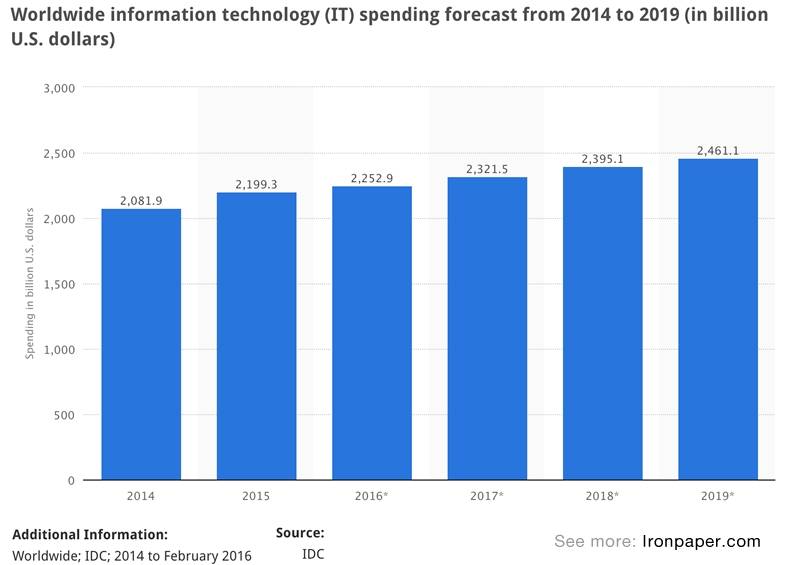
The imminent rollout of 5G and the increasing demand for unlimited data usage is emerging a new behaviour coupled with the influence of digital giants for mobile users. The 5G network is the next generation wireless communication which is almost five times faster than 4G/LTE and the potential highest data rates could reach up to 20Gbps.

The real time high definition streaming is an emerging trend for watching live performance on the mobile device. As per the Department of Communications and the Arts’ Bureau of Communications in Australia, the 5G networks will add up to an additional $2000 in GDP per person in the Australian economy.

Data is growing at an exponential pace which is doubling almost every two years. Therefore, the requirement of database and statistical tools is exploding day by day. More than 80% of the businesses are using databases and data warehouses for storing historical data. Historical data is a prerequisite for predictive analytics. Business managers need to build reports or dashboards for real time performance.

By 2020, the accumulated volume of big data will increase to 44 trillion GB.



**What is a database?**

It is a collection of records organized in a way which can be easily accessed and updated by users. A database can store structured data in a table format usually resides in relational databases (RDBMS).

Data can be human input or machine generated. Multiple users can update data, delete data or generate reports for analysis. The format is searchable by human to generate queries or can be automated through algorithms.

Structured Query Language (SQL) enables those queries within relational databases such as ERP system, inventory control, product names, price, customer or supplier ID, transaction number or customers transactions with their email, phone numbers or address.

Some relational databases such as Customer Relationship management (CRM) still store data which is not structured. However, most CRM databases are structured. All structured data resides in relational databases and datawarehouses.

A database can also store unstructured data such as NoSQL or semi-structured within a non-relational environment without a schema. It may contain textual or non-textual files, emails, websites data, social media data, mobile data, chat, IM, phone recording, audio, MP3, MS office documents, productivity applications, Internet of Things (IOT), satellite imagery or scientific data regarding space exploration, seismic imagery, oil & gas exploration, sensor data, traffic , weather or oceanographic sensors. These data resides in applications, NoSQL databases, data warehouses or data lakes.

Mature analytics tools are mostly developed for more structured data whereas mining unstructured data is still developing. Tool like Alteryx or SaS Viya are available for analyzing unstructured data through machine learning models such as classification, Natural Language Processing (NLP), pattern sensing, sentiment analysis or text-mining algorithms.

You need ETL to extract and analyze your data based on various sources such as applications, databases or warehouses.

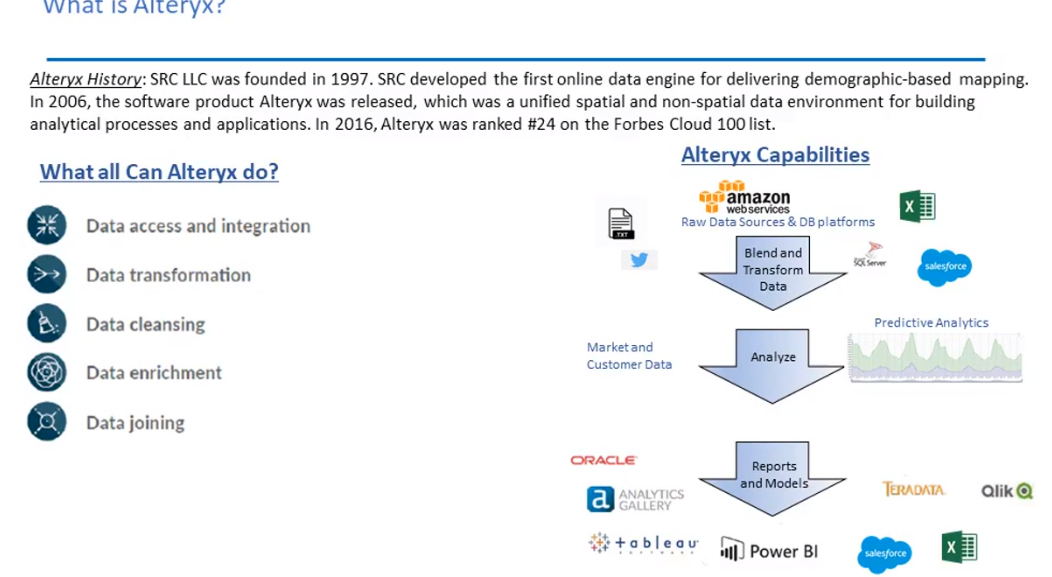
**What is ETL?**

ETL (extract, transform, load) is used to blend data from various sources and often used to build data warehouses. As part of the process, data is initially extracted from various sources through integration, consolidate and convert them into an analyzable format and stored into a data warehouse or other systems.

ETL provides the context for the business through consolidation improving the time productivity without writing codes or scripts and is evolving due to emerging integration requirements such as streaming data. While SQL is the most common method of accessing and transforming data within a database, ETL uses those data to transform into new format as per the business rules.

Self service data preparation is a fast-growing trend which requires the power of data accessing, blending and transforming data into the hands of analytics professionals. ETL supports BI platforms, master data management (MDM) hubs, transactional systems, operational data stores and the cloud.

The most famous ETL tools in the market are Alteryx, SAS, Informatica, IBM Datastage, SAP, KNIME and SSIS (SQL Server Integration Services).



Alteryx can collect raw data from cloud or database platforms to blend and transform data. It then can perform predictive analytics based on market or customer data and at the final stage can be easily integrated with business intelligence tools such as tableau, PowerBI, Oracle, Teradata, Qlick or Excel for visualization.

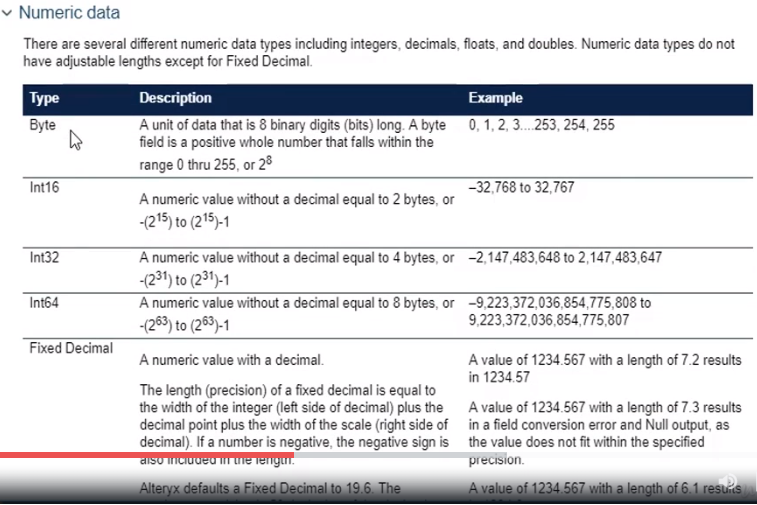
Alteryx can access databases and integrate transform, cleaning, join data and enrich data through creating new variables for analysis. It helps you for data processing, application development, report writing and building models for decision making.

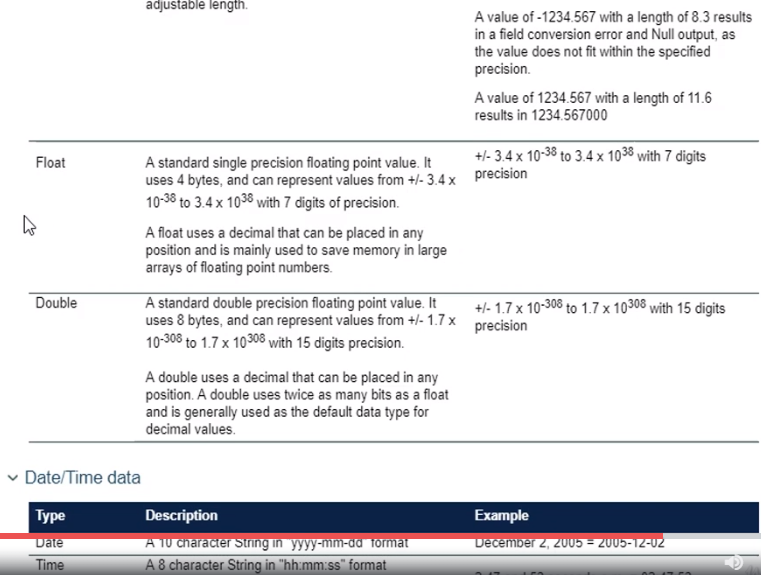
Alteryx designer can do data blending where analysts can access all data source required for analytics. It provides intuitive workflows to deliver data blending and data preparation capabilities for predictive analytics. Analysts can save their workflow or tasks in various files format or can create analytical apps for various businesses. Analysts can also prepare spatial analytics with the capabilities of geo coding, cleansing and delivering special critical insights.

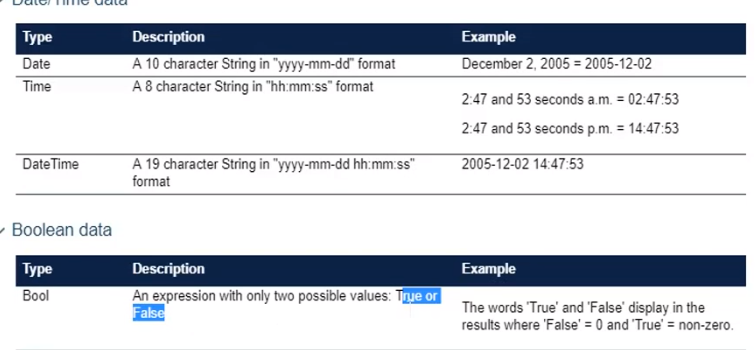
Popular ETL functionality tools in Altyrx available such as Append Fields, Input data, Data cleansing, Select, Sample, Filter, Formula, Sort, Join, and Union.

Analytics functionality or statistical tools such as ARIMA, Decision Tree, Lift Chart, ETS, TS Forecast, Weighted Average, Cross Tab, Linear Regression, Logistic Regression, Naïve Bayes Classifier, Network Analysis, stepwise and Summarize.









We need to understand the types of variables in our dataset. The primary data types are quantitative and qualitative. Quantitative variables can be categorized into two groups –

Continuous (Infinite number of values potentially, and each value is distinct) – sales dollar value, weight of patients, per capita income

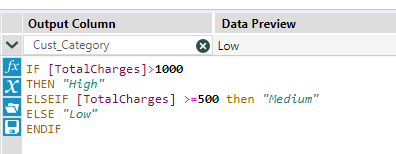
Discrete (Finite number of values) – can be numeric such as number of members in families or it can be categorical such as male or female, red or blue, rating between 1 to 5,

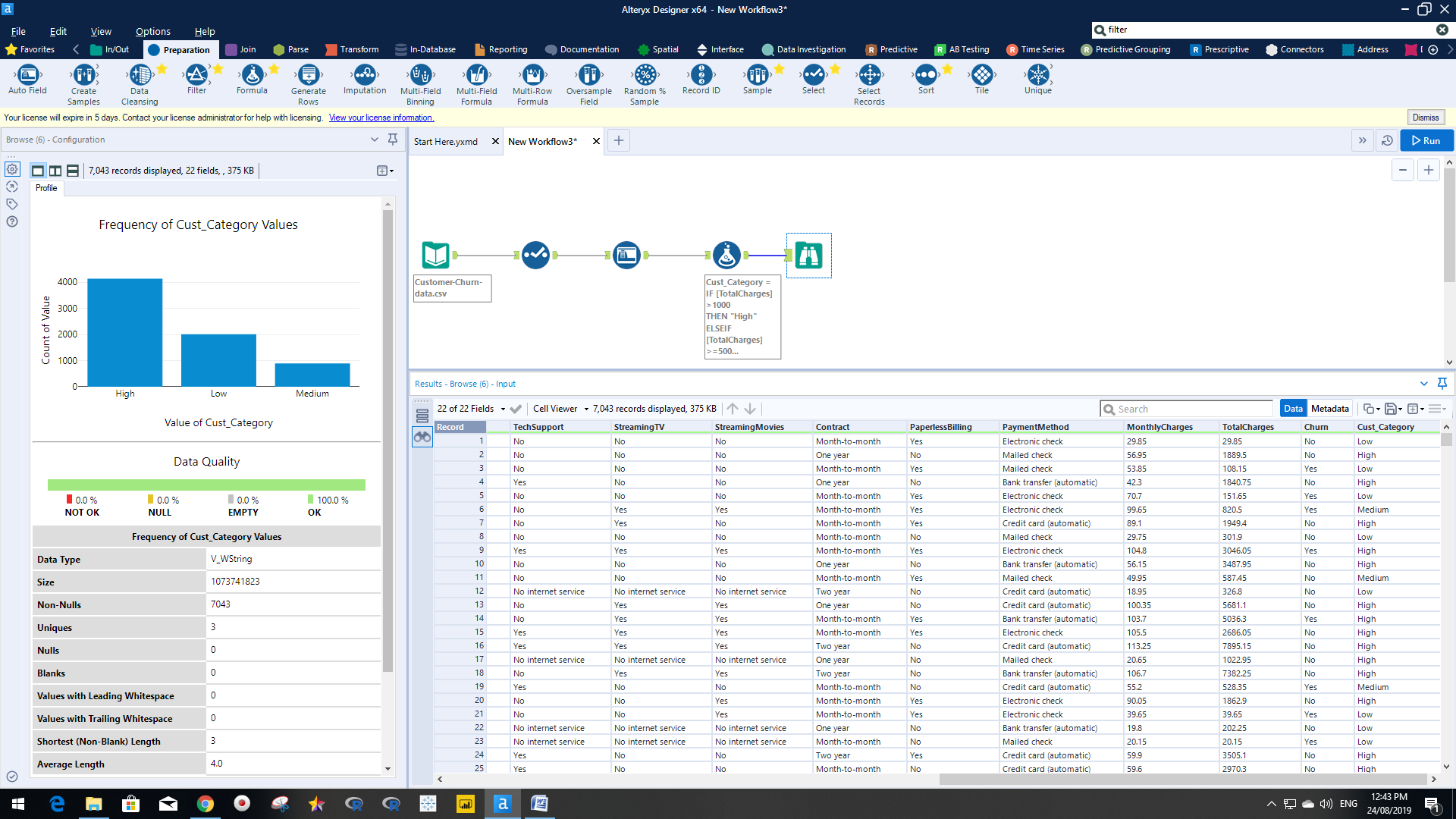
Qualitative variables cannot be quantified in numbers. There are two categories – nominal and ordinal. Nominal variables can not be ordered such as blood group, gender or colour. Ordinal variables can be ordered such as social economic group, winning position in an election, income level categories.

Lets go back to our Alteryx canvas. We will be doing a case study on a telecommunication company which has multiple variables and we will also create additional variables to extend our understanding aor further deep dive analysis.

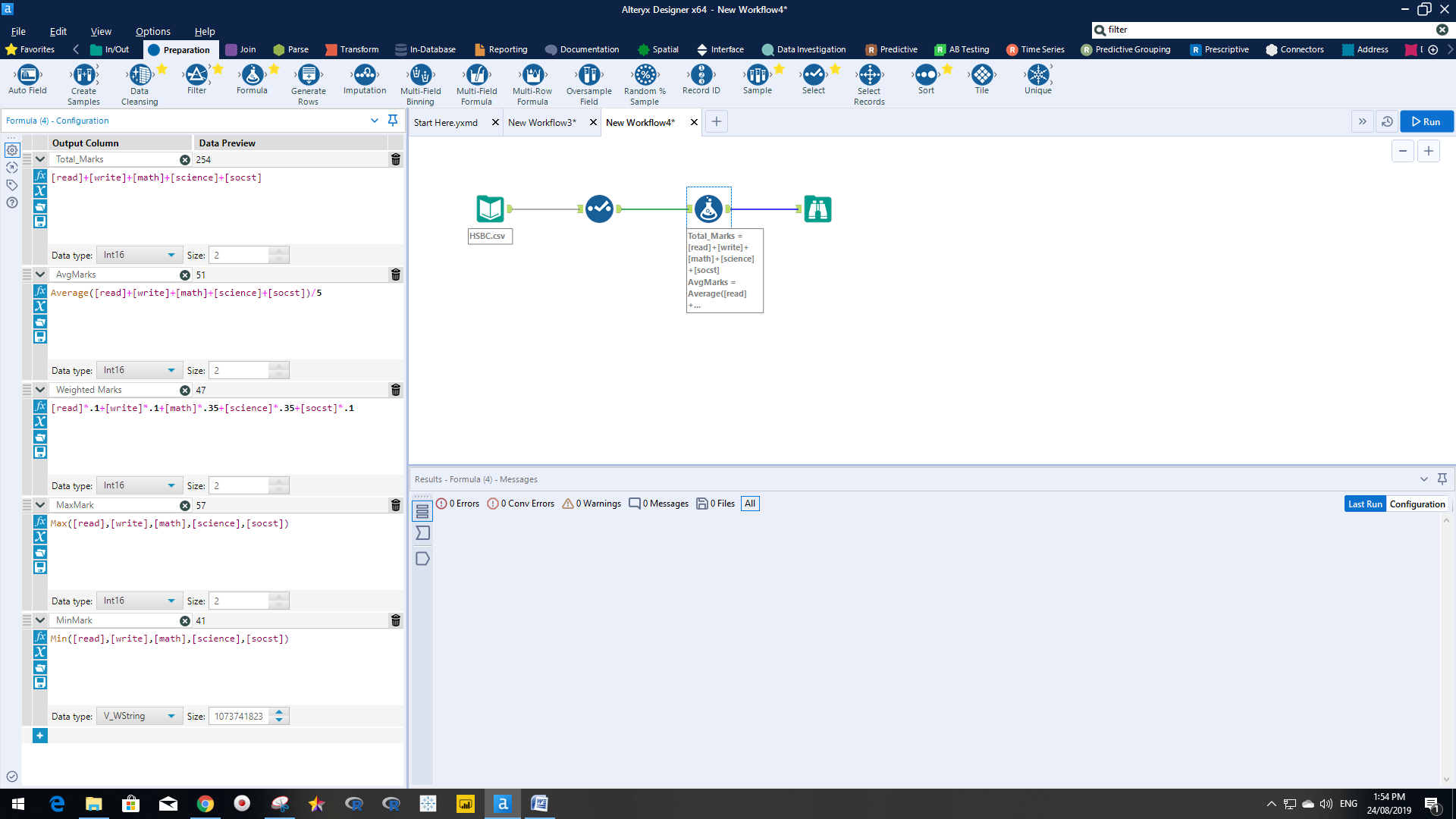
At the first phase, we have checked our data types and changed the numeric values such as monthly charges, totalcharges from string to double. A double uses a decimal that can be placed in any position. A double uses twice as many bits as a float and is generally used as the default data type for decimal values.

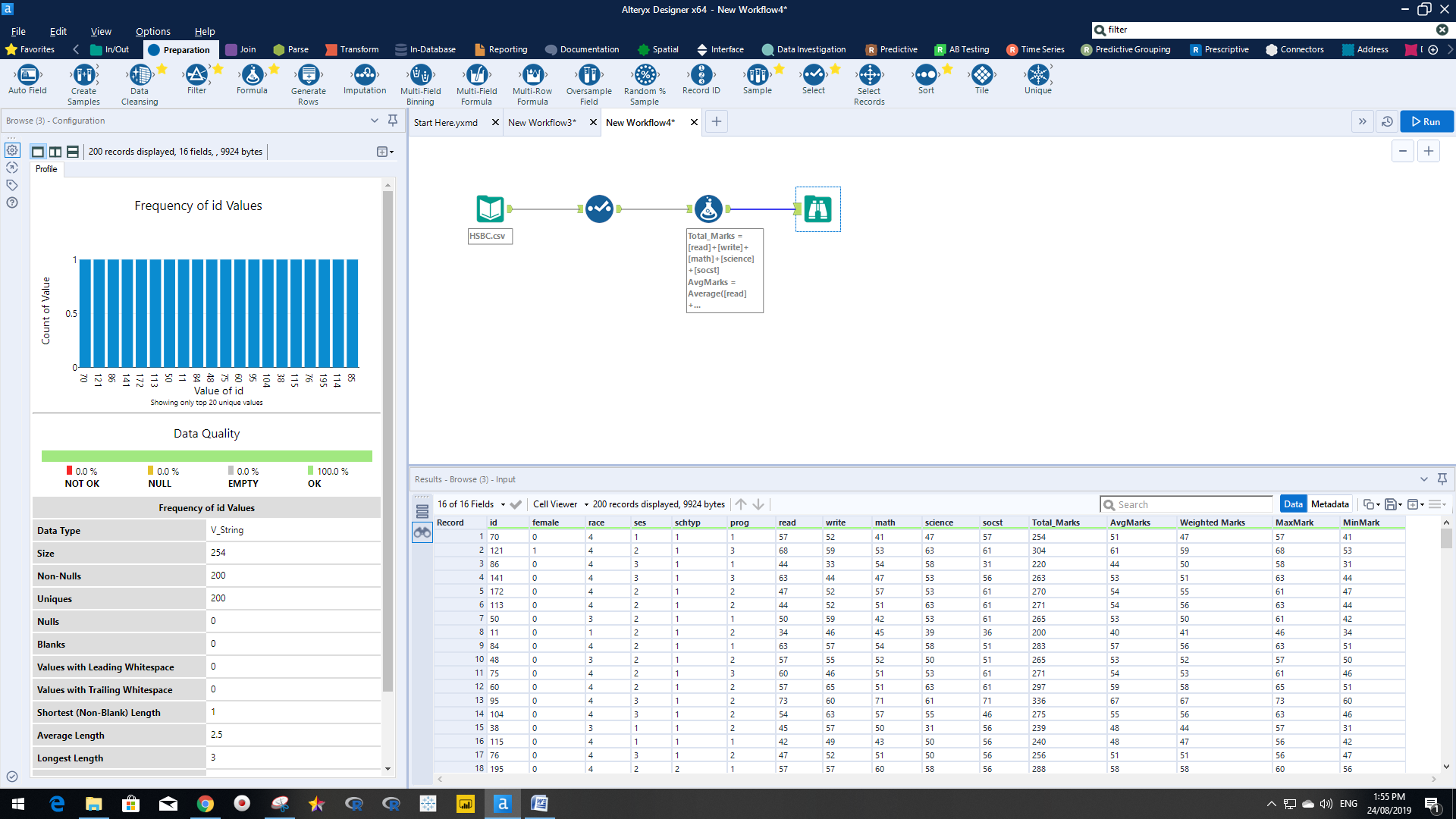
We want to create an additional variable called Cust\_Category with below conditions and run. The last column has now created new variables with three segments high, medium and low.





In the below example, we have data for student Id and their scoring result per subject. We have created multiple formula expressions to create new columns based on their score on each subject. In the bottom left hand corner, we have created formula expressions such as total score per student, average score, weighted marks (we have given more weight to science and low weight to language and social science), maximum and minimum.





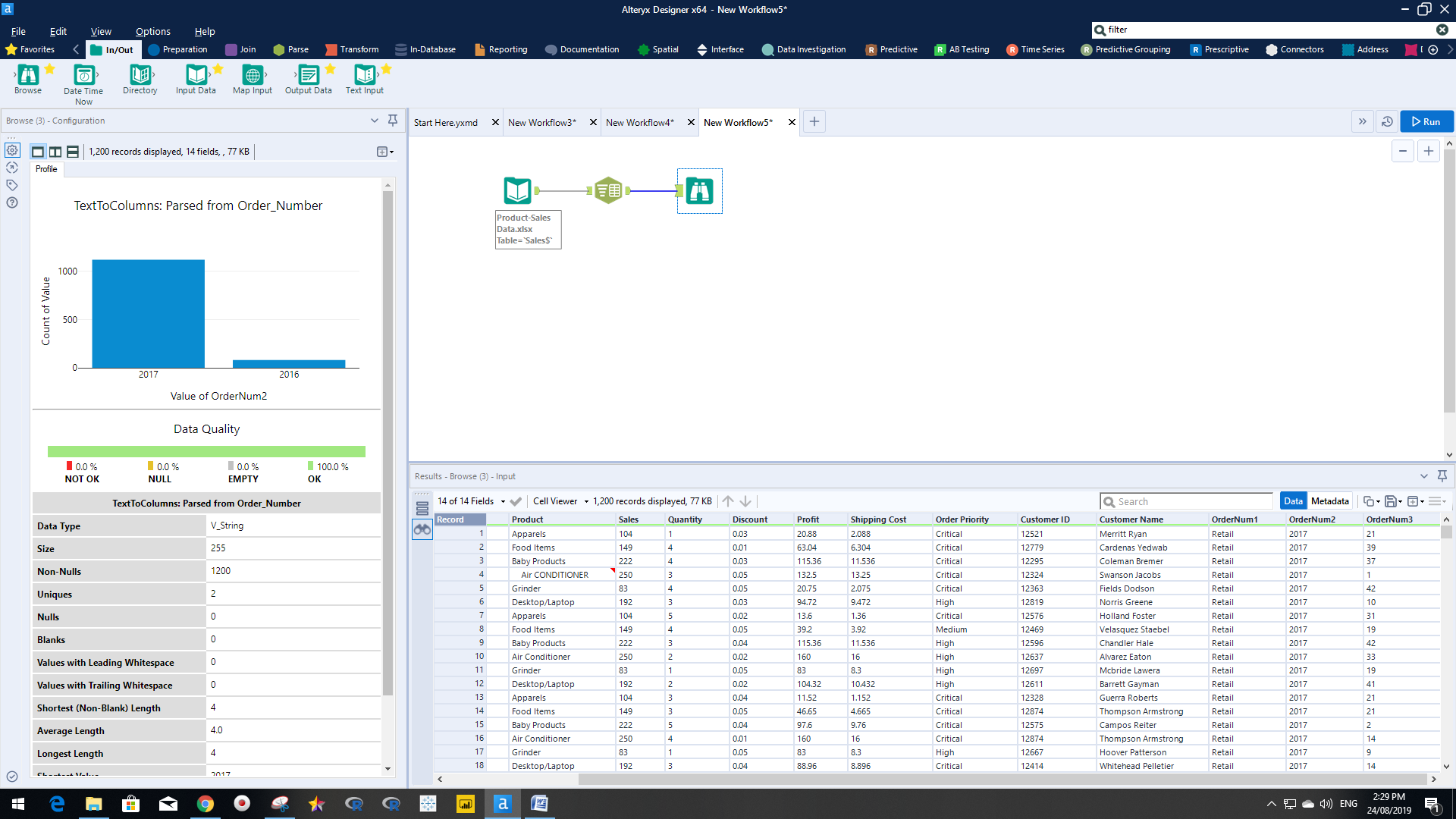
We will work on retail data where as part of the exercise we have to separate the order number into three difference columns.



We have gone to ‘Parse’ and chosen ‘Text to column’ tool and selected 3 columns with Output root name OrderNum and run. The output result came up with three additional columns OrderNum1, OrderNum2, and OrderNum3. You can even create two columns based on your requirements removing first or last characters.

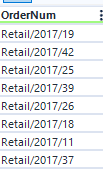
You can also use substring function in the formula tool to separate only the first seven characters using the formula Substring([Order\_Number],7).

The Product column we wanted to show in the uppercase and using trim function (to remove if there are leading spaces before the character and trailing spaces after the last character). We have also created a new column called ProductNew, using the formula [(uppercase(Trim([Product]))] called nested function.

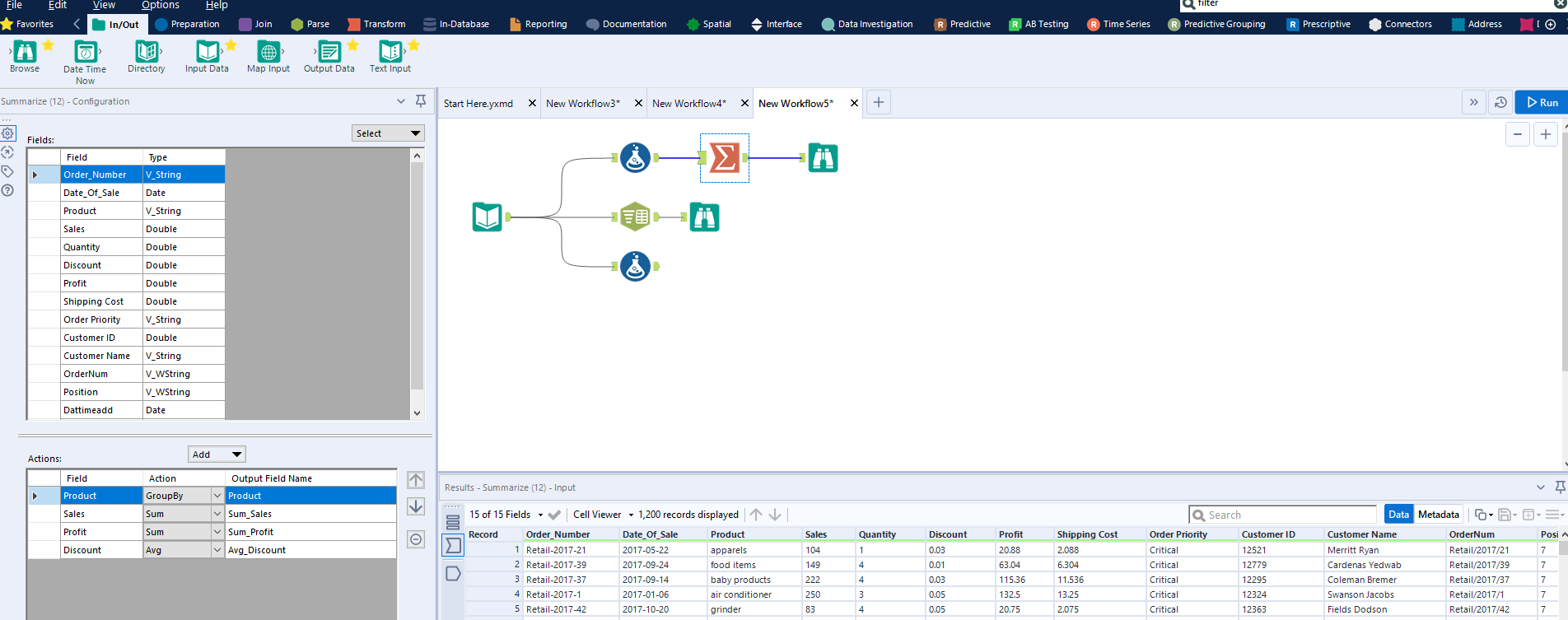


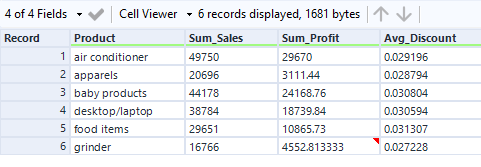
There is a function is called replace. We have now replaced the “-“to “/” of the OrderNumber column with the replace function.

Replace([Order\_Number],"-","/")

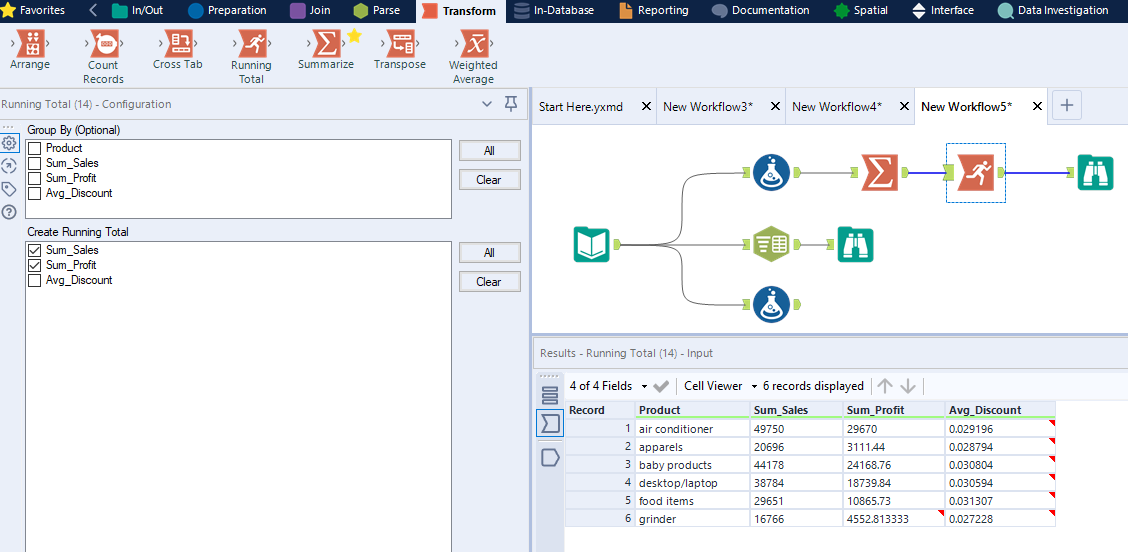
 

We will apply summarize tool to group products and give us total sales and profit with average discount.

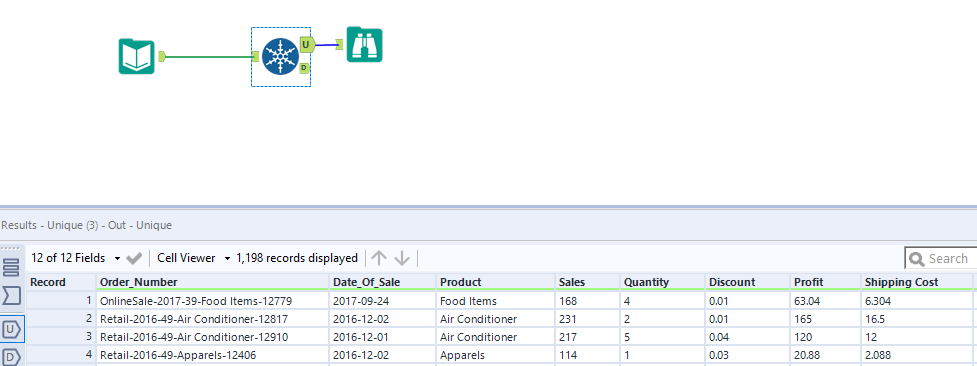


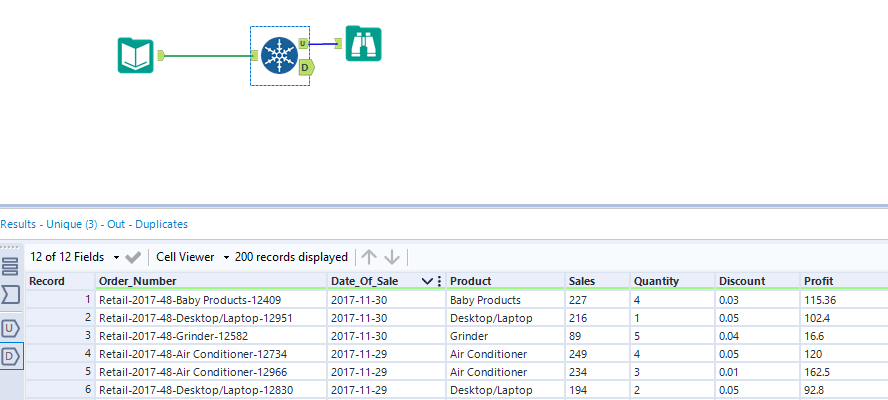


Now, if you want to show the cumulative sum then you need to use Running Total tool. In the below, running total configuration, we are not going to select product in the ‘Group By’ configuration as we have already performed the aggregation through ‘Group By’. We are not able to do apply ‘group by’ function on the ‘Group By’ output.

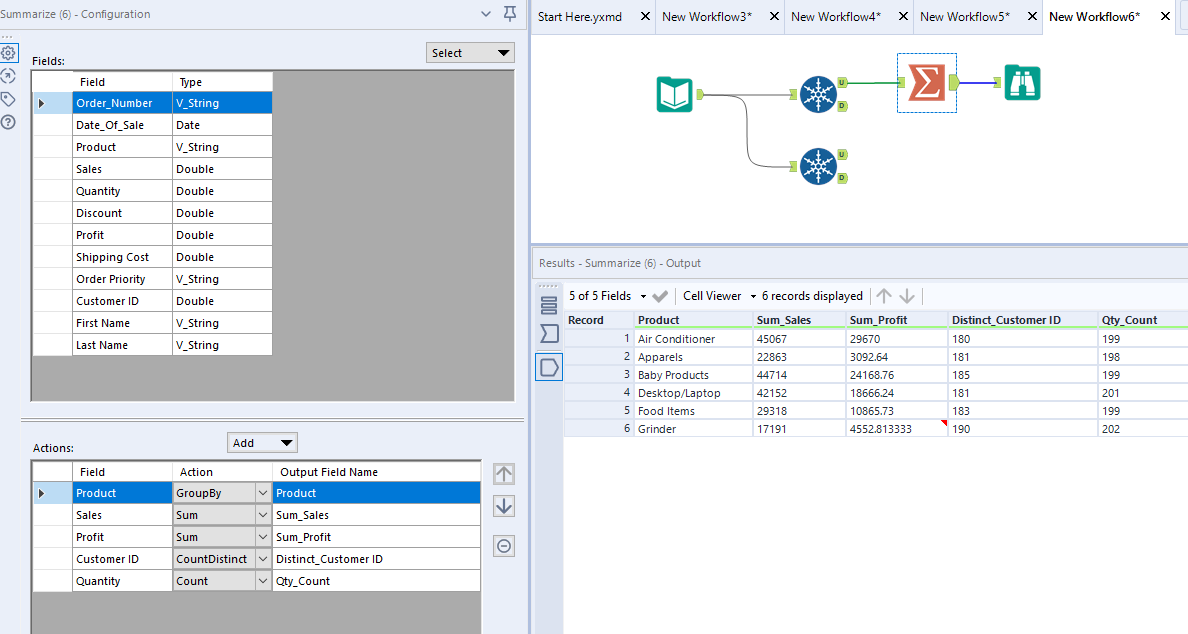


Now, we will discuss one of the key areas all data analysts need to be careful about the duplicate values. Duplicate values can inflate the observation. Alteyrx provides the tool called ‘Unique’ to remove all duplicates. The Unique tool shows the data only unique (U) and only duplicate values (D).

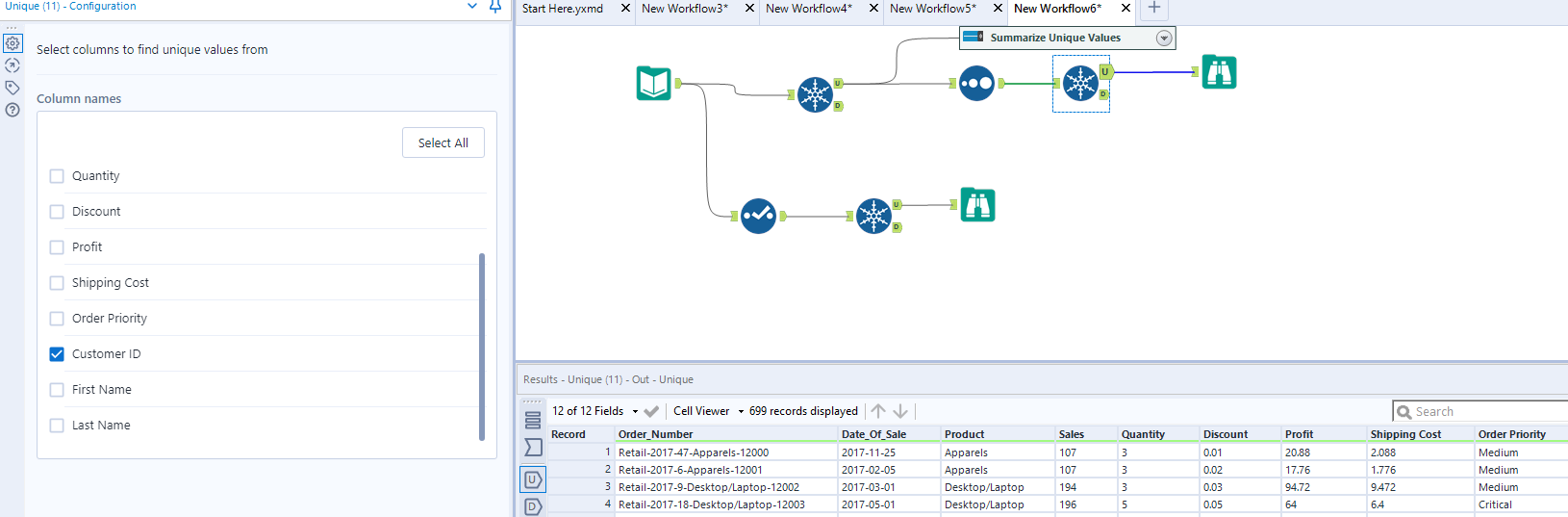




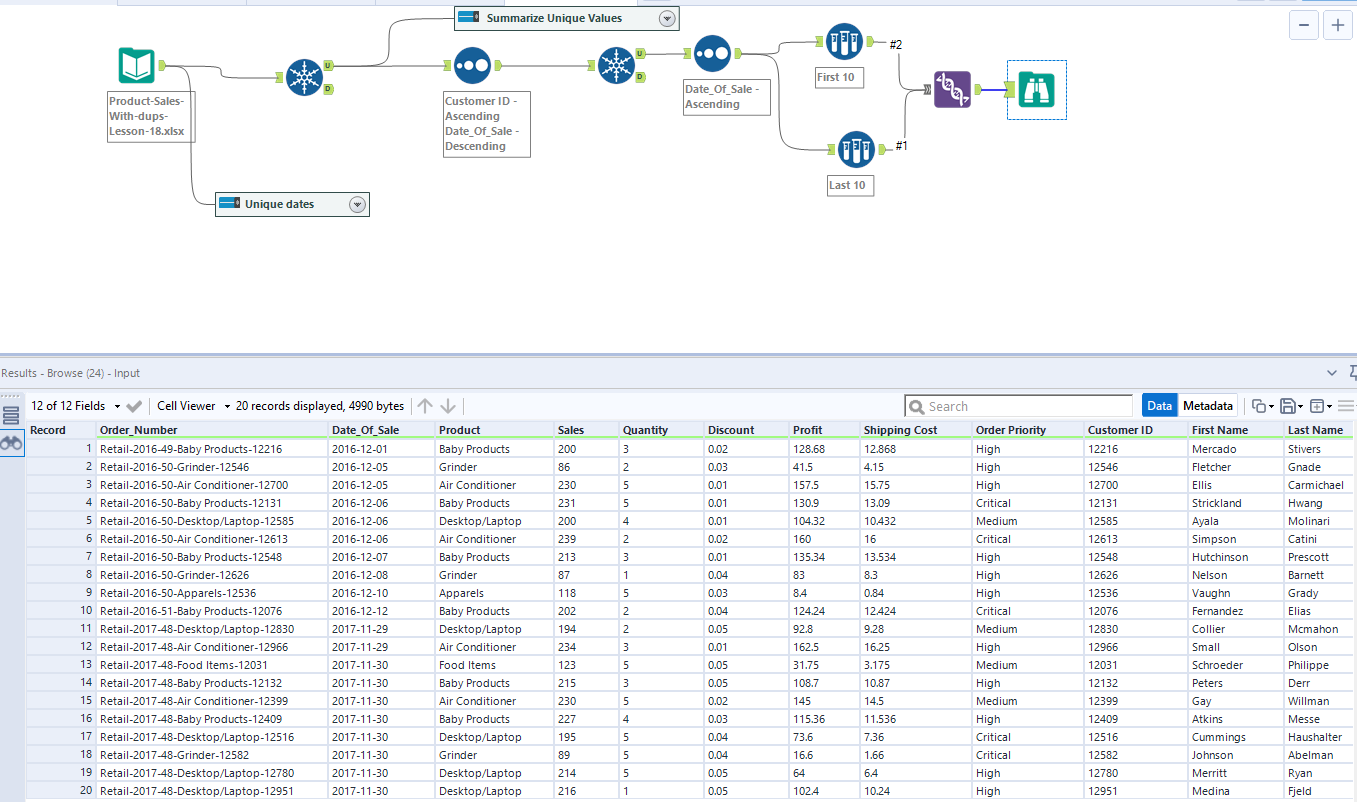
Once we have removed duplicate values, we can do the aggregate analysis grouped by products and sales or profit analysis. We need to use Summarize tool in this case.



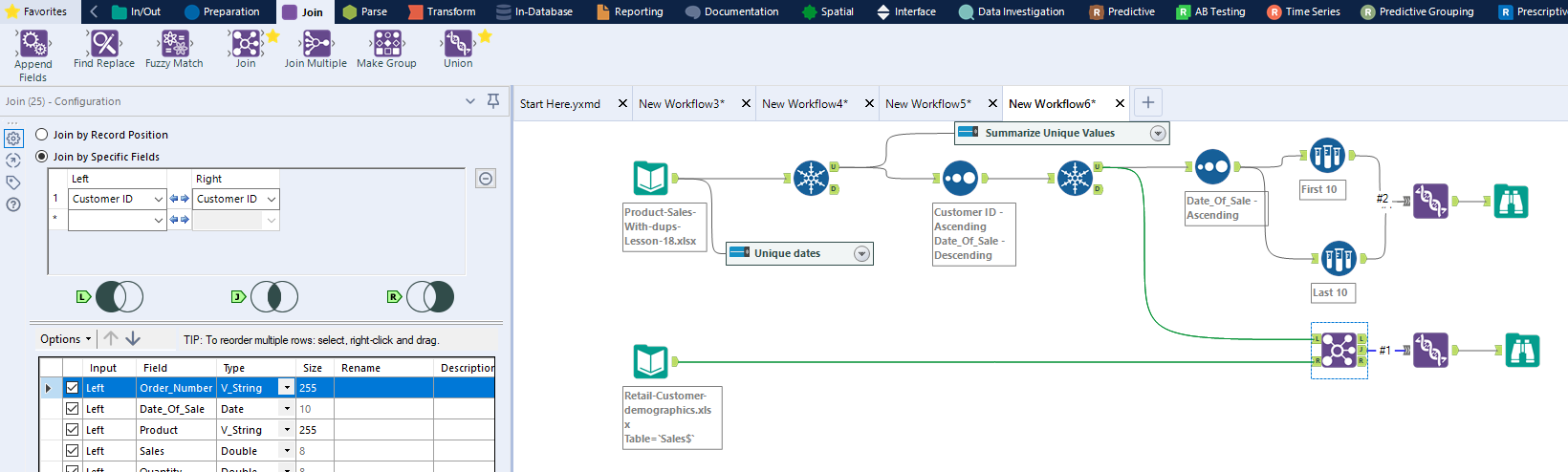
In the below example, we want to see the unique customer ID’s and their activity. We have sorted unique data by customer ID and Date\_Of\_Sales. Next, we have used the Unique tool to see data selecting only Customer ID and we have found 699 rows. This analysis will help data analysts to see the active customers’ recent purchase behaviour. Marketing/sales department can promote discounts or offer new products based on the recent transactions for retaining existing customers or future growth strategy.



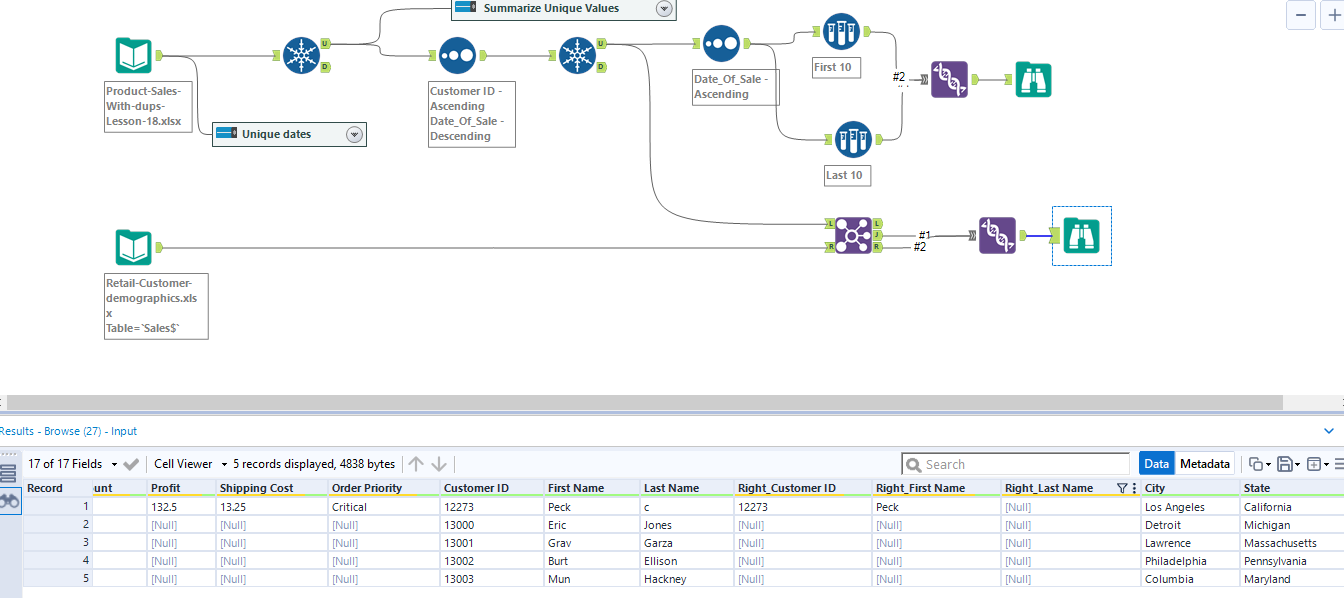
Now, the marketing team wants to see the top 10 (mostly active) and bottom 10 customers (mostly inactive). We will use the ‘Sample’ tool to see the first 10 rows and the last 10 rows. And finally we chose the ‘Union’ tool for a final consolidated table with total number of 20 rows.



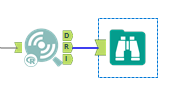
In the below canvas, we have created a workflow pulling our demography file on the canvas and joined by Customer Id with the unique values of the product sales file. In this workflow, customer ID’s will be distributed between the J and the R output nodes. The J node contains customers ID’s who appear in the both tables while the R node contains ID’s which only in the retail-customer-demography table. Now, we have created the join by adding a Union tool which connects the J and R nodes to it as inputs.

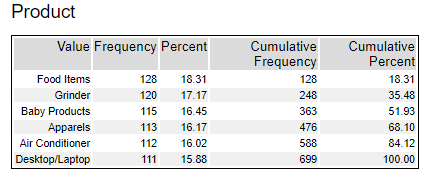


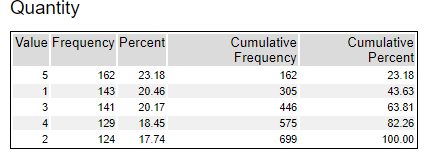
Our right outer join has four customers in the retail-customer-demography table which are missing in the product-sales table. A data analyst needs to investigate with sales or operations department who are putting information in the CRM system regarding those missing values.



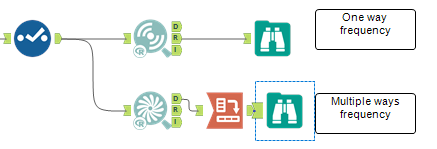
We now want to see the frequency of our product and quantity from our unique data table. Drag the frequency tool on your canvas and you will see the categorical variables are appearing on the configuration panel.

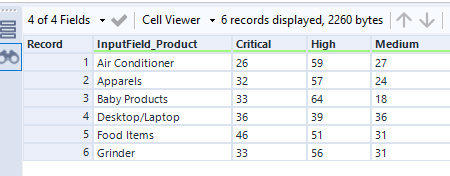






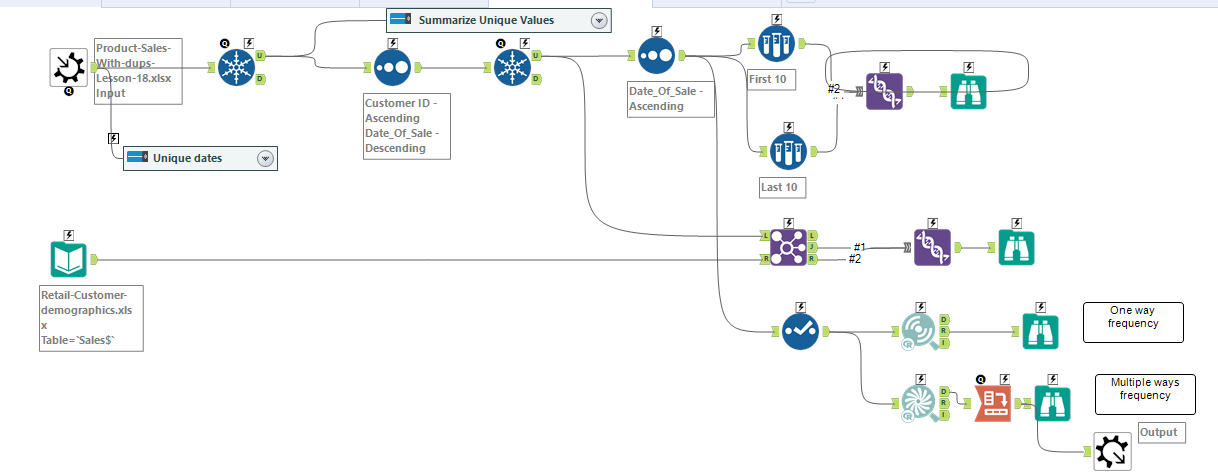
Now we will run frequency with multiple variables (N way frequency) using Contingency tool and cross tab to see the number of order priorities based on critically per product category.



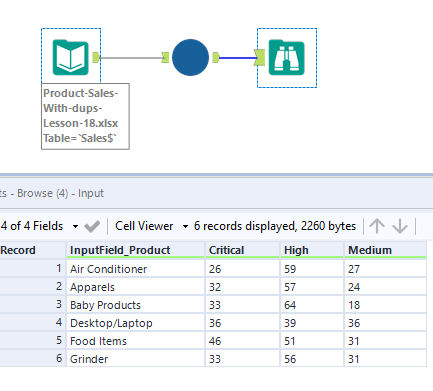


A macro is a workflow or group of tools built into a single tool that can be inserted into another workflow. We can save the whole workflow as file type .yxmc

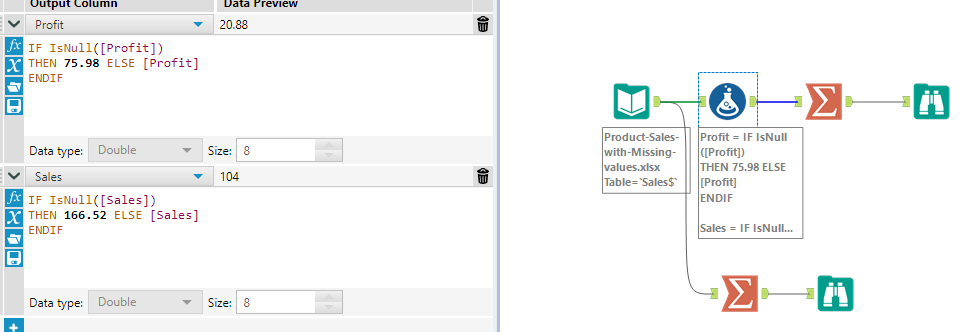
You can select your data file and right click and select ‘Convert to Macro Input’ and the output you can type on the search bar ‘Macro Output’ tool.



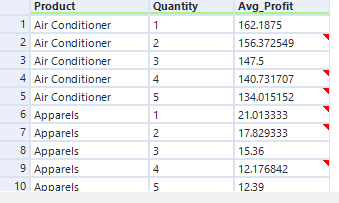
We have connected the macro we have created before and now connected with the data file. The output populated same as the previous result. You do not need to create a workflow every time, just use the saved macro when data table is updated with new data.

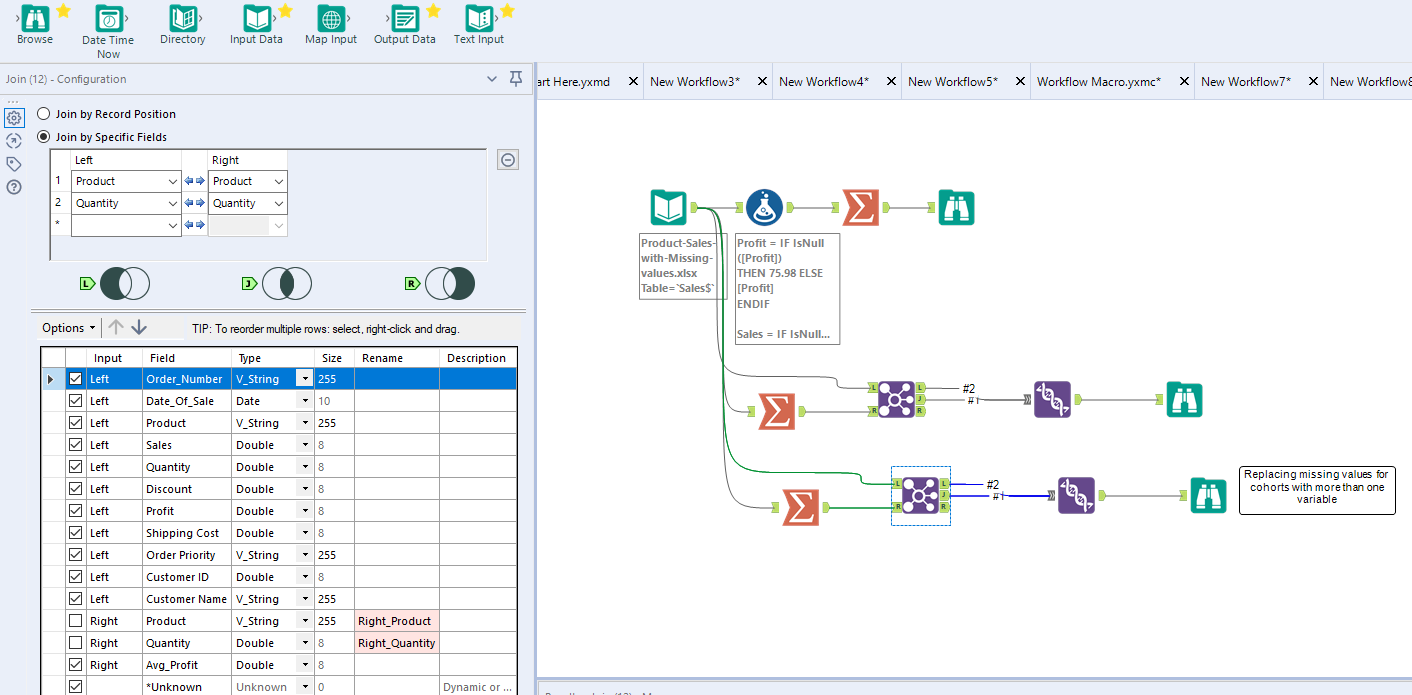


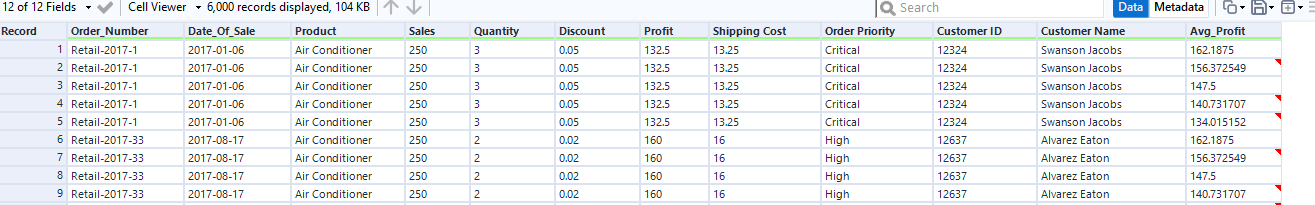
Now we will discuss the possible ways to treat missing values. If we have missing values in profit column, we can average all the sales and then use the average sales values in null fields. This is a simplistic way of filing all missing values.



We can now also replace missing values with multiple variables. In the below, we have joined data by product and quantity (various quantity volume sold had different profit average). Therefore, the missing values (null) in the profit column are replaced by the average profit grouped by product and quantity.







**What is Analytics?**

Analytics is exploration and interpretation of data to find out more meaningful and insightful patterns that can be applied to detect business problems and predict with descriptive and predictive methodologies to improve the business performance. Analytics can be classified into four broader categories:

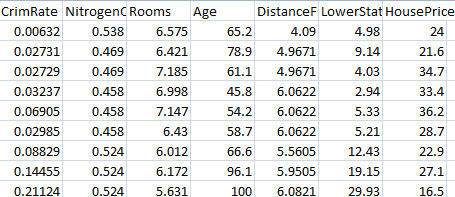
Descriptive analytics (what happened?)

Diagnostic analytics (why it happened?)

Predictive analytics (what will happen?)

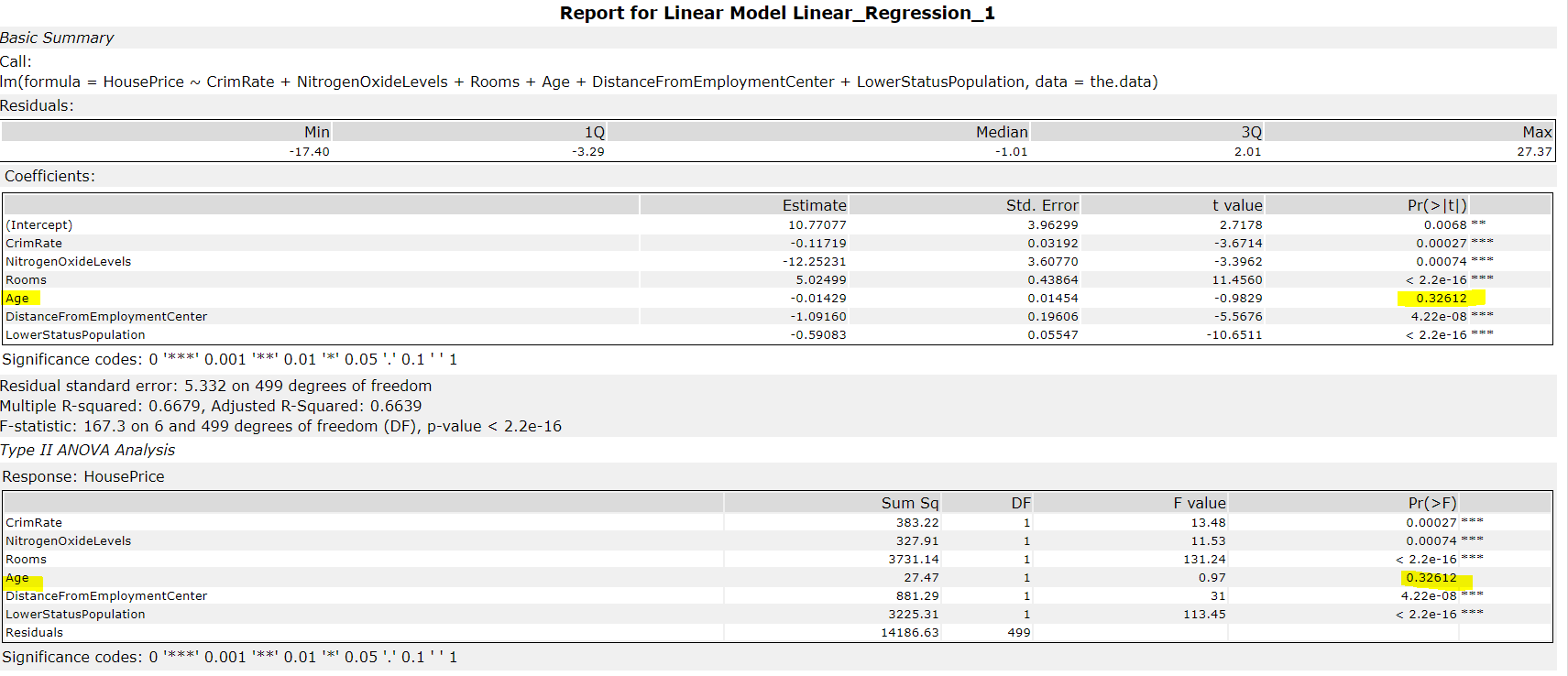
Prescriptive Analytics (what should happened?)

We will perform a liner regression model on multiple variables. Liner regression is used for predicting continuous variables. We want to see how the target variable (house price), we are trying to predict the relationship with multiple predictor independent variables.

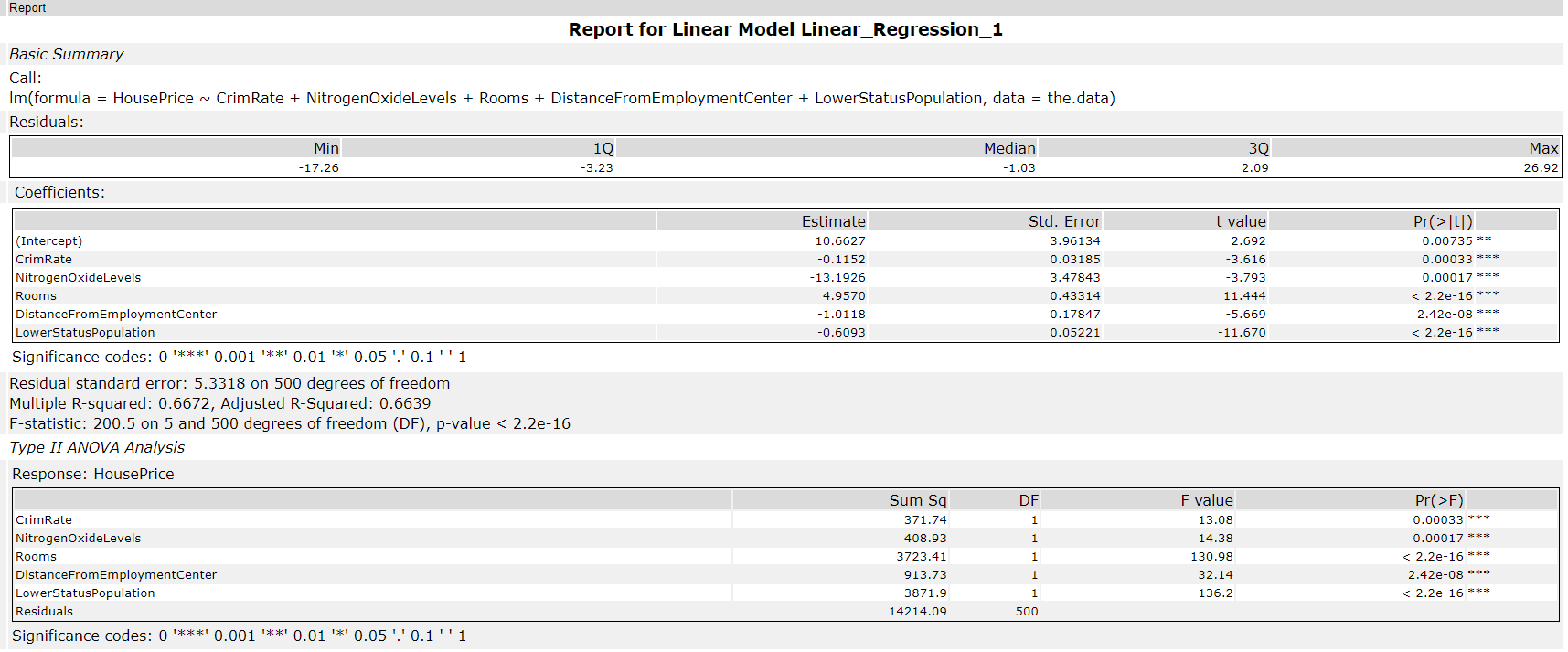


The R is liner regression report and O stands for Liner regression output in the Linear regression tool.

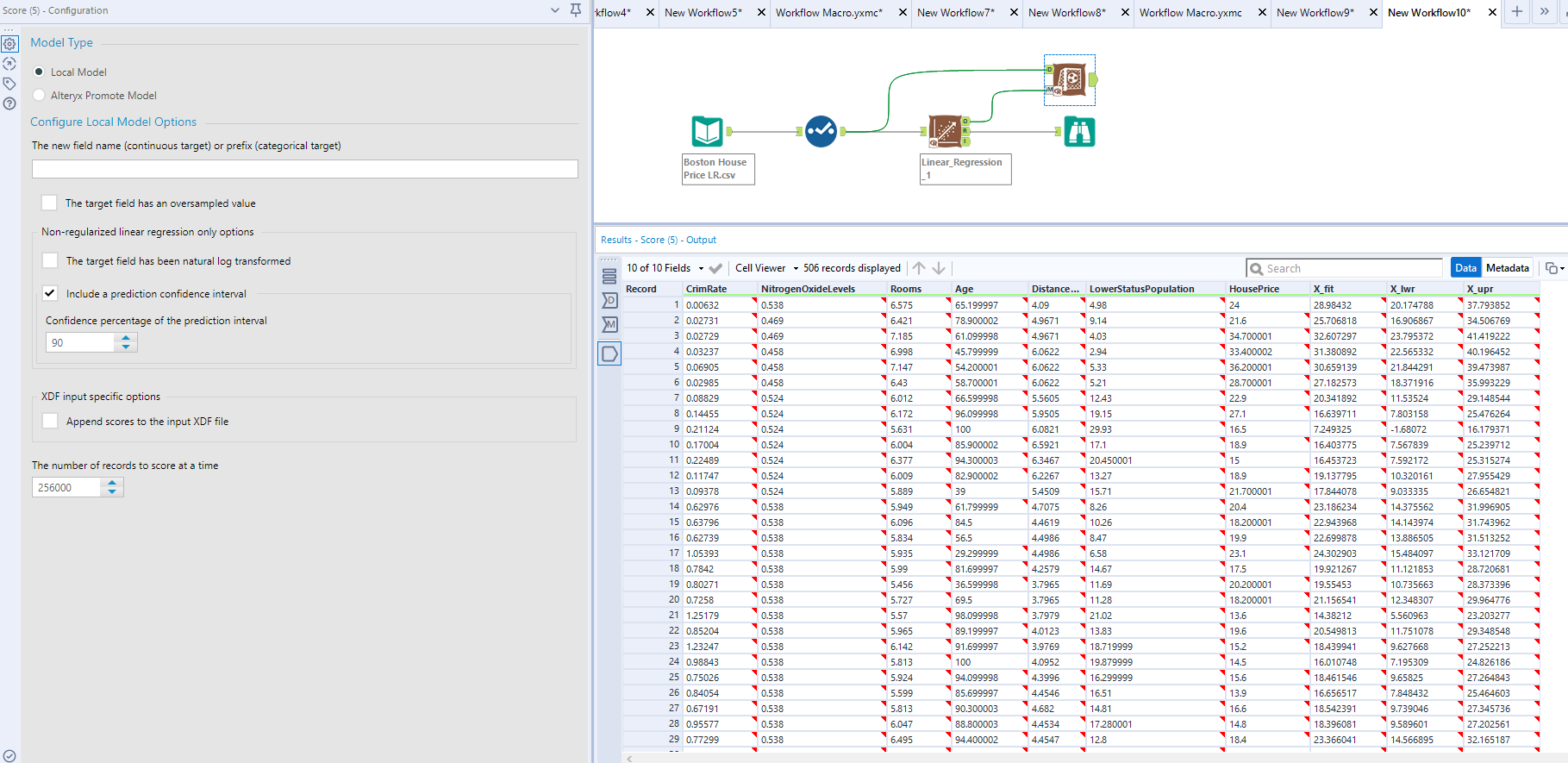
In our first model from the populated model created by R programming in the background, we can take only those variables which are significant (p value is more than 5%). Therefore, we need to remove the Age independent variable form our model as age does not have any impact on the price of any house. The R square is .66 which means the model accuracy is 66% which can go up to maximum 100%



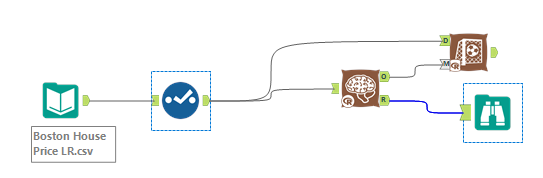
The new model excluding the age variable where we can see stars which represents the significant level.

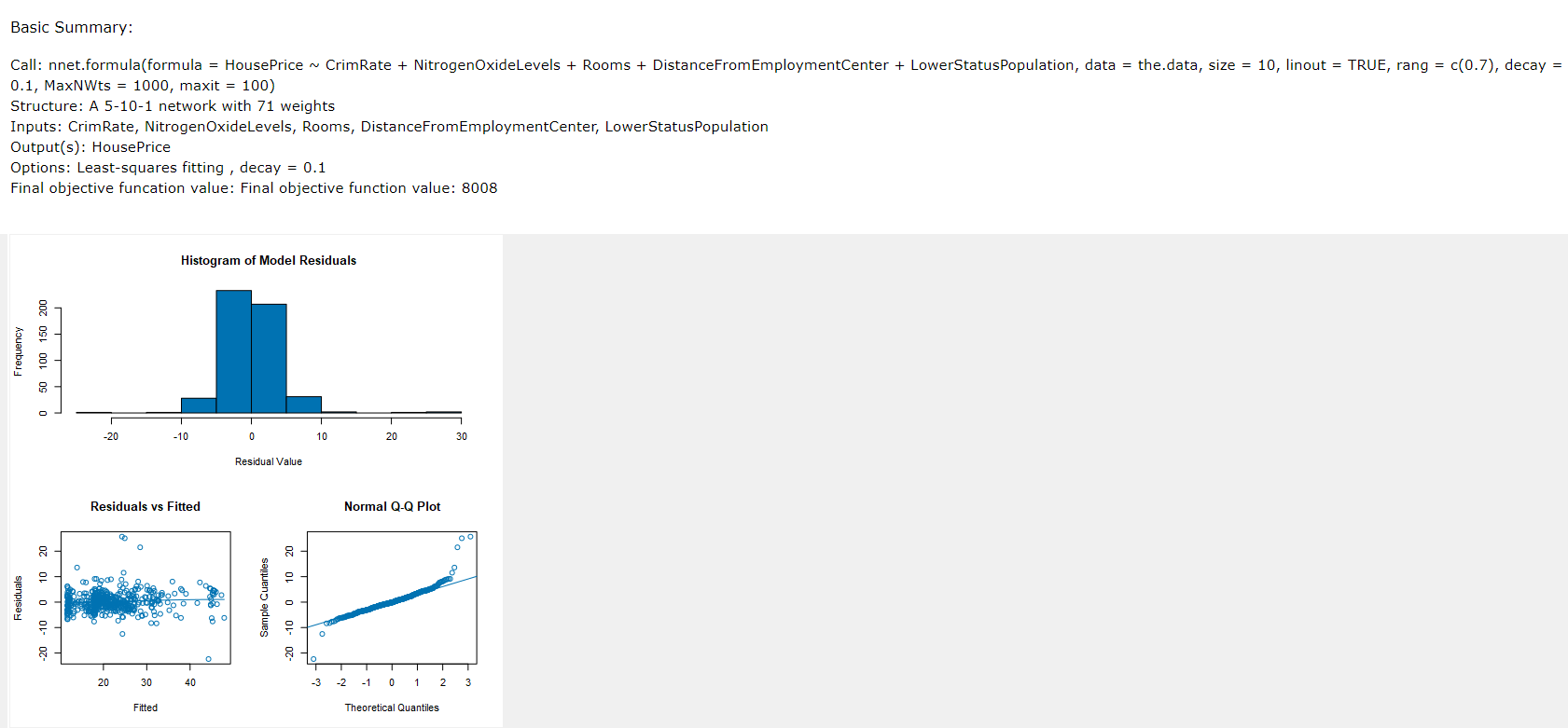


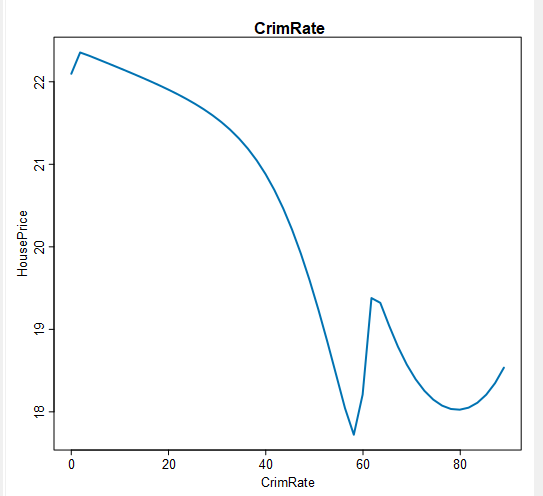
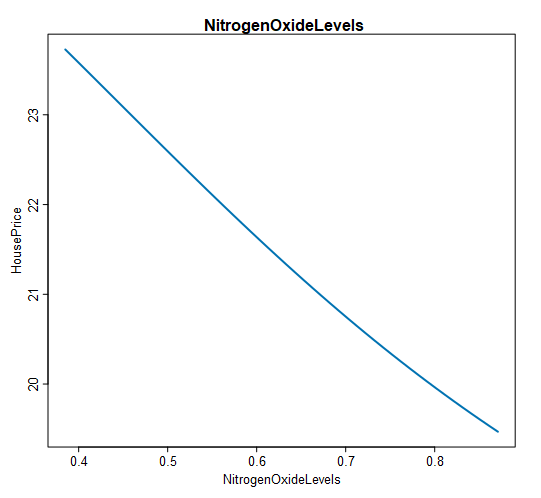
Now, we have added the tool Scoring with 90% confidence level to generate the scoring for house prices.

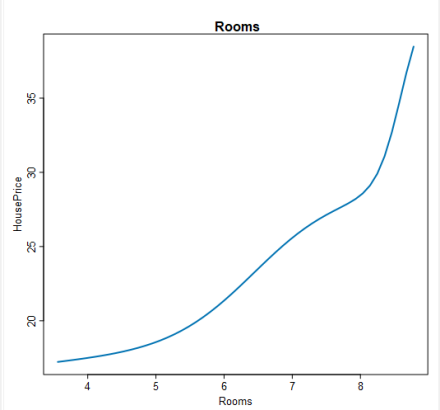
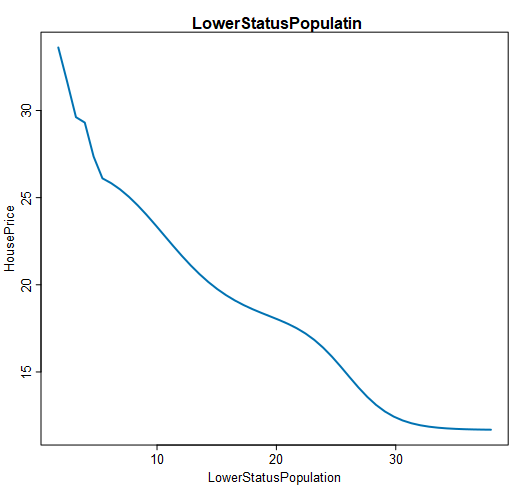
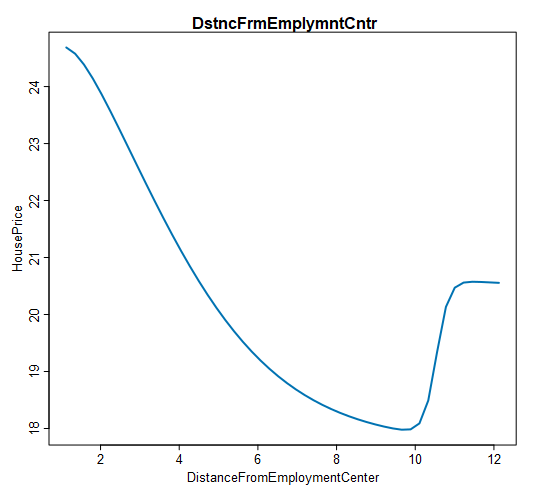


We have made a neural network model based on the house price data.

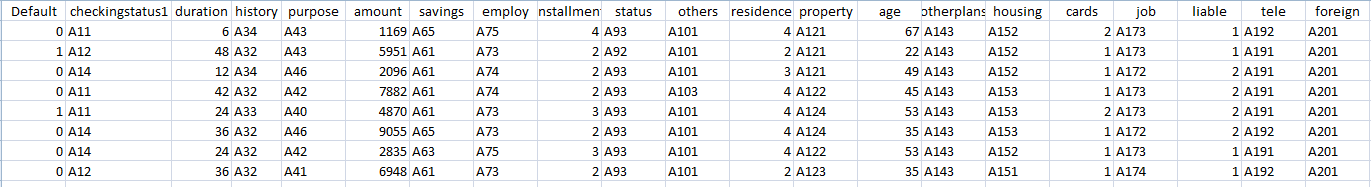




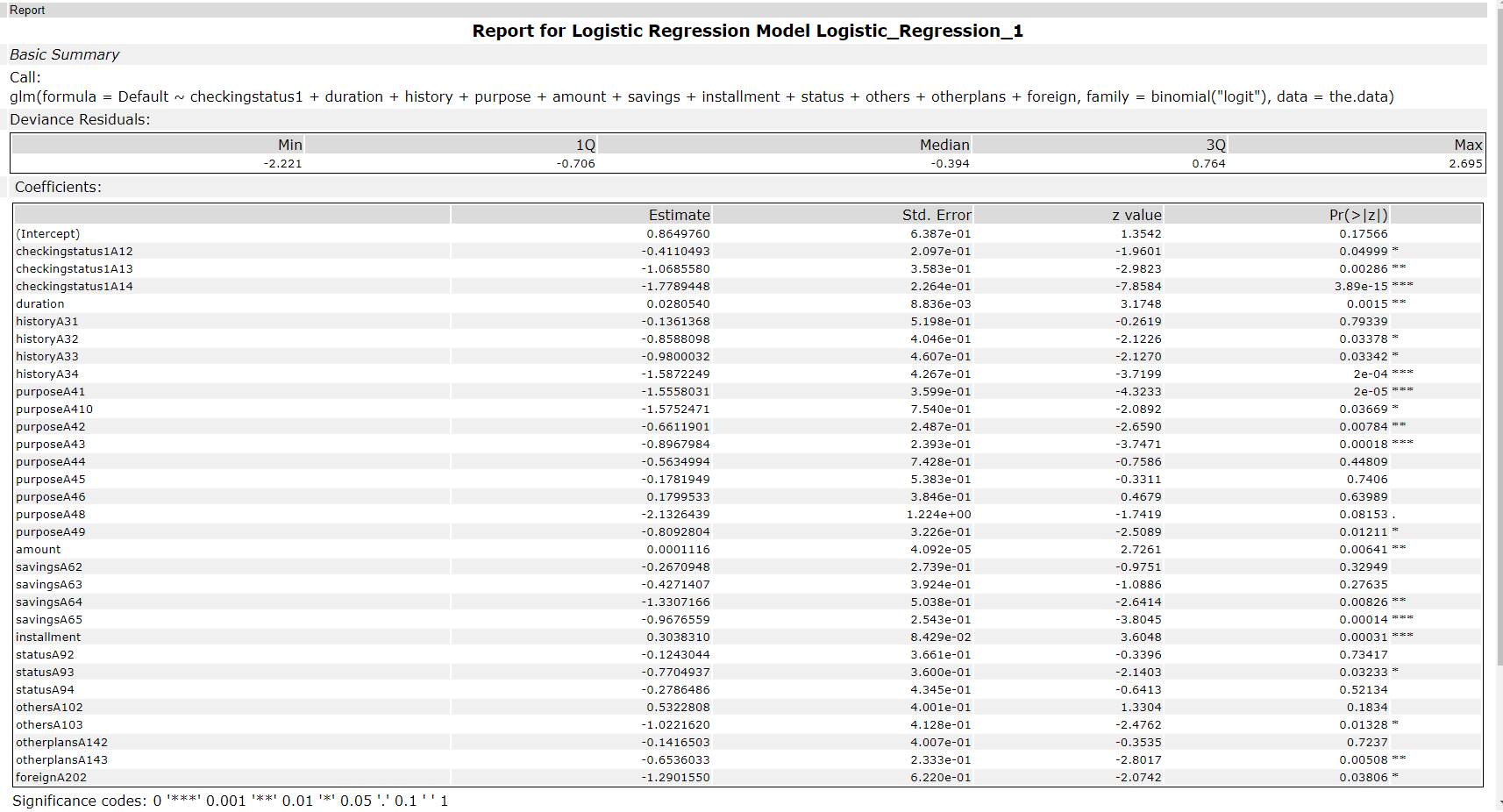
 

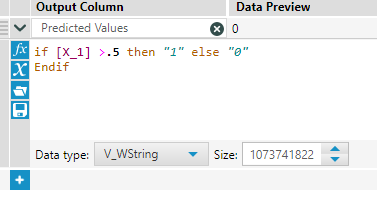
Now we will create a logistic regression model where target variable is categorical or binary outcome. In our below example, we have a bank’s customer data. The default column represents binary outcome which means if a customer is not in default ‘0’ or ‘1’ when in default. We also have other independent variables (X variables) which will help us to score our predicted value given the default variable (Y variable) through the logistic model.

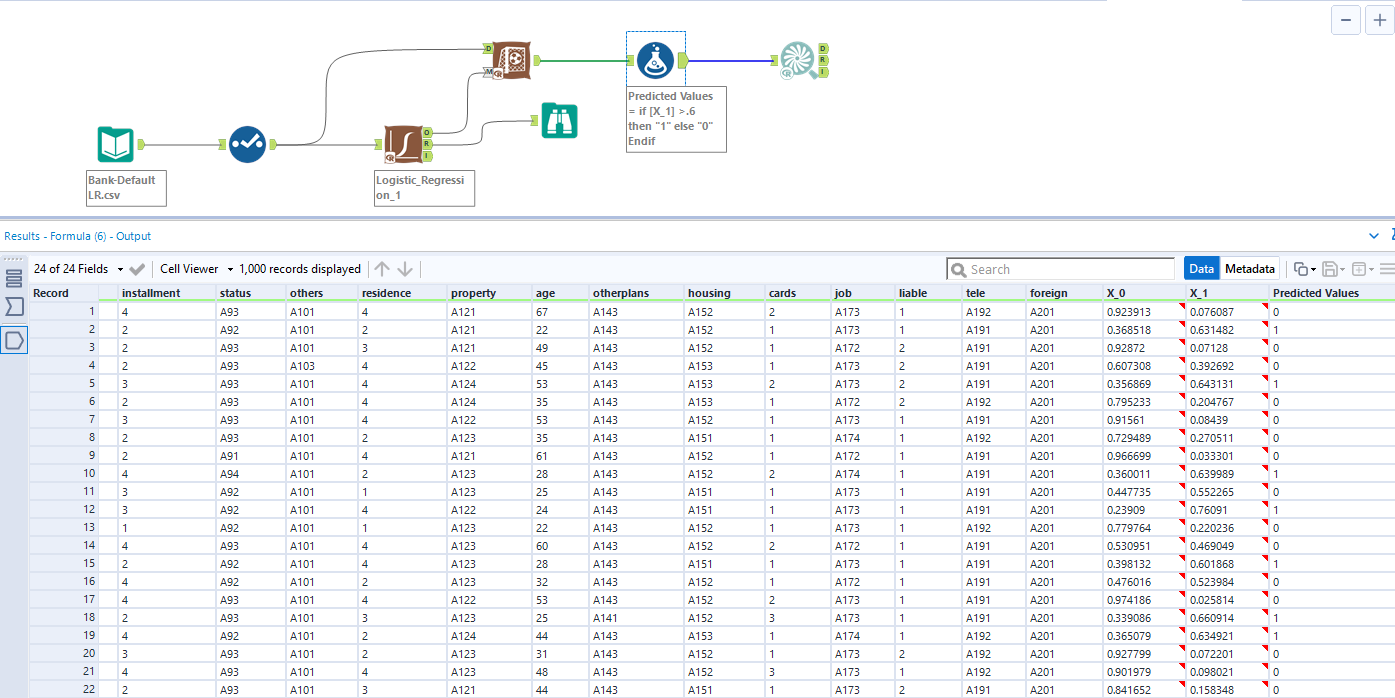


We have performed the logistic regression and the same way we have removed all the insigficant variables where the p value is more than .05 and got the below output.

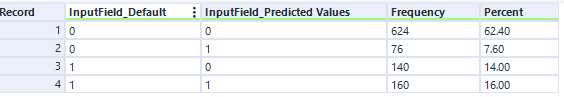


Now we will be adding Scoring tool and it has created two columns, X\_0 (the probability of the customer will not be in default) and X\_1 (the probability of the customer will be in default). The total of two columns will be 100%. We are going to make a rule that if the value in X\_1 column of each field is more than .5 then the customer will be in default and the output has appeared in another new column called ‘Predicted values’ which has created based on the rules mentioned above to determine the customer will be in default or not.





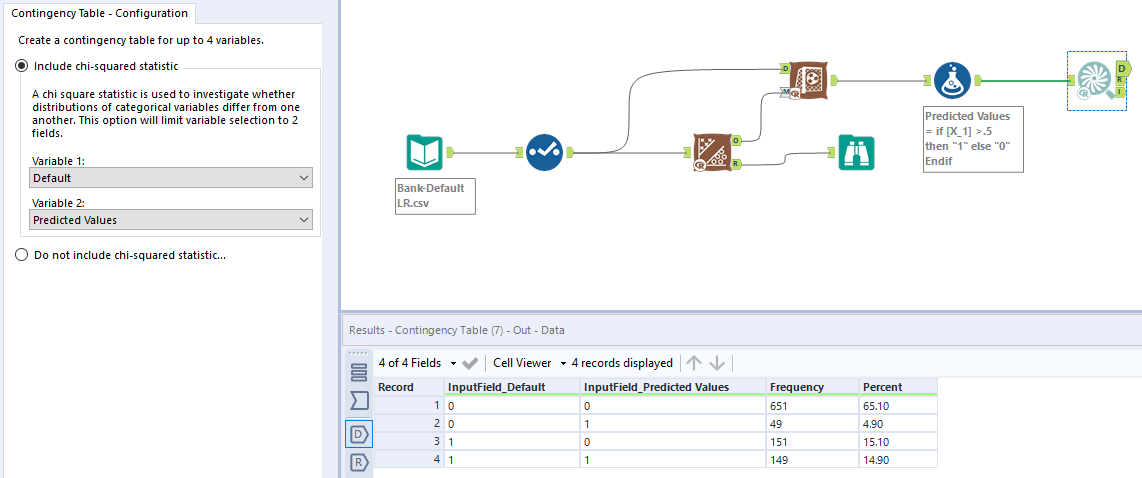
We can run a contingency table on the default (actual) and the predicted value (predicted) and we can see the accuracy of the model.



The accuracy score for our model = (624+160)/1000 = 78%

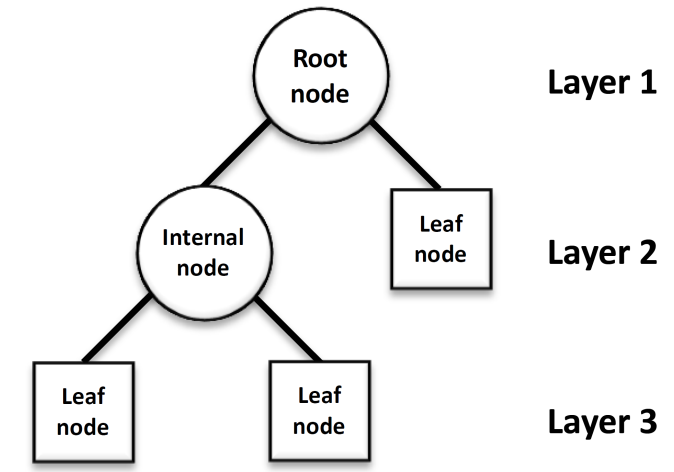
We will now discuss two most popular machine learning algorithm, support vector machine (SVM) and Decision tree. SVM is a classification algorithm which is generally used to predict the categorical variables. In SVM, the data is partitioned on the basis of a hyperplane. The partitioned data is two different groups where each one is representing a prediction of Y variable. An ideal hyperplane have all data points farther away from each other. A hyperplane can be liner or non-linear (like polynomial or Sigmoid function).

We have performed SVM algorithm below just replacing the Logistic Regression tool and in this case our accuracy has come bit higher to 80%.



We can take this SVM as a final model in our production environment.

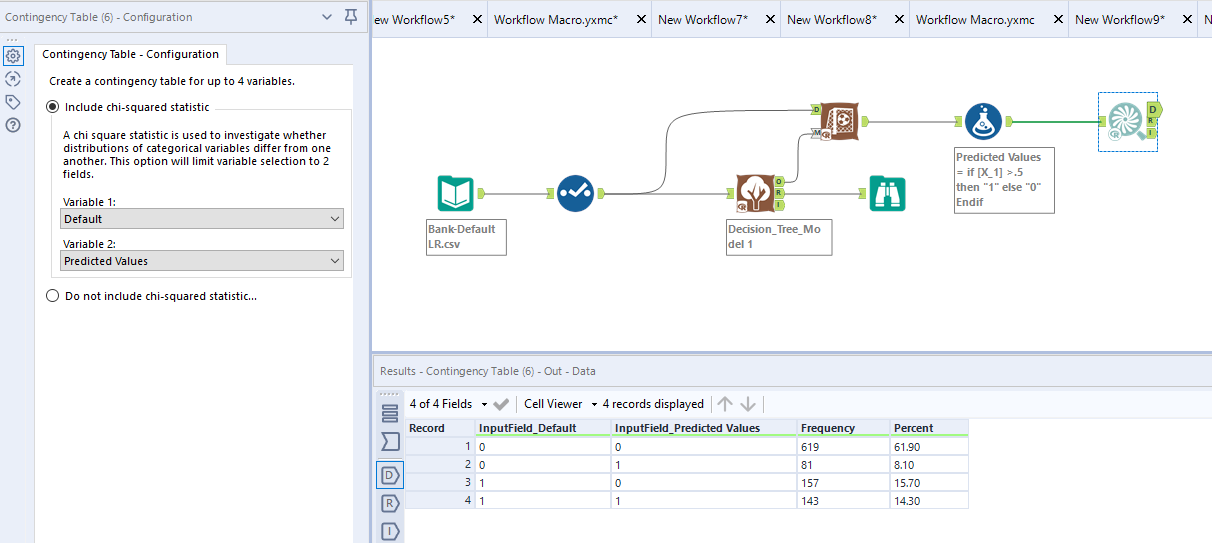
A decision tree is an another machine learning algorithm which uses a tree like model of decisions and their possible consequences including event outcome by chance. Root node represents the top node which asks a question or performs a logical test on each attribute and goes to the next layer through branches to internal nodes. If the logical test finds the parameter to pass through another layer of validation then it goes to the next level down. Finally, it has to find the leaf node for the outcome.



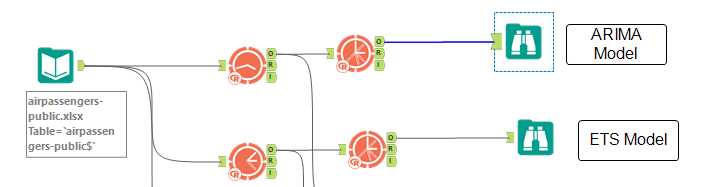
The decision tree plot of our data



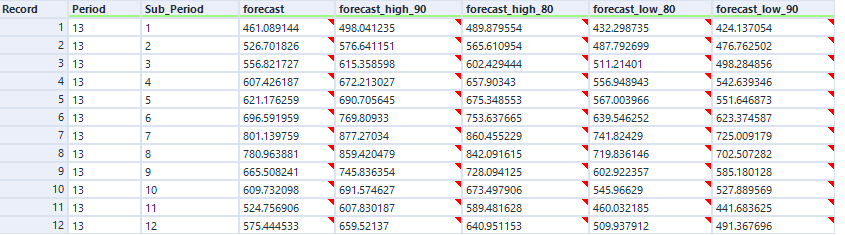
The accuracy score we have found with the decision tree model is 76.3% which means decision tree model is giving us lesser accuracy comparative to the logistic regression model and the SVM model.



We will do ARIMA and ETS forecasting models for the next 12 months based on the historical data of number of passengers will be travelling. We will get the forecasting results in the range of 90% and 80% confidence intervals.



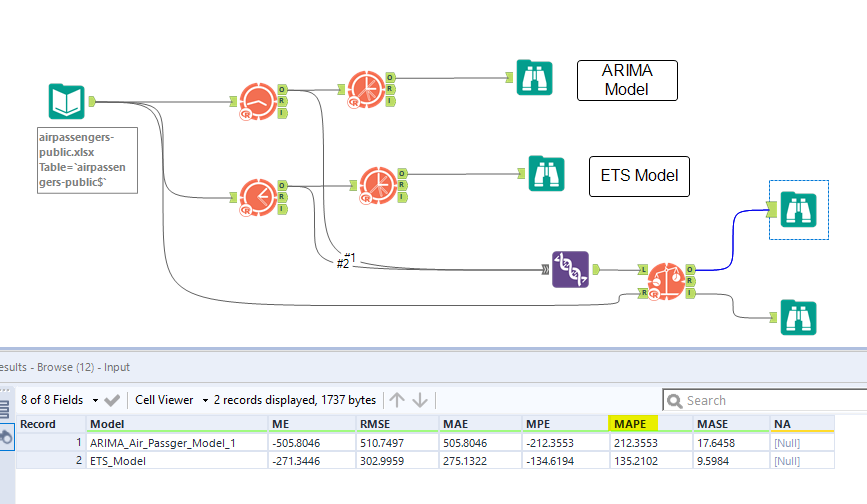
ARIMA



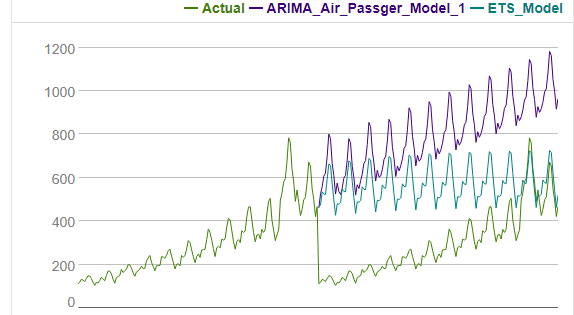
ETS



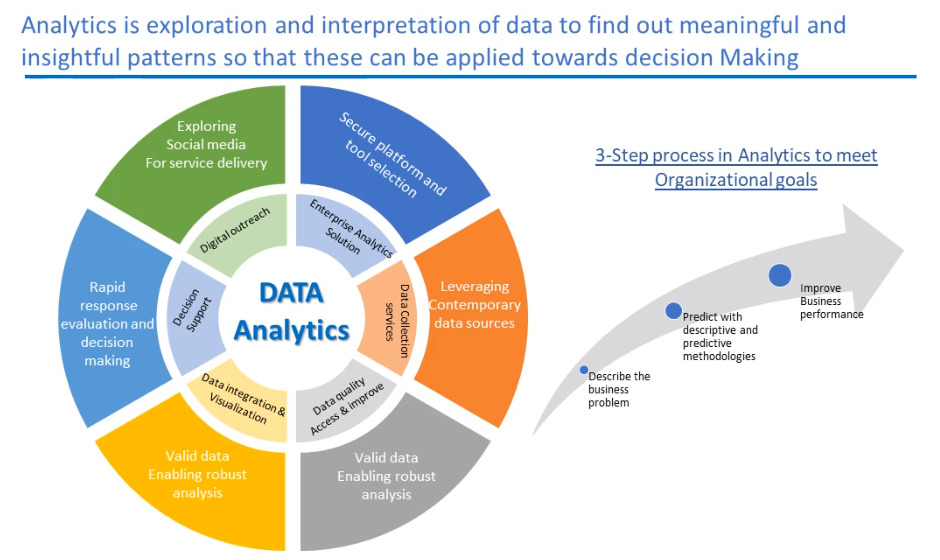
Now we will combine both models to find the best forecast data in respect to the air passenger data with the TS Compare tool. The TS compare tool compares the two time series modeling techniques. The output compares the MAPE (mean absolute percentage error) of the outcome.

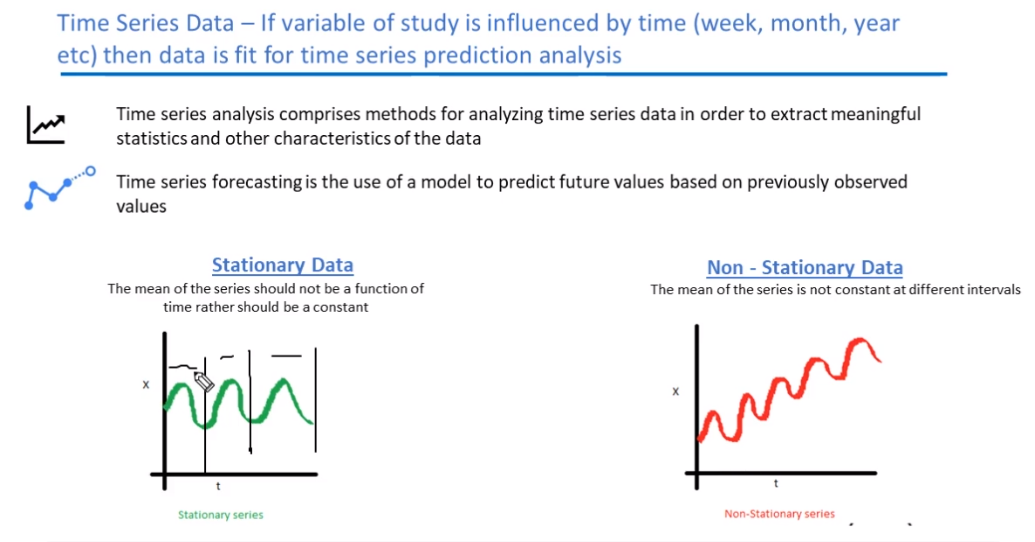


We will look into the MAPE and ETS has the lowest value. Since MAPE is a measure of error, high numbers are bad and low numbers are good. Therefore, we will choose ETS model for our better prediction accuracy.









Global Constants: Make a dynamic process with input  
constants as dynamic

