

「第7期(2016年度)AITCオープン活動発表会」

AITC女子会 発表～はじめての データ分析

チーム:「ペット、飼ってますか？」

吉井(NUL)・石倉(NUL)

野田・遠山・稲田(TIS)

概要



犬の飼育率は県によって違います。

なぜ違うのか、ペットを飼おうと決めるとき人は何の影響を受けているのか？

国勢調査や気象などの公開データを使って傾向を探ってみました。

<http://inuneko-magazine.com/column/entertainment/tips/dog-lover-ranking>

作業の流れ

1. テーマ選定

自治体の公開データで「気になる項目」をピックアップ

2. 情報収集

仮説を立てるための情報をネットで収集

3. 仮説を立てる

チームディスカッションで仮説を出し合って検討

4. データの収集

データをネットで収集して持ち寄り

5. データの分析

Rでデータを分析→結果をチームで考察

6. 見直しと再分析

対象データの見直しと追加→結果をチームで考察

1. テーマ選定

- オープンデータ、自治体サイトなどを見てテーマ探し
- 横浜市ポータルで見つけた「犬」をテーマに選定
 - データの集めやすさから、分析対象は横浜市でなく全国版(都道府県別データ)に決定

犬

い				
家	家出	遺棄	イギリス館	医師
いじめ	移出・移入	いずみ野線	委託児童	1時間値
1住宅当たり	1世帯当たり人員	1世帯当たり延べ面積	1日当たり乗降人員	1日平均乗車人員
市場(いちば)	市立公園	一戸建	一世帯あたりの人員	一般会計
一般住宅用地	一般世帯	一般団体	移転	緯度
井戸	移動体通信	移動前・移動後の住所地	犬	犬の保護
違反	イベント	入込観光客数	医療	医療施設静態調査
医療費	衣料品	医療保険	飲酒	飲酒運転
飲食店	インターネット	インフルエンザ	インフレ	

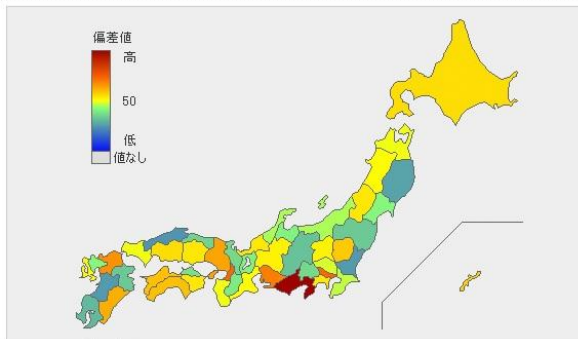
<http://www.city.yokohama.lg.jp/ex/stat/index-k.html>

2. 情報収集

- 仮説を立てるために、ペット飼育の実態、動機、課題・・・の情報をネットで収集(図は一例)
 - － 数値データ、アンケート
- 自治体等の実態調査（下記は千葉県のもの）

■都道府県別統計とランキングで見る県民性:
ペットショップ店舗数ランキング
<http://todo-ran.com/t/kiji/21135>

ペットショップ店舗数 [2014年第一位 静岡県]



■I LOVE DOGな都道府県ランキング発表！

<https://prtimes.jp/main/html/rd/p/000000070.000002816.html>

※出典は廣濟堂「いぬねこマガジン」(現在はサイト閉鎖)

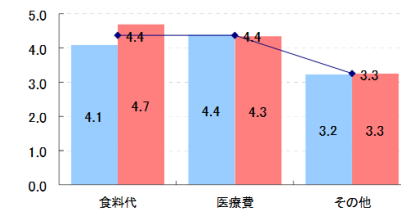
<http://inuneko-magazine.com/column/entertainment/tips/dog-lover-ranking>

■自治体等の実態調査（下記は千葉県のもの）

https://www.pref.chiba.lg.jp/eishi/pet/doubutsu/documents/shiyoujittai_kekka.pdf

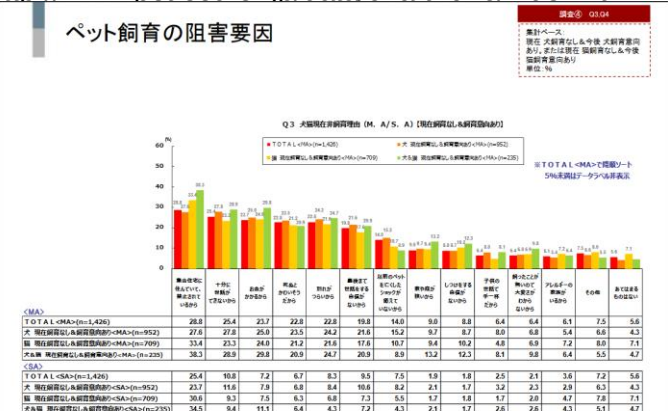
犬の飼育経費(万円)

■ 非人口集中地区平均 ■ 人口集中地区平均 ◆ 経費平均



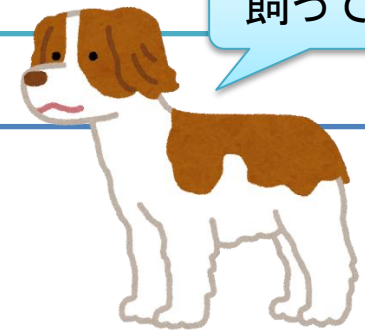
■日本ペットフード協会 飼育実態調査(アンケート)

<http://www.petfood.or.jp/data/chart2016/index.html>



3. 仮説を立てる

- 人が「犬を飼うかどうか」を決めるとき、こういった要因の影響を受けているのか？



- 立てた仮説
 - 年収(収入): 多いほど犬を飼う？
 - 家族構成: 子供がいると犬を飼う？／一人暮らしだと飼う(癒し)？
 - 住宅環境: 広いほど飼う？／マンションより戸建のほうが飼う？
 - 自然環境: 散歩しにくいと飼わない？(雪国など)

4. データの収集

- データ収集の観点
 - 仮説と関連性がありそうな公開データ
 - 間接的な関連もあり得るので、項目は少し広めに収集
 - 「意外な関連性」が発見できそうなデータ
 - 県別データで、犬の飼育頭数と分布が似た項目をピックアップ

4. データの収集

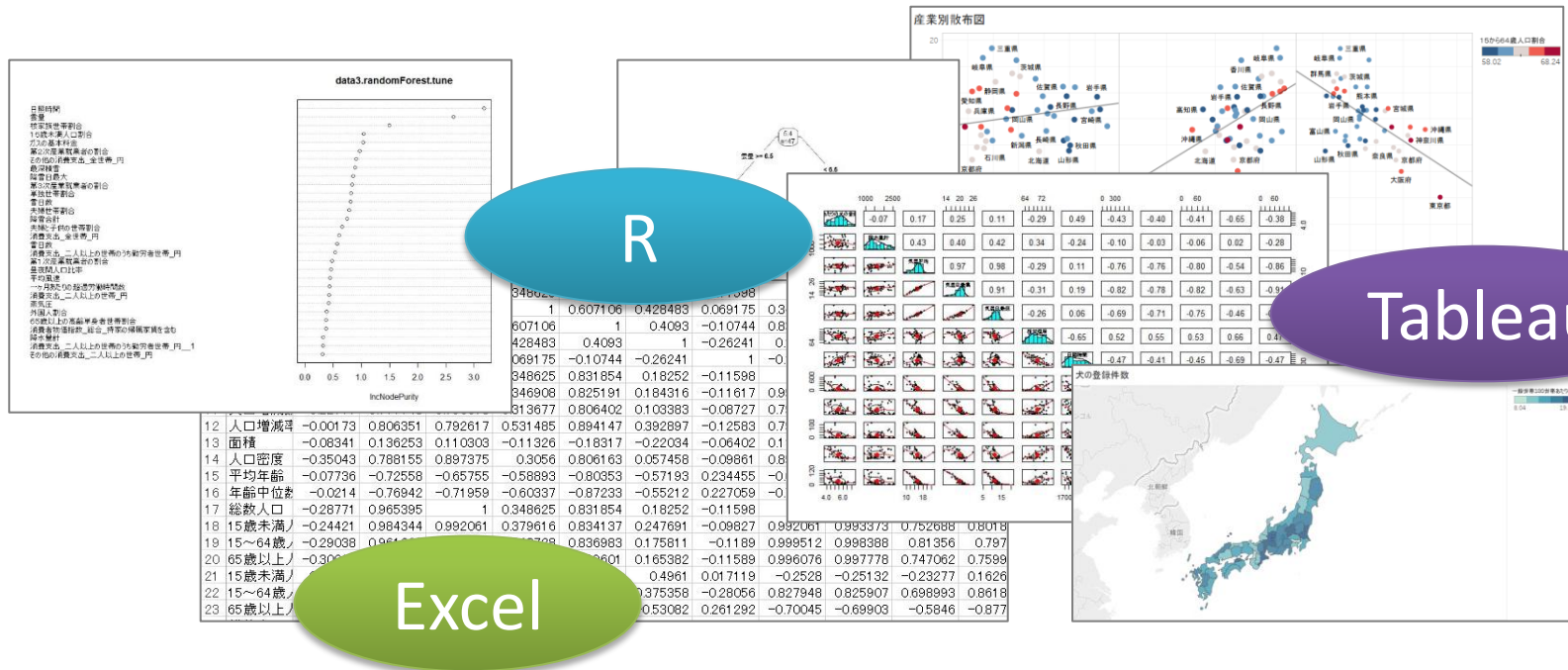
- 対象データ(県別)
 - 目的変数: 人口100人あたりの犬の登録数
 - 説明変数: 集めたデータから選抜
 - 人口・世帯(世帯種別、世帯構成、・・・)
 - 労働・賃金(年収、労働時間、持家比率、消費・・・)
 - 気象データ(日照時間、降水量、・・・)

データ入手元

- 政府統計ポータルe-Stat
(<https://www.e-stat.go.jp/SG1/estat/eStatTopPortal.do>)
- 気象庁(<http://www.data.jma.go.jp/gmd/risk/obsdl/index.php>)
元データは観測所単位なので、各県庁所在地に一番近い観測所のデータを県の値とする。
欠損値は隣接県の値で代替
- 都道府県格付研究所(<http://grading.jpn.org/index.html>)

5. データの分析

- 基本統計量、地図マッピング・・・データの姿の観察
- 散布図行列、相関行列・・・特徴的なデータの把握
- 決定木、ランダムフォレスト・・・寄与度の確認
- 寄与度が上位の項目の検討(チームディスカッション)



5. データの分析

初回の結果

- 仮説の中で「人口100人あたりの犬の登録数」と最も相関が強いのは「自然環境」
- 年収は無相関

当初の仮説	データに見る関連性	データ項目と相関係数
年収	× 無相関	サラリーマン平均年収 (-0.08)
家族構成	○ 弱い相関	15歳未満人口割合 (0.23)
自然環境	◎ やや強い相関	日照時間 (0.71) 雲量 (-0.67)
家の広さ	? 判断不能	(使えるデータがない)



（気象が一番大きい要因・・・なのか？）



うちじゃ犬は飼えないの！
雲量が多いんだから

6. 見直しと再分析

- ここまでの観察結果、仮説との整合性を踏まえて、分析の方針を見直し
 1. 目的変数の見直し
犬は個人ではなく「家庭」で飼うもの
→目的変数を「人口xxx人あたりの犬の登録数」から「xxx世帯あたりの犬の登録数」に変更
 2. 説明変数の拡充
 - 世帯構成(子供がいるか)のデータを追加
 - 世帯収入のデータを追加

6. 見直しと再分析

見直しの結果

- 目的変数を「人口あたり」から「世帯数あたり」に変えたことで説明変数の寄与度が入れ替わった
- 自然環境の影響が弱まり、産業構造・家族構成が上位に

《ランダムフォレストで見た、寄与度上位の項目》

「人口100人あたりの犬の登録頭数」

データ	相関係数
日照時間	0.71
雲量	-0.67
核家族世帯割合	0.39
15歳未満人口割合	0.23
ガスの基本料金	0.50



「100世帯あたりの犬の登録頭数」

データ	相関係数
第3次産業就労者の割合	-0.55
単独世帯の割合	-0.41
第2次産業就労者の割合	0.34
18歳未満がいる世帯の割合	0.32
第1次産業就労者の割合	0.33

チームの結論

- 犬の飼育率は人口（個人）単位ではなく、世帯単位で見るほうが妥当と思われる
- 「人口あたり」の分析結果：
 - 晴れている地域ほど、犬をよく飼う（非常に強い要因）
- 「世帯あたり」の分析結果：
 - 都市部ほど犬を飼わない傾向がある
 - 子供がいる世帯ほど犬を飼う傾向がある
 - 一人暮らしだと犬をあまり飼わない傾向がある



- 「ペット」について、もっとやってみたいこと
 - 大型犬／小型犬など区分ごとの分析（データが揃わず断念）
 - 犬以外のペットの分析（データが揃わず断念）
 - 日本以外の調査
 - 都道府県より細かい単位での分析。データの収集と加工は大変だが、もっと見えてくるものがあるのでは
 - もう一歩踏み込んで、ビジネスにつながる分析。たとえばペットショップの出店推薦や販売予測など
- 「分析」について、今後やってみたいこと
 - 時系列の分析
 - 複数相関などで深掘り
 - 仮説→検証サイクル（今回のテーマは検証がないため）

- 苦労した点
 - Rのバージョンの違いでつまづいた
- 学んだこと
 - データ集めの苦労、加工の手間
 - Rのパッケージなど、新しいものが学べた
 - 目的変数の設定(=分析の切り口)はとても重要
- その他
 - グループディスカッションが楽しかった！



最新情報は
こちらをご参照ください



<http://aitc.jp>



<https://www.facebook.com/aitc.jp>



ハルミン

AITC非公式イメージキャラクター