DATA 180 Recap

Amanda Tran & Chloe Ho

Topics to be covered:

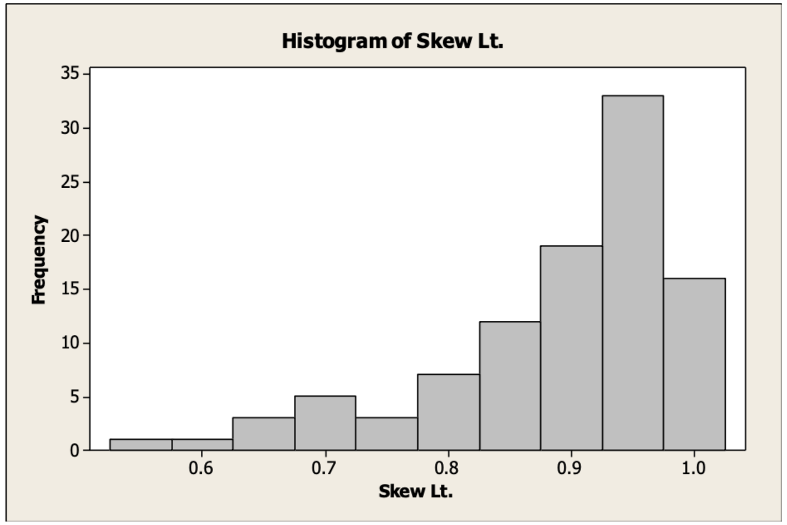
VISUALIZATIONS

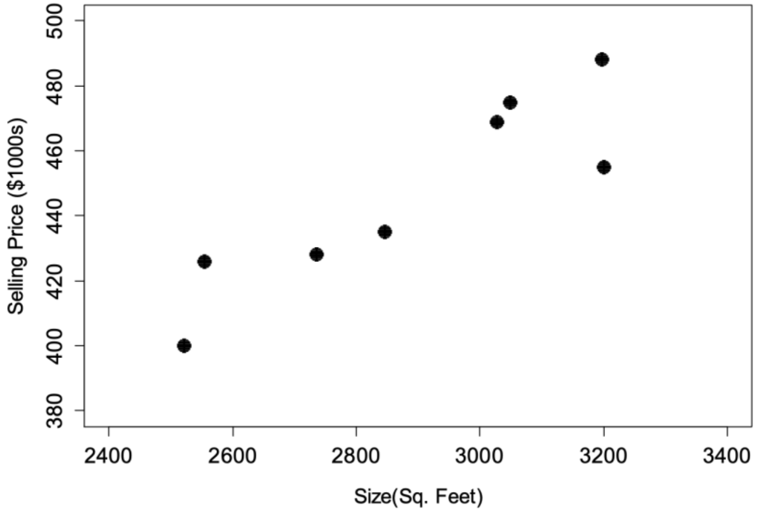
DATA
WRANGLING

UNSUPERVISED
LEARNING

SUPERVISED
LEARNING

1. Visualizations – Numerical Variables





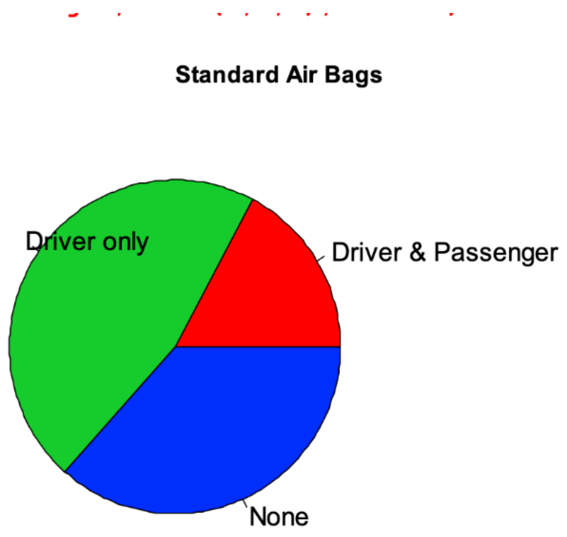
Single Numerical Variable: Histogram

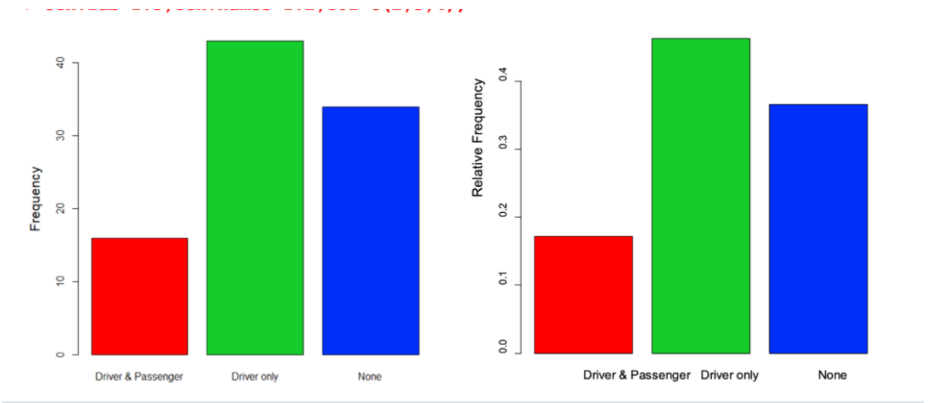
- Mean, median, mode
- Overall shape of the data set (e.g., symmetric or skewed)
- Presence of (1) gaps in the data set, (2) outliers

Paired data visualization: Scatterplot

1. Visualizations – Categorical Variables

- **Key word: frequency/relative frequency**
- **Visualization examples:** Bar plots & pie charts
- Alternatives… for when there are too many categories? (i.e., balloon plots, mosaic plots)





2. Data Wrangling

- **Key Data Wrangling Techniques:**
  - Data Importing: Using functions like read.csv(), read.table()
  - Data Cleaning: Handling missing values, outliers, and incorrect data types
  - Data Transformation: Reshaping data, combining datasets, creating new variables
  - Data Exporting: Saving the cleaned and transformed data for further analysis

- **Why wrangle data?**

- **Data Wrangling Trivia**

4. Supervised Learning

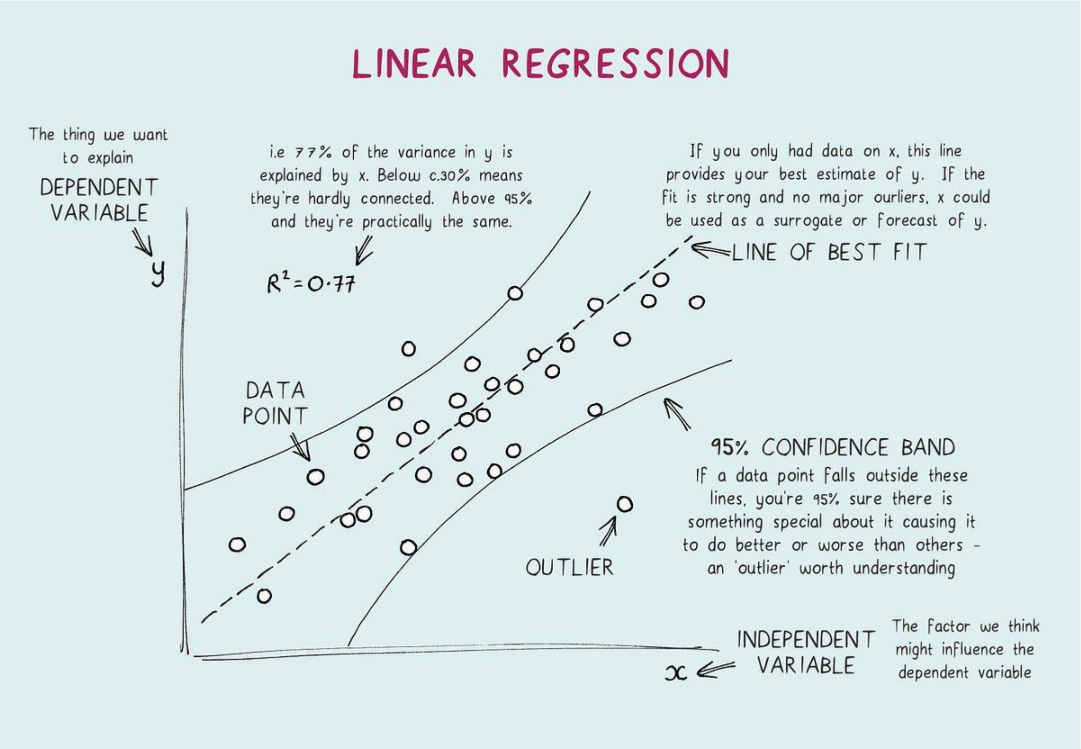
✓ Observations are classified into ***predictor*** and **response** variables

Primary goal: modelling the relationship between a set of predictors and a response variable.

Covered material:

- Linear regression
- Logistic regression

Simple Linear Regression


LINEAR REGRESSION
The thing we want to explain
DEPENDENT VARIABLE
y
i.e 77% of the variance in y is explained by x. Below c.30% means they're hardly connected. Above 95% and they're practically the same.
$R^2 = 0{\cdot}77$
If you only had data on x, this line provides your best estimate of y. If the fit is strong and no major ourliers, x could be used as a surrogate or forecast of y.
LINE OF BEST FIT
DATA POINT
95% CONFIDENCE BAND
If a data point falls outside these lines, you're 95% sure there is something special about it causing it to do better or worse than others - an 'outlier' worth understanding
OUTLIER
INDEPENDENT VARIABLE
The factor we think might influence the dependent variable
x

- SLR is a way for predicting a response Y on the basis of a single predictor variable X.
  - Y could be numeric, binary/categorical (coded as integers)
- Easily expandable to Multiple Linear Regression where there exist multiple predictors Xi.
- Eg. $\hat{y} = b_0 + b_1x$

Logistic Regression

- Logistic regression estimates the probability of an event occurring, such as "voted" or "didn’t vote", based on a given dataset of independent variables.
- Since the outcome is a probability, the dependent variable is bounded between 0 and 1.
- *What's the fundamental difference between logistic regression and linear regression?*

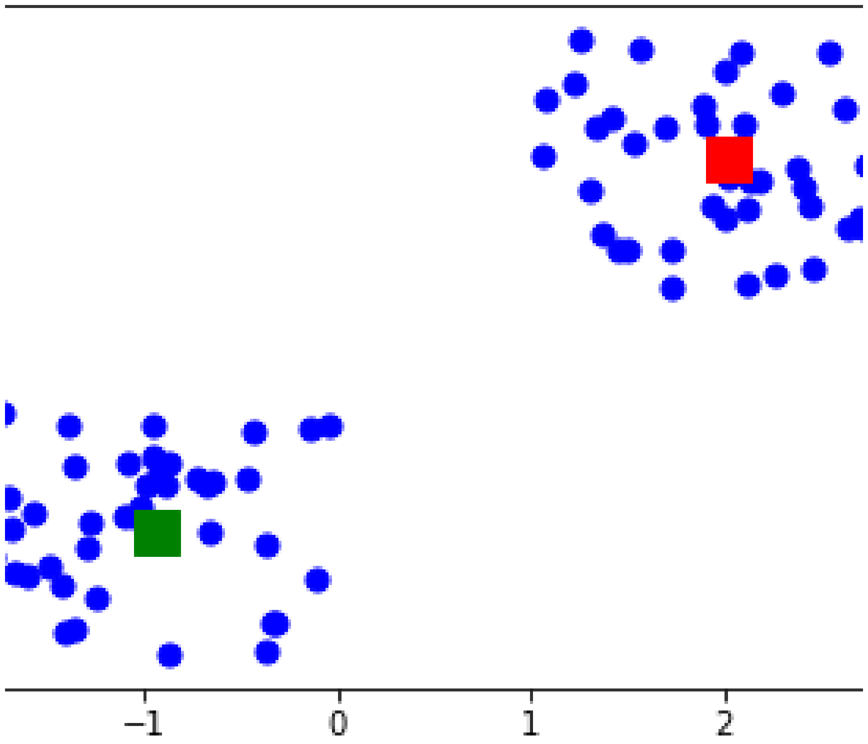
3. Unsupervised Learning

For every observation x_i that we observe, we do not observe a response variable y_i.

Goal: finding patterns in the data set.

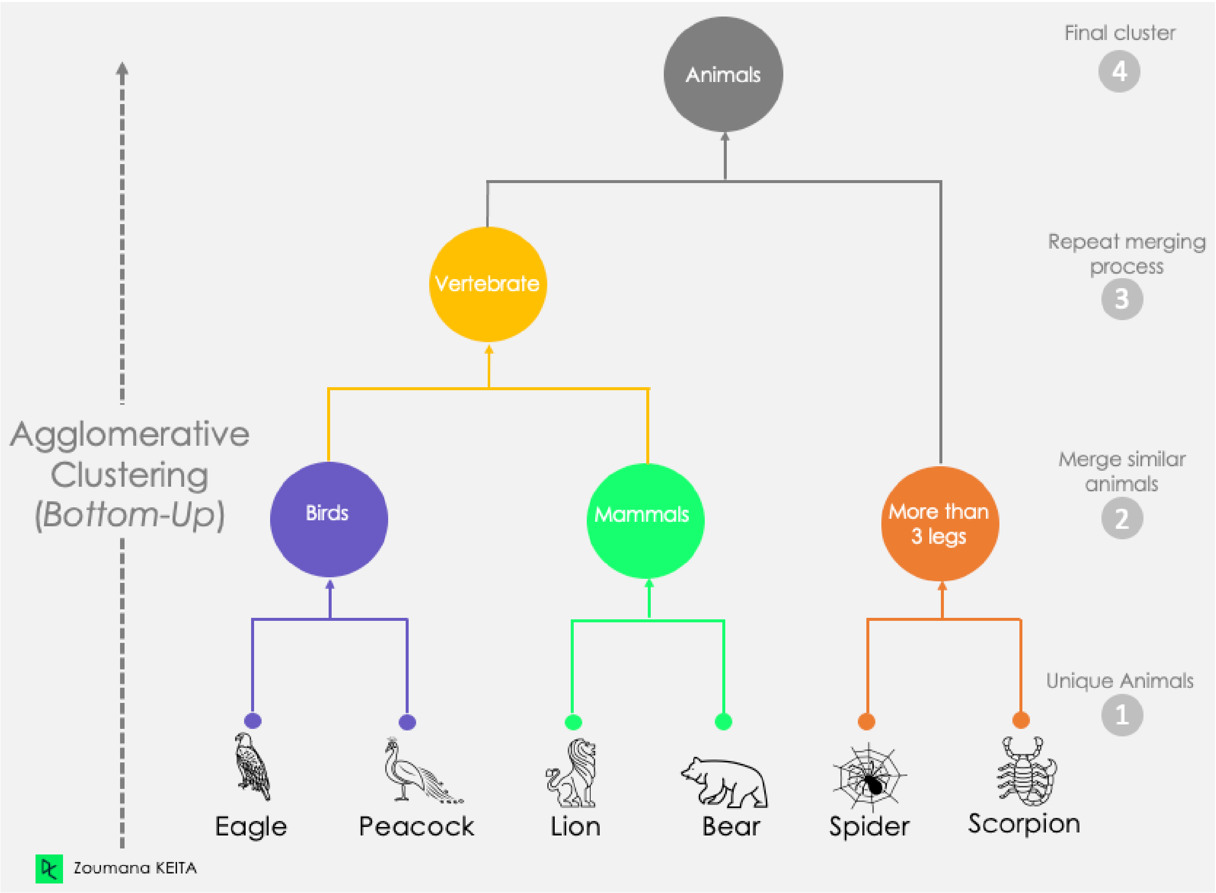
Covered technique:

- K-means clustering
- Hierarchical clustering



K-means clustering

- You’ll define a target number k, which refers to the number of centroids you need in the dataset. A centroid is the imaginary or real location representing the center of the cluster.
- In other words, the K-means algorithm identifies *k* number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible.



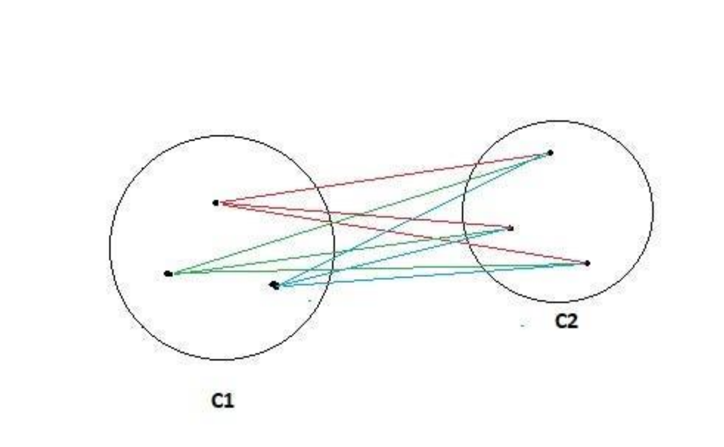
Hierarchical clustering

The basic algorithm of Agglomerative is straight forward:

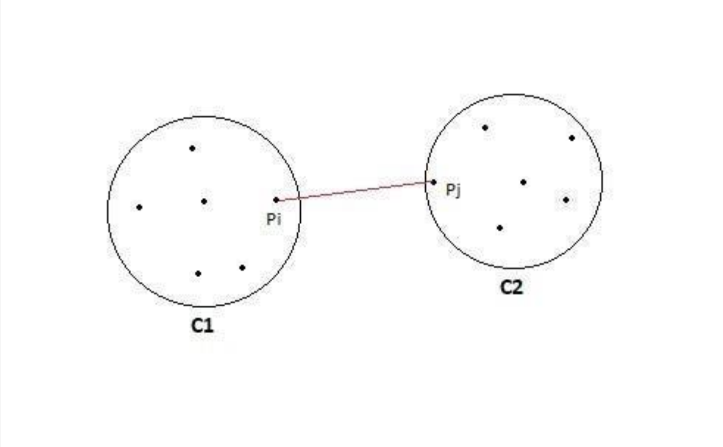
- Compute the proximity matrix
- Let each data point be a cluster
- Repeat: Merge the two closest clusters and update the proximity matrix
- Until only a single cluster remains

Hierarchical clustering

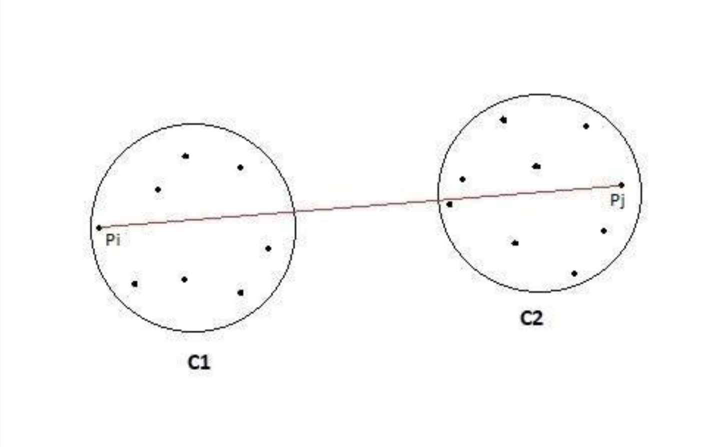
How do we calculate the similarities between two clusters?



MIN (Single Linkage)



MAX (Complete Linkage)



GROUP AVERAGE (Centroid Linkage)