Exploratory data analysis (EDA)

Contents

Lecture learning goals

By the end of the lecture you will be able to:

1. Visualize missing values.
2. Define what EDA is on a conceptual levels both for numerical and categorical variables.
3. Use repeated plot grids to investigate multiple data frame columns in the same plot.
4. Visualize correlations.
5. Visualize counts of categorical variables.

Required activities

Before class:

- This 10 min video on EDA concepts.

After class:

- Review the lecture notes.
- Interactive flow chart for which visualization to choose. Not really anything to read, but go there to look around and bookmark it to use as a reference when needed throughout this course and others.

```python
import altair as alt
import pandas as pd

# Save a vega-lite spec and a PNG blob for each plot in the notebook
alt.renderers.enable('mimetype')
# Handle large data sets without embedding them in the notebook
alt.data_transformers.enable('data_server')
```

```
DataTransformerRegistry.enable('data_server')
```

4.1. Missing values

4.1.1. Py

First we are going to load in some weather data and simulate a faulty sensor by assigning some NaN values manually.

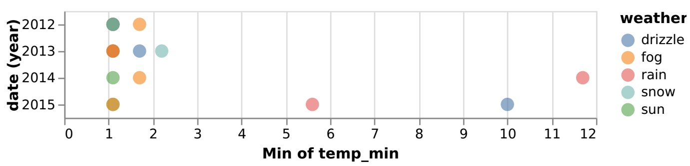
```
from vega_datasets import data
import numpy as np


df = data.seattle_weather()
# Artificially introduce some NaNs to simulate a sensor that can't measure below freezing
df.loc[df['temp_min'] <= 0.6, 'temp_min'] = np.nan
df.loc[df['temp_max'] <= 0.6, 'temp_max'] = np.nan
df
```

	date	precipitation	temp_max	temp_min	wind	weather
0	2012-01-01	0.0	12.8	5.0	4.7	drizzle
1	2012-01-02	10.9	10.6	2.8	4.5	rain
2	2012-01-03	0.8	11.7	7.2	2.3	rain
3	2012-01-04	20.3	12.2	5.6	4.7	rain
4	2012-01-05	1.3	8.9	2.8	6.1	rain
...	...	...	...	...	...	...
1456	2015-12-27	8.6	4.4	1.7	2.9	fog
1457	2015-12-28	1.5	5.0	1.7	1.3	fog
1458	2015-12-29	0.0	7.2	NaN	2.6	fog
1459	2015-12-30	0.0	5.6	NaN	3.4	sun
1460	2015-12-31	0.0	5.6	NaN	3.5	sun

1461 rows × 6 columns

If we were to visualize this data in order to answer the the question “what what the minimum temperature each year for the difference types of weather?” it might look something like this:

```
alt.Chart(df).mark_circle(size=100, opacity=0.6).encode(
    y='year(date):N',
    x='min(temp_min)',
    color='weather'
)
```



One immediate drawback with this visualization is that there could be overlapping dots in the exact same spot which make it difficult to see the exact values from each weather type. This chart still reveals something peculiar about our data: it looks like the dots are neatly arranged in a grid around the lowest temperature for each year. Almost too neat… this might be due to some binning on the data or maybe that the resolution of the sensor is low. There is also probably something going where the lowest values are clamped to just over one degree celsius, that seems quite odd as we would expect some variation in the yearly minimum temperature as well. This could be due to some error in how the data was pre-processed or maybe that the sensor mechanics don’t tolerate values under freezing and it stops measuring during that time.

In general when we read in data, it is a good idea to perform some EDA on the missing values before starting to do it on the rest of the data, regardless of whether we have already observed a weird pattern or not. `df.info()` can show us how many missing values there are, but we don’t get any info on their pattern.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1461 entries, 0 to 1460
Data columns (total 6 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   date           1461 non-null   datetime64[ns]
 1   precipitation  1461 non-null   float64
 2   temp_max       1456 non-null   float64
 3   temp_min       1345 non-null   float64
 4   wind           1461 non-null   float64
 5   weather        1461 non-null   object
dtypes: datetime64[ns](1), float64(4), object(1)
memory usage: 68.6+ KB
```

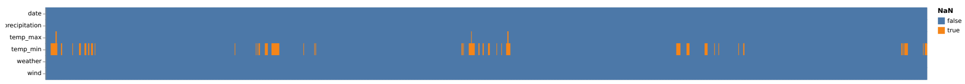
To see if there is a pattern in the missing values, the simplest we could do would be to use pandas built in styling functions. This produces quite long output, so I have truncated it here, but if you were to scroll through it you would see that there seems to be a periodicity to the data and since this data set is ordered by the date the measurement was collected it seems like this periodic variation is through time.

```
df[:40].style.highlight_null()
```

	date	precipitation	temp_max	temp_min	wind	weather
0	2012-01-01 00:00:00	0.000000	12.800000	5.000000	4.700000	drizzle
1	2012-01-02 00:00:00	10.900000	10.600000	2.800000	4.500000	rain
2	2012-01-03 00:00:00	0.800000	11.700000	7.200000	2.300000	rain
3	2012-01-04 00:00:00	20.300000	12.200000	5.600000	4.700000	rain
4	2012-01-05 00:00:00	1.300000	8.900000	2.800000	6.100000	rain
5	2012-01-06 00:00:00	2.500000	4.400000	2.200000	2.200000	rain
6	2012-01-07 00:00:00	0.000000	7.200000	2.800000	2.300000	rain
7	2012-01-08 00:00:00	0.000000	10.000000	2.800000	2.000000	sun
8	2012-01-09 00:00:00	4.300000	9.400000	5.000000	3.400000	rain
9	2012-01-10 00:00:00	1.000000	6.100000	nan	3.400000	rain
10	2012-01-11 00:00:00	0.000000	6.100000	nan	5.100000	sun
11	2012-01-12 00:00:00	0.000000	6.100000	nan	1.900000	sun
12	2012-01-13 00:00:00	0.000000	5.000000	nan	1.300000	sun
13	2012-01-14 00:00:00	4.100000	4.400000	nan	5.300000	snow
14	2012-01-15 00:00:00	5.300000	1.100000	nan	3.200000	snow
15	2012-01-16 00:00:00	2.500000	1.700000	nan	5.000000	snow
16	2012-01-17 00:00:00	8.100000	3.300000	nan	5.600000	snow
17	2012-01-18 00:00:00	19.800000	nan	nan	5.000000	snow
18	2012-01-19 00:00:00	15.200000	nan	nan	1.600000	snow
19	2012-01-20 00:00:00	13.500000	7.200000	nan	2.300000	snow
20	2012-01-21 00:00:00	3.000000	8.300000	3.300000	8.200000	rain
21	2012-01-22 00:00:00	6.100000	6.700000	2.200000	4.800000	rain
22	2012-01-23 00:00:00	0.000000	8.300000	1.100000	3.600000	rain
23	2012-01-24 00:00:00	8.600000	10.000000	2.200000	5.100000	rain
24	2012-01-25 00:00:00	8.100000	8.900000	4.400000	5.400000	rain
25	2012-01-26 00:00:00	4.800000	8.900000	1.100000	4.800000	rain
26	2012-01-27 00:00:00	0.000000	6.700000	nan	1.400000	drizzle
27	2012-01-28 00:00:00	0.000000	6.700000	nan	2.200000	rain
28	2012-01-29 00:00:00	27.700000	9.400000	3.900000	4.500000	rain
29	2012-01-30 00:00:00	3.600000	8.300000	6.100000	5.100000	rain
30	2012-01-31 00:00:00	1.800000	9.400000	6.100000	3.900000	rain
31	2012-02-01 00:00:00	13.500000	8.900000	3.300000	2.700000	rain
32	2012-02-02 00:00:00	0.000000	8.300000	1.700000	2.600000	sun
33	2012-02-03 00:00:00	0.000000	14.400000	2.200000	5.300000	sun
34	2012-02-04 00:00:00	0.000000	15.600000	5.000000	4.300000	sun
35	2012-02-05 00:00:00	0.000000	13.900000	1.700000	2.900000	sun
36	2012-02-06 00:00:00	0.000000	16.100000	1.700000	5.000000	sun
37	2012-02-07 00:00:00	0.300000	15.600000	7.800000	5.300000	rain
38	2012-02-08 00:00:00	2.800000	10.000000	5.000000	2.700000	rain
39	2012-02-09 00:00:00	2.500000	11.100000	7.800000	2.400000	rain

We could create a more effective visualization of this using Altair, but it would be a bit involved as you can see below. The reason we are melting the data is so that we can perform this viz for all the columns since we don't know where the EDA values are when we do EDA. If you are only interested in one column, you could skip the melting part.

```
alt.Chart(
    df.isna().reset_index().melt(
        id_vars='index'
    )
).mark_rect().encode(
    alt.X('index:O', axis=None),
    alt.Y('variable', title=None),
    alt.Color('value', title='NaN'),
    alt.Stroke('value')  # We set the stroke which is the outline of each rectangle in the heatmap
).properties(
    width=df.shape[0]
)
```



Aha! Here we can clearly see the periodic pattern in the data. There is a predictable reoccurrence of the missing values. Again, remember that this data is ordered by time even thought we are using the index when plotting it to make sure this technique works with dataframe that do not have a date column.

We could explore what this patterns is due to by sorting the data by another variable. Let's test our hypothesis that the sensor can't measure low temperatures. If this is true, we should see that all the temp_min values bunch up to the left since this is also where the lowest temp_max temparatures are (a day with a low max temperature is more likely to have a low min temperature).

```
alt.Chart(
    df.sort_values(
        'temp_max',
        ignore_index=True
    ).isna().reset_index().melt(
        id_vars='index'
    )
).mark_rect().encode(
    alt.X('index:O', axis=None),
    alt.Y('variable', title=None),
    alt.Color('value', title='NaN'),
    alt.Stroke('value')
).properties(
    width=df.shape[0]
)
```

There does indeed seem like the missing values are related to low daily temperatures! (to the far right you can see the values that are NaN for both the max and min temperature, NaNs are sorted after all numerical values in pandas).

We could have used the sorting technique with the pandas styler as well:

```
# Just showing the first 50 to save space
df.sort_values('temp_max')[:50].style.highlight_null()
```

	date	precipitation	temp_max	temp_min	wind	weather
708	2013-12-09 00:00:00	0.000000	1.100000	nan	1.300000	sun
705	2013-12-06 00:00:00	0.000000	1.100000	nan	4.700000	sun
704	2013-12-05 00:00:00	0.000000	1.100000	nan	2.600000	sun
14	2012-01-15 00:00:00	5.300000	1.100000	nan	3.200000	snow
384	2013-01-19 00:00:00	0.000000	1.100000	nan	1.900000	drizzle
1428	2015-11-29 00:00:00	0.000000	1.700000	nan	0.900000	fog
15	2012-01-16 00:00:00	2.500000	1.700000	nan	5.000000	snow
386	2013-01-21 00:00:00	0.000000	2.200000	nan	1.100000	drizzle
707	2013-12-08 00:00:00	0.000000	2.200000	nan	2.200000	sun
378	2013-01-13 00:00:00	0.000000	2.200000	nan	1.500000	sun
377	2013-01-12 00:00:00	0.000000	2.800000	nan	2.000000	sun
1064	2014-11-30 00:00:00	0.000000	2.800000	nan	4.400000	sun
765	2014-02-04 00:00:00	0.000000	2.800000	nan	4.700000	sun
376	2013-01-11 00:00:00	0.000000	2.800000	nan	1.900000	drizzle
379	2013-01-14 00:00:00	0.000000	3.300000	nan	1.300000	sun
387	2013-01-22 00:00:00	0.000000	3.300000	nan	0.600000	drizzle
1094	2014-12-30 00:00:00	0.000000	3.300000	nan	3.600000	sun
768	2014-02-07 00:00:00	0.000000	3.300000	nan	4.200000	sun
385	2013-01-20 00:00:00	0.000000	3.300000	nan	2.100000	drizzle
375	2013-01-10 00:00:00	0.300000	3.300000	nan	2.100000	snow
383	2013-01-18 00:00:00	0.000000	3.300000	nan	1.300000	drizzle
365	2012-12-31 00:00:00	0.000000	3.300000	nan	2.000000	drizzle
16	2012-01-17 00:00:00	8.100000	3.300000	nan	5.600000	snow
1095	2014-12-31 00:00:00	0.000000	3.300000	nan	3.000000	sun
382	2013-01-17 00:00:00	0.000000	3.900000	nan	1.000000	drizzle
770	2014-02-09 00:00:00	0.500000	3.900000	nan	2.400000	fog
352	2012-12-18 00:00:00	3.300000	3.900000	nan	5.300000	snow
1455	2015-12-26 00:00:00	0.000000	4.400000	nan	2.500000	sun
1063	2014-11-29 00:00:00	3.600000	4.400000	nan	5.300000	fog
349	2012-12-15 00:00:00	5.300000	4.400000	nan	5.100000	snow
1456	2015-12-27 00:00:00	8.600000	4.400000	1.700000	2.900000	fog
5	2012-01-06 00:00:00	2.500000	4.400000	2.200000	2.200000	rain
703	2013-12-04 00:00:00	0.000000	4.400000	nan	1.600000	sun
1065	2014-12-01 00:00:00	0.000000	4.400000	nan	2.200000	sun
13	2012-01-14 00:00:00	4.100000	4.400000	nan	5.300000	snow
364	2012-12-30 00:00:00	0.000000	4.400000	nan	1.800000	drizzle
1454	2015-12-25 00:00:00	5.800000	5.000000	2.200000	1.500000	fog
12	2012-01-13 00:00:00	0.000000	5.000000	nan	1.300000	sun
56	2012-02-26 00:00:00	1.300000	5.000000	nan	3.400000	snow
764	2014-02-03 00:00:00	0.000000	5.000000	nan	4.300000	sun

	date	precipitation	temp_max	temp_min	wind	weather
77	2012-03-18 00:00:00	3.600000	5.000000	nan	2.700000	rain
702	2013-12-03 00:00:00	0.000000	5.000000	nan	5.600000	sun
1452	2015-12-23 00:00:00	6.100000	5.000000	2.800000	7.600000	fog
59	2012-02-29 00:00:00	0.800000	5.000000	1.100000	7.000000	snow
710	2013-12-11 00:00:00	0.000000	5.000000	nan	0.800000	sun
1457	2015-12-28 00:00:00	1.500000	5.000000	1.700000	1.300000	fog
363	2012-12-29 00:00:00	1.500000	5.000000	3.300000	1.700000	rain
366	2013-01-01 00:00:00	0.000000	5.000000	nan	2.700000	sun
1098	2015-01-03 00:00:00	0.000000	5.000000	1.700000	1.700000	fog
718	2013-12-19 00:00:00	0.000000	5.000000	nan	2.100000	sun

The missing values in our case were directly related to the column itself, low `temp_min` values meant that they were missing. Values could also be missing seemingly randomly, or depending on another column in the data frame. These three types of missing values are sometimes called:

- Missing not at random - Depends on the value in the column itself
- Missing at random - Depends on the value in another column
- Missing completely at random - Doesn't seem to depend on any other *measured* variable

The underlying reasons for missing values can be quite domain specific, e.g. a medical doctor might be missing blood pressure measurements for younger adults than for older adults since that information might be more valuable for older people. In this case we would say that the values are missing at random since there is no relation between the actual blood pressure reading and whether a value goes missing but there is a relation to the patients age which would be another variable in our dataset (it is more important you remember these principles than memorizing the exact names). There are more discussion of these concepts in this article.

4.1.2. R

There is no styler function in dplyer like in pandas, so we will go straight to showing a summary and creating the heatmap:

```
%load_ext rpy2.ipython
```

```
%%R -i df
library(tidyverse)

df |> glimpse()
```

```
/home/joel/miniconda3/envs/531-test/lib/python3.10/site-packages/pandas/core/arrays
/datetimes.py:2236: PytzUsageWarning: The zone attribute is specific to pytz's interface;
please migrate to a new time zone provider. For more details on how to do so, see
https://pytz-deprecation-shim.readthedocs.io/en/latest/migration.html
  values, tz_parsed = conversion.datetime_to_datetime64(data.ravel("K"))
```

```
── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──
✔ ggplot2 3.3.6      ✔ purrr   0.3.5
✔ tibble  3.1.8      ✔ dplyr   1.0.10
✔ tidyr   1.2.1      ✔ stringr 1.4.1
✔ readr   2.1.3      ✔ forcats 0.5.2
── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
Rows: 1,461
Columns: 6
$ date          <dttm> 2012-01-01, 2012-01-02, 2012-01-03, 2012-01-04, 2012-01…
$ precipitation <dbl> 0.0, 10.9, 0.8, 20.3, 1.3, 2.5, 0.0, 0.0, 4.3, 1.0, 0.0,…
$ temp_max      <dbl> 12.8, 10.6, 11.7, 12.2, 8.9, 4.4, 7.2, 10.0, 9.4, 6.1, 6…
$ temp_min      <dbl> 5.0, 2.8, 7.2, 5.6, 2.8, 2.2, 2.8, 2.8, 5.0, NaN, NaN, N…
$ wind          <dbl> 4.7, 4.5, 2.3, 4.7, 6.1, 2.2, 2.3, 2.0, 3.4, 3.4, 5.1, 1…
$ weather       <chr> "drizzle", "rain", "rain", "rain", "rain", "rain", "rain…
```

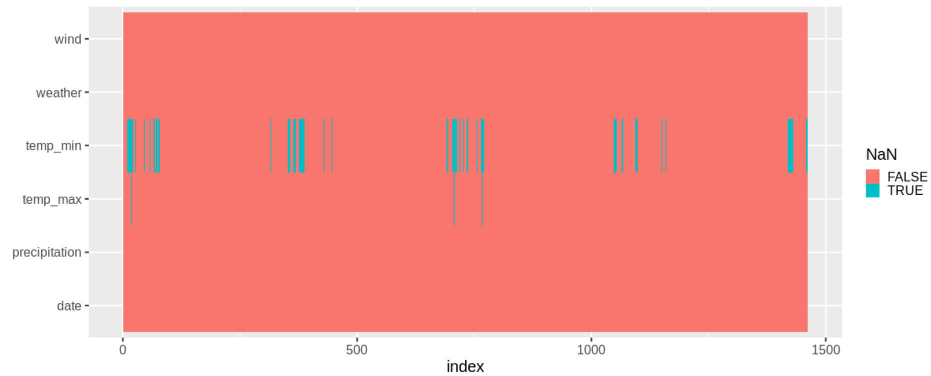
There is no overview of NAs in `glimpse`, instead we can see the NAs in `summary`:

```
%%R
df |> summary()
```

```
      date                      precipitation       temp_max
 Min.   :2012-01-01 00:00:00   Min.   : 0.000   Min.   : 1.10
 1st Qu.:2012-12-31 00:00:00   1st Qu.: 0.000   1st Qu.:10.97
 Median :2013-12-31 00:00:00   Median : 0.000   Median :15.60
 Mean   :2013-12-30 23:20:54   Mean   : 3.029   Mean   :16.50
 3rd Qu.:2014-12-31 00:00:00   3rd Qu.: 2.800   3rd Qu.:22.20
 Max.   :2015-12-31 00:00:00   Max.   :55.900   Max.   :35.60
                                                NA's   :5
    temp_min           wind          weather
 Min.   : 1.100   Min.   :0.400   Length:1461
 1st Qu.: 5.600   1st Qu.:2.200   Class :character
 Median : 8.900   Median :3.000   Mode  :character
 Mean   : 9.055   Mean   :3.241
 3rd Qu.:12.800   3rd Qu.:4.000
 Max.   :18.300   Max.   :9.500
 NA's   :116
```

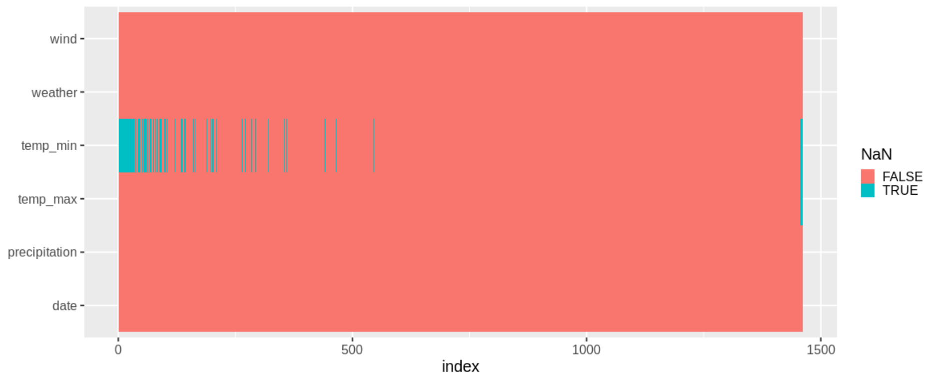
We use the same strategy to visualize NaNs in ggplot as in Altair where we first pivot_longer the dataframe:

```
%%R -w 1200
theme_set(theme_gray(base_size = 20))

df |>
    map_df(is.na) |>
    mutate(index = row_number()) |>
    pivot_longer(-index) |>
    ggplot(aes(x = index, y = name, fill = value)) +
        geom_raster() +
        labs(y = '', fill = 'NaN')
```



Sorted by max temp:

```
%%R -w 1200
theme_set(theme_gray(base_size = 20))

df |>
    arrange(temp_max) |>
    map_df(is.na) |>
    mutate(index = row_number()) |>
    pivot_longer(-index) |>
    ggplot(aes(x = index, y = name, fill = value)) +
        geom_raster() +
        labs(y = '', fill = 'NaN')
```



4.2. Repeating the same plots for multiple dataframe columns

4.2.1. Py

Previously we have made subplots via faceting, which creates one subplot per unique value in a categorical column and displays the same numerical columns in all the subplots/facets. This view of multiple subsets in the data is often called a trellis plot or plot of small multiples. Here, we will see how we can create subplots that each display all the data points, but is repeated for different columns in the data.

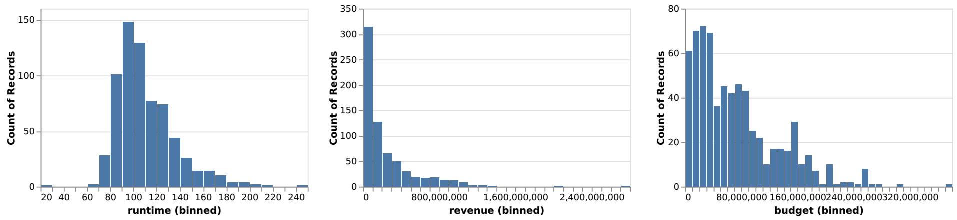
We specify which columns we want to use via the `.repeat` method, and where we want to use them via `alt.repeat`. For this plot, we keep the y-axis constant, and repeat the plot for different x-axis columns.

```
url = 'https://raw.githubusercontent.com/joelostblom/teaching-datasets/main/movies.json'
movies = pd.read_json(url)[['runtime', 'budget', 'revenue', 'genres', 'countries']]
movies
```

	runtime	budget	revenue	genres	countries
0	100	94000000	940335536	[Animation]	[United States of America]
1	143	140000000	655011224	[Fantasy]	[United States of America]
2	87	75000000	527068851	[Animation]	[United States of America]
3	151	200000000	1065659812	[Fantasy]	[United States of America]
4	106	70000000	101371017	[Fantasy]	[United States of America]
...	...	...	...	...	...
675	107	100000000	519876949	[History]	[United Kingdom, United States of America]
676	86	50000000	66913939	[Animation]	[United States of America]
677	100	22000000	194647323	[Fantasy]	[United States of America]
678	72	3500000	3775000	[Animation]	[United States of America]
679	109	10000000	4073489	[History]	[United Kingdom, United States of America]

680 rows × 5 columns

```
# You could extract the columns like this, but we will write them out for clarity here
# numeric_cols = movies.select_dtypes('number').columns.tolist()
alt.Chart(movies).mark_bar().encode(
     alt.X(alt.repeat(), type='quantitative', bin=alt.Bin(maxbins=40)),
     y='count()',
).properties(
    width=300,
    height=200
).repeat(
    ['runtime', 'revenue', 'budget']
)
```



4.2.2. R

There is not “repeat” function in ggplot, so we would first need to pivot all data frame variables into the same column and then use faceting on this new column name.

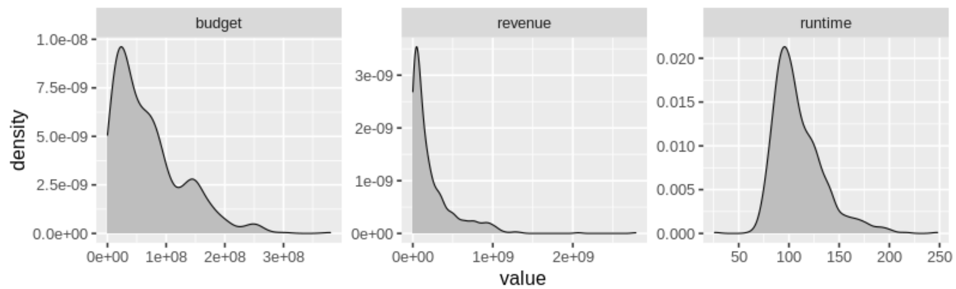
```
%%R

library(rjson)
library(tidyverse)
theme_set(theme_gray(base_size = 16))


url <- 'https://raw.githubusercontent.com/joelostblom/teaching-datasets
/main/movies.json'
movies <- fromJSON(file = url) %>%
    as_tibble() %>%
    select(runtime, revenue, budget, genres, countries) %>%
    unnest(-c(countries, genres))
glimpse(movies)
```

```
Rows: 680
Columns: 5
$ runtime   <dbl> 100, 143, 87, 151, 106, 119, 115, 178, 179, 201, 132, 101, 1…
$ revenue   <dbl> 940335536, 655011224, 527068851, 1065659812, 101371017, 7627…
$ budget    <dbl> 9.40e+07, 1.40e+08, 7.50e+07, 2.00e+08, 7.00e+07, 2.50e+07, …
$ genres    <named list> "Animation", "Fantasy", "Animation", "Fantasy", "Fant…
$ countries <named list> "United States of America", "United States of America…
```

```
%%R

movies_long <- movies %>%
    select_if(is.numeric) %>%
    pivot_longer(everything())  # To get only some columns we can write c(revenue,
budget)

movies_long
```

```
# A tibble: 2,040 × 2
   name        value
   <chr>       <dbl>
 1 runtime       100
 2 revenue 940335536
 3 budget   94000000
 4 runtime       143
 5 revenue 655011224
 6 budget  140000000
 7 runtime        87
 8 revenue 527068851
 9 budget   75000000
10 runtime       151
# … with 2,030 more rows
# ℹ Use `print(n = ...)` to see more rows
```

```
%%R -w 800 -h 250

movies_long %>%
    ggplot(aes(x = value)) +
        geom_density(fill = 'grey') +
        facet_wrap(~name, scales = 'free')
```

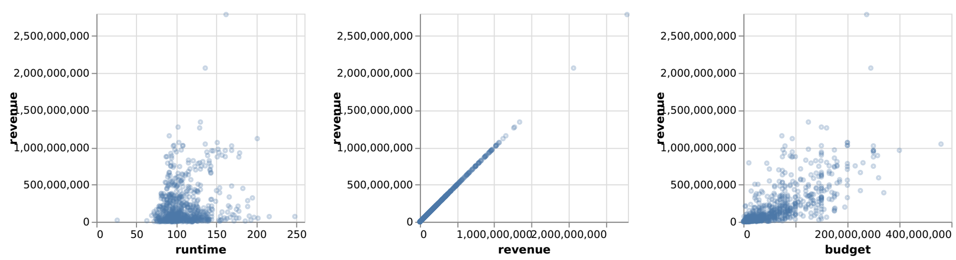


4.2.3. Repeating on two axes

Py

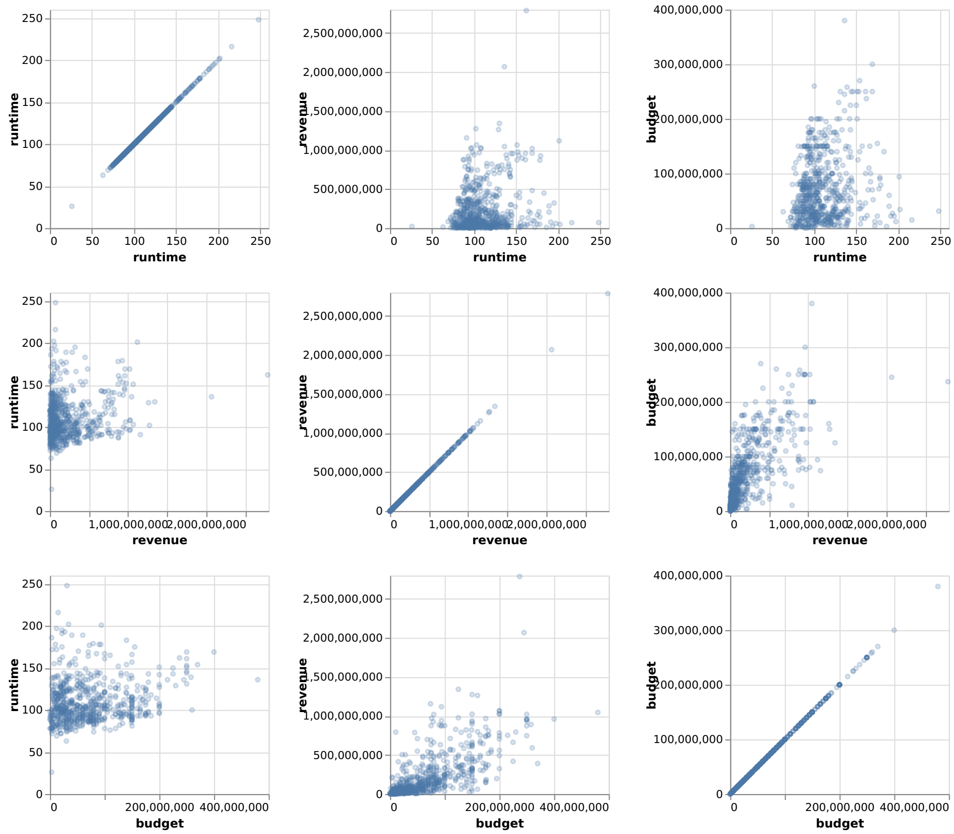
To efficiently repeat over both the axis, we can use the `column` and `row` parameter to `repeat`. These work similarly as they do for `facet`, distributing the list of data frame variables over only one axis, which we specify with `alt.repeat` to the x and y axis.

```
alt.Chart(movies).mark_point(opacity=0.3, size=10).encode(
     alt.X(alt.repeat(), type='quantitative'),
     y='revenue',
).properties(
    width=200,
    height=200
).repeat(
    ['runtime', 'revenue', 'budget']
)
```

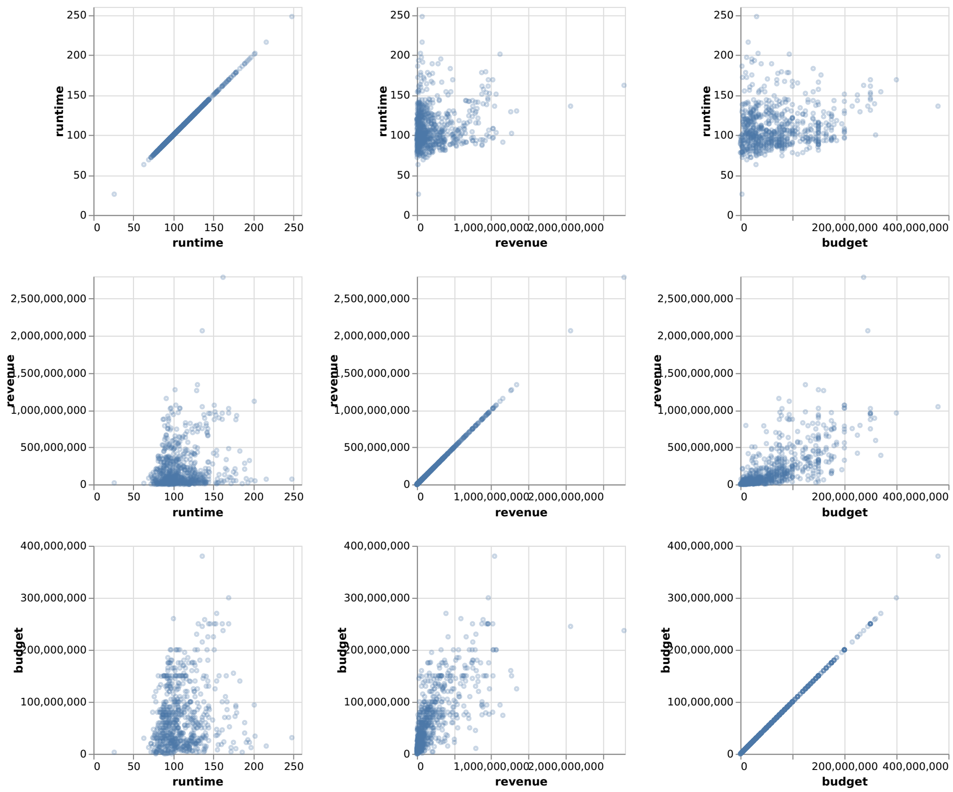


When we don't give any named parameters to `repeat` and `alt.repeat`, they use the default, which is `repeat` and then you can wrap with `columns=2` if you want more than one row. We could use this technique to create a what is usually called a scatter plot matrix (or pairplot), where we repeat numerical columns on both axes to investigate their pairwise relationships.

```
alt.Chart(movies).mark_point(opacity=0.3, size=10).encode(
     alt.X(alt.repeat('row'), type='quantitative'),
     alt.Y(alt.repeat('column'), type='quantitative')
).properties(
    width=200,
    height=200
).repeat(
    column=['runtime', 'revenue', 'budget'],
    row=['runtime', 'revenue', 'budget']
)
```




```
alt.Chart(movies).mark_point(opacity=0.3, size=10).encode(
     alt.X(alt.repeat('column'), type='quantitative'),
     alt.Y(alt.repeat('row'), type='quantitative')
).properties(
    width=200,
    height=200
).repeat(
    row=['runtime', 'revenue', 'budget'],
    column=['runtime', 'revenue', 'budget']
)
```

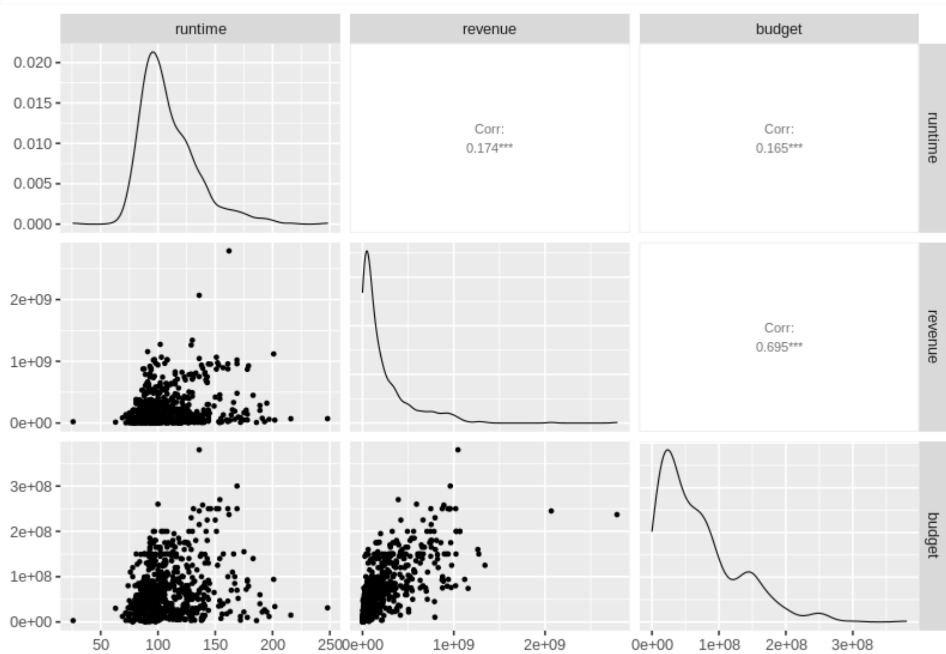


R

This would not work if both axes are numerical, but luckiy there is an extension packagge for ggplot called GGally, which can be used here. It also plots the density curves on the diagonal and the correlations on top.

```
%%R -w 800 -h 550
library(GGally)
GGally::ggpairs(movies %>% select_if(is.numeric), progress = FALSE)
```

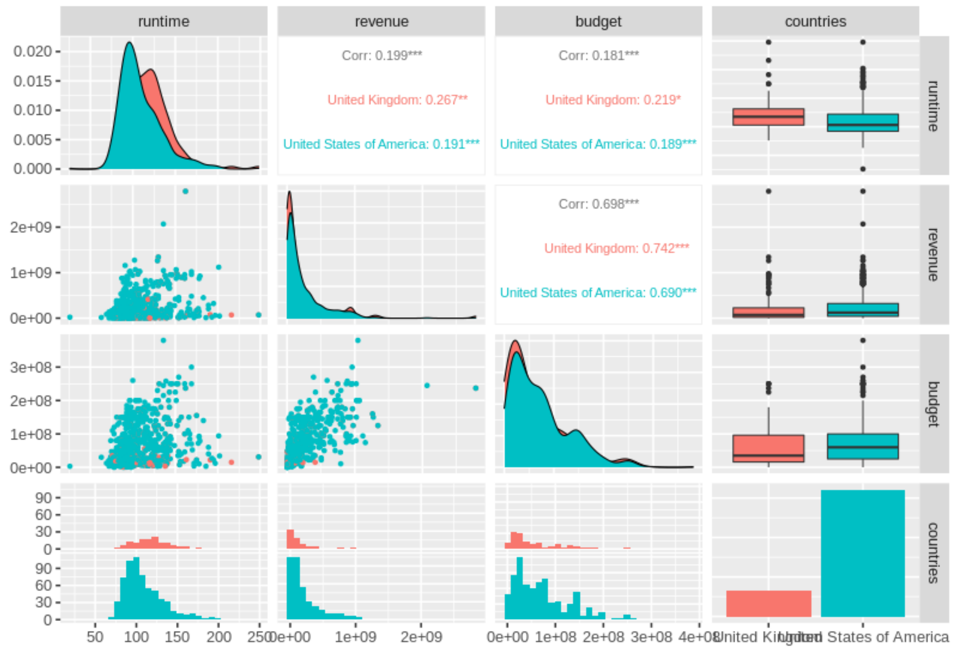
```
R[write to console]: Registered S3 method overwritten by 'GGally':
  method from
  +.gg   ggplot2
```



This function can also show info on categorical variables.

```
%%R -w 800 -h 550

free_countries <- movies %>% unnest(countries) %>% select(!genres)
GGally::ggpairs(free_countries, aes(color = countries), progress=FALSE)
```

```
R[write to console]: `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

R[write to console]: `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

R[write to console]: `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Plotting multiple categories is also possible and will include counts of the combinations of categories as in the next plot. Customizing these plots is a bit cumbersome, [this blog post has some good examples](), and you can see one below too.

```
%%R -w 800 -h 550

free_both <- movies %>% unnest(genres) %>% unnest(countries)
GGally::ggpairs(
    free_both,
    aes(color = countries),
    progress = FALSE,
    lower = list(continuous = GGally::wrap('points', alpha = 0.3, size=0.1, color =
'steelblue')),
    diag = list(continuous = GGally::wrap('densityDiag', fill = 'coral', alpha = 0.4)))
```

```
R[write to console]: `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

R[write to console]: `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

R[write to console]: `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

R[write to console]: `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

R[write to console]: `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

R[write to console]: `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

4.3. Correlation plots

4.3.1. Py

A good complement to a scatter plot matrix is a correlation plot which helps formalize the correlation between numerical variables. Using “spearman” correlation instead of “pearson” allows us to detect non-linear relationships better.

This can be done in pandas via the styler (requires matplotlib).

```
movies.corr().style.background_gradient()
```

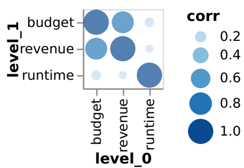
	runtime	budget	revenue
runtime	1.000000	0.165387	0.174192
budget	0.165387	1.000000	0.694802
revenue	0.174192	0.694802	1.000000

We could do the same in Altair but it would require some more work

```
corr_df = (
    movies
    .corr('spearman')
    .abs()                      # Use abs for negative correlation to stand out
    .stack()                    # Get df into long format for altair
    .reset_index(name='corr'))  # Name the index that is reset to avoid name collision
corr_df
```

	level_0	level_1	corr
0	runtime	runtime	1.000000
1	runtime	budget	0.147367
2	runtime	revenue	0.106208
3	budget	runtime	0.147367
4	budget	budget	1.000000
5	budget	revenue	0.758968
6	revenue	runtime	0.106208
7	revenue	budget	0.758968
8	revenue	revenue	1.000000

```
alt.Chart(corr_df).mark_circle().encode(
    x='level_0',
    y='level_1',
    size='corr',
    color='corr')
```

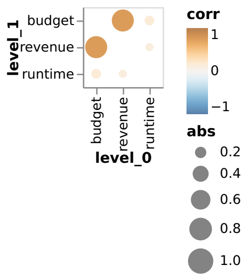


With some effort, we can make it look nicer and remove the correlation on the diagonal since that is just with the same column.

```
corr_df =
movies.select_dtypes('number').corr('spearman').stack().reset_index(name='corr')
corr_df.loc[corr_df['corr'] == 1, 'corr'] = 0  # Remove diagonal
corr_df['abs'] = corr_df['corr'].abs()
corr_df
```

	level_0	level_1	corr	abs
0	runtime	runtime	0.000000	0.000000
1	runtime	budget	0.147367	0.147367
2	runtime	revenue	0.106208	0.106208
3	budget	runtime	0.147367	0.147367
4	budget	budget	0.000000	0.000000
5	budget	revenue	0.758968	0.758968
6	revenue	runtime	0.106208	0.106208
7	revenue	budget	0.758968	0.758968
8	revenue	revenue	0.000000	0.000000

```
alt.Chart(corr_df).mark_circle().encode(
    x='level_0',
    y='level_1',
    size=alt.Size('abs', scale=alt.Scale(domain=(0, 1))),
    color=alt.Color('corr', scale=alt.Scale(scheme='blueorange', domain=(-1, 1)))
)
```

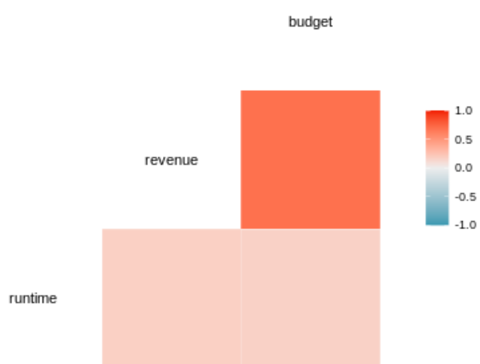


The domain argument sets the extent of the colorscale (from a perfect negative to a perfect positive correlation), we will talk more about colorscale later in the course.

4.3.2. R

We could do the separate calculation in ggplot as well, but there is also a special function in GGally, which does the computation and visualizes it as a heatmap.

```
%%R -w 600 -h 350
GGally::ggcorr(movies)
```

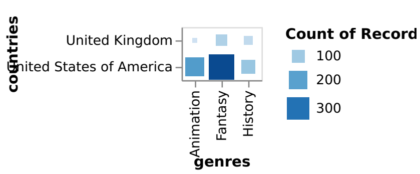


4.4. Counting combinations of categorical groups

4.4.1. Py

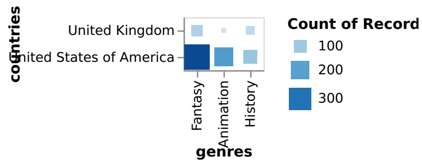
Counting categoricals is helpful to get and overview of where most observations lie. Often the size and color can be used for this, similar to the correlation plot.

```
boom_both = movies.explode('countries').explode('genres')
alt.Chart(boom_both).mark_square().encode(
    x='genres',
    y='countries',
    color='count()',
    size='count()')
```



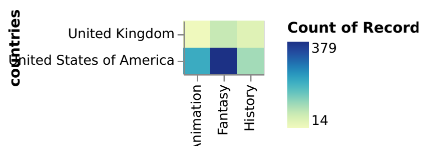
As with bar charts of counts, it is often helpful to sort these categorical plot according to which combinations have the most counts. Here we can sort based on the `color` or `size` encoding, just as we have sorted based on `x` and `y` before (the effect of this sorting is more notable in larger charts).

```
boom_both = movies.explode('countries').explode('genres')
alt.Chart(boom_both).mark_square().encode(
    x=alt.X('genres', sort='-color'),
    y=alt.Y('countries', sort='color'),
    color='count()',
    size='count()')
```



A heatmap could be created as well, but it can be a little harder to see the difference here sine we are not using size for the count, just color variations.

```
alt.Chart(boom_both).mark_rect().encode(
    x='genres',
    y='countries',
    color='count()')
```



By Joel Ostblom