 BONBUTTERY 

Data Exploration with dplyr

Hands-on Output 1: How many columns do the dataset have?

37 columns

Code:

```
> ncol(df)
[1] 37
```

Hands-on Output 2: How many non-numerical variables do the dataset have?

3 variables (County ID, State, and County)

Hands-on Output 3: What is the difference between observations and variables?

Normally, variables are those encoded in columns and the observations are those inputted in rows. Variables are considered to contain values that measure attributes that are similar (e.g. Age, County, Total Population). On the other hand, observations are those that contain values that are measured on the same unit such as an Individual (person), a city, etc. across all the tributes (aka the variables) (Weber, 2021).

Hands-on Output 4: Which State and County has the lowest population?

Texas and Loving County with a Total Population of 74

Code:

```
> df_state_county_pop<-df%>%select(State, County, TotalPop)
> df_state_county_pop%>%arrange(TotalPop)
```

Hands-on Output 5: What is the ratio of those that drive to work and walk? Create a new column called **Drive/Walk** and arrange the output in descending order based on the new column. Show the code.

```
> df_drive_walk
    Drive Walk Ratio_Drive_Walk
1    66.3  0.0              Inf
2    89.0  0.0              Inf
3    82.6  0.0              Inf
4    75.1  0.0              Inf
5    89.1  0.0              Inf
6    83.7  0.0              Inf
7    82.3  0.0              Inf
8    79.8  0.0              Inf
9    80.0  0.0              Inf
10   78.6  0.0              Inf
11   84.1  0.0              Inf
12   78.2  0.0              Inf
13   96.1  0.0              Inf
14   94.2  0.0              Inf
15   84.9  0.1        849.00000
16   84.4  0.1        844.00000
17   84.3  0.1        843.00000
18   81.4  0.1        814.00000
19   80.7  0.1        807.00000
```

Code:

> df_drive_walk <- df %>% select(Drive, Walk)

> df_drive_walk <- df %>% select(Drive, Walk) %>% mutate(Ratio_Drive_Walk = Drive/Walk)

> df_drive_walk <- df %>% select(Drive, Walk) %>% mutate(Ratio_Drive_Walk = Drive/Walk) %>% arrange(desc(Ratio_Drive_Walk))

Exercises:

Answer the following questions and show the code.

1. Which state has the highest income per capita?

New York has the highest Income per capita

Code:

> df_state_incomepercap <- df %>% select(State, IncomePerCap)

> df_state_incomepercap <- df %>% select(State, IncomePerCap) %>% arrange(desc(IncomePerCap))

2. In the State of Ohio, which county has the highest percentage of Asians?

Delaware County of Ohio has the Highest Percentage of Asian with 5.5%

Code:

> df_state_asian <- df %>% select(State, County, Asian)

> df_state_asian <- df %>% select(State, County, Asian) %>% filter(State == "Ohio")

> df_state_asian <- df %>% select(State, County, Asian) %>% filter(State == "Ohio") %>% arrange(desc(Asian))

3. In States that have Drive greater than 60%, which State and County has the lowest income recorded?

Puerto Rico and Adjuntas Municipio has the LOWEST Income recorded with a total of 11680

Variables:

Drive >60%, State, County, Income (lowest)

Code:

> df_state_county_drive_income <- df %>% select(State, County, Drive, Income)

> df_state_county_drive_income <- df %>% select(State, County, Drive, Income) %>% filter(Drive>60)

> df_state_county_drive_income <- df %>% select(State, County, Drive, Income) %>% filter(Drive>60) %>% arrange(Income)

4. Create a new column that calculates the ratio of White from Black population. Store the results in a variable and the other columns are limited to State, County, TotalPop, White, Black.

	State	County	TotalPop	White	Black	Ratio_White_Black
1	Alabama	Autauga County	55036	75.4	18.9	3.9894180
2	Alabama	Baldwin County	203360	83.1	9.5	8.7473684
3	Alabama	Barbour County	26201	45.7	47.8	0.9560669
4	Alabama	Bibb County	22580	74.6	22.0	3.3909091
5	Alabama	Blount County	57667	87.4	1.5	58.2666667
6	Alabama	Bullock County	10478	21.6	75.6	0.2857143
7	Alabama	Butler County	20126	52.2	44.7	1.1677852
8	Alabama	Calhoun County	115527	72.7	20.4	3.5637255
9	Alabama	Chambers County	33895	56.2	39.3	1.4300254
10	Alabama	Cherokee County	25855	91.8	5.0	18.3600000
11	Alabama	Chilton County	43805	80.4	9.5	8.4631579
12	Alabama	Choctaw County	13188	56.3	42.1	1.3372922
13	Alabama	Clarke County	24625	53.0	45.7	1.1597374
14	Alabama	Clay County	13407	80.2	14.7	5.4557823
15	Alabama	Cleburne County	14939	92.7	2.8	33.1071429
16	Alabama	Coffee County	51073	71.0	17.1	4.1520468
17	Alabama	Colbert County	54435	78.8	15.9	4.9559748
18	Alabama	Conecuh County	12649	50.3	46.3	1.0863931
19	Alabama	Coosa County	10955	65.3	33.2	1.9668675

Variables:

White, Black, State, County, Totalpop

Code:

> df_white_black <- df %>% select(State, County, TotalPop, White, Black)

> df_white_black <- df %>% select(State, County, TotalPop, White, Black) %>% mutate(Ratio_White_Black = White/Black)

REFERENCES

Weber. (2021, April 25). *What are variables and observations in a data set? – Machine learning future*. BRAINBI | The Analytics Bot. https://www.brainbi.dev/2021/04/25/what-are-variables-and-observations-in-a-data-set-machine-learning-future/