MBAS901: Assessment 3

Exploratory and Predictive Analysis

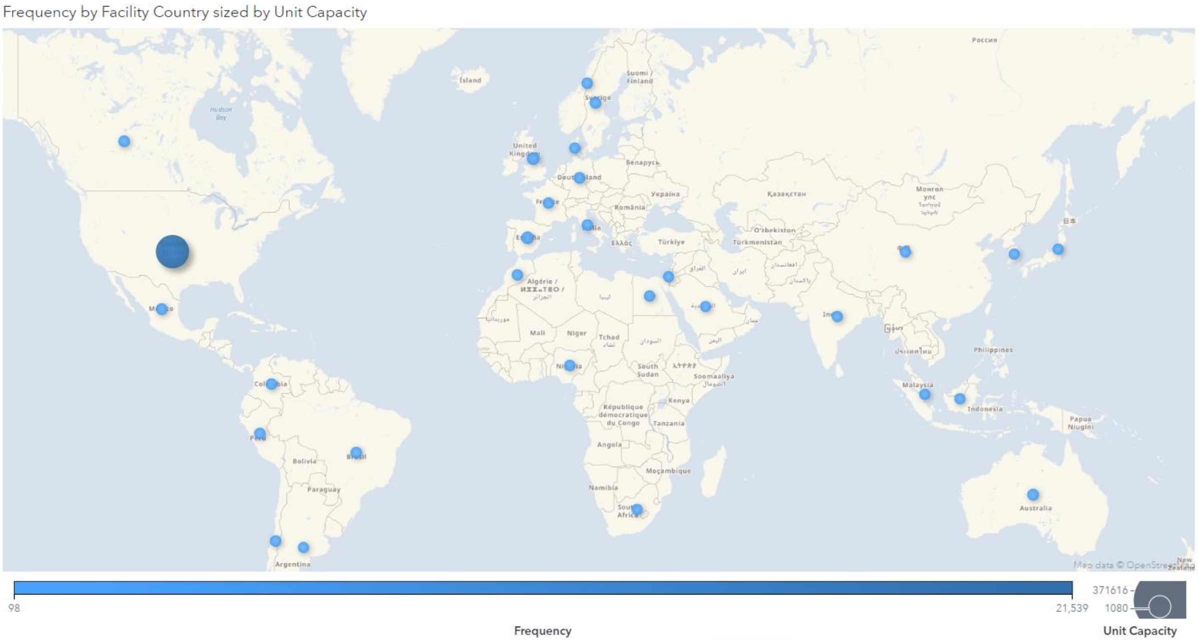
Name: Chinmay Datar

Student No.: 6956361

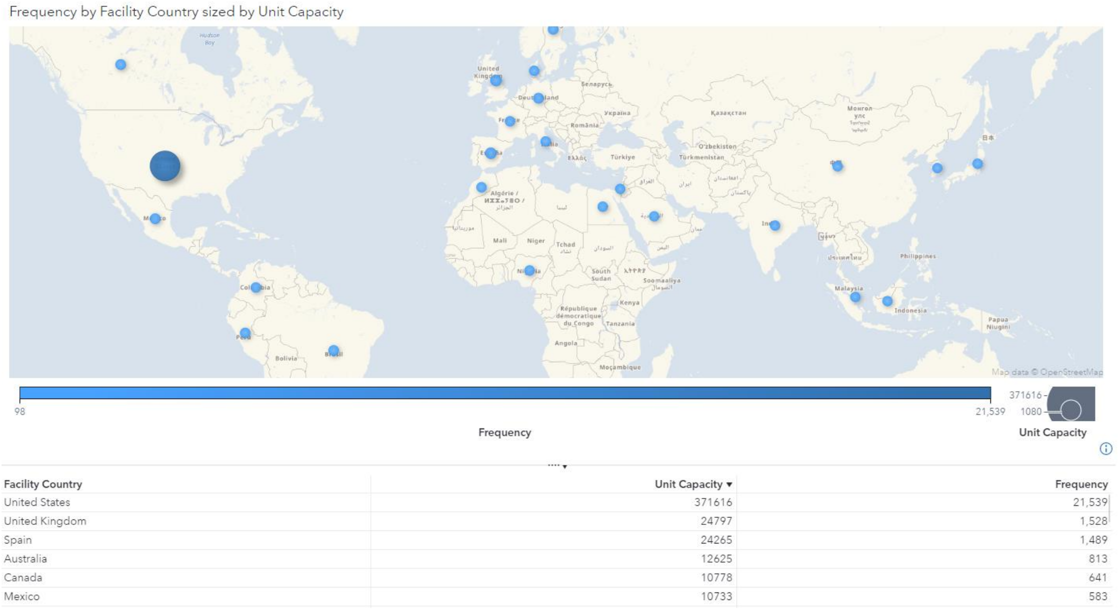
Task 1: Exploratory Data Analysis

Q1. On a geographic map, show the countries where toy facilities are located. Size of the bubble should be the total unit capacity.

A. Geographic map showing countries with toy facilities:



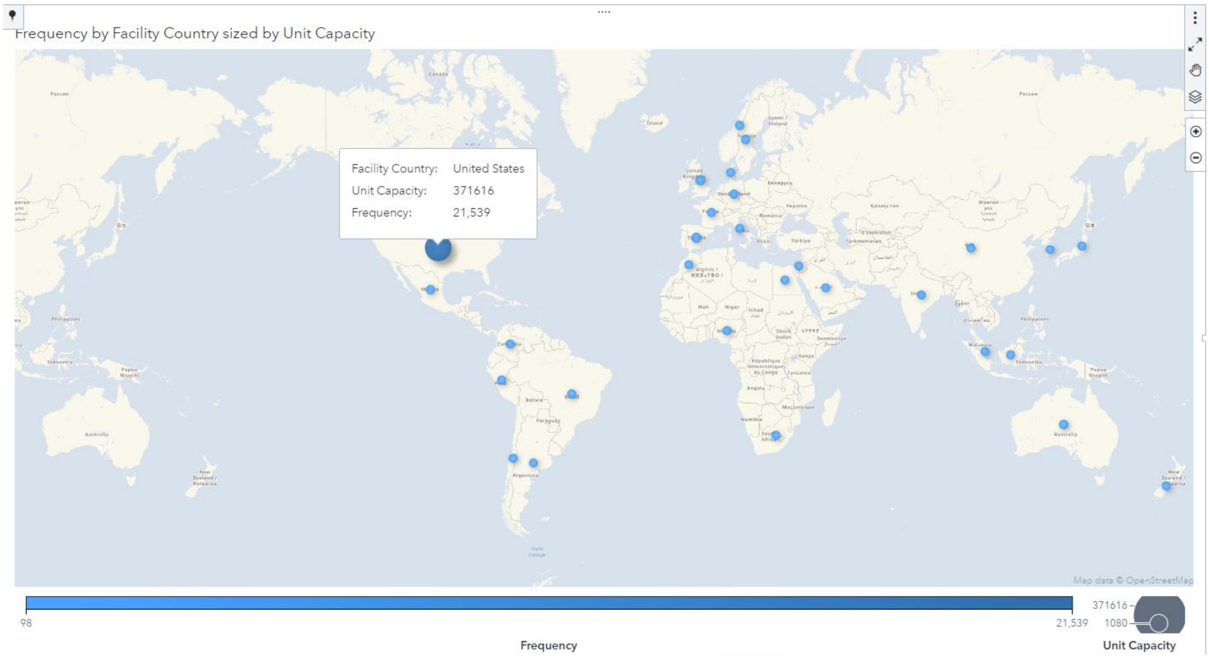
The above geographic map shows the countries where the facilities are located, and the bubble is sized by the unit capacity. It is evident United States has the higher unit capacity than any other country with capacity of 371,616 followed by United Kingdom with capacity of 24,797.



Facility Country	Unit Capacity ▾	Frequency
United States	371616	21,539
United Kingdom	24797	1,528
Spain	24265	1,489
Australia	12625	813
Canada	10778	641
Mexico	10733	583

Q2. What is the total unit capacity in the United States and in Australia?

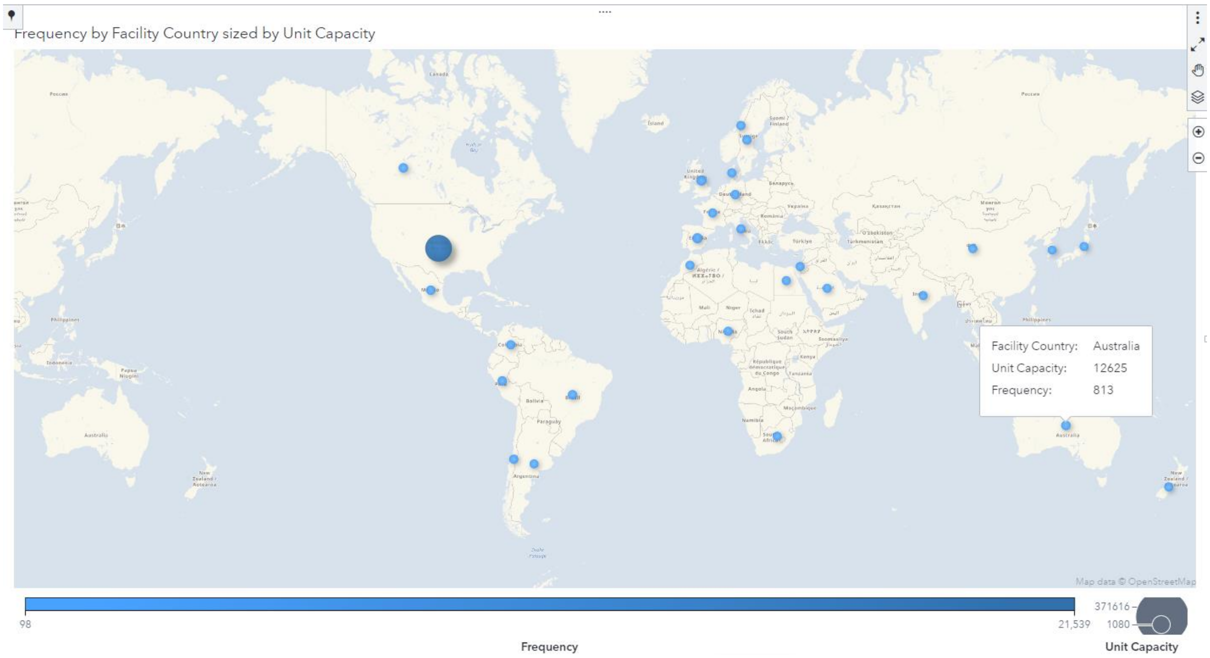
A.



Frequency by Facility Country sized by Unit Capacity

Facility Country: United States
Unit Capacity: 371616
Frequency: 21,539

The unit capacity of United States is 371,616.



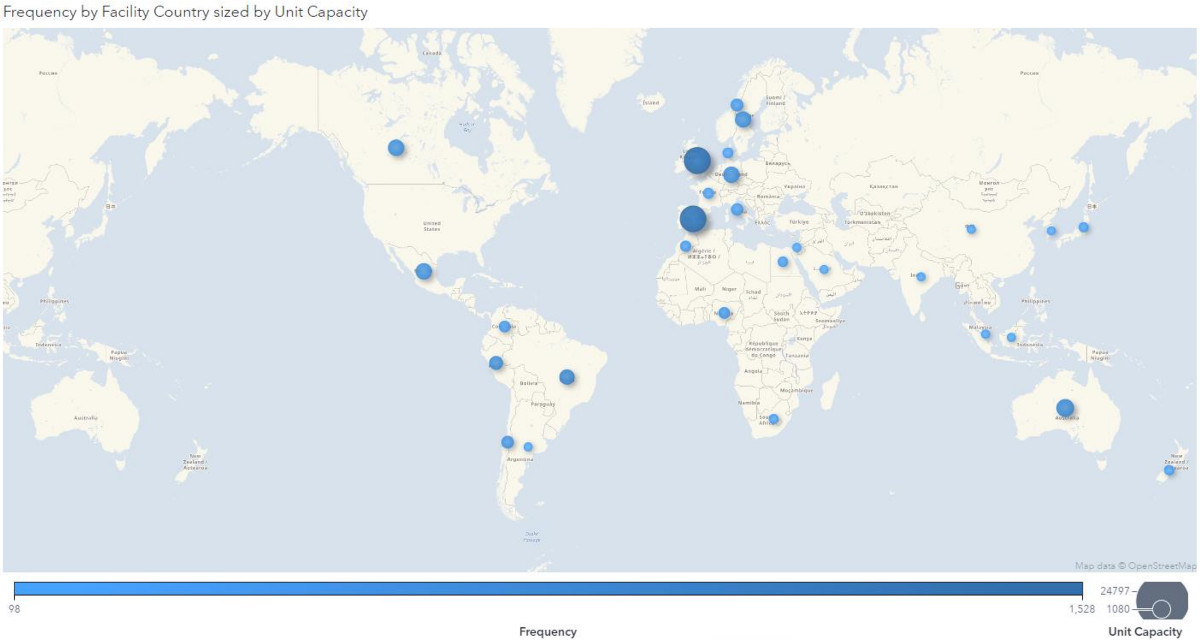
Frequency by Facility Country sized by Unit Capacity

Facility Country: Australia
Unit Capacity: 12625
Frequency: 813

The unit capacity of Australia is 12,625.

Q3. Temporarily remove United States from the map you prepared in Q1. Show the updated map (without US). Which country has the second-largest unit capacity after United States.

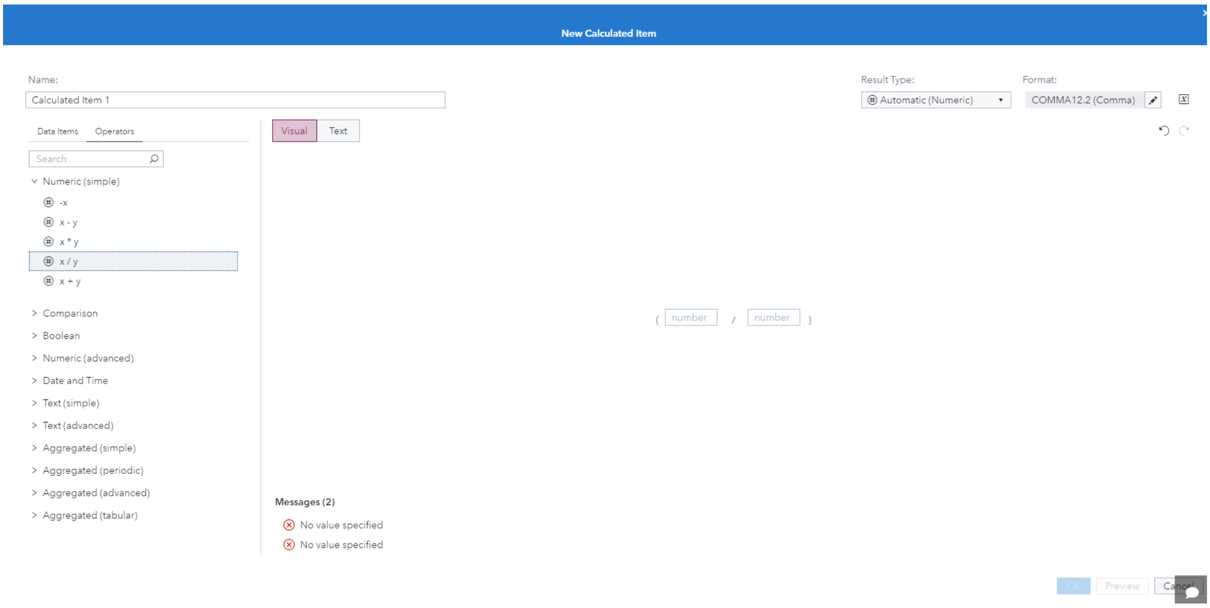
A. Updated map after filtering United States:



The second largest unit capacity after United States is United Kingdom with capacity of 24,797, higher than Spain with capacity of 24,265.

Q4. Many countries have more than one toy facility. Further, most facilities have more than one unit manufacturing toys. As one would expect, majority of these units do not operate at full capacity. Assuming the actual usage of the units is provided by 'Unit Actual' variable and the total unit capacity is provided by 'Unit Capacity' variable, calculate the 'Capacity Utilisation Ratio' and store values in a new variable. Show how you created this calculated item by taking a screenshot of the appropriate SAS Viya window. Generate a histogram of the new variable and copy/export it into your answer script. Interpret the histogram.

A. Steps to calculate Capacity Utilisation Ratio:



New Calculated Item

Name: Calculated Item 1

Result Type: Automatic (Numeric)

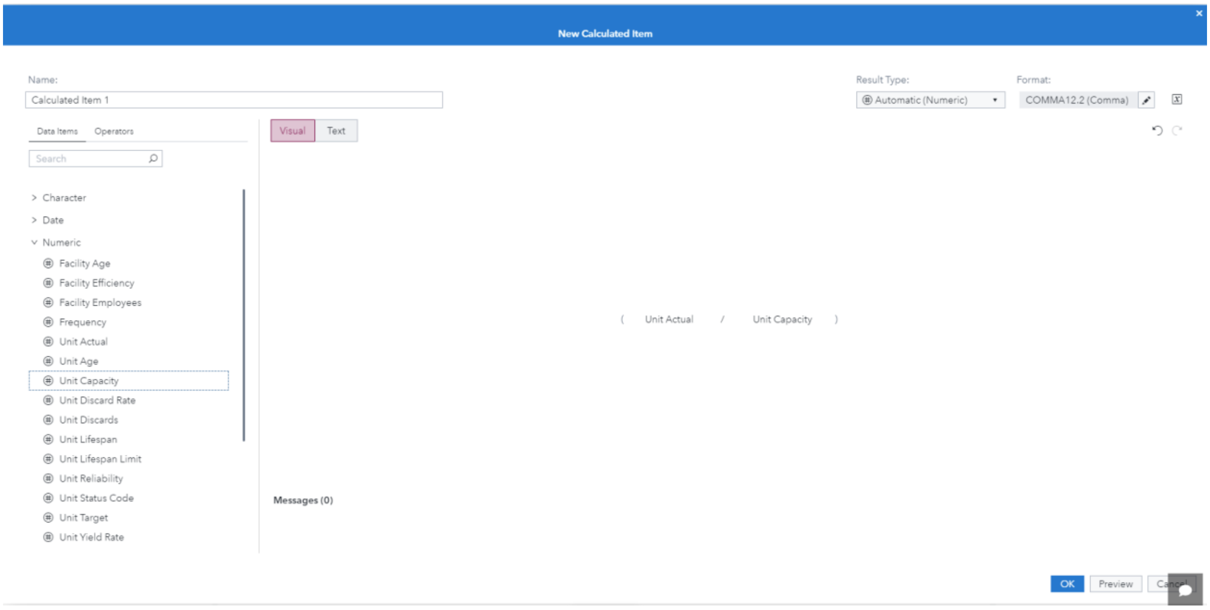
Format: COMMA12.2 (Comma)

(number / number)

Messages (2)

No value specified

No value specified



New Calculated Item

Name: Calculated Item 1

Result Type: Automatic (Numeric)

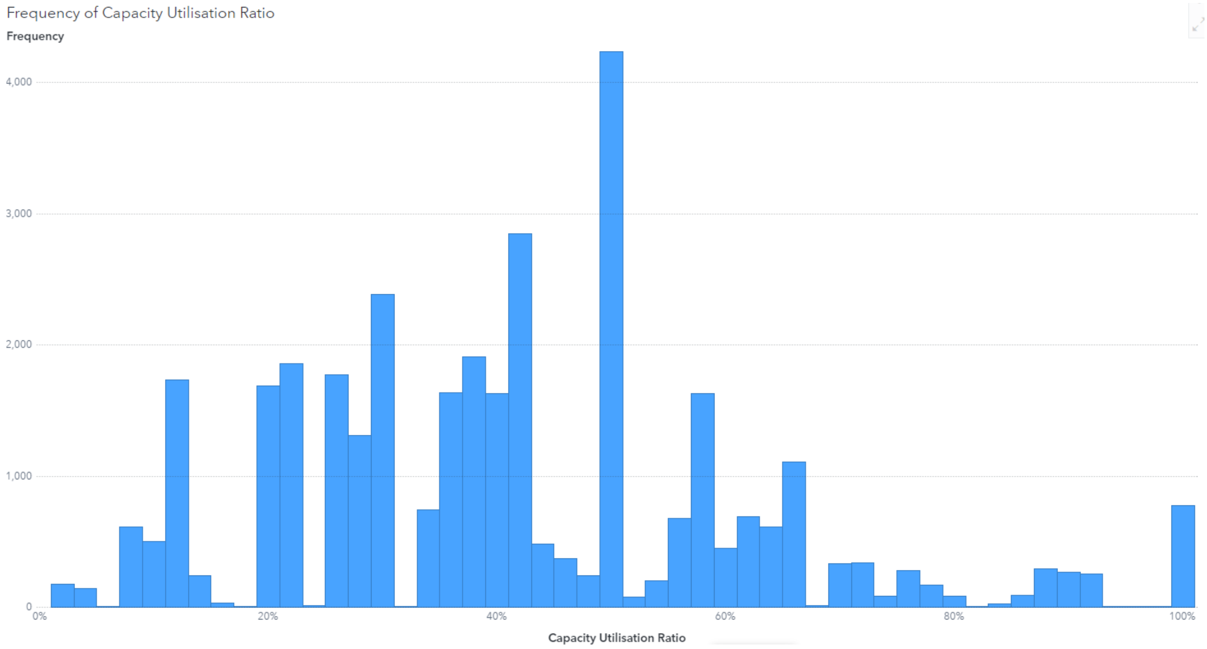
Format: COMMA12.2 (Comma)

(Unit Actual / Unit Capacity)

Messages (0)

The division operator is used and then the variables are assigned to it to create the new calculated item.

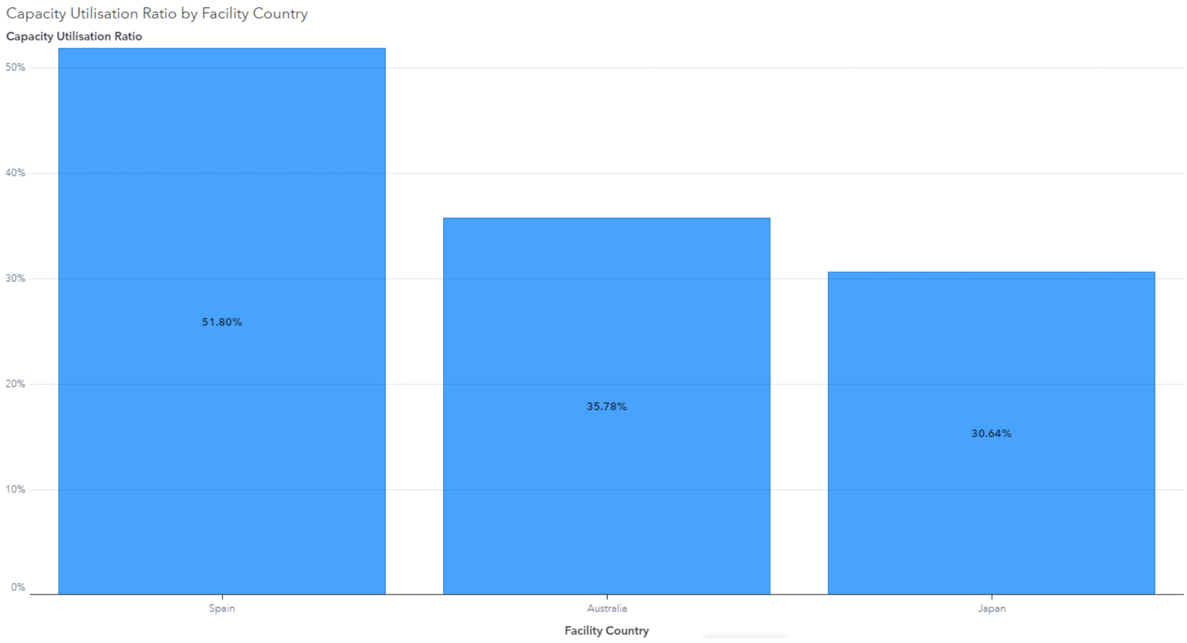
Histogram for Capacity Utilisation Ratio:



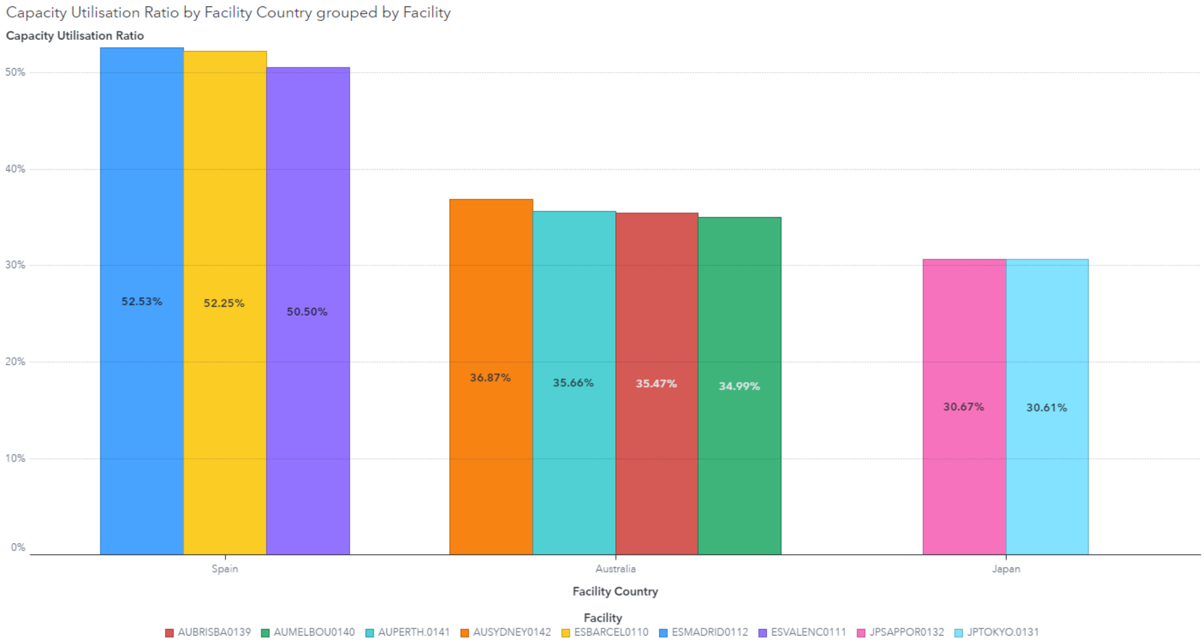
The above histogram shows distribution of capacity utilisation ratio illustrating how many facilities are working at a certain utilisation ratio. We can see that more than 4000 facilities work at 50% capacity followed by almost 3000 facilities working at 44% capacity. We can observe that majority of the facilities work at 60% or lower capacity while only a few facilities work above 70% capacity with only about 800 facilities working at 100% capacity. This makes sense as most facilities don't work at maximum capacity but at an optimum level in order to maximize life of the machineries and only produce products as per the demand forecast.

Q5. Prepare a bar chart to show the average 'Capacity Utilisation Ratio' by facility for each country. Use a filter to show only Spain, Australia, and Japan in this bar chart. Copy/export the chart into your answer script. Interpret your chart.

A. The bar graph showing Capacity Utilisation Ratio for Spain, Australia and Japan:



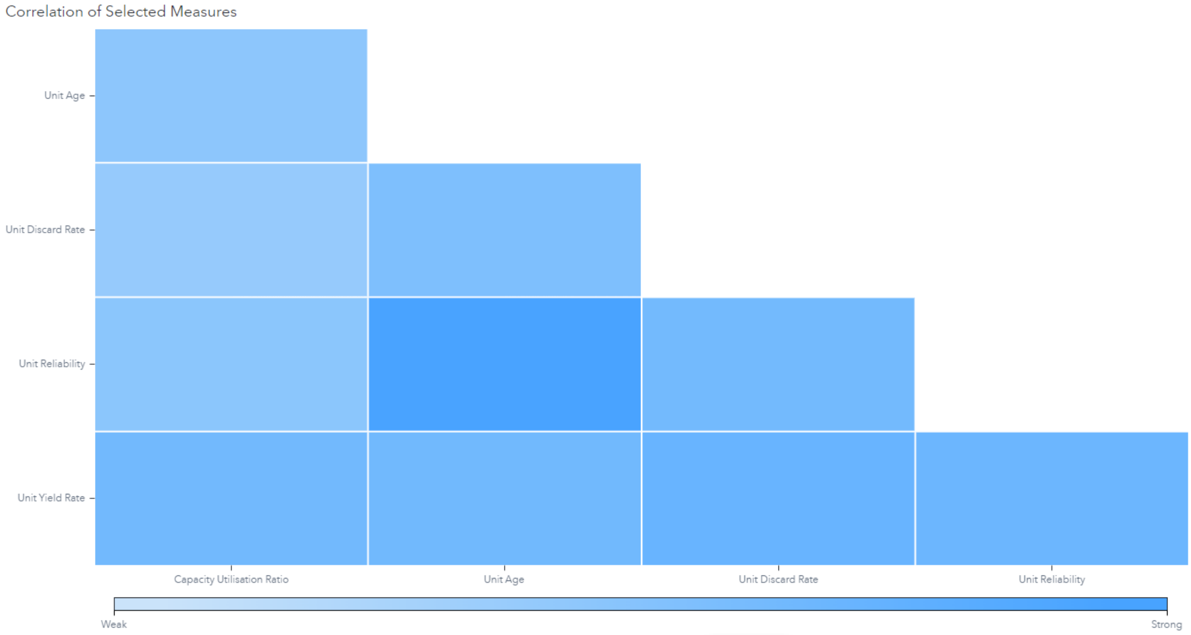
We can see in the above bar chart that Spain is the only country among the three countries to have a capacity utilisation of over 50% while Australia and Japan's utilisation is between 30% and 40%.



We can conclude that Spain is making better use of their capacity compared to the other two countries. Tokyo has the lowest capacity utilisation of 30.61%.

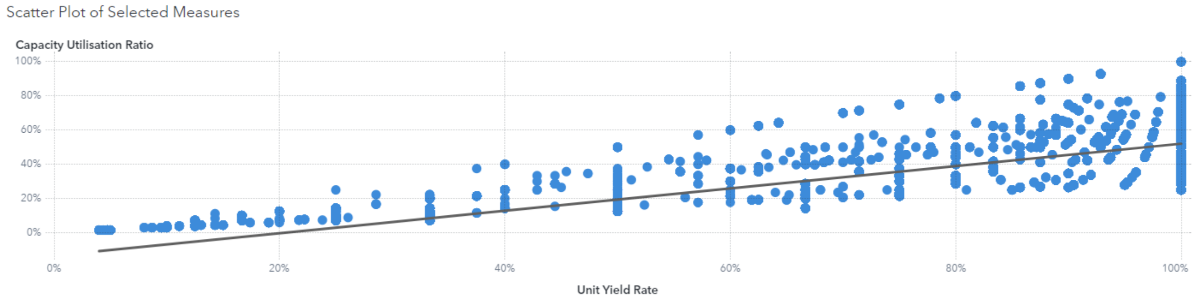
Q6. There are many factors that could explain the variation observed in the Unit Capacity Utilization Ratio. Identify two such factors and demonstrate how these two factors explain the variation in Unit Capacity Utilization Ratio with the help of two charts and associated interpretation.

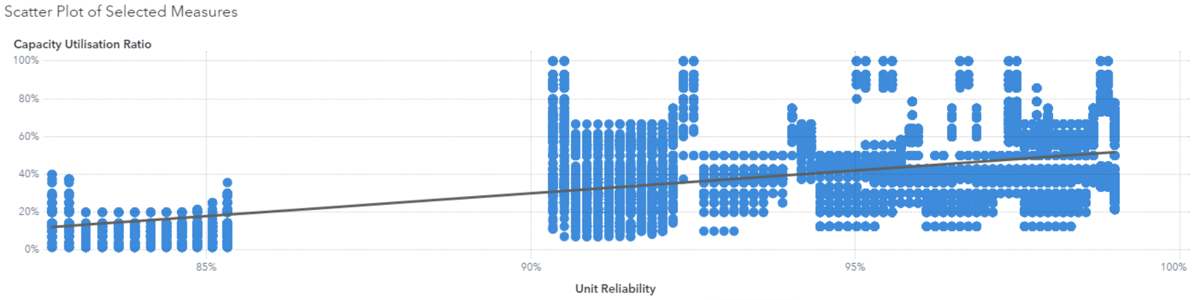
A. To find the factors affecting the Capacity Utilisation Ratio we use a correlation matrix. It is observed that Unit Yield Rate, Unit Reliability and Unit Age.

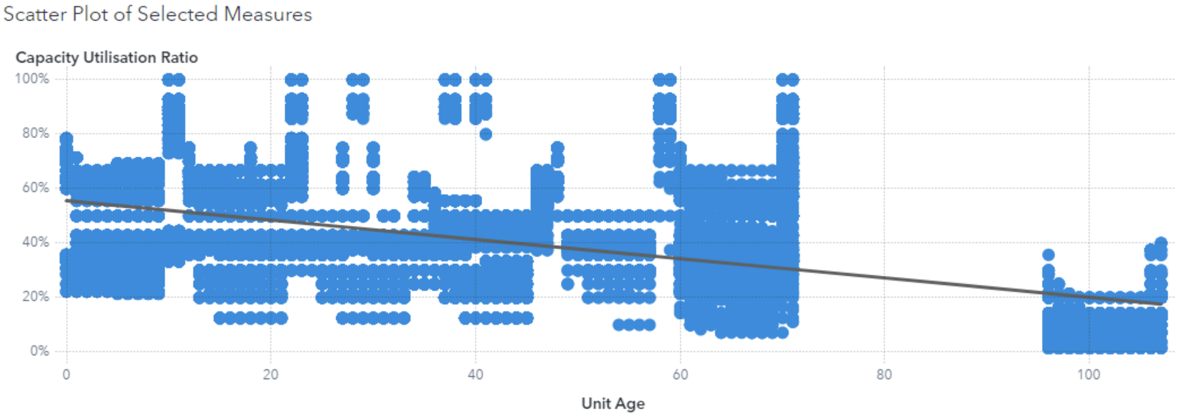


Explanatory Factors	Correlation value	Strength
Unit Yield Rate	0.6434	Strong
Unit Reliability	0.4528	Moderate
Unit Age	-0.4449	Moderate

Scatter plots:





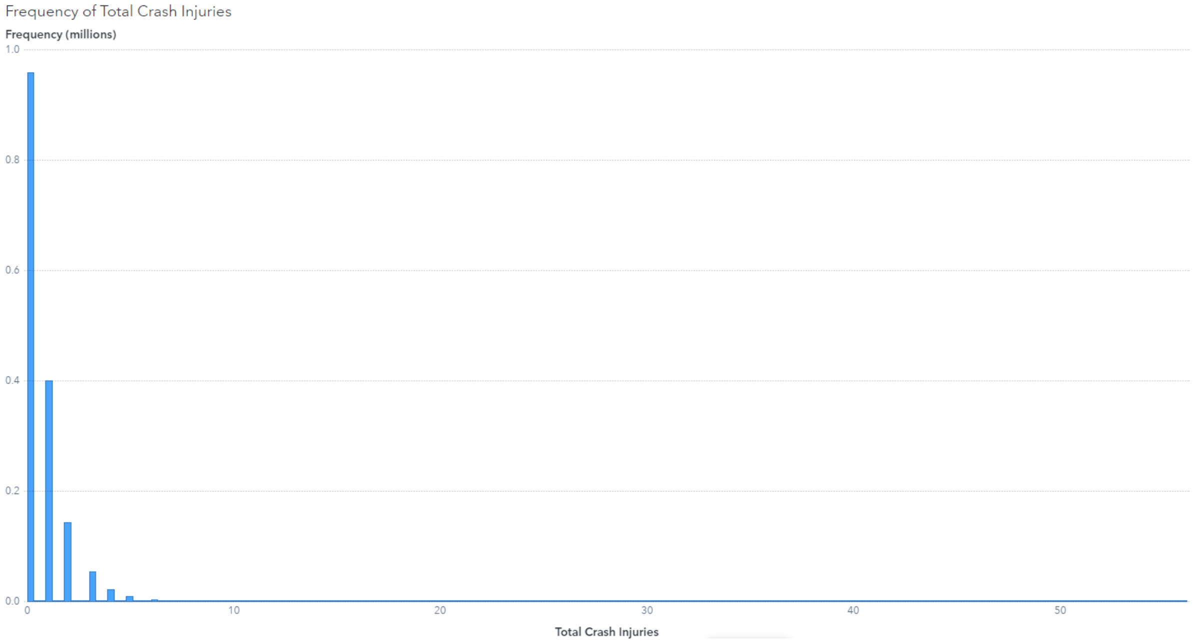


It can be observed that Unit Yield Rate and Unit Reliability are positively correlated while Unit age is negatively related. Units with higher Yield Rate and Reliability tend to have higher Capacity Utilisation ratio and vice-versa.

Task 2: Predictive Data Analytics

Q1. Note the variable 'Total Crash Injuries' provide a number of injuries associated with every accident. In SAS Viya, prepare a histogram showing the distribution of Total Crash Injuries. What can you say about the distribution of crash injuries?

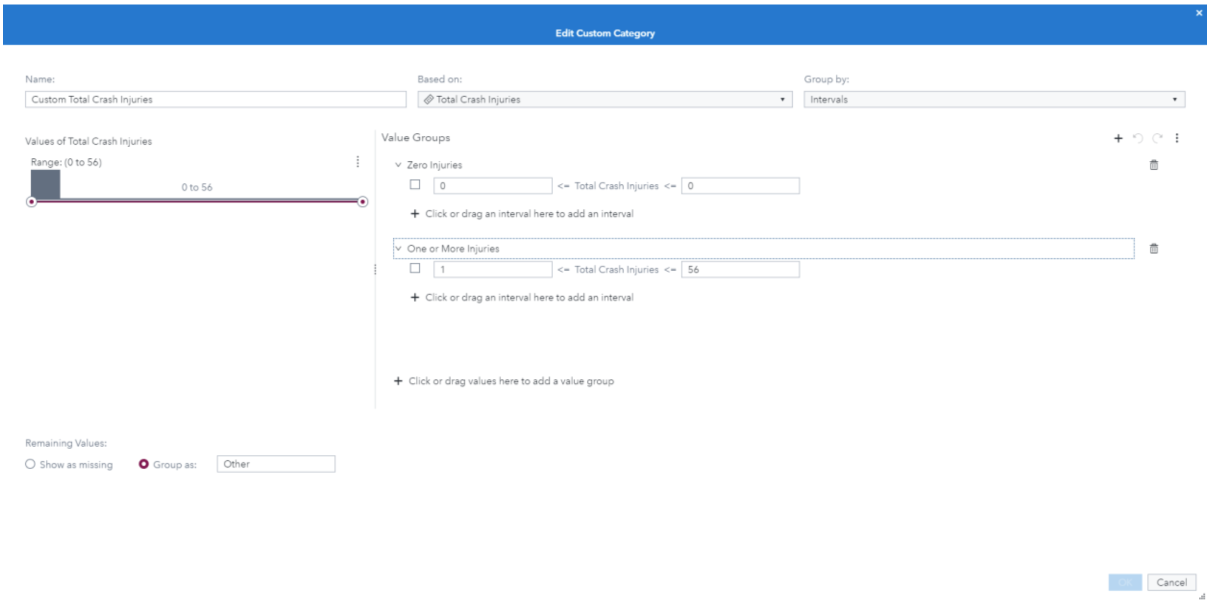
A.



The above histogram shows distribution of Total Crash Injuries ranging between 0 to 56. The histogram is heavily skewed to the right. It is evident that there are about a million crashes with no injuries.

Q2. Create a new custom category variable based on 'Total Crash Injuries' variable. This new custom category variable should contain two categories only. One category is injuries equal to zero, while the other category is for crashes with one or more injuries. Visualise the frequency of the two new categories you just created on a bar chart. How many crashes report zero injuries?

A. Steps to create custom category:



Edit Custom Category

Name: Custom Total Crash Injuries

Based on: Total Crash Injuries

Group by: Intervals

Values of Total Crash Injuries

Range: (0 to 56)

0 to 56

Value Groups

Zero Injuries

□ 0 <= Total Crash Injuries <= 0

+ Click or drag an interval here to add an interval

One or More Injuries

□ 1 <= Total Crash Injuries <= 56

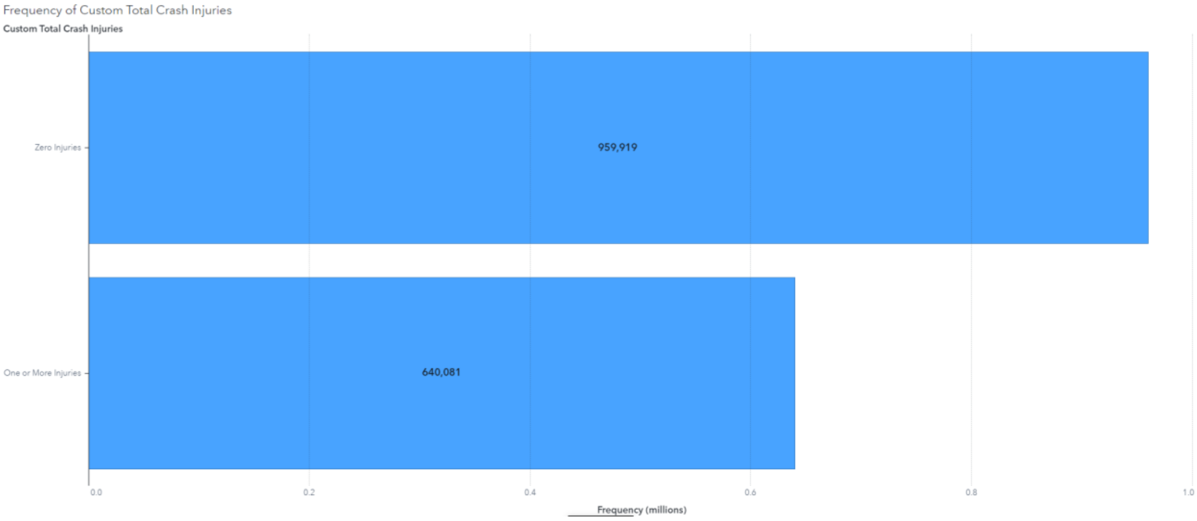
+ Click or drag an interval here to add an interval

+ Click or drag values here to add a value group

Remaining Values:

○ Show as missing ● Group as: Other

Frequency Distribution of the new category Custom Total Crash Injuries:

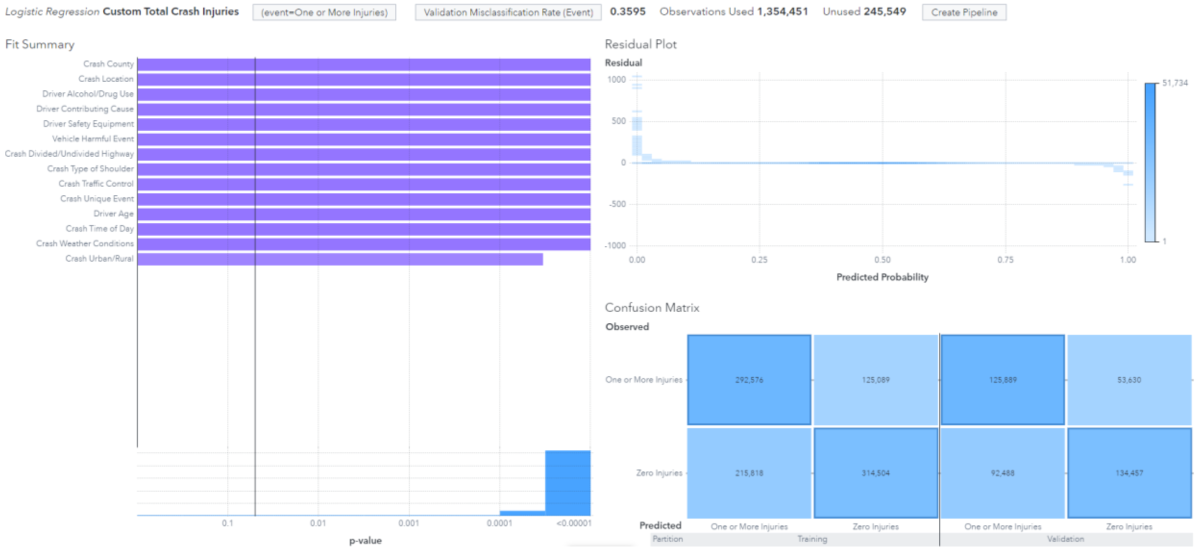


There are a total of 959,919 crashes with zero injuries.

Q3. In Q2, you created a new categorical variable with only two values (binary). Your task now is to develop two models that can predict the value this target variable takes, given other explanatory variables. In other words, you attempt to predict if a crash is going to result in injuries (or not) given other important variables. What are the two models (or techniques) you can use to predict this target variable? Create one model to predict the target variable you created in Q2. Assess this model's accuracy. What are the most important variables in predicting this target variable? Create the second model to predict the target variable. Assess this model’s accuracy. What are the most important variables identified by the model to predict the target variable? Compare the performance of the two model. Report and discuss the results of your comparison. Which model is the champion?

A. Since the target variable Custom Total Crash Injuries is a category, we use classification models. We use ‘Logistic Regression’ and ‘Decision Tree’ models for this purpose. The data is split into training and validation set: 70% training and 30% validation.

Logistic Regression:



Logistic Regression **Custom Total Crash Injuries** (event=One or More Injuries) Validation Misclassification Rate (Event) **0.3595** Observations Used **1,354,451** Unused **245,549** Create Pipeline

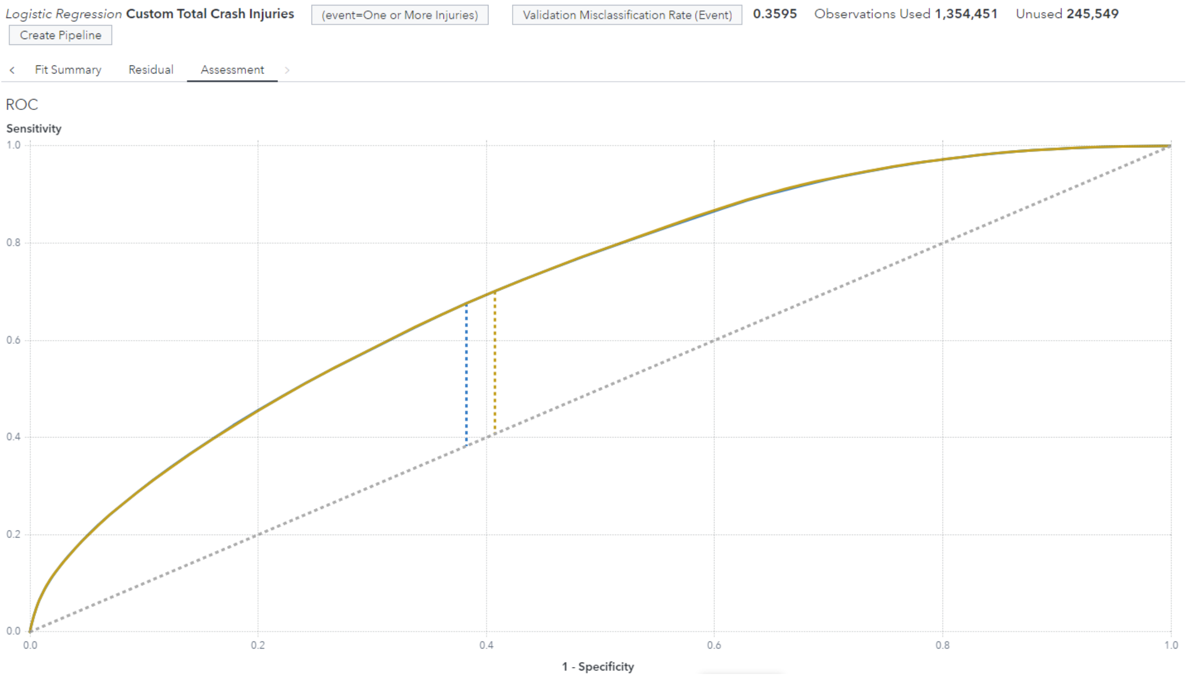
Observed	Training: One or More Injuries	Training: Zero Injuries	Validation: One or More Injuries	Validation: Zero Injuries
One or More Injuries	292,576	125,089	125,889	53,630
Zero Injuries	215,818	314,504	92,488	134,457

The misclassification rate on validation data is 0.3595 which is considerably low. One can also look at the Confusion Matrix and that more than 60% of the crash injuries are rightly predicted.

The only measure variable of importance is Driver Age. Rest of the important variables for the model are categorial variables as follows:

Variable (Categorical)	**Importance**
Crash County	Refers to crashes in local administrative divisions
Crash Divided/Undivided Highway	The type of highway related to the crash
Crash Location	Whether the crash took place in a car park, bridge or at an intersection of roads etc.
Crash Time of Day	Time of the day when crash happened (morning, noon, evening or night)
Crash Traffic Control	Type of traffic control
Crash Type of Shoulder	Road conditions (paved, unpaved etc.)
Crash Weather Conditions	Weather conditions that might cause the crash

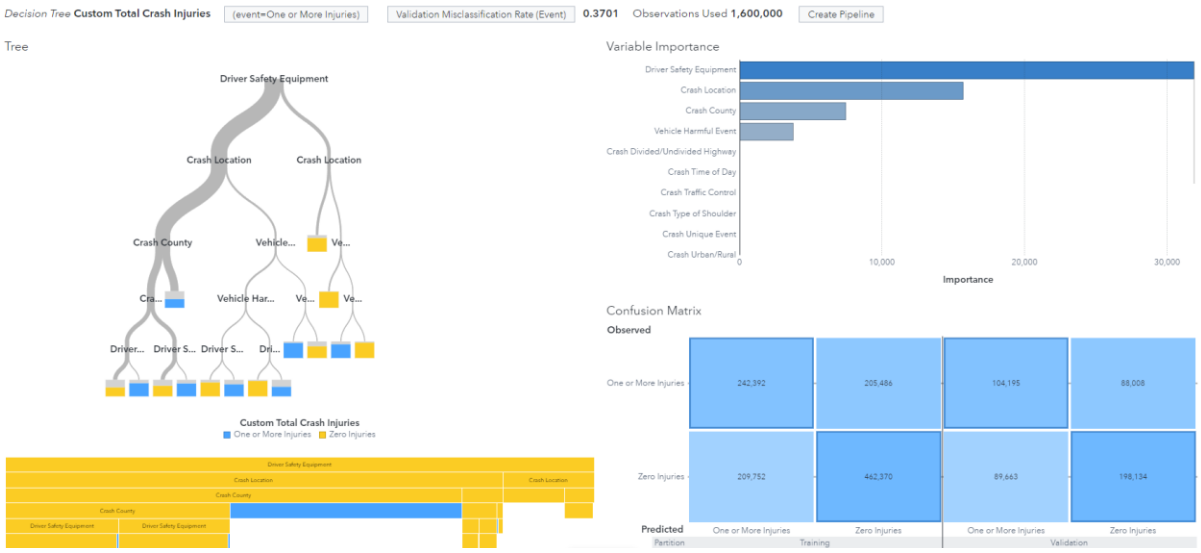
Driver contributing cause	Indicates the possible reason for the crash
Driver safety equipment	Driver's safety measures
Vehicle Harmful Event	Refers to the type of crash
Driver Alcohol/Drug use	If the driver was under the influence by alcohol/drugs
Crash Divided/ Undivided Highway	Crashes on divided or undivided highway



Logistic Regression **Custom Total Crash Injuries** (event=One or More Injuries) Validation Misclassification Rate (Event) **0.3595** Observations Used **1,354,451** Unused **245,549**

The cut-off this model is 0.44 for the validation set which means any prediction probability above this value will surely lead to an accident.

Decision Tree:

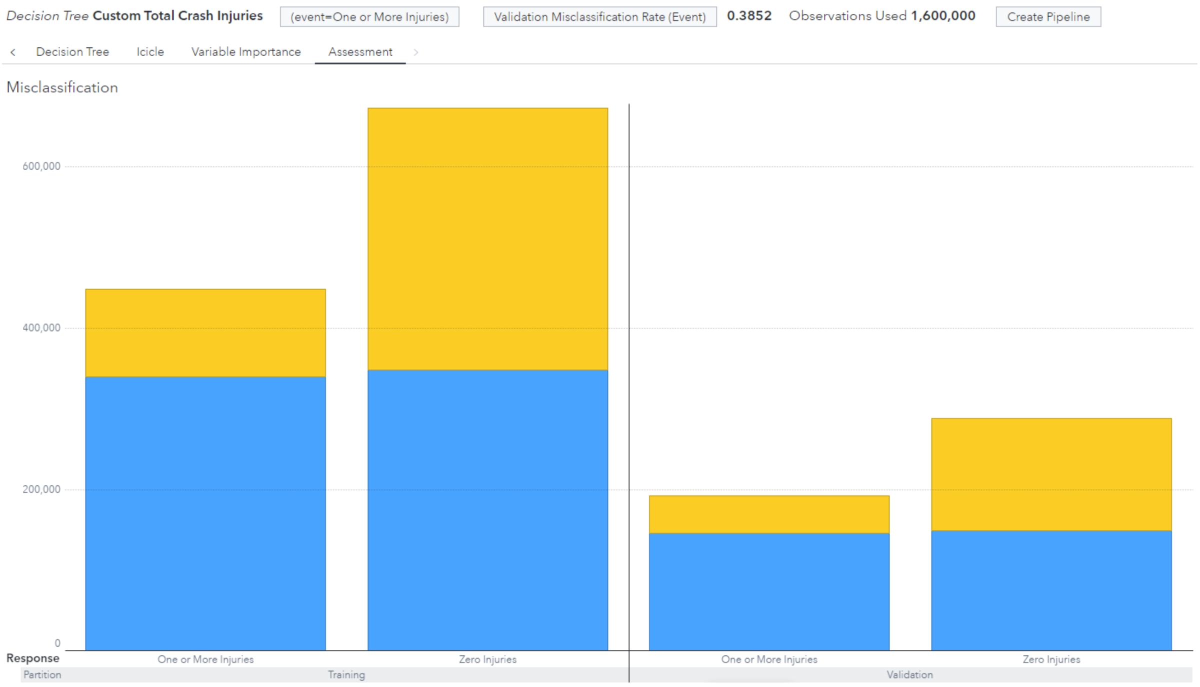


Decision Tree **Custom Total Crash Injuries** (event=One or More Injuries) Validation Misclassification Rate (Event) **0.3701** Observations Used **1,600,000**

Confusion Matrix

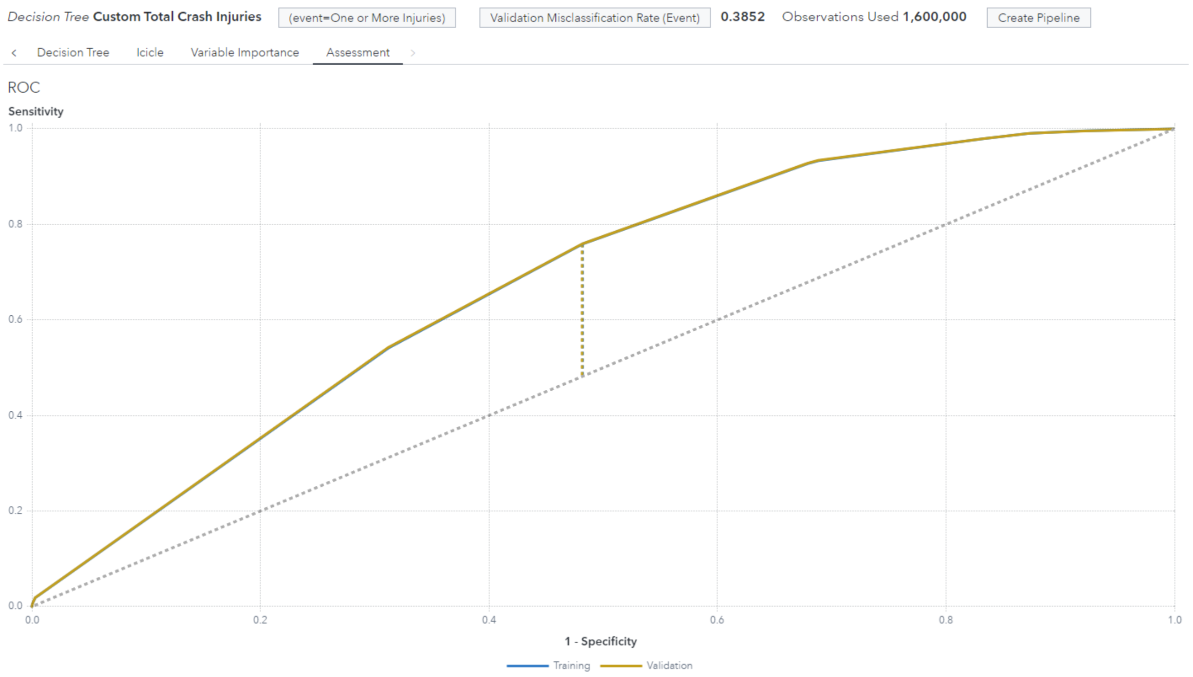
Observed	One or More Injuries	Zero Injuries	One or More Injuries	Zero Injuries
One or More Injuries	242,392	205,486	104,195	88,008
Zero Injuries	209,752	462,370	89,663	198,134
Predicted	One or More Injuries	Zero Injuries	One or More Injuries	Zero Injuries
Partition	Training		Validation	

The misclassification rate of this model on validation rate is 0.3701 which is higher than the previous model.



Decision Tree **Custom Total Crash Injuries** (event=One or More Injuries) Validation Misclassification Rate (Event) **0.3852** Observations Used **1,600,000** Create Pipeline

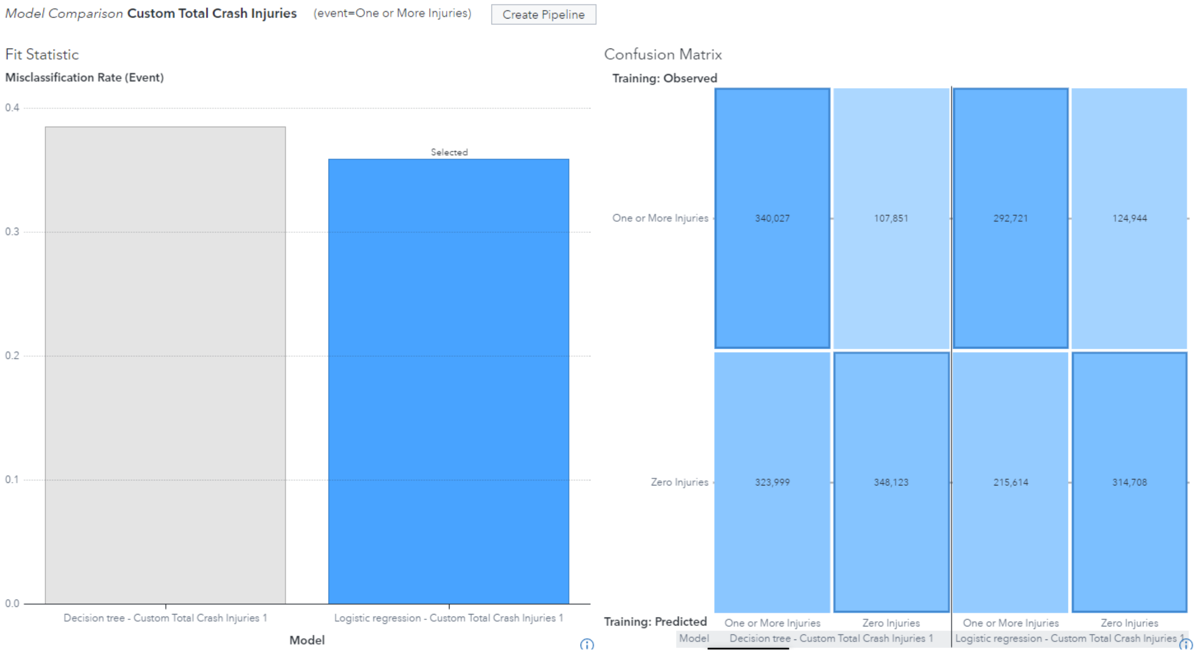
It is evident from the misclassification graph that the model predicts 'One or More Injuries' more accurately then zero accuracies. The cut-off for this model is 0.42.



Decision Tree **Custom Total Crash Injuries** (event=One or More Injuries) Validation Misclassification Rate (Event) **0.3852** Observations Used **1,600,000** Create Pipeline

The most important variable for this model is ‘Driver Safety Equipment’ followed by ‘Location’, ‘County’ and ‘Vehicle Harmful Event’.

Comparison:

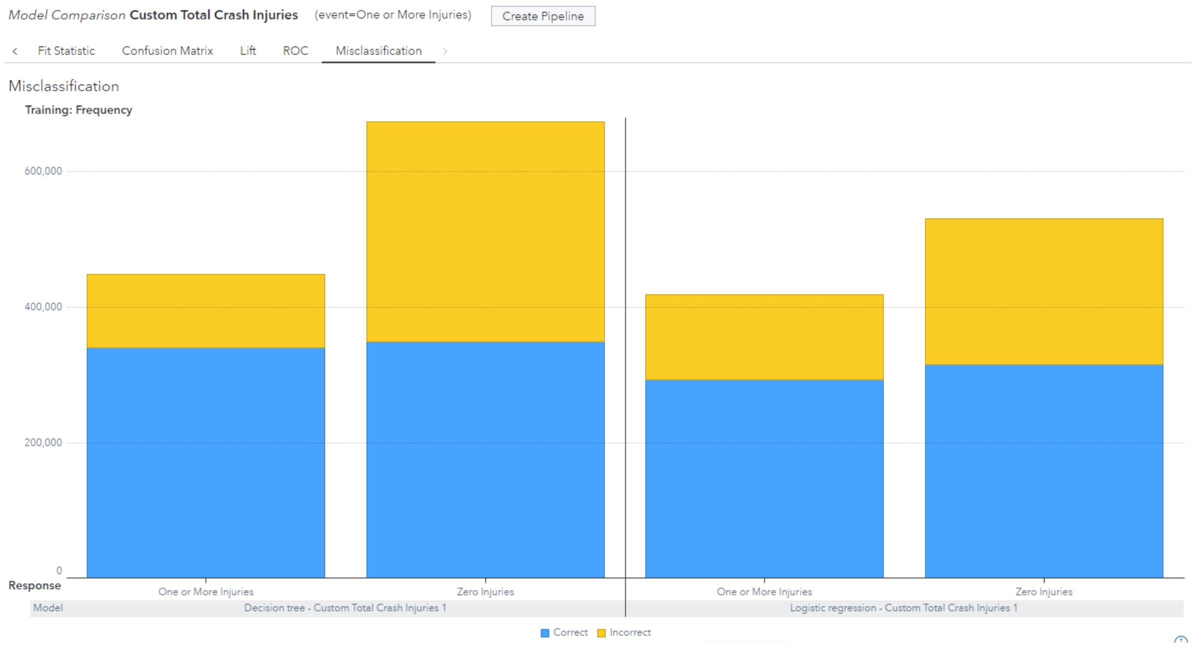


Model Comparison **Custom Total Crash Injuries** (event=One or More Injuries)

Create Pipeline
Fit Statistic
Misclassification Rate (Event)
0.4
0.3
0.2
0.1
0.0
Selected
Decision tree - Custom Total Crash Injuries 1
Logistic regression - Custom Total Crash Injuries 1
Model
Confusion Matrix

Training: Observed	Decision tree - Custom Total Crash Injuries 1		Logistic regression - Custom Total Crash Injuries 1	
Training: Predicted	One or More Injuries	Zero Injuries	One or More Injuries	Zero Injuries
One or More Injuries	340,027	107,851	292,721	124,944
Zero Injuries	323,999	348,123	215,614	314,708

From the above comparison it is clear that Logistic Regression model is better of the two models.



Model Comparison **Custom Total Crash Injuries** (event=One or More Injuries)

Even though the prediction of Decision Tree for ‘One or More Injuries’ is higher than Logistic Regression model the overall accuracy of the Logistic Regression model is better for the both the categories with misclassification rate of 0.3592.

Taking the ROC curves and misclassification rate into account it is clear that the Linear Regression model overpowers the Decision Tree model. The Confusion matrix tells us the same story. Hence Logistic Regression model is the clear choice and the champion model.