

Mini-Project in Business Intelligence

Spring 2009, 22-04-2009

By Morten Gryning

Introduction

The precise definition of Business intelligence today varies a lot, depending on what methodology or textbook one use. The core elements are somewhat the same: extract, query, analysis and presentation, but the different approaches, and especially what areas weights highest, are different. This paper will use the methodology described in [Kimball]. This approach concludes that to have a successful business intelligence program it depends on three fundamental concepts:

- Focus on the business.
- Dimensionally structure the data that's delivered to the business via ad hoc queries or reports.
- Iteratively develop the overall data warehouse environment in management able lifecycle increments rather than attempting a galactic Big Bang.

The last concept is not that important while writing this paper, but the other two are really essential. For part 1 the data has to be dimensionally structured, to make its use intuitive to the business users. For part2 the data analysis has to focus on the business needs rather than what's actually possible with the data. Of course there still has to be an alignment; it's impossible to present something for which there is no data.

About this paper

As the project assignment paper is not very precise, I have made some decisions regarding what to include in this project. Considering that the maximum length of this paper is set to 10 pages, I have decided not to go into topics covered in the first part of [Kimball], and thus included in the Kimball methodology. There topics are the Managing the project/program, collecting requirements, technical architecture and creating the architectural plan. The reason for this is that

1. They are not relevant for part1, since all information I have about the company, is a picture of a transaction slip.
2. They are not relevant for part2, since that part use arbitrary data selected by the student¹. Of course it's essential to understand the topics, but writing about them here, will add no value.
3. Using space on there topics will reduce the depth of the other topics.

Part 1

This part will complete the 4-step process for designing dimensional models based on the slip from J-Crew displayed on the project assignment paper. Before beginning the 4-step process, dimensional modeling will be briefly explained.

Dimensional modelling

Dimensional modellings is a logical design technique for structuring data so that it's intuitive to business users and delivers fast performance². The fundamental requirement is simplicity, because it allows users, and especially business users without much technical background, to better understand the database structure. Dimensional modelling can be compared to normalize modelling, also called Third normal form (3NF) modelling. 3NF modelling is a design technique to eliminate data redundancies, which is done by divide the data into many discrete entities, each of which becomes a

¹ If I selected data from a real-life existing company, the situation would be different.

² [Kimball p. 234]

table in a relational database³. It's worth noticing that both dimensional modelling and 3NF is a technique; the amount of information present is the same using both techniques.

Dimensional modelling divides the world into measurements and context. Measurements are captured by the business processes and their supporting operational source systems⁴ and are referred to as facts. The context is divided into dimensions which describe the measurements in the fact table. The facts table and dimensions are related using the primary key for the dimensions, which are surrogate keys. As a rule of thumb, one business process results in one facts table with numeric measurements surrounded by a halo of dimension tables containing the textural context. This star-like structure is also called a star-join. Using the dimensional modelling technique, the fact table can become very large, as it contains all the measurements of the business process. Furthermore it's good to remember that every foreign key in the fact table must match to a primary key in the corresponding dimensional table, thus the facts tables keys should never be null.

Four-step dimensional design process

This section will describe and execute the four step dimensional design process. The approach chosen is to first do the four step process, and then redesign the dimensional model to support the additional design issues and questions.

Step 1 – Choose the business process

As mentioned in the section above, one business process links to one fact table⁵ which contains measurements for that business process according to the selected grain. So, the first step is to identify a business process or measurement event to be modelled. A business process is defined as an activity in the company/organisation that adds values to the company. Business processes are a fundamental building block of the dimensional data warehouse⁶.

Since the background for part1 is a slip from a sale, the business process chosen is the sales transaction. When a sales transaction occurs, all information about the transaction defined by the grain has to be stored in our database.

Step 2 – Declare the grain

After the business process has been chosen, the level of detail in the fact table for the selected business process has to be declared. Declaring the grain ensures an understanding of what exactly a measurement in the fact table represents. [Kimball] suggest that the grain is selected so the fact table is designed at the lowest level of detail available. The reason for this is that atomic data is the most expressive and most flexible. It's possible through various methods to sum up the atomic data to a higher level, but it's impossible to go the other way.

In this paper, the grain in the fact table is an individual transaction for a customer in a J-Crew store at a specific date and time. This means, that a group of rows in the fact table (*one for each product sold identified by the transaction*) contains information about a sales transaction, and the attributes in the fact table should support this information. Specifying exactly what attributes are valid for the sales transaction will be done in the next step.

³ [Kimball p. 236]

⁴ [Kimball p. 235]

⁵ It can link to more than one, but this is a rule of thumb.

⁶ [Kimball p. 298]

Step 3 - Identify the dimensions

Having stated the level of detail for the fact table, and identified a business process, the next step is to decide the dimensions applicable to the fact table at the selected grain. To help me in this process the table below lists all the information available from the J-crew slip on the assignment paper together with groups that I've assigned them to. Every dimension should match the grain.

Store_Name	Store_Address	Store_Telephone	Store_Area	Store_Number	...
Employee_Name	Employee_Associate				
PRO_Name	PRO_Number	PRO_Price	PRO_Id		
PRO_Promotion					
Date	Date_Time				

These are the five dimensions I've identified that match the grain: Store, Employee, Product, Promotion, and Date.

The information listed is not adequate for the dimensions, but represents the amount of information which can be extracted directly from the slip. Beside these five dimensions there's also some information from the slip which is stored directly in the fact table without a direct related dimensions. This information is identified in step 4.

Step 4 – Identify the facts

The last step in the dimensional modelling process is to identify the facts from the business process. Declaring the fact table grain establishes the foundation for determining the appropriate dimensions⁷. The list below contains the identified facts: *Subtotal, Total, Approval transaction code, Transaction identify, Unique ticket number, Payment type, Total products sold*.

The attribute payment type is not in a dimension. This is because it's currently not used for anything else than identify how the customer pays for the products. If there was a good way to identify the customer, which there isn't in this case, then the payment type could be added to a customer dimensions.

The quantity of a product sold is not included anywhere either. The reason for this is that the grain indirectly specifies that a row in the fact table contains information about a specific item sold in a J-Crew store at a specific time: Quantity would therefore always be one. Total_Product_Sold is a valid attribute, because it contains information about that total number of items sold when the transaction occurs.

Additional design issues and questions

This section will give answers to the design issues and questions. See Four-step dimensional design process, which mentions that the approach taken is to first design the dimensional model based on the slip alone, and then classify there issues and questions as additional business user demands.

How would you handle sales on a given day, before Christmas, during evening rush periods, from first shift and from 9 PM on a given date until 7 AM on the following

For business users to be able to query the dimensional model for both date and time, I've decided to add an extra dimension called time. This dimension will contain information about the time the

⁷ [Kimball p. 301]

transaction occurred, together with information about shift number and if the time is identified as a specific rush period⁸. My decision regarding an extra dimension is based on that the additional context adds value when one wants time related information. This dimension can then be used to see sales during rush periods and between shifts. Furthermore the addition of a time dimensions is aligned with the Kimball approach which mention that the time of day should be treated as a dimensions only if there are meaningful textual descriptions for periods within the day, such as the lunch hour rush or third shift.⁹ To support the display of sales before Christmas, the date dimensional will include information about fiscal year, month and day, and a holiday indicator that, e.g., indicates if a specific day is prior or after Christmas. This is also aligned with the approach in [Kimball] which state that for date dimensions calendar navigation should be driven through the date dimensions table, not through hard coded application logic¹⁰. In short, the date dimensions is extended with information about holidays and other indicators, and a time dimensions is added with information about work shifts and an indicator for rush periods.

What if there are over 50 characteristics of products or stores?

Since that I have already identified both a product dimension and a store dimension, it is not a problem if there are over 50 characteristics for a specific product or store. This is due to two things

1. There exist a one-to-many relationship between the fact table and the dimension table for both products and store.
2. Adding more characteristics for a product or store, would result in simply adding an additional column in the dimensional table, which would just serve as a contextual description. The fact table would still only reference the dimensions table through a surrogate key, so as far as the size of the fact table, it would not matter.

Is there a customer dimension?

As mentioned earlier, it's not possible to identify the customer. Therefore it would be impossible to have a customer dimension, because there would be no distinct relations between the fact table and the customer dimension.

How do you handle the unique ticket number (000389005500...)?

The unique ticket number has the following elements in it: transaction, store, date and employee. Because of the unique characteristic of the ticket number, I consider it as a degenerate dimension and store it in the fact table. The ticket number can then be used as the glue that holds the line item rows together¹¹. It would be pointless to make a dimension for the unique ticket number, because there would be no associated context in that dimension. Also, the unique ticket number will serve as the primary key for the fact table. Besides using the ticket number as a primary key to glue the rows together, the number can also be used to link back to operational systems for compliance, auditing, or researching data quality concerns¹².

Is there an employee dimension for the cashier?

⁸ This requires that the business users can agree on, when the evening rush is ;)

⁹ [Kimball p. 255]

¹⁰ [Kimball p. 254]

¹¹ [Kimball p. 257]

¹² [Kimball p. 257]

Because there exist context about the employee, and the employee can be directly identified at the transaction time, I've chosen to make an employee dimension. Currently this dimensions contains attributes for the employee name and his or hers associate number. The dimension can, like all other, be extended with e.g. contact information about the employee. To support this, I've added the attributes employee telephone and address, even though the information is not present on the slip.

How do you handle the VAT (MOMS)/sales tax amount?

When handling the tax I've considered three approaches. The first is to store the amount of tax in the fact table together with subtotal and total amount. However, I find that this would not be optimal because one can consider the tax amount to be used for filtering and labeling, and therefore it would be better to have the tax amount in a dimension, and then relate to that dimension in the fact table. The next approach is to make a separate dimension for the tax amount. This is still better than the first approach but there really isn't much context to add besides the tax amount number. Therefore I've chosen to add the tax amount to the store dimension. The reason is that it makes sense that every store has a tax attribute, and this approach does not include adding an extra dimension.

What about payment type?

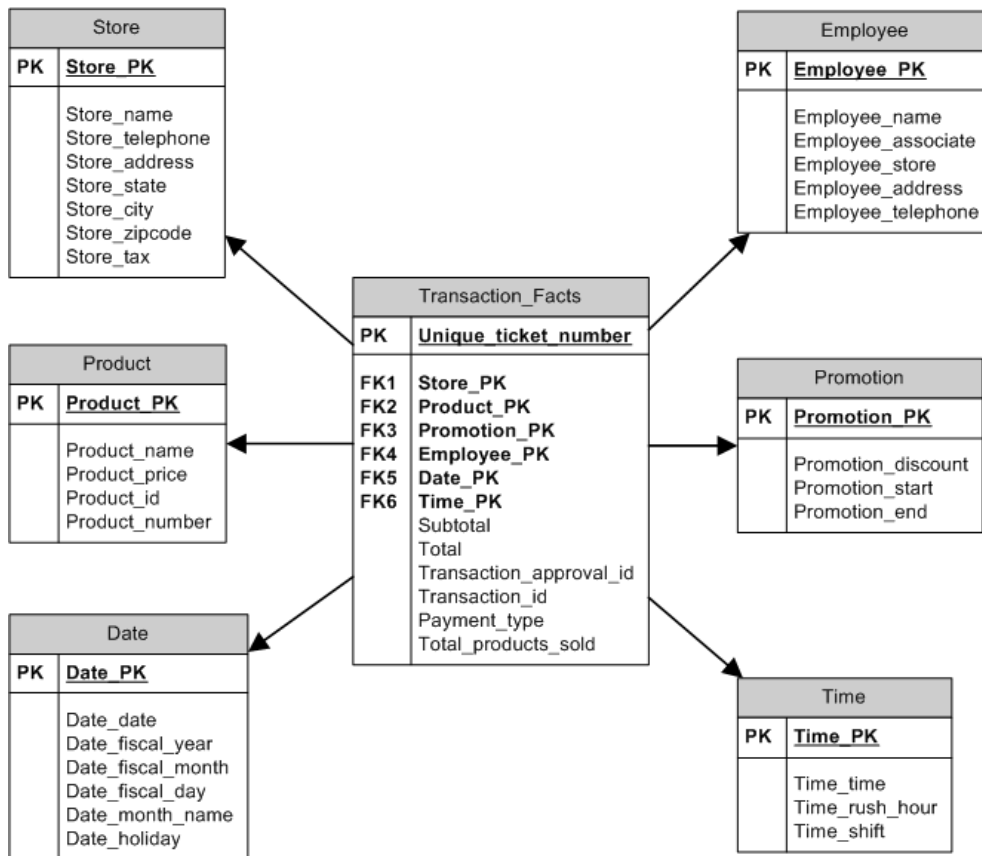
Even though payment type follows the characteristics to be inserted in a dimensional model: Discretely valued and used for filtering or labeling, I've chosen to store it directly in the fact table. This is because it doesn't belong to any of the dimensions identified so far, and because it currently have no associated context. In the future of the program, when a dimension like payment would be meaningful to have, payment type could be an attribute in that dimension.

Is there a promotion dimension?

Because every store can have different product discounts I've chosen to have a promotion dimension. The primary key for this dimension will be the same as that of the product dimension and the dimension will have attributes that describe the discount for a selected product, the date when the discount was first offered, and the date hen the discount will end.

Final model

After completing the four step model, and considering that additional design question, the model below shows the resulting dimensions model. The model contains 6 identified dimensions and the fact table. The primary key usedin the fact table is the unique_ticket_number, which functions as the glue that binds the different products brought together for each transaction.



Part 2

While the first part had focus on dimensional modeling and design, this part will focus on analyzing data using business intelligence tools and displaying the results to the end-users.

The data set and associated company

Healthy Foods (HF) is a medium sized company which aims to sell healthy products. The strategy is to differentiate from the competitors by using IT¹³ and give the customer a good service. The company has retail stores around the USA in 3 different proximity states. To optimize their sales of fruit, they have experienced with different prices for oranges in their different stores. Currently the company sells 2 different kinds of orange (O1, O2) which vary in size and quality. The experiment is a beach-head, and if the management gets some value from it, they will price the oranges and other fruit, according to the business intelligence reports. As part of their technology based strategy, the company collects the following data during each day: A number identifier for the current day (1-6), the quantity sold of the different oranges O1 and O2, the daily price for the oranges and a store identifier. So far the pricing experiment has run for 1½ month¹⁴.

To help the management judge if they gained some value from the experiment, they stated some question they wanted to investigate. The assignment for the business intelligence team is to produce

¹³ And in extend Business Intelligence

¹⁴ The dataset used is oranges in the Sas sample data sets

one or more reports which give answers to the stated questions. The questions are the following:

- Assuming that each day is as good as another to sell oranges.
 - Is there an average best price for O1 and O2 as to maximize the profit gained by selling these products? If we are unsure, for example when opening a shop in a new state, how should we price the products?
 - If we have many oranges in stock which rotten soon, or a very little quantity left, is it, with some precision, possible to predict what price we need to sell them for, to get them sold quickly/not run out?
- Assuming that every day is not equal, what is the best price for O1 and O2 for each of the days, to maximize the profit when we consider all shops the same?

Question 1

Since the numbers of different prices are relatively low, I've chosen to show the correspondence between the selected price, and the average amount of money gained for each price by using bar charts.

The reason for choosing the amount of money gained is based on the fact that you could sell all your oranges if you gave them away free; the quantity sold is not interesting. However the amount of profit gained from selling them is what the price is evaluated at. The first step in the analysis is therefore to compute how much money the company gain each day by multiplying the quantity and price. The predicted result is that a price set too low would result in either people thinking the quality is below-average, and thus buying less, or that many are sold, but the gain is low because they each cost very little. A price too high would result in that people thinking that the oranges are too expensive. Also, since the dataset is relatively small, data from all stores are used.

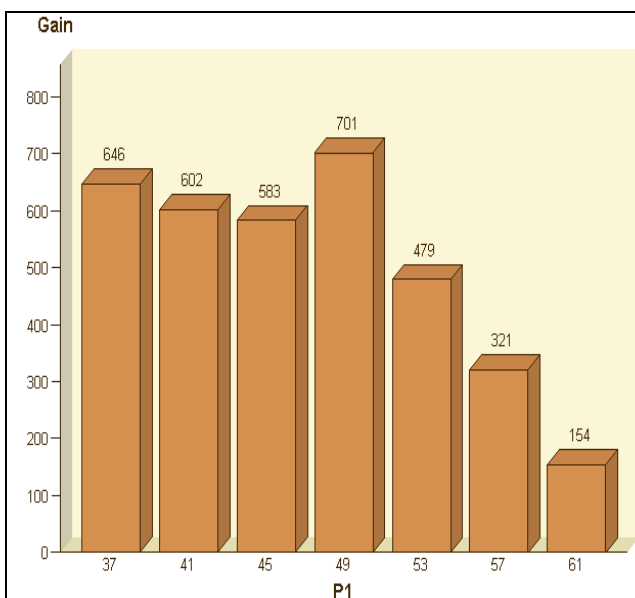


Illustration 1: The relation between the price for orange type 1 and the average daily gain from selling these products.

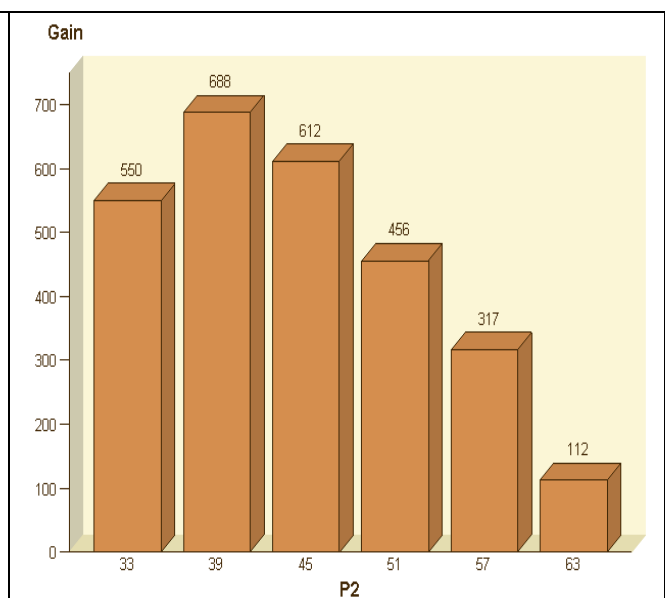


Illustration 2: The relation between the price for orange type 2 and the average daily gain from selling these products.

As predicted for the dataset, illustration 1 and 2 indicates that the best price is in the middle for O1. For O2, which is priced cheaper, the best price is almost the cheapest. To answer the management question, the best price to maximize profit, as shown by illustration 1 and 2, is 49 for orange type 1, and 39 for orange type2. To show the relationship, I've concluded that the bar charts is easy to understand without much technical background, and therefore sufficient as a report to the management showing the relationship between the price and gain.

Question 2

If factors about the quantity matters more in relations to the profit, that is, if the company makes more money on getting rid of the oranges quickly or price them high, how should they price them considering the amount left?

To give a solid answer to this question, I've chosen to see if there exist a linear or non-linear regression between the quantity and amount sold. If such a regression exists, the company can use it to predict how to price the oranges, based on how many there's left in stock. The below illustration shows a linear regression for the price for O2 related to the average quantity sold for that price.

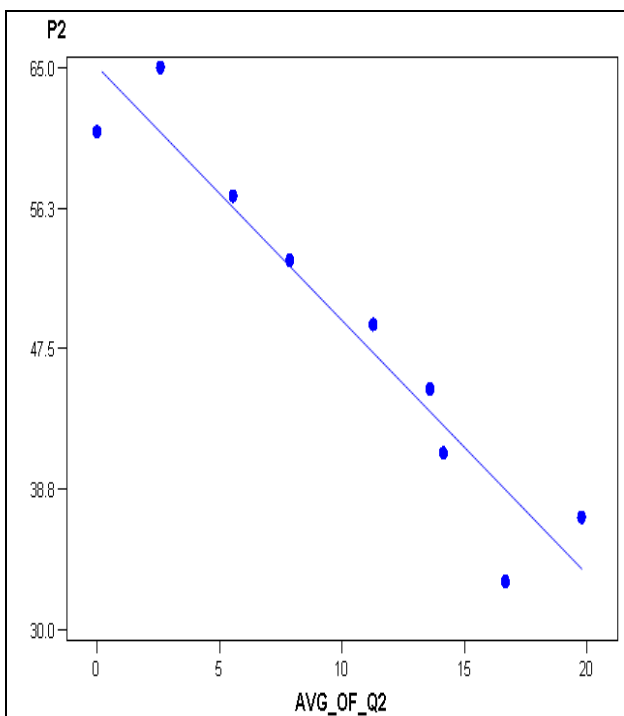


Illustration 3: The relationship between the price (P2) and average quantity sold of O2 (Avg_Of_Q2)

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	875.67861	875.67861	72.70	<.0001
Error	7	84.32139	12.04591		
Corrected Total	8	960.00000			

Root MSE	3.47072	R-Square	0.9122
Dependent Mean	49.00000	Adj R-Sq	0.8996
Coeff Var	7.08311		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	65.06397	2.21093	29.43	<.0001
AVG_OF_Q2	1	-1.58061	0.18538	-8.53	<.0001

Illustration 4: The parameters for the linear regression and information about how well the model fits the dataset.

As the numbers and graph shows, there exist a good relationship between the quantity sold and the price using linear regression. Since the error rate is sufficient low (*R-Square is close to 1.0*¹⁵), I've chosen not to try with non-linear regression. Illustration 3 shows that a lower price results in more oranges sold, which is rather obvious. However the illustration and associated numbers also tell

¹⁵ An R-Square value of 1.0 means that the regression line perfectly fits the data, while a value of 0 means that it doesn't fit the data at all.

exactly *how* to price the oranges. Using the parameter estimates, the company can price the oranges compared to the stock amount, and thus maximize the chance they will sell the oranges before they rotten/still have some left, as to maximize the profit.

Of course, the company wants to know if there's a relationship between the higher priced orange type O1, and the quantity sold as well. The approach for investigating this is the same as for O2.

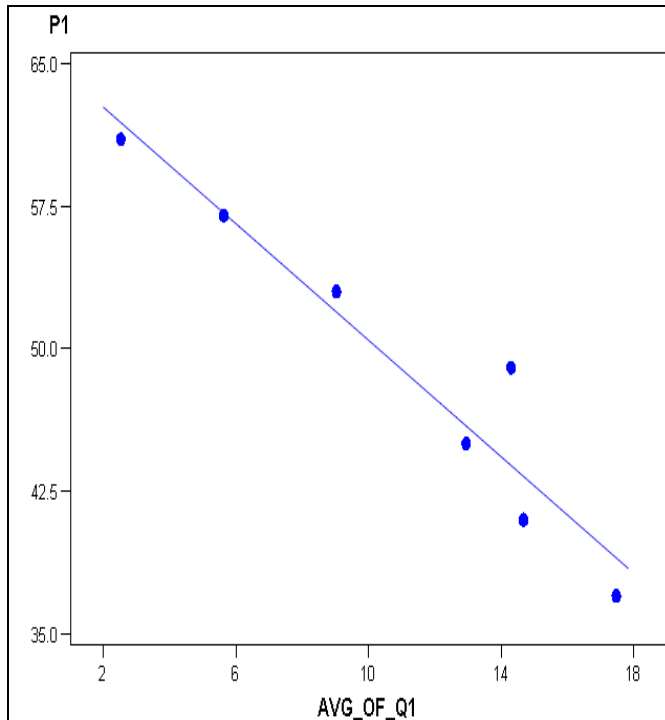


Illustration 5: The relationship between the price (P1) and average quantity sold for O1 (Avg_Of_Q1)

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	409.46155	409.46155	53.12	0.0008
Error	5	38.53845	7.70769		
Corrected Total	6	448.00000			

Root MSE	2.77627	R-Square	0.9140
Dependent Mean	49.00000	Adj R-Sq	0.8968
Coeff Var	5.66586		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	65.76240	2.52789	26.01	<.0001
AVG_OF_Q1	1	-1.53192	0.21018	-7.29	0.0008

Illustration 6: The parameters for the linear regression and information about how well the model fits the dataset.

As illustration 5 and 6 shows, there exist a similar relationship between the quantities sold and price; very similar to the previous relationship for O2. Again, the R-Square error is close to 1.0, which means that the model fits the dataset well. This can also be verified graphically.

To answer the question stated by the business management, illustration 3 and 5 shows an understandable graphical view of the relationship which can easily be understood. I would not show the numbers to the business users, but they can be attached if the business wants further clarification of the model.

Question 3

The last question assumes that each day is different when selling oranges. Maybe some people tend to buy food in the start of the week, or are willing to pay more in the weekends.

Because the days also are to be taken into account, this creates some challenges for reporting the findings. This is due to the fact that there are 3 dimensions: price, profit gain and day number. Using a bar chart like in question 1 is sufficient for reporting the relation between price and profit, but the third dimension results in, that the bar charts are to be either divided into 6 (one for each day) or shown as a 3rd dimensional graph.

I've chosen to show the relationship (*if any*) as a 3rd dimensional graph. The reason is that the limited number of different prices, and different days, should keep the graph understandable.

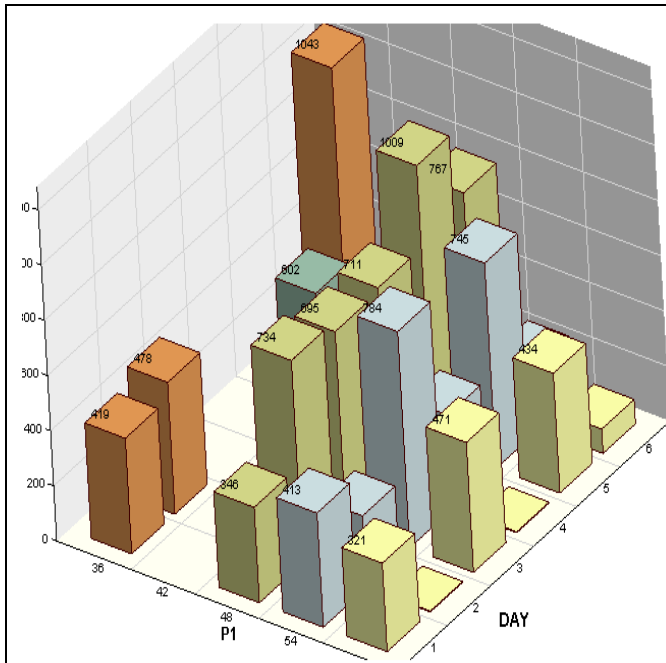


Illustration 7: The relationship between the price and profit day for each day 1-6 for orange type 1

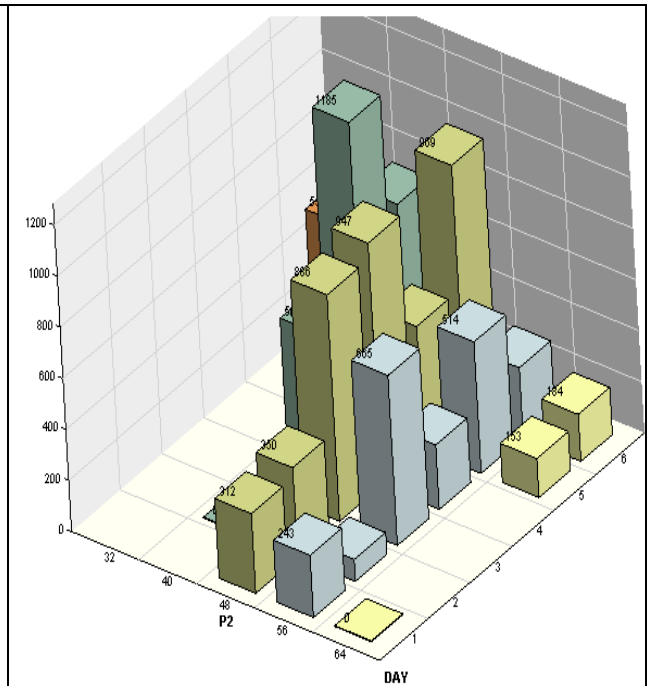


Illustration 8: The relationship between the price and profit for each day 1-6 for orange type 2

The axis are day, price and profit gained. The graphs give a good picture of how to price the oranges. However, since it can be hard to get a precise overview, I would attach the 6nd dimensional graphs for each orange type so that the business management could look at the specific days if they wanted more details.

Illustration 7 shows that the average best price found in question 1 fits very well for most days. Day 6 on the other hand, which is Saturday, is different in that the lowest price results in the highest profit gained. For orange type 1, the oranges should be priced according to the average best price found in question 1 for days Monday to Friday, and priced much lower (36) at Saturday.

Unlike illustration 7, the average best price doesn't fit so well for illustration 8. This is because day 5 has a high profit gain at a lower than average price, which influences the data when the days aren't considered. For O2, the best price is the same as for O1 for the days [1;4] and 6, and then lower for day 5.

Conclusion for part 2

Using the SAS enterprise guide as Business Intelligence tool I've come up with answers to the three questions the management wanted to investigate. I've concluded that for each question, the graph shown would be the main arguing material, and understandable for the management, but in some cases assisted by additional attachments for further precision.

Literature

[Kimball]

The Data Warehouse Lifecycle Toolkit, second edition, 2008

ISBN: 978-0-470-14977-5