# Use Skimr for Data Quality
## Exploratory Data Analysis

```
-- Data Summary ------------------------
                           Values
Name                       Piped data
Number of rows             574
Number of columns          6
_____________________
Column type frequency:
  Date                     1
  numeric                  5
_____________________
Group variables            None

-- Variable type: Date -----------------------------------------------
# A tibble: 1 x 7
  skim_variable n_missing complete_rate min        max
* <chr>              <int>         <dbl> <date>     <date>
1 date                   0             1 1967-07-01 2015-04-01
  median     n_unique
* <date>        <int>
1 1991-05-16      574

-- Variable type: numeric --------------------------------------------
# A tibble: 5 x 11
  skim_variable n_missing complete_rate      mean       sd       p0
* <chr>              <int>         <dbl>     <dbl>    <dbl>    <dbl>
1 pce                    0             1   4820.    3557.      507.
2 pop                    0             1 257160.   36682.   198712
3 psavert                0             1      8.57     2.96      2.2
4 uempmed                0             1      8.61     4.11      4
5 unemploy               0             1   7771.    2642.     2685
      p25      p50      p75     p100 hist
*   <dbl>    <dbl>    <dbl>    <dbl> <chr>
1   1578.    3937.    7626.   12194.
2 224896   253060   290291.  320402.
3      6.4      8.4     11.1     17.3
4      6        7.5      9.1     25.2
5   6284     7494     8686.   15352
>
```

The Data Quality Report from skimr

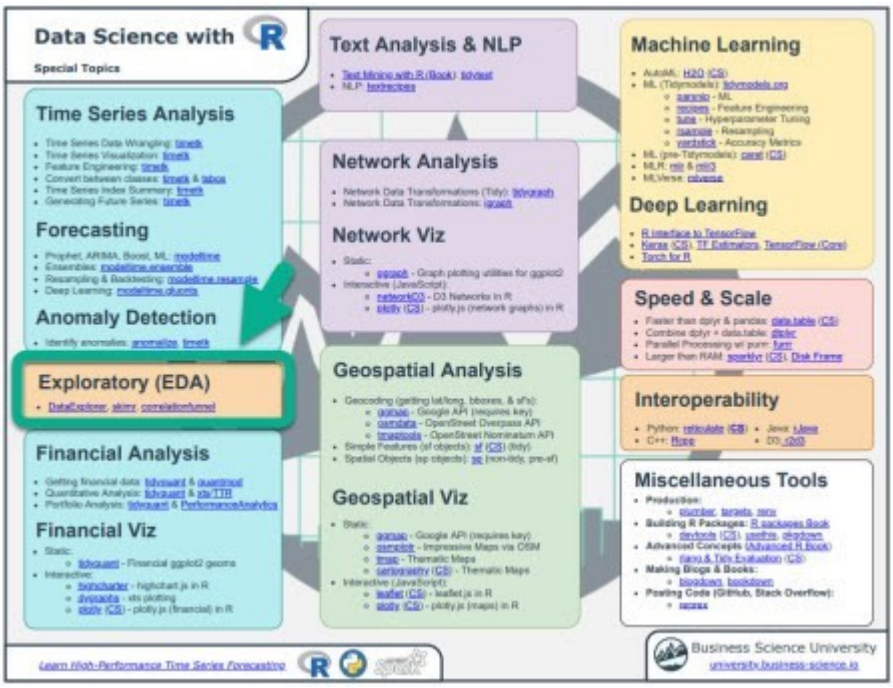
# Rapid Data Quality Checks in R
## Automatic Data Quality Reporting

**Data Scientists spend 80% of their time** understanding data, exploring it, wrangling and preparing for analysis.

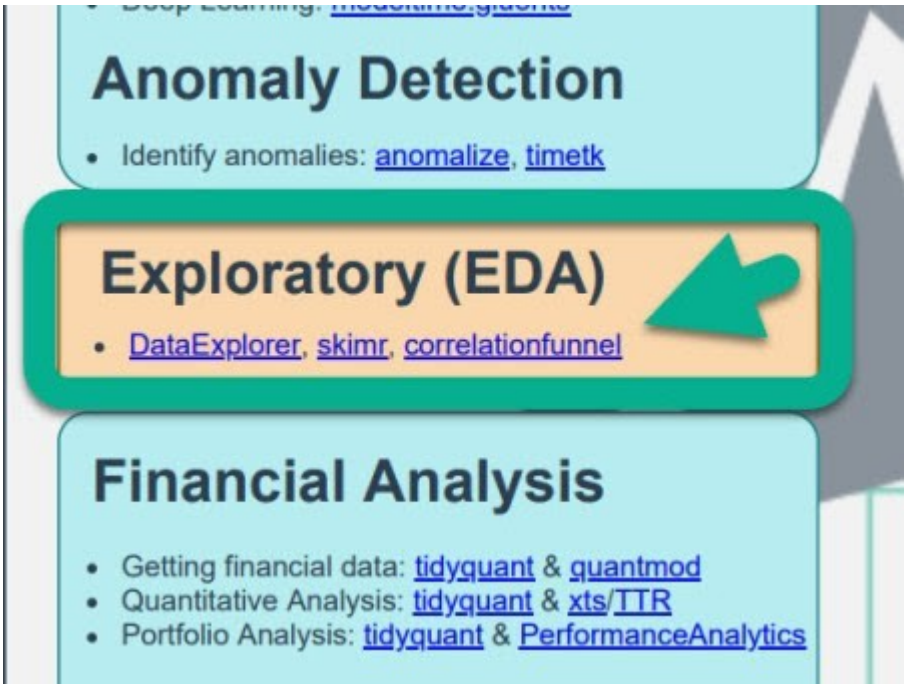
This is way too long!

**We can speed this up.** One tool I use in EVERY SINGLE DATA PROJECT is called skimr. It's my go-to.

PRO TIP: I've added **links to skimr and two more SUPER-IMPORTANT R PACKAGES FOR EDA** on Page 3 of my Ultimate R Cheatsheet. 👇




You can use my Ultimate R Cheatsheet to help you learn R. It consolidates the most important R packages (ones I use every day) into one cheatsheet. Here's where skimr is located.




## How Skimr Works
### Automatic Data Quality Reporting

One of the coolest features of Skimr is the ability to **create a Data Quality Report in 1 line of code**. This automates:

- Date Profiling
- Works with Numeric, Categorical, Text, Date, Nested List Columns, and even Dplyr Groups

Ultimately, this saves the Data Scientist SO MUCH TIME. ⏳

# Missing Data, Categorical & Numeric Reporting (Starwars)

The “starwars” data set has a 87 starwars characters with various attributes. This is a messy data set containing a lot of missing values and nested list-columns.

```
21
22 # * Starwars / Missing Data & Lists ----
23 starwars %>% skim()
```

**Overall Data Summary**
Number of Rows/Columns, Data Types by Column, Group Variables.

```
-- Data Summary ------------------------
                           Values
Name                       Piped data
Number of rows             87
Number of columns          14
___________________________
Column type frequency:
  character                8
  list                     3
  numeric                  3
___________________________
Group variables            None
```

**Character Summaries**
Missing / completion rate, number of unique observations, and text features.

```
-- Variable type: character ------------------------------------------------------------
# A tibble: 8 x 8
  skim_variable n_missing complete_rate   min   max empty n_unique whitespace
* <chr>            <int>         <dbl> <int> <int> <int>    <int>      <int>
1 name                 0         1         3    21     0       87          0
2 hair_color           5         0.943     4    13     0       12          0
3 skin_color           0         1         3    19     0       31          0
4 eye_color            0         1         3    13     0       15          0
5 sex                  4         0.954     4    14     0        4          0
6 gender               4         0.954     8     9     0        2          0
7 homeworld           10         0.885     4    14     0       48          0
8 species              4         0.954     3    14     0       37          0
```

**List Summaries (nested column)**
Number of unique elements in each list.

```
-- Variable type: list -----------------------------------------------------------------
# A tibble: 3 x 6
  skim_variable n_missing complete_rate n_unique min_length max_length
* <chr>            <int>         <dbl>    <int>      <int>      <int>
1 films                0             1       24          1          7
2 vehicles             0             1       11          0          2
3 starships            0             1       17          0          5
```

**Numeric Summaries**
Missing/completion rates and distributions.

```
-- Variable type: numeric --------------------------------------------------------------
# A tibble: 3 x 11
  skim_variable n_missing complete_rate  mean    sd    p0   p25   p50   p75  p100 hist
* <chr>            <int>         <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <chr>
1 height               6         0.931 174.   34.8    66 167     180 191     264 ▁▁▅▇▂
2 mass                28         0.678  97.3 169.     15  55.6    79  84.5  1358 ▇▁▁▁▁
3 birth_year          44         0.494  87.6 155.      8  35      52  72     896 ▇▁▁▁▁
```

# Time Series Reporting (Economics)

The “economics” data set has a date feature called “Date” and several numeric features. We’ll focus on the date feature.

```
# * Economics / Time Series (Date) ----
economics %>% skim()
```

**Date Summaries**
Missing/completion rates, min/max dates, and the number of unique dates.

```
-- Variable type: Date -------------------------------------------------------------------
# A tibble: 1 x 7
  skim_variable n_missing complete_rate min        max        median     n_unique
* <chr>            <int>         <dbl> <date>     <date>     <date>        <int>
1 date                 0             1 1967-07-01 2015-04-01 1991-05-16      574
```

# Grouped Time Series Reporting (Economics Long)

The “economics_long” data set has been pivoted so each time series from “economics” is stacked on top of each other – perfect for a groupwise skim analysis.

```
# * Economics Long / Grouped Data ----
economics_long %>% group_by(variable) %>% skim()
```

**Grouped Date Summaries**
Each of these are provided by group: Missing/completion rates, min/max dates, and the number of unique dates.

```
-- Data Summary ------------------------
                           Values
Name                       Piped data
Number of rows             2870
Number of columns          4
_____________________
Column type frequency:
  Date                     1
  numeric                  2
_____________________
Group variables            variable

-- Variable type: Date -----------------------------------------------------------------
# A tibble: 5 x 8
  skim_variable variable n_missing complete_rate min        max        median     n_unique
* <chr>         <chr>        <int>         <dbl> <date>     <date>     <date>        <int>
1 date          pce              0             1 1967-07-01 2015-04-01 1991-05-16      574
2 date          pop              0             1 1967-07-01 2015-04-01 1991-05-16      574
3 date          psavert          0             1 1967-07-01 2015-04-01 1991-05-16      574
4 date          uempmed          0             1 1967-07-01 2015-04-01 1991-05-16      574
5 date          unemploy         0             1 1967-07-01 2015-04-01 1991-05-16      574
```