

三国志で学ぶデータ分析

ill-identified

Japan.R, 07.12.2019, Updated at 09.12.2019

自己紹介

twitter: ill-identified



- (~2014): 学生 (経済学)
- (~2017): だいたい SAS エンジニア
- (2018~): アドテク, 機械学習エンジニア
- 詳しい経歴: [LinkedIn](#)

主語の大きい話をするのが好きです

- ・『計量経済学と機械学習の関係』
- ・『ベイズ統計とは何なのか.』
- ・『「AI の正体は最小二乗法」記事を読み解く』
- ・『TeX と Word のどちらが文書作成しやすいのか』
- ・『Python でデータ分析するのに適したグラフツール 3 選』

第一回 桃園に R を祭り 分析の議を説く

アンケートによって決定



ill-identified
@ill_Identified

例によってアンケート
2019/12/7 の #JapanR の LT で期待しているテ
ーマ

(今回はスケジュールが厳しいのでアンケート
が反映されない可能性があります)

30% 実用的なデータ分析の事例

28% 経済学と関係のある何か

14% 時空間の統計モデル

28% 何か一発芸をやれ

64票・最終結果

図 1: 当初 Long 枠が空いていたので LT から繰り上がり

こういう話をします

- 「実践的なチュートリアル」としての発表
 - 特定の題材に基づいて
 - なおかつ応用範囲の広いものになるよう
 - R の実践的な使い方を紹介する
 - あまりむずかしいことはやらないように

技術キーワード

- スクレイピング
- 名寄せ処理
- 画像認識 (?), 自然言語処理 (?)
- 機械学習 (多変量解析?)

全てを詳解するのは無理なので原稿見て

Q: なぜ三国志か?

A. 完全にその場の思いつき

目録

第一回: イントロダクション

第二回: 三国志の背景と今回の目的

第三回: `rvest` と `tidyverse` による前処理

第四回: 機械学習を利用した名寄せ処理

第五回: `skimr` と `ggplot2` による結果の提示

第六回: まとめ

第二回 三国を大いに論じ 奇謀を用いて立つ

三国志とは

1. 歴史書

- 西晉時代、陳寿作『正史三国志』
- 2世紀末の東漢～魏晉時代の正史

2. 中国文学

- 元-明代に史書と説話から創作『三国志演義』
- 義を演ずる = 儒教道徳心の布教目的 [10]

3. 日本文学

- (1, 2) を元に作者が独自の翻案・脚色
- 吉川英治作品(1939-43)[9]が有名
- 陳舜臣(1974-77)[5]¹、北方謙三(1996-98)[4]、
宮城谷昌光(2004-13)[8]

¹『インド三国志』も面白いですね

大衆文化の三国志

1. 映像作品:

- 人形劇三国志 (1982)
- 中国での多くの映画・TV ドラマ

2. 漫画:

- 横山光輝『三国志』(1971-1987): 吉川英治版に準拠
- 李學仁・王欣太『蒼天航路』(1994-2005)
- 他, 『一騎当千』『恋姫+無双』

3. コーエーテクモ (光栄) 『**三國志**』 (1985-2016)

- 三国志をモチーフにした「歴史シミュレーションゲーム」

創作と史書での扱いの差異 1/4

カ ユウ
華雄 (? - 191)

史書

- 「**胡轸**の配下として**孫堅**軍に討たれた」のみ

創作

- 董卓**配下の猛将として、逆に**孫堅**を撃退
- しかし**关羽**の噛ませ犬役

創作と史書での扱いの差異 2/4

カンコウ
関興 (? - ?)

史書

- 「父**関羽**の死後、将来を嘱望されるも数年後病死」のみ

創作

- 父の仇討ちに成功し、数度の北伐で活躍

創作と史書での扱いの差異 3/4

ソウシン

曹真 (? - 231)

史書

- 諸葛亮の北伐に対する防衛を指揮し、二度退ける

創作

- 終始諸葛亮に翻弄され、最期は罵倒され憤死

創作と史書での扱いの差異 4/4

リツウ 李通 (168-209) 史書

- 曹操の本拠地の南境を守り抜く

創作

- バチョウ
馬超の噛ませ犬役
- 眉毛が太い (蒼天航路)

何がいいたいか

- ・史書・創作で矛盾した展開が多数
- ・何が真実か・史実かは問題ではない
- ・三国志の人物像がどう変わってきたか
- ・ゲーム『三國志』の数値検証

ようやく R の話

- こんな流れで**ほぼ全て R** でやりました

1. データの取得と前処理

- **rvest, tidyverse**: データのスクレイピングと整形
- 名寄せ処理
 - 手作業・ドメイン知識
 - クラスタリング

2. **skimr, ggplot2**: 様々な切り口からデータを見る。

第三回 rvest インターネット 互聯網を智取し tidyverse 前処理を力斬す

rvest によるスクレイピング

- css セレクタまたは xpath で抽出

```
1 read_html("https://...") %>%
 2   html_node(css = "table».HOGE") %>%
 3   html_table()
```



ソースごとに異なるフォーマット

- 一般ユーザの非公式な一覧表を参考にした
 - 作品ごとにフォーマットが違う
- 次の 2 つが特に複雑な構造

1つのセルに複数の項目が凝集

ID	名前	字	ヨミ	統率	武力	知力	政治	誕生	寿命	相性	義理	野望	性格	奮奮	突突	騎走飛	斎連速	蒙樓閣	井衝投象	造石罠救	混戻心幻	罵鼓治妖	声舞療術
あ	阿会喃		アカナン	66	73	30	42	190	3	62	8	4	猪突	○×○	xxx	xxx	xxx	xxx	xxxx	xxxx	xxxx	xxxx	
い	韋昭	弘嗣	イショウ	18	17	68	74	204	6	131	11	6	剛胆	xxx	xxx	xxx	xxx	xxx	xxxx	○xxx	○xxx	xxxx	
い	伊籍	機伯	イセキ	25	24	73	85	162	5	77	10	3	冷静	xxx	xxx	xxx	xxx	xxx	xxxx	xx○x	xx○x	x○xx	
尹	尹賞		インショウ	51	54	60	67	194	6	72	6	5	冷静	xxx	xxx	xxx	○xx	xxx	○xxx	xxxx	xxxx	○xxx	
尹	尹大目		インダイモク	5	9	33	51	211	5	38	8	4	慎重	xxx	xxx	xxx	xxx	xxx	xxxx	xxxx	xx○x	xxxx	
尹	尹默	思潜	インモク	13	17	65	78	183	4	80	7	4	慎重	xxx	xxx	xxx	xxx	xxx	xxxx	○x○x	xxxx	xxxx	
う	于禁	文則	ウキン	82	76	72	57	159	5	22	8	9	冷静	○xx	x○○	x○x	○xx	x○x	x○xx	xxxx	xxxx	xxxx	
于	于誼		ウキン	67	73	42	36	204	3	126	10	3	猪突	○○x	xxx	xxx	xxx	○xx	xxxx	xxxx	xxxx	○xxx	
え	衛力	カン	エイカ	69	53	81	79	220	7	31	7	10	慎重	xxx	xxx	xxx	○xx	xxx	○xxx	xxxx	○xxx	○xxx	
袁	袁遵	伯業	エイイエイ	61	43	71	76	150	5	95	12	7	剛胆	○xx	xxx	xxx	xxx	xxx	xx○x	xxx○	xxxx	x○xx	
袁	袁胤		エンイン	27	18	42	43	163	3	140	7	7	慎重	xxx	xxx	xxx	xxx	xxx	xxxx	xxxx	xxxx	xxxx	
閻	閻宇	文平	エンエイ	70	69	46	54	209	4	50	1	12	慎重	xxx	xxx	xxx	○xx	xxx	○xxx	xxxx	○xxx	○xxx	
吉	吉肥	暗あ	ヤマシタ	65	55	64	72	176	6	101	9	5	慎重	xxx	xxx	xxx	○xx	xxx	○xxx	○xxx	xxxx	xxxx	

図 2: 『三國志 9』の人物一覧を掲載したページ

2段構成 + セルの塗りつぶしで表現

武将名	字	特技										義理	勇猛	相性	誕生	登場	没年	寿命	口調	格付け	
		統率	武力	知力	政治	合計	兵科	戦法	収集	人脈	監視										
あかいなん	-	商才	耕作	名士	兵心	練兵	収集	人脈	監視	補修											
阿会鳴	-	65	74	26	33	198	騎兵	攻撃強化	2	2	62	190	217	225	36	威厳男	★				
いせき	きはく	商才	耕作	名士			収集		弁舌												
伊籍	機伯	29	24	80	86	219	弓兵	破壊力弱化	3	0	77	162	189	226	65	策士男	★★				
いんしょう	-	商才					監視														
尹賞	-	51	44	62	66	223	弓兵	弓攻撃強化	2	0	72	194	213	260	67	丁寧男	-				
いんもく	しせん	商才				収集										兵器					
尹默	思潜	26	15	66	77	184	槍兵	知力上昇	2	0	80	183	212	239	57	能吏男	★				
うきん	ぶんそく					練兵	監視					水練		攻城							
子禁	文則	83	78	74	57	292	弓兵	弓軍強射	1	1	22	159	184	221	63	勇将男	★★				
えいかん	はくぎょく		名士				弁舌					遠射		兵器							
衛瓘	伯玉	69	46	79	78	272	弓兵	弓攻撃強化	2	0	31	220	239	291	72	策士男	★				
えんいん	-		耕作																		
袁胤	-	32	14	39	41	126	槍兵	防御強化	2	0	140	163	184	199	37	丁寧男	-				
えんき	けんえき					収集															
袁熙	顕奕	66	51	63	65	245	弓兵	射程弱化	2	0	101	176	190	207	32	策士男	-				

図 3: 『三國志 12』の人物一覧ページ

tidyverse を使いましょう

- 日本語資料充実
 - 『整然データとは何か | Colorless Green Ideas』
 - 『データハンドリング | Kazutan.R』
 - 『データラギングリングチートシート』
 - 『heavywatal』
 - 『tidyr 1.0.0 の新機能 pivot_() / tidyr-pivot』
- 日本語処理のため, `stringi` も必要



tidyverse で整形

```
1 df9 <- filter(sources, title==9)$html[[1]] %>%
  read_html %>% html_node("table") %>% html
  _node("table") %>% html_table(header=T)
  %>% as_tibble
2 df9 <- filter(df9, ID!="ID") %>% mutate_all(na
  _if, "") %>% fill(ID) %>% mutate_at(.vars=
  vars(統率, 武力, 知力, 政治, 誕生, 寿命,
  相性, 義理, 野望), .fun=as.integer)
  %>% rename(name=名前)
3 df9 <- df9 %>% dplyr::select(-奮奮奮戰鬪迅, -
  突突突破進撃, -騎走飛射射射, -齊連連射
```

整形結果

- **tidyverse** で整然化
- 変数の標準化はしない
 - 作品ごとにルールが違う
 - 値の範囲は一貫して 1~100
- **7,115 件/1,120 名の人物データ**

```
# A tibble: 7,115 x 4
  title order name   data
  <chr> <int> <chr> <list>
1 1      1     伊籍  <tibble [1 x 6]>
2 1      2     于禁  <tibble [1 x 6]>
3 1      3     袁胤  <tibble [1 x 6]>
4 1      4     袁熙  <tibble [1 x 6]>
5 1      5     袁紹  <tibble [1 x 6]>
6 1      6     袁尚  <tibble [1 x 6]>
7 1      7     袁術  <tibble [1 x 6]>
8 1      8     袁譚  <tibble [1 x 6]>
9 1      9     閻闓  <tibble [1 x 6]>
```

第四回

某三度顧み策を決し 機械學習名を薦む

名寄せ処理

- 人物ごとに集計したい
- 誤記・表記のゆらぎ問題
 - 個人作成の入力ミス多いリスト
 - そもそも原典でも誤記・表記ゆらぎアリ
 - 同姓同名にも注意
- まず手作業で修正
 - 時代の異なる人物を除外 (**179 件**)
 - 漢字の使われてない文字を修正 (**122 件**)
 - 3 文字以上の人名のみ検査し修正 (**19 件**)

手動作業 1/3: 三国志以外の登場人物

- 8以降では隠し要素として他の時代・地域の人物も存在
 - 例: 管仲・楽毅, 李信, 劉邦・項籍, 高長恭, 岳飛, **成吉思汗**, 秦良玉, **織田信長**, 糸芸爪覽
 - 来年の新作には『銀河英雄伝説』も
- 高能力は評価に影響大
- 原則: 『正史三国志』『三国志演義』に関係する人物のみ対象
- **179件除外**

手動作業 2/3: 漢字が使われてない人名

- shift_jis に含まれていない字の代用
 - 「竜」は現在の環境でも非対応
- 122 件の修正箇所

正	別表記
チョウコウ 張 郎	張 [合 β], 張コウ
リュウシュン 劉 琅	劉・(文字化け)
ソンワン 孫竜	孫ワン

表 1: 非漢字の表記のゆらぎ例

手作業 3/3: 字数の多い人名のみ確認

- 当時の中国人名は 2 字が多い
- 19 件の修正箇所発見

正	別表記	解説
キヨショウ 許 劄	許子将	あざな 字と混同
キンカンサンケツ 金環三結	金環結	ハードの制約?
シュクユウ 祝 融	祝融夫人	表記ゆらぎ
シン ギ ロク 秦宜祿	秦誼	原典由来の誤記
ケイドウエイ 邢道榮	刑道榮	非 SJIS 漢字

データの品質管理

- 2字でも同様の表記ゆらぎの可能性
- 手作業だけでは辛い
- きりがない...
- 「機械学習」でなんとかする

画像認識 (?) + 教師なし学習

- 文字画像に対して**教師なし学習**する
 - 32x124 でビットマップ出力
- 問題: 対応フォントがない人名
 - RStudio に孫輩を書き込むと**エディタがバグる**
 - 花園明朝B にグリフ収録
- **plot()** で文字を描画し, ビットマップファイルに書き出し, ピクセルを行列として読み込む

文字を伸ばして幅を統一する

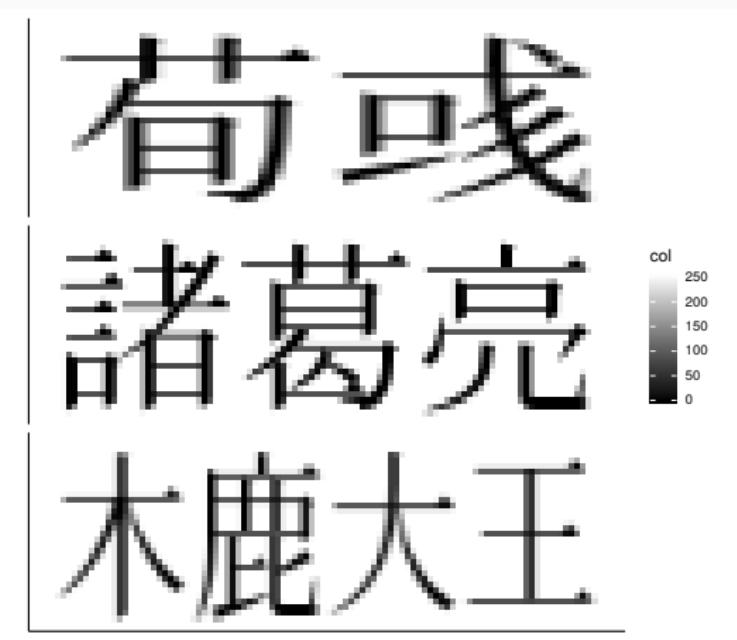


図 5: ピクセル表示の例

類似度の計算

- テンプレートマッチングというらしい [6]
 - 画像の特徴量を作り、類似度（距離）を計算
(factoextra)
- 以下 2 種類の計算方法で試す

$$s(\mathbf{x}, \mathbf{y}) := \begin{cases} \sqrt{\sum_{d=1}^D (x_d - y_d)^2} & \text{(ユークリッド)} \\ \sum_{d=1}^D |x_d - y_d| & \text{(マンハッタン)} \end{cases}$$

特徴量の計算

1. ピクセル情報をそのまま使う

- $32 \times 128 = 4096$ 次元
- 末端が空白になるので実質 4025 次元

2. 行・列ごとに情報を圧縮 ([3] の方法)

- 白・黒の変化の回数 (微分), 黒領域の割合 (積分)
- $(32 + 128) \times 2 = 320$ 次元
- 実質 317 次元
- パッケージはないので tidyverse で

• 両者似たような結果になった

文字形状類似度の結果

name1	name2	Manhattan	Euclidean
于糜	于糜	4.94	4.79
車胄	車胄	4.87	4.93
王凌	王凌	4.81	5.03
夏侯威	夏侯咸	4.73	4.79
吳鋼	吳綱	4.65	4.79
薛翊	薛珝	4.59	4.58
邢道榮	刑道榮	4.58	4.00
全禕	金禕	4.55	4.47
王匡	士匡	4.52	4.45
劉瓊	劉瀆	4.46	4.47

表 3: M 類似度上位 10 件, 誤字を強調

特に紛らわしい一例

正	誤	解説
シャチュウ	車胃	「胃」の下
カンイ	関彝	「米糸」と「米分」
ショウカイ	鍾会	鐘ではない

表 4: 発見できた紛らわしい表記のゆらぎ例

こんな分かるか!

似ているが別人

キンイ 金禪	ゼンイ 全禪
カコウイ 夏侯威	カコウカン 夏侯咸
オウキョウ 王匡	シキョウ 士匡
トウガイ 鄧艾	トウシ 鄧芝
カンカイ 桓楷	カンカイ 桓階

表 5: 「誤検知」された別人物

こんな分かるか!

名寄せ処理の改良の余地

- 誤検知も少なからず
- 王匡/土匡, 朱異/王異の誤検知は避けたい
- $4096 \rightarrow 320$ 次元削減でも影響なし
- より効果的な特徴量のとり方がある?
- 誤字は部首や音の似た字で起こる?
- **なんもわからん**

DBpedia で名寄せ二重チェック

- DBpedia = Wikipedia を構造化
 - SPARQL で取得可能
 - R なら SPARQL パッケージ
- Wikipedia のカテゴリで条件付けて取得
- これも ground-truth なデータではない
 - 「卑弥呼」は三国志の人物か?
 - DBpedia の更新頻度の問題

(こっちのが簡単では?)

第五回

名を寄せ skimr 再び上表し

ggplot2 像に見えて英雄を論ず

登場武将の変遷をグラフで表す

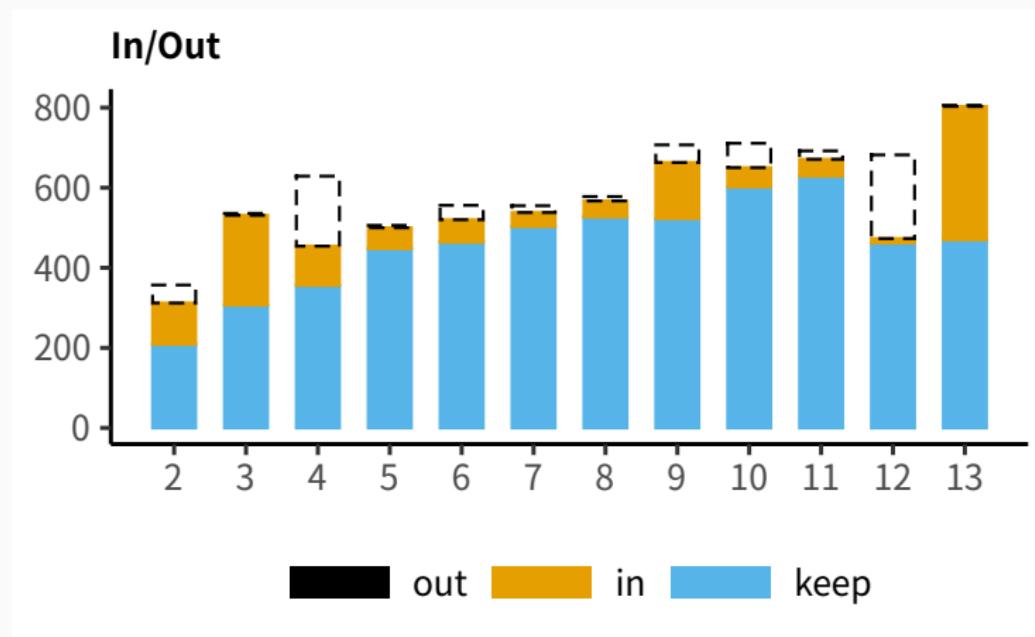


図 6: 登録・除外フロー

skimr で要約統計量を表示

- ・今日はスライドにするには多すぎる

title	variable	min	p25	p50	p75	max	mean	sd	skewness	kurtosis
1	武力	15.00	36.00	57.50	78.75	100.00	57.37	24.63	-0.00	1.80
2	武力	11.00	41.00	61.00	74.00	100.00	58.78	21.33	-0.09	2.13
3	武力	15.00	52.00	64.00	71.00	100.00	61.40	17.08	-0.46	3.05
4	武力	13.00	49.25	66.00	75.00	100.00	61.39	20.25	-0.56	2.63
5	武力	7.00	44.00	67.00	76.00	100.00	60.26	22.74	-0.64	2.47
6	武力	16.00	42.75	62.50	73.00	100.00	58.36	20.00	-0.32	2.27
7	武力	11.00	45.25	63.00	74.00	98.00	58.79	20.37	-0.50	2.26
8	武力	10.00	46.00	65.00	72.00	100.00	58.68	20.83	-0.70	2.59
9	武力	0.00	36.00	65.00	72.00	100.00	55.63	24.35	-0.66	2.33
10	武力	1.00	39.00	64.00	73.00	100.00	56.43	23.37	-0.67	2.45
11	武力	1.00	33.50	64.00	73.00	100.00	55.05	24.79	-0.57	2.16
12	武力	2.00	37.00	66.00	75.00	100.00	57.09	25.15	-0.66	2.31
13	武力	1.00	34.50	64.00	72.00	100.00	55.09	23.90	-0.58	2.21

表 6: skimr による表 (一部)

skimr の特徴

- **summary()** よりデフォルトの表示見やすい
- **group_by()** すればグループ集計
- 日本語情報少ない: 『niszet 氏のスライド』
 - 今は更に仕様が変わっている...

```
1 my_skim <- skim_with(numeric = sfl(  
    skew = skewness, kurto = kurtosis,  
    hist = NULL), append = T)  
2 DATA_FRAME %>% group_by(title) %>% my  
    _skim()
```

(発表後追記) 要約統計量

- min-max 正規化で揃えるべき

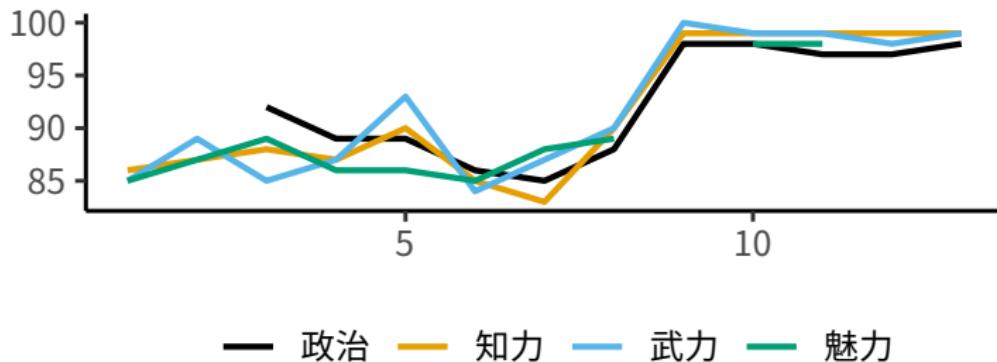


図 7: 作品ごとにレンジ (最大 - 最小) に幅がある

演義で活躍の盛られた人物の評価

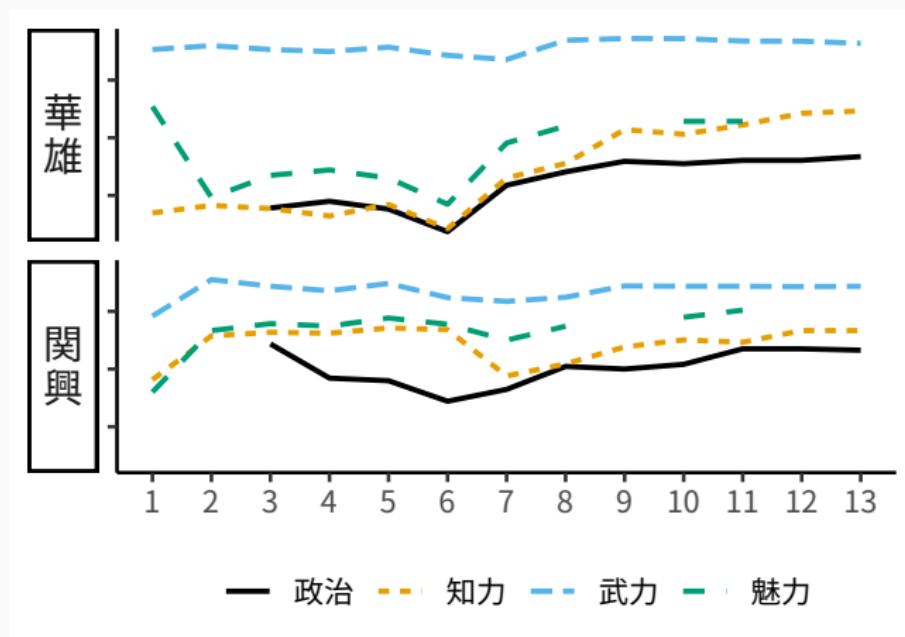


図 8: 両者ともあまり低下していない?

演義で扱いの悪い人物の評価

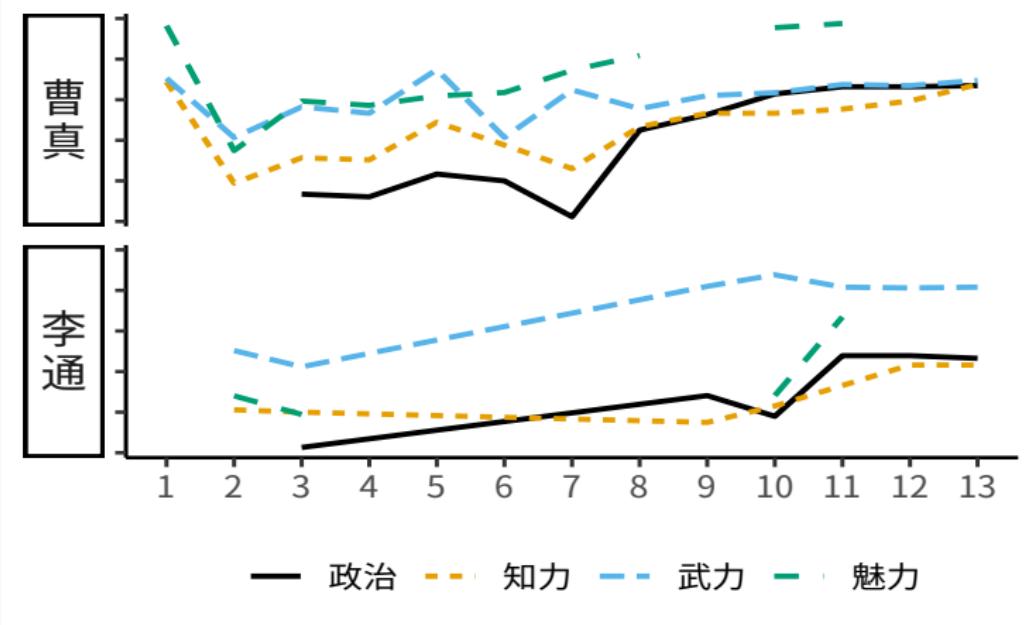
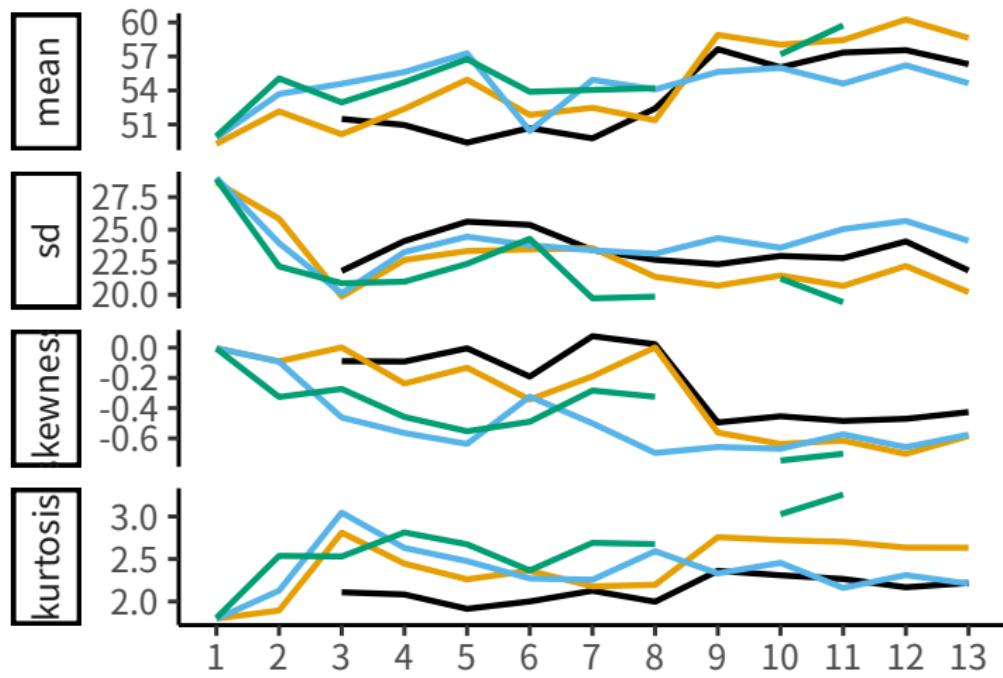


図 9: 後期の作品ほど評価が向上している?

(発表後追記) 全体の傾向を見る

- 分散の減少・非対称性の増加



シリーズごとの分布をバイオリン図で

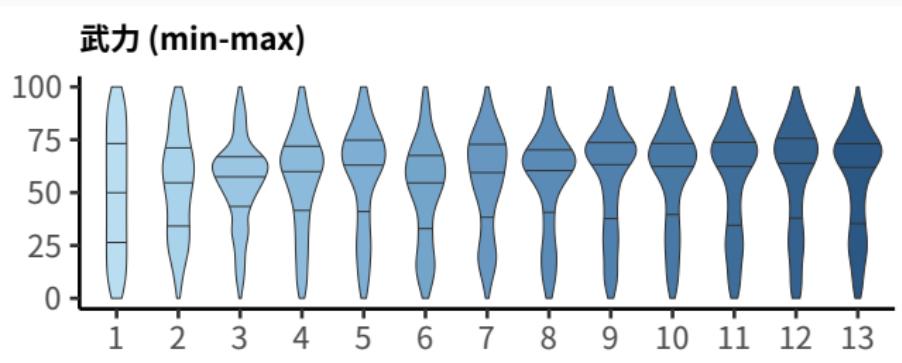


図 11: 後期の作品につれ分布が変化していく

ggplot2 (+ggthemes) で書ける

```
1 ggplot(df_center,
2         aes(x = title, y = 武力,
3               fill = title)) +
4   geom_violin(draw_quantiles = c(0.25, 0.5,
5                                0.75)) +
6   scale_fill_continuous_tableau(guide = F)
```

- カテゴリカルな色分けは `_colorblind()`

どういう傾向か

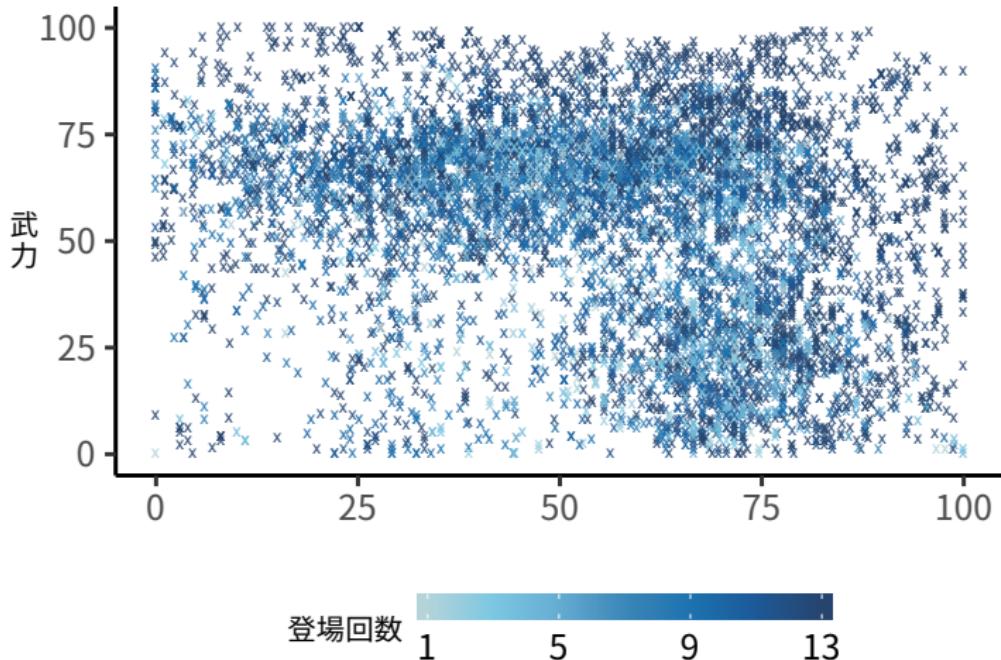


図 12: シリーズ登場回数との 3 軸

(発表後追記) ステータス分布の一極集中

主要ステータスの平均値で見せたほうが分かりやすかった

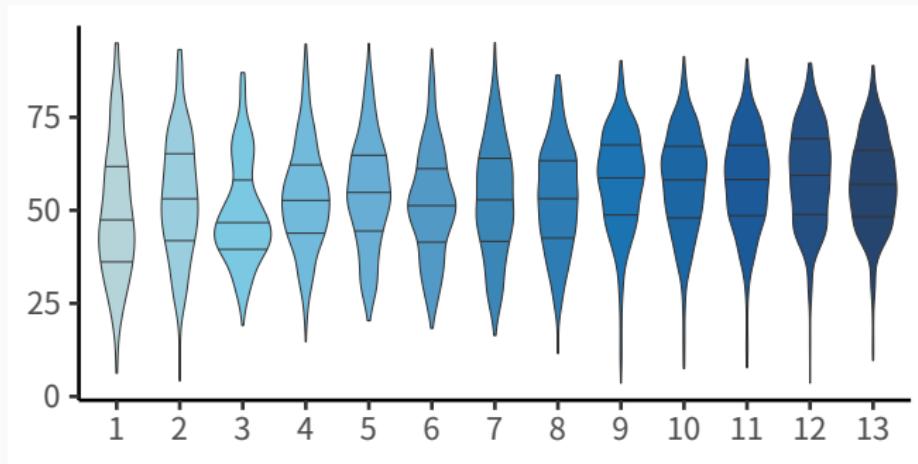


図 13: 平均値で見ると一極化している

(発表後追記) バイオリン図の分散と尖度

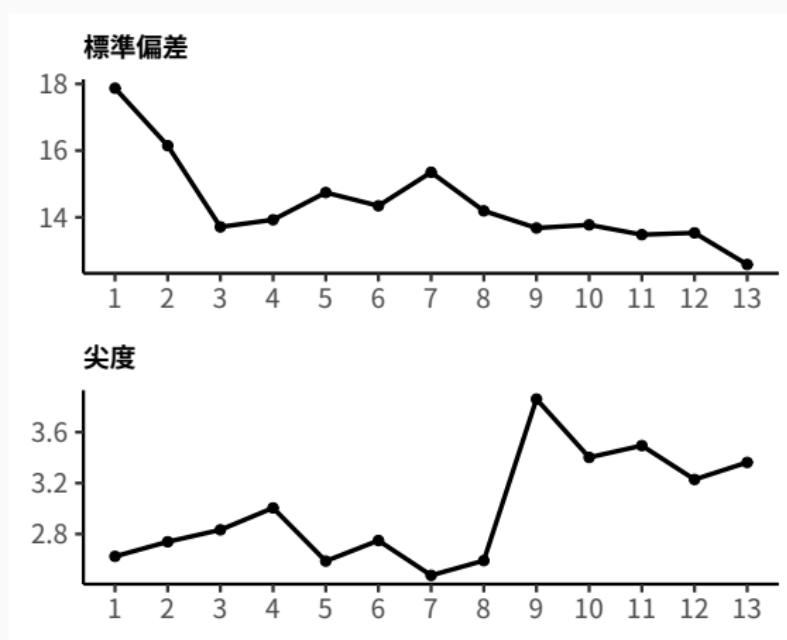


図 14: 明らかに一極集中している

(発表後追記) 主成分分析で見てみる

- 取得できる全能力情報を**主成分分析**
- `stat::prcomp()` と
`factoextra::fviz_pcabiplot()`
- 主成分分析はスケーリングに注意
 - 『学力テストの主成分分析のバイプロット - 裏 RjpWiki』
 - 今回は全て**標準化**

(発表後追記) 主成分バイプロット

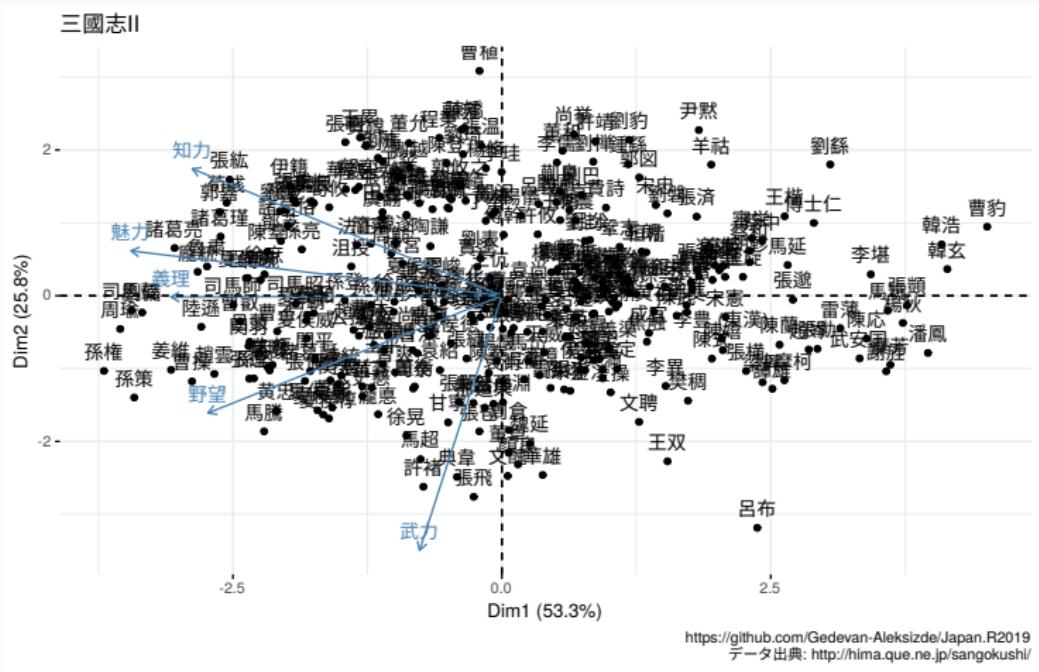


図 15: 三國志 2 の分布

(発表後追記) 三国別主要ステータス平均

曹操(魏), 劉備(蜀), 孫權(呉)に最も近い勢力別

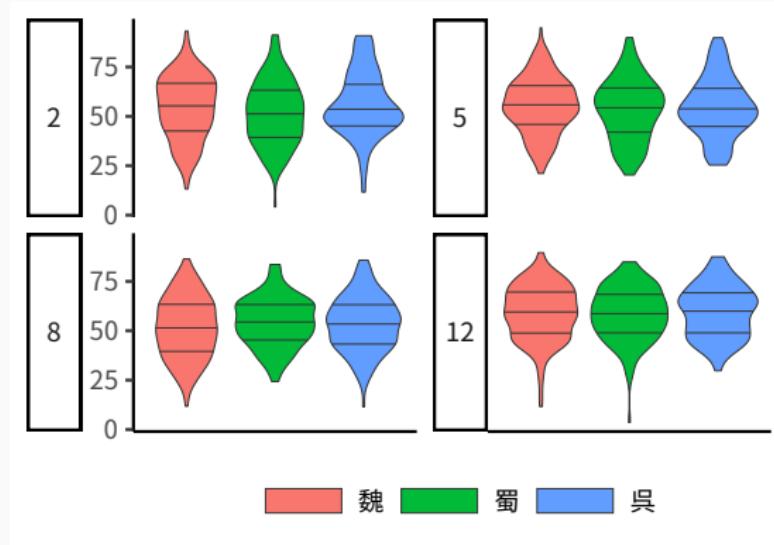


図 16: 後期の作品ほど三国とも分布の裾が細くなる

補足: グラフ作りにもルールがある

- [1] は `ggplot2` のコードもあり初心者向け
- 内容が近い日本語の教科書もある [7]

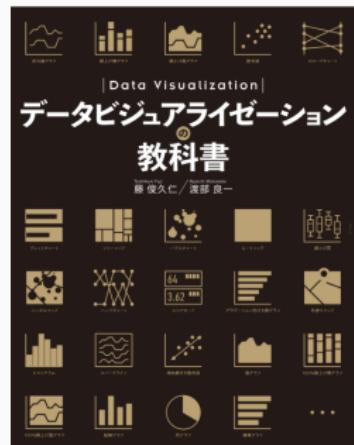
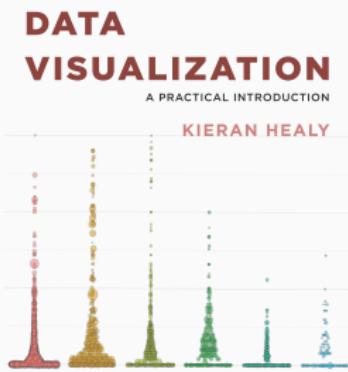
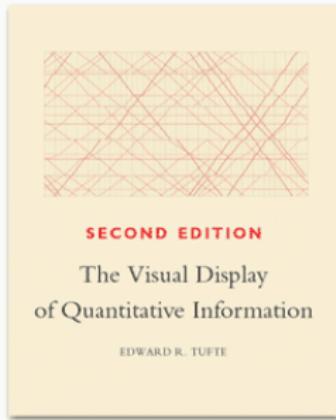


図 17: グラフの教科書 [2, 1, 7]

補足: ggplot2 の手引き

- 日本語資料充実
- 『ggplot2 に関する資料 | Kazutan.R』
- 『ggplot2 - きれいなグラフを簡単に合理的に』

第六回

是を以て新謀献じ

分析一統に帰す

(発表後修正) 検証結果のまとめ

- 主要能力値だけを見ると一極集中が進んでいる
- 分布形状が大きく変化
- 一方で後期では能力値の極端に低い者も
- 特定の作品がどう影響しているかは言えない

まとめ

- ほぼ全ての作業を R でやった:
 - スクレイピング: **rvest**
 - データの整形要約: **tidyverse**, **skimr**
 - 多変量解析: **factoextra**, **tidyverse**
 - グラフ作成: **ggplot2**, **ggthemes**
- R でやってないこと:
 - このスライドと原稿の作成
 - Rmarkdown はレイアウト部分ほぼ \LaTeX 依存のため
 - Hmisc::latex() は便利

劇終

THE END

参考文献 i

- [1] Healy, Kieran (2018) *Data Visualization: A Practical Introduction*, Princeton, NJ: Princeton University Press, retrieved from [here](#).
- [2] Tufte, Edward R. (2001) *The Visual Display of Quantitative Information*, Cheshire, Conn: Graphics Press, 2nd edition.

参考文献 ii

- [3] 鴨下隆志・奥村健一・高橋和仁・増村正男・矢野宏 (1998) 「文字認識におけるマハラノビスの距離による判定の研究」, 『品質工学会』, 第 6 卷, 第 4 号, 39–45 頁, retrieved from *here*.
- [4] 北方謙三 (1996) 『三国志』, 角川春樹事務所.
- [5] 陳舜臣 (1974) 『秘本三国志』, 文藝春秋.

参考文献 iii

- [6] 糟谷勇児・山名早人 (2006) 「二種類の SVM を用いたオンライン類似数式文字識別」, 『電子情報通信学会技術研究報告. PRMU, パターン認識・メディア理解』, 第 105 卷, 第 614 号, 55–60 頁, 2 月, retrieved from *here*.

参考文献 iv

- [7] 藤俊久仁・渡部良一 (2019) 『データビジュアライゼーションの教科書』, 秀和システム, 東京, retrieved from *here*, OCLC: 1103469309.
- [8] 宮城谷昌光 (2004) 『三国志』, 文藝春秋.
- [9] 吉川英治 (1939) 『三國志』, 大日本雄辯會講談社.

参考文献 v

- [10] 渡辺義浩 (2011) 『三国志: 演義から正史,
そして史実へ』, 中央公論新社, 東京,
OCLC: 752021927.