

AUDIT DATA ANALYTICS

Easy Fast Powerful

Stewart Li

2023-08-08

https://github.com/stewartli/datong_uni

WHY SHOULD YOU CARE?



https://github.com/stewartdl/datong_uuu

DATA

```
1 df_r <- readxl::read_excel(here::here('raw/Financial_Sample.xlsx')) %>%
2 janitor::clean_names() %>%
3 select(-c(month_number:last_col()))
4 head(df_r[, 1:5], 2)
```

```
# A tibble: 2 x 5
  segment    country product discount_band units_sold
  <chr>      <chr>   <chr>        <dbl>
1 Government Canada Carretera     1618.
2 Government Germany Carretera    1321.
```

```
1 df_py = r.df_r
2 df_py.head(2)
```

```
segment    country product ...   cogs   profit   date
0 Government Canada Carretera ... 16185.0 16185.0 2014-01-01
1 Government Germany Carretera ... 13210.0 13210.0 2014-01-01
```

```
[2 rows x 13 columns]
```

```
1 conn <- DBI::dbConnect(duckdb::duckdb(), 'finsample.duckdb')
2 DBI::dbWriteTable(conn, "finsample", df_r)
3 DBI::dbListTables(conn)
4 DBI::dbGetQuery(conn, "SELECT * FROM finsample LIMIT 3;")
5 DBI::dbDisconnect(conn)
```

```
1 cp finsample.duckdb /c/Users/Stewart_Li/Desktop/
2 mv /mnt/c/Users/Stewart_Li/Desktop/finsample.duckdb .
3
4 ./duckdb
5 .open finsample.duckdb
6 .tables
7 from finsample limit 3;
```

https://github.com/stewartri/datong_uni

ESY

File	Home	Insert	Paged Layout	Formulas	Data	Review	View	Developer	Q	Tell me	A _z	?
C11	:	x	v	f _k	=IF(C7>1100,100,0)+IF(C8>12,100,0)+IF(C9>=C7,100,0)+IF(C10>700,100,0)+IF(C6-C5<35,100,0)+IF(AND(C7>1100,C8>12,C9==C7,C10>700,C6-C5<35),250,0)							
1	Agent's monthly bonus calculations											
2	Real estate agent											
3	Tenants											
4	Property address											
5	Date property placed on market											
6	Move-in date per lease agreement											
7	Monthly rent amount											
8	Length of rental agreement (in months)											
9	Deposit amount											
10	Tenant's credit rating											
11												
12												

```

1 df %>%
2   group_by(discount_band) %>%
3   count(country, segment, product) %>%
4   mutate(pct = n / sum(n)) # A tibble: 372 × 6
# Groups: discount_band [4]
#> #> #> #> #> #>
#>   discount_band country segment product <chr> <dbl> <int> pct
#>   <dbl> <chr> <chr> <dbl> <dbl>
#> 1 1 High Canada Channel Partners Amarilla 2 0.00816
#> 2 2 High Canada Channel Partners Carretera 1 0.00408
#> 3 3 High Canada Channel Partners Paseo 2 0.00816
#> 4 4 High Canada Channel Partners VTT 1 0.00408
#> 5 5 High Canada Enterprise Amarilla 1 0.00408
#> 6 6 High Canada Enterprise Carretera 1 0.00408
#> 7 7 High Canada Enterprise Paseo 1 0.00408
#> 8 8 High Canada Enterprise VTT 3 0.0122
#> 9 9 High Canada Enterprise Velo 1 0.00408
#> 10 10 High Canada Government Amarilla 3 0.0122
#> # i 362 more rows

```

https://github.com/stewartri/datong_uni

FAST

Excel can handle 1,048,576 rows
and 16,384 columns at a time.

```
1 skimr::skim(df_r)
2 table1:=table1(~sale_price + discounts | segment, data = df_r)

1 df_r %>%
2 group_by(units_sold_cut = cut(units_sold, 3)) %>%
3 summarise(across(is.numeric), tibble::lst(min, max, me

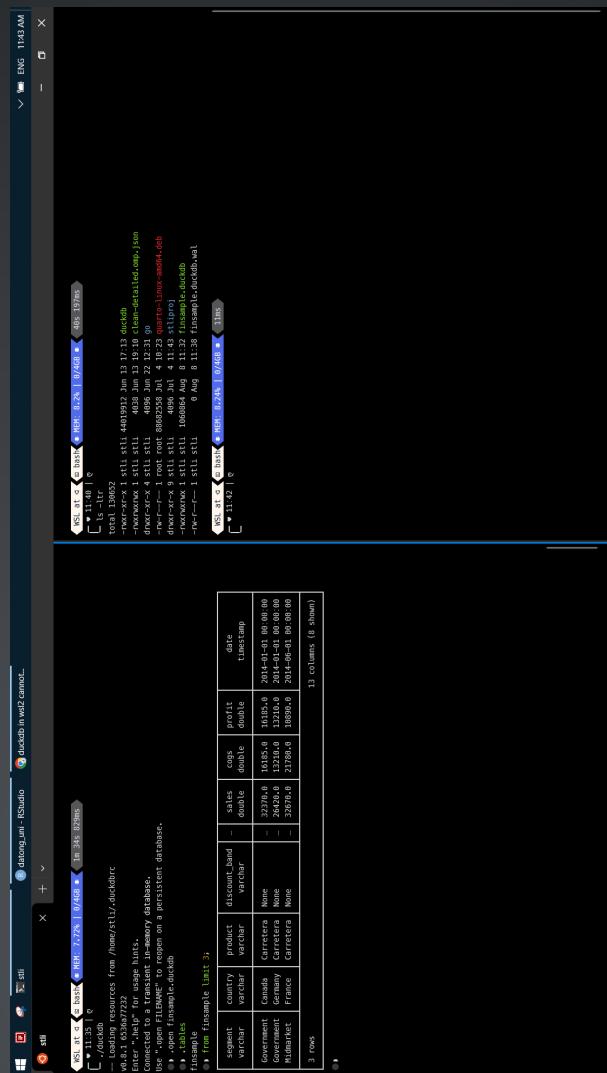
# A tibble: 3 × 25
units_sold_cut      units_sold_min units_sold_max
<dbl>                <dbl>          <dbl>
1 (196,1.63e+03]    200            1630.
2 (1.63e+03,3.06e+03] 1642           2996
2281.
3 (3.06e+03,4.5e+03] 3165           4492.
3747.

# i 21 more variables: manufacturing_price_min <dbl>,
#   manufacturing_price_max <dbl>, manufacturing_price_mean
<dbl>,
#   sale_price_min <dbl>, sale_price_max <dbl>, sale_price_mean
<dbl>,
#   gross_sales_min <dbl>, gross_sales_max <dbl>,
gross_sales_mean <dbl>,
#   discounts_min <dbl>, discounts_max <dbl>, discounts_mean
<dbl>,
#   sales_min <dbl>, sales_max <dbl>, sales_mean <dbl>, cogs_min
<dbl>,
```

POWERFUL

You need learn VBA, M,.Net, and
C# to extend Excel.

R Python



https://github.com/stewartli/datong_uni

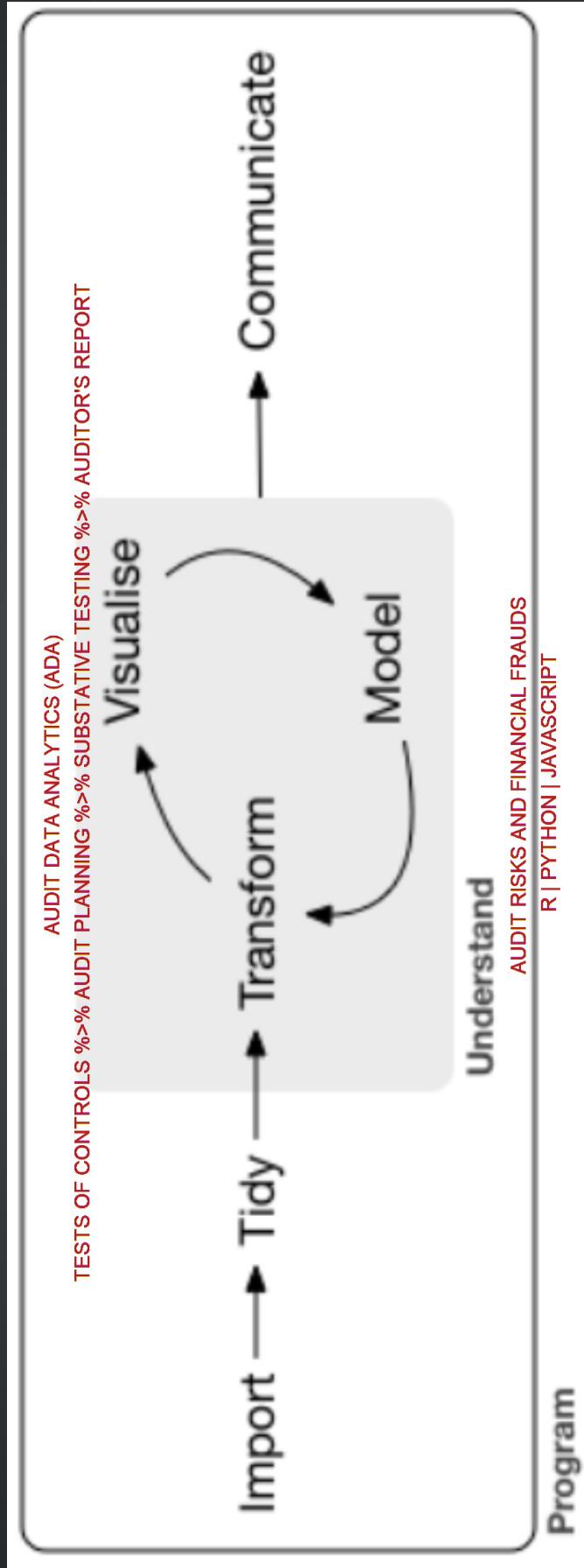
Audit workflow

https://github.com/stewartli/datong_uni

Audit workflow

https://github.com/stewartli/datong_uni

AUDIT WORKFLOW



- Cross-check: re-calculation of data.
- EDA: overview of data set.
- Test: analytical procedures.
- Report: tables, etc.

https://github.com/stewartli/datong_uni

CROSS-CHECK

```
1 df %>%
2   mutate(check_gross_sales = units_sold * sale_price - gross_sales,
3   check_profit = round(units_sold * sale_price - discounts - cogs - profit, 2) ) %>%
4   summarise(across(c(check_gross_sales, check_profit), sum))
```

```
# A tibble: 1 × 2
#>   check_gross_sales check_profit
#>       <dbl>        <dbl>
#> 1           0            0
```

EDA

```
1 df %>%
2   group_by(week(date)) %>%
3   summarise(across(where(is.numeric), tibble::lst(min, max, mean)) )

# A tibble: 12 x 25
`week(date)` units_sold_min units_sold_max units_sold_mean
<dbl>          <dbl>          <dbl>          <dbl>
1            1        384        4251        1938.
2            5        260        2844        1575.
3            9        259        2993        1526.
4           13        510        4492.       2254.
5           18        200        2918        1479.
6           22        448        2907        1476.
7           26        367        4026        1981.
8           31        381        2832        1734.
9           35        218        2993        1541.
10          40        214        2996        1436.
11          44        263        2954        1730.
12          48        266        2852        1479.

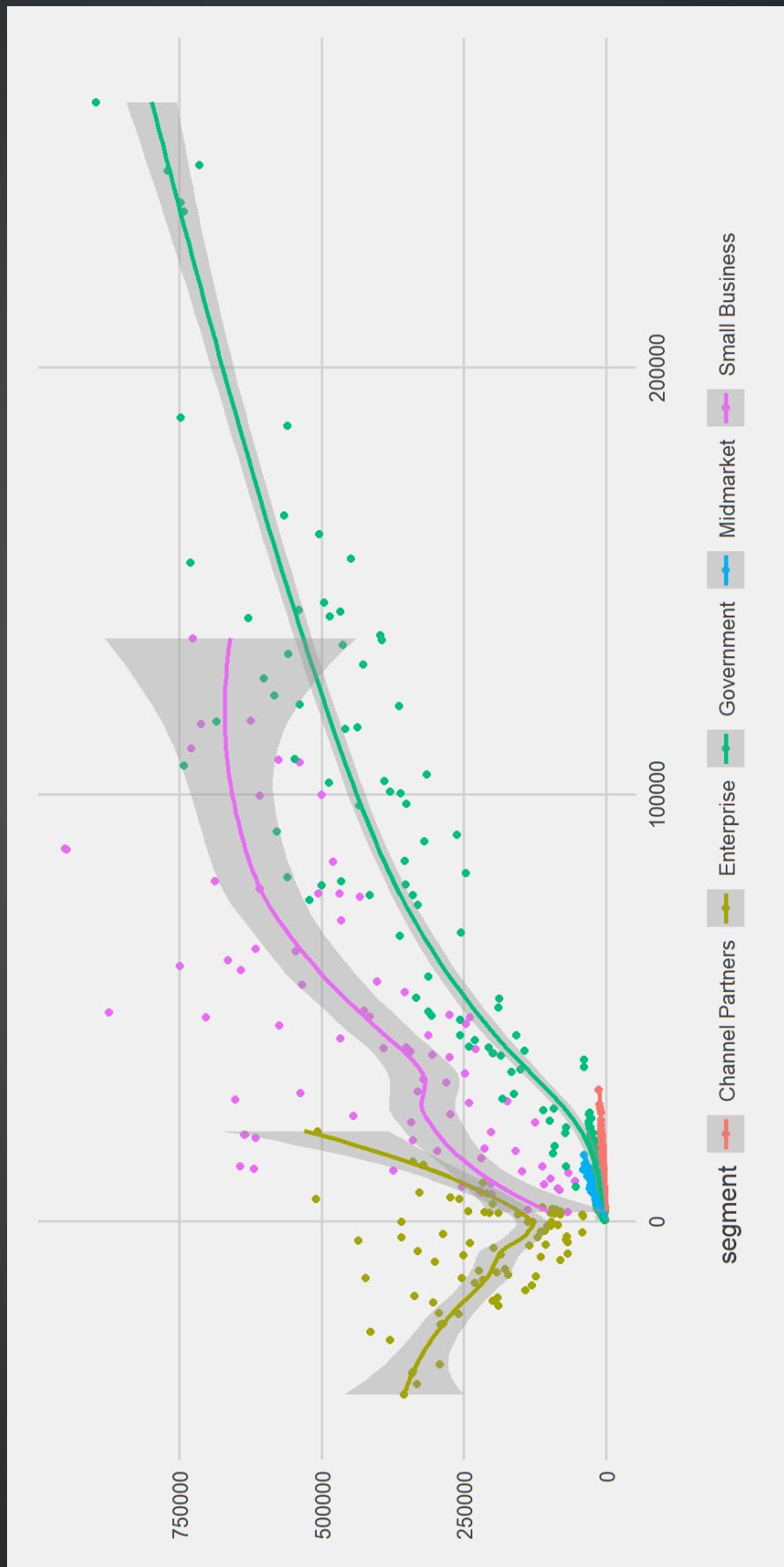
# i 21 more variables: manufacturing_price_min <dbl>,
# manufacturing_price_max <dbl>, manufacturing_price_mean <dbl>,
# sale_price_min <dbl>, sale_price_max <dbl>, sale_price_mean <dbl>,
# gross_sales_min <dbl>, gross_sales_max <dbl>, gross_sales_mean <dbl>,
# discounts_min <dbl>, discounts_max <dbl>, discounts_mean <dbl>,
# sales_min <dbl>, sales_max <dbl>, sales_mean <dbl>, cogs_min <dbl>,
# cogs_max <dbl>, cogs_mean <dbl>, profit_min <dbl>, profit_max <dbl>, ...
```

https://github.com/stewartli/datong_uni

TREND

```

1 df_r %>%
2   ggplot(aes(profit, cogs, color = segment)) +
3     geom_point() + geom_smooth() +
4     ggthemes::theme_fivethirtyeight()
  
```

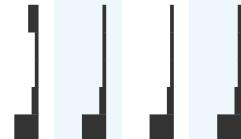
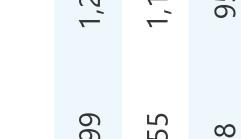
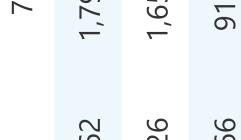


https://github.com/stewartli/datong_uni

REPORT

```
1 myskimr <- skimr::skim_with(numeric = skimr::sfl(p25 = NULL, p50 = NULL, p75 = NULL))
2 myskimr(select(df_r, where(is.numeric) & dplyr::matches("sale|^c")) ) %>%
3 select(-skim_type) %>%
4 mutate(across(where(is.numeric), formatable::comma, digits = 0)) %>%
5 gt() %>%
6 opt_stylize(style = 6, color = "blue", add_row_striping = TRUE)
```

```
skim_variable n_missing complete_rate numeric.mean numeric.sd numeric.p0 numeric.p100 numeric.hist
```

sale_price	0	1	118	137	7	350	
gross_sales	0	1	182,759	254,262	1,799	1,207,500	
sales	0	1	169,609	236,726	1,655	1,159,200	
cogs	0	1	145,475	203,866	918	950,625	

CONTACT ME  

https://github.com/stewartli/datong_uni

https://github.com/stewartli/datong_uni