
</talentlabs>
Bivariate Analysis

# Agenda

1. What is Bivariate analysis and why is it so important?
2. Recap - Covariance and Correlations
3. Quantitative vs Quantitative
4. Quantitative vs Qualitative
5. Qualitative vs Qualitative Analysis
7. Statistical analysis using t-tests and anova

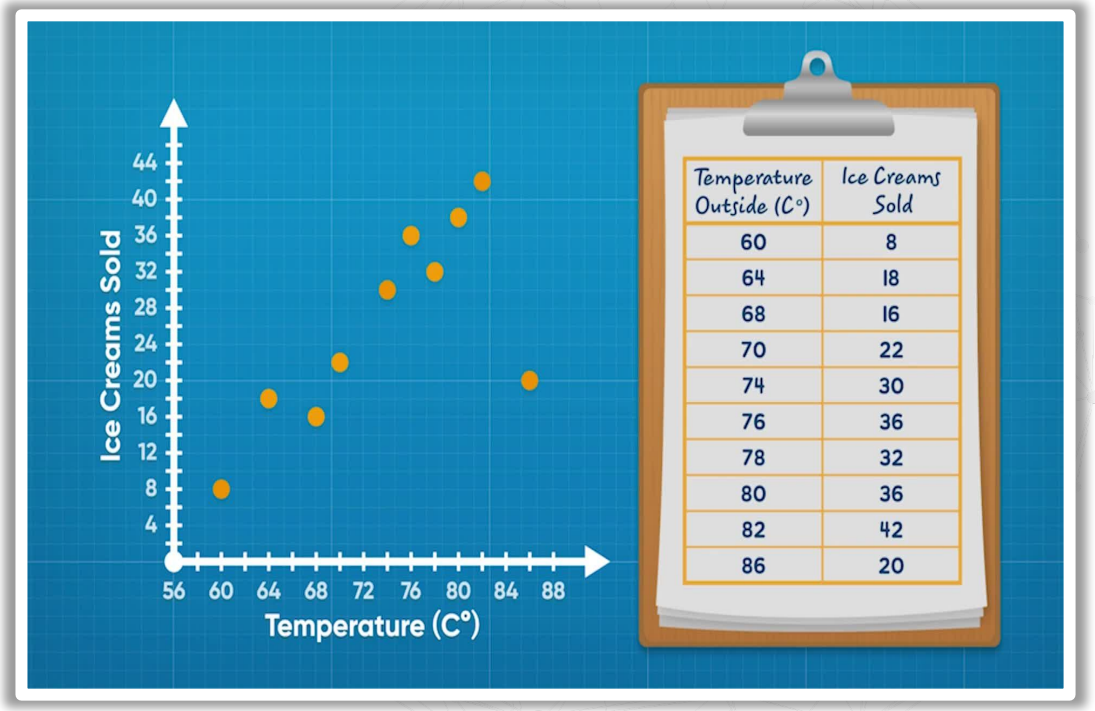





</talentlabs>

# What is Bivariate Analysis?

- It involves the analysis of two variables
- purpose of determining the empirical relationship between the two variables
- Want to test a hypothesis of association or relationship?
- The results from bivariate analysis can be stored in a two-column data table.



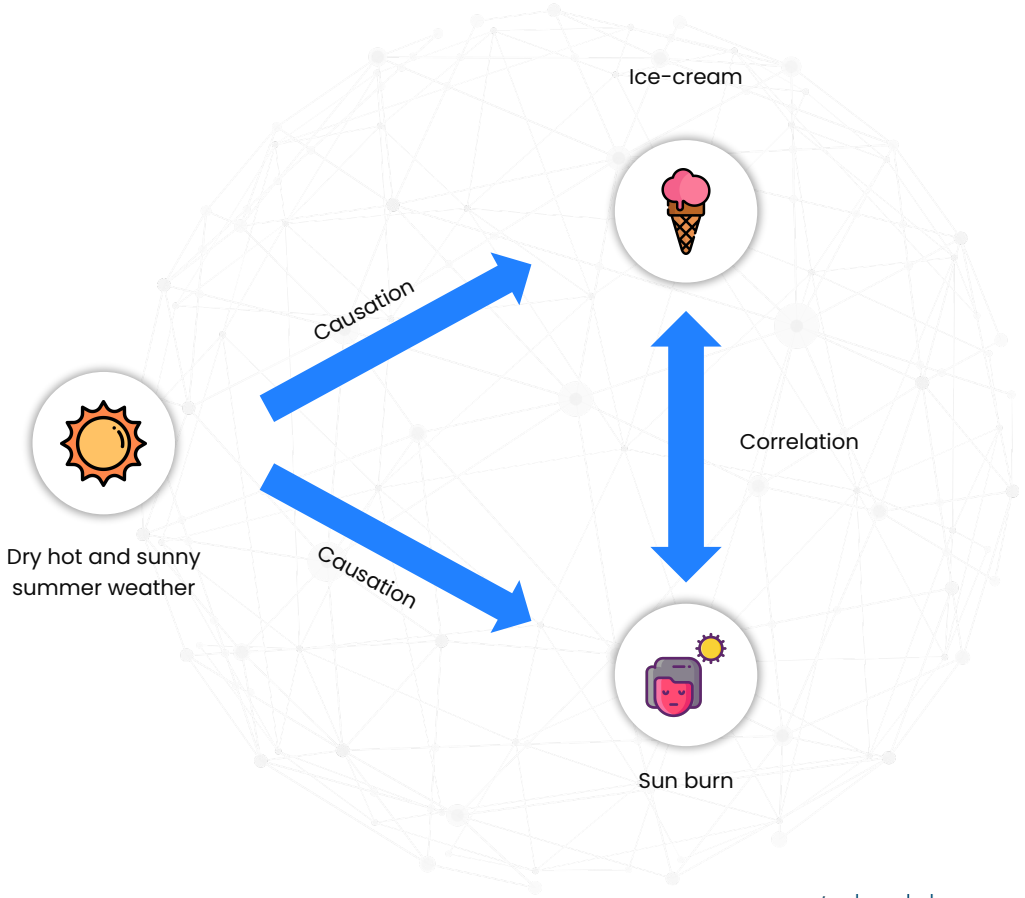





</talentlabs>

# Statistics to represent relationships

- Covariance is a statistic that measures the size of the relationship between two variables.
- Correlation is a measure that determines the strength of the relationship between two variables.
- A myth:  correlation = causation.



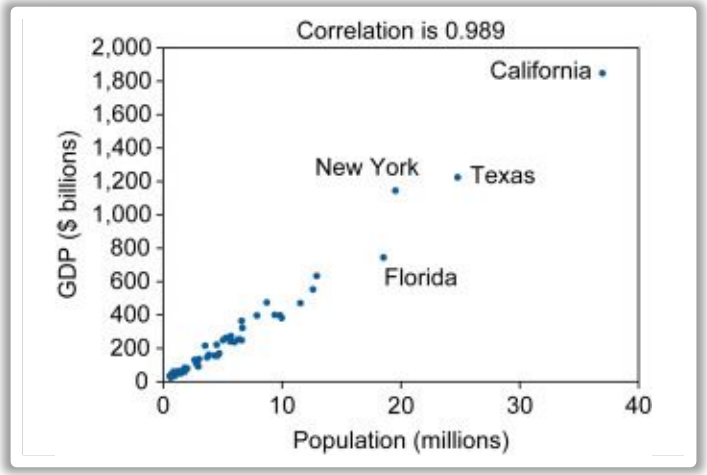


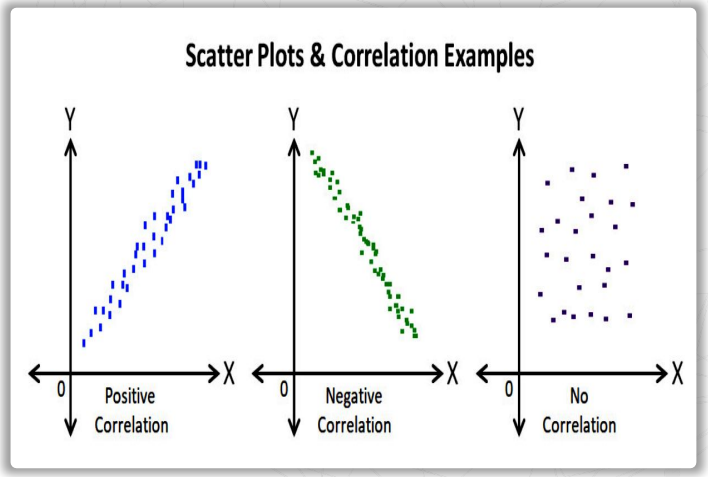



</talentlabs>

# The Scatterplot

- Relationship between numerical or continuous variables.
- **The graphic representation of the relationship between the two variables coming from a bivariate data set.**
- Think of them as the graphic representation of two data sets which have been put into place by dedicating each axis in **the plot to a different variable.**
- **Here, majorly we look into plotting data points without connectivity**



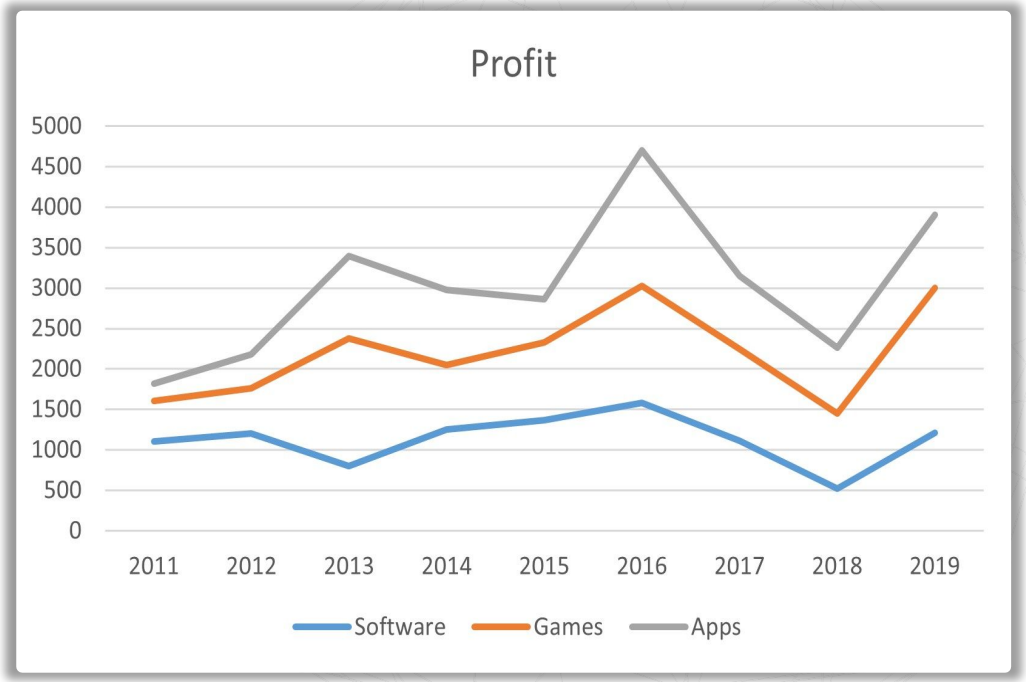





</talentlabs>

# The Line plot

- Relationship between numerical or continuous variables.
- **The graphic representation of the relationship between the two variables coming from a bivariate data set.**
- Here, **majorly we look into plotting data points with connectivity, but we can't see the points themselves.**
- More of trend line plots



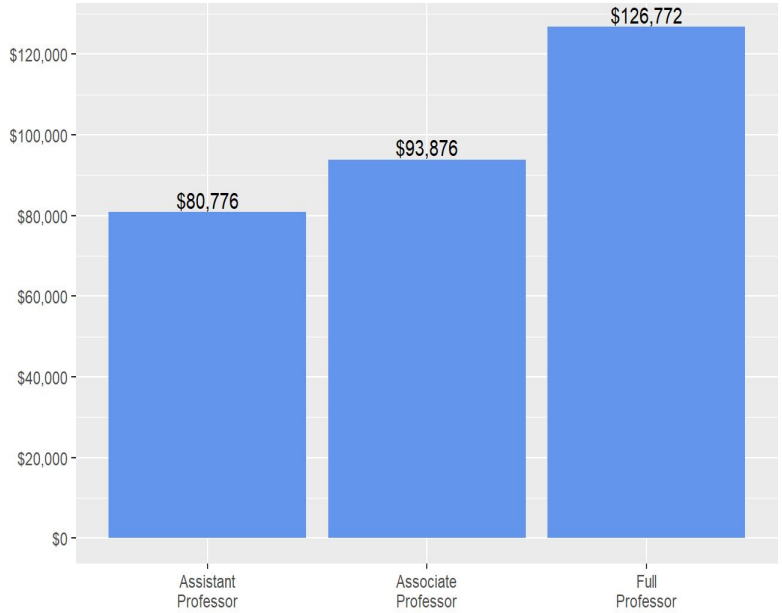




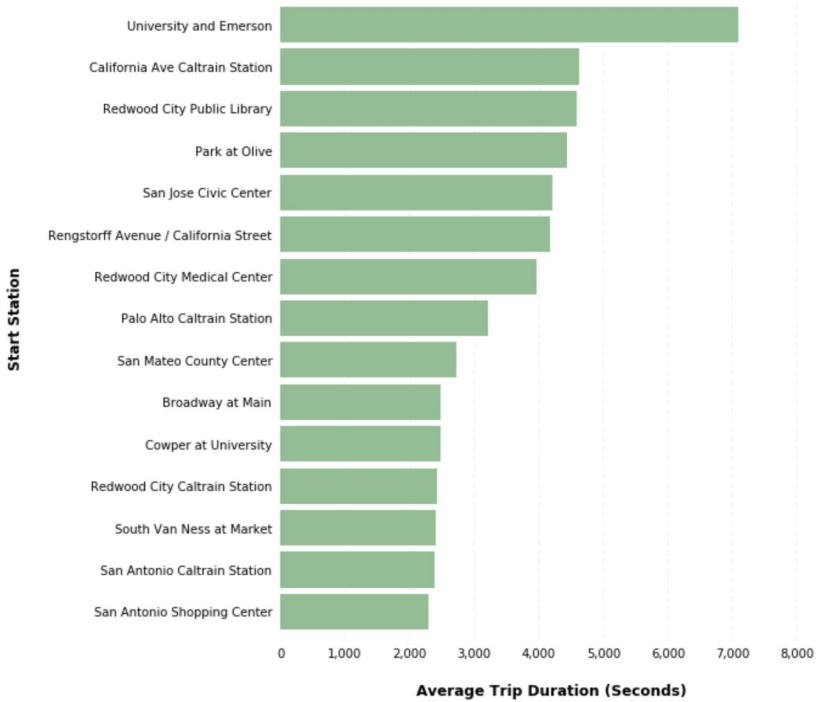

</talentlabs>

# Bar Charts

**9 Month average academic salary**
By Professor Level




**Trip duration**
By Station



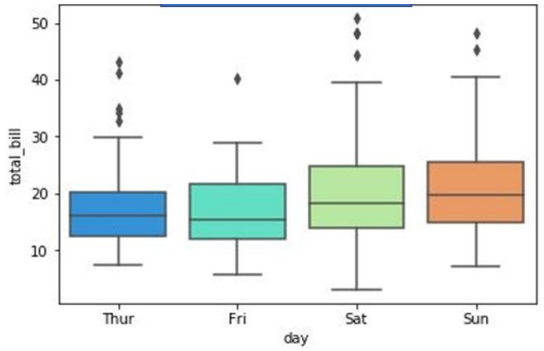




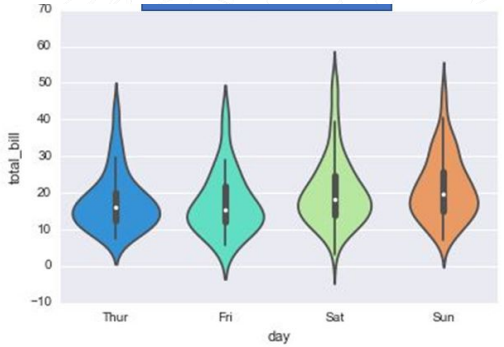

</talentlabs>

# Box plots

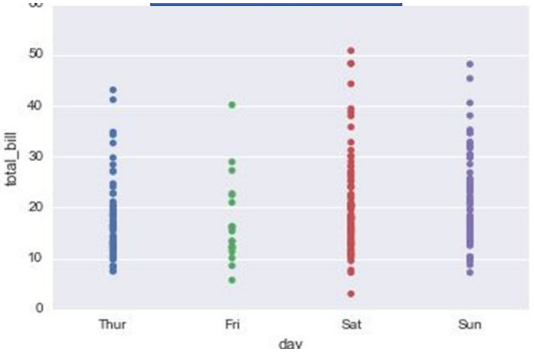
**Box plot**



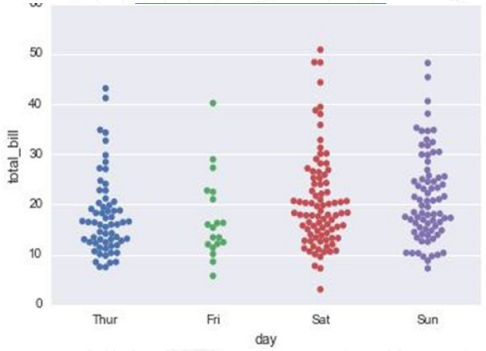

**Violin plot**




**Strip plot**




**Swarm plot**





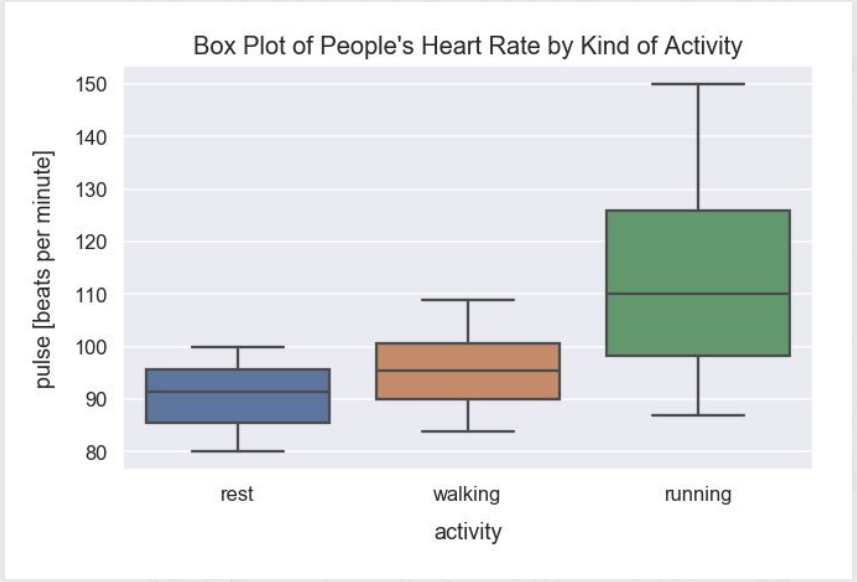



</talentlabs>

# Boxplot –Two variable example

- What? - I want to know how the distribution of heart rate differs for people resting, walking and running.
- Assumption that with more exercise activity, the median heart rate increases.

| Id | Diet | Pulse | Time | Kind |
|---|---|---|---|---|
| 5 | Low fat | 91 | 30 min | Rest |
| 21 | Low fat | 93 | 1 min | Running |
| 27 | No fat | 100 | 1 min | Running |
| 21 | Low fat | 110 | 30 min | Running |
| 4 | Low fat | 80 | 1 min | Rest |






</talentlabs>

# Cross Tabulations

- Both variables are categorical
- It is used to count between categories, or get summaries between two categories.

| | Cust | Categ_X | Categ_y |
|---|---|---|---|
| 0 | 1 | a | a |
| 1 | 1 | a | b |
| 2 | 1 | b | a |
| 3 | 1 | b | b |
| 4 | 2 | a | a |
| 5 | 3 | a | a |
| 6 | 3 | a | b |
| 7 | 3 | b | a |
| 8 | 3 | b | b |
| 9 | 4 | b | b |
| 10 | 5 | b | b |

| Categ_y | a | b | All |
|---|---|---|---|
| Categ_X | | | |
| a | 3 | 2 | 5 |
| b | 2 | 4 | 6 |
| All | 5 | 6 | 11 |

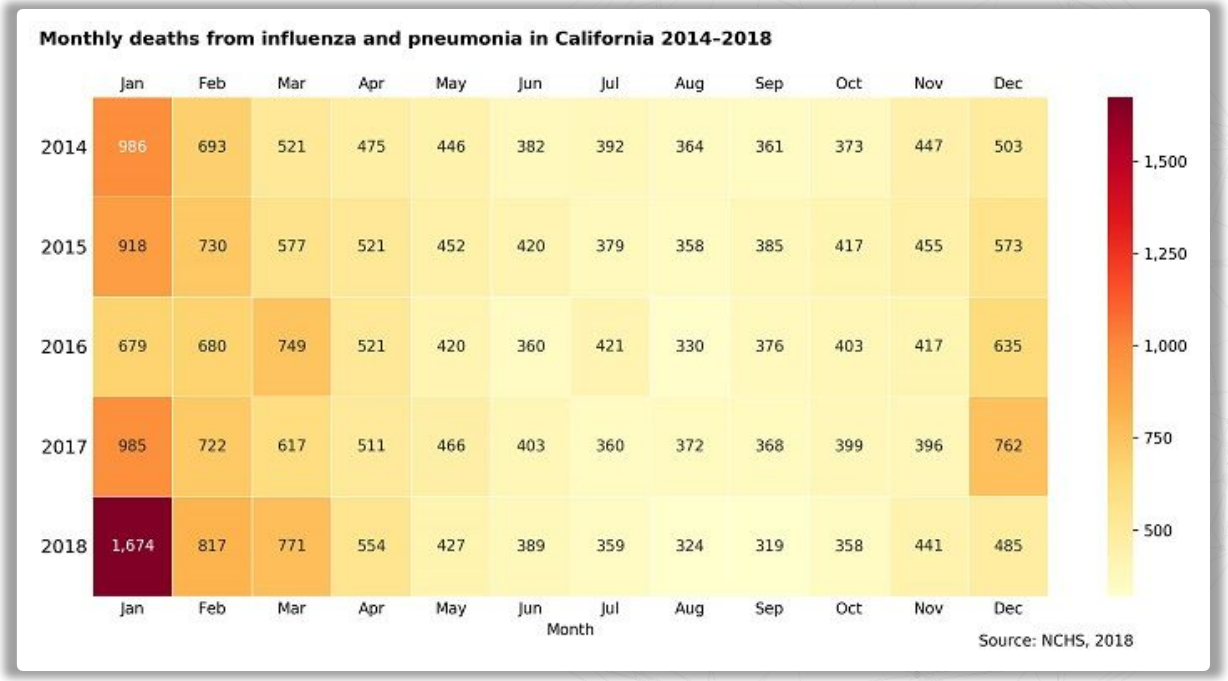




</talentlabs>

# Heat Maps in contingency tables

Contingency table with colors showing strength in counts/frequencies





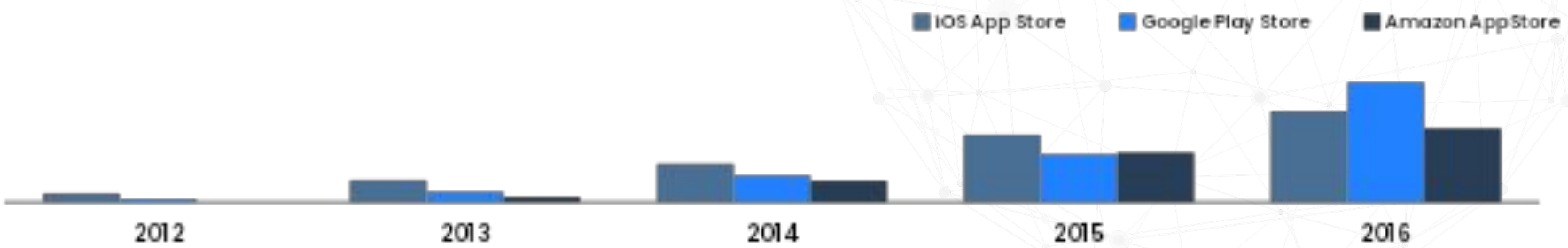



</talentlabs>

# Bar Charts with two variables

**Distribution of population in South Africa**
By Provinces


**App Publishing trend**
2012 – 2016





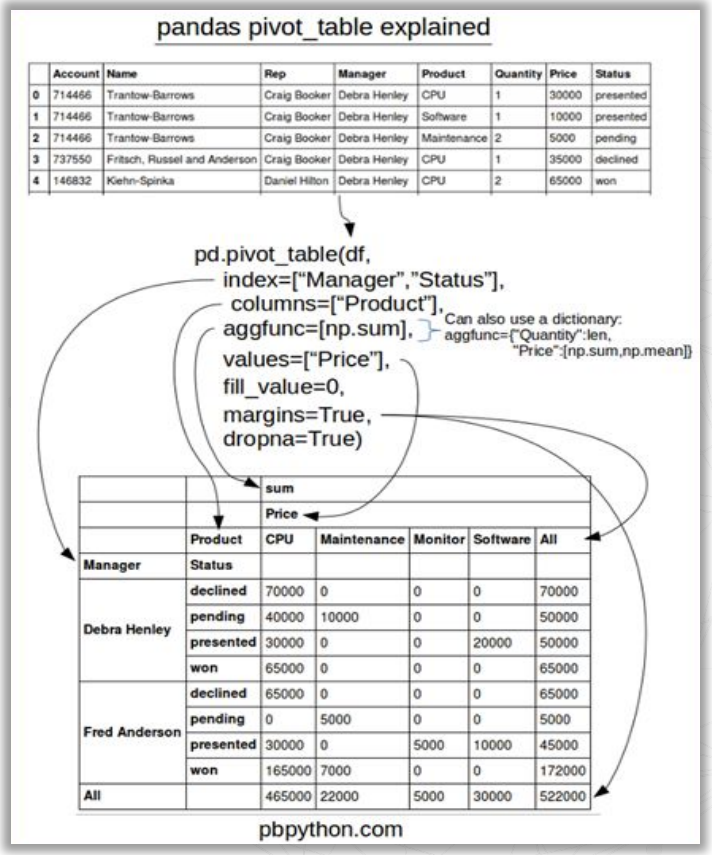



</talentlabs>


</talentlabs>
Thank you

# 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>