

Past Exams

Data Management and Visualization



SoftEng
<http://softeng.polito.it>

Version 1.0.4
© Diego Monti, 2022








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:

-  **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.

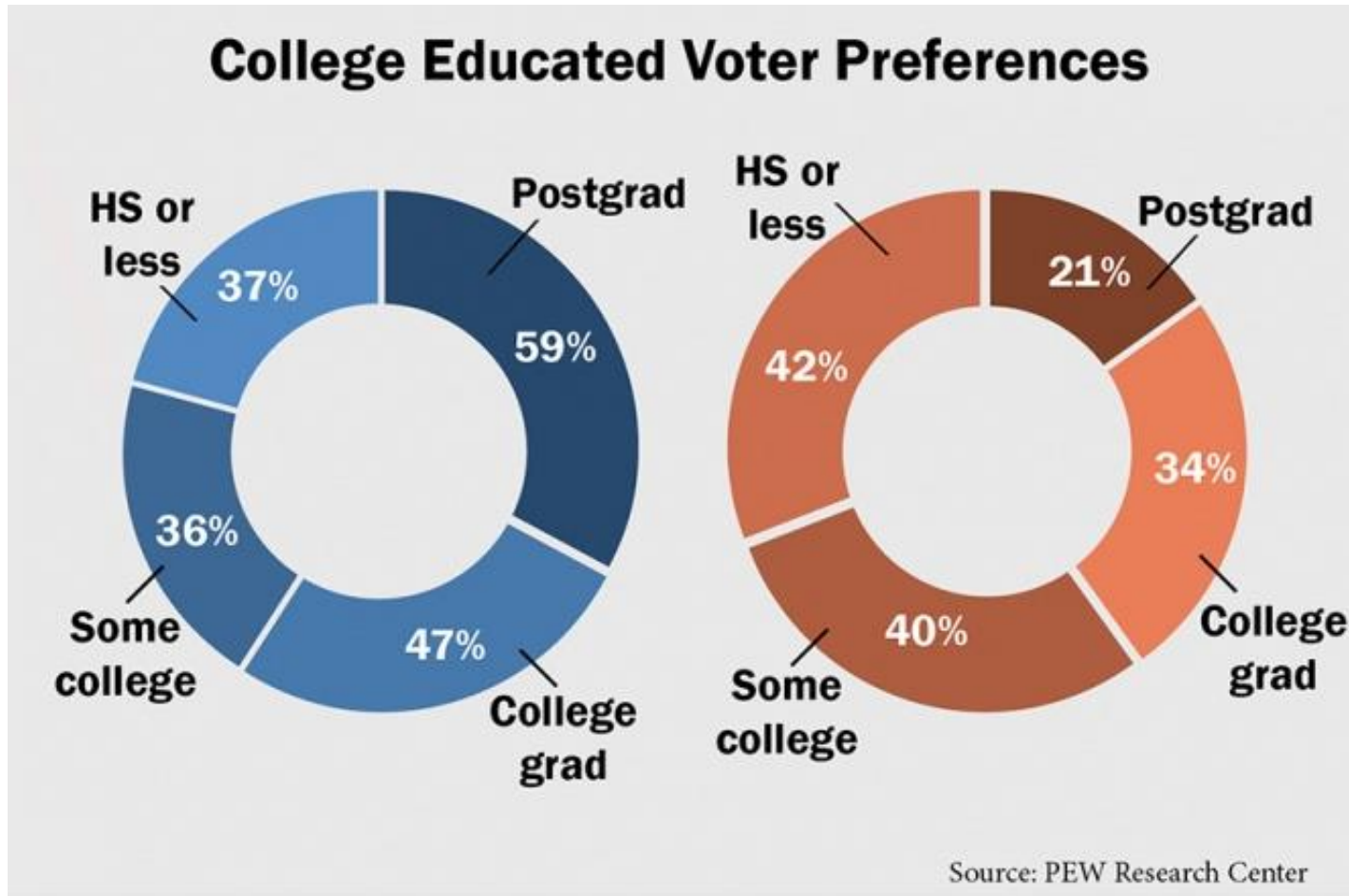
Your fair use and other rights are in no way affected by the above.

EXAM OF 2020-01-31



SoftEng
<http://softeng.polito.it>

Visualization



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 "College-educated 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 interpret the missing 20%
Consistency	Partly	Sum of percentages not equal to 100%
Currency	Yes	Presumably, data is from 2016, but that is ok
Credibility	Yes	Source is reported as a polling company
Understandability	Yes	Data is easy to understand
Precision	Yes	Single % point precision is reasonable

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.

Visual Proportionality

- Sectors representing 59% and 42% have almost the same size.
- Moreover, areas and angles are generally not perceived accurately.

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 associations of red to Republican party and blue to Democratic party.
- The labels are placed close and connected to the items they describe.
- The educational levels are increasing (thus ordered): one would expect them to be encoded as color with increasing intensity, which is not the case.

Data structure

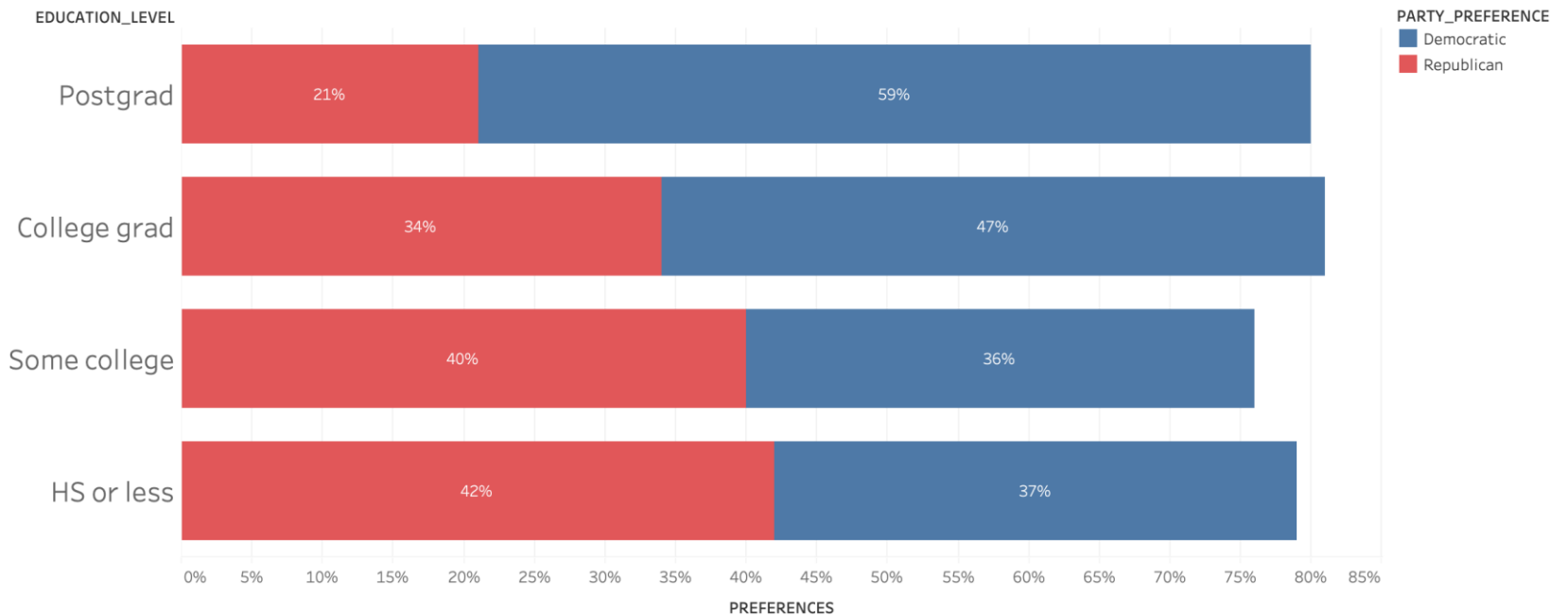
Field	Dim./Measure	Description
EDUCATION_LEVEL	Dimension	Level of education of respondent
PARTY_PREFERENCE	Dimension	Party preferred by responded, either Dem or Rep
PREFERENCES	Measure	Percentage of respondent expressing preference for that party

Schema #1

Schema	Details
Columns	SUM(PREFERENCES)
Rows	EDUCATION_LEVEL
Graph type	Bar
Color	PARTY_PREFERENCE
Size	Default
Label	Default

Redesign #1

Bar Chart Preferences Stacked



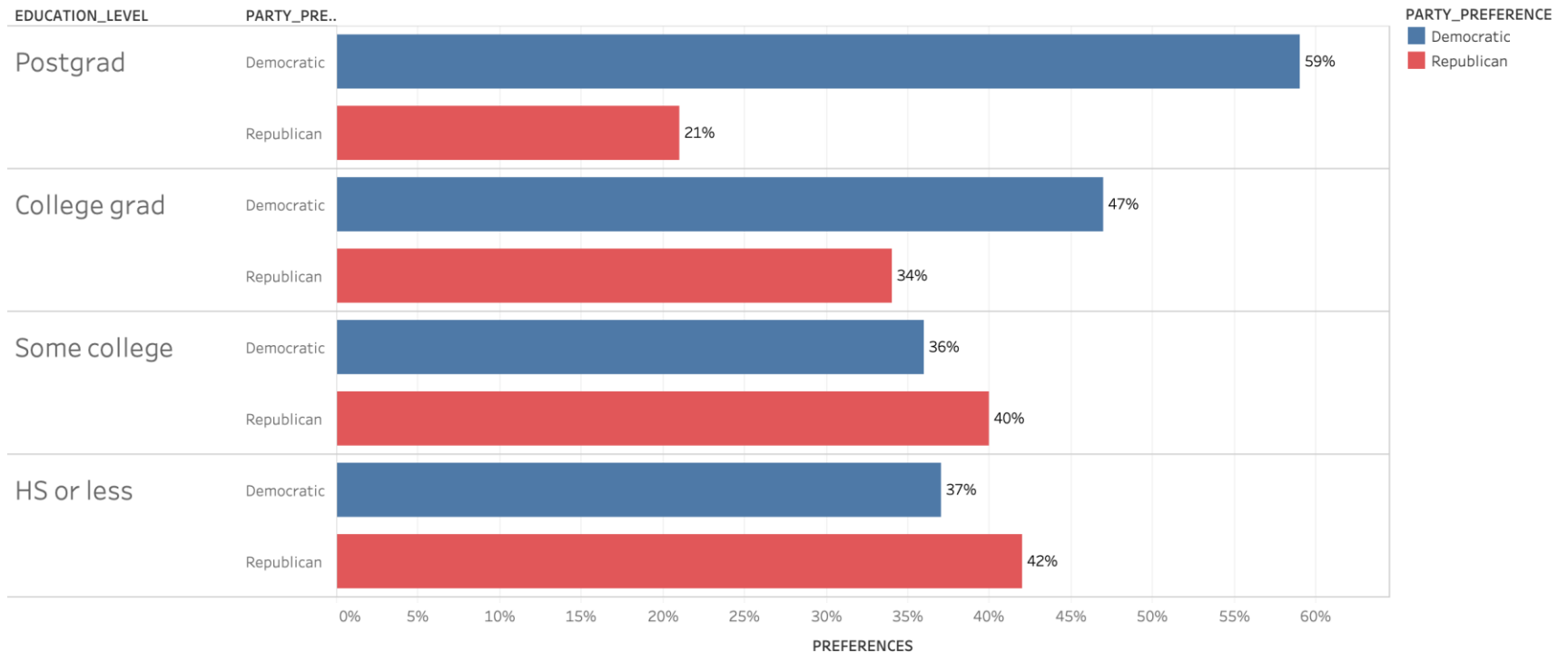
Sum of PREFERENCES for each EDUCATION_LEVEL . Color shows details about PARTY_PREFERENCE. The marks are labeled by sum of PREFERENCES.

Schema #2

Schema	Details
Columns	SUM(PREFERENCES)
Rows	EDUCATION_LEVEL, PARTY_PREFERENCE
Graph type	Bar
Color	PARTY_PREFERENCE
Size	Default
Label	SUM(PREFERENCES)

Redesign #2

Bar Chart Preferences Sided



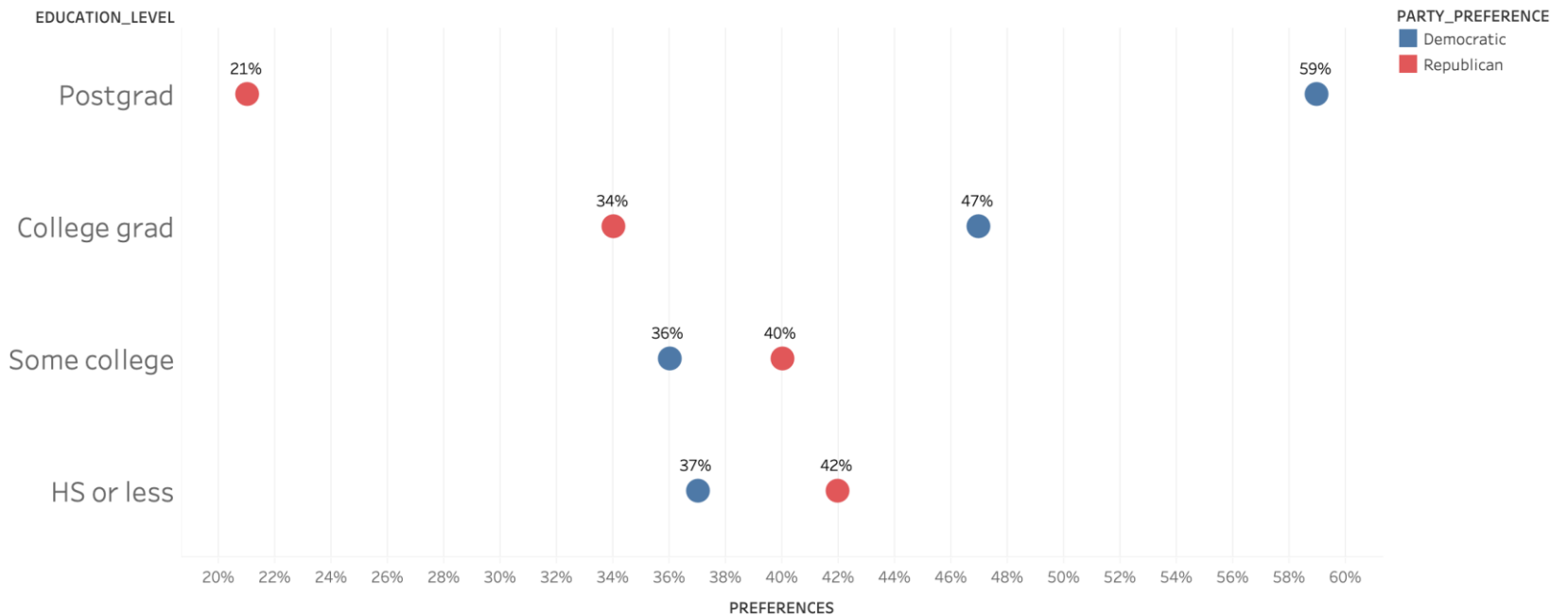
Sum of PREFERENCES for each PARTY_PREFERENCE broken down by EDUCATION_LEVEL . Color shows details about PARTY_PREFERENCE. The marks are labeled by sum of PREFERENCES.

Schema #3

Schema	Details
Columns	SUM(PREFERENCES)
Rows	EDUCATION_LEVEL
Graph type	Circle
Color	PARTY_PREFERENCE
Size	Default
Label	SUM(PREFERENCES)

Redesign #3

Dot Plot Preferences



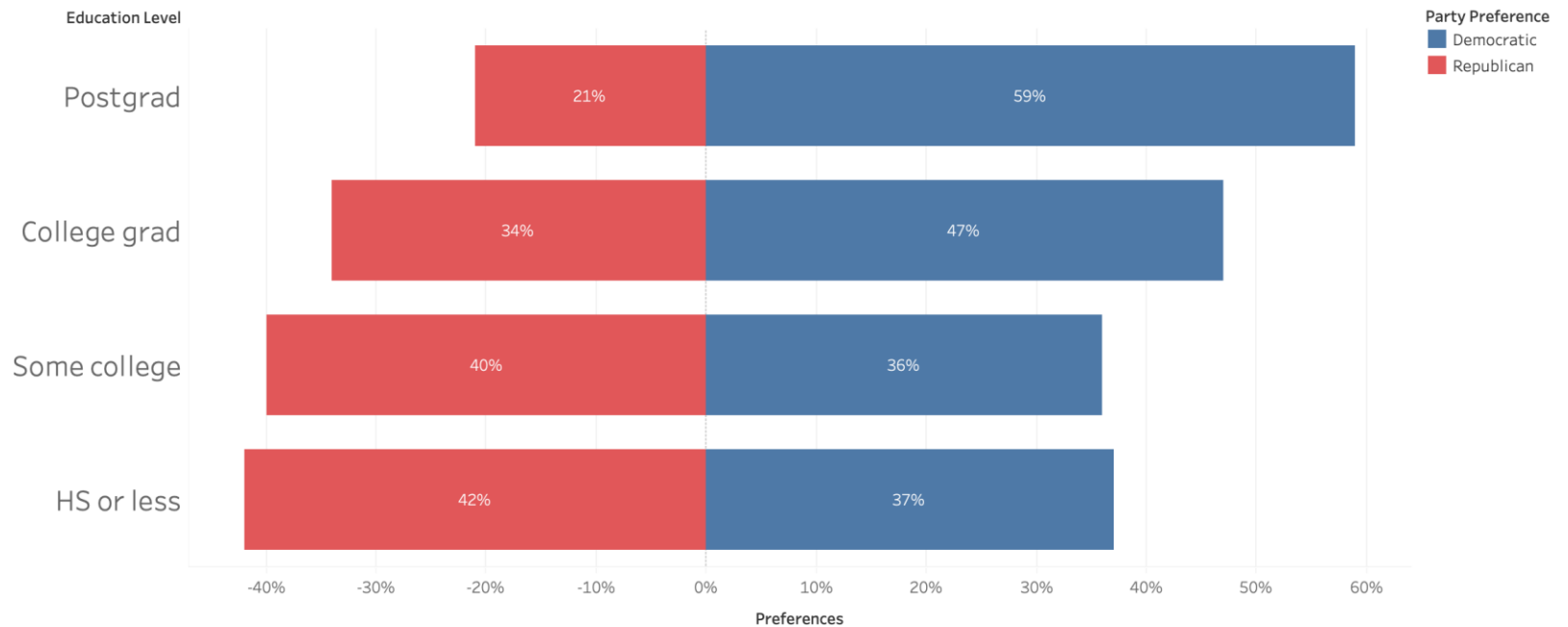
Sum of PREFERENCES for each EDUCATION_LEVEL . Color shows details about PARTY_PREFERENCE. The marks are labeled by sum of PREFERENCES.

Schema #4

Schema	Details
Columns	SUM(iif(PARTY_PREFERENCE=="Democratic",1,-1)*[PREFERENCES])
Rows	EDUCATION_LEVEL
Graph type	Bar
Color	PARTY_PREFERENCE
Size	Default
Label	SUM(PREFERENCES)

Redesign #4

Bar Chart Preferences Stacked Diverging



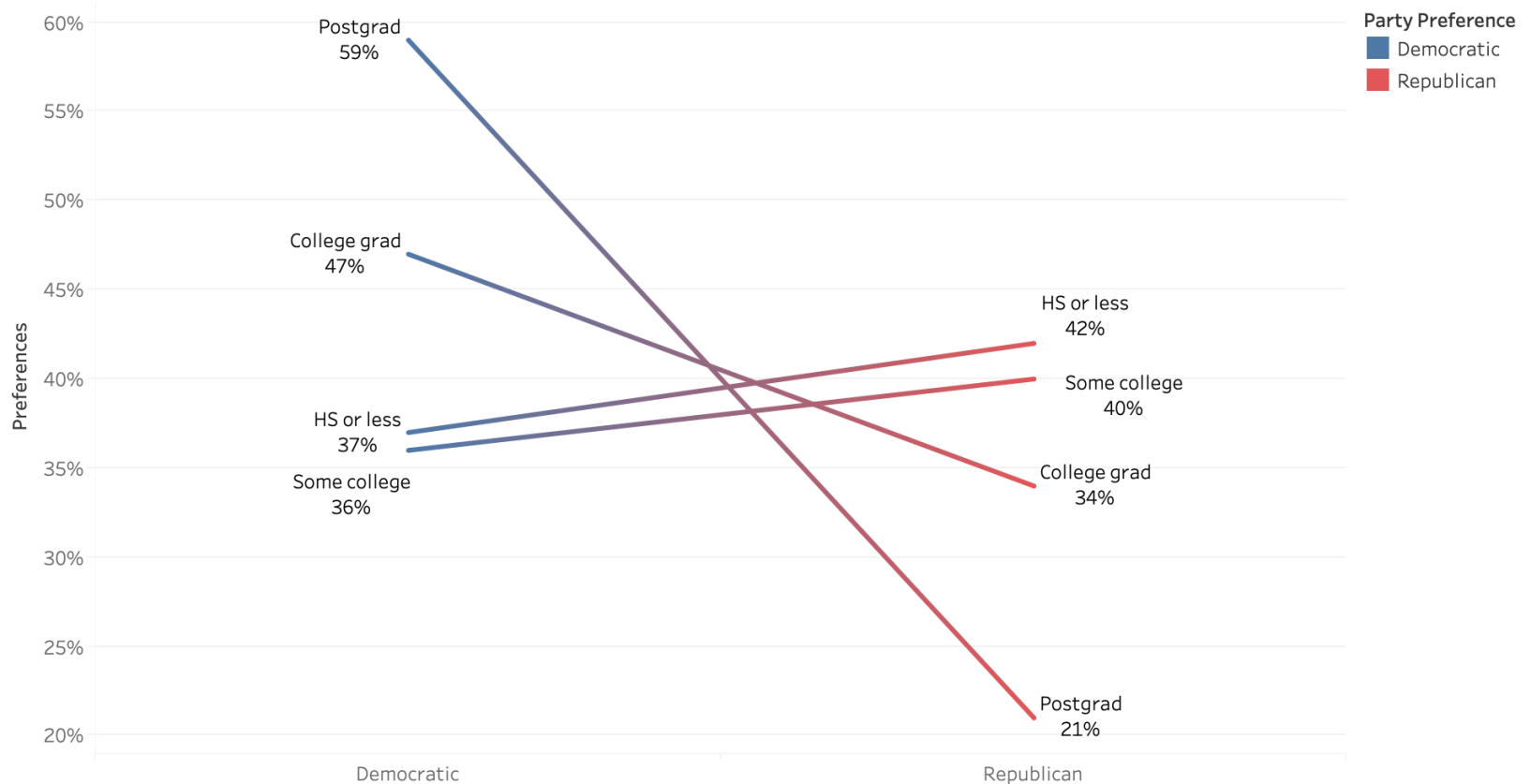
$\text{SUM}(\text{iif}([\text{Party Preference}] = \text{"Democratic"}, 1, -1) * [\text{Preferences}])$ for each Education Level. Color shows details about Party Preference. The marks are labeled by sum of Preferences.

Schema #5

Schema	Details
Columns	PARTY_PREFERENCE
Rows	SUM(PREFERENCES)
Graph type	Line
Color	PARTY_PREFERENCE
Size	Default
Label	EDUCATION_LEVEL, SUM(PREFERENCES)

Redesign #5

Slope Chart Preferences



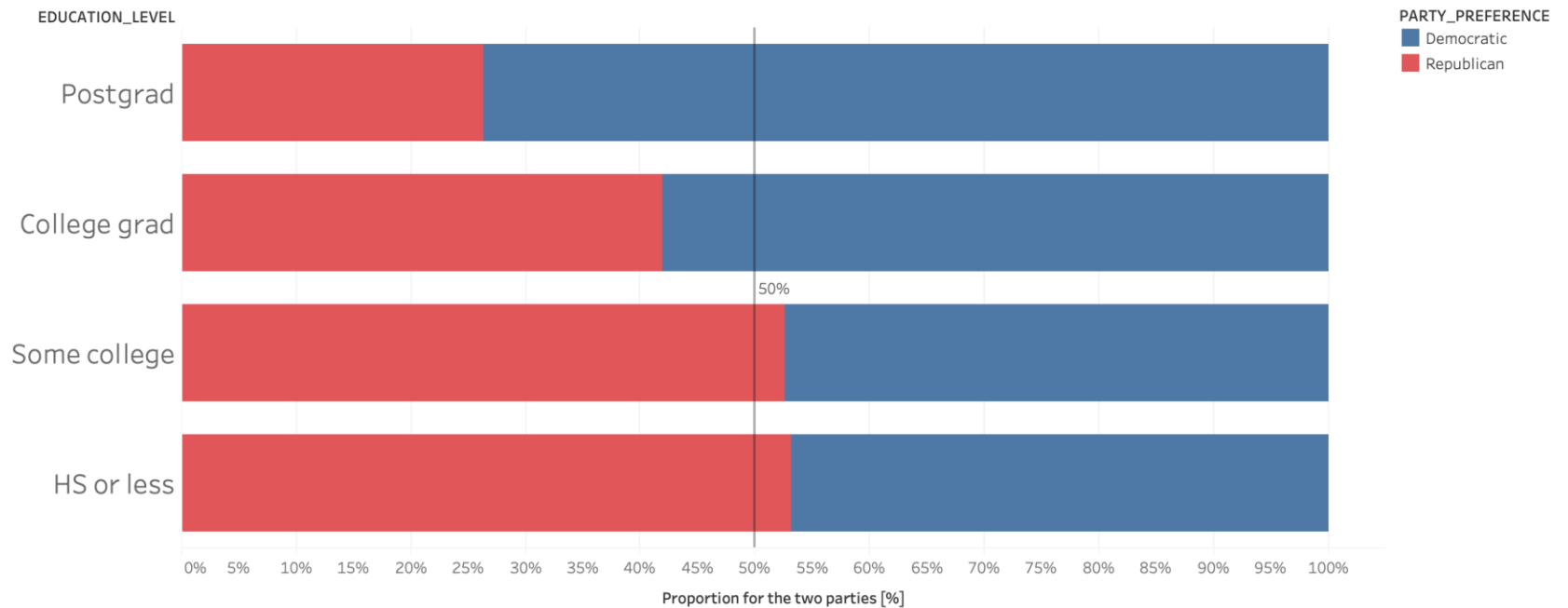
The trend of sum of Preferences for Party Preference. Color shows details about Party Preference. The marks are labeled by Education Level and sum of Preferences.

Schema #6

Schema	Details
Columns	SUM(PREFERENCES) (% 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.

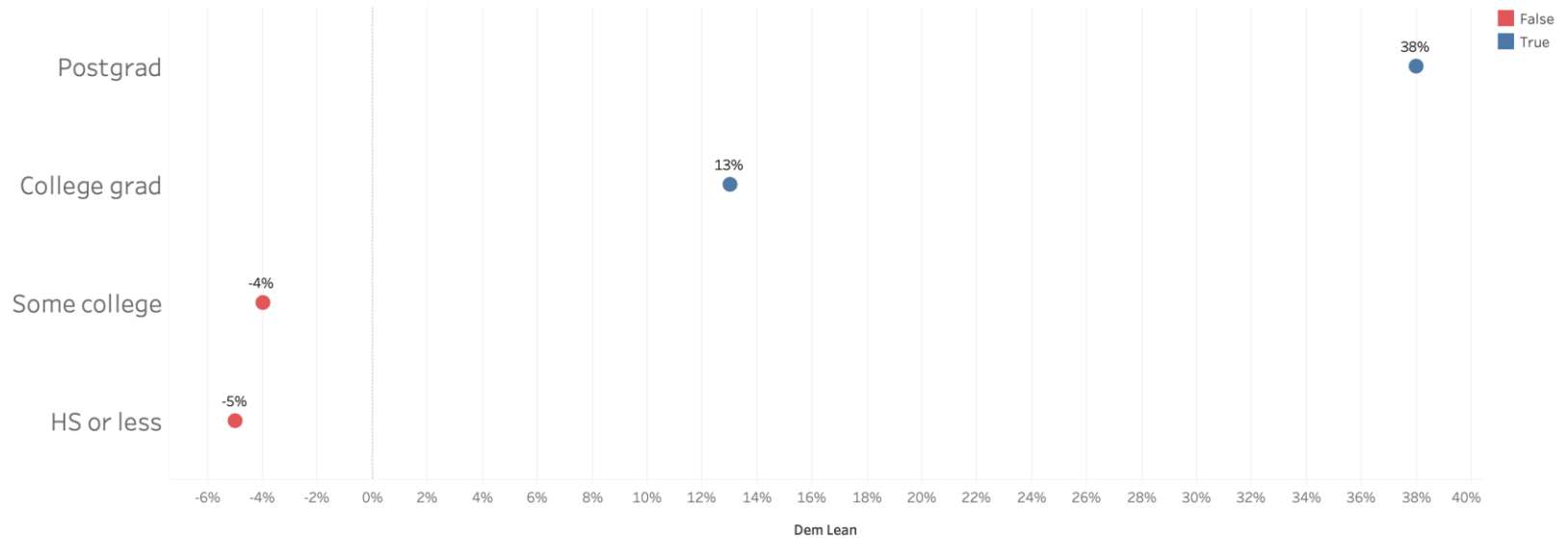
Schema #7

Field	Dim./Measure	Description
DemIndicator	Measure	$\text{iif}(\text{PARTY_PREFERENCE} = \text{"Democratic"}, 1, 0)$
DemLean	Measure	$\text{sum}(\text{PREFERENCES} * [\text{DemIndicator}]) - \text{sum}(\text{PREFERENCES} * (1 - [\text{DemIndicator}]))$

Schema	Details
Columns	DemLean
Rows	EDUCATION_LEVEL
Graph type	Circle
Color	PARTY_PREFERENCE
Size	Default
Label	DemLean

Redesign #7

Dot plot Dem Lean



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

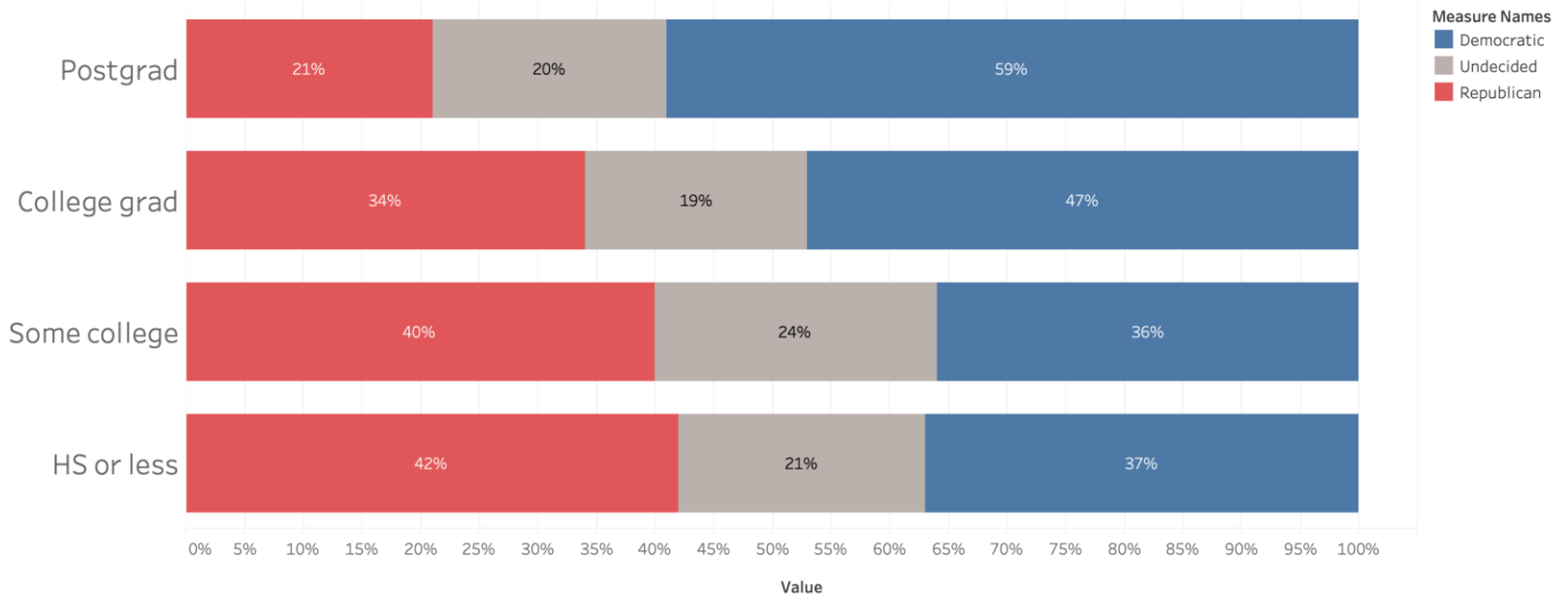
Schema #8

Field	Dim./Measure	Description
DemPref	Measure	$\text{DemIndicator} * [\text{Preferences}]$
RepPref	Measure	$(1 - \text{DemIndicator}) * [\text{Preferences}]$
Undecided	Measure	$1 - \text{SUM}(\text{PREFERENCES} * [\text{DemIndicator}]) - \text{SUM}(\text{PREFERENCES} * (1 - [\text{DemIndicator}]))$

Schema	Details
Columns	$\text{SUM}(\text{DemPref}), \text{SUM}(\text{Undecided}), \text{SUM}(\text{RepPref})$
Rows	EDUCATION_LEVEL
Graph type	Bar
Color	Measure Names
Size	Default
Label	Default

Redesign #8

Bar Chart Preferences Stacked w/Undecided



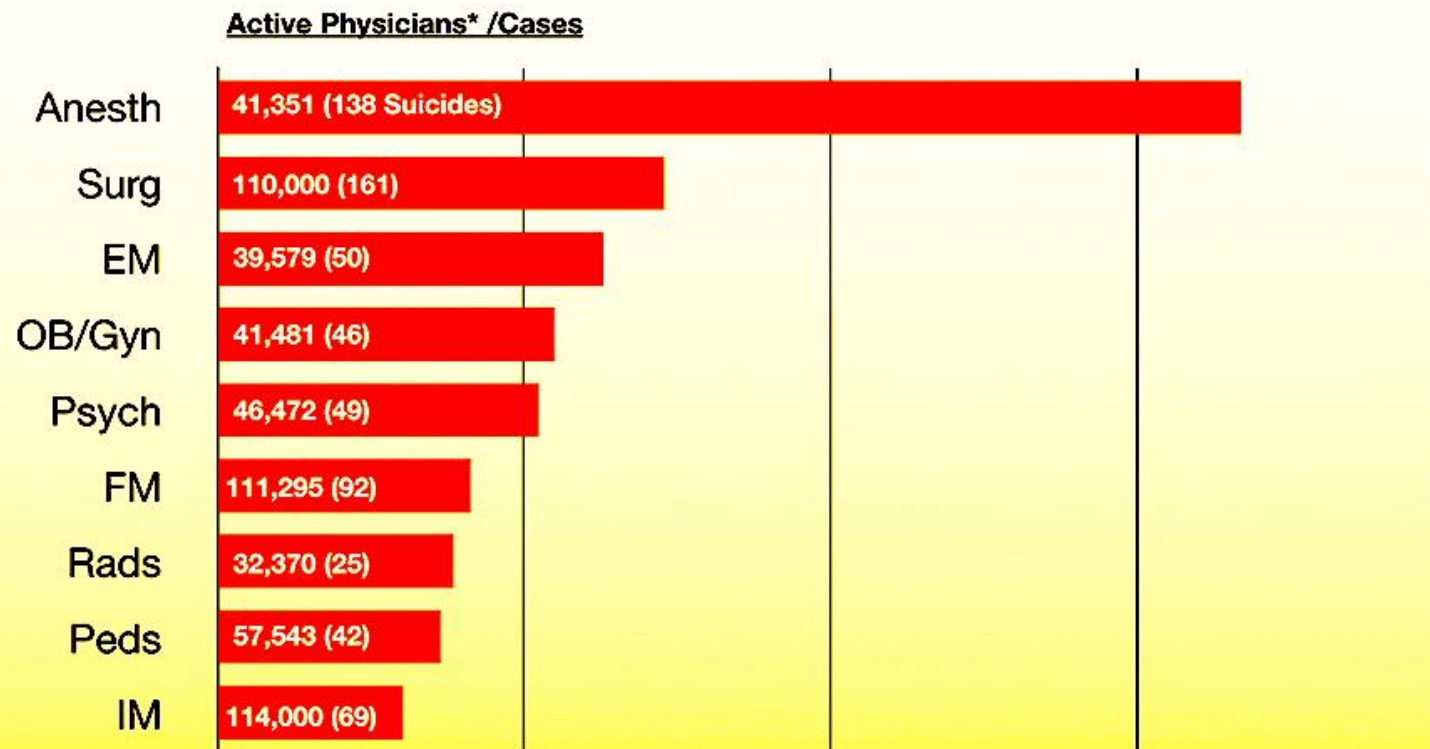
Democratic, Undecided and Republican for each EDUCATION_LEVEL . Color shows details about Democratic, Undecided and Republican. The marks are labeled by Democratic, Undecided and Republican.

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/>

Analysis

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 number 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

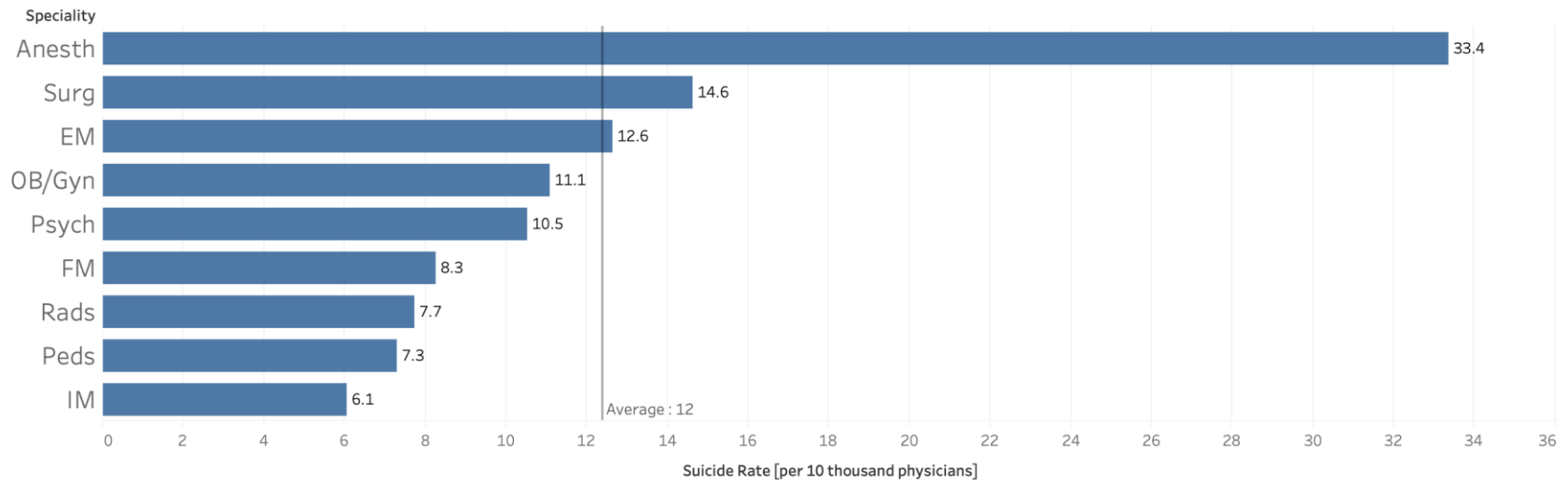
Field	Dim./Measure	Description
SPECIALITY	Dimension	Different medical specialties
ACTIVE_PHYSICIANS	Measure	Number of active physicians in that specialty
SUICIDES	Measure	Number of suicides among physicians in that specialty

Schema #1

Schema	Details
Columns	$\text{SUM}(\text{SUICIDES}) / \text{SUM}(\text{ACTIVE_PHYSICIANS}) * 10000$
Rows	SPECIALITY
Graph type	Bar
Color	Default
Size	Default
Label	$\text{SUM}(\text{SUICIDES}) / \text{SUM}(\text{ACTIVE_PHYSICIANS}) * 10000$

Redesign #1

Bar - Rate PTT



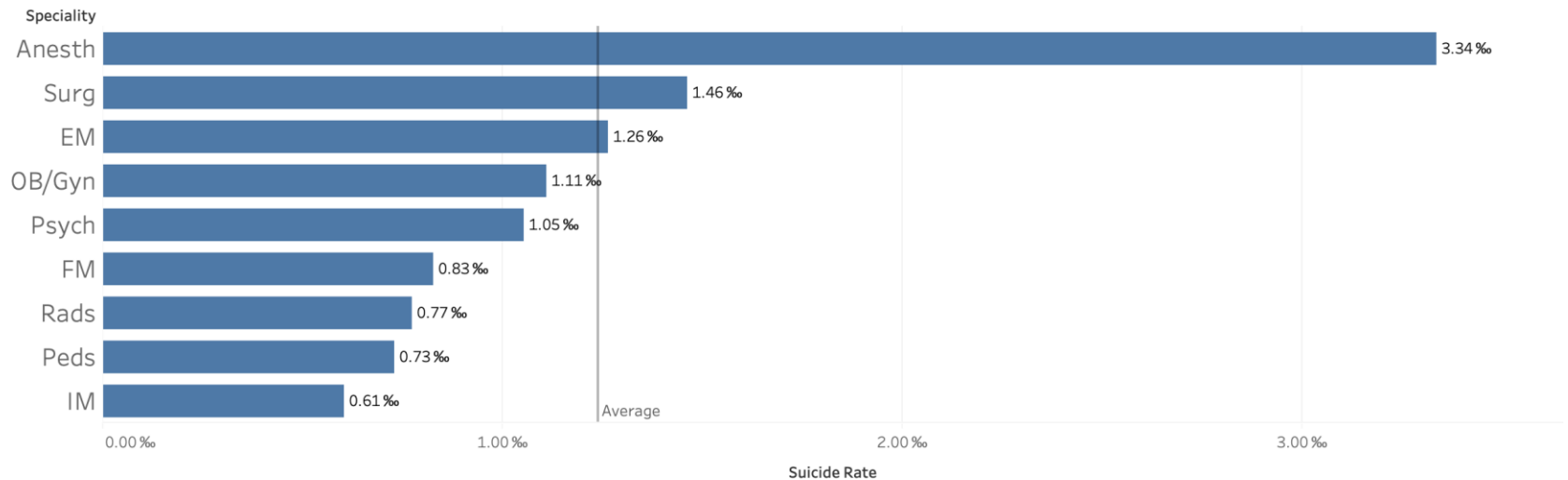
Sum of Suicide Rate [per 10 thousand physicians] for each Speciality. The marks are labeled by sum of Suicide Rate [per 10 thousand physicians].

Schema #2

Schema	Details
Columns	SUM(SUICIDES) / SUM(ACTIVE_PHYSICIANS)
Rows	SPECIALITY
Graph type	Bar
Color	Default
Size	Default
Label	SUM(SUICIDES) / SUM(ACTIVE_PHYSICIANS)

Redesign #2

Bar - Rate



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

Schema #3

Schema	Details
Columns	SUM(SUICIDES) / SUM(ACTIVE_PHYSICIANS)
Rows	SPECIALITY
Graph type	Shape
Color	Default
Size	Default
Label	SUM(SUICIDES) / SUM(ACTIVE_PHYSICIANS)

Redesign #3

Dots - Rate

Speciality

Anesth

Surg

EM

OB/Gyn

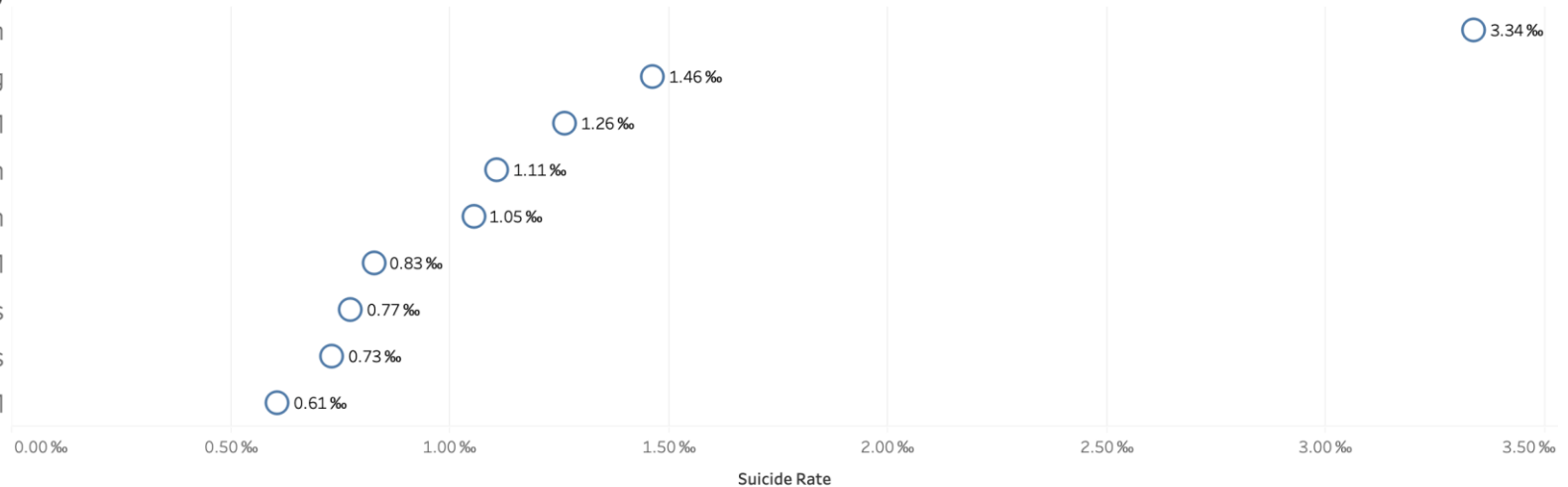
Psych

FM

Rads

Peds

IM



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

Schema #4

Schema	Details
Columns	SUM(SUICIDES) / SUM(ACTIVE_PHYSICIANS)
Rows	–
Graph type	Shape
Color	Default
Size	Default
Label	SUM(SUICIDES) / SUM(ACTIVE_PHYSICIANS), SPECIALITY

Redesign #4

DotStrip - Rate



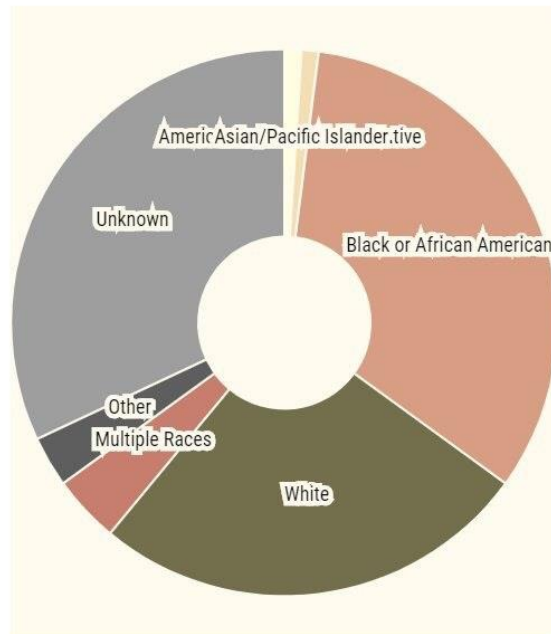
EXAM OF 2020-06-18

Visualization

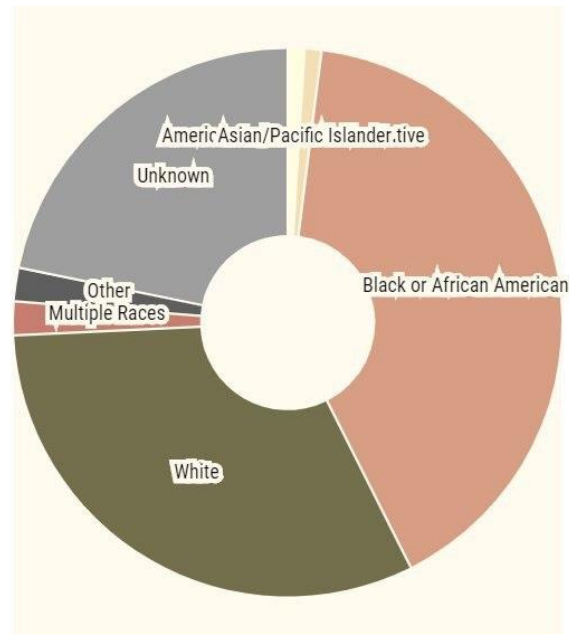
COVID-19 cases and deaths by race in Michigan

American Indian or Alaska Native Asian/Pacific Islander Black or African American
White Multiple Races Other Unknown

Percentage of Overall Cases by Race



Percentage of Deceased Cases by Race



Source: Michigan Department of Health and Human Services

Made with Flourish

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").

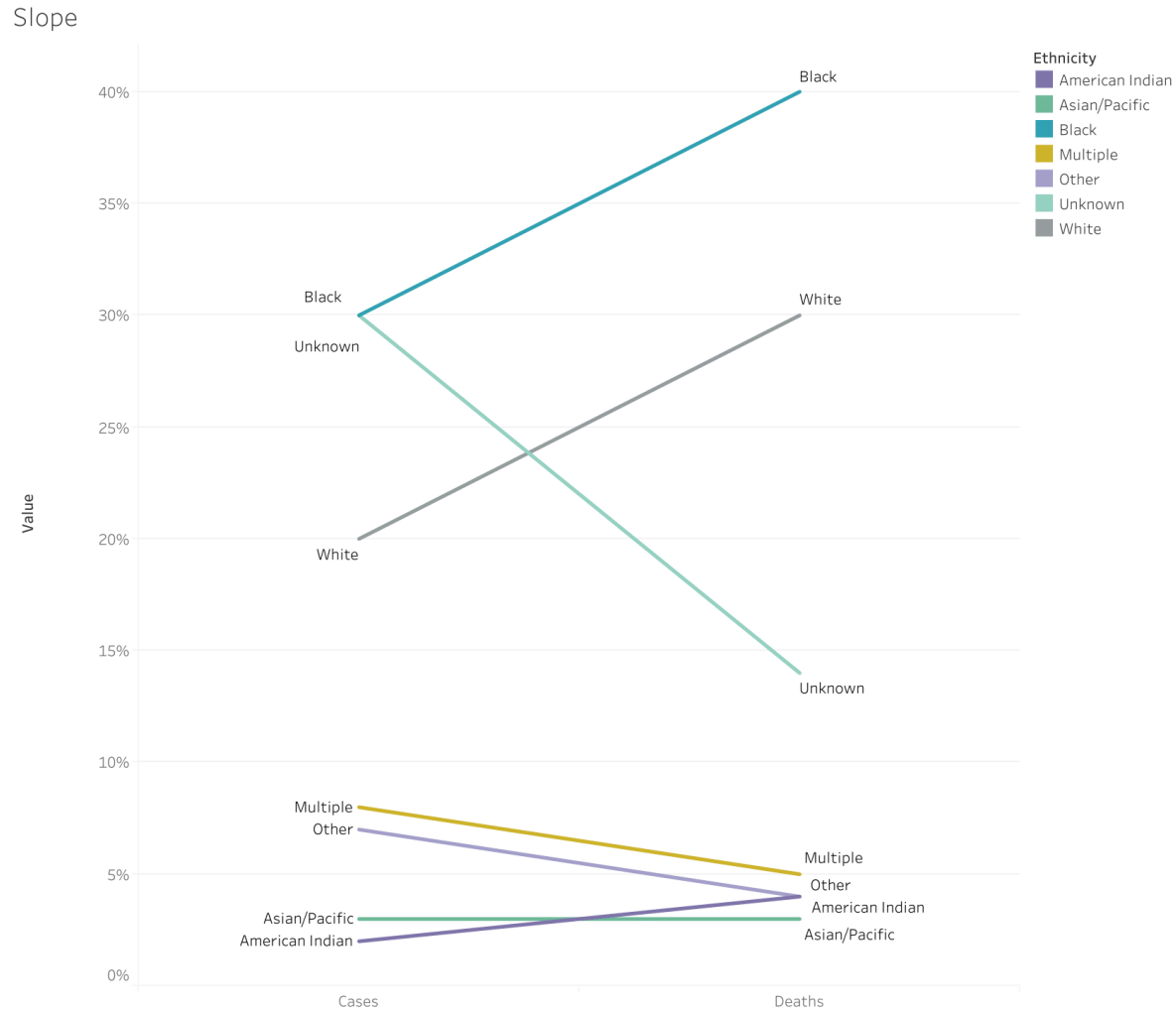
Data structure

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

Schema #1

Schema	Details
Columns	Measure Names
Rows	Measure Values
Graph type	Line
Color	Race
Size	Default
Label	Race

Redesign #1

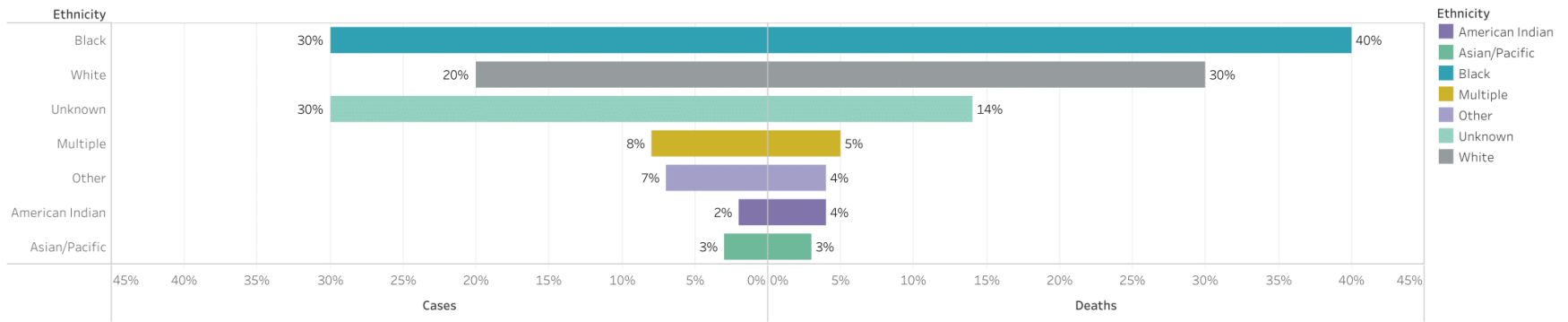


Schema #2

Schema	Details
Columns	SUM(Cases), SUM(Deaths)
Rows	Race
Graph type	Bar
Color	Race
Size	Default
Label	SUM(Cases), SUM(Deaths)

Redesign #2

Diverging Barchart

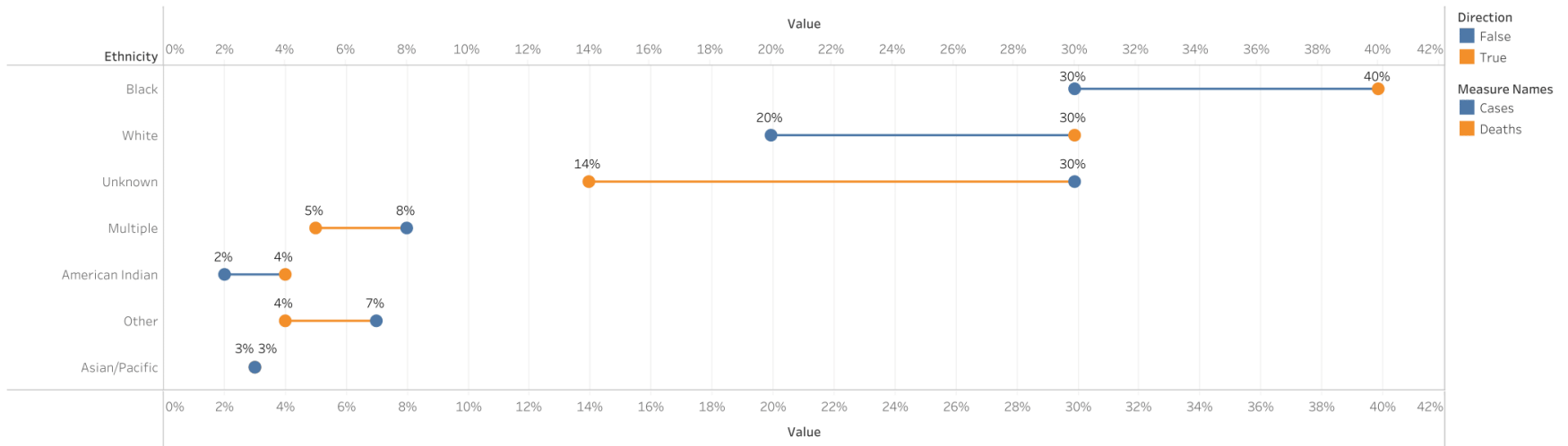


Schema #3

Schema	Details
Columns	Measure Values, Measure Values
Rows	Race
Graph type	Line, Shape
Color	Measure Names
Size	Default
Label	Measure Values

Redesign #3

Dumbbell

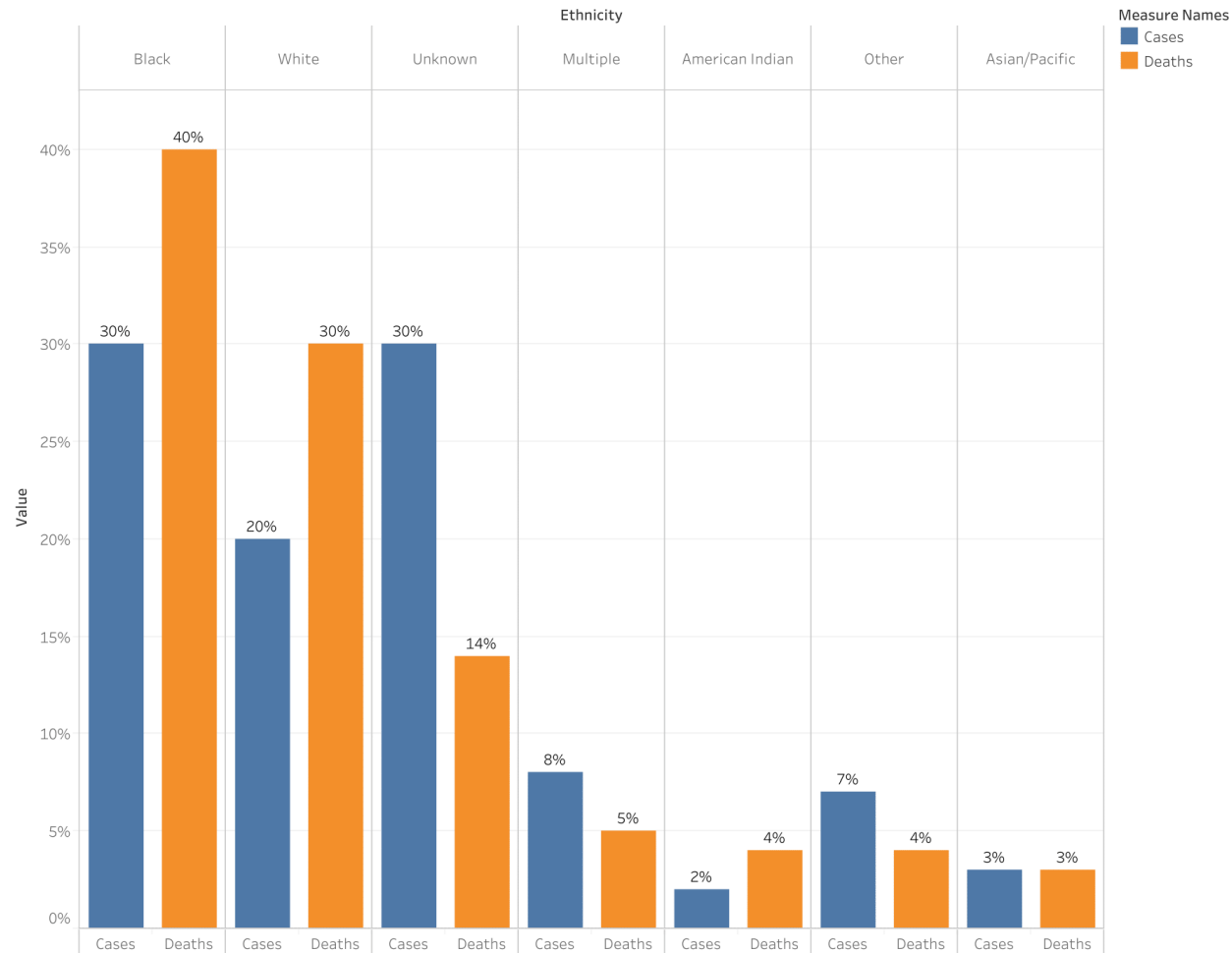


Schema #4

Schema	Details
Columns	Race, Measure Names
Rows	Measure Values
Graph type	Bar
Color	Measure Names
Size	Default
Label	Measure Values

Redesign #4

Paired bars

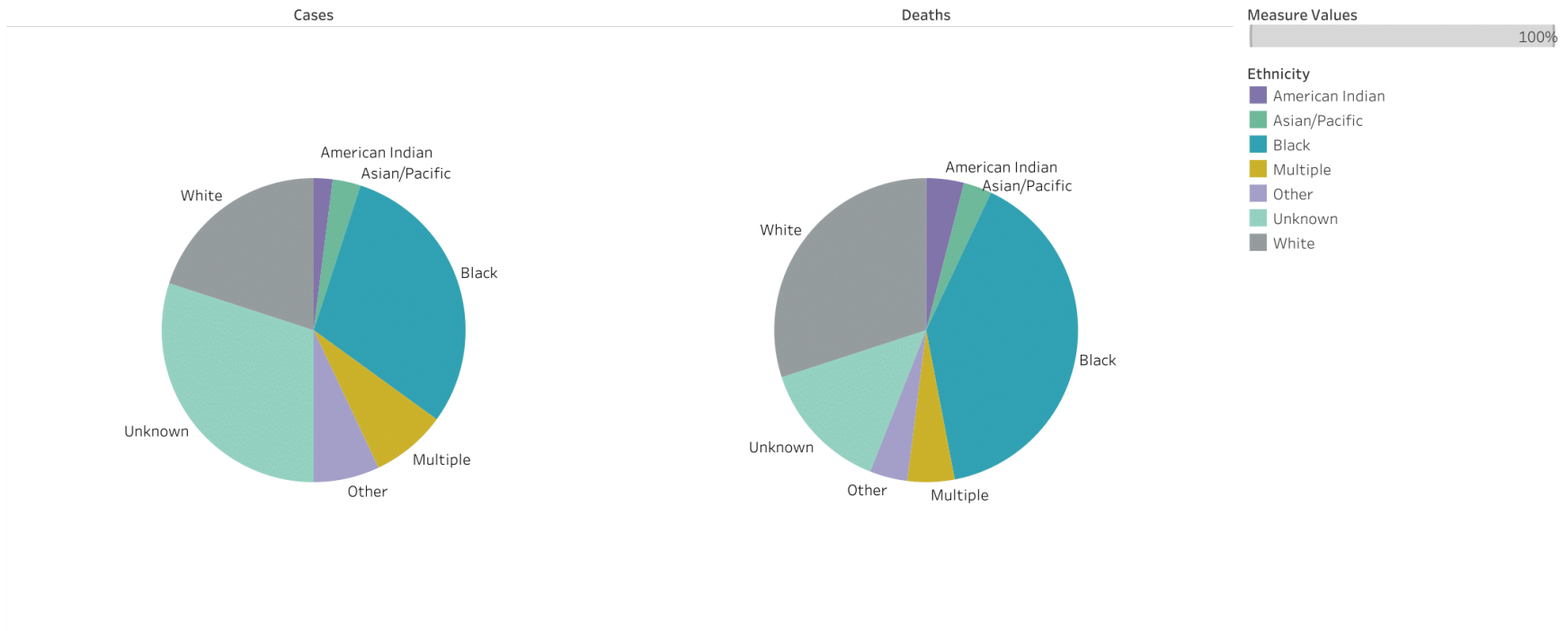


Schema #5

Schema	Details
Columns	Measure Names
Rows	–
Graph type	Pie
Color	Race
Size	Measure Values
Label	Race

Redesign #5

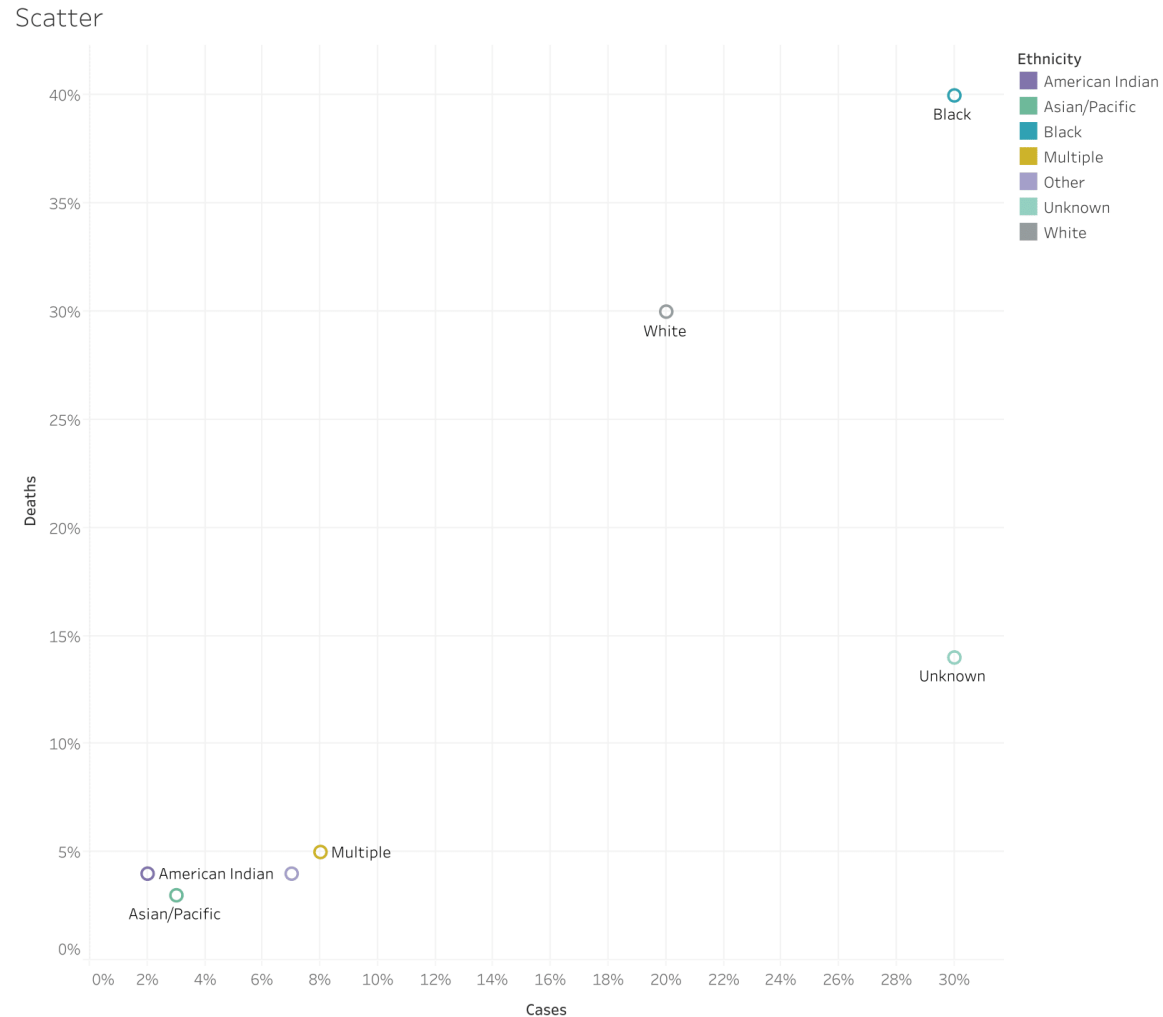
Pies



Schema #6

Schema	Details
Columns	SUM(Cases)
Rows	SUM(Deaths)
Graph type	Shape
Color	Race
Size	Default
Label	Race

Redesign #6

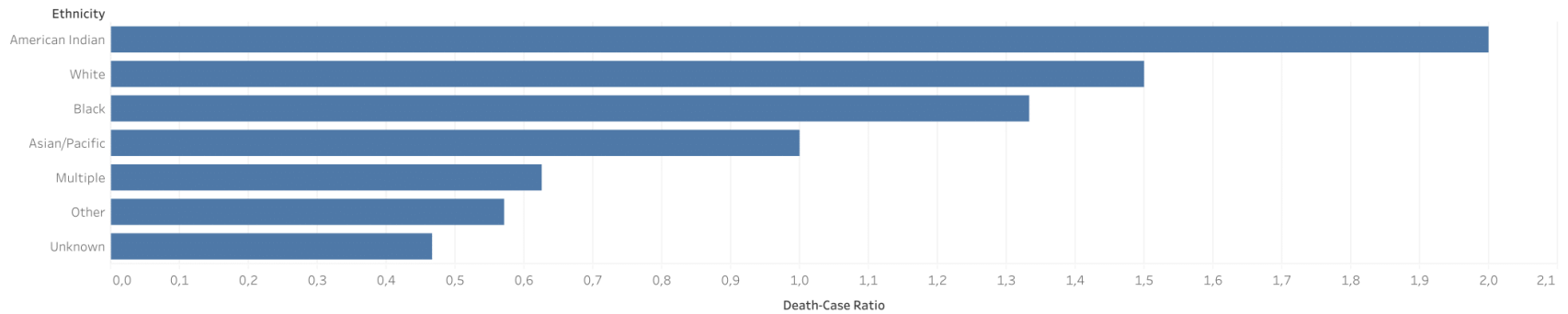


Schema #7

Schema	Details
Columns	SUM([Deaths]/[Cases])
Rows	Race
Graph type	Bar
Color	Default
Size	Default
Label	Default

Redesign #7

Bar Chart Death ratio



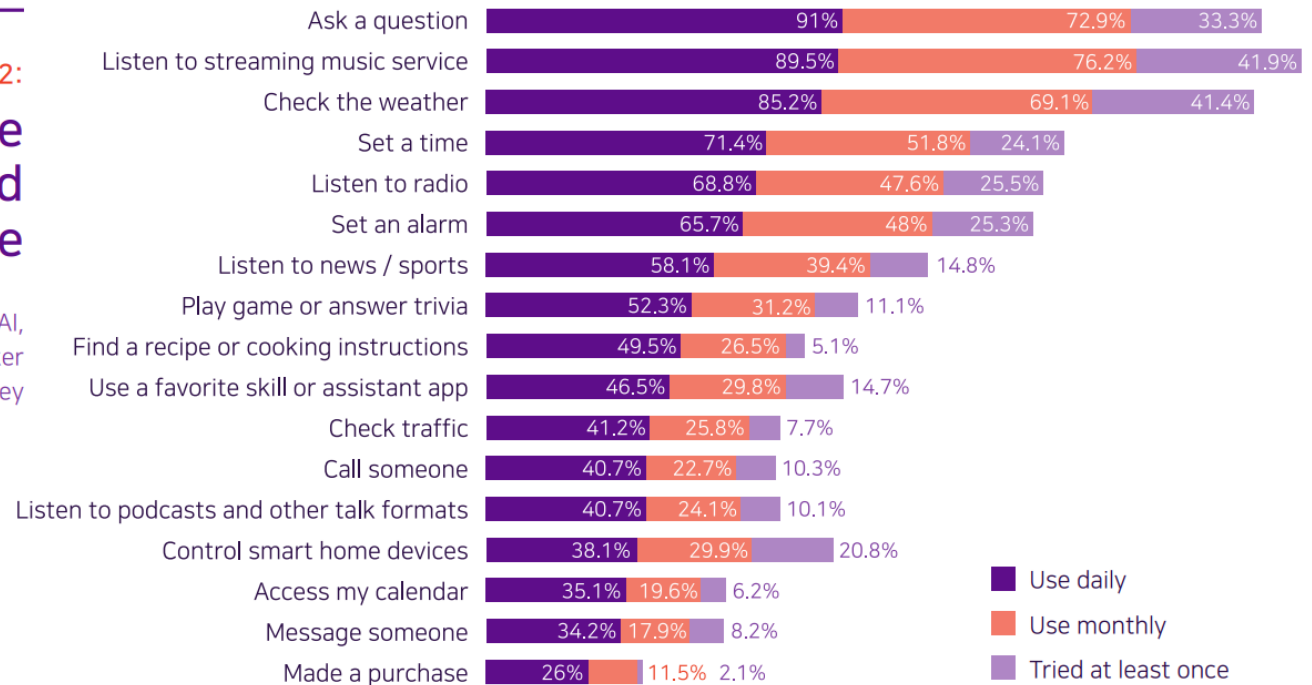
EXAM OF 2020-09-11

Visualization

Image 2:

Uses of voice assistants and frequency of use

Source: Voicebot AI,
2018 Smart Speaker
Use Case Survey



Analysis

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

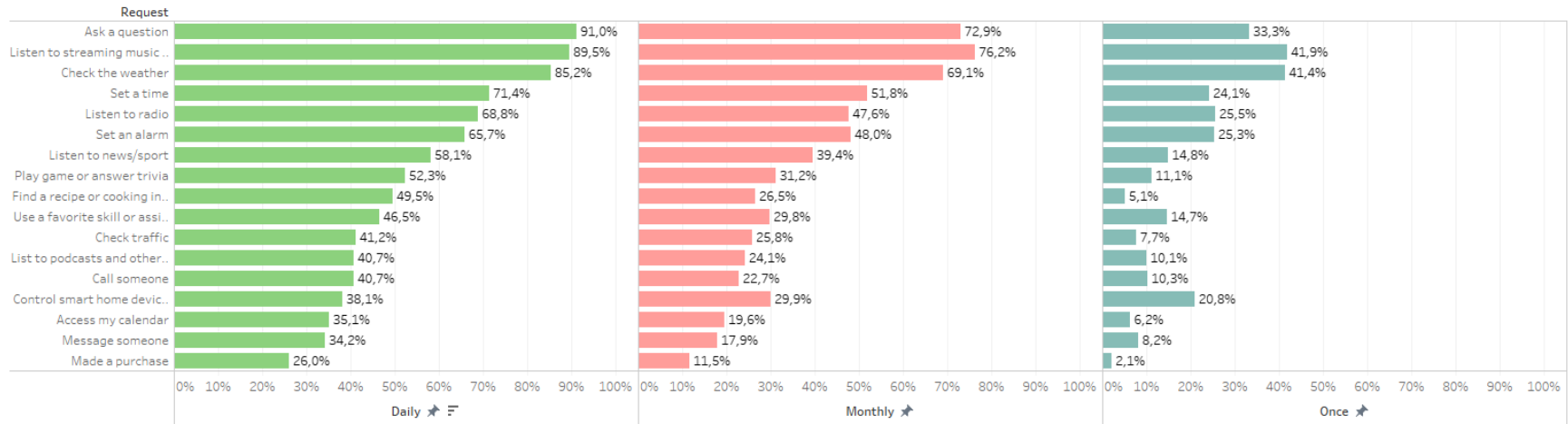
Field	Dim./Measure	Description
USE_DAILY	Measure	Percentage of the daily use
USE_MONTHLY	Measure	Percentage of the monthly use
TRIED_ONCE	Measure	Percentage of used at least once
REQUEST_TYPE	Dimension	The different categories of requests

Schema #1

Schema	Details
Columns	SUM(USE_DAILY), SUM(USE_MONTHLY), SUM(TRIED_ONCE)
Rows	REQUEST_TYPE
Graph type	Bar
Color	Three different colors, one for each use
Size	Default
Label	SUM(USE_DAILY), SUM(USE_MONTHLY), SUM(TRIED_ONCE)

Redesign #1

Uses of voice assistants and frequency of use



Schema #2

Schema	Details
Columns	SUM(USE_DAILY), SUM(USE_MONTHLY), SUM(TRIED_ONCE)
Rows	REQUEST_TYPE
Graph type	Circle
Color	Three different colors, one for each use
Size	Default
Label	SUM(USE_DAILY), SUM(USE_MONTHLY), SUM(TRIED_ONCE)

Redesign #2

Uses of voice assistants and frequency of use

