
</talentlabs>
Multivariate Analysis

Agenda

- More variables – lets talk multivariate
- Heatmaps
- Scatterplots
- Grouped Boxplots
- Grouped and Stacked Bar charts
- Three dimensional plots
- Cluster Analysis
- Principal Component Analysis

</talentlabs>

Why Multivariate Analysis?

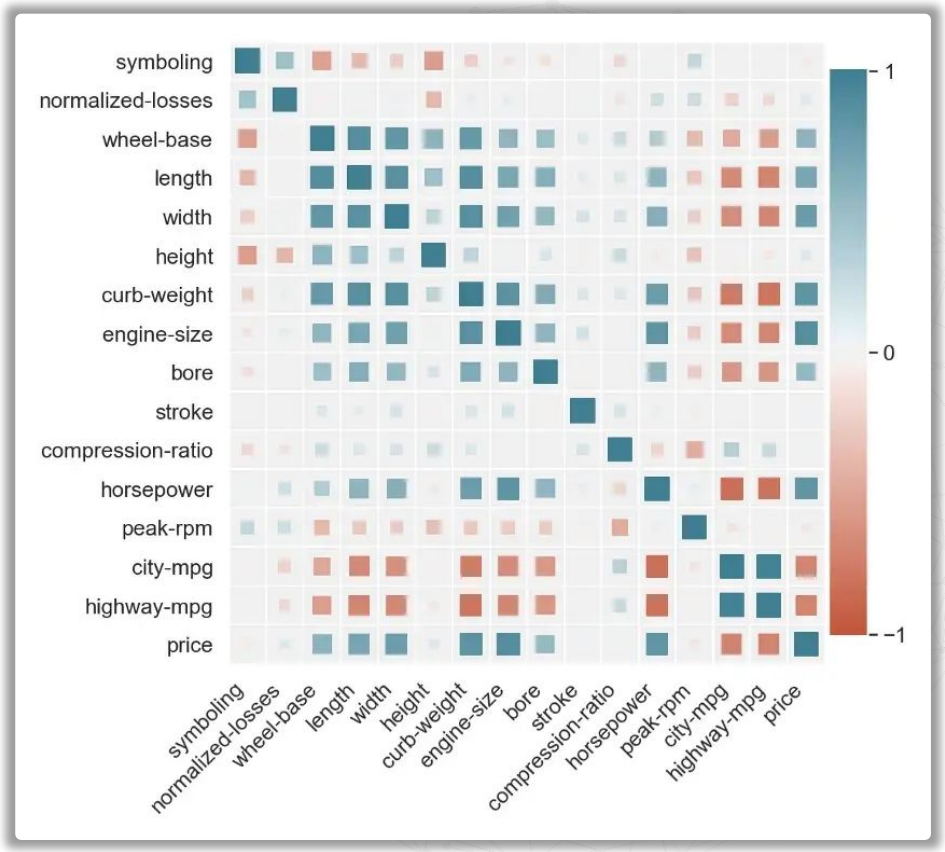
Data reduction or structural simplification	This helps data to get simplified as possible without sacrificing valuable information. This will make interpretation easier.
Sorting and grouping	When we have multiple variables, groups of “similar” objects or variables are created, based upon measured characteristics.
Investigation of dependence among variables	The nature of the relationships among variables is of interest. Are all the variables mutually independent or are one or more variables dependent on the others?
Predict Relationships between variables	Must be determined for the purpose of predicting the values of one or more variables based on observations on the other variables.
Hypothesis construction and testing	Specific statistical hypotheses, formulated in terms of the parameters of multivariate populations, are tested. This may be done to validate assumptions or to reinforce prior convictions.

</talentlabs>

Heatmaps for Numerical Variables

- Automobiles data - containing various characteristics of a number of cars.
- Green means positive, red means negative. The stronger the color, the larger the correlation magnitude.
- Size shows the magnitude or the measure of correlation coefficient

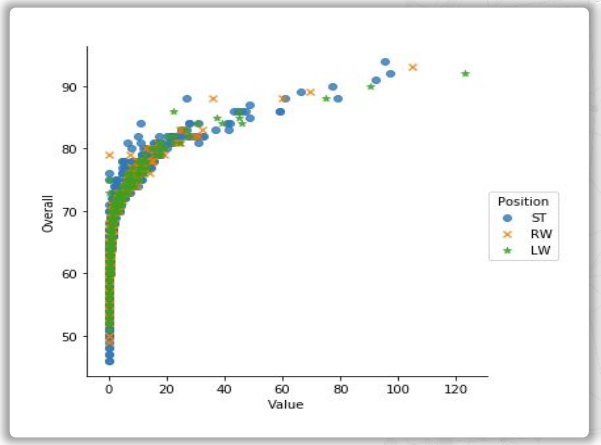
size(c1, c2) ~ abs(corr(c1, c2))

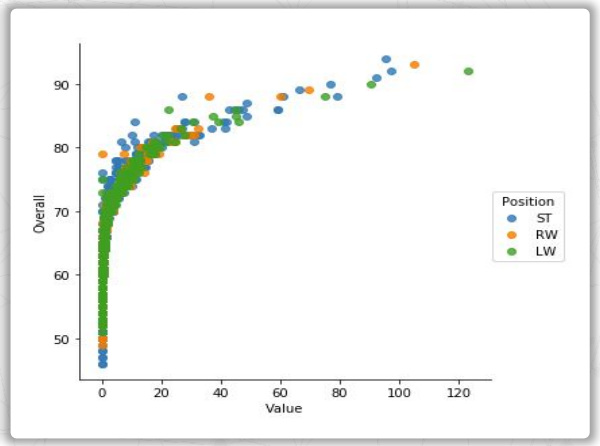


</talentlabs>

Multivariate Scatterplots

- Suppose that we are interested in seeing which type of offensive players tends to get paid the most: the striker, the right-winger, or the left-winger.
- This scatterplot uses three visual variables. The horizontal position (x-value) tracks the Value of the player (how well they are paid). The vertical position (y-value) tracks the Overall score of the player across all attributes. And the color (the hue parameter) tracks which of the three categories of interest the player the point represents is in.
- Another example visual variable is shape. Shape controls, well, the shape of the marker.

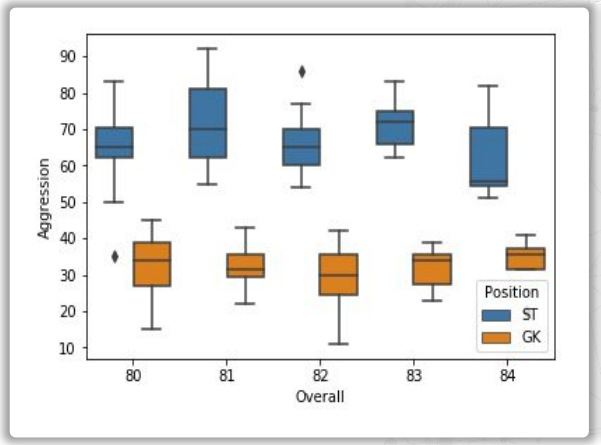


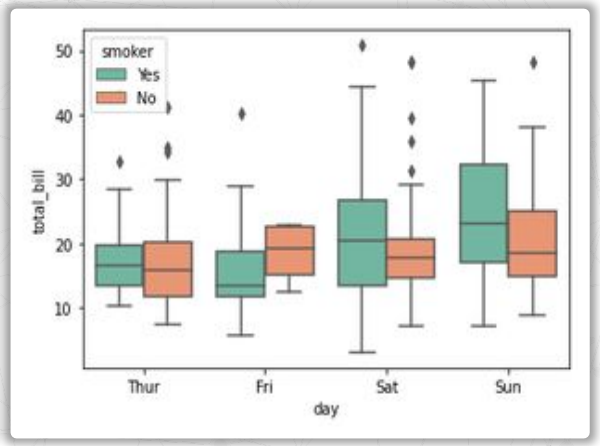


</talentlabs>

Multivariate Boxplots

- Suppose we're interested in the following question: do Strikers score higher on "Aggression" than Goalkeepers do?
- As you can see, this plot demonstrates conclusively that within our datasets goalkeepers (at least, those with an overall score between 80 and 85) have much lower Aggression scores than Strikers do.
- In this plot, the horizontal axis encodes the Overall score, the vertical axis encodes the Aggression score, and the grouping encodes the Position.
- Grouping is an extremely communicative visual variable: it makes this chart very easy to interpret. However, it has very low cardinality: it's very hard to use groups to fit more than a handful of categorical values.

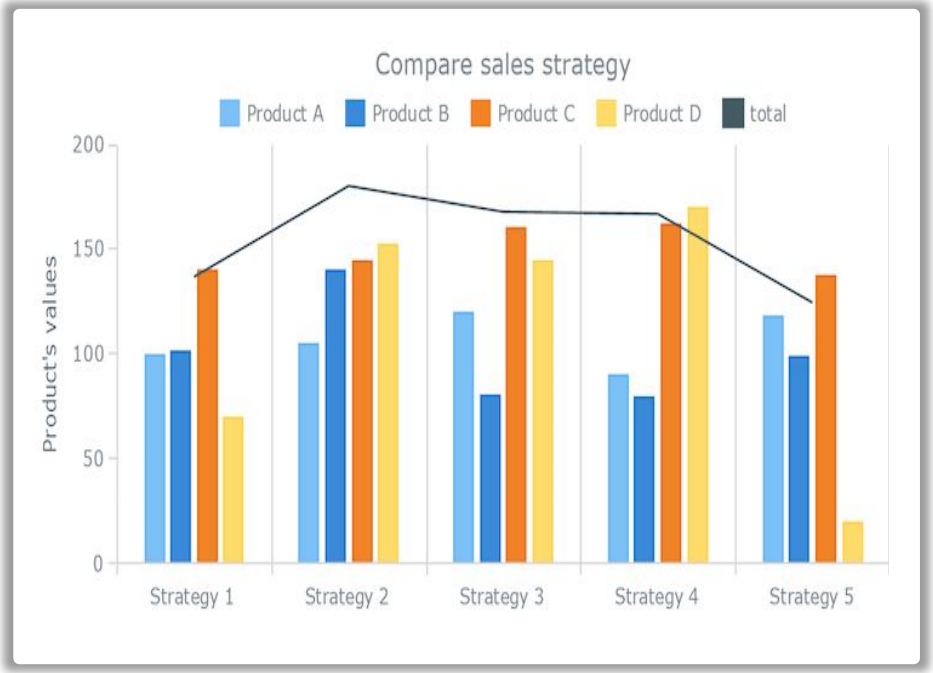




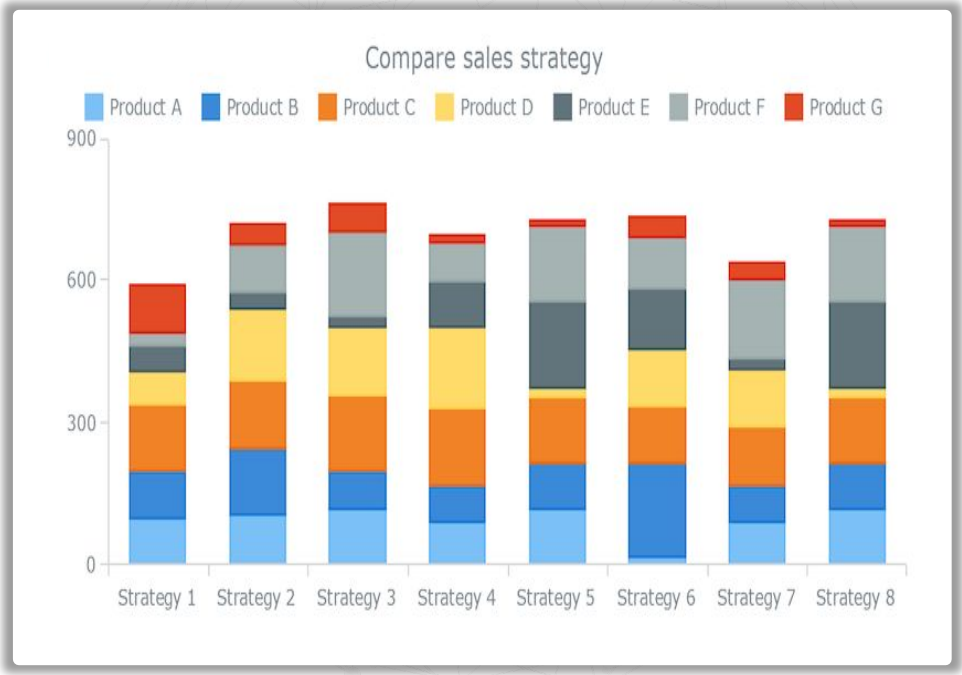
</talentlabs>

Stacked and Grouped Bar charts

Sales Strategy across product lines using Grouped Bar Chart



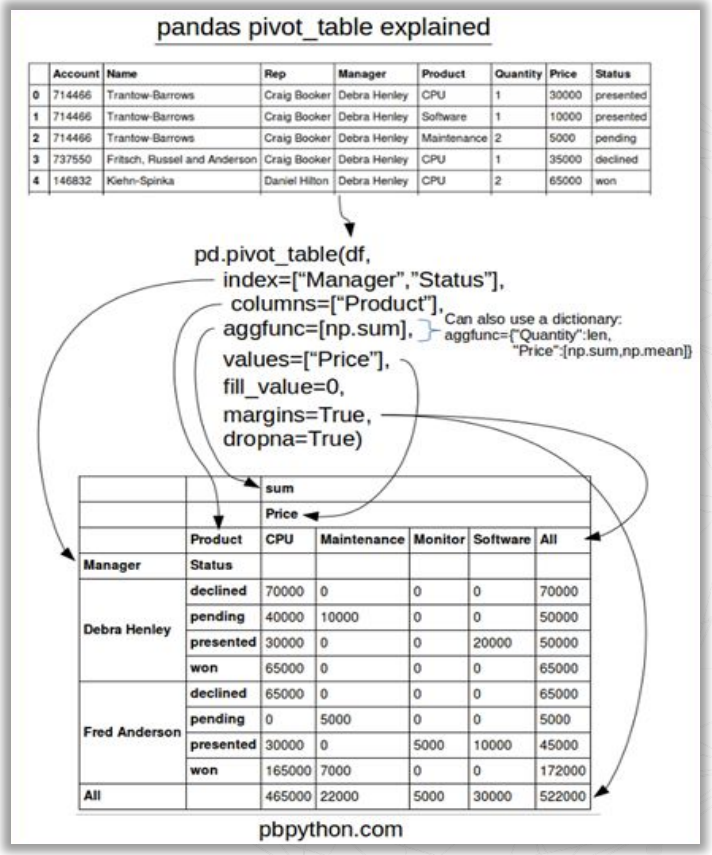
Sales Strategy across product lines using Stacked Bar Chart



</talentlabs>

Pivot tables

- Wrangle the data around the pivot to analyze better.
- Derived from Excel.
- Shift the data and aggregate it around the pivot.
- Explains better than normal dataframes



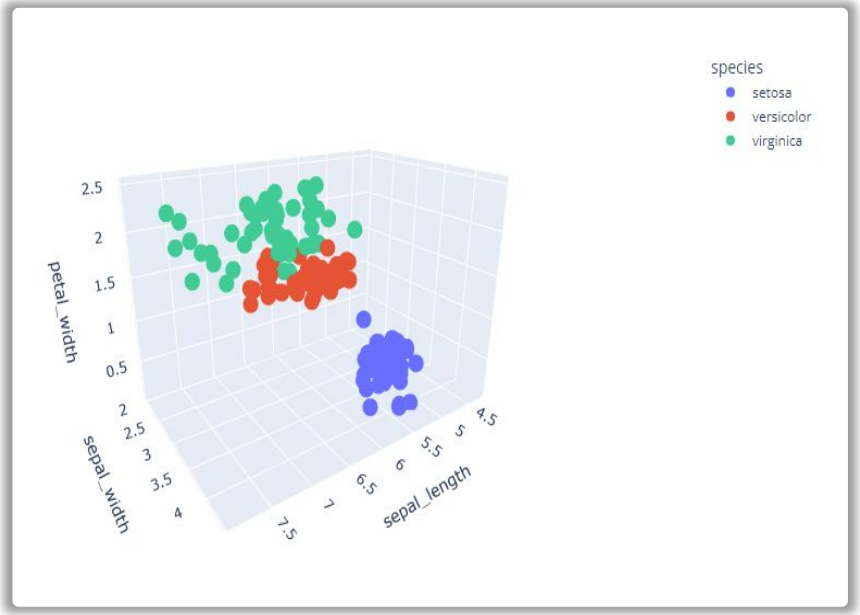
	Account	Name	Rep	Manager	Product	Quantity	Price	Status
0	714466	Trantow-Barrows	Craig Booker	Debra Henley	CPU	1	30000	presented
1	714466	Trantow-Barrows	Craig Booker	Debra Henley	Software	1	10000	presented
2	714466	Trantow-Barrows	Craig Booker	Debra Henley	Maintenance	2	5000	pending
3	737550	Fritsch, Russel and Anderson	Craig Booker	Debra Henley	CPU	1	35000	declined
4	146832	Kiehn-Spinka	Daniel Hilton	Debra Henley	CPU	2	65000	won

		sum				
		Price				
	Product	CPU	Maintenance	Monitor	Software	All
Manager	Status					
Debra Henley	declined	70000	0	0	0	70000
	pending	40000	10000	0	0	50000
	presented	30000	0	0	20000	50000
	won	65000	0	0	0	65000
Fred Anderson	declined	65000	0	0	0	65000
	pending	0	5000	0	0	5000
	presented	30000	0	5000	10000	45000
	won	165000	7000	0	0	172000
All		465000	22000	5000	30000	522000

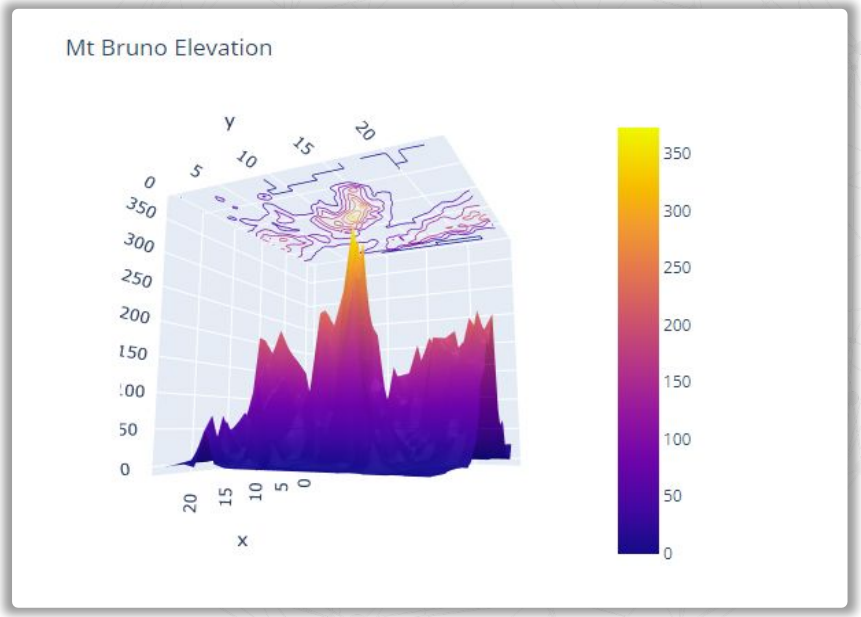
</talentlabs>

Three dimensional visualizations

Iris flower species with different characteristics



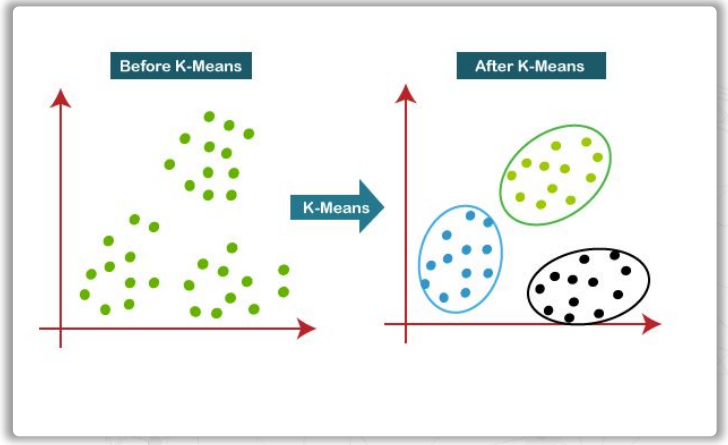
Surface plot of Mt Bruno's Elevation

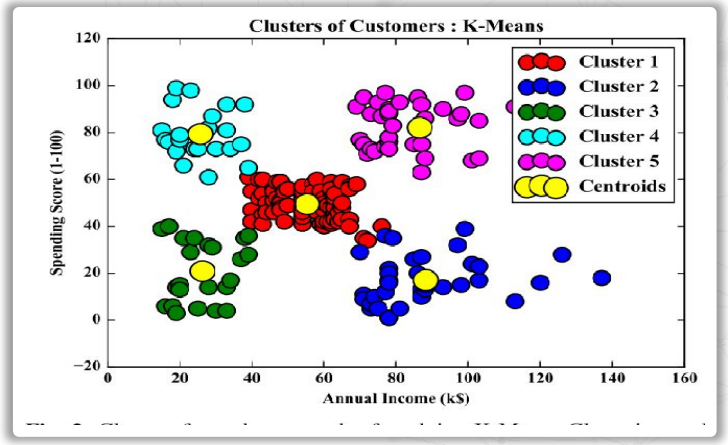


</talentlabs>

Cluster Analysis

- Find groups of similar samples/people.
- Homogenous within cluster and heterogenous between the clusters.
- Retail companies often use clustering to identify groups of households that are similar to each other.
- Streaming services often use clustering analysis to identify viewers who have similar behavior.





</talentlabs>


</talentlabs>
Thank you