SI649-20winter Lab 3 -> Altair II

Overview

We will continue working with the dataset from the article “The Dollar-And-Cents Case Against Hollywood’s Exclusion of Women”--don't worry, we'll switch soon. We'll focus on **transformation** and build the following charts:

1. Line chart within a range
2. Heatmap with additional annotations
3. Bar chart with additional annotation.
4. Bar Chart with fold
5. Ugliest chart

For this lab, please write Altair code to answer the questions. In many situations you could also solve the problem using Pandas. However, we want code that can be deployed without using Python so it's better practice to just do as much as we can in Altair. You can complete the entire lab without writing any pandas transformation.

It's fine if your visualization looks slightly different from the example (e.g., getting 1.1 instead of 1.0, use orange instead of red)

Lab Instructions (read the full version on the handout of the previous lab)

- Save, rename, and submit the ipynb file (use your username in the name).
- Run every cell (do Runtime -> Restart and run all to make sure you have a clean working version), print to pdf, submit the pdf file.
- For each visualization, we will ask you to write down a "Grammar of Graphics" plan first (basically a description of what you'll code).
- If you end up stuck, show us your work by including links (URLs) that you have searched for. You'll get partial credit for showing your work in progress.
- There are many bonus point opportunities in this lab.

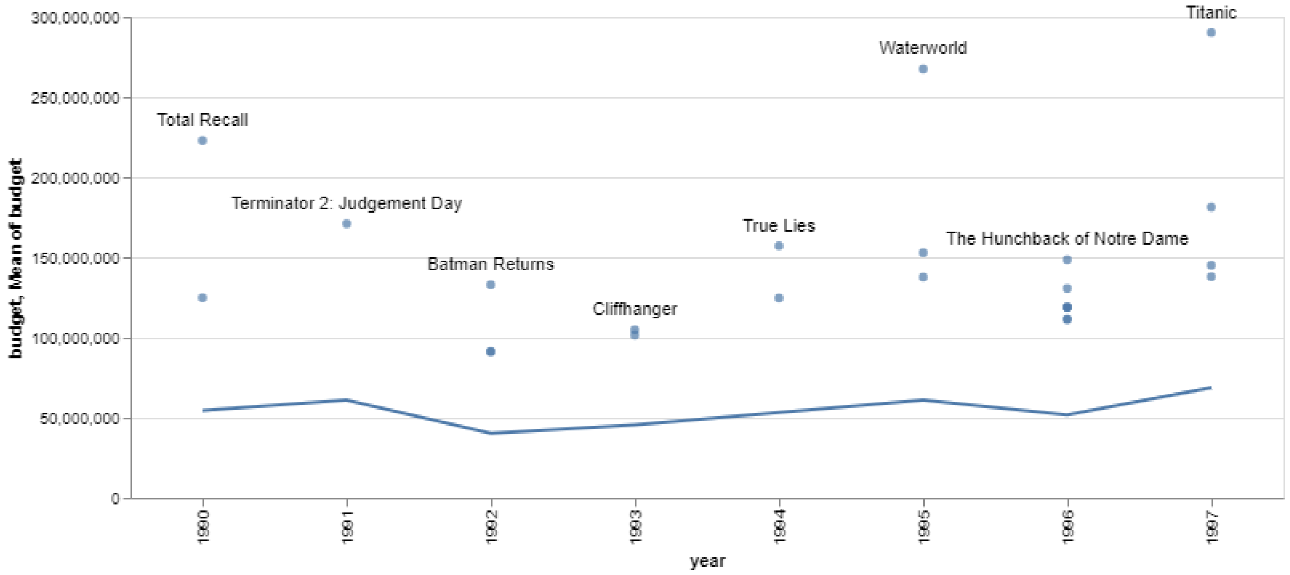
```
# imports we will use
import altair as alt
import pandas as pd
datasetURL="https://raw.githubusercontent.com/LiciaHe/SI649/master/week3/movie_after_1990.csv"
movies_test=pd.read_csv(datasetURL, encoding="latin-1")
```

Here's the processed Bechdel test dataset.

```
movies_test.sample(2)
```

	Unnamed: 0	year	title	binary	budget	dom_gross	int_gross	rating	country	language	test_result	country_binary	roi_dom	roi_int
171	171	2012	The Vow	PASS	8866236	126844056.0	200511006.0	6.8	United States	English	Passes Bechdel Test	U.S. and Canada	14.306415	8.308706
1061	1061	2003	The Matrix Reloaded	PASS	160793043	356471453.0	935102616.0	7.2	United States	English	Passes Bechdel Test	U.S. and Canada	2.216958	3.598608

Visualization 1: Line chart within a range



Description of the visualization:

We want to see a visualization of movies that were significantly above budget relative to the mean. We'd like to know the name of the top movie that year but also to have a sense of how many other outliers (we'll defined those as at least 2x the mean budget) were produced that year.

- plot a line chart using year and average budget for movies between 1990 and 1998.
- plot dots to represent movies with budgets that are at least 2 times bigger than the mean budget of that year. (e.g., if the mean budget of 1990 is 50M, plot movies made in 1990 whose budgets are at least 100M)
- For each year, annotate the title of the movie with the highest budget of that year.

Visualization 1 Plan:

TODO: edit this cell to write your visualization plan. You can write in altair syntax, in full sentence, or in bullet points, whichever way that helps you to plan your chart.

line chart

- Describe the encoding rules:

> TODO:
>
> Mark: mark_line
>
> Data: movies_test
>
> Encoding: alt.X('year:O'), alt.Y('mean(budget):Q')

- Describe the transformations:

> TODO:
>
> transform_filter(year >= 1990 & <1998)

dot chart

- Describe the encoding rules:

> TODO:
>
> Mark: mark_circle
>
> Data: movies_test
>
> Encoding: alt.X('year:O'),alt.Y('budget:Q')

- Describe the transformations:

> TODO:
>
> transform_joinaggregate(#Calculate the mean budget by year,#Calculate budget of each movie by title)

transform_filter(#Filter by those movies that are above the mean by 2,#Years between 1990 -1998)

text annotation

- Describe the encoding rules:

> TODO:
>
> Mark: mark_text
>
> Data: movies_test
>
> Encoding: text= alt.Text('title:O')

- Describe the transformations:

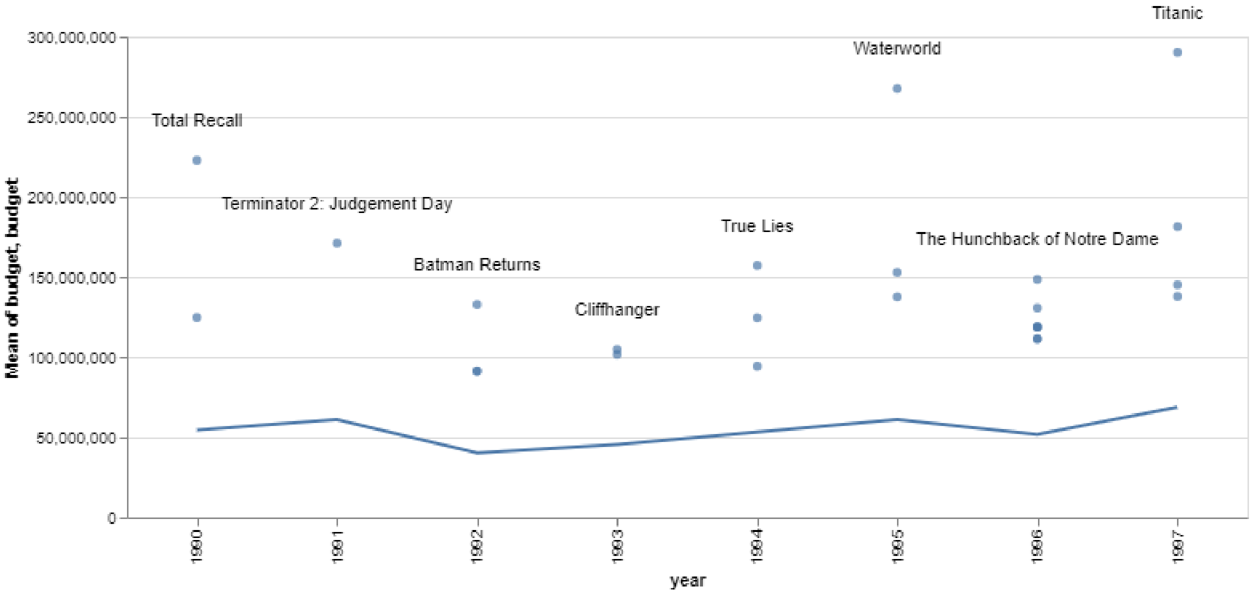
> TODO:
> transform_window(#Sort by budget movie #Get a rank (best to worst) , descending order,#Group by year)
> transform_filter(#Report only the first in each category)

Replicate Vis 1

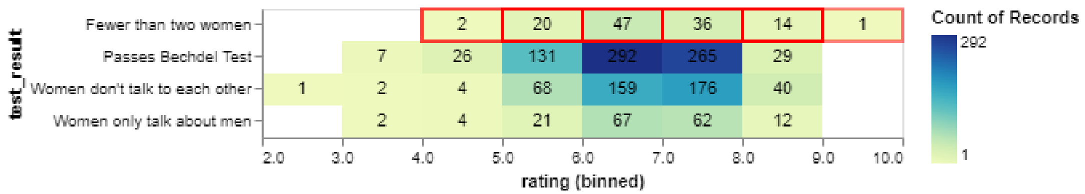
Hint:

- Line chart is the simplest one, you can start from there.
- Think about the difference between transform_aggregate and transform_joinaggregate

```
#TODO: Replicate visualization 1
line_chart=alt.Chart(movies_test).mark_line().encode(
    alt.X('year:O'),
    alt.Y('mean(budget):Q')
).transform_filter(
    # & for and, | of or, > < = !=
    (alt.datum.year >= 1990) & (alt.datum.year <1998)
).properties(
    width=700,
    height=300

)

dot_chart = alt.Chart(movies_test).transform_joinaggregate(
     #Calculate the mean budget by year
    mean_budget_all='mean(budget)',
    groupby=['year']
).transform_joinaggregate(
    #Calculate budget of each movie  by title
    budget_movie='sum(budget)',
    groupby=['title']
).transform_filter(
    #Years between 1990 -1998 inclusive
    (alt.datum.year >= 1990) & (alt.datum.year <1998)
).transform_filter(
    #Filter by those movies that are above the mean by 2
    alt.datum.budget_movie> alt.datum.mean_budget_all*2
).mark_circle().encode(
    alt.X('year:O'),
    alt.Y('budget:Q')
)


text = dot chart.transform window(
    #Sort by budget movie
    #Get a rank (best to worst) , descending order
    #Group by year
  sort=[alt.SortField('budget_movie', order = 'descending')],
  the_rank='rank(budget_movie)',
  groupby = ['year']
).transform_filter(
    #Report only the first (this is the highest in each category)
    alt.datum.the_rank == 1
).mark_text(
    #Make text prettier
    align='center', #left, right, center position of text with respect to dot mark
    baseline='top', #top , middle, bottom
    dy= -30 # Nudges text to up/down (y axes)  so it doesn't appear on the  mark
).encode(
    #Display title info
    text= alt.Text('title:O')
    )

#MERGE
line_chart+dot_chart+text
```



Visualization 2 heatmap with additional annotations



Description of this visualization:

We want to produce a heatmap contrasting different types of bechdel test results to the IMDB scores of those movies. Each cell in the heatmap will tell us how many movies (not normalized here, we want raw count) are in that category. The author of the article wants to point out some property of the really well-regarded movies so wants those highlighted.

- Plot a heatmap with rating and test result. Encode the count of movies as color. Remove the category "dubious".
- For each cell, add text to indicate the count of movies.
- If a test_result category has at least one movie whose rating is higher than 9, highlight that entire category. You can highlight by adding a layer of heatmap that is not filled.

Visualization2 Plan:

TODO: edit this cell to write your visualization plan. You can write in altair syntax, in full sentence, or in bullet points, whichever way that helps you to plan your chart.

heatmap

- Describe the encoding rules:

> TODO:
> Mark: mark_rect
> Data: movies_test
> Encoding: alt.X('rating', bin = True),alt.Y('test_result'), color=('count(rating)')

- Describe the transformations:

> TODO:
> transform_filter(alt.FieldOneOfPredicate(field='test_result', #remove dubious))

text annotation

- Describe the encoding rules:

> TODO:
> Mark: mark_text

> Data: movies_test
> Encoding: text = 'count(rating)',color= alt.value('black')

- Describe the transformations:

> TODO:
> No transformation

highlight

- Describe the encoding rules:

> TODO:
> Mark: mark_rect
> Data: movies_test
> Encoding: alt.X('rating', bin = True),alt.Y('test_result'), color = alt.Color('count(rating)')

- Describe the transformations:

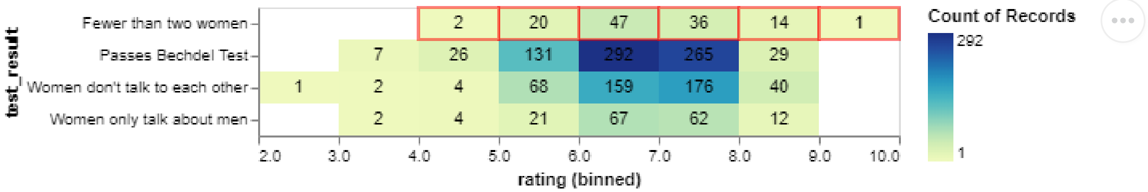
> TODO:
> transform_filter(alt.FieldOneOfPredicate(field='test_result', oneOf=""]))

Replicate Vis 2

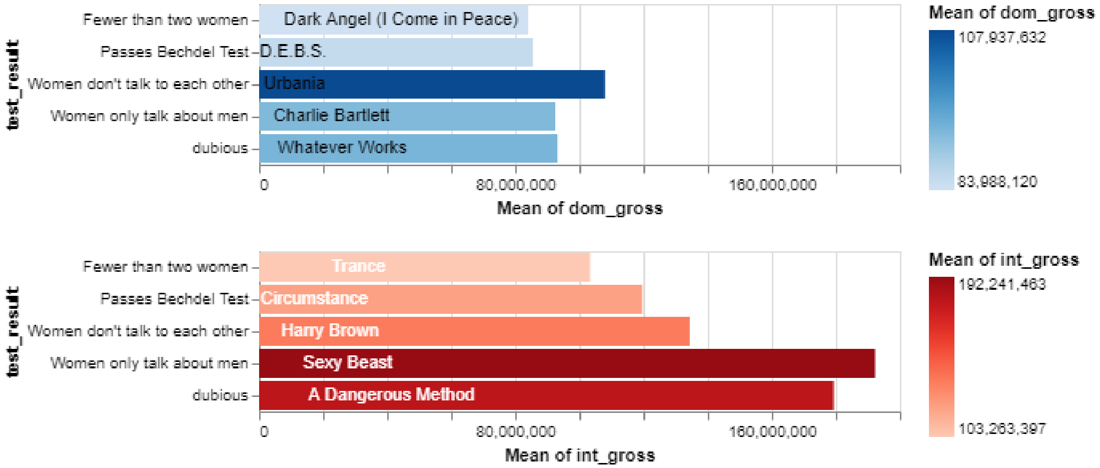
Hint:

- When you pass the heatmap to your text chart, you are passing all of the heatmap's encoding settings as well, which include the color setting.
- How do you translate "at least one movie with rating > 9 " into code?

```
#TODO: Replicate vis 2
heat_map = alt.Chart(movies_test).mark_rect().encode(
    alt.X('rating', bin  = True),
    alt.Y('test_result'),
    color = alt.Color('count(rating)')
).transform_filter(
        alt.FieldOneOfPredicate(field='test_result', oneOf=["Passes Bechdel Test","Women only talk about men","Women don't talk to each other","Fewer than t
        )


text = heat_map.mark_text(
    #Make text prettier
    #align='center', #left, right, center position of text with respect to dot mark
    #align='center', #left, right, center position of text with respect to dot mark
    baseline='middle', #top , middle, bottom
    #dx=-33, # Nudges text to up/down (y axes)  so it doesn't appear on the  mark

).encode(#Display title info

    text = 'count(rating)',
      color= alt.value('black')   # And if it's not true it sets the bar steelblue.
)


highlight = alt.Chart(movies_test).mark_rect(stroke='red').encode(

    alt.X('rating', bin  = True),
    alt.Y('test_result'),
    color = alt.Color('count(rating)')

).transform_filter(

    alt.FieldOneOfPredicate(field='test_result', oneOf=["Fewer than two women"]),

)


(heat_map+highlight+text)
```



Visualization 3: Bar chart with additional annotations



Description of this visualization: We want to contrast the domestic and international gross based on bedchel test passing category. We also want an example movie for each category so we're going to pick the 10th most popular one.

- Plot a bar chart *for movies made in US and Canada*, using the test result and mean of domestic gross.
- For test result category (i.e., each bar), find out the movie whose domestic gross ranks as the 10th lowest (sort by ascending order, the rank = 10). Annotate the title of this movie as text.
- BONUS: Plot a similar bar chart *for movies made internationally* using the international gross. Concatenate the international and domestic charts charts. Make sure they share the same x axis but have independent color scales.

Visualization3 Plan:

TODO: edit this cell to write your visualization plan. You can write in altair syntax, in full sentence, or in bullet points, whichever way that helps you to plan your chart.

bar chart

- Describe the encoding rules:

  > TODO:
  > Mark: mark_bar
  > Data: movies_test
  > Encoding: alt.X('mean(dom_gross)'),alt.Y('test_result'),color=('mean(dom_gross)')

- Describe the transformations:

  > TODO: your answer here
  > transform_filter(# filter by country_binary['U.S. and Canada'])

text annotation

- Describe the encoding rules:

> TODO:
> Mark: mark_text
> Data: movies_test
> Encoding: text= alt.Text('title:O'), color= alt.value('black')

- Describe the transformations::

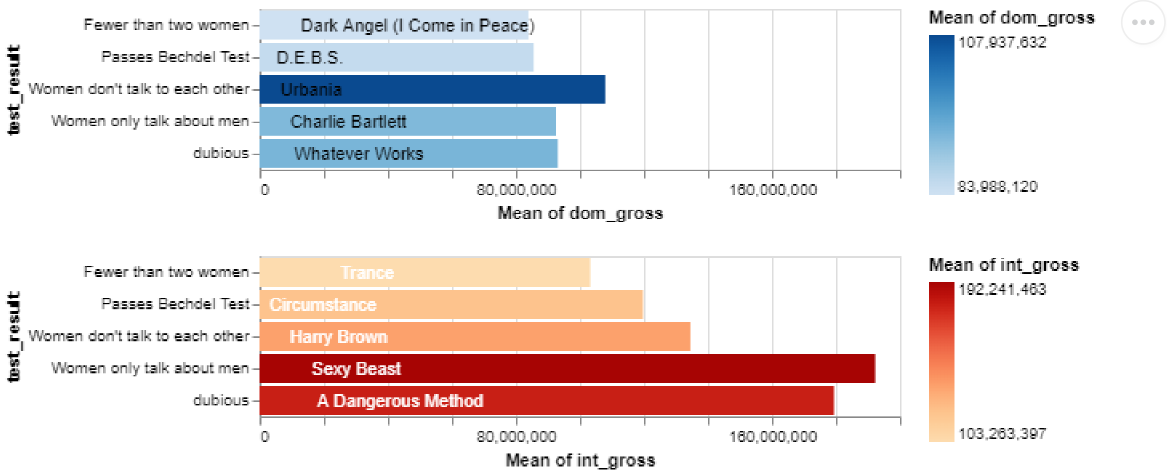
> TODO:
> transform_window(#Sort by budget movie, #Get a rank (best to worst) , #descending order, #Group by year)
> transform_filter(#Filter by country_binary 'U.S. and Canada') transform_filter(#Report only the first element of Rank (highest))

Replicate Vis 3

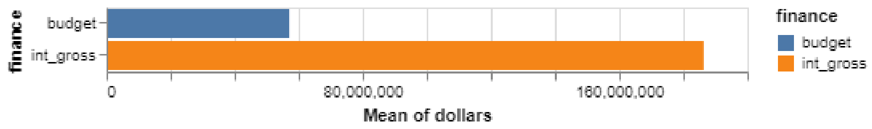
Hint:

- You want to generate the rank using a window transformation.
- It's fine if text annotations don't align perfectly with your bars.
- There are several transformations in this task. If your chart looks slightly different from ours, you are still likely to get full points.

```
#TODO: Replicate Vis 3
bar_chart= alt.Chart(movies_test).mark_bar().encode(
    alt.X('mean(dom_gross)'),
    alt.Y('test_result'),
     color=alt.Color('mean(dom_gross)')

).transform_filter(
    # & for and, | of or, > < = !=
    (alt.datum.country_binary=='U.S. and Canada')
)

text = bar_chart.transform_window(
    #Sort by budget movie
    #Get a rank (best to worst) , descending order
    #Group by year
    sort=[alt.SortField('dom_gross', order = 'ascending')],
    the_rank='rank(dom_gross)',
    groupby = ['test_result']
).transform_filter(
    # & for and, | of or, > < = !=
    (alt.datum.country_binary=='U.S. and Canada')
).transform_filter(
    #Report only the first (this is the highest in each category)
    alt.datum.the_rank == 10
).mark_text(
    #Make text prettier
    align='left', #left, right, center position of text with respect to dot mark
    baseline='top', #top , middle, bottom
    dx= 10, # Nudges text to up/down (y axes)  so it doesn't appear on the  mark
    dy = -5

).encode(
    #Display title info
    text= alt.Text('title:O'),
    color= alt.value('black')

    )

bar_chart2= alt.Chart(movies_test).mark_bar().encode(
    alt.X('mean(int_gross)'),
    alt.Y('test_result'),
     color=alt.Color('mean(int_gross)', scale=alt.Scale(scheme='orangered'))

).transform_filter(
    # & for and, | of or, > < = !=
    (alt.datum.country_binary=='International')
)

text2 = bar_chart2.transform_window(
    #Sort by budget movie
    #Get a rank (best to worst) , descending order
    #Group by year
    sort=[alt.SortField('int_gross', order = 'ascending')],
    the_rank='rank(int_gross)',
    groupby = ['test_result']
).transform_filter(
    # & for and, | of or, > < = !=
    (alt.datum.country_binary=='International')
).transform_filter(
    #Report only the first (this is the highest in each category)
    alt.datum.the_rank == 10
).mark_text(
    #Make text prettier
    align='left', #left, right, center position of text with respect to dot mark
    baseline='top', #top , middle, bottom
    dx= 5, # Nudges text to up/down (y axes)  so it doesn't appear on the  mark
    dy = -5
).encode(
).encode(
    #Display title info
    text= alt.Text('title:O'),
    color= alt.value('white')

    )

blue = (bar_chart+text)
red = (bar_chart2+text2)

combined= alt.vconcat(blue,red).resolve_scale( x="shared", color= "independent")
combined
```



Visualization 4: Bar chart



Description of this visualization:

- Plot a bar chart that enables the comparison between mean budget and mean international gross across the whole dataset.
- Two bars should be colored differently

Visualization4 Plan:

TODO: edit this cell to write your visualization plan. You can write in altair syntax, in full sentence, or in bullet points, whichever way that helps you to plan your chart.

- What kind of transformation do I need to plot this bar chart?

  > TODO: your answer here
  > Get only the necessary varibales wide= ["title","budget","int_gross"] Next convert wide to long format long = movies_wide.melt()

- What kind of encoding do I want to use??

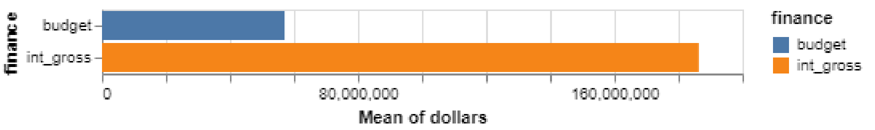
  > TODO:
  > x = alt.X('mean(dollars):Q'),
  > y = alt.Y('finance:N')
  > color = alt.Color('finance:N')
  > title
  > var_name='finance',
  > value_name='dollars'

Replicate Vis 4

Hint:

- Do I need a long form or wide form data? Which type of data do I have?

```
# TODO: replicate vis 4

#This is the current wide data (only the variables that we truly need)
movies_wide=movies_test.loc[:,["title","budget","int_gross"]]

#This is the data frame in which we have Budget/Int_Gross as tags
movies_long=movies_wide.melt('title', var_name='finance', value_name='dollars')

finance_chart= alt.Chart(movies_long).mark_bar().encode(
    x = alt.X('mean(dollars):Q'),
    y = alt.Y('finance:N'),

    color = alt.Color('finance:N')
)

finance_chart
```



5. Ugly Visualization

For this task, we're going to ask that you make the worst(!), ugliest, chart possible. Please use the movies dataset as a starting point, but you can add to it if you want.

You don't have to do it programatically. You can start with the Altair version or go back to Tableau, Excel, or whatever you want... You can download the chart and edit it in photoshop, illustrator, inkscape... any software you want to use. Please don't just turn in just a blank screen... :) Your chart should express **something** useful. Just do it badly.

After making the chart, please upload it to canvas (jpg or png or pdf).

When you're done crafting your terrible chart, use the following cell to tell us what perception rules you are violating that make your chart really bad!

TODO:

The chart that I have created a bar chart that shows the median rating/and revenue according to director gender and movie category (comedy) for each country in the data set. This graph's purpose is to show the revenue/rating of films who pass the test and see if the gender and country of the director has an impact on the results. Some violations: there is no title, no description of what is going on, too much information the viewer is overwhelmed with an insane amoun of information, there is no principal focus (not a specific order of imporatnce for what it pertains the variables). The graph has no apparent finding; the purpose of it its unclear.

This is the end of lab 3.

Please run all cells (Runtime->Restart and run all), and

1. save to PDF (File->Print->Save PDF -> landscape, shrink to 80%)
2. save to ipynb (File -> Download .ipynb)

Rename both files with your uniqname: e.g. uniqname.pdf/ uniqname.ipynb

Upload both files AND the ugliest chart to canvas