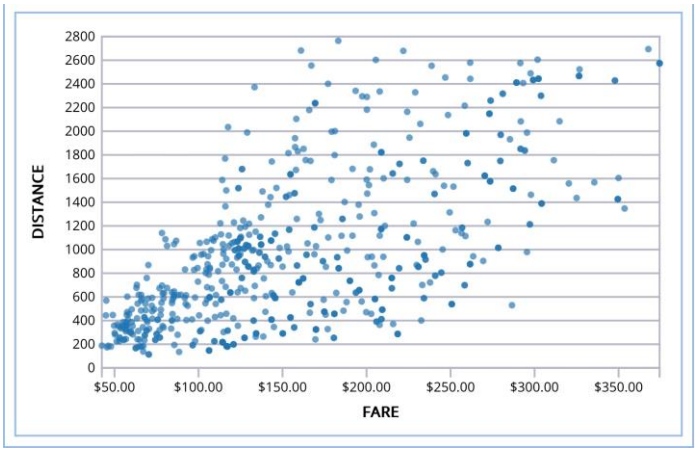

Figure 26. DISTANCE and FARES

Relationship between VACATION and FARE: In the distribution of fares for holiday routes (1) and non-holiday routes (0), the fares for holiday routes are generally high, and it is concluded that holiday routes are a positively correlated factor affecting fares. As shown in the hypothesis, the demand for holiday routes (1) is much higher than that for non-holiday routes (0), and according to the relationship between supply and demand, increasing fares within a reasonable range will not reduce consumer demand.

DISTANCE, S_POP, E_POP, SW and VACATION have the most significant impact on ticket prices. Among them, the effects of DISTANCE and SW are particularly significant, and the scatter plots explain the main trends in fares. Other factors such as HI and SLOT may need to be analyzed in conjunction with more variable factors.

2.Based on the settings and findings of the model from item A, build another model using the available (in your opinion) variables only.

Based on Excel file of Model A, remove all report sheets and remove STDPartition Sheet. Back to Rescaling sheet, click Data Science – Partition, only select(S_INCOME,E_INCOME,S_POP,E_POP,DISTANCE,VACATION_DUMMY_CODE, FARE).After this, we get new STDPartition below.

Record ID	S_INCOME	E_INCOME	S_POP	E_POP	DISTANCE	VACATION_DUMMY_CODE	FARE
Record 1	0.60851557	0.26894662	0.3331282	0.0105056	0.07472019	0	64.11
Record 5	0.64758243	0.62933124	0.8642021	0.7864369	0.18792688	0	85.47
Record 8	0.49031167	0.62933124	0.156271	0.7864369	0.30452978	1	116.54
Record 11	0.35379714	0.62933124	0.1293336	0.7864369	0.7509415	1	158.2
Record 12	0.36201184	0.62933124	1	0.7864369	0.83584652	0	228.99

Figure 27. STDPartition only available (in your opinion) variables

3.Use this model to predict the average fare using only the available (in your opinion) data from the record in item A.2.

Metric	Value
SSE	571981.98
MSE	2243.0666
RMSE	47.361024
MAD	38.701425
R2	0.6083927

Record ID	Prediction: FARE
Record 1	249.765255

Figure 28. Predict Result

4.Compare performance of this model with performance of the model from item A.

In terms of variable selection: the subset 12 model, by using scatterplot to analyse the variable factors, eliminates irrelevant variables and retains only the core variables that have the most obvious influence on ticket prices, which achieves the effect of optimising the fit of the model. As for the Part B model: although more variables were removed, there was no rigorous analysis on variable selection, and as a result, some variable factors that had a greater impact on fares were mistakenly deleted, which led to a reduction in the performance of the model.

For Model complexity, compared with the subset 12 model, the Part B model is oversimplified and does not include enough variables to explain the reasons for fare changes, resulting in the predictive power of the whole model being much lower than that of the subset 12 model.

In Conclusion,Compared with the part B model, the subset 12 model with lower SSE, MSE and higher R^2 proves that the subset 12 model is better in terms of fitting, error control, practicality and interpretability of the whole model. Therefore, we believe that the linear model of subset 12 is more suitable as the final model for fare prediction, although the model of Part B is not recommended to be used based on the consideration of the performance of the model despite its close prediction value.