Northhwind Database Statistical Analysis

Akshay Ghalsasi

Flatiron Project 3

Data

A database schema diagram showing the following tables and their columns:

Product

Column Name	Condensed Type	Nulla...
ProductID	int	No
ProductName	nvarchar(40)	No
SupplierID	int	Yes
CategoryID	int	Yes
QuantityPerUnit	nvarchar(20)	Yes
UnitPrice	money	Yes
UnitsInStock	smallint	Yes
UnitsOnOrder	smallint	Yes
ReorderLevel	smallint	Yes
Discontinued	bit	No

Order

Column Name	Condensed Type	Nullable
OrderID	int	No
CustomerID	nchar(5)	Yes
EmployeeID	int	Yes
OrderDate	datetime	Yes
RequiredDate	datetime	Yes
ShippedDate	datetime	Yes
ShipVia	int	Yes
Freight	money	Yes
ShipName	nvarchar(40)	Yes
ShipAddress	nvarchar(60)	Yes
ShipCity	nvarchar(15)	Yes
ShipRegion	nvarchar(15)	Yes
ShipPostalCode	nvarchar(10)	Yes
ShipCountry	nvarchar(15)	Yes

Employee

Column Name	Condensed Type	Nullable
EmployeeID	int	No
LastName	nvarchar(20)	No
FirstName	nvarchar(10)	No
Title	nvarchar(30)	Yes
TitleOfCourtesy	nvarchar(25)	Yes
BirthDate	datetime	Yes
HireDate	datetime	Yes
Address	nvarchar(60)	Yes
City	nvarchar(15)	Yes
Region	nvarchar(15)	Yes
PostalCode	nvarchar(10)	Yes
Country	nvarchar(15)	Yes
HomePhone	nvarchar(24)	Yes
Extension	nvarchar(4)	Yes
Photo	image	Yes
Notes	ntext	Yes
ReportsTo	int	Yes
PhotoPath	nvarchar(255)	Yes

Customer

Column Name	Condensed ...	Nul...
CustomerID	nchar(5)	No
CompanyName	nvarchar(40)	No
ContactName	nvarchar(30)	Yes
ContactTitle	nvarchar(30)	Yes
Address	nvarchar(60)	Yes
City	nvarchar(15)	Yes
Region	nvarchar(15)	Yes
PostalCode	nvarchar(10)	Yes
Country	nvarchar(15)	Yes
Phone	nvarchar(24)	Yes
Fax	nvarchar(24)	Yes

Supplier

Column Name	Condensed Type	Nullable
SupplierID	int	No
CompanyName	nvarchar(40)	No
ContactName	nvarchar(30)	Yes
ContactTitle	nvarchar(30)	Yes
Address	nvarchar(60)	Yes
City	nvarchar(15)	Yes
Region	nvarchar(15)	Yes
PostalCode	nvarchar(10)	Yes
Country	nvarchar(15)	Yes
Phone	nvarchar(24)	Yes
Fax	nvarchar(24)	Yes
HomePage	ntext	Yes

OrderDetail

Column Name	Condensed Type	Nullable
OrderID	int	No
ProductID	int	No
UnitPrice	money	No
Quantity	smallint	No
Discount	real	No

CustomerCustomerDemo

Column Name	Condens...	Nullable
CustomerID	nchar(5)	No
CustomerType...	nchar(10)	No

EmployeeTerritory

Column Name	Condensed ...	N...
EmployeeID	int	No
TerritoryID	nvarchar(20)	No

CustomerDemographic

Column Name	Condens...	Nul...
CustomerTypeID	nchar(10)	No
CustomerDesc	ntext	Yes

Category

Column Name	Condensed Type	Nullable
CategoryID	int	No
CategoryName	nvarchar(15)	No
Description	ntext	Yes
Picture	image	Yes

Territory

Column Name	Condensed ...	N...
TerritoryID	nvarchar(20)	No
TerritoryDescript...	nchar(50)	No
RegionID	int	No

Region

Column Name	Condens...	Null...
RegionID	int	No
RegionDescription	nchar(50)	No

Shipper

| Column Name	Condensed ...	Null...
ShipperID	int	No
CompanyName	nvarchar(40)	No
Phone	nvarchar(24)	Yes

Questions

- 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?

- Do sales get better year on year? Does the month when the products were ordered matter?

- Is there a standout employee in terms of the distribution from which quantities sold are drawn ? Are there employees who are shipping their orders significantly late?

- Are there regions that are growing or decreasing in overall sales ? Are there product categories that are growing or decreasing in overall sales?

Q1

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?

OrderId	ProductId	UnitPrice	Quantity	Discount
10248	11	14.0	12	0.0
10248	42	9.8	10	0.0
10248	72	34.8	5	0.0
10249	14	18.6	9	0.0
10249	51	42.4	40	0.0

Log transformed data at various discount levels
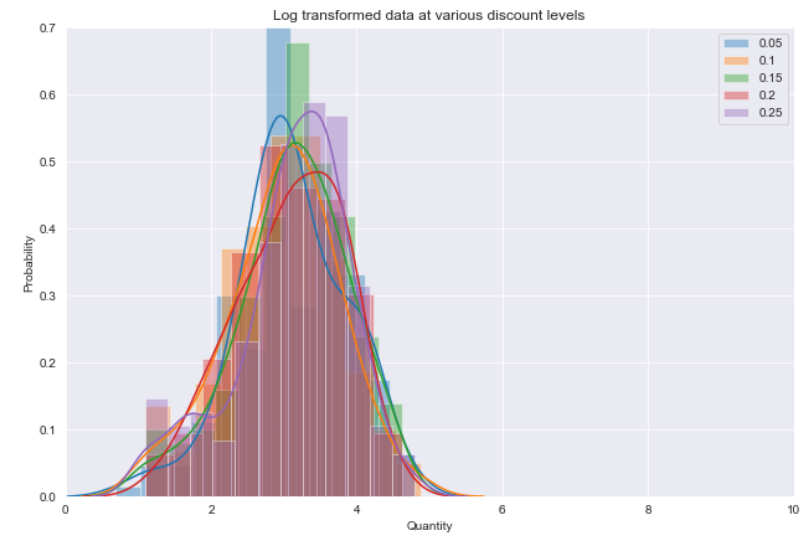
Q1 - Conclusions

- Considering individual discount levels, discounted products at 0.05, 0.15, 0.2, and 0.25 levels are drawn from a a different distribution compared to the undiscounted quantities and have higher means. There is no concrete distinction in the distributions between 0.05, 0.15, 0.2 and 0.25 levels. 0.1 discount level needs to have more data points for us to potentially reject the null hypothesis

- At the individual product level there is no evidence that discount levels matter, However that is most likely due to low statistics per product and per discount level.

Discount	Welch	KS	Power
0.05	0.000017	0.001530	0.891
0.10	0.022100	0.153000	0.403
0.15	0.000022	0.002000	0.886
0.20	0.000097	0.005980	0.810
0.25	0.000051	0.000082	0.869

Q2

- Do sales get drawn from a different distribution every year? Does the month when the products were ordered matter?
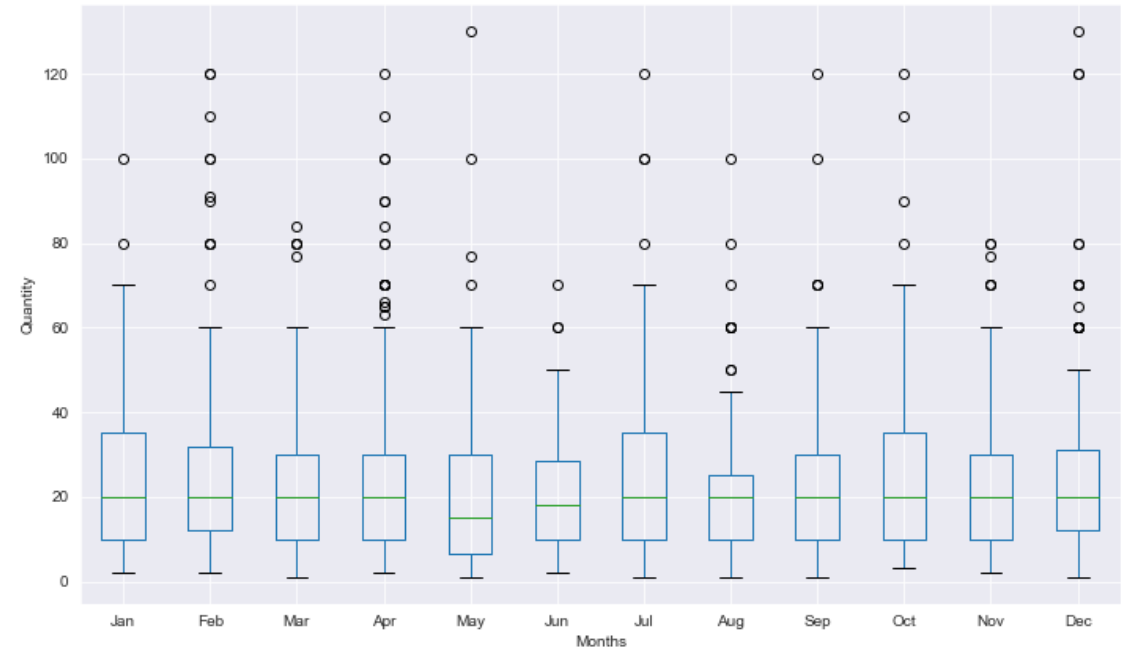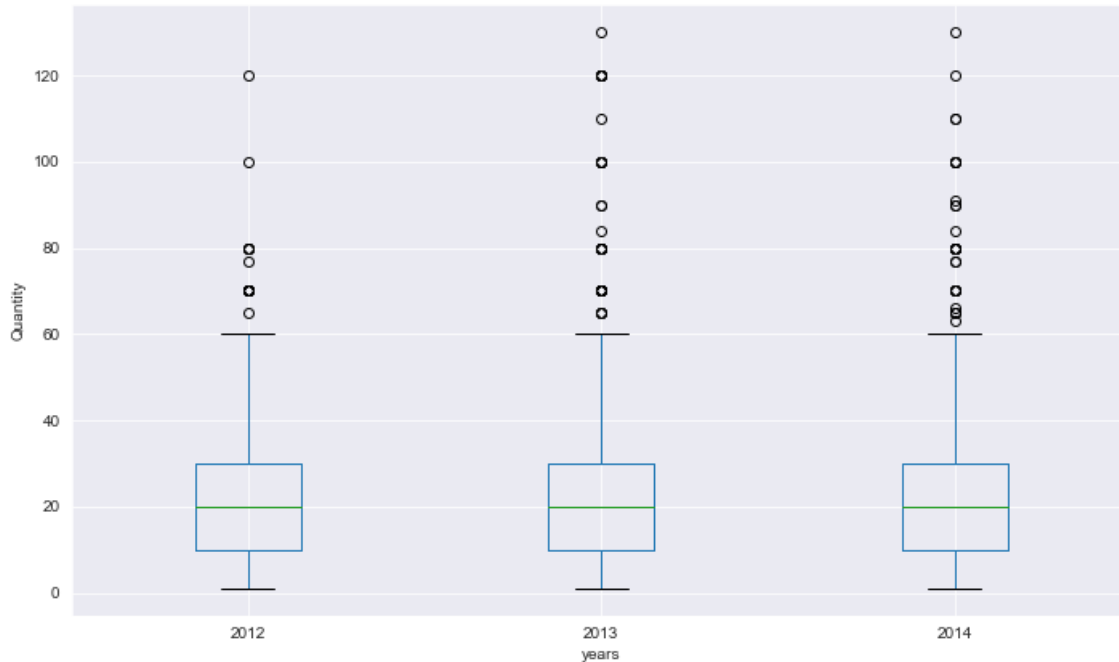
Q2-Conclusions

- The sales do not seem to depend on year and the quantities seem to be drawn from the same underlying distribution.

- For overall sales the month of May shows a definite decrease in quantities purchased
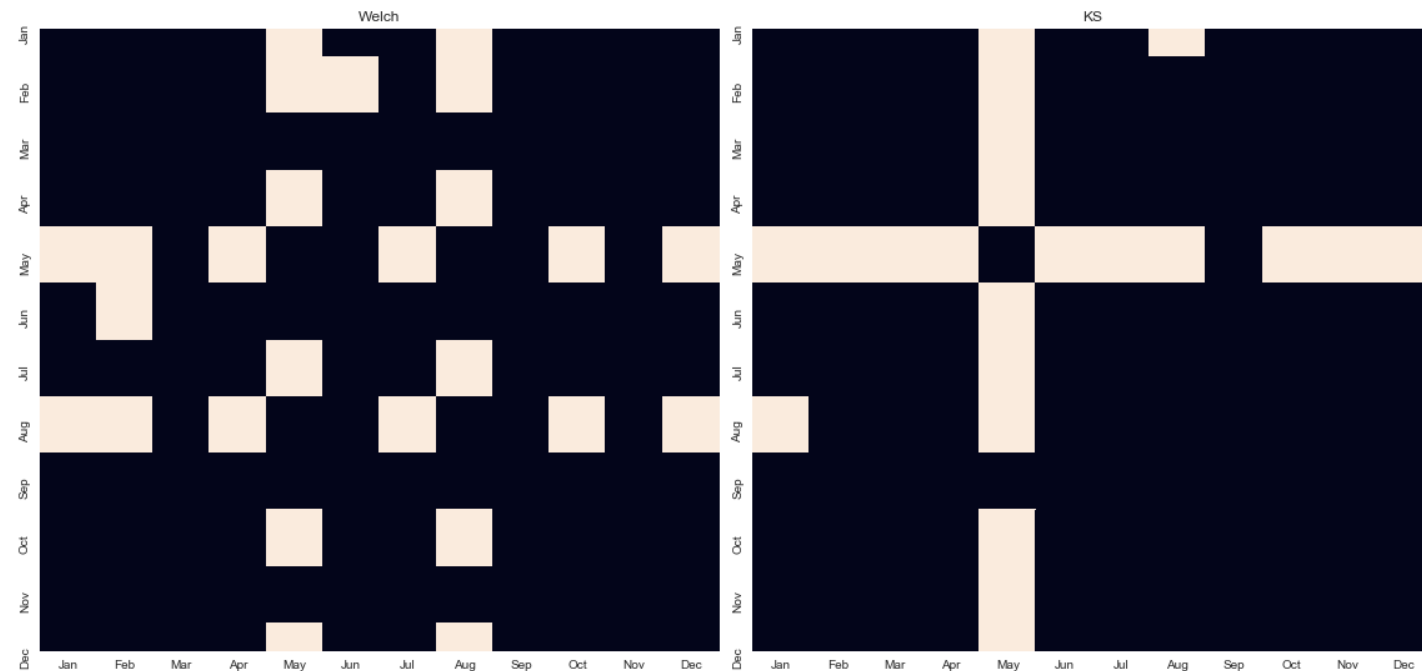
Q3 and Conclusions

- Is there a standout employee in terms of the distribution from which quantities sold are drawn ? Are there employees who are shipping their orders significantly late?

Conclusions

- There is insufficient evidence that the quantities sold between employees come from different distributions

- Employees 2,4,7,9 statistically have a bad record of shipping on time compared to Employee 5
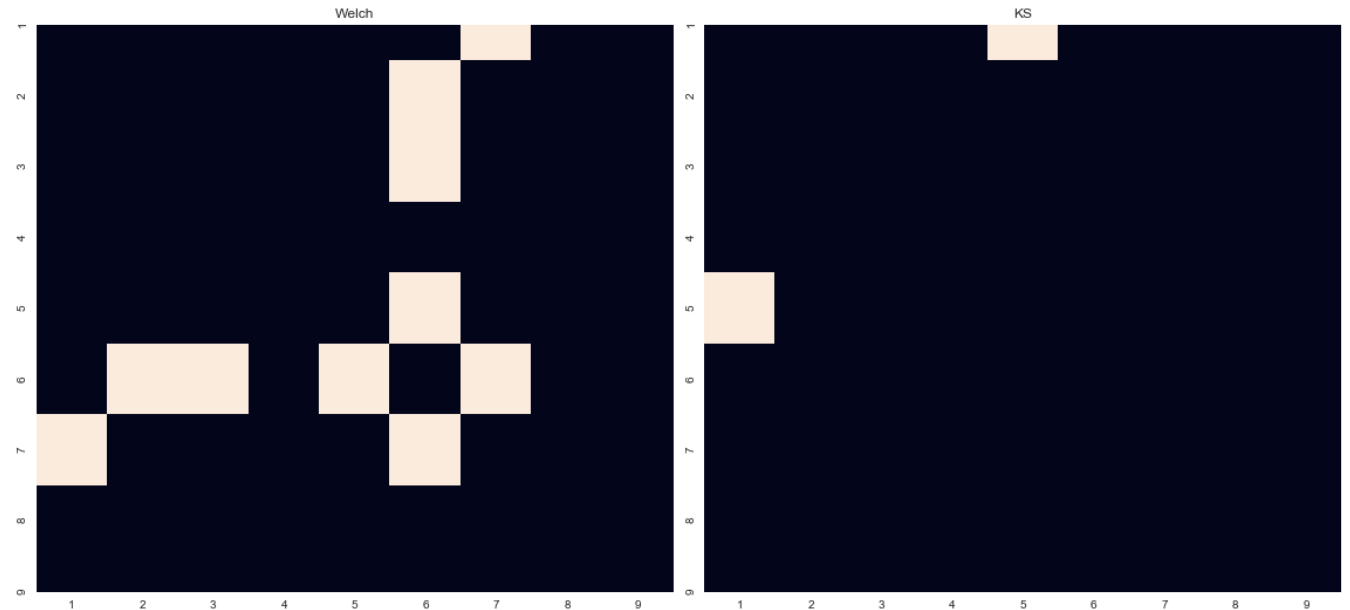
Q4 and Conclusions

Q4 Are there regions that are growing or decreasing in overall sales ? Are there products that are growing or decreasing in overall sales

Conclusions

- All regions show a growth comparing year 2012 and 2014

- All products show a growth comparing year 2012 and 2014

Region	p_2012	p_2014	z-statistic
Western Europe	0.015649	0.034579	40.454319
South America	0.004175	0.012373	30.588491
Central America	0.000818	0.001130	3.283594
North America	0.007328	0.019786	36.290699
Northern Europe	0.002423	0.005969	18.429383
Scandinavia	0.000818	0.001190	3.870762
Southern Europe	0.001776	0.003979	13.770165
British Isles	0.004270	0.007088	12.409567
Eastern Europe	0.000180	0.000557	6.654723

Category	p_2012	p_2014	z-statistic
Dairy Products	0.007455	0.014269	22.126396
Grains/Cereals	0.002083	0.007446	26.854013
Produce	0.002118	0.004652	14.744752
Seafood	0.005046	0.014782	33.390538
Condiments	0.003774	0.007429	16.397729
Confections	0.005307	0.012869	26.845962
Beverages	0.007151	0.019444	36.230211
Meat/Poultry	0.003792	0.005759	9.416188