Day 3 Notes - Data Cleanup and Extraction

# Demo of the day

Word associations

<https://projector.tensorflow.org/>

Notebook : <https://github.com/elephantscale/cool-demos>

Word2vec 1 and 2

# Dealing with Missing Values

**Example 1: What will you replace these missing values with?**

| **Person id** | **weight** | **height** |
| --- | --- | --- |
| 1 | 50 | 150 |
| 2 | 60 | 160 |
| 3 | ??? = average(mean) / median | 170 |
| 4 | 75 | ??? |

**Example 2 : how will we deal with these missing values**

| **city** | **year** | **rainfall** |
| --- | --- | --- |
| San Jose | 2020 | 10 |
| Seattle | 2020 | 50 |
| San Jose | 2019 | 14 |
| Seattle | 2019 | 70 |
| San Jose | 2018 | ???  mean/median  Distribution  Mean or median of SJ  0 |
| Seattle | 2018 | ??? |

**Example 3 : How will you deal with this one**

| **city** | **year** | **rainfall** |
| --- | --- | --- |
| San Jose | 2020 | 10 |
| ??? | 2020 | 50 |
| San Jose | 2019 | 14 |
| Seattle | 2019 | 70 |
| ??? | 2018 | 12 |
| Seattle | 2018 | 60 |

Options

* Drop the data (you will lose some data)
* Substitute
  + 0
  + Mean / median

# Data Cleanup

Start with this lab first

Repo : <https://github.com/elephantscale/python-data-science-workshop>

Notebook : data-cleanup-1

And then do lab-5

**Deep copy**



# Lab 5 - data cleanup

To display numbers in human readable format

pd.options.display.float\_format = '{:,.2f}'.format

**Plot gallery**

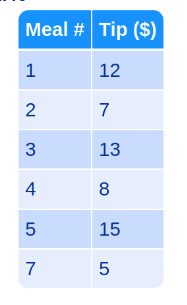
<https://matplotlib.org/stable/gallery/index.html>

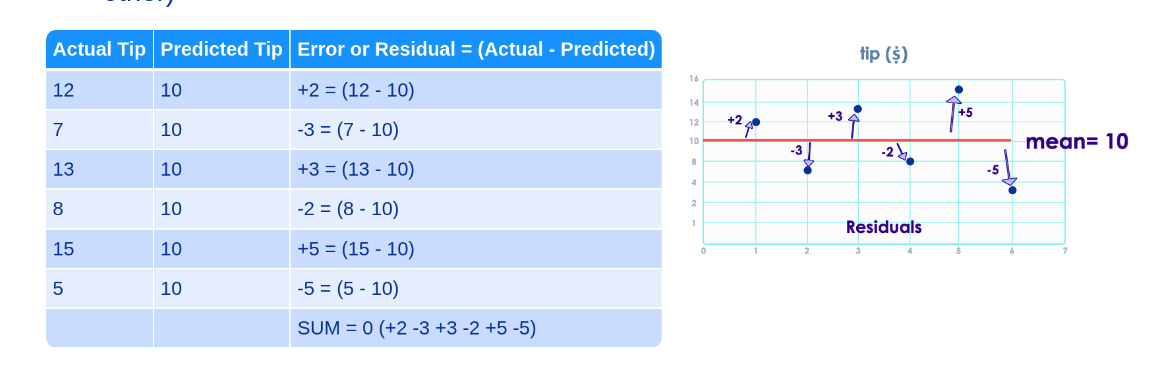
<https://chart-studio.plotly.com/~>

# Data Exploration

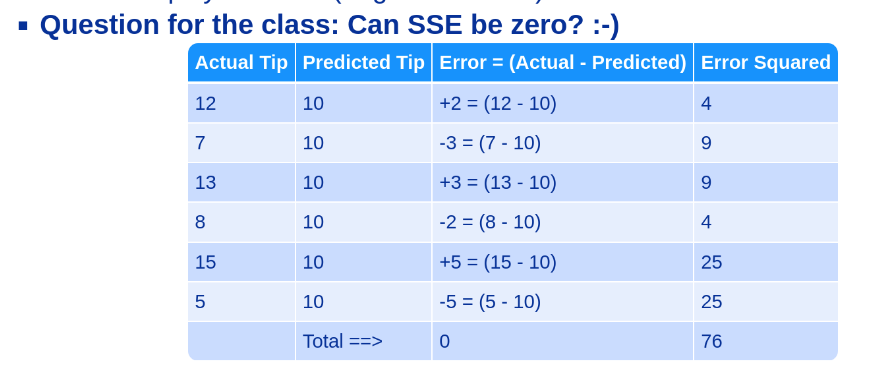
Start with this notebook : <https://github.com/elephantscale/python-data-science-workshop/blob/main/exploration/explore-house-sales.ipynb>

# Errors and Residuals





So to avoid positive negative cancellation, square the error



**heteroskedasticity**

<https://www.investopedia.com/terms/h/heteroskedasticity.asp>

**Homoscedasticity**

<https://www.statisticshowto.com/homoscedasticity/>

**Cook's distance**

<https://www.statisticshowto.com/cooks-distance/>

<https://www.youtube.com/watch?v=xc_X9GFVuVU> - watch first 4 mins

**Understanding Q-Q plots**

<https://data.library.virginia.edu/understanding-q-q-plots/>

<https://www.youtube.com/watch?v=okjYjClSjOg> - with some music :) - Watch first 3 mins

<https://www.youtube.com/watch?v=X9_ISJ0YpGw>