Question 1: The CEO would like to see the company's monthly profit, for each month, as compared to the same month of the previous year to judge the year over year improvements.

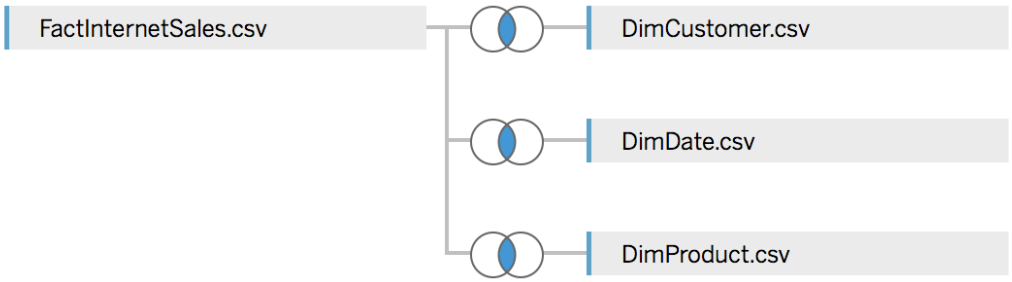

Figure 1: Dimensional tables needed for answering question 1, along with their alignment

Figure 1 is the showing the alignment of tables before initiating the analysis of the data. It is essential to align the table before analysis in order to validate that primary keys and foreign keys are in sync and the inner join between the two tables is possible. The reason for such validation is the fact that data analysis is done across tables so that meaning information can be derived.
In the data analysis challenge, the architecture of the data is in the form of snowflake, wherein the center table is the called the FACT table and all the linked tables are called DIMENSIONAL tables. Dimensional tables contain the primary key and attributes relevant to the dimension. Fact tables contain the foreign key of all the dimensional tables along with the transactional data concerning all dimensional tables.
As shown in figure 1, there are 2 dimensional tables viz. *DimCustomer, DimProduct, and DimDate* linked via inner join to *FactInternetSales* table.

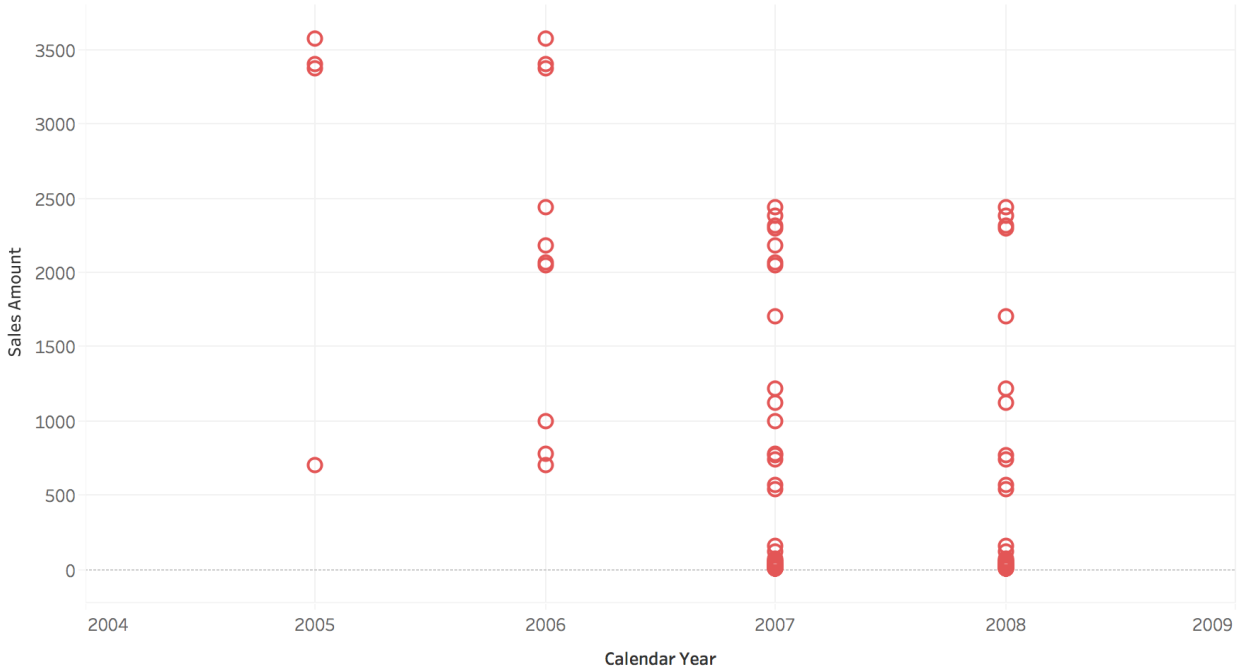

Figure 2: Change in Sales Amount over year

What we need to show ?

We need to show that there is an increment or decrement in monthly profit over the years. Before showing analysis of varying profit for the organization, we show the intuition for such analysis. Figure 2, shows the change in sales amount over calendar year. Sales amount is an essential visual factor to judge the change in profit in the organization. We observe that company did a lot of improvement in sales in the year 2006, 2007 and 2008 as compared to 2005. Though the sales amount didn't reach those levels which were attained in 2005 and 2006 but the closeness of the circles in the plot shows high traffic in sales. Also the sales didn't remain active throughout every month of the year.

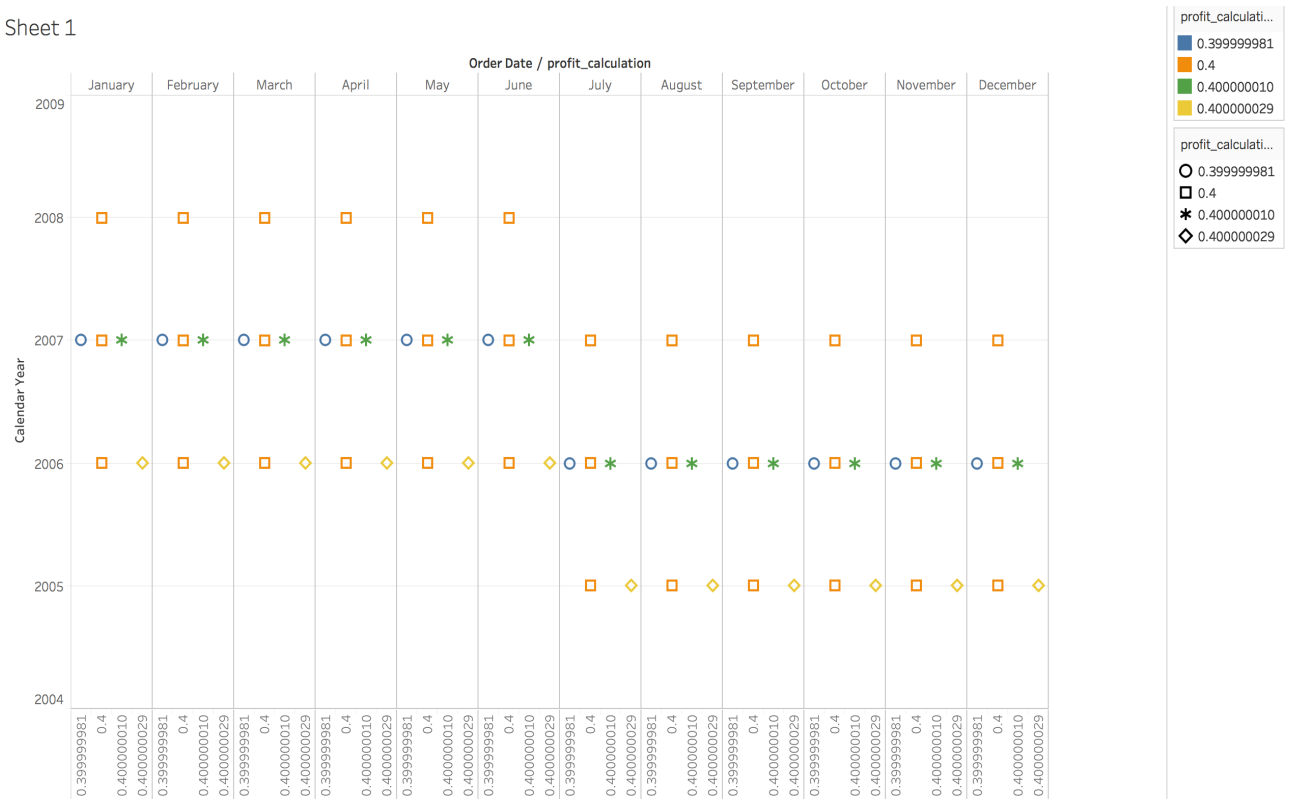

Figure 3: Profit and Loss movement over calendar year and over months

Figure 3, visualizes the movement in profits and loss over months and year. We formulated the equation for profit as:

$$Profit = \frac{ListPrice - DealersPrice}{ListPrice}$$

After carefully analyzing the data on the bases of profit, we saw that the company has a achieved a maximum profit of 40% over the years {2005,2006,2007, 2008, 2009}. When we stretched our profit margin to 41%, we observed that company hasn't achieved that much profit over 5 years. Figure 3, shows that, year 2008, didn't have any improvement in profit over the months. In the year 2007, the company showed very minute change in profit over the changing months. The highest profit percentage of 40.0000029% was achieved in the year 2006 and 2005.
From figure 3, we observed that company's profit was saturated to 40% from July 2007 to June 2008.

Question 2: The manager of the Canadian region would like to compare the sales of each product subcategory (except the bike category) in her region versus all other regions in the company. She would also like to see the trends of each of these subcategories month over month.

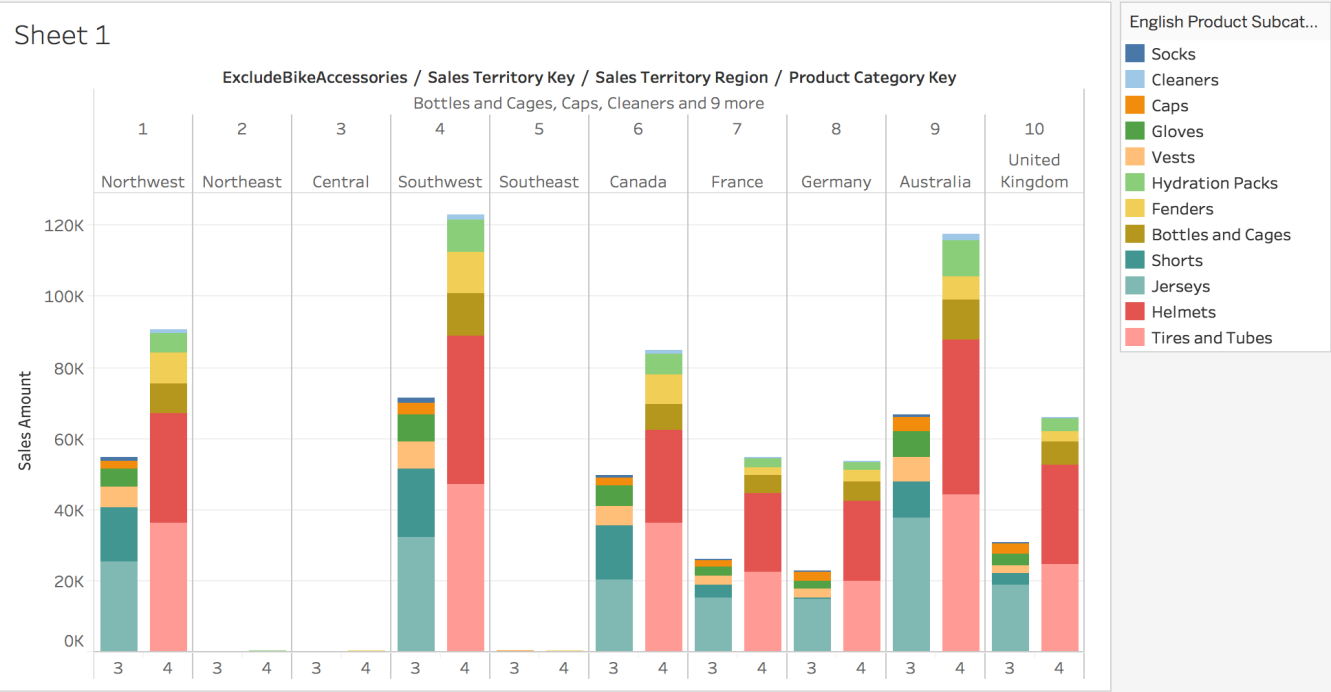

Figure 4: Sales of each product subcategory (excluding bike) across regions

Category 3 contains:
{Jerseys, Shorts, Vests, Gloves, Caps, Socks}

Category 4 contains:
{Tires and Tubes, Helmets, Bottles and Cages, Fenders, Hydration Packs, Cleaners}

The company has 10 regions worldwide. These regions are Northwest, Northeast, Central, Southwest, Southeast, Canada, France, Germany, Australia and United Kingdom. Figure 4, shows varying sales of category 3 and category 4 across 10 regions with focus on Canadian region. Considering the positive side of Canadian sales, we see that sales of items in category 3 and category 4 in Canada are higher than Northeast, Central, Southeast, France, Germany and United Kingdom. Whereas, the sales of these items are less as compared to Southwest, Australia and Northwest.

Suggestion: The difference in sales between Canada and {Southwest, Australia and Northwest} is approximately 40K. Proper gender and demographic based advertising of items in category 3 and 4 can help Canada achieve higher sales than its competitive regions. Width of each color bar implies which product need marketing and in which region.

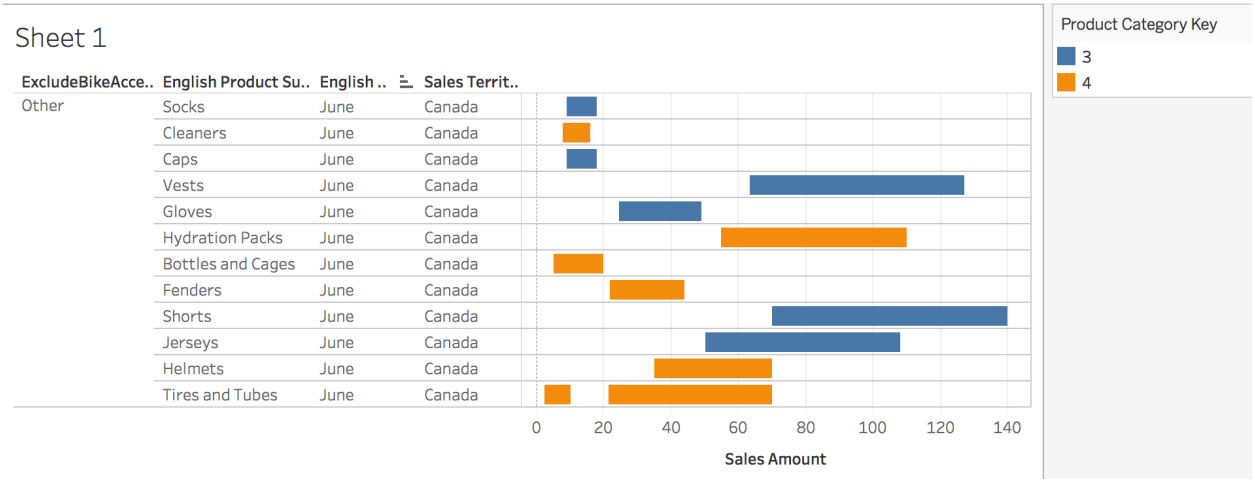

Figure 5: Subcategory sales in Canada over months

Coming to the second part of question concerned with sales of subcategories in Canada, we are amazed to see that sales of items in category 3 and 4 was done in month of June whereas there is no sign of sales in other months. From figure 5, we observe that majority of sales is attributed to Helmets, Shorts, Jerseys, Tires and Tubes, Fender, Hydration Packs and Gloves. The distribution of items doesn't show class imbalance which is a positive observation for Canadian sales.

Suggestion: Socks, Cleaners, Caps, Tires and Tubes needs advertising in Canada.

Question 3: The same Canadian manager would also like to run some promotions in her region specifically to sell some more bicycles*. Help her identify who/what/where might help make the most successful promotions.

* There is no category as bicycles in dataset. The analysis is done on category: Bike.

- **Where:** Canadian manager needs to run promotions in her region for bikes. Canada being a very large region, promotions should be targeted to specific cities in side Canada.
- **What:** Canadian manager needs to know what are attributes that can help in promoting the sales of bikes. We saw days of cost, manufacturing, order quantity, number of children and income are some of the attributes that can help in promoting the sales of the bikes.
- **Who:** Canadian manager needs to know who are the group of people that can be targeted for promotion. We observed that gender and commuted distance based grouping of people aid in promoting bikes sales.

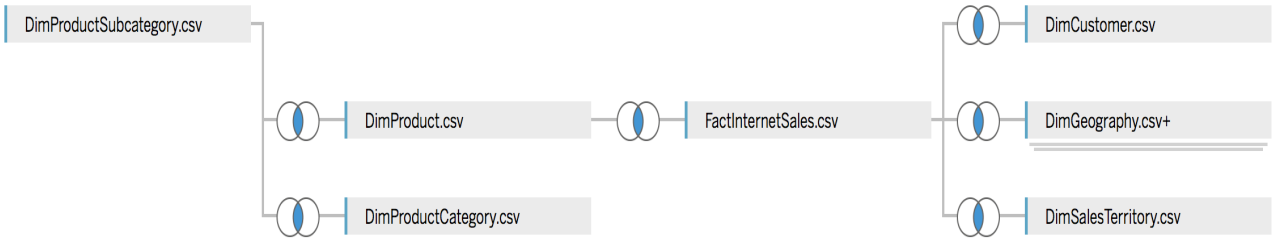

Figure 6: Dimensional tables needed for answering question 2, along with their alignment

Figure 6, shows which dimensional tables are needed for providing analysis to Canadian CEO so that he can find sectors for promoting bicycle sales. Since, we are analyzing category bikes and its subcategories, so *DimProductSubcategory* is the first table to look for. This table is linked to *DimProduct* and *DimProductCategory* so that we can analyze at category level and subcategory levels. We are linking to *FactInternetSales* table for finding sectors (attributes) where promotion can be done. This will answer "what" part of the question. We link this table to *DimGeography* and *DimSalesTerritory* for answering "where" part of the question. There is also link to *DimCustomer* for answering "who" part of the question.

Answering "where" part of the question

In this section we will identify those cities in Canada which have attributed to low sales in bikes.

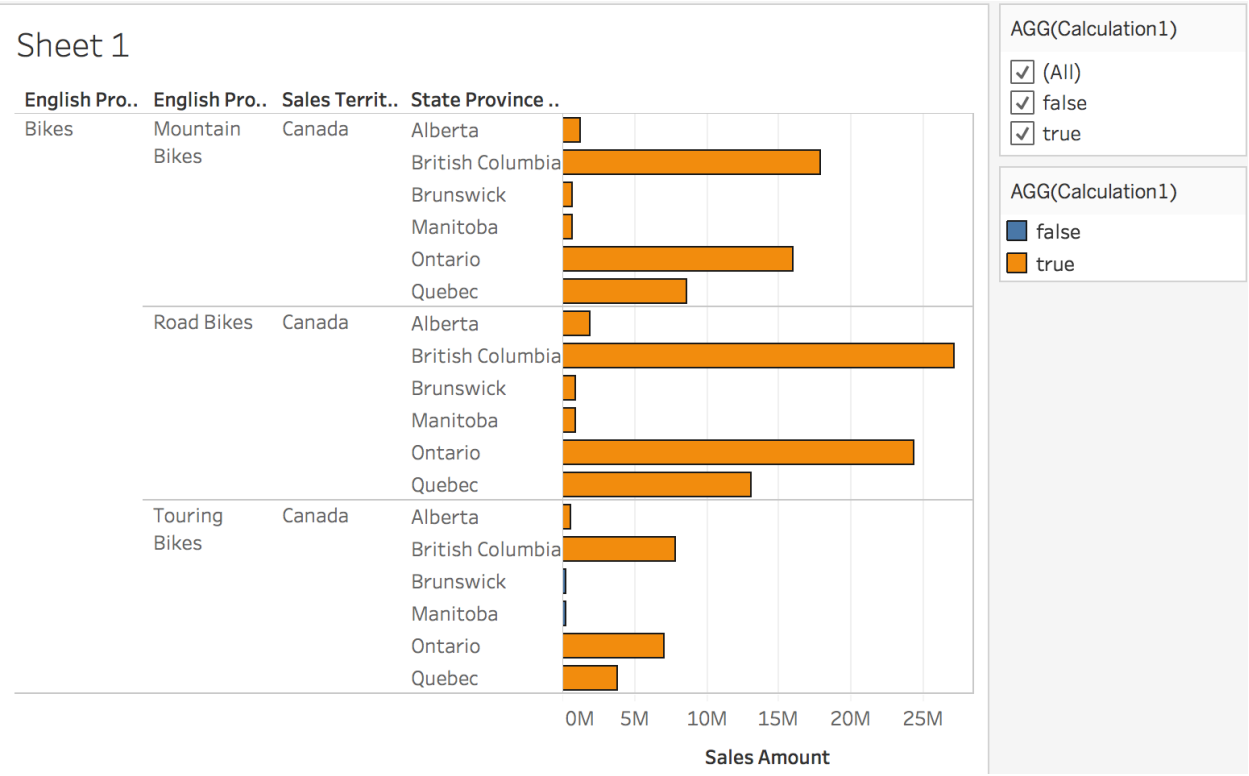

Figure 7: Sales of subcategory of bikes over different states in Canada

From figure 7, we identify that under subcategories Mountain bikes, Road bikes and Touring bikes, *Brunswick and Manitoba* are the two cities which should be targeted for improving the sales of bikes in Canada region. It is evident from the figure that cities viz. Brunswick and Manitoba showed poor sales under subcategory: Touring bikes, hence have been listed as **"false"** based on the minimum threshold on sales amount. Though other cities like Alberta should be also be considered for improving sales of bikes.

Answering "what" part of the question

First, we will consider the *income and Number of Children at Home* as the two probable sectors for giving up thrust to the sales of bikes.

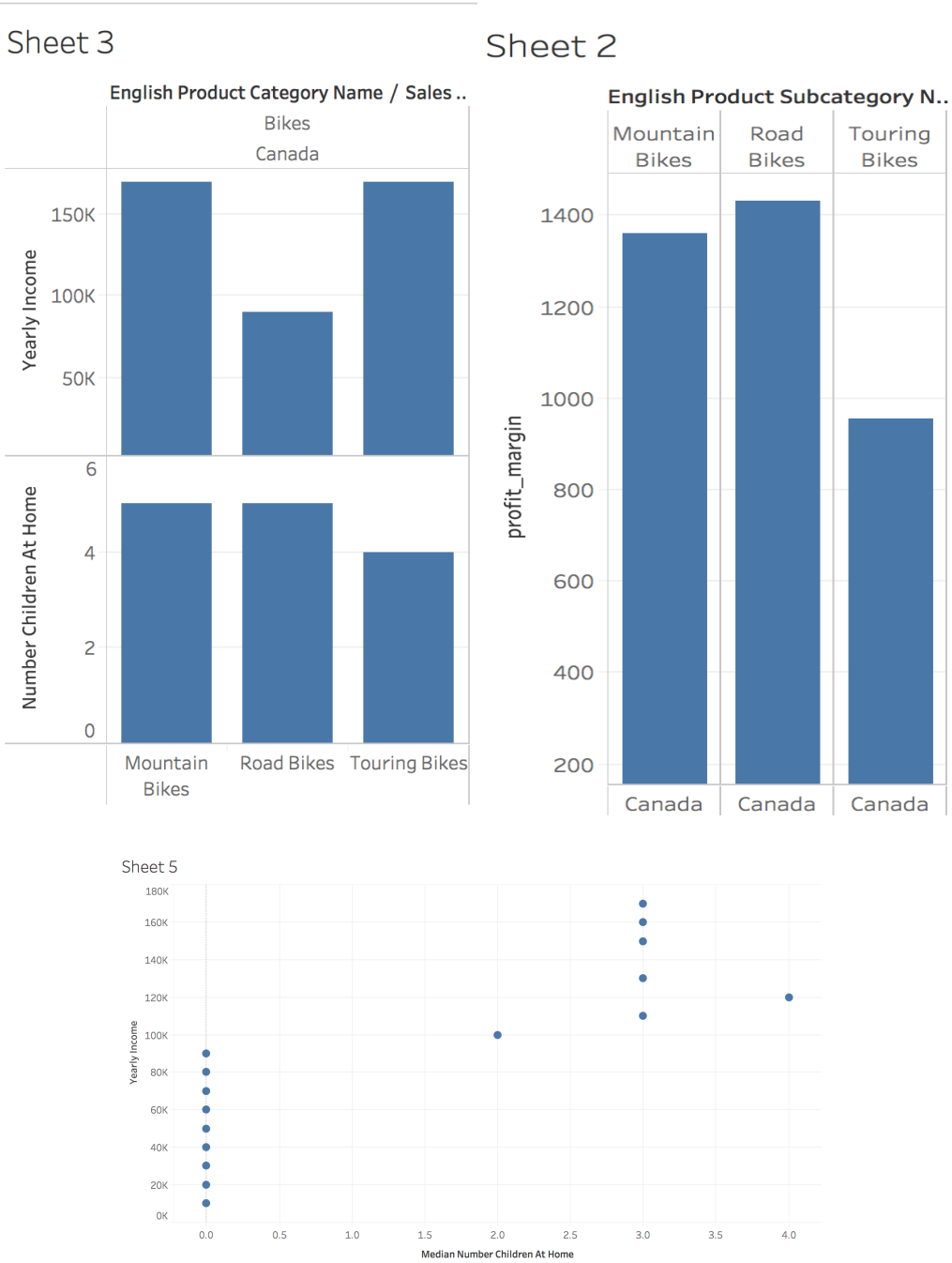

Figure 8: a. Bikes distribution over Number of Children at Home and Yearly Income; b. Profit margin; c. Yearly income v/s median number of children at Home

We can observe from figure 8a, that families with yearly income over 150K prefer mountain bikes and touring bikes over road bikes. When we look at figure 8b, we see that profit margin is high for road bikes as compared to mountain and touring bikes. **Hence, the Canadian bicycle dealer**

should promote selling of road bikes over mountain and touring bikes. This will increase the profit for the company. Considering figure 8a, we observe that families having number of children above 4 prefer road bikes and mountain bikes over touring bikes. The company can promote touring bikes among children by organizing tours. We were unable to find an attribute of safety with respect to bikes. "Safety", can be one reason why touring bikes are not popular among children. Figure 8c, shows that families earning below 100K have zero children, hence we can't target promotions in those families. We also observe that families earning 100K and above have 2, 3 or 4 children on median scale. Since, these families love bikes as evident from figure 8a, we can bring up promotional events for such families.

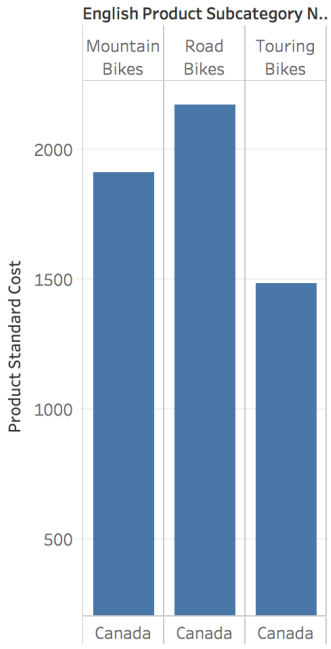

Figure 9: Standard cost of bikes in Canada region

From figure 8a, families with income over 150K are fond of touring bikes over road bikes and standard cost of touring bikes are less (figure 9). Hence increasing the sales of touring bike can also help in compensating the sales from road bikes. This task is more feasible as compared to road bikes.

Answering the "who" part of the question

In this section we are concentrating on gender and need based promotional events. Bikes are keeps human being fit and active throughout the day. Many people commute miles on bikes for touring, mountaineering and road trip. Here we are targeting both males and females. Focusing on the need of the people have seen tremendous rise in sales of a product and growth of the organization. Primarily, we are focusing on gender based promotional event.

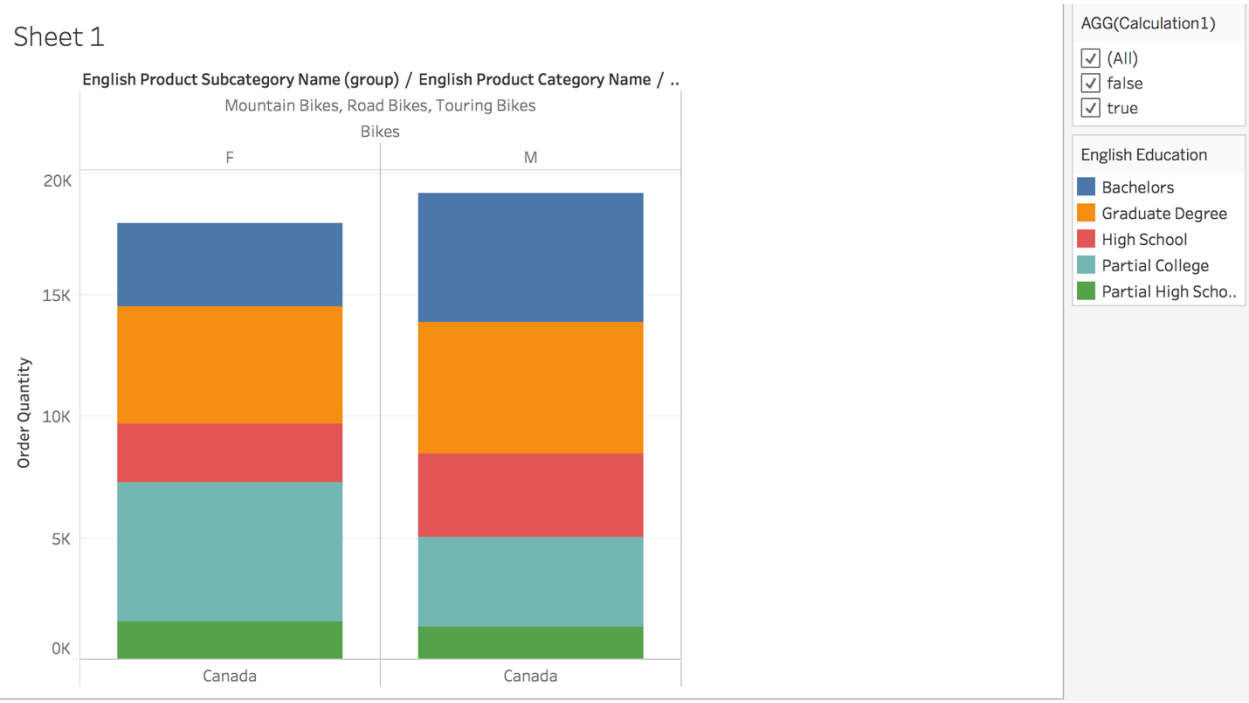

Figure 10: Order Quantity of bikes among educated males and females in Canada

So far we have talked about "where" and "what" part regarding promotion of bike sales. Those factors are essential but till we don't cover individual or group of people, our analysis remains weak. In figure 10, we show that order quantity for males if more than females. This shows that sales can increase if we focus towards male community. But the difference between male and female order quantity is less, this makes us move towards both communities with equal probability. We observed that :

- Within these communities we see that population in *Partial High School* have low order quantity as compared to other classes in *English Education*. Hence, the company can organize promotional events for this population.
- *High school girls* have low order quantity as compared to *High School boys*.
- *Partial College girls* have high order quantity as compared to *Partial College boys*.
- Bikes are popular among *Bachelor boys* as compared to *Bachelor girls*.

These identified points can help in setting up promotional events for different classes of educated individuals (male and female).

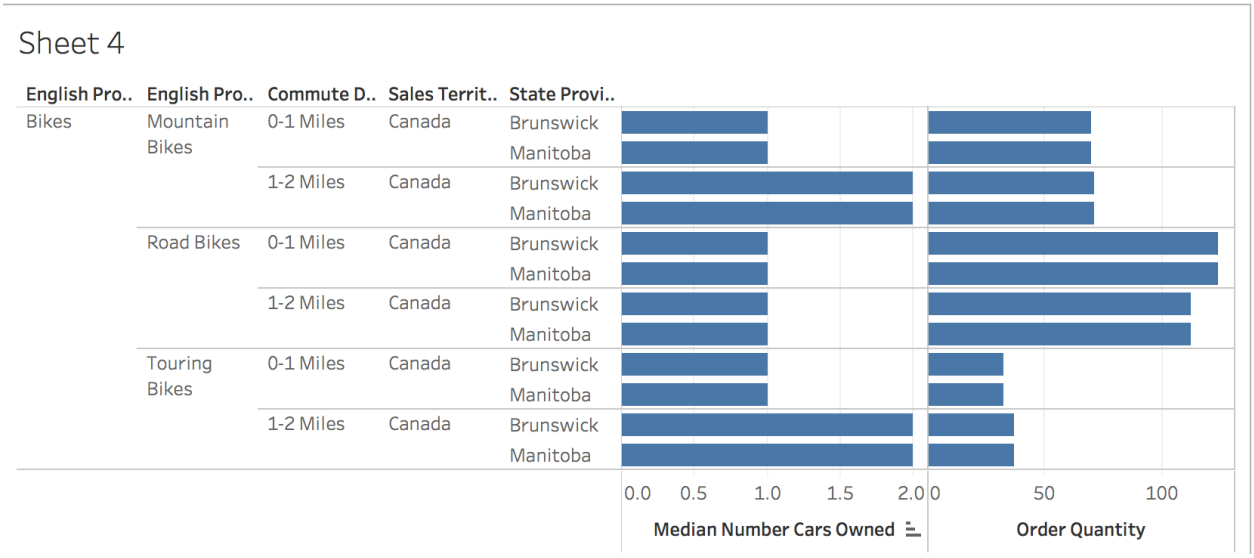

Figure 11: Order quantity and Number of cars along with commuted distance (0-1, 1-2 miles) based bikes distribution

From figure 10, we observe that males and females in different subcategories of *Education Class* differ in terms of quantity of bikes ordered. This is because of their requirement. Some people have to travel less than a mile which they can do by walking. As a result, they don't order bikes. Such a behavior motivated us to look for distribution among *Order Quantity, Number of Cars Owned* and *Commuted Distance*. Figure 11, reveals that, **why people in Brunswick and Manitoba have order less bikes as compared to other regions in Canada? See, under *Touring Bike → (1-2 or 0-1 Miles) → Canada → (Brunswick or Manitoba) ; very low quantity of bikes are ordered.***

From figure 11, we noted that people have 1 car on the median scale, tend to order large quantity of bikes. On the contrary, people having 2 cars on the median scale, tend to order less quantity of bikes. Hence, families travelling 1mile or more, need to considered for promoting the sales of the bikes.

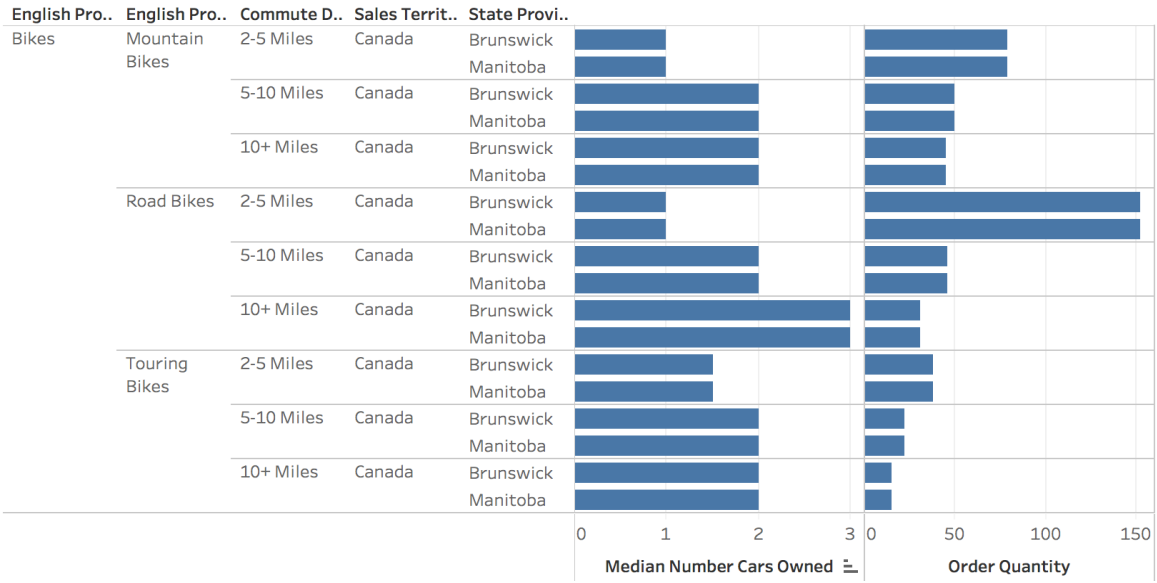

Figure 12: Order quantity and Number of cars along with commuted distance (2-5, 5-10, 10+ miles) based bikes distribution

What we concluded from figure 11, becomes even more evident after seeing figure 12. Figure 12, talks about families who commute more than 2 miles and how they behave towards ordering bikes? **From both the figures 11 & 12, it is convincing that people with 1 car on median scale, tend to order road bikes more than mountain and touring bikes.** It is good for company's business as profit margin (figure 8b) is large for road bikes.
In figure 12, following the path, *Touring Bikes → (2-5, 5-10 or 10+ Miles) →Canada→ (Brunswick or Manitoba) ;* very low order quantity ordered. Similarly, people commuting more than 5 Miles tend to order less road bikes and mountain bikes as they prefer cars over bikes.
Hence, people having 2 to 3 cars on the median scale can be targeted for promoting the sales of bikes.

Moving deeper into wheel, now we need to see how many people order bikes based on their occupation and marital status. We did outlier detection for validating our results.

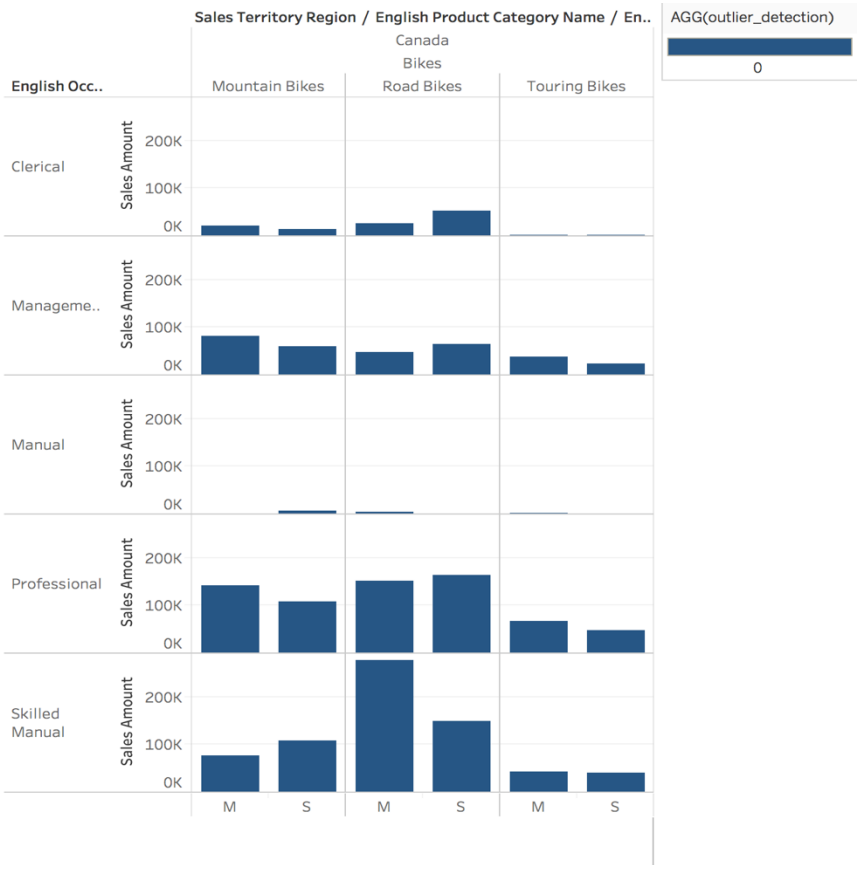

Figure 13: Analyzing bikes sales on occupation and marital status based distribution. M: Married, S: Single

It is apparent from figure 13, people involved in *Manual* occupation didn't contributed much to sales of the bikes irrespective of their marital status. **Around 6700** people are involved in this occupation. The company need to promote bikes among these people using some kind of promotional discount. Along with people in *Manual* occupation, we need to target people in *Clerical* and *Management* occupation because their contribution to sales very less as compared to people involved in *Skilled Manual* and *Professional* occupation.
It is noticeable that there is not much difference across marital status among people working in *Professional, Manual, Management* and *Clerical* occupation except *Skilled Manual*.

In *Skilled Manual* occupation, contribution to sales by married individual is more than single individual. Since, yearly income of *Skilled Manual* is above 50K, company can foresee profit, if they target these group of people.

Question 4: Free for all. The director team has given you free reign to dig into the dataset and make some interesting reports that will help to analyze trends and make prudent business decisions. Do your best to impress them and help them to easily understand where the strengths and weaknesses are.

In this section we will be discussing on some deeper analysis that can be done on the business dataset of the company. So far we have done analysis on

- **Profit.**
- **Canadian sales across different subcategories, seen their trends over months.**
- **Finally, within Canada we identified cities, company attributes and community which company need to target in order to increase the sales of bikes.**

Before beginning with trends and decision making, we will layout some measures to find those attributes which should be targeted and which should not be. The measures which we are incorporating are:

- T-test (two tails): Example: Are the sales across Region 1 and Region 2 different? The test will return "Different", if the Null Hypothesis is rejected at 95% confidence and "Same" otherwise. [1]

$$p_{value} = 2 * \left(0.5 * e^{\left(-1.2 * measure^{1.3}\right)}\right)$$

- Outlier Detection: Outlier is a value that is distant from other observations. Either variability in measurement or occurrence of error at the time of recording of data, are few sources of outliers.

$$Outlier = mean - 2 * std\_dev$$

For trend analysis we are performing **Moving Average** and for making prudent business decision, we are performing **Forecasting**.

T-Test

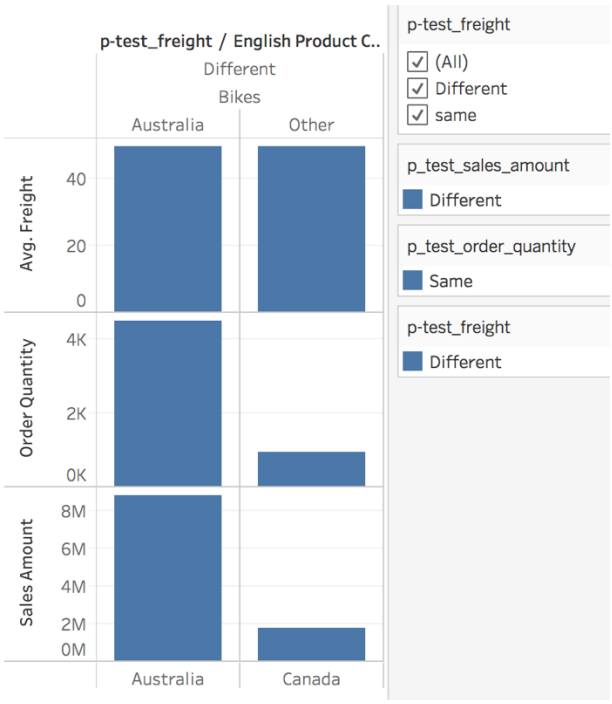

Figure 14: p-value based analyzing which attribute varies across regions

Amazing information provided by figure 14 is that distribution of *Order Quantity* across Canada and Australia is same based on p-value. The inference we get from this graph is that whatever rules of promotion we create based on trend in *Order Quantity* for Canada region, we can use the same rule for Australia. Whereas such is not case for *Sales Amount* and *Average Freight Cost*.

Outlier Detection

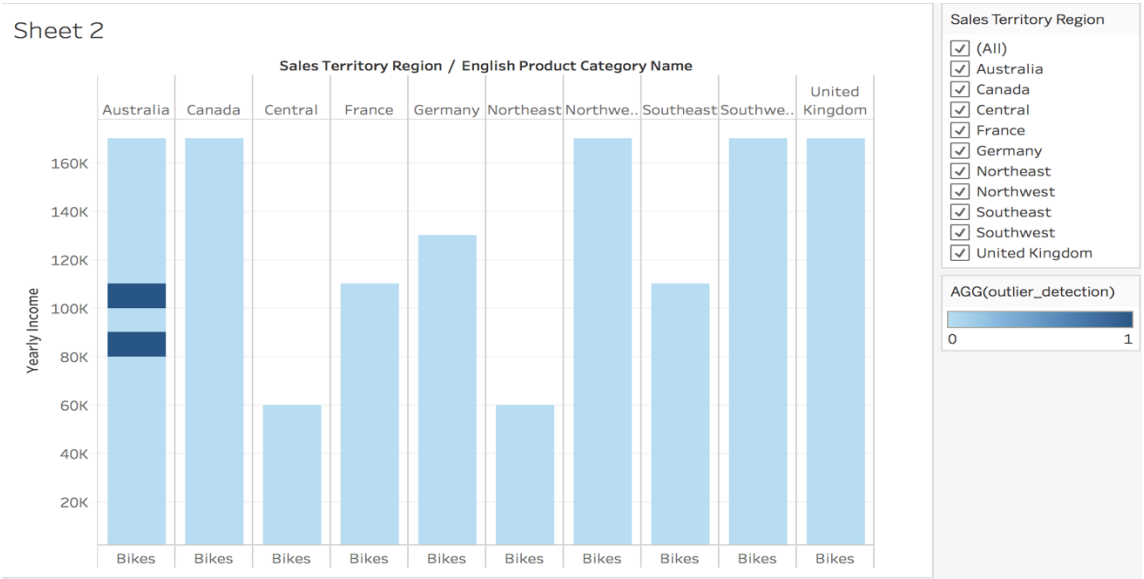

Figure 15: Outlier Detected in data on yearly outcome

There is a presence of outlier in the data which become noticeable when we did outlier detection on *Bike* sales in the region of *Australia*. Now the question is why we specifically did outlier detection in the region of *Australia?*

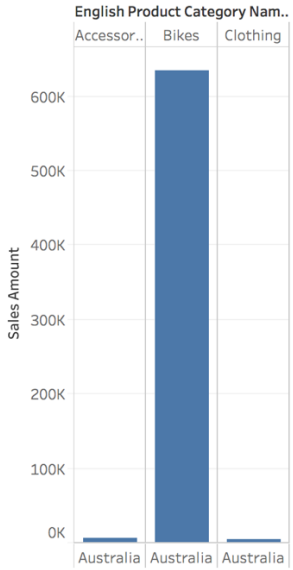

Figure 16: Class imbalance in *Sales Amount* in the region of Australia

We specifically chose *Australia* for outlier detection because we obtained class imbalance in *Sales Amount* distribution for bikes as compared to *Accessories* and *Clothing*. This is evident from figure 16.

Trend Analysis

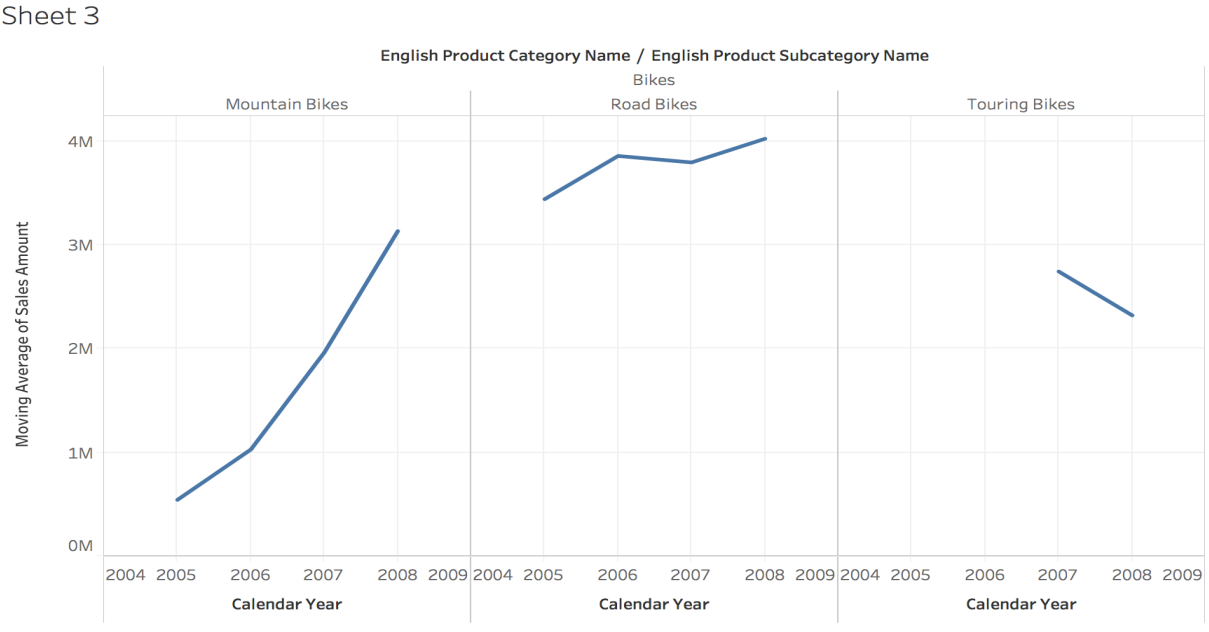

Figure 17: Moving Average on Sales over Calendar year

Figure 17, provide information on "How sales of *Mountain Bikes*, *Road Bikes* and *Touring Bikes* vary over years irrespective of the region?" We noticed following points:

- With changing years, increase in sales of *Mountain Bikes* become more steep. More steepness signifies high rise in *Sales Amount*. **This is the strength of the company.**
- With changing years, increase in sales of *Road Bikes* is abrupt. From figure 17, we notice that there is rise in sales from 2005 to 2006. There is decline in sales from 2006 to 2007 after which there is low steep rise in sales from 2007 to 2008. This movement can be seen in figure 3. The profit declined from 40% to 39%. **The company needs to make target specific rules in order to increase profit.**
- Considering the sales of *Touring Bikes*, we notice decline between 2007 and 2008. Also we noticed that company didn't have much sales for touring bikes. **This is a weak point for the company.**

Decision Making

For increasing profit, a company should be prudent in decision making. Decision making is always for the future. We can make effective decision if identify those attributes which will vary least in coming year. In other words, we should identify low variance (error) attributes.

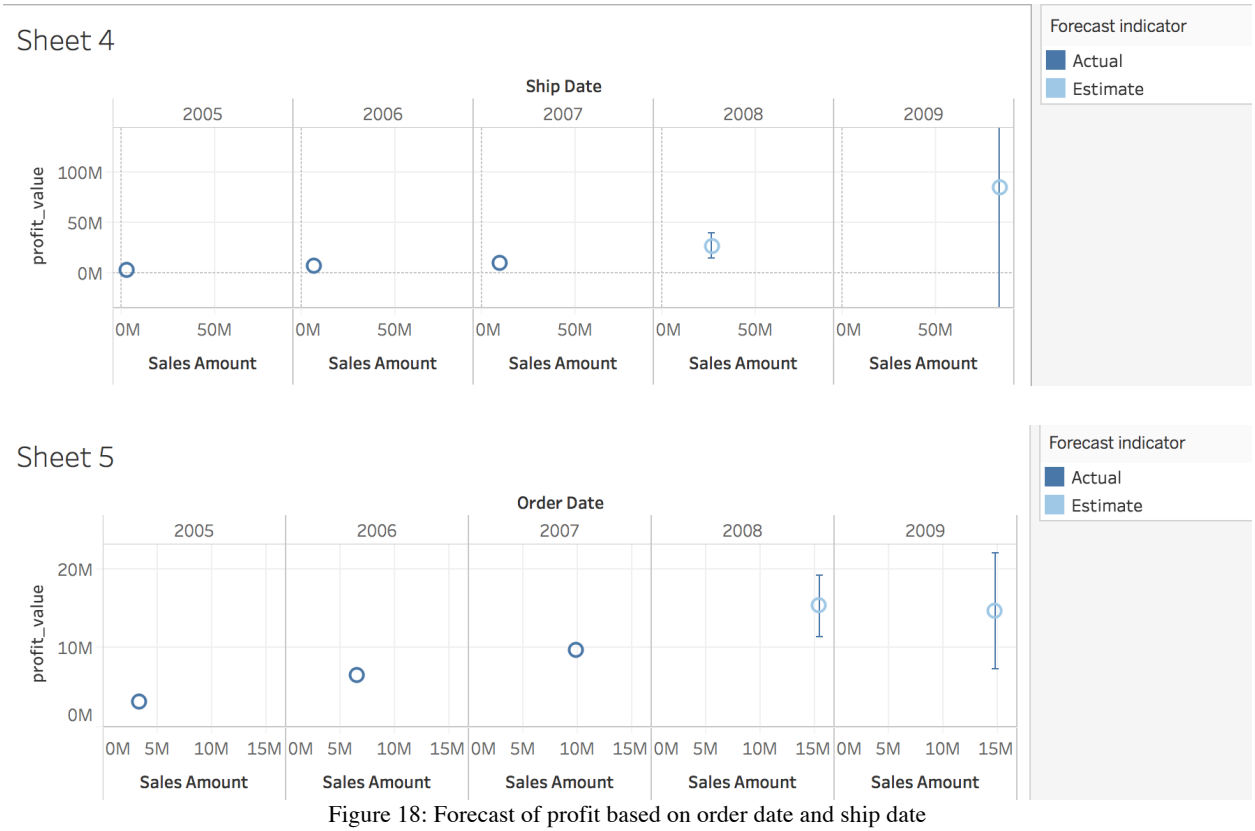

Figure 18: Forecast of profit based on order date and ship date

Based on the forecast shown in the figure 18, we observe following:

- Decision made on *Sales Amount* based on shipment date will tend to get wrong because the error bar for the 2009 forecast is very large.
- Decision made on *Sales Amount* based on order date will tend to get right because the error bar for the 2009 forecast is very low.

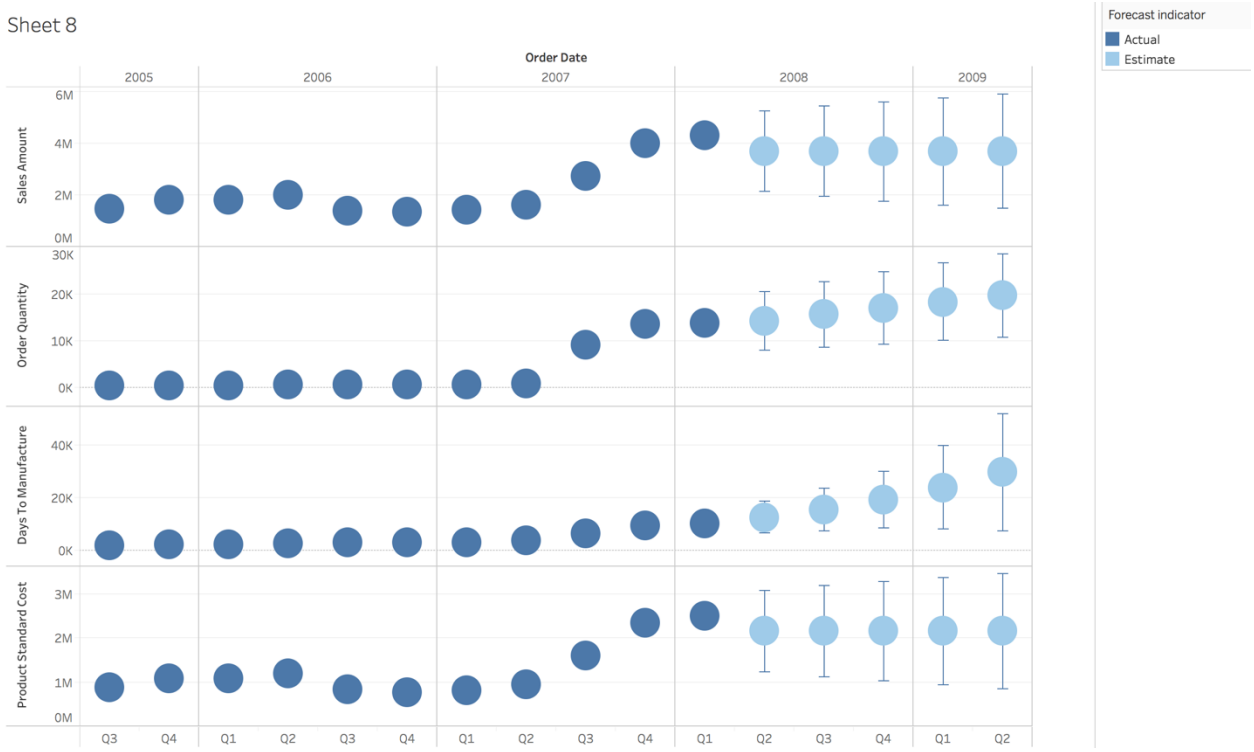

Figure 19: Forecasting of *Product Standard Cost, Days to Manufacture, Order Quantity, Sales Amount* quarterly for 2008 & 2009

In figure 18, we did identify which attribute should be taken for forecasting based on error bars (uncertainty variance). In figure 19, we did forecasting on some of the essential attributes which we identified as critical for incrementing the sales of bikes in Canada region.

Observations :

- **Considering *Sales Amount*, there will be an early quarter decline in sales for the year 2008 after which it will saturate.** There is a good amount of uncertainty in the forecast as seen by the error bars. Uncertainty value increases in 2009. We notice that the sales from 2nd Quarter of 2008 to 2nd Quarter of 2009 will lie between 1.8M to 6M.
- **Considering *Order Quantity*, there will be a rise in the value of this attribute throughout 2008 and 2009. This seems to be strengthening for the company's profit.** The variation in the error bar is low as compared to *Sales Amount*, hence company can make decisions based on *Order Quantity*.
- **Considering *Days to Manufacture*, there will be a rise in the value of this attribute throughout 2008 and 2009. This is a weak point for the company and will affect its sales.** The increase in error bar from 2008 to 2009 signifies increase in uncertainty. So, the company can't lay its decision based on this attribute.

- **Considering *Product Standard Cost,* there will be an early decline in the cost of the product for the year 2008 after which it will saturate**. We noticed that similar behavior in forecast is shown by *Product Standard Cost* and *Sales Amount*. Hence, trends in cost of the product will be guided by *Sales Amount*. This is a positive point for the company. Company cannot make decision based on this attribute.