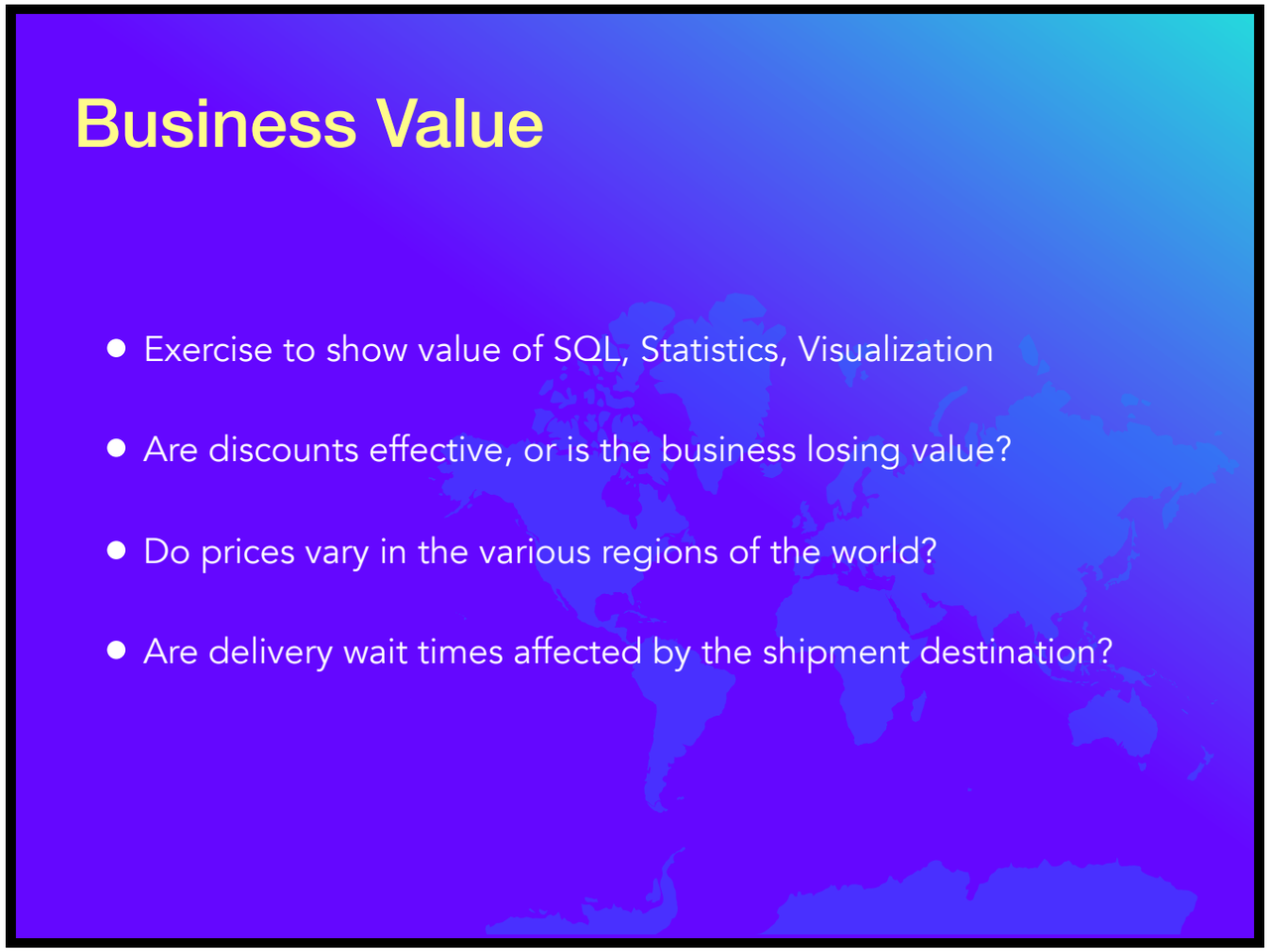
Statistical Insights:
Northwind Traders Database*

* A *very* fictional dataset

By Valentina Fontiveros
2019

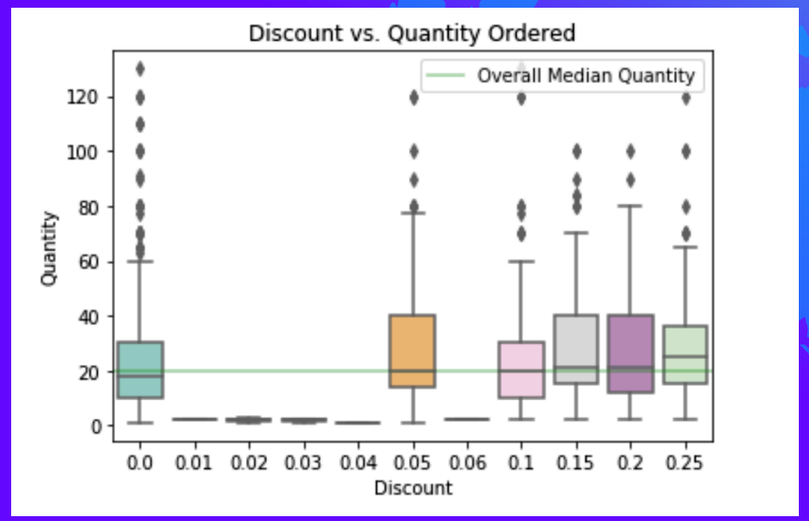
Hello, my name is Valentina and I will present to you today the results of several hypothesis tests done on the classical Northwind Traders Database.



The purpose of doing this project is to show the business value of integrating database information with statistical and visualization analyses.
The three main questions asked in this project are:

- Are discounts effective in incentivizing bulk ordering, or is the business losing value?
- Do prices vary in the various regions of the world?
- Are delivery wait times affected by the shipment destination?

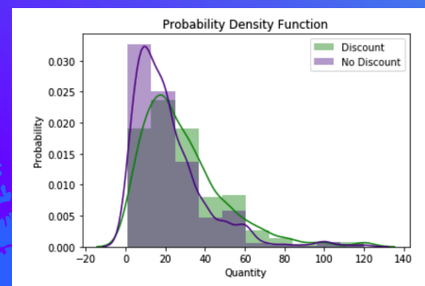
Do discounts guarantee bulk orders?



From a cursory look to mean quantities ordered, it would seem that only a discount of 25% would result in a significant incentive to change the median quantity ordered by customers. However, visualizing these relationships do not provide the necessary resolution to answer the question authoritatively.

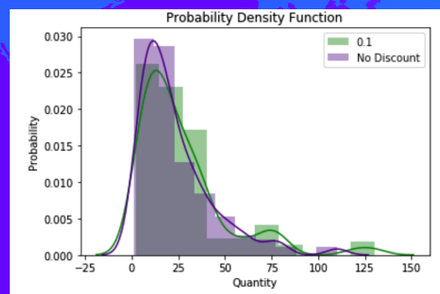
Do discounts guarantee bulk orders?

Discount	Mean P value 1000 trials	Significant?
5%	0.03	✓
10%	0.20	✗
15%	0.03	✓
20%	0.06	✗
25%	0.03	✓



✓

Ttest_indResult(statistic=7.592263710187593, pvalue=5.217529356034328e-14)

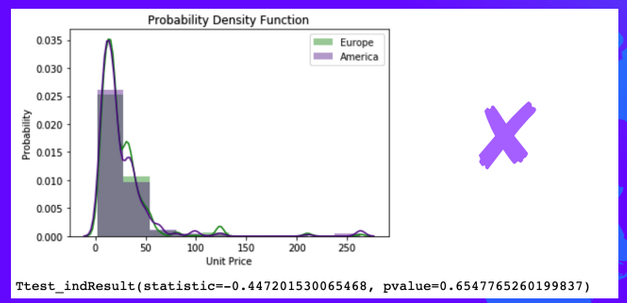


✗

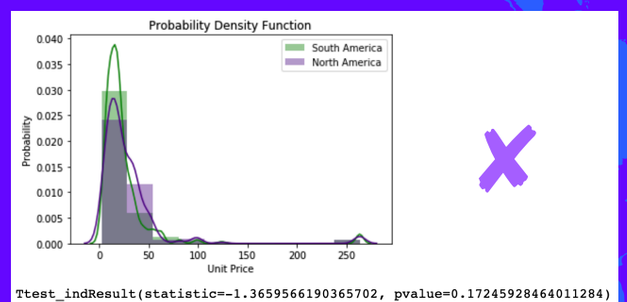
Ttest_indResult(statistic=1.3264263804621745, pvalue=0.18557842581651013)

Instead, data scientists perform what we call t-tests: we compare two samples, in this case quantities ordered with a discount and no discount, and we determine how statistically different they are by using mathematical formulae. In this example, we found that discounts do change the quantities ordered. However, when we look closer at various discount brackets, we see that some are more effective than others. Discounts of 10% and 20% are not statistically different as the no-discount case, whereas discounts of 5%, 15%, and 25% are. This finding merits further research!

Does shipment destination affect price?



Ttest_indResult(statistic=-0.447201530065468, pvalue=0.6547765260199837)

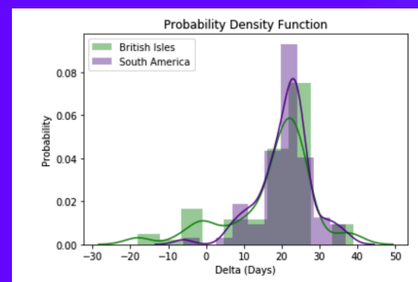


Ttest_indResult(statistic=-1.3659566190365702, pvalue=0.17245928464011284)

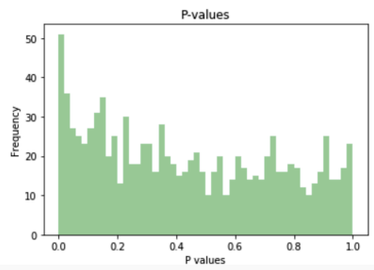
COMPARING	P VALUE	Significant?
EUROPE VS. AMERICAS	0.65	✗
NORTH VS. SOUTH AMERICA	0.17	✗

Another question we asked was: does shipment destination affect price? We looked at various regions in the world, and the surprising conclusions was that, in fact, they are not affected. This is frankly unrealistic from a business perspective and may be a consequence of using a synthetic dataset!

Does Geography Affect Delivery Times?



Ttest_indResult(statistic=-2.202425756106206, pvalue=0.02918224062520203)



Mean p-value for all experiments: 0.43620651248311854

?

What's going on here?

Sample size matters!

The last question we asked was, does geography affect delivery times? Similarly to the previous finding, this answer was also a clear-cut negative. However, when comparing delivery times between South America and the British Isles, we got an interesting result. A requirement for two sample t-tests is to use samples with the same number of entries. On this dataset, deliveries to the British Isles were limited, less than 50. Therefore, the sample size was quite limited. There was high instability when computing p-values. In some instances, the p-value would come out lower than 0.05, which is the significance threshold used. In the majority of trials, however, the p-value was much higher than the significance threshold, resulting in an average p-value of 0.44 for 1000 trials. The importance of this is that 1) We need to review sample sizes in automated analyses and that 2) some questions require executing various t-test trials to reach a valid conclusion.

Conclusions

- Not all discount levels significantly affect order quantity.
- All significant discounts ended in # 5
- Shipment destination does not affect price.
- Shipment destination does not affect delivery wait times.

While this is most likely a synthetic dataset, we found some interesting insights:
Not all discount levels significantly affect order quantity.
All significant discounts ended in # 5
Shipment destination does not affect price.
Shipment destination does not affect delivery wait times.

Future Work

- Repeat a similar analysis using a real corporate dataset.
- Explore the relationship between discount numbers and quality bought (is there a psychological link?)

For future work, it would be interesting to repeat this analysis with a real corporate dataset.
If in fact, the discount information is valid, researching the potential psychological link to various discount numbers would be of value for the marketing department.

Thank you! (By the way this is the original logo of the fictional corporation created by Microsoft).