Messy_data_exploration

```
rm(list=ls())
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to beco
```

```
anscombe_quartet = readRDS("anscombe quartet.rds")
```

Task 1

Summarise the data demographic

```
str(anscombe_quartet)
```

```
tibble [44 x 3] (S3: tbl_df/tbl/data.frame)
 $ dataset: chr [1:44] "dataset_1" "dataset_1" "dataset_1" "dataset_1" ...
 $ x      : num [1:44] 10 8 13 9 11 14 6 4 12 7 ...
 $ y      : num [1:44] 8.04 6.95 7.58 8.81 8.33 ...
```

```r
anscombe_quartet %>%
  group_by(dataset) %>%
  summarise(
    mean_x    = mean(x),
    mean_y    = mean(y),
    min_x     = min(x),
    min_y     = min(y),
    max_x     = max(x),
    max_y     = max(y),
    crrltn    = cor(x, y)
  )
```

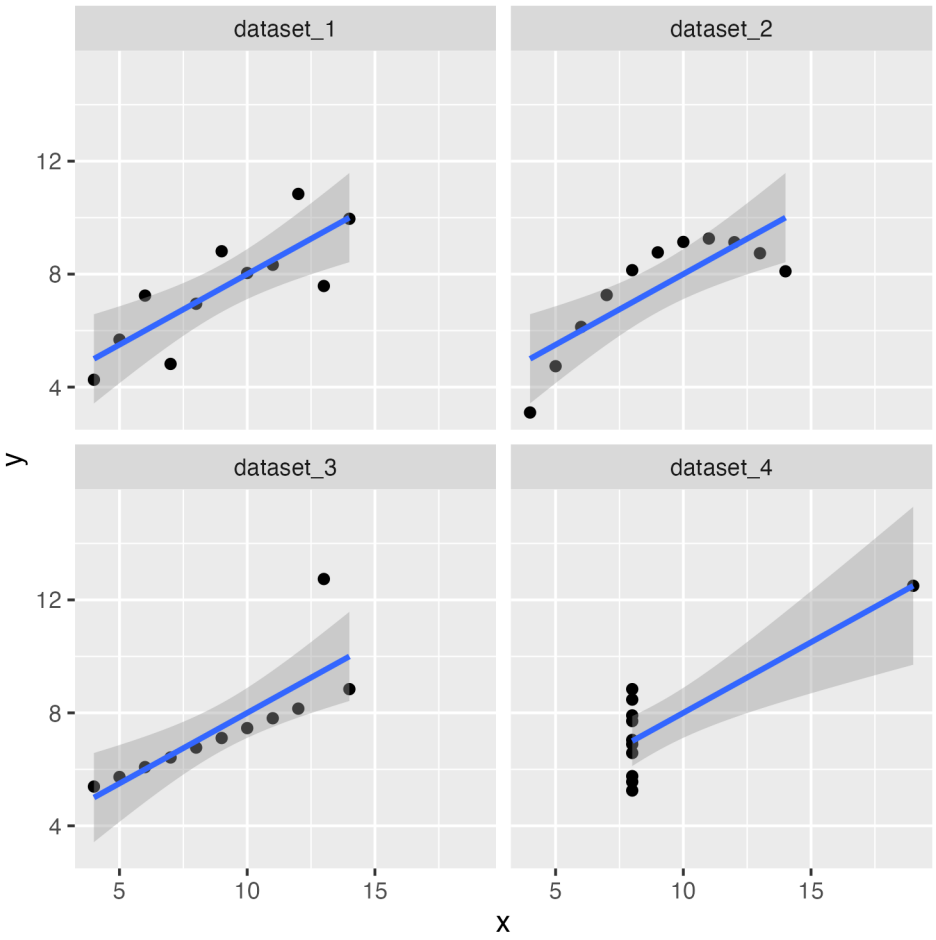
```
# A tibble: 4 x 8
  dataset   mean_x mean_y min_x min_y max_x max_y crrltn
  <chr>      <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl>  <dbl>
1 dataset_1      9   7.50     4  4.26    14 10.8   0.816
2 dataset_2      9   7.50     4  3.1     14  9.26  0.816
3 dataset_3      9   7.5      4  5.39    14 12.7   0.816
4 dataset_4      9   7.50     8  5.25    19 12.5   0.817
```

The four data sets have the similar correlation coefficient, but have diferrent min and max.

Now I want to plot the data:

```r
library(ggplot2)
T1 <- ggplot(anscombe_quartet, aes(x=x,y=y)) +
  geom_point() +
  geom_smooth(method = "lm",formula = "y ~ x") +
  facet_wrap(~dataset)
ggsave("T1_plot.png", plot = T1, width = 5, height = 5, dpi = 300)
```

The output image is saved as T1_plot and inserted below



Although they have similar regression coefficient, they are very different datasets. The first data set is more linear regression. The second is like a projection. Third one has an outlier. Fourth one has an outlier around 20 and the rest is around 6. Only the first dataset, or the third dataset excluding the outlier can be modeled appropriately with regression. Hence the summary table might be misleading.

Task 2

now I want to load in the second dataset

```
datasaurus_dozen = readRDS("datasaurus_dozen.rds")
```

I want to explore the different dataset.

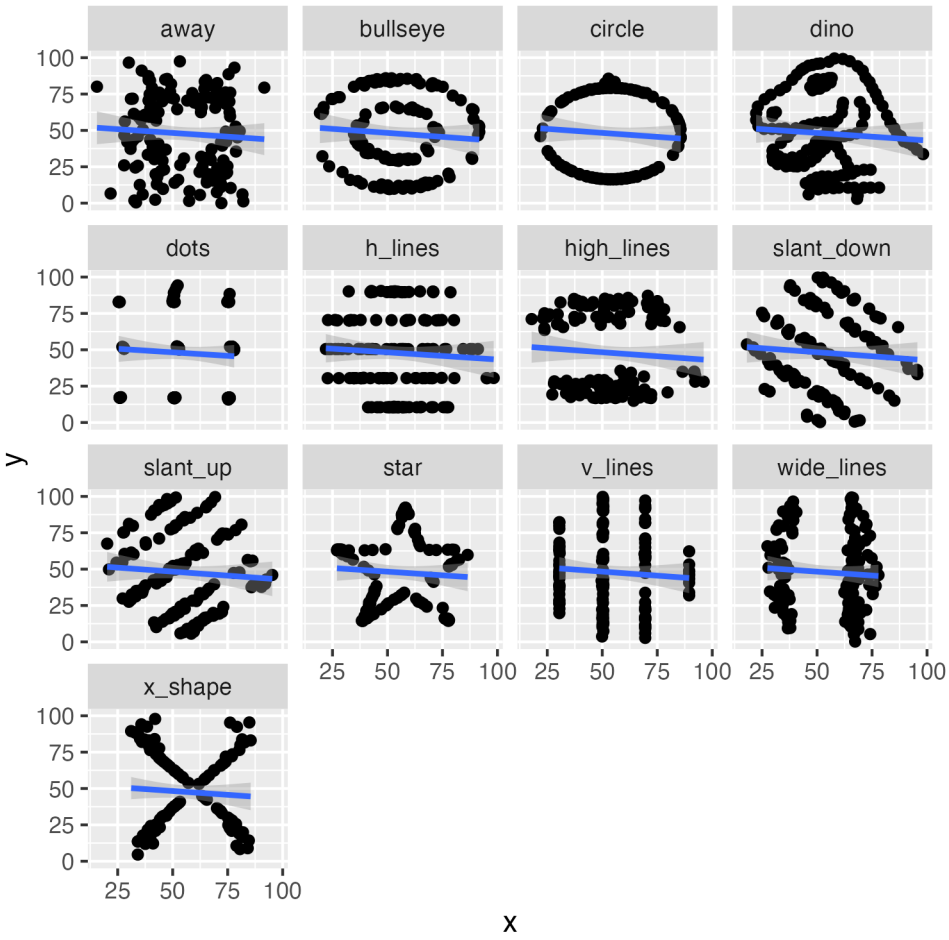
```r
result <- datasaurus_dozen %>%
  group_by(dataset) %>%
  summarise(
    mean_x    = mean(x),
    mean_y    = mean(y),
    min_x     = min(x),
    min_y     = min(y),
    max_x     = max(x),
    max_y     = max(y),
    crrltn    = cor(x, y)
  )

knitr::kable(result, digits = 2, caption = "T2 Summary Statistics")
```

Table 1: T2 Summary Statistics

dataset	mean_x	mean_y	min_x	min_y	max_x	max_y	crrltn
away	54.27	47.83	15.56	0.02	91.64	97.48	-0.06
bullseye	54.27	47.83	19.29	9.69	91.74	85.88	-0.07
circle	54.27	47.84	21.86	16.33	85.66	85.58	-0.07
dino	54.26	47.83	22.31	2.95	98.21	99.49	-0.06
dots	54.26	47.84	25.44	15.77	77.95	94.25	-0.06
h_lines	54.26	47.83	22.00	10.46	98.29	90.46	-0.06
high_lines	54.27	47.84	17.89	14.91	96.08	87.15	-0.07
slant_down	54.27	47.84	18.11	0.30	95.59	99.64	-0.07
slant_up	54.27	47.83	20.21	5.65	95.26	99.58	-0.07
star	54.27	47.84	27.02	14.37	86.44	92.21	-0.06
v_lines	54.27	47.84	30.45	2.73	89.50	99.69	-0.07
wide_lines	54.27	47.83	27.44	0.22	77.92	99.28	-0.07
x_shape	54.26	47.84	31.11	4.58	85.45	97.84	-0.07

The correlation is very similar, as shown in the table

```r
T2 <- ggplot(datasaurus_dozen, aes(x=x,y=y)) +
 geom_point() +
 geom_smooth(method = "lm",formula = "y ~ x") +
 facet_wrap(~dataset)
ggsave("T2_plot.png", plot = T2, width = 5, height = 5, dpi = 300)
```


away
bullseye
circle
dino
dots
h_lines
high_lines
slant_down
slant_up
star
v_lines
wide_lines
x_shape
y
x
0
25
50
75
100