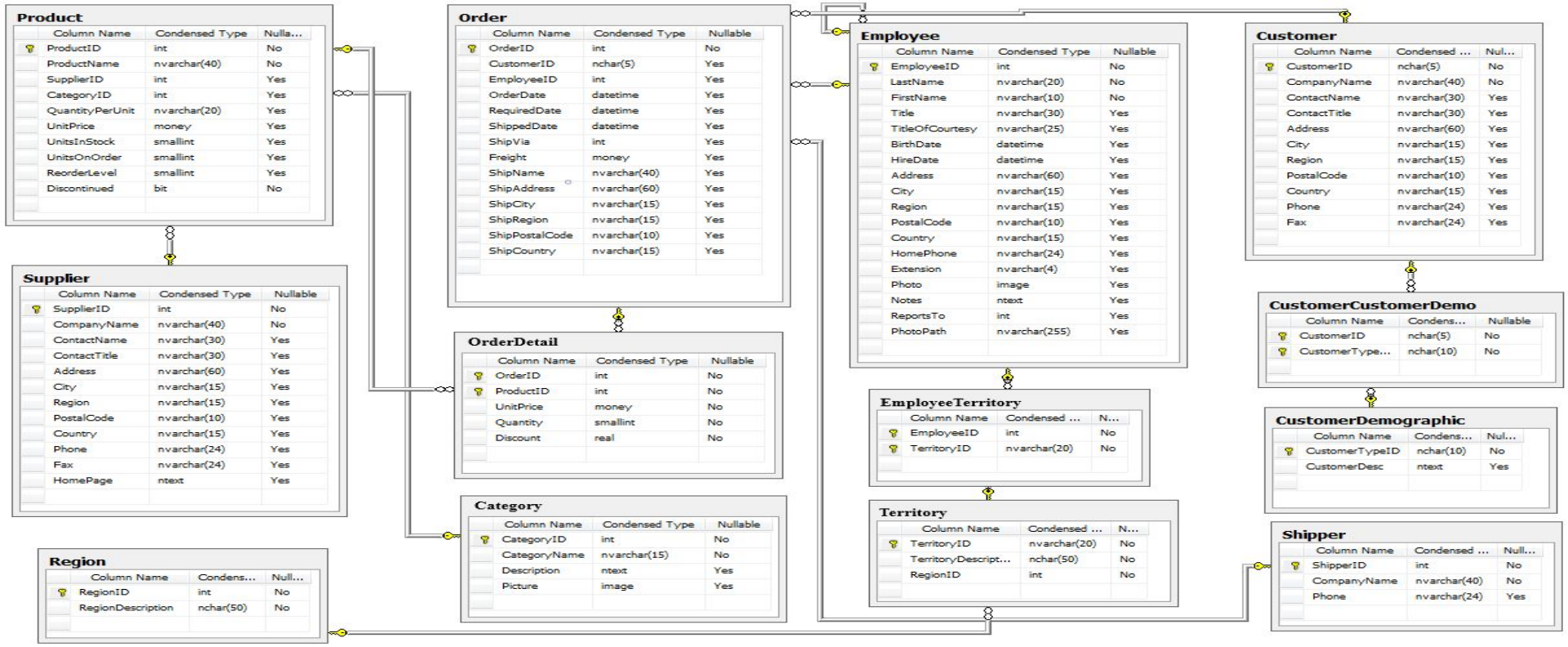
Hypothesis testing project

Northwind Database

By Luigi Fiori

We have been hired as consultants from the Alpha Company in order to increase their overall Revenue.
We have been asked to check their Database and to find valid solutions to increase their competitivity and efficiency in the Market.
The Database that we will be using is a free-source Google Database called: Northwind
Firstly in our approach we will be inspecting the Database.



TERMINOLOGY

- **Null Hypothesis:** is the conservative hypothesis(e.g. there's no variation in what we're observing)
- **Alternative Hypothesis:** states that the result is smaller, greater or different than the value in the Null Hypothesis.
- **P-value**: is the probability of an event happening due to random chance
- **Effect size:** tells us how 'big' the difference is between 2 observed groups.

After a first exploration of the Database we want to focus our analysis on 6 different questions:

1. Does discount amount have a statistically significant effect on the quantity of a product in an order? If so, at what level(s) of discount?

2. Is it December, considering the Christmas time period in it, the month with the highest revenue?

3. Does discount amount have a statistically significant effect on the revenue of a product in an order? If so, at what level(s) of discount?

4. There is a statistically significant difference on shipping time between different shipper companies?

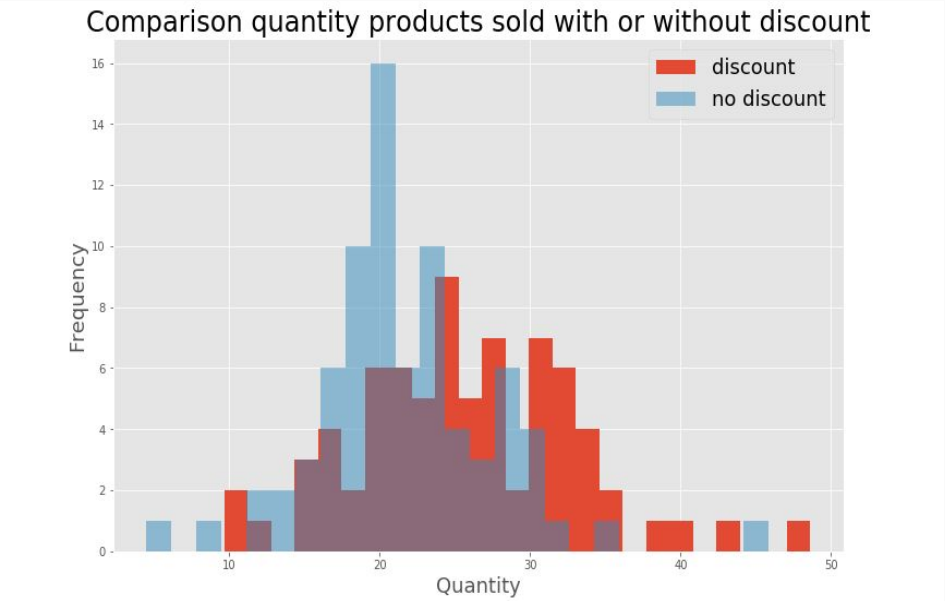
5. There is a statistically significant difference in terms of Revenue for different countries?

6. There is statistically significant difference in terms of Revenue between different categories?

Let’s start to answer it!

1. Does discount amount have a statistically significant effect on the quantity of a product in an order? If so, at what level(s) of discount?

Null Hypothesis = there is no statistically difference in quantity for different discounts
Alternative Hypothesis = there is statistically difference in quantity for different discounts

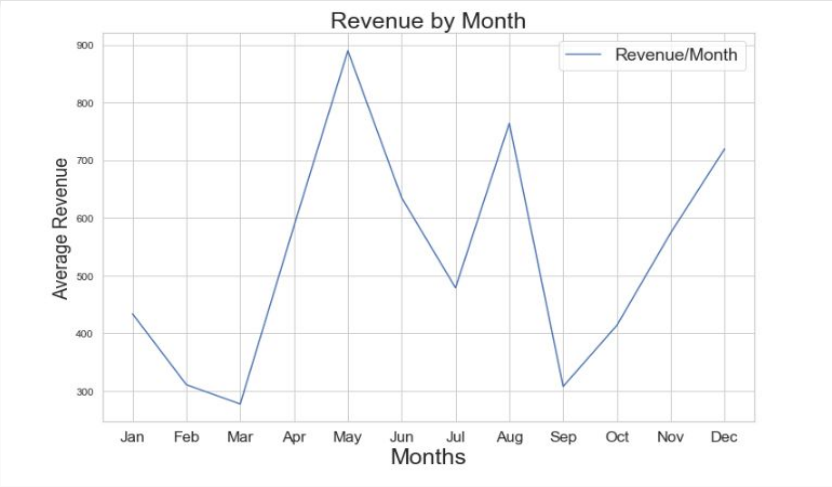


Discount	P-value	Null H.	Cohen's
5%-10%	0.26	Fail	0.12
10%-20%	0.33	Fail	-0.1
5%-25%	0.59	Fail	0.05

There' no statistically significant difference for different levels of discount.
There's no difference in quantity sold when applying different discounts.
Our advice is, to use a 5% discount, such that being the one with the highest effect size, and the smallest discount, leaves us with the highest profit.

2. Is it December, considering the Christmas time period in it, the month with the highest revenue?

Null Hypothesis = Revenue for December <= Revenue other month
Alternative Hypothesis = Revenue for December > Revenue other month

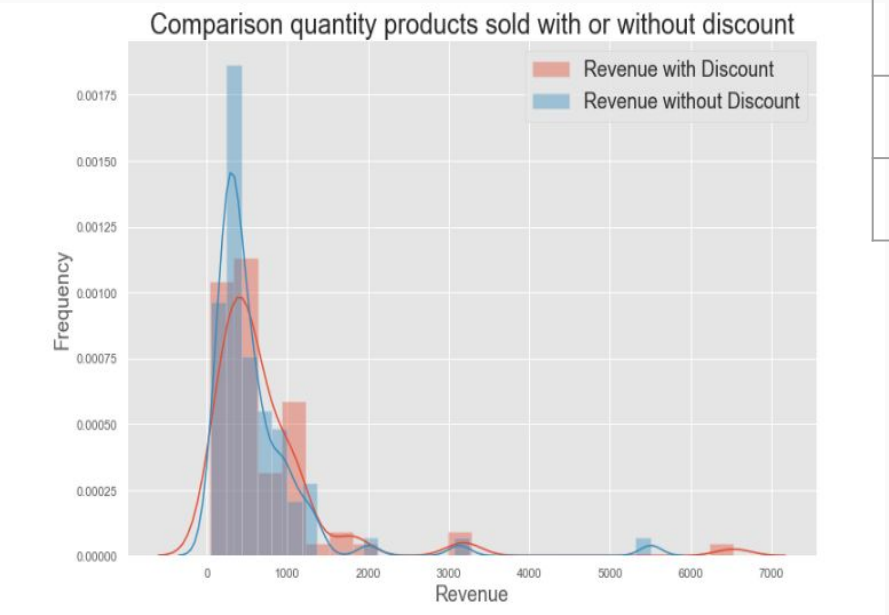


	Month	P-value	Null Hypothesis	Cohens d
4	5	2.587298e-01	Failed	0.210380
7	8	7.235247e-01	Failed	0.061308
5	6	5.983209e-01	Failed	-0.113364
3	4	2.030295e-01	Failed	-0.190250
10	11	1.838974e-01	Failed	-0.229959
6	7	3.153316e-02	Reject	-0.381989
0	1	8.368373e-03	Reject	-0.413633
9	10	4.994054e-03	Reject	-0.479868
8	9	3.410447e-04	Reject	-0.629506
1	2	6.892920e-06	Reject	-0.730884
2	3	2.164679e-07	Reject	-0.805945

As we can see we failed to reject our hyphothesis for April, May, June, August and November.
This means that our Revenue is actually higher on these months compared to December.
Our advice is to focus our product strategies during these months, May furthermore gives us the highest effect size.
In fact could probably be more difficult trying to increase the overall Revenue during December, being the potential competition much higher and obstinate.

3. Does discount amount have a statistically significant effect on the revenue of a product in an order? If so, at what level(s) of discount?

Null Hypothesis = there is no statistically difference in revenue for different discounts
Alternative Hypothesis = there is statistically difference in revenue for different discounts



Discount	P-value	Null H	Cohen's
5%-20%	0.04	Reject	0.20
20%-25%	0.04	Reject	- 0.19

We reject our null hypothesis for 5% and 20% discounts and 20% and 25% discounts.
This means that there's statistical significant difference in terms of Revenue taking in consideration these 2 different levels of discount.
Our advice is to focus our attention on 5%-20% , the d Cohen's effect size is in fact significantly higher and the possibly strategies leaving us with more profit compared to 20%-25%.

4. There is a statistically significant difference on shipping time between different shipper companies?

Null Hypothesis = there is no statistically difference between different companies in terms of shipping time
Alternative Hypothesis = there is statistically difference between different companies in terms of shipping time


```
1 print(pairwise_tukeyhsd(v, labels, 0.05))
```

```
  Multiple Comparison of Means - Tukey HSD,FWER=0.05
=====================================================
   group1       group2   meandiff lower  upper  reject
-----------------------------------------------------
federal_ship speedy_ship  0.9739  0.0934 1.8544  True
federal_ship united_ship  1.4712  0.639  2.3033  True
speedy_ship  united_ship  0.4973  -0.336 1.3305 False
-----------------------------------------------------
```

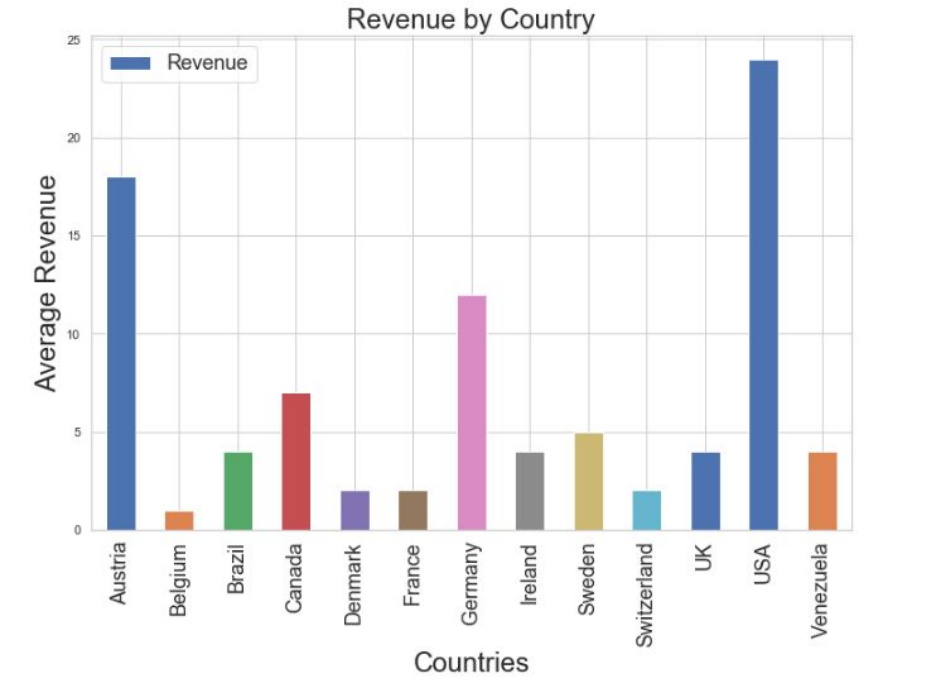
The test shows clearly that Federal Shipping is more efficient .

We advice the management to understand why the others companies are taking longer and in case to ask for a discount or to switch all the orders to the Federal shipper.

5. There is a statistically significant difference in terms of Revenue for different countries?

Null Hypothesis = there is no statistically difference between Europe and North-America in terms of Revenue

Alternative Hypothesis = there is statistically difference between Europe and North-America in terms of Revenue



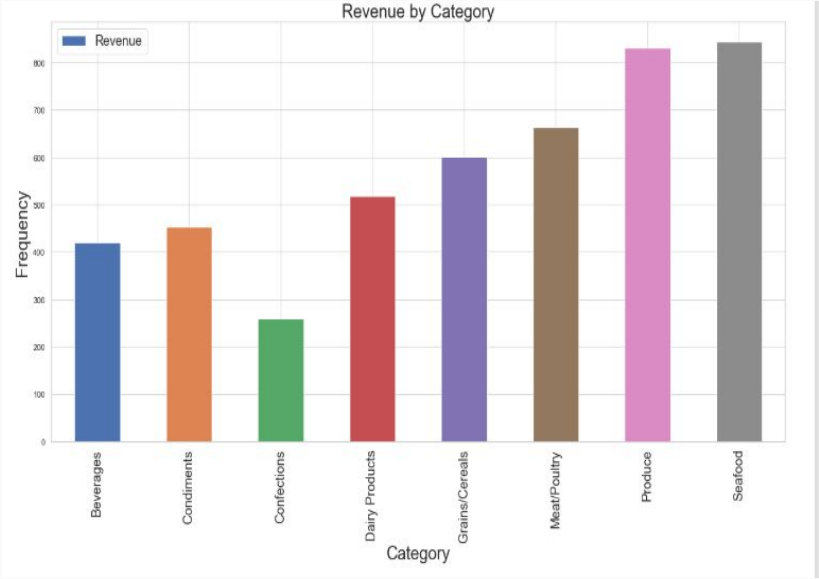
P-value = 0.0

We reject the Null Hypothesis for 2 Groups, Europe and North-America, compared in terms of Revenue.

The Profit is the same in proportion no matter which Group we take in consideration.

6. There is statistically significant difference in terms of Revenue between different categories?

Null Hypothesis = there is no statistically difference between different categories in terms of Revenue
Alternative Hypothesis = there is statistically difference between different categories in terms of Revenue



Group 1	Group 2	Null Hypothesis
Beverages	Produce	Reject
Beverages	Seafood	Reject
Condiments	Seafood	Reject

From our test results that there's a statistically significant difference between some of the categories.
Comes out, in fact, that the categories with the highest Revenue are Seafood, Produce and Meat/Poultry.
In particular we recommend to focus our production on Produce.
Being the description of this category "Dried fruit and bean curd" we know that are easier to conserve and ship compared to Seafood and Meat/Poultry.

Future work

In future, having more time, we could explore more in depth some aspects of our Dataset like the impact of each category on each Country.
Would be really interesting also having more data about the cost of each spedition for both Shippers and Suppliers to check the efficiency.
Finally would be interesting to get a more in depth analysis about the Revenue of each Country.

Thanks for watching!