RESPONSES TO AN ANALYST CANDIDATE TEST

http://competitiveanalytics.com/employment/business-intelligence-analysts/

Exercise 1: Data Prep Qualtrics Survey Parsing

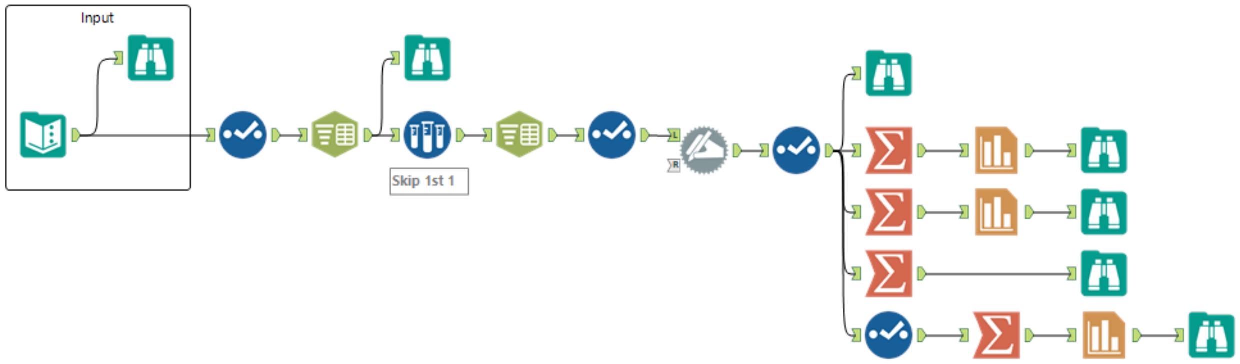
Introduction

The source contains results from a download tool. The data is concatenated in the "DownloadData" field.

The objective is to parse the data for analysis.

Methodology (Alteryx workflow)

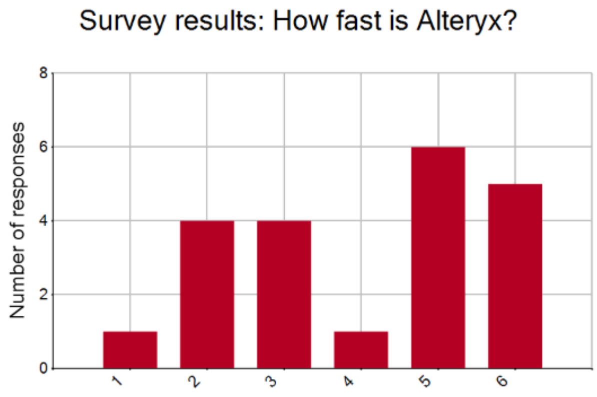
<u>Tools</u>: Select, Text to Columns, Sample, Dynamic Rename, Summarize, Charting

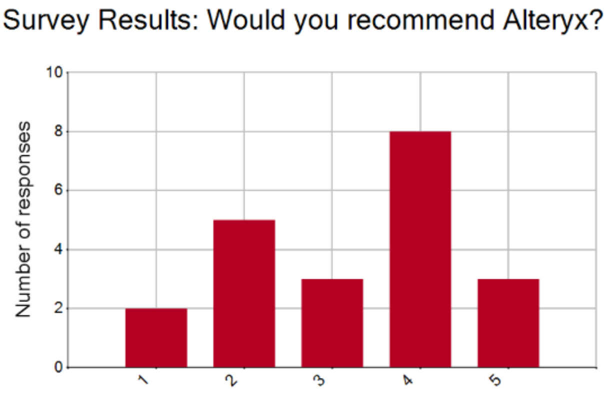


1. Select the DownloadData column.
2. Split the DownloadData field into rows using Text to Columns and the new line ("/n") delimiter.
3. Skip the 1st row using the Sample tool.
4. Split the data into 15 columns using Text to Columns and the comma delimiter.
5. Select only the relevant fields (first 12 split columns).
6. Rename the fields with the first row of data using the Dynamic Rename tool.
7. Change survey responses from text (string) to numeric (int16) in order to calculate mean, etc.
8. Summarize and chart the survey results.

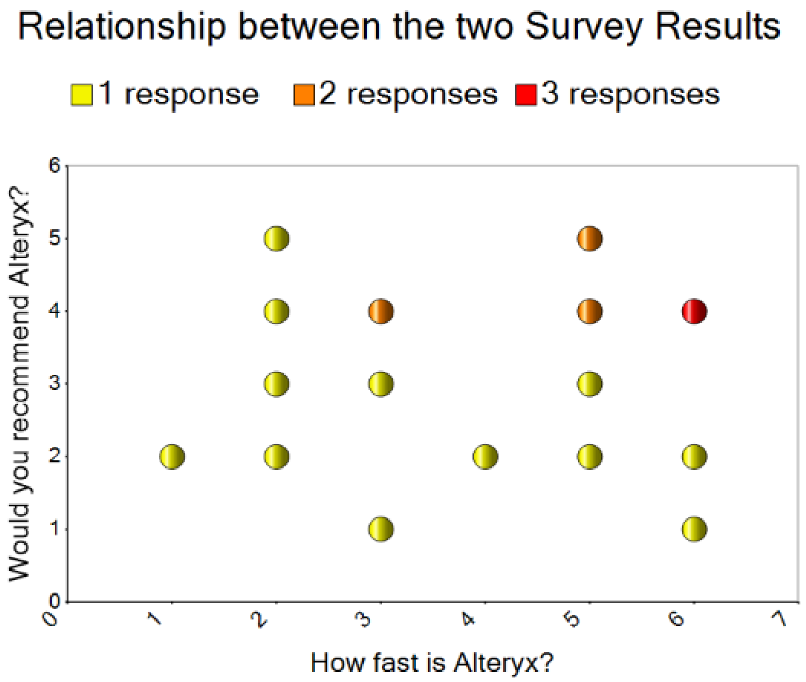
Results

There are 21 responses to two questions regarding Alteryx. The average responses regarding Alteryx' speed are 4 (mean) and 5 (median). The average response regarding whether one would recommend Alteryx or not was 3.2 (mean) and 4 (median). However, we do not know the rating scale.





A scatter chart of the survey results does not show any clear correlations between responses to the two questions.

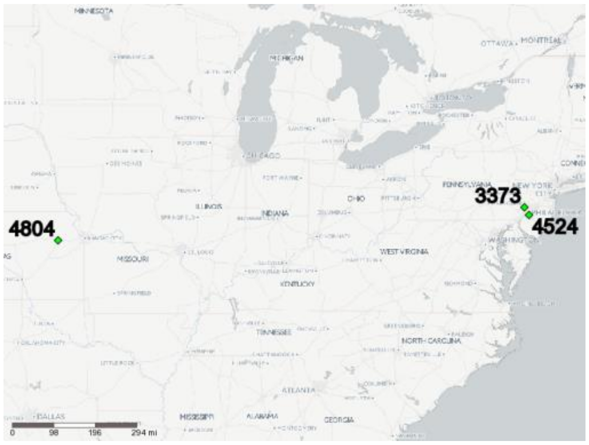


With a greater number of responses, a positive correlation might be observed since speed is most likely one of the criteria used for software recommendation.

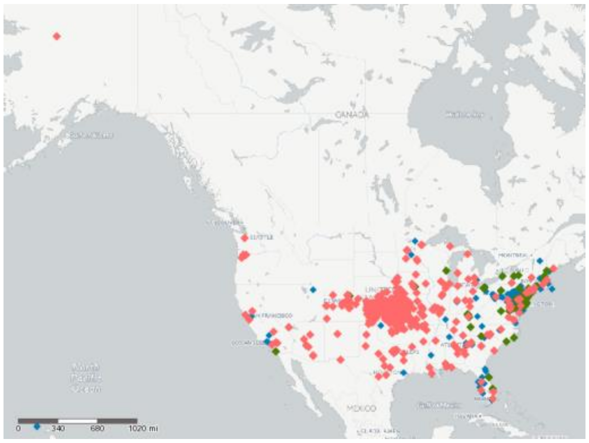
Exercise 2: Spatial Trade Area Creation

Introduction

There are 3 stores located near Kansas City (Store ID: 4804) and Philadelphia (ID 3373 and 4524).



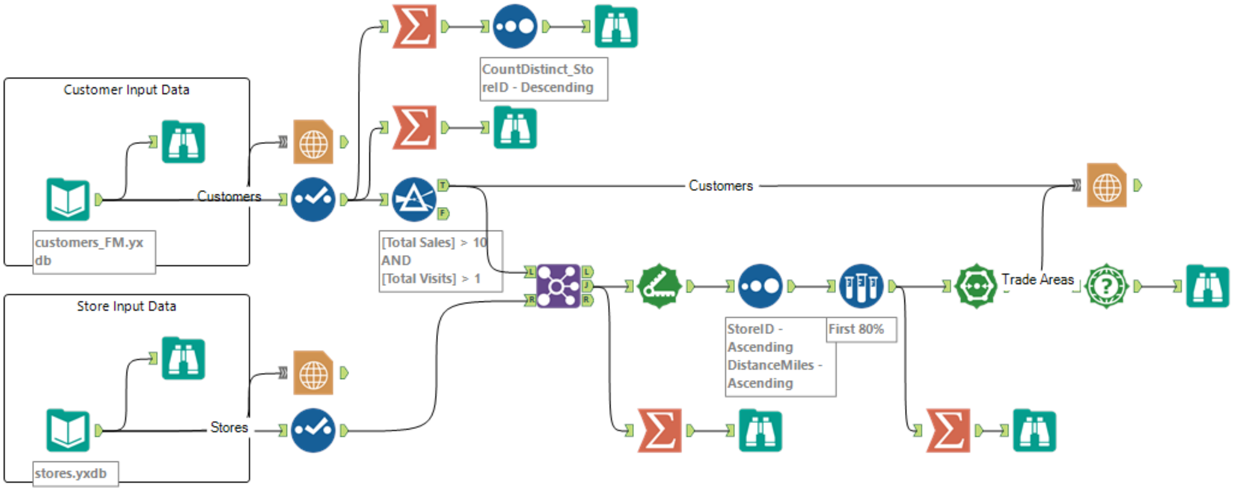
There are 18,538 unique customers located throughout the US, including Alaska and Hawaii.



The goal is to determine the convex hull trade area of each store that encapsulates 80% of the customer base who spend more than $10 and have had more than 1 visit for each store.

Methodology (Alteryx workflow)

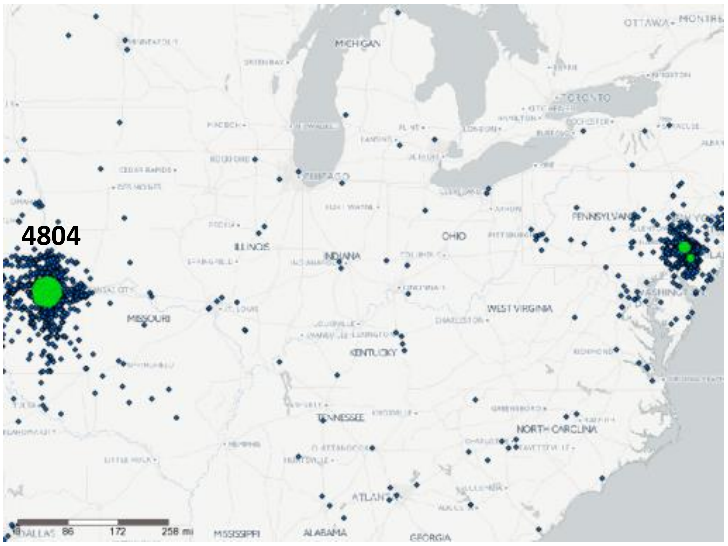
<u>Tools</u>: Filter, Select, Join, Distance, Sort, Sample, Poly-Build (Convex Hull), Spatial Info, Report Map


Customer Input Data
customers_FM.yxdb
Customers
Store Input Data
stores.yxdb
Stores
CountDistinct_StoreID - Descending
[Total Sales] > 10 AND [Total Visits] > 1
Customers
StoreID - Ascending DistanceMiles - Ascending
First 80%
Trade Areas

1. Filter customers with more than $10 in sales and more than 1 visit. We are left with 14,923 pairs of customer-store. Note: Two customers (ID 27718927 and 40882578) visit two stores.
2. Join the customer and store database (inner join). Note: Change the type of Store ID in both databases from Double to Integer in order to avoid roundup issues when joining on that field.
3. Calculate the distance between each customer and the store they visit.
4. Sort customers by increasing distance to store, and sample the first 80% for each store.
5. Create a convex hull polygon centered around each store using the Poly-Build tool.
6. Extract the area square miles for each trade area using the Spatial Info tool.
7. Use the Report Map tool to visualize the locations of customers, stores, and their trade areas.

Results

Map showing the location of ‘qualifying’ customers (in blue) and each store’s trade area (in green)

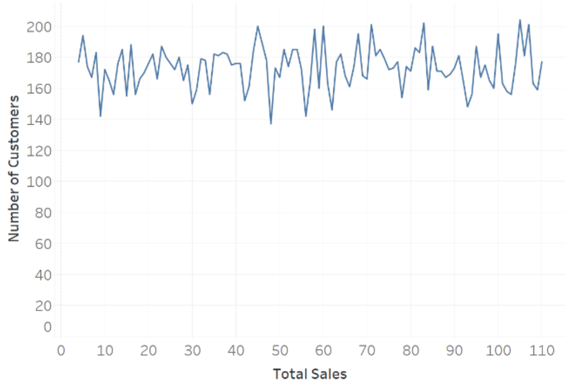


Store ID	Trade Area	All ‘qualifying’ customers			Closest 80% ‘qualifying’ customers		
		Customers	Sales	Visits	Customers	Sales	Visits
3373	173 sqm	4,333	290,542	12,288	3,466	232,369	9,826
4524	60 sqm	2,815	190,790	8,016	2,252	152,713	6,426
4804	1,747 sqm	7,775	526,360	22,079	6,220	419,514	17,635
Total	**1,980 sqm**	**14,923**	**1,007,692**	**42,383**	**11,938**	**804,596**	**17,635**

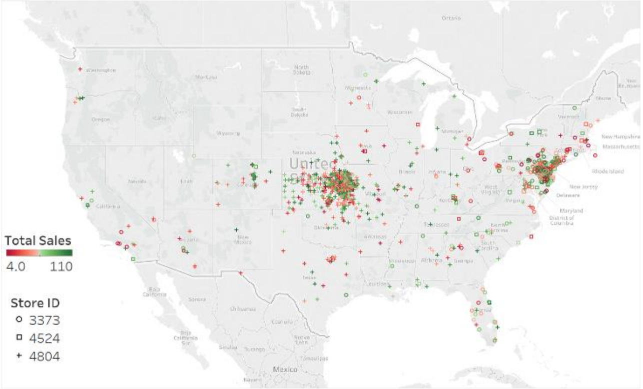
Discussion

The 80% closest customers represent about 80% of sales and visits because the distribution of sales is approximately uniform (between $4 and $108), and customers with different sales amounts are distributed relatively uniformly around each store.

Distribution of Total Sales



Distribution of Customers per Location, Total Sales and Store



** Charts made with Tableau*

The Kansas City store has a much larger trade area than the Philadelphia stores because it has more customers in a less dense area. Population density is much lower in Kansas City (about 1,500 people per sqm) than in Philadelphia (about 11,000 people per sqm).

Alternative trade areas could be determined using drive time instead of distance.

Exercise 3: Preparing Sales Data for Visualization

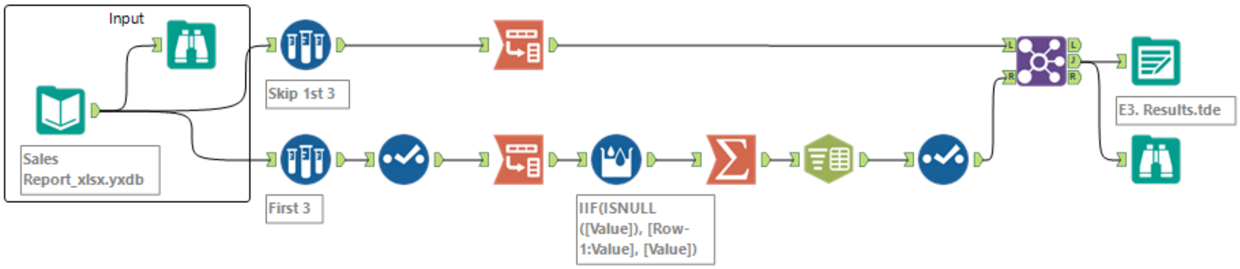
Introduction

The source contains sales data for two months, two reporting types and various classes of businesses in a crosstab format.

The objective is to transform the data into a format that can easily be visualized in Tableau.

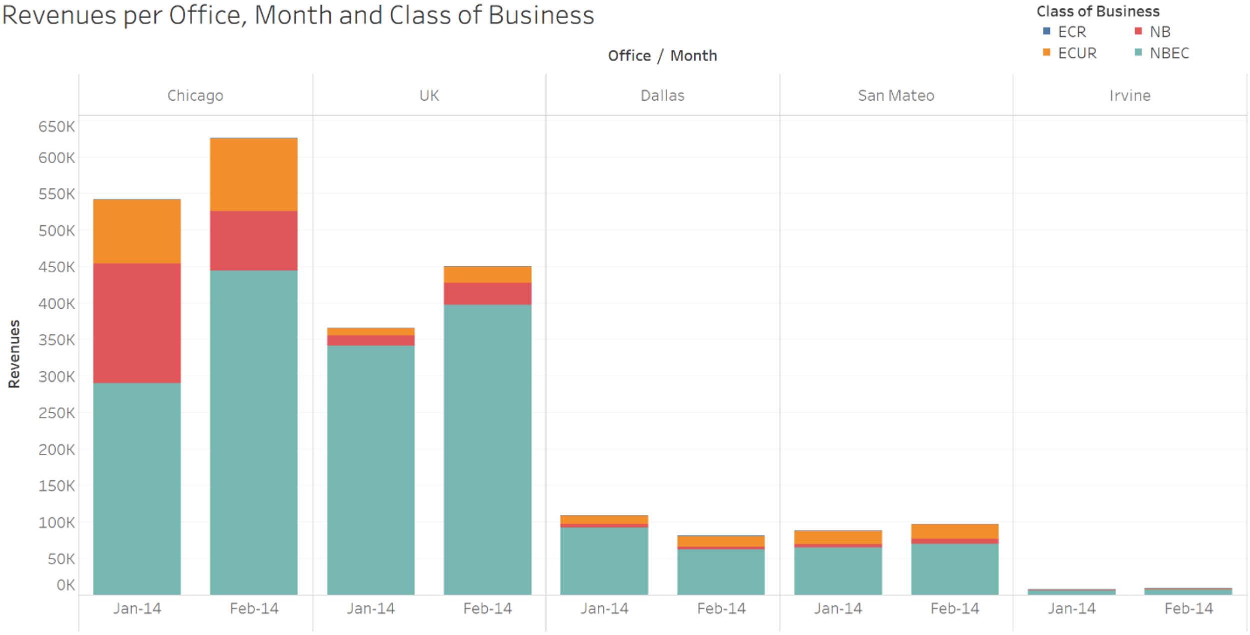
Methodology (Alteryx workflow)

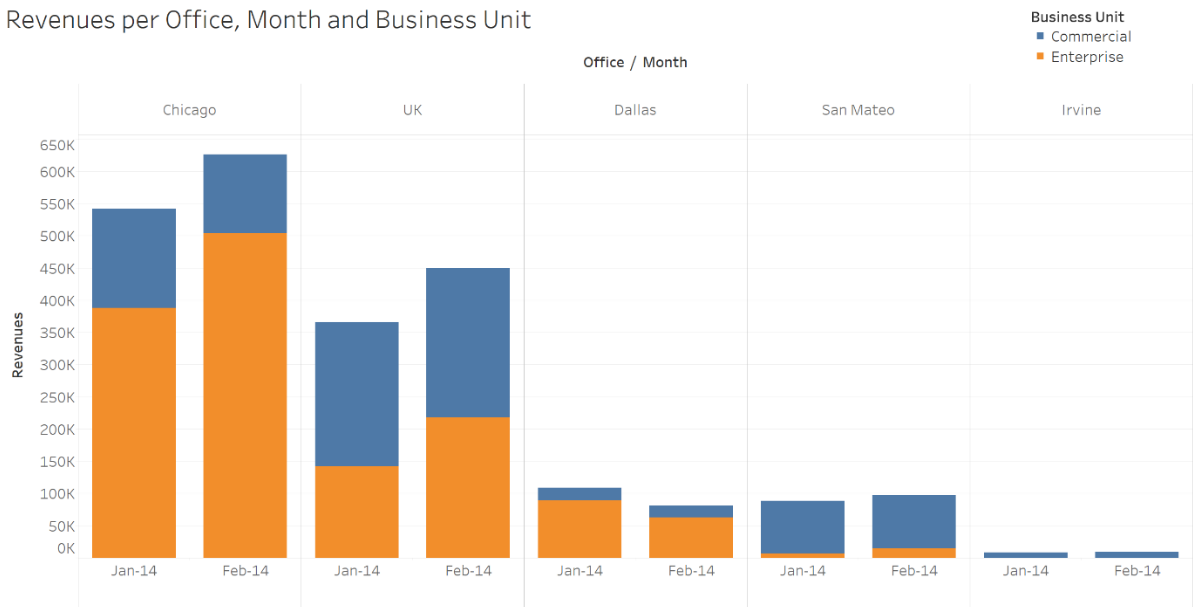
Tools: Sample, Select, Transpose, Multi-Row Formula, Summarize, Text to Columns, Union



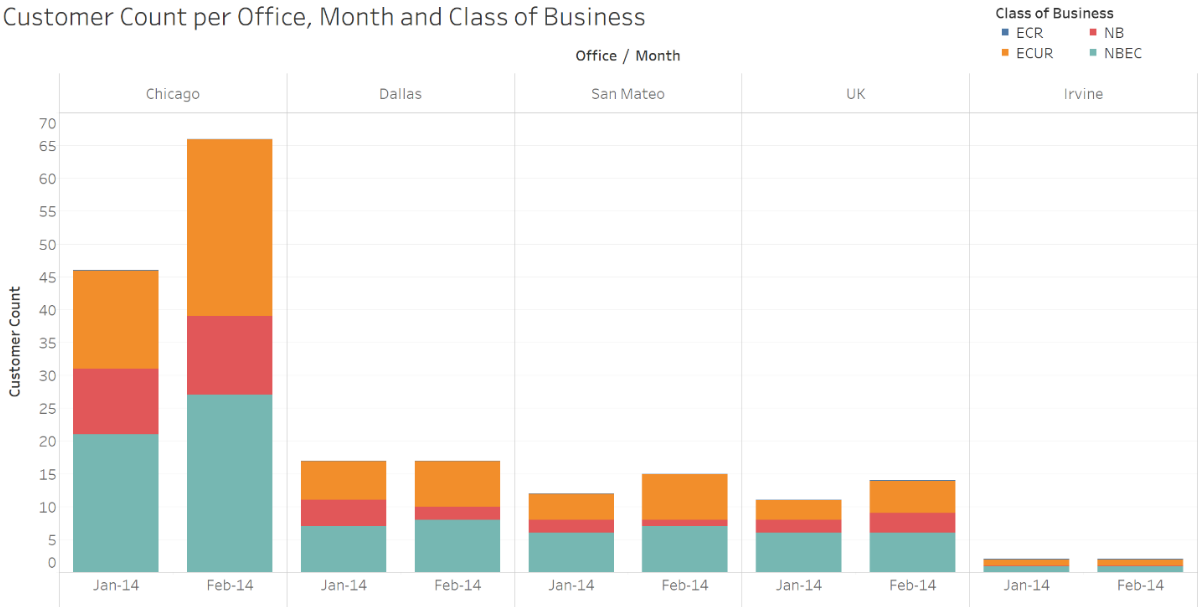
1. Separate the data from the headers (first 3 rows).
2. Transpose the data and the headers (except the first 3 fields).
3. Fill missing values in the headers using Multi-Row Formula:
   IF ISNULL([Value]) THEN [Row-1:Value] ELSE [Value] ENDIF
4. Group and concatenate headers using the Summarize tool.
5. Create new columns from the concatenated fields (month, type, class of business) using Text to Columns and Select.
6. Finally, join back the data and the edited headers (inner join). Change the type of Value from text (string) to numeric (integer).
7. Export the joined table to a Tableau Data Extract for visualization.

Results





Revenues per Office, Month and Business Unit



Customer Count per Office, Month and Class of Business

Discussion

The Tableau visualizations clearly show the following:

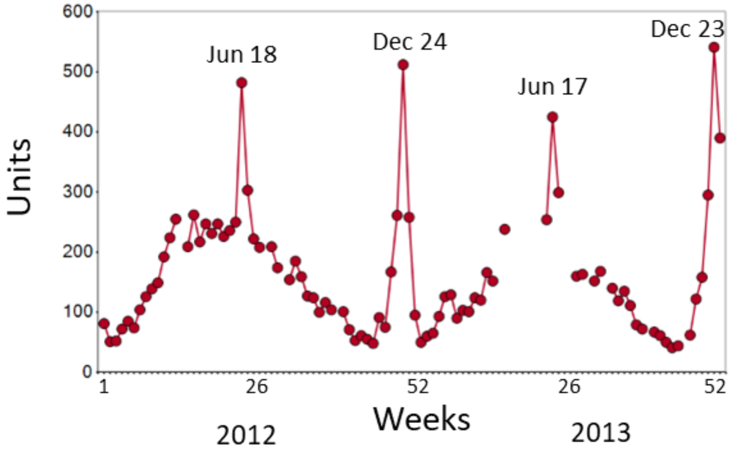
- The largest offices in terms of revenues are Chicago (~$600k) and the UK (~$450k), followed by Dallas and San Mateo (~$100k), and finally Irvine (~$10k).
- The UK office has about 4 times more revenues than the Dallas and San Mateo office even though it has slightly less customers.
- Total revenues increased from January to February for all offices but Dallas.
- NBEC businesses are the largest source of revenues for all offices. In terms of customer count however, there are about as many ECUR as NBEC businesses.
- Enterprise business units are the largest sources of revenues for the Chicago, UK and Dallas offices, while Commercial business units are more important for the San Mateo and Irvine offices.

Further analysis could include calculating average revenues per customer for each office, month, class of business and business unit, and analyzing the 51 different accounts.

Exercise 4: Time Series Forecasting

Introduction

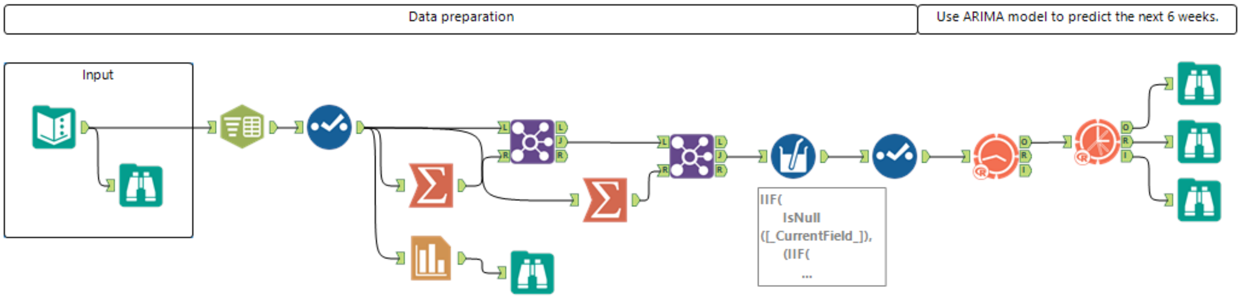
The source contains weekly data on number of units sold in 2012 and 2013. We can observe sharp peaks in the middle of June and end of December for both years.



The objective is to fill the missing data and forecast the number of units for the next 6 weeks (i.e., Jan-Feb 2014).

Methodology (Alteryx workflow)

<u>Tools</u>: Text to Columns, Select, Summarize, Join, Multi-Field Formula, ARIMA, TS Compare, TS Forecast



```
IIF(
   IsNull
([_CurrentField_]),
   (IIF(
   ...
```

Data Preparation

1. Transform the Fiscal Month column from String to Integer using Text to Columns and Select.
2. Calculate monthly and annual averages using the Summarize tool.
3. Replace NULL values by the monthly average, or annual average if the monthly average is also NULL, using Multi Field Formula:

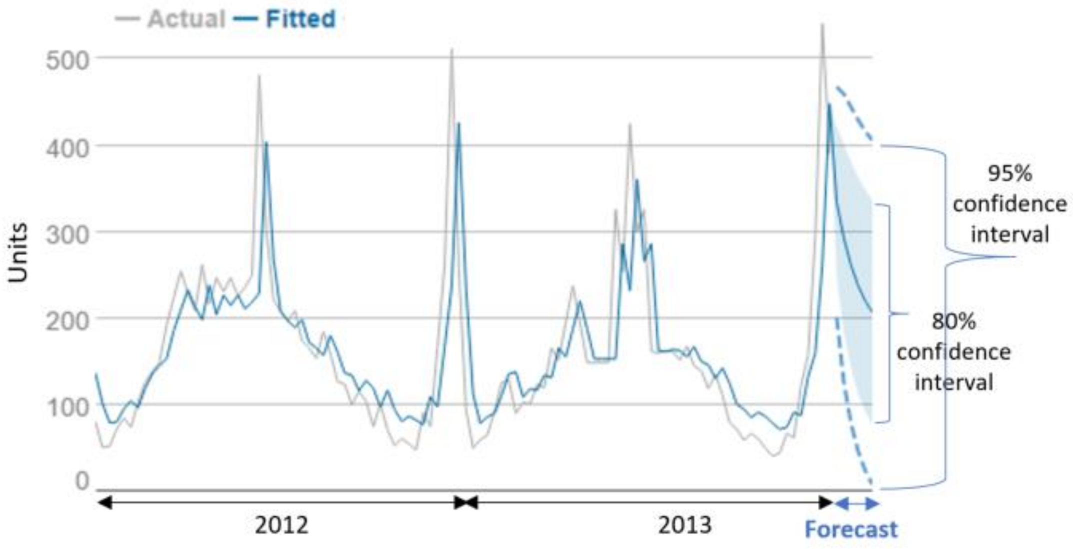
```
IIF(      ISNULL([_CurrentField_]),
          (IIF(ISNULL([Monthly_Avg_Units]), [Annual_Avg_Units], [Monthly_Avg_Units])),
          [_CurrentField_]
  )
```

Forecast

4. Use ARIMA model and TS Forecast to predict the number of units for the next 6 weeks.

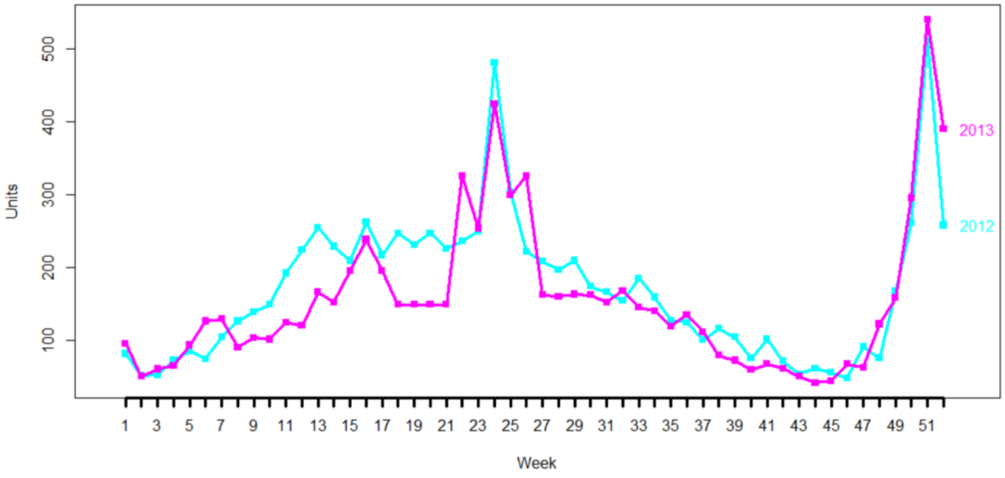
Results

1. **Forecasts:** The ARIMA model forecasts the following values for the next 6 weeks.

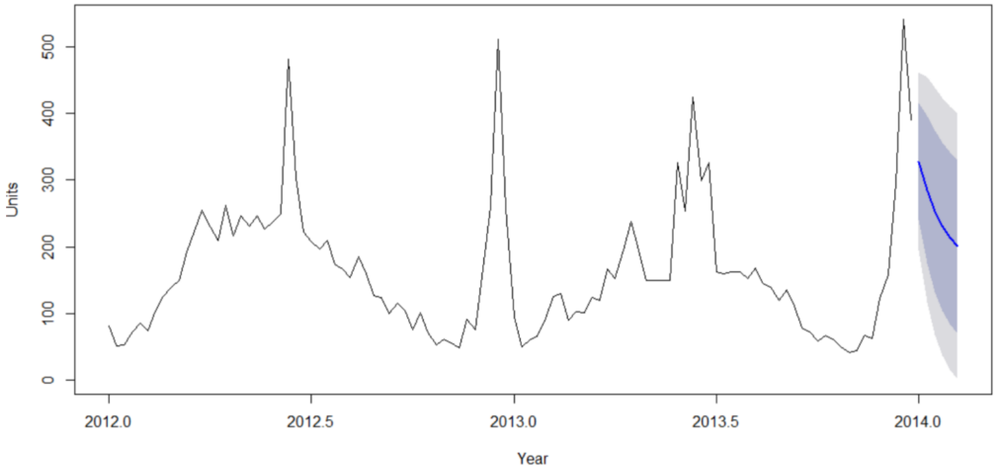


2. Similar analysis conducted directly in R:

Seasonal decomposition



Forecast from ARIMA model



R code:	Comments

units <- c(81,51,52,72,85…)
units.ts <- ts(units, start=c(2012, 1), end=c(2013, 52), frequency=52)
	# Create the time series

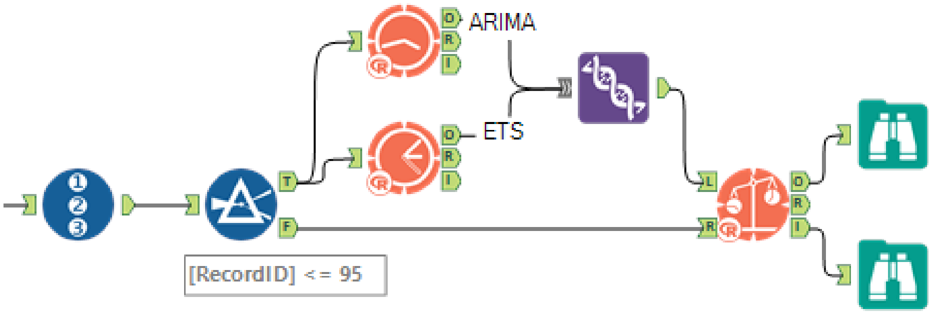
library(forecast)
seasonplot(units.ts, ylab = "Units", year.labels = TRUE, labelgap = 2, col = c(5,6), pch = 15, lwd = 3)
	# Seasonal decomposition

fit.arima <- auto.arima(units.ts)
forecast(fit.arima, 6)
plot(forecast(fit.arima, 6))
	# Train the model
	# Predict next 6 future values
	# Plot

Discussion

The ARIMA model seems to capture the drop in sales after the end of December peak.

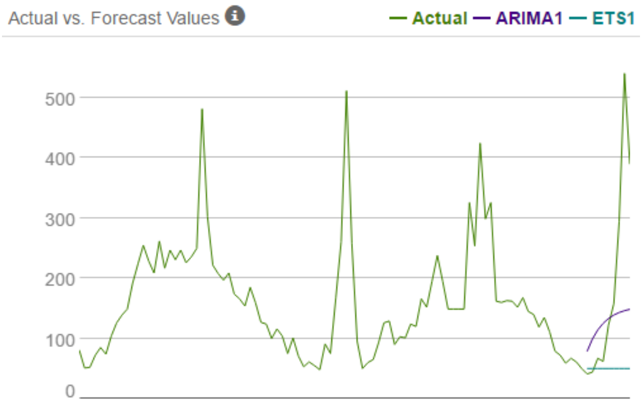
Comparison of models: In general for time series forecasting, we can first compare the performance of different models on known data. For example in our case, we could compare the performance of ARIMA and ETS for the last two months of 2013 for which we have actual data. The methodology after data preparation is as follows:



1. Assign a unique identifier to each record and filter out the last 2 months of 2013 (IDs 96 to 104).
2. Run the ARIMA and ETS models in parallel on the same input data, and union the outputs.
3. Compare the forecasts of each model with the actual data using TS Compare.

The TS Compare tool shows that the ARIMA model is better suited (lowest errors except for MAPE). The models do not capture well the December peak and the forecasts are much lower than the actual values.

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ARIMA	66	166	113	-22	66	2.9
ETS	141	220	144	44	52	3.7



The choice of the ARIMA model thus seems appropriate.