Data Modeling and Visualisation

CCT490H5F - Social Data Analytics

Professor Alex Hanna

October 20, 2016

FiveThirtyEight

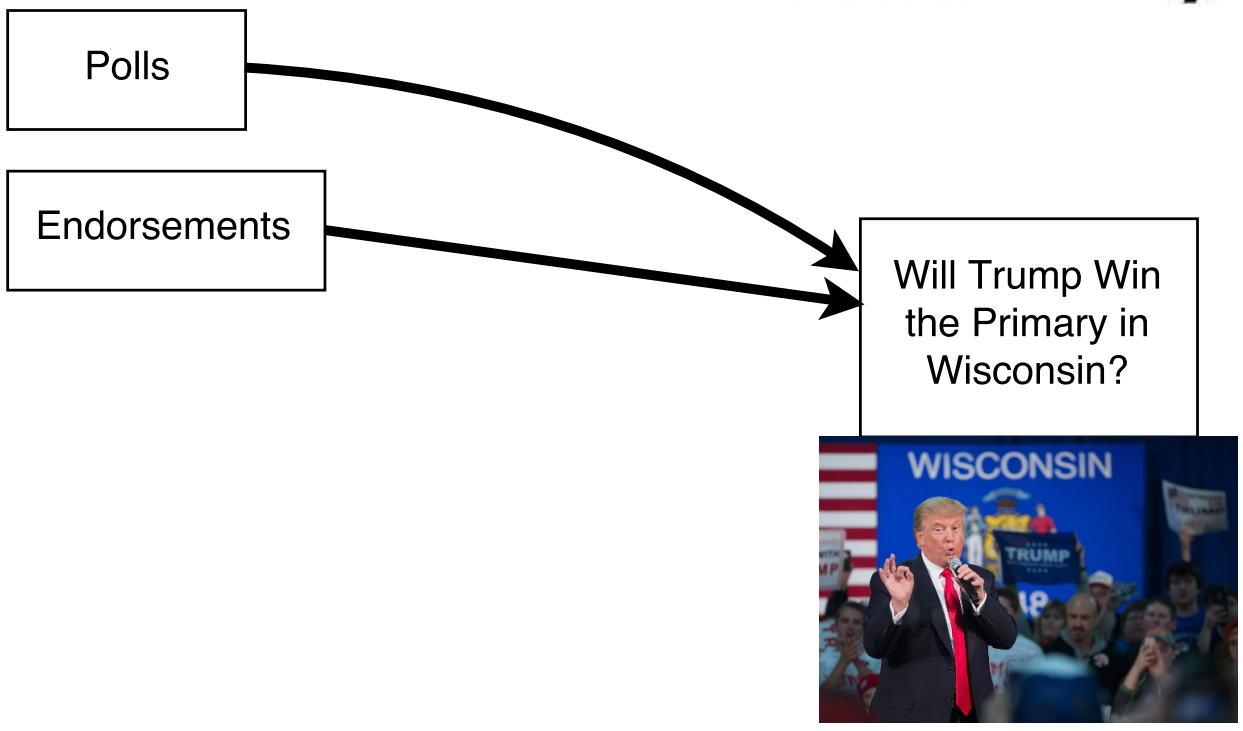
Will Trump Win the Primary in Wisconsin?

FiveThirtyEight

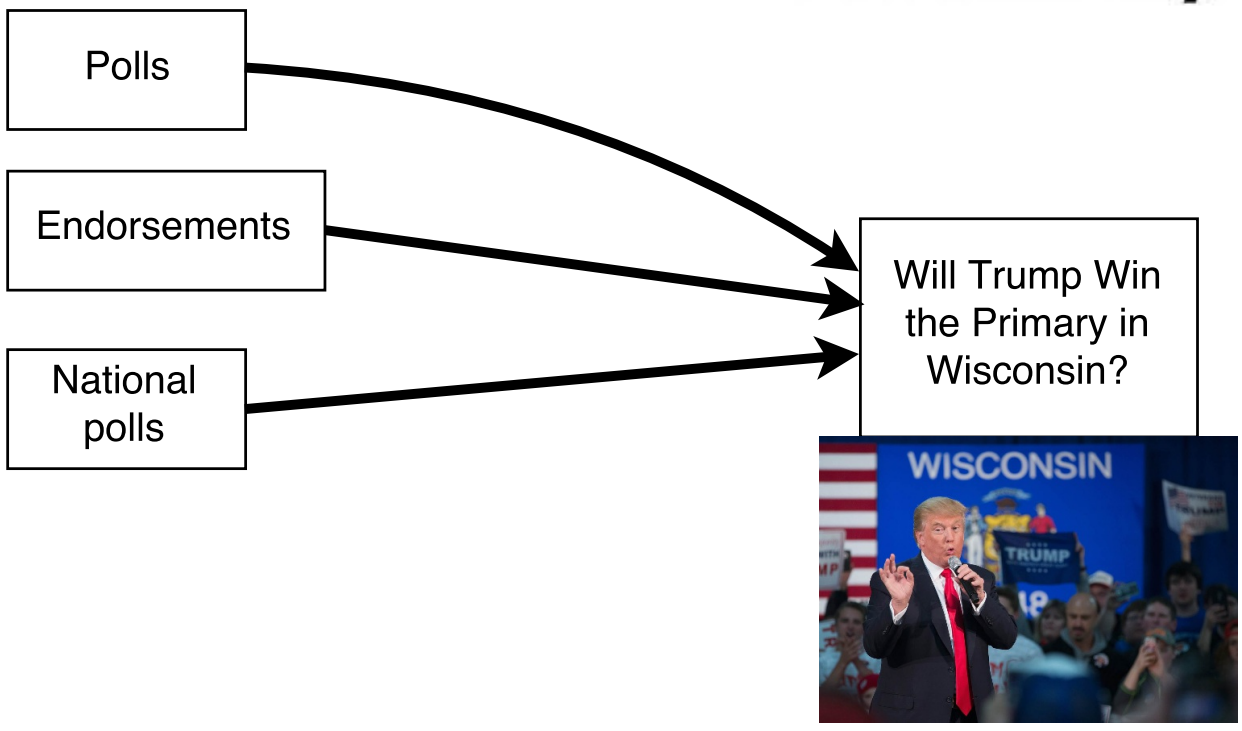
Polls

Will Trump Win the Primary in Wisconsin?

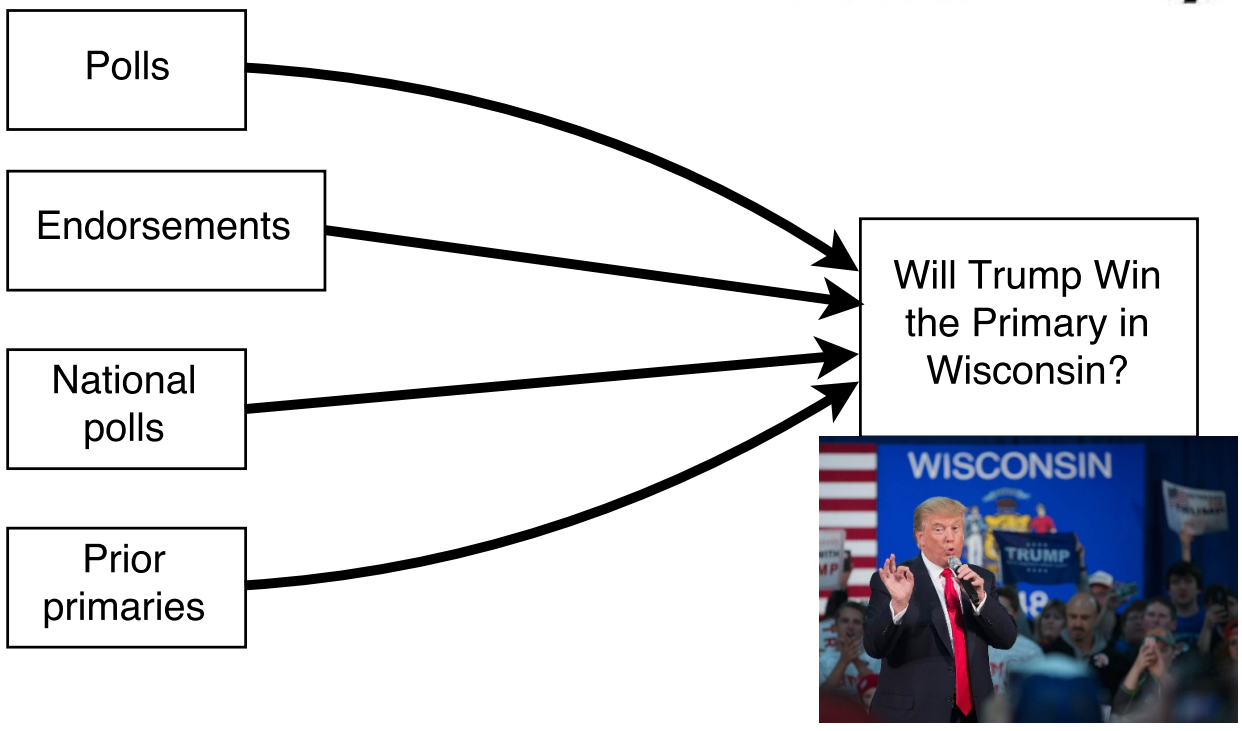
FiveThirtyEight



FiveThirtyEight



FiveThirtyEight



Data Modeling

A *model* is a simplified representation of reality.

Data Modeling

A *model* is a simplified representation of reality.

A *mathematical model* is a representation of reality using numbers.

Why have a model?

- *Description*: You want to summarise data.
- *Explanation*: You want replicate the working of the world with existing data.
- *Prediction*: You want to forecast the future from past data.

Why have a model?

- *Description*: You want to summarise data.
- *Explanation*: You want replicate the working of the world with existing data.
- *Prediction*: You want to forecast the future from past data.

Why have a model?

- *Description*: You want to summarise data.
- *Explanation*: You want replicate the working of the world with existing data.
- *Prediction*: You want to forecast the future from past data.

Why have a model?

- *Description*: You want to summarise data.
- *Explanation*: You want replicate the working of the world with existing data.
- *Prediction*: You want to forecast the future from past data.

Modeling is a *data reduction* process.

Why have a model?

- *Description*: You want to summarise data.
- *Explanation*: You want replicate the working of the world with existing data.
- *Prediction*: You want to forecast the future from past data.

Modeling is a *data reduction* process.

“All models are wrong but some are useful.” - George Box, statistician

Example: More Tweets, More Votes

DiGrazia et al. 2013. “More Tweets, More Votes: Social Media as a Quantitative Indicator of Political Behavior”

The more times a 2010 US House candidate was mentioned, the more likely it is they will be elected.

Other things which affect votes

- Incumbency
- Ideological leaning
- Age
- Education
- Gender
- Race
- Media markets

Major variables in MTMV

Critical variables for this analysis

- *Dependent variable*: Republican percent of the vote share (`vote_share`)
- *Independent variables*
  - Republican percent of Twitter mention share (`mshare`)
  - Republican incumbency (`rep_inc`)

Major variables in MTMV

Critical variables for this analysis

- *Dependent variable*: Republican percent of the vote share (`vote_share`)
- *Independent variables*
  - Republican percent of Twitter mention share (`mshare`)
  - Republican incumbency (`rep_inc`)

Major variables in MTMV

Critical variables for this analysis

- *Dependent variable*: Republican percent of the vote share (`vote_share`)
- *Independent variables*
  - Republican percent of Twitter mention share (`mshare`)
  - Republican incumbency (`rep_inc`)

Major variables in MTMV

Critical variables for this analysis

- *Dependent variable*: Republican percent of the vote share (`vote_share`)
- *Independent variables*
  - Republican percent of Twitter mention share (`mshare`)
  - Republican incumbency (`rep_inc`)

Description of MTMV Data: Crosstabs

In [71]:
```
import pandas as pd
import numpy as np
df_mtmv = pd.read_csv("data/mtmv_data_10_12.csv", index_col = 0)
```

In [59]:
```
## vote share and mention share mean
## by Republican incumbency
gr_mtmv = df_mtmv.groupby('rep_inc')
gr_mtmv[['vote_share', 'mshare']].mean()
```

Out[59]:

	vote_share	mshare
rep_inc		
0	42.119264	39.360672
1	67.516408	72.170127

In [60]:
```
## vote share and mention share standard deviation
## by Republican incumbency
gr_mtmv[['vote_share', 'mshare']].std()
```

Out[60]:

	vote_share	mshare
rep_inc		
0	13.850496	27.930878
1	7.005758	28.911278

Explanation of MTMV Data: Correlation

```
In [8]: from scipy.stats.stats import pearsonr

        print(pearsonr(df_mtmv['mshare'], df_mtmv['vote_share'])[0])

        0.508867322507
```

Pearson correlation: measure of the linear dependence between two variables X and Y. Ranges from [-1, 1].

Explanation of MTMV Data: Linear Regression

```
In [9]: from statsmodels.formula.api import ols

model = ols("vote_share ~ rep_inc + mshare + pct_white + \
             pct_college + med_hhinc + pct_female", df_mtmv).fit()
model.summary()
```

	coef	std err	t	P>|t|	[95.0% Conf. Int.]
Intercept	8.2099	24.490	0.335	0.738	-39.936 56.356
rep_inc	18.3989	1.008	18.257	0.000	16.418 20.380
mshare	0.0543	0.015	3.639	0.000	0.025 0.084
pct_white	0.4735	0.026	18.028	0.000	0.422 0.525
pct_college	-0.3384	0.073	-4.619	0.000	-0.482 -0.194
med_hhinc	0.1132	0.051	2.211	0.028	0.013 0.214
pct_female	0.0418	0.464	0.090	0.928	-0.870 0.953

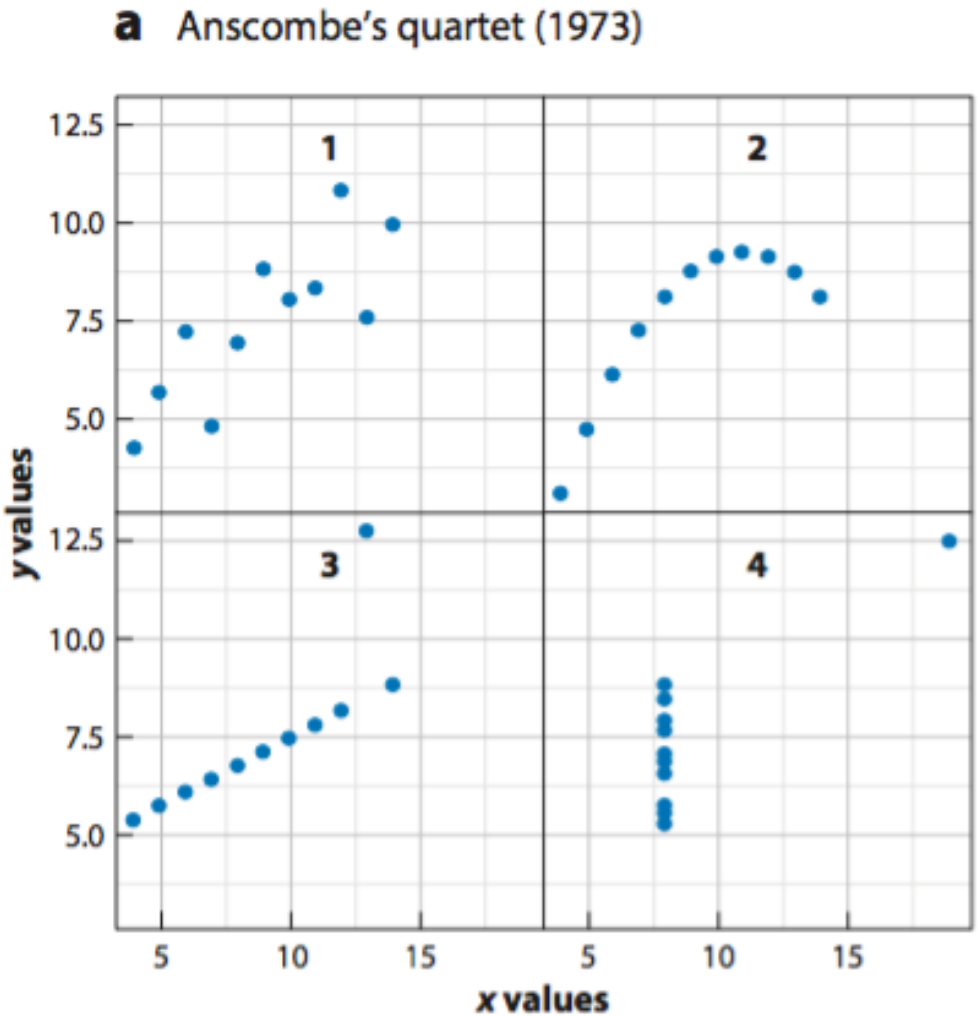
Regression: statistical technique which models the relationship between multiple variables.

Exercise: Build your own model!

Visualisation

Purposes of visualisation

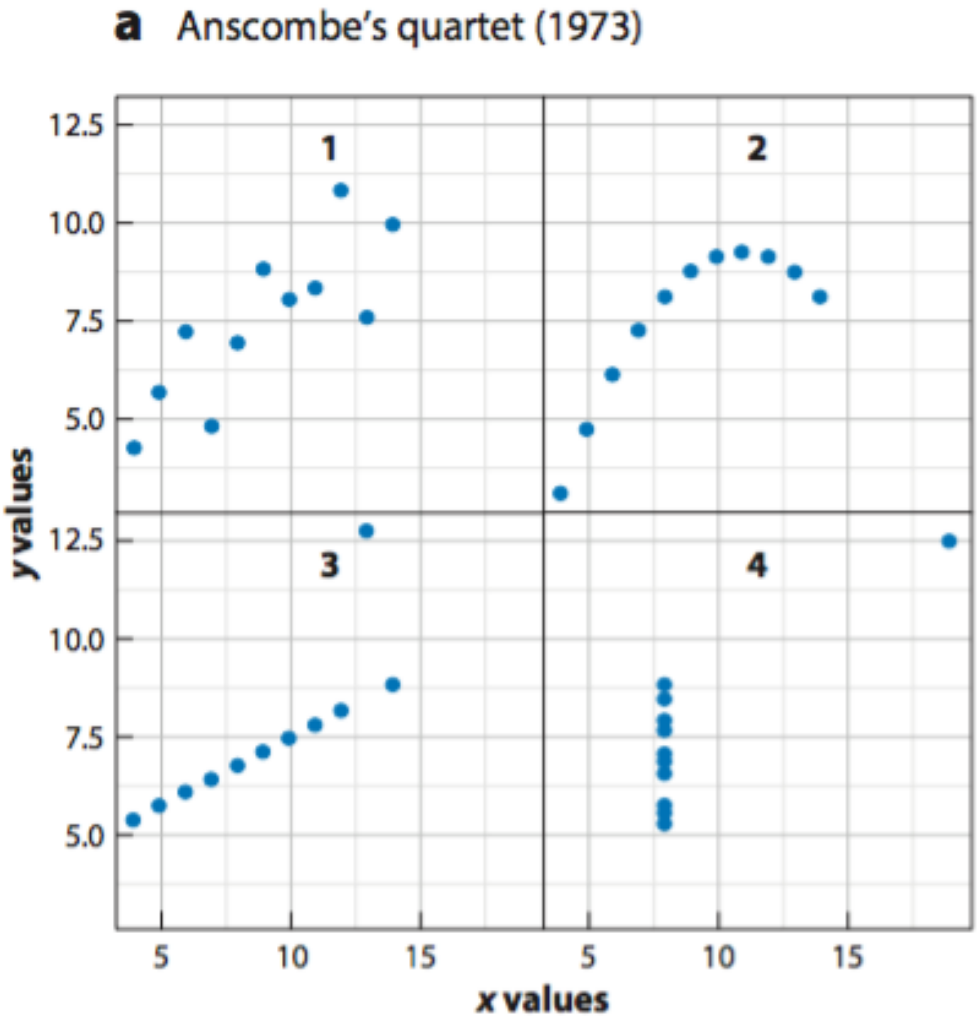
- Exploring data
- Confirming model
- Presenting results



Visualisation

Purposes of visualisation

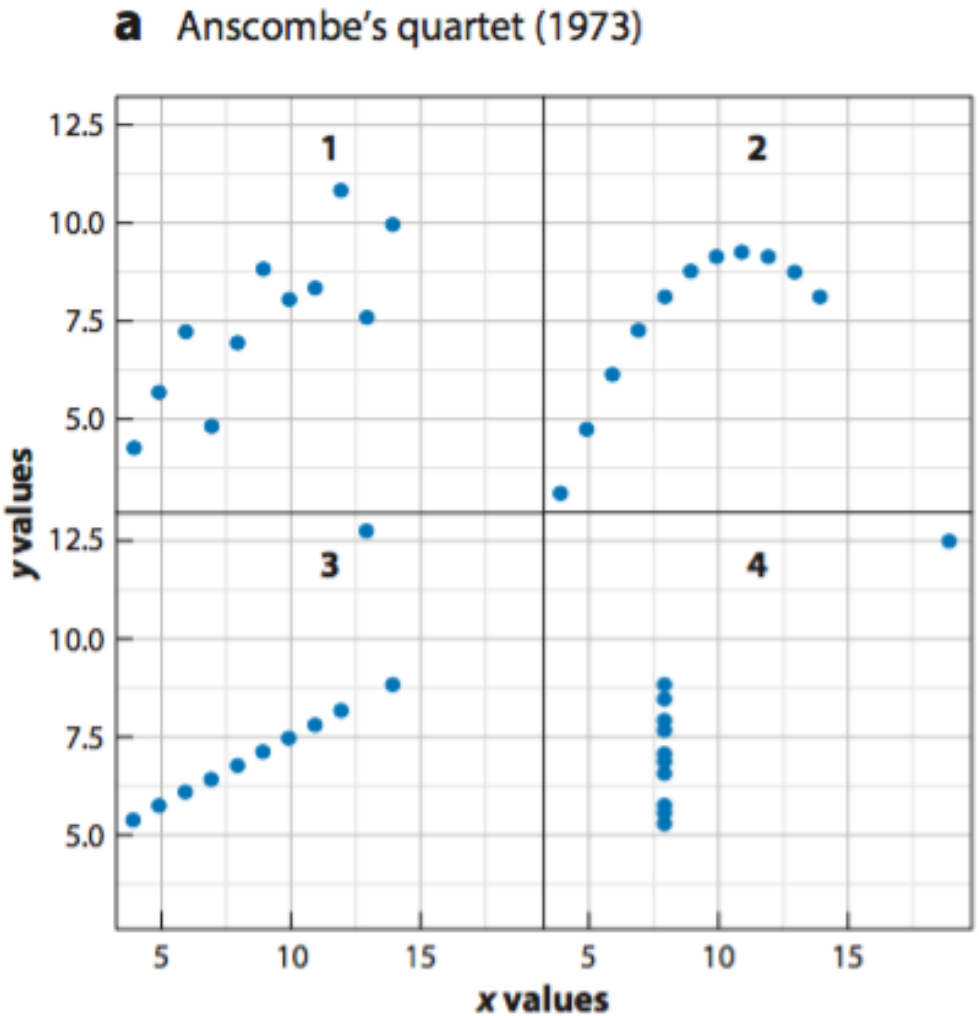
- Exploring data
- Confirming model
- Presenting results



Visualisation

Purposes of visualisation

- Exploring data
- Confirming model
- Presenting results



Univariate visualisations

- Understand the variable beyond mean, median, standard deviation, etc.
- Should be first part of exploring data
- Types
  - Histogram
  - Density

Univariate visualisations

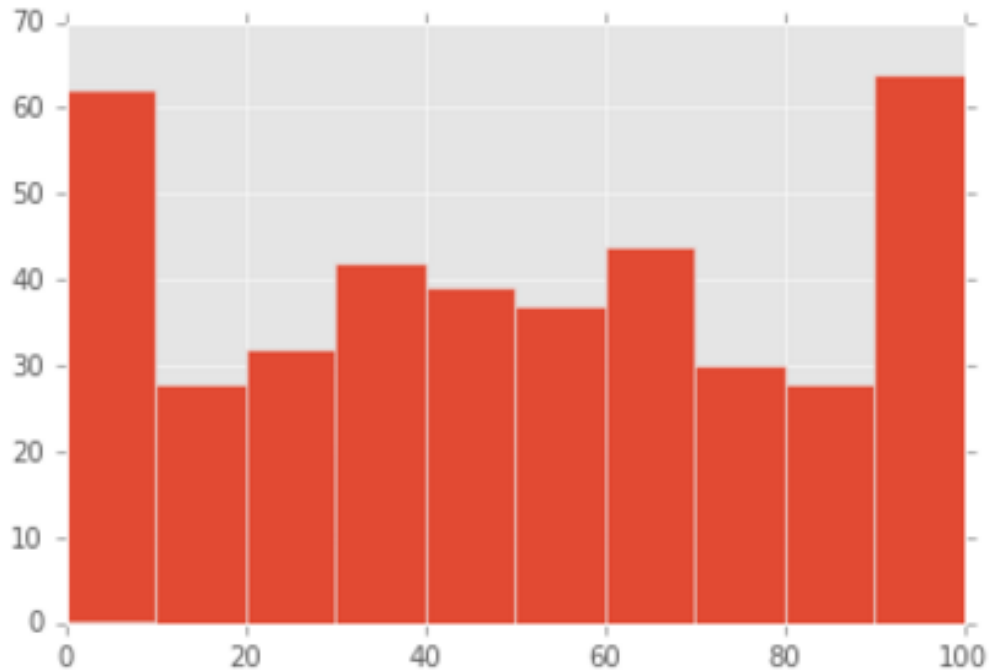
- Understand the variable beyond mean, median, standard deviation, etc.
- Should be first part of exploring data
- Types
  - Histogram
  - Density

Univariate visualisations

- Understand the variable beyond mean, median, standard deviation, etc.
- Should be first part of exploring data
- Types
  - Histogram
  - Density

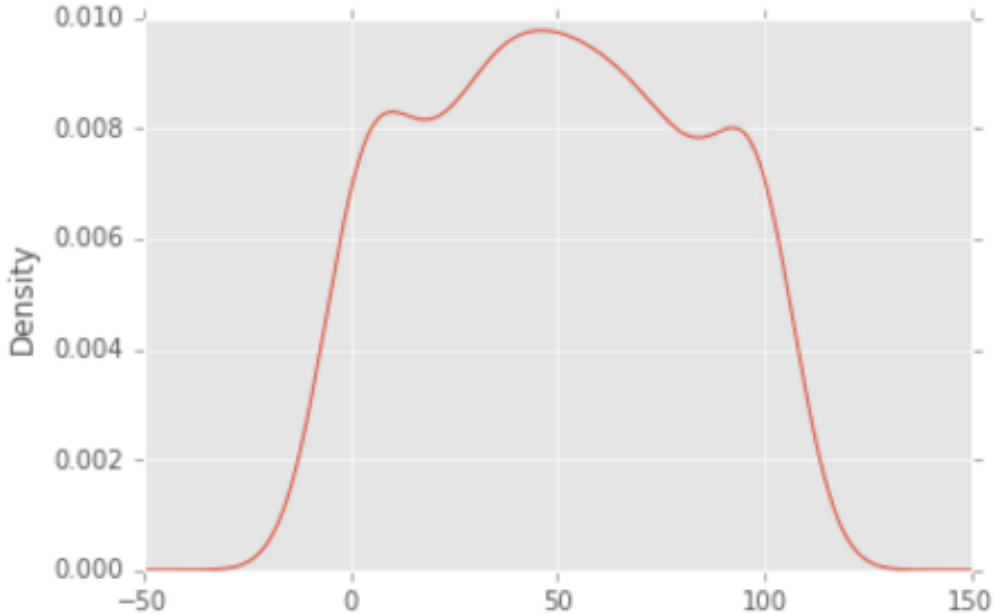
Univariate visualisations

- Understand the variable beyond mean, median, standard deviation, etc.
- Should be first part of exploring data
- Types
  - Histogram
  - Density

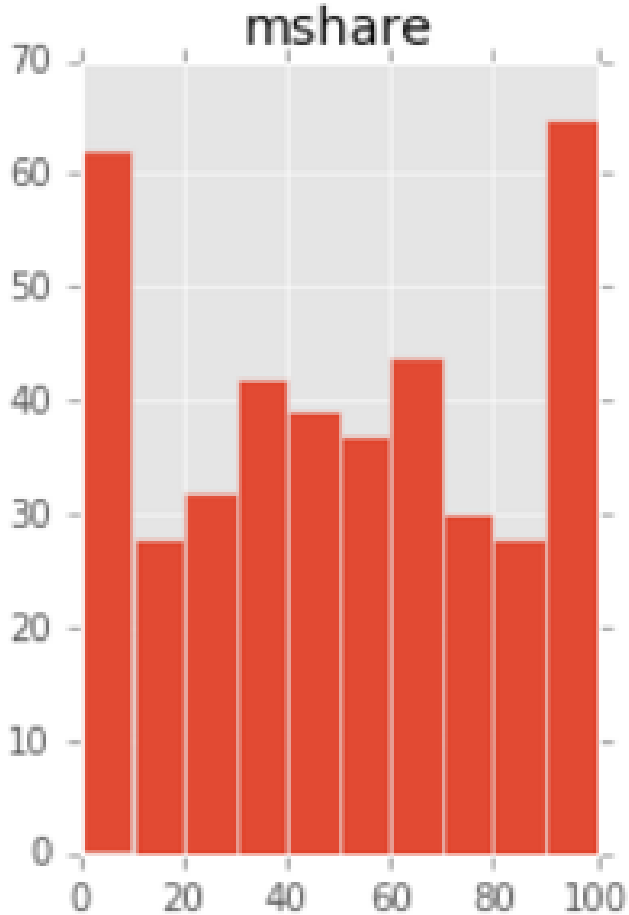


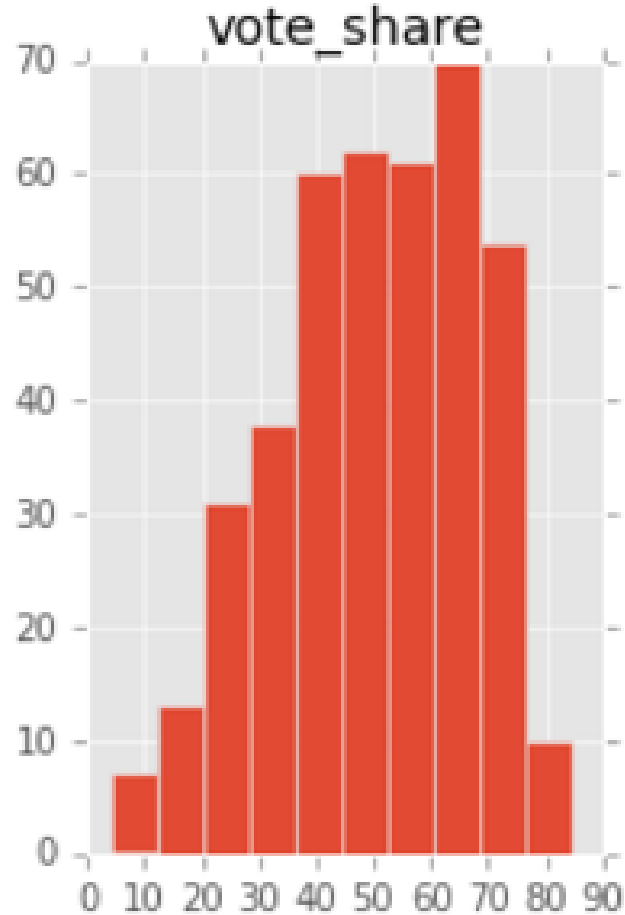
Univariate visualisations

- Understand the variable beyond mean, median, standard deviation, etc.
- Should be first part of exploring data
- Types
  - Histogram
  - Density



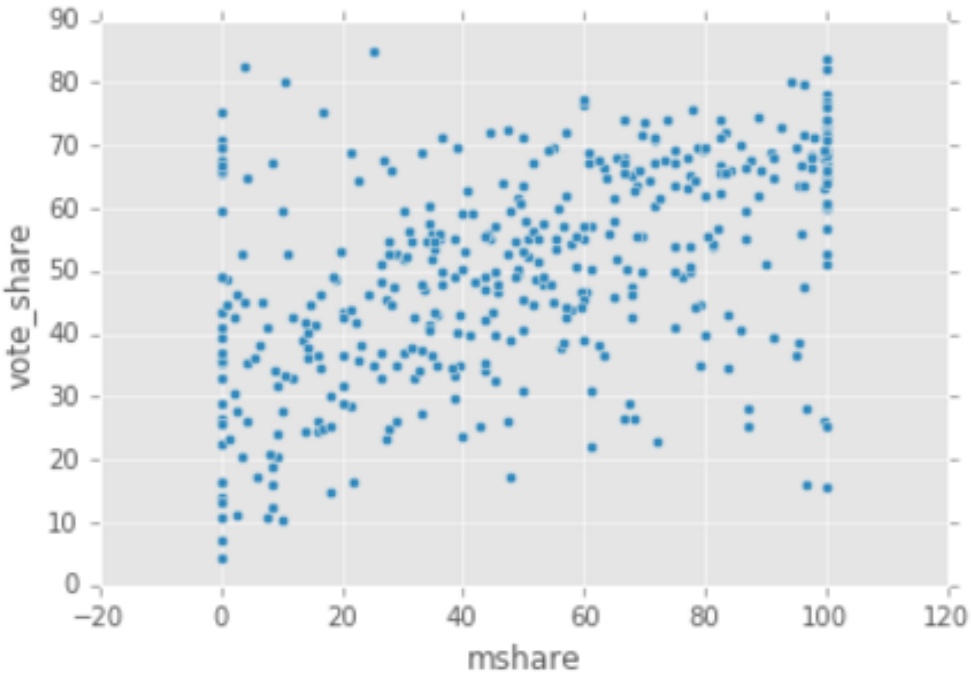
Univariate visualisations: Comparing variables





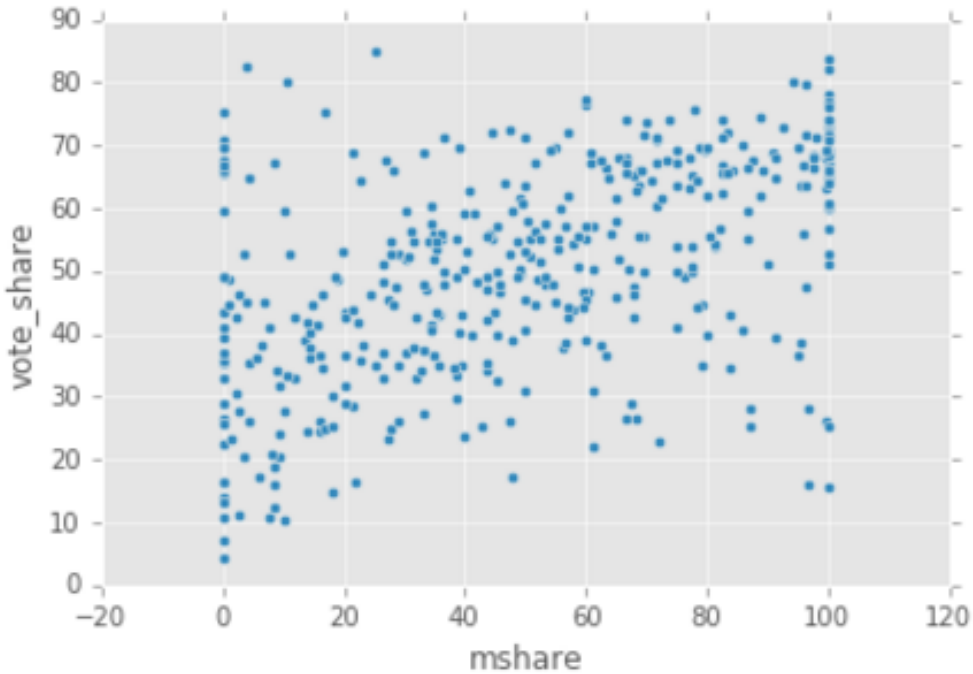
Bivariate and multivariate visualisations

- Understanding variable relationships
- First part of model exploration

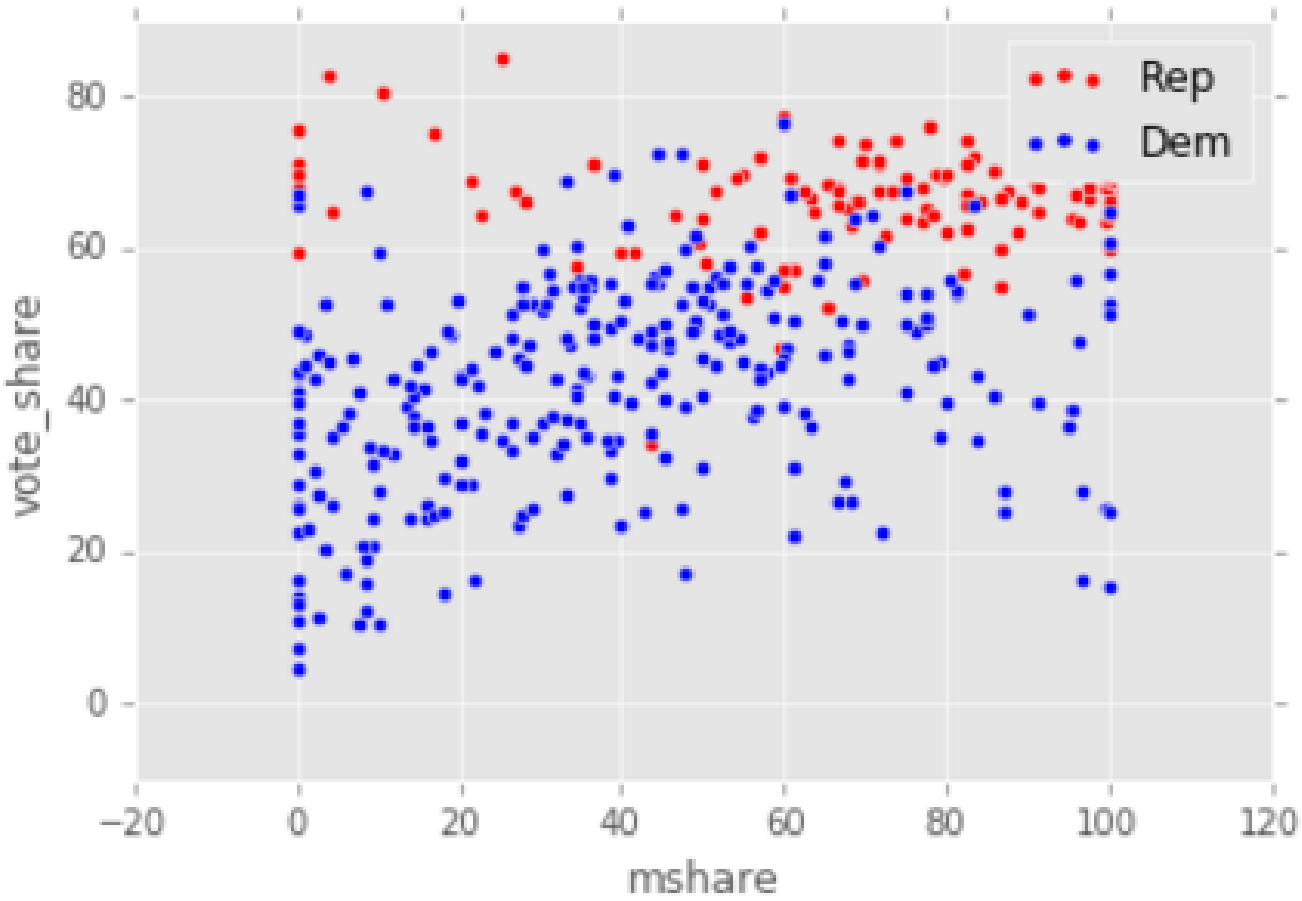


Bivariate and multivariate visualisations

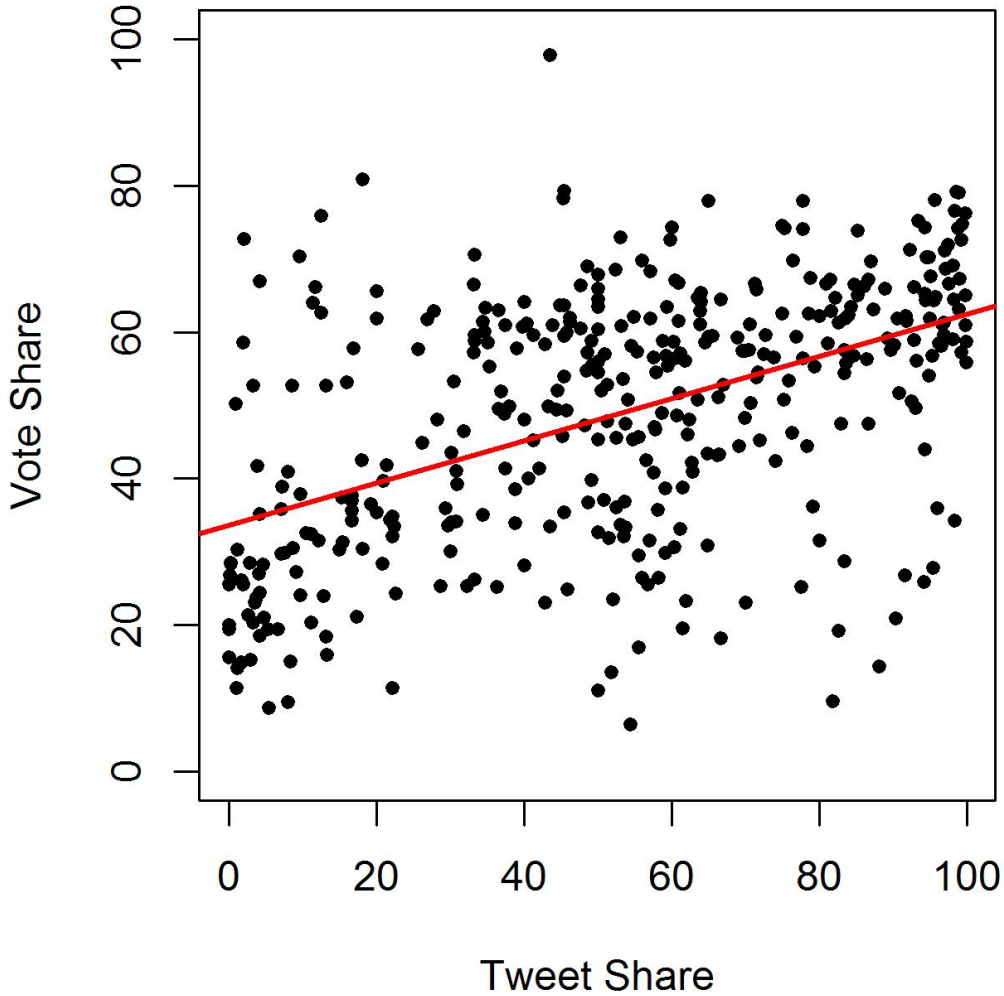
- Understanding variable relationships
- First part of model exploration



Multivariate visualisation: Adding color



Model confirmation



guessthecorrelation.com