## Past Exams

## Data Management and Visualization

Version 1.0.2
© Diego Monti, 2021
(CC) $(\underset{\text { BY NC }}{(-)}$

This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.
To view a copy of this license, visit
http:/ / creativecommons.org/licenses/by-nc-nd/4.0/.
You are free: to copy, distribute, display, and perform the work
Under the following conditions:

(1)
Attribution. You must attribute the work in the manner specified by the author or licensor.

Non-commercial. You may not use this work for commercial purposes.
No Derivative Works. You may not alter, transform, or build upon this work.

- For any reuse or distribution, you must make clear to others the license terms of this work.
- Any of these conditions can be waived if you get permission from the copyright holder.

Your fair use and other rights are in no way affected by the above.

## EXAM OF 2020-01-31

## Visualization

## College Educated Voter Preferences



Source: PEW Research Center

Source: https://thedailytexan.com/2016/10/21/college-educated-voters-will-back-clinton-according-to-survey

## Analysis

Analyze the above graph that was published by "The Daily Texan" on Oct 21,2016 in an article entitled "Collegeeducated voters will back Clinton, according to survey". Please remember that the conventional color for the Democratic Party in US is blue and for the Republican Party it is red.

## Question

- The question is fairly clear: how does preference for the two party varies as educational level changes?


## Data quality

- The quality of the data is reasonably suitable to answer the question.
- The percentages of each donut do not sum to $100 \%$ because they refer to different wholes.
- The values should be summed by education level and we can assume there is roughly $20 \%$ of "undecided".


## Data quality

| Characteristic | Adequate | Comments |
| :--- | :--- | :--- |
| Accuracy | Yes | Percentage numbers |
| Completeness | Partly | Not sure about how to <br> interpret the missing $20 \%$ |
| Consistency | Partly | Sum of percentages not <br> equal to $100 \%$ |
| Currency | Yes | Presumably, data is from <br> 2016, but that is ok |
| Credibility | Yes | Source is reported as a <br> polling company |
| Understandability | Yes | Data is easy to understand <br> Single \% point precision is <br> reasonable |
| Precision |  |  |

## Visual Proportionality

- The proportionality is completely altered by the wrong use of data: pie/donut MUST be used to represent part-whole relationships only.
- Percentages close to $50 \%$ are not half donut as one would expect.
- Sectors representing 59\% and 42\% have almost the same size.


## Visual Proportionality

- Moreover, areas and angles are generally not perceived accurately.
- The educational levels are increasing (thus ordered): one would expect them to be encoded as color with increasing intensity, which is not the case.


## Visual Utility

- All elements in the graph convey useful information.
- One might argue the thick lines separating the sectors could be removed, though colors are very similar and removing the lines could introduce a clarity issue.
- The background color is not really useful but being uniform does not represent an issue.


## Visual Clarity

- The color coding relies on the implicit (for US citizens) associations of red to Republican party and blue to Democratic party.
- The labels are placed close and connected to the items they describe.


## Data structure

| Field | Dim./Measure | Description |
| :--- | :--- | :--- |
| EDUCATION_LEVEL | Dimension | Level of education of <br> respondent |
| PARTY_PREFERENCE | Dimension | Party preferred by <br> responded, either Dem <br> or Rep |
| PREFERENCES | Measure | Percentage of <br> respondent expressing <br> preference for that <br> party |

## Schema \#1

## Schema

Columns
Rows
Graph type
Color
Size
Label

## Details

SUM(PREFERENCES)
EDUCATION_LEVEL
Bar
PARTY_PREFERENCE
Default
Default

## Redesign \#1

Bar Chart Preferences Stacked
EDUCATION_LEVEL
Democratic
$\square$ Republican


Sum of PREFERENCES for each EDUCATION_LEVEL. Color shows details about PARTY_PREFERENCE. The marks are labeled by sum of PREFERENCES.

## Schema \#2



Columns
Rows

Graph type
Color
Size
Label

## Details

SUM(PREFERENCES)
EDUCATION_LEVEL, PARTY_PREFERENCE

## Bar

PARTY_PREFERENCE
Default
SUM(PREFERENCES)

## Redesign \#2

Bar Chart Preferences Sided


Sum of PREFERENCES for each PARTY_PREFERENCE broken down by EDUCATION_LEVEL. Color shows details about PARTY_PREFERENCE. The marks are labeled by sum of PREFERENCES.

## Schema \#3

## Schema

Columns
Rows
Graph type
Color
Size
Label

## Details

SUM(PREFERENCES)
EDUCATION_LEVEL

## Circle

PARTY_PREFERENCE
Default
SUM(PREFERENCES)

## Redesign \#3

## Dot Plot Preferences



College grad

47\%

Some college
36\% 40\%

-
$37 \% \quad 42 \%$
HS or less


Sum of PREFERENCES for each EDUCATION_LEVEL. Color shows details about PARTY_PREFERENCE. The marks are labeled by sum of PREFERENCES.

## Schema \#4



Columns

Rows
Graph type
Color
Size
Label

## Details

SUM(iif(PARTY_PREFERENCE=="De mocratic",1,-1)*[PREFERENCES]) EDUCATION_LEVEL Bar
PARTY_PREFERENCE
Default
SUM(PREFERENCES)

## Redesign \#4

Bar Chart Preferences Stacked Diverging
Education Level


SUM(iif([Party Preference]=="Democratic",1,-1)*[Preferences]) for each Education Level. Color shows details about Party Preference. The marks are labeled by sum of Preferences.

## Schema \#5

## Schema

Columns
Rows
Graph type
Color
Size
Label

## Details

## PARTY_PREFERENCE

## SUM(PREFERENCES)

Line
PARTY_PREFERENCE
Default
EDUCATION_LEVEL, PARTY_PREFERENCE

## Redesign \#5

Slope Chart Preferences


The trend of sum of Preferences for Party Preference. Color shows details about Party Preference. The marks are labeled by Education Level and sum of Preferences.

## Schema \#6

| Schema | Details |
| :--- | :--- |
| Columns | SUM(PREFERENCES) <br> (\% of total by row) |
| Rows | EDUCATION_LEVEL |
| Graph type | Bar |
| Color | PARTY_PREFERENCE |
| Size | Default |
| Label | Default |

## Redesign \#6

Bar Chart Preferences Proportion

\% of Total PREFERENCES for each EDUCATION_LEVEL. Color shows details about PARTY_PREFERENCE. Percents are based on each row of the table.

## Schema \#7

| Field | Dim./Measure | Description <br> iif(PARTY_PREFERENCE <br> $==$ "Democratic", 1,0) |
| :--- | :--- | :--- |
| DemIndicator | Measure | sum(PREFERENCES*[De <br> mIndicator])- <br> sum(PREFERENCES*(1- <br> [DemIndicator])) |
| DemLean | Measure | Details |
| Schema |  | DemLean |
| Columns | EDUCATION_LEVEL |  |
| Rows | Circle |  |
| Graph type | PARTY_PREFERENCE |  |
| Color | Default |  |
| Size | DemLean |  |
| Label |  |  |

## Redesign \#7

Dot plot Dem Lean


Dem Lean for each EDUCATION_LEVEL. Color shows details about FavorDem. The marks are labeled by Dem Lean.

## Schema \#8

| Field | Dim./Measure | Description |
| :--- | :--- | :--- |
| DemPref | Measure | DemIndicator *[Preferences] |
| RepPref | Measure | (1-DemIndicator) *[Preferences] |
| Undecided | Measure | 1-SUM(PREFERENCES*[DemIndicator])- <br> SUM(PREFERENCES*(1-[DemIndicator])) |



Columns

Rows
Graph type
Color
Size
Label

## Details

SUM(DemPref), SUM(Undecided), SUM(RepPref)
EDUCATION_LEVEL
Bar
Measure Names
Default
Default

## Redesign \#8

Bar Chart Preferences Stacked w/Undecided


Democratic, Undecided and Republican for each EDUCATION_LEVEL. Color shows details about Democratic, Undecided and Republican. The marks are labeled by Democratic, Undecided and Republican.

## EXAM OF 2020-02-14

## Visualization

## 1103 Physician Suicides By Specialty

* Active Physicians Based On 2016 AAMC Physician Specialty Data Report


Source: https://www.idealmedicalcare.org/1103-doctor-suicides-13-reasons-why/

Analyze the above graph that was published on a medical blog in 2018.

## Question

- The question is clearly defined: what is the incidence of suicides among different medical specialties?


## Data quality

- Accuracy: Partly, number of Surg and IM are too round to be accurate.
- Completeness: Yes, we assume all specialties are reported.
- Consistency: No, the sum of suicides is not 1103 as reported in the title; number of physicians is from 2016, but suicides are presumably on a wider time frame.


## Data quality

- Currency: Partly, data is from 2018 (2016 the active).
- Credibility: Yes, data seem to come from trustable sources.
- Understandability: No, the length of the bar encodes neither the suicide cases nor the numner of active physicians. The value is the suicide rate (suicides/physicians).


## Data quality

- Precision: Yes, precision seems reasonable for the purpose.


## Visual Proportionality

- Assuming the encoded values are the ratios, apparently the representation is proportional.


## Visual Utility

- The gradient background is not useful.
- The strongly bright colors are not useful and may lead to sight fatigue.
- The vertical lines are not much useful without a proper axis.


## Visual Clarity

- The data reported in each bar might support the comprehension. Though we miss the value that is encoded in the bar lengths: it is not immediate to compute the ratios and compare them (e.g. 50 / 39579 ).
- The note above the graph is clearly misleading because it suggest the reciprocal of the rate is used.


## Data structure

| Field | Dim./Measure | Description |
| :--- | :--- | :--- |
| SPECIALITY | Dimension | Different medical <br> specialties |
| ACTIVE_PHYSICIANS | Measure | Number of active <br> physicians in that <br> specialty |
| SUICIDES | Measure | Number of suicides <br> among physicians in <br> that specialty |

## Schema \#1

| Schema | Details |
| :--- | :--- |
| Columns | SUM(SUICIDES) / <br> SUM(ACTIVE_PHYSICIANS) * 10000 |
| Rows | SPECIALITY |
| Graph type | Bar |
| Color | Default |
| Size | Default |
| Label | SUM(SUICIDES) / <br> SUM(ACTIVE_PHYSICIANS) * |

## Redesign \#1



Sum of Suicide Rate [per 10 thousand physicians] for each Speciality. The marks are labeled by sum of Suicide Rate [per 10 thousand physicians].

## Schema \#2

| Schema | Details |
| :--- | :--- |
| Columns | SUM(SUICIDES) / <br> SUM(ACTIVE_PHYSICIANS) |
| Rows | SPECIALITY |
| Graph type | Bar |
| Color | Default |
| Size | Default |
| Label | SUM(SUICIDES) / <br> SUM(ACTIVE_PHYSICIANS) |

## Redesign \#2



Sum of Suicide Rate for each Speciality. The marks are labeled by sum of Suicide Rate.

## Schema \#3

| Schema | Details |
| :--- | :--- |
| Columns | SUM(SUICIDES) / <br> SUM(ACTIVE_PHYSICIANS) |
| Rows | SPECIALITY |
| Graph type | Shape |
| Color | Default |
| Size | Default |
| Label | SUM(SUICIDES) / <br> SUM(ACTIVE_PHYSICIANS) |

## Redesign \#3



Sum of Suicide Rate for each Speciality. The marks are labeled by sum of Suicide Rate.

## Schema \#4

Columns

Rows
Graph type
Color
Size
Label
I

##  <br> Schema

## Details

## SUM(SUICIDES) /

SUM(ACTIVE_PHYSICIANS)
Colums
?
(

## Redesign \#4

## DotStrip - Rate

| 0.61\% |  | $7 \%$ | 1.05\% | 1.26\% | 1.46\% |
| :---: | :---: | :---: | :---: | :---: | :---: |
| IM |  | ds | Psych | EM | Surg |
| $\bigcirc$ |  | 0 | $\bigcirc$ | $\bigcirc$ | 0 |
|  | 0.73\% | 0.83 | 1.11 |  |  |
|  | Peds | FM | OB/G |  |  |

$3.34 \%$
Anesth
○
$2.00 \%$
$2.50 \%$
$3.00 \%$

Sum of Suicide Rate. The marks are labeled by sum of Suicide Rate and Speciality.

## EXAM OF 2020-06-18

## Visualization

## COVID-19 cases and deaths by race in Michigan

American Indian or Alaska Native $\square$ Asian/Pacific Islander $\square$ Ultack or African American
White $\square$ Unknown

Percentage of Overall Cases by Race Percentage of Deceased Cases by Race


Source: Michigan Department of Health and Human Services

## Analysis

Analyze the above graph comparing the number of COVID-19 cases and deaths by race in Michigan.

## Question

- The question is clearly defined and it deals with the number of cases of COVID-19 compared with the number of deaths of COVID-19 by ethnicity.


## Data quality

- Accuracy: It is impossible to evaluate because the data is not available.
- Completeness: The data is not complete at all, as it is missing.
- Consistency: It is not clear the difference among "multiple races", "other", and "unknown". The "unknown" slice could report inconsistent values between the cases and the deaths.


## Data quality

- Currency: Obviously, data is from the first half of 2020, but there is no information about the currency of the data.
- Credibility: The source seems reliable.


## Data quality

- Understandability: We do not know how data were measured.
- Precision: Precision is not appropriate because the data is not available.


## Visual Proportionality

- We cannot say because the data is not available. Perceptual proportionality of arcs and areas is usually problematic.


## Visual Utility

- The text "made with..." is not useful.
- Double labels can be removed, it is better to use direct labeling.
- The background color is useless.


## Visual Clarity

- Some labels on the pie chart cannot be read ("American Indian" and "Asian").


## Data structure

| Field | Dim./Measure | Description |
| :--- | :--- | :--- |
| RACE | Dimension | Ethnicity of the person <br> that was affected by <br> COVID-19 |
| CASES | Measure | Percentage of overall <br> cases by ethnicity |
| DEATHS | Measure | Percentage of <br> deceased cases by <br> ethnicity |

## Schema \#1

## Schema

Columns
Rows
Graph type
Color
Size
Label

## Details

Measure Names
Measure Values
Line
Race
Default
Race

## Redesign \#1

Slope


0\%

Ethnicity
American Indian

- Asian/Pacific
- Black

Multiple
Other
$\square$ Unknown
$\square$ White

## Schema \#2



Columns
Rows
Graph type
Color
Size
Label

## Details

SUM(Cases), SUM(Deaths)
Race
Bar
Race
Default
SUM(Cases), SUM(Deaths)

## Redesign \#2

Diverging Barchart


## Schema \#3

## Schema

Columns
Rows
Graph type
Color
Size
Label

## Details

Measure Values, Measure Values
Race
Line, Shape
Measure Names
Default
Measure Values

## Redesign \#3

Dumbbell


## Schema \#4



Columns
Rows
Graph type
Color
Size
Label

## Details

Race, Measure Names
Measure Values
Bar
Measure Names
Default
Measure Values

## Redesign \#4



## Schema \#5

| Schema | Details |
| :--- | :--- |
| Columns | Measure Names |
| Rows | - |
| Graph type | Pie |
| Color | Race |
| Size | Measure Values |
| Label | Race |

## Redesign \#5

Pies

> Cases

Deaths
Measure Values

Ethnicity
American Indian
Asian/Pacific
Black
Multiple
Other
Unknown
$\square$ White

## Schema \#6



Columns
Rows
Graph type
Color
Size
Label

## Details

SUM(Cases)
SUM(Deaths)
Shape
Race
Default
Race

## Redesign \#6

Scatter

Ethnicity

- Asian/Pacific
$\square$ Black
Multiple
$\square$ Other
U Unknown
$\square$ White


## OAmerican Indian

 OAsian/Pacific

## Schema \#7



Columns
Rows
Graph type
Color
Size
Label

## Details

SUM([Deaths]/[Cases])
Race
Bar
Default
Default
Default

## Redesign \#7

Bar Chart Death ratio
Ethnicity


## EXAM OF 2020-09-11

## Visualization



Analyze the above graph comparing the frequency of use of voice assistants (e.g: Alexa, Siri...) by request type.

## Question

- What is the relation between the frequency of use of voice assistants and the (popularity of | most popular | most asked) category of the request?


## Data quality

- Accuracy: data are comparable, and the values are reasonable.
- Completeness: data are complete, several categories are reported.
- Consistency: the percentages of some frequencies are probably overlapped; they cannot be summed.


## Data quality

- Currency: data are referred to the year 2018, so it is reasonably up to date.
- Credibility: the source is reported, and it seems trusted.
- Understandability: data are understandable, but it is better to report absolute numbers instead of percentages.


## Data quality

- Precision: precision is up to the first decimal digit and it is appropriate.


## Visual Proportionality

- The bars are proportional to the associated values. The total bar is proportional to the sum of the percentages, but they cannot be summed because the frequencies are overlapped.


## Visual Utility

- Almost all visual elements are useful, but the bar at the top-left and the legend "Image 2:".


## Visual Clarity

- The second and the third type of bars are difficult to compare, because they are not aligned.
- Colors are too bright.
- The legend is difficult to read via color-codes.


## Data structure

| Field | Dim./Measure | Description |
| :--- | :--- | :--- |
| USE_DAILY | Measure | Percentage of the daily <br> use |
| USE_MONTHLY | Measure | Percentage of the <br> monthly use |
| TRIED_ONCE | Measure | Percentage of used at <br> least once |
| REQUEST_TYPE | Dimension | The different <br> categories of requests |

## Schema \#1

## Schema

## Columns

## Rows

Graph type
Color
Size

## Details

SUM(USE_DAILY), SUM(USE_MONTHLY), SUM(TRIED_ONCE)
REQUEST_TYPE
Bar
Three different colors, one for each use
Default
SUM(USE_DAILY), SUM(USE_MONTHLY), SUM(TRIED_ONCE)

## Redesign \#1

Uses of voice assistants and frequency of use


## Schema \#2

## Schema

## Columns

Rows
Graph type
Color
Size

## Details

SUM(USE_DAILY), SUM(USE_MONTHLY), SUM(TRIED_ONCE)
REQUEST_TYPE
Circle
Three different colors, one for each use
Default
SUM(USE_DAILY), SUM(USE_MONTHLY), SUM(TRIED_ONCE)

## Redesign \#2

Uses of voice assistants and frequency of use


