Table 1

title	audience_rating	critic_rating	dir_age
“Backrooms”	74	89	“20-29”
“Mother Mary”	60	69	“40-49”
“The Bone Temple”	88	92	“30-39”
“The Invite”	89	96	“40-49”
“Tony”	92	93	“40-49”

Table 2

title	audience_rating	critic_rating	dir_20-29	dir_30-39	dir_40-49
“Backrooms”	74	89	True	False	False
“Mother Mary”	60	69	False	False	True
“The Bone Temple”	88	92	False	True	False
“The Invite”	89	96	False	False	True
“Tony”	92	93	False	False	True

Figure A

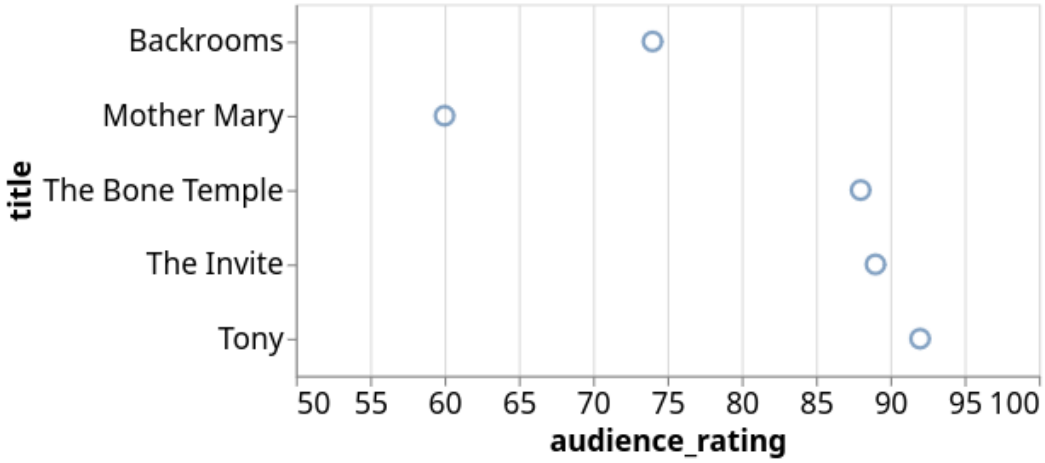


Figure B

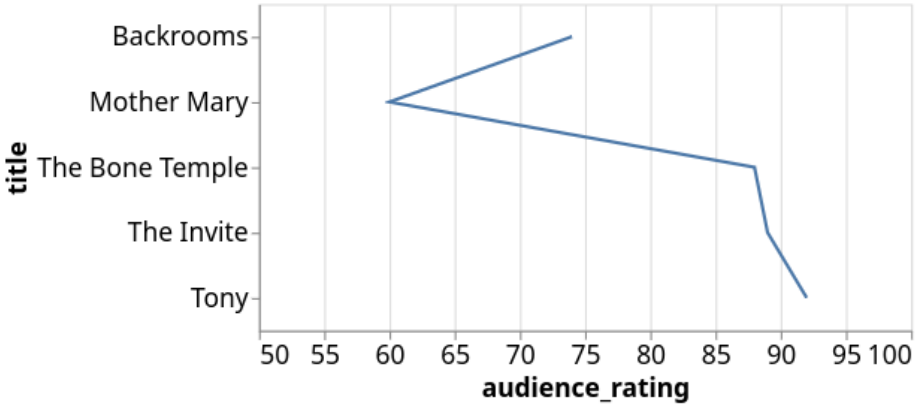


Figure C

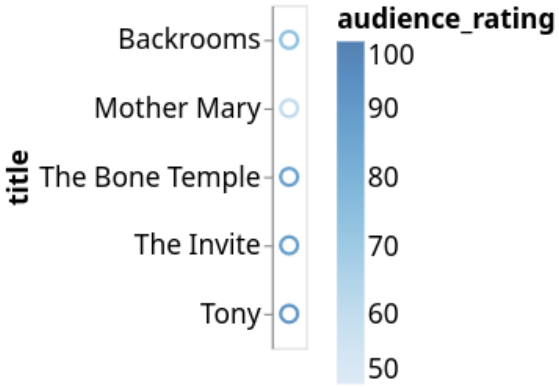


Figure D

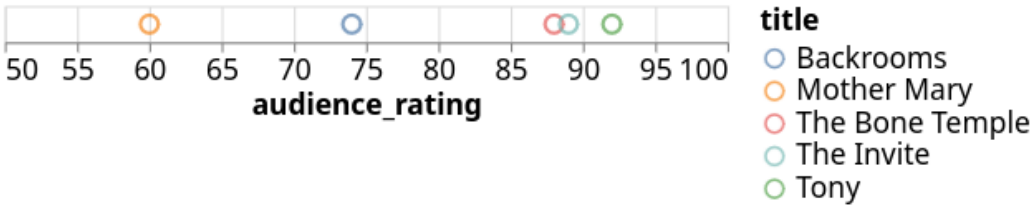
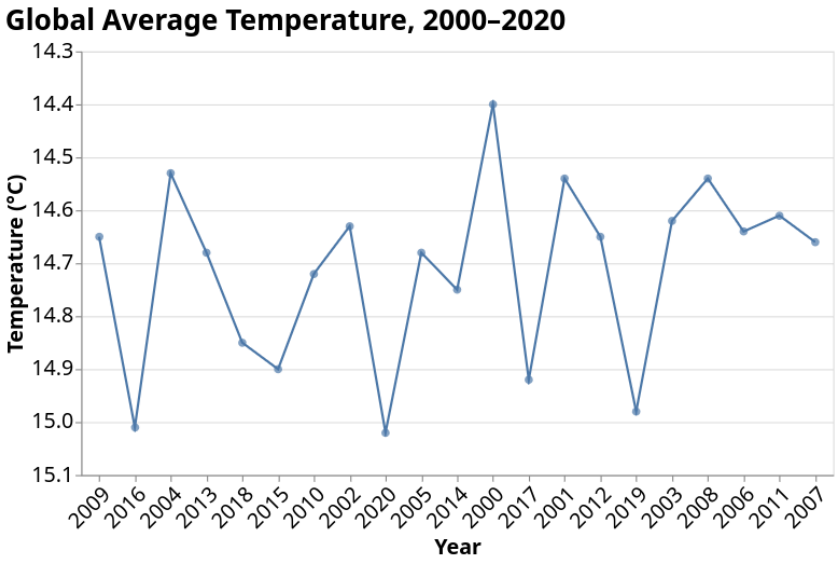


Figure E



Q1 Fill in the following Vega-lite specification so that it produces **Figure B** using Table 1 as data

```
{
      mark: {type: line },
      data: { values: table_1 },
      encoding: {

      x: {

            field: audience_rating,

            type: quantitative,

            "scale": { "domain": [50, 100] }
       },

       y: {

            field: title,

            type: nominal,

       }
   }
}
```

We tried to score this question relatively generously with respect to the exact wording used to describe the different blanks (e.g. “y” and “y-axis” were both accepted). In isolation, there is a little ambiguity as to whether the ratings (which in this case are Rotten Tomatoes scores) are best thought of as quantitative or ordinal. In this code however, the way the domain is specified for the scale does imply that the variable is quantitative.

Q2 Fill in the following Vega-lite specification so that it produces **Figure C** using Table 1 as data

```
{
      mark: {type: point },
      data: { values: table_1 },
      encoding: {

      color: {

            field: audience_rating,

            type: quantitative,

            "scale": { "domain": [50, 100] }
       },

       y: {
```

Name: _____ netid: _____

```
            field: title,

            type: nominal,

        }
    }
}
```

Q3: Select all statements that you believe accurately describe Figures A-D.

- ☐ Assuming the same x-axis domain/scale, Figure A would be better as a bar chart.
- ☑ Figure A uses effective visual encodings.
- ☑ Figure B violates the principle of expressiveness since it encodes relationships that are not in the data
- ☐ Figure C violates the principle of expressiveness
- ☐ Figure C uses effective visual encodings
- ☐ Figure D violates the principle of expressiveness due to overplotting issues

The first option is false, since a *bar chart* with a non-zero axis starting point is deceptive.
The second option is true. Figure A encodes both rating and title using position, which tops Mackinlay's effectiveness rankings for the respective data types. While there may be other critiques of Figure A, the choice of visual encodings is largely effective.
The third option is also true. Using a line mark to connect categorical data is a violation of expressiveness, since it implies some kind of continuity between categories that doesn't exist.
The fourth and fifth options are false. Figure C's color encoding abides by the expressiveness principle since it does convey all relationships in the data, and doesn't convey relationships absent from the data. However, a color encoding is a poor choice for a quantitative variable, particularly when there are other encoding channels available to use.
The sixth option is false. While two data-points do slightly overlap in Figure D, this generally wouldn't be considered overplotting, since all points are still clearly visible.

Q4: Select all statements you believe accurately describe Figure E.

- ☐ Figure E is deceptive, because the y-axis range does not include 0
- ☑ Figure E is deceptive, because the x-axis is not ordered chronologically
- ☑ Figure E is deceptive, because it uses an inverted y-axis
- ☐ Figure E is not deceptive

The first option is false. It is not inherently deceptive to start an axis at a non-zero value. In the case of global climate data, zero Celsius is not a plausible temperature value. Since, a one degree change is significant in terms of global warming, zooming in on the relevant range is appropriate.
The second option is true. For temporal data, non-chronological ordering of dates violates convention, and can easily confuse a viewer not paying close attention.
The third option is true. By convention, most readers will assume that a higher position on the y-axis corresponds to a warmer temperature, particularly when the axis line appears below the

Name: _____ netid: _____

plotted data. Using an inverted y-axis is deceptive.
The fourth option is false since the second and third options are true.

Q5: Select all scenarios in which you believe it is appropriate to scale a variable logarithmically.

- ☐ When you are more interested in making additive comparisons between data points than multiplicative comparisons
- ☑ When you are visualizing data that spans multiple orders of magnitude
- ☐ When you want to make a deceptive visualization
- ☐ When you have data with negative and positive values

The first option is false. The opposite is true – logarithmic scaling means that points of equal ratio are plotted at equal distances from one another. This makes log scales more suitable for multiplicative comparisons than additive ones.
The second option is true. Scaling by log base 10 can be (loosely) thought of as plotting the number of digits in a number rather than the number's raw value. This makes it well-suited for datasets where, say, many data points have values in the range 1-10, and a few have values 100+.
The third option was intended to be false. There is nothing inherently deceptive about log scaling. However, because the question wording was somewhat vague, and it could be possible to produce a deceptive visualization by log-scaling a variable without communicating it clearly, we accepted both checked and unchecked as correct here.
The fourth option is false. You cannot take the log of a negative number (or zero).

Q6: Suppose you created a scatter plot of Table 1 in which `critic_rating` is visualized on the x-axis and `audience_rating` on the y-axis. If you wanted to add an encoding to visualize `dir_age`, which of the following would be most appropriate? Choose one option.

- ○ Color Hue (base color)
- ○ Volume
- ○ Color Value (color lightness)
- ○ Shape

The intended correct option here was color value, under the assumption that dir_age is best thought of as an ordinal variable. Amongst the answer choices, color value tops Mackinlay's effectiveness rankings for ordinal variables. Shape cannot really represent ordinal relationships and volume does not make sense for a two dimensional scatter-plot. However, depending on how the dir_age variable is constructed in the dataset, there is an argument that it is better treated as categorical than ordinal. For instance, if we suspect that there are some strong cohort effects that influence the relationship between age and the other variables, dir_age would make less sense to treat ordinally. In this case, amongst the options color hue tops Mackinlay's rankings, so we decided to allow it as an answer option as well.

Name: _____ netid: _____

Q7: Which of the following statements best describes Tables 1 and 2? Choose one option.

- ○ Table 1 aligns with tidy data principles, Table 2 does not
- ○ Table 2 aligns with tidy data principles, Table 1 does not
- ○ Both tables violate tidy data principles equally
- ○ Both tables align with tidy data principles equally well

Table 2 violates tidy data principles since it spreads what should be a single variable (dir_age) across three separate columns. If we wanted to add a director over 50 to the dataset, we'd need to create a new column. In Table 1, each column is its own distinct variable.

Q8: (Fill in the blank) Suppose you applied an `aggregate` transformation to Table 1. Specifically, you apply an `average` operation on the `critic_rating` field, and grouped by `dir_age`.

The resulting table would have **3** rows.

An aggregate transformation produces a new table where there is one row for each distinct value of the variable being grouped by. In this case, there are three distinct values in the dataset for dir_age, so the final table has three rows.